

Network Working Group  
Internet Draft  
Expires: March 2007

Peng He  
Gregor v. Bochmann  
University of Ottawa  
September 21, 2006

A Novel Framework for Inter-area MPLS Optimal Routing  
draft-he-ccamp-optimal-routing-00.txt

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on March 21, 2007.

Abstract

We propose a novel framework for inter-area MPLS optimal routing. The key to our proposal lies in deploying an overlaid star optical network in the OSPF backbone area and introducing the concept of "virtual area border routers" (v-ABRs). Compared with other proposals, our framework can provide globally optimized inter-area routing and has very good compatibility to existing traditional IP/MPLS routers.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

Table of Contents

- 1. Terminology.....2
- 2. Introduction.....3
  - 2.1. What's The Problem.....3
  - 2.2. Current Approaches.....4
- 3. The Novel Framework for Inter-area MPLS Optimal Routing.....5
  - 3.1. The Basic Idea.....5
    - 3.1.1. Overview of Agile All-Photonic Network (AAPN).....5
    - 3.1.2. Deploying AAPN in the OSPF Backbone Area.....6
  - 3.2. The Routing-Info Component.....7
  - 3.3. The Path Computation Component.....8
  - 3.4. The Signalling Component.....8
    - 3.4.1. Path Message in the Head-End Area.....9
    - 3.4.2. Path Message in the Backbone Area.....9
    - 3.4.3. Path Message in the Tail-end Area.....10
    - 3.4.4. Resv Message.....10
- 4. Further Considerations.....11
- 5. Generalization of the Proposed Routing Framework.....12
- 6. Security Considerations.....13
- 7. IANA Considerations.....13
- 8. Conclusions.....13
- 9. Acknowledgments.....13
- 10. References.....14
- Author's Addresses.....15
- Intellectual Property Statement.....15
- Disclaimer of Validity.....16
- Copyright Statement.....16
- Acknowledgment.....16

1. Terminology

- OSPF: Open Shortest Path First
- CSPF: Constraint-based Shortest Path First
- LSP: Label Switched Path
- LSR: Label Switched Router

LSDB: Link State Database

RSVP: Resource Reservation Protocol

AAPN: Agile All-Photonic Network

## 2. Introduction

Currently, several carriers have multi-area networks, and many other carriers that are still using a single IGP area may have to migrate to a multi-area environment as their network grows and approaches the single area scalability limits [RFC4105]. Hence, it would be useful and meaningful to extend current MPLS TE (Traffic Engineering) capabilities across IGP areas to support inter-area resources optimization. That is why RFC4105 [RFC4105] was recently published to define detailed requirements for inter-area MPLS traffic engineering and ask for solutions.

In this draft, we consider the Open Shortest Path First (OSPF) [RFC2328, RFC3630] IP routing protocol, which is commonly used for routing within a single administrative domain and adopted by MPLS and GMPLS (Generalized MPLS) (with extensions).

### 2.1. What's The Problem

OSPF supports large networks through multiple OSPF areas: one backbone area (Area0) surrounded by non-backbone areas. Area border routers (ABR) are located at the border between the backbone and the non-backbone areas.

An inter-area connection normally starts in a non-backbone area, traverses a backbone area, and terminates in another non-backbone area. MPLS TE mechanisms that have been deployed today by many carriers are limited to a single IGP area and can not be expanded to multi-areas directly. The limitation comes more from the routing and path computation components than from the signalling component. This is basically because the OSPF/OSPF-TE hierarchy limits topology visibility of head-end LSRs (Label Switch Routers) to their area, and consequently head-end LSRs can no longer run a CSPF algorithm to compute the shortest constrained path to the tail-end, as CSPF requires the whole topology information in order to compute an end-to-end shortest constrained path.

For an example, Figure 1 shows a common multi-area network and we suppose RT1 in Area1 is the source node while RT6 in Area2 is the destination. Generally speaking, a non-backbone area (e.g., Area1 in Figure 1) often has multiple ABRs (existing points). One ABR might be

much closer to the destination of a requested MPLS connection than another. Because the head-end node does not have the entire topology, it does not know which ABR is the best choice. In Figure 1, how could R1 choose an optimum ABR in Area1 to the destination RT6? Through local optimization, R1 may select ABR2 to be on the path, but how does ABR2 know what the best path is to go to RT6? Although local optimization can be done in each of the respective areas along the inter-area path (RT1 to RT6), the simple summation of the three local optimizations does not necessarily lead to a global optimization. What many carriers want is to optimize their resources as a whole. Therefore, the question of how to implement inter-area routing with global optimization guarantee is a key issue in inter-area traffic engineering.

## 2.2. Current Approaches

Most current approaches for inter-area routing center on the "how-to" issue, that is, how to find out an inter-area route (not optimal and not dynamic) and how to build up this path through the inter-area signalling process. The per-area approach uses a two-step method to compute an inter-area route: find out a "loose inter-area route" first through topology aggregation/abstraction, then resolve the loose route into a strict path, area by area. Actually, this per area approach would always lead to sub-optimal resource utilization. Another segment approach divides an inter-area path into two segments, one in Area1 and one in Area0 & Area2 (see Figure 1 for a rough look). Optimal routing of the 1st segment is done first by the head-end LSR; then based on the 1st segment, a far-end ABR (e.g., ABR5) computes the 2nd segment. Obviously, this approach can not achieve global route optimization either. PCE (Path Computation Element)-based approaches [PCE-ARC] are the only category of approaches that can provide global optimization. But this needs building up an independent overlay external PCE network that covers all the areas, and defining and implementing many associated new protocols.

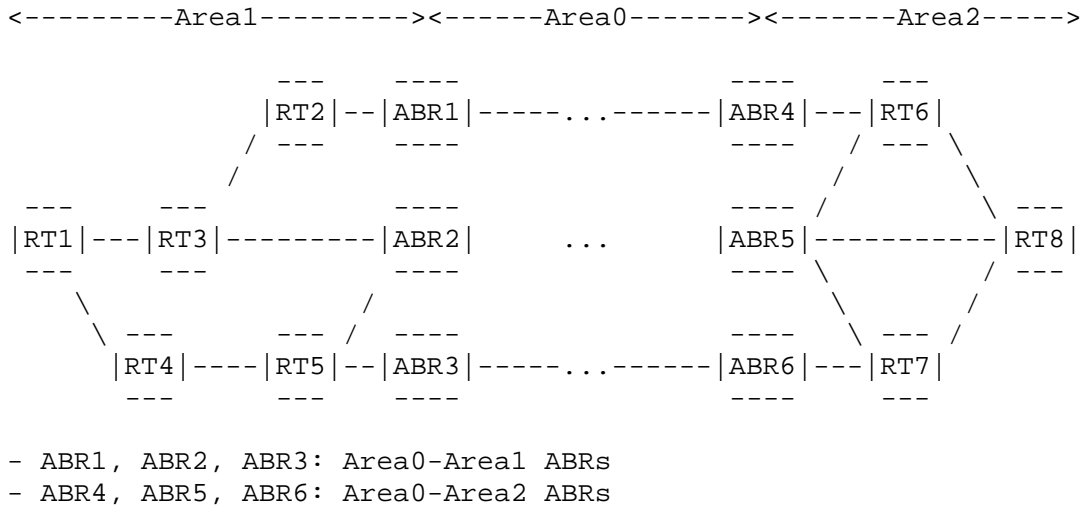


Figure 1 : A common network with multiple OSPF areas.

### 3. The Novel Framework for Inter-area MPLS Optimal Routing

#### 3.1. The Basic Idea

In this draft, we propose a novel framework that deploys an overlaid star optical network in the backbone area so as to implement global resource optimization in an efficient and distributed manner. The star topology can help to achieve globally-optimized routing (as explained later), and the overlaid star can provide high reliability.

AAPN (Agile All-Photonic Networks) is a representative example of overlaid star optical networks. Hence we use AAPN to represent overlaid star networks in the following context of this draft.

##### 3.1.1. Overview of Agile All-Photonic Network (AAPN)

As shown in Figure 2, an AAPN [AAPN-ARC] consists of a number of hybrid photonic/electronic edge nodes connected together via several (not less than two) load-balancing core nodes and optical fibers to form an overlaid star topology. Note that there are no direct physical links among these cores. By introducing concentrating devices, AAPN can support up to 1024 edge nodes [AAPN-TOP]. Each core node contains a stack of bufferless transparent photonic space switches (one for each wavelength). A scheduler at each core node is used to dynamically allocate timeslots over the various wavelengths to each edge node. An edge node contains a separate buffer for the traffic destined to each of the other edge nodes. Traffic aggregation is performed in these buffers, where packets are collected together

in fixed-size slots (sometimes called bursts) that are then transmitted as single units across the AAPN via optical links. At the destination edge node the slots are partitioned, with reassembly as necessary, into the original packets that are sent to the outside routers. The term "agility" in AAPN describes its ability to deploy bandwidth on demand at fine granularity, which radically increases network efficiency and brings to the user much higher performance at reduced cost.

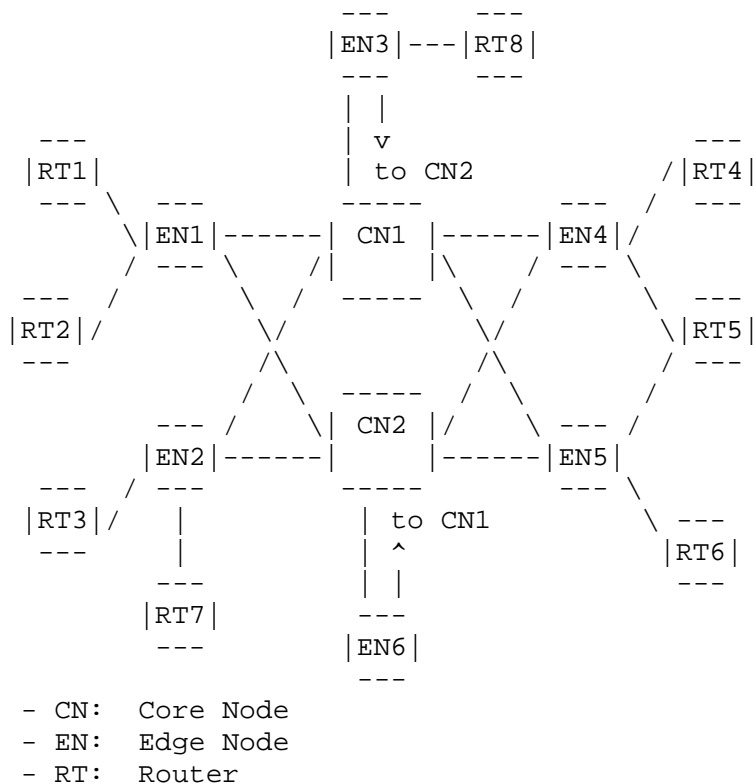


Figure 2 : Agile All-Photonic Network (AAPN) Overlaid Star Topology.

### 3.1.2. Deploying AAPN in the OSPF Backbone Area

AAPN is more suitable to be used in multi-area network environment due to its agility at the core and large capacity. The direct and natural way to deploy an AAPN in the OSPF backbone area is like this: the core nodes are located in the middle of Area0 and the edge nodes act as ABRs at the border between Area0 and other non-backbone areas. However, in this scheme, inter-area routing with global optimization still can not be guaranteed. Therefore, we adopt a novel way of deploying OSPF over an AAPN that interconnects several OSPF areas to

provide such guarantee, as shown in Figure 3. This "overlapped" OSPF architecture is fundamental for our inter-area MPLS optimal routing framework. In details, our proposed framework consists of three main components, namely the routing-info, path computation and signalling components, as described in the following sections.

Note: due to the symmetric architecture of AAPN (see Fig. 2), we use the "bundle" [RFC4201] concept to further reduce the overhead traffic to the outside. That is, all the links from one edge node to the core nodes are exported as one TE link. Similarly, the overlaid core nodes in AAPN are, if necessary, exported as one core node, named as "the core" (see Figure 3).

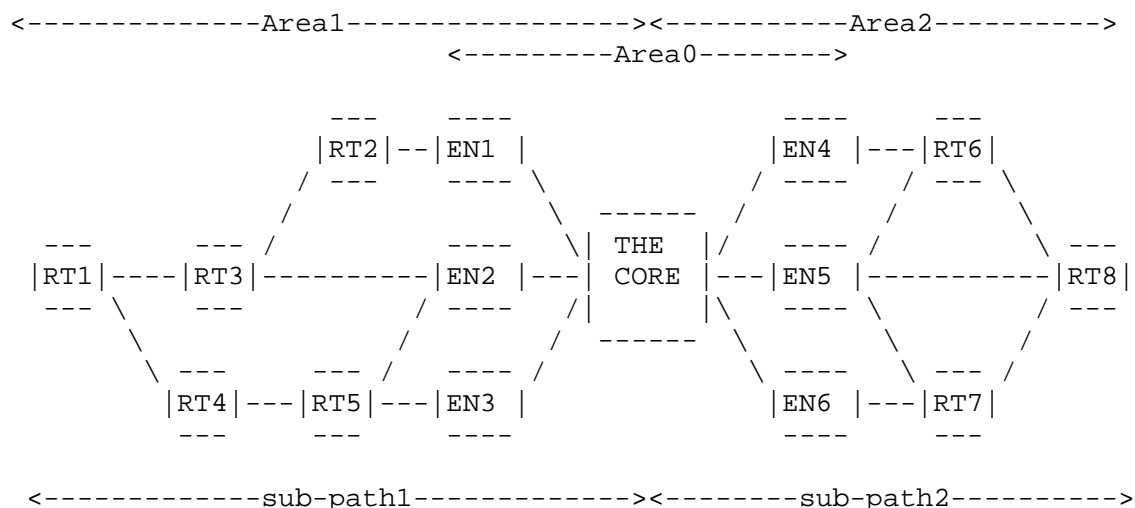


Figure 3 : Our framework deploys AAPN as backbone area (Area0).

### 3.2. The Routing-Info Component

This component is responsible for the discovery of the TE topology, which can be ensured through OSPF [RFC2328] and OSPF-TE [rfc3630]. As shown in Figure 3, after deploying AAPN as backbone area, we further expand the OSPF non-backbone areas a little step so that there is an overlap between Area0 and each expanded non-backbone area. Then the AAPN edge nodes located in the overlap (see the top of Figure for the overlap), together with their direct TE links to the core, and the associated part of the core, belong to both the Area0 and a non-backbone area. In such a scenario, legacy routers in a non-backbone area see related AAPN edge nodes as normal internal IP/MPLS routers, see the AAPN TE links as normal internal links and see the associated part of the core as the (only) ABR of their non-backbone area. In other words, a legacy router sees what it can see in its area about

the core as an ABR, which we call a virtual-ABR (v-ABR). For each legacy router in an expanded non-backbone area, the exchange and distribution of routing/TE information is just like in any other standard OSPF/OSPF-TE area.

While within the Area0 (within the AAPN), the AAPN edge nodes that belong to the same non-backbone area can be organized as a big virtual router and one edge node in each virtual router is selected as the head of this virtual router. Then the area-specific reachability information is exchanged among these heads and distributed to each edge node per area and then outside routers. Therefore, it is the head edge node that actually performs the functions of the virtual-ABR, that is, distributing area-specified reachability information. Note that only the reachability (not TE) information, which is enough for our framework, is exchanged among virtual routers, and hence among non-backbone areas.

### 3.3. The Path Computation Component

In our framework, an inter-area LSP can be considered consisting of two segments as shown in the bottom Figure 3: one in the head-end (expanded) area (sub-path1) and one in the tail-end (expanded) area (sub-path2). The core connects these two segments/sub-LSPs to form a complete inter-area LSP.

The most interesting thing is that local routing optimization (through CSPF) with both of these two sub-LSPs can lead naturally to a globally-optimized inter-area LSP. As seen in Figure 3, this is due to the particular star topology of the AAPN architecture and the load-sharing core nodes that can be viewed as one single virtual router (v-ABR) from the outside.

The local routing optimization in the head-end area can be performed by the source LSR, which takes the TE topology and LSP constraints as input. While in the tailend area, local routing optimization is done by one of the edge nodes in the area (see next section for details). Obviously, dynamic inter-area routing can be implemented in our proposed framework.

### 3.4. The Signalling Component

This component is responsible for the establishment of the LSP along the computed path. In Figure 3, consider the case that a source LSR (RT1) in Area1 wants to set up a LSP to a destination LSR (RT8) in Area2. RT1 must first compute an optimized path to the virtual-ABR of Area1 (v-ABR1) through CSPF, and then signal this establishment



request to the network. RSVP with TE extension (RSVP-TE) [RFC2205, RFC3209] can be used as the signalling protocol.

#### 3.4.1. Path Message in the Head-End Area

As shown in Figure 4, RT1 starts the signalling process by creating a RSVP Path message with two objects inserted, namely LABEL\_REQUEST Object (LRO) to request a label binding for the path, and EXPLICIT\_ROUTE object(ERO) to indicate the computed explicit path (with one sub-object per hop). However, RT1 has to use the loose ERO sub-objects for the hops outside Area1. In Figure 3, the ERO specifies the explicit path as RT1->RT3->EN2->v-ABR1->RT8, where RT8 is a loose ERO sub-object. Then, RT1 sends the Path message to the next hop defined in the ERO, which is RT3.

RT3 (a non AAPN node) receives the Path message and processes it as defined by RSVP-TE:

1. Checks the message format to make sure everything is OK,
2. Performs admission control to check the required bandwidth,
3. Stores the "path state" from the Path message in its local Path State Block (PSB) to be used by the reverse-routing function, and
4. If successful, deletes the 1st sub-object (itself) in the ERO and forwards the Path message according to the new 1st sub-object (next hop) in the ERO, in our case, EN2.

#### 3.4.2. Path Message in the Backbone Area

EN2, an AAPN edge node, receives the Path message from RT3 and checks the contained ERO. If EN2 finds that the IP address of the 2nd sub-object in the ERO is a v-ABR and the 3rd sub-object (with the loose attribute) is beyond Area1, then EN2 has the task of resolving the loose sub-object into strict ones. In our case, there is one loose sub-object, RT8, which represents the destination of the requested LSP. Although EN2 can not find a strict path from v-ABR1 to RT8 by itself, it knows who can. First, by checking the inter-area reachability information and internal parameters, EN2 finds out which group of edge nodes (also which associated v-ABR) locates in the same non-backbone area as RT8. In Figure 4, these are EN4, EN5 and EN6 (v-ABR2). Second, it selects an edge node among them randomly, e.g., EN4. In the third step, EN2 removes the first two sub-objects (itself and v-ABR #1) from the ERO of the original received Path message, and inserts v-ABR2 at the top, then forwards the modified Path message to EN4.

When EN4 receives the Path message and finds that the 1st sub-object in the received ERO is v-ABR2, together with a loose second sub-object, RT8, it knows that it should find an explicit path between these two sub-objects. As shown in Figure 3, EN4 is capable to do the resolving work because EN4 and RT8 reside in the same expanded area, Area2. EN4 finds the optimized explicit path v-ABR2->EN5->RT8. EN4 then replaces the ERO object in the received Path message with a new ERO object that stores the resolved explicit route (EN5->RT8). Finally, EN4 forwards the new modified Path message to EN5 as if it were forwarded from EN2 by using EN2's data (IP address, etc.). We call this process a Path message handoff. At the same time, EN5 also sends an acknowledge message (containing the resolved path) to EN2 (Figure 5). From the above handoff process, we can see that only the area-specific reachability (not TE) information needs to be exchanged among areas. In our proposal, TE information is organized within each area.

#### 3.4.3. Path Message in the Tail-end Area

The edge node EN5 receives the Path message and believes it is from EN2. Since all the sub-objects in the received ERO are strict, EN5 processes this Path message in a standard way, just as RT3 did in Area1, and then forwards the processed Path message to RT8.

#### 3.4.4. Resv Message

When the destination, RT8, gets the Path message, it responds to this establishment request by sending a RSVP Resv message. The purpose of this response is to have all routers along the path perform the Call Admission Control (CAC), make the necessary bandwidth reservations and distribute the label binding to the upstream router.

As defined in standard RSVP-TE [RFC3209], the label is distributed using the Label Object in the Resv message. The labels sent upstream become the output labels for the routers receiving the label object. The label that a router issues upstream become the inbound label used as the lookup into the hardware output tag table. The reservation-specific information is stored in the local reservation state block (RSB) of each router.

When the AAPN edge node EN5 receives the Resv message from downstream (RT8), it starts internal AAPN signalling to ask the core to set-up a connection from EN2 to EN5 (omit v-ABR1&2). If bandwidth is available for this connection, the core informs both EN2 and EN5. EN5 then sends a Resv message to EN1. Note that it is an internal choice of the AAPN to select a label, for instance, a timeslot number, a wavelength, or a normal MPLS label.



the 1st half or by an edge node for the 2nd half of LSP) or globally by the head-end LSR (end-to-end re-optimization).

Regarding inter-area QoS, there is not much work left. Current single area QoS mechanisms [RFC2475, RFC 3270] can be expanded directly to multiple areas and to AAPN.

As seen in Figure 3, our proposal keeps OSPF's hierarchical structure and just expands non-backbone areas a little. Hence the scalability of our proposal is as good as OSPF/OSPF-TE [RFC2328, RFC3630].

Our proposed framework also supports diversely-Routing of inter-area TE LSPs, as required in RFC4105 [RFC4105]. As shown in Figure 3, diversely routing can be deployed in Area 1 and 2 independently and optimally, and then the AAPN residing in the backbone area connects them together.

## 5. Generalization of the Proposed Routing Framework

The routing concepts discussed in this draft are based on the assumption that there are a number of edge nodes (that are connected with other routers - through traditional Internet technology - and belonging to the same OSPF area) and these edge nodes can establish optical connections between one another in an agile manner and can adjust the bandwidth of each connection in an agile manner according to the varying bandwidth that is required by the IP traffic. We think any agile optical switching technology (burst switching, TDM (Time-Division-Multiplexing), or routed wavelength (with less bandwidth flexibility)) may be used.

Our proposal is not limited to AAPNs, it is actually applicable in a much larger context. The fundamental ideas abstracted from our proposal are: (1) a "load-symmetrical" network (optical mostly) as backbone, (2) overlap between backbone and non-backbone areas, and (3) virtual-ABR. A load-symmetrical optical network is a network that can provide one or several optical connections for each edge node pair (source-destination pair) and the load among the several optical connections of each edge node pair is balanced. Hence the "bundle" concept [5] can be used and a single core can represent the optical network topology.

Note that "load-symmetrical" does not mean the loads among any distinct edge node pairs are balanced; it only refers to the load-balancing within the connections of one edge node pair. Load symmetrical networks do not need to have a symmetrical physical network topology, although a symmetrical physical network topology

(such as AAPN] and PON (Passive Optical Networks)) can be made load-symmetrical easily.

## 6. Security Considerations

This document does not introduce new security issues beyond those inherent in MPLS TE [RFC3209, rfc3630].

## 7. IANA Considerations

This informational document makes no requests for IANA action.

## 8. Conclusions

Based on deploying a load-symmetrical network (such as an overlaid star optical network (AAPN)) in multi-area networks, we propose a novel framework that aims to implement optimal inter-area MPLS routing.

Compared with other inter-area routing proposals, our proposal has two distinguishing characteristics:

1. Our proposal can provide globally-optimized inter-area routing;
2. There will be no change, hardware or software, on existing traditional IP/MPLS routers in the peripheral OSPF areas to implement our proposal.

Furthermore, our proposal is not limited to AAPNs; it is actually applicable to any load-symmetrical (optical) network with arbitrary physical network topology. Indeed, our proposal can be considered as a highly competitive candidate that has the potential to become a total solution to Inter-area MPLS Traffic Engineering [RFC4105].

## 9. Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and industrial and government partners, through the Agile All-Photonic Networks (AAPN) Research Network. This work is part of an AAPN Partnered Research Project sponsored by TELUS. The authors thank Rainer Iraschko for useful discussions.

## 10. References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC2328] J. Moy, "OSPF Version 2", RFC 2328, April, 1998.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Service", RFC 2475, December 1998.
- [RFC3209] Awduche, D., et al. "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December, 2001.
- [RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.
- [RFC3630] Katz, D., Yeung, D., Kompella, K., "Traffic Engineering Extensions to OSPF Version 2", RFC 3630, September, 2003.
- [RFC4105] Le Roux, et al., "Requirements for Inter-Area MPLS Traffic Engineering", RFC 4105, June 2005.
- [RFC4201] K. Kompella et al. "Link Bundling in MPLS Traffic Engineering (TE)" RFC 4201, October, 2005.
- [PCE-ARC] Farrel, A., "A Path Computation Element (PCE) Based Architecture", draft-ietf-pce-architecture-04 (work in progress), January 2006.
- [AAPN-ARC] G. v. Bochmann, M.J. Coates, T. Hall, L. Mason, R. Vickers and O. Yang, "The Agile All-Photonic Network: An architectural outline", Proc. Queen's University, Biennial Symposium on Communications, 2004, pp.217-218.
- [AAPN-TOP] L.G. Mason, A. Vinokurov, N. Zhao and D. Plant, "Topological Design and Dimensioning of Agile All Photonic Networks", Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 50, Issue 2 (February 2006), Pages: 268 - 287, 2006.

Author's Addresses

Peng He  
School of Information Technology and Engineering (SITE)  
University of Ottawa  
800 King Edward Avenue  
Ottawa, Ontario  
K1N 6N5 Canada

Phone: 1-613-5625800 ext. 2191  
Email: penghe@site.uottawa.ca

Gregor v. Bochmann  
School of Information Technology and Engineering (SITE)  
University of Ottawa  
800 King Edward Avenue  
Ottawa, Ontario  
K1N 6N5 Canada

Phone: 1-613-5625800 ext. 6205  
Email: bochmann@site.uOttawa.ca

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at [ietf-ipr@ietf.org](mailto:ietf-ipr@ietf.org).

#### Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

#### Copyright Statement

Copyright (C) The Internet Society (2006).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

#### Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.



