Optical network control plane for multi-domain networking

Fagertun, Anna Manolova; Ruepp, Sarah Renée; Buron, Jakob Due; Ellegård, Lars; Dittmann, Lars

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Optical Network Control Plane for Multi-domain Networking

Anna Vasileva Manolova

04.12.2009

Networks Technology and Service Platforms
DTU Fotonik
Technical University of Denmark
Building 343
2800 Kgs. Lyngby
DENMARK

Abstract

This thesis focuses on multi-domain routing for Traffic Engineering and survivability support in optical transport networks under the Generalized Multi-Protocol Label Switching (GMPLS) control framework.

First, different extensions to the Border Gateway Protocol for multi-domain Traffic Engineering are designed and evaluated. Throughout the thesis three extensions are proposed: an end-to-end aggregated TE metric per destination, an extension for multi-path dissemination, and an AS-disjoint path selection modification. It is shown that simple TE metrics applied within the BGP path selection process are not enough for efficient TE in mesh multi-domain networks. Enhancing the protocol with multi-path dissemination capability, combined with the employment of an end-to-end TE metric proves to be a highly efficient solution. Simulation results show good performance characteristics of the proposed extensions, such as providing lower connection blocking and stable protocol overhead. Furthermore, different export policies for multi-path dissemination with the proposed BGP enhancements are designed and evaluated. Simulation results indicate that the amount of disseminated paths per destination is not as essential for improved network performance as the length of the provided paths.

Second, the issue of multi-domain survivability support is analyzed. An AS-disjoint BGP extension is proposed and its performance evaluated under single multi-domain link failure. It is shown that providing AS-disjoint paths is beneficial not only for resilience support, but also for facilitating adequate network reactions to changes in the network, which trigger BGP protocol re-convergence. By notifying the proper network elements for changes in the network (e.g. failures), the connection blocking can be significantly reduced.

Furthermore, novel restoration mechanisms, which provide differentiated failure handling, are proposed and evaluated. It is shown that the applied routing protocol and the topology of the multi-domain network have very strong influence on the efficiency of the applied restoration techniques.

Finally, different challenges of the integration of the GMPLS control framework with the novel Optical Burst Switching technology are analyzed. Existing integration architectures are discussed and categorized and solutions for different protocol extensions are proposed.

Resume

Denne afhandling omhandler optiske netværk i relation til Generalized MultiProtocol Label Switching (GMPLS) protokol suiten. Specielt fokuseres på hvordan multi-domæne routning kan yde robusthed og intelligent trafikstyring.

Først, designes og evalueres de forskellige udvidelser til Border Gateway Protocol for multi-domæne Traffic Engineering. I afhandlingen er tre udvidelser foreslået: en aggregeret destinationsstyret end-to-end metode, en udvidelse til multi-kanal informationdistribution, og en AS-disjunkt modifikation til ruteudvælgelse. Det er vist, at simple TE målinger der anvendes inden for BGP ruteudvælgelseprocesen er utilstrækkelige til effektiv TE i meshed multi-domæne net. En forbedring af protokollen med multi-path informationsdistribution, kombineret med anvendelse af end-to-end TE information viser sig at være en særdeles effektiv løsning. Simuleringsresultaterne viser god ydeevne med de foreslåede udvidelser, såsom lavere sandsynlighed for blokering af forbindelser og et stabilt protokoloverhead. Desuden er forskellige eksport politikker for multi-kanal informationsdistribution med de foreslåede BGP ekstraudstyr er designet og evalueret. Simuleringsresultaterne viser, at mængden af informationsdistribution per destination er ikke så afgørende for bedre netværkets ydeevne som længden af de givne ruter.

For det andet, analyseres spørgsmålet om "overlevelsesevne" i multi-domæne netværk. En AS-disjunkt BGP udvidelse er foreslået og resultaterne er vurderet ved enkelt multi-domæne linkfejl. Det er vist, at benyttes AS-disjunkte ruter gavner dette ikke kun modstandsdygtigheden mod fejl, men det faciliterer også tilstrækkelig imødegåelse af af forandringer i netværket, som udløser BGP protokollens re-konvergens. Ved at notificere de korrekte netkomponenter om ændringer i nettet

(f.eks fejl), kan blokeringssandsynligheden reduceres betydeligt.

Herudover er en nyudviklet reetableringsmekanisme foreslået og evalueret, som yder differentieret håndtering af fejl. Det er vist, at den anvendte routingsprotokol og topologien af multi-domæne netværk har meget stor indflydelse på effektiviteten af de anvendte reetableringsteknikker.

Endelig, analyseres forskellige udfordringer for integrationen af GMPLS kontrolmekanismer med en nyudviklet Optisk Burst Switching teknologi. Eksisterende arkitekturer for sådan integration er diskuteret og kategoriseret og løsninger til forskellige protokol udvidelser er foreslået.

Acknowledgments

First, I would like to thank my main supervisor Professor Lars Dittman, for giving me the opportunity to carry out my Ph.D. project in the *Networks Technology and Service Platforms* group at DTU Fotonik. It has been a real pleasure working with him and learning many new things, among which numerous exciting aspects within the optical networking area. I would also like to thank my supervisor team: Sarah Ruepp, Jakob Buron and Lars Ellegaard, for the fruitful discussions, the support and the encouragement they gave me during my Ph.D. project.

Next, I would like to thank everybody I have collaborated with: Jose Marzo and Eusebi Calle from the University in Girona, Ricardo Romeral from US3M in Madrid and the research group from Scuola Superiore Sant'Anna in Pisa. Thank you for the great collaboration, the interesting discussions, the good ideas, for believing in me and for supporting me. Special thanks to Lars Staalhagen for being my personal OPNET guru.

Special gratitude to all my colleagues from the *Networks Technology and Service Platforms* group for the great time together, the superb working atmosphere they all create, the very much appreciated Friday seminars and for being good friends and supporters.

Finally, special thanks to all my friends, which made sure I do have life out of the office. Thanks to my parents and to Jens for loving me, for being there for me in my worst periods and for making me laugh even when it felt like crying. Special gratitude to Eli and Mira for their support and for making sure I do not forget Bulgarian language.

List of Publications

This Ph.D. project has resulted in 13 peer-reviewed publications, presented in international conferences, and 5 publications currently under review.

Peer-reviewed conference publications:

[1] **A. Manolova**, S. Ruepp, and L. Dittmann, "Performance Comparison of Multi-domain Routing Schemes in GMPLS Networks with BGP," in *IEEE Proc. 17th International Conference on Photonics in Switching (PS)*, September 2009.

[2] N. Sambo, I. Cerutti, A. Giorgetti, P. Castoldi, R. Munoz, S. Ruepp, R. Casselas, R. Martinez, and **A. Manolova**, "Restoration GMPLS-based Wavelength Switched Optical Networks with Limited Wavelength Converters," in *IEEE Proc. 17th International Conference on Photonics in Switching (PS)*, September 2009.

[3] S. Ruepp, A. Koster, N. Andriolli, and **A. Manolova**, "Prioritizing Connection Requests in GMPLS-Controlled Optical Networks," *IEEE Proc. 17th International Conference on Photonics in Switching (PS)*, September 2009.

[4] **A. Manolova**, E. Calle, S. Ruepp, J. Marzo, and L. Dittmann, "Location-based Restoration Mechanism for Multi-domain GMPLS Networks," in *IEEE Proc. International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, July 2009.

[5] **A. Manolova**, S. Ruepp, and L. Dittmann, "TE-enhanced Path Selection for QoS Provisioning in Multi-domain GMPLS Networks,"

in *IEEE Proc. Optical Fiber Communication Conference/National Fiber Optic Engineering Conference (OPF/NFOEC)*, March 2009.

[6] **A. Manolova**, S. Ruepp and L. Dittmann, "On the Efficiency of BGP-TE Extensions for GMPLS Multi-domain Routing," in *IEEE Proc. 13th Conference on Optical Network Design and Modeling (ONDM)*, February 2009.

[7] L. Xiaohua, S. Ruepp, **A. Manolova** and L. Dittmann, "Survivability Enhancing Routing Scheme for Multi-domain Network," in *IEEE Proc. GLOBECOM 2008*, December 2008.

[8] J. Buron, S. Ruepp, H. Wessing, N. Andriolli, **A. Manolova**, and L. Dittmann, "Wavelength Converter Placement in Optical Networks with Dynamic Traffic," in *Proc. Asia Pacific Optical Conference (APOC)*, October 2008.

[9] L. Xiaohua, S. Ruepp, L. Dittmann, and **A. Manolova**, "OPNET Model for Multi-domain Routing with Enhanced Survivability," in *Proc. OPNETWORK 2008*, August 2008.

[10] **A. Manolova**, J. Buron, S. Ruepp, L. Dittmann, and L. Ellegaard, "Modeling Contention Resolution Strategies in Optical Burst Switched Networks," in *Proc. OPNETWORK 2007*, August 2007.

[11] **A. Manolova**, J. Buron, S. Ruepp, L. Dittmann, and L. Ellegaard, "Segmentation-based Path Switching Mechanism for Reduced Data Losses in OBS Networks," in *Proc. 11th International Conference on Optical Networking Design and Modeling (ONDM), Lecture Notes in Computer Science 4534*, May 2007.

[12] **A. Manolova**, S. Ruepp, J. Buron, L. Dittmann, and L. Ellegaard, "Advantages and Challenges of the GMPLS/OBS Integration," in *Proc. VI GMPLS Workshop*, April 2007, pp.133-144.

[13] J. Buron, S. Ruepp, and **A. Manolova**, "Teaching Cost-Effective and Resilient Network Design with OPNET WDM Guru," in *Proc. OPNETWORK 2006*, August 2006.

In review publications:

1 **A. Manolova**, R. Romeral, and S. Ruepp, "Enhancing Network Performance Under Single Link Failure with AS-disjoint BGP Extension", submitted to 4th WSEAS International Conference on CIRCUITS, SYSTEMS, SIGNAL and TELECOMMUNICATIONS.

2 **A. Manolova**, and S. Ruepp, "Export Policies for Multi-Domain WDM Networks", submitted to OFC/NFOEC 2010.

3 **A. Manolova**, E. Calle, S. Ruepp, J. Marzo, and L. Dittmann, "Restoration Management in Multi-domain Networks: Survey and Enhanced Mechanisms", submitted to IEEE Communications Magazine, Network & Service Management Series.

4 S. Ruepp, et al. "The Road to 100 Gigabit Ethernet", submitted to Infocom 2010.

5 **A. Manolova**, E. Calle, S. Ruepp, J. Marzo, and L. Dittmann, "Restoration in Multi-Domain GMPLS-based Networks", submitted to Journal of Computer Communications.

List of Acronyms

ARO	Associated Route Object
AS	Autonomous System
ASON	Automatically Switched Optical Network
BRPC	Backward-Recursive Path Computation
BHP	Burst Header Packet
BGP	Border Gateway Protocol
CP	Control Plane
CCG	Control Channel Group
DCG	Data Channel Group
DB	Data Burst
DP	Data Plane
E2E	End-to-End
E-NNI	External Network-to-Network Interface
ERO	Extended Route Object
FDL	Fiber Delay Line
FF	First Fit
GMPLS	Generalized Multi-Protocol Label Switching

GCC	GMPLS Control Channel
IGP	Interior Gateway Protocol
IETF	Internet Engineering Task Force
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
I-NNI	Internal Network-to-Network Interface
IS-IS-TE	Intermediate System to Intermediate System - Traffic Engineering
IP	Internet Protocol
LMP	Link Management Protocol
LS	Label Set
LSA	Link State Advertisement
LIB	Label Information Base
LSP	Label Switched Path
L2E	Local-to-End
MED	Multi-Exit Discriminator
MPLS	Multi-Protocol Label Switching
MRIT	Minimum Route Advertisement Interval Timer
NLRI	Network Layer Reachability Information
OBS	Optical Burst Switching
OBGP	Optical Border Gateway Protocol
OCS	Optical Channel Switching
OIF	Optical Internetworking Forum
OPS	Optical Packet Switching

OSPF-TE	Open Shortest Path First - Traffic Engineering
OXC	Optical Cross-Connect
PDPC	Per-Domain Path Computation
PCE	Path Computation Element
PCEP	Path Computation Element Communication Protocol
PCC	Path Computation Client
PNNI	Private Network-to-Network Interface
PPRO	Primary Path Route Object
PXC	Photonic Cross-Connect
QoS	Quality of Service
RRO	Recorded Route Object
RSVP	Resource ReserVation Protocol
RSVP-TE	Resource ReserVation Protocol with TE extensions
SLA	Service Level Agreement
SN	Shortest New
SRLG	Shared Risk Link Group
TE	Traffic Engineering
TDM	Time Division Multiplexing
TLV	Type-Length-Value
UI	Update Interval
UNI	User-Network Interface
VPN	Virtual Private Network
VT	Virtual Topology
WDM	Wavelength Division Multiplexing

Contents

Chapter 1

Introduction

1.1 General motivation

The unprecedented boom of different applications and services offered by communication networks during the past 20 years have put a huge burden on the transport network technologies, pushing them to their limits. Introducing optical fibers as transmission medium has proven to be highly effective solution due to the enormous amount of data which can be carried over very long distances. The never ceasing demand for diverse Quality of Service (QoS) provisioning, dynamic network operation, flexibility and network reliability has pushed forward the need for reconfigurability and dynamics at the optical transport level. Thus, dynamically reconfigurable optical networks have become the new trend within the telecommunications industry with both network operators and equipment vendors investing in research activities within the areas of novel optical transmission systems and next generation networking paradigms.

Today, almost every aspect of our every-day life is highly dependent on the proper and secure operation of numerous interconnected communication networks. While the common goal among different network providers is to satisfy the needs of their clients, a common trend among them remains the information preservation within network boundaries. With the advent of new services, the need for more open and effective cooperation between the network providers arises. In particular, the process of dynamic and cost efficient service provisioning across multi-

ple service providers, is currently under investigation by many industry and academia research groups. Relying on simple bi-lateral agreements between domains cannot guarantee globally optimized objectives. Thus, developing a cooperative framework for Traffic Engineering (TE) and QoS provisioning is accepted to be the right way towards achieving optimal service provisioning and network operation.

The problem of automatic cross-domain TE is relatively new. Several standardization bodies work on the development of protocols and mechanisms for facilitating dynamic TE with minimal human intervention. The main problem in this field comes from the very strict requirements which network providers pose on a potential collaborative framework. Furthermore, the existing mechanisms and protocols for multi-domain networking meet certain challenges when applied in Next Generation networks, due to a mismatch between the specifics of the Next Generation networks architectures and the design objectives of the current protocols.

This Ph.D. thesis elaborates on the establishment of such a cooperative framework. In particular it investigates the limitations which some existing protocols and widely-used mechanisms for TE pose on the establishment of such a framework, and explores several novel solutions for TE and QoS provisioning.

1.2 Structure of the thesis

The main focus of this Ph.D. study is on the control of multi-domain optical transport networks. Additionally, several other research topics related to single-domain TE and QoS provisioning have been included. The following areas have been part of the research activities: multi-domain routing and connection provisioning, single- and multi-domain TE and QoS support, network survivability, control plane integration, optical packet and burst switching. Since the main topic of the Ph.D. project is related to multi-domain inter-networking, only work related to this topic is presented in this thesis.

The remainder of the thesis is organized as follows: Chapter 2 gives background information related to optical transport networks, standards and architectures, general multi-domain networking principles, protocols and solutions, and outlines the main scope of the thesis. Chapter 3 evalu-

ates the main challenges the Border Gateway Protocol (BGP) protocol meets when applied in connection-oriented networks. Simple TE extensions to the protocol are evaluated via simulations. In Chapter 4 further extensions to the BGP protocol for multi-domain TE are proposed and evaluated. Comparison with the simple TE extensions is illustrated via simulations. Chapter 5 focuses on BGP extensions for survivability in multi-domain connection-oriented networks, whereas Chapter 6 investigates the problem of multi-domain restoration specifically. Chapter 7 elaborates on the challenge of multi-domain control plane integration with respect to the relatively new optical burst switched technology. Chapter 8 summarizes the main work carried out during this thesis and gives concluding remarks. Two Appendixes are included. Appendix A gives general description of the simulation model, developed during this Ph.D. study for evaluating the efficiency of the different proposals. Appendix B gives details of the different topologies used during the simulations.

Fig. 1.1 illustrates the main research topics, the corresponding chapters where they are discussed, and the related personal publications.

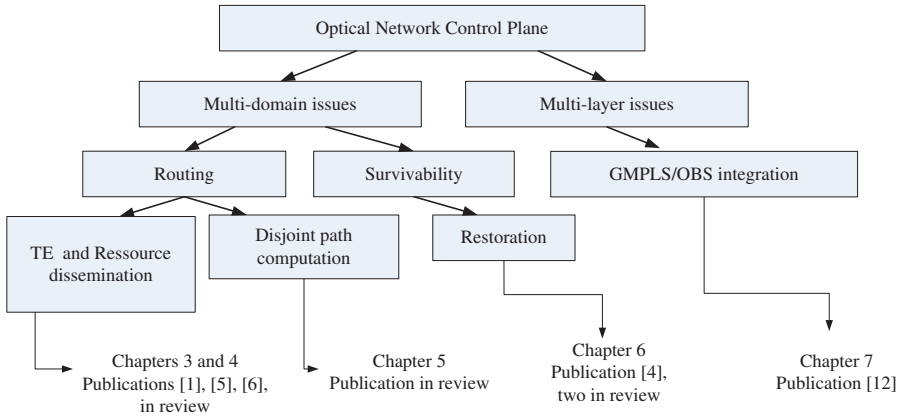


Figure 1.1: Thesis overview: covered topics, related chapters and corresponding publications.

1.3 Personal contributions

This section outlines my personal contributions to all publications, published during the Ph.D. project.

[1] : I developed the used models for the performance evaluation, carried out the simulations, wrote the paper and presented it at the conference. All authors participated in the editing of the paper by providing comments and suggestions.

[2] : I contributed to the editing of the paper by providing comments.

[3] : I contributed to the editing of the paper by providing comments and presented the paper at the conference.

[4] : The idea of applying differentiated failure handling based on the location of the failure along the path was my idea. I implemented the simulation model and carried out the simulations for the verification of several different multi-domain restoration mechanisms. I wrote the sections of the paper related to the analysis of the applicability of restoration in multi-domain environments and provided the simulation results. I presented the paper at the conference. All authors contributed with comments and feedback to the paper.

[5] : This paper evaluates different methods for TE with standard BGP in connection oriented networks and proposes the usage of end-to-end aggregated TE metric. I developed the model, used for the simulation results, I identified the challenges for the application of the BGP protocol in connection-oriented networks and performed the performance evaluation study. I wrote the paper and presented it at the conference. All authors contributed with comments to the paper.

[6] : Based on the analysis presented in the previous publication, I developed the Enhanced BGP protocol. I implemented it and carried out all performance evaluation simulations. I wrote the paper and presented it in the conference. All authors contributed to the paper by providing feedback.

[7] : I was consulting the first author on the topics related to the multi-domain routing process. I gave suggestions and corrections to the presented model in the paper with respect to feasibility and protocol applications. I provided feedback and comments to the paper and I presented the paper at the conference.

[8] : I contributed to the editing of the paper by providing comments and suggestions.

[9] : This paper is based on [7] and presents details of the model used in that work. I contributed with feedback and suggestions to the structure and the content of the paper.

[10] : This paper is based on [11] and presents details of the model, used in the simulations. I wrote the paper and presented it at the conference. All authors contributed with feedback to the paper.

[11] : I designed the contention resolution scheme and implemented it in an event-driven simulator. I carried out all performance evaluation tests, I wrote the paper and presented it at the conference. All authors contributed with suggestions during the implementation phase, and with feedback and comments for the paper.

[12] : I summarized the current state of the art in the field of GMPLS/OBS integration and presented a thorough analysis of the potential challenges in the process of the integration. I made several suggestions for protocol extensions and proposed different network architectures, based on different integration scenarios. I wrote the paper and presented it at the conference. All authors contributed with comments and feedback.

[13] : I worked together with the other authors on some of the work, presented in the paper. I provided feedback and comments for the paper.

My personal contributions in the papers under review are:

1 The original idea for the BGP extensions is from Ricardo Romeral, who is co-author of the paper. During a collaboration activity with him I implemented the BGP extensions. Furthermore, I identified

several problems with the algorithm and enhanced it to compute link-disjoint paths from end-to-end. The application scenarios for the BGP extensions were my idea as well. I designed and run all simulations related to the work and carried out the main analysis of the results. The work presented in this paper is the basis for Chapter 5.

2 This work is an extension of the work presented in [6]. I performed the design of the policies, carried out the simulations and wrote the paper. S. Ruepp participated in the review process and gave feedback for improvements.

3 This work is a continuation of the research, presented in [4]. I extended the set of restoration mechanisms based on different criteria (e.g. the load of the failed link). I performed the experiments and analyzed the results. All authors contributed to writing the paper.

4 I contributed with comments and feedback on the paper.

5 This paper is extension of paper [4]. I performed the additional experiments, analyzed the results and wrote the paper. All authors contributed with comments and suggestions.

Chapter 2

Optical Network Control Plane

2.1 Introduction

According to [14] a contemporary optical network can be divided in three planes: transport(data), management and control (see Fig. 2.1). Each plane carries out specific functions in order to provide and sustain connectivity between end points of the network. The transport plane (also referred to as Data Plane (DP)) is responsible for the physical transfer of data; the management plane is responsible for managing the network and its services (Operation and Maintenance functions, Billing, Inventory management, etc.); the Control Plane (CP) off-loads the traditional management plane from the complexity of the dynamic establishment and maintenance of connections in the network.

From the CP point of view optical networks have undergone significant development during the past 15 years by increasing the flexibility and the complexity of the supported functionalities. The main trends in this development are towards smaller supported granularity of the carried traffic and towards dynamic, fully distributed network operation [15]. In order to achieve the former goal, technologies such as fast Optical Channel Switching (OCS), Optical Burst Switching (OBS) and Optical Packet Switching (OPS) are being developed. Three main standardization bodies work towards achievement of the latter goal: the International Telecommunication Union - Telecommunication Standardization

Sector (ITU-T), the Internet Engineering Task Force (IETF) and the Optical Internetworking Forum (OIF). ITU-T has developed a reference architecture for Automatically Switched Optical Network (ASON) [16] which defines the main components in a dynamically reconfigurable optical transport network and the main interfaces between them: User-Network Interface (UNI), External Network-to-Network Interface (E-NNI) and Internal Network-to-Network Interface (I-NNI). IETF, on the other hand, has developed a set of protocols under a common framework for control of any type of switching technology called Generalized Multi-Protocol Label Switching (GMPLS) [17]. Under this framework several protocols cooperate in order to provide the needed functionality defined for dynamic connection provisioning regardless of the underlying networking technology. The OIF provides an inter-operability forum for matching the architectural requirements set by ITU-T and the protocols specified by IETF.

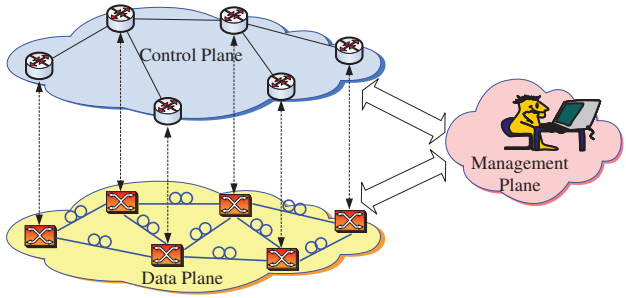


Figure 2.1: Network planes: Data, Control and Management.

This chapter covers the specifics of the different control plane solutions with special focus on multi-domain networking. Section 2.2 describes the GMPLS architecture and its operation, and outlines the main points of the ASON specification and the OIF implementation agreements. Section 2.3 focuses on different aspects, related to the area of multi-domain networking. Requirements for multi-domain connection provisioning and summary of the current state of the art in the area are provided. Section 2.4 gives a short description of the Border Gateway Protocol (BGP) needed for the discussions throughout the thesis. The general framework considered in this thesis is outlined in Section 2.5.

2.2 GMPLS and ASON

GMPLS and ASON are two control plane architectures defined from different standardization bodies, which are designed to facilitate dynamic service provisioning in transport networks. This section outlines the basic specifics of both solutions and how they relate to each other.

2.2.1 GMPLS

GMPLS is an IETF defined framework for distributed network control regardless of the underlaying data plane technology. It is an extension of the Multi-Protocol Label Switching (MPLS) framework [18], which was designed to provide Traffic Engineering (TE) in Internet Protocol (IP) networks. Under the GMPLS architecture, the network performs four main functions [14], which must work in cooperation in order to provide dynamic connection establishment between end points in the network:

- Neighbor discovery - allows all network elements to automatically discover their connectivity (i.e. their neighbors);
- Routing - finds a path to a specific destination. In GMPLS networks source routing is one of the main enablers for traffic engineering, thus the function of *path computation* is an integral part of the routing functionality;
- Signalling - determines the interaction between network elements in the process of connection establishment and management;
- Resource Management - allows network elements to perform local resource management.

The GMPLS framework defines a suite of IP-based protocols which are responsible for the main CP functions in the network (see Fig. 2.2):

- Routing function: performed by the Open Shortest Path First - Traffic Engineering (OSPF-TE) protocol [19] or the Intermediate System to Intermediate System - Traffic Engineering (IS-IS-TE) protocol [20];
- Signalling function: performed by the Resource ReserVation Protocol with TE extensions (RSVP-TE) [21];

- Link management: performed by the the Link Management Protocol (LMP) [22].

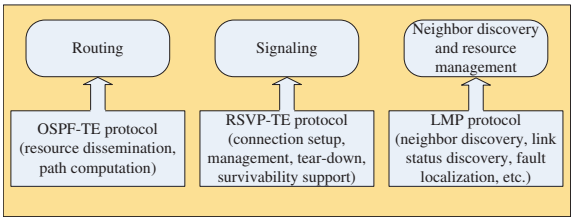


Figure 2.2: GMPLS functional blocks and protocols.

The established connections within a GMPLS network are called Label Switched Paths (LSPs) where, depending on the specific data plane technology, the label can indicate one of the following:

- Time slot - if the transport network is a Time Division Multiplexing (TDM) network such as SONET/SDH;
- Wavelength - if the transport network is a lambda-switched network such as a Wavelength Division Multiplexing (WDM) network;
- Port or Fiber - if the traffic is switched based on the port number or the fiber identifier (e.g. in a Photonic Cross-Connect (PXC)).

This means that the labels in GMPLS are directly associated with the physical resources. GMPLS also defines a strict hierarchical model, based on the supported data plane technologies (referred to as switching types). This allows tunneling of LSPs, similar to the MPLS label stacking. The GMPLS hierarchy is depicted on Fig. 2.3.

For operation in heterogeneous network environments GMPLS defines three service models for inter-operation between networks: peer, overlay and hybrid [23]. The main differences lie in the level of trust between the involved networks. The peer model is the so called *unified service model*, where the end-to-end LSP establishment (or service provisioning) is performed across different networks and there is a full visibility of the topology and the resources from all network elements. In the overlay model (also called *domain service model*), strict layering

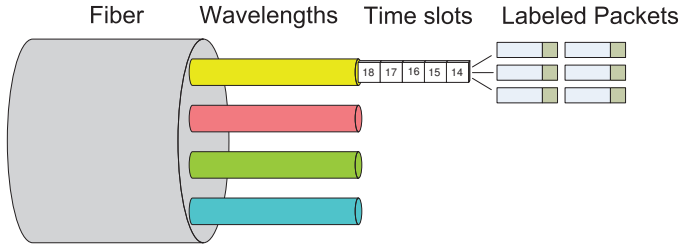


Figure 2.3: GMPLS hierarchy.

of the network is enforced and the communication between the layers is carried out via a well-defined interface. Upper-layer elements may request services from a lower-layer network, which in turn may request services from a lower-layer network. The relationships between layers are strictly client-server and no topological information is allowed to be distributed across the inter-layer interface. The hybrid model, also called *augmented model*, allows for some topology and resource information to be passed between network layers for improved network performance and TE. Limited and well-controlled information exchange is established between networks according to the level of trust.

2.2.2 ASON

ASON is a specification for dynamically reconfigurable optical transport network described in the ITU-T's recommendation G.8080/Y.1304 [16]. This recommendation gives the general control plane components and the architecture for dynamic service provisioning in optical networks. A key feature of ASON is the strict separation of the network elements and the definition of clear interfaces between them for communication and information exchange. Three interfaces, also called reference points, are defined (see Fig. 2.4):

- User-to-Network Interface (UNI)
- Internal Network-to-Network Interface (I-NNI)
- External Network-to-Network Interface (E-NNI)

Depending on the level of trust between the communicating elements in the network, different amounts of routing and signaling information are allowed to cross the reference points. The UNI is an untrusted interface for requesting services and no routing information is allowed to be disseminated. The I-NNI is an internal network interface between elements within one administrative domain. It is with the highest trust level and thus, any type of information is allowed to cross. The E-NNI can be with various levels of trust and the amount of information to be disseminated depends on the involved interconnected domains. In general, it is assumed that no topological or network state information is allowed to cross an E-NNI interface [24] when it connects different operators, and only limited (abstracted) topological information is allowed to cross if the E-NNI interface is between domains belonging to the same network provider.

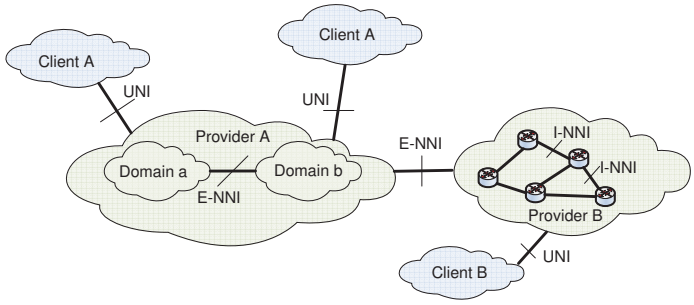


Figure 2.4: ASON architecture and reference points.

Another characteristic of the ASON architecture is the separation of the network(s) in routing areas and the creation of a routing hierarchy (see Fig. 2.5) [25]. Each routing area is controlled by a *routing controller* which only has knowledge of the topology and state information in its own subnetwork.

The ASON architecture is based on client-server relationships between routing areas, subnetworks and domains. It is protocol agnostic and requires clear independence of the intra-domain protocols applied in each network segment.

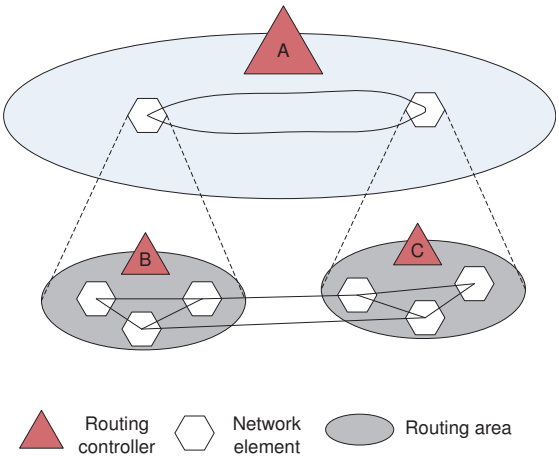


Figure 2.5: ASON routing hierarchy.

2.2.3 ASON vs. GMPLS

At a first sight, it might seem as if ASON and GMPLS are competing technologies. They both define functional specifications for dynamic connection establishment in connection-oriented networks. Nevertheless, they are quite different. The ASON architecture is a general specification which defines structural and functional components and generic interfaces between them with no specific focus on protocols. Moreover, it is defined only for optical networks. GMPLS, on the other hand, provides protocols which perform some of the specified functions within the ASON architecture. Furthermore, GMPLS is independent of the underlying technology, i.e. it is not specifically designed for optical networks. Based on the specifics of GMPLS and ASON, both solutions are considered complementary and there are working groups established within IETF, which focus on mapping the ASON requirements and the GMPLS protocol suite.

The OIF forum is an organization which brings together both solutions by developing implementation agreements which extend and modify the protocols from the GMPLS suite to be applicable across ASON interfaces. Both the routing and the signaling protocols are extended:

- UNI-related agreements: since no routing information is allowed to traverse the UNI interface, only signaling protocol extensions have been defined in [26], which are based on the extension of the RSVP-TE protocol.
- E-NNI-related agreements: RSVP-TE based E-NNI signaling specification is provided in [27] and OSPF-TE based E-NNI routing specification is provided in [28]. The routing specification is only for the intra-carrier scenario.
- I-NNI-related agreements: the OIF does not specify any protocol extensions or specifications related to the I-NNI interface.

Furthermore, special guidelines for interworking between ASON and GMPLS domains with respect to the signalling functionality are provided in [29].

2.3 Multi-domain networking

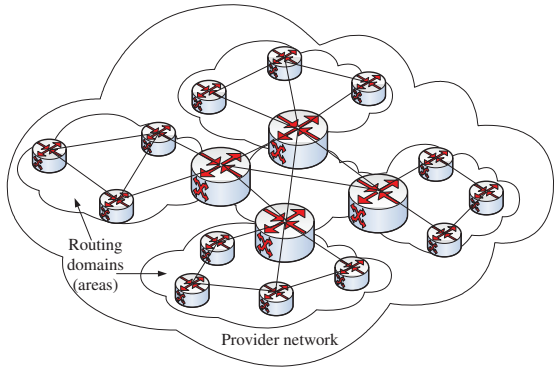
This section presents background information related to multi-domain networking, connection management, routing and path computation. Analysis of existing solutions for multi-domain routing in connection-oriented networks is presented along with the main requirements for such routing, defined by the standardization bodies in the field.

2.3.1 Multi-domain networks

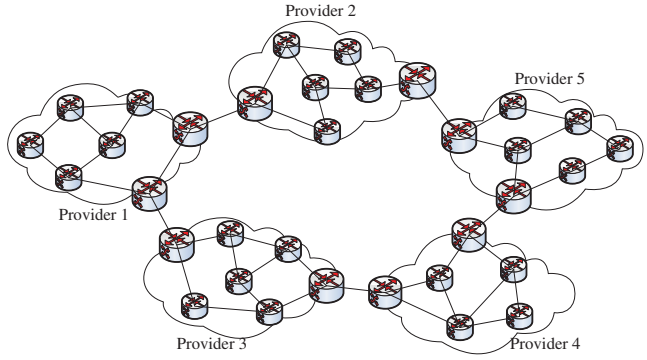
For scalability reasons and ease of management a network is often divided into smaller regions [14]. The division could be at the data plane and/or at the control plane level depending on the goal of the separation and the local management policies. The separate divisions are called sub-networks, areas or domains and the overall network is referred to as multi-domain or multi-area network. Interconnecting different Autonomous Systems (ASs)[1] also results in a multi-domain global network (e.g. the Internet). According to the IETF, a domain is considered to be *"any collection of network elements within a common sphere of address*

[1] A network under a common administration, applying a unified routing policy is referred to as Autonomous System.

management or path computational responsibility" [30]. According to the ASON specification, a domain represents *"a collection of entities that are grouped for a particular purpose"* [31]. Based on this, dividing a domain internally into different routing areas makes it a multi-domain network. The same holds for any interconnection of different service providers, which also forms a multi-domain network. Different multi-domain network types are depicted on Fig. 2.6.



(a) Multi-domain single provider - division for routing scalability.



(b) Multi-provider network - division based on ownership and administrative responsibility.

Figure 2.6: Examples of multi-domain networks.

Regardless of the type of the multi-domain network, one common principle can be outlined - there is a strict information preservation

requirement within domain boundaries. The limitation of information distribution beyond domain borders may be due to scalability, confidentiality and/or security reasons. The extent of the limitation depends on the specific type of the multi-domain environment. If domains are routing areas within one AS, the information preservation is done mainly for scalability reasons and the trust between the domains is high. If the domains are separate network operators the separation of the network is based on policy and the trust between the domains is very low, thus the information preservation is very high. This differentiation is very important when the specifications and the requirements for inter-domain connection provisioning are laid out. The level of information filtering at the borders of the domains poses very stringent requirements on the connection management procedures and the applied algorithms and mechanisms in the network as a whole.

2.3.2 Connection management

The main purpose of a telecommunication network is to transport data from one point (source) to another (destination). In GMPLS/ASON networks this is done via pre-established connections, i.e. no client data is sent in the network before a connection is established between the source and the destination points of the communication. In GMPLS networks these connections are called Label Switched Paths and are managed using the RSVP-TE protocol. The protocol is responsible for notifying all nodes along the path of a connection for the incoming traffic, where each node is responsible for allocating the needed resources for the connection. Fig. 2.7 illustrates the process in a single-domain environment. The same protocol is used also to tear down, modify or recover a connection.

In a multi-domain environment the process of establishing a connection is the same. One modification in the application of the RSVP-TE protocol is the filtering of some information at the border nodes, such as the Recorded Route Object (RRO) [32], which may potentially disclose sensitive topological information [30]. Three different inter-domain signaling methods for establishing LSPs across domain borders are specified in [30] (see Fig. 2.8):

- Continuous [21,33] - one RSVP-TE session manages the connection request;

- Nesting [34] - a higher-level LSP, established separately, is used to nest the current LSP request. Nodes in the nesting domain do not react to the RSVP-TE messages for the current LSP request;
- Stitching [35] - separate RSVP-TE sessions are established per LSP segment within each domain.

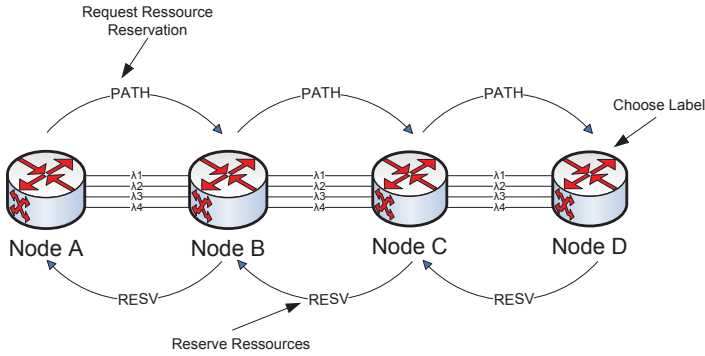


Figure 2.7: Resource reservation with RSVP-TE in GMPLS networks.

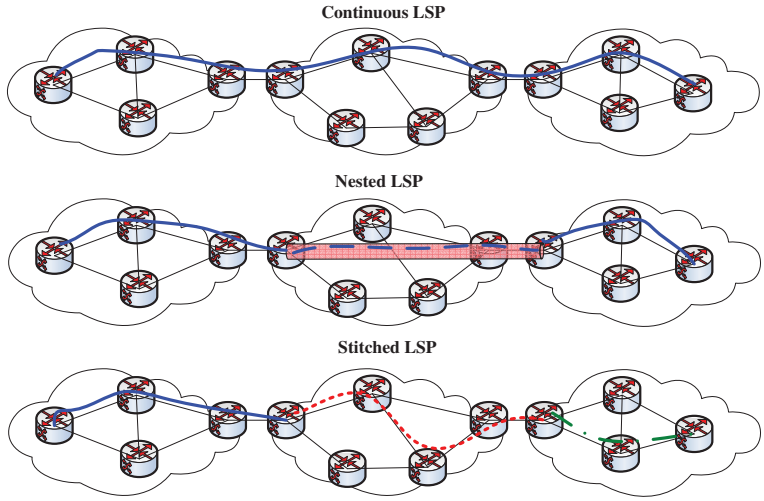


Figure 2.8: Multi-domain LSP establishment methods.

2.3.3 Routing and path computation

One of the fundamental functional blocks in the CP of a network is the routing. Routing relates to the process of finding a path from one node in the network to another. Depending on the applied network technology, different routing paradigms can be used. In general, there are three main routing paradigms - hop-by-hop routing, source routing and hierarchical routing [14]. Hop-by-hop routing is typical for connectionless networks such as the Internet, where each data unit (i.e. IP packet) is routed independently and routing decisions are taken at every hop (node) along the path of a packet. Source routing is typical for connection-oriented networks where the source node specifies explicitly the path of a packet by indicating a list of nodes to be traversed. This gives a higher level of control on the head-end of a connection on the path. A mixture of these two routing paradigms is also possible as specified in [17]. In large-scale networks hierarchical routing is often employed for scalability reasons. In hierarchical routing, the network is separated in areas, where each area has one dedicated routing controller responsible for the routing within the area. The routing controllers can also be grouped in separate areas, forming a higher level of the hierarchy. Such grouping and separation can be done recursively, forming multiple hierarchical levels. Routing information can be distributed up or down the hierarchy depending on the network configuration and local policies.

The process of path computation is paramount for TE and QoS provisioning [36]. This function provides a sequence of nodes to be traversed by a connection (i.e. LSP), based on some requirements with respect to the quality of the path. Typical requirements are available bandwidth, total delay, cost, etc. There are many algorithms for constrained-based path computation (also referred to as QoS routing), all of which aim at satisfying different objectives [14].

Depending on the applied routing paradigm, different protocols can be employed in a network to support the functionality. For example the BGP protocol exclusively supports the hop-by-hop routing paradigm [37], even though it provides sequence of ASs which could be considered as a "loose" hop path towards a destination. The OSPF protocol, on the other hand, supports both routing paradigms.

Path computation is typically performed based on a graph representation of the network topology. This implies that in order to perform

constrained-based path computation, the entity responsible for this function must have a view of the topology (i.e. of the nodes and links in the network and how they are interconnected). Furthermore, some link and node parameters (such as available bandwidth, delay or processing capabilities) are also required [14]. This information is disseminated within a domain using the intra-domain routing protocol, which typically is a link-state protocol. All algorithms for constrained-based path computation require topology dissemination for deriving a full or abstracted view of the network.

In multi-domain networks though, the process of path-computation can be severely obstructed due to the applied routing protocol. Currently, in the Internet, there is only one inter-AS routing protocol - the BGP. This protocol does not disseminate link-state information and thus, cross-domain path computation is hindered. If a multi-area network is considered, which is also a multi-domain network (see the beginning of this section), a link-state protocol can be used for disseminating abstracted topological information per area. In this case, a network graph is available but the performed path computation cannot guarantee the quality of the computed path due to the loss of information during the abstraction process [14,36].

2.3.4 Requirements for multi-domain routing

As presented in the previous subsection, the multi-domain environment poses specific challenges to the routing and the path-computation functions. Both standardization bodies, IETF and ITU-T, define strict requirements for routing in multi-domain environments based on their respective control plane technologies - GMPLS and ASON. Furthermore, in [24] specific requirements for multi-domain service provisioning from the carriers' point of view are detailed.

Some of the IETF's requirements for inter-domain GMPLS networking, specified in [38], are:

1. Reachability information exchange beyond domain borders includes: node ID, interface address and interface ID.

2. TE information exchange beyond domain borders includes: interface switching capability (including free, reservable, unreserved

bandwidth), bandwidth encoding type, Shared Risk Link Group (SRLG), and protection type.

Among the ITU-T's requirements for multi-domain routing, specified in [25][2], are the following:

1. The inter-domain routing protocol must be independent of the intra-domain routing protocol.

2. The routing protocol must support hierarchical routing and must be able to disseminate an abstracted view of each domain.

3. The information exchange between routing domains must be subject to policy constraints.

The requirements from different network operators outlined in [24] are:

1. The inter-domain routing protocol must support policy-based routing.

2. No information containing network topology details must be allowed to cross domain boundaries.

3. Only reachability information, next routing hop and service capability information should be exchanged over external interfaces.

4. The routing protocol should be able to disseminate abstracted domain topology information.

From the presented requirements two main conclusions can be darwn. First, the requirement for information preservation is very strong among the network operators. Second, in order to support constrained-based path computation and TE across domains, some link-state information is required to be shared between all domains. These two requirements pose a very challenging trade-off. Link-state information is needed, but network operators are strongly opposing any dissemination of domain topology and state information. Moreover, from the providers' perspective policy enforcing functionality is even more fundamental than TE provisioning aided by constrained-path computation. Thus, protocols which provide policy control are preferable [24].

[2]Note that these requirements are for multi-domain, *intra-carrier* networks.

2.3.5 Existing solutions for routing - overview and analysis

Given the control plane frameworks [16, 17] and the respective routing requirements [25,38], both standardization bodies work towards the development of a suitable routing framework/protocol. Currently, the most promising solutions are the OIF's E-NNI specification, which is based on the ASON architecture, and the IETF's Path Computation Element (PCE) architecture [36]. Furthermore, some researchers work in the area of applying BGP for multi-domain optical routing even though it is considered an ineligible protocol for that purpose by some authors [39, 40].

Next, the different solutions are briefly described, their strongest characteristics and some unsolved issues are discussed.

E-NNI

The OIF has specifies protocol extension to the OSPF-TE protocol in [28] in support of multi-domain ASON routing, following the specifications outlined in [25]. Since the solution is based on OSPF-TE, it is clearly a link-state solution. A routing hierarchy is created and each domain is represented by an abstracted topology with virtual links between domain border nodes. The higher level abstracted topology is disseminated in standard manner with the OSPF-TE protocol and a graph representation of the abstracted multi-domain network topology is obtained by all routing controllers. Then, standard path computation techniques can be applied. Fig. 2.9 illustrates the operation. All routing controllers obtain the aggregated topology (Level 1) and can compute paths across domains.

The strongest feature of this approach is the compliance with the ITU-T's specification and the application of standard protocols and methods for path computation. The biggest drawback is the fact that the graph, on which path computation is performed, is aggregated. Inaccuracy from aggregation leads to suboptimal solutions. The inaccuracy may be based not only on the aggregation method but also due to protocol re-convergence delays. Furthermore, this approach reveals network state information per domain which is heavily opposed by network operators. The E-NNI specification is itself defined only for intra-carrier application. In inter-carrier scenarios (i.e. inter-AS networking), the re-

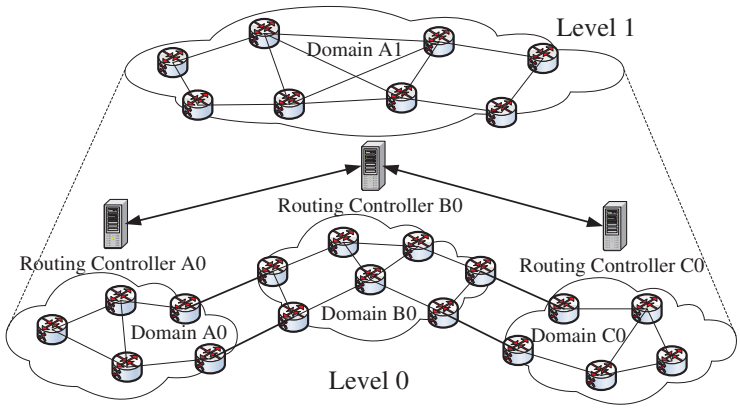


Figure 2.9: E-NNI multi-domain (intra-carrier) routing operation.

quirements for topology and state information preservation are much more stringent [24].

PCE

The IETF is currently developing an extended framework for path computation in large-scale and multi-domain/area networks. The PCE architecture [36] is a complete framework for constrained-based path computation for TE purposes in connection-oriented networks. It is applicable in both single- and multi-domain scenarios. As a concept, it is very close to the E-NNI architecture - there is one (or several) routing controllers per domain, called Path Computation Elements (PCEs). These entities are solely responsible for computing paths in the network. Each node requesting a path computation is called Path Computation Client (PCC). There is a special protocol - Path Computation Element Communication Protocol (PCEP) [41], which is designed for communication between PCCs and PCEs. The PCEs have access to a topology database which contains full topology and state information related to the domain to which the element belongs. Based on this information, the PCEs are capable of computing very complex constrained-based paths based on the requirements given by the PCCs. Fig. 2.10 illustrates the operation of the PCE element.

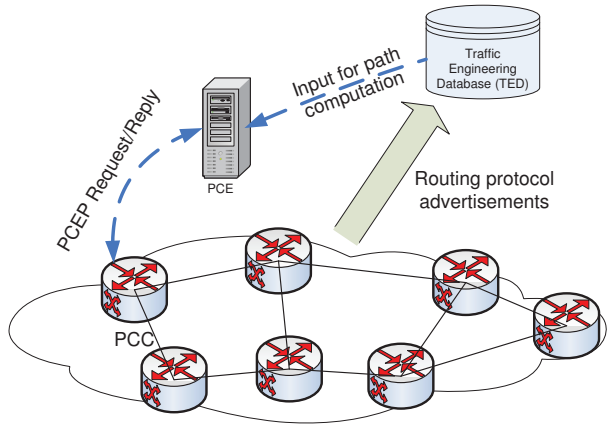
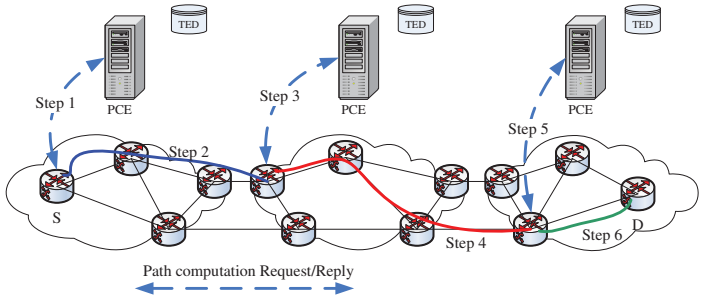


Figure 2.10: Path computation with PCE.

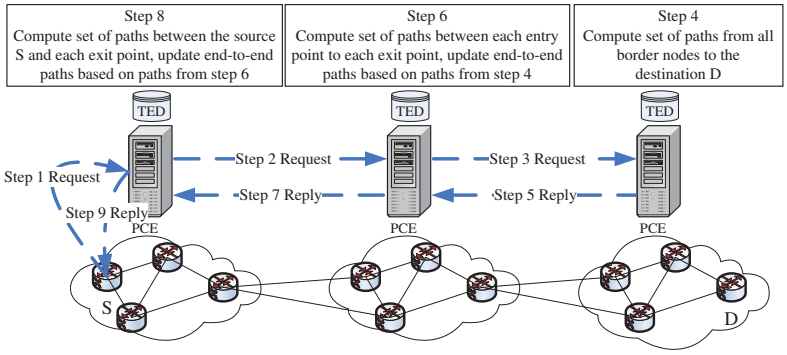
Two multi-domain path computation techniques are defined within the PCE architecture: the Per-Domain Path Computation (PDPC) [42] and the Backward-Recursive Path Computation (BRPC) [43]. In the PDPC technique, no inter-PCE communication is required between the PCEs of different domains. The BRPC on the other hand, requires co-operation between the PCE elements in the different domains. Fig. 2.11 illustrates the techniques.

The advantages of the PCE architecture are numerous. It provides flexibility and scalability in large-scale networks. It supports constrained-based path computation for TE and at the same time provides full domain privacy. No topological nor network state information is disseminated between PCE elements. Furthermore, the PCE architecture supports various policy enforcing techniques, which is one of the main requirements of the network operators as outlined in [24]. Even though the original PCE architecture does not explicitly define support for hierarchical routing [36], there is ongoing research work on the applicability of the PCE in hierarchical structures [44]. Such approach fits very well within the multi-domain routing requirements specified for ASON networks and the requirements identified by network operators.

Nevertheless, several challenges with the application of the PCE architecture can be outlined. First, the scalability of the approach is still



(a) Per Domain Path Computation.



(b) Backward Recursive Path computation.

Figure 2.11: Multi-domain path computation within the PCE architecture.

under investigation by numerous research groups. The authors of [36] state that "*PCE is not considered to be a solution that is applicable to the entire Internet.*", but the exact scalability limitation is still under investigation. The second challenge lies in the reachability dissemination and in the identification of valid AS-paths (or domain sequences). The PCE architecture assumes manual configuration of AS-paths or paths, derived from BGP advertisements. Thus, most of the work carried out in this area focuses on flat linear topologies (i.e. only one AS-path is assumed). However, having only one path per destination limits the efficiency of the path computation, since the chosen AS-path may not be the optimal one. Last, the PCE architecture is a relatively new proposal and its level of maturity is very low. Currently, mostly proprietary

implementations of PCE can be found within single provider networks for inter-area MPLS networking. Very few implementations exist, that involve many providers interconnected in complex mesh topologies [45], especially for GMPLS networks. Furthermore, all implementations use predefined linear domain sequences, i.e. the efficiency of the PCE has not been tested in multi-domain mesh networks taking full advantage of the connectivity between the domains. From a test-bed perspective not many large-scale PCE test-beds can be found (e.g. the ADRENALINE [46], the KDDI-PCE [47]). There exist even less inter-operability demonstrations, based on PCE architecture, which is mainly due to inconsistencies in the implementations and the inclusion of proprietary defined shortcuts and functions.

OBGP

Optical Border Gateway Protocol (OBGP) is a protocol extension, designed for routing and light-path provisioning in optical networks [48]. More specifically, it provides customer control over the light-path provisioning through an optical cloud by allowing customers to control the provider's cross-connects. This solution is a proprietary designed BGP extension by Canarie, Inc. [49]. Some of the features of the solution are the support for light-path trading, facilitating customer's control over the peering relations with other customers, and allowing customers to create their own proprietary topologies throughout the provider's optical cloud.

In essence, the OBGP solution is facilitating optical peering between client networks connected to one optical cloud (provider). Client networks are allowed to create short-range light-paths within the cloud and take over the control of the provider's cross-connects by making some of the ports part of their own network (see Fig. 2.12). It requires Optical Cross-Connects (OXCs) to be configured as virtual BGP peers and necessitates the usage of centralized entities called Light-path Route Arbiters which manage the light-path establishment. It is obvious that this solution does not fit either to the GMPLS LSP establishment framework, since it uses the OBGP protocol itself to signal the light-paths [50], or to the ASON architecture. Nevertheless, the solution is implemented and operational within Canarie's CA*net4 network in Canada.

Clearly, the inter-operability between OBGP and GMPLS/ASON networks would be hindered due to the differences in the applied net-

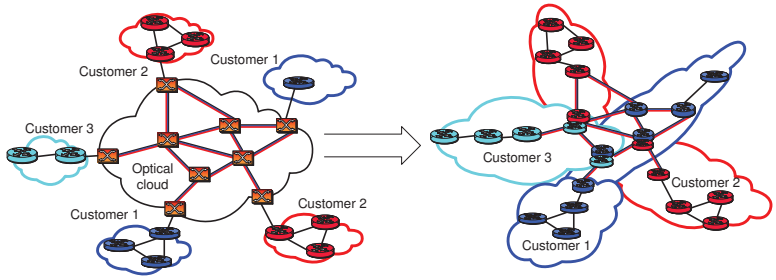


Figure 2.12: Optical BGP operation: the provider delegates the management of the cross-connects to the clients, which create peering relations within the cloud.

working principles. Giving customers control over the providers infrastructure goes strongly against the standard telecommunication principles. Requirements from providers strictly prohibit topology and network state information exchange with clients [16, 24], let alone giving them control over the network's elements. The basic GMPLS networking principle divides signaling from routing functionalities, whereas in OBGP these are carried out with the same protocol. As a result, the OBGP protocol has remained a proprietary solution, without being standardized.

2.3.6 Resilient multi-domain networks

Along with the requirement for multi-domain connection provisioning comes the necessity for multi-domain survivability. Transport networks carry huge amounts of information and failures greatly affect the network performance and the operators' revenue. While the area of intra-domain resilience has been extensively studied throughout the years, the multi-domain resilience area has received very little attention. There are many unsolved issues. Starting from the problem of limited topology visibility, which hinders computation of disjoint paths across multiple domains, going through the issues of failure notification and responsibility for recovery, and ending up with the need for differentiated failure handling frameworks - the area of multi-domain resilience offers many interesting challenges.

In this thesis, some of the pending problems in the area are analyzed

and novel approaches for failure notification and handling are proposed for multi-domain GMPLS networks. The discussions throughout Chapter 5 and Chapter 6 focus on the issues of providing disjoint paths for failure recovery and for enhanced network performance under failure conditions when the BGP protocol is used as a multi-domain routing protocol. By means of analysis and simulation results, interesting observations regarding the dependencies between different topological parameters and the efficiency of different recovery techniques, are made. A thorough review of the state of the art in the area is given at the beginning of each chapter.

2.3.7 Multi-layer network integration

Network layering is a standard practice for reducing complexity. In telecommunication networks, the different supported switching technologies are often referred to as layers. A layer network is defined by [31] as "*the complete set of access groups of the same type which may be associated for the purpose of transferring information*". A multi-layer network, according to [51], is a network which uses a unified GMPLS control plane and comprises of interconnected elements which perform different types of switching functionality (e.g. packet and TDM switching performed in the same transport network). The concept of layering is thus orthogonal to the concept of domain partitioning [31].

Nevertheless, interconnecting domains employing different switching capabilities and using autonomous CPs is an integral part of the multi-domain inter-networking problem. Thus, part of the research work in this thesis is focused on integrating networks that employ different switching technologies. In particular, the problem of incorporating the GMPLS CP within a novel switching technology such as the OBS, is investigated. Since this research area is considerably new, an extensive analysis of the state of the art is presented in Chapter 7, along with a broad discussion on different aspects and challenges of such multi-layer integration.

2.4 The Border Gateway Protocol

One of the main contributions of this thesis is the extension and modification of the BGP protocol for multi-domain GMPLS networking. Thus,

it is important to present the basics of BGP, its operation and main characteristics. This section focuses on general BGP protocol description and outlines the current methods and tools for TE with BGP. Since BGP is explicitly designed for interconnection of ASs in the Internet, all described TE methods are focused on TE within the context of the Internet (i.e. packet-switched networks).

2.4.1 Standard BGP - protocol description

BGP is the inter-domain routing protocol for use in the Internet. It is a path-vector protocol, i.e. it disseminates reachable destinations (called Network Layer Reachability Information (NLRI)) along with the path (sequence of ASs) to be traversed to reach them. One of the main design principles of BGP is to limit the amount of disseminated information in order to achieve scalability. Thus, only reachable IP prefixes (often aggregated) along with characteristics of the path to reach a prefix are disseminated via UPDATE messages. The protocol also gives numerous options for policy enforcement which aid the network operators in managing their networks. The operation of the protocol can be divided in two: IP prefix dissemination and path selection. Both parts are controlled by local policies.

There are two types of BGP sessions - i-BGP (internal BGP) between BGP speakers from the same domain and e-BGP (external BGP) between speakers from neighboring domains. BGP nodes communicate via pre-established TCP connections, where all BGP nodes within a domain are interconnected, i.e. there is a full mesh connectivity. Whenever a BGP speaker receives a path towards a destination, the previous advertisement, regarding the destination at hand, is implicitly withdrawn. Then, a decision process is run, based on the received path attributes, in order to decide whether this path should be installed in the common IP routing table. Before the path evaluation process begins, local policies are applied to the path and its attributes. The path selection process is based on the following criteria [37] (see also Fig. 2.13(a)):

1. Highest Local Preference
2. Shortest AS_PATH attribute
3. Lowest Origin type

4. Lowest Multi-Exit-Discriminator (MED) value
5. e-BGP learned paths over i-BGP learned ones
6. Lowest Interior Gateway Protocol (IGP) cost towards the BGP next hop
7. Lowest router identifier
8. Lowest IP address neighbor

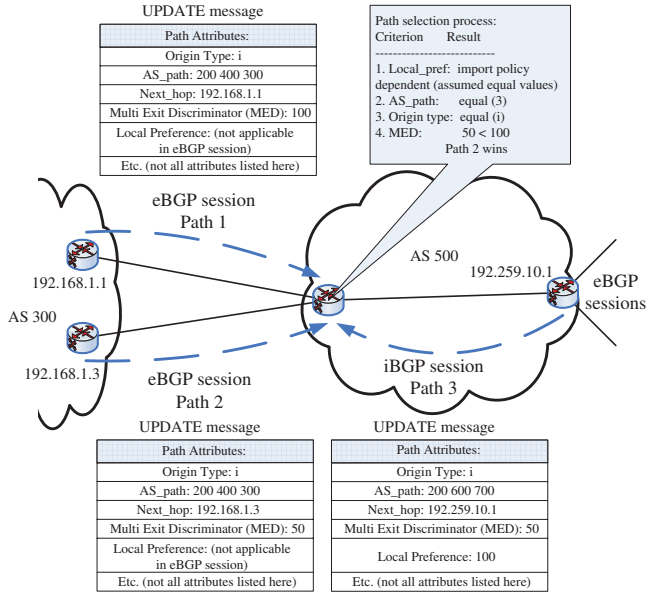
After a path has been chosen, it is installed in the IP routing table and depending on local policies, is disseminated further to other BGP neighbors. Fig. 2.13(b) illustrates the general process of BGP.

BGP is primarily concerned with finding the shortest path towards a destination ("shortest" in terms of number of ASs on the way). Most of the decision criteria are related to intra-domain traffic engineering, whereas the last two criteria do not have any TE significance and are used solely for tie-breaking. The decision process provides the network operators with a powerful tool to control the network's resources. Nevertheless, the applied policies in each domain are only locally valid and a global coordination between several domains very rarely exists. Thus, one of the biggest challenges BGP meets today is suboptimal path computations due to the lack of global coordination and variances between different protocol implementations [52].

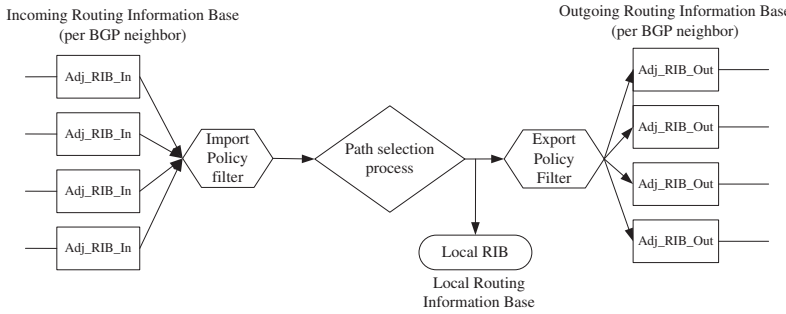
Following from the BGP specific operation, it is clear that BGP does not support path computation. Instead, BGP performs path selection. This fundamental feature of the protocol is very important and will be thoroughly discussed throughout this thesis (see Chapter 3 and Chapter 4). Furthermore, the BGP extensions and modifications proposed in this thesis are closely related to this particular characteristic.

2.4.2 Existing work on Traffic Engineering with BGP

BGP is a well-established protocol and has been used in the Internet for over 30 years. The development of the protocol has followed the need for enhanced functionality and improved scalability and stability during the years. The current version of the protocol (BGP-4) provides several options for inter-domain TE in the Internet. Most of the applied methods



(a) BGP path selection example.



(b) BGP operation.

Figure 2.13: BGP path selection and general operation.

have only local significance, i.e. they operate only between two domains which have particular bi-lateral agreements specified beforehand. Two main local objectives can be met by the existing BGP TE methods:

- Control of outgoing traffic;

- Control of incoming traffic.

Each of these can be accomplished by one of the following TE methods:

- Local Preference manipulation: controls the outgoing traffic from a multi-homed domain [53] based on local policy;

- IGP weight: controls the outgoing traffic using the $6th$ decision criteria of the path selection process, which chooses the outgoing router based on the shortest path from the BGP speaker to its neighbors;

- Multi-Exit Discriminator (MED) attribute: controls the incoming traffic by indicating to a neighbor the preferable entry point in the domain. This can be ignored from neighbors so efficiency is guaranteed only after coordination with the neighbor [54];

- AS_PATH prepending: controls the incoming traffic by inflating the length of the AS_PATH attribute on less preferable incoming links. Such technique can be very efficient between two domains since the AS_PATH length is the second decision criteria, but might have unexpected consequences globally [54];

- Prefix splitting: controls the incoming traffic by announcing a more specific prefix via the most preferred incoming link. This technique relies on the fact that a router would always prefer the longest matching prefix and has only local inter-domain influence. It is applicable only between two domains with at least two inter-domain links between them.

- Communities: controls the incoming traffic by conveying additional information about the path in a Community attribute [55], which influences the decision process in the neighboring BGP router. The community values are typically used between customer networks and their transit providers. They facilitate finer control of the incoming traffic for the customer networks and require bi-lateral agreements in order to coordinate the values of the Community attributes and the corresponding local policies. Three Community values are being standardized [55]:

- NO_EXPORT: indicating that the received route must not be exported to peers outside the BGP confederation;
- NO_ADVERTISE: indicating that the received route must not be advertised to other BGP peers;
- NO_EXPORT_SUBCONFED: indicating that the received route must not be advertised to external BGP peers, even though they are in the same confederation.

Depending on the bi-lateral agreements between two domains, many different Community values could be defined and applied, e.g. changing the local preference of the router, receiving the path [56], or enforcing AS_PATH prepending.

All presented techniques are rather primitive and are based on manipulating the BGP settings and local policies for controlling the ingoing and/or outgoing traffic. Thus, global objectives are almost impossible to enforce since all ASs need to coordinate the manipulation of certain local policies. This is not impossible, but is very difficult to achieve because such manipulations are static, require manual interventions (BGP is typically manually configured) and dynamic optimizations is not supported. Furthermore, all presented techniques are related to control of reachability information dissemination, not to engineering end-to-end paths based on TE parameters.

2.5 Scope of the thesis

This section outlines the starting point for the discussions carried out in this thesis. Some basic assumptions, clarifications and the general context in which the research subjects have been discussed is presented.

First, in this work it is assumed that the multi-domain network consists of separate ASs where for each AS the CP and DP topologies coincide. Furthermore, each AS is a GMPLS controlled lambda-switched network (or WDM network), i.e. the TE metrics of interest are wavelength availability and/or hop-count. The only exception is related to the discussions in Chapter 7, where the focus is on multi-layer integration rather than multi-domain TE. Unlike most of the existing work in the field, which considers linear multi-domain topologies, in this work a

general mesh multi-domain topology is used. Furthermore it is assumed that all domains participate in a cooperative way in order to provide cross-domain TE. The relationships between the domains are transit, i.e. each domain offers transit services to its neighbors and there are no client-server relationships.

The terms *optical transport network*, *network provider*, *GMPLS network* and *carrier network* are used interchangeably and relate to an AS (or simply a domain). The main focus of the discussions is on optical (connection-oriented, lambda-switched) networks, but some of the proposals made throughout the thesis are applicable in any GMPLS switching technology and for heterogenous networks.

The proposals outlined in this work are not intended to be applied in the global Internet. Instead, the application scenario is for a group of cooperating transport networks as in the PCE architecture [36], i.e. they are intended to be applied within a group of networks with common agreements. This is based on the observation that applying simple bilateral agreements between domains is not enough in order to support end-to-end QoS provisioning across multiple domains.

Third, in the presented work, the terms Traffic Engineering and Quality of Service (QoS) provisioning are used interchangeably, even though there is a slight difference between them. According to [14] TE relates to "*measurement, modeling, characterization and control of traffic to achieve specific performance objective*". QoS provisioning relates to the process of establishing paths, which meet certain criteria, most often specified from the client. Thus, the process of TE does not necessarily lead to QoS provisioning. For example, a path optimizing resource utilization might lead to longer delay for establishing the path and for transmitting client traffic. Thus, the process of TE in this case is related to the provider's objectives, not the client's requests. Nevertheless, the two processes are tightly bound since QoS provisioning is achieved by applying Traffic Engineering methods.

Chapter 3

Supporting TE with BGP in Multi-Domain GMPLS Networks

3.1 Introduction

The focus of this chapter is on enhancing Traffic Engineering (TE) in multi-domain optical networks by extending and modifying the Border Gateway Protocol (BGP). First, in Section 3.2 the motivation behind using BGP for optical multi-domain routing and the considered routing model are discussed. The section describes the possible problems in applying standard BGP in GMPLS networks. The design and implementation of a TE extension is presented in Section 3.3. Different aspects of the efficiency of the proposed TE-attribute are illustrated by means of simulations in Section 3.4. Section 3.5 concludes the chapter.

3.2 Optical multi-domain routing with BGP

This section presents the motivation for applying BGP in GMPLS networks and outlines the main differences between the routing paradigms applied in connection-oriented (optical) and packet-switched (Internet)

networks. Then, the routing model[1] is outlined and the main challenges for the applicability of the BGP protocol are analyzed.

3.2.1 Motivation

In Chapter 2 different solutions for multi-domain routing in GMPLS networks are presented. The ASON-compliant E-NNI specification for using OSPF-TE across domain borders is defined for domains under the same administrative entity (one carrier), since OSPF-TE runs only between areas belonging to the same provider due to privacy considerations. Other existing hierarchical routing protocols such as Private Network-to-Network Interface (PNNI) [57] may comply with the requirement for hierarchical routing but do not comply with the requirement for independence of the routing protocols employed at the different levels of the hierarchy. BGP is the only true inter-provider routing protocol [58], but since it is a path-vector protocol, which does not support any type of TE information dissemination or link-state dissemination, it is regarded as an ineligible candidate for GMPLS multi-domain routing [39,40]. Nevertheless, it has strong features (see Section 2.4) which make it the de-facto standard for multi-domain routing in the Internet today. Being both deployed world-wide and very well-known, BGP can potentially be used for routing in other types of networks which employ an IP-centric control plane. The experience, of almost 30 years, with this protocol gives great insight into the processes of interconnecting domains with strong privacy protection requirements, by enabling global interworking. The policy enforcement and the proven scalability are two of the merits which advocate consideration of BGP as part of the multi-domain routing and provisioning framework for future transport networks. Moreover, BGP complies with some of the requirements for multi-domain TE in optical networks [25], such as independence of the multi-domain routing protocol from the intra-domain routing protocol, and strong policy support capabilities.

[1] This routing model is used for the discussions in all chapters throughout the thesis, excluding Chapter 7.

3.2.2 Routing model

Some of the existing solutions for multi-domain routing in optical networks, presented in Chapter 2, are based on intra-domain approaches and mechanisms applied on an inter-domain level by abstracting the domains to a level of a node or a virtual mesh [16, 59]. This approach is understandable, considering the specifics of the LSP provisioning process. The need for strict path computation and QoS guarantees, in order for a connection to meet Service Level Agreement (SLA) requirements, necessitates link-state dissemination between domains. These approaches stretch the well-known intra-domain algorithms and methods across domains and seek to achieve end-to-end optimality.

The routing model considered in this work, on the other hand, is closer to the typical Internet routing model, where BGP forms a second level of routing hierarchy, providing higher-level AS-paths augmented with global TE metrics. An independent intra-domain routing protocol is responsible for the routing within each domain. Under this model, BGP is not used for path computation, but only for high-level path selection. Fig. 3.1 presents a possible implementation where the standard BGP is used for higher-level (inter-domain) routing and OSPF-TE for low-level (intra-domain) routing. Border nodes obtain only the next-hop towards a destination via BGP, the segments between ingress and egress border nodes are computed either by OSPF-TE explicitly or a standard hop-by-hop routing is used. RSVP-TE messages carry partial Extended Route Object (ERO) per segment, provided by the routing protocols.

Two types of routing can be applied within each level of the considered hierarchy: hop-by-hop and source-routing. Hop-by-hop is the typical routing for the best-effort Internet, whereas source-routing is typically used for Quality of Service (QoS) provisioning and TE (e.g. in GMPLS networks) since it provides explicit control of the source node over the whole path. Considering the requirements for routing flexibility and support of different provisioning paradigms [30], the routing model proposed here can support different levels of traffic engineering capabilities, listed in Table 3.1.

For the models with hop-by-hop intra-domain routing, the only option for TE is to employ high-level TE by applying standard or novel BGP TE methods. Using standard BGP methods (see Section 2.4) is not dynamic enough and is based on local policy enforcement and bi-

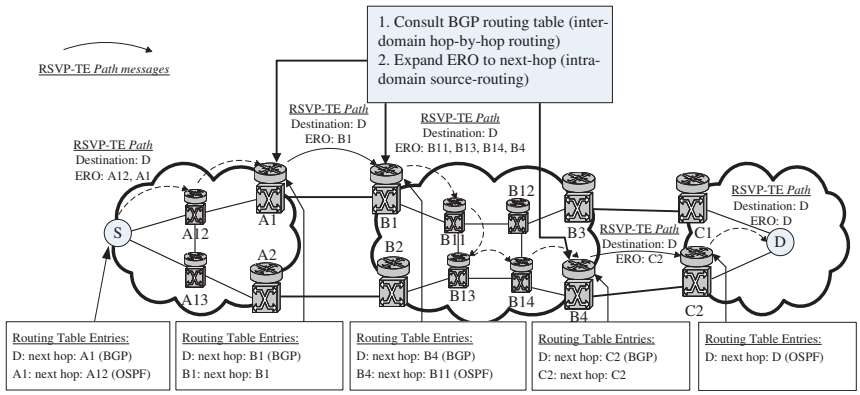


Figure 3.1: Two-level hierarchical routing model with BGP for multi-domain GMPLS connection provisioning.

Inter-domain routing	Intra-domain routing	Inter-/Intra- domain Traffic Engineering capabilities
hop-by-hop	hop-by-hop	Limited intra-domain TE, Static inter-domain TE based on standard BGP methods (MED, Communities), no global coordination
hop-by-hop	source-routing	Dynamic intra-domain TE, Static inter-domain TE, based on standard BGP methods (MED, Communities), no global coordination
source-routing	hop-by-hop	Limited intra-domain TE, Dynamic inter-domain TE based on explicit paths, computed at the higher level of the routing hierarchy
source-routing	source-routing	Dynamic TE at all levels of the routing hierarchy

Table 3.1: Levels of TE capabilities based on the applied routing paradigm.

lateral agreements between domains. Furthermore, global coordination is difficult to achieve, since it needs to be included in the inter-provider agreements between each two interconnected domains. Using novel me-

thods explicitly designed for optical networking (e.g. the proposals in Chapter 4 and Chapter 5), can significantly improve the LSP provisioning performance.

The work presented in this chapter is focused on novel methods for high-level TE within the outlined routing model. Applying the BGP modifications proposed here provides flexibility in the LSP provisioning process by supporting different path computation mechanisms. The considered routing model is compliant with both multi-domain path computation techniques specified by IETF [42, 43]. Since both techniques require a pre-defined sequence of Autonomous Systems (ASs) to the destination, applying the proposals outlined in this chapter can potentially enhance the operation of these techniques by providing them with AS-paths with better TE state at the time of LSP request.

The scope of the disseminated routing information is limited only to the infrastructure of the transport networks, i.e. a typical overlay model is considered [17]. The focus is on managing and controlling the resources of the optical transport network, which could consist of domains employing different transport technologies, and the disseminated reachability information is not the overlay client networks, but rather the Edge nodes of the considered optical domain. Fig. 3.2 presents the scope of the routing and provisioning model. This model is a general one, where the transport domains are not in client-server relationships. Instead, all domains provide transit service for all their neighbors.

Considering the outlined scope of the BGP routing protocol, the entities involved in the BGP communication (see Section 2.4) are the border nodes interconnecting the transport domains and the edge nodes communicating with the client networks. In this way, the head-ends of the Label Switched Paths (LSPs) have access to the multi-domain connectivity table and can choose directly end-to-end cross-domain paths.

3.2.3 Optical versus Internet routing principles

When designing the needed modification in the BGP protocol it is important to analyze the differences between the routing paradigms used in the Internet, for which BGP was initially designed, and in GMPLS networks, and to outline the resulting implications for the adoption of the BGP protocol in the optical environment. The main differences are [58]:

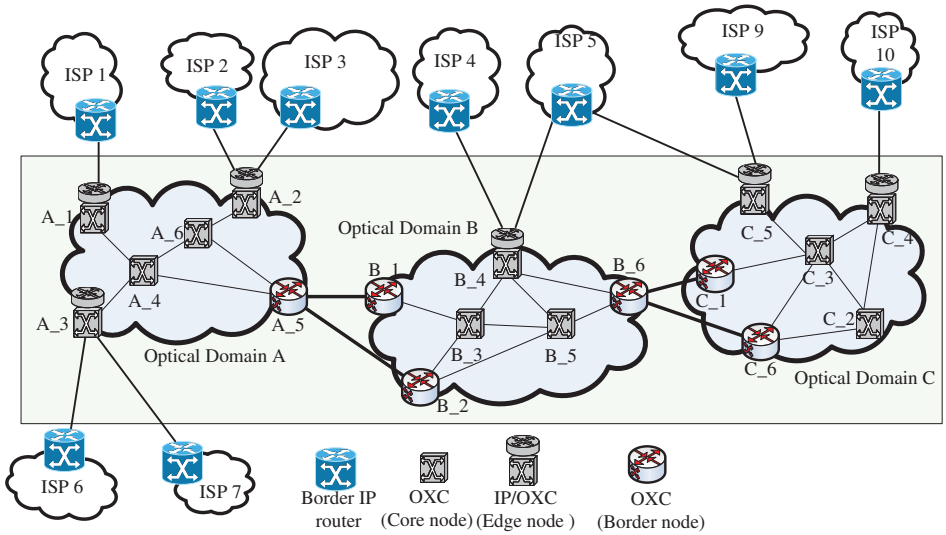


Figure 3.2: Routing scope of BGP in optical transport networks.

- The type of networks - The Internet is a packet based network and the optical networks are connection-oriented networks. This necessitates different routing paradigms to be used. In IP networks, each client packet is routed hop-by-hop, whereas in optical networks data is forwarded end-to-end, using pre-established connections (paths). The underlying routing paradigm in IP networks requires that routers use the same topological view of the network and the same routing algorithm in order to avoid routing loops. In optical networks, on the other hand, each router can use different path computation algorithms. This does not create routing loops for the client traffic because the traffic is explicitly routed. Even though the GMPLS architecture allows for loose-hop routing of the connection requests, source routing is considered a paramount TE feature.

- Data/Control plane relationship - In GMPLS networks the Control and Data planes are separated whereas in IP networks they are integrated. In IP networks when a failure in the control plane

occurs, this affects the client traffic as well, because for each packet both processes of routing and forwarding need to be performed on a per node basis. In optical networks the routing and the data forwarding functions are separated, just as the control and the data planes are. This implies that a failure or error in the control plane does not affect existing connections, only new connection requests.

- Survivability support - The connection-oriented networking paradigm requires path diversity for survivability support. In IP networks, each client packet is routed based on the current topological view of the router processing the packet, which means packets can be deflected on alternative paths dynamically. In optical networks, an alternative path needs to be pre-established in order not to lose client traffic at times of failure.

- Routing information dissemination - Constrained-based path computation is considered paramount for TE and QoS support in GMPLS networks. This requires the dissemination of additional link-state, node-capability and other information (for details see [58]). In conventional IP networks the main routing goal is to maintain a consistent topology view for hop-by-hop forwarding based on simple administrative costs.

Being an IP routing protocol, designed for packet-switching in large scale multi-AS networks, BGP features certain characteristics which are at odds with the main requirements for optical routing. When analyzing these features, the outlined routing model presented in Section 3.2.2 is considered. The proposed BGP extensions in this thesis do not seek to provide all required functions for optical multi-domain connection provisioning solely with the BGP protocol. Instead, the goal is to enhance its operation so that it can be used as an efficient high-level routing protocol within the specified earlier routing model.

3.2.4 Challenges in applying BGP for GMPLS routing

BGP has proven to be scalable and efficient for inter-domain routing - it has become the *de facto* routing standard in the Internet. Nevertheless, the fundamental differences between the optical and the IP networks (see the previous section) turn some of the very attractive BGP features into

drawbacks. The first major challenge with BGP is the lack of end-to-end TE information dissemination. Limiting the amount of disseminated information per reachable destination and applying aggregation whenever possible has made BGP a very scalable protocol capable of sustaining stable operation under thousands of interconnected ASs. Applied in optical networks though this advantage turns into a challenge since GMPLS nodes need TE information in order to perform constrained-based path computation for TE and QoS provisioning. The existing options for conveying limited TE information across borders with BGP (see Section 2.4) have only local significance and are often ignored unless special bi-lateral agreement exists between the interacting domains. However, even if such local TE is performed the path selection process in BGP makes paths with shorter AS-path preferable over paths with better MED metric (see Section 2.4). Applied in connection-oriented networks this problem leads to the problem of hiding feasible paths [60].

Another challenge with BGP is the slow protocol convergence, caused by path exploration and chattiness among the BGP routers. BGP is a path vector protocol and under network change conditions it may take minutes until a consistent view of the overall network is created in all BGP speakers. A direct implication for GMPLS networks is worsened survivability. In case of a failure huge amounts of traffic could be lost unless path protection is applied across the network. Path protection requires link-disjoint paths per destination, which cannot be supplied by standard BGP due to the one-path-per-destination policy of the protocol. Methods for computing link-disjoint paths using the same AS chain exist [61, 62], but require complex disjoint computations in each domain and specific, additionally disseminated, information. If no such methods are employed, waiting for BGP to re-converge and provide a stable path to an affected destination, in case of a failure, can take very long time.

The last challenge for BGP can be found in the path dependency phenomenon. BGP speakers advertise only the route to a destination which they use for traffic forwarding themselves, because BGP supports only the destination-based forwarding paradigm [37]. As a result, some multi-domain links become overloaded and others remain with low utilization of available resources, especially when there are multiple links between domains. Furthermore, paths with acceptable TE state can be hidden from upstream domains. Fig. 3.3 illustrates the problems related

to path dependency. When node 33 in AS 3 performs path selection, it advertises to AS 4 only one path to the Destination - the one it uses itself. Thus, traffic from AS 4 to AS 1 cannot be redirected via AS 2 because AS 3 itself does not use this path. As a result, the inter-domain link 32-22 will not be utilized efficiently, whereas link 31-11 will be carrying all traffic from AS 4 and AS 3 to AS 1.

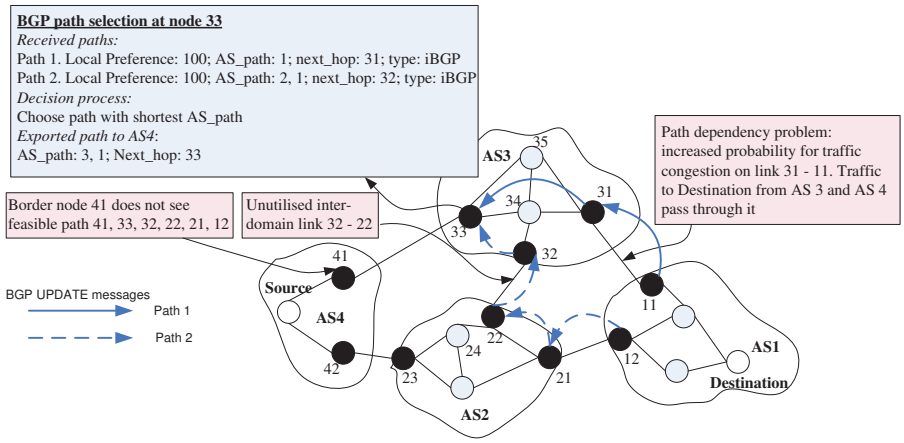


Figure 3.3: Path dependency problems with BGP applied for GMPLS networks.

3.3 BGP TE-attribute

Several literature sources elaborate on an extended TE-attribute for BGP [63–65]. The only solution, proposed for GMPLS networks, is from [65], but the application of the attribute is only for Layer 1 Virtual Private Networks (VPNs). Depending on the actual application and role of the BGP protocol within the provisioning process, different information can be conveyed with a potential BGP TE-attribute. In this section a proposal for the contents and the structure of a BGP TE-attribute is given, which is based on the routing model described in Section 3.2.2.

3.3.1 TE-related information

As presented in Chapter 2, there is a difference between Traffic Engineering and QoS provisioning. Nevertheless, these two terms are tightly bound. TE-attributes in GMPLS networks refer to GMPLS-specific link and node (interface) attributes, which are necessary for the proper engineering of the end-to-end path of a connection. Since next generation transport networks are expected to be complex heterogenous networks, very specific information such as the switching capability of a node or the encoding type of the link, is needed in order to achieve efficient TE. A complete description of the GMPLS TE-attributes is given in [17]. Taking into account the specifics of the GMPLS technology the following attributes can be provided within a BGP TE-attribute for proper end-to-end path engineering [38,58,65,66]:

- Protection type
- Shared Risk Link Group (SRLG)
- Interface switching capability
- Bandwidth
- Encoding type

The proposed attributes in the literature follow the format of the TE extensions for the intra-domain routing protocols (OSPF-TE in particular). This implies that these characteristics are related to link-state protocols. BGP, on the other hand, is a path vector protocol. The applicability of link-state characteristics in the path vector protocol brings up confusion and limits the applicability of the BGP TE-attribute as seen in [65]. Providing link-related information within the BGP protocol is suitable for inter-domain TE involving only 2 neighboring domains as in [66], where tunnels are pre-established between border routers. Another field of application could be for cross-layer dissemination, where TE paths established on a lower layer appear as links in the upper layer (under the overlay GMPLS model). Nevertheless, providing any of the presented characteristics can aid the process of engineering the end-to-end path of a connection especially in heterogenous multi-domain GMPLS networks, where domains with diverse TE capabilities are interconnected.

3.3.2 QoS-related information

The process of providing QoS corresponds to complying with specific connection parameters requested from the clients. Providing a connection (LSP) with certain parameters such as available bandwidth, protection type, and/or delay, necessitates constrained path computation. Computing and providing such a path is related to TE in the network. Thus, providing QoS metrics within a path attribute is necessary for traffic engineering. Apart from the presented attributes above, typical QoS metrics such as end-to-end delay, available bandwidth (reservable wavelengths), blocking probability, etc., can also be included in a BGP TE-attribute. Disseminating information regarding the physical quality of paths is also important for impairment-aware routing. Suggestions for including such attributes exist for the standard BGP application in the Internet [63,64] and for the OBGP protocol [67]. So far, there have been no suggested QoS-related BGP extensions strictly for GMPLS networks.

3.3.3 Integrated BGP TE-attribute

Considering the routing model described at the beginning of this chapter and the actual application of the BGP protocol within it, a combined TE metric which supports both QoS provisioning and high-level TE is proposed here. Since the role of BGP is to provide AS-paths for reachable destinations, the TE and QoS information conveyed in the TE-attribute has to be related to the complete end-to-end path from ingress to egress node (see Fig. 3.2). This implies that the BGP TE-attribute is disseminated per path, which spans multiple domains, not per virtual link as in [66] or within a single provider network as in [65]. This retains the path-vector character of the higher level routing protocol.

A novel proposal the BGP TE-attribute, consisting of several metrics related to different aspects of the end-to-end quality and characteristics of the path, is suggested here. In general, the TE-attribute can carry any subset and combination of the outlined QoS-related and TE-related metrics, depending on the agreements within the cooperating domains, which employ the BGP extension (see the assumptions in Section 2.5). Considering the scalability of the BGP protocol, the contents of the TE-attribute should be coordinated with the network architecture (heterogenous or homogenous multi-domain environment), the set of supported

QoS metrics between the domains and the provided TE capabilities.

Conveying TE characteristics

Providing TE characteristics such as switching capability, SRLG, protection type and/or encoding type is important for proper end-to-end path design especially in heterogenous environments[2]. In such cases the end-to-end TE characteristics of a path cannot be represented as a single aggregated value (see Fig. 3.4). Instead, the TE characteristic of a path can be presented as a tuple of values, corresponding to the characteristics of each passed domain. For example, on Fig. 3.4 a path complying with the requirement for consistent encoding type might pass through domains using different switching capabilities, in which case the head-end router needs to know in advance how many switching-capability adaptations the client traffic will go through. This can determine the cost of establishing the path, or the end-to-end quality. Nevertheless, providing explicit information per domain is very beneficial for fine tuning of TE characteristics in the network (e.g. for survivability management).

Conveying QoS metrics

Disseminating QoS metric(s) across domain borders can potentially harm the confidentiality preservation requirement between domains. In particular, domains should not disseminate to other domains the TE state (QoS metrics and/or topology information) of their neighbors. This is the reason why under the standard BGP protocol exporting the MED attribute to other neighbors is prohibited, i.e. the MED metric is shared only between two direct neighbors. Following this requirement, the QoS metrics need to be aggregated per path instead of provided in a tuple per segment as done with the TE characteristics. Applying an aggregated QoS metric per path indicates the complete state of the path[3], without disclosing sensitive topological and/or network state information to non-neighbor domains.

[2] In this work, a heterogenous inter-domain environment refers to interconnection of domains that are homogenous themselves, but may differ in their TE parameters from their neighbors.

[3] Referred to as *TE state* further in this thesis.

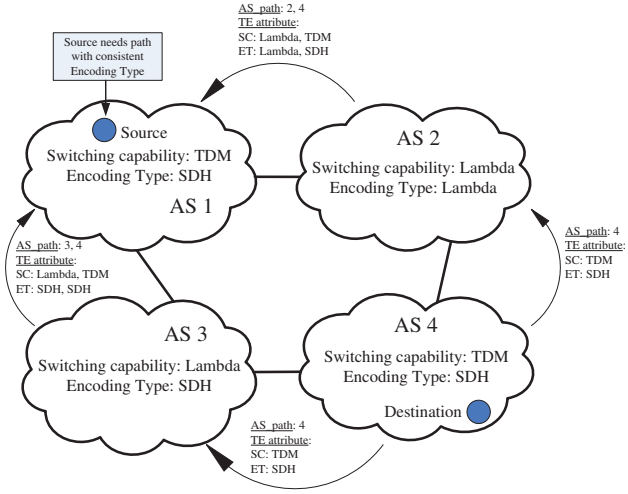


Figure 3.4: Reachability dissemination across heterogenous neighboring domains with BGP TE-attribute.

Aggregating QoS metrics can be done in different ways, depending on the metric itself. Some metrics are additive such as hop-count or end-to-end delay. Others are bottleneck-type, such as bandwidth (or wavelength) availability. Some impairment metrics are nonlinear and depend on the current state of the network. How these metrics are updated along the path has to be standardized in order to achieve a consistent view of the end-to-end quality of the path.

3.3.4 BGP TE-attribute operation

The BGP path selection procedure performs local TE within the tie-breaking set of rules [37], i.e. choosing a path with good QoS metric is not the first decision criteria. In particular, the decision steps which refer to local TE are: i) between two neighboring domains - the MED comparison and ii) within a domain - the lowest Interior Gateway Protocol (IGP) metric (see Section 2.4). Thus, paths with better QoS metric(s) can be hidden, due to longer AS-path attributes [60]. This drawback can be avoided by modifying the BGP path selection procedure to consider the conveyed QoS metric as a first decision criteria.

Fig. 3.5 illustrates the operation of the proposed TE-attribute. Only

the QoS-related information is presented where several metrics are considered. The advertised destination is C2 in Domain C. Border node C1 calculates the set of QoS metrics (in the example these are wavelength availability (WA), hop count, cost and delay), and includes them in the TE-attribute of the UPDATE messages sent to neighbors B2 and D2 (MSG.1). Border nodes B2 and D2 obtain the QoS metrics from the received UPDATE messages, calculate the QoS metrics from themselves to the next hops (obtained from the UPDATE messages again) and update the end-to-end path state, depending on the metrics type. Msg.2 and Msg.3 are generated by including the updated path state (the new end-to-end QoS metrics) in the TE-attributes. Border node B1 obtains Msg.2 and Msg.3, accesses the QoS metrics from the TE-attributes, calculates the QoS metrics of the path segments towards nodes B2 and D2, updates the end-to-end path state, runs a decision process by considering one or more specific QoS metrics as a first decision step and chooses one path. Msg.4, which contains the aggregated path state of the chosen path in the TE-attribute, is generated. Border node A2 preforms a similar procedure and chooses one path (among Msg.3 and Msg.4) with the best end-to-end QoS metric. The procedure for TE-attribute updating at each node is outlined in Algorithm 1.

The application of the proposed attribute is for dynamic networks, where the TE state of a path changes rapidly[4]. Thus the TE state (the end-to-end QoS metrics) of paths needs to be updated in a way that assures satisfactory performance. Different strategies for TE state update can be applied:

- Timer-based - the TE-state update is triggered periodically;
- Threshold-based - the TE-state update of a path is triggered when a link is in overloaded or under-loaded condition;
- Combined.

BGP TE-attribute format

The proposed format of the BGP TE-attribute is a standard Type-Length-Value (TLV) format used for parameter encoding in the GMPLS

[4]Relative to the static transport network operation today.

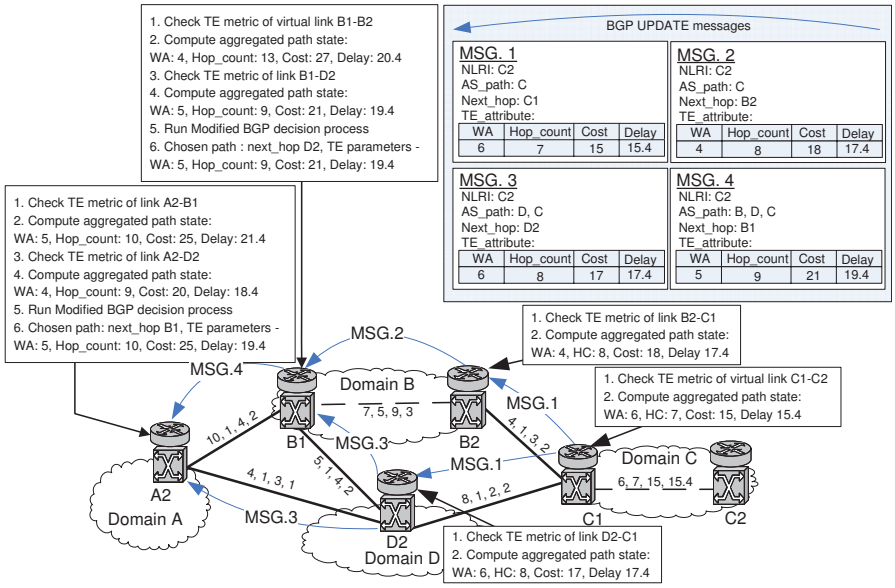


Figure 3.5: Conveying end-to-end QoS metric with BGP TE-attribute.

protocols. The *Type Code* value for the attribute is a matter of standardization. Two sub-TLV fields may be present: one for the QoS-related metric(s) and one for the TE-related characteristic(s) (see Fig. 3.6). The format of the sub-TLVs is also a recursive TLV structure. The contents of the TE-attribute sub-TLVs are:

- QoS metrics sub-TLV: The values in this sub-TLV must be QoS end-to-end path related metrics. The representation of each metric is in a <type,value> format, so no special ordering of the fields is necessary. Possible code types and values are presented below:

Type code	Value (metric)
1	Delay
2	Hop count
3	Reservable Bandwidth
4 to N	Different Impairment parameters

- TE characteristics sub-TLV: The values listed in this sub-TLV must

begin
1. Obtain TE metric from received UPDATE packet, identify type (additive, bottleneck, etc.)
2. Obtain next hop from received UPDATE packet
3. Compute TE metric from current node to next hop
if *Next hop not reachable* **then**
Ignore the Update;
end
4. Update Received TE metric:
if *Metric is additive* **then**
$newTEmetric = oldTEmetric + currentTEmetric$
else if *Metric is bottleneck* **then**
$newTEmetric = min\{oldTEmetric, currentTEmetric\}$

5. Perform Path selection:
if *Current path TE metric is better* **then**
Retain current path
else if *Current path TE metric is the same* **then**
Apply standard BGP path selection
else
Disseminate new path to neighbors, install new path in routing table

end

Algorithm 1: Procedure at receiving new UPDATE message

represent the characteristics of the individual ASs (their E-NNI interfaces), listed in the AS-path attribute. Thus, there needs to be a correspondence between both attributes. For each entry in the AS-path array presented in the AS-path attribute there is a corresponding TE characteristic sub-TLV, each of which consists of TLVs describing the TE characteristics outlined earlier in this section.

Type Code: BGP-TE	Length (bytes)	
Type: QoS metric(s)	Length (elements)	QoS sub-TLV
Type: Delay	Value	
Type: Bandwidth	Value	
Type: Hop Count	Value	
…….	…...	
Type: TE characteristic(s)	Length (bytes)	TE characteristics sub-TLV
AS indicator 1	Length (elements)	
Type: Switching capability	Value	
Type: Encoding type	Value	
…….	…...	
AS indicator 2	Length (elements)	
Type: Switching capability	Value	
Type: Encoding type	Value	
…….	…...	
AS indicator N	Length (elements)	
Type: Switching capability	Value	
Type: Encoding type	Value	
…….	…...	

Figure 3.6: TE-attribute format.

3.4 Performance evaluation of the proposed BGP TE extension

In this section the performance of the proposed BGP-TE extension in a dynamic multi-domain wavelength-switched network is evaluated. Providing end-to-end TE-attribute solves only one of the BGP drawbacks, outlined in Section 3.2.4 - the lack of TE dissemination capabilities. Thus, the focus of the performance comparison is mainly on the blocking of LSP requests. Blocking of LSP requests can happen due to two main reasons: lack of resources and/or lack of visibility, caused by routing loops during BGP convergence or by failures.

First, the performance of the proposed TE-attribute is compared to the standard approach for local inter-domain TE via MED compa-

rison [37]. Standard BGP operation is used as a baseline case. Then, four different TE-state update strategies are tested - two timer-based and two threshold-based. The created BGP overhead during protocol re-convergence for TE state update is evaluated as well.

3.4.1 Simulation setup

The proposed BGP TE-attribute was implemented in the event-driven simulator OPNET [68]. Details on the implementation can be found in Appendix A. Two different topologies are tested[5]: a general mesh topology (*Topology 1* Fig. B.1) and a Pan-European topology (Fig. B.4) from the COST 266 project [69]. The TE metric of interest is wavelength availability. Since it is a bottleneck-type of metric, the value of the metric per path indicates the lowest wavelength availability on the path. Resource ReserVation Protocol with TE extensions (RSVP-TE) [21] is used for LSP setup. BGP provides only the next-hop towards a destination, thus the intra-domain path is calculated at the time of LSP request. RSVP-TE uses the standard Label Set (LS) attribute with First Fit (FF) wavelength assignment policy.

The connection holding time is an exponentially distributed value with a mean of 7200 seconds (i.e. 2 hours). For the mesh topology, there are 20 wavelengths per link, where each link has a delay of $1ms$. For the Pan-European topology, link delays depend on the actual geographical position of the nodes, with 100 wavelengths per link. The values for the BGP Minimum Route Advertisement Interval Timers (MRITs) are set according to the specification in [37], i.e. 30 seconds for eBGP and 5 seconds for iBGP.

3.4.2 Simulation results

Three sets of results are presented here. First, the blocking ratio of LSP requests in the general mesh topology is evaluated in three cases:

- Standard BGP with no TE (referred to as BGP);
- Standard BGP with local TE by using the Multi-Exit Discriminator (MED) comparison with the "Always-compare" policy [37] (referred to as BGP-AC);

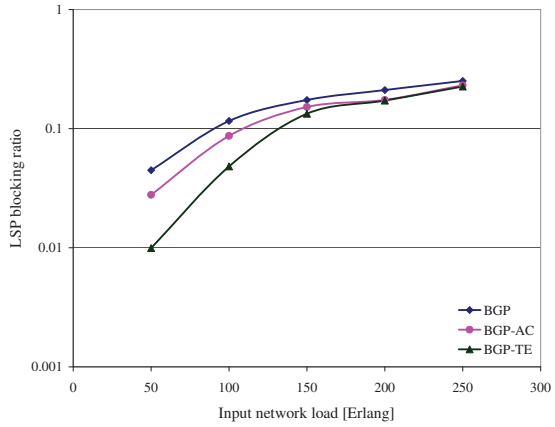
[5] All tested topologies can be found in Appendix B.

- An enhanced BGP, employing the proposed TE-attribute as a first path selection criteria (referred to as BGP-TE).

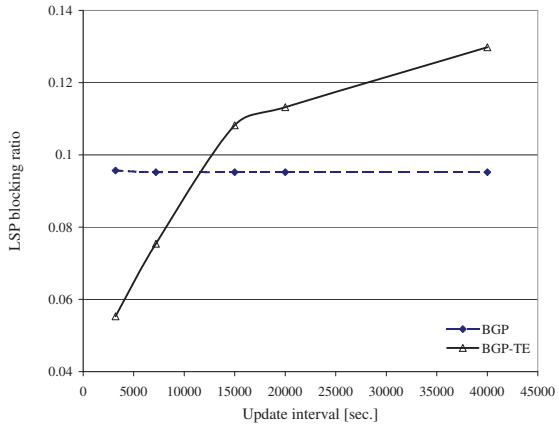
The relationship between the efficiency of the BGP TE-attribute and the Update Interval (UI) (the interval for updating the TE state of the paths) is evaluated for a fixed traffic load in the network. Finally, the performance of the TE-attribute under different TE information update strategies is presented for the Pan-European topology.

LSP blocking ratio

For this set of results the BGP MRITs timers were disabled. This was done in order to observe the effect of the BGP TE-attribute in general. Under such an operation all changes in the TE state of a path are disseminated via the BGP UPDATE messages. Fig. 3.7(a) presents the relationship between the blocking ratio for LSP requests and the input traffic load for three cases: no TE (BGP), local TE (BGP-AC) and end-to-end TE (BGP-TE). The UI for TE information update is 7200 seconds. As can be expected, using the BGP TE-attribute provides a lower blocking ratio for the requests, compared to both the standard BGP and the local TE with BGP cases. At high loads the benefit of using the BGP TE-attribute decreases, since all domains become equally congested and no better path can be found. Fig. 3.7(b) shows how the blocking ratio changes when the update interval is increased for a fixed load in the network for the BGP-TE and the standard BGP schemes. The longer the period between updates, the lower the efficiency of the BGP TE-attribute is. In fact, the results show that the BGP TE-attribute is efficient only when the connections have durations, which are comparable to the TE information update interval. Using long update intervals leads to using the same AS-path for many LSP requests, which overloads the inter-domain links, especially in transit domains. This is due to the path dependency characteristic of the BGP protocol. When the update interval is shorter, the BGP changes the AS-paths for the requests more often and thus, performs load balancing. This decreases the blocking of requests, but results in higher cost in terms of signaling overhead, as illustrated further in this section.



(a) Blocking ratio vs. Input network load.



(b) Blocking ratio vs. TE metric update interval.

Figure 3.7: Performance evaluation of BGP TE-attribute.

TE metric update strategies

As it was shown in the previous section, the benefit of applying the BGP TE-attribute is present only when the connection duration is comparable to the TE metric UI. For very large networks this is highly inefficient, especially when the BGP MRIT timers are enabled due to the very slow protocol re-convergence. During this period many LSP requests are lost.

This can be observed for the Pan-European topology (see Fig. 3.8). The normalized input load is calculated according to Equation 3.1.

$$A = \frac{D}{MIT * LC} \quad [Erlang], \tag{3.1}$$

where A is the normalized input load per node, D is the average connection duration, MIT is the Mean-Interarrival Time between LSP requests and LC is the capacity of a link in number of wavelengths.

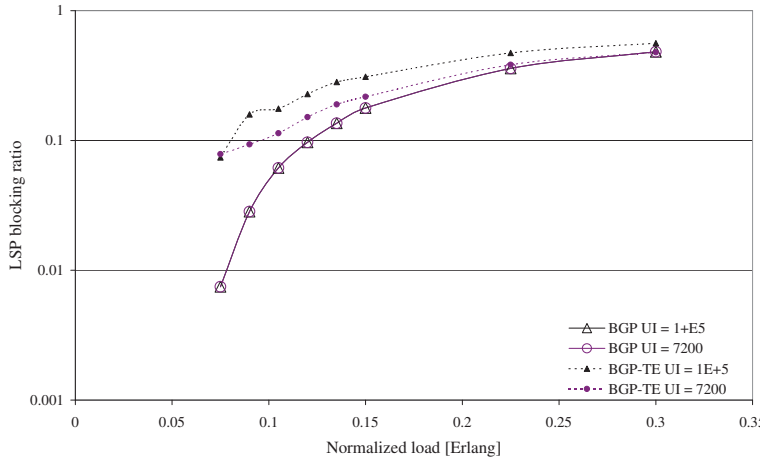


Figure 3.8: LSP blocking ratio for the Pan-European network.

As it can be expected, the UI's effect is negligible on the standard BGP, since it does not use the end-to-end TE information during the path selection process. For the BGP-TE scheme, the shorter UI brings lower LSP blocking, but still much higher than the standard BGP without TE extensions. This is due to the slow protocol convergence, which contributes to the increased amount of blocking due to lack of visibility. Furthermore, BGP-TE changes the AS-paths based on the distributed TE metric. This, combined with the path dependency, leads to overloading a small amount of inter-domain links and to moving this congestion to another set of inter-domain links every TE update cycle.

In order to improve the performance of the BGP TE-attribute, different update triggers can be used, which aim at minimizing the negative

effect of the re-convergence period on the LSP establishment. Four different strategies are evaluated further, two time-based and two load-based:

- Time-based, synchronous (TS) - the update of the TE metrics is governed by a global UI timer. This is the base-line policy used in all results, presented so far (including for Fig. 3.8).

- Time-based, asynchronous (TA) - the UIs for the reachable destinations in the different domains are asynchronous, i.e. each domain is managed by its own UI timer. This is done in order to avoid possible synchronization in the update process.

- Load-based, synchronous (LS) - the update of the TE metric depends on the load of the inter-domain links. When the state of a link changes from an overloaded to lightly loaded condition or vice versa, the reachable destinations in the affected domain initiate TE metric update. The link load is checked synchronously on all inter-domain links according to a global UI timer.

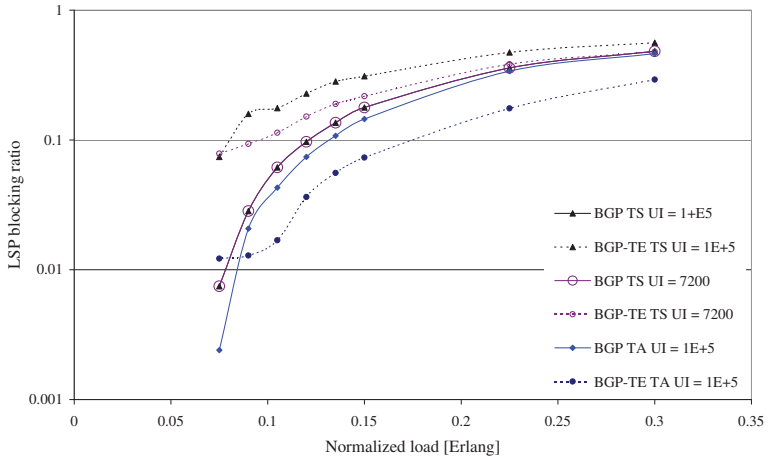
- Load-based, asynchronous (LA) - as with the LS strategy but each domain checks the load on its inter-domain links independently.

The load-based solutions are combined threshold/timer-based solutions, thus threshold values and corresponding timers must be carefully chosen so that the created overhead and the achieved performance are in satisfactory balance. For the presented results, the TE-metric update is triggered every time the load on a link exceeds 2/3 of the capacity and the link was lightly loaded at the previous TE update cycle, or the load goes below 1/3 of the capacity and the link was highly loaded at the previous TE update cycle. Two timer values have been used for the UI: $1 * 10^5$ seconds and 7200 seconds.

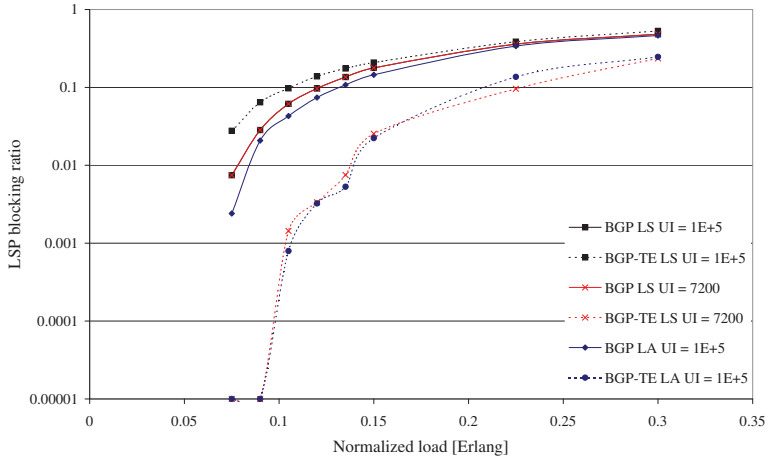
Fig. 3.9 presents the LSP blocking ratio versus the normalized input load per node for the different strategies. Under the time-based schemes, the BGP-TE scheme benefits more from eliminating synchronization than from frequent TE state updates. Even at high values of the UI, the TA scheme performs better than the TS scheme at lower UI value. This is due to the fact that when updates are not simultaneous, only some source/destination pairs change their AS-paths at a time, which reduces the negative effect of the route flapping. The effect

of the asynchronous schemes on the standard BGP is due to the changes within the intra-domain TE state which affect the BGP decision process (see the decision process described in Section 2.4). For the load-based schemes, the BGP-TE outperforms the standard BGP operation in terms of LSP blocking when the TE state is updated more often. Furthermore, eliminating synchronization within the load-based scheme also improves performance. In general, synchronous TE state update results in moving congestion from one part of the network to another. This is referred to as *route flapping*. If the updates of the paths are done asynchronously, only few source/destination pairs change their paths at a given period of time, which leads to better load balancing in the network.

Fig. 3.10 shows the amount of generated BGP overhead during protocol re-convergence for the BGP-TE scheme during the whole simulation duration. For the TS scheme under $UI = 7200$ seconds the overhead is around $9 * 10^5$ UPDATE packets, which is not illustrated so that the scale of the changes in the rest of the schemes is more visible. The time-based schemes are considerably less dependent on the amount of input load in the network, whereas for the load-based schemes the higher the load is - the bigger the overhead is. This is due to the fact that, at higher loads, more inter-domain links change their state frequently beyond the given limits. The smaller the UI is, the more often the status of the links is checked and updated if needed. Thus, higher overhead is generated at lower UI value and at high traffic loads. Furthermore, using the load-based schemes yields smaller overhead than the time-based schemes because the TE state of paths is changed only when needed, whereas under the time-based schemes the status is always updated.



(a) Time-based strategies.



(b) Load-based strategies.

Figure 3.9: LSP blocking ratio vs. Normalized input load for different TE-state update strategies for the BGP-TE mechanism.

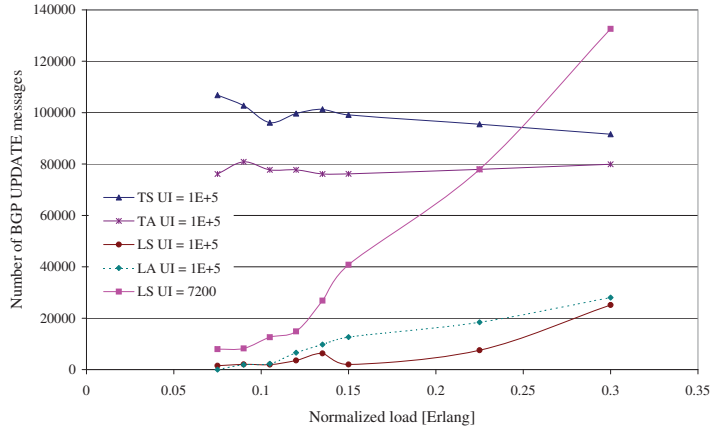


Figure 3.10: Number of BGP UPDATE messages vs. Normalized load for the BGP-TE scheme.

3.5 Conclusion

In this chapter the main driving forces and motivation for the design of a BGP TE-attribute in GMPLS multi-domain networks are presented. First, proposals regarding the BGP TE-attribute content and the mechanism for its distribution and operation are outlined. Then, the efficiency of employing such TE-attribute are analyzed by means of simulations. The presented results show that the employment of a BGP TE-attribute alone is not enough for adequate QoS provisioning and TE in multi-domain GMPLS networks, especially when the average connection duration in the network is short. This is due to the need for frequent TE metric updates, which leads to increased protocol overhead. It is also shown that employing end-to-end TE metric for path selection is highly dependent on the chosen TE metric update interval and the applied update strategy. The inherent path dependency hinders the efficient application of the end-to-end TE-attribute because it leads to flapping congestion from one part of the network to another when long TE state update intervals are used. Applying threshold-based schemes improves the performance compared to using time-based schemes because the TE state of paths is updated only when needed. Avoiding simultaneous updates of the TE state of many paths also improves performance, because it re-

sults in a more efficient distribution of the load among the inter-domain links, which are often the bottleneck links in multi-domain topologies.

The presented results indicate that the lack of TE information dissemination within BGP is not the biggest drawback of the protocol. Instead, the path dependency is a greater challenge, since it impedes the efficient applicability of the end-to-end TE metric for path selection.

Chapter 4

Enhanced BGP Protocol

4.1 Introduction

As discussed in Chapter 3 , the proposed BGP TE-attribute solves only one of the outlined BGP drawbacks, namely the lack of TE information dissemination. It was also shown that providing TE information alone is not enough to achieve a robust and efficient multi-domain routing. Since BGP is a path-vector protocol, its behavior depends heavily on the network load (see Section 3.4.2). Furthermore, proper operation parameters (e.g. update interval, threshold values) need to be carefully configured in order to provide the needed network performance.

In this chapter a modified BGP protocol is proposed, which aims at solving the drawbacks of the standard BGP protocol outlined in Chapter 3, Section 3.2.4. The proposal for BGP enhancement is twofold. First, BGP is re-designed to support multi-path dissemination by disabling the path selection phase of the protocol operation. Second, the set of path attributes is extended with a *Border_node_sequence* attribute which is needed for multi-path BGP operation. For end-to-end TE the BGP TE attribute described and evaluated in Chapter 3 is employed.

The main design decisions along with a step-by-step explanation of the operation of the proposed BGP modifications are given in Section 4.2. The performance of the proposal is evaluated by means of simulations in Section 4.3. Section 4.4 concludes the chapter.

4.2 Enhanced BGP protocol

This section presents the design, operation and some specific aspects of a modified BGP protocol, referred to as *Enhanced BGP*.

4.2.1 Design

BGP performs two main functions: path selection and path dissemination. Path selection is performed at each BGP speaker during BGP convergence and may be performed several times per destination. Most of the problems BGP encounters are due to the path selection procedure, which slows down convergence and results in path dependency. Thus, under the Enhanced BGP proposal, BGP speakers only disseminate policy-compliant paths to their neighbors, without performing path selection. When a path is received in a node, it is stored in a Virtual Topology (VT) database[1]. If the path is eligible for export to neighbors (depending on local export policies), it is sent further after proper TE attribute update. At the end of the path-dissemination process (i.e. the BGP convergence) each head-end router has several paths towards each destination, each of which has a specific TE metric provided by the proposed earlier BGP TE-attribute. At the time of LSP request the head-end router can choose the best path among all disseminated paths.

To allow this modification operation, a *Border_node_sequence* attribute is required. Since the standard BGP supports only the destination-based routing (i.e. hop-by-hop routing), having multiple paths per destination cannot be supported in a straightforward manner. Providing a list of strict border nodes along a path, which can be directly used as a loose Extended Route Object (ERO) in the RSVP-TE *PATH* message, facilitates multi-path operation. Moreover, this modification brings the BGP routing closer to the GMPLS-compliant source-routing paradigm. Using only the AS_PATH sequence, which is disseminated by the standard BGP, is not enough for supporting multi-path BGP dissemination in topologies with multiple links between neighboring domains. Since the intra-domain routing protocols do not keep track of the TE state of the inter-domain links as standard, a source node would not have enough

[1] This implies that the standard "implicit path withdraw" (see Section 2.4) is also disabled.

information to decide the correct exit node/link from its own domain[2] Thus, using the *Border_ node_ sequence* attribute can provide the needed exit node from the domain. Furthermore, this attribute gives a higher control of the head-end of a connection over which inter-domain links are to be used.

In summary, the design of the modifications is directly aimed at solving the four main drawbacks of the BGP protocol outlined in Section 3.2.4. The utilization of the suggested earlier BGP TE-attribute provides TE information dissemination; the multi-path dissemination aids the processes of survivability support and load-balancing; the lack of path selection within the BGP operation decreases the negative effects of the path dependency and the slow protocol convergence on the LSP provisioning process during BGP convergence. Furthermore, after protocol convergence each BGP speaker has several paths per destination, which are policy compliant and can be used for LSP establishment. Even if BGP re-convergence is required (e.g. for TE metric update) this does not lead to loss of visibility, as it was observed for BGP-TE in Chapter 3.

4.2.2 Operation

The proposal in this chapter fits within the last group of routing options provided in Table 3.1. Both levels of the routing hierarchy employ source-routing, and thus have dynamic TE capabilities. The lower level uses the current intra-domain state to compute explicit path between each pair of border nodes, whereas the higher level uses the TE-enhanced BGP updates to select the path with the best end-to-end metric. The operation of the proposal is depicted in Fig. 4.1. The advertised destination is in Domain C. Nodes B1 and D1 from domains B and D, respectively, will eventually receive several advertisements for the destination. Instead of choosing the best among them, they only re-distribute the received paths further to their neighbors (after applying local export policies). In this way, the source node will receive multiple paths, each supplemented with a TE-attribute and a *Border_ node_ sequence* attribute. Using these multiple advertisements the source node can create an abstracted VT of the multi-domain network (as shown on Fig. 4.1) and choose a path with an appropriate TE metric at the time of LSP request.

[2] Unless it is explicitly provided by the network administrator.

In order to minimize the effect of inaccurate TE state information, the state of paths must be updated regularly. Since the Enhanced BGP does not perform path selection and is not governed by the Minimum Route Advertisement Interval Timers (MRITs)[3], the delay for TE metric update depends only on the propagation delays of the links and the packet processing delays in the nodes.

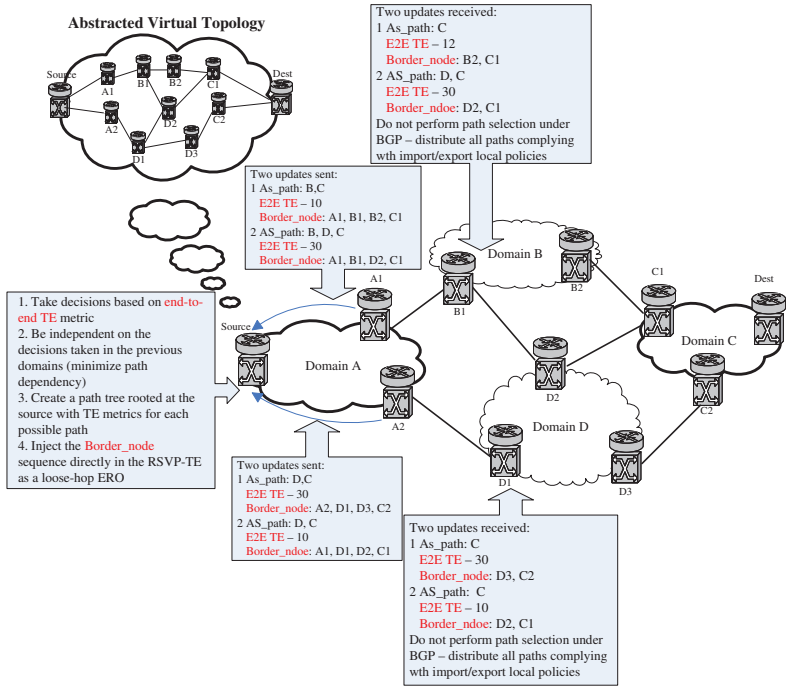


Figure 4.1: Enhanced BGP operation.

4.2.3 Export policies

A very important aspect of the Enhanced BGP protocol is the applied export policies, since they affect greatly the scalability of the solution. In a large network with many inter-domain links and many BGP speakers, distributing every possible path would create enormous overhead. Thus,

[3]For details on the necessity of the MRIT timers see [37].

the amount of exported paths needs to be carefully controlled. Different export policies can be applied. Depending on the local TE goals (e.g. avoid paths which use many local resources) and the negotiated inter-domain agreements (e.g. do not export more than 3 paths for a given reachable destination to a neighbor) different set of rules can be applied in each BGP speaker.

Six different export policies are proposed here:

- Policy 1: BGP speakers export paths with longer AS_PATH attribute only if the path provides strict AS_PATH disjointness to all of the already distributed paths;
- Policy 2: BGP speakers export only the first 5 of the paths they receive, but are allowed to use all paths received from neighbors;
- Policy 3 - exclusive: BGP speakers export only paths which are AS-disjoint to all paths distributed so far in the process;
- Policy 4 - exclusive: BGP speakers export only paths which are Border node disjoint to all paths distributed so far in the process;
- Policy 3 - inclusive: BGP speakers export paths which are AS-disjoint to at least one of the paths distributed so far in the process;
- Policy 4 - inclusive: BGP speakers export paths which are Border node disjoint to at least one of the paths distributed so far in the process.

The benefits of having AS-disjoint paths for multi-domain TE are mainly related to survivability. Providing AS-disjoint paths facilitates the process of link-disjoint path computation. If two disjoint AS-paths are available per destination, the complex link-disjoint calculations need to be done only for the source and the destination domains. Furthermore, using AS-disjoint paths for load balancing guarantees a good balance in the usage of the scarce inter-domain resources.

The first policy aims at providing AS-disjoint paths to destinations and at the same time at minimizing the path length (i.e. the end-to-end delay). The second policy is focused solely on scalability and aims at minimizing the amount of distributed paths. The goal of evaluating the last four policies is to see if loosening up the tight restrictions

in the number of distributed paths improves the network performance. The *exclusive* policies will distribute less paths since a path needs to be disjoint to all previously distributed paths. This can potentially result in no AS-disjoint paths in the network at all. When applying the *inclusive* policies, more paths are distributed which increases the chances for finding AS-disjoint pairs of paths and for using more paths for load balancing. A drawback is the increased overhead.

Applying export policies inevitably leads to certain level of path dependency, but unlike the standard BGP, where only one path is distributed per destination, Enhanced BGP supports multiple paths per destination, which decreases the negative effects of the path dependency outlined in Section 3.2.4.

4.3 Performance evaluation of the Enhanced BGP proposal

4.3.1 Simulation setup

The performance of the proposed Enhanced BGP was evaluated via simulations in the event-driven simulator OPNET [68]. Two different topologies are tested: a general mesh topology (*Topology 1* Fig. B.1) and a Pan-European topology (Fig. B.4) from the COST 266 project [69]. The TE metric of interest is wavelength availability. RSVP-TE is used for LSP setup. BGP provides a list of border routers via the *Border_node_sequence* attribute. The segments of the LSP within each domain are calculated at the time of LSP request processing. RSVP-TE uses the standard Label Set (LS) attribute with First Fit (FF) wavelength assignment policy.

The duration of the requests is an exponentially distributed value with a mean of 7200 seconds (i.e. 2 hours). For the general mesh topology, there are 20 wavelengths per link, where each link has delay $1ms$. For the Pan-European topology the links' delays depend on the actual geographical position of the nodes, with 100 wavelength per link. The values for the BGP MRITs are set according to the specification in [37], i.e. 30 seconds for eBGP and 5 seconds for iBGP.

4.3.2 Simulation results

Several results are shown for illustrating the efficiency of the Enhanced BGP. First, the performance in terms of LSP request blocking is shown for the general mesh topology. Then the effect of the Update Interval (UI) is evaluated. The amount of BGP overhead during TE state updateing, as well as the multi-domain link utilization, are examined. Finally, the efficiencies of applying different TE metric update strategies and of using different export policies are examined for the Pan-European network. For all simulation results, the normalized load is calculated according to Equation 3.1.

LSP blocking ratio

The performance of the three different methods for multi-domain TE with BGP in the general mesh topology is compared here. The evaluated cases are:

i) *BGP-TE case 1* which uses the TE-attribute as a first decision criteria as presented in Chapter 3 (i.e. AS_PATH comparison is done after TE metric comparison);

ii) *BGP-TE case 2* which uses the TE metric as a first tie-breaking criteria in a traditional BGP path selection process (i.e. AS_PATH comparison is done before TE metric comparison);

iii) *Enhanced BGP.*

Fig. 4.2 presents the LSP blocking ratio versus the normalized input network load per node for all evaluated cases for two different values of the UI.

As can be seen, the *Enhanced BGP* protocol achieves the lowest LSP blocking ratio. Minimizing the path dependency makes the application of the TE metric as a first[4] decision criteria much more efficient than it is in the *BGP-TE* cases. Furthermore, the *Enhanced-BGP* uses several paths to a destination and effectively performs load balancing, which contributes to the lower blocking ratio. An interesting observation to

[4] The TE metric is actually the only used criteria under the Enhanced BGP operation.

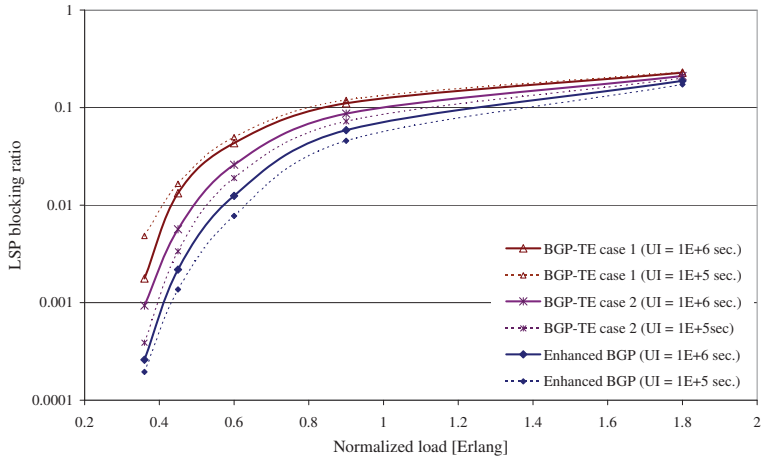


Figure 4.2: LSP blocking ratio vs. Normalized input load per node for two different TE metric update intervals.

be made is that the *BGP-TE case 1* performs worse than the *BGP-TE case 2*. This result is due to the route flapping inherent from the standard BGP path exploration process. Since under *BGP-TE case 2* paths are selected mainly based on the length of the AS_PATH attribute (TE metric comparison is done after that), the route flapping is not that pronounced because the AS_PATH attribute is a more stable attribute.

Applying a shorter UI for the TE state update yields lower blocking ratio which is expected, since a shorter UI provides more up-to-date information. An exception is observed for the *BGP-TE case 1* where the shorter UI yields higher blocking ratio. This is due to the combination of path exploration, during which nodes loose visibility to destinations or create routing loops, and ongoing changes in the value of the TE metric during the protocol re-convergence. The latter is due to the fact that during protocol re-convergence new requests enter the network and old connections expire, which makes the TE metric dynamically change in short periods. Moreover, the values used for the UI are significantly higher than the average connection length. Thus, improvement of the LSP blocking ratio, similar to Fig. 3.7(a), is not observed. In order to compare the true potential of the investigated BGP TE methods the three mechanisms were compared by using the best settings for the UI

observed so far: UI = 7200 for the *BGP-TE case 1* (following the result presented on Fig. 3.7(a)), $UI = 1 * 10^5$ for the *BGP-TE case 2* and $UI = 1 * 10^6$ for the *Enhanced BGP*. The result is presented on Fig. 4.3. Even at the shortest possible UI for the *BGP-TE case 1* scheme it cannot achieve as good performance as the *Enhanced BGP* at the longest UI tested in the simulations.

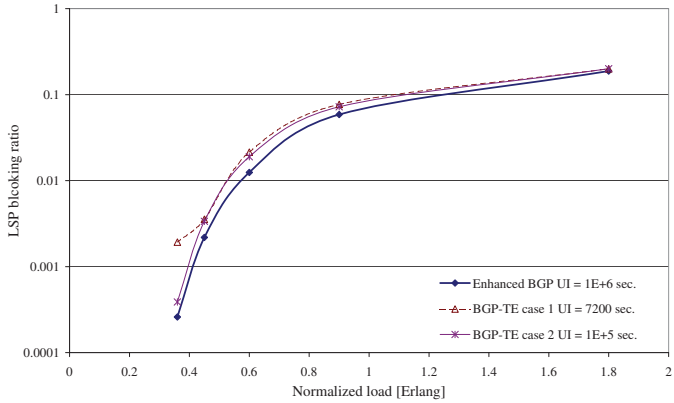


Figure 4.3: Performance comparison for the best observed settings.

BGP overhead and inter-domain link utilization

The efficiency of the proposed *Enhanced BGP* protocol can also be validated by using different metrics in addition to the LSP blocking ratio parameter. Here two such parameters are used - the overhead, created for TE-metric update in the network, and the effect on the inter-domain link utilization. The created overhead for TE state updates is a good measure of the efficiency of the proposal and of the cost which must be payed for the achieved performance enhancement in the network. The link utilization is very important since the inter-domain links are the ones, which get saturated faster especially in transit domains in topologies with a well-defined core area. The performance of the *Enhanced BGP* is compared only to the performance of the *BGP-TE case 1* scheme because the *BGP-TE case 2* scheme has the drawback of hiding paths with better TE metrics behind longer AS_PATH metrics (as discussed in [60]).

Fig. 4.4 illustrates the number of UPDATE messages exchanged during one cycle of TE metric update under *BGP-TE case 1* and *Enhanced BGP*. The considerably higher amount of overhead for the *BGP-TE case 1* scheme comes from the path-dependency problem. Standard BGP experiences route flapping and chattiness even in the contemporary Internet where end-to-end TE metrics are not conveyed. Adding the TE metric in the path selection and increasing the dynamics of the network cause extensive route flapping. Furthermore, there is no clear relation between the load in the network and the observed overhead. For the *Enhanced BGP* on the other hand the overhead is a constant value and is considerably lower. This is due to the fact that no path selection is performed during BGP re-convergence. Only notifications for the TE state of already distributed paths are propagated, which results in relatively constant amount of UPDATE messages at every cycle of the TE metric update. Furthermore, the low dependence of the overhead from the input load is beneficial for better Control Plane (CP) dimensioning and is an indicator for stable protocol operation.

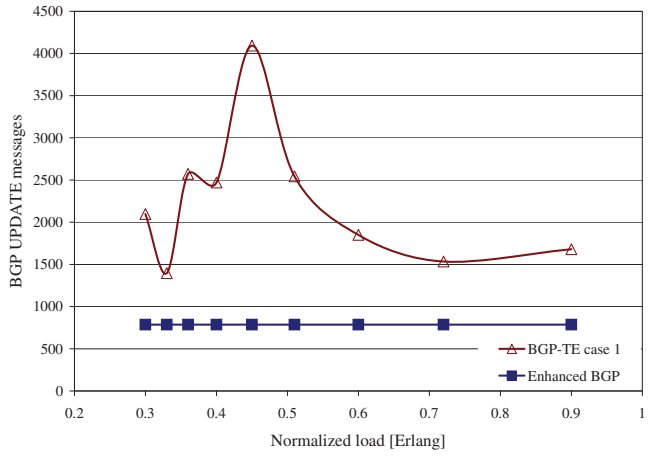
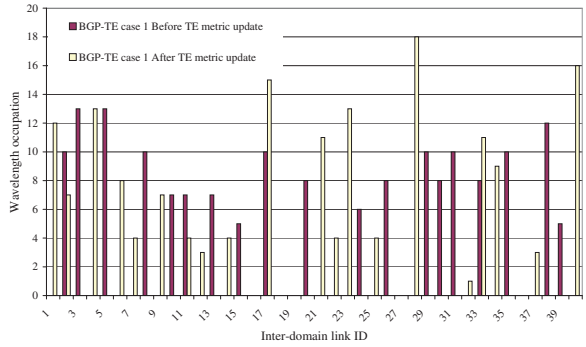


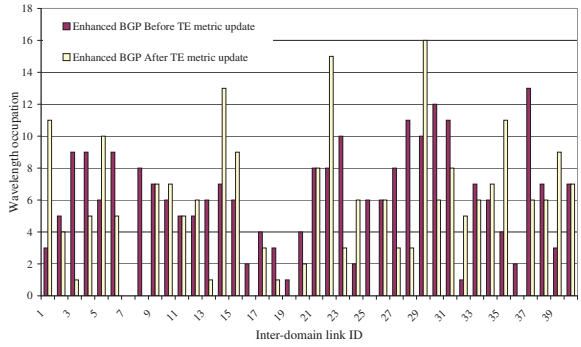
Figure 4.4: BGP overhead under protocol re-convergence for *BGP-TE case 1* and *Enhanced BGP* schemes.

Fig. 4.5 illustrates the inter-domain utilization per link before and after TE metric update for *BGP-TE case 1* and *Enhanced BGP* schemes. The route flapping explained earlier is clearly visible on Fig. 4.5(a). Most

multi-domain links are used either only before or after TE metric update (or BGP re-convergence). Moreover, five of the links are not used at all. This clearly illustrates the negative effect of the path dependency on the network performance. The load balancing feature of the *Enhanced BGP* can be observed on Fig. 4.5(b), where it can be seen that most of the multi-domain links are used both before and after TE metric update. Thus, the *Enhanced BGP* provides better utilization of the scarce inter-domain link resources.



(a) BGP-TE case 1.



(b) Enhanced BGP.

Figure 4.5: Occupied resources per inter-domain link before and after TE metric update.

TE metric update strategies

In this section the update strategies, presented in Chapter 3, are applied for the *Enhanced BGP* scheme and are compared to the performance of the *BGP-TE case 1* (referred to as *BGP-TE* further on) under the same strategies for the Pan-European (COST 266) network.

Fig. 4.6(a) illustrates the results for the time-based strategies. It is interesting to observe that the value of the Update Interval (UI) has a negligible effect on the performance of the *Enhanced BGP* scheme. Under synchronous updates all paths get updated within very short interval (the advertisements of the Enhanced BGP are not affected by the MRITs), and the source nodes change their paths almost simultaneously. This results in the same process of route flapping, as observed under the *BGP-TE* scheme. However, since nodes have multiple paths to choose from and the used UI are still relatively long compared to the average connection duration, the negative effects of the route flapping are almost negligible. Hence the performance is not affected. Eliminating synchronization brings much more significant improvement of the performance, because eliminated synchronization combined with decreased path-dependency leads to much more effective load balancing due to slower changes in the load of the multi-domain links. This results in lower blocking of requests.

Fig. 4.6(b) illustrates the results for the load-based strategies. As expected, the more often the TE state of the paths is checked and updated, the better the performances of both evaluated protocol extensions are, with the *Enhanced BGP* outperforming the simple *BGP-TE* extension. Under the load-based update strategy the performance of the *Enhanced BGP* is improved by more frequent state updates, whereas under the time-based scheme it was not affected by the value of the UI. This is due to the differences in the operation of the update schemes. The load-based update schemes lead to better load balancing because BGP re-convergence (i.e. path changes) are triggered only if needed. Checking the load of the links more often leads to more up-to-date TE information distribution and thus, more effective load balancing. Under the time-based update scheme triggering more frequent update of paths only leads to simultaneous change in the set of used paths, i.e. to more frequent route flapping.

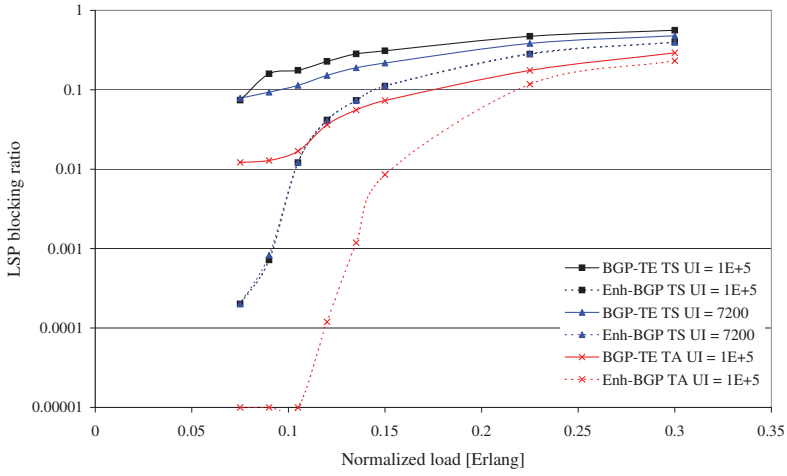
The presented results indicate that for the *Enhanced BGP* protocol

the performance under the time-based update scheme is improved more significantly by eliminating synchronization, whereas with the load-based update scheme, the performance is more significantly improved by performing more frequent TE state checks. The *BGP-TE* protocol, on the other hand, experiences almost equal performance improvement by eliminating synchronization and by applying frequent TE state updates under the link-load update strategies. This is due to the path-dependency problem, which is still present in the *BGP-TE* protocol.

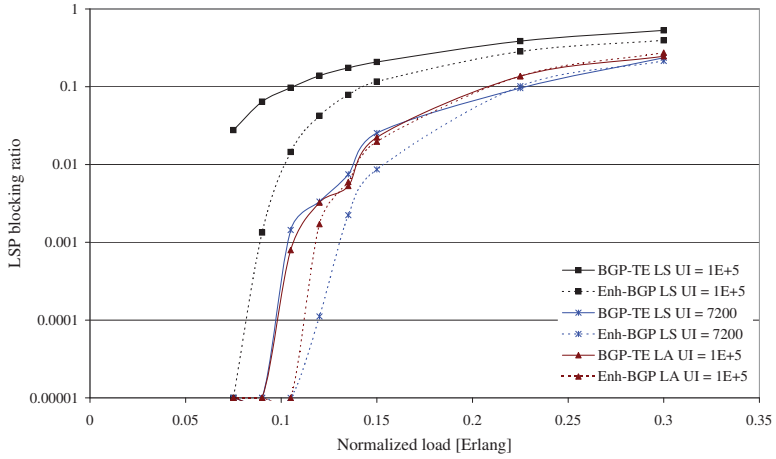
Both protocol extensions bring performance enhancements in terms of lower LSP blocking by applying load-balancing and by distributing more evenly the load on the multi-domain links. However, the *Enhanced BGP* performs much better because of the eliminated path-dependency which contributes to much more efficient load balancing and thus, lower blocking ratio.

Fig. 4.7 illustrates the BGP overhead during TE state update. The results for the amount of updates with UI = 7200 seconds for both the *BGP-TE* and the *Enhanced BGP* are not shown. For the *BGP-TE* the total overhead is above $9 * 10^5$ (as presented earlier in Chapter 3) and for the *Enhanced BGP* it is a little below $3 * 10^5$, but for clarity they are not depicted. As expected, the *Enhanced BGP* introduces lower overhead for the time-based schemes. This is due to the decreased path dependency and eliminated chattiness between the BGP peers. For the load-based schemes the *Enhanced BGP* results in higher protocol overhead because there are more paths to be updated in the network per update cycle. Thus, there is a clear tradeoff between the achieved lower connection blocking and the generated protocol overhead with the load-based schemes.

An important result is the lack of dependency on the network load for the *Enhanced BGP* protocol under the load-based strategies. This indicates protocol stability and predictable protocol behavior which is very helpful for the proper dimensioning of the CP.

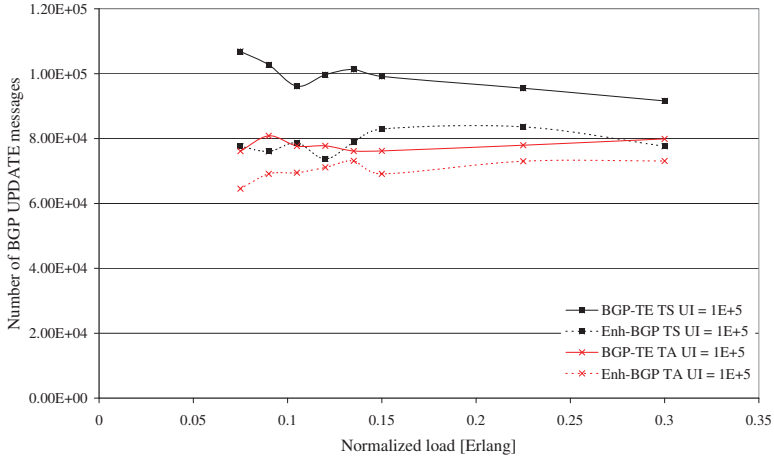


(a) Time-based strategies.

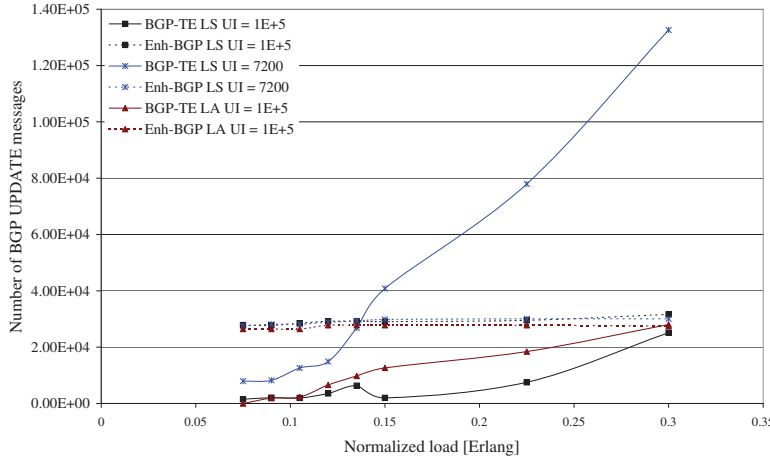


(b) Load-based strategies.

Figure 4.6: LSP blocking ratio vs. Normalized load for different TE-state update strategies.



(a) Time-based strategies.



(b) Load-based strategies.

Figure 4.7: Number of BGP UPDATE messages vs. Normalized load.

Export policies

The export policies for the *Enhanced BGP* regulate the amount of paths to be distributed. It is clear that flooding every possible path in a large topology such as the Pan-European one is not feasible due to the huge overhead created in the network. Here, the efficiency of the six policies presented in Section 4.2.3 is evaluated in terms of achieved LSP blocking ratio and the ability to provide AS-disjoint pairs of paths to destinations. Furthermore, the created overhead for distributing the paths (i.e. creating the VTs) is examined for the different strategies.

Fig. 4.8 shows the distribution of source/destination pairs with and without AS-disjoint paths in their VT databases. The results are grouped by the ability of the policy to provide AS-disjoint paths in five categories:

i) pairs without AS-disjoint paths;

ii) pairs with only one pair of AS-disjoint paths;

iii) pairs with only two pairs of AS-disjoint paths;

iv) pairs with only three pairs of AS-disjoint paths;

v) pairs with four or more pairs of AS-disjoint paths.

The policies which provide the least cases where no AS-disjoint paths could be found and the most cases where more than three AS-disjoint paths could be found per destination, are the *inclusive* policies. On the other hand, the worst policies are the *exclusive* ones, providing the least amount of AS-disjoint paths to destinations. In fact, almost no edge nodes get AS-disjoint pair of paths under *Policy 3 - exclusive* and *Policy 4 - exclusive*. This is due to the strong restriction of the policies - paths are distributed to neighbors only if they are AS-disjoint or Border node disjoint to all previously distributed paths. Considering the Pan European topology, which possesses a clearly established core (in Germany), the ability to provide such set of paths is very limited.

Fig. 4.9 depicts the amount of BGP UPDATE messages needed for the path distribution in the network and the average amount of paths per destination in all edge nodes. As expected, the *inclusive* policies provide the largest amount of paths per destination, but this comes at the expense of a huge protocol overhead. Nevertheless, the *inclusive*

policies seem to be more appropriate for survivability support, since they provide many AS-disjoint pairs of paths, and for load balancing, since they provide the biggest amount of paths per destination.

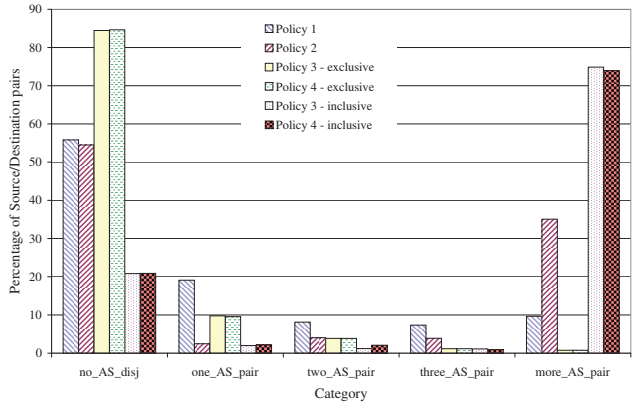


Figure 4.8: Percentage of source/destination pairs with and without AS-disjoint paths between them.

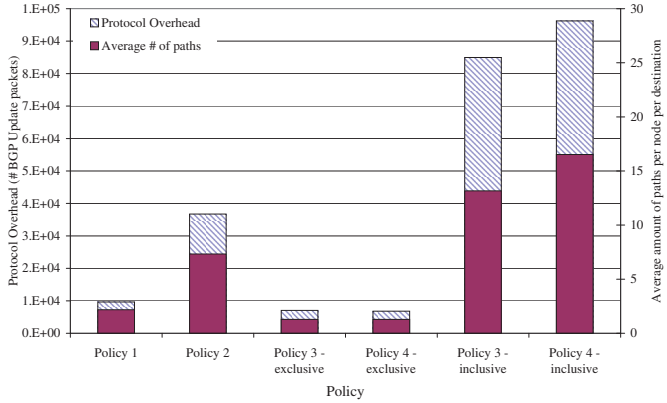
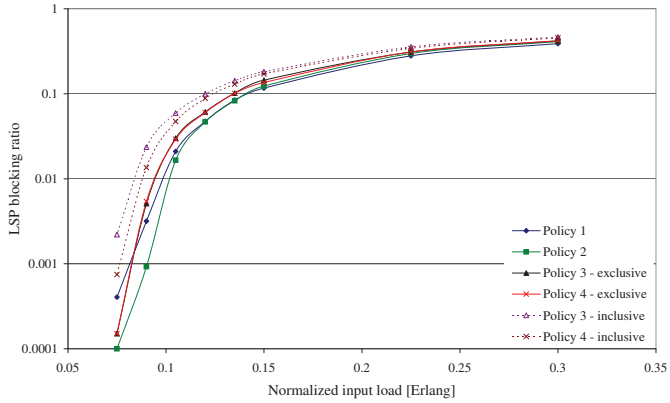


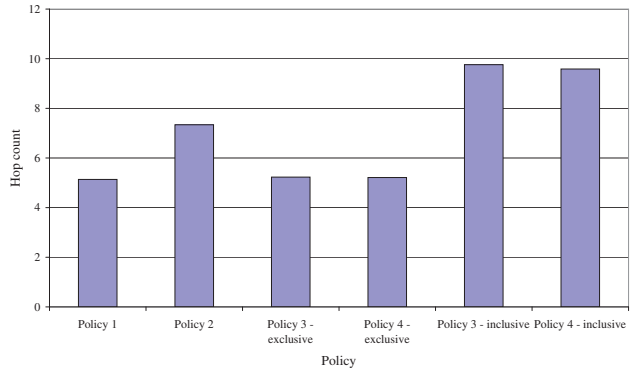
Figure 4.9: Protocol overhead and average number of paths per destination per node for the tested policies.

Fig. 4.10(a) illustrates the LSP blocking ratio under failure-free network operation for the different policies. The TE state of the entries in

the VTs are updated every $5*10^4$ seconds and at time of LSP request the entry with the best TE state is chosen. Surprisingly, the *inclusive* policies have the worst performance, even though they provide on average more than 20 paths per destination among which nodes can choose for load balancing. The reason for that performance is the average length of the provided paths, depicted on Fig. 4.10(b). Longer paths increase the probability for resource contention, which results in higher blocking of the LSP requests under the *inclusive* policies.



(a) LSP blocking ratio vs. Normalized input load for all tested export policies.



(b) Average path length in hops for all tested export policies.

Figure 4.10: LSP blocking ratio and average path length for all export policies.

4.4 Conclusion

This chapter presents novel BGP extensions for multi-domain TE in GMPLS-controlled networks. The extensions are designed explicitly to solve four major drawbacks the standard BGP experiences when applied to connection-oriented networks: lack of TE information dissemination, path dependency, lack of path diversity and slow protocol convergence. The efficiency of the proposal is illustrated by means of simulation results, which clearly show that the *Enhanced BGP* protocol provides lower LSP blocking and lower protocol overhead compared to the simple TE extension presented in Chapter 3. Furthermore, the proposed *Enhanced BGP* shows stable operation, ability to support different survivability mechanisms by providing AS-disjoint paths, and low dependency of the created protocol overhead from the network load. By providing multiple paths per destination and by decreasing the path dependency, the *Enhanced BGP* protocol achieves good load balancing and improved inter-domain link utilization.

Extending the functionality of the BGP protocol comes at a certain cost. Careful considerations must be made in order to maintain the scalability of the protocol by applying efficient export policies. They must be designed taking into consideration the overall multi-domain topology and connectivity in the network. Results show that there is a trade-off between the number of distributed paths and their quality. This implies that efficient multi-domain TE and QoS provisioning cannot be based solely on bi-lateral agreements between domains. Global coordination is required in order to achieve TE across multiple transport networks.

Chapter 5

BGP Enhancement for AS-disjoint Path Selection

5.1 Introduction

The routing requirements, that were taken into consideration when the Border Gateway Protocol (BGP) protocol was designed, and the requirements of the next generation transport networks are very different. QoS support, reliability requirements and adequate support for a dynamic network environment have not been envisioned as requirements of the BGP protocol. In next generation networks though, these requirements are paramount and necessitate extended information exchange between domains. Optical networks can transport a lot of information per second, hence failures affect the performance of the network heavily. Each domain should have enough information to adequately react to failures but the current version of BGP does not provide such information.

In this chapter, an extension of the BGP protocol which allows for the selection of two AS-disjoint paths per destination, is proposed and evaluated. Two paths are AS-disjoint when they do not share any AS (domain), except the source and the destination domains. Obtaining two disjoint paths per destination is beneficial not only for survivability, but also for load balancing and network performance enhancement, in case of failures when no resiliency mechanisms are applied. The operation of the mechanism, as well as the benefits of having two disjoint AS-paths for enhanced network performance under single inter-domain link failure,

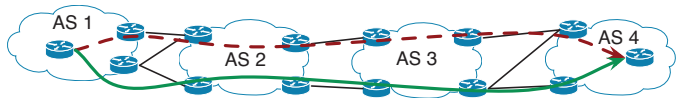
are illustrated by means of simulations.

The chapter is organized as follows: Section 5.2 and Section 5.3 outline the problem and the related work in the area respectively. Section 5.4 gives details on the proposed BGP extension. Section 5.5 focuses on the potential performance enhancements of the suggested BGP modification in multi-domain network environment. Section 5.6 presents the simulation case study and the obtained results. Conclusions are drawn in Section 5.7.

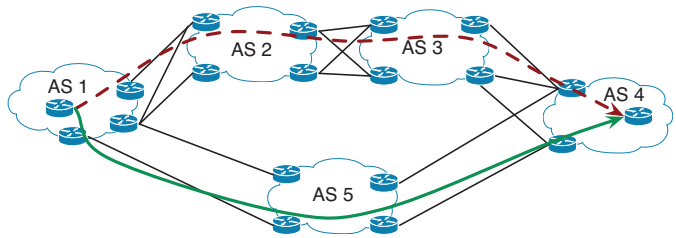
5.2 Disjoint path computation in multi-domain networks

In some scenarios, especially multi-AS ones, it is not possible to obtain the complete graph of the network without flooding the network with sensitive information, which is unacceptable because of the strong privacy protection policies between the Autonomous Systems (ASs). Thus, it is impossible to apply standard methods for disjoint path computation using the complete network graph such as the Suurballe algorithm [70].

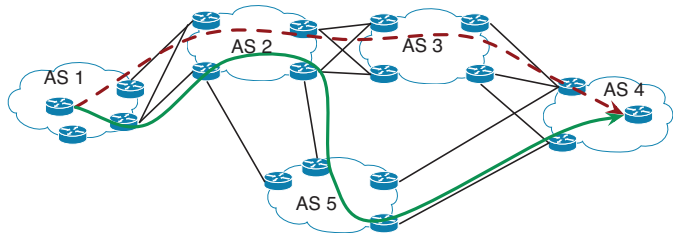
The methods for solving the multi-AS disjoint path computation problem can be divided in three categories (see Fig. 5.1). The first category uses standard BGP information or manual configuration to find the AS-path to a destination and tries to compute two disjoint paths along the obtained AS-path, i.e. it shares ASs. Solutions from this category necessitate sharing of information between the domains or employing new protocols as in the PCE architecture [36], or new protocol extensions such as the Primary Path Route Object (PPRO) [62] and the Associated Route Object (ARO) [61] for the RSVP-TE protocol. A drawback is that the applied optimizations in these mechanisms can be done only within one AS-path, and this limits their efficiency. Furthermore, an AS failure or disconnection between two ASs on the path cannot be recovered using such approaches. The second category of solutions provides two AS-disjoint paths between the source and the destination nodes. Therefore, two Label Switched Paths (LSPs) can be established via standard signalling protocols along each AS-path. The third category provides partially AS-disjoint paths, i.e. only part of the paths share the same ASs. This type of solution requires sharing of more information in order to obtain the disjoint AS-paths and does not provide the same level of



(a) Disjoint paths using the same AS-path.



(b) Disjoint path using disjoint AS-paths.



(c) Mixed solution.

Figure 5.1: AS-disjoint path solution categories.

protection against different types of failures as the solutions from the second category.

This chapter proposes a solution within the second category. The designed algorithm for AS-disjoint path selection[1] uses BGP extensions. The mechanism provides two AS-disjoint policy-compliant paths per reachable destination in a multi-domain network where ASs are multi-homed. Using the proposed solution in connection-oriented networks does not exclude the use of other advanced schemes for optimal path computation, such as e.g. the PCE approach. BGP and PCE are completely inter-operable since PCE elements need an AS-path in order to calculate an optimal path for each LSP request. Furthermore, calculating two disjoint paths using different AS-paths reduces the complexity of

[1] Note that the proposed BGP extension is not used for path computation, but for AS-disjoint path selection.

the joint path computation process since the complex disjointness calculations need to be performed only for the source and the destination domains. The proposed modified BGP is used as a complementary routing protocol which provides a higher-level path specified by AS numbers, whereas PCE can be used internally in each domain for optimal path computation.

5.3 Related work

Several proposals for BGP modifications for multi-path dissemination can be found in the literature. Kushman et al. [71] have proposed R-BGP, a modification of BGP that allows the sending of alternative paths per destination to downstream neighbors. Two BGP peers exchange the standard AS_PATH attribute and an alternative AS_PATH attribute, called a *failover path*. Using this solution the number of lost packets decreases significantly during BGP re-convergence. There are two drawbacks though. First, the forwarding process in all involved routers must be changed, because a decision of which path (normal or failover) to be used is taken on a per packet basis. Second, in order to use the failover path an extra BGP session via the failover path must be established with the neighbor. There are several differences between R-BGP and the solution proposed in this chapter. First, R-BGP is designed for packet-switched networks, whereas the proposal presented here is for connection-oriented networks. In GMPLS networks source routing is a fundamental feature, thus the head-ends of all connections require an extended view of the overall path, not only of the next hop. Second, the Kushman et al. [71] proposal supports protection only against link failures between neighboring domains. The proposed algorithm in this chapter is more general and offers survivability support for link failure, node failure and even entire AS failure.

Other proposals, as Bhatia et al. [72] or Walton et al. [73], try to eliminate the BGP route oscillations by sending extra information in the BGP UPDATE messages. Walton proposes to send several paths, not necessarily disjoint, for the same destination using a *Path Identifier* attribute. Bhatia proposes to use *Multiple-Hop* capability to report more than one Next-Hop for the same reachable destination to a BGP peer. The goal of these proposals is to reduce or eliminate the well-known BGP

route oscillations. The proposed BGP enhancement in this chapter does not seek to eliminate route oscillation during BGP re-convergence, but rather to eliminate (or minimize) the effect of the oscilations on the operation of the network.

5.4 Obtaining disjoint AS-paths with BGP

BGP is a path vector protocol, i.e. the created routing table contains the destination, the next hop towards the destination and the sequence of ASs to reach the destination. These are distributed via BGP UPDATE messages between BGP peers. Interior routing information is not shared across domain borders via BGP, so different ASs have no interior information about other ASs. Limiting the shared information is done for scalability purposes as well as to avoid disclosure of sensitive information to other domains. Thus, the received BGP information by an AS is aggregated as much as possible and only one path to a destination is chosen and further distributed to other BGP peers[2]. BGP peers choose paths according to a special decision procedure described in RFC 4271 [37]. Further on in this chapter, paths chosen under the standard BGP operation are referred to as *primary AS-paths*.

The proposed BGP enhancement is a concurrent modified BGP decision procedure which obtains a disjoint AS-path to the primary AS-path, referred to as *backup AS-path*. This backup path can be used for resilience purposes, load balancing or routing of LSP requests during BGP protocol re-convergence. The proposal necessitates three new Routing Information Bases (RIBs)[3]: Adj-RIB-Disj-In, Loc-RIB-Disj and Adj-RIB-Disj-Out. In practice a backup route can be identified by a flag in the existing RIBs. The proposed modified BGP decision procedure comprises of three phases as follows:

1. When a BGP entity receives an UPDATE message for a backup AS_path from a peer, the route is added to the Adj-RIB-Disj-In and a preference is assigned. Upon a route addition or change in the Adj-RIB-In or Adj-RIB-Disj-In, phase two is triggered.

[2] Note that aggregation of destinations is a common practice in BGP, in which case only one path is distributed per aggregated destination [37].

[3] See Fig. 2.13(b) for reference.

2. For the destination under update, the best route disjoint to the one in the Loc-RIB from all available routes to that destination in all Adj-RIB-In and Adj-RIB-Disj-In is selected. The selected route is included in the Loc-RIB-Disj, which keeps all the BGP backup routes used locally. If this implies a route change in Loc-RIB-Disj, apply phase 3.

3. After a change in the Loc-RIB-Disj, the new route undergoes a policy filtering process and is included in the selected Adj-RIB-Disj-Out. UPDATE messages are sent further.

Under a few scenarios it is possible that a BGP speaker cannot obtain disjoint routes to a destination, e.g. there is a trap in the topology of ASs, or the existing disjoint path is not policy-compliant. This can be solved by changing the primary path by manual configuration of the local preferences or by adjusting the local policies.

Note that the received backup routes are in Adj-RIB-Disj-In not in Adj-RIB-In and thus, they are not selectable as primary routes by the normal BGP decision procedure. This is a desirable behavior since backup routes might create loops if the usual hop-by-hop routing of the standard BGP is used. Due to the specifics of the proposed algorithm and the hop-by-hop routing paradigm, enforced by the BGP protocol, source routing is needed in order to use the backup paths [7]. This is necessary because in some cases two neighboring ASes choose each other as next-hop for their backup paths and in other cases they choose other neighboring ASes. This is topology dependent and cannot be predicted. Thus, in order to use the backup path, the responsible border node must apply source routing for forwarding LSP requests on that path. Possible solutions for source routing on the inter-AS level with BGP are proposed in [7] and [6] (see Chapter 4).

Considering the operation of the proposed BGP enhancements, the scalability of the protocol is not seriously affected. The amount of stored data is at most twice the amount of data stored in BGP speakers under standard BGP operation, since there are only two paths per destination.

Fig. 5.2 shows an example of how the proposed BGP enhancement works. Fig. 5.2 a) shows the RIB's content for destinations in AS 5 in all other ASs and the selected AS_PATH (marked with solid line). These paths are obtained using the standard BGP process. As Fig. 5.2 b)

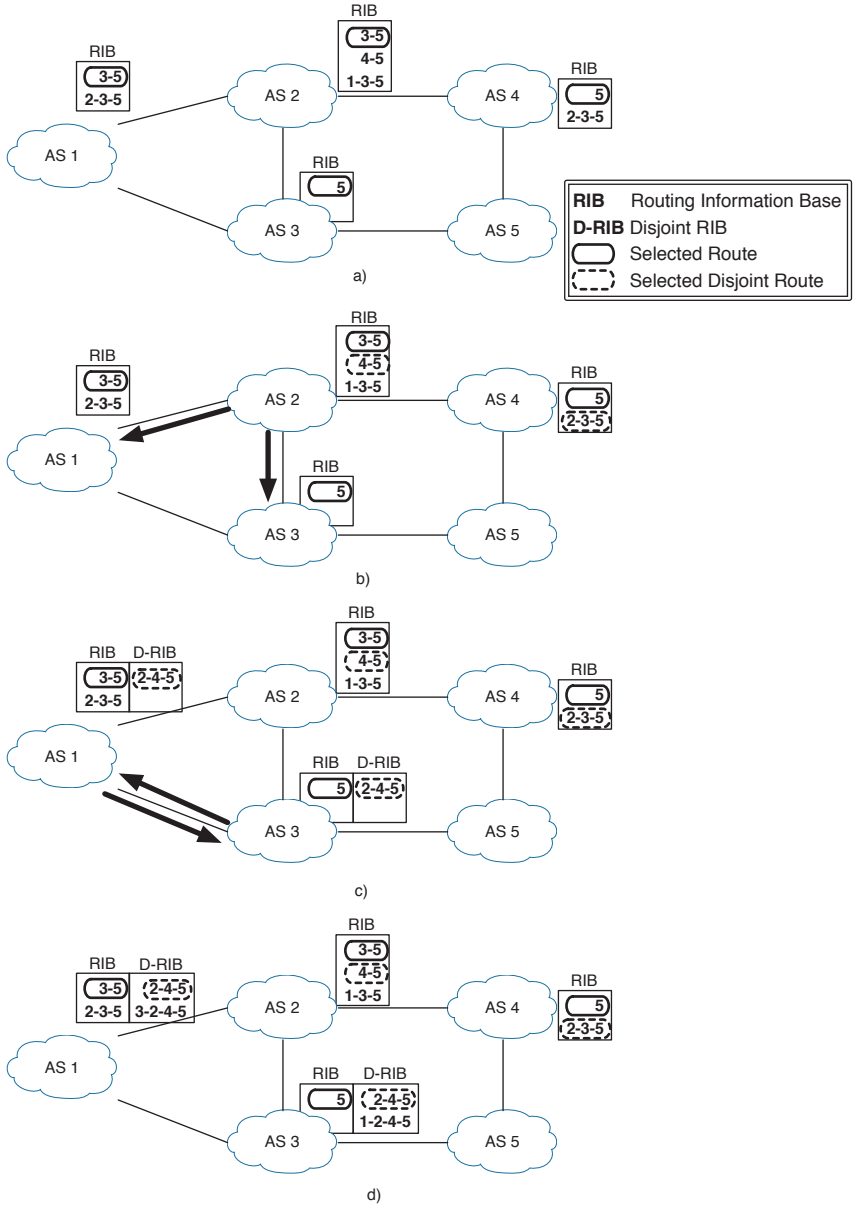


Figure 5.2: Modified BGP example.

shows, all ASs which can select a disjoint AS-path to the primary path among all available paths in their Adj-RIB-Ins do that (paths marked with dashed line), and send an UPDATE message with the disjoint AS_PATH attribute to their peers (solid arrows). ASs which receive new disjoint AS_PATH attribute include it in the Adj-RIB-Disj-In and the selection mechanism is activated (Fig. 5.2 c)). The new selected route is sent further until all ASs have a disjoint AS-path, as shown in Fig. 5.2 d).

5.5 Network performance enhancement

In this work, three performance aspects are considered. First, an analysis of the benefit of having two disjoint paths per destination, with respect to the loss of traffic, is done. Since the BGP protocol re-convergence takes significant time, this results in high loss of traffic on existing connections and thus, in degraded network performance. Then, the effect of applying connection restoration for the affected LSPs is evaluated. Utilizing the disjoint paths for restoring the affected connections at the time of failure, without being affected by the BGP re-convergence delay or route oscillations, is demonstrated. The third analyzed aspect is the performance of the dynamic multi-domain network in terms of blocking of future connection requests under BGP protocol re-convergence, when no resiliency mechanisms have been applied. During the BGP re-convergence, some nodes loose visibility of destinations or loops are created. This increases the LSP connection request blocking.

5.5.1 Failure recovery

There are two main approaches for providing resilience in a network - protection and restoration. Protection is the process of establishing a backup path, disjoint to the primary path, before the failure occurs. Restoration is the process of re-establishing an affected connection after failure using an alternative path. Depending on the applied survivability mechanism, different procedures are applied in case of a failure. In general, the total recovery time for protection and for restoration can be given by the following approximate equations [74]:

$$T_{recovery}^{Protection} = T_d + T_n + T_{sw}, \tag{5.1}$$

$$T_{recovery}^{Restoration} = T_d + T_n + T_{setup} + T_{sw}, \tag{5.2}$$

where T_d is the time to detect the failure, T_n is the time to notify the node, responsible for failure recovery, T_{setup} is the time to set up the backup LSP and T_{sw} is the time to switch the traffic over to the backup path.

Improving network performance under link failure can be achieved by minimizing the time to recover from the failure. In case of protection, the time to recover is typically much shorter than in case of restoration because in the latter case, the node initiating the backup path setup must have an alternative path. If no pre-computed disjoint paths are available, the T_{setup} will typically include the BGP re-convergence time. If an AS-disjoint alternative path is available, the T_{setup} can be drastically reduced. Thus, in both cases (protection and restoration) the availability of AS-disjoint primary and backup paths is clearly beneficial for resilience support in the multi-domain network.

5.5.2 Failure notification strategies

In case of a link failure it is important to inform the correct network elements in order to minimize the impact of the failure on the network performance. In a multi-domain scenario there is still no consensus whether a failure should be signalled all the way to the head-end of a connection or if it should be handled locally. For single domain operation the head-end of the connection decides the protection method [75]. For multi-domain networks though this is unclear due to the diverse policies applied in the ASs and their capabilities for survivability support. In order to evaluate this, the enhancement of the BGP protocol proposed in this chapter is applied and the blocking ratio of connection requests after an inter-domain link failure is analyzed using the following notification strategies:

- *No notification*: The BGP protocol re-converges without notifying anybody of the failure. All LSP requests which cannot be routed due to lack of visibility or routing loops in this period are dropped.

- *Local notification*: Only the border nodes of the domains which detect the failure are notified. The border nodes route the upcoming LSP requests using the backup paths obtained by the proposed BGP enhancement. If a routing loop occurs (in case a domain uses its upstream neighbor for the backup path) the requests are dropped at the upstream node[4]. No BGP re-convergence is done.

- *Head-end notification*: The head-ends of the connections are notified that they must use their corresponding backup paths obtained using the proposed extensions. In this case no routing loops are possible and LSP blocking occurs only due to lack of resources. No BGP re-convergence is done.

- *Mixed strategies*: The LSP requests are routed on the backup paths during the BGP protocol re-convergence (using either the Head-end or the Local notification) and when the BGP protocol converges, the subsequent connection requests are routed on the new primary paths.

The actual failure notification can be performed in several ways, e.g. by the RSVP-TE Notify message or by extending the BGP Keep Alive messages. Here, the RSVP-TE Notify message is employed, which is used to propagate a list of affected destinations in case of a link failure. The scope of propagation of the message depends on the applied notification procedure as described above.

5.6 Simulation case-study

In order to evaluate the efficiency of the proposal, the BGP extensions were implemented in the event driven simulator tool OPNET [68]. The behavior of the protocol was analyzed for two different topologies: a general mesh topology (*Topology 2* Fig. B.2) and a Pan-European topology (Fig. B.4) from the COST 266 project [69]. For the Pan-European topology each country represents a domain, the intra-domain topologies of the separate domains are randomly generated and have no more than 4 nodes acting as sources/destinations. In total there are 46 source/destination nodes in 22 domains interconnected via 42 inter-domain links.

[4]Due to loop-detection mechanism within the RSVP-TE implementation.

Hop count is used as TE metric for the intra-domain routing protocol. No specific import/export policies are applied, i.e. it is assumed that all domains offer transit services to all their neighbors. The values for the Minimum Route Advertisement Interval Timers (MRITs) are set according to the specification in [37], i.e. 30 seconds for eBGP and 5 seconds for iBGP.

Two different sets of simulations are presented. First, the outcome of the extended BGP protocol, illustrating its ability to provide AS-disjoint paths, is shown. The second set of simulations shows the benefits of applying the proposed BGP protocol extension in case of a single inter-domain link failure in a dynamic multi-domain connection-oriented network. For these simulations the settings presented in Table 5.1 are used. The wavelength continuity constraint is assumed. For LSP signaling the RSVP-TE protocol is used. The wavelength assignment at the destination node is random among all free wavelengths along the path.

Average connection Duration:	600 seconds		
Load Range	High	Medium	Low
MIT*	20 sec.	40 sec.	80 sec.
Wavelengths	30	40	40
Resulting Load per node	1 Erlang	0.375 Erlang	0.1875 Erlang

* Mean Inter-arrival time for LSP requests.

Table 5.1: Simulation settings.

5.6.1 AS-disjoint path computation

Fig. 5.3 depicts the selected paths for two randomly selected pairs of nodes in *Topology 2* (Fig. B.2). Using the AS-disjoint scheme all source/destination pairs find AS-disjoint paths. For the presented result the primary path from domain 3 to 7 is: $3 \rightarrow 5 \rightarrow 7$ and the disjoint path is: $3 \rightarrow 2 \rightarrow 4 \rightarrow 7$. For the second case (domain 4 to 10) the primary path is: $4 \rightarrow 2 \rightarrow 3 \rightarrow 9 \rightarrow 10$ and the disjoint one is: $4 \rightarrow 7 \rightarrow 8 \rightarrow 11 \rightarrow 10$.

For the COST 266 topology though, the proposed mechanism does not provide all source/destination pairs with AS-disjoint paths. 24.25% of the sourde/destination pairs cannot obtain AS-disjoint paths. This is due to the inherent path dependency of the BGP protocol (see comments

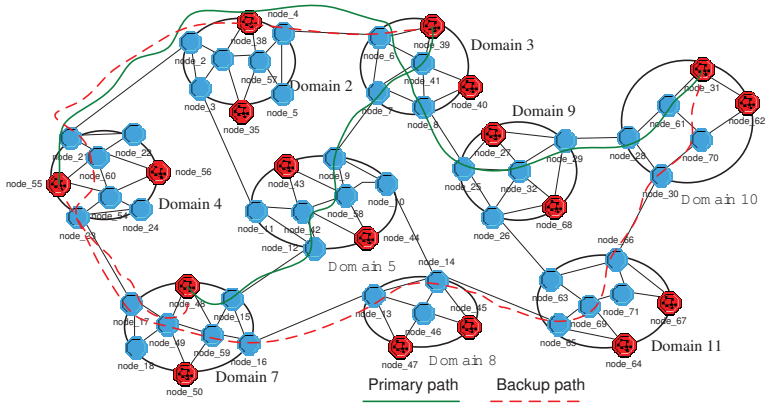


Figure 5.3: Illustration of selected AS-disjoint paths.

through Chapter 3 and Chapter 4). Furthermore, the standard BGP path selection applies tie-breaking rules which are not TE-oriented, such as "lower router ID", "lowest neighbor IP address" and "oldest path" and as a result, trap topologies are created.

5.6.2 Traffic loss under BGP re-convergence

Here, the effect on the loss of traffic on existing connections under BGP re-convergence for the Pan European topology is evaluated. The performance is evaluated under high input network load (46 Erlang) and the performance metric is the BGP re-convergence time. Without having pre-established or pre-computed backup paths, during this period, source nodes cannot re-establish the affected LSPs and the traffic on them is lost. The re-convergence times of the BGP protocol in 13 different randomly chosen inter-domain links failures are presented on Fig. 5.4. The lost traffic under BGP re-convergence is proportional to the time to re-converge the BGP protocol plus the time to re-establish the affected connections on the new paths. Since the time to re-converge is the dominating factor the time required for re-establishing the affected connections is not considered. Each connection has a capacity of 10 Gb/s.

From the results it can be seen that for the failed links the convergence time varies between 8 and 20 minutes. Considering the bit rate of the affected connections and the number of affected LSPs, a link failure

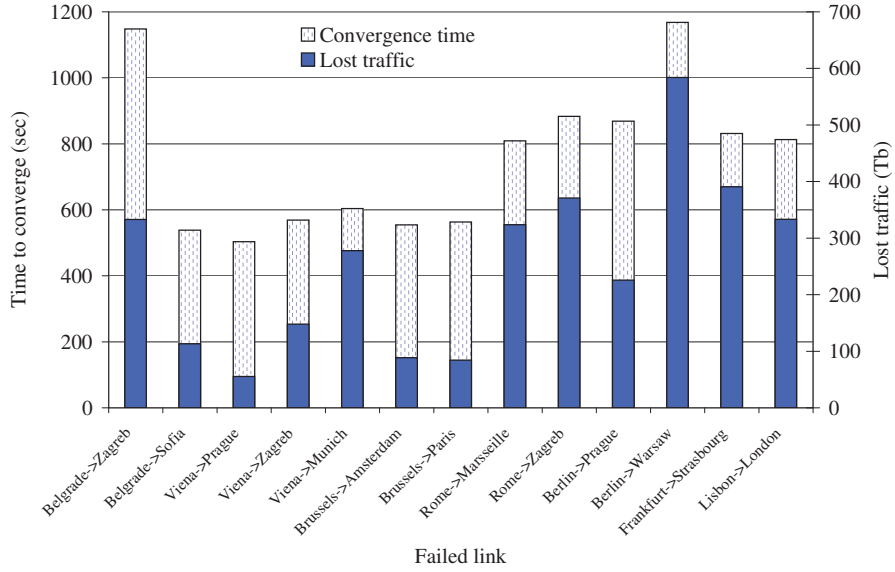


Figure 5.4: BGP re-convergence time and corresponding loss of traffic for 13 link failures.

results in loss of traffic between approximately 55 and 580 Tera bits. This shows the importance of applying a survivability mechanism in the network. If pre-selected backup paths were available (for protection or restoration), this could save a lot of revenue for the network operators. Using the proposed mechanism for deriving AS-disjoint backup paths can significantly decrease the amount of lost traffic in the network, since the source nodes do not have to wait for the BGP protocol to re-converge.

5.6.3 Failure notification for future LSP requests

Here, the importance of a proper failure notification method for reducing blocking of LSP requests, under BGP re-convergence, is evaluated for the Pan-European topology. The first tested case is a failure of the link, experiencing the highest traffic loss in the previous experiment (i.e. Berlin - Warsaw, see Fig. 5.4). The considered traffic conditions are medium and low load, where traffic is generated only between the nodes which have both primary and backup paths. The results for the LSP blocking ratio using the different notification strategies are presented on Fig. 5.5.

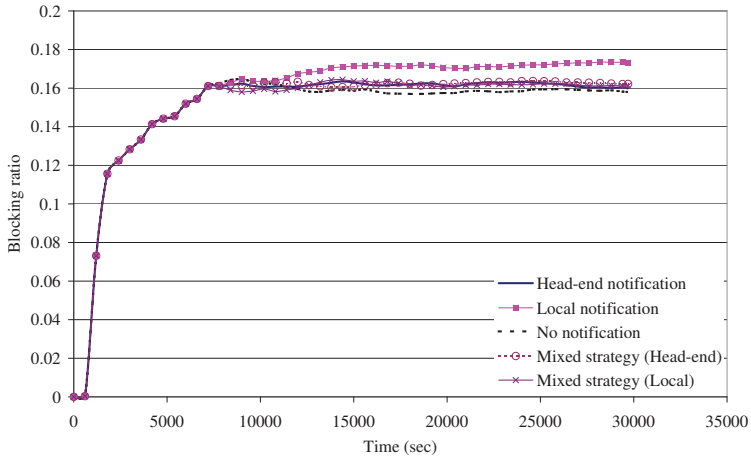
As can be seen, the LSP blocking ratio for the whole network is the highest for the *Local notification* strategy. This implies that the strategy of preserving the failure information locally is not always the best choice. Applying the *Local notification* scheme may yield longer paths for the LSPs, which will result in increased blocking probability. For the medium loaded network, the remaining strategies perform almost equally well. This is due to the fact that under more loaded conditions, the LSP blocking is dominated by the lack of resources. Thus, the difference between the schemes is difficult to observe. At the low loads though, the *Mixed strategy (Head-end)* performs the best. Under this scheme, the LSP requests are routed from the Head-end onto their backup paths during the BGP re-convergence, and onto their new primary paths after re-convergence. The achieved improvement, compared to the *Local notification* strategy, is about 50%.

Fig. 5.6 illustrates the blocking ratio of two flows[5] in the case of two different link failures for all tested notification mechanisms: flow England → Hungary with failed link Berlin - Warsaw and flow Poland → Greece with failed link Belgrade - Sofia. The mixed strategy is *Mixed strategy (Head-end)* and the network is under medium load. The goal is to investigate how different failures affect individual flows and how the tested notification schemes perform on a per flow basis.

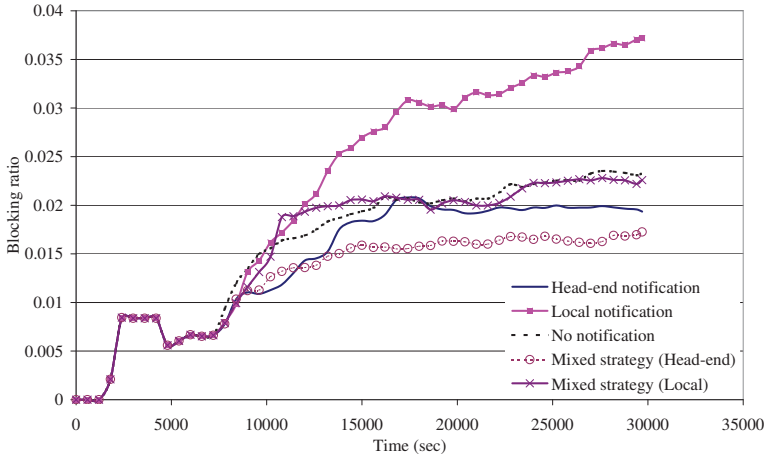
The first thing to be noted is that the tested strategies affect the blocking ratio of flows differently. For the first flow (England → Hungary) it can be seen that informing the head-end of the connections (England in this case) brings significant LSP blocking improvement (around 50% better than the *Local notification*). For the second flow though, the *Local notification* yields the lowest blocking ratio (around 30% better than the *Head-end notification*). This calls for the development of schemes, which handle affected flows in a differentiated manner.

The second interesting result is that the mixed strategy does not have the best performance, as could be expected. This is due to the fact that in the failure cases at hand the new paths for the affected source/destination pairs are different than the backup paths obtained before the failure and in these particular cases, they yield a higher blocking ratio. This implies that configuring the BGP routers of a certain domain taking

[5]Here a flow refers to all of connection requests between a fixed source/destination pair.

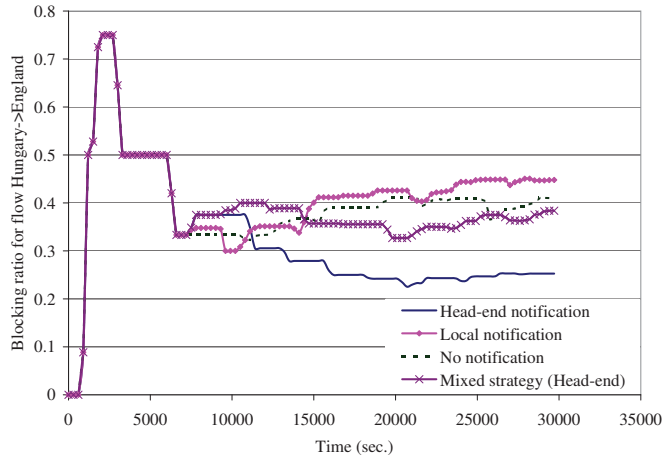


(a) Medium network load (17.25 Erlang).

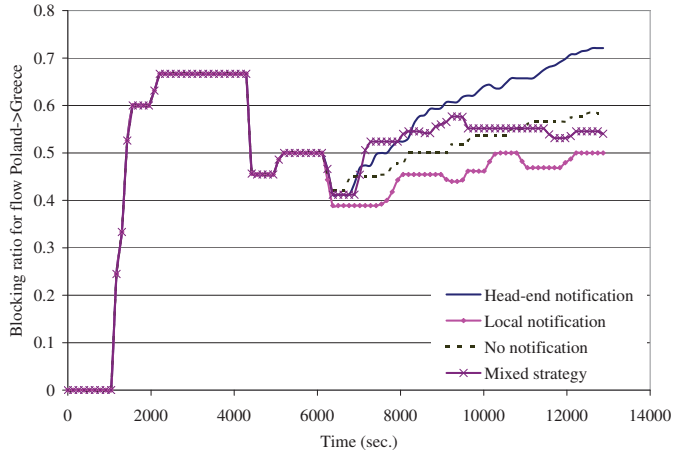


(b) Low network load (8.6 Erlang).

Figure 5.5: LSP blocking ratio for different notification strategies.



(a) Flow England → Hungary, failed link Berlin - Warsaw.



(b) Flow Poland → Greece failed link Belgrade - Sofia.

Figure 5.6: LSP blocking ratio for different notification strategies for two monitored flows.

only the bi-lateral agreements with the neighbors into account is not enough to obtain the best performance in a multi-domain environment. Global coordination is required in order to provide end-to-end QoS of connections crossing multiple domains.

In order to verify the result discussed above more flows in different failure cases were examined. The results are presented in Table 5.2 and illustrate clearly the need for differentiated failure handling. One of the most illustrative examples is the failure of link Amsterdam-Glasgow and the affected flows from Czech Republic and Belgium to England (see Fig. 5.7 to follow the discussion). For the same failure, the Local and the Head-end notifications have completely the opposite effect on the affected flows. Two observations can explain the result. First, the backup path used locally by Amsterdam is via London which is a very highly loaded link. Amsterdam-London is on the shortest path towards Portugal from Germany, Denmark and the rest of the northern countries. The backup path from the head-end nodes in Belgium is via France, where link Paris-London is less loaded. As a result, the Head-end notification is more efficient for the Belgium $\rightarrow$ England flow. Second, the backup path used from the head-end of the Czech Republic $\rightarrow$ England flow is very long, since Czech Republic is far away from England, thus the blocking probability for the flow is higher. In fact, it is higher than the blocking of the flow if Amsterdam re-routes the requests locally. Thus, the Local notification is more effective in this case. For the flow from Hungary to Italy with failed link Zagreb-Belgrade the higher blocking ratio of the Head-end notification is due to the fact that the backup path from head-end nodes in Hungary pass via Germany. This is an effective core of the topology and the probability for blocking there is very high. The local backup path for Zagreb is via Greece which is relatively less loaded area in the network. Thus, the Local notification is more efficient. In general, whenever the backup paths pass via more loaded links, the notification mechanisms perform poorer. A possible parallel can be drawn between the position of the failed link, the provided backup paths from the BGP enhancement and the efficiency of the notification strategies. The effect of the topology on the efficiency is thus significant.

Fig. 5.8 illustrates another relationship, which can be outlined - between the load on the failed link and the efficiency of the Local and the Head-end notification strategies. The more loaded the failed link

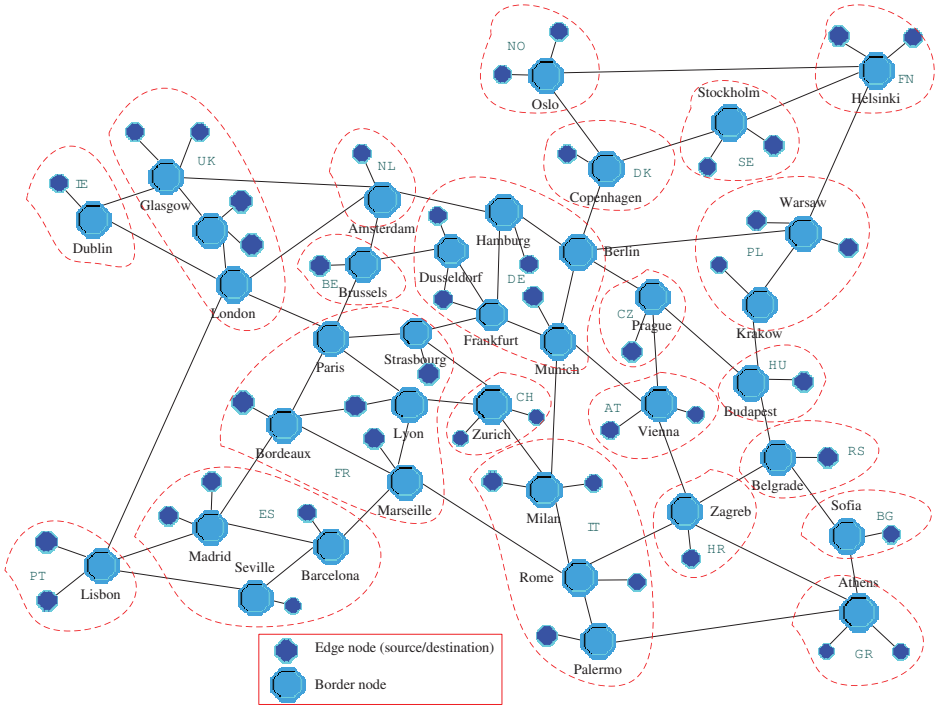


Figure 5.7: COST 266 Pan-European topology.

is, the less effective the Local notification is, compared to the Head-end notification. This is due to the fact that when many connection request flows are affected, re-routing them locally, possibly using the same inter-domain backup link, becomes more difficult. Furthermore, the used backup paths usually carry working traffic as well. Applying Head-end notification results in better load distribution, because different flows use different backup paths. Hence, the efficiency of the Head-end notification is higher than the efficiency of the Local notification at high loads of the failed links.

Failed link	Affected flow	Blocking ratio for the flow	
		Local notification	Head-end notification
Amsterdam - Glasgow	Czech Republic → England	0.45	0.52
	Belgium → England	0.23	0.19
Zagreb - Belgrade	Hungary → Italy	0.175	0.45
Bordeaux - Madrid	Holland → Spain	0.35	0.28
	Belgium → Spain	0.54	0.46
Strasbourg - Frankfurt	Czech Republic → Spain	0.67	0.64

Table 5.2: LSP blocking ratio for different flows and different failure cases.

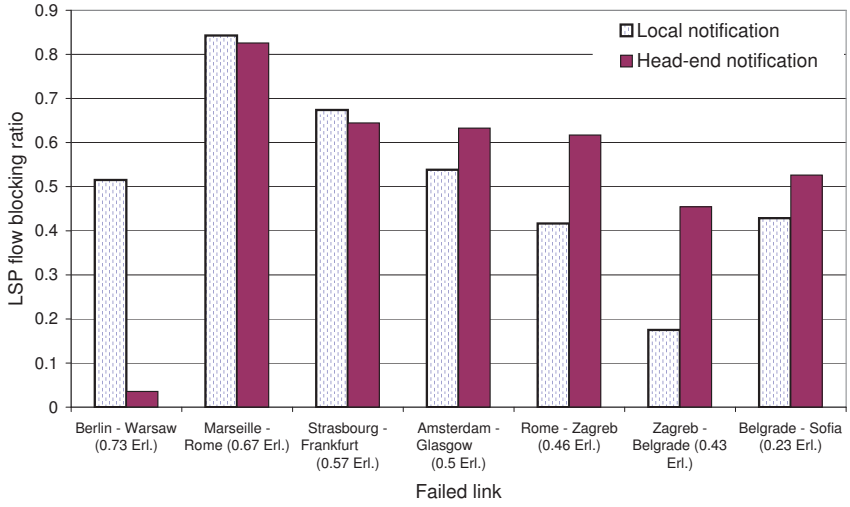


Figure 5.8: LSP flow blocking ratio vs. the load of a failed link for Local and Head-end notification strategies.

5.7 Conclusion

In this chapter an extension of the standard BGP protocol for obtaining AS-disjoint paths in multi-domain GMPLS networks is proposed. The main focus is on the potential benefits of applying the proposed mechanism for improving network performance in the case of inter-domain

link failures. The presented simulation results illustrate that employing AS-disjoint paths for reestablishing affected LSP connections after link failures can potentially save huge amounts of traffic. With BGP re-convergence times within tens of minutes, this implies a lot of saved revenue.

Furthermore, it is shown that applying a proper failure notification strategy can considerably lower the blocking ratio of new LSP requests. Different failures affect LSP request flows in different ways due to the position of the failed links along the affected LSP paths and the overall multi-domain topology and connectivity. This calls for the development of a framework for a differentiated approach to failure notification and handling. The presented results imply that the load condition of the failed link as well as its position in the network should be taken into account when deciding on the notification mechanism.

Chapter 6

Multi-Domain Restoration

6.1 Introduction

The new generation high speed networks are able to transport huge amounts of information, thus the consequences of a failure become more pronounced and network reliability becomes a key requirement. A crucial aspect in developing a fault management system is the routing and establishment of backup paths. This can be achieved either prior to failure (protection), where the connection backup paths are pre-established, or after a failure (restoration), where the backup paths are created and routed in reaction to network failures. Protection schemes require pre-planned and pre-established protection paths, which increase the needed capacity in the network, whereas restoration techniques do not require additional allocation of protection resources given that they rely on the existing infrastructure and use the available spare capacity pool [74].

The area of multi-domain survivability is relatively new with most of the work being focused on protection mechanisms. This chapter, on the other hand, explores the area of multi-domain restoration. In particular, inter-domain link failures are the main focus of the presented work. Section 6.2 presents a short overview of general survivability concepts, followed by a survey of the state of the art in the area of survivability in multi-domain networks. In Section 6.2.3, motivation for applying restoration in multi-domain networks is given, along with a detailed analysis of the most pending challenges in the area. Enhanced restoration techniques, based on two standard single-domain restoration methods, are

proposed in Section 6.3. Section 6.4 presents the simulation case study, results and discussions. Conclusions are drawn in Section 6.5.

6.2 Survivability in multi-domain networks

In this section a review of existing survivability schemes is presented. Then a survey of existing protection and restoration proposals for multi-domain networks is given. Based on this study, the main challenges and motivations for applying restoration mechanisms in multi-domain networks are presented.

6.2.1 Survivability in intra-domain networks

There is much literature on protection and restoration schemes in intra-domain network scenarios. A classical way to classify protection methods is the well-known N:M notation, where N is the number of backup paths used to protect M working paths [76]. Depending on how resources are assigned to a backup path, other classifications can be defined: i) the allocated resources can be dedicated to a working path, or shared by a set of working paths, where independent single failures are supposed, and ii) the resource allocation process can be carried out in advance (protection) or after the failure occurrence (restoration).

Three main recovery strategies exist (see Fig. 6.1): End-to-End (E2E), local (or span) and Local-to-End (L2E). Under E2E recovery, a Label Switched Path (LSP) is recovered from the head-end to the destination. This requires a link-disjoint path computation to be performed. Under local recovery, only the nodes adjacent to the failed link are involved in the recovery of the failure by establishing a local backup path between them. Under L2E recovery the upstream node, detecting a failure, recovers the connection from itself to the destination. With local and L2E strategies, no explicit link-disjoint path is required. These strategies are applicable in both protection and restoration mechanisms. Even though the efficiency of these strategies is very well-investigated in intra-domain scenarios, their applicability in multi-domain cases is not straightforward. Some suitable protection/restoration proposals for intra-domain failures cannot be directly applied to multi-domain networks due to certain constraints imposed by the applied local policies in the domains

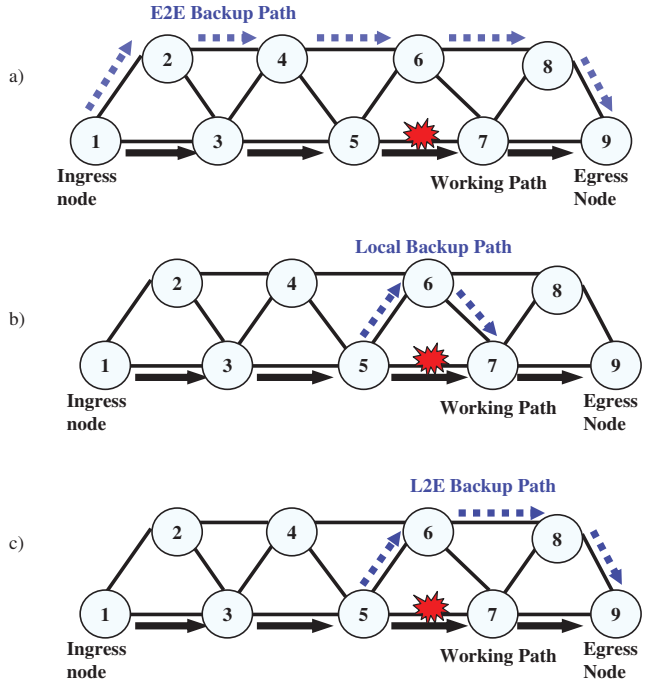


Figure 6.1: Basic recovery strategies.

and the specifics of the multi-domain environment. In the next sections challenges, solutions and novel proposals offering survivability in multi-domain networks are introduced.

6.2.2 Survivability in multi-domain networks

The lack of published research work on multi-domain resilience is mainly due to the assumption that a failure in a given domain should be handled within the domain [77]. Furthermore, it is also considered that the resilience mechanisms should not be different to those in the single domain case. However, this is only true in some scenarios according to [77], where a classification of the multi-domain resilience problems is presented. Three specific failure cases pose special challenges: inter-domain link failure, border node failure and a full domain failure. In these cases the survivability mechanism involves more than one domain,

which requires coordination between the domains, additional intelligence and protocol extensions.

In Table 6.2.2, a classification of the main multi-domain recovery proposals is presented. They are sorted into protection or restoration methods. One of the first protection proposals is from [78], where protection is independently developed in each domain by using end-to-end backup paths. These paths are merged at the domain boundaries building Inter-Domain Boundary Bypass Tunnels. In [79], both internal failures and a single gateway failure are protected by a combination of local and end-to-end pre-established protection paths. An extension of the Resource ReserVation Protocol with TE extensions (RSVP-TE) is included to support signaling of multiple ingress/egress points in a domain. In [80] and [81] new enhancements are presented allowing shared backups without flooding the information on working and backup paths. These techniques are based on p-cycles and multi-path routing. Once again, new extensions of RSVP-TE and the Path Computation Element (PCE) architecture are presented to support such new mechanisms.

However, there is less literature on restoration applied in multi-domain scenarios. In [77] the MPLS fast-rerouting is extended for the multi-domain scenario. In [82] a proposal addressing the limitation of information exchange among the domains allowing the utilizations of end-to-end restoration algorithms is presented. In [83] a new proposal to use p-cycles, enhancing some previous proposals (described in [82]) is introduced. The main contribution of this paper is the adaptation top-cycle computations based on a new topology aggregation model.

6.2.3 Motivation and challenges for multi-domain restoration

In this work restoration schemes are investigated, because they have certain advantages which could be very beneficial for multi-domain environments. The following can be outlined as the strongest motivation points for applying restoration in multi-domain networks:

- Fast recovery and suitable resource consumption: Providing backup protection paths results in fast failure recovery but is costly in terms of occupied bandwidth. Restoration, on the other hand, offers a suitable compromise between fast recovery and reasonable

Reference	Network Environment/ Technology	Failure Recovery Methods	Description and main contributions
[78]	MPLS	Protection	Independent protection mechanisms (end-to-end) within individual domains and merged at the domain boundaries (Inter-Domain Boundary Bypass Tunnel)
[79]	GMPLS / Optical	Protection	Internal failures and a single gateway failure; Path structure set up; New RSVP-TE object (GSRO)
[80]	GMPLS / Optical	Protection	End-to-end backup paths; Multi-domain p-cycles and multi-domain multipath routing.
[81]	GMPLS / Optical	Protection	P-cycles; Two sub-problems: higher level inter-domain protection and lower level intra-domain protection; Extensions of RSVP-TE and the PCE architecture
[77]	GMPLS / Optical	Restoration	Link failures; MPLS fast rerouting; Label stacking multi-domain.
[82]	None defined	Restoration	Shared backups; Multi-domain Information exchange; Topology aggregation.
[83]	Optical	Protection	P-cycles; New topology aggregation model

Table 6.1: Classification of main multi-domain recovery proposals.

resource consumption. For services, which do not require fast recovery, restoration techniques provide good compromise between cost and efficiency.

- Suitable in dynamic network scenarios: Restoration offers very high flexibility as it acts just after the failures in a dynamic way. Res-

toration is also the only way to handle unexpected and unplanned massive failures, natural disasters, cascade nodes failures, etc.

- Increased flexibility: Extending the applicability of restoration techniques across domain boundaries gives network operators the flexibility to offer different types of services to clients with different survivability requirements. Applying a combination of protection and restoration techniques for different classes of service is a beneficial strategy for every operator. Furthermore, implementing restoration techniques allows interconnecting domains with various survivability support capabilities. Some domains may not be capable of providing reliability services, whereas others may have a local policy of handling failures internally.

Some intra-domain protection/restoration proposals cannot be directly applied to multi-domain networks due to certain constraints imposed by the applied local policies in the domains. The main challenges to deploy multi-domain restoration are:

- Limited visibility of the topology: This problem comes from the limited visibility of the nodes regarding the multi-domain connectivity and the full network topology. The requirement for preservation of topological and state information within the domain is based on the strong confidentiality preservation policies between domains. This topology filtering has a very important implication - a border node in one domain would generally have information only about reachable destinations (edge nodes) in the other domains, but this border node has no information about core or other border nodes. Therefore, unless there are parallel links between border nodes, they are unable to apply local restoration techniques for failed inter-domain links.

- No standardized multi-domain routing: Currently there is no standard for multi-domain routing in GMPLS networks. Several approaches are being evaluated: the Border Gateway Protocol (BGP) [37], the PCE [36] and the External Network-to-Network Interface (E-NNI) routing specification (only for intra-carrier application) [28]. Applying BGP requires protocol re-convergence after failure in order to obtain the restoration path, since BGP provides only one

path per destination. Furthermore, a major challenge with BGP is the inability to support constraint based routing. Employing PCE for path computation also includes a delay for restoration path computation. PCE has the advantage of providing constrained-based path computation guaranteeing the QoS requirements. The E-NNI approach also needs protocol re-convergence at different levels of the applied hierarchy. It is obvious that the slower the path computation is, the longer it takes to restore a failed connection. A possible solution is relying on pre-computed link disjoint backup paths. Both the PCE approach and some BGP extensions [6] (see also chapters 4 and 5) offer solutions for obtaining such paths.

- Confidentiality preservation policies: This challenge is due to the confidentiality preservation policies regarding the network state. Since inter-domain link failures are considered it is still unclear how much information the domains are willing to share regarding the inter-domain links and how far the failure notification should be propagated. A typical approach is to confine the failure notification just within the domains adjacent to the failed link [79].

6.3 Restoration mechanisms for multi-domain networks

In this section, the two standard restoration methods E2E and L2E, are discussed in multi-domain scenario. Then, two enhanced restoration strategies based on the selective application of E2E and L2E recovery per failed link or per affected connection are proposed.

6.3.1 Standard restoration mechanisms applied in multi-domain networks

The following assumptions are considered as a basis for the investigated recovery framework. There is no requirement for parallel links between domains or between border nodes. Path diversity is assumed to be provided at an AS level via manual configuration, modified BGP operation or PCE. Each entity, responsible for the multi-domain routing is assumed to have two disjoint paths to each reachable destination. This is

referred to as LSP restoration with route pre-computation in the literature [84]. Topology preservation requirements between domains are considered, i.e. a node in one domain does not have visibility of the topology of another domain (refer to Section 6.2.3 for details). The nodes responsible for initiating restoration are either the border nodes of the domain adjacent to the failed link, or the head-end of the connections. Considering the outlined routing constraints and failure scenarios (inter-domain link failure), from the standard restoration techniques only the E2E and the L2E restorations are possible, since border routers only have knowledge of final destinations.

6.3.2 Enhanced restoration management

Here, two novel restoration techniques based on the standard L2E and E2E methods are proposed. Their basic operation is outlined and some challenges are discussed.

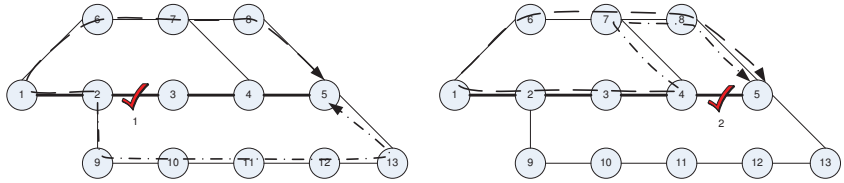
Simple location-based restoration (SLBR)

Previous research work on intra-domain recovery suggests that there is a close correlation between the location of the failed link and the availability of a connection, traversing that link [85]. In a multi-domain scenario, considering inter-domain link failure, the same relation can be expected. An intuitive suggestion could be that a failed link closer to the destination would be faster recovered with the L2E technique, whereas a failed link closer to the source - with the E2E technique. In well-connected networks the length of the backup paths are often close to the length of the working paths. This leads to the assumption that the L2E technique would recover a connection faster, if the failed link is closer to the destination (see Fig. 6.2 for example in a single-domain case). For a failure closer to the source, on the other hand, the E2E recovery could be more efficient with respect to the consumed resources for recovery. Thus, the first proposal in this chapter is a location-dependent restoration, based on the location of the failed link along the affected LSP path.

The actual decision process at the node detecting the failure is relatively simple:

- if the failed link is close to the destination, apply L2E restoration;

- else, notify the head-end of the connection to perform E2E restoration.



(a) Failure at the beginning of the path. (b) Failure at the end of the path.

Figure 6.2: Location-dependent failure handling. Lines indicate signaling options for L2E and for E2E recovery.

Shortest New restoration (SN)

A typical objective in recovery strategies is improved resource utilization. In order to minimize the additional resources consumed during restoration, compared to the resources consumed for the working path, the restoration decision can be based on the amount of potentially consumed resources. Considering a wavelength-division multiplexing (WDM) network, the resources are occupied wavelengths. The total resource consumption of a LSP would be the number of links it traverses. This can be expressed in terms of number of hops. The decision taken in the node detecting the failure is:

- if $D(s,i) + D_{res}(i,d) > D_{res}(s,d)$, apply E2E;
- else, apply L2E

where i indicates the node, detecting the failure, s indicates the source node, d indicates the destination node, $D(i,s)$ is the distance in hops from i to s, D_{res} is the length of the restoration path in hops.

Operation example

The operation of the proposed restoration schemes is divided into two main phases:

I Pre-failure actions: in this phase, all nodes along the path of an LSP need to obtain certain information, which can help them to take the most appropriate decision at the time of the failure. The information can vary and will be discussed further.

II Post-failure actions: in this phase, the node detecting the failed inter-domain link must make a decision - either to restore the connection directly, by applying L2E restoration, or to inform the head-end of the failed LSP, to apply E2E restoration. The decision is made based on the information disseminated in the previous phase.

Fig. 6.3 illustrates the operation of the mechanisms. In Phase I of the recovery mechanism the source node signals the working LSP and supplies the additional information needed for the enhanced mechanisms. Each border node on the way adds the link position and/or the length of the head-end restoration path as additional information in the LSP state database. In Phase II, depending on the applied restoration mechanism, the node detecting the failure takes a recovery decision. For the indicated failure on Fig. 6.3 if the location-dependent technique is used, L2E restoration will be performed. If the objective is to minimize resource consumption, then the E2E restoration should be applied.

6.3.3 Implementation aspects

The proposed restoration techniques rely on additional information in order to take the proper decision at the time of failure. This section is focused on the needed additional information, on the required protocol extensions, and on some practical issues involved in the process of implementing the schemes.

Both proposed schemes need to obtain the actual location of the failed link along the path of the affected connection. This can be done by examining the Recorded Route Object (RRO) of the RSVP-TE messages used for working LSP establishment. A possible challenge may occur if filtration of the RRO object is applied at domain borders. Depending on the level of trust between the domains the RRO is either filtered or encoded, which obstructs the process of obtaining the amount of hops to the destination and the source. A solution to this problem is to use an additional field in the RSVP-TE messages that keeps track of

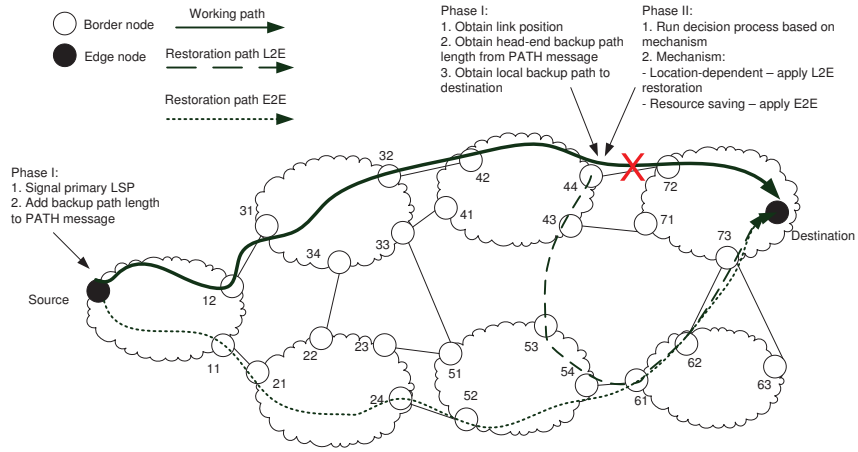


Figure 6.3: Enhanced restoration techniques in multi-domain scenario. For clarity of the figure the intra-domain topologies are not depicted.

the hop-count from the source and the destination respectively. Fig. 6.4 illustrates the options for obtaining the location of an inter-domain link. Both proposed options do not take into account the actual link propagation delays, which brings inaccuracy during the decision process at the time of failure. This can be avoided by using a different technique for obtaining the actual delays between nodes: time stamping. In this case all domains should use global time synchronization.

In the decision process of the resource preserving restoration scheme additional information is used: the length of the backup path which the head-end would use during restoration. According to the initial assumptions, irrespective of the applied multi-domain routing technique, all routing entities have two disjoint paths (if such exist) to each reachable destination. Both PCE and some modifications of BGP provide options for obtaining such paths (for example the proposals in Chapter 4 and Chapter 5). Thus, the head-end of a LSP can provide this information in an extended field in the RSVP-TE PATH message. The accuracy of this information can vary, depending on the used multi-domain routing protocol. It can indicate the number of Autonomous Systems (ASs) on the way (as in standard BGP), the number of border routers (as in Chapter 4 and Chapter 5), or a total hop count (as in PCE).

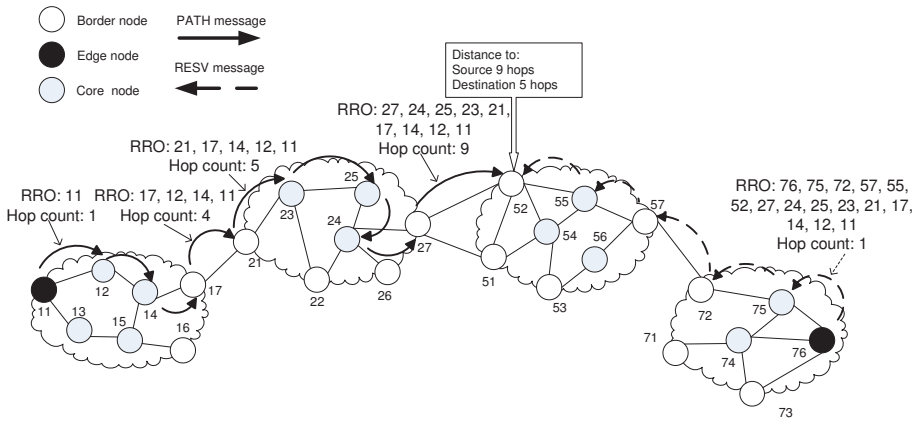


Figure 6.4: Obtaining the link position along the path of the working LSP: manipulation of RRO or using hop-count extensions.

As it can be seen, the needed extensions are not complex and the scalability of the involved protocols is not harmed regardless of the disseminated additional information.

6.4 Simulation study

The efficiency of the discussed restoration techniques was examined via simulations in an event-driven simulator for four different topologies: the 14-node NSFNET (Fig. B.6), the NOBEL Pan-European network (Fig. B.5), and a modified NOBEL topology (Fig. B.3) and the COST 266 Pan-European topology (Fig. B.4). The NSFNET is a network with low nodal degree and short average connection length (in terms od hops), where each domain has between one and two source/ destination nodes. The NOBEL and the COST 266 networks have high nodal degree and an effective core around domain DE (Germany). The domain borders follow the boundaries of the countries, with up to four source/destination nodes per domain. The modified NOBEL topology was created by adding links and deleting nodes and domains. The goal was to provide a more uniform topology with high nodal degree and no effective core.

RSVP-TE is used as a resource reservation protocol and the proposed modified BGP from Chapter 5 as multi-domain routing protocol. Using

the extensions from Chapter 5 means that the length of the protection paths (needed for the Shortest New (SN) mechanism) are in terms of border nodes on the way to the destination. The modeled networks are Wavelength Division Multiplexing (WDM) transport networks with 150 wavelengths per link. Traffic is uniformly distributed between all source/destination nodes and the LSP request inter-arrival time is exponentially distributed. All requests have average duration of 600 seconds. No wavelength conversion is applied during working path setup. Since head-ends can choose among all their unoccupied resources for recovery, in order to carry out a fair comparison limited wavelength conversion is introduced at the points of repair when local restoration is performed. Wavelength assignment is random and is carried out at the destination node based on the RSVP-TE label set. Each signaling packet is a subject to 1 ms processing delay per node.

Two performance metrics are investigated:

- Resource Overbuild (RO) - the difference between the occupied resources by the affected LSP before the failure and after the recovery.
- Recovery success ratio (RSR) - the ratio between the number of affected LSPs and the number of successfully restored LSPs.

The restoration delay was not considered as a performance metric since in well-connected networks with relatively equal-length primary and backup paths (as the Pan-European ones), where the hop-count in terms of ASes is considered as a TE metric, the L2E recovery will almost always be the fastest one. Exceptions are possible depending on the actual link lengths in the network and depending which links are failed during simulations.

Table 6.2 gives the general LSP blocking ratio and the average link utilization for all topologies at high and medium load ranges. The exact input load per node is also indicated. This table is only informative and gives the reader the notion of how loaded the networks are at the different tested load ranges.

Table 6.4 presents the results for the Resource Overbuild and Table 6.4 presents the results for the Recovery Success Ratio for the different restoration techniques at medium and high network loads for all tested topologies. Since the goal of this work is the general exploration of the

Network	Input load per node [Erlang]	Average link utilization	Average LSP blocking ratio
NSFNET	1	54%	0.17
	0.6	40%	0.014
NOBEL	0.7	39%	0.26
	0.4	32%	0.148
Modified NOBEL	0.55	40%	0.09
	0.3	21.5%	0
COST 266	0.7	42%	0.27
	0.4	34%	0.155

Table 6.2: Operation conditions per network per load range.

efficiency of different restoration techniques, first the performance of the standard restoration techniques is analyzed. Then, the performance of the novel techniques, proposed in this Chapter, is discussed.

Failed link (affected connections*)	NSFNET							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Node_10 – Node_3 (77-122)	2.64	0.81	2.17	0.81	2.36	0.76	2.03	0.76
Node_8 – Node_9 (31-42)	3	0.93	2.48	0.93	3	1	2,6	1
Average over different links	2.92	0.74	2.37	0.74	2.68	0.67	1.97	0.67
Failed Link (affected connections*)	NOBEL							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Strasbourg – Frankfurt (144-146)	0.16	-0.39	-0.08	-0.52	0.22	-0.45	0.04	-0.34
Amsterdam – Hamburg (130-155)	2.08	1.36	1.50	1.62	1.56	1.54	1.60	1.43
Berlin - Prague (85-91)	0.93	0.83	0.69	0.33	0.60	0.50	0.32	0.23
Berlin – Warsaw (55-64)	2	1.3	1.61	0.99	2	0. 8	1.38	0.59
Failed link (affected connections*)	Modified NOBEL							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Barcelona – Rome (91-130)	2.85	1.59	1.87	1.1	2.15	1.39	1.98	1.20
Berlin – Warsaw (42-72)	1.4	1.33	1.56	1.28	1.30	0.79	1.13	1.18
Failed link (affected connections*)	COST 266							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Berlin – Warsaw (131-144)	2.2	0.31	1.87	0.44	2.45	0.29	2.05	0.48
Vienna - Prague (9-20)	2	1	1.33	1	2	1	1.2	1
Rome – Zagreb (48-82)	2	-0.5	1.86	-0.22	1.86	0.33	1.88	0.68
* The numbers for the affected connections are in the format (Medium load – High load)								

Table 6.3: Resource Overbuild for all tested topologies and restoration mechanisms.

Failed link (affected connections*)	NSFNET							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Node_10 – Node_3 (77-122)	0.54	0.69	0.61	0.69	0.23	0.3	0.26	0.3
Node_8 – Node_9 (31-42)	0.77	0.93	1	0.93	0.26	0.57	0.36	0.57
Average over different links	0.73	0.89	0.83	0.89	0.25	0.49	0.36	0.49
Failed Link (affected connections*)	NOBEL							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Strasbourg – Frankfurt (144-146)	0.68	0.54	0.74	0.65	0.63	0.49	0.71	0.58
Amsterdam – Hamburg (130-155)	0.44	0.23	0.41	0.38	0.30	0.13	0.27	0.23
Berlin - Prague (85-91)	0.63	0.78	0.70	0.80	0.64	0.57	0.64	0.69
Berlin – Warsaw (55-64)	0.45	0.60	0.51	0.72	0.27	0.42	0.36	0.56
Failed link (affected connections*)	Modified NOBEL							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Barcelona – Rome (91-130)	0.82	0.87	0.83	0.90	0.41	0.48	0.48	0.52
Berlin – Warsaw (42-72)	0.86	0.56	0.83	0.80	0.72	0.30	0.64	0.58
Failed link (affected connections*)	COST 266							
	Medium Load				High Load			
	L2E	E2E	SLBR	SN	L2E	E2E	SLBR	SN
Berlin – Warsaw (131-144)	0.48	0.32	0.51	0.34	0.31	0.19	0.26	0.2
Vienna - Prague (9-20)	0.38	0.5	0.38	0.5	0.27	0.54	0.45	0.54
Rome – Zagreb (48-82)	0.33	0.16	0.29	0.18	0.35	0.15	0.3	0.19
* The numbers for the affected connections are in the format (Medium load – High load)								

Table 6.4: Recovery Success Ratio for all tested topologies and restoration mechanisms.

6.4.1 L2E vs. E2E restoration

The efficiency of the restoration techniques is different for the different topologies. Thus, they are discussed separately.

NSFNET For the NSFNET the E2E mechanism provides on the average the highest Recovery Success Ratio and the lowest Resource Overbuild. This is due to the following two facts. First, the network is with low nodal degree, which makes local backup paths longer especially for failures in the middle of the network. Second, the design of the NSFNET results in very well-balanced traffic distribution. The link occupation at the moment of failure can be seen on Fig. 6.5 which illustrates the percentage of links which are loaded up to 30%, between 30% and 60% and above 60%. The X-axis represents input load per node. As it can be seen, the load in the network is evenly distributed among most of the links, 50% of the links are loaded between 30% and 60% at high loads. Moreover, there are not many links which are loaded up to their limit (above 90% utilized) and non of the links is not used at all. Thus, if the L2E technique is used to restore failed connections, all of them need to be accommodated on the local backup path, which is already in use by normal traffic and this saturates the link. On the other hand, if the E2E technique is used, the load is distributed more evenly between the backup paths of the affected source/destination pairs, i.e. E2E provides better load balancing in the network. With respect to the RO, considering the low nodal degree and the relatively long backup paths, it is expected that the E2E mechanism will result in the lowest RO. The results obtained for the NSFNET are consistent across different load ranges and failed links and thus, an averaged result is also presented.

Pan-European networks From a single-domain perspective, if a network is with high nodal degree and the backup paths are comparable in length to the working paths, then it can be expected that the L2E scheme would perform worse than the E2E when the load of the affected link is high. When many connections need to be reestablished locally, the local backup path, which also carries normal working traffic, would congest. The observed behavior in both the NOBEL and the COST 266 networks though is exactly the opposite - the higher the load on the failed link,

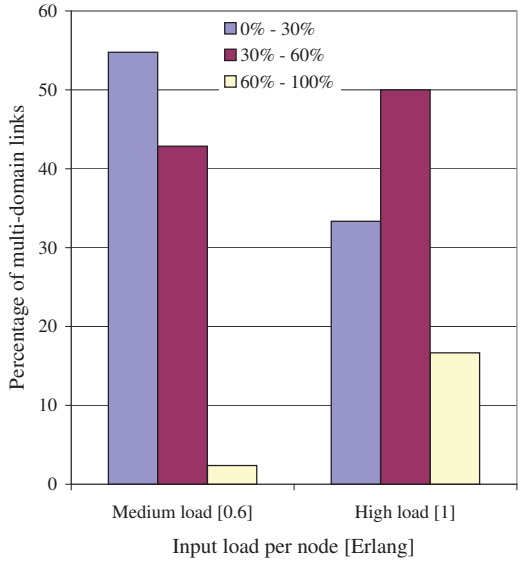


Figure 6.5: Link load distribution at medium and high loads before failure for NSFNET.

the better the performance of the L2E scheme is. This is due to the following reasons. First, most of the working paths which pass through domain DE (the effective core) use the same inter-domain links (due to the path-dependency of BGP illustrated in Chapter 3). Second, the network is with high connectivity especially in the core area. As a result, if an inter-domain link close to the core is heavily loaded, the local backup path would be lightly loaded; and vice versa, if a link is lightly loaded, the backup path would carry a lot of working traffic.

For the NOBEL topology, since all tested failures are for links connected to domain DE, then if the link is heavily loaded, the L2E restoration is more efficient than the E2E and vice versa. For example, for links Amsterdam-Hamburg and Strasbourg-Frankfurt in the NOBEL topology the L2E RSR is higher than the E2E RSR, whereas for the relatively lightly-loaded Berlin-Prague and Berlin-Warsaw the E2E RSR is higher than the L2E one for the medium load range.

For the COST 266 topology the same results are observed: for the heavy loaded link Warsaw-Berlin the L2E RSR is higher than the E2E RSR, whereas for the low-loaded link Vienna-Prague, the result is oppo-

site. An important aspect to be pointed out is the fact that in the Pan-European topologies many multi-domain links remain unutilized. For the COST 266 network 10.29% of the links are not used at all, whereas almost 15% are heavily loaded. The COST 266 link occupation can be seen on Fig. 6.6.

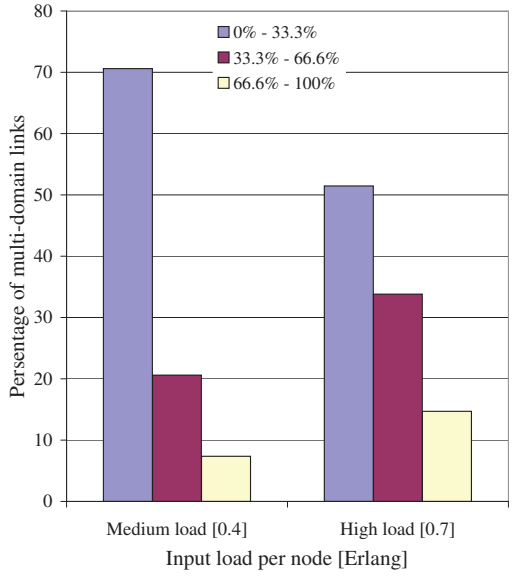


Figure 6.6: Link load distribution at medium and high loads before failure for COST 266.

For the Modified NOBEL topology, the topology design has eliminated the effective core area around domain DE and the inter-domain traffic is more evenly distributed in the network. The behavior of the L2E and E2E can be expected to be close to a single-domain scenario, since the negative effects of the BGP path-dependency are decreased, i.e. the L2E technique would be more effective if the amount of affected connections is not high. As seen from the results, when a heavyly loaded link fails, the L2E restoration performs worse than the E2E and vice versa.

With respect to the RO, the obtained results follow the expectation of lower RO for the E2E scheme due to the good connectivity of the networks and the relatively equal length of working and backup paths.

Interesting result is the negative RO in some cases. This is due to the fact that BGP chooses paths based on the shortest AS-path, regardless of how many actual hops the path consists of. A path with 3 domains, each domain having one border node is considered "longer" than a path with 2 domains, each having 2 border nodes. For example, the path from Austria to Netherlands is 2 AS-hops long (Austria-Germany and Germany-Netherlands), but in fact there are 4 hops: Vienna-Munich, Munich-Frankfurt, Frankfurt-Hamburg and Hamburg-Amsterdam). Thus, a longer AS-path may result in a shorter hop-count. As a result, a backup path maybe shorter hop-wise than a working path which explains the negative RO values.

6.4.2 Shortest New(SN) and Simple Location-Based Restoration(SLBR) mechanisms

The design goal behind the SN mechanism is minimizing recovery resources, which is exactly what the scheme does. The SN mechanism does not provide improvement in terms of RSR but it indeed provides a lower RO in almost all cases, especially at high network loads and link failures with heavy traffic. Inaccuracies in the estimation of the potential consumed resources may occur due to the mechanism used for obtaining the additional information needed for restoration (see Section 6.3).

Regarding SLBR, the scheme clearly provides a good balance between RO and RSR, especially for the cases where the L2E provides the highest recovery success and the E2E provides the best RO. In most of the tested cases for all topologies the RSR is between the RSR for the L2E and the E2E schemes. The same is valid for the RO. In some cases (link Strasbourg-Frankfurt) the SLBR even outperforms both the L2E and the E2E with respect to RSR for all load ranges.

The improvement in the network performance under the SLBR and the SN mechanisms is based on the differentiated failure handling. Under the SLBR and SN schemes, local and head-end restoration are applied selectively. This contributes to a better load balancing in the multi-domain network which increases the probability for successful restoration. In fact, under the SN mechanism head-end recovery is applied 50% of the cases in the NOBEL network and 90% of the cases in the COST 266 network. This is due to the fact that the working and the backup paths

in these two networks are of comparable lengths, which make head-end restoration the most effective when resource consumption is considered. For the SLBR mechanism the head-end restoration is applied in close to 50% of the cases for both topologies which explains the balance between the obtained performance metrics.

6.5 Conclusion

This chapter explores the area of survivability management in multi-domain networks. After presenting the state of the art in survivability in multi-domain networks, the main challenges in applying restoration mechanisms are identified.

The main goal of the work presented here is the evaluation of the efficiency of different restoration techniques. The performance of two standard restoration techniques, namely local-to-end and end-to-end, is presented. Furthermore, two novel techniques are presented and evaluated. The first one is based on the location of the failed link with respect to the head-end of the connection and the second one is based on minimizing the occupied resources for recovery. Simulation results show that the applied routing protocol (BGP in this case) highly influences the efficiency of all restoration techniques. Furthermore the parameters of the network topology as well as the load of the failed link also affect the performance.

In networks with low nodal degree, such as NSFNET, end-to-end restoration performs the best due to the short end-to-end backup paths and the relatively even load distribution in the network. For networks with high nodal degree, such as NOBEL and COST 266, the performance of the mechanisms varies with the load of the network and the location of the failed links. In particular, heavy-loaded links in the core of the network can be successfully recovered locally, without distributing failure notifications through the whole network. The path-dependency inherent in the BGP operation leaves some core links unutilized, which improves the performance of the local-to-egress restoration method.

The presented results indicate that with few simple protocol extensions it is possible to create differentiated failure recovery strategies. The proposed Shortest New mechanism focuses on minimizing resource consumption and the Simple Location-based Restoration balances the

performance of the recovery success ratio and the resource overbuild metrics, which are typically in a trade-off relationship.

The obtained results indicate a need for the development of a differentiated failure management framework, which supports recovery based not only on traffic priority, but also on the abilities of the intermediate domains to support recovery, the multi-domain connectivity and the load distribution in the network. Such a framework would facilitate the interconnection of heterogeneous domains with various survivability support capabilities and would increase the potential service portfolio of cooperating network operators.

Chapter 7

GMPLS Control Plane for OBS Networks

7.1 Introduction

The Generalized Multi-Protocol Label Switching (GMPLS) framework [17] was designed to support all existing switching technologies at the time of its creation. Almost in parallel though a new switching paradigm was developed - the Optical Burst Switching (OBS). Thus, the specifics of its operation have never been included in the GMPLS specification. As the OBS technology developed, it became clear that it is indeed a viable solution for certain applications, e.g. various applications in grid computing networks and for metro networks [15,86]. This development brought with it the need for OBS control plane specification and for integrating OBS and legacy networks under a common control framework.

This chapter explores the work in the area of GMPLS/OBS integration. Most of the existing literature dates from before the original research work, which was carried out during this Ph.D. project and was published in [12]. Nevertheless, some very good and valuable work has been published afterwards. This chapter is based on the work presented in [12] and is further extended to cover the new work in the area.

The chapter is organized as follows. Section 7.2 gives a short background description of the OBS technology. Section 7.3 introduces the motivation for the GMPLS/OBS integration. Section 7.4 presents a

general GMPLS-based OBS architecture and the relationship between the corresponding functional blocks in both technologies. In Section 7.6 the existing GMPLS/OBS integration approaches are presented and analyzed. Section 7.5 identifies and elaborates on the different advantages of the GMPLS/OBS integration and the challenges, which require due consideration. The chapter is concluded in Section 7.7.

7.2 Optical Burst Switching

OBS is an intermediate switching paradigm between Optical Channel Switching (OCS) and Optical Packet Switching (OPS) [87]. In an OBS network, the Ingress nodes assemble several client packets[1] into Data Bursts (DBs), which they send in the network preceded by a special Burst Header Packet (BHP) without waiting for an acknowledgment from the Egress node. This leads to connectionless, one-way resource reservation paradigm. The BHP contains all needed resource reservation information (burst length, priority, wavelength, etc.), it uses separate dedicated wavelength channel and is electronically processed in each intermediate node. By the time the DB reaches a core node, its switching matrix is configured and the burst cuts-through the node without Optical-Electro-Optical conversion (see Fig. 7.3). This specific way of operation results in bypassing the existing bottleneck in the electronic routers and relaxes the requirements for optical buffering. OBS is highly dynamic and achieves a high degree of statistical multiplexing. Additionally, the technology provides complete transparency of the switching node regarding the code, the format and the speed of the client traffic, and the used upper-layer protocols by providing all-optical transport of the client data. A thorough OBS technology description and discussion are available in [88].

Being a young technology, OBS lacks a well-defined control plane. The general trend in this area of research is towards the design of a flexible and efficient IP-centric control plane for OBS networks [89, 90]. The rationale behind this initiative is the current trend of intelligence migration towards the IP layer [91] justified by the achieved progress in

[1] Assuming IP over OBS, but not excluding other packet or frame oriented technologies.

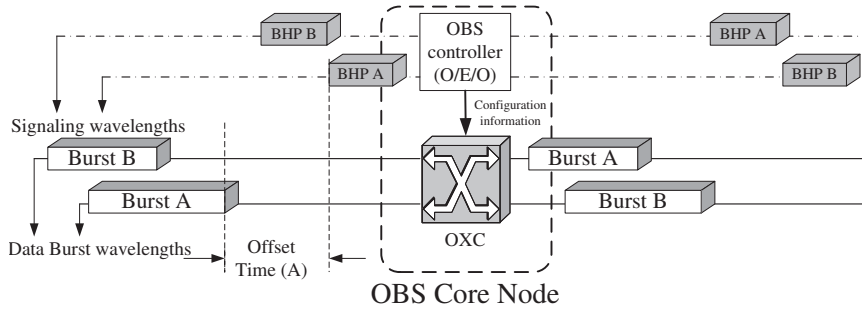


Figure 7.1: Basic OBS paradigm.

standardizing the IP-based GMPLS control plane. GMPLS, being technology independent and suitable for control of both packet and circuit switched networks, seems to be an eligible candidate for the control of OBS networks. The main challenge in applying GMPLS in the OBS control plane is the fact that GMPLS was never designed to support burst switching. This brings along some integration impediments which need thorough discussion and consideration.

Up until now, there are only few research publications, focusing on the issue of integrating GMPLS and OBS control planes. The objective of this chapter is twofold. First, to clarify the main reasons behind the GMPLS integration into the OBS control plane, to identify some of the most important challenges of the GMPLS/OBS integration process, and to discuss different approaches for coping with them. Second, to provide a comprehensive description and classification of the existing solutions for GMPLS/OBS integration, to analyze their advantages and drawbacks, and to elaborate on some required protocol extensions.

7.3 Motivation for GMPLS/OBS integration

There are two main motives for integrating the OBS CP within the GMPLS framework. The first one is based on the fact that OBS networks still do not have a well-defined control plane and the second one is based on the similarities between the technologies.

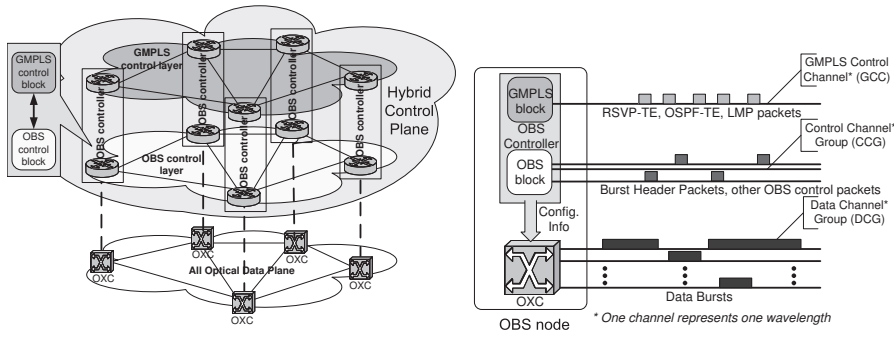
Currently, there is a trend towards the development of IP-centric

control plane for the optical transport infrastructure. In fact, there is "*a consensus (among the researchers) that the IP routing and signaling protocols can be adapted for the optical network control*" [92]. This is as a result of the advanced deployment and standardization of the GMPLS control framework [17, 18]. Since GMPLS is technology independent framework, it is the preferred control technology for Next Generation Transport networks, including for OBS networks. The authors of [92] argue that: "*incremental extensions of existing protocols for reuse may not be the best choice in terms of software complexity and overhead*". Nevertheless, the level of standardization of the GMPLS protocols is fairly advanced. The migration of intelligence towards the IP layer and the prediction for IP dominance in the near future justify the efforts for adapting the existing protocols to the needs of the optical transport networks. The main objectives of this process must be reduced complexity and increased efficiency, as well as realistic and future-proof perspectives. Adapting the GMPLS protocol suite to the OBS technology will alleviate the process of standardizing the OBS technology by providing it with some missing functionalities in the control plane (e.g. addressing, routing, link management). It will also speed up the deployment of OBS networks due to the higher level of compatibility with existing networks, controlled by the GMPLS paradigm.

Apart from eliminating the need for defining a completely new control plane for the OBS network, the GMPLS/OBS integration is also somehow natural due to the similarities between the technologies. The explicit separation of the data and the control planes and the use of source-routing as essential routing method facilitates the integration of both technologies. Furthermore, GMPLS is the predominant solution for the integration of both packet and circuit switched technologies, whereas OBS is de-facto an integrated packet/circuit switched technology.

7.4 Integrated GMPLS/OBS control plane architecture

The specifics of the OBS technology imply that the OBS network architecture can be presented as comprising two separate overlaying networks: a transparent all-optical network for DB transportation and a hybrid control network for control information exchange [89]. Integrating the



(a) GMPLS/OBS network layering. (b) GMPLS/OBS channel separation.

Figure 7.2: Integrated GMPLS-based OBS architecture.

GMPLS framework into the OBS control plane will further divide the control network in two parts: GMPLS control part and OBS control part (Fig. 7.2 (a)). This architecture suggests the existence of three groups of channels [93] (Fig. 7.2 (b)): Data Channel Group (DCG) for DBs, Control Channel Group (CCG) for BHPs and other OBS-related packets, and GMPLS Control Channel (GCC) for GMPLS-related packets.

There are two main functional blocks which need to exist in an Optical Control Plane (CP) for basic network operation - routing (for resource dissemination and path computation) and signaling (for resource reservation). Fig. 7.3 presents the relationship between the OBS-CP functional blocks and the GMPLS functional blocks. It can be seen that some of the OBS-specific functions are not provided by the GMPLS CP, such as contention resolution or offset-time management, and that other functions overlap - the routing and the signaling for resource reservation. Depending on the adopted integration scenario (see the discussions in Section 7.5), the level of overlap between the functional blocks will be different.

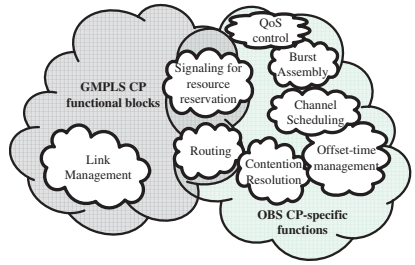


Figure 7.3: Integrated GMPLS/OBS control plane functions.

7.5 GMPLS/OBS integration scenarios

In this section, classification of the existing solutions for GMPLS/OBS integration is presented. The solutions are grouped in two categories. In the first category, the GMPLS control plane is decoupled from the OBS control plane functions and both work at separate levels independently. This model is referred to as the *overlay model*. The second category is an integrated model type where the signalling for resource reservation is only performed by the GMPLS CP. Detailed description and analysis of the solutions is presented further.

7.5.1 Overlay model

Under this model the GMPLS and the OBS CP are in a client-server relationship. In this case either the GMPLS CP defines the topology over which the OBS network operates, or the OBS network is the underlying transport network and it provides services to the client GMPLS networks.

OBS as a client

In this group of solutions [94–96] the main approach is to use GMPLS for defining the OBS network topology, i.e. the OBS network appears to be an overlay network over a GMPLS-controlled WDM network. Under this scenario, the signaling functional block on Fig. 7.3 is separated in two and each CP performs signaling for resource reservation, but at different levels.

The first work on this integration approach is from [94]. The authors suggest a general architecture for GMPLS control of OBS networks and outline needed modifications in the GMPLS protocols. Under their suggestion, the labeled unit is the BHP and the BHPs and their corresponding DBs are treated as client-traffic, i.e. they both reside in the data plane of the GMPLS controlled WDM network. According to the model, the BHPs are normal label-switched packets, i.e. it is assumed that a two-way signaling for end-to-end label distribution (i.e. Label Switched Path (LSP) setup) has been performed in advance. Additional aspects of link bundling, routing and link management are also considered.

The authors of [95] acknowledge that the strongest feature of the GMPLS framework is the flexibility and manageability of LSPs. Following this and similar to [94], they propose to use the GMPLS CP for defining the OBS network topology by setting up explicit light-paths, but without explicit resource allocation per client flow. From a technical point of view, this scheme proposes that the GMPLS CP defines the Virtual Topology which an overlay OBS network uses to transport bursts (see Fig. 7.4). Under this scheme, if two OBS nodes need to communicate they can do that only over predefined paths which are set-up via the normal GMPLS procedures. During the signaling procedure the involved nodes exchange label identifiers which are to be used as labels for the BHPs. When there are more LSPs established between two OBS edges, they can choose to use any available pre-established LSP, optimally the one that in the moment provides lowest blocking probability. The proposal also suggests that there is no explicit resource reservation per client traffic flow during the GMPLS signaling (as outlined in [94]) in order not to contradict the statistical multiplexing feature that enables OBS to achieve good resource utilization.

Another proposal within this category comes from Guo et al. [96]. The authors propose a multi-layered architecture, where OBS networks are clients of a GMPLS controlled OCS network. The OBS networks use optical channels provided by a circuit-switched GMPLS network to interconnect the OBS border nodes. Unlike the proposal in [94], the authors use the GMPLS network to create only the inter-domain LSPs over which the OBS client domains communicate. The authors elaborate on the definition of different extension to the RSVP-TE protocol as well as definition of new fields in the OSPF-TE protocol messages.

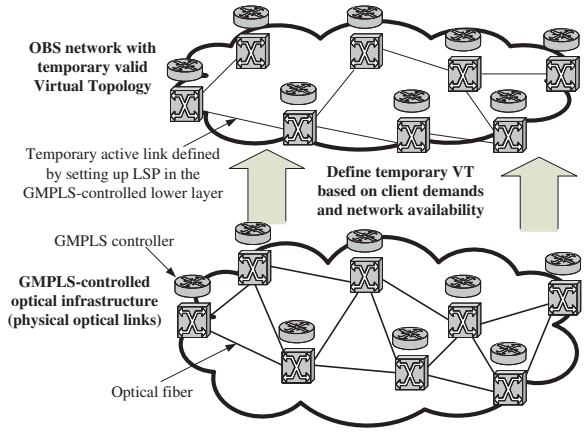


Figure 7.4: GMPLS controlled optical infrastructure providing and managing the virtual topology to be used by OBS flows.

All of the proposed integration schemes [94–96] are designed for IP-over-WDM integration and suggest the same application of GMPLS signaling: to define (and respectively manage) the topology of the OBS network. This can be seen as a layered approach: the lower layer is a circuit-switched optical network, which resources are controlled by the GMPLS CP. On top of that optical infrastructure is the OBS network, which is allowed to use specific paths and wavelengths for transmission only when there has been established at least one explicit LSP. Establishing the LSPs to connect the OBS nodes serves as means for exchanging label identifiers which are used for labeling the BHPs traveling between these OBS nodes (similar to Multi-Protocol Label Switching (MPLS) networks where label exchange is not equivalent to strict physical resource reservation). This approach is valid as it fits the GMPLS hierarchical model (see Chapter 2). However, some inherent OBS drawbacks, such as contention due to lack of strict resource reservations, cannot be avoided. Such integration cannot guarantee strict QoS provisioning per client flow (see the discussions in Section 7.6.1). Nevertheless, the approach supports flexible traffic handling and the applicability of well-defined survivability techniques within the GMPLS framework.

GMPLS as a client

Under this option for GMPLS/OBS integration the OBS network is used as a generic transport solution for support of various switching technologies. This approach is similar to the previous one, with the only difference being that the OBS network is used as the underlying transport network, connecting domains of different switching capabilities. This approach is described in [97] and comes closest to the idea of integrating all known services under the OBS framework. The proposals in [94–96] are designed for IP-over-WDM integration whereas this approach assumes any type of client network, including wireless access networks. For regular bursts the authors propose the same operation as in [94–96]. For periodic bursts, i.e. constant bit-rate connectivity, and wavelength channels the traditional GMPLS signaling for resource reservation could be used. However, no specific details possible protocol extensions are given. It is only suggested that GMPLS should be modified in order for this model to be implemented. This solution is depicted on Fig. 7.5, where the OBS network is the underlying transport network and the OBS CP is kept separate and independent of the CPs of the overlay legacy networks. Under this proposal, the OBS signaling adapts according to the signals' requirements, as suggested in [97], and there are LSPs established prior to client data transfer.

7.5.2 Integrated model

Under this integration scheme the GMPLS CP seeks to mimic the functionalities needed for OBS operation. This integration option is only briefly discussed in [12]. It is depicted on Fig. 7.6, where a modified GMPLS signaling is used to perform the requirements for DB signaling. Under this approach, no LSP establishment is performed prior to client data transfer and the BHPs are not labeled. In fact, under this proposal, the GMPLS signaling is responsible for the actual resource reservation, i.e. the GMPLS framework is extended for support of one-way resource reservation. Such an approach facilitates the horizontal integration between GMPLS-controlled legacy networks and the OBS transport network. Besides the required changes in the GMPLS resource reservation procedure, a specific adaptation function at the edge of the OBS network for translation of the standard GMPLS information into OBS-compliant

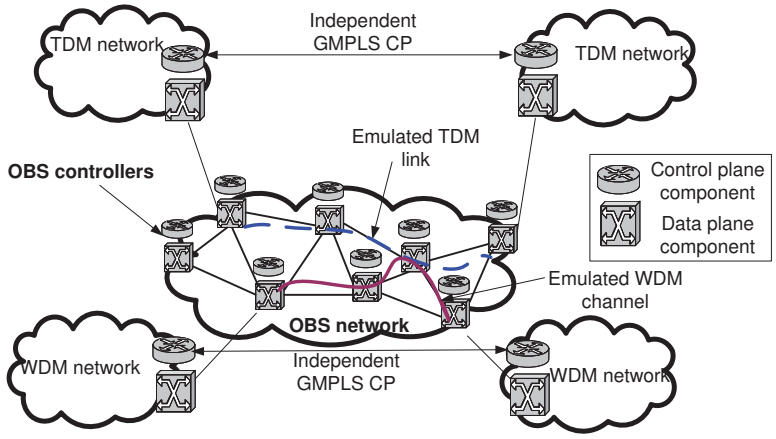


Figure 7.5: OBS controlled optical core infrastructure providing different services in order to connect different types of GMPLS controlled clients.

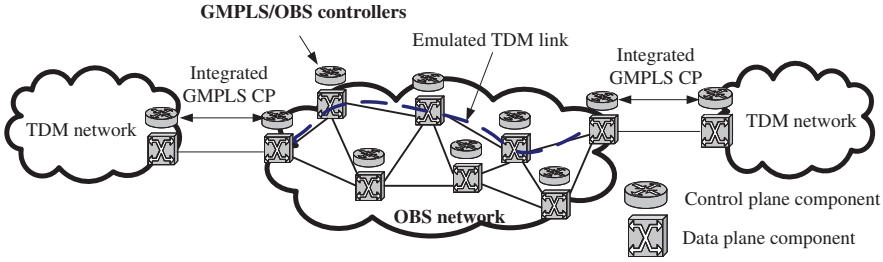


Figure 7.6: Integrated GMPLS/OBS model employing modified GMPLS for OBS signaling.

information is needed as well. The implications of this approach are analyzed in Section 7.6.1, where possible GMPLS protocol extensions and modifications, which can facilitate DB signaling using the standard GMPLS protocols, are suggested.

7.6 GMPLS/OBS integration considerations

This section focuses on different integration issues with respect to the two overlapping functional blocks from Fig. 7.3. The main integration issue

arises in the area of resource reservation, because both the GMPLS-CP and the OBS-CP have such functional block. The routing function, on the other hand, poses less issues because OBS does not have a technology-specific method or protocol for routing. The only routing requirement is the application of source-routing [88]. Thus, any link-state routing protocol that supports constrained-based path computation can be employed. The generality of the GMPLS framework does not limit the possible integration solutions, but creates a range of opportunities for solving the potential integration issues. The decision, which exact solution to be applied, depends on the application of the OBS network and the targeted network performance characteristics.

7.6.1 Signaling issues

Control channels and control packets

In the basis of GMPLS lies the idea of complete decoupling of the control and the data planes. This decoupling is not only logical, but also physical. GMPLS control packets do not need to use the same infrastructure as the data packets, they can use a completely separate network. In OBS networks, on the other hand, the BHPs and their corresponding DBs use the same optical fiber, i.e. the DBs using certain optical fiber have their BHPs on the same fiber, just on a separate wavelength channel. This difference brings up the question: will there be a separate control network for the GMPLS protocols, or will this control information use the same dedicated control channels (see Fig. 7.2 (b)). Two options are possible: to have two separate networks (GMPLS- CP and OBS-CP) or to share the wavelength channels for the GMPLS and the OBS control messages. Between two OBS nodes there can be as many OBS control channels as optical fibers, but only one GMPLS control channel is needed. The decision of where and how to configure the GCC becomes an issue not only from an implementation point of view but also, from a management and control perspective. Separating the GCC completely necessitates the maintenance of two control networks. Furthermore, assigning a whole wavelength for the GMPLS control information might waste capacity if the information cannot use the entire bandwidth. The implementation of the GCC and the CCG is an important aspect which depends on the applied integration model (see Section 7.5). If the overlay model is ap-

plied, then the GCC and the CCG should be implemented separately. If the integrated model is applied, a common data communication network can be used for all CP-related messages, both OBS and GMPLS.

Integrating GMPLS into the OBS control plane also brings up the question about the format of the control messages. Since there is no standard for the OBS-related control messages, it is possible to use the format of the GMPLS-related control packets. In [98] and [99] it is proposed to extend and piggy-back the GMPLS keep-alive/hello messages, or to modify the existing packet formats so that they can carry the needed reservation parameters. However, no details on possible extensions or analysis of possible inter-operability issues are presented. Here, three options for transporting OBS-related information within the RSVP-TE PATH message under the integrated model presented in Section 7.5.2 are evaluated. The there options are[2]:

1. To modify the original objects in the *Path* message (Fig. 7.7 a));
2. To piggy-back the *Path* message, i.e. to define new optional objects without modification of the original objects (Fig. 7.7 b));
3. To define a new Resource ReserVation Protocol (RSVP)-TE message type (Fig. 7.7 c)).

There are several reasons that advocate for the definition of a new Resource ReserVation Protocol with TE extensions (RSVP-TE) message type. First, modification of the original *Path* message (Fig. 7.7 a)) leads to compatibility issues during the message processing procedures, especially if the processing node does not support OBS. The second option, for piggy-backing the *Path* message (Fig. 7.7 b)), avoids this problem. However, there are several objects that are not needed for OBS support but are mandatory *Path* message objects. This implies increased signaling overhead. Instead, a new RSVP-TE message type can be designed (Fig. 7.7 c)), which explicitly identifies OBS resource reservation. This makes it easier to recognize if a message is OBS-related or is a *Path* message, needed for actual LSP establishment, and facilitates the simultaneous existence of connection-less and connection-oriented [88] OBS

[2]Note that the modified RSVP-TE messages do not serve as means for *label* distribution in this case.

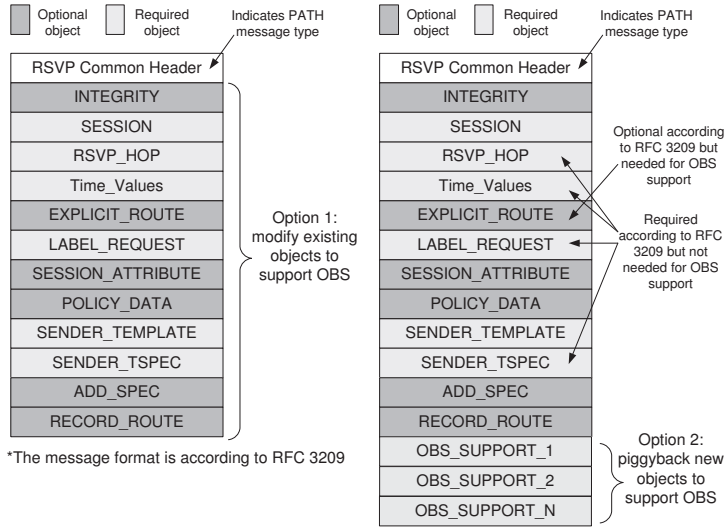
networks on the same infrastructure. Furthermore, QoS provisioning and multi-vendor inter-operability are also alleviated.

GMPLS signaling in OBS networks

An open question is whether to include the GMPLS signaling for LSP establishment prior to data transfer. The origin of this problem is the fundamental difference between the resource reservation paradigms in both technologies. GMPLS relies on end-to-end connection setup prior to data transfer in order to guarantee delivery and different levels of QoS. On the other hand, the main advantage of the OBS technology is its one-way reservation procedure, which minimizes the end-to-end delays and improves the bandwidth efficiency by introducing statistical multiplexing of the wavelength channels.

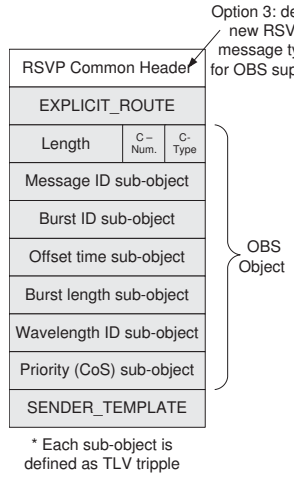
All proposals within the overlay model category (see Section 7.5) assume the existence of preestablished LSPs for the flow of BHPs per source-destination pair. However, the specifics of the OBS technology does not allow for the most powerful TE tool of the GMPLS framework, which is QoS guarantees via LSP establishment, to be fully utilized. The main problem is that setting up LSPs in OBS networks under the overlay model does not guarantee resource reservation for the client traffic. In fact, the resource reservation per client traffic flow is still carried out via the OBS signaling on a hop-by-hop principle. Another issue is that the fast packet forwarding, which is one of the biggest advantages of the label switching technology, cannot be utilized. This is because even though the BHPs are labeled and the output port can easily be determined by a simple look-up in the Label Information Base (LIB), the BHPs still needs to wait until the scheduler of the node performs the resource reservation. Specific information related to the resource reservation process needs to be read from the BHP and eventually modified (offset-time, wavelength, etc.). There is also a third feature offered by the GMPLS control framework which cannot be utilized. From survivability point of view, if there is a failure in the control plane of a GMPLS controlled network, the data plane should be unaffected[3], i.e. all existing connections continue to be operational. In OBS networks this is not the case since the CCG is also a part of the control plane. If a failure happens there and the BHPs

[3] This comes from the decoupling of the data and the control planes.



(a) Modifying the *Path* message for OBS support.

(b) Piggybacking the *Path* message for OBS support.



(c) Defining new OBS message type for RSVP-TE.

Figure 7.7: Possible extensions of the RSVP-TE protocol for OBS support.

cannot be delivered, the corresponding data bursts will be lost as well.

Despite the missed advantages, it is clear that setting up LSPs prior to data transfer has one very strong side: simplified traffic handling. Characterizing flows in a simple way and handling them in fast and efficient manner is a basic tool of traffic engineering. identical requirements. Apart from being technologically simpler, using the label-based forwarding brings the advances achieved in the area of fast protection and restoration to the OBS domain. Another strong point is the compatibility with other MPLS/GMPLS controlled networks which undoubtedly facilitates multi-vendor and inter-domain inter-operability.

Generic labels in OBS networks

An important question in the GMPLS/OBS integration is what will the *label* signify? Labels are used in each label switching capable node to identify the input-output port relationship via LIB look-up, i.e. the label shows the binding between the port at which the data enters and the port at which the data leaves the node. Most of the proposed integration scenarios (see Section 7.5) suggest the BHPs to be the labeled entity, i.e. the label is not associated with the actual client traffic. This implies that the OBS CP is actually an MPLS network. The only difference is that the packets (BHPs) are processed additionally instead of simply label-forwarded. The label can signify the <input port, output port> binding, the priority level of the client flow, the range of wavelengths which can be used, or other resource reservation information. In fact, almost all needed resource reservation information can be encoded in the label. This significantly reduces the BHP processing time and the signaling overhead, and increases the throughput of the network. A possible drawback is decreased flexibility in the OBS network due to the needed fixed parameters such as burst length or used wavelength.

7.6.2 Routing issues

Maintaining up-to-date information for resource availability in the network is essential for traffic engineering. The responsibility for resource availability dissemination and topology state maintenance falls on the Routing functional block (see Fig. 7.3). Under the GMPLS framework, OSPF-TE is responsible for this function. Disseminating Link State

Advertisements (LSAs) in OBS network poses several challenges. One of the characteristics of the OBS technology is that it offers statistical multiplexing of the network's resources. At the same time, optical buffering via Fiber Delay Lines (FDLs) is very limited and in most cases even excluded as a possibility for contention resolution. This brings up the question what exactly will be advertised in the LSAs. Since the resource reservation happens at the granularity of a whole wavelength and resources are reserved for a limited duration, it will not be appropriate to advertise number of available wavelengths. A possible solution is to advertise the percentage of occupied resources for a given time period. In this way, an Edge node will know how much a given link and node are loaded and will be able to calculate a less congested path for the data bursts. Additionally, the availability of FDLs and wavelength converters must also be advertised.

Advertising the availability of each wavelength channel is not practical because of the huge number of wavelength channels, that can be supported per fiber. The concept of link bundling can help for scalability improvement. Data channels between two neighboring nodes can be aggregated into one logical link and advertised as one TE link with specific characteristics. For example, if there are N optical fibers between nodes A and B, all wavelength channels can be grouped in TE links according to their color, i.e. all N red wavelengths are included in the bundled red TE link between these two nodes. When the status of this link is advertised, the percentage of occupied red wavelengths is indicated, and thus an Edge node is aware of the availability of red wavelengths between nodes A and B. Such a scheme is beneficial also when the wavelength continuity constraint must be obeyed. However, it does not give any guarantees that a DB will be accepted even if the load level of the TE link is low, because of the asynchronous nature of the DB transportation mechanism.

7.7 Conclusion

This chapter introduces a very young problem that is still awaiting attention from the research community: the integration of the GMPLS framework within the OBS technology. The main motivation behind the need to integrate GMPLS and OBS are presented, along with thorough

elaboration on the strongest advantages, which can be achieved from such integration.

From the presented analysis it is clear that integrating the GMPLS control framework in OBS networks is beneficial. The GMPLS protocol suite supplements the missing functionalities in the conventional OBS control plane. It has the potential to enhance the scalability and the survivability of the OBS networks, and to improve their management. Nevertheless, not all benefits of GMPLS can be integrated in the OBS network. The main reason for this is that there cannot be strict QoS guarantees based on the GMPLS framework. Instead, OBS-specific techniques for QoS provisioning must be applied.

Different integration scenarios exist in the literature, depending on the application of the OBS paradigm and the underlying network architecture. The relation between GMPLS and OBS can either be an overlay type or an integrated type. In the overlay model the CPs are separated and the relation between them is hierarchical, following the standard GMPLS architecture. Such an integration scenario requires minimal, if any at all, changes in either of the technologies. This approach is suitable for GRID networking applications such as E-science and E-learning, where overlay networks need dynamic virtual topology re-configurations. Under the proposed integrated model the GMPLS protocol suite must be modified in order to fit the OBS CP functionalities. This approach requires more extensive changes of the GMPLS control framework and is applicable within single provider networks, where the CP is integrated and different transport technologies are applied in different parts of the network. Possible application is in metro-networks, which interconnect heterogenous access environments.

Adapting the GMPLS protocols to the OBS environment may raise some challenges, but is neither impossible nor inappropriate. Since both technologies are constantly being improved, they can be adjusted for a combined GMPLS-based OBS network. The main question is to what extent the technologies need to be modified. One of the main principles, on which they both are built, is flexibility. This provides different solutions for the outlined problems and gives the network designers a broad scope of possibilities to create highly flexible, scalable and well-performing OBS networks.

Chapter 8

Conclusion

This Ph.D. thesis focuses on several open issues in the area of multi-domain networking. The main investigated problems are related to protocol extensions for TE and QoS routing across multiple domains, survivability support (failure notification and recovery), and integrating novel switching paradigms within the GMPLS framework. Optical transport networks owned by different providers, employing the GMPLS control framework, have been considered. This thesis investigates the performance of different protocol extensions and modifications in multi-domain mesh network topologies.

The protocol of choice for the discussions in this thesis is the Border Gateway Protocol (BGP). Even though it is currently not favored for GMPLS multi-domain routing, it was illustrated that the protocol can provide potential benefits in the process of multi-domain connection establishment. In the presented work, it was shown that BGP can be successfully integrated within the GMPLS routing principles, by using BGP as a higher-level path selection protocol and by modifying the protocol to meet the routing requirements for connection-oriented networks. It was illustrated that using end-to-end TE information per path can aid the process of correct path selection. Furthermore, since the TE information is aggregated, it does not disclose sensitive topology and/or state information for the domains on the way. By means of simulations it was demonstrated that using just a simple end-to-end TE metric per path within the standard BGP protocol is not enough for providing good network performance, especially in networks with highly dynamic traffic

patterns and short connection durations.

The lack of TE dissemination within the standard BGP protocol is not its biggest drawback. Indeed, it was shown that the main challenges for the applicability of the protocol are its path dependency characteristic and the lack of path diversity. By modifying the protocol further into a path distribution protocol (referred to as *Enhanced BGP* throughout the thesis), enhanced with an end-to-end TE metric per path, it was illustrated that the performance of the network in terms of blocking ratio of requests improves significantly.

With respect to the efficiency of the proposed BGP extensions, two specific processes were thoroughly investigated: the process of updating the TE metric in a timely and efficient manner for supporting adequate TE, and the process of disseminating multiple paths with the *Enhanced BGP* in a scalable way. Avoiding synchronization in the process of updating the TE state of paths increases the performance of the evaluated routing schemes. Furthermore, introducing threshold-based procedures reduces the signaling overhead for TE-metric update in the network. Disseminating multiple paths per destination can inevitably lead to worse scalability. Thus, special export policies were defined for the *Enhanced BGP* proposal. It was shown that there is a fine trade-off between the number of disseminated paths and the achieved performance in terms of blocking of connection requests. In particular, using many long paths may increase the probability of finding disjoint paths for survivability support, but it also leads to increased blocking because of the paths being longer.

Another topic of this thesis is multi-domain survivability support. The main focus is on multi-domain link failures in mesh topologies, which necessitates domain cooperation. The open questions of failure notification and differentiated failure handling were investigated. The performance of a BGP extension, designed solely to derive two AS-disjoint paths per destination, was evaluated. It was illustrated that providing AS-disjoint paths is beneficial not only for recovery purposes in failure scenarios, but also for enhanced network performance when no survivability mechanisms have been employed in the network. Regarding the efficiency of different failure notification strategies, results indicate that depending on the position and the load the failed link different notification mechanisms have different impact on the connection blocking in the

network after link failure. More precisely, it was confirmed via extensive simulations that if more loaded links are failed, head-end notification is more effective for reducing blocking probability during routing protocol re-convergence.

Throughout this thesis, it was also argued that restoration strategies are necessary for providing different levels of service to customers. Thus, the efficiency of several restoration approaches has been evaluated. Applying restoration based on the proximity of the failed link to the source or the destination of a connection is a relatively simple approach for differentiated failure handling. Nevertheless, it showed the potential to provide a good balance between the provided restoration success ratio and the resource overbuild in the network. Focusing on the resource consumption during recovery, a mechanism for minimizing the recovery overbuild was proposed. Even though it does not provide a performance enhancement in terms of recovery success ratio, it provides better resource consumption under failure recovery. It was illustrated that depending on the network topology and the AS-paths, provided by the BGP protocol, different connections have better chance to be successfully recovered by different recovery approaches. In particular, if the failed link is in the core of the topology, it is more efficient to apply local-to-end restoration. This is due to the path dependency inherent to the BGP protocol, which leads to low utilization of multi-domain links in the core, and thus, increases the probability for successful local restoration.

The last topic covered is the integration of the Optical Burst Switching (OBS) technology into the GMPLS control framework. Due to the specifics of the OBS technology, it is difficult to integrate it into the GMPLS framework and to utilize all benefits the framework offers. In particular, strict QoS guarantees provided by explicit resource reservation cannot be achieved in a straightforward manner. Many aspects need to be considered in the course of potential integration including the exact integration and application scenarios, which may lead to either an overlay integration approach, or to peer integration approach.

From the presented results in this thesis one main conclusion can be drawn. In order to support TE and QoS provisioning across multiple domains, an advanced cooperative framework is required. A simple solution such as over-provisioning could improve multi-domain network

performance, but with the development of highly dynamic and bandwidth intensive services such a solution will not be efficient. Exchanging simple aggregated metrics is also a possible approach, but it was shown to be effective only for networks with very long connection durations. In addition, relying only on bi-lateral agreements between domains is inefficient on a global scale. In order to obtain global optimal performance, domains need to apply a more sophisticated cooperative framework which also takes network parameters in account, not only client QoS requirements.

The requirement for global cooperation is even more pronounced when network survivability is considered. The next generation transport infrastructure will comprise of heterogenous domains with different survivability capabilities. Providing a flexible framework for differentiated failure handling based on the abilities of the domains to support different protection/restoration mechanisms will inevitably improve the quality of the provided services. If network providers want to extend their service portfolio in a flexible and cost-efficient manner, they need to start cooperating on a higher level. Sharing limited state information and cooperating in case of failures can indeed improve the network performance, so the strict privacy preservation requirements need to be loosened up.

With the development of novel applications and services, which require multi-domain cooperation for efficient TE and QoS provisioning, it is important to evaluate the applicability of existing protocols and solutions. Taking advantage of the maturity of the BGP protocol and its strong features, such as policy enforcement and privacy preservation, it has the potential to be an attractive solution for facilitating inter-networking between multiple transport providers. Applying the proposed BGP extensions contributes to the development of a highly dynamic and automatic framework for multi-domain networking under the GMPLS control plane architecture. Furthermore, the performed investigations within the field of multi-domain resiliency contribute to the better understanding of the limitations the multi-domain environment poses to the standard resiliency approaches. The presented work reveals important dependencies, which need to be taken into account when novel survivability frameworks for heterogenous environments are designed.

Appendices

Appendix A

Model Description

The proposals described in chapters 3, 4, 5 and 6 were implemented in the event driven simulator OPNET [68]. This appendix gives details on the implementation, the relevant models and model parameters, and detailed description of the used and proposed mechanisms and algorithms.

A.1 Node model

Two types of nodes were used in the simulated networks - normal nodes (Edge, Core or Border) and domain configuration nodes. The domain configuration nodes are used for management purposes and topology discovery, and there is one per domain. Each normal node consists of two processes: traffic generator and packet dispatcher (see Fig. A.1). The traffic generator is responsible for generating connection requests (RSVP-TE PATH messages) based on specific traffic generation parameters (given as simulation parameters). Connection requests are generated according to a Poisson distribution, where each connection had exponentially distributed duration (connection holding time). The packet dispatcher is the parent process which handles incoming packets (both BGP-related and RSVP-TE-related) and different events such as failure handling, initiation of AS-disjoint path computation, etc. This process creates two types of child processes: BGP child process for each BGP neighbor the node has and RSVP-TE child process for each LSP request. Each node can be one of three types:

- Edge node: generates traffic and participates in the BGP process;
- Border node: does not generate traffic, and participates in the BGP process;
- Core node: it does not generate traffic and does not participate in the BGP process.

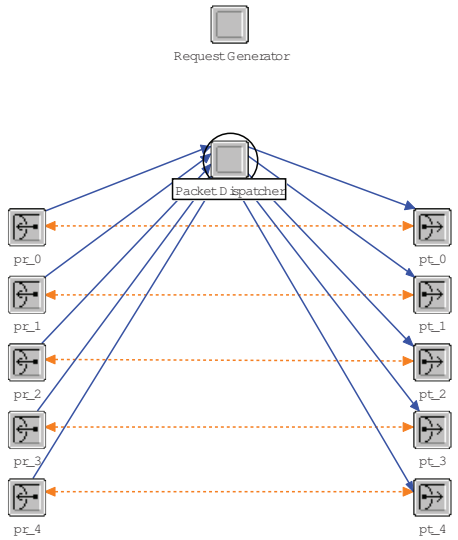


Figure A.1: Node model.

The process model of a packet processor is given on Fig. A.2. Each node, participating in the BGP process creates child processes to handle the BGP sessions with the node's neighbors. Furthermore, each LSP request is handled by a separate child process, created when PATH message (connection request) is received.

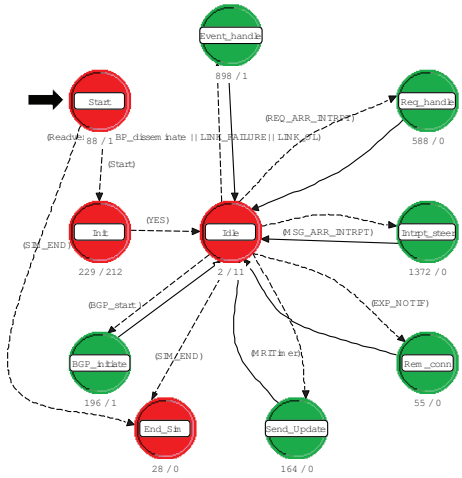


Figure A.2: Packet dispatcher process model.

A.2 BGP model

The model of the BGP process is given on Fig. A.3. The implemented BGP state machine is very simplified and contains only generic path selection method. All functions, related to the calculation of the TE state of a path, checks for path eligibility, management of the Adj_Rib_In, Adj_Rib_Out and the Local_Rib databases and Virtual Topology (VT) management belong to the BGP process. The BGP model is responsible for creating the multi-domain routing tables, i.e. the intra-domain routing is out of the scope of this part of the model. Upon network initiation each Edge node advertises itself as a reachable destination, creates UPDATE messages and sends them to all i-BGP neighbors in the domains. Upon receiving of an UPDATE message each BGP speaker checks the eligibility of the advertisement[1] and processes the path according to the path selection described in Section 2.4.

[1] An advertisement is eligible if it does not create routing loops, in terms of AS hops, and if the next node, indicated by the next_node attribute, is reachable.

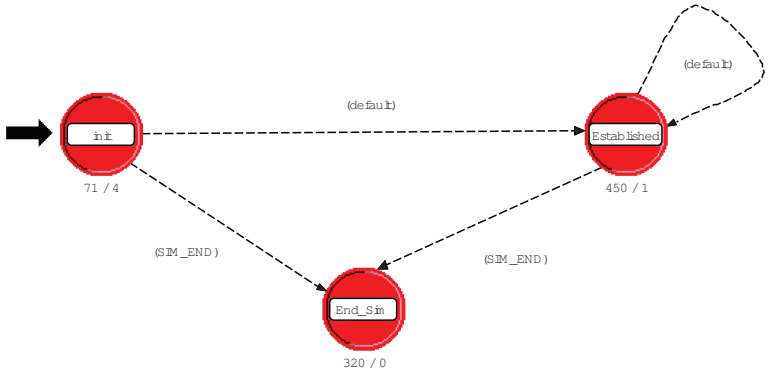


Figure A.3: BGP process model.

A.3 RSVP-TE model

Each connection request is handled by a RSVP-TE process model, which is enhanced for restoration management. The process model can be seen on Fig. A.4.

A.3.1 Failure free operation

Under failure free operation the RSVP-TE model follows the standard connection establishment procedure specified in [21]. Upon receiving a *PATH* message if the request is satisfiable (i.e. the outgoing label set is not empty and if the node is destination, a label has been found) the state-machine enters one of two possible states, depending on the node, processing the request. If the node is the destination for the request, the state machine enters the *Unres wait* state, else, it enters the *Res wait* state. Upon receiving a *RESV* message if the node is the source of the connection, the process enters the *Established* state, else it enters the *Unres wait* state. When the connection expires, the state machine in the source node enters state *Expired*, sends *Path_ Tear* message to the downstream node and destroys the process. If a node receives any type of Error message or a *Path_ Tear* message, the state machine goes in state *Closed*, performs all needed resource deallocation procedures, sends the required messages to down-/up- stream nodes and destroys the process.

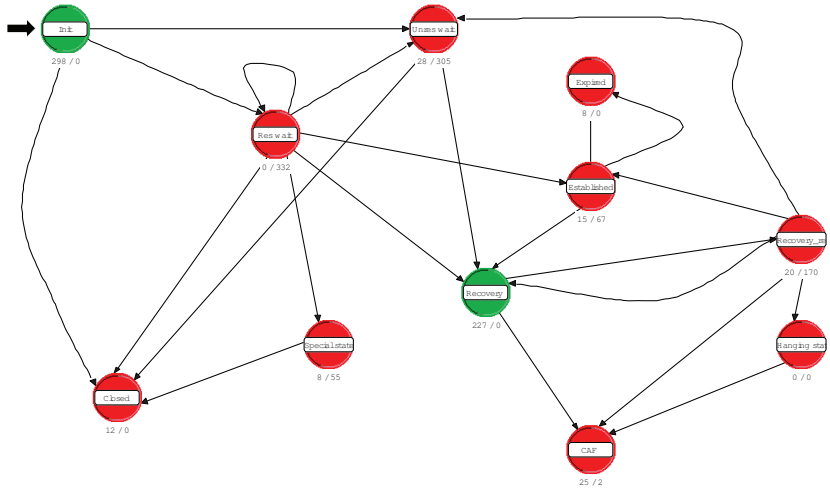


Figure A.4: RSVP-TE child process.

A.3.2 LSP restoration operation

Depending on the position of a node, with respect to the failed link (upstream or downstream) and the current state of the state machine, different procedures are executed. Fig. A.5 illustrates the options and shortly outlines the procedures.

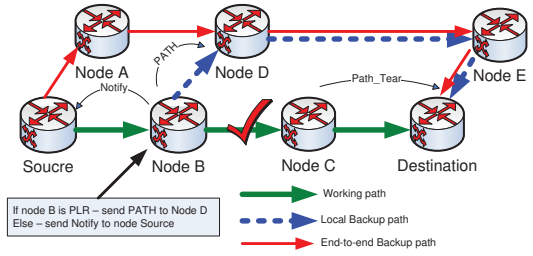


Figure A.5: Procedures in case of failure.

The node is adjacent to the failed link in downstream direction

In this case, the node is responsible for performing span release. If the process model is in *Unres wait* state, then the node releases any occupied resources and sends downstream a *Path_ Tear* message. Note, in this way the downstream nodes have no information for the failure in the network. If the process is in *Res wait* state, this means that the connection is still in process of establishment. The process enters the *Special state* where it waits for any notification from the downstream nodes. Irrespective of the received message in this state, the state machine enters state *Closed*, sends *Path_ Tear* messages if needed and destroys the process.

The node is adjacent to the failed link in upstream direction

In this case, the node is the one responsible for the failure handling. Depending on the applied method for recovery the node either performs local restoration or simply sends a failure notification upstream and performs span release. If the node is the point of local repair the state machine enters in state *Recovery*. In this state, new connection establishment is initiated (this is the recovery segment). If the newly computed outgoing label set is empty, then no recovery can be performed and the state machine enters state *CAF* where the process is destroyed. If the label set is not empty the state machine enters state *Recovery_ res* where the node waits for recovery confirmation. If RESV message is received, then the restoration path has been established successfully and the state machine enters state *Established*, if the node is the source of the initial connection (i.e. E2E restoration has been performed), or state *Unres wait*, if the node has been an intermediate node (i.e. L2E restoration has been performed). If the restoration was unsuccessful, i.e. an error message is received, and recovery retry is activated (controlled by a global simulation parameter), then the state machine enters state *Recovery* again and initiates new connection establishment. In the rare occasion of connection expiry during recovery, the state machine enters state *Hanging state*, where the notification for the establishment of the restoration path is awaited. Regardless of the result of the restoration the state machine enters state *CAF*, sends notification messages for span release if needed and destroys the process.

A.4 Model attributes

The attributes presented in Table A.4 can be given as initial parameters for simulations. They can be divided in two main groups: node and simulation attributes. Node attributes specify certain attributes per node and can have different values for different nodes. Simulation attributes are global and have the same value for all nodes. Setting up different attributes create diverse simulation scenarios.

ATTRIBUTE	POSSIBLE VALUES	DESCRIPTION
Node attributes		
AS number	Topology specific, integer	Specifies to which domain the node belongs
Node ID	Topology specific, integer	Uniquely identifies a node in the network
Node type	Edge/Core/Border	Uniquely specifies the role of the node in the topology
Number of WC*	Unlimited, integer	Number of Wavelength Converters per node
Simulation attributes		
BGP type	Normal/Enhanced	Specifies the BGP operation mode (see Chapter 2 and Chapter 3)
Backup path compute timer	Unlimited, double	Specifies when the disjoint path selection process should begin under the AS-disjoint BGP scheme
Compute Backup Paths	Enabled/Disabled	Indicates if AS-disjoint paths should be computed
Duration	Unlimited, double (sec.)	Specifies the mean value of the duration of the connections
E2E	Enabled/Disabled	Indicates if E2E (Enabled) or L2E (Disabled) restoration and failure notification is applied
Failed link A, Failed link B	Topology specific, integer	Indicates the link to fail by its end points
Failure time	Unlimited, double	Indicates the moment of link failure
Initial Interval	Unlimited, double	Indicates the Update Interval for the timer-based TE update schemes
LSP restoration	Enabled/Disabled	Indicates if Restoration should be applied after failure
Label Preference	NO/LS/Suggested Label	Indicates the label preference scheme for the RSVP-TE protocol operation
Link Overload Trigger	Enabled/Disabled	Indicates if a threshold-based TE metric update should be performed
MED	Disabled/Always Compare/ Deterministic/Modified	Specifies the procedure for handling the TE metric under BGP path selection
Mean Inter-arrival time	Unlimited, double (sec.)	Indicates the mean inter-arrival time for LSP requests
Restoration Retry	Enabled/Disabled	Indicates if Restoration retry should be performed
Re-convergence after failure	Enabled/Disabled	Indicates if the BGP protocol should be allowed to re-converge after failure
Shortest New	Enabled/Disabled	Indicates if Shortest New restoration should be activated
Simple Restoration	Enabled/Disabled	Indicates if the SLBR should be activated
TE metric	Hop count/Delay/ Wavelength availability	Indicates the type of TE metric used in the multi-domain network
Wavelength Assignment	Random/Fist Fit	Indicates the type of Wavelength assignment for the RSVP-TE protocol
Wavelengths per link	Unlimited, integer	Specifies the amount of wavelengths per link
eMRIT, iMRIT	Unlimited, double (sec.)	Indicates the values for the MRIT timers

Table A.1: Node and simulation attributes.

Appendix B

Topologies

This appendix presents the topologies, used for the simulations throughout this thesis. The parameters of all topologies are presented in Table B.1 at the end of the appendix.

B.1 Artificial topologies

Topology 1 (Fig. B.1) and *Topology 2* (Fig. B.2) were designed to be with high nodal degree, high inter-domain connectivity, relatively symmetric structure and no specified core area. *Topology 3* (Fig. B.3) is based on the NOBEL topology (see Fig. B.5) and is created by adding links and deleting domains for creating more uniform topology with no core area.

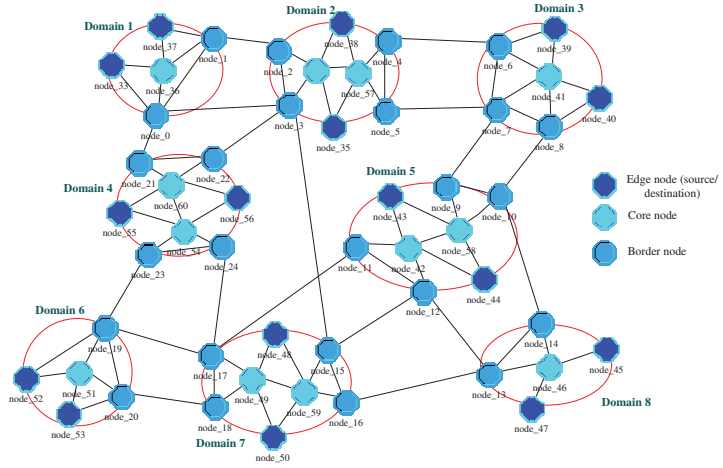


Figure B.1: Topology 1.

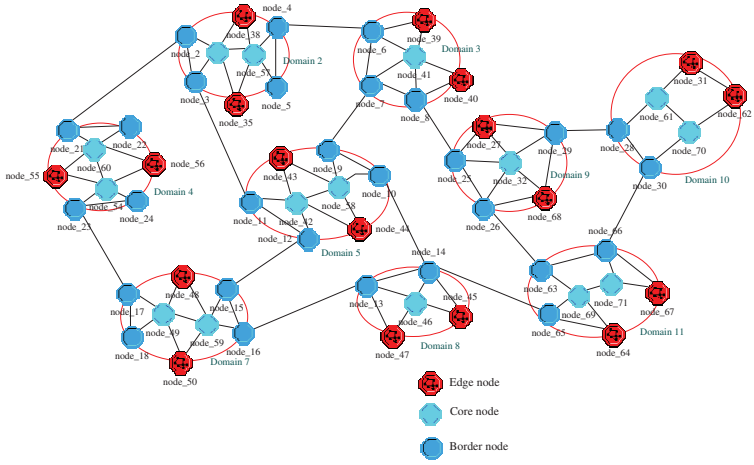


Figure B.2: Topology 2.

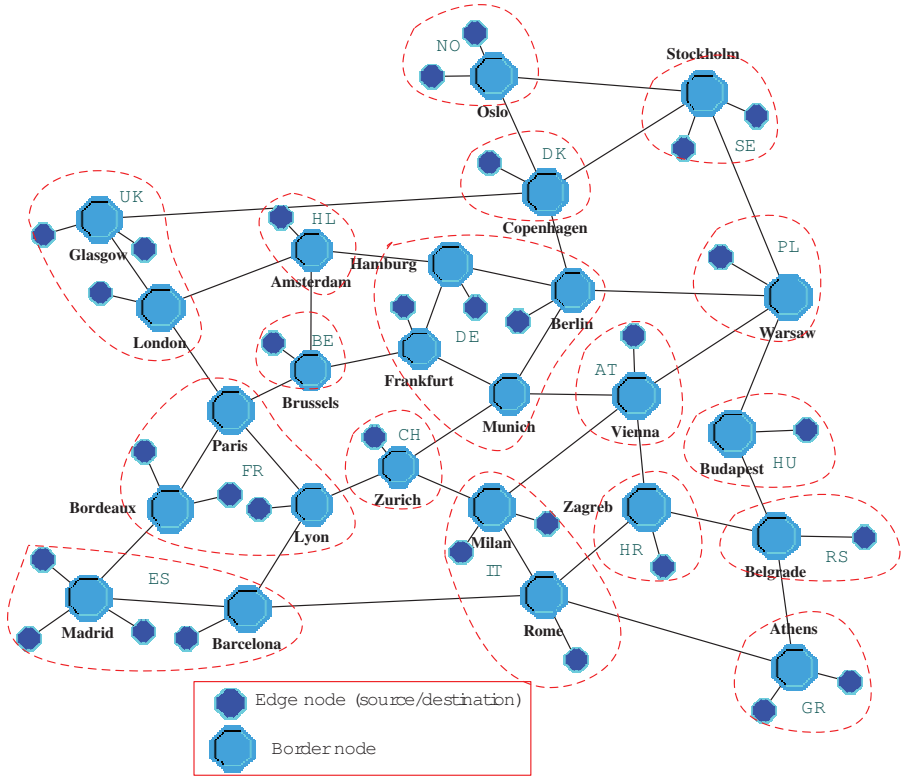


Figure B.3: Topology 3 - modified NOBEL topology.

B.2 Pan-European topologies

Two Pan-European topologies were used: the COST 266 topology [69] (Fig. B.4) and the NOBEL topology [100] (Fig. B.5). For both topologies, each country is a separate domain. The intra-domain topologies for each domain are randomly generated and have between one and four edge nodes (source/destination nodes). Both topologies have an effective core area around domain Germany (DE).

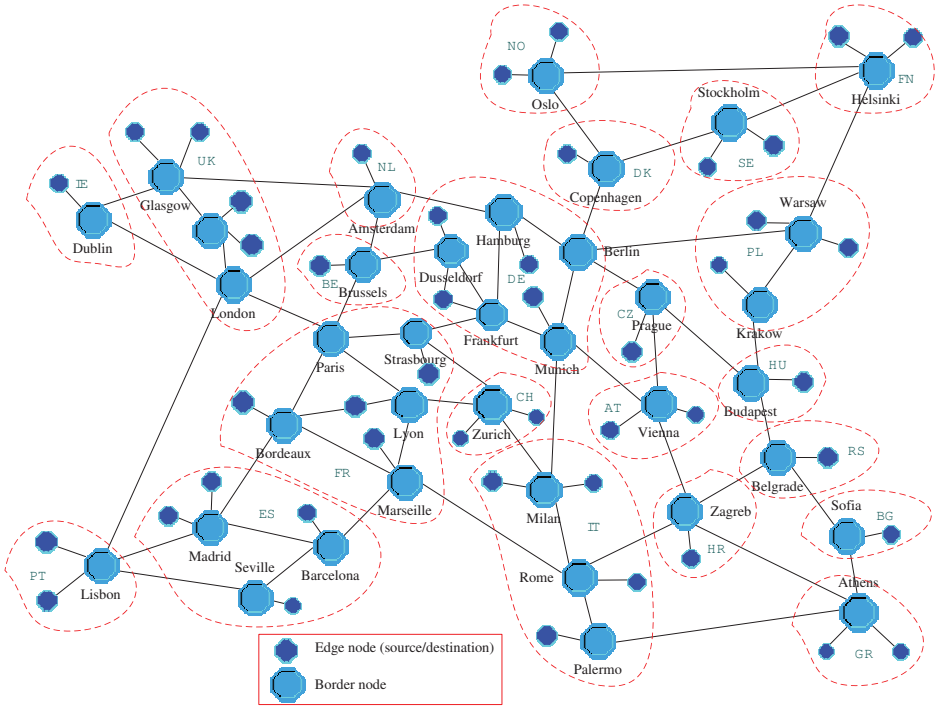


Figure B.4: COST 266 Pan-European topology.

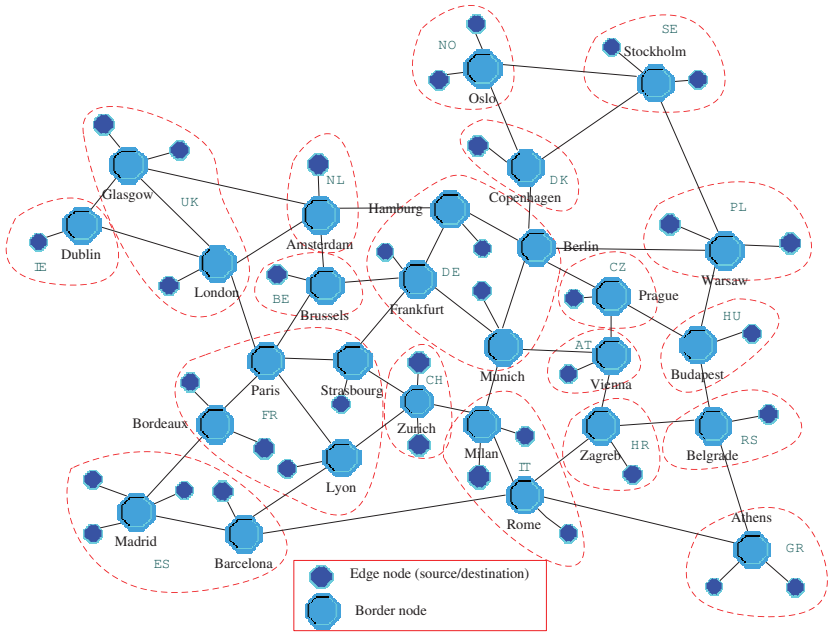


Figure B.5: NOBEL Pan-European topology.

B.3 NSFNET

This multi-domain topology (Fig. B.6), based on the NSFNET, is created by randomly grouping some of the nodes into separate domains. Each domain has one or two source/destination nodes.

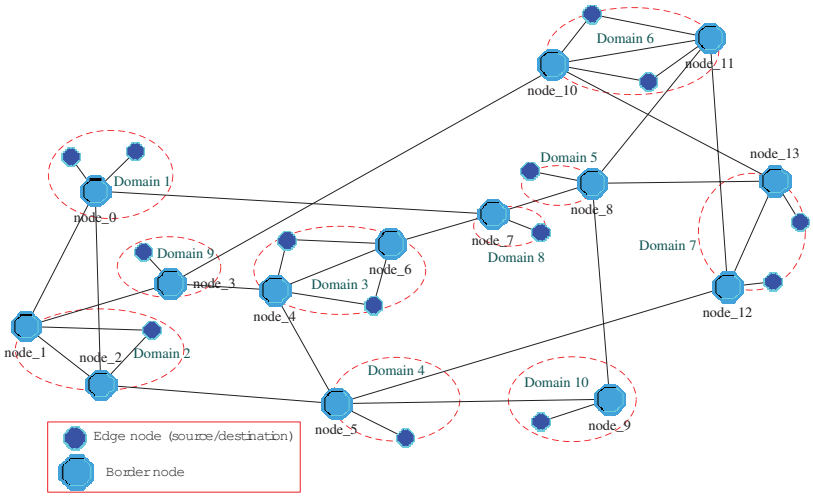


Figure B.6: NSFNET multi-domain topology.

Topology	# Domains	# Multi domain links	# Source/ Destination nodes	# Core nodes	# Border nodes	Average nodal degree*	Average domain degree**
Topology 1	8	18	16	12	25	3.39	3
Topology 2	9	13	18	15	29	3.17	2.89
Topology 3	17	28	31	0	25	3.04	3.29
COST 266	22	40	46	1	36	3.166	3.18
NOBEL	19	34	36	0	28	3.14	3.15
NSFNET	10	17	14	0	14	3	2.9

* Indicates the average connectivity of the border nodes.
** Indicates the average connectivity of the domains.

Table B.1: Topology parameters.

Bibliography

[1] A. Manolova, S. Ruepp, and L. Dittmann, "Performance comparison of Multi-domain Routing Schemes in GMPLS networks with BGP," in *IEEE Proc. 17th International Conference on Photonics in Switching (PS)*, Sept. 2009.

[2] N. Sambo, I. Cerutti, A. Giorgetti, P. Castoldi, R. Munoz, S. Ruepp, R. Casselas, R. Martinez, and A. Manolova, "Restoration GMPLS-based Wavelength Switched Optical Networks with Limited Wavelength Converters," in *IEEE Proc. 17th International Conference on Photonics in Switching (PS)*, Sept. 2009.

[3] S. Ruepp, A. Koster, N. Andriolli, and A. Manolova, "Prioritizing Connection Requests in GMPLS-Controlled Optical Networks," in *IEEE Proc. 17th International Conference on Photonics in Switching (PS)*, Sept. 2009.

[4] A. Manolova, E. Calle, S. Ruepp, J. Marzo, and L. Dittmann, "Location-based restoration mechanism for multi-domain GMPLS networks," in *IEEE Proc. International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, July 2009.

[5] A. Manolova, S. Ruepp, and L. Dittmann, "TE-enhanced path selection for QoS provisioning in multi-domain GMPLS networks," in *IEEE Proc. Optical Fiber Communication Conference/National Fiber Optic Engineers Conference (OFC/NFOEC)*, March 2009.

[6] ——, "On the efficiency of BGP-TE extensions for GMPLS multi-domain routing," in *IEEE Proc. 13th Conference on Optical Network Design and Modelling (ONDM)*, Feb. 2009.

[7] L. Xiaohua, S. Ruepp, A. Manolova and L. Dittmann, "Survivability enhancing routing scheme for multi-domain network," in *IEEE Proc. GLOBECOM 2008*, Dec. 2008.

[8] J. Buton, S. Ruepp, H. Wessing, N. Andriolli, A. Manolova, and L. Dittmann, "Wavelength converter placement in optical networks with dynamic traffic," in *Proc. Asia Pacific Optical Conference (APOC)*, Oct. 2008.

[9] L. Xiaohua, S. Ruepp, L. Dittmann and A. Manolova, "OPNET Model of Multi-domain Routing with Enhanced Survivability," in *Proc. OPNETWORK 2008)*, Aug. 2008.

[10] A. Manolova, J. Buron, S. Ruepp, L. Dittmann and L. Ellegaard, "Modeling contention resolution strategies in Optical burst switching networks," in *Proc. OPNETWORK 2006*, Aug. 2006.

[11] ——, "Segmentation-based path Switching Mecanism for Reduced Data Losses in OBS Networks," in *Proc. 11th International Conference on Optical Networking Design and Modeling (ONDM), Lecture Notes in Computer Science 4534*, May 2007.

[12] A. Manolova, S. Ruepp, J. Buron, L. Dittmann and L. Ellegaard, "Advantages and Challenges of the GMPLS/OBS Integration," in *Proc. VI GMPLS Workshop*, Apr. 2007, pp. 133–144.

[13] J. Buron, S. Ruepp, and A. Manolova, "Teaching Cost-effective and Resilient Network Design with OPNET WDM Guru," in *Proc. OPNETWORK 2007*, Aug. 2007.

[14] G. Bernstein, B. Rajagopalan and D. Saha, *Optical Network Control: Architecture, Protocols and Standards*. Addison-Wesley, 2004.

[15] B. Mukherjee, *Optical WDM Networks*. Springer - Optical Networks Series, 2006.

[16] ITU-T Rec. G.8080/Y.1304, "Architecture for the automatically switched optical network (ason)," 2003.

[17] E. Mannie ed., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture," *RFC 3945*, Oct. 2004.

[18] E. Rosen, A. Viswanatan, and R. Callon, "Multiprotocol Label Switching Architecture," *RFC 3031*, Jan. 2001.

[19] D. Katz, K. Kompella and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2," *RFC 3630*, Sept. 2003.

[20] H. Smit and T. Li, "Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)," *RFC 3784*, June 2004.

[21] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels," *RFC 3209*, Dec. 2001.

[22] J. Lang, Ed., "Link Management Protocol (LMP)," *RFC 4204*, Oct. 2005.

[23] A. Farrel and I. Bryskin, *GMPLS: Architecture and Applications*. Morgan-Kaufmann Publishers, Elsevier, 2006.

[24] Y. Xue, Ed., "Optical Network Service Requirements," *draft-ietf-ipo-carrier-requirements-05.txt*, Dec. 2002.

[25] ITU-T Rec. G.7715/Y.1706, "Architecture and requirements for routing in the automatically switched optical networks," 2003.

[26] The OIF Forum, "User Network Interface (UNI) 2.0 Signaling Specification, OIF-UNI-02.0-RSVP-RSVP Extenstions for User Network Interface (UNI) 2.0 Signaling," Feb. 2008.

[27] ——, "OIF E-NNI Signaling Specification, IA ♯ OIF-E-NNI-Sig-02.0," April 2009.

[28] ——, "External Network-Network Interface (E-NNI) OSPF-based Routing - 1.0 (Intra-Carrier) Implementation Agreement, OIF-ENNI-OSPF-01.0," Jan. 2007.

[29] ——, "OIF Guideline Document: Signaling Protocol Interworking of ASON/GMPLS Network Domains, IW ♯ OIF-G-Sig-IW-01.0," June 2008.

[30] A. Farrel, J.-P. Vasseur, and A. Ayyangar, "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering," *RFC 4726*, Nov. 2006.

[31] ITU-T Rec. G.805, "Generic functional architecture of transport networks," 2000.

[32] J.-P. Vasseur, Z. Ali, and S. Sivabalan, "Definition of a Record Route Object (RRO) Node-Id Sub-Object," *RFC 4561*, June 2006.

[33] L. Berger, Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions," *RFC 3473*, Jan. 2003.

[34] K. Kompella and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)," *RFC 4206*, Oct. 2005.

[35] A. Ayyangar, K. Kompella, J.-P. Vasseur, and A. Farrel, "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)," *RFC 5150*, Feb. 2008.

[36] A. Farrel, J.-P. Vasseur, and J. Ash, "A Path Computation Element (PCE) - Based Architecture," *RFC 4655*, Aug. 2006.

[37] Y. Rekhter, T. Li, and S. Hares, "A Border Gateway Protocol 4 (BGP-4)," *RFC 4271*, Jan. 2006.

[38] T. Otani, S. Okamoto, S. Okamoto, and W. Imajuku, "GMPLS Inter-domain Traffic Engineering Requirements," *draft-otani-ccamp-interas-gmpls-te-07.txt*, June 2008.

[39] D. Papadimitriou, L.Ong, J. Sadler, S. Shew and D. Ward, "Evaluation of Existing Routing Protocols against Automatic Switched Optical Network (ASON) Routing Requirements," *RFC 4652*, Oct. 2006.

[40] R. Douville, JL. Le Roux, JL. Rougier, and S. Secci, "A Service Plane over the PCE Architecture for Automatic Multidomain Connection-Oriented Servcies," *IEEE Comm. Mag.*, vol. 46, no. 6, pp. 94 – 102, June 2008.

[41] JP. Vasseur, and J.L. LeRoux, "Path Computation Element (PCE) Communication Protocol (PCEP)," *RFC 5440*, March 2009.

[42] J.-P. Vasseur, A. Ayyangar, and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)," *RFC 5152*, Feb. 2008.

[43] J.-P. Vasseur, R. Zhang, N. Bitar, and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths," *RFC 5441*, Apr. 200.

[44] Y. Xue, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in Multi-Domain Multiprotocol Label Switching Traffic Engineering and Generalized Multiprotocol Label Switching," *draft-king-pce-brpc-app-00.txt*, Sept. 2009.

[45] D. Amzallag, A. Farrel, and D. King, "Selecting Domain Paths in Inter-Domain MPLS-TP and MPLS-TE Networks," in *12th Annual MPLS Conference*, Oct. 2009.

[46] R. Martínez, R. Muñoz, M. Requena, J. Sorribes, J. Comellas, and G. Junyent, "ADRENALINE Testbed: architecture and implementation of GMPLS-based network resource manager and routing controller," in *IEEE/Create-Net Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM)*, March 2006.

[47] R. Casellas, T. Tsuritani, S. Okamoto, R. Martínez, and R. Muñoz, "Experimental Field-Trial of Multi-domain PCE-based Path Computation for OSNR-aware GMPLS enabled translucent WSON," in *35th European Conference and Exhibition on Optical Communication (ECOC)*, Sept. 2009.

[48] F. Parent, M. Blanchet, and B. St-Arnaud, "Optical BGP(OBGP): InterAS lightpath Provisioning," *ietf-draft-parent-obgp-01.txt*, March 2001.

[49] CANARIE Inc., "Canada's advanced research and innovation network," http://www.canarie.ca/en/home.

[50] M. Francisco, S. Simpson, L. Pezoulas, C. Huang, I. Lambadaris, and W. St-Arnaud, "Interdomain routing in optical networks," in *Optical Networking and Communications (OptiComm), SPIE*, vol. 4599, Aug. 2001, pp. 120–129.

[51] K. Shiomoto, D. Papadimitriou, JL. Le Roux, M. Vigoureux, and D. Brungard, "Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)," *RFC 5212*, July 2008.

[52] D. McPherson and V. Gill, "BGP MULTI EXIT DISC (MED) Considerations," *RFC 4451*, March 2006.

[53] S. Uhlig, O. Bonaventure, and B. Quoitin., "Interdomain Traffic Engineering with minimal BGP Configurations," in *Proc. 18th International Teletraffic Congress (ITC)*, Sept. 2003.

[54] J. Winick, S. Jamin, and J. Rexford, "Traffic Engineering Between Neighboring Domains," *available at http://www.cs.princeton.edu/ jrex/papers/interAS.pdf*, 2002.

[55] R. Chandra, P. Traina, and T. Li, "BGP Communities Attribute," *RFC 1997*, Aug. 1996.

[56] E. Chen and T. Bates, "An Application of the BGP Community Attribute in Multi-home Routing," *RFC 1998*, Aug. 1996.

[57] The ATM Forum, "Private Network-Network Interface Specification Version 1.1," *PNNI 1.1*, April 2002.

[58] G. Bernstein, V. Sharma, and L. Ong, "Interdomain Optical Routing," *Journal of Optical Networking*, vol. 1, no. 2, pp. 80 – 92, Feb. 2002.

[59] Q. Liu, M.A. Kok, N. Ghani, V.M. Muthalaly, and M. Wang, "Inter-Domain Provisioning in DWDM Networks," in *IEEE Proc. GLOBECOM 2006*, Nov. 2006.

[60] M. Yannuzzi, X. Masip-Bruin, S. Sanchez, J. Domingo-Pascual, A. Orda, and A. Sprintson, "On the challenges of establishing disjoint QoS IP/MPLS paths across multiple domains," *IEEE Comm. Mag.*, vol. 44, no. 12, pp. 60–66, Dec. 2006.

[61] A. DÆAchille, M. Listanti, U. Monaco, F. Ricciato, D. Ali, and V. Sharma, "Diverse Inter-Region Path Setup/Establishment," *draft-dachille-diverse-inter-region-path-setup-01.txt*, Oct. 2004.

[62] J.P. Lang, Y. Rekhter, and D. Papadimitriou, "RSVP-TE Extensions in support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS)- based Recovery," *RFC 4872*, May 2007.

[63] G. Cristallo, and C. Jacquenet, "Providing Quality of Service Indication by the BGP-4 Protocol: the QOS_NLRI attribute," *draft-jacquenet-qos-nlri-05.txt*, Dec. 2003.

[64] L. Xiao, J. Wang, K-S. Lui, and K. Nahrsted, "Advertising Interdomain QoS Routing Information," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 10, pp. 1949 – 1964, Dec. 2004.

[65] H. Ould-Brahim, D. Fedyk, and Y. Rekhter, "BGP Traffic Engineering Attribute," *RFC 5543*, May 2009.

[66] A. Muchanga, L. Wosinska, F. Orava, and J. Haralson, "Requirements for Interdomain Routing in Optical Networks," in *Proc. OFC/NFOEC'05*, March 2005.

[67] M. Yannuzzi, X. Masip-Bruina, S. Sánchez-Lópeza and E. Marín-Torderaa, "OBGP+: An improved path-vector protocol for multi-domain optical networks," *Optical Switching and Networking*, vol. 6, pp. 111 – 119, April 2009.

[68] OPNET Technologies, Inc., http://www.opnet.com.

[69] R. Inkret et al., "Advanced Infrastructure for Photonik Networks, Extended final report of COST Action 266," http://www.ufe.cz/dpt240/cost266/.

[70] J.W. Suurballe, "Disjoint Paths in a Network," *Networks*, vol. 4, no. 2, pp. 125 – 145, June 1974.

[71] N. Kushman, S. Kandula, D. Katabi, and B. Maggs, "R-BGP: Staying Connected in a Connected World," in *Proc. 4th USENIX Symposium on Networked Systems Design & Implementation*, April 2007.

[72] M. Bhatia, J.M. Halpern, and P. Jakma, "Advertising Multiple Next_Hop Routes in BGP," *draft-bhatia-bgp-multiple-next-hops-01.txt*, Aug. 2006.

[73] D. Walton, A. Retana, E. Chen, and J. Scudder, "Advertisement of Multiple Paths in BGP," *draft-walton-bgp-add-paths-06.txt*, July 2008.

[74] J.-P. Vasseur, M. Pickavet and P. Demeester, *Network Recovery, Protection and Restoration of Optical, SONET-SDH, IP, and MPLS*. Morgan-Kaufmann Publishers, Elsevier, 2004.

[75] P. Pan, G. Swallow, and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels," *RFC 4090*, May 2005.

[76] J. L. Marzo, E. Calle, C. Scoglio, and T. Anjah, "QoS On-Line Routing and MPLS Multilevel Protection: a Survey," *IEEE Comm. Mag.*, vol. 41, no. 10, pp. 126–132, Oct. 2003.

[77] D. Larrabeiti, R. Romeral, I. Soto, M. Uruena, T. Cinkler, J. Szigeti, and J. Tapolcai, "Multi-domain issues of resilience," in *Proc. 7th International Conference on Transparent Optical Networks*, July 2005, pp. 375–380.

[78] C. Huang and D. Messier, "A Fast and Scalable Interdomain MPLS Protection Mechanism," *Journal of Communications and Networks*, vol. 6, no. 1, March 2001.

[79] D. Staessens, D. Colle, U. Lievens, M. Pickavet, P. Demeester, W. Colitti, A. Nowe, K. Steenhaut, and R. Romeral, "Enabling High Availability over Multiple Optical Networks," *IEEE Comm. Mag.*, vol. 46, no. 6, pp. 120–126, June 2008.

[80] J. Szigeti, L. Gyarmati, and T. Cinkler, "Multidomain shared protection with limited information via MPP and p-cycles," *Journal of Optical Networking*, vol. 7, no. 5, pp. 400–409, Apr. 2008.

[81] J. Szigeti, R. Romeral, T. Cinkler, and D. Larrabeiti, "P -cycle Protection in multi-domain optical networks," *Photonic Network Communications*, vol. 17, no. 1, pp. 35–47, Feb. 2009.

[82] Z. Gao, and H. Naser, "End-to-end shared restoration algorithms in multi-domain mesh networks," in *IEEE Symposium on Computers and Communications (ISCC)*, July 2008, pp. 411–416.

[83] H. Drid, S. Lahoud, B. Cousin, and M. Molnar, "Survivability in multi-domain optical networks using p-cycles," *Photonic Network Communications*, Sept. 2009.

[84] D. Papadimitriou, and E. Mannie, "Analysis of Generalized Multi-Protocol Label Switching (GMPLS)-based Recovery Mechanisms (including Protection and Restoration)," *RFC 4428*, March 2006.

[85] J. Segovia, E. Calle, P. Vila, and J. Marzo, "Topology-focused availability analysis of basic protection schemes in optical transport networks," *Journal of Optical Networking*, vol. 7, no. 4, Feb. 2008.

[86] H. Lin, and W. Chang, "Integration of Differentiated Services in Optical Burst Switching Metro Ring Networks," in *31st IEEE Conference on Local Computer Networks*, Nov. 2006, pp. 327–334.

[87] C. Qiao, and M. Yoo, "Optical burst switching (OBS) - a new paradigm for an optical Internet," *Journal of High Speed Networks*, vol. 8, no. 1, pp. 69–84, March 1999.

[88] J.P. Jue, and V.M. Vokkarane, *Optical Burst Switched Networks*. Springer Science + Business Media, Inc., 2005.

[89] Y. Xiong, and M. Vandenhoute, "Control Architecture in Optical Burst-Switched WDM networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 10, Oct. 2000.

[90] F. Farahmand, J. Rodrigues, and J.P. Jue, "A Layered Architecture for Supporting Optical Burst Switching," in *Proc. AICT'05*, 2005.

[91] G. Hjalmtysson, J. Yates, S. Chauduri, and A. Greenberg, "Smart Routers - Simple Optics: An Architecture for the Optical Internet," *IEEE J. Lightw. Technol.*, vol. 18, no. 12, Dec. 2000.

[92] C. Xin, Y. Ye, T. Wang, S. Dixit, and C. Qiao, "On an IP-Centric Optical Control Plane," *IEEE Comm. Mag.*, vol. 39, no. 9, pp. 88–93, Sept. 2001.

[93] X. Yang, K. Long, X. Yang, Y. Zhang, and G. Liu, "A General Framework for GMPLS-Based OBS Networks," in *Proc. Network Architectures, Management, and Applications III, SPIE*, vol. 6022, Dec. 2005.

[94] K. Long, Z. Yi, Y. Xin, X. Yang, and H. Liu, "Generalized MPLS (GMPLS) architecture's extensions for Optical Burst Switch network," *draft-long-gmpls-obs-00.txt*, Nov. 2005.

[95] P. Pedroso, J. Sole-Pareta, D. Careglio, and M. Klinkowski, "Integrating GMPLS in the OBS control palne," in *Proc. ICTON'07*, vol. 3, July 2007, pp. 1–7.

[96] H. Guo, T. Truritani, Y. Yin, T. Otani, and J. Wu, "Proposal of a Multi-layer Network Architecture for OBS/GMPLS Network Interworking," in *Proc. Network Architectures, Management, and Applications V, SPIE*, vol. 6784, Nov. 2007.

[97] C. Qiao, W. Wei, and X. Liu, "Extending Generalized Multiprotocol Label Switching (GMPLS) for Polymorphous, Agile, and Transparent Optical Networks (PATON)," *IEEE Comm. Mag.*, vol. 44, no. 12, pp. 104–114, Dec. 2006.

[98] C. Qiao, "Labeled Optical Burst Switching for IP-over-WDM Integration," *IEEE Comm. Mag.*, vol. 38, no. 9, pp. 104–114, 2000.

[99] J. K. Choi, M. Kang, "Research Activities on Optical Burst Switching at OIRC," in *IEEE Proc. GLOBECOM 2003*, Dec. 2003.

[100] IST project NOBEL, "Nobel - next generation optical networks for broadband european leadership," obtained from http://sndlib.zib.de.