MESCAL

Management of End-to-end Quality of Service
Across the Internet at Large

IST-2001-37961

D1.3: Final specification of protocols and algorithms for inter-domain SLS management and traffic engineering for QoS-based IP service delivery

Document Identifier: MESCAL/WP1/UniS/D1.3/final	
Deliverable Type: Report **Deliverable Nature:** Public	**Contractual Date:** 30 June 2005 **Actual Date:** 30 June 2005

Editor:	Ning Wang, University of Surrey
Authors:	*FTR&D:* P. Morand, M. Boucadair, T. Coadic, P. Levis *TRT:* R. Egan, H. Asgari *UCL:* D. Griffin, J. Griem, J. Spencer *UniS:* M. Howarth, N. Wang, S. Georgoulas, K.H. Ho, P. Flegkas, P. Trimintzios, G. Pavlou *Algo:* P. Georgatsos, E. Mykoniati, I. Liabotis, T. Damilatis
Abstract:	This standalone document is the final WP1 deliverable of the MESCAL project, and is the result of activities AC1.1-1.3. It illustrates the MESCAL business model, functional architecture, and also specifies enhanced algorithms and protocols that enable inter-domain Quality of Service (QoS) across the Internet, including in particular: • Algorithms and protocols that enable SLS establishment between peers and invocation of service instances across domains; • Offline inter-domain and intra-domain traffic engineering algorithms; • QoS enhancements to BGP (q-BGP) for dynamic inter-domain traffic engineering; • Path Computation Server (PCS) and its use for inter-domain QoS constrained path computation; • Traffic enforcement and inter-domain monitoring; • Multicast SLSs and traffic engineering algorithms.
Keywords:	Inter-domain, Quality of Service (QoS), Meta-QoS classes, SLS, traffic engineering, q-BGP, PCS, admission control

Copyright © MESCAL Consortium:

France Telecom Research and Development	FTR&D	Co-ordinator	France
Thales Research and Technology	TRT	Principal Contractor	UK
University College London	UCL	Principal Contractor	UK
The University of Surrey	UniS	Principal Contractor	UK
Algonet SA	Algo	Principal Contractor	Greece

Project funded by the European Community under the
"Information Society Technology" Programme (1998-2002)

Executive Summary

This standalone document is the final WP1 deliverable of the MESCAL project, and it includes all the results from activities AC1.1-1.3. The overall objective of MESCAL is to propose and validate scalable, incremental solutions that enable the flexible deployment and delivery of inter-domain Quality of Service (QoS) across the Internet. The project has validated its results through prototypes, and evaluated the overall performance through simulations and prototype testing. This document specifies the MESCAL business model, functional architecture and the supporting algorithms and protocols that enable inter-domain Quality of Service (QoS) across the Internet. These supporting algorithms, mechanisms and protocols are summarised as follows:

- The Meta-QoS class (m-QC) concept;
- Algorithms and protocols that enable Service Level Specification (SLS) establishment (including ordering and order handling) between peers;
- Algorithms that integrate inter- and intra- domain SLS management with traffic engineering, defining the data that needs to be passed between the SLS handling functional blocks, traffic forecast, and traffic engineering components;
- Algorithms for online SLS invocation handling (i.e. inter-domain admission control);
- Algorithms for inter-domain traffic forecast;
- Offline inter-domain and intra-domain traffic engineering algorithms. Two inter-domain provisioning cycles are described: a longer-timescale cycle in which pSLS requirements are determined by the traffic engineering algorithms and then negotiated with peer domains, and a shorter-term cycle in which inter-domain bandwidth is invoked within the framework of existing pSLSs;
- Dynamic inter-domain traffic engineering algorithms and protocols, including QoS enhancements to BGP (q-BGP);
- Path Computation Server (PCS) and its supporting communications protocol (PCP) for inter-domain MPLS traffic engineering;
- Inter-domain traffic monitoring and enforcement;
- Multicast SLS definitions and multicast traffic engineering algorithms.

In describing the algorithms and protocols, the structure of the document reflects the highest-level view of the MESCAL functional architecture. Each functional block of the final functional architecture is decomposed, and its interfaces and behavioural specification described. The final experimentation results from simulation and tested that correspond to the above algorithms/protocols are provided in D3.2 [D3.2].

Copyright © MESCAL Consortium, June 2005

Table of Contents

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

Detailed Table of Contents

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

List of Figures

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

Copyright © MESCAL Consortium, June 2005

List of Tables

Copyright © MESCAL Consortium, June 2005

1 INTRODUCTION

1.1 Background

This is the final document that has been produced as part of Work Package 1 of the EU IST MESCAL project. The overall objective of MESCAL is to propose and validate scalable, incremental solutions that enable the flexible deployment and delivery of inter-domain Quality of Service (QoS) across the Internet. MESCAL will validate its results through prototypes, and evaluate the overall performance through simulations and prototype testing.

MESCAL adopts a phased approach and the technical work is split into three work packages (WPs):

- WP1, *Specification of Functional Architecture, Algorithms and Protocols,* is responsible for defining business models and the generic, multi-domain, multi-service IP QoS functional architecture for inter-domain QoS delivery. Based on these models WP1 will develop algorithms and protocols for negotiation and establishment of inter-domain service level specifications (SLSs), and will enhance and extend inter-domain traffic engineering (TE) mechanisms and routing protocols, including the required interactions with intra-domain functionality. WP1 will also define algorithm test requirements. Based on implementation experience and experimental results fed back from WP2 and WP3, later activities within WP1 will validate the initial specifications and derive enhancements as appropriate.

- WP2, *System Design and Implementation,* is responsible for undertaking basic enhancements of experimental Linux-based routers and developing simulation tools to model the general inter-domain and QoS requirements of the project. Based on the specifications from WP1, WP2 will specify the engineering approach, conduct detailed implementation design and finally implement both testbed prototypes and simulation environments.

- WP3, *Integration, Validation and Experimentation,* is responsible for configuring the required experimentation infrastructure and for conducting validation and performance evaluation activities on the prototypes and simulators developed by WP2 according to the test requirements identified by WP1. Experimentation will be executed both in the MESCAL testbed (with the support of extended development environments at other partners' premises) and using the simulators.

1.2 Role of WP1 and this deliverable

WP1, *Specification of Functional Architecture, Algorithms and Protocols*, comprises three activities. In the first, AC1.1, *Inter-domain Business Models and System Architecture,* business models have been defined and an overall functional architecture for inter-domain QoS-based services has been developed, starting from the requirements, assumptions and state of the art in this area. The second activity, AC1.2, *Algorithm and Protocol Specification,* starts from the functional architecture produced in AC1.1 and will specify algorithms and protocols for: peer SLS establishment and invocation of service instances across domains; QoS enhancements to BGP; consideration of alternative, novel approaches (e.g. link state-based); integrated inter- and intra-domain SLS management and traffic engineering; multicast SLSs and traffic engineering; impact of IPv6 on traffic engineering possibilities; and information models, algorithms and protocols for an overall policy-driven system approach. The third activity, AC1.3, *Enhancements to Algorithms and Protocol Specifications*, will produce modifications and enhancements to the AC1.2 algorithms and specifications, based on feedback from simulation and implementation experience in WP2 and WP3.

This document includes all the results of activity AC1.3, *Enhancements to Algorithms and Protocol Specifications*. The document includes the following principal components:

- Final MESCAL business model and functional architecture description;

Copyright © MESCAL Consortium, June 2005

- Final description on service planning and QoS capabilities exchange;
- Final version of algorithms and protocols that enable SLS establishment between peers and invocation of service instances across domains;
- Final version of offline inter-domain and intra-domain traffic engineering algorithms;
- Final analysis of Integrated and Decoupled traffic engineering approaches (inter-relationships between Inter-domain and Intra-domain TE);
- Final specification of q-BGP and PCS;
- Newly added solutions for inter-domain monitoring;
- Final description on multicast SLSs and traffic engineering algorithms, in addition to unicast capabilities.

1.3 Structure of this document

The rest of this document is thus structured as follows:

- Section 2, *The MESCAL business model*, defines the principal actors in QoS-based service delivery across multiple domains. In particular, the model defines the terms "customer" (the target recipient of QoS services) and "provider" (entities responsible for the offering and provisioning of QoS-based services).
- Section 3, *Assumptions and requirements*, documents the requirements for QoS-related services from the perspectives of both the customers and providers defined in Section 2. The requirements are drawn from current business practices and market needs as understood by the project partners. The customer requirements cover QoS characteristics, subscription, invocation, verification, and multicast requirements. The provider requirements cover QoS extension across multiple domains, efficient path discovery and negotiation, verification, scalability, resilience, incremental deployment, ease of deployment, accounting, and multicast requirements.
- Section 4, *The MESCAL QoS service model (definitions)*, presents the MESCAL model for Internet QoS-based services and defines the terms used in the model. It includes the specification of appropriate notions, entities and the relationships and associations between them, which are thought pertinent to the issue of definition and provisioning of QoS-based services in the Internet, across multiple Provider domains. The MESCAL model defined in this Section extends the model used in TEQUILA so as to cover QoS-based services that span multiple autonomous systems (ASs), rather than the domain of a single Internet Service Provider.
- Section 5, *Inter-domain QoS Issues*, identifies and discusses a number of issues related to Inter-domain QoS delivery, focusing in particular on inter-domain peering arrangements, service guarantees, traffic engineering, scalability and multicast. The MESCAL consortium partners have considered these issues during the process of developing both the MESCAL functional architecture and the solution options that implement QoS delivery in accordance with this functional architecture.
- Section 6, *The MESCAL Functional Architecture*, defines the functional architecture that will be used in MESCAL. The functional architecture consists of five top-level blocks: service planning and QoS capabilities exchange, traffic engineering, SLS management, traffic enforcement, and monitoring and assurance. These blocks are further decomposed in the Section.

Copyright © MESCAL Consortium, June 2005

- Section 7, *Solution Space*, defines three QoS-based service options, which respectively provide Loose, Statistical and Hard QoS guarantees. The Loose service option enables a provider to offer customers access to differentiated transport services. The Statistical service option provides customers access to inter-domain QoS services with firmer guarantees than the Loose option, based primarily on qualitative guarantees. The Hard service option provides customers with inter-domain QoS services with strict performance guarantees based on quantitative levels. Section 7 then proceeds to describe three solution options that provide QoS-based services, each of these solution options corresponding to a service option, and is in accordance with the MESCAL functional architecture.

- Section 8, *Service Planning and QoS Capabilities Exchange*, provides a decomposition of each functional block in the Service Planning functional group. The Section describes updated approaches for defining service offerings (including the planning of l-QCs and e-QCs), and for providing traffic demand estimates to the Traffic Forecast functional block. The Section also describes the mechanisms by which a provider advertises QoS capabilities to peer providers and in turn discovers their capabilities; the structure of these advertisements is also defined.

- Section 9, *SLS Management*, provides a decomposition of each functional block in the SLS Management functional group. The Section begins by defining the formats of the Service Subscription Specification (SSS) and SLS for Inter-domain unicast and multicast traffic. The Section then proceeds to present algorithms and protocols for pSLS ordering and order handling. This includes the definition of an enhanced Inter-domain Service Negotiation Protocol (SrNP), based on the Intra-domain SrNP originally defined in [TEQUI, D1.4]. The Section then presents algorithms for dynamic invocation and invocation handling (i.e. admission control) for real-time traffic.

- Section 10, *Traffic Engineering*, provides a decomposition of each functional block in the TE functional group, including the detailed interfaces with other functional blocks. The Section starts by defining new terminology relevant to Inter-domain TE, and by discussing the interactions between Inter-domain and Intra-domain TE including the relationship between resource provisioning cycles for Inter- and Intra-domain TE. The Section then proceeds to describe how traffic forecasting supports the integration of Inter-domain SLS management with inter-domain TE as a component of the resource provisioning cycle. It then proceeds to present novel offline traffic engineering algorithms (both Inter-domain and Intra-domain). The Section then presents dynamic Inter-domain TE algorithms and protocols, discussing issues in QoS enhancements to BGP (q-BGP) and describing proposed modifications to the protocol. The final specification of Path Computation Server (PCS) Communication Protocol (PCP) is also proposed.

- Section 11, *Traffic Enforcement*, describes changes in the Data Plane functional blocks that are a consequence of the Inter-domain QoS functions described in Sections 3-5. The areas covered include QC enforcement (classification and traffic conditioning), IP/MPLS forwarding (principally the influence of q-BGP on the forwarding information base (FIB) and its interaction with the routing information base (RIB)), and PHB enforcement.

- Section 12, *Inter-domain monitoring*, provides the proposed monitoring system across multiple domains, which assists in: (1) verifying whether the QoS performance guarantees committed in c/pSLSs are in fact being met, and (2) traffic optimisation according to short-to-medium term changes as well as providing measurement information for long-term planning in order to optimise network usage and avoid undesirable conditions.

- Section 13, *Multicast*, integrates the multicast components of the MESCAL functional architecture, describing the functions and interfaces of each functional block. Mechanisms/ algorithms for multicast traffic engineering and service differentiation are described.

- Finally, the appendix includes the three Internet drafts for PCE related specifications.

Copyright © MESCAL Consortium, June 2005

2 MESCAL BUSINESS MODEL

2.1 Introduction

An initial specification of the MESCAL business model was presented in deliverable D1.1. This was subject to refinement as the problem area and its solutions were studied in more detail during the specification phase of the algorithms and protocols for the components that constitute the overall system architecture. This Section presents the final version of the MESCAL business model incorporating the refinements identified during the detailed design, implementation and experimental work of the project.

The business model assumed by MESCAL is illustrated in Figure 1. The model depicts from the perspectives of MESCAL the stakeholders involved in the chain of QoS-based service delivery in the Internet.

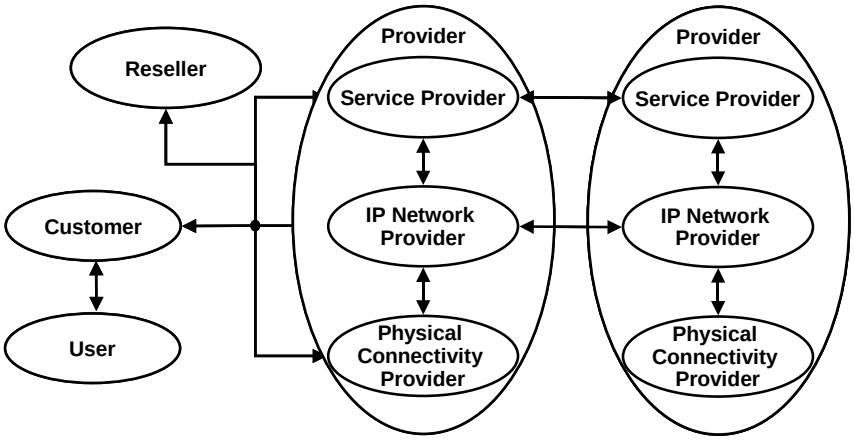


Figure 1. The MESCAL business model from D1.1

The broad classes of business relationships described by this model are those identified between the different entities involved in interdomain QoS delivery. The focus of the MESCAL project is on the interactions between the set of IP Network Providers (ISPs) involved in the end-to-end delivery of QoS-based IP services, i.e. across multiple domains. A large number of ISPs can be involved in the provision of global IP connectivity services. The necessary business relationships and roles between the set of IP Network Providers is analysed in some detail in deliverable D1.4 [D1.4]. This is summarised in the following subsections as is the definition of the role of the other stakeholders that form the overall MESCAL business model.

2.2 Customers and users

A *'Customer'* (subscriber) denotes an entity, which has the legal ability to subscribe to QoS-based services offered by *'Providers'*. *'Customers'* are the target recipients of QoS-based services. They interact with *'Providers'* (or *'Resellers'*, see below) following a customer-provider paradigm, with the purpose to 'buy' services to meet their communication needs and requirements.

A *'User'* is an entity (human being or a process from a general perspective), which has been named by a *'Customer'* and appropriately identified by a *'Provider'* for actually requesting/accessing and using the QoS-based services bought by the *'Customer'*. The use of the services should be in line with the terms and conditions agreed in the SLA between the *'Customer'* and the *'Provider'*. In essence, *'Users'* are the end-users of the services and they can only exist in association with a *'Customer'*.

Copyright © MESCAL Consortium, June 2005

2.3 IP network providers

'IP Network Providers' offer QoS-based plain IP connectivity services, that is services, which provide reachability between hosts in the IP address space. Such *'Providers'* must own and administer an IP network infrastructure. For connecting customers to their IP infrastructure, *'IP Network Providers'* may interact with separate *'Access Providers'* – a role which isn't explicitly covered in a separate entity in the business model but may be considered to be a provider with the physical connectivity provider role only. Alternatively, customers could be connected through means/facilities provided by the *'IP Network Providers'* themselves.

'IP Network Providers' may be differentiated according to the geographical span of their IP network infrastructure. As such, we may distinguish between small, medium and large *'IP Network Providers'*, with this distinction being relative (to a given area size) rather than absolutely defined. For example, considering a continental area, small, medium, large *'IP Network Providers'* may be thought as regional (covering specific cities of a country), national (covering a specific country), continental (covering specific countries of the continent) respectively.

Based on this distinction the current business model of the best effort Internet is built around a three tier hierarchy, with the business relationships between the providers being determined by their relative position in this hierarchy [HUST, D1.4]. In order to provide access to the global Internet, *'IP Network Providers'* must interact with each other; there cannot be a single provider offering global Internet coverage. Currently, in the best-effort Internet, there exist two forms of distinct relationships between *'IP Network Providers'* for traffic exchange, underlined by respective business agreements: *peering* and *transit*. Peering is termed as the business relationship, whereby *'IP Network Providers'* reciprocally provide only access to each other's customers. Peering is a non-transitive relationship. Peering is a mutual agreement between *'IP Network Providers'* to exchange data between themselves, normally for no fee or charge. Transit is the business relationship, whereby one transit provider provides access to all destinations in its routing table (could be global Internet) to another *'IP Network Provider'* for a charge. It should be clarified that the term 'peering' is also used in this document to denote that two providers interact with each other for the purpose of expanding the topological scope of their offered services, under any business relationship which may govern this interaction; it should be not be taken that this implies a specific peering business relationship as defined above.

The MESCAL solution adopts a *hop-by-hop, cascaded model* for the interactions between NPs both at the service and network (IP) layers. Service layer interactions result in the establishment of service agreements between NPs, *pSLSs* in MESCAL terminology, aggregating customer service traffic, which need to be supported by appropriate service management and traffic engineering capabilities per provider domain as well as by BGP-based interactions at the IP layer for QoS inter-domain routing purposes.

The type of inter-domain relationships and interactions impacts the service negotiation procedures, the required signalling protocols, the QoS binding, and path selection. The following approaches are considered in detail in deliverable D1.4 [D1.4]:

- The *centralised* approach where a Network Provider negotiates *pSLSs* directly with an appropriate number of downstream providers to construct an end-to-end QoS service. With this approach, service peers are not necessarily BGP peers.
- The *cascaded* approach where a NP only negotiates *pSLS*s with its immediate neighbouring provider/s to construct an end-to-end QoS service. With this approach, service peers are also BGP peers.
- The *hub* approach, which is similar to the centralised approach, where the Service Provider (SP), as a distinct entity from NP, is the central point that negotiates and establishes *pSLSs*.
- The *hybrid* approach, which is the mixture of centralised and cascaded approaches.

Copyright © MESCAL Consortium, June 2005

Within the MESCAL project, the first two major approaches have been considered for further study in order to construct end-to-end QoS-based services across the Internet at large scale. A single point of control for the service instances is the compelling feature of the centralised approach. The use of the centralised approach for more than a few interconnected NPs would be increasingly difficult to manage. Providers would prefer to offer services which reflect current Internet structure and for whom the use of the centralised approach would be inappropriate in many instances. Such providers would probably consider using the cascaded approach, which reflects the loosely coupled structure of Internet. Within the context of MESCAL, we focus on and provide solutions using the cascaded approach.

D1.4 concludes that the cascaded approach makes it possible to build IP QoS services on a global basis while only maintaining contractual relationships with adjacent operators. Hence, this approach is more scalable than the centralised approach.

Deliverable D1.4 also contains a chapter dealing further with business relationships and financial settlements between *IP Network Providers*. As service accounting, billing and marketing aspects are outside the scope of MESCAL, viability from business perspectives is addressed at the level of business relationships between NPs and related financial settlements for exchanging QoS traffic; accounting and data collection methods, charging, rating and pricing models are not addressed explicitly.

Two business cases have been identified for a MESCAL-enabled QoS-aware Internet; one for providing services based only on qualitative QoS guarantees and one for additionally providing services based on statically guaranteed quantitative QoS metrics. In both cases, services relying on hard QoS guarantees could also be provided, however not for the mass market because of scalability limitations inherent in the technical solution. The qualitative-QoS Internet business case directly corresponds to the three-tier, hierarchical model currently in place, whereas the statistical-QoS Internet business case advocates a flat Internet, where the business relationships between ISPs are of the same type, which is not affected nor dictated by the tier levels the ISPs may reside. In the flat Internet, the net flow of money always follows the flow of traffic. In the hierarchical Internet, assuming that a tier 1 ISP must always be involved, the net flow of money follows the flow of traffic until a tier 1 ISP is reached, at which point on, the net flow of money goes against the traffic.

2.4 Service providers

'Service Providers' offer higher-level QoS-based services encompassing both connectivity and informational aspects e.g. telephony, content streaming services. As opposed to *'IP Network Providers', 'Service Providers'* may not necessarily own and administer an IP network infrastructure; they need to administer the necessary infrastructure required by the provisioning of the offered services e.g. VoIP gateways, IP video-servers, content distribution servers. As such, for fulfilling the connectivity aspects of their services, *'Service Providers'* may rely on the connectivity services offered by *'IP Network Providers'*. In this sense, *'Service Providers'* interact with *'IP Network Providers'* following a customer-provider paradigm on the basis of respective agreements (SLAs). Furthermore, for expanding the geographical scope and augmenting the portfolio of the services offered, *'Service Providers'* may interact with each other on a peer-to-peer or a strict customer-provider basis.

2.5 Physical connectivity providers

'Physical Connectivity Providers' offer physical (up to the link layer) connectivity services between protocol-compatible equipment in determined locations. It should be noted that the connectivity services may also be offered in higher layers (layer-3 e.g. IP), however these services are mainly between specific points as opposed to the IP connectivity services offered by *'IP Network Providers'* which may be between any points in the IP address space. *'Physical Connectivity Providers'* are distinguished into two main categories according to their target market: *'Facilities Providers'* and

Copyright © MESCAL Consortium, June 2005

'Access Providers'. These types of Providers could be seen as distinct stakeholders. One special case of a *'Facilities Provider'* is an Internet eXchange Point (IXP). An IXP is a physical network infrastructure operated by an entity with the purpose of facilitating the exchange of Internet traffic between *IP Network Provider* domains. Any *IP Network Provider* that is connected to an IXP can exchange traffic with any other *IP Network Providers* connected to the same IXP, using a single physical connection to the IXP, thus overcoming the scalability problem of individual interconnection links. Deliverable D1.4 contains a detailed discussion on the role of IXPs and their implications on the MESCAL solutions.

The services of *'Facilities Providers'* are mainly offered to *'IP Network Providers'* to provide the required link-layer connectivity in their IP network infrastructure or to interconnect with their peers as discussed previously. As such, *'IP Network Providers'* may interact with *'Facilities Providers'* following a customer-provider paradigm on the basis of respective agreements (SLAs). These interactions are analysed further from the perspective of dynamic network provisioning where an *IP Network Provider* may dynamically determine and request capacity between its IP routers from the underlying *Physical Connectivity Provider*. *'Facilities Providers'* may be differentiated according to the type of technology they rely upon (e.g. optical fibre, satellite, antennas), deployment means (terrestrial, submarine, aerial) and their size in terms of geographical span and customer base. The technological means for provisioning optical networks are reviewed in some detail in chapter 3 of deliverable D1.4 on optical network technologies and their implication on MESCAL.

The preliminary business model as described in D1.1 did not adequately capture the relationships between '*Facilities Providers'*. This is corrected in the current view. A single *Facility Provider* may interconnect the *IP Network Providers*, as in the case of an IXP, or where a single national carrier provides a leased line between them. This is depicted in Figure 2. Alternatively, private peering could be achieved through international connections through a chain of *Facility Providers*. In this case there would be separate *'Physical Connectivity Providers'* who cooperate and interwork to provide the end-to-end physical layer capability. The latter case could be captured with an additional arrow between separate *'Physical Connectivity Providers'* as shown in Figure 3.

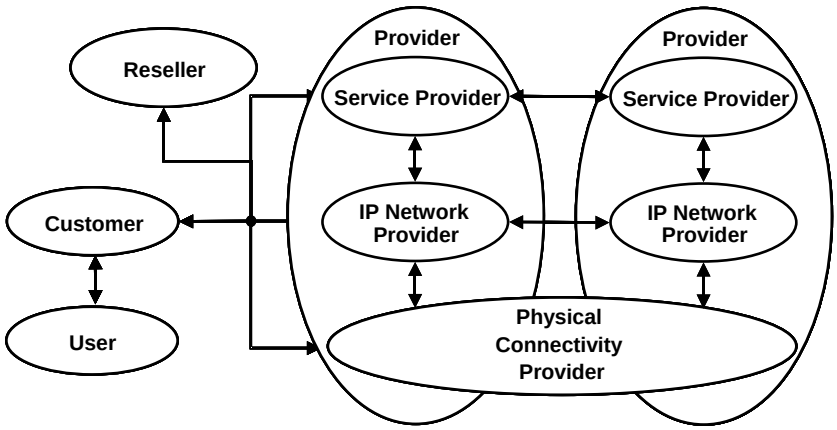


Figure 2. Revised MESCAL business model – Common Physical Connectivity Provider

Copyright © MESCAL Consortium, June 2005

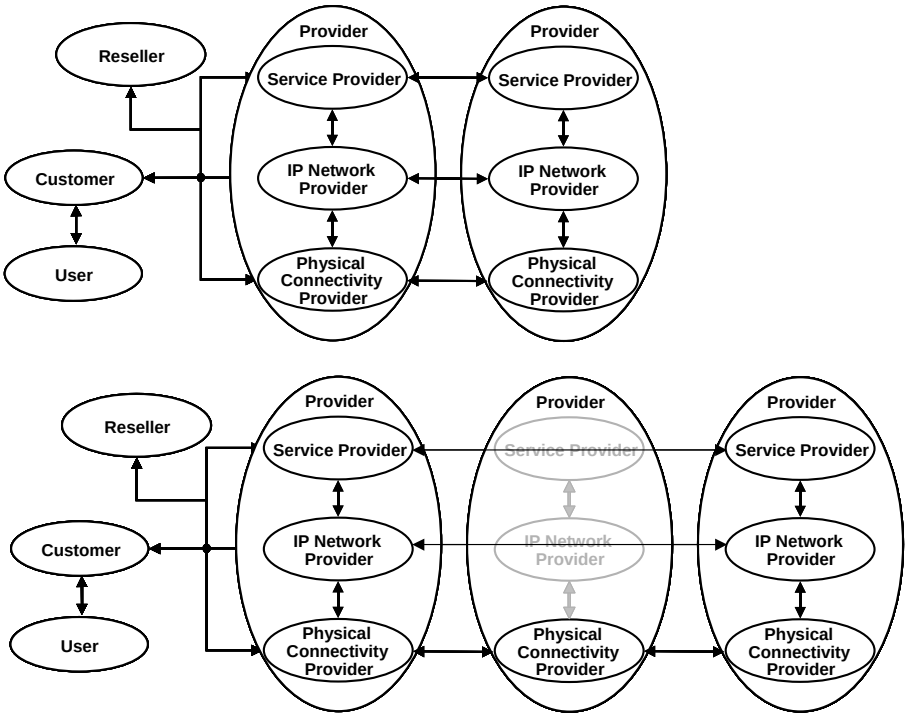


Figure 3. Revised MESCAL business model – Interworking between Physical Connectivity Providers

'Access Providers' offer services for connecting *'Customer'* premises equipment to the appropriate (*'Service'* or *'IP Network'*) *'Providers'* equipment. They own and administer appropriate infrastructure e.g. cables, concentrators. They may be differentiated according to the type of technology they employ e.g. POTS, FR, ISDN, xDSL, WLAN, Ethernet, as well as their deployment means and their size in terms of covered geographical area and customer base.

2.6 Resellers

'Resellers' are intermediaries in offering the QoS-based services of the *'Providers'* to the *'Customers'*. In essence, *'Resellers'* offer market-penetration services (e.g. sales force, distribution/selling points) to *'Providers'* for promoting and selling their QoS-based services in the market.

Copyright © MESCAL Consortium, June 2005

3 ASSUMPTIONS AND REQUIREMENTS

3.1 MESCAL Assumptions

"The Internet is tremendously diverse, complex, and dynamic. Nothing is ``typical''!"

Vern Paxson (IRTF chair),

The MESCAL project, which aims at deploying end-to-end quality of service at large scale, i.e. across multiple domains, relies on certain assumptions allowing a better understanding of the problem area to be addressed and thus a clear definition of the required solutions and mechanisms to be pursued.

Regarding network aspects, MESCAL assumes that providers in the Internet employ IP-based networks with DiffServ and/or MPLS capabilities for their intra-domain needs. We assume that the QoS capabilities of a given domain can be described as a limited set of well-known performance characteristics (typically one-way transit delay, inter-packet delay variation, packet loss). The QoS capabilities are tightly coupled with the constraints of the topological infrastructure. This means that the provider will engineer the network so that QoS capabilities holding from any ingress point to any egress point of its domain.

These assumptions allow the specification effort of the project to take advantage of the standardised but flexible IP DiffServ framework to build solutions for inter-domain QoS. Moreover, MESCAL considers that each provider enforces its own traffic engineering policies for its intra-domain needs.

MESCAL assumes that there is NO "Internet God". Therefore, a given provider cannot have direct service contracts with all Internet actors; Therefore, MESCAL needs to seek for solutions, which will have to rely on agreements between providers based on what is available and deployed in the different domains of the Internet.

The domain granularity assumed by MESCAL is the AS and/or set of ASs managed by the same providers. In the rest of this document when referring to interactions between ASs it may be implied to be between different providers. This is not an assumption and it does not mean that the MESCAL solution will only be applicable between ASs that belong to different administrative authorities. The MESCAL solution will be applicable even to cases where the ASs are under the same administration, i.e. provider. It is envisaged that an approach, which handles the inter-provider case, will be directly applicable to the intra-provider inter-AS (if and only if these ASs are adjacent). In the latter case some further optimisations may be applicable, and they will be studied as extensions to the general case.

MESCAL distinguishes two kinds of customers, leading to the definition of two kinds of service contracts:

- pSLS, for inter provider relationships (service-peering)
- cSLS, for end-customers (end-customer – provider relationship)

When peering, a provider wants to extend the network services it provides to its end-users within its own domain (AS scope) at a larger scale. Thus a pSLS can be viewed as a permission to send and/or receive certain quantities of traffic with contractual guarantees (destinations, throughput, QoS constraints…).

It is argued any financial settlement structure is robust only where a retail model exists that is relatively uniform in both its nature and deployment and encompasses the provision of services on an end-to-end basis [Huston99]. In MESCAL we assume the uniform financial settlement model. In this model the QoS signal initiator (i.e. cSLS) undertakes to bear the cost of the entire end-to-end traffic flow associated with the QoS signal. This is a retail model where the application initiator undertakes to fund the entire cost of data transit associated with the application.

Note that funding the entire end-to-end cost as described above does not necessarily assume a centralised model were the QoS initiator has to pay all the intermediate providers. This model is

Copyright © MESCAL Consortium, June 2005

analogous to the end-to-end retail models of the telephony, postal, and freight industries. In such a model, the participating agents are compensated for the use of their services through a financial distribution of the original end-to-end revenue, and a logical base for inter-agent financial settlements (i.e. pSLSs) is the outcome [Huston99]. Note that the service cost is out of scope of MESCAL.

As far as end- customers are concerned, a clear distinction is made between mass market and enterprise customers, as their service needs and habits differ -implying different kinds of cSLS's and, possibly, different solutions:

- On one hand, mass-market customers are known to require some quality of service “in general” (i.e. when they decide they need QoS, for any service they might want to use at the time, to/from wherever these services are offered in the Internet). Thus cSLSs in this case will rather provide loose guarantees, encompassing all possibilities that can be required by these customers.
- On the other hand, enterprise customers are known to require quality of service for specific usage (i.e. at certain times, for specific services, to given destinations and/or from specific sources, with accurate constraints per service/destination). Thus, cSLSs in this case will provide strong guarantees, contractually enforced by an accurate definition of the customers' requirements.

Finally, the following general assumptions/constraints are made in order to build solutions that are adequate and deployable in the Internet.

- Networks should be ready to convey inter-domain QoS traffic before cSLS agreements are negotiated (as is the case with inter-domain routing).
- The MESCAL proposal does not make any assumptions on the applications that will use the QoS capabilities, allowing in for the support of unanticipated applications.
- Whenever a QoS route to destinations is not available, the best effort route may be used as an alternative.

3.2 Customer and Provider Requirements

3.2.1 Introduction

Increasing the deployment of QoS-based services across the Internet requires a large set of providers to cooperate. This cooperation raises a number of complex challenges for Internet operators, not only due to the complexity of the technical issues to be solved but also due to the lack of appropriate standardised contractual agreements and automatic negotiation mechanisms between providers. To this aim, MESCAL will design a suitable inter-domain IP QoS architecture and appropriate solutions.

A first step to achieve the above task is to list the requirements related to the actors involved in the QoS delivery chain. This section aims at identifying requirements from both providers' and customers' perspectives. Therefore, the proposed MESCAL solution will address such requirements.

Then, for evaluating an Internet QoS solution against a specific requirement, the solution will have to:

- Indicate what level of support it offers for each of the requirements: F (Full) means that all implications of the requirement can be fulfilled, M (Medium) means that a significant part of the requirement can be satisfactorily met, L (Low) otherwise.
- Give an explanation of the above rating, i.e., whatever the result of the evaluation; concrete features must be put forward to justify the rating. Especially, when the solution is said to 'Fully' meet the requirement, a detailed justification of all the points tackled by the requirement description should be provided.

The listed requirements are described as follows: For each requirement a general definition is given. Then, a description of its applicability in the Inter-domain QoS delivery context is provided, by outlining the extent at which the requirement will be considered by the project.

Copyright © MESCAL Consortium, June 2005

3.2.2 Provider requirements

3.2.2.1 Introduction

This section presents the set of provider requirements that MESCAL should address to ensure end-to-end QoS delivery. The purpose is to give an exhaustive and precise definition of requirements against which different solutions will be judged and their applicability evaluated.

3.2.2.2 Description of requirements

3.2.2.2.1 P1: Extend the geographical scope of its QoS services

Definition: The ability for a provider to furnish a level of inter-domain QoS equivalent to the one it can offer to its customers for intra-domain traffic.

Applicability to Inter-domain QoS delivery context:

The MESCAL solution should ensure that a provider is enabled to have at its disposal QoS offers, spanning beyond its domain i.e. across multiple ASes, with the levels of QoS being coherent with the ones it is able to offer to its customers for intra-domain traffic.

More specifically, the intent is to enable a provider to extend its QoS classes (notion of e-QC) over multiple domains, apart from its own, thus enabling the provider to offer reachability to networks beyond its own domain with QoS parameters similar/close to what it could provide within its own domain.

This requirement breaks down into the following two non-exclusive cases, regarding how the expandability of the QoS span of a provider is meant:

- Limited expandability: The provider is able to offer QoS reachability only to specific networks outside its domain. In this case, different QoS levels may apply to different networks. That is, a particular QoS level may only be experienced when reaching a specific destination network.
- Unlimited expandability: The provider is able to offer QoS reachability to (almost) any destination in the Internet, much like as today reachability is offered in the Internet at best-effort QoS levels. The offered QoS levels apply to all destinations.

The above cases are distinguished because they refer to different business models and because they may require different technical solutions e.g. in the first case it may be better to build inter-domain VPNs (e.g. MPLS-based), whereas in the latter case it may be better to build a QoS-aware IP layer across the Internet.

Obviously, the above cases depend on corresponding capabilities of other providers. As such, the requirement of expanding the geographical scope of QoS services in a provider domain entails the following sub-requirement: What are the QoS reachability capabilities assumed to exist in the other providers? The MESCAL solution options should clearly identify the QoS reachability capabilities assumed by the other provider domains.

3.2.2.2.2 P2: Find QoS partners quickly and easily

Definition: The ability to easily and quickly determine the appropriate partners (from a business perspective) for expanding the scope of QoS services i.e. with which to establish pSLSs and the way to achieve that.

Applicability to Inter-domain QoS delivery context:

There are two aspects contained in this requirement.

- Offered QoS Class discovery: Solutions should provide appropriate means to enable providers to discover feasible Offered QoS Classes.

Copyright © MESCAL Consortium, June 2005

- pSLS negotiation: Once a path and QoS values to reach a destination are chosen by a provider, the means to set up the required pSLS(s) should be rapid and easy. This means that the process for establishing pSLSs should be feasible in the sense that it should follow accepted business practices, well-defined, involving finite steps and based on commonly understood notions. Relevant automated means are also desired for speeding-up the process. A provider may need to set up pSLSs with direct peers, or with a remote AS, or pSLSs may need to be established between two remote ASs upon request from a third-party AS. In the two latter cases, the information necessary can be provided by the means used for QoS path discovery, as described above.

3.2.2.2.3 P3: Verify the fulfilment of the contract

Definition: The ability to check that what is provided conforms to what has been stated contractually.

Applicability to Inter-domain QoS delivery context:

The solution must enable conformance verification of the actual service against the contractual expectations. This should be true for both cSLSs and pSLSs. In either case, the networks' configurations and policies derived by the MESCAL system must ensure that the QoS parameters negotiated in the contract are respected. Some tools or monitoring points must be available to check the conformance of the measured QoS service towards what has been negotiated.

Related to the above, the solution must state relevant tools and information, which are assumed to be provided by other providers.

3.2.2.2.4 P4: Accounting, charging and billing

Definitions:

- Accounting: Technical process of collecting usage records from network nodes such as sender, receiver or router.
- Charging: Transforming the usage records into monetary units and associating them with the user's identity.
- Billing: Collecting charging records, summarising their charging content, and delivering a bill to a customer including an optional list of detailed charges per user, per service.

Applicability to Inter-domain QoS delivery context:

Not considered by MESCAL.

3.2.2.2.5 P5: Scalability

Definition: Ability for the system to function effectively and keep its performance in desired levels, as the size of the parameters influencing its behaviour increase. In other words, the proposed MESCAL solution should be able to keep its performances unaffected whatever the size of domain span, which could be expressed in terms of number of participating domains (and routers), whatever the number of (c/p)SLSs to be dynamically negotiated and invoked. Performances of the system should also be kept unchanged whatever the volume of the QoS-related information that will be propagated across domains, and without affecting the overall stability and (access) availability of the IP networks themselves.

Applicability to Inter-domain QoS delivery context:

The scalability of the MESCAL solution should be evaluated. This entails the assessment of the complexity of the decision-making components.

Typical size parameters to take into account include:

- Per AS: average number of peers, average number of QCs
- Globally: number of participant ASs, number of required/established pSLSs, number of e-QCs, and number of cSLSs.

Copyright © MESCAL Consortium, June 2005

3.2.2.2.6 P6: Manageability

Definition: Ability for the system to be managed easily.

Applicability to Inter-domain QoS delivery context:

There are two main domains covered, which must be tackled by MESCAL, in this area:

- Configuration
  - The base configuration, which is intrinsic to the solution, must be tolerable and automation must be provided.
  - The configuration induced by the enforcement of a newly agreed pSLS must not be too heavy, nor make the system unstable (even briefly).
  - The impact of an external modification (for instance, a modification of an intra-domain QC) must be limited, and must not leave the system unstable (even briefly).
- Monitoring
  - The system must offer specific points of visibility for monitoring and feedback purposes (different from the traditional ones, SNMP MIBs, COPS PIBs…etc)

3.2.2.2.7 P7: Resiliency

Definition: Ability for the system to recover from a failure by repairing itself automatically without having to restart the service.

Applicability to Inter-domain QoS delivery context:

Within MESCAL, this means among others that, in case of failure (e.g. link rupture, router breakdown), the system must be able to find/propose another path of equivalent QoS for the impacted destinations. This operation must ensure that all active flows are automatically redirected correctly (e.g. no routing loops) with a minimum of disruption. Notably, a renegotiation of the cSLS conditions former to the failure must be avoided, the system being responsible for providing a satisfactory alternative.

One particular aspect concerning resiliency is security. The following questions should be addressed:

- Does the system present points, which could be exploited by hackers? Are there well-known possible points of failure, whose malfunction could lead to an unavailability of the system?

3.2.2.2.8 P8: pSLS management flexibility

Definition: Degree of freedom for an AS to modify its pSLSs.

Applicability to Inter-domain QoS delivery context:

pSLSs should be viewed as managed entities. As such, providers should be given means for requesting, establishing, modifying and deleting pSLSs. Caution should be taken to ensure that the modification of pSLSs do not disturb, but is in accordance with the requirements of other pSLSs relying on the pSLSs under modification.

In case of pSLS deletion, means must be provided to ensure the coherence and stability of the system, notably the good handling and management of pSLSs that were relying on the deleted pSLS (in a cascading approach). Possible solutions are for instance: forbid the deletion, notification to peers so that they modify their pSLSs before deletion is completed…

3.2.2.2.9 P9: Deployment easiness

Definition: How long and difficult it would take to have all the building blocks ready for operation, that is to say, to begin actual inter-domain communications with QoS activated.

Applicability to Inter-domain QoS delivery context:

Copyright © MESCAL Consortium, June 2005

Easiness of deployment depends on a number of parameters, such as: number of new protocols required, degree of adherence of the proposed solutions to the market and capabilities of commercially available routers, magnitude of required modifications to existing protocols, impact on intra-domain routing, impact on inter-domain routing and required conformance of other providers with the proposed solutions. The MESCAL solution(s) should clearly identify and describe such aspects.

3.2.2.2.10 P10: Backward compatibility

Definition: The risk and impact on the infrastructure already in place, when deploying the MESCAL solutions.

Applicability to Inter-domain QoS delivery context:

In order to achieve the goals pursued by MESCAL, proposed solutions are likely to introduce more or less modifications on the existing infrastructures. The MESCAL approach should provide the adequate guarantees as far as the backward compatibility issue is concerned, not only for allowing a smooth migration, but also to prevent existing infrastructures from being unusable and instable.

Among other criteria, the following are considered as important to judge the fulfilment of this requirement:

- The impact on the intra-domain routing process must be as limited as possible.
- The impact on the inter-domain routing process must be as limited as possible.
- When in operation, the MESCAL system must not introduce instability neither on the network itself, nor on the already deployed and running services.

3.2.2.2.11 P11: Applicability to business model

Definition: To what business case(s) the solution is applicable.

Applicability to Inter-domain QoS delivery context:

As there are different business cases in offering Internet QoS services, the MESCAL solutions should be clearly positioned as to which type of business cases they can address.

Different business cases can be seen along the following views:

- Customer view:

  A typical mass-market customer, is potentially interested in accessing any kind of service in any location in the Internet and at any time.

  A typical enterprise customer, is focussed on a well- known and limited set of services whose location, duration, QoS constraints, can be perfectly defined.

- Provider view:

  The provider wishes to extend its own QoS services to external users at specific or any network in the Internet.

3.2.2.2.12 P12: Multicast aspects

Definition: Support for delivering multicast-based IP services in the Internet.

Applicability to Inter-domain QoS delivery context:

It is important to evaluate the impact of supporting multicast-based services on the features and performance of the approach along the following lines:

- Does the multicast support imply major changes or add-ons to the unicast model?
- Does the multicast service address all aspects (and customers) listed in the business model
- How to manage the replicated multicast traffic within the network?

Copyright © MESCAL Consortium, June 2005

- How to avoid imposing significant impacts on the underlying IGMP, PIM-SM, MBGP protocols, as well as core router architecture for including DiffServ aware multicast services?
- How to handle the scalability issues concerning QoS deployment?
  - Low overhead for group/QoS state maintenance within core networks.
  - No traffic conditioning capability within DiffServ core routers.

3.2.3 Customer Requirements

3.2.3.1 Introduction

This section presents requirements from the perspectives of the customers of QoS-based Internet services. The requirements are drawn from current business practices and market needs as understood by the project partners. The requirements pose corresponding requirements to the providers of QoS-based Internet services, which in turn need to be taken into account by the solution proposed by MESCAL.

Considering a provider, the term "customer" is taken to denote either an end-customer (recipient of QoS services), or another provider. Unless explicitly stated to denote a particular type of customer, the term "customer" is used to denote either of these types of customers.

3.2.3.2 Customer Requirements details

3.2.3.2.1 C1: Characteristics of QoS Services

Definition: Ability of customers to send/receive traffic with end-to-end QoS guarantees to/from destinations in the Internet.

Applicability to Inter-domain QoS delivery context:

This general requirement can break down into the following requirements:

- On the topological scope of the services: Customers should be able to send/receive traffic to/from specific and/or any destination in the Internet. That is, given the sites of a particular customer, the customer should be able to:
  - Send traffic with end-to-end QoS guarantees to specific destinations i.e. only to destinations, which have been a-priori agreed with the provider.
  - Send traffic with end-to-end QoS guarantees to any possible destination; of course, at the time of actually requesting the service, the destination should be clearly specified in the IP address space.
  - Receive traffic with end-to-end QoS guarantees from specific sources.
  - Receive traffic with end-to-end QoS guarantees from any possible sources.
- On the QoS: The QoS guarantees should refer to well-defined performance metrics reflecting the quality of the service from the customer's perspective. At the network layer, these metrics should reflect the packet transfer quality e.g. throughput, one-way transit delay, inter-packet delay variation, and packet loss. Note that, since MESCAL is concerned with connectivity QoS-based services only, these network-level metrics also make sense from customer perspectives. The end-to-end QoS guarantees should be clearly specified, commonly understood and mutually agreed by the customers and the providers. Related to this requirement are the following requirements:
  - The QoS could be quantitatively specified e.g. by means on certain bounds on related performance metrics.
  - The QoS could be qualitatively specified e.g. relatively to other QoS levels by means of appropriate qualifications such as golden, silver, bronze QoS levels.

Copyright © MESCAL Consortium, June 2005

- Customers should be able to freely choose their QoS-based services according to their actual needs. Customers should be ideally offered with a choice of QoS-services, even similar services at different QoS levels. However, when the service is actually requested, its QoS levels should be clearly and unambiguously defined.

The above requirements are distinguished because they refer to different types of customers, in terms of their requirements in using QoS services; therefore corresponding to different business cases. Some customers may know in advance the type of QoS services they require, whereas some others may not.

From a provider's perspective the above requirements yield the following requirements:

- The SLSs (pSLSs or cSLSs) underlying the offering of QoS-based services should be able to:
  - Capture the QoS characteristics of both upstream and downstream traffic (with respect the premises of a customer),
  - Specify the QoS characteristics quantitatively and/or qualitatively, and
  - Leave appropriate degrees of freedom in specifying the destinations and/or the QoS-levels of the QoS services, as required for covering the diverse needs of the customers, obviously according to the service provisioning capabilities of the provider.
- To be able to expand the geographical span of the offered QoS services beyond the provider domain –refer to corresponding provider requirement P1 in section 3.2.2.2.1.

3.2.3.2.2 C2: Dynamic Service Subscription

Definition: Ability of customers to dynamically subscribe and unsubscribe to QoS services, as per their communication needs.

Applicability to Inter-domain QoS delivery context:

Subscriptions should not be taken for granted as long-lived service contracts. Subscriptions may well be short-lived e.g. for a weekend. In fact, given the multi-service, multi-provider nature of the telecommunications market and the dynamic nature of customer needs –not all customers may know in advance their QoS service needs-, the ability to establish SLSs is a key aspect of service offering. Some customers may be more attracted by such dynamic service offerings compared to static, monolithic offerings, as their service needs continuously evolve.

From a provider's perspective, this requirement yields the following requirements:

- Providers should provide means for enabling customers to modify and terminate existing service contracts (SLSs).
- Providers should provide means for enabling customers to subscribe to QoS services on-demand and for short time periods, upon customers' requests.

Automated means for enabling subscription e.g. through the Web and for handling subscription requests e.g. service configuration/activation means, could facilitate the satisfactory fulfilment of the above requirements.

3.2.3.2.3 C3: Service Invocation

Definition: Ability of customers to invoke i.e. to actually request QoS services. Services are invoked by the users, within the subscription profiles (as described in the SLSs) agreed between the customer and the provider.

Applicability to Inter-domain QoS delivery context:

This requirement entails the following:

- Customers should be able to invoke the services either explicitly or implicitly. Explicit invocation will probably yield the use of an explicit QoS signalling protocol. Implicit invocation does not

Copyright © MESCAL Consortium, June 2005

require the explicit use of a QoS signalling protocol; users can initiate their flows at any time, as long as the corresponding streams adhere to agreed subscribed profile.

- Customers should be provided with appropriate means to invoke their QoS services. These means should be in accordance with the QoS service specifications i.e. should be able to convey the required information for identifying the particular QoS service requested, as specified by MESCAL.

From provider perspectives, this requirement yields the following requirements:

- Providers should be able to support both explicit and implicit service invocations. As for the former case is concerned, providers should be able to support the termination and handling of appropriate QoS signalling protocols.
- In either case, the invocation means should be capable of conveying the MESCAL QoS SLSs; either as part of the QoS signalling protocol used or through the information included in the IP header. The conveyed information should help providers in unambiguously identifying the MESCAL-conforming requested QoS service and the customer requesting it.
- Providers should provide for automated means in authenticating and authorising a (implicitly or explicitly) request of a QoS-based service.

3.2.3.2.4 C4: Verify the fulfilment of the contract

Definition: Ability of customers to assess on-line that the invoked services are provided in accordance to the agreed QoS levels.

Applicability to Inter-domain QoS delivery context:

Customers should be able to check that the quality of the services they have subscribed to is in accordance with what they have agreed with the provider. This requires that they should be provided with appropriate self-monitoring tools.

From a provider's perspective, this requirement yields the following requirements:

- Providers should provide customers with appropriate monitoring tools, enabling the customers to assess the QoS of the services they request.
- Providers should cater for appropriate means for receiving and analysing customer complaints with respect to the received services.

3.2.3.2.5 C5: Multicast Aspects

Definition: Ability of customers to initiate and/or participate to multicast groups with some QoS guarantees.

Applicability to Inter-domain QoS delivery context:

- Since almost all the multicast services are receiver oriented, the following is from the perspectives of multicast receivers (group members):
- Receivers desire to receive specific multicast traffic from the subscribed group, and hence the functionality of source filtering is needed to avoid delivering unwanted multicast traffic.
- Receivers should be able to specify their QoS requirements individually, i.e. different recipients could specify different QoS levels via receiver-oriented cSLSs for multicast traffic.

Copyright © MESCAL Consortium, June 2005

4 THE MESCAL QOS SERVICE MODEL (DEFINITIONS)

4.1 Introduction

Current business practices prove that there is not (cannot be) a single provider offering global coverage of the whole Internet. As such, providers need to interact between them so as to expand the geographical scope of the services they offer. Considering QoS-based services, these interactions may not exclusively occur at the network (IP) layer; they may also occur at the service layer on the basis of specific service agreements.

In the above scenery, this document introduces the MESCAL Internet service model, which aims at laying down the notions, entities and relationships between them, pertinent to the issue of definition and provisioning of QoS-based services in the Internet, across multiple Provider domains. In other words, the MESCAL Internet service model presents the informational architecture/the basic 'service vocabulary' for building/defining Internet QoS-based services.

From another angle, the MESCAL service model outlines the requirements of Internet QoS-based services from an informational viewpoint. As such, it sets the functional targets of the service offering and provisioning functionality, while it also presents the necessary abstractions in the service layer around which this functionality needs to be designed.

The MESCAL model relies on the QoS service model proposed by TEQUILA [TEQUI], [Trimin03], for QoS-based intra-domain services. The MESCAL model extends the TEQUILA model to cover QoS-based services spanning the whole Internet, rather than a domain of a particular provider.

4.2 Notions and Entities

This section presents the notions and entities of the MESCAL Internet QoS model.

4.2.1 QoS-based Services

4.2.1.1 Definitions

The term *service* denotes, from customer perspectives, a specific offering made by a provider, which (offering) should clearly and unambiguously describe what it offers and the terms and conditions under which it could be used. Equivalently, from provider perspectives, a service denotes a subset of the provider's domain capabilities with a clear description of the what's and how's regarding its use by customers or third parties in general.

The term *QoS-based service, or just QoS service* denotes a service that is believed to entail a sort of added value to customers e.g. matching application and customer usage requirements.

The current trend in service offering is contract-based. Services are offered on the basis of the so-called *Service Level Agreements (SLAs)*. SLAs are established between customers and providers and describe the characteristics of the service and the mutual responsibilities of each party (customer, provider) for using/providing the service. In SLA-based service offering then, on one hand services should be described comprehensively enough so that can be understood by the customers, on the other hand providers should ensure that the characteristics of the services, as depicted in the SLAs, are indeed provided as agreed. SLAs may also be established between two providers -with one provider acting in a customer role and the other in a customer role- to back-up agreements at service level for expanding the geographical span of their services (see also section 4.2.1.2). SLAs between providers extend the notion of peering business agreements that exist today between providers for mutually exchanging traffic at given rates, or even without any financial settlement [Huston99]. Obviously, in a QoS-based Internet such agreements do not present a viable model; they need to include description of service characteristics, accounting and billing aspects, hence the need for SLAs.

The term *Service Level Specifications (SLSs)* denotes the technical characteristics of a given service in the context of an SLA. The technical characteristics of a service refer to the network level

Copyright © MESCAL Consortium, June 2005

provisioning aspects of the service e.g. request, activation and delivery aspects from network perspectives. Non-technical service provisioning aspects such as billing and payment aspects, are not part of SLSs; they are part of the overall SLA. SLSs are integral part of SLAs, and conversely SLAs include SLSs.

MESCAL is concerned with SLSs. Service accounting and billing aspects are outside the scope of MESCAL investigation.

4.2.1.2 On SLSs – cSLSs and pSLSs

MESCAL distinguishes two types of SLSs (and subsequently SLAs): *cSLSs* established between customers and providers and *pSLSs* established between providers.

The providers between which pSLSs are established may not necessarily be interconnected. In the general case, a provider (acting in a customer role) may establish pSLSs with a remote provider (acting in a provider role), should the latter be appropriately located and contacted.

The term *peering providers* is used to denote providers, which are interconnected; and, the term *service-peering providers* is used to denote providers between which pSLSs have been established.

The following operations on SLSs (and SLAs) should be allowed: *establishment of new SLSs, modification and termination of already established SLSs*. To this end, appropriate means should be provided, including informational models for describing SLSs and well defined manual and/or automated procedures for discovering, requesting and agreeing on the establishment, modification and termination of SLSs. Such procedures should provide for negotiation semantics/primitives for overcoming the limitations of a monolithic 'yes/no' type of interaction. The Service Negotiation Protocol (SrNP) specified by TEQUILA [TEQUI] is an example of such automated negotiation means.

Two styles in requesting and subsequently establishing SLSs can be distinguished: *restricted SLS request style* and *unrestricted SLS request style*. Under the restricted SLS request style, a requestor (customer or provider acting in a customer role) requests from a provider the establishment of SLSs, which refer only to currently offered services. Under the unrestricted request style, a requestor may request from a provider the establishment of SLSs referring to services that may need additional capabilities than the ones provided by the currently offered services. In a sense, the unrestricted request style is equivalent to the restricted request style with an addition of the nature 'please send any other request to the marketing department'.

The above differentiation is necessary for capturing different business strategies instigating the establishment of SLSs as well as decisions regarding the services to be offered. For instance, providers could allow for an unrestricted style, as a means to 'grasp' needs for future services. Furthermore, this differentiation is helpful for deriving requirements for negotiation procedures and associated logic.

All the above aspects on SLSs and their establishment, are deemed essential in the arena of service provisioning in the Internet, where in addition to advances in the network (IP) layer, appropriate 'hooks' for capturing business level objectives and policies need to be catered for.

4.2.1.3 MESCAL Service Focus - Connectivity Services

MESCAL is concerned with *QoS-based connectivity services.* A connectivity service is a 'get-through'/ 'traverse' service for reaching particular destination(s) from specific source(s) in the IP address space. The QoS aspects of connectivity services mainly refer to the quality at which the user-transmitted IP datagrams are transferred by the network between user-ends. Higher-level, informational, application-specific services e.g. streaming or video-on-demand services are outside of the scope of MESCAL. Note, that the latter services usually have a connectivity dimension, which if not provisioned properly, would lead the whole service not be provisioned at all. Therefore, connectivity services should be studied first, before moving to higher-level services.

MESCAL distinguishes QoS-based connectivity services into *elementary and complex connectivity services*. Elementary connectivity services are strictly point-to-point and unidirectional, whereas complex connectivity services may be multi-point-to-multi-point and bi-directional. As such, complex

Copyright © MESCAL Consortium, June 2005

connectivity services encompass a number of elementary connectivity services as appropriate to the context of the connectivity service itself; equivalently, elementary connectivity services can be viewed as the 'connectivity legs' (the 'nucleus') of complex connectivity services. Typical examples of (complex) connectivity services include VPN, Internet access, server access services. Complex connectivity services constitute the connectivity services actually offered to the customers, whereas elementary connectivity services can only exist in the context of these services and as such are not offered to customers. As such, the term connectivity service used throughout this document implies a complex connectivity service.

The distinction between complex and elementary connectivity services is deemed helpful for decomposing the provisioning of connectivity services from the perspectives of a provider. Furthermore, this distinction may be used to facilitate the specification of SLSs. As a complex connectivity service is comprised by a number of elementary connectivity services, its SLS may be comprised by the SLSs of its constituent elementary services; therefore, SLSs may only be specified for elementary connectivity services. In line with this view, TEQUILA has specified its SLS template for intra-domain QoS-based connectivity services [Goder02].

Figure 4 depicts the QoS-based service hierarchy as assumed by MESCAL; from higher-level, informational, application-specific services to complex and elementary connectivity services.

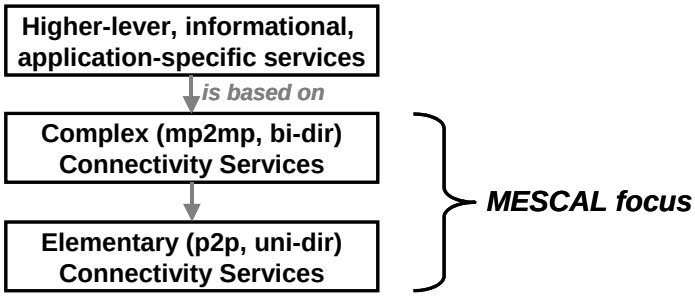


Figure 4: QoS-based service hierarchy and MESCAL focus.

Based on the roles identified in the MESCAL business model, QoS-based connectivity services are offered by the so-called 'IP Network Providers', who own and administer an IP network infrastructure including customer access means. As already said, 'IP Network Providers' need to interact between them so as to expand the geographical scope of the QoS services they offer. These interactions may occur at a network (IP) and/or service layer on the basis of respective pSLSs. The following sections introduce appropriate notions underlying these interactions.

Throughout this document the term service denotes a connectivity service and the term provider an 'IP Network Provider', unless otherwise specified.

4.2.2 QoS-classes

4.2.2.1 Definitions

As outlined in the previous section, QoS-based services reflect and need to be supported by corresponding 'capabilities' of the provider domains across the Internet. As such, the following definitions are put forward:

A *QoS-class (QC)* denotes a basic network-wide *QoS transfer capability* of a Provider domain.

A QoS transfer capability is a set of attribute-value pairs, where the attributes express various packet transfer performance parameters such as one-way transit delay, packet loss and inter-packet delay variation (jitter), and their values have the meaning of upper bounds on them. Considering the statistical nature of the packet transfer performance parameters, the corresponding attributes may not be invariant; rather, they could refer to specific time-intervals, denoting moving averages and/or percentiles or inverse percentiles (confidence levels for being below a given threshold). Furthermore, the attribute values (bounds) may not be absolutely defined; they may be qualitatively defined

Copyright © MESCAL Consortium, June 2005

relatively to the corresponding values of other QoS-classes. In essence then, a QoS-class is a set of packet *transfer performance parameters* (attributes) associated with specific *performance targets* (values).

It should be noted that QoS-classes are not services per se; their definition does not entail service provisioning semantics and aspects e.g. activation modes, user identification and usage requirements. The concept of QoS-class could be compared to the notion of Per Domain Behaviours (PDBs) – debated in the DiffServ workgroup of the IETF in the recent past.

QoS-classes are associated with a number of constraints, which denote conditions for their time- and topology-wise availability. Time-related constraints are expressed in period(s) of day/week/month during which the QoS-class can be (or cannot be) made available. Topological constraints are expressed in terms of reachable domain boundaries (e.g. IP network prefixes) between which the QoS-class can be (or cannot be) made available.

Considering a provider's domain, the provisioning of a QoS-class may solely rely upon the domain's own network engineering abilities, those related to routing and resource (bandwidth and buffer) management, which result by combining the elementary IP DiffServ QoS capabilities with intelligent traffic engineering functions and related policies. In addition to the domain's own engineering abilities, the provisioning of a QoS-class end-to-end may rely upon the QoS transfer capabilities (QoS-classes) provided by other provider domains, should the latter could be made known and used; hence the necessity of interactions between providers.

We distinguish between *local-QoS-classes* and *extended-QoS-classes*. Namely, given a provider domain:

A l*ocal-QoS-class (l-QC)* denotes a basic network-wide QoS transfer capability that can be provided by means employed in the provider domain itself. Evidently, the domain boundaries appearing in the topological constraints of an l-QC should belong to the boundaries of the provider domain.

An *extended-QoS-class (e-QC)* denotes a basic network-wide QoS transfer capability that can be provided by means employed not only in the provider domain but also utilising appropriate means in other (service-peering) provider domains. In other words, an e-QC is provided by combining the QoS transfer capabilities (QoS-classes) of the provider domain with appropriate capabilities (QoS-classes, l-QC or e-QC) of other provider domains. The domain boundaries appearing in the topological constraints of an e-QC could be outside the boundaries of the provider domain, thus extending the topological scope of the QoS transfer capabilities of the provider domain.

The above distinction is required for capturing the notion of 'QoS capabilities' across domains, upon which QoS-based Internet services are/could be built. In a sense, this distinction is analogous to the intra-/inter-domain distinction that usually applies in the context of the Internet.

Hereafter, the term QoS-class (QC) denotes either a local or an extended QoS-class, unless it is explicitly said to mean a particular one of the two.

4.2.2.2 Comparisons between QoS-classes

By comparing the values of the corresponding QoS-class performance parameters, an ordering relationship could be defined amongst QoS-classes. The following definitions are put forward:

A QoS-class A is said to be *"at least as good as"* a QoS-class B, conversely QoS-class B is said to be *"at most as good as"* a QoS-class A, denoted by *"$A \geq B$"* if and only if the values of *all* corresponding performance parameters of the QoS-classes A and B are accordingly ordered. The 'accordingly' qualification refers to the nature of the QoS-class performance parameters (attributes) as discussed in the previous section (moving averages, percentiles etc.). For instance, if the QoS-class attributes denote averages over the same period of time, their values (bounds) should be ordered according to the $\leq$ relationship, whereas if the QoS-class attributes denote inverse percentiles for a given threshold, their values should be ordered according to the $\geq$ relationship. Obviously, the attribute values to compare should be expressed in the same or convertible units.

Copyright © MESCAL Consortium, June 2005

Similarly we define the case that a QoS-class A is *"better"* than a QoS-class B, conversely that QoS-class B is *"worse"* than a QoS-class A, denoted by "A > B".

The above definitions could be extended to define a *lexicographical ordering* between QoS-classes. In this case the performance attributes of the QoS-classes should be appropriately prioritised viewing QoS-classes as ordered vectors of performance parameters; the first co-ordinate reflecting the most significant performance parameter and so on. Then, QoS-classes can be ordered by checking the values of the corresponding attributes per co-ordinate, not by checking the values of all corresponding attributes as in the previous definitions.

It should be noted that the defined *ordering relationship is partial*, not total, meaning that not every pair of QoS-classes can indeed be compared. For instance, this could be the case when the corresponding attributes of the QoS-classes to compare are of not of similar nature e.g. averages over different time periods or averages versus percentiles, making the comparison of their values infeasible. Alternatively, such cases could appear when QoS-classes are not compared lexicographically and *some* of the values of corresponding QoS-class performance parameters are accordingly ordered, whereas *some* others are not.

Because the QoS-class ordering relationship is partial, there might be a number of *"best"* or *"worst"* QoS-classes instead of a single such element, even if the set of QoS-classes is finite.

4.2.2.3 Types of values of QoS-classes

Orthogonal to the interpretation and the nature of the QoS-class performance parameters (attributes) as discussed in section 4.2.2.1, the values -and subsequently the QoS-classes- may be distinguished into different types according to how these values are assumed. The values may be nominal or actual. Nominal values are set/deduced theoretically, whereas actual values are set/deduced from operational practices. Both nominal and actual categories of values are subject to the specific business policies and operational practices of the particular Provider administration regarding service provisioning. Table 1 presents possible types of nominal and actual values.

QoS-class parameter values		
Category	**Type**	**Description**
Nominal – Set	Targeted	Values set as objectives for engineering the network, setting the targets of the off-line traffic engineering functions that dimension the network. These values are deduced by the requirements of widely deployed applications (cf. the notions of Meta-QoS-Class and global-QoS-class below) and/or market needs.
Nominal – Deduced	Engineered	Values yielded as a result of the off-line traffic engineering algorithms run to dimension the network so as to be able offer QoS-classes at their 'targeted' values. These values take into account the characteristics of the physical network configuration and topology, and their validity is subject to the errors inherent in the mathematical models used. These values should be as least as good as the corresponding 'targeted' type values.
Actual – Set	Offered	Values as assigned by the actual service offering activities i.e. values deemed appropriate for creating competitive service offerings to third parties (customers or providers). That is, these values are exported in the SLSs. Considering that QoS-based services should be in accordance with the capabilities of the domain, these values should be primarily at least as good as what is deemed 'attractive' to customers, while close to the corresponding 'engineered' or 'targeted' type values. These values may change as the corresponding policies for service offering change. They may be assigned either in absolute terms or qualitatively, relatively to the corresponding values of other QoS-classes.
Actual – Deduced	Measured	Values yielded by actual measurement during network operation. These values may be in any relation with the previous types of values. Ideally, they should be –on average- over a sufficiently large timescale, less than the corresponding engineered types of values and should not violate (at all) the 'offered' values. They may be used for validating and/or advertising the performance of the network. These values change as network traffic conditions change.

Table 1: QoS-class parameter value types.

In the above cases where the QoS-class parameters values (bounds) can be set and not deduced (i.e. 'targeted' and 'offered' type cases), the determination of appropriate values is subject to relevant business policies regarding service provisioning, taking into account requirements of well-known applications/services (cf. the notions of 'Meta-QoS-Class' and 'global-QoS-class' below), perceived

Copyright © MESCAL Consortium, June 2005

user needs and current/emerging market trends. The deduced QoS-class parameter values (i.e. 'engineered' and 'measured' type cases) are influenced by policies at network operation level. E.g. 'engineered' values may be influenced by policies determining the desired network-wide load balancing levels and 'measured' values are subject to the policy-set measurement parameters.

Obviously, the number of QoS-classes supported by a provider domain corresponds to the number of distinct values, which are actually set to the QoS-class performance parameters.

Based on the identified QoS-class parameter value types (cf. Table 1), the following terminology is introduced:

Targeted-QoS-class (t-QC), *engineered-QoS-class (eng-QC)*, *offered-QoS-class (o-QC)*, *measured-QoS-class (m-QC)* denotes a QoS-class where the values of its performance parameters are of 'targeted', 'engineered', 'offered', 'measured' type, correspondingly. The following statements are true regarding the relationship of these QoS-class types:

- By definition, there should be: o-QC ≤ t-QC ≤ eng-QC
- While eng-QC ≤ m-QC, the traffic-related objectives of traffic engineering are satisfied.
- When t-QC ≤ m-QC < eng-QC, re-engineering of (parts of) the network has to be considered.
- When o-QC ≤ m-QC < t-QC, the network must be re-engineered urgently.
- When m-QC < o-QC, the service contracts cannot be fulfilled anymore and the network must be re-engineered or even additional resources to be brought in.

4.2.2.4 Offering and Using QoS-classes

As QoS-classes reflect capabilities, this section addresses the question of 'what can these capabilities be used for?' or equivalently, 'how can these capabilities be used by third parties?'

Considering a Provider domain, QoS-classes may be used in either (not exclusively) of the following two cases:

- *For offering QoS-based services* to customers or other providers. In this case, the values of the QoS-classes may be of 'offered' or 'targeted' types (cf. Table 1).

  QoS-classes are building blocks for offering and provisioning QoS-based connectivity services – not the services themselves. Conversely, QoS-based connectivity services should be mapped to QoS-classes. In essence, from the perspectives of service offering, QoS-classes express the transfer quality aspects of the QoS-based connectivity services; and, from the perspectives of service provisioning, QoS-classes segregate the network QoS-space into a number of distinct classes, aggregating user QoS traffic accordingly. In this respect, the notion of QoS-classes sets the traffic-related objectives of the traffic engineering functions, prompting for approaches such as the ones following the Bandwidth Constraints model in the context of DiffServ-aware MPLS traffic engineering [Lefau03a] -the Russian Dolls Model [Lefau03b], the Maximum Allocated bandwidth Model [Lefau03c]- or the TEQUILA 'initially plan then take care' approach [Trimin01]. It should be stressed that the notion of QoS-classes does not necessarily prompt for hard bandwidth reservations per QoS-class in the network, as for instance in the TEQUILA approach.

- *For 'pure' informational purposes* that is, for announcing the QoS transfer capabilities of the provider domain. In this case, QoS-classes are announced 'as is' i.e. without service semantics. The values of the QoS-classes may be of 'targeted' or 'offered' or 'measured' types. Announcements could be done through various means, protocol- and/or platform-based, either periodically or asynchronously based on well-defined triggering conditions.

  Capability announcements are mainly targeted at service-peering providers, since QoS-classes do not bear service semantics, which are of interest to customers. They could also be targeted at customers, being provided as part of an agreed service. Providers might find useful to announce their QoS-classes –QoS transfer capabilities- for attracting service-peering providers for the purpose of increasing their revenue-earning sources (volumes of terminating and transiting QoS

Copyright © MESCAL Consortium, June 2005

traffic), furthermore for expanding the reach of the supported QoS-class capabilities on a mutual basis.

The substantial difference between the above cases lies in the implications incurred for the provider domain. In the first case, the provider is formally obliged to honour the terms and conditions underlying the offering of their services (SLS/SLAs). In the second case, the provider does not assume such formal obligations, as it (the provider) is not bound to any agreement, though it needs to uphold its announcements for the sake of its integrity and reputation.

Conversely, considering the cases above, QoS-classes supported by a provider domain can be used by other providers or customers in either of the following two cases:

- *Contentedly,* through corresponding QoS-based services. In this case, the use of QoS-class capabilities is done implicitly (indirectly) and is bound to mutual agreements underlying service offerings (cf. pSLSs, section 4.2.1). As such, QoS-class capabilities may be used with the guarantees underlying the offering of the corresponding service (cf. section 4.2.2.5).
- *Non-contentedly,* following related capability announcements. As long as a provider domain announces QoS-class capabilities, other provider domains or customers can use directly these capabilities i.e. not through the establishment of SLS/SLAs. In this case, the use of announced QoS-class capabilities is not bound to any agreement and it is on a 'to-do-my-best' basis.

The following point is worth discussing. The 'non-contentedly' use case does not necessarily imply that providers offer their QoS network resources for free. This kind of use case may happen on the basis of mutual business agreements between providers for exchanging aggregate traffic, as they exist today. The 'contentedly' use case extends these 'aggregate traffic exchange agreements', to agreements regarding the exchange/usage of traffic at certain QoS characteristics; these agreements are substantiated in corresponding pSLSs/pSLAs.

The 'non-contentedly' use case can be seen as a special case of the 'contentedly' use case when the services guarantees (cf. section 4.2.2.5), as depicted in the SLS/SLAs, are very loose –even non-existent. As such, without loss of generality *it is considered that QoS-classes can only be used in the context of QoS-based services i.e. in the context of SLSs/SLAs, which may or may not bear service guarantees*.

4.2.2.5 QoS-based Service Guarantees and QoS-classes

When applied to an offered service, the term *QoS-based service guarantees*, or *QoS service guarantees* for short, denotes the guarantees with which the quality aspects of the offered service can be provided from provider perspectives. These quality aspects differentiate similar services amongst them.

Considering QoS-based connectivity services, the focus of MESCAL, we view that QoS service guarantees consist of the following parts:

- *Performance guarantees*, which reflect the quality of the transfer of the user-transmitted datagrams in the context of the service. Considering that QoS-classes are the building blocks of QoS-based services (cf. discussion in previous section), these guarantees directly correspond to the values of (bounds on) the performance parameters of the QoS-class(es), which the offered service is based on.
- *Bandwidth guarantees*, expressed as an upper limit, in bandwidth units, on the user traffic injected in the network up to which the agreed service performance guarantees can be given.
- *Grade of service* denoting the probability of getting through the network valid (according to subscription profile) service requests.

The above types of QoS service guarantees should be reflected in the c/pSLSs, underlying the offering of QoS-based services. It is the responsibility of the provider offering the services to ensure that the above guarantees can be gracefully provided -not significantly violated.

Copyright © MESCAL Consortium, June 2005

The above classification of QoS service guarantees is in accordance to the view of the 'ippm' workgroup of the IETF, which does not consider bandwidth as a performance parameter. Furthermore, it is in line with the template proposed by TEQUILA [Goder02] for describing SLSs for QoS-based connectivity services.

It should be noted that the definition of QoS-classes prompts for hard or statistical/probabilistic QoS service performance guarantees, depending on the nature of the QoS-class performance parameters (attributes); as already outlined (cf. section 4.2.2.1), these attributes may be invariant, or they could be of statistical nature e.g. percentiles.

4.2.2.6 Provisioning of QoS-classes

It should be noted that the extent (confidence) at which the QoS-classes can be gracefully provisioned i.e. their performance targets –upper bounds on their performance parameters- can be safely met is not considered part of the definition of the QoS-class itself. This aspect entails service semantics (cf. previous section), which are not assumed by QoS-classes.

Therefore, the issue of being able to gracefully provision QoS-classes should be seen only in connection to the way QoS-classes are made available by the provider domain for use, as outlined in section 4.2.2.4). If QoS-classes are used for offering QoS-based services, QoS-class performance targets should be sufficiently met so that service performance guarantees (as specified in the SLSs) are not violated. If QoS-classes are used for announcing domain's capabilities, QoS-class performance targets should be met to the extent deemed necessary for the announcements to be valid.

The provisioning of QoS-classes to the extent desired falls into the realm of the domain's QoS delivery capabilities, combining the DiffServ elementary (nodal) QoS-enabling mechanisms with intelligent traffic engineering functions for QoS-based routing and resource management. In addition, in the case of extended-QoS-classes, QoS-class provisioning is also dependent on the corresponding capabilities of service-peering provider domains, which in turn are dependent on the corresponding capabilities of their service-peering domains and so on. Obviously, the existence of pSLSs between provider domains increases the confidence at which extended-QoS-classes could be provisioned in each domain.

For feasibility, manageability and scalability reasons, the QoS-classes should be pre-determined and fairly restricted in number; otherwise, the likelihood of not being able to manage effectively their provision would prohibitively increase. The fact that the values (bounds) of the performance parameters of the QoS-classes may be set in accordance to known application/service requirements (see following sections) contributes to this direction.

4.2.3 Meta-QoS-Classes

Although there has been much work done in Quality of Service (QoS) field over the last decade, little work has been undertaken to provide guidance on how to deploy QoS throughout the whole Internet. This section introduces a new concept in order to ease and guide the deployment of inter-domain QoS delivery services which could be potentially made accessible to a large Internet community independently of the service coverage of the involved network access providers.

4.2.3.1 Current inter-domain QoS deployment assessment

Based on current best practices, we can hardly say that QoS (if over-provisioning isn't considered as a part of QoS management) has been currently deployed inter-domain and even intra-domain in Service Providers' networks. The Internet remains an interconnection of best effort networks. The only worldwide transport service usable throughout the Internet is the best effort service. For instance there are currently no activated means at IP level for a video content provider to make it possible for their ready to pay customers to access the service via a performance guaranteed transport at large scale.

Copyright © MESCAL Consortium, June 2005

4.2.3.2 Requirements

An inter-domain QoS delivery solution *should* take into account some requirements that would prevent QoS techniques and architecture to impair the spirit in which the Internet has been devised since its early days. The idea is of course not to refuse any evolution in the Internet paradigm just because the Internet is as it is. The intention is to keep the features the great majority of people can agree on, because these features are deemed worth to be preserved for the good of citizens. The priority isn't necessary about technical and financial considerations. We should preserve the facility to spread Internet access, the facility to welcome new applications and the possibility to communicate from any point to any other points.

From this angle, the list of requirements *should* encompass:

- Networks *should* be ready to convey inter-domain QoS traffic before customers can initiate end-to-end SLS negotiations (just like inter-domain routing is);
- The solution *must* not, to the greatest extent possible, preclude unanticipated applications;
- A best effort route *must* be available when no QoS route is known;
- Best effort delivery *must* survive QoS;
- The solution *should* not rely on the existence of a centralised entity that have the knowledge and the control of Internet (*an Internet God*).

4.2.3.3 A basic QoS inter-domain problem: binding l-QC

4.2.3.3.1 Problem statement

A given Service Provider offers QoS-based services to its customers. The span of these services is limited to its network boundaries. On the other hand, this Service Provider is aware that many other Service Providers, scattered in the Internet, provide also QoS-based services to their customers. From a centric view, this Service Provider wants to benefit from the QoS infrastructure made up with all the QoS-enabled networks, to expand its QoS-based services to customers beyond the scope of its own network.

4.2.3.3.2 Who is a given Service Provider going to trust?

Let's consider a QoS AS path used by clients of a given provider to reach remote destinations. This provider can have strong agreements with its immediate neighbours, but what visibility of agreements between farther ASs? If one AS of the path does not respect its commitment, how can this provider know it? Even if it knows it what can it do? If its directly peered AS guarantees end-to-end performances and it complains to it, what will the neighbour do? Complain to the following AS? That will complain to the following AS? That will…?

The Conclusion is that the following sensible assumption has to be made: *each provider should trust only what its own peered neighbours guarantee for the crossing of their own networks*.

4.2.3.3.3 Using only local information to bind l-QCs

In order to provide QoS-based services, an AS implements l-QCs. Service extension to other ASs, on a low level (with regard to OSI layers), means l-QC extension outside the scope of a single AS. Then, knowing l-QCs capabilities advertised by its service peers, the basic technical question a provider has to face is: "*on what basis shall I bind my l-QC to their l-QC?*". Given one of local l-QCs what is the best match? What will be the criteria to choose one binding?

A Service Provider knows very little about agreements more than one AS hop away. These agreements can change and it is hard to have an accurate visibility of their evolutions. Therefore the provider *should* take the decision to bind one of its l-QCs to one of its AS neighbour l-QCs based solely on:

- What it knows about its own l-QCs

Copyright © MESCAL Consortium, June 2005

- What it knows about its AS neighbour l-QCs

A Service Provider *shouldn't* use any information related to what happens more than one AS hop away during the process of QC binding. It *should* try to find the best match between its l-QCs and its AS neighbour l-QCs. That is to say, it *should* bind one of its l-QC with the neighbour l-QC that has the closest performances (idea of extending l-QC). The result is that any QoS AS path is the concatenation of sheer local binding decisions.

4.2.3.3.4 What will ensure the AS path consistency?

At this stage, we can be confronted with a problem of QoS AS path consistency. If there's systematically a slight difference between the upstream l-QC and the downstream l-QC we can wind up with a significant slip between the first and the last l-QC. Therefore we must have a means to ensure the consistency and the coherency of a whole QoS AS path. The idea is to have a classification tool *that says two l-QCs can be bound together if, and only if, they are classified in the same category*. We call Meta-QoS-Class each category of this l-QC taxonomy.

From this viewpoint: *two l-QCs can be bound if, and only if, they correspond to the same Meta-QoS-Class.*

4.2.3.4 The Meta-QoS-Class concept

4.2.3.4.1 Meta-QoS-Class based on a worldwide common understanding of application QoS needs

The underlying philosophy behind Meta-QoS-Class concept relies on a worldwide common understanding of application QoS needs. Wherever end-users are connected they more or less use the same kinds of applications in quite similar business contexts. They also experience the same QoS difficulties and are likely to express very similar QoS requirements to their respective providers. Globally confronted with the same customers' requirements, providers are likely to define and deploy similar l-QCs, each of them being particularly designed to support applications of the same kind of QoS constraints. There are no particular objective reasons to consider that a Service Provider located in Japan would design a "VoIP" l-QC with short delay, low loss and small jitter while another Service Provider located in the US would have an opposite view. Applications impose constraints on the network, independently of where the service is offered in the Internet.

Therefore, even if we strongly believe there is no Internet God, we consider that:

There is a Customer God and he invented the Meta-QoS-Class concept.

It should be understood that a Meta-QoS-Class is actually an abstract concept. It is not a real l-QC implemented in a real network.

4.2.3.4.2 Meta-QoS-Class definition

A Meta-QoS-Class could be defined with the following attributes:

- A list of services (e.g. VOIP) the Meta-QoS-Class is particular suited for.
- Boundaries for each QoS performance attribute (one-way transit delay, one-way transit variation delay –jitter-, loss rate). In addition, a priority value could be assigned to each QoS performance attribute. For the sake of preserving the service objectives, the Meta-QoS-Class definition should also indicate if a given QoS Performance attribute is "Mandatory" or "Optional". Note, that several levels could be defined for these boundaries depending on the size of the network provider (regional, national, etc.)
- Constraints on traffic to put onto the Meta-QoS-Class (e.g. only TCP-friendly).
- Constraints on the ratio: resource for the class to traffic for the class.

Copyright © MESCAL Consortium, June 2005

A given Meta-QoS-Class followed by the same Meta-QoS-Class should equal the same Meta-QoS-Class (invariance).

4.2.3.4.3 What's in and out of a Meta-QoS-Class?

Only a limited set of Meta-QoS-Classes should be defined. Each AS classifies its own l-QCs based on Meta-QoS-Class. An l-QC from an AS can be bound only with a neighbour l-QC that refers to the same Meta-QoS-Class. Hereafter some precisions about the Meta-QoS-Class:

- *A Meta-QoS-Class typically bears properties relevant to the crossing of one and only one AS.* However this notion can be extended in a straightforward manner to the crossing of several AS, as long as we consider the set of AS as a super and single AS.
- A Meta-QoS-Class doesn't describe the way to implement an l-QC. It is not a real l-QC. It is a classification tool for implemented l-QC.
- The Meta-QoS-Class concept is very flexible with regard to new unanticipated applications. A new unanticipated application could drive a new Meta-QoS-Class. According to the end-to-end principle a new unanticipated application should have very little impact on existing l-QC, but this issue doesn't concern Meta-QoS-Classes per se, it is the problem of l-QC design and engineering.
- A hierarchy of Meta-QoS-Classes can be defined for a given type of service (e.g. VoIP with different quality levels). A given l-QC can be suitable for several Meta-QoS-Classes (even outside the same hierarchy). Several l-QCs in a given AS can be classified as belonging to the same Meta-QoS-Class. Private chains of interconnections, outside the scope of a global reachability, can do whatever they want i.e. bound to the Meta-QoS-Class constraint.

The DiffServ concept of Per-Domain Behaviour (PDB) should not be confused with the Meta-QoS-Class concept. The two concepts share the common characteristic of specifying some QoS performance values. However the two concepts don't exactly overlap. The two concepts differ in their purposes. The objective for the definition of a PDB is to help implementation of QoS capabilities within a network and that the objective for a Meta-QoS-Class is to help agreement negotiation between Service Providers. A PDB is closer to an l-QC than to a Meta-QoS-Class.

In summary the interest of Meta-QoS-Class concept is threefold as listed hereafter:

- Gives guidance for l-QC binding;
- Allows relevant l-QC bindings with no knowledge of the distant AS agreements;
- Enforces coherency and consistency in a QoS AS path with no knowledge of the complete chain of ASs.

4.2.3.5 The fundamental use case: the QoS Internet as a set of Meta-QoS-Class planes

In this section, we describe an Internet QoS model based on the Meta-QoS-Class concept. The purpose of this model is to build a QoS-enabled Internet, which keeps, as much as possible, the openness of the existing best effort Internet, and more precisely conforms to the requirements expressed above in4.2.3.2. In this model, the resulting Internet appears as a set of parallel Internets or Meta-QoS-Class planes. Each Internet is devoted to serve a single Meta-QoS-Class. Each Internet consists in all the l-QCs bound according to the same Meta-QoS-Class. When an l-QC maps several Meta-QoS-Classes it belongs to several Internets. The user can select the Internet that is the closest to his needs as long as there is currently a path available for the destination.

We assume that in a Meta-QoS-Class plane, because we want to stay close to the Internet paradigm, all paths were to a reasonable extent, born equal. Therefore, the problem of path selection amounts to: *Do your best to find one path, as best as you can, for the selected Meta-QoS-Class plane.* This sounds like the traditional routing system used by the Internet routers. Therefore we can rely on a BGP-like

Copyright © MESCAL Consortium, June 2005

protocol for the path selection process. By destination, q-BGP selects and advertises one path for each Meta-QoS-Class plane.

For a given Meta-QoS-Class plane, when there is no available path to a given destination, the only way for a datagram to travel to this destination is to use another Meta-QoS-Class plane. The only Meta-QoS-Class plane available for all destinations is the best-effort Meta-QoS-Class plane (also known as "the Internet"). There's no straightforward solution to change from one plane to another on the fly. So, there's no straightforward way to span a Meta-QoS-Class plane hole by a best-effort bridge.

This solution gives only loose administrative guarantees, however as long as all actors (especially, all service peers involved in the QoS AS path) do their job properly, the actual level of guarantee will be what is expected.

This solution stands only if l-QC "Meta-QoS-Class"-based binding is largely accepted and proceeded.

4.2.3.6 Proposal for a set of Meta-QoS-Classes

We propose to define five Meta-QoS-Classes:

- Premium Meta-QoS-Class
- Gold Meta-QoS-Class for TCP-friendly traffic
- Gold Meta-QoS-Class for non TCP-friendly traffic
- Best effort Meta-QoS-Class
- Cool Meta-QoS-Class

Below some examples of basic groupings which are given for the sake of clarification and not to recommend a particular configuration:

- Internet with the five Meta-QoS-Classes;
- Internet with only the first four Meta-QoS-Classes;
- Internet with only the last two Meta-QoS-Classes.

We define some parameters for each Meta-QoS-Class in the following sub-paragraphs. These parameters are: Targeted use, Performance, Constraint on the flows and Resources.

The values for the performance parameters have not been set yet. They should be derived from the knowledge of the application needs and the knowledge of the performances of the main Service Providers' networks.

4.2.3.6.1 Premium Meta-QoS-Class

- Targeted use: mission critical applications
- Mandatory QoS parameters are: loss, jitter and delay
- Performance: very low delay, very low jitter, no loss
- Constraint on the flows: some sort of admission control and possibly shaping to enforce the resource requirement.
- Resources: on each output interface, the traffic for the class is always much smaller than the bandwidth reserved for the class (EF based). The resources must always absorb the traffic with no loss even with bursted aggregates.

Copyright © MESCAL Consortium, June 2005

4.2.3.6.2 Gold Meta-QoS-Class (Two classes)

- Targeted use: sensitive applications split into two different classes TCP-friendly traffic and non TCP-friendly traffic. We differentiate two classes because since we allow diagram deletion a mix of TCP and non-TCP flows could put TCP flows at a disadvantage.
- Mandatory QoS parameters are: loss, jitter and delay
- Performance: low delay, low jitter, low loss
- Constraint on the flows: TCP friendly traffic for the TCP-friendly Class traffic.
- Resources: on each output interface, the traffic for the class can be greater than the bandwidth reserved for the class (AF based) the delta has a direct impact on the loss rate.

4.2.3.6.3 Best Effort Meta-QoS-Class

- Targeted use: current applications
- Performance: no guarantees however the measured values should not be too bad
- Constraint on the flows: no constraint
- Resources: the ratio resource for the class to traffic for the class must not be too small.

4.2.3.6.4 Cool Meta-QoS-Class

- Targeted use: any delay requirement applications
- Performance: no guarantees
- Constraint on the flows: services that don't care at all about delay (may be partly because very cheap)
- Resources: the resources reserved for this class must be very small compared to the other classes (included the traditional Best Effort). The ratio resource for the class to traffic for the class can be very small.

4.2.3.7 Next steps

4.2.3.7.1 Thorough definition

Some work should be undertaken to refine the definition of a Meta-QoS-Class. Some parameters should be more deeply investigated. For example: how exactly should a service be described? What are the sub-attributes? How should the performance characteristics be described? Do we need a parameter for availability? How to define it?

In the basic parameters we gave, a Meta-QoS-Class appears for a customer both as a way to convey a certain type of application (for instance video traffic) and as a way to get some guarantees in terms of one-way transit delay, one-way transit variation delay and Loss rate. It would be worthwhile to investigate in these two approaches and to decide whether we should privilege one of them or keep the two of them.

4.2.3.7.2 Standardising Meta-QoS-Classes

Each Service Provider must have the same understanding of what a given Meta-QoS-Class is about. A global agreement (a.k.a. standards) is needed. This agreement could be typically reached in an international standardisation body. There must be also a mean to certify the l-QC classification made by an AS conforms to the Meta-QoS-Class standards. So the Meta-QoS-Classes standardisation effort should go along with some investigation on conformance testing requirements.

Copyright © MESCAL Consortium, June 2005

4.2.3.7.3 Meta-QoS-Class outclassing procedures

Additional techniques should be investigated in order to allow a fructuous usage of the Meta-QoS-Class hierarchy such as the outclassing.

4.2.3.7.4 Security

Security is a main concern in a QoS-enabled Internet. Flows entering an AS and requesting QoS are likely to arrive from any AS and to be destined to any AS. So, it is of primary importance for a Service Provider to be able to filter the flows whose requests are not legitimate. Some investigations must be conducted in this direction. The Meta-QoS-Class concept opens the possibility of QoS services potentially reachable from any Internet position. Consequently, the menace of a spurious attack grows accordingly.

4.2.4 Global-QoS-Classes

Global-QoS-classes (g-QCs) are QoS-classes, where the values of the performance parameters (considered of 'offered' type, cf. Table 1) express the desired transfer requirements of widely deployed (globally known) services/applications. Typical examples of global-QoS-classes could be VoIP-QoS-class or High-Quality-Video-QoS-class.

For a given widely deployed service/application, a number of corresponding global-QoS-classes could be defined, depending on the nature (e.g. average, percentile) of the QoS-class attributes expressing transfer performance parameters.

Based on the QoS-class ordering relationship defined in section 4.2.2.2, global-QoS-classes can be arranged hierarchically, and multiple QoS-classes could adhere to a specific global-QoS-class.

Similar to the notion of Meta-QoS-Class, the notion of global-QoS-class may be useful for determining the values of QoS-classes as well as for providing the grounds for 'grouping' (mapping between, see below) QoS-classes of different Provider domains.

4.3 The MESCAL Internet QoS Service Model

Summarising the concepts and the notions presented in the previous section, the MESCAL model for Internet QoS-based services is shown in Figure 5.

Copyright © MESCAL Consortium, June 2005

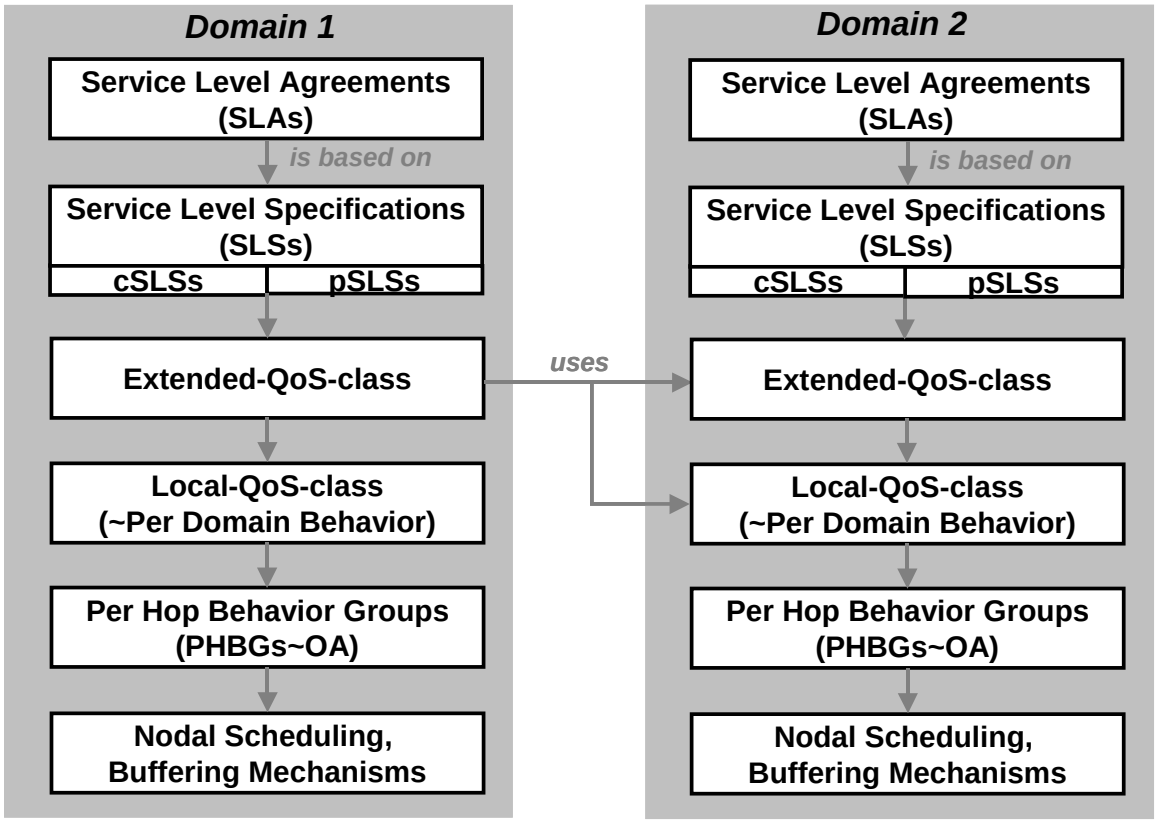


Figure 5: The MESCAL Internet QoS service model.

The MESCAL QoS service model is a layered peer model.

The essence of the model is the notion of QoS-class introduced in the previous section. Considering a provider domain, QoS-classes abstract the elementary nodal QoS enabling capabilities into sets of network-wide packet transfer capabilities, which are deemed appropriate to support the connectivity requirements of QoS-based services and applications. The notion of the QoS-class provides the necessary abstraction level for (a) building QoS-based services and (b) for linking service-peering provider domains to the end of expanding the geographical scope of their QoS-based services, independently of the underlying network-level capabilities, even technologies, employed in the different provider domains.

In particular, the layered aspect of the model refers to within a provider domain; through a 'is-based-on' relationship builds from the elementary nodal QoS enabling capabilities (IP DiffServ is assumed) to SLAs. The peering aspect of the model refers to between two provider domains; through a 'uses' relationship between QoS-classes (in the sense of section 4.2.2.4) allows different providers to combine their QoS transfer capabilities to the benefit of extending their QoS-based services beyond their geographical span.

4.4 Operations for Building Internet QoS-based Services

Following the concepts and notions of the proposed QoS-based service model, this section outlines suitable operations, called *QC-operations*, which need to be performed by provider domains to the end of building QoS-classes, and therefore corresponding QoS-based services, spanning beyond the reach of their domain (cf. extended-QoS-classes, section 4.2.2.1). It should be stressed that the purpose of QC-operations is to build extended-QoS-classes, not to actually provision –fulfil, assure- extended-QoS-classes. The identification of such operations is useful for a number of reasons:

- It puts the proposed concepts and notions into a sort of 'functional order', thus contributing to the validation of the model from functional perspectives.

Copyright © MESCAL Consortium, June 2005

- It contributes to the drawing of a functional architecture, per and across provider domains, for QoS-based service provisioning/delivery in the Internet. The identified operations should be reflected in appropriate functional blocks and/or protocol features.
- It introduces appropriate terminology against which different solutions for QoS-based service delivery in the Internet could be described and compared. Such different solutions may employ different means –functions, algorithms and protocols- in realising the identified QC-operations.

The following point is worth noting. QC-operations prompt for distinguishing and functionally decoupling the required functionality (traffic engineering and service management functionality) per provider domain for QoS-based service delivery, into intra- and inter-domain. QC-operations primarily imply inter-domain functionality as they target at building external-QoS-classes. Intra-domain type of functionality is mainly implied by the delivery of local-QoS-classes, which are taken for granted from the perspectives of QC-operations.

Considering a provider domain wishing to provide e-QCs onwards, from its domain to destinations outside its domain, the identified QC-operations are depicted in Figure 6 and described in the following sections.

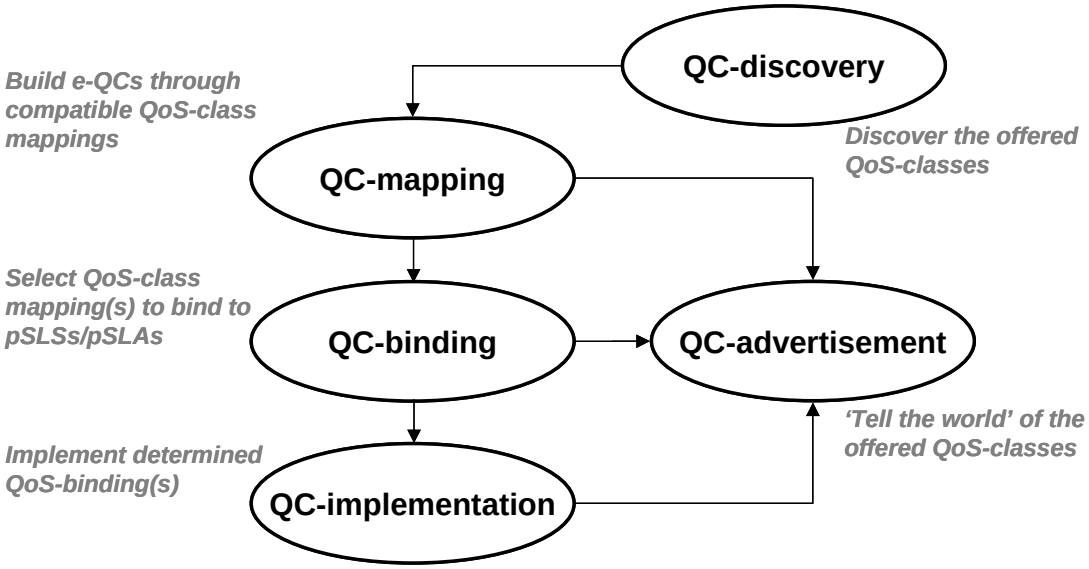


Figure 6: MESCAL QoS-class operations.

4.4.1 QC-advertisement

Through the QC-advertisement operation a provider domain informs other providers of its QoS-class capabilities. QoS-classes may be advertised at various levels as deemed appropriate by relevant policies of the provider. They may be advertised when are first conceived as a result of the marketing and service planning activities of the provider or during when the necessary actions for building them (agreements, configurations) are being taken or after they can be actually supported and provided.

As outlined in section 4.2.2.4, QoS-classes can be made known to other providers through (one or both of) the following two methods: by advertising corresponding QoS-based services (cf. pSLSs, section 4.2.1) and/or by appropriate capability announcement means. Without loss of generality, it is assumed that the advertised QoS-classes are of offered-QoS-class type (cf. section 4.2.2.3).

The means for other providers to actually use the QoS-classes advertised by a given provider as well as any other information deemed appropriate to accompany QoS-class information (e.g. topological scope constraints, corresponding Meta-QoS-Class) are assumed that they are conveyed as part of QC-advertisement method employed. Given a provider domain, the means for actually using the advertised QoS-classes should be such that they can be feasibly realised at a packet level through standard capabilities of the IP layer (see discussion in section 4.4.5).

Copyright © MESCAL Consortium, June 2005

4.4.2 QC-discovery

Through the QC-discovery operation a provider domain is able to locate and find out the QoS-classes offered by other provider domains. The discovery means should be in accordance to the means employed by providers to advertise the QoS-classes they offer.

4.4.3 QC-mapping

Through the QC-mapping operation a provider domain sees how to build extended-QoS-classes, that is QoS transfer capabilities with reach beyond its domain. This is done by determining suitable -according to the performance characteristics of the extended-QoS-class to be built- combinations of the domain's own capabilities (local-QoS-classes) with the QoS-class capabilities offered by other provider domains. The latter capabilities are made known through the QC-discovery operation (cf. section 4.4.2). The combinations might be based on any grounds of compatibility deemed appropriate by the provider domain to build the extended-QoS-class e.g. based on Meta-QoS-Classes equivalence or global-QoS-class conformance criteria. To this end, the QC-mapping operation may entail a *QC-classification* process, whereby a provider domain may classify its local-QoS-classes against widely accepted service categories e.g. Meta-QoS-Classes.

It should be noted that for an extended-QoS-class deemed necessary to be provided, a number of combinations could be potentially made. For example, this may be the case when the provider domain provides more than one local-QoS-class for the same Meta-QoS-Class. The QC-mapping operation determines a subset of the compatible combinations that could be possibly made. The term *QoS-mapping* is used to denote a 'compatible' QoS-class combination determined by the QC-mapping operation for building a particular extended-QoS-class capability.

The operation is primarily instigated by the business policies of the provider domain determining the performance characteristics of the extended-QoS-classes that need to be provided and various constraints regarding combination/service peering options.

The QC-mapping operation is denoted by the symbol '$\rightarrow$'.

4.4.4 QC-binding

As already outlined, the QC-mapping operation in a provider domain may result into a number of possible QoS-mappings for building a particular extended-QoS-class. In the general case, these mappings may involve a number of different local-QoS-classes each combined with a number of offered-QoS-classes from other -one or more- provider domains.

Through the QC-binding operation, a provider domain decides which of the possible QoS-mappings determined for building an extended-QoS-class will be used for actually providing this extended-QoS-class. The selection of using a QoS-mapping is substantiated by negotiating corresponding pSLSs/pSLAs with the provider of the offered-QoS-class pertinent to the QoS-mapping; thus 'binding' the local-QoS-class with the offered-QoS-class to the terms and conditions underlying the use of the offered-QoS-class. In other words, the QC-binding operation selects a subset of QoS-mappings to cast them into pSLSs/pSLAs with the corresponding service-peering providers. The term *QoS-binding* is used to denote a QoS-mapping for which a pSLS/pSLA with a service-peering provider has been established.

QoS-binding selection should take into account the provisioning requirements of the extended-QoS-class (e.g. in terms of maximum targeted bandwidth and cost) as well as the constraints underlying the use of the offered-QoS-classes as set by their providers (e.g. availability, cost). The latter constraints could be made available through the QC-advertisement operation. In any case, they are deemed as subjects of negotiation.

It should be noted that the QC-binding operation might result in a number of QoS-bindings for a given extended-QoS-class. QoS-bindings with the same service-peering provider may differ in the local-QoS-class and subsequently in the offered-QoS-class they use. Alternatively, QoS-bindings may differ when established with different service-peering providers. Providers may find such multiplicity

Copyright © MESCAL Consortium, June 2005

advantageous for avoiding to be bound to a specific QoS-capability of a particular service-peering provider and/or exploit the merits of dynamic, multi-path routing –note that different bindings imply different intra- and inter-domain routes in general.

Related to the above, the decision as to which of the established QoS-bindings will be *put in effect* in the network for actually implementing an extended-QoS-class as well as related routing/forwarding decisions fall into the realm of (inter-domain) traffic engineering. For instance, depending on the capabilities of the IP layer and corresponding policies, a provider domain may decide to put in effect only one of the determined bindings at a time, switching to another one should appropriate conditions warrant so. Or, a provider domain may decide to put in effect all determined bindings and employ a dynamic routing scheme with or without multi-path and load distribution features.

Once in the context of an extended-QoS-class the appropriate bindings have been determined, established with service-peering provider domains and effected in the network, the extended-QoS-class capability can actually be provided. The provider domain may make known this capability to other provider domains or customers by defining appropriate offered-QoS-classes and advertising them through the QC-advertisement operation.

The QC-binding operation is denoted by the symbol '⊕'.

4.4.5 QC-implementation

Through the QC-implementation operation, a provider domain implements at the network layer a QoS-binding. The operation encompasses only the necessary configurations at the IP layer required for the appropriate treatment of the packets. As stated in the previous section, routing and forwarding issues are outside the scope of the QC-operations.

Considering a provider domain offering a given QoS-class, which corresponds to an extended-QoS-class of the domain actually implemented through a particular QoS-binding, which in turn, by definition, involves a local-QoS-class of the domain and an offered-QoS-class of a service-peering domain, the QC-implementation operation encompasses the following aspects:

- Identification of the QoS-class according to which the packets entering the provider domain should be treated.
- Enforcement of the corresponding local-QoS-class in the provider domain.
- Enforcement of the use of the corresponding offered-QoS-class in the service-peering provider domain, which at the end corresponds to a local-QoS-class in that domain.

The above aspects should be realised based on the capabilities of the network layer –IP DiffServ/MPLS-capable routers are assumed.

Packets entering a provider domain are identified as belonging to an offered-QoS-class of the domain based on their IP header information. Similarly, the means for a provider domain to enforce the use of a QoS-binding-related QoS-class in a service-peering provider domain should be based on information contained in the IP header. For scalability reasons, the IP header information used for these purposes, should not be too fined-grained e.g. specific to customer contracts (cSLSs). Information facilitating traffic aggregation should be used e.g. DSCP.

Enforcement of local-QoS-classes is realised by mapping them (that is, the classified packets) to an OA (Ordered Aggregate). An OA corresponds to the notion of PHBG, the QoS building block of IP DiffServ domains, which prescribe to particular types of nodal packet treatment (EF, AF1-4, BE). At a packet level, an OA corresponds to specific information in the IP header, the so-called DSCP and is realised through appropriately configured scheduling (and buffering) mechanisms available at the network nodes. Note that different provider domains may use different DSCPs for the same OA.

The choice of OA per local-QoS-class should be made in accordance to the targets (bounds) of its performance parameters. When a provider domain institutes its local-QoS-classes, a set of possible OAs is associated with them for their implementation; the first OA denotes the most appropriate OA and the other OAs denote alternatives of superior performance e.g. EF could be set as an alternative of

Copyright © MESCAL Consortium, June 2005

an AF1 OA which is deemed the most appropriate to implement a local-QoS-class. The QC-implementation operation determines which of the associated OAs is 'best' to be used for enforcing the local-QoS-classes, according to actual network status and state conditions.

The above aspects may be realised through the classification and marking mechanisms prescribed by the IP DiffServ architecture, or by setting-up LSPs across domains or a combination of them. The actual realisation means are left to the individual solutions for Internet QoS-based service delivery.

Copyright © MESCAL Consortium, June 2005

5 INTER-DOMAIN QOS ISSUES

5.1 Introduction

This chapter discusses a number of issues that arise in Inter-domain QoS delivery. The topics addressed cover all aspects of the MESCAL project, including peering arrangements, service guarantees, traffic engineering, scalability and multicast. The objective is to provide background information on and to explore the intrinsic aspects of each topic. Later chapters discuss these issues in the context of a solution for delivering Inter-domain QoS.

5.2 Inter-domain Peering

5.2.1 Cascaded vs. Centralised Approach

Within the MESCAL project, two major approaches have been considered to establish a consistent set of inter-domain peering agreements in order to construct end-to-end QoS-based services across Internet at large scale:

- The cascaded approach where a provider only negotiates pSLSs with its immediate neighbouring provider/s to construct an end-to-end QoS service. With this approach, service peers are also BGP peers.
- The centralised approach where a provider negotiates directly with an appropriate number of downstream providers to construct the service. With this approach, service peers may not be BGP peers.

The following two sections provide a description of these two approaches. It should be noted that the type of inter-domain peering impacts the service negotiation procedures, the required signalling protocols, the QoS binding, and path selection.

5.2.1.1 The Cascaded Approach

In the cascaded approach, the QoS peering agreements are between BGP peers, but not between providers more than "one hop away". This type of peering agreement is used to provision the QoS connectivity from a customer/domain to reachable destinations when crossing several domains.

Figure 7 gives an overview of the operations in this approach. The domain AS5 supports an intra-domain QoS capability (l-QC1). AS4 supports an intra-domain QoS capability (l-QC2) and is a BGP peer of AS5. AS4 and AS5 negotiate a contract (pSLS3) that enables customers of AS4 to reach destinations in AS5 with a QoS (e-QC1). This process can be repeated recursively to enable AS3 to also reach destinations in AS4 and AS5, but at no point do AS3 and AS5 negotiate directly.

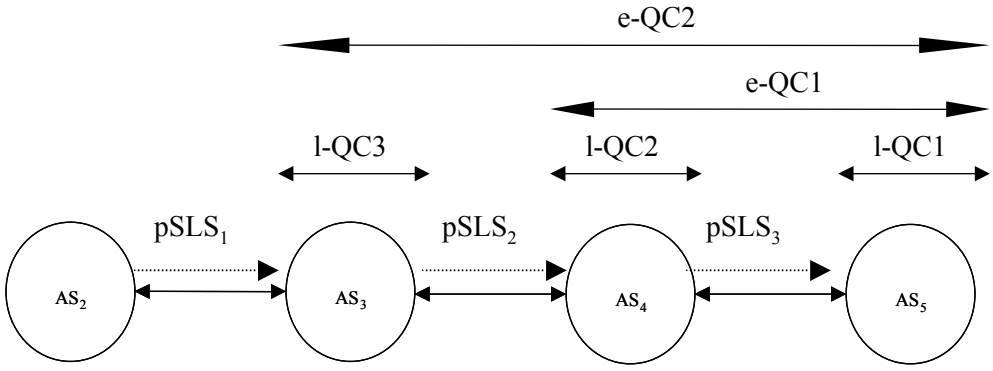


Figure 7: Cascaded Approach.

Copyright © MESCAL Consortium, June 2005

5.2.1.2 The Centralised Approach

The centralised approach disassociates pSLS negotiations from the existing BGP peering arrangements. The originating domain knows the end-to-end topology of the Internet and establishes pSLSs with a set of potential domains (neighbours, transit, and distant ASs) in order to reach a set of destinations, to offer end-to-end QoS-based services.

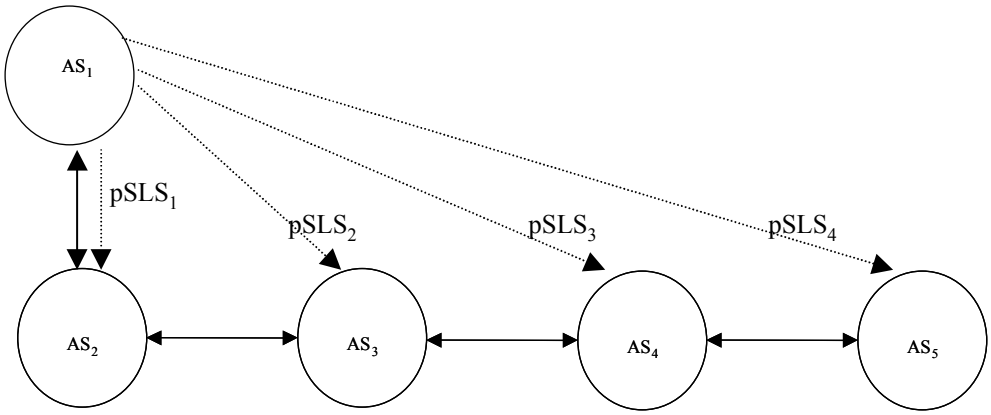


Figure 8: Centralised Approach.

The centralised approach presents an alternative to the cascaded approach providing a high degree of flexibility at the service negotiation level, but it may create deployment/scalability concerns.

Within the context of MESCAL project, we focus only on the Cascaded approach and the MESCAL solution presented in Chapter 7 is based on this approach.

5.2.2 Passive and On-demand Peering

5.2.2.1 Passive pSLS negotiation

The cascaded approach can be characterised as follows:

- The pSLS is only negotiated between two adjacent ASs, i.e. autonomous systems whose ASBR routers have established eBGP peering relationships,
- Services that are constructed by cascaded pSLSs are dependent on what has been negotiated in the downstream cascaded AS chain.

One of the concerns with the cascaded approach is that it is passive, insofar as an AS cannot directly control the QoS negotiation beyond its adjacent AS. This can be a problem if one of the ASs in the cascaded chain is not motivated to create a pSLS with a peering domain – perhaps it is not aware of the business opportunity – so the end-to-end cascaded chain cannot be created.

In order to address this problem, the pSLS On Demand is proposed and explained in the following section.

5.2.2.2 pSLS On Demand

The idea of pSLS On Demand is that an AS can request a target AS to establish a particular pSLS with one of its adjacent ASs. This mechanism assumes that the target AS can offer the desired QoS capability – perhaps it has already advertised the QoS offering – but it has not negotiated pSLSs based on its capabilities.

Figure 9 shows a scenario where it is not possible to build an end-to-end QoS agreement due to the lack of an appropriate pSLS between the AS_3 and AS_4.

Copyright © MESCAL Consortium, June 2005

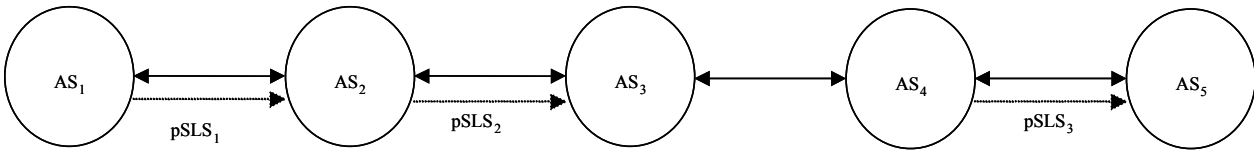


Figure 9: Passive pSLSs.

If AS_2 and/or AS_1 identify a business opportunity to build an end-to-end agreement to destinations in AS5, they can solicit AS_3 to establish the appropriate pSLS with AS4. The contents of the $pSLS_{21}$ contract include a pointer to the resulting contract $pSLS_o$.

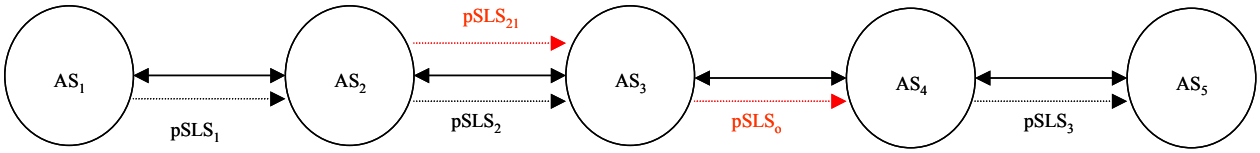


Figure 10: pSLS On Demand.

5.3 Inter-domain Service Guarantees

5.3.1 Inter-domain Service Options

It is possible to consider several service options that might be offered by inter-domain QoS-enabled IP networks, each with their particular requirements for QoS performance guarantees. For example:

- A service option that targets customers requiring differentiated network services whatever the destination of their traffic is within the domain. This service offering is only provided if the amount of the QoS traffic remains within certain limits compared to the rest of the best-effort traffic. Since the provider does not have the prior knowledge about the traffic destinations, the overall sum of flows must remain lower than the amount of resources provisioned (in the links and network elements) by the provider for this purpose. The dimensioning of the network is performed statistically and control of network resource usage may rely on monitoring. If the network dimensioning and control are performed appropriately, the end-user can expect that the QoS-enabled traffic sent in the network will reach its final destination with some loose QoS guarantees (better than best-effort). It should be noted that it would be difficult to provide strict bandwidth guarantees due to the statistical nature of the service.
- An alternative service option is dedicated to customers requiring strict QoS offering including end-to-end performance and bandwidth guarantees. To provide this service option, especially for the end-to-end bandwidth guarantee, it is mandatory to reserve the appropriate resources along the end-to-end path (over booking can also be an option) and to control the path. Traffic engineering techniques are normally used for this purpose. For example, in MPLS-enabled networks, end-to-end LSPs are established for which appropriate resources have been reserved.

The project must identify the types of inter-domain QoS Service that it will provide (see Section 7.2), as it will have a significant impact on the MESCAL solution.

5.3.2 Bandwidth Guarantees

End-to-end bandwidth guarantees can be provided to customers on an inter-domain basis but the issue requires careful consideration. For example, it is difficult to provide bandwidth guarantees to customers if the destination of the traffic is not known in advance (e.g., for services such as Internet access). However, it is possible to provide bandwidth guarantees if the destination of the traffic is known in advance, as for services such as VPNs.

Copyright © MESCAL Consortium, June 2005

5.4 Inter-domain Traffic Engineering

Traffic engineering is the means to optimise the use of available resources. Such optimisation inevitably involves the control of *outgoing, incoming* and *within the network* traffic flow. The first two are collectively regarded as inter-domain traffic engineering, while the latter as intra-domain.

The following operations are considered as Traffic Engineering (TE) issues:

- Define, provision and control local QCs (*l-QCs)*
- Reduce high variance in link bandwidth utilisation per QC
- Control of the outgoing/incoming traffic
  - Balance the traffic among external links
  - Prefer some links over the others

5.4.1 Peer Provider Selection problem

One problem that a provider is facing is the choice of adjacent ASs with which pSLSs will be negotiated. We name this problem as Peer Provider Selection.

The criteria for the selection are:

- The advertised QCs from the various ASs
- Economic criteria
  - Cost of link,
  - Cost of traffic, i.e. bandwidth per QC
- Business-oriented constraints
- The advertised network reachability information

This is in fact a cost optimisation problem. Note that the result of this selection will be more than one pSLSs for the same o-QC technical (traffic engineering) and economical reasons. These will be utilised later for load balancing.

5.4.2 Controlling the Outgoing Traffic

Load sharing is an important part of traffic engineering because it allows the traffic to spread among different paths and different classes and thus achieve better resource utilisation. We can achieve the maximum utilisation of the inter-domain resources by controlling the outgoing inter-domain traffic.

Note that the problem of optimising the utilisation of network resources requires controlling both the inter- and intra-domain resources. In this section the focus is on the inter-domain issues, but we will also elaborate on intra-domain resource control wherever is appropriate.

5.4.2.1 Load sharing based on different destination prefixes

Being able to control the load of the egress links and pSLSs on the granularity of different routes towards different destination address prefixes is an issue to be solved by traffic engineering techniques. This is to control the outgoing traffic on a per prefix basis so as to optimise the use of the egress resources.

In this simplest form of load balancing scenario we assume a single egress point, a single egress link, and a single egress pSLS that will be used to route traffic towards a specific destination prefix. Then for offering a particular o-QC we must balance the load between e-QCs (i.e., QC-binding to be put in effect)), which satisfy the requirements of the o-QC, i.e. choose an egress link/pSLS, for each destination prefix based on the different pSLSs. That means we do not require multiple simultaneous paths to be in effect for the same destination.

Copyright © MESCAL Consortium, June 2005

This exit point selection can be formulated and solved as an optimisation problem, where the objective is to optimise the utilisation of each egress pSLSs.

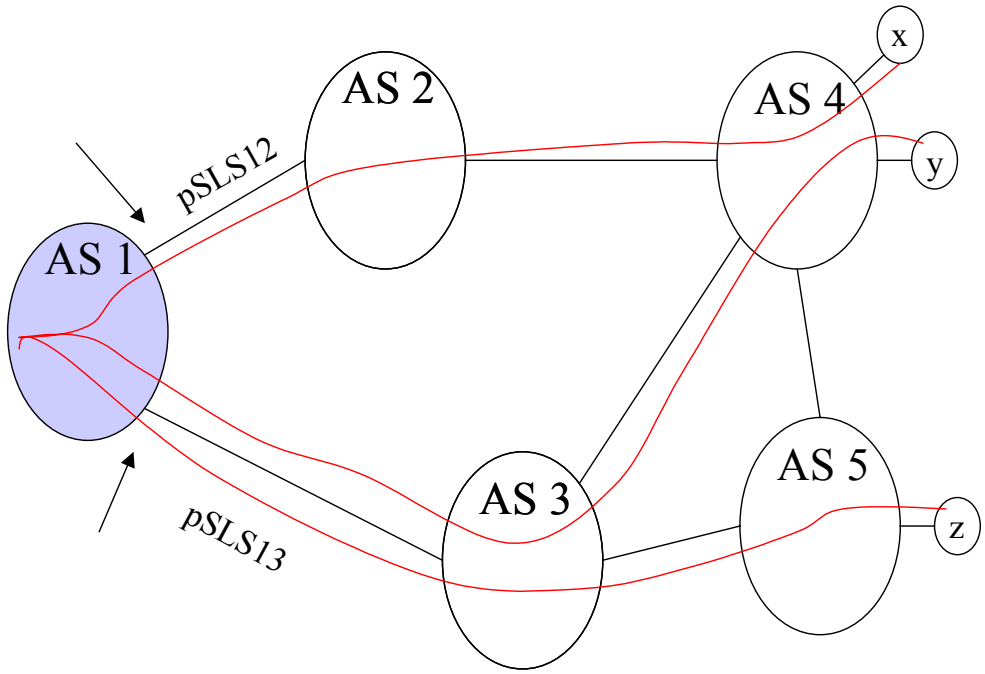


Figure 11: Balancing based on different destination prefixes.

In the simple example shown in Figure 11, we assume that pSLS12 and pSLS13 are compliant with the same o-QC and that pSLS12 has half the bandwidth of pSLS13, thus from the AS1 point of view it needs to route twice as much traffic through AS1-AS3 link as the AS1-AS2 link. In this example, the simplest way to achieve this splitting ratio is to route traffic for the prefixes towards the AS3 twice as much as the AS2. The enforcement of such load sharing decision can be done with the enforcement of specific routing policies either by fixing the path or introducing policy rules to dynamic routing protocols.

5.4.2.2 Multi-path load balancing for the same destination prefix

Load balancing on a per destination prefix basis only is not very flexible and thus the resulting engineering solution may overload one or another egress pSLSs while others are under-utilised. This situation can be improved by allowing more than one path towards the same destination prefix. In the following we will describe all possible load-balancing scenarios that can be taken when we have multiple paths towards a destination. The discussion will include intra-domain load balancing actions which are not tailored only to multi-path load balancing described in this section, but some of them may also be used in conjunction with the single path load sharing case discussed in the previous subsection.

Copyright © MESCAL Consortium, June 2005

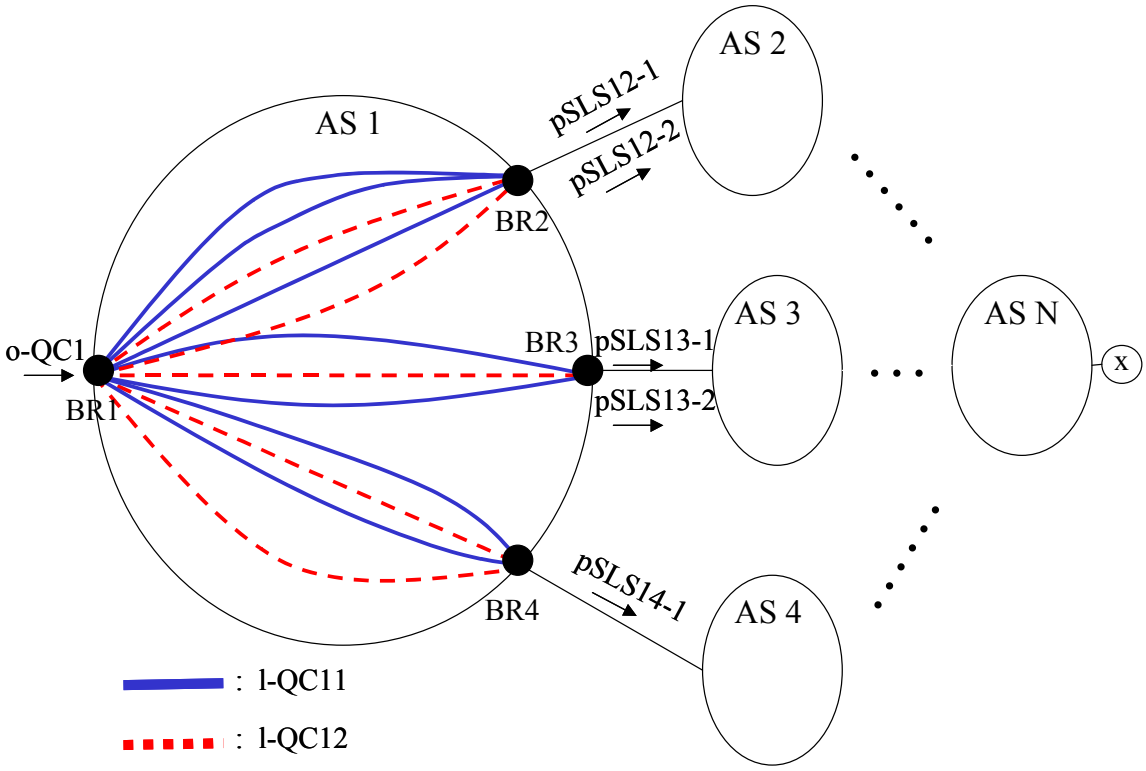


Figure 12: Load balancing possibilities (example 1).

In MESCAL, we can identify multiple levels of load balancing. In order to offer an o-QC1, we may have multiple combinations (QC bindings), which achieve the required performance characteristics of that o-QC. Among all these possibilities we have to decide:

1. Statically (*offline*) which QC-bindings and routes, i.e. e-QCs, are going to in effect for offering that o-QC (a reminder here that the o-QC is offered to an upstream AS *via* the agreement in a pSLS) so as to optimise the utilisation of resources.
2. If the previous case chooses to have more than one alternative bindings and routes in effect, then dynamically (*online*), based on measurements, we can decide for each flow which of the alternatives to use, so as to optimise the use of resources.

Static load sharing is required so that we can configure the control mechanisms in order to enforce the traffic engineering decision. The timescale is that of the Resource Provisioning Cycle (RPC) of the AS. Dynamic load sharing is required when we want to more accurately reflect on the traffic fluctuations. Note that dynamic TE is not at the per-packet timescale but rather on a per-flow or multiple flows in order of minutes.

With refer to Figure 12 and assuming that all the alternatives shown there are compatible, i.e. *as good as*, for offering o-QC1, load balancing (both *offline* and *online*) can be applied at multiple levels:

- Choosing the egress point, e.g. choosing one of the BR2, BR3, or BR4, and the egress link (see Figure 12).
- Choosing between the (potentially) multiple pSLSs, e.g. between pSLS12-1, pSLS12-2 (of course all the o-QCs of the adjacent ASs included in the pSLSs bound with the l-QCs must be “at least as good” as the offered o-QC).
- Choosing between the Local QCs (*l-QCs)*, e.g. between *l-QC11* and *l-QC12*
- Choosing between the potentially multiple paths of the chosen *l-QC*.

Note that some of the above options may not be available. As we move from static to dynamic load balancing the options may be reduced.

Copyright © MESCAL Consortium, June 2005

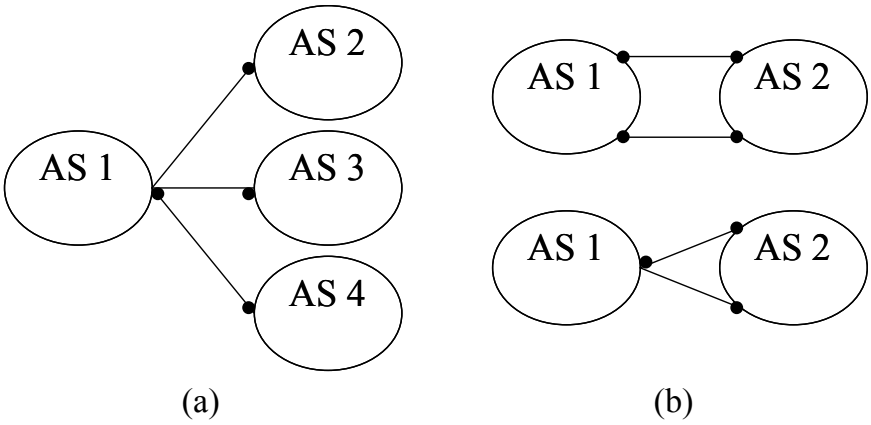


Figure 13: Choosing egress point or next-hop AS different from choosing link.

The first point in inter-domain load balancing deserves a little bit further discussion. The selection of the egress point does not mean that we necessary choose the next-hop AS and by choosing the next-hop AS doesn't mean that we choose the egress link. For example, as shown in Figure 13, in (a) when we have multiple peering at the same egress router (usually the case of multi-homed domains [Rekht95]) by choosing the egress point does not necessarily mean that we choose the next hop AS, and in (b) when we choose the next hop AS does not necessarily mean we choose the output egress link. It should be noted that in the case (a) the egress node have multiple interfaces each connected to an AS via interconnection links and in the second case (b) the AS connected to the next hop AS via multiple interfaces and interconnection links.

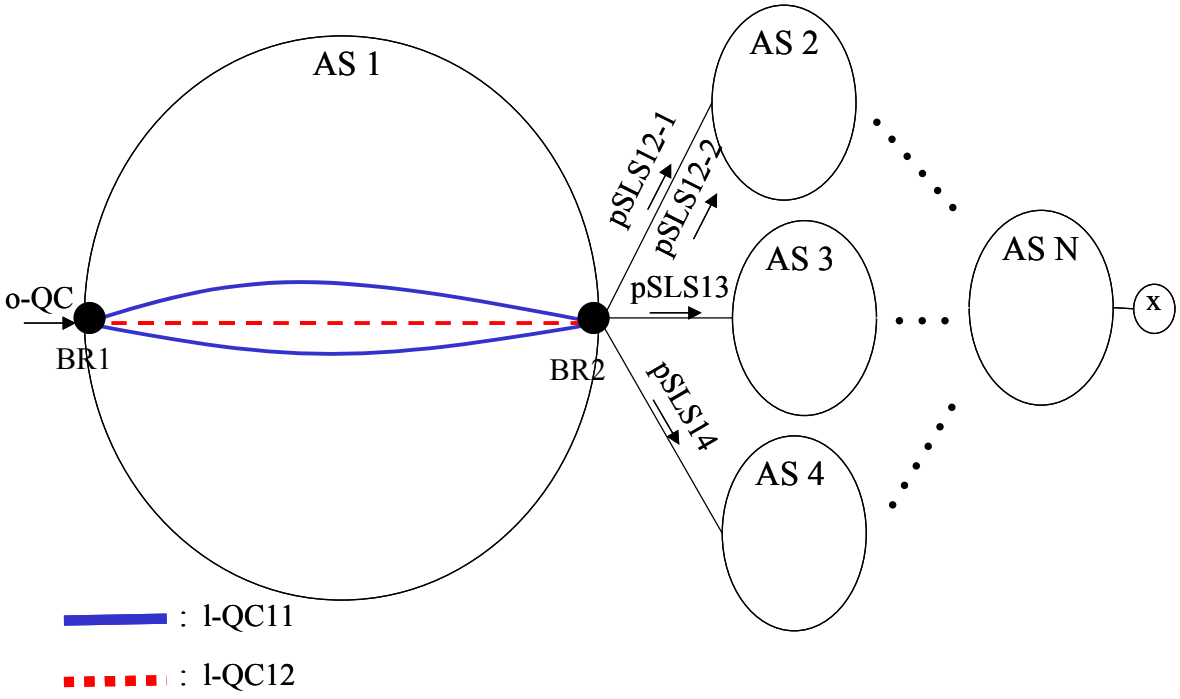


Figure 14: Load Balancing possibilities example 2.

In Figure 14, we can see that the inter-domain load balancing does not include the choice of the egress point, but includes the choice of the egress link (i.e., an output interface in the egress point/node) and the choice of the pSLS to be used. This is the scenario where an AS is multi-homed. The load balancing in this case is: a) among the interconnection links i.e., AS1-AS2, AS1-AS3, AS1-AS4, b) between the pSLSs of the chosen link, e.g. pSLS12-1, pSLS12-2, c) between the multiple l-QCs that can be bound with external o-QCs so to define multiple e-QC to comply with the offering the same o-

Copyright © MESCAL Consortium, June 2005

QC, and d) between the (potentially) multiple internal paths of the chosen l-QC. Combinations of situations as in both example 1 and example 2 may also exist.

5.4.3 Routing Aspects

The means to implement the various TE decisions is to control the routing. Even if we do not allow for balancing all the traffic as described in the previous section, routing has to be controlled in order to adhere to the QCs, both internally (l-QCs) and externally (pSLSs, i.e. o-QCs).

In this section, we will describe the requirements and the possible implementation mechanisms for controlling the inter-domain aspects of routing. Controlling intra-domain routing in order to achieve certain objectives is a very important issue, which has been studied extensively in the past [TEQUI] [Fortz00] and is not the main focus of MESCAL.

In the cascaded approach, inter-domain routing has the following aspects:

- Choose the egress point from the AS
- Choose the next hop AS
    - Choose among the possible links

5.4.3.1 Requirements for the inter-domain route selection process

There are three requirements from inter-domain traffic engineering:

1. It must be QC-aware
2. It must be constrained by the pSLS agreements
3. It should support load balancing capabilities for different destination prefixes

The first requirement says that the routing decision should be aware of the fact that the traffic that will be routed based on a particular QC. Thus the decision for routing may be different, and effectively is to have a different routing policy for each of the supported QCs. Being able to differentiate the traffic flow between different QCs is very important for the performance of the end-to-end QC. This requirement includes another important aspect that we need to "inject" somehow the o-QC information into the routing information distribution process. For example, if BGP is the protocol used to distribute the routing information, then we need to have the appropriate attributes for disseminating the QC information, which will be processed by the BGP peers.

The second requirement states that the possible egress points for specific QCs are only the ones for which we have agreed some pSLS with a downstream peering AS. This means that even if we have classical NLRI information (i.e. for best effort traffic) through some peering AS, we cannot use that AS as the next domain for QoS traffic, if we do not have a pSLS for a using particular external o-QC. A consequence of this requirement is that each time we agree on a new pSLS with a downstream peer we need to make this information available to the route selection process.

The third requirement reflects the discussion of section 5.4 on balancing the load over the multiple egress points in order to avoid overloading some of them, while others are under-loaded. This balancing is performed over *different destination prefixes*.

As a secondary requirement for routing:

4. It should support load balancing over multiple egress paths (as described in the previous section) for the *same destination prefix*.

Although the fourth requirement is important when we want to perform traffic engineering, we leave it as a "should", indicating that it is important but not a mandatory feature. Ideally we would like to have the flexibility to perform load balancing over non-equal cost paths with non-equal sharing ratios, but, if this is not easy to support from the implementation point of view, then we can make use of equal-cost traffic splitting. The exact load balancing capabilities are of great importance when we are to devise the Traffic Engineering algorithms. It is envisaged that there will be a trade-off between the

Copyright © MESCAL Consortium, June 2005

additional required protocol changes and the flexibility and optimality that can be achieved by the traffic engineering processes.

The above can be realised through static or dynamic routing schemes (cf. Discussion in Section 5.4.3.3).

5.4.3.2 Propagation of Inter-domain QC routing information

The proposal in MESCAL is to keep BGP as the basic means for propagating Network Level Reachability Information (NLRI) on per QC basis. The basis for this work will be the IETF initiative for defining the QOS_NLRI [Crist02] together with the work which allows the advertisement of multiple routes towards the same destination prefix [Walton02].

BGP provides the means to influence to whom and from which ingress points the routing updates are to be sent/received by filtering the updates according to some policies. These policies will be enforced so that to advertise the QoS reachability only to peers with which we have agreed on with some pSLS.

The appropriate attributes, which describe the QCs must be defined and included in the advertisements. This may be a source of some scalability concerns since it will lead to increase the routing table size depending on the number of the supported QoS classes, which in the MESCAL solution is bound by the number of DSCPs i.e. 64.

The QC information is propagated by BGP whenever a pSLS is agreed. One can allow for the QC information to change more dynamically, e.g. at each Resource Provisioning Cycle (RPC), in order to achieve some kind of QC performance monitoring. Although this feature may be quite useful for the engineering (e.g. load balancing) of upstream ASs, it may constitute sources of instabilities. This option has to be examined in greater detail in the context of the project to assess the potential instabilities.

The implementation overhead is related to the definition and manipulation of the attribute to carry the QC information. Note that modifying a BGP route selection process may be risky as the BGP Finite State Machine (FSM) may be affected accordingly.

5.4.3.3 Enforcing the inter-domain routing control policies

The basic requirements of inter-domain traffic engineering can be met with either fixed or dynamic path routing solutions. As stated previously, in all cases BGP is used, for the dissemination of QC-aware NLRI information.

The route selection algorithm should take into account the QC information and it should perform load balancing on the exit links (and pSLSs) for traffic destined to different prefixes. The latter is the first case for load balancing as described in section 5.4.2. Load balancing over multiple paths to the same destination prefix is an extra non-mandatory feature.

5.4.3.3.1 Fixed path routing

Enforcing inter-domain traffic engineering policy for statically fixing the path can be implemented in two methods:

1. If we assume the IGP-EGP model enforcing the fixed path routing, it means just to add the routing information into the BGP, whenever a pSLS is agreed (including renegotiations). We have to inject the new route and enforce the appropriate policies so that to advertise only the selected routes. In addition the BGP route selection algorithm has to be overwritten to ignore route changes advertisements (i.e. fixing the path).
2. The second method is to implement the fixed path routing decisions by using the mechanism proposed by CISCO in the Internet Draft, “Inter-AS MPLS Traffic Engineering” [Vasse03]. Here, we describe the use of “Scenario 1: Per AS Traffic Engineering Path Computation” solution as described in the draft. The use of “scenario 2: Path Computation Server” solution is for further study.

Copyright © MESCAL Consortium, June 2005

In the solution of Scenario 1, there must be the support of inter-AS TE paths, spanning more than one domain. In some cases, this solution can be used to support the establishment of end-to-end TE paths. The MESCAL solutions described in Chapter 7 are based on the cascaded approach, which require service peering relationships *only* between adjacent domains. Thus, we will not take advantage of the full spectrum of the Cisco's solution capabilities, but rather we use the mechanism specified in the Scenario 1 where we set-up TE paths towards and up to the first adjacent AS. This is shown in Figure 15, where the TE paths are set-up to the first ASBR of the adjacent AS. Note that although the Label Switched Paths (LSP) are set-up to a ASBR2 and ASBR4, this does not impose any administrative problems since the label switching operation can stop at egress ASBR, i.e. ASBR1 and ASBR3, due to the penultimate hop label popping feature of MPLS.

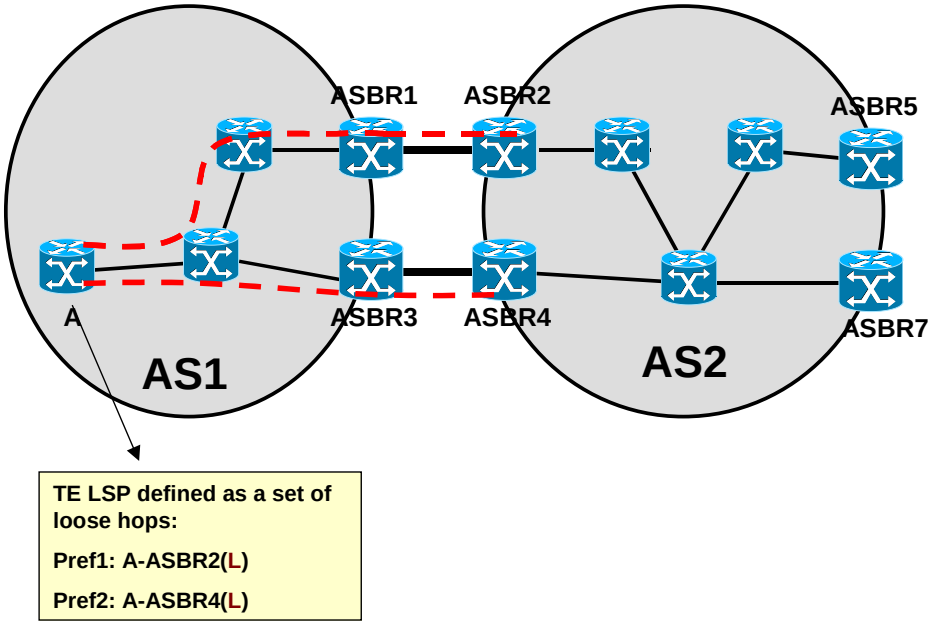


Figure 15: Facilitating the CISCO inter-AS solution scenario 1 proposal.

The CISCO solution proposes to flood the TE information related to the ASBR-ASBR link(s) even though there is no IGP enabled over those links. This allows the TE DB (Data Base) in each router to include TE information (TE metric, bandwidth, etc.) for the ASBR-ASBR links and thus to the potential head-end Label Switched Routers (LSRs). Since it is required for the inter-domain traffic engineering to be QC-aware, this means that the TE information must be on a per class of service, e.g. per {TA}PSC according to the definition provided by [LeFau03a] where TA is Traffic Aggregate and PSC is PHB Scheduling Class..

There are three important considerations in order to use this mechanism for enforcing the inter-domain TE decisions. We need to be able to manipulate the per-QC TE information (metric) of the ASBR-ASBR links flooded by the link state IGP. The second consideration is that we may need to change or overwrite the result of the CSPF (Constraint-based Shortest Path First) algorithm for computing the TE route. The final consideration has to do with implementation of the load balancing over multiple paths. This is supported since the solution allows for the definition of multiple paths, even for the same destination address prefix, and the ability to map traffic onto the multiple paths [Zhang03].

5.4.3.3.2 Dynamic path routing

Static path routing can support the required functionality in order to enforce the inter-domain TE and QoS policies discussed in this document. However, such routing schemes are hindered by lack of adaptability to changing topological and/or load changes and the restricted potential in achieving load-balanced states and thus optimising network utilisation, as compared to dynamic and/or multi-path

Copyright © MESCAL Consortium, June 2005

routing schemes. The following questions should better be addressed by the employed routing scheme: - how to learn QoS path failures? - and how to respond to such failures?

A dynamic inter-domain routing protocol, i.e. BGP, extended to convey QC-related information (we name this protocol as q-BGP), can be used to answer the above questions, since link failure detection is an implicit capability of IP routing protocols like BGP. QC-related information can be conveyed by the BGP UPDATE messages based on the results of the engineering processes of a RPC. It is expected that the supported QCs in one AS will not differ considerably from one RPC to the next, and thus the information injected into BGP will not change considerably.

In general, the changes to BGP will be similar to the first case (see section 5.4.3.3.1) for implementing the fixed path routing approach. The BGP path selection algorithm needs to change in order to take into account the QC information. The details of the q-BGP extensions mentioned above and the potential instabilities of the dynamic routing behaviour will be studied in the course of the project, and the starting point will be the initiative for the QOS_NLRI [Crist02] attribute, which is able to convey QoS information between domains and the work which allows the advertisement of multiple routes towards the same destination prefix [Walton02]..

5.5 QoS Issues

5.5.1 The "QC splitting" Problem

Figure 16 shows an example where pSLSs have been established between adjacent domains allowing each domain to send QoS-enabled traffic to its peering partner for crossing the Internet. Different l-QCs have been defined and deployed within each domain.

Users C1 and C2 requested red and blue e-QCs where each ordered set of l-QCs (red and blue) represents an e-QC i.e., red e-QC: (QC12, QC21, QC54, QC74) and blue e-QC: (QC11, QC21, QC51, QC71). If the DSCP field in the IP packet header is used for QoS-signalling across domains, both red and blue e-QCs are mapped to QC21 at AS2. At the AS2 egress point/s, we will encounter a splitting problem in that it will not be possible to distinguish the red and blue packets, based on DSCP values, so as to re-mark them with their individual DSCP values for onward transmission.

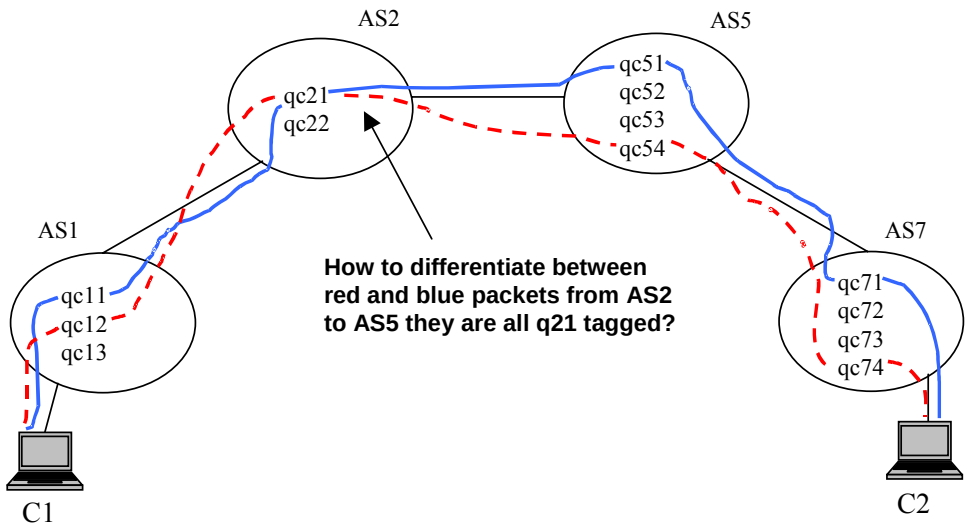


Figure 16: The QC splitting.

The QC splitting problem arises when a provider binds more than one o-QC of a service peering domain to one of its l-QCs. The issue is that:

Copyright © MESCAL Consortium, June 2005

"What should be and how to determine the appropriate DSCP marking for the datagrams forwarded to AS5?"

Any inter-domain QoS solution must overcome the QC splitting problem, while controlling the amount of state information that must be stored at each ASBR. To solve the splitting problem, the egress point at the AS could use one of the following mechanisms:

- By using full-set or sub-set of the 5-tuple (source and destination IP addresses, protocol, source and destination ports) and the DSCP to be used in the AS. Building such a list/table would assume that all cascaded ASs should know about c/pSLSs (at the aggregated level) and/or the QoS classes supported by the neighbouring domains.
- By employing a source route descriptor, embedded in the IP packet (*e.g.* IP source route options), which would explicitly state the ordered set of DSCP. This descriptor would be populated by the source or by a device close to the source.
- By using virtual QCs, i.e. map the flows belonging to the red and blue e-QC to different DSCPs in order to avoid the splitting problem. The treatment of both flows within the AS must be the same i.e., the same PHB and intra-domain route will be used for both.

5.5.2 IPv6 Issues

It is an objective of the MESCAL project that its solution should be applicable to both IPv4 and IPv6 networks. This is facilitated by the common approach to DSCPs for example, as discussed below. However, there are some IPv6 features mentioned below that can be exploited to enhance further the proposed solution.

The definition of QoS (cf. DiffServ) has been integrated in the specification of the IPv6 protocol. [RFC2460] defines the 8-bits field called "Traffic Class" allowing services differentiation as defined in [RFC2474] of Ipv4. This field, commonly known as the DiffServ (DS) byte, is composed of two parts like in IPv4 (DSCP and the two ECN bits). Therefore, the 8-bit Traffic Class field in the IPv6 header is to identify and distinguish between different classes or priorities of IPv6 datagrams.

In other words, the Traffic Class field in the IPv6 header is intended to allow similar functionality to be supported in IPv6 as in IPv4 DS field bits.

IPv6 facilitates traffic engineering approaches that are not possible with IPv4. For example,

- Exploiting of the Flow Label field: The Flow Label field is a 20-bit field included in every IPv6 datagram header. Datagrams are labelled by the source to identify a flow. An intermediate router can use this value to apply a specific treatment to the datagrams. To enable flow-specific treatment, flow state needs to be established along the path from the source to the destination. Within the context of the proposed MESCAL solution, this field could find an interesting applicability. For instance, one possibility offered by the Flow Label could consist in using it as an extended DSCP field using 20-bit length.
- Defining extension header(s): In IPv6, optional network-layer information is encoded in separate headers that may be placed between the IPv6 header and the upper-layer header of a datagram. There are a small number of such extension headers, each identified by a distinct Next Header value. New headers can be defined in order to implement a new service or option, without modifying the core IPv6 protocol specification. Regarding inter-domain QoS, some information exchanges or some mechanisms could take advantage of this IPv6 feature.

5.5.3 Ingress/Egress Conditioning

Solutions for inter-domain QoS that require complex traffic conditioning at ingress/egress points need to be aware of the capabilities/limitations of the high speed ASBRs. Otherwise, QoS solutions may place more functional demands on these routers that cannot be feasibly sustained. MESCAL will take this factor into account when assessing its solution.

Copyright © MESCAL Consortium, June 2005

When crossing multiple domains, each flow must be treated based on the l-QCs selected for that flow in each domain. To the end of eliminating scalability problems (see discussion in next section), aggregate-level information contained in the IP header, notably the DSCP field, should be used for the QC-signalling. It may be necessary to re-mark the packet's DSCP at the ingress point of the AS to the appropriate DSCP value (l-QC) and re-mark to the another DSCP value at the egress point of AS. Figure 17 shows this operation, where packets arriving at the transit domain with DSCP1 are re-marked with DSCP45 for transit and re-marked to DSCP1 at the egress.

Current routers are capable of re-marking DSCPs and/or performing traffic conditioning at ingress/egress interfaces but on high-speed interfaces, the process for traffic conditioning functionality must be simplified. .

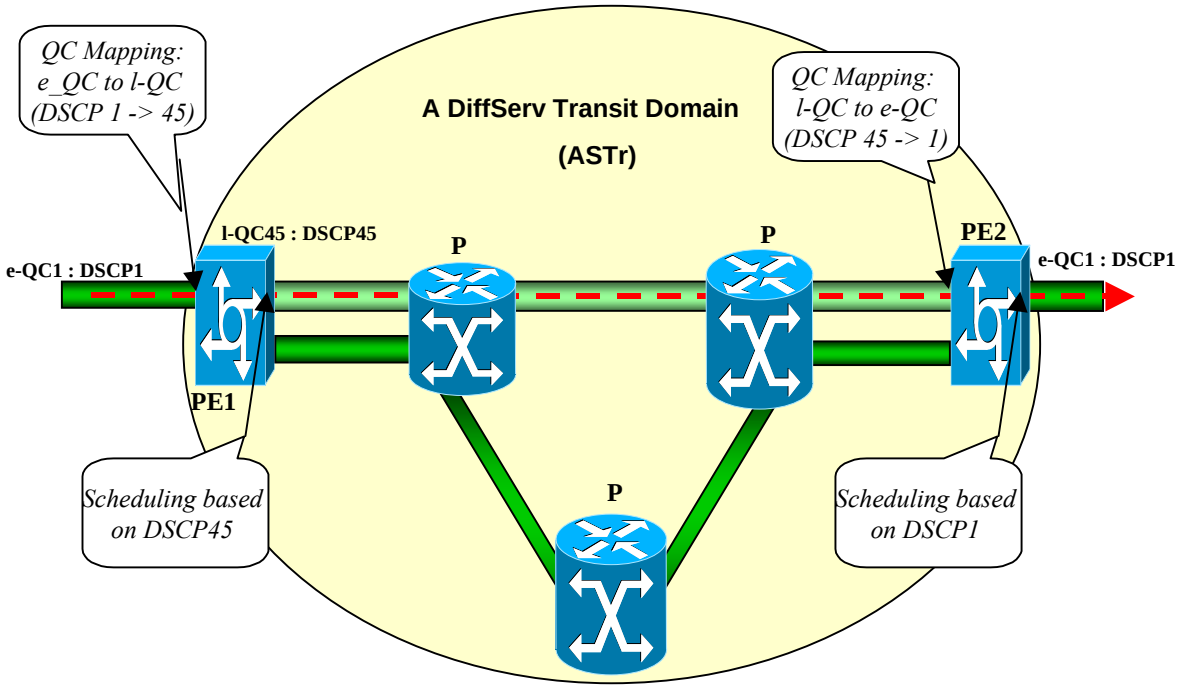


Figure 17: Ingress/Egress traffic conditioning.

5.6 Scalability & Complexity Issues

5.6.1 QC Implementation Issues

With the QC enforcement we mean the process of implementing QC-bindings (cf. Section 4.4.5) – classifying, enforcing, and forwarding of QoS-enabled packet to the correct paths (IP routes or LSP paths), as appropriate to the o-QC treatment that these packets should receive per domain. QC enforcement takes place at the data-plane after the c/pSLSs have been established. Specifically, QC-bindings are realised by downloading appropriate information for setting up the traffic classification and marking mechanisms of the DiffServ-capable routers and QoS-based packet forwarding is performed by accessing the QoS-based routing tables. QC-signalling is performed across all domains using DSCPs/MPLS-EXPs.

The following section explores different options regarding the use of IP header information in realising QC enforcement. These options justify the expected trade-off between increased flexibility in implementing QC-bindings and corresponding forwarding decisions, and per-packet processing overhead in the routers.

Copyright © MESCAL Consortium, June 2005

5.6.1.1 QC Implementation in MPLS-Based Networks

In tunnel-based solutions such as MPLS, the process of QoS-based packet routing must take place at the head-end of the tunnels. Provision must also be made at ingress boundary routers for QC enforcement when inter-AS MPLS is used. QoS-based routing is to direct specific traffic to the specific tunnel. In addition, traffic belonging to MPLS tunnels should receive different PHB treatment along the tunnel path depending on their QCs. The MPLS-EXP is the only visible field along the path to be used for service differentiation and for directing each tunnel's traffic to specific queues/schedulers. The issue is whether a unified set of EXP definitions is used across all domains or there is a need to remark the EXP at the ingress point of each AS.

5.6.1.2 QC Implementation in IP-Based Networks - Scenarios

Routing protocol should normally provide information for packet forwarding by taking into account the packet's associated l-QC. But before inter-domain QoS packet forwarding occurs, the packet's DSCP must be mapped and set to an appropriate value. Both IGP and EGP protocols for routing purposes should be QC-aware. In the following scenarios, we take into account the actions required at the AS ingress boundary routers for QC implementation - QC enforcement and packet forwarding.

In Figure 18 to Figure 21, each customer network administratively belongs to its directly connected ISP/AS (e.g., N1 to ISP/AS1). It is assumed that unique l-QCs are used in each domain. Thus, two distinct flows originated from two different sources within an AS (e.g., N0 and N1 in Figure 18) using the same e-QC and destined for a destination AS, will exactly traverse the same AS path along the route to the destination.

IP address aggregation at the network-level (i.e., network prefix/length tuple) is used in the QC enforcement scenarios described in the following sections, in order to prevent to work with individual IP addresses. Further address aggregation within an AS is possible if the network addresses belong to the AS is aggregated to a higher level of aggregation and resulted to another address tuple (i.e., prefix/length). The prefix/length tuple is available to BGP routers that can be used for QC enforcement.

5.6.1.2.1 Unified QoS Classes (l-QCs) Across All Domains (Scenario 1)

In this scenario, each l-QC has the same DSCP in every domain. While this has benefit of requiring very little ingress/egress conditioning at the domain boundaries, except at the customer ingress point, the scenario is neither realistic nor flexible. It is unlikely that network operators would agree to such constrained l-QC/DSCP mapping. Additionally, the binding flexibility is severely constrained. For example, in Figure 18, QoS binding cannot be achieved between different QCs (l-QC10 and l-QC11).

QoS-based packet forwarding need to be performed based on source address, destination address and DSCP. Source inspection for packet forwarding is required because traffic from different sources going to the same destination may transit different paths based on the e-QCs.

Copyright © MESCAL Consortium, June 2005

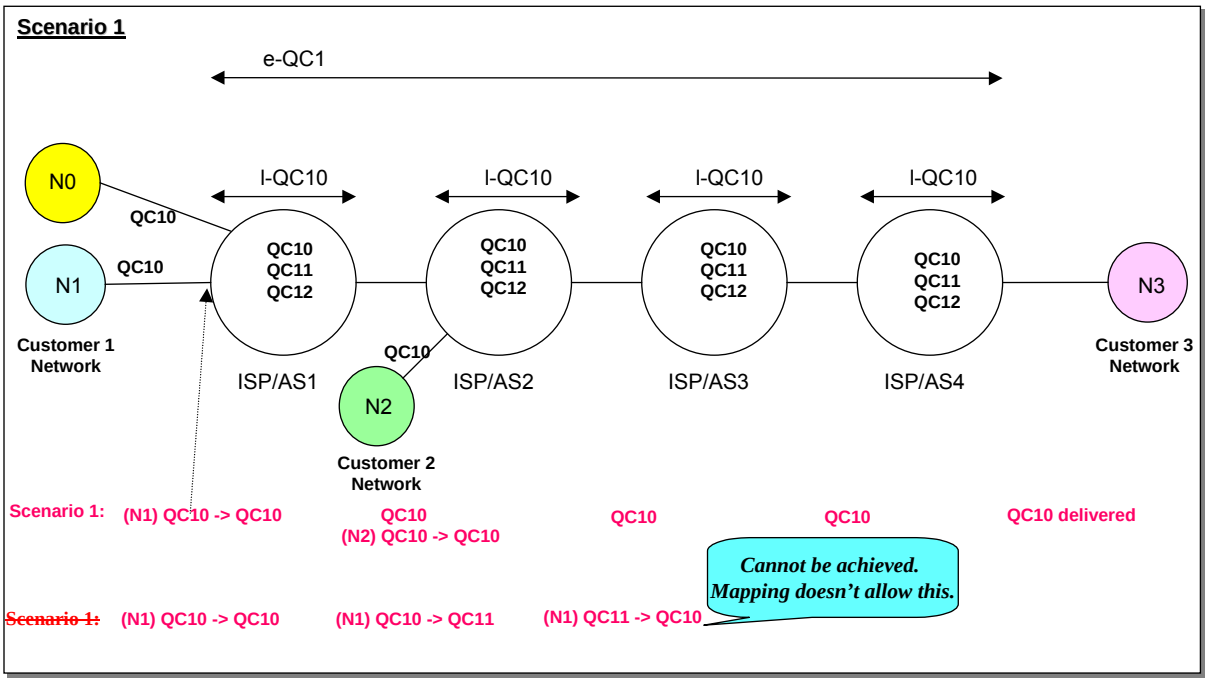


Figure 18: QC Implementation in IP-Based Networks - Scenario 1.

5.6.1.2.2 Direct l-QC Mapping (Scenario 2)

Scenario 2 differs from Scenario 1 in that DSCP manipulation is introduced at the domain ingress nodes. Each domain uses its own QoS definition using DSCP to differentiate them. Even with the introduction of this function, the binding we are still very constrained.

QoS binding is restricted as in Scenario 1 as shown in Figure 18. QC enforcement is performed on direct l-QC mapping and QoS-based packet forwarding is similar to Scenario 1.

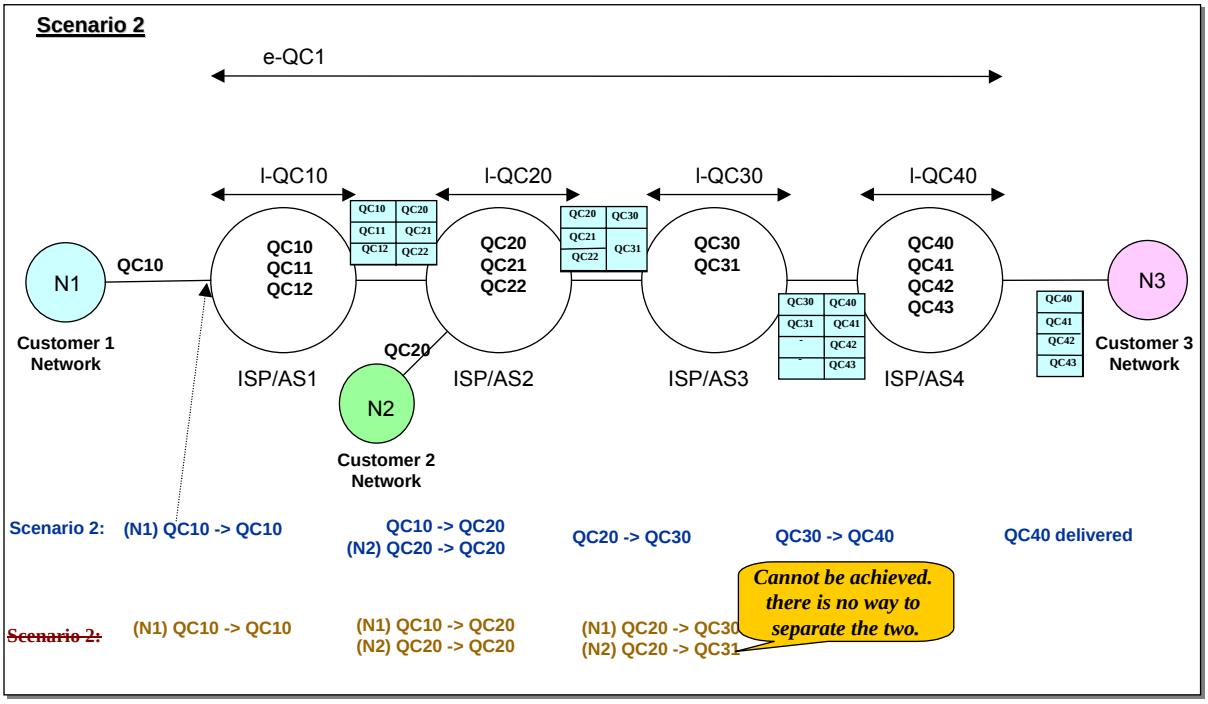


Figure 19: QC Implementation in IP-Based Networks - Scenario 2.

Copyright © MESCAL Consortium, June 2005

5.6.1.2.3 QC Implementation by Using Destination Network Address & Packet's DSCP (Scenario 3)

In Scenario 3 at the ingress point of each domain, QC implementation is performed based on the destination address and DSCP. Packet having the same destination address and DSCP are mapped to the same l-QC. However, a problem arises if, as shown in Figure 20, customers (N1, N2) will reach destination (N3) using QoS classes e-QC1 & e-QC2 respectively. The QC splitting problem, described in Section 5.5.1, arises. Introducing destination address processing in this scenario is a complexity concern, although this problem can be overcome by additional DSCP manipulation, at the domain egress point. A solution is proposed in Section 7.3.2 that eliminates the dependency on destination addresses as part of QC enforcement, but with limitation on the number of e-QC that a domain can support.

At ingress point of each domain, packets are examined by looking at the destination address and embedded DSCP. Packets having the same destination network addresses and DS code points are mapped to the same l-QCs. But packets coming from different sources (i.e., different ASs) possibly require different l-QCs to be mapped to, based on their e-QCs, and may use different routes to the destination. Thus, Destination Network Address & Packet's DSCP) are not adequate for proper QC implementation.

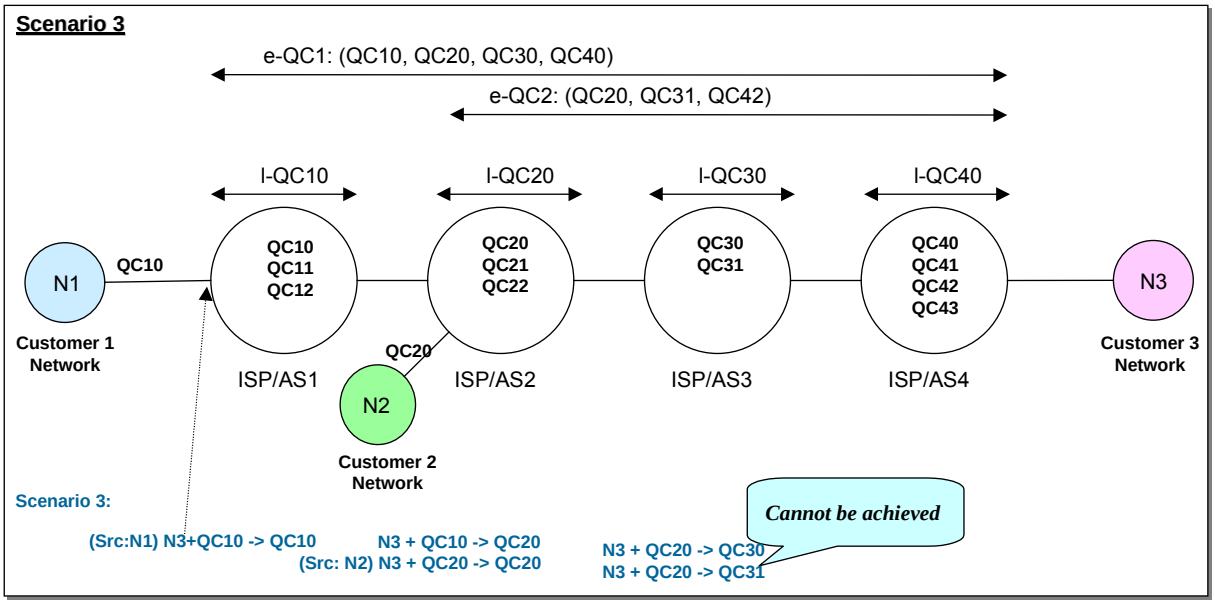


Figure 20: QC Implementation in IP-Based Networks - Scenario 3.

5.6.1.2.4 QC Implementation by Using Source & Destination Network Addresses and Packet's DSCP (Scenario 4)

The method introduced in this scenario provides total flexibility for QoS mapping and binding across all domains.

This scenario uses aggregate SLS characteristics and requires information on source network address, destination network address, l-QC used in the preceding AS and the l-QC that the packet is going to map to (see Figure 21) for QC enforcement. This requires full packet inspection (source address, destination address, DSCP) which is costly to ingress border routers.

Copyright © MESCAL Consortium, June 2005

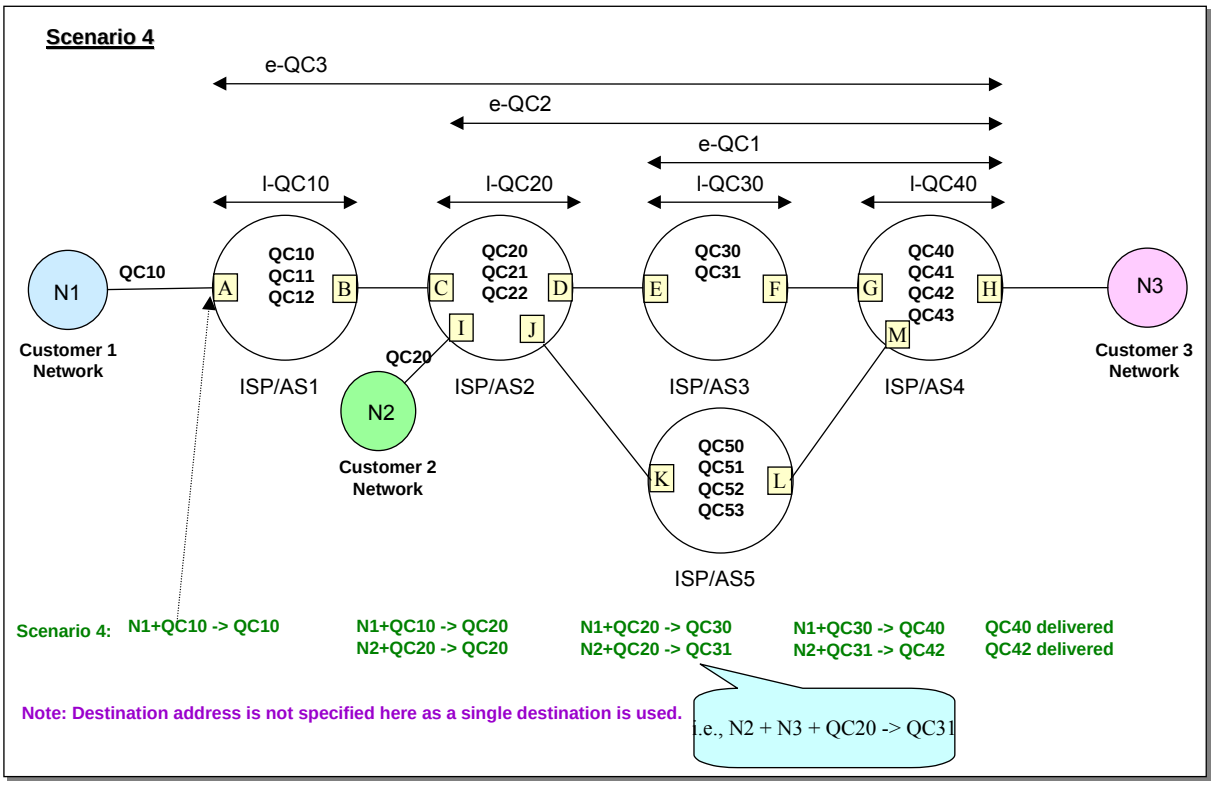


Figure 21: QC Implementation in IP-Based Networks - Scenario 4.

To perform QC enforcement on an IP packet sent from a customer in AS1-N1 to a customer in AS4-N3, the following actions are required across domains:

1- Border routers must be aware and act based on the tuple: (source AS1-N1, destination AS3-N3, DSCP embedded in the packet header, the l-QC to mapped to). In Scenario 4, the preceding AS is used instead of AS1-N1.

2- Router A in ISP1 maps the customer QoS class to QC10 by using three tuple (source IP address, destination IP address, packet's DSCP)

3- Router C maps QC10 to QC20 by using tuple (AS1-N1, AS3-N3, QC10)

4- Router E maps QC20 to QC30 by using tuple (AS1-N1, AS3-N3, QC20)

5- Router G maps QC30 to QC40 by using tuple (AS1-N1, AS3-N3, QC30)

6- Packet is directed to N3's customer via G and H routers.

Figure 22 shows an example for look up process at boarder router of an AS.

Copyright © MESCAL Consortium, June 2005

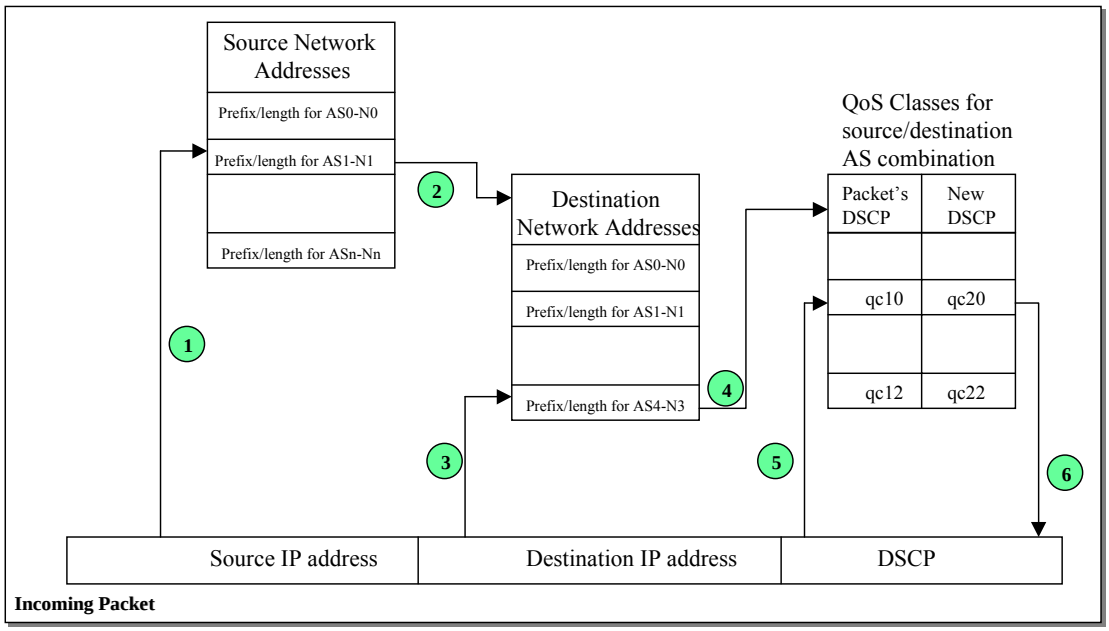


Figure 22: QoS class table lookup at router C of AS2.

BGP update messages contains a list of <prefix, length> tuples that indicate the list of destinations that can be reached via a BGP speaker. The update message also contains the path attributes, which include such information as the degree of preference for a particular route and the list of ASs that the route has traversed. In order to provide the information for QC enforcement and packet forwarding, the border router's (e.g. C) outgoing interface (similar to tunnel interface in the MPLS environment) or the outgoing boarder router (e.g., D) in the AS2 can be specified as part of this lookup. Consequently, the QC enforcement and packet forwarding can happen in this process.

5.6.2 QC Mapping & Binding

Each domain may offer a large number of l-QCs with different performance characteristics. Since there can be a large number of domains, a large number of possible/potential QoS mapping/binding can be found to satisfy the e-QCs performance targets at finer granularities. This can increase the number of possible paths to provide the e-QCs performance targets. An approach that creates many different e-QCs and possible paths may create complex routing issues and also degrade the routers' performance. This must be avoided by any proposed solutions. The number of e-QCs offered across the domains should be limited in order to avoid having complex routing tables, degrading the router performance, etc. To this end, the project has devised the notion of Meta-QoS-Classes to make this issue more manageable.

5.6.3 BGP

Any enhancement to the BGP protocol needs to be assessed for scalability and stability.

The use of BGP to carry QoS capability information between domains may lead to increasing the size or the complexity of the routing tables, as discussed in Section 5.4.3.2. The increase of the Internet routing table size has been a continual concern to routing manufacturers and the IETF. MESCAL needs to assess the consequences of its solutions on this aspect of BGP operation.

The stability of BGP is also an aspect that the project must address insofar as the frequency of updates caused by the MESCAL solution.

Copyright © MESCAL Consortium, June 2005

5.7 Bidirectionality

The MESCAL solutions allow QoS-based IP delivery service between end-points spanning a substantial number of domains. The general requirements of providing bi-directional services with, possibly different, QoS assurances in the forward and reverse paths should be considered.

In the cascaded approach adopted by the project, each Network Provider (NP) or ISP forms *pSLS* contracts with adjacent NPs. Thus, the QoS peering agreements are only between BGP peers. This process is repeated recursively to provision QoS to reachable destinations that may be several domains away. Figure 23 shows an example for end-to-end uni-directional QoS service implementation using the cascaded approach. Each NP/ISP administers its own domain and the inter-connection links that it is responsible for. For example in Figure 23, ISP1 is responsible for the network provisioning and resource allocation in AS1 including the configuration of both *"**a**"* and *"**b**"* interfaces.

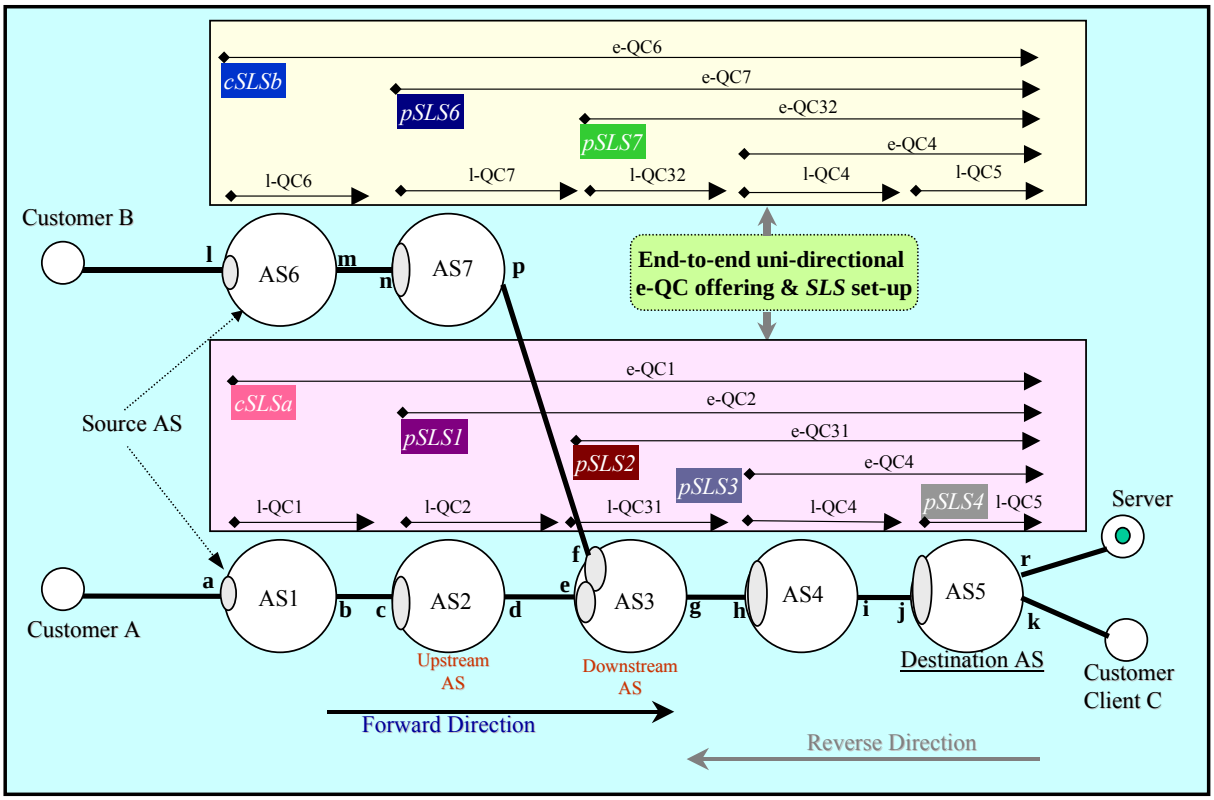


Figure 23. End-to-end uni-directional QoS service implementation

5.7.1 Bi-directionality in Statistical Guarantees Solution Option (2)

There are some fundamental problems to be solved in order provide bi-directional services with solution option 2. There are two methods to tackle the problem of providing QoS enabled path in reverse direction. The first method extends the single cascade with bi-directional capabilities. The second method employs a unidirectional cascade in each forward and reverse direction to build bi-directional services.

5.7.1.1 Method 1: Bi-directional pSLSs

One possible solution for setting up a reverse path is to negotiate *pSLSs* in reverse direction between peer ASs with an open destination scope (*). An open scope is necessary when considering that as the e-QC is sold on, it can become part of a new e-QC, the scope and QoS parameters of which cannot be known by the Destination AS. To allow the upstream AS to offer the e-QC to further upstream ASs without the need for amending the scope of pre-existing downstream pSLSs every time the scope

Copyright © MESCAL Consortium, June 2005

changes, the (*) is required. This potentially solves the bi-directionality problem at the pSLS level, but it raises some issues in implementing the e-QCs and invoking the service as discussed further in deliverable D1.4 [D1.4].

Alternatively, bi-directionality could be tackled by employing e-QC enabled *c/pSLSs* in the forward direction and l-QC enabled *pSLSs* with no explicit e-QC binding in reverse direction. This could have scalability problems in some specific scenarios, however. For example it is possible that two different streams of return traffic originated from a destination to a source may use the same l-QC in one of the transit domains. This creates a splitting problem at the egress point of that domain. v-QCs can be used within the domain to differentiate the streams at the egress point of the domain but it implies that a v-QC is needed per p_rSLS unless additional state information is used for inspecting and classifying packets. Both implications raise scalability issues.

5.7.1.2 Method 2: Multiple unidirectional cascades

This method allows the establishment of uni-directional *SLSs* for sending traffic only. Bi-directionality is left to the application layer to resolve. The suitable e-QCs have to be set-up separately by the source and destination ASs. There is no guarantee that a suitable e-QC for the return path will exist for any given forward e-QC, except by virtue of a "customer God", who ensures that suitable reverse path e-QCs exist in the destination AS, based on application requirements. This method would potentially provide the environment for having bi-directional services using the cascaded approach in both directions.

In the case where a client wants to receive traffic from the server with a given QoS (e.g., to download a file), the client must contact the server at the application layer with a request to send traffic to the client. The QoS requirements of the sending traffic as well as the billing details are also agreed between the two. The application layer communication between customers or client/server will need a way to describe and agree on the QoS levels to be used in each direction. This could be done by exchanging details of the specific e-QCs they have subscribed to in their respective *cSLSs*, or it could be done at a more abstract level in a customer language without exposing exactly how this is mapped to the e-QCs/*cSLSs*/QoS parameters they have with their respective ISPs.

5.7.2 Bi-directionality in Loose Guarantees Solution Option (1)

In this solution option, an AS advertises the Meta-QoS-classes it supports within its administrative domain. Other domains can make pSLS arrangement with this domain to make use of offered Meta-QoS-classes . Thereafter, each domain can find out whether it can reach certain destinations in a Meta-QoS-class plane through q-BGP updates it receives. *pSLSs* agreed between two domains are not tied with certain destinations as in solution option 2. Hence, as *pSLSs* are uni-directional and they are established for transporting traffic in forward direction, p_rSLS can be established for transporting traffic in reverse direction. The scopes for handling QoS of these two *pSLSs* are the same i.e., Meta-QoS-classes within the domain.

There might be a different Meta-QoS-class requirement in reverse direction than forward direction. To address this, there can be an application level communication between the two parties (customers) involved in order to specify the QoS requirements in either direction.

5.7.3 Bi-directionality in Hard Guarantees Solution Option (3)

Neighbouring domains establish pSLSs between themselves. q-BGP runs between the domains, which already have established pSLSs. Solution option three uses q-BGP to announce PCS unique identifiers across the Internet in order for "option-3" ASs to be able to discover a path towards every AS having a PCS. Therefore, when an AS wants to establish an LSP between 2 addresses, its PCS calculates a PCS-path towards the destination AS, and it is up to each AS in the PCS-path to establish the LSP. At the service/application level, when originating AS wants to establish an LSP to a destination ASs, there must be an agreement between the two ASs (PCSs). This agreement specifies both the tail-end address of the LSP, the PCS identifier of the destination AS and this is also used to verify the

Copyright © MESCAL Consortium, June 2005

existence of service contract exists between the two. In order to have bi-directional communication, *pSLS* and *p_rSLS* can be set-up the same as solution option 1. Thus, based on these SLS, LSPs can be created in forward and reverse directions in order to build bi-directional services.

5.7.4 Conclusion

This Section has given an overview of the problems and solutions discussed in deliverable D1.4 [D1.4] for providing bi-directional services with the three MESCAL solution options. The main issue is how to construct the QoS-enabled reverse path for return traffic. For solution option 2 we believe the most feasible solution for providing bi-directionality is the use of multiple cascades for uni-directional e-QCs. Providing bi-directional services in solution options 1 and 3 causes less complication, because *pSLSs* are based on the Meta-QoS-class concept without specific end-to-end performance guarantees or predefined service scope in terms of reachable destinations.

A general conclusion for all solution option 1 and 2 is the requirement for service/application level signalling between the communicating parties. This is to find-out about the Meta-QoS-class plane for reverse direction, information for billing and admission control in solution option 1, to specify the desired sink for return traffic for the Destination AS and the l-QC/e-QC for return traffic, information for billing and admission control in solution option 2. In solution option 3, service level communication is also required to pass to source AS head-end of LSP and possibly PCSID of that domain and destination AS with tail-end of LSP and possibly PCSID of that domain and necessary information for authentication and billing purposes.

5.8 Multicast Implications

5.8.1 Multicast Service Models

Proper selection of multicast service models is a vital prerequisite for successful development in provisioning QoS-enabled multicast services in the Internet. It has been argued that the service model of IP multicast [Deeri88] was originally defined without an explicit objective in commercial services, which is one of the major reasons for its slow deployment [Diot00]. IP multicast, also known as Any Source Multicast (ASM), is an open group service model in that there are no mechanisms that restrict hosts from sending data to a group, or receiving data from it. In summary, the traditional IP multicast is lacking sophisticated group management. Source Specific Multicast (SSM) [Holbr03] is proposed as a closed group service model, and it has received more and more attractions ever since its birth. Compared with ASM, SSM has its own advantages in multicast source management and implementation scalability.

5.8.2 Multicast Service Level Specification (mSLS)

In IP multicast, group members are always anonymous to the multicast source. Moreover, almost all the multicast applications are receiver initiated other than sender based. Concerning QoS requirements, it is individual group members that request different service levels based on their own capabilities. These characteristics require that the Service Level Specification for QoS aware multicast applications should not be borrowed directly from the unicast scenario that is purely source based. How to define and implement multicast oriented Service Level Specification is one of the most important issues in the relevant deployment.

5.8.3 Multicast routing

Using PIM-SM [Fenner03a] with the aid of MBGP [Bates00] / MSDP [Fenne03b] has once been recognised as a promising near-term solution to the deployment of IP multicast services in the Internet. However, whether this is a valid argument is still under debate today. In this approach, PIM-SM caters for the construction of intra-domain multicast trees, and MSDP has the functionality of discovering active sources located in different domains so that these intra-domain trees can be connected together

Copyright © MESCAL Consortium, June 2005

to form a unique inter-domain tree. MBGP is the multi-protocol extension to BGP4 that allows incongruent routes for unicast and multicast traffic across multiple domains. BGMP [Thale03] / MASC [Rados00] was first proposed as a long-term solution for internet-wide multicast routing, but it has not seen any significant progress in practical development till now. On the other hand, IGMPv3 [Cain02] and PIM-SMv2 have been adapted to support the SSM service model, with capabilities of source filtering and explicit group join. As far as the MESCAL project is concerned, it is also an important issue to select a proper routing infrastructure from existing schemes (MSDP/PIM, BGMP/MASC and SSM) for further QoS deployment.

It should also be noted that, since none of the existing multicast routing protocols support QoS aware routing, some adaptation/extensions would become necessary to achieve this capability. One of the typical issues is Reverse Path Forwarding (RPF) checking that is used to detect loops in multicast tree. In PIM-SM, if a multicast packet does not come from the interface, which is used to deliver unicast traffic towards the source, it will be discarded. However, the paths computed by QoS routing mechanism are often not the shortest one, and hence QoS multicast tree construction will fail if the conventional RPF checking is performed.

5.8.4 Multicast Group Management

In the IP multicast, IGMP is used to notify Designated Routers (DR) on active receivers for each group. A new group membership report will trigger the underlying routing protocol for sending the corresponding join request, while redundant reports for the same group will be suppressed. This will not be the case when individual receivers demand heterogeneous QoS requirements for the same group session. As a result, mechanisms for handling QoS-aware group membership reports will also be investigated in the MESCAL project. On the other hand, multicast group member admission control will become a new functionality of group management, and this is another important issue for successful QoS provisioning.

5.8.5 Multicast Scalability

It has also been deemed that scalability is one of the significant obstacles that hamper fast development in multicast services. This issue exists not only in the inter-domain semantics such as AS-level source discovery and class D address allocation, but also in the group state maintenance at the level of router implementation. In the MESCAL project, when we consider the QoS enabled multicast services, efforts should be targeted at minimising extra impacts on both cases. The proposed solution should not significantly worsen the current multicast scalability problems, so that the corresponding implementation is too complicated to be achieved.

Copyright © MESCAL Consortium, June 2005

6 MESCAL FUNCTIONAL ARCHITECTURE

This Section introduces the functionality required for the provision of inter-domain QoS services from the perspective of a single provider. The functional architecture analyses the overall problem of providing inter-domain QoS and decomposes it into a set of finer grained components. One of the objectives of this exercise is to aid the development of the MESCAL solutions by breaking it down into manageable entities while maintaining a holistic view of the overall issues to be solved. In essence it is a divide and conquer exercise.

The MESCAL functional architecture was initially proposed in deliverable D1.1 [D1.1]. This Section revisits the architecture in the light of the detailed specification, implementation and experimentation activities that have subsequently taken place and culminating in the specifications in the main sections of this document. Each of the function blocks is analysed detail in Sections 8 to 12 of this deliverable, where algorithms and protocols to implement the required functionality are proposed.

6.1 Functional architecture overview

Figure 24 shows the MESCAL functional architecture showing the interactions between functional blocks at a high level. The arrows depict the direction of the main flow of information between functional blocks, generally implying a configuration or the invocation of a method in the direction of the arrow. Figure 24 also shows the interactions between providers and between customers and providers. The downstream provider on the right of the figure shows only the components directly involved in service peer interactions. An upstream provider is also implied on the left hand of the figure, although not shown explicitly. Interactions with upstream providers are a mirror of those shown with the downstream provider on the right.

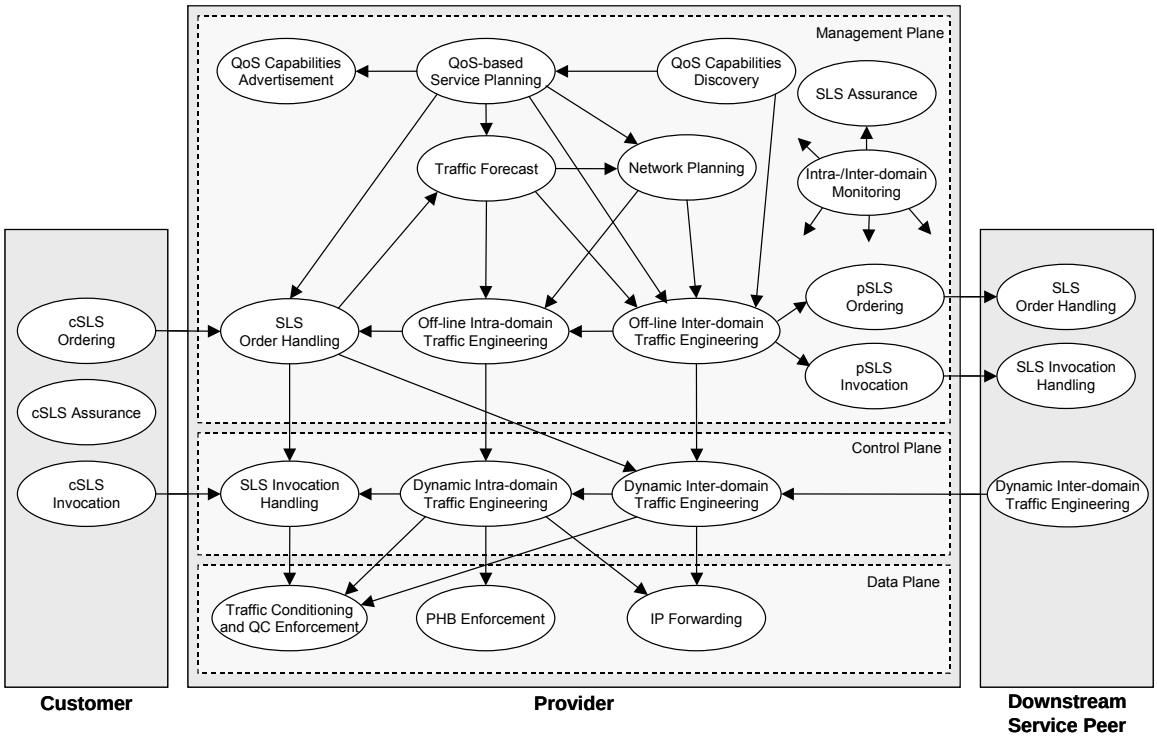


Figure 24. The MESCAL functional architecture

The data plane is responsible for per packet treatment within packet arrival epochs. The control plane covers intra- and inter-domain routing, SLS invocation handling – including authentication, authorisation and admission control – dynamic resource management – including load distribution and capacity management functions. Typically, control plane functions are embedded within network equipment although they are not involved in packet-by-packet decisions.

Copyright © MESCAL Consortium, June 2005

The management plane is off-line functionality, typically located outside of the network elements in management servers. The management plane functions are responsible for planning, dimensioning and configuring the control and data planes and interacting with customers and service peers to negotiate contracts. While management plane functions are not as dynamic as control and data plane functions they are by no means static. Within the MESCAL system there is a continual background activity within the management plane at the epochs of the so-called resource provisioning cycles (RPCs). There are two RPCs in MESCAL – the intra-domain RPC which involves off-line intra-domain Traffic Engineering, and the inter-domain RPC which involves off-line inter-domain Traffic Engineering. The latter may be further decomposed into a *Binding Selection Cycle* and a *Binding Activation Cycle* (see Section 10. The RPCs aim at proactively optimising network resources to meet predicted demand and to build in sufficient spare capacity to avoid the burden of reconfiguring the network for each and every SLS subscription or renegotiation, without the inefficiencies and costs associated with massively over provisioning resources.

While the architecture describes the full set of functions required for a provider to participate in the end-to-end provision of QoS-based IP services by no means does it prescribe the implementation means by which they will be realised – within network equipment or in external management servers, with automated or manual processes. This is a matter for each provider. While the full set of functional blocks (or their equivalent) are expected to be in place in downstream providers, MESCAL does *not* assume that *automated* processes will always implement all blocks. This deliverable proposes algorithms suitable for deployment as automated processes in the traffic engineering and service management functional blocks but it is also possible to deploy much of the management plane through manual processes, at the cost of reduced responsiveness or flexibility. For some of the service options identified in this deliverable, the algorithms or manual processes required to implement the functionality might be trivial. For instance, the *loose guarantees service option* does not require explicit admission control functionality in the SLS Invocation Handling block, and the QC Mapping, Binding and Activation processes are simplified due to its adoption of well-known Meta-QoS-Classes and the restriction to bindings only with the same Meta-QoS-Class in service peer domains.

The following subsections identify the major aspects of the functionality contained within each of the blocks shown in and highlight the changes to the functional architecture that have been made since D1.1.

6.2 QoS-based Service Planning, QoS Capabilities Discovery and Advertisement

QoS-based Service Planning encompasses all the higher level business related activities responsible for defining the services that the provider should offer to its customers and service peer providers. These are specified according to the business objectives of the provider, and include l-QCs within the scope of its own network and e-QCs combining its local QoS-based services with those offered by its service peers.

Prior to any pSLS agreement with a neighbouring provider, a provider discovers the QoS capabilities, capacities, destination prefixes and costs of potential service peer providers thanks to the *QoS Capabilities Discovery* functional block. Once l-QCs and e-QCs have been defined and engineered (by Intra- and/or Inter-domain TE) the *QoS Capabilities Advertisement* block is responsible for promoting the offered services so that its customers and service peer providers are aware of its offerings. It is envisaged that a variety of advertising means will be used, ranging from digital marketplaces or other automated peer-to-peer processes to conventional techniques such as salespersons, newspapers and word of mouth.

6.3 Off-line Traffic Engineering

Traffic Forecast is responsible for aggregating and forecasting traffic demand. During a provisioning cycle, the set of subscribed cSLSs and pSLSs are retrieved from *SLS Order Handling* and an aggregation process derives a traffic matrix with the demand per ordered aggregate between ingress

Copyright © MESCAL Consortium, June 2005

and egress points of the domain (ASBRs). The demand matrix is used by the intra- and inter-domain traffic engineering processes to calculate and provision the local and inter-domain resources needed to accommodate the traffic from established SLSs as well as those anticipated to be ordered during the provisioning cycle.

Binding Selection is the process of combining l-QCs of the local domain with o-QCs of other domains, learned through *QoS Capabilities Discovery*, to construct potential e-QCs that meet the service requirements defined by *QoS-based Service Planning*. It should be noted that *Binding Selection* might result in a number of QoS-bindings for a given e-QC. QoS-bindings with the same service-peering provider may differ in the l-QC and subsequently in the o-QC they use. Alternatively, QoS-bindings may differ when established with different service-peering providers.

Binding Activation is responsible for mapping the predicted traffic matrix to the inter-domain network resources (once pSLSs have been established), satisfying QoS requirements while aiming at optimising the use of network resources across AS boundaries. *Binding Activation* decides which of the established QoS-bindings will be *put in effect* in the network for implementing an e-QC together with the associated routing constraints for those e-QCs. The QC-bindings in effect will be enforced through routing decisions as well as configurations of the *Traffic Conditioning and QC Enforcement* block, e.g. configuring the egress ASBR to perform DSCP remarking for realising a QC-binding. The latter configuration can be made directly to the egress router or passed through *Dynamic Inter-domain Traffic Engineering*.

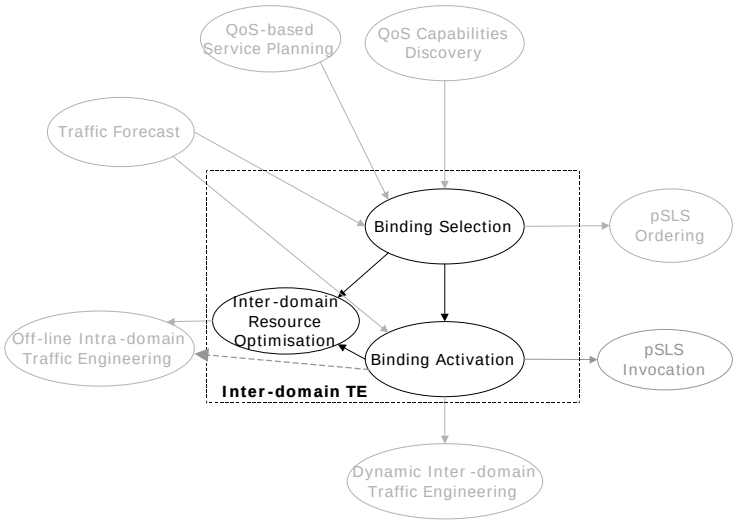


Figure 25. Decomposition of the Offline inter-domain TE

Although in the initial Functional Architecture described in D1.1 [D1.1] the inter-domain Traffic Engineering system was decomposed into the *QC Mapping*, *Binding Selection* and *Binding Activation* blocks, after a more detailed study of the functionality of the blocks as well as the corresponding algorithms, a change in this decomposition was decided which is depicted in Figure 82. First, the functionality of the *QC Mapping* functional block was considered too lightweight to justify a single functional block and was incorporated in the *Binding Selection* block as the first step of its algorithm. Moreover, both *Binding Selection* and *Binding Activation* have to run an optimisation algorithm, which will decide on the most optimal resource allocation in terms of inter- as well as intra-domain cost in order to satisfy a predicted traffic demand. This resource allocation could be either for the establishment of the pSLSs for the next *Binding Selection* period or for the allocation of the inter-domain resources for the next provisioning cycle.

Consequently, we have included in the Traffic Engineering System an *Inter-domain Resource Optimisation* block, which realises the algorithm described above and is called both *by Binding Selection* and *Binding Activation*. Of course, the input to the algorithm will be different when called by *Binding Selection* and when called by *Binding Activation* since in the first case a traffic demand for a longer period will be passed as input to the algorithm while in the latter case a shorter term prediction of the traffic demand will be the input.

Copyright © MESCAL Consortium, June 2005

Moreover, when Binding Activation triggers the *Inter-domain Resource Optimisation* algorithm the allocation of resources is constrained by the already established pSLSs while *Binding Selection* has to consider different hypothetical scenarios of pSLSs in order to decide which one of them leads to a more optimal solution in terms of resource utilisation and at the same time satisfying the traffic demand.

Off-line Intra-domain Traffic Engineering computes the intra-domain network configuration in terms of routing constraints and PHB capacity requirements in order to satisfy the predicted traffic demand at intra-domain RPC epochs.

The off-line intra-domain TE block has been further decomposed into two sub-components: *Resource Optimisation* and *Network Reconfiguration Scheduler*. The *Network Reconfiguration Scheduler* is the control system for the Offline Intra TE block. It has two main purposes, handling computation requests to *Resource Optimisation* (Resource Provisioning Cycles, Inter-domain Traffic Engineering "what if" queries, etc) and scheduling the reconfiguration of the network using link weight settings computed by the *Resource Optimisation* block. The *Resource Optimisation* block contains the OSPF link weight optimisation algorithm. It is a passive block, until called by the *Network Reconfiguration Scheduler* at which point it collects a traffic demand matrix and a network topology and computes an optimal set of link weights. Computed weights are deposited in a link weight database inside the Offline Intra-domain Traffic Engineering block, until they are put into operation in the network by the *Network Reconfiguration Scheduler*.

The interactions required between off-line inter- and intra-domain TE and the options for coupling/decoupling the inter- and intra-domain RPCs are analysed in detail in Section 10.

6.4 Dynamic Traffic Engineering

Dynamic Inter-domain Traffic Engineering runs within an inter-domain RPC and is responsible for inter-domain routing e.g. q-BGP advertisement, q-BGP path selection and for dynamically performing load balancing between the multiple paths defined by the static component based on real-time monitoring information changing appropriately the ratio of the traffic mapped on to the inter-domain paths.

Dynamic Intra-domain Traffic Engineering is the dynamic management layer as defined in TEQUILA [TEQUI]. This includes the intra-domain routing algorithms, e.g. QoS-enhanced OSPF, together with other dynamic algorithms to manage the resources allocated by *Off-line Intra-domain Traffic Engineering* during the system operation in real-time, in order to react to statistical traffic fluctuations and special arising conditions within an intra-domain RPC. It basically monitors the network resources and is responsible for managing the routing processes dynamically as well ensuring that the capacity is appropriately distributed among the PHBs.

6.5 SLS Management

The SLS Management functionality can be split into two parts: (a) the part responsible for the contracts offered by the provider to its customers, i.e. the end-customers and interconnected providers, and (b) the part responsible for the contracts requested by the provider from its peer providers. The resulting functional components are named "*SLS Order Handling*" and "*SLS Ordering*" respectively. While the ordering process establishes the contracts between the peering providers, the invocation process is required to commit resources before traffic can be exchanged, with "*SLS Invocation Handling*" and "*pSLS Invocation*" providing the necessary functionality.

SLS Order Handling is the functional block implementing the server side of the SLS negotiation process. Its job is to perform subscription level admission control. The *Off-line Intra-domain Traffic Engineering* block will provide *SLS Order Handling* with the resource availability matrix (RAM) which indicates the available capacity of the engineered network to accept new SLS orders – both within the AS and on any inter-domain pSLSs it has with neighbouring ASs. *SLS Order Handling* will negotiate the subscription of both cSLSs and pSLSs – they will be (largely) treated in the same way.

Copyright © MESCAL Consortium, June 2005

SLS Order Handling maps incoming SLS requests onto the o-QCs it can offer and investigate whether there is sufficient intra- and inter-domain capacity, based on the RAM for that o-QC.

pSLS Ordering is the client side of the pSLS negotiation process. During an inter-domain RPC *Binding Selection* may identify the need for new pSLSs with service peers. *pSLS Ordering* implements the decisions of the *Binding Selection* algorithms and undertakes the negotiation process.

The *pSLS Invocation* function block is responsible for invoking pSLSs with peer domains. The pSLSs have already been subscribed through an ordering process between *pSLS Ordering* and *SLS Order Handling*. Optionally *pSLS Invocation* may be directly invoked by *Dynamic Inter-domain Traffic Engineering* to cater for fluctuations in traffic demand which are significantly different to those forecasted and used by *Binding Activation* for the current RPC. Whether or not this should trigger a new binding activation cycle by involving *Binding Activation* and *Inter-domain Resource Optimisation* is a topic for further study.

Admission control is needed to ensure that the network is not overwhelmed with traffic when the network adopts a policy of overbooking network resources at the subscription level. *SLS Invocation Handling*, containing the admission control algorithm, receives signalling requests from customers or peer providers for cSLS and pSLS invocations respectively. *SLS Invocation Handling* checks whether the invocation is conformant to the subscribed SLS and whether there is sufficient capacity in the local AS and also on the inter-domain pSLSs in the case of SLSs that are not terminated locally.

6.5.1 Monitoring and SLA Assurance

Monitoring is responsible for both node and network level monitoring through both passive and active techniques. It is able to collect data at the request of the other functional blocks and asynchronously notify the other functional blocks when thresholds are crossed on both elementary data and derived statistics.

For simplicity in the diagram the full set of interactions with *Monitoring* is not depicted, however *SLS Invocation Handling*, *Dynamic Inter-/Intra-domain Traffic Engineering* and *pSLS Invocation* blocks continually use monitored data in order to operate. The less dynamic *Off-line Inter-/Intra-domain Traffic Engineering* functions as well as *Traffic Forecast* use monitored network statistics at RPC epochs. *Traffic Forecast* uses historical data to improve the accuracy of future traffic matrix estimates.

Inter-domain monitoring could take several forms: monitoring inter-domain links (pSLS) only; monitoring end-to-end performance across several ASs through loop-backs or remote probes for one-way measurements; collection of data generated by service peers (possibly through BGP advertisements, or through another monitoring data exchange protocol). Alternatively third part auditing may be a more acceptable means for both monitored and monitoring ASs.

SLS Assurance compares monitored performance statistics to the contracted QoS levels agreed in the SLSs to confirm that the network or service peer-networks are delivering the agreed service levels.

6.5.2 Traffic Conditioning and QC Enforcement, PHB Enforcement and IP Forwarding

Traffic Conditioning and QC Enforcement is responsible for packet classification, policing, traffic shaping and DSCP marking according to the conditions laid out in previously agreed SLSs and the invocation of those SLSs. At ingress routers the *Traffic Conditioning* function is responsible for classifying incoming packets to their o-QC and subsequently mark them with the correct DSCP for the required l-QC. At the egress router the *QC Enforcement* function may need to remark outgoing packets with the correct DSCP as agreed in the pSLS with the service peer. In other words *QC Enforcement* is responsible for implementing the data-plane binding from l-QC to o-QC of the service peer. Note that *QC Enforcement* is not responsible for selecting the correct peer AS: this is decided by q-BGP (part of the *Dynamic Traffic Engineering* blocks in Figure 24), therefore *QC Enforcement* does not implement the full QC mapping/binding process in the data plane.

Copyright © MESCAL Consortium, June 2005

PHB Enforcement represents the queuing and scheduling mechanisms required to be present in order to realise the different PHBs with the appropriate configuration as defined by the TE related blocks.

IP Forwarding represents the functionality needed to forward IP datagrams based on the information maintained in the corresponding FIBs. Optionally, IP forwarding may also include mechanisms to perform multipath load balancing.

6.6 Interactions between SLS Management and Dynamic Inter-domain Traffic Engineering

This Section describes the relationship between *SLS Management* and q-BGP *when we have an agreement either for a new or an updated pSLS*. Note that these interactions are not the *only* interactions between the MESCAL management functions and q-BGP, for example traffic engineering decisions will also control and influence the q-BGP machinery.

The rest of this Section is organised as follows. We review briefly the structure of pSLSs and the functionality of q-BGP. This review is at an abstract level, since pSLSs and q-BGP are defined in Sections 9 and 10.5.1 respectively. The second part of this Section is devoted to the actual exchange of information between pSLS and q-BGP. We discuss **where, what**, **when** and **who** is responsible for the information exchange.

6.6.1 Review of pSLS and qBGP

A pSLS contains the following constituents that have been agreed between two ASs as part of the *SLS Order Handling* function:

- A defined offered QoS Class, o-QC (required for all solution options);
- Reachability information: a set of destination addresses to which this o-QC is valid (required for statistical and hard solution options; not required for loose solution option);
- A bandwidth (i.e. a data rate, in units of bits/second) that defines the rate at which, traffic may be sent within the terms of this pSLS, possibly including a traffic profile (required for statistical and hard solution options; not required for loose solution option);
- Time schedule (required for all solution options).

It is anticipated that in a case where there are multiple links between two ASs, then for each link we will in general have different values for some of the constituent parameters enlisted above. For example, the bandwidth may be different, or the reachable address prefixes may be different for different peering links. This is addressed by assigning separate pSLSs to each link.

q-BGP will perform inter-domain path selection based on QC-related information and path availability information. As described in Section 10.5.1, q-BGP allows exchange of QoS Service Capabilities, QC identifier, and QoS performance characteristics.

6.6.2 Interactions

6.6.2.1 Introduction: principal entities in pSLS-q-BGP interaction

The pSLS – q-BGP interaction is illustrated with the pair of autonomous systems shown in Figure 26. Each AS contains a management node, denoted X and Y respectively (we assume one per AS; discussions of backup nodes are outside the scope of this discussion). For the pSLS agreement between AS1 and AS2, X is responsible for performing the *pSLS Ordering* function, and Y is responsible for the *SLS Order Handling* function. Nodes X and Y are thus responsible for agreeing the pSLS (or pSLSs) between AS1 and AS2.

The other entities in scope here are:

- Upstream AS ingress node(s) (i.e. A in Figure 26);

Copyright © MESCAL Consortium, June 2005

- Upstream AS egress node(s) (i.e. B in Figure 26);
- Downstream AS ingress node(s) (i.e. C in Figure 26).

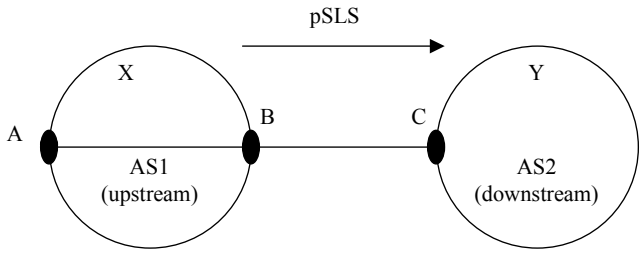


Figure 26. Two adjacent autonomous systems

Between B and C there is an exterior q-BGP protocol flow (e-q-BGP) and between A and B there is an interior q-BGP (q-iBGP) session. We assume that a pSLS has just been agreed (either a new or a revised old one) between AS1 and AS2. In the following we will elaborate on the interaction required between the management functions and q-BGP, based on the following: *where* do we foresee this interaction, *what* is the information included in that interaction, *when* this interaction happens, and finally *who* is responsible to perform that interaction.

6.6.2.2 Where

Routing advertisements are propagated from AS2 to AS1 using e-q-BGP. These advertisements must include some QoS information that is part of the agreed pSLSs between the two ASs. In general, the two domains will filter out any other advertisement that is not part of an agreement. Thus, after a pSLS is agreed, whether new or revised, both parties should enable the exchange of such advertisements.

We therefore conclude that interaction between pSLS information and q-BGP is required at the following locations:

- At the ingress nodes of the downstream AS (i.e. C in Figure 26), to implement a policy that enables the related q-BGP advertisements towards the upstream AS;
- At the egress nodes of the upstream AS (i.e. B in Figure 26), to implement a policy that allows (stops filtering out) the related q-BGP advertisements.

Additionally, when a new pSLS is agreed, the upstream node within the AS (i.e. A in Figure 26) has to know about the new available resources in order to use them in the egress selection process. The Interior q-BGP (q-iBGP) within an AS, between A and B in our example, will provide the appropriate reachability and QoS information. If the domain's approach is that bandwidth information is not carried in i-q-BGP then there are two ways for the internal nodes, like A, to "learn" that information. Either we run an IGP with Traffic Engineering (TE) extended LSAs including inter-domain links as TE-links (as proposed in [Vass03]), or the management node X could pass the pSLS bandwidth information directly to ingress node(s) A.

6.6.2.3 What

Having identified *where* the information is exchanged, we will now look into *what* is the required information to be exchanged.

6.6.2.3.1 Policy filters

As outlined in Section 6.6.2.2 the downstream AS must advertise q-BGP reachability information to the specific addresses included in the pSLS or to "all addresses" in the case where reachability information is not specified in the pSLS. Thus the appropriate policies that allow these advertisements should be conveyed to the downstream AS ingress nodes (C in our example). Similarly, the upstream AS must allow these advertisements to be accepted and not filtered out, and further allow them to

Copyright © MESCAL Consortium, June 2005

propagate into q-iBGP after applying the path selection algorithm. Therefore the appropriate policy for allowing in (i.e. stop filtering out) the related advertisements should also be downloaded to the q-BGP process in the appropriate node(s), router B in our example.

6.6.2.3.2 QoS attributes

Routing advertisements are propagated from AS2 to AS1 using e-q-BGP. These advertisements must include some QoS attribute. This QoS information is closely related to the pSLSs agreed between the two ASs. Thus, the information passed from a management node to q-BGP must be the agreed o-QC. Note that this does not necessarily mean that the o-QC is the actual attribute included in q-BGP, but rather that the q-BGP advertised information needs to be related somehow with the agreed o-QC. The only requirement for this relationship is that the o-QC values must be the worst-case upper bound for the relevant q-BGP QoS attribute values, thus allowing some flexibility in what is actually advertised into q-BGP. The decision of the actual parameters that constitute the q-BGP QoS attribute are for further study: for example, in a simple case we can just copy the appropriate o-QC values into the QoS attribute fields, and still be compliant with worst-case upper bound requirement.

This o-QC information is also required for the policy filters described in Section 6.6.2.3.1, and therefore o-QC is required at both the upstream AS egress nodes and the downstream AS ingress nodes.

6.6.2.3.3 Reachability information

Reachability information, i.e. specific address prefixes, is required both as part of the policy filter information and also for injection into q-BGP. For the former reason, it is therefore required at both upstream AS egress nodes and at downstream AS ingress nodes. If the information about specific address prefixes is not part of the pSLS agreement, then it is assumed to be “wildcard”, that is the equals all the address prefixes to which there is reachability with the best-effort class.

6.6.2.3.4 Bandwidth

As discussed in the last paragraph of Section 6.6.2.2, bandwidth availability on the egress link for a particular QC is required for TE functions within the upstream AS, i.e. AS1 in our example. One of the TE functions that require this information is the egress path selection process of the ingress nodes of the upstream AS (e.g. node A). In Section 6.6.2.2 we described a number of alternatives of how this information becomes available to ingress nodes, and one of these alternatives included using q-BGP as that means. In the rest of this Section we will assume that the preferred alternative is q-BGP, and will discuss how and where this bandwidth must be injected into q-BGP.

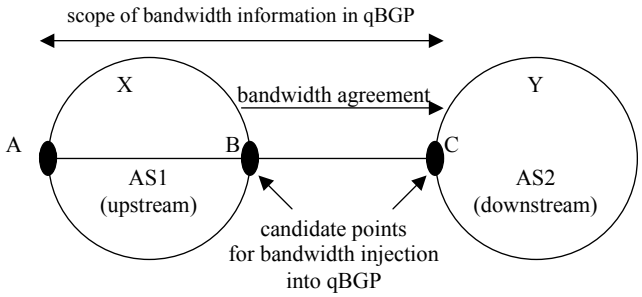


Figure 27. The case of bandwidth in q-BGP

If q-BGP is used to propagate pSLS bandwidth within the upstream domain, the scope of this propagation is **only** between the ingress node of the downstream AS, i.e. node C, and all the ingress nodes of the upstream AS, e.g. node A, see Figure 28. There are two principal alternatives as to where bandwidth is injected into q-BGP if this policy is adopted. One is at the egress point of the upstream

Copyright © MESCAL Consortium, June 2005

AS of the agreement, i.e. node B in the example, and the other alternative is at the ingress node of the downstream AS, i.e. node C of our example.

We propose to choose the latter alternative for two reasons. First, because node C already is responsible for setting the QoS attributes of the q-BGP advertisements towards node B. Second, this alternative gives us the ability to perform dynamic TE with q-eBGP at node B, in addition to the TE for egress selection with q-iBGP at node A (see Figure 28). Figure 28 extends our model to the case of multiple links between the upstream and downstream AS: node B can use bandwidth information propagated using q-eBGP to select either path BC or BE.

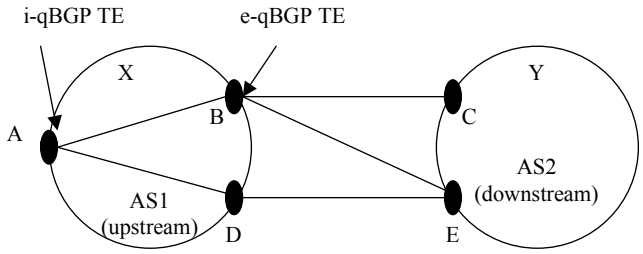


Figure 28. Illustration of q-iBGP and q-eBGP dynamic traffic engineering

6.6.2.3.5 Summary

Table 2 summarises **what** information needs to be passed from the pSLS related functions to q-BGP:

Upstream AS egress nodes (e.g. B)	**Downstream AS ingress nodes (e.g. C)**
Policy Filter (**allow in**) for q-BGP advertisements	Policy Filter (**allow out**) for q-BGP advertisements
o-QC	o-QC
Reachability (destination addresses)	Reachability (destination addresses)
	Bandwidth

Table 2. Summary of data transferred from pSLSs to q-BGP

6.6.2.4 When

The information related to a pSLS that needs to be exchanged between the management functions and q-BGP, as identified in Section 6.6.2.3, needs to be conveyed to the q-BGP machinery each time:

- A pSLS is created, modified or deleted (i.e. part of the *SLS Order Handling* function); or
- Some bandwidth is dynamically invoked within the given pSLS as part of the *SLS Invocation Handling* function (based on the restrictions discussed in the last paragraphs of Sect. 6.6.2.2 and 6.6.2.3).

6.6.2.5 Who

The interactions discussed in this Note are between the pSLS related blocks and the (edge) nodes that support q-BGP. In the MESCAL functional architecture [D1.1] these are *pSLS Ordering* and *pSLS Invocation* in the case of an upstream AS (e.g. AS1 in Figure 1) and *SLS Order Handling* and *SLS Invocation Handling* in case of a downstream AS (e.g. AS2 in Figure 1).

Copyright © MESCAL Consortium, June 2005

The offline SLS management blocks are assumed to reside in some management server nodes, X and Y in our example, while the dynamic functions regarding the pSLS invocations are implemented in the edge routers, B and C in the example. The communication between the SLS management nodes and the q-BGP routers can be implemented using any standardised management protocol e.g. SNMP or any other proprietary means e.g. Telnet/CLI.

6.7 Network Provisioning Cycle

6.7.1 Network Planning and Provisioning

Network Planning is defined as the off-line processes that are responsible for determining the type, quantity and geographical location of the physical resources required by an IP Network Provider conduct its business by offering IP connectivity services to meet the predicted demand of its customers. According to the role of the IP Network Provider, as defined in the MESCAL business model, the physical resources in question include, points of presence, IP routers and the communications links interconnecting them, as well as mother equipment required for the operation of an IP network, such as management servers.

Network Provisioning is defined as the processes responsible for ensuring that the physical resources are deployed as planned and with the appropriate physical configuration. This is distinct from *Traffic Engineering,* which is responsible for managing the distribution of traffic, optimising the use of the deployed physical resources and ensuring QoS in a cost effective manner. In the MESCAL functional architecture TE is involved with the soft configuration of existing physical resources, which will be accomplished by setting and modifying OSPF weights, PHB bandwidth, q-BGP route selection parameters as well as dynamically creating and updating the RIBs, FIBs etc.

Many network management activities, including traffic engineering, can be achieved automatically through configuring equipment via network management interfaces. The MESCAL solutions for SLS management and traffic engineering aim to deploy intelligent algorithms to meet this goal. On the other hand, the implementation of network planning decisions through network provisioning processes usually involves manual installation or configuration of physical equipment. This is clearly not something that can be automated, although it is possible to generate trouble tickets and work schedules this way. One aspect of network provisioning that could be achieved automatically, however, is the creation and modification of the transport capabilities of underlying physical networks to provide the required connectivity between the routers of the IP network.

The MESCAL business model assumes *Physical Connectivity Providers* (*Facilities Providers)* provide link layer pipes (e.g. electrical, optical, satellite) to interconnect the *IP Network Providers'* routers. Chapter 3 of deliverable 1.4 [1.4] considers the underlying transport network provided by Physical Connectivity Providers and how they may be interfaced to IP Network Providers offering one or more MESCAL service options.

Network provisioning can occur at a range of time scales. On a monthly scale new IP peering agreements will cause the network planner to request new or additional physical connectivity between IP Network peers. On a short time scale the Intra- and Inter- Domain provisioning cycles could cause the creation of new links and/or the modification of existing links' capacities via a management plane protocol, e.g. XML a la TEQUILA/MESCAL SrNP or control plane signalling, e.g. RSVP-TE, LCAS, (see D1.4).

6.7.2 Optical network technologies for dynamic network provisioning

The current most widespread approach to optical networking is the provisioning of static wavelengths within fibres in WDM (Wavelength Division Multiplexing) systems. Static point-to-point links provide fixed paths for wavelengths between two geographic locations. Network configuration is performed through manual configuration or via electrical switching. Electrical domain switching is not however fast enough for new applications and emerging line speeds and therefore new all-optical approaches to wavelength switching are being developed. To support the physical connectivity

Copyright © MESCAL Consortium, June 2005

demands of MESCAL solution options at the fastest possible provisioning speeds with the least restrictions (capacity granularity, enforced topology, hierarchy etc.), it is envisioned that intelligent dynamic optical network would be required. While SDH and many other Layer 2 protocols could support the capacity requirement, their switching and transmission bandwidth limits are being approached. The emerging technologies considered in D1.4 are GMPLS (Generalised Multi-Protocol Label Switching) and ASON (Automatically Switched Optical Networks). These technologies provide an overlapping set of features that could be used in future networks to provide all optical dynamically re-configurable capacity.

GMPLS is the union of existing MPLS solutions, MPLambdaS (MPλS) and label switching through TDM networks. MPLambdaS provides for the configuration of optical forwarding as well as features associated with MPLS such as label nesting and link bundling. The only possible interoperability issue is that of capacity granularity and the ability to multiplex clients and sub-divide the bandwidth of each wavelength. Used together with link/LSP bundling in GMPLS or link aggregation in ASON, or VCat it would be possible to efficiently allocate fine-grained capacity up to very high speeds (multiple wavelengths).

The organisational separation of IP and optical networks would mean that there is no direct link between MESCAL Solution Options 1 and 2 and DWDM networks, and therefore any of the technologies considered in D1.4 would be suitable for the dynamic provisioning of bandwidth. MESCAL Solution Option 3's use of MPLS however would allow for a closer integration as the IP network providers PCSs could now directly interface to the optical network's PCSs for faster more efficient provisioning.

6.7.3 Network Provisioning in the MESCAL functional architecture

The hierarchical relationship between functional components and the development of the concept of "plan then take care" for the management and control of QoS in IP networks was developed in TEQUILA [TEQUI,D1.4] and adopted by MESCAL in D1.1. Network planning and provisioning fits into this hierarchy as follows:

Hierarchy of Management/Control functionality: "plan then take care":

- Service Planning
  Defines services to be offered based on perceived customer demand and business objectives
- Network Planning/Provisioning
  Provisions sufficient physical network resources to meet service requirements
- Resource Provisioning/Traffic Engineering
  Configures the physical network, based on subscriptions
- Dynamic Traffic Engineering
  Dynamically adjusts network configuration based on actual traffic and network state (within limits imposed by off-line TE)
- Packet scheduling/forwarding
  Implements decisions of higher-layer algorithms in the data plane in real time

MESCAL has defined *Resource Provisioning Cycles*, both intra- and inter-domain to configure the network to meet perceived service demands (see section 10). These currently assume that the physical network is fixed, although the TE functional blocks are assumed to raise alarms to the off-line network planning processes when they are unable to accommodate the traffic demands within the existing physical network by soft configuration alone.

Thanks to the emerging capabilities of modern optical networks it is now possible to conceive of network resources (link bandwidth) being provisioned dynamically, which could be exploited by a *Network Provisioning Cycle* within the MESCAL functional architecture. This would involve algorithms deployed within a *Network Planning* functional block, which may be invoked by *QoS-*

Copyright © MESCAL Consortium, June 2005

based Service Planning, *Traffic Forecast*, or could be triggered by *Off-line Intra- or Inter-domain TE* when they are unable to satisfy the traffic demands within the existing resources.

6.7.4 Relationships between Network Planning and Traffic Engineering Algorithms

Traffic Engineering is assumed to operate within the constraints of the existing physical network. A common TE/planning algorithm for optimising physical and logical resources is not considered as this has major implications on the MESCAL TE algorithms and introduces too many degrees of freedom – not in the spirit of "plan then take care". This is further justified by the fact that for the majority, if not all, operators, the underlying transport networks, such as SDH or DWDM, support a number of client networks, such as PSTN or leased lines, in addition to their IP network offerings. Furthermore the networks are often operated by different administrative divisions. A common TE policy across both client and server networks is unlikely to be deployed as the transport infrastructure has to be optimised for the demands made by all clients and not just the IP networks.

In a similar way to *Offline Inter-domain TE* interacting with *Off-line Intra-domain TE*, as described in Section 10, to achieve a loosely-coupled optimisation of both inter- and intra-domain resources, Network Planning may need to interact with intra- and inter-domain TE to investigate "what-if" scenarios before committing to buying new physical resources.

6.8 Other functions and capabilities

The functional architecture covers those capabilities necessary for deploying and operating inter-domain QoS services. A provider may need other more general support functions such as fault and configuration management, but as these are not an explicit part of the inter-domain QoS provision problem they are not covered in this architecture. The role of dynamic network provisioning and the role of a network provisioning cycle is analysed in Section 6.7. As a result a *Network Planning* block has been added to Section 6.7 to demonstrate where this functionality is positioned. It should be noted that this study was limited to the level of updating the functional architecture. Detailed studies of interactions between *Network Planning* and underlying *Physical Connectivity Providers* or algorithms for optimising the deployment of *physical* resources, e.g. optical networks, are out of scope of MESCAL.

It is envisioned that rather than being entirely hard-coded at development or installation time, the behaviour of many of the MESCAL functions and algorithms can be influenced at run time by a *Policy Management* infrastructure. Policies are expected to cover the *SLS Management* and *Traffic Engineering* functional blocks. There are no explicit functional blocks shown to handle multicast services. As described in Section 6.10 it is assumed that multicast functionality distributed over several of the blocks and only two additional blocks have been identified: *Dynamic Group Management* and *RPF Checking*. These are introduced in Section 6.10 and are discussed in detail in Section 13.

For most providers, an important aspect of providing service differentiation is the means for charging appropriate rates for different service levels. Metering, rating, billing and other commercial aspects of QoS delivery are outside of the scope of MESCAL and are therefore not part of the specified functionality. The issues associated with financial settlements according to the various business models for interactions between network providers have been studied, however, and an analysis of the implications on the MESCAL solutions is documented in deliverable D1.4 [D1.4].

6.9 Functional Architecture Interaction Scenario

In this section we describe the interactions between the components of the functional architecture for creating a new inter-domain QoS-based service. In the following paragraphs the numbers in parenthesis refer to the numbered interactions in Figure 29.

Copyright © MESCAL Consortium, June 2005

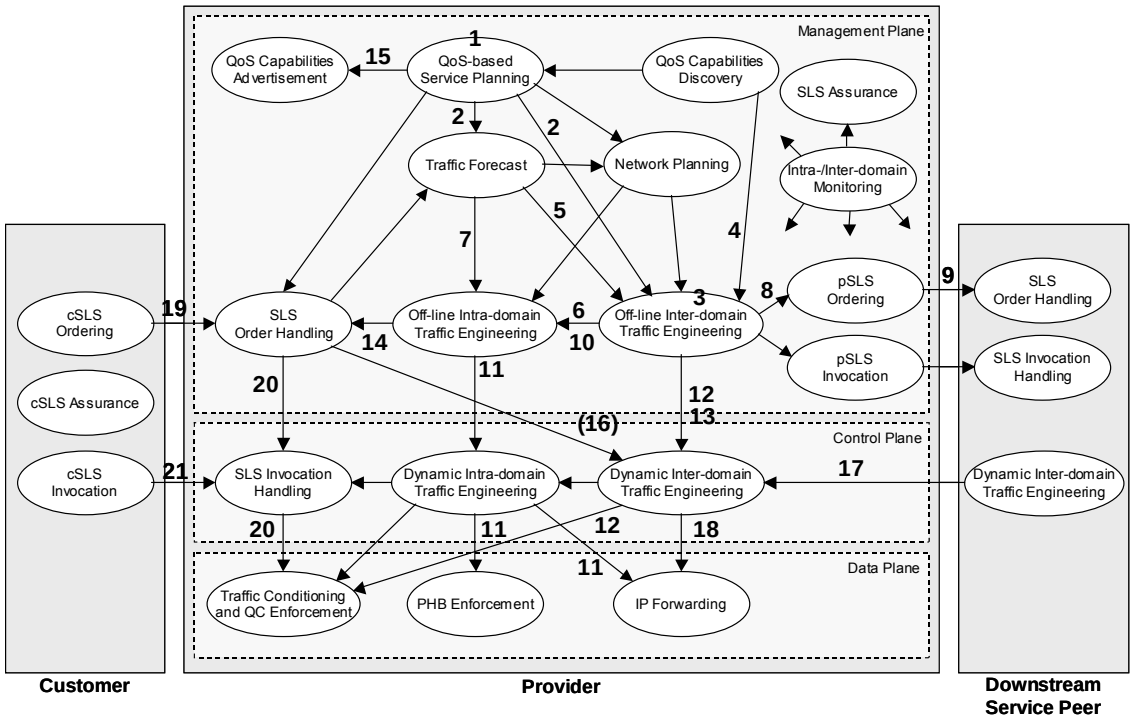


Figure 29 Functional architecture scenario

QoS-based Service Planning (1) identifies a new inter-domain service that could be offered to its customers, e.g. for viewing high quality streamed video from a set of servers located in remote ISPs. The business planning functions will specify the technical parameters of the e-QC (bandwidth, delay, etc.) that could be formed from combinations of its existing l-QCs and o-QCs already offered by its peers (which it will be aware of via the QoS Capabilities Discovery block) to the remote destinations. Part of this process will also determine the expected demand from its customers and the cost constraints, including the cost of provisioning l-QC capabilities and capacities as well as the price it is prepared to pay its peers for the o-QCs.

The e-QC QoS parameters, the set of required destinations, and cost constraints are passed to Off-line Inter-domain Traffic Engineering (2) to trigger a new Inter-domain RPC. The anticipated demand is passed to Traffic Forecast to generate a new traffic matrix for this RPC (2).

Off-line Inter-domain TE first of all invokes its Binding Selection algorithms (3) to discover suitable bindings of l-QCs and o-QCs. It will discover appropriate peer ISPs and their available o-QCs – destinations, QoS parameters, and cost – via the QoS Capabilities Discovery Block (4). After selecting feasible l-QC-o-QC bindings, it will run its Inter-domain Resource Optimisation algorithm to find the most suitable bindings (and the bandwidths of the required pSLSs) that meet all of the traffic demands specified in the traffic matrix (5) from Traffic Forecast (including the new demands for the new service).

While Inter-domain TE is focussing on optimising inter-domain resources – the QC bindings and the pSLSs with peers – it is necessary to ensure, first of all, that there are sufficient intra-domain resources in terms of l-QC capacities between the (anticipated) customers and the selected egress routers, and, secondly, that the intra-domain configuration to meet the selected inter-domain bindings is not sub-optimal. It may be the case that the second or third most optimal set of inter-domain bindings may offer a better overall intra- and inter-domain solution. For these reasons, the Inter-domain Resource Optimisation algorithms will present candidate solutions to Off-line Intra-domain TE (6). Off-line Intra-domain TE algorithms will then calculate the intra-domain cost of the proposed inter-domain solution based on the intra-domain traffic forecast matrix (7) raised by the candidate solutions (including existing demands as well as those anticipated by the new service to the proposed egress routers).

Copyright © MESCAL Consortium, June 2005

Off-line Inter-domain TE will select several candidate solutions – the most optimal considering intra- and inter-domain costs, as well as back-up solutions – which will be negotiated with its peers via the pSLS Ordering functional block (8).

PSLS Ordering initiates a negotiation process with the candidate peer ISPs for the pSLSs as previously determined by Off-line Inter-domain TE (9). It is the responsibility of pSLS Ordering to enforce a transaction in the case of negotiations for multiple pSLS, some of which may fail if the bandwidth to certain destinations is unavailable or the cost is too high, for example.

Once the pSLSs have been negotiated and agreed, Off-line Inter-domain TE triggers Off-line Intra-domain TE (10) to configure the selected intra-domain solution. Intra-domain TE will configure revised OSPF metrics and PHB capacities (or LSPs in the case where MPLS-TE is used for intra-domain provisioning) and deploy these in the routers via the Dynamic Intra-domain TE functions (11).

Off-line Inter-domain TE will configure the egress routers with the correct DSCP mappings for the selected l-QC to o-QC bindings (12). It will also configure the q-BGP processes in the Dynamic Inter-domain TE blocks (13) with appropriate policies for processing the q-BGP messages that will subsequently arrive from the downstream peer ASs where new pSLSs have been established.

Off-line Inter-domain and Intra-domain TE will also forward to SLS Order Handling the Inter- and Intra-domain Resource Availability Matrices for the current configuration (14). These will allow SLS Order Handling to determine whether there is capacity for future c/pSLS subscriptions from customers or upstream peer ISPs. Also at this point the QoS-based Service Planning functions will advertise the new e-QC capabilities to upstream ISPs and potential customers via the QoS Capabilities Advertisement functions (15).

In the downstream peer ISPs, once a new pSLS has been agreed, SLS Order Handling will configure the q-BGP processes (16) to forward q-BGP announcements to its new customer ASs for the destinations and o-QCs that are subject to the new pSLS.

q-BGP announcements will subsequently be received from the downstream ASs (17). The Dynamic TE processes, which include the e- and i-q-BGP speakers, will select appropriate inter-domain routes according to the policies they were previously configured with by Off-line Inter-domain TE (18). From this point on the ISP is able to forward packets to remote destinations with the required QoS, however the ISP's customers (end customers as well as upstream ISPs) must first establish SLSs to use these capabilities.

A customer wishing to subscribe to the new inter-domain service will initiate a c- or pSLS negotiation with SLS Order Handling (19). The latter will consult the Resource Availability Matrix and the repository of already subscribed SLSs to determine whether there is sufficient capacity for the request. Once the SLS has been agreed, the traffic conditioners in the ingress routers will be configured for the new SLS (20). In the case of an end-customer, when a policy of SLS over-booking is deployed in the ISP, each flow which is part of the overall pSLS subscription will be signalled (21) via the SLS Invocation Handling components in the ingress routers where admission control algorithms will determine whether there is sufficient capacity to avoid QoS deterioration.

As s- and pSLS subscriptions change over time the current intra- and/or inter-domain configurations may not be sufficient to allow future anticipated demands. In this case Traffic Forecast will initiate new intra- and/or inter-domain RPCs. These may result in modified intra-domain resources (e.g. OSPF metrics, PHB configurations), modifications to existing pSLSs (e.g. increasing or reducing bandwidth), or even brand-new QC bindings with new peers that will require new pSLSs to be established. In the case where existing pSLSs will be modified, the Binding Activation sub-component of Off-line Inter-domain TE will trigger the pSLS Invocation processes.

Copyright © MESCAL Consortium, June 2005

6.10 MESCAL multicast functional architecture

6.10.1 Overview

The proposed multicast functional architecture (shown in Figure 30) is consistent with the overall MESCAL scenario, and most of the components can be included in or mapped onto the blocks in the general architecture. In this way the corresponding implementation can be compatible with its unicast counterpart. From this point of view, the multicast architecture is not new nor is it independent of the general MESCAL model. For simplicity we do not include all the functional blocks in the overall architecture, but only illustrate the components that should be necessarily associated with multicast services. Meanwhile some new blocks are appended exclusively for multicast services (with * inside the block). On the other hand, there is one difference in defining service peers in the figure: we name the server side ISP (the right most part in the figure) the *upstream* provider instead of a downstream one because the multicast traffic is flowing in the opposite direction of the unicast flows. This implies that the domain level multicast SLS ordering/handling is always from the receiver to the source.

MSLS Order Handling is a subset of *SLS Order Handling* in the general architecture, and it is responsible for subscription level admission control on multicast customers. The most distinguishable aspect from its unicast counterpart is that the functional block negotiates with multicast group members/receivers instead of data sources. The Offline Multicast TE block will provide mSLS Order handling the resource availability of the engineered network for multicast traffic such that the later is able to decide whether to accept new mSLS requests for receiving multicast data. This type of mSLS requests can come from both local multicast customers and the ISP's peering neighbours.

Offline Multicast TE can be further divided into intra- and inter-domain parts, which are respectively embedded in the corresponding offline TE blocks in the general MESCAL architecture. The task of this functional block is to map the demanded multicast flows onto the physical network resources and configure these resources in order to accommodate the forecasted traffic from both local customers and peering ISPs. Furthermore, in order to achieve end-to-end QoS requirements across domains, the QC mapping and binding selection/activation process still apply to the multicast scenario, and there should be minimum, if not no direct impact on the conventional mechanisms for unicast traffic. The process of Offline Multicast TE is also in a centralised manner within an AS during each RPC.

MpSLS Ordering is included in *pSLS Ordering* in the general architecture, and it interacts with the mSLS Order Handling block in the upstream service peer. Specifically, this block takes the responsibility of negotiating new multicast pSLSs with the upstream ISP, and this negotiation is based on the binding selection algorithms from the offline multicast TE block.

Dynamic Group Management can be appended to the *cSLS invocation handling* in the general architecture specifically for multicast services. In order to ensure that the network is not overwhelmed with multicast traffic resulted from the policy of over-reserving resources at the subscription level, admission control should be introduced in group management for rejecting excessive join requests on new group sessions. Moreover, this functional block should also have the capability of dealing with heterogeneous QoS requirements from members who subscribe to a common group session.

Similar to the offline scenario, *Dynamic multicast routing* can be regarded as part of the Dynamic TE blocks in the general architecture, and it has the functionality of constructing and updating real time multicast trees according to the group membership dynamics. When the Designated Router (DR) receives an IGMP membership report, the task is how to deliver the QoS join request towards the source, such that a feasible path can be found to carry the multicast traffic to the receiver. Moreover, this block should also provide capabilities of dynamic traffic engineering such as bandwidth conservation and load balancing etc.

mpSLS Invocation basically has the similar functionality to the corresponding *pSLS Invocation* in the general functional model. The only difference is that the interaction is with the upstream ISP in terms of the usage of multicast pSLS dynamics from receiver peer's perspective.

Copyright © MESCAL Consortium, June 2005

PHB enforcement for multicast services is contained in its counterpart in the general architecture and it mainly considers how to treat multicast packets with proper PHBs at the core network. Compared to the unicast scenario, multicast packets can be replicated at any branching point where two or more join requests are merged together. How to treat replicated packets destined to group members with heterogeneous QoS requirements becomes a new issue. As it is known that conventional multicast trees are recorded through group state maintenance within the network, how to enable these trees to exhibit multiple PHBs without significantly extending core router forwarding architecture is another issue to be coped within this block.

Multicast forwarding is part of *IP forwarding* in the general architecture, and it basically has two tasks: first, when a multicast packet arrives at the incoming interface, the router should replicate it and forward the packets on all the outgoing interfaces where group join requests are received. Second, at each outgoing interface the replicated packets should be treated with proper PHBs that correspond to the original QCs expressed in the join requests from downstream group members. The behaviour of multicast forwarding should also obey the reverse path forwarding (RPF) rule.

RPF checking is a packet-level mechanism for avoiding loops dedicated to multicast traffic delivery. At each multicast router, if the packet is not received from the interface on the shortest path back to the source, this packet will be silently dropped. This guarantees that multicast traffic is always forwarded along the shortest path from the source to individual group members. In the MESCAL solutions, even if QoS routing is to be used in multicast tree construction, the multicast RPF checking mechanism should still take effects as a necessary constraint for multicast forwarding.

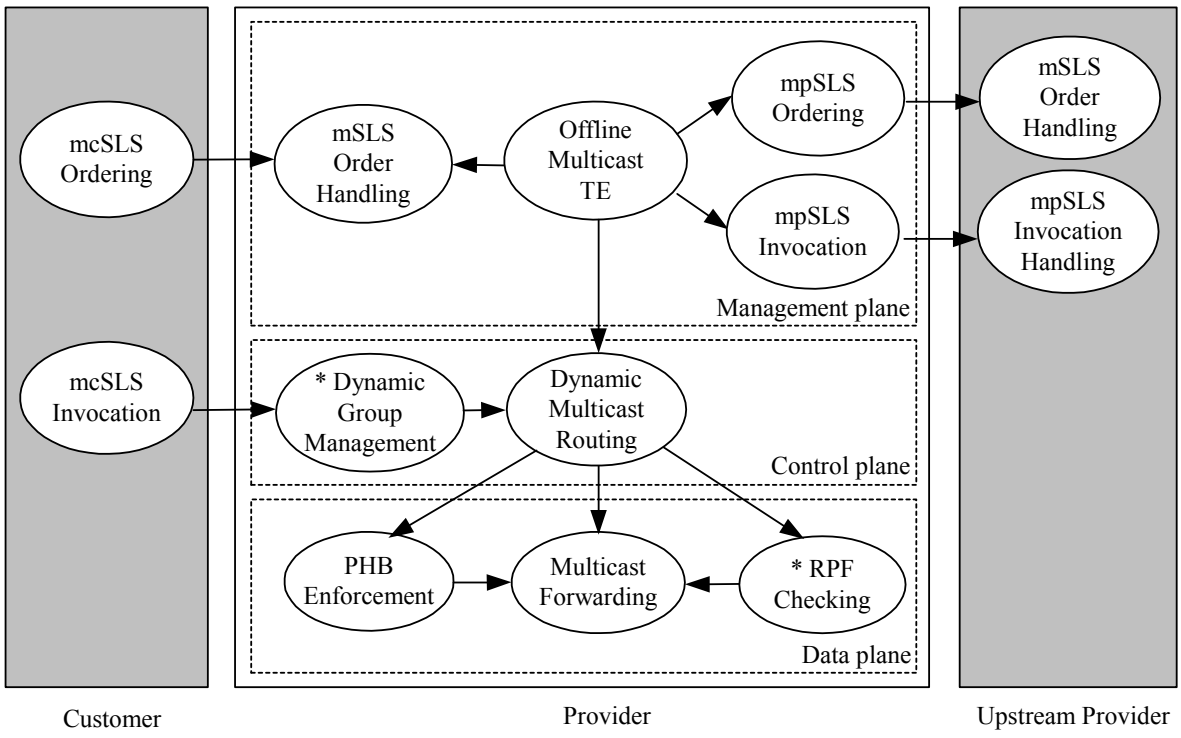


Figure 30. MESCAL multicast functional architecture

Copyright © MESCAL Consortium, June 2005

7 SOLUTION SPACE

7.1 Introduction

Based on the customer requirements listed in section 3.2.3, the MESCAL project has identified, at a high level, two major end-users categories. These categories differ at the level of the QoS guarantees they require, the topological scope of their SLSs and by the permanence of their communications requirements.

Residential customers may subscribe to IP services such as VoIP, video on demand and broadcasting services. These users may want to reach any available destination at any time without being tied to a single destination, or limited set of destinations at subscription time. The duration of the communications between one of these end-users and a *specific* content provider or peer customer may be short (just the duration of a service transaction for instance) and the frequency of interactions can be sparse. In the case of peer-to-peer file sharing applications or premium web-browsing, for example, the total sum of the communications requirements from one customer to a large number of destinations may be relatively long lived with a dense frequency of interactions.

On the other hand, *corporate* customers, may request specific, strong guarantees for supporting particular mission- or safety-critical applications and services, such as IP VPNs, Virtual Leased Lines, corporate VoIP services, remote control of equipment such as control of robot arms or surgical instruments. These requirements are usually to a limited, small set of destinations, the relationship between the communicating entities is long-lived and the frequency of interactions is usually dense.

These two categories can be seen as two extremes: the residential customer wants to communicate with all destinations with better-than-best-effort service levels, while the corporate customer wants a point-to-point pipe to a named destination with hard upper bounds on QoS and a constant bandwidth. Obviously these are two extreme cases and a range of customer categories could be identified between these two, such as the customer requiring hard upper bounds on delay to a large but limited set of destinations with statistically guaranteed throughput.

From a contractual viewpoint these requirements introduce some variations in the way the following SLS parameters are handled:

- Topological scope: which is "any" for residential customers but is usually a limited set of specific destinations for corporate business customers.
- End-to-end QoS guarantees: residential customers may have only loose requirements which could be captured in qualitative parameters while corporate customers may require explicit hard guarantees with specific values for the upper bounds on loss, delay and jitter, for example.
- End-to-end bandwidth guarantees: corporate customers require at least a statistical guarantee, if not a hard peak-rate allocation, of the bandwidth specified in its SLSs. Residential customers may be content with best effort bandwidth availability or may require some statistical guarantees, but they are unlikely to be willing to pay the premiums associated with peak rate end-to-end bandwidth reservations.

It is intuitively obvious that end-to-end hard QoS performance and bandwidth guarantees cannot be offered to all Internet users with the level of dynamics that characterises the large number of residential customers. This is mainly due to scalability reasons: IntServ was widely seen as unscalable even *within* domains, for example. In order satisfy the requirements of the aforementioned customer categories MESCAL has specified a solution space encompassing three main service options.

These service options are discussed in . Note that a given provider could support all or only a subset of these service options. In section 7.3 we provide the details of the MESCAL solution, evaluate its conformance against the provider and customer requirements and map it to the MESCAL functional architecture.

Copyright © MESCAL Consortium, June 2005

7.2 Service Options

Previous chapters have described the Inter-domain QoS requirements that the MESCAL solution must meet, from both provider and customer perspectives. MESCAL has identified three service options characterised by the level of guarantee they can provide:

- The *Loose Guarantees* service option, which globally aims at providing better Internet-based services, but doesn't provide any strong guarantees.
- The *Statistical Guarantees* service option, which offers QoS performance guarantees for specific destinations and which allows some loose end-to-end bandwidth guarantees.
- The *Hard Guarantees* service option, which improves the above option with strong end-to-end bandwidth guarantees.

These service options provide distinct and different service characteristics, which enable providers to meet the requirements of a diverse range of customers, see Table 3, below.

	Service Options		
Characteristics	Loose	Statistical	Hard
E2E QoS Performance	Qualitative	Qualitative/Quantitative (statistical guarantee)	Quantitative
E2E Bandwidth	No guarantee	Statistical guarantee	Guaranteed
Topological Scope	Any reachable destination	Specific destinations	Specific destinations

Table 3: MESCAL Service Options

The MESCAL Loose service option enables a provider to offer customers access to differentiated transport services, where each differentiated service is related to a Meta-QoS-Class. It is envisaged that providers throughout the Internet will implement a small number of well-known Meta-QoS-Classes. Inter-domain QoS services are then created by constructing paths across those domains that support a particular Meta-QoS-Class. In effect, a set of parallel "internets" are deployed, each offering service levels associated with a specific Meta-QoS-Class. The guarantees associated with the Loose service are restricted to qualitative services, although it is anticipated that the characteristics of each Meta-QoS-Class based service will be based on common application requirements, for example VoIP. The Loose service option does not provide any end-to-end bandwidth guarantees because the option enables any destination to be reached, without prior identification in the cSLS/pSLS. The objective of the Loose service option is to address the requirements of a large population of users, while keeping the network engineering as simple as possible by supporting relaxed service guarantees.

The MESCAL Statistical service option provides customers access to inter-domain QoS services with firmer guarantees than the Loose option. The Statistical service option is able to provide a qualitative QoS service, although quantitative services where values for packet delay and loss are specified can also be offered. Additionally, an end-to-end bandwidth guarantee is provided within statistical bounds. An Inter-domain QoS service based on the MESCAL Statistical option is created by constructing paths across domains that are able to guarantee their QoS capabilities. QoS services can either be constructed to meet specific quantified QoS constraints or the Meta-QoS-Class approach can be used for offering qualitative services. A distinguishing feature of this service option is that the guarantees are statistical. It is a policy decision for each provider to decide the level of the guarantee that it wants to offer and it is to be expected that QoS services with firmer guarantees will require higher allocation of resources in the provider's network.

The MESCAL Hard service option provides customers with strict inter-domain performance guarantees. The Hard service option is targeted at providing services with quantitative QoS and bandwidth guarantees with a high probability of fulfilment. An Inter-domain QoS service based on the MESCAL Hard option is created by constructing paths across domains that are able to guarantee their

Copyright © MESCAL Consortium, June 2005

QoS capabilities to the required level. It is envisaged that network resources will have to be permanently allocated for this service and consequently, the MESCAL Hard service option is suitable for services that can justify the high costs that will be associated with the service. The Hard service option will be appropriate for a small number of added-value services, such as critical business services.

7.3 The MESCAL Solution

The purpose of this section is to describe the MESCAL solution to supporting the three identified service options. The MESCAL solution is directly mapped to the Functional Architecture, for each of the service options and is conformant with both the customer and provider requirements, which have been identified in section 3.2. This section provides the detailed description of all required QC-operations in order to achieve the objectives of each of the aforementioned service options.

Based on the service options described above, the MESCAL project has designed three solution options that target three different end-users categories:

- The Loose Guarantees solution option (LGSO): this solution option aims at providing an implementation of the Loose Guarantees service option. This option allows having some QoS treatment when this is possible. No strict guarantees are assumed by this option.
- The Statistical Guarantees solution option (SGSO): this solution option is based on the statistic service option.
- The Hard Guarantees solution option (HGSO): this solution option gives hard guarantees to the customers in terms of QoS treatment and bandwidth.

7.3.1 Loose Guarantees Solution Option

The LGSO aims at providing an implementation of the Loose Guarantees service option, which has been introduced above.

A light version of this solution option is also presented. This version changes the way the mapping operation is done and makes the signalling operation less heavy than the non-light version.

7.3.1.1 Use of Meta-QoS-Class concept

The underlying philosophy of meta-QoS-classes relies on the assumption that wherever end-users are connected they use similar applications in similar business contexts. Customers also experience the same QoS difficulties and are lead to express similar QoS requirements to their respective service providers.

within this service option, we assume that providers define and deploy similar classes of service because they are in general confronted with the same customers requirements. These classes target to support applications, which have similar QoS constraints. There is no reason to consider that a provider in Japan for example would design a "Voice Over IP" l-QC with short delay, low loss and small jitter while another one in Germany would have a completely different view. Therefore, constraints are implicitly imposed by applications to the network, independently of where the service is consumed or accessed.

The *Meta-QoS-Class* concept is actually an abstract concept. It is not a real l-QC provisioned in real networks. A *Meta-QoS-Class* is defined to serve dedicated services (e.g. VoIP) and can specify a set of boundaries for pertinent QoS performance attributes. This point is a key funding aspect for the LGSO.

In addition, the *Meta-QoS-Classes* could inherit from each other as follows:

Copyright © MESCAL Consortium, June 2005

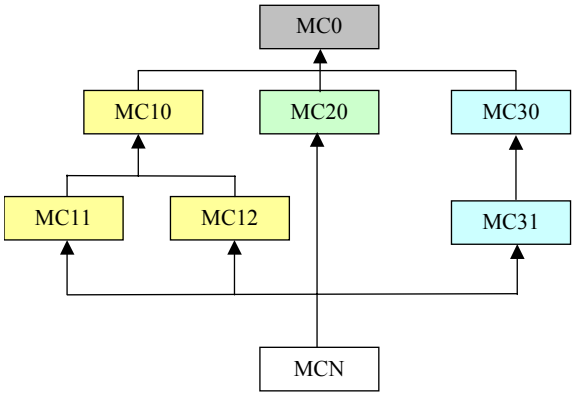


Figure 31: *Meta-QoS-Class* inheritance example diagram

In this example, MC0 represents the BE effort *Meta-QoS-Class* and MCN is the "impossible to get" neutral element. Each branch of the tree is designed to be suitable for different QoS applications. The categories of applications are generic and are described in terms of network performance (mainly in terms of sensitivity to delay, jitter, loss or any other network performance characteristic which can be qualitative and/or quantitative).

If several grades of QoS are considered for an application category, *Meta-QoS-Classes* can be defined to form a hierarchical tree. In this particular example, this means that MC11 would also be suitable for conveying flows requesting MC10 and MCN could potentially be used for any kind of traffic since it represents the neutral element. This hierarchical ordering of *Meta-QoS-Class* is an assumption and, at this stage, it is still uncertain whether branch splitting (MC11 and MC12 for instance) should be conceptually kept in future specification.

7.3.1.2 QC-classification

Within the LGSO, each provider must classify its l-QCs with regard to *Meta-QoS-Classes*. This is denoted by *QC-classification* process. This operation occurs each time a new l-QC is designed or an existing one is re-engineered. An l-QC can potentially satisfy several *Meta-QoS-Classes*.

For instance, a provider could have defined:

- `l-QC20`: satisfies MC0, noted `l-QC20[MC0]`,
- `l-QC21`: satisfies MC10 and MC20, noted l-QC`21[MC10,MC20]`
- `l-QC22`: satisfies MC11, noted l-QC`22[MC11]`.

In the *Light approach*, an l-QC can satisfy one and only one *Meta-QoS-Class*. *Meta-QoS-Classes* inheritance properties cannot be used. *Meta-QoS-Class* concept is only used for mapping and binding purpose.

7.3.1.3 QC-mapping

From a business perspective, a provider can logically express the need to extend its own classes of service across the Internet. In particular, this means that a flow originated in the provider's domain, with an indication of the requested class of service, should experience similar and coherent treatment when crossing the set of autonomous systems up to its final destination. Therefore, providers must establish peering contracts (pSLSs) in order to extend their QoS capabilities.

Before the establishment of any *pSLS*, the provider requesting the pSLS must proceed to a *QC-mapping* in order to identify the whole set of potentially compatible bindings between its own l-QCs and the remote's o-QCs with the objective to extend the scope of its services beyond its boundaries.

Copyright © MESCAL Consortium, June 2005

Within the LGSO, the QC-mapping concerns only the *Meta-QoS-Classes* that the provider decides to extend. This compatibility-mapping criterion is ensured by the *Meta-QoS-Class* concept. Two classes are declared to be compatible for mapping if they belong to the same *Meta-QoS-Class*, directly or by inheritance.

For achieving this QC-mapping the neighbour AS must indicate to the requestor AS if it supports each of the requested *Meta-QoS-Classes*.

In the *Light approach*, the requesting provider will consider all possible mappings between each of its l-QCs **with only one** of the remote o-QC providing that the remote o-QC belongs to either the same or a better *Meta-QoS-Class*.

7.3.1.4 QC-binding

In the context of the MESCAL solution for supporting the loose guarantees service option, QC-binding concerns only the *Meta-QoS-Classes* the requesting AS decides to extend. The QC-binding process becomes very simple and can be summarised as a binary assessment: *does the peering partner support the requested Meta-QoS-Class or not?* In this case, there can be a very limited number of combinations.

At the end of this process, several l-QCs from the requesting AS can be potentially used for transporting datagrams that belong to the same *Meta-QoS-Class*. On its side, the peering provider can choose to select only one or several of its compatible l-QCs to fulfil the contractual terms of the pSLS.

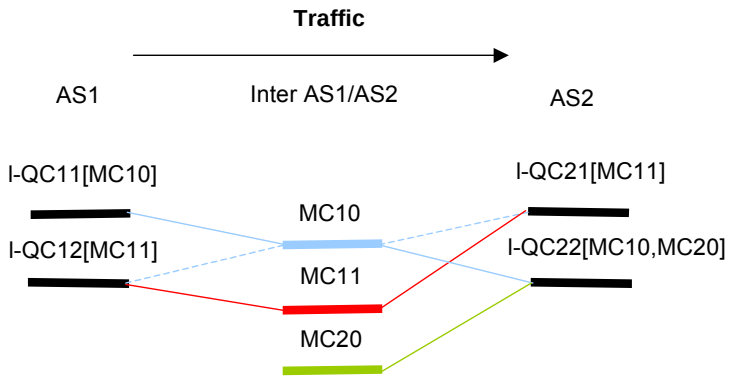


Figure 32: Example of the QC-binding operation

In Figure 32 we show an example of the QC-binding operation. Within AS1, as a result of the QC-classification operation, the MC10 traffic can be assigned to l-QC11 or l-QC12. The MC11 traffic can only be transported with l-QC12. MC20 is not supported by AS1. In AS2, as a result of the QC-classification operation, the MC10 traffic can be assigned to l-QC21 or l-QC22. The MC11 traffic can only be assigned to l-QC21. MC20 is transported by QC22.

In the above example, at the highest level, the QC-binding leads AS1 to exchange MC10 and MC11 traffic. In detailed the following binding has been achieved:

- l-QC11 ==> l-QC21 or l-QC22
- l-QC12 ==> l-QC21 or l-QC22

Depending on the *Meta-QoS-Class*, one of the 4 possible bindings is used.

Copyright © MESCAL Consortium, June 2005

In the *Light approach* the same figure would become:

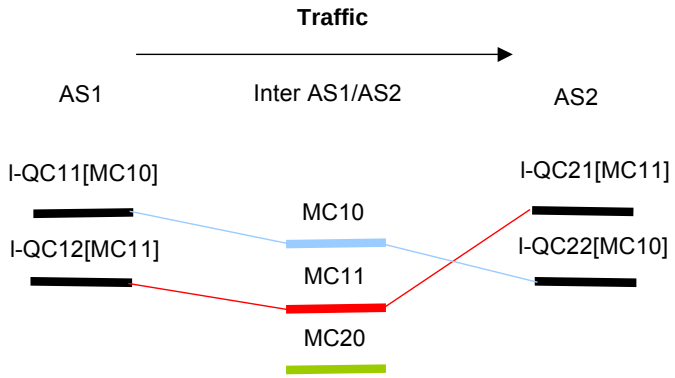


Figure 33: Example of the QC-binding operation with the Light approach

In AS1, as a result of the QC-classification operation, the MC10 traffic can be assigned to l-QC11. MC11 traffic can only be assigned to l-QC12. MC20 is not supported by AS1.

In AS2, as a result of the QC-classification operation, the MC10 traffic can be assigned to QC22. MC11 traffic can only be transported with QC21. MC20 is not supported by AS2.

In the above example, the QC-binding leads AS1 to exchange MC10 and MC11 traffic. In detailed the following binding has been achieved:

- l-QC11 ==> l-QC21
- l-QC12 ==> l-QC22

7.3.1.5 QC-implementation

7.3.1.5.1 QC-Indication

An important aspect of this approach is that *Meta-QoS-Classes* are used to indicate the requested QoS treatment across the Internet. A *Meta-QoS-Class* indicator is used both intra-domain and inter-domain. This could be a global value agreed by all providers or a local value understandable by two adjacent eBGP peers. The DSCP can be used for this purpose with the limitation of 64 values.

In intra-domain, the end-user submits a datagram with an indication of the requested *Meta-QoS-Class*. The first provider's router chooses an appropriate l-QC for transporting this datagram within the domain (since several l-QCs can potentially satisfy the same Meta-QoS-Class). This l-QC is used cross the domain and the QoS of service experienced by this datagram is compliant with that l-QC. Nevertheless, the *Meta-QoS-Class* indicator is kept in the datagram.

When the datagram reaches a domain boundary, the l-QC indicator cannot be used anymore in the remote domain and the *Meta-QoS-Class* indicator is used instead. The receiving provider then uses its own l-QC to transport the datagram up to its border router in the domain. Using a Meta-QoS-Class indication allows splitting an l-QC while avoiding the QC-splitting problem.

In the *Light approach,* there is no *Meta-QoS-Class* signalling indicator. The end-user submits a datagram using an l-QC indicator. The egress AS is supposed to indicate the remote ingress l-QC that will be used by the ingress AS, thanks to the DSCP field of the IP datagram. By definition of the mapping and splitting processes, there is no possible QC-splitting.

It should be noted that *Meta-QoS-Class* indication allows outclassing traffic (i.e. treat the traffic within a better MC) when crossing an external domain because the Meta-QoS-Class indicator is transported end-to-end by the datagram. When exiting the remote domain, the datagram can be transported by a more appropriate remote l-QC, as originally requested by the end-user.

Copyright © MESCAL Consortium, June 2005

In the *Light approach,* outclassing is also supported but once a datagram has been outclassed it cannot go back to its originally requested *Meta-QoS-Class* since the datagram doesn't convey such indicator.

In Figure 34, l-QCij have been classified as follows:

- AS1: l-QC11[MC10], l-QC12[MC11]
- AS2: l-QC21[MC11], l-QC22[MC10, MC20]
- AS3: l-QC31[MC11], l-QC32[M20], l-QC33[MC10]

In order to keep the figure simple, *Meta-QoS-Class* MC0 is not shown.

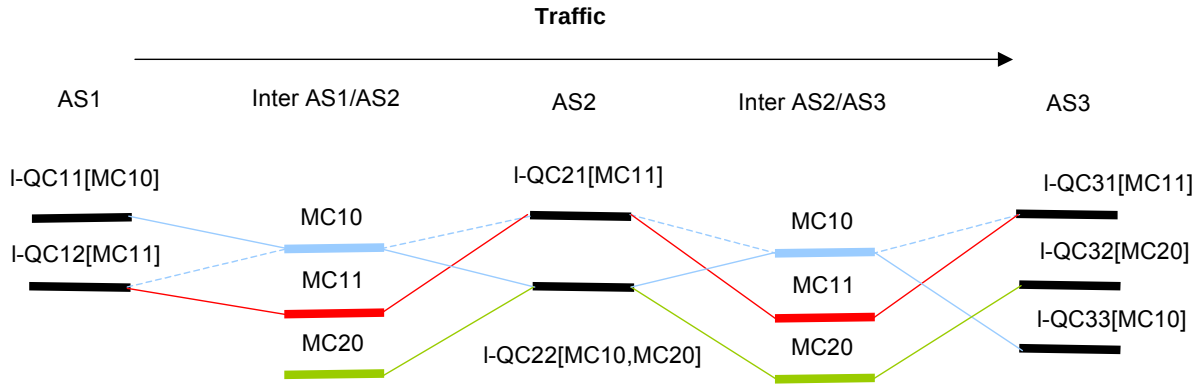


Figure 34: QC bindings in the name of *Meta-QoS-Classes*

Considering in AS1 an IP datagram marked with l-QC11 in the name of *Meta-QoS-Class* MC10 (hereafter noted l-QCij{MCx}), l-QC11 can be bound to l-QC21 or l-QC22 since those two classes are respectively mapped to MC11 (which inherits from MC10 in this example) and MC10. l-QC22 binding is probably the optimal binding for MC10 but l-QC21 is valid too. The choice to use l-QC21 in AS2 outclasses the traffic sent by AS1 for *Meta-QoS-Class* MC10. Outclassed bindings have been indicated with dotted lines.

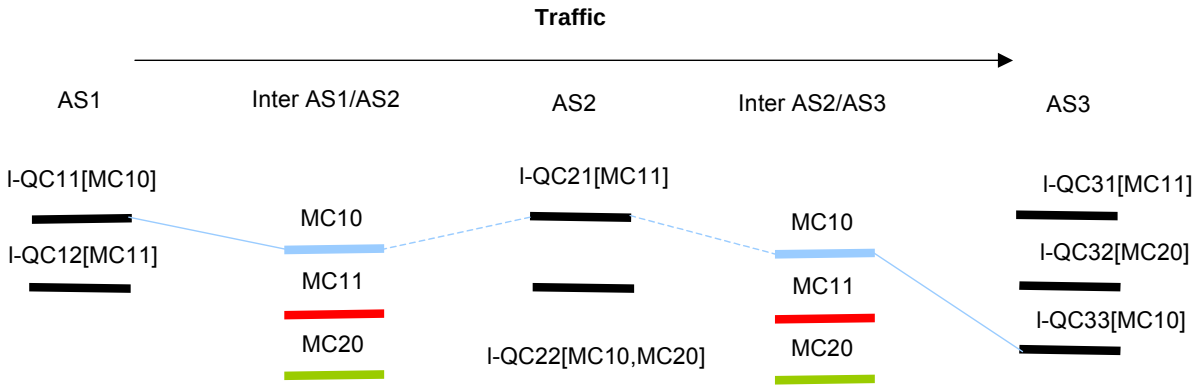


Figure 35: Temporarily outclassing example

In Figure 36, AS1 has deployed three l-QCs. One of them, l-QC11, has been declared (l-QC-classification operation) as a member of a particular *Meta-QoS-Class*. In the name of this *Meta-QoS-Class*, QC bindings have been achieved iteratively across all ASes. All ASes have gone through the same process, no matter the order in which the bindings have been established. The resulting "l-QC" bindings, for this particular *Meta-QoS-Class* are depicted in red (bold for black and white restitution support).

Copyright © MESCAL Consortium, June 2005

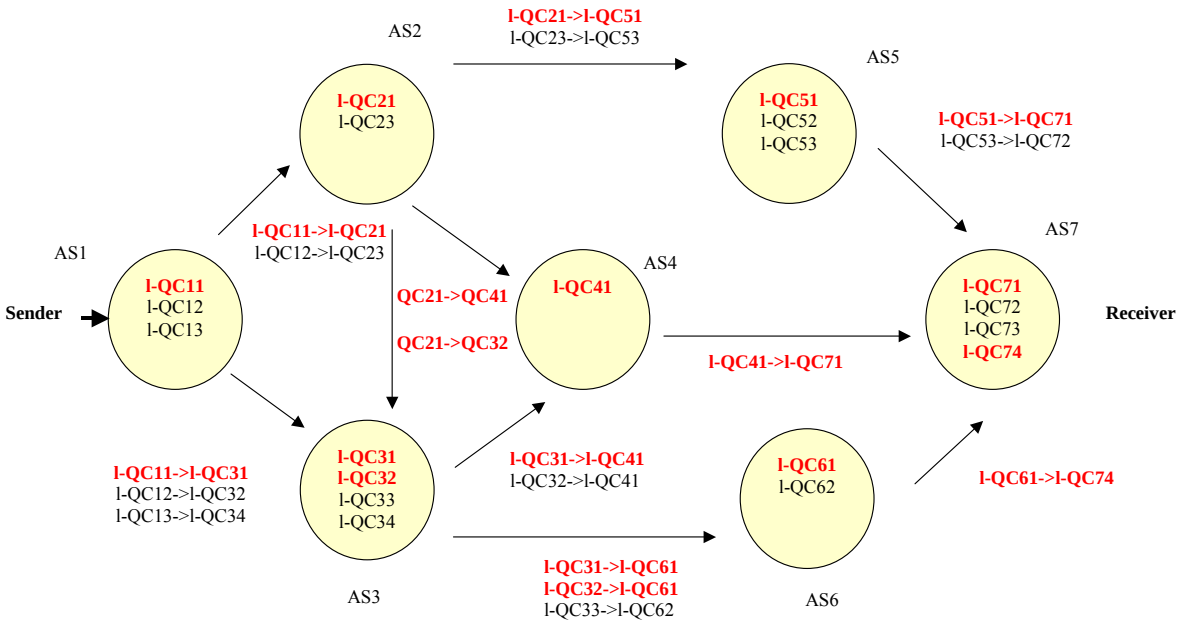


Figure 36: Following QC11 through contractual cross binding

From AS1 perspective, l-QC11 has been extended throughout the whole topology. Any sender from AS1 can reach any receiver anywhere through l-QC11 extension. At this stage, there are several possible paths from the sender to the receiver following l-QC11 extension. We'll see in the paragraph "Intra-domain and inter-domain routing aspects" how we propose to select only one path.

Figure 36 shows a connected topology. This solution option is interesting only if these bindings become common practice, so that each provider can see its own l-QCs extended throughout almost the whole Internet. However, we may reasonably expect some holes even if this solution option is largely and globally spread. The figure shows unidirectional bindings but it should be possible to establish bi-directional bindings.

7.3.1.5.2 Intra-domain and inter-domain routing aspects

7.3.1.5.2.1 Inter-domain routing: path selection

In this approach, the Internet appears as a set of parallel *Meta-QoS-Class* planes. Each *Meta-QoS-Class* plane consists of all the l-QCs bound in the name of the same *Meta-QoS-Class*. When an l-QC maps different *Meta-QoS-Classes* then it belongs to all the different *Meta-QoS-Class* planes.

We assume that in a *Meta-QoS-Class* plane, all paths are, to a reasonable extent, treated equally. Therefore, the problem of path selection amounts to: do your best to find one path for each *Meta-QoS-Class*. We rely on a BGP-like protocol for the path selection process. We call this protocol q-BGP, this protocol selects and advertises one path for each *Meta-QoS-Class* plane per destination.

When, for a given *Meta-QoS-Class* plane, there is no path available to a destination, the only way for a datagram to travel to this destination is to use another *Meta-QoS-Class* plane from start. The only *Meta-QoS-Class* plane available for all destinations is the best-effort *Meta-QoS-Class* plane (also known as "the Internet"). There's no straightforward solution to change from one plane to another on the fly. So, there's no straightforward way to span a *Meta-QoS-Class* plane hole by a best-effort bridge.

When a datagram enters an AS, the AS must know in which *Meta-QoS-Class* plane it belongs to in order to retrieve the egress point selected by q-BGP and also to apply the correct l-QC. QC-indication as described in 7.3.1.5.1 applies here: each datagram should convey an indicator of the *Meta-QoS-Class* it refers to.

Copyright © MESCAL Consortium, June 2005

7.3.1.5.2.2 Intra-domain routing: from the AS ingress point to the AS egress point

The intra-domain routing should also take into account the *Meta-QoS-Class* concept.

In a domain, each router will have to maintain one routing plane per *Meta-QoS-Class*. Indeed, since an l-QC can belong to several *Meta-QoS-Classes*, the same l-QC may be used for transporting traffic on behalf of different *Meta-QoS-Classes*. Egress points, for a same destination but for different *Meta-QoS-Classes,* may be different even if the same l-QC is used for crossing the domain up to the egress point. Intra-domain routing must be achieved on the destination and the *Meta-QoS-Class* indicator.

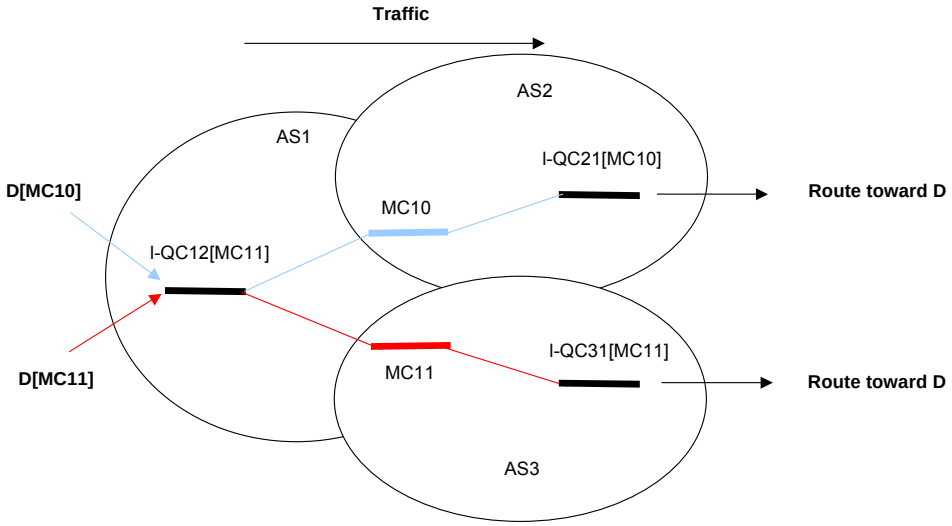


Figure 37: Example of an l-QC belonging to several Meta-QoS-Class

In the above example, two sources want to reach destination "D". D is not located in AS3 and AS2. A source (blue one) uses the Meta-QoS-Class MC10 and the other the MC11 (red one). They use l-QC12 to cross the first domain but the egress point of the domain is different. Routing cannot rely only on the destination address but on both the destination address and the MC. When an end-user or an external peering provider injects some traffic in the domain, the first provider router is responsible for selecting the l-QC to use for reaching the egress point. The chosen l-QC must support the requested *Meta-QoS-Class*.

In the *Light approach,* an l-QC can belong to only one *Meta-QoS-Class*. When a border router learns a destination "D" (because a pSLS exists) on behalf of QC-binding enforced, the q-BGP path selection process selects the most appropriate egress point and made it known to the intra-domain routing within the l-QC routing plane concerned with this the QC-binding. If the domain binds several l-QC on the same remote l-QC, the learned destination is flooded into the corresponding l-QC planes. Consequently, there is one routing plane per l-QC and routing must take into account both the l-QC and the destination address.

7.3.1.6 IPv6 support

This approach does not use any specific IPv4 capabilities other than the DSCP field in order to signal the l-QC to use. If this solution option can be implemented using IPv4 it should also be supported by an IPv6 infrastructure.

Since it might be easier to code the MC indicator in one of the IPv6 options, this protocol may be convenient for MESCAL LGSO.

7.3.1.7 QoS Guarantees

in the context of the LGSO approach, QoS is achieved thanks to a cascade of pSLS. If previously established pSLS are cancelled, any cSLS relying on those contracts becomes invalid. Network

Copyright © MESCAL Consortium, June 2005

accessibility cannot be ensured, and "holes" can appear anywhere at any time. The provider who establishes a cSLS cannot constrain a remote provider to maintain pSLS for its own needs.

QoS, which is experienced by the end-user, can be variable. In fact, the inter-domain route selection can change at any time for example when a pSLS is cancelled or when a route is no more accessible. In those cases, the path changes (if there is more than one) and the new end to end QoS performance characteristics values can be different from the previous ones.

Nevertheless, we can know, at any time, the QoS values associated with a route for a given destination if we add a reporting functionality to the q-BGP protocol. This mechanism would compute, in a hop-by-hop manner, the QoS attributes {D, J, L} attached to each AS path and advertises outside its domain(s). When a given AS (myAS) receives from one of its service peers the announcement: *Meta-QoS-Class* + (AS path) + (QoS value) + destination, and selects this path, it advertises: *Meta-QoS-Class* + {myAS,AS path} + (myQoS ⊗ QoS value) + destination.

Bandwidth guarantees cannot be supported since the final destinations are not known in advance.

7.3.1.8 Scalability

When q-BGP is employed the volume and rate messages exchanged can become much more important than with the current BGP (especially when several meta-QoS-class planes are activated). Several distinct routing and forwarding tables should be activated and maintained per router. This number will depend on the number of supported *Meta-QoS-Classes*.

ASBR routers will have to swap DSCP values according to binding rules driven by established pSLS. Shaping and policing will probably impact the router forwarding performances.

Deployment of the QoS Internets can be gradual and assumes a close cooperation of adjacent providers.

7.3.1.9 Deployment issues

A new QoS aware inter domain routing has to be specified, developed and validated. This extra features could be implemented using the current inter domain routing protocol particularly BGP.

IGPs will compute routes based on the destination prefix information AND *Meta-QoS-Class* (l-QC for *Light approach*) (probably QoS routing planes identified by a couple {QC, MC}).

In order to implemented this extra features (e.g. QoS aware routing), routers will have to be updated. Therefore, introduction of such new services might be risky and slow due to its impact on existing infrastructure.

7.3.1.10 Requirements on pSLSs

Within this solution option, a pSLS should be considered as a permission to send some amount of traffic, towards any destination, within the context of a given *Meta-QoS-Class*.

Before establishing any pSLS, an INP shall qualify its l-QC in verifying their compliance with *Meta-QoS-Classes*. Only l-QCs for which a *Meta-QoS-Class* membership has been stated, are eligible to be extended across the Internet. This is necessary to ensure the service consistency requirement.

pSLS will likely be negotiated with some contractual maximum bandwidth per *Meta-QoS-Class* (l-QC binding in the case of the *Light approach*). Consequently, the upstream AS should make sure it doesn't send more data than it is allowed to. The downstream AS must police the incoming traffic so that it fits in the contracted traffic envelope.

The routers automatically choose the path. pSLS invocation and contractual bandwidth consumption will be hard to achieve.

Copyright © MESCAL Consortium, June 2005

7.3.1.11 Implications for cSLSs

Within this solution option, a cSLS should be considered as permission to send some maximum agreed quantity of traffic, towards any destination, within the context of a given *Meta-QoS-Class*.

Network accessibility through a *Meta-QoS-Class* plane is never permanently ensured.

The implicit versatility of QoS value shall be indicated. Informational values can be provided by the reporting functionality added to q-BGP. These values can't be contractual.

cSLSs don't need to explicitly state in advance the destination points.

The result is a best-effort QoS service. Normally clients should get the level of quality they need. But, we can't guarantee there will be no disruption or big fluctuation in the QoS they receive.

7.3.1.12 On demand inter-domain pSLS interactions

As described above, this approach allows a set of MC routing planes to be built dynamically; QoS information is exchanged within each plane for route computation purposes, with the final objective of selecting optimal QoS paths that meet average customer application needs.

Thus, if a remote domain does not support an appropriate pSLS that extends a given *Meta-QoS-Class,* it may imply, from a local domain perspective, the introduction of possible holes in the address space within the corresponding *Meta-QoS-Class* plane.

In order to solve this issue, one of the potential solutions is to make use of an “On Demand” pSLS feature to request the establishment of the missing *Meta-QoS-Class* extension class near the domains where these "QoS holes" exist.

The reasons why a remote domain may have no pSLS established for extending a Meta-QoS-Class plane are mainly of 2 categories:

- The remote domain cannot do it: because no DiffServ architecture has been deployed in its domain or extended MESCAL protocols and mechanisms are not available in its domain. Nothing can be done in that case. This domain can only be reached or crossed on a best effort basis.
- The remote domain doesn't want to do it because he hasn't identified yet any valid business reason for doing it.

In this latter case, it is suggested that the provider anyway proceeds to a QC-mapping and QC-binding operation and activates, at its domain's boundaries, specific q-BGP functions (to be specified) allowing to advertise lifeless o-QC (lo-QC) he would be ready to implement (QC-implementation) if some interest was shown by external providers In turns, these lo-QCs could be used by external domains and propagated using q-BGP. From a single domain standpoint, q-BGP could announce:

- Either a lifeless QoS reachability for a given destination within a Meta-QoS-Class plane with the corresponding lo-QC
- Or an e-QC and a possible lo-QC when this lo-QC would have been selected by q-BGP if this lo-QC hadn't been a virtual one.

Thus, thanks to this mechanism, external domains can become aware of possible capabilities of a remote domain and can now identify this domain quickly so that an On Demand pSLS can be requested easily and a negotiation cycle started.

7.3.1.13 Applicability to the Business Model

The business target covered by this approach is the residential market. It is suitable for service providers who are willing to benefit from network-wide differentiated services for improving their existing services or as a leverage to create new ones. This can be the case of web-based services (e-learning, e-training, consultation services…) or video-on-demand for instance for which some categories of end-users are ready to pay to get better services. The approach does not constrain

Copyright © MESCAL Consortium, June 2005

customers to specify the final destination of the traffic in the cSLS (or pSLS between providers). The address space, which can be reached within a Meta-QoS-Class (or l-QC) plane, depends on the number of established pSLSs between providers. All Internet users would consequently not be able to request such services until it is globally deployed.

The basic approach is resilient, scalable and respects the underlying philosophy, which guided the elaboration of the Internet. But the QoS guarantees it provides are loose since:

- QoS performance associated with an e-QC can change at any time since the Inter-domain path can change.

It is impossible to provide end-to-end bandwidth guarantees. The traffic matrix can be very stochastic (destination addresses and routes followed) and network engineering can only be achieved on a statistical basis.

7.3.2 Statistical Guarantees Solution Option

This section presents how to build the required capabilities in order to be able to support end-to-end QoS classes (QCs), and it focuses on the required inter-AS interactions. These classes can be used to offer end-to-end services with some statistical guarantees.

7.3.2.1 Introduction

Each domain is engineered to support some Quality of Service classes, also known as Per Domain Behaviours (PDBs) [Nichols01].

The engineering of QoS classes includes the provisioning of network resources in terms of routing and bandwidth management (including scheduling and buffer resources) for implementing the required Per-Hop-Behaviours (PHBs). This provisioning can be done either by an automaton (e.g. [Trimin01]) which defines the appropriate provisioning directives and enforces them to the network elements, or through human static configuration. Even in the latter case there may be tools, which aid the human administrators to take the provisioning decisions (e.g. [Feldm00]). We have to mention that in this engineering for provisioning process, we include the over-provisioning engineering model. In this solution option the desired behaviour of some class is based on allocating link bandwidth, which is well above the maximum average requirements for that class (common practice is to keep it the utilisation below 50%). In the latter engineering model still some basic differentiation between classes is assumed to exist, but the over-provisioning factor between the classes may vary according to the significance of the class (e.g. a premium class may be over-provisioned to always below 10% utilisation).

Note that this solution option does not take into account the access network QoS capabilities in the forwarding path. These capabilities can be incorporated into this solution option either if the first hop ISP takes into account the QoS capabilities of the customer's access network, or the access network itself plays the role of an AS, as this role is defined by this solution option.

The timescales in which these engineered classes are realised and possibly changed, are at the level of a Resource Provisioning Cycle [Trimin03], which is from few hours to the level of weeks, depending on the operating procedures of the providers. This is the medium-to-long timescale traffic engineering as defined by the IETF [Awduc02]. Normally these classes are not expected to change considerably from one provisioning cycle to another because the provider will have agreements based on these classes which impose some restrictions on the supported classes. A provider will always try to enforce these classes by setting them as the engineering target QoS classes (see below for more details).

QoS classes are differentiated within an AS by using a different DSCP (Differentiated Services Code Point) value in the appropriate octet (Type Of Service TOS → IPv4, Traffic Class → IPv6) of the IP header. This DSCP marking is then used to classify the packets into (ordered) traffic aggregates which are processed (buffered and forwarded, typically) according to different PHBs, depending on the class.

The solution option described in this section makes use of a concept the Virtual QC, in addition to the QC concepts presented in section 4.2.2.

Copyright © MESCAL Consortium, June 2005

Virtual QC (v-QC): this is a virtually introduced engineered QoS class. Within an AS, the differentiation of packets into l-QCs is implemented using a different DSCP for each l-QC, which then maps onto the one PHB. This means there exists a "1-1" mapping between a DSCP an l-QC and a PHB. If we relax this "1-1" mapping, and allow for "N-1" mappings, i.e. *n* DSCPs mapped to the same PHB, it would be as if *n*-1 additional l-QCs were introduced. We call these additional l-QCs, virtual QCs (v-QCs). Note that the mechanism to support v-QCs already exists since the DiffServ standard supports this "N-1" mapping from DSCPs to PHBs. The need for introducing these v-QCs, the rules for their introduction, and their use in this solution option is going be discussed in the following sections (see section 7.3.2.4). Because a v-QC is at the same level as a l-QC, in the rest of this document we may use the term l-QC for both of them and will differentiate only when necessary.

7.3.2.2 The Cascaded Solution for Statistical Guarantees

The essence of the MESCAL solution to service option 2, i.e. offering services with some statistical guarantees, can be summarised as follows:

- The end-to-end QCs are built based on the cascaded model, i.e. by service peering between adjacent ONLY domains.
- It supports statistical end-to-end guarantees both in terms of QoS parameters and in terms of bandwidth.
- The solution requires the pSLS to valid for specific address prefixes. A pSLSs includes the required QoS class, bandwidth both with some probabilistic guarantee, for some specific destinations. The service peer AS that signs to a pSLS, undertakes the responsibility to adhere to all the agreed requirements, within the error margin given by the probabilistic guarantees terms.
- An AS that wishes to offer a particular o-QC to a destination prefix, is allowed to use MORE THAN ONE e-QCs, i.e. many internal l-QCs and many external o-QCs, as long as the offered o-QC constraints are met.
- Mapping and binding are allowed on an N-M basis. This means that, in order to build a given o-QC which satisfies business objectives, the solution option allows the mapping process to produce a set of e-QCs formed with different l-QCs and different external o-QCs. Constraints on both can be imposed by the business objectives. These objectives MAY (not necessarily though) facilitate the Meta-QoS-Class concept.
- The total number of offered o-QCs, both e-QCs and l-QCs, is constraint to be NO MORE THAN 64, since the Differentiated Service Code Point (DSCP) is the means to indicate both internally and externally one QC in an IPv4 realm. This controls the scalability of the solution unless IPv6 was introduced in the network.
- The QC splitting problem is tackled with the v-QC, an engineering approach that facilitates the fact that we can configure the routers so that several different DSCP markings refer to the same Per-Hop Behaviour.
- Inter-domain routing is pSLS constrained, i.e. the established pSLSs influence the routing. Inter-domain routing is also QoS-enabled, i.e. it is able to compute different paths for different QCs. There is no other mandatory requirement from routing.

7.3.2.3 QC Advertisement

QC advertisement is not mandatory for this solution option, but QoS-related information needs to be propagated throughout the (peering) domains. This means that this solution option can use the advertised o-QC information of adjacent AS but it is not a requirement to have such advertisements that is we can have a fully operational solution even without advertisements.

In the following sections it will be clear that QC advertisements are only necessary when we want to make a mapping between the QCs so as to find the ones that are compatible and then based on some

Copyright © MESCAL Consortium, June 2005

logic request a pSLS. In the case we do not have explicit advertisements, the logic which decides how to request a pSLS will base the decision only on the requirements set locally and then the pSLS receiver will do a "best match" counter proposal during the pSLS negotiations.

7.3.2.4 QC mapping

We have QC mappings at two levels. The first level is mapping within the AS, between the local l-QCs or e-QCs and the o-QC. The second mapping is an external mapping between the l-QCs and/or e-QCs of one AS with o-QCs of the adjacent ASs. In this solution option in the case where we do not have any QC advertisement or discovery the QC mapping step is omitted.

Both mappings are dictated by the fact that an AS wants to extend its own local l-QCs to prefixes that can be reached by traversing other ASs. In the example shown in Figure 38, AS1 wants to extend the offering of QCs to addresses located outside of the domain, in this example to addresses located in AS2. Somehow AS1 needs to communicate (see section on pSLSs) the requirements it has in terms QoS. The mappings described in this section can be either the result of an agreement between the two adjacent AS, in the case of AS1 does not know the o-QCs of AS2, or could be done before any agreement based on the information that AS1 has about the o-QCs of AS2. In this example we are showing the mapping being done with the o-QCs of AS2. These o-QCs maybe composed of many e-QCs or l-QCs, in both cases there will be an l-QC applied internally to AS 2. The latter are the l-QCs shown in the figure and thus they represent either a single used l-QC or the first part of an e-QC.

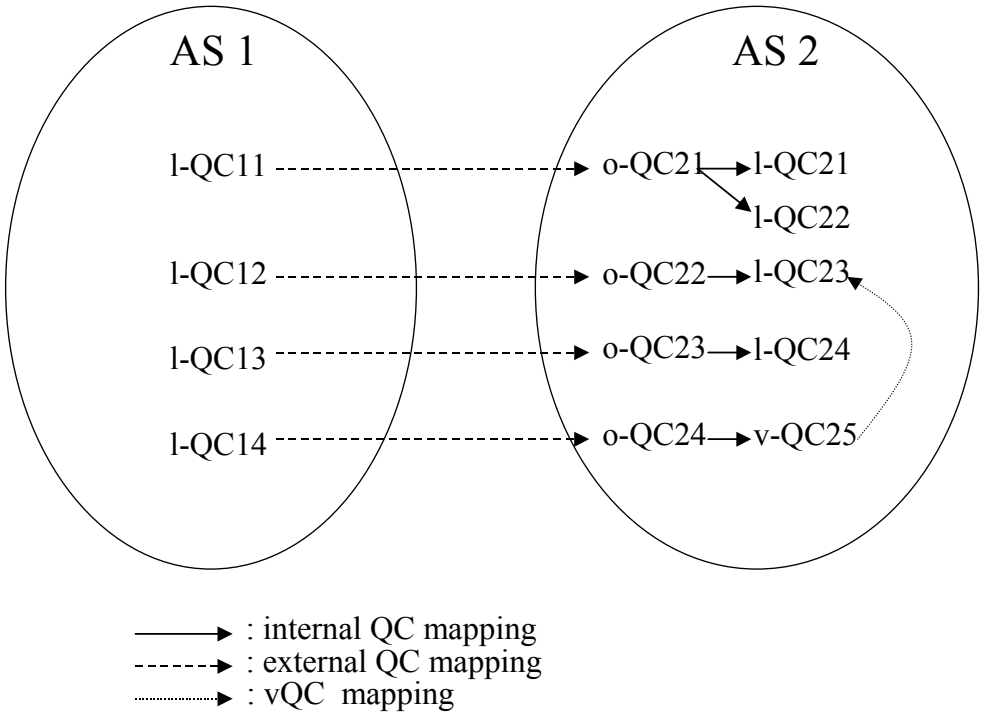


Figure 38: QC Mapping example

The internal mapping between the o-QCs and the QCs of a given AS is a "1-N" mapping. This means that the ISP has the freedom to provision any number (less than 64 in an IPv4 scenario, since they have to be uniquely identified by a DSCP) of l-QCs but only offer some of them to the peer ASs. For example o-QC21 in AS2 is mapped to both l-QC21 and l-QC22. The rules for such mapping is that all the l-QCs mapped to the same o-QC must be "compatible" with each other and the o-QC, and in addition the following must be hold:

$$\forall QC_i \rightarrow o\text{-}QC, \quad o\text{-}QC \geq QC_i \qquad (1)$$

i.e. each l-QC used to support an o-QC it must be *at least as good as* the o-QC it is mapped to. The reason for having additional l-QCs used by same o-QC is for reasons like load sharing or offering a

Copyright © MESCAL Consortium, June 2005

much better QC to internal customer VPNs. It is not compulsory for every l-QC to be mapped to an o-QC. When more than one l-QC is used, then there must be some static or dynamic load balancing of the o-QC traffic to the various supporting l-QCs. Note that in the example shown in the figure we do not show the internal mapping of l-QCs to o-QCs within the AS1.

The AS, e.g. AS1, that requires the extension of its own l-QCs to the addresses supported by the peer AS, e.g. AS2, may request to map an o-QC for which the receiving AS2 does not have any advertise an offered QC. For example l-QC14 does not have a compatible l-QC within AS2. In this case AS2 may refuse this mapping and this will create a "hole" in the end-to-end QoS, and therefore AS1 will only be able to support this class for its own addresses. On the other hand, AS2 may want to offer a mapping to AS1 for one of AS2 l-QCs, e.g. l-QC23, which is *at least as good as* the l-QC14.

In the latter case, we may have the splitting problem, see section 5.5.1. This problem will only occur when the traffic is going to exit AS2 towards another AS, and in this case AS2 will not be in a position to know which part of the l-QC23 aggregate was from l-QC12 and which from l-QC14. One solution to the splitting problem is to allow only merging of l-QCs and never splitting. In many cases this solution may not be acceptable, since the end-to-end classes, which were merged at some point, will tend to be the same as the path includes more ASs.

We propose a more general solution to the splitting problem by introducing the notion of a virtual QC (v-QC). For example when AS2 receives a request for mapping the l-QC14 from AS1, and realises that the closest local QC support this request is l-QC23, for which there already exists a mapping to another offered QC, i.e. o-QC22, it will introduce a new o-QC, o-QC24, which is identified by a unique DSCP. The role of this new DSCP is only to differentiate between the two o-QCs since the corresponding PHB received by the packets of both classes will be the same, i.e. that of QC23, but the two classes will be distinct at every egress point of AS2.

The above describes unidirectional mappings from AS1 to AS2. Similarly, AS2 will request information from its peer AS1 to extend its own l-QCs to the addresses supported by AS1, thus providing unidirectional mappings from AS2 to AS1. The approach for mapping will be similar to the one described above. At the end of the day we will have mappings for both directions.

7.3.2.4.1 Mapping with Meta-QoS-Classes

An observant reader may have noticed that with the mapping procedures discussed so far, if everybody accepts their peers requests for all o-QCs mappings by introducing v-QCs, we will end up with a large number of required l-QCs within each AS. At the steady state of the overall mapping procedure, the number of o-QCs within each and every AS, will be the same to that of the AS which has the maximum requirements in o-QCs. This number has to be bounded by 64, since the DSCP value must uniquely identify each o-QC within an AS, it may still be big enough to introduce complexity and scalability concerns in negotiations, provisioning, and routing functions.

In order to further reduce the total number of o-QCs and at the same time adhere to some globally well-known and defined classes, we make use the notion of the Meta-QoS-Class, see section 7.3.1.1.

Copyright © MESCAL Consortium, June 2005

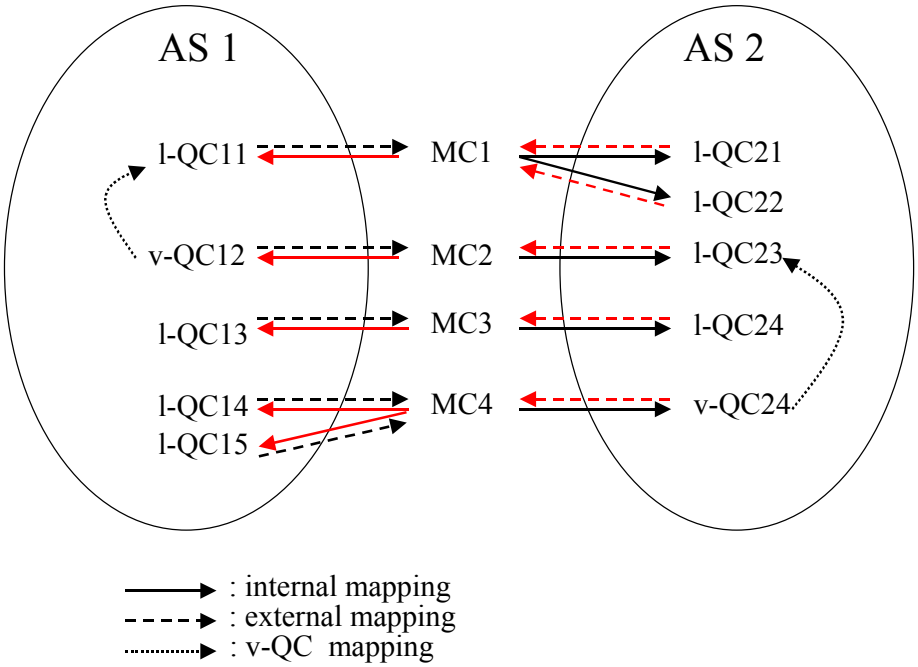


Figure 39: Mapping example with Meta-QoS-Classes

In Figure 39 we show the mapping example presented in the previous section with the use of MCs. The figure shows the mappings from both sides, with blue corresponding to the mapping agreement for traffic from AS1 to AS2, and the red for the mapping agreement in the opposite direction. Note that in this example since we are showing the mappings in both directions we can observe the QCs within AS1, which constituted the QCs of the example shown in Figure 39. We can observe that, as with the l-QC21 and l-QC22, the l-QC14 and l-QC15 are mapped internally to the same MC. The actual mapping of incoming traffic to l-QC14 or l-QC15 can either be done statically or dynamically with load balancing between the different l-QCs, which can be relied on load sharing criteria and implemented at the with a hashing function criteria.

7.3.2.5 QC binding

QC binding is the application of the bind operator "$\oplus$" between QCs, in order to define an e-QC. The ultimate target is to have at each ASi a precise definition of the e-QCs that are available. That e-QC can then be offered, i.e. become an o-QC, to the other upstream ASs. In general an o-QC can be either an l-QC or an e-QC.

The binding between the QCs is done in a *cascaded* fashion. This binding is the recursive definition of e-QCs at ASi, as follows:

$$e\text{-}QC^0 = l\text{-}QC^0 \tag{2}$$

$$e\text{-}QC^i = l\text{-}QC^i \oplus o\text{-}QC^{i-1} \tag{3}$$

That is an $e\text{-}QC^0$ at the home AS of the address prefix is defined to be a local $l\text{-}QC^0$ of that AS. And then recursively define the $e\text{-}QC^i$ of an ASi is the binding result of a local $l\text{-}QC^i$ of that AS, and an offered $o\text{-}QC^{i-1}$ of the previous ASi-1. This cascaded definition of QCs is the main characteristic this solution option.

According to the definition for e-QC as given above, if we bind different l-QCs internally with the same external o-QC then the resulting e-QCs will be different, similarly if we bind the same l-QC internally with the different external o-QCs the resulting e-QCs will be different. This solution option

Copyright © MESCAL Consortium, June 2005

does not restrict these bindings, and they are all allowed, thus it allows N-M bindings. Restrictions can only apply based on the policies of the domain.

When based on marketing and business objectives, the service planning functionalities an AS*i* decides to offer an o-QC towards some destination, this o-QC will have specific characteristics. It may be the case that more than one e-QC are able to comply with the requirements of the specified o-QC. So the $o\text{-}QC^i$ can utilise all the $e\text{-}QC^i_k$ which are at least as good as it:

$$o\text{-}QC^i \rightarrow e\text{-}QC^i_j \tag{4}$$

such that

$$e\text{-}QC^i_j \leq o\text{-}QC^i, \quad \forall j \tag{5}$$

The actual offering of the o-QC happens when this is included in pSLS, i.e. when AS*i* becomes the downstream AS for an AS*i*+1. In this case AS*i* has to make some selection about which bindings in effect, that is to choose which of the compliant $e\text{-}QC^i_j$ will be used for offering that o-QC. In the simplest case a single $e\text{-}QC^i_k$ can be the chosen one. In a more complex scenario there may be some policy to have more than one in effect, so as to allow for some dynamic load balancing between the locally used l-QCs and the agreed with the downstream o-QCs, as those are bound in the definition of each of the e-QCs.

If the exists such a load sharing functionality as discussed above it will have to take into account the utilisation of the various $l\text{-}QC^i_j$ s and $o\text{-}QC^{i-1}_j$ s bound as in **Error! Reference source not found.** to the $e\text{-}QC^i_j$ s which belong to the subset of the e-QCs which are compliant to $o\text{-}QC^i$. The implementation of the load balancing decision, i.e. splitting ratios and mapping of traffic could be done in two ways. Either at the forwarding level based on some hashing function on the fields of the IP header, or at a higher level based on assignment of SLSs to each of $e\text{-}QC^i_j$ s bindings. In any case this load balancing could be considered in combination with the load sharing options discussed in section 5.4.2.

Summarising, the QC binding operation includes the following sub operations:

- Out of all possible QC mappings we need to select the ones for which will establish pSLSs
- When we are to offer an o-QC and accept a pSLS from an upstream AS, the pSLSs with our downstream ASs must be in place. At this point we need to decide out of these pSLSs which are the ones that will be used for offering the particular e-QC.

And finally, the actual QC values used with binding operator "$\oplus$" are the ones decided by any dynamic load balancing algorithm.

7.3.2.6 QC Implementation

The basic assumption of this solution option was that, within an AS, the packets belonging to a QC are uniquely identified by the DSCP value marking in the IPv4 TOS, or IPv6 Traffic Class fields. Packets of the same QC have the same DSCP marking.

Between ASs, this solution option proposes to use the same field, i.e. the DSCP, to signal the o-QC mapping. The exact DSCP values that will be used to signal the o-QC requirements between the ASs are defined between the ASs during the agreement request-negotiation process.

Copyright © MESCAL Consortium, June 2005

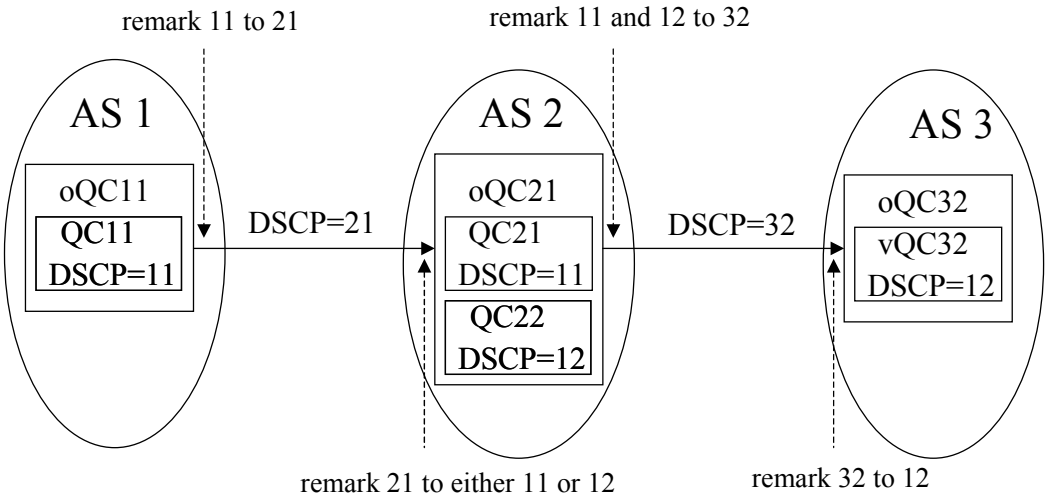


Figure 40: QC implementation example

In Figure 40 we show a QC implementation example between 3 ASs. AS1 internally maps the o-QC11 to QC11, which is identified by the DSCP value 11, AS2 maps internally QC21 and QC22 to o-QC21 and uses DSCP values 11 and 12 respectively. The external mapping between o-QC11 and o-QC21 is signalled with DSCP value 21, that is when traffic marked with 11 leaves AS1 it is remarked to 21, and when traffic marked with 21 enters AS2 it is remarked, either statically or dynamically (for load balancing) to 11 or 12. Similarly, the traffic which leaves AS2 and is marked 11 or 12 is remarked to 32 in order to obey the external mapping of o-QC21 to o-QC32.

7.3.2.7 Requirements on pSLSs

Thus far we have assumed that one AS is able to use a peer AS's offered QC. This ability is defined in a pSLS (peer-SLS). The purpose of this section is to identify at a high level the required fields in a pSLS for the solution option in consideration. The details of the pSLS structure will be subject of further research within the MESCAL project.

This solution option requires building QoS agreements only with direct peering ASs. Thus the pSLSs will be requests for agreements with the management/administrative entities of the peer ASs. The number of offered o-QCs that an AS is supporting is defined by the Marketing and Business objectives of the ISP, and, in this solution option, is constrained by the total number of DSCPs, i.e. 64.

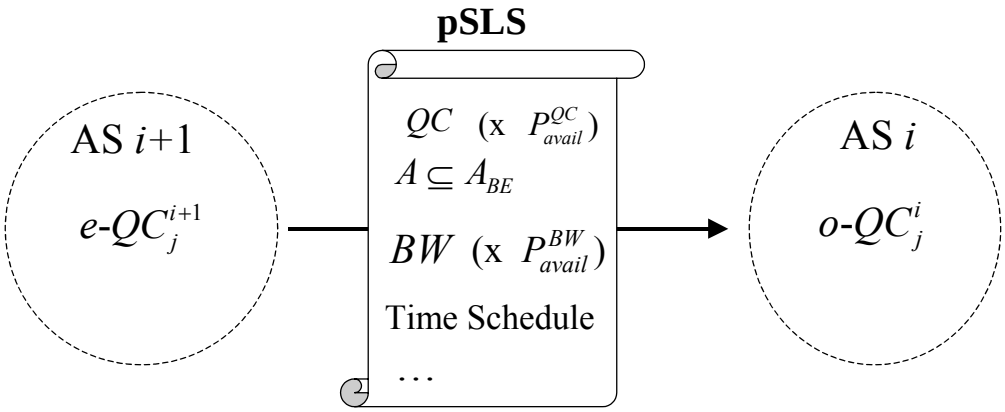


Figure 41 Abstract required fields in a pSLS

With reference to Figure 41, we assume that AS i+1 wants to create an $e\text{-}QC_j^{i+1}$, by extending its $l\text{-}QC_j^{i+1}$, that spans to addresses other than the ones managed by itself. We distinguish between two cases:

Copyright © MESCAL Consortium, June 2005

1. There have been QC advertisements, and thus AS *i*+1 knows the o-QCs offered by AS *i*.
2. There have not been any QC advertisements, or there were "marketing-language" advertisements, e.g. "I offer a low delay QC", without specific values in the advertised QCs.

In both cases the decision for creating the $e\text{-}QC_j^{i+1}$ is driven by the Business/Marketing policies of AS *i*+1. The difference is that in the first case the actual *QC* value that will be *requested* in the pSLS is decided by choosing the most appropriate from the ones that have been offered and advertised by the adjacent AS, while in the second case, the request is arbitrary and the requested AS will find the one of its offered QCs which is the closest to the request.

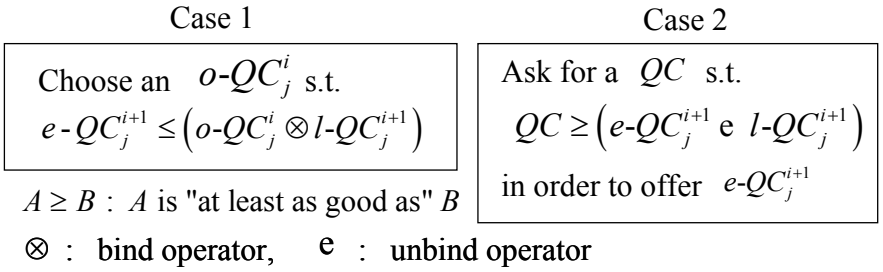


Figure 42: Two cases for requesting the QC in a pSLS

In Figure 42 we show in pseudo-code the difference between the two cases for requesting a pSLS mentioned above. In the first case the requester chooses to ask for the o-QC of AS *I* which bounds with the local QC that can produce an e-QC which is at least as good as the required e-QC. In the second case, it is free to request whatever it wants, which is the QC which is at least as good to the target e-QC unbound to the local QC. Note that in both cases after negotiations and based on the pSLS receiver AS's policies, the end result in the pSLS agreement is one of the offered QCs, e.g. $o\text{-}QC_j^i$, of AS *i*.

AS *i*+1 knows the list of address prefixes A_{BE} that it can reach via AS *i*, e.g. from the routing protocol information used to compute Best Effort (BE) routes, the pSLS should include a list of addresses prefixes $A \subseteq A_{BE}$, to which the requested e-QC should be available. AS *i* will know the answer to this question based on its own addresses and on the pSLS that has been established with its peers.

The bandwidth is another important factor that should exist in a pSLS. This request will be based on the predicted requirements of AS *i*+1 for that peering connection. This information will help AS *i* to provision its own domain. It is expected that bandwidth is going to be renegotiated more often than the other fields of the pSLS. The time schedule of the required QoS is also another important parameter in the pSLS.

We envisage that there will be a number of negotiations between the peers before an agreement is reached. The end agreement will have to include the DSCP value that will signal the class mapping. In addition the whole agreement, and/or individual fields, e.g. e-QC, BW, may have an availability factor.

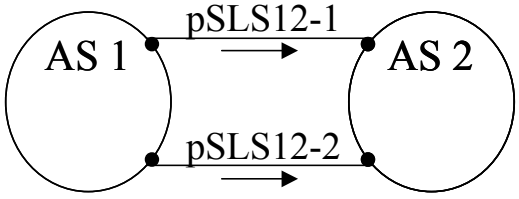


Figure 43: Peering at more than one point

Copyright © MESCAL Consortium, June 2005

If we have more than one peering points between two ASs we require to have a different pSLS for each of them. Even if we want to use the same o-QC from both peering points we need to have different pSLSs.

Note such a pSLS is the basic building block of agreements between two ASs. This means that this may be only part of the final agreement that needs to be negotiated, and may include more than one of the basic building block pSLSs. Thus the amount of required communication and negotiation can be significantly reduced, since we they will be done on higher level than the basic pSLS. For example in Figure 43 AS1 may choose to put pSLS12-1 and pSLS12-2 into the same agreement and request for a combination, and in this case AS2 will be in position to make alternative suggestions, especially as far as the total bandwidth is concerned.

Summarising, the following are the required fields in a basic pSLS:

- A QC, which corresponds to a peering offered QC, e.g. oQC_j^i (*Which quality?*)
    - An associated probability with which the QC is available, P_{avail}^{QC}
    - The DSCP which signals the agreed QC
- The address prefixes which are covered by the pSLS, $A \subseteq A_{BE}$ (*Where to?*)
- The bandwidth available for use BW (*Which quantity?*)
    - An associated probability with which this bandwidth is available, P_{avail}^{BW}
- The time schedule

7.3.2.8 Scalability

This solution option follows the cascaded model for building end-to-end QoS and constrains the maximum number of offered QCs to be no more than the maximum allowable DSCPs. In the following we will make a first attempt to estimate scalability imposed by the solution option in the QC management and routing decision process, as well as the routing dissemination process.

When this solution option is used in an IPv4 or IPv6 realm, it does not constrain the possible combinations between the QCs. But it allows in an IPv4 realm only 64 QCs to be offered per AS. This means that in the worst case the possible combinations for offering a single QC is 64xN where N is the number of peer ASs, and thus 64^2xN for offering all the possible QCs. This number will has to be multiplied for each AS pair with number, K, of peering points between that pair of ASs. So, the scalability factor for supporting and offering the maximum number of QCs is in the worst case 64^2xNxK, assuming that K is the maximum number of different peering points between two ASs.

The number given above is the scalability factor for the inter-intra domain routing decision processes (including load balancing) as well as the QC management ones. This means that in the worst case the routing information that is handled today will have to be multiplied by scalability factor 64^2xNxK. This is the worst case that assumes that all peer ASs offer the maximum (64) number of QCs and that we have pSLSs with all our peers ASs in order to use all their offered QCs. Also it assumes that all ASs peer with the AS in question at the maximum number of points, K. For the routing information dissemination process (section) the scaling factor is 64.

We can see that the solution option, apart from the constant factor, scales with the number of peer ASs and the number of peering points for each peer AS. Since the larger, e.g. tier-1, tier-2, the ISPs the more the peering points and thus smaller ISPs can handle the burden for supporting the QoS. We can also observe that the scale factors are only first-degree polynomial to the number of peering points and peers ASs, thus avoiding an exponential growth.

Note that this scalability assessment is at an abstract level, and does not include any considerations about the IPv6 realm case. After the exact algorithms have defined will be in position to make a more detailed assessment of the scalability factors.

Copyright © MESCAL Consortium, June 2005

7.3.3 Hard Guarantees Solution Option

The level of QoS guarantees reached with the above solution options is not satisfactory for all corporate business services for which strong guarantees must be provided.

In particular, such categories of end-users would request:

- Guaranteed QoS performance
- Bandwidth reservation

In order to satisfy these requirements it is necessary to elaborate a solution, which enhances the MESCAL solution for loose and statistical guarantees service options, in two ways.

- The first action is to fix the inter-domain path so that the QoS performance of an e-QC cannot change.
- The second step is to provision the requested bandwidth all along the path. This reservation must be achieved, in a coordinated manner, within all crossed domains and at the boundaries of these domains.

Those two constraints imply that the final destination of the traffic is known in advance to the providers and will become mandatory information for all cSLS established within this context.

Within the previous sections we defined a "QoS enabled shared IP network". Thus, in the same way additional functionalities were built on top of best effort network, we propose to extend the basic QoS approach we defined in section 7.3.1 with additional functionalities. Inter-domain MPLS TE is a good candidate since it entails most of the features we need and it has strong business support.

MESCAL introduces QoS consideration in the existing MPLS TE approach and embeds its IP based approach in such a way the QoS MPLS TE solution can greatly benefit from the underlying infrastructure to make easier the computation and the establishment of QoS constrained LSPs.

7.3.3.1 QoS and LSP considerations

In a best-effort environment, the establishment of a best-effort LSP between two extremities (identified by their respective IP address) is only constrained by the existence of an inter-domain path (learned via IGP/BGP) providing that network resources are available.

Within an intra-domain QoS context, each LSR is configured to support the PHB corresponding to the l-QCs defined by the provider for its domain. If we ignore resource availability observations (bandwidth for instance) each datagram conveyed within an LSP will be handled according to the l-QC it requests, whatever the path it follows. In particular, this means that a given LSP could be multi-coloured, that is it could be used to transport datagrams requesting different l-QC from one point to another.

In the inter-domain QoS context, it is not possible anymore to assume that a single LSP will cross a set a domain in which compatible l-QC will have been defined. All l-QC defined by the provider requesting the LSP may possibly have been extended up to the target destination termination of the LSP but the path to follow can be different.

Some provider's customer will likely ask for multi-coloured LSPs (or EXP-Inferred-PSC LSPs as defined in RFC3270). Sometimes it will be possible to aggregate all requested l-QC traffic in the same LSP, sometimes not. Several LSPs will have to be used. Thus, a service offering will have to take into account a possible multiplication of the number of LSPs to establish, in order to satisfy customers' requirements. The solution will consequently have to support the ability to compute multi-coloured path for an LSP with some options allowing returning either one multi-coloured path or several mono/multi-coloured paths as the result of a QoS path computation.

Copyright © MESCAL Consortium, June 2005

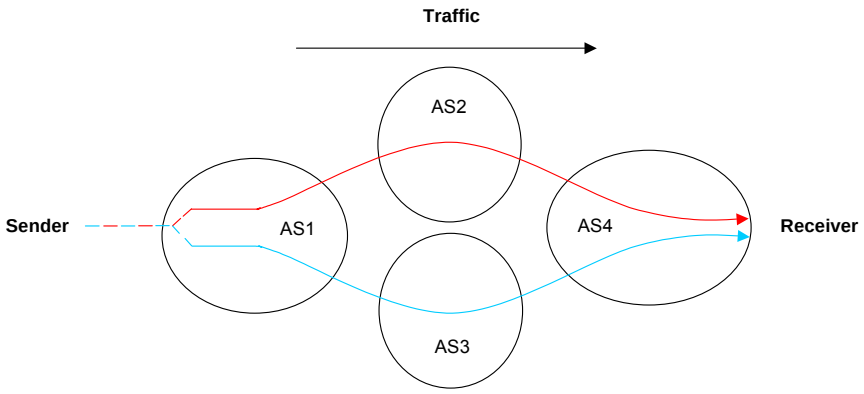


Figure 44: Multi mono-coloured LSP

At the service access point, in a situation where several mono-coloured would have been created, (first LSR), the injection of the traffic in the correct LSP would have to take into account the destination address of the datagram **and** the requested l-QC.

In the above figure, traffic splitting occurs at the boundary of the first domain but it could be imagined that this splitting is achieved at an inter-domain boundary, but this case is for further studies.

7.3.3.2 Working overview

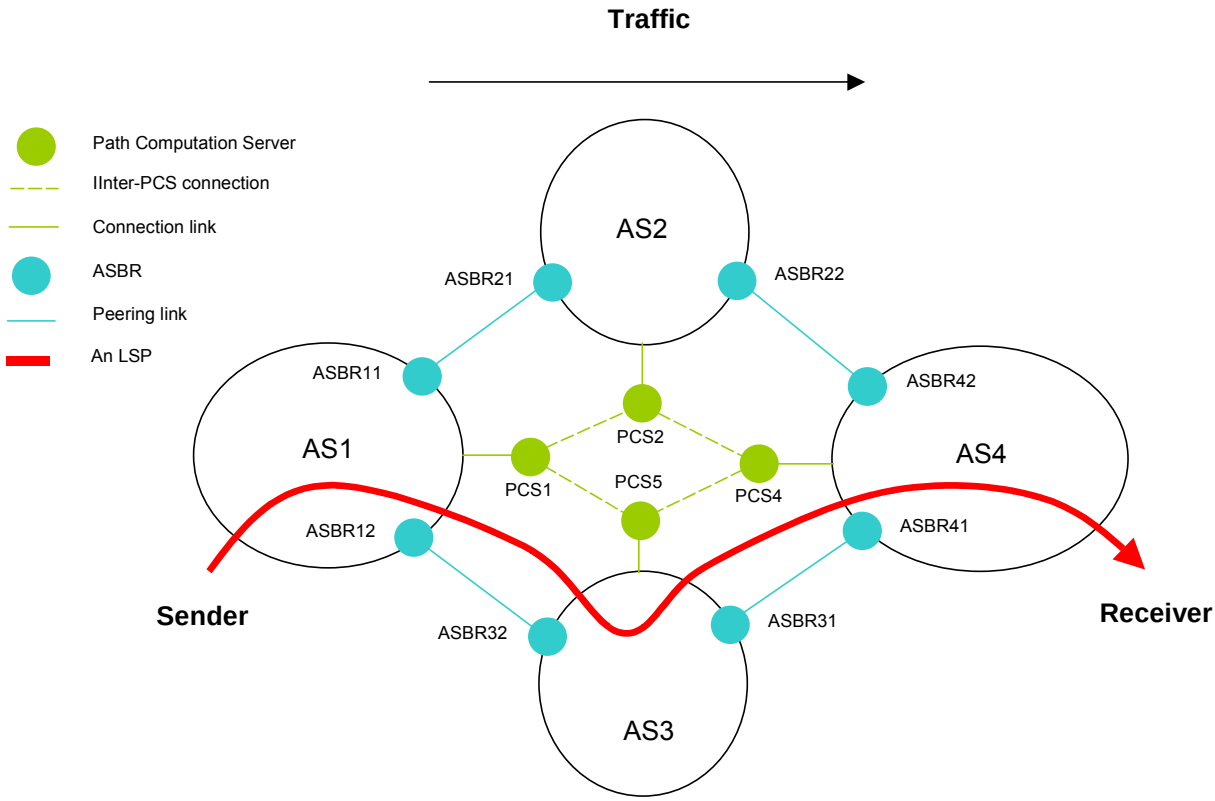


Figure 45: Working overview

Each domain is assumed to have some l-QC deployed. q-BGP is running between domains which have agreed to established a pSLS. Each domain receives, per Meta-QoS-Class plane, the set of destinations that can be reached within each *Meta-QoS-Class* plane it supports, together with some aggregated QoS performance information. A full q-iBGP mesh between all ASBRs of a domain has been set-up so that destination learnt by a peering ASBR can be propagated to the other ASBR of the same domain. QoS

Copyright © MESCAL Consortium, June 2005

routes learned by q-BGP are made known of the q-IGP in place in each domain so that a datagram can be routed up to the correct egress point within a *Meta-QoS-Class* plane.

A PCS (Path Computation Server) is present in each domain and receives q-BGP announcements from all ASBR of the domain. Thus, the PCS can know all destinations that can be reached within a *Meta-QoS-Class* plane together with their associated QoS performance characteristics. Moreover, each PCS establishes a session with the neighbours PCS of the external peering domains for which pSLS have been contracted. Communications between PCS occur within the best-effort Meta-QoS-Class plane thanks to the activation of Inter PCE Communication Protocol (PCP).

For creating an inter-domain QoS constrained LSP, the domain which requests the establishment of the LSP asks its local PCS to compute an inter-domain path satisfying a set of QoS constraints. This first PCS selects one possible path among the set of possible alternatives and identifies the next-hop domain. It then verifies that appropriate resources are available in its own domain and set-up administrative pre-reservation in the management system of the domain. Then it contacts the next hop PCS in the external domain, requesting a path computation between its peering ASBR and the termination address of the inter-domain LSP. This second PCS performs the same computation as the first one did and the procedure is iteratively repeated up to the last PCS. If a path satisfying all requirements is found, each PCS returns the QoS path to follow as a list of LSR. Each intra-domain sub-path is concatenated with the result received and when the last result reaches the originating PCS the whole path is available. A PCS can try several alternatives before sending back any path error computation. If a PCS in the AS path returns an error the path cannot be computed and therefore the LSP cannot be established. Otherwise, an RSVP TE LSP paths set up message is sent by the head end with the computed path.

When the RSVP TE RESV message is returned, some outsourcing admission control should be done at each inter-domain boundary in conjunction with information stored by PCS in the management system, for security, provisioning and accounting purposes.

7.3.3.3 QoS path computation

As briefly described above, path computation is distributed between a set of cascaded PCS. At a high and preliminary description level PCS communicate together thanks to the activation of PCP:

7.3.3.3.1 Finding an egress point

Since it receives q-BGP routes or at least it accesses to ASBR RIBs , each PCS knows all the available QoS routes for reaching a particular destination within a *Meta-QoS-Class* plane. The LSP is not constrained to follow the path selected by q-BGP and can also follow an alternative QoS path. For selecting a path, the PCS can rely on the number of domain hops and/or on the QoS performances of each corresponding e-QC, or any other administrative local policy enforced.

Comparing e-QC is not an easy task. It is suggested that this comparison is achieved using the definition of the *Meta-QoS-Class* itself, which is supposed to particularly optimise one of the performance parameters of a QoS-Class (if a given *Meta-QoS-Class* has been defined for delay sensitive kind of application it can try to optimise its researches using this performance parameter). Thus, if can classify the learned paths according to this specific performance parameter and choose, from this perspective, the best egress point.

If the requested LSP is multi-coloured, it must select a path supporting all the requested *Meta-QoS-Classes*. If it finds more than one, the choice of the path can become much more tricky. When possible (e.g. when requested *Meta-QoS-Classes* belong to the same hierarchical branch), it is suggested to exploit the fact that *Meta-QoS-Classes* are hierarchically organised and to base the selection process on the highest *Meta-QoS-Class*. When Meta-QoS-Classes belong to different branches of the hierarchical tree there is no evident selection criterion. In such a case, it is suggested that the requesting PCC, indicates the priority order that will be used by PCS for searching a path.

Copyright © MESCAL Consortium, June 2005

If no path can be found the PCS must return an error. If a multi-coloured LSP was requested, one could imagine that the PCS could proceeds to an LSP splitting, providing it found different paths to reach the destination within the requested *Meta-QoS-Class* plane.

If an LSP (or one of the *Meta-QoS-Class* aggregated flows within the LSP) requests some bandwidth protection, the PCS can ignore the loss rate performance parameter of the corresponding *Meta-QoS-Class*. Indeed, considering bandwidth will be successfully provisioned for that LSP, no datagram loss will occur providing that the end-user respected the related service contact and doesn't sent more traffic for a given class that what was agreed in the cSLS.

7.3.3.4 QoS path establishment

When a PCS computes a possible inter-domain QoS path, it will closely interact with its inter-domain management system. Indeed, PCS interactions will not only find a QoS path but will also verify that necessary resources are available and can be reserved. In order to achieve this goal, each PCS should have an accurate view of availability of the requested bandwidth:

- Along the path followed to cross its domain (intra-domain).
- At the inter-domain boundary the QoS path is supposed to use.

A given PCS can have several pending queries in progress. Resources requested by those queries will very likely interfere and simultaneously ask for common resources along the same path. Consequently, PCS must take care of that and must register in their management system a PRE-RESERVATION-INTENT of the corresponding resources. When path computation requests are received, the state of each related resource should change to PRE-RESERVATION to indicate that the corresponding network resources will be engaged soon in an LSP set-up, or freed if a request is received. Since the effective realisation of the LSP is done via RSVP TE, there must be some very close interaction between PCS and RSVP TE mechanisms so that the distributed transaction can be monitored and error cases tracked (if the RSVP TE Path message is never sent for instance, PCS corresponding states should be freed). Each time an RSVP TE Resv crosses a domain boundary, some interaction between RSVP TE and PCS must occur (thanks to COPS for RSVP for instance) so that PCS can definitely register that resources have been effectively reserved and used.

The inter-working between the set of PCS and technical set-up mechanisms should be considered as a distributed transaction. LSP disassembling and breakdown should be considered in the same way.

This also means that each PCS will have to keep track of all individual inter-domain established LSP, which will be assigned a global and unique identifier.

7.3.3.5 Bandwidth reservation considerations

The basic MESCAL approach doesn't allow customers to request for bandwidth guarantees. But, as part of the QC-implementation process, the provider has, in some way, to allocate a given maximum bandwidth for each of the *Meta-QoS-Classes* it supports: in its own network but also at its boundaries when pSLS have been established.

Copyright © MESCAL Consortium, June 2005

Figure 46: Bandwidth Repartition per MC

The figure above illustrates this bandwidth repartition using 3 *Meta-QoS-Classes* within this example. This provisioning must be achieved on all links of the domain and at each inter-domain peering edge. At the peering edge, these maximum bandwidths are those, which have been agreed in the pSLS. Within the network their values are at the discretion of the provider and reflect an optimal balancing of business and traffic engineering objectives.

In addition, the QoS MPLS TE end-to-end extension, introduces additional engineering issues which are depicted below:

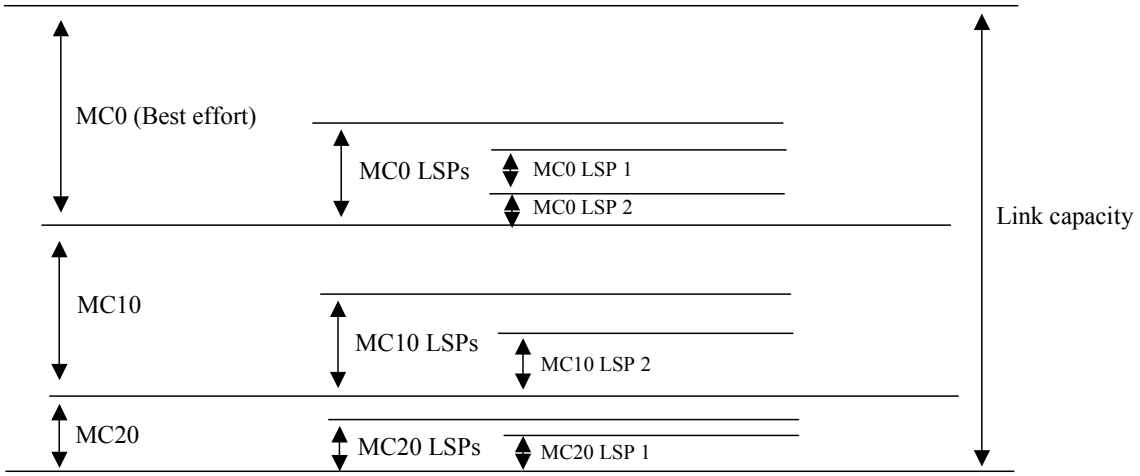


Figure 47: LSPs BW Reservation across multiple MCs

On a same link, for a given *Meta-QoS-Class*, both non-protected and protected traffics live together. These traffics can be IP traffic issued within the scope of the basic approach or within its extended scope (QoS MPLS TE). The traffic conditioning mechanism must be able to handle them in parallel in a consistent manner.

MCx LSPs, represents the maximum bandwidth which can be allocated to the MCx bandwidth protected traffic. Doing this prevents the non-protected traffic to fall into a starvation situation. MCx LSPs, represents also the amount of bandwidth that is always available to the MCx protected traffic. Depending on the level of control the provider has over its network, the maximum bandwidth that can be allocated to protected LSPs, should either be considered as an administrative maximum upper-bound or be enforced with appropriate mechanisms. Protected LSPs should be handled in such a way they never experience any datagram loss. In fact, if traffic policing is correctly achieved at the ingress point of the LSP, no loss should be observed for bandwidth-protected traffic. The remaining

Copyright © MESCAL Consortium, June 2005

bandwidth is used by the IP traffic. The minimum bandwidth, which can become available, is MCx bandwidth minus MCx LSPs bandwidth. When no LSPs are established, the non-protected IP traffic can potentially use all the MCx bandwidth.

7.3.3.6 QoS guarantees

With this end-to-end approach the guarantees provided are end-to-end QoS performances. In addition, if the LSP requests it, bandwidth guaranties can be provided.

7.3.3.7 Terms of cSLS

The cSLS agreed between the provider and the end-user will have to specify:

- The destination edge of the traffic
- The maximum bandwidth of each LSP (Or per MC)
- The *Meta-QoS-Classes* requested, together with their maximum bandwidth and, for each of them:
  - The guaranteed bandwidth, if any is requested, with must be smaller or equal than the maximum requested bandwidth for the *Meta-QoS-Class*.
- The nature of the LSP (mono or multi-coloured)

7.3.3.8 Terms of pSLS

In addition to QoS contractual terms stated in the basic section of pSLS, this end-to-end approach leads to introduce specific contractual parameters.

- an agreement between the parties allowing the requestor to dynamically establish inter-domain LSP
- a list of possible destination restrictions (probably handled by the PCS of each domain)
- a specification of the maximum bandwidth dedicated for each *Meta-QoS-Class* (including the best-effort one)
- a specification of the maximum bandwidth dedicated to bandwidth guaranteed LSP, within each *Meta-QoS-Class*.
- plus any other appropriate clauses such as the maximum number of LSP that can be requested, or the maximum bandwidth per LSP…

7.3.3.9 On demand inter-domain pSLS interactions

When computing a path, the PCS can fail for intra-domain and/or inter-domain reasons. Those failures, in normal operations, will be mainly due to the lack of resources. In such a situation, a path, which would have been the optimal path, would not be established. Identification of the domain where the path computation failed, together with the associated reasons, would be of a real added value for providers in order to improve the service they offer, thanks to an appropriate remote pSLS (re) negotiation request.

One way for achieving that is to rely on the Path Computation Protocol, which could be improved to return all the path alternatives which were tried but which failed. Doing so, the requesting provider would be aware of the reasons of the failure and possibly interact with the remote failing AS.

The remote AS, confronted to multiple requests, from external domain, could objectively consider a possible modification of some of its pSLSs based on objective business incitements.

7.3.3.10 IPv6 support

No major specific IPv6 issues were raised excepted MPLS TE support within an IPv6 environment.

Copyright © MESCAL Consortium, June 2005

7.3.3.11 Scalability

Clearly, if this solution option were deployed for all Internet users, it would not be scalable at all. But, this solution option has been designed to support the hard guarantees service option, which is mainly dedicated for mission critical applications, and so, targets corporate users and/or added-value services providers. Since the solution effectively reserves appropriate network resources across multiple-domains, providers pricing policies will be consequently adapted and will naturally regulate the usage of this service option. It will be deployed only when the interested future/potential customers will show clear demands. This is the reason why it is not expected that a large number of inter-domain LSPs be deployed, which would lead to non-scalable deployments in terms of number of LSP to be maintained and engineered. No full-mesh of LSPs is expected nor considered.

In each domain, the number of requests each PCS will have to answer will consequently be limited and thus, PCS systems should treat a predictable and a reasonable number of requests. Path computation is made easier thanks to the use of q-BGP, which advertises the QoS performance associated with each selected path. The number of inter-PCS queries might become important when the bandwidth criteria cannot be satisfied (note that this is classical client/server behaviour and no over computing is added), but this could be improved using some specific q-BGP extensions reporting the available bandwidth which can still be reserved toward a destination.

7.3.3.12 Applicability to Business Model

Thanks to the QoS MPLS TE extension, corporate business can be targeted:

- QoS performances become stable because the path followed by a given LSP is now fixed.
- Bandwidth guarantees can be offered, because it becomes possible for the provider to allocate appropriate resources (and no more) all along the path followed by the LSP.

In this option, pSLS becomes very strict and each crossed provider's domain must commit to respect all terms of the pSLS, since LSPs have an end-to-end meaning.

7.4 Interoperability of MESCAL service options

7.4.1 Introduction

Several MESCAL service options have been defined in the deliverable D1.1. Each service option [D1.1] provides different QoS-based services guarantees and is supported by a dedicated solution option. . Part of the technical means and protocols used to implement and to deploy each solution option can sometimes differ. In the contrary, some of these solution options can make use of common techniques and/or protocols, but their use can vary depending on the context of each individual solution option (for example a dynamic inter-domain protocol could be used for all solution options but the information to be carried by this routing protocol messages could differ). These differences can be sensitive and can become critical when deploying more than one solution option within the same AS (*in the rest of this Section, we will refer to this as the Co-existence Scenario*) or when extending the scope of a given solution option through an AS supporting a different solution option (*in the rest of this Section, we will refer to this as Inter-working Scenario*). Interoperability issues raised during this study together with potential solutions for solving them will very likely introduce new requirements that in turn will impact the solution option themselves including protocols and algorithms. Both service and technical considerations are taken into account when dealing with these co-existence and the inter-working scenarios.

This Section summarises the issues discussed within the chapter 4 of the D1.4 deliverable [D1.4] which develops those co-existence and inter-working scenarios in more details. Major problems raised during the study of these aforementioned scenarios have been highlighted. Only major issues are presented and restated here. No arguing is developed in this Section. A summary of the recommendations proposed in D1.4 is also provided.

Copyright © MESCAL Consortium, June 2005

7.4.2 Service considerations

The main purpose of these service considerations section is to qualify and classify the equivalent service option resulting from the interconnection of two ASs operating different service options, independently of any technical inter-working considerations. Only inter-working service scenarios providing an upstream service options with a better service (in terms of guarantees and not the QoS performance characteristics) are considered. This service evaluation phase will decrease the number of possible scenarios that need to be studied deeply from a technical angle.

This classification effort leads to the conclusion that only the following scenarios are valid from an inter-working service perspective:

- Extending the loose service option through the statistical service option,
- Extending the loose service option through the hard service option,
- Extending the statistical service option through the hard service option.

This service logic isn't strictly respected within the technical discussions. The motivation is to be able to address transit scenarios (*the transit scenario could be defined as follows: a given AS that enables a service option x could cross one or more ASs that offers different service option y in order to join remote service option x clouds*) and traffic bi-directionality issue.

7.4.3 Co-existence scenario

This scenario consists at examining the implications of the existence of several service options in the same autonomous system. For this purpose we examine the impact of the deployment of each service option on the network infrastructure and we qualify the compatibility of these functions between service options (For more details refer to [D1.4], Chapter 5). The basis of this comparison relies upon the technical description of the three solution options provided in D1.1.

The four discussed scenarios are:

- Co-existence of the loose and the statistical solution options in the same AS,
- Co-existence of the loose and the hard solution options in the same AS,
- Co-existence of the statistical and the hard solution options in the same AS,
- Co-existence of the all solution options within the same AS.

The discussion about the above scenarios focuses on the following:

- Main technical divergence issues between the considered solution options,
- A brief description of the problems raised,
- A list of recommendations to fix these problems,
- The adopted solution(s).

Copyright © MESCAL Consortium, June 2005

	Subjacent concepts in conflict	*Encountered problems*	*Recommendations*	*Adopted solution(s)*
Co-existence of the loose and the statistical solution options in the same AS	• Use of meta-QoS-class concept • Use of q-BGP • Bandwidth management • Contractual guarantees • Information contained in pSLS	• Differentiate the intra-domain path and the egress point per solution option • Usage of routes learned via q-BGP • Usage of common and shared network infrastructure for both solution options (Multiple Solution Option Management Problem, MSOMP)	• Use different ranges of DSCP values for the two solution options. • Build a management system able to handle simultaneously the two solution options on top of a common and shared network infrastructure. • The q-BGP process must have a means to separate announcements per solution option so that it can process each announcement according to the service option it belongs to.	• Use different ranges of DSCP values for the two solution options. • Build a management system able to handle simultaneously the two solution options on top of a common and shared network infrastructure. • The q-BGP process must have a means to separate announcements per solution option so that it can process each announcement according to the service option it belongs to.
Co-existence of the loose and the hard solution options	• Use of q-BGP	• Usage of routes learned via q-BGP for the two solution options. Possible routing inconsistencies, inefficiency of inter-PCS communications.	• Differentiate q-BGP updates per service option. • At a peering point, the activation of the hard service option must be conditioned by the activation of the loose service option. • Build a management system able to handle simultaneously the two solution options on top of a common and shared network infrastructure.	• Differentiate q-BGP updates per service option. • Build a management system able to handle simultaneously the two solution options on top of a common and shared network infrastructure.
Co-existence of the statistical and the hard solution options,	• Use of meta-QoS-class concept • Use of q-BGP • Information contained in pSLS	• Usage of common and shared network infrastructure for both solution options • Usage of routes learned via q-BGP	• Use different ranges of DSCP values for the two solution options. • Build a management system able to handle simultaneously the two solution options on top of a common and shared network infrastructure. • Differentiate q-BGP updates per service option.	• Use different ranges of DSCP values for the two solution options. • Build a management system able to handle simultaneously the two solution options on top of a common and shared network infrastructure. • Differentiate q-BGP updates per service option.
Co-existence of all solution options	• Use of meta-QoS-class concept • Use of q-BGP • Information contained in pSLS	• Differentiate the intra-domain path and the egress point per solution option • Usage of routes learned via q-BGP • Usage of common and shared network infrastructure for all solution options	• Use a dedicated range of DSCP values for each solution option. • Build a management system able to handle simultaneously all solution options on top of a common and shared network infrastructure. • Differentiate q-BGP updates per service option.	• Use a dedicated range of DSCP values for each solution option. • Build a management system able to handle simultaneously all solution options on top of a common and shared network infrastructure. • Differentiate q-BGP updates per service option.

Copyright © MESCAL Consortium, June 2005

7.4.4 Inter-working scenario

The inter-working scenario deals with technical problems encountered when extending a given solution option through an AS offering different solution option(s). Thus, the following scenarios have been studied:

- Extending the loose service option through the statistical service option,
- Extending the loose service option through the hard service option,
- Extending the statistical service option through the hard service option.

The following table highlights the major problems:

Copyright © MESCAL Consortium, June 2005

	Encountered problems	*Bi-directionality problems*	*Recommendations*	*Adopted solution(s)*
Extending the loose service option through the statistical service option	• What is the methodology for inserting an o-QC in a given meta-QoS-class? • Who will adapts the q-BGP announcements • Usage of routes learned via q-BGP: if no indication is inserted in q-BGP messages, any solution option could pretend that this route is valid from a service point of view. As a result, this could generate black holes in the Internet. • The bandwidth management isn't optimal since: in the loose solution option side policing is done per meta-QoS-class and in the statistical solution option shaping is achieved on a per pSLS basis.	• How to transform a route learned from a loose service option in an o-QC that remains compatible with the solution option? • Differentiate the o-QC that are built thanks to a pure statistical solution option pSLS, and the ones that are bought from an AS offering loose solution options	• Specify a methodology the statistical solution option should follow in order to adapt o-QCs to the meta-QoS-class concept of the loose solution option • Solve the bandwidth management problem • When the two solution options need to coexist in the same AS: • Differentiate q-BGP announcements per solution option • Use different range of DSCP per option for a given PDB. • Deploy the loose solution option when the statistical one is offered	• Differentiate q-BGP announcements per solution option
Extending the loose service option through the hard service option	• Usage of routes learned via q-BGP: if no indication is inserted in q-BGP messages, any solution option could pretend that this route is valid from a service point of view. As a result, this could generate black holes in the Internet.		• Make mandatory the deployment of the loose solution option when the hard solution option is offered • Adopt a single channel signalling: this consists in introducing a dedicated flag in q-BGP messages that will indicate the presence of "hard solution option 3" holes in a given path. • Adopt the double signalling channel. This could be achieved in at least two ways: • Duplicate the q-BGP announcements and indicate the service option it serves • As far as the hard solution option is considered, an AS will announce only its PCS thanks to the use of an identifier (PCSID) associated with QoS performance characteristics.	• Adopt the PCSID signalling
Extending the statistical service option through the hard service option	• Usage of routes learned via q-BGP (if q-BGP is activated by the AS offering statistical service option): if no indication is inserted in q-BGP messages, any solution option could pretend that this route is valid from a service point of view. As a result, this could generate black holes in the Internet.		• Adopt the double signalling channel. This could be achieved in at least two ways: • Duplicate the q-BGP announcements and indicate the service option it serves • As far as the hard solution option is considered, an AS will announce only its PCSs thanks to the use of an identifier (PCSID) associated with QoS performance characteristics.	• Adopt the PCSID signalling

Copyright © MESCAL Consortium, June 2005

8 SERVICE PLANNING AND QOS CAPABILITIES EXCHANGE

8.1 Introduction

The *Service Planning and QoS Capabilities Exchange* functional group [D1.1, Section 6.1] describes the business decisions-related part of the Management plane of the MESCAL functional architecture. The functional block this group is divided to are highlighted in Figure 48 below. These two components are the *QoS based Service Planning* and the *QoS capabilities Advertisement and Discovery* functional blocks.

The planning component's purpose is to aid the AS's decision-making parties by providing them with informational statistics and planning analysis, and to automate the process of enforcing the business decisions, such as new service offerings, by triggering the relevant components of the MESCAL system. The advertisement and discovery components augment the functionality of the planning component by automating the process of QoS capabilities exchange between this AS and its potential business partners –i.e. other ASs.

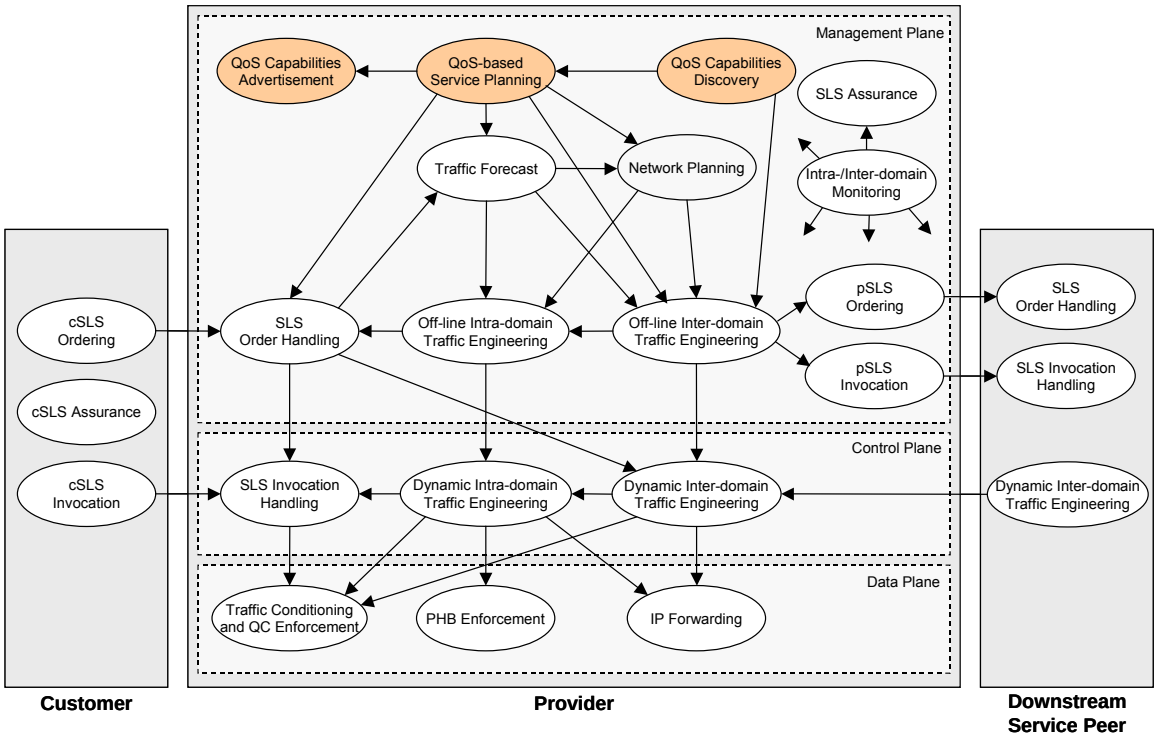


Figure 48. Service Planning and QoS Capabilities Exchange

The interfaces -internal and external- of the Service Planning and QoS Capabilities Exchange functional group are specified in this section. The scope of the functionality of each block will be analysed appropriately.

8.2 QoS-based Service Planning

8.2.1 Objectives

The main focus of the *QoS-Based Service Planning* functional block is the business decisions making processes. Its goal is to facilitate these processes by offering statistics and projections concerning the

Copyright © MESCAL Consortium, June 2005

opportunities of QoS service offerings by the AS, and after a decision has been made to ensure its enforcement by the service provisioning mechanisms of the AS.

Figure 49 presents the QoS based Service Planning component together with the other functional components with which it interacts, within the same AS employing the MESCAL functional architecture. The behaviour of this component is influenced by marketing and business policies.

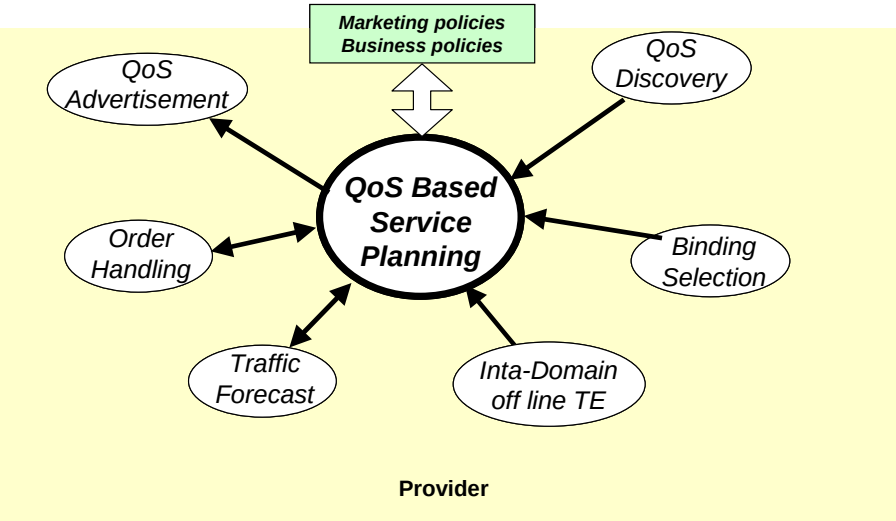


Figure 49. QoS-based Service Planning

8.2.2 Interface Specification

According to the MESCAL functional architecture *QoS-Based Service Planning* implements interfaces with several other components internal to the architecture, but no external interface:

- *QoS Capabilities Discovery*. Through this interface planning receives advertisements of pSLSes offerings by other ASs. These pSLS advertisements contain all the information inherent to pSLS technical aspects plus the terms and conditions for their purchase including cost. Not all advertisements that reach the discovery component are propagated to planning. Filters decided by planning and enforced by discovery will filter out irrelevant advertisements. In addition planning can use this interface to request an active search for pSLS offerings of specific attributes.
- *QoS Capabilities Advertisement*. Through this interface planning mandates the advertisement of the decided service offerings, pSLSes and cSLSes. The information passed to the *QoS Capabilities Advertisement* component is the service offering parameters, restrictions on these parameters, indicative costs and the desired advertisement campaigns.
- *SLS Order Handling*. Through this interface planning receives logs of conducted negotiations for the deduction of statistics concerning the requests of services issued to the AS and the negotiation outcomes. Planning also configures the *SLS Order Handling* function block so as to be able to handle requests for the newly introduced QoS service offerings. This configuration includes the cSLS and pSLS templates, offering restrictions –in terms of SLS parameters such as e-QCs, supported destinations, invocation and assurance means per service offering etc.–, admission logic policies and cost for each service. In the case of newly offered cSLSes the cSLS offering web-server is appropriately configured for handling the ordering of these new services by the end consumers.
- *Traffic Forecast*. Through this interface planning receives the currently valid traffic demand forecast factors per each offered service per each class differentiating usage behaviour. It also receives the established subscriptions. Planning uses this interface to configure traffic forecast algorithms with initial traffic demand predictions for the new services decided to be offered. These predictions will be refined by forecast mechanisms later on based on usage statistics.

Copyright © MESCAL Consortium, June 2005

- *Binding Selection*. Through this interface planning assists in the establishment of new pSLSes by expressing to *Binding Selection* the list of e-QCs that the domain wishes to offer, and giving to *Binding Selection* the set of l-QC capabilities of the AS, and setting the policies regulating this extension.
- *Offline Intra-Domain TE*. Planning receives the calculated resource availability matrix (RAM) each time a new cycle commences.
- Marketing / business logic. Through this interface planning informs the business decision making parties of the AS of statistics concerning service planning and of the results of its algorithms, analysing this data with the aim of deducing the optimum service offerings. Planning receives configuration policies influencing the outcome of its algorithms and approval of its decisions or direct orders, by business marketing authority, for the enforcement of specific service offerings.

8.2.3 Behaviour Specification

The functionality of the *QoS-based Service Planning* functional block aims to deliver the following results:

- Produce a set of potential pSLS and cSLS offerings including indicative costs, offering terms and restrictions as well as proposed advertisement policies. These results will be used by the marketing and business authorities that will make the final decision.
- Offer statistics deduced from data, collected by the operation of several components of the MESCAL architecture, concerning planning and aiming to facilitate the business authorities decisions.
- Realise the approved or ordered service offerings by appropriately configuring *SLS Order Handling*, *Binding Selection, QoS Capabilities Advertisement* and *Traffic Forecast* functional blocks.
- Decide on the potential e-QCs to be offered and policies influencing the binding decisions that will realise these e-QCs. The business authority must approve these decisions first.

8.3 QoS Capabilities Discovery and Advertisement

8.3.1 Objectives

The objectives of the QoS Capabilities Discovery component are to inform the planning component of the advertised capabilities of other ASs by filtering received messages or by performing search for requested QoS connectivity capabilities, initiated by planning queries.

The objectives of the QoS Capabilities Advertisement component is to advertise the QoS connectivity services offered by this provider, as decided by planning, to other ASs and to end consumers.

Figure 50 presents the *QoS Capabilities Advertisement and Discovery* function block of a provider along with its interactions with the internal functional components of the MESCAL architecture and the interactions with external entities such as a virtual market for QoS connectivity and other ASs.

Copyright © MESCAL Consortium, June 2005

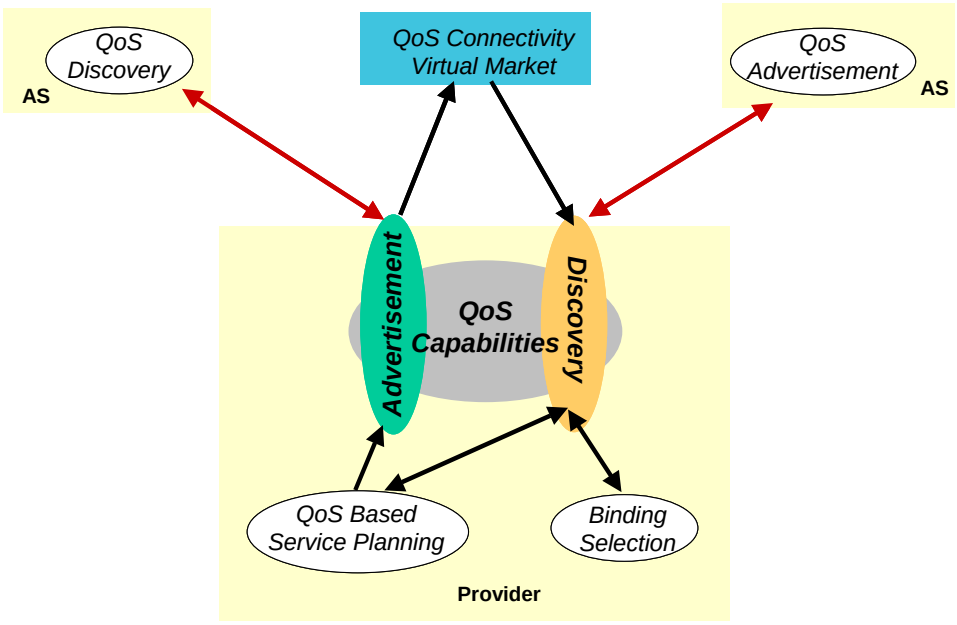


Figure 50. Advertisement and Discovery

8.3.2 Interface Specification

QoS Capabilities Discovery and Advertisement functionality can be divided to the advertisement related and the discovery related process. Each process implements external interfaces for communicating with other ASs or third parties acting as brokers and internal interfaces to exchange information with the planning and the binding components as defined by the MESCAL functional architecture.

8.3.2.1 External Interface

- QoS connectivity Virtual Markets.
- The Advertisement process implements an interface that allows it to publicise advertisements of the provider's offered QoS connectivity services. These advertisements can be customised and targeted to specific consumer groups, also they can be passive, discovered by consumers search, or active, shipped to the consumers.
- The Discovery process implements an interface that allows it to subscribe in order to receive selectively -by setting the desired filters- advertisements of QoS connectivity services offered by other providers. Through this interface it could also launch searches for service offerings fulfilling certain desired criteria or even publicise open requests for specific services.
- Known ASs.
- The Advertisement process implements an interface that allows it to communicate directly with the QoS discovery component of other ASs. These ASs maintain relationships with this provider so their contact point is well known. Through this interface appropriate advertisements according to the relationship between the provider and each AS can be shipped.
- The Discovery process implements an interface that allows it to communicate directly with the QoS advertisement component of other ASs. Same case as before, the ASs maintain relationships with this provider. Through this interface the provider receives advertisements within the limitations imposed by the relationships between the ASs, and also the provider can submit queries for desired services.

Copyright © MESCAL Consortium, June 2005

8.3.2.2 Internal Interface

- *QoS-based Service Planning*. Through this interface QoS advertisement and discovery receives this AS' service offerings and their related advertisement policies. Also the interest of planning for specific service offers is expressed, to be translated to advertisement filters by discovery process. In addition requests for active searches for service offerings of specific attributes may be received. Service offerings acquired on demand or filtered from the received advertisements are forwarded to planning through this interface
- *Binding Selection*. Through this interface *Binding Selection* requests and receives the available QoS connectivity offerings -in terms of offered pSLSes (in particular, the o-QC and destination address prefixes)- that fulfil certain desired criteria.

8.3.3 Behaviour Specification

The functionality of the *QoS Capabilities Advertisement and Discovery* function block aims to deliver the following results:

- Advertisement of the service offerings of the provider through publication to relative virtual markets. This publication is done under specific terms dictating the targeted consumers and the advertisement methods employed. A suitable technology for the implementation of such functionality is the emerging technology of web services including WSDL base language for defining advertisements and UDDI protocol for realising the communications.
- Discovery of available service offerings from other ASs, satisfying certain criteria though virtual markets. The discovery involves the subscription for receiving desired advertisements, the active search for fitting service offerings and the publication of service requests. The implementation of this functionality could be based on web services, as mentioned before, since both advertisement and discovery of services through virtual marketplaces is an undivided system.
- Advertisements of the service offerings of the provider directly to other associated providers. The advertisement process must maintain a list containing all the known ASs, their contact points and their relationship with this provider which determines the relative advertisement policies. Each new service offering, direct advertisement decided by planning can be realised by using this data. A suitable technology for implementing this direct communication is the SOAP protocol -also employed by web services- and for implementing the list of contacts is current database technology.
- Discovery of available service offerings from other ASs directly. This includes the direct reception of advertisement and their filtering, based on criteria dictated by the needs of planning, and the direct querying of other ASs for their offered services. This functionality is realised using the maintained list of ASs as before and could be implemented using the same technologies.

Copyright © MESCAL Consortium, June 2005

9 SLS MANAGEMENT

9.1 Introduction

The SLS Management functionality is a major part of the management plane of the MESCAL functional architecture as introduced by deliverable D1.1 [D1.1]. By SLS Management we name the functionality responsible for handling the technical contracts -modelled as cSLS and pSLS- that specify the QoS connectivity services offered or acquired by the provider. SLS Management is essential for providing QoS connectivity services because it undertakes the task of establishing agreements that will allow the provider to expand its network's QoS connectivity beyond its domain, as well as the task of responding to the requests for services from the provider's customers taking into account predictions of the capacity of the network and business directives.

The SLS Management functionality can be split into two parts: (a) the part responsible for the contracts offered by the provider to its customers, i.e. the end-customers and interconnected providers, and (b) the part responsible for the contracts requested by the provider from its peer providers. The resulting functional components are named "*SLS Order Handling*" and "*SLS Ordering*" respectively. The communication between these components is based on a protocol especially designed for the conduction of the required negotiations for service purchase, the SrNP (Service Negotiation) protocol. While the ordering process establishes the contracts between the peering providers, the invocation process is required to commit resources before traffic can be exchanged, with "*SLS Invocation Handling*" and "*pSLS Invocation*" providing the necessary functionality.

The MESCAL functional architecture is presented in Figure 51 below with the components realising the SLS Management functionality highlighted.

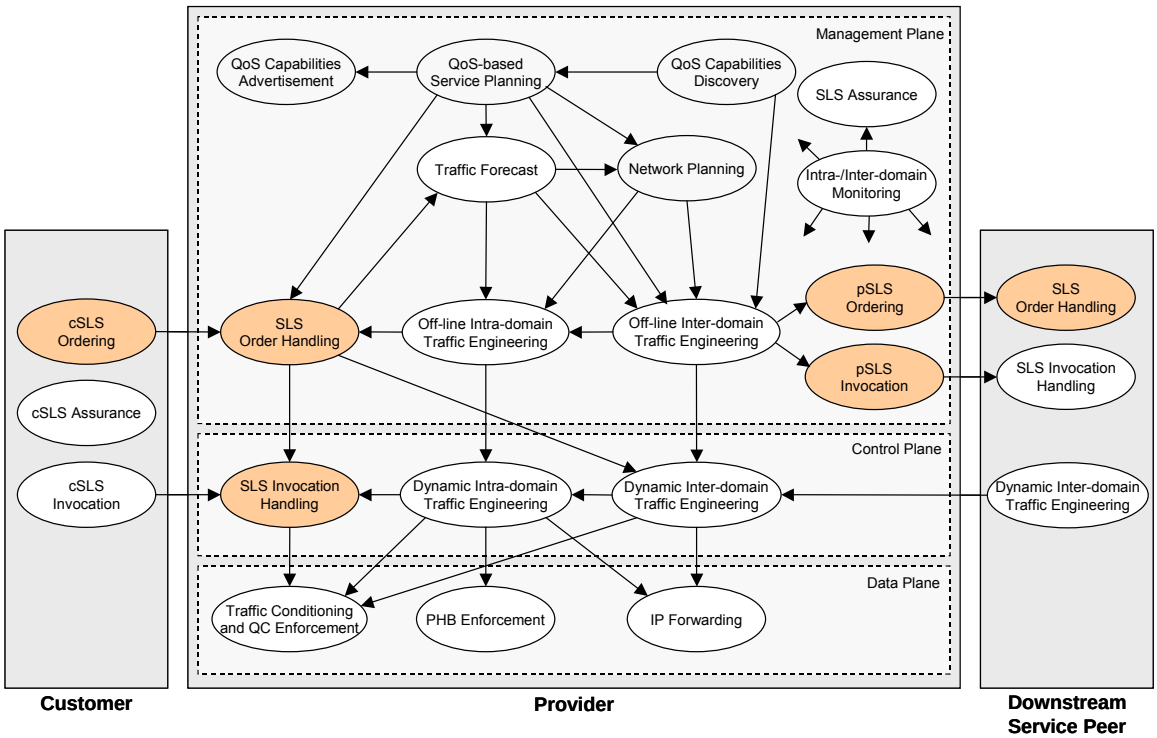


Figure 51. The MESCAL functional architecture

This section contains the following contents. Specifications for pSLSes are given in section 9.2. The Service Negotiation Protocol (SrNP) for inter-domain QoS is described in section 9.3. The MESCAL functional architecture function blocks are then detailed: *SLS Order Handling* in Section 9.4, *pSLS*

Copyright © MESCAL Consortium, June 2005

Ordering in Section 9.5, *cSLS Ordering* in section 4.6, *pSLS Invocation* in section 9.6 and finally SLS *Invocation Handling* in Section 9.7.

9.2 cSLS and pSLS specifications

9.2.1 Introduction

The current trend in service offering is agreement (contract)-based. The term *Service Level Agreement (SLA)* is widely used to denote such an agreement. It describes the characteristics of the service offering and the mutual responsibilities of the parties involved for using/providing the offered service. The term *Service Level Specifications (SLS)* is used to denote the technical characteristics of the service offered in the context of an SLA. The service technical characteristics refer to the provisioning aspects of the service e.g. request, activation and delivery aspects from network perspectives. Non-technical service provisioning aspects such as billing and payment aspects, are not part of the SLS; they are part of the overall SLA. SLS are an integral part of a SLA, and conversely a SLA includes SLS.

MESCAL is concerned with SLS; service accounting and billing aspects are outside the scope of investigation. As the MESCAL solution for QoS delivery in the Internet adopts a hop-by-hop, cascaded model of interactions between providers, both at the service and network (IP) layer, we distinguish two types of SLS (and subsequently of SLAs):

- *cSLS*, established between end-customers and providers, and
- pSLS, established between providers with the purpose to back-up agreements at a service level for expanding the geographical span of their services.

The definition of c/pSLS, from an informational viewpoint, is the main theme of this chapter. Specifically, this chapter specifies suitable templates, set of parameters with clear semantics, for completely describing the contents of c/pSLS. We firmly believe that there is a need for standardising c/pSLS to the benefit of Internet service deployment and provisioning. Standardised c/pSLS would provide a common informational basis for the interactions between end-customers and providers and between providers, as well as for building the required service provisioning functionality; thus, enabling the automation of the respective processes.

The essence of our specification work is to look at pSLS under two angles: (a) as agreements between providers for QoS-traffic exchange, pertinent to the particular relationships holding in the business model for QoS provisioning in the Internet and (b) as QoS-based services offered by a provider. Clearly, there is a strong interrelation between these two aspects and each poses its own requirements on the content of pSLS. Analysing these requirements, suitable templates are specified. Note that this does not apply to cSLS, which mainly encompass QoS-based service offering aspects. As such, without loss of generality, we focus on pSLS; cSLS are quite similar.

The section is organised as follows. First, the requirements underlying our specification work are outlined along line the different types of pSLS, which can be distinguished depending on the business context they are to be established in. Subsequently, by viewing that c/pSLS represent the QoS-based connectivity services offered by providers, templates for describing in detail all aspects of QoS-based services are specified. While these templates present an open, detailed pSLS model, suitable condensed, summarised pSLS models are then specified, as required by the specific requirements posed by the different types of business relationships between providers.

9.2.2 Types of pSLS and Specification Requirements

This section is specific to pSLS, not to cSLS.

pSLS form the basis of the agreements between providers for traffic exchange in the Internet. In essence, pSLS extend, to the end of QoS traffic exchange, the respective agreements that exist today between the providers in the best-effort Internet -for transiting or inter-exchanging traffic- [HUST]. As

Copyright © MESCAL Consortium, June 2005

such, they should be in-line with the specific context of traffic exchange, which is implied by the particular business relationships holding between providers.

The business model, relationships and financial settlements between providers in today's best-effort Internet as well as in the MESCAL-enabled QoS-aware Internet are described and discussed from QoS perspectives in MESCAL deliverable D1.4 [D1.4, chapter 6]. Based on the analysis therein, the MESCAL solution advocates two basic business cases:

- A business case for the provisioning of QoS-based services relying only on loose QoS guarantees - qualitatively expressed performance targets, and no bandwidth guarantees.
- A business case for the provisioning of QoS-based services relying on statistical guarantees for quantitative performance targets and bandwidth, in addition to qualitative QoS guarantees.

It should be noted that in either of the above business cases, services relying on hard QoS guarantees could also be provided, by establishing MPLS-based tunnels (LSPs) between specific points in the Internet; however, this service is not for the mass market because of scalability limitations inherent in the technical solution.

The qualitative-QoS Internet business case directly corresponds to the three-tier, *hierarchical* model currently in place in the best-effort Internet. In this case, the business relationships between providers are completely determined by their relative positioning in the hierarchy; the following types are distinguished:

- *The MESCAL pSLS-based customer-provider* relationship, whereby, one provider -said to be acting in a provider-business-role- provides the QoS Internet connectivity service, as seen by its domain, to the other provider -said to be acting in a customer-business-role. Usually, this type of business relationship is between providers belonging to different levels of the three-tier Internet model, with the provider in the lower tier being a customer of the provider in the upper tier.
- *The MESCAL pSLS-based peer-to-peer* relationship, whereby, the providers mutually agree to exchange QoS traffic between their domains; not transiting traffic to their providers or to other peer-to-peer providers, although the latter could be a possibility. This relationship is a kind of 'short-cut' to prevent traffic flowing into the upper tiers and, usually, is between providers of similar size -belonging to the same tier.

The statistical-QoS Internet business case advocates a *flat* Internet, where the business relationships between providers are not affected nor dictated by the relative positioning of the providers in the three-tier hierarchy; we propose the following common type of business relationships between providers:

- *The MESCAL pSLS-based (upstream)-QoS-proxy* relationship, whereby, either of the providers may request from the other provider to provide a transit QoS-based connectivity service to (a subset of) anywhere the latter provider can reach in the Internet with this QoS. The provider offering the transit QoS service would have built its QoS reach capabilities based on similar agreements with (some of) its directly attached providers, which in turn would have built their own QoS reach capabilities based on similar agreements with (some of) their own adjacencies and so on. Therefore, each provider in a chain of QoS-proxy relationships established in the same direction appears as kind of a 'proxy' of the providers further along this direction. This type of business relationship is of a strong transitive nature, while is not following a strict customer-provider business paradigm; it could be thought as being the QoS Internet counterpart of a call-termination agreement in the PSTN and VoIP business world.

Based on the above discussion, the following different types of pSLS are distinguished:

In the hierarchical Internet business case:

- *Provider loose QoS Internet access pSLS* – suitable for customer-provider business relationships, offered by providers wishing to undertake a provider-business-role
- *Provider loose QoS tunnels in the Internet pSLS* – as above

Copyright © MESCAL Consortium, June 2005

- *Peer loose QoS traffic inter-exchange pSLS* – suitable for providers wishing to establish corresponding peer-to-peer business relationships
- *Peer loose QoS tunnel extension pSLS* (the term extension is meant from/to this domain to/from another domain) – as above

In the flat Internet business case:

- *Proxy statistically guaranteed QoS Internet access pSLS* – suitable for QoS-proxy business relationships, offered by providers wishing to provide a QoS transit service
- Proxy statistically guaranteed QoS tunnels in the Internet pSLS – as above

The pSLS identified above differ each other in the type of the offered QoS guarantees, the directionality and topological scope of the traffic flows, according to the business case and the particular context of business relationship they refer.

In customer-provider business relationship, pSLS have the connotation of agreements for the provider in the customer-business-role to 'join in' (send and receive traffic) the QoS-aware Internet as seen by the other provider. They imply a bi-directional flow of QoS traffic and can only offer qualitative QoS guarantees to all destinations that can be reached by the provider in the provider-business-role.

pSLSes between peer-to-peer providers have the connotation of mutual agreements for the exchange of QoS traffic from one provider domain to the other provider domain. They imply a bi-directional flow of QoS traffic and offer qualitative QoS guarantees within the scope of the provider domains.

In the upstream-QoS-proxy business relationship, pSLS have the connotation of agreements for the provider offering the pSLS, pSLS-provider, to deliver QoS traffic from the other provider, pSLS-requestor, to (a subset of) the destinations that can be reached from the pSLS-provider with this QoS. They imply a unidirectional flow of QoS traffic, from the pSLS-requestor to the pSLS-provider and may offer statistical and/or qualitative QoS guarantees to certain destinations in the Internet -those reachable by the pSLS-provider.

Once pSLS are in place allowing providers to establish inter-domain QoS tunnels, these providers could offer to their end-customers cSLS with hard QoS and bandwidth guarantees.

Up to now, by viewing pSLS as agreements underlying the business relationships between providers, we analysed their intrinsic aspects regarding the characteristics of the QoS traffic flows that they imply. A number of different types of pSLS are required, as a result of the different types of business relationships that may hold between providers in a MESCAL-enabled QoS-aware Internet. Subsequently, the pSLS information specification task has to meet the following challenge:

- Provide for a common, 'well-known and understood' vocabulary to describe pSLS contents in a way that can satisfactorily fulfil the following two diverse requirements:
  - capture the essential aspects of the agreements between providers for QoS traffic exchange as implied by their business relationships, to the benefit of facilitating provider interactions and therefore service deployment in the Internet–different pSLS are bound to exist; while at the same time
  - considering that pSLS express QoS-based service offers, create a stable informational basis for building service management and traffic engineering functions, to the benefit of automated service provisioning and graceful delivery; although there may be a number of different pSLS, they should be supported by a common set of functions.

To the above end, MESCAL first specifies a general, open, detailed service model for describing pSLS as well as any QoS-based service; thus fulfilling the latter requirement. Subsequently, by appropriately restricting and/or summarising the information identified in this open service model, suitable models for describing the different types of pSLS, as identified per business case, are specified; thus fulfilling the former requirement. It should be noted that our specifications concentrate on service connectivity aspects; aspects such as service accounting, monitoring, billing and payment are not included. All these are presented in the subsequent sections.

Copyright © MESCAL Consortium, June 2005

9.2.3 A General Model for pSLS and QoS-based Services

By viewing pSLS as QoS-based service offers, this section specifies a general model for describing the technical (connectivity) aspects of such services, which are required for their provisioning and need to be agreed upon the provider and its end-customers or its peering providers i.e. their SLS.

Our work draws from the SLS template specification work of the IST TEQUILA project [TEQUI]. TEQUILA specified a service management and traffic engineering framework for intra-domain QoS provisioning [GODE02a], [TEQUI,D1.4], [GODE02], which prompts for standardisation of the notion of SLS, proposing a standard template. The proposed *SLS template (SLS-T)* is considered as the nucleus of IP QoS-based services. Broadly speaking, recognising that from connectivity perspectives, QoS-based services may be comprised of several 'connectivity legs' (e.g. between several sites), SLS-T describes the technical characteristics of a single 'connectivity leg' -topology, IP flows, transfer quality characteristics, traffic compliance criteria. The connectivity aspects of a service then, are the collection of suitable SLS-T's bound to the same customer and the same access/usage means and characteristics. As such, the TEQUILA service management framework has specified the notion of *SSS-T (Service Subscription Structure Template)*, which may contain a number SLS-Ts, to describe the whole of the connectivity aspects of a QoS-aware IP connectivity service.

The SSS-T is the general, open, detailed model for pSLS and in general for SLS of any QoS-based service, adopted by MESCAL. The particular instances of the SLS-T and SSS-T templates for a particular QoS-based service are simply denoted by SLS and SSS. Note that in the context of specifications, the term SLS is meant as a unidirectional connectivity leg of a QoS-aware service, whereas, in any other context, this term denotes the general technical characteristics of a QoS-aware service; the latter corresponds to a SSS in the context of specifications.

The following sections present the TEQUILA-based SLS and SSS templates, highlighting the enhancements and clarifications that need to be made according to the MESCAL inter-domain perspectives.

9.2.3.1 SLS Template Specifications

SLS-T is specified against the following information elements (clauses), which are described in the following:

- SLS Identification
- Scope
- Flow Identification
- Traffic Conformance (Envelope)
- Excess Treatment
- Performance Guarantees
- Agreement Type

9.2.3.1.1 SLS Identification

A key, uniquely identifying the SLS in the context of a SSS; it is set by the provider.

9.2.3.1.2 Scope

The *Scope* clause explicitly identifies the geographical/topological region over which the QoS policy, as specified by this SLS, is to be enforced by indicating the boundaries of that region. It includes the following attributes:

- *Ingress,* indicating the entry point of the region over which SLS is to hold
- *Egress,* indicating the exit point of the region over which SLS if to hold

Copyright © MESCAL Consortium, June 2005

The *Ingress* and *Egress* attributes can take the following values:

<interface identifier | set of interface identifiers | label | set of labels | any>, where:

"|" denotes an exclusive OR, "label" denotes a mutually agreed upon identifier uniquely identifying a set of interface identifiers or a set of remote destinations, i.e. interfaces not directly attached to the provider and "any" is logically equivalent to unspecified.

The *Ingress/Egress* interface identifier may be an IP address or a layer-two identifier in case of Ethernet or unnumbered PPP-based access links in case of Point-to-Point Protocol or any other well-defined identifier uniquely determining a boundary link as defined in [RFC-2475].

The definition of *Ingress/Egress* is necessitated by the fact that providers cannot provide for QoS guarantees over the aggregation/distributions networks that usually intermediate between end-customers and provider domains. The values of the *Ingress/Egress* attributes may be deduced by other SLS or SSS attributes or by the traffic engineering functions of the provider. Usually, in the case of inter-domain services offered to end-customers and agreements with peering providers, one of these attributes corresponds to the interface of a customer access or an interconnection link, while the other attribute is left unspecified or set to an appropriate label denoting at high-level the points where liability for QoS policy enforcement ends for reaching a specific set of destinations or the boundaries of a particular domain (ASs). As an example, in the case of Internet access services offered to end-customers the value of the *Ingress* of the upstream SLS could be deduced by the customer information and the value of the *Egress* would be "any"; the latter could be refined to denote the interface of a particular inter-domain link by the traffic engineering functions, however, this is an internal matter, being not subject of agreement. In the case of VPN services offered to end-customers, both these attributes should be clearly specified.

The *Ingress/Egress* should not be confused with the characteristics of the flows entitled to receive the treatment of this SLS (cf. *Flow Identification* clause, below). They are quite distinct in semantics. Ingress/Egress, if specified, imply that the SLS traffic will have to pass through these points (interfaces); the issue is discussed in more detail in section 9.2.3.1.3.

The following combinations of Ingress, Egress values are allowed:

(1,1) - implying an one-to-one communication; we call the SLS as *a pipe SLS*

(1,N) - one-to-many communication (N>1); we call the SLS as a *hose SLS*

(1,any) - one-to-any communication; we call the SLS as an *unspecified hose SLS*

(N,1) - many-to-one communication (N>1); we call the SLS as a funnel SLS

(any,1) - any-to-one communication; we call the SLS as an *unspecified funnel SLS*

Because SLSes in this template are assumed unidirectional QoS-based connectivity (legs of) services, the above taxonomy excludes the many-to-many communication (M, N); either *Ingress* or *Egress* attributes must be specified to exactly one interface identifier. Many-to-many communication can be achieved at the level of SSS, where a number of SLS are combined.

In case where the specified sites are remote, their mapping to the provider's domain boundary links is subject to the contracts and routing decisions in place. Hence, internally in the provider and transparently to the agreed SLS, traffic to remote sites may be merged over one boundary link or split to many boundary links turning a hose to a pipe, or a pipe to a hose, etc.

9.2.3.1.3 Flow Identification

The *Flow Identification* (*Flow Id*) clause defines the stream of IP datagrams, at an IP level, for which, the QoS policy, as specified by this SLS, is to be enforced. It includes the following attributes:

- *Differentiated Services Information*, specifying possible values of the DSCP field in the IP header for characterising the packets entitled to the SLS; it can take the following values: <DSCP value | set of DSCP values | any>

Copyright © MESCAL Consortium, June 2005

- *Source Information,* specifying possible values of the source IP address field in the IP header for characterising the packets entitled to the SLS; it can take the following values: <source IP address | set of source IP addresses | source IP prefix | set of source IP prefixes | any>
- *Destination Information*, specifying possible values of the destination IP address field in the IP header for characterising the packets entitled to the SLS; it can take the following values: <destination IP address | set of destination IP addresses | destination IP prefix | set of destination IP prefixes | any>
- *Application Information*, specifying possible values of application-related fields in the IP header for characterising the packets entitled to the SLS; it can take the following values: <sets of protocol number, source port, destination port combinations | any>

The term "any" appearing above is logically equivalent to all.

Usually, each SLS must always have a single *Flow Id* clause with specified information along the above attributes. This is dependent on the nature of service; for instance, this clause may not be specified in the case of services for QoS tunnel set-up.

In essence, the *Flow Id* clause provides the necessary information for classifying the packets at the provider inbound link (cf. *Ingress*). The necessary information is included to enforce either an Aggregate (BA)- or a Multi-Field (MF)-based classification. In case of MF-classification, all above attributes may be specified; this classification may depict micro-flows as well as aggregate macro-flows. In case of BA-classification, the *Differentiated Services Information* attribute i.e. DSCP information must be specified, while the other attributes must not be specified. For scalability and performance reasons, especially for inter-domain services and related agreements, a BA-based classification is highly recommended; one should avoid fine-grained classifications or classifications based on multiple fields, even though aggregate traffic.

It should noted that the DSCP-value(s) specified in this clause, has(have) as such nothing to do with the DSCP-marking of packets inside the domain. The information included in this clause is solely to the purpose of identifying the traffic belonging to the contract underlying the SLS; therefore is agreement-specific, not behaviour/engineered-capabilities specific nor revealing. Following classification, packets may be remarked by the provider to the appropriate DSCP, as required to receive the QoS treatment specified by the SLS. In the case of inter-domain services, the packets when leaving the domain may need to be remarked again to the DSCP corresponding to the SLS established with the determined next-hop peering provider, where they may need to be remarked again to the appropriate domain-specific DSCP to receive the required QoS treatment and so on.

Finally, the relationship between the *Scope* and *Flow Id* SLS information and their implications to routing are discussed. In general, if only *Flow Id* is specified and the *Ingress/Egress* are unspecified, or specified at a high-level by means of a label denoting the boundaries of a domain where QoS enforcement liability ends for reaching specific destinations, then, this is taken that there is no a-priori assumption about the actual *Ingress/Egress* points that the traffic will cross. Indeed, it is the responsibility of the provider to define the most appropriate route through its intra and inter-domain traffic engineering and routing policies. Thus, in this case, the *Ingress/Egress* information, which in this case is not an explicit part of the SLS, is implicitly derived by the routing policy of the provider. On the other hand, if both *Flow Id* and *Ingress/Egress* are explicitly specified, say by the pairs (DSCP, IP-src, IP-dest) and (IP-ingr, IP-egr) respectively, then, it is taken, that IP packets, adhering to the *Flow Id* information, must follow the route (IP-src, ..., IP-ingr, ..., IP-egr, ..., IP-dest). Conclusively, the information under the *Scope* and *Flow Id* clauses has different semantics, although in some cases unspecified information in one clause could be implicitly derived by the specified information in the other clause. Further, when information in both clauses is specified, this poses requirements on routing in that: the specified *Ingress/Egress* in the *Scope* clause should always be en-route of the packets specified in the *Flow Id* clause, in other words, packets must always be routed through the *Ingress/Egress* points, if these are specified.

Copyright © MESCAL Consortium, June 2005

9.2.3.1.4 Traffic Conformance (Envelope)

The *Traffic Conformance* clause describes the criteria (characteristics) that the traffic injected in the provider domain should comply with, in order to get the QoS guarantees specified by the *Performance Guarantees* clause. In essence, this clause sets the sufficient conditions at a traffic-rate level, that is, for the flows of the packets, entitled to the SLS (cf. Flow Id clause), to receive the specified QoS. It includes the following attributes:

- *Traffic Conformance Algorithm*, specifying the type of the mechanism, which is used to unambiguously identify the packets which comply with the traffic conformance criteria and those which do not, called the "in" and "out" of profile packets, respectively.
- *Traffic Conformance Parameters*, a set of parameters required as input by the *Traffic Conformance Algorithm;* generally speaking, these parameters express the traffic conformance criteria in terms of rate (bandwidth) thresholds.

Basically, this clause includes the information required for configuring the traffic conditioners at the provider edges or border gateways for controlling the traffic injected in the provider domain.

Examples of *Traffic Conformance Algorithms* are: leaky bucket, token bucket, combined token bucket with peak, a two-rate three-colour marker scheme and an MTU-based scheme. Associated *Traffic Conformance Parameters* include: peak rate, token bucket rate, bucket depth and maximum transfer unit (MTU).

9.2.3.1.5 Excess Treatment

The *Excess Treatment* clause describes how excess traffic, i.e. out-of-profile traffic will be processed. The process takes place after the application of the *Traffic Conformance Algorithm* (cf. *Traffic Conformance* clause). It includes the following attributes:

- *Action*, specifying the action to be taken for the excess traffic; it can take the following values: <drop (default) | shape | remark>.
- *Action Parameters*, a set of parameters that may be required by the action taken e.g. for remarking, the alternative performance guarantees must be specified and for shaping, the buffer size of the shaper.

9.2.3.1.6 Performance Guarantees

The *Performance Guarantees* clause describes the guarantees on packet transfer performance parameters (metrics) that the provider (agrees to) offers to the packets entitled to the SLS (cf. *Flow Id* clause) within the limits of the SLS geographical/topological span (cf. *Scope* clause). The guarantees to be given are subject to the SLS traffic conformance criteria (cf. *Traffic Conformance* clause); guarantees are given for each of the conformance levels, in case of a multi-level *Traffic Conformance Algorithm*, whereas for out-of-profile no particular guarantees can be given. This clause includes the following self-evident attributes, corresponding to the packet transfer performance metrics against which performance guarantees are given.

- *Delay Guarantees,* specifying the guarantees for the one-way packet delay as measured between specific ingress and egress points crossed by the entitled SLS traffic.
- *Jitter Guarantees*, similar to the above.
- *Loss Guarantees*, specifying the guarantees for the packet loss probability; this is defined as the ratio of the lost in-profile packets between specific ingress and egress points and the injected in-profile packets at ingress.
- *Throughput Guarantees,* specifying the guarantees for rate of the traffic delivered, that is, as measured at a specific egress point, counting all packets entitled to the SLS. Note that all packets, independently of their conformance level (in/out-of-profile) contribute to measuring the delivered throughput. Indeed, if a customer (only) wants throughput guarantee for its traffic, then he/she

Copyright © MESCAL Consortium, June 2005

does not care whether in- or out-profile packets are dropped, but is only interested in the overall throughput of its generated packet stream.

It may not be necessary for all above attributes to be specified.

The following aspects underlying the semantics of the above attributes are worth noting:

Performance guarantees can only have meaning within a certain topological scope (cf. *Scope* clause), which is usually designated by couples of ingress and egress points; this scope should be well-defined and understood by both the provider offering the SLS and the customer –end-customer of peering provider.

Delay, jitter and packet loss guarantees refer to the in-profile traffic, conforming traffic injected in the domain, whereas throughput guarantees refer to the overall traffic hit the provider boundary.

The following types of performance guarantees are distinguished: *quantitative* and q*ualitative*. The guarantees to a particular performance parameter are said to be quantitative, if they can be expressed in quantitative, numerical, values. Otherwise, they are said to be qualitative; possible qualitative values, as appropriate as per performance parameter, may include: high, medium, low or red, yellow, green. The quantification of the relative difference between the qualitative values is a matter of provider's policy e.g. 'high' could be twice good as 'medium', which in turn is twice as good as 'low'.

Quantitative performance guarantees are expressed as maximum (worst-case) bounds or as (sets of) percentiles or inverse percentiles, indicating also the granularity period of the associated measurements. The meaning of the values of qualitative performance guarantees and/or their relative difference should be clear to the customers, while it should be backed-up with relevant historical performance data.

Similarly, we can distinguish between *quantitative* and q*ualitative* SLSes. A SLS is said to offer quantitative guarantees, if all the specified guarantees are quantitative; otherwise, is said to offer qualitative guarantees and in the case where no guarantees are specified, the SLS is said to be a *best-effort* SLS.

Finally, the following relationships and dependencies must hold between the information defined in this clause and the information under the *Traffic Conformance* and *Excess Treatment* clauses:

Quantitative delay/jitter/loss guarantees cannot be given unless a rate-based *Traffic Conformance Algorithm* is specified, that is, such guarantees can only be given for in-profile traffic and as such, explicit bandwidth constraints on the offered traffic must have been defined.

For in-profile traffic, loss and throughput guarantees are equivalent and only one of them should be specified. The same holds for qualitative guarantees.

Related to the above, quantitative throughput guarantees, in addition to quantitative loss guarantees, can only be given if excess traffic is remarked, not dropped or shaped.

If quantitative throughput guarantees are only given, then the *Traffic Conformance Algorithm* may not be specified. However, the provider may still wish to protect its domain by requesting for the specification of a *Traffic Conformance Algorithm* e.g. setting a bucket token mechanism to operate such that the average rate of the traffic injected in the domain to almost equal to the guaranteed throughput rate.

9.2.3.1.7 Agreement Type

The *Agreement Type* clause specifies the type of the agreement pursued on the particular SLS. A SLS may be *requested* or *offered* allowing for SLAs between peer providers. From the point of view of each party in a SLA between peer providers, an SLS is requested for the traffic to traverse the provider's domain and another SLS is offered for the traffic to be routed through the peer provider's domain.

Copyright © MESCAL Consortium, June 2005

9.2.3.2 SLS Template XML Modelling

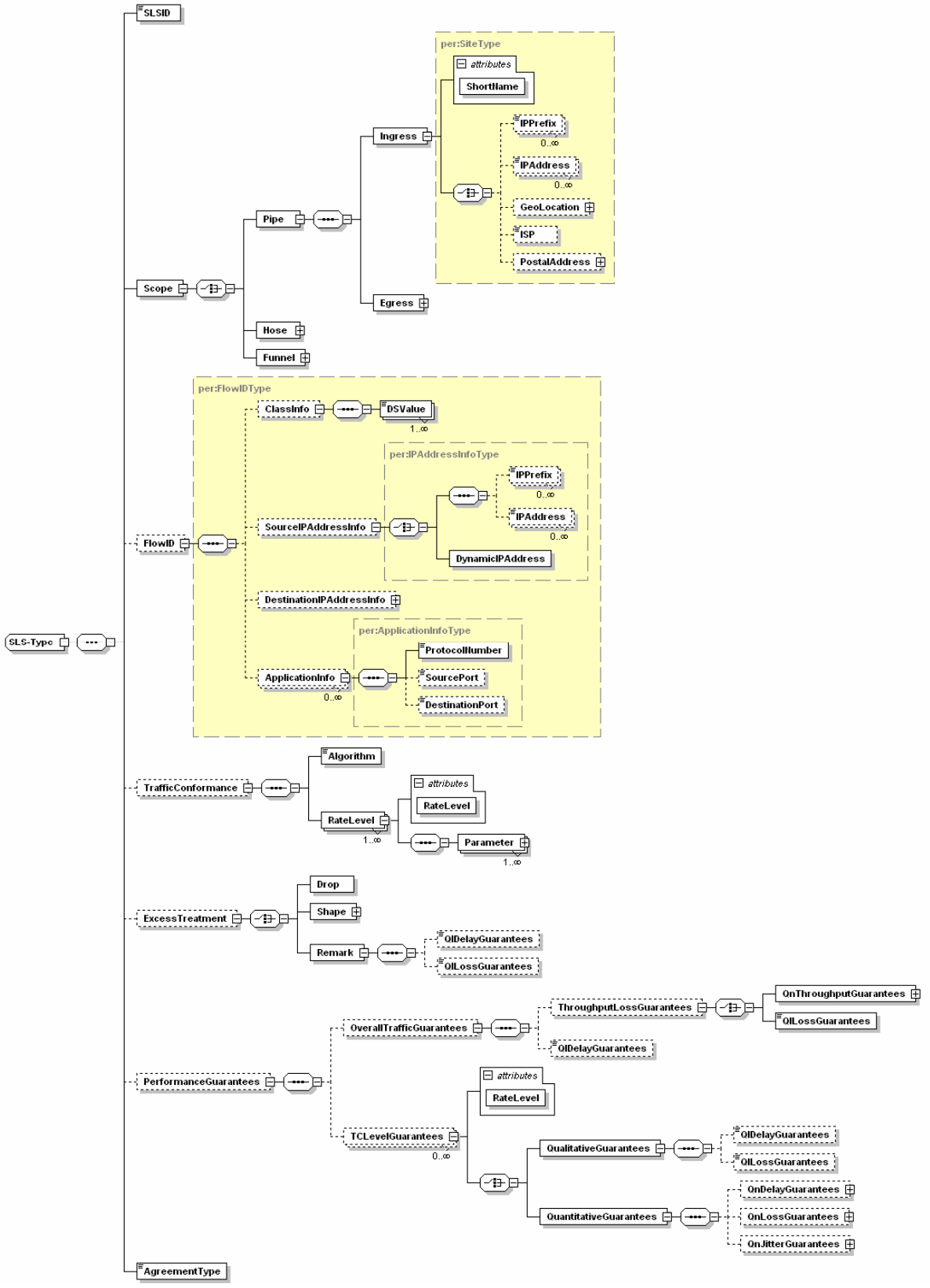


Figure 52. SLS-T XML Schema

Copyright © MESCAL Consortium, June 2005

Element	Description
SLSID	The SLS identification key (see section 9.2.3.1.1).
Scope	The topological boundaries of the SLS (see section 9.2.3.1.2).
Scope.Ingress/Egress	Directly attached sites may be identified by a postal address or the IPPrefixes/IPAddresses of the customer site given that the access link they are attached to is already known to the system. An IP Address may explicitly point to the boundary link interface. Geographical locations, ISP names and/or IPPrefixes/IPAddresses may point to remote sites, the mapping of which to boundary links is subject to the contracts and routing decisions in place. GeoLocation attributes: Country, Region, City PostalAddress attributes: Street, Number, AreaCode, Country
FlowID	The characteristics of the flows entitled to use the SLS (see section 9.2.3.1.3).
FlowID.ClassInfo	Characterisation of flows based on their class as this is inferred by the ToS/DS-byte of the IP header.
FlowID.SourceIPAddressInfo/ DestinationIPAddressInfo	Characterisation of flows based on source and destination IP addresses.
FlowID.ApplicationInfo	Characterisation of flows based on application information specified by the protocol number, source and destination ports.
TrafficConformance	The conformance criteria that the traffic should meet to enjoy the performance guarantees of the SLS (see section 9.2.3.1.4).
TrafficConformance.Algorithm	The traffic conformance algorithm and parameters to apply per conformance rate level.
ExcessTreatment	How traffic not conforming to the conformance criteria should be treated by the provider (see section 9.2.3.1.5).
ExcessTreatment.Drop	Denotes that non conforming traffic will be dropped as the default treatment.
ExcessTreatment.Shape	The parameters of the shaper, if shaping is the desired excess treatment.
ExcessTreatment.Remark	The qualitative guarantees desired for non conforming traffic, if remarking is the desired excess treatment.
PerformanceGuarantees	The guarantees offered to conforming traffic in terms of packet transfer characteristics (see section 9.2.3.1.6).
OverallTrafficGuarantees	The performance guarantees for the overall traffic, including non conformant traffic in case traffic conformance clause is specified.
TCLevelGuarantees	The performance guarantees for a particular traffic conformance level.
QuantitativeGuarantees	Quantitative guarantees on packet delay (*QnDelayGuarantees*), loss (*QnLossGuarantees*), jitter (*QnJitterGuarantees*) and throughput (*QnThroughputGuarantees*) Attributes: UpperBound, Quantile, TimeInterval
QualitativeGuarantees	Qualitative guarantees (e.g. *premium*/*gold*) on packet delay (*QlDelayGuarantees*) and loss (*QlDelayGuarantees*).
AgreementType	The type of pursued agreement, SLSes are by default requested, however they may also be offered in the context of SSSs between peer providers (see section 9.2.3.1.7).

Table 4. SLS-T XML Elements

9.2.3.2.1 Logical Validation Rules

SLST-L1: The ingress and egress points specified in the topological scope of the SLS should be reachable by the network.

SLST-L2: If specified, the IP addresses included in *SourceIPAddressInfo* and *DestinationIPAddressInfo* should be reachable from *Ingress*, *Egress* specified in the *Scope* of SLS.

SLST-L3: If *ExcessTreatment* is specified, then *TrafficConformance* must also be specified.

Copyright © MESCAL Consortium, June 2005

SLST-L4: The number of *RateLevel* elements under *Algorithm* under *TrafficConformance* and *TC Level Guarantees* elements under *PerformanceGuarantees* must be the same.

SLST-L5: *OverallTrafficGuarantees* may be specified only in the following exclusive cases, and not in any other case: (a) when *TrafficConformance* is not specified, or (b) when *TrafficConformance* is specified and *ExcessTreatment* is specified to *Remark* (not *Shape*).

SLST-L6: The information that may be specified under *FlowID* should be unique amongst all subscriptions of the subscriber (for ensuring that packets can be differentiated per SLS, per subscription of the subscriber).

SLST-L7: If in the *Scope* clause a customer site is specified (by *PostalAddress* or by *IPAddress/IPPrefix*) and the access means are *dialup* (see SSST-I2 interpretation rule in section 9.2.3.4.2) then the router the customer site is attached to must have dialup capabilities.

SLST-L8: If a *Site* element within the *Scope* clause is network server or internet gateway (see SLST-I2 bellow) then the site of the opposite end must either contain a *GeoLocation* element or indicate a customer site with known geographical location.

9.2.3.2.2 Interpretation Rules

In a valid SLS-T XML document:

SLST-I1: If no site is specified in the ingress/egress of the *Scope* clause then wildcard internet is assumed (*.*.*.*).

SLST-I2: If a *Site* element within the *Scope* clause has no *IPAddress*, *IPPrefix*, *GeoLocation*, *ISP* or *PostalAddress* elements then the site points to any dialup server, or to a specific internet gateway or network server, as indicated by the site's *ShortName* attribute set to *modem*, *internet gateway* and *network server* respectively. The internet gateway or network server router is determined based on the area of the site of the opposite end of the *Scope* clause (*Ingress/Egress*). For example, if the egress site of an SLS is an internet gateway, then the area where the SLS ingress is located determines which internet gateway and corresponding router will be used to route the traffic for this SLS.

SLST-I3: If *TrafficConformance* is specified and *ExcessTreatment* is not specified, then a *drop* action is assumed for treating the excess traffic.

SLST-I4: If *TrafficConformance*, *ExcessTreatment* and *PerformanceGuarantees* are not specified, then a best-effort packet treatment is assumed.

SLST-I5: In case *OverallTrafficGuarantees* and *ExcessTreatment* are specified then the QoS-class is set according to the guarantees specified in *OverallTrafficGuarantees* and the *Remark* parameter is ignored; otherwise (i.e. if *OverallTrafficGuarantees* is not specified), then if the remarking parameter is specified, it denotes the overall traffic guarantees to be given and if not, a best effort packet treatment is assumed.

9.2.3.3 SSS Template Specifications

SSS-T is specified in terms of the following information elements (clauses), which are described in the following:

- Subscriber Info
- Subscription Id
- Set of SLS
- Invocation Means
- User Info
- Grade of Service

Copyright © MESCAL Consortium, June 2005

- Activation Info
- Schedule
- Availability Guarantees
- Reliability Guarantees

9.2.3.3.1 Subscriber Info

The *Subscriber Info* clause includes the required information to uniquely identify a customer, an end-customer or a peering provider requesting a QoS-based service from the provider. Once the service agreement is in place, the customer becomes a subscriber to the provider.

9.2.3.3.2 Subscription Id

A key, uniquely identifying the agreements established by the provider. It is set by the provider.

9.2.3.3.3 Set of SLS

The set of SLSes, that is, the connectivity legs composing the QoS-based service. Each SLS should be correct and valid instances of the SLS-T, as specified in section 9.2.3.1.

9.2.3.3.4 Invocation Means

The *Invocation Means* clause describes the procedures and related information for invoking the service.

A service can be invoked either *implicitly*, directly as a result of the establishment of the respective agreement, or *explicitly* based on a well-defined signalling protocol e.g. RSVP, SIP or the PCP protocol specified by MESCAL for requesting the establishment of inter-domain QoS MPLS-based tunnels (LSPs).

We further distinguish two types of explicitly invoked services, which need to be supported by suitable signalling protocols: *on-demand* and *partially*. On-demand service invocation denotes a request for using the service as a whole, whereas partial service invocation denotes a request for using certain resources/characteristics associated with the service e.g. bandwidth, number of QoS tunnels; obviously, within the constraints of the overall resources/characteristics agreed in the SLS comprising the service.

Partial invocation is particularly useful for managed bandwidth services, allowing customers to dynamically request that portion of service bandwidth, which they happen to require. In the context of inter-domain agreements, such an invocation method may also be useful, as it would facilitate a more accurate and effective provisioning of the required inter-domain resources (cf. the dynamic pSLS establishment functionality specified in the MESCAL solution). Furthermore, this type of invocation particular suites to the pSLS for establishing QoS MPLS-based tunnels (LSPs); in this case, the service resources are the LSPs. Partial invocation could be alternatively carried out through a sequence of modifications of established service agreements, which obviously presents a burden both to providers and customers.

9.2.3.3.5 User Info

The *User Info* clause includes the required information for uniquely identifying the users of the subscriber who are entitled to invoke the service e.g. user id, password. Obviously, this clause should be specified only in the case of explicitly invoked services.

9.2.3.3.6 Grade of Service

The *Grade of Service* clause describes the guarantees for getting through service invocations.

Copyright © MESCAL Consortium, June 2005

The specification of such guarantees depends on the nature of the service, the invocation type (on-demand, partial) and the capabilities/policies of the provider. Generally speaking, guarantees for on-demand invoked services could be described in terms of the following parameters: minimum number of simultaneous sessions and acceptance percentage beyond that minimum number. Guarantees for partially-invoked services could be described in terms of a set of confidence levels for using a specific percentage of the totally agreed service resources – as in the SLSes. Get through guarantees could be given in quantitative or qualitative terms.

Obviously, this clause should be specified only in the case of explicitly invoked services.

9.2.3.3.7 Activation Info

The term *activation* denotes the appropriate configurations and provisions that need to be undertaken in the provider domain for making the service available to the customer so that its users can use the service. Service activation is an internal process.

Certain services require that they are activated by the provider in a particular way, so that the (users of the) customer can use them. Examples of such services include: MPLS VPNs offered to end-customers and services for transiting QoS traffic that require the exchange of q-BGP messages as in the MESCAL solution. Therefore, the way (method, not how) that the service will be activated may be an essential aspect of service offering and a subject for agreement.

The *Activation Info* clause includes the required information for describing the method according to which the service is to be activated -made available to the customer for use. It includes the following attributes:

- *Activation Method*, describing the particular method according to which, the service should be made available to the customer for use. It could be specified in terms of a set of URLs or protocols e.g. BGP, q-BGP or possible technologies e.g. MPLS VPN and related parameters.
- *Activation Verification Procedures*, describing the procedures necessary to be undertaken for ensuring that the service has been activated as specified above. Its specification is outside the scope of MESCAL investigation, as it relates to the issue of service assurance.

If this clause is specified, the agreement is considered to be in effect according to the agreed *Schedule* only after the successful undertaking of the involved *Activation Verification Procedures*.

9.2.3.3.8 Schedule

The *Schedule* clause describes the time period during which the service should be made available to the customer, in other words the time constraints for using the service. It includes the following attributes:

- Start Time
- Termination Time
- *Hours*, specifying a range of hours of the specified *Days* of the *Months* during which, the service should be made available to the customer.
- *Days*, specifying a range of days of the specified *Months* during the specified *Hours* of which, the service should be made available to the customer.
- *Months*, specifying a range of months during the specified *Hours* of the *Days* of which, the service should be made available to the customer.

The specification of the exact semantics of the *Start Time* and *Termination Time* or the need for other similar attributes is for further study. The benefit of such information to service agreement management and inter-domain traffic engineering needs to be investigated thoroughly (e.g. could be beneficial in that agreement flapping could be avoided), while the implication on service negotiations and activation needs to be assessed.

Copyright © MESCAL Consortium, June 2005

9.2.3.3.9 Availability Guarantees

The *Availability Guarantees* clause includes a single attribute denoting the probability of the service to be made available to the customer as required according to the agreed terms and conditions –in the SSS. It may be specified quantitatively as a percentage or qualitatively e.g. high, medium, low.

9.2.3.3.10 Reliability Guarantees

The *Reliability Guarantees* clause describes guarantees for reliably providing the service during its lifetime. Its specification is outside the scope of MESCAL investigation.

9.2.3.4 SSS Template XML Modelling

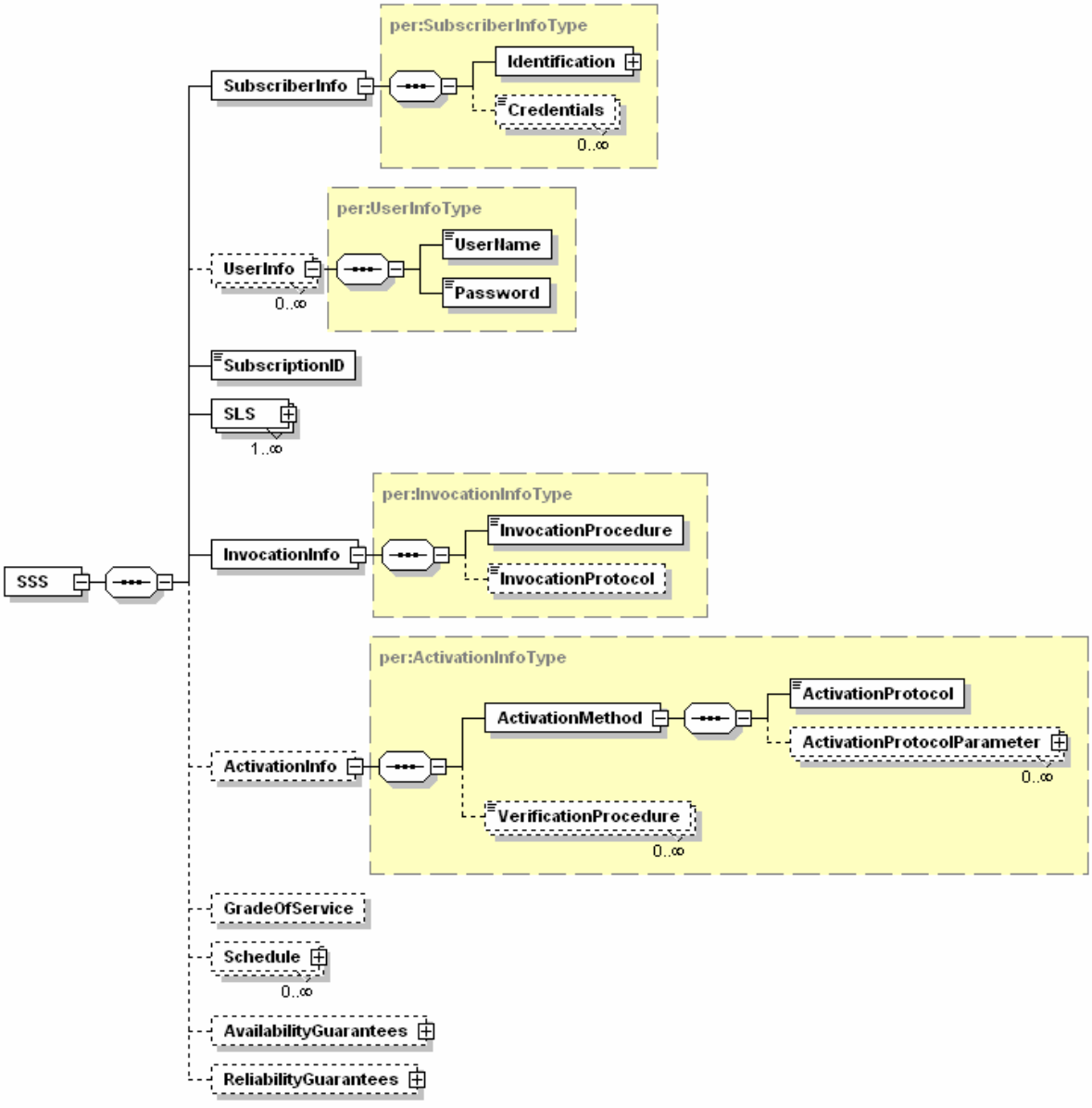


Figure 53. SSS-T XML Schema

Copyright © MESCAL Consortium, June 2005

Element	Description
SubscriberInfo.Identification	Identity of the subscriber (see section 9.2.3.3.1). Attributes: LegalName, ShortName, VATNumber, AddressInfo
SubscriberInfo.Credentials	Credentials (e.g. e-signature) of the subscriber.
UserInfo.UserName	Identifier of a user entitled to invoke the service (see section 9.2.3.3.5).
UserInfo.Password	Password of a user entitled to invoke the service.
SubscriptionID	The subscription identification key (see section 9.2.3.3.2).
SLS	The 'connectivity legs' composing the service (see section).
InvocationInfo.InvocationProcedure	The procedure to invoke the service, *Implicit*, *On-Demand* or *Partial* (see section 9.2.3.2).
InvocationInfo.InvocationProtocol	The protocol for invoking the service (*PCP*, *SIP*, *RSVP* etc.).
ActivationInfo.ActivationMethod	The method to make the service available to the customer (see section 9.2.3.3.7), specified by the ActivationProtocol (*q-BGP* etc.) and related ActivationProtocolParameters.
ActivationInfo.VerificationProcedure	A procedure undertaken to verify the successful activation of the service.
GradeOfService	Get-through service guarantees (see section 9.2.3.3.6). For on-demand invoked services it is extended to GradeOfOnDemandService with: Attributes: MinimumSimultaneousSessions, AcceptancePercentage (for further sessions), MaximumSimultaneousSessions For partially invoked services it is extended to GradeOfPartialService with: Attributes: Set of: BandwidthPercentage (of the bandwidth specified in the SLSes), AcceptancePercentage
Schedule	The time period during which the SLS should be made available to the customer (see section 9.2.3.3.8). Attributes: Set of: StartHour, EndHour, StartDay, EndDay, StartMonth, EndMonth, Year
AvailabilityGuarantees	Guarantees for the service to be available when requested for the first time (see section 9.2.3.3.9). Attributes: AvailabilityPercentage
ReliabilityGuarantees	Guarantees on the reliability for provisioning the service during its life-time (see section 9.2.3.3.10). Attributes: MaximumDownTimes (per year), MaximumRepairTime

Table 5. SSS-T XML Elements

9.2.3.4.1 Logical Validation Rules

SSST-L1: *SubscriptionID* must be unique amongst all subscriptions (of all subscribers).

SSST-L2: *SLSID* of the SLS-T (see section 9.2.3.1) must be unique amongst the SLSes for the subscription they are part of.

SSST-L3: *UserInfo* may be specified when *InvocationProcedure* is specified to *On-Demand* or *Partial* and not in other cases.

SSST-L4: *InvocationProtocol* must be specified when *InvocationProcedure* is specified to *On-Demand* or *Partial* and not in other cases.

SSST-L5: When specified, *InvocationProtocol* must be compatible with the protocols supported by the network for the specified *InvocationProcedure*.

Copyright © MESCAL Consortium, June 2005

SSST-L6: When specified, *ActivationProtocol* and associated *ActivationProtocolParameters* must be compatible with the capabilities of the network.[1]

SSST-L7: *GradeOfOnDemandService* must be specified when *InvocationProcedure* is specified to *On-Demand* and not in other cases.

SSST-L8: *GradeOfPartialService* may be specified when *InvocationProcedure* is specified to *Partial* and not in other cases.

SSST-L9: If *InvocationProcedure* is not *Implicit*, then the following must hold: (a) *UserInfo* must be unique across subscriptions of other subscribers, and (b) if the invocation protocol does not convey reference information to the SSS-T instance the invocation is related to (e.g. SubscriptionId)), then *UserInfo* must be unique across subscriptions of the same subscriber as well.

9.2.3.4.2 Interpretation Rules

In a valid SSS-T XML document:

SSST-I1: If *AvailabilityGuarantees.AvailabilityPercentage* is not specified then the minimum configured availability percentage is assigned to the service.

SSST-I2: When on the first SLS the *FlowID.SourceIPAddress* is *dynamic* and the *Scope.Ingress.ShortName* is set to a) *modem*, the access means are set to *mobile* denoting dialup access from any dialup server, b) *network server* or *internet gateway*, the access means are set to dialup denoting access from a specific network server or internet gateway. In all other cases, access means are set to *leased line*.

SSST-I3: If *Schedule* is not specified, it is assumed that the service should be made available all day every day.

SSST-I4: If *GradeOfPartialService* is not specified, the following values are assumed: *BandwidthPercentage*→100%, *AcceptancePercentage*→100%.

9.2.4 pSLS Models

In this section we present suitable models for the different types of pSLS identified by MESCAL (see section 9.2.2) according to the different types of business relationships between providers.

While the SLS-T and SSS-T templates, as specified in the previous sections, present an open, detailed model for describing the technical aspects (from connectivity perspectives) of general QoS-based services, including pSLS, there is need for more condensed, summarised pSLS models, for a number of reasons:

- pSLS need to reflect the specific context of the particular business relationship; not to be expressed in general terms, which might create ambiguity and confusion.
- The service fill-in (subscription) process at the abstraction of SSS-T might be tedious from customer perspectives, because SSS-T specification entails the specification of a number of attributes.
- There are a number of engineering incompatibilities between the values of the attributes of an SSS-T that the customers might not be aware. Therefore, customers might get frustrated as the provider would turn down their service requests, simply because they were not formed correctly.

To the above end, we increase the abstraction level of the SSS-T attributes (as appropriate to the pSLS context), by introducing the so-called *group-alias attributes*. A group-alias attribute is strictly associated with some SSS-T attributes, providing 'alias' for these SSS-T attributes. By definition, there

[1] For q-BGP in particular, the activation protocol parameters, i.e. the QoS characteristics conveyed by q-BGP, must be validated against the q-BGP capabilities supported per QoS-class according to service planning. The QoS-class is deduced by the performance guarantees clause in the SLS (see section 9.2.3.1.6).

Copyright © MESCAL Consortium, June 2005

must be a one-to-one mapping between the values of a group-alias attribute and the values of the SSS-T attributes that is used to alias. Evidently, through the alias mapping function, group-alias attributes can be used to eliminate invalid combinations of SSS-T attribute values. Group-alias attributes may be complex, in the sense of containing other group-alias attributes.

The XML schema of the types of pSLS identified by MESCAL (see section 9.2.2) is presented in Figure 54– Figure 56. Table 6 describes group-alias attributes included in the pSLSes. Note that the *Subscriber Info*, *Subscription ID*, *Schedule*, *Availability Guarantees* and *Reliability Guarantees* attributes are common to both pSLS and SSS-T models. Other SSS-T attributes such as *User Info*, *Invocation Info*, *Activation Info* and *Grade of Service* are included in the pSLSes or filled with default values depending on the service characteristics. The translation of the pSLS models to SSS-T and SLS-T attributes is outlined in Table 6.

Group-Alias Attribute	Description
Interconnection Point	It aliases a *Scope Site* (see section 9.2.3.2) which must indicate a unique network edge interface, used to identify the boundary link over which the pSLS is to be provided. The site used as interconnection point must be a customer site, identified either by its *Postal Address*, or the *IP Prefix(es)* of the customer's subnetwork, or the *IP Address* of the attached interface.
Destination Nets	It aliases the other end of *Scope*; *Interconnection Point* being the *Ingress* point, *Destination Nets* will be the *Egress* points and vice-versa. The *Destination Nets* is identical to an *Ingress/Egress* attribute; it may be any combination of *Scope Sites* of any form.
(Upstream/Downstream) Connectivity	It provides alias to *Traffic Conformance*, *Excess Treatment* and *Performance Guarantees* SLS-T attributes. *Connectivity* can be quantitative or based on meta-QoS classes (see section 4.2.3). A detailed description of the *Connectivity* group-alias attribute is provided in section 9.2.4.1.
Class Flow ID	In combination with *Connectivity* it provides alias to *Flow ID*. The *Class Flow ID* blocks the use of multi-field (MF) filters for flow classification. It corresponds to the *Flow ID* narrowed to allow only the configuration of the *Class Info* element.

Table 6. pSLS Group-Alias Attributes

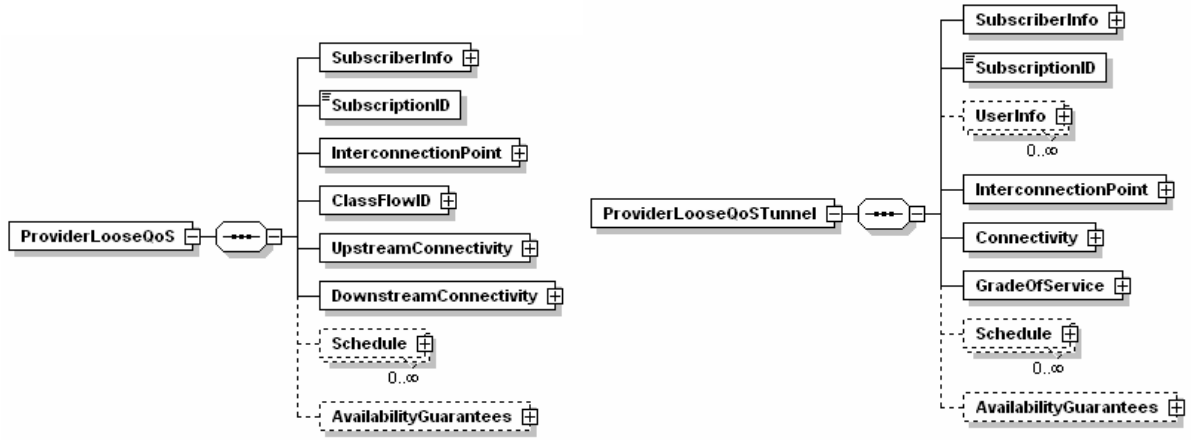


Figure 54. Provider Loose QoS pSLSes XML Schema

Copyright © MESCAL Consortium, June 2005

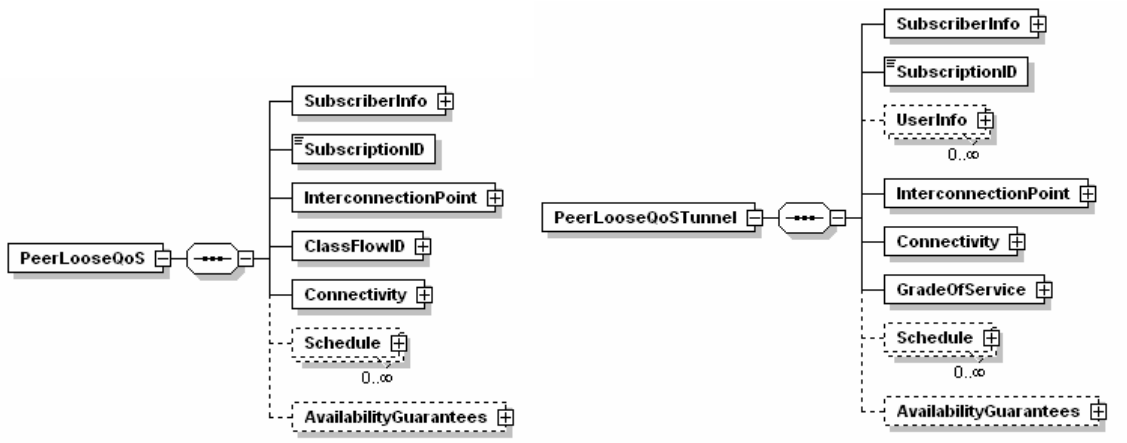


Figure 55. Peer Loose QoS pSLSes XML Schema

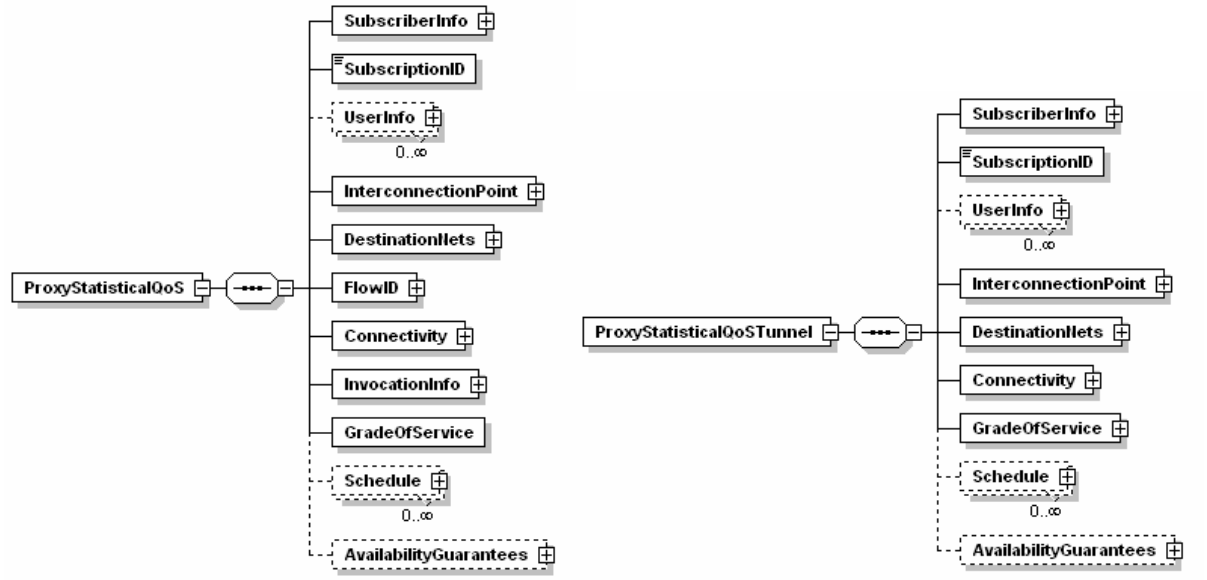


Figure 56. Proxy Statistical QoS pSLSes XML Schema

Copyright © MESCAL Consortium, June 2005

	SSS-T					SLS-T					
	User Info	**Invocation Info**	**Activation Info**	**Grade of Service**	**SLSes (requested, offered)**	**Scope Ingress**	**Scope Egress**	**Flow ID**	**Traffic Conformance**	**Excess Treatment**	**Performance Guarantees**
ProviderLooseQoS		implicit	q-BGP, (ConveyParameter, QoSAllSupported)		req:2	[Interconnection Point]		[Class FlowID]	→[Meta Class Connectivity]		
ProviderLooseQoSTunnel	[*]	partial, PCP	q-BGP, (ConveyParameter, QoSAllSupported), (ConveyParameter, PCSid)	[partial.*]	req:2	[Interconnection Point]					→ [Meta Class Connectivity]
PeerLooseQoS		implicit	q-BGP, (ConveyParameter, QoSAllSupported)		req:1 off:1	[Interconnection Point]		[Class FlowID]	→[Meta Class Connectivity]		
PeerLooseQoSTunnel	[*]	partial, PCP	q-BGP, (ConveyParameter, QoSAllSupported), (ConveyParameter, PCSid)	[partial.*]	req:1 off:1	[Interconnection Point]					→ [Meta Class Connectivity]
ProxyStatisticalQoS	[*]	[*]	q-BGP, (ConveyParameter, QoSAllSupported)	[*]	req:1	[Interconnection Point]	[Destination Nets]	[*]	→[Quantitative Connectivity]		
ProxyStatisticalQoSTunnel	[*]	partial, PCP	q-BGP, (ConveyParameter, QoSAllSupported), (ConveyParameter, PCSid)	[partial.*]	req:1	[Interconnection Point]	[Destination Nets]				→ [Quantitative Connectivity]

Table 7. Translation of pSLS Models to SSS-T and SLS-T

Copyright © MESCAL Consortium, June 2005

9.2.4.1 Connectivity Group-Alias Attribute

The *Connectivity* group-alias attribute is common to all pSLSes and is used to alias the *Traffic Conformance*, *Excess Treatment* and *Performance Guarantees* SLS-T attributes (see section 9.2.3.1). Hence, it abstracts the configuration of the packet transfer characteristics offered by the service and the profile the customer traffic must adhere to for being granted these characteristics. This way, only the supported combinations and value ranges are exported, the complexity of the configuration is reduced and the customer is facilitated.

There are two types of connectivity supported in the pSLSes specified in Mescal, connectivity specified through quantitative values and connectivity based on the specified meta-QoS classes (see section 4.2.3). The *Connectivity* group-alias XML schema is presented in Figure 57. Table 8shows the translation of the specified *Connectivity* types and values to the corresponding SLS-T attributes.

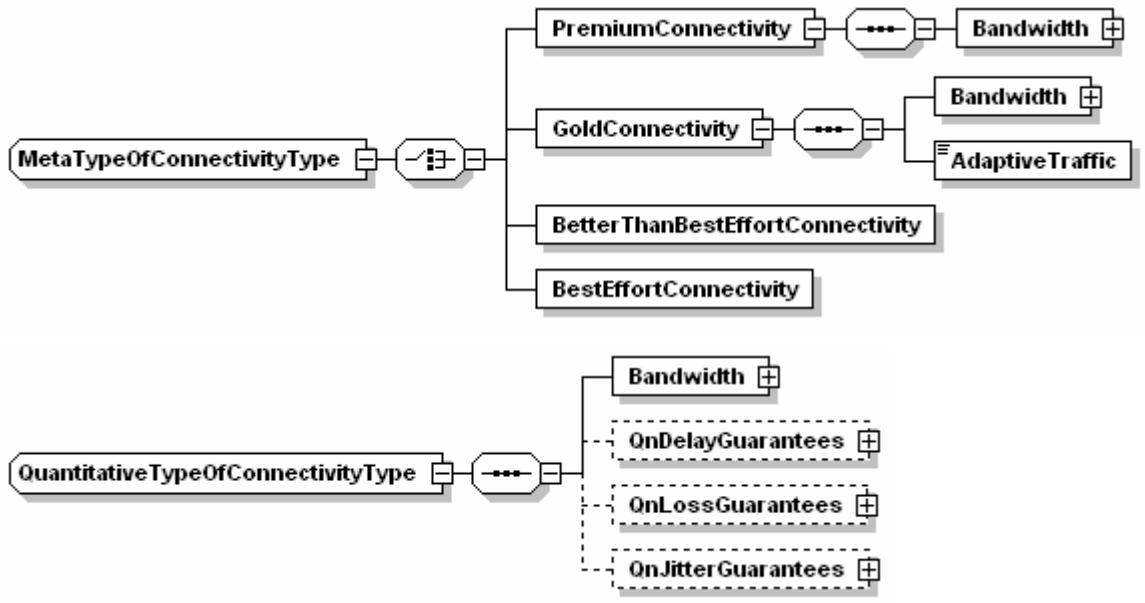


Figure 57. Connectivity XML Schema

	Flow ID Application Info	Traffic Conformance	Excess Treatment	Performance Guarantees
Premium		Token Bucket, [Bandwidth]	Shape, [Bandwidth]	TC Level Qualitative Guarantees Ql Delay = premium Ql Loss = premium
Gold	[Adaptive]	Token Bucket, [Bandwidth]	Remark	TC Level Qualitative Guarantees Ql Delay = gold Ql Loss = gold
Better than Best Effort				Overall Guarantees Ql Delay = better-than-BE QlLoss = better-than-BE
Best Effort				
Quantitative		Token Bucket [Bandwidth]		TC Level Quantitative Guarantees [Qn Delay], [Qn Loss], [Qn Jitter]

Table 8. Translation of Connectivity Group-Alias Attribute to SLS-T

Copyright © MESCAL Consortium, June 2005

9.3 Service Negotiation Protocol

9.3.1 Introduction

Quality of service (QoS) delivery across the Internet provides new business opportunities, but also presents new challenges. Nowadays, QoS-based services are offered on the basis of the so-called *Service Level Agreements (SLAs)*, which set the terms and conditions on behalf of both providers and customers in providing and requesting services, respectively.

It is without doubt that flexibility and automation are the 'names of the game' in QoS-based service provisioning. Providers have to be able to offer quickly new services according to market needs, while ensuring that their networks are appropriately configured to efficiently deliver the quality requirements of the services offered. At the same time, customers should be able to find out the offered QoS-based services and subsequently request/modify the level of QoS they desire according to their actual needs.

As flexibility requirements are high, SLAs are not necessarily monolithic contracts -agreed once, valid forever. Although this might hold for a number of business cases e.g. between peer wholesalers, SLAs might well be short-lived agreements e.g. SLAs between customers and a provider for pay-per-view services over the weekend. Furthermore, SLAs should be seen as 'living documents', in the sense of being modified, on a common agreement basis, according to actual customer needs and providers' availability or related policies. In line with this view of SLAs, is the widely accepted distinction between static and dynamic SLAs [BLAK98], [NIC99].

In the above scenery then, automated means for agreeing on SLAs (establishment, modification, deletion) will pave the way towards flexible and automated QoS-based service provisioning. At the same time it will constitute an important step in the evolution of the Internet itself. Compared to agreeing on SLAs in a sort of a manual fashion (e.g. through fax, e-mail, post), as it is mainly the practice of today, automated establishment of SLAs has a number of benefits to both providers and customers. For providers, it reduces operational costs, contributes to an integrated, fully automated service provisioning process and increases the level of attraction of the offered services. For customers, it increases their flexibility in requesting and accessing services by reducing the required time. Furthermore, automated means for SLA agreement opens-up new ways, and businesses, in promoting QoS-based services in the Internet. E.g. through Web-based service portals, where customers can view existing service offerings and agree on SLAs for desired services according to actual needs.

To automate the process of agreeing on SLAs new protocols are required. These protocols should enable customers and providers or peer providers to automatically negotiate between each other with the purpose to finally agree on a SLA. We call these protocols SLA negotiation protocols.

We view that there is a clear distinction between SLA negotiation protocols and QoS-signalling or reservation or QoS-enabled session control protocols (e.g. RSVP, SIBBS, SIP, H.323, PPP). Specifically, we view that SLA negotiation protocols are used for agreeing on SLAs, whereas, QoS-signalling or reservation or session control protocols are used for signalling/requesting the level of QoS that customers require, should respective SLAs with the providers have been agreed. SLA negotiation protocols operate at service subscription epochs, where customers subscribe to the desired services offered by the providers, and QoS-signalling or reservation or session control protocols operate at service invocation epochs, where the users of the customers (subscribers) call for the services to which have been subscribed. The distinction between service subscription and invocation is required mainly for AAA (authentication, authorisation and accounting) purposes i.e. for checking conformance of user service requests against agreed profiles, which is essential in SLA-based service provisioning. This view largely follows current business practices; it is also in line with the principles of the service management framework presented in [GOD02a].

In the above spirit, a protocol, the Service Negotiation Protocol (SrNP), for SLA negotiations is presented. It should be noted that SrNP is not specific to the particular contents of SLAs, nor it is specific to particular transport, policy or information exchange protocols. Furthermore, SrNP is completely decoupled from the negotiation logic -the logic, per negotiating party, for conducting

Copyright © MESCAL Consortium, June 2005

negotiations- offering to it, through clear interfaces, the necessary primitives required for enabling negotiations. These features increase protocol applicability, and as such, SrNP could be used for establishing any type of agreements (e.g. on price-lists) in a general e-commerce context.

9.3.2 Negotiation Protocol Requirements

SrNP design is driven by the following requirements. It should be noted that these requirements are drawn from our own experience and objectives, as requirements for negotiation protocols have not yet been commonly agreed.

Functional requirements

- The negotiation protocol should provide for primitives to enable the process of negotiations between two or more parties (negotiating parties). Generally speaking, the negotiation process is a process where several parties are seeking for an agreement on a number of commonly understood issues e.g. on a SLA.
- There should be a clear distinction between the primitives offered by the negotiation protocol and the negotiation logic. The term 'negotiation logic' denotes the logic according to which negotiations are conducted on behalf of each negotiating party. Negotiation logic should be specific to each negotiation party, being subject to its business policies and operational capabilities, and as such, is considered application- and domain-specific. In essence, the negotiation protocol should provide a service to the negotiation logic i.e. it should be seen as a layer based on which application-specific negotiation logic could be built.
- The negotiation protocol should not duplicate but complement the functionality of existing, widely deployed, standardised protocols. For SLA agreements (QoS negotiation), the corresponding negotiation protocols should not duplicate (aspects of) the functionality of existing QoS-signalling or reservation or session control protocols (e.g. RSVP).
- The targets of the negotiation logic for the particular negotiation process may not be static; rather they may change depending on factors outside the particular negotiation process. This is the case for example for the MESCAL pSLS Ordering function (see section 9.5) which performs collective negotiations with multiple providers to establish a set of pSLSes with minimum total cost while satisfying the collective capacity criteria set by the traffic engineering. It is therefore mandatory that the negotiation protocol allows for the negotiations to continue further beyond the first candidate agreement found satisfactory by all parties, as long as there are parties with related pending issues to resolve.
- The negotiation protocol should lead at convergent negotiation processes. Appropriate mechanisms should be provided at protocol layer, for ensuring that the negotiation process can terminate successfully or unsuccessfully in finite steps and in a reasonable time period, as deemed necessary by (the negotiation logic of) each of the negotiating parties.
- The negotiation protocol should be independent of the underlying transport and network protocols. In fact, it should be able to operate with multiple such protocols.

Non-functional requirements

- The negotiation protocol should provide for secure and reliable communication.
- The negotiation protocol should be expandable in terms of additional negotiation primitives.
- The negotiation protocol should be able to support a number of simultaneous active negotiation processes.

It is clear, that the above requirements contribute to the openness and therefore the applicability of negotiation protocols; they are not specific to SLA negotiations and they could apply to any negotiation protocol.

Copyright © MESCAL Consortium, June 2005

9.3.3 Negotiation Model

The following assumptions underline the negotiation model to which SrNP has been designed to apply.

It is assumed that the negotiation process involves *two parties* only; one acting in a server role, called the *server*, and the other acting in a client role, called the *client*. The roles are exclusive to the parties that is, a party cannot act in both roles in the context of a particular negotiation process. Following the usual distinction between client and server roles (client requests, server responds), in the context of a negotiation process, these roles are distinguished in that agreements can only be pursued by the client towards the server. This distinction is in line with the semantics underlying a customer-provider relationship between two interacting parties. Valid client-server tuples of negotiating parties could be: customer-provider (intra-domain SLA negotiations) or provider-provider (inter-domain SLA negotiations). Note that the provider-customer tuple is also considered valid. For instance, this case may arise in situations where the provider deems necessary to renegotiate SLAs with some customers for improving or lowering the quality of the subscribed services.

It is assumed that the issues under negotiation can be described in a form of a *document*. The target of the negotiation process is then for the negotiating parties to come to an *agreement regarding the content of* (information included in) the document.

Furthermore, it is assumed that all negotiating parties have a common understanding of the semantics and syntax of the information included in the document as well as means for constructing, extracting and manipulating the information in the document. In line with the requirements presented in the previous section, document/information format, construction and manipulation are not of concern to the protocol, but rather of the negotiation logic. Evidently then, SrNP is not specific to any SLA format or to the content of SLAs. It is general enough to apply to negotiating any issues, provided that these issues can be appropriately described in the form of a commonly understood document.

Finally, it is assumed that authentication and authorisation with respect to negotiation aspects are not of concern to SrNP; they are of concern to the negotiation logic that SrNP services. It is also assumed that SrNP uses the services of a reliable and secure transport protocol.

9.3.4 SrNP Overview

SrNP is an *application-layer*, *session-oriented* protocol allowing for sessions to:

- establish an agreement,
- modify an established agreement, and to
- delete an established agreement

SrNP sessions are initiated by the client.

Agreement Establishment Session
Generally speaking, the negotiation process for establishing an agreement is an iterative process, whereby the negotiating parties exchange their views/requirements on the issues under negotiation until an agreement is reached.

SrNP follows a *client-server*, *multi-dialogue-based* approach for realising the necessary interactions between the negotiating parties towards establishing an agreement. Specifically:

First, the client *connects* to the server to initiate a session for negotiating the establishment of an agreement. Within a session, the client initiates *options*. An option is an independent negotiation track allowing for pursuing a particular variation of the issues under negotiation. Agreement will be eventually established on only one of the open options.

Within each option the client issues *proposals* and the server responds by either *accepting* the proposals or by issuing *revisions*. Proposals and revisions convey the client's and the server's views/requirements on the issues under negotiation, respectively. Through revisions, the server is

Copyright © MESCAL Consortium, June 2005

enabled to respond to the client views/requirements not in a monolithic 'agree/do not agree' manner but, in a flexible, in the spirit of 'I could agree provided that/even if', manner indicating the points of argumentation and suggesting possible alternatives. It is up to the negotiation logic of the client to determine whether to adhere or not to adhere to the suggested revisions in subsequent proposals.

Accepted proposals and concrete revisions can be *cooled* by the client as an indication of provisional agreement; the option then is considered closed unless either party decides to drop it. The option can be *dropped* by the client at any time, while be the server it can be dropped only as a response to a request. However, a negotiation session may be *quitted* at any time by either party.

Eventually the client calls for an *agreement* on either a previously provisionally agreed proposal or on a new one. If the server *accepts* it the negotiation process concludes successfully, otherwise the client may call for another *agreement* until either the agreement is reached or either party decides to abort. When accepting the call for agreement the server also includes the received proposal as a form of 'hand-shaking'.

The client and the server interact in two different levels, in the session and in the option level. Messages in the session level affect the status and the behaviour of the protocol in the option level.

Within an option the client and the server interact in a dialogue (half-duplex) manner; once a party sends information to the other party, the party is blocked until a *valid response* from the other party is received. Specifically, once the client sends a proposal, it is blocked until it receives a revision, an acceptance or a rejection from the server. Similarly, once the server replies to the last proposal, it is blocked until it receives an alternative proposal or a signal of provisional agreement or a rejection from the client. Once the client issues a rejection, the option is dropped.

At the session level the client may at any time initiate options, initiating thus another parallel dialogue (multi-dialogue nature). A request for agreement occurs at the session level and results in pausing open options. In case of failure to reach the agreement, negotiation over the open options may resume as normal. Either party may quit session at any time; the protocol terminates the negotiation process from this party, without waiting any further response from the other party.

To ensure graceful operation, SrNP does not allow a party to be blocked forever, waiting to receive a valid response from the other party. To this end, when a party sends information to the other party within an option, SrNP requires that the party must specify a maximum tolerable time period willing to wait for the other party to respond back. SrNP will reject the negotiation process on behalf of the sending party, if during the specified maximum tolerable time period no valid response from the other party is received. In addition to avoiding communication blocking, this mechanism of SrNP has the intuitive counterpart of 'sent information is only valid for a specific time period'. In addition, the negotiation session is governed by a maximum idle time allowed between any actions in the options level.

SrNP also offers the negotiation features of 'last word', 'need more time' and 'need clarifications'. Specifically:

SrNP provides the means to either party for forcing the negotiations to conclude allowing for a last round. The last round can be invoked by the client by sending a last proposal. If the server replies with a definite agree or reject answer to the received proposal the negotiations conclude. The server is also entitled to reply with a last revision and the negotiations will conclude with the definite answer of the client to the last revision. If the server invokes the last round the client must reply with a definite answer, either to call for agreement on its last proposal or to abort negotiations altogether. In addition to its intuitive counter part ('last word'), this feature offers a lever for enforcing the termination of a negotiation process in finite steps.

SrNP allows either party to request more time beyond the maximum tolerable time period for providing its response to the last received position within a negotiation track. The other party may reply by accepting to grant more time, not necessarily equal to the requested time extension, or by rejecting to extend the tolerable period. This can be useful for example when the server sees that an agreement is likely to be reached shortly after the elapse of the time period specified by the client, or when the client waits for an answer to other correlative negotiation processes in order to evaluate its

Copyright © MESCAL Consortium, June 2005

margin to consent to the server position, or in the case of human interaction for negotiation decision making, etc.

Finally, SrNP provides means to ask for clarifications at the negotiation logic level. In case a proposal or a revision is not concrete with the criteria of the server's/client's negotiation logic, i.e. when important negotiation issues are not sufficiently specified, then clarifications may be requested. The other party then should reply with an appropriately modified proposal/revision.

Agreement Modification Session
During this session, SrNP operates similarly to the agreement establishment session outlined above. In this case, the first proposal to be sent by the client denotes the agreement modifications that the client wishes to make.

Agreement Deletion Session

Once the client has successfully initiated a session for deleting an already established agreement, SrNP allows for the server to respond by either accepting or rejecting the agreement deletion request. The decision for accepting or rejecting the deletion request is taken by the server negotiation logic.

9.3.5 SrNP Messages and Interface

9.3.5.1 Protocol Messages

The SrNP messages reflect the negotiation primitives offered by the protocol to the negotiation logic. SrNP messages are distinguished into *client* or *server session* or *option* level messages.

Following the multi-dialogue nature of SrNP, during a negotiation session initiated by the client for establishing/modifying/deleting an agreement, client and server messages are exchanged alternately within the context of an option (one after the other); server messages are sent in response to client messages and vice versa.

Furthermore, SrNP dictates that messages must be exchanged in a particular order, reflecting the natural evolution of a negotiation process. That is, given a message sent by a party, the other party can respond only with specific messages, which SrNP regards as *valid responses* to the message sent. Subsequently, the party, which has sent a message and received a valid response by the other party, can only send specific messages corresponding to the valid responses of the response-message received, and so on until the negotiations are terminated.

The SrNP messages are described in Table 9 and their parameters in Table 10.

Message		Description	Valid Responses
SrNP Client Messages			
session	SessionInit	It requests the initiation of a session for negotiating the establishment, modification or deletion of an agreement. It is the first message that the client must send.	SessionAccepted, Quit
	BindProposal	A proposal that signals a call for agreement to conclude the negotiations.	AgreeProposal, RejectBindProposal, Quit
	LastProposal	A proposal that forces the server to either consent to the position of the client carried by the message, quit, or state one last alternative position. LastProposal signals the start of the last negotiation round.	AgreeProposal, LastRevision, Quit

Copyright © MESCAL Consortium, June 2005

	Quit	It indicates that the client is not satisfied by the responses of the server and does not wish to pursue a better deal any further and, as such, is not willing to continue the negotiations. When reliably delivered to the server, the protocol terminates at both ends concluding unsuccessfully the negotiations.	
option	Proposal	It carries the client's requirements/views on the issues under negotiation, described in a document, which must be commonly understood by the negotiating parties. When used with a new option identifier it initiates a new option. This message is exchanged during negotiations for establishing or modifying an agreement. The carried document is constructed by the client's negotiation logic and the semantics and syntax of the information included in it are transparent to the protocol. This -or the LastProposal message- is the first message that the client must send after the negotiation session has been established.	Revision, LastRevision, Accept, Reject, Clarify, Time, Quit
	Cool	It indicates that the client wishes to establish a provisional agreement on the last server position - either an accepted client proposal or a concrete server revision. A provisional agreement notifies the server that its last position is a satisfactory candidate to finally agree upon sometime later, depending on the availability of other options in this or other correlated sessions with this or with other servers. With this message the option is considered closed; it can only be dropped allowing for no other position exchanging over it.	Reject, Quit
	ForgetIt	It indicates that the option is dropped by the client, either as a result of non satisfactory responses from the server or after a prior provisional agreement because the client now is offered other better options.	Quit
	Time	It requests an extension of the time allotted to the client for responding in the last message from the server.	TimeGranted, Reject, Quit
	TimeGranted	It signals that a previously requested time extension is granted to the server for responding in the last client proposal.	Revision, LastRevision, Accept, Reject, Quit
	Clarify	It requests clarifications at the negotiation logic level on the lastly received document.	Revision, LastRevision, Reject, Quit
SrNP Server Messages			
session	SessionAccepted	It confirms the client's request to initiate a negotiation session for agreement establishment, modification or deletion (cf. SessionInit message).	Proposal, BindProposal, LastProposal, Quit
	AgreeProposal	It indicates that the server consents to the last received proposal on the issues under negotiations (cf. BindProposal/LastProposal messages) and an agreement is reached. The message should carry the last received document by the client as a form of 'hand-shaking'.	
	RejectBindProposal	It indicates that the server rejects the call for agreement received by the client. However, further negotiations may proceed as normal.	Proposal, BindProposal, LastProposal, Quit

Copyright © MESCAL Consortium, June 2005

	LastRevision	A revision that forces the client to either call for agreement or quit negotiations. LastRevision signals the start of the last negotiation round.	BindProposal, Quit
	Quit	It indicates that the server does not wish to continue negotiations any further. When reliably delivered to the client, the protocol terminates at both ends concluding unsuccessfully the negotiations.	
option	Revision	It carries the server's counter-requirements/views on the issues under negotiation, should the server cannot accept (some of) the respective client's requirements/views as last received (cf. Proposal message). Server's counter-requirements/views are described in a document, constructed by the server's negotiation logic, which must be commonly understood by the negotiating parties. The semantics and syntax of the information included in the document are transparent to the protocol.	Proposal, BindProposal, LastProposal, Cool, ForgetIt, Time, Clarify, Quit
	Accept	It indicates that the server accepts the last received client's requirements/views on the issues under negotiations (cf. Proposal/LastProposal messages).	Proposal, BindProposal, LastProposal, Cool, ForgetIt, Time, Quit
	Reject	It indicates that the server rejects the last received client's requirements/views on the issues under negotiations (cf. Proposal message) entirely without presenting counter-requirements/views.	Quit
	Time	It requests an extension of the time allotted to the server for responding in the last message from the client.	TimeGranted, ForgetIt, Quit
	TimeGranted	It signals that a previously requested time extension is granted to the client for responding in the last server position.	Proposal, BindProposal, LastProposal, Cool, ForgetIt, Quit
	Clarify	It requests clarifications at the negotiation logic level on the lastly received document.	Proposal, ForgetIt, Quit

Table 9. SrNP protocol messages

SrNP Header Parameters – common to all messages	
SessionId	Unique identifier of the negotiation session. It is a compound identifier composed by a unique identifier assigned by the client SrNP engine and a unique identifier assigned by the server SrNP engine.
MessageId	Unique identifier of the sent messages in the locality of a party. It is set by the SrNP engine.
InResponseTo	The MessageId of the message to which this message is sent as a response. By examining this parameter a party is able to determine whether a received valid response is correct i.e. it truly corresponds to the last message sent by this party. It is set by the SrNP engine.
TimeToRespond	The maximum time period, in seconds, that a party sending a message is willing to wait for the other party to respond. If this period elapses, the protocol terminates. It is specified by the negotiation logic.
OptionId	Unique identifier of the option to which the message refers to. It may not be present to some session level messages. It is specified by the SrNP engine.

Copyright © MESCAL Consortium, June 2005

Message	Time ToRespond	OptionId	Parameter	Description
SrNP Client Message Parameters				
SessionInit	√		SessionType	The type of the negotiation session the client wishes to initiate. It takes three possible values: newAgreement, modifyAgreement, deleteAgreement.
			ConctactAddr	The client contact address where the server will send subsequent SrNP replies.
			MaxIdleTime	The maximum idle time, in seconds, allowed between any actions from the client in the options level. If this period elapses, the protocol terminates.
			AgreementId	In cases of sessions for modification/deletion of established agreements, it is the unique identifier of the agreement to be modified or deleted as set by the server (cf. AgreedProposal message). In cases of sessions for establishing a new agreement, it is left unspecified.
BindProposal	√	√	Document	The document describing the client's requirements/views on the issues under negotiations. It is constructed by the client's negotiation logic.
LastProposal	√	√	Document	As in BindProposal.
Quit			Reason	The reason for terminating the negotiation process. It is specified by the client's negotiation logic.
Proposal	√	√	Document	As in BindProposal.
Cool	√	√		
ForgetIt		√	Reason	The reason for dropping the option. It is specified by the client's negotiation logic.
Time	√	√	TimeRequest	The additional time requested by the client for responding in the last message received by the server.
TimeGranted	√	√	TimeGranted	The additional time granted by the client to the last time extension request sent by the server.
Clarify	√	√	Document	The document requesting clarifications on the previously received server's requirements/views.
SrNP Server Message Parameters				
SessionAccepted	√			
AgreeProposal		√	Document	The last document received with a call for agreement describing the client's requirements/views on the issues under negotiation (carried by a BindProposal/LastProposal message).
			AgreementId	A unique identifier of an agreement It is specified by the server's negotiation logic.
RejectBindProposal	√	√	Reason	The reason for rejecting the last call for agreement issued by the client. It is specified by the server's negotiation logic.

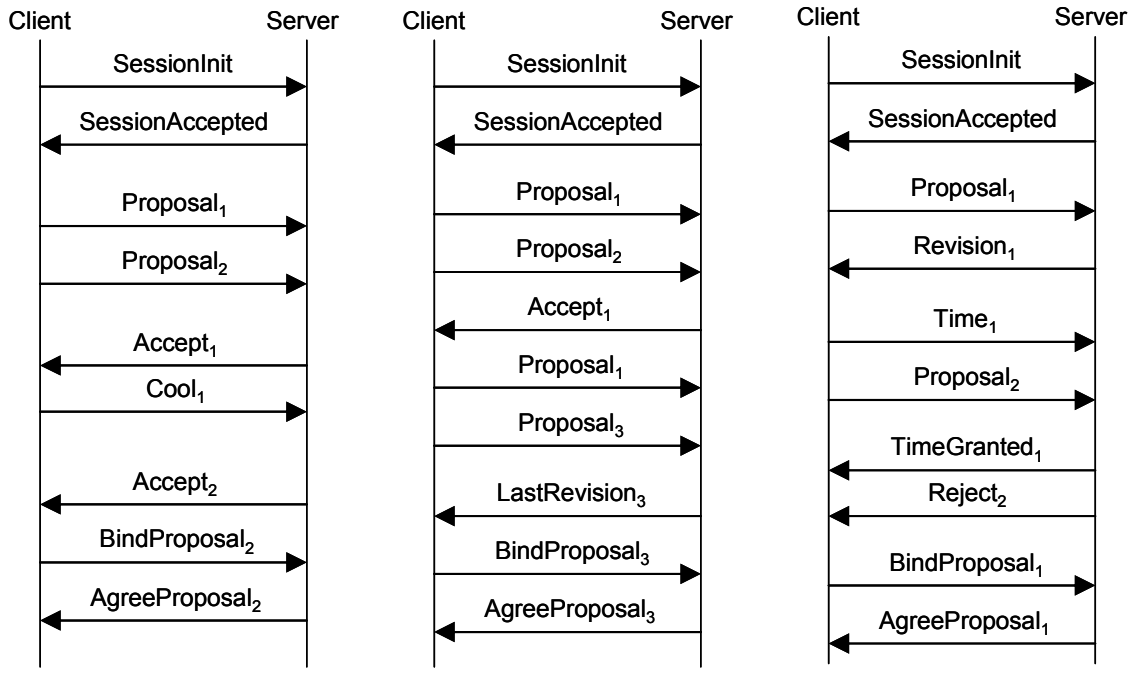
Copyright © MESCAL Consortium, June 2005

LastRevision	√	√	Document	It describes the server's counter-requirements/views on the issues under negotiation based on the respective client's requirements/views, as last received by a Proposal/LastProposal message.
Quit			Reason	The reason for terminating the negotiation process. It is specified by the server's negotiation logic.
Revision	√	√	Document	As in LastRevision.
Accept	√	√		
Reject		√	Reason	The reason for dropping the option. It is specified by the server's negotiation logic.
Time	√	√	TimeRequest	The additional time requested by the server for responding in the last message received by the client.
TimeGranted	√	√	TimeGranted	The additional time granted by the server to the last time extension request sent by the client.
Clarify	√	√	Document	The document requesting clarifications on the previously received client's requirements/views.

Table 10. Parameters of the SrNP protocol messages

9.3.5.2 Message Sequence Charts (MSCs)

Figure 58 and Figure 59 depict the exchange of SrNP protocol messages in a number of typical negotiation processes initiated for agreement establishment or modification.



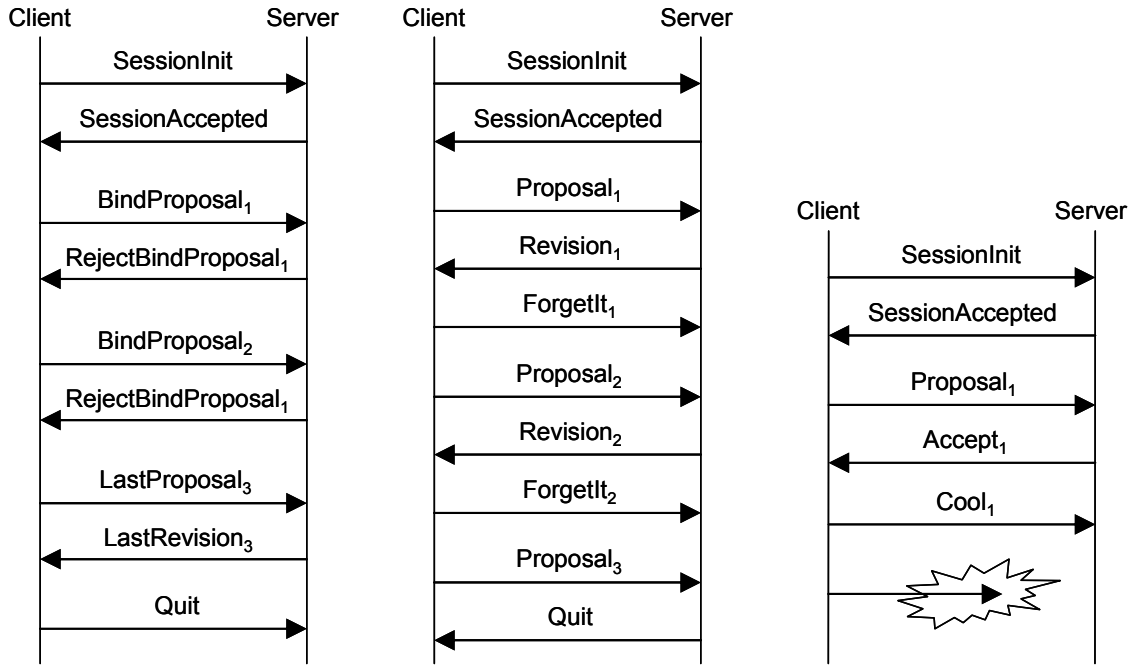
(a) Client explores different options before calling for agreement

(b) Server forces the client to conclude negotiations

(c) Client is granted extra time to explore more options

Figure 58. MSCs of successful negotiations

Copyright © MESCAL Consortium, June 2005



(a) Client aborts negotiations after several failed attempts to establish agreement

(b) Server aborts negotiations without trying last round

(c) Transport layer failure, timers expire at boths sides and the protocol terminates

Figure 59. MSCs of unsuccessful negotiations

9.3.5.3 Interface Messages

The interface that SrNP offers to applications realising negotiation logic is described in a technology-independent manner through a set of messages depicted in Table 11.

SrNP interface messages are self-explained and mainly correspond to the SrNP protocol messages presented previously. The prefix "Send" denotes the 'pull'-part of the SrNP interface allowing the application to send a protocol message to the other party, whereas the prefix "Forward" denotes the 'push'-part of the SrNP interface notifying the application of a protocol message received from the other party.

The `ForwardException` message notifies the application on abnormal protocol termination. This can occur (a) when the maximum tolerable time period that a party has specified to wait for the other party to respond elapses without receiving such a response, (b) on transport service failures and (c) on unexpected application behaviour. The latter occurs when the negotiation logic of a party responds to the last message received from the other party with a not valid response message or with multiple valid response messages.

SrNP Client Interface Message	SrNP Server Interface Message
`SendSessionInit`	`SendSessionAccepted`
`SendBindProposal`	`SendAgreeProposal`
`SendLastProposal`	`SendRejectBindProposal`
`SendQuit`	`SendLastRevision`
`SendNewProposal`	`SendQuit`
`SendProposal`	`SendRevision`
`SendCool`	`SendAccept`
`SendForgetIt`	`SendReject`
`SendTime`	`SendTime`

Copyright © MESCAL Consortium, June 2005

SendTimeGranted	SendTimeGranted
SendClarify	SendClarify
ForwardSessionAccepted	ForwardSessionInit
ForwardAgreeProposal	ForwardBindProposal
ForwardRejectBindProposal	ForwardLastProposal
ForwardLastRevision	ForwardQuit
ForwardQuit	ForwardProposal
ForwardRevision	ForwardCool
ForwardAccept	ForwardForgetIt
ForwardReject	ForwardTime
ForwardTime	ForwardTimeGranted
ForwardTimeGranted	ForwardClarify
ForwardClarify	ForwardProtocolException
ForwardProtocolException	

Table 11. SrNP interface messages

9.3.6 SrNP Finite State Machine

We distinguish between client and server, and session level and option level SrNP FSMs. In the context of a particular negotiation process, there is a single instance of a client and server session FSM at client and server sides respectively. As such, at a client side there will be as many client FSMs as the negotiation processes initiated by the client, and at a server side there will be as many server FSMs as the negotiation processes of the connected clients. Associated with each session FSM there are as many option FSMs as the number of open options of the particular session.

9.3.6.1 Timers

- *OptionTimer*: SrNP updates this timer whenever a negotiating party sends or receives a message to/from the other party. Its value is the maximum time period the party sending the message can possibly wait for the other party to respond, determined by its negotiation logic and carried in the TimteToRespond field of the SrNP message header (see Table 10) The timer expires if no *valid* (cf. rightmost column of Table 9) and *correct* (cf. MessageId header parameter, Table 10) message from the other party is received or sent as a response from/to the other party. At this point, SrNP terminates the negotiation process. Evidently, by utilising this mechanism SrNP avoids communication blocking, while ensures that the negotiation process will terminate in finite steps and time.
- *SessionTimer*: SrNP updates this timer each time an option message is received or sent and sets its expiration time to the maximum of the running OptionTimer expiration times, plus the MaxIdleTime parameter of the SessionInit message (see Table 10). Upon expiration the negotiation process is terminated at both sides SrNP. This way SessionTimer ensures that the negotiation process will terminate in case of lack of activity in the option level due to either communication layer failure or the client's negotiation logic obstruction.

9.3.6.2 Events

Considering a particular negotiating party (client or server), in addition to the protocol and interface messages (see Table 9 and Table 11), the following events are also considered by the SrNP FSMs:

- *TimerExpired*: It is fired whenever the OptionTimer or the SessionTimer expire.
- *TransportError*: It is fired whenever SrNP is notified by the underlying transport services of failures in communicating with the other party.

Copyright © MESCAL Consortium, June 2005

The above events are considered protocol operation exceptions and are encapsulated into the ForwardProtocolException protocol interface message (see Table 11).

In addition, internal events generated by the client session level to drive the option level FSMs are:

- *Freeze*: It is fired when a BindProposal message is sent and results in freezing the option FSMs to only consider valid state transitions due to message received from the server. In other words it prohibits any activity from the client negotiation logic once the latter has sent a BindProposal and until it receives a valid response or a timer expires.
- *Continue*: It is fired when a valid response to a previous BindProposal message is received and allows the open option FSMs to return to normal operation.

9.3.6.3 SrNP Client FSM

The client session FSM includes the following states:

- *Initialisation (INIT):* In this state the protocol performs the necessary initialisation communication between the negotiation parties before proceeding to the actual negotiations.
- *Idle (IDLE):* In this state the protocol is active only in the option level and the control is to the corresponding option FSMs.
- *Bound Wait for Peer (BWFP):* In this state the protocol waits for a definite answer from the server to a previously sent call for agreement and blocks the negotiation logic from sending any other message.
- *Last Wait for Peer (LWFP):* In this state the protocol has entered the last negotiation round and it's the server's turn to either give a definite answer or provide a last revision.
- *Last Wait for Negotiation Logic (LWFNL):* In this state the protocol has entered the last negotiation round and it's the client's turn to make a call for agreement or quit the negotiations.
- *Last Bound Wait for Peer (LBWFP):* In this state the protocol has entered the last negotiation round and it's the server's turn to either consent to the call for agreement or quit the negotiations.

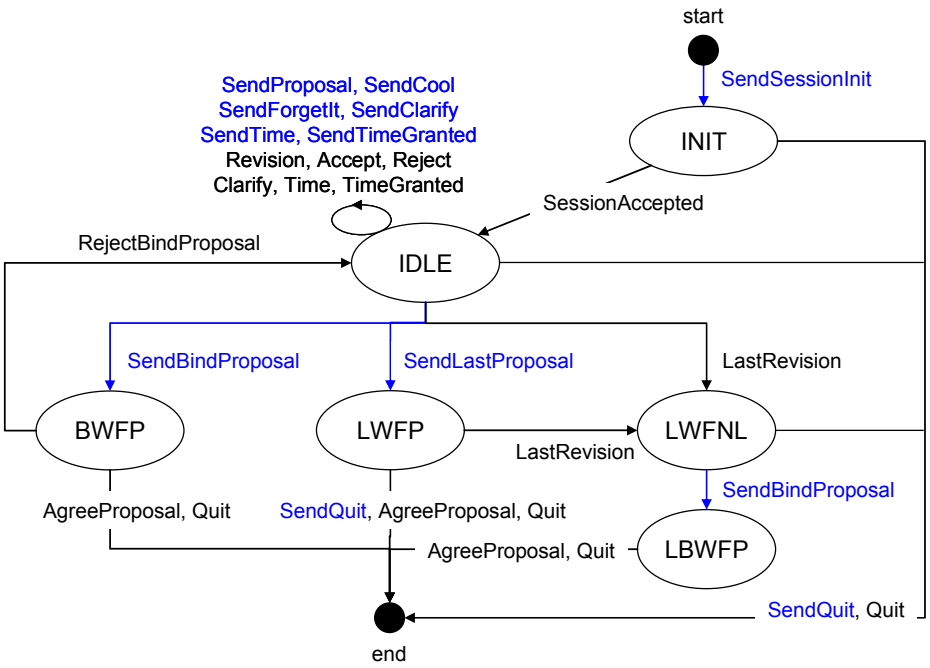


Figure 60. The SrNP client FSM session level state transition diagram

Copyright © MESCAL Consortium, June 2005

State	Event	Actions	Next State
INIT	SessionAccepted	If message is on time then forward message, update SessionTimer, update state; otherwise forward protocol exception and abort session.	IDLE
	SendQuit, Quit	Send/forward message and abort session.	
IDLE	SendProposal, SendCool, SendForgetIt, SendClarify, SendTime, SendTimeGranted, Revision, Accept, Reject, Clarify, Time, TimeGranted	Pass control to the corresponding option FSM. If message is deemed on time and valid then update SessionTimer (see section 9.3.6.1).	IDLE
	SendBindProposal	If message is on time then freeze other options, send message, update OptionTimer, update SessionTimer, update state; otherwise forward protocol exception and abort session.	BWFP
	SendLastProposal	If message is on time then drop all other options, send message, stop SessionTimer, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LWFP
	LastRevision	If message is on time then drop all other options, forward message, stop SessionTimer, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LWFNL
	SendQuit, Quit	Send/forward message and abort session.	
BWFP	AgreeProposal	If message is on time then forward message; otherwise forward protocol exception. Abort session.	
	RejectBindProposal	If message is on time then drop option, resume other options, forward message, update SessionTimer, update state; otherwise forward protocol exception and abort session.	IDLE
	Quit	Forward message and abort session.	
LWFP	LastRevision	If message is on time then drop all other options, forward message, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LWFNL
	AgreeProposal	If message is on time then forward message; otherwise forward protocol exception. Abort session.	
	SendQuit, Quit	Send/forward message and abort session.	
LWFNL	SendBindProposal	If message is on time then drop other options, send message, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LBWFP
	SendQuit, Quit	Send/forward message and abort session.	
LBWFP	AgreeProposal	If message is on time then forward message; otherwise forward protocol exception. Abort session.	
	Quit	Forward message and abort session.	

Table 12: The SrNP client session FSM state transition table

The client option FSM includes the following states:

- *Wait for Peer (WFP):* In this state the protocol waits to receive from the server a response to the last sent message determined by the client's negotiation logic for the particular option. While at

Copyright © MESCAL Consortium, June 2005

this state, the protocol blocks client's negotiation logic in the sense that it discards as invalid any other message that may come from the client's negotiation logic.

- *Wait for Negotiation Logic (WFNL):* In this state the protocol waits for the client's negotiation logic to determine and formulate its response to the last message received from the server within the context of the particular option. While at this state, any other option message that may come from the server is blocked in the sense that is not forwarded to the client's negotiation logic.
- *Bound Wait for Peer (BWFP):* This state is functionally identical to WFP with the exception of its transition behaviour. Here, the protocol allows for updating the state of an open option upon receiving a message from the server, despite the fact that the client is blocked waiting for an answer to a previously sent call for agreement. Thus, in case of rejection of the call for agreement, negotiation on open options can resume as normal.
- *Bound Wait for Negotiation Logic (BWFNL):* In this state the client's negotiation logic is blocked until the reception of an answer to a previously sent call for agreement.
- *Cooled (COOLED):* In this state the option is provisionally agreed and closed; subsequently it may only be dropped by either party, but no further negotiated.

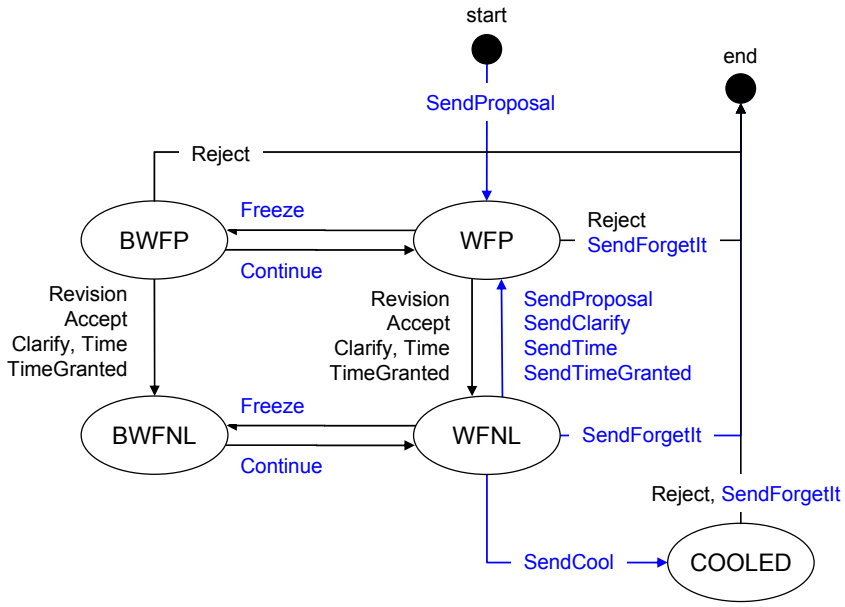


Figure 61. The SrNP client FSM option level state transition diagram

State	Event	Actions	Next State
WFP	Revision, Accept, Clarify, Time, TimeGranted	If received message is valid and on time then forward message, update OptionTimer, update state; otherwise forward protocol exception and drop option.	WFNL
	Freeze	Update state.	BWFP
	Reject, SendForgetIt	If message is on time send/forward message; otherwise forward protocol exception. Drop option.	
WFNL	SendProposal, SendClarify, SendTime, SendTimeGranted	If message is valid and on time then send message, update OptionTimer, update state; otherwise forward protocol exception and drop option.	WFP
	Freeze	Update state.	BWFNL

Copyright © MESCAL Consortium, June 2005

	SendCool	If message is valid and on time then send message, update OptionTimer, update state; otherwise forward protocol exception and drop option.	COOLED
	SendForgetIt	If message is on time send message; otherwise forward protocol exception. Drop option.	
BWFP	Revision, Accept, Clarify, Time, TimeGranted	If message is valid and on time then forward message, update OptionTimer, update state; otherwise forward protocol exception and drop option.	BWFNL
	Continue	Update state.	WFP
	Reject	If message is on time forward message; otherwise forward protocol exception. Drop option.	
BWFNL	Continue	Update state.	WFNL
COOLED	Reject, SendForgetIt	If message is on time send/forward message; otherwise forward protocol exception. Drop option.	

Table 13: The SrNP client option FSM state transition table

9.3.6.4 SrNP Server FSM

The server session FSM includes the following states:

- *Initialisation (INIT):* In this state the protocol performs the necessary initialisation communication between the negotiation parties before proceeding to the actual negotiations.
- *Idle (IDLE):* In this state the protocol is active only in the option level and the control is to the corresponding option FSMs.
- *Bound Wait for Negotiation Logic (BWFNL):* In this state the protocol waits for a definite answer from the server's negotiation logic to a call for agreement previously received from the client.
- *Last Wait for Peer (LWFP):* In this state the protocol has entered the last negotiation round and it's the client's turn to make a call for agreement or quit the negotiations.
- *Last Wait for Negotiation Logic (LWFNL):* In this state the protocol has entered the last negotiation round and it's the server's turn to either give a definite answer or provide a last revision.
- *Last Bound Wait for Negotiation Logic (LBWFNL):* In this state the protocol has entered the last negotiation round and it's the server's turn to either consent to the call for agreement or quit the negotiations.

Copyright © MESCAL Consortium, June 2005

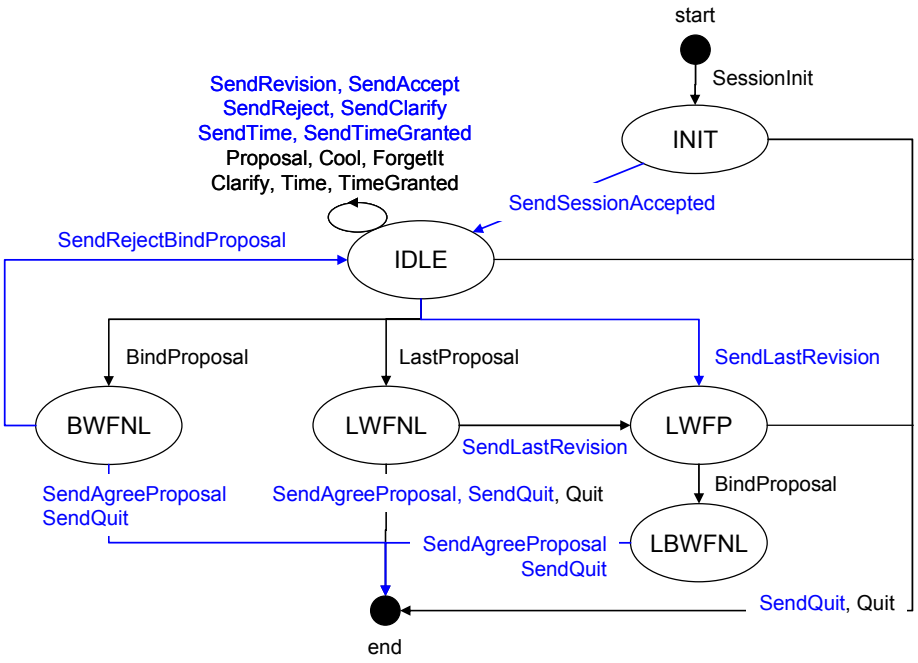


Figure 62. The SrNP server FSM session level state transition diagram

State	Event	Actions	Next State
INIT	SendSessionAccepted	If message is on time then send message, update SessionTimer, update state; otherwise forward protocol exception and abort session.	IDLE
	SendQuit, Quit	Send/forward message and abort session.	
IDLE	SendRevision, SendAccept, SendReject, SendClarify, SendTime, SendTimeGranted, Proposal, Cool, ForgetIt, Clarify, Time, TimeGranted	Pass control to the corresponding option FSM. If message is deemed on time and valid then update SessionTimer (see section 9.3.6.1).	IDLE
	BindProposal	If message is on time then forward message, update OptionTimer, update SessionTimer, update state; otherwise forward protocol exception and abort session.	BWFNL
	LastProposal	If message is on time then drop all other options, forward message, stop SessionTimer, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LWFNL
	SendLastRevision	If message is on time then drop all other options, send message, stop SessionTimer, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LWFP
	SendQuit, Quit	Send/forward message and abort session.	
BWFNL	SendAgreeProposal	If message is on time then send message; otherwise forward protocol exception. Abort session.	

Copyright © MESCAL Consortium, June 2005

	SendRejectBindProposal	If message is on time then send message, drop option, update SessionTimer, update state; otherwise forward protocol exception and abort session.	IDLE
	SendQuit	Forward message and abort session.	
LWFNL	SendLastRevision	If message is on time then drop all other options, send message, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LWFP
	SendAgreeProposal	If message is on time then send message; otherwise forward protocol exception. Abort session.	
	SendQuit, Quit	Send/forward message and abort session.	
LWFNP	BindProposal	If message is on time then drop other options, forward message, update OptionTimer, update state; otherwise forward protocol exception and abort session.	LBWFNL
	SendQuit, Quit	Send/forward message and abort session.	
LBWFNL	SendAgreeProposal	If message is on time then send message; otherwise forward protocol exception. Abort session.	
	SendQuit	Send message and abort session.	

Table 14: The SrNP server session FSM state transition table

The server option FSM includes the following states:

- *Wait for Peer (WFP):* In this state the protocol waits to receive from the client a response to the last sent message determined by the server's negotiation logic for the particular option. While at this state, the protocol blocks server's negotiation logic in the sense that it discards as invalid any other message that may come from the server's negotiation logic.
- *Wait for Negotiation Logic (WFNL):* In this state the protocol waits for the server's negotiation logic to determine and formulate its response to the last message received from the client within the context of the particular option. While at this state, any other option message that may come from the client is blocked in the sense that is not forwarded to the server's negotiation logic.
- *Cooled (COOLED):* In this state the option is provisionally agreed and closed; subsequently it may only be dropped by either party, but no further negotiated.

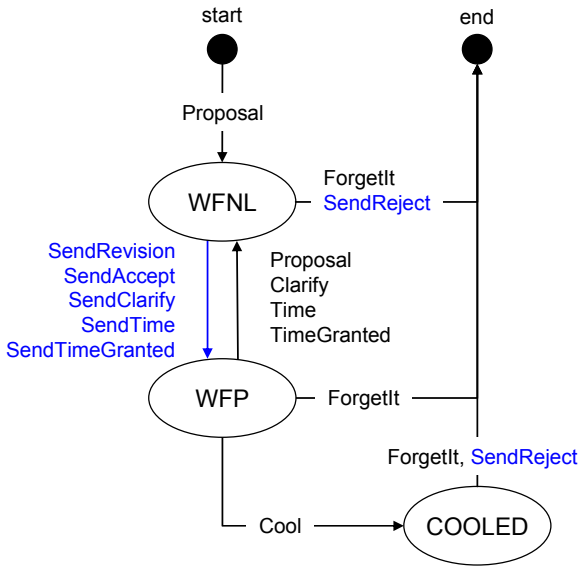


Figure 63. The SrNP server FSM option level state transition diagram

Copyright © MESCAL Consortium, June 2005

State	Event	Actions	Next State
WFNL	SendRevision, SendAccept, SendClarify, SendTime, SendTimeGranted	If received message is valid and on time then send message, update OptionTimer, update state; otherwise forward protocol exception and drop option.	WFP
	SendReject, ForgetIt	If message is on time send/forward message; otherwise forward protocol exception. Drop option.	
WFP	Proposal, Clarify, Time, TimeGranted	If message is valid and on time then forward message, update OptionTimer, update state; otherwise forward protocol exception and drop option.	WFNL
	Cool	If message is valid and on time then forward message, update OptionTimer, update state; otherwise forward protocol exception and drop option.	COOLED
	ForgetIt	If message is on time forward message; otherwise forward protocol exception. Drop option.	
COOLED	SendReject, ForgetIt	If message is on time send/forward message; otherwise forward protocol exception. Drop option.	

Table 15: The SrNP server option FSM state transition table

9.4 SLS Order Handling

9.4.1 Objectives

The goal of the *SLS Order Handling* function block, as clearly declared by its name, is to handle the orders of the customers of the AS -end consumers and peering ASs- concerning the offered services. The orders come in the form of pSLS, cSLS requests, imprinted using XML documents.

Figure 64 presents the SLS Order Handling functional component and the functional components with which it interacts, within the same AS and across neighbouring ASes employing the MESCAL functional architecture.

Copyright © MESCAL Consortium, June 2005

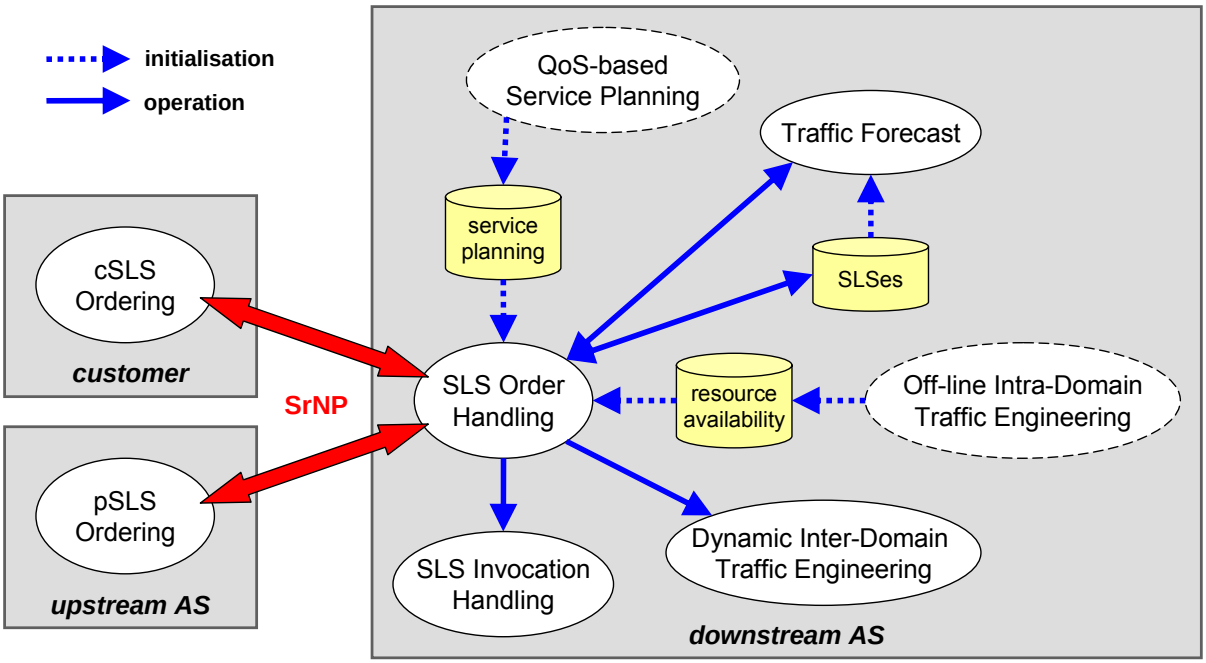


Figure 64: SLS Order Handling

9.4.2 Interface Specification

The SLS Order Handling component implements two kinds of interfaces: The external interface for communicating with the peering ASs and the customers and a set of internal interfaces for communicating with other components of its own AS, as defined by the MESCAL architecture.

Notation and basic abstractions underlying the interactions of the SLS Order Handling over its interfaces are described in the next sections followed by the interfaces specification.

9.4.2.1 Notation

Notation used in the following sections includes the following conventions:

`{ }` as in {x}, denotes a set of x

`.` as in x.y, denotes membership, y is an attribute or a function of x

`::` as in x :: y z, denotes definition, x is defined as a type with attributes y and z

`|` as in x :: (y | z), denotes logical XOR, x is defined as a type with attributes either y or z.

9.4.2.2 Destination Groups

In an effort to facilitate the communication among providers of reachable destinations per pSLS offering, we introduce the notion of destination groups. A destination group may include several IP address prefixes, grouped under a descriptive label. We identify two ways to divide the Internet address space, division based on geological topology and division based on ISPs (see Figure 65).

Copyright © MESCAL Consortium, June 2005

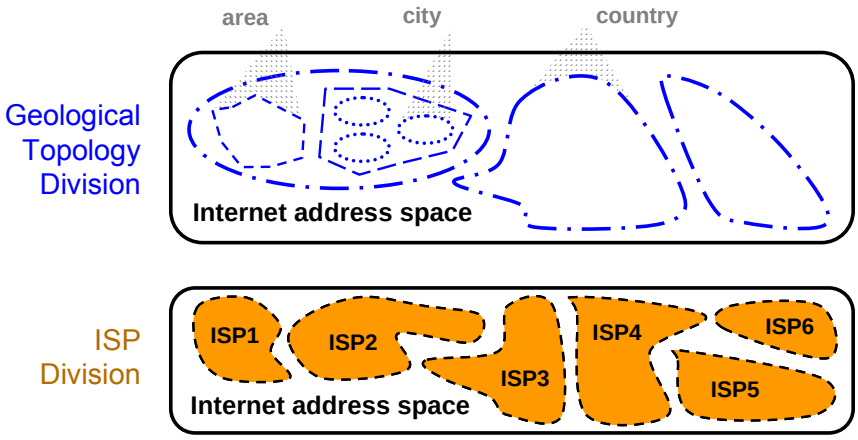


Figure 65: Internet Address Space Grouping

We assume the existence of universally known destination group labels together with their mapping to IP address prefixes. In fact, organisations such as Regional Internet Registries ARIN, RIPE, APNIC and LACNIC are responsible for maintaining consistency on IP address prefixes allocation to ISPs. Moreover, tools and protocols for maintaining a database with IP address prefixes mapping to geological locations defined in standardisation bodies (ISO, FIPS) are widely available.

In addition to universally known labels, each provider may create proprietary labels, grouping any combination of universally known labels and/or IP address prefixes.

In any case, a destination group will be translated into a set of IP address prefixes:

```
region :: {city}
country :: {region}
geo-location :: country | region | city
dest-group :: dest-group-id (geo-location | ISP-id) {IP-addr-prefix}
dest :: dest-group-id | IP-addr-prefix
```

The relationships among IP prefixes and the derived relationships among destination groups are depicted in Figure 66.

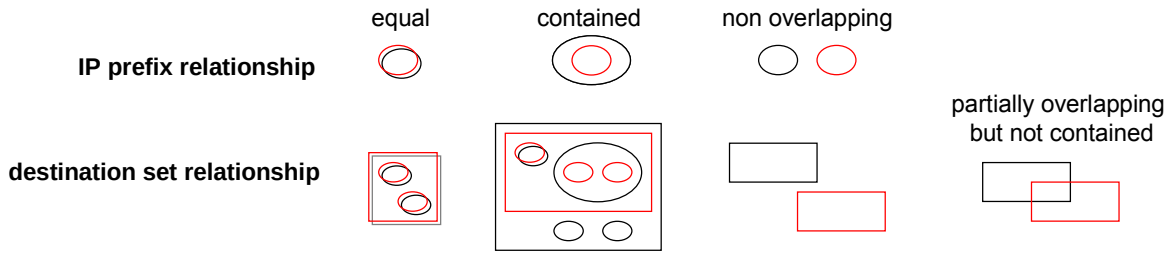


Figure 66: Destination Group Relationships

9.4.2.3 External Interface to Ordering components

SLS Order Handling acts as a provider within the ordering framework (see section 9.5.2). It negotiates pSLSes with the *pSLS Ordering* function of upstream ASs or cSLSes with the *cSLS Ordering* customer function over *Service Negotiation Protocol* (see section 9.3).

SrNP specifies the necessary common vocabulary and sets the negotiation rules that will lead to unambiguous and converging negotiations. In addition it undertakes the transportation of the messages –in form of XML documents– between the negotiating parties.

Copyright © MESCAL Consortium, June 2005

The information exchanged over this interface is in terms of pSLS or cSLS requests in the form of XML documents adhering to the offered by the AS pSLS, cSLS templates and carried over SrNP messages (see sections 9.2, 9.3). The destination groups requested or offered in the context of an SLS are modelled using the SiteType XML element (see Figure 52). The output of this interface will be the response of the SLS Order Handling component with pSLS and cSLS counter-offers, or with definite accept/reject replies to the received requests following the SrNP primitives.

9.4.2.4 Internal Interface to QoS-based Service Planning

Based on the business objectives and the findings of the *QoS Capabilities Discovery* function, the *QoS-based Service Planning* constructs the targeted local-QoS-classes (lQCs), extended-QoS-classes (eQCs), destination groups and services, against which the network resources are dimensioned by the traffic engineering functions.

```
lQC :: lQC-id {qos-attr}
eQC :: (eQlC-id {qos-attr-alias}) | (eQnC-id {qos-attr})
eQlC-impl :: {impl-order lQC}
service :: srv-type eQC {dest} {q-BGP-attr} cost-formula
```

lQCs are defined in terms of quantitative values of well known QoS attributes (delay, loss, etc.). eQCs are defined either qualitatively or quantitatively. Qualitative eQCs are mapped to one, or a list of prioritised alternative lQCs to be implemented by in the intra-domain scope. *Premium* qualitative eQC for example may be mapped to the lQC1 with zero loss, minimum delay and jitter, while *gold* eQC may be mapped to lQC2 (QoS lower than lQC1) impl-order = 1 and to lQC1 impl-order = 2.

Supported services are defined per service type, service types distinguished by business relationships and solution option type, such as the pSLS models described in section 9.2.2. Per service are also specified the supported q-BGP attributes (e.g. average one way delay, jitter, available rate etc.) updated by dynamic traffic engineering and advertised by q-BGP speaker nodes for the particular service. Finally the cost formula applicable to the service is specified per service. Cost formulas are defined arbitrarily, depending on the business objectives of the provider.

The above mentioned service planning information is provided to the SLS Order Handling component upon initialisation.

9.4.2.5 Internal Interface to Off-line Intra-domain Traffic Engineering

Through this interface SLS Order Handling receives as input information upon initialisation the established downstream pSLSes (offered pSLSes) and the associated offered QoS-classes (oQCs), the QoS-bindings in effect and the Resource Availability Matrix (RAM), used by the service translation and subscription admission control functions. In fact, we distinguish between inter-domain and intra-domain resource availability, the first directly derived from the capacity of the downstream pSLSes, the latter calculated by the off-line intra-domain traffic engineering and ensured by the dynamic intra-domain traffic engineering. The above mentioned information is expressed as:

```
off-pSLS :: egress-link {dest} oQC capacity
QC-binding :: eQC lQC oQC
iRAM-entry :: ingress-router egress-router lQC capacity
```

9.4.2.6 Internal Interface to Traffic Forecast

The admission control function within the SLS Order Handling component invokes the *Demand Derivation* function of the *Traffic Forecast* component to perform its calculations.

In addition, the established SLSes are stored and maintained in the *SLS Repository* from which they are retrieved by the *Traffic Forecast* component for calculating the forecasted demand at the beginning of a Resource Provisioning Cycle (RPC).

Copyright © MESCAL Consortium, June 2005

9.4.2.7 Internal Interface to Dynamic Inter-domain Traffic Engineering

Through this interface SLS Order Handling notifies the dynamic inter-domain traffic engineering component upon SLS establishment for undertaking the necessary actions for the activation of the newly established SLS. Such actions would be the update of q-BGP configuration at AS boundary routers or, for the hard guarantees solution option, the configuration of the *Path Computation System* (PCS) information of the new peer provider to be used in subsequent *Path Computation Protocol* messages for the establishment of hard guaranteed LSPs. q-BGP at AS boundary routers is provided with the QoS attributes and values to advertise (e.g. average one way delay, jitter, etc.), the interface of the neighbour AS to advertise to and the DSCP for the service over the particular boundary link.

9.4.2.8 Internal Interface to SLS Invocation Handling

Upon SLS establishment SLS Order Handling configures the admission logic of the *SLS Invocation Handling* component to take into account and serve accordingly the new subscription.

9.4.3 Behaviour Specification

For fulfilling its objectives SLS Order Handling should cater for the following functionality.

SLS Order Handling *conducts negotiations with customers*. The SLS Order Handling component implements the role of a provider within the ordering and negotiation framework, conducting negotiations with the *Ordering* components of the customers or the peer providers requesting the establishment of a cSLS/pSLS respectively. pSLS authoring and translation functions implement the manipulation of the negotiated documents based on the SLS XML schemas (see section 9.2). The SLS Order Handling negotiation logic may handle many simultaneous negotiation sessions.

SLS Order Handling *performs subscription admission control*. This functionality is responsible for deciding on the faith of the requested SLSes. Admission control is two-fold, authorisation-based and resource-based. Authorisation-based admission control undertakes validity checks of the requested pSLS against the service offerings configured by service planning, e.g. check if the requested eQC is supported for the requested destination groups etc. Resource-based admission control checks whether the anticipated traffic of the requested plus the existing SLSes can be accommodated over the available resources. The anticipated traffic is calculated using functions of the *Traffic Forecast* component and the available resources are provided by the *Off-line Intra-domain Traffic Engineering* component upon initialisation. In case that the SLS cannot be accepted, admission control attempts to calculate alternative feasible SLSes, close to the one requested. These SLSes are proposed to the customer as counter-offers.

SLS Order Handling is responsible for *SLS establishment*. It triggers the necessary functions for activating the newly agreed SLSes, so that the customer will be enabled to use the service he/she has subscribed for. This function entails the storage of the SLS to a commonly accessible repository and then, the notification of the *SLS Invocation* and of the *Dynamic Inter-domain Traffic Engineering* components for appropriately activating the SLS, shipping the necessary q-BGP updates and configuring the *Patch Computation Server* (PCS).

9.4.4 Resource-Based Subscription Admission Control Algorithm

The goal of resource-based subscription admission control is to determine whether the demand implied by the existing plus the requested SLS can be sustained by the network resources. The admission control algorithm must minimise false negatives, i.e. the rejection of a service request despite the sufficiency of resources to accommodate it. To this end, the subscription admission control algorithm builds upon the assumption of an underlying dynamic traffic engineering system capable to achieve the optimum distribution of traffic to network resources at any time. The load balancing, dynamic link scheduling adaptation or other dynamic traffic engineering mechanisms required to achieve this are not of concern to the subscription admission control function.

Copyright © MESCAL Consortium, June 2005

To fulfil its goal, resource-based admission control needs to model the network resources and the demand implied by the services.

We distinguish between inter-domain and intra-domain network resources. Inter-domain resources are specified by the QoS-aware capacity for reaching specific destinations at each boundary link of the domain, limited by the established downstream pSLSes (see Figure 67). Intra-domain resources are provided at the *internal Resources Availability Matrix* (iRAM) calculated and enforced by the traffic engineering system. The availability estimate is expressed per *internal Traffic Trunk* (iTT).

```
iTT :: ingress-router egress-router lQC
off-pSLS :: egress-link {dest} oQC capacity
iRAM-entry :: iTT capacity :: ingress-router egress-router lQC capacity
QC-binding :: eQC lQC oQC
```

Inter-domain and intra-domain resources are combined to form an abstract network of two-hop paths, the first hop corresponding to the iTT and the second hop to the downstream pSLS. The capacity of the links is specified then in the iRAM and downstream pSLSes for the first and second hop respectively.

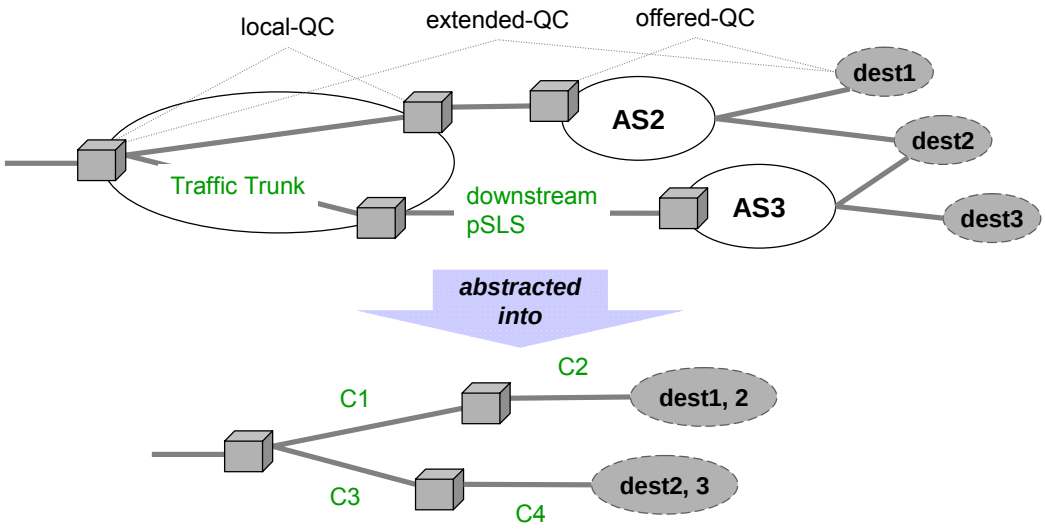


Figure 67: Network Resources Model for Subscription Admission Control

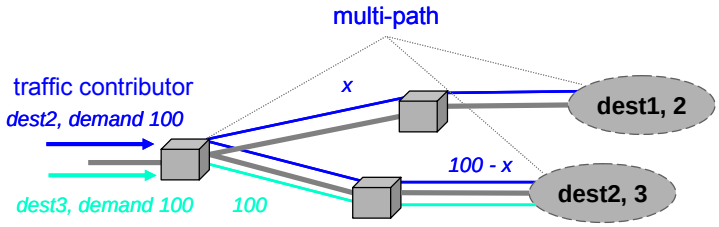


Figure 68: Demand Model for Subscription Admission Control

The following notions underline the representation of the demand in the resource-based subscription admission control algorithm (see Figure 68):

```
traffic-contributor :: ingress-router {dest} eQC demand
eTT :: ingress-link {dest} eQC
multi-path :: traffic-contributor {iTT off-pSLS}
```

- *Traffic Contributor*: A traffic contributor describes a QoS macro-flow entering a particular ingress edge router at the provider's domain, heading to certain destinations anywhere in the Internet and

Copyright © MESCAL Consortium, June 2005

requiring a specific eQC treatment. Traffic demand is associated with a traffic contributor denoting the volume of traffic for the macro-flow.

- *Multi-Path*: A multi-path captures a path of one ingress point and many egress points (hose), adequate for accommodating a traffic contributor in terms of reaching the required destinations with the required eQC. The ingress point of the multi-path is the ingress-router of the traffic contributor while the egress points are the set of downstream (offered) pSLSes such that:

  a) There is a QoS-class binding for the eQC of the traffic contributor to the oQC of each downstream pSLS, with some capacity allocated for the corresponding lQC at the iTT between traffic contributor ingress node and downstream pSLS boundary link.

  b) The union of the destinations reached by the downstream pSLSes is a superset of the destinations of the traffic contributor.

  c) The destinations of each pSLS are not fully contained in the destinations of another pSLS in the same multi-path.

After the translation of a c/pSLS to the contained SLSes and their authorisation-based validation, resource-based admission control invokes the *Demand Derivation* function of *Traffic Forecast* component (see section 10.2.3.2) to calculate the anticipated demand for the requested and the existing c/pSLSes in terms of eTT-SLS (see **Step #1** in section 10.2.3.2.2). The anticipated demand is then aggregated over the eTT-SLSes with identical ingress router, eQC and equal destination groups to form a traffic contributor.

Alternative multi-paths per traffic contributor are calculated based on the QoS-class bindings in effect, the existence of iTT first-hop links for the lQCs in the bindings and the matching of destinations.

After setting up all the possible multi-paths per traffic contributor, admission control needs to deduce whether there is a splitting of the demand of each traffic contributor in its alternative multi-paths, such that demand from all traffic contributors can be accommodated without overloading any link of the abstract network. To this end, a linear equation system is build with an equation per link of the abstract network. The left part of the equation is the parameterised demand of the traffic contributors with a multi-path over the link and the right part the capacity of the link. For example, the linear system for the network resources and traffic contributors depicted in Figure 67 and Figure 68 is the following:

$$\begin{cases} x < C1 \\ x < C2 \\ 100 + 100 - x < C3 \\ 100 + 100 - x < C4 \end{cases} \Rightarrow \begin{cases} x < \min(C1, C2) \\ 200 - x < \min(C3, C4) \end{cases}$$

The admission control algorithm then applies a simplex-based method to deduce whether the linear system has solutions, in which case the service request can be accepted.

Based on the above notions and terminology, the general problem of pSLS admission control can be formulated as an optimisation as follows:

Let:

T: be the set of traffic contributors; traffic contributors are usually indexed by $i \in T$.

$b^{(i)}$: be the total traffic demand of traffic contributor i, $i \in T$.

$P^{(i)}$: be the set of multi-paths of traffic contributor I, $i \in T$; each multi-path in the set is noted by $mp_j^{(i)}$, $j \in P^{(i)}$

$x_j^{(i)}$: be a variable associated with each multi-path, $mp_j^{(i)}$, representing the load (portion of total contributor's demand) to be accommodated in the multi-path, as determined by the optimisation problem (see below).

Copyright © MESCAL Consortium, June 2005

E: the set of links in the abstract (2-link-hop) network constructed from the resource availability estimates and the QC-bindings determined by TE; each link in this abstract (resource availability) network is usually denoted by $\ell \in E$.

C_ℓ: the capacity of link $\ell \in E$; as previously outlined, for the first-hop links their capacity expresses the availability of the engineered network to accommodate QoS traffic intra-domain (per l-QC) and for the second-hop links the capacity of the pSLSes 'bought' to accommodate QoS inter-domain traffic.

The optimisation problem:

Minimise

$f(x_j^{(i)}, i \in T, j \in P^{(i)})$, where f is a linear function with respect to $x_j^{(i)}$

Subject to the constraints:

Total demand constraints:

$$\sum_{j \in P^{(i)}} x_j^{(i)} = b^{(i)}, i \in T \qquad \text{--as many equations as the traffic contributors}$$

Link capacity constraints:

$$\sum_{i \in T} \sum_{j \in P^{(i)}} \delta_j^{(i)}(\lambda) x_j^{(i)} \le C_\lambda, \ell \in E \qquad \text{--as many as the links participating the multi-paths}$$

where $\delta_j^{(i)}(\lambda)$ is an indicator function, being 1 if the multi-path j of contributor i passes link ℓ and 0 otherwise.

Non-negativity flow constraints:

$x_j^{(i)} \ge 0, i \in T, j \in P^{(i)}$

By solving the above linear optimisation problem (e.g. with the simplex method), one can determine whether there is a feasible solution satisfying the above constraints and if yes, the optimum demand distribution. If a feasible solution does not exist, then the requested pSLS cannot be safely (given the availability estimates) accommodated by the engineered network.

9.5 pSLS Ordering

9.5.1 Objectives

The role of the *pSLS Ordering* functional block is to establish the set of pSLS agreements, the most advantageous to the AS with respect to traffic engineering and business objectives. The *Binding Selection* block places a collective pSLS order (see Figure 69). pSLS Ordering executes the order after conducting negotiations over the SrNP protocol (see section 9.3) with the SLS Order Handling blocks of the downstream ASes related to the pSLSes in the pSLS order.

Copyright © MESCAL Consortium, June 2005

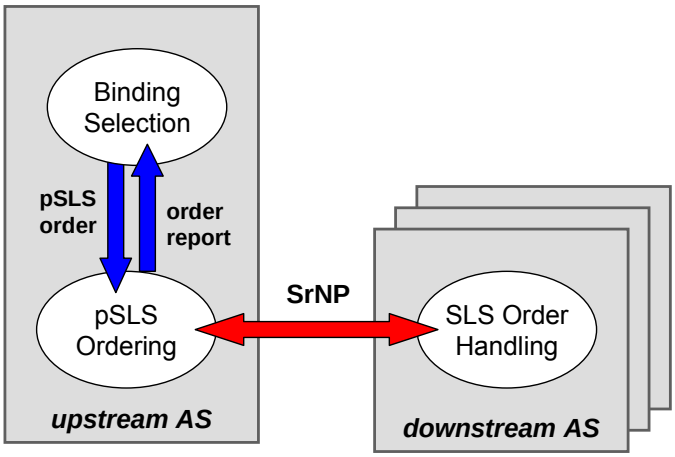


Figure 69. pSLS Ordering

Offline traffic engineering algorithms calculate the pSLSes to order, based on QoS announcements of the peer providers. However, this information may be insufficient for the particular needs of the provider or obsolete at the moment of actually ordering the pSLS because e.g. the service resources are no longer available or because the provider has changed its price list. Hence, traffic engineering can benefit from a non-monolithic pSLS ordering process, such that it can go beyond have-it or not-have-it result and explore alternatives based on well defined negotiation directives.

Departing from the requirements of pSLS Ordering we investigated the issues around a generic ordering and negotiation framework presented in the following sections.

9.5.2 Ordering and Negotiation Framework

The ordering and negotiation framework is built over the service negotiation protocol (SrNP) (see section 9.3).

We distinguish the roles of the *customer*, the *negotiation agent* and the *provider*. The customer compiles the order and passes it to the negotiation agent together with negotiation strategies, agreement restrictions and preferences; we call this enhanced order a *Negotiation Plan*. The negotiation agent undertakes the negotiations and the establishment of agreements with the providers involved in the order.

9.5.2.1 Negotiation Plan

An order may contain multiple order items. Each order item constitutes a *Negotiation Target* (see Figure 70 and Figure 71), an agreement for the negotiation agent to pursue with the provider of the particular order item. Each negotiation target is specified as a set of *Negotiation Issues* and each negotiation issue can be evaluated with either scalar or ordered discrete values. *Tolerance criteria* may be specified for the negotiation issue. For example, a pSLS is a negotiation target specified by the meta-QoS-class, the bandwidth, the cost etc. Possible tolerance criteria would be a limited set of desired meta-QoS-class values, a limited range of required bandwidth, a maximum cost, etc. In addition, an initial value to pursue at the first negotiation round may be provided per issue as guidelines.

Multiple negotiation targets to be collectively ordered form a *Negotiation Packet*. A negotiation packet cannot be ordered if it doesn't meet the *packet acceptance criteria*. Acceptance criteria are expressed as restrictions on values of negotiation issues of the contained negotiation targets, aggregated across the targets. For example, the total cost of a packet is evaluated as the sum of the cost of each individual target and may be restricted to a maximum value. Acceptance criteria may also take the form of arithmetic expressions with parameters the aggregated values, or they may be expressed conditionally; conditions are also expressed as arithmetic expressions on aggregated values.

Copyright © MESCAL Consortium, June 2005

Multiple alternative negotiation packets form a *Negotiation Package*. The negotiation packets within a negotiation package are pursued together and may overlap on some negotiation targets. Among the packets meeting their acceptance criteria the best one is selected based on the package *selection criteria* and finally ordered. If none of the contained packets meets its acceptance criteria then the package fails. Selection criteria are expressed as prioritised maximise/minimise instructions on aggregated target values. For example, the packet of pSLSes to order may be selected to be the one maximising the total bandwidth, provided that the packet total cost does not exceed the maximum value specified in its acceptance criteria.

The *Negotiation Pool* is a negotiation package within which all the possible combinations of the contained targets will be pursued instead of only specific negotiation packets. A combination may contain any number of the contained targets, starting from all single targets up to the combination of all targets. The *combination acceptance criteria* and *combination selection criteria* will be applied to every combination as in the negotiation package.

In case of failure of the negotiation package or the negotiation pool, an alternative package or pool may be pursued following a *Negotiation Plan*. A negotiation plan is a collection of packages and pools, organised as failover options. Transition to options when the package or pool fails is decided based on *transition criteria*. Transition criteria are expressed as conditional expressions on aggregated values resulting in either pursuing further with another plan or pool, interrupting for human interaction to make the decision or stop without executing an order.

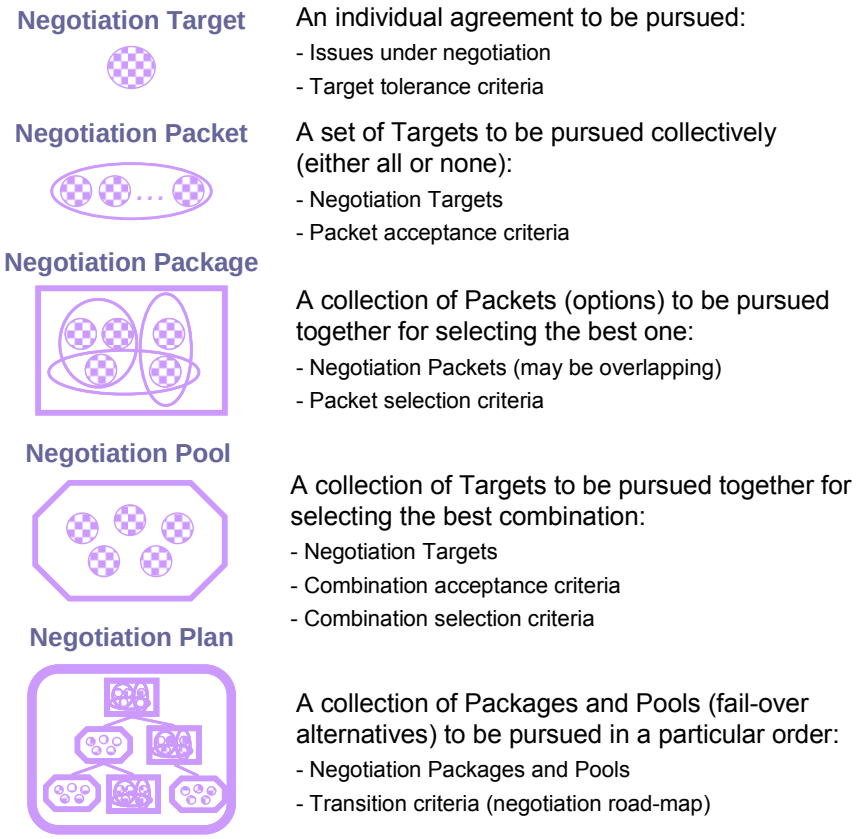


Figure 70. Negotiation Plan Elements

Copyright © MESCAL Consortium, June 2005

```
<negotiation plan>       ::= <plan identifier> 'START FROM' <negotiation node>
<negotiation node>       ::= (<negotiation package> | <negotiation pool>) <transition statement>

<negotiation pool>       ::= <pool identifier> 'ACCEPTANCE CRITERIA' <acceptance criteria>
                             'SELECTION CRITERIA' <selection criteria> 'CONTAINS' <negotiation issue>+
<negotiation package>    ::= <package identifier> 'CRITERIA' <selection criteria> 'CONTAINS' <negotiation packet>+
<negotiation packet>     ::= <packet identifier> 'CRITERIA' <acceptance criteria> 'CONTAINS' <negotiation target>+
<negotiation target>     ::= <target identifier> 'CRITERIA' <complex criteria> 'CONTAINS' <negotiation issue>+
<negotiation issue>      ::= <issue identifier> 'CRITERIA' <tolerance criteria>

<selection criteria>     ::= <preference statement>*
<acceptance criteria>    ::= <acceptance statement>
<complex criteria>       ::= <acceptance statement>
<tolerance criteria>     ::= <scalar> | <discrete numeric> | <discrete string>

<transition statement>   ::= <if clause> <transition statement> ['ELSE' <transition statement>] | <transition instruction>
<transition instruction> ::= 'STOP' | 'PROCEED TO' <negotiation node> | 'INTERRUPT'
<preference statement>   ::= <precedence> ('MAXIMISING' | 'MINIMISING') <arithmetic expression>
<acceptance statement>   ::= <if clause> <acceptance statement> ['ELSE' <acceptance statement>] | <acceptance instruction>
<acceptance instruction> ::= 'ACCEPT' | 'REJECT' | 'INTERRUPT'
<if clause>              ::= 'IF' <boolean expression> 'THEN'

<precedence>             ::= <unsigned integer>
<scalar>                 ::= ['FROM' <number>] ['TO' <number>] 'STARTING FROM' <number>
<discrete numeric>       ::= 'IN' <numeric sequence> 'STARTING FROM' <number>
<discrete string>        ::= 'IN' <string sequence> 'STARTING FROM' <string>
<numeric sequence>       ::= <number> (',' <number>)*
<string sequence>        ::= <string> (',' <string>)*

<plan identifier>        ::= 'PLAN::'<identifier>
<package identifier>     ::= 'PACKAGE::'<identifier>
<packet identifier>      ::= 'PACKET::'<identifier>
<target identifier>      ::= 'TARGET::'<identifier>
<issue identifier>       ::= 'ISSUE::'<base parameter>

<variable>               ::= <aggregated parameter> | <base parameter>
<aggregated parameter>   ::= <aggregation function> '(' <base parameter> ')'
<aggregation function>   ::= 'SUM' | 'AVG' | 'MIN' | 'MAX' | 'VAR' | 'STDEV'
<base parameter>         ::= <identifier>

<boolean expression>     ::= (<boolean factor> 'OR')* <boolean factor>
```

Copyright © MESCAL Consortium, June 2005

```
<boolean factor>          ::= (<boolean negation> 'AND')* <boolean negation>
<boolean negation>        ::= ['NOT'] <boolean primary>
<boolean primary>         ::= '(' <boolean expression> ')' | <relation> | <logical value>
<relation>                ::= <arithmetic expression> <relational operator> <arithmetic expression>

<arithmetic expression>   ::= (<arithmetic term> <adding operator>)* <arithmetic term>
<arithmetic term>         ::= (<arithmetic factor> <multiplying operator>)* <arithmetic factor>
<arithmetic factor>       ::= (<arithmetic primary> <power operator>)* <arithmetic primary>
<arithmetic primary>      ::= '(' <arithmetic expression> ')' | <unsigned number> | <variable>

<number>                  ::= [<sign>] <unsigned number>
<unsigned number>         ::= <decimal number> ['E'<integer>]
<decimal number>          ::= <unsigned integer> ['.'<unsigned integer>]
<integer>                 ::= [<sign>] <unsigned integer>
<unsigned integer>        ::= <digit>+

<string>                  ::= <alphanumeric>+
<alphanumeric>            ::= <digit> | <letter>

<logical value>           ::= 'TRUE' | 'FALSE'
<relational operator>     ::= '<' | '>' | '=' | '<=' | '=>' | '<>'

<adding operator>         ::= '+' | '-'
<multiplying operator>    ::= '*' | '/' | '÷'
<power operator>          ::= 'POWER' | 'SQUARE'

<sign>                    ::= '+' | '-'
<digit>                   ::= '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9'
<letter>                  ::= 'a' | 'b' | 'c' | 'd' | 'e' | 'f' | 'g' | 'h' | 'i' | 'j' | 'k' | 'l' | 'm' | 'n' | 'o' |
                              'p' | 'q' | 'r' | 's' | 't' | 'u' | 'v' | 'w' | 'x' | 'y' | 'z' | 'A' | 'B' | 'C' | 'D' |
                              'E' | 'F' | 'G' | 'H' | 'I' | 'J' | 'K' | 'L' | 'M' | 'N' | 'O' | 'P' | 'Q' | 'R' | 'S' |
                              'T' | 'U' | 'V' | 'W' | 'X' | 'Y' | 'Z'
```

Figure 71. Negotiation Plan EBNF specification

Copyright © MESCAL Consortium, June 2005

9.5.2.2 Ordering and Service Negotiation Protocol

The negotiation agent acts as an SrNP client pursuing a negotiation target within the context of an SrNP session with the provider acting as an SrNP server. A negotiation target may be part of several negotiation packages or combinations in negotiation pools. As such, the values for its negotiation issues required to satisfy the different packet/pool acceptance criteria may differ, depending also on the other values of the other targets of the packet/combination it is part of. For each such variation of a negotiation target, an independent SrNP option may be pursued. Eventually, only one of the alternative packets or combinations will be ordered, hence only one of the open options per SrNP session will be established.

9.5.3 pSLS Ordering Case Study

9.5.3.1 Interface with Traffic Engineering

The pSLS Ordering component is merely an instrument for negotiating and establishing pSLSes. The decision on the pSLSes the AS needs to establish and on the terms of their negotiations belongs to the *Binding Selection* function. The Binding Selection function calculates the anticipated cost of alternative pSLSes combinations. Instead of ordering only the anticipated optimum pSLSes combination, several alternatives can be fed to the pSLS Ordering, increasing thus the possibility to achieve the actually optimum pSLSes combination.

The input received by the Binding Selection could then be in the form of a negotiation plan containing one negotiation package with selection criteria the minimisation of the actual cost. The negotiation package would contain as many negotiation packets as the anticipated optimum combinations the Binding Selection deems worth to explore. Each such negotiation packet will contain the pSLSes of the combination and may have restrictions on the total cost, the minimum QoS, the minimum bandwidth etc., as well as individual restrictions per pSLS.

In fact, a pSLS will be provided as a set of parameters such as boundary link identifier or equivalent, destination address prefix(es), o-QC etc. (see 10.4.1.3.5) based on which the pSLS Ordering will generate an XML document complying with the pSLS XML schemas (see section 9.2.4).

After the termination of the negotiations pSLS Ordering reports the results back to the Binding Selection. In the case of success the finally established pSLSes are reported, in case of failure the reasons are reported along with the negotiation logs. These logs can be used for statistics, helpful for various decision making processes of the AS e.g. planning. Also, logs are the proof of the negotiations results in case of disputes with the peering ASs.

9.5.3.2 Behaviour Specification

pSLS Ordering implements the role of a negotiation agent for pursuing optimum collective agreements on pSLSes through multi-party negotiations with the *SLS Order Handling* components of the downstream ASs acting as ordering providers. pSLS authoring and translation functions implement the manipulation of the negotiated documents based on the pSLS XML schemas (see section 9.2.4).

The negotiation logic operates on a negotiation plan containing a negotiation pool and based only on the bandwidth and the cost of a pSLS. It adopts a transactional mechanism, evaluating the pSLSes not individually but as part of a combination. To this end, negotiation logic operates in rounds.

A round is initiated by the pSLS Ordering sending appropriate client SrNP messages towards the downstream ASs and is completed when receiving responses for every sent message following the multi-dialogue nature of SrNP. Before proceeding to the next round the negotiation logic translates the replies received by the providers based on the type of the exchanged SrNP messages and the contained documents (see Table 16). Then, it evaluates the candidate orders and forms the pSLS requests to pursue in the following negotiation round.

Copyright © MESCAL Consortium, June 2005

SrNP sent message	document	SrNP reply message	reply document	reply type
Proposal	Interrogative[2]	Reject		Rejected
		Revision	Same bandwidth	Answered
			Different bandwidth	Ignored
		Accept		Invalid
	Concrete	Reject		Rejected
		Revision	Same bandwidth	Revised cost
			Different bandwidth	Ignored
		Accept		Accepted
BindProposal		AgreeProposal		Accepted
		RejectBindProposal		Rejected
Cool		Reject		Rejected

Table 16. SrNP messages interpretation

The following sections provide a formal description of the pSLS Ordering problem definition and the negotiation logic algorithm.

9.5.3.3 Problem Definition

A negotiation plan is provided containing one negotiation pool with a number of targets.

Each target is a pSLS with a neighbouring AS, described by two negotiation issues: a) the amount of bandwidth and b) the cost per bandwidth unit. Note that this could be generalised to any commodity in place of bandwidth. Tolerance criteria may be provided for the bandwidth per pSLS, expressed by an admissible area [$tmin_i$, $tmax_i$] for i-th target. In addition it may be provided the initial value to pursue for bandwidth $tinit_i$.

The acceptance criteria for the negotiation pool are defined by the total amount of bandwidth `bwtotal` to pursue from all targets in a combination and optionally the maximum acceptable cost `costtotal`. The selection criterion for the negotiation pool is the minimisation of the total cost.

Finally, the number of negotiation rounds `maxrounds` the pSLS Ordering function may go through at maximum before concluding may be provided.

The objectives of the pSLS Ordering function are then to converge to the most beneficial admissible combination without exceeding the given maximum number of rounds.

9.5.3.4 pSLS Ordering Algorithm

The pSLS Ordering algorithm is presented in Figure 72 and Figure 73.

[2] An interrogative vs. a concrete pSLS document does not specify a value for the cost.

Copyright © MESCAL Consortium, June 2005

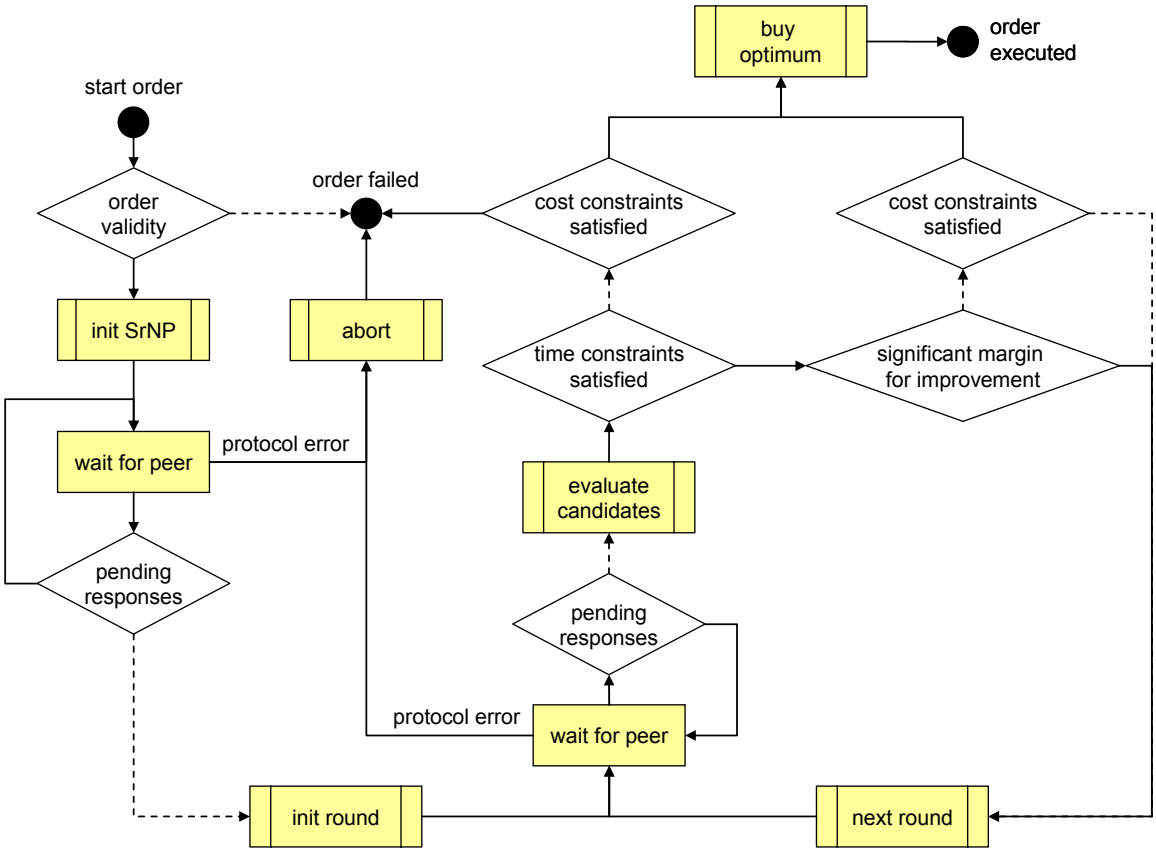


Figure 72. pSLS Ordering algorithm activity diagram

order validity

Order is valid if $\sum(tmin_{(j)})$ <= bwtotal <= $\sum(tmax_{(j)})$

init SrNP

Send SrNP SessionInit message for each target.

wait for peer

Blocks until it receives an SrNP message.

abort

Send SrNP Quit message for each target.

pending responses

There are no pending responses when there is a received message for every sent SrNP SessionInit, Proposal or BindProposal.

init round

Send new option SrNP Proposal message per target with bandwidth set to $tinit_{(i)}$ for targets with $tinit_{(i)}$ specified, otherwise set to

$tmin_{(i)} + (bwtotal - \sum(max(tmin_{(j)},tinit_{(j)})*((tmax_{(i)}-tmin_{(i)})/\sum(tmax_{(j)}-tmin_{(j)})))$.

evaluate candidates

Server response interpretation

- Get all server responses per target (accepted, rejected, ignored sent client proposals and server concrete revisions, see Table 15).
- Assume infinite cost for rejected and ignored proposals.
- Form $(bw_{(i)(j)},cost_{(i)(j)})$ cost formula points for each target *i* and bandwidth in server responses.
- Assume that $cost_{(i)(j)}$ per bandwidth unit holds for $(bw_{(i)(j-1)},bw_{(i)(j)}]$ range (cost formula discrete area).

Copyright © MESCAL Consortium, June 2005

Evaluate candidate orders against acceptance criteria

- Get all combinations of cost formula discrete areas of different targets (an order).
- For each order validate that:

$\sum(\max(bw_{(i)(j-1)}, tmin_{(i)})) <= bwtotal <= \sum(\min(bw_{(i)(j)}, tmax_{(i)}))$

- Discard invalid orders.
- For each order if $\sum(bw_{(i)(j)}) = bwtotal$ mark order as confirmed and calculate confirmed cost as $\sum(bw_{(i)(j)} * cost_{(i)(j)})$.
- For each non confirmed order calculate estimated cost as

$\sum(bw_{(i)(j)} * cost_{(i)(j)}) + \sum(\min(bw_{(k)(j)}, (bw_{(k)(j-1)} + bwstep)) * cost_{(k)(j)}) + (bwtotal - \sum(bw_{(i)(j)} * cost_{(i)(j)}) + \sum(\min(bw_{(k)(j)}, (bw_{(k)(j-1)} + bwstep)) * cost_{(k)(j)})) * cost_{(l)(j)}$,

where i-targets cost less than l-target, which in turn costs less than k-targets.

Candidate orders selection

- Select only the confirmed orders with minimum confirmed cost and discard the other confirmed.
- Select only the non confirmed orders with estimated cost less than the minimum confirmed cost and discard the rest orders.

time constraints satisfied

Time constraints are satisfied when number of rounds is not greater than maxrounds.

cost constraints satisfied

Cost constraints are satisfied there is a confirmed order with total cost not greater than costtotal.

significant margin for improvement

Significant margin for improvement is considered to exist when the minimum estimated cost over the minimum confirmed cost is greater than costdelta.

next round

- Send SrNP Cool message for every option in selected confirmed orders.
- Send SrNP Cool message for i-targets options in selected non confirmed orders.
- Send new option SrNP Proposal message for k-/l-targets in selected non confirmed orders.
- Send SrNP ForgetIt message to close all other open options requiring response from the client.

buy optimum

Send SrNP BindProposal message for each option participating in one of the orders with minimum confirmed cost.

Figure 73. pSLS Ordering algorithm pseudo-code

Copyright © MESCAL Consortium, June 2005

9.6 pSLS Invocation

9.6.1 Objectives

The *pSLS Invocation* function block is an offline component that is responsible for invoking pSLSs with peer domains. The pSLSs have already been subscribed through an ordering process between *pSLS Ordering* and *SLS Order Handling*. The ordering process establishes the overall boundaries and performance guarantees of the transport agreement between the peering domains, but an invocation process is required before user traffic can be passed between the domains. The invocation may request that the entire bandwidth that has been ordered is committed or it may request that only a portion is committed. Invocations may occur at intervals that are shorter than the Resource Provisioning Cycle epoch.

9.6.2 Interface specification

This section describes the interaction of the pSLS Invocation functional block with other functional blocks, as specified in the MESCAL functional architecture [D1.1]. Figure 74 shows the interfaces related with SLS Invocation.

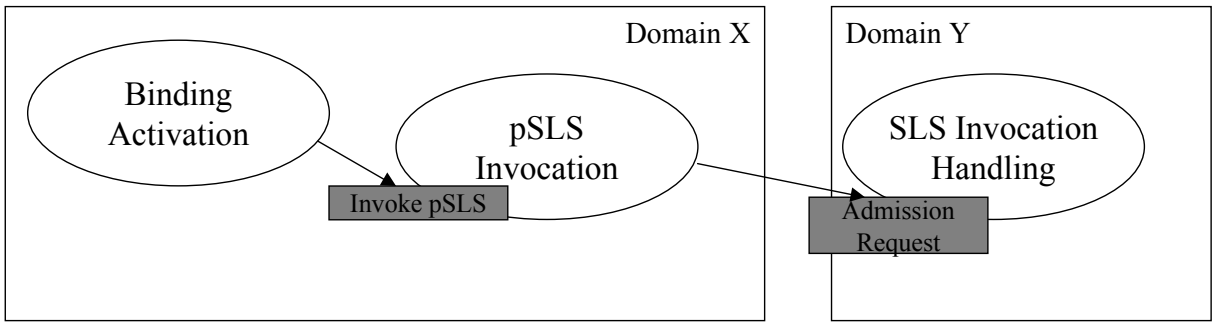


Figure 74. pSLS Invocation

- **Binding Activation to pSLS Invocation**

Invoke pSLS (eRAM)

Binding Activation determines the optimum arrangement of Inter-domain pSLSs to satisfy the TE requirements for a particular time interval and passes the information to pSLS Invocation. The eRAM identifies the peering domain, by specifying an egress interface, for each pSLS along with the required bandwidth.

- **pSLS Invocation to SLS Invocation Handling**

Admission Request

Each pSLS Invocation requires a request to be forwarded to the *SLS Invocation Handling* process in the peering domain, which is responsible for performing admission control.

Copyright © MESCAL Consortium, June 2005

9.6.3 Behavioural specification

Binding Activation determines the Inter-domain TE solution, as a set of pSLSs, that need to be established with peering domains. The pSLS requirements are conveyed as a set of entries in an eRAM. The eRAM contains the necessary information to enable pSLS Invocation to invoke the pSLSs, although it may be supplemented with pSLS identifiers, which are known from the prior pSLS subscription/handling process.

pSLS Invocation is responsible for identifying the appropriate peer domain and requesting that the peer domain admit the pSLS. If successful, *Binding Activation* is informed and the domain is prepared to transport QoS traffic over this path. If unsuccessful, *Binding Activation* is informed and it is *Binding Activation's* responsibility to decide on the next action.

It is possible that a set of pSLSs may have to be regarded as atomic by pSLS Invocation, which means that all invocations in the set must be invoked if the action is to be considered successful.

The negotiation between *pSLS Invocation* and *SLS Invocation Hand*ling can be based on the SrNP protocol described previously, although a simplified version would be sufficient as invocation requires less complex negotiation than the subscription/handling process.

9.7 SLS Invocation Handling

9.7.1 Introduction

SLS Invocation Handling (Admission Control) is an online component that is responsible for controlling the amount of traffic injected into the network so that conformant users achieve predefined performance objectives, as these are specified in the (c/p)SLSes. The term 'users' can correspond to either individual customers, e.g. home users, universities, organisations, or to entire provider domains. In the former case, cSLSes describe the required performance guarantees, whereas in the latter case, pSLSes describe the required traffic treatment. SLS Invocation Handling is a necessity in QoS-enabled networks and should act proactively so as to prevent saturation of the available resources and its consequences before they actually happen. Depending on the QoS guarantees that need to be provided and other factors, such as type (real-time or elastic) and aggregation level of carried traffic (cSLSes or pSLSes), overbooking ratios etc, different QCs will require different admission control policies and algorithms. Admission control needs also to take into account potential statistical multiplexing gain and interactions between the QCs that share the same links. Additionally the type of QoS parameters declared in the SLSes (e.g. peak rate only or other traffic descriptors) will greatly influence the employed admission control schemes.

9.7.2 Objectives

The main objective of SLS Invocation Handling is to guarantee that, once admitted, customer/domain service requests will receive the pre-agreed QoS treatment with the agreed guarantees for their entire duration, as these are described in the corresponding SLSes, without causing any downgrade to the already established services. An additional objective of SLS Invocation Handling is to optimise the use of network resources.

Note that SLS Invocation Handling must be capable of achieving these two objectives under any offered traffic conditions in the context of the statistical and hard guarantees solution option. Loose SLS Invocation Handling is required for the loose guarantees solution option.

9.7.3 Interface specification

This section describes the interaction of the SLS Invocation Handling functional block with other functional blocks, as specified in the MESCAL functional architecture [D1.1], through events, messages or signals. Figure 75 shows the interfaces related with SLS Invocation Handling.

Copyright © MESCAL Consortium, June 2005

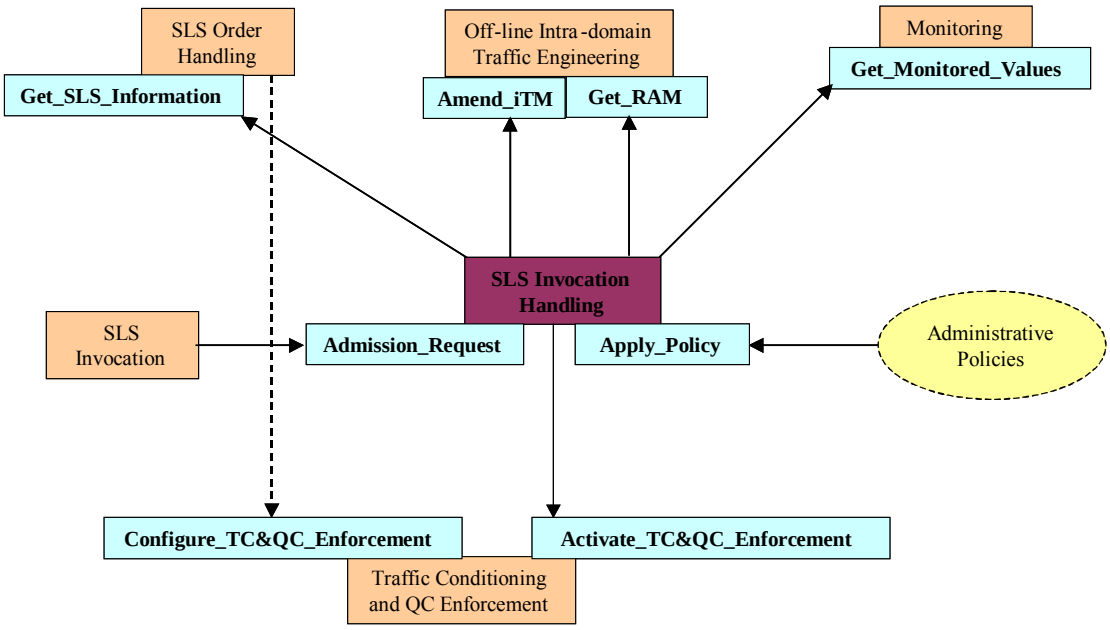


Figure 75 The SLS Invocation Handling interfaces.

SLS Invocation to SLS Invocation Handling

Admission_Request

This method will be called by SLS Invocation to initiate the SLS Invocation Handling process.

SLS Order Handling to SLS Invocation Handling

Get_SLS_Information

This method will be called by SLS Invocation Handling to request and get SLS information from the SLS Order Handling functional block regarding the number and type of the subscribed SLSes. The SLS subscription process should take into account the available resources, as these are expressed in the Resource Availability Matrix (RAM), output by the Off-line Intra-domain Traffic Engineering functional block.

SLS Order Handling to Traffic Conditioning and QC Enforcement

Configure_TC&QC_Enforcement

This method will be called by SLS Order Handling to configure the traffic conditioners and bind the l-QCs that will carry the traffic of the subscribed SLSes with the appropriate l-QC/o-QCs of the peering domain, if traffic needs to be carried to destinations that cannot be reached within the domain.

Off-line Intra-domain Traffic Engineering to SLS Invocation Handling

Get_RAM

This method will be called by SLS Invocation Handling to request and get the Resource Availability Matrix (RAM) from the Off-line Intra-Domain Traffic Engineering. RAM, which is derived from the internal RAM (iRAM) and the external RAM (eRAM) provides estimates of the available resources end-to-end for all employed QCs or Meta-QoS-Classes. The estimates are provided as a range of values allowing for potential resource sharing among the employed QCs or Meta-QoS-Classes.

Copyright © MESCAL Consortium, June 2005

Monitoring to SLS Invocation Handling

Get_Monitored_Values

This method will be called by SLS Invocation Handling to request and get information from Monitoring regarding the actual state of the network by means of real-time measurements. The measured entities can be bandwidth, one-way packet delay, packet delay variation and packet loss rate.

Administrative Policies to SLS Invocation Handling

Apply_Policy

This method will be called by the Administrative Policies block to apply predecided policies that will influence the SLS Invocation Handling decision-making process.

SLS Invocation Handling to Traffic Conditioning and QC Enforcement

Activate_TC&QC_Enforcement

This method will be called by SLS Invocation Handling to trigger the activation of the appropriate Traffic Conditioning and QC Enforcement following every successful SLS invocation request.

SLS Invocation Handling to Off-line Intra-domain Traffic Engineering

Amend_iTM

This method will be called by SLS Invocation Handling to trigger the recalculation of the internal Traffic Matrix (iTM), in case of excessive subscribed SLSes invocation rejections. This situation could occur when the available resources are oversubscribed, or the demands of the subscribed SLSes, in terms of bandwidth, are underestimated.

9.7.4 Behavioral specification

SLS Invocation Handling is performed at the network ingress routers and is responsible for accepting/rejecting SLS invocation requests on the behalf of an entire network domain or a sequence of domains, if the destination address of the traffic, points to another domain. The SLS Invocation process can be initiated either explicitly, e.g. through RSVP signalling, or implicitly. Upon receipt of an SLS Invocation request, the SLS Invocation Handling block will use the following inputs and derive the following outputs, as these are depicted in Figure 76.

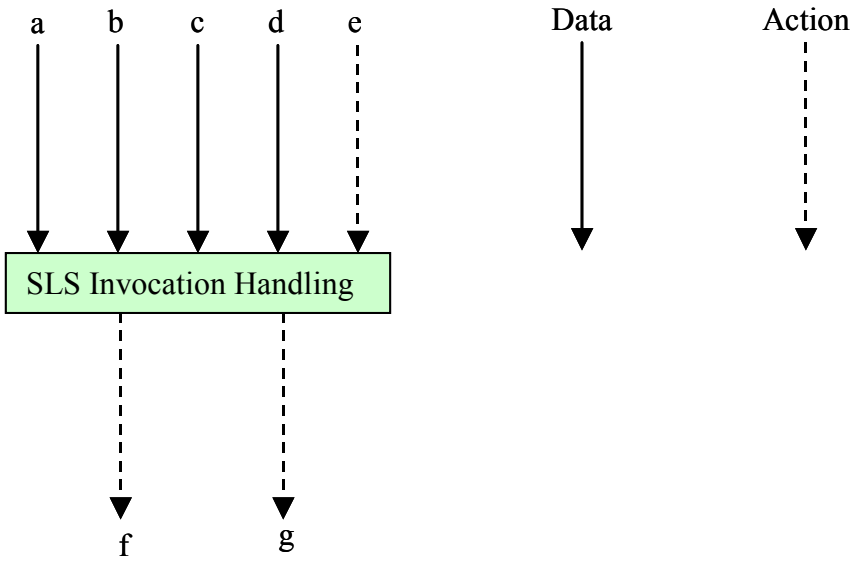


Figure 76 SLS Invocation Handling Inputs/Outputs.

Copyright © MESCAL Consortium, June 2005

9.7.4.1 Inputs/Outputs

Input data:

a. *SLS information*

SLSes may include the following:

Parameter Group	Description
Customer/user identifier	Identifies the customer or the user for Authentication, Authorisation and Accounting (AAA)
Flow descriptor	Identifies *the packet stream* of the contract by e.g. specifying a packet filter (DSCP, IP source address, etc).
Service Scope	Identifies the geographical region *where* the contract is applicable by e.g. specifying ingress and egress interfaces.
Service Schedule	Specifies *when* the contract is applicable by giving e.g. hours of the day, month, year
Traffic descriptor	Describes the traffic envelope through e.g. a token bucket, allowing identification of in- and out-of-profile packets
QoS Parameters	Specifies the QoS network guarantees offered by the network to the customer for in-profile packets including delay, jitter, packet loss and throughput guarantees.
Excess Treatment	Specifies the treatment of the out-of-profile packets at the network ingress edge including dropping, shaping and re-marking.

b. RAM

RAM should give an estimate of the available resources for the various reachable destination prefixes (end-to-end) for all employed QCs or Meta-QoS-Classes. The estimates could be expressed as a range of bandwidth values, taking into account potential resource sharing between classes and intra/inter-domain reachable destination prefixes.

c. Monitored values

Monitored values are needed for measurement-based SLS Invocation Handling approaches. They depict the network condition in real-time and can involve bandwidth, packet loss, delay and delay variation measurements.

d. Number and type of already invoked SLSes

Model-based SLS Invocation Handling processes require knowledge of the number and type of established services at all times. This means that services must signal, apart from their initiation, their termination. If this is not feasible, an alternative is to employ a time-out period, based on the activity of a service, as an indication of its termination.

Input Actions:

e. Policies

Predefined policies that influence the SLS Invocation Handling decision. Policies can determine facts, such as the level of conservativeness the administrator is willing to enforce with respect to the satisfaction of the predefined performance objectives, the preferential treatment of some applications (e.g. real-time) compared to others, and the treatment of invocations that don't correspond to subscribed services.

Copyright © MESCAL Consortium, June 2005

Output actions:

f. Traffic Conditioning and QC enforcement

Upon an SLS Invocation admission, at ingress routers, actions such as packet classification, policing, shaping and DSCP marking, according to the conditions laid out in the previously agreed SLSes, are taken. QC enforcement is responsible for implementing the binding of the employed l-QC to the service peer o-QC, in case the destination prefix of the service requires inter-domain transition.

g. Traffic Matrix Amendment

Excessive service rejection rates may indicate oversubscription of the resources or underestimation of the bandwidth requirements of the subscribed SLSes. In all cases, an iTM recalculation might be needed for the next Resource Provisioning Cycle.

9.7.4.2 Process Description

Upon an SLS Invocation request, the SLS Invocation Handling will use SLS information to check whether the initiating user, either customer (cSLS) or domain (pSLS), is authorised to request the specific service. If there exists a subscription for the requested service, the SLS Invocation request will always be considered for admission. If there does not exist a subscription for the requested service, then policies will determine whether the request will be further considered or immediately dropped. The applied policies will take into account facts, such as the type of traffic the service request will send and the current state of the network (e.g. if the available resources for the QC that will be employed for carrying the traffic of the requested service are more than a threshold, then consider, otherwise reject) If a request is decided to be considered for admission, through monitoring (measurement-based approach) and/or by using traffic descriptors (model-based approach), and by taking into account the information of the RAM, the available resources for the o-QC or Meta-QoS-Class that will be used to carry the traffic injected by the service are estimated. Another approach for indirectly estimating the available resources is by sending a stream of probing packets end-to-end (endpoint approach). If the resources are adequate to support the service, it is admitted and the appropriate Traffic Conditioning and QC enforcement actions are triggered. If the resources are not adequate, the service request is rejected. Note, that policy reasons may require a service request to be rejected even if the available resources are adequate. The service rejection rate, especially of requests corresponding to subscribed SLSes, needs to be maintained, since excessive values may indicate traffic engineering and oversubscription problems.

9.7.5 SLS Invocation handling issues

9.7.5.1 Case Studies

SLS Invocation Handling will need to consider and discriminate the following options, and different approaches and algorithms might be required for each possible case:

Real-time traffic vs Elastic traffic

Real-time (UDP controlled) and elastic traffic (TCP controlled) have different QoS requirements and exhibit different traffic patterns. Furthermore, the invocation of elastic traffic flows is mainly implicit (HTTP traffic) whereas for real-time traffic is explicit. Therefore, an SLS Invocation Handling approach that suits one type of traffic may not achieve satisfactory performance for the other type of traffic.

Peak rate allocation vs Statistical multiplexing allocation

In the first approach, for bandwidth allocation, each SLS is allocated bandwidth equal to its declared peak rate whereas in the second approach we allow for resource sharing between the SLSes belonging to the same o-QC/Meta-QoS-Class. The first approach can provide harder QoS guarantees but can lead to poor network utilisation.

Copyright © MESCAL Consortium, June 2005

cSLSes vs pSLSes

cSLSes and pSLSes may differ greatly regarding the aggregation level and the characteristics of traffic. Therefore, an SLS Invocation Handling approach that suits one type of SLSes may not achieve satisfactory performance for the other type of SLSes.

Interactions between classes

SLS Invocation Handling needs to take into account potential interactions between different classes in terms of resource sharing and relative priority in terms of scheduling.

Overbooking vs Non-Overbooking

The two cases may require different treatment depending on the employed SLS Invocation Handling Approach for providing the same QoS guarantees for the invoked and admitted service requests.

Endpoint vs Measurement-based vs Traffic descriptor-based approaches for determining available resources

The employed SLS Invocation Handling algorithms will be greatly determined by the approach used to estimate the available resources. For the endpoint approach, the estimation is based on the calculation of some metrics on streams of probing packets. For the measurement-based approach, the estimation of resources is based on real-time measurements of the actual network traffic, whereas for traffic descriptor-based approaches the estimation relies totally on the declared traffic descriptors included in the SLSes.

Bi-directionality issues

SLS Invocation Handling for bi-directional services, such as Voice-over-IP will need to take into account not only the state of the forward path, that is the end-to-end path that will be used by the SLS originated traffic, but additionally the state of the return path, that is the end-to-end path for the receiver initiated traffic.

9.7.5.2 Examples

With respect to the aforementioned cases, the SLS Invocation Handling approaches as presented in [D1.4] and [GEOR04, GEOR05] address the following:

In [D1.4]:

- Real-time traffic
- Peak rate allocations
- cSLSes
- No interactions between classes
- Overbooking
- Measurement-based approach (based on congestion indications –green/red states)
- Bi-directionality not taken into account

In [GEOR04, GEOR2]:

- Real-time traffic
- Statistical multiplexing allocations
- cSLSes
- No interactions between classes
- Overbooking
- Combined Measurement and Traffic Descriptor-based approach
- Bi-directionality not taken into account

Copyright © MESCAL Consortium, June 2005

9.7.6 SLS Invocation Handling Algorithms

In general we can separate the SLS traffic based on its responsiveness to congestion. The categorisation, which is widely accepted today, is that between elastic and non-elastic traffic. The latter is also known as real-time traffic. Examples of elastic traffic sources are sources that use TCP as the transport protocol, while real-time traffic sources are applications that use UDP without any congestion control at the application level.

In the following we will look into algorithms for admission control of real-time and elastic traffic sources.

9.7.6.1 Intra-domain Real-time Traffic Admission Control

In this section we consider intra-domain admission control for real-time traffic. We define as real-time traffic sources, the ones which have a strict small delay requirement and a bounded, not necessarily too low, packet loss rate (PLR) requirement. In a Diffserv domain, the PHB used for this traffic will be the Expedited Forwarding (EF). We assume that such traffic will be aggregated to form one or more real-time traffic aggregates, and that the traffic from the sources that composes each traffic aggregate will receive the same treatment over the entire domain. The delay requirement of the traffic aggregate can be taken into account in the provisioning stage, i.e. by appropriately setting small queues and by manipulating the routing process to choose appropriate paths. Our assumption related to packet loss, is that packets are expected to be lost only at the first point of aggregation (ingress link), which, according to [AKE03], is currently considered as the most probable congestion point (bottleneck link) of a domain. We assume that further downstream inside the domain, real-time traffic aggregates are provisioned in a peak rate manner. This is feasible since, as stated in [IANN01], in a common network configuration, backbone links are over-provisioned. Low jitter is also a requirement for real-time traffic, but according to [BON], jitter can remain controlled in successive multiplexing queues as long as the flows are shaped to their nominal peak rate at the network ingress. Furthermore, the deployment of non-work conserving scheduling in routers for the EF PHB can be beneficial for controlling jitter [MOW98].

We also assume that the interior of the Diffserv domain has been provisioned and engineered in this way in order to support the real-time traffic aggregates. As a result of the provisioning process, and taking into account the routing behaviour, at each ingress node, we can have an estimate of the minimum bandwidth available for the real-time traffic aggregate from that ingress to each of the corresponding egress nodes. This available bandwidth is the basis for our admission control scheme, which is employed at the edge (ingress) node of the first Diffserv aggregation point for accepting a traffic source on behalf of the entire network domain. Our assumptions imply that our admission control scheme does not induce any states in the core network, which is desired for scalability and resilience reasons, and it is also proven to be a resource-efficient approach if resilience against network failures is required [MEN04].

We will now present our Admission Control scheme, which is a combination of Measurement-based and *a priori* Traffic descriptor Admission Control -we will be referring to it as MTAC. We use real-time measurements of the actual load and also we employ the traffic descriptors of a *reference source* model in order to account for traffic heterogeneity. Our scheme is applicable to real-time sources that are able to provide even only a single traffic descriptor, their peak rate. Given the diversity of Internet-based applications that have real-time requirements, the use of more complex traffic descriptors in admission control, as stated in [FLO96], to accurately characterise source traffic, is neither necessary nor plausible. Therefore, we assume that the only available traffic descriptor to use is the source's peak rate. This traffic descriptor is easy to police and even if not available, for sources described by a token bucket filter (r,b) an estimate $\hat{p}$ of it can be derived [FLO] using the equation:

$$\hat{p} = r + b/U \quad (1)$$

where U is a user-defined averaging period.

Copyright © MESCAL Consortium, June 2005

Our scheme uses the bufferless statistical multiplexing approach. Bufferless multiplexing is very attractive for real-time traffic since it ensures that the traffic experiences minimal delay. In addition, the dynamics leading to an overload event in a bufferless system are much simpler than those of a buffered system [TSE97]. The main disadvantage of using a buffer is that overflow probability depends significantly on flow characteristics [ROB98] and can only be tightly controlled if these characteristics are known. Moreover, in this case, provisioning needs to account for statistical variations in the traffic mix as new flows arrive and others terminate. On the other hand, buffered multiplexing allows higher utilisation for the same loss rate [ROB98] but, as stated above, requires more complex traffic management and is not as robust with respect to flow characteristics as bufferless multiplexing. We need to stress that bufferless multiplexing, is, of course, just an abstraction [ROB98]. For packetised traffic, as in IP networks, a small buffer for packet scale queuing is needed to take into account coincident packet arrivals from distinct flows [BON].

According to [GUE], when the effect of statistical multiplexing is significant, the distribution of the stationary bit rate can be accurately approximated by a Gaussian distribution. In [SHROF] it is suggested that the aggregation of even a fairly small number of traffic streams is usually sufficient for the Gaussian characterisation of the input process. In that case, the effective bandwidth of the multiplexed sources is given by:

$$C \approx m + \alpha'\sigma \quad with \quad \alpha' = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)} \quad (2)$$

where m is the mean aggregate bit rate, σ is the standard deviation of the aggregate bit rate and ε is the upper bound on allowed loss probability.

9.7.6.1.1 SLS Invocation Handling Logic

In a Diffserv domain we assume that the real-time traffic aggregate is provisioned and engineered in such a way that at minimum C_{total} bandwidth is available edge-to-edge for the l-QC that will be used to carry the traffic of the real-time traffic aggregate. Every time a source wants to establish a service instance, it signals this to the ingress node through some resource reservation protocol. A similar assumption can be made for the service termination. If the latter is not explicitly signalled, an alternative option could be to use a time-out period as an indication of the service termination. In any case, at each point in time, the MTAC process at each ingress point knows the number of active sources.

When a new service request arrives, we need to decide whether or not to allow the source to send traffic using the real-time traffic aggregate resources until the known egress point. Initially, we need to calculate an appropriate time period, the *measurement window*, within which we need to take and use measurements for bandwidth usage estimations. The measured parameters are the mean rate of the offered load, $M_{measured}$, and the variance of the offered load, $\sigma^2_{measured}$, at the output queue of the ingress node. Having the measurements and the peak rate p_{new} of the new source, and by making the worst case assumption that the new source will be transmitting at its peak rate, we compute the estimated bandwidth C_{est} as follows:

$$C_{est} = M_{measured} + p_{new} + \alpha'_{PLR}\sqrt{\sigma^2_{measured}} \quad (3)$$

where α'_{PLR} is computed as in (2), based on the target PLR bound of the real-time traffic aggregate. This value C_{est} is the estimated bandwidth used in the admission control criterion.

9.7.6.1.2 Measurement Window Estimation

We define the measurement window w, as the time interval within which the offered load is taken into account for deriving the required measurements. In a similar fashion to [BELEN], we use the following expression for the measurement window:

$$w = \max(DTS, w') \quad (4)$$

Copyright © MESCAL Consortium, June 2005

In (4), DTS represents the Dominant Time Scale. DTS is the most probable time scale over which overflow occurs. In [SHROF], the authors describe a systematic way to derive DTS using real-time measurements, with the assumption that the input process to the multiplexing point in the network is Gaussian. This is by definition our assumption when employing (2), therefore we use this method in order to estimate the DTS. DTS, as computed in [SHROF], is a function of the mean rate, the variance of the offered load and the output buffer size. The reader should recall that even though we employ the bufferless multiplexing approach, a small output buffer is still required for packet scale queuing, as explained in the previous section. This value for the output buffer is involved in the estimation of the DTS.

Let w' represent the mean inter-departure delay [GROS1], defined as follows (Little's formula):

$$w' = \frac{h_{avg}}{N_{active}} \quad (5)$$

where N_{active} is the number of simultaneously active sources and h_{avg} is their average duration.

Since we assume that the service establishment and termination is signalled to the ingress nodes, the average duration of the sources can be easily obtained and updated.

That is we select as measurement window the mean inter-departure delay, i.e., the time interval within which the system can be considered stationary -no flow departures-, unless this time interval is not long enough to capture the time-scale fluctuations of the aggregate traffic stream. This can happen in case of long-range dependent traffic. In this case and in order to enable the network to react to these traffic fluctuations, we use DTS as the value of the measurement window.

9.7.6.1.3 The Precaution Factor (PF)

Before deriving the admission control criterion, there are two important issues that challenge the effectiveness of any admission control scheme and, therefore, need to be taken into account:

(a) the traffic source heterogeneity, and

(b) the effect of measurement errors.

In order to account for these two issues, we introduce *a precaution factor (PF)*, which we involve in the admission control criterion and we give a heuristic formula for *PF* with which we address these two important issues.

As mentioned, the first issue is that the aggregate traffic stream might have characteristics that do not suit the effective bandwidth formula (2). This, for instance, can happen if the stream is composed of a small number of very bursty connections with high peak rates and low utilisations [GUE].

To account for this, we use an exponential ON/OFF source, with mean and standard deviation (m_{ref}, σ_{ref}) as a model source for engineering reasons (*reference source*). The reason for the specific selection is that exponential ON/OFF sources are representative models for VoIP traffic, which is likely to be a big part of the traffic carried by real-time traffic aggregates and their traffic characteristics suit the effective bandwidth formula (2). Furthermore, exponential ON/OFF sources are short-range dependent, which means that their traffic characteristics are more easily captured within the given measurement window. We define as *reference trunks* (T_{ref}) the number of simultaneously established reference sources that can fit in C_{total}, according to (2), for a given bound on packet loss rate.

When a new request arrives, having measured the mean rate $M_{measured}$ and the variance $\sigma^2_{measured}$ of the offered load, we calculate the number N_m of the reference sources, whose aggregate mean rate is equal to or greater than $M_{measured}$. We also calculate the number N_σ of the reference sources, whose

Copyright © MESCAL Consortium, June 2005

aggregate variance is equal to or greater than $\sigma^2_{measured}$. That is, N_m and N_σ satisfy the following relationships:

$$N_m = \left\lceil \frac{M_{measured}}{m_{ref}} \right\rceil and\ N_\sigma = \left\lceil \frac{\sigma^2_{measured}}{\sigma^2_{ref}} \right\rceil \quad (6)$$

Having estimated N_m and N_σ, we compute their mean value N_{ref}:

$$N_{ref} = (N_m + N_\sigma)/2 \quad (7)$$

This value represents a rough estimate of the number of reference sources that produce, within the measurement window, load with characteristics (mean rate and variance) similar to the ones measured. To compensate for the above, we set *PF* to be proportional to the quantity (N_{ref} / T_{ref}).

The second issue that needs to be taken into account with the precaution factor is the effect of measurement errors. As shown in [GROS1], the *certainty equivalence* assumption, i.e., that the measured parameters represent the real traffic, can heavily compromise the performance of an admission control scheme. The stringent the PLR requirement, the easier it is to violate it due to measurement errors. In the case where only aggregate bandwidth information is available through measurements, as in our scheme, the degradation in performance can be mainly attributed to errors in the estimation of the variance [GROS2]. With non-negligible probability the variance can be significantly underestimated. To compensate for the measurement uncertainty, we proceed as follows: given (2), for a specific target PLR level ε, we set *PF* to be proportional to the quantity $\frac{\sqrt{-2\ln(\varepsilon) - \ln(2\pi)}}{\sqrt{-2\ln(\varepsilon_{ref}) - \ln(2\pi)}}$.

That is, we inflate the part of equation (2) that relates to the variance estimation, based on a reference PLR level. By setting ε_{ref} to be larger than ε, we ensure that the more stringent the PLR requirement, the greater the value of this quantity. This reference PLR can be set by policy to adjust the conservativeness of the MTAC scheme.

Combining the two aforementioned quantities, the final expression for the precaution factor that can be used is:

$$PF = (N_{ref} / T_{ref}) * \frac{\sqrt{-2\ln(\varepsilon) - \ln(2\pi)}}{-2\ln(\varepsilon_{ref}) - \ln(2\pi)} \quad (8)$$

We also set $PF = 1$ whenever the equation above results to be less than 1. That means that we use PF in a conservative way in the admission control criterion. The precaution factor can be considered as a tuning parameter. Even though we derive *PF* somehow heuristically, based on intuition rather than strict mathematical analysis, one should take into account that all admission control schemes employ additional admission tuning parameters [GROS2, BRES00] because it is not possible to completely decouple performance from traffic characteristics.

9.7.6.1.4 The Admission Control Criterion

Given the allocated bandwidth for the real-time traffic aggregate from edge-to-edge is C_{total}, and having computed the value for C_{est}, employing the precaution factor and the measurement window, the admission control criterion becomes:

$$\begin{aligned} &If\ \ (C_{est} \times PF) \le C_{total,}\ \ admit \\ &If\ \ (C_{est} \times PF) > C_{total,}\ \ reject \end{aligned} \quad (9)$$

Copyright © MESCAL Consortium, June 2005

9.7.6.1.5 Scalability issues

MTAC does not raise any major scalability issues. Per-flow state information and signalling is only required at ingress nodes, where the number of flows is relatively small, and the core nodes need only to perform forwarding, not violating, therefore, the basic Diffserv paradigm. No feedback information is required from the core nodes since MTAC is based on provisioned Diffserv information. Therefore, the core nodes can be totally unaware of any kind of signalling.

Furthermore, MTAC is completely distributed. That means that one instance MTAC runs at each ingress node, serving real-time time traffic from that ingress node to the corresponding egress node, independently from the MTAC instances running at other ingress nodes. That means that no coordination between the ingress nodes is required and the performance, regarding scalability, of MTAC is independent of the intra-domain topology.

9.7.6.2 Inter-domain Real-time Traffic Admission Control

For the inter-domain traffic case, since peering links at the border routers between neighbouring domains are often bottlenecks [AKE, RAS, AKA99] and also the source of some of the greatest costs for network operators [MON02], they cannot be considered over-provisioned. Therefore, admission control also needs to take into account the state of these links before deriving the admission control decision.

In order to do so, we extend our MTAC scheme to take into account the state of these links –we will be referring to it as e-MTAC (extended MTAC). We make the assumption that the status information of the peering inter-domain links can be conveyed upon request to the corresponding ingress nodes, where the admission control decision is made.

Since we assume that the core is over-provisioned, packet losses are expected to occur only at the ingress links and at the peering (inter-domain) links.

As it can be easily easily proved [LIMA04], packet loss rate parameters are multiplicative. That means that for a set of sources that traverse a sequence of links, $l_i, i = 1,...,N$ with packet loss rates PLR_i, the total packet loss rate PLR_{total} can be approximated using the expression:

$$PLR_{total} = 1 - \tilde{\prod}_{i=1}^{N}(1 - PLR_i) \quad (10)$$

which, in turn becomes additive for low values of PLR_i:

$$PLR_{total} = \sum_{i=1}^{N} PLR_i \quad (11)$$

For our case, this means that the total packet loss rate for the inter-domain real-time traffic sources is:

$$PLR_{total} = PLR_{ingress} + PLR_{egress} \quad (12)$$

where $PLR_{ingress}$ is the allowed packet loss rate at the corresponding ingress link and PLR_{egress} is the allowed packet loss rate at the corresponding egress link. The value of PLR_{total} should be set so that when combined with the packet loss rate value of the o-QC from the neighbour downstream domain it can satisfy the end-to-end packet loss requirement of the real-time traffic flows.

The functionality of e-MTAC for inter-domain traffic, taking into account the above discussion can be divided on two parts; one related to the first-hop ingress link and one related to the peering inter-domain link.

Copyright © MESCAL Consortium, June 2005

9.7.6.2.1 First-hop Ingress Link

The functionality is as described in section 9.7.6.1. The difference is that the packet loss rate involved in equations (3) and (8) is this case is $PLR_{ingress}$ and the demanded bandwidth value inside the domain derived by e-MTAC is: $C_{dem\_ingress} = C_{est\_ingress} * PF_{ingress}$.

9.7.6.2.2 Inter-domain Link

The functionality is as described in section 9.7.6.1. The main difference relates to the expression of the measurement window for the inter-domain link. Since per-flow state is not kept at egress nodes we cannot involve equation (5) in the expression for the measurement window. Therefore for the inter-domain link the expression used for the measurement window *w* is:

$$w = DTS \quad (13)$$

Also the packet loss rate involved in equations (3) and (8) is this case is PLR_{egress} and the demanded bandwidth value on the inter-domain link derived by e-MTAC is: $C_{dem\_egress} = C_{est\_egress} * PF_{egress}$.

9.7.6.2.3 The Admission Control Criterion

Given the allocated bandwidth for the inter-domain real-time traffic aggregate from edge-to-edge inside the domain is $C_{total\_ingress}$ and the shared allocated bandwidth for the inter-domain real-time traffic aggregates on the inter-domain link is C_{total_egress}, the admission control criterion becomes:

$$\begin{array}{l} If \quad (C_{dem\_ingress} \pounds C_{total\_ingress}) \; and \; (C_{dem\_egress} \pounds C_{total\_egress}), \; admit \\ \qquad\qquad otherwise, \quad reject \end{array} \quad (14)$$

9.7.6.2.4 Scalability Issues

Similar to MTAC, e-MTAC does not raise any major scalability issues. Per-flow state information and signalling is required at ingress nodes and the core nodes only perform forwarding. No feedback information is required from the core nodes. Therefore, the core nodes can be totally unaware of any kind of signalling. In addition to ingress nodes, egress nodes also need to be aware of signalling in order to convey the status information of the peering inter-domain links to the corresponding ingress nodes, where the admission control decision is made.

e-MTAC is also distributed and does not require any coordination between ingress nodes. It requires only coordination between the ingress node where an instance of e-MTAC runs and the corresponding egress nodes through which the inter-domain traffic –entering through this specific ingress node- exits the domain. Its performance, regarding scalability, is independent of the intra-domain topology and depends only on the inter-domain link topology.

9.7.6.3 Elastic Traffic

The algorithm proposed for dealing with SLS Invocation Handling for real-time traffic cannot be applied to elastic traffic because the traffic pattern of TCP controlled traffic deviates greatly from the Gaussian assumption. For elastic traffic new algorithms must be derived taking into account previous work in this field, as in [FRED01], [PRATT], [ROB2], [MAS], [JOVE], [CHAIT], [CHARZ].

Copyright © MESCAL Consortium, June 2005

10 TRAFFIC ENGINEERING

10.1 Inter-domain TE terminology

10.1.1 Introduction

The purpose of this section is to clarify some of the technical terms involved in the description of the interactions between the various MESCAL functional blocks.

Figure 77 provides a graphical representation of the possible types of traffic, with respect to origin and destination, that may traverse an AS. It provides a basis for discussing and defining technical terms in the following Section.

The four types of traffic shown are as follows:

- (1) Traffic both originating and terminating in AS1;
- (2) Traffic originating in AS1 and terminating in a downstream AS;
- (3) Traffic originating in an upstream AS and terminating in a downstream AS;
- (4) Traffic originating in an upstream AS and terminating in AS1.

We define traffic that is originated and terminated within the same domain as *intra-domain traffic* (cases (1) and (4)), while traffic that terminates at a remote (downstream) domain is *inter-domain traffic* (cases (2) and (3)).

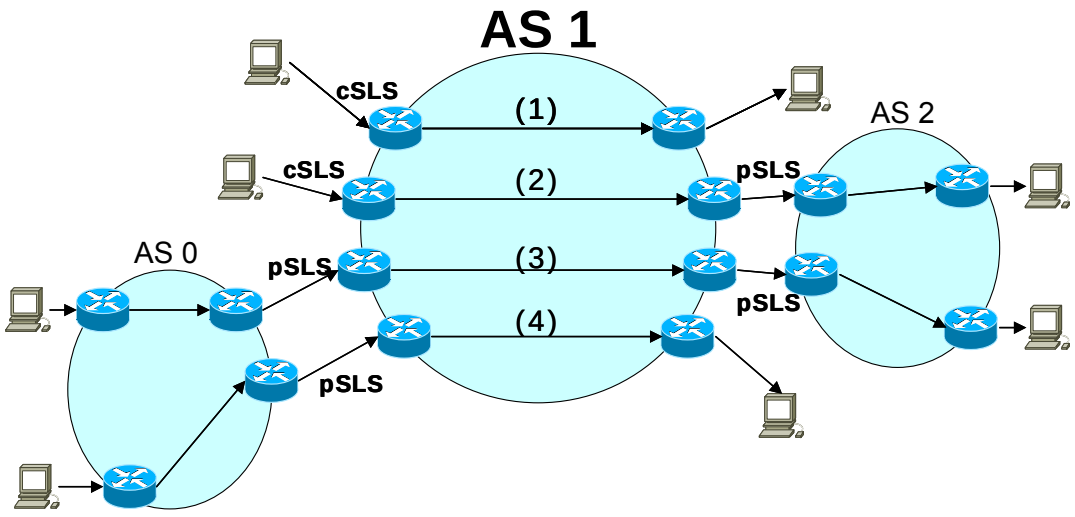


Figure 77. Traffic origination/destination cases

10.1.2 Definitions

10.1.2.1 $pSLS_{in}$ and $pSLS_{out}$

For a specific AS A:

- $pSLS_{in}$ denotes a pSLS that the domain A provides (or offers) to another AS B. That means that the domain A has received a request from domain B for the establishment of a pSLS.

Copyright © MESCAL Consortium, June 2005

- $pSLS_{out}$ denotes a pSLS that is provided to domain A by another domain B. That means that domain A has requested the establishment of a pSLS with domain B.

These terms are illustrated in Figure 78.

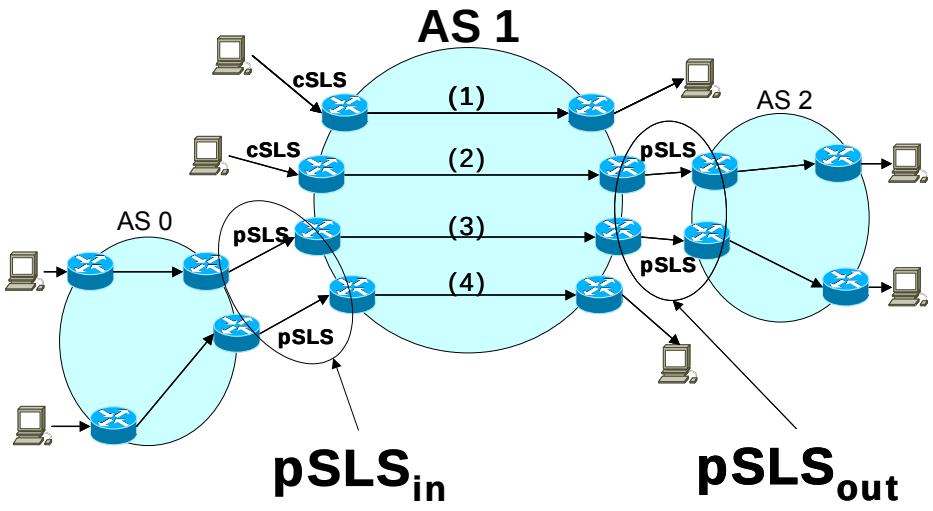


Figure 78. The $pSLS_{in}$ and $pSLS_{out}$ sets in a domain

10.1.2.2 eTM and iTM

For a specific AS A:

- eTM (external Traffic Matrix) denotes the traffic matrix that depicts the bandwidth requirements for all employed QCs, from each ingress interface (border router) to all the destination prefixes outside domain A. That is:

  eTM entry = [ingress i/f, {destination prefixes}∉A, bw(min,max), e-QC]

  Considering Figure 78, the eTM represents the aggregate bandwidth requirements for traffic shown in types (2) and (3). Note that in the definition, {} represents the *set-of* semantics.

- iTM (internal Traffic Matrix) denotes the traffic matrix that depicts the bandwidth requirements for all employed QCs, from each ingress interface to all egress interfaces of domain A. The egress interfaces need not be the termination points of the traffic. That is:

  iTM entry = [ingress i/f, {egress i/f }∈A, bw(min,max), l-QC]

  Considering Figure 78, iTM represents the aggregate bandwidth requirements for traffic shown in all types (1)-(4).

10.1.2.3 eRAM, iRAM and RAM

For a specific AS A:

- eRAM (Resource Availability Matrix) represents the available resources, the bandwidth buffer (bw_buffer), for all QCs for all destination prefixes outside the domain A. eRAM must also specify the egress interface(s) (inter-domain links) that are used for providing these resources and additionally either the splitting ratio of the traffic on these egress interfaces or the mapping of the reachable destination prefixes on these egress interfaces. That is:

  eRAM entry = [ingress i/f, {egress i/f}+ratio, {destination prefixes}∉A, bw_buffer, {l-QC}, e-QC]

Copyright © MESCAL Consortium, June 2005

- iRAM represents the available resources (bw_buffer) for all (local) QCs, from each ingress interface to all egress interfaces of domain A. The egress interfaces can be both termination points and transit points of traffic. That is:

  iRAM entry = [ingress i/f, {egress i/f}+ratio, bw_buffer, l-QC]

The term "bandwidth buffer" is defined below.

10.1.2.4 Bandwidth buffer

The bandwidth buffer, bw_buffer, is the bandwidth allocation computed by the traffic engineering processes. It is defined as the partitioning of the total available bandwidth between some *end points* for a given *QoS class*. The partitioning of the available bandwidth for a single QoS class reflects the provisioning decisions for that class. The bandwidth partitioning clearly defines the limits for admission control. The number of partitions depends on the policies used within the traffic engineering algorithms. In a simple scenario the total available bandwidth for a QoS class might be partitioned into two ranges, one that reflects the maximum bandwidth requirements of the currently subscribed SLSs, and the other that corresponds to either the provisioned bandwidth for future SLS subscriptions, or any overbooking ratios. For example a given bw_buffer=(10Mbps, 15Mbps) means that the total provisioned bandwidth for a particular QoS class is 15Mbps, from which 10Mbps corresponds to the requirements of currently subscribed flows.

The definition above assumes the bw_buffer is for a single *QoS class* between two *end points*. In the case where the bw_buffer is within an eRAM, the QoS class is the e-QC, and the end-points are the ingress interface and destination prefixes, of the corresponding eRAM entry. In the case where the bw_buffer is within an iRAM, the QoS class is the l-QC, while the end points are the ingress and the egress interfaces of the corresponding iRAM entry.

10.2 Traffic Forecast

10.2.1 Objectives

The main objectives of *Traffic Forecast* (TF) are:

- To forecast QoS traffic demand, based on existing and anticipated subscriptions, c/pSLS, related historical data combining subscriptions and network usage, and related business policies e.g. sale targets. This is required for the traffic engineering (TE) functions to appropriately dimension the network in terms of required intra- and inter-domain resources.
- To assess the validity of the forecasted traffic demand against actual usage statistics and based on various alarms coming from the service management and traffic engineering functions, and determine the cases where the current forecasts are no longer valid, triggering the process of revising them.
- To support the required interactions at RPC (resource provisioning cycles) epochs with the traffic engineering functions.

The main outcome of TF, of interest to MESCAL, is the production of the so-called:

- *Internal Traffic Matrix (iTM)*, presenting forecasted QoS traffic demand between network ingress and egress interfaces (border routers) of the domain, and the
- *External Traffic Matrix (eTM)*, presenting forecasted QoS traffic demand between network ingress interfaces and destination prefixes outside the domain.

Traffic demand is expressed in terms of bandwidth units. For scalability reasons, the QoS traffic needs to be of aggregate nature. The iTM and eTM are specified from TE perspectives in section 10.1.2.2.

Copyright © MESCAL Consortium, June 2005

10.2.2 Interface Specification

The interfaces of Traffic Forecast are shown in Figure 79 and briefly described in the following.

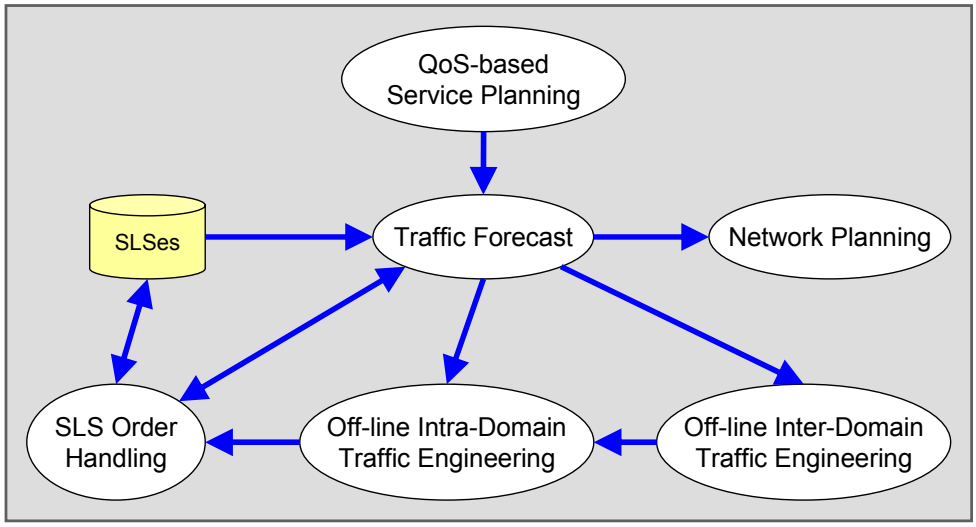


Figure 79. Traffic Forecast

Traffic Forecast receives per service offering the anticipated QoS traffic demand from *QoS-based Service Planning*, as seen by business perspectives, e.g. as a result of sale targets or newly launched marketing campaigns. In particular, the following information is passed:

- Traffic demand for currently supported and envisaged services between particular sets of source and destination groups (see section 9.4.2.2). Traffic demand in this case, may either be expressed directly, as a specific amount of bandwidth, or indirectly, as the number of new expected customers.
- Policy parameters related to the functionality of the Traffic Forecast process.

Traffic Forecast receives from *SLS Order Handling* information on negotiated services; specifically:

- Established service agreements, c/pSLSes during the current and previous RPCs.
- Negotiation logs during the current RPC and previous RPCs.
- Various alarms on thresholds defined on the rate at which, service agreements are requested and established and related historical data.

Traffic Forecast provides to the *Network Planning* and *Traffic Engineering* components the traffic matrices it produces. This is to the end of ensuring that the local and inter-domain resources will be planned and engineered so that to effectively and gracefully accommodate established c/pSLSes as well as those anticipated to be ordered during the current provisioning cycle.

Network Planning aspects are outside the scope of MESCAL investigation. With respect to the two traffic matrices of interest to MESCAL outlined in the previous section, *Traffic Forecast* outputs the eTM to the *Off-line Inter-domain Traffic Engineering* component and the iTM to the *Off-line Intra-domain Traffic Engineering* component. While eTM calculation is solely based on the existing and anticipated population of c/pSLS subscriptions and related historical data, the calculation of the iTM is additionally based and bound on the outcome of the *Off-line Inter-domain Traffic Engineering* component. In fact, this interaction is of an iterative nature, internal to the traffic engineering algorithms (hence, not shown in the figure). All these interactions occur at RPC epochs.

Last, *Traffic Forecast* interacts with the network monitoring services to retrieve appropriate usage statistics to the end of validating its current forecast. This kind of interactions is not shown in Figure 79; network monitoring is outside the scope of MESCAL.

Copyright © MESCAL Consortium, June 2005

10.2.3 Behavioural Specification

10.2.3.1 Functional Decomposition

Figure 80 presents the internal functional architecture of the *Traffic Forecast* component, together with the interactions with the 'rest of the world' identified in the previous section.

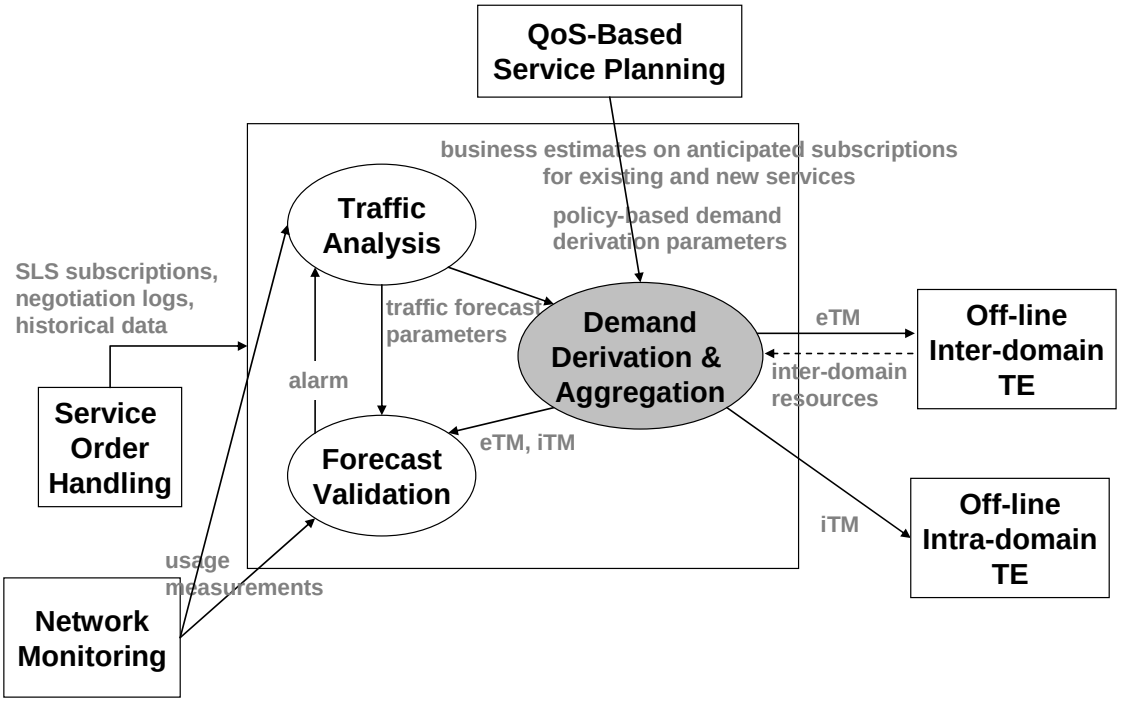


Figure 80. Functional decomposition of Traffic Forecast

The *Traffic Analysis* function is responsible for calculating the so-called *traffic forecast parameters*, which will allow the Demand Derivation and Aggregation functions to calculate the required traffic matrices. This component is based on historical data regarding service requests/subscriptions and network usage. By employing appropriate statistical test and inference methods, together with ad-hop means, it determines the traffic forecast parameters, which correlate static aspects of the traffic implied by a population of service subscriptions, to traffic volumes offered to the network. The traffic forecast parameters include: the different service classes that could be distinguished from traffic forecast perspectives and associated multiplexing factors and aggregation weights; defined in section 10.2.3.2.1. These notions have been commonly used since the traditional telecom business world.

The *Demand Derivation and Aggregation* functions are responsible for deriving the traffic demand implied by a specific population of service subscriptions, based on the forecast parameters produced by the Traffic Analysis function. Specifically, service subscriptions are classified into the corresponding service classes determined by Traffic Analysis. Then, Demand Derivation function calculates the expected demand per QoS class for each service, subsequently aggregated by the Demand Aggregation, based on multiplexing factors and aggregation weights specified by the Traffic Analysis component. Note that this process is linear, thanks to the semantics of the forecast parameters.

The *Forecast Validation* function is responsible for assessing the validity of the traffic matrices produced by Demand Derivation and Aggregation. Validity is checked against actual traffic developments, by comparing the forecasted demand as specified in the traffic matrices with suitable extrapolations, in the medium to long term, derived by actual measurements of offered QoS traffic per service class. Furthermore, the component utilises feedback from the service order handling function regarding trends in the rate of service offer requests and established agreements. It also utilises

Copyright © MESCAL Consortium, June 2005

feedback from the traffic engineering functions regarding significant under-utilisation or overloading of the resources of the dimensioned (against the traffic matrix under test) network.

MESCAL focuses only on the Demand Derivation and Aggregation functions. The aspects involved in the other two components fall outside the scope of investigation of the project.

10.2.3.2 Demand Derivation and Aggregation

10.2.3.2.1 Notions and Terminology

The notation introduced in section 9.4.2.1 will be used in the following. The traffic forecast parameters considered are:

- *Service Class (SC)*: The notion of service classes has been introduced to cope with the user diversity in service usage. Service classes distinguish offered services, c/pSLSes, based on their technical characteristics (e.g. invocation method, topological scope) according to the levels of statistical convergence observed in their usage patterns. As such, given a certain service class, it is considered that the users of the c/pSLSes of this class have the same service usage habits; hence valid multiplexing factors (see below) could be determined.
- *Multiplexing Factor (MF)*: For a given service class, a multiplexing factor is defined to be a proportion that consistently can relate the total subscribed traffic demand of a subscription population to the actual traffic peak that this population offers to the network. The validity of multiplexing factors should have been statistically verified over multiple observation periods involving subscription populations of different size. Given the variability in Internet service users, we take that a unique multiplexing factor per service class cannot be safely estimated; therefore, we assume two values for safely specifying a multiplexing factor: a minimum and a maximum, denoted by *Mfmin*, *Mfmax*, respectively.
- *Aggregation Weight (AW)*: For a given service class, the aggregation weight is the relative contribution of the service class's actual traffic peak to the peak of the corresponding offered total traffic. This notion is necessitated by the fact that the BHT (busy hour time, the period where traffic peak happened) of different service classes occurs in different time periods.

The following notions underline the functionality of the Demand Derivation and Aggregation functions.

- *TT (traffic trunk)*: a traffic commodity that a provider domain must serve, therefore must be dimensioned against. It does not denote a commodity offered by the network; in fact, TTs multiplex the offered commodities c/pSLS. With respect to the MESCAL QoS terminology, a TT is defined as a QoS-class in a certain topological scope. We naturally then, distinguish between *internal TTs (iTTs)* and *external TTs (eTTs)* as follows:

```
iTT :: ingress-router egress-router lQC
eTT :: ingress-link {dest} eQC
```

  where *dest* denotes a group of IP prefixes (see section 9.4.2.2) anywhere in the Internet. Based on the above definitions eTM is calculated in terms of eTTs and iTM in terms of iTTs.

10.2.3.2.2 eTM, iTM Calculation

This section outlines the process of calculating eTM and iTM (see Figure 81). This process executes at the beginning of an RPC.

Copyright © MESCAL Consortium, June 2005

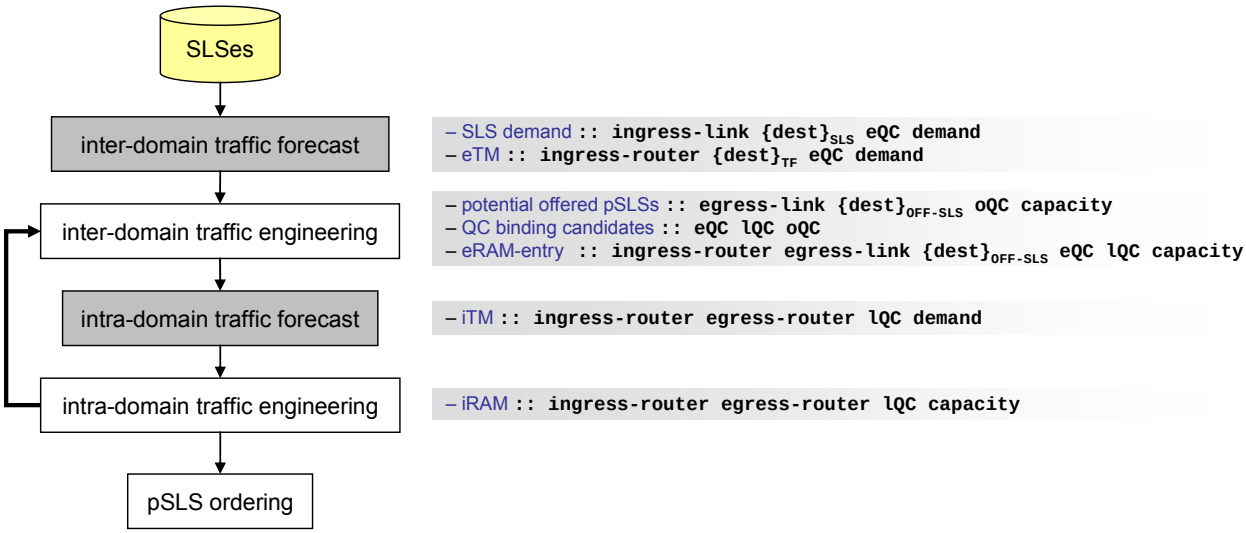


Figure 81: Traffic Matrices Calculation

Consider a population of c/pSLS subscriptions; this population may correspond to the union of already established and anticipated e.g. by market estimates c/pSLS. We assume that the c/pSLSes have been mapped to SSS/SLSes by the *Service Order Handling* component.

Step #1. Inter-domain traffic forecast demand derivation

```
For each c/pSLS (SSS)
     Determine service class and associated traffic forecast parameters (MF, AW)
     For each contained SLS
          Get the SLS ingress link and {dest}
          For each QoS level (eQC and traffic conformance)
               Create eTT-SLS for this SLS ingress link, eQC and {dest}
               Calculate and assign Maximum Contracted Demand from traffic
               conformance to the created eTT-SLS (assuming a fluid-flow model)
               Find Anticipated Demand per eTT-SLS (divide by MF and multiply by
AW)
```

Step #2. Inter-domain traffic forecast demand aggregation

```
Group eTT-SLS with identical ingress router, eQC and equal or fully contained
destination groups (see section 9.4.2.2) to form eTTs
For each eTT
     Find eTT Anticipated Demand summing the Anticipated Demand per merged eTT-SLS
```

Step #3. Off-line inter-domain traffic engineering

Off-line Inter-domain Traffic Engineering component produces a solution to accommodate eTM traffic demand towards the upstream service-peering domains. This solution is defined in terms of the offered pSLSes to establish with the peering domains, of the QoS-class binding candidates and of the *extended Resource Availability Matrix* (eRAM). The eRAM captures the resources assigned by the traffic engineering algorithm to a particular ingress, eQC pair to reach certain destinations out of a particular egress node and over a particular lQC.

Step #4. Intra-domain traffic forecast

Bound to the calculations of the *Off-line Inter-domain Traffic Engineering* component, the traffic forecast function reduces to aggregating the resources allocated per eRAM entry to corresponding iRAM entries. The output of intra-domain traffic forecast will be fed to *Off-line Intra-domain Traffic Engineering* which has the objective to find an optimum intra-domain resources distribution for supporting the allocated inter-domain resources in eRAM.

```
Group eRAM-entries with identical ingress router, egress-router and lQC to form
iTTs, for each iTT
     Find iTT Anticipated Demand summing the capacity per merged eRAM-entry
```

Copyright © MESCAL Consortium, June 2005

Step #5. Off-line intra-domain traffic engineering

Off-line Intra-domain Traffic Engineering distributes intra-domain resources to accommodate the iTM traffic demand based on the QoS-class binding candidates. The overall inter-domain and intra-domain solution is evaluated resulting in either a new iteration pursuing a better solution or in proceeding to implement the best solution, invoking among others the *pSLS Ordering* component to establish the selected offered pSLSes.

10.3 Traffic Engineering interactions

10.3.1 Decomposition of Offline Inter-domain Traffic Engineering

Figure 82 shows the decomposition of the Offline Inter-domain Traffic Engineering component of the MESCAL functional architecture.

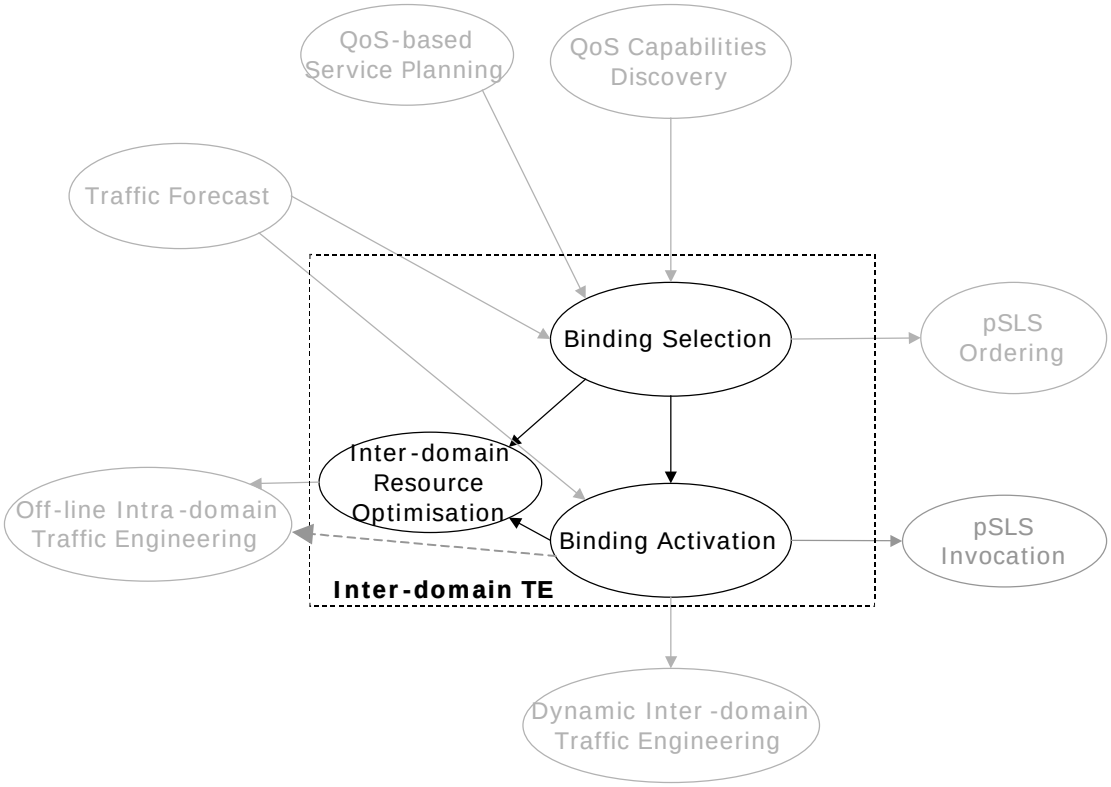


Figure 82. Decomposition of Offline Inter-domain TE

10.3.2 Resource Provisioning Cycles

10.3.2.1 Definitions

We define two terms as follows: the Intra-domain resource provisioning cycle (RPC) is the sequence of network resource dimensioning functions performed *within* a domain, while Inter-domain RPC is the sequence of network resource dimensioning functions performed *between* adjacent domains. Both Intra-domain RPC and Inter-domain RPC functions are performed at regular periods.

Copyright © MESCAL Consortium, June 2005

In the case of Inter-domain RPC, we need to further define two different cycles: the Binding Selection Cycle and the Binding Activation Cycle (Figure 83[3]).

The Binding Selection Cycle concerns the period when the Binding Selection component decides inter-domain resource usage, and determines which pSLSs to establish with the domain's peer ASs; the Binding Selection function block then commands the pSLS Ordering function block to negotiate these pSLSs with peer domains.

The Binding Activation Cycle is the period between two successive network resource dimensioning enforcements between adjacent domains, when this resource dimensioning is constrained by established pSLSs.

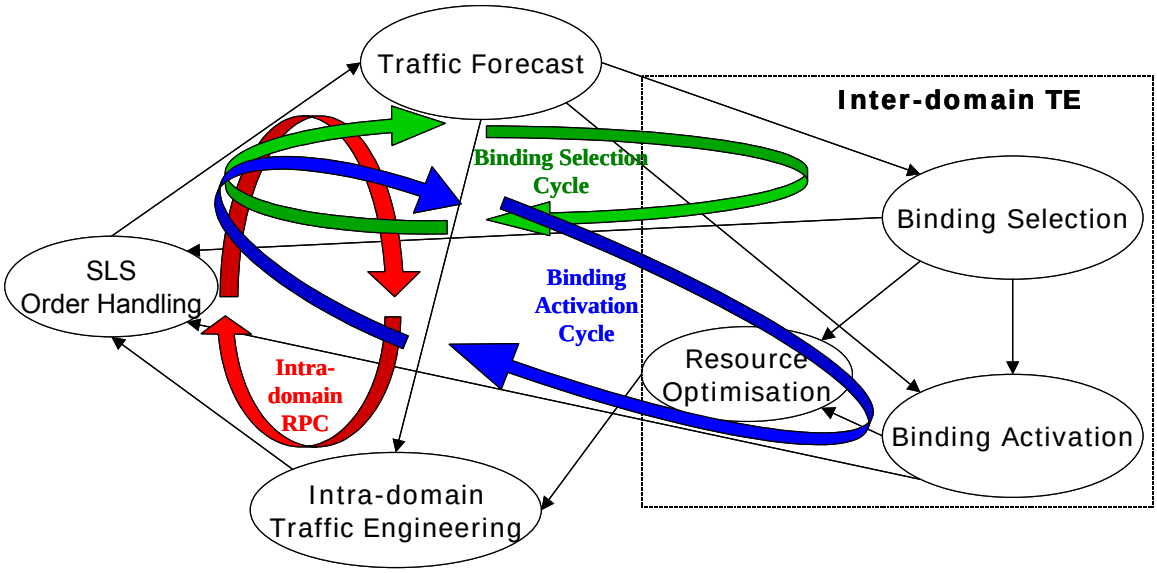


Figure 83. Resource Provisioning Cycles

10.3.2.2 Issues

The Binding Selection Cycle occurs less frequently and with a longer timescale than the Binding Activation Cycle; for example the Binding Selection Cycle may run monthly while the Binding Activation Cycle may run daily. Although it would provide a more optimal solution with regards to resource utilisation when the periods of both the Binding Selection Cycle and the Binding Activation Cycle are equal, it is not expected that ASs will change their established pSLSs with their peers every time they enforce a new intra/inter-domain resource configuration. In consequence, the MESCAL approach allows domains flexibility by assuming that the Binding Activation Cycle occurs more frequently than the Binding Selection Cycle. Every time a Binding Selection Cycle runs, a Binding Activation Cycle is also triggered so that the inter-domain configuration selected by Binding Activation reflects the current set of pSLSs.

[3] Note that the arrows from Binding Selection and Binding Activation to SLS Order Handling do not appear in the overall functional architecture figure. They are illustrated here because the eRAM/iRAM are assumed to be sent via the Resource Optimisation and Intra-domain offline TE function blocks

Copyright © MESCAL Consortium, June 2005

We also argue that the periods of intra-domain and inter-domain RPC should be equal. The justification of this argument is given by analysing the consequences of the following possible combinations:

- Case 1: Intra-domain RPC occurs more often than Binding Activation Cycle

  In this case, intra-domain traffic changes more dynamically than inter-domain traffic. However, triggering intra-domain network dimensioning more frequently and independently from inter-domain dimensioning is not an optimal solution. This is because, since the inter-domain dimensioning remains unchanged, the resources allocated within the domain for inter-domain traffic remain intact. This gives more stringent link capacity constraints for Intra-domain network dimensioning subsystem to engineer the network for intra-domain traffic. With these constraints, the Intra-domain network dimensioning subsystem may not produce a network configuration that achieves optimal resource utilisation. As a result, only a sub-optimal intra-domain configuration can be achieved.

- Case 2: Binding Activation Cycle occurs more often than Intra-domain RPC

  This case is similar to case 1. Inter-domain traffic engineering may not be able to select an optimal egress point for inter-domain traffic due to the stringent link capacity constraints. In this case, the stringent capacity constraints are a result of fixed link resource allocations for the intra-domain traffic. As a result, only a sub-optimal inter-domain configuration can be achieved.

- Case 3: Equal Binding Activation Cycle and Intra-domain RPC periods

  When both the Binding Activation Cycle and Intra-domain RPC periods are equal, a network configuration that yields optimal resource utilisation for intra-domain and inter-domain traffic simultaneously is possible, by taking the latest iTM and eTM into consideration. In this case, compared to case 1 and 2, a more optimal intra-domain and inter-domain resource utilisation can be achieved.

10.3.3 Decoupled and integrated approaches to Inter- and Intra-domain TE

The purpose of this section is to introduce two potential approaches for inter-domain resource optimisation, namely decoupled and integrated resource optimisation. In decoupled resource optimisation, the algorithms that perform Inter-domain resource allocation and Intra-domain resource allocation run independently. The algorithm proceeds by iterating between inter- and intra-domain resource allocation. In comparison, the integrated resource optimisation approach considers both inter- and intra-domain resources at the same time.

In principle, the integrated approach will provide a more optimal system configuration since it is taking account of many variables simultaneously. However, the decoupled approach allows algorithms for inter- and intra-domain traffic engineering to be considered separately, and it is the view of the MESCAL team that this approach will initially lead to more fruitful insights into inter-domain QoS engineering. In this Section, we compare the two approaches. However, the algorithms that have been developed and which are described in Section 10.4.3 assume the decoupled approach. For each approach, we provide a brief introduction, and describe the inputs and outputs, and give a process description.

Copyright © MESCAL Consortium, June 2005

10.3.3.1 Decoupled Inter-domain Resource Optimisation

Figure 84 shows the generic decoupled *Inter-domain Resource Optimisation* approach. The description of this approach is as follows.

Two function blocks are shown in the figure: *Inter-domain Resource Optimisation* and *Offline Intra-domain TE*. The function of *Offline intra-domain TE* is to compute the intra-domain network configuration and dimension resources between edges within a network.

Initially, *Inter-domain Resource Optimisation* accepts some input data and actions from *Binding Activation* or *Binding Selection*. It is noted that *Inter-domain Resource Optimisation* is responsible for mapping customer inter-domain traffic to appropriate egress points/pSLSs while satisfying the traffic with QoS requirements. In order to satisfy the traffic with QoS requirements, both inter-domain and intra-domain must have sufficient resources to accommodate the traffic. Since the availability of inter-domain resource (i.e. $pSLSs_{out}$) has been produced by *Binding Selection* and is passed to *Inter-domain Resource Optimisation* through *Binding Activation*, *Inter-domain Resource Optimisation* can simply check whether or not a specific $pSLS_{out}$ has sufficient resource to accommodate the traffic with QoS requirements. However, *Inter-domain Resource Optimisation* is not able to determine whether resources within the domain are sufficient or not. Since *Offline Intra-domain TE* is responsible for routing traffic within the network towards specified destination or egress points, it is able to return the resulting intra-domain resource availability to *Inter-domain Resource Optimisation,* which in turn can select an appropriate inter-domain TE solution. Thus, communication between *Inter-domain Resource Optimisation* and *Offline Intra-domain TE* is necessary.

Basically, there are two reasons to establish communication between *Inter-domain Resource Optimisation* and *Offline Intra-domain TE*. The first one is that, as discussed in the previous paragraph, *Offline Intra-domain TE* can help to check whether there is sufficient intra-domain resource to accommodate the traffic with QoS requirements. The second reason is that *Inter-domain Resource Optimisation* can assign egress points to inter-domain traffic as a presumed solution and request *Off-line intra-domain TE* to indicate the resulting intra-domain resource availability and utilisation after both intra-domain and inter-domain traffic (egress points have been identified from that presumed solution) are being routed in the network. Specifically, this evaluates the impact of that presumed solution when routed with intra-domain traffic on the intra-domain resource utilisation. This reason is supported when the objective of *Inter-domain Resource Optimisation* is to optimise not only inter-domain resource utilisation but also intra-domain resource utilisation.

With the above reasoning, *Inter-domain Resource Optimisation* queries *Offline Intra-domain TE* to perform intra-domain resource checking and to provide a resulting intra-domain resource utilisation by giving a presumed inter-domain TE solution (eRAM). The resulting intra-domain resource availability and utilisation returned from off-line intra-domain TE will help *Inter-domain Resource Optimisation* to determine an optimal inter-domain TE solution while meeting the target optimisation objectives. *Inter-domain Resource Optimisation* may consult with *Offline Intra-domain TE* a number of times in order to produce a single or multiple optimal inter-domain TE solutions.

From *Inter-domain Resource Optimisation* point of view, *Offline Intra-domain TE* is a black box which only provides interfaces to accept input, make decisions and produce output. The *Offline Intra-domain TE* can be any existing intra-domain TE solutions such as OSPF together with Constrained Shortest Path First (CSPF) or the TEQUILA intra-domain TE subsystem [TEQUILA]. In this decoupled approach, *Inter-domain Resource Optimisation* and *Offline Intra-domain TE* operate separately but a relationship by functional calls and parameters passing is established between them.

The next section describes input, output and processes of the generic decoupled *Inter-domain Resource Optimisation* approach.

Copyright © MESCAL Consortium, June 2005

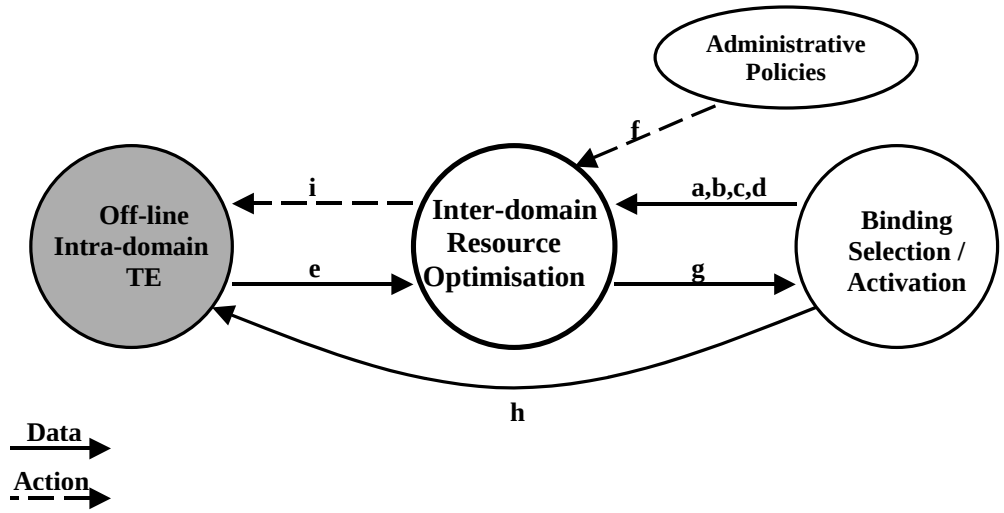


Figure 84. Decoupled Inter-domain Resource Optimisation

10.3.3.1.1 Decoupled Inter-domain Resource Optimisation input and output

This section describes input and output of the decoupled *Inter-domain Resource Optimisation*. Input and output are further divided into two types: data and action. Action requires some controls or feedback. The letter used in each item below corresponds to the one in Figure 84 and the target of usage (either *Binding Selection* or *Binding Activation*) is specified in a square bracket. If the target of usage is not specified, the corresponding item is applicable to both the *Binding Selection* and *Binding Activation*.

Input data

a) [*Binding Selection*] *eTM* -- The extended traffic matrix produced by *Traffic Forecast* and it is passed to *Inter-domain Resource Optimisation* through *Binding Selection*. The time scale of this eTM is large.

[*Binding Activation*] *eTM* -- The extended traffic matrix produced by *Traffic Forecast* and it is passed to *Inter-domain Resource Optimisation* through *Binding Activation*. The time scale of this eTM is small.

b) [*Binding Activation*] *$pSLSs_{out}$* -- A set of established $pSLSs_{out}$ produced by *Binding Selection*. The required fields include destination prefix, o-QC, bandwidth availability and egress interface.

c) [*Binding Selection*] One or more sets of options for sets of l-QCs, o-QCs and egress node IDs.

[*Binding Activation*] *QC mapping compatibility* -- A corresponding QC mapping compatibility for each $pSLS_{out}$. The mapping compatibility describes a set of eligible l-QCs maps to a specific o-QC.

d) Number of solutions to be returned.

e) *iRAM:* The internal resource availability matrix that specifies estimates of the availability of the engineered network to accommodate QoS traffic between edges in the network.

Copyright © MESCAL Consortium, June 2005

Input action

f) *Policies*: customised policies that may affect inter-domain resource optimisation decisions.

Output data

g) [*Binding Selection*] A number of *Network configuration* according to the parameter of *number of solutions to be returned*. Each network configuration is associated with a cost value to access its quality.

[*Binding Activation*] *eRAM(s)* -- The extended resource availability matrix produced by *Inter-domain Resource Optimisation*. It is an inter-domain traffic engineering solution that specifies estimates of the availability of the inter-domain resources (e.g. $pSLSs_{out}$) to accommodate QoS traffic towards the upstream service-peering domains. Each solution is associated with a cost value to access its quality. *Inter-domain Resource Optimisation* produces a set of eRAM(s) (i.e. inter-domain TE solutions) corresponds to the same input data according to the parameter of *number of solutions to be returned*.

h) [*Binding Activation*] *eRAM*: The best inter-domain TE solution selected by *Binding Activation* ($eRAM \in eRAM(s)$). Off-line intra-domain TE has to be informed this selected solution in order to select the corresponding intra-domain configuration.

Output action

i) *Intra-domain resource query*: A "what-if"-type query which includes a solution of *Binding Selection* or *Binding Activation* as a parameter and asks for internal network resource availability if the input solution is presumably selected/activated.

10.3.3.1.2 Decoupled Inter-domain Resource Optimisation process

This section briefly describes how the decoupled *Inter-domain Resource Optimisation* works in general. The process of decoupled *Inter-domain Resource Optimisation* is divided into following steps (words in bold are data or actions that have been defined in the previous section):

Calls from Binding Selection

1. Inter-domain *Resource Optimisation* receives **eTM, One or more sets of options for sets of l-QCs, o-QCs and egress node IDs** and the **number of solutions to be returned** from *Binding Selection* as input data and **policies** (if any) as an input action.
2. The decision of *Inter-domain Resource Optimisation* may depend on intra-domain resource availability and utilisation. *Inter-domain Resource Optimisation* sends an **intra-domain resource query** to *Offline Intra-domain TE* requesting the resulting intra-domain resource availability and utilisation assuming a given *Binding Selection* solution were to be selected.
3. *Offline Intra-domain TE* takes the solution as input and then uses its optimisation algorithms to compute a network configuration for both intra-domain and inter-domain traffic routed within the network. *Offline Intra-domain TE* answers the query by providing **iRAM** as output to *Inter-domain Resource Optimisation*.
4. *Inter-domain Resource Optimisation* takes iRAM into consideration (together with other factors) to determine whether the considered solution is an optimal solution while achieving the target resource utilisation objectives.
5. *Inter-domain Resource Optimisation* may produce multiple solutions by repeating the step from 2 to 4 with different *Binding Selection* solutions as input.

Copyright © MESCAL Consortium, June 2005

6. *Inter-domain Resource Optimisation* outputs a set of solutions to *Binding Selection* for which to select and implement the best one.

Calls from Binding Activation

1. *Inter-domain Resource Optimisation* receives eTM, a set of pSLSsout , the corresponding QC mapping compatibility and the number of solutions to be returned from *Binding Activation* as input data and policies (if any) as an input action.
2. The decision of *Inter-domain Resource Optimisation* may depend on intra-domain resource availability and utilisation. *Inter-domain Resource Optimisation* sends an intra-domain resource query to *Offline Intra-domain TE* requesting the resulting intra-domain resource availability and utilisation assuming a given inter-domain TE solution (eRAM) were to be put into effect or activated.
3. *Offline Intra-domain TE* takes this given eRAM as input and then uses its optimisation algorithms to compute a network configuration for both intra-domain and inter-domain traffic routed within the network. *Offline Intra-domain TE* answers the query by providing iRAM as output to the *Inter-domain Resource Optimisation*.
4. *Inter-domain Resource Optimisation* takes iRAM into consideration (together with other factors) to determine whether the considered inter-domain TE solution is an optimal solution while achieving the target resource utilisation objectives.
5. *Inter-domain Resource Optimisation* may produce multiple optimal inter-domain TE solutions by repeating the step from 2 to 4 with different eRAM as input.
6. *Inter-domain Resource Optimisation* outputs a set of solutions (eRAM(s)) to *Binding Activation* for which to select and implement the best one.
7. Finally, *Binding Activation* informs *Offline Intra-domain TE* of the selected inter-domain TE solution (eRAM∈eRAM(s)) for which to select the intra-domain configuration corresponds to that selected solution.

10.3.3.2 Integrated Inter-domain Resource Optimisation

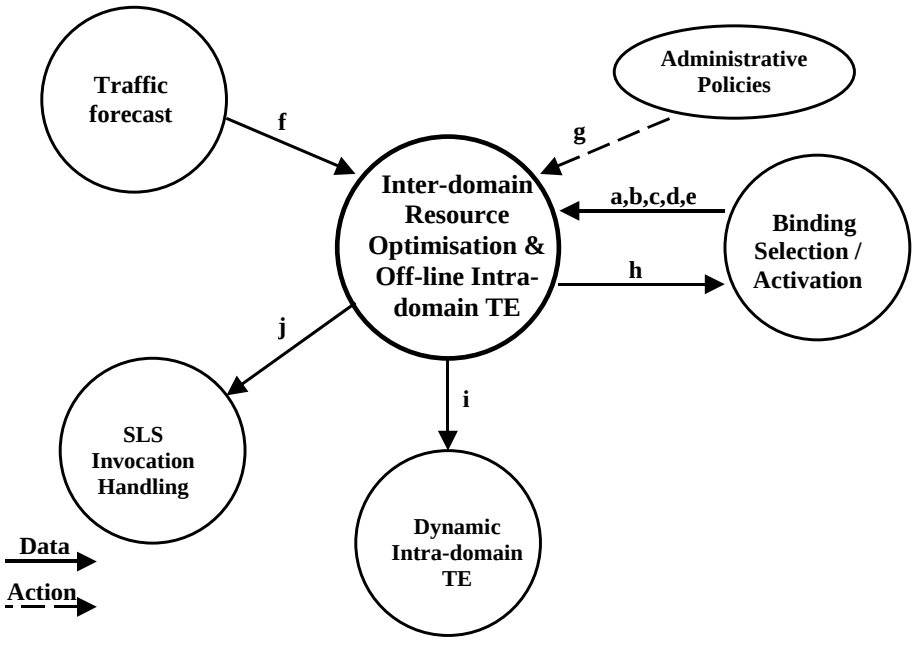


Figure 85. Integrated Inter-domain Resource Optimisation

Copyright © MESCAL Consortium, June 2005

Figure 85 shows the generic integrated *Inter-domain Resource Optimisation* approach. This approach integrates *Inter-domain Resource Optimisation* with *Offline Intra-domain TE* to produce a complete traffic engineering solution for both intra-domain and inter-domain traffic simultaneously. The integrated approach differs from the decoupled approach in that the integrated approach, unlike the decoupled approach, does not presumably assign egress points to inter-domain traffic and consult with *Offline Intra-domain TE* to obtain the resulting intra-domain resource utilisation. Instead, the integrated approach performs inter-domain TE *collaboratively* with intra-domain TE to produce a complete traffic engineering solution simultaneously (this includes egress point/$pSLS_{out}$ selection and traffic routing between edge nodes in the network).

Note that, in the integrated approach, since inter-domain traffic is routed in the network where some resources are consumed, each decision on routing inter-domain traffic may affect the decision of routing intra-domain traffic in the network since both types of traffic share resources within the network and the capacity of those resources are constrained. This also holds for the opposite case where each decision on routing intra-domain traffic can affect the decision of egress point selection for inter-domain traffic.

10.3.3.2.1 Integrated Inter-domain Resource Optimisation input and output

This section describes input and output of the integrated *Inter-domain Resource Optimisation*. Input and output are further divided into two types: data and action. Action requires some controls or feedback. The letter used in each item below corresponds to the one in Figure 85 and the target of usage (either *Binding Selection* or *Binding Activation*) is specified in a square bracket. If the target of usage is not specified, the corresponding item is applicable to both *Binding Selection* and *Binding Activation*.

Input data

a) [*Binding Selection*] *eTM* -- The extended traffic matrix produced by *Traffic Forecast* and it is passed to *Inter-domain Resource Optimisation* through *Binding Selection*. The time scale of this eTM is large.

   [*Binding Activation*] *eTM* -- The extended traffic matrix produced by *Traffic Forecast* and it is passed to *Inter-domain Resource Optimisation* through *Binding Activation*. The time scale of this eTM is small.

b) [*Binding Activation*] $pSLSs_{out}$ -- A set of established $pSLSs_{out}$ produced by *Binding Selection*. The required fields include destination prefixes, o-QC, bandwidth availability and egress interface.

c) [*Binding Selection*] One or more sets of options for sets of l-QCs, o-QCs and egress node IDs.

   [*Binding Activation*] *QC mapping compatibility* -- A corresponding QC mapping compatibility for each $pSLS_{out}$. The mapping compatibility describes a set of eligible l-QCs maps to a specific o-QC.

d) Number of solutions to be returned.

e) [*Binding Activation*] *eRAM*: The selected inter-domain TE solution, eRAM∈eRAM(s), which is determined by *Binding Activation.*

f) *iTM:* The internal traffic matrix that includes traffic routed between edges in a network. Note that this iTM does not contain any inter-domain traffic since its egress point has not been selected yet.

Copyright © MESCAL Consortium, June 2005

Input action

g) *Policies*: customised policies that may affect *Inter-domain Resource Optimisation* decisions.

Output data

h) [*Binding Selection*] A number of *Network configuration* according to the parameter of *number of solutions to be returned*. Each network configuration is associated with a cost value to access its quality.

[*Binding Activation*] *eRAM(s)* -- The extended resource availability matrix produced by *Inter-domain Resource Optimisation*. It is an inter-domain traffic engineering solution that specifies estimates of the availability of the inter-domain resources (e.g. $pSLSs_{out}$) to accommodate QoS traffic towards the upstream service-peering domains. Each solution is associated with a cost value to access its quality. *Inter-domain Resource Optimisation* produces a set of eRAM(s) (i.e. inter-domain TE solutions) corresponds to the same input data according to the parameter of *number of solutions to be returned*.

i) [*Binding Activation*] *iRAM:* The internal resource availability matrix that specifies estimates of the availability of the engineered network to accommodate QoS traffic between edges in the network.

j) [*Binding Activation*] *RAM:* RAM should give an estimate of the available resources for the various reachable destination prefixes (end-to-end) for all employed QCs or Meta-QoS-Classes. The estimates could be expressed as a range of bandwidth values, taking into account potential resource sharing between classes and intra/inter-domain reachable destination prefixes.

10.3.3.2.2 Integrated Inter-domain Resource Optimisation process

This section briefly describes how the integrated *Inter-domain Resource Optimisation* works in general. The process is divided into following steps: (words in bold are data or action that have been defined in the previous section)

Calls from Binding Selection

1. The integrated *Inter-domain Resource Optimisation* receives **eTM**, **One or more sets of options for sets of l-QCs, o-QCs and egress node IDs** and **iTM** as input data from *Binding Selection* and *Traffic Forecast*, and **policies** (if any) as input action.
2. The integrated *Inter-domain Resource Optimisation* takes the input and computes a complete traffic engineering solution simultaneously for both intra-domain and inter-domain traffic. This includes egress point/$pSLS_{out}$ selection, traffic routing between edge nodes in the network. The integrated inter-domain resource optimisation may produce a set of optimal inter-domain and intra-domain traffic engineering solutions.
3. The integrated *Inter-domain Resource Optimisation* outputs a set of solutions to *Binding Selection* for which to select and implement the best one.

Calls from Binding Activation

1. The integrated *Inter-domain Resource Optimisation* receives **eTM**, a set of **$pSLSs_{out}$**, the corresponding **QC mapping compatibility** and **iTM** as input data from *Binding Activation* and *Traffic Forecast*, and **policies** (if any) as input action.
2. The integrated *Inter-domain Resource Optimisation* takes the input and computes a complete traffic engineering solution simultaneously for both intra-domain and inter-domain traffic. This includes egress point/$pSLS_{out}$ selection, traffic routing between edge nodes in the network. The integrated *Inter-domain Resource Optimisation* may produce a set of optimal inter-

Copyright © MESCAL Consortium, June 2005

domain and intra-domain traffic engineering solutions. Each solution consists of eRAM with corresponding iRAM (or collectively called RAM).

3. The integrated *Inter-domain Resource Optimisation* outputs a set of optimal inter-domain TE solutions (**eRAM(s))** to *Binding Activation* for which to select and implement the best solution. *Binding Activation* returns the selected inter-domain TE solution, **eRAM**∈eRAM(s), to the integrated *Inter-domain Resource Optimisation*.

4. According to the selected eRAM, the integrated *Inter-domain Resource Optimisation* selects the corresponding **iRAM**. The iRAM is output to dynamic intra-domain TE for which to configure the network while **RAM** (the selected eRAM plus its corresponding iRAM) is output to *SLS Invocation Handling* for which to control the amount of traffic injected into the network.

10.3.3.3 Algorithms for Decoupled and Integrated Optimisation

We propose greedy heuristic algorithms to evaluate the effectiveness of the decoupled and the integrated optimisation. Note that the proposed algorithms have only considered bandwidth as the QoS metric based on the assumption that delay and jitter can be converted into effective bandwidth.

10.3.3.3.1 Algorithms for Decoupled Optimisation

The greedy-based heuristic algorithm for the *decoupled* approach works as follows:

1. Sort all eTM traffic flows in a descending order according to their bandwidth requirement. Consider the first traffic flow in that order and assign it to a feasible inter-domain link that meets the bandwidth requirement while incurring the lowest cost. Update the inter-AS resource availability and repeat the selection for the next traffic flow until all eTM traffic flows have been considered. This step refers to as inter-domain TE by which we mean egress point selection.

2. Based on the sorted eTM traffic flows produced by step 1, select a minimum cost route that satisfies the bandwidth demand for the first traffic flow between the associated ingress and egress routers. Update network resource availability and repeat this route selection for the rest of the traffic flow in that sorted order. This step refers to as intra-domain TE by which we mean intra-domain route optimisation.

10.3.3.3.2 Algorithms for Integrated Optimisation

The greedy-based heuristic algorithm for the *integrated* approach works as follows:

1. Sort all eTM traffic flows in a descending order according to their bandwidth requirement.

2. Consider the first ordered traffic flow, identify a set of inter-domain links together with their corresponding intra-AS routes that satisfy the bandwidth demand. Calculate a cost based on the utilisation of each possible combination of inter-domain link and intra-domain route were the traffic flow to be assigned to them. Among these possibilities, select the one with the minimum cost. Update the inter-domain and intra-domain resource availability and repeat the selection for the next flow until all the traffic flows have been considered.

The primary difference between the algorithms proposed for the two approaches is that the decoupled approach algorithm divides egress router and intra-domain route selection into two successive phases while the integrated approach algorithm finds both egress router and intra-domain route simultaneously for each traffic flow by considering their conditions (i.e. cost).

10.3.4 Off-line inter-domain TE cases

This section explores all the possible cases for off-line inter-domain traffic engineering. Readers are referred to D1.1 section 5.4 for relevant detail [D1.1]. The following cases are applied to each inter-domain traffic (i.e. aggregated traffic based on ingress router and destination prefix).

Copyright © MESCAL Consortium, June 2005

10.3.4.1 Single/Multiple egress point selection

Under inter-domain routing, it is common that a destination prefix can be reached through multiple egress points in a network. Service and Network Providers thus have to select appropriate egress points to egress inter-domain traffic. Two variant of egress point selection are deduced. For ***single egress point selection***, only a single egress point is selected for each destination prefix. Thus, all traffic towards a destination prefix, regardless of using which ingress routers, will always egress from the same egress point. In practice, single egress point selection is used when Service and Network Providers always have a preferable egress point over the others for each destination prefix. However, it is possible to improve network resource utilisation by allowing multiple egress points for each destination prefix. In this case, ***multiple egress point selection***, all traffic towards a destination through a designated ingress router will egress from a selected egress point. As a result, multiple egress point selection allows resource load balancing and the assignment of an optimal egress point to each aggregated inter-domain traffic based on ingress router and destination prefix.

10.3.4.2 Single/Multiple $pSLS_{out}$ selection

Although an egress point is selected, there may still be multiple $pSLSs_{out}$ that are offered by the same or different service peering providers and are attached with the egress point towards a destination prefix. Service and Network Providers thus have to select appropriate $pSLSs_{out}$ to egress traffic to the destination prefix with end-to-end guarantees. Two variant of $pSLS_{out}$ selection are deduced. For ***single $pSLS_{out}$ selection***, only a single $pSLS_{out}$, among all the $pSLSs_{out}$ that are attached with each egress point, is selected. In this case, load balancing between multiple $pSLSs_{out}$ on each egress point is not allowed. However, single $pSLS_{out}$ selection exists for policy and managerial reasons and when the cost to sign a $pSLS_{out}$ is high. On the other hand, ***multiple pSLS selection*** allows more than one $pSLS_{out}$ to be selected among all the $pSLSs_{out}$ that are attached with each egress point. In this case, load balancing between multiple $pSLSs_{out}$ on each egress point is allowed.

10.3.4.3 Single/Multiple l-QC selection

To provide an extended QC, a Service or Network Provider has to bind its l-QC with the QC offered by its service peering provider. There may be multiple appropriate l-QCs to complete the binding. ***Single l-QC selection*** allows only a single l-QC bind to each offered QC, while ***multiple l-QC selection*** allows more than one l-QC for the reason of intra-domain load balancing.

10.3.4.4 Single/Multiple intra-domain route selection

When an egress point has been chosen, a path is then found between the designated ingress router and the chosen egress point. By taking a selected l-QC into consideration, there may be multiple paths to the egress point. ***Single intra-domain route selection*** allows only a single path selected between the ingress router and the egress point, while ***multiple intra-domain route selection*** allows multiple paths for the reason of intra-domain load balancing.

10.3.4.5 QoS parameters consideration

We consider a set of binding activation problems by considering various QoS parameters. Three possible binding activation problems can be deduced:

- ***Bandwidth constrained binding activation***: Map the predicted traffic matrix to the inter-domain network resources, satisfying bandwidth requirements while aiming at optimising the use of network resources.

Copyright © MESCAL Consortium, June 2005

- ***Delay constrained binding activation***: Map the predicted traffic matrix to the inter-domain network resources, satisfying delay requirements while aiming at optimising the use of network resources.

- ***Jitter/Loss constrained binding activation***: Map the predicted traffic matrix to the inter-domain network resources, satisfying jitter/loss requirements while aiming at optimising the use of network resources.

10.3.4.6 Traffic engineering scenarios

All the potential problems defined in the Sections 10.3.4.1 to 10.3.4.4 can be combined to form a set of traffic engineering scenarios. The justification is that:

- The selection of egress point does not mean that pSLS will also be chosen and vice versa. The reader is referred to D1.1 section 5.4.2.2 for more detail [D1.1].

- The selection of both egress points and $pSLSs_{out}$ means that an exit point to egress traffic towards a destination prefix has been identified. The next step is to bind l-QCs to these egress points and pSLSs.

- The selection of l-QCs may be supported by multiple intra-domain paths between the ingress and egress point.

Thus, there are maximum 2^4 possible traffic engineering scenarios formed by the combination of problems defined in Sections 10.3.4.1 to 10.3.4.4. Note that, however, some of the scenarios may not be available for each of the three MESCAL solution options. The number of scenarios becomes larger if we consider more QoS parameters. We formulated three binding activation problems in section 3.5 and these problems, when arbitrary combined together, form a multi-constrained problem. Thus, this produces a set of single/multi-constrained problems by combination. We denote N by the number of problem combinations in section 3.5. In this case, the maximum number of possible traffic engineering scenarios becomes $N \cdot 2^4$.

Copyright © MESCAL Consortium, June 2005

10.4 Offline Inter-domain TE

10.4.1 Binding Selection

10.4.1.1 Introduction

The *Binding Selection* function block is part of the Offline Inter-domain Traffic Engineering function. The component that implements these functions will run at the Binding Selection Cycle epoch. *Binding Selection* is expected to run less frequently than *Binding Activation*.

The functions of *Binding Selection* may be divided into three principal areas, outlined in the following three paragraphs.

The first function of *Binding Selection* is to compute potential e-QCs, each consisting of bindings of l-QCs in the local domain with the o-QCs of downstream peers. The process of identifying this list of e-QCs will take into account business-related constraints (policies) when generating combinations of l-QCs and o-QCs. The list of e-QCs may also take into account simple engineering constraints, such as destination addresses that can only be reached through a single egress. This function of *Binding Selection* is essentially the function originally specified in [D1.1] for the *QC Mapping* function block.

The second function of *Binding Selection* is to select a set of e-QCs that meets the QoS requirements of the forecast inter-domain traffic and which makes optimal use of network resources both within the domain and on the inter-domain links. It achieves this by interfacing with the function block responsible for computing (and optimising) the Inter-domain configuration, *Inter-domain Resource Optimisation*, and through it the function block responsible for computing the optimal intra-domain network configuration, *Offline Intra-domain TE*.

The third function of *Binding Selection* is to command the *pSLS Ordering* function block to negotiate pSLSs with peers, and to identify ranges of parameters (such as bandwidth, one-way transit delay, cost) and groups of pSLSs which have to be ordered or negotiated.

10.4.1.2 Objectives

The overall objectives of *Binding Selection* are to select a set of e-QCs that meet the QoS requirements of the forecast inter-domain traffic while taking into account the inter- and intra-domain configuration and traffic demands, to identify an optimal set of pSLSs that support these e-QCs, and to command the *pSLS Ordering* function block to negotiate these pSLSs.

10.4.1.3 Interface specification

Figure 86 shows the interfaces related to Binding Selection.

Copyright © MESCAL Consortium, June 2005

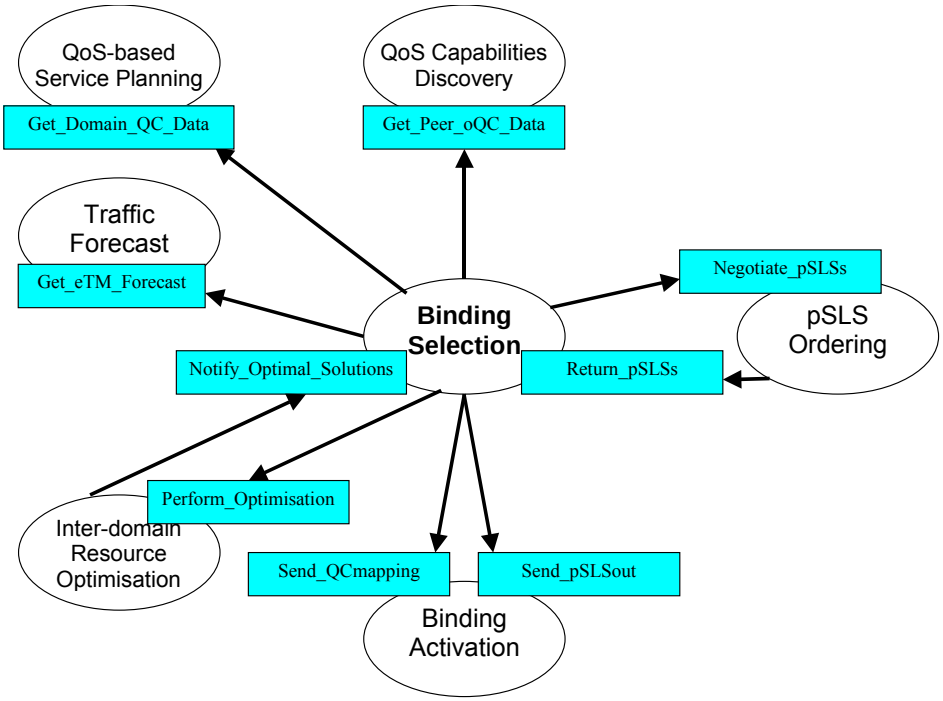


Figure 86. Binding Selection interfaces

10.4.1.3.1 From QoS Capabilities Discovery to Binding Selection

Get_Peer_oQC_Data (o-QC)

This interface provides *Binding Selection* with a list of all the o-QCs offered by peers. For each o-QC, the following information is required:

- AS identifier (or egress node interface ID);
- Destination address prefix(es);
- o-QC parameters (attribute/value pairs);
- Time schedule that the o-QC is or will be available.

10.4.1.3.2 From QoS-based Service Planning to Binding Selection

Get_Domain_QC_Data (l-QC, e-QC)

This interface provides to *Binding Selection* (a) the l-QCs that the domain provides to its customers and service peer providers within the scope of its network, and (b) a list of the e-QCs to be used to build o-QCs that the domain will offer, as defined by higher business-related activities and objectives.

For each l-QC the following information is required:

- l-QC;
- Time schedule that the l-QC is or will be available.

For each e-QC the following information is required:

- Destination address prefix(es);
- e-QC;
- Time schedule that the e-QC is or will be available.

Copyright © MESCAL Consortium, June 2005

10.4.1.3.3 From Traffic Forecast to Binding Selection

Get_eTM_Forecast (e-TM)

This interface provides the predicted traffic demand for this Resource Provisioning Cycle (in the form of the e-TM) to *Binding Selection*. The principal difference between the e-TM passed to each of *Binding Selection* and *Binding Activation* is that the former component plans for traffic flows over a longer timescale than the latter: consequently the e-TM used by *Binding Selection* is a longer-term forecast than that used by *Binding Activation*.

For each aggregate flow the following information is required:

- Ingress node ID;
- Input data rate (bandwidth);
- Destination address prefix(es);
- e-QC;
- Time schedule.

No information about committed resources is returned by *Binding Selection* to *Traffic Forecast*. This compares with *Binding Activation*, which returns an updated set of eRAM.

10.4.1.3.4 Interface with Inter-domain Resource Optimisation

This interface is bi-directional. *Binding Selection* passes to *Resource Optimisation* a specific configuration of flows including proposed l-QCs and egress node interface IDs. *Resource Optimisation* computes the best *n* inter- and intra- domain configurations, and then returns each network configuration together with its cost function value.

Perform_Optimisation (e-TM, pSLS_options, n)

The parameters passed from *Binding Selection* to *Inter-domain Resource Optimisation* are:

- Aggregate traffic flows (e-TM);
- One or more sets of options for inter-domain configuration, comprising:
    - Egress node interface ID;
    - Data rate (bandwidth);
    - Destination address prefix(es);
    - Time schedule;
    - l-QC and o-QC options;
- *n*, the number of solutions to be returned.

Notify_Optimal_Solutions (configuration, cost)

The parameters returned from *Inter-domain Resource Optimisation* to *Binding Selection* are as follows for each of the *n* solutions returned:

- The network configuration (for example, egress node interface IDs, selected l-QC and o-QCs);
- The cost of this configuration.

The difference in the two interfaces between *Binding Selection* and *Resource Optimisation* on the one hand, and between *Binding Activation* and *Resource Optimisation* on the other hand is due to one of the fundamental differences between the two binding components. *Binding Selection* is trying to define a range of "best" configurations so that it can give *pSLS Ordering* a range of pSLSs to negotiate. *Binding Activation* on the other hand operates at a shorter provisioning cycle, does not negotiate pSLSs, and uses only pre-existing pSLSs, and therefore is only interested in finding the single best configuration for its current predicted set of traffic flows.

Copyright © MESCAL Consortium, June 2005

It should be noted that in order to compute the cost of the inter-domain traffic, *Resource Optimisation* will in turn call the *Offline Intra-domain TE* function block, which requires knowledge of both inter- and intra- domain traffic.

10.4.1.3.5 Interface with pSLS Ordering

This interface is bi-directional.

Negotiate_pSLSs (candidate_pSLSs)

Binding Selection passes to *pSLS Ordering* sets of parameters; each set forms the basis of a pSLS to be negotiated with a downstream peer domain. Each negotiation may be either a new pSLS, modifications to an existing pSLS, or a cease of a pSLS. The parameters include the following, which are the principal components of a $pSLS_{out}$ from the traffic engineering perspective:

- Egress node interface ID and/or downstream AS identifier (AS ID);
- Destination address prefix(es);
- Required data rate (bandwidth);
- Required o-QC;
- Time schedule.

The parameters may be in the form of required single values; or a range of values (e.g. min, max, mean); or qualitative measures. Values and negotiation margins of parameters may be defined.

The information passed to *pSLS Ordering* will also include logic that provides a set of negotiating parameters. Examples of negotiating logic include (a) only one of several identified pSLSs need be successfully negotiated; (b) if one pSLS is successfully negotiated with a certain set or range of parameters then parameters of other(s) are changed; (c) a list in descending order of priority list, with alternative sets of pSLSs (which can be negotiated if the initial set is not agreed by peers).

Return_pSLSs (pSLS_data)

The return value passed by *pSLS Ordering* to *Binding Selection* is a statement of which pSLSs were successfully negotiated (as new, changed or ceased) and the agreed parameter values. The return value should also include explicit statements where pSLSs failed to be negotiated, and a reason code for each such failure (examples of this might be: insufficient bandwidth available; o-QC withdrawn; revised business-level policy).

10.4.1.3.6 From Binding Selection to Binding Activation

Send_QCMapping (l-QC,o-QC)

Send_pSLSout (pSLS)

This interface provides the output of the *Binding Selection* function block to *Binding Activation*. The output consists of the following information for each aggregate flow recorded in the e-TM:

- The expected components of a $pSLS_{out}$ (Egress node interface ID; Data rate (bandwidth); Destination address prefix(es); o-QC; Time Schedule);
- Mapping between l-QC and o-QC.

10.4.1.3.7 Additional parameters

Policies: l-QC / o-QC combination policies.

Topology information: network topology, in particular the inter-domain link data rates (bandwidth) and their QoS parameters.

Copyright © MESCAL Consortium, June 2005

10.4.1.4 Algorithm description

In this Section we describe algorithms for the three principal functions of *Binding Selection*, namely:

- QC mapping: identification of binding candidates of l-QCs and o-QCs;
- Selection of l-QCs and o-QCs that satisfy the required e-QCs, by means of inter- and intra-domain configuration computation and optimisation;
- *pSLS Ordering* negotiation.

10.4.1.4.1 Nomenclature

L	The set of l-QCs supported by this domain
l	One such l-QC, $l \in L$
Q	The set of o-QCs supported by neighbouring domains
q	One such o-QC, $q \in Q$
E	The set of e-QCs that the domain offers
e	One such e-QC, $e \in E$
N	The set of different e-QC mapping options (binding candidates) identified by the QC Mapping function within *Binding Selection*.
n	One such mapping option, $n \in N$
j	egress router interface
k	Destination address prefix
$f(j,q)$	Inter-domain link cost on egress router interface j using o-QC q (passed from *QoS Capabilities Discovery*)
$b(q,j)$	Inter-domain link bandwidth assigned to o-QC q on egress router interface j, and therefore the bandwidth that is assigned to a particular pSLS
bc	set of binding candidates $\{ l, q \}$ used for a particular run of *Inter-domain Resource Optimisation* (c.f. N, which contains *all* possible binding candidates)
S	Set of pSLSs
s	pSLS, $s \in S$, $s=(j, k, b(q,j), q)$
Ψ	Cost function (generated by *Inter-domain Resource Optimisation*)

10.4.1.4.2 Problem formulation: QC mapping

10.4.1.4.2.1 Formulation

The objective of QC mapping is to identify all combinations of l-QC (supported by this domain) and o-QC available to a particular destination address prefix (offered by one or more neighbouring peers) that will satisfy e-QCs that are both (a) supported by this domain and (b) have at least one entry in the e-TM for the current Resource Provisioning Cycle to an address in the range of the given destination address prefix. This process is performed separately for each time period / time schedule range. It is to be understood that the e-TM for Binding Selection includes predicted and estimated aggregate traffic flows valid over the long timescale used in Binding Selection.

This algorithm also includes the process of "QC Classification" [D1.1 Section 4.4.3], that is, the classification of l-QCs into meta-QoS-classes (M-QCs). Here, the QC mapping function determines all the possible/compatible combinations between the l-QCs and the M-QCs offered by peer domains. Since in general more than one l-QC can be mapped to a single M-QC, and one l-QC may belong to more than one M-QC, there could be many possible bindings of l-QC with o-QCs that support a given M-QC. These need to be identified, and then passed to the l-QC / o-QC selection process. In the case

Copyright © MESCAL Consortium, June 2005

where there is a M-QC that is not supported within the domain by a l-QC then the domain cannot offer this M-QC to its peers, and the domain thus appears as a "black hole" in that M-QC plane.

The algorithm presented below is applicable to statistical and hard solution options.

10.4.1.4.2.2 Algorithm

The QC mapping algorithm is presented below as pseudo-code.

/* apply QC-classification to l-QCs: maps the set L into a new set L' */

for (each time schedule period)

/* Select e-QCs that are required in the current time schedule range from the e-TM; record also the destination address prefixes (we assume that if an e-QC is in the e-TM then it must be an e-QC offered by this domain) */

set $X = N =$ null

for (each entry in the e-TM that is in the current time schedule period)

add e-TM's e-QC e_k and destination address prefix k_i to the set X

for (each e-QC e_k in set X)

/* Sort through all l-QC / o-QC combinations to form binding candidates; accept those where (a) the binding candidate satisfies the required e-QC; and (b) the destination address prefix that can be supported with the o-QC includes the destination address prefix required by one or more e-TM entries */

for (each $l' \in L'$ and each $q \in Q$)

calculate $e_{candidate} = l' \oplus q$ /* [D1.1] Section 7.3.2.5 */

for (each destination address prefix k_i supported by e_k)

if ($e_{candidate} \geq e_k$ and $k_i \leq$ any destination address prefix supported with q) add [e_k, l', q] to N;

end (for loop)

end (for loop)

end (for loop)

/* Apply any business policies for further elimination of l-QC / o-QC combinations */

apply policies to N

/* Error check */

for (each combination of e-QC and destination address prefix)

if (count of accepted binding candidates for this e-QC and destination address prefix) == 0 then flag error;

end (for loop)

end (for loop)

10.4.1.4.3 Problem formulation: l-QC and o-QC selection

10.4.1.4.3.1 Formulation

The objective of this algorithm is to select from the set of l-QC / o-QC combinations identified by QC mapping a set of l-QCs and pSLSs (i.e. o-QCs, bandwidth, egress node ID and destination address prefixes) that minimises the cost of inter-domain and intra-domain network resources, while meeting

Copyright © MESCAL Consortium, June 2005

the QoS requirements for the traffic forecast contained in the e-TM for the current Resource Provisioning Cycle.

The algorithm needs to take into account the following issues:

- Providing to *pSLS Ordering* a sufficient range of good parameters to provide the ability to negotiate pSLSs.
- Dealing with pSLSs inherited from the preceding RPC. The algorithm needs to establish whether any of these pSLSs should be ceased or changed, taking account of the need to provide a balance between on the one hand the stability and lifetime of negotiated pSLSs and on the other hand the optimal use of network resources.
- Algorithm complexity and runtime.
- Sensitivity of the proposed solution to variations in actual traffic demand during the RPC. For example, if the actual traffic were say 5% higher than planned, would the proposed solution still be optimal or near-optimal?

10.4.1.4.3.2 Description of algorithm

The basic algorithm is as follows: for each combination of l-QC and o-QC for each e-QC (i.e. for each of the binding candidates) passed from QC mapping, *Resource Optimisation* is run. The e-TM and the set of existing pSLSs is passed to *Resource Optimisation* in each of these runs. The optimisation algorithm returns a cost function Ψ and a mapping of e-TM flows to pSLSs in the form of e-RAMs. Intra-domain information such as assigned l-QCs are also returned, but only in the form of state information to provide a baseline configuration for when *Binding Selection* and *Resource Optimisation* run at the next Resource Provisioning Cycle.

Two variations of the algorithm are proposed at this stage:

- *Binding Selection* passes to *Resource Optimisation* both existing pSLSs and a set of new candidate pSLSs; *Resource Optimisat*ion returns a cost function Ψ for the given configuration of pSLSs;
- *Binding Selection* passes to *Resource Optimisation* only existing pSLSs; *Resource Optimisation* calculates (for the predicted traffic flows in the e-TM) a set of additional pSLSs, and returns both the cost function Ψ for the new of existing and additional pSLSs, and the set of additional pSLSs.

To enable changes to existing pSLSs to be included in the algorithm, a *perturbation* approach is used. Once a notionally minimum cost function Ψ has been established for the existing set of pSLSs, each pre-existing pSLS is in turn *relaxed* (i.e. removed from the set of existing pSLSs for this run only) and *Resource Optimisation* is re-run. Once this has been repeated for each pSLS individually, if relaxing one of the pSLSs has resulted in a significant improvement in the cost function, then this pSLS relaxation is accepted, and the perturbation is re-run to see if a further improvement in the cost function can be achieved by further relaxation. If the relaxation does not result in a significant improvement in the cost function, the perturbation stage of the algorithm terminates. The definition of "significant" will be refined during later analysis: it will be chosen as a compromise between on the one hand the desirability of maintaining stable pSLSs that are not changed at every RPC, and on the other hand ensuring a responsive low cost network that is capable of meeting QoS requirements for the traffic flows that it carries. An initial view of "significant" might be a 5% improvement in the cost function.

The QC mapping algorithm presented below is applicable to statistical and hard solution options. For the loose solution option, where binding arithmetic is not applicable, the binding candidates used are not { *l'*, *q* } but { *l'*, *M* } where *M* is the M-QC and *l'* is one of the l-QCs that has been classified in the same M-QC *M*.

Copyright © MESCAL Consortium, June 2005

10.4.1.4.3.3Algorithm

The l-QC and o-QC Selection algorithm is presented below as pseudo-code.

/* *Binding Selection* algorithm for case *where Resource Optimisation* creates candidate pSLSs */

/* Run *Resource Optimisation* for each combination of binding candidates */

while (not all binding candidate combinations exhausted) do

set bc = null

set S = all existing pSLSs

for (each $e_k \in N$) select a binding candidate { l' , q } where q is supported by at least one existing pSLS, and add to the set bc

call (res_optimisation (parms passed: e-TM, S, bc, $f(j,q)$; parms returned: Ψ, e-RAM, new_pSLSs))

end (while loop)

for the configuration whose returned cost is the lowest value:

set $\Psi_{lowest} = \Psi$

select returned e-RAM configuration whose returned cost Ψ is the lowest value

set $bc_{min} = bc$

set $S = S$ + new_pSLSs

/* Perturb the existing pSLSs about the operating point that is bc_{min} */

set significant_change = true

set $\Psi_{perturbed} = \Psi_{lowest}$

set $S' = S$

while (significant_change) do

for (count:=1; count<=|S'|; count++)

set $S'' = S' - S'$[count] /* remove the 'count'th element from S */

call(res_optimisation (parms passed: e-TM, S'', bc_{min}, $f(j,q)$; parms ret'd: Ψ_{ret}, e-RAM, new_pSLSs))

if ($\Psi_{ret} < \Psi_{perturbed}$)

set $\Psi_{perturbed} = \Psi_{ret}$ /* Record the new lowest cost */

set $S_{ret} = S''$ + new_pSLSs

end (for loop)

if ($\Psi_{perturbed} < (1\text{-}x) \cdot \Psi_{lowest}$)

set $\Psi_{lowest} = \Psi_{perturbed}$ /* New lowest value of cost found */

set $S' = S_{ret}$ /* S' is the set of pSLSs that give this lowest cost */

else

set significant_change = false /* change in Ψ is not "significant" */

end (while loop)

Copyright © MESCAL Consortium, June 2005

10.4.1.4.4 Problem formulation: pSLS negotiation

Once the lowest cost functions have been computed by the l-QC / o-QC selection algorithm, the combinations of parameters that result in the lowest value or values of the cost function Ψ are used as the basis for changing the pSLSs (new, changes to existing, or ceases), as is now described.

Two options are proposed that allow a narrower or wider range of negotiating stances for pSLS Ordering when new pSLSs are required. These two options are as follows:

- Simple approach: take the e-RAM and pSLSs corresponding to that of the lowest cost Ψ. Information about new or changed pSLSs is passed to *pSLS Ordering* for negotiation of pSLSs, and any pSLSs which are to be ceased are similarly passed to *pSLS Ordering*.
- Variation around an operating point: several sets of results for e-RAMs and pSLSs are used, for example all those configurations where the cost Ψ is within *x*% of the lowest .

10.4.1.4.4.1 Algorithm

The first option is easily calculated.

The second option is approached as follows. For each pair [egress node id *j* , o-QC *q*], for which any pSLS is specified in any configuration whose Ψ is within *x*% of the lowest, we determine the range (i.e. the minimum and maximum) of values of required bandwidth *b(q,j)*, destination address prefixes *k*, and time schedules. Depending on domain policies, a number of negotiating positions could then be established. For example, maximum flexibility (and probably maximum cost) could be incurred by negotiating a set of pSLSs each of which has the maximum range of bandwidth, destination address prefixes and time schedules. A more conservative approach would be to negotiate only the pSLSs which are required by the lowest cost Ψ, and applying a provisioning margin to the required bandwidth.

The impact of negotiating position will be considered and assessed during WP2 and WP3.

10.4.2 Binding Activation

10.4.2.1 Introduction

Binding Activation is an offline component that runs at Binding Activation Cycle epochs and produces an inter-domain traffic engineering solution (i.e. the established QC-bindings which has been put in effect for inter-domain traffic) at each short time scale.

10.4.2.2 Objectives

Binding Activation has two objectives:

- To indicate the optimal resource usage of each $pSLS_{out}$ produced by the inter-domain resource optimisation functional block. The necessity of this indication is due to the optimal resource usage of each $pSLS_{out}$ may be different from that originally defined in the $pSLS_{out}$. As a result, *Binding Activation* has to inform pSLS invocation how much resources will be actually invoked or used.
- To select the best among multiple inter-domain traffic engineering solutions. This is because the inter-domain resource optimisation may produce a set of optimal inter-domain traffic engineering solutions but only one is selected. *Binding Activation* determines and selects the best solution.

The selected inter-domain TE solution is enforced through routing decisions as well as configurations of the *Traffic Conditioning and QC Enforcement* function block, e.g. configuring the egress ASBR to perform DSCP remarking for realising an inter-domain TE solution.

Copyright © MESCAL Consortium, June 2005

10.4.2.3 Interface specification

This section describes the interaction of the *Binding Activation* function block with the others through events, messages or signals. Figure 87 shows the interfaces related with *Binding Activation*.

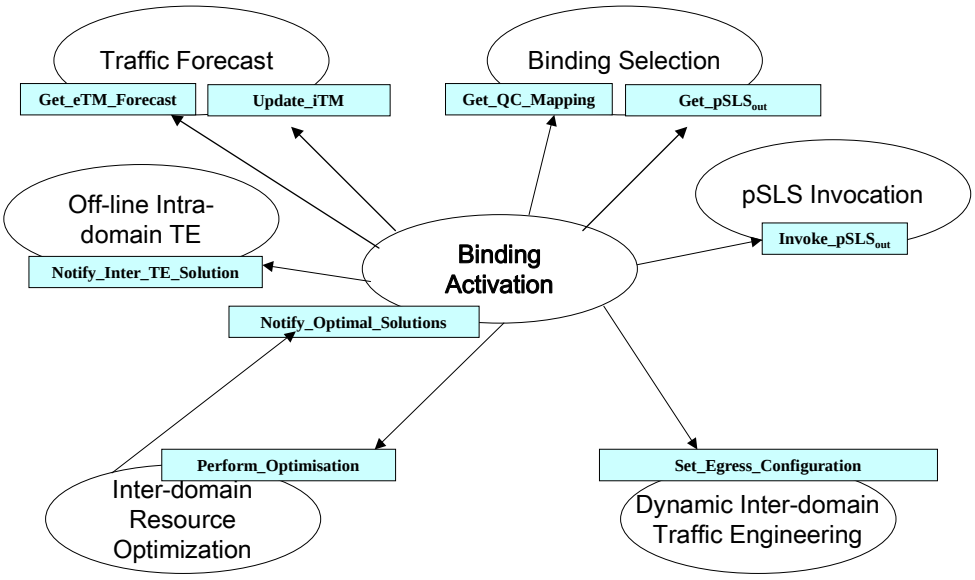


Figure 87. Binding Activation interfaces

- **Binding Activation to Traffic Forecast**
  Get_eTM_Forecast (Forecast parameters)

  This method will be called by *Binding Activation* to request and get the predicted extended traffic matrix from traffic forecast. Required traffic matrix parameters include destination prefix, a designated ingress router and requested QoS. Traffic is aggregated at each ingress router based on destination prefix.

  Update_iTM (eRAM)

  This method will be called by *Binding Activation* to pass the *Binding Activation* decision/inter-domain traffic engineering solution (i.e. eRAM) to traffic forecast for which to update the iTM.

- **Binding Activation to Binding Selection**
  Get_pSLS$_{out}$ (none)

  This method will be called by *Binding Activation* to request and get a set of pSLS$_{out}$ from Binding Selection.

  Get_QC_Mapping (none)

  This method will be called by *Binding Activation* to get a corresponding QC mapping compatibility for each received pSLS$_{out}$. The QC mapping compatibility describes a set of eligible l-QCs mappings to a specific o-QC.

- **Inter-domain Resource Optimisation to Binding Activation**
  Notify_Optimal_Solutions (eRAM(s))

Copyright © MESCAL Consortium, June 2005

This method will be called by inter-domain *Resource Optimisation* to notify *Binding Activation* a set of optimal inter-domain traffic engineering solutions (i.e. eRAM(s)).

- **Binding Activation to Inter-domain Resource Optimisation**
  Perform_Optimisation (optimisation parameters and data)

  This method will be called by *Binding Activation* to invoke inter-domain resource optimisation by giving a set of optimisation parameters and data.

- **Binding Activation to Dynamic Inter-domain TE**
  Set_Egress_Configuration (eRAM)

  This method will be called by *Binding Activation* when the decision on which of the established QoS-bindings will be put in effect in the network for implementing e-QC has been made. The method is to pass the management directives to Dynamic Inter-domain Traffic Engineering for which to set up a configuration to realise the decision from binding activation (e.g. using BGP policies).

- **Binding Activation to pSLS Invocation**
  Invoke_pSLS$_{out}$ (eRAM)

  This method will be called by *Binding Activation* to indicate pSLS invocation functional block how much resources are estimated to be invoked or used.

- **Binding Activation to Off-line Intra-domain TE**
  Notify_Inter_TE_Solution (intra-domain configuraiton)

  This method will be called by *Binding Activation* to indicate offline intra-domain TE which intra-domain TE solution (i.e. the intra-domain configuration) has been selected. The purpose of this notification is to enable *Offline Intra-domain TE to* physically configure the network resources, thereby enabling the selected resource allocation.

10.4.2.4 Input and output

This section describes input and output of *Binding Activation*. Input and output are further divided into two types: data and action. Action requires some controls or feedback. The letter used in each item below corresponds to the one in Figure 88.

Input data

a) *eTM*: The extended traffic matrix produced by Traffic Forecast. Inter-domain traffic is aggregated based on ingress router and destination prefix. Each entry in eTM includes a QC, an ingress router, a destination prefix and an aggregated QoS demand. The time scale of this eTM is smaller than the eTM that has been input to binding selection.

b) *pSLSs$_{out}$*: A set of established pSLSs$_{out}$ produced by *Binding Selection*. The required fields include destination prefix, o-QC, bandwidth availability and egress interface.

c) *QC mapping compatibility:* A corresponding QC mapping compatibility for each pSLS$_{out}$. The mapping compatibility describes a set of eligible l-QCs maps to a specific o-QC.

d) *eRAM(s):* The extended resource availability matrix produced by inter-domain resource optimisation. It is the inter-domain traffic engineering solution that specifies estimates of the availability of the inter-domain resources (e.g. pSLSs$_{out}$) to accommodate QoS traffic towards the upstream service-peering domains. *Binding Activation* may receive a set of inter-domain TE solutions (i.e. eRAM(s)) from inter-domain *Resource Optimisation*.

Copyright © MESCAL Consortium, June 2005

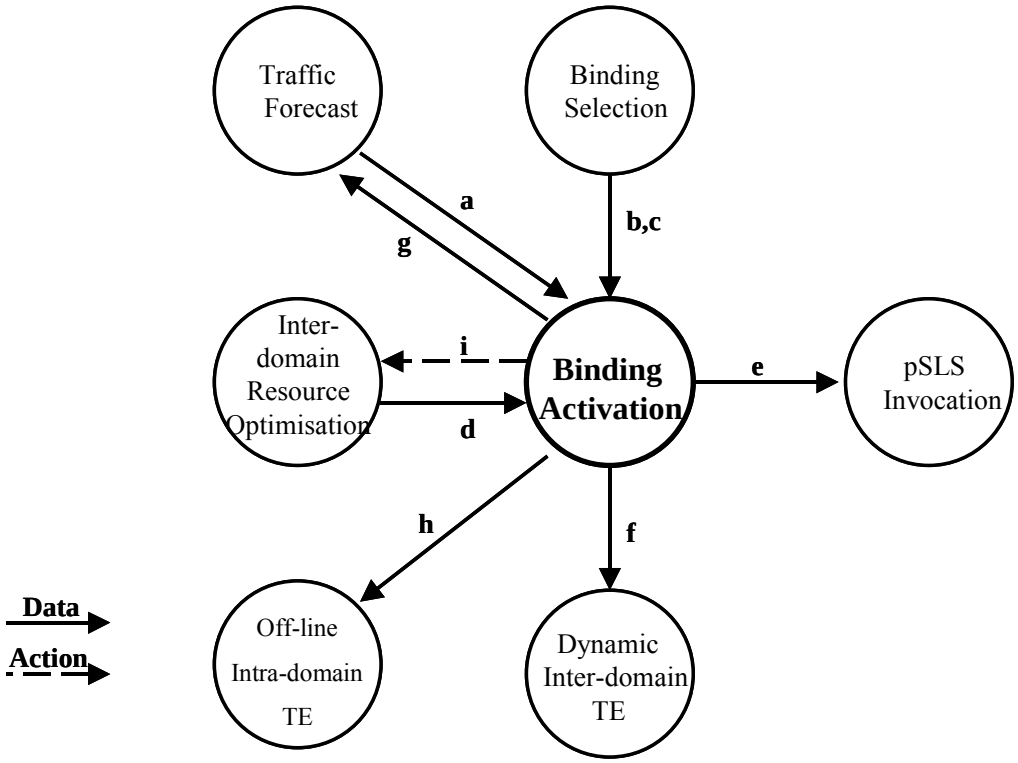


Figure 88. Binding Activation: Input, process and output

Output data

e) *eRAM*: It is the selected inter-domain TE solution among the set produced by inter-domain *Resource Optimisation* (i.e. eRAM∈eRAM(s)). This eRAM is passed to pSLS invocation to indicate how much resources will be actually invoked or used.

f) *eRAM:* It is equivalent to the output data (e). This eRAM is passed to *Dynamic Inter-domain Traffic Engineering* for which to enforce the *Binding Activation* decision/inter-domain TE solution.

g) *eRAM*: The eRAM is passed to *Traffic Forecast*, which will use it to update the iTM (the purpose of this is to notify *Traffic Forecast* of which configuration has been selected by *Binding Activation*).

h) *eRAM*: The eRAM is passed to *Offline Intra-domain TE*. This allows *Offline Intra-domain TE* to select its intra-domain configuration corresponding to this selected inter-domain traffic solution, and to configure the physical network.

Output action

i) *Start_Inter_Rsr_Optimisation*: An action to start the inter-domain *Resource Optimisation*. The following data is also passed together with the action to the inter-domain *Resource Optimisation*:

- *eTM*
- $pSLS_{out}$*={egress interface, o-QC, bandwidth, destination prefix}*
- *QC mapping compatibility={o-QC, {l-QC}}*

Copyright © MESCAL Consortium, June 2005

10.4.2.5 Process summary

This section briefly describes how *Binding Activation* works in general. The process is divided into following steps: (words in bold are data and actions that have been defined in the previous section)

1. *Binding Activation* receives **eTM**, a set of $\mathbf{pSLSs_{out}}$ and the corresponding **QC mapping compatibility** as input from traffic forecast and binding selection.
2. *Binding Activation* invokes inter-domain *Resource Optimisation* by sending a **Start_Inter_Rsr_Optimisation** action. The data together with the action is also passed to inter-domain *Resource Optimisation*.
3. Inter-domain *Resource Optimisation* takes the input data and produces a set of optimal inter-domain traffic engineering solutions (i.e. **eRAM(s)**). The solutions are returned to *Binding Activation*.
4. *Binding Activation* receives a set of optimal inter-domain traffic engineering solutions from inter-domain *Resource Optimisation* and selects the best one. The best solution (i.e. **eRAM**$\in$eRAM(s)) is then passed to
   - *Traffic Forecast* for which to update the iTM as egress points have been identified for each inter-domain traffic.
   - *pSLS Invocation* for which to indicate how much resources will be invoked or used.
   - *Dynamic Inter-domain Traffic Engineering* for which to realise the inter-domain TE solution by configurations.
   - *Offline Intra-domain Traffic Engineering* for which to select the corresponding intra-domain configuration according to the inter-domain TE solution.

10.4.2.6 Algorithm description

The *Binding Activation* algorithm is presented below as pseudo-code.

/* Call *Resource Optimisation* */

read in (e-TM, *S*, *QC_mappings*)

call (res_optimisation (parms passed: e-TM, *S*, *QC_mappings*, *f(j,q)*; parms returned: Ψ, e-RAM, *SLS_bandwidths*))

/* Call *pSLS Invocation* */

call (pSLS_invocation (*S*, *SLS_bandwidths*)

Copyright © MESCAL Consortium, June 2005

10.4.3 Inter-domain Resource Optimisation

10.4.3.1 Objectives

Inter-domain *Resource Optimisation* computes an optimal inter-domain traffic engineering solution, taking the predicted inter-domain traffic matrix (e-TM) and intra/inter-domain resources as input. It may produce multiple optimal inter-domain TE solutions and returns the solutions to *Binding Activation* or *Binding Selection* which in turn will select and implement the best one.

The objective of inter-domain *Resource Optimisation* is to map the predicted inter-domain traffic matrix to the inter-domain network resources, satisfying QoS requirements while aiming at optimising the use of network resources within or across AS boundaries.

10.4.3.2 Interface specification

This section describes the interaction of inter-domain *Resource Optimisation* function block with the others through events, messages or signals. Figure 89 shows the interfaces related with inter-domain *Resource Optimisation*.

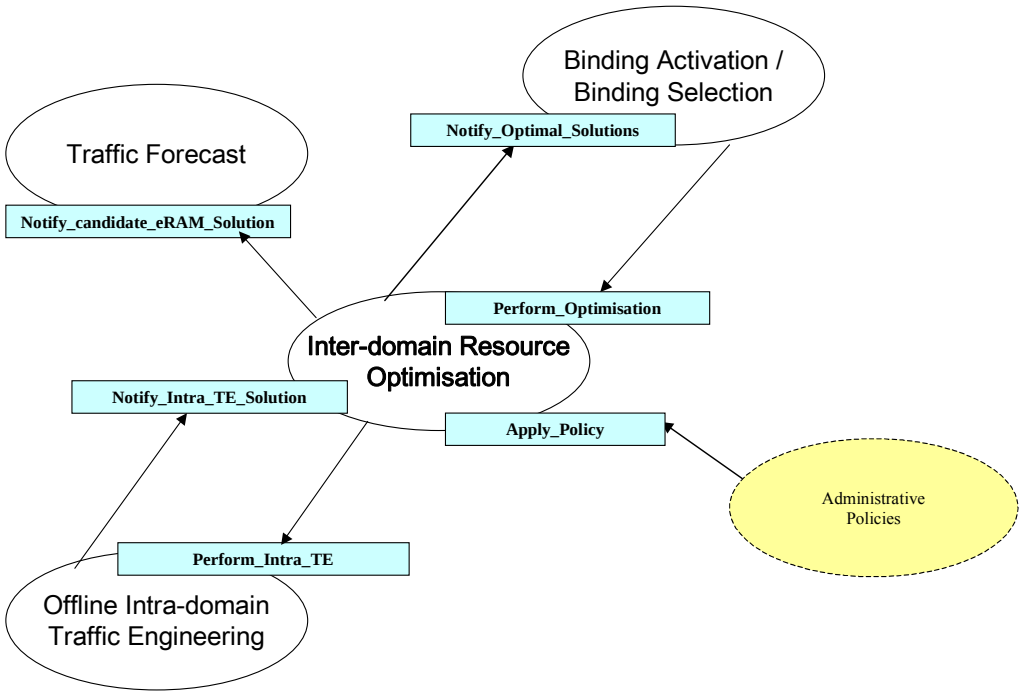


Figure 89. Inter-domain Resource Optimisation interfaces

- **Binding Activation / Selection to Inter-domain Resource Optimisation**
  Perform_Optimisation (optimisation parameters and data)

  This method will be called by *Binding Activation* or *Binding Selection* to invoke inter-domain *Resource Optimisation* by giving a set of optimisation parameters and data.

- **Inter-domain Resource Optimisation to Binding Activation / Selection**
  Notify_Optimal_Solutions (eRAM(s))

  This method will be called by inter-domain *Resource Optimisation* to return to *Binding Activation* or *Binding Selection* a set of optimal inter-domain traffic engineering solutions (eRAM(s)).

Copyright © MESCAL Consortium, June 2005

- **Administrative Policies to Inter-domain Resource Optimisation**
  Apply_Policy (function + parameters)

  Any administrative policies that can affect the decision-making of inter-domain *Resource Optimisation*

- **Inter-domain Resource Optimisation to Traffic Forecast**
  Notify_Candidate_eRAM_Solution (eRAM)

  This method will be called by *Inter-domain Resource Optimisation* to notify to *Traffic Forecast* a draft or candidate eRAM configuration (*Traffic Forecast* will in turn use this to calculate a corresponding iTM and pass the iTM to *Offline Intra-domain Traffic Engineering* for calculation of cost, intra-domain resource availability and utilisation).

- **Inter-domain Resource Optimisation to Offline Intra-domain TE**
  Perform_Intra_TE (eRAM flag)

  This method will be called by *Inter-domain Resource Optimisation* to notify *Offline Intra-domain Traffic Engineering* of a request to calculate intra-domain resource availability and utilisation. *Offline Intra-domain Traffic Engineering* will use the given candidate inter-domain TE solution (eRAM) passed to *Traffic Forecast* (Note that *Offline Intra-domain TE* obtains the iTM from *Traffic Forecast*, allowing the Intra-domain algorithm to calculate the iRAM).

- **Offline Intra-domain TE to Inter-domain Resource Optimisation**
  Notify_Intra_TE_Solution (Intra domain cost Φ', and *intra-domain configuration)*

  This method will be called by *Offline Intra-domain Traffic Engineering* to return to inter-domain *Resource Optimisation* the intra-domain cost Φ and (optionally) the corresponding traffic engineering solution.

10.4.3.3 Algorithm description

We use a decoupled (i.e. not integrated) approach, to select for each of the aggregate inter-domain traffic flows the optimum egress router selection point and the optimum o-QC.

For Inter-domain TE we define "optimum" to mean:

- Avoiding overloading some inter-domain resources while others are underloaded, and
- Minimising the total cost of using all the inter-domain resources; and
- Minimising the intra-domain cost.

Note that the details of how the intra-domain cost is optimised is not considered further here, since our decoupled / distributed approach means that we consider separately inter-domain and intra-domain optimisation.

10.4.3.3.1 Nomenclature

I	Set of ingress routers
i	One such ingress router
J	Set of egress router interfaces (one router may have several links to different peer domains, and we need to distinguish between these various links)
j	One such egress router interface
K	Set of destination prefixes
k	One such destination prefix
c(j,q)	Egress capacity on router interface *j* allocated to o-QC *q*

Copyright © MESCAL Consortium, June 2005

p(j,q) Allowed overbooking ratio on router interface *j* for o-QC *q*;
p(j,q) = {max booked data rate on link *j* for o-QC *q* } / *c(j,q)*

b(q,j) Inter-domain link bandwidth assigned to o-QC *q* on egress router interface *j*, and is therefore the bandwidth that is assigned to a particular pSLS

x(i,k,e,j) =1 if aggregate traffic flow *t(i,k,e)* uses egress router *j*, else =0 otherwise

f(j,q) Inter-domain link cost on egress router interface *j* using o-QC *q*

Φ Intra-domain cost for aggregate flows, calculated from values returned from *Intra-domain TE* (i.e., Φ' and/or the intra-domain configuration, see Section 10.4.3.2

Ψ Inter-domain cost function

Q The set of o-QCs

q One such o-QC, $q \in Q$

E The set of e-QCs

e One such e-QC, $e \in E$

T The set of aggregated flows described in the External Traffic Matrix (eTM)

t(i,k,e) One such flow, $t \in T$, being of data rate (=bandwidth) *t*, from ingress *i* to destination address *k*, with e-QC *e* (note there may be two or more flows from *i* to *k*, with different e-QCs *e*)

S Set of pSLSs (may use subscripts: $S_{existing}$; S_{new})

s pSLS, $s \in S$, *s=(j, k, b(q,j), q)*

w(e,q) =1 if e-QC *e* is bound with o-QC *q*, =0 otherwise

10.4.3.3.2 Problem formulation

10.4.3.3.2.1 Binding Selection / Activation implications on Resource Optimisation

For *Binding Selection*: not all pSLSs are fixed; we can select one or more o-QCs for each egress router / destination prefix combination *(j,k)*, and the bandwidth is not fixed.

For *Binding Activation*: all pSLSs are fixed, but the bandwidth made available to a given pSLS at an egress router interface is assumed to be variable within some range $b_{min}(q,j)$ to $b_{max}(q,j)$.

10.4.3.3.2.2 Formulation

We seek to identify the "best" (i.e. least cost to our domain) set of e-QCs for each of the |*M*| values in the eTM.

Based on the above discussion, the objectives of inter-domain *Resource Optimisation* are as follows:

- Minimise the maximum inter-domain link cost, i.e. minimise $\max_{j \in J} \sum_{q \in Q} f(j,q)$
- Minimise the total cost of all the inter-domain links, i.e. minimise $\sum_{j \in J} \sum_{q \in Q} f(j,q)$
- Minimise the intra-domain cost, Φ

subject to the following constraints:

- Inter-domain egress capacity constraint for each link *j*: the sum of bandwidths assigned to all the QoS classes (or pSLSs) on each egress router interface must not exceed the interface's bookable capacity *p(j,q)c(j,q)* summed over all QoS classes. In other words, $\forall j$, $\sum_{q \in Q} b(q,j) \le \sum_{q \in Q} p(j,q)c(j,q)$

Copyright © MESCAL Consortium, June 2005

- pSLS capacity assignment constraint: the sum of the aggregate traffic flows of each o-QC must not exceed the bandwidth assigned to that pSLS: $\forall q, j, k$, $\sum_{k \in K} \sum_{i \in I} t(i,k,e)x(i,k,e,j)w(e,q) \le b(q,j)$
- Bandwidth range limits constraint: for *Binding Activation* the constraint is: $b_{\min}(q,j) \le b(q,j) \le b_{\max}(q,j)$. For *Binding Selection* the constraints are as follows: b_{min}=0; $\sum_{S_{existing}+S_{new}} b_{\max}(q,j) \le \sum_{q \in Q} c(j,q)$
- Intra-domain constraints: are not considered here since they are dependent on the *Offline Intra-domain TE* functional block.

10.4.3.3.2.3Inter-domain cost functions

The analysis is presented in terms of a generic cost function *f(j,q)*. In testing our algorithms, we employ two specific cost functions:

- A pSLS cost model that considers the cost incurred by neighbouring domains of providing a given level of QoS. For a single pSLS, the cost is given by $C_s \sum_{i \in I} t(i,k,e)x(i,k,e,j)w(e,q)$ where C_s is the cost per unit bandwidth of pSLS *S*. The total pSLS cost Ω summed across all pSLSs is therefore $\Omega = \sum_{S_{existing}+S_{new}} C_s \sum_{i \in I} t(i,k,e)x(i,k,e,j)w(e,q)$.
- An inter-domain link utilisation cost function that considers explicitly the objective of minimising the overall inter-domain link utilisation. This reduces maximum queuing delays and allows for statistical fluctuations in traffic beyond that forecast in the traffic matrix. We use a cost function such as that defined in [FORT02a], whose cost for a single link is $\theta(x)$ where x is the link utilisation. The inter-domain link utilisation cost function Θ is then $\Theta = \sum_{j \in J} \theta(x_j)$.

The generic inter-domain cost function *f(j,q)* can then be either the pSLS cost Ω or the inter-domain link utilisation cost function Θ or their sum, or in general, some other function.

10.4.3.3.2.4Analysis

We observe that our formulation of the Inter-domain resource optimisation problem is similar to the well-studied Generalised Assignment Problem (GAP). Note that the GAP is known to be NP-hard. In the following, we will show the equivalence of our problem with GAP for the case of resource allocation during Binding Activation. As a result, our problem is also NP-hard. For resource allocation during Binding Selection, the problem is even more difficult.

The formulation of the GAP can be described in general terms as follows:

Let I = {1,2,..,m} be a set of agents, and let J={1,2,…,n} be a set of jobs. For $i \in I$, $j \in J$, define γ_{ij} as the cost of assigning job *j* to agent *i* (or assigning agent *i* to job *j*), ρ_{ij} as the resource required by agent *i* to perform job *j*, and β_i as the resource availability (capacity) of agent *i*. Also χ_{ij} is a 0-1 variable that is 1 if agent *i* performs job *j* and 0 otherwise. The mathematical formulation of the GAP is:

Minimise $$\sum_{i \in I} \sum_{j \in J} \gamma_{ij} \chi_{ij} \quad (1)$$

Subject to $$\sum_{i \in I} \chi_{ij} = 1, \forall j \in J \quad (2)$$

Copyright © MESCAL Consortium, June 2005

$$\sum_{j \in J} \rho_{ij} \chi_{ij} \leq \beta_i, \forall i \in I \quad (3)$$

$$\chi_{ij} \in \{0,1\}, \forall i \in I, \forall j \in J \quad (4)$$

Equation (2) ensures that each job is assigned to exactly one agent and (3) ensures that the total resource requirement of the jobs assigned to an agent does not exceed the resource capacity of the agent.

In the following, we describe the transformation process of our problem to GAP.

- Each pSLS with bandwidth *b(q,j)* maps to a GAP agent with capacity β_i.
- Each eTM aggregated traffic flow of data rate or bandwidth *t(i,k,e)* maps to a GAP job of size ρ_{ij}.
- The cost of assigning each aggregated traffic flow to each pSLS is a function of inter- and intra-domain resource utilisation, and so some combination of *f(j,q)* and Φ corresponds to the GAP cost γ_{ij}.
- The assignment variable *x(i,k,e,j)* represents whether an aggregate traffic flow uses egress router *j* and therefore corresponds to the GAP variable χ_{ij}.

For the case of resource optimisation during Binding Activation, our problem is thus essentially the GAP of assigning each aggregated traffic flow to exactly one pSLS, so that the total cost of assignment is minimised and each pSLS does not exceed its capacity assignment constraint.

For the case of resource optimisation during Binding Selection, the assignment problem is more complicated than GAP. This is because while the set of jobs (i.e. aggregate traffic flows) is known, the set of agents (corresponding to pSLSs) is not defined: it is the task of the *Binding Selection* function block working with *Resource Optimisation* and *pSLS Ordering* to select and negotiate pSLSs. Thus resource optimisation for Binding Selection is a harder task than GAP.

10.4.3.3.3 Problem solution approaches

In describing our algorithms, we consider here two variations of parameters passed to and from functions that call *Resource Optimisation*. These variations match the two variations described in *Binding Selection* (Section 10.4.1) and are as follows:

1. *Resource Optimisation* is passed a set of existing pSLSs and a set of new candidate pSLSs, and returns simply a cost function Ψ for the given configuration of pSLSs;
2. *Resource Optimisation* is passed only the existing pSLSs, but calculates any optimum set of additional pSLSs; it returns both Ψ and the new pSLSs to the calling function.

In the case of calls from *Binding Activation*, the set of pSLSs is fixed, but the bandwidth made available to a given pSLS is assumed to be variable. This can therefore be treated as a special case of Variation 2, where the only pSLS parameter to be optimised is the bandwidth.

10.4.3.3.3.1 Random assignment

This algorithm is proposed simply to provide a baseline against which the performance of other algorithms can be compared.

For Variation 1, the algorithm consists of the following: an eTM aggregate flow is randomly selected from the eTM, and a binding candidate that satisfies this flow (i.e. a binding candidate for the eTM flow's e-QC) is randomly chosen. The eTM flow is then randomly assigned to a pSLS that (a) meets the binding candidate's o-QC *q*, (b) supports the required destination address prefix *k*, and (c) has sufficient spare bandwidth to accommodate the flow's data rate *t(i,k,e)* (the pSLS capacity assignment constraint). If any flow cannot be assigned, the solution is discarded and a further attempt at randomly assigned traffic flows is made. Once the traffic flows have been assigned to pSLSs, the intra-domain

Copyright © MESCAL Consortium, June 2005

cost Φ and the inter-domain cost $\sum_{j\in J}\sum_{q\in Q} f(j,q)$ are calculated, and a weighted cost function, say $\Psi = \alpha\Phi + (1-\alpha)\sum_{j\in J}\sum_{q\in Q} f(j,q)$, is calculated and returned. If the Inter-domain assignments cannot all be satisfactorily accommodated after *N* attempts (value of *N* to be decided) the algorithm fails.

Variation 2 proceeds as for Variation 1, with the following modification. If a flow cannot be assigned, in the case of a call from *Binding Selection*, an additional pSLS (*j, k, b(q,j), q*) that allows the flow to be assigned is randomly generated. Specifically, the pSLS parameters *k* and *q* are assigned to match the flow, *j* is randomly selected from the set of information provided by the *QoS Capabilities Discovery* function block from those domain peers that can reach destination address *k* with o-QC *q*, and the pSLS bandwidth *b(q,j,)* is randomly selected in a range between the flow's required bandwidth *t(i,k,e)* and the available spare provisionable bandwidth on the egress interface $\sum_{q\in Q} p(j,q)c(j,q) - \sum_{existing pSLSs+new pSLSs} b(q,j)$ where the summation is over the existing and new pSLSs at egress interface *j*. This available spare provisionable bandwidth corresponds to the Inter-domain egress capacity constraint. If the required bandwidth exceeds the available spare provisionable bandwidth, then a different pSLS is randomly generated; if a pSLS cannot be generated after *N* attempts (value of *N* to be decided) the algorithm fails. In the case of a call from *Binding Activation*, bandwidth is assigned in a pSLS as required up to the maximum for that pSLS, while meeting the Inter-domain egress capacity constraint. If the constraint cannot be met, a different pSLS is used; if the constraint cannot be met on any pSLS, the algorithm fails.

10.4.3.3.3.2 Brute force

This algorithm is proposed to determine the lowest cost function Ψ, although it is expected to be slow running and not a viable solution in practice. It is however intended to provide a useful baseline for comparing the behaviour and efficiency of algorithms in relatively small networks.

For Variation 1, the algorithm consists of the following. Traffic flows in the eTM are assigned to pSLSs. The constraints are tested (Inter-domain egress capacity constraint, pSLS capacity assignment constraint, bandwidth range limits constraint), and if the configuration fails any constraint it is discarded. If the constraints are met, the intra-domain cost Φ and the inter-domain cost $\sum_{j\in J}\sum_{q\in Q} f(j,q)$ are calculated, and a weighted cost function, say $\Psi = \alpha\Phi + (1-\alpha)\sum_{j\in J}\sum_{q\in Q} f(j,q)$, is calculated. This process is in turn repeated for every combination of assignment of traffic flows to pSLSs. The configuration that gives the lowest cost function is then the optimum configuration, and its parameters are returned.

Variation 2 would require a brute force attack on selecting a number of new pSLSs, assigning values to each of the pSLS parameters (*j, k, b(q,j), q*). For calls from *Binding Selection*, this could be achieved if each parameter consisted of one of a finite number of values, and this can be approximated as follows:

- *j*: can take the value of each egress router interface ID;
- *k:* to be any of the destination address prefixes that peer domains have advertised as being reachable from egress router interface *j* with o-QC *q*, according to information provided to *Resource Optimisation* by the *QoS Capabilities Discovery* function block.
- *b(q,j)*: if *z* new pSLSs are being generated for a single egress interface *j*, then share the available spare provisionable bandwidth equally between the *z* new pSLSs at that interface. That is, set $b(q,j) = \left(\sum_{q\in Q} p(j,q)c(j,q) - \sum_{existing pSLSs+new pSLSs} b(q,j)\right)/z$.

Copyright © MESCAL Consortium, June 2005

- *q*: can take the value of each o-QC advertised by peer domains and provided to *Resource Optimisation* by the *QoS Capabilities Discovery* function block.

For calls from *Binding Activation*, only the parameter *b(q,j)* needs to have a value assigned to it. A simple model is to take the provisionable bandwidth $\sum_{q \in Q} p(j,q)c(j,q)$, divide it into *z* equal sized portions, and allocate bandwidth to pSLSs in multiples of $\left(\sum_{q \in Q} p(j,qc(j,q) \right) / z$.

Allowing for the above sets of parameters, Variation 2 proceeds as described in Variation 1.

10.4.3.3.3.3Genetic algorithm

Genetic algorithms provide a heuristic mechanism for solving complex optimisation problems ([Chu96], [ERIC02], [LIN03]). Each potential solution to a problem is represented by a set of values known as a chromosome; for example, in our case the chromosome might consist of the assignment of each aggregate traffic flow in an eTM to a particular egress router or a particular pSLS. The chromosome is composed of individual genes; in our case a gene would be the assignment of a single aggregate traffic flow to its egress router or pSLS. The genetic algorithm heuristic comprises the following. An initial population of *N* randomly generated chromosomes is generated. Each of these solutions is then used as the configuration of the system under investigation (in our case, the inter-domain configuration), and a fitness function is calculated that quantifies the "goodness" of the solution represented by chromosome. Once the fitness function has been calculated for all *N* chromosomes a new generation of chromosomes is produced, as follows.

The chromosome population is divided into three sections: the best, the medium, and the worst. The best chromosomes are passed unchanged to the next generation. The poorest chromosomes (i.e., those with the worst cost function) are discarded. Processes of crossover and mutation (described below) are applied to the best and medium chromosomes to generate new chromosomes for the next generation. This process results in a new population of *N* chromosomes, and the process of generating a new generation is repeated until convergence. The chromosome with the best fitness function is then the best (or fittest) chromosome, and represents the best obtained configuration of the system under investigation.

In the crossover process, two chromosomes are randomly selected, one from the best section of the population, and one from the medium section. Genes are randomly selected from each of the chromosomes to generate a new chromosome. The probability of selecting from the fitter parent chromosome (i.e. from the "best" section) is called the crossover probability, p_c. In mutation, genes are randomly changed, with some mutation probability p_m.

The effectiveness and convergence rate of the genetic algorithm depends on the values of *N*, p_c and p_m. Previous research suggests typical satisfactory values to be $150 \le N \le 300$, $0.5 \le p_c \le 0.8$, and $0.001 \le p_m \le 0.1$ [Lin,03].

Our proposed work improves on and differs from existing genetic algorithms in the following respects:

- Genetic algorithms have not previously been used for Inter-domain traffic engineering;
- We consider chromosomes composed of multiple *different* gene types (for Variation 2);
- We consider variable numbers of genes in a chromosome as part of our optimisation process (for Variation 2).

For Variation 1 described above, each gene assigns an eTM aggregate traffic flow to a given pSLS.

For Variation 2 described above, there are two types of gene: (a) a gene that assigns an eTM aggregate traffic flow to a given pSLS (as for Option 1); and (b) a gene that contains a pSLS configuration. In the case of a call from *Binding Selection*, all parameters in the pSLS are part of the gene, i.e. *(j, k, b(q,j,k,), q)*. In the case of a call from *Binding Activation*, the pSLS parameters that are variable are limited to the bandwidth, *b(q,j,k)*. Since part of the optimisation process for *Binding Selection* is in

Copyright © MESCAL Consortium, June 2005

Variation 2 is the number of additional pSLSs (as well as their values), we allow the number of type (b) genes to be variable when the function is called by *Binding Selection*.

Let the fitness function be a combination of the intra-domain cost Φ and the inter-domain cost $\sum_{j\in J}\sum_{q\in Q} f(j,q)$. For example, we could take a weighted sum of these two components, $\Psi = \alpha\Phi + (1-\alpha)\sum_{j\in J}\sum_{q\in Q} f(j,q)$, and then seek to minimise the fitness function as we iterate the genetic algorithm.

If any of the constraints described in Section 10.4.3.3.2.2 are not met by a particular chromosome then we set the fitness function = 0 for that chromosome.

10.4.3.3.3.4 Heuristic algorithms

We propose three greedy-based heuristic algorithms for *Inter-domain Resource Optimisation*. These are a greedy-cost heuristic, a greedy-penalty heuristic, and a greedy-random heuristic. These are now described.

10.4.3.3.3.4.1 Greedy-cost heuristic

This sorts the eTM aggregate traffic flows in descending order based on their bandwidth requirements and selects one at a time in that order.

Step 1: We evaluate each of the pSLSs individually to determine its feasibility for the eTM traffic. We refer to this step as *pre-selection*. Pre-selection is based on the information such as destination address prefix and available bandwidth of pSLSs. The pSLS *q* is feasible if it meets the required destination address prefix *k* and has sufficient spare bandwidth to accommodate the flow's data rate *t(i,k,e)* (satisfy the pSLS capacity assignment constraint).

Step 2: Among a set of feasible pSLSs identified in step 1, we select a pSLS with the minimum cost. The cost of assigning the eTM traffic to a pSLS is determined by an objective function and the function may include the optimisation of one or more objectives defined in 1.1.4.4, for example, $\Psi = \alpha\Phi + (1-\alpha) f(j,q)$. The eTM traffic under consideration is then assigned to the selected pSLS.

Step 3: We search for a second eTM traffic and repeat step 1 to 3. We iterate until all the eTM traffic have been considered.

10.4.3.3.3.4.2 Greedy-penalty heuristic

It is possible that assigning an eTM traffic to a pSLS in different orders generates different selection scenarios. For example, if we assign the eTM traffic *t(i,k,e)*=2 in the first place, we can assign it greedily to pSLS *q* with the cost equals to 6; if we delay allocating it for a while, however, pSLS *q* may not have sufficient bandwidth because its bandwidth has been allocated to other eTM traffic and the original eTM traffic has to be assigned to pSLS *q'* with the cost equals to 12; in this case, we are penalising the consumption of additional bandwidth and we use *plt* to refer to the penalty value. A penalty-based algorithm aims to minimise the cost by placing eTM traffic in certain order according to *plt*. We propose a similar algorithm called Greedy-penalty heuristic as follows. Such an algorithm is also used to solve the Generalised Assignment Problem (GAP).

Step 1: For each eTM traffic, we measure the desirability of assigning it to each feasible pSLS that meets the required destination address prefix *k* and has sufficient spare bandwidth to accommodate the flow's data rate *t(i,k,e)* (satisfy the pSLS capacity assignment constraint). The desirability is the cost of assigning the eTM flow to a specific pSLS on an egress interface, for example, $\Psi = \alpha\Phi + (1-\alpha) f(j,q)$. In this case, the smaller the desirability, the better for the selection.

Step 2: Compute *plt* for each unassigned eTM traffic, which is the different between the desirability of the eTM traffic's best and second best selection (i.e. the two pSLSs which yield the smallest desirability). If there is only one feasible pSLS to accommodate the eTM traffic, we need to assign the

Copyright © MESCAL Consortium, June 2005

eTM traffic to it. Otherwise, this currently feasible pSLS may become unfeasible afterwards, having been assigned to accommodate other eTM traffic, which leads to insufficient bandwidth so that we would reject the eTM traffic. In this case, we set *plt* to infinite.

Step 3: Among all unassigned eTM traffic, the one yielding the largest *plt* is placed with its best selection. If multiple eTM traffic have the same largest penalty, they are placed in the order of decreasing bandwidth requirement.

Step 4: We iterate step 1 to step 4 until all the eTM traffic have been considered.

10.4.3.3.3.4.3 Greedy-random heuristic

As with the Greedy-cost heuristic, this sorts the eTM traffic in descending order based on their bandwidth requirements and selects one at a time in that order. It is identical to Greedy-cost heuristic except that the selection of pSLS is done at random, with uniform probability among all the feasible pSLSs. We consider this algorithm as the behaviour of the current BGP for solving our selection problem. The current non-TE BGP will select an egress router with respect to bandwidth information completely at random.

10.4.3.3.3.4.4 Cost function

The three proposed heuristic algorithms use cost function to determine the cost of assigning eTM traffic to pSLSs. According to the cost value, the most appropriate pSLS is assigned to each eTM traffic flow. We consider the following potential cost functions (also called selection policies):

- ***Closest-egress-first***: selects the pSLS with which the associated egress interface is the closest to the ingress router where the eTM traffic enters the ISP domain, using hop count as cost function. If there are several such pSLSs, the one with the maximum intra-domain path bottleneck bandwidth is selected. If there are several such pSLSs with the same bottleneck bandwidth, the one with the maximum inter-domain link available bandwidth is selected. This variation of the traditional shortest path policy provides a widest-shortest solution.
- ***Widest-egress-first***: selects the pSLS with which the associated egress interface has the maximum bottleneck bandwidth on the path leading to the ingress router (i.e. using the maximum bottleneck bandwidth of intra-domain path as cost function). If there are several such egress routers, the one with the minimum hop count is selected. If there are several such pSLSs with same minimum hop count, the one with the maximum inter-domain available bandwidth is selected. This variation of the traditional widest-path policy provides a shortest-widest solution.
- ***Least-loaded-egress-first***: selects the pSLS which has the minimum loading (i.e. using the loading of pSLS as cost function)
- ***Shortest-dist-egress-first***: We define two distance factors for intra-domain and inter-domain resources to quantify their available bandwidth, and then define a distance of selecting the pSLS from the ingress router. The distance factor of intra-domain link *l* is defined as

$$DF\_Intra(l) = \frac{1}{\left(bw_{intra}^{l} - t(i,k,e)\right)^{\alpha}}$$

where bw_{intra}^{l} is the available bandwidth of intra-domain link *l*. The distance factor of the pSLS *p* associated with egress interface *j* that accommodates the eTM traffic is defined as

$$DF\_Inter(j,q) = \frac{1}{\left(bw_{inter}^{j,q} - t(i,k,e)\right)^{\alpha}}$$

where $bw_{inter}^{j,q}$ is the available bandwidth of the pSLS *q* (i.e. the available bandwidth of *b(q,j,k)*) associated with egress interface *j*. The parameter α in the distance factor represents the degree

Copyright © MESCAL Consortium, June 2005

to which a lightly loaded resource is favoured over a congested resource. This observation leads us to a hint that we could find an optimal α such that shortest-dist-egress-first yields the best performance on a specific objective. We define the distance of selecting the pSLS q associated with egress interface j from the ingress router where the customer traffic enters to be

$$\sum_{l \in Intrapath} DF\_Intra(l) + DF\_Inter(j,q)$$

where l is the link used by the selected intra-domain path *Intrapath* between the ingress router and the egress interface j where the considered pSLS associated with, and the optimality clause for shortest-dist-egress-first is to select the egress router with the minimum distance.

We summarise the characteristic of each egress router selection policy as follows. Closest-egress-first gives high priority to limit the hop count that a traffic flow must traverse. The motivation for choosing closest-egress-first as the weight cost function is to minimise the total resource consumption on assigning eTM traffic to pSLSs. The widest-egress-first gives high priority to spread the load evenly among the links within the domain. The motivation is to achieve resource load balancing in the network. The least-loaded-egress-first gives high priority to balance the load among inter-domains links of the domain. Shortest-dist-egress-first dynamically balances the impact of hop count and path load on making the selection decision. Table 17 shows the objective focus of each selection policy. The objectives refer to the items defined in section 10.4.3.3.2.2.

Selection policy	Objective focus
Closest-egress-first	Minimise the intra-domain cost
Widest-egress-first	Minimise the intra-domain cost
Least-loaded-egress-first	Minimise the maximum inter-domain link cost Minimise the total cost of all the inter-domain links
Shortest-dist-egress-first	Minimise the maximum inter-domain link cost Minimise the total cost of all the inter-domain links Minimise the intra-domain cost

Table 17 Objective focus of selection policies

Copyright © MESCAL Consortium, June 2005

10.5 Dynamic Inter-domain TE

10.5.1 q-BGP

10.5.1.1 Introduction

The deployment of Internet was a success thanks to a fruitful cooperation between several categories of actors especially service providers, standardisation bodies, regulators and equipment manufacturers. Network Providers have deployed standard inter-domain routing protocols in order to convey reachability information between their domains. The existence of such standard protocols has facilitated interconnection between distinct autonomous systems and then allowed to reach remote destinations located beyond the boundaries of a single INP.

Nowadays, the panorama of required information to be exchanged between Network Providers, explicitly with their respective domains, is different from what could be exchanged thanks to existing inter-domain routing protocols. In other words, reachability information should be richer than what is exchanged via current routing protocols and should provide routers with pertinent information in order to help the route selection decision-making process. Such information could be for instance the QoS that will be experienced along a given path. From this perspective, it is obvious that Network Providers have to evolve and update the protocols (not necessary change the core of the operational mode of these protocols but exploits as much as possible extensibility capabilities of existing protocols) they are used to deploy in their domains in order to meet the new requirements and then to be able to offer new sophisticated added value services.

QoS delivery services are seen as a part of future Internet services (*see [ATKI03], the IAB has qualified these services as critical*). In order to offer such services, network infrastructures (network devices capabilities, protocols, management tools, etc.) must be updated to offer this type of services. From a pure control plane viewpoint, modifications need to be brought to existing signalling and routing protocols. Particularly from an inter-domain standpoint, this is critical since Network Providers should deploy means to convey QoS-related information between their domains so that QoS-based services could have a world-wide scope and then be accessible for a large set of customers in the world *(the notion of "world" doesn't mean geographical location but the affiliation to any Service and Network provider)*. This could be considered as an important risk since Network Providers have to ensure backward compatibility with existing protocols and deployed technologies. This risk must be considered carefully when designing a solution claiming to meet the new requirements.

In this section, we describe a proposal that aims at exchanging QoS-related information between adjacent ASs. The QoS-related information exchange occurs either at the service level or at the routing level. The place this exchange occurs depends on the deployed inter-domain QoS delivery solution. Two groups of QoS delivery solutions have been identified and are detailed hereafter:

- The first group of solutions requires propagating only an identifier that has been agreed during the pSLS negotiation phase. Additional QoS performance characteristics were negotiated but not exchanged in the routing level. In the rest of this document this will be denoted by *group-1*.
- The second group requires the propagation of a set of QoS performance characteristics associated with an identifier. The nature of the QoS-related information to be exchanged has to be agreed in the pSLS negotiation phase. In the rest of this document this will be denoted by *group-2*.

This section describes a proposal that benefits from the extensibility capability offered by the Border Gateway Protocol (BGP) [RFC1771] and that meets a set of generic requirements described below. The aforementioned proposal could apply for any kind of inter-domain QoS delivery solution that is based on an exchange of QoS-related information between domains. In particular, within the context of MESCAL, the proposal meets "Solution Options"-specific requirements.

Copyright © MESCAL Consortium, June 2005

The MESCAL Solution Options requirements on QoS-related information will be studied in details and then each Solution Option will be classified according to the aforementioned grouping.

This section is organised as follows: Sub-section 10.5.1.3 lists goals and needs of a means allowing exchange of QoS-related information. Sub-section 10.5.1.4 details the MESCAL Solution Options requirements and the nature of required QoS-related information to be carried in q-BGP messages. Sub-section 10.5.1.5 presents q-BGP specifications in terms of messages and route selection process.

10.5.1.2 Definitions

Within this section the following terms are used as defined below:

- QoS-related information can be expressed in terms of one-way delay, inter-packet delay variation, loss rate, DSCP marking, or a combination of these parameters;
- "QoS service"-related attributes: denotes dedicated q-BGP attributes for the usage of a given QoS service;
- Inter-domain QoS delivery solution is used to denote an inter-domain system that aims at offering inter-domain QoS services.

10.5.1.3 Objectives and Needs

Most of ISPs use the Border Gateway Protocol (BGP) protocol in order to interconnect their ASs with the ones of their peers. It is the unique inter-domain routing protocol that is used in the Internet. Several proposals aiming to enhance the capabilities of BGP to carry more information than what had been included in the BGP specifications have been proposed (e.g. [KEY04] and [EXTC05]). The goal of this section is to describe how BGP could be used as a means to convey QoS-related information between adjacent autonomous systems especially within the context of MESCAL while taking into account the three Solution Options detailed in [D1.1]. The proposed solution is generic and could be applied to any kind of inter-domain QoS delivery solution that is based on the exchange of QoS-related information.

MESCAL Solution Options rely on an exchange of QoS-related information that takes place between adjacent ASs. This exchange occurs at the service level and at the routing level. It consists in negotiating QoS guarantees during the pSLS negotiation phase and then in propagating them (or part of them). The means of exchanging QoS-related information should meet a set of generic requirements: (1) It should be dynamic, scalable and (2) should be able to propagate topology changes without any significant impact on the existing best-effort based network infrastructure.

This section aims at identifying and describing these requirements and examines if they are applicable for all the solution options since each MESCAL solution option could require specific QoS messages and route selection process.

This document doesn't intend to detail the BGP protocol specification or its operational mode. For more information about BGP, the reader can refer to [RFC1771] and other related IDR working group [IDR] RFCs. This document will focus only on QoS-related information that needs to be conveyed by BGP messages and the use of this information by the route selection process.

10.5.1.4 Towards a QoS-inferred BGP

The purpose of this section is to identify the specific requirements of the three solution options described in [D1.1] in order to convey relevant QoS information between autonomous systems. In addition, and from a routing/signalling perspective, we identify additional requirements that are mandatory for each solution option to become operational. From BGP standpoint, the intent is to identify (1) the possible lacks, (2) the information to be carried in the BGP messages and (3) required modifications in order to meet the solution options requirements. These requirements will be taken carefully into account when designing a solution that will apply for any inter-domain QoS delivery solution that is *mainly* based on an exchange of QoS-related information between service providers'

Copyright © MESCAL Consortium, June 2005

domains. In other words, the purpose is to identify the group (See Introduction) each solution belongs to.

In the rest of this document, the resulting modified BGP will be denoted by q-BGP (for QoS inferred BGP). Both the route selection process and BGP attributes will be considered.

10.5.1.4.1 Analysis of the three MESCAL Solution Options needs

10.5.1.4.1.1 The Loose Guarantees Solution Option

10.5.1.4.1.1.1 The Loose Guarantees Solution Option assumptions

The LGSO relies deeply on the use of both q-BGP protocol and the meta-QoS-class concept. The resulting QoS-enabled Internet can be viewed as a set of parallel meta-QoS-class planes running distinct instances of inter-domain routing protocol.

When a service agreement exists, the service peers exchange (at the service level) QoS information about their reachability scopes in term of adhering to meta-QoS-class planes. These agreements impact the routing policy filters and grant the remote service peer to benefit from its neighbour's inter-domain QoS capabilities. Additional policies could be negotiated, for example to restrict the set of authorised destinations or to restrict the routes to be advertised to this peer.

Reminder:

- *The prior establishment of pSLSs conditions the exchange of inter-domain connectivity information per meta-QoS-class.*
- *Each meta-QoS-class is identified by a well-known identifier*
- *Each AS achieves a DSCP swapping operation: the egress AS changes its l-QC DSCP into the agreed Meta -QoS-Class DSCP and the ingress AS point changes the meta-QoS-Class identifier into its l-QC DSCP)*
- *Each AS announces to its service peers the network prefixes that can be reached within each meta-QoS-class plane*

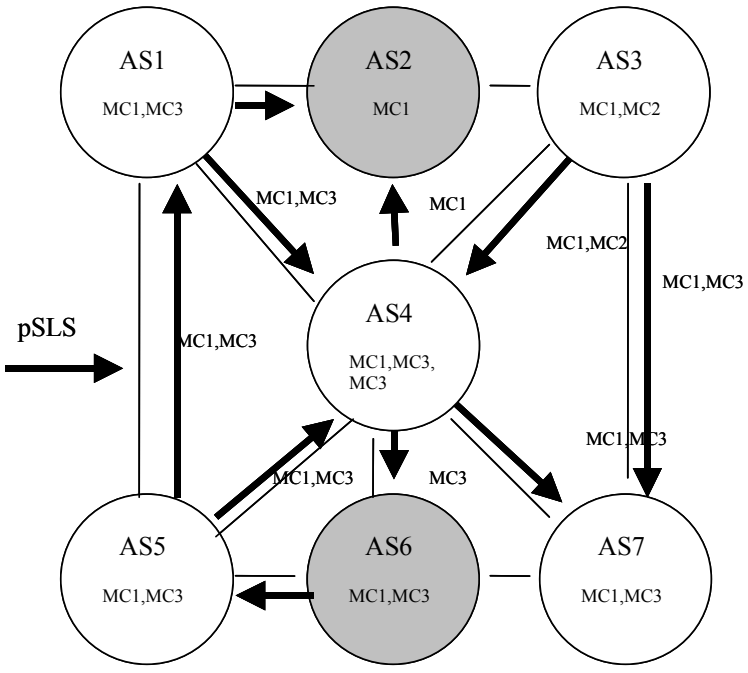


Figure 90. Reachability information exchange *a la* loose solution option

Consequently, BGP UPDATE messages must include a meta-QoS-class identifier, so that each message can be processed within the context of the corresponding meta-QoS-class plane.

Since inter-domain paths depend on the meta-QoS-class used to signal the requested quality of service treatment, it becomes necessary to store the information associated with an individual update in a

Copyright © MESCAL Consortium, June 2005

Routing Information Base (RIB) instance dedicated to the meta-QoS-class the update is intended to. Handling as many Routing Information Bases (RIB) as available meta-QoS-classes also requires that a route selection process runs for each RIB instance in order to select routes that will be stored in the Forwarding information Bases (FIB).

When transiting through a set of ASs, the QoS treatment experienced by a datagram is "*consistent*" in all transited ASs. The notion of "*consistent*" denotes the fact that the treatment received by IP packets in each AS conforms to the corresponding meta-QoS-class definition. It doesn't mean that the QoS characteristics applied to the datagrams when crossing the different ASs are the same but means that they are similar. From this perspective, conveying a meta-QoS-class identifier in q-BGP announcements could be sufficient to learn end-to-end QoS paths. In this case, the BGP route selection process could be kept unchanged but it would not select an optimal path since QoS characteristics resulting from the concatenation of each l-QC encountered along the path would not be present and thus not considered by the selection process. If this information was inserted in q-BGP update messages it could be advantageously taken into account by the q-BGP route selection process to select an optimal path. This indeed would enable to tune more precisely the route selection process in order to select routes according to more sophisticated routing policies.

QoS-related information inserted in q-BGP update messages is intended to facilitate the selection of the best possible end-to-end route. But this information could be of different nature. It could be administratively enforced. In that case it would not change too frequently. Or, it could be much more dynamic (result of a measurement for instance) and in that case the frequency of changes could be much higher.

Administrative setting of QoS values could be achieved either statically or dynamically. If these values are set statically, the behaviour of q-BGP will be static and the route selection process will choose the same route. The QoS-related information doesn't bring major added-value to the final behaviour of the route decision making process and freezes the state of the inter-domain routing. Nevertheless, in the case of QoS performance characteristics values are set administratively or dynamically, providers will deploy mechanisms that monitor the network and then guide the setting of these values. q-BGP will be provided by accurate information in order to select the optimal path. The frequency between two q-BGP router configuration operations in an administrative scheme should not be too small and could be very small in the dynamic scheme. The difference between the administrative and the dynamic setting are summarised in the following figure:

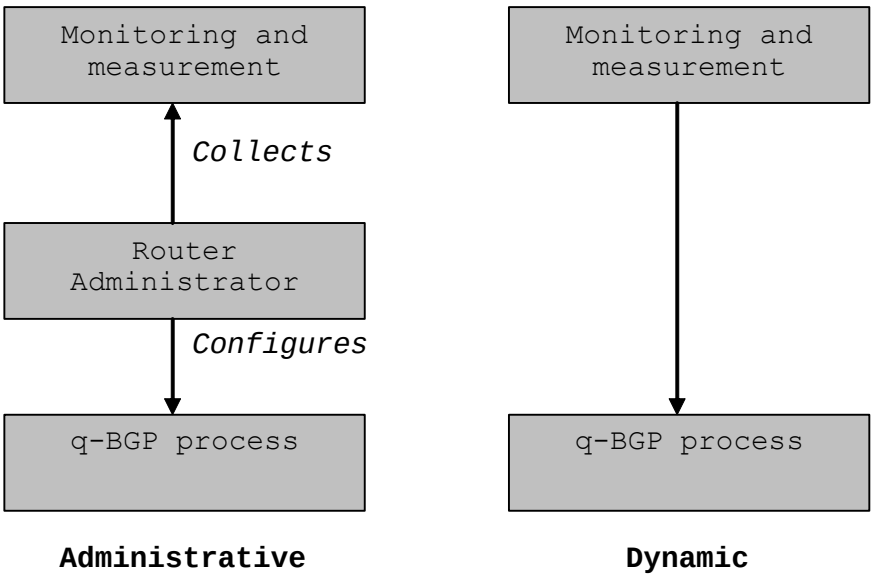


Figure 91. Administrative and dynamic QoS configuration schemes.

Copyright © MESCAL Consortium, June 2005

As a consequence, three sub scenarios need to be studied:

- Scenario 1: q-BGP carries only meta-QoS-class identifiers;
- Scenario 2: q-BGP carries meta-QoS-class identifiers *AND* end-to-end QoS information:
  - Scenario 2-1: QoS information are administratively enforced;
  - Scenario 2-2: QoS information is dynamic and reflects the real status of the network.

10.5.1.4.1.1.2 Only MC identifier is propagated

***Assumption**:* BGP messages carry only meta-QoS-class identifiers.

***Impact on BGP**:* In this context, the route selection process remains the same as the classical BGP one. The main selection criterion is still the AS path length. For each meta-QoS-class plane, a route selection should be executed. Note that the use of AS_PATH as unique criterion to select a path could lead to non optimal routes, because of there is no differentiation between big and small ASs. A variant of this mode is to include a description of the diameter of the AS.

***Example**:* Let's consider the example illustrated in Figure 92

Thanks to the pSLSs established between the different domains involved in this example, AS6 can reach prefixes located in AS2 within MC1 plane thanks to several paths:

1. AS5, AS1
2. AS5, AS4
3. AS5, AS1, AS4

AS6 can choose either the first or the second path since those ones are the shortest (i.e. a smaller AS_PATH attribute). The final selection will be based on the local routing policies enforced by AS2.

***Summary**:* The characteristics that can be put forward for this scenario are:

- No modification of the BGP route selection process.
- The selected paths are guaranteed to be in the same meta-QoS-class plane.
- The route selection process doesn't necessarily select the optimal path.
- In case of problem (l-QC over-charged in an AS along the path leading to a meta-QoS-class criterion break) there is no way to detect and correct the failure.

Copyright © MESCAL Consortium, June 2005

Operational mode:

- *Each service peer adds a meta-QoS-Class identifier to its BGP announcements in order to identify the meta-QoS-class plane the announcements is intended to (this information is represented by MCi)*
- *Upon reception of these q -BGP announcements, the receiving AS (listening ASBR) proceeds to the classification of each announcement and queues it in the related RIB*
- *In this case, the per meta -QoS-class plane route selection decision -making remains identical to the classical route selection process*

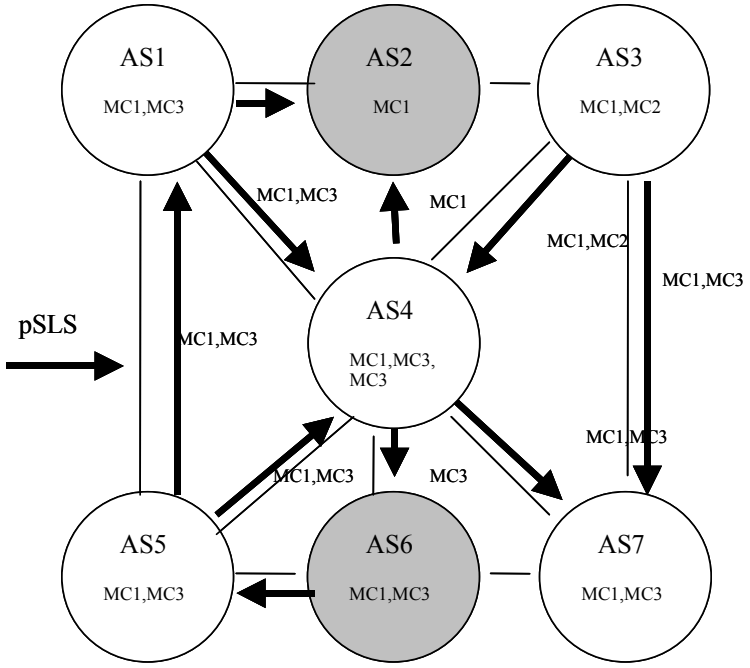


Figure 92. Only meta-QoS-class identifiers are carried in the q-BGP messages.

10.5.1.4.1.1.3 Administrative QoS-related information are propagated together with an identifier

Assumption*:* q-BGP carries meta-QoS-class identifiers AND end-to-end administrative QoS information. This QoS information is the concatenation of the characteristics of the l-QCs experienced along the AS path.

The assumption of 10.5.1.4.1.1.1 applies here, i.e. QoS information must be exploited to perform route selection. This enables administrators to define precise policies that will lead to select the best route (according to administrators' criteria). This leads to a significant modification of the route selection process, as it must now take into account the QoS information to choose the optimal route, depending on the policy enforced by the administrator per meta-QoS-class plane.

In the example illustrated by Figure 93, thanks to pSLSs established between the different domains involved, AS3 can reach prefixes located in AS7 within meta-QoS-class MC1 plane via several paths, explicitly:

- e-QC137
- e-QC134

The AS7 has to decide which path to activate. This is done by comparing the two e-QCs and choosing the best one (thanks to a well-known or proprietary QoS class comparison logic)

Copyright © MESCAL Consortium, June 2005

Operational mode:

- *e-QCijk refers to end-to-end QoS characteristics announced by ASk to ASj in the MCi meta-QoS-class plane.*
- *For each network prefix announcement, the AS must associate a meta-QoS-class identifier AND an e -QC parameter. The e -QC parameters result from the concatenation of the QoS performance characteristics of l-QC of the ASs traversed by the AS path. This is represented in the figure by {MCi, e-QCijk}*
- *Upon the reception of q -BGP announcements, each AS computes the resulting e -QC parameters in concatenating the QoS performance characteristics of its l-QC with those of the received e-QC.*
- *The route selection process mainly consists in selecting an inter -domain path that optimizes end-to-end QoS characteristics of the route for the meta-QoS-class, which is considered.*

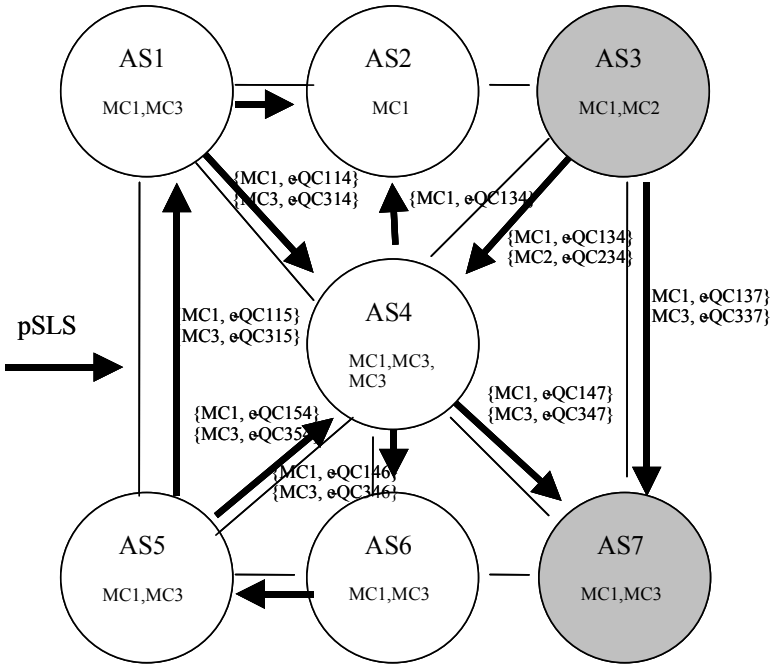


Figure 93. Use of meta-QoS-class identifier and end-to-end QoS characteristics.

The main characteristics of this approach are:

- The selected paths are guaranteed to be in the same meta-QoS-class plane;
- The modifications to the route selection process are important;
- Selection of an optimal QoS route can be achieved with more accuracy;
- In case of problem (l-QC over-charged in an AS along the path leading to a meta-QoS-class criterion break) there is no way to detect and correct the failure unless network administrators change the configuration related to a given meta-QoS-class.

We can observe that, if QoS information is administratively enforced, the route selection process will always make the same decision in normal operation conditions during the interval of validity of the configuration. The main task of the administrator is to decide when q-BGP process configuration should be modified.

10.5.1.4.1.1.4 Dynamic QoS information is available

The assumptions are similar to those of the former scenario except that QoS-related information becomes dynamic and results of an active measurement protocols/procedures. The principles regarding

Copyright © MESCAL Consortium, June 2005

the route selection process and the ability for administrators to apply specific policies mentioned in the previous section apply also in this scenario.

Operational mode:

- *e-QCijk refers to end-to-end QoS characteristics announced by ASk to ASj in the MCi meta-QoS-class plane.*
- *For each network prefix annou ncement, the AS must associate a meta -QoS-class identifier AND an e -QC parameter. The e -QC parameters result from the concatenation of the QoS performance characteristics of l-QC of the ASs traversed by the AS path. This is represented in the figure by {MCi, e-QCijk}*
- *Each AS runs an active measurement protocol that provides the real-time QoS performances parameters of its own l-QCs*
- *Upon the reception of q -BGP announcements, each AS computes the resulting e -QC parameters in concatenating the measured QoS performance characteristics of its l-QC with those of the received e-QC.*
- *The route selection process mainly consists in selecting an inter -domain path that optimizes end -to-end QoS characteristics of the route for the meta-QoS-class, which is considered.*

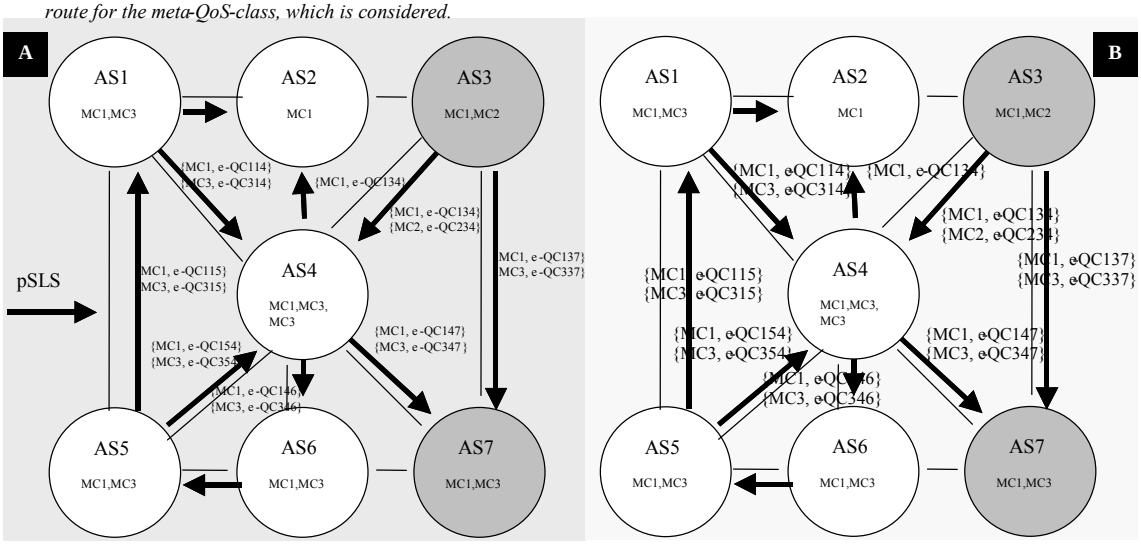


Figure 94. Use of meta-QoS-class identifier and dynamic end-to-end QoS characteristics

Let's consider the following example (Figure 94) where AS3 wants to join prefixes located in AS7 within the meta-QoS-class MC1. In case "*A*" we suppose that e-QC134 is better than e-QC137, then the route that will be chosen is the one that follows the e-QC134. In the case "*B*" we suppose that e-QC137 is better than e-QC134, and then the route that will be chosen is the one that follows the e-QC137 since it is now better than e-QC134. The changes of the e-QC parameter's values could be frequent. Therefore distinct routes could be chosen depending on the result of e-QC comparison logic.

The fact that the QoS-related information is regularly updated provides an important advantage compared to the two other scenarios. Indeed, it allows the detection of a break in the meta-QoS-class plane guarantee paradigm. Therefore, the route selection process can perform another choice that will ensure the traffic will still be forwarded in the required meta-QoS-class plane. Thus, conveying dynamic QoS information brings a real advantage, which is not present with administratively enforced QoS information. Nevertheless, the updates will be more dynamic and will impact the convergence of the BGP and the stability of the routing tables.

The characteristics of this scenario could be summarised as follows:

- The selected paths are guaranteed to be in the same meta-QoS-class plane;
- The modifications to the route selection process are important;
- In case of problem (l-QC over-charged in an AS along the path leading to a meta-QoS-class criterion break) the regular update of the QoS information enables to select another path that will remain in the same meta-QoS-class plane.
- This approach can lead to oscillation phenomena that will affect the stability of the QoS-enabled Internet.

Copyright © MESCAL Consortium, June 2005

10.5.1.4.1.1.5 Summary

Table 18 can be considered as the set of recommendations applying to the different contexts and objectives.

Attributes conveyed in BGP	*Features*	*Impact on BGP*
MC	• Ensures that route remains in the same meta-QoS-class plane. • Doesn't compute an optimal path with regards to end-to-end QoS performance characteristics	• No impact on classical route selection process • Slight modifications to BGP protocol • Modification of the information contained in the RIB (that becomes q-RIB) • Duplication of routing process and associated RIBs (one q-RIB per meta-QoS-class)
MC and administrative QoS information	• Ensures that route remains in the same meta-QoS-class plane. • Compute an optimal path with regards to end-to-end QoS performance characteristics	• Modification of the information contained in the RIB (that becomes q-RIB). • Duplication of routing process and associated RIBs (one q-RIB per meta-QoS-class). • Modifications to BGP protocol. New attributes have to be defined. • Impact on classical route selection process. Route selection process relies on QoS information
MC and dynamic QoS information	• Ensures that route remains in the same meta-QoS-class plane. • Compute an optimal path with regards to end-to-end QoS performance characteristics • Detect an overloading of a given l-QC.	• Duplication of routing process and associated RIBs (one q-RIB per meta-QoS-class). • Modifications to BGP protocol. New attributes have to be defined. • Impact on classical route selection process. Route selection process relies on QoS information

Table 18. Summary of the loose solution option recommendations and requirements

As a result of this analysis, MESCAL has decided to adopt the second scenario that aims at announcing meta-QoS-class identifiers together with administrative QoS performance characteristics.

The dynamic scheme of setting QoS values is discarded since it will lead to oscillation phenomena that will perturb the traffic and could lead to a loss of portion of traffic when changing from an old path to the new selected one.

10.5.1.4.1.2 The Statistical Guarantees Solution Option

Within the context of the SGSO, the use of q-BGP can be considered as optional. Indeed, each pSLS contains an exhaustive description of the destination prefixes attached together with their associated QoS performance characteristics and guarantees. The management system of each AS stores this information. Remote destination network prefixes are known because these prefixes are part of the of the pSLS negotiation phase and before their effective activation. Consequently, knowledge of the next hop ASs for a given destination (and then the choice of a path to a given destination) can be directly deduced from pSLS information maintained by the management plane. Considering that a routing protocol generally achieves two main elementary functions, which are: (1) routes discovery and (2) route selection it can be concluded that q-BGP (for this solution option) doesn't bring any real added-value to the inter-domain routes discovery function since those routes are already known and enforced

Copyright © MESCAL Consortium, June 2005

by the management plane of the peering ASs. Then, inter-domain *routing discovery could be completely management-based* since pSLSs include routing-related information.

Nevertheless q-BGP can be advantageously used as a means to enhance the effectiveness of this Solution Option especially by:

- Propagating inter-domain routes within the domain, thanks to q-iBGP;
- Signalling dynamically the effective availability of the routes;
- Providing a means to select dynamically alternative routes in case of failure: this requires establishment of multiple pSLSs allowing reaching the same destinations with similar o-QC;
- Making potentially easier the deployment of some load balancing features between paths that serve the same destination(s) and in which traffic will experience similar o-QC. Note that this latter point will not be developed and will be kept for further studies.

Within the scope of the SGSO, routing and forwarding are DSCP based. When q-eBGP is activated between two service peers' domains, q-eBGP updates must convey, in addition to the network prefix, a DSCP value, which indicates to the upstream AS, the DSCP value to use in order to benefit from the QoS performance guarantees attached to the o-QC a given destination belongs to.

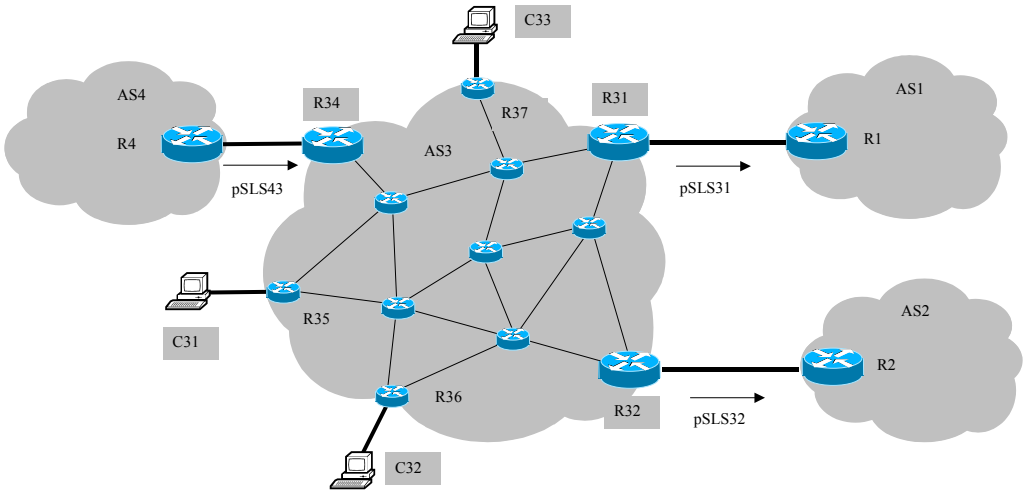


Figure 95. The statistical solution option operational mode

In Figure 95, AS3 can deploy a maximum of N l-QCs quoted l-qc-i where i can vary between 1 and N (N<64). In order to benefit from a particular QoS treatment, a customer attached to AS3 signals the requested QoS using the appropriate DSCP value. This DS code point is bound to an appropriate AS3's l-qc. Note that several DSCP values can be bound to the same l-qc in order to solve the QC splitting problem.

Let's assume now that AS3 buys the same destination prefix "D" from AS2 and AS1 with some QoS guarantees thanks to pSLS31 and pSLS32. In this example QoS guarantees are supposed to be almost the same and AS3 decides to bind o-qc31 and o-qc32 with one of its own l-qcs: l-qc1. As a consequence of these two bindings two new e-QCs are now available. Their respective QoS characteristics are too close:

- e-qc-1-31=l-qc-1 ⊕ o-qc-31
- e-qc-1-32=l-qc-1 ⊕ o-qc-32

From a commercial perspective, they could be sold as a same o-qc: o-qc1-31-32

Copyright © MESCAL Consortium, June 2005

Once pSLSs have been activated, AS1 and AS2 send q-eBGP updates indicating that destination D is available and should be signalled using a particular inter-domain DSCP value. In the example above these values are DSCP31 for AS1 and DSCP32 for AS2. The prefix D is announced to all AS3 routers. Thus, R31 and R32 send q-iBGP updates to all AS3 routers. These updates are only relevant for the DSCP routing plane corresponding to l-qc-1. Each update received from a downstream AS must consequently be interpreted by the ASBR. R31 will have to consider, for each announcement, the network prefix, the associated inter-domain DSCP value and the identity of the q-eBGP speaker. Using this information, it can retrieve from the management plane, the o-qc this announcement belongs to and gets in return the bindings details. In our case, these binding details will indicate to the two ASBR that q-iBGP update for destination D must be done within DSCP1 plane. Learned DSCP value must be accordingly swapped to its new intra-domain value.

As a consequence of this processing, R35, R36 and R37 have the same q-RIB information concerning destination D and have now to select a route toward this destination.

Destination	*DSCP plane*	*q-BGP Next Hop*
D	DSCP-1	R31
D	DSCP-1	R32

Table 19. A very simplified q-RIB example

All these paths are equivalent in term of QoS performance. For an AS3 customer, there is no difference joining D via R31 or R32. Indeed, each router can select any of these routes since, by construction, they all provide the same QoS performance guarantees. But, in the above example R36 would certainly select R32, and R37 would select R31, if the network provider would apply a "hot potato" policy.

As a consequence q-BGP doesn't need that QoS information be carried in updates for feeding the route selection process. The decision to announce a destination within a DSCP plane is enforced by the binding process. This is a 100% administrative decision. Resulting QoS characteristics of each o-QC is perfectly known, stored and maintained by the management plane. It doesn't change if the pSLS remains unchanged.

The network provider (via the management plane) should ensure that all announcements, within a given DSCP plane, concerning a given destination D, must be seen as a "similar to" o-QC even if they are really achieved with different remote o-QCs. It is the responsibility of the management plane to ensure this consistency. But this is more a binding than a q-BGP constraint.

From this perspective, the per-DSCP plane route selection process remains unchanged compared to the existing BGP selection process.

In addition, in order to illustrate the complexity of managing pSLSs that could leads definition of services to reach a given destination with similar QoS performance characteristics, let's consider the example of Figure 96:

Copyright © MESCAL Consortium, June 2005

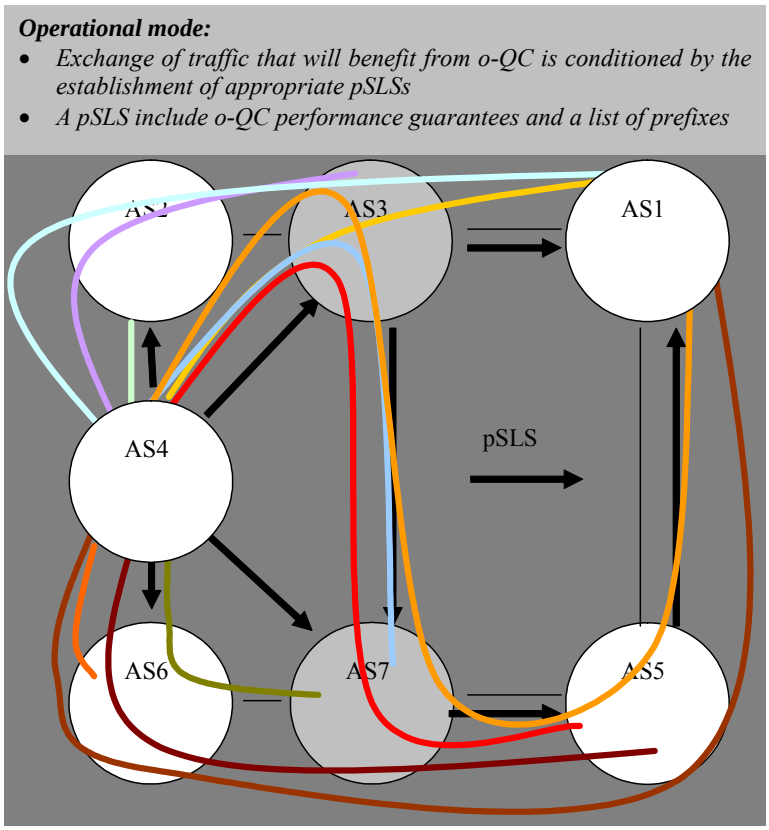


Figure 96. The statistical solution option operational mode-bis

AS4 has to manage at least 11 o-QCs that should be differentiated in order to achieve advanced task like load balancing. Some issues are to be solved like the insufficiency of DSCP range that could be used. Some ideas could be put forward as the use of an o-QC identifier that will identify a given o-QC obtained thanks to a pSLS negotiation.

From this perspective, it is obvious that the route selection process for the SGSO is less complex than the LGSO one and minor modifications are to be added to BGP. Nevertheless, offline traffic engineering functionalities *are* complex and have to provide the dynamic inter-domain routing with (in a static or automatic fashion):

- The configuration of routing policies
- The configuration of LOCAL_PREF
- The configuration of the load balancing-related policies: between pSLS, between similar o-QC …

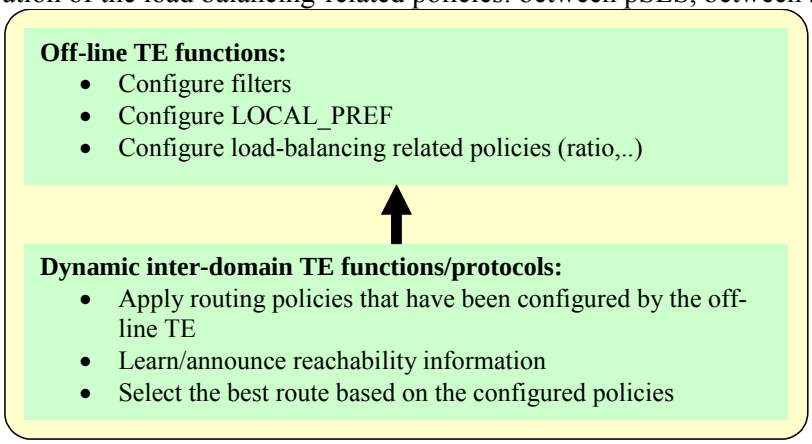


Figure 97. Interaction between off-line TE and dynamic inter-domain TE.

Copyright © MESCAL Consortium, June 2005

As a conclusion, it can be stated that in order to be able to support the SGSO, q-BGP:

- Must associate to each update the DSCP value corresponding to the agreed o-QC the announcement belongs to;
- Must run one routing decision process per DSCP plane;
- Can keep the standard routing decision algorithm at least as long as load-balancing and dynamic inter-domain bandwidth constraints are not considered.

10.5.1.4.1.3 The Hard Guarantees Solution Option

The HGSO exploits q-BGP announcements as a means to learn IP address of end-points in distant domains in order to build end-to-end LSPs. [D1.1] deliverable specifies that the HGSO uses q-BGP announcements in order to learn new destinations per meta-QoS-class. However in Section 7.4, new solutions have been proposed in order to decrease the size of the routing tables when q-BGP is used as a means to convey QoS-related information.

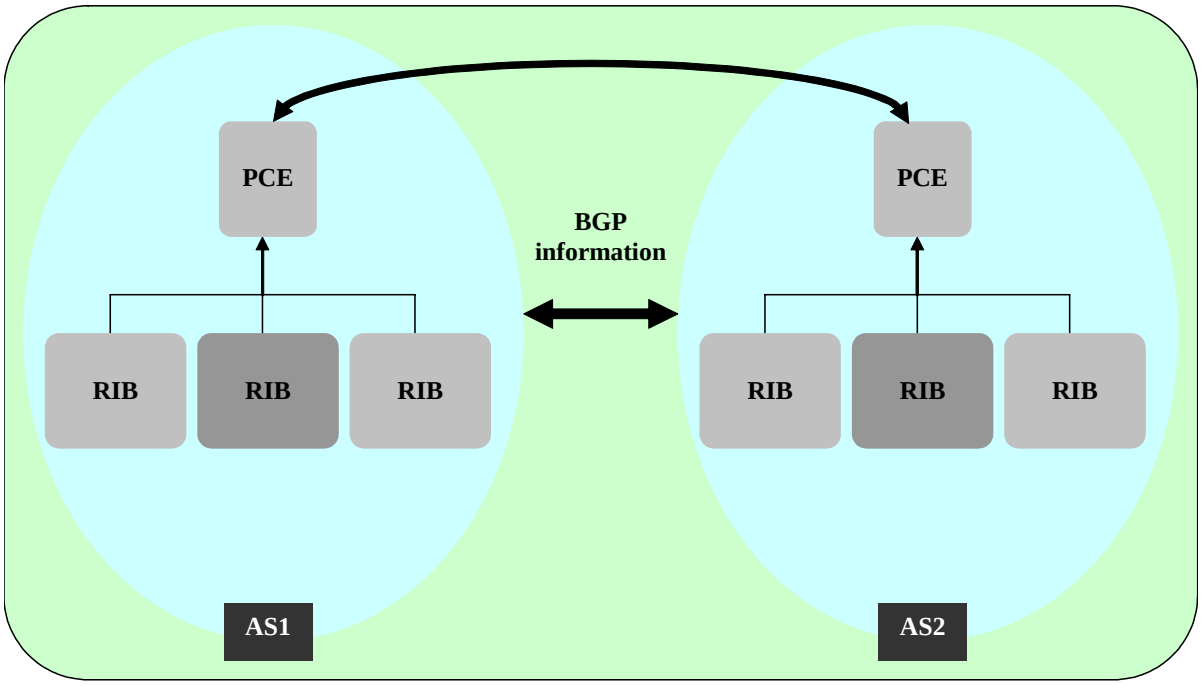


Figure 98. Inter PCE communication.

The two proposals were introduced in order to solve issues resulting of the inter-working between these two Solution Options:

- *The single signalling channel*: in this proposal, the same q-BGP announcements are used by the LGSO and HGSO. The activation of the Hard Guarantees Service Option at a given peering point is conditioned by the activation of the Loose Guarantees Service Option. Route filtering is common for the two solution options. HGSO holes encountered along an inter-domain path are signalled. This information is ignored by the LGSO but is taken into account by the PCE of each domain in order to compute an inter-domain LSP. Under these conditions, q-BGP behaviour is the same for the two Solution Options and the requirements stated in 10.5.1.4.1 are applicable in the context of the HGSO also.
- *The double signalling channels*: in this proposal, we introduce a mechanism to distinguish q-BGP announcements of each Solution Option. A Solution Option identifier inserted in q-BGP updates achieves this. At a given peering point, the Hard and the Loose Guarantees Service Option can be activated independently of each other. This discrimination doesn't induce a difference on q-BGP behaviour but only indicates to which Solution Option q-BGP announcements are intended for. HGSO specific information can differ according to two variants which are discussed below:

Copyright © MESCAL Consortium, June 2005

- *Announcement of potential LSP termination end-point addresses (routers' loopbacks or interfaces)*: in this case the announcement only differs by the value of the Solution Option identifier. HGSO specific information is ignored by the LGSO. The same rules than for the LGSO are applied for building those announcements. External HGSO routes are not taken into account when building the q-FIB. Only, intra-domain interface and loopback addresses of the routers must be present in the q-FIB. Then, the q-BGP behaviour in the HGSO is the same as for the LGSO one. From this angle, the requirements stated in 10.5.1.4.1 are applied also in this case.
- *Announcement of only the Path Computation Service identifiers (PCSID)*: in this case, we decrease the number of q-BGP announcements that are reduced to one announcement per AS. This case is similar to the previous one since these announcements are still differentiated from the LGSO ones. The requirements stated in 10.5.1.4.1 are applied also in this case. Note that in this case the related information could not be stored in the FIBs. This particularity should be taken into account in the implementation phase when considering forwarding aspects.

10.5.1.4.2 Towards a q-BGP convergent solution

10.5.1.4.2.1q-BGP behaviours

The q-BGP protocol should, as far as possible, be able to operate independently of deployed inter-domain QoS delivery solutions. q-BGP should be able to support all kind of solutions based on an exchange of QoS-related information. Within the MESCAL context, q-BGP must more specially meet the requirements of group-1 and group-2 solutions (as defined in the Introduction of this section). q-BGP should then be unique but could have distinct behaviours depending on the requirements and goals of each solutions group.

q-BGP behaviour depends deeply on the nature of the QoS-related information carried by its messages. If q-BGP messages carry only a QC identifier (this identifier could be a DS code-point or a proprietary identifier), offline traffic engineering functions are certainly complex but the q-BGP route selection process complexity is reduced. This complexity increases when a set of QoS characteristics are associated with each QC identifier. The route selection process can use either the QC-identifier for all solutions that take part of group-1 or the QC-identifier and QoS performance characteristics for solutions belonging to group-2. Figure 99 summarises these behaviours:

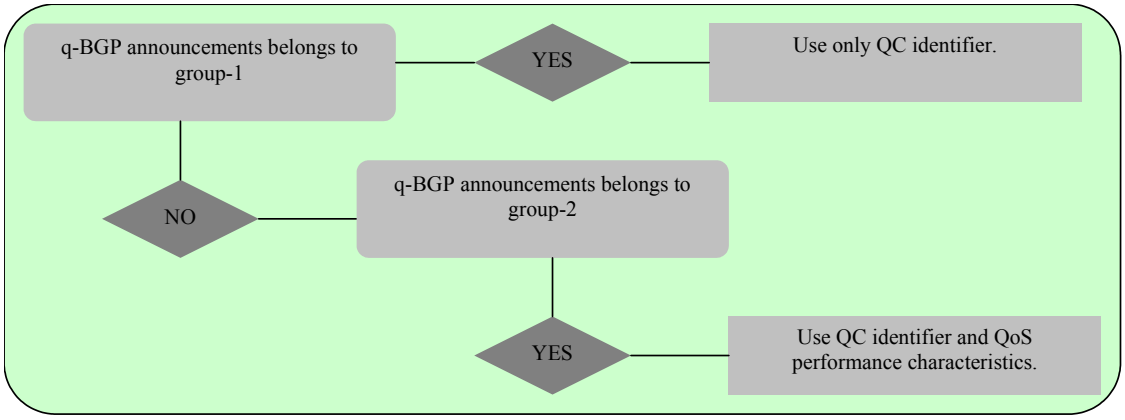


Figure 99. Route selection process required information per group.

From this standpoint, q-BGP protocol should be able to detect the group it serves. Then, it is required to introduce an additional step in the above diagram consisting at exchanging QoS service capabilities supported by each AS (q-BGP speaker). Therefore, Figure 99 becomes as shown in Figure 100:

Copyright © MESCAL Consortium, June 2005

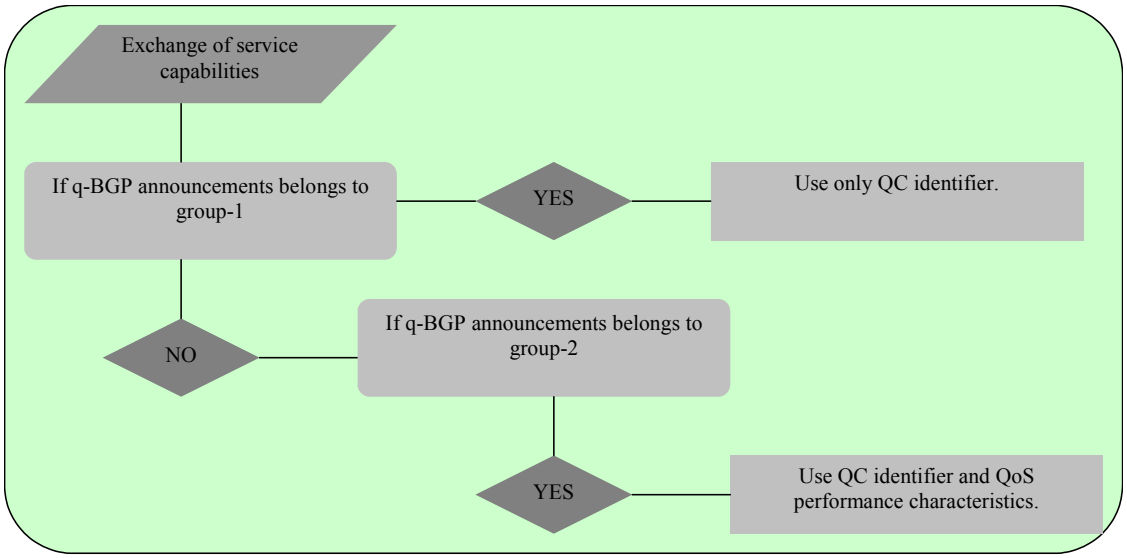


Figure 100. Towards convergent q-BGP-bis.

The purpose of the first step of the above diagram is to enable q-BGP peers to exchange the QoS service capabilities they support. Thus, q-BGP neighbours can ensure that they are able to understand each other when sending a new message, and then avoiding inopportune BGP closing of session.

10.5.1.4.2.2Applicability to MESCAL solution options

The above discussions (see sub-section 10.5.1.4.1) have revealed some common issues, related to the exchange of QoS information, between the Loose and the Hard Guarantees Solution Options that q-BGP could solve in a similar manner. In particular, q-BGP messages should carry a QoS-class identifier together with the QoS performance characteristics associated to the destination network prefix. However, in the case of the Statistical Guarantees Solution Option, q-BGP should carry only a DSCP identifier.

This could be summarised in Figure 101:

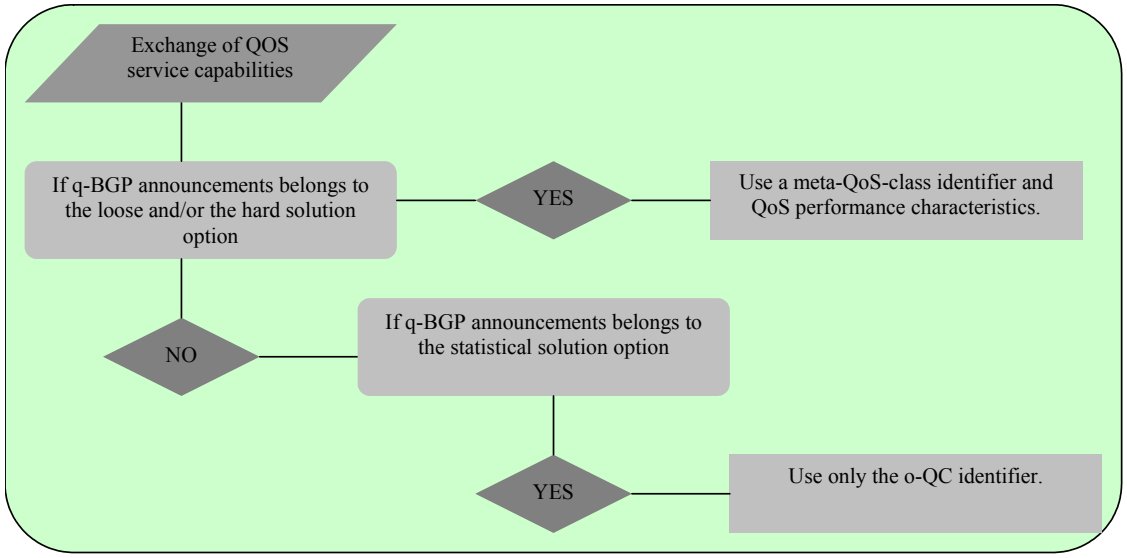


Figure 101. q-BGP in case of MESCAL solution options.

Copyright © MESCAL Consortium, June 2005

10.5.1.5 q-BGP specification

The discussions above have shown that q-BGP needs to carry some pertinent information according to the group it serves. This information is listed below:

- *QoS Service Capabilities*: this is motivated by the fact that peering entities need to ensure each other of their QoS service capabilities in order to avoid peering disruptions when a new service option is activated. Each q-BGP router has to indicate the solution group(s) it can serve and thus indicating what kind of information can potentially be carried by its messages. This will be achieved thanks to the capability optional attribute defined in [RFC3392].
- *QC identifier*: This identifier will be used to differentiate the extended QCs (i.e. between meta-QoS-class planes, o-QCs) that have been bought to service peers.
- *QoS performance characteristics*: are a set of QoS parameters like loss rate, one-way packet delay and one-way delay variation.

10.5.1.5.1 QoS Service capabilities

It is useful for a q-BGP peer to know the capabilities of a q-BGP neighbour with respect to the q-BGP protocol extensions. Capabilities exchange is achieved thanks to the specification of a new optional parameter. This parameter is included in the optional parameters of the OPEN message of a q-BGP session.

In order to indicate that a given inter-domain QoS delivery solution (in the context of MESCAL, we speak about Solution Options) belongs to a given group (either group-1 or group-2), we introduce a new parameter called *QoS Service Capabilities*.

A q-BGP speaker should use this capability advertisement in order to indicate the group to which an offered inter-domain QoS delivery solution belongs to, so that its peers can deduce if they can use the 'QoS service'-related attributes with this q-BGP peer.

The fields of this optional parameter are set as follows:

- The *capability code* field is set to a value between 128 and 255 as described in [RFC2434];
- The *capability length* is set to 2;
- The *capability value* field is encoded as shown in Figure 102:

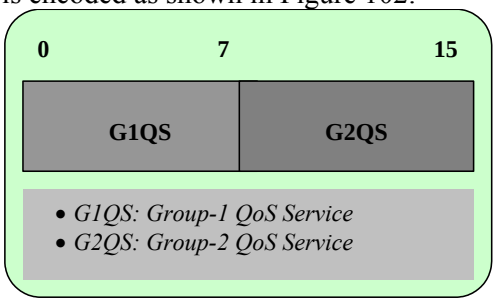


Figure 102. QoS service capability attribute.

  - The first octet is set to 0xFF if an offered inter-domain QoS delivery solution that belongs to group-1 is supported (in the context of MESCAL, if a given domain offers the statistical solution option);
  - The second octet is set to 0xFF if an offered inter-domain QoS delivery solution that belongs to group-2 is supported (in the context of MESCAL, if a given domain offers the loose and/or the hard solution options).

The way the “capability value” filed is set could be more elaborated than this first proposal. Sub groups could be defined with options.

Copyright © MESCAL Consortium, June 2005

10.5.1.5.2 QoS Class identifiers

10.5.1.5.2.1 Overview

In order to identify the inter-domain QC plane a q-BGP message belongs to, a dedicated field in q-BGP messages will be introduced. This field is called "*QoS Class identifier*". This field carries the information about the PDB, meta-QoS-class or o-QC (depending on deployed inter-domain QoS delivery solution) that is used in the downstream AS. The value of this field conforms to what has been agreed between two service peers during pSLS negotiation phase. Note that QC identifiers could be different than the DSCP code point.

The proposed field length is an octet and it is inserted in the QOS_NLRI attribute.

10.5.1.5.2.2 MESCAL Specific requirements on QC identifiers

As mentioned above, the QC identifier' purpose is to indicate to service peers either a meta-QoS-class plane or an o-QC a given q-BGP announcement belongs to. Note that a specific range of QC identifiers has to be assigned for meta-QoS-class usage so as to allow global use of the meta-QoS-class concept and then global and unified usage of this concept. If not, the value of the meta-QoS-class identifiers will be negotiated and agreed between two service peers.

Within the context of the LGSO and the HGSO, 64 possible values of meta-QoS-class identifiers are sufficient since the number of identified meta-QoS-class (at least for the MESCAL project) is less than that. Only 4 or 5 meta-QoS-classes are judged pertinent to be standardised. Definition of new meta-QoS-classes is possible since the concept is open and is basically based on the applications' requirements.

For the SGSO, and in a large-scale environment, 64 could be easily consumed. This could be a handicap for this Solution Option. Aggregation methods and pertinent service objectives are to be considered carefully within this Solution Option.

10.5.1.5.3 QoS-related information

In order to convey QoS-related information, we adopt the [CRIS05] proposal that consists at introducing a new optional attribute called QOS_NLRI attribute as the starting point. This attribute is modified in order to meet the requirements elaborated above. The format of the QoS_NLRI is different depending on the group of solutions it serves.

10.5.1.5.3.1 QOS_NLRI attribute

10.5.1.5.3.1.1 QOS_NLRI attribute for Group-1

As described above, both group-1 and group-2 solutions need to exchange a QC identifier. This identifier is used to differentiate between the extended QCs that have been bought to service peers. Especially, this is the unique additional information that must be carried by q-BGP messages. Therefore, we introduce a new QoS_NLRI attribute for group-1 solutions. The format of this QoS_NLRI attribute is as follows:

Copyright © MESCAL Consortium, June 2005

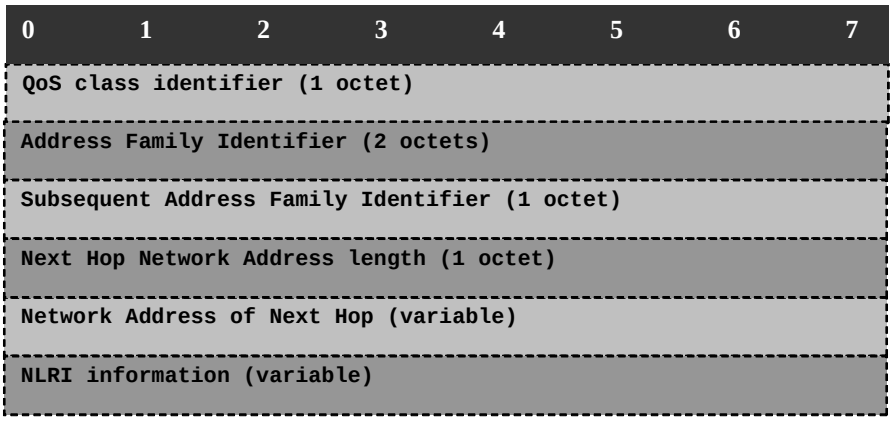


Figure 103. QoS_NLRI attribute for group-1.

The meaning of the fields of the group-1 QOS_NLRI attribute is as follows:

- QoS class identifier: this field indicates the QC identifier as described in [RFC2474];
- Address Family Identifier (AFI): this field carries the identity of the Network Layer protocol associated with the Network Address that follows;
- Subsequent Address Family Identifier (SAFI): this field provides additional information about the type of the prefix carried in the QOS_NLRI attribute;
- Next Hop Network address length: the length of next hop network address;
- Network address of Next Hop: this field contains the network address of the next router on the path to the destination prefix;
- Network Layer Reachability Information: This variable length field lists the NLRI information for the feasible routes that are being advertised by this attribute. The next hop information carried in the QOS_NLRI path attribute defines the Network Layer address of the border router that should be used as the next hop to the destinations listed in the QOS_NLRI attribute in the UPDATE message.

10.5.1.5.3.1.2 QOS_NLRI attribute for Group-2

Some modifications are added to the [CRIS05] proposal in order to meet the group-1 specific requirements listed in the sections above. The modifications are twofold:

- Information carried by this attribute:
  - The [CRIS05] proposal allows to send only one QoS performance characteristic per q-BGP announcement. This limitation has been relaxed within this specification since it might be necessary to carry a list of QoS performance characteristics in a single q-BGP UPDATE message;
  - Information about QC identifiers: unlike the [CRIS05] proposal, this specification allows to propagate information about extended QCs that are pre-negotiated between service peers. Thus PDB, meta-QoS-class and/or o-QC identifiers are announced by q-BGP thanks to QOS_NLRI attribute;
  - The [CRIS05] proposal adopts the multiple paths [WALT02]. The basic q-BGP specification focuses on single path, but we define also a QoS_NLRI for multiple path purposes. Nevertheless, we don’t describe in depth the usage of such attribute;
  - The PHB identifier has been removed from the list of possible "QoS Information Code" because of the existence of "QoS Class identifier".

Copyright © MESCAL Consortium, June 2005

- The format of the QoS_NLRI attribute:
  - Add a new field called "QoS Information length": the purpose of this field is to indicate the number of QoS performance characteristics that are enclosed in a q-BGP UPDATE message.
  - The lengths of "QoS Information code" and "QoS Information Sub-code" have been reduced to 4 bits in order to reduce the total length of the QOS_NLRI attribute. This is also motivated by the fact that 2^4 values are sufficient to indicate this information.

This attribute is encoded as shown in Figure 104.

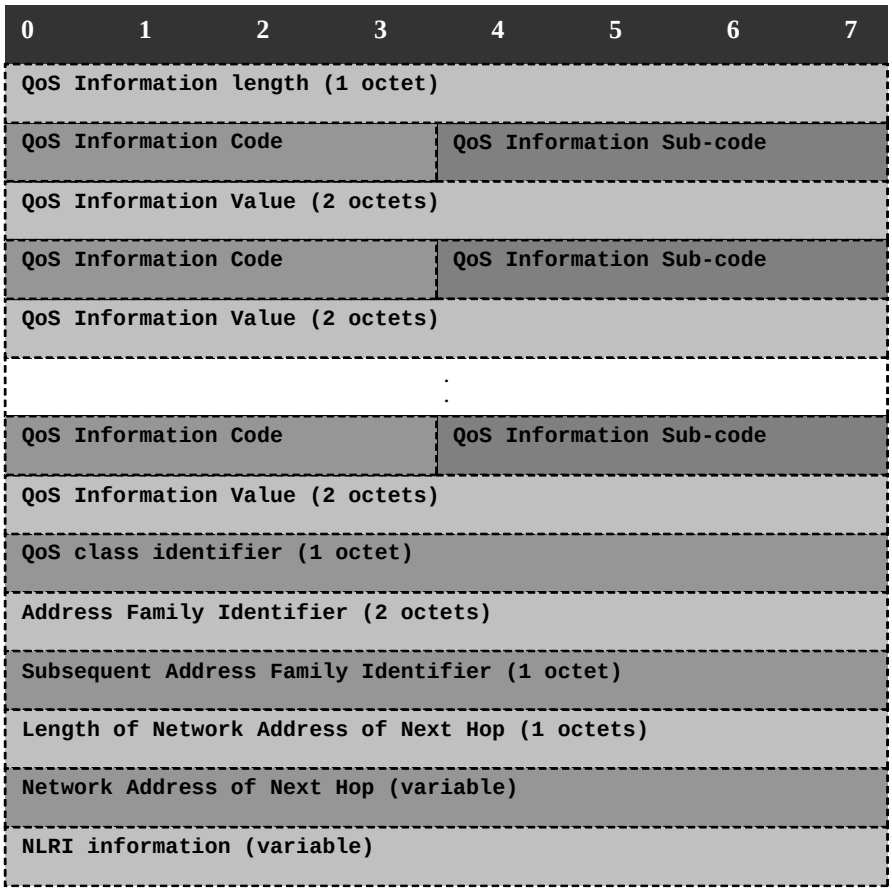


Figure 104. QoS_NLRI attribute for group-2.

The meaning of the fields of the QOS_NLRI attribute is defined below:

- QoS information length: this field carries the number of the QoS information Code that will be sent by the BGP speaker in a single q-BGP UPDATE message.
- QoS information Code: this field identifies the type of QoS information:
  - (0) Reserved
  - (1) Packet rate
  - (2) One-way delay metric
  - (3) Inter-packet delay variation

Copyright © MESCAL Consortium, June 2005

- QoS information Sub-code: this field carries the sub-type of the QoS information. The following sub-types have been identified:
  - (0) None
  - (1) Reserved rate
  - (2) Available rate
  - (3) Loss rate
  - (4) Minimum one-way delay
  - (5) Maximum one-way delay
  - (6) Average one-way delay

  Table 20 summarises the compatible (code, sub-code) pairs (a red cell refer to an invalid pair).

	0	1	2	3	4	5	6
0							
1							
2							
3							

Table 20. Compatible (code, sub-code) pairs

- QoS information value: this field indicates the value of the QoS information. The corresponding units depend on the instantiation of the QoS information code.
- QoS class identifier: this field indicates the QC identifier as described in [DS].
- Address Family Identifier (AFI): this field carries the identity of the Network Layer protocol associated with the Network Address that follows.
- Subsequent Address Family Identifier (SAFI): this field provides additional information about the type of the prefix carried in the QOS_NLRI attribute.
- Length of Next Hop' Network address: this field carries the length of next hop's network address;
- Network address of Next Hop: this field contains the network address of the next router on the path to the destination prefix.
- Network Layer Reachability Information: This variable length field lists the NLRI information for the feasible routes that are being advertised by this attribute. The next hop information carried in the QOS_NLRI path attribute defines the Network Layer address of the border router that should be used as the next hop to the destinations listed in the QOS_NLRI attribute in the UPDATE message.

10.5.1.5.3.2q-BGP provisioning

q-BGP should be configured appropriately in order to meet service needs. Especially, q-BGP process should be provisioned with the following information:

- l-QCs descriptions: at least the DSCP values used within the domain to signal l-QCs. Additional QoS performance characteristics values related to each l-QC are also needed especially in the case of deployment of inter-domain QoS delivery solutions that belongs to group-2;
- List of QC bindings: in other words, the list of q-BGP instances that should be activated and consequently the list of q-RIB/q-FIB that should be created. This could be identified by {M-QC/o-QC/e-QC DSCP value, l-QC DSCP value} pairs;

Copyright © MESCAL Consortium, June 2005

- Additional policies like the network prefixes to announce for each M-QC/o-QC/e-QC plane.

10.5.1.5.3.3 Processing QoS_NLRI attribute

As described above, q-BGP peers could exchange their respective capabilities through capability negotiation procedure. As a consequence, q-BGP peers will conclude if they both support QoS_NLRI attribute or not. If a q-BGP speaker doesn't support capability negotiation feature, it will be hard to know in advance its behaviour when receiving QoS_NLRI attribute. Therefore, three scenarios should be examined in order to describe the processing of QoS_NLRI attribute by a q-BGP speaker.

- Scenario 1: If a q-BGP speaker does not support capability feature, no QoS_NLRI should be sent to this peer.
- Scenario 2: If a q-BGP speaker does not support QoS Service Capability, no QoS_NLRI should be sent to this peer.
- Scenario 3: Both q-BGP peers support QoS Service Capability. In this case, q-BGP peers could use QoS_NLRI attribute. The variant of QoS_NLRI attribute that will be used depends on the nature of the deployed inter-domain QoS delivery solution, either it is a group-1 or group-2.
    - When sending a QoS_NLRI attribute, the local q-BGP speaker SHOULD set the QC identifier field to the identifier of extended QC on the corresponding inter-domain link. In addition, if it is a group-2 solution and if the q-BGP peer supports group-2 QoS delivery solution, the local q-BGP speaker SHOULD set the value of "QoS Information value" field(s).
    - When receiving a QoS_NLRI attribute, q-BGP speaker applies its inbound policies to grant the received announcement depending on QC binding list. The local q-BGP speaker gets the value of the "QoS Class Identifier" enclosed in the QoS_NLRI of the received announcement and checks if there is a binding entry.
        - If there is no entry in the binding list: the local q-BGP speaker drops the received announcement.
        - If there is an entry in the binding list: the local q-BGP speaker updates the values of "QoS Information value" enclosed in the QoS_NLRI with the local QC ones.

10.5.1.5.3.4 MESCAL Considerations

As already mentioned in this document, QoS performance characteristics will be only provided when either the loose or the hard solution option is deployed. In the case of the statistical solution option, QoS performance characteristic might not be propagated by q-BGP messages.

As far as the loose and the hard solution options are concerned, the QoS-related information characterises the QoS performance of the route within the meta-QoS-class specified by the value of the attached QoS Class Identifier field.

Within the context of inter-domain QoS delivery solutions that make use of the meta-QoS-class concept, a priority property will be associated to each QoS performance characteristic. For example: for loss sensitive meta-QoS-classes a value indicating a high priority could be assigned to loss rate parameter. This usage and knowledge of this priority value is part of the definition of the meta-QoS-class and is suppose to become well known from providers. Therefore, there is no need to propagate these priority properties in q-BGP messages.

10.5.1.5.3.5 Multiple paths

[WALT02] proposes a mechanism that allows the advertisement of multiple paths for the same prefix without the new paths implicitly replacing any previous ones. This is achieved thanks to the use of an arbitrary identifier that will identify (in addition to the prefix) a given path. The QoS_NLRI attributes become as follows:

Copyright © MESCAL Consortium, June 2005

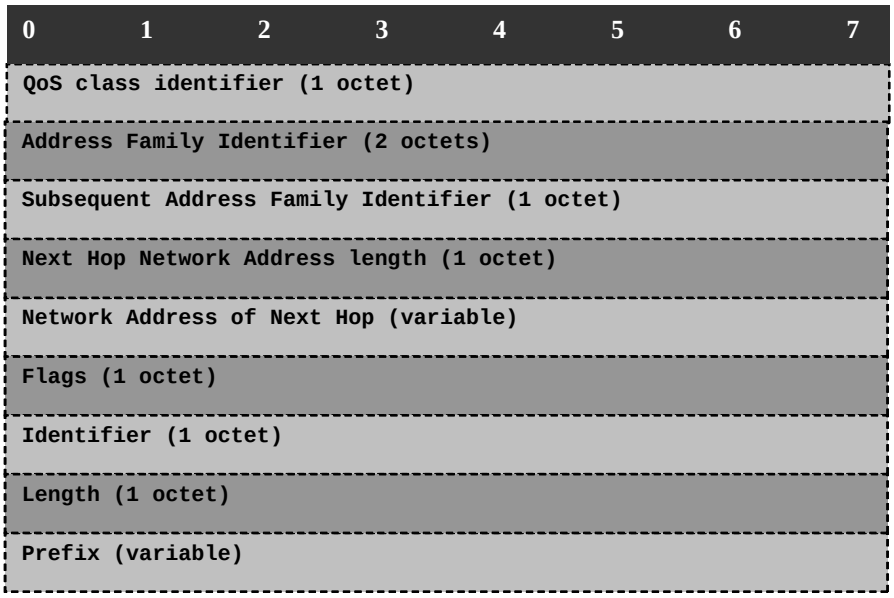


Figure 105. QoS_NLRI attribute for group-1 (multiple paths).

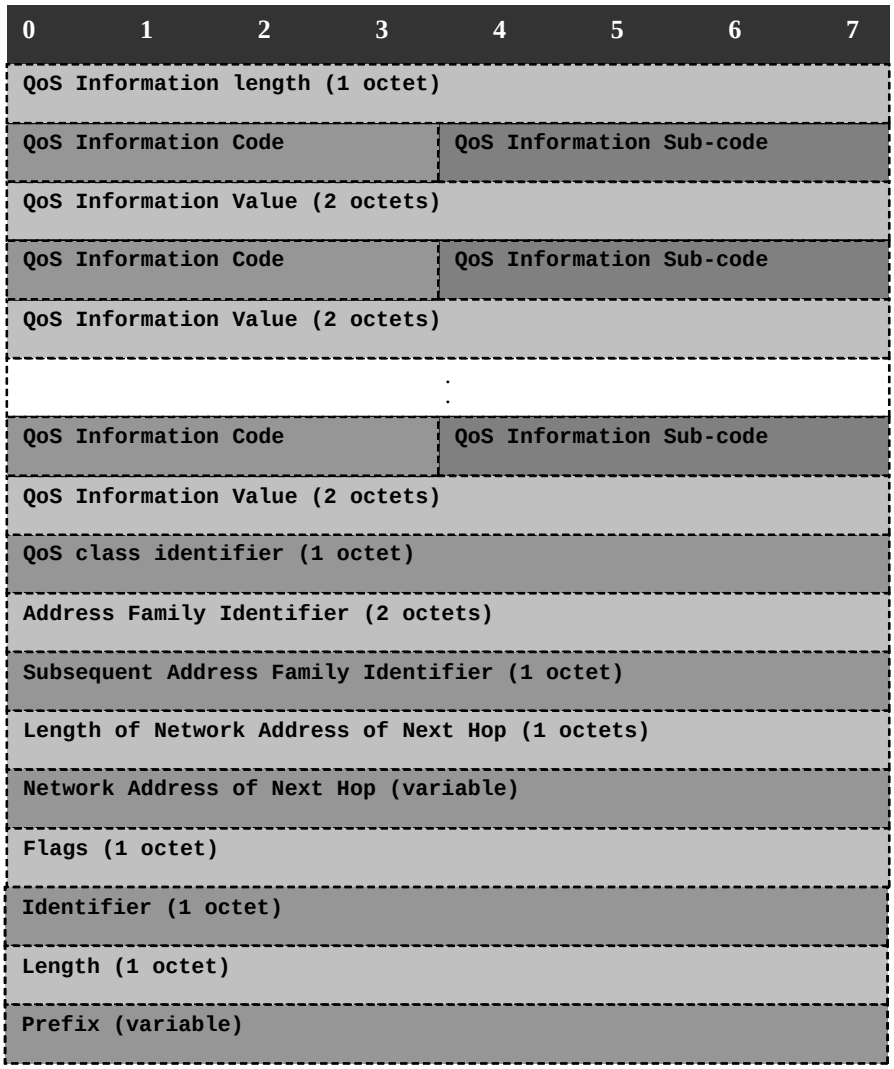


Figure 106. QoS_NLRI attribute for group-2 (multiple paths).

Copyright © MESCAL Consortium, June 2005

The meanings of the QOS_NLRI attribute fields are the same as the ones described in 10.5.1.5.3.1.1 and 10.5.1.5.3.1.2, except the following fields:

- Flags, Identifier, Length and Prefix: the meaning of these fields is described in [WALT02].

10.5.1.5.4 MESCAL specific change: Service options differentiation

As a result of the study about the inter-working of the MESCAL solution options, it was decided that a clear differentiation had to be introduced in order to separate the q-BGP announcements in order to solve signalling problems (see [D1.4]). This could be implemented by dedicating specific community values for each solution option.

The BGP community attribute was added to BGP in order to simplify the configuration of complex routing policies. It is an optional and non-transitive attribute. The community attribute is a list of community values that are between 0x00000000 and 0xFFFFFFFF. The ranges 0x00000000 through 0x0000FFFF and 0xFFFF0000 through 0xFFFFFFFF are reserved. The well-known values are:

- *0xFFFFFF01*: NO-EXPORT. If a BGP speaker receives a route with this community value, the BGP speaker must not export that route beyond its local AS.
- *0xFFFFFF02*: NO-ADVERTISE. If a BGP speaker receives a route with this community value, it must not re-advertise this route beyond the local router.
- *0xFFFFFF03*: NO-EXPORT-SUBCONFED. This is similar to the NO-EXPORT except that it is used in the context of the confederations.

[RFC1997] states: "***The rest of the community attribute values shall be encoded using an autonomous system number in the first two octets. The semantics of the final two octets may be defined by the autonomous system (e.g. AS690 may define research, educational and commercial community values that may be used for policy routing as defined by the operators of that AS using community attribute values 0x02B20000 through 0x02B2FFFF).***"

In order to be aligned with this recommendation and also to allow the distinction between the supported service options, only the second octet of the community value will be used to indicate the service option. The first octet will encode the AS number (NAS). Thus, we adopt the following structure:

- *NAS:01*- this means that the AS identified by the number NAS supports the loose solution option.
- *NAS:02*- this means that the AS identified by the number NAS supports the statistical solution option.
- *NAS:03*- this means that the AS identified by the number NAS supports the hard solution option.

10.5.1.5.5 Route selection process

10.5.1.5.5.1 Classical Route selection process

The BGP specification [RFC1771] has defined a decision process for the selection of the routes that will be installed in the local RIB. This process is responsible for the:

- Selection of routes to advertise to BGP listeners located in the local speaker's autonomous system
- Selection of routes to advertise to BGP listeners located in neighbouring autonomous systems
- Route aggregation and route information reduction.

This process takes into account the BGP attributes, which can impact the selection of the routes. [RFC1771] specifies a set of attributes that could be used as tie-breaker in the context of the route selection process. The following attributes are the most used:

Copyright © MESCAL Consortium, June 2005

- LOCAL_PREF: this attribute is used at the beginning of the route selection process. It is a well-known discretionary attribute that is used by a BGP speaker to inform BGP peers from its own autonomous system of the originating speaker's degree of preference for an advertised route.
- MED: this attribute is an indicator of which local entrance point an AS would like a peering AS to use. This attribute isn't suitable to break the tie between two equal paths learned from distinct ASs.
- IGP Metric: this metric could be used to influence the choice of the path to put in the local RIB.

Hereafter the BGP Path selection process as commonly understood and implemented:

- Prefer largest Local Preference.
- If same Local Preference prefer the route that the specified router has originated.
- If no route was originated prefer the shorter AS path.
- If all paths are external prefer the lowest origin code (IGP<EGP<INCOMPLETE).
- If origin codes are the same prefer the path with the lowest MED.
- If path is the same length prefer the External path over Internal.
- If path is learn via the same type of BGP advertisement (external or internal), prefer the path with lowest IGP cost
- Prefer the route with the lowest IP address value for BGP router ID.

This process may vary from a vendor to another. For instance, the Cisco implementation adds a new metric called "weight" that is used to choose the best path.

10.5.1.5.5.2 Modified Route selection process

As far as QoS-related information is conveyed in BGP UPDATE messages, the route selection process should take into account this information in order to make a choice and make a tie-break between equal paths and determine the one(s) to be stored in the local FIB. This process could differ between solutions that belong to group-1 or group-2.

10.5.1.5.5.2.1 Group-1 Route Selection Process

For all inter-domain QoS-delivery solution that belongs to group-1 (example: the statistical solution option) only the identifier of the extended QC is to be taken into account in order to choose a path that will be stored in the local RIB. The unique modification to be added to the classical route selection process is to identify routes that serve the same destination with similar e-QCs. Local policies could be configured by each INP in its ASBRs. These policies depend on the usage of pSLS as decided by the INP. INPs could define their own policies.

Thus, the pseudo code of the modified route selection process will be as follows:

```
1. Identify the received routes that serve the same destination
2. Consider the routes with similar o-QCs
3. Apply local policies (prefer a given origin AS, cost,…).
4. If only one route has been returned
     Store this route in the RIB
5. If more than one route has been returned
     Apply the classical BGP route selection process.
```

Copyright © MESCAL Consortium, June 2005

10.5.1.5.5.2.2 Group-2 Route Selection Process

10.5.1.5.5.2.2.1 Generic Route Selection Process

For all inter-domain QoS-delivery solutions that belong to group-2 (examples: LGSO, HGSO), q-BGP UPDATE messages carry QoS performance characteristics together with a QC identifier. q-BGP route selection process should exploit enclosed QoS performance characteristics in order to determine the path that will be stored in the local RIB. Modifications that should be added to the classical route selection process are at least:

- Identify routes that serve the same destination in the same QC plane;
- Select a route that optimises QoS performance characteristics.

Therefore, the new route selection process becomes:

1. Identify routes that serve the same destination
2. Consider routes that have the same QoS class identifier
3. Compare the QoS performance characteristics associated with resulting routes with respect to a well know comparison logic
4. Return the route that optimises the QoS performance characteristic
5. if more than one route has been returned, apply the classical BGP route selection process

10.5.1.5.5.2.2.2 Variants of QoS Comparison Logic

10.5.1.5.5.2.2.2.1 Priority-based route selection process

In the case of the Loose Guarantees Solution Option, the comparison logic could be based on the use of the priority value that has been affected to each QoS performance characteristic. The priority ordering of the QoS performance characteristics is well known and commonly understood by all INPs because it is a part of the definition of the meta-QoS-class. The philosophy of this method is: "*find the route that optimises the highest priority QoS-related information*".

Therefore, the pseudo code of the route selection process algorithm becomes as follows:

1. Identify routes that serve the same destination
2. Consider routes that have the same QoS class identifier
3. Consider the QoS performance characteristic that has the highest priority, and return the routes that optimise that QoS performance characteristic
    i. If only one route is returned store this route in the local RIB
    ii. If more than one route are returned
        1. Exclude the QoS performance characteristic that has been used in the step 3 from the list of QoS performance characteristics.
            a. If there is no remaining QoS performance characteristics, go to step 4
            b. Else, go to step 3
4. if more than one route has been returned, apply the classical BGP route selection process

Copyright © MESCAL Consortium, June 2005

- ***Example:***

Let suppose that a q-BGP router has received the following routes for reaching the same destination. Each of these routes is associated with a set of QoS performance values as follows:

- R1: minimum one way delay=150ms, Loss rate=5%
- R2: minimum one way delay=120ms, Loss rate=2%
- R3: minimum one way delay=100ms, Loss rate=3%
- R4: minimum one way delay=200ms, Loss rate=8%

If the q-BGP router is configured to prioritise minimum one way delay, the selected route is R3. But if the q-BGP router is configured to prioritise loss rate, the selected route is R2.

- ***Route selection consistency:***

Let suppose now that a q-BGP router has received from its peers, the following routes for reaching the same destination P1. The received routes enclose the following QoS performance values as detailed below:

- Route R1: QoS1=90ms, QoS3=150ms, QoS4=5%
- Route R2: QoS2=30ms, QoS3=153ms, QoS4=1%
- Route R3: QoS1= 120ms, QoS2=100ms, QoS3= 60ms, QoS4=3%
- Route R4: QoS2=90ms, QoS3= 50ms, QoS4=8%

The aforementioned routes enclosed different QoS performance characteristics. The issue is how to compare these routes and how to ensure that the selected route is consistent with the service needs. This problem could be caused by a mis-configuration or because an INP doesn't support a given QoS parameter characteristic. This risk in both cases is that INPs will not advertise QoS-related information since if only one AS in the chain doesn't implement a QoS parameter; it will introduce a QoS hole in all routes that it will advertise and then impact the decision of leaf ASs.

In order to solve this issue, at least two solutions could be considered:

- ***Solution 1***: Add a new control level in the definition of m-QC. This consists in defining mandatory and optional attributes. A "Mandatory QoS information" is a parameter that must be present in the QoS_NLRI. In the case it is missing the announcement will be dropped by q-BGP receiver. The announcement isn't dropped if an "Optional QoS Information" is missing. Nevertheless, the problem of ensuring route selection consistency when optional parameters are missing is unsolved. For solving this case, one of options details under Solution 2 bullet could be implemented.
  It is obvious that all INP should have to same understanding of the mandatory and the optional parameters in order to preserve a consistent treatment in all traversed ASBRs. Within LGSO this is guaranteed thanks to the use of meta-QoS-class concept.
- ***Solution 2***: No additional control level is introduced in the meta-QoS-class definition (all parameters are optional). In this second solution, the risk is that service providers won't advertise routes with required QoS information that should be used as guidance to meet service needs. As a consequence, group-2 solutions become as group-1 ones because there is no control regarding the enclosed QoS information. Nevertheless, when a QoS parameter is missing, three options could be considered.
  - ***Option 1***: Discard unvalued routes but keep them all if they are all unvalued. In this case, the priority criterion is respected and the comparison between routes is consistent. But, the risk is that some destinations could be unreachable if received routes don't enclose higher priority QoS performance characteristics.

Copyright © MESCAL Consortium, June 2005

- ***Option 2***: Replace unvalued QoS information with the upper boundaries of the missing parameter as defined by m-QC. The inconvenient of this option is how to set the missing values. This inconvenient of this solution is that this could be seen as an encouragement to not send valued QoS performance characteristics. The convergent situation could be that the announced QoS values are not better than upper boundaries of QoS performance characteristics as defined per m-QC.
- ***Option 3***: Always keep routes with unvalued parameter, but perform selection for the remaining routes. The route selection process compares between routes that have valued the QoS parameter used as criterion in a given step. If there still have equal routes, all routes are considered and the route selection process checks for the route that encloses the next QoS parameter to be used as criterion of its selection even if these routes were not present in the previous step. The inconvenient of this option is that the priority criterion isn't satisfied.

Figure 107 summarises the three options detailed above.

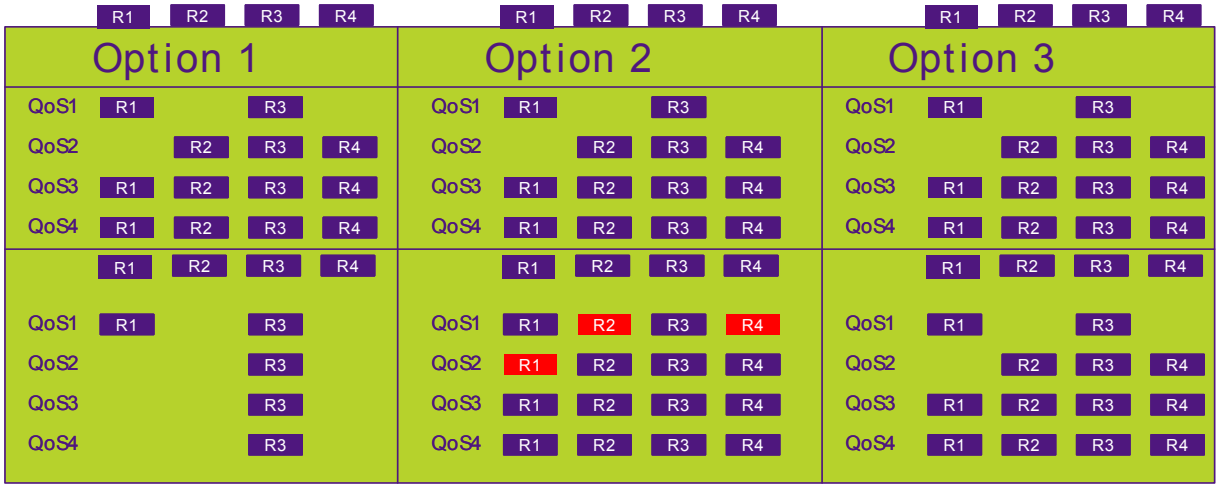


Figure 107. Example of route decision-making.

MESCAL project adopted the solution 1 together with option 1. As a consequence, the updated pseudo code of the route selection process is as follows:

```
1. Identify routes that serve the same destination
2. Group routes having the same QoS class identifier
3. For each route group,
        i. If the number of remaining routes is not null, check
           for each route, if all mandatory QoS performance
           characteristics are valued
            i. If yes, add this route to remaining routes and go to
               step 4
            ii. If no, drop this route
4. Consider the remaining routes
        i.  If the number of remaining routes is not null,
            1. Consider the QoS performance characteristic that
               has the highest priority, and return the routes
               that optimise that QoS performance characteristic
```

Copyright © MESCAL Consortium, June 2005

a. If only one route is returned store this route in the local RIB

b. If more than one route is returned
   - Exclude the QoS performance characteristic that has been used in the step 4.i.1 from the list of QoS performance characteristics.
     1. If there is no remaining QoS performance characteristics, go to step 5
     2. Else, go to step 4.i.1

ii. If the number of remaining routes is null, go to step 5

5. If more than one route has been returned, apply the classical BGP route selection process

---

10.5.1.5.5.2.2.2.2 Random based route selection process

This variant of route selection process assumes that all INPs have classified correctly their l-QCs with regard of m-QCs. QoS-related information are only seen as a means to share load between several routes and not a way to enhance optimal path selection process. In this solution, no priority is assigned to QoS parameters. The pseudo code of this route selection process is as follows:

---

1. Identify routes that serve the same destination
2. Consider routes that have the same QoS class identifier
3. Choose randomly one QoS performance characteristic

i. If there is no route that encloses the selected QoS performance characteristic, exclude this QoS performance characteristic and go to step 3

ii. Else, return the routes that optimise that QoS performance characteristic

1. If only one route is returned store this route in the local RIB
2. If more than one route are returned

a. Exclude the QoS performance characteristic that has been used in the step 3 from the list of QoS performance characteristics.

b. If there is no remaining QoS performance characteristics, go to step 4

c. Else, go to step 3

4. if more than one route has been returned, apply the classical BGP route selection process

---

The advantage of this route selection process is that the route selection process won't return the same route. This could be considered as a mechanism to prevent from saturating the same route.

Copyright © MESCAL Consortium, June 2005

10.5.1.5.5.2.2.2.3 Weight-based route selection process

In this Section, we discuss additional approaches for comparing sets of QoS values. Consider two QoS tuples *X* and *Y*. These tuples consist of both the attributes (e.g. delay, jitter, loss rate) and their values (which may in general be either quantitatively or qualitatively expressed). The tuples may for example be QoS advertisements exchanged by the QoS Capabilities Announcement functional block; or they may be QoS values negotiated through the pSLS Ordering functional block; or they may be QoS announcements exchanged through q-BGP. Let the tuples consist of QoS attributes *A, B* and *C*, and let the QoS tuple *X* have the values (A_x, B_x, C_x) and let QoS tuple *Y* have the values (A_y, B_y, C_y).

Then to compare the two QoS tuples X and Y, a number of mechanisms can be adopted. To generalise the discussion, here we assume that "*P*>*Q*" means that *P* is **better** than *Q*, irrespective of whether we are comparing bandwidth values (where a higher numerical value represents a better level of QoS) or delay values (where a lower numerical value represents a better level of QoS). The mechanisms are as follows:

- **Lexicographical ordering**: the QoS attributes are compared in strict order. Thus if $A_x > A_y$ then *X* is better than *Y*, irrespective of the relative values of B_x, B_y, C_x or C_y. If $A_x = A_y$ then the second QoS attributes are compared: if $B_x > B_y$ then *X* is said to be better than *Y*. This approach is currently implemented in q-BGP. This approach always produces a result by computing one tuple to be better than the other (or the two tuples to be equal), but tends to be based only on the higher priority attributes while ignoring the values of the lower priority attributes (See section 10.5.1.5.5.2.2.2.1)
- **Simultaneous comparison**: *X* is better than *Y* if $A_x > A_y$ **and** $B_x > B_y$ **and** $C_x > C_y$. Similarly, *Y* is better than *X* if $A_y > A_x$ **and** $B_y > B_x$ **and** $C_y > C_x$. This approach was defined in [D1.1] Section 4.2.2.2. This approach does not define a result if some of the QoS attributes *A, B, C* of one tuple are better than the second tuple but some of the QoS attributes are worse. For example, if $A_x > A_y$ while $B_y > B_x$ then the result of the comparison of *X* with *Y* is undefined.
- **Weighted ordering**: the QoS attributes are normalised to create dimensionless values, and summed. This results in a single value for each QoS tuple, which can be compared to determine which tuple is better. The dimensionless values could additionally be weighted so as to prefer one attribute over others. The advantage of this approach is that it potentially allows a wider range of QoS metrics to be fairly compared, for example "a low delay route with reasonable bandwidth". The metric is computed as follows:
    - Each QoS parameter is normalised to the range {0..1} where 0 represents worst and 1 represents best. If say R_x and R_y are two values of a parameter such as delay (where a lower numerical value represents a better level of QoS) then their normalised values are $\left(\frac{\min(R_x, R_y)}{R_x}\right)$ and $\left(\frac{\min(R_x, R_y)}{R_y}\right)$ respectively. On the other hand, if S_x and S_y are two values of a parameter such as bandwidth (where a higher numerical value represents a better level of QoS) then their normalised values are $\left(\frac{S_x}{\max(S_x, S_y)}\right)$ and $\left(\frac{S_y}{\max(S_x, S_y)}\right)$ respectively.
    - Optionally, each of the *n* normalised QoS parameters in each of the QoS tuples is multiplied by a weight w_i where $\sum_{i=1}^{n} w_i = 1$ and the same weights are used in each of the QoS tuples.
    - The weighted normalised QoS parameters are summed.

Copyright © MESCAL Consortium, June 2005

- The best QoS tuple is the one whose sum of normalised QoS parameters is the highest.

It should be noted that for the lexicographical ordering and weighted ordering mechanisms, the order of the QoS attributes is guided by the service objectives, and should ideally be applied consistently between different domains: this point is illustrated in the next Section. In addition, for weighted ordering, the weights of the QoS attributes should ideally be agreed amongst all domains, and an approach such as the meta-QoS class provides a mechanism for this.

- Example: application to QoS routing:

We now illustrate the use of the three QoS tuple comparison mechanisms by comparing the behaviour of the q-BGP route selection process under these three mechanisms. Figure 108 shows a set of six domains that connect two end hosts. Each domain supports a pair of QoS characteristics, namely delay and bandwidth. For each domain, the delay and bandwidth are as shown. We assume for simplicity that the inter-domain links have zero delay and infinite bandwidth.

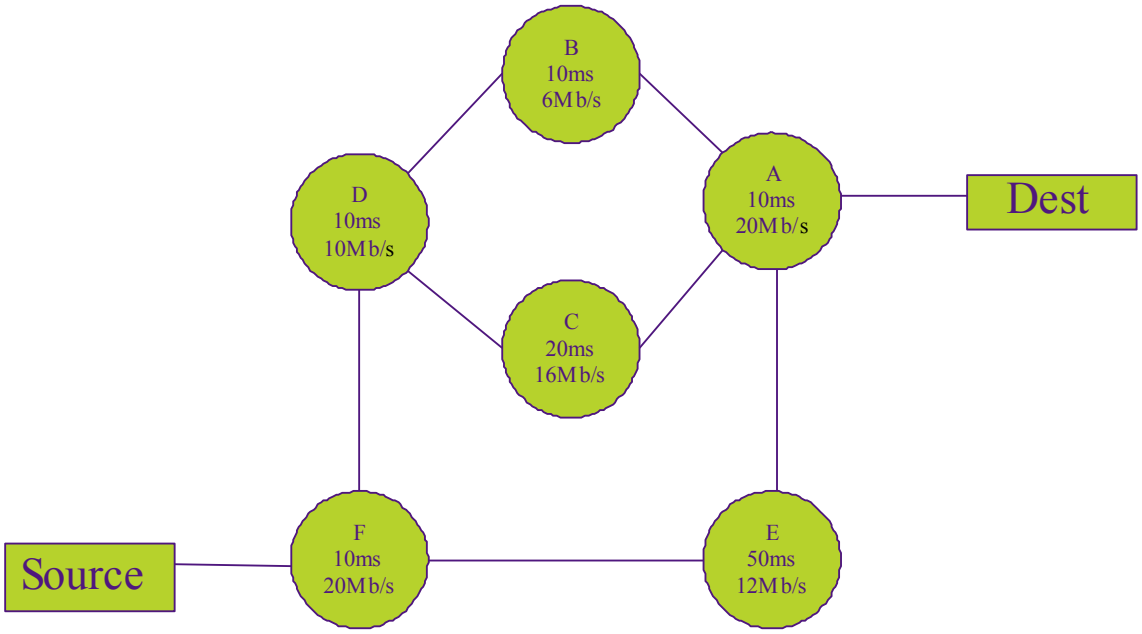


Figure 108. q-BGP route selection scenario.

Paths and their QoS characteristics can be computed using the various comparison mechanisms defined in Section 10.5.1.5.5.2.2 as follows:

- Lexicographical ordering of (delay, bandwidth) values, all ASs implementing same policy: in this case, delay is the most important attribute when selecting a path and bandwidth is only considered in the event that two path options have the same delay. This option results in the highest QoS path being the top route, F – D – B – A, with path QoS values delay = 40ms, bandwidth = 6Mbit/s.
- Lexicographical ordering of (bandwidth, delay) values, all ASs implementing same policy: in this case, bandwidth is the most important attribute when selecting a path and delay is only considered in the event that two path options have the same bandwidth. This option results in the highest QoS path being the bottom route, F – E – A, with path QoS values delay = 70ms, bandwidth = 12Mbit/s.
- Lexicographical ordering of values, with each AS implementing its own routing selection policy. In this case, the "highest QoS path" depends on which policy each AS implements. This approach provides no QoS optimisations because each domain is effectively using QoS

Copyright © MESCAL Consortium, June 2005

parameters as local preference metrics. The important domains whose policy affects the final result are the two where a path choice is made: ASs F and D. This gives us four sub-results:

- Both F and D implement delay as highest priority: selected path is F – D – B – A (delay = 40ms, bandwidth = 6Mbit/s);
- F implements delay, D implements bandwidth as highest priority: selected path is middle route F – D – C – A (delay = 50ms, bandwidth = 10Mbit/s);
- F implements bandwidth, D implements delay as highest priority: selected path is F – E – A (delay = 70ms, bandwidth = 12Mbit/s);
- Both F and D implement bandwidth as highest priority: selected path is F – E – A (delay = 70ms, bandwidth = 12Mbit/s).

- Simultaneous comparison: here, decisions are made at domains D and F. At D, the choice is between path B-A (delay=20ms, bandwidth=6Mbit/s) and path C-A (delay=30ms, bandwidth=16Mbit/s). Since B-A has the better delay but C-A has the better bandwidth, the simultaneous comparison is unable to select the better path, and it is therefore necessary to employ some tiebreak mechanism such as a lexicographical ordering. If for example D tie-breaks using the ordering (delay, bandwidth) then path B-A has the better QoS and is selected. At F, the choice is then between D-B-A (delay=30ms, bandwidth=6Mbit/s) and E-A (delay 60ms, bandwidth=12Mbit/s). Again, the tiebreak mechanism is required; if F also uses the ordering (delay, bandwidth) then path F – D – B – A is selected, with path QoS values delay = 40ms, bandwidth = 6Mbit/s.
- Weighted ordering: initially assume that both delay and bandwidth are given equal weighting, and the requirement is thus for "reasonable delay with reasonable bandwidth". Then at D, where the top path B-A is (20ms, 6Mbit/s) and the middle path C-A is (30ms, 16Mbit/s). The normalised values are respectively (1.0, 0.375) and (0.67, 1.0). The sums are respectively 1.375 and 1.67, so the latter path C-A is selected. At F, the path D-C-A is (40ms, 10Mbit/s) and the path E-A is (60ms, 12Mbit/s). The normalised values are respectively (1.0, 0.83) and (0.67, 1.0). The sums are respectively 1.83 and 1.67, so path D-C-A is selected. Thus the best path for "reasonable delay with reasonable bandwidth" is F – D – C – A.

  Weighting can also be introduced: let the weights be 0.6 for delay and 0.4 for bandwidth; we can call this "good delay with reasonable bandwidth". Then at D, the weighted normalised values are for B-A (0.6, 0.15) and for C-A (0.402, 0.4); these respectively give sums of 0.75 and 0.802, so path C-A is selected. At F, the normalised values are for D-C-A (1.0, 0.83) and for E-A (0.67, 1.0). The weighted normalised values are therefore respectively (0.6, 0.33) and (0.402, 0.4). The sums are respectively 0.93 and 0.80, so the best path for "good delay with reasonable bandwidth" is F – D – C – A.

10.5.1.5.5.2.2.2.4 Discussion

Lexicographical ordering provides an approach that focuses on one QoS metric (the highest priority) at the expense of lower priority parameters. For example, considering delay as the highest priority parameter might produce a path with a very low bandwidth; and conversely considering bandwidth as the highest priority parameter might produce a path with high delay. To some extent this can be ameliorated by introducing a "precision" so that two QoS values are said to be identical if they are within some defined percentage of each other; this then allows lower priority QoS metrics to be considered in path comparison.

Simultaneous comparison provides an approach that can be combined with another approach such as lexicographical ordering to compare path QoS metrics.

Weighted ordering provides an approach that allows a mix of QoS values to be taken into account in path selection. By varying the weights, paths can in principle be selected that provide a mix of QoS metrics, rather than paths that are focused primarily on a single primary QoS metric. These paths may then be more closely aligned with the QoS requirements of some applications.

Copyright © MESCAL Consortium, June 2005

In all cases, it is important that policies be applied consistently across all domains. Failure to do this result in the scenario illustrated above where each domain implements its own routing selection policy; the QoS of the resulting path is then not predictable.

10.5.1.5.5.2.2.3 HGSO-specific behaviour

q-BGP route selection process is the same for the HGSO and LGSO Solution Options. In the case this route selection process is priority-based, some requests could not be issued by a PCE because there is no route installed in the q-RIB that satisfy the requested QoS constraints.

In order to illustrate this behaviour, let consider the following scenario described in Figure 109. On each link are indicated the average (A-OWD) and the maximum (M-OWD) one-way delay included in the q-BGP announcement message of the prefix P1 in the MQC1 plan. These QoS characteristics are updated by each AS depending on the characteristics of the l-QCs corresponding to the MQC 1. The average one-way delay and the maximum one-way delay of the l-QC used in the AS1 are respectively set to 5 and 10. The others l-QC characteristics are not useful in this example.

The AS1 owns two routes for the destination P1:

- a route via AS2 with the following QoS characteristics: A-OWD=60, M-OWD=100
- a route via AS3 with the following QoS characteristics: A-OWD=40, M-OWD=200

The priority of the average one-way delay is the highest and the route via AS3 is selected by the AS1. This route is then propagated to the AS0 with the updated QoS characteristics (A-OWD=45, M-OWD=210)

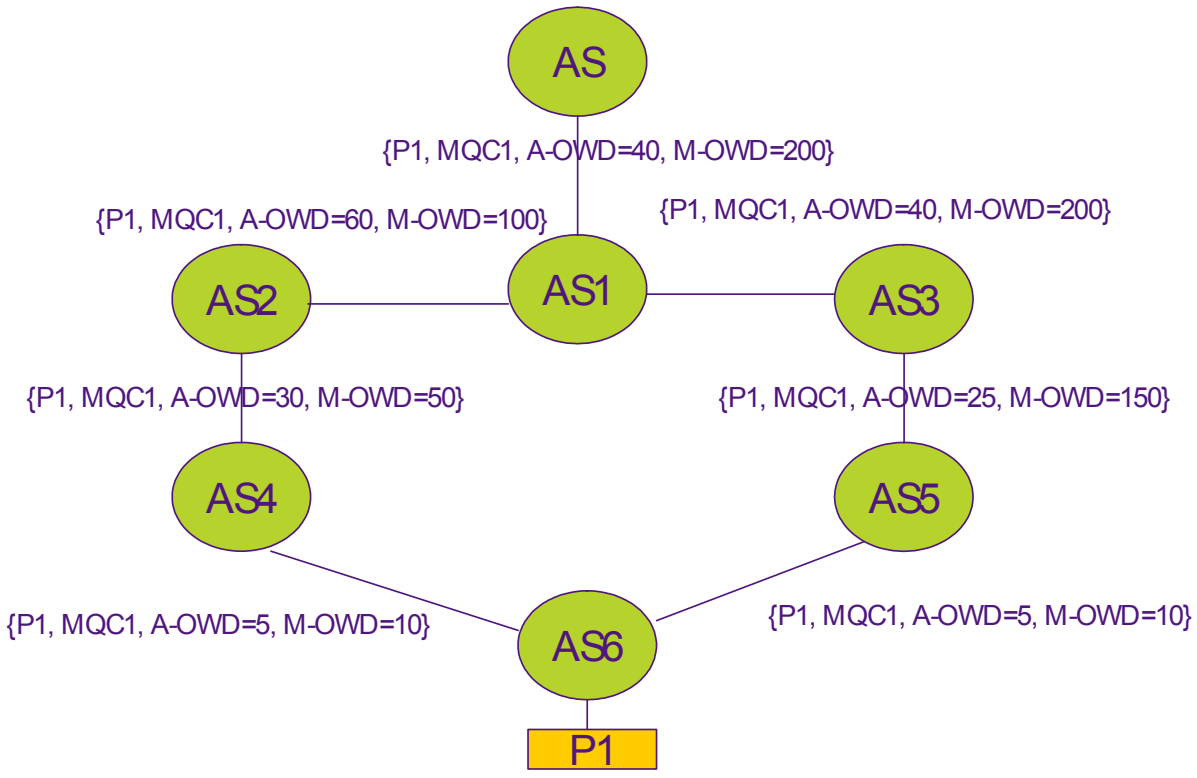


Figure 109. Example HGSO.

In this case, a computation request with the average and maximum one-way delay constraints set to 65 and 110 would be rejected by the PCE of the AS0. Indeed, if the PCE invoke its routing interface with the inter-domain routing process, the result will be negative since the stored route (A-OWD=45, M-OWD=210) in the local_RIB of ASBR doesn't satisfy the QoS constraints requested by the PCE. In this case, the PCE won't send a PCP request to AS1 because there is no advertised route that satisfies

Copyright © MESCAL Consortium, June 2005

the required QoS constraints, even if AS1 has stored in its adj_RIB_in an alternative route (via AS2) that satisfies these QoS constraints. If the PCE of AS0 sends this PCP request to AS1, it could succeed (depending on resources availability). The issue here is how to let AS0 knows about the existence of this alternative route.

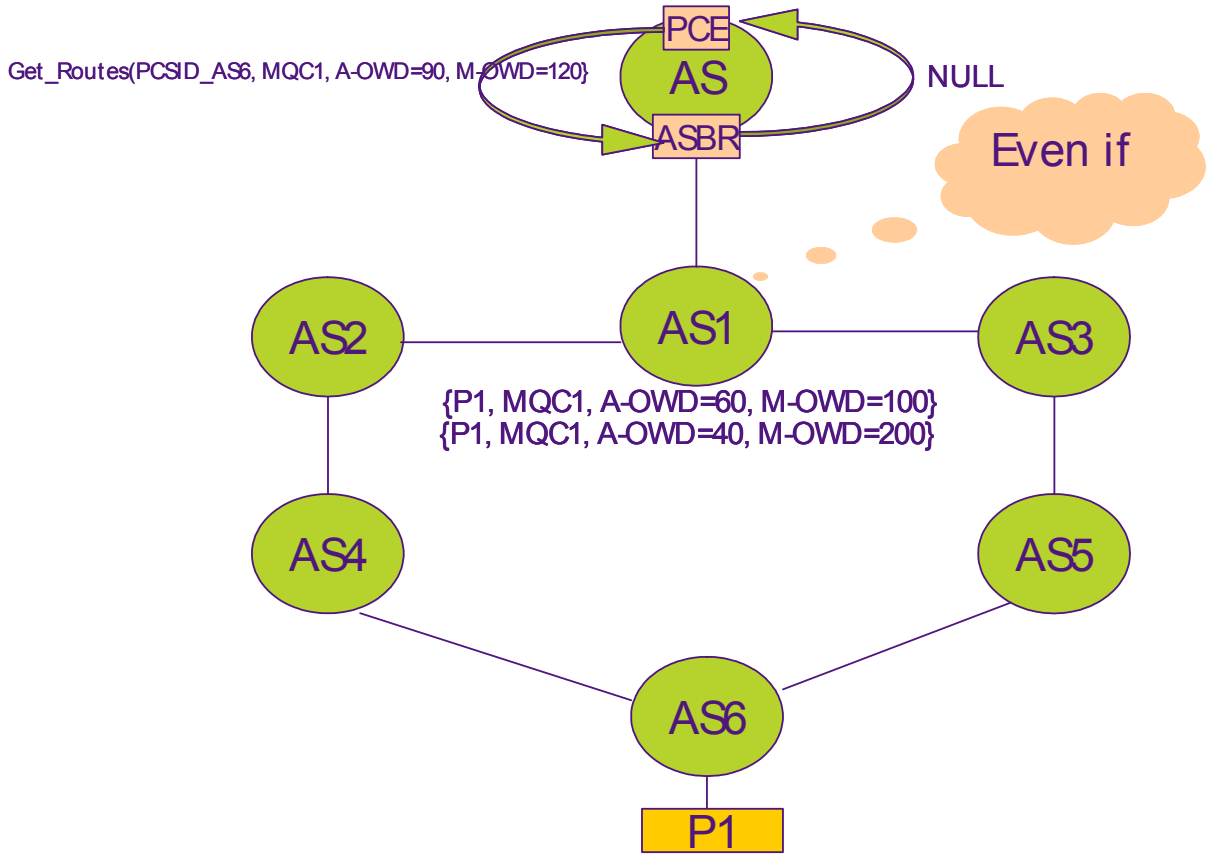


Figure 110. Example HGSO (bis).

In order to solve this issue, several solutions could be adopted. Hereafter a description of each solution:

- ***Solution 1***: Allow multiple-paths and announce all available q-BGP routes for a given prefix. In this scenario, q-BGP announcements will increase linearly. Nevertheless, the PCE has to trust on q-BGP announcements it will receive and must build its requests depending on received q-BGP announcements. The PCE philosophy in this scenario is: "*It is inutile to ask for something else better than what you have in your routing tables!"* In this case, the amount of requests that will fail due to QoS-related information is null.
- ***Solution 2***: Define a new q-BGP and/or PCP command to request alternative routes. In this scenario, a PCE or an ASBR can request its peer to send a copy of its Local_Rib containing routes towards a given destination. Based on received information, a PCE can create new requests, the risk that a request fail is equal to zero. But In order to achieve this goal, a new PCP/BGP message should be defined.
- ***Solution 3***: Announce a single route with/without an AS_PATH enclosing optimal values per QoS performance parameter. In this case, a provider announces to its peers the optimal values of each QoS performance parameter per m-QC. An AS path is prepended to this announcement. The way the AS path is constructed could be random or based on AS_SET. The spirit of this scenario is "*tune your requested QoS information based on this optimal values*". An adjacent PCE must be aware that those optimal values may not be ALL satisfied by a single route. Therefore, some PCP requests could be rejected due to the requested QoS performance values. In addition, a loop problem has to be taken into account when no AS is prepended.

Copyright © MESCAL Consortium, June 2005

- ***Solution 4***: Announce the best route with its QoS-related information and prepend the optimal values per QoS performance parameter per m-QC. This scenario is an enhancement of the previous one. The loop problem is obsolete in this case. In addition, the PCE tuning process is guided by the QoS values associated with the selected route, the optimal ones and the common understanding of m-QC. Although this enhancement, there still have some requests that will fail due to requested QoS performance values.
- ***Solution 5***: Announce a single route enclosing optimal value and an AS_PATH per QoS performance parameter. In this case, a provider announces to its peers the optimal values of each QoS performance parameter per m-QC. An AS path is prepended per QoS performance parameter. The spirit of this scenario is "*tune your requested QoS information based on this optimal values and this AS_PATHs*". An adjacent PCE is aware that those optimal values may not be ALL satisfied by a single route. As a consequence some PCP requests will be rejected due to the requested QoS performance values. Finally, we note that in there is no loop problem in this scenario.
- ***Solution 6***: Try all service peers' PCE with a high priority assigned to the AS that announced the best q-BGP route. This last scenario is "brute force" one. The spirit of this scenario is "*Query your q-RIB, if no route found try all your PCE peers*". Within this scenario, the number of PCE requests increase linearly. Therefore, the role of q-BGP becomes marginal. Note that timers setting should be carefully achieved.

In order to prevent from these kinds of behaviours, hereafter some generic guidelines:

- The PCE tuning process should be guided by the common understanding of m-QC.
- If there is a big gap between requested QoS performance values and available ones, the PCE administrator/user should ensure that there is not a need of changing the m-QC plane and use another m-QC instead.

10.5.1.5.6 Additional features

Other concepts need to be studied, such like MC-aggregation in order to reduce the volume of exchanged information.

Copyright © MESCAL Consortium, June 2005

10.5.2 Path Computation System

10.5.2.1 Inter PCE COMMUNICATION PROTOCOL

10.5.2.1.1 Terminology

This document makes use of the following terms:

- Path Computation Element (PCE): an entity that is responsible for computing/finding inter/intra domain paths for establishing LSPs. This entity can simultaneously act as client and a server. Several PCEs could be deployed in a given AS.
- Path Computation Client (PCC): a PCE acting as a client. This entity is responsible for issuing path computation requests that fulfil the Service Management constraints for the establishment of inter/intra domain LSPs.
- Path Computation Server (PCS): a PCE acting as a server. This entity is responsible for handling path computation requests in order to satisfy remote PCC constraints.
- Path Computation System (PCS): A system that uses PCE as a means to compute inter domain QoS constrained LSP paths.
- High-level service: is the service using a PCE-based system as an underlying infrastructure (an inter-domain QoS VPNs service for instance).
- High-level service customer: is a customer that subscribes to a High-level service.
- SLS Management: This management entity is responsible for SLS-related activities, including pSLS ordering (i.e. establishing contracts between peers) and SLS invocation (i.e. committing resources before traffic can be admitted).
- Domain: within this document it denotes an Autonomous system.

10.5.2.2 Introduction

The level of Quality of Service (QoS) guarantees offered by Internet Network Providers (INP) using a pure IP-based traffic engineering (TE) solution, other than overbooking, is not yet satisfactory for all corporate business services, for which strong guarantees must be provided. For this type of customers, hard QoS performance and bandwidth guarantees are considered as the major requirements. Currently, these requirements can be satisfied within a single domain or across several interconnected domains managed only by a single INP. However, it becomes very challenging when these domains are managed by different INPs.

MESCAL has specified three Solution Options that target distinct categories of customers and that provide different services guarantees. This section focuses on the Hard Guarantees Solution Option (a.k.a Solution Option 3 or HGSO), which has been designed to offer strict QoS guarantees for the corporate market (i.e. Hard Guarantees Service Option in conformance with MESCAL terminology).

This section presents the Path Computation Element (PCE), its interactions with the MESCAL functional blocks and provides a description of a first version of the inter PCE Communication Protocol (PCP). This specification has been realised to serve a basic QoS-constrained path computation without any enhanced features.

The following discussions rely on the inter-domain QoS signalling solution described in section 5.6.2.5.2 from D1.4 [D1.4].

10.5.2.3 Reminder

The Hard Guarantees Solution Option (HGSO) makes use of a dedicated entity called PCE (Path Computation Element), which is responsible for finding/computing an inter-domain path satisfying a

Copyright © MESCAL Consortium, June 2005

set of QoS performance guarantees in order to establish inter-domain QoS-constrained tunnels/LSPs. The computation of this path is distributed and requires communication between PCEs from different domains. The communication between two PCE entities is enabled thanks to the activation of Inter PCE Communication Protocol (PCP). Once "computed", the path is provided to the RSVP-TE/MPLS machinery of the head-end LSR, which can establish an inter-domain LSP that will follow the inter-domain path provided by the PCE.

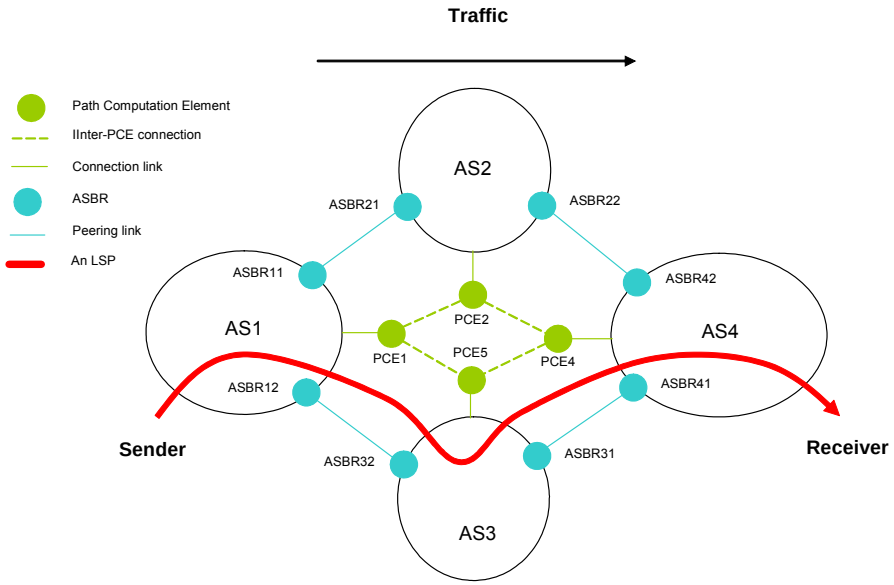


Figure 111-Overview

In Figure 111 above, each domain is assumed to support a set of meta-QoS-classes implemented by a set of l-QCs. In addition, “Hard Guarantees Service Option”-based pSLSs have been established between adjacent domains and q-BGP is consequently running between service peers.

A PCE at least is present in each domain that supports the HGSO. The PCE has an interface with routing processes (this could be implemented by allowing that a PCE receives q-iBGP announcements from all the ASBRs of its domain). This routing interface allows the PCE to know per *meta-QoS-class* plane, all the remote ASs that support the HGSO together with the associated QoS performance guarantees associated with an inter-domain path.

Each time a “Hard Guarantees Service Option”-based pSLS is established, the domains exchange their respective PCE information (name, IP address, identifiers, authentication information…) so that they can communicate.

In order to create an inter-domain QoS LSP, the domain which requests the establishment of the LSP asks its PCE(s) to compute an inter-domain path satisfying QoS constraints, expressed in term of *Meta-QoS-Class* availability along the path together with bandwidth guarantee per Meta-QoS-Class and optional constraints such as maximum end-to-end delay, etc. The first PCE selects one possible path among the set of path candidates and identifies the next-hop domain. It then verifies that the appropriate resources are available in its own domain and sets up administrative pre-reservations in the management system of its domain. Then it contacts the next hop PCE, requesting a path computation between the next hop ASBR and the termination address of the inter-domain LSP. This second PCE performs the same computation as the first one and the procedure is iteratively repeated up to the last PCE. If a path satisfying all requirements is found, each PCE returns the path received from the responding PCE concatenated with the sub-path it computed. When the last result reaches the originating PCE the whole path is available.

Copyright © MESCAL Consortium, June 2005

10.5.2.4 Interactions with MESCAL functional blocks

From the HGSO management point of view, the key service provided by the PCE is mainly path computation service, which consists in finding an inter-domain path satisfying a set of QoS constraints. Additional services like finding paths in which all inter-domain links have backup links could be supported by the PCE. But those services are outside the scope of this document.

Within the MESCAL functional model, the PCE is "only" responsible for computing an inter-domain QoS path. The implementation of the service (whether it is automated or not) and the creation of inter-domain LSP results from the cooperation of distinct functional blocks, including management plane blocks, control plane blocks and data plane blocks.

The PCE does not itself trigger the establishment of inter-domain LSPs, but provides inter-domain paths, when those are available. In particular, it is un-aware of business considerations but the HGSO management is. The PCE provides an interface for the higher-level functional blocks so that they can ask for path computation when necessary. It communicates with other remote PCEs thanks to the activation of PCP protocol. The PCE could request additional services from other functional blocks as illustrated in Figure 112.

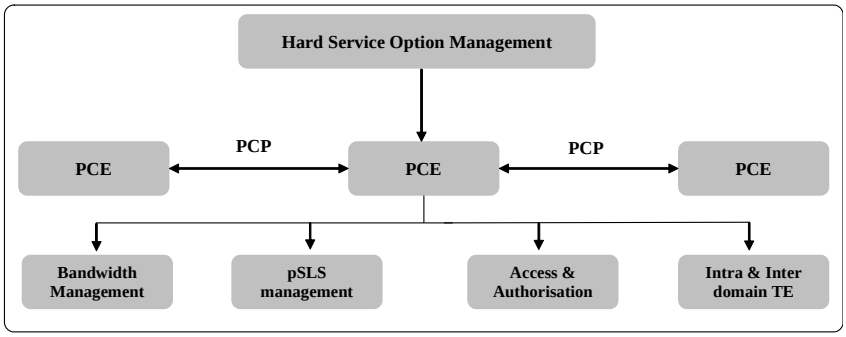


Figure 112-PCE interfaces

A pSLS established within the scope of the HGSO is considered as a right to request/establish inter-domain LSPs. The destination and the number of LSPs are not known in advance and could not be part of the pSLS. The pSLS indicates only the upper-boundaries that the upstream AS is allowed to use, in terms of meta-QoS-class that can be used for establishing inter-domain LSP and in terms of maximum bandwidth associated to each meta-QoS-class. The pSLS does not reserve any network resources in advance. Resources are actually allocated when an inter-domain LSP is set up.

However, it is difficult to establish such a contract in advance especially when the LSP path is not known. Thus, the sequence of operation for establishing an LSP should be:

- Compute inter-domain path candidate(s);
- Negotiate inter-domain QoS guarantees along the path for this particular LSP using information returned by the path computation;
- Establish the LSP once final contractual QoS guarantees terms have been end-to-end agreed.

The establishment of this contractual QoS guarantees negotiation can be difficult to achieve and can take some time. In particular, the risk is not negligible that the resources that were available when the PCE performed the path computation are no longer available along the path when the cascaded QoS guarantees negotiation (we will refer to this in the rest of the document by: "contract") are agreed, because others LSPs have used the corresponding resources.

In order to solve this issue it is necessary that the PCE of each domain makes an administrative reservation of the corresponding resources and indicates the characteristics of the path. This information is registered by the management plane, which triggers in parallel the creation of a provisional contract referencing the technical characteristics of the future LSP. Subsequent path computation requests may be impacted because the management plane removes these resources from

Copyright © MESCAL Consortium, June 2005

the available overall network resources. This provisional contract is valid for a limited time, which is the minimum date reported by each domain along the path. If the date exceeds this limit the provisional contract can be removed from the management systems, and related administrative network resources must be relaxed.

It is the responsibility of the management plane of each domain to cooperate in agreeing the exact financial terms and additional clauses of this contract, including its duration. Each domain knows the entry and the exit point of the LSP within its own domain and consequently knows both the upstream and downstream ASs to deal with. This validation procedure should ideally be automated to speed up the process and could integrate pricing negotiation. The way that the other blocks of the management plane deal with this automation is out the scope of this document.

Thus, once the contract is validated, the path computed by the PCE can be provided to the head-end LSR, which effectively sets up the LSP. Note that each ingress point of each domain should activate some outsourced policy functions that would allow RSVP-TE to get an authorisation right from the management platform.

The PCE interacts also with the intra and inter-domain TE blocks to retrieve routing information that is used to compute an inter-domain path satisfying expressed QoS constraints. An interface must be made available to the PCE so that it can access to this information. Note that both intra and inter-domain routes must be made available to the PCE

In addition, for access control and authorisation purposes, the PCE must be provided with access to the list of other PCEs from which it will accept requests. This list is updated each time a “Hard Guarantees Service Option”-based pSLS is agreed by the provider.

10.5.2.5 PCE discovery

As described above, the “Hard Guarantees Service Option”-based pSLS indicates the IP address of the service peer PCE(s). This information is stored in the SLS Management Systems of each INP. As described in [BOUC05][D1.4], instead of announcing all potential tail-end addresses in q-BGP, only an identifier is announced via q-BGP. It is called the Path Computation Service Identifier (PCSID). This particular q-BGP announcement is identified by a well-known community value and is represented by a routable IP address, which can be different from the real IP address of the PCE.

q-BGP announcements of PCSID will ease to discover the set of remote ASs supporting HGSO service and associated end-to-end QoS-related information for reaching them. In order to compute a path towards a specific domain supporting the HGSO service, the local PCE chooses a route that serves the PCSID of that domain and extracts from the AS_PATH attribute the AS number of the next hop ASBR. Then, the local PCE queries its SLS Management system and gets back the PCE's IP address of the next neighbouring PCE to contact. Finally, the local PCE forms and forwards a path computation request to the next PCE. The process is iteratively repeated until the request reaches the PCE of the target AS identified by its PCSID.

10.5.2.6 The PCE Communication protocol (PCP)

10.5.2.6.1 Overview

This section describes the communication protocol used between two PCE in order to collaborate for computing inter-domain QoS constrained paths, this protocol is called inter-PCE Communication Protocol (PCP).

The main characteristics of the PCP protocol are as follows:

- The protocol employs a client/server model in which a PCE can both act as a client and/or a server at the same time. A PCE Client (PCC) sends requests, cancels and receives responses.

Copyright © MESCAL Consortium, June 2005

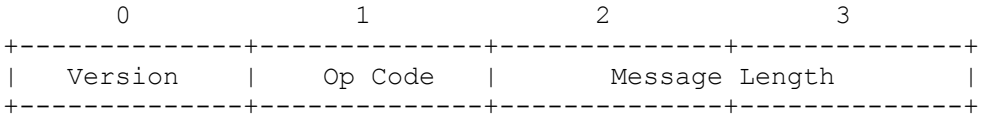
- The protocol uses TCP as its transport protocol for reliable exchange of messages between PCEs. Therefore, no additional mechanisms are necessary for reliable communication between two PCEs.
- In this version, PCP does not provide any message level security for authentication, message replay protection, and integrity. However, PCP can reuse existing protocols for security such as IPSEC [RFC2401] or TLS [RFC2246] to authenticate and secure the communication channels between PCEs.
- The current PCP protocol provides the service for supporting only a basic path computation function. In particular it does not support additional path computation constraints, or provide enhanced reporting features in the case of path computation failure.

10.5.2.6.2 PCP messages

This section describes the PCP message formats and objects exchanged between PCEs.

10.5.2.6.2.1 Common header

Each PCP message consists of the PCP header followed by a list of arguments depending on the nature of the operation.



Global note: //// implies field is reserved, set to 0.

The fields in the header are:

Version: 8 bits. PCP version. Current version is 1.

Op Code: 8 bits. The PCP operations are:

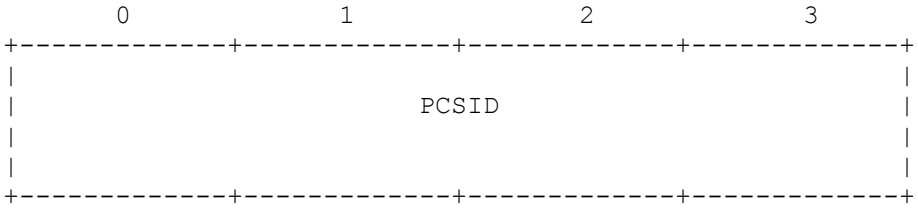
1 = OPEN	(OPN)
2 = ACCEPT	(ACP)
3 = CLOSE	(CLO)
4 = REQUEST	(REQ)
5 = RESPONSE	(RSP)
6 = PATH-ERROR	(ERR)
7 = CANCEL	(CCL)
8 = ACKNOWLEDGE	(ACK)
9 = KEEP-ALIVE	(KA)

Message Length: 16 bits

This is the size of the message in octets, which includes the standard PCP header and all encapsulated objects. Messages MUST be aligned on 4 octet intervals.

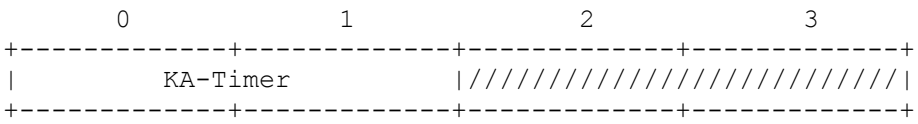
Copyright © MESCAL Consortium, June 2005

10.5.2.6.2.2 OPEN message



The message contains only one argument. This PCSID is propagated by q-BGP between the domains. This is a routable IPv4 or IPv6 address identifying a domain supporting the HGSO service. This PCSID MUST been inserted by the PCE in the OPEN message of a PCP session. The size of the PCSID is 4 for IPv4 and 16 bytes for IPv6 addresses.

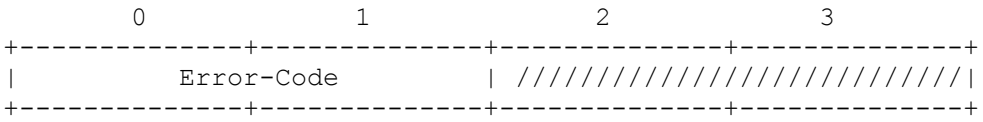
10.5.2.6.2.3 ACCEPT message



KA-Timer (Keep-Alive Timer): The argument of the accept message is a 2 octets integer value which represents a timer value expressed in units of seconds. This timer value is treated as a delta. KA-Timer is used to specify the maximum time interval over which a PCP message MUST be sent by the two communication entities. The range of finite timeouts is 1 to 65535 seconds represented as an unsigned two-octet integer. The value of zero implies infinity.

10.5.2.6.2.4 CLOSE message

The close message contains an error code indicating the reason of the close of the session.



Error-Code:

1 = Shutting Down

2 = Bad Message Format

3 = Incorrect identifier

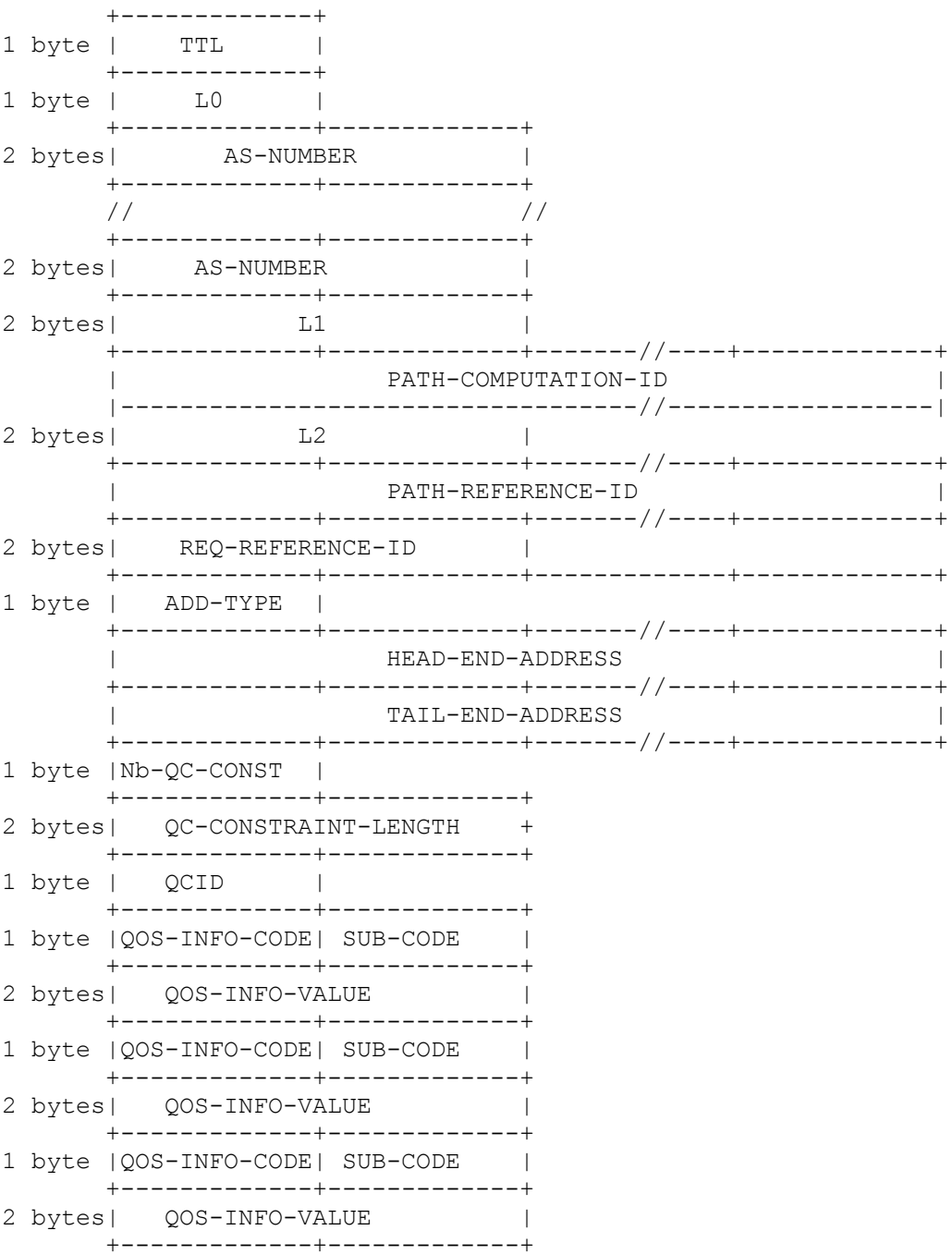
4 = Unable to process

5 = Protocol error

Copyright © MESCAL Consortium, June 2005

10.5.2.6.2.5 REQUEST message

The Request message is sent by the PCC for computing an inter-domain path.



- TTL: is the maximum number of ASs that can be crossed by the path. This field is decremented by one each time a PCE handles a request.
- L0: is a 1-byte length field. It represents the number of ASs that have been already crossed.
- AS-NUMBER: is a 2 bytes length field representing an AS number. The first AS-NUMBER value of the list is the AS-NUMBER of the PCE that initialised the path computation.
- L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.
- PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. It is suggested, at least in this first specification, that this

Copyright © MESCAL Consortium, June 2005

identifier is computed using the PCSID of the domain, concatenated with the date and an identifier that will be computed by the first requesting PCE each time a request will have to be issued. Across PCE reboots, this identifier MUST be unique. This PATH-COMPUTATION-ID will be replicated in all subsequent request initiated by the PCEs along the path.

- L2: is the length in bytes of the PATH-REFERENCE-ID. Size of this field is 2 bytes.
- PATH-REFERENCE-ID: is a variable-length field. It is an identifier that represents a pre-agreement between the head and the tail-end domain that allows the PCE from the terminating domain to accept or reject the path computation request.
- REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to identify the REQUEST. It allows making the difference between several REQ messages issued for different path computation (with the same PATH-COMPUTATION-ID) between two neighbouring ASs interconnected via multiple links.
- ADD-TYPE: indicates the nature of the IP addresses of the tail-end and head-end termination:
  - 1 = IPv4
  - 2 = IPv6
- HEAD-END-ADDRESS: is the head-end address of the future LSP represented in the form HEAD-END@PCSID. This is a couple of IPv4 or IPv6 address. The first address of the couple identifies a loopback or an interface address of a network element, the second element is the PCSID of the domain owning the previous address.
- TAIL-END-ADDRESS: is the tail-end address of the future LSP represented in the form TAIL-END@PCSID. This is a couple of IPv4 or IPv6 address. The first address of the couple identifies a loopback or an interface address of a network element, the second element is the PCSID of the domain owning the previous address.

This above parameters MUST be present in each REQUEST and in the same order.

- NUMBER-OF-QC-CONSTRAINT: represents the number of QoS class constraints the PCE must take into account when computing a path. A QoS class constraint contains a QoS-Class-Identifier followed by additional constraints. The size of this filed is 1 byte.
- QC-CONSTRAINT-LENGTH: is the length in bytes of the QoS-Class-Constraint that follows. The size of this field is 2 bytes.
- QOS-CLASS-IDENTIFIER: is an identifier of a QoS-class (could be a DSCP value). The size of the field is 1 byte.
- QOS-INFO-CODE: this field identifies the type of QoS-related information. The size of this field is 4 bits.
  - (0) Reserved
  - (1) Packet rate
  - (2) One-way delay metric
  - (3) Inter-packet delay variation
- QOS-INFO-SUB-CODE: this field carries the sub-type of the QoS-related information. The following sub-types have been identified. The size of this field is 4 bits.
  - (0) None
  - (1) Reserved rate
  - (2) Available rate
  - (3) Loss rate

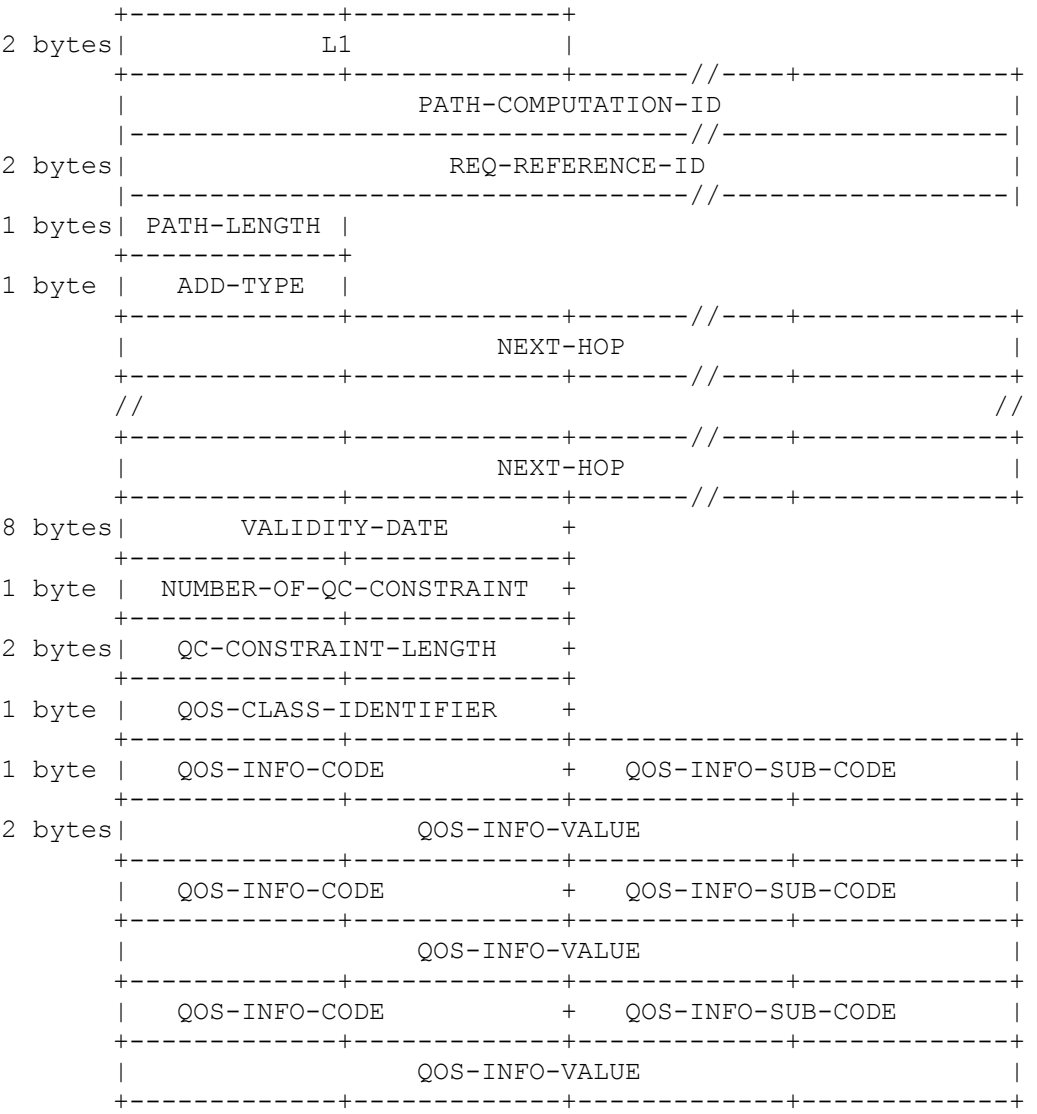
Copyright © MESCAL Consortium, June 2005

- (4) Minimum one-way delay
- (5) Maximum one-way delay
- (6) Average one-way delay

- QOS-INFO-VALUE: this field indicates the value of the QoS information. The corresponding units depend on the instantiation of the QoS information code.

10.5.2.6.2.6 RESPONSE-PATH message

This message is sent back when a path has been successfully computed.



- L1: is the length in bytes of the PATH-COMPUTATION-ID. The size of this field is 2 bytes.
- PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. This value of this identifier must be the same as the one provided by the REQUEST.
- REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.
- PATH-LENGTH: indicates the number of next hops that form the path. The size of this filed is 1 byte.
- ADD-TYPE: indicates the nature of the IP addresses in the PATH. The size of this filed is 1 byte.

Copyright © MESCAL Consortium, June 2005

- 1 = IPv4
- 2 = IPv6

- NEXT-HOP: IP address of a next hop that is part of the computed path. The size of this field depends on the nature of the IP address.
- VALIDITY-DATE: represents the GMT date after which the computed path returned will not be valid. The size of this field is 8 bytes.

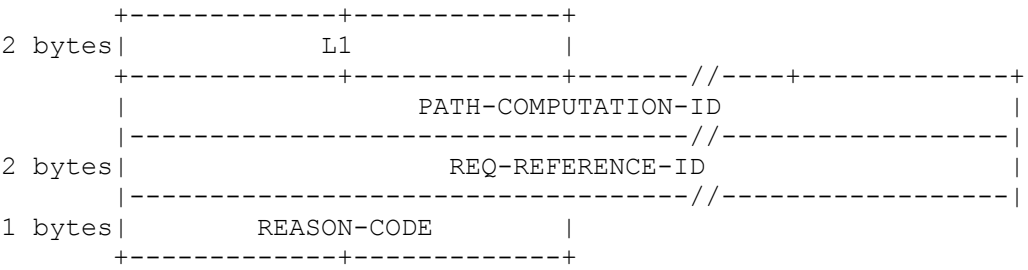
These above parameters MUST be present in each RESPONSE and in the same order.

The other parameters have the same meaning than for the REQUEST except:

- QOS-INFO-VALUE: represents the QoS guarantees of the path, for this particular QoS-INFO-CODE parameter (delay, jitter, etc.) between the ingress ASBR of the responding PCE and the tail-end of the path.

10.5.2.6.2.7 PATH-ERROR message

This message is sent back when a path could not be computed.

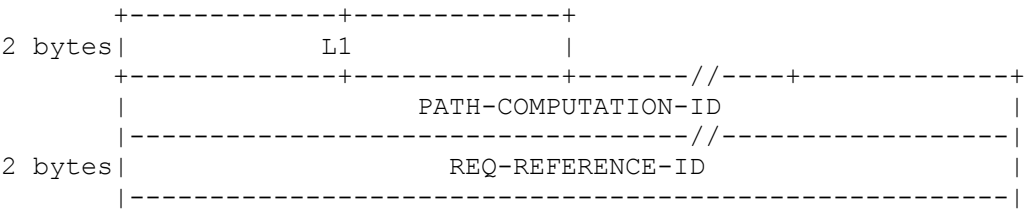


- L1: is the length in bytes of the PATH-COMPUTATION-ID. The size of this field is 2 bytes.
- PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. This identifier MUST be the same as the one enclosed in the REQUEST.
- REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.
- REASON-CODE: indicate the reason of the failure. Identified failure are:
  - 1 = No resource available
  - 2 = Path reference error
  - 3 = Abnormal termination
  - 4 = PATH-COMPUTATION-ID already used
  - 5 = TTL expired
  - 6 = Loop detected
  - 7 = Request already handled

Copyright © MESCAL Consortium, June 2005

10.5.2.6.2.8 CANCEL message

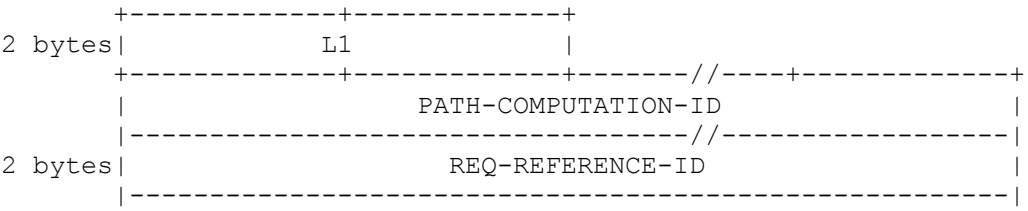
This message is sent by a PCC or a PCS when a path computation is to be cancelled.



- L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.
- PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. This identifier MUST be the same as the one embedded in the REQUEST message.
- REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.

10.5.2.6.2.9 ACKNOWLEDGE message

This message is sent by a PCC to a PCS to confirm the reservation of the path. This feature is particularly used when a PCC launches multiple REQUEST messages during its path computation phase.



- L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.
- PATH-COMPUTATION-ID: is globally unique value that identifies a path computation occurrence. It is a variable-length field. This identifier must be the same as the one provided by the REQUEST.
- REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.

10.5.2.6.2.10 KEEPALIVE message (KA)

Message exchanged between two PCEs to maintain TCP session when no other messages are exchanged. This message has no argument.

10.5.2.6.3 Exchange of PCP messages

10.5.2.6.3.1 Communication

The PCP protocol uses a single persistent TCP connection between a PCC and a remote PCS. The PCS MUST listen on a well-known TCP port number (to be defined). The PCC is responsible for initiating the TCP connection to PCS. The location of the remote PCS is deduced and retrieved from the management plane blocks during the path computation process or at PCE boot via the pSLS management block.

Copyright © MESCAL Consortium, June 2005

10.5.2.6.3.2 OPEN (OPN)

An OPN message MUST be sent before any other message exchange. As part of the open message, the PCC provide its PCSID, which allows the server to identify the client. It can also use this information to retrieve the client context near its management plane. Only one OPN message can be issued at a time. If the PCS receives malformed message it MUST close the session using the appropriate error code.

10.5.2.6.3.3 ACCEPT (ACP)

The ACP message is used to positively respond to the OPN message from the PCC. This message will return to the PCC a KA-timer value object indicating the maximum acceptable intermediate time between the generations of messages by the PCEs. The KA-timer value is determined by the PCS and is specified in seconds. If the PCS refuses the PCC open message, it will instead issue a CLOSE message.

10.5.2.6.3.4 CLOSE (CLO)

The CLOSE message can be issued by either the PCC or the PCS to notify to its adjacent PCE that it is no longer available. The Error code is included to describe the reason for the close.

When issuing a CLOSE both the PCC and the PCS MUST delete all internal states related to this PCP session. Additionally, all pending requests MUST be cancelled in order to release all pending resources reservations that have been established. PATH-ERROR or CANCEL message MUST be sent depending on requests' state.

10.5.2.6.3.5 REQUEST (REQ)

A request is issued by a PCC when it has found a potential path toward the target final destination. This request can be issued as a consequence of a request received from another domain it has agreement with or from its own Service Management Plane.

When the service request comes from a remote PCC, the server achieves the following tasks:

(0) If the receiving TTL is zero the PCS MUST discard the request. If not, the receiving PCS, decrements by one the TTL value. If the updated value of TTL is equal to zero, the request is rejected if the PCE is not the last one in the chain. In addition the PCS examines the AS-PATH included in the received REQ and reject it if it finds its own AS number in the list. This mechanism allows avoiding possible loops.

(1) It checks if the PATH-COMPUTATION-ID of the received REQ is already associated to a pre-contract or contract for the same requester. If this is the case, it returns a PATH-ERROR message with a reason-code = 4. It checks if the PATH-COMPUTATION-ID and the REQ-REFERENCE-ID of the received REQ are already associated to a pre-reservation record concerning the same requester. If a pre-reservation is found, it returns a PATH-ERROR message with a reason-code = 4.

(2) The PCS examines the HEAD-END-ADDRESS and the TAIL-END-ADDRESS parameters present in the request. The HEAD-END-ADDRESS MUST indicate a valid entry point in its domain. If not, the PCS returns a PATH-ERROR with an appropriate reason value.

(3) Then it extracts the PCSID from the TAIL-END-ADDRESS and parses the QoS constraints provided by the request message.

(4) The PCS achieves some policing and verifies that the request constraints will not exceed the resources negotiated in the related pSLS. If resources are exceeded, the PCS returns a PATH-ERROR message. If resources are available, the PCE pre-reserves the corresponding resources near the management plane.

(5) If the PCS recognises its own PCSID in the TAIL-END-ADDRESS, it considers the PATH-REFERENCE-ID otherwise it jumps to step (6). If this identifier is known from its management plane, the request is accepted and processing continues on (51). Otherwise the PCS returns a PATH-ERROR message with a reason-code = 2.

Copyright © MESCAL Consortium, June 2005

(51) The PCS computes an intra-domain path and verifies the availability of the resources along this internal path. If available, the PCS interacts with its management plane and creates a context, which triggers the administrative reservation of the resources. When interacting with the management blocks, the PCS MUST provide all information necessary to identify the sub-path it selected. In particular it MUST provide the PATH-COMPUTATION-ID, the REQ-REFERENCE-ID, the ingress point ASBR address used in its domain and the termination point in its domain. The PCS sends a RESPONSE-PATH message back to the requesting PCC. If resources are not available a PATH-ERROR message is generated.

(6) It then queries the dynamic inter-domain traffic-engineering block with the retrieved PCSID and the list of requested QoS-classes. The dynamic inter-domain TE block returns the available q-BGP announcements. The PCS then verifies whether it can find a next-hop ASBR, which announces the PCSID within the requested QoS-class. If cannot find it the procedure stops and a PATH-ERROR message is returned back to the requesting entity with an appropriate reason-code value.

(7) If one or several next-hops are found, the PCS examines the QoS performance guarantees of the announcements and compare the values with the requested ones. If it doesn't understand one of the requested QoS constraints, a PATH-ERROR message is sent back. Otherwise, QoS constraints are successively compared to those received from q-BGP. All next-hops propagating the set of announcements satisfying the required QoS constraints are kept.

(8) For each possible next hop ASBR the PCS checks if there are enough available resources at the inter-domain links. In particular if some bandwidth guarantees are required the PCS checks if the administrative maximum bandwidth agreed during the pSLS negotiation phase will not be exceeded. If resources are not available the ASBR is left on side and the next ASBR in the list is considered. If resources are available, the PCS pre-reserves the corresponding resources near the management plane. At this stage, the management plane doesn't create any contract since we are not sure that an end-to-end path exists. This pre-reservation can be taken into account by the PCS for subsequent requests. It can use it as a lock and delay the incoming requests or introduce the pre-reservations in its resource availability computation according to the local policy enforced. When interacting with the management blocks, the PCS must provide all information necessary to identify the sub-path it selected. In particular it must provide the PATH-COMPUTATION-ID, the REQ-REFERENCE-ID, the ingress point address of its domain and the ingress point address of the next domain. This latter information can be used by the management plane to identify the upstream and downstream involved domains.
The PCS computes an intra-domain path and verifies the availability of the resources along this internal path. If resources are available, the sub-path is valid and the PCE forms a new REQUEST message which is sent to the PCS of the remote domain owning the next-hop ASBR. It adds its own AS number to the existing list. If internal resources are not available, the PCS discard the pre-reservation and considers the next hop ASBR in the list. When building the request the PCC keeps the PATH-COMPUTATION-ID, the PATH-REFERENCE-ID, the TAIL-END-ADDRESS unchanged. The initial HEAD-END-ADDRESS is replaced by the address of the ingress next-hop ASBR identified during the path computation. The QoS constraints characteristics are modified in order to take into account the QoS performance guarantees provided by the domain.

(9) If QoS constraints cannot be satisfied for any of the ASBR, the PCS returns a PATH-ERROR message.

Note that it is quite possible that several next hops ASBR can satisfy the requested constraints. In such a case the PCS can process one next-hop ASBR at a time or several in parallel. For one incoming request, there can be multiple simultaneous outgoing requests towards different PCS. If several requests are sent toward the same neighbour, for a same PATH-COMPUTATION-ID, the REQ-REFERENCE-ID MUST be different.

Copyright © MESCAL Consortium, June 2005

10.5.2.6.3.6 RESPONSE (RSP)

A RESPONSE message is sent by a PCS in response to a request issued by a PCC. RSP messages are sent back when a valid end-to-end path has been computed. The RSP message MUST be initiated by the tail-end domain.

When a valid end-to-end path has been computed, the PCS of the last domain on the path, forms a RSP message. It first inserts the original PATH-COMPUTATION-ID. Then its forms a path argument that MUST contains the IP address of the tail-end LSP and the IP address interface of the ingress ASBR supporting that path. It MAY insert between these two extremities, the IP address of additional hops. It MAY also indicates the date after which the path will not be valid anymore because administratively reserved resources will have been relaxed. Then, it MUST indicate QoS guarantees it provides between the ingress ASBR and the tail-end address of the LSP. The RSP message is then sent to the requesting PCC.

On receipt, the PCC adds its own intra-domain sub-path to the list. It does not indicate the next-hop ASBR since this latter has already been inserted by the downstream PCS. This sub-path can be a strict or loose. It also modifies the QoS guarantee parameters so that they reflect the QoS guarantees it can provide for its own part of the path. The VALIDITY-DATE MUST be modified so that the value indicates now the smaller date between the date received in the RSP message and the date reported by the management plane.

If the PCC sent multiple REQUEST messages in parallel, it MAY wait for a RSP or ERR message for all the requests it sent. If the PCC got multiple RSP messages it MUST select only one and inform the un-selected PCS that they can cancel their reservation. It forms CANCEL messages, sends them to the appropriate PCS and cancels its own pre-reservation for the corresponding requests. If the PCC doesn't wish to wait for a reply, it can send a CANCEL message at any time.

10.5.2.6.3.7 ACKNOWLEDGE (ACK)

The ACK message is used by PCE to confirm to its management plane that the resources needed for the path referenced by PATH-COMPUTATION-ID and REQ-REFERENCE-ID present in the message need to be reserved. It allows the management plane to create a contract based on information previously stores by the PCE during the computation phase. If no ACK is received, no contract is created and the negotiation at the management level will fail. If for some reasons, no ACK were received, the VALIDITY-DATE will be used and the administrative pre-reservation automatically removed for that path. ACK messages are only accepted if they arrive after the server has issued a RSP, otherwise they are ignored.

10.5.2.6.3.8 CANCEL (CCL)

A CANCEL message can be sent by PCC and PCS. CCL messages can be generated during the normal path computation cycle but also in case of an abnormal termination of a PCE to PCE communication. If a PCS, received a CCL message from a PCC, it MUST form new CCL messages and forward a CCL message to each PCS to which it sent a REQ and for which it did not received any positive or negative reply. Once this has been achieved it MUST delete all its internal states referencing the request identified by the PATH-COMPUTATION-ID and REQ-REFERENCE-ID indicated in the message. If the PCE has no pending request concerning this PATH-COMPUTATION-ID and REQ-REFERENCE-ID, it can optionally query its management plane to retrieve a possible existing contract referenced by this PATH-COMPUTATION-ID and delete it. Just before deleting this contract, it can form a new CCL message and forward it to the next PCS in the path. If it does not, the VALIDITY-DATE will be applied.

The same procedure applies if the PCS detects a communication problem with one of its PCC. In that case, the PCS issues CCL messages for all pending request received from this PCC.

Copyright © MESCAL Consortium, June 2005

10.6 IP-based Intra-domain TE

10.6.1 Introduction

10.6.1.1 Overall Objectives

The purpose of intra-domain Traffic Engineering is to configure the intra-domain network in such a way that it satisfies the requirements of the traffic forecast. The forecast provides Intra-domain Traffic Engineering with demands for ingress, egress pairs and QoS constraints. The Intra-domain Traffic Engineering functional block is responsible for the distribution of this traffic among the available network resources as efficiently as possible while honouring the given QoS constraints.

The functional architecture diagram in Figure 113 highlights the *Offline Intra-domain Traffic Engineering* block.

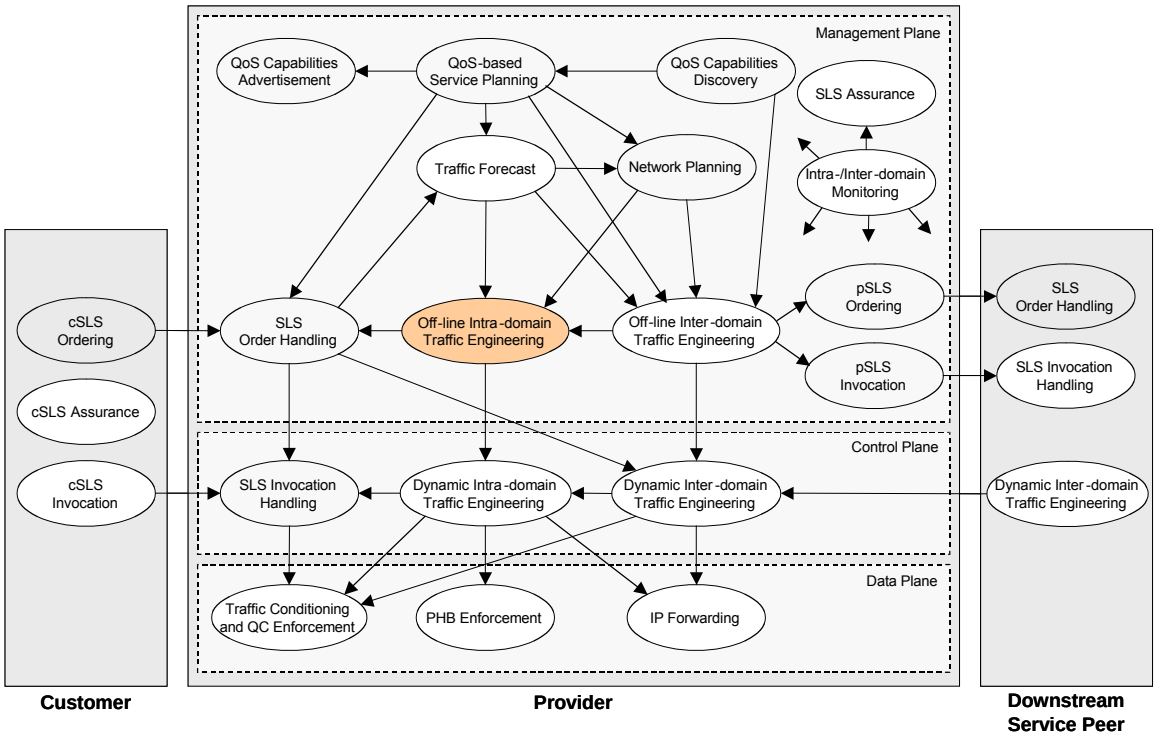


Figure 113. MESCAL Functional Architecture, highlighting Offline Intra-domain TE

Ideally, the distribution of traffic on the available network resources is carried out without any implementation constraints. However, the IP based Traffic Engineering described in this chapter is based on optimising OSPF link weights. Link weight based shortest path routing is not as flexible as an optimal routing solution. This is because the routing is limited to the shortest path algorithm that is run individually by each router of the network. OSPF is using only the destination address and link weights to route packets, whereas more differentiated routing decisions would require more information. A mechanism allowing for more flexibility is MPLS, however, it comes at the cost of introducing state information into the network, as well as loss of the self management capability (routing) of IP networks.

The IP traffic engineering algorithm proposed in this chapter is based on the assumption that the flexibility of MPLS like solutions is not required to efficiently support quality of service, but that the efficient optimisation of OSPF link weights allows sufficient control.

Copyright © MESCAL Consortium, June 2005

In addition to the OSPF link weight traffic engineering, it is proposed to support one weight setting per DiffServ code point (DSCP) on each link. This is not to be confused with the DiffServ Per Hop Behaviour (PHB), which is a queuing behaviour defined by as part of the Differentiated Services [RFC2475]. The DSCP values can be mapped onto a PHB as well as serve other purposes, such as provide additional routing information. When considering the DSCP value as additional routing information, it is possible to route all traffic aggregates one DSCP "routing plane" differently to that of other DSCP routing planes.

The Offline Intra-domain IP Traffic Engineering block is also responsible for adapting the network to predicted changes in demand. These changes include an increase in demand during office hours and evenings. When considering different QoS enhanced services, changes in service may also occur during these hours, VoIP during office- and streaming Video during evening hours. Further possibilities are surges in demand, in the event of a public holiday where resources could be committed away from office applications. A different type of change could be the failure of an important network link, for which a backup configuration is available. In summary, offline Intra-domain TE can put into effect predicted, pre-computed scenarios. These pre-computations are carried out by the offline-TE during its idle times between Resource Provisioning Cycles. It is important to point out that offline intra-domain IP TE cannot manage sudden changes in demand or topology that where not previously predicted without the processing needed to compute a set of link weights. Therefore, while the link weight computation is not dynamic, the algorithm responsible for recognising the need for network adaptation and for effecting weight modifications is. As discussed in detail in section 10.6.2.3, real link weight computation is computationally expensive and so it is difficult to conceive of more dynamicity in the actual weight computation.

This study has two purposes. To improve the IP based intra-domain traffic engineering results from the IST-Tequila project and to take a fresh approach at the intra-domain routing problem in the light of the inter-domain traffic engineering developed in IST-MESCAL. Special emphasis is placed on the optimisation of interactions between inter-domain and intra-domain traffic engineering.

10.6.1.2 Decomposition of Functionality

Figure 114 shows the internal structure of the Offline Intra-domain Traffic Engineering block and its relationship with neighbouring blocks. There are two sub-components, *Resource Optimisation* and *Network Reconfiguration Scheduler*.

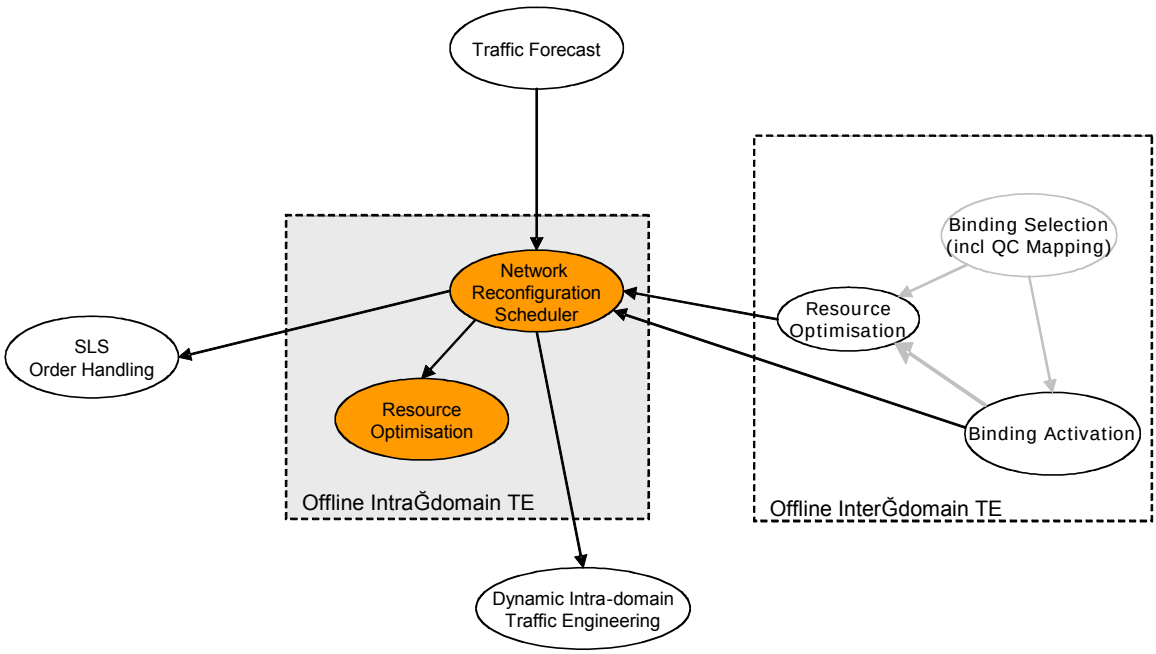


Figure 114. Decomposed Intra-domain Traffic Engineering

Copyright © MESCAL Consortium, June 2005

The *Resource Optimisation* block contains the OSPF link weight optimisation algorithm. It is a passive block, until called by the *Network Reconfiguration Scheduler* at which point it collects a traffic demand matrix and a network topology and computes an optimal set of link weights. Computed weights are deposited in a link weight database inside the Offline Intra-domain Traffic Engineering block, until they are put into operation in the network by the *Network Reconfiguration Scheduler*.

The *Network Reconfiguration Scheduler* is the control system for the Offline Intra TE block. It has two main purposes, handling computation requests to *Resource Optimisation* (Resource Provisioning Cycles, Inter-domain Traffic Engineering "what if" queries, etc) and scheduling the reconfiguration of the network using link weight settings computed by the *Resource Optimisation* block. Requests for network reconfiguration that have not been computed previously are passed to the *Resource Optimisation* together with information on where to retrieve traffic demand matrices and network topologies for the link weight computation. The *Network Reconfiguration Scheduler* uses its scheduling capabilities in order to predict (or be alerted to) the periodic changes in demand, for which pre-computed network configurations are available in the link weight database. When the scheduler is alerted to yet unknown changes in demand or topology it invokes *Resource Optimisation*, the results of which are implemented in the network but are also stored for possible reoccurrences in the future. This way the scheduler learns about the dynamic behaviour of the network, improving its effectiveness with time.

10.6.2 Resource Optimisation

10.6.2.1 Objectives

- Compute a set of OSPF link weights given a *Traffic Forecast* and while honouring QoS constraints and optimising network capacity.
- Support "what if" scenarios to *Offline Inter-domain Traffic Engineering* in order to optimise inter- intra-domain interactions.
- Provide network configuration scenarios to *Network Reconfiguration Scheduler* for pre-computation.

10.6.2.2 Interface Specification

Figure 115 enlarges the decomposed Offline Intra-domain Traffic Engineering, providing a clear definition of interfaces required internally.

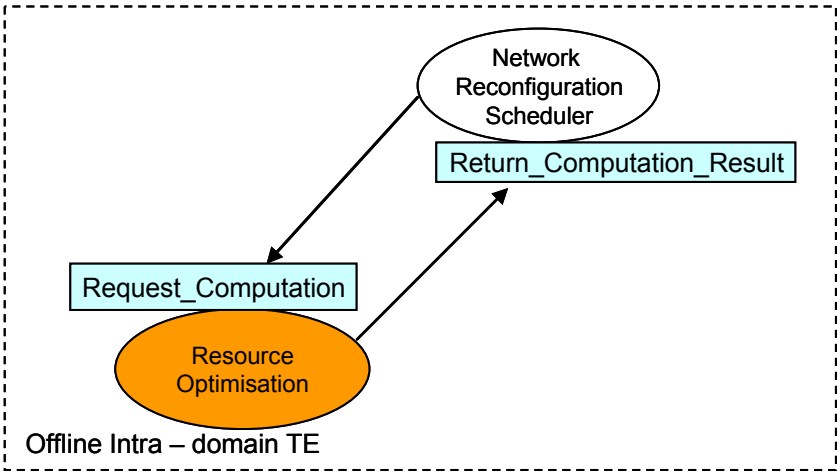


Figure 115. Interactions of the Resource Optimisation block

Copyright © MESCAL Consortium, June 2005

- **Resource Optimisation & Network Reconfiguration Scheduler**
  Request_Computation(handle to iTM)

  The Network Reconfiguration Scheduler requests the computation of link weight optimisation for a particular network topology and demand matrix. Location of both topology and demand matrix have to be specified by the Network Reconfiguration Scheduler at the time of the request (locations for these may be databases of *Traffic Forecast* or databases for "what-if" scenarios inside *Offline Intra-domain Traffic Engineering*).

  Return_Computation_Result()

  Once the Resource Optimisation has completed the request, it stores the information in the link weight database and notifies the Network Reconfiguration Scheduler.

10.6.2.3 Algorithm Description

10.6.2.3.1 Motivation

Distributing traffic demands on OSPF networks is a challenge, because of the indirectness of the traffic engineering problem. Finding optimal paths for each traffic demand as for MPLS is only half of the problem, the other half is the implementation of these paths with OSPF link weights. While MPLS allows the explicit pinning of a route between any two nodes and effectively switches packets according to this configuration, OSPF relies on individual routing decisions taken at each node. These are based solely on destination IP address and the networks link weight metrics that determine the shortest path towards the destination address. It is therefore clear that mapping an "optimal" demand distribution as calculated for an MPLS network onto an OSPF network is not an achievable goal in most practical cases. Instead, the algorithms for finding good paths and for translating these paths into link weights, have to go hand in hand, iteratively searching for better link weights to spread the load across the network.

Several link weight optimisation techniques to achieve this have been proposed. There are heuristic, genetic and hybrid (memetic) algorithms for solving the link weight computation. Most proposed algorithms follow slightly different optimisation criteria, minimum average utilisation, maximised capacity or a combination of those and other criteria weighted in a cost function. A much cited link optimisation algorithm was developed by Bernard Fortz and Mikkel Thorup and has been published in [FORT00], this paper is also proving the NP-hardness of the link weight optimisation problem. Several iterations and improvements followed in [FORT02a, FORT02b and others]. The algorithm aims to maximise the networks free capacity and minimise the number of heavily congested links. Several experimental studies followed, aimed at improving the heuristic weight setting algorithm described in [FORT00]. Algorithms described in [Eric02, BURIO02, Riedl] are all exploring genetic and memetic heuristics. More recently, [WANG04] uses the DSCP for their k-set traffic engineering. This paper mainly targets use of the routing planes for non-QoS enabled traffic engineering. Similarly, [PSEN05] presents Multi-topology OSPF routing, again, using the DSCP field to facilitate differentiation between the different MT-topologies.

All IP link weight optimisation algorithms discussed so far are aimed at some form of network utilisation optimisation. None are explicitly designed to satisfy quality of service constraints of the traffic. The algorithms designed as part of the IST-Tequila [TEQUI,D1.4] project are specifically targeted at traffic engineering in QoS aware networks and support DSCP aware routing.

10.6.2.3.2 Optimisation Algorithm Objectives

- The primary objective is to distribute the demand projected by *Traffic Forecast* in such a way that all QoS constraints are honoured for as long as the demands do not exceed their specified bandwidth.
- In addition to the primary objective it is important to maximise the available capacity within the network, while ensuring that this capacity is well balanced across the network so that arising demands can be satisfied without extensive reconfiguration of the link weights. Thus there is a trade off between balancing and optimising free capacity.

Copyright © MESCAL Consortium, June 2005

10.6.2.3.3 Routing Planes

As outlined in the introduction, the constraint on hop counts introduces a limitation of the OSPF based traffic engineering. The correlation of traffic on the same route. The problem is that all traffic flowing from one ingress point to the same egress point, must follow the same route through the network. It follows that the hop count for all this traffic has to be the same (with the exception of ECMP [RFC2328], which will be discussed later). If any of the traffic trunks on this source destination pair have a strict hop-count-limit, all others have to be treated in the same way. Hence a premium service is provided to some traffic whether it has been paid for or not. From a traffic engineering perspective, this means that if many source-destination pairs are part of the same traffic trunk with strict limits, all of the traffic on the network effectively becomes premium traffic and the choices of performing traffic engineering become limited. A further complication to this problem becomes apparent when taking into account that all traffic from any ingress to the same egress might merge into an aggregate flow somewhere in the network. If this is the case, then hop count constraints from all these ingress nodes to the same egress node has to be taken into account when modifying the route of such an aggregate flow.

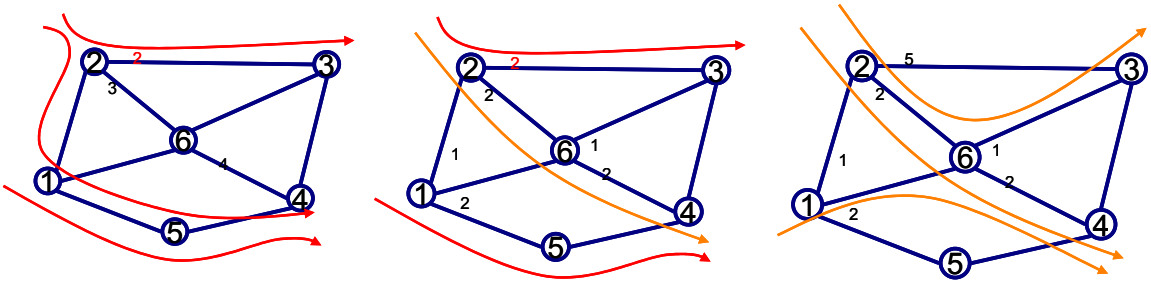


Figure 116. Routing Planes

In order to avoid this problem, routing planes are introduced as an addition to OSPF routing (illustrated in Figure 116). This is essentially (Type of Service) TOS based routing and is done by marking traffic with similar hop count constraint characteristics (l-QC) with the same DSCP on ingress. As shown in Figure 116, each DSCP marked traffic aggregate is routed individually, by a different instance of OSPF. Several instances of OSPF are operating in parallel this way, each with an independent set of link weights (small numbers next to nodes).

As a result of this approach, it is now possible to route traffic with different hop count and other QoS constraints independently from one another. It is proposed that this will provide the flexibility required to allow the link weight optimisation algorithm to find efficient solutions, but will have to be proven through simulation.

10.6.2.3.4 Link Weight Computation Algorithm Outline

This section gives a quick introduction to the basic steps of the link weight computation algorithm, the activity diagram in Figure 117 illustrates the process. As each item is described in more detail later on in this chapter, more complexity is introduced and the order of the steps is changed. At this point, the emphasis is placed on clarity.

Copyright © MESCAL Consortium, June 2005

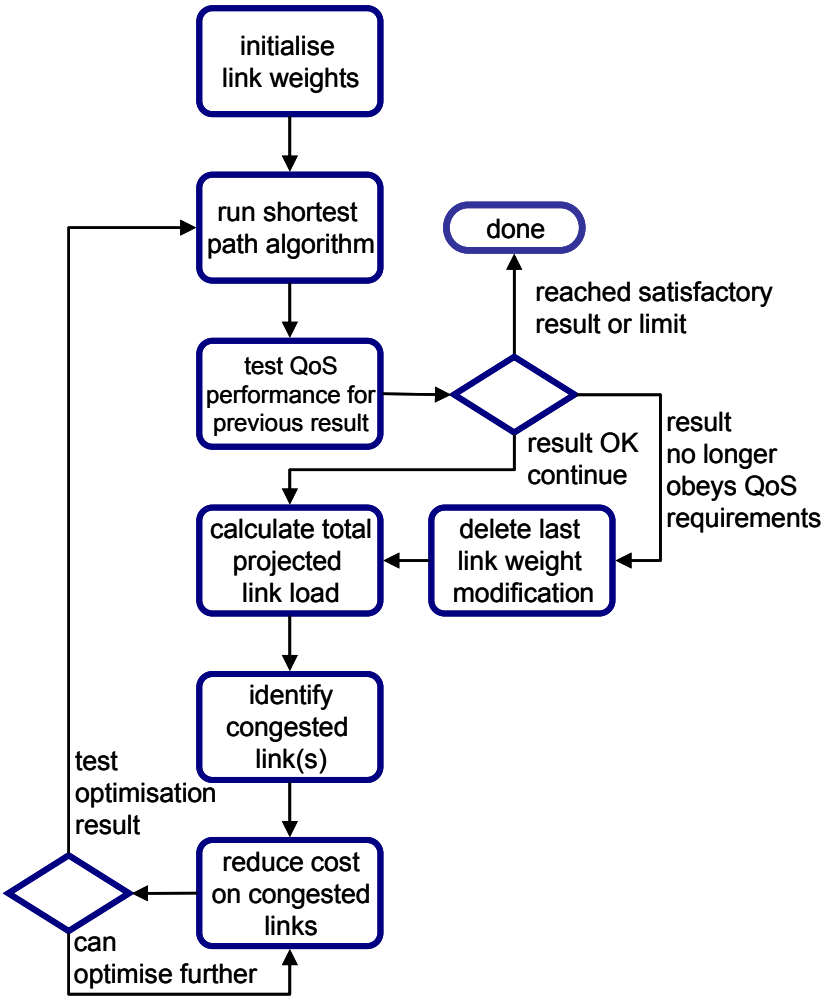


Figure 117. Link Weight Optimisation Flow Chart

1. **Initialise the network** by setting some arbitrary (e.g. unit value) weights for each link and each DSCP routing plane. This defines the initial condition of the weight setting algorithm.
2. **Calculate the shortest path tree** for each ingress-egress pair that has a traffic load. This has to be repeated for each DSCP routing plane. Check the hop count limits that each routing plane has to obey.
3. **Evaluate the total link load** for all links and hence identify the most congested link. Each DSCP routing plane has to be taken into account individually, although the total link load is a single integer figure representing total effective link load across all planes. Traffic for each PHB constitutes differently to the effective link load.
4. **Identify high cost links.** This step involves taking the effective link loads and computing the cost function. The cost function defines on what the optimisation algorithm places most emphasis. For example, high cost links can be those that are congested, those that have high delay or those that are highly under-utilised.
5. **Reduce the load on the most congested link.** This is the most crucial task of the algorithm and its many options and tradeoffs will be discussed in detail.
6. **Re-compute steps 1 to 5** and test to see if step 4 was successful. If unsuccessful, ban the solution and compute steps 1 to 4 again, if successful recomputed steps 1 to 4 until no more improvement can be achieved or until the computation is stopped.

Copyright © MESCAL Consortium, June 2005

10.6.2.3.5 Routing Model Definitions

In order to define a weight manipulation algorithm and test its performance, it is essential to set up some definitions. As in [RFC2702] the network is modelled as a directed graph, $G = (N, E)$ where the nodes $n \in N$ and links $l \in E$ represent routers and links between routers. A link *l* has capacity *c(l)* indicating the amount of traffic that *l* can accommodate. The traffic is given in form of a demand matrix *D*, representing the amount of traffic flowing on a path between any nodes (s, d). This demand matrix is provided by traffic forecast, which in turn is calculated based on historic data and intra/inter-domain traffic demands. It can be expected that many of the demands (s, d) are zero, as not every ingress/egress pair has traffic flowing between it. So the routing problem is to distribute the traffic from non zero $D(s, d)$ across the network evenly. The load on a link is given as x_l, this is the sum of all demands $D(s, d)$ using the link *l*. The utilisation of *l* is then given by $x_l / c(l)$. For the weight allocation, a weight ω is assigned to each DiffServ Code Point $h \in H_l$ on each *l*, so that $\omega_{h,l}$ denotes a unique weight.

10.6.2.3.6 QoS Constraints

The network has to meet the QoS constraints of each flow. Constraints of delay, jitter and packet loss probability are imposed on each Per Hop Behaviour and are the same throughout a single routing plane.

- In order to ensure an upper limit on queuing delay, a maximum queue length has to be defined per PHB. The queue length limit is enforced by dropping excess packets and so both queuing delay and the arising jitter can be enforced by imposing a hop count constraint**.**
- Similarly, it can be demonstrated (e.g. in [TEQUI,D1.4]) that packet loss probability and the end-to-end delay can be seen as a hop count constraint problem, if some simplifying assumptions are made and link loads do not exceed the planned amounts.
- As the queue length has an upper bound, packet loss has to be accommodated through distribution of link loads and prevention of congested links with the link weight calculation algorithm. This can only be ensured if the traffic demands do not exceed the ones specified in the *Traffic Forecast* that was used for the link weight calculation. Some degree of excess link load may be tackled by over specifying the demands for the link weight calculation. However, an admission control policy is unavoidable for quality of service enforcement, to police that the traffic does not exceed the planned capabilities of the network.
- Since hop count constraints do not currently feature in the routers shortest path algorithms, the hop count limit has to be enforced by the IP Traffic Engineering algorithm. As previously discussed, routing planes ensure that appropriate maximum hop count values apply for all groups of similar traffic (l-QC).

10.6.2.3.7 Initialising the Network

This step is important, because it defines the initial conditions for the weight setting algorithm. Better initial conditions should lead to faster and better convergence. The apparent choices are random, unit or inverse capacity weight settings. In operational networks, inverse capacity weights are often used. Their advantage is that OSPF now automatically favours the larger links towards the core of the network and so inverse capacity link weights are a crude but practical approach to traffic engineering. It has been shown in [FORT02a] that initial weight settings based on inverse link capacity, greatly improve the convergence time of the algorithm. However, for the case of this algorithm, unit weight settings may be more suitable. Although inverse capacity weight settings are tailored towards accelerating the flow of traffic towards the network core, the links delay and congestion characteristics are not taken into account. Additionally, inverse capacity weight settings place a large emphasis on one particular solution or local minimum in the solution map and gives an algorithm less chance of escaping this minimum.

Copyright © MESCAL Consortium, June 2005

10.6.2.3.8 Calculating the Shortest Path Trees

Principally this needs to be done for each load bearing ingress egress tuple. However, because each DSCP routing plane has a unique set of link weights, the shortest path tree has to be calculated for each ingress egress tuple and for each DSCP thus multiplying the routes that have to be calculated by a maximum of 64. It is essential that the algorithm for calculating the path is exactly the same as that of the routers in the network, since all weight setting is based on these paths. After performing this step, hop count constraints have to be confirmed for each *D(s,d)*. If a previous iteration of the algorithm has caused a hop count limit to be exceeded, the solution has to be discarded.

10.6.2.3.9 Evaluating the Total Link Load and Cost

The total link load is calculated by distributing the traffic from each ingress egress tuple over the shortest paths calculated in the previous step. Link loads from each routing plane and the physical capacity of each link are used to give the total free capacity of each link. Each PHB has different delay and loss characteristics, which are expressed by associating an *equivalent* bandwidth to PHBs rather than the physical bandwidth. For each PHB h of a set H_l of PHBs on a link with bandwidth allocation $x_{l,h}$, the equivalent bandwidth can be expressed as a function $f_{l,h}(x_{l,h})$ increasing in $x_{l,h}$ and greater for any given $x_{l,h}$ with higher priority h. The total equivalent load of each link is then $\sum_{h \in H_l} f_{l,h}(x_{l,h})$. Assuming that *c(l)* is not the same for all links,

$$L_e = \sum_{h \in H_l} f_{l,h}(x_{l,h}) \Big/ c(l)$$

is the normalised utilisation of the link.

In order to arrive at an overall cost function, the statement has to be extended to reflect a cost per link which can then be summed over $l \in E$.

$$\Phi = \sum_{l \in E} \Phi_l(L_e) = \sum_{l \in E} \Phi_l \left(\frac{\sum_{h \in H_l} (f_{l,h}(x_{l,h}))}{c(l)} \right)$$

The cost function should be convex and increasing to avoid highly congested links. The actual function Φ_l could be approximating an exponential curve with discrete values as defined in [FORT00]. In addition to equivalent bandwidth, the cost function can also define per class emphasis on links with large available capacity as well as low delay characteristics of the link. This way, special class specific characteristics can be honoured. In order to do this, terms can be added to the $f_{l,h}(x_{l,h})$ for each class that express the cost-behaviour of a class with respect to e.g. spare capacity on the link or transmission delay.

10.6.2.3.10 Reduce the load on the most congested link

In order to minimise the cost function, those links with the largest contribution to the total cost have to be identified. Hence $l_c = \max_{l \in E} \Phi_l(L_e)$ identifies such a link. Before modifying any weights, it is necessary to choose a candidate traffic flow passing though l_c which to modify. Depending on the type of modification and the amount of load, it may be good to choose either a small or a large flow, however, a cost for each sub flow can be calculated to aid the choice. There are several DSCP routing planes to select from. As long as the initial condition of the algorithm is based on a good weight setting, it may be a good option to leave traffic mapped to high priority PHBs and modify first the lower priority traffic. This ensures that high priority traffic stays on the large capacity links and it also guarantees least disruption of this traffic in case of a traffic engineering event such as a Resource

Copyright © MESCAL Consortium, June 2005

Provisioning Cycle (see Section10.6.3.3.1). The high priority traffic will not be as affected if its paths do not change.

The choices made here on DSPC routing plane and size of traffic flow represent a small subset of the available options, detailed simulations of the IP Traffic Engineering system should reveal the most suitable choices.

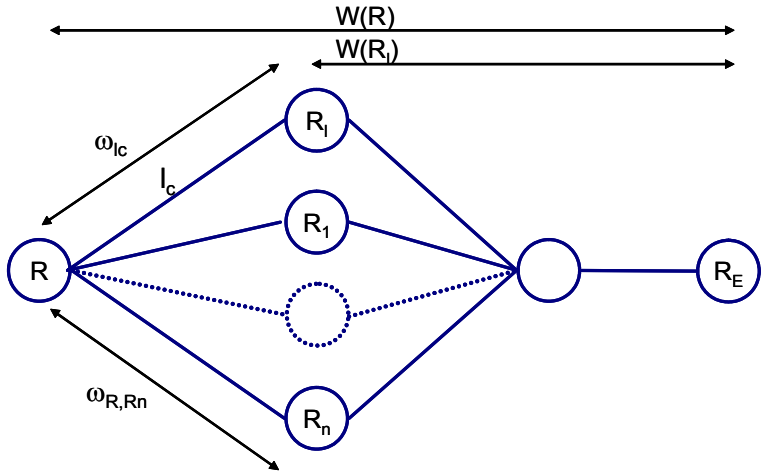


Figure 118. Reducing Loads

As in Figure 118, let R be the origin node of link l_c, and let R_l be the destination node of l_c. Let *W(R)* be the sum of the weights on the shortest path from R to some egress node R_E. Also let R_n be a neighbouring node to R of a set of neighbours, B. Then,

$$W(R) = W(R_l) + \omega_{l,h} \leq W(R_n) + \omega_{(R,R_n),h} \text{ for all } R \in B$$

The objective now is the reduction of the load on l_c, by redistribution of its load onto other neighbouring links.

Modifying a single link weight.

For a single weight modification, the neighbourhood of node R is searched in order to locate a neighbouring node R_n such that

$$W(R_n) + \omega_{R,R_n} < W(R_{l_c}) + \omega_{l_c} + \varphi \text{, where } \varphi \text{ is a weight adjustment}$$

It is sensible to choose the neighbour where φ is the smallest value in all of B. The reason is that the adjustment of a weight may cause other routes on R_n to change. By choosing the smallest weight change, the probability of such unwanted changes is kept low.

Copyright © MESCAL Consortium, June 2005

Balancing the loads.

Large aggregated flows develop naturally in networks with shortest path algorithms, as a result of traffic aggregation from multiple ingress points towards a particular egress point. As soon as these different traffic flows meet on a node, they become a single aggregate flow. A network consisting of such large aggregates is difficult to optimise. However, it is possible to split these flows between neighbouring nodes by making use of equal cost multi path which causes a split of the traffic when more than one shortest path route is available. By setting the *W(R)* equal to some or all of $W(R_n) \in B$ so that all those R_n lie on the shortest path. In order to do this,

$$W(R_1) + \omega_{(R,R_1)} = W(R_2) + \omega_{(R,R_2)} =$$
$$\ldots = W(R_n) + \omega_{(R,R_n)}$$

should all be set equal. It is important that

$$W(R_n) + \omega_{(R,R_n)} \leq W(R_y) + \omega_{(R,R_y)}, \text{ where } R_n \in B, R_y \notin B$$

holds true for this modification, as a reduction in weights may cause loading of a link that is already heavily loaded. The drawback of this approach is that a carefully set up multi path can be damaged when further iterations of the algorithm modify paths nearer the sink of the ECMP traffic. These modifications could cause for some of the paths to be no longer on the shortest path, or for one of them to become shorter than all others and attracting all of the traffic. Care has to be taken in order to prevent these scenarios from occurring by e.g. applying the ECMP rules only in the last iterations, or as a last resort. It could also be conceived of keeping a database with all the ECMP weight changes to remember which links have been modified.

Some undesirable consequences of link weight modifications.

Although modifications made to link weights should, in general, succeed at redistributing the traffic flow in question, they may have undesirable side effects. A full view of changes that a link weight modification has caused, becomes visible after the shortest path algorithm has been run on the new set of weights and costs have been calculated for each link. Two problems may occur:

- Traffic shifted away from one link has caused congestion on a neighbouring link.
- The route modification has caused the hop count constraint of the rerouted traffic to be exceeded. This could be the case for the traffic that was chosen for rerouting and also for other traffic that was rerouted though the weight change.

If neither of the two cases holds, the weight modification can be accepted and the algorithm continues. If a problem occurs, the weight modifications have to be discarded (and also banned from being chosen again) and other, different modifications have to be identified to redistribute the traffic.

10.6.2.3.11 Re-computation of steps 1 to 5

The iteration of the algorithm has two purposes, to check on the effectiveness of the changes made and to continue making changes if necessary. Re-computation should continue until no further improvement can be made or until a certain number of iterations has passed. The number of iterations could also be controlled by a cost function analysing the time taken to achieve further optimisation. A threshold could then be defined to stop the algorithm when improvement is too slow.

10.6.2.3.12 Improving the Heuristics Effectiveness

It is well known that weight optimisation algorithms can be computationally expensive and this dilemma is reflected in several mechanisms to reduce the complexity of computation. Several techniques to reduce computational complexity have been implemented, among them hash tables that record previously encountered weight constellations. At each iteration an entry is created. If at a later

Copyright © MESCAL Consortium, June 2005

time, the same constellation is encountered again, it can be skipped without computation cost. The table is based on an symmetric XOR metric, so that any one change in link weights does not require a recompilation of the whole table.

Another means for saving computation is the evaluation of more than one link weight setting per optimisation cycle, by changing the weights to more than one routing plane simultaneously. This can also be useful in order to displace more traffic of a congested link, if the traffic of one single class is not sufficient to decongest the link.

In order to “kick start” an optimisation that fails to improve through many steps, the link weight set can be perturbed by randomly changing a percentage of the weights on all routing planes. This allows exploration of other regions of search space.

Copyright © MESCAL Consortium, June 2005

10.6.3 Network Reconfiguration Scheduler

10.6.3.1 Objectives

- Dynamic reconfiguration of network configuration to adapt to changes in
  - traffic demand
  - network topology
- Reconfiguration according to schedules
- Effecting the network configuration of the offline Intra-domain Traffic Engineering

10.6.3.2 Interface Specification

Figure 119 enlarges the relevant parts of the architecture, in order to allow a clear definition of interfaces required.

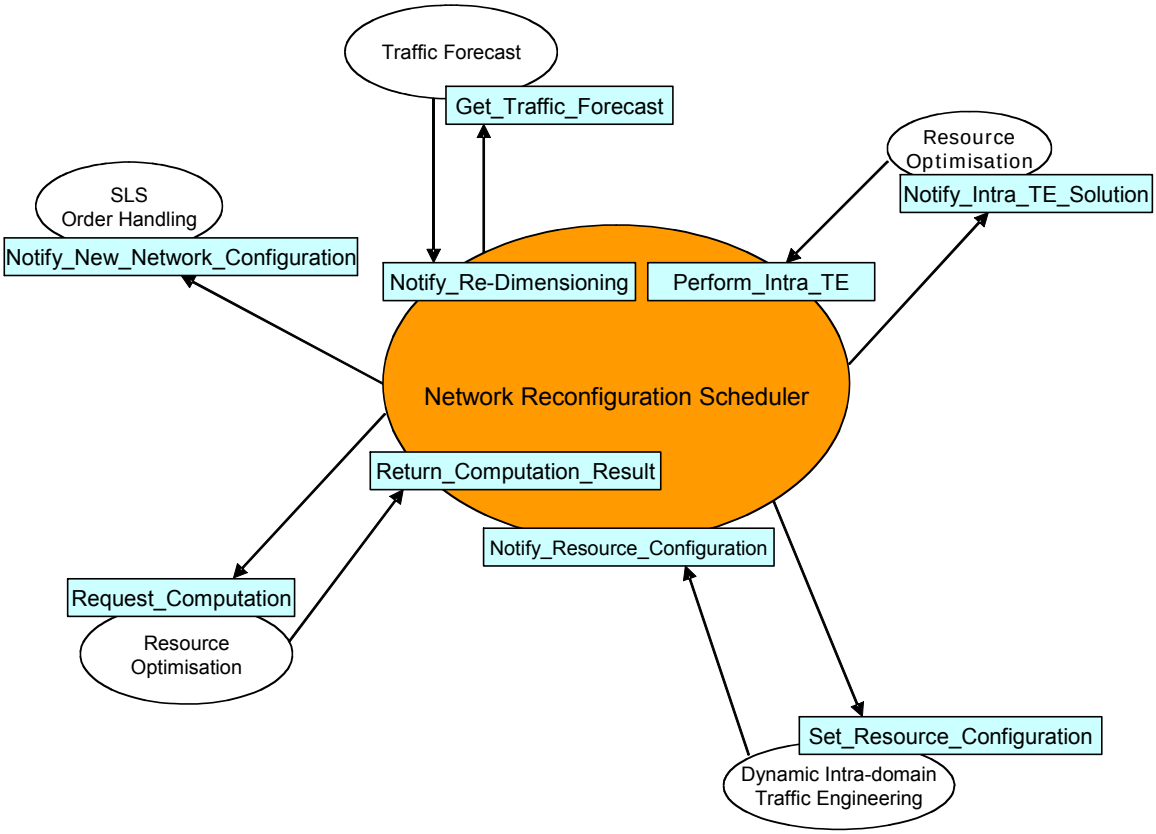


Figure 119. Interactions of the Network Reconfiguration Scheduler Block

- **Traffic Forecast & Network Reconfiguration Scheduler**
  Get_Traffic_Forecast(handle to iTM)

  This method is called by *Resource Reconfiguration Scheduler* in order to request information on where to locate (inside a database) the current traffic forecast. (Note that *Resource Optimisation* will be passed the information so that it can retrieve the forecast for computation.) This method can also be used to retrieve a computed hypothetical iTM' for an inter-domain TE "what if" scenario

Copyright © MESCAL Consortium, June 2005

Notify_Re-Dimensioning()

Traffic Forecast will call this method in order to notify *Resource Reconfiguration Scheduler* that the current network configuration is not satisfactory to accommodate the demands and that a Resource Provisioning Cycle is needed.

- **Network Reconfiguration Scheduler to SLS Order Handling**
  Notify_New_Network_Configuration(iRAM)

  This method will be called by *Resource Reconfiguration Scheduler* in order to notify *SLS Order Handling* that a new network configuration is available.

- **Network Configuration Scheduler & Dynamic Intra-domain Traffic Engineering**
  Notify_Resource_Configuration()

  This method will notify the *Network Reconfiguration Scheduler* that the dynamic resource management was not able to accommodate the demand with its current configuration or that link failures have occurred.

  Set_Resource_Configuration (Link weight matrix ω_o)

  This method will be called by *Network Reconfiguration Scheduler* passing the scheduling information, link weights and ECMP settings to be configured for each router order to affect a change in the network.

- **Inter-domain Resource Optimisation & Network Configuration Scheduler**
  Perform_Intra_TE (handle to iTM')

  This method will be called by inter-domain *Resource Optimisation* to consult *Network Configuration Scheduler* on the resulting intra-domain resource availability and utilisation if a given inter-domain TE solution is activated. For this, a handle to a modified iTM is passed that includes the additional, hypothetical inter-domain demands.

  Notify_Intra_TE_Solution (iRAM', cost Φ, (full intra-domain TE configuration))

  This method will be called by *Network Reconfiguration Scheduler* to return the intra-domain traffic engineering solution or resource availability (intra-domain configuration, e.g. iRAM') to inter-domain *Resource Optimisation*. In addition, a cost for network reconfiguration is passed should the presented solution be implemented. At this stage, it is left for further study to determine whether the iRAM provides sufficient information to the inter-domain TE or whether the full intra-domain configuration needs to be passed.

- **Binding Activation to Network Configuration Scheduler**
  Notify_Inter_TE_Solution (handle to selected solution)

  This method will be called by *Binding Activation* to indicate to off-line intra-domain TE which inter-domain TE solution (i.e. eRAM∈eRAM(s)) has been selected. The purpose of this notification is to enable the *Network Configuration Scheduler* to physically configure the network resources, thereby enabling the selected resource allocation.

- **Resource Optimisation & Network Reconfiguration Scheduler**
  Request_Computation(handle to iTM)

  The Network Reconfiguration Scheduler requests the computation of link weight optimisation for a particular network topology and demand matrix. Location of both topology and demand matrix have to be specified by the Network Reconfiguration Scheduler at the time of the request (locations for these may be databases of *Traffic Forecast* or databases for "what-if" scenarios inside *Offline Intra-domain Traffic Engineering*).

Copyright © MESCAL Consortium, June 2005

Return_Computation_Result()

Once the Resource Optimisation has completed the request, it stores the information in the link weight database and notifies the Network Reconfiguration Scheduler.

10.6.3.3 Algorithm Description

10.6.3.3.1 Modifying link weights in an operational Network

There are problems associated with the modification of link weights in an operational networks. Each single link weight update has to be flooded as a link state advertisement which causes re-computation of the routers forwarding tables. This is a disruptive process with the amount of disorder introduced into the network being a function of the number of modified link weights. It is therefore desirable to modify as few link weights as possible, as few times as possible. An example for a technique to limit the number of weight changes at each iteration is presented in [FORT02a] as an extension to the link weight setting algorithm. The disruption caused by link weight modification, is a deterrent from optimising to frequently and careful considerations have to precede an optimisation decision.

There are two different causes for network optimisation to become necessary: Intra-domain Resource Provisioning Cycles and dynamic events that occur on smaller timescales, within an Intra-domain RPC.

Resource provisioning cycles of the intra-domain and inter-domain TE always coincide and can cause extensive reconfiguration of the network according to long term changes in demand. However, it is also possible for a resource provisioning cycle to cause little reconfiguration in the intra-domain TE, if the demands can still be satisfied with little or no link weight modification. Although a more optimal solution may be found through an extensive reconfiguration, it is not always carried out because of the additional cost of network disruption. Network disruption can therefore be expressed as a threshold cost, causing the network to stay in a suboptimal state until reconfiguration is unavoidable. The value of the threshold is proportional to the size of the network (time to convergence after disruption) and number of link weight modifications necessary (amount of disruption introduced). Finding solutions with as few weight changes as possible is therefore beneficial.

Smaller, dynamic changes in demand that occur within the Intra-domain RPC (as outlined in the next section) have to cause minimal disruption to the network and are based on one or few link weight changes. This method will be applied to optimise for new concentrated demands like pSLS traffic as well as link failures, etc. Fluctuations in daily demand could also be addressed in this way, but it may be beneficial to take these into account by optimising weight settings for multiple demand matrices, because this technique does not require any periodic weight changes. The technique aims at selecting a weight setting solution that is good for a set of demand matrices that reoccur periodically. Further studies on the dynamic events occurring within the RPC, will be studied in more detail in the course of the project.

10.6.3.3.2 Dynamic events

The *Network Configuration Scheduler* is responsible for effecting link weight modifications that have already been computed. A small subset of events that can be foreseen in this way are

- Changes in demand
    - yearly demand changes
        - public holidays
        - new applications
    - weekly demand changes
        - weekends, public holidays
    - daily demand changes
        - office hours, etc

Copyright © MESCAL Consortium, June 2005

- extraordinary events
  - Christmas, new years eve (this does not include unforeseen disaster events such as earthquakes, although some of its implications could be foreseen and effected in an emergency case)

- Changes in Network topology
  - Link failure
  - New Links

These events with their different time scales can be used to produce demand matrices for the *Network Configuration Scheduler* which has two functions.

- Schedule events to be pre-calculated by the offline Intra-domain Traffic Engineering
- Effect link weight changes when they become necessary (RPCs)

Some simple algorithms for the dynamic traffic engineering component are shown below, these are straight forward and do not require further explanation.

```
/*link failure
if link failure or scheduled event signalled do
      check database for pre-computed configuration
            if database entry exists do
                  effect link weight modification
            else
                  call offline intra-TE RPC
                  add resulting configuration to database
            end
end
/*RPC event
if resource provisioning cycle completed in offline intra-domain TE do
      effect link weight modification
end
```

Given time, the *Network Configuration Scheduler* accumulates a large database of ideal network configurations for many scenarios. One could conceive of more sophisticated learning algorithms that accumulate useful network configuration scenarios more quickly. However, all scenarios become outdated when a large change in network topology occurs. While this might only have an effect when high capacity links are added or removed, it means that the *Network Configuration Scheduler* needs to adapt its scenarios if possible and discard them if they cannot be adapted. It may be possible to re-optimise an "old" network configuration by passing it to *Resource Optimisation* together with the new topology. This should consume less time than re-computing from scratch, although this is dependent on the extent of the changes. At this point the problem is left for further study.

10.6.3.3.3 Discussion of the interactions with the Inter-domain Traffic Engineering

The *Inter-domain Resource Optimisation* function block queries the *Resource Configuration Scheduler* for "what-if" scenarios in order to allow a choice of the best pSLS options not only based on inter-domain cost considerations, but also intra-domain resource availability and reconfiguration cost.

Copyright © MESCAL Consortium, June 2005

The *Inter-domain Resource Optimisation* passes information to *Traffic* Forecast that allows it to calculate and pass to intra-domain *Resource Optimisation* a projected iTM' for each "what if" scenario, including the additional inter-domain demands. The return value from the intra-domain *Resource Optimisation* should consist of a cost Φ and optionally an intra-domain configuration (e.g. iRAM'). Two scenarios are possible,

- The *Resource Configuration Scheduler* finds that the projected iTM' passes the threshold for causing a resource provisioning cycle. The intra-domain *Resource Optimisation* is executed, and the resulting cost and iRAM' is passed back to inter-domain *Resource Optimisation*. The cost function, should include the extra cost for necessary network configuration.
- If the intra-domain *Resource Configuration Scheduler* finds that the threshold is not crossed, it will return the iRAM' and the difference in cost between this projected iTM' and the iTM that the last resource provisioning cycle was based upon. Although the cost of all current network disorder is passed this way, it should be significantly less than a case involving a new weight setting from the intra-domain *Resource Optimisation*. This is because of the cost raised for disrupting the network configuration. This low cost raises the question if such "what if" scenarios based on no network reconfiguration should be passed back at all, and if so, if they should be limited to low bandwidth or short term pSLSs. It would be highly undesirable if a suboptimal pSLS is chosen on the grounds of a low intra-domain cost, an advantage that will disappear by the time the next resource provisioning cycle takes place.

Copyright © MESCAL Consortium, June 2005

11 TRAFFIC ENFORCEMENT

This Section presents the interactions and the behaviour of the data plane Functional Blocks (FBs) assumed by the MESCAL system: *Traffic Conditioning & QC Enforcement*, *PHB Enforcement*, and *IP Forwarding*. MPLS forwarding is also considered specifically for the hard guarantees solution option but it is not implemented in the testbed. The interactions of these FBs with the rest of the MESCAL FBs are defined. The behaviour specification of any FB explains the specification of the functions that are essential for the deployment of that specific FB.

11.1 Traffic Conditioning & QC Enforcement

11.1.1 Objectives

With the *Traffic Conditioning and QC Enforcement* function block we mean the process of realising the results of c/pSLS agreements and intra/inter domain TE functions in classifying, conditioning, and QoS class enforcement to traffic streams as appropriate to the o-QC treatment that these streams should receive per domain. These processes takes place at the data-plane after the c/pSLSs have been established, activated, and invoked. They are realised by downloading appropriate information for setting up the traffic classification and conditioning mechanisms to the DiffServ-capable routers. QC-signalling is performed across all domains using DSCPs. Traffic conditioning & QC enforcement must be performed:

At Network Edges (customer ingress point): Customer traffic at the edge routers should be classified in order to capture and reflect the negotiated *cSLSs*. In addition, suitable traffic profiles derived from the negotiated *cSLSs* should be enforced on the classified traffic before it actually enters the provider's network. The *SLS Invocation Handling* function block calculates the appropriate traffic classification and conditioning configuration parameters and downloads them to the Network Elements (NE) via the *Traffic Conditioning* function block. This block should also provide capabilities for statistics information retrieval with respect to the traffic classification and conditioning results.

The Dynamic inter-domain TE functional block provides configurations to the *Traffic Conditioning and QC Enforcement* function block for configuring the ingress NE to possibly perform DSCP re-marking for realising an intra-domain TE solution.

At the domain boundaries (ASBRs): Peer provider traffic at the border routers should be re-marked to the appropriate DSCP depending on the l-QC treatment it receives in the domain. The traffic should be classified in order to capture and reflect the negotiated *pSLSs* for traffic conditioning. In addition suitable traffic profiles derived from the negotiated *pSLSs* should be enforced on the classified traffic before it actually enters the domain. The *SLS Invocation Handling* function block calculates the appropriate traffic classification and conditioning configuration parameters and downloads them to the ASBRs via the *Traffic Conditioning* function block.

The *Dynamic Intra-domain TE* function block provides configurations to the *Traffic Conditioning and QC Enforcement* function block for configuring the egress ASBR to perform DSCP remarking for realising an inter-domain TE solution.

11.1.2 Interface Specification

11.1.2.1 Traffic Conditioning & QC Enforcement Interface to SLS Invocation Handling

The *Traffic Conditioning & QC Enforcement* (TC-QC) interface to *SLS Invocation* use basic concepts specified and documented in [RFC3290]. The model of different successive traffic conditioning elements contained in traffic conditioning blocks is adopted [RFC3289]. The output of each TC-QC element should be associated with the input of its subsequent element, which could be another traffic

Copyright © MESCAL Consortium, June 2005

conditioning element. This way a full sequence of successive elements inside a TC-QC can be specified.

The *Traffic Conditioning & QC Enforcement* interface to *SLS Invocation* is defined as follows:

TC-QC_NewTC (TC-QC_{ID}, Interface, Direction)

Creates a new TC-QC to be applied in the ingress or the egress *Direction* of the specified *Interface*.

TC-QC_Classifier (ClassID, TC-QC_{ID})

Creates a new classifier.

TC-QC_Filter (FilterID, ClassID, Precedence, Type, Parameters)

Creates a new filter as part of the classifier *ClassID*. The filter *Type* (BA, MF, other) with related *Parameters* Specific to *Type*. The *FilterID* will be applied to the *ClassID* with the *Precedence*.

TC-QC_Meter (MeterID, Type, Parameters)

Creates a new meter with meter *Type* that specifies an Average Rate, EWMA (Exponential Weighted Moving Average), Token Bucket, or other. The associated *Parameters* (e.g., average time interval) are specific to *Type.*

TC-QC_Marker (MarkerID, Type, Parameters)

Creates a new marker. The marker *Type* is DSCP/EXP and the related *Parameters* are specific to *Type*.

TC-QC_Shaper (ShaperID, Profile Parameters, Buffer Size)

Creates a new shaper. The *Profile Parameters* describe the profile the traffic is shaped to and the *Buffer Size* specifies the maximum queue length.

TC-QC_Dropper (DropperID, Type, Parameters)

Creates a new dropper. The dropper *Type* (WRED, other) and the related *Parameters* are *Type* specific.

TC-QC_GetStats (ElementID)

Returns statistics for *ElementID*. The type of statistics depends on the TC-QC element.

TC-QC_DeleteElement (ElementID)

Deletes the element identified by *ElementID*.

11.1.2.2 Traffic Conditioning & QC Enforcement Interface to Dynamic Inter-and Intra- domain TE

The *Dynamic Inter-domain TE* decision is enforced through configuration of the *Traffic Conditioning and QC Enforcement* function block, by configuring the ingress/egress ASBR to perform DSCP remarking. This is based on the fact that the TC-QC has been already created and TC-QC_Marker is called.

A similar interface as above is used for TC-QCs at the ingress point of a domain.

Copyright © MESCAL Consortium, June 2005

11.2 PHB Enforcement

This *PHB Enforcement* function block represents the required queuing and scheduling mechanisms for realising different PHBs associated with NE interfaces with appropriate configuration as determined by the related TE blocks. This block is responsible for implementing the mechanisms needed to provide differential forwarding treatments to traffic passing through the NEs based on the DiffServ specifications. The *PHB Enforcement* block manipulates the NE's native scheduling and queuing management mechanisms in order to enforce the parameters and the bandwidth and buffer sharing rules and policies, defined by the Dynamic Intra/Inter TE functional blocks, and hence satisfy the requirements in terms of throughput, delay, jitter and loss. *PHB Enforcement* should provide capabilities for PHB selection, PHB prioritisation, bandwidth and buffer resources allocation and excess resources sharing rules.

11.2.1 Interface Specification

Dynamic Inter-domain TE provides information for PHBs associated to the egress ASBR for realising inter-domain TE solution while *Dynamic Intra-domain TE* provides information for PHBs associated to the NEs within the domain for realising intra-domain TE solution. The interfaces of *PHB Enforcement* function block to these two function blocks are specified as below.

11.2.1.1 PHB Enforcement Interface to Dynamic Inter-domain Traffic Engineering

The following interface is used as *PHB Enforcement* interface to *Dynamic Inter-domain TE* for providing information to be used in ASBR type of NEs.

PHBEnf_NewPHB (PHBID, Interface ID, Name, Scheduling Class, Priority)

Creates a new PHB for an associated *Name* (e.g., EF) to be enforced in the interface identified by the *Interface ID*. The *Scheduling Class* parameter defines the scheduling class the PHB belongs to. The *Priority* parameter determines the priority with which the packets of this PHB will be served given that the resources allocated to each PHB are properly granted.

PHBEnf_MapDSCP (PHBID, DSCP)

Maps the *DSCP(s)* to the PHB identified by the *PHBID*. The packets marked with *DSCP* will be serviced by the *PHBID*.

PHBEnf_AllocateResources (PHBID, [Reserved Bandwidth], [Reserved Buffer], [Excess Bandwidth], [Excess Buffer], [Average Time, List of {Threshold, Dropping Probability}])

Allocates the scheduling resources to the *PHBID*. *Reserved Bandwidth/buffer:* the minimum amount of bandwidth/buffer is allocated to *PHBID*. *Excess Bandwidth/buffer:* the excess bandwidth/buffer can be used by PHBID in case of other PHBs' temporal under-use. *Average Time:* the time period over which the average queue size is calculated. A list of *Threshold* and *Dropping Probability* pairs determines the algorithmic dropping behaviour applied to *PHBID* for each virtual queue (i.e., WFQ, CBWFQ).

PHBEnf_ActivatePHB (PHBID)

Activate the *PHBID*. This activation results in downloading the PHB configuration parameters to the NE and enforcing the PHB.

PHBEnf_DeactivatePHB (PHBID)

Deactivates the *PHBID*. This deactivation results in releasing the PHBs allocated resources. These resources are still considered unavailable when the resources are checked for allocating to other PHBs.

Copyright © MESCAL Consortium, June 2005

PHBEnf_DelPHB (PHBID)

Deletes the *PHBID*. The allocated resources of *PHBID* are freed.

PHBEnf_GetStatistics (PHBID)

Returns the packets and bytes serviced by the *PHBID*, as well as the dropped packets and bytes for each defined *Threshold* and *Dropping Probability* pair.

A Similar interface is used as *PHB Enforcement* interface to *Dynamic Intra-domain TE* for providing information to be used in NEs within a network domain.

11.2.2 Behavioural Specification

11.2.2.1 Description of Functions

PHB Enforcement performs the four functions namely as Configuration, Verification, Enforcement, and Statistics. The Configuration function maintains the configuration and state information for each defined PHB. It is triggered based on the specific request for a configuration change. According to this request and the state of the PHB, it triggers the Verification and the Enforcement functions. The *PHB Enforcement* function block should check the requested resources against the available resources in the NE before granting resources to a PHB. The Verification function of PHB FB is triggered by the Configuration function. It calculates the available resources. It decides and its output is directed to the Configuration function as grant or rejection of requested resources for the PHB. *PHB Enforcement* should download the related information to the scheduling and buffer management mechanisms of the NE. An Enforcement function performs this task. When a configuration is successfully downloaded to the NE, the Enforcement function replies to the Configuration function, otherwise an enforcement error will occur. The Statistics function is for calculating statistics per PHB. This function inquires the NE, gathers the necessary information and then calculates the packets & bytes serviced and the packets & bytes dropped by the specific PHB.

11.3 IP Forwarding

Both IGP and EGP protocols for routing purposes must be QC-aware. Routing protocol normally provide information for packet forwarding by taking into account the packet's associated l-QC. At the edge of autonomous domains, the *Traffic Conditioning & QC Enforcement* FB re-marks the packet's DSCP to an appropriate value with regard to the l-QC specified for the traffic stream.

The objective of the *IP Forwarding* function block is to maintain the Forwarding Information Base (FIB) of the router. A FIB stores all the routes, which have been selected/installed by the IP Routing processes that have been activated in the router and according to the results of each route calculation that has been launched by the activation of dynamic routing protocols. It is important to note that, in any case, the route selection is a decision which is made by Dynamic Intra/Inter-domain TE blocks. This decision is based on the q-BGP information received, *pSLS* agreements, the QoS requirements that have been expressed by the appropriate parameters values in each SLS, that being processed by the MESCAL system.

It is possible for a router to keep many FIB tables, for example within the context of Meta-QoS-class deployment where there may be one FIB per l-QC that belongs to an m-QC.

It should be noted that In testbed configuration and for simplicity, Intra-domain QoS-based routing protocols are kept out of scope. Instead static rout are employed.

11.3.1 Interface Specification

Dynamic Inter-domain TE provides routing information for the egress ASBR for realising inter-domain TE paths while Dynamic Intra-domain TE provides routing information for the NEs within the domain for realising intra-domain TE paths. The interfaces of IP forwarding to these two function blocks are specified as below.

Copyright © MESCAL Consortium, June 2005

11.3.1.1 IP Forwarding Interface to Dynamic Inter- and Intra- domain TE

The following interface is used as the *IP Forwarding* interface to *Dynamic Inter-domain TE* for installing route-related information to ASBR type of NEs.

IP-FIB_CreateRouteEntry (Destination Prefix, DSCP, Next Hop Interface)

Creates a route entry where:

The *Destination Prefix* determines the destination to be reached using this route entry.

The *DSCP* is the DSCP value of the packets that use this route.

The *Next Hop Interface* determines the outgoing interface the packets will be forwarded for reaching the destination.

IP-FIB_DeleteRouteEntry (Destination Prefix, DSCP)

Deletes the route entry identified by both the *Destination Prefix* and *DSCP*.

IP-FIB_GetRouteEntryStats (Destination Prefix, DSCP)

Returns the number of bytes and the packets transmitted with regard to the route entry identified by *Destination Prefix*, *DSCP*.

IP-FIB_GetOutputInterfaceStats (Interface)

Returns the number of bytes and the packets received (and destined) by the output *Interface*.

Similar interface is used as *IP Forwarding* interface to *Dynamic Intra-domain TE* for installing route-related information to NEs within a network domain.

11.4 MPLS forwarding

The objective of the Data Plane block with respect to MPLS is to forward packets on LSPs. Dynamic Intra/Inter-domain TE must pass down the loose LSP path in order for head-end NE to request for the establishment of the LSP tunnels across the networks. The label distribution protocol should cross the boundary of domains for setting-up LSP end-to-end. The loose path used by a given tunnel at any point in time is already determined based on tunnel resource requirements and network resources such as bandwidth negotiated through *pSLSs* between domains. A packet crossing the MPLS-enabled network travels on a single tunnel that connects the ingress to the egress points across multiple domains.

Customer traffic at the LSP head-end should be classified in order to capture and reflect the negotiated *cSLSs*. The traffic profiles derived from the negotiated *cSLSs* should be enforced on the classified traffic before it actually enters the LSP. The *SLS Invocation Handling* function block should download appropriate traffic classification and conditioning configuration to ingress NE via the *Traffic Conditioning & QC Enforcement* function block. MPLS EXP field at the edge of the network can be set and change the ASBRs based on c/pSLS agreements.

Copyright © MESCAL Consortium, June 2005

12 INTER-DOMAIN MONITORING

12.1 Objective of Monitoring

Quality of service monitoring is becoming crucial to service providers and IP Network Providers (INP) for providing QoS-based services and service assurance and for managing network resources at both intra- and inter-domain levels. Intra-domain measurements are performed in a single/multiple autonomous system where monitoring and measurement processes and realisation are under the control of a single administration. Inter-domain measurements are performed between two domains, which may not be under the same administrative authority.

In addition, users require network performance statistics, as network performance has a direct impact on the perceived quality of the application viewed by the users. The performance requirements of a customer's service is described in an SLA and consequently its SLS part [GODE02a]. Two types of SLS are distinguished, customer SLSs (cSLSs), and peer SLSs (pSLSs) [D1.2]. pSLSs support aggregated traffic serving many customers. Different cSLSs will have different QoS requirements as they depend on the type of services offered.

A QoS-based monitoring system should provide information for end-to-end service assurance and resource management in:

1. Assisting network providers to verify whether the QoS performance guarantees committed in c/pSLSs are in fact being met. In-service verification of traffic and performance characteristics per service type may be required.
2. Assisting network providers to make provisioning decisions for optimising the usage of network resources (both at intra and inter-domain levels) according to short to medium term changes, as well as providing measurement information for long-term planning. For this, it is necessary to perform QoS-based resource monitoring at traffic class, node, path, and network levels.
3. Assisting customers to verify the fulfilment of their subscribed cSLSs.

12.1.1 Background and Related work

There are a number of working groups in the Internet Engineering Task Force (IETF) related to measurements and monitoring such as Remote MONitoring (RMON), IP Performance Metrics (IPPM), Real-Time Flow Measurement (RTFM), IP Flow Information Export (IPFIX), and Packet Sampling (PSAMP) [PSAMP]. These working groups are defining metrics, developing a common IP traffic flow measurement technology, and specifying a standard set of capabilities for sampling packets through statistical and other methods respectively.

There are numerous monitoring tools, such as the RIPE Test Traffic Measurement (TTM) [TTM], NetFlow from Cisco, SFlow, NIMI (National Internet Measurement Infrastructure) [PAXS98], Network Analysis Infrastructure (NAI) [NLANR], cflowd, RTG high-performance SNMP statistics monitoring system, Skitter, NeTraMet, CoralReef, and Beluga of CAIDA (Cooperative Association for Internet Data Analysis) [CAIDA], and so on.

There has been some work at the intra-domain level to use measurement information for tackling network performance degradation and managing congestion in operational networks as well as addressing service level monitoring, for example NetSope [FELD00], Rondo [ALBE91], and others. These measurement tools and systems collect, analyse and visualise forms of Internet or Intranet traffic data such as network topology, traffic load, performance, and routing. An intra-domain QoS monitoring system was developed in the TEQUILA project for IP-based networks offering IP connectivity services and under the control of a single network provider [ASGA04], [ASGA03].

There has also been some work on monitoring and measurements at inter-domain level, by European research projects [IST]. The objective of the IST-INTERMON project has been to develop an integrated inter-domain QoS monitoring, analysis and modelling system to be used in the multi-domain Internet infrastructure for the purpose of planning, operational control and optimisation [ELIS04], [HOFM04]. The proposed solution assumes that a centralised manager negotiates monitoring operations with each domain along the service delivery path. This results in a scalability problem for the INTERMON system as the inter-domain network expands. The focus of the IST-MoMe project has been the enhancement of inter-domain real-time QoS architectures with integrated

Copyright © MESCAL Consortium, June 2005

monitoring and measurement capabilities. The objective of the IST-SCAMPI project was to develop an open and extensible network monitoring architecture for the Internet including a passive monitoring adapter at 10 Gbps speeds, and other measurement tools to be used for denial-of-service detection, SLS auditing, quality-of-service, traffic engineering, traffic analysis, billing and accounting [COPP04]. IST-LOBSTER is a follow on project to SCAMPI aimed at deploying an advanced pilot European Internet Traffic Monitoring Infrastructure based on passive monitoring sensors at speeds from 2.5Gbps and possibly up to 10Gbps [COUT00]. The IST- 6QM [IST] is working towards measurement technologies for Quality of Service in IPv6 networks by developing a system with the required functions for QoS measurement, such as packet capturing, precise time-stamping, data collection, QoS metrics derivation and result presentation.

12.2 Monitoring in Multi-domain environment

Different networks have different monitoring architectures/tools. Currently, every provider monitors its own domain and decides what to measure, how to conduct measurements, and what to do with its measurement data in terms of processing, analysis and visualisation. For inter-domain and inter-provider QoS monitoring, a number of aspects must be considered. One important aspect is the co-operation of providers in the service delivery chain. It is essential for providers to co-operate based on an agreed common framework formulating some technical aspects including the metrics to be measured, the configuration of monitoring elements and service, the execution of measurements, the composition of results in an appropriate way, and the exchange of measurement data between providers. These pre-requisite conditions for performing any inter-domain monitoring and measurements are listed and explained briefly below: It should be noted that some business related aspects should also be considered in the framework including verification of metrics, resolution of disputes/violation, etc.

Co-operation of providers: Each provider performs monitoring in its domain. Monitoring operations could also be conducted by a third party that exchanges measurement information with interested providers. For end-to-end performance/traffic monitoring, providers must share measurement information. The monitoring information should be available to other domains. Therefore, monitoring in multi-domain environment requires co-operation of network providers. This co-operation must be formalised somehow through a negotiation process between providers resulting in an agreement. This agreement should set a common measurement framework based on the measurement needs and how to conduct the measurements and what to measure.

Metrics to measure: Both performance and traffic related metrics should be considered in the framework. Relevant performance metrics in IP networks include one-way delay, packet delay variation, one-way packet loss and traffic related measurements including traffic load and throughput. Measurement types and metrics are discussed in section 3.

Configuration of monitoring service: The common measurement framework must specify how to conduct the monitoring operation through a monitoring specification, which gives the guidelines for configuring monitoring elements and services. The monitoring elements must be configured appropriately based on the monitoring specification. Through the agreement, providers should agree on how to conduct the measurements in terms of data format, form of data collection, sampling periods, monitoring frequency, synthetic packet sizes, etc. Providers should also agree on the type of measurements they perform such as periodic measurements, random measurements, alarms, etc.

Composition of results: The monitoring agreement must specify how to compose or compare the measurement results as providers may have different measurements tools. Through the agreement, providers should agree on how to aggregate measurements, how to concatenate measurements, how to summarise the result and what should be the summarisation periods, reporting schedule, etc. The objective is to provide a reliable, consistent and accurate view across different domains. A common approach to reduce the monitoring overhead should also be used. Processing, aggregating, sampling, or filtering the raw data into accurate and reliable statistics and reducing the amount of data near to the observation point (source) are key functions to scalable dynamic measurement operations. This should minimise the amount of information exchange between domains.

Copyright © MESCAL Consortium, June 2005

Exchange of monitoring data: The monitoring data should be shared through a communication protocol or other mechanism. Data must be exchanged reliably and securely, because without ensured privacy no provider would agree to share its data. .
It should be noted that the framework must take the following factors into account:

- Network Providers own and administer their own IP networking infrastructures. The framework should not require any direct access to probes/test agents in another provider's domain.
- Providers do not allow other providers to control their network elements .
- Framework should only specify the required information to exchange between domains, because Providers tend to disclose as little domain-internal information as possible.
- Framework should not over-specify the tests and mandate very specific test requirements.

12.3 Measurement types and metrics

A monitoring system should offer measurement granularity levels from macro-flow, at QoS class levels, to micro-flows. Monitoring and measurements can be performed at flow level, interfaces links, nodes, node-pairs, or QoS-based routes. The MESCAL QoS classes are defined for different solution options in order to offer quantitative (hard) guarantee, qualitative (statistical) guarantee, or very limited/no (loose) guarantee. Since monitoring is a costly task, it is reasonable that the loose guarantee services are not monitored with the same granularity as hard or statistical guaranteed services.
In general, the passive metrics that should be measured by the monitoring infrastructure are as follows:

- Offered load per cSLS at ingress point of source domain and throughput per cSLS at egress point of destination domain. This is done for service monitoring purposes.
- Offered load at ingress border router of a domain and throughput at egress border router of same domain per pSLS/QoS class. This is done for service monitoring purposes.
- PHB throughput at edge/core/border routers for traffic engineering purposes.
- PHB packet discards in packets per second.
- If MPLS technology is used, LSP offered load at the head-end of the tunnel and LSP throughput at the tail-end of the tunnel.

The active metrics should be measured by the monitoring system are as follows for both service monitoring and traffic engineering purposes:

- One-way packet delay of QoS-based path either end-to-end or per AS hop, LSP, and PHB.
- One-way packet loss at QoS-based path either end-to-end or per AS hop, LSP, and PHB.
- Delay variations experienced at QoS-based path either end-to-end or per AS hop, LSP, and PHB levels.

12.4 Monitoring system architecture

This section describes a large-scale inter-domain QoS monitoring framework for use in a multi-domain and multi provider IP-based networking environment. This framework specifies three types of QoS monitoring components operating at different levels i.e., node, network and service levels. Any QoS-based inter-domain monitoring infrastructure should actually allow co-operation between different providers while maintaining the authority, confidentiality, and full control of each provider over its underlying resources.

12.4.1 Monitoring System Components

A Monitoring infrastructure should have three distinct entities in order to fulfil the stated requirements. These entities are namely Node, Network, and Service Level Monitors.
Node Monitors perform node-related measurements. They should be able to perform active measurements between one node and any other node in the network at path or per hop level as well as passive monitoring. A Node Monitor collects measurement results from either meters or probes located at routers through passive or active monitoring agents. Node Monitors are configured with information about the variable to be monitored, the sampling/summarisation periods and the threshold parameters.
Network Monitor performs domain-wide post-processing of measurement data using a library of statistical functions. This monitoring sub-system utilises network-wide performance and traffic

Copyright © MESCAL Consortium, June 2005

measurements collected by all the Node Monitors in order to build a physical and logical network view at intra-domain level (i.e., the view of the routes that have been established over the network). It also collects network-wide monitoring information regarding the INP's inter-domain links and resources.
Service Monitor performs customer/provider related service level monitoring, auditing, and reporting. Each NP can have a Service Monitor for its own purposes. Service Monitor must keep track of the compliance of the level of service provided to the customer SLS instances. It utilises information provided by the node and network monitoring entities of the networks involved in the end-to-end chain of QoS delivery.
The Service Monitor can be in charge of exchanging configuration information, signalling messages and monitoring information between different domains that share monitoring results by means of a protocol or other means.
A proposal for signalling for QoS measurement in multi-domain environment is described in [17].

12.4.2 QoS Peering Models and Monitoring System

There are many models for the interconnection of providers' to offer global QoS services. The type of inter-domain peering impacts the inter-domain monitoring functions. The proposed QoS monitoring system functions are further explored below for two QoS peering models: namely source-based and cascaded models.

12.4.2.1 Monitoring System in the Source-based[4] Model

In the *source-based* model, an INP establishes QoS peering agreements directly with a number of downstream providers to offer an inter-domain QoS service. This is achieved by making peering agreements with a chain of INPs so as to create a service within a scope beyond its boundaries. In this model, the INP is the central point which takes the responsibility for the overall service management including service monitoring of any given customer end-to-end service instance.
In the source-based model, one INP as the central entity establishes business relationships with all AS domains to be involved in the service delivery chain. Therefore, it is possible for the source INP to get monitoring information directly from each domain. Hence, as pSLSs along the chain are explicitly established by the source INP, it is able to locate the domain/s which are violating the service agreement, if an end-to-end service violation occurs. Figure 120 shows the monitoring system infrastructure that can operate where the source-based model is employed for QoS peering.
In this model, a hop-by-hop method, at the AS level, can be used for calculating end-to-end performance results. In hop-by-hop method, monitoring between two edge nodes of a domain for specific pSLS or QoS class can be carried in order to determine the performance status of the domain. The source INP, then uses these per AS measurements to calculate the end-to-end result.

[4] Known as Centralised Model in D1.4

Copyright © MESCAL Consortium, June 2005

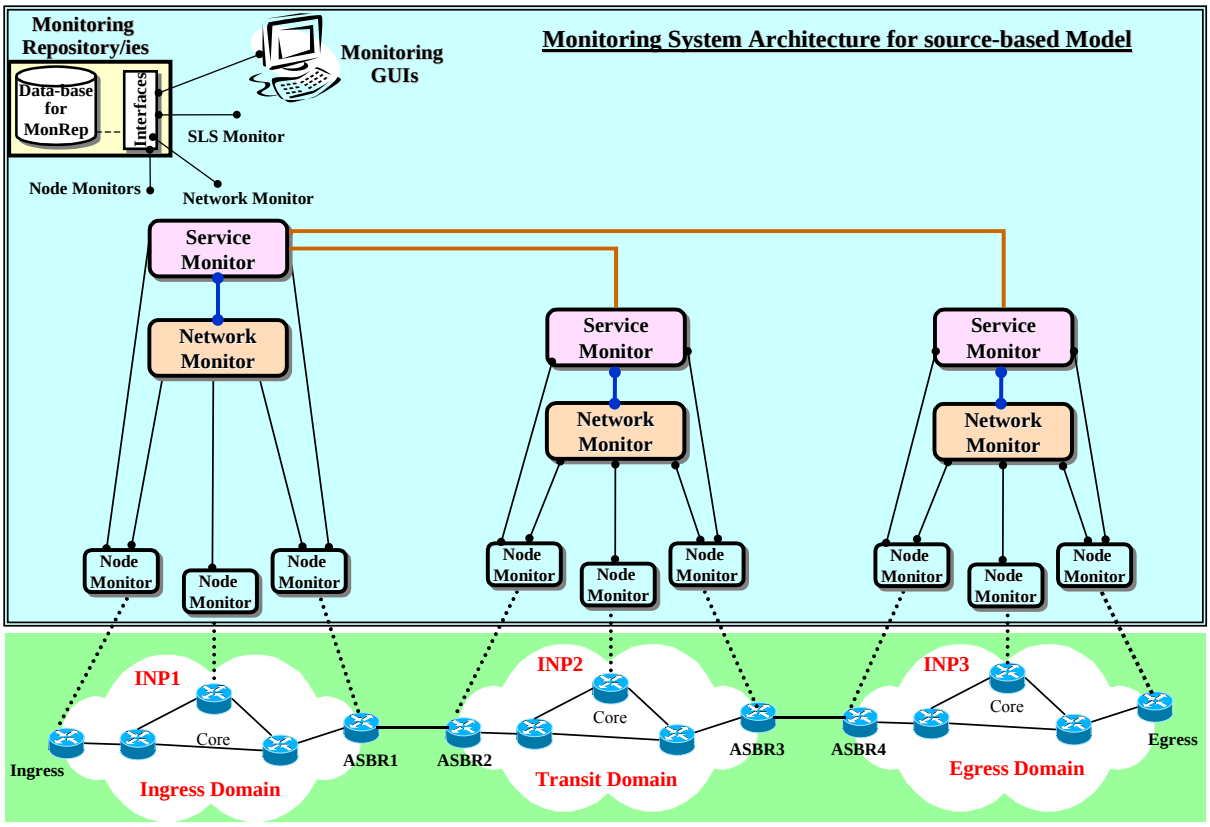


Figure 120: Monitoring system architecture for the source-based model.

Node Monitors are configured to perform the measurements on PHBs and QoS-enabled paths. Paths allow control over routing of traffic flows requiring specific QoS within a domain. IP engineered paths (i.e., IP routes, LSPs) are used to carry aggregate user traffic belonging to several SLSs with similar performance requirements. Measurements can be from the PHBs at the border routers for inter-domain or edge/core routers for intra-domain resource provisioning and traffic engineering purposes. In addition, performance measurements can be carried out per AS hop i.e., between the ingress border router of a domain to the egress border router of the domain for a specific QoS route. End-to-end performance measurements between ingress and egress points of two remote domains (ingress at AS1 and egress at AS3 as shown in Figure 120) can be carried out for specific QoS routes. Passive monitoring (offered load and throughput) for cSLS can be performed at the ingress and egress points. Passive monitoring for pSLS can be performed at the border routers (e.g., ASBR2 for offered load and ASBR3 for throughput as shown in Figure 120).

The Network monitor in each domain can deduce per AS/domain performance view by analysing PHB and route related measurements. PHB QoS performance measurements can be used for managing inter-domain links and the link buffer space. The intra/inter-domain Traffic Engineering sub-system can use the measured performance of the various routes in order to do route management, load balancing, and dimensioning.

The Service Monitor uses the measurement data, which is collected by Network Monitors and Node Monitors, and composes the data for its domain and will pass the related information to the source INP if it is already configured to do so. This information includes path level performance related measurements and SLS specific traffic related statistics. Since each service type has specific requirements, different metrics may need to be measured for each service type.

The source INP knows the end-to-end routes (i.e., transit AS hops). pSLS Monitoring can be performed in source-based model for the sake of source INP. The Service Monitor can request retrieval of the required monitoring information from the Network Monitor and Node Monitors in the border nodes as there exists business relationship between the source INP and each INP along the end-

Copyright © MESCAL Consortium, June 2005

to-end path.
There are a number of factors to be considered for the monitoring system when a source-based QoS peering model is used:

- The source INP needs the topological information at large scale in order to perform end-to-end monitoring operation.
- The source INP must initiate the monitoring operation and it will receive per AS performance measurement results.
- The source INP composes per AS performance results obtained from the INPs in the end-to-end chain in order to built an end-to-end view.
- As source INP knows the destination INP, it can easily obtain the throughput measurements from destination INP directly as there is a direct business relationship between the two.
- Source INP can also initiate end-to-end performance measurements between its ingress node and the destination INP egress node as it is capable of receiving the measurement information from the destination domain directly.
- pSLSs are explicitly established for the source INP along the delivery chain. If an end-to-end service violation occurs, it is possible for the source INP to identify the domain/s, which are not providing the promised service (pSLS).

12.4.2.2 Monitoring System in the Cascaded Model

In the *cascaded* model, an INP only establishes QoS peering agreements with its immediate neighbouring provider/s to construct an end-to-end QoS service. Figure 121 shows the monitoring system infrastructure for the cascaded model. In the cascaded peering model, the business relationship is between the source INP and its immediate adjacent INPs. There is no peering relationship between source INP and transit INPs.

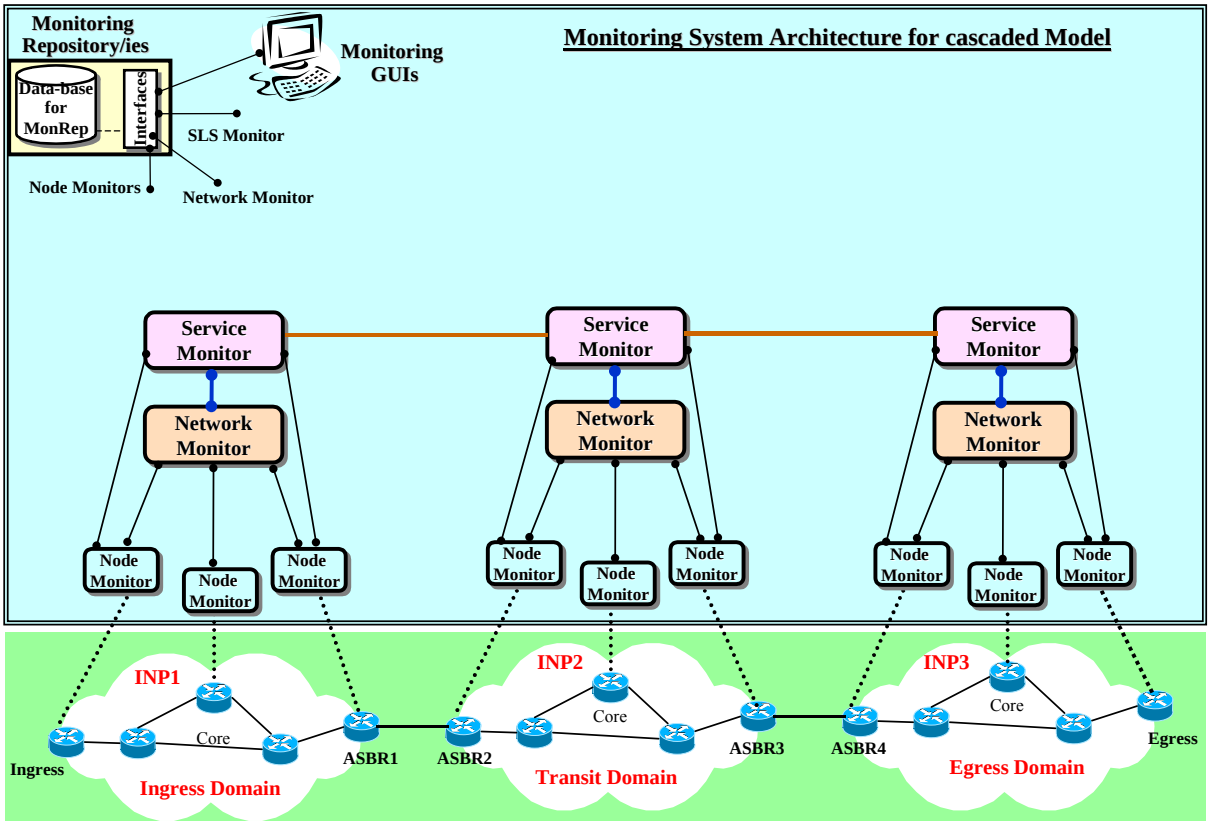


Figure 121: Monitoring system architecture for the cascaded model.

The cascaded model is adapted as the preferred model in MESCAL, as it reflects the loosely coupled

Copyright © MESCAL Consortium, June 2005

structure of Internet. In MESCAL, three solution options are defined offering loose, statistical and hard QoS guarantees. The following section describes the implications of each solution option on monitoring.

In loose QoS guarantee, the Open Scope Cascaded Model is used for QoS peering. This model relies on the cascaded model and the use of the meta-qos-class (*m-QC)* concept. Setting-up *pSLS*s with open scope (i.e., no explicit destination address/reachability information) and no distinct performance characteristics but simple compliance with well-known m-QC behaviours between adjacent INPs is the defining feature of this model. In this model, there are no end-to-end QoS guarantees defined and consequently there is no need to build *e-QC*s, which are the fundamental differences between this model and 'strict scope cascaded model'. This model does not provide any end-to-end bandwidth guarantees because it enables any destination to be reached, without prior explicit indication in the *pSLS*. Each domain is engineered to support a number of local QoS classes (i.e. *l-QC*s). These *l-QC*s are mapped to globally well-known *m-QC*s. Each AS advertises the *m-QC*s that it supports in its administrative domain. Other domains can make *pSLS* arrangement in cascaded fashion with this domain to make use of offered *m-QC*s. Although, inter-domain routing is *pSLS* constrained, each domain can find out whether it can reach certain destinations in an *m-QC* plane through a BGP-like protocol (q-BGP) [D1.2].

With no end-to end scope at the time of pSLS set-up in cascaded fashion, the types of measurements, which are feasible to be carried out are as follows:

- Offered load at ingress border router of a domain and throughput at egress border router of the domain per pSLS/QoS-enabled route. This information is passed to the adjacent upstream domain with which this INP has pSLS agreements.
- Performance/traffic related measurements per AS hop at pSLS/ l-QC (at path level) granularity.

These measurements can be propagated through a measurement protocol or via q-BGP to upstream domains to build end-to-end views on m-QC enabled routes and services.

It may not be possible to have end-to-end traffic related measurements as there is no end-to-end scope and bandwidth guarantees and there are business agreements only between adjacent neighbours, but not with providers more than "one AS away".

In statistical QoS guarantee, the Strict Scope Cascaded Model is used for QoS peering. Each INP makes *pSLS* contracts with the immediately adjacent INPs. Thus, the QoS peering agreements are between adjacent neighbours, but not with providers more than "one AS away". This type of peering agreement provides the QoS connectivity from a customer to reachable destinations that may be several domains away. Setting-up *pSLS*s with defined scope and distinct performance characteristics between adjacent INPs is the compelling feature of this model. Each INP in the chain needs to know its adjacent neighbours and the status of related interconnection links. In addition, each INP needs to know the *e-QC*s advertised by its neighbouring domains for binding with its own *l-QC*s in order to implement its own *e-QC*s, which will be subsequently advertised to its customers and upstream domains.

The monitoring system operation is effected by a number of factors as below:

- The source INP does not need the topological information at large scale in order to launch end-to-end monitoring operation. This makes the monitoring system more scalable.
- The source INP must initiate the monitoring operation and it will receive measurement results only from its adjacent domain/s.
- The end-to-end route, in all probability, may not be known to the source INP in order for it to build a complete view of the transit route, as the use of q-BGP is not essential for the inter-domain network operation when this type of peering is applied. Only the last INP in the chain (egress AS) may be known to the source INP (strict cascaded approach) and it can establish business relationship with that INP for monitoring purposes only. This relationship may be required for cSLS throughput monitoring at the INP premises. Any monitoring information from other domains will come to the source INP through the transit domains.
- Each transit INP must compose the results obtained from its adjacent INP. Source INP can then build an end-to-end view.
- pSLS throughput monitoring can only be performed at the egress point of downstream domain. As the pSLSs are merged, it will not be possible to perform pSLS throughput monitoring at end-to-

Copyright © MESCAL Consortium, June 2005

end level. If any service degradation (c/pSLSs) is detected, it will be difficult to locate exactly the domain responsible for the service violation. Any violation has to be directed to the next INP domain in the chain by the upstream INP.

Node Monitors initiate the measurements and collects information on PHBs and QoS routes. This can be from the PHBs at the border routers for inter-domain resource provisioning/allocation purposes. In addition, it can be performance-related information per AS i.e., between ingress border router of a domain to the ingress border router of its immediate downstream domain for specific QoS route. Passive monitoring (offered load and throughput) for cSLS can be performed at the ingress and egress points. Passive monitoring for pSLS can be performed at the border routers (e.g., ASBR2 for offered load and ASBR3 for throughput as shown in Figure 121).

The Network monitor at each domain can compose and deduce an end-to-end performance view by analysing PHBs and routes related measurements and measurement information received from downstream domains.

The Service Monitor uses the measurement data that is collected by the Network and Node Monitors, and combines the data i.e., path level performance related measurements and SLS specific traffic related statistics. This information is sent upstream to the adjacent domain/s upon request of the upstream domain either for the INP to construct end-to-end view for its own use or combine with its own measurement data and pass it up for the upstream domain/s utilisation.

In hard QoS guarantee, the Open Scope Cascaded Model is used for QoS peering. Instead of IP routes, MPLS tunnels are constructed to carry QoS-based traffic. As the tunnel request carries both the head-end and tail-end of the LSP tunnel and Path Computation Elements (PCEs) are used to compute the tunnel path, it is possible to locate the destination INP and establish direct contractual agreements for monitoring purposes. The types of measurements to be carried out are as follows:

- LSP offered load at the head-end of the tunnel by the source INP and LSP throughput at the tail-end of the tunnel by the destination INP.
- One-way packet delay, packet loss and delay variations per LSP.

Copyright © MESCAL Consortium, June 2005

13 MULTICAST

13.1 Multicast cSLS/pSLS

13.1.1 Introduction

Multicast cSLS/pSLS (mcSLS/mpSLS) are customer/provider service level specifications regarding multicast services with an ISP. Due to the inherent distinguishes in service model from unicast, the definition and specification of multicast cSLS/pSLS should also be different. The objectives of defining mcSLS/mpSLS are summarised as follows:

- From customer's viewpoint (mcSLS):
  - To allow for multicast end users (i.e., group members) express their individual QoS requirements in receiving multicast traffic from sources located in local/remote domains,
  - To set up an agreement on the maximum volume of multicast traffic each end user is allowed to receive. That means, the bandwidth consumption of each group that is subscribed by the customer should not exceed the upper limit that is specified in the mcSLS with the ISP. This implies that the bandwidth negotiation is based on per (S, G) group.
- From ISP's viewpoint (mpSLS):
  - To specify the QoS requirement (e.g., scope of domain level reachability for individual QoS classes) expressed to the upstream ISP in terms of multicast flows it is going to receive. Based on this type of mpSLS, the downstream ISP is able to set up mcSLS with its own multicast customers as well as to offer further mpSLS with other directly peering ISPs who request multicast transit services from it. Through this type of cascaded manner of mpSLS ordering/handling, QoS aware multicast service can be deployed globally,
  - To set up an agreement on the maximum volume of aggregated multicast traffic the requesting peer is allowed to receive. That means, the total bandwidth consumption of the multicast traffic by the requesting ISP should not exceed the upper limit that is specified in the SLS with its upstream peer.

Apart from the above objectives, the target mcSLS/mpSLS design should also take the some scalability considerations. Given a specific remote source S, multicast members attached to an ISP can express QC requirements on any group session provided by S. Therefore, the total number of mcSLS can be n times that of the corresponding unicast services between the two parties, where n is the number of groups rooted at S. This scalability issue should be taken into account especially when Source Specific Multicast (SSM) service model is adopted.

13.1.2 McSLS/mpSLS Specification

This document addresses the multicast pSLS (mpSLS) specification, and the related issues, e.g., business interactions. We also discuss the set-up procedure of mpSLS across multiple ASes, together with the guidelines for the corresponding mq-BGP configuration.

According to the business model defined in the MESCAL project, it is always the case that data source pays for the traffic it injects into the network. This implies that multicast group members should be charged by the source (i.e., multicast content provider), and this charge is split into two parts: the charge for the data content of the source, and the traffic delivery from the source to the group member within the network. The multicast source should then pay the related ISP the latter part of the cost for the data treatment. The business interactions among the three parties can be illustrated in Figure 122. The arrows with solid line in the figure point out the direction of payment between the business parties. According to this business relationship, multicast pSLSes should be constructed from the

Copyright © MESCAL Consortium, June 2005

source and span to all the ASes with potentially interested receivers. In this scenario, the mpSLS ordering and its associated handling are from the traffic sender to the receivers. However, since multicast services are always receiver oriented (i.e., multicast trees are usually constructed by individual group joins), the actual invocation of this type of mpSLS will have to be receiver initiated accordingly. The detailed description of this mpSLS ordering/handling will be presented in the upcoming sections. Note that the business contract of the receiver with the Content Provider is out of the scope of this document.

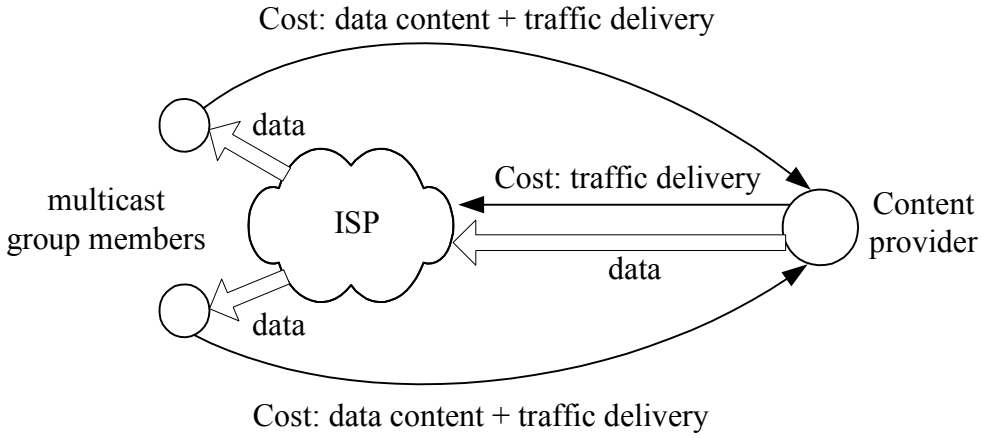


Figure 122 Business relationship in multicast services

13.1.2.1 Attributes

A valid mpSLS between adjacent ASes should have the following components:

1. Service Description: same with its counterpart in conventional pSLS;
2. Connectivity Type: same with its counterpart in conventional pSLS;
3. Inter-connection Point: same with its counterpart in conventional pSLS;
4. Source address: the address of the multicast source, which is not contained in the conventional pSLS for unicast traffic. It is used during the interactions between mpSLS and mq-BGP for proper setting of NEXT_HOP information at mq-BGP speakers (specified later);
5. Destination Nets: The address prefix of ASes where potential receivers are attached. This attribute also tells the scope of the multicast service the ISP is going to provide.

13.1.2.1.1 MpSLS Ordering and Order Handling

When a content provider wants to deploy inter-domain multicast services, it should first decide the geographical scope of the provided service. This procedure can be achieved through some out-of-band mechanism of contacting the first hop AS of potential receivers before any mpSLS ordering. During this negotiation, the first hop ASes of potential receivers may ask for some QoS requirements on the multicast traffic their attached customers are going to receive. Thereafter, when the eQC is built up based on mpSLS ordering, it will be associated with the address prefix of the ASes attached with these potential receivers (i.e., Destination Net attribute). At the same time, the eQC construction will also take the QoS requirements from the boundary ASes into account, and this is based on the previous out-of-band negotiation with s. Similar to the unicast scenario, the mpSLS ordering/handling also take place in a cascaded style from the boundary of the multicast service (i.e., the first hop AS of receivers) back to the source's AS. As the ordering of the mpSLS proceeds, the corresponding eQCs also expand hop by hop at the AS level from the multicast service boundary ASes and finally converge at the

Copyright © MESCAL Consortium, June 2005

source AS. As a result, the hosts located in any domain belonging to this AS-level eQC tree are able to subscribe to the multicast service provided by the source s. In this situation, any intermediate ASes (e.g., AS2) can also offer the multicast service to their local customers without offline contact with s in the first place. However, these internal ISPs are not able to place any specific request on the QoS for their local receivers.

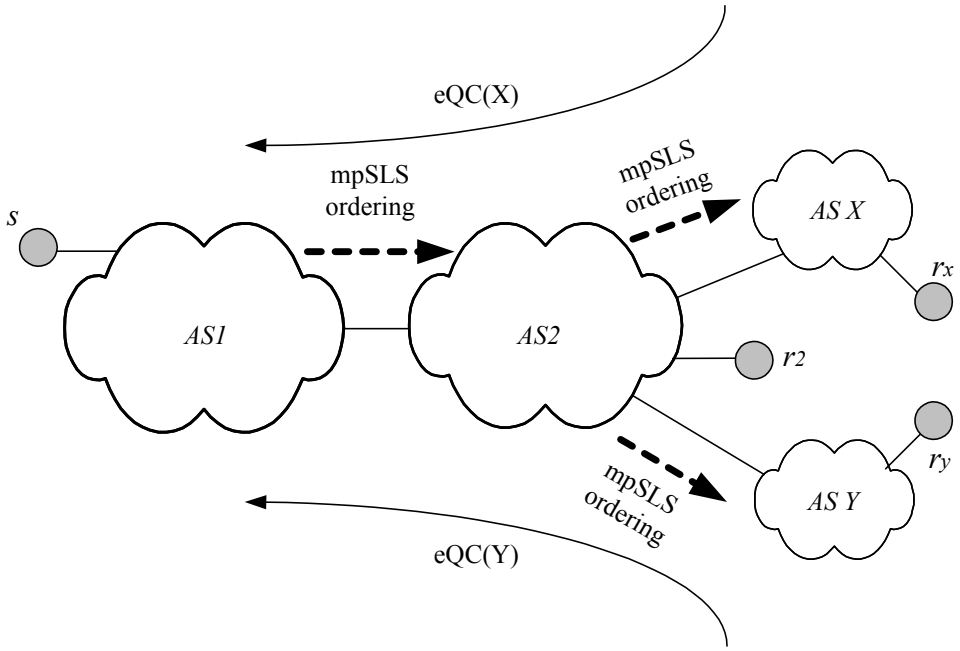


Figure 123 mpSLS ordering

13.1.2.1.2 MpSLS invocation

Due to the receiver initiated service model of SSM, mpSLS will be invocated from the receiver side accordingly. This means that mpSLS are triggered only when an inter-domain group join request takes place within any on-tree (in terms of mpSLS) domain. It should be noted that the group join packet travels along the AS-level path of the mpSLS chain in the direction from the receiver side back to the source s. Therefore the triggered multicast traffic is guaranteed to flow back to the receiver's domain with desired eQC. We still take Fig. 2 as an example for illustration. We assume that host r_x subscribes to the group rooted at the source s by sending an inter-domain group join. This join packet will travel along the same AS path of eQC(X), and within each AS (e.g., AS2), the join packet will follow the directed intra-domain path of engineered lQC that is bound to eQC(X). In this scenario, multicast traffic flowing back to r_x from s will be treated with eQC(X) on the multicast tree constructed by the join packet. On the other hand, for any intermediate AS in the eQC path, its local hosts (e.g., r_2 in AS2) are also allowed to join the multicast group with some QoS treatment. If AS2 is not willing to utilise these "bypass" eQCs, it can contact the source and let AS1 order a brand new mpSLS, such that a dedicated eQC2 will be built between AS1 and AS2 for its own customers.

During the phase of mpSLS invocation, it is an issue how to guarantee that join packets are delivered along the reversed path such that the corresponding multicast traffic can be treated with the correct inter-domain eQC and intra-domain lQC. To solve this problem, it is required that the configuration for NEXT_HOP of individual routers (including core routers and border routers) should be in a completely reversed direction against unicast traffic. Detailed description on this configuration will be presented in the next section.

Copyright © MESCAL Consortium, June 2005

13.1.2.2 MPSLS-MQ-BGP/M-ISIS interactions

During the phase of mpSLS ordering and handling, it is important to specify how mq-BGP routers are configured according to the QoS requirements of the mpSLS. Thereafter when the mpSLS invocation takes place, it can be guaranteed that the packet of join request is able to follow the engineered path such that the multicast flow can be served with proper eQC in the reserved direction. In the following two sub-sections, we discuss how mq-BGP and M-ISIS speakers are configured inter- and intra-domain wide according to the result of mpSLS ordering.

13.1.2.2.1 MpSLS/mq-BGP interaction

The most distinguished difference between uncast and multicast services in this aspect is how to configure q-BGP speakers for the ordered pSLS. Since the actual multicast forwarding path is decided by the join request from receiver side back to the source, it is required that the NEXT_HOP information back to the source should be correctly configured for delivering join requests according to the ordered mpSLS. In Figure 124 we assume AS1 is the upstream of AS2 in terms of multicast traffic (not join request!). When AS1 is ordering an mpSLS from AS2 based on the eQC advertised from remote multicast receiver prefix, it should decide which egress router is used to bind its lQC to the eQC. We assume that it wants to use intra-domain path A→B for multicast traffic delivery, then the NEXT_HOP attribute for the source s in router B should be configured to be router A. It should be noted that this configuration is in the reversed direction of unicast scenario (NEXT_HOP in A for the destination prefix points to B). Meanwhile, AS1 should additionally instruct AS2 on its own mq-BGP NEXT_HOP configuration. In Fig. 3, the NEXT_HOP for s in router F ("ingress router" for the join request") in AS2 should point to B in AS1. This configuration of router F is the task of mpSLS order handling in the downstream AS2. As a result, when the mpSLS is invocated, the join request towards the source s is able to follow the path F→D→B→A. It is worth mentioning that this type of NEXT_HOP configuration is completely mpSLS driven, and it is different from the traditional case where NEXT_HOP configuration is based on the advertised NLRI for destination prefixes.

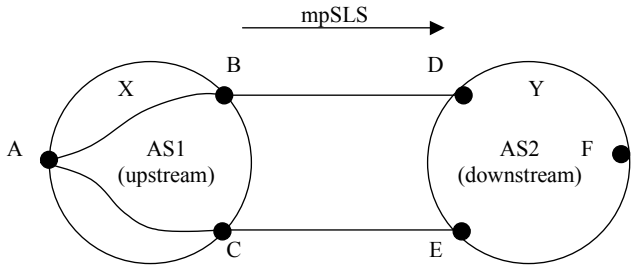


Figure 124 Two adjacent autonomous systems

13.1.2.2.2 MpSLS/M-ISIS interaction

The interaction between mpSLS and intra-domain routing configuration is similar to its inter-domain counterpart. We take Multi-topology extension to ISIS (M-ISIS) that can support dedicated multicast routing as an example. The key idea is that the link weight should be properly set for delivering the join request edge-to-edge, so that the multicast traffic in the reversed direction can be later on treated with the engineered lQC that is bound to external eQC according to the ordered mpSLS. In Fig. 3, the weight for all the links in AS1 should be configured for allowing the join request to be delivered from B to A, while the explored path used for multicast traffic delivery from A to B is able to conform to the engineered lQC.

Copyright © MESCAL Consortium, June 2005

13.2 Offline Intra-domain Multicast Traffic Engineering (OMTE-Intra)

13.2.1 Introduction

The objectives of Traffic Engineering (TE) may include: (1) efficient network dimensioning so that customer traffic demands can be satisfied while also keeping bandwidth consumption to a minimum; (2) control of traffic routes for achieving overall load balancing in the network. For multicast flows, bandwidth conservation is regarded as one of the most important tasks in traffic engineering, and this problem can be formulated into the directed Steiner tree problem, which is NP-Complete. If additional end-to-end QoS constraints and other TE objectives as load-balancing capability are embedded into the objective, the problem becomes even more complicated.

Despite the progress achieved for unicast services, traffic engineering for multicast services remains largely a dark area, especially in the IP layer. Recent research works have focused on Multi-Protocol Label Switching (MPLS) based online multicast traffic engineering, with the purpose of minimising multicast flow interferences [KODIA03]. Scalability becomes an issue if MPLS explicit routing is adopted for multicast traffic engineering, given that a huge number of labels could be consumed for tree maintenance. Despite some research efforts on aggregating multicast traffic for reducing group states at the expense of extra bandwidth consumption [FEI01], mature solutions are still missing nowadays. On the other hand, pure IP, i.e. hop-by-hop routing approaches, present the following difficulties for multicast traffic engineering. First, the PIM-SM protocol uses the underlying unicast routing table for the construction of receiver-initiated multicast trees, and hence it is difficult to decouple multicast traffic engineering from its unicast counterpart. Bandwidth optimisation for multicast traffic can be formulated as the directed Steiner tree problem, which is NP-complete. The enforcement of Steiner trees can be achieved through packet encapsulation and explicit routing mechanisms such as MPLS. However, this approach lacks support from IP layer protocols, such as PIM-SM, due to RPF in the underlying multicast routing protocols. In PIM-SM, if multicast packets are not received on the shortest path with which unicast traffic is delivered back to the source, they are discarded for avoiding traffic loops. Given the inherently difference in shape between the shortest path tree used by PIM-SM and the optimised Steiner tree, the engineered multicast traffic for bandwidth optimisation through Steiner trees could result in RPF check failure.

The MESCAL solution will basically consider IP based multicast traffic engineering mechanisms including bandwidth consumption as well as load balancing. Also, the relevant task is decomposed into intra- and inter-domain parts. Since the proposed solution will be based on the existing routing protocols such as PIM-SM/MBGP, special considerations should be taken on RPF checking failure. Finally, given the fact that common network links are usually shared by both unicast and multicast traffic, it is desirable to provide a unified traffic engineering mechanism for the two type of services simultaneously, and this aspect will also be investigated in the project.

13.2.2 Interface Specifications

Next_Hop_Update(OMTE-INTRA to DMR)

The function parameters provided from OMTE-INTRA to DMR are listed as follows:

- Source address prefix;
- Ingress router address for inter-domain group join;
- NEXT_HOP router address for intra-domain group join;
- Bandwidth availability on each intra- and inter-domain link;
- Reachability information on local and remote source prefixes.

Copyright © MESCAL Consortium, June 2005

- Configured QoS parameters such as delay, delay variation, jitter, loss probability etc.

13.2.3 Behavioural Specification

As we have mentioned, the basic task of OMTE-INTRA is to optimise multicast traffic with QoS guarantees such as bandwidth and delay constraints that have been agreed in mSLSes. As far as intra-domain multicast services are concerned, traffic engineering also involves minimising overall bandwidth consumption within the network. Here we propose an IP layer TE approach for achieving this objective. Issues of inter-domain multicast TE will be included in our future research work. For intra-domain OMTE-INTRA, in order to support dedicated mechanism for multicast traffic engineering, we base our approach on the Multi-topology ISIS (M-ISIS) routing protocol, which enables independent routing for unicast and multicast traffic.

13.2.3.1 M-ISIS based multicast TE

The conventional OSPF and IS-IS protocols only have a mono-viewpoint of the weight of each link in the network, and this influences path selections for both unicast and multicast traffic. In contrast, M-ISIS provides the original IS-IS protocol with the additional ability of viewing the weight of each link independently for different IP topologies. According to [PRZYG03], M-ISIS can support up to 128 different IP topologies. For multicast traffic, the Multi Topology identifier (MT-ID) of value 3 in M-ISIS is currently dedicated to the multicast RPF topology, i.e., the RPF table for PIM-SM can be populated using a set of independent link weights with MT-ID equal to 3. With this multi-topology capability of viewing link weights, it becomes possible that PIM-SM based multicast routing is completely decoupled from the underlying routing table for unicast traffic.

Figure 125 illustrates the basic framework of OMTE-INTRA through optimised M-ISIS link weight setting. In a similar fashion to the unicast scenario, the network topology and the forecasted traffic demand from each multicast group are obtained as the input parameters for calculating the optimised link weights. After the link weights are computed through offline algorithms, they are configured in the network that runs the M-ISIS routing protocol with MT-ID equal to 3, which is dedicated to the multicast RPF table construction. Subsequently, each M-ISIS aware router computes shortest path trees according to this set of link weights and decides the NEXT_HOP router for a specific IP address/prefix. This type of NEXT_HOP information populates the multicast RPF table. When a PIM-SM join request is received, the router simply looks up the RPF table and finds the proper NEXT_HOP for forwarding the packet. In addition, the multicast forwarding information base (FIB) is dynamically updated for the incoming interface (iif) and outgoing interface (oif) list of each group.

Finally, it is worth mentioning that the M-ISIS link weight configuration can be performed on per QC basis, so that the NEXT_HOP to a specific address/prefix can be different in individual QC tree constructions. This makes it possible that PIM-SM join requests for different QCs are able to follow different paths when they are delivered towards the same source. Detailed description will be presented in the DMR section.

Copyright © MESCAL Consortium, June 2005

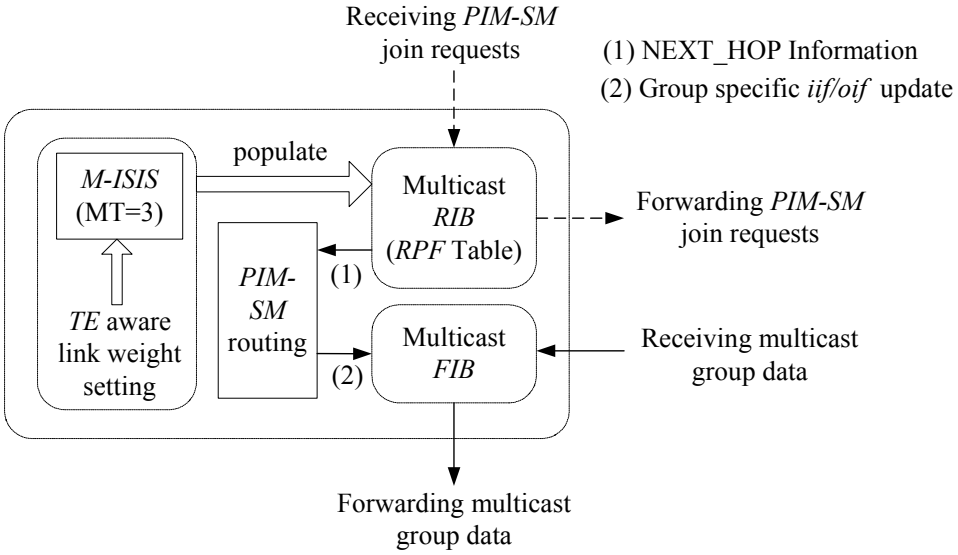


Figure 125 M-ISIS based multicast traffic engineering

13.2.3.2 The OMTE-INTRA Algorithm

The objective of QoS-aware multicast TE can be formulated into the constrained Steiner tree problem. In the IP-layer based approach (i.e., hop-by-hop routing), the enforcement of engineered PIM-SM path selections is via setting proper link weights for the underlying unicast routing protocols. In our proposed approach, PIM-SM follows the shortest path based on the manually-set link weights of M-ISIS, whereas the resulting multicast tree is in effect a constrained Steiner tree with minimum number of links involved. This implies that minimum bandwidth resources are consumed, subject to the end-to-end QoS requirement. In Figure 126, we illustrate a simple instance of the M-ISIS based multicast TE without considering any QoS requirements. We assume that node A is the RP of group X that contains member nodes E, F and G. If PIM-SM is routing in terms of hop-counts, the total bandwidth consumed is 6 units assuming that every link has equal cost 1, as shown on Figure 126(a). If we set the link weight for the underlying unicast routing protocol according to Figure 126(b), only 4 units of bandwidth will be consumed using PIM-SM shortest path routing. In the rest of this section, we will take bandwidth capacity as a typical constraint to the original multicast TE problem, but we also provide a generalised template for satisfying additional QoS requirements.

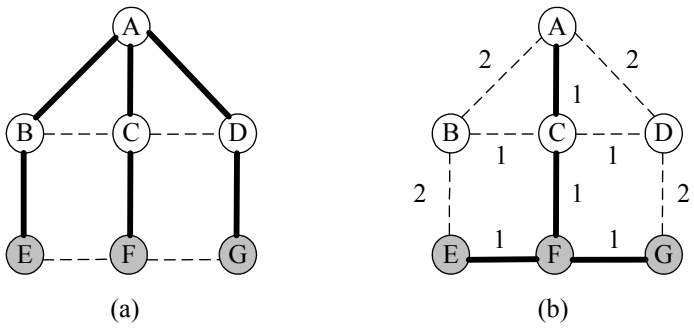


Figure 126 OMTE-INTRA by setting M-ISIS link weights

Copyright © MESCAL Consortium, June 2005

13.2.3.2.1 Problem Formulation

The following is the integer-programming formulation for computing hop-count Steiner trees for each multicast group with the objective of minimising overall bandwidth consumption. By setting the group-specific binary variables $x_{ij}^{g,k}$ and y_{ij}^{g} for each link (i,j), a set of optimised multicast trees with minimum number of links is obtained, which implies that minimum bandwidth consumption is achieved. We first present some definitions below:

G — Total number of active multicast groups;

r_g — Root node (RP) of group g;

V_g — Multicast member (receiver) set for group g;

D_g — Bandwidth demand for group g traffic on each link;

C_{ij} — Bandwidth capacity of link (i,j);

y_{ij}^{g} — Equal to 1 if link (i,j) is included in the multicast tree for group g;

$x_{ij}^{g,k}$ — Equal to 1 if link (i,j) is on the unique elementary path from the RP node r_g of group g to the group member node k in the multicast tree;

The integer programming problem of computing a set of Steiner trees with minimum overall bandwidth consumption is formulated as:

Minimise

$$\sum_{g=1}^{G} \sum_{(i,j)\in E_g} D_g \times y_{ij}^{g}$$

Subject to

$$\sum_{h\in V} x_{ih}^{g,k} - \sum_{j\in V} x_{ji}^{g,k} = \begin{cases} 1 & i = r_g \\ -1 & i = k, \quad k \in V_g \\ 0 & i \neq r_g, i \notin V_g \end{cases} \qquad (1)$$

$$x_{ij}^{g,k} \leq y_{ij}^{g} \qquad (i,j) \in E, k \in V_g \qquad (2)$$

$$x_{ij}^{g,k} = 0,1 \qquad (i,j) \in E, k \in V_g \qquad (3)$$

$$y_{ij}^{g} = 0,1 \qquad (i,j) \in E \qquad (4)$$

The variables to be determined are $x_{ij}^{g,k}$ and y_{ij}^{g} for every link $(i,j) \in E$. Constraint (1) ensures one unit of multicast flow from r_g to every group member node $k \in V_g$. Constraint (2) guarantees that the amount of flows along link (i,j) must be zero if this link is not included into the multicast tree for group g. Finally $x_{ij}^{g,k}$ and y_{ij}^{g} are confined to zero-one variables in constraints (3) and (4).

As we have mentioned before, the enforcement of the above set of hop-count Steiner trees can be achieved through explicit routing techniques such as MPLS on per group basis. However, the paths in the Steiner tree from r_g to individual group members $k \in V_g$ might not completely overlap with the shortest paths between them. This means that multicast traffic flowing on the Steiner tree will be discarded due to the network layer RPF checking failure in PIM-SM if the packets are not received from the correct interface on the shortest path back to the source. In order to apply the above

Copyright © MESCAL Consortium, June 2005

programming model to IP layer solutions, we introduce a unified M-ISIS link weight w_{ij} for each link (i, j), and by properly setting those link weights it is guaranteed that the path from r_g to each node $k \in V_g$ is the shortest path according to this set of weights, i.e., $\sum_{(i,j)\in P_k^g} x_{ij}^{g,k} \times w_{ij}$ is confined to be minimum, where P_k^g is the path on the hop-count Steiner tree from r_g to node $k \in V_g$. This is a necessary condition for successful RPF checking in PIM-SM routing, since all the paths in the hop-count Steiner tree are the shortest ones in terms of those link weights. It is also worth emphasising that the weight setting for each link is unified for all multicast groups, because the RPF table populated from M-ISIS is not group specific. In summary, the problem is formulated as follows: To set a unified M-ISIS weight for each link such that the shortest path tree for any group according to this set of link weights is in effect a Steiner tree in terms of hop-counts, which include minimum number of network links.

Still, we can append any QoS requirement to the above problem. For example, if the constraint of bandwidth capacity is considered, one extra constraint is to be added:

$$\sum_{g=1}^{G} y_{ij}^{g} \times D_g \le C_{ij} \qquad (i, j) \in E \qquad (5)$$

This constraint ensures that the overall bandwidth consumption on each link should not exceed the bandwidth capacity. The following OMTE-INTRA algorithm is proposed to solve this bandwidth constrained multicast traffic engineering problem, but it can be easily adapted for any other QoS requirement such as end-to-end delay.

13.2.3.2.2 A Genetic Algorithm Based Solution

The basic working mechanism of a Genetic Algorithm can be described as follows. First, a series of random solutions are obtained as the initial generation of chromosomes in the population. Thereafter, improved offsprings evolve iteratively from the parents by calculating their fitness. Chromosomes with higher fitness have higher probabilities of being inherited by the next generation. In each iteration, a new generation of approximations is created through the process of parent selection and reproduction. This is specifically achieved through genetic operators such as crossover and mutation. Finally, after a predefined number of generations, or the performance of fitness has reached its convergence, the resulting chromosome with the best fitness is selected as the final solution.

- Encoding and initial population

In a similar fashion to [FORTZ00], in our GA approach each chromosome is represented by a link weight vector $< w_1, w_{2,} ... w_{|E|} >$ where |E| is the total number of links in the network. The default value of each weight is within the range from 1 to 65535 (MAX_WEIGHT). In our experiments we define the value of MAX_WEIGHT to be 64 instead for reducing the search space. On the other hand, the population size is set to 100, with the initial values inside each chromosome randomly varying from 1 to MAX_WEIGHT. In addition to these randomly generated chromosomes, we add the solution of using hop-count as the link weight into the initial population.

- Fitness evaluation

Chromosomes are selected according to their fitness. In our approach, the QoS constraint (e.g., bandwidth capacity of each link) has also been included into the fitness function, such that a feasible

Copyright © MESCAL Consortium, June 2005

solution can be found. The fitness of each chromosome can be defined to be a two-dimension function of overall network load (l1) and excessive bandwidth allocated to saturated links (l2).

$$fitness = f(l1, l2) = \frac{\lambda}{\alpha \times l1 + \beta \times l2} \quad (6)$$

where α, β, λ are manually configured coefficients.

In equation (6) l1 and l2 are expressed as follows:

$$l1 = \sum_{g=1}^{G} \sum_{(i,j) \in E_g} D_g \times y_{ij}^g \quad (7)$$

$$l2 = \sum_{(i,j) \in E} \omega_{ij} \times \sum_{g=1}^{G} (D_g \times y_{ij}^g - C_{ij}) \quad (8)$$

where

$$\omega_{ij} = \begin{cases} 0 & if \quad \sum_{g=1}^{G} D_g \times y_{ij}^g \le C_{ij} \\ 1 & else \end{cases} \quad (9)$$

Procedure Computing_Fitness(Chromosome i)

Begin

Set the weight of each link in the network according to the gene values in chromosome i;

For each multicast Group g

Compute the shortest path tree T_g rooted at r_g, and spanning to all the group members in V_g;

For each link (i, j) in T_g

Update link load L_{ij} according to the bandwidth demand D_g of group g;

Load1 = 0; Load2 =0;

For each link (i, j) in the network

Load1= Load1 + L_{ij};

If $L_{ij} > C_{ij}$

Load2= Load2 + (L_{ij} - C_{ij});

Return fitness= $f(Load1, Load2)$;

End

Figure 127 Fitness calculation

Copyright © MESCAL Consortium, June 2005

- Crossover and mutation

According to the basic principle of Genetic Algorithms, chromosomes with better fitness value have higher probability of being inherited into the next generation. To achieve this, we first rank all the chromosomes in descending order according to their fitness, i.e., the chromosomes with high fitness (lower overall load) are placed on the top of the ranking list. Thereafter, we partition this list into two disjoined sets, with the top 50 chromosomes belonging to the upper class (UC) and the bottom 50 chromosomes to the lower class (LC). During the crossover procedure, we select one parent chromosome C_U^i from UC and the other parent C_L^i from LC in generation i for creating the child C^{i+1} in generation $i+1$. Specifically, we use a crossover probability threshold $K_C \in [0,0.5)$ to decide the genes of which parent to be inherited into the child chromosome in the next generation. We also introduce a mutation probability threshold K_M to replace some old genes with new ones.

Procedure Crossover(C_U^i **,** C_L^i **)**

Begin

For all genes $j = 1,...,|E|$

Generate $r = random\ [0,1]$;

If $r > K_C$

$C^{i+1}(j) = C_U^i(j)$;

Else if $r > K_M$

$C^{i+1}(j) = C_L^i(j)$;

Else

$C^{i+1}(j) = random[1, MAX\_WEIGHT]$

End For

Return C^{i+1};

End

Figure 128 Crossover and mutation

13.3 Offline Inter-domain Multicast Traffic Engineering (OMTE-Inter)

13.3.1 Introduction

In this section we extend Offline Multicast Traffic Engineering (OMTE) to the inter-domain scenario, and currently this is in the preliminary stage. Same as its intra-domain counterpart, we still focus on plain IP based solutions to avoid scalability issues in MPLS networks. In order to extend OMTE to inter-domain, PIM-SM needs not only to cooperate with multi-topology IGP such as M-ISIS, but also inter-domain routing protocols. As we have specified in D1.1, Multi-protocol BGP (MBGP) is the candidate for this role. From routing point of view, the M-ISIS link weight optimisation is responsible for intra-domain path selection, while the task of MBGP configuration is to control optimally how multicast traffic can be injected into the local domain. As it has been proposed for inter-domain

Copyright © MESCAL Consortium, June 2005

unicast traffic engineering, inter-domain traffic can be tuned through tweaking BGP policy metrics such as *local-preference*, *AS-Path* and *Multi-Exit-Discriminator (MED)* etc. These techniques can be also applied to MBGP for optimising inter-domain multicast traffic.

In this section we introduce our design of offline inter-domain multicast traffic engineering on top of the MBGP+M-ISIS routing infrastructure. Specifically we propose two distinct approaches: (1) single ingress router selection through configuration of local preferences, and (2) Hot Potato Routing (HPR) for multiple ingress router selection. In both approaches, the objective is to achieve efficient bandwidth utilisation within the network and at the same time to provide bandwidth guarantees with the constraint of inter-domain link capacities where the bottleneck of the Internet is often located. It should be noted that these two approaches are specifically for Source Specific Multicast (SSM), which allows individual designated routers to send direct group join requests towards individual sources even if they are located in different ASes.

13.3.2 Interface Specifications

Same as section 13.2.2

13.3.3 Behavioural Specification

13.3.3.1 Single Ingress Router Selection

13.3.3.1.1 Problem Description

If a common source for a set of multicast groups can be reached via more than one border routers, then the selection of which as the ingress router can also influence the relevant TE performance, e.g., intra-domain bandwidth consumption and inter-domain link utilisation. We illustrate this by showing a simple example concerning only one group in Figure 129. Assume that the address prefix which contains source S for group (S, G) can be reached via both ingress A and B, and also that the intra-domain IGP (e.g., M-ISIS) link weight setting is hop-count based. In this scenario, if we select ingress B by setting higher local-preference value than that of A for the address prefix of S, the resulting intra-domain multicast tree is shown in the left diagram in Figure 129. The total amount of bandwidth consumed is 7 units, assuming one unit of bandwidth is consumed on each multicast tree link. On the other hand, if ingress A is selected only 6 units of bandwidth is consumed, as the corresponding multicast tree is shown on the right diagram of Figure 129.

To enforce the selection, the local-preference for the address prefix that includes the selected ingress router should be assigned with higher value than any other candidate through which the source prefix can be reached. Finally, it should be noted that the ingress router selection for inter-domain multicast traffic engineering should be source specific other than group specific. This means that if multiple groups share a single source, the selected ingress router should be consistent among all groups.

Copyright © MESCAL Consortium, June 2005

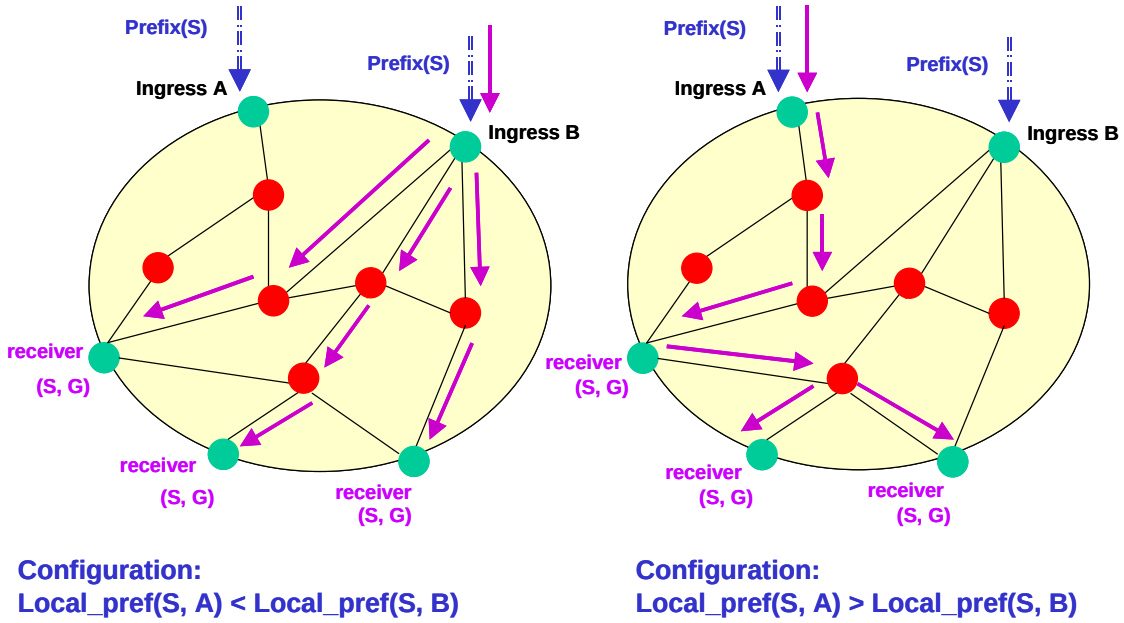


Figure 129 Single ingress router selection

13.3.3.1.2 Greedy Single Ingress Router Selection (GSIRS)

In this section we propose a simple greedy heuristic for the single ingress router selection with the aim of optimising intra-domain bandwidth resources. Moreover, we notice that inter-domain links are normally found to be the bottleneck of the Internet, and hence we put the bandwidth capacity of inter-domain links as the optimisation constraint. The pseudo code of the greedy heuristic is given in Figure 130.

Procedure Single_Ingress_Selection()

Begin

Configure IGP routing to be hop-count based by setting all link weights to be 1;

For each remote source $s \in PREFIX(S)$

Aggregate bandwidth demand for all groups;

For each PREFIX(S)

min = INFINITY;

For each ingress I that can reach PREFIX(S)

If (inter-domain link capacity is sufficient)

Set I as the current ingress router for all groups associated with PREFIX(S);

Calculate overall intra-domain bandwidth consumption sum;

If (min > sum) min = sum**;**

Set the highest local-preference value to the ingress $I_{\min}$ with bandwidth consumption equal to min;

Copyright © MESCAL Consortium, June 2005

Update inter-domain link bandwidth on I_{min};

End

Figure 130 Single ingress router selection heuristic

13.3.3.2 Multiple Ingress Router Selection

13.3.3.2.1 Hot Potato Routing

Hot Potato Routing (HPR) has been often adopted for routing inter-domain unicast traffic, with the objective of achieving lowest bandwidth consumption. The basic idea of HPR is to deliver unicast traffic out of the local domain through the closest egress router according to IGP costs. To achieve this, each BGP speaker examines all the internal peers that can reach the remote destination prefix, and then selects the one with the least distance in terms of IGP link metrics. In this scenario, the simplest approach is to set the IGP link weight to be 1 so that the IGP path reflects effectively the total number of hops towards each egress router. According to the BGP route selection procedure, in order to enable hot potato routing, all the BGP policy parameters that have higher priority than the IGP cost metric should be configured to be equal. That is to say, only if multiple inter-domain routes have equal values in local-preference, AS paths, MED and also all the advertisements are received via the same type of BGP updates (either iBGP or eBGP), HPR may take effect in making routing decisions.

13.3.3.2.2 Problem Description

In the context of inter-domain multicast traffic engineering, the IGP link weight optimisation has two impacts. First, to represent Steiner trees into shortest path trees, which has already been studied in the intra-domain multicast sections. Second, to enable multiple ingress router selection for a common source, if there exist multiple potential ingress routers through which this particular source can be reached.

We take Figure 131 as an example to illustrate the difference between single and multiple ingress routers selection. On the left part of the figure, ingress B is selected as the unique ingress for (S, G) group (e.g., by setting higher value of local-preference for S), with overall intra-domain bandwidth consumption being 6 units (of course there exist some other single ingress configurations with equally good results). On the other hand, if ingresses A and B are configured as equally good routes so as to enable HPR, then a proper IGP link weight setting to enable multiple ingress router selection is possible, and one of them is shown on the right part of Figure 131, with overall intra-domain bandwidth consumption being 5 units.

Copyright © MESCAL Consortium, June 2005

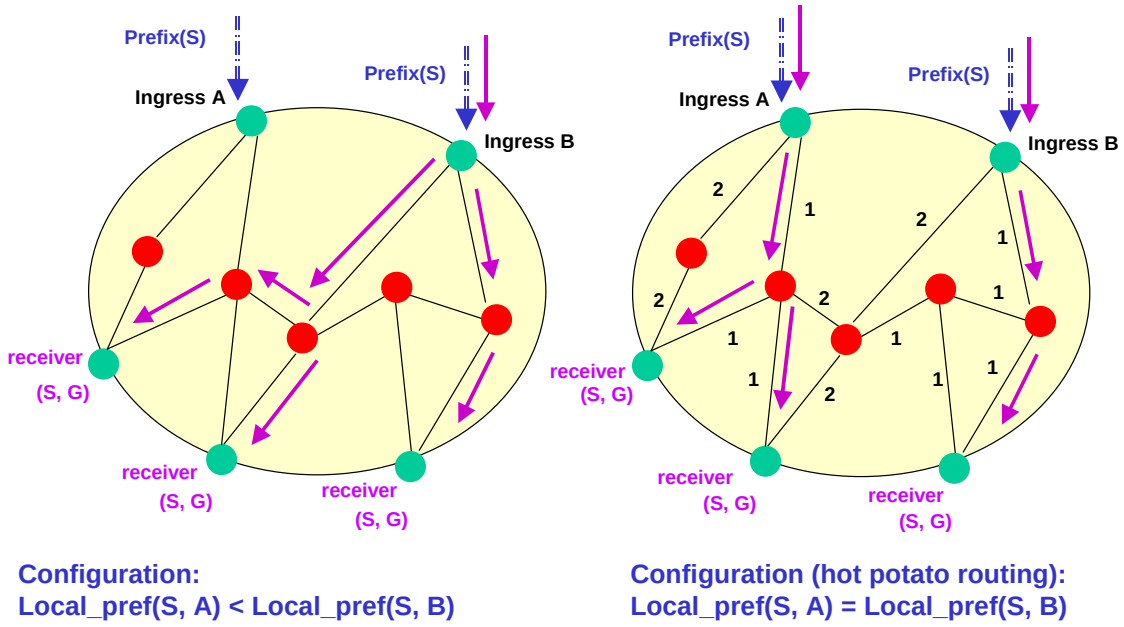


Figure 131 Single vs. Multiple ingress router selection

13.3.3.2.3 Conventional HPR

In conventional hot potato routing, the link weight of the multi-topology IGP is configured to be hop-count based. Thereafter, each designated router is able to find its nearest border router to send its PIM-SM group join requests towards remote sources. In this case, multicast traffic flows can reach the DR with minimum number of hops within the domain. However, this does not mean that the overall bandwidth consumption within the domain is minimised. Another shortcoming is that, bandwidth resources could be consumed excessively on inter-domain links where bottlenecks are liable to form. In order to solve these problems, we propose in the next section a more intelligent approach to perform restricted HPR routing for multicast traffic.

13.3.3.2.4 A Genetic Algorithm based Solution

In effect, further optimisation can be done on top of the conventional HPR through IGP link weight optimisation. First, intra-domain bandwidth conservation can be still improved. Second, load balancing on inter-domain links can be also achieved through more intelligent ingress router selection. We extend our original GA based OMTE-INTRA solution for the inter-domain scenario. As the basic operation of genetic algorithms has already been described in the previous section, it will not be repeated here.

Compared to OMTE-INTRA, there are some important issues that need to be addressed in the inter-domain scenario. One important concern is that, if individual receivers for the same source/group use different border routers to send their join requests, this results in the situation that the multicast traffic will be injected via more than one border routers, thus consuming more bandwidth on inter-domain links which is deemed to be precious network resources. In order to avoid uncontrolled over-consumption of bandwidth on inter-domain links, HPR should be applied in a restricted fashion to eliminate unnecessary choice of many border routers. In Figure 132, both of the two configuration examples are able to achieve minimum intra-domain bandwidth consumption. However, the configuration on the left side uses two separate ingress routers, which results in double consumption of bandwidth on inter-domain links. In comparison, the IGP link weight configuration on the right side tries to confine the merging point of the multicast tree within the local network, thus the bandwidth on the inter-domain link attached with ingress B is conserved. From Figure 131 and Figure 132 we can notice that OMTE-INTRA needs an efficient control mechanism to balance the trade-off between

Copyright © MESCAL Consortium, June 2005

intra-domain and inter-domain network resources. Our proposed strategy is, (1) unless significant bandwidth resources can be conserved, multiple ingress router selection has lower preference than single ingress router selection, and (2) load balance across inter-domain links has higher priority than the corresponding overall bandwidth consumption.

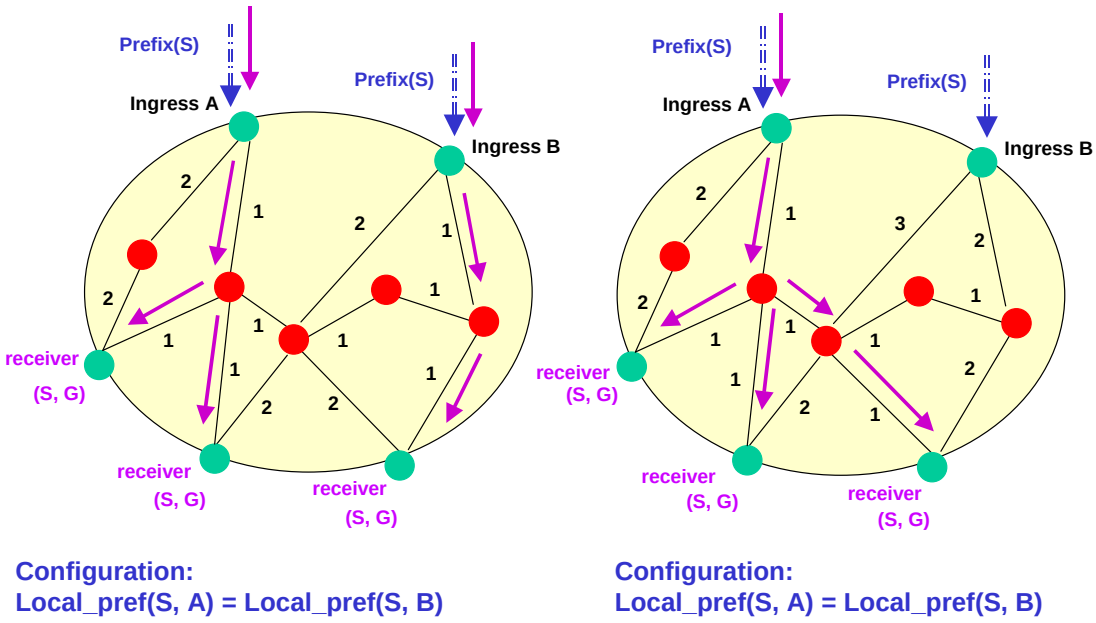


Figure 132 Restricting unnecessary ingress router selection

13.4 Dynamic Group Management (DGM)

13.4.1 Introduction

The basic task of dynamic group management can be summarised into two parts: (1) to efficiently handle group membership dynamics with heterogeneous QoS requirements, and (2) to enforce admission control on multicast receivers to avoid network congestion due to overwhelming group subscriptions.

In the current best-effort based multicast services, Internet Group Management Protocol (IGMP) has the responsibility of managing multicast group dynamics. When a new group membership report is received, the Designated Router (DR) will trigger the PIM-SM protocol for sending the join request towards the source. In case that multicast group members demand different QoS requirements, IGMP should be extended to be QoS-aware such that the underlying routing protocol can be triggered to explore a proper path for the required QoS class. Moreover, special considerations should be taken when receivers with different QoS demands are attached to the common broadcast network. In this scenario, actions should be taken to prevent the receiver with lower QoS demand from accessing the same multicast data content but with higher QoS treatment that is requested by other members who might be charged more for his higher QoS services.

In the unicast scenario, admission control applies to external data sources during the SLS invocation period to prevent network congestions. When receiver-initiated multicast services are considered, sender oriented admission control is far from sufficient, since packets can be replicated anywhere due to multiple join requests from group members. This aspect explains why admission control should be performed at the receiver side when the multicast SLS is activated. In effect, network bandwidth can be over-reserved during the offline TE phase for efficient utilisation of resources, and this incurs possible congestion when overwhelming multicast SLSs are invocated simultaneously. If the DR receives a join request with the QoS requirement that the network resources cannot handle, this request should be rejected. To achieve this, the current IGMP should be extended such that excessive group

Copyright © MESCAL Consortium, June 2005

membership reports are suppressed and the underlying routing protocol will not be triggered for join request delivery at the DR in time of congestion.

13.4.2 Interface Specification

Group_Member_Invocation(DGM to DMR)

When the join request from a new group member is notified in DGM, the following parameters should be passed to DMR:

- Source address;
- Group address;
- Specification of QoS requirements (e.g. QCs) by the new receiver.

13.4.3 Behavioural Specification

In our proposed DGM solution, one or more class D addresses in the SSM address range 232/8 are used to encode each QoS class (QC) provided by the ISP (see Table 21). In such a situation, the interpretation of the SSM address tuple (S, G) becomes straightforward: S identifies the address of the information source and G stands for the QC level (we name it QoS channel) that is available from S. The advantage of this scheme is that QoS requirement handling for individual receivers is translated into multicast group management, which can be directly fulfilled using IGMPv3 on a broadcast LAN. On the other hand, encoding QCs into SSM group address solves scalability problems in terms of QoS state maintenance at DiffServ core routers during dynamic multicast tree construction, and this will be specified in the DMR section. It should be noted that any class D address that does not belong to 232/8 is not considered to have such functionality. In effect, the maximum number of QCs in DiffServ is restricted by 6 bits of the DSCP field, and the allocation of 64 dedicated class D addresses will not cause any scalability problem in the usage of SSM address range that contains 2^{24} addresses.

If the mapping between SSM group addresses and QCs is strictly one-to-one, then each source can be only involved in one group per QC. Since the QoS channel is source specific, it is impossible for a single source with a unique IP address S to send multiple data streams with different contents. In the classic SSM model, an information source can be simultaneously involved in multiple groups because (S, G1) and (S, G2) are completely independent with each other. To avoid this restriction, we suggest that the mapping between SSM group address and QCs should be m-to-1, which enables a single source to be able to send traffic multiple groups (up to m) within a particular QC.

SSM group address	QoS Class
G1 (232.*.*.*)	QC1
G2(232.*.*.*)	QC2
…	…
Gn(232.*.*.*)	QCn

Table 21 SSM QC encoding table (one-to-one mapping)

The management of group join requests with heterogeneous QC demands is as follows. Once an end user wants to subscribe to a multicast service provided by the source S in a desired QoS channel (i.e., QC), it will send an IGMPv3 (S, G) group membership report to its Designated Router (DR), where G is the associated group address mapped to that QC. On receiving this report, the DR will send an (S, G) join request towards S if this is the first (S, G) membership report appeared on the LAN. In case that multiple receivers subscribe to one multicast session provided by S but with different QCs, IGMPv3 should handle these membership reports independently since they contain different SSM

Copyright © MESCAL Consortium, June 2005

group address. Finally, it is worth mentioning that admission control for multicast receivers is also on per (S, G) channel basis. That means, receiver-oriented mcSLS invocations for the same multicast source S but with different QC requirements are performed independently.

13.5 Dynamic Multicast Routing (DMR)

13.5.1 Introduction

Dynamic multicast routing refers to the procedure of multicast tree construction with dynamic group membership updates (i.e., mcSLS activation) with heterogeneous QoS class requirements. The basic task is (1) to build the deliver tree(s) that satisfies the QoS demand of all the attached receivers, and (2) dynamic path selection for bandwidth conservation and load balancing purposes. This functionality can be split into intra- and inter-domain parts, which respectively corresponds to the traffic engineering components in the offline blocks located in the management plane.

In the intra-domain scenario, the PIM-SM routing protocol constructs multicast trees based on the underlying unicast routing table. Traditionally this routing table for both unicast and multicast is populated by intra-domain unicast protocols such as OSPF and ISIS. To decouple multicast routing from the unicast realm, Multi-topology ISIS [PRZYG03] is extended from ISIS and it can provide dedicated routing decisions for PIM-SM tree construction. As a result, it is possible that the routing process can be dynamically managed specifically for multicast QoS demands and traffic engineering purposes. In case of statistical traffic fluctuations within each RPC, the adapted PIM-SM can make dynamic routing decisions to avoid further potential congestions.

At the inter-domain level, multi-protocol BGP (MBGP) [BATES00] is currently used for advertising multicast source reachability information which gives input to the inter-domain PIM-SM group join towards remote sources. Within each RPC, if there is significant domain-level topology or resource availability changing, the QoS-aware MBGP (qMBGP) is responsible for advertising updated reachability information (e.g., MP_REACH_NLRI and MP_UNREACH_NLRI) such that PIM-SM is able to dynamically decide the path for inter-domain join request delivery. The most challenging issues in this scenario include: (1) the extension of qMBGP from MBGP for the eQC-aware reachablity information of inter-domain multicast sources and (2) the corresponding online adjustment of multicast path selection across multiple domains.

Another important issue of DMR is the multicast tree construction with different QCs that serve heterogeneous QC requirements. In DiffServ networks, there are two strategies for building multicast trees that support heterogeneous receivers. First, a single tree exhibiting all QCs can be constructed for each group session, and branches with lower QCs can be directly grafted onto the part of the tree that has higher QC treatment. We name this strategy the hybrid tree approach. Another solution is to build a dedicated multicast tree for each QC, which means that k trees are needed for a particular group where k is the total number of QCs the ISP is providing within the network. We name this approach per QC trees. The difference between the two types of the trees is shown in Figure 133. In this figure we assume that QC(i) is in higher priority than QC(j) if $i < j$. The advantage of per QC trees lies in its simplicity in implementation and management, while the hybrid tree has its virtue in bandwidth and group state conservation. The choice between the two strategies is one of the basic issues that the DMR block is going to address.

Copyright © MESCAL Consortium, June 2005

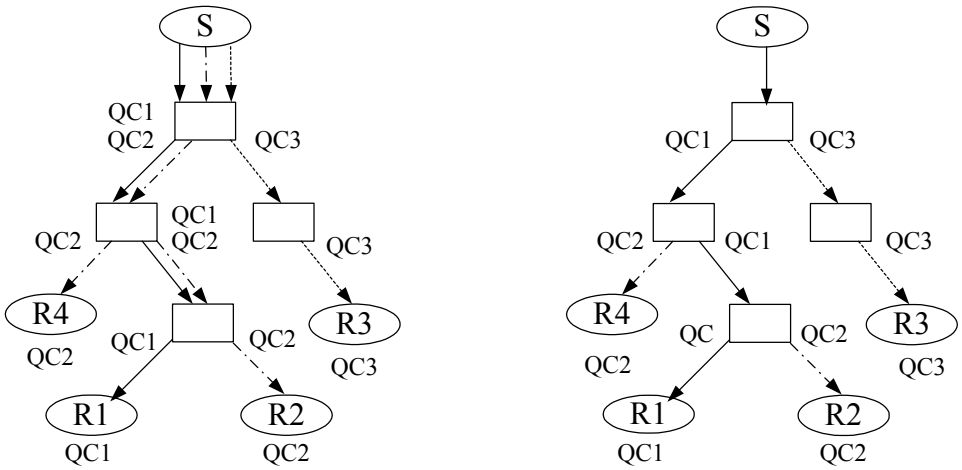


Figure 133 Per-QC tree vs. hybrid tree

13.5.2 Interface Specification

(1) Group_State_Update(DMR to MF)

This function basically informs the Multicast Forwarding block to create a new group state when a join request has received at a node for the first time. The MF will install the new (S, G) state as well as its iif and oif (i.e., the interface from which this join request has been received). The parameters include:

- Source address;
- Group address;
- Address of the iif;
- Address of the new oif.

(2) Oif_List_Update(DMR to MF)

This function basically informs MF to update the outgoing interface list (oif) when a new join request packet is received at a node that has already obtain the group state (i.e., on-tree router). Hence the MF will instruct the core router to forward multicast data packet on the new oif. The parameters include:

- Source address;
- Group address;
- Address of the new oif.

(3) PHB_State_Update(DMR to PE)

When a join request is received at each core router, not only the corresponding oif list should be updated, but also the associated PHB that is responsible for the multicast data packet treatment. The task of this function is to update at each oif the new PHB state with the dynamics of received group joins having different QC requirements. It should be noted that how this state is updated depends on the dynamic routing policy, i.e., whether per-PHB trees or a hybrid tree is adopted. The parameters include:

- Source address;

Copyright © MESCAL Consortium, June 2005

- Group address;
- Address of the new oif;
- The PHB used for requested QC;
- Routing policy ID (per QC tree vs. Hybrid tree).

(4) Iif_Update(DMR to RC)

Due to the network resource dynamics within each RPC, it is possible that the incoming interface (iif) for a particular source prefix is dynamically modified for the purpose of online bandwidth optimisation. This function provides RC the updated incoming interface for RPF checking. The parameters include:

- Source address (prefix);
- The new interface that is configured on the shortest path back to the source (prefix).

13.5.3 Behavioural Specification

We apply per QC trees in DMR. The most distinct advantage of this strategy is that bandwidth resources are much easier in provisioning for different QCs, because tree construction in one QC does not interact with any other QC in bandwidth consumption. From routing point of view, since M-ISIS can provide different routing table for multiple QCs, the PIM-SM join request for different QCs may follow different join paths towards the same source, thus resulting different tree shapes. According to [PRZYG03], M-ISIS can provide up to 128 different IP topologies, this means that the proposed scheme can support 128 QCs in maximum if M-ISIS is exclusively used for multicast services.

13.5.3.1 Intra-domain DMR

The construction of per QC trees is as follows. Once an end user wants to subscribe to the multicast service rooted at source S in a desired QC, it will send an IGMPv3 (S, G) group membership report to its Designated Router (DR), where G is the associated group address mapped to that QC. On receiving this report, the DR will send a (S, G) join request towards S if the invocation has been admitted. This join request packet will either be intercepted by an on-tree router with the same (S, G) state or arrive at S itself.

In DMR, how to enable PIM-SM join requests with different group addresses (i.e., different QCs) to follow different join paths for achieving per QC tree constructions is a key issue. In the conventional SSM routing with best effort service, the underlying routing table is not group specific, but exclusively source specific. In this case, there should be additional mechanism needed for differentiating multiple routing topologies for different QCs. As we have mentioned in OMTE-INTRA, M-ISIS is used to provide different routing tables for each QC. Hence, it is required that a mapping mechanism is used to link the group address carried in PIM-SM join packets to a specific M-ISIS routing table within a network. During the group join phase, when a router receives an (S, G) join request, it first finds the corresponding routing table by mapping the group address G into an M-ISIS MT-ID. Thereafter, the router looks up the routing table with that MT-ID and finds the NEXT_HOP for the source S. Finally, it forwards the (S, G) join packet on the interface associated with that specific NEXT_HOP. If there exist multiple NEXT_HOP entries leading towards the same source, the router should be allowed to deliver join requests in an ECMP style for load balancing purpose. From the above description, we notice that one extra mapping list should be maintained within each router such that group address can be linked to a specific QC topology identified by a dedicated MT-ID of the M-ISIS routing protocol.

Copyright © MESCAL Consortium, June 2005

Apart from the relationship between group address and MT-ID, considerations should also be taken on the interactions between group addresses and DSCPs, and this will be illustrated in the PHB enforcement section.

13.5.3.2 Inter-domain DMR

Inter-domain DMR is supported by Multi-protocol BGP (MBGP). Similar to the unicast scenario, inter-domain group joins are based on QoS aware MP_REACH_NLRI configuration at border routers within each AS. Within an RPC, binding selection for multicast services decides one or multiple MBGP edge routers for delivering inter-domain join requests towards a remote source address/prefix. During the mpSLS invocation period, group join packets can dynamically choose different edge routers decided by binding selection according to the instant QoS conditions and bandwidth consumption. As a result, the constructed inter-domain multicast tree can be dynamically adjusted due to the different join paths.

One of the challenges in handling inter-domain QoS delivery lies in the fact that ISPs have heterogeneous DiffServ configuration policies. For example, each DiffServ domain might have different number of QCs, and meanwhile the mapping between group address (i.e., QC identification in the PIM-SM join request) and the M-ISIS MT-ID is not necessarily consistent in all domains for the purpose of flexibility. This requires that the locally selected group address for QC identification should be made known to foreign domains, which is similar to the DSCP usage in unicast services. To achieve inter-domain per QC trees, we propose a swapping mechanism for group address at the border router of adjacent domains. Assume that a group join request needs to travel two adjacent domains A and B to reach the group source S, and it is required that the corresponding multicast flow should be treated with a particular eQC formed by lQC(A) and lQC(B) in the two domains respectively. In this case, the group address identifying lQC(A) in domain A should be changed into the address that corresponds to lQC(B) in its adjacent domain. This is because, for lQC(A), the ISP of domain A maps group address G(A) to MT-ID(A) for delivering the join packet within the local AS, but G(A) might not be recognised in domain B, or this address is mapped to some other lQCs using different routing tables. Hence, a swapping table should be agreed between two domains, so that each AS is able to perform correct local mapping between group address and MT-ID for the construction of inter-domain per QC trees.

13.6 PHB Enforcement (PE)

13.6.1 Introduction

Since multicast data packets can be replicated at core routers, additional issues arise for the corresponding PHB Enforcement. First, the join request from group members should be able to inform core routers in the tree about the associated QoS classes, so that the latter is able to enforce corresponding PHBs at the outgoing interfaces leading to heterogeneous receivers. Hence the first requirement of multicast PHB enforcement is to extend PIM-SM join request packet for inclusion of QoS class requirements that can be met with the configured PHBs. Second, as multicast trees are maintained through group states at core routers, if the outgoing interfaces of a router are associated with different QoS classes for the same (S, G) group, the maintained group state should also be extended for the associated QC information. We name this extra information at each outgoing interface QC state. It can be inferred that this type of state extension should take place at each outgoing interface of the (S, G) group.

The functional block of PHB Enforcement only provides some mechanism for supporting multicast services, and it does not output any input/output interface to other blocks.

Copyright © MESCAL Consortium, June 2005

13.6.2 Interface Specification

The PHB Enforcement interfaces are shown in Figure 135.

13.6.3 Behavioural Specification

SSM group address is used for carrying QC requirements from group members during the group join procedure. When multicast data packets are delivered backwards along each QC specific multicast tree, the PHB treatment is still based on Diffserv DSCP value. In our proposed scheme, QC states mentioned in section 13.5.1 are not needed to be maintained at core routers, because they can be directly reflected by (S, G) group states. This means that the issue of scalability at core routers is avoided. More detailed analysis on this can be found in [WANG04]. However, considerations should be taken on how the group address in a multicast packet is translated into a proper DSCP value for multicast data treatment. To solve the problem, we need a table at edge routers to map SSM group addresses (identifying QCs) into a specific DSCP value. The relationship of group addresses and DSCP values is somehow different from its relationship with MT-ID (see section 13.4.3). The most distinguished point is that: the mapping between group address and DSCP only takes place at edge routers, while each core router must know how to map one group address into a proper MT-ID in M-ISIS based PIM-SM routing.

For intra-domain multicast services, the edge router attached to the source is responsible for the translation of group address into a specific DSCP. When the edge router attached to a source receives a multicast data packet, it will mark the DSCP value of this packet coming from the source according to the pre-configured mapping table between group addresses and DSCPs, which should be agreed in the SLS between the source and the ISP. For example, when the edge router attached to source S receives an (S, G) multicast packet, the router first checks the locally maintained mapping table, and it then finds the DSCP value that is associated with group address G. Finally the edge router uses this value to mark the (S, G) multicast packet that will be injected into the network. When a core router that is already on the (S, G) tree receives another (S, G) join request from a new interface, it simply duplicates the multicast packet and forwards on the new oif. It should be noted that the DSCP value of the new packet is automatically inherited from the incoming packet, and this guarantees that each (S, G) tree is also a QC specific tree.

In the inter-domain scenario, group address swapping at border routers of adjacent domains is also needed for correct DSCP usage, with the reason being the same with MT-ID described in section 13.4.3. As it is shown in Figure 134, since group address swapping is performed at edge routers, inter-domain DSCP swapping is not necessary because the DSCP value to be used in the local domain can be obtained from the mapping of the swapped group address at the local edge router. We should also emphasise that the group address in both join packets and multicast data packet should both undergo such type of swapping in the inter-domain scenario. As a summery, group address swapping is double folded: first it enables inter-domain per QC tree construction by means of local mapping with MT-ID during the group join procedure (routing), and second, it enables correct PHB enforcement by means of local mapping with DSCP value (scheduling).

Copyright © MESCAL Consortium, June 2005

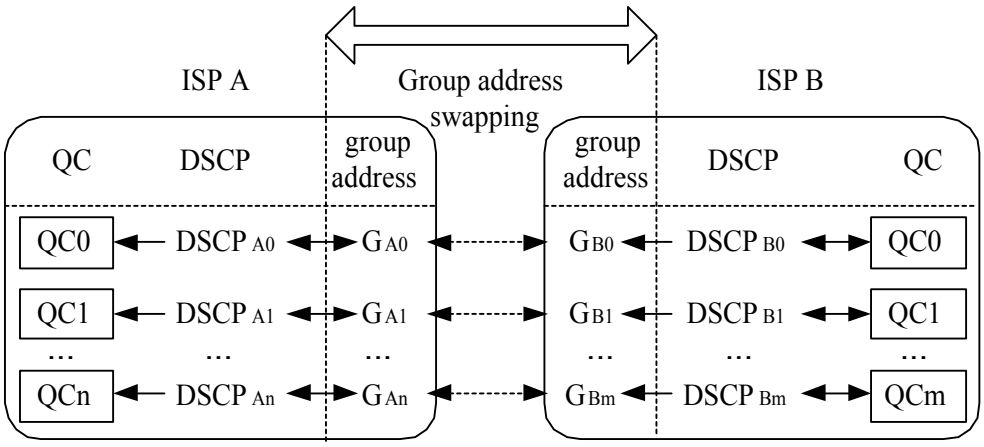


Figure 134 Inter-domain group address swapping

13.7 Multicast Forwarding (MF)

13.7.1 Introduction

The multicast forwarding part in the MESCAL project is not appended with additional functionalities compared to the conventional forwarding mechanism for multicast packets. On the other hand, the treatment of replicated packets with different PHBs is described in the PHB enforcement block.

13.7.2 Interface Specification

(1) PHB_Lookup(MF to PE)

During multicast data transmission, this function returns to MF how to schedule the data packets with proper PHBs at each outgoing interface. This action is based on the PHB state maintained at the PE block. The parameters include:

- PHB state

(2) Iif_Lookup(MF to RC)

This function returns valid incoming interface to MF. If the data packet is not coming from this returned interface, it will be discarded. The parameters include:

- Source address/prefix;
- The address of the valid iif for the source (prefix).

13.7.3 Behavioural Specification

MESCAL will not study the Multicast Forwarding any further.

13.8 RPF Checking (RC)

RPF checking is a simple but efficient mechanism for preventing traffic loops during packet delivery in IP multicast and SSM. Since the MESCAL solution for multicast services will be based on the SSM

Copyright © MESCAL Consortium, June 2005

model, RPF checking should be retained. On the other hand, there will be no extra adaptations on the RPF checking itself for QoS support.

MESCAL will not study the RPF checking any further.

13.9 The Overall Diagram

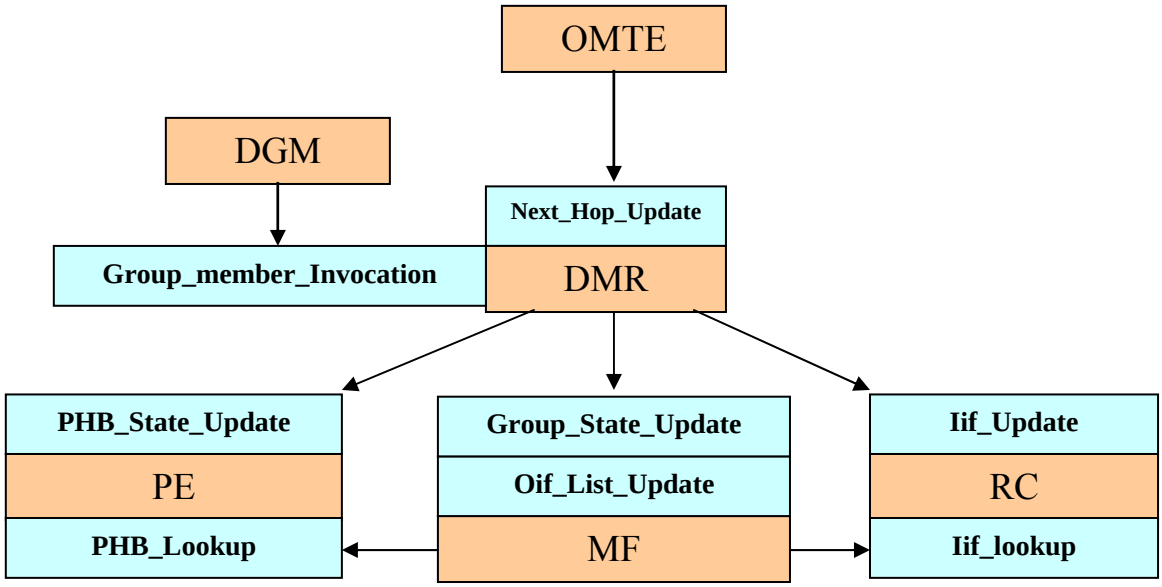


Figure 135 Overall diagram

Copyright © MESCAL Consortium, June 2005

14 REFERENCES

[ABN01] Abarbanel, B., Venkatachalam, S., BGP-4 support for Traffic Engineering, draft-abarbanel-idr-bgp4-te-00.txt, September 2000

[AHUJA93] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, *Network Flows: Theory, Algorithms and Applications*, Prentice Hall, 1993

[AKA99] Akamai "Internet Bottlenecks: the Case for Edge Delivery Services", Technical Report, Akamai White Paper 1999.

[AKE03] A.Akella, S.Seshan and A.Shaikh "An Empirical Evaluation of Wide-Area Internet Bottlenecks", ACM Internet Measurement Conference 2003.

[ALBE01] J. L. Alberi, Ta Chen, S. Khurana, A. Mcintosh, M. Pucci, and R. Vaidyanathan, "Using Real-Time Measurements in Support of Real-Time Network Management," RIPE-NCC 2nd Workshop on Active and Passive Measurements (PAM2001), Amsterdam, April 2001

[ASGA04] A. Asgari, R. Egan, P. Trimintzios, G. Pavlou, "Scalable Monitoring Support for Resource Management and Service Assurance", IEEE Network Magazine, Dec./Nov. 2004, Vol. 18, No. 6, pp. 6-18.

[ASGA03] A. Asgari, P. Trimintzios, M. Irons, G. Pavlou, and R. Egan, "Building Quality of Service Monitoring Systems for Traffic Engineering and Service Management," Journal of Network and Systems Management (JNSM), Vol. 11, No. 3, December 2003.

[ASP00] B. Jaruzelski, R. M. Lake, F. M. Riberio, ASP101: *Understanding the Application Service Provider Model*, Booz.Allen & Hamilton, 2000, www.bah.com.

[ATKI03] Atkinson, R., Floyd, S., "IAB Concerns & Recommendations Regarding Internet Research & Evolution", draft-iab-research-funding-01.txt, July 2003.

[AUKIA00] P. Aukia, et al., *RATES: A Server for MPLS Traffic Engineering*, IEEE Network Magazine, vol. 14, no. 2, pp. 34-41, March 2000 [BATES00] T. Bates et al, "Multiprotocol Extensions to BGP4," RFC 2858, June 2000.

[BAIN] A.Bain, P.Key, *Modelling the Performance of Distributed Admission Control for Adaptive Applications*

[BAKER00] F. Baker, C. Iturralde, F. le Faucher and B. Davie, *Aggregation of RSVP for IPv4 and IPv6 Reservations*, Internet Draft, draft-ietf-issll-rsvp-aggr-02.txt, Mach 2000.

[BALLA97] A. Ballardie, *Core Based Trees (CBT version 2) Multicast Routing*, *RFC* 2189, September, 1997

[BATES98] T. Bates et al, *Multiprotocol Extensions to BGP-4*, *RFC* 2283, February 1998

[BEGD02] Bernet, Y., Elfassy, N., Gai, S. and D. Dutt, *RSVP Proxy*, draft-ietf-rsvp-proxy-03, expired Sept. 2002.

[BELE02] S.Belenki "An Enforced Inter-Admission Delay Performance-Driven Connection Admission Control Algorithm," *ACM SIGCOMM*, April 2002, Vol. 32, No. 2, pp. 31-41.

[BGRP] P. Pan, E, Hahne, and H. Schulzrinne, *BGRP: A Tree-Based Aggregation Protocol for Inter-domain Reservations*, Journal of Communications and Networks, Vol. 2, No. 2, June 2000, pp. 157-167.

[BGRP+] Stefano Salsano (ed.), *Inter-domain QoS Signalling: the BGRP Plus Architecture*, Internet Draft <draft-salsano-bgrpp-arch-00.txt>, Expired November, 2002.

Copyright © MESCAL Consortium, June 2005

[BGRP-fm] P. Pan, E. Hahne, H. Schulzrinne, *BGRP: A Framework for Scalable Resource Reservation*, Internet Draft <pan-bgrp-framework-00.txt> Bell Labs/Columbia Uni., Expired July, 2000.

[BGRP-per] Eugenia Nikolouzou, et al., BGRPP: *Performance evaluation of the proposed Quiet Grafting mechanisms*, Internet Draft <draft-nikolouzou-bgrpp-sim-00.txt>, Expired January 2003.

[BHATT00] S. Bhattacharyya et al, *A Framework for Source-Specific IP Multicast Deployment*, draft-bhattach-pim-ssm-00.txt, July 2000

[BIANC03] G. Bianchi et al, *QUASIMODO: Quality of Service-aware Multicasting Over DiffServ and Overlay Networks*, IEEE Network, special issue on multicasting, January/February, 2003

[BLAK98] S. Blake, et al, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[BLESS02] R. Bless, K. Wehrle, *IP Multicast in Differentiated Services Networks*, <draft-bless-DiffServ-multicast-05.txt>, work in progress, November 2002

[BON01] Bonaventure, O., *Using BGP to distribute flexible QoS information*, draft-bonaventure-bgp-qos-00.txt, February 2001

[BON01] T.Bonald, A.Proutiere and J.W.Roberts "Statistical Performance Guarantees for Streaming Flows using Expedited Forwarding," *IEEE INFOCOM 2001*, Vol. 2, pp.1104-1012.

[BONA02] T.Bonald, S.Oueslati-Boulahia and J.Roberts "IP traffic and QoS control: the need for a flow-aware architecture", World Telecommunications Congress, September 2002.

[BOUC05] Boucadair M., Morand P., "PCE discovery via Border Gateway Protocol", draft-boucadair-pce-discovery-01.txt, Work in progress, May 2005

[BREIT02] Y. Breitbart, M. Garofalakis, A. Kumar and R. Rastogi, *Optimal Configuration of OSPF Aggregates*, In Proc. of IEEE INFOCOM02, New York, USA, June 2002

[BRES00] L. Breslau, S. Jamin and S. Shenker "Comments on the Performance of Measurement-Based Admission Control Algorithms", IEEE INFOCOM 2000.

[BRES03] T.Bressoud and R.Rastogi "Optimal Configuration for BGP Route Selection", IEEE INFOCOM 2003.

[BURIO02] Buriol, L.S., M.G.C. Resende, C.C. Ribeiro, and Mikkel, "A Memetic Algorithm for OSPF Routing," *Proceedings 6th INFORMS Telecom*, pp. 187-188, 2002.

[CAIDA] Cooperative Association for Internet Data Analysis (CAIDA) projects, for more information visit http://www.caida.org. [CAMAR02] Camarillo G. et al., *Integration of Resource Mangement and SIP,* IETF Internet Draft <draft-ietf-sip-manyfolds-resource-07.text>, April 2002

[CAO00] Z. Cao, Z. Wang, and E. Zegura, *Performance of Hashing-based Schemes for Internet Load Balancing*, In Proc. of IEEE INFOCOM 00, pp. 332-341, March 2000

[CARLB97] K. Carlberg et al, *Building Shared Trees Using a One-to-many joining Mechanism*, *ACM* Computer Communication Review, January 1997, pp5-11

[CDN01] B. Krishnamurthy, C. Wills, Y. Zhang, *On the Use and Performance of Content Distribution Networks*, ACM SIGCOMM Internet Measurement Workshop 2001.

[CENTIN00] C. Centinkaya and E. Knightly, *Scalable Services via Egress Admission Control*, In Proceedings of IEEE INFOCOM'00, Tel Aviv, Israel, March 2000.

[CETINK] C.Cetinkaya, E.W. Knightly, *Egress Admission Control*

Copyright © MESCAL Consortium, June 2005

[CHARZ01] J.Charzinski "Problems of Elastic Traffic Admission Control in an HTTP Scenario", *Proceedings of IWQoS* 2001.

[CHAIT02] Y.Chait, C.V.Hollot, V.Misra, D.Towsley, H.Zhang and J.Lui "Providing Throughput Differentiation for TCP Flows Using Adaptive Two Color Marking and Multi-Level AQM," *IEEE Infocom* 2002, New York, NY, 23-27 Jun. 2002.

[CHEN00] S. Chen et al, *A QoS-Aware Multicast Routing Protocol*, in proc. IEEE INFOCOM 2000, Vol. 3 pp.1594-1603

[CHEN98] S. Chen, K. Nahrstedt, *An Overview of Quality-of-Service Routing for the Next Generation High-Speed Networks: Problems and Solution*", IEEE Network Magazine, vol. 12, no. 6, pp. 64-79, November 1998

[CHOI01] B. Kyu Choi, R.Bettati, *Endpoint Admission Control: Network Based Approach*, 21 International Conf. On Distributed Computing Systems, Phoenix, April 2001

[CONOV99] Conover, J., *Policy-based Network Management*, Network Computing, http://www.networkcomputing.com/1024/1024f1.html, 1999

[CRI01] G. Cristallo, C. Jacquenet, *Providing Quality of Service Indication by the BGP-4 Protocol: the QOS_NLRI attribute*, <draft-jacquenet-qos-nlri-04.txt>, March 2002

[CHU96] P.C. Chu and J.E. Beasley, "A genetic algorithm for the generalised assignment problem," *Computers Ops Res*., Vol. 24, No.1, pp.17-23, 1997.

[Cisco] Open Shortest Path First, http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/ospf.htm.

[COPP04] J. Coppens, E.P. Markatos, J. Novotny, M. Polychronakis, V. Smotlacha & S. Ubik; "SCAMPI - A Scaleable Monitoring Platform for the Internet"; Proceedings of the 2nd International Workshop on Inter-Domain Performance and Simulation (IPS 2004), Budapest, Hungary, 22-23 March 2004.

[COUT00] A. Couturier, "Singnallingfor QoS measurement", IETF Internet Draft, draft-couturier-nsis-measure-00.txt, expired November 2003.

[CRIS05] Cristallo, G., Jacquenet, C, " Providing Quality of Service Indication by the BGP-4 Protocol: the QOS_NLRI attribute", <draft-jacquenet-qos-nlri-05.txt>, work in progress.

[Damian01] Damianou, N., Dulay, N., et al., *The Ponder Policy Specification Langauge*, in Proc. of the IEEE Workshop on Policies for Distributed Systems and Networks (Policy 2001), Bristol, U.K., January 2001

[Dee01] S. E. Deering, *Multicast Routing in a Datagram Internetwork*, Ph.D. thesis, Stanford University, Dec 1991

[Deering89] S. Deering et al, *Host Extensions for IP Multicasting*, *RFC* 1112, Aug. 1989

[Deering96] S. Deering et al, *The PIM Architecture for Wide-Area Multicast Routing*, *IEEE/ACM* Transactions on Networking, Vol. 4, No. 2, Apr. 1996, pp 153-162

[Devalla] B.Devalla et al., *Adaptive Connection Admission Control for Mission Critical Real-Time Communication Networks*

[DIFF-PIB] M. Fine, K. McCloghrie, J. Seligson, K. Chan, S. Hahn, C. Bell, A. Smith, F. Reichmeyer, *Differentiated Services Quality of Service Policy Information Base*, draft-ietf-DiffServ-pib-09.txt, June 2002

[DMTF] http://www.dmtf.org

[DRAFT-FLOWLABEL] J. Rajahalme, A. Conta, B. Carpenter, S. Deering, *IPv6 Flow Label Specification*, draft-ietf-ipv6-flow-label-04.txt, December 2002

Copyright © MESCAL Consortium, June 2005

[DRAFT-MBGP] T. Bates, R. Chandra, D. Katz, Y. Rekhter, *Multiprotocol Extensions for BGP-*, draft-ietf-idr-rfc2858bis-02.txt

[DRAFT-QOS-FLOW] H. Jagadeesan, T. Singh, *A Radical Approach in providing Quality-of-Service over the Internet using the 20-bit IPv6 Flow Label field*, draft-jagadeesan-rad-approach-service-01.txt, March 2002

[D1.1] Paris Flegkas et al., "Specification of Business Models and a Functional Architecture for Inter-domain QoS Delivery", MESCAL Deliverable D1.1, 19 June 2003.

[D1.2] M. Howarth, et al., "Initial specification of protocols and algorithms for inter-domain SLS management and traffic engineering for QoS-based IP service delivery and their test requirements," MESCAL project deliverable D1.2, available at ", available at: http://www.mescal.org/, January 2004.

[D1.4] H. Asgari et al., "Issues in MESCAL inter-domain delivery: technologies, bidirectionality, interoperability and financial settlements," MESCAL Deliverable 1.4, 30 January 2004.

[D3.2] MESCAL Deliverable D3.2

[Elek00] V. Elek et al., *Admission Control Based on End-to-End Measurements*, In Proc. of IEEE INFOCOM'00, Tel Aviv, Israel, March 2000.

[Elwal01] A. Elwalid, C. Jin, S H. Low, and I. Widjaja, *MATE: MPLS Adaptive Traffic Engineering*, In Proc. of IEEE INFOCOM2001, pp. 1300-1309, Alaska, USA, April 2001

[ETSI] European Telecommu), ETSI ES 201 915-1 *Open Service Access; Application Programming Interface (API); Part 1: Overview*, related standards available at http://portal.etsi.org

[ELIS04] Elisa Boschi, Salvatore D'Antonio, Giorgio Ventre, "Inter-domain Communication and Data Exchange", Inter-domain Performance and Simulation Workshop, Budapest, March 2004

[ERIC02] M. Ericsson, M.G.C. Resende and P.M. Pardalos, "A genetic algorithm for the weight setting problem in OSPF routing," *J. Combinatorial Optimization*, Vol. 6, No. 3, pp.299-333, 2002.

[EXTC05] R. Srihari, D. Tappan,Y. Rekhter, "BGP Extended Communities Attribute", draft-ietf-idr-bgp-ext-communities-08.txt, February 2005.

[Falou98] M. Faloutsos et al, QoSMIC: Quality of Service Sensitive Multicast Internet protocol, proc. ACM SIGCOMM 1998, pp144-153

[Farina98] D. Farinacci et al, Multicast Source Discovery Protocol (MSDP), Internet Draft, draft-farinacci-msdp-*.txt, Jul. 1998

[Fauch02] F. Le Faucheur, et al. Requirements for support of Diff-Serv-aware MPLS Traffic Engineering, IETF Internet draft, draft-ietf-tewg-diff-te-reqts-05.txt, work in progress, June 2002

[Feld01] A. Feldmann and J. Rexford, IP Network Configuration for Intradomain Traffic Engineering, IEEE Network Magazine, vol. 15, no. 5, pp. 46-57, September 2001

[Fenner02] B. Fenner, Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised), draft-ietf-pim-sm-v2-new-06.txt, December 2002

[Fenner97] W. Fenner, Internet Group management Protocol, version 2, RFC 2236, Nov. 1997

[Flegk02] Flegkas, P., Trimintzios, P., Pavlou, G., A Policy-based QoS Management System for IP DiffServ Networks, IEEE Network Magazine, March/April 2002

Copyright © MESCAL Consortium, June 2005

[FRWK-PIB] M. Fine, K. McCloghrie, J. Seligson, K. Chan, R. Sahita, S. Hahn, A. Smith, F. Reichmeyer, Framework Policy Information Base, draft-ietf-rap-frameworkpib-09.txt, June 2002

[Fu02] Xiaoming Fu, et al, Analysis on RSVP Regarding Multicast, Internet Draft, draft-fu-rsvp-multicast-analysis-01.txt), Expires April 2003.

[FEI01] Aiguo Fei et al, "Aggregated Multicast with Inter-Group Tree Sharing," *Proceedings of Third International Workshop on Networked Group Communications (NGC2001),* UCL, London, UK, November 7-9, 2001.

[FELD00] A. Feldman, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "NetScope: Traffic Engineering for IP Networks," IEEE Network Magazine, Vol. 14, No. 2, pp. 11-19, March/April 2000.

[FLO96] S.Floyd "Comments on Measurement-based Admission Control for Controlled-Load Services", July 1996, Lawrence Berkeley Laboratory Technical Report.

[FORT00] Fortz, B., and M. Thorup, "Internet traffic engineering by optimizing OSPF weights," paper presented at INFOCOM 2000. *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings*. IEEE, Vol.2, 2000.

[FORT02a] Fortz, B., and M. Thorup, "Optimizing OSPF/IS-IS weights in a changing world," *IEEE Journal Selected Areas in Communications*, May, 20, 756-767 2002.

[FORT02b] Fortz, B., J. Rexford, and M. Thorup, "Traffic engineering with traditional IP routing protocols," *IEEE Communications Magazine*, 40, 118-124 2002.

[FRED01] S.B.Fredj, S.Oueslati-Boulahia, and J.W.Roberts "Measurement-based Admission Control for Elastic Traffic," in *Proceedings of ITC17*, Sept. 2001.

[Gibbens99] R.J. Gibbens, F.P. Kelly, *Distributed connection acceptance control for a connectionless network*, 16th International Teletraffic Congress, Edinburgh, June 1999

[GEOR04] S.Georgoulas, P.Trimintzios and G.Pavlou "Joint Measurement- and Traffic Descriptor-based Admission Control at Real-Time Traffic Aggregation Points," to appear in *ICC 2004*.

[GEOR05] S.Georgoulas, P.Trimintzios, G.Pavlou and Kin-Hon Ho "Measurement-based Admission Control for Real-time Traffic in IP Differentiated Services Networks", IEEE ICT 2005.

[GODE02a] Danny Goderis et al., "Service Level Specification Semantics, Parameters and Negotiation Requirements," draft-tequila-diffserv-sls-02.txt, January 2002.

[GODE02b] D. Goderis et al, "A Scalable Service-Centric IP Quality of Service Architecture for Next Generation Networks", NOMS 2002.

[GODE02c] D. Goderis, S. Van D. Bosch, Y. T'joens, O. Poupel, C. Jacquenet, G. Memenios, G. Pavlou, R. Egan, D. Griffin, P. Georgatsos, L. Georgiadis, and P. Van Heuven, "Service Level Specification Semantics and Parameters," Internet draft: draft-tequila-sls-02.txt, Expired August 2002.

[GROS99] M.Grossglauser and D.Tse "A Framework for Robust Measurement-Based Admission Control", IEEE/ACM Transactions on Networking, June 1999.

[GROS03] M.Grossglauser and D.Tse "A Time-Scale Decomposition Approach to Measurement-Based Admission Control", IEEE/ACM Transactions on Networking, August 2003.

[GUER91] R.Guerin, H.Ahmadi, and M.Naghshieh "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 7, pp.968-98, September 1991.

[GUN92] R.Guerin, and L.Gun "A Unified Approach to Bandwidth Allocation and Access Control in Fast Packet-Switched Networks," *IEEE INFOCOM* 1992, Vol. 1, pp. 1-12.

Copyright © MESCAL Consortium, June 2005

[Hanna99] S. Hanna, B. Patel, M. Shah, *Multicast Address Dynamic Client Allocation Protocol (MADCAP)*, *RFC* 2730, Dec. 1999

[Hoag98] Hoagland, J. A., Padney, R., et al., *Security Policy Specification using a Graphical Approach*, UC Davis, Computer Science Department

[Holbro02] H. W. Holbrook, *Using IGMPv3 and MLDv2 For Source-Specific Multicast*, draft-holbrook-idmr-igmpv3-ssm-03.txt, November 2002

[Holbro99] H. W. Holbrook, D. R. Cheriton, *IP Multicast Channels: EXPRESS Support for Large-scale Single-source Applications*, Proc. *ACM SIGCOMM'99*

[HOFM04] U. Hofmann, I. Miloucheva, and T.Pfeiffenberger, "INTERMON: Complex QoS/SLA Analysis in large scale Internet environment", WISICT 2004, Proceedings of the Winter International Symposium on Information and Communication Technologies, Cancun, Mexico, January 5th-8th, 2004.

[HUST] G. Huston, "ISP Survival Guide," John Wiley and Sons 1998. ISBN 201-3-45567-9

[IETF] Internet Engineering Task Force (IETF www.ietf.org, related work groups and drafts).

[IPDR] ipdr.org, *Service Specifications*, related documents, http://www.ipdr.org

[IPTE-ACC]M. Boucadair, *An IP Traffic Engineering PIB for Accounting purposes"*, "draft-boucadair-ipte-acct-pib-01.txt, December 2002

[IPTE-CT] C. Jacquenet, *A COPS client-type for IP traffic engineering*, draft-jacquenet-ip-te-cops-04.txt, January 2003

[IPTE-PIB] M. Boucadair, C. Jacquenet, *An IP Traffic Engineering Policy Information Base*, draft-jacquenet-ip-te-pib-02.txt, June 2002

[Ish01] K. Ishiguro, T. Takada, *Traffic Engineering Extensions to OSPFv3"*, "draft-ishiguro-ospfv3-01.txt, October 2002

[Ivars] I.M. Ivars, G. Karlsson, *PBAC: Probe-Based Admission Control*

[IANN01] G.Iannaccone, M.May, and C.Diot "Aggregate Traffic Performance with Active Queue Management and Drop from Tail", Computer Communications Review, July 2001.

[IDR] http://www.ietf.org/html.charters/idr-charter.html.

[IST] European IST (Information Society Technologies) research projects, for more information visit: www.cordis.lu/ist/. Specifically for IST-INTERMON visit: www.ist-intermon.org/, for IST-MoMe visit: www.ist-mome.org/, and for IST-SCAMPI visit: www.ist-scampi.org/, for IST-6QM visit: http://www.6qm.org/index.php.

[Jain] The Parlay Group, The Jain APIs: Integrated May 2002, related documents available at http://java.sun.com/products/jain/

[Jia97] X. Jia et al, Group Multicast Routing Algorithm by Using Multiple Minimum Steiner Trees, Computer Communications 20 (1997) pp750-758

[Jia98] X. Jia, A Distributed Algorithm of Delay-bounded Multicast Routing for Multimedia Applications in Wide Area Networks, IEEE/ACM Transactions on Networking 6(6): 828-837 (1998)

[Johnston03] Johnston A. et al, Session Initiation Protocol Private Extension for an OSP Authorization Token, IETF Internet Draft draft-johnston-sip-osp-token-04.txt, February 2003

[Juniper] Juniper, Junos software documentation

[JOVE01] L.Fabrega, T.Jove, A.Bueno, J.L.Marso "An Admission Control Approach for Elastic Flows in the Internet," *9th IFIP Working Conference on Performance Modelling and Evaluation of ATM & IP Networks*, Budapest, June 2001.

Copyright © MESCAL Consortium, June 2005

[Kau02] Kaur, H., Kalyanaraman, S., *A Connectionless Approach to Intra- and Inter-Domain Traffic Engineering*, 2nd New York Metro Area Networking Workshop, September 2002.

[Kelly] F.P. Kelly, P.B. Key, S. Zachary, *Distributed Admission Control*

[Knight99] E.W. Kinghtly, N.B. Shoft, *Admission Control for Statistical QoS: Theory and Practice*, IEEE Network, March 1999

[Kohli99] Kohli, M., Lobo, J., *Policy-based Management of telecommunication Networks*, in the Proc. of IEEE Policy Workshop (Policy99), HP Labs, Bristol, U.K., 1999

[Kompe93] V. P. Kompella et al, *Two Distributed Algorithms for multicasting multimedia Information*, Proc. *IEEE ICCCN'1993*, pp343-349

[Kou81] L. Kou et al, *A Fast Algorithm for Steiner Trees*, Acta Inormatica 15 (1981) pp141-145

[KPP93] V. P. Kompella et al, *Multicast Routing for Multimedia Communication*, IEEE/ACM Transaction on Network 1993, pp286-292

[Kummar99]S. Kummar et al, *The MASC/BGMP Architecture for Inter-domain Multicast Routing*, Proc. *ACM SIGCOMM'99*

[KEY04] Keyur Patel, Susan Hares, "Aspath Based Outbound Route Filter for BGP-4", draft-ietf-idr-aspath-orf-07.txt, December 2004

[KODIA03] M. Kodialam et al, "Online Multicast Routing with Bandwidth Guarantees: A new Approach Using Multicast Network Flow," *IEEE/ACM* Trans. on Networking, Vol. 11, No. 4, pp676-686, 2003.

[Lobo99] Lobo, J., Bhatia, R., et al., *A Policy Description Language*, in Proc. of AAAI, Orlando, Florida, 1999

[Low00] C. P. Low et al, *An Efficient Algorithm for Group Multicast Routing Problem with Bandwidth Reservations*, Computer Communications, Vol. 23(18) (2000) pp1740-1746.

[Low02] C. P. Low et al, *On Finding Feasible Solutions for the Delay Constrained Group Multicast Routing Problem*, *IEEE* Transactions on computers, vol. 51 No. 5 (2002) pp581-588

[Lupu97] Lupu, E., Sloman, M., *Towards a Role Based Framework for Distributed Systems Management*, Journal of Networks and Systems Management (JNSM), Vol. 5, no. 1, 1997

[LIMA04] S.Lima, P.Carvalho and V.Freitas "Distributed Admission Control for QoS and SLS Management", Journal of Network and Systems Management, September 2004.

[LIN03] X.-H. Lin, Y.-K. Kwok and V.K.N. Lau, "A genetic algorithm based approach to route selection and capacity flow assignment," *Computer Communications*, Vol. 26, pp.961-974, 2003.

[LOBST] IST-LOBSTER Home Page, at http://www.ist-lobster.org/about/objectives.html.

[MAS99] L.Massoulié and J.W.Roberts "Arguments in Favour of Admission Control for TCP Flows," In *proc. ITC 16*, Edinbourg, June 1999.

[MAY01] G.Iannaccone, M.May, and C.Diot "Aggregate Traffic Performance with Active Queue Management and Drop from Tail," *Computer Communications Review*, July 2001.

[MEN04] M.Menth "Efficient Admission Control and Routing for Resilient Communication Networks", PhD Thesis, University of Wurzburg, July 2004.

[MESCAL] The IST MESCAL project; website: www.mescal.org.

[Marr96] Marriot, D., Sloman, M., *Implementation of a Management Agent for Interpreting Obligation Policy*, in Proc. of the seventh IFIP/IEEE International Workshop on

Copyright © MESCAL Consortium, June 2005

Distributed Systems: Operations & Management (DSOM'96), L'Aquila, Italy, October 28-30, 1996.

[Martin02] Martinez, P., Brunner, M., et al., *Using the Script Mib for Policy-based Configuration Management*, in the Proc of the IEEE Network Operations and Management Symposium (NOMS'02), Florence, Italy, April 2002.

[Mayer00] D. Mayer, P. Lothberg, *GLOP Addressing in 233/8*, *RFC* 2770, Feb. 2000

[Mitra99a] D. Mitra, and K. G. Ramakrishnan, *A Case Study of Multiservice, Multipriority Traffic Engineering Design for Data Networks*, In Proc. IEEE GLOBECOM 99, pp. 1087-1093, Brazil, December 1999

[Mitra99b] D. Mitra, J.A. Morrison and K.G. Ramakrishnan, *Virtual Private Networks: Joint Resource Allocation and Routing Design*, In Proc. IEEE INFOCOM 99, USA, March 1999

[Mortier00] R. Mortier, et al., *Implicit Admission Control*, IEEE Journal on selected areas in communications, vol. 18, no. 12, December 2000

[Moy94] J. Moy, *Multicast Extensions to OSPF*, *RFC* 1584, Mar. 1994

[MSK+02] J. Manner, T. Suihko, M. Kojo, M. Liljeberg, K. Raatikainen, *Localized RSVP*. Internet Draft draft-manner-lrsvp-01.txt, Expires July 2003.

[Mykon03] E.Mykoniati et al., *Admission Control for Providing QoS in DiffServ IP Networks: The TEQUILA approach* IEEE Communications Magazine, January 2003

[MON02] T. Monk "Inter-domain Traffic Engineering: Principles and Case Examples", INET 2002.

[MOW98] M.Mowbray, G.Karlsson and T.Kohler "Capacity Reservation for Multimedia Traffics", Distr. Syst. Eng., 1998.

[MYK03] E.Mykoniati, C.Charalampous, P.Georgatsos, T.Damilatis, D.Goderis, P.Trimintzios, G.Pavlou, and D.Griffin "Admission Control for Providing QoS in Diffserv IP Networks: The TEQUILA Approach," *IEEE Communications Magazine*, January 2003, Vol. 41, No. 1, pp. 38-44.

[NIC99] K. Nichols, V. Jacobson, L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," RFC 2638, July 1999.

[NLANR] Network Analysis Infrastructure (NAI) projects of National Laboratory for Applied Network Research (NLANR), for more information visit: http://mna.nlanr.net/infrastructure.html.

[PAXS98] V. Paxson, J. Mahdavi, A. Adams,, and M. Mathis "An Architecture for Large-Scale Internet Measurement" IEEE Communications Magazine, vol. 36 no. 8, pp. 48-54, August 1998.

[PRAT00] R.Mortier, I.Pratt, C.Clark, and S.Crosby "Implicit Admission Control," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 12, December 2000.

[PRZYG03] T. Przygienda et al, "*M-ISIS: Multi Topology (MT) Routing in IS-IS,*" draft-ietf-isis-wg-multi-topology-06.txt, March 2003 work in progress.

[PSAMP] Internet Engineering Task Force (IETF), for information on the various IETF working groups including DiffServ, IPPM, IPFIX, RTFM, PSAMP etc., visit: www.ietf.org.

[PSEN05] P. Psenak, et. al., "Multi-Topology (MT) Routing in OSPF", Network Working Group, Internet-Draft, April 2005

[Pelsser02] Cristel Pelsser (FUNDP), Olivier Bonaventure (UCL), RSVP-TE extensions for inter-domain LSPs, Internet draft, Work in Progress, October 2002.

Copyright © MESCAL Consortium, June 2005

[PolicyWG] http://www.ietf.org/html.charters/policy-charter.html

[Poppe00] F. Poppe, et al. Choosing the Objectives for Traffic Engineering in IP Backbone Networks Based on Quality-of-Service Requirements, In Proc. Workshop on Quality of future Internet Services (QofIS'00), pp. 129-140, Germany, September 2000

[PPVPN-PIB] Y. El Mghazli, BGP/MPLS VPN Policy Information Base, draft-yacine-ppvpn-2547bis-pib-01.txt, July 2002

[QPIM] Snir, Y., Ramberg, Y., et al., Policy QoS Information Model, internet-draft, IETF, Novemeber 2001

[RFC 2597] J. Heinanen, et. el., *Assured Forwarding PHB Group*, June 1999.

[RFC 2638] K. Nichols, V. Jacobson, L. Zhang, *A Two-bit Differentiated Services Architecture for the Internet*, July 1999.

[RFC 3086] K. Nichols, B. Carpenter, *Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification*, April 2001.

[RFC 3140] S. Brim, B. Carpenter, F. Le Faucheur, *Per Hop Behavior Identification Codes*, June 2001.

[RFC 3246] B. Davie, et al, *An Expedited Forwarding PHB (Per-Hop Behavior)*, March 2002.

[RFC 3317] K. Chan, R. Sahita, S. Hahn, K. McCloghrie, *Differentiated Services Quality of Service Policy*, Informational RFC, March 2003.

[RFC1771] Y. Rekhter, T. Li, *A border gateway protocol 4 (BGP-4)*, RFC 1771, March 1995

[RFC2205] Braden, R., Zhang, L., Berson, S., Herzog, S. and S. Jamin, *Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification*, RFC 2205, Sep 1997.

[RFC2207] L. Berger and T. O'Malley, *RSVP Extensions for IPSEC Data Flows*, RFC 2207, September 1997.

[RFC2328] J., Moy, *OSPF Version 2*, RFC 2328, April 1998.

[RFC2380] Berger, L., *RSVP over ATM Implementation Requirements*, RFC 2380, August 1998.

[RFC-2460] S. Deering, R. Hinden, *Internet Protocol, Version 6 (IPv6) Specification*, RFC 2460, Standards Track, December 1998

[RFC-2474] K. Nichols, et. al., *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers*, December 1998

[RFC2475] S. Blake, et. al., *An Architecture for Differentiated Services*, December 1998.

[RFC2573] Yavatkar, R., Pendarakis, D., Guerin, D., *A Framework for Policy Based Admission Control*, Informational RFC 2753, January 2000

[RFC2702] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, J. McManus, *Requirements for Traffic Engineering Over MPLS,* RFC 2702, September 1999.

[RFC2740] R. Coltun, D. Ferguson, J.Moy, *OSPF for IPv6*, RFC2740, December 1999.

[RFC2745] Terzis, A., Braden B., S. Vincent, and L. Zhang, *RSVP Diagnostic Messages*, RFC 2745, January 2000.

[RFC2746] Terzis, A., Krawczyk, J., Wroclawski, J. and L. Zhang, *RSVP Operation Over IP Tunnels*, RFC 2746, January 2000.

[RFC2748] Durham D., et al, *The COPS (Common Open Policy Service) Protocol*, IETF RFC 2748, January 2000

Copyright © MESCAL Consortium, June 2005

[RFC2814] Yavatkar, R., Hoffman, D., Bernet, Y., Baker, F. and M. Speer, *SBM (Subnet Bandwidth Manager): A Protocol for Admission Control over IEEE 802-style Networks*, RFC 2814, May 2000.

[RFC2842] R. Chandra, J. Scudder, *Capabilities Advertisement with BGP-4*, RFC 2842, Standards Track, May 2000

[RFC2842] R. Chandra, J. Scudder, *Capabilities Advertisement with BGP-4*, RFC 2842, Standards Track, May 2000

[RFC2961] Berger, L., Gan, D., Swallow, G., Pan, P. and F. Tommasi, *RSVP Refresh Reduction Extensions*, RFC 2961, April 2001.

[RFC2996] Bernet, Y., *Format of the RSVP DCLASS Object*, RFC 2996, November 2000.

[RFC3060] Moore, B., Elleson, E., et al., *Policy Core Information Model – Version 1 Specification*, Standard-Tracks RFC 3060, IETF, February 2001.

[RFC-3084] K. Chan, J. Seligson, D. Durham, S. Gai, K. McCloghrie, S. Herzog, F. Reichmeyer, R. Yavatkar and A. Smith, *COPS Usage for Policy Provisioning*, RFC 3084, March 2001

[RFC3107] Y. Rekhter (Juniper Networks), E. Rosen (Cisco Systems, Inc.), *Carrying Label Information in BGP-4*, Network Working Group, RFC 3107, IETF, May 2001.

[RFC-3159] K. McCloghrie, M. Fine, J. Seligson, K. Chan, S. Hahn, R. Sahita, A. Smith, F. Reichmeyer, *Structure of Policy Provisioning Information (SPPI)*, RFC 3159, August 2001

[RFC3175] F. Baker, C. Iturralde, F. Le Faucheur, B. Davie, *Aggregation of RSVP for IPv4 and IPv6 Reservations*, RFC 3175, Sept. 2001.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V. and G. Swallow, *Extensions to RSVP for LSP Tunnels*, RFC 3209, Dec. 2001.

[RFC3219] Rosenberg J. et al, *Telephone Routing over IP (TRIP),* IETF RFC 3219, January 2002

[RFC3261] Rosenberg J. et al, *SIP: Session Initiation Protocol,* IETF RFC 3261, June 2002

[RFC3270] F. Le Faucheur (ed), L. Wu, and et al*, Multi-Protocol Label Switching (MPLS) Support of Differentiated Services*, RFC 3270, May 2002.

[RFC3272] D. Awduche, et al. *Overview and Principles of Internet Traffic Engineering*, IETF Informational RFC-3272, May 2002

[RFC3460] Moore, B., *Policy Core Information Model Extensions (PCIMe)*, Standards-Track RFC 3460, IETF, January 2003

[Ribe01] Ribeiro, C., Zuquete, A., et al., *SPL: An access control language for security policies with complex constraints*, in the Proc. of Network and Distributed System Security Symposium (NDSS'01), San Diego, California, 2001

[Rie01] A., Riedl, D., Schupke, *A Flow-Based Approach for IP Traffic Engineering Utilizing Routing Protocols With Multiple Metric Types***,** Sixth INFORMS Telecommunications Conference, March 2002

[Rouska97] G. N. Rouskas, et al, *Multicast Routing with End-to-end Delay and Delay Variation Constraints*, *IEEE* Journal on Selected Areas in Communications 15(3): 346-356 (1997)

[RSVP] Zhang, L., Deering, S., Estrin, D. and D. Zappala, *RSVP: A New Resource Reservation Protocol*, IEEE Network, Volume 7, Pages 8-18, Sep 1993.

[RIED02] Riedl, A., "A Hybrid Genetic Algorithm for Routing Optimization in IP Networks Utilizing Bandwidth and Delay Metrics," *Proceedings IEEE Workshop on IP Operations and Management (IPOM)*, Dallas, USA, 2002.

Copyright © MESCAL Consortium, June 2005

[ROB98] J.W.Roberts and L.Massoulié. "Bandwidth Sharing and Admission Control for Elastic Traffic," In *Proc. ITC Specialist Seminar*, Yokohama, October 1998.

[Salsano01] Salsano S., *COPS Usage for DiffServ Resource Allocation (COPS-DRA*", IETF Internet Draft <draft-salsano-cops-dra-00.txt>, October 2001

[Sargen] Susana Sargento et al., *Call Admission Control in IP networks with QoS support*

[SIBBS] QBone Signalling Design Team - *Final report*, http://qos.internet2.edu/wg/documents-informational/20020709-chimento-etal-qbone-signaling/

[SIG1] Manner (ed.), J. X. Fu, *Analysis of Existing Quality of Service Signalling Protocols*, <draft-ietf-nsis-signalling-analysis-00.txt> Expires April, 2003.

[SIG2] H.de meer, et al., *Analysis of Existing QoS solutions*, <draft-demeer-nsis-analysis-03.txt>, Expires May 2003.

[Sinnreich00] Sinnreich H., Donovan S., Rawlins D., *Inter-domain IP Communications with Qos, Authroization and Usage Reporting,* IETF Internet Draft <draft-sinnreich-sip-qos-osp-01.txt>

[Slom94] Sloman, M., *Policy Driven Management For Distributed Systems*, Journal of Network and Systems Management, Vol. 2, No. 4, pp. 333-360, Plenum Publishing, December 1994

[Slom99] Sloman, M., Lupu, E., *Policy Specification for Programmable Networks*, in Proc. of the 1st International Conference on Active Networks, Berlin, Germany, June 1999

[Stoica99] I. Stoica and H. Zhang, *Providing Guaranteed Services without per Flow Management*, in Proceedings of ACM SIGCOMM'99, Cambridge, MA, August 1999.

[Stone01] Stone, G. N., Lundy, B., et al., *Network Policy Languages: A Survey and a New Approach*, IEEE Network Magazine, pp 10-21, Jan/Feb 2001

[Stream01] D. Wu, et. al., *Streaming Video, over the Internet: approaches and Directions*, IEEE Transcations on circuits and systems for video technology, Vol. 11, No. 1, Feb 2001.

[Strieg01] A. Striegel, G. Manimaran, *A scalable approach for DiffServ Multicasting*, proc. *IEEE ICC* 2001

[Suri01] S. Suri, et al., *Profile-based Routing: A New Framework for MPLS Traffic Engineering*, In Proc. of the 2nd International Workshop on Quality of future Internet Services (QofIS'01), pp. 138-157, Portugal, September 2001

[SHRO03] D.Eun and N.Shroff "A Measurement-Analytic Approach for QoS Estimation in a Network Based on the Dominant Time Scale," *IEEE/ACM Transactions on Networking*, April 2003, Vol. 11, No. 2, pp. 222-235.

[Taka80] H. Takahashi, A Mastuyama, *An Approximate Solution for the Steiner Problem in Graphs*, Math. Japonica 6, pp573-577

[Thaler00] D. Thaler et al, *The Internet Multicast Address Allocation Architecture*, *RFC* 2908, Sept. 2000

[Thaler02] D. Thaler, *Border Gateway Multicast Protocol (BGMP): Protocol Specification*, draft-ietf-bgmp-spec-03.txt, July 2002

[Thom01] Michael Thomas, *Analysis of Mobile IP and RSVP Interactions*, draft-thomas-seamoby-rsvp-analysis-00.txt, Issued Oct. 2002.

[TINA] Telecommunications Information Network Architecture (TINA) Consortium, related documents available at www.tinac.org, 1996-2000.

[TMF] TeleManagement Forum, *Telecom Operations Map*, March 2000; HYPERLINK "http://www.tmforum.org" www.tmforum.org

Copyright © MESCAL Consortium, June 2005

[TMN] ITU-T TMN Recommendation M.3400 *TMN Telecommunication Management Network*, M.3200 TMN Management Services, and related recommendations.

[Trimin01] P. Trimintzios et al., *A Management and Control Architecture for Providing IP Differentiated Services in MPLS-based Networks,* IEEE Commun. Mag., vol. 39, no. 5, May 2001.

[Tsch02] Hannes Tschofenig, *RSVP Security Properties*, Internet Draft draft-tschofenig-rsvp-sec-properties-00.txt, Expired Nov. 2002.

[TEQUILA] IST TEQUILA. www.ist-tequila.org.

[TEQUI,D1.4] P. Van Heuven et al., "D1.4: final architecture, protocol and algorithm specification," TEQUILA, EU IST Project IST-1999-11253, 30 April 2002.

[TSE97] D.Tse and M.Grossglauser "Measurement-based Call Admission Control: Analysis and Simulation", IEEE INFOCOM 1997.

[TSE99] M.Grossglauser and D.Tse "A Framework for Robust Measurement-Based Admission Control," *IEEE/ACM Transactions on Networking*, June 1999, Vol. 7, No. 3, pp.293-309.

[TTM] Test Traffic Measurements (TTM) project of RIPE (Réseaux IP Européens) Network Coordination Centre (NCC), for more information visit: http://www.ripe.net/ttm/.

[Uhlig00] Steve Uhlig, Olivier Bonaventure, On the Cost of using MPLS for inter-domain traffic, COST263 workshop, September 2000.

[Vasseur00] J.P.Vasseur, et. al., Definition of an RRO node-id subobject, Internet Draft, draft-vasseur-mpls-nodeid-subobject-00.txt, Work in Progress, February 2003

[Vasseur01] J.P. Vasseur, Y. Ikejiri, Reoptimization of an explicit loosely routed MPLS TE paths, Internet Draft, draft-vasseur-mpls-loose-path-reopt-01.txt, Work in Progress, February 2003

[Vasseur02] J.P. Vasseur, et. al., RSVP Path computation request and reply messages, Internet Draft, draft-vasseur-mpls-computation-rsvp-03.txt, June 2002

[Vasseur03] Jean-Philippe Vasseur, Raymond Zhang, Inter-AS MPLS Traffic Engineering, Internet Draft, draft-vasseur-inter-as-te-00.txt, Work in Progress, February, 2003

[Veltri02] Veltri, L., Salsano., S and Papalilo, D., SIP Extensions for QoS Support, IETF Internet draft <draft-veltri-sip-qsip-01.txt> , October 2002

[Vida02] R.Vida et al, Multicast Listener Discovery Version 2 (MLDv2) for IPv6, draft-vida-mld-v2-06.txt, November 2002

[W3C] World Wide Web Scenarios, June 2002; related documents available at http://www.w3.org

[Waitz88] D. Waitzman, C. Partridge, S. Deering, *Distance Vector Multicast Routing Protocol (DVMRP)*, *RFC* 1075, Nov. 1988

[Wang01] Z. Wang, Y. Wang, and L. Zhang, *Internet Traffic Engineering without Full Mesh Overlaying*, In Proc. of IEEE INFOCOM 2001, Alaska, April 2001

[WALT02] Walton, D., et al., "Advertisement of Multiple Paths in BGP", draft-walton-bgp-add-paths-01.txt, Work in Progress, November 2002.

[WANG04] Jun Wang, Yaling Yang, Li Xiao, and Klara Nahrstedt, ``Edge-based Traffic Engineering for OSPF Networks," Elsevier Journal on Computer Networks, Vol. 48/4, pp. 605-625, 2005.

[Xia01] Xiao, L., Shan-Lui, K., Wang, J., Nahrstedt, K., *QoS Extension to BGP*, UIUCDCS-R-2002-2295, September 2002.

[Xuan] D. Xuan et al., *Utilization-Based Admission Control for Real-Time Applications*

Copyright © MESCAL Consortium, June 2005

[Yang01] D. Yang et al, *MQ: An Integrated Mechanism for Multimedia Multicasting*, IEEE Transactions on multimedia, vol.3, no.1, March 2001

[Zhang] Zhi-Li Zhang et al., *Decoupling QoS Control from Core Routers: A novel Bandwidth Broker architecture for scalable support of Guaranteed Services*

[Zhang03] Raymond Zhang, JP Vasseur, *MPLS Inter-AS Traffic Engineering requirements*, Internet draft, draft-zhang-mpls-interas-te-req-02.txt, Work in Progress, February 2003

[Zhu95] Q. Zhu, et al, *A Source Based Algorithm for Delay-constrained Minimum-cost Multicasting*, in proc. of IEEE INFOCOM'95

Copyright © MESCAL Consortium, June 2005

15 APPENDIX – INTERNET DRAFTS ON PCS

15.1 Path Computation Service discovery via Border Gateway Protocol

PCE Working Group
Internet Draft

M. Boucadair (Ed.)
P. Morand (Ed.)
France Telecom R&D

Document: draft-boucadair-pce-discovery-01.txt
Category: Standards Track

May 2005

Path Computation Service discovery via Border Gateway Protocol
< draft-boucadair-pce-discovery-01.txt >

Status of this Memo

This document is an Internet-Draft and is subject to all provisions of section 3 of RFC 3667 [RFC3667]. By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she become aware will be disclosed, in accordance with RFC 3668 [RFC3668].

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on November 2005.

Abstract

This draft describes a simple mechanism that ease discovery of remote Autonomous Systems (AS) supporting inter-domain MPLS-based constrained tunnels service (this service is also denoted by Path Computation Service (PCSv)) thanks to the use of Path Computation Elements (PCEs). Remote ASs could be managed by a single or distinct Internet Network Providers (INP).
Particularly, this draft describes how Border Gateway Protocol (BGP) is used to announce Path Computation Service unique identifiers

Copyright © MESCAL Consortium, June 2005

across the Internet in order for other PCEs to be able to discover a path towards every AS supporting this Path Computation Service.

Table of Contents

1. Contributors

- Hamid Asgari (Thales Research and Technology)
- Panagiotis Georgatsos (Algonet)
- David Griffin (University College London)
- Micheal Howarth (University of Surrey)
- Noel Cantenot (France Telecom)

2. Changes since last version:

The main changes occurred in this version are:
- Rewording of several sections of the draft

3. Terminology

This memo makes use of the following terms:

- Path Computation Element (PCE): an entity that is responsible for computing/finding inter/intra domain paths for establishing LSPs. This entity can simultaneously act as client and a server. Several PCEs could be deployed in a given AS.

Copyright © MESCAL Consortium, June 2005

- o Path Computation Client (PCC): a PCE acting as a client. This entity is responsible for issuing path computation requests that fulfill the Service Management constraints for the establishment of inter/intra domain LSPs.

- o Path Computation Server (PCS): a PCE acting as a server. This entity is responsible for handling path computation requests in order to satisfy PCC constraints.

- o High-level service: is the service using a PCE-based system as an underlying infrastructure (an inter-domain QoS VPNs service for instance)

- o High-level service customer: is a customer that subscribes to a High-level service.

- o pSLS: A provider SLS is an SLS established between two Internet Network Providers (INP) with the purpose of extending the geographical span of their service offers.

- o SLS Management: This management entity is responsible for SLS-related activites, including pSLS ordering (i.e establishing contracts between peers) and SLS invocation (i.e committing resources before traffic can be admitted)

- o q-BGP: QoS-inferred BGP. A modified BGP protocol that takes into account QoS information as input for its route selection process.

- o Domain: within this draft it denotes an Autonomous system.

4. Introduction

4.1. General

Recently, several proposals describing the use of a Path Computation Element (PCE) as additional element to existent IP network entities have been submitted to the IETF. The main objective of introducing a PCE element is to ease computation of constrained paths in sophisticated schemes like inter-domain (both in intra-provider or inter-provider) and then driving the establishment of inter-domain LSPs.

A framework for establishing and controlling Multi-Protocol Label Switching Protocol (MPLS) and Generalized MPLS (GMPLS) Label Switching Paths (LSPs) in multi-domain networks has been defined in [CCAMP-FWRK]. The notion of domain in this framework draft encloses both Interior Gateway Protocol (IGP) areas and Autonomous System (AS) contrary to the current draft that restricts the notion of domain to a single AS.

Copyright © MESCAL Consortium, June 2005

Another draft that proposes a solution to compute inter-domain constrained paths has been submitted to the IETF [INTERAS-PCE]. This draft takes into account the inter-provider specific service considerations. In addition, the draft [INTERAS-PCP] describes a new protocol allowing communication between two PCEs located in different domains in order to compute inter-domain paths satisfying a set of constraints.

All aforementioned drafts require a Path Computation Service (PCSv) discovery function that allows discovery of remote ASs supporting this type of service (the path computation service could be implemented by one or several PCE elements) together with theirs associated capabilities like QoS capabilities, inter-domain bandwidth, reachable IP prefixes, type of links, etc. Discovery of such capabilities could also be passive and be restricted to a simple service advertisement (like web-pages). PCSv locations and associated capabilities discovery depends on providers search. We will refer to this method as passive discovery method.

It is evident that passive method allows finding remote PCSv locations and their associated capabilities, but this information is not usable alone within a distributed PCE architecture, when a set of end-to-end constraints must be satisfied. Therefore, computation of end-to-end constraints must be achieved based on advertised individual PCE capabilities. The knowledge of the PCE path is then mandatory in order to deduce the end-to-end capabilities.

In this draft, we present a simple method that allows discovery of remote PCSv with their associated capabilities. This method will also help the PCE decision-making process to choose the next PCE to contact in order to optimize paths towards a given destination.

4.2. Structure of the draft

This draft is structured as follows:

- Section 5 gives an overview of the service approach;
- Section 6 argues on the need of PCSv discovery functions;
- Section 7 presents a solution proposal for PCSv discovery.

5. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

Copyright © MESCAL Consortium, June 2005

6. PCE discovery within a single domain

Within a single domain, discovery of PCSv location and capabilities could be achieved for instance thanks to the activation of the Service Location Protocol (SLP, [RFC2608]). This protocol allows discovery of activated services that uses client/service architecture. SLP defines the same framework for all services.

In order to use SLP as a means to discover PCSvs, a PCE Service Type Template SHOULD be defined.

7. Overview of the service approach

Neighboring domains establish pSLSs (a pSLS is an enhanced SLS agreement between two providers. SLS template is defined in [SLS]) between them in order to have appropriate rights to request establishment of LSPs. An inter-domain routing protocol runs between the domains (for instance Border Gateway Protocol (BGP, [RFC1771])). The LSP creation request is propagated downstream to appropriate PCEs. The requests include the AS's ASBR and the tail-end address of the LSP. This procedure is repeated until the request reaches the destination PCE.

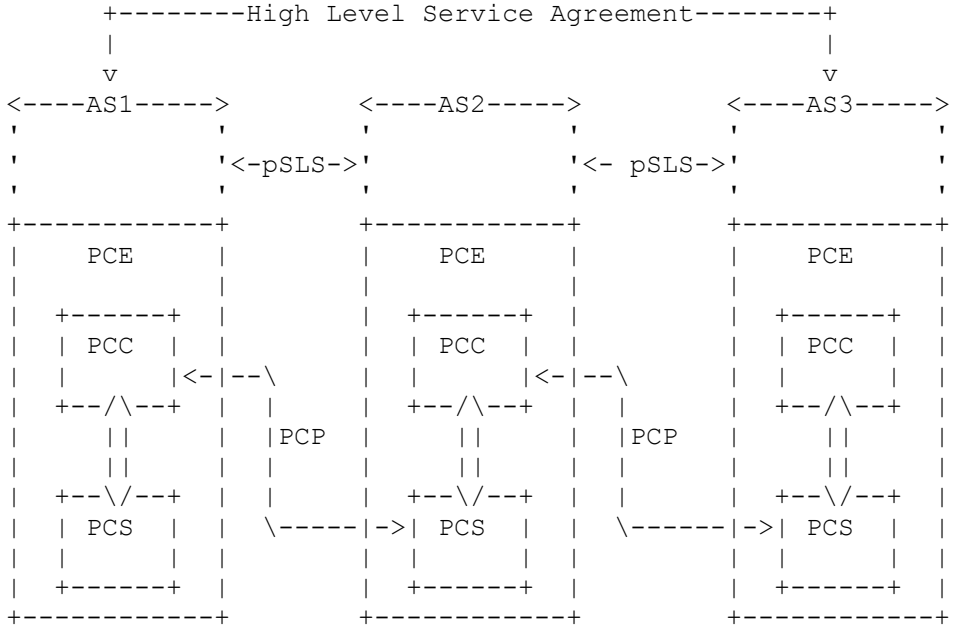


Figure 1: Service Overview

After authenticating the identity of LSP originating PCE, the destination PCE send a reply message back to the downstream domain's

Copyright © MESCAL Consortium, June 2005

PCE accepting the request and include the LSP loose path (destination, ASBR) addresses in the message. The next downstream domain's PCE does the same and adds its own relevant ASBR addresses to the loose path. The originating PCE inserts its intra-domain path and then initializes an RSVP reservation request for LSP establishment using the returned loose path.

At the service/application level (in order to differentiate this service from extending scope of IP connectivity service, we will denote it as high level service), when originating AS wants to establish an LSP to a destination in a remote ASs, there MUST be an agreement between the two ASs.

8. Service Advertisement and Discovery

Within this draft, we make a difference between the Service Advertisement and Discovery (SAD) and PCSv discovery function. SAD is a function that is achieved before establishing a service agreement between two peers. The SAD operation consists mainly at advertising/learning from/to the rest of the Internet the capabilities supported by a given AS in term of offered services (like Inter-domain LSP establishment service). PCSv advertisement is conditioned by the existence of a pSLS between two peers.

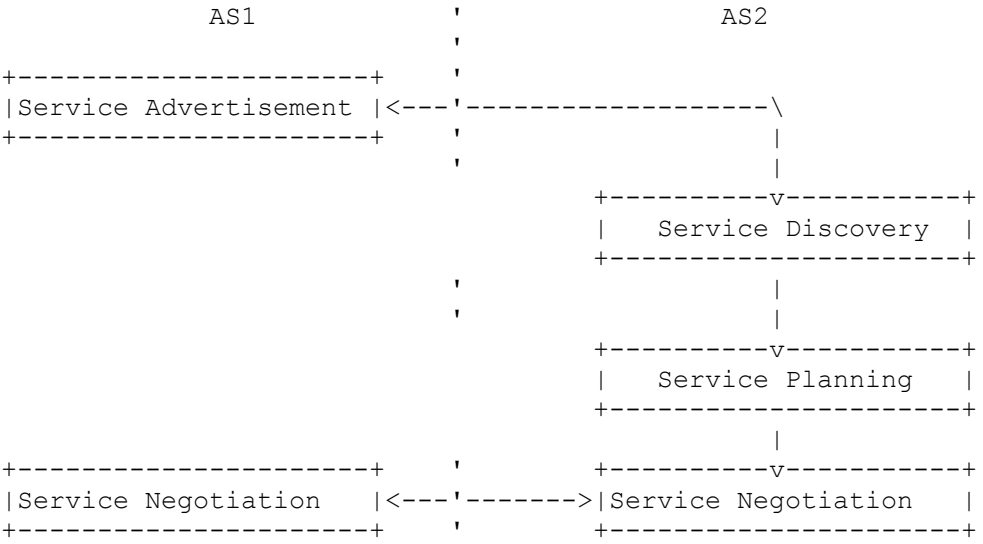


Figure 2: Service Advertisement and Discovery

Local Service Discovery block is responsible for finding remote offered services that is an essential input for Service Planning block. This functional block is responsible for choosing from discovered offered services the ones that will be used in order to

Copyright © MESCAL Consortium, June 2005

build it own services. Thus, a negotiation process SHOULD start and an SLS MAY be agreed between the two parties.

During service negotiation between two Service Providers, they MAY exchange their PCE reachability information and associated capabilities. Theses capabilities could include the following:

- o Supported Computation algorithms
- o Types of Constraints (e.g. QoS)
- o Set of attributes for a given constraint (one-way delay, one-way delay variation…)
- o Support of P2MP path computation techniques,

As a consequence each INP has a full knowledge of the PCE capabilities of its adjacent providers.

9. Why PCE discovery is needed

Path Computation elements are responsible for finding inter-domain paths satisfying a set of constraints (like QoS performance guarantees) to establish inter-domain constraint-based LSPs. The computation of this path is distributed and needs PCEs from different domains to communicate. Communication between two PCE entities is enabled thanks to inter PCE Communication Protocol (PCP) [INTERAS-PCP].

When receiving a request from the "High-Level" Service Management to compute/find a path towards a given tail-end address, the local PCE has to determine the next PCE to contact. In the worst case, local PCE can contact all its neighboring PCEs that are known to Service Management System. Nevertheless, it has no criteria to choose between those PCEs the next PCE to be contacted in order to send its path computation request. The risk of a request failure is then important.

In order to help the PCE decision-making process to choose the next PCE to be contacted, local PCE need to discover remote PCSvs reachable beyond the immediate neighbor PCEs. This information will help the next hop PCE decision. PCE need at least access to intra and inter-domain Routing Information Bases (RIB) in order to check the reachability status of destination prefixes if are propagated thanks to routing protocols.

10. Solution for PCSv discovery

Within this draft, we assume that during service negotiation phase between two peers, they MUST exchange IP addresses of their PCE(s). SLS Management Systems of the two peers MUST store this information.

Copyright © MESCAL Consortium, June 2005

In order to help the PCE computation process, routing information MUST be made available for the PCE. Thus, reachability information associated with capabilities (like QoS intra and/or inter-domain capabilities) SHOULD be propagated in the routing level. In the case of QoS-based service, each potential tail-end address (practically all routers interfaces) SHOULD be announced in all offered QoS Class plans (i.e. as many as used DSCP values). As a consequence, routing tables sizes will drastically increase.

From this perspective, instead of announcing all potential tail-end addresses in BGP, only an identifier needs to be announced. It is called the Path Computation Service Identifier (PCSID). This particular BGP announcement is identified by a well-known community value (to be defined be IANA) and is represented by a routable IP address, which can be different from the real IP address of the PCE.

As a consequence, this particular route SHOULD NOT be installed in the Forwarding Information Base (FIB) since this PCSID is not necessarily the IP address of the PCE.

BGP announcements of PCSID will ease to discover the set of remote ASs supporting the inter-AS MPLS-based constrained tunnels service together with their associated end-to-end capabilities for reaching them. In order to compute a path towards a specific domain supporting this inter-AS MPLS-based constrained tunnels service, the local PCE chooses a route that serves the PCSID of that domain and extracts from the AS_PATH attribute the AS number of the next hop ASBR. Then, the local PCE queries its SLS Management system and gets back the PCE's IP address of the next neighboring PCE to contact. Finally, the local PCE forms and forwards a path computation request to this next PCE. The process is iteratively repeated until the request reaches the PCE of the target AS identified by its PCSID.

This solution decreases the number of BGP announcements that are reduced to one announcement per AS.

11. IANA Considerations

The solution proposed in this draft uses a well-know community attribute value that SHOULD be attributed by IANA [RFC2434] in order to facilitate recognition of BGP announcements that announce PCSv and associated capabilities.

12. Security Considerations

This additional draft does not change the underlying security issues in the existing BGP-4 protocol specification [RFC2385].

Copyright © MESCAL Consortium, June 2005

13. References

[RFC3667] Bradner, S., "IETF Rights in Contributions", RFC 3667, February 2004

[RFC3668] Bradner, S., "Intellectual Property Rights in IETF Technology", RFC 3668, February 2004

[CCAMP-FWRK] Farrel, A., Vasseur, JP., Ayyangar, A., "A Framework for Inter-Domain MPLS Traffic Engineering", draft-ietf-ccamp-inter-domain-framework-00.txt, August 2004

[INTERAS-PCE] Boucadair, M., Morand, P., "A Solution for providing inter-AS QoS tunnels", draft-boucadair-pce-interas-01.txt, May 2005

[INTERAS-PCP] Boucadair M., Morand P. and al., "Inter-AS PCE Communication Protocol", draft-boucadair-pcp-interas-01.txt, May 2005

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

[RFC2608] Guttman, E., Perkins, C., Veizades, J. and M. Day, "Service Location Protocol, version 2", RFC 2608, July 1999

[SLS] Goderis D., T'Joens Y., Jacquenet C., Memenios G., Pavlou G., Egan R., Griffin D., Georgatsos P., Georgiadis L., Heuven P.V., "Service Level Specification Semantics and parameters", draft tequila-sls-02.txt, Work in progress.

[RFC1771] Rekhter, Y., Li T., "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[RFC2434] Alvestrand, H. and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 2434, October 1998

[RFC2385] Heffernan, A., "Protection of BGP sessions via the TCP MD5 Signature Option", RFC 2385, August 1998

14. Acknowledgments

Part of this work is funded by the European Commission, within the context of the MESCAL (Management of End-to-End Quality of Service Across the Internet At Large, http://www.mescal.org) project, which is itself part of the IST (Information Society Technologies) research program.

Internet Draft PCE Discovery via Border Gateway Protocol May 2005

Copyright © MESCAL Consortium, June 2005

The authors would also like to thank all the partners of the MESCAL project for the fruitful discussions.

15. Author's Addresses

Mohamed Boucadair
France Telecom R & D
42, rue des Coutures
BP 6243
14066 Caen Cedex 4
France
Phone: +33 2 31 75 92 31
Email: mohamed.boucadair@francetelecom.com

Pierrick Morand
France Telecom R & D
42, rue des Coutures
BP 6243
14066 Caen Cedex 4
France
Email: pierick.morand@francetelecom.com

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTSOR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNETENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR

Copyright © MESCAL Consortium, June 2005

IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Statement

Copyright (C) The Internet Society (2005). This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

Copyright © MESCAL Consortium, June 2005

15.2 A Solution for Providing Inter-AS MPLS-based QoS Tunnels

PCE Working Group M. Boucadair (Ed.)
IETF Internet Draft P. Morand (Ed.)
Document: draft-boucadair-pce-interas-01.txt France Telecom R&D
Proposed Status: Informational May 2005
Expires: November 2005

A Solution for providing inter-AS MPLS-based QoS tunnels
< draft-boucadair-pce-interas-01.txt >

Status of this Memo

This document is an Internet-Draft and is subject to all provisions of section 3 of RFC 3667 [RFC3667]. By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she become aware will be disclosed, in accordance with RFC 3668 [RFC3668].

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on November 2005.

Abstract

This document describes a solution for providing inter-AS MPLS-based Quality of Service (QoS) tunnels. This solution makes use of Path Computation Elements (PCE) as a means to compute inter-domain constraint-based paths. Service considerations and agreements between two IP Network Providers (INP) implementing this solution are also described.

Copyright Notice

Copyright © MESCAL Consortium, June 2005

Copyright (C) The Internet Society. (2005)

Table of Contents

1. Contributors

- Hamid Asgari (Thales Research and Technology)
- Noel Butler (Thales Research and Technology)
- Panagiotis Georgatsos (Algonet)
- David Griffin (University College London)
- Micheal Howarth (University of Surrey)

2. Changes since last version:

The main changes occurred in this version are:
- Add new contributor to the draft
- Rewording of several sections of the draft

3. Terminology

This memo makes use of the following terms:

Copyright © MESCAL Consortium, June 2005

- Path Computation Element (PCE): an entity that is responsible for computing/finding inter/intra domain MPLS tunnels (LSPs). This entity can simultaneously act as client and a server. Several PCEs can be deployed in a given AS.

- Path Computation Client (PCC): a PCE acting as a client. This entity is responsible for issuing path computation requests that fulfill the Service Management constraints for the establishment of inter/intra domain LSPs.

- Path Computation Server (PCS): a PCE acting as a server. This entity is responsible for handling path computation requests including neighboring PCC constraints.

- High-level service: the service employing a PCE-based system as the underlying infrastructure for creating e.g., inter-domain QoS VPNs.

- High-level service customer: the customer that subscribes to a High-level service.

- pSLS(provider SLS): A provider level SLS, which is established for specific QoS class between two Internet Network Providers (INP) for exchanging traffic in the Internet with the purpose to expand the geographical span of their offered services. pSLSs are meant to support aggregate traffic and they are assumed to exist prior to any agreements with end customers.

- SLS Management: a service level management system, which includes service ordering (i.e for establishing contracts between peer providers) and service invocation (i.e for committing resources before traffic can be admitted)

- q-BGP: QoS-inferred BGP. A modified BGP protocol that takes into account QoS information as input for its route selection process.

- Domain: within this document it denotes an Autonomous System.

4. Introduction

4.1. General

The level of Quality of Service (QoS) guarantees offered by INPs using a pure IP-based traffic engineering (TE) solution, other than overbooking, is not yet satisfactory for all corporate business services, for which strong guarantees must be provided. For this type of customers, hard QoS performance and bandwidth guarantees are considered as the major requirements.

Copyright © MESCAL Consortium, June 2005

Currently, these requirements can be satisfied within a single domain or across several interconnected domains managed only by a single INP. However, it becomes very challenging when these domains are managed by different INPs. Each INP defines and deploys its own QoS policy within the scope of its domain(s), utilizes its proprietary TE functions, etc.

Providing QoS-based services within the scope of the Internet requires the collaboration among INPs in order to offer this type of services. This document aims at proposing a solution that will facilitate the introduction of such services in an Inter-provider environment. Service considerations are also taken into account.

4.2. Assumptions

In this document, we assume the following:

- An AS can announce a given prefix in several QoS planes; each of these QoS planes being identified across inter-domain links by a unique DiffServ Code Point (DSCP);
- Each announcement, except best effort ones, contains values of a set of QoS parameters that characterizes the likely end-to-end QoS to be experienced for reaching a given prefix (we call this end-to-end QoS as aggregated QoS);
- The way aggregated QoS values are computed is out of the scope of this document;
- Adjacent ASs agree on the DSCP values to use in order to signal a given QoS class at inter-domain links (we call these DSCP values inter-domain DSCPs);
- Every AS has the freedom to bind an inter-domain DSCP to a local DSCP within its domain, which identifies its local QoS class for signalling a QoS planes in its domain;
- An AS supporting the inter-domain MPLS-based QoS tunnels service, owns a Path Computation Service Identifier(s) (PCSID), which is a routable IP address. This IP address is not necessarily the IP address(es) of the PCE(s) of the domain. These PCSIDs are announced per QoS plane basis, by an inter-domain routing protocol, together with the plane's aggregated QoS values;
- ASs can discover other ASs supporting the inter-domain MPLS-based QoS tunnels service by receiving inter-domain routing protocol announcements. These announcements provide an AS with

Copyright © MESCAL Consortium, June 2005

the end-to-end QoS characteristics of the path towards any prefix of the remote AS owning the PCSID.

4.3. Draft structure

The structure of this document is as follows:

- Section 5 describes the inter-AS PCE model.
- Section 6 discusses the service considerations.
- Section 7 highlights PCE functions.
- Section 8 explains the PCE discovery function.
- Section 9 gives an overview of the PCP protocol that is used for communication between PCEs.
- Section 10 and 11 describe routing issues.
- Finally, section 12 presents some advance features in order to enhance PCE service.

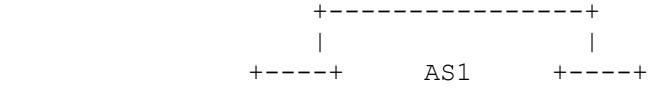
5. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

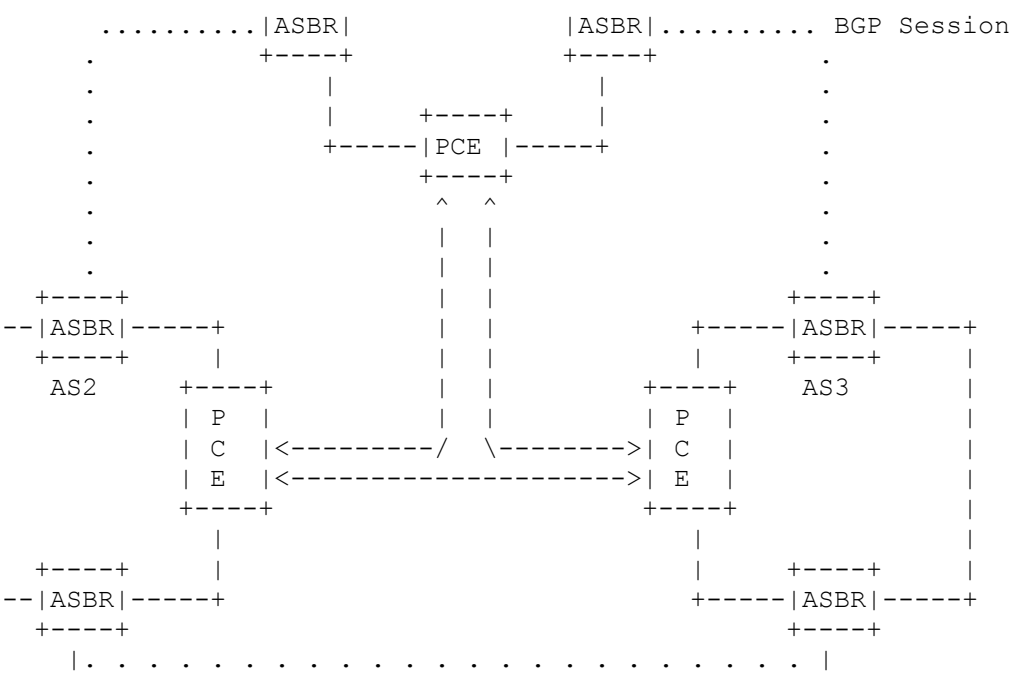
6. Inter-AS PCE model

A Path Computation Element (PCE) is responsible for finding an inter-domain path satisfying a set of constraints (e.g. specific QoS performance guarantees requested by a customer), in order to establish inter-domain constraint-based LSPs.

In an inter-provider environment, the computation of this path is necessarily distributed and required communication between PCEs of different domains. Communication between PCE entities is achieved via the PCE Communication Protocol (PCP, [INTERAS-PCP]). Once computed, the path is provided to the RSVP-TE/MPLS machinery of the head-end Label Switching Router (LSR), which can request/establish an inter-domain Label Switching Path (LSP) that will follow the inter-domain path provided by the PCE.



Copyright © MESCAL Consortium, June 2005



...: Peering links
---: inter-PCE session

Figure 1: Inter-AS PCE scenario

In the above figure, we assume that each domain operates a set of QoS-classes. Each INP deploys its own local QoS-classes with specific QoS characteristics. QoS-classes in a domain have not necessarily the same QoS characteristics with QoS-classes in the other domains. We also assume that service level QoS-based contracts (pSLSs) have been established between adjacent domains. BGP is running between these adjacent ASs. Each AS learns, the set of destinations its peer domains can reach, together with end-to-end QoS performance characteristics.

When a pSLS is established, the domains exchange their respective PCE information (name, IP address, identifiers, authentication information, etc.).

In order to create an inter-domain constraint-based LSP, the domain which requests the establishment of the LSP asks its PCE to compute an inter-domain path satisfying a set of constraints, (for example bandwidth guarantee per QoS-Class, maximum end-to-end delay, etc.)

The first PCE selects one possible path among the set of alternatives and identifies the next-hop domain. It then verifies that appropriate resources are available in its own domain and sets up administrative pre-reservations in the management system of its domain. Then it

Copyright © MESCAL Consortium, June 2005

contacts the next hop PCE, requesting a path computation between the next hop AS Border Router (ASBR) and the termination address of the inter-domain LSP. This second PCE performs the same computation as the first one did and the procedure is iteratively repeated up to the last PCE. If a path satisfying all requirements is found, each PCE returns the path received from the responding PCE concatenated with the sub-path it computed. Finally, the originating PCE receives the whole computed path.

7. Inter-provider Service Considerations

A pSLS MUST be established between two neighbors in order to grant to the requestor the appropriate rights for requesting and/or establishing inter-domain LSPs (see figure below). Nevertheless, information like destination and number of future LSPs to be established is not known in advance. The pSLS indicates only the upper-boundaries that the upstream AS is allowed to use, in terms of QoS-class identifiers that can be used at the inter-domain for establishing inter-domain LSP and in terms of maximum bandwidth associated to each QoS-class.

The pSLS agreement does not reserve any network resources in advance. Resources are only allocated when an inter-domain LSP is set up. The management plane of each downstream domain along the path SHOULD be aware of the existence of those LSPs together with their associated QoS guarantees.

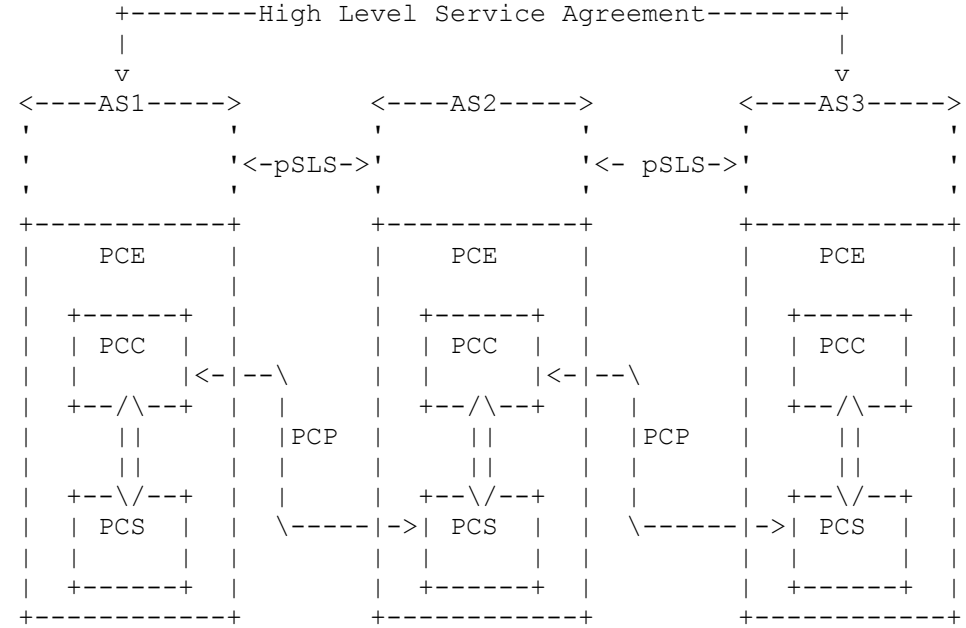


Figure 2: Service Overview

Copyright © MESCAL Consortium, June 2005

However, it is difficult to establish such a contract in advance especially when the LSP path is not known in advance. Thus, the sequence of operation for establishing an LSP should be:

- o Compute inter-domain path candidate(s);
- o Select and request an inter-domain path for this particular LSP using information returned by the PCE,
- o Establish the LSP once final negotiation terms have been agreed end-to-end between PCEs of adjacent domains.

The establishment of this cascade of negotiations can be difficult to achieve and can take some time. In particular, the risk is not negligible that the resources that were available when the PCE performed the path computation are no longer available along the path when the cascaded negotiation terms are agreed, because others LSPs have used the corresponding resources.

In order to solve this issue it is necessary that the PCE of each domain makes an administrative reservation of the corresponding resources and indicates the characteristics of the path. This information is registered by the management plane, which triggers in parallel the creation of a provisional contract referencing the technical characteristics of the future LSP. Subsequent path computation requests may be impacted because the management plane removes these resources from the available overall network resources. This provisional contract is valid for a limited time period, which is the minimum of time periods each specified and reported by a domain along the path. If time period expires, the provisional contract can be removed from the management systems, and related administrative network resources have to be informed.

It is the responsibility of the management plane of each domain to cooperate in agreeing the exact financial terms and additional clauses of this contract, including its duration. Each domain knows the entry and the exit point of the LSP within its own domain and consequently knows both the upstream and downstream ASs to deal with. This validation procedure SHOULD ideally be automated to speed up the process and could integrate pricing negotiation. The way that the other blocks of the management plane deal with this automation is out the scope of this document.

Thus, once the pre-contract is validated, the path computed by the PCE can be provided to the head-end LSP, which effectively sets up the LSP. Note that every ingress point of each domain SHOULD activate some outsourced policy functions that would allow RSVP TE to get an agreement from the management system.

8. Path Computation Element functions

Copyright © MESCAL Consortium, June 2005

The main function provided by a PCE is to contribute to the overall path computation by computing part of the end-to-end inter-domain path satisfying a set of constraints. The management plane could call other elementary services such as requesting a path computation for informational purposes or canceling a request in progress for instance.

The deployment and the maintenance of the LSP connectivity require cooperation of several functional entities from different planes. Within this document, the PCE is only responsible for computing an inter-domain constraint-based path. The implementation of the service (whether it is automated or not) and the creation of inter-domain LSP results from the cooperation of functional blocks, control plane blocks and data plane blocks aren't described in this document.

The PCE does not itself trigger the establishment of any inter-domain LSP, but provides inter-domain paths, if they are available. Unlike the management plane, it is not aware of business considerations A PCE entity provides an interface used by the service management block to request a path computation. It communicates with other remote PCE thanks to the PCP protocol and requests additional services from other functional blocks as illustrated in the figure below:

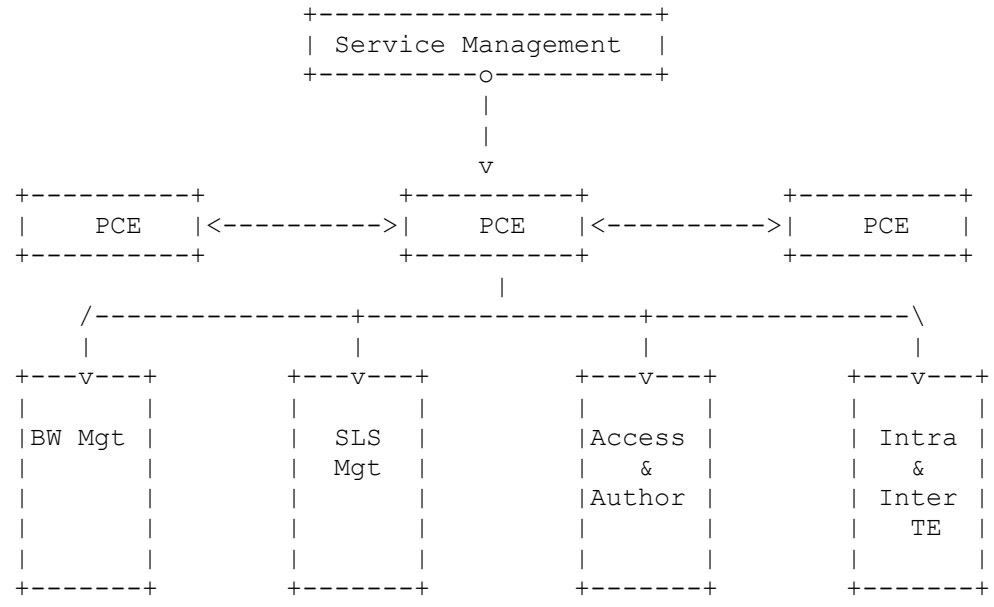


Figure 3: PCE Interactions

The PCE interacts with the dynamic routing processes to retrieve routing information that is used to compute an inter-domain path satisfying the expressed constraints. An interface MUST be made available to the PCE so that it can access routing information. Note that both intra and inter-domain routes MUST be made available to the PCE.

Copyright © MESCAL Consortium, June 2005

In addition, for access control and authorization purposes, the PCE MUST be provided with access to the list of other PCEs from which it will accept requests. This list is updated each time new pSLSs are negotiated by the INP.

9. PCE discovery

Within this document, we assume that during the service negotiation phase the peers exchange the IP addresses of their respective PCE(s). This information is stored in the SLS Management Systems of each INP.

As described in [PCE-DISCOVERY], instead of announcing all potential tail-end addresses in BGP, only an identifier is announced via BGP. It is called the Path Computation Service Identifier (PCSID). This particular BGP announcement is identified by a well-known community value (to be defined by IANA) and is represented by a routable IP address, which can be different from the real IP address of the PCE.

As a consequence, this particular route SHOULD NOT be installed in the Forwarding Information Base (FIB) since this PCSID is not necessarily the IP address of the PCE.

BGP announcements of PCSID will ease to discover the set of remote ASs supporting the inter-AS MPLS-based QoS tunnels service and associated end-to-end QoS-related information for reaching them. In order to compute a path towards a specific domain supporting this inter-AS MPLS-based QoS tunnels service, the local PCE chooses a route that serves the PCSID of that domain and extracts from the AS_PATH attribute the AS number of the next hop ASBR. Then, the local PCE queries its SLS Management system and gets back the PCE's IP address of the next neighboring PCE to contact. Finally, the local PCE forms and forwards a path computation request to the next PCE. The process is iteratively repeated until the request reaches the PCE of the target AS identified by its PCSID.

10. PCE to PCE Communication

A PCE can act as a client (Path Computation Client, PCC) or a server (Path Computation Server, PCS). The PCC is responsible for issuing Path Computation requests. The PCS is responsible for handling requests received from PCCs.

The PCP (Path Computation Protocol) is a simple query and response protocol that is used for communication between PCE entities, i.e., PCC and PCS.

Copyright © MESCAL Consortium, June 2005

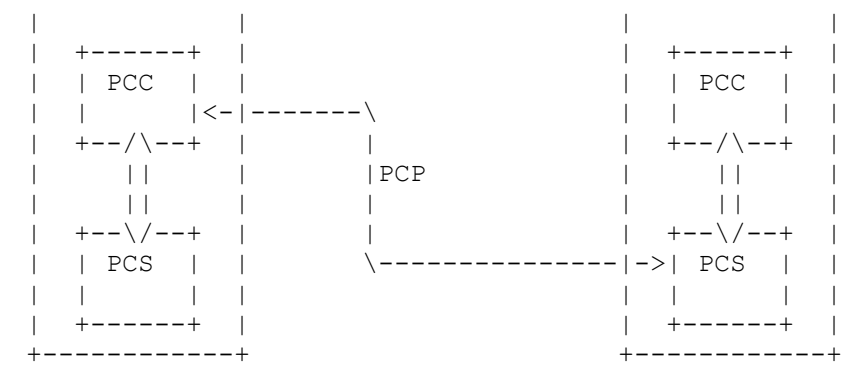


Figure 4: PCC to PCS communication

The main characteristics of the PCP protocol are:

- The protocol employs a client/server model in which a PCE can both act as a client and/or a server at the same time. A PCE Client (PCC) sends requests, cancellation and receives responses.

- The protocol uses TCP as its transport protocol, providing the reliable exchange of messages between PCEs.

- In its first version, PCP does not provide any message level security for authentication, message replay protection, and integrity. However, PCP can reuse existing protocols for security such as IPSEC [RFC2401] or TLS [RFC2246] to authenticate and secure the channel between two PCE.

- The current PCP protocol provides the service for supporting only a basic path computation function. In particular it does not support the service for additional path computation constraints, or provide enhanced reporting features in the case of path computation failure.

11. Routing considerations

11.1. Assumptions

We assume in this document that PCE addresses are only announced in a few QoS-Class planes. Addresses of LSR/LER interfaces could be announced in the best effort plane. This reduces the number of BGP announcements to one announcement per PCE per AS. By setting a well-known community value, we specify announcements that serve PCEs. These are not regarded as routes and are not stored in the FIBs.

11.2. Finding inter-domain LSP paths

Copyright © MESCAL Consortium, June 2005

In order to find an inter-domain path, the PCE MUST be provided with a set of attributes including the information describing head-end and tail-end of the LSP and the performance requirements for the LSP. The aforementioned information MUST include the loopback IP address of its LSR, and the IP Address of the PCE of its domain (notation is IPADDRESS@PCSID). This information MUST also include the performance guarantees required for the inter-domain constraint-based LSP. This information MAY encompasses the requested QoS-class(es) so that the set of collaborating PCE can compute a path that will cross a set of domain satisfying the expressed constraints.

It can also contain, per QoS-class, additional QoS performance guarantees that the PCE must take into account. These performance guarantees include guaranteed end-to-end delay, jitter, loss rate and/or bandwidth. Note that these parameters can differ depending on the type of requested QoS-class, and they MAY not all be present in the LSP set-up request. If included in the path computation request they MUST be taken into account by the PCE. If the PCE doesn't recognize a given QoS parameter, the PCE MUST stop its computation and MUST return an error (PCP Error Message).

When computing a path, a PCE interacts with other blocks of the management plane. In particular, it checks the availability of the resources within the boundaries of its domain. If the resources are available, and the sub-path (path between the ingress point of its domain and a potential ingress point of a given domain) conforms to the path constraints requested, it MUST inform the management plane of a pre-reservation concerning this path. This information can then be taken into account when processing other path computation requests. Once this operation succeeds, the request is propagated to the next domain PCE, which has been selected by the PCE of this domain.

Note that the performance guarantees requested MUST be updated before forwarding it to the next domain by taking into account the performance figures already observed along the computed sub-path plus the performance figures within its own domain. Therefore, PCE MUST be aware of the above performance figures of the QoS-classes.

The requesting PCE MUST use the QoS-class identifier they agreed during the pSLS negotiation phase in order to signal a given QoS-class.

If an end-to-end LSP has to be re-engineered because the associated constraints have changed in terms of QoS-class requested, bandwidth, delay, etc., a new end-to-end path needs to be computed. In order to improve its chances of finding a valid path, the requestor can specify that the path for which the request is issued will replace a previously established LSP. For doing so, the requestor can indicate the reference of the path corresponding to this LSP. A PCE can release, during the path computation process, the resources

Copyright © MESCAL Consortium, June 2005

corresponding to the former LSP, if the new path follows part of the former path. This reference is stored in the management plane of each domain and is generated by the initial requestor. This reference is globally unique.

The ability to address such additional constraints can be interesting in the case of backup LSPs, so that the PCE can compute a path along a different route. These considerations are out of scope of this document.

12. Communication between PCE and LSR/LER for initiating LSP set-up

Communication between PCE and an LER/LSR could be achieved thanks to the use of Common Open Policy Service protocol (COPS, [RFC2748]). An RSVP client-type could be used in order to convey configuration data resulting from the computation operation executed by a PCE. Specification of RSVP configuration data is out of scope of this document.

13. Advanced features

13.1. Exclusion of specific ASs from the path

If a PCE in the chain wants to exclude particular AS(s) from the path, additional constraints (that can be expressed using the AS number of the excluded domain/s) MUST be added to the request message body and MUST be propagated downstream.

13.2. Feedback

When computing a path, the PCEs can fail to find a path for a number of reasons. These failures, in normal operations, will be mainly due to the lack of resources, or not meeting the requested QoS requirements. In such a situation, a path, which would have been the optimal path, would not be established. Identifying the domain/s where the path computation failed, together with the reasons, would be of a real added value for providers in order to improve their efficiency.

A proposal for achieving this is to rely on the Path Computation Protocol, which could be improved to return all the path alternatives which were tried but have failed. In doing so, the requesting provider would be aware of the reasons of the failure and possibly interact with that AS where the path computation failed and aborted.

The AS (where the path computation failed), faces with multiple requests, from external domains, could consider a possible

Copyright © MESCAL Consortium, June 2005

modification of some of its peering agreements based on business objectives.

14. Security Considerations

This document does not change the underlying security issues in the PCP and BGP protocols specifications. It is recommended that a security protocol like IPSec or TLS to be activated in order to protect PCP sessions.

15. References

[RFC3667] Bradner, S., "IETF Rights in Contributions", RFC 3667, February 2004

[RFC3668] Bradner, S., "Intellectual Property Rights in IETF Technology", RFC 3668, February 2004

[RFC2385] Heffernan, A., "Protection of BGP sessions via the TCP MD5 Signature Option", RFC 2385, August 1998

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

[INTERAS-PCP] Boucadair M., Morand P., "Inter PCE Communication Protocol", draft-boucadair-pcp-interas-01.txt, May 2005

[PCE-DISCOVERY] Boucadair M., Morand P., "PCE discovery via Border Gateway Protocol", draft-boucadair-pce-discovery-01.txt, Work in progress, May 2005

[RFC2401] Atkinson R., "Security Architecture for the Internet Protocol", RFC 2401, August 1998

[RFC2246] Dierks T., Allen C., "The TLS Protocol", RFC 2246, January 1999

[RFC2748] Boyle, J., Cohen, R., Durham, D., Herzog, S., Raja, R. and A. Sastry, "The COPS (Common Open Policy Service) Protocol", RFC 2748, January 2000.

16. Acknowledgments

The authors would also like to thank all the partners of the MESCAL (Management of End-to-End Quality of Service Across the Internet At Large, http://www.mescal.org) project for the fruitful discussions.

Copyright © MESCAL Consortium, June 2005

17. Author's Addresses

Mohamed Boucadair
France Telecom R & D
42, rue des Coutures
BP 6243
14066 Caen Cedex 4
France
Phone: +33 2 31 75 92 31
Email: mohamed.boucadair@francetelecom.com

Pierrick Morand
France Telecom R & D
42, rue des Coutures
BP 6243
14066 Caen Cedex 4
France
Email: pierick.morand@francetelecom.com

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED,

Copyright © MESCAL Consortium, June 2005

INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Statement

Copyright (C) The Internet Society (2005). This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

Copyright © MESCAL Consortium, June 2005

15.3 Inter PCE Communication Protocol

Network Working Group M. Boucadair (Ed.)
P. Morand (Ed.)
Internet Draft France Telecom R&D
Document: draft-boucadair-pce-comm-proto-00.txt May 2005
Category: Standards Track

Inter PCE Communication protocol
draft-boucadair-pce-comm-proto-00.txt

Status of this Memo

This document is an Internet-Draft and is subject to all provisions of section 3 of RFC 3667 [RFC3667]. By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she become aware will be disclosed, in accordance with RFC 3668 [RFC3668].

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on November 2005.

Abstract

This draft describes a new protocol allowing communication between two Path Computation Elements (PCEs) located in different domains in order to compute inter-domain paths satisfying a set of QoS constraints. This protocol could also be used for intra-domain purposes.

Table of Contents

Copyright © MESCAL Consortium, June 2005

1. Contributors

- Hamid Asgari (Thales Research and Technology)
- Panagiotis Georgatsos (Algonet)
- David Griffin (University College London)
- Micheal Howarth (University of Surrey)
- Thibaut Coadic (France Telecom)

2. Changes since last version:

The main changes occurred in this version are:
- Add new contributor;
- Rewording of several sections of the draft;
- Correction of some typos.

Copyright © MESCAL Consortium, June 2005

3. Terminology

This memo makes use of the following terms:

- Path Computation Element (PCE): an entity that is responsible for computing/finding inter/intra domain paths for establishing LSPs. This entity can simultaneously act as client and a server. Several PCEs could be deployed in a given AS.

- Path Computation Client (PCC): a PCE acting as a client. This entity is responsible for issuing path computation requests that fulfill the Service Management constraints for the establishment of inter/intra domain LSPs.

- Path Computation Server (PCS): a PCE acting as a server. This entity is responsible for handling path computation requests in order to satisfy PCC constraints.

- High-level service: is the service using a PCE-based system as an underlying infrastructure (an inter-domain QoS VPNs service for instance)

- High-level service customer: is a customer that subscribes to a High-level service.

- pSLS: A provider SLS is an SLS established between two Internet Network Providers (INP) with the purpose of extending the geographical span of their service offers.

- SLS Management: This management entity is responsible for SLS-related activities, including pSLS ordering (i.e establishing contracts between peers) and SLS invocation (i.e committing resources before traffic can be admitted)

- q-BGP: QoS-inferred BGP. A modified BGP protocol that takes into account QoS information as input to for its route selection process.

- Domain: within this draft it denotes an Autonomous system.

4. Introduction

Nowadays, services are deployed on the same infrastructure (best-effort shared IP network) on which more elaborate functionalities rely for providing enhanced network services. Especially those intended for specific corporate customers or providers needs. These extra functionalities were introduced because the basic IP approach failed to support those added-value services or was not considered to be efficient enough.

Copyright © MESCAL Consortium, June 2005

MPLS is a technical solution that has been successfully deployed by a large number of providers for supporting connection-oriented services such as IP VPN services for which traffic isolation is an important criterion. Then, the solution evolved to encompass QoS issues, and Traffic engineering functions were then progressively introduced. Up to now, some providers have deployed MPLS TE but only within their own domains.

Extending the scope of offered intra-domain services (like QoS-based services), using MPLS as infrastructure, to the Internet scale is conditioned by the cooperation between service providers. Several proposals have been proposed within the IETF in order to deal with this issue but only from inter-AS point of view (see for example [INTERAREA-REQ], [INTERAS-REQ], [PCE-ARCH] and [PCE-FWK]).

Inter-provider issues need to be studied further in order to build a complete end-to-end solution.

Draft [INTERAS-PCE] describes a solution that could be implemented in order to offer end-to-end services. This solution requires a close cooperation between distinct Path Computation Elements (PCE) that are located in distinct domains.

This draft describes a protocol to use for communication between two Path computation Elements.

The structure of this draft is as follows:
- o Section 5 presents an overview of the overall service approach;
- o Section 6 lists characteristics of the PCP protocol;
- o Sections 7 and 8 detail the PCP messages and operations.

5. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

6. Overview of overall service approach

Neighboring domains establish pSLSs between themselves. An inter-domain routing protocol runs between the domains. This inter-domain routing protocol is used to announce PCE unique identifiers [PCE-DISCOVERY] across the Internet in order for other PCEs to be able to discover possible paths towards every AS having a PCE. Therefore, when an AS wants to establish an LSP between 2 addresses, its PCE forms a path computation request containing the HEAD-END-ADDRESS and the TAIL-END-ADDRESS defining the future LSP. In addition to the IP address of the head and the tail of the LSP, each X-END-ADRESS contains also the PCE unique identifier of the AS these IP addresses

Copyright © MESCAL Consortium, June 2005

belong to. Using information reported by BGP the PCE identifies possible paths that reach the target AS identified by its PCE unique identifier. It then selects one of these paths and forms a new request, which is sent to the neighboring PCE it selected along that path.

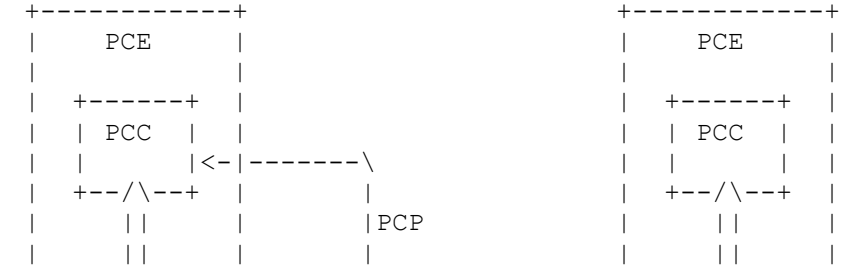
The path computation request is propagated downstream to the appropriate PCEs and is repeated until the request reaches the destination PCE. Each PCE along the path ensures that the constraints expressed by the request are satisfied. Each PCE is responsible for computing both the intra- and inter-domain sub-path and to ensure that resources are available and will remain available until the LSP is effectively created. If for some reasons the path computation aborts, all resources must be relaxed.

After authenticating the identity of LSP requester (originating) PCE, the destination PCE sends a reply message back to the upstream domain's PCE accepting the request. The LSP sub-path (from the ingress ASBR and the final destination) is inserted in the message. The next upstream domain's PCE does the same adding its own relevant sub-path to the overall loose or strict path. At the end of the chain, the originating PCE does also the same. An end-to-end path has thus been computed. The originating PCE is now in a position to provide the service request handler with appropriate information (end-to-end inter-domain path) allowing an RSVP reservation to be issued for the establishment of the LSP.

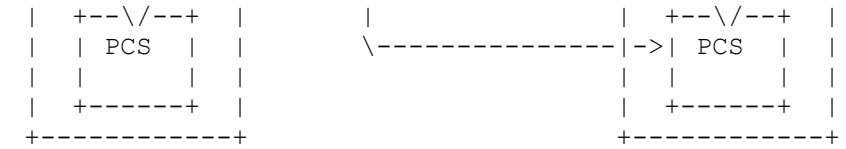
At the service/application level, when an originating AS wants to establish an LSP towards a destination ASs, there MUST exist a preliminary agreement between the two ASs (Service providers owning these PCEs). This agreement specifies both the tail-end and head-end address of the LSP, together with the PCE unique identifier of the originating and destination AS. This allows only agreed LSP to be established.

7. PCE to PCE communication

A PCE can act as a client (PCC) or a server (PCS). A PCC is responsible for issuing requests. PCS is responsible for handling requests received from PCCs.



Copyright © MESCAL Consortium, June 2005



PCP protocol is used for communication between a PCC and a PCS.

PCP is a simple query and response protocol that can be used between PCE entities to collaborate for computing an inter-domain QoS constrained path.

The main characteristics of the PCP protocol include:

- The protocol employs a client/server model in which a PCE can both act as a client and/or a server at the same time. A PCE Client (PCC) sends requests, cancellation and receives responses.

- The protocol uses TCP as its transport protocol for reliable exchange of messages between PCE. Therefore, no additional mechanisms are necessary for reliable communication between two PCEs.

- In its first version, PCP does not provide any message level security for authentication, message replay protection, and integrity. However, PCP can reuse existing protocols for security such as IPSEC [RFC2401] or TLS [RFC2246] to authenticate and secure the channel between two PCEs.

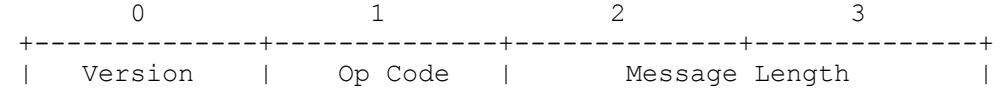
- The current PCP protocol provides the service for supporting only a basic path computation function. In particular it does not support additional path computation constraints, or provide enhanced reporting features in the case of path computation failure.

8. PCP messages

This section discusses the PCP message formats and objects exchanged between PCE entities.

8.1. Common header

Each PCP message consists of the PCP header followed by a number of arguments depending on the nature of the operation.



Copyright © MESCAL Consortium, June 2005

Global note: //// implies field is reserved, set to 0.

The fields in the header are:

Version: 8 bits. PCP version. Current version is 1.

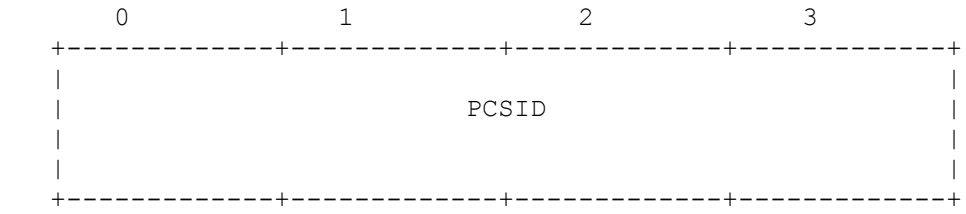
Op Code: 8 bits. The current defined PCP operations are:

```
1 = OPEN                (OPN)
2 = ACCEPT              (ACP)
3 = CLOSE               (CLO)
4 = REQUEST             (REQ)
5 = RESPONSE            (RSP)
6 = PATH-ERROR          (ERR)
7 = CANCEL              (CCL)
8 = ACKNOWLEDGE         (ACK)
9 = KEEP-ALIVE          (KA)
```

Message Length: 16 bits

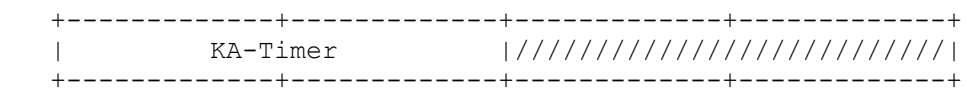
This is the size of the message in octets, which includes the standard PCP header and all encapsulated objects. Messages MUST be aligned on 4 octet intervals.

8.2. OPEN message



The message contains only one argument. This PCSID is propagated by BGP between the domains. This is a routable IPv4 or IPv6 address identifying a PCS of a domain. This PCSID must be inserted by the PCE opening a PCP session. The size of the PCSID is 4 or 16 bytes.

8.3. ACCEPT message



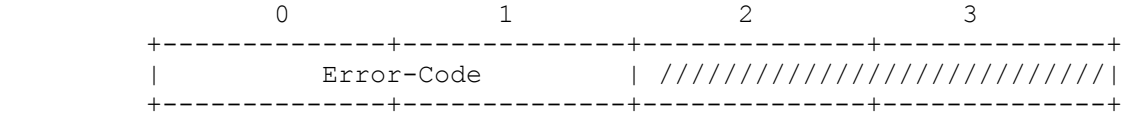
o KA-Timer (Keep-Alive Timer): The argument of the accept message is a 2 octets integer value which represents a timer value expressed in units of seconds. This timer value is treated as a delta. KA-Timer is used to specify the maximum time interval over which a PCP message MUST be sent by the two communication

Copyright © MESCAL Consortium, June 2005

entities. The range of finite timeouts is 1 to 65535 seconds represented as an unsigned two-octet integer. The value of zero implies infinity.

8.4. CLOSE message

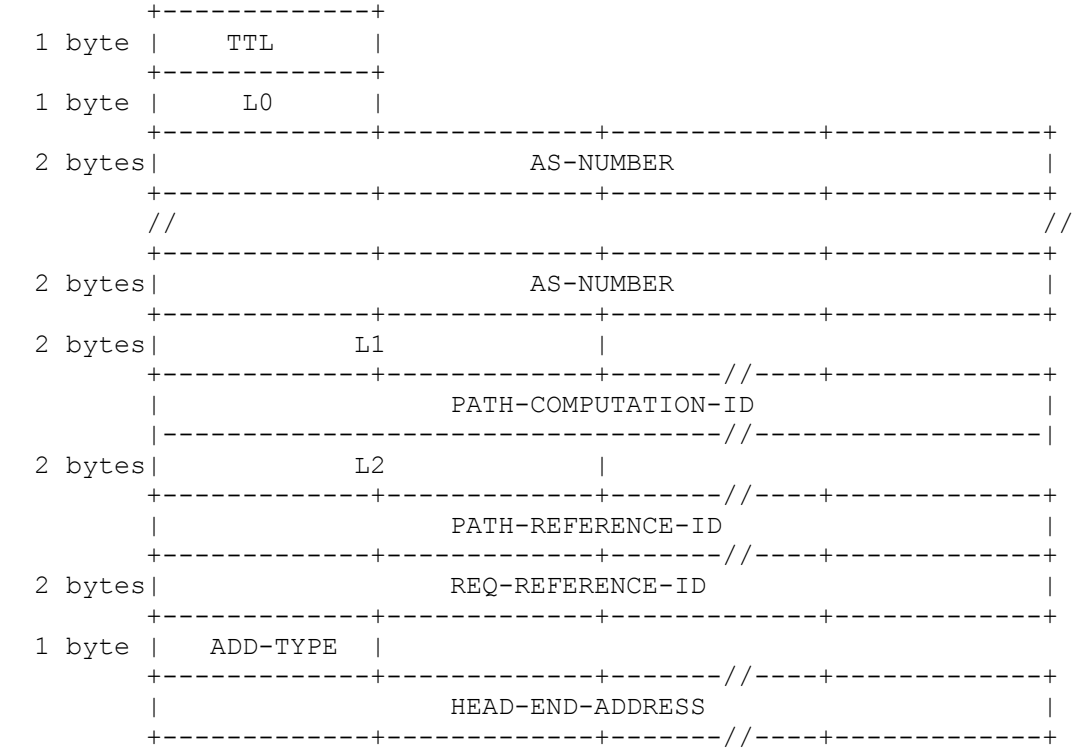
The close message contains an error code indicating the reason of the close of the session.



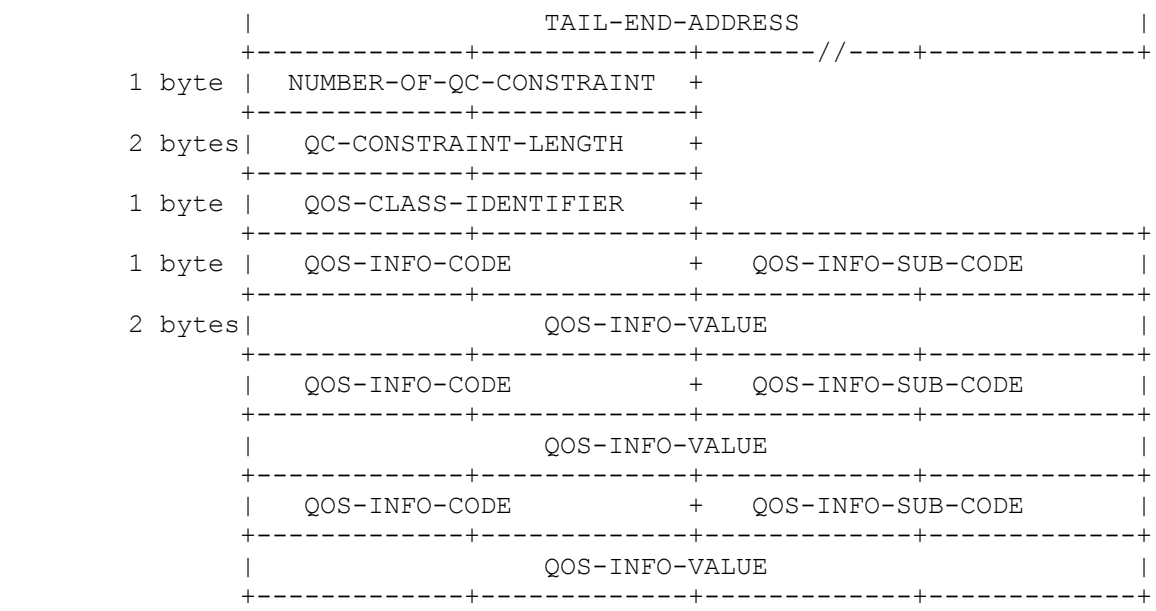
Error-Code:
1 = Shutting Down
2 = Bad Message Format
3 = Incorrect identifier
4 = Unable to process
5 = Protocol error

8.5. REQUEST message

The Request message is sent by the PCC for computing and inter-domain path.



Copyright © MESCAL Consortium, June 2005



- TTL: is the maximum number of ASs that can be crossed by the path. This field is decremented by one each time a PCS issues a request.

- L0: is a 1-byte length field. It represents the number of ASs that have already been crossed.

- AS-NUMBER: is a 2 bytes length field representing an AS number. The first AS-NUMBER value of the list is the AS-NUMBER of the PCC that initialized a path computation.

- L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.

- PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. It is suggested, at least in this first specification, that this identifier is computed using the PCSID of the domain, concatenated with the date and an identifier that will be computed by the first requesting PCC each time a request will have to be issued. Across PCC reboots, this identifier must be unique. This PATH-COMPUTATION-ID will be replicated in all subsequent request initiated by the PCEs along the path.

- L2: is the length in bytes of the PATH-REFERENCE-ID. Size of this field is 2 bytes.

- PATH-REFERENCE-ID: is a variable-length field. It is an identifier that represents a pre-agreement between the head and

Copyright © MESCAL Consortium, June 2005

the tail-end domain that allows the PCS from the terminating domain to accept or reject the path computation request.

- REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to identify the REQUEST. It allows making the difference between several REQ issued for different path computation (but same PATH-COMPUTATION-ID) between two neighbor ASs interconnected via multiple links.

- ADD-TYPE: indicates the nature of the IP addresses of the tail-end and head-end termination:
  - 1 = IPv4
  - 2 = IPv6

- HEAD-END-ADDRESS: is the head-end address of the future LSP represented in the form HEAD-END@PCSID. This is a couple of IPv4 or IPv6 address. The first address of the couple identifies a loopback or an interface address of a network element, the second element is the PCSID of the domain owning the previous address.

- TAIL-END-ADDRESS: is the tail-end address of the LSP represented in the form TAIL-END@PCSID. This is a couple of IPv4 or IPv6 address. The first address of the couple identifies a loopback or an interface address of a network element, the second element is the PCSID of the domain owning the previous address.

These above parameters MUST be present in each REQUEST and in the same order.

- NUMBER-OF-QC-CONSTRAINT: represents the number of QoS class constraints the PCS must take into account when computing a path. A QoS class constraint contains a QoS-Class-Identifier (like a DSCP value) followed by additional constraints. The size of this filed is 1 byte. This field in not really necessary in this first version of the specification but it could become useful if additional path constraints were included in the request.

- QC-CONSTRAINT-LENGTH: is the length in bytes of the QoS-Class-Constraint that follows. The size of this field is 2 bytes.

- QOS-CLASS-IDENTIFIER: identifies a particular QoS-class. The size of the field is 1 byte.

- QOS-INFO-CODE: this field identifies the type of QoS information. The size of this field is 4 bits. This code could be:
  - (0) Reserved
  - (1) Packet rate

Copyright © MESCAL Consortium, June 2005

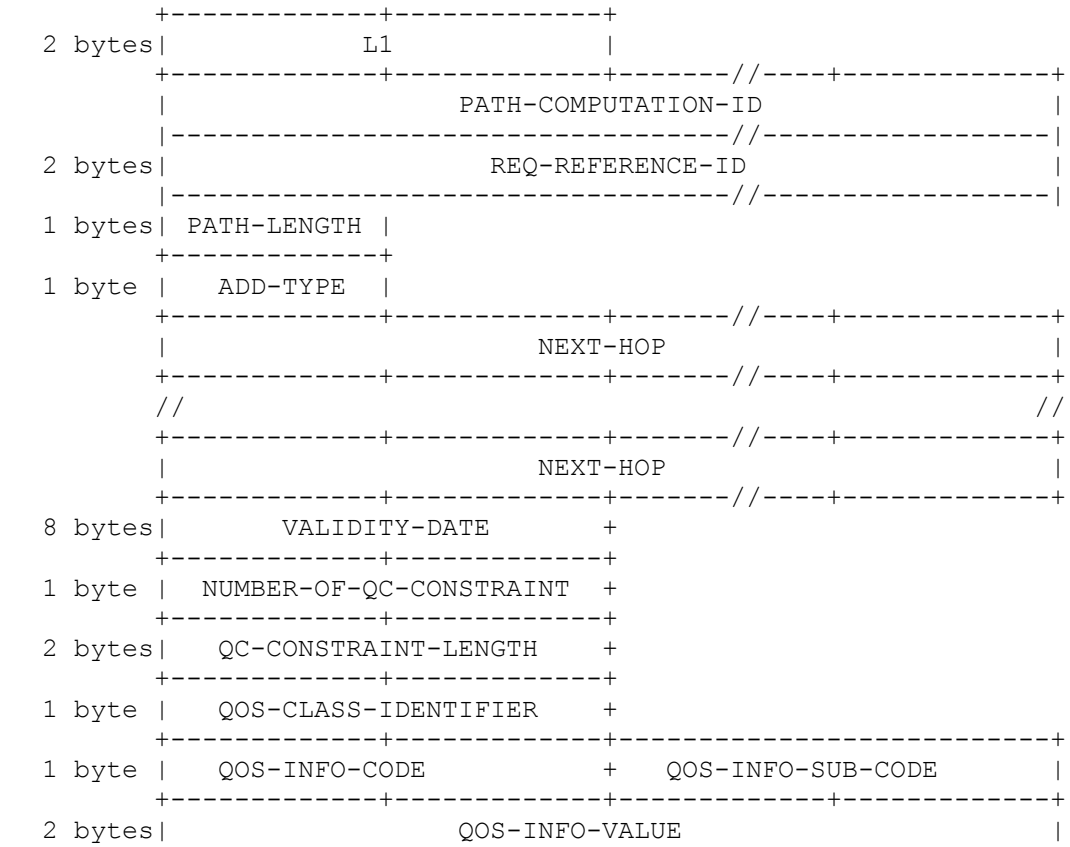
- o (2) One-way delay metric
- o (3) Inter-packet delay variation

- o QOS-INFO-SUB-CODE: this field carries the sub-type of the QoS information. The following sub-types have been identified. The size of this field is 4 bits.
  - o (0) None
  - o (1) Reserved rate
  - o (2) Available rate
  - o (3) Loss rate
  - o (4) Minimum one-way delay
  - o (5) Maximum one-way delay
  - o (6) Average one-way delay

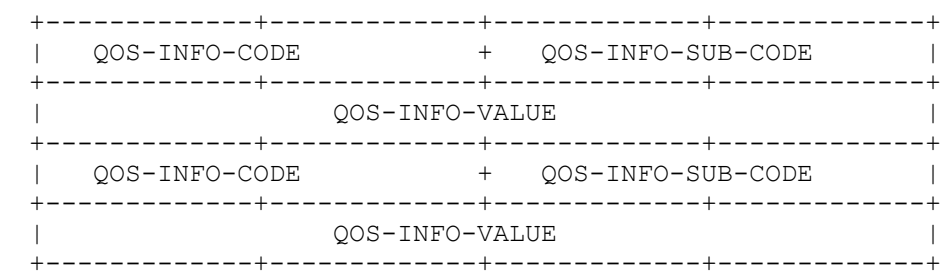
- o QOS-INFO-VALUE: this field indicates the value of the QoS information. These are the constraints that the PCE should respect. The corresponding units depend on the instantiation of the QoS information code.

8.6. RESPONSE-PATH message

This message is sent back when a path has been successfully computed.



Copyright © MESCAL Consortium, June 2005



- L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.
- PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. This value of this identifier MUST be the same as the one provided by the REQUEST.
- REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.
- PATH-LENGTH: indicates the number of next hops that form the path. The size of this filed is 1 byte.
- ADD-TYPE: indicates the nature of the IP addresses in the PATH. The size of this filed is 1 byte.
  - 1 = IPv4
  - 2 = IPv6
- NEXT-HOP: IP address of a next hop that is part of the computed path. Size of this field depends on the nature of the IP address.
- VALIDITY-DATE: represents the GMT date after which the computed path returned will not be valid. The size of this field is 8 bytes.

These above parameters MUST be present in each RESPONSE and in the same order.

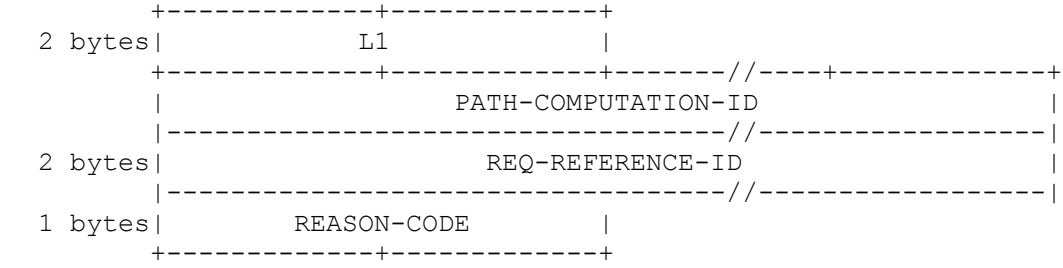
The other parameters have the same meaning than for the REQUEST except:

- QOS-INFO-VALUE: represents the QoS guarantees of the path, for this particular QoS-INFO-CODE parameter (delay, jitter,…) between the ingress ASBR of the responding PCE domain and the tail-end of the path.

8.7. PATH-ERROR message

Copyright © MESCAL Consortium, June 2005

This message is sent back when a path could not be computed.

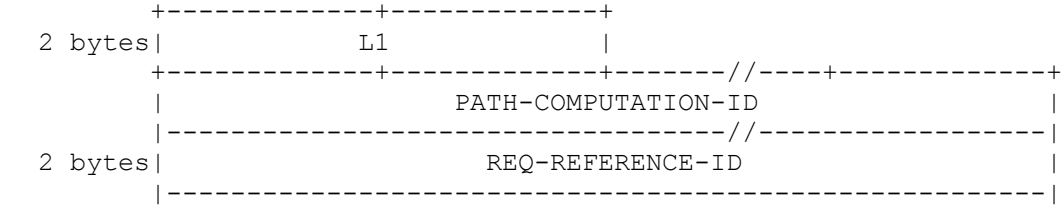


- o L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.
- o PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. This identifier MUST be the same as the one provided by the REQUEST.
- o REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.
- o REASON-CODE: indicate the reason of the failure. Identified failure are:

  1 = No resource available
  2 = Path reference error
  3 = Abnormal termination
  4 = PATH-COMPUTATION-ID already used
  5 = TTL expired
  6 = Loop detected
  7 = Request already handled

8.8. CANCEL message

This message is sent by a PCC or a PCS when a path computation must be cancelled.



- o L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.

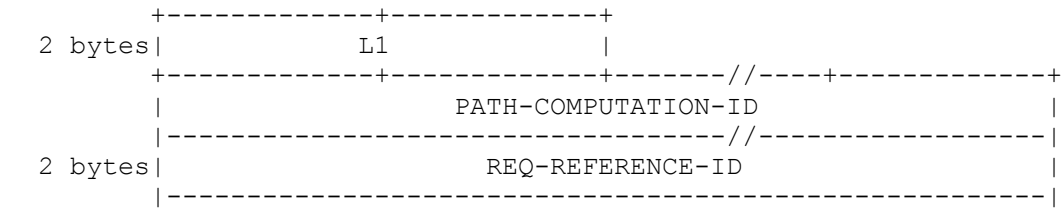
Copyright © MESCAL Consortium, June 2005

- o PATH-COMPUTATION-ID: is a globally unique value that identifies a path computation occurrence. It is a variable-length field. This identifier MUST be the same as the one provided by the REQUEST.

- o REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.

8.9. ACKNOWLEDGE message

This message is sent by a PCC to a PCS to confirm the reservation of the path. This feature is particularly used when a PCC launches multiple REQUEST messages during its path computation phase.



- o L1: is the length in bytes of the PATH-COMPUTATION-ID. Size of this field is 2 bytes.

- o PATH-COMPUTATION-ID: is globally unique value that identifies a path computation occurrence. It is a variable-length field. This identifier MUST be the same as the one provided by the REQUEST.

- o REQ-REFERENCE-ID: is a 2 bytes length field representing an unsigned integer. This field is used to reference the initial REQUEST.

8.10. KEEPALIVE message (KA)

Message exchanged between two PCEs to maintain TCP session when no other messages are exchanged.

This message has no argument.

9. Exchange of PCP messages

9.1. Communication

The PCP protocol uses a single persistent TCP connection between a PCC and a remote PCS. One PCE server implementation per server MUST listen on a well-known TCP port number (to be defined). The PCC is responsible for initiating the TCP connection to the PCS. The location of the remote PCS is deduced and retrieved from the management plane blocks during the path computation process or at PCS

Copyright © MESCAL Consortium, June 2005

boot via the SLS management block. PCE can have crossed communication; some are acting as a client role, others as a server role.

9.2. OPEN (OPN)

An OPN message MUST be sent before any other message exchange. As part of the open message, the PCC provide its PCSID, which allows the server to identify the client. It can also use this information to retrieve the client context near its management plane. Only one OPN message can be issued at a time.

If the PCS receives malformed message it MUST close the session using the appropriate error code.

9.3. ACCEPT (ACP)

The ACP message is used to positively respond to the OPN message from the PCC. This message will return to the PCC a KA-timer value object indicating the maximum acceptable intermediate time between the generation of messages by the PCEs. The KA-timer value is determined by the PCS and is specified in seconds.

If the PCS refuses the PCC open message, it will instead issue a CLOSE message.

9.4. CLOSE (CLO)

The CLOSE message can be issued by either the PCC or the PCS to notify the other that it is no longer available.

The Error code is included to describe the reason for the close.

When issuing a CLOSE both the PCC and the PCS MUST delete all the internal states related to this PCP session. Additionally, all pending requests MUST be cancelled in order to free as much as possible all pending resources reservations that could have been established. PATH-ERROR or CANCEL message must be sent depending on requests' state.

9.5. REQUEST (REQ)

A request is issued by a PCC when it has found a potential path toward the target final destination. This request can be issued as a consequence of a request received from another domain it has agreement with or from its own service management plane.

When the service request comes from a remote PCC, the server achieves the following tasks:

Copyright © MESCAL Consortium, June 2005

(0) If the receiving TTL is zero the PCS MUST discard the request. The receiving PCS, decrements by one the received TTL value. If the TTL is equal to zero, the request is rejected if the PCS is not the last PCS in the chain. In addition the PCS examines the AS-PATH included in the received REQ and reject it if it finds its own AS number in the list. This mechanism allows avoiding possible loops when a limited set of QoS constraints are provided in the request.

(1) It checks if the PATH-COMPUTATION-ID of the received REQ is already associated to a pre-contract or contract for the same requester. If this is the case, it returns a PATH-ERROR message with a reason-code = 4. It checks if the PATH-COMPUTATION-ID and the REQ-REFERENCE-ID of the received REQ are already associated to a pre-reservation record concerning the same requester. If a pre-reservation is found, it returns a PATH-ERROR message with a reason-code = 4.

(2) It considers the HEAD-END-ADDRESS and the TAIL-END-ADDRESS parameters present in the request. The HEAD-END-ADDRESS MUST indicate a valid entry point in its domain. If not, the PCS returns a PATH-ERROR with an appropriate reason value.

(3) Then it extracts the PCSID from the TAIL-END-ADDRESS and parses the QoS constraints provided at part of the request message. It has thus identified all QoS-class required together with their associated QoS constraints.

(4) The PCS achieves some policing and verifies that the request constraints will not exceed the resources negotiated in the pSLS. If resources are exceeded, the PCS returns a PATH-ERROR message. If resources are available, the PCS pre-reserves the corresponding resources near the management plane.

(5) If the PCS recognizes its own PCSID in the TAIL-END-ADDRESS, it considers the PATH-REFERENCE-ID otherwise it jumps to step (6). If this identifier is known from its management plane, the request is accepted and processing continues on (51). Otherwise the PCS returns a PATH-ERROR message with a reason-code = 2.

(51) The PCS computes an intra-domain path and verifies the availability of the resources along this internal path. If available, the PCS interacts with its management plane and creates a context, which triggers the administrative reservation of the resources. When interacting with the management blocks, the PCS MUST provide all information necessary to identify the sub-path it selected. In particular it MUST provide the PATH-COMPUTATION-ID, the REQ-REFERENCE-ID, the ingress point ASBR address used in its domain and the termination point in its domain. The PCS sends a RESPONSE-PATH message back to the

Copyright © MESCAL Consortium, June 2005

requesting PCC. If resources are not available a PATH-ERROR message is generated.

(6) It then queries the dynamic inter-domain traffic-engineering block with the retrieved PCSID and the list of requested QoS-classes. The dynamic inter-domain TE block returns the available BGP announcements. The PCS then verifies whether it can find a next-hop ASBR, which announces the PCSID within the requested QoS-class. If cannot find it the procedure stops and a PATH-ERROR message is returned back to the requesting entity with an appropriate reason-code value.

(7) If one or several next-hops are found, the PCS examines the QoS performance guarantees of the announcements and compare the values with those requested in the request. If it doesn't understand one of the requested QoS constraints, PATH-ERROR message is sent back. Otherwise, QoS constraints are successively compared to those received from q-BGP. All next-hops propagating the set of announcements satisfying the required QoS constraints are kept. The others are left on side.

(8) For each possible next hop ASBR the PCS checks if there are enough available resources available at the domain boundaries. In particular if some bandwidth guarantees are required the PCS checks if the administrative maximum bandwidth agreed during the pSLS negotiation phase will not be exceeded. If resources are not available the ASBR is left on side and the next ASBR in the list is considered. If resources are available, the PCS pre-reserves the corresponding resources near the management plane. At this stage, the management plane doesn't create any contract since we are not sure that an end-to-end path exists. This pre-reservation can be taken into account by the PCS for subsequent requests. It can use it as a lock and delay the incoming requests or introduce the pre-reservations in its resource availability computation according to the local policy enforced. When interacting with the management blocks, the PCS must provide all information necessary to identify the sub-path it selected. In particular it must provide the PATH-COMPUTATION-ID, the REQ-REFERENCE-ID, the ingress point address of its domain and the ingress point address of the next domain. This latter information can be used by the management plane to identify the upstream and downstream involved domains.

- (81) The PCS computes an intra-domain path and verifies the availability of the resources along this internal path. If resources are available, the sub-path is valid and the PCE forms a new REQUEST message which is sent to the PCS of the remote domain owning the next-hop ASBR. It adds its own AS number to the existing list. If internal resources are not available, the PCS discard the pre-reservation and considers

Copyright © MESCAL Consortium, June 2005

the next next hop ASBR in the list. When building the request the PCC keeps the PATH-COMPUTATION-ID, the PATH-REFERENCE-ID, the TAIL-END-ADDRESS unchanged. The initial HEAD-END-ADDRESS is replaced by the address of the ingress next-hop ASBR identified during the path computation. The QoS constraints characteristics are modified in order to take into account the QoS performance guarantees provided by the domain.

(9) If QoS constraints cannot be satisfied for any of the ASBR, the PCS returns a PATH-ERROR message.

Note that it is quite possible that several next hops ASBR can satisfy the requested constraints. In such a case the PCS can process one next-hop ASBR at a time or several in parallel. For one incoming request, there can be multiple simultaneous outgoing requests towards different PCS. If several requests are sent toward the same neighbor, for a same PATH-COMPUTATION-ID, the REQ-REFERENCE-ID must be different. Nevertheless, this feature can lead to scalability issues and needs further investigations.

9.6. RESPONSE (RSP)

A RESPONSE message is sent by a PCS in response to a request issued by a PCC. RSP messages are sent back when a valid end-to-end path has been computed. The RSP message MUST be initiated by the tail-end domain.

When a valid end-to-end path has been computed, the PCS of the last domain on the path, forms a RSP message. It first inserts the original PATH-COMPUTATION-ID. Then its forms a path argument that MUST contains the IP address of the tail-end LSP and the IP address interface of the ingress ASBR supporting that path. It MAY insert between these two extremities, the IP address of additional hops. It MAY also indicates the date after which the path will not be valid anymore because administratively reserved resources will have been relaxed. Then, it MUST indicate QoS guarantees it provides between the ingress ASBR and the tail-end address of the LSP. The RSP message is then sent to the requesting PCC.

On receipt, the PCC adds its own intra-domain sub-path to the list. It does not indicate the next-hop ASBR since this latter has already been inserted by the downstream PCS. This sub-path can be a strict or loose description. It also modifies the QoS guarantee parameters so that they reflect the QoS guarantees it can provide for its part of the path. This is achieved in the same way as for the request, but it is an "addition" operation if we consider the delay, for example. The VALIDITY-DATE MUST modified so that the value indicates now the smaller date between the date received in the RSP message and the date reported by the management plane.

Copyright © MESCAL Consortium, June 2005

If the PCC sent multiple REQUEST messages in parallel, it MAY wait for a RSP or ERR message for all the requests it sent. If the PCC got multiple RSP messages it MUST select only one and inform the un-selected PCS that they can cancel their reservation. It forms CANCEL messages, sends them to the appropriate PCS and cancels its own pre-reservation for the corresponding requests. If the PCC doesn't whish to wait for a reply, it can send a CANCEL message at any time.

The PCS can send the consolidated RES message to the requesting PCC after sending ACK message to the PCS it decided to keep in the path.

9.7. ACKNOWLEDGE (ACK)

The ACK message is used by PCS to confirm to its management plane that the resources needed for the path referenced by PATH-COMPUTATION-ID and REQ-REFERENCE-ID present in the message need to be reserved. It allows the management plane to create a contract based on information previously stores by the PCS during the computation phase. If no ACK is received, no contract is created and the negotiation at the management level will fail. If for some reasons, no ACK were received, the VALIDITY-DATE would be used and the administrative pre-reservation automatically removed for that path. ACK messages are only accepted if they arrive after the server has issued a RSP otherwise they are ignored.

9.8. CANCEL (CCL)

A CANCEL message can be sent by PCC and PCS. CCL messages can be generated during the normal path computation cycle but also in case of an abnormal termination of a PCE to PCE communication.

If a PCE, acting as a server for the PCP session, received a CCL message from the PCC, it MUST form new CCL messages and forward a CCL message to each PCS to which it sent a REQ for which it did not received any positive or negative reply. Once this has been achieved it MUST delete all its internal states referencing the request identified by the PATH-COMPUTATION-ID and REQ-REFERENCE-ID indicated in the message. If the PCE has no pending request concerning this PATH-COMPUTATION-ID and REQ-REFERENCE-ID, it can optionally query its management plane to retrieve a possible existing contract referenced by this PATH-COMPUTATION-ID and delete it. Just before deleting this contract, it can form a new CCL message and forward it to the next PCS in the path. If it does not, the VALIDITY-DATE will be applied.

The same procedure applies if the PCE server detects a communication problem with one of its PCC. In that case, the PCS issues CCL messages for all pending request received from this PCC.

When a PCE, acting as a client for the PCP session, received a CCL message from a PCE server, this indicates that a PCS along the path towards the target destination has experienced communication problems

Copyright © MESCAL Consortium, June 2005

leading to close a PCP communication. In such a case, each PCC cancels all the internal states referencing this PATH-COMPUTATION-ID and forward this indication to the upstream client PCS up to the initial requestor.

10. Security Considerations

PCP is a communication protocol that is used between two PCEs. No security mechanisms are defined in this PCP specification. It is recommended that a security protocol like IPSec or TLS MUST be activated in order to protect PCP sessions.

11. References

[RFC3667] Bradner, S., "IETF Rights in Contributions", RFC 3667, February 2004

[RFC3668] Bradner, S., "Intellectual Property Rights in IETF Technology", RFC 3668, February 2004

[INTERAREA-REQ] Le Roux, J., Vasseur, JP, Boyle, J., "Requirements for Inter-Area MPLS Traffic Engineering ", draft-ietf-tewg-interarea-mpls-te-req-03.txt, November 2004

[INTERAS-REQ] Zhang, R., Vasseur, JP., et. al., "MPLS Inter-AS Traffic Engineering requirements ", draft-ietf-tewg-interas-mpls-te-req-09.txt, September 2004

[PCE-ARCH] Ash, J., Farrel, A., Vasseur, JP., "Path Computation Element (PCE) Architecture", draft-ietf-pce-architecture-00.txt, March 2005

[PCE-FWK] Farrel, A., Vasseur, JP., Ayyangar, A., "A Framework for Inter-Domain MPLS Traffic Engineering", draft-ietf-ccamp-inter-domain-framework-01.txt, February 2005

[INTERAS-PCE] Boucadair, M., Morand, P., "A Solution for providing inter-AS QoS tunnels", draft-boucadair-pce-interas-01.txt, May 2005

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

[PCE-DISCOVERY] Boucadair, M., Morand, P., "PCE Discovery via Border Gateway Protocol", draft-boucadair-pce-discovery-01.txt, May 2005

[RFC2401] Atkinson R., "Security Architecture for the Internet Protocol", RFC 2401, August 1998.

Copyright © MESCAL Consortium, June 2005

[RFC2246] Dierks T., Allen C., "The TLS Protocol", RFC 2246, January 1999

12. Acknowledgments

The authors would also like to thank all the partners of the MESCAL (Management of End-to-End Quality of Service Across the Internet At Large, http://www.mescal.org) project for the fruitful discussions.

13. Author's Addresses

Mohamed Boucadair
France Telecom R & D
42, rue des Coutures
BP 6243
14066 Caen Cedex 4
France
Phone: +33 2 31 75 92 31
Email: mohamed.boucadair@francetelecom.com

Pierrick Morand
France Telecom R & D
42, rue des Coutures
BP 6243
14066 Caen Cedex 4
France
Email: pierick.morand@francetelecom.com

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary

Copyright © MESCAL Consortium, June 2005

rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Statement

Copyright (C) The Internet Society (2005). This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

Copyright © MESCAL Consortium, June 2005