

STRONGEST – Document

## Deliverable D3.2

### Next generation transport networks: efficient solutions for OAM, control, and traffic admittance

Version and Status	Version 2.0, final	
Date of issue:	31.12.2010	
Distribution:	Public	
Author(s):	Name	Partner
	Berechya, David	NSN-I
	Bincoletto, Luca	TI
	Botham, Paul	BT
	Broniecki, Ulrich	ALUD
	Casellas, Ramon	CTTC
	Castoldi, Piero	CNIT
	Corliano Gabriele	BT
	Cugini, Filippo (editor)	CNIT
	Di Giglio, Andrea	TI
	Garcia Argos, Carlos	TID
	Giorgetti, Alessio	CNIT
	Georgiades, Michael	PRIMETEL
	González de Dios, Oscar	TID
	Gutkind, Dana	NSN-I
	Iovanna, Paola	TEI
	Javier Jimenez Chico, Francisco	TID

	Lautenschläger, Wolfram	ALUD
	Loffredo, Tullio	TI
	Maier, Guido	CNIT
	Marchetti, Loris	TI
	Margaria, Cyril	NSN-G
	Martinez, Ricardo	CTTC
	Milbrandt, Jens	ALUD
	Morro, Roberto	TI
	Muñoz, Raul	CTTC
	Muñoz del Nuevo, Fernando	TID
	Paolucci, Francesco	CNIT
	Pulverer, Klaus	NSN-G
	Rambach, Franz	NSN-G
	Sfeir, Elie	NSN-G
	Siracusa, Domenico	CNIT
	Zema, Cristiano	TEI
Checked by:	Vezzoni, Emilio	VECOMM
	Di Giglio, Andrea	TI
Approved by:	Iovanna, Paola (WP3 leader)	TEI

## **Abstract**

The STRONGEST WP3 on “end-to-end solutions for efficient networks” aims at providing an efficient end-to-end control plane architecture for single/multi domain, single/multi region and single/multi carrier networks.

The activities reported in this deliverable focus on medium-term network scenario, which addresses the inter-working between heterogeneous GMPLS-controlled networks.

In particular, this document reports the activities carried out in the context of: (1) OAM parameters and mechanisms, (2) control plane architectures, solutions and proposed extensions (with specific target on the PCE architecture) and (3) end-to-end services and traffic admittance solutions.

## Table of Contents

<b>Table of Contents</b>	<b>4</b>
<b>Executive summary</b>	<b>8</b>
<b>1 Introduction</b>	<b>9</b>
<b>2 OAM parameters and mechanisms</b>	<b>10</b>
2.1 Path Packet Loss Ratio	10
2.1.1 The nature of packet loss	10
2.1.2 Existing solutions and known problems	12
2.1.3 Loss rate prediction	13
2.2 BFD Tools: performance evaluation	14
2.3 Multi-domain and multi-carrier end-to-end OAM	17
2.3.1 Proposed OAM Mechanism	17
2.3.1.1 MP functionalities	18
2.3.1.2 OAM flows	20
2.4 OAM mechanisms: considerations and future work	22
<b>3 Control Plane Architectures, Solutions and proposed Extensions</b>	<b>24</b>
3.1 Control plane in a single-domain scenario	24
3.1.1 IGP scalability - motivation	24
3.1.2 IGP scalability - analysis	25
3.2 Control plane in a multi-domain scenario: hierarchical PCE	28
3.2.1 Proposed Architecture	28
3.2.2 Proposed Control Plane Extensions	31
3.2.3 Case study and performance evaluation in multi-domain WSON	32
3.2.4 A multi-domain hierarchical PCE-based system for Domains Topology creation	36

3.3	Control plane in a multi-carrier scenario	40
3.3.1	Multi-domain PCE-based architectures	40
3.3.2	BRPC-based mechanism for MPLS-TP / WSON networks	41
3.3.3	Hierarchical path vector protocol for multi-carrier networks	44
3.3.4	Confidentiality in multi-carrier PCE-based networks	46
3.3.4.1	Confidentiality issues in PCEP	47
3.3.4.2	Proposed Policy-based Architecture	47
3.3.4.3	Authorization policy implementation	49
3.3.4.4	Experimental assessment.	50
3.3.4.5	Conclusions and future work.	51
3.4	Control plane in a multi-layer scenario	51
3.4.1	Multi-layer PCE-based architecture	51
3.4.1.1	Simulation results	53
3.5	Specific issues in path computation	55
3.5.1	Point-to-Multipoint (P2MP)	55
3.5.1.1	Core Tree Computation Procedures	56
3.5.1.2	Sub Tree Computation Procedures	57
3.5.2	Topology summarization method	57
3.5.3	Management and adaptation of PCE algorithms	63
3.5.3.1	Motivation	63
3.5.3.2	PCE algorithm management alternatives	64
3.6	RACS/PCE integrated architecture in inter-domain/inter-carrier scenario	67
3.6.1	Proposed architecture in inter-domain/inter-carrier scenario	67
3.6.2	GMPLS Control Plane considerations	70
3.7	PCEP extensions for GMPLS networks	71
3.7.1	Proposed PCEP extensions and standardization update	72

3.7.2	PCEP Extensions for Temporary Reservation of Path Resources	75
3.7.2.1	PCEP Proposed Extensions	76
3.7.2.2	Application case: Multiple LSP Restoration	77
3.8	Future Works	79
<b>4</b>	<b>End-to-end Services set up and Traffic Admittance</b>	<b>81</b>
4.1	Service Definition	81
4.1.1	Business drivers and requirements	81
4.1.2	Service Reference Model	83
4.1.3	Implications for STRONGEST	86
4.2	Mapping of End-to-End Services	87
4.3	Inter-Domain end-to-end QoS and OAM signaling over heterogeneous domains	91
4.3.1	Signaling Considerations and SoA for Inter-Domain QoS and OAM functionality communication	91
4.3.2	Signaling Requirement Considerations for Inter-Domain Networks	92
4.3.3	NSIS suitability for the STRONGEST architecture	93
4.4	Application of RACS/PCE architecture	94
4.5	Delay-based Metric Abstraction for OIF E-NNI	95
4.5.1	TE Metric for delay-critical applications	95
4.5.2	PCE-RC Architecture	96
4.5.3	Metric Abstraction Schemes	99
4.5.4	Conclusions	104
4.6	Considerations and future works on end-to-end services	104
<b>5</b>	<b>Conclusions</b>	<b>106</b>
<b>6</b>	<b>References</b>	<b>107</b>
6.1.	STRONGEST Publications	107
6.2.	Informative references	107



**STRONGEST**  
*Scalable, Tunable and Resilient Optical  
Networks Guaranteeing Extremely-high  
Speed Transport*

**Next generation transport  
networks: efficient solutions for  
OAM, control, and traffic  
admittance**

**D32 2.0.doc**

<b>7. Document History</b>	<b>111</b>
<b>8. Acronyms</b>	<b>112</b>

## Executive summary

The activities reported in this deliverable, carried out within STRONGEST WP3, are focused on the medium-term network scenarios which address the inter-working between heterogeneous GMPLS-controlled networks. In particular, on the basis of the reference scenario described in the STRONGEST deliverable D3.1, the following three main aspects have been addressed: (1) OAM parameters and mechanisms, (2) control plane architectures, solutions and proposed extensions and (3) end-to-end services and traffic admittance solutions.

Chapter 2 focuses on OAM mechanisms, providing considerations and novel effective solutions to (i) predict, (ii) monitor, (iii) quantify and (iv) certify SLA degradation in packet transport networks due to packet loss induced by congestion, physical impairments or network failures.

Chapter 3 addresses control plane architectures, providing innovative solutions aiming at improving the effective provisioning of application services while guaranteeing an efficient utilization of network resources, specially, in multi-domain scenarios. In particular, several procedures and solutions are proposed in the context of the PCE-based architecture (including hierarchical PCE), hierarchical routing protocols and architectures, abstraction schemes, PCEP extensions and RACF-based solutions.

Chapter 4 then applies some of the main innovative solutions defined within WP3 on specific network scenarios, aiming at supporting advanced end-to-end services. Innovative solutions are presented in the context of OAM, RACF-based traffic admittance, and OIF E-NNI and abstraction schemes for services with strict delay constraints.

The main scientific publications and standardization documents produced during the first year of STRONGEST WP3 activities are also highlighted within the Reference section, thus demonstrating the relevant impact of the proposed technical solutions within the research and standardization communities.



## 1 Introduction

This document includes the main considerations and results that have been achieved within STRONGEST WP3 in the context of the medium-term reference scenarios which include single/multi domain, single/multi carrier and single/multi-layer networks. These scenarios, as well as the reference architectures and the state-of-the-art of the current solutions are reported in the STRONGEST deliverable D3.1. On the basis of D3.1, this document reports on the main WP3 activities and innovative solutions proposed for improving the performance of OAM, control plane and e2e service delivery.

OAM activities, reported in Chapter 2, target the ability of a packet transport network to support services with guaranteed and strict service level agreements (SLAs) while reducing its operational costs. The focus is on the prediction and measurement of SLA degradation (e.g., due to packet loss). Innovative solutions, considering both BFD-based and Y1731-based procedures, are then investigated and proposed to improve the possibility to detect and certify SLA degradation, also for commercial purposes.

Control plane considerations and solutions are reported in Chapter 3. As reported in D3.1, currently available multi-domain solutions are not able to guarantee the adequate level of performance in traffic engineering and scalability, and several mechanisms are still inadequate or missing. In this document, the main focus is on single/multi domain and single/multi carrier solutions with significant emphasis on the PCE-based architectures. First, considerations on routing protocol and path computation are reported to assess the scalability performance of a large scale single routing area/domain. Then, innovative multi-domain procedures and solutions are proposed and evaluated, with particular emphasis on the promising hierarchical PCE architecture. Solutions are also proposed and evaluated in the specific scenario of multi-carrier networks where confidentiality needs to be preserved. Novel control plane mechanisms are proposed in the context of point-to-multi-point, WSON and multi-layer networks. In addition, specific issues for path computation (e.g., on the management of different computation algorithms) and RACF-PCE integration are considered and successfully addressed.

The final goal of any transport network and, more specifically, on the technical solutions in OAM and control plane, is to provide the infrastructure for efficient network service provisioning. Chapter 4 proposes the mapping between application to service classes and to transport technologies, by successfully applying some of the solutions defined within the STRONGEST project to the provisioning of QoS-guaranteed services.

## 2 OAM parameters and mechanisms

Effective OAM (Operations, Administration and Maintenance) procedures are required to support services with guaranteed and strict Service Level Agreements (SLAs) in GMPLS-controlled networks, while reducing their operational costs.

OAM terminology, requirements, and state-of-the-art mechanisms for performance monitoring (loss/degrade of received information) and fault management (detection and localization) have been reported in [D3.1]. In particular, [D3.1] summarizes the two competing standardization proposals and solutions either based on BFD-tools or derived from [Y.1731]. Both solutions are considered and investigated. In particular, work is ongoing within the STRONGEST project aiming at defining effective procedures for failure detections and localization in packet transport networks, including both major and minor causes of SLA degradation.

In this chapter, SLA degradation due to packet loss is investigated at first. Three causes are considered: (i) network congestion, (ii) link transmission quality degradation and (iii) major failures. The first two causes of SLA degradation typically induce sporadic packet loss while the latter induces continuous packet loss with major implications in service provisioning.

Then this chapter reports considerations on loss rate prediction and measurement, and investigates the suitability of existing OAM tools to provide effective performance monitoring and failure localization. In particular, the performance of the BFD tool is experimentally analyzed in case of minor SLA degradations.

Finally an innovative and reliable mechanism derived from [Y.1731] is proposed and described to efficiently provide failure localization and measure the fault duration.

### 2.1 Path Packet Loss Ratio

#### 2.1.1 The nature of packet loss

Circuit switched network technologies like SDH or OTH implement sophisticated quality monitoring of Bit Error Rate (BER). Monitoring of the BER per section (link) and per path (end-to-end) allows for early detection of physical equipment degradations, their localization, isolation (protection switching) and proactive maintenance. BER estimation in OTH is based on parity check sums in the protocol overhead (section monitoring fields – SM, and path monitoring fields – PM in the ODU header). The bit error rate in operational SDH/OTH networks is typically better than 10<sup>-12</sup>.

In packet switched networks the packet loss probability is a similar quality indicator. Particularly for connection oriented technologies like MPLS the corresponding

OAM functions for link and path monitoring are requested with the same intention as in SDH/OTH networks – early detection of degradations and proactive intervention. Nevertheless there are remarkable differences between bit error rate and packet loss probability, which must not be ignored when designing OAM functions for connection oriented packet networks.

First of all, packet loss is not a direct translation of bit error rate. Packet loss due to bit errors can be estimated by the following example: Ethernet packets are protected by a CRC32 check sum that enables the receiver to identify corrupted packets with 1 to 3 bit errors. Corrupted packets are dropped as a whole. A single packet consists of  $n=12000$  bit. Hence the drop probability due to CRC failures is  $P \cong n \cdot BER$ , which would be  $P=12000 \times 10^{-12} \approx 10^{-8}$ . In contrast to these quite low loss figures, the observed packet loss probability in the operational Internet today is  $10^{-4}$  up to  $10^{-2}$ , which is 4 orders of magnitude higher! Obviously packet drop in the Internet is dominated by effects other than BER.

Packet drop in congested networks is the sole result of a temporary resource deficit. If a packet arrives at some networking device, the device should immediately process or forward the packet, or at least put it into a temporary storage. If at the given moment the storage is full, the packet gets lost, no way out. It does not matter if the space was there just before or just after the incident, or in some average. The drop decision is a singular event, just one clock cycle. Particularly the drop cannot be avoided by further processing or signaling, it would be too late. The only way to avoid packet drops is a network operation in a way, where the resource deficit is avoided beforehand. But this cannot be done locally in the affected device. In other words, there is no such thing like a built-in “do not drop” feature.

Packet forwarding is a rather uncoordinated process. Traffic engineering, admission control or similar activities are directed on the integral amount of packets that could be arriving. The arrival of a particular packet, however, is more or less random and unpredictable. In consequence the resource occupation in network devices is random, too. The danger of a resource deficit depends on the average amount of arriving traffic, the traffic load, and on the extent, how far the actual traffic fluctuates around the mean. It is commonplace that the load is different from link to link and that it is changing over time, and so are the fluctuations. If we now explore the packet loss probability by direct observation or as deduced from load and volatility, it must be clear that we observed the past of a user population (that created the traffic), and extrapolate its activity into the future. This is a fundamental difference with respect to the bit error rate (BER) where we intrinsically observe the past of some network devices (fibers, amplifiers, lasers) and extrapolate their physical properties to the future. The difference is obvious: The device parameters are purposely built to be constant, but possibly slowly degrading. Extrapolation is straightforward. In contrast, for the user population of a particular network link, the observed activity must be sufficiently large to get at least some reliable statistical data. On the other hand the observation and prediction periods must be sufficiently short to fulfill the invariance hypothesis for the extrapolation.

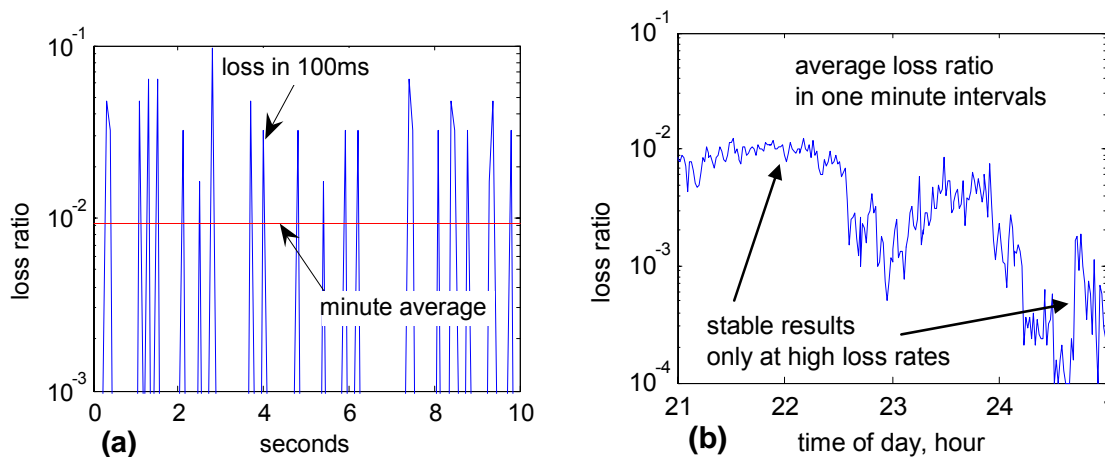
Conclusion: Packet loss occurs due to physical impairment (BER) and due to temporary resource deficit (congestion). BER induced packet loss dominates uncongested links. But it can be neglected as soon as congestion occurs. Packet loss due to congestion

is load dependent. Its acquisition and prediction times are not arbitrary. They are inherently a compromise between statistical relevance of the observed activity (the longer the better) and the assumption of unchanged conditions (the shorter the better). Signaling protocols for packet loss probabilities have to take these temporal limitations into account.

### 2.1.2 Existing solutions and known problems

The packet loss probability can be observed directly by periodic read out of packet counters at ingress and egress of links or paths [Y.1731]. Alternatively the gaps in packet number sequences can be counted at the receiving end of a connection – given the packets include reliable sequence numbers. Both methods are applicable in different circumstances but they yield more or less the same result: the fraction of lost packets out of a certain amount of offered traffic during an interval of time. The default time interval is set to 100ms in [Y.1731].

Unfortunately the extrapolation of results is difficult for several reasons: (i) The better a connection is (i.e. low loss probability), the less loss events are encountered. It takes extremely long time to acquire sufficiently large samples. (ii) Connections with zero losses do not tell anything about how far they are away from the loss limit. (iii) Losses are clustered in bursts [Bor98]. Measurements converge even slower than expected. Short term deviations from long term mean are arbitrary large. For illustration we show results from our own Internet measurements. These results were obtained by the sequence number method on an arbitrary path from a University campus network to a residential home access.



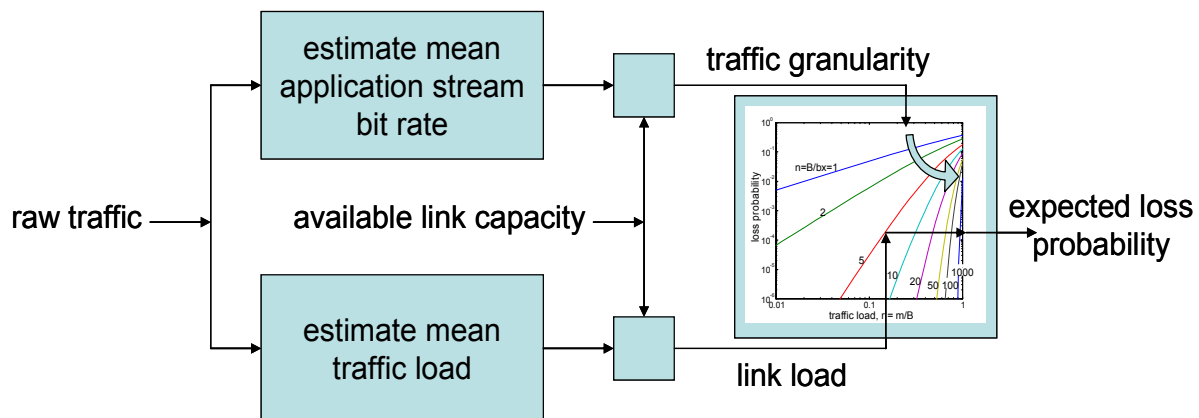
**Figure 1: Internet packet loss measurements - Losses are not uniformly distributed**

The observed loss rates of Figure 1a show the unstable statistics over short time intervals of 100ms. Most of the intervals exhibit no losses at all, while some intervals suffer from loss bursts with an order of magnitude higher loss rate than on average. The bursty nature of loss events disappears when measured over larger one minute intervals, Figure 1b. But then we are faced with another problem: The loss statistics are stable during the rush hour until 22:30. After midnight, however, loss incidents are so rare that even minute intervals do not collect sufficiently large samples. The resulting values do change from minute to minute by an order of magnitude or more.

In the following we propose two alternative methods for the analysis of packet loss processes. The first one is applicable to congestion induced packet loss. It tries to overcome the problem of slow convergence due to the burstiness of the loss process by taking into account the whole traffic fluctuation and not only the congested periods. The second proposal is addressing the uncongested case, where physical impairments dominate the loss process. In particular it aims at defect detection in the context of existing and standardized network hardware without specific OAM functionality, e.g. native Ethernet.

### 2.1.3 Loss rate prediction

In [LAU08] we introduced a method that calculates the expected packet loss probability of a link derived from its actual load and the known end-user access link capacities. Recent work suggests that the knowledge of end-user access link capacities can be substituted by an estimate of the mean application stream bit rate of aggregated traffic. This way it could be sufficient to observe the evolution of short term load and its variance to calculate the expected loss probability.



**Figure 2: Prediction of loss probability based on actual traffic and available capacity**

Figure 2 shows a block diagram of the proposed solution. Link load and traffic granularity are both derived from raw traffic that is actually traversing the link. As such both of them are statistical estimates rather than exact measurements. In that the accuracy of our method does not differ from the direct acquisition of loss rates as explained above. On the other hand the estimation takes into account the whole traffic with all its fluctuations and not only the rare overflow events as in the direct acquisition. For this reason our method should be 2 to 4 orders of magnitude faster than the direct loss rate acquisition.

The indirect packet loss prediction is applicable per link. It can be continuously calculated locally in intermediate nodes, independent of the assignment of connections or LSPs etc. To get a path packet loss ratio, the particular link specific values need to be concatenated by an appropriate signaling protocol. In the simplest case this could be an extension to an existing path signalization, e.g. ping. This would be interesting for a continuous path monitoring with continuously flowing traffic. In the case of newly to be established LSPs the signaling could include a load forecast for the envisaged connection.

Then each intermediate node could include the announced load into the loss calculation according to Figure 2.

We will investigate the methods as drafted above by an experimental signaling protocol for end-to-end path packet loss ratio prediction.

## 2.2 BFD Tools: performance evaluation

In [RFC5880], the Bidirectional Forwarding Detection (BFD) protocol is defined as a protocol intended to detect faults in the bidirectional path between two forwarding engines or interfaces, potentially with very low latency. BFD has been mainly designed to provide fast fault detection on media not equipped with native OAM mechanisms (e.g., Ethernet). BFD can also be applied in a (multi-domain) Label Switched Path (LSP) to verify the liveness of the end-to-end connection. BFD resorts to LSP Ping to establish the BFD session, which consists in the exchange of simple *Hello* packets. BFD has two operating modes. In the first *Asynchronous* mode, the systems periodically send BFD Control packets to one another, and if a number of consecutive packets are not received by the other system, the session is declared to be down. In the second *Demand* mode, once a BFD session is established, a system may ask the other system to stop sending BFD packets, except when the system needs to verify connectivity explicitly, in which case a short sequence of BFD packets is exchanged, and then the far system quiesces. An adjunct to both modes is the *Echo* function. When the Echo function is active, a stream of BFD Echo packets is transmitted in such a way as to have the other system looping them back through its forwarding path. If a number of packets of the echoed data stream are not received, the session is declared down.

In this study, we investigate for the first time the use of BFD not only as a mechanism to detect link or interface failures (which trigger the BFD down state), but also to verify the end-to-end SLA and to discover minor issues affecting the end-to-end connection. Indeed, in multi-domain networks controlled by different operators, the detailed verification of the transmission quality on any link of a multi-domain LSP is practically unfeasible for a single operator and additional mechanisms are required to evaluate the end-to-end SLA. The goal is to discover minor malfunctioning which do not trigger BFD down states (and in turn recovery schemes) but might induce some sporadic packet loss and the worsening of the provided SLA (particularly if multiple links along the same LSP are affected by minor malfunctioning).

To assess the BFD performance as a tool for SLA verification, the test-bed depicted in Figure 3 has been set up. It includes four commercially-available network nodes (LSR1-LSR4) connected through Gigabit Ethernet optical links (1000BaseLX). An LSP has been activated between LSR1 (Ingress) and LSR4 (Egress). Along the LSP, a BFD session has been activated: BFD Asynchronous mode with no Echo function (the sole configuration supported) has been configured on LSR1 having LSR4 as BFD session destination. Thus, two BFD flows have been practically activated: one from LSR1 to LSR4 along the LSP and one from LSR4 to LSR1 in the reverse direction. Both the BFD Transmitter and Receiver timers have been configured to  $T_C=100$  ms (minimum supported value  $T_{C-min}=50$ ms). Note that the Receiver Timer is configured at LSR1 but refers to the transmitter timer at LSR4, the session destination. Each BFD flow occupies 5.6 kb/s. To introduce some real physical degradation, a Variable Optical Attenuator (VOA) has been

inserted in the link from LSR1 to LSR2. A Network Generator and Analyzer (NGA) has been connected to LSR1 and LSR4 to send traffic along the LSP. The NGA traffic was activated only to measure the end-to-end packet loss rate and then disabled to focus just on BFD packets.

Several tests have been performed, considering also different BFD parameters. Here the following two tests are reported.

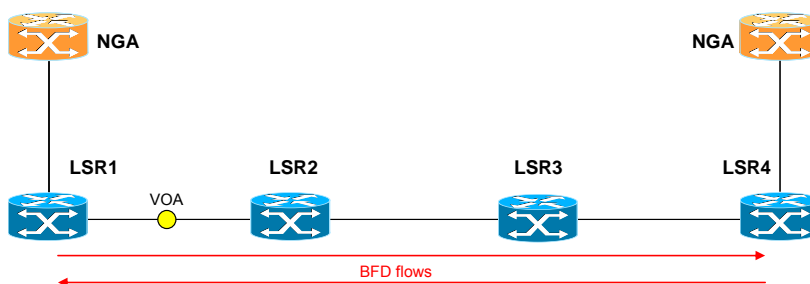
In the first test, no degradation was introduced by the VOA.

In 1 hour time, LSR1 sent  $N_{LSR1 \rightarrow LSR4} = 36939$  BFD packets along the LSP towards LSR4 and received  $N_{LSR4 \rightarrow LSR1} = 37027$  BFD packets from LSR4. The larger value of received packets is due to the independence of the two BFD flows in the considered BFD configuration (no echo function). The average time interval between two consecutive received BFD packets at LSR1, as expected, was  $T_{avg} = 100ms$ . The measured minimum and maximum intervals were  $T_{min} = 61ms$  and  $T_{max} = 140ms$  respectively (similar statistics for received BFD packets at LSR4).

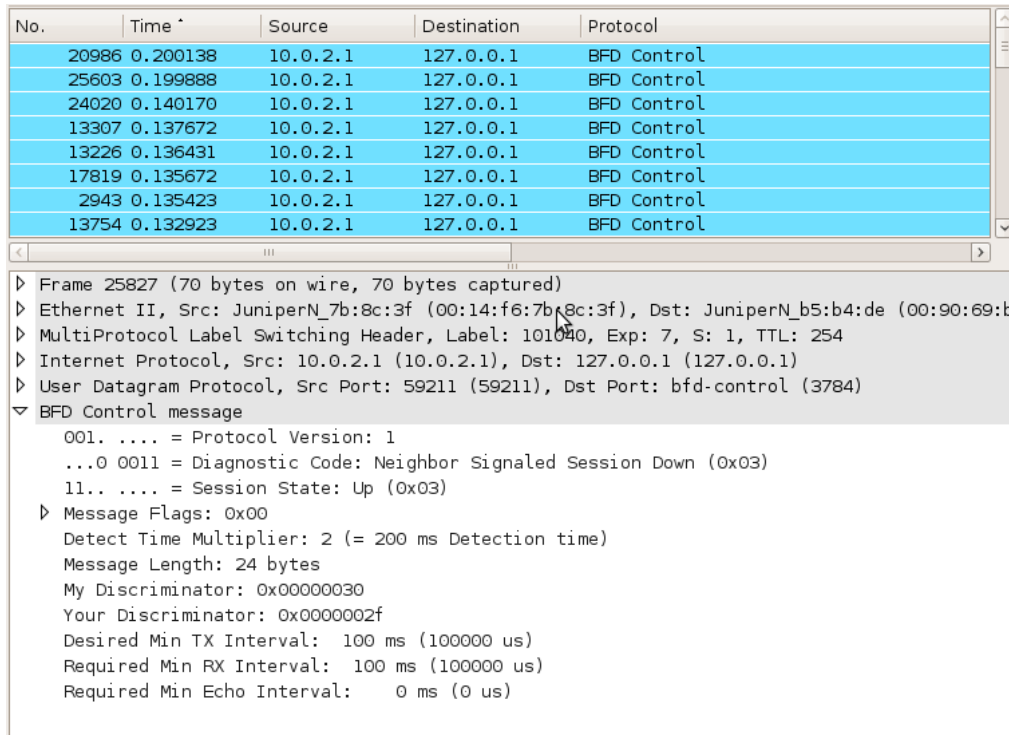
In the second test, the VOA has been tuned to introduce significant degradation on link LSR1→LSR2. The NGA traffic measured a packet loss rate of around  $10^{-5}$ .

The test, supposed to last one hour, abruptly terminated after 22 minutes when the link LSR1-LSR2 was torn down by the attached network nodes, thus inducing the BFD session termination. Before that instant, two time intervals between consecutive packets were discovered at LSR4 having values higher than the previous maximum value ( $T_{max} = 140ms$ ):  $T'_{max} = 200.1ms$  and  $T''_{max} = 199.8ms$ . Figure 4 shows the capture of some BFD packets received at LSR4 and ordered with respect to the time interval from the previous received BFD packet.

The result of two missing BFD packets in the direction LSR1→LSR4 was confirmed by the network interface at LSR2 which revealed two incoming packets with "Framing Error". At LSR1 no differences were identified with respect to the first test. Other tests under the same scenario confirmed the obtained results.



**Figure 3: Test-bed**



No.	Time	Source	Destination	Protocol
20986	0.200138	10.0.2.1	127.0.0.1	BFD Control
25603	0.199888	10.0.2.1	127.0.0.1	BFD Control
24020	0.140170	10.0.2.1	127.0.0.1	BFD Control
13307	0.137672	10.0.2.1	127.0.0.1	BFD Control
13226	0.136431	10.0.2.1	127.0.0.1	BFD Control
17819	0.135672	10.0.2.1	127.0.0.1	BFD Control
2943	0.135423	10.0.2.1	127.0.0.1	BFD Control
13754	0.132923	10.0.2.1	127.0.0.1	BFD Control

```
Frame 25827 (70 bytes on wire, 70 bytes captured)
Ethernet II, Src: JuniperN_7b:8c:3f (00:14:f6:7b:8c:3f), Dst: JuniperN_b5:b4:de (00:90:69:b5:b4:de)
MultiProtocol Label Switching Header, Label: 101040, Exp: 7, S: 1, TTL: 254
Internet Protocol, Src: 10.0.2.1 (10.0.2.1), Dst: 127.0.0.1 (127.0.0.1)
User Datagram Protocol, Src Port: 59211 (59211), Dst Port: bfd-control (3784)
BFD Control message
  001. .... = Protocol Version: 1
  ...0 0011 = Diagnostic Code: Neighbor Signaled Session Down (0x03)
  11.. .... = Session State: Up (0x03)
  Message Flags: 0x00
  Detect Time Multiplier: 2 (= 200 ms Detection time)
  Message Length: 24 bytes
  My Discriminator: 0x00000030
  Your Discriminator: 0x0000002f
  Desired Min TX Interval: 100 ms (100000 us)
  Required Min RX Interval: 100 ms (100000 us)
  Required Min Echo Interval: 0 ms (0 us)
```

**Figure 4: Capture at LSR4. Received BFD packets are ordered as a function of the interval of the previous received BFD packet**

### Considerations and future work

The second test confirmed that some malfunctioning (e.g., sporadic BFD packet loss) not only demonstrates a degradation in the provided SLA, but might also be an indicator for subsequent major failures.

Due to the independence between BFD flows, no considerations can be derived from the comparison between the amount of transmitted and received BFD packets. Thus, if no Echo function is enabled, BFD packet loss can be detected only at the destination node and no information can be derived at the source node. This complicates the management of BFD statistics and might require additional procedures to enable the utilization of BFD as a tool for SLA verification.

Even if the generation of BFD packets suffers from significant jitter (as experienced in the considered implementation), the interval between two consecutive received BFD packets is sufficient to identify SLA degradation in terms of packet loss. This confirms that BFD-based mechanisms can be used as an effective tool for SLA verification in a multi-carrier environment.

Future work will consider multiple LSPs running BFD on the same multi-domain meshed network infrastructure. The objective is to apply possible strategies and correlations between received BFD statistics aiming at identifying possible locations of malfunctioning. The analysis will take into account scalability issues in terms of both number of activated flows and BFD rate.



## 2.3 Multi-domain and multi-carrier end-to-end OAM

This chapter describes a novel OAM mechanism that provides reliable means to allocate failures and to measure the failure duration in the Multi Carrier scenario. This mechanism focuses on major failures (e.g., link failure) and allows the different carriers that operate services together to agree on the failure reason and on the out-of-service period. This chapter will elaborate the OAM system that has been described in [D3.1].

This innovative OAM mechanism addresses technical issues with significant commercial implications, particularly in the multi-carrier scenario. This mechanism adds further features to the existing OAM system, that currently deals with fault handling and SLA monitoring, but ignores other aspects with fundamental commercial impact.

### 2.3.1 Proposed OAM Mechanism

The mechanism which is proposed and described here can be considered as an add-on mechanism working on top of the current various OAM standards.

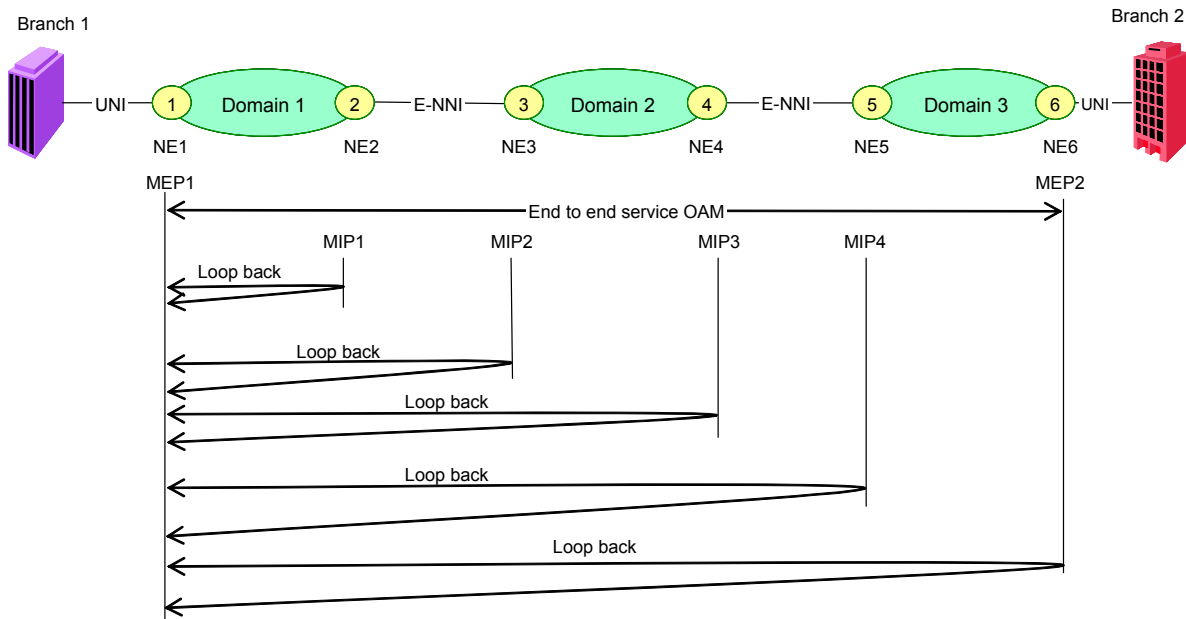
As described in D3.1, OAM for packet networks is defined by several standard bodies e.g. ITU-T, IETF and IEEE. These various standards are often tightly coupled, and have had a mutual effect on each other. The ITU-T and IETF have both defined OAM mechanisms for MPLS LSPs, IEEE and ITU-T for Ethernet networks.

We propose to use the OAM LB (Loop Back) as defined by the ITU-T Y.1731 for this purpose because it is suitable for the transport network as it is the basis for MPLS-TP OAM. In a later phase the other OAM standards will adopt this mechanism.

Y.1731 defines a protocol for OAM of Ethernet based networks, and it is also the basis for MPLS-TP OAM which is currently under study in IETF.

The loopback (LB) function verifies connectivity with a peer MEP or MIP. LBM messages are generated from the MEPs (Maintenance End Point) to each MIP (Maintenance Intermediate Point) along the service path and to the peer MEP. For this mechanism we propose a new mode of LB that allows LB messages to be generated periodically (other than the existing ITU-T Y.1731 where the LB is triggered by the management upon demand).

MIPs will be assigned to each inter domain connection point as described in Figure 5



**Figure 5: OAM network diagram**

Figure 5 describes an example network of Multi Carrier scenario (assuming each domain is operated by a different carrier). This network includes three domains, 4 MIPs and 2 MEPs. In this example the ME is a service between MEP1 and MEP2. MEP1 sends Loopback messages to the MIPs and to its peer MEP (MEP 2).

The following subsections elaborate the proposed OAM process and the functionalities of the MEPs and the MIPs.

### 2.3.1.1 MP functionalities

#### MEP Functionality

Maintenance End Point (MEP) is one of the end points of a Maintenance Entity (ME) which can be either a point-to-point or a point- to-multipoint relationship between two or more MEPs. MEP can initiate OAM messages and respond to them. MEP and ME are defined according to Y.1731 and IEEE 802.1ag terminology and have equivalents in the other standards.

#### LoopBack Messages

LoopBack Messages (LBMs) are generated by the MEP in proactive mode to all the MIPs along the service path and to the peer MEP. The period of the consecutive LB messages is configurable. Since the aim of this mechanism is to determine who is the domain that causes the failure' and the duration of the failure for commercial purposes, then the period of the consecutive LBs can be few seconds.

- **Maintenance Entity Group Level (MEG Level)** at which the MEP exists, MEG level defines the hierarchical level of the OAM.
- **Unicast address** of a remote MIP or the peer MEP to which LBM is intended.

- **Transaction ID/Sequence Number:** Transaction ID/Sequence Number is a 4-octet field that contains the transaction ID/sequence number for the LBM. The receiver (MIP or peer MEP) is expected to copy the Transaction ID/Sequence Number in the LBR PDU
- **Data** – Optional element whose length and contents are configurable at the MEP. The contents can be a test pattern and an optional checksum. etc. This Data field is used for the multi carrier OAM:
  - **The status** of the ME (signed with the MEP private key), when OK, means that up to the time that is indicated in the Real Time Stamp (see below) the service was delivered OK.
  - **Real Time stamp** (signed with the MEP private key) contains Year, Month, Hour, Minutes, Seconds
  - **MEP Identifier** (signed with the MEP private key), each MP has a unique identifier in the multi carrier network (the identifier method is out of the scope of this document, such a method can be found in FP7 Project ETNA)
- **Priority** – Identifies the priority of frames with Unicast LBM information.
- **Drop Eligibility** – Identifies the eligibility of frames with LBM information to be discarded when congestion conditions are encountered. When MIP or the peer MEP receive LB they reply with LB Reply (LBR)

#### Logging of LoopBack Replay messages

The LBR messages that are received from the MIPs and from the peer MEP can be used by the originator MEP in case of failure to locate the failure (i.e. to determine which segment causes the failure) and to prove to the other carriers how it is responsible for the failure.

For that purpose the MEP logs the last LBR messages from the MIPs and from the peer MPP.

#### MIP and Peer MEP functionality

Maintenance Intermediate Point (MIP) is a point between two MEPs, which is able to respond to OAM frames, but does not initiate them. MIP functionality is assigned to each inter domain connection point as described in Figure 5. The MIPs receive LBM and respond with LBR

In this context peer MEP functions like MIP respond to the LBMs that were generated by the MEP.

#### LoopBack Reply Messages

LoopBack Reply Messages (LBRs) are generated by MIPs or a peer MEP in response to the LBMs that are sent by a MEP:

- **Maintenance Entity Group Level (MEG Level)** at which the MIP or the peer MEP exist - MEG level defines the hierarchical level of the OAM.
- **Unicast address** of the MEP that has generated the original LBM.
- **Transaction ID/Sequence Number** - a 4-octet field that contains the transaction ID/sequence number for the LBM. The receiver (MIP or peer MEP) is expected to copy the Transaction ID/Sequence Number in the LBR PDU
- **Data** – Optional element whose length and contents are configurable at the MEP. The contents can be a test pattern and an optional checksum, etc. This Data field is used for the multi-carrier OAM to register:
  - **Status** of the ME (signed with the MIP private key), copied from the received LBM.
  - **Real Time stamp** (signed with the MEP private key), containing Year, Month, Hour, Minutes, Seconds
  - **MIP or peer MEP Identifier** (signed with the MEP private key), each MP having unique identifier in the multi-carrier network (the identifier method is out of the scope of this document, while it is dealt by the FP7 Project ETNA)
- **Priority** – Identifies the priority of frames with Unicast LBM information.
- **Drop Eligibility** – Identifies the eligibility of frames with LBM information to be discarded when congestion conditions are encountered. When MIP or the peer MEP receive LB they reply with LB Reply (LBR)

#### Logging of LoopBack messages

The LBM messages that are received from the MEP can be used by the MIP in case of failure to prove that the service is provided in good condition till a certain time.

For that purpose the MEP logs the last LBR messages from the MIPs and from the peer MEP.

### **2.3.1.2 OAM flows**

#### Ongoing Flow

This sub section describes the OAM flow and how the failure cause and the failure duration are proven.

The MEP generates LBM periodically, the configured period should allow detection and calculation of out of service duration with a reasonable resolution (e.g. seconds)

The Data field in the LBM PDU contains:

1. Status condition of the ME segment between the MEP and the MIP/peer MEP: "OK" if the previous LBR were received OK or "not OK" in case of missing previous LBRs
2. Real Time Stamp of the LBM transmission
3. MEP's unique identifier.

These parameters are digitally signed with the private RSA (Rivest Shamir Adelman) key of the carrier.

A MIP or the peer MEP sends LoopBack Reply (LBR) as a response to the LBM, it copies the Transaction ID/Sequence Number from the LBM and inserts the following information to the Data field in the LBR PDU:

1. The status of the ME (copied from the received LBM)
2. Real Time stamp
3. MIP or peer MEP unique Identifier

These parameters are digitally signed with the private RSA key of the carrier.

The MIP logs the Data Field content from the last received LBMs

The MEP logs the Data Field content from the last received LBRs from the MIPs and the peer MEP.

### Failure Event

The OAM mechanism can be used for settling commercial issues between the carriers. In Multi Carrier services, one carrier is the Retail Provider (i.e. sell the service to the end customer), while several other Wholesale carriers sell segments of the service to the Retail Provider. Usually the Retail Provider is the carrier that is connected to the end customer (e.g. the carrier that operates Domain1 in Figure 5). Several terminologies are used by different standard bodies to define this commercial relationship; here we use the terms Retail Provider and Wholesale Provider.

### Root cause

In case of out-of-service, the Retail Provider requests compensation for out-of-service from the domain that causes the failure (e.g. the carrier that operates Domain3 in Figure 5). He can prove that the root cause is domain 3 and not the other domains by showing the logged LBR messages (e.g. from MIP1, MIP2, MIP3 and MIP4 in Figure 5) with timestamp and signature, while the Retail Provider can prove that the segments provided by domain 1 and 2 were OK. The authentication of the MIP messages can be checked by the wholesale provider of domain 3 by using the public keys of the other domains.

### Fault duration

Fault duration is an important factor to determine the inter carrier compensation in case of out-of-service, because the retail provider (requesting compensation for the out-of-service duration from the carrier that is responsible for that failure, e.g. domain 3 in Figure 5) can calculate the out-of-service period from the logged LBRs. On the other hand the retail provider can not claim for durations longer than those calculated from the logged LBRs

## **2.4 OAM mechanisms: considerations and future work**

This chapter presented the main achievements carried out within the STRONGEST project in terms of OAM procedures. Studies are reported in the context of multi-domain packet transport networks to show the importance of loss rate and measurements to predict SLA degradations. Then, two innovative OAM mechanisms are presented to address both minor and major sources of SLA degradation (e.g., due to network congestion or physical impairments and link failures, respectively). The former mechanism exploits the existing BFD protocol to detect sporadic packet loss. Experimental validation of the proposed mechanism has been reported, showing the capability to perform effective SLA monitoring. The latter solution, derived from Y.1731, proposed protocol enhancements to efficiently provide failure localization and reliably measure the fault duration for commercial implications.

Within the STRONGEST project, future works will continue to address major and minor sources of packet loss inducing SLA degradation.

Within the scope of the STRONGEST project we will further investigate the different approaches for quantification of the path packet loss ratio. In particular we will implement an experimental protocol for the concatenation of link loss prediction values. The primary aim of the activity is a quantitative comparison of our newly proposed method with existing solutions. The central questions are to which extent the predicted values coincide with the afterwards registered real packet losses, how fast predictions converge and if applications can draw a benefit from this earlier quality advertisement if compared to plain packet loss registration.

Then, within the STRONGEST project, both BFD-based and Y.1731-based solutions as well as novel solutions exploiting NSLP signaling framework will be considered.

In particular, in the context of BFD (or alternative Hello protocols), strategies and correlations between received BFD statistics related to different LSPs will be implemented, aiming at identifying possible locations of malfunctioning. The analysis will take into account scalability issues in terms of both number of activated flows and BFD rate.

Specific studies will be carried out by considering the NSLP signaling framework for OAM purposes. This framework will need to support the OAM functionality defined within the STRONGEST project as well as control plane, E-NNI interface and architectural requirements, to ensure that it can support the proposed data plane scenarios. Initially a solution for the medium term data plane architecture will be attempted eventually leading to long term architecture considerations.

Finally, specific studies will be performed in the context of WSON scenarios, i.e. to enable effective monitoring of physical impairments and perform reliable failure localization.

### **3 Control Plane Architectures, Solutions and proposed Extensions**

In [D3.1], control plane and PCE-based architectures and solutions were summarized in the context of the considered single/multi-domain, single/multi-region and single/multi-carrier reference scenarios. In particular, [D3.1] reported the reference architectures, the significant standardization activities and the main research studies relevant for the STRONGEST project. In addition, [D3.1] specified the requirements and reference control plane architecture for the STRONGEST project.

In this chapter, preliminary innovative procedures and solutions are proposed and analyzed in the context of the reference STRONGEST control plane architecture specified in [D3.1]. The goal is to enable the implementation of the identified architectures and solutions by proposing specific control plane solutions and procedures (including novel protocol extensions and operational techniques).

The PCE-based architectures are widely investigated. Innovative procedures are defined and evaluated in the context of Hierarchical PCE, GMPLS translucent networks and WSON, multi-layer networks and point-to-multi-point scenarios. Moreover, solutions are proposed to provide the PCE with updated reachability and TE information, and to improve the efficiency of path computation techniques, e.g. by exploiting temporary reservations or by selecting ad-hoc routing algorithms. Additional studies are provided to combine path computation capabilities with admission control functionalities (e.g., G-RACF). Specific solutions have been also identified in the context of multi-carrier networks, where confidentiality needs to be preserved. They include the proposal of an alternative hierarchical path-vector routing protocol, the definition of a policy-enabled PCE architectures, and the implementation of BRPC procedures encompassing Path Key mechanisms.

#### **3.1 Control plane in a single-domain scenario**

##### **3.1.1 IGP scalability - motivation**

A major goal of traffic-engineering (TE) is to facilitate efficient and reliable network operation while simultaneously optimizing network resource utilization and traffic performance. Two approaches have been introduced to enable TE for label-switched paths (LSPs). One is a control plane architecture based upon a centralized *path computation element (PCE)* and the other is an architecture based upon distributed PCE (i.e., path computation is performed at ingress nodes). Both approaches can rely on a *routing protocol extension* to retrieve TE information.

In the PCE-based architecture, a PCE may be a network node or component which takes responsibility for collecting TE information and performing optimal LSP computation in response to a path computation client upon a connection request. The motivations for a PCE-based architecture are its high compatibility with the existing network model and the ability to use distributed centres of information or computational capability.



In particular, we might consider two approaches for PCE implementation and two approaches for collecting TE information:

- PCE: either distributed (D) i.e. in each network node, or centralised (C) i.e. unique for the whole area/domain
- TE information: collection via either IGP (I), or management-based mechanisms (M)

In (I), TE extensions are advertised within the routing protocols. Adding TE features into the routing phase can exploit IGP robustness with no additional network component required to manage the traffic-engineering information. However, such enhancement also complicates the IGP behaviour, as network states change frequently upon the dynamic traffic-engineered LSP set up and release, so the network is more easily driven from stable to unstable operating regimes.

Internal processing delays in IGP implementations impact the speed at which updates propagate in the network, the load on individual routers and the time needed for both intra-domain and inter-domain routing to re-converge, following an internal topology or configuration change. Reliable performance hinges on routing stability, a low convergence time indicating a stable configuration because the network can quickly come back to steady state when perturbed. Any sort of service level agreement (SLA) or quality assurance depends on routing stability. Once a router has done its shortest-path tree (SPT) calculation, it has to install all the routes in its RIB/FIB (Routing/Forwarding Information Base), introducing an additional delay.

Typical IGPs deployed in today's IP networks were originally designed for best-effort IP packets. Nowadays, following the widespread deployment of real time applications such as VoIP and the common use of Virtual Private Networks, much tighter SLAs are required, leading to sub-second convergence requirements.

The generic question is therefore whether sub-second link-state IGP convergence can be easily met on a large-scale network (perhaps 1000s of nodes) without compromise on stability.

### **3.1.2 IGP scalability - analysis**

To understand the convergence process, we need detailed measurements to determine the time required to perform the various operations of a link state protocol on currently deployed routers. A typical IGP convergence may be characterised as the sum of components for an individual router:

1. link failure detection time
2. time to originate the Link State Advertisement (LSA) describing the new topology after the link failure
3. flooding time from the node detecting the failure to the re-routing nodes that must perform a FIB update to bring the network in a consistent forwarding state
4. shortest-path tree computation time
5. time to update the RIB/FIB on the main CPU
6. time to distribute the FIB updates to the line cards in the case of a distributed router architecture

**Table 1: Convergence Time Components [FRA]**

COMPONENT	SYMBOL	TIME (ms)
detection	D	20
LSA origination	O	10
flooding	F	30
shortest-path tree (~1000 nodes)	SPT	30
RIB/FIB update	RIB	500
FIB distribution to linecards	DD	50

Some typical values are shown in Table 1, while the following general observations may also be made regarding specific components:

- **D:** use of Packet over SDH/SONET links in SP backbones and hence the ability to detect a link failure in a few tens of milliseconds is a major enabler of sub-second IGP convergence; most router interconnects benefit from very fast failure detection without any compromise on stability
- **O:** to achieve both rapid and stable convergence, dynamic (rather than static) timers have been introduced to adaptively control the LSA generation process; this ensures fast exchange of routing information when the network is stable and moderate routing protocol overhead when the network is unstable, thus allowing the network to settle down; origination time is then extremely fast without any compromise on stability
- **F:** flooding time from the failure node to the re-routing nodes depends on the sum at each hop of the propagation and IGP processing time; fast flooding has been introduced to bypass processing LSAs that describe a new link-state change event, reacting only to Refresh and TE LSAs; time to flood one LSA is then negligible compared to the sub-second convergence objective
- **SPT:** similarly, IGPs may be tuned such that when the network is stable, their timers will be short and they will react within a few milliseconds to any network topology changes; in times of network instability, however, the SPT timers will increase in order to throttle the rate of response to network events; this scheme ensures fast convergence when the network is stable and moderate routing protocol processing overhead when the network is unstable; shortest- paths for a network of 1000 nodes (large by current standards) may be computed in tens of milliseconds, without any compromise on stability
- **RIB:** RIB/FIB update duration is linearly dependent on the number of modified prefixes; introducing prefix prioritisation solves this problem, the important prefixes being updated first so worst-case RIB/FIB update duration scales based on a much smaller number (the number of important, rather than total) IGP prefixes
- **DD:** router implementation is optimised to allow for the parallel execution of the routing table update on the central CPU and distribution of modifications to the linecards, so this “distribution delay” DD is typically only tens of milliseconds

The computational complexity of a typical implementation of the shortest-path algorithm in an  $n$  node network is of order  $n \log(n)$ . In the range of interest, it seems clear that RIB/FIB update will dominate any small variation of shortest-path calculation time with  $n$ , as depicted in Figure 6. While in the wider network context there will be additional convergence time variation according to propagation times, failure scenarios etc., this overall conclusion is expected to remain true.

Furthermore, experimental results [HUA] demonstrate that a typical IGP with TE extensions requires additional time to converge. In particular, introducing per wavelength availability and continuity constraints may cause severe convergence time (perhaps, an order of magnitude increase) and link state advertisement scalability concerns.

The overall conclusion, from an operator perspective, is that this provides strong motivation for PCE as a distributed solution to the routing (processing) problem in a competitive, traffic-engineering environment. More specifically, using an IGP to collect TE information within a PCE scheme (“D+I” in the nomenclature introduced above) seems likely to incur scalability issues in larger networks. This also motivates the studies within the STRONGEST project (also included in this document) on alternative scalable techniques to provide the PCE with updated TE information.

A key additional focus for future work will be the desire for optimality in path computation and likely tradeoffs with information needs. For example, a centralized (stateful) PCE could implement effective shared protection whereas other distributed or stateless PCE-based solutions might provide poorer performance. However, a stateful PCE may suffer from scalability issues, so the relative merits of each scheme will need to be further explored.

### IGP Convergence Time

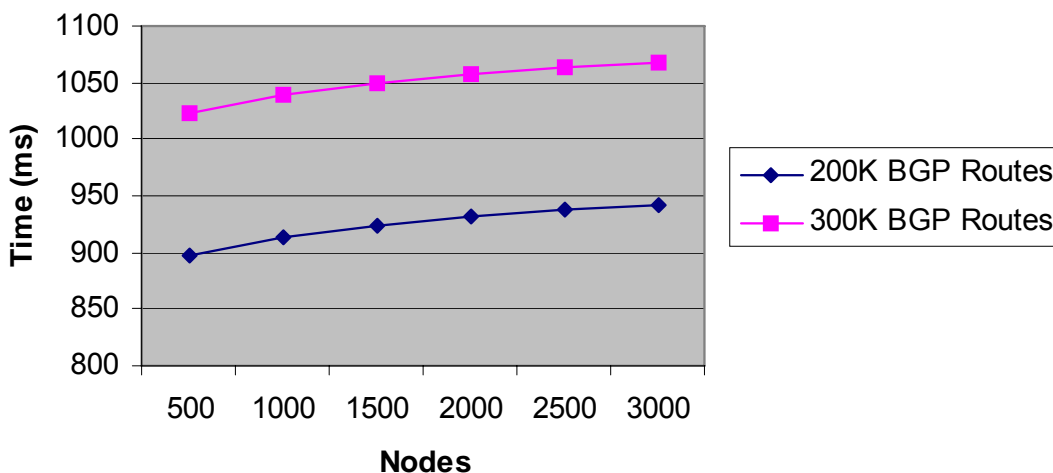


Figure 6: Convergence Time Variation

## 3.2 Control plane in a multi-domain scenario: hierarchical PCE

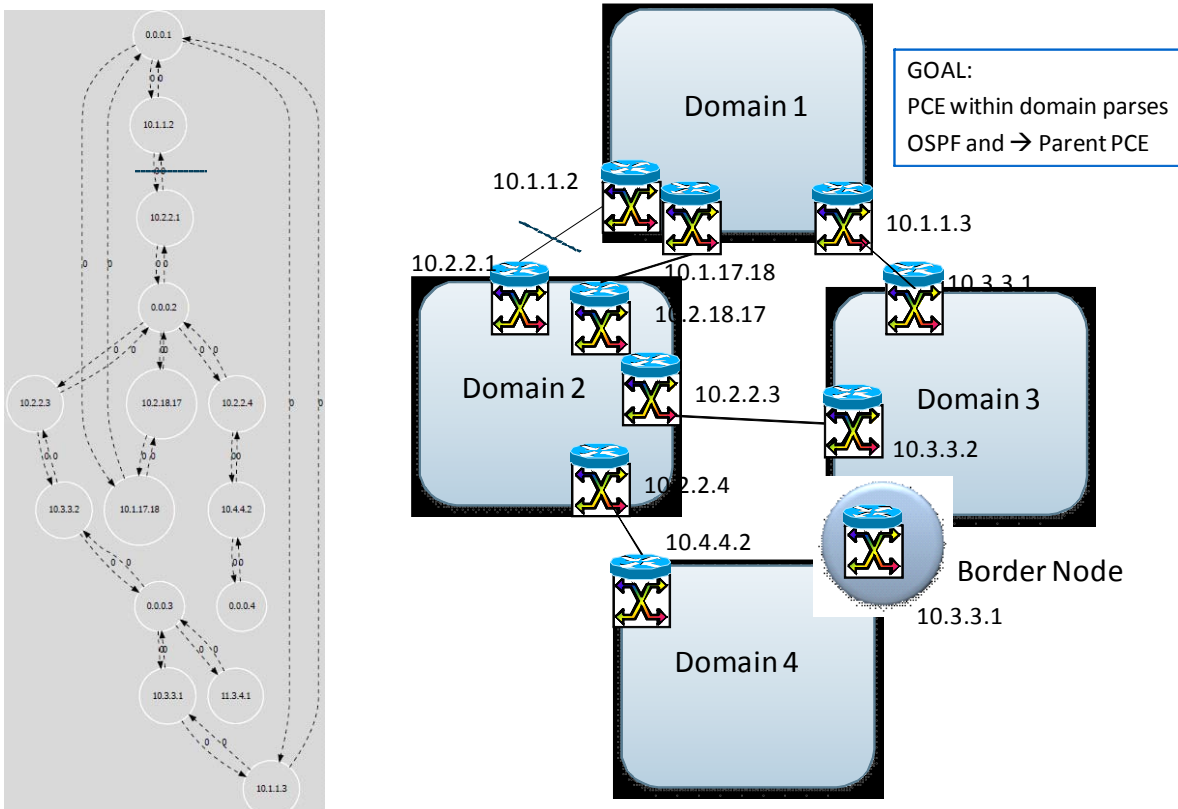
Although standard solutions have been proposed, targeting simplicity or path optimality, the sequence of domains is assumed to be known and the general problem of end to end path computation with arbitrary domain meshes remains open. Recently, work started in order to adapt existing hierarchical schemes to the IETF PCE context (i.e., hierarchical PCE or H-PCE [H-PCE]).

In this scope, the notion of hierarchy implies that, from the point of view of (control) functional entities, a group of entities such as PCEs may be interconnected hierarchically, with well defined interfaces and functionalities at each level defining a tree-like structure. From the point of view of the network topology, a multi-domain network can be seen as a network graph where domains are the new nodes and inter-domain links become the new graph links. In short, hierarchical PCE refers to a family of functional architectures where collaborating PCEs present a hierarchy relationship (such as parent-child), defining, for example, the number of levels, the functionalities of entities at a given computation hierarchy level (e.g., domain selection, intra-domain path computation) and the corresponding trust model (i.e., restricted to parent-child or between siblings). The concept of topology aggregation and summarization is coupled to the notion of hierarchy, which enables global computation while ensuring scalability by defining how nodes and links within a domain are synthesized in order to reduce the number of topology elements. Common approaches involve summarizing a domain as a set of virtual links (e.g. a mesh between all domain entry/exit node-pairs) or a virtual node, along with the inter-domain links providing domain connectivity. This is complex, since not only attributes in terms of bandwidth, traffic engineering metric or delay need to be considered, but also shared risk link groups, protection capabilities, etc. Moreover, there may be additional restrictions such as wavelength continuity in optical networks. More details about possible topology summarization algorithms are given in chapter 3.5.2.

In the Hierarchical PCE architecture, a single *parent* PCE is responsible for inter-domain path computation (e.g., to determine the sequence of domains to traverse), while in each domain a local *child* PCE performs intra-domain path computation. The goal of the solution is the definition of multi-domain (MD) and multi-vendor solutions involving hierarchical PCE, including the adaptation of PCE/GMPLS requirements for the MD context, and targeting MD path computation with technology specific requirements. This involves flexible topology attribute dissemination (PCE as ASON hierarchical routing controllers) while meeting operators requirements (topology confidentiality, well defined interfaces, etc.). The main outputs for this will be Control Plane functional / protocol network architectures and procedures and the corresponding PCEP extensions.

### 3.2.1 Proposed Architecture

The proposed approach is based on and extends an ongoing IETF draft [H-PCE]. It involves the construction of a domain topology within the parent PCE, composed of virtual nodes representing intra-domain connectivity plus border nodes and inter-domain links to allow end-to-end computation, as detailed in the following (see, for example, Figure 7).



**Figure 7: Parent TED in H-PCE, example with 4 domains**

The main aspects of the solution are:

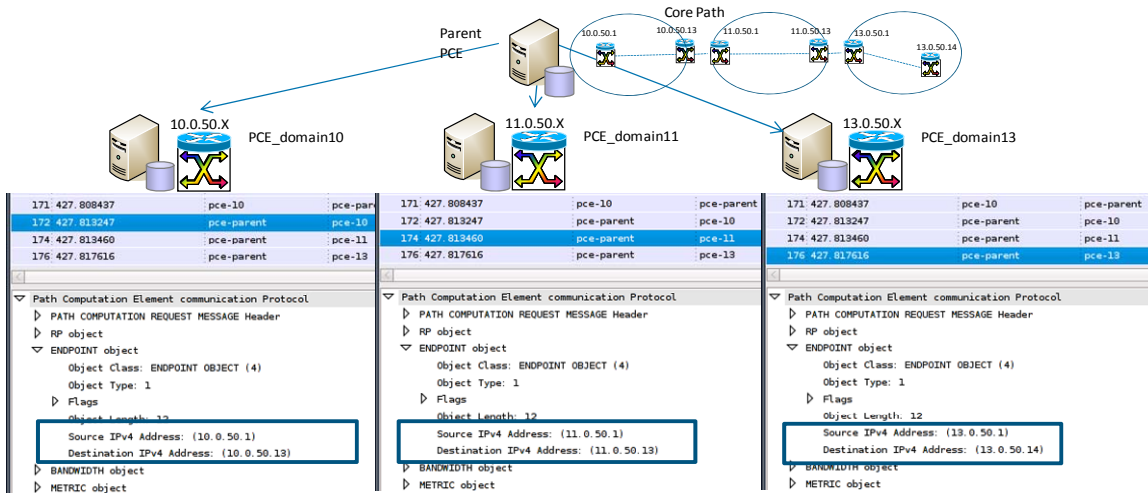
**Hierarchy management:** child PCEs are provisioned with the address of the parent PCE. A persistent PCEP session is initiated by the child PCE, which announces the set of domains for which it is responsible.

**Topology and Reachability management:** child PCEs construct their Traffic Engineering Database (TED) based on passive and stateful inspection of OSPF-TE TLVs. Filtered/selected topology elements (including border nodes and inter-domain links) are notified to the Parent PCE using the PCEP adjacency along with aggregated reachability, for example, using classless inter-domain routing (CIDR) for endpoint localization.

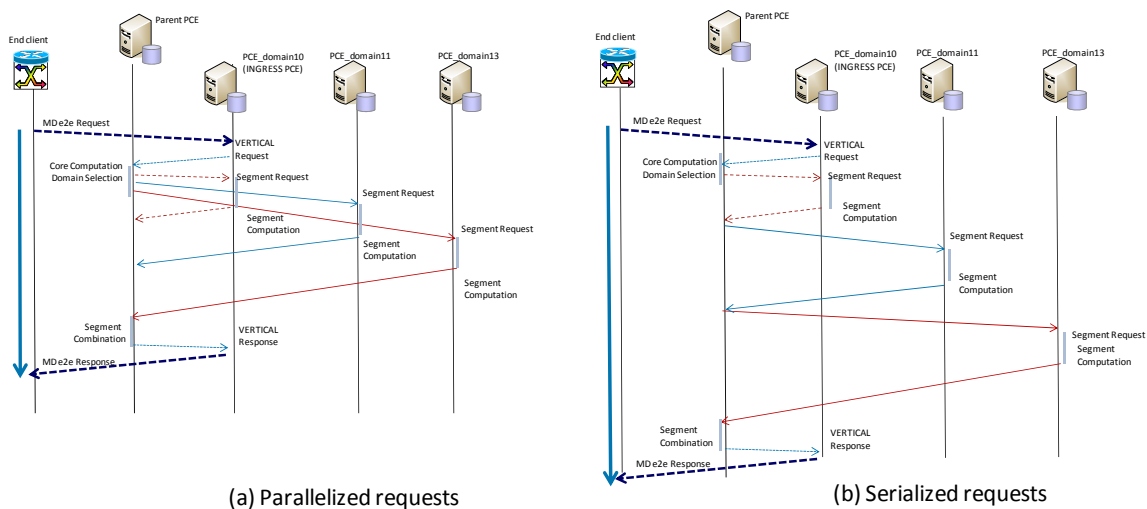
**Parent topology:** to the parent PCE, child domains are completely opaque and appear as star-hub networks, with virtual links from dynamically learnt border nodes

**Path Computation:** involves two steps. First, the end-to-end request is forwarded to the parent PCE, which carries out the domain sequence and inter-domain link selection based on the aforementioned parent topology using a Dijkstra based algorithm. This Core Path includes the domains and the selected border nodes/links. Second, the parent PCE requests segment expansions (Figure 8) to the children PCEs (segment computation delegation, using an OSNR-aware algorithm). Segments are then concatenated to form the end-to-end path. One of the main benefits of the approach is that the parent PCE is able to

parallelize segment expansion requests notably reducing latency (e.g., using threads) rather than one after the other (see Figure 9 for 3-domains).



**Figure 8: Segment computation delegation in H-PCE**



**Figure 9: Parallelized vs. serialized segment requests in H-PCE**

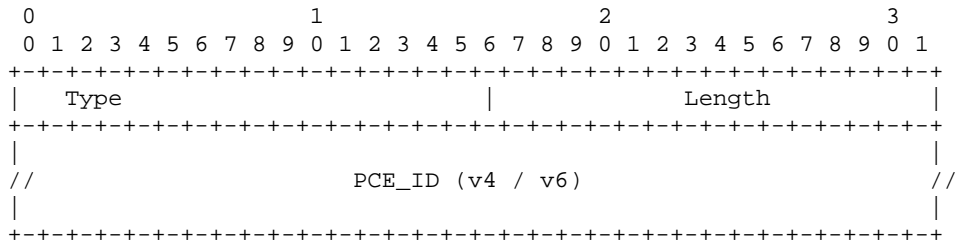
**Border Nodes and Inter-AS links:** border nodes (ABRs and ASBRs) are learnt from Summary and External OSPF-TE LSAs and forwarded to the parent. The Inter-AS-TE-v2 LSA is used as defined in [RFC5250], which contains the Remote AS Number sub-TLV and IPv4 Remote ASBR ID Sub-TLV.

### 3.2.2 Proposed Control Plane Extensions

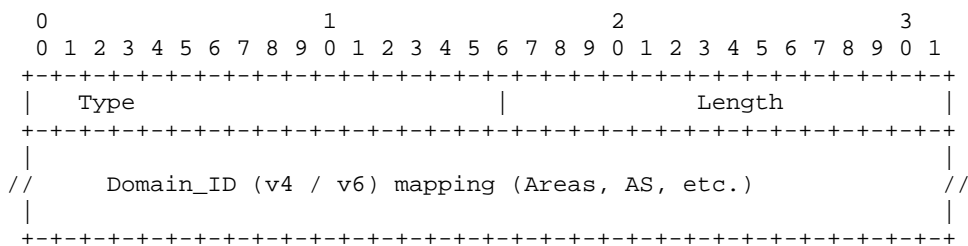
We propose new objects and TLVs that allow the mapping of PCEs to Domains, topology summarization, border nodes and inter-domain links announcement and reachability dissemination: a PCE\_ID and a new generic domain identifier, covering either Autonomous Systems or OSPF-TE areas. TLVs are included in children OPEN object and in Notification messages. The latter are used to announce reachability, re-using ERO sub-objects and allowing specifically IPv4 and IPv6 CIDR prefixes for the domain endpoints and to wrap topology elements in OSPF-TE TLVs (notably the Inter-AS links).

The proposed new TLVs are:

**PCE\_ID TLV** – To identify a PCE regardless of its interfaces AND to differentiate it with respect to a regular PCC. Both short (IPv4) and long (IPv6) formats



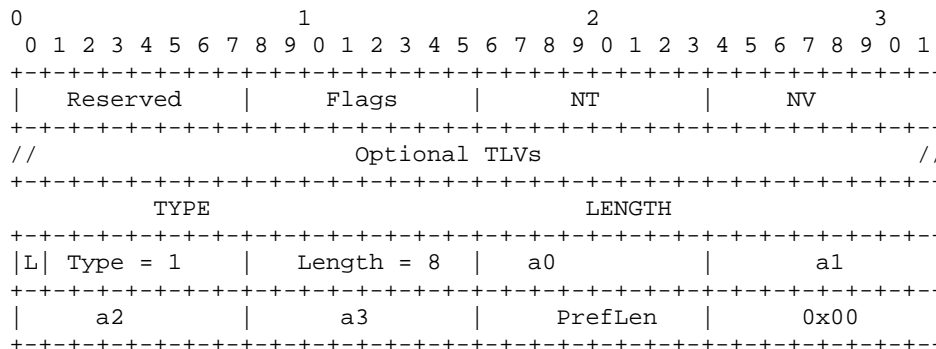
**DOMAIN\_ID TLV** -- Identify a domain, with both short (IPv4) and long (IPv6) formats. It should support all notions of “domain” (areas, AS, 4-octet AS, etc) with mapping into 4 or 16 bytes



These TLVs are used in both Open and Notification messages. Within the Open message, it is used to bind PCE and domains. In Notifications, we opt to re-use PCNf without RP to include “VENDOR\_NOTIFICATIONS” or “OSPF-TE TLV”, where to embed OSPF-TE TLVs. This approach supports both the border “node” model such as OSPF-TE ABR and the Border “links” model such as Inter-AS links. For the advertisement of OSPFv2 inter-AS TE links, a new Opaque LSA, the Inter-AS-TE-v2 LSA, is defined in this document. The Inter-AS-TE-v2 LSA has the same format as “Traffic Engineering LSA”, both Type 10 Opaque LSA [RFC5250] if the flooding scope is to be limited to within the single IGP area and Type 11 Opaque LSA [RFC5250] if the information is intended to

reach all routers (including area border routers, ASBRs, and PCEs) in the AS. This extension includes Remote AS Number Sub-TLV and IPv4 Remote ASBR ID Sub-TLV

Notably, new notification types were defined to include REACHABILITY\_NOTIFICATION with a REACHABILITY\_LIST\_TLV with route sub-objects:



The main use case is to add IPv4 prefix - IPv6 prefix – AS [RFC3209] - Unnumbered Interface ID [RFC3477] to announce ENDPOINTS within a domain. With this approach, a parent PCE may locate an ENDPOINT by best match prefix.

In this section we have presented the main aspects of a H-PCE based solution. The basics of the approach are being implemented in WP4 and are being proposed as a new work item draft at IETF 79 Beijing. In addition, the innovative solutions proposed in this study will appear in the proceedings of OFC 2011 Conference [Casellas11].

### 3.2.3 Case study and performance evaluation in multi-domain WSON

In this study, different provisioning schemes enabled by the Hierarchical PCE architecture are considered and evaluated through simulations in terms of the overall network resource utilization and network scalability.

The considered network scenario is a multi-domain WSON. A separate OSPF-TE instance runs in each domain advertising detailed wavelength availability information of intra-domain WDM links. Therefore, each child PCE resorts to a detailed TED to compute the edge-to-edge segments. The parent PCE resorts to a Hierarchical TED (i.e., H-TED). H-TED stores wavelength availability information of inter-domain WDM links. Therefore, the parent PCE operates on a topology made of edge nodes, inter-domain links, and opaque domains (a full mesh topology of virtual intra-domain links is considered). OIF E-NNI routing (extended to carry detailed wavelength information on inter-domain links) could be used to build and maintain the H-TED. In this study, the parent PCE does not exploit multiple PCEP requests to child PCEs to retrieve the metrics of possible edge-to-edge intra-domain LSP segments.



Two schemes enabled by the Hierarchical PCE architecture are analyzed:

- In the first scheme, called *Lightweight H-PCE Path Computation (LHPC) Domain*, the parent PCE only provides the sequence of domains.
- In the second scheme, called *LHPC Edge*, the parent PCE provides the sequence of the domain edge nodes to be traversed by the inter-domain LSP.

Moreover, for each scheme, two different levels of information detail are considered in the H-TED stored by the parent PCE:

- The H-TED with aggregated information (i.e., AGG H-TED) stores the available bandwidth (i.e., number of available wavelengths) along each inter-domain link.
- The H-TED with detailed information (i.e., DET H-TED) stores the status (i.e., available/reserved) of every wavelength along each inter-domain link.

The parent PCE selects the sequence of domains (or the sequence of edge nodes) by computing the shortest route in terms of number of traversed inter-domain links. In case of multiple equal cost routes: (i) with AGG H-TED the path with the largest number of available wavelengths on its most congested inter-domain link is selected; (ii) with DET H-TED the path that can accommodate the largest number of wavelength-continuous LSPs considering inter-domain links is selected. In both cases, all wavelengths of the virtual intra-domain links are always considered as available.

LSPs are established using the provisioning steps detailed in Figure 10.

- Step 1: the PCC (e.g., source node) sends a PCEP PCReq message to its child PCE, asking for an LSP from node A to node N.
- Step 2: if the LSP destination node is inside the local domain the child PCE computes the path and replies to the source node with a PCEP PCRep message including the strict Explicit Route Object (ERO) with the list of nodes to be traversed; otherwise the child PCE forwards the PCReq to the parent PCE.
- Step 3: the parent PCE performs LHPC schemes using the H-TED and returns a PCRep message to the child PCE including (i) the ERO with the sequence of edge nodes (i.e., A-D-E-H-I-N) if LHPC Edge scheme is used or (ii) the sequence of domains (i.e., 1-2-3) if LHPC Domain scheme is used.
- Step 4: the child PCE computes the strict path towards the next selected edge/domain using the information stored in the local TED and the ERO received by the parent PCE. As an example, if the parent PCE provides the aforementioned list of edge nodes, the child PCE can select the paths A-B-C-D-E or the path A-P-Q-D-E, both using the inter-domain link D-E. Conversely, if the parent PCE only provides the domain sequence, the child PCE can also select the path A-P-Q-T using the inter-domain link Q-T. In both cases the child PCE returns a PCRep message to the source node with an ERO including the selected path.

- Step 5: the source node starts the RSVP-TE signaling sending a Path message along the selected path up to the edge node of the next domain, node E in Figure 10
- Step 6: in turn, each ingress edge node freezes the Path message and asks the local child PCE for a strict path towards the next domain using a PCReq message.
- Step 7: when the child PCE receives the PCReq message it computes the strict path as in step 4 (i.e., E-F-G-H-I in domain 2 and I-L-M-N in domain step 3 and replies with a PCRep message.
- Step 8: when the edge node receives the PCRep message, the previously frozen Path message is updated with the received ERO and forwarded.
- Step 9: once the Path message reaches the destination node, wavelength assignment is performed based on the Label Set object included in the received Path message. The Label Set object is updated during the signaling phase by each intermediate node so that when it reaches the destination it lists the wavelengths that are available on the whole path. After wavelength assignment, a RSVP-TE Resv is sent backward up to the source node effectively reserving the selected wavelength.

*Simulation results.* The considered LHPC schemes are evaluated by means of simulations using a custom built event-driven C++ simulator. The considered multi-domain WSONs are depicted in Figure 11 with 72 nodes and 139 bidirectional WDM links with 32 wavelengths per direction. The whole network is divided in 9 domains. Each child PCE is co-located within a domain node, the parent PCE is co-located with the child PCE of domain D7. The traffic is uniformly distributed among node pairs and LSPs arrive following a Poisson process. The mean inter-arrival time is fixed to 100 s. The child PCEs TED contains detailed wavelength availability information. Wavelength assignment is first-fit.

The proposed LHPC schemes are compared against two classic solutions:

In the *BGP* solution the network is still divided in domains, but there is not a parent PCE. Therefore, child PCEs route inter-domain LSPs toward the next domain retrieved in the statically filled up BGP routing table.

In the *single-domain* solution the network is considered as a single-domain and a single OSPF-TE instance runs on the whole network.

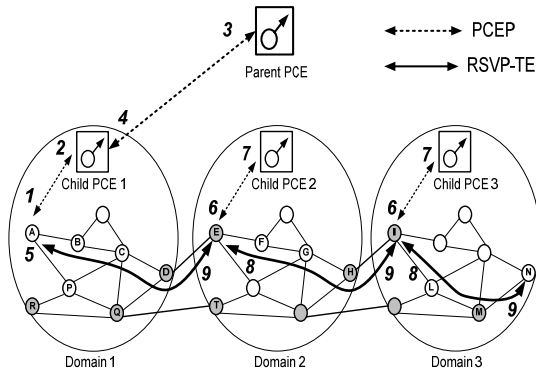


Figure 10: H-PCE-based LSP provisioning

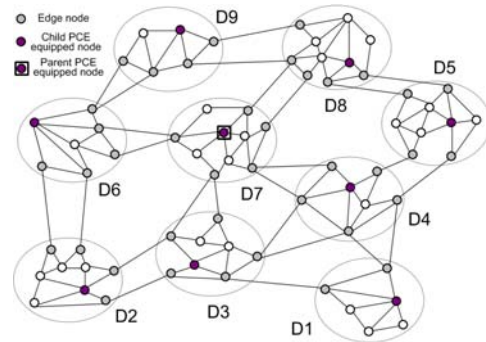


Figure 11: WSON test network topology

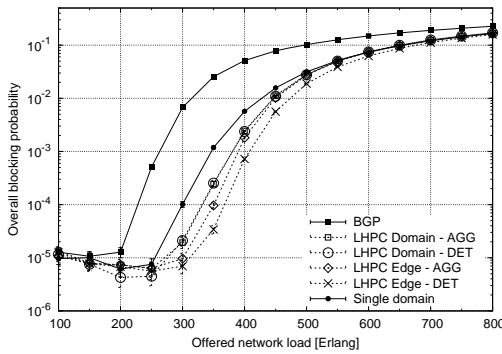


Figure 12: LSP blocking probability

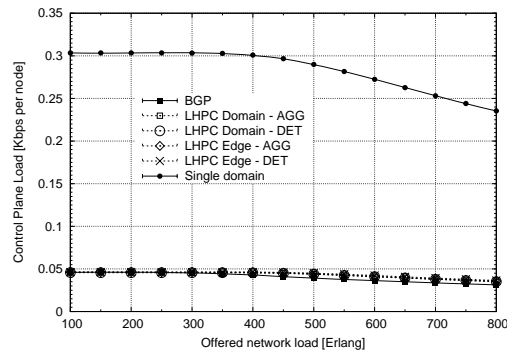


Figure 13: Control plane overhead.

Figure 12 shows the blocking probability versus the offered network load achieved by each considered scheme. At very low loads, when backward blocking dominates (i.e., contention mainly occurs during resource reservation), all schemes provide similar blocking. At higher loads, the LHPC schemes significantly reduce the blocking with respect to the BGP solution. Indeed in the BGP solution, inter-domain LSPs are routed on a deterministic sequence of domains and without considering wavelength availability information of remote domains or inter-domain links. Also the single-domain solution provides worse blocking with respect to the LHPC schemes. This result is quite surprising because, in single-domain, detailed wavelength availability information of all network links is always available. Two reasons mainly motivate this result: (i) inter-domain links, that topologically represent a bottleneck, are more intensively used, finally resulting in a higher blocking; (ii) just shortest routes are considered in this study, i.e. blocking is experienced if the shortest path (including all the equal cost shortest paths) does not provide any available wavelength from source to destination.

Among LHPC schemes, if the parent PCE only provides the sequence of domains, an H-TED with detailed information on inter-domain links does not provide benefit. Indeed, with both schemes (LHPC with AGG and DET H-TED) the edge node is selected by the child PCE utilizing information stored in the local TED.

Lower blocking is obtained when the parent PCE provides the list of edge nodes to use. In this case, the availability of detailed information on inter-domain links provides significant benefit. Indeed, in this case the parent PCE specifies the single inter-domain links to be used, thus maximizing the probability of finding a wavelength available on the whole inter-domain path.

The best performance is achieved by the LHPC scheme with DET H-TED.

Figure 13 shows the overall control plane load expressed in terms of average Kbps switched per node (RSVP-TE, OSPF-TE and PCEP messages are considered). The figure shows that all the considered LHPC schemes generate a control plane load similar to the one generated by the BGP solution, thus guaranteeing network scalability. Conversely, the single-domain solution generates a large amount of OSPF-TE messages due to the increased dimension of the routing area.

*Conclusions and future work.* This study for the first time evaluated the use of the H-PCE architecture in multi-domain WSONs. Simulation results showed that the use of a hierarchical PCE architecture enables the computation of effective sequences of domains to be exploited by per-domain procedures. The best results have been achieved by a proposed scheme which provides also the list of edge nodes computed on the basis of the detailed wavelength availability information on inter-domain links.

The innovative solutions proposed in this study will appear in the proceedings of OFC 2011 Conference [Giorgetti11].

Future works will consider other multi-domain PCE-based procedures (e.g., BRPC) and the opportunity for a parent PCE to perform multiple PCEP requests to child PCEs (to retrieve the metrics of edge-to-edge LSP segments). The latter option might improve the overall network utilization, but might also introduce remarkable scalability issues and delay into the inter-domain LSP provisioning.

### **3.2.4 A multi-domain hierarchical PCE-based system for Domains Topology creation**

The most relevant inter-domain path computation procedures currently standardized ([RFC5441] and [RFC5152]) either assume the pre-determined knowledge of the domains chain to be traversed or rely on an uncoordinated sequence of intra-domain path computations (using loose hop routes and the auto-discovery of the next Border Elements to be traversed in order to continue the path computation in the neighbor domain).

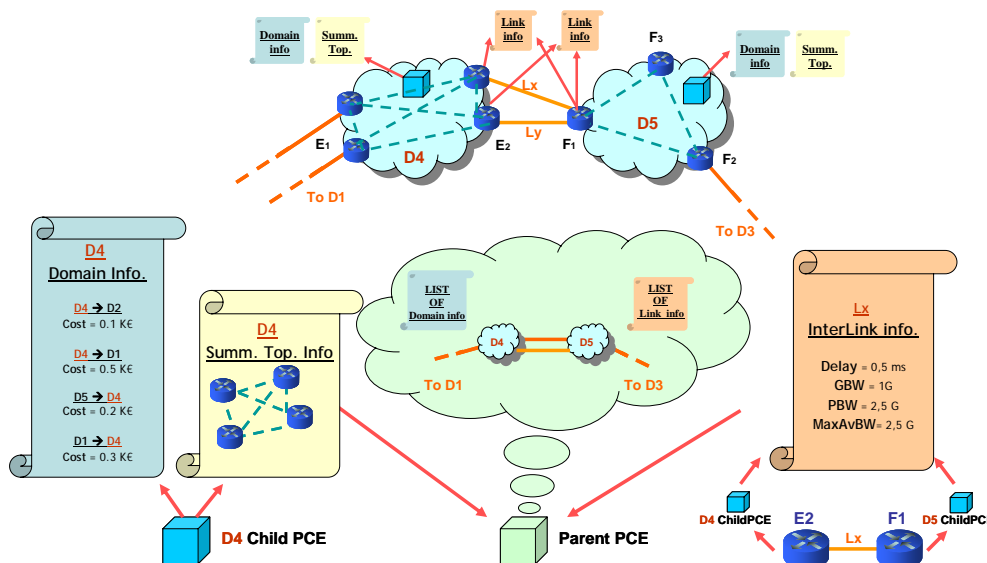
The system described in this section is being deployed within the STRONGEST WP3 with the aim of overcoming these limitations by means of a virtual inter-domain rough topology with summarized information, to be constructed and used by a hierarchical PCEs structure.

For sake of simplicity, and without loss of generality, the description of the proposed method reported in this document relates to a multi-domain hierarchical PCE-based architecture where there is supposed to be at least a "ChildPCE" (C-PCE) for each

domain, having an intra-domain scope and at least a “ParentPCE” (P-PCE), having an inter-domain scope. According to [RFC 4655] such entities can be centralized, distributed, implemented in a border router or as separated engine.

In order to create an overall multi-domain topology, all the C-PCEs send to the P-PCE (e.g. using PCEP protocol [RFC5440], with extensions where needed) the following information, as shown in Figure 14:

- a set of operator-driven and/or administrative information, named “**Domain\_Info**”, collected by the C-PCE(s) of a given domain Dx. Such high-level information describes only some economic and/or administrative domain’s characteristic (e.g. administrative and/or economic costs to go from the considered domain Dx to a domain Dy according to SLAs). Internal information, if any, would be described in the Summ\_Top\_Info;
- a set of common parameters, named “**InterLink\_Info**”, describing the inter-domain links that connect the considered domain with its neighbors (e.g. the service-oriented parameters defined in section 3.5.2, the link load, etc);
- a summarized topology of the managed domain, named “**Summ\_Top\_Info**” with summarized resources. Such summarized topology can be generated according to any method (e.g. the one described in section 3.5.2), but it should include as minimum information at least the Border Elements (BEs) of the domain (i.e. the nodes of the domain having interfaces with other domains and/or other regions) and at least a connection traversing the domain.

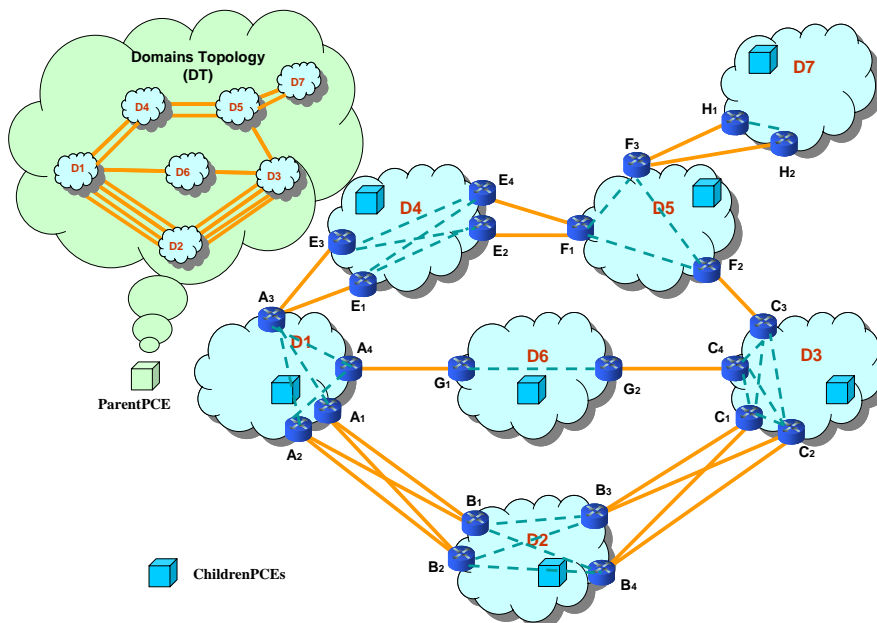


**Figure 14: Domain\_Info, InterLink\_Info and Summ\_Top\_Info**

Using Domain Info parameters, the summarized intra-domain topologies, and InterLink Info parameters, the P-PCE creates an inter-domain virtual topology, named Domains Topology (DT), as shown in the example of Figure 15.

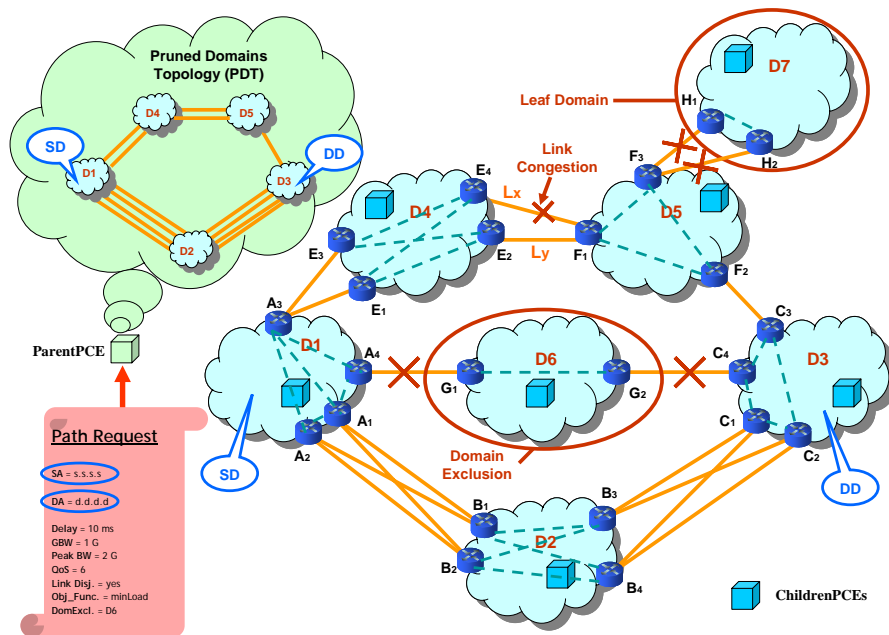
Such Domain Topology will be made only by domains (seen as black boxes or as a set of summarized BE-BE connection), their Border Elements and Inter-domain links (InterLinks) connecting them so it would be quite stable; therefore, it would be updated only if some updated parameters are received from some C-PCE of some domain. Moreover, not every event in the network (i.e. setup/release of an end to end connection, link failures, link or node congestions, etc.) will significantly impact the summarized view of the inter-domain topology. As a consequence, the updating process is driven by a thresholds system that reduces the amount of messages flooding in the network by filtering minor changes in the resources status and by advertising only the relevant ones.

All the steps above described allow the P-PCE to be aware of a summarized view of the entire multi-domain topology independently of a path request (i.e. they are performed off-line). As a matter of fact, an incoming Path Request would be served referring at the last updated version of DT, that is representative of the network status at the moment of the Path Request delivery and that is immediately available to the P-PCE.



**Figure 15: Domains Topology creation**

When a Path Request comes, the P-PCE exploits some fields of the PathReq message to prune some domains and/or some InterLinks from the last updated version of DT, obtaining a Pruned Domains Topology (PDT).



**Figure 16: Domains Topology pruning**

In Figure 16 some possible causes of Domain pruning (i.e. an Explicit Domains Exclusion in the PathReq and a leaf-connected domain that is neither the source one nor the destination one) and some possible causes of InterLink pruning (i.e. the congestion of some InterLinks and/or summarized links that makes available less resources than the ones requested) are shown. The resulting Pruned Domains Topology has only 5 domains and 11 InterLinks (instead of the 7 domains and 17 InterLinks of the DT).

Once created the PDT, the P-PCE can perform the following set of possible path computations, depending on the way the DT is constructed:

- **Computation of a summarized E2E path directly** – the P-PCE computes a suitable summarized E2E path directly;
- **Computation of a summarized E2E path in collaboration with C-PCEs** – The P-PCE computes a summarized E2E path in collaboration with C-PCEs, by means of a request/response mechanism for the computation of intra-domain summarized topologies;
- **Computation of an enhanced domains chain** – The P-PCE computes only a suitable sequence of domains (enhanced with the information of the BEs and the Inter-domain link to be used) for the E2E path, according to the received request.

The path computation method and the mechanism to perform it would be chosen according to the following considerations.

If the C-PCEs advertised summarized topologies for the domain they manage, then the P-PCE has enough information to compute a suitable summarized E2E path directly. As a matter of fact, P-PCE can choose the sequence of domains, the BEs, the inter-domain links and the intra-domain connections between the selected BE-BE couples to be used. C-PCEs have only to “translate” the summarized BE-BE connections in the correspondent intra-domain paths (already pre-computed or dynamically computed). In other words, the P-PCE performs E2E path computation, but leaving the details to the C-PCEs.

If the C-PCEs advertised only the minimum information for the domain they manage (i.e. only Domain\_Info, InterLink\_Info and Summ\_Top\_Info parameters with only BEs), then the P-PCE has not enough information to perform E2E path computation directly. Therefore, a first option is to compute the E2E path in collaboration with C-PCEs. In this case, the P-PCE would ask to the C-PCEs of the possible involved domains (i.e. only the domains of the PDT) to provide the summarized topologies, specifically computed according to the path request. After all C-PCEs have provided such information, then the P-PCE would be able to perform a suitable summarized E2E path computation. A second option is to compute just a suitable sequence of domains (enhanced with the information of the BEs and the Inter-domain link to be used) for the E2E path, according to the received request. In this case, the P-PCE would ask to the C-PCEs of the involved domains to complete the E2E path computation according to one of the standardized/under standardization methods ([RFC5623], [RFC5520], [King\_H-PCE], etc.).

The above described method is composed by two phases: the first one is performed off-line, dealing with the construction of an asynchronous domain topology, described with quite stable parameters (i.e. administrative, economic and summarized ones) and updated with a threshold mechanism, allow the P-PCE. The second one is performed after the receipt of an inter-domain path request and deals with the pruning of the domain topology. The resulting pruned topology can therefore be used to compute an E2E inter-domain summarized path (directly or in collaboration with C-PCEs of the domains not affected by the pruning step) or to compute a suitable and enhanced sequence of domains.

### **3.3 Control plane in a multi-carrier scenario**

#### **3.3.1 Multi-domain PCE-based architectures**

In the GMPLS architecture, a domain can be defined as a collection of network elements within a common sphere of address management or path computational responsibility such as an IGP area or an Autonomous Systems. The applicability of the Path Computation Element (PCE) [RFC4655] for the computation of such paths is discussed in [RFC5671], and the requirements placed on the PCE communications Protocol (PCEP) for this are given in [RFC5862].

Label switched routers (e.g., optical connection controllers or OCCs in WSON) have full topology visibility within their domain boundaries and limited visibility of the other domains, usually as aggregated information (e.g., reachability). Consequently, in traditional source routing approaches, a source OCC is not able to compute, autonomously, an end-to-end inter-domain path with the same control and degree of TE as for an intra-area path.



In this context, two methods are applicable for inter-domain path computation, the per-domain path computation method and the path computation element (PCE)-based path computation method.

*Per-domain path computation method:* the source OCC determines the next domain and the ingress within that domain. Then, it computes the corresponding path segment to the domain boundary, obtaining a strict explicit route object (ERO) within its own domain and appending to it (a list of) loose hops for the neighbor domain toward the destination. Next, the path computation moves to the ingress OCC of the next domain and so forth until the destination domain. During the signaling phase, the OCC at each boundary domain expands the ERO. As a result, this simple method generally precludes the computation of a shortest inter-domain path in an end-to-end lightpath perspective.

*PCE-based path computation method:* this method assumes that a domain chain (succession of transit TE domains from source to destination) is known in advance. The method relies on dedicated PCEs, which collaboratively compute an inter-domain optimum path along the given domain chain. Each PCE is responsible for the path computation within its domain. Such an architecture is motivated by the complexity of path computation in large, multi-domain, multi-region, or multi-layer networks, and that of advanced (e.g., protection-enabled) algorithms and heuristics, which may eventually require dedicated computational resources and cooperation between network domains. The new architecture raises new challenges regarding the feasibility and applicability of the PCE in general, and in GMPLS controlled WSONs in particular. This includes, for example, the Wavelength Continuity Constraint (WCC), which may significantly degrade performance if not addressed correctly.

Most research efforts on the PCE-based multi-domain path computation seem targeted to improve or extend the backwards recursive path computation (BRPC) [RFC5441] procedure which is currently the one that meets best the operator and supplier requirements in terms of complexity and network information hiding.

### **3.3.2 BRPC-based mechanism for MPLS-TP / WSON networks**

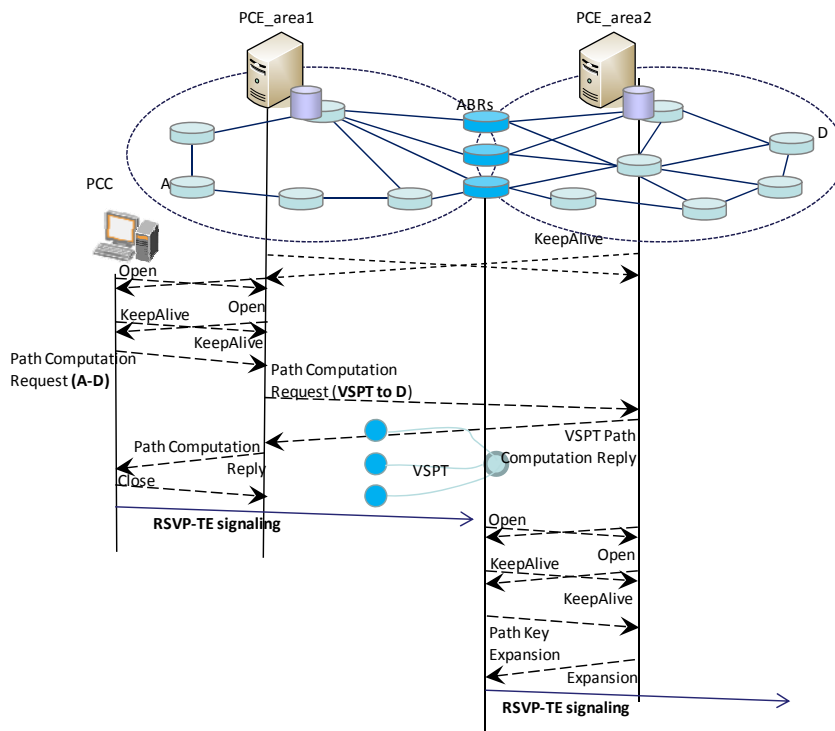
All-optical and translucent wavelength switched optical networks (WSON) are a key element in current and future core transport architectures. Multi-domain path computation is one of the main drivers behind the adoption of Path Computation Element (PCE) based schemes. Nonetheless, network operators have stringent requirements on topology confidentiality. In this sense, two constraints need to be jointly addressed: path optimality and network topology confidentiality. To this end, the IETF has published several PCE Protocol (PCEP) extensions: for the former, the use of the Backwards Recursive Path Computation (BRPC) [RFC5441], which is proposed as a more efficient approach with respect to the (sub-optimal) per-domain [RFC5152] one. For the latter, enabling Path Keys can indeed preserve confidentiality. The Path Key [RFC5520] mechanism replaces route segments (ERO sub-objects) in PCEP Reply messages with tokens that are transported transparently outside the relevant domains and that can be expanded (converted back to regular EROs) by border nodes within the concerned domain during signaling. The combined architecture includes the Path Key mechanism with a few important notes:

- Whether a PCE replies with a detailed ERO rather than a Path Key is determined by policies. Only nodes within the same area may request a Path Key expansion. Keys (16 bits) are managed in a hash table within the PCE. Allocated Keys are reserved and available for retrieval for a (configurable) period of 30 minutes.
- The RSVP-TE Connection Controller detects a Path Key sub-object within the ERO and requests an expansion at the PCE whose ID is within the object. In a multi-domain lightpath provisioning spanning N domains there are N-1 consecutive Path Key expansions at each ABR. Except the last domain, expanded EROs also contain downstream Path Key objects.

Deploying Path Key based mechanisms may have a noticeable impact on set up delay, although arguably within acceptable limits. Book-keeping of allocated keys increases state within the PCE, since for each request, up to  $N_{\text{ABR}}$  keys need to be allocated and stored along with the ERO for at least 30 minutes.

BRPC enables the computation of an optimal (with regard to a given metric) end-to-end path in the presence of multiple exit/entry nodes, recursively pruning a tree of paths from the domain entry points towards the destination (Virtual Shortest Path Tree or VSPT). BRPC has been enabled as follows (see Figure 17):

- The PCEP session between PCEs is persistent; the adjacency is pre-configured in order to avoid the handshake between PCEs at each request.
- Downstream BRPC path EROs are encoded as PATH\_KEY objects with 16-bit Path Keys. A PREFIX ERO sub-object is pre-pended to the Path Key one in order to identify the ABR the given path applies to.
- Metrics (TE, Hop Count, OSNR value) are forwarded within the BRPC VSPT, allowing flexible combination of metrics for a given notion of “optimality”.



**Figure 17: Combining BRPC and Path Keys in PCE based multi domain Path Computation**

Since each PCE performs RWA, without other considerations / extensions to ensure cross-domain wavelength continuity, a 3R regenerator (or a wavelength converter) needs to be systematically allocated at each ABR during lightpath provisioning, since Path Keys do not convey information on the available wavelengths on downstream domains. To overcome this, it is possible to delegate WA to signaling, or to extend PCEP and the BRPC with wavelength (label) availability information, by means of Labelset object within PCEP path attributes. Additionally, a naive BRPC implementation may overload the ABR with the shortest path, exhausting its regenerator pool, especially if the number of regenerators is “low” with regard to the potential number of LSPs crossing the ABR. For this, as an innovative solution, we extend BRPC, considering the number of available regenerators, pruning from the VSPT any ABR with no available regenerators

*Conclusions:* We have detailed the combined architecture of multi-domain (OSPF-TE areas) lightpath provisioning with PCE-based path computation for OSNR-aware GMPLS-enabled translucent WSON combining BRPC and Path Keys for optimality and topology confidentiality preservation. We have proved its feasibility, allowing optimal paths as regards the per-domain method, yet meeting operator requirements concerning topology non-disclosure.

The innovative solutions proposed in this study have been published in the proceedings of ECOC 2010 Conference [Casellas10].

### **3.3.3 Hierarchical path vector protocol for multi-carrier networks**

The PCE architecture has to be injected with multi-domain information for domain sequence pre-computation. Various methods have been considered to provide this information to PCEs. On this basis the Optical Internetworking Forum (OIF) has defined the Network-to-Network Interface (E-NNI), addressing the multi-domain single-carrier scenario [ENNI]. OIF E-NNI adopts link-state routing solution, which may not represent the best solution in terms of confidentiality in a multi-carrier scenario. As a matter of fact, with E-NNI all domains obtain the detailed view of the whole inter-domain network resources.

To this aim, to preserve both confidentiality and scalability, a hierarchical instance of a path-state protocol dedicated to Traffic Engineering (TE) information can be adopted. In particular, a hierarchical instance of an inter-domain routing protocol exploiting part of the code of the Border Gateway Protocol is proposed to operate within a restricted set of authorized domains at the PCE level. This solution has been named Hierarchical BGP (HBGP). It is worth noting that this solution is not intended to replace BGP, which is still operating between the routers of the network and is not modified. Rather, HBGP is intended as a candidate alternative to the OIF E-NNI solution in the context of multi-carrier networks.

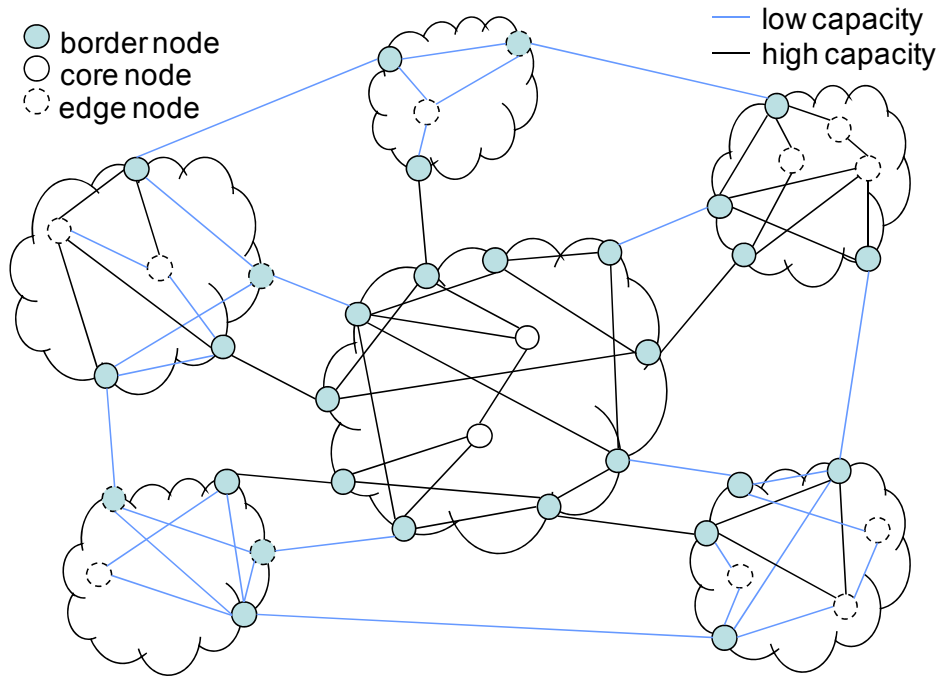
We initially refer to a network scenario with a PCE-architecture that is operating in a limited set of domains working in a peering relationship. In terms of confidentiality, this solution restricts the multi-domain information that can be received by a node, to the network view provided by the adjacent domains and does not allow the detailed view of all network resources. Moreover, a path-vector view of the network resources is compatible with the view provided by the PCE architecture, which is based on path information exchanged between adjacent domains. The sequence of domains to be crossed by an inter-domain connection is pre-computed by the PCE. Then a specific procedure computes the final detailed end-to-end path. In this study we have selected the Per-Domain procedure (PD) [RFC5440] to perform this task.

The aforementioned integrated HBGP-PCE approach enables the implementation of multiple schemes that exploit the various configurations in different ways. Thus, two different solutions are considered: HBGP-BW-PD-MA and HBGP-R-PD-MA [Buzzi10]. Both solutions exploit the PD procedure, but while in the first one the sequence of domains to be crossed is computed on the basis of announcements regarding inter-domain bandwidth (BW) availability information, the second one randomly (R) selects the sequence of domains. Both announce multiple routes per each domain or network prefix and perform multiple attempts (MA) along the sequence of domains.

We intend to focus on solutions that aggregate multiple inter-domain links between the same pairs of adjacent domains and announce such multiple resources as a single link. We classify them as solutions with Aggregated multiple Inter-domain Links (AIL). Thus, the topological information is not completely disclosed. Our goal is to understand what is the penalty in terms of routing efficiency which has to be paid for such abstraction, compared to a fully-disclosed link-state approach (single carrier multi-domain scenario, E-NNI standard).

To this account, we performed the following simulation experiment by which an ALL approach is compared to routing considering the inter-domain links between the same pairs of adjacent domains as separated. We have identified this second approach by suffix SIL (Separate Inter-domain Links).

The reference network is depicted in Figure 18.



**Figure 18: Reference network**

The nodes generating traffic are the edge nodes and the nodes of the central domain. There are links with high capacity (64 channels) and links with low capacity (16 channels).

The comparison between ALL and SIL is carried out in terms of inter-domain blocking probability. Results are shown in Table 2:

**Table 2: Blocking probability vs offered load; first scenario:**

<b>Offered load per node</b>	<b>HBGP-BW-PD-MA-AIL</b>	<b>HBGP-BW-PD-MA-SIL</b>	<b>HBGP-R-PD-MA-AIL</b>	<b>HBGP-R-PD-MA-SIL</b>
<b>10</b>	<b>0.004</b>	<b>0.007</b>	<b>0.003</b>	<b>0.003</b>
<b>20</b>	<b>0.059</b>	<b>0.064</b>	<b>0.068</b>	<b>0.067</b>
<b>30</b>	<b>0.163</b>	<b>0.158</b>	<b>0.173</b>	<b>0.167</b>

<b>40</b>	<b>0.268</b>	<b>0.26</b>	<b>0.269</b>	<b>0.262</b>
<b>50</b>	<b>0.364</b>	<b>0.359</b>	<b>0.362</b>	<b>0.357</b>

Simulation results have a confidence interval of 1%. We also deployed a different scenario in which all the inter-domain links are set with high capacity (64 channels) and the nodes of the central domain produce a traffic 4 times bigger than in the first scenario. These results are shown in Table 3.

**Table 3: Blocking probability vs offered load; second scenario**

<b>Offered load per node</b>	<b>HBGP-BW-PD-MA-AIL</b>	<b>HBGP-BW-PD-MA-SIL</b>	<b>HBGP-R-PD-MA-AIL</b>	<b>HBGP-R-PD-MA-SIL</b>
<b>10</b>	<b>0.075</b>		<b>0.072</b>	<b>0.059</b>
<b>20</b>	<b>0.288</b>	<b>0.271</b>	<b>0.281</b>	<b>0.259</b>
<b>30</b>	<b>0.444</b>	<b>0.426</b>	<b>0.431</b>	<b>0.409</b>
<b>40</b>	<b>0.542</b>	<b>0.514</b>	<b>0.524</b>	<b>0.502</b>
<b>50</b>	<b>0.6048</b>	<b>0.569</b>	<b>0.595</b>	<b>0.563</b>

By these results we notice that the SIL based solutions provide better results in terms of blocking probability against the AIL ones, but this improvement is rather limited (at maximum around 4%). This means that hiding the information about the full state of the inter-domain links is not that detrimental for the end-to-end path computation.

However, it should be noted that these preliminary results have been obtained under uniform traffic conditions, with equal bandwidth per circuit for each request. We can conjecture that in less symmetric conditions, characterized by variable bandwidth connection requests, the gap between the SIL solutions and the AIL ones can increase.

### **3.3.4 Confidentiality in multi-carrier PCE-based networks**

The Path Computation Element (PCE) Architecture has been proposed to provide effective resource utilization in multi-domain network scenarios, while potentially guaranteeing an adequate level of information confidentiality [RFC5441]. Typical confidential information includes details on intra-domain network resources, congested network portions, node architectural limitations and constraints, the ability/inability to support advanced network services, recovery schemes, and QoS-guaranteed applications. In [RFC5440], the need for effective access policies to avoid malicious utilizations of the PCE Protocol (PCEP) procedures is identified. In particular, the objective is to prevent that a Path Computation Client (PCC) belonging to a different domain might perform bogus or false computation requests, thus discovering important confidential information inside other

domains. Although the subject is of great relevance to the successful implementation of traffic engineering across multiple carriers, no discussions or implementation solutions have been proposed so far in the literature.

### **3.3.4.1 Confidentiality issues in PCEP**

In inter-domain path computations procedures, PCE and PCC will not exchange strict explicit list of traversed intra-domain hops and each confidential path segment will be expressed in an encrypted form, i.e. through an identifier called Path-key.

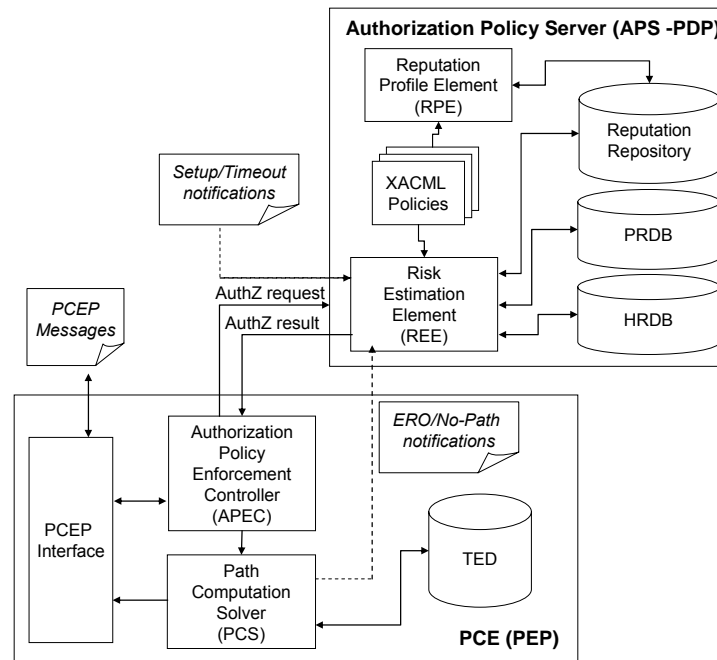
However, several PCEP parameters might be maliciously utilized to break confidentiality. Path computations requesting small values of bandwidth do not usually induce a confidentiality issue while very high values should require some careful treatment, or even an immediate rejection, since they may allow the discovery of bottlenecks in case of negative reply (i.e., No-path). This is aggravated by the presence of additional constraints such as diversity and bi-directionality, which might implicitly reveal topological limitations or node architectural constraints. Metric values returned to a PCC might be used to infer intra-domain topological information. In addition, the backward nature of the inter-domain PCEP procedures [RFC5441, RFC5376] allows the requesting domain to retrieve information without providing any detail about its own resources. This is particularly critical in the case of the Backward Recursive PCE-based Computation (BRPC [RFC5441]), where a tree between border nodes and the destination is returned together with the computed metric values. Furthermore, correlations among different path computations might introduce additional risks. For example, multiple independent requests, targeting destinations located in the same geographical area, providing positive replies under certain constraints and negative replies under different constraints (e.g., link and SRLG disjointness) or in different time periods, could practically reveal lack/availability of intra-domain resources or performance (metrics). Requests for which a PCE provided a positive reply, might not be followed by the related set up procedure (i.e., signaling messages). On the one hand, this could refer to a truthful need to identify the optimal path along alternative routes controlled by different operators. In this case, just one route will be eventually set up, while the others will be discarded upon the expiration of a predefined timeout (e.g., ten minutes). On the other hand, expired path computations might be considered as an attempt to discover confidential information. Also the time period between a positive reply and the related connection set up or timeout should be carefully treated, since a burst of requests could take place without being eventually set up.

### **3.3.4.2 Proposed Policy-based Architecture**

The proposed architecture, elaborated within the STRONGEST project, enables policy-based authorization schemes, as shown in Figure 19. A PCE is equipped with a PCEP interface to handle PCEP communication and with a Path Computation Solver (PCS) to perform path computations. The architecture encompasses two additional new elements: a local Authorization Policy Enforcement Controller (APEC) and a centralized Authorization Policy Server (APS).

APEC is introduced to filter the incoming inter-domain computation requests and perform basic authorization evaluations through simple permit/deny conditions specified in the form of access lists. APS is introduced to run, when needed, more sophisticated

authorization policies based on a joint analysis of (i) the reputation associated to the requesting domain (maintained in a Reputation Repository); and (ii) the risk to confidentiality related to incoming and previous requests. To accomplish the latter task, APS resorts to two additional databases per adjacent domain: a Pending Request Database (PRDB) and a History Request Database (HRDB). PRDB caches all pending requests, i.e., those just received or those for which the path computation has been provided but neither the connection set up nor the expiration timeout has occurred. HRDB stores all the details of the completed path computations handled by APS for each requesting domain, limited within a reasonable period of time (e.g., six months), and specifying also whether: (i) the path computation failed, (ii) the request has been successfully computed and set up or (iii) a timeout has occurred upon the successful computation.



**Figure 19: Policy-based authorization: architecture**

Communication between APEC and APS is achieved through the exchange of specific authorization messages. This can be implemented using Security Assertion Markup Language (SAML) carried by Simple Object Access Protocol (SOAP) over Hypertext Transfer Protocol (HTTP) and following the approach based on Policy Decision Point (PDP) and Policy Enforcement Point (PEP) [Toktar04, Dem09]. To maintain PRDB and HRDB databases, APS is notified with information on the final status (set up or timeout) of the computed path, e.g. through Simple Network Management Protocol (SNMP) notifications from the NMS.

The decoupling of the authorization evaluation performed by APEC (involved in all inter-domain path computations) and possibly by APS (when complex evaluation is required) is introduced to better address the scalability requirements of the overall authorization scheme.



### 3.3.4.3 Authorization policy implementation

Upon the arrival of an inter-domain path computation request, the PCEP interface forwards it to the local APEC which begins the overall authorization procedure:

Step 1: APEC first evaluates the request parameters and decides whether to immediately reject the authorization request (e.g., excessive bandwidth) or proceed with the evaluation. In the former case, go to step 7.

Step 2: APEC evaluates the request on the basis of simple access lists. Requests not determining risks to confidentiality (e.g., negligible bandwidth requirements) are directly forwarded to PCS (Step 6), while the remaining are passed to APS for more careful authorization evaluation.

Step 3: APS computes the confidentiality risk  $r_k$  of the incoming request as a combination of three different contributions. The first contribution (weight  $w_k$ ) refers to the parameters included within the request itself. Coefficients are introduced to assess the risk of each parameter. They are defined according to the domain network conditions (e.g., topology, availability of alternative routes, node constraints). For example, the highest coefficient might be associated with path diversity. The weight  $w_k$  is then obtained as a combination of the coefficients associated to the requested parameters. The second and third contributions (weights  $w_{pk}$  and  $w_{hk}$ ) are introduced to account for the correlation between the incoming request and the requests previously elaborated by APS and stored in PRDB and HRDB. Correlation coefficients are then introduced to assess such risk. Numerical examples are reported in Figure 20, where the highest coefficients are assigned to requests targeting same destination areas and to expired path computations (additional details can be found in [Paol-Ecoc10]).

Step 4: APS evaluates the computed risk  $r_k$  together with the reputation  $R$  of the requesting domain.  $R$  is computed on the basis of the information stored in HRDB.  $R$  is updated by a value proportional to  $w_k$  (see in Figure 20) when the  $k$ -th request is inserted in the database.  $R$  worsens upon: (i) a negative path computation (i.e., No-path) for which the domain was the destination domain; (ii) a positive path computation which expired after the timeout; (iii) a rejected path computation for policy violation by APS. Conversely,  $R$  improves upon: (i) a positive computation followed by the related set up or (ii) the erase of an aged request from HRDB. The computed value of  $R$  is then associated to one among four different adjacent domain states: 'good', 'poor', 'critical', 'bad'. According to the values of  $R$  and  $r_k$ , different authorization messages are sent to APEC (Step 5), which performs one of the following:

5a). Authorize the incoming request to be forwarded to the PCS. Go to step 6.

5b). Temporarily deny the incoming request because of temporary confidentiality issues. A temporary deny is introduced to deal with critical bursts of requests, marked as pending and for which the final result is unknown in terms of set up or tear down. Such requests are rejected because they might represent, at that moment, an attack to confidentiality. Go to step 7.

5c) Deny the incoming request because of an excessive risk. Go to step 7.

Step 6. The PCS performs the required path computation and the positive or negative result is returned to the PCEP interface.

Step 7. The PCEP interface returns to the PCC the result of the path computation in the form of:

7a) PCRep message with path computation failure (i.e., No-path) or with the computed path.

7b) PCErr for temporary authorization failure.

7c) PCErr for critical authorization failure.

Step 8. Update of the APS databases: the pending request is moved from PRDB to HRDB. Then, the domain Reputation is updated taking into account the overall data stored in HRDB, including the computed risk values. If the domain Reputation goes to critical values, the PCEP session between the domains is eventually closed. Warning messages are specified to allow the requesting domain to become aware of its risky position.

Reputation	$R = c_{NP} \sum_H w_{k, NP} + c_{ERR} \sum_H w_{k, ERR} + c_T \sum_H w_{k, T} - c_{SET} \sum_H w_{k, SET}$			
	$c_{NP}$ : NO-PATH	$c_{ERR}$ : ERROR	$c_T$ : TIMEOUT	$c_{SET}$ : SETUP
	1	1	2	1.1
Request: Risk level	$w_k = \prod_i c_k^i$			
	$C_k$ : {No_value, Yes_value}			
	-Bidirectional: {1, 2}	-LocalProtection: {1, 1.5}		
	-BRPC: {1, 4}	-Link/Node disjoint SVEC: {1,6}		
	-Egress: {1, 2}	-SRLG disjoint SVEC: {1, 7}		
	-Critical BW: {1, 3}			
Risk Correlation: PRDB	$w_{pk} = \sum_{PRDB} w_k$	$w_k = w_{k, ENTRY} \cdot C_D \cdot C_M$		
	$c_D$ : END-POINTS	$c_M$ : METRIC		
	-Transit:1	-No Metric:1		
	-Egress:1.5	-Metric:1.5		
	-Egress, same area: 2			
Risk Correlation: HRDB	$w_{hk} = \sum_{HRDB} w_k$	$w_k = w_{k, ENTRY} \cdot C_D \cdot C_M \cdot C_T$		
	$c_D$ : END-POINTS	$c_M$ : METRIC	$c_T$ : EVENT	
	-Transit:1	-No Metric:1	-No-path:1	
	-Egress:1.5	-Metric:1.5	-Timeout:2	
	-Egress same area:1.7			
	-Egress same area, different attributes: 2			

No.	Time	Source	Destination	Protocol	Info
5	2.625290	193.205.83.105	193.205.83.110	PCEP	OPEN MESSAGE
7	2.623606	193.205.83.110	193.205.83.105	PCEP	OPEN MESSAGE
9	2.623898	193.205.83.105	193.205.83.110	PCEP	KEEPALIVE MESSAGE
10	2.623926	193.205.83.110	193.205.83.105	PCEP	KEEPALIVE MESSAGE
11	2.624306	193.205.83.105	193.205.83.110	PCEP	PATH COMPUTATION REQUEST MESSAGE
13	4.982162	193.205.83.110	193.205.83.105	PCEP	PATH COMPUTATION REPLY MESSAGE
22	12.982808	193.205.83.105	193.205.83.110	PCEP	PATH COMPUTATION REQUEST MESSAGE
29	15.298894	193.205.83.110	193.205.83.105	PCEP	ERROR MESSAGE → TCVE
40	23.388343	193.205.83.105	193.205.83.110	PCEP	PATH COMPUTATION REQUEST MESSAGE
47	25.223439	193.205.83.110	193.205.83.105	PCEP	ERROR MESSAGE → CCVE
53	33.223849	193.205.83.105	193.205.83.110	PCEP	PATH COMPUTATION REQUEST MESSAGE
55	35.917694	193.205.83.110	193.205.83.105	PCEP	ERROR MESSAGE → CCVE
57	35.918084	193.205.83.110	193.205.83.105	PCEP	NOTIFICATION MESSAGE → Warning
60	40.917769	193.205.83.110	193.205.83.105	PCEP	CLOSE MESSAGE → Bad Reputation

Figure 20. Example: reputation, risk weights Figure 21: PCEP message exchange

### 3.3.4.4 Experimental assessment.

The APS behavior has been specified through a set of first-applicable XACML policies [Toktar04, Dem09]. A custom implementation of the APS has been realized based on JAVA code elaborating XACML descriptions. According to the database size, the overall time required to complete the authorization procedure is between 1 and 3 seconds. However, it is important to notice that APS gets involved just on a small fraction of the requests, while the majority are simply handled by the access lists implemented at the local APEC, with negligible additional delay.

The overall architecture implementation includes also novel PCEP objects. The current PCEP specification defines a specific type of Error called Policy Violation [RFC5440].

The proposed scheme requires the definition of two additional Policy Violation error values: the Temporary Confidentiality Violation Error (TCVE) and the Critical Confidentiality Violation Error (CCVE) to respectively specify the temporary and definitive policy violation.

In addition, upon the update of the domain reputation to an insufficient value, a Close message might be generated with the novel Compromised Client Reputation reason.

This message forces the PCEP Session termination. Warning notification messages have been also implemented as shown in **Figure 21**

### **3.3.4.5 Conclusions and future work.**

A two-step policy-based authorization scheme is proposed to increase the level of security in PCEP-based inter-domain computations. The scheme relies on access lists to avoid the direct discovery of critical information and on XACML policies to avoid malicious correlations among different PCEP requests. Some details of the implementation are provided, including novel PCEP objects.

The innovative solution proposed in this section has been presented at the conference ECOC 2010 [Paol-ECOC10].

## **3.4 Control plane in a multi-layer scenario**

### **3.4.1 Multi-layer PCE-based architecture**

Path Computation Element (PCE) Architecture has been defined to compute and provide effective Traffic Engineering solutions. An accurate and timely PCE TE Database (TED) is then required. Traditionally, the PCE TED has been retrieved from a link state routing protocol (e.g., OSPF-TE). However, the amount of TE information to account for may be extremely high, particularly in the case of detailed WSON information [draft-lee] or, as considered in this study, of Forwarding Adjacency Label Switch Paths (FA-LSPs) information in GMPLS multi-layer networks (MLN) [RFC5212]. The advertisement of this kind of information through a routing protocol may determine convergence and scalability issues and may affect the processing, storage and communication performance of network nodes. In [draft-lee], three alternative methods to create and maintain a PCE TED are investigated: (1) nodes send local information to all PCEs; (2) nodes send local information to an intermediate server that will relay it to all PCEs; (3) nodes send local information to at least one PCE and have the PCEs share this information with each other. In [draft-lee], due to the informational nature of the document, no implementation details are provided. No practical solutions have been proposed so far, especially in the context of GMPLS MLN. In this study, we focus on the implementation of a slightly upgraded version of the third method, where, for reliability purposes, TE information is sent to at least *two* PCEs.

The considered GMPLS multi-layer network (MLN) scenario, in which the proposed method is applied, consists in a two-layer network, where layers are characterized by the same Interface Switching Capability (ISC) and are referred to as lower (e.g., ISC of type Lambda Switching Capable (LSC)) and upper (e.g., ISC of type Packet Switching Capable (PSC)). In compliance with the GMPLS MLN specifications [RFC5212]: (i) a *single* GMPLS control plane instance is considered; (ii) a Label Switched Path (LSP) starts and ends at the same layer (i.e., ISC); (iii) once an LSP is established at the lower layer from one layer border node to another, it can be used as a data link in the upper layer. Furthermore, an LSP at the lower layer can be advertised as a TE Link and exploited in the path computation of LSPs originated by different nodes. Such TE Link is referred to as FA-LSP. An FA-LSP has the special characteristic that it does not require the set up of a routing adjacency (peering) between its end points. At the upper layer, the FA-LSPs compose the Virtual Network Topology (VNT) provided by the lower layer. The VNT facilitates the path computation of LSPs in MLN since it describes the resources at a single layer. To address scalability and reliability requirements, multiple PCEs per layer are typically considered:  $L\text{-PCE}_i$  and  $U\text{-PCE}_j$  are responsible for path computations at the lower and upper layer respectively ( $2 \leq i \leq M$ ,  $2 \leq j \leq N$ ). The PCE TEDs are retrieved from the routing protocol, e.g. by listening to the OSPF-TE advertisement. It is an implementation decision whether or not the whole VNT is advertised and made available in the path computation of LSPs originated by different nodes.

Two main implementation schemes have been discussed and considered so far. In the first scheme, here referred to as No-FA, the FA-LSPs are not advertised at the upper layer. In No-FA, upon connection request from source  $s$  to destination  $d$ ,  $U\text{-PCE}_j$  performs the path computation by exploiting just FA-LSPs starting at node  $s$  and terminating at node  $d$ . If this computation fails because of lack of resources,  $U\text{-PCE}_j$  requests  $L\text{-PCE}_i$  to compute a new segment or path at the lower layer, which is then exploited to complete the path computation request. In the second scheme, here referred to as CP-FA, the FA-LSPs are advertised in the control plane and the whole VNT is exploited in the path computations of LSPs originated by different nodes. As in No-FA, also in CP-FA when the path computation fails,  $U\text{-PCE}_j$  requests  $L\text{-PCE}_i$  to compute a new segment or path at the lower layer. For both schemes, if also the path computation performed by PCE-L fails, the request is rejected. On the one hand, the full availability of the FA-LSPs provided by CP-FA with respect to No-FA may improve the overall network resource utilization since it allows the implementation of effective grooming policies. On the other hand, the additional advertisement required by CP-FA may significantly affect the control plane and network stability and scalability. It is worth noticing that intermediate schemes may be adopted, e.g. based on the advertisement of the sole FA-LSPs guaranteeing a certain amount of available bandwidth. However, such schemes would not be able to provide lower control plane load than No-FA or better resource utilization than CP-FA.

The scheme proposed in this study exploits both (i) the No-FA scheme and (ii) an alternative method for TED creation. The scheme, called DP-FA, resorts to a Designated PCE to which TE information is propagated through a *separate* instance of OSPF-TE. Differently from typical control plane instances, in DP-FA, all L-PCEs first elect a Designated PCE (DP) and a Backup Designated PCE (DP) per layer. In a two-layer network, this corresponds to the election of an L-DP and an L-BDP at the lower layer, a U-DP and a U-BDP at the upper layer. The list of eligible PCEs is provided through manual configuration or by exploiting the automatic discovery procedures defined in [RFC5088].

The procedure for DP and BDP elections is the one defined for the Designated Router (DR) and Backup DR in OSPF-based broadcast networks. As for BDR, BDP is elected for reliability purposes: in case the elected DP fails, it becomes the new DP. A separate instance of OSPF-TE adjacency is established between each PCE and the elected DP (and BDP) belonging to the same layer. In addition, adjacencies are established between DP (and BDP) belonging to adjacent layers. Two main procedures are defined.

First, L-DP is elected to be responsible for the advertisement of the FA-LSPs to the upper layer (i.e., to U-DP).

Second, U-DP is elected to be responsible for the flooding of the TE information received from L-DP to the other U-PCEs.

In this way, the exchange of TE information (e.g., the full set of FA-LSPs) is performed without forming the full mesh of adjacencies among all PCEs and thus avoiding the resultant chaotic and inefficient flooding of many copies of the same LSA. In addition, differently from typical OSPF-based TE information exchange, the exchange of TE information is proposed to be implemented in a unidirectional way: no TE information is returned during the adjacency set up or upon changes in the resource utilization (i) from the U-DP to the L-DP (and L-BDP) and (ii) from the U-PCEs to the U-DP (and U-BDP).

### 3.4.1.1 Simulation results

The performance of the three considered schemes have been evaluated through a custom event-driven simulator. Two different two-layer network topologies are considered: a Ring and a Pan-European [Cug05] network with  $N_{Ring}=N_{Euro}=17$  nodes and  $L_{Ring}=17$  and  $L_{Euro}=32$  bidirectional links respectively. In both scenarios, each link carries  $W=40$  wavelengths at 10Gb/s. Network nodes are equipped with both LSC and PSC interfaces.  $T=20$  tunable transponders at the PSC layer are considered in each node to serve as adaptation/termination capacity. The two considered networks represent the two typical scenarios where blocking is first achieved because of lack of available wavelengths (i.e., the Ring) or lack of available transponders (i.e., the Pan-Euro).

Each unidirectional traffic request, uniformly distributed among all PSC layer interfaces, requires the set up of  $b=1$  Gb/s bandwidth guaranteed LSP. Requests are provisioned, in all the schemes, following the *MinTH* grooming policy presented in [Zhu03]. Simulation points are depicted with the confidence interval at 95% confidence level.

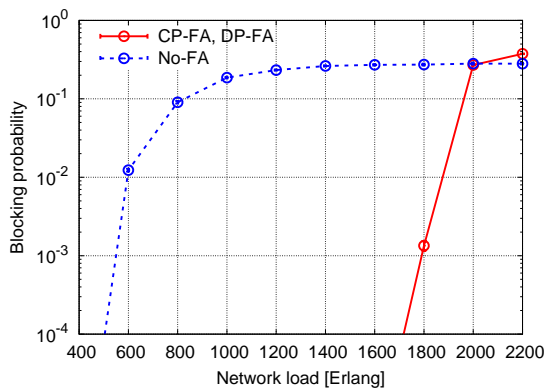
Figure 22 and Figure 23 show the blocking probability as a function of the offered load of the three schemes in the two considered networks. Despite the different scenario and link usage, the schemes provide similar relative performance. All curves show a steep increase, rapidly passing from low to high blocking as soon as the critical resources (either links or transponders) are exhausted. Results show that CP-FA and DP-FA, by exploiting all available FA-LSPs, significantly outperform No-FA in terms of blocking probability.

Figure 24 and Figure 25 show the control plane load expressed in terms of LSA announcements per second as a function of the network load. Results show that No-FA and DP-FA provide the same control plane load which is significantly lower with respect to CP-FA. At low network loads the amount of advertised LSAs is higher since each set up/teardown of an FA-LSP induces the flooding of  $n$  LSC LSAs, where  $n$  is the number of

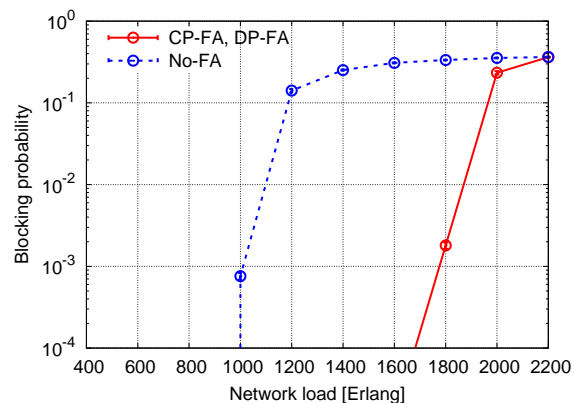
traversed links. With the increase of the load, on the one hand, as shown by No-FA and DP-FA curves, the lower layer becomes almost stable and few LSC LSAs are propagated. On the other hand, with CP-FA the amount of advertised LSAs increases since one PSC LSA is advertised upon each connection set up or release. At very high network load, where blocking becomes excessively high, the amount of advertised LSAs in CP-FA slightly decreases: LSPs are typically established along non-shortest routes exploiting multiple FA-LSPs, which in turn tend to remain more stable.

*Conclusions.* In this study, a Designated PCE Election procedure is proposed for PCE TED creation in multilayer PCE-based control plane architectures. The proposed procedure, by moving the PSC LSA information exchange on an out-of-band communication between PCEs, enables the effective exchange of TE information referred to Forwarding Adjacencies LSPs between PCEs operating at different switching layer. Simulation results show that the proposed scheme guarantees lightweight control plane loads without affecting the overall network resource utilization, as provided by schemes announcing the whole set of available resources.

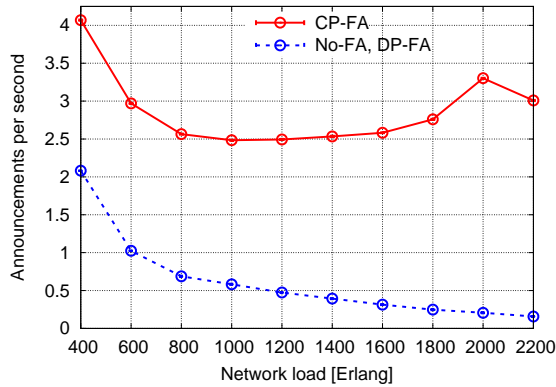
The innovative solution proposed in this section has been presented at the conference ECOC 2010 [Cugini10].



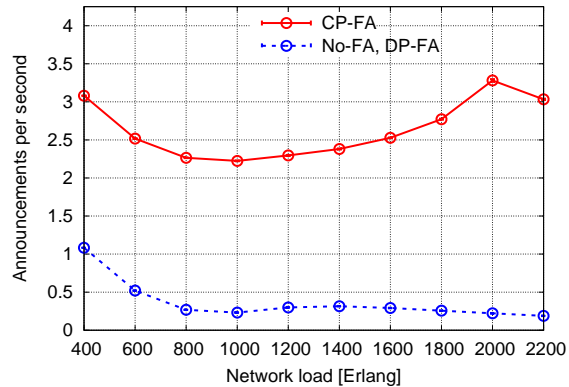
**Figure 22: Blocking Probability - Ring network**



**Figure 23: Blocking probability - PanEuro**



**Figure 24: Control plane load – Ring network**



**Figure 25: Control plane load – PanEuro network**

### 3.5 Specific issues in path computation

#### 3.5.1 Point-to-Multipoint (P2MP)

Multicast services are increasingly demanded for high-capacity applications such as multicast Virtual Private Networks (VPNs), IP- television (IPTV) which may be on-demand or streamed, and content-rich media. The ability to compute constrained Traffic Engineering Label Switched Paths (TE LSPs) for point-to-multipoint (P2MP) LSPs in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains has been identified as a key driver for the adoption of PCE based path computation.

Work is ongoing within STRONGEST WP3 to describe how multiple PCE techniques can be combined to address the requirements. These mechanisms include the use of the per-domain path computation technique specified in [RFC5152], extensions to the backward recursive path computation (BRPC) technique specified in [RFC5441] for P2MP LSP path computation in an inter-domain environment, and a new procedure for core-tree based path computation defined in this document. These three mechanisms are suitable for different environments (topologies, administrative domains, policies, service requirements, etc.) and can also be effectively combined.

As discussed in [RFC4461], a P2MP tree is a graphical representation of all TE links that are committed for a particular P2MP LSP. In other words, a P2MP tree is a representation of the corresponding P2MP tunnel on the TE network topology. A sub-tree is a part of the P2MP tree describing how the root or an intermediate P2MP LSPs minimizes packet duplication when P2P TE sub-LSPs traverse common links. As described in [RFC5671] the computation of a P2MP tree requires three major pieces of information. The first is the path from the ingress LSR of a P2MP LSP to each of the egress LSRs, the second is the traffic engineering related parameters, and the third is the branch capability information.

Generally, an inter-domain P2MP tree (i.e., a P2MP tree with source and at least one destination residing in different domains) is particularly difficult to compute even for a distributed PCE architecture. For instance, while the BRPC recursive path computation may be well-suited for P2P paths, P2MP path computation involves multiple branching path segments from the source to the multiple destinations. As such, inter-domain P2MP path computation may result in a plurality of per-domain path options that may be difficult to coordinate efficiently and effectively between domains. That is, when one or more domains have multiple ingress and/or egress border nodes, there is currently no known technique for one domain to determine which border routers another domain will utilize for the inter-domain P2MP tree, and no way to limit the computation of the P2MP tree to those utilized border nodes.

It is assumed that, due to deployment and commercial constraints (e.g., inter-AS peering agreements), the sequence of domains for a path (the path domain tree) will be known in advance. The algorithms to compute the optimal large core tree are outside scope, but a basic approach is shown in the following.

### 3.5.1.1 Core Tree Computation Procedures

The following extended BRPC based procedure can be used to compute the core tree. First, using the BRPC procedures to compute the VSPT(i) for each leaf BN(i),  $i=1$  to  $n$ , where  $n$  is the total number of entry nodes for all the leaf domains. In each VSPT(i), there are a number of  $P(i)$  paths. When the root PCE has computed all the VSPT(i),  $i=1$  to  $n$ , take one path from each VSPT and form a set of paths, we call it a PathSet(j),  $j=1$  to  $M$ , where  $M=P(1) \times P(2) \dots \times P(n)$ . Next, for each PathSet(j), there are  $n$  S2L (Source to Leaf BN) paths and form these  $n$  paths into a Core Tree(j). There will be  $M$  number of Core Trees computed, so apply the OF to each of these  $M$  Core Trees and find the optimal Core Tree.

Note that the application of BRPC in the aforementioned procedure differs from the typical one since paths returned from a downstream PCE are not necessarily pruned from the solution set by intermediate PCEs. The reason for this is that if the PCE in a downstream domain does the pruning and returns the single optimal sub-path to its parent PCE, BRPC ensures that the ingress PCE will get all the best optimal sub-paths for each LN (Leaf Border Nodes), but the combination of these single optimal sub-paths into a P2MP tree is not necessarily optimal even each S2L (Source-to-Leaf) sub-path is optimal. Without trimming, the ingress PCE will get all the possible S2L sub-paths set for LN, and eventually by looking through all the combinations, and taking one sub-path from each set to build one p2mp tree, it finds the optimal tree.

The proposed method may present a scalability problem for the dynamic computation of the Core Tree (by iterative checking of all combinations of the solution space), specially with dense/meshed domains. Considering a domain sequence  $D1, D2, D3, D4$ , where the Leaf border node is at domain  $D4$ , PCE(4) will return 1 path. PCE(3) will return  $N$  paths, where  $N$  is  $E(3) \times X(3)$ , where  $E(k) \times X(k)$  denotes the number of entry nodes times the number of exit nodes for that domain. PCE(2) will return  $M$  paths, where  $M = E(2) \times X(2) \times N = E(2) \times X(2) \times E(3) \times X(3) \times 1$ , etc. Generally speaking the number of potential paths at the ingress PCE is given by  $Q = \text{prod } E(k) \times X(k)$ . Consequently, it is expected that the Core Path will be typically computed offline, without precluding the use of dynamic, online mechanisms such as the one presented here, in which case it SHOULD be



possible to configure transit PCEs to control the number of paths sent upstream during BRPC (trading trimming for optimality at the point of trimming and downwards).

### 3.5.1.2 Sub Tree Computation Procedures

Once the core tree is built, the grafting of all the leaf nodes from each domain to the core tree can be achieved by a number of algorithms. One algorithm for doing this phase is that the root PCE will send the request for the path computation to the destination(s) directly to the PCE where the destination(s) belong(s) along with the core tree computed from the previous. This approach requires that the root PCE manage a potentially large number of adjacencies (either in persistent or non-persistent mode), including PCEP adjacencies to PCEs that are not within neighboring domains.

A first alternative would involve establishing PCEP adjacencies that correspond to the PCE domain tree. This would require that branch PCEs forward requests and responses from the root PCE towards the leaf PCEs and vice-versa. Finally, another alternative would use a hierarchical PCE (H-PCE) architecture. The "hierarchically" parent would require sub tree path computations.

### 3.5.2 Topology summarization method

Hierarchical architectures are based on the summarization of lower layers' topologies, which are presented to the higher layers in a simplified way. That resources' virtualization allows to both improve network scalability by masking the lower layers' topology complexity and to keep confidentiality in multi-carriers contexts.

As a consequence, PCE-based hierarchical architectures, such the ones considered within the STRONGEST project, should rely on the ability of automatically build a summarized view of domains topologies, composed by virtualized elements. Such Summarized Topologies should provide a simplified way to represent the network, reducing the number of nodes and links (and so improving scalability of the overall inter-domain system), while preserving the information needed to correctly perform an inter-domain Path Computation according to a set of service-oriented parameters.

Given a certain domain, one of the considered methods should be the creation of a summarized topology composed by only its Border Elements –BE- (i.e. the nodes of the domain having interfaces with other domains and/or other regions), interconnected by a full mesh of equivalent summarized links (i.e. virtual links representing a set of paths with certain common characteristics/parameters). Other kinds of topologies (e.g. comprising also some virtual internal node, other aggregation methods than full mesh, etc.) should be also considered, but we considered them out of scope for this stage of the work.

Particularly, for each virtual link interconnecting a couple of BEs, a domain should advertise the following parameters:

- the average delay to cross the domain using the considered virtual link;
- the minimum guaranteed bandwidth on such virtual link;

- the maximum available bandwidth on such virtual link;
- the peak bandwidth of such virtual link;
- a set of optional parameters.

**Table 4: Set of proposed common parameters**

Common parameter	Mandatory	Packet-switched Domain	Wavelength-switched Domain
<b>Delay</b>	<b>Yes</b>	= sum of processing time along all nodes and transmission time along all links of a path between border elements (transmission time can be also ignored in case of minor values)	= sum of transmission times along all links in a path between border elements
<b>Guaranteed bandwidth (GBW)</b>	<b>Yes</b>	= link with the minimum guaranteed bandwidth along a path between border elements	= bandwidth allocated to the wavelength channel (e.g. 2.5G, 10G, 40G)
<b>Peak bandwidth (PBW)</b>	<b>Yes</b>	= link with the minimum link capacity along a path between border elements	= bandwidth allocated to the wavelength channel (e.g. 2.5G, 10G, 40G)
<b>Max Available bandwidth (MaxAvBW)</b>	<b>Yes</b>	= maximum value of the guaranteed bandwidth for a set of paths between two border elements	= maximum value of the guaranteed bandwidth for a set of paths between two border elements
<b>Recovery Scheme</b>	<b>No</b>	=the recovery scheme(s) performed to protect a path between border elements	=the recovery scheme(s) performed to protect a path between border elements
<b>Admin. color</b>	<b>No</b>	= an administrative ID used to group some paths between border elements having similar characteristics	= an administrative ID used to group some paths between border elements having similar characteristics

Table 4 describes an advantageous set of common parameters (peak bandwidth, guaranteed bandwidth, delay and some optional ones) and how these are derived from parameters collected from a domain. Particularly, mandatory parameters are essentially service-based parameters (e.g. UNI parameters), with the aim to have homogenous service-oriented metrics across domains which uses different technologies; however, optional parameters (e.g. technology-specific ones, QoS related ones) are also allowed. Attributes related to operator's requirements, such as SRLGs, administrative colors, economic metrics, etc.) should also be covered, but again we considered them out of scope for this stage of the work.

As a matter of fact, to recap different technologies under a common "umbrella" is of a paramount importance when an end-to-end service must cross different domains and every domain has its own peculiar policies and technology-related behaviors and parameters.

Several intra-domain paths can satisfy the parameters' ranges which define each virtual link. All these paths are therefore grouped in a "Class Basket", representing the class of equivalence for the advertisement of the connectivity ensured by the virtual link.

Some ranges of values are considered for delay or guaranteed bandwidth, so the paths belonging to the same baskets have the considered parameter (i.e. delay or guaranteed bandwidth) within the considered range of values. The number of paths belonging to the same basket depends on the level of configurability (e.g. lambda switching, tunable transponder, etc.) of the elements in the optical nodes. The higher is the level, the higher is the possibility to have more paths for the same virtual connection (i.e. the same basket).

The definition of the basket is an asynchronous operation, with respect to the multi-domain routing; therefore, each domain can apply a proper policy/routing strategy to define the internal paths for the same basket. As a matter of fact, an entity working on a multi-domain scope (e.g. an inter-domain PCE) would work on the virtual (summarized) topology that is represented as homogeneous links and parameters, while each domain would apply its peculiar internal routing policies. Technological specific constraints (e.g. wavelength continuity, physical impairments) are considered in the intra-domain routing to define which resources belong to the basket, but they are not considered in the of multi-domain routing.

Such summarized view enables a domain to advertise a single connectivity instead of the whole pattern of paths which are hidden outside the domain. Three main criteria can be considered to define the paths belonging to a basket:

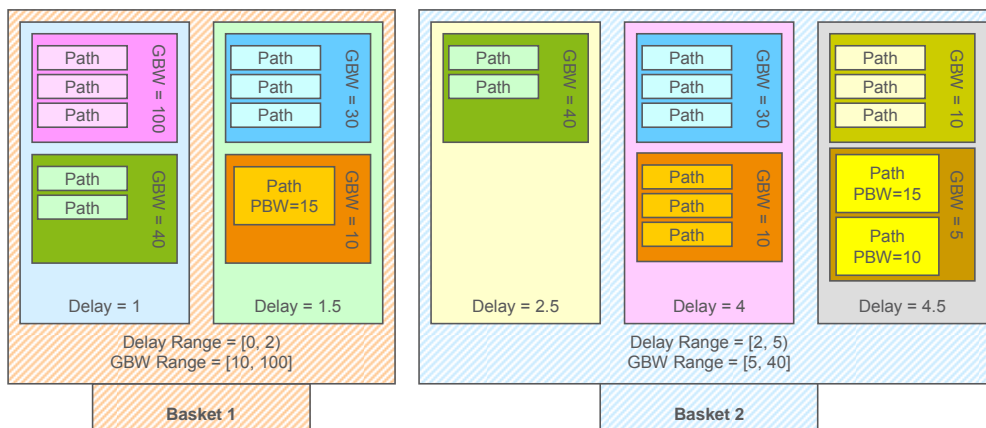
- **Fully pre-planned connectivity.** The domain computes in advance all the possible paths which are compliant with the basket “policies”.
- **Partially pre-planned connectivity.** The domain computes off-line in advance only a subset of such paths, while other ones are computed on-line dynamically.
- **Fully dynamic connectivity.** All paths are computed on-line dynamically.

Obviously each domain can independently enforce its own strategy to fill the baskets, according to the domain’s technology (e.g. due to longer setup time, optical paths are often pre-planned) and policies. Moreover, once defined the common parameters to be summarized, each domain can group its internal paths according to different criteria, organizing them also in more complex schemes (see section 4.5, [Paol-JOCN10] and [Iov-Bot-DBA\_MD]). As an example, a domain can define its baskets according to a single most valued parameter, having a set of baskets, while another domain can define its baskets according to two or more criteria simultaneously, having matrix of baskets, etc.). However, the result of the topology summarization would be a set of virtual links summarizing the connectivity between two BEs with the same set of common parameters, independently of how they were summarized by the correspondent domain.

For the sake of simplicity, let’s refer to the simple case of summarizing the connectivity between only two border elements of a domain, where the “Fully pre-planned connectivity” case is considered and the basket definition is restricted only to the delay. In that case the domain’s policy would build a set of baskets according to the following steps:

1. a set of baskets are defined according to some delay ranges (i.e. the delay is considered as the most valued parameter for defining the baskets);

2. the summarization algorithm scans the connectivity between the two Border Elements and found all the possible paths connecting them (i.e. the fully pre-planning is performed);
3. paths are first sorted by delay, from the lower delay value to the higher one, and then stored in the different baskets according to the delay range they belong to;
4. among the paths belonging to the same basket, paths with the same delay are sorted by the guaranteed bandwidth from the higher value of guaranteed BW to the lower one (i.e. the GBW is considered as the second most valued parameter for defining the baskets);
5. on equal terms for delay and bandwidth ranges, the paths are sorted by peak bandwidth, from the higher value of peak bandwidth to the lower one (i.e. the peak bandwidth is considered as the third most valued parameter for defining the baskets);



**Figure 26: Class Baskets representing the connectivity between two Border Elements**

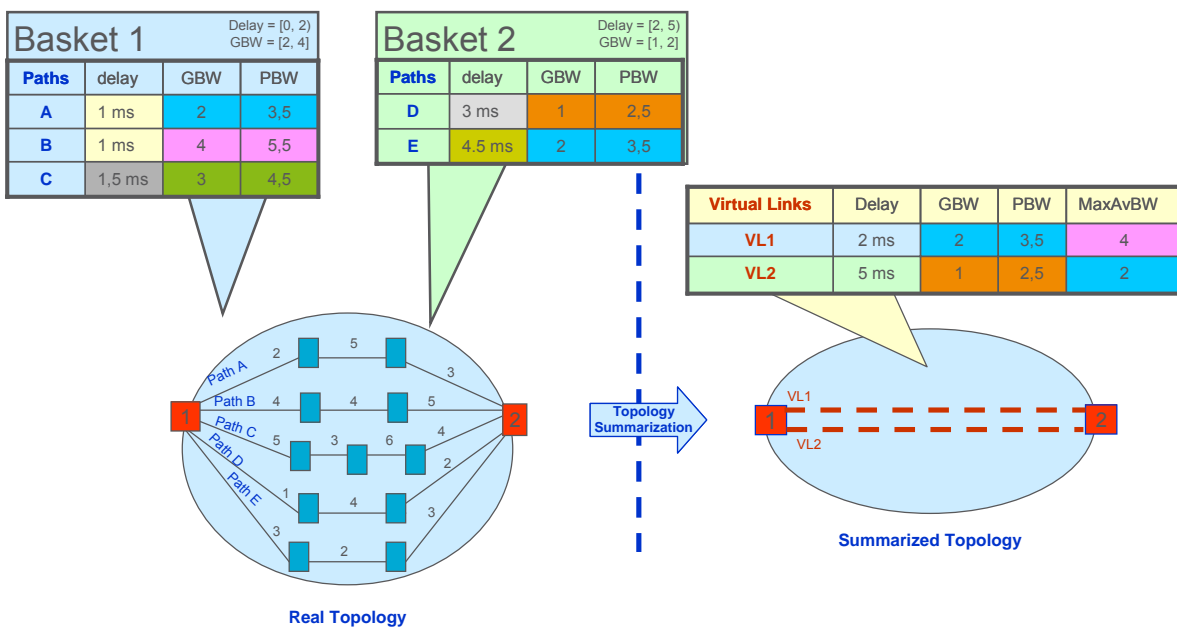
As a result, a set of baskets will be defined and filled, in order to be advertised as a single virtual link summarizing all the paths belonging to the basket. Figure 26 shows an example where 22 paths between a couple of border elements are grouped into 2 baskets, defined for the delay ranges [0ms, 2ms) and [2ms, 5ms) and having GBW ranges of [10G, 100G] and [5G, 40G] respectively.

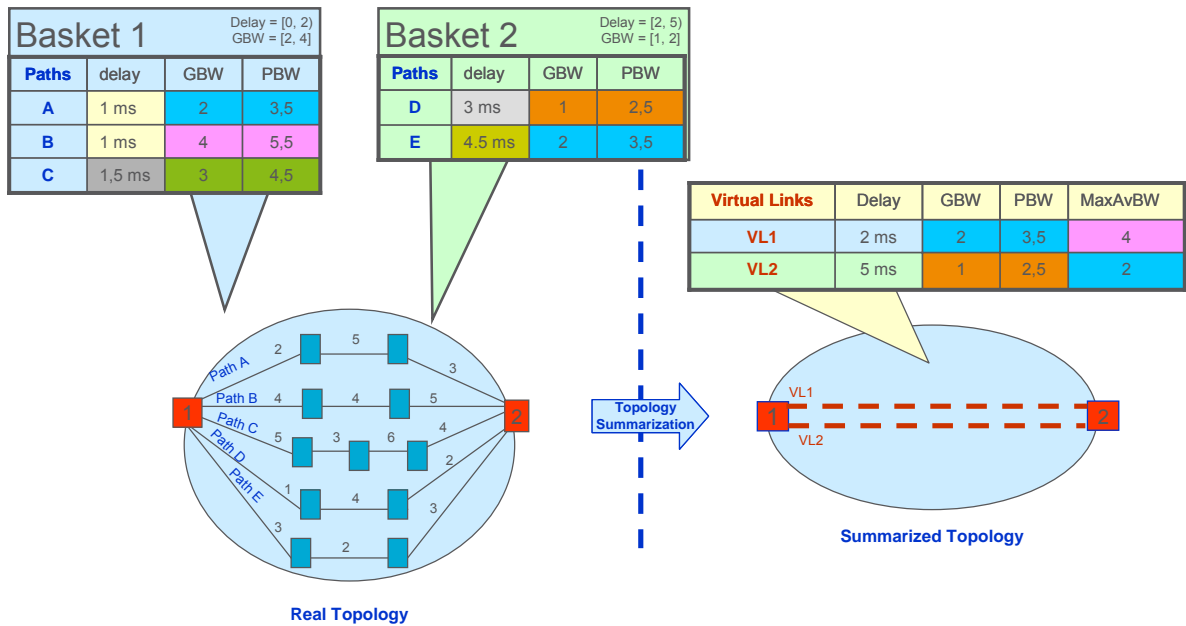
Once defined and filled the baskets representing the connectivity between all the Border Elements, the domain's summarized topology would be computed as follows:

- for each basket representing the connectivity between a given couple of border elements a virtual link is advertised;
- for each virtual link, the delay parameter is advertised as the upper boundary of the delay range that defines the correspondent basket;
- for each virtual link, the GBW parameter is advertised as the minimum of the GBW values among all the paths belonging to the correspondent basket;

- for each virtual link, the PBW parameter is advertised as the PBW value of the path from which the GBW value was derived;
- for each virtual link, the MaxAvBW parameter is computed as the maximum of the GBW values of all the paths belonging to the correspondent basket (i.e. the maximum allocable bandwidth for the considered connection).

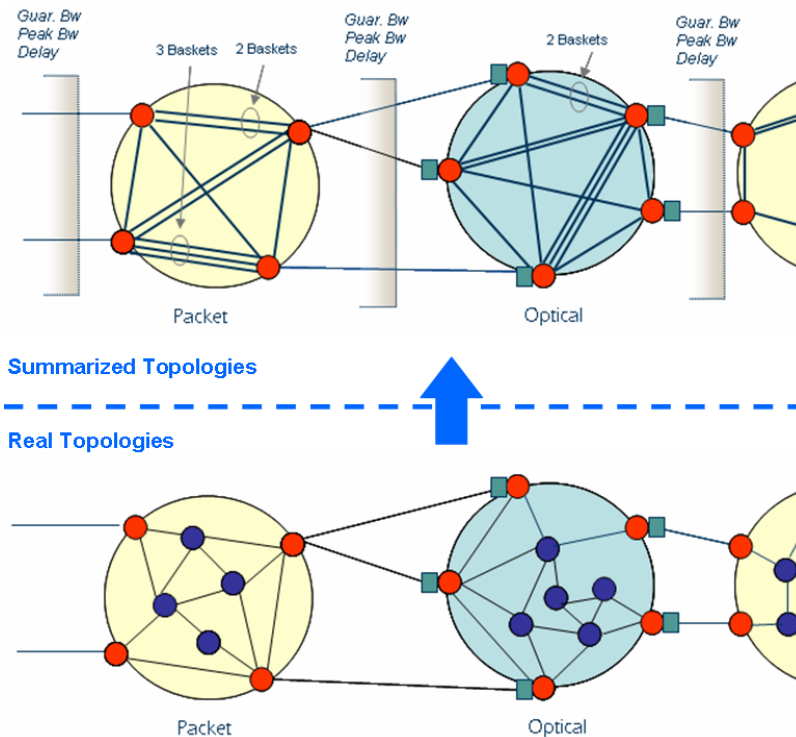
The values of parameters describing a given virtual link remain unchanged until in the correspondent basket there will be at least a path that can satisfy them. That way the updating rate is significantly reduced.





reports an example of summarization for a domain having 5 paths connecting its two Border Elements. The five paths are shown in the domain's Real Topology, together with the available bandwidth values for each link. Two ranges of delay are considered, so two baskets are defined and two virtual links are advertised in the summarized topology. It is worth to notice that the paths to be summarized need not to be necessarily disjoint, therefore, within the updating policies, several paths may change after a successful establishment.

Finally, Figure 27 depicts the outcome of the summarization procedure applied to different domains.



**Figure 27: Summarization of different domains**

After the process, the connectivity of each domain is resumed by virtual links interconnecting the border elements of the domain itself. Independently on which is the technology of each domain and which is the strategy to define the baskets, the final summarization is homogeneous as end to end topology crossing all domains.

A possible application of the above considered scenario is a multi-domain hierarchical PCE-based architecture where there is supposed to be at least a “ChildPCE” for each domain, having an intra-domain scope and at least a “ParentPCE”, having an inter-domain scope. The Children PCEs generate a summarized view of their own domain’s topology with summarized resources and send it to the parent PCE. The Parent PCE is therefore made aware of a summarized view of the entire multi-domain topology which makes it able to choose the best chain of domains to be traversed and to perform a suitable end-to-end inter-domain path computation.

### 3.5.3 Management and adaptation of PCE algorithms

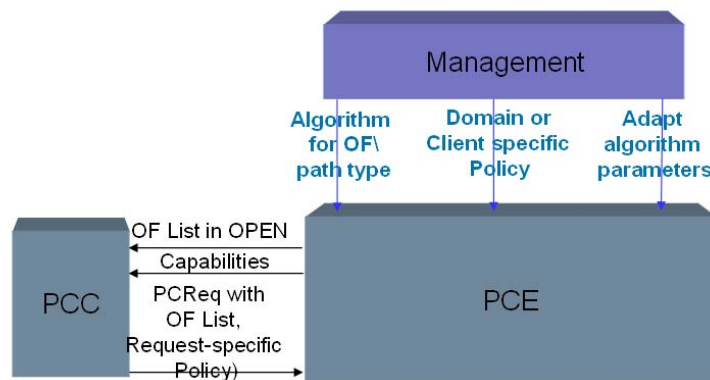
#### 3.5.3.1 Motivation

PCE may support several path computation algorithms, and usually has a default algorithm to run upon path computation requests. However it may apply various algorithms for different path requirements or network conditions. In order to gain more flexibility and manage the network, specifically the PCE, in a more efficient way, it is suggested in this chapter to allow network operators to configure the path computation algorithm from management. Configuring and fine-tuning the algorithm has major advantages of

personalizing the network, achieving better performance, as different networks have different requirements. The concept covered in this chapter is an addition to current PCE documentation and standardization.

The scope of this chapter is managing and affecting the algorithm from management, which shall be covered in all of CLI\SNMP\Web interfaces. The influence of PCC on the algorithm through selecting OF\policy\constraints is not related to this chapter, and neither contradicts it. Moreover, PCC should not care or know the algorithms names which run at PCE side. Therefore publishing of supported algorithms within PCE capabilities is not necessary. The idea of this contribution is to allow configuration or tuning the algorithm.

All configuration and adaptation alternatives are not mandatory. They are depicted in this chapter as a tool for better control and for a smarter PCE. The configurations described hereby shall not affect already computed paths. Only new paths requests will use the new configured algorithm.



**Figure 28: Managing PCE algorithms**

### 3.5.3.2 PCE algorithm management alternatives

PCE algorithm is proposed to be configured from management in 3 ways:

- A) Translation of an OF or a path type to a specific algorithm
- B) Applying a policy
- C) Tuning of specific parameters of an algorithm

All are depicted in

**Figure 28** and further explained.



### A) Translation of an OF or a path type to a specific algorithm

This option is mainly used when deploying a PCE in the network, during which operators may configure the following:

#### 1. Default algorithm

When PCE does not have a specific instruction which algorithm to run it uses the default algorithm.

#### 2. Algorithm for P2P/P2MP

The difference between path types brings to various algorithms which some of them are more suitable for P2P while others do better with P2MP. In addition, it is well known that different P2MP algorithms may result with different cost of multicast trees. Operators may have the strength to influence in such case.

#### 3. Algorithm for Multi-domain paths

This algorithm should be based on the agreed metric between operators governing the inter-domain path.

#### 4. Algorithm for Multi-layer paths

In multi-layer network a link or a node may be traversed twice. Such behavior of the algorithm is different than in a single layer path algorithm.

#### 5. Algorithm per specific Objective Function (OF)

The OF is a set of one or more optimization criteria to bear in mind when computing the TE LSP(s). A list of standard OFs appears in RFC 5541, while other OFs can be defined in other papers. OFs for example could be: Minimum Cost Path, Minimum Load Path, Minimum aggregate Bandwidth Consumption, etc.

The PCE supports one or more OFs. One of them is the default OF. Generally, an OF implies an algorithm. However, one OF may be mapped to more than one path algorithm (e.g.: for adding more advanced algorithms to the PCE). This configuration option mandates the usage of a specific algorithm when a certain OF is requested by the PCC.

For example, an OF for disjoint pair of paths, may be treated by operators differently in certain areas. Fully disjoint paths are not always desirable. In areas with low probability of failures operators would prefer partially disjoint paths rather than fully disjoint paths. It helps reducing the dollar cost using less disjoint resources in the pair of paths.

PCC may ask a specific OF or policy in the PCReq (Path Computation Request) and hereby implicitly influence the algorithm. Yet, this chapter focuses on local explicit configuration.

Another case for algorithm per OF could be in a certain domain where network is WSON. Then the algorithm may include physical layer validation.

The mentioned configurations may be required also in runtime, upon operators' experience (being convinced that other algorithms may do better after benchmarking current behavior vs. other potential configurations) or after PCE upgrade (with more advanced algorithms).

#### B) Applying a policy - instructing a certain algorithm in certain conditions

Policies may be managed by the network operator. This configuration option allows the operator to apply a domain-specific or client-specific policy which override the algorithm, in case certain conditions take place.

The following example uses the terms "Policy Condition" and "Policy Action" from RFC 3198 (Policy-Based Management).

Policy Condition:

If number of P2MP requests exceeds N

Or

PCE CPU load > X%

Policy Action:

Use less memory consumption algorithm for P2P and OF i

P2MP path computations are very CPU intensive. If the operator wants to fulfill the large number of P2MP requests, by using such policy the PCE is instructed to override the algorithm for P2P requests and OF i related requests. This policy action may result in less optimal paths for P2P and OF i, but may reduce the memory load and accept more path requests during this overloaded period of time.

#### C) Adaptation of specific parameters of an algorithm

The last option allows adaptation of the algorithms, in a way that only the values of algorithms' parameters are being changed but the algorithms are not replaced. For example, after adding new links to an existing network, the operator may decide to lower the cost of neighbor links. The algorithm parameters are well known to the operator, and the configuring function only includes the algorithm identifier, the list of parameters to be updated, and their new values. Also here any modification applies only for new computations.

All of the proposed three mentioned alternatives can be easily standardized and will be further offered in a MIB.

## 3.6 RACS/PCE integrated architecture in inter-domain/inter-carrier scenario

### 3.6.1 Proposed architecture in inter-domain/inter-carrier scenario

In the present section, a complete Control-Layer and Control-Plane architecture for the inter-domain/inter-carrier STRONGEST reference scenario 3 (see [STR-D31], Section 2.3) is presented and analyzed. Such proposal extends, in a multi-domain and multi-carrier scenario, the architecture presented in [STR-D31], Section 4.3. Main objectives are:

- The analysis and the possible extensions of the existing standard RACS inter-domain architecture and interfaces
- The integration of PCE-RACS functionalities in an inter-domain and inter-carrier scenario

The proposed architecture aims at addressing the aforementioned requirements in terms of routing, signaling and path computation, and describes an extension of traditional GMPLS and PCE solutions, in order to implement complete control framework architecture, supporting value added services in a NGN.

Two main inter-domain networking scenarios can be identified:

- **Intra-carrier:** where multiple (routing) domains are interconnected within the same administrative domain
- **Inter-carrier:** where multiple administrative domains are interconnected to each other

The first, intra-carrier, scenario was already analyzed in [STR-D31], and the Service-based Policy Decision Function SPDF<sup>1</sup> was introduced as the element managing, among other elements, the inter-domain interconnections. Objective of the present study is the second, inter-carrier, scenario.

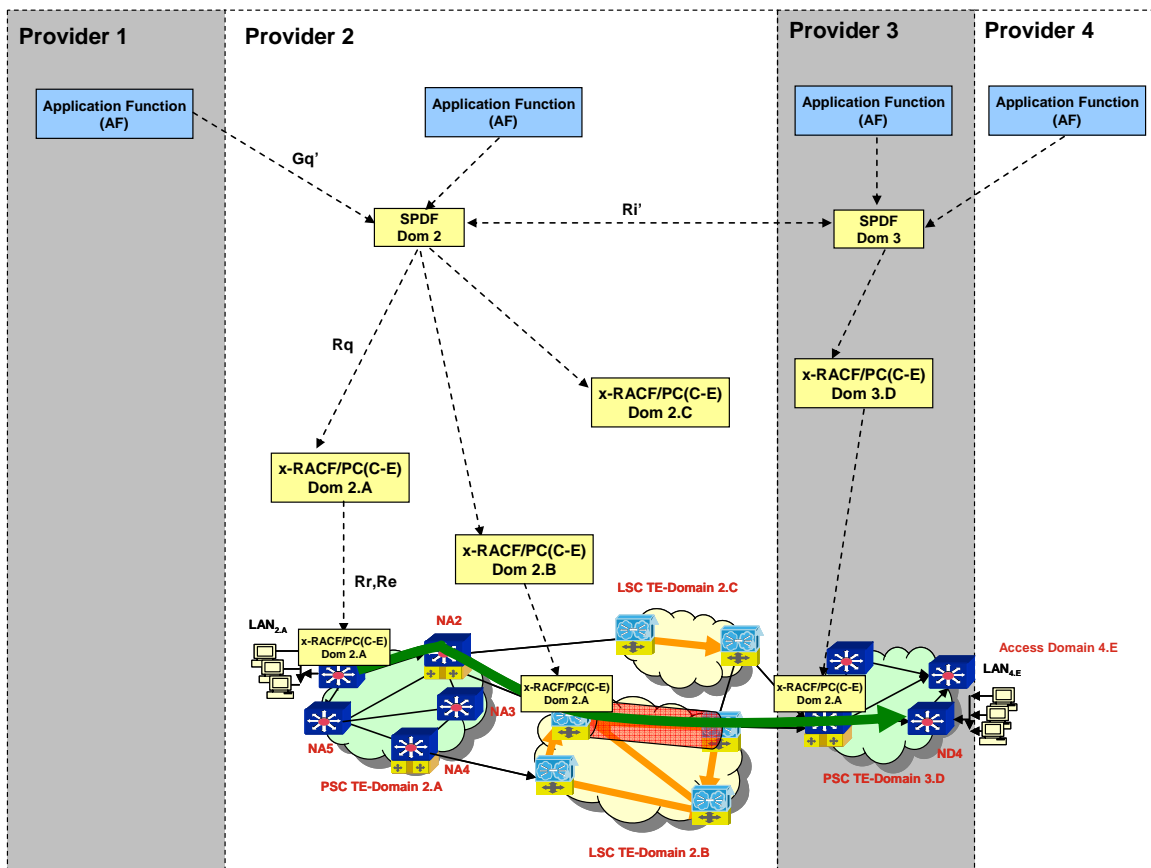
In the ETSI/TISPAN RACS architecture, the only element provided with inter-carrier capabilities is the SPDF. In particular, as depicted in Figure 29, the SPDF may provide the following interfaces for inter-domain and inter-carrier communication:

- **The Gq' interface towards the Application Function (AF).** Such inter-carrier scenario is applicable to:

---

<sup>1</sup> In STR-D31 a single SPDF interconnected with many x-RACF/PCE modules was presented. Moreover, many SPDF could also be interconnected with each other, by the Rd' inter-domain and intra-carrier interface. Such scenario is useful, for example, for geographic redundancy purposes

- A service/network provider lacking in the control framework. In such case no admission control or resource reservation would be performed by the provider without the control framework (e2e model is not applicable). Such scenario is reported in Figure 29 for Provider 4.
- A Service Provider delivering services in a network managed by another Network Provider with a complete control framework. Such scenario, supporting e2e model, is reported in Figure 29 for Provider 1. Such scenario, with a clear network and service provider separation, will be further detailed in Chapter 4.
- **The Ri' interface towards another SPDF.** Such inter-carrier scenario, reported in Figure 29 for Providers 2 and 3, is applicable to providers supplied with a control framework. Therefore, each carrier involved should own at least the SPDF module. However, in order to perform e2e admission control and resource reservation, and, in our case study, in order to perform e2e path computation, each carrier should own a complete G-RACS framework, as described in [STR-D31] (Section 4.3)



**Figure 29 – Possible inter-carrier scenarios**

In order to support inter-carrier communication, the functional elements of the G-RACS architecture, already presented in [STR-D31], should be extended as in the following:

- The **SPDF** module:
  - Shall be able to process the inter-carrier AF and SPDF user data flow transport (and OSS path provisioning) requests, with resource constraints
    - The ETSI TISPAN RACS architecture supports data flows declared as 5-uple, and routed transport scenarios, with NAT support. G-RACS architecture should support many flow declarations and transports, at different layers (e.g. pseudo-wires). In a inter-carrier scenario, 5-uple for user data flow definition could not be applicable, due to possible overlapping of address plans
    - The inter-domain resource constraints supported could be in terms of peak and guaranteed bandwidth, maximum delay and jitter, media type, priority, security
  - Shall be able to admit or reject inter-domain requests by its Policy Decision Function, based on service policies and constraints defined for inter-carrier connections, for example in terms of:
    - Inter-Carrier connectivity (inclusions/exclusions), cost...
    - Inter-Carrier resources (bandwidth, delay, jitter, class of service, priority...)
    - Inter-Carrier path selection (e.g. shared path selection, or dedicated path setup)
  - Shall collect information about the provider's interconnections and build an inter-carrier topology model. For example:
    - Each remote provider with an interconnection agreement with the local provider could be modeled as a virtual router
    - Each provider interconnection could be modeled as a link
    - Each link could have an associated cost, based on service types and constraints, and various inter-carrier policies/agreements
    - Default routes for providers not directly interconnected on the basis of an agreement could be provided
  - Should implement an algorithm to identify the next provider to contact, based e.g. on the destination provider, the service type and constraints requested

- Shall trigger the admission control, resource reservation and path computation procedures, both in the local domains (addressing the directly managed multi-layer x-RACF/PCE modules, as already stated in [STR-D31]), and in the next provider to contact (addressing the interconnected SPDF modules), in order to satisfy an inter-carrier service request from both an AF, or another provider
- If e2e GMPLS path setup and modification is used (see the following sub-section), the upper **x-RACF/PCE** modules, directly connected to the SPDF:
  - should generate and collect the inter-carrier path computation keys, in order to mask the local ER sequences
  - should perform the translation between the inter-carrier path computation keys and local ER sequences, when triggered by an inter-carrier path setup request from the GMPLS control plane
- If e2e GMPLS path setup and modification is used (see the following sub-section), the lower **x-RACF/PCE** modules, distributed in LSRs:
  - should collect inter-domain routing and TE information (e.g. by a proper path vector routing protocol, as BGP, extended with TE information)
  - should request an inter-carrier path computation key translation to its reference upper x-RACF/PCE module, when triggered in the E-NNI
- If hop-by-hop GMPLS path setup and modification is used (see the following sub-section), x-RACF/PCE modules will operate as described in [STRD3.1].

### 3.6.2 GMPLS Control Plane considerations

In the present section, the following G-RACS (x-RACF/PCE) and GMPLS Control Plane interaction options, for path setup and modification, are presented:

- **End-to-end path setup and modification**, wherein only the G-RACS in the source domain/provider triggers the GMPLS Control Plane, and the GMPLS signaling is spread towards the destination domain, through the E-NNI interfaces. Such scenario requires a proper GMPLS Control Plane E-NNI interaction, in terms of:
  - Multi-vendor and multi-technology interworking
  - Security and confidentiality
  - Routing and TE information sharing and summarization
- **Hop-by-hop path setup and modification**, wherein each G-RACS involved in the e2e inter-carrier path setup, or modification, triggers the GMPLS Control Plane for an intra-domain/carrier path setup. In such scenario:

- There is no need of sharing inter-carrier TE information over the E-NNI
- Intra-domain/carrier path setups/modifications could be performed in parallel, for a faster operation
- X-RACF/PCE modules need to enforce IP data flows into the pseudo-wires
- Different models can be applied for the shared link between border routers

### 3.7 PCEP extensions for GMPLS networks

This section covers the PCEP extensions to take into account GMPLS networks. It discusses the requirements, proposed solutions and current issues. It will end with an update on the current standardization activities.

As described in STRONGEST deliverable D3.1, PCEP required extensions can be summarized as follows.

When requesting a path calculation the client can provide and require more information than in packet-based networks. In GMPLS networks this is possible and sometimes necessary due to policies for the client to use explicit routing. The client should be able to indicate in a route calculation request which granularity in terms of node, link or label the client wants to apply to the request. This involves whether the path should be composed of only high level network elements, such as just nodes and/or links, or include also labels to apply. In tandem with requesting label information a client can also put restrictions on labels to be allocated, so label restriction are added as routing constraints. This includes the wavelength continuity constraint that is typical of Wavelength Switched Optical Networks (WSON). Label restrictions for the endpoint are also to be considered as separate restrictions.

Depending on the switching technology used by the control plane the bandwidth requirements are expressed in a more detailed and technology specific way. This different representation is needed to correctly identify the resources to be used, in such a way that the client can express the detailed switching technology traffic specification (covering at least Ethernet, packet, SDH/Sonet/ OTN/DWDM specifications) in the route calculation request, and also specify what are the constraints (inclusions, exclusions) on the switching technology to be considered during path computation.

Finally some switching technology specific constraints like optical signal performance have to be considered in the request. This is an indication exchanged between the client and the server on the path calculation properties. This additional information in the request is mirrored in the response.

The Explicit label control implies that the explicit label or label ranges can be present in the calculated path. For multi-layer and multi-region routes this can take the form of an explicit indication of the layers boundaries: that is, in a MRN, the nodes that are at the boundary of a region change, so the signaling controller may easily and seamlessly trigger the related procedures such as the establishment of a forwarding adjacency (FA) to the remote boundary node.

Finally the path calculation response can indicate which technology specific routing calculation (for WSON the optical quality checks) were applied and/or verified.

### **3.7.1 Proposed PCEP extensions and standardization update**

The required PCEP extensions do not cover, yet, all the requirements. Within the STRONGEST project, and in turn within IETF, activities are ongoing to fill this gap.

The IETF WG responsible for PCEP standardization is relying on the signaling WG for the WSON signaling aspects. So the more generic requirements (GMPLS, Multi-domain and multi-layer) were the main focus of the standardization work.

The required extensions are covered in the following IETF documents (version at the time of writing):

1. draft-ietf-pce-gmpls-pcep-extensions (version01)  
"This memo provides extensions for the Path Computation Element communication Protocol (PCEP) for the support of GMPLS control plane."
2. draft-ietf-pce-inter-layer-ext(version04)  
"MPLS and GMPLS networks may be constructed from layered service networks. It is advantageous for overall network efficiency to provide end-to-end traffic engineering across multiple network layers through a process called inter-layer traffic engineering. PCE is a candidate solution for such requirements. The PCE communication Protocol (PCEP) is designed as a communication protocol between Path Computation Clients (PCCs) and PCEs. This document presents PCEP extensions for inter-layer traffic engineering."
3. draft-gonzalezdedios-pce-reservation-state (version 00)  
"This document proposes an extension to the PCEP protocol to allow the PCC to request the PCE to block or reserve the resources computed in a path request of a TE LSP for subsequent requests for a certain time."
4. draft-zhang-pcep-hierarchy-extensions (version 00)  
" The hierarchical Path Computation Element (PCE) architecture defined in [PCE-HIERARCHY-FWK] allows the optimum sequence of domains to be selected, and the optimum end-to-end path to be derived through the use of a hierarchical relationship between domains. This document defines the Path Computation Element Protocol (PCEP) extensions for the purpose of implementing hierarchical PCE procedures which are described in [PCE-HIERARCHY-FWK]."
5. draft-zhang-ccamp-gmpls-h-lsp-mln (version 02)  
" This specification describes the hierarchy LSP creation models in the Multi-Region and Multi-Layer Networks (MRN/MLN), and provides the extensions to the existing protocol mechanisms described in [RFC4206], [RFC4206bis] and [MLN-EXT] to create the hierarchy LSP through multiple layer networks."



In order to support the indication of which information should be returned by the path calculation, the PCEP RP object is extended with a new flag indicating if node, link or label should be returned in the response, this flag is an indication and it can be overridden by the PCE.

The detailed traffic specification requires the extensions of 2 PCEP objects: BANDWIDTH and LOAD-BALANCING. The BANDWIDTH is indicating the total bandwidth requested (or existing in case of re- optimization) by the request and the LOAD-BALANCING optionally indicate if several diverse path can be returned by indicating the number of path and the minimum bandwidth allowed on a path.

Following strictly the definition of the object, it is not allowed to change their length, for example by adding TLVs. Those restrictions require the definition of new PCEP objects, namely GENERALIZED-BANDWIDTH and GENERALIZED-LOAD-BALANCING.

Those new objects have the same semantic as their RFC5440 counterparts, but allow the following:

- It is possible to have a different bandwidth for forward and reversion direction (which would be required by a control plane supporting RFC 5467 (RSVP asymmetric BW))
- All the currently definition RSVP-TE traffic specification can be represented
- Variable-length traffic specification is supported
- TLVs can be added for future extensions

In PCEP the endpoints involved in a Path Computation (e.g. source and destination nodes) are indicated in the ENDPOINT object. The existing types defined in RFC 5440 and RFC 6006(P2MP) do only allow to have both endpoints addressed either by IPv4 or IPv6 addresses. No other information is provided for the endpoint. In order to fully support all GMPLS endpoints then ENDPOINT object should be able to accept:

- Different type of ingress/egress endpoint
- Unnumbered endpoints
- Existing endpoints (IPv4 /IPv6)
- P2MP endpoint specification, including P2MP source and leaves
- Endpoint-specific restrictions like label, label range or suggested label

This is realized in the PCEP protocol by using a new object type for the ENDPOINTS object, which makes used of the following TLVs:

- Endpoints : IPV4, IPV6, unnumbered
- Restrictions : label request, label , label set and suggested label

The label request is used to scope the label information to the correct switching layer.

The body of the object consist of a set TLV following the following grammar (for end-to-end request, the point to multipoint is detailed in the draft-ietf-pce-gmpls-pcep-extensions document).

```
<generalized-endpoint-tlvs> ::=
    <endpoint>[<endpoint-restrictions>]
    <endpoint>[<endpoint-restrictions>]

<endpoint> ::=
    <IPV4-ADDRESS> | <IPV6-ADDRESS> | <UNNUMBERED-ENDPOINT>

<endpoint-restrictions> ::=
    <LABEL-REQUEST>
    <label-restriction>[<endpoint-restrictions>]

<label-restriction> ::= ((<LABEL><UPSTREAM-LABEL>) | <LABEL-
SET> | <SUGGESTED-LABEL-SET>)[<label-restriction>]
```

The first endpoint and optional endpoint-restriction is the ingress endpoint and the second endpoint is the egress endpoint (followed by restrictions)

The restriction concerning the end-to-end request (not just the endpoints) are present as PCEP object with TLVs, reusing the LABEL-REQUEST and LABEL-SET TLV from the ENDPOINT object.

The same objects are used in the response to indicate the label range to be used for signaling.

These extensions have been successfully implemented and verified. The related innovative results and considerations have been published in the proceedings of ECOC 2010 [Munoz10].

Multi-layer aspects are covered in the draft-ietf-pce-inter-layer-ext, which defines new objects (INTER-LAYER, SWITCH-LAYER and REQ-ADAPT) respectively indicating if multi-layer or mono-layer path computation is desired or not, which switching layer to consider or exclude, and finally what is the requested adaptation on the endpoint in case the endpoints are in a different layer.

The METRIC object is also extended to allow consideration of the number of adaptation as route calculation criteria.

The layer boundaries are represented in the ERO, as signaled in RSVP. To indicate to a non-IGP capable node the layer boundaries a new Sub-TLV of the ERO object is defined : the SERVER\_LAYER\_INFO sub-TLV.

This Sub-TLV contains the label-request and traffic specification at the switching layer border node. This allows any node in the path to detect when to create server layer LSP and what is the next hop for a given switching layer.

Finally the document draft-gonzalezdedios-pce-reservation-state addresses the problem of multi-domain path computation and resource reservation. Without explicit resource reservation contention condition on resources are very likely to occur in multi-domain networks, as the window between the start of computation and the IGP update can be large depending on the type of network and resources managed.

All the above mentioned documents have an established and solid base, but are still in an early stage of the standardization process. Some of them are already WG documents, while others still need to be adopted as official drafts.

The current standardization activities need to continue in order to provide a consistent picture (at present several drafts overlap with different angles on similar problems) and, finally, the adoption as a Standard by the IETF.

### **3.7.2 PCEP Extensions for Temporary Reservation of Path Resources**

According to RFC4655, a PCE can be either stateful or stateless. In the former case, there is a strict synchronization between the PCE and not only the network states (in terms of topology and resource information), but also the set of computed paths and reserved resources in use in the network.

In other words, the stateful PCE utilizes information from the TED as well as information about existing LSPs in the network when processing new requests. However, the maintenance and synchronization of a stateful database can be non-trivial, not only because it should verify the actual establishment of the computed paths, but also because it might not be the unique element to compute paths. Moreover, maintaining such a stateful database does not seem to be a function of the PCE, but rather of an NMS.

On the other hand, a stateless PCE does not keep track of any computed path, and each set of request(s) is processed independently of each other. With a stateless PCE, there is a 'potential window of TED inaccuracy', where a stateless PCEs may compute paths based on current TED information, which could be out of sync with actual or potential network state changes given by other recent PCE-computed paths.

For example, some sources for this potential TED inaccuracy are:

- Control Plane link latencies, increasing: a) the time required for a PCC to obtain the paths after a successful computation, requiring several Round-Trip-Times (RTT) as per TCP; b) the set up delay and c) the time it takes for the PCE to update the local TED given IGP update times.

- IGP (i.e. OSPF-TE) may operate with timers for LSA updates, to avoid excessive control plane overhead.
- Concurrent requests that arrive during the time window, between a response is sent and the LSP is set up and the topology changes flooded. Even for very fast networks with low latency, there may be 'batched' requests: several path computation requests within a PCReq message or, in dynamic restoration without pre-planning, several LSPs that need to be rerouted avoiding a failed link.
- Local PCE contention, where the PCE needs to concurrently serve path computation requests and update the LSA (e.g. parsing OSPF-TE LSA updates). A PCE implementation may need to find a trade-off, when synchronizing access to the local TED: favor OSPF-TE parsing which means that some path computations are slightly delayed to allow an 'update' to be processed, or give strict priority to computation requests.

In consequence, a stateless PCE may assign the same (or a subset of the same) resources to several requests, which may result in contention and degraded network performance. The effects are detected late, typically during path signaling, causing path blocking and excessive crank-backs and retries.

In this light, a limited form of statefulness is useful to improve PCE functionality in situations in which the local TED might not be up to date, or in the case of concurrent requests where most of the LSPs are computed before the end of the set-up of the LSPs when the TED is updated. The PCE can retain some context from the resources assigned to Path Requests during a certain period of time, so that it avoids suggesting the use of the same resources for subsequent TE LSPs.

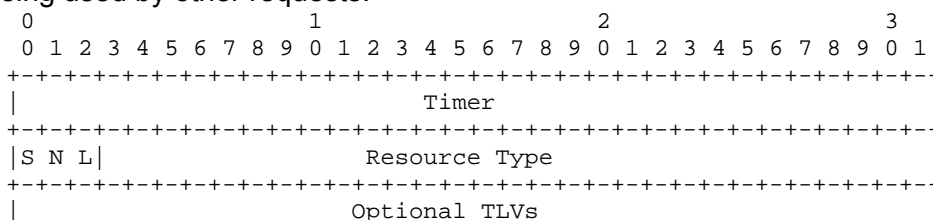
We propose here for the first time a method, developed inside STRONGEST WP3, that is an extension to the PCEP protocol to allow the PCC to request the PCE to block or reserve the resources computed in a path request of a TE LSP for subsequent requests for a certain time.

### 3.7.2.1 PCEP Proposed Extensions

We propose new Objects and TLVs to support the reservations.

#### RESERVATION object

The RESERVATION object indicates the intention of the PCC to set up the requested path and request the PCE to reserve the resources of the computed path to avoid being used by other requests.

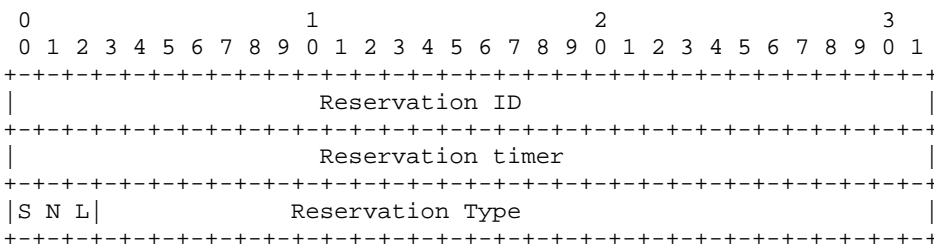


...  
 +-----+

Timer is the value in ms of the time that the resources should be blocked. Resource Type indicates the type of resource to be reserved (Bandwidth, PSC, L2SC, ..., Time Slot, Sonet/SDH TDM, Tributary Slot (G709 OTN ODU-k TDM), Wavelength (G709 OTN OCh or WSON LSC).

RESERVATION\_CONF Object

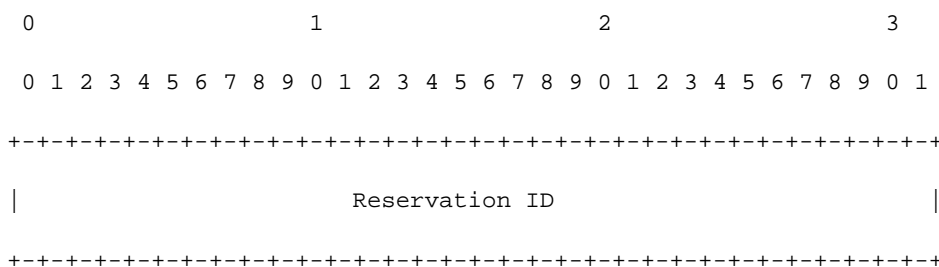
The RESERVATION\_CONF object is optional. The RESERVATION\_CONF object indicates that the PCE has reserved the resources of computed path to avoid being used by other requests. The RESERVATION\_CONF object is sent in the PCRep.



Timer is the value in ms of the time that the resources are blocked. The PCE may decide to apply a value different from the one requested by the PCC.

RESERVATION\_ID TLV

The TLV indicates the reservation ID.



**3.7.2.2 Application case: Multiple LSP Restoration**

One of the most challenging scenarios for a PCE-based architecture is the one of massive restoration. In the event of a network failure affecting a high number of LSPs (e.g. a fiber cut), a PCE could potentially receive a significant amount of restoration requests in a short period of time. One of the various challenges in this scenario is the fact that the PCE needs to sequentially perform multiple independent path computations.

In order to evaluate the potential benefits of the proposed mechanism and extensions, a simulation study has been performed. An event-based simulation tool (i.e. Omnet++ V4.1) was used to model the complete network and process the events, as it eliminates dead times and reduces significantly the simulation time. The 14-node NSFNet backbone was used as a reference topology for the simulation scenario. As a simple traffic matrix, 1x10Gbps LSP is established between every pair of nodes. For routing and wavelength assignment, a typical Shortest Path plus First Fit approach was followed. The control plane communication network was emulated by introducing delays for control plane messages between neighboring nodes. For simplicity, it was considered a similar control plane delay for messages between ROADMs, and that every node had a similar delay to the PCE, as well.

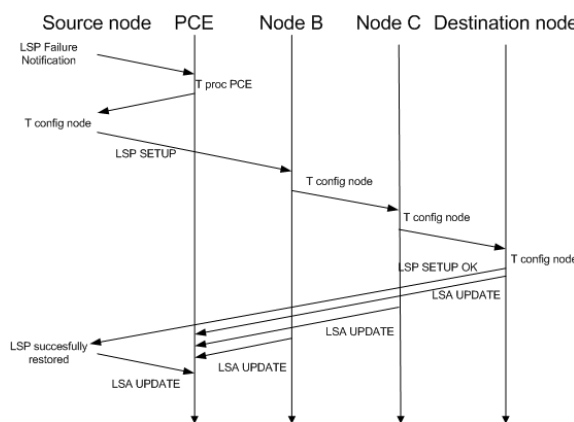
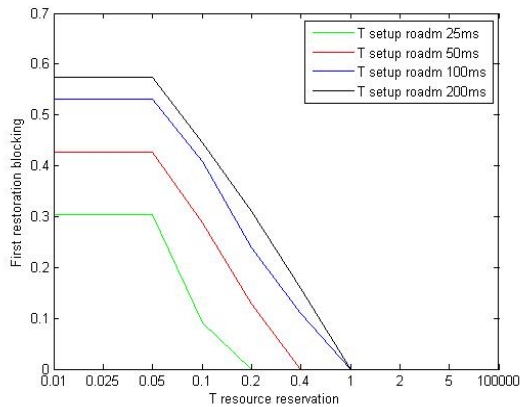


Figure 30: Control plane messages workflow

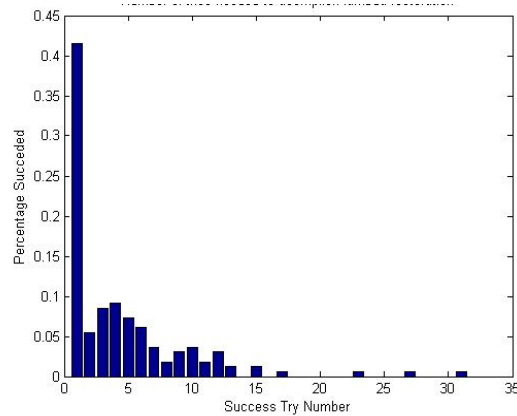
The simulation is triggered by a link cut affecting several LSPs (in our case, 27 LSPs), which could be perfectly accommodated by using alternative paths and lambdas. The results presented in the next section show the average results for the simulation of six different failures. After the link failure, the source node of each of the affected LSPs will independently trigger a Path Computation Request to the PCE demanding for an alternative path. The total time for path restoration will include the round trip delay to the PCE ( $T_{pce}=10ms$ ), the time to compute the path ( $T_{proc\_PCE}=50ms$ ), the time to configure each ROADM in the path ( $T_{setup\_roadm}$ , which varies from 10ms to 200ms) plus the time spent for the interchange and process of control messages ( $T_{msg}=1ms$ ). It must be noted that the longer the LSP, the higher the number of ROADMs, and therefore the longer the total establishment time. For the sake of simplicity, the TED is automatically updated after LSP set up confirmation.

Figure 31 shows the restoration blocking probability for different values for the resource reservation period (ranging from 0 to 1 second). It can be observed that the blocking ratio (i.e. signaling fails in the first attempt) depends on the relation between the LSP set up time and the reservation time. On one side, the higher the set up time, the higher the blocking probability. On the other side, the higher the reservation time, the lower the restoration blocking probability. It can be concluded that a value of 1 second for the resource reservation timer is enough to avoid any potential collision. It can be perceived that there is a significant amount of colliding path computation requests (around 60%), and that there are a significant number of requests with a very high number of retries. In other

words, the picture measures the number of retries needed for different connections, and illustrates the ineffectiveness of a non context aware massive recovery. Of course, the number of retries could be reduced by means of retry timers, but this will not significantly enhance the performance.



**Figure 31: Blocking Probability vs Resource Reservation Timer (without retry).**



**Figure 32: Number of attempts needed to accomplish restoration for the different connections.**

Summarizing the simulation findings, the results show that keeping context information for just one second can reduce by a 30% the average recovery time per connection, as well as a 50% of reduction in the total recovery time, when the ROADM set up time is higher than 100ms. The benefit comes from the fact that it avoids the computation of a significant amount of colliding path requests. It must be noted that these benefits would be negligible if the path set up time is low enough.

### 3.8 Future Works

In the scope of the control plane solutions for the STRONGEST mid-term scenario, a significant amount of work will be dedicated to the topic of the applicability of path computation elements in multi-layer and multi-domain networks, extending current results where appropriate and with the additional consideration of new collaborative methods or methods that can benefit from the particular PCE architecture. In this sense, the hierarchical PCE remains a key topic, with specific emphasis not only in the architecture (functional, protocol and physical ones) but also considering the topology aggregation methods. Studies will include more complex topology summarization criteria (not only full mesh of virtual intra-domain links, additional summarized parameters (comprising operators-driven ones, with relative algorithm improvements) and the overall framework for topology summarization and end-to-end path computation in inter-domain and inter-carrier scenarios.

The coupling between the PCE and the G-RACF will also be covered, focusing on the actual interfaces and concrete scenarios.

The concept of reservations in the PCE will be further studied, focusing on the multiple PCE Case where reservations need to be synchronized and multi-domain (hierarchical PCE, VSTP computation) and multi-layer reservations (multiple resource types in the same request). The use of absolute timestamps for reservations is also considered for the multi-domain reservations.

Simulation studies will be also reported to evaluate potential benefits of the PCE architecture in terms of separating path computation and path selection. This will involve comparative analysis of distributed, real-time PCE solutions against static offline path computation. The emphasis, also from the point of view of operator perspectives, will be on scalability in a realistic, large-scale network environment (thousands of nodes), where sensitivity to (topology) information disclosure is paramount, implying strong practical constraints on what can be achieved.

Studies will be reported in the context of hierarchical routing protocols: OIF E-NNI for single-carrier scenarios and the proposed hierarchical path-state vector for multi-carrier networks. For example, with reference to the latter protocol, additional studies will be provided aiming at evaluating the advantages on routing efficiency of the following options: a) advertising multiple inter-domain links as separated, rather than as a single aggregated link; b) propagating information of Maximum Required Bandwidth at LSP setup; c) propagating information of Intra-domain bottlenecks (without disclosing intra-domain link details). The HBGP-PCE architecture studied in D3.2 will be also considered in order to make comparisons with hierarchical PCE.

From the point of view of the actual PCEP protocol, STRONGEST WP3 will address refinements in the context of PCEP extensions for GMPLS, extensions for P2MP path computation, extensions for resource reservation and extensions for H-PCE. This part of the work will be carried out in tight cooperation with the IETF PCE working group, where several of the STRONGEST Proposals have been presented.

Specific studies will be also provided in the context of heterogeneous networks, e.g. WSON encompassing different bit-rates and modulation formats as well as ROADM architectural constraints.



## 4 End-to-end Services set up and Traffic Admittance

This chapter reports on the main considerations and results achieved within STRONGEST WP3 in the context of end-to-end services and traffic admittance.

The classification of end-to-end services is first reported, highlighting the business drivers and major users' requirements. Then, a mapping of end-to-end services over the MPLS technology is proposed.

This chapter then reports on the specific implementation of some OAM and control plane solutions elaborated within STRONGEST and described in the previous chapters that are here applied to guarantee the provisioning of end-to-end services. Specific implementation issues are addressed and innovative solutions are then derived and evaluated.

For example, innovative RACF-based solutions are proposed and applied to enable the possibility for applications to ask for QoS services without having to know how the request will be handled by the different network portions and network providers.

OIF E-NNI is then applied in the context of services with strict-delay constraints. In particular, a novel PCE-RC architecture is utilized together with different innovative summarization schemes, specifically designed for multi-carrier networks.

### 4.1 Service Definition

#### 4.1.1 Business drivers and requirements

Network operators are part of a value chain where networks underpin their offer to customers. Within this business context, operating and evolving networks is complicated by the existence of uncertainties in both customer demand and operator offer. Customer demand is unpredictable at all timescales and is affected by factors such as market conditions, usage patterns (when, how often, how long for) and the nature of communications (arrivals, durations, sizes and required capacity fluctuations). This spectrum of uncertainties creates a whole range of effects that must be managed.

Pressures to increase margins by reducing operating costs have led operators to formulate a network technology strategy whereby cost reduction is to be achieved by replacing circuit networks with packet networks, thereby benefitting from increased efficiency gain and simplified management. Over the past few years, typical operator strategy for networks has been to deploy a global open network ("multi-service platform") able to deliver any type of service to customers. Over time, it has emerged that migrating from old to new technologies and platforms requires more care than anticipated, especially in terms of the trade-off between service delivery and operational efficiency.

Extra functionality is required to "harmonise" packet networks with the strict performance requirements typical of voice and private network services. Bridging this gap has been approached via two opposite but converging directions.

- One approach (Table 5) involves assigning additional resource control functionality to packet networks, in order to improve their supported performance characteristics to the point of being able to support applications with less tolerant constraints; this leads to the emergence of a new type of network that can deliver a well-defined Service Level Agreement (SLA) and take calculated risks in doing so; these networks deliver predictable performance with very high probability, while maintaining high efficiency; typical performance characteristics include bounded rate, bounded delay and jitter, bounded loss rate, in-order delivery and pre-established routes
- A second direction (Table 6) involves assigning additional resource control functionality to the application, in order to relax strict performance requirements, re-designing (or emulating) the original application to run adequately over a native packet rather than circuit network; this new class of applications have relaxed the intolerant constraints, making them more suitable for packet networks

**Table 5: Typical performance characteristics for different network service types**

Type of service	Service invocation timescales	Supported applications	Typical performance characteristics	Example of technology	Resource control
<b>Type #1</b>	Both slow and fast set-up	<ul style="list-style-type: none"> <li>• Private circuits</li> <li>• PSTN Voice</li> </ul>	<ul style="list-style-type: none"> <li>• Fixed rate</li> <li>• Constant delay</li> <li>• Ordered delivery</li> <li>• Pre-existing route</li> </ul>	<ul style="list-style-type: none"> <li>• SDH</li> <li>• PONs</li> <li>• WDM</li> <li>• PSTN</li> </ul>	Connection-oriented circuit-switching, based on provisioned paths + possibly simple resource control (e.g. CAC)
<b>Type #2</b>	No set-up	<ul style="list-style-type: none"> <li>• Internet traffic</li> </ul>	<ul style="list-style-type: none"> <li>• Variable rate</li> <li>• Variable delay</li> <li>• Out-of-order delivery</li> <li>• No pre-existing route</li> </ul>	<ul style="list-style-type: none"> <li>• IP</li> <li>• MPLS</li> <li>• Ethernet</li> </ul>	Connectionless packet-switching with no resource control
<b>Type #3</b>	Slow set-up	<ul style="list-style-type: none"> <li>• VPNs</li> <li>• Circuit emulation</li> <li>• ATM CBR/VBR/UBR</li> </ul>	<ul style="list-style-type: none"> <li>• Guaranteed min rate</li> <li>• Guaranteed max delay</li> <li>• Ordered delivery</li> <li>• Pre-existing route</li> </ul>	<ul style="list-style-type: none"> <li>• IP</li> <li>• MPLS(-TE)</li> <li>• Ethernet (PBB-TE)</li> <li>• ATM</li> </ul>	Connection-oriented packet-switching, based on provisioned paths + simple RC (e.g. differentiation, simple CAC)
	Fast set-up	<ul style="list-style-type: none"> <li>• Voice</li> <li>• Media</li> </ul>	<ul style="list-style-type: none"> <li>• Guaranteed min rate</li> <li>• Guaranteed max delay</li> <li>• Out-of-order delivery</li> <li>• No pre-existing route</li> </ul>		Connectionless or connection-oriented packet-switching, based on un-provisioned paths + non-trivial resource control (CAC, path

					selection)
--	--	--	--	--	------------

**Table 6: Typical performance requirements for different application types**

Type of application	Description	Examples	Typical performance requirements
<b>Class #1 INTERACTIVE</b>	Apps interested in a fixed rate with tight requirements on latency and loss. Typically intolerant to any loss and high delay.	Voice, interactive media, circuit emulation, ATM CBR/VBR	Guaranteed min rate, guaranteed max delay, very low jitter, no loss
<b>Class #2 GUARANTEED</b>	Apps interested in a small range of rates (typically min-max range) with some (possibly complex) requirements on latency and loss. Tolerant to some small amounts of delay or loss	Media streaming, VPNs, ATM UBR	Low loss, high rate within traffic envelope
<b>Class #3 BEST EFFORT</b>	Apps interested in the shortest time to completion, but that can cope with any rate that achieves that (TCP-like rate adaptation, Internet traffic, web, email, file transfer, telnet)	Web browsing	Minimum time-to-completion

### 4.1.2 Service Reference Model

The future Internet is expected to be more agile, scalable, secure and reliable. Meanwhile, we have witnessed the unprecedented development and growth of new applications and services in recent years ranging from location-based services, social networking, cloud computing and peer-to-peer applications. High-speed broadband penetration and the ongoing growth of Internet traffic among residential and business customers have already placed a huge bandwidth demand on the underlying telecommunications infrastructure. Traffic patterns have been propelled from voice- and text-based services to user-generated interactive video services.

Future real-time video communications and dynamic video content is expected to ultimately test the network more than pre-recorded video content. Such rapidly emerging applications with different requirements and implications for the future Internet design pose a significant set of problems and challenges.

There is no generally accepted standard for Quality of Service (QoS) classes and parameters, so the implementation of network QoS will be operator-dependent, following (e.g.) the methodology in [NIK]. Essentially, parameters for QoS can be defined with respect to packet forwarding on layers 3 and 4:

- Capacity and throughput
- End-to-end packet loss
- Packet delay / transfer delay and jitter (transfer delay is related to the transmission of the application information, whereas packet delay means the delay for the transmission of a packet)

QoS differentiation refers to the fact that services can be distinguished by the different requirements they have in order to have successful operation. Typically services are characterised by specific values for minimum/maximum bandwidth, availability, security, frame delay, jitter, loss, error rate, priority and buffering. For example, a Voice over IP service requires a minimum bandwidth and expects a strict range of frame delay, jitter and loss, otherwise the service becomes unusable. The intention of QoS specifications is to use network mechanisms, such as IntServ or, DiffServ in order to deliver predictable service levels such that service requirements can be fulfilled. The first mechanism is based on resource reservation and implies that states are maintained in every intermediate node of the flow and the second uses a per-packet stateless approach based on priority bits (DiffServ code points) marked in IP packets.

Another basic subset of QoS parameters is related to Quality of Resilience (QoR):

- Service availability
- Recovery time
- Maximum outage time

Video and audio are always-on services that cannot accept unpredictable network recovery timeouts and best-effort QoS implementations. Unpredictable behaviour can result in a user perception of poor video quality and eventually increase customer churn. Operators require a highly available infrastructure foundation that will enable them to build their brand equity through the flawless achievement of the required service level guarantees. Such a foundation should be based on products designed to exceed the most stringent reliability demands of service providers, with hardware and software architectures designed for maximum uptime. It should provide millisecond-level service recovery or restoration mechanisms at the path, link, node and network levels, for the infrastructure control, forwarding and management planes. Security is a further key element in ensuring service continuity or non-stop services, providing the operator with mechanisms that guarantee an assured user experience while containing denial of service and theft of service attacks.

At the network level, typical SLA committed figures [SPR] are:

- Round Trip Delay < 100ms
- Packet Loss 0.1%
- Jitter < 2ms

From a user perspective, the guideline application requirements in Table 7 have been suggested [NOB]. Service names are largely descriptive, based simply on differences in the parameter values shown. In many cases, both fixed and mobile applications might be offered over the STRONGEST transport infrastructure, perhaps with similar QoS

requirements. However, bandwidth requirements are often lower for mobile applications (web browsing, videoconference, gaming, etc.) than for fixed applications due to limitations in wireless access bandwidth, terminal screen size and resolution. More generally, the widespread coverage of broadband access networks may well favour increased nomadism, introducing more dynamic and unpredictable traffic.

In the next phase of the STRONGEST service definition, WP3 will seek to coordinate traffic volumes and service mixes from the wider project to create a consolidated view of future service needs and performance requirements.

**Table 7: User Application Requirements**

Type of service	peak down (Mbps)	peak up (Mbps)	mean down (Mbps)	mean up (Mbps)	Max delay (ms)	Max jitter (ms)	Packet loss	Blocking prob.
Broadband type 0 Mobility	0.384	0.128	0.033	0.006	*	*	*	*
Broadband type 1	1	0.3	0.086	0.082	*	*	*	*
Broadband type 2	2	0.512	0.1468	0.14	*	*	*	*
Broadband type 3 Mobility	3.2	0.384	0.1101	0.10496	*	*	*	*
Broadband type 4	10	0.8	0.2293	0.21867	*	*	*	*
Broadband type 5	10	10	2.733	2	*	*	*	*
Broadband type 6	20	0.8	0.27	0.24	*	*	*	*
Broadband type 7	50	6	1.72	1.64	*	*	*	*
Broadband type 8	50	50	11	11	*	*	*	*
Broadband type 9	100	10	2.87	2.733	*	*	*	*
Broadband type 10	100	100	23	23	*	*	*	*
Video Broadcast 0 (Mobility TV)	0.384	0	0.256	0	< 2000	< 40	< 3 E-3	< 0.1%
Video Broadcast 1 (SDTV mpeg2)	6	0	6	0	< 2000	< 40	< 3 E-3	< 0.1%
Video Broadcast 2 (SDTV mpeg4)	3	0	3	0	< 2000	< 40	< 3 E-3	< 0.1%
Video Broadcast 3 (SDTV mpeg2)	20	0	20	0	< 2000	< 40	< 3 E-3	< 0.1%
Video Broadcast 4 (SDTV mpeg4)	10	0	10	0	< 2000	< 40	< 3 E-3	< 0.1%

VoIP	0.008	0.008	0.008	0.008	< 70	< 20	< 3 E-3	< 0.1%
Grid computing	0.512	0.128	0.0358	0.009	< 200	< 50	< 1 E-4	< 0.1%
Gaming 1 Mobility	0.04	0.04	0.03	0.03	< 50	< 10	< 5 E-2	< 0.1%
Gaming 2	0.25	0.25	0.2	0.2	< 50	< 10	< 5 E-2	< 0.1%
Videoconference 1 Mobility	0.03	0.03	0.026	0.026	< 100	< 10	< 3 E-3	< 0.1%
Videoconference 2	0.128	0.128	0.1	0.1	< 100	< 10	< 3 E-3	< 0.1%
Videoconference 3	3	3	2.2	2.2	< 100	< 10	< 3 E-3	< 0.1%
Telemedicine 1	2	10	1.73	8.67	< 40	< 10	< 1 E-4	< 0.1%
Telemedicine 2	1	4	0.2	0.8	< 200	< 50	< 5 E-3	< 1%
Domotics, E- Business	0.064	0.064	0.042	0.042	< 200	< 50	< 5 E-3	< 1%
Business Data 1	10	10	2.6	2.6	< 50	< 5	< 5 E-3	< 1%
Business Data 2	100	100	26.4	26.4	< 50	< 5	< 5 E-3	< 1%
Business Data 3	1000	1000	264	264	< 50	< 5	< 5 E-3	< 1%
Business Data 4	10000	10000	2640	2640	< 50	< 5	< 5 E-3	< 1%
SAN 1 (Back- up/Restore)	400	400	324	324	< 500	< 50	< 5 E-2	< 1%
SAN 2 (Storage On Demand)	1000	1000	810	810	< 10	< 1	< 5 E-3	< 1%
SAN 3 (Asynchronous Mirroring)	400	400	324	324	< 200	< 50	< 5 E-2	< 1%
SAN 4 (Synchronous Mirroring)	2000	2000	1620	1620	< 10	< 1	< 5 E-3	< 1%

### 4.1.3 Implications for STRONGEST

The STRONGEST challenge is to effectively bridge the gap between packet networks and strict performance requirements, with additional complications and constraints due to operating in a multi domain/region/vendor environment.

From an operator perspective, it is realistic to consider services which may require strict delay constraints. In particular, financial services (e.g. stock exchanges, financial information providers) do have very stringent delay constraints, demanding connectivity

services of guaranteed and minimum latency. This may impact, for example, on the metric to be used within the multi-domain control plane or the objective function to be considered in the PCE Architecture.

Furthermore, operators normally prefer not to disclose bandwidth or internal topology information [BBF], with obvious implications for schemes like OIF ENNI routing which includes the possibility to advertise reservable bandwidth information via the routing algorithm.

The next phase of STRONGEST service definition will require coordination of traffic volumes and service mixes across the wider project to help create a consolidated view of future service needs and performance requirements.

## 4.2 Mapping of End-to-End Services

In the STRONGEST reference scenarios [D3.1], there are typically two network layers involved in an end-to-end connection: a packet transport layer like MPLS, and a circuit switched layer like WSON, or more general, layers of different switching capabilities. It is assumed that all layers are coordinated by a unified GMPLS control plane. Traffic forwarding is done along Label Switched Paths (LSPs) that are set-up and torn-down by the control plane across all involved layers. Despite the intentional generalization it must not be ignored that the switching capabilities of the different layers are inherently incompatible to each other. This results in the rule that LSPs start and terminate always in one and the same layer. Moreover, an LSP in one layer can use lower layers only in the expression of tunnels, i.e. a separate LSP in the lower layer is being set-up to host one or more LSPs in the higher layer. The lower layer LSP acts virtually as an additional link in the higher layer network. In this section we discuss which instances are there to trigger the different LSP types, which kind of information they have, how fast they are able to respond, etc.

We introduce following classification of LSPs:

By **requestor**: In general it is said an LSP set-up is triggered by means of a User to Network Interface (UNI). The crucial question is, however, who or which entity is sitting behind the UNI. Is it really an end-user, who starts a particular application that in turn requests a connection? Let's call this a **session controlled LSP**. It is typically a precise just-in-time request, e.g. a voice channel. Or is it rather a (sub-) network administrator, who represents the joined demand of a group of end-users? In this case the requestor has only statistical knowledge of when and how much to request. Statistical data takes some time for acquisition and cannot change immediately. We call this an **administrator controlled LSP**. It is typically set-up pro-actively, with large fluctuation reserves included, and infrequent adaptations, e.g. a remote campus interconnection, VPN.

By **persistence**: A similar distinction as the above, but not necessarily identical, is the question, when a particular LSP is requested. Is it requested immediately just before a particular transmission is due – an **on demand LSP**? It is assumed that an on demand LSP is also released immediately after the transmission has finished. Or is it a **pre-established LSP** that is set-up pro-actively, long before real transmission occurs? It is

assumed that pre-established LSPs are never released or at least live for weeks, months, or years.

By **resource allocation**: An LSP in general is a forwarding topic (addressing). However, in the case of a particular transmission, the actually available resources along the path could be not sufficient to fulfill the request. Depending on whether or not we associate resources with the LSP, we distinguish **LSPs with reservation** and **LSPs without reservation**. Only LSPs with reservation can give transmission guarantees. At the other hand reservations can be sold only once. The instantiation of a reserved LSP ties resources that are not anymore available to further LSPs. Circuit switched LSPs mandatorily imply reservation. An allocated circuit cannot be used by any other LSP at the same time. Packet switching technologies could be both, reserved and unreserved (MPLS-TP vs. MPLS).

The combinations of the categories have remarkable features and prominent service representatives. Please note, the example services and technologies are not necessarily really GMPLS controlled. However, they fulfill the same business case as a corresponding GMPLS solution could do.

**Pre-established LSP with reservation**: Example: Digital line service, leased line, VPN. Technologies: MPLS-TP, SDH/OTH, WDM. Only dedicated connections, limited network coverage. Full network coverage would require full mesh of LSPs. Full mesh is impossible, due to the bandwidth explosion of the reservations in large network domains. Pure statistical multiplexing, low load on narrow LSPs. But, bit rate capacity can be guaranteed.

**Pre-established LSP without reservation**: Example: MPLS network. Full network coverage by full mesh of LSPs between all nodes of a particular network domain, e.g. an operator core network. Statistical multiplexing between competing LSPs. No transmission guarantee, best effort. Not applicable for circuit switched layers due to the inherent resource allocation in circuit switching. Even though inherently a best effort technology, the packet loss ratio advertizing of chapter 2.1 could be applicable and useful for this kind of LSP.

**On-demand LSP with reservation**: Example: Bandwidth on demand service (classical ISDN telephony, but also Video conferencing). Full network coverage due to the "on-demand" provisioning. Once established, LSPs have guaranteed bit rate capacity. But, admission blocking could result in pure user acceptance. The reservation process is a network wide statefull transaction with pure scalability (network wide bottleneck). Admission blocking ratio and duration of an LSP set-up cycle are key success factor for user acceptance.

**On-demand LSP without reservation**: No explicit LSP services known yet, but TCP 3-way handshake could serve as a performance model. No bit rate guarantee. But also no admission blocking. LSP set-up is a stateless transaction, independent of other LSPs, decentralized implementation is possible. The set-up cycle is in the range of a single round trip time (RTT). Up to this point the on-demand LSP without reservation offers no additional benefit over a pure IP solution. But, the path packet loss ratio signaling of chapter 2.1 could be applicable to this kind of service.



Despite the combination of persistence and resource allocation we also investigate the combination of requestor and persistence:

A session controlled LSP should be “on-demand” at the highest network layer. However, all lower layer LSPs, that are hosting the on-demand LSP, are typically pre-established. (e.g. ISDN calls: The 64kbps channels are switched on-demand. The hosting PDH/SDH/SONET trunks are pre-established.) If the session controlled LSP is combined with a reservation (the only useful application at a first glance), than the hosting lower layer LSPs must be reserved, too, which raises the full mesh problem of pre-established LSPs with reservation. The workaround would be multi hop routing in the higher layer (cf. international telephone exchange). This would relax the need for a full mesh at lower layers. There is some academic research to do lower layer LSP set-up also on-demand, however, up to now there is no practical proof whether or not this attempt is competitive, robust, and scalable.

Administrator controlled LSPs are most likely pre-established. Their capacity is requested based on statistically collected demand values plus some more or less precisely calculated fluctuation overhead. The statistical data could be slowly varying, e.g. due to a general trend, or daily or weekly periods. This opens the opportunity for an automatic “on-demand” capacity adaptation up to “on-demand” allocation of additional links. The Dynamical Optical Bypass activity in [D3.1] is an example. Nevertheless it must not be ignored that the “on-demand” allocation of resources requires sufficient (idle running) bandwidth reserves in the hosting lower network layers. Similar proposals should proof, to which extend the “on-demand” capacity adaptation is better than the immediate allocation of all available resources, no matter, if they are needed now or later.

The following tables summarize the classification

**Table 8: Classification of LSPs by requestor**

<b>Session controlled LSP</b>	<b>Administrator controlled LSP</b>
Precise bandwidth request	Bandwidth forecast based on statistical data, includes idle running reserve to cope with actual traffic fluctuation
Bandwidth deficit is out of scope (application problem, not a network issue)	Bandwidth deficit is quality parameter (fluctuation reserve too small)

**Table 9: Classification of LSPs by duration**

<b>On-demand LSP</b>	<b>Pre-established LSP</b>
Just in time set-up and tear-down	In advance allocation, infrequent adaptations
Admission blocking ratio is part of the service quality	Admission blocking is out of scope (investments, not a network control issue)

**Table 10: Classification of LSPs by resource allocation**

<b>reserved LSP</b>	<b>unreserved LSP</b>
transmission guarantee	Best effort but applicable to prioritization, resource or quality advertizing
Admission blocking	Non blocking
Circuit switching, traffic engineered packet switching	Packet switching only (packet, frame, burst, macro frame, etc.)

**Table 11: combination of the categories resource allocation and duration**

	<b>Reserved LSP</b>	<b>Unreserved LSP</b>
<b>Pre-established</b>	<ul style="list-style-type: none"> <li>• e.g. leased line</li> <li>• bit rate is guaranteed</li> <li>• low statistical multiplexing gain (low load on many narrow links)</li> <li>• restricted coverage (dedicated point-to-point connections)</li> <li>• full mesh impossible (in non-trivial network)</li> </ul>	<ul style="list-style-type: none"> <li>• e.g. MPLS LSP</li> <li>• no guarantee, best effort</li> <li>• statistical multiplexing gain with other LSPs on same resource</li> <li>• full network coverage possible, potentially full mesh</li> <li>• prioritization and quality advertizing applicable</li> </ul>
<b>On-demand</b>	<ul style="list-style-type: none"> <li>• bandwidth on demand</li> <li>• cf. ISDN network</li> <li>• bit rate guaranteed</li> <li>• risk of admission blocking</li> <li>• set-up is a statefull transaction (network wide bottleneck), pure scalability</li> </ul>	<ul style="list-style-type: none"> <li>• ?</li> <li>• no guarantee</li> <li>• no admission blocking</li> <li>• stateless transaction, good scalability</li> <li>• prioritization applicable</li> <li>• quality advertizing possible at set-up (routing) and during LSP lifetime</li> </ul>

### **4.3 Inter-Domain end-to-end QoS and OAM signaling over heterogeneous domains**

This section investigates how legacy, SoA and standardized protocols can be adopted in STRONGEST's proposed architecture. It discusses particularly the work and signaling protocols proposed by IETF but also takes into consideration the ITU and OIF requirements. It also proposes how the NSIS suite of protocols may be utilized with the development of an NSLP application for OAM to support related mechanisms and functionality.

#### **4.3.1 Signaling Considerations and SoA for Inter-Domain QoS and OAM functionality communication**

Signaling across administrative domains is commonly associated with QoS based on the Resource Reservation Protocol (RSVP). This, however, is one of many signaling applications, which include signaling for middleboxes, signaling for label distribution MPLS just to name a few. Requirements for Signaling Protocols across different network and heterogeneous environments, has been discussed in [RFC 3726]. [RFC5151] describes Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions for Inter-Domain MPLS and GMPLS Traffic Engineering and proposes protection and recovery of Inter-Domain TE LSPs and fast recovery support using MPLS-TE fast reroute. [RFC 3209] on RSVP-TE proposes a number of extensions to RSVP, for allowing the establishment of explicitly routed label switched paths using RSVP as a signaling protocol resulting is the instantiation of label-switched tunnels which can be automatically routed away from network failures, congestion, and bottlenecks.

IETF's NSIS WG has been working towards a generic IP-based signaling framework. The NSIS protocol suite [RFC 4080] consists of a two layer model with a lower

generic transport layer, termed NTLP (NSIS Transport Layer Protocol), and a higher layer containing functionalities specific to a particular signaling application called NSLP (NSIS Signaling Layer Protocol). This allows the support of signaling of different types of services including QoS, Middlebox traversal (e.g. NATs, Firewalls) etc. A concrete NTLP protocol has been developed in [RFC5971] called GIST (General Internet Signaling Transport). So far the focus within the WG has been on QoS and as a result the development of QoS-NSLP (NSIS Signaling Layer Protocol (NSLP) for Quality-of-Service Signaling) as generic model for carrying end-to-end QoS signaling in IP networks [RFC5974]. The idea is that each network along an end-to-end path should implement a corresponding QoS Model that interprets the requests and guides the appropriate behaviours of the RMF (Resource Management Function) module of a QoS-NSLP aware NSIS node in a comprehensible manner to the network ensuring the delivery of the desired QoS. QoS NSLP is similar in concept to decoupling RSVP [RFC2205] from the IntServ architecture [RFC2210].

[RFC5976] on Y.1541-QOSM: Model for Networks Using Y.1541 Quality-of-Service Classes and [RFC5977] on RMD-QOSM: The NSIS Quality-of-Service Model for Resource Management in Diffserv are interesting examples which have recently been developed within the NSIS WG. [RFC5975] describes a QSPEC Template for the Quality-of-Service NSIS Signaling Layer Protocol (NSLP).

[RFC 4377] Describes the OAM requirements for MPLS Networks. These include Detection of Label Switched Path Defects, Diagnosis of a Broken Label Switched Path, Path Characterization, Service Level Agreement Measurements, Frequency of OAM, Alarm Suppression, Aggregation and Layer Coordination, Support for OAM Interworking for Fault Notification, Error Detection and Recovery, Standard Management Interfaces, Detection of Denial of Service Attacks, Per-LSP Accounting Requirements.

[RFC 4726] Describes a framework for establishing and controlling Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineered (TE) Label Switched Paths (LSPs) in multi-domain networks. As part of the advanced functionality it discusses inter-domain OAM, how collaboration between PCEs or domain boundaries might be required in order to provide end-to-end OAM especially where topology confidentiality is strong and also raises the issue of ensuring that end-points support the various OAM functionalities. It also mentions how different signaling mechanisms may need refinement to [RFC4379] to gain full end-to-end visibility. [RFC4379] on Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures describes a simple and efficient mechanism to detect plane failures in MPLS LSPs.

### **4.3.2 Signaling Requirement Considerations for Inter-Domain Networks**

As discussed in [D2.1] with regard to the IP over SDH and static DWDM scenarios although the scalability of the control plane is good due to the use of an IP hierarchical architecture, it does not perform satisfactorily in terms of survivability assurance, QoS, support for VPNs, data plane scalability, power consumption and CAPEX. Also in the IP over WSON (independent layers scenario) although multilayer optimization is expected to significantly reduce total network costs, one of the main drawbacks is that E2E service provisioning and fault detection are complex due to the lack of E2E signaling and OAM protocols.

As discussed in D3.1 [D3.1], IETF and ITU-T provide a set of requirements with regard to the architecture and functionality for packet transport networks (i.e. MPLS-TP), considering such items as OAM, network services, and underlying networks. Specifically, OAM functions must be self-sufficient and compatible with the bidirectional nature of connections. The main OAM mechanisms required by the joint ITU-T – IETF working group for fault management are: Continuity check / verification, Alarm suppression, Lock indication, Diagnostic test, Trace-route, and Remote defect indication. The main OAM mechanisms required by the joint ITU-T – IETF working group for performance monitoring are: Packet loss measurement and Delay measurement.

On the other hand MEF, focusing on the OAM service, has specified the following list of requirements: service OAM should discover other elements in the Metro Ethernet Networks; service OAM should monitor the connectivity status of other elements (active, not-active, partially active); performance monitoring should estimate Frame Loss Ratio Performance, Frame Delay Performance, and Frame Delay Variation Performance; in a multi-domain environment OAM frames should be prevented from “leaking” outside the appropriate OAM domain to which they apply; the OAM frames should traverse the same paths as the service frames; the OAM should be independent of but allow interoperability with the underlying transport layer and its OAM capabilities; the OAM should be independent of the application layer technologies and OAM capabilities.

Based on the above signaling and network requirements for QoS and OAM and in comparison with RSVP-TE we attempt in the following section to see the suitability of the NSIS suite of protocols as a possible solution to the problem.

### **4.3.3 NSIS suitability for the STRONGEST architecture**

Two advantages of the NSIS protocol suite over RSVP signaling is that NSIS protocols can be used in different parts of the network, for different needs, without the need for end-to-end deployment and secondly the signaling is intended for more purposes than just QoS resource reservation. As compared to RSVP-TE the NSIS signaling supports a variety of possible triggers from different parts of the network and it may initiate the signaling from hosts, domain boundary nodes (edge nodes), interior domain nodes etc. Although two NSIS peer nodes which communicate directly are said to be one hop away from each other, however this does not imply that it corresponds to a single IP hop but it could be a much longer distance away. Either of the NSIS nodes might store some temporary state information about the other, monitoring status, however there is no assumption that they will establish a long-term signaling connection between themselves. This will support the flexible dynamic parts of the interconnected networks. NSIS supports both path-coupled and path-decoupled signaling. In the case of path-coupled, signaling messages are routed only through the NSIS nodes that are on the data path. In the path-decoupled case, signaling messages are routed to nodes that are not necessarily on the data path but are aware of it. The advantage of path decoupled signaling is the ease of additional functionality deployment without upgrading any of the routers in the data plane e.g. to support authorization or accounting. NSLP also supports uni- and bi-directional operation support of the same session e.g. a voice call. The correlation of the signaling for the two flow directions will be carried out using NSLP, and NTLP will be used to bundle the messages together. Furthermore the decoupling of signaling with respect to transport layer

protocol as part of the same suite of protocols makes NSLP/NTLP very suitable for signaling between Heterogeneous Domains.

NSLP offers the possibility to achieve aggregation of the context of individual flow signaling, hence the merit of supporting scalability when it comes to performance issues. For example in the QoS NSLP it is possible to add together the resources specified in a number of separate reservations. Bypassing Intermediate Nodes is another major issue under consideration in the NSIS framework for the reason that not all NSIS intermediate nodes are related to a particular signaling application and should be traversed at the lowest level possible. This work could be advantageous when it comes to the STRONGEST objective to reduce the number of involved nodes.

Based on the above benefits and operation description of the NSIS suite of protocols it is worth further investigating its suitability for QoS and OAM associated signaling in the STRONGEST architecture.

#### **4.4 Application of RACS/PCE architecture**

As stated in STRONGEST deliverable D3.1, QoS and admission control can be offered to end-to-end application requests towards the control layer. The RACS control layer allows applications to ask for a set of resources, without having to deal with transport network details (transport technology used, domains involved, ...). Using the control layer we achieve a logical separation between application services and transport-layer services; a single application request can generate different network requests and can involve different kinds of networks and domains.

The RACS control layer can offer Traffic Admission also in inter-carrier and inter-domain scenarios. In particular, the SPDF may provide the following interfaces for inter-domain/carrier communication:

- the Gq' interface between AF and SPDF;
- the Ri' interface between two SPDFs.

A detailed description of the two scenarios is presented in Chapter 3.

The usage of these interfaces enables the possibility for service applications to ask for QoS from point A to point B without having to know how the request will be handled by the different network portions and network providers. In the following of this section both interconnection scenarios are analyzed from an application point of view.

The interface between AF and SPDF is suitable in case of a separation between service provider and network provider. The service application that needs QoS asks for traffic admission to the service provider that will manage the request and will interact with the involved network provider(s) in order to satisfy the application request. The main advantage of this scenario is the logical separation between application services and transport services: the mapping between a QoS request made to the service provider's AF can be completely different from the one sent from the service provider's AF to the network provider's SPDF. This logical division can be useful in order to have a different granularity.

The interface between SPDFs enables the interconnection between control layers. It is suitable in inter-carrier scenarios with portions of the network belonging to different network providers. In this scenario, the control layer that handles the service application request will interact with other control layers in order to fulfill the QoS request from point A to point B. The advantages of this scenario are:

- network topology separation: each control layer has to know only network details of the managed transport networks;
- in case more paths are available, more requests can be forwarded to different control layers: different network providers can be asked for resources and the best path can be chosen;
- the interconnection between control layers can be handled using both network policies (lower cost, bandwidth availability, existing path already established, ...) and non-network policies (contracts between network providers, time policies, type of traffic constraints, ...).

In conclusion, QoS control and traffic admittance, based on RACS control layer, can be easily offered to end-to-end services in a multi-carrier scenario with different operators and networks using control layer(s).

## 4.5 Delay-based Metric Abstraction for OIF E-NNI

In this study, multi-domain multi-carrier networks running OIF E-NNI routing are considered for the provisioning of services with strict delay constraints. Three open issues are addressed: (i) the choice of the TE Metric type to apply in OIF E-NNI routing; (ii) the integrated architecture encompassing Routing Controller (RC) and PCEs; (iii) the TE Metric abstraction scheme to compute the TE Metric value to advertise for virtual intra-domain links. This section reports on a specific implementation of the abstraction schemes described in Chapter 3 applied to the context of end-to-end services with strict delay constraints in multi-carrier networks.

### 4.5.1 TE Metric for delay-critical applications

OIF E-NNI routing defines three different approaches to represent an intra-domain topology, the *abstract node*, the *pseudo-node* and the *abstract link* model. In this study, we focus on the latter model where a full-mesh of virtual intra-domain links is considered between border nodes. Each virtual intra-domain link is then advertised through a TE-LSA. Differently from typical intra-domain OSPF-TE implementations, in the context of multi-carriers, TE-LSAs may not detail virtual intra-domain reservable bandwidth information due to confidentiality reasons. Thus, two TE parameters are mainly considered in the constraint-based routing: (i) the Interface Switching Capability Descriptor (ISCD) which describes the switching capability of the link (e.g., fiber-switched, lambda-switched, packet-switched) and includes also the Max LSP Bandwidth information; (ii) the TE Metric. The TE metric is an additive value that can be related to different network attributes. However, there is no established consensus about the attribute that may be utilized in hierarchical multi-domain multi-carrier scenarios. Possible candidate attributes are hop count, domain

hop count, end-to-end delay, end-to-end delay jitter, link availability, or a combination of them.

The hop count, often used as TE metric in intra-domain OSPF-TE routing, presents a noticeable concern on domain confidentiality, since a non-trivial portion of the real topology may be inferred.

The domain hop count refers to the number of traversed domains. This option preserves confidentiality and scalability, however it is a quasi-static parameter which may provide poor TE performance. Moreover, this metric is used by BGP and its utilization within OIF E-NNI routing would achieve very limited improvements in comparison with BGP-based routing.

More complex metrics may be obtained as a combination of different network parameters (e.g., link availability, administrative cost). However, such metrics introduce a significant concern of measurability, agreement and trustiness among administrative carriers.

The end-to-end delay represents one of attributes mostly considered in the literature [RFC2679]. However, manual entry or specific dedicated hardware monitors may be required to measure, collect and announce end-to-end delay values. The main limitation affecting this attribute is that it is hard to perform and predict accurate end-to-end delay measurements in case of congestion. However, QoS-guaranteed services are supposed to be supported by networks implementing specific policies aiming at avoiding congestion. In such scenario the end-to-end delay of multi-domain paths, characterized by long distance, is measurable and equals the propagation delay with good approximation. Moreover, path delay values are verifiable upon service set up, thus guaranteeing trustiness and SLA verification. The delay presents a direct relationship with services and represents a critical benchmark parameter for a large set of applications (e.g., real-time interactive video, VOIP, financial trading, search engines, interactive gaming). In case of networks with short distances (e.g., metro networks) the delay is less significant and the TE metric assumes a quasi-static value. Conversely, the delay is particularly suitable for multi-domain networks, where long and highly spread distances may lead to noticeably different values of delay among paths having the same source and destination nodes.

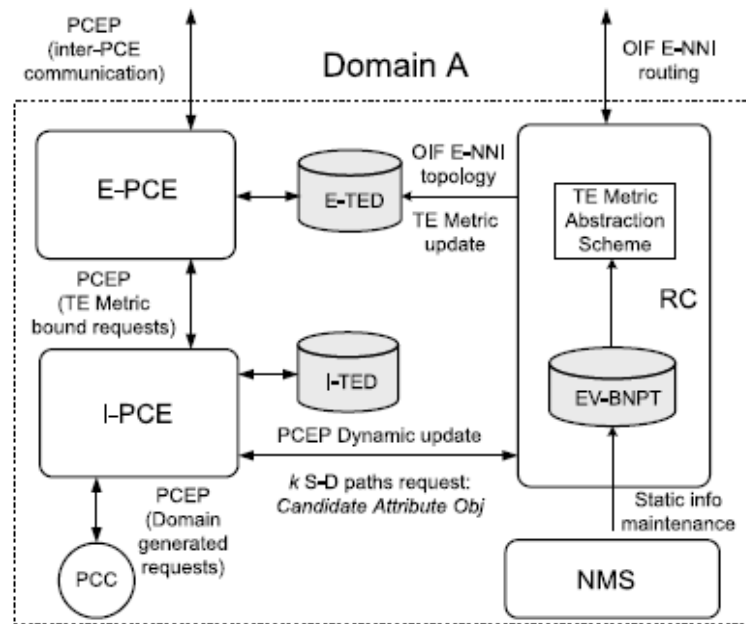
In this study, specifically targeting services having strict delay constraints and crossing multi-carrier networks, the end-to-end delay is then considered as the most suitable OIF E-NNI TE Metric.

## 4.5.2 PCE-RC Architecture

A joint PCE and RC architecture is proposed and elaborated within the STRONGEST project in the context of OIF E-NNI multi-carrier networks to enable the provisioning of delay-critical applications. It consists of an RC and two PCEs per domain with separated functions, as depicted in Figure 33. An Interior-PCE (I-PCE) and an Exterior-PCE (E-PCE) are functional elements responsible for the intra-domain and hierarchical inter-domain path computations, respectively. They retrieve TE information from Interior/Exterior TE Database (I-TED, E-TED). In particular, the I-TED stores the internal topology learned through the intra-domain routing protocol (e.g., OSPF-TE). The E-



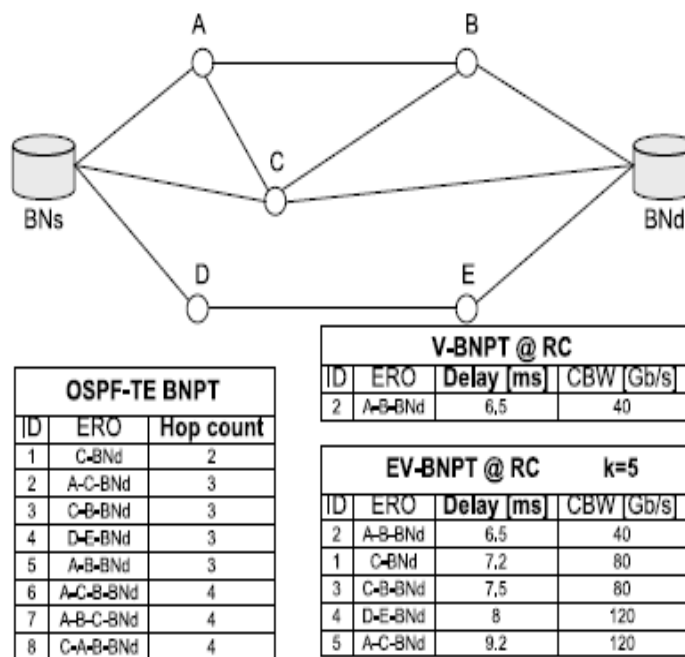
TED is created and maintained by RC and stores the topology learned through E-NNI instance peering. RC and I-PCE, I-PCE and E-PCE, and E-PCEs, communicate through PCE Protocol (PCEP). In case of system co-location (e.g., I-PCE and E-PCE), dedicated Application Programming Interfaces (API) may be used instead of PCEP.



**Figure 33: RC-PCE integrated architecture**

*Virtual intra-domain link advertisement.* To perform the advertisement of virtual intra-domain link information, the RC acts as Path Computation Client (PCC) and it requests the I-PCE to perform the path computations having the BN pairs as end-points. However, the specific issue of computing multiple paths with the same end points needs to be addressed. Indeed, by exploiting the current PCEP protocol specifications, the I-PCE returns just the computed least cost (i.e., in terms of TE Metric) path for each BN pair. The RC stores such information in a Virtual Border Node Path Table (V-BNPT) and advertises the related computed attributes to other domains through the E-NNI routing instance, including the TE Metric (i.e., delay), for each BN-BN path. However, such information, disseminated as TE metric, imposes that subsequent inter-domain LSP requests have to satisfy the minimum intra-domain delay constraint. This constraint may significantly affect the overall intra-domain network resource utilization since it typically imposes that the path used for the subsequent inter-domain LSP provisioning is just the one associated to the advertised minimum value. In order to overcome such limitation, we propose to exploit the PCEP candidate attribute object recently introduced in [draft-imp] for different purposes (i.e., impairments in WSON). This PCEP extension enables the PCC to request the PCE to compute a set of  $k > 1$  paths per BN pair. In addition, we introduce an extended version of the V-BNPT table, called Extended Virtual BNPT (EV-BNPT). EV-BNPT includes, for each BN pair, the delay-ordered list of BN-BN paths with information about available bandwidth and static additional information collected from the Network Management System (NMS) (e.g., link delay, mileage, capacity). Such information may be used by a TE Metric abstraction scheme in order to derive the desired value of TE metric to be announced. PCEP communication is utilized between the RC (acting as PCC) and the I-PCE to

populate the EV-BNPT. It includes two different procedures: 1) static information maintenance and 2) dynamic information update. The former procedure is triggered in case of physical network changes: link information entries may be inserted or removed. The latter procedure is triggered periodically or by NMS when connections are established or torn down, thus causing allocated network resource variation. In this case the request for the set of  $k$  paths is sent by the RC to the I-PCE through a PCReq message including the candidate attribute object. The I-PCE returns the paths Explicit Routing Object (ERO) list and the value of the maximum available bandwidth of the paths. The values of end-to-end delay and capacity of the paths are computed based on the ERO list and the static information set collected by NMS. The dynamic update of EV-BNPT depends on the implemented abstraction scheme policy and it may trigger a virtual intra-domain link TE metric change. In this case, the local E-TED and RC are updated with the new TE value and a TE-LSA is generated towards the confederated RCs. The OSPF-TE BNPT, stored within the I-PCE, contains the paths, described through the ERO sequence, ordered as a function of the intra-domain TE metric (e.g., hop count). The VBNPT, by using current standard PCEP, just retrieves and stores the (single) result of the path computation between nodes BNs and BNd with the mentioned TE constraint (e.g., the shortest path in terms of delay). The EV-BNPT is able to store up to  $k$  BNs-BNd paths ordered as a function of the delay. In this way the abstraction scheme can be applied to all the stored paths to manage a range of possible TE metrics to advertise. In Figure 34 an example of virtual intra-domain link summarization is shown.



**Figure 34: Virtual intra-domain link summarization**

*Inter-domain path computation.* When a PCC (e.g., a node, a management system) requests a QoS-guaranteed connection provisioning, it sends a PCReq message to its I-PCE, which forwards the request to the E-PCE in case of inter-domain provisioning. The E-PCE, based on the information collected by the RC through E-NNI flooding (i.e, virtual intra-domain and inter-domain links topology) and stored in the E-TED, computes the

traversed domains sequence by identifying a path composed of virtual intra-domain links and inter-domain links such that the total TE metric satisfies the requested TE constraint. In this case, the E-PCE starts the standard inter-PCE communication with the next involved domain by sending a PCReq message including the BNs sequence in the Include Route Object (IRO). The E-PCEs forward the request up to the destination domain E-PCE. The destination E-PCE maps the request to an intra-domain request between the last BN and the destination node performed to the I-PCE. The I-PCE ERO reply is encrypted through a Path-Key object, thus preserving confidentiality, and inserted, together with the BN-destination metric, within the PCRep message.

During the path computation process, each transit E-PCE, upon the PCRep reception, sends a PCReq message to its I-PCE asking for a BN-BN path including the required TE attributes (e.g., bandwidth) and the TE metric constraint. Such constraint is the current TE Metric value announced by the RC. This metric is included as TE Metric Bound (i.e., with the B flag activated) in the Metric Object. The I-PCE computes a path subject to the advertised delay constraint. The reply is sent to the E-PCE, which encrypts the ERO within a path-key object and includes the cumulated metric within the Metric object. At the end of the inter-PCE communication procedure, the source E-PCE collects the path-key list and is able to check whether the total cumulated metric satisfies the requested end-to-end delay. In case of acceptable value, the resources along the path are reserved by the RSVP-TE protocol including the encrypted path-key list.

The twofold PCE architecture provides a complete separation between the actual intra-domain resources (stored in the OSPF-TE BNPT and in the I-TED) and the resources advertised to other domains (stored in the E-TED). Moreover, the EV-BNPT update mechanism enables the implementation of TE metric abstraction schemes providing full control on the advertised E-NNI parameters and updates.

### 4.5.3 Metric Abstraction Schemes

The proposed TE metric abstraction schemes consider, for each virtual intra-domain links, the set of possible intra-domain paths between all the border nodes. Within each set, several sub-sets are identified having similar border-to-border delay. The advertised TE metric will assume only the delay values characterizing the subsets. The subsets are ordered according to the delay value. For each subset, the schemes dynamically evaluate the amount of additional bandwidth made available between the considered border nodes with respect to the whole bandwidth provided by the subsets having lower delay. These bandwidth values are dynamically updated according to network conditions. Three main figures of merit can be considered to assess the performance of the TE metric abstraction schemes. The first figure of merit refers to the OIF E-NNI control plane scalability, e.g. the rate  $R$  of the advertised TE-LSA. In general, abstraction schemes should not change too often the advertised values in order not to overload the control plane and jeopardize the advertisement trustiness and efficiency (note that one domain might generate excessive LSA flooding, thus potentially impacting control plane stability in external domains). The second figure of merit refers to the intra-domain network resource utilization, e.g. the intra-domain blocking probability  $P_b$ . In fact, when a  $BN_i$ - $BN_j$  virtual intra-domain link is advertised with a specific TE Metric value  $d_{adv}(BN_i, BN_j)$ , subsequent multi-domain requests between  $BN_i$  and  $BN_j$  have to be provisioned by applying the intra-

domain routing constraint of the advertised delay value thus excluding possible longer paths and potentially inducing network congestion on the eligible paths.

The third figure of merit refers to the advertised service level. The lower the advertised TE Metric value  $d_{adv}(BNi, BNj)$ , the better the service level (i.e., delay) proposed through OIF E-NNI advertisement to customers belonging to different domains. Low TE values would attract the provisioning of high quality services from other domains, thus increasing revenues. To this purpose, a parameter  $\rho$  is defined as the ratio between the advertised TE Metric  $d_{adv}(BNi, BNj)$  and the related minimum available TE Metric  $d_{min}(BNi, BNj)$ , averaged out of all virtual intra-domain links. The lower the  $\rho$  value, the better the advertised service. According to the three aforementioned figures of merit, two abstraction schemes can first be identified as bounds.

The first scheme, called *MinMetric* (MinM), advertises the minimum available TE Metric value, i.e. the delay value of the first sub-set of paths with non-null available bandwidth. With MinM, a new TE-LSA Update is generated if: (i) a new TE Metric lower than the advertised value becomes available, (ii) the advertised sub-set becomes empty and the one with higher delay is advertised, (iii) the whole set becomes empty or return available. MinM represents the lower bound in terms of the advertised service level, which is always the optimal available ( $\rho = 1$ ).

The second scheme, called *MaxMetric* (MaxM), advertises the maximum TE Metric value of the set.

MaxM generates a new TE-LSA Update only when the whole set becomes empty or return available. MaxM represents the upper bound in terms of the advertised service level, which is the worst possible value. Furthermore, MaxM represents the lower bound in terms of generated TE-LSA, since the TE metric is kept constant to the maximum value.

The following schemes are then proposed.

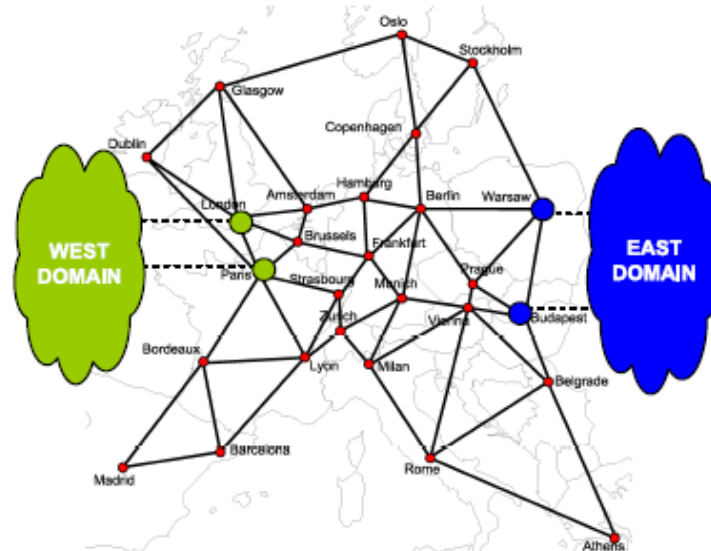
*K-path scheme.* The *K-path* scheme advertises the TE Metric of the K-th delay-ordered sub-set of paths having non-null available bandwidth. The TE-LSA update generation utilizes the same criteria illustrated for MinM scheme. The K-path represents a generalization of the MinM scheme, where MinM is K-path with K=1.

*Delay-Bandwidth Aware abstraction schemes (DBA).* The proposed Delay-based Bandwidth-aware Abstraction Scheme (DBA), advertises the TE Metric value associated to the group of paths which present a suitable proportion between the provided bandwidth and the related BN-BN delay. A parameter  $h(BNi-BNj)$  is introduced to account for the available cumulated bandwidth of the considered sub-sets and the ratio between the minimum available delay and the highest delay of the subsets. The aim is to achieve an effective trade-off such that the advertised TE Metric is able to support a reasonable amount of LSPs providing high service level. In DBA, two cases determine a new advertisement: the emptiness of the advertised sub-sets and the availability of certain amount of bandwidth on sub-sets having a delay lower than the advertised one.

Two different versions of the DBA abstraction scheme are considered: the *Control-oriented DBA* (C-DBA) scheme and the *Service-oriented DBA* (S-DBA) scheme. The two schemes differ in the adopted threshold on the parameter  $h(BNi-BNj)$ . In C-DBA, the

threshold is configured in such a way that the TE-LSA updates generation is kept limited and a higher number of paths are considered available for routing. Conversely, in S-DBA, the threshold enables a more frequent TE-LSA update generation and a better announced service level.

The considered abstraction schemes have been evaluated by means of simulations. The considered network is a Wavelength Switched Optical Network (WSO). The simulation scenario comprises the Pan-European network topology depicted in **Figure 35** consisting of 27 optical nodes and 55 bidirectional WDM links with capacity  $W = 40$  wavelengths along each direction. Each wavelength channel has a capacity  $C=10$  Gb/s and the Max LSP Bandwidth is set to one wavelength channel. The considered network acts as a single routing and transparency domain. It is attached to 2 adjacent domains by means of 2 BNs per domain. Thus, 8 unidirectional virtual intra-domain links are advertised to external domains, not considering the links between the BNs connected to the same adjacent domain (e.g., the London-Paris and the Budapest-Warsaw links in **Figure 35**). The simulator includes the I-PCE and the RC equipped with EV-BNPT, together with the extended PCEP and RSVP-TE. Incoming connections request a one-wavelength LSP within the considered network. The LSPs arrival process is a Poisson one and is uniformly distributed between intra-domain and inter-domain requests. Intra-domain requests consider any possible node pair as LSP end-points, while inter-domain requests consider BN pairs as the LSP end-points. The LSP inter-arrival and holding times follow a negative exponential distribution. The I-PCE utilises detailed bandwidth information and applies the wavelength continuity constraint and the least fill policy among paths including all paths within one hop from the shortest path. Moreover, for inter-domain LSPs requests only, the additional routing constraint introduced by the specific abstraction scheme is applied, thus limiting the routing selection to paths satisfying the currently advertised TE metric.

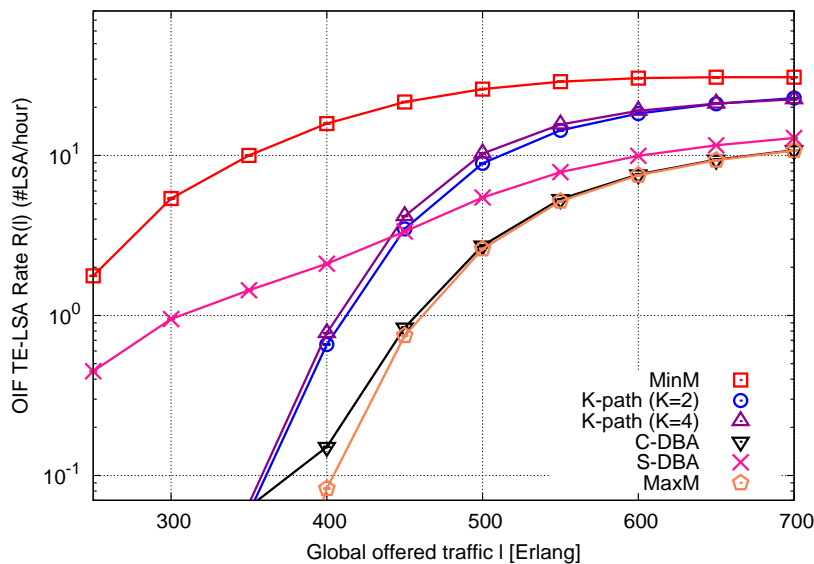


**Figure 35: Pan-European topology**

Once routing is computed, RSVP-TE is utilized for the LSP resource allocation and set up. The EV-BNPT update procedure is triggered when an LSP is newly established or

torn down. The EV-BNPT update may trigger the RC to generate TE-LSA update messages, based on the utilized abstraction scheme.

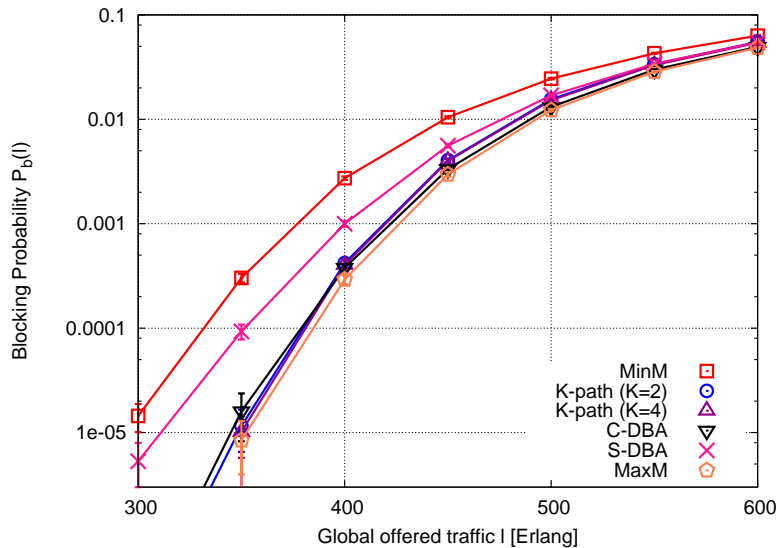
**Figure 36** shows the rate  $R$  (generated TE-LSA per hour) of the considered schemes. TE-LSA refresh messages are not measured. Results show that MinM provides the highest rate at any load. This occurs due to frequent advertisements of new available paths having the current minimum delay. The K-path schemes achieves better results (i.e., lower rate) at low and medium loads due to multiple path clustering. The effect of the path clustering is more evident as  $K$  increases. However, at high loads, the value of  $R$  obtained by the K-path scheme tends to the one obtained by MinM. This demonstrates the inefficiency of K-path at critical loads, due to bounce back advertisement which keeps the announced TE metric continuously unstable. The MaxM scheme provides the lowest  $R$  at any load, because it advertises only full and available virtual intra-domain link events. DBA schemes provide intermediate performance, in particular C-DBA provides low rate, while S-DBA provides a slightly relative higher rate but still limited and with a lower slope with respect to MinM and K-path. Both C-DBA and S-DBA tend to the lower bound represented by MaxM. This means that, as the load increases, the DBA schemes trigger a limited amount of TE-LSA updates and the joint delay-bandwidth aware threshold mechanism prevents the TE metric from bounce back change advertisements, since only available/unavailable virtual intra-domain link events are triggered. In particular the C-DBA scheme provides a reduced amount of TE-LSAs, thus keeping OIF control plane more stable.



**Figure 36: Generated TE-LSA rate**

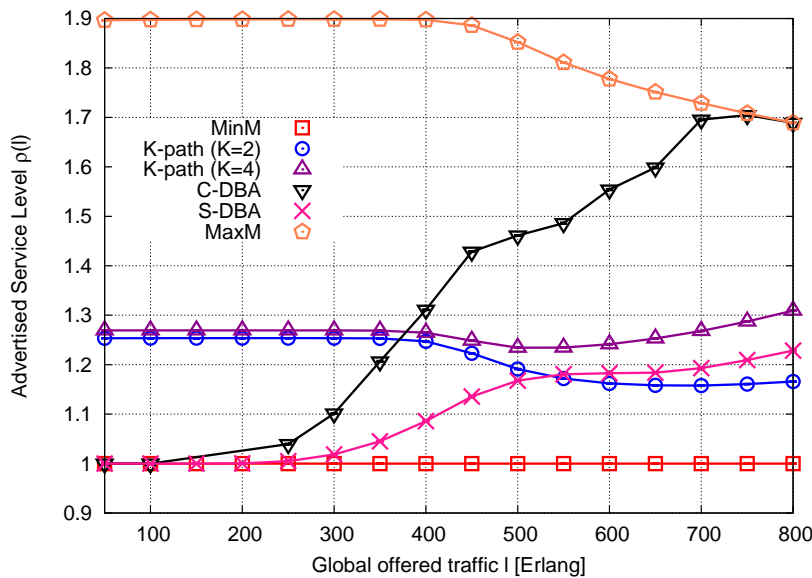
Blocking probability performance, depicted in **Figure 36**, show that the MinM scheme suffers from a too strict routing constraint, being the shortest delay-based path the sole eligible for routing. K-path improves blocking by allowing more paths to be eligible for routing: as  $K$  grows the blocking is only slightly reduced. The MaxM scheme presents the lowest blocking at any load, since all the paths are eligible for routing and network resources are better utilized, thus reducing congestion. C-DBA scheme equals K-path

blocking rate, while S-DBA experiences an intermediate block between MinM and K-path at medium loads, while assumes the same value as the load increases.



**Figure 37: Blocking probability**

**Figure 38** shows the service level advertised to external domains, expressed in terms of  $\rho$ . Results first confirm that the bound MinM provides the best performance at any load (MinM always advertises the lowest available delay). K-path service level is kept constant but never optimal, either at low loads, since the best delay is never advertised. In particular, K-path advertises a worse service level at low loads and suffers from control plane instability at high loads. The MaxM scheme provides the worst performance, because it announces the highest possible TE-metric. DBA schemes provide optimal service level results at low and medium load. In particular, S-DBA provides a service level very close to 1 even at high and very high loads, thus confirming that this specific DBA scheme is designed to offer a noticeable service level at any load. This occurs because at very high loads the probability that a link is full is very high, the threshold is never reached due to lack of resources, and the advertisement of the best path does not produce blocking and strongly reduces  $\rho$ . Conversely, C-DBA performance decreases as network load increases. This is due to the higher threshold with respect to S-DBA which implies that a higher TE metric is announced especially at high loads.



**Figure 38: Advertised service level**

### 4.5.4 Conclusions

This study proposed a novel PCE-RC architecture for inter-domain and inter-carrier routing based on the OIF E-NNI framework. The choice of the end-to-end delay as TE metric to be disseminated through hierarchical OIF routing instance and the twofold PCE structure enables QoS-based inter-domain path computation preserving intra-domain confidentiality and trustiness. Furthermore, the implementation of the EV-BNPT inside the RC drives the use of several TE metric abstraction schemes aiming at guaranteeing a high advertised service level. Different TE schemes were evaluated, including upper and lower bounds. The proposed DBA abstraction schemes, implementing a joint delay-bandwidth threshold, are then designed to announce the most suitable TE metric in terms of intra-domain resource utilization, control plane stability and offered service. Simulation results showed, on the one hand, the ability of S-DBA scheme to offer an effective service level while keeping block and control plane load within acceptable values. On the other hand the C-DBA scheme is demonstrated to guarantee a considerable control plane stability, even in highly dynamic contexts, and a low network block while guaranteeing acceptable service level at low and medium traffic loads.

Part of this innovative study has been published in [Paol-JOCN10]. The paper also includes further details including the formal definition of the abstraction schemes and additional references and simulation results.

### 4.6 Considerations and future works on end-to-end services

This chapter summarized the relevant application services providing the mapping to the service classes proposed and considered within the STRONGEST project. In addition, some of the innovative studies addressed in the previous chapter have been here investigated and applied within specific network scenarios. Innovative results have been



presented in the context of OAM, RACF-based traffic admittance, OIF E-NNI and abstraction schemes for services with strict delay constraints.

Future work will continue to apply and evaluate relevant STRONGEST control plane solutions and procedures in the specific context of end-to-end services with strict QoS requirements.

In addition, considerations will be provided on E2E services to better understand differentiation mechanisms, in terms of identifying/defining boundaries between commodity and high-margin services. In conjunction with traffic volumes and service mixes provided by Work Package 2, this will serve to crystallize performance requirements for the STRONGEST network architecture and hence drive research activities particularly in the OAM and Control Plane work areas.

Concerning control layer activities, current work about fixed and mobile control layer convergence will be taken into account. This implies an analysis of 3GPP PCC architecture and related mapping with IETF PCE architecture.

## 5 Conclusions

The present deliverable summarized the first year activities of STRONGEST WP3 in the context of the medium-term network scenarios which address the inter-working between heterogeneous GMPLS-controlled networks.

On the basis of the reference architectures and open issues specified within the STRONGEST Deliverable D3.1, this document detailed considerations and proposed innovative solutions in the context of (i) OAM parameters and mechanisms for packet transport; (ii) control plane architectures, solutions and proposed extensions and (iii) end-to-end services. Relevant results are reported in the field of OAM and multi-domain single/multi carrier control plane solutions, with particular focus on the PCE-based architectures. Some of the presented results have been also published in scientific and standardization documents, thus demonstrating the validity of the proposed technical solutions and the significant impact of the STRONGEST project within the research and standardization communities.

## 6. References

### 6.1. STRONGEST Publications

[Casellas10]	R. Casellas, R. Martínez, R. Muñoz, T. Tsuritani, L. Liu, M. Tsurusawa, "Lab-Trial of Multi-Domain Lightpath Provisioning with PCE Path Computation combining BRPC and Path-Key Topology Confidentiality in GMPLS Translucent WSON networks", ECOC 2010
[Casellas11]	R. Casellas, R. Muñoz, R. Martínez, "Lab Trial of Multi-Domain Path Computation in GMPLS Controlled WSON Using a Hierarchical PCE", OFC2011
[Cugini10]	F. Cugini, N. Andriolli, G. Bottari, P. Iovanna, L. Valcarenghi, P. Castoldi, "Designated PCE Election Procedure for Traffic Engineering Database Creation in GMPLS Multi-Layer", ECOC 2010
[Giorgetti11]	A. Giorgetti, F. Paolucci, F. Cugini, P. Castoldi, "Hierarchical PCE in GMPLS-based Multi-Domain Wavelength Switched Optical Networks" OFC 2011
[Munoz10]	R. Muñoz, R. Casellas, R. Martínez, "Experimental Evaluation of Dynamic PCE-based Path Restoration with Centralized and Distributed Wavelength Assignment in GMPLS-enabled Transparent WSON networks", ECOC 2010, September 2010
[Paol-ECOC10]	F. Paolucci, F. Cugini, B. Martini, M. Gharbaoui, L. Valcarenghi, P. Castoldi, "Preserving Confidentiality in PCEP-based Inter-domain Path Computation", ECOC 2010
[Paol-JOCN10]	F. Paolucci, F. Cugini, P. Iovanna, G. Bottari, L. Valcarenghi, P. Castoldi, "Delay-Bandwidth Aware (DBA) Metric Abstraction Schemes for OIF E-NNI Multi-domain Traffic Engineering", Journal of Optical Communications and Networking (JOCN), Vol. 2, Issue 10, pp. 782-792, 2010
[draft-ietf-pce-gmpls-pcep-extensions]	C. Margaria et al., "PCEP extensions for GMPLS", Internet Draft <a href="http://tools.ietf.org/wg/pce/draft-ietf-pce-gmpls-pcep-extensions/">http://tools.ietf.org/wg/pce/draft-ietf-pce-gmpls-pcep-extensions/</a>

### 6.2. Informative references

[Annex I]	Annex I to the Contract
[Annex II]	Annex II to the Contract
[BBF]	<a href="http://www.broadband-forum.org/technical/download/ipmpls/IPMPLSForum19.0.0.pdf">http://www.broadband-forum.org/technical/download/ipmpls/IPMPLSForum19.0.0.pdf</a>
[Bor98]	M. S. Borella, D. Swider, S. Uludag, and G. B. Brewster, "Internet Packet Loss: Measurement and Implications for End-to-End QoS," Proceedings, International Conference on Parallel Processing, August 1998.
[Buzzi10]	L. Buzzi, M. Conforto Cardellini, D. Siracusa, G. Maier, F. Paolucci, F. Cugini, L. Valcarenghi, P. Castoldi, "Hierarchical Border Gateway Protocol (HBGP) for PCE-based

	Multi-domain Traffic Engineering”, ICC IEEE International Conference on Communications, Cape Town, South Africa, 2010.
[Cug05]	F. Cugini, N. Sambo, N. Andriolli, A. Giorgetti, L. Valcarengi, P. Castoldi, E. Le Rouzic, J. Poirrier, J. of Lightweight Technology, Vol. 26, No. 19, Oct 2008
[D3.1]	STRONGEST Deliverable D3.1, August 2010
[Dem09]	Y. Demchenko, M. Cristea, C. DeLaat, “XACML Policy Profile for Multi-Domain Network Resource Provisioning and Supporting Authorization Infrastructure” IEEE POLICY Conf., 2009
[draft-gonzalezdedios-pce-reservation-state]	PCEP Extensions for Temporary Reservation of Computed Path Resources and Support for Limited Context State in PCE <a href="http://tools.ietf.org/html/draft-gonzalezdedios-pce-resv-res-context-state-00">http://tools.ietf.org/html/draft-gonzalezdedios-pce-resv-res-context-state-00</a>
[draft-ietf-pce-inter-layer-ext]	Extensions to the Path Computation Element communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering <a href="http://tools.ietf.org/wg/pce/draft-ietf-pce-inter-layer-ext/">http://tools.ietf.org/wg/pce/draft-ietf-pce-inter-layer-ext/</a>
[draft-imp]	Y. Lee et al, “A Framework for the Control of Wavelength Switched Optical Networks (WSON) with Impairments”, draft-ietf-ccamp-wson-impairments-04, Oct 2010
[draft-lee]	Y. Lee and G. Bernstein, draft-lee-pce-ted-alternatives-02.txt, May 2009
[draft-zhang-pcep-hierarchy-extensions]	Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE) <a href="http://tools.ietf.org/id/draft-zhang-pcep-hierarchy-extensions-00.txt">http://tools.ietf.org/id/draft-zhang-pcep-hierarchy-extensions-00.txt</a>
[ENNI]	External Network-Network Interface (E-NNI) OSPF-based routing -1.0 (Intra-Carrier) implementation agreement, Optical Internetworking Forum, OIF, Jan 2007
[FRA]	P. Francois et al., “Achieving sub-second IGP convergence in large IP networks”, ACM SIGCOMM Computer Communication Review, 2005
[H-PCE]	King, D. and A. Farrel, "The Application of the PCE Architecture to the Determination of a Sequence of Domains in MPLS & GMPLS", Work In Progress, 2010.
[HUA]	S. Huang et al., “An experimental analysis on OSPF-TE convergence time”, 2008
[Iov-Bot-DBA_MD]	F. Paolucci, F. Cugini, P. Iovanna, G. Bottari, P. Castoldi, “Delay-Based Bandwidth-Aware Topology Abstraction Scheme for OIF E-NNI Multi-Domain Routing”, Proc. of OFC/NFOEC 2010, San Diego, CA, USA, March 21-25, 2010.
[king_H-PCE]	D. King et A. Farrel, “The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS”, draft-king-pce-hierarchy-fwk-05, September 2010
[LAU08]	W.Lautenschlaeger, W.Frohberg, “Bandwidth Dimensioning in Packet-based Aggregation

	Networks,” 13th International Telecommunications Network Strategy and Planning Symposium, Networks2008, Budapest, 2008
[NIK]	E. Nikolouzou et al., “Network services definition and deployment in a differentiated services architecture”, 2002
[NOB]	Nobel2, D1.1, “Architectural vision of network evolution”, 2006
[RFC2679]	G. Almes, S. Kalidindi, and M. Zekauskas, “A One-way Delay Metric for IPPM,” Internet Engineering Task Force (IETF), RFC 2679, September 1999.
[RFC4461]	Yasukawa, S., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC 4461, April 2006.
[RFC4655]	Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
[RFC4875]	Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol – Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
[RFC5088]	JL. Le Roux et al, “OSPF Protocol Extensions for Path Computation Element (PCE) Discovery”, RFC 5088, January 2008
[RFC5152]	Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", February 2008.
[RFC5212]	K. Shiomoto et al., “Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)”, RFC 5212, Jul. 2008
[RFC5376]	Bitar, N., Zhang, R., and K. Kumaki, "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)", RFC 5376, November 2008.
[RFC5376]	N. Bithar, R. Zhang, and K. Kumaki, “Inter-AS requirements for the Path Computation Element Communication Protocol (PCECP)”, RFC 5376, Nov. 08.
[RFC5440]	J-P. Vasseur and J. LeRoux, “Path Computation Element (PCE) Communication Protocol (PCEP)”, RFC 5440, March 2009.
[RFC5441]	JP. Vasseur, Ed. “A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths”, RFC 5441, April 2009.
[RFC5441]	J-P. Vasseur, et al., “A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths”, RFC 5441, April 2009.
[RFC5520]	Bradford, R., Ed., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC5520, April09

[RFC5541]	JL. Le Roux, JP. Vasseur, Y. Lee "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, April 2009.
[RFC5623]	E. Oki, T. Takeda, JL. Le Roux et A. Farrel, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, September 2009
[RFC5862]	Yasukawa, S. and A. Farrel, "PCC-PCE Communication Requirements for Point to Multipoint Multiprotocol Label Switching Traffic Engineering (MPLS-TE)", RFC 5862, June 2010.
[RFC5880]	D. Katz, D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC5880, June 2010
[SPR]	<a href="http://www.sprint.com/business/resources/mpls_vpn.pdf">http://www.sprint.com/business/resources/mpls_vpn.pdf</a>
[Toktar04]	E. Toktar, E. Jamhour, C. Maziero "RSVP Policy Control using XACML", IEEE POLICY 2004
[Y1731]	ITU Recommendation ITU-T Y.1731, OAM functions and mechanisms for Ethernet based networks
[zhao-id-p2mp-md]	PCE-based Computation Procedure To Compute Shortest Constrained P2MP Inter-domain Traffic Engineering Label Switched Paths, draft-zhao-pce-pcep-inter-domain-p2mp-procedures-06
[Zhu03]	H. Zhu, et al., "A Novel Generic Graph Model for Traffic Grooming in Heterogeneous WDM Mesh Networks", IEEE/ACM Transactions on Networking, April 2003
[draft-zhang-ccamp-gmpls-h-lsp-mln]	GMPLS-based Hierarchy LSP creation in Multi-Region and Multi-Layer Networks <a href="http://tools.ietf.org/id/draft-zhang-ccamp-gmpls-h-lsp-mln-02.txt">http://tools.ietf.org/id/draft-zhang-ccamp-gmpls-h-lsp-mln-02.txt</a>

## 7. Document History

Version	Date	Authors	Comment
0.01	29/09/2010	C. Zema	D3.2 template distribution
0.04	21/10/2010	C. Zema, F. Cugini	D3.2 ToC 1 <sup>st</sup> proposal
0.06	12/11/2010	C. Zema, F. Cugini	D3.2 1 <sup>st</sup> draft (incomplete)
0.08	07/12/2010	C. Zema, F. Cugini	D3,2 draft for quality check
0.09	16/12/2010	E. Vezzoni	Quality checked version
1.00	17/12/2010	C. Zema, F. Cugini, A. Di Giglio, E. Vezzoni	D3.2 pre-final draft after audioconference
1.1	22/12/2010	A. Di Giglio, E. Vezzoni	D3.2 version for final editing
1.2	23/12/2010	F. Cugini	D3.2 version for WP3 feedback
1.3	28/12/2010	WP3, F. Cugini, E. Vezzoni	Final draft for G.A. approval
2.0	31/12/2010	General assembly	Approved version

## 8. Acronyms

ABR	Area Border Router
AS	Autonomous System
ASBR	Autonomous System Border Router
BER	Bit Error Rate
BRPC	Backwards Recursive Path Computation
CIDR	Classless Inter-Domain Routing
CRC32	cyclic redundancy check (32bit)
ENNI	External Network to Network Interface
ERO	Explicit Route Object
FIB	Forwarding Information Base
GIST	General Internet Signaling Transport
GMPLS	Generalized Multi Protocol Label Switching
GBW	Guaranteed Bandwidth
H-PCE	Hierarchical PCE
IGP	Interior Gateway Protocol
ISDN	Integrated Services Digital Network
LB	Loop Back
LBM	Loop Back Message
LBR	Loop Back Replay
LSA	Link State Advertisement
LSP	Label-Switched Path
MaxAvBW	Maximum Available Bandwidth



ME	Maintenance Entity
MEP	Maintenance End Point
MIP	Maintenance Entity
MP	Maintenance Point
MPLS	Multi Protocol Label Switching
MPLS-TP	MPL S Transport Profile
NSIS	Next Step In Signaling
NSLP	NSIS Signaling Layer Protocol
NTLP	NSIS Transport Layer Protocol
OAM	Operation, Administration, and Maintenance
OCC	Optical Connection Controller
OIF	Optical Internetworking Forum
OPEX	Operational expenditures
OSNR	Optical Signal to Noise Ratio
OTH	Optical Transport Hierarchy
P2MP	Point to Multi Point
PBW	Peak Bandwidth
PCC	Path Computation Client
PCE	Path Computation Element
PCEP	Path Computation Element (Communications) Protocol
PDH	Plesiochronous Digital Hierarchy
PDU	Protocol Data Unit
QoR	Quality of Resilience
QoS	Quality of Service
RIB	Routing Information Base
RMF	Resource Management Function

RSA	Rivest Shamir Adelman
RTT	Round Trip Time
RWA	Routing and Wavelength Assignment
SDH	Synchronous Digital Hierarchy
SLA	Service Level Agreement
SLBA	Service Loop Back Level Agreement
SPT	Shortest Path Tree
SRLG	Shared Risk Link Group
TCP	Transport Control Protocol
TE	Traffic-Engineering
TLV	Type Length Value
TCO	Total cost of ownership
VSPT	Virtual Shortest Path Tree
WCC	Wavelength Continuity Constraint
WDM	Wavelength Division Multiplexing
WP	Work Package
WSON	Wavelength Switched Optical Network