PCE QoS Tools and Related Scalability in WDM Networks

Joana Sócrates Dantas[a,b], Davide Careglio[b], Regina Melo Silveira[a], Wilson Vicente Ruggiero[a], Josep Solé-Pareta[b]

[a] *University of São Paulo, LARC - Computer Networks and Architecture Laboratory, Brazil*

[b] *Technical University of Catalonia (UPC), c/ Jordi Girona, 1-3, 08034 Barcelona, Catalonia, Spain*

ABSTRACT

The virtual circuit characteristic of MPLS and GMPLS architectures has enabled the implement of Traffic Engineering (TE) and Differentiated Services (Diff-Serv) profile to networks. Both schemes are possible by the application of preemption policies and priority of path assignment. With the advent of Path Computation Element (PCE) the routing task can be centralized and properly (re-)designed to accommodate MPLS' and GMPLS' Diff-Serv and TE characteristics. However, as a centralized system, the complex characteristics of heterogeneous networks and the extra data involved on TE and Diff-Serv constrains and requirements for path computations may overload the PCEs' processing capacity and radically increase the number of PCEP control messages exchanged, jeopardizing the scalability of the system in general. In this article we present an analysis on the current standardized PCE tools in a WDM network in terms of the number of control messages exchanged impact on the system and its optimization performance.

1. INTRODUCTION

As networking advance technically the variety and amount of requirements service providers deal with proportionally increases. In such a scenario Traffic Engineering (TE) mechanisms may be used by service providers as an efficient tool for managing and balancing resources throughout a network and also as an assistance to provide Quality of Service (QoS) adequately conforming to Service Level Agreements (SLA) [1].

One of the diverse requirements that TE helps attend is the varied performance guarantees demanded by multimedia applications. Due to their intrinsic virtual circuit switched characteristics MPLS and GMPLS networks support routing paths to be specified and therefore qualified according to their ability to satisfy diverse QoS requirements. Differentiated Service (Diff-Serv) architecture permits traffic flows to be classified into classes and managed following specific rules so that higher priority traffic can receive preferential service [2].

Diff-Serv aggregated to MPLS networks aid advanced planning procedures in order to provide specific QoS guarantees [2]. This combination, of Diff-Serv and MPLS resulted in the Diff-Serv aware TE (DS-TE), a scalable solution to support QoS and efficiently balanced network's use. The DS-TE is implemented in MPLS and GMPLS network through distributed constrained path computation.

In heterogeneous multi-domain and multi-layer networks, however, constrained-based path computation becomes extremely complex and resource invoking. Path Computation Element (PCE) architecture was developed as a possible solution to this issue [3]. In PCE architecture dedicated devices centralizes the function of constrained based path computation releasing the burden from the MPLS or GMPLS nodes. Furthermore PCEs in different layers or domains may collaborate with each other to perform an optimal end-to-end path computation.

In the PCE architecture path computation is no longer responsibility of MPLS or GMPLS nodes, therefore, the DS-TE features of virtual circuit switched networks had to be supported by the PCE architecture. Indeed some tools regarding DS-TE are described in [3].

Nonetheless, the PCE technology faces scalability issues when the networks get too large or too complex and a multitude of PCEs and Path Computation Clients (PCCs) from different layers and domains are involved. This scalability issue is mainly related to the large number of control messages exchanged through PCE communication Protocol (PCEP) between PCEs as well as between PCEs and PCCs in order to calculate an optimum end-to-end path.

Researches concerning the scalability of PCEs in large heterogeneous networks have been developed with approaches to multi-domain routing schemes, bulking of requests per message and preemption policies. In our research we focus on the use of the Diff-Serv and TE tools of PCE architecture to verify their impact on scalability issues on a WDM network.

This paper is organized as follows: in Section 2 we briefly describe DS-TE MPLS, and the conforming characteristics of the PCE architecture that are the focus of our study; in Section 3 we explain the simulation technique and the scenarios studied; Section 4 presents results and performance analysis and Section 5 concludes the work.

978-1-4673-2229-4/12/$31.00 ©2012 IEEE

2. BACKGROUND AND RELATED WORK

2.1 MPLS DS-TE

The virtual circuit characteristic of MPLS and GMPLS architectures makes them a perfect playground for applying TE and Diff-Serv.

The MPLS DS-TE in [4] presents the concept of Class-type that differentiates the treatment that would be given to groups of connection belonging to the same class. The Class-Type defines how the available bandwidth will be distributed among the demands.

In addition, a mechanism that may be used as a helpful tool in a MPLS where traffic is treated differentially is pre-emption which can also be applied for TE purposes [5]. The pre-emption mechanism defines, according to traffic and demands priorities, if a demand for a connection may pre-empt a traffic trunk and use its resources. The TE-ClassType is the combination of the connection's ClassType and its pre-emption priority values. The pre-emption values are composed by set-up priority and holding priority, the values vary from 0 to 7, being 0 the highest priority and 7 the lowest priority. A demand for a Label Switched Path (LSP) with smaller numeric value (higher priority) set-up may pre-empt an active LSP with larger numeric value (lower priority) holding priority. These values are set locally by network management, but they must follow a simple rule: the set-up value cannot be of lower priority than the holding value. So if a LSP has a set-up priority of x, its holding priority must be $<= x$ [6]. This rule should be observed in order to prevent that one LSP may be pre-empted by another LSP which it has recently pre-empted causing a continuous mutual chain of pre-emption.

In a network where the pre-emption is an available and applied mechanism, the route calculation must follow some directives in order to achieve less network instability or to better serve higher importance traffic. One possibility on route calculation is to first attempt to calculate the route for all the currently available bandwidth and only then, if there is contention for resources, calculate a route considering the resources that would be available from the eventual tear down LSPs. In the second possibility the best route is calculated for the higher priority set-up value requests always considering eventual released resources [6].

Whenever the tear down of a connection occurs, the topology state information database must be updated with the released resource from that LSP in all the links from the path it was assigned to. This update is of great importance on the path computation because LSP may release enough resources throughout the other links on selected path.

There are two main approaches for a pre-emption mechanism, the locally and the globally aware pre-emption. The local pre-emption only considers information related to link being analysed. Global pre-emption, on the other hand, considers topology information of all links of calculated possible paths in order to optimise the selection of LSPs to be pre-empted, aiming in a lesser as possible number of tear down connections [6].

Global pre-emption may also influence on the occurrence of a cascading effect. A cascading pre-emption effect is when one pre-empted LSP, in order to be reconnected, causes further LSPs pre-emption. Pre-emption cascading causes network instabilities and should be avoided when possible [5].

Moreover, the global pre-emption policy adopted is of extreme importance as it directly impacts the efficiency of the network routing scheme in terms of overall blocking probability. A path selection may be considered optimum if it generates less number of pre-emption or if, in total, from all the subsequent pre-emption, it results in less pre-empted bandwidth [6].

2.2 PCEP control messages

The PCEP defines the communication between PCEs and PCCs and between PCEs. When there is a demand for a new connection and, consequently, a computation regarding the route, it is necessary to allocate the wavelength and the resource availability in the light-path. A PCC sends a Path Computation Request (PCReq) message to the PCE [3]. A PCReq message may contain some requirements used as constraints in the path computation task.

To perform the constraint based path calculation the PCE relies on accessing information from a Traffic Engineering Database (TED) containing updated state of the network's topology.

When a path is successfully computed, a Path Computation Reply (PCRep) is sent by the PCE to the PCC informing the path via an Explicit Route Object (ERO). Besides the route, the PCE also selects the wavelength that will be assigned throughout the path, and informs it in the reply message [7].

If a path could not be calculated under the stipulated constrains, a PCRep containing a NOPATH object is replied and, optionally the attributes that could not be met can be informed [3]. The PCC may, eventually, send a new PCReq for the demand that was not attended maintaining the original attributes requested or altering their values to meet the information comprised by the reply message.

The constant and intense exchange of control messages that may occur in the PCE architecture is one of the reasons for scalability issues of PCEs in large networks. Whenever a constrained path is not computed successfully, a new request for the same demand is likely to be sent to the PCE. Therefore the efficiency of the constrained based path computation impacts on the number of requests generated and processed and, thus, on the final number of control messages exchanged.

Part of this efficiency may be altered by different routing techniques, the frequency of TED update and pre-emption policies approaches.

An approach that can diminish the total number of PCEP messages exchanged is to bundle multiple requests in one PCReq message or multiple path calculation results, EROs or NOPATH objects, in one PCRep message in case the sender or destination, respectively, are the same [3]. In our study we focus on the sequence of request processing and the pre-emption mechanisms and therefore do not consider the bundle requests per message possibility.

2.3 PCE DS-TE tools

In order to properly accommodate the DS-TE characteristics of MPLS using the PCE architecture, the IETF has defined some requirements to be respected by the PCE when computing a path. The requirements for the constrained based path computation are defined as objects present in the PCReq message. The Request Parameters Object (RP) contains a Priority field (Pri field) with possible values varying from 1 to 7, the highest value being equivalent to the highest priority and so on. These values may be used by the PCE scheduler to determine in what order among many requests that specific request would be processed. Therefore the higher the Pri value of a demand more likely it will be processed before other requests.

The Pri field conforms to [8] stipulates that the PCE protocol allows a PCC to determine the priority of a path computation request. This prioritization, however, is supposed to be implemented by autonomous system's local policy following a particular strategy. The network manager must use the priority values in a consistent way throughout the various PCC comprised by the network [3].

The LSP Attributes (LSPA) object contains required attributes for the establishment of a LSP. It contains, among other requirements, the set-up and holding priority attributes of the LSP requested [3]. This requirement is not necessarily considered by a PCE but, when it is, a request for an LSP with higher set-up priority may preempt a current LSP connection with lower holding priority value and use its released resources.

This preemption is performed at the GMPLS control level and coordinated by the Resource Reservation Protocol (RSVP). The PCE, however, calculates the constrained based path regarding this possibility of preemption and considering, as available resources, the resources that would be released from the tear-down connection.

To be re-established the pre-empted path must send a new path computation request which could eventually pre-empt another inferior value holding priority active path, and so on.

Set-up and holding priority attributes are defined locally, and when a path computation request crosses domains these attributes are usually ignored by a different domain PCE. In [9], a proposed Class-Type data base would contain mappings of Class-Types containing preemption priority information according to clients' SLA so path demands would have a priority value equivalence as they cross different domains.

In a network where a preemption policy is adopted all PCEs comprised by such a network must consider the set-up and holding priority values in the LSPA object, and, furthermore, the TED at each PCE must contain information on those values for all active flows.

The three PCE DS-TE values aforementioned are independently set by network management. In our study we have analyze the impact of the Pri field values on the preemption performance and the results from variations of the Pri values in combination with preemption priorities values.

To the best knowledge of the authors no comparison study has been published, so far, regarding the priority of processing combined with preemption values in PCE architecture.

3. CONSIDERED SCENARIO

We tested our proposal in a Java built ad-hoc simulator which network topology was based on the physical topology of the Nobel-German reference network [10] part of the network references collection from SNDlib at the Zusse Institut. The network contains 17 nodes and 26 bi-directional links each of which consisting of 20 channels of 20 Gbps aggregated capacity. The demands and their values were collected from actual connections averaged from random days during the years of 2004 and 2005 [10], it consists of 288 demand matrices with an average of 250 demands each. Each demand matrix is referred to one time interval.

The demand value from the demands matrices is equivalent to the current connection bandwidth usage at time of measurement. The actual resource amount reclaimed per demand was calculated as the difference between the demand value at one interval of time and the same ID demand at subsequent interval. In case the result is negative we consider that resources were released and the release bandwidth amount is aggregated to the current capacity of a link/wavelength that serves a connection with the same ID.

The routing algorithm used is based on OSPF; it applies the Dijkstra's algorithm on the links that have enough available capacity for the demand requirement at hand. The shortest distance is counted as number of links crossed. The wavelength is assigned in a first fit basis. A path is first calculated with the available resource capacity of the links, only then, if there is no resource available, the routing mechanism consider the possible LSPs to preempt and the resources released by those.

When preempting a connection, the resources used by that connection in all links being used by it are released. Preemption is performed locally which means only the information of the link being analyzed at the time is considered, links are preempted in order of processing until the required resource is achieved.

We assume there is only one PCE to calculate all requests and each node work as a PCC with direct communication with the PCE. Each pair source-destination connection request is sent in one request message and each ERO or NOPATH object is also sent in one reply message. The requests are cued for processing in order of arrival in the NO LSPA and NO PRI scenario and in the order determined by the Pri field values in the other scenarios.

Due to the fact it was not possible to find information regarding the priority of service of demands, we have defined values for the set-up priorities for each request with random values uniformly distributed from 0 to 7 inclusive. In order to avoid chain of mutual preemptions, as mentioned in Sec. 2, the holding values were randomly assigned but with values above or equal to the set-up values for the same request.

Every time a path is not successfully computed, the emission of one ERO message is counted. The request that was not served enters a list of non-served requests and will be later re-sent and reprocessed increasing the number of requests per interval of time, therefore increasing the number of messages for the next interval of time. The same happens to preempted demands they enter the non-served requests list and become new requests. At each interval of time, the number of demands in the non-served list is counted and considered as demands not served for that interval. Preempted connections do not receive an ERO reply message but, they are counted as not being served at that period of time.

We started the scenario with zero usage at all links, meaning all links had their individual total capacity of 400 Gbps available at interval of time 0. As requests are being processed, and connections are being established according to demands resource requirements, the network load is increasing. As it is possible to observe in Fig. 1, at around the interval of time 30, out of the 200 from total running time, the network usage increase is less steep. For the purpose of our studies the results presented in Sec. 4 refers to the simulation time interval gap from 100 to 200 time units.

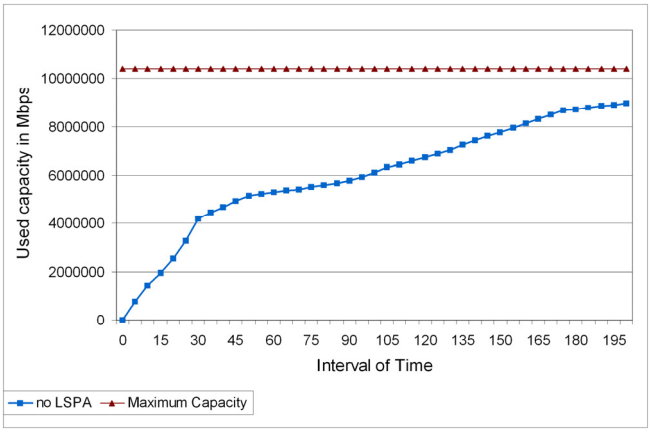

Figure 1. Total network used capacity and total maximum network capacity per interval of time.

Each one of the scenarios evaluated consisted of different combinations of values of the 3 DS-TE components present on the PCE architecture. The description of those scenarios and the abbreviation used in the graphics are as follows:

- No LSPA: no preemption attributes and no priority of request processing are used.
- LSPA: only preemption priorities (set-up and holding priority values) are considered, no priority for request processing is considered.
- Pri = Setup: Pri values are set as the same value as the set-up priority for each demand request.
- Pri = Hold: Pri values are set as the same value as the holding priority for each demand request.
- Pri = Average: Pri values are set as the average of holding and setup priority values for each demand request.
- Pri = Diff: Pri values are set as the difference between the setup and the holding priority values for each demand request.

The metrics used as results for performance evaluation are: number of demands processed per interval of time, number of ERO messages exchanged per interval of time, number of NOPATH messages exchanged per interval of time, number of demands not served per number of demands processed per interval of time, amount of served bandwidth per Set-up priority and holding priority per demands served per interval of time, number of tear-down connections per interval of time amount of tear-down bandwidth per interval of time.

The results are compared for the network performance in the different scenarios on the same topology, same interval of time and the same demand matrix.

4. RESULTS AND ANALYSIS

The simulation results were mapped into the 7 graphs represented by Figs. 2 to 8.

Figure 2 shows the number of extra path computation requests compared to the original number of demands per interval of time for the 6 scenarios. The original number of demands refers to the demands obtained from the reference data and not considering possible regenerated requests from unsuccessful path computations or tear down demands.

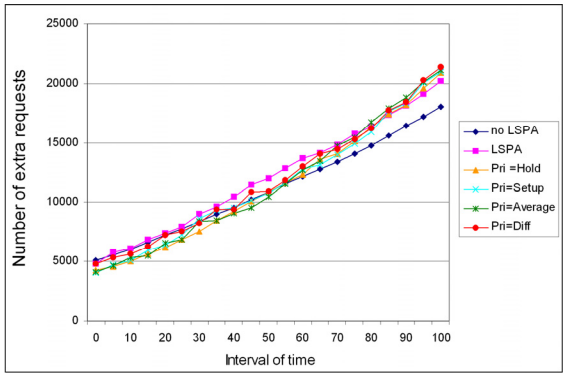

Figure 2. Number of extra requests messages compared to original number of demands per interval of time.

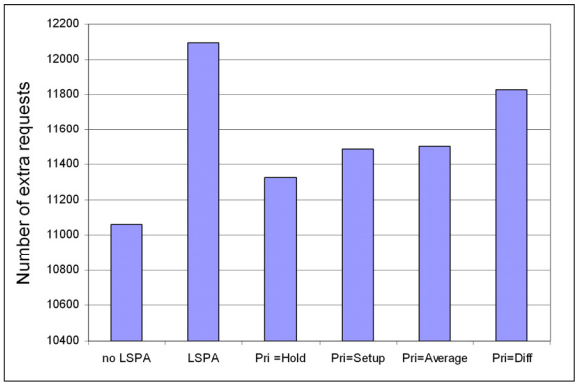

Figure 3. Average number of extra requests messages compared to original number of demands.

It is possible to observe that in all six scenarios the number of over generated requests increases as time overlaps. Such increase in the number of path computation requests is a result of the recurring requests from demands that were previously declined connection because a path with requested attributes could not be computed.

In Fig. 3 the difference of values for each of the scenarios is more prominently illustrated, the graph presents the average number of extra request messages. Due to the frequent preemption of connections the extra number of requests is more accentuated on the scenarios that use preemption priority policies, but this number decreases as different types of ordering of requests processing are implemented. When the Pri field value equals holding priority value, for instance, the number of requests is decreased since less preemption occurs because demands with higher holding priority values are accommodated first.

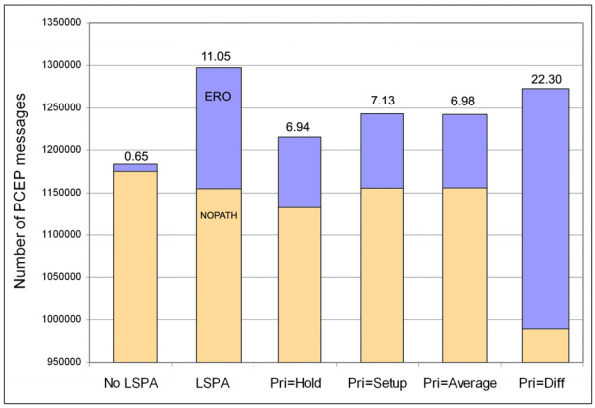

Figure 4. Total number PCEP control messages, ERO messages and NOPATH messages exchanged per scenario. The number above the bar refers to the percentage of ERO messages per total PCEP messages.

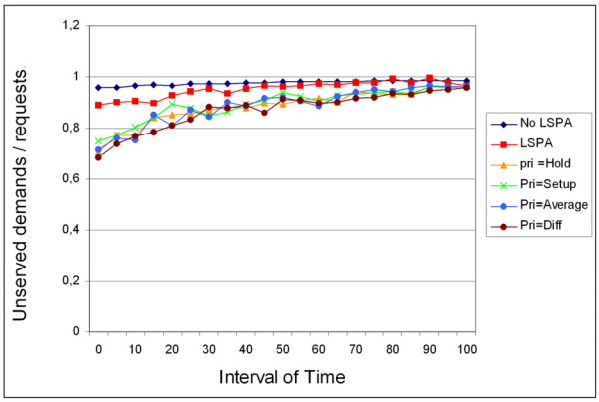

Figure 5. Number of un-served demands per number of requests ratio per interval of time.

An increasing number of reply messages occur as a natural consequence of the higher number of requests. Figure 4 shows the total number of ERO messages (upper area in columns) and NOPATH messages (lower area in columns). The number above each column refers to the percentage of ERO messages for total number of reply messages. The higher percentage of ERO reply messages is a result of the demands that are accommodated due to the preemption of another connection.

The proportion of un-served demands per requests also tends to increase as the network load increases with overlap of time. However, with lower level of network resources usage the scenarios where priority of request processing is implemented the number of un-served demands is lessen as can be observed in Fig. 5.

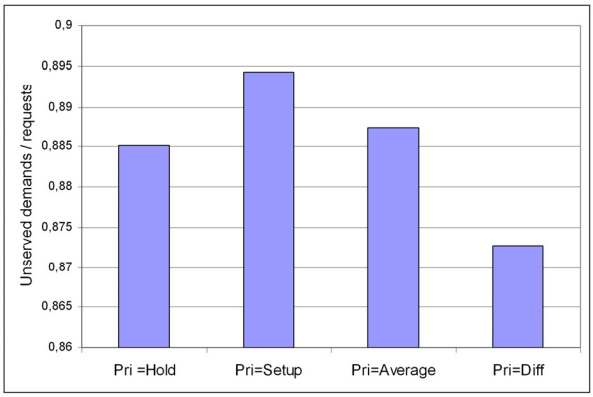

Figure 6. Average un-served demands per number of requests ratio.

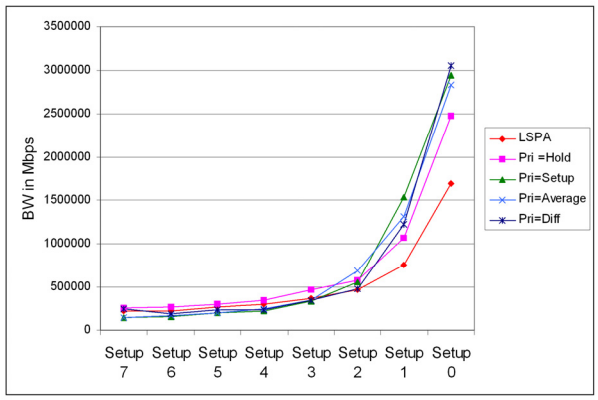

Figure 7. Total bandwidth distribution per set-up priority values.

Figure 6 presents the difference in averaged ratios of demands un-served per number of requests for the different values of priorities of request processing. It is possible to note that the value referred as Pri = Diff has the lower proportion of un-served demands per requests.

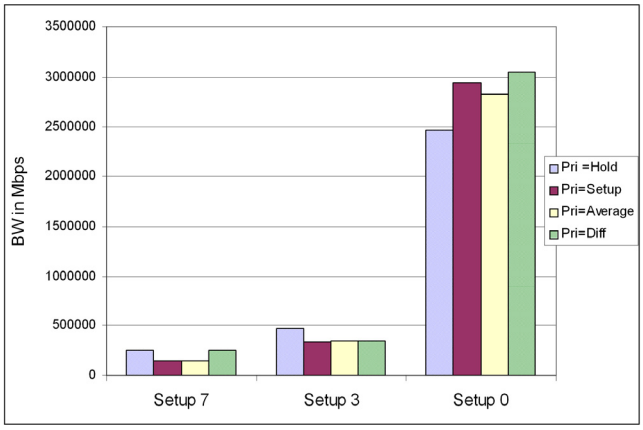

Figure 8. Total bandwidth distribution per Setup priority value.

The policy of ordering priority of request processing also determines how the available bandwidth will be distributed for each class of service. Figure 7 presents the total bandwidth in Mbps assigned to demands belonging to each setup priority value. In this figure it is possible to note that the use of the Pri field tends to be favorable to the assignment of bandwidth to higher priority setup values.

The results presented on Fig. 8 focuses on the scenarios where the Pri field is implemented. The graph shows the total distribution of bandwidth for 3 values of setup priorities, the lower value 7, the higher value 0 and an intermediate value 3. As it would be expected the Pri value equal to the holding priority value assigns less bandwidth to higher setup priority values, whether in the Pri = Diff scenario even more bandwidth is assigned to the highest setup priority value than in the scenario where Pri value equal setup value, this is due to the fact that Pri = Diff scenario serves a higher total number of demands.

5. CONCLUSION

This paper presents the results from the comparison of different uses and values of the current standardized DS-TE tools of PCE architecture applied in a WDM network. The analysis of simulation results shows a considered difference in performance of the diverse combination of values of the Pri field and set-up and holding priorities values that can be adopted in a WDM network.

It was possible to note that the combined use of preemption policies and the Pri field with value set to the difference of set-up value and the holding priority value is the most beneficial combination of values in terms of performance. Once a preemption policy is adopted the adequate value set of the Pri field have direct repercussion on the total number of PCEP messages exchanged and the total number of demands accepted.

From this result we believe it is justified the implement and use of priority policies in requests processing ordering on a PCE architecture whenever available in multi-domain networks in any autonomous system.

For future work in the area of PCE scalability in WDM networks we intend to analyze the impact on numbers of messages exchanged for other MPLS differentiated service mechanism: the bandwidth distribution according

to Class-Types. The use and policies of the PCE Pri-field maybe be used in tandem with different preemption policies, and a study on the combination of the two schemes may be useful to assess the variation of performance results from the diverse combined possibilities.

ACKNOWLEDGMENT

The authors would like to thank Capes Brazilian PhD student scholarship. This work has been partially supported by the Spanish Ministry of Science and Innovation through the DOMINO project (TEC2010-18522).

REFERENCES

[1] S. Dasgupta, J. C. de Oliveira, J.P. Vasseur: Path-computation-element-based architecture for interdomain MPLS/GMPLS traffic engineering: Overview and performance, *IEEE Network*, vol. 21, no. 4, pp. 38-45, Jul. 2007.

[2] L. Atzori, F. D'Andreagiovanni, C. Mannino,T. Onali: An algorithm for routing optimization in Di_Serv-aware MPLS networks, *Antonio Ruberti Technical Reports*, Department of Computer and System Sciences, 2010.

[3] J. Vasseur, J.L. Le Roux, eds.: Path computation element (PCE) communication protocol (PCEP)", *IETF RFC 5440*, Mar. 2009.

[4] F. Le Faucheur , W. Lai: Requirements for support of differentiated services-aware MPLS traffic engineering, *IETF RFC 3564*, Jul. 2003

[5] J. de Oliveira, J. P. Vasseur, L. Chen, and C. Scoglio: Label switched path (LSP) preemption policies for MPLS traffic engineering, *IETF RFC 4829*, Apr. 2007

[6] S. Kaczmarek and K. Nowak: Performance evaluation of preemption algorithms in MPLS networks, *Int. J. Electr. Telecommun.*, vol. 57, no. 2, pp. 169-175, Jul. 2011.

[7] Y. Lee, G. Bernstein, J. Martensson, T. Takeda, T. Tsuritani: PCEP requirements for WSON routing and wavelength assignment, *IETF*, draft-ietf-pce-wson-routing-wavelength-04.txt, Oct. 2011.

[8] J. Ash and J.L. Le Roux: Path computation element communication protocol generic requirements, *IETF RFC 4657* Sep. 2006.

[9] J.S. Dantas, D. Careglio, R.M. Silveira, W.V. Ruggiero, J. Sole-Pareta: PCE algorithm for traffic grooming and QoS in multi-layer / multi-domain IP over WDM networks, in *Proc. 13th Int. Conf. Transp. Opt. Netw. (ICTON 2011)*, Stockholm, Sweden, Jun. 2011.

[10] S. Orlowski, M. Pióro, A. Tomaszewski, R. Wessäly: SNDlib 1.0-survivable network design library, in *Proc. 3rd Int. Netw. Optim. Conf. (INOC 2007)*, Spa, Belgium, Apr. 2007.