Protection and Routing over Agile All-Photonic Networks (AAPN)

Peng He

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the PhD degree in name of program

Electrical Engineering

Ottawa-Carleton Institute for Electrical and
Computer Engineering
School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

University of Ottawa

Abstract

Routing and Protection over Agile All-Optical Networks (AAPN)

by Peng He

The term “agility” in optical networks describes the ability to deploy bandwidth on demand at fine granularity that allows carriers to deploy services rapidly. An overlaid-star all-photonic WDM network, called the Agile All-Photonic Network (AAPN), can provide such agility through multiplexing over each wavelength in the time domain. The AAPN consists of a number of hybrid photonic/electronic edge nodes connected together via several load-balancing bufferless transparent core nodes and optical fibers to form an overlaid star topology. An AAPN can potentially provide an efficient high bandwidth/high performance core transport network for carriers. Hence, it is very important to design and position AAPN to support widely-deployed IP/MPLS architecture and protocols. This thesis deals with routing and protection of MPLS flows over AAPNs, especially in multi-domain (OSPF multi-area or inter-AS) network environments.

For the routing problem, several AAPN configurations seen by the Internet routers are proposed to solve the scalability issue when deploying AAPN within one IP/MPLS network. Furthermore, a novel inter-area routing framework is proposed which can provide dynamic and optimal inter-area routing in an efficient and scalable way with full backward compatibility with existing OSPF routers. In addition, an AAPN-based Internet Exchange (AIX) architecture with traffic engineering capacity is also developed to provide inter-AS optimal routing.

For the protection problem, instead of using link or path protection, an inter-area shared segment protection approach is proposed, which can take advantage of the above inter-area routing framework. Through sharing, the network resources are used in an efficient way. Through segment-based protection, the recovery time is reduced in case of failures.

By generalizing and extending this AAPN-based routing and protection framework, a general inter-domain (AS or area) traffic engineering architecture, called Star-TE, is proposed. Star-TE satisfies the requirements for inter-area and inter-AS traffic engineering defined by IETF in RFC 4105 and RFC 4216.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYM

AAPN	*Agile All-Photonic Network*
ABR	*Area Border Router*
AS	*Autonomous System*
ASBR	*AS Border Router*
BGP	*Border Gateway Protocol*
BRPC	*Backward Recursive Path Computation procedure*
CE	*Customer Edge Equipment*
CSPF	*Constraint Shortest Path First*
GCO	*Global Concurrent Optimization*
GMPLS	*Generalized MPLS*
LDP	*Label Distribution Protocol*
LSP	*Label Switched Path*
LSR	*Label Switched Router*
MPLS	*Multi-Protocol Label Switch*
OBS	*Optical Burst Switch*
OSI	*Open Systems Interconnection*
OSPF	*Open Shortest Path First*
OTDM	*Optical Time Division Multiplex*
OXC	*Optical Cross-connect*
PBB	*Provider Backbone Bridges*
PBB-TE	*PBB Traffic Engineering*
PBT	*Provider Backbone Transport*
PCE	*Path Computation Element*
PE	*Provider Edge Equipment*
QoS	*Quality of Service*
RSVP	*Resource Reservation Protocol*
SSP	*Shared Segment Protection*
TDM	*Time Division Multiplex*
T-MPLS	*Transport MPLS*
TE	*Traffic Engineering*
VPN	*Virtual Private Network*
VSPT	*Virtual Shortest Path Tree*
WDM	*Wavelength Division Multiplex*

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my supervisor, Dr. Gregor v. Bochmann, for his guidance, support and encouragement throughout my study at the University of Ottawa.

I would like to specially thank Dr. Peng Cheng, Dr. Jun Zheng and Qiao Ying for their insightful comments to and countless discussions on this thesis work. I am also indebted to colleagues in the AAPN project and the Distributed Systems Research group for their help and support.

I would like to thank Bilan my mother, Dengxuan my father, He Wen my father-in-law, Lier Chen my mother-in-law, Jie Wen my sister-in-law, my super cute niece Ella, Jing Wen my wife and my son Albert. I love you all infinitely.

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

1. Introduction

1.1. Problem Statement

Most telecommunication carriers, including Telus, Bell Canada, AT&T, MCI and British Telecom, are migrating to an IP based converged network for provisioning multi-services (data, voice, video, etc.). In such IP networks, Multi-Protocol Label Switching (MPLS) is adopted to enable Traffic Engineering (TE) and support Virtual Private Networks (VPN). Together with Diffserv, MPLS can also provide Quality of Service (QoS) support.

An emerging overlaid-star all-photonic WDM network, called the Agile All-Photonic Network (AAPN) [Bochmann2004, Mason2005, Bochmann2007], can potentially provide an efficient high-bandwidth core transport network solution for carriers. Hence, it is very important to design and position AAPN to support widely-deployed IP/MPLS architecture and protocols.

The AAPN consists of a number of hybrid photonic/electronic edge nodes connected together via several load-balancing buffer-less transparent core nodes (photonic space switches) and optical fibers to form an overlaid star topology. By introducing concentrating devices, AAPN can support up to 1024 edge nodes [Mason2005]. The term “agility” in AAPN describes its ability to deploy bandwidth on demand at fine granularity (time-slot instead of wavelength), which radically increases network efficiency and brings to the user much higher performance at reduced cost [Bochmann2004].

Since the AAPN provides transparent optical transmission between all pairs of edge nodes, the architecture, as seen by the Internet routers within the surrounding networks would normally be a completely interconnected mesh. Hence the straightforward use of the OSPF (Open Shortest Path First) routing protocol used for IP and MPLS leads to scalability problems when deploying a large AAPN within an IP/MPLS network. For instance, with 1000 edge nodes in an AAPN, this would involve 10^6 links, which is a number much too large to be handled by normal Internet routers. Hence in this thesis, we study first how to solve the scalability issues for deploying an AAPN within a single IP network.

AAPN is actually more suitable to be used in multi-domain (OSPF area or AS) network environment due to its agility at the core and large capacity. Hence in this thesis, we study mainly how to deploy AAPN in multi-domain networks. In such environments, we focus on inter-domain traffic engineering, that is, how to provide optimal inter-domain routing and associated protection in a scalable and efficient way for MPLS flows over the AAPN.

Note that we study the Open Shortest Path First (OSPF) [RFC 2328] and Border Gateway Protocol (BGP) [RFC 1771] IP routing protocols, which are commonly used for routing within and among administrative domains, respectively. In the rest of this thesis, we use the term "domain" to represent an OSPF area or an AS (Autonomous System). The definitions of OSPF area and AS are as follows:

Area: OSPF area. OSPF is a hierarchical routing protocol that supports large networks through multiple OSPF areas: one backbone area (Area 0) surrounded by non-backbone areas. Area border routers (ABR) are located at the borders between the backbone and the non-backbone areas, and distribute summarized routing information between the areas.

Autonomous System (AS): controlled by one service provider; routers within an AS communicate routing information to each other using an Interior Gateway Protocol (IGP), namely OSPF, or IS-IS. An AS shares routing information with other AS using BGP. Autonomous System Border Routers (ASBRs) are used to connect to another AS via one or more physical links that interconnect ASes.

1.2. Motivation and Objective

The work in this thesis is motivated by the goal to find simple and efficient mechanisms for deploying an AAPN in an existing legacy IP/MPLS. We consider two closely related problems: *protection* and *routing*. Furthermore, for both of these problems we should consider various network environments, e.g., single-layer vs. multi-layer, intra-domain vs. inter-domain. Since efficient routing in a large multi-domain network supporting QoS requirements and global fast recovery from failures is still an unsolved research topic, we hope that our results provide promising solutions in this area.

1.3. List of Contributions in this Thesis

The main contributions of this thesis are the followings.

1. A routing architecture for MPLS flows over AAPN that solves the scalability issue when deploying a large AAPN in single-area networks. (Refer to Section 3.1)
2. An inter-area MPLS dynamic routing framework that provides optimal inter-area routing in a scalability and efficient way on the basis of deploying an AAPN as the backbone area in multi-area networks. (Refer to Section 3.2)
3. Inter-area protection schemes that provide shared (efficient) and segmented (fast) protection for MPLS flows over AAPN under both the commonly-used single-failure assumption and a weakened single-failure (multiple failures) assumption we defined for multi-area networks. (Refer to Chapter 4)
4. A novel Internet eXchange (IX) Architecture based on an AAPN that can guarantee the inter-AS (inter-ISP) optimal routing while keeping scalability and the confidentiality for the TE information in multi-domain environments. (Refer to Section 3.3)
5. A general inter-domain (AS or area) traffic engineering architecture, called Star-TE, that generalizes and extends our AAPN-based routing and protection work, and satisfies the requirements for inter-area and inter-AS traffic engineering defined by IETF in RFC 4105 and RFC 4216. (Refer to Section 5.2)
6. A segmented PCE (Path Computation Element) architecture that avoids the inherent scalability and robustness issues in the original IETF PCE architecture and can be considered as a promising inter-domain traffic engineering solution for large-scale meshed multi-domain networks. (Refer to Section 5.3.3)
7. Other contributions include: delay-performance analyses for AAPN with two proposed bandwidth allocation schemes (called first-fit scheme and first-fit with random scheme), a novel analytical performance model for time-space-switched all-optical networks, and optimization analysis of deploying passive optical time slot interchanger (POTSI) in All-Optical Network. (Refer to Section 6.2)

1.4. Outline of Thesis

The thesis is organized as follows. Chapter 2 gives an introduction to AAPN and the literature review of various routing and protection schemes for MPLS over optical networks. Chapter 3 focuses on routing architectures for MPLS flows over AAPN in single-area, multi-area, multi-AS scenarios. The performance of proposed inter-domain traffic engineering framework is assessed by discrete event simulation. Chapter 4 proposes the inter-area shared segment protection schemes that provide efficient and fast protection for MPLS flows over AAPN. Chapter 5 generalizes and extends our AAPN-based work into a general inter-domain traffic engineering architecture, called Star-TE, and presents more applications of Star-TE in traffic engineering with further discussion. Chapter 6 concludes the thesis and identifies areas for further research.

2. Background

2.1. AAPN Overview

As shown in Fig. 1, an Agile All-Photonic Network (AAPN) [Bochmann2004, Mason2005, Bochmann2007] consists of a number of hybrid photonic/electronic edge nodes connected together via several (at least two) load-balancing core nodes and optical fibers to form an overlaid star topology. By introducing concentrator devices, AAPN can support a large number of edge nodes (up to 1024) [Mason2006]. Each core node contains a stack of bufferless transparent photonic space switches (one for each wavelength). In order to avoid optical memory and optical header recognition (hence no E-O-E conversion within the core switches), as required by certain forms of burst switching, AAPN uses synchronous slot-by-slot switching with fixed-sized slots (e.g., 10µs per timeslot) and the arrival of the slots at the input ports of the switch must be synchronized with the slot switching period controlled by the switch. If in a mesh network, where a slot transmitted by an edge node must possibly traverse several switching nodes, the propagation delay on the links between several switching nodes must be precisely adjusted to be a multiple of the slot duration. Since this is difficult to realize, AAPN adopted a star topology with a single core switch. In this case, the only synchronization requirement is that the edge node transmits the next slot at such a time that it arrives at the core switch in time for the beginning of a slot period. This can be realized by a relatively simple synchronization protocol between the edge nodes and the core. Furthermore, in order to cope with the failure of a core node, and in order to increase the overall capacity of the network, AAPN finally uses an architecture of overlaid stars, as shown in Fig. 2.1.

A scheduler at each core node is used to dynamically allocate timeslots over the various wavelengths to each edge node. Each edge node contains a separate buffer for the traffic destined to each of the other edge nodes. In these buffers, packets are collected together in fixed-size slots 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 (e.g., timeslots instead of a whole wavelength, in the order of 100Mbps instead of 10 Gpbs), which radically increases network efficiency and brings to the user much higher performance at reduced cost.

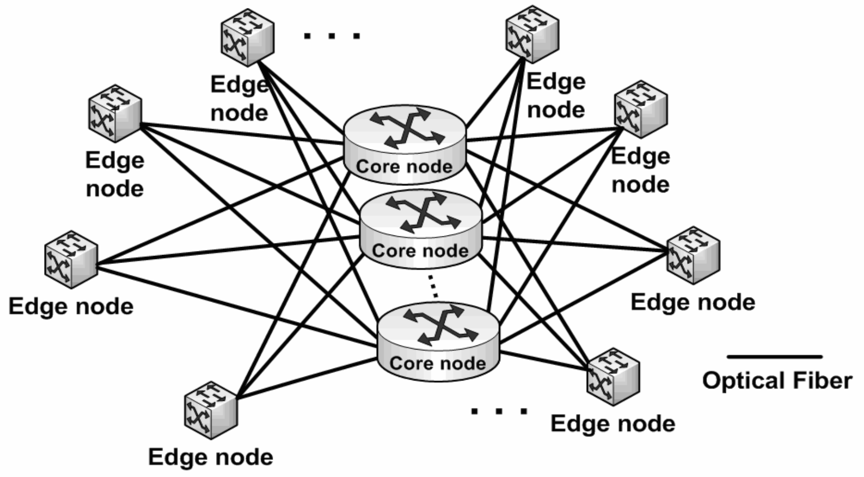


Figure 2.1: AAPN Overlaid-Star Topology

Generally speaking, an AAPN is a wavelength-division-multiplexed (WDM) network that consists of several overlaid stars formed by edge nodes that aggregate traffic, interconnected by bufferless optical core nodes that perform fast switching in order to provide bandwidth allocation in sub-wavelength granularity. On the other hand, an AAPN can also be viewed as a distributed switch with potentially large geographical coverage. It contains four key ingredients: (1) (optical) switching core: rapidly reconfigurable switching at the core, (2) intelligent edge: control and routing functionality concentrated at the edge nodes that surround the switching core, (3) AAPN internal optical fiber connecting edge and core nodes. AAPN adopted OTDM (optical time domain multiplexing) on fibers. (4) Overlaid star topology for reliability and increased bandwidth.

The main competition for an AAPN, on the future networking market would probably be electronic Internet routers or switches. However, we believe that the conceptual simplicity of an AAPN and the power of optical transparent switching without E-O-E conversion will give some advantage to the AAPN approach. We also note that the optical lightpaths provided by an AAPN are protocol and rate-independent at the physical level, which facilitates network evolution.

2.2. MPLS Overview

MPLS [RFC 3031] operates at an OSI (Open Systems Interconnection) model layer between Layer-2 (data link layer) and Layer-3 (network layer), and thus is often referred to as a "Layer-2.5" protocol. MPLS is a packet label-based switching technique. Packets are assigned a 20-bit label as they enter a MPLS capable IP network. Subsequent packet treatment in the network is based on the label only, e.g., these MPLS-labeled packets are switched after a Label Lookup/Switch instead of a lookup into the IP table of a router.

The labeling of a packet allows the use of advanced forwarding techniques. A packet entering the network at a particular router can be labeled differently than the same packet entering the network at a different router. As a result, some kind of policy routing can easily be made. Since MPLS decouples forwarding from routing, it is able to support a large variety of routing policies that are either difficult or impossible to support with just conventional network layer forwarding.

MPLS uses the Label Distribution Protocol (LDP) (RFC 3036) or RSVP (RFC 2205) to exchange the label and LSP (Label Switched Path) binding between Label Switching Routers (LSRs). A LSP is defined as a sequence of labels from an ingress LSR to an egress LSR. LSPs are very similar to unidirectional ATM virtual circuits. The route taken by a LSP between two LSRs can be the same as the conventional network layer route, or the sender LSR can specify an explicit route for this LSP (an explicit route is specified as a sequence of hops rather than being determined by conventional layer-three routing algorithms on a hop-by-hop basis). Thus, apart from conventional IP routing facilities, MPLS can use the routing technique called explicit routing, which can support policy routing and traffic engineering. An explicit route needs to be specified at the time that labels are assigned and does not have to be specified with each IP packet.

MPLS is in use in large "IP Only" networks, and is standardized by IETF in RFC 3031. In practice, MPLS is mainly used to forward IP datagram and Ethernet traffic. Major applications of MPLS are traffic engineering and MPLS VPN (Visual Private Network).

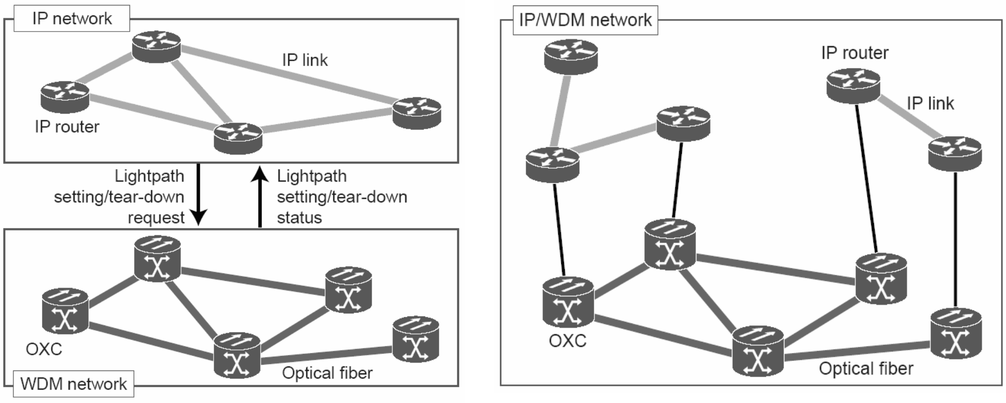
2.3. Previous Work on Routing of IP/MPLS over WDM Optical Networks

Future networks will typically be carrying Internet (IP) traffic – enhanced with MPLS functionality – over optical networks that provide data transmission between IP/MPLS routers. This section reviews the previous work on routing of IP/MPLS over optical networks, both for intra- and inter-domain (area or AS) scenarios.

2.3.1. IP/MPLS Routing over WDM Networks: Intra-Domain Scenario

The basic components of the routing functionality are the collection of information on the network topology and a routing algorithm that determines a route for given source, destination and QoS (Quality of Service) requirements. In a general environment of IP/MPLS over optical networks, the optical network is composed of OXCs (Optical Cross-connect) interconnected by fiber links, and IP/MPLS routers are connected to the OXCs through wavelength ports comprising of optical transmitters and receivers. Under this framework, which is considered by most of the literature, OXCs and fiber links between them form a *physical topology* in the optical layer; the IP/MPLS routers and lightpaths set up between them form a *virtual or logical topology* in the IP/MPLS layer.

There are generally three different inter-networking models between these two kinds of topologies: namely *peer*, *augmented*, and *overlay* models for IP/MPLS over optical networks [Koo2004]. One of the most important differences among these models is the type of information shared between the IP/MPLS layer and the optical layer. In the overlay model (Fig. 2.2a), no network information is shared between these layers. Hence routing is done separately in the two layers with their own signaling and control planes. Under this network model, the IP layer and WDM layer work in a client-server model. In the peer model (Fig. 2.2b), the topology and other network information (e.g., routing information and link state information) are shared between the layers, and a unified routing mechanism, e.g., GMPLS (Generalized MPLS), controls the whole network. The augmented model is a compromise between the peer model and the overlay model. This means that the augmented model shares some part of the information, such as the reachability information, between the layers according to certain agreements, and the two layers are managed independently like in the overlay model.



(a) Overlay Model

(b) Peer Model

Figure 2.2: Models of IP/MPLS over Optical Networks [Koizumi2006]

2.3.1.1. Routing in the overlay model

Two similar and simple routing schemes, based on the overlay model, were proposed in [Ye2001, Koo2004]. When a LSP request arrives, the network first finds a route for the request over the residual capacity on the current logical topology. If no available route exists, it then tries to open a new lightpath directly (one hop) [Ye2001] or indirectly (more than one hop) [Koo2004] between source and destination LSRs of the request.

As a variant of the overlay model, the study in [Koizumi2005] built up the IP/MPLS logical topology by using both the logical links and “virtual link”. A virtual-link is a special logical link that is not configured as a lightpath, but can be activated as a lightpath according to the request. If a virtual-link is selected as part of a LSP, the lightpath corresponding to the virtual link is established right away.

2.3.1.2. Routing in the peer model

In the peer model, an integrated graph or auxiliary graph, where both the physical and logical links coexist, is usually adopted to solve the routing/TE problems. The auxiliary graph model has also been used in a dynamic grooming study [Zhu2003]. The graph is usually a layered graph, in which each layer represents one wavelength.

The authors in [Kodialam2001, Zheng2003, Liu2004] provided similar peer-mode routing schemes: the idea is to assign the cost values to both the logical and physical links in the integrated graph first, and then calculate a least cost route for the request using the shortest path algorithm (e.g., Dijkstra's algorithm). The only difference between various routing schemes is the definition of the cost function. It is reported in [Koo2004] that a simple scheme [Zheng2003] just using the physical hops as the cost function performs better (in terms of LSP blocking probability) than the one in [Kodialam2001] which adopted a much more complex cost function. A similar phenomenon was also observed in [Liu2004].

2.3.1.3. Routing in the augmented model

In [Koo2004], an augmented-model-based routing scheme was proposed. It was assumed in this scheme that only a summary of capacity information from the optical layer is shared with the IP/MPLS layer. The novelty of this scheme is considering the constraint on the number of (optical) ports per LSR in the WDM layer, which is usually assumed to be infinite in other schemes.

Loosely speaking, it is believed that the augmented model has the potential to benefit from the advantages of both the overlay and peer models. However, there is still no promising routing algorithm(s) proposed for this model, and there is still little understanding of what kind of information would be most helpful for making LSP routing decisions [Koo2004].

2.3.2. IP/MPLS Routing over WDM Networks: Inter-Domain Scenario

2.3.2.1. Why Inter-area/AS Routing

Currently, several carriers have multi-area networks, and many other carriers that are still using a single OSPF area may have to migrate to a multi-area environment as their networks grow and approach the single-area scalability limit. Hence, it would be useful and meaningful to extend the current MPLS TE capabilities, which are still limited within one OSPF area, across several OSPF areas to support inter-area resource optimization. That is why RFC 4105 was recently published to define detailed requirements for inter-area MPLS traffic engineering and to ask for solutions.

The practical interest in inter-AS end-to-end MPLS routing with guaranteed QoS is also increasing. This is due to the surging VoIP traffic and large VPNs between sites hosted by different carriers (Inter-domain/inter-provider scenario). And some Internet service providers (ISPs) maintain different legacy ASes after corporate acquisitions or mergers (intra-provider inter-AS scenario) [Ricciato2005]. That is why RFC 4216 was published to define detailed requirements for MPLS inter-AS traffic engineering requirements and to ask for solutions.

2.3.2.2. Why It is Difficult

However, having efficient routing in a large multi-area/AS network that obeys QoS requirements is a yet unsolved research topic. The major challenge for inter-area routing is the ***scalability*** of routing information. To make routing scalable, the TE information that an area advertises about itself and learns from other areas must be very small. This limits the TE visibility of the head-end LSR essentially to only its own area, and consequently it can no longer run a CSPF (Constrained Shortest Path First) algorithm to compute the shortest path to the tail-end, as the CSPF algorithm requires the whole TE topology information. The scalability issue is even severer in inter-AS routing as multi-AS networks are usually larger networks. Besides, another big challenge for inter-AS routing is the ***confidentiality*** of traffic engineering information. This is because an AS is normally operated by one ISP who competes against other ISPs and hence does not want to reveal its sensitive internal information concerning topology, link capacities, traffic volumes, and routing parameters to its competitors. Only a very limited amount of data (e.g., AS reachability information) is made available to other ASes by means of the BGP protocol.

Scalability and confidentiality limit the quantity and content of inter-domain cooperation. It is unknown yet, even theoretically [Feamster2003, Shrimali2007, Tomaszewski2007], to what extent limited cooperation among domains can still provide global optimality. [Winnick2002] suggests conducting negotiations on traffic engineering information between a pair of neighboring ASes though BGP protocol extensions, but did not clarify the contents and frequency of such negotiations. [Shrimali2007] adopts the method of "Nash Bargaining and Decomposition" to deal with cooperation between domains, but they focus only on two domains and no encouraging results are found. [Tomaszewski2007] presents a decomposable ILP mathematical model for the multi-domain routing optimization problem, but their model is still at a very theoretical level, and does not work well in most cases. In addition, in such a large real-time process, traditional linear programming techniques may not be suitable for flexible trading of constraints and objectives in dynamic networks.

Concerning inter-domain protection, especially shared protection (to save the network resource), it is even more difficult than the inter-domain routing. This is because shared protection needs more TE information than routing. A big part of the TE information for shared protection is about the "protection relationship" among links, e.g., how many links protect this link, and how many links (or nodes) are protected by this individual link, and who are they. This information is part of the confidential information in each domain and should not be disclosed to other domains. On the other hand, beyond the scalability and confidentiality, one extra factor that any inter-domain protection approach has to consider is the speed of protection, as customers usually require fast protection. Multi-domain networks are usually large-scale networks, in which end-to-end path protection that works well in single-domain networks, may not be suitable.

Most of the current proposed practical inter-domain routing approaches fall into three categories, namely TE (traffic engineering) abstraction/aggregation, PCE-based architecture, and Per-domain approaches with or without partially-extending-TE-visibility, as explained in the followings.

2.3.2.3. TE Abstraction/Aggregation

TE abstraction/aggregation approaches [Zhu2003, Saad2004, Thiongane2005, Guo2007] usually adopt a two-step/layer approach to compute an inter-domain route: first find out a "loose inter-domain route" through topology aggregation/abstraction, then resolve the loose route into a strict path, domain by domain. In general, the TE abstraction techniques refer to either virtual link/tunnel/trees or virtual nodes. The virtual link/tunnel/tree is like "You can reach this destination along this link with these characteristics". The virtual node represents a sub-network as a virtual switch, but it might be deceptive since there is no easy way to advertise its "limited cross-connect capabilities" (due to internal blocking in the abstracted sub-network). Actually, the two-step approach would lead to sub-optimal resource utilization and a precise topology aggregation/abstraction always needs very frequent updates which further raise scalability issues. Hence this category of approaches is considered somewhat unpractical and does not gain strong support in the IETF.

2.3.2.4. PCE-based approaches proposed by IETF

As defined in [RFC 4655], a Path Computation Element (PCE) is an entity, a node or a process, that is capable of computing a TE LSP based on a given network graph (together with computational constraints) in response to a path computation request from a LSR or an application (e.g., from a network management system). A domain could have one or more PCEs to compute intra-domain paths in either a centralized or a distributed (cooperative) fashion. For inter-domain path computation, the cooperation among the PCEs of different domains is a must. Currently, the PCE-based architecture can compute optimal inter-domain TE LSPs if the domain sequence to traverse is given [Vasseur2007b]. And the optimality is guaranteed through each PCE by advertising a candidate optimal path for each entry point, which leads to an exhaustive search.

In [Matsuura2007], the authors propose a hierarchically distributed PCE (HDPCE) architecture to cooperatively create appropriate IPTV trees for multi-domain users. It actually implements in the PCE the previously-discussed (Section 2.3.2.3) two-step/layer TE abstraction/aggregation approach by PCE.

PCE-based approaches separate path computation from signaling, and hence can support more complex routing (i.e., multi-layer, multi-domain). But this is at the cost of building up a kind of multi-domain path computation plane (or overlay network), composed of the PCEs and associated PCE–PCE communication processes, and corresponding modifications on current routing/signaling protocols, e.g., OSPF-TE, RSVP-TE. Meanwhile, the performance of PCE-based routing mechanism is still unclear. The potential scalability issues of the PCE-based architecture, when deployed in large-scale networks, have not yet been considered.

2.3.2.5. Per-domain approach and Partially-extending-TE-visibility

Per-domain approach is a straight-forward method to compute inter-domain paths in a domain by domain fashion (see Fig.2.3), usually triggered by a signaling process. Partially-extending-TE-visibility is to extend the LSRs’ TE-visibility further to the inter-domain links. This is to improve the chance of successful signaling along the next domain in case of resource shortage or unsatisfied constraints on inter-domain connectivity and to reduce the signaling crankback. As illustrated in Fig.2.3, there are two kinds of partial-visibility-extensions, namely *outgoing-view-extension* [Chen2007] if only the TE information of the outgoing direction connectivity from the domain borders is advertised inside, and *incoming-view-extension* [Miyamura2004a, Miyamura2004b, Otani2007] if only the TE information of the incoming direction connectivity to the domain borders is advertised inside. Actually, the per-domain approach or either of the two extended per-domain approaches separates an inter-domain path into several segments (see Fig. 2.3) and then does the path computation segment by segment. Hence this category of approaches can not guarantee the findings of optimal inter-domain paths [Vasseur2007a].

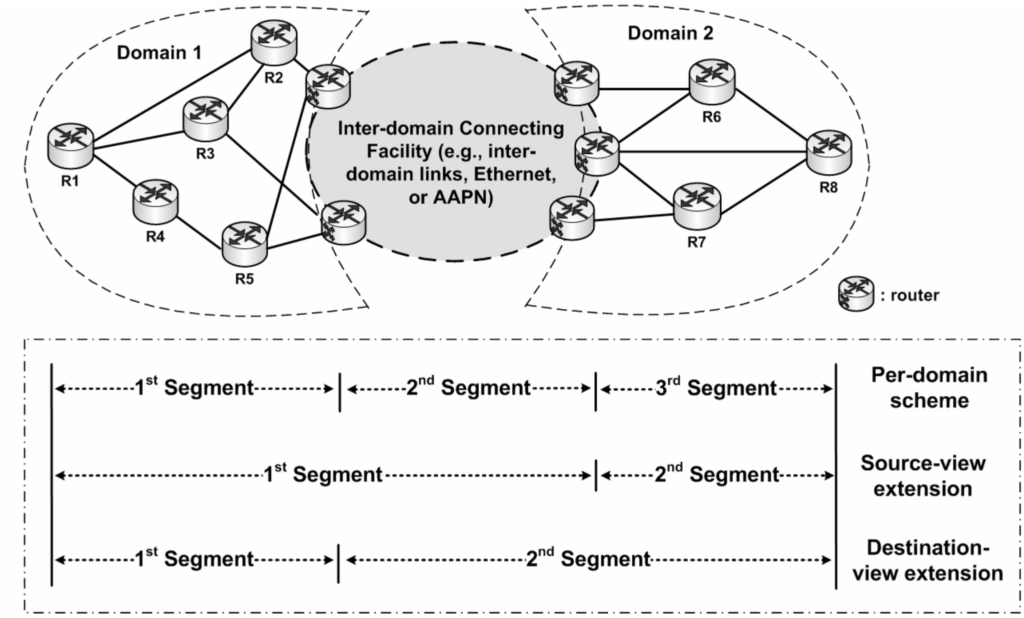


Figure 2.3: Per-domain approach and two partially-extending-TE-visibility approaches.

2.4. Previous Work on Resilience of IP/MPLS over WDM Optical Networks

Although networks are now becoming more and more reliable (due to the software and hardware upgrading), there are still frequent network failures that are becoming a cause for concern. There is no doubt that current and future optical networks will carry a tremendous amount of traffic, including voice and video flows, along with regular Internet traffic. Hence a failure in an optical network will have a disastrous effect: affecting millions or billions of users. This is the main reason why resilience (recovering from failure) remains a major research topic for next generation optical networks.

2.4.1. Resilience Techniques Overview

The resilience techniques can be classified, in general, into two categories: *protection* and *restoration*.

In protection, backup path(s) are established and spare capacity is reserved for them at the time the working path is set up, in another words, before any network failures happen. In restoration, only after a network failure has taken place, backup path(s) are computed based on updated topology and resource information, and established in real-time while the spare capacity is allocated to them dynamically.

Obviously, protection will yield the fastest recovery and highest availability but may cost more resources, since it pre-allocates spare capacity for pre-established backup paths to react to the failures. On the other hand, restoration has high resource efficiency but it is typically slower than protection and can not guarantee full restoration of the affected traffic. This is because real-time backup path establishment may involve dynamic route calculation and spare capacity allocation, which may not always be successful due to the dynamic nature of the network traffic (especially in heavy network load scenario).

As shown in Figure 2.4, the resilience techniques can also be grouped by their protected scope: either link/node-, or segment- or end-to-end (E2E) path-oriented. In path-oriented methods, traffic is recovered along backup paths (physically disjointed from the corresponding working paths) between source and destination node pairs for each connection that traverses the failed links. In link/node/segment-oriented methods, traffic is recovered around failed links/nodes. On the other hand, protection can be also classified according to its accessibility to the back-up network resources: either 1+1 dedicated, or *M:N* shared (*N* working paths share the resource of *M* backup paths). In 1+1 dedicated protection, the traffic is transported simultaneously on both working path and backup path; whereas in *M:N* shared protection, the traffic is switched over from a working path to a backup path only after the occurrence of a failure. Hence, the backup resources can be used by some low priority pre-emptive traffic in the absence of a failure.

Restoration can be further classified according to the computation style of the backup paths: it could be in real-time right after the failure, or pre-planned. This means that the backup paths are calculated already but the corresponding resource is not yet reserved.

When a failure occurs, one of the backup paths will be selected based on the topology and resource information at that time.

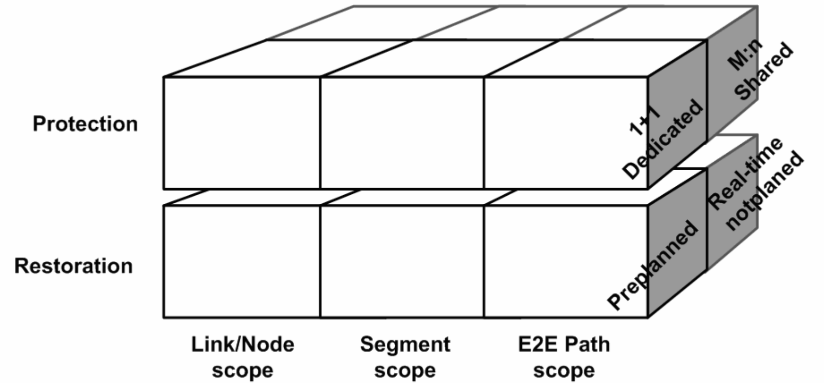
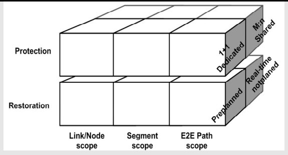
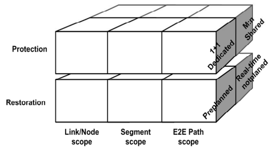
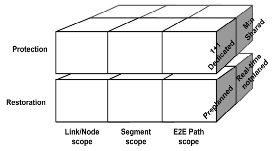
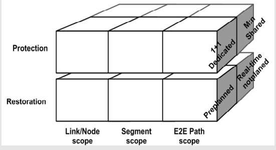
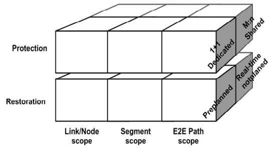
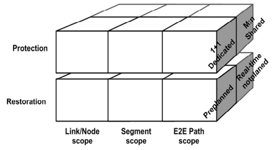
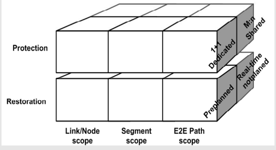
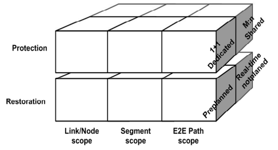
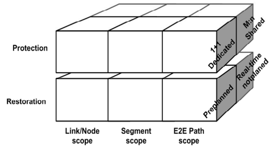
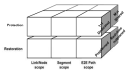


Figure 2.4: Classification of Resilience Techniques (single layer, single area)

TABLE 2.1: OVERALL CLASSIFICATION OF RECOVERY TECHNIQUES

	Intra-area	Inter-area	Inter-AS
MPLS Layer Recovery			
Optical Layer Recovery			
Multi-Layer Recovery			
(Note:  in each cell represents the Resilience Techniques illustrated in Figure 2.4)			

Besides, as seen in Table 2.1, depending on which network layer(s) performs the protection and restoration, resilience techniques can be classified as single-layer (e.g., MPLS layer, or optical layer) and multi-layer (e.g., MPLS over Optical) mechanisms. In addition, based on the applicable-range, resilience techniques can also be classified into intra-area, multi-area, or inter-domain (or multi-domain). Only the techniques corresponding to the first column of Table 2.1 (intra-area) have been studied extensively.

2.4.2. Intra-Domain MPLS Layer Recovery

Due to the relatively slow fault recovery mechanism of IP, which may take several seconds to minutes to recover from a failure [Larrabeiti2005], G/MPLS technology is involved to provide fast recovery around a failure point in tens of milliseconds. This is comparable to SDH/SONET recovery, and hence makes G/MPLS satisfy the reliability requirements of optical networks [Huang2004].

Classification of MPLS layer recovery mechanisms follows the same spirit of Fig. 2.3, but it can be simplified into two criteria [RFC 3469]: rerouting (restoration-type) vs. protection switching (protection-type), and local (link/node scope) repair vs. global (end-to-end path scope) repair.

The intent of global (path) repair is to protect against the failure of any link or node or any segment of the working path, whereas the intent of local repair is to protect against a single link or node failure. In global repair, the recovery (rerouting or protection switching) is always activated on an end-to-end basis, irrespective of where a failure occurs. But in local repair, it is usually the upstream node of the failure point who initiates the recovery actions; hence the amount of overall recovery time is minimized. In the MPLS layer, generally speaking, recovery schemes based on rerouting mechanisms prefer local repair [Chem1999, Yoon2001], while schemes based on global repair prefer protection switching.

The study in [Huang2002] presented an end-to-end MPLS path protection scheme and used signaling from the point of failure to inform the upstream LSRs that a path has failed. The novelty of this scheme is to establish a Reverse Notification Tree (RNT) to distribute the fault and recovery notifications efficiently to all ingress nodes that may be hidden due to label merging operations along the path. The RNT is normally a multipoint-to-point tree, where the PSLs (Path Switch LSR) become the leaves of the trees and an appropriated chosen PML (Path Merge LSR) is the root. Unlike schemes that treat each individual LSP independently, the RNT allows for only one (or a small number of) signaling message on the shared segments of all the LSPs [Huang2004].

Haskin’s recovery scheme [Haskin2000] adopted protection switching with global repair. When a LSR detects a failure, it switches the (working) incoming traffic by linking the upstream portion of the working path to the downstream portion of the recovery path. As presented in Figure 6, when LSR7 fails, the working traffic is rerouted along the recovery path LSR5–3–1–2–4–6–8–9; whereas in Huang’s scheme above [Huang2002], the recovery path is LSR1–2–4–6–8–9. The advantages of Haskin’s scheme are fast recovery time and almost no packet loss during link/node failure, while the main drawback is inefficiency in terms of bandwidth utilization.

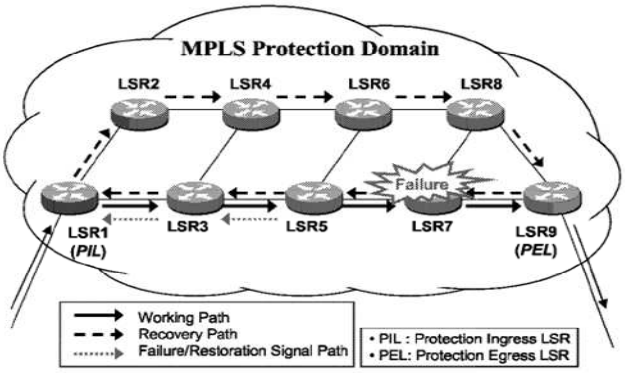


Figure 2.5: Path recovery: Haskin’s scheme [Ahn2002].

Ahn’s scheme [Ahn2002] uses a rerouting model with local repair. The novelty of the scheme is introducing the concept of a “candidate-PML” (Path Merge LSR), which is any LSR along the working path that can be used as a PML. The basic idea is that when a network failure occurs, the immediate-upstream LSR of the failure starts to calculate the least-cost recovery path of all possible alternative paths between itself and each downstream candidate-PML so as to increase the recovery probability.

Following the same line of thinking, the authors in [Colle2001] proposed a rerouting mechanism called fast topology-driven constraint-based rerouting (FTCR), where the first available upstream LSR rather than the original LSR is responsible for rerouting. While in [Zheng2004], a multi-initiation rerouting mechanism was proposed. In this mechanism, each of the nodes along the working path initiates a restoration process upon the detection or notification of a link failure. In each of the processes, the initiating node attempts to dynamically establish a backup path to work around the failed link. The destination node, acting as a coordinator, finally chooses one of these multiple restoration processes.

Besides, [Pan2000, RFC 3469] introduced several extensions to RSVP-TE so as to enable the signaling for local protection to establish, maintain, and switchover bypass tunnels. The proposed bypass tunnel is defined as a LSP that backs up a set of working LSP-segments by making use of a label stack. This local protection (also referred as *Local Fast Reroute* in [Huang2004]) is based on pre-established bypass tunnels; hence the efficiency of resource utilization is poor (you need to pre-establish many bypass tunnels) although it has fast recovery time.

2.4.3. Intra-Domain Recovery in Meshed Optical Networks

In general, classification of optical-layer resilience mechanism/techniques follows the same spirit of Fig. 2.3. However, as we discussed previously, protection-based mechanisms usually result in a high resource redundancy and poor network throughput (especially in the dedicated-protection as in the case of SONET/SDH networks) although it provides 100% recovery from any single failure. In restoration-based mechanisms, on the other hand, the network resource utilization is efficient but at the price of longer restoration time and the risk of no available backup path when a failure happens. As we know, the resource allocation in the optical layer is coarse, e.g., a lightpath occupies one wavelength. Hence most recovery schemes in the optical layer prefer shared protection that can achieve 100% restorability while significantly reducing the redundancy in terms of network capacity consumption. Restoration techniques are desired that can decrease the restoration time while increasing the restorability at the same time. Optimization methods such as Integer Linear Programming (ILP) are usually employed to solve the problem of shared protection resource provisioning. Note that in shared protection, various backup paths are allowed to share resources subject to the shared risk link group (SRLG) constraint. According to the SRLG constraint, resources cannot be shared by backup paths whose working paths can fail simultaneously. The single-link failure model is usually assumed as fault model.

2.4.3.1. Path-based Shared Protection

The intent of path-based shared protection in optical networks is to find spare resources that are physically disjoint from the corresponding working paths, over which the working traffic could be rerouted to backup paths during any failure of network elements along the working paths. The task of finding shared protection paths for a specific group of working paths is referred to as “survivable routing” (technically, it is diverse routing since it needs to find out the disjoint working and backup path pairs) or “spare capacity allocation”. These two are the same because survivable routing looks for the least cost route (working path or working and backup path pair), while the efficient spare capacity allocation is usually considered in the definition of the cost function of the routing algorithms.

The straightforward scheme for diverse routing is a two-step approach, also referred as active path first (APF) [Xu2002, Tapolcai2003]. The APF scheme derives the working path using Dijkstra’s algorithm at the first step, then finds out the backup path in the residual network topology. However, APF only guarantees the best path for the working path, not the working-and-backup path pair. Sometimes in some network topologies, there is even no existing disjoint backup path for a derived working path. This means the working and backup paths should be computed jointly.

In [Tapolcai2003], an improved two-step-approach, iterative two-step-approach (ITSA), was proposed. Basically, ITSA runs the above two-step-approach iteratively: inspecting *k*th-shortest paths between each S-D pair one after the other until the least cost working and shared protection path pair is found. The link cost metric is defined to be roughly proportional to the required bandwidth of the connection.

Further considering shared spare capacity allocation, a novel link cost metric is defined in [Xu2002]: the cost for the working path to take link *j* is determined by the maximum spare capacity among all the other links in the network protecting link *j*. The purpose of this new link metric is to force the working path to traverse through links yielding smaller maximum non-sharable spare capacity. Hence the backup path can have a better chance of finding sharable spare capacity.

The study in [Ho2004b] provides a modified Dijkstra’s algorithm, namely Maximum Likelihood Relaxation (MLR), which jointly takes the link cost and the number of links (hops) with sufficient sharable spare capacity into consideration. It is reported in

[Ho2004b,c] that ITSA can achieve the best performance in terms of blocking probability, while PBC takes the least amount of computation time. MLR initiates a compromise between these two.

Besides, a *M:N* style shared protection is proposed in [Maach2004] for time slotted optical networks. This scheme relies on the fact that the working traffic of a request may be composed of many flows going through different physical paths. Therefore traffic granularity to be protected is reduced and the protection could be achieved more efficiently and cost effectively by provisioning one backup (e.g., *M*=1) and sharing it among these flows (*N* flows). A very similar idea, called “Self-Protecting Multipaths”, was provided in [Menth2004].

2.4.3.2. Segment-based Shared Protection

Actually, segment-based protection can be considered as a generalized form of link-based protection (if a segment equals to a link) and path-based protection (if a segment equals to a path). A segment of the working path together with its corresponding backup path is called a protection domain (Figure 2.6). Normally, compared with path-based shared protection, segment-based shared protection can improve capacity efficiency (note this is true in shared not dedicated segment-based protection) and minimize restoration time. On the other hand, segment-based protection increases the complexity of design and implementation. The key lies in the questions: how to optimally (NP-hard in most cases), or near-optimally assign the protection domains along the working path while considering network resource sharing at the same time?

In [Su2001] the authors proposed an algorithm to find the working path first and then its backup path segments. By introducing a new light-weight aggregated link cost metrics termed “buckets”, the algorithm focuses on constructing the backup paths with minimal wavelength consumption. Each bucket corresponds to a failure event, and the “height” of the bucket, indicates the protection wavelengths that are reserved on the link for that failure event.

The authors in [Qiao2002] proposed an Integer Linear Programming (ILP) formulation for performing segment-based shared protection according to the working path that is already determined. The algorithm is characterized by the effort of inspecting all possible allocations of protection domain and all possible numbers of protection domains. But the advantage of this algorithm is limited if the working path is not selected well. That means, just like routing with path shared protection (as described above), the computation of working and backup segments should be performed jointly. Based on this idea, the routing algorithm of [Qiao2002] was improved in [Xiong2003], which is based on a heuristic ILP formulation. Together, two scaling parameters are introduced in the algorithm to avoid the nonlinearity possibly induced when multiple states of spare capacity along each link are considered. However, how to select optimal values for these two parameters is still an open issue.

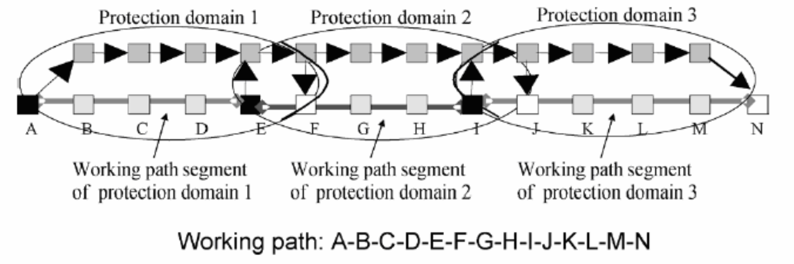

Protection domain 1
Protection domain 2
Protection domain 3
A
B
C
D
E
F
G
H
I
J
K
L
M
N
Working path segment of protection domain 1
Working path segment of protection domain 2
Working path segment of protection domain 3
Working path: A-B-C-D-E-F-G-H-I-J-K-L-M-N

Figure 2.6: Illustration of SLSP [Ho2004a].

The most promising segment-based shared protection is the SLSP (Short Leap Shared Protection) scheme proposed in [Ho2002]. The main idea of SLSP is to subdivide a working path into several overlapping segments as shown in Figure 7. The overlap between adjacent protection domains is for the purpose of protecting node failure along a working path. Each protection domain has a PSN (Path Switch Node) and a PMN (Path Merge Node), which switches over and merges back the affected traffic during a failure, respectively. A following heuristic algorithm called Cascaded Diverse Routing (CDR) is introduced in [Ho2004a]:

Step 1: Select the *S* shortest alternate paths from the *L*th-shortest paths in terms of hop count for each node pair in the network.

Step 2: Define a series of PSL–PML pairs along each alternate path with a fixed distance *D*. Note: Steps 1 and 2 can be performed off-line.

Step 3: As a connection request arrives, the ITSA algorithm [Tapolcai2003] is invoked upon a set of PSL–PML pairs along an alternate path. Then the ITSA algorithm is iteratively performed on each alternate path until a feasible solution is derived.

By tuning the values of certain parameters called *S*, *L* and *D*, it is reported in [Ho2004a] that CDR performs better than other schemes.

2.4.3.3. Active Restoration

The study in [Azim2004] proposed a restoration scheme, the so-called active restoration scheme, where a working path is protected by (pre-defined but not reserved) multiple backup paths that start from the source node of the path to the nodes along the working path, respectively. Upon a failure along the working path, the node immediate-downstream to the failure probes the availability of the pre-defined backup routes supported by each downstream node of the failure, and the first available backup route will be taken to restore the affected traffic [Azim2004]. Similar ideas were also presented in [Zheng2004, Tan2005, Ahn2002].

2.4.3.4. p-Cycle

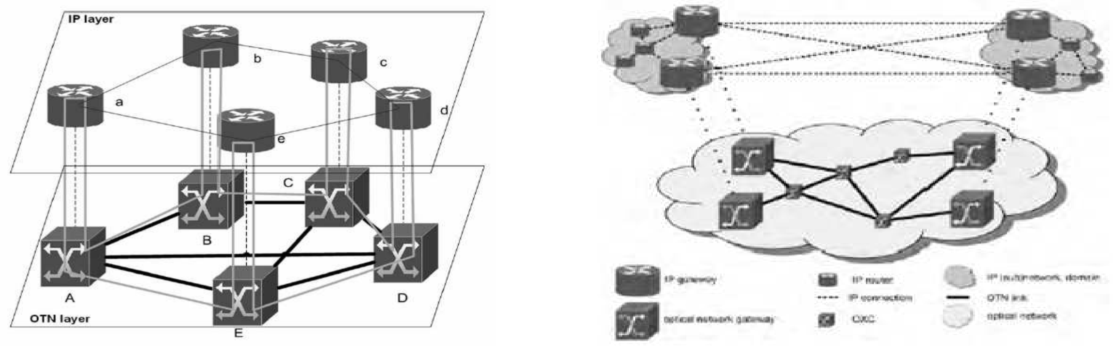
The p-cycle [Grover2000] is a cyclic, pre-calculated, pre-assigned, closed path with a certain amount of allocated spare capacity [Blouin2003]. It provides protection for any link that has both end nodes on the cycle as either an on-cycle link or a straddling link [Grover2000]. The p-cycle combines the advantages of the ring and of the mesh: it realizes ring-like recovery speed while retaining the capacity efficiency of the mesh-based methods. As a shared link protection method, p-cycle has high capacity efficiency and shorter fault detection and shorter re-routing time than path oriented methods.

Regarding the p-cycle allocation/assignment methods, there are generally two kinds of approaches: ILP-based [Schupke2002, Grover2002], which are difficult and time-consuming, and heuristic (iterative-based) solution [Doucette2003]. In [Doucette2003], a p-cycle is decided as follows: first to identify a set of candidate p-cycles, then to search iteratively for improvements on those cycles through various operations, finally to obtain a cycle with high efficiency and all working capacity of the network being protected.

At the end of this section, we should point out that the basic ideas in most of the above optical layer recovery techniques can also be applied in the MPLS layer, e.g., adopting segment-based shared protection in MPLS networks while considering restoration latency by limiting the length of the protection domain [Li2002], or building up MPLS layer p-cycle [Kang2003], etc.

2.4.4. Multi-Layer Recovery

It is evident that resilience schemes at the WDM or MPLS layer have their own pros and cons. Recovery at the WDM layer has the advantages that the average protection and restoration time is relatively small (1-100ms) and the granularity of switch is coarse. However, the recovery schemes at the WDM layer cannot resolve failures in a higher layer, such as router faults and service degradation in the IP layer. On the other hand, the recovery at the IP/MPLS layer results in better resource utilization and offers finer-grained service to different traffics, but it is slower and less scalable than its counterparts in the WDM layer. The study in [Qin2003, Pickavet2006] indicated that recovery at multiple layers is necessary to reach high availability compared with single-layer-only recovery. Finding efficient mechanisms to coordinate the various resilience techniques at different layers of the network, together with the corresponding spare capacity allocation at different layers in an efficient way, are the main objectives in multi-layer recovery.



(a) (b)

Figure 2.7: Multi-layer Resilience (a) MPLS LSR and OXC one-to-one match [Pickavet2006]; (b) MPLS LSR and OXC non one-to-one match [Staessens2006]

Roughly, multilayer recovery schemes can be classified as uncoordinated, sequential and integrated according to the inter-working between the IP/MPLS and optical layers [Colle 2002, Pickavet2006].

a) Uncoordinated Approach

In this most straightforward multi-layer recovery approach, recovery actions are deployed in multiple layers without any coordination, resulting in parallel recovery actions at distinct layers. Obviously, the main drawback of this approach is inefficient resource utilization due to the duplicated recovery of a single failure.

b) Sequential Approach

In the sequential approach, the logical network topology remains unchanged (static) at the time of a failure and no specific actions are initiated to modify it. The coordination of the recovery mechanisms at the network layers normally follows a sequential way, bottom-up [Colle2002] or top-down escalation.

The bottom-up strategy starts the recovery actions in the lowest detecting layer and escalates upwards only when the failure-affected traffic could not be recovered by the lower layer. The top-down escalation initiates the recovery actions from the highest detecting layer and escalates downwards only if the failure-affected traffic could not be handled by the higher layer. But since a lower layer has no easy way to detect on its own whether a higher layer was able to restore the traffic, the implementation of this strategy is somewhat more complex and currently not applied.

c) Integrated Approach

The integrated approach takes into account the combined knowledge of resource and topology information in both the IP/MPLS and optical layers to deploy recovery actions. This combined topology knowledge can be represented in a single integrated graph, by assigning appropriate cost to the edges of the graph. Survivable routing can be performed just like it is performed in the previous peer model (Section 2.2.1.2). Following this line of thinking, an integrated (peer model) shared path protection scheme was proposed and studied through simulation in [Liu2004]. The novelty of this scheme is its definition of the cost function: through an adjustable parameter, *k*, in the cost function, people can control the preference of minimum physical resources (number of hops per path) and load balance. The idea is based on the following observation: when the network load is low, minimizing the hops of physical links on a path is most important in routing; when network load is high, load balancing becomes more important. Similar work has also been done in [Zheng2003].

2.4.5. Inter-Domain Recovery

2.4.5.1. MPLS/GMPLS Inter-Domain Recovery

For inter-area/AS recovery, MPLS/GMPLS schemes prefer to use path (or sub-path, segment) protection because it is, generally speaking, difficult to find a backup path/sub-path in real time after the failure for a long inter-area or inter-domain working LSP. Hence the diverse path computation is important in these schemes.

a). Primary Path Route Object (PPRO)-based Scheme [Lang2004]

Lang et al. computed (and hence installed) the primary/working and backup paths subsequently (in separate phases) to support the end-to-end GMPLS-based path recovery. For this purpose, they proposed a new route object, called Primary Path Route Object (PPRO), in addition to the original Path Route Object (PRO) in standard RSVP-TE.

The working LSP is signaled with the standard RSVP-TE procedure. The record route object (RRO) is then activated in the working path RESV message with the role of collecting the detailed node-level path information of the established working LSP and reporting it back to the head-end node. Subsequently, the head-end node starts the setup phase of the backup LSP with a new downstream PATH message including an additional PPRO that records the working path to avoid overlap with the working path.

b). Associated Route Object (ARO) Scheme [Ricciato2004, 2005]

The ARO-based scheme computes the working and backup paths jointly. Hence, it is superior to the PPRO-based one: higher success rate and lower total cost [Ricciato2005]. The ARO-based scheme is accomplished by using the ARO in conjunction with the ERO (Explicit Routing Object) during the first PATH message. The key point of this scheme is to assume that each domain border node is able to jointly compute a pair of diverse paths from any pair of border nodes toward the remote destination.

When the working path PATH message goes to the destination node, the working path has been fully computed and partially signaled (PATH phase), and the backup path has been completely computed and collected in the ARO. The backup path can be returned in the RESV messages to the head node, which will start the subsequent installation of the backup LSP.

c). Inter-area SRLG-disjoint multi-segment protection Scheme in GMPLS networks [Miyamura03]

[Miyamura03] proposed a two-segment mechanism for providing path protection in multi-area GMPLS networks, with SRLG considerations.

Suppose we want to set up a connection from node S to node D in Fig. 2.8 with path protection. In the first step, node S asks an ABR in its own area (e.g., ABR2 in Area 1) to compute both the working sub-path (e.g., S→ABR 2) and backup sub-path (S→ABR 1). In the second step, ABR 2 sends the Path Computation Request message to one of the ABRs in Area 2 (e.g., ABR 5). Then ABR 5 can find a pair of SRLG-disjoint paths optimally from ABR 1 and ABR 2 to node *D* as ABR 5 has the routing information of both Area 0 and Area 2. However, this scheme is still sub-optimal because the summation of routing optimization in Area 1 and routing optimization in both Area 0 and Area 2 does not lead to the global routing optimization in overall network (Area 1 + Area 0 + Area 2).

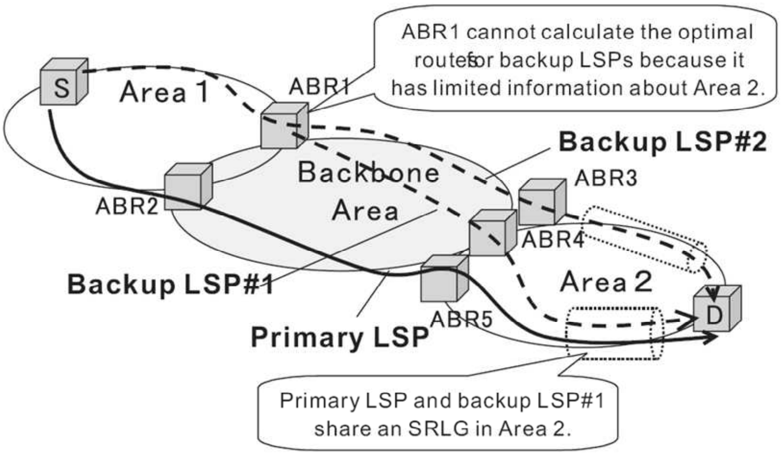

ABR1 cannot calculate the optimal routesfor backup LSPs because it has limited information about Area 2.
S
Area 1
ABR1
Backup LSP#2
Backbone Area
ABR3
ABR2
ABR4
Area 2
Backup LSP#1
D
ABR5
Primary LSP
Primary LSP and backup LSP#1 share an SRLG in Area 2.

Figure 2.8: Inter-area SRLG-disjoint and multi-segment protection Scheme [Miyamura03]

d). Bypass Tunnel Scheme for Inter-domain Restoration [Huang2003, 2004]

This scheme allows the working and backup LSPs to be concatenated at the domain boundaries so as to combine the advantages of fast local repair at the domain boundaries and resource-efficiency of end-to-end path protection within domains.

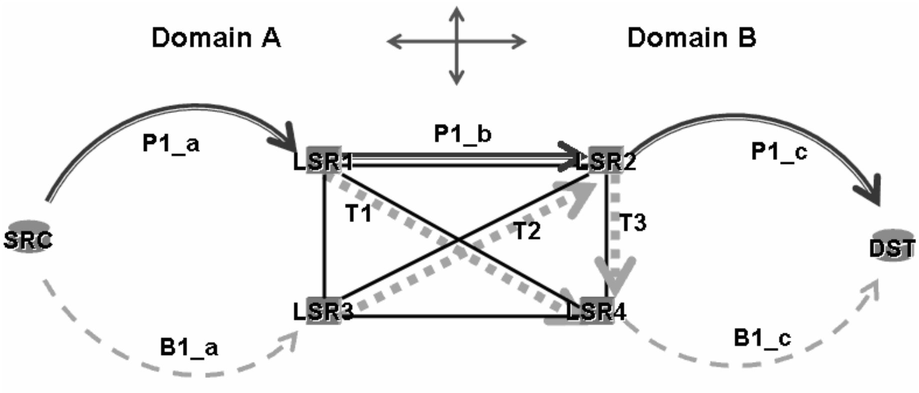


Figure 2.9: Inter-domain MPLS Protection [Huang04]

In Fig. 2.9, three bypass tunnels (*T1*, *T2*, *T3*) are setup manually at domain boundaries. A working LSP, *P1*(*P1_a*→*P1_b*→*P1_c*), is hence protected by three segments: backup sub-path *B1_a* in (source) Domain *A*, backup sub-path *B1_c* in (destination) Domain *B*, and the three bypass tunnels. Suppose *P1* fails at the destination domain, the traffic can go *P1_a*→*LSR1*→*P1_b*→*LSR2*→*T3*→*LSR4*→*B1_c* or *P1_a*→*LSR1*→*T1*→*LSR4*→*T3*→ *B1_c*. Compared with MPLS end-to-end path protection, in which the source node always acts as PSR (path switch LSR), the bypass tunnel scheme has less restoration time [Huang2003, 2004]. This is because the domain boundary LSRs (*LSR1* and *LSR2* in Fig. 2.9) can also act as PSR in some failure cases. Meanwhile, the bypass tunnel scheme has more efficient resource utilization through using the working path as much as possible when performing recovery after the failure.

d) PCE-based approach [Vasseur2003, 2004]

Only the PCE-based approach can provide global routing optimization, at least theoretically. The PCE-based optimal-routing mechanisms, described in Section 2.2.2.5, can be extended in a straightforward way for diverse/disjoint path computation. The only difference is that PCE needs to compute a least cost path pair (working and backup) now. In order to ensure a globally optimal solution, as in Figure 4, PCE has to advertise an optimal path pair for each possible pair of entry points. Obviously, if the number of domain gateway nodes is large, the route computation overhead of the PCE-PCE communication becomes heavy, which may degrade the overall performance of the PCE-based approaches [Ricciato2005].

2.4.5.2. Optical Layer Inter-Domain Recovery

a) Segment-based Protection (per-segment approach discussed in Sec.2.2.2.3)

In [Akyamac2002], the authors considered a segment-based inter-domain protection scheme. The end-to-end inter-domain lightpath consists of smaller lightpath segments that are routed between the gateway nodes. Routing and protection is strictly limited to each domain, thus the backup paths for local failures will be contained in the domain and will not traverse into the neighboring domains. This protection scheme assumes that the gateway nodes will never fail.

b) Path Shared Protection

The recovery scheme in [Thiongane2005] used the layered approach (Sec.2.2.2.4) to compute working and backup path pair. For a new request, the working path is routed first, and then the (shared) backup path is computed based on the network graph after removing the working path, both using Dijkstra's algorithm.

c) Inter-domain p-cycle

The p-cycle-based inter-domain recovery scheme in [Farkas2005] is basically a two-layer approach. It decomposes the multi-domain resilience problem into two sub-problems, namely the higher-level inter-domain protection, and the lower level intra-domain protection. Building p-cycles at the higher level is to handle the failures for the inter-domain links. At the lower level, traditional protection schemes can be applied as intra-domain protection. However, this higher level p-cycle is not able to handle node failures and might lead to inefficient use of network resources [Farkas2005].

d) Several Resilience Mechanisms in a multi-domain IP-over-Optical Environment [Staessens2006]

The study in [Staessens2006] is the one most close to my own research topic. It considered the end-to-end recovery in multi-domain IP networks interconnected by an optical domain. The end-to-end recovery is composed of difference segments (IP network recovery, gateway recovery and optical network recovery). This paper focused on the gateway-to-gateway (gateway->Optical domain->gateway) recovery and assumed the recovery within the domains is provided. The following generic end-to-end recovery techniques (Fig. 2.10) are studied quantitatively:

- *No optical protection* (Fig. 2.10a): working and corresponding backup IP connections are set up without any extra protection in the optical domain.
- *Optical protection of both IP connections* (Fig. 2.10b): there is extra protection in optical domain, both to working and backup IP connections.
- *Optical protection for the working paths* (Fig. 2.10c): there is extra protection in the optical domain, but only for the working IP connections.
- *Dynamic restoration* (Fig. 2.10d): The optical network is capable of setting up lightpaths at will in case of failure; this means that providing restoration-type recovery in the optical domain so as to use the network resource in a very efficient way.

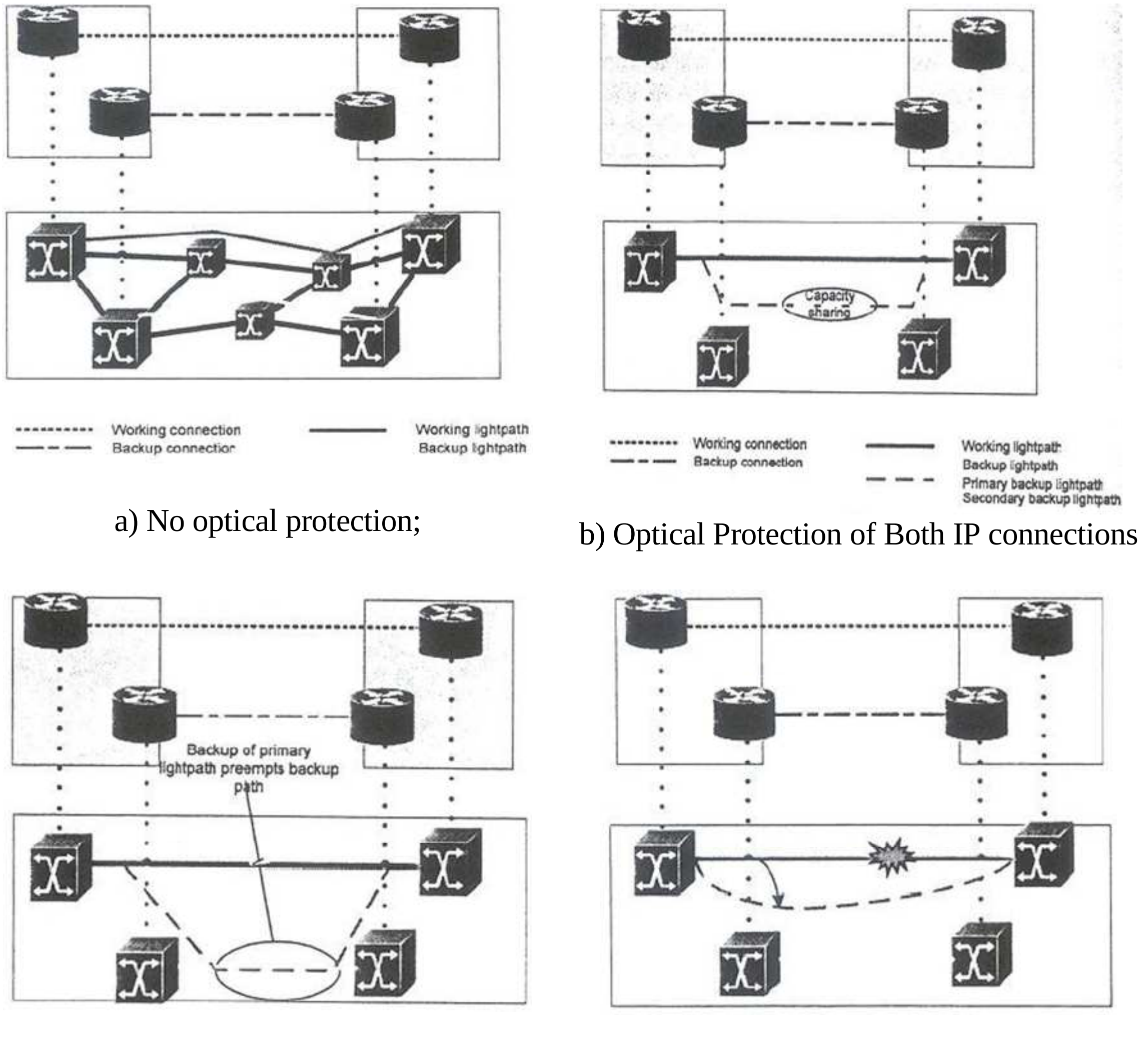
The study in [Staessens2006] gives hints and is helpful to people doing research in multi-layer multi-domain recovery. The results are somehow preliminary and a little bit rough. The work is interesting, although many deeper issues are still open, e.g., routing optimization is never considered.

2.4.5.3. Multi-layer Inter-Domain Recovery

As far as we know, there is no promising solution published for the resilience issues in multi-layer multi-domain network environments. Two general reviews on resilience in a multi-layer multi-domain network environment were given in [Larrabeiti2005] and [Demeester2005], together with the classification of some issues related to it. In [Larrabeiti2005], it is further pointed out that network resilience based on fast re-routing might be achieved, loosely speaking, by using alternative disjoint multi-domain backup paths through other domains together with inter-domain link protection strategies. And intelligent usage of MPLS label stacking might add scalability to the establishment of LSPs between non-directly connected domains.

In [Demeester2005], the author suggested that multi-domain p-cycle might be useful for resilience in a multi-domain environment, since a p-cycle is established in advance and shares the capacity (resource) without knowing all the working and backup paths.

The study in [Miyamura2004-2] touched the resilience issue for inter-region multi-layer networks through providing a multi-layer disjoint path selection algorithm, which is a simple extension of their work in [Miyamura2004a]. The multi-layer routing is implemented in a simple way: finding the least-cost working and backup path pair first at the G/MPLS virtual topology; if not successful, then finding the least-cost path pair at the optical layer.



a) No optical protection;

b) Optical Protection of Both IP connections

c) Optical Protection for the working path

d) Dynamic Protection

Figure 2.10: Several generic end-to-end recovery techniques in [Staessens2006]

3. Routing of MPLS flows Over AAPN

An Agile All-Photonic Network (AAPN) [Bochmann2004, Mason2005, Bochmann2007], with an overlaid-star topology, can potentially provide an efficient high bandwidth/high performance core transport network solution for carriers. Hence, it is very important to design and position AAPN to support the IP/MPLS architecture and protocols. Deploying AAPN in an IP/MPLS network environment needs signaling and routing information exchange between the routers surrounding the AAPN. Particularly, we study the Open Shortest Path First (OSPF) [RFC 2328] IP routing protocol, which is commonly used for routing within a single administrative domain.

OSPF is a link-state routing protocol that is used by MPLS and GMPLS (Generalized MPLS) (with extensions). Each OSPF-running router exchanges LSAs (link state advertisements) through a reliable flooding mechanism to build up and synchronize its link state database (LSDB) with the database of other nodes in the network. The LSDB thus becomes a complete representation of the network topology and resource information (OSFP with TE extensions, OSPF-TE [RFC 3630]). Based on it, each router can use the shortest-path-first (SPF) algorithm to compute its routing table, or run a constraint-shortest-path-first (CSPF) algorithm to perform source routing. OSPF-TE and RSVP-TE (Resource Reservation protocol with TE extensions) [RFC 3209] are fundamental for MPLS TE as they are used to compute and establish explicitly routed LSPs (label-switched paths) whose paths follow a set of TE constraints.

OSPF is a hierarchical routing protocol that supports large networks through multiple OSPF areas: one backbone area (Area 0) surrounded by non-backbone areas. Area border routers (ABR) are located at the border between the backbone and the non-backbone areas, and distribute summarized routing information among the areas.

In this chapter, we consider two scenarios to deploy an AAPN in an IP/MPLS network environment, namely within one OSPF area networks and within several OSPF areas networks. Since a large portion of the anticipated connections will need to traverse both the backbone area and the non-backbone areas, we focus on the second scenario, in which the proposed inter-networking framework can implement inter-area MPLS Traffic Engineering in an efficient and distributed manner.

3.1. Solving the Scalability Issue when Deploying AAPN within a Single OSPF Area Network

The first scenario to deploy AAPN in IP/MPLS networks is in a single OSPF/OSPF-TE area (as shown in Fig. 3.1).

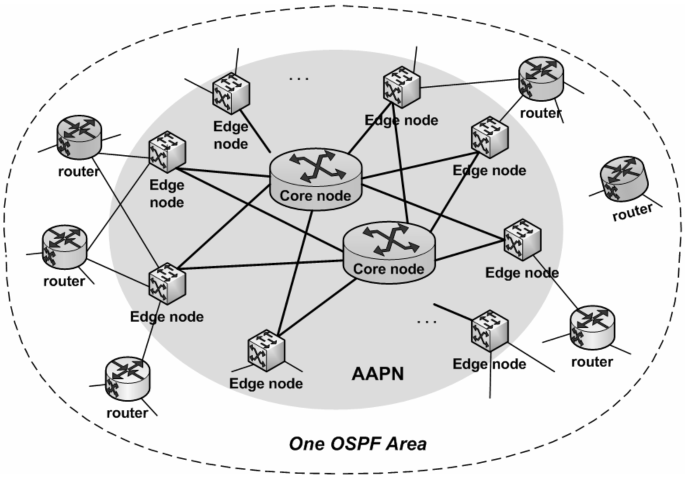


Figure 3.1: Deploying an AAPN is in a single OSPF/OSPF-TE area network

3.1.1. The Problem: Scalability Issue

Since the AAPN provides an $N \times N$ interconnection structure for the N edge nodes of the AAPN architecture, the straightforward usage of a routing protocol like OSPF leads to scalability problems since the value of N could be very large (e.g., around 1000 and OSPF has to deal with $N \times (N-1)$ links). Hence we need to consider what aspect of the AAPN topology should be exported to the IP/MPLS world and how to organize the related routing information exchange. In addition, the exported topology should be:

- As simple as possible (to reduce routing protocol traffic, routing calculation and the size of the link-state database)
- Provide a good match with the fault model of the AAPN (e.g., link/edge node/core-node failure)
- Meet the traffic engineering requirement (not fully opaque to the outside)

Note: due to the symmetric architecture of AAPN (see Fig. 2.1 and Fig. 3.1), we use the "bundle" concept to further reduce the overhead traffic to the outside. As illustrated in Fig. 3.2, 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, called "the core".

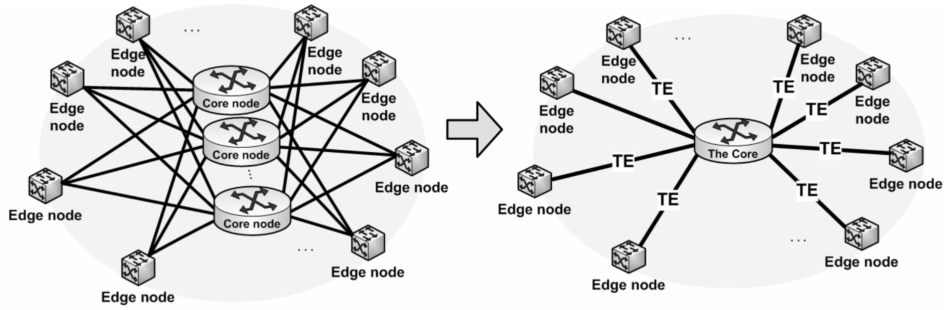


Figure 3.2: Abstraction of AAPN: TE links and the core.

3.1.2. Description of Proposed Solutions

We propose four approaches described as follows:

1) Full-Mesh (Fig. 3.3a)

The whole AAPN (edge nodes, links, and core nodes) is exported as a full-mesh network to the outside (Fig. 3.3a). Within the AAPN, permanent connections are set up between all edge node pairs for routing information exchange, and may also be used for data exchange; while additional connections for data transmission may be established on demand. Each AAPN edge node behaves as an IP/MPLS router and the core nodes are invisible to the outside.

2) Core-Star (Fig. 3.3b)

A star topology (the core surrounded by *N* edge nodes) is exported to the outside IP world. Each edge node maintains a two-way permanent connection only with the core for routing information exchange. Data connections will be established on demand. Each AAPN edge node behaves as an IP/MPLS router and the core is visible from outside as an IP/MPLS router.

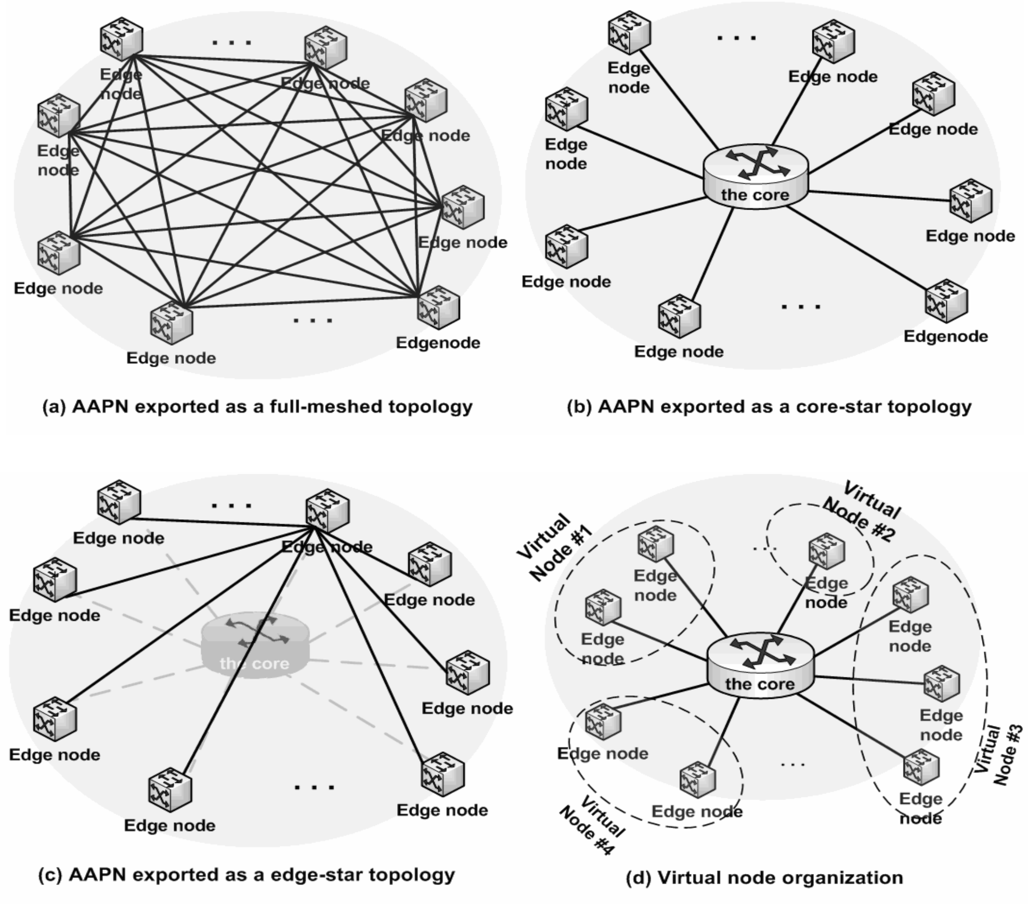

Edge node
the core
Virtual Node #1
Virtual Node #2
Virtual Node #3
Virtual Node #4
Edgenode
(a) AAPN exported as a full-meshed topology
(b) AAPN exported as a core-star topology
(c) AAPN exported as a edge-star topology
(d) Virtual node organization

Figure 3.3: Exported Topologies of Agile All-Photonic Network

3) Edge-Star (Fig. 3.3c)

As an alternate to the core-star topology, AAPN can also be exported as a star topology where one edge node is surrounded by the other *N-1* edge nodes. The major differences with the core-star topology are the following: (a) each edge node maintains a two-way permanent connection only with one particular edge node (not the core) for routing information exchange, and (b) each edge node behaves as an MPLS-capable IP router but the core node is invisible from outside.

TABLE 3.1: COMPLEXITY ANALYSIS OF MESH AND STAR TOPOLOGIES

	Full-mesh	Core-Star	Edge-Star
# of 1-way connections to be maintained	$N\times(N-1)$	$2N$	$2(N-1)$
# of router-lsas flooded within aapn after a single connection failure	O(N^2)	O(N)	O(N)
# of router-lsas flooded within aapn after a single link/edge node failure	O(N^3)	O(N)	O(N)

Table 3.1 compares the above three exported topologies. Full-Mesh has a severe scalability problem when *N* is large: there are too many connections to set-up and maintain and hence there is a heavy load of control traffic although it is shared among all the edge nodes. The star topologies are much simpler. However, the load at the center of the star (the core or the central edge node) would become much heavier than at the other edge nodes when *N* is big.

4) Virtual Router Organization (Fig. 3.3d)

To find a balance between the simplicity of the exported topology and the load-sharing of control traffic, we propose the concept of a virtual router (VR) to organize the routing information exchange in the AAPN in a hierarchical manner. A VR represents a collection of co-located (or near-located) edge nodes and part of the core node switching capability (see dotted line in Fig. 3.3d). A VR is viewed as one IP/MPLS router, and these VRs, together with the core node, can form a virtual star architecture, thus reducing the number of paths among these "routers” and simplifying exported network topology, as compared with the Core-Star topology.

In an extreme case, a single VR may include all the edge nodes (there is no need for a core node any more), and the whole AAPN can be seen as one big router. In another extreme, a VR may just contain a single edge node. Generally speaking, to reduce the routing protocol traffic, the size of the VR should be big. But the TE requirements may push for smaller VR sizes (fine granularity). Hence the VR size is normally a balance between the above two extreme cases. Meanwhile, the VR-based topology is scalable since adding an edge node in a VR will not affect the whole exported topology.

The VR Organization implies a two-layer organization of the routing information exchange: (1) within each VR domain, and (2) between the VRs (still within the AAPN).

The communication between the VRs can adopt the architecture of Full-Mesh, Core-Star or Edge-Star. Among these possibilities, we recommend Edge-Star or Core-Star because it achieves the balance of simple exported topology and load-sharing, and has the smallest number of permanent connections to be set up.

Each VR has a head (a designated edge node, possibly with a designated backup node). When an edge node finds a routing update from its neighbor router(s), it reports the update to its head node. The head checks the update, aggregates it, if possible, and forwards it to the heads of other virtual routers. Those heads then distribute the update to the edge nodes that are member of their respective VR domain. A simple internal routing cooperation protocol, like the one in [Chou2002], can be used for this purpose within each VR domain. Note that the forwarding table of each edge node includes both the forwarding information per local external (non AAPN) output port and information for forwarding through the AAPN network.

3.1.3. Short Summary

Based on all the above analysis, we have the following conclusions when OSPF with a single area (or any other non-hierarchical routing protocol) is to be deployed over a network including an AAPN:

- A very small AAPN can be viewed as single big IP router with MPLS capability.
- A small or medium-sized AAPN (with a few tens of edge nodes) can be viewed as a full-mesh or a star topology where each edge node is viewed as an IP router with MPLS capability.
- A large AAPN should use hierarchical information exchange using the concept of virtual routers (VR) interconnected in a VR-star topology, as explained above.

3.2. An Inter-Area Optimal Routing Framework by Deploying Several OSPF Areas over an AAPN

3.2.1. The Problem

An inter-area connection normally starts in a non-backbone area, traverses the 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 signaling component. This is so 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, Fig. 3.4 shows a common multi-area network and we suppose *r1* in Area *x* is the source node while *r6* in Area *y* is the destination. Generally speaking, a non-backbone area (e.g., Area *x* in Fig. 3.4) 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 Fig. 3.4, how could *r1* choose an optimum ABR in Area *x* to the destination *r6*? 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 *r6*? Although local optimization can be done in each of the respective areas along the inter-area path (*r1* to *r6*), 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.

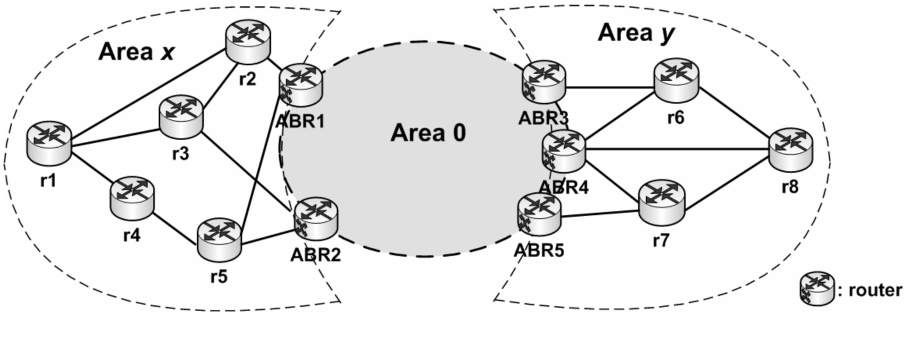


Figure 3.4: A common inter-area network ($x \neq y, xy \neq 0$)

3.2.2. A Novel Inter-area Optimal Routing Framework

The direct and natural way to deploy an AAPN in a multi-area network is shown in Fig. 3.5, where the core nodes are located in the middle of Area 0 and the edge nodes act as ABRs at the border between Area 0 and non-backbone areas. However, in this scheme, inter-area routing with global optimization still can not be guaranteed. Therefore, we propose a novel approach/framework, shown in Fig. 3.6, which can provide such guarantee. Our proposed framework consists of three main components, namely *the routing-information*, *path computation* and *signaling components*:

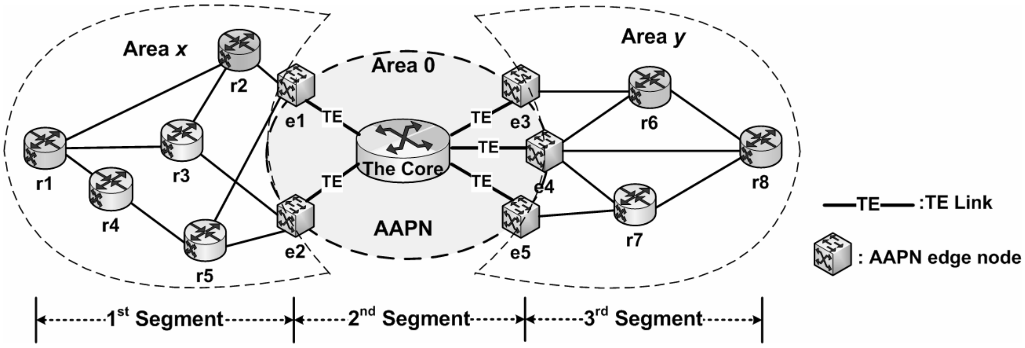


Figure 3.5: Directly deploying an AAPN as the backbone area

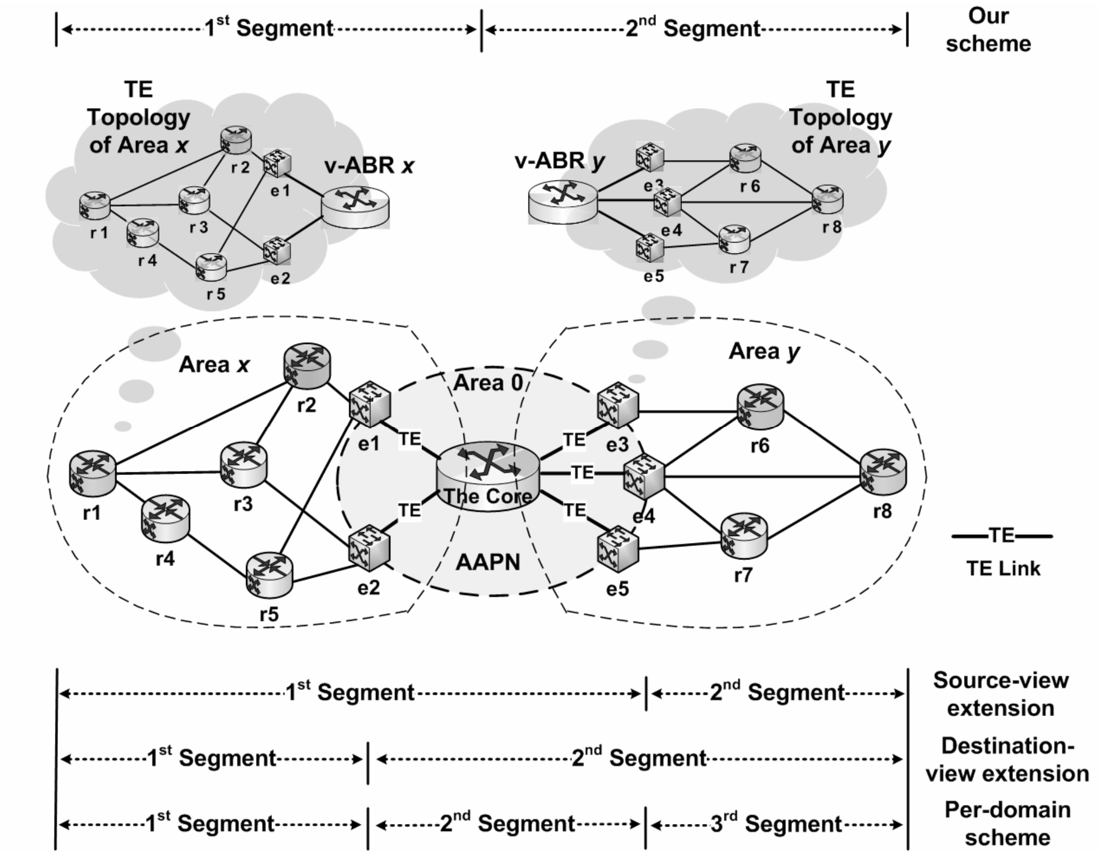


Figure. 3.6: The inter-area optimal routing framework we proposed deploys AAPN as backbone. We also illustrate the per-domain approach and its two extensions.

3.2.2.1. The Routing-Information Component

This component is responsible for the discovery and export of the TE topology of the AAPN. As seen in Fig. 3.6, we expand the OSPF non-backbone areas a little so that there is an overlap between Area 0 and each expanded non-backbone area. Then the AAPN edge nodes located in the overlap, together with their direct TE links to the core and the associated part of the core, belong to both the Area 0 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 its 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 with TE capability.

3.2.2.2. The Path Computation Component

In our framework, an inter-domain LSP can be considered consisting of two segments (instead of three, shown in Fig. 3.5) as shown at the top Fig. 3.6: one in the head-end (expanded) area and one in the tail-end (expanded) area. The core connects these two segments to form a complete inter-domain LSP.

The most interesting thing is that local routing optimization (through CSPF) by each of these two segments can lead naturally to a globally-optimized inter-area LSP. As seen in Fig. 3.6, 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. In other words, our proposed architecture presents the advantage that optimal end-to-end routes can be easily established by simply concatenating optimal routes to/from the core, which can be determined by the source and destination sub-areas independently of one another. The problem of finding optimal end-to-end routes can in general only be solved by considering global knowledge; in our architecture with a virtual router, no global knowledge is required, only the local routing information within each area. In addition, dynamic inter-area routing is implemented also.

It is worth noting that we use a TE abstraction technique (i.e., all the links from a given edge node to all the core nodes are assumed to be load-balanced and are abstracted as a single TE link). However, this link-load-balance assumption is hard to strictly enforce in practice. Therefore, our theoretical inter-area global optimality, strictly speaking, will become “near-optimal” in most practical cases.

3.2.2.3. The Signaling Component

This component is responsible for the establishment of the LSP along the computed path. In Fig. 3.6, consider the case that a source LSR (e.g., *r1*) wants to set up a LSP to a destination LSR (e.g., *r8*). *r1* must first compute an optimized path to the v-ABR of Area *x* through CSPF, and then signal this establishment request to the network.

Shown in Fig. 3.7, *r1* starts the signaling process by creating a RSVP Path message including an EXPLICIT_ROUTE object (ERO) [RFC 3209] to indicate the computed explicit path (with one sub-object per hop). However, *r1* has to use the loose ERO sub-objects for the hops outside Area x. In Fig. 3.7, the ERO specifies the explicit path as *r1->r3->e2->v-ABR x->r8*, where *r8* is a loose ERO sub-object. Then, *r1* sends the Path message to the next hop defined in the ERO, which is *r3*. *r3* receives the Path message and processes it as follows:

- checks the message format to make sure everything is OK,
- performs admission control to check the required bandwidth,
- stores the “path state” from the Path message in its local Path State Block (PSB) [RFC 2205] to be used by the reverse-routing function, and
- 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, e2.

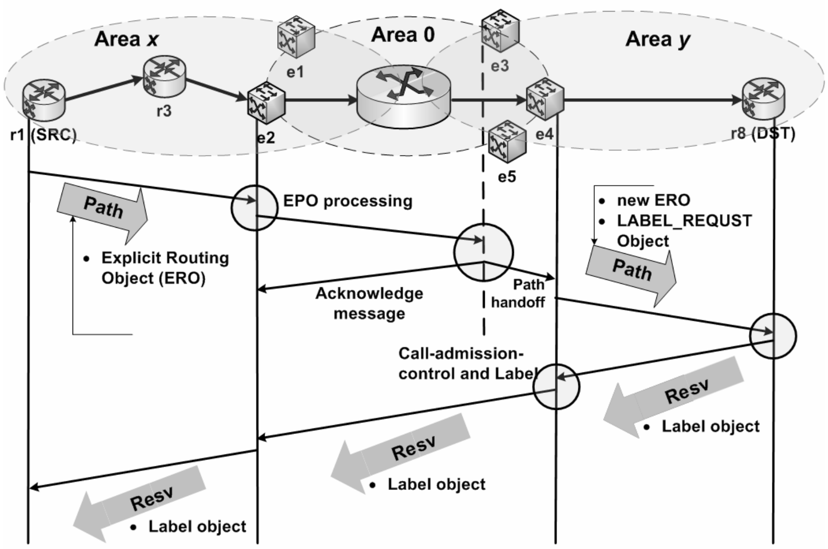


Figure 3.7: Inter-Area LSP Signaling Process that follows RFC 3209

e2, an AAPN edge node, receives the Path message from *r3* and checks the contained ERO. If *e2* finds that the IP address of the 2nd sub-object in the ERO is v-ABR *x* and the 3rd sub-object (with the loose attribute) is beyond Area *x*, then *e2* has the task of resolving the loose sub-object into strict ones. In our case, there is one loose sub-object, *r8*, which represents the destination of the requested LSP. Although *e2* can not find a strict path from v-ABR *x* to *r8* by itself, it knows who can. First, by checking the inter-area reachability information and internal parameters, *e2* finds out which group of edge nodes (also which associated v-ABR) are located in the same area as *r8*. In Fig. 3.7, these are *e3*, *e4* and *e5* (v-ABR *y*). Second, it selects an edge node among them randomly, e.g., *e3*. In the third step, *e2* removes the first two sub-objects (itself and v-ABR *x*) from the ERO of the original received Path message, and inserts v-ABR *y* at the top, then forwards the modified Path message to *e3*.

When *e3* receives the Path message and finds the 1st sub-object in the received ERO is v-ABR *y*, together with a loose second sub-object, *r8*, it knows that it should find an explicit path between these two sub-objects. As shown in Fig. 3.7, *e3* is capable to do the resolving work because *e3* and *r8* reside in the same expanded area, Area *y*. *e3* finds the optimized explicit path as: *v-ABR y->e4->r8*. *e3* then replaces the ERO object in the received Path message with a new ERO object that stores the resolved explicit route (*e4->r8*). Finally, *e3* forwards the new modified Path message to *e4* as if it were forwarded from *e2* by using *e2*'s data (IP address, etc.). We call this process a Path message *handoff*. At the same time, *e3* also sends an acknowledge message (containing the resolved path) to *e2* (Fig. 3.7). From the above handoff process, we can see that only the area-specific reachability (not TE) information needs to be exchanged among different areas. In our inter-area optimal routing framework, TE information is organized within each area independently. Edge node *e4* receives the Path message and believes it is from *e2*. Since all the sub-objects in the received ERO are strict, *e4* processes this Path message in a standard way, just as *r3* did in Area x, and then forwards the processed Path message to *r8*.

When the destination, *r8*, 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. The Resv message makes its way upstream (Fig. 3.7), hop by hop, and when it reaches the source LSR, *r1*, the inter-ISP path is setup: *r1->r3->e2->v-ABR x->v-ABR y->e4->r8*. Thus, a globally-optimized inter-domain TE LSP is set-up. It can be maintained or torn-down just as any normal LSP.

3.2.3. Further Discussions

As we can see, our proposal can provide globally-optimized inter-area dynamic routing and does not require any changes on existing traditional IP/MPLS routers, hardware or software (good backward compatibility). Furthermore, there is no node having global TE information. Instead, the TE information is distributed on per-area basis and only area-specific reachability (not TE) information is exchanged among areas. Global optimization is achieved through cooperation and interaction between AAPN edge nodes in different areas (Path message handoff). In addition, for the 2nd half of an inter-area LSP (in the tail-end area), the optimized routing computation is done randomly by an AAPN edge node in the tail-end area. Hence, load-sharing among these edge nodes is achieved.

Under our proposed framework, inter-area routing can be dynamic. In addition, re-optimization of an inter-area TE LSP can also be implemented, either locally within an area (by the head-end LSR for 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).

As seen in Fig. 3.6, 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.

Regarding the information complexity, we denote the number of areas by A and the number of edge nodes in each area by E (assuming that each area has the same number of edge nodes). In order to compute inter-area paths, the amount of additional TE information that a normal router in an area has to maintain in *enhanced* per-domain approaches (e.g., source-view-extension approach or destination-view-extension approach) is of size $O(A \times N_C \times E)$ (where N_C is the number of core nodes in AAPN); while it is $O(E)$ in our routing framework, which is much smaller and independent of the number of areas and the number of core nodes in AAPN.

3.3. A Novel Internet eXchange (IX) Architecture with Traffic Engineering Capability

Besides the multi-area network scenario, an AAPN can also be used in multi-AS networks. Particularly, we propose a new Internet Exchange architecture based on AAPN to implement MPLS inter-AS traffic engineering.

3.3.1. Internet eXchange (IX) in Multi-AS/Multi-ISP Networks

The Internet is a worldwide multi-AS network. As the Internet grows, the Internet eXchange (IX) plays an important role in supporting the Internet backbone. This is because an Internet Exchange is a place where Internet Service Providers (ISPs) (normally Autonomous IP Systems (AS)) can interconnect their networks and exchange Internet traffic with each other. The exchanging of inter-ISP/inter-AS traffic on an IX is known as “peering” [Ams]. The direct way to implement peering between two ISPs’ networks is to build physical links connecting them. However, this will lead to an *n*-squared scalability problem if the number of these ISPs, *n*, is large. By adopting an Internet Exchange in the middle to inter-connect these ISPs’ networks (see Fig. 3.8), the *n*-squared issue can be solved. In addition, the ISPs can set up peering with each other in an efficient way through the IX.

There are many large and fast growing Internet Exchanges in the world, either non-profit or commercial, e.g., AMS-IX (Amsterdam Internet Exchange [Ams]), Japan Internet Exchange [Jap], Switch and Data (U.S.) [S&D], etc. In Europe, there are now more than 30 IXes and over 1,600 connected networks to these IXes [EIXA]. In May 2001, Euro-IX

(European Internet Exchange Association) was formed with the intention to further develop, strengthen and improve the Internet Exchange community [EIXA].

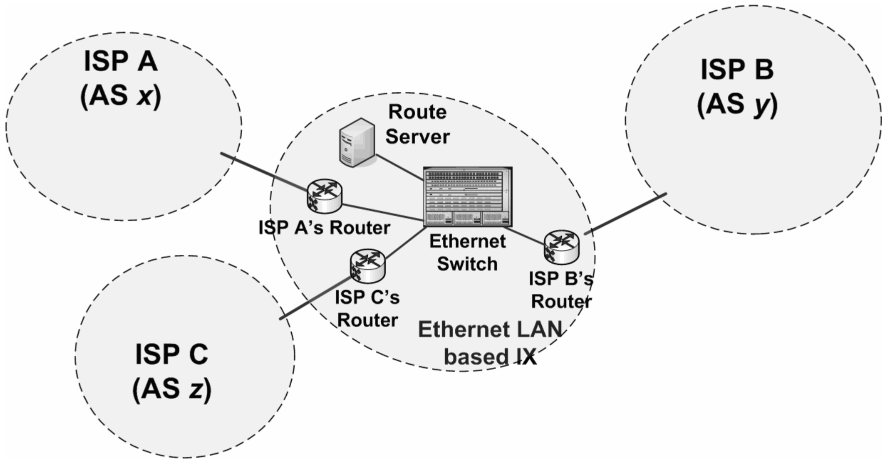


Figure 3.8: An Example of Internet Exchange (IX)

Normally, only the BGP routing protocol is allowed in an IX, thus the traffic over the IX is exchanged based on BGP routes. An ISP's local traffic is not allowed to pass through the IX. The majority of IXes opted for a layer-2 switched Ethernet LAN architecture, while only a few IXes use ATM or FDDI. Meanwhile, several new architectures for IX have been proposed, e.g., IPv6 IX [Morelli2005], MPLS-IX [Nak2002], photonic IX [Shake2005], etc. However, there are two common potential drawbacks in the above IX architectures:

1) Complex IX internal routing: large amount of core routers/switches in an IX will greatly increase the internal complexity of an IX, e.g., routing issues among these cores. Hence it is not easy to extend current IX's capability.

2) Lack of TE capability: none of the above IX architectures considers inter-ISP traffic engineering. For instance, in the layer-2 Ethernet IX, the widely-used VLAN configuration is static, and does not adapt to traffic change; there is no tunnel technology in the Ethernet shared switching infrastructure, no point-to-point connection; strict QoS guarantee can not be provided.

3.3.2. AIX: AAPN based Internet Exchange

AAPN is suitable to be deployed as an Internet Exchange (we call an AAPN-based IX *AIX*, see Fig. 3.9) with the following advantages compared to the existing IX architectures:

1) *Flexible and Distributed Access*: As shown in Fig. 3.9, due to the simple and geographically distributed topology of AAPN, an AAPN based IX (AIX) can provide access (through its edge nodes) to customer ISPs at exactly their local locations. Furthermore, an ISP can have several access points to the AIX through the different edge node of the AAPN. In this way, the inter-ISP traffic can be distributed widely and balanced (avoid the “bottlenecks”) within the ISP’s network, which also increases the reliability of inter-ISP peering.

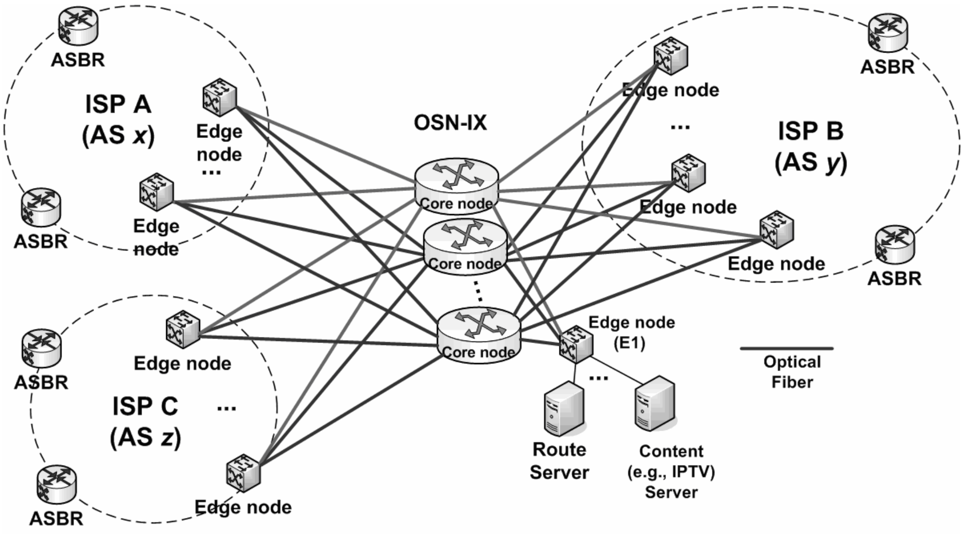


Figure 3.9: AIX: AAPN based Internet Exchange

2) *Nearly Unlimited Capacity with Good Scalability*: The switching capacity or throughput of an AAPN-based IX, C_{AIX} , can be calculated as the follows:

$$C_{AIX} = B \times W \times P \times N_C \text{, where,} \tag{3.0}$$

- B is the bandwidth of a wavelength over an AAPN internal fiber connecting AAPN core node and edge node. Typically, it is 10Gbit/s.
- W is the number of wavelengths per AAPN fiber.
- P is the number of switch ports (fibers) per AAPN core node.
- C is the number of core nodes in the AAPN.

Let W be 20, P be 64 (maximal value [Mason2006]), N_C be 5, C_{AIX} of such an AAPN is 64Terabit/s. An AAPN can scale gracefully and continuously in capacity to keep pace with the growth in demand of customers by increasing the number of core nodes and/or wavelengths per fiber.

3) *Good Resilience*: We consider an AAPN as a switch only conceptually; it does not mean we can use a powerful single physical switch to replace AAPN. This is because adopting a single core switch will concentrate all the risks of an IX into a single point. The more powerful the switch, the higher is the risk. In the AIX architecture, things are just the opposite: the higher the capacity (more core nodes), the more reliable is the AIX. This is due to the fact that in the AAPN architecture, the risk is distributed among several independent (not-so-powerful) core nodes, and the overlaid star topology lets them back-up each other. Besides, an AIX can increase its capacity by installing extra core nodes in a graceful and natural way without introducing any complex internal routing, as is the case for other IX architectures.

3.3.3. AIX's Traffic Engineering Framework

The most attractive feature of AAPN-based IXes should be its traffic engineering capability. The optimal routing framework proposed in Section 3.2, which we initially designed for inter-area traffic engineering, can also be applied to the AAPN-based IX architecture with the following modifications:

- The AAPN-based IX interconnects various ASes (not areas);
- The AAPN's edge nodes, located in various ASes, will exchange BGP-specific routing information (e.g., AS path) with each other;
- The virtual router that represents the AAPN becomes a virtual AS Border Router (v-ASBR, not v-ABR);
- An inter-AS TE LSP starts in an AS, traverses the AAPN, and terminates in another AS. (similar to inter-area TE LSP).

The three essential components of the optimal routing framework proposed in Section 3.2, namely routing info, path computation, and signaling components, can be applied to the AIX architecture in the multi-AS network environment with the above changes (see Fig. 3.10). Thus, the AAPN-based IX obtains the inter-AS traffic engineering capability and the inter-AS TE LSP routing can be performed optimally, dynamically and automatically without leaking AS internal confidential information to other ASes.

Meanwhile, no AS needs to maintain the global TE information, hence the scalability of AIX architecture is good.

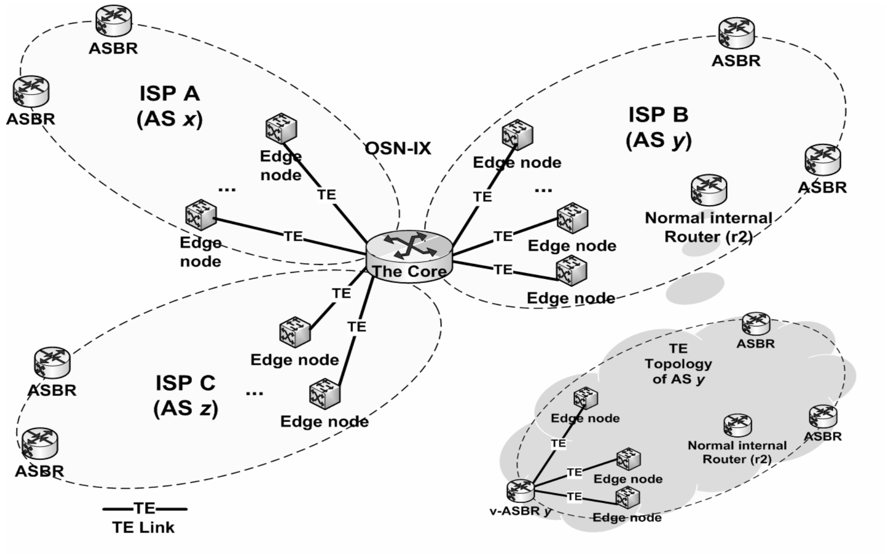


Figure 3.10: AIX’s TE framework (similar to Figure 3.6)

3.3.4. Route Service in AIX Architecture

Besides the traditional full-mesh style, AAPN-based IX offers an efficient alternative to build up BGP sessions internally (iBGP sessions within each ISP AS) and externally (eBGP sessions among the ISP ASes) to exchange inter-ISP reachability information. As illustrated in Fig. 3.11, there is a physical route server connected to an edge node (e.g., *E1*) that is co-located with core nodes. The route server provides route service to both the iBGP sessions and the eBGP sessions. Each individual ISP AS has one route reflector (RR) [RFC 4456] instance running in the route service server to offer an alternative to the logical full-mesh requirement of iBGP sessions within the ISP ASes. The RR instance actually acts as a concentrated focal point; multiple BGP routers of an AS can thus peer with it rather than peer with each other in a full mesh style (to avoid the *n*-square scalability issue). For eBGP sessions among peered ASes, the exchange of the reachable IP prefixes offered by each AS is done within the route server all by software. There is no explicit eBGP sessions among the ASes inter-connected by the AIX. Compared with the methods in current IXes, this is more efficient and manageable. Note that for the prefix exchange with other outside ASes that are not directly inter-connected through AIX, the standard BGP procedures must be followed, e.g., eBGP sessions build up as in Fig. 3.11. Please note that in order to increase the reliability of the route server, there could be another *backup* route server either connected to the same edge node as the active route server, or connected to another edge node which is the backup of the co-located (e.g., E1 in Fig. 3.11) edge node for higher reliability requirement.

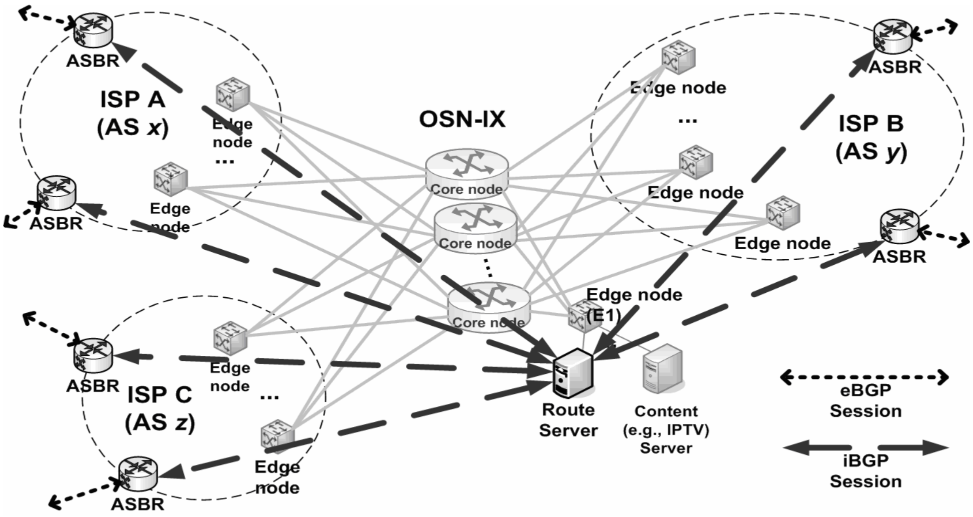


Figure 3.11: Router service within AIX

3.4. Performance of AAPN-based Architecture in Inter-Domain Traffic Engineering

We now study the performance of our AAPN-based inter-domain (area or AS) optimal routing framework and compare it by simulation with the following existing inter-domain TE approaches:

- *Per-domain approach*, which computes the inter-domain path in a domain-by-domain fashion starting from the head-end domain (see the bottom of Fig. 3.6).
- *Outgoing-view-extension [Chen2007]*. As shown at the bottom of Fig. 3.6, this approach extends the source node's TE visibility so that it can view its own domain

and the whole AAPN in order to compute the first segment of an inter-domain path. Then the second segment is computed by an ingress gateway (edge) node of the tail-end domain.

- *Incoming-view-extension [Miyamura2004a, Miyamura2004b, Otani2007]*. Also shown at the bottom of Fig. 3.6, this approach defines two segments in the opposite way compared to the outgoing-view-extension approach. The source node can only view its own domain to compute the first segment. The gateway (edge) nodes in the tail-end domain can view its own domain and the whole AAPN to compute the second segment.
- *Global Knowledge (ideal case)*, in which each inter-domain routing decision is made on the basis of global knowledge of real-time TE information. We use this case as a benchmark.

Simulation experiments were conducted on a 27-node ladder-like network, which is the extended version of the topology adopted in [Miyamura2004a, Miyamura2004b]. As seen in Fig. 3.12, two domains (area or AS) are interconnected through an AAPN. We suppose that the call requests arrive at the network following a Poisson process, and the call holding time is exponentially distributed. We further assume that all the inter-domain source-destination node pairs have the same traffic load, and also all the intra-domain node pairs. We assume that 60% of the overall network traffic as inter-domain traffic. We call the links within each domain the normal links and assign to all the same capacity. We use a time-slot (e.g., 100Mbps) as the basic capacity unit on each normal link and the AAPN internal fiber links. We take the overall (intra- and inter-domain) call blocking probability as our performance metric. The simulation time is set long enough to achieve a sufficiently small 95% confidence interval in all the simulation trials.

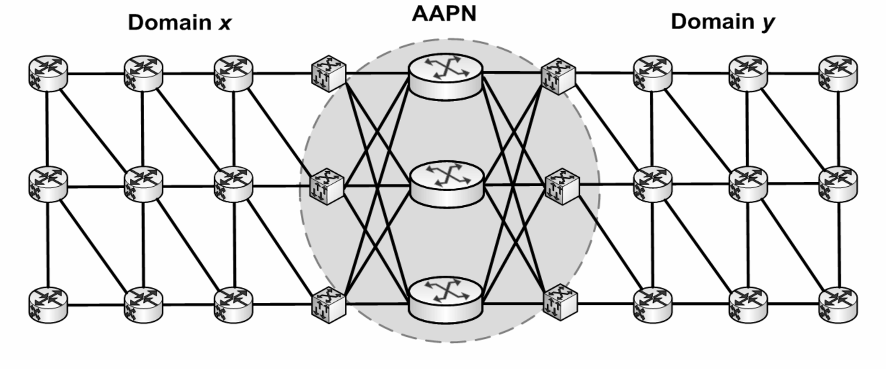


Figure 3.12: Network topology used for simulation (27 nodes and 120 directional links)

3.4.1. Single Path Routing

For single path routing, we assume the bandwidth requirement of each call connection is one single time-slot. Least-cost routing is adopted for path selection, where the cost of a path is defined as the sum of the costs of all the links along the path, and the link cost is defined as the inverse of the residual bandwidth of the link. A call is accepted only when there exists a path with enough available bandwidth. In the first experiment (Fig. 3.13), we fix the bandwidth of the AAPN internal fiber links (100 time-slots) while varying the bandwidth of the normal links in the two domains. This is for the purpose of exploring the effective range of each compared TE approach. We increase the capacities of the normal links gradually (Fig. 3.13) to simulate the phenomenon that the blocking of inter-domain calls is mainly due to lack of capacity in the AAPN. Similarly, by decreasing the capacities of normal links, we simulate the phenomenon of inter-domain blocking caused by the lack of capacity in the head-end and/or tail-end domains. (Note that many situations could lead to capacity lack in the real world, e.g., dynamic change of intra/inter-domain traffic, link failures, or not well-engineered network capacity, etc.). Since we use the ideal case as the benchmark for performance comparison, we further adjust the traffic amount so that the blocking probability of the ideal case is kept at around 1% for each given normal link capacity. That’s why the blocking curve of the ideal case is a flat line in Fig. 3.13.

The per-domain approach performs worst among the compared approaches. In Fig. 3.13, we notice the blocking curve of the per-domain approach is a near-flat line with the highest values of blocking probability among all the approaches. This is due to its path computation mechanism: domain by domain, sequentially. Referring to the bottom of Fig. 3.6, each of the three segments of an inter-domain path is computed only on the basis of its own domain’s TE information. The starting node of the second or third segment is thus “blindly” determined by the previous segment. If there was a “bottleneck” in the AAPN or tail-end domain, the per-domain approach can not avoid it.

Our approach performs best (except for the ideal case) among the compared approaches. As seen in Fig. 3.13, the blocking curve of our approach is flat and very close to the curve of the ideal case in the full value range of normal link capacity. This also shows the robustness property (wide effective range) of our approach to the change of network environment, e.g., dynamic traffic change, link failure, etc. The small performance difference to the ideal curve is due to the use of approximated (e.g., abstracted/ aggregated) TE information in the AAPN.

As illustrated clearly in Fig. 3.13, the source-view-extension and destination-view-extension approaches have opposite behaviors/effective ranges under varying normal link capacities. When the normal links have the major contribution (lower capacities) to the inter-domain call blocking, the destination-view-extension approach performs better than the source-view-extension approach. Referring to the bottom of Fig. 3.6, this is because the computation of the first segment of an inter-domain path in the source-view-extension approach does not consider any TE information of the destination domains. For a blindly-given ingress border node of the tail-end domain, it is not easy, sometimes impossible, to work around the bottleneck links in the tail-end domain. While for the destination-view-extension approach, although the first segment is computed without any information of the tail-end domain, the ingress border node of the tail-end domain can be chosen to a large extent freely. This is because the second segment in the destination-view-extension approach includes the AAPN and the AAPN can connect any egress border node of the head-end domain to any ingress border node of the tail-end domain if the capacity permits. For the same reason, in an extreme case where all the inter-domain blocking is due to the head/tail-end domain (the very left end of Fig. 3.13), we observe that (1) the blocking curves of the destination-view-extension approach, the ideal case, and our approach merge; (2) the blocking curves of the source-view-extension approach and per-domain approach merge.

When increasing the capacities of the normal links (right-hand-side of Fig. 3.13), the AAPN internal fiber links become the major contributors to the inter-domain call blocking. Then the source-view-extension approach starts to work better than the destination-view-extension approach. Similar as above, this is due to the fact that the computation of the first segment of an inter-domain path in the source-view-extension approach considers the TE information of the AAPN so that the “bottleneck” in the AAPN can be avoided, while the destination-view-extension approach can not do that. Again, in another extreme scenario where the inter-domain call blocking is fully due to the AAPN (the very right end of Fig. 3.13), the blocking curves of the ideal case and the source-view-extension merge. While the curve of our approach is still close to the ideal one since our approach uses approximated TE information of the AAPN to compute the inter-domain path.

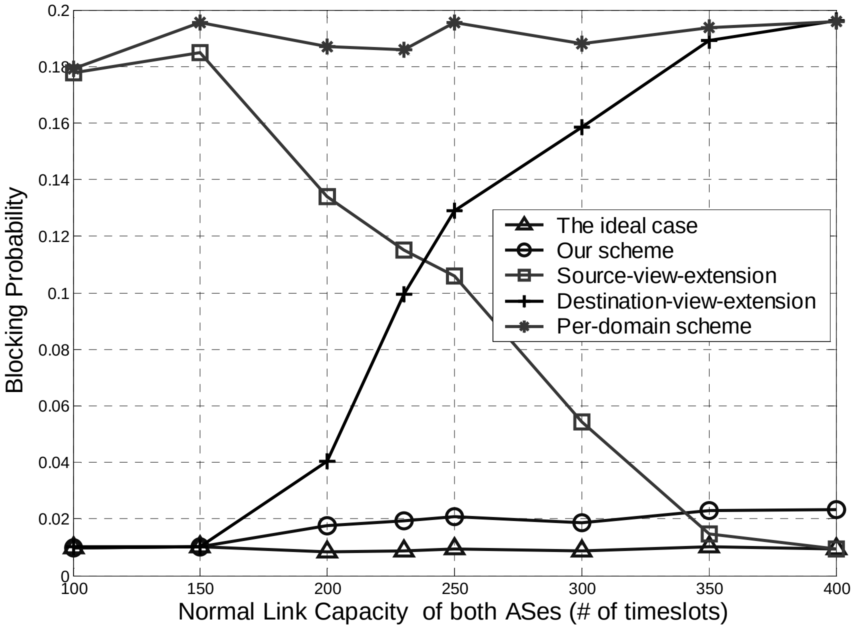


Figure 3.13: Overall network blocking probability of single path routing under varying normal link capacities. (95% confidence interval: ±0.05%)

In our second experiment, we study the blocking performance of the compared TE approaches under varying traffic loads (Fig. 3.14-3.16) and with three selected normal link capacities referring to the various ranges in Fig. 3.13). As seen in Fig. 3.14-3.16, when the traffic grows, the blocking probability increases in all the compared approaches. Our approach always works very well as the traffic loads change and remains close to the ideal case. The per-domain approach always performs worst, which is expected and mainly due to its path computation mechanism. For the three normal link capabilities, the source- and destination-view-extension approaches have various behaviors, that is, source-view-extension performs better in Fig. 3.15, worse in Fig. 3.16, and the two approaches perform close in Fig. 3.14. All these observations coincide with Fig. 3.13 and fully follow our analysis to Fig. 3.13.

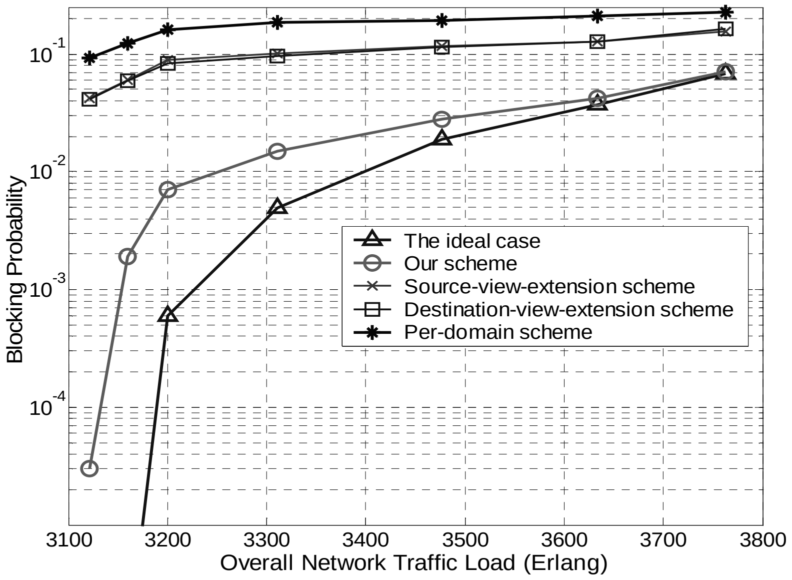


Figure 3.14: Overall network blocking probability of single path routing under various traffic loads with normal link capacity in both domains as 240 timeslots.

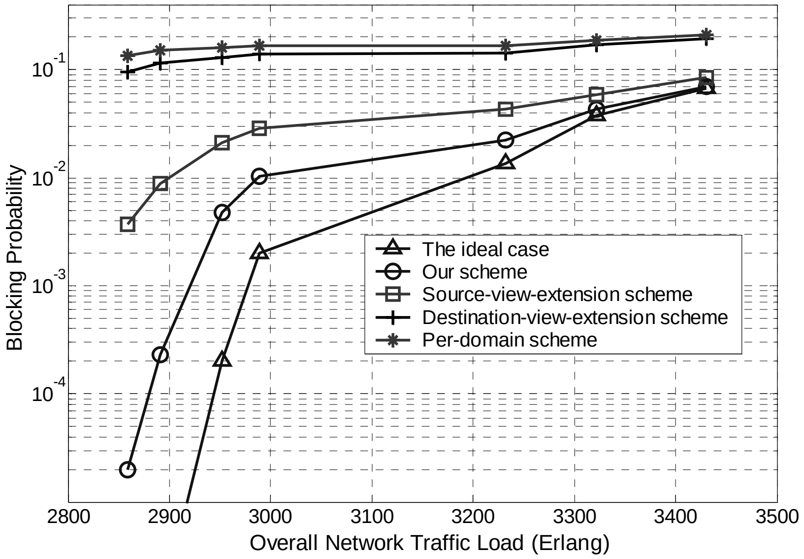


Figure 3.15: Overall network blocking probability of single path routing under various traffic loads with normal link capacity in both domains as 190 timeslots.

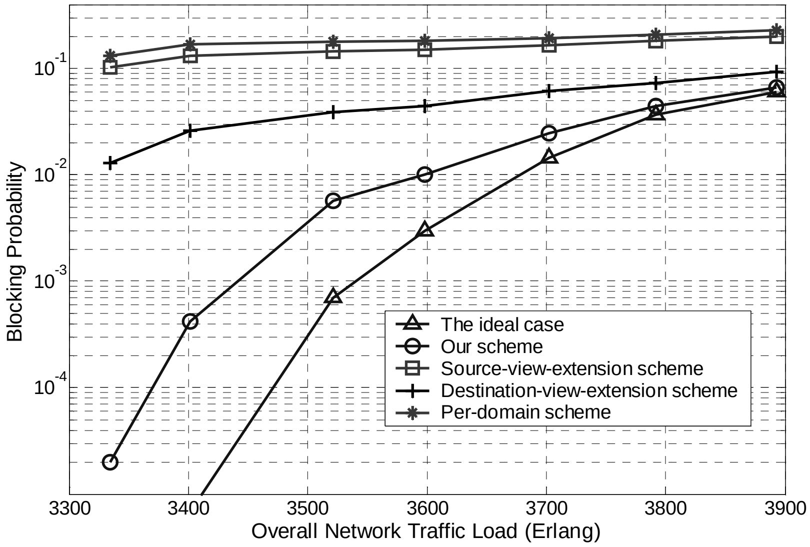


Figure 3.16: Overall network blocking probability of single path routing under various traffic loads with normal link capacity in both domains as 300 timeslots.

3.4.2. Diverse Routing

The purpose of diverse routing is load-sharing or end-to-end protection. We define the path diversity as link disjointness in the head-end/tail-end ISP domains, and (edge-, core-) node disjointness in the AAPN. Each call requires two diverse paths with the same bandwidth requirements (one timeslot for each), and the call is accepted only when both the two diverse paths are available. We still adopt least-cost routing in which the path cost is the sum of the costs of the two diverse paths of a call. The link cost is the same as before. In the simulation, note that the two diverse paths or diverse path segments in each domain are computed simultaneously (not sequentially) to avoid the well-known “trap problem” in diverse path computation.

The simulation results of diverse routing are presented in Fig. 3.17, which is similar to Fig. 3.13. We notice that the performance of our approach is very close to the ideal case. It shows that the Star-TE applied AAPNs work very well not only for inter-domain single-path routing, but also for inter-domain diverse-path routing. Meanwhile, when fixing the link (normal link and fiber link) capacity and varying the traffic load as in Fig. 3.18, the phenomena very similar to Fig. 3.14 is observed.

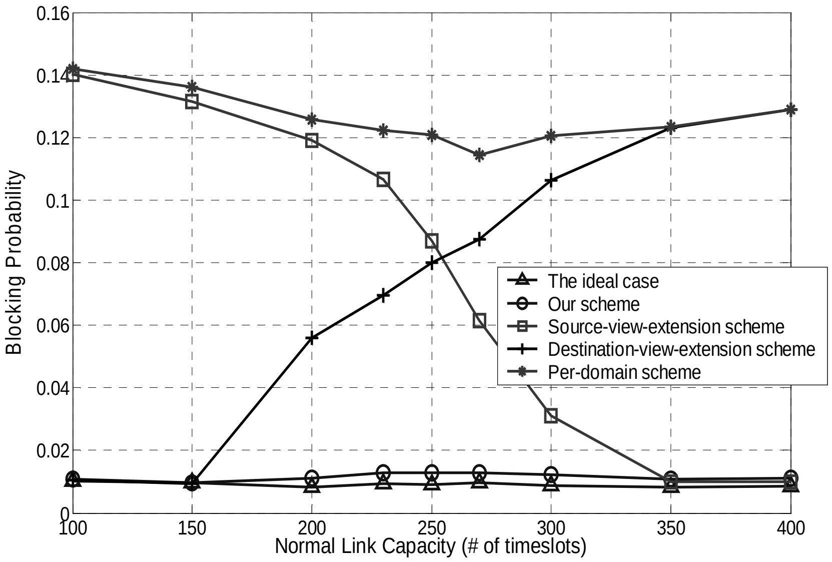


Figure 3.17: Overall network blocking probability of diverse path routing under varying normal link capacities. (95% confidence interval: ±0.05%)

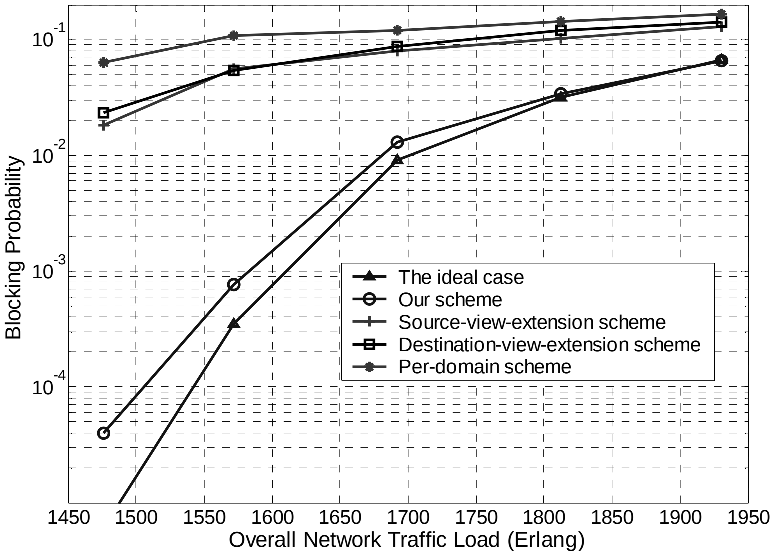


Fig. 3.18: Overall network blocking probability of diverse path routing under various traffic loads with normal link capacity in both domains as 250 timeslots.

3.4.3. The Amount of AAPN’s Core Nodes and Edge Nodes

Based on the topology in Fig. 3.12, we simulate two scenarios of upgrading the AAPN (increase its capacity) by gradually adding core or edge nodes (and associated fiber links), starting with two core nodes. We fix the capacity of internal fiber links (including the new added ones). We also fix the number of edge or core nodes when upgrading the other one, then adjust the traffic amount and capacities of normal links such that the blocking probability of the ideal case in the upgraded configuration is kept around 1% (as a benchmark). Table 3.2 and 3.2 list the blocking probability difference between our AAPN-based proposal and the ideal case under various amounts of core/edge nodes:

TABLE 3.2: THE DIFFERENCE OF THE BLOCKING PROBABILITY FOR OUR AAPN-BASED PROPOSAL AND THE IDEAL CASE WITH 3 AAPN EDGE NODES (PER DOMAIN)

# of Core Nodes	**2**	**3**	**5**	**7**	**9**	**12**	**15**
Blocking Difference	1.4%	1.0%	0.8%	0.6%	0.5%	0.5%	0.5%

As shown in Table 3.2, the blocking probability differences are small in general. As the number of core nodes increases, the blocking of our AAPN-based inter-domain routing framework is getting closer to the ideal case. This is because more core nodes would bring more chances to connect two edge nodes. Meanwhile, we also notice that the trend of “closing to the ideal case” is in saturation when the number of core nodes becomes large, e.g., 7.

When upgrading AAPN by adding edge nodes, shown in Table 3.3, we notice that the blocking differences increases slowly. This is due to the fact that our AAPN-based inter-domain routing framework adopted approximated TE information and more edge nodes lead to less precise TE information. On the other hand, more edge nodes increase the successful chance for an inter-domain call request. Hence, when the amount of edge nodes reaches a certain value (e.g., five in Table 3.3), there is also a “saturation” phenomenon as in Table 3.2. Meanwhile, in the AAPN architecture, more edge nodes will unavoidable require more core nodes, which will, on the other hand, balance the performance. Afterall, AAPN is suitable to provide inter-connectivity among domains; it would be better to distribute the AAPN edge nodes over many domains, instead of putting too many edge nodes in one or a few domains.

TABLE 3.3: THE DIFFERENCE OF THE BLOCKING PROBABILITY FOR OUR AAPN-BASED PROPOSAL AND THE IDEAL CASE WITH 3 AAPN CORE NODES

# of Edge Nodes →	**2**	**3**	**4**	**5**	**6**
Blocking Difference →	**0.5%**	**1.0%**	**1.7%**	**2.0%**	**2.1%**

3.5. Analytical Model for Blocking Probability

In this section, we develop an analytical performance model for the blocking probability of our inter-area optimal routing framework (Section 3.2). It also applies to the inter-AS scenario in Section 3.3. Our performance model is different from most of the previous analytical work on network performance modeling in two aspects:

- It is a model of *inter-domain* dynamic routing AND there is no node in the multi-domain network that has global TE information. Note that if the nodes have global information, then it is equal to the single-domain network case. Although the analytical modeling for single-domain networks has been studied extensively, there are very few results on inter-domain networks.
- We consider the *overlapping* among inter-domain routes (via shared common links) employed by an inter-domain source-destination node pair. The overlapping is due to the fact that the inter-domain routes in multi-domain networks may consist of a much larger number of hops and there are typically a large number of possible routes between source and destination nodes. Most of the previous work avoided considering the overlapping among routes used by the same source-destination node pair mainly because this can simplify the analytical work by assuming the routes (of an source destination node pair) are independent to each other [Chung1993, Mitra1993, Greenberg1997, Li2004, Chu2005]. Overlapping greatly complicates the performance analysis and

blocking probability computation. Just a few papers take into account the overlapping among routes [Greenberg1997, Liu2004b], but they only focus on routing in single-domain networks.

3.5.1. Assumptions

- (**A1**): We consider that a multi-area network typically consists of a non-backbone area (Area 1, arbitrary topology) inter-connected with another non-backbone area (Area 2, arbitrary topology) through an AAPN as the backbone area (see Fig. 3.19 for an example).
- (**A2**): We only consider inter-area traffic in the above multi-area network.
- (**A3**): The optimal inter-area routing framework we developed in Section 3.2 applied to the above multi-area network.
- (**A4**): Calls for a node pair arrive according to an independent stationary Poisson process. The duration of each call is exponentially distributed with a mean of one unit ($1/\mu = 1$).
- (**A5**): We use a time-slot as the basic unit for the capacity of either normal links in the non-backbone area or AAPN internal fiber links. Each call requires a full time-slot on each link of its path.
- (**A6**): Following our inter-area dynamic routing framework, an inter-area route is divided into two segments with one in each (extended) area, respectively (see Fig. 3.20 for an example). We assume the selecting of each segment of an inter-area route in its (extended) area is independent to the selecting in another (extended) area.
- (**A7**): Routes in an inter-area route-set are disjoint either in one of the two extended areas, or both (see Fig. 3.20).
- (**A8**): We assume that each AAPN internal fiber has one wavelength and the same amount of time-slots on that wavelength. The time-slot assigned to a route is chosen uniformly randomly from the set of idle time-slots. The assumption makes all time-slots identical and the analysis tractable.

- (**A9**): The time-slot continuity constraint must be met since there is no time-slot-interchanger in the AAPN core nodes.

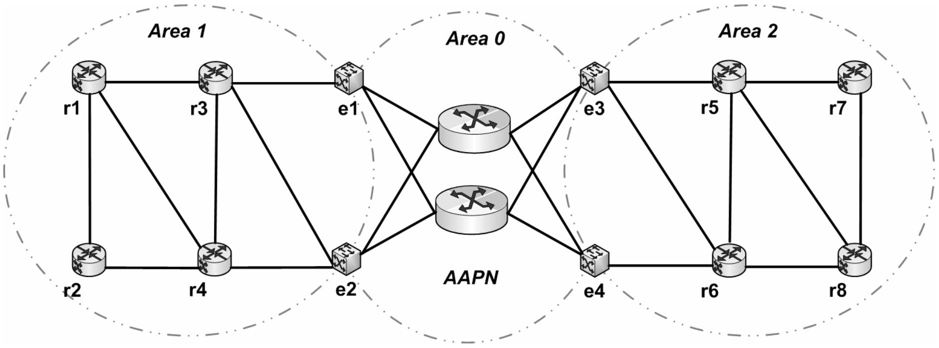


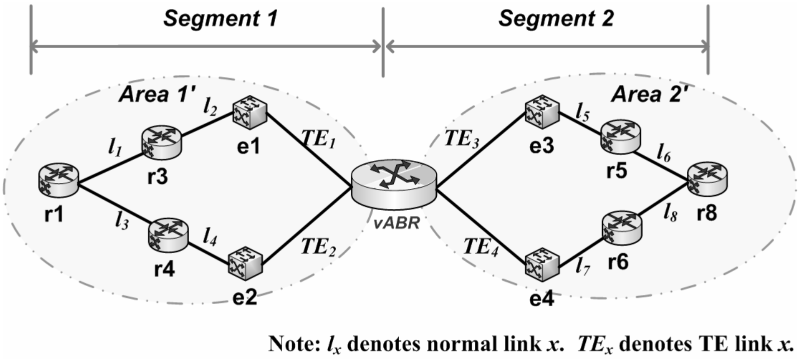
Figure 3.19: An example of multi-area network.

3.5.2. Notations

As a general rule, we use superscripts to indicate a route, or a route-set associated with a source-destination node pair, while subscripts are needed to indicate a particular link along a route, or a particular route in a route-set. We define the following notations:

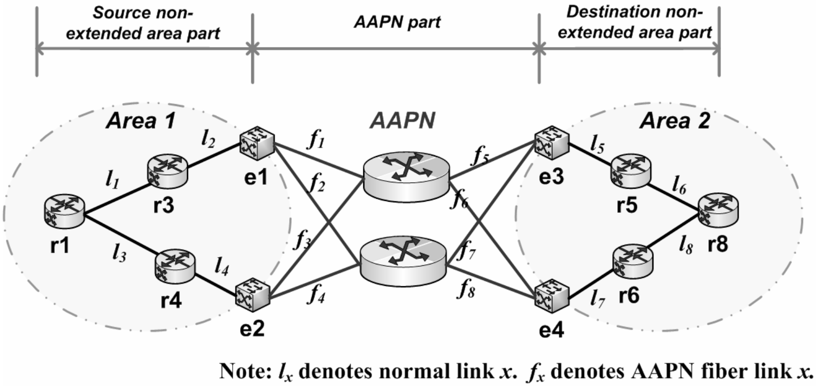
- N : the set of all the normal nodes in Area 1 and Area 2.
- $\Gamma(m,n)$: indicator function, $m,n \in \mathrm{N}$. $\Gamma(m,n)=0$ indicates node m and node n are not in the same area.
- N_C : the number of core nodes in AAPN.
- $B_{m,n}$: the blocking probability of the inter-domain source-destination node pair node m to node n, $\Gamma(m,n)=0$.

According to our inter-area dynamic routing framework (Section 3.2), an inter-area route is selected first at the *logical-view topology* (non-backbone areas are extended), where the AAPN internal fibers are abstracted as TE links and the overlaid core nodes are abstracted as one virtual-ABR (vABR). Fig. 3.20 illustrates the logical-view of an inter-area source-destination node pair in Fig. 19, *r1* to *r8*. As shown in Fig. 3.20, an *r1*-to-*r8* route consists of two parts, one in (extended) Area 1 and one in (extended) Area 2, connected by vABR (virtual Area Border Router). Actually, the route-set associated with the *r1*-to-*r8* node pair in the logical-view topology can be considered as a full combination of the two *intra-area* route-sets, one is *r1*-to-vABR, and another one is vABR-to-*r8*. We assumed that the routes within either of the two intra-area route-sets are *disjointed* **(A7),** which is reasonable and adopted by most previous work on network performance modeling [Chung1993, Mitra1993, Greenberg1997, Li2004, Chu2005]. But the inter-domain routes in an inter-area route-set (e.g., *r1*-to-*r8*) could overlap with each other and the amount of inter-domain routes in a route-set is the product of the amounts of routes in the two related intra-area route-sets (e.g., to-vABR and from-vABR).



Note: l_x denotes normal link *x*. TE_x denotes TE link *x*.

Figure 3.20: Logical level view for a source-destination node pair: *r1* to *r8*.



Note: l_x denotes normal link *x*. f_x denotes AAPN fiber link *x*.

Figure 3.21: Physical-view for a source-destination node pair: *r1* to *r8*

- $\overline{RS_{m,n}}$, $RS_{m,n}$: the inter-area route-sets from node *m* to node *n*, $\Gamma(m,n)=0$, in the logical-view topology and physical-view topology, respectively. $\left\|\overline{RS_{m,n}}\right\|$, $\left\|RS_{m,n}\right\|$ are the amount of routes in the associated route-sets, respectively.
- $\overline{RS_{m,vABR}}$: the intra-area route-sets from node *m* to *vABR* in the logical-view topology.

- $\overline{RS_{vABR,m}}$: the intra-area route-set from *vABR* to node *m* in the logical-view topology. $\left\|\overline{RS_{m,n}}\right\| = \left\|\overline{RS_{m,vABR}}\right\| \times \left\|\overline{RS_{vABR,n}}\right\|$ and $\left\|RS_{m,n}\right\| = \left\|\overline{RS_{m,vABR}}\right\| \times \left\|\overline{RS_{vABR,n}}\right\| \times N_C$.
- $\overline{R^i_{m,n}}$, $R^i_{m,n}$: the *i*th route in route-set $\overline{RS_{m,n}}$ and $RS_{m,n}$, respectively.
- $\overline{R^i_{m,vABR}}$, $R^i_{m,vABR}$: the *i*th logical-view route in the route-set $\overline{RS_{m,vABR}}$ and its physical-view ***mapping***, respectively. Note that $\left\|\overline{R^i_{m,vABR}}\right\| < \left\|R^i_{m,vABR}\right\|$.
- $L^{i,k}_{m,vABR}$: *k*th link along its physical-view mapping $R^i_{m,vABR}$.
- $R^i_{m,vABR} = \left\{ \underbrace{l_1,\dots,l_{\left\|R^i_{m,vABR}\right\|-N_C}}_{\text{normal links}}, \underbrace{l_{\left\|R^i_{m,vABR}\right\|-N_C+1},\dots,l_{\left\|R^i_{m,vABR}\right\|}}_{\text{AAPN fiber links}} \right\}$.
- $R^i_{vABR,n} = \left\{ \underbrace{l_1,\dots,l_{N_C}}_{\text{AAPN fiber links}}, \underbrace{l_{N_C+1},\dots,l_{\left\|R^i_{m,vABR}\right\|}}_{\text{normal links}}, \right\}$.
- $\overline{R^i_{m,n}} = \left\{ \overline{R^j_{m,vABR}}, \overline{R^k_{vABR,n}} \right\}$.

We use the example illustrated in Fig. 3.20 and 3.21 to explain the above notations. In Fig. 3.20 and 3.21, we have: $N_C = 2$, $\left\|\overline{RS_{r1,r8}}\right\| = 4$, $\left\|RS_{r1,r8}\right\| = 8$, $\left\|\overline{RS_{r1,vABR}}\right\| = 2$, $\left\|\overline{RS_{vABR,r8}}\right\| = 2$. $\overline{R^1_{r1,vABR}} = \left\{ \underbrace{l_1,l_2}_{\text{normal links}}, \underbrace{TE_1}_{\text{TE link}} \right\}$, and its physical-view mapping, $R^1_{r1,vABR}$, is $\left\{ \underbrace{l_1,l_2}_{\text{normal links}}, \underbrace{f_1,f_2}_{\text{fiber links}} \right\}$. $\overline{R^2_{r1,vABR}} = \left\{ \underbrace{l_3,l_4}_{\text{normal links}}, \underbrace{TE_2}_{\text{TE link}} \right\}$, and its physical-view mapping, $R^2_{r1,vABR}$, is $\left\{ \underbrace{l_3,l_4}_{\text{normal links}}, \underbrace{f_3,f_4}_{\text{fiber links}} \right\}$. Similarly, we have $\overline{R^1_{vABR,r8}} = \left\{ \underbrace{TE_3}_{\text{TE link}}, \underbrace{l_5,l_6}_{\text{normal links}} \right\}$, and its physical-view mapping, $R^1_{vABR,r8}$, is $\left\{ \underbrace{f_5,f_6}_{\text{fiber links}}, \underbrace{l_5,l_6}_{\text{normal links}} \right\}$. $\overline{R^2_{vABR,r8}} = \left\{ \underbrace{TE_3}_{\text{TE link}}, \underbrace{l_7,l_8}_{\text{normal links}} \right\}$, and its physical-view mapping, $R^2_{vABR,r8}$, is $\left\{ \underbrace{f_7,f_8}_{\text{fiber links}}, \underbrace{l_7,l_8}_{\text{normal links}} \right\}$.

- X_j, $X^i_{m,vABR}$, $X^i_{vABR,n}$: the state of link *j*, route $\overline{R^i_{m,vABR}}$, and route $\overline{R^i_{vABR,n}}$. Link state is defined as the free/available capacity of the link in terms of time-slots. Route state is the least free capacity (in terms of time-slots) among all the links along the route. Note that computing the state of a logical-view route is actually computing the state of the associated mapped physical-view route.
- $\Lambda^i_{m,n}$: the probability that the inter-area route $\overline{R^i_{m,n}}$ is selected among its route set $\overline{R_{m,n}}$. Note that $\sum_i \Lambda^i_{m,n} = 1$. Similarly, we have $\Lambda^j_{m,vABR}$ and $\Lambda^k_{vABR,n}$.
- $\Omega^i_{m,n}$: the probability that there is at least one bandwidth unit available along the inter-area route $\overline{R^i_{m,n}}$.
- C_N : the capacity of a normal link in terms of time slots. We assume all the normal links have the same capacity to simply the expression.
- C_A : the capacity of an AAPN internal fiber link in terms of time slots.
- $q_{l,t}$: the probability that exactly *t* time-slots are idle on link *l*, i.e., $q_{l,t} = \Pr\{X_l = t\}$.
- $\alpha_{l,t}$: the carried-traffic arrival rate on link *l* given link state as $X_l = t$.
- $\lambda_{m,n}$: the offered load from node m to node n, $\Gamma(m,n) = 0$.

3.5.3. Blocking Probability of Each Source-Destination Pair

In our dynamic routing environment, the end-to-end inter-area blocking probability from node *m* to node *n* ($\Gamma(m,n) = 0$) should be computed theoretically according to Equation (3.1). However, since the event of selecting an inter-area route is, generally, not independent of the event that the route is in an available state, the computation directly according to Equation (3.1) becomes very difficult, especially when a route has many hops and/or high capacity [Liu2004b].

$$B_{m,n} = 1 - \sum_{\substack{\text{all the routes}\\ \text{from } m \text{ to } n}} \Pr\left\{\text{an inter-area route is selected AND it is in available state}\right\} \tag{3.1}$$

Hence we adopt an approximation that was widely used in other papers [Chung1993, Mitra1993, Greenberg1997, Li2004, Chu2005]: ignoring the dependency of the event that a route is selected and the fact that it is in an available state, and assuming that these probabilities are *independent* (**A10**). Then we have

$$\begin{aligned} B_{m,n} &= 1 - \sum_{i=1}^{\left\|\overline{RS_{m,n}}\right\|} \left(\Pr\left\{ \overline{R^{i}_{m,n}} \text{ is selected } \right\} \times \Pr\left\{ \overline{R^{i}_{m,n}} \text{ is in available state} \right\} \right) \\ &= 1 - \sum_{i=1}^{\left\|\overline{RS_{m,n}}\right\|} \left(\Lambda^{i}_{m,n} \times \Omega^{i}_{m,n} \right) \end{aligned} \tag{3.2}$$

Following our inter-area dynamic routing framework, an inter-area path is selected in the (extended) Area 1 and (extended) Area 2 independently (**A6**) (Fig. 20). Then we have

$$\begin{gathered} \Lambda^{i}_{m,n} = \Lambda^{j}_{m,vABR} \times \Lambda^{k}_{vABR,n}, \text{ where} \\ i \in \left[1, 2, ..., \left\|\overline{RS_{m,n}}\right\|\right],\ j \in \left[1, 2, ..., \left\|\overline{RS_{m,vABR}}\right\|\right], k \in \left[1, 2, ..., \left\|\overline{RS_{vABR,n}}\right\|\right]. \end{gathered} \tag{3.3}$$

Note that we adopt the well-know LLR (Least-Loaded Routing) dynamic routing scheme to select a route in each intra-area route-set, that is, the intra-area route with the maximal route state will be selected in its intra-area route-set. The LLR scheme is used in almost all the previous papers on network performance modeling with dynamic routing [Chung1993, Mitra1993, Greenberg1997, Li2004, Liu2004b, Chu2005].

Based on the assumption (**A6**) and our inter-area optimal routing framework, we use Equ. (3.3) to compute the probability that an inter-area route is selected among the overlapped routes in an inter-area route-set. Although there are two papers [Greenberg1997] and [Liu2004] that considered the overlapping among routes, their computation techniques are either only valuable to the sequential routing [Greenberg1997] or not covering all the overlapping patterns in our case. Hence they can not be used by us.

$$\begin{aligned}\Lambda^{j}_{m,vABR} &= \Pr\left\{\overline{R^{j}_{m,vABR}} \text{ is selected in } \overline{RS_{m,vABR}}\right\} \\ &= \sum_{w=0}^{\min(C_N,C_A)}\left(\Pr\left\{X^{j}_{m,vABR}=w\right\}\times\Pr\left\{X_{all\ other\ routes\ in\ \overline{RS_{m,vABR}}}\le w\right\}\right) \\ &= \sum_{w=0}^{\min(C_N,C_A)}\left(\Pr\left\{X^{j}_{m,vABR}=w\right\}\times\prod_{h=1,h\neq j}^{\left\|\overline{RS_{m,vABR}}\right\|}\Pr\left\{X^{h}_{m,vABR}\le w\right\}\right) \\ &= \sum_{w=0}^{\min(C_N,C_A)}\left(\Pr\left\{X^{j}_{m,vABR}=w\right\}\times\prod_{h=1,h\neq i}^{\left\|\overline{RS_{m,vABR}}\right\|}\sum_{v=0}^{w}\Pr\left\{X^{h}_{m,vABR}=v\right\}\right)\end{aligned} \tag{3.4}$$

Similar, we have

$$\Lambda^{k}_{vABR,n} = \sum_{w=0}^{\min(C_N,C_A)}\left(\Pr\left\{X^{k}_{vABR,n}=w\right\}\times\prod_{h=1,k\neq k}^{\left\|\overline{R_{vABR,n}}\right\|}\sum_{v=0}^{w}\Pr\left\{X^{h}_{vABR,n}=v\right\}\right) \tag{3.5}$$

In order to guarantee the availability of an inter-area path, all the three parts (namely *source non-extended area part*, *AAPN part*, *destination non-extended area part*, see Fig. 19) along it must be available at the same time. In addition, for the AAPN part, time slot continuity must be held and there should be at least one optical path (an edge node to a core node to another edge node) available. Hence $\Omega^{i}_{m,n}$ is computed based on the physical-view mapping ($R^{j}_{m,vABR}$ and $R^{k}_{vABR,n}$) of $\overline{R^{i}_{m,n}}$. Then we have

$$\Omega^{i}_{m,n} = \underbrace{\left[\prod_{h=1}^{\left\|R^{j}_{m,vABR}\right\|-N_C}\left(1-q_{L^{j,h}_{m,vABR},0}\right)\right]}_{source\ non\text{-}extended\ Area}\times\underbrace{\left[1-\prod_{f=1}^{N_C}\left(\sum_{x=0}^{C_A}\sum_{y=0}^{C_A}\begin{matrix}\Pr\left\{X_{a,b}=0\mid X_a=x,X_b=y\right\}\times\\ \Pr\left\{X_a=x\right\}\times\Pr\left\{X_b=y\right\}\end{matrix}\right)\right]}_{AAPN}\times \tag{3.6}$$

$$\underbrace{\left[\prod_{h=N_C+1}^{\left\|R^{k}_{vABR,n}\right\|}\left(1-q_{L^{k,h}_{vABR,n},0}\right)\right]}_{destination\ non\text{-}extended\ Area},\quad \text{where } a=L^{j,\left\|R^{i}_{m,vABR}\right\|-N_C+f}_{m,vABR} \text{ and } b=L^{k,f}_{vABR,n}.$$

Note that both a and b are AAPN internal fiber links.

$\Pr\left\{X_{a,b}=0\mid X_a=x,X_b=y\right\}$ is the probability that there is no common free time-slots along fiber link a and b (considering time-slot continuity) given x free time-slots in link a and y free time-slots in link b. This probability can be computed according to Equ. (3.7) [Birman1996] and just let z be zero.

$$\Pr\left\{X_{a,b}=z\mid X_a=x,X_b=y\right\}=\binom{y}{x}\times\left(\prod_{i=1}^{x}\frac{z-i+1}{C_A-i+1}\right)\times\left(\prod_{i=1}^{y-x}\frac{C_A-z-i+1}{C_A-x-i+1}\right) \tag{3.7}$$

3.5.4. Link State Probability

The number of idle time-slots on link *l*, either a normal link or a fiber link, can be viewed as a birth-death process. The arriving and serving behavior on link *l* forms an M/M/N/N system. Since all the states in the associated Markov chain are ergodic, the equilibrium state distribution of the chain can be derived as follows.

$$q_{l,t} = \frac{C_l(C_l\text{ -}1)...(C_l\text{ -}t+1)}{\alpha_{l,1}\alpha_{l,2}...\alpha_{l,t}} \times q_{l,0}\,,\ if\ \ t\text{=1,2,3, ..., } C_l\,;\ \ q_{l,t}=0,\ if\ \ t>C_l\,; \tag{3.8}$$

$$q_{l,0} = \left[1+\sum_{t=1}^{C_l}\frac{C_l(C_l-1)...(C_l-t+1)}{\alpha_{l,1}\alpha_{l,1}...\alpha_{l,t}}\right]^{\text{-}1},\ \ C_l = C_N\ \ or\ \ C_A\,. \tag{3.9}$$

Given the link states, we can compute the route state, $X^{\,j}_{m,vABR}$, according to Equ. (3.10, 3.11). The very similar two equations for the computation of route state $X^{\,k}_{vABR,n}$ can also be easily derived in the same way.

$$\Pr\left\{X^{\,j}_{m,vABR}=x\right\}=\sum_{x_1=0}^{C_N}\cdots\sum_{x_{\left\|R^j_{m,vABR}\right\|-N_C}=0}^{C_N}\ \sum_{y_1=0}^{C_A}\cdots\sum_{y_{N_c}=0}^{C_A}\left\{\begin{array}{c}\Phi\left(x;x_1,...,x_{\left\|R^j_{m,vABR}\right\|-N_C},y_1,...,y_{N_c}\right)\times\\ \underbrace{\prod_{h=1}^{\left\|R^j_{m,vABR}\right\|-N_C}q_{L^{j,h}_{m,vABR},x_j}}_{\text{normal links in }R^j_{m,vABR}}\times\underbrace{\prod_{h=1}^{N_C}q_{L^{j,h}_{m,vABR},y_k}}_{\text{AAPN fiber links in }R^j_{m,vABR}}\end{array}\right\} \tag{3.10}$$

where,

$$\Phi\left(x;x_1,...,x_{\|r\|-N_C},y_1,...,y_{N_c}\right)=\begin{cases}1,\ \ if\ \ \min\left(x_1,...,x_{\left\|R^j_{m,vABR}\right\|-N_C},\max\left(y_1,...,y_{N_c}\right)\right)=x\\ 0,\qquad\qquad\qquad\qquad\qquad\qquad otherwise\end{cases} \tag{3.11}$$

3.5.5. State-dependent Link Arrival Rate

$\alpha_{l,t}$ is the carried-traffic arrival rate on link *l* given link state as $X_l=t$. It is contributed by the carried traffic loads of all the inter-area routes that pass link *l*. That is, it is computed as the summation of the original external offered arrival rates *thinned* by blocking on the other links, hence knows as the *reduced load* method [Chung1993, Liu2004b].

$$\alpha_{l,t} = \begin{cases} \sum\limits_{\forall m \in N} \sum\limits_{\substack{\forall n \in N, \\ m \neq n, \\ \Gamma(m,n)=0}} \sum\limits_{\substack{i=1, \\ l \in R^i_{m,n}}}^{\left\|\overline{R_{m,n}}\right\|} \left(\lambda_{m,n} \times \Lambda^i_{m,n} \times \Omega^i_{m,n} \left(X_l = t \right) \right), \ if \ \ t \neq 0; \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad if \ \ t = 0. \end{cases} \tag{3.12}$$

$\Lambda^i_{m,n}$ is computed according to Equ. (3.3-3.5). $\Omega^i_{m,n}\left(X_l = t\right)$ is the conditional probability that route $\overline{R^i_{m,n}}$ is in an available state given the state of the link *l* ($l \in \overline{R^i_{m,n}}$) as *t* and the traffic will traverse link *l*. See Fig. 3.19 and 3.20 as reference, $\Omega^i_{m,n}\left(X_l = t\right)$ is computed according to Equ. (3.13-3.14).

$$\begin{aligned} &\Omega^i_{m,n}\left(X_l = t, t > 0\right) \\ &= \left[\prod_{h=1, L^{j,h}_{m,vABR} \neq l}^{\left\|R^j_{m,vABR}\right\| - N_C} \left(1 - q_{L^{j,h}_{m,vABR},0}\right) \right] \times \left[1 - \prod_{f=1}^{N_C} \left(\sum_{x=0}^{C_A} \sum_{y=0}^{C_A} \begin{matrix} \Pr\left\{X_{a,b} = 0 \mid X_a = x, X_b = y\right\} \times \\ \Pr\left\{X_a = x\right\} \times \Pr\left\{X_b = y\right\} \end{matrix} \right) \right] \times \\ &\left[\prod_{h=N_C+1}^{\left\|R^k_{vABR,n}\right\|} \left(1 - q_{L^{k,h}_{vABR,n},0}\right) \right], \qquad if \ \ l \in \text{source non-extended area part of } \overline{R^i_{m,n}}, \end{aligned} \tag{3.13a}$$

where $a = L^{j, \left\|R^i_{m,vABR}\right\| - N_C + f}_{m,vABR}$ and $b = L^{k,f}_{vABR,n}$

$$\begin{aligned} &\Omega^i_{m,n}\left(X_l = t, t > 0\right) \\ &= \left[\prod_{h=1}^{\left\|R^j_{m,vABR}\right\| - N_C} \left(1 - q_{L^{j,h}_{m,vABR},0}\right) \right] \times \left[1 - \prod_{f=1}^{N_C} \left(\sum_{x=0}^{C_A} \sum_{y=0}^{C_A} \begin{matrix} \Pr\left\{X_{a,b} = 0 \mid X_a = x, X_b = y\right\} \times \\ \Pr\left\{X_a = x\right\} \times \Pr\left\{X_b = y\right\} \end{matrix} \right) \right] \times \\ &\left[\prod_{h=N_C+1, L^{i,k}_{vABR,n} \neq l}^{\left\|R^k_{vABR,n}\right\|} \left(1 - q_{L^{k,h}_{vABR,n},0}\right) \right], \ if \ \ if \ \ l \in \text{destination non-extended area part of } \overline{R^i_{m,n}}, \end{aligned} \tag{3.13b}$$

where $a = L^{j, \left\|R^i_{m,vABR}\right\| - N_C + f}_{m,vABR}$ and $b = L^{k,f}_{vABR,n}$.

$$\begin{aligned} &\Omega^i_{m,n}\left(X_l = t, t > 0\right) \\ &= \left[\prod_{h=1}^{\left\|R^j_{m,vABR}\right\| - N_C} \left(1 - q_{L^{j,h}_{m,vABR},0}\right) \right] \times \theta^i_{m,n}\left(X_l = t\right) \times \left[\prod_{h=N_C+1}^{\left\|R^k_{vABR,n}\right\|} \left(1 - q_{L^{k,h}_{vABR,n},0}\right) \right], \end{aligned} \tag{3.13c}$$

$if \ \ l \in$ AAPN part of $\overline{R^i_{m,n}}$, where

$$\theta_{m,n}^{i}\left(X_l = t\right) \tag{3.14}$$

$$= \begin{cases} \left(1 - \left(\sum_{y=0}^{C_A} \Pr\left\{X_{a,b} = 0 \mid X_a = t, X_b = y\right\} \times \Pr\left\{X_b = y\right\}\right)\right) \times \dfrac{1}{N_C}, & if\ \ l = a \\ \left(1 - \left(\sum_{x=0}^{C_A} \Pr\left\{X_{a,b} = 0 \mid X_a = x, X_b = t\right\} \times \Pr\left\{X_a = x\right\}\right)\right) \times \dfrac{1}{N_C}, & if\ \ l = b \end{cases},$$

where $a = L_{m,vABR}^{j,\left\|R_{m,vABR}^{i}\right\|-N_C+f}$, $b = L_{vABR,n}^{k,f}$, $f \in \left[1, 2, ..., N_C\right]$.

Note that the $\dfrac{1}{N_C}$ in Equ. (3.14) is due to the assumption that given an edge node pair (one in the source area and one in the destination area), the core node connecting these two edge nodes is selected randomly. We make this assumption in order to simplify the complex computation of Equ. (3.13) in a reasonable way.

3.5.6. The network-wide blocking probability

The network-wide inter-area end-to-end blocking probability is calculated as follows:

$$B = \frac{\sum_{\forall m \in N} \sum_{\forall n \in N, \Gamma(m,n)=0} \lambda_{m,n} \times B_{m,n}}{\sum_{\forall m \in N} \sum_{\forall n \in N, \Gamma(m,n)=0} \lambda_{m,n}} \tag{3.15}$$

, where $B_{m,n}$ is computed according to Equ. (3.2).

3.5.7. Computation

From the above analysis, a set of non-linear coupled equations has been obtained for the computation of blocking probabilities. An iterative algorithm can be developed accordingly to find the solution by repeated substitution. The method of iterative substitution is described as follows:

Step 1. For all the inter-domain source-destination node pairs, initialize $\overline{\overline{B_{m,n}}}$ to zero. For all the links, initialize $\alpha_{l,0} = 0$ and set $\alpha_{l,t}$, $t > 0$ to an arbitrary value.

Step 2. Determine the link state of all the links $q_{l,t}$, using Equ. (3.8-3.9).

Step 3. Calculate route states for all the intra-area routes using Equ. (3.10-3.11).

Step 4. Calculate $\Lambda^{i}_{m,n}$ for all the inter-area routes using Equ. (3.3-3.5). Calculate $\Omega^{i}_{m,n}$ for all the inter-area routes using Equ. (3.6-3.7).

Step 5. Update $\alpha_{l,t}$, $t > 0$ for all the links using Equ. (3.12-3.14).

Step 6. Calculate $B_{m,n}$ for all the inter-domain source-destination node pairs using Equ. (3.2). If $\max\limits_{\substack{\forall m \in N \forall n \in N, \\ \Gamma(m,n)=0,}} \left| B_{m,n} - \overline{\overline{B_{m,n}}} \right| / \overline{\overline{B_{m,n}}} < \varepsilon$ then terminate. Otherwise, let $\overline{\overline{B_{m,n}}} = B_{m,n}$ and go back to Step 2. (ε is a very small real number that determines the number of iterations and the precision of the obtained results.)

3.5.8. Numerical Results

We demonstrate the accuracy of our analytical techniques by comparing analytical results with simulation results. Simulation results are plotted along with 95% confidence intervals estimated by long enough simulation time. For the analytical results, the iterative algorithm terminates when all blocking probability values have converged within 10^{-6} ($\varepsilon = 10^{-6}$). Both the analytical work and simulation experiments are conducted on the 14-node network topology shown in Fig. 3.19 (14 nodes with 48 directional links) and a 10-node network topology (by removing nodes *r1, r2, r7* and *r8* in Fig. 3.19). Both the topologies are organized as two non-backbone areas inter-connected by an AAPN. All the other simulation configurations are almost the same as our previous simulation work (least-cost routing) in Section 3.4, with the only exception that the cost of a route, R , is defined as $\left(\min(N_C, N_A) - X_R\right)$, (X_R is the state of route R), which is actually the LLR scheme.

Fig. 3.22 and Fig. 3.23 demonstrate the numerical results (blocking performance) in two networks obtained from the proposed analytical models. As expected, the network-wide blocking probabilities (both from the analytical model and the simulation) increase as the network load becomes heavier.

Fig. 3.22 and Fig. 3.23 compare the numerical results from the analytical model to those from simulation experiments over the 10-node network and the 14-node network, respectively. As we can see, the numerical results in the two figures conform closely to the simulation results and follow the trend of the simulation results. This exhibits the accuracy of our analytical model. Furthermore, numerical results from the analytical model are getting closer to the simulation results as the load is increased. The reason for this is that our analytical model adopts the classic reduced load approximation (see Equ. (3.12) in Section 3.5.5), and it is known that this approximation technique gives accurate results as the load is increased [Chung1993].

In addition, we also observe that the analytical model overestimates the blocking probabilities. This can be explained by the independency assumption of route selection (A10) we made in this analytical model. The independency assumptions ignore the correlation between a route being selected and its current state and also the correlation of selecting intra-area routes in one area and in another area. If the analytical model considers these correlations, then a more accurate estimation of the blocking probabilities can be obtained but with much higher computational complexity.

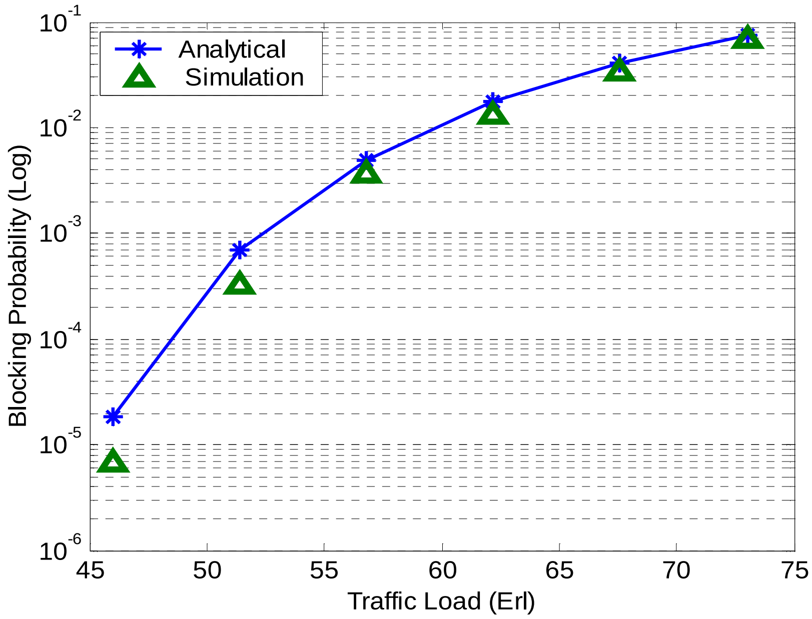


Figure 3.22: Network blocking probability over the 10-node network. (The capacity of a normal link is 18 time-slots; the capacity of an AAPN fiber link is 12 time-slots.)

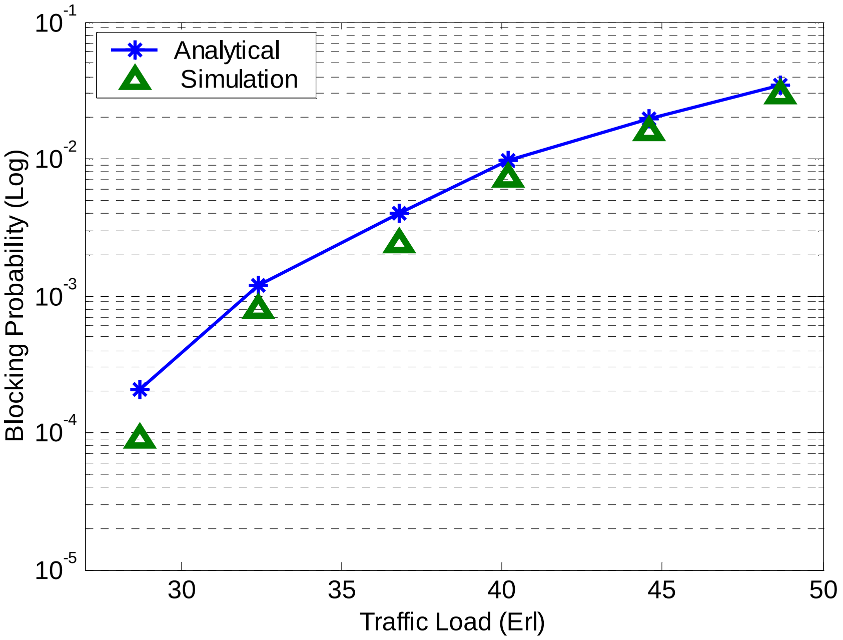


Figure 3.23: Network blocking probability over the 14-node network. (The capacity of a normal link is 10 time-slots; the capacity of an AAPN fiber link is 8 time-slots.)

4. Inter-Area Shared Segment Protection of MPLS Flows over AAPN

Optimal routing and associated protection, which are the two key issues of traffic engineering, become much more difficult in multi-area networks than in single area networks due to the inter-area information scalability and confidentiality issues. Furthermore, it is even severe for inter-area protection since the protection needs even more information compared to the routing.

In the previous chapter (Chapter 3), by deploying an agile all-photonic network (AAPN) as the backbone area, we developed a novel routing architecture that can provide globally-optimized inter-area routing with good compatibility to existing traditional IP/MPLS routers, but the protection issues were not considered yet. In this chapter, we focus on inter-area shared link and node protection in multi-area networks with an agile all-photonic backbone. On the basis of our inter-area optimal routing architecture (proposed in Section 3.2), we propose and develop a scalable inter-area shared segment-based protection (SSP) framework, which consists of three components, namely

1) the segment protection schemes (for the strict and a weakened single failure assumptions),

2) the management of supporting routing information and

3) a related signaling process.

Through *sharing*, we can utilize the network resource in an efficient way. Through *segment*-based protection, we can reduce the recovery time for inter-area connections. In addition, segment-based protection can help us to develop distributed routing information management to avoid the scalability issues related to multi-area networks. Meanwhile, for an inter-area connection, our SSP (segment-based protection) schemes provide not only link protection but also protection to failures of the "key" nodes along the working path. The key nodes include the edge and core nodes that act as ABR/v-ABR. This corresponds to the requirements of RFC 4105.

4.1. Proposed Schemes for Inter-Area Shared Segment-Based Protection

We consider dynamic (e.g., on-line fashion) inter-area shared segment protection routing that aims to optimally identify an inter-area working path and associated backup paths for each arriving connection request.

4.1.1. Inter-Area Shared Segment Protection Scheme (IASSP)

Our scheme belongs to the active path first (APF) approach and we adopt the single-failure assumption. When an inter-area connection request (with protection requirement) comes in, the following steps are performed:

- An optimal inter-area working path for this connection request is determined (using our optimal routing framework in Section 3.2).
- The working path is then divided into two *overlapping* (at the AAPN) protected half-paths (see Fig. 4.1 top).
- For each protected half-path, different protection techniques (link-, segment-, or path-based) can be applied independently.
- An extra optical cross-connection between the two AAPN edge nodes along the working path, but through a different core node, can be setup to provide further and instant protection against the failure of the core node or one of the two optical links along the working path. It is called "nested" protection.

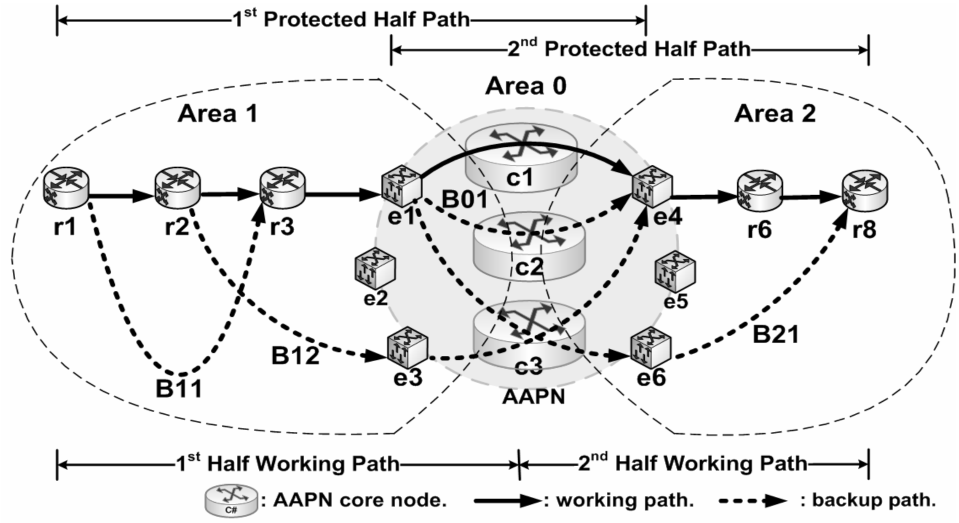


Figure 4.1: the Inter-Area shared segment-based protection scheme

As shown in Fig. 4.1, for an inter-area connection request (*r1* to *r8*), suppose the optimal working path is *r1->r2->r3->e1->v-ABR->e4->r6->r8* (solid arrow line in Fig.4.1). For the first protected half-path (*r1->r2->r3->e1->v-ABR->e4*), the source node, *r1*, which is in Area 1, is in charge of computing the associated optimal backup path(s). But according to our inter-area optimal routing framework (Section 3.2), the farthest node that *r1* can see is *v-ABR*, not *e4*. Hence we need *e4* to "act" as v-ABR, which means to export the necessary routing information to *r1* so that the backup paths can be computed optimally without breaking our inter-area routing architecture. We call this as "handoff-exporting". Similarly, we need *e1* to "act" as *v-ABR* when computing the backup path(s) for the 2nd protected half-path.

Now suppose the first protected half path uses segment-based protection, as shown in Fig. 4.1, and the associated backup paths are *B11* and *B12* with the branch nodes *r1* and *r2*, and the merge nodes as *r3* and *e4*, respectively. *B11* and *B12* protect the normal links along the working path from *r1* to *e1*. *B12* also protects edge node *e1*. Suppose the second protected half-path uses path-based protection, then the associated protection path may be *B21* protecting the links from *e4* to *r8* and *e4*. *B01* is the nested protection path protecting optical links *e1->c1, c1->e4* and the core node *c1*. When a failure occurs, the first notified branch node will activate the shared backup path and then switch the traffic from the working path to the backup path.

4.1.2. Backup Bandwidth Sharing

Backup bandwidth sharing is an efficient way to reduce recovery resource utilization. The idea is to let backup paths share network resources when the working LSPs that they protect are physically disjoint (i.e., link, node, SRLG, etc.). There is nothing special in our framework for backup sharing in the MPLS part of a non-backbone area. Whereas for the AAPN part, as presented in the above example, for one inter-area connection request, there are in total four cross-connections involved in the AAPN domain: one for the working path, one for nested protection, and two for half-path protection. These backup optical cross-connections are setup within the AAPN to provide fast protection for inter-area connections; they all follow the time-slot constraint (since no time-slot interchanger exists at the core nodes of AAPN). Fig. 4.2 below illustrates the scenarios of backup cross-connections sharing in AAPN:

- Parallel case: 1 and 4 can share backup cross-connections, such as *e3->c2->e4* and/or *e1->c2->e6*.
- Same tail-end: 1 and 3 can share backup cross-connections, such as *e3->c3->e4* or *e2->c3->e4* (nested).
- Same head-end: 1 and 2 can share backup cross-connections, such as *e1->c3->e6* or *e1->c3->e4* (nested).
- Nested protection: 1 and 5 can share nested backup cross-connection, such as *e1->c2->e4*.

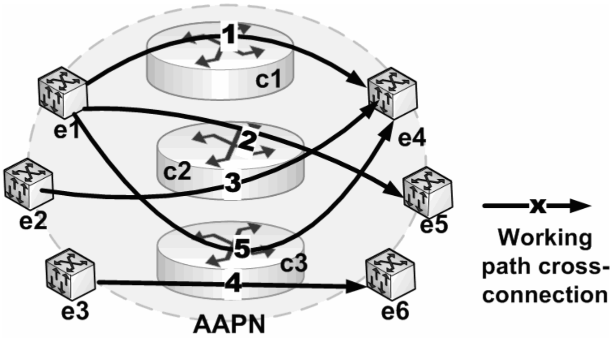


Figure 4.2: Backup Cross-connection sharing in AAPN

4.1.3. IASSP under a Weakened Single-Failure Assumption

Multi-area networks are normally large-scale networks, in which the commonly-used “single-failure” assumption becomes unrealistic. Hence we propose a weakened single-failure assumption for multi-area networks. As illustrated in Fig. 4.3, the modified single-failure assumption assumes:

- At any given time, there will be at most one failure within each circle (area) shown in Fig. 4.3.

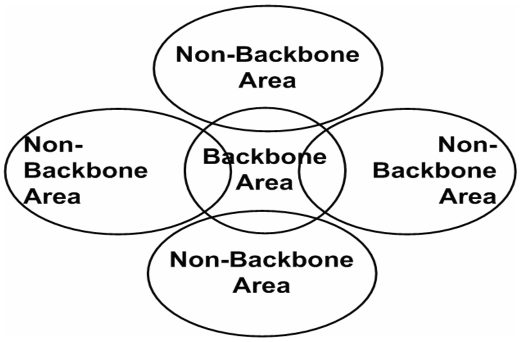


Figure 4.3: Weakened single-failure assumption

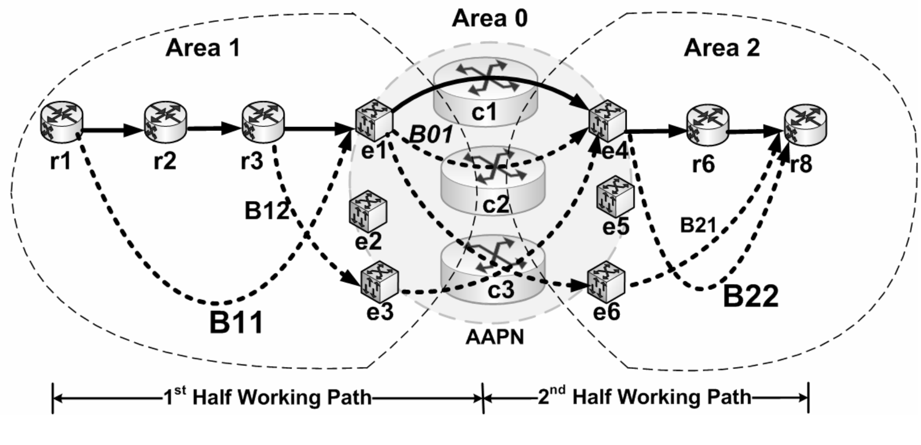


Figure 4.4: IASSP under the weakened single-failure assumption.

Under this assumption, multiple failures could happen simultaneously. Our proposed protection scheme can still work, just with two minor modifications as follows:

- For the first protected half-path, there must be one backup path that ends at the edge node along this path (see *B11* ending at e1 in Fig. 4.4).
- For the second protected half path, there must be one backup path that starts at the edge node along this path (see *B22* starting from e4 in Fig. 4.4).

It is worth mentioning that other inter-domain protection schemes in [Miyamura2004a, Huang2004, Thiongane2005] do not consider the multi-failure scenario, hence they will not work under our weakened single-failure assumption. The protection scheme in [Huang04] can work under our weakened single-failure assumption, but no backup bandwidth sharing was considered.

On the other hand, in order to make our proposed schemes work in the real world, we need to distribute and manage the routing information (see Section 4.2) so that the nodes can compute the paths according to our schemes and we also need a related signaling process (see Section 4.3) so that the nodes can successfully install the computed paths.

4.2. Routing Information Management

We use the word “routing” to indicate both the working path selection and the backup path selection. Normally, routing information management can be classified according to whether it provides complete information (e.g., global per-flow or per-link information) or partial information (e.g., part of the complete information). In multi-area networks, the former may not be practical due to the scalability issue. Hence we adopt a partial routing information management scheme described in [Qiao2001] and expand it to the case of the multi-area networks and node protection requirement, while treating the routing with complete information as the ideal case for comparison.

4.2.1. General Notations

We define the following notations:

- B_l , R_l : the total occupied backup bandwidth, and the residual free bandwidth on link l , respectively.
- d : the bandwidth requirement of an inter-area request.

- W_m^l : the total working bandwidth on/passing-through m (link or edge/core node) protected by link l $(l \neq m)$.
- B_l^m : the total backup bandwidth occupied on link l used to protect m (link or edge/core node) $(m \neq l)$.
- Data set $WSet(m) = \{\langle W_m^l, l, m\rangle \mid l \neq m\}$.
- $\overline{W}_m = \max\{W_m^l \mid W_m^l \in WSet(m)\}$.
- Data set $BSet(l) = \{\langle B_l^m, m, l\rangle \mid m \neq l\}$.

Consider the overlap part between a non-backbone area and the AAPN (see Fig. 4.1 and 4.4). We identify three kinds of links there, namely physical links (l^P), TE links (l^T), and virtual links (l^V). Physical links are individual AAPN optical links connecting edge nodes and core nodes. The TE links are bundles of the AAPN physical links exported to the MPLS non-backbone area. A virtual link is like a "tunnel" from an edge node in one area through a core node to another edge node in another area. It includes all the bandwidths occupied by the existing working and backup paths traversing it. By adopting the virtual tunnel/link concept to manage the AAPN internal routing information, we can avoid maintaining the per-timeslot backup information which is due to the timeslot continuity constraint in AAPN.

4.2.2. Routing with Complete Information (ideal case)

We adopt the least cost routing for path selection, where the cost of a path is defined as the sum of the costs of all the links along the path.

1) *Finding first the least cost inter-area working path.* The link cost function for working path computing is:

- *For a normal link l*

$$\begin{cases} 1/R_l, \ if \ d \leq R_l \\ \infty, \quad otherwise \end{cases} \tag{4.1}$$

- *For an AAPN virtual link l^V*

$$\begin{cases} \dfrac{1}{R_{l^V}} + 0.5 \times \dfrac{d \times \overline{W}_{l^V}}{M}, \ if \ d \le R_{l^V}; M = \max\limits_{\forall l^V}(\overline{W}_{l^V}) \\ \infty, \qquad otherwise \end{cases} \tag{4.2}$$

The $0.5 \times \dfrac{d \times \overline{W}_{l^V}}{M}$ part in Equ. (4.2) was inspired from the concept of potential backup cost (PBC) proposed in [Qiao2001] to make the performance of shared protection routing outperform even the ILP model. By involving PBC, we can consider the potential impact of selecting a working path on the future backup paths. This is very necessary particularly in AAPN due to its symmetric topology.

2) *Based on the determined working path, computing the least cost backup paths*. We denote AB_l^{WPi} as the additional bandwidth required on link l to protect a working path segment (or half-path) WPi. Its exact value in the backup bandwidth sharing scenario ($AB_l^{WPi} \le d$), is

$$AB_l^{WP_i} = \max_{m \in WPi} \{0, W_m^l + d - B_l\} \tag{4.3}$$

We then define the same link cost function of normal and virtual links for backup path computation as:

$$\begin{cases} 0, & if \ AB_l^{WPi} = 0 \\ \dfrac{1}{R_l}, & if \ 0 < AB_l^{WPi} \le R_l \\ \infty, & if \ AB_l^{WPi} > R_l \end{cases} \tag{4.4}$$

4.2.3. Routing with Partial Information

In this scenario, the routing information is distributed among the nodes in the network and no one maintains a global and complete view of the multi-area network.

1) *Routing procedures: similar procedures are adopted with the following changes*:

<u>For selecting an inter-area working path</u>: following our inter-area optimal routing framework proposed in Chapter 3 (see Section 3.2 and Fig. 3.6), use Equation (4.1) to compute the 1st and 2nd half working paths and then use Equation (4.2) to decide which core node to connect these two half paths.

For selecting a backup path: same link cost function as Equ. (4.4) except AB_l^{WPi} is over-estimated [Qiao2001] by

$$AB_l^{WPi} = \max_{m \in WPi}\left\{0, \bar{W}_m + d - B_l\right\} \tag{4.5}$$

2) *Link state* $\left\{R_m, \bar{W}_m, B_m\right\}$ *and data sets* $WSet(m)$, $BSet(l)$. Similar as [Qiao2001], we define $\left\{R_m, \bar{W}_m, B_m\right\}$ as the link state but a general one, since in our case *m* could be a normal link, TE link, virtual link, AAPN edge node or core node, depending on which node the link state is stored in. $WSet(m)$ is used to generate $\bar{W}_m$; while $BSet(l)$ is to adjust the actual amount of additional backup bandwidth after path determination as in [Qiao2001]. In general, the link state is updated through the OSPF-TE flooding mechanism within each area and the two data sets are updated by the RSVP-TE signaling process during call set-up.

3) *Routing information maintained at each normal node*. Similar as in [Qiao2001]: the link state for each link (normal links and TE links) in the non-backbone area, and data sets $WSet(l)$ and $BSet(l)$ for each local outgoing link of the normal node.

4) *Routing Information Maintained at each Edge Node*. On the non-AAPN side, the same as the normal node; whereas on the AAPN side, each edge node needs to maintain necessary internal AAPN routing information so as to export the link state $\left\{\boldsymbol{R}_{l^T}, \boldsymbol{B}_{l^T}, \bar{\boldsymbol{W}}_{l^T}\right\}$ of its two TE links (to/from the core) to the normal nodes in the same non-backbone area. The necessary AAPN internal routing information at each edge node includes:

- $WSet(e_i)$, where e_i is the edge node itself;
- Link state, $WSet(l^V)$ and $BSet(l^V)$ for each local outgoing virtual link;
- Copy of the link states of each local incoming virtual link.

5) *Exporting* $\left\{\boldsymbol{R}_{l^T}, \bar{\boldsymbol{W}}_{l^T}\right\}$ *through OSPF-TE Flooding*. Each edge node can derive the first two elements of the link state of its two TE links from its own AAPN internal routing information:

- Let R_{l^T} be the maximal link residual bandwidth among the physical links represented by this TE link.

- Let $\overline{W}_{l^T}$ be $\overline{W}_{e_i}$, where e_i is the edge node of this TE link. Since $\overline{W}_{l^T} \leq \overline{W}_{e_i}$, we thus can avoid exporting $\overline{W}_{e_i}$.

6) *Edge node handoff exporting* $\{\boldsymbol{B}_{l^T}\}$ *through RSVP-TE*. Observing Fig. 4.1, suppose a working path traverses the virtual link *e1->c1->e4*. When computing the associated shared backup path(s) according to Equ. (4.4, 4.5) in area 1, we notice that only the information about virtual links *e2->ck->e4*, *e3->ck->e4* ($k = 1, 2, 3$) are useful. That means the value of $\{B_{l^T}\}$ is actually working-path-dependent, and can be determined only after a working path is determined. Hence we have to use the *handoff* exporting mentioned earlier (Section 4.1.1) to export $\{B_{l^T}\}$, and export only to the source node (for the 1st protected half paths) or edge node (for the 2nd protected half path) through the transmission of RSVP-TE message (instead of OSPF-TE flooding) to compute the backup paths for the 1st or 2nd protected half paths. We approximate each B_{l^T} by the sum of B_{l^V} of all the related useful virtual links after a working is determined.

4.3. Related Signaling Process

We use a simple two-phase signaling scheme, which is fully based on the RSVP-TE protocol [RFC 3209, RFC 4873], to setup an inter-area LSP and its backup paths subsequently. As an example, we consider a request for inter-area connection with protection requirement from *r1* to *r8* in Fig. 4.1.

4.3.1. Signaling Phase I: Working Path Set-up

The signaling process in Phase I is almost the same as the one in our AAPN-based inter-area routing framework (see Section 3.2.2.3), which is a PATH↔RESV message "round-trip" for the inter-area working path set-up. The only difference is when the RESV message arrives at *e4*, through the handoff exporting, *e4* attaches related $\{B_{l^T}\}$ (to-*e4* direction) to the RESV message going back to *r1*.

4.3.2. Signaling Phase II: Backup Path build-up

Signaling phase II is another PATH↔RESV “round-trip” process. After *r1* receives the RESV message (including $\{B_{l^T}\}$) from Signaling Phase I, it computes the optimal shared backup path(s) for the 1st protected half path and then starts Phase II by sending a PATH message that includes:

- One primary ERO (Explicit_Route Object) [RFC 4873]: list of the explicit end-to-end inter-area working path.
- One or more SEROs (Secondary ERO [RFC 4873]): list of the backup path(s) for the first protected half path.

1) *PATH message processing.* The PATH message propagates along the working path until a node finds itself a branch node by checking the SEROs in the PATH message. The node then uses the related SERO and other information in the received PATH message to create a new PATH message and send it out; the original message still traverses the working path while the new one traverses along a backup LSP from this branch node to the related merge node (following the standard LSP setup procedures). When the original PATH message arrives at *e1*, *e1* exports necessary $\{B_{l^T}\}$ (from-*e1* direction) through a PATH message to *e4*. *e4* can thus compute the backup path(s) for the 2nd half protected path. After that, the same procedures as for the first protected half path are followed to set up the backup LSP(s) of the second protected half path.

2) *RESV message processing.* There are two kinds of RESV messages now: one for the working path and others for various backup paths. During the transmission of these RESV messages, the local routing information ($WSet(m)$, $BSet(l)$ and hence link state) at each passed node is updated. When the RESV message of the working path arrives at a branch node, it will not be propagated upstream until the branch node receives the RESV message of the backup LSP starting from itself. Thus, when *r1* receives the RESV message of the working path, it means that all the related backup paths are set up.

4.3.3. Complexity

As we can see, the complexity of the information updates for our protection schemes after building up an inter-area connection is in the order of the number of edge nodes (not the square of the number of edge nodes, as in [Thiongane2005]).

4.4. Performance Evaluation

We now study the performance of our segment-based shared protection framework through simulation. Simulation experiments are conducted on a 21-node 3-area ladder-like network (Fig. 4.5) (instead of 27 node network as in Section 3.4 to save simulation time), which is the extended version of the topology adopted in [Miyamura2004a]. In the simulations, the call requests arrive to the network following a Poisson process, and the call holding time is exponentially distributed. We assume that all the inter-area source-destination node pairs have the same traffic load, and also all the intra-area node pairs. A call request is accepted only when both the working path and backup paths are available.

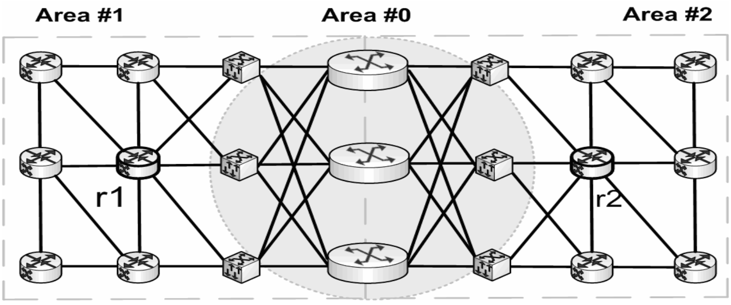


Figure 4.5: Multi-Area Network topology used for simulation (21 nodes and 96 directional links in total)

4.4.1. Blocking/Rejection Probability Analysis

There could be several IASSP (Inter-Area Shared Segment Protection) schemes, namely IASSP-CS, IASSP-PS, and IASSP-PM, where C stands for complete information, P for partial information, S for single-failure assumption and M for weakened single failure assumption. We compare our IASSP schemes with the ISDR scheme proposed in [Miyamura2004a].

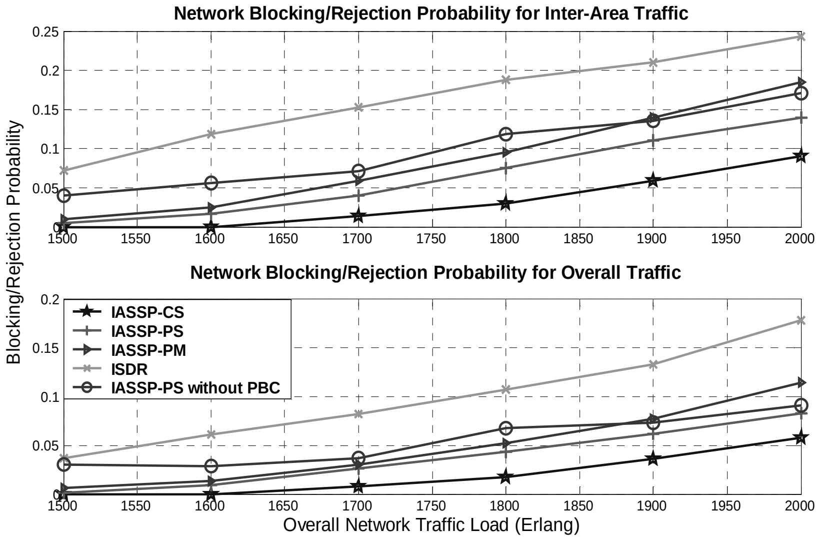


Figure 4.6: Blocking probabilities of various dynamic inter-area protection schemes with 95% confidence interval as +/- 0.1%. 60% of the overall network traffic is inter-area traffic.

As seen in Fig. 4.6, IASSP-CS has the best performance in term of blocking probability. This is reasonable since it has the complete information when doing the routing. ISDR has the worst performance, which is partially due to its non-optimal routing and partially due to less backup bandwidth sharing for inter-area routing. The IASSP-PS scheme performs closely to IASSP-CS in general, which shows the routing information management we developed works quite well. IASSP-PS outperforms IASSP-PM but not so much. This is because IASSP-PS has more flexibility when selecting backup paths. It also shows that IASSP-PM achieves multi-failure protection without great performance degrading. Fig. 8 also shows the necessity of involving PBC (see Equation (4.2)) into the link cost function (see the curve of IASSP-PS without PBC).

IASSP-PM can be considered as a special case of IASSP-PS. But it has two distinguished features, namely isolation and security. By *isolation*, we mean that it isolates an inter-area working path into three “big” segments: two MPLS segments in the head- and tail-end areas and one AAPN segment in the middle (see Fig. 4.4). Each segment can use various protection techniques, fully independently. Thus the opportunity for backup bandwidth sharing in each segment is increased. By *security*, we mean that in IASSP-PM each normal node has no information about nodes outside its own area and each AAPN edge node has no information about any normal node outside its area. These two features make IASSP-PM very attractive for inter-AS protection.

4.4.2. Protection Bandwidth Cost Ratio

We define the protection bandwidth cost ratio as the percentage of the average overall backup bandwidth to the average overall working bandwidth at a fixed network blocking probability. On the basis of the results in Table I, we notice that the schemes we proposed have lower cost ratios (e.g., better backup bandwidth sharing efficiency) than the ISDR scheme. In addition, the ratios of IASSP-PS and IASSP-PM are considerably similar to that of IASSP-CS.

TABLE 4.1 PROTECTION BANDWIDTH COST RATIO OF SCHEMES

Blocking Prob.	**IASSP-CS**	**IASSP-PS**	**IASSP-PM**	**ISDR**
1% →	59%	66%	78%	110%
10% →	55%	62%	73%	105%

4.4.3. Backup Bandwidth Sharing within AAPN

To study the maximal efficiency of backup cross-connections sharing in AAPN (see Fig. 4.2), we remove all the normal nodes in the topology of Fig. 4.5 except *r1* in Area 1 and *r2* in Area 2. IASSP-CS is chosen for the evaluation. As seen in Table II, the protection cost ratio decreases (e.g., sharing efficiency increases) as the number of edge/core nodes increases. But the speed of the decreasing becomes slow when the number of edge and core node both reach 6.

TABLE 4.2 PROTECTION BANDWIDTH COST RATIO IN AAPN

# of edge nodes per area	**# of core nodes**	**Protection Bandwidth Cost Ratio**
2	**2**	100%

3	**3**	67%
4	**5**	50%
6	**6**	40%
12	**12**	33%

4.5. Summary

We studied the resilience issue of MPLS flows over an agile all-photonic star WDM network (AAPN) in this chapter. Based on our previous inter-area optimal routing architecture, we presented a dynamic inter-area MPLS shared segment protection framework consisting of:

1. The IASSP schemes consider both single-failure and multi-failure (weakened single-failure) scenarios;

2. Distributed management of partial routing information greatly reduces the scalability issue in multi-area networks with link and key node protection;

3. A related signaling process consistent with RSVP-TE.

Meanwhile, our framework requires little change on existing traditional IP/MPLS routers to implement it. The simulation results show that our protection schemes have a performance similar to the case with complete routing information and outperform greatly the inter-area protection scheme described in [Miyamura2004a]. Indeed, together with our previous inter-area optimal routing architecture (Section 3.2), we can now provide an attractive MPLS inter-area traffic engineering solution that satisfies the requirements defined in RFC 4105. The details will be discussed in the next chapter.

Furthermore, IASSP-PM, our protection scheme under the weakened single-failure assumption, shows great potential to be a solution for inter-AS protection as discussed in Section 4.4.1. Hence we re-name IASSP-PM as IA-SSP (Inter-AS Shared Segment Protection) and choose it as the protection scheme for our AIX (AAPN-based Internet Exchange) architecture. Now we have an inter-AS traffic engineering solution for Internet Exchanges that satisfies the requirements defined in RFC 4216. The details will be discussed in the next chapter.

5. Generalization, Extension and Applications

In this chapter, we consider Overlaid-Star Networks (OSN), a generalization of AAPN, and extend our AAPN-based inter-domain traffic engineering architecture to OSN-based architectures. We show that our OSN-based traffic engineering architecture has several important applications in addition to acting as backbone in multi-area networks and as Internet exchange in multi-AS networks. It has the potential to play a key role in inter-provider, inter-customer, inter-technology, inter-domain and multi-layer traffic engineering. This is mainly due to the fact that our OSN-based architecture does not suffer from the two fundamental issues in inter-domain traffic engineering, namely information scalability and confidentiality, which have not been solved by the other existing approaches.

5.1. Overlaid-Star Networks (OSN)

5.1.1. Definition of Overlaid-Star Networks

We now define an overlaid-star network (OSN) as a network that comprises edge nodes interconnected by core nodes that function independently from each other to form an overlaid-star (also called composite-star) topology. Generally speaking, an OSN can be viewed as a distributed switch with potentially large geographical coverage. It contains three key ingredients:

- Rapidly reconfigurable switching at the core,
- Intelligent edge: control and routing functionality concentrated at the edge nodes that surround the switching core,
- Overlaid star topology for reliability and increased bandwidth.

Obviously, the AAPN is an OSN-type network. Besides, two other representative examples of overlaid-star networks are: Ethernet Star Networks with TE Capability, and PetaWeb [Vickers2000].

5.1.2. Ethernet Overlaid-Star Networks with TE Capability

It is possible to implement the overlaid-star network by Ethernet switches with TE capability. Here we define a TE capable Ethernet switch to include the following aspects:

- PBB-TE (Provider Backbone Bridging – Traffic Engineering) technology enabled, and
- Routing protocol (e.g., OSPF-TE) and signaling protocol (e.g., RSVP-TE) support.

PBB-TE or what was formerly known as Provider Backbone Transport (PBT) [Nortel2007] is a new technology concept that enables "connection-oriented" end-to-end Ethernet tunnels to be created in order to make the Ethernet an ISP class transport network. Technically, PBB-TE uses the existing Ethernet technologies of VLAN tagging (IEEE 802.1Q), Q-in-Q (IEEE 802.1ad) and MAC-in-MAC (IEEE 802.1ah), but disables flooding/broadcasting (MAC learning) and the spanning tree protocol (STP). The packets are forwarded based on VLAN ID (VID) and destination MAC address. The PBB-TE tunnels are set up either manually by a management plane or dynamically through a full or partial implementation of the GMPLS control plane [Fedyk2007]. PBB-TE is thus intended to be used in connection-oriented network applications. PBB-TE is now undergoing ratification in the standards bodies.

As shown in Fig. 5.1, several TE capable Ethernet switches are organized into an overlaid star topology: core Ethernet switches surrounded by edge Ethernet switches with 10Gbit/s Ethernet over optical fibers. There is E-O-E conversion at the core switches, in opposition to the AAPN. Core and edge Ethernet switches use RSVP-TE to set up on-demand PBB-TE tunnels (edge-core-edge).

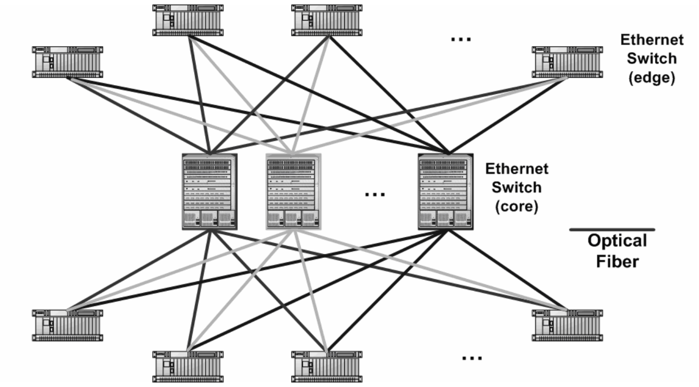


Figure 5.1: Ethernet version of OSN

It is important to compare the Ethernet overlaid-star network with an important Ethernet switch cluster architecture, called SMLT (Split Multi-Link Trunk) [Nortel2001a] (illustrated in Fig. 5.2). SMLT is a method that allows two aggregation (core) switches to appear logically as a single device to edge switches that are dual homed to the aggregation switches (Fig. 5.2). The aggregation switches are interconnected through an IST (Inter Switch Trunk) on which they exchange link state and addressing information. IST is used only on two aggregation switches. The edge switches require no knowledge of whether they are connected to a single switch or to two switches. Hence in SMLT the intelligence is concentrated only on the aggregation switches.

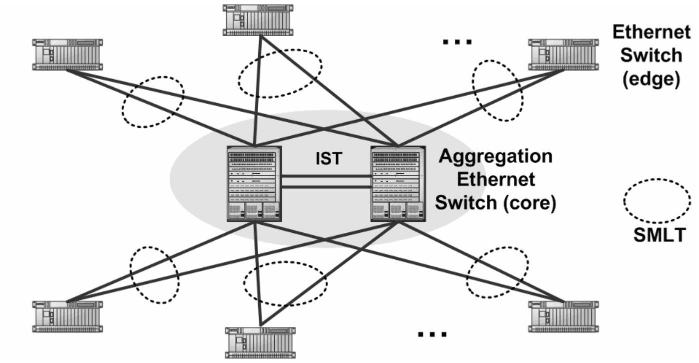


Figure 5.2: SMLT based Ethernet Star Topology (IST: Inter-Switch Trunk)

The Ethernet overlaid-star network is more scalable than the SMLT architecture. When the number of aggregation switches exceeds two in the SMLT architecture, people have to build complex switch clusters at the core. This would lead to both severe scalability issues and complex internal routing among the switches in the cluster. But this will not happen in the Ethernet overlaid-star network, where the core switches are fully independent (no IST) and all the intelligence is distributed among the edge nodes.

5.1.3. PetaWeb

The PetaWeb architecture, proposed by Nortel Networks [Vickers2000], is another example of overlaid-star networks. PetaWeb is designed to scale to a capacity of several Petabits per second, as well as to thousands of edge nodes with a global geographic coverage so as to be a candidate for the future Internet infrastructure. It is based on the use of a variety of adaptive switching cores and universal edge switches. The PetaWeb has core nodes that may operate in channel (wavelength) switching mode or burst switching mode. The edge nodes must be adapted to interact with various core nodes to support both the connectionless and connection-oriented services.

5.2. Star-TE: An Inter-domain Traffic Engineering Architecture

We extend our work on inter-domain traffic engineering (Chapter 3 and Chapter 4) from AAPN to OSN and name this OSN-based inter-domain traffic engineering *architecture* as ***Star-TE***.

Star-TE is a novel inter-domain MPLS traffic engineering architecture on the basis of **(1) deploying an OSN as the physical inter-connecting facility and (2) applying the OSN with the traffic engineering framework which we initially designed for inter-area traffic engineering (Chapter 3 and 4)**.

The traffic engineering framework is the main part of Star-TE, which, in general, comprises three main components, namely the routing-information, path computation and signaling components. As we described in Section 3.2.2, the basic idea of this TE framework is to configure the OSN (e.g., AAPN in Section 3.2.2) in such a way that it is seen by connected local networks (domains) to be a star with a (virtual) node (e.g., ABR or ASBR) at the place of the core. Such a virtual node does not exist at the OSN core node, but the edge nodes may project such a vision to the other nodes in the connected local networks. Each connected local network can extend its TE visibility up to the virtual core node with almost no impact on its TE information scalability. This configuration presents the advantage that optimal end-to-end routes can be easily established by simply concatenating optimal routes to/from the core, which can be determined by the source and destination connected local networks independently of one another. This problem of finding optimal end-to-end routes can in general only be solved by considering global knowledge; in our architecture with a virtual core node, no global knowledge is required, only the local routing information within each connected local network.

To establish protection paths for data flows that require high reliability, instead of using link or path protection, Star-TE adopts the protection approach (in Chapter 4) using shared segment protection to take advantage of its multi-domain routing framework with the virtual core node.

Star-TE is a general inter-domain traffic engineering architecture, which imposes no constraints on the mechanism of organizing/maintaining the TE information in each domain (e.g., either in a centralized fashion or in a distributed fashion), no constraints on the techniques for computing paths in each domain (PCE based or any others), and no constraints on the mechanism of setting-up the inter-domain path (e.g., automatically by signaling protocol, RSVP-TE, LDP-TE, or by the management plane). It is worthy mentioning that Star-TE can be directly deployed in GMPLS networks. This is because GMPLS [RFC 3945] is based on TE extensions to MPLS (e.g., OSPF-TE/IS-IS-TE, RSVP-TE/LDP-TE, LMP). GMPLS separates the control and data planes, but this has no impact on deploying the Star-TE in GMPLS-enabled networks.

We now discuss the global optimality vs. per-domain criteria. In the case of inter-area TE, the same optimization TE criteria is usually adopted for all the OSPF areas (e.g., based on the number of path hops, bandwidth consumption, etc.) [Vasseur2007b]. In contrast, in the case of Inter-AS TE, there might be scenarios in which different ASes chose different criteria to determine/define their own TE optimality. In such a situation, in order to have a meaningful path computation, it may be necessary to perform criterion normalization (or equivalence-mapping) between the ASes. For instance, as indicated in [Vasseur2007b], the Service Providers need to agree on a common normalized TE optimization criterion and use this criterion for “global” optimal path computation. Note that Star-TE is suitable to any of the above cases. This is due to the fact that in Star-TE optimal end-to-end routes can be easily established by simply concatenating optimal routes to/from the core, which can be determined by the source and destination areas/ASes independently of one another. Hence the optimization criterion, e.g., the “cost” (in normalized cost unit), could be calculated differently within source and destination areas/ASes.

5.3. Applications of Star-TE

5.3.1. MPLS Inter-Area Traffic Engineering

RFC 4105 defines the requirements for Inter-Area MPLS Traffic Engineering. In Table 5.1, we check if Star-TE satisfies RFC 4105 when deployed in a multi-area network scenario.

TABLE 5.1: RFC 4105 CHECKLIST FOR STAR-TE

Requirements in RFC 4105	**Satisfied**
Inter-Area MPLS TE Operations and Interoperability (interoperate seamlessly with current intra-area MPLS TE mechanisms).	**Yes**
Inter-Area TE-LSP Signaling: The solution MUST allow for the signaling of inter-area TE LSPs, using RSVP-TE.	**Yes**
Path Optimality	**Yes**
Inter-Area MPLS-TE Routing: avoiding avoid any dynamic-TE-topology-related information from leaking across areas, even in a summarized form.	**Yes**
Inter-Area MPLS-TE Path Computation: the solution should support more than one path computation method; it should allow the operator to select by configuration, and on a per-LSP basis, the desired option.	**Yes**
Support inter-area (signaling)crankback routing	**Yes**
Support of Diversely-Routed Inter-Area TE LSPs	**Yes**
Intra/Inter-Area Path Selection Policy: the solution should allow IGP area crossing to be enabled/disabled, on a per-LSP basis, for TE LSPs whose head-end and tail-end reside in the same IGP area.	**Yes**

Re-optimization of Inter-Area TE LSP	**Yes**
Rerouting of Inter-Area TE LSPs	**Yes**
Fast Recovery of Inter-Area TE LSP	**Yes**
Hierarchical LSP Support	**Yes**
Hard/Soft Preemption	**Yes**
Backward Compatibility	**Yes**
Complexity and Risks: The proposed solution SHOULD not introduce complexity to the current operating network to such a degree that it would affect the stability and diminish the benefits of deploying such a solution over service provider networks.	**Yes**
Capability to share bandwidth among inter-area backup LSPs protecting independent facilities.	**Yes**

From the above Table 5.1, we conclude that Star-TE satisfies RFC 4105 when deployed in a multi-area network scenario.

5.3.2. Internet Exchange: MPLS Inter-AS Traffic Engineering

RFC 4216 defines the requirements for Inter-AS MPLS Traffic Engineering. In Table 5.2, we check if Star-TE satisfies RFC 4216 when deployed in a multi-AS network scenario.

TABLE 5.2: RFC 4216 CHECKLIST OF STAR-TE

Requirements in RFC 4216	**Satisfied**
The proposed solution SHOULD allow the provisioning of a TE LSP at the Head/Tail-end with end-to-end Resource Reservation Protocol (RSVP) signaling (eventually with loose paths) traversing across the interconnected ASBRs, without further provisioning required	**Yes**

along the transit path.	
The solution SHOULD allow the set-up of an inter-AS TE LSP that complies with a set of TE constraints and follows an optimal path.	**Yes**
Support of Diversely Routed Inter-AS TE LSP	**Yes**
Support of Re-Optimization	**Yes**
Fast Recovery Support Using MPLS TE Fast Reroute	**Yes**
Scalability and Hierarchical LSP Support	**Yes**
Complexity and Risks: The proposed solution(s) SHOULD NOT introduce unnecessary complexity to the current operating network to such a degree that it would affect the stability and diminish the benefits of deploying such a solution over service provider networks.	**Yes**
Backward Compatibility	**Yes**

From the above Table 5.2, we conclude that Star-TE satisfies RFC 4216 when deployed in a multi-AS network scenario.

5.3.3. Segmented PCE Architecture

The PCE architecture has been proposed by IETF [RFC 4655] to compute MPLS/GMPLS inter-domain (area or AS) paths. Recently, IETF suggested two inter-domain path computation techniques for PCE: the per-domain method [Vasseur2007a] and the Backward Recursive Path Computation (BRPC) method [Vasseur2007b]. As discussed previously and proved by simulation in Section 3.4, the per-domain path computation technique is suboptimal. In addition, it is quite challenging to compute a set of diverse inter-domain paths by the per-domain technique. The BRPC path computation technique can compute optimal inter-domain TE LSP if the domain sequence that the path will traverse is given. Hence in the rest of this section, we focus on the PCE Architecture with the BPRC technique.

BPRC relies on the collaboration between PCEs through the PCEP protocol [Vasseur2007c]. As illustrated in Fig. 5.3, the path computation is performed by each PCE (along the domain sequence) computing a Virtual Shortest Path Tree (VSPT) and passing the computed VSPT in a backward recursive fashion from the destination to the source domain. The root of each VSPT is always the destination while the link in each VSPT representing the shortest path between the border nodes of a domain and the destination LSR. The VSPT is *re-computed* when another domain is passed on the basis of 1) the optimal paths from each entry point of the domain to each exit point and 2) the VSPT of the downstream domain [Vasseur2007b]. After a round trip (Query->Respond) path computation process, shown in Fig. 5.3, the computed TE LSP is signaled using the standard REVP-TE protocol, which is another round-trip (Path->Resv) signaling process.

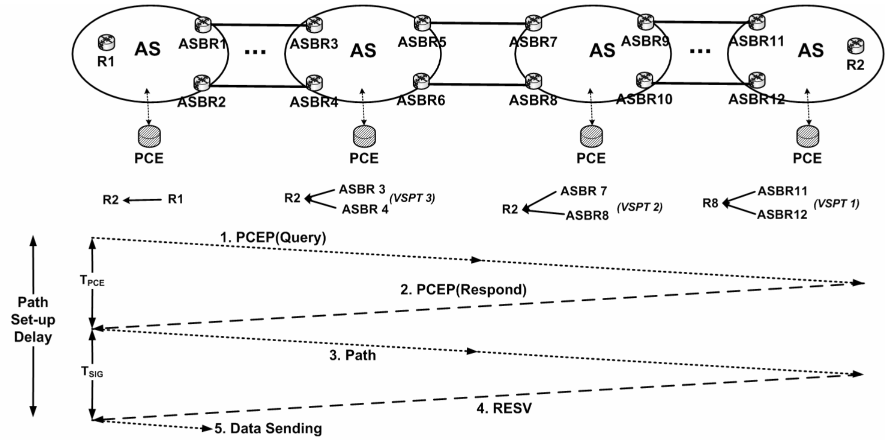


Figure 5.3: BRPC inter-domain path computation technique proposed by IETF

5.3.3.1. Potential Drawbacks of the PCE Architecture with BRPC Technique

When the *number of domains* in the given domain sequence and/or the *number of border nodes* in each domain become large, the PCE architecture with BRPC technique has the following inherent drawbacks:

1) *High load and complexity of path computation.* A long domain sequence will involve many PCEs in the path computation processes. Many border nodes will cause the exchange of a large amount of information between PCEs. For instance, suppose each domain in Fig. 5.3 has E border nodes on each side, then the information exchanged

between PCEs has the size of $O(E)$ for single path routing; and in size of $O\left(\frac{E!}{m!(E-m)!}\right)$ for m end-to-end diverse paths routing.

2) *Long path set-up delay*. We define the path set-up delay as the duration starting from a call request arriving at a source LSR and ending when this LSR begins to send data. Suppose the number of domains that a path has to traverse is N (including the source and destination domains). As shown in Fig. 5.3, the path set-up delay, T_{set-up}^{BRPC}, of the BRPC technique can be computed as follows:

$$T_{set-up}^{BRPC} = T_{PCE} + T_{SIG} = N \times t_{PCE} + N \times t_{SIG} \tag{5.1}$$

where T_{PCE} is the round-trip PCE path computation delay defined as the duration starting when the source node of the call sends a Query message to a PCE until the source node receives the Respond message from the PCE. T_{SIG} is the round-trip RSVP-TE signaling delay defined as the duration starting when the source node of the call sends a Path message until the source node receives a Resv message. t_{PCE} and t_{SIG} are round-trip PCE path computation delay and round-trip RSVP-TE signaling delay within a given domain, respectively. We assume that the PCE path computation delay is the same in all domains, same assumption to the RSVP-TE signaling delay.

As shown in [6], if the path set-up delay is comparable with (or greater than) the average network-wide call inter-arrival time, $T_{inter-arrival}$, the call blocking probability will dramatically increase. But usually $T_{set-up} < T_{inter-arrival}$. However, note that $T_{inter-arrival} \propto \frac{1}{N}$ and $T_{set-up} \propto N$. As N increases, sooner or later, it will reach a value, at which $T_{set-up} \geq T_{inter-arrival}$. This value of N, denoted as N_{limit}, should be considered as the maximal number of domains (in a linear direction) that the PCE architecture with BPRC technique can support, which is an important parameter for provisioning PCE in practice.

3) *Robustness and flexibility issue*. BRPC path computing technique requires that PCE is implemented in each domain along the inter-domain path. If one PCE fails during the path computation processes, all the ongoing computation processes in this PCE will have to be repeated from the beginning. Meanwhile, in practice, some operators may choose not to implement PCEs within their network that have to cooperate with other PCEs outside their network.

All the above drawbacks show that the PCE architecture with BPRC technique suffers from scalability and robustness issues when deployed in large networks. As far as we know, no solution to this problem has been found. To solve this problem, we propose a *segmented PCE architecture*.

5.3.3.2. Segmented PCE Architecture for Large-scale Networks

We call a PCE architecture that deploys one or several Star-TE (e.g., as backbone area or Internet Exchange) a Segmented PCE architecture. As shown in Fig. 5.4, a long domain sequence is divided into two segments by putting a Star-TE in the middle. We propose two options for path computation and signaling in the segmented PCE architecture. *Option one* is to reduce the round-trip PCE computation delay. Shown in Fig. 5.4, the optimal paths in Segment 1 and Segment 2 can be computed nearly “in parallel” without impact on the global optimality. The BRPC path computation technique can be applied to the two segments independently. Now consider a general case of a sequence of $N\ (N > M)$ domains which is divided into $M+1$ equal segments by M Star-TE. We have:

$$T_{set-up}^{Method\ One} = T_{PCE}^{Method\ One} + T_{Signaling} = \frac{N}{M} \times t_{PCE} + N \times t_{Signaling} \tag{5.2}$$

Option two is to further reduce the round-trip signaling delay. When the PATH message arrives at an edge node of the OSN, the edge node will return a RESV message to the source node (as shown in Fig. 5.4). When the source node receive this RESV message it knows that first segment has been setup. After waiting some additional time Δt (Δt could be zero), the source node will start to send data. The purpose of Δt is to make sure that the 2nd segment is ready when the data arrives at the edge node. Still consider the general case, the overall path set-up delay of Option Two is

$$T_{set-up}^{Method\ Two} = T_{PCE}^{Method\ One} + T_{Signaling}^{Method\ Two} == \frac{N}{M+1} \times \left(t_{PCE} + t_{Signaling}\right) + \Delta t = \frac{T_{set-up}^{BRPC}}{M+1} + \Delta t \tag{5.3}$$

As M increases, $T_{set-up}^{Method\ Two}$ decreases, hence the call blocking due to long set-up delays can be reduced. Please note that the value of Δt is actually determined by the maximal size of the segments involved; it also must take into account the speed of the signaling through the different segments which may depend on the load of the networks. Hence it should be very careful to select the value of Δt .

Table 3 summarizes the properties of the original and segmented PCE architectures concerning the inter-domain path set-up delay and the diverse path computation. It shows the advantages of deploying the Star-TE within the original PCE architecture. As mentioned in Section 5.3.3.1, the inter-domain path set-up delay of the original PCE architecture is the sum of two round-trip delays, namely the PCE path computation delay and the RSVP signaling delay. Our segmented PCE architecture with Option One reduces the path computation delay to $\frac{1}{M+1}$ of the original value (Equ. 5.2). In addition, Option Two further reduces the RSVP signaling delay into $\frac{1}{M+1}$ of the original value plus Δt (Equ. 5.3). Hence for a multiple Star-TE case, the more Star-TEs are included in the path, the more the delay is reduced.

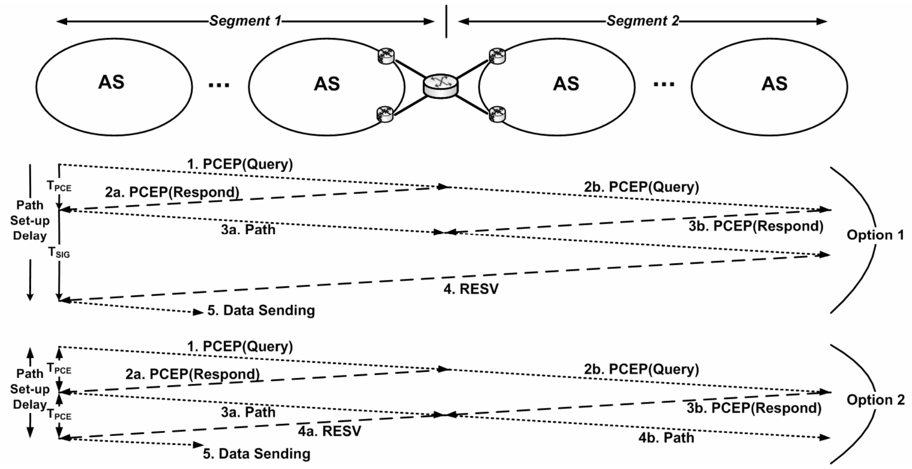


Figure 5.4: Segmented PCE-based architecture

Table 5.3: Comparison of Original PCE Architecture and Segmented PCE Architecture

Several PCE Architectures	**Inter-Domain Path Set-up Delay**	**Guarantee of finding out *m* end-to-end divers paths**
Normal IETF PCE Architecture: Per-Domain	$N \times t_{PCE} + N \times t_{SIG}$	No
Normal IETF PCE Architecture: BRPC	$N \times t_{PCE} + N \times t_{SIG}$	Yes with $O\left(\frac{E!}{m!(E-m)!}\right)$ information exchanged between domains

Segmented PCE Architecture: per-segment with Option One	$\frac{N}{M+1}\times t_{PCE}+N\times t_{SIG}$	Yes with $O(m)$ information exchanged between segments
Segmented PCE Architecture: per-segment with Option Two	$\frac{N}{M+1}\times\left(t_{PCE}+t_{SIG}\right)+\Delta t$	Yes with $O(m)$ information exchanged between segments

Please note that the advantages of segmented PCE architecture are not limited to a linear topology. Actually, in any general network topology, a source-to-destination path, at the domain level will always be a linear domain sequence, to which our approach can be applied. Also note that due to the star topology, a Star-TE can actually segment any domain sequences passing through the OSN. However, it is not clear how to choose one or a few "strategic" locations in a large-scale meshed multi-domain network to deploy Star-TE, as shown in Fig. 5.4 for an example. This problem can be solved offline and is one of our future work.

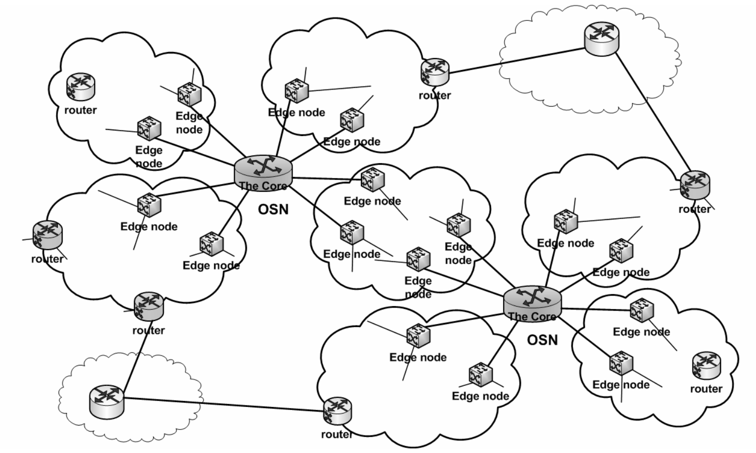


Figure 5.5: An example of deploying two Star-TE in a meshed multi-domain networks.

5.3.4. Global Concurrent Optimization (GCO)

Global Concurrent Optimization (GCO), defined by IETF [Lee2007], is one of the key applications of inter-domain traffic engineering, which computes a set of TE paths concurrently. A GCO path computation should simultaneously consider the entire topology of the network and the complete set of existing LSPs so as to optimize or re-optimize the resource utilization on the basis of the whole network [Lee2007]. The need for a global concurrent path computation usually arises in the following situations: (1) re-optimize the existing networks; this is when the global network resources become fragmented after LSPs joining in and leaving over time, and the network might no longer provide the optimal use of the available capacity. (2) Re-route a set of TE LSPs in the event of catastrophic network failures.

The nature of GCO needs online processing. However, online global concurrent optimization within the current approaches, e.g., PCE architecture with BPRC, does not scale well and the major bottleneck is not just the path computation itself but the bulk of data exchanged, synchronization issues, failures during re-optimization, and so on. In a large-scale network, GCO would likely affect the network stability and significantly diminish the benefits of deploying PCEs.

The Star-TE or segmented PCE architecture can solve this problem. It avoids the exchange and synchronization of a substantially large amount of data by dividing a global concurrent optimization problem into several relatively small and independent sub-GCO problems, one per segment with just limited cooperation and interactions among these sub-processes.

5.3.5. CE-PE Connecting Facility

Today, customers expect triple play services through BGP/MPLS IP-VPNs (Virtual Private Network) [RFC 4364]. And their requirements for end-to-end QoS and session management of applications are increasing. As illustrated in Fig. 5.6, customers want the service provider to provide a service that guarantees a bandwidth from a local CE (Customer Edge Equipment) to a remote CE through the network (C-TE LSP, see Fig. 5.6). Furthermore, they also want an end-to-end host-to-host service with bandwidth guaranteed (see for example Fig. 5.6, from a host *H1* in a local customer site to another host *H2* in remote customer site but within the same VPN).

A recently published Internet draft [Kumaki2007] lists the detailed requirements from the viewpoint of customers. The main items are the end-to-end resource optimization and the scalability considerations, that is, the end-to-end (host-to-host) optimal routing while keeping the information scalability and confidentiality (e.g., customer networks should not share internal information with the service provider).

A PCE based approach to the above problem has been proposed [Vasseur2007d]. The basic idea is to extend the MP-BGP [RFC 2858] protocol so as to convey TE characteristics of the PE-CE links in order to extend the visibility of the Traffic Engineering Database to those links. Using the BPRC technique [Vasseur2007b], the PCEs located in service provider network can then compute optimal C-TE LSPs (see Fig 5.6) which may require specific services such as bandwidth guarantees and fast path protection in a MPLS VPN environment.

However, although the approach in [Vasseur2007d] can compute optimal C-TE LSPs through PCEs, it has two drawbacks: one is the potential scalability issue of the TE database when the service provider supports a large number of VPN customers. Another one is that this approach can not compute the customer-required host-to-(remote)host optimal path (end-to-end) since the service provider's TE database can never include the internal TE information of customer networks.

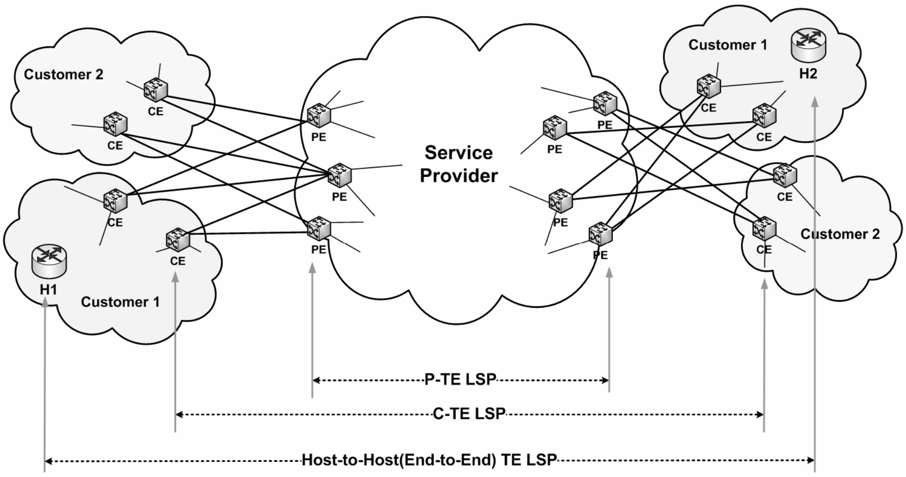


Figure 5.6: CE-to-CE Reference Model [Kumaki2007]. There are two VPNs in the above reference model, Customer 1 VPN and Customer 2 VPN. For reliability reasons, each CE needs to connect at least two PEs. (PE: Provider Edge Equipment; C-TE LSP: Customer Traffic Engineering Label Switched Path. P-TE LSP: Provider Traffic Engineering Label Switched Path.)

If the service provider uses Star-TE as the inter-connecting facility between PEs and CEs, as shown in Fig. 5.7, then both the optimal customer-required host-to-(remote)host paths and the optimal C-TE LSPs can be computed, just like the case of Star-TE in the multi-area or multi-AS scenario that is discussed in Chapter 3. Instead of v-ABR or v-ASBR, the core in the OSN is exported as a single virtual CE (v-CE) to each connected customer network and one virtual PE (v-PE) to the service provider network.

There will be no scalability issue in the above Star-TE based approach. As seen in Fig. 5.7, each customer network only needs to maintain the TE information from its own network up to the core of the connected OSN, so does the service provider network. Furthermore, the TE information database of the service provider network is almost *independent* to the number of connected customer networks. Only reachability information is exchanged between the customer and service provider networks. Hence the information confidentiality of the Star-TE approach is also good. Meanwhile, the reliability of Star-TE is guaranteed by its overlaid-star topology.

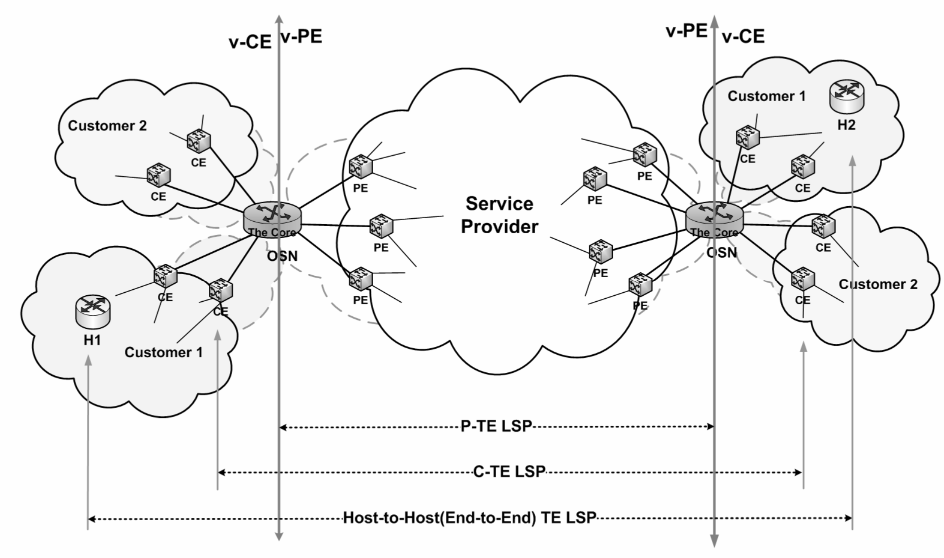


Figure 5.7: Star-TE based CE-PE inter-connecting facility with TE capability.

5.3.6. Universal and Optimal Network TE Inter-Connector

To build up a network, especially a large network, carriers usually use IP/MPLS technology; or furthermore, like a few operators in Japan, experienced with the GMPLS technology to construct their networks. Nowadays, carriers have two extra technology choices beyond the MPLS/GMPLS: one is PBT (or called PBB-TE, IEEE 802.1Qay) standardized by IEEE. (see Section 5.1.2. or [Nortel2007]) Another one is T-MPLS (Transport MPLS) [TPACK2007] standardized by ITU-T. T-MPLS can be considered as a new formulation of MPLS, designed specifically for application in packet transport networks. It builds upon IP/MPLS technology and standards, but offers a simpler implementation, where features not relevant to connection-oriented applications are removed. In addition, in order to lower the carriers' operational expenses, T-MPLS has the some enhancements to the original MPLS, e.g., engineered point-to-point bi-directional LSPs, end-to-end LSP protection together with advanced OAM support for optimal control of transport network resources. T-MPLS is formulated in conjunction with today's circuit-based transport networks (e.g., SONET/SDH), following the same architectural, management and operational models. It is thus intended to provide an optimum evolution path for many carriers to a packet-based future [TPACK2007].

When any new technology emerges, it is impossible to eliminate the "old" technologies over night, especially if the "old" technologies are still working well. Actually, the biggest value of PBT or T-MPLS is that they give carriers more choices. Before, carriers had almost only one choice, namely IP/MPLS. Both PBT and T-MPLS are under standardization now and they still have a long way to go to become as strong/mature as IP/MPLS. For instance, what about the control plane of PBT and T-MPLS to provide dynamic TE capability? Therefore, for a relative long time in the future, all these technologies, IP/MPLS, GMPLS, PBT, T-MPLS, will *co-exist* in networks. Different carriers may choose different technologies. But how to inter-connect the various-technologies-based networks "smoothly" and "intelligently" will be a problem that both the carriers and venders have to seriously deal with. By smoothly, we mean that a network based on one technology should not feel the difference when communicating with other networks that are implemented in different technologies. By intelligently, we mean that global optimal resource utilization can still be achieved among the various-technologies-based networks, that is, *inter-technology traffic engineering*.

Star-TE can be a candidate solution to the inter-technology traffic engineering. As shown in Fig. 5.8, we use an OSN to inter-connect four networks, which are implemented through various technologies. With Star-TE, the core in the OSN can be exported as a virtual Ethernet switch to the PBT based network, as a virtual T-MPLS node to T-MPLS based network, as a virtual MPLS router to MPLS based network, and as a virtual GMPLS node to GMPLS based network. Thus the “smoothly” problem can be solved. Within each network, traffic engineering can be done up to the virtual switch/node/router. Then similar to the applications of Star-TE in inter-area or inter-AS traffic engineering, two local optimization (one in each local network) can be merged by the OSN into a global inter-network optimization. Hence the “intelligently” problem is also solved. The detailed and related protocol design, extension or “mapping” among various technologies is still for further study.

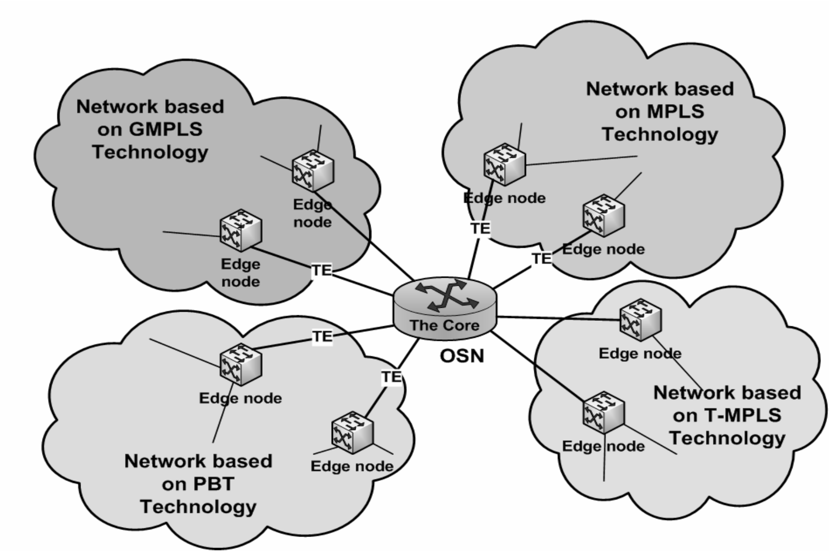

Network based on GMPLS Technology
Edge node
Edge node
TE
Network based on MPLS Technology
Edge node
TE
TE
Edge node
The Core
OSN
TE
TE
Edge node
Edge node
Network based on PBT Technology
Edge node
Network based on T-MPLS Technology
Edge node

Figure 5.8: Adopting a Star-TE as a universal inter-connecting facility among networks with various technologies

5.4. Beyond OSN: Flat-Star Networks

It is also worthy mentioning that although the Star-TE was originally designed for the OSN architecture, it can also be applied over other architecture, e.g., a router-cluster in the core surrounded by edge routers (see Fig. 5.9). We call this kind of architectures Flat-Star Networks (FSN).

Star-TE can be applied to a FSN when the FSN is used to inter-connect domains (areas or ASes), e.g., multi-area or multi-AS (internet Exchange) scenarios. Contrary to the OSN case, the edge routers have no idea of the router-cluster, they just think that they are connected to a single router. When applying Star-TE, the edge router can adopt the "bundle" concept to export all the links connected to/from the router-cluster as a single TE link. The router-cluster then acts as a single (virtual) ABR (Area Border Router) or a (virtual) ASBR (AS Border Router) to the edge routers and other the routers within each domain (area or AS). Here the "acting" means transferring and exchanging sufficient routing and signaling information with the domain-internal routers. Each virtual ABR or virtual ASBR could be implemented as an instance in the router-cluster. The inter-domain reachability information is exchanged among these instances within the router-cluster. Hence the control software in the router cluster might be complex.

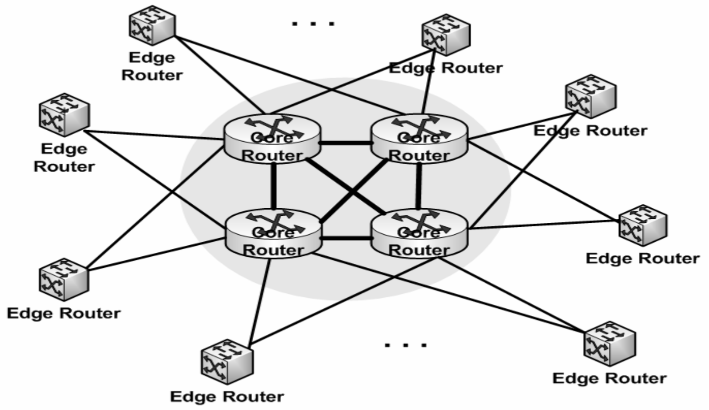


Figure 5.9: Sample of Flat-Star Network (FSN) (Note that the shadow area is the router cluster as the core). Note that the IX (Internet Exchange) architectures of Fig. 3.8 – 3.9 are a special case of this architecture.

We compare the scenarios of applying Star-TE to OSN and FSN in Table 5.4. The biggest difference of the two cases is the location of intelligence, either distributed among the edge nodes (e.g., OSN) or concentrating at the core (e.g., FSN). Apparently from Table 5.4, we can see that OSN is better than FSN in aspects of scalability, reliability, and confidentiality. However, if the inter-domain traffic is not huge, Star-TE applied FSN is still attractive since it can use existing physical devices (router/switch clusters) and existing technologies (e.g., SMLT [Nortel2001a], RSMLT [Nortel2001b], VRRP [RFC 3768], etc.).

TABLE 5.4: COMPARISON OF APPLYING STAR-TE TO OSN AND FSN

	Star-TE over OSN	**Star-TE over FSN**
Inter-domain optimal routing	Yes	Yes
Intelligence (Routing and signaling, acting as ABR /ASBR)	Distributed among edge nodes and implemented by edge nodes	Concentrated and implemented at the core
Reliability	Good	Depend on the architecture of the

		router cluster
Scalability	Good	Not so Good (due to complex internal routing within the cluster)
Confidentiality	Good	Depends on the cluster control software

5.5. Summary

5.5.1. What is a Good Inter-Domain Traffic Engineering Solution?

Based on RFC 4105 and RFC 4216, a “good” inter-domain traffic engineering solution can be summarized as follows:

- First, it should be an automatic process running in the control plane of the network (e.g., (G)MPLS, or some others), computing inter-domain paths in real-time, and dynamically adapting to changing traffic and link availability conditions.
- Second, it should be a network-wide distributed optimization process that is distributed over the network domains.
- Third, each domain should be responsible for optimizing its own routing, and only a limited cooperation between domains should be needed without disclosing the internal information of the involved domains.

However, till now, as we mentioned in Section 2.3.2.2, despite some initial theoretical trials [Winnick2002, Shrimali2007, Tomaszewski2007] and much practical work (the most promising one is the IETF’s PCE architecture [RFC 4655] but it requires a giving domain sequence first, and it has potential scalability problems), an effective, scalable and reliable distributed optimization that satisfies the above three criteria for inter-domain traffic engineering has yet not been found. It is still not clear to what extent the TE information should be exchanged and synchronized among cooperative domains for obtaining the global optimality without breaking the ***scalability*** and ***confidentiality*** constraints [Tomaszewski2007]. In addition, the issue of long path set-up delay in large-scale networks has to be considered in inter-domain traffic engineering. As discussed in Section 5.3.3.1, path set-up delays may lead to high blocking probability if the delay is long enough.

5.5.2. Star-TE Could be a Good Inter-Domain Traffic Engineering Solution

We believe that Star-TE has the potential to be a ***good*** inter-domain traffic engineering solution basing on the following facts:

- With Star-TE, the inter-domain global optimization process is distributed over a set of collaborating network domains (physically inter-connected by the OSN/FSN of Star-TE). Each domain is responsible for optimizing routing in its own domain locally up to the virtual node in Star-TE. Hence the *scalability* of Star-TE is good.
- Only a limited cooperation (e.g., exchanging reachability information) among domains is needed. No confidential TE information will leak from one domain into the others. Hence the information *confidentiality* of Star-TE is good.
- Star-TE automatically merges the domain-wide local optimizations into a global optimization. Hence the inter-domain *optimality* is guaranteed by Star-TE.
- The local optimization process in each domain runs independently (to each other) and near-simultaneously. As the local optimization is inherently a part of the global optimization, there would be a possibility that the signaling process can start even before the whole inter-domain path computation process is finished. In other words, the signaling process can start “earlier”. Thus the whole path set-up delay (defined in Section 5.3.3.1) can be reduced. Note that only the Star-TE has this distinctive property. None of other existing inter-domain TE approaches has it.

- Star-TE is a very flexible inter-domain traffic engineering architecture. It has no limits on the local path computation mechanism (e.g., either PCE or non-PCE), routing information distribution and signaling processes (e.g., centralized or distributed, in control plane or in management plane) in each domain.
- Especially, as we discussed in Section 5.3.3.3, when combining Star-TE with the PCE architectures proposed by IETF, we obtained a segmented PCE architecture that can be a good TE solution for meshed large-scale multi-domain networks. The segmented PCE architecture does not have the scalability and robustness issues inherent in the original PCE architecture. In addition, Star-TE is only needed to be deployed in one or several “strategic” points (as shown in Fig. 5.5) of the networks.

5.5.3. The Value of Simulation Results for Inter-Domain Single-Path Routing (in Fig. 3.13) and Diverse-Path Routing (in Fig. 3.17)

In this chapter, we extend our inter-domain traffic engineering work in Chapter 3 and Chapter 4 from the AAPN to a generalized form, OSN. In this context, we should ***emphasize*** the value of the simulation results (the “X”-like curves) in Fig.3.13 (single-path routing) and Fig. 3.17 (diverse-path routing) in the context of OSN. All the results in Fig. 3.13 and Fig. 3.17 still hold true in the context of OSN. Meanwhile, these results have two significant and important meanings to inter-domain traffic engineering, as explained in the following.

5.5.3.1. The Effective Range

It is normally considered that the outgoing-view-extension and incoming-view-extension techniques are beneficial improvements to the commonly-used per-domain path computation technique. IETF has taken the above two techniques as two main mechanisms in inter-domain traffic engineering for distributing TE information from outside the domain (area or AS) to the inside. The related standardization work (e.g., protocol extensions) in IETF is ongoing [Chen2007, Otani2007]. But it is unknown yet under what condition(s) these two extension techniques work well or when one performs better than another.

The results in Fig.3.13 and Fig. 3.17 give the answer to the above question. The two extension techniques have different effective ranges (for single-path and diverse routing). Outside their effective ranges, they are the same to the original per-domain path computation mechanism. No similar results have been found yet.

5.5.3.2. NBMA Media among Domains

IETF only considers the scenario that multiple domains are inter-connected through direct physical inter-domain links now. If the inter-domain connecting facility is not individual domain-to-domain physical links, but the NBMA (Non-Broadcasting Multi-Access) media, e.g., Ethernet, ATM, or AAPN, then what is the mechanism of TE information distribution between the domains (area or AS) to inside? Again, no answer to this question has been found yet.

Based on the results in Fig.3. 13 and Fig. 3.17, we believe that the “virtual core node” concept in Star-TE can be used to answer the above question: distributing both the outgoing and incoming TE information of the NBMA facility into the connected domains but up to the virtual core node. We work on related protocol extensions work in order to publish another Internet Draft in the IETF.

6. Conclusions

6.1. Summary

It is expected that an agile all-photonic network (AAPN) would be used in a metropolitan context or as a wide-area network for interconnecting many local Internet networks containing many users and servers, and probably also as an Internet backbone. Hence, it is very important to design and position AAPN to support existing widely-deployed IP/MPLS architecture and protocols. This thesis focuses on the routing and protection of MPLS flows over the agile all-photonic networks (AAPN).

Since an AAPN with *N* edge nodes provides *N* x *N* logical links among the edge nodes, the straightforward use of the OSPF routing protocol used for IP and MPLS leads to scalability problems. This thesis has therefore proposed that the AAPN configuration seen by the Internet routers should be a star with a (virtual) router at the place of the core. Such a router does not exist at the AAPN core node, but the routers at the edge nodes may project such a vision to the other routers in the connected local Internets. This routing architecture may be implemented by the routers associated with the edge nodes in such a way that they present this virtual router to the other routers in their local environment, and communicate with the routers associated with the other edge nodes by a specially adapted routing protocol that takes into account the bandwidth allocation in the AAPN and other traffic engineering parameters.

In the case that OSPF is used within the whole AAPN environment, the multi-area option of OSPF may be adapted in such a manner that the backbone area (Area 0) collapses into the virtual core router and the other independent (non-backbone) areas extend up to the virtual core router (virtual Area Border Router). This configuration present the advantage that optimal end-to-end routes can be easily established by simply concatenating optimal routes to/from the core, which can be determined by the source/destination non-backbone area independently of one another. The problem of finding optimal end-to-end inter-area routes can in general only solved by considering global knowledge; in our routing architecture with a virtual core router, no global knowledge is required, only the local routing information within each non-backbone area is needed. Simulation studies have confirmed that our routing architecture performs very well, close to the ideal case (knowing the global knowledge), and outperforms other practical inter-domain TE approaches, both in inter-area single-path routing and diverse routing.

This virtual routing IP/MPLS architecture that we proposed is also useful for establishing protection paths for data flows that require high reliability. Instead of using link or path protection, this thesis points out that a protection approach using shared segment protection can take advantage of the multi-area routing architecture with a virtual core router. As in the case of optimal inter-area path selection, the optimization issues for the protection path can be handled independently in the source and destination areas. Simulation studies have been done to evaluate the efficiency of this protection path allocation scheme. They confirmed that this scheme leads to low blocking probabilities and efficient sharing of protection bandwidth between multiple working paths.

Our optimal routing architecture together with the **elaborated** protection schemes could be a promising solution for MPLS inter-area traffic engineering. In addition, we extended the routing and protection schemes developed in the context of multi-area OSPF to the case where several Internet domains (ASes) are interconnected. We showed that the virtual routing architecture and the protection schemes can also be applied to the case where several ISP networks (ASes) are interconnected by a star-like interconnection structure, as sometimes used for so-called Internet Exchanges (IX). An AAPN could thus be used for the realization of an Internet Exchange, which we call AIX (AAPN-based IX). AIX, as a novel IX architecture, for the first time, has the capability of optimal inter-ISP traffic engineering, which can not be provided by other known IX architectures in a scalable manner. We show the outstanding and distinguished advantages of AAPN when deploying it as an inter-domain (area or AS) inter-connecting infrastructure, which we believe could be one of the major applications of AAPN.

By generalizing the AAPN as the OSN (Overlaid-Star Network), or even further, as the FSN (Flat-Star Network), this thesis proposes a brand-new inter-domain traffic engineering architecture, called *Star-TE*, which is based in fact on the routing and protection methods we developed for optimal traffic engineering in the inter-area and inter-domain contexts over star-like architectures. It can not only be used when an AAPN is adopted as inter-connecting structure, but also when a hardware architecture of overlaid stars (OSN) or even a flat star (FSN), realized for instance by fast Ethernet technology, is used. After checking the RFC 4105 (Requirements for MPLS Inter-Area Traffic Engineering) and RFC 4216 (Requirements for Inter-AS Traffic Engineering), we conclude that the Star-TE satisfies all the requirements for MPLS inter-domain (area or AS) traffic engineering defined by IETF.

As we know, the main objective of inter-domain traffic engineering is to have (optimal) efficient routing (including the associated protection) in a large multi-domain network that obeys QoS requirements and features global fast recovery. But this is a yet unsolved problem, even theoretically. The difficulty comes from two aspects: one is the conflict between routing optimality and routing information scalability, another is the conflict between the routing optimality and routing information confidentiality. For optimal inter-domain routing, as a general rule, the overall optimization process should have global TE information. But this is obviously impossible in large-scale multi-domain networks, which requires the optimization process to be distributed over the network domains (e.g., one optimization sub-process per domain). Each domain should only be responsible for optimizing its own routing, and only a limited cooperation/interaction between domains would be needed. Scalability considerations limit the *quantity* of the interaction between domains to be small; confidentiality considerations limit the *contents* of the interaction between domains which should not disclose domain-specific internal information. However, it is still not clear what kind of information should be exchanged between domains to implement global optimization, even theoretically [Tomaszewski2007].

The most attractive existing solution to inter-domain traffic engineering is the PCE-based architecture proposed by IETF [RFC 4655]. By exchanging so-called “VSPT” (Virtual Shortest Path Tree) [Vasseur2007b] between neighbor domains, PCE’s BRPC (Backward Recursive Path Computation) path computation technique can compute optimal inter-domain paths given a domain sequence that the path will traverse. But the BRPC technique still has a potential scalability issue when deployed in large-scale (long domain sequence, multi-geography) networks, and/or when multiple end-to-end paths are computed.

All the above problems can be solved (at least to a big extent) by deploying Star-TE among domains in one or several “strategic” locations. The best property of Star-TE is that it can naturally “merge” (like a “glue”) local optimizations of domains directly into global optimization. Star-TE solves the scalability issue by adopting the “virtual router” concept so as to separate the IP/MPLS control plane into each individual and independent domain. Star-TE has no requirement on the domain-local path computation mechanism (e.g., PCE, AAPN-based, or any others) in each domain. Only reachability information is needed to be exchanged among the domains connected by Star-TE. Hence there is no confidentiality issue here. In addition to the MPLS inter-area and inter-AS optimal traffic engineering, Star-TE can implement several key and important traffic engineering applications in an easy and scalable manner, which can not be provided by any other existing traffic engineering approaches. These extra applications include: inter-domain global concurrent optimization, host-to-(remote)-host optimal routing in MPLS VPN environment, universal inter-connect architecture for sub-networks deployed through various technologies (e.g., MPLS, GMPLS, PBT, T-MPLS), etc.

Furthermore, combing the best points of PCE and Star-TE architectures, this thesis proposes the segmented-PCE architecture that does not has the potential scalability problems as the original PCE proposal. We believe that the segmented-PCE architecture could be a promising candidate for a “good” inter-domain traffic engineering solution as discussed in Section 5.5.1.

6.2. Overview of Thesis Contributions

This thesis brings the following contributions:

1. **Routing Architecture of MPLS Flows over AAPN**.

The main contribution of this thesis work lies in the optimal routing architecture for MPLS flows over AAPN. I proposed and designed[1] several schemes to improve the scalability when deploying AAPN in single OSPF area networks.

AAPN is more suitable to be used in multi-area network environment due to its agility at the core and large capacity. Based on deploying AAPN as the backbone in multi-area networks, I proposed and developed a novel and scalable framework[1,2] that implements inter-area MPLS traffic engineering with the two distinguishing characteristics, namely

[1] Peng He and Gregor von Bochmann, “Routing of MPLS flows over an agile all-photonic star network”, in Proc. International Conference on Communication Systems and Applications (CSA 2006), July, 2006. pp. 138-144.

[2] Peng He and Gregor von Bochmann, “A Novel Framework for Inter-area MPLS Optimal Routing”, Internet Draft, draft-he-ccamp-optimal-routing-00.txt, Sept., 2006.

global routing optimality guarantee and good backward compatibility with existing OSPF routers. This is described in Section 3.2.2.

2. **Inter-Area Shared Segmented Protection of MPLS Flows over AAPN**

I proposed[1] inter-area shared segment protection schemes for MPLS flows over AAPN under the traditional single-failure assumption and the weakened single-failure (multiple failures in multi-area networks) assumption we proposed for multi-area networks, respectively. The protection schemes take advantage of the flexibility of shared segment protection and the architecture of the AAPN with several overlaid core nodes for providing reliability over the AAPN but also (at the same time) over the Internet networks (areas) that are connected through the AAPN. I also developed a related signaling procedure for establishing working and protection paths, together with an inter-area routing/protection information management scheme. The simulation studies confirm the very favorable performance characteristics of the protection schemes compared with other applicable protection approaches. This is described in Section 4.4.

3. **A Novel Internet eXchange (IX) Architecture based on AAPN: AIX**

I proposed a novel Internet Exchange (IX) architecture, namely AIX[2], which adopts an AAPN as an IX. AAPN can be considered as a "distributed switch", which combines the advantages of the network and switch. Compared to other IX architectures, e.g., LAN-based IX, MPLS IX, Photonic IX, etc., AIX has good properties of scalability, resilience, and widely distributed access points. Particularly, for the first time, AIX introduce traffic engineering (TE) into the IX world. Based on the TE framework we developed for AIX, AIX can provide optimized dynamic inter-AS/ISP (Internet Service Provider) routing while requiring no change, hardware or software, on existing traditional IP/MPLS routers. I have shown by simulation that our TE framework outperforms several existing inter-AS TE schemes. Part of the above inter-area shared segment protection can be applied to AIX, thus we can have a complete inter-AS traffic engineering solution (optimal routing and protection) for AIX. This is described in Section 3.3.

4. **Star-TE: A Promising General Solution to Inter-Domain Traffic Engineering**

[1] Peng He and Gregor von Bochmann, "Inter-Area Shared Segment Protection of MPLS Flows over Agile All-Photonic Star Networks", IEEE 2007 Globecom conference, November, 2007.

[2] Peng He and Gregor von Bochmann, "OSN-IX: A Novel Internet eXchange (IX) Architecture based on Overlaid-Star Networks", submitted to NGI2008 conference.

As the generalized extension to the AAPN-based inter-area and inter-AS traffic engineering frameworks, I proposed and developed[1] a novel inter-domain optimal traffic engineering architecture, called Star-TE, on the basis of the overlaid-star networks (OSN). Star-TE adopts the concept of virtual router (v-ABR or v-ASBR) to separate inter-domain control plane so as to solve the inter-domain scalability and confidentiality issues while keeping the global optimality guarantee. This is described in Section 5.2.

I also proposed the Segmented PCE architecture that solves the inherent scalability and robustness issues in the original PCE architecture proposed by IETF. With Star-TE deployed inside, the PCE architecture can be used in large-scale, multi-geography, multi-provider, and multi-domain networks. This is described in Section 5.3.3.

Simulation results showed that the performance of Star-TE is very close to the ideal global knowledge case. Furthermore, the simulation results also indicated the particular effective conditions/ranges for the outgoing-view-extension and incoming-view-extension techniques that are considered as key improvement to the commonly-used per-domain path computation technique. Based on this result, I believe that the "virtual core router" concept in Star-TE can be used to answered the question, to a large extend, of how to export TE information when the inter-domain connecting facility is a NBMA (non-broadcast multi-access, ATM, Ethernet, etc.) media, a problem that was not solved yet.

During the last three years, I also worked on the following research topics:

1. **Delay Performance Analyses for an Agile All-Photonic Star Network**

Dr. Cheng Peng and I proposed two analytical models[2], called first-fit model (by Dr. Peng) and first-fit with random model (by myself), to analyze the delay performance for the AAPN and made the simulation to verify the model. It is shown that, if a bandwidth allocation algorithm keeps allocating free bandwidth (i.e. the bandwidth that is not allocated to any requests), the allocation algorithm may achieve a good delay performance especially in long-haul networks. This is not described in this thesis.

2. **Blocking Analysis for Time-Space Switched All-Optical Networks**

[1] Peng He and Gregor von Bochmann, "Overlaid-Star Networks for Inter-Area and Inter-AS MPLS Traffic Engineering", submitted to OSA JON journal.

[2] Cheng Peng, Peng He, Gregor v. Bochmann and Trevor J. Hall, "Delay performance analysis for an agile all-photonic star network", 5th International IFIP-TC6 Networking Conference, Coimbra, Portugal, May 15-19, 2006, Proceedings. Lecture Notes in Computer Science 3976 Springer, 2006, pp. 368-378.

Dr. Bin Zhou and I proposed and developed a new analytical model[1] based on the inclusion-exclusion principle from combinatorics, for evaluating the blocking performance of time-space switched optical networks with fixed routing and random wavelength/timeslot assignment. This model can be used to analyze networks with arbitrary topologies and traffic patterns. The accuracy of the proposed analytical model is validated through simulations. In the work, Dr. Zhou and I developed the model together, and I also focus on the related simulation work. This is not described in this thesis.

3. **Optimization Analysis of Optical Time Slot Interchangers in All-Optical Network**

Mr. Hassan Zeineddine and I studied optimization issues of deploying passive optical time slot interchangers (POTSI), a simplified form of the Optical Time Slot Interchanger (OTSI), in all-optical networks. We conducted a comparison[2] between POTSI and the various OTSIs noted in the literature in terms of fiber length, crossbar size, and number of switching operations. Furthermore, we proposed an optimized form of POTSI, Limited-Range POTSI (POTSI-LR), whose capability is limited to switching a timeslot to a subset of nearby timeslots in the frame instead of all possible timeslots. Meanwhile, we investigated the sharing of POTSI-LR as opposed to dedicating one device to each ongoing link. Through analytical and simulation results, we showed that deploying shared limited-range POTSIs can achieve blocking probabilities very close to those of dedicated full-interchanging-range POTSIs. Precisely, the POTSI sharing-percentage can be as small as 20% of the nodal degree together with an interchanging-range as small as 30%; and hence, the overall cost and crossbar complexity can be substantially reduced while still maintaining close to optimal performance gain. These results can be used to guide the design of OTSIs for optical networks. In this work, Zeineddine and I proposed the sharing optimal POTSI architecture and I also focused on the analytical analysis work. This is not described in this thesis.

[1] Bin Zhou, Peng He, and Gregor von Bochmann, “Blocking Analysis For Time-Space Switched All-Optical Networks”, Proc. Intern. Conf. on Optical Communication Systems and Networks (IASTED), Banff, Canada, July 2004.

[2] H. Zeineddine, P. He and G. v. Bochmann, Optimization analysis of optical time slot interchangers in all-optical networks, Proc. Intern. Conf. on Optical Communication Systems and Networks (IASTED), Banff, Canada, July 2006.

6.3. Future Work

This thesis leaves the following aspects for further study:

1) **Further study of the segmented-PCE architecture**. We believe the segmented-PCE architecture proposed in this thesis has the great potential to be a “good” solution for inter-domain traffic engineering. Further research work on this topic includes:

- Thorough performance comparison between the segment-PCE and original PCE architecture in meshed large-scale multi-domain networks;
- Given a meshed large-scale multi-domain network, and a traffic matrix (intra-domain and inter-domain), how to choose “strategic” location(s) to deploy Star-TE? Any tradeoff between the cost of deploying Star-TE and the network performance improvement?
- How to compute an optimal domain sequence in a segmented PCE architecture?
- Further study on the possible usage of Star-TE (and/or Segmented PCE) for addressing the complexity of traffic engineering in large, multi-vendor, multi-technology or multi-domain networks.

2) Study **what possible role the Star-TE enabled Internet Exchange (IX) can play in the commercial framework proposed by IPSphere Forum**. The IPsphere Forum is an international, industry-wide, non-profit association of IT, telecommunications, and networking companies with the mission of enabling the "Business of IP" by developing an open multi-stakeholder multi-geography web-services framework for the rapid creation and automated deployment of IP-based services [IPSF]. The IPsphere Forum proposes the addition of a business layer. The ISPs advertise the services that they support in the business layer. However, these services must be composed or inter-connected so as to provide end-to-end services. Before providing a new service between a source and destination that are located in different ISP networks, the list of providers that this new service has to cross must be determined. The complexity of this problem is similar to the complexity of inter-domain optimal routing that we have studied in this thesis. Hence we believe that the Star-TE enabled Internet Exchange can play an important role in the IPsphere business framework since an IX usually inter-connects several ISP networks and a Star-TE enabled IX inter-connects these ISP networks optimally for traffic engineering. But

how to merge Star-TE enabled IXes (also including existing IXs) into the IPsphere framework (e.g., functionality extensions, protocol extensions, etc.) and related performance issues remain as open problems.

3) **Define a protocol between edge nodes for inter-area (OSPF) and inter-AS (BGP) routing and protection**. The work in [Chou2002] can be used as a reference.

4) **Implementation of Star-TE in the AAPN prototype**. It would be very valuable and interesting to implement the Star-TE in the AAPN prototype [Bochmann2007] so as to evaluate the performance of Star-TE in this near-realistic prototype environment, either with multi-area or multi-AS configuration.

References

[**Ams**] Amsterdam Internet Exchange, Holland, http://www.ams-ix.net.

[**Ahn2002**] G. Ahn and J. Jang, "An Efficient Rerouting Scheme for MPLS-based Recovery and Its Performance Evaluations", Telecommunication Systems, Vol. 3, pp.481-495, 2002.

[**Akyamac2002**] A. A. Akyamac, S. Sengupta and J.-F. Labourdette, "Reliability in Single-domain vs. Multi-domain Optical Mesh Networks", National Fiber Optic Engineers Conference (NFOEC) 2002, Dallas (USA), September 2002.

[**Azim2004**] Mohamed Mostafa A. Azim, Xiaohong Jiang, M. M. R. Khandker, Susumu Horiguchi, and Pin-Han Ho, "Active Lightpath Restoration in WDM Networks", Journal of Optical Networking , Vol. 3, No. 4, pp. 247-260, April, 2004.

[**Azim2006**] Mohamed Mostafa A. Azim, Xiaohong Jiang and Susumu Horiguchi, "Performance Analysis for Active Restoration-Based Optical Networks Incorporating the Correlation among Backup Routes", Proceeding of IEEE 2006 ICC conference, Volume 6, June 2006, pp. 2483 - 2488.

[**Birman1996**] A. Birman, "Computing approximate blocking probabilities for a class of all-optical networks", IEEE JSAC, vol. 14, pp. 852-857, Jun, 1996.

[**Blouin2003**] F. J. Blouin, A. Sack, W. D. Grover, H. Nasrallah, "Benefits of pcycles in a mixed protection and restoration approach", Proc. Fourth International Workshop on the Design of Reliable Communication Networks (DRCN 2003), Banff, Alberta, Canada, 19-22 Oct. 2003, pp. 203-211.

[**Bochmann2004**] G.v. Bochmann, M.J. Coates, T. Hall, L. Mason, R. Vickers and O. Yang, "The Agile All-Photonic Network: An architectural outline", in Proc. 22nd Bien. Symp. on Comm., Kingston, Canada, 2004, pp. 217-218.

[**Bochmann2007**] Gregor v. Bochmann, "Design of an agile all-photonic network", Proc. SPIE Asia-Pacific Optical Communications Conference (APON2007), Wuhan, China, November, 2007.

[**Chen1999**] T.M. Chen and T.H. Oh, "Reliable services in MPLS", IEEE Communications Magazine (December 1999), pp. 58–62.

[**Chen2007**] Mach Chen, Renhai Zhang, “OSPF-TE Extensions in Support inter-AS MPLS/GMPLS Traffic Engineering”, draft-ietf-ccamp-ospf-interas-te-extension-02.txt, Internet Draft, November, 2007.

[**Chou2002**] M. Chou, J. Wybenga, and B. C. Kang, “A Routing Coordination Protocol in a Loosely-Coupled Massively Parallel Router”, in Proc. of HPSR02, May 2002, pp. 52-57.

[**Chu2005**] Xiaowen Chu, Bo Li, “Dynamic Routing and Wavelength Assignment in the Presence of Wavelength Conversion for All-Optical Networks”, IEEE/ACM Transactions on Networking, vol. 13, No. 3, pp. 704-715, June, 2005.

[**Chung1993**] S. Chung, A. Kashper, and K. W. Ross, “Computing approximate blocking probabilities for large loss networks with state-dependent routing”, IEEE/ACM Trans. Networking, vol. 1, pp. 105–115, Feb. 1993.

[**Colle2001**] D. Colle et al., “MPLS Recovery Mechanisms for IP-over-WDM Networks”, Special Issue on IP over WDM and Optical Packet Switching, Photonic Net. Commun., vol. 3, no. 1, pp. 23-40, Jan. 2001.

[**Colle2002**] D. Colle et al., “Data-Centric Optical Networks and Their Survivability”, IEEE JSAC, vol. 20, no. 1, pp. 6-20, Jan. 2002

[**Demeester2005**] Piet Demeester, "Recovery in Multi-layer, Multi-domain & Multi-service Core Networks", tutorial in The 5th International Workshop on Design of Reliable Communication Networks, October 16, 2005.

[**Doucette2003**] J. Doucette, D. He, W. D. Grover, 0. Yang, "Algorithmic approaches for efficient enumeration of candidate p-cycles and capacitated p-cycle network design", Proc. of Fourth International Workshop on the Design of Reliable Communication Networks (DRCN 2003), Banff, Alberta, Canada, 19-22 Oct. 2003, pp. 212-220.

[**EIXA**] European Internet Exchange Association, http://www.euro-ix.net.

[**Farkas2005**] Farkas, A.; Szigeti, J.; Cinkler, T.; “P-cycle based protection schemes for multi-domain networks”, Proceedings of 5th International Workshop on Design of Reliable Communication Networks, 2005. (DRCN 2005). 16-19 Oct. 2005, Page(s): 8 pp.

[**Farrel2006**] Adrian Farrel, et al., “A Path Computation Element (PCE) Based Architecture”, draft-ietf-pce-architecture-05.txt, April 2006.

[**Feamster2003**] N. Feamster, J. Borkenhagen and J. Rexford, “Guidelines for Inter-domain Traffic Engineering”, ACM SIGCOM Computer Communications Review, vol.33, no. 5, pp. 19–30, October 2003.

[**Fedyk2007**] Don Fedyk, David Allan, et al., “GMPLS control of Ethernet PBB-TE”, Internet draft, draft-fedyk-gmpls-ethernet-pbb-te-02.txt, November, 2007.

[**Girard1989**] A. Girard and M. A. Bell, “Blocking evaluation for networks with residual capacity adaptive routing”, IEEE Trans. Commun., vol. 37, pp. 1372–1380, Dec. 1989.

[**Greenberg1997**] A. G. Greenberg and R. Srikant, “Computational techniques for accurate performance evaluation of multirate, multihop communication networks”, IEEE/ACM Trans. Networking, vol. 5, pp. 266–277, Feb. 1997.

[**Groebbens2005**] A. Groebbens et al., “Logical topology design for IP rerouting: ASONs versus static OTNs”, Photonic Netw. Commun., 2005.

[**Grover 2000**] W. D. Grover, D. Stamatelakis, "Bridging the ring-mesh dichotomy with p-cycles", Proc. IEEE/VDE Workshop on Design of Reliable Communication Networks (DRCN 2000), Munich, Germany, April 2000, pp. 92-104.

[**Grover2002**] W. D. Grover, J. Doucette, "Advances in optical network design with p-cycles: Joint optimalization and pre-selection of candidate p-cycles", Proc. of IEEE/LEOS Summer Topicals 2002, Mont Tremblant, PQ, July 2002, pp. 49-50.

[Grover2003] W. D. Grover, “Mesh-based survivable transport networks: options and strategies for optical, MPLS, SONET and ATM networking”, Prentice Hall PTR, Upper Saddle River, New Jersey, Aug. 2003.

[**Guo2007**] Aihua Guo, et al., “Interdomain traffic engineering in ASON/GMPLS controlled multilayer optical networks”, Journal of Optical Networking, vol.6., pp. 719-727, 2007.

[Hall2005] Trevor J. Hall, Sofia A. Paredes, Gregor v. Bochmann. "An agile all-photonic network", in Proceedings of the International Conference on Optical Communications and Networks, ICOCN 2005; Bangkok, Thailand, 14-16 December 2005, pp. 365-368.

[He2006a] P. He and G. v. Bochmann, “Routing of MPLS flows over an agile all-photonic star network”, in Proceedings of IASTED International Conference on Communication Systems and Applications (CSA 2006), July, 2006, pp. 138-144.

[He2006b] P. He and G. v. Bochmann, “A Novel Framework for Inter-Area MPLS Optimal Routing”, Internet Draft, draft-he-ccamp-optimal-routing-00.txt, September, 2006.

[**He2007a**] P. He and G. v. Bochmann, “Inter-area shared segment protection of MPLS flows over agile all-photonic star networks”, IEEE GLOBECOM 2007.

[**He2007b**] P. He, G.v. Bochmann, “Overlaid-Star Networks for Inter-Area and Inter-AS MPLS Traffic Engineering”, submitted to OSA JON journal.

[**He2007c**] P. He and G.v. Bochmann, “A novel Internet eXchange (IX) architecture based on overlaid-star all-optical networks”, submitted for NGI2008 conference.

[**Haskin2000**] D. Haskin and R. Krishnan, “A method for setting an alternative label switched paths to handle fast reroute”, IETF Draft, draft-haskin-mpls-fast-reroute-05.txt, Nov. 2000.

[**Ho2002**] P.-H. Ho and H. T. Mouftah, “A framework of service guaranteed shared protection for optical networks”, IEEE Commun. Mag., vol. 40, pp. 97–103, Feb. 2002.

[**Ho2004a**] Pin-Han Ho, János Tapolcai, and Tibor Cinkler, “Segment Shared Protection in Mesh Communications Networks with Bandwidth Guaranteed Tunnels”, IEEE/ACM transactions on networking, vol. 12, no. 6, pp. 1105-1118, December, 2004.

[**Ho2004b**] P.-H. Ho and H. T. Mouftah, "On optimal diverse routing for shared protection in mesh WDM networks", IEEE Trans. Reliability, vol. 53, pp. 216-225, June 2004.

[**Ho2004c**] P.-H. Ho and H. T. Mouftah, “Shared Protection in Mesh WDM Networks”, IEEE Communications Magazine, vol. 42, issue: 1, pp. 70-76, January, 2004.

[**Hsu2001**] Ching-Fang Hsu, Te-Lung Liu, Nen-Fu Huang, “Performance of Adaptive Routing Strategies in Wavelength-Routed Networks”, Proc. of IEEE International Conference on Performance, Computing, and Communications, pp. 163-170, 2001.

[**Huang2002**] C. Huang et al., “Building reliable MPLS networks using a path protection mechanism”, IEEE Commun. Mag., vol. 40, issue: 3, pp. 156-162, Mar. 2002.

[**Huang2004**] C. Huang and D. Messier, "A Fast and Scalable Inter-Domain MPLS Protection Mechanism", Journal of Communications and Networks, Vol.6, No.1, pp. 375 – 380, March 2004.

[**IPSF**] IPsphere Forum, http://www.ipsphereforum.org/

[**Jap**] Japan Internet Exchange, Japan, http://www.jpix.ad.jp

[**Kang2003**] Jianghui Kang, Martin J. Reed, “Bandwidth Protection in MPLS Networks Using p-Cycle Structure”, Proc. of Design of Reliable Communication Networks (DRCN) 2003, pp. 356-362, October 19-22, 2003.

[**Kodialam2001**] M. Kodialam and T. V. Lakshman, “Integrated dynamic IP and wavelength routing in IP over WDM networks”, in Proc. of INFOCOM, vol.1, pp. 358-366, 2001.

[**Koizumi2005**] Yuki Koizumi, Shin'ichi Arakawa, and Masayuki Murata, "On the integration of IP routing and wavelength routing in IP over WDM networks", in Proceedings of SPIE APOC, pp. 20-29, Nov. 2005

[**Koizumi2006**] Yuki Koizumi, "Cross-layer traffic engineering in IP over WDM networks", Master's thesis, Graduate School of Infromation Science and Technology, Osaka University, Feb. 2006.

[**Kolarov1995**] Kolarov, A. and Hui, J., "On computing Markov decision theory-based cost for routing in circuit-switched broadband networks", Journal of Network and Systems Management, vo1.3, no.4, 1995, pp.405-426.

[**Koo2004**] S. Koo, G. Sahin, and S. Subramaniam, "Dynamic LSP provisioning in overlay, augmented, and peer architectures for IP/MPLS over WDM networks", in Proceeding of IEEE INFOCOM, vol. 1, pp. 523-532, Mar. 2004.

[**Kumaki2007**] K. Kumaki, R. Zhang, "Requirements for supporting Customer RSVP and RSVP-TE Over a BGP/MPLS IP-VPN", Internet Draft, draft-kumaki-l3vpn-e2e-rsvp-te-reqts-05.txt, November 19, 2007.

[**Lang2004**]J. P. Lang, Y. Rekter, and D. Papadimitriou, "RSVP-TE Extensions in Support of End-to-End Generalized Multiprotocol Label Switching (GMPLS)-Based Recovery," Internet draft, draft-ietf-ccamp-gmpls-recovery-e2e-signaling-02.txt, 2004.

[**Larrabeiti2005**] David Larrabeiti, Ricardo Romeral, Ignacio Soto, et al., "Multi-Domain Issues of Resilience", Proceedings of 2005 7th International Conference on Transparent Optical Networks, vol. 1, pp. 375- 380, July 2005.

[**Lee2002**] SuKyoung Lee, David Griffith, and Nah-Oak Song, "An Analytical Approach to Shared Backup Path Provisioning in GMPLS Networks", Proc. of IEEE MILCOM'02, vol. 1, pp. 75– 80, 7-10 Oct. 2002.

[**Lee2007**] Lee, et al., "Path Computation Element Communication Protocol (PCECP) Requirements and Protocol Extensions In Support of Global Concurrent Optimization", Internet-Draft, draft-lee-pce-global-concurrent-optimization-04.txt, May, 2007.

[**Li2002**] L. Li et al., "Routing Bandwidth Guaranteed Paths with Local Restoration in Label Switched Networks", IEEE JSAC., vol. 23, issue 2, pp. 437 – 449, Feb. 2005.

[**Liu2004a**] Xin Liu , Qingji Zeng, Yun Wang, "Multi-Layer Recovery Mechanism for IP over WDM Networks", Proceedings of SPIE Network Architectures, Management, and Applications, vol. 5282, 2004.

[**Liu2004b**] Mingyan Liu and John S. Baras, “Fixed Point Approximation for Multirate Multihop Loss Networks with State-Dependent Routing”, IEEE/ACM Transactions on Networking, vol.12, No. 2, pp. 361-374, April, 2004.

[**Maach2004**] Abdelilah Maach, Gregor v. Bochmann, Hussein Mouftah, "Shared Protection for Time Slotted Optical Networks", Proc. of Third IEEE International Symposium on Network Computing and Applications (NCA'04), pp. 333-336, 2004.

[**Morelli2005**] Mario Morelli et al., “An IPv6 Internet Exchange Model. Lessons from Euro6IX project”, Proceedings of the 2005 Symposium on Applications and the Internet Workshops (SAINT-W’05), pp. 50-53, Jan., 2005.

[**Mason2006**] L.G. Mason, A. Vinokurov, N. Zhao and D. Plant, "Topological design and dimensioning of agile all photonic networks”, Elsevier Computer Networks Journal, vol. 50, No. 2, pp.268-287, 2006.

[**Matsuura2007**] Hiroshi Matsuura, Naotaka Morita, Isami Nakajima, “Hierarchically Distributed PCE for Flexible Multicast Traffic Engineering”, IEEE Globelcom2007.

[**Menth2004**] Michael Menth, Andreas Reifert, Jens Milbrandt, “Self-Protecting Multipaths - A Simple and Resource-Effcient Protection Switching Mechanism for MPLS Networks”, Proc. of IFIP Networking 2004, pp. 526-537.

[**Mitra1993**] D. Mitra, R. Gibbens, and B. D. Huang, “State-dependent routing n symmetric loss networks with trunk reservations I”, IEEE Trans. Commun., vol. 41, pp. 400–411, Feb. 1993.

[**Miyamura2004a**] T. Miyamura, T. Kurimoto, M. Aoku and A. Mis- awa, “An inter-area SRLG-disjoint routing algorithm for multi-segment protection in GMPLS networks”, Proc. ICBN 2004, April, 2004.

[**Miyamura2004b**] “A Multi-layer Disjoint Path Selection Algorithm for Highly Reliable Carrier Services”, Proc. Globecom 2004, vol. 3, pp. 1974- 1978, Dec., 2004.

[**Miyamura2005**] T. Miyamura, T. Kurimoto, M. Aoki, and S. Urushidani,“ A disjoint path selection scheme based on enhanced shared risk link group management for multi-reliability service”, Proc. Globecom 2005, Volume: 4, 28 Nov.-2 Dec. 2005.

[**Nak2002**] Nakagawa, I.; Esaki, H.; Nagami, K., “A design of a next generation IX using MPLS technology”, Proceedings of the 2002 Symposium on Applications and the Internet (SAINT 2002), pp. 238 – 245, 2002.

[**Nortel2001a**] Nortel SMLT White Paper, http://www.nortel.com/products/01/passport/ 8600_rss/collateral/ nn108460-060304.pdf.

[**Nortel2001b**] Nortel RSMLT white paper, http://www.nortel.com/products/01/passport/ 8600_rss/collateral/nn107680-031804.pdf.

[**Nortel2007**] Nortel PBT (Provider Backbone Transport) White Paper, http://www.nortel.com/solutions/collateral /nn115500.pdf

[**Otani2007**] T. Otani, K. Ogaki, S. Okamoto, "GMPLS Inter-Domain Routing in support of inter-domain links", Internet Draft, draft-otani-ccamp-gmpls-routing-interlink-01.txt, November, 2007.

[**Pan2002**] P. Pan et al., "Fast reroute extensions to RSVP-TE for LSP tunnels", IETF Draft, draft-ietf-mpls-rsvp-lsp-fastreroute-00.txt, Jan., 2002.

[**Papadimitriou2003**] D. Papadimitriou, E. Mannie, D. Brungard, S. Dharanikota, J. Lang, G. Li, B. Rajagopalan, and Y. Rekhter, "Analysis of Generalized MPLS-based Recovery Mechanisms (including Protection and Restoration)", Internet Draft, draft-papadimitriou-ccamp-gmpls-recovery-analysis-03.txt, May, 2003.

[**Pickavet2006**] Pickavet, M.; Demeester, P.; Colle, D.; Staessens, D.; Puype, B.; Depre, L.; Lievens, I., "Recovery in multilayer optical networks", Journal of Lightwave Technology, Volume 24, Issue 1, 2006.

[**Qiao2001**] C. Qiao and D. Xu, "Distributed Partial Information Management (DPIM) schemes for survivable networks - Part I, Part II", Proc. of IEEE INFOCOM, Apr. 2001.

[**Qiao2002**] D. Xu, Y. Xiong, and C. Qiao, "Protection with Multi-Segments (PROMISE) in Networks with Shared Risk Link Groups (SRLG)", 40th Annual Allerton Conf. Commun., Control, and Comp., 2002.

[**Qin2003**] Yang Qin, Mason, L., Ke Jia, "Study on a joint multiple layer restoration scheme for IP over WDM networks", IEEE Network, vol. 17, Issue 2, pp. 43-48, March, 2003.

[**RFC 1771**] A border gateway protocol 4 (BGP- 4)", March, 1995.

[**RFC 2205**] Resource ReSerVation Protocol (RSVP), September, 1997.

[**RFC 2328**] OSPF Version 2, April, 1998.

[**RFC 2858**] Multiprotocol Extensions for BGP-4, June, 2000.

[**RFC 3031**] Multiprotocol Label Switching Architecture, January, 2001.

[**RFC 3036**] LDP Specification, October, 2001. (obsoleted by RFC5036, October, 2007)

[**RFC 3209**] RSVP-TE: Extensions to RSVP for LSP Tunnels, December, 2001.

[**RFC 3469**] Framework for MPLS based recovery, February, 2003.

[**RFC 3630**] Traffic Engineering (TE) Extensions to OSPF Version 2, September, 2003.

[**RFC 3768**] Virtual Router Redundancy Protocol (VRRP), April, 2004.

[**RFC 4105**] Requirements for Inter-Area MPLS Traffic Engineering, June, 2005.

[**RFC 4216**] MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements, November, 2005.

[**RFC 4655**] A Path Computation Element (PCE)-Based Architecture, August, 2006.

[**RFC 4873**] GMPLS Segment Recovery, May, 2007.

[**Ricciato2004**] F. Ricciato, U. Monaco, and A. D'Achille, "A Novel Scheme for End-to-End Protection in a Multi-Area Network", IPS '04, Budapest, Hungary, Mar. 2004.

[**Ricciato2005**] F. Ricciato, U. Monaco, and A. D'Achille, "Distributed Schemes for Diverse Path Computation in Multi-domain MPLS Networks", IEEE Communications Magazine, Volume 43, Issue 6, Page(s): 138 –146, June, 2005.

[**Saad2004**] Tarek Saad and Hussein T, Mouftah, "Inter-Domain Wavelength Routing in Optical WDM Networks", 11th International Telecommunications Network Strategy and Planning Symposium, IFIP Networking 2004, Page(s): 391 – 396, June 2004.

[**Schupke2002**] D. A. Schupke, C. G. Gruber, A. Autenrieth, "Optimal configuration of p-cycles in WDM networks," Proc. IEEE International Conference on Communications (ICC 2002), New York City, NY, 28 April - 2 May, 2002, vol. 5, pp. 2761 - 2765.

[**Shake2005**] I. Shake, et al., Tsukishima, and W. Imajuku, " Experiments on optical link capacity adjustment for photonic IX", 31st European Conference on Optical Communications (ECOC2005), Tu.3.4.3, 2005.

[**Shrimali2007**] G. Shrimali, A. Akella and A. Mutapcic, "Cooperative Inter-Domain Traffic Engineering Using Nash Bargaining and Decomposition", Proc. of IEEE INFOCOM 2007, Page(s):330 – 338, May 2007

[**Sriram2003**] Sriram, Kotikalapudi; Griffith, David W.; Lee, SuKyoung; Golmie, Nada, "Backup Resource Pooling in (M:N)^^{n} Fault Recovery Schemes in GMPLS Optical Networks", Proceedings of SPIE , October 13-17, 2003 , Dallas, TX - October 01, 2003.

[**Su2001**] C. Su and X. Su, "An On-line Distributed Protection Algorithm in WDM Networks", Proc. of IEEE ICC 2001, vol.5, pp. 1571 – 1575, 2001.

[**Staessens2006**] Dimitri Staessens, Leen Depr´e, Didier Colle, et al., “A Quantitative Comparison of Some Resilience Mechanisms in a Multi-domain IP-over-Optical Environment”, Proc. of IEEE ICC 2006, vol. 6, pp. 2512 – 2517, June 2006.

[**S&D**] Switch and Data, United States, http://www.switchanddata.com.

[**Tan2005**] X. Tan, O. Liang, and W. Cheng , “An analytical framework for performance of different fault restoration policies with QoS constraints in MPLS networks”, The IEEE Conf. on Local Computer Networks 30th Anniversary (LCN'05), pp. 226-233, 2005.

[**Tapolcai2003**] J. Tapolcai and T. Cinkler, “On-line Routing Algorithm with Shared Protection in WDM Networks”, ONDM, Budapest, Hungary, Feb. 2003.

[**Thiongane2005**] B. Thiongane, D.L. Truong , "Shared Path Protection in Multi-domain Optical Mesh Networks", in Proceeding of IASTED Computer and Communication Network, Marina del Rey, CA, USA, Oct. 24-26, 2005, pp.138-145.

[**Torab2006**] Payam Torab and Bijan Jabbari, “On Cooperative Inter-Domain Path Computation”, Proc. of IEEE Symposium on Computers and Communications (ISCC 2006), pp. 511–518, 26-29 June 2006.

[**Tomaszewski2007**] Artur Tomaszewski, Michał Pi´oro, Mariusz Mycek, “Distributed Inter-Domain Link Capacity Optimization for Inter-Domain IP/MPLS Routing”, IEEE Globelcom2007, November, 2007.

[**TPACK2007**] TPACK, “T-MPLS: A New route to carrier Ethernet”, http://www.tpack.com/fileadmin/user_upload/Public_Attachment/T-MPLS_WP_v2_web.pdf, June, 2007.

[**Vasseur2007a**] Vasseur, J., "A Per-domain path computation method for establishing Inter-domain Traffic Engineering (TE) Label Switched Paths (LSPs)", Internet Draft, draft-ietf-ccamp-inter-domain-pd-path-comp-04 (work in progress), February, 2007.

[**Vasseur2007b**] Vasseur, J., "A Backward Recursive PCE-based Computation (BRPC) procedure to compute shortest inter-domain Traffic Engineering Label Switched Paths", Internet Draft, draft-ietf-pce-brpc-04 (work in progress), March, 2007.

[**Vasseur2007c**] JP. Vasseur and JL. Le Roux, “Path Computation Element (PCE) communication Protocol (PCEP) ", Internet Draft, draft-ietf-pce-pcep-09.txt, Nov., 2007.

[**Vasseur2007d**] JP. Vasseur, Gargi. Nalawade, and K. Kumaki, “An MP-BGP protocol extension to advertize TE-related PE-CE link information”, Internet Draft, draft-vasseur-ccamp-ce-ce-te-03, November 16, 2007.

[**Vickers2000**] R Vickers and M Beshai, “PetaWeb Architecture”, Networks 2000-toward Natural Networks: the 9th International Telecommunication Network Planning Symposium, Toronto, Canada, 10-15 Sept. 2000.

[**Wang2002**] Dongmei Wang, John Strand, Jennifer Yates, “OSPF for Routing Information Exchange Across Metro/Core Optical Networks”, Optical Networks Magazine, Vol. 3, Issue 5, 2002.

[**Winnick2002**] J. Winnick, S. Jamin, J. Rexford, “Traffic Engineering Between Neighboring Domains”, Tech. Report, July, 2002.

[**Xiong2003**] Y. Xiong, D. Xu, and C. Qiao, “Achieving fast and bandwidth efficient shared-path protection”, J. Lightwave Technol., vol. 21, pp. 365–371, Feb. 2003.

[**Xu2002**] D. Xu, C. Chunming, and Y. Xiong, “An Ultra-Fast Shared Path Protection Scheme - Distributed Partial Information Management, Part II”, in Proc. of 10th IEEE International Conference on Network Protocols, Page(s): 344 – 353, 12-15 Nov. 2002.

[**Xu2004**] K. Xu, Z Duan, Z. L. Zhang, J. Chandrashekar, “On Properties of Internet Exchange Points and their Impact on AS Topology and Relationship”, IFIP Networking 2004, Athens, Greece.

[**Ye2001**]Y. Ye, C. Assi, S. Dixit, and M. A. Ali, “A simple dynamic integrated provisioning/protection scheme in IP over WDM networks”, IEEE Communications Magazine, vol. 39, Issue 11, pp. 174–182, November 2001.

[**Yoon2001**] S. Yoon et al., “An efficient recovery mechanism for MPLS-based protection LSP”, in Proc. of Joint 4th IEEE International Conf. on ATM (ICATM 2001), pp. 75–79, Seoul, Korea, April 2001.

[**Zheng2003**] Q. Zheng and G. Mohan, “An efficient dynamic protection scheme in integrated IP/WDM networks”, in Proc. IEEE ICC2003, vol.2, pp. 1494-1498, May 2003.

[**Zheng2004**] Jun Zheng, Baoxian Zhang, Mouftah, H.T., “Dynamic path restoration based on multi-initiation for GMPLS-based WDM networks”, 2004 IEEE International Conference on Communications, vol. 3, pp.1639 – 1643, 20-24 June 2004.

[**Zhu2003**] Y. Zhu, A. Jukan and M. Ammar, "Multi-Segment Wavelength Routing in Large-Scale Optical Networks", in Proc. of IEEE ICC2003, vol.2, pp. 1381–1385, 11-15 May 2003.