ABSTRACT

ON IP OVER WDM BURST-SWITCHED LONG HAUL AND METROPOLITAN AREA NETWORKS

by
Jingxuan Liu

The IP over Wavelength Division Multiplexing (WDM) network is a natural evolution ushered in by the phenomenal advances in networking technologies and technical breakthroughs in optical communications, fueled by the increasing demand in the reduction of operation costs and the network management complexity. The unprecedented bandwidth provisioning capability and the multi-service supportability of the WDM technology, in synergy with the data-oriented internetworking mechanisms, facilitates a common shared infrastructure for the Next Generation Internet (NGI).

While NGI targets to perform packet processing directly on the optical transport layer, a smooth evolution is critical to success. Intense research has been conducted to design the new generation optical networks that retain the advantages of packet-oriented transport prototypes while rendering elastic network resource utilization and graded levels of service.

This dissertation is focused on the control architecture, enabling technologies, and performance analysis of the WDM burst-switched long haul and Metropolitan Area Networks (MANs). Theoretical analysis and simulation results are reported to demonstrate the system performance and efficiency of proposed algorithms.

A novel transmission mechanism, namely, the Forward Resource Reservation (FRR) mechanism, is proposed to reduce the end-to-end delay for an Optical Burst Switching (OBS)-based IP over WDM system. The FRR scheme adopts a Linear Predictive Filter and an aggressive reservation strategy for data burst length prediction and resource reservation, respectively, and is extended to facilitate Quality of Service (QoS) differentiation at network edges. The FRR scheme improves

the real-time communication services for applications with time constraints without deleterious system costs.

The aggressive strategy for channel holding time reservations is proposed. Specifically, two algorithms, the success probability-driven (SPD) and the bandwidth usage-driven (BUD) ones, are proposed for resource reservations in the FRR-enabled scheme. These algorithms render explicit control on the latency reduction improvement and bandwidth usage efficiency, respectively, both of which are important figures of performance metrics.

The optimization issue for the FRR-enabled system is studied based on two disciplines - addressing the static and dynamic models targeting different desired objectives (in terms of algorithm efficiency and system performance), and developing a "crank back" based signaling mechanism to provide bandwidth usage efficiency. The proposed mechanisms enable the network nodes to make intelligent usage of the bandwidth resources.

In addition, a new control architecture with enhanced address resolution protocol (E-ARP), burst-based transmission, and hop-based wavelength allocation is proposed for Ethernet-supported IP over WDM MANs. It is verified, via theoretical analysis and simulation results, that the E-ARP significantly reduces the call setup latency and the transmission requirements associated with the address probing procedures; the burst-based transport mechanism improves the network throughput and resource utilization; and the hop-based wavelength allocation algorithm provides bandwidth multiplexing with fairness and high scalability. The enhancement of the Ethernet services, in tandem with the innovative mechanisms in the WDM domain, facilitates a flexible and efficient integration, thus making the new generation optical MAN optimized for the scalable, survivable, and IP-dominated network at gigabit speed possible.

ON IP OVER WDM BURST-SWITCHED LONG HAUL AND METROPOLITAN AREA NETWORKS

by
Jingxuan Liu

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2004

APPROVAL PAGE

ON IP OVER WDM BURST-SWITCHED LONG HAUL AND METROPOLITAN AREA NETWORKS

Jingxuan Liu

Dr. Nirwan Ansari, Dissertation Advisor — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Teunis J. Ott, Dissertation Co-Advisor — Date
Professor of Computer Science, NJIT

Dr. Edwin Hou, Committee Member — Date
Associate Professor of Electrical and Computer Engineering, NJIT

Dr. Joseph Leung, Committee Member — Date
Distinguished Professor of Computer Science, NJIT

Dr. James McHugh, Committee Member — Date
Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Jingxuan Liu

Degree: Doctor of Philosophy

Date: May 2004

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2004
- Master of Science in Computer Application,
  Beijing University of Posts and Telecommunications, Beijing, P.R. China, 1999
- Bachelor of Science in Computer Communications,
  Beijing University of Posts and Telecommunications, Beijing, P.R. China, 1996

Major: Computer Science

Presentations and Publications:

J. Liu and N. Ansari, "A New Control Architecture With Enhanced ARP, Burst-Based Transmission, and Hop-Based Wavelength Allocation for Ethernet-Supported IP-Over-WDM MANs," *IEEE Journal on Selected Areas in Communications*, Accepted.

J. Liu, N. Ansari, and T. Ott, "FRR for Latency Reduction and QoS Provisioning in OBS Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 21, No. 7, pp. 1210-1219, Sept. 2003.

J. Liu and N. Ansari, "On Aggressive Resource Reservation for OBS Systems," *IEE Proceedings on Communications*, Vol 150, No. 4, pp. 233-238, Aug. 2003.

L. Zhu, N. Ansari, and J. Liu, "Throughput of HighSpeed TCP in Optical Burst Switching Networks," *IEEE Communications Letters*, Submitted.

J. Liu and N. Ansari, "The Impact of the Burst Assembly Interval on the OBS Ingress Traffic Characteristics and System Performance," *IEEE International Conference on Communications*, (ICC 04), Jun. 20-24, 2004, to be presented.

J. Liu and N. Ansari, "A Bandwidth Enhancement Mechanism for FRR-Enabled OBS Networks," *Proceedings of IEEE 2003 Workshop on High Performance Switching and Routing (HPSR)*, pp.135-139, Jun.24-27, 2003.

J. Liu and N. Ansari, "Determining the Channel Holding Time Adjustment for the FRREnabled OBS Networks," *Proceedings Of the 37th Annual Conference on Information Science And Systems (CISS 03)*, March 12-14, 2003.

J. Liu and N. Ansari, "Forward Resource Reservation for QoS Provisioning in OBS Systems," *Proceedings of IEEE Globecom2002*, pp. 2777-2781, Nov. 17-21, 2003.

J. Liu and N. Ansari, "Class-Based Dynamic Buffer Allocation for Optical Burst-Switched Networks," *Proceedings of IEEE 2002 Workshop on High Performance Routing and Switching*, (HPSR 2002), pp. 295-299, May 26-29, 2002.

G. Nong, J. Liu, N. Ansari, and S. Zhang, "Improving the Uplink Saturated Throughput for DCF-enabled Wireless Networks," in preparation.

To my whole-heartedly beloved Mom and Dad,
Fenglan Hou and Changxiang Liu,
Whose love and belief
Are my sources of happiness and strength;
Whose unwavering support and inspriation
at every step of my life,
Lift me on the wings of thoughts.

谨以此文献给我亲爱的父亲母亲 - -

感谢你们赋予我生命，引导我人生

你们深厚的爱是我幸福的源泉

你们的信任放飞我思想自由远翔

我坚强，只为有最眷恋的港湾

依泊我成长的欢乐与心伤

我骄傲，只为是你们温柔臂挽中的宠儿

将爱怜与叮嘱--我生命之羹--贪享

ACKNOWLEDGMENT

First of all, I owe my deepest gratitude to my advisor and mentor, Dr. Nirwan Ansari. This work would not have been possible without his indispensable efforts and involvement. I appreciate his insightful guidance, substantial assistance, and enthusiastic encouragement at every step of my progress. An excellent advisor and unstinting friend, he has ironed out the focus of my research work, and has fostered me with courage and confidence to aim high. I would also thank him for the time, efforts, and patience which paved the way for me to conduct the research without uncertainty. My academic experience would not be as fruitful today but for his protection.

I am most pleased to acknowledge my co-advisor, Dr. Teunis Ott, for his valuable discussions and suggestions. He has always willingly given his time and insights when I need helps, advice, and suggestions at each stage of the research work. His excellent lectures provided me the foremost knowledge on the Internet.

My cordial appreciation goes to Dr. Joseph Leung for his open-minded leadership that allows me to conduct interdisciplinary research across the departmental boundary. Without his unprecedented support, I could not fullfil my interested research under the guidance of Dr. Ansari and Dr. Ott.

I am also indebted to my committee members, Dr. Edwin Hou and Dr. James Mchugh, for their being interested in this work and their heedful reading of the dissertation. They provide the support that I am very pleased to acknowledge.

My thankfulness also goes to the support from the New Jersey Commission on Higher Education via the NJI-TOWER project, and my fellow colleagues at the Advanced Networking Laboratory who kept up my colorfulness of the daily life.

Last, though not least, words can never express my gratitude to my parents, my husband, and my sisters. Their love and passion have been a great source of encouragement and incentive that made this endeavor possible.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF ACRONYMS

ADM	Add Drop Multiplexer
ARP	Address Resolution Protocol
ATM	Asynchronous Transmission Mode
BCU	Burstification Control Unit
BEFRR	Bandwidth Enhanced Forward Resource Reservation
BER	Bit Error Rate
BHP	Burst Header Packet
BUD	Bandwidth Usage Driven
CDF	Cumulative Distribution Function
CoS	Class of Service
DR	Delayed Reservation
DXC	Digital Cross-connect
E-ARP	Enhanced Address Resolution Protocol
FDL	Fiber Delay Lines
FFT	Fast Fourier Transform
FGN	Fractional Gaussian Noise
FRR	Forward Resource Reservation
GbE	Gigabit Ethernet
GMPLS	Generalized Multiprotocol Label Switching
IP	Internet Protocol
JET	Just-Enough-Time
JIT	Just-In-Time
LAM	Local Address Mapping
LAN	Local Area Network
LMS	Least Mean Square

LIST OF ACRONYMS
(Continued)

LPF	Linear Predictive Filter
LS-burst	Limited Size Burst
MAC	Media Access Control
MAN	Metropolitan Area Network
MIB	Management Information Base
MPLS	Multiprotocol Label Switching
MTIT	Multitoken Interarrival Time
NFRR	None Forward Resource Reservation
GNI	Next Generation Internet
NS-burst	Non-limited Size Burst
OADM	Optical Add Drop Multiplexer
OAMP	Operation, Administration, Maintenance, and Provisioning
OBS	Optical Burst Switching
OC	Optical Channel
OEO	Optic-Electric-Optic
OPS	Optical Packet Switching
OTN	Optical Transport Network
OXC	Optical Cross-Connection
PDF	Probability Density Function
QoS	Quality of Service
RAM	Remote Address Mapping
RMS	Root Mean Square
RWA	Routing and Wavelength Assignment
SCU	Switching Control Unit

LIST OF ACRONYMS
(Continued)

SDH	Synchronous Digital Hierarchy
SLA	Service Level Agreement
SNR	Signal to Noise Ratio
SONET	Synchronous Optical Network
SPD	Success Probability Driven
TAG	Tell-and-Go
TAW	Tell-and-Wait
TDM	Time Division Multiplexing
VF	Void Filling
WDM	Wavelength Division Multiplexing

CHAPTER 1

INTRODUCTION

1.1 IP Over WDM

The IP over Wavelength Division Multiplexing (WDM) network is a natural evolution ushered in by the phenomenal advances of networking technologies and technical breakthroughs in optical communications, fueled by the increased demands for the elimination of unnecessary network layers that will lead to a vast reduction in the cost and complexity of the network [1, 2, 3]. A growing concensus about the network evolution is that the Next Generation Internet (NGI) will be an IP-based WDM network.

1.1.1 Wavelength Division Multiplexing Networks

WDM is an approach to unleash the potential fiber bandwidth. In a WDM system, the optical transport spectrum is carved up into a number of non-overlapping wavelength bands. Multiple wavelengths of light are transmitted simultaneously over a single fiber optic line, with each wavelength supporting one or multiple communication channels [4]. Each wavelength can be individually transported and routed through the network, and independently recovered by wavelength-selective components. WDM enables the utilization of a significant portion of the available fiber bandwidth, and is expected to be one of the methods of choice for future ultra-high bandwidth multi-channel systems.

Some technological breakthroughs enabling the reach of WDM systems include, but not limited to:

- Special fibers with nonzero dispersion characteristics
- Erbium-doped fiber amplifier

- The tunable laster diode operating around 1550nm
- In-fiber Bragg grating

These technological breakthroughs, in tandem with the emergence of optical networking devices, such as Optical Add Drop Multiplexers (OADMs) and Optical Cross-connects (OXCs), enable the WDM technology to inroad from a bandwidth provider to a networking solution. Its multi-channel concurrent transport capability, together with the accommodation for transparent traffic delivery, makes the WDM-enabled network an ideal platform to deliver the traffic of mixed type applications.

While the existing optical communications technologies make a good start, the WDM-enabled optical transport network is at the same time a great asset and a great challenge as well. New architectures and protocols that fully exploit the optical bandwidth capacity, coincide with the current and foregoing service requirements, and retain the advantages of existing transmission mechanisms, are of the essence to improve the overall value of optical networks.

1.1.2 The Evolution of IP Over WDM

In late 1990s, one of the characteristics of typical IP over optical networks is that data payload is piggybacked over traditional Time Division Multiplexing (TDM) based optical transport mechanism [5]. Such a network consist of four layers: the IP layer for network interconnection, the ATM layer renders traffic engineering, the SONET/SDH layer facilitates augmented transport and network survivability, and the WDM layer provides transmission capacity.

The network evolution has been impacted by several driving forces, which can be summarized into the following categories:

- Development of networking technologies and optical communications
- Dominating traffic trends in terms of traffic volume and traffic types

- Internet economy and market requirements

First and foremost, some important development at the IP layer expedites the evolution of network architectures. Faster and denser IP routers continue to evolve and spread to network edges, enabling the trunk speed to match the aggregation level of the optical transport layer. Meanwhile, Service Level Agreement (SLAs), QoS, and legacy integration are being provisioned by routers with increasing functionalities and protocols, such as the Multiprotocol Label Switching (MPLS) protocol. MPLS enables the network to provide layer 2 features on a layer 3 network, therefore providing engineers a solution that they need to guarantee the quality performance and traffic engineering of the IP network without resorting to ATM and its inherent complexity and scalability issues. Along with the integration of IP routing (intelligent to forward datagrams) and the ATM switching (high speed and high capacity connectivity), the ATM function of traffic engineering is being absorbed into the IP layer.

At the same time, with the availability of fast-service provisioning and network survivability mechanisms in the optical domain, and with the basic unit of transport bandwidth shifting from time slot to optical channel, the WDM layer advances beyond a bandwidth provider, absorbing the transport capacity of SONET/SDH in the optical optical domain. In much the same way that Digital Cross-connects (DXCs) emerge to manage network connectivity at the electrical layer, OXCs emerge to manage connectivity at the optical layer. In addition, flexible and reconfigurable OADMs have become an integral element of WDM networks to add and/or drop wavelengths at the intermediate nodes.

Therefore, what has been four layers converges to the two-layer model, whereby the IP network is built directly on the WDM optical infrastructure, while the intermediate layers, including the ATM layer and the SONET layer, are bypassed (Figure 1.1).

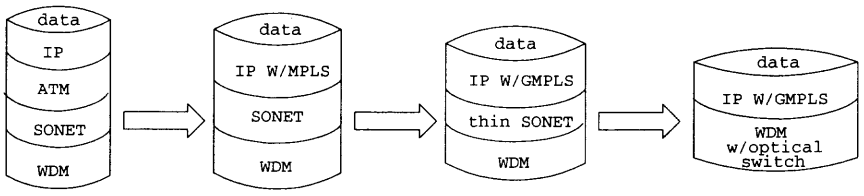

Figure 1.1 The evolution of the IP over WDM integration.

Second, the evolution of the two-layer model is fueled by the growth of Internet traffic. Besides the exponential growth of bandwidth demand due to internet and intranet communications, the shift in the service mix from circuit traffic to data-dominated traffic is well underway. As data traffic is increasing at a tremendous rate, and this trend is likely to continue in the future, it has been widely recognized that the new generation Internet should be optimized for data prototypes, and a true full-service network will emerge [1, 2, 3].

The emergence of multi-type applications, such as data, voice, and video-conferencing, requires more increased-capacity, data-aware infrastructures than found in traditional multi-layer solutions. The IP over WDM infrastructure functions as a new breed of flexible, scalable, multi-service delivery platforms that offers the cost-effective insurance needed to exploit the increasing capabilities of WDM networks and the mature data transport technologies, and to accommodate the realities of the future network churn.

Third, besides the network technologies and the application requirements, the Internet economy and commerce revenue have direct impact on the network evolution. From the network Operation, Administration, Maintenance, and Provisioning (OAMP) point of view, the multi-layer model involves more vendor integration, multiple network management systems, and increased capital and operational cost. From the data plane point of view, the advanced ATM features at the same time introduce shortcomings, e.g., the cell tax on variable-length IP packet. In addition, SONET, which is a voice-oriented transport solution rooted in telephony, demon-

strates inefficiency and inflexibility as the packet-centered traffic pattern prevails. The graded bandwidth provisioning and the Optic-Electric-Optic (OEO) conversion required by the ADMs and DXCs also make SONET costly.

The IP over WDM infrastructure combines gigabit and terabit IP routers with WDM switching and transmission systems to create an optimized optical transport network. Such integration features the advantages of reduced network management overlay, maximized interoperability, and minimized number of service interfaces.

Based on the two-layer model, the IP over WDM integration can be viewed as an underlying all-optical layer upon which either IP routers lead to higher user-oriented protocol layers, or OEO intermediate nodes lead to the next island (Figure 1.2).

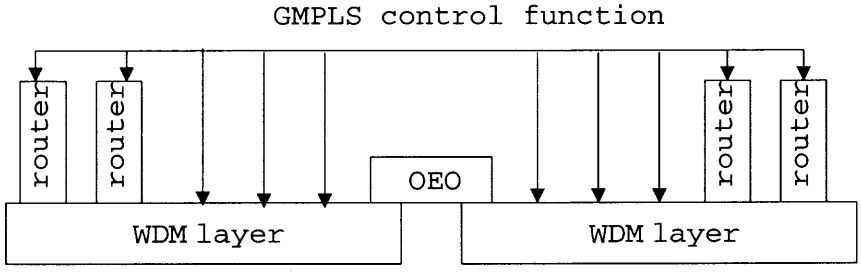

Figure 1.2 The IP over WDM topology.

1.1.3 Architecture Model

The development of IP over WDM integration is an evolutionary process, which is generally classified into three generations [2]:

- IP over point-to-point WDM
- IP over re-configurable WDM
- IP over switched WDM

In the IP over point-to-point WDM model, IP routers are directly connected with each other using multi-wavelength fibers. The WDM links provide only transmission capability, while the network OAMP functionalities and the traffic engineering are typically performed in IP layers. The IP over point-to-point WDM architecture is deemed as the first realization of optical networks based on multiple wavelengths. The network topology is fixed, and the network configurations are all static. Such an integration places new demand on the service layer for through-traffic networking and survivability, on top of the already growing service layer requirement.

Beyond this point-to-point WDM framework, the next evolutionary step in reliable, scalable transport networks will be full functionality optical networking, whereby besides the transmission capability, the WDM layer also participates in network control and management, and supports traffic engineering capability, such as routing and switching, QoS, scalability, protection/restoration, etc.

In the IP over reconfigurable WDM architecture, the interfaces of the IP routers are connected to the ports of OXCs, which are by themselves interconnected in a mesh configuration with multi-wavelength fiber links. By appropriately configuring the OXCs, a given router interface can be connected to any other interfaces of any other router. As a result, the neighboring routers for a given router interface are configurable.

The IP over reconfigurable WDM architecture establishes the future-ready ubiquitous infrastructure for full-functionality optical transport networking. It enables the migration from rigid and separately managed overlay networks to streamlined architectures which enable the set-up and tear-down of optical connections. By interconnecting the networks at the optical layer, the network OAMP complexity and the associated costs are reduced or eliminated, and the service-layer scalability, restoration and survivability are provisioned.

In the IP over switched WDM network, the WDM layer directly supports the on-demand traffic switching capability, as opposed to simply supporting ingress-to-egress lightpaths. As such, it enables a much finer grain sharing than re-configurable WDM system does, and addresses the requirements of the core network in a real-time fashion, particularly in the areas of traffic engineering, QoS, connection management, and restoration/protection.

The IP over switched WDM network is still evolving, with the focus on building up the IP-based full-functionality Optical Transport Networks (OTN) (in that besides the transmission capability, the WDM layer also provides multiplexing, supervision, and survivability for a wide range of client signals). IP over switched WDM networks are deemed as the natural choice for NGI owing to, from the networking perspective, its essential advantages such as efficient and cost effective capacity expansion, flexible optical-channel bandwidth management, and survivability mechanisms to support improved reliability of data networks.

1.2 Switching Technologies for Next Generation OTNs

Optical switches are referred to as the fiber interconnection devices which integrate a combination of algorithms, protocols and signaling mechanisms to support optical channel provisioning, routing, and restoration. The intelligent and flexible IP over WDM networks require bit-rate and protocol-independent optical switches. The switching technologies for optical networks fall under three broad categories: wavelength switching, optical packet switching, and optical burst switching.

1.2.1 Wavelength Switching

In the wavelength switching approach systems, lightpaths are set up between sources (ingress nodes) and destinations (egress nodes) via nodes equipped with OXCs (or wavelength routers). At each OXC, the output wavelength (at an output port) to

which an incoming signals is routed at any given time is determined solely based on the input wavelength (and input port) carrying the signal. Accordingly, wavelength switching is a form of circuit switching. Under distributed signaling, two-way reservation is needed, whereby a source node sends out a control packet to make a reservation, and then waits for an acknowledgment to come back before transmitting data [6, 7, 8].

The wavelength switching approach requires no optical buffer at the intermediate nodes of the core network, and enables transport transparency in the optical domain. It also features the advantages of simple implementation. However, the wavelength switched network presents low bandwidth usage efficiency, coarse-granularity traffic engineering, and limited flexibility for capacity expansion, survivability, and restoration. A lightpath takes up an entire wavelength on each link along the source-destination path, resulting in low bandwidth utilization when carrying bursty traffic streams (i.e., IP traffic). Meanwhile, when the number of wavelengths is not enough to support the full mesh connectivity, load distribution in the network may be uneven given that the traffic intensity varies over time.

1.2.2 Optical Packet Switching

In the Optical Packet Switching (OPS) approach, the switching unit at the intermediate nodes of the core network supports the store-and-forward functionality, and the network traffic is transmitted and switched in the form of optical packets. The payload is transported along with its control header without setting up a lightpath in advance [9, 10, 11, 12].

The OPS approach facilitates per-packet granularity traffic engineering, thus rendering a finer degree of service flexibility for the IP over WDM integration (e.g., bandwidth sharing, traffic balance, restoration options, and contract duration).

In the absence of photonic devices that perform signaling processing and lightpath configuration in the optical domain, however, the OPS approach requires OEO conversions at the intermediate nodes. The tight coupling in time between the payload and the header as well as the store-and-forward nature of packet switching requires each optical packet to be buffered at every intermediate node, resulting in the demand for a large amount of optical buffers (e.g., Fiber Delay Lines (FDL)) and the signal regeneration technique. Owing to the variations in the processing time of the packet header at the intermediate nodes, optical packet switching also requires stringent synchronization and the complicated control that goes with it. Another problem inherent to the OPS approach is that the sizes of the data packets are usually too small given the high bandwidth of optical channels, thus resulting in relatively high control overhead.

The OPS technology is still evolving and it remains to see if it will mature in the future to become commercially viable.

1.2.3 Optical Burst Switching

Optical Burst Switching (OBS) is a technology devised based on consideration of the above technologies, taking into account of both the advantages and the potential problems. In an OBS system, the transport and switching granularity is the individual data burst, which may contain multiple IP packets. It is deemed as a sound and promising mechanism for the IP over WDM infrastructure, with high wavelength usage efficiency and low requirements for optical buffers [13, 14, 15, 16].

OBS Principles The basic ideas underlying an OBS system are twofold:

- Transport and switching of user traffic in the burst granularity
- Decoupling of the data payload and the corresponding Burst Header Packet (BHP)

First, multiple IP packets with the same destination and attributes (e.g., QoS requirements) may be transmitted and switched as an entity, namely, a burst. Transporting and switching the traffic in a burst granularity is a solution to compensate for the time constraint of directly switching individual IP packets at optical routers induced by the mismatch between the transmission capability of WDM fibers and the processing capability of the electronic control plane. Therefore, the OBS approach reduces the control processing overhead, and enables much finer traffic engineering than that in the wavelength switching method.

Second, each data burst is preceded by a control header, which is transmitted in a different optical channel from those for data traffic. A BHP is processed at each and every intermediate node in the core network to reserve resources and set up a switching path, while the corresponding data payload is transported throughout the network transparently, without the interpretation and examination of the data format or bit rate at the intermediate nodes. Such physical decoupling between the data payload and the control headers maintains the desirable property of optical transparency for data bursts, and leads to a better synergy of both the mature electronic technologies and advanced optical technologies.

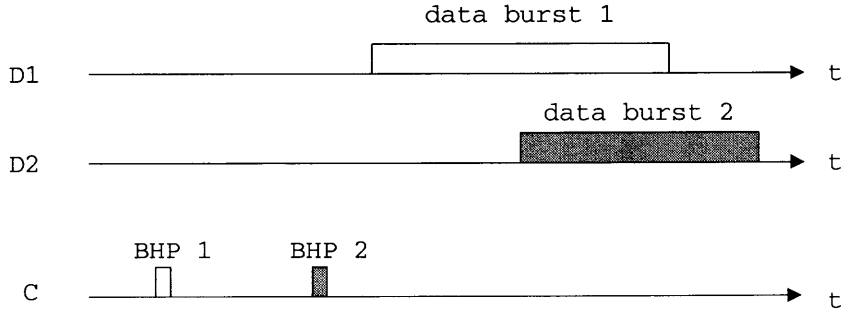

Figure 1.3 An OBS transmission example.

Figure 1.3 is a simplified example of the transmission mechanism of an OBS system, where D_i $(i = 1, 2)$ represents the i-th data channel, and C represents the control channel.

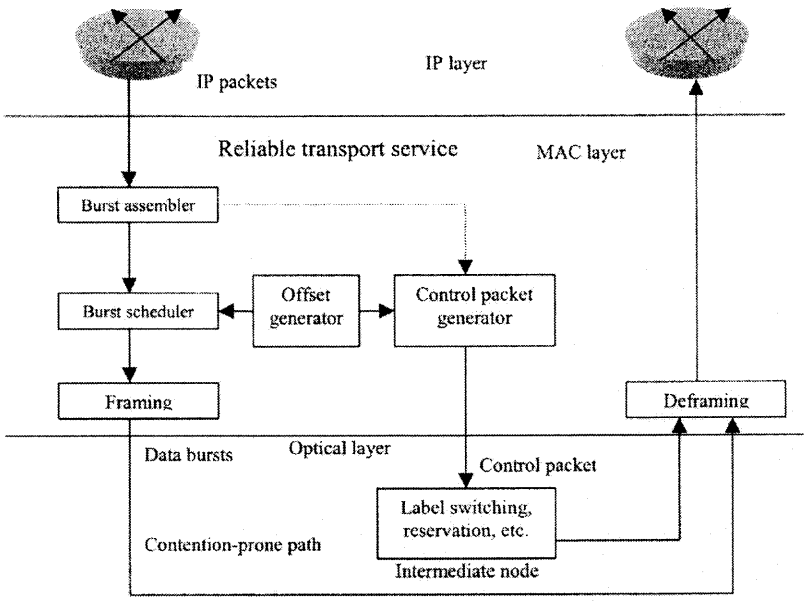

Figure 1.4 The OBS system in the IP over WDM integration.

OBS in IP Over WDM Integration Figure 1.4 presents the primary mechanisms and functionalities of an OBS-embedded IP over WDM system. A variety of solutions have been proposed in literature for some of the essential functionalities.

To transport and switch traffic in the burst granularity, one of the distinctive procedure of the OBS system is a burst assembly process, whereby multiple IP packets are assembled into an entity, namely, a data burst. Such an assembling process is usually termed as the burstification procedure [13]. Burstification is performed at the network ingress nodes, where a Burstification Control Unit (BCU) resides and coordinates the assignment and transmission of data channels and control channels.

Several burst assembly mechanisms exist in literature [13, 17, 18]. Generally speaking, there are two categories with respect to burst assembly: Limited Size burst (LS-burst) and Non-Limited Size burst (NS-burst). These methods distinguish from each other in the way new IP packets are treated when they arrive at an edge router while a datum is already being transmitted. In the LS-burst approach, the new

packets have to be queued and formulated into a new data burst, while in the NS-burst method the new packets are considered as part of the current burst. With the LS-burst method, the length of burst is known when the data transmission begins, and the length value is proportional to the burst assembly time and the input bit rate. In the NS-burst, the duration of the burst is not known when the data transmission begins, since it is subjected to change while the current data burst is being transmitted.

Another important concept in the OBS system is the offset between the data burst and the corresponding BHP. The offset-related problem concerns two facts–the generation mechanism of the offset values, and the management of the offset values. The major function of the offset value is to compensate the mismatch between the electronic processing of the BHPs and the all-optical transport of the data burst supported, and to reduce the contention between control headers. The offset value should be at least enough for the BHP to set up an all-optical lightpath before the data payload enters the intermediate node. Various offset determination algorithms have been proposed [19, 20, 21].

Signaling protocols for OBS systems are based on two alternative schemes: Tell-and-Wait (TAW) and Tell-and-Go (TAG) [22]. While the former features a two-way reservation, the TAG scheme uses the one-way signaling, i.e., at the ingress node, a control packet is sent out and after a fixed delay (offset time), without waiting for confirmation from the network, its data burst is transmitted. Some popularly discussed proposals of this category are the Just-In-Time protocol (JIT) and the Just-Enough-Time (JET) one. Comparing with the TAW alternative, the TAG mechanism holds the advantage of reduced end-to-end delay associated with the round-trip signaling transmission for lightpath reservation in long-haul networks. For more details on the TAG-based signaling protocol and the related work, interested readers are referred to [21, 22, 23, 24] and references therein.

Reservation schemes of OBS systems differentiate from each other depending on how an intermediate switch node is made aware of the beginning and the ending of a burst. Four main reservation schemes are discussed in literature [21]: I) Explicit Setup and Explicit Release; II) Explicit Setup and Estimated Release; III) Estimated Setup and Explicit Release; and IV) Estimated Setup and Estimated Release. These variants result in different complexities of hardware requirements, and different amount of time that the switching elements are reserved for an individual burst.

An improvement of reservation approach II is termed as Delayed Reservation (DR) [25], i.e., the resources at intermediate nodes are reserved for the incoming data payload from its arrival time, and are released (torn down or timed-out) at its departure time, determined from the arrival time of the BHP and the burst length. This approach enables a BHP to reserve resources for a more precise duration that corresponds to the burst length, and delivers efficient bandwidth utilization and high system throughput. A BHP in this scenario has the knowledge of its payload, including ingress/egress node identification, and the data burst length.

Design Objectives The OBS approach outperforms OPS and wavelength switching with less control overhead and better bandwidth utilization. Many OBS-specific issues have to be addressed before this technology becomes practical and efficient [26, 27, 28]. The interested issues include but not limited to:

- Burstification and offset management
- Transport mechanisms and signaling protocols
- Resource utilization and contention resolutions
- QoS provisioning and optimization.

1.3 Ethernet-supported IP over WDM Metropolitan Area Networks

More bandwidth than ever before can now be used for long-haul transport and service provisioning, as fiber installations and WDM equipment increase the capacity and lower the cost per bit for these networks. The Metropolitan Area Network (MAN) environment, however, has lagged behind in the availability of low-cost fast service provisioning using WDM.

1.3.1 WDM in Metropolitan Area Network

A Metropolitan Area Network (MAN) is defined as the part of the network that interfaces the end users and the backbone long-haul networks [29]. WDM technology, which is easy to justify on engineering and economic grounds in long distance networks, has been slower in development in metropolitan networks, although its long-term prospects are not in dispute.

Historically, the metropolitan area service requirements have been dominated by voice services, which have led to today's MANs based primarily on a ring topology with hubbed traffic patterns and SONET transport equipment. The pervasiveness of SONET networking in such network scenarios is undeniable. It offers rapid and predictable reliability and network protection, as well as management and alarming features. The increasing traffic demands in the metropolitan area have been satisfied by increasing SONET channel bit-rate, or the number of fibers connecting the nodes.

The circuit-based transport solutions of SONET, however, do not quite match the requirements of today's metropolitan area networks. On one hand, in pure SONET access arcitectures, the ADMs map the full user bandwidth into a SONET tributary wthout statistical aggregation in the access node. The SONET cross-connect function in the access node does provide some level of aggregation, but only at the discrete granularities that SONET supports [4]. On the other hand, with the aforementioned trends of traffic growth, the new generation optical MANs

demand for flexibility and efficiency. They must accommodate a wide variety of protocols and interfaces in network applications which are becoming increasinly multi-service in nature, and support more data-savvy integrated optical access solutions that both meet the current and future network demand projections and offer the required service-level flexibility and functionality with QoS. In addition, SONET establishes the network connectivity relying on ADMs and DXCs, which require O-E-O conversion, a costly technology.

Among the myriad of architecture choices, the compelling technological and economic benefits of WDM are becoming attractive. The WDM technology approaches the metro bottleneck problem with a clean slate. Its multi-channel concurrent transport capability and the accommodation for transparent traffic delivery make the WDM-enabled network an ideal platform to deliver the traffic of mixed type applications. Meanwhile, the WDM layer further enhances the service and bandwidth scalability of each of these strategies [30].

1.3.2 Ethernet Beyond Local Access Networks (LANs)

While the new generation WDM-enabled transport network is targeted for direct IP packet processing to reduce the complexity of multiplayer architectures, a smooth evolution is critical to success. Intermediate steps are necessary to support IP datagrams onto optical channels. Alternative forms of intermediate steps have been proposed for the metropolitan area environment, among which Ethernet holds great promise as the connectivity solution enabling a graceful migration from the current voice-oriented MAN prototype into a world optimized for packets [31, 32].

Ethernet has evolved over the past decade from a simple shared Medium Access Control (MAC) to a full-duplex switched network. With more than 90 percent of Internet traffic originating from Ethernet-based LANs, and with the emergence of High-speed Gigabit Ethernet and 10 Gigabit Ethernet products, Ethernet is emerging

as the protocol of choice for carrying IP traffic in the metropolitan and even long haul networks. For both enterprise customers and carriers, it is advantageous to preserve the native customer Ethernet data frame rather than to terminate it and remap its payload into another layer 2 protocol (e.g., PPP) for transport. The unprecedented bandwidth supportability of WDM technology, in tandem with the packet-oriented Ethernet prototype, facilitates a common shared infrastructure, thus making a new generation of optical MAN optimized for scalable, survivable, and IP-dominated networks at gigabit speeds possible.

However, Ethernet, while a natural fit for data traffic, lacks the flexible MAC mechanisms to manage the access across multiple users in the WDM prototype. Native Ethernet protocols need extensions or support from other technologies in terms of scalability, QoS, resiliency, OAMP, and so on. The above problems are broadly referred to as metro Ethernet scability issues [33].

1.4 Outline

The foregoing challenges make it increasingly important to design the new generation networks which retain the advantages of the data-oriented transport mechanisms while rendering elastic network resource utilization and graded levels of services. This dissertation is focused on the WDM burst-switched long haul and metropolitan area networks, covering the control architecture, transport mechanisms, enabling technologies, algorithm design, together with the performance analysis and system optimization issues that arise therefrom. Theoretical analysis and simulations results are reported to demonstrate the system performance and algorithm efficiency.

The rest of the dissertation is organized as follows. Chapter 2 investigates the timer-based burst assembly algorithms and their impact on the system performance at the network ingress nodes. Chapter 3 proposes an innovative transmission scheme, namely the Forward Resource Reservation (FRR) scheme, for latency reduction in the

OBS-based IP over WDM systems. Chapter 4 is focused on the aggressive resource reservation strategies. System optimization mechanisms in terms of bandwidth usage efficiency are discussed in Chapter 5. Chapter 6 presents on the Ethernet-supported WDM burst-switched networks, proposing a novel control architecture with Enhanced Address Resolution Protocol (E-ARP), a burst-based transmission mechanism, and a hop-based wavelength allocation algorithm. Concluding remarks and the focus of future research are given in Chapter 7.

CHAPTER 2

THE IMPACT OF THE BURST ASSEMBLY INTERVAL ON THE OBS INGRESS TRAFFIC CHARACTERISTICS

This chapter addresses the burstification interval scaling problem when the timer-based LS-burst assembly methods, including the periodic and the non-periodic alternatives, are employed. We investigate the impact of the burstification interval on the burst traffic characteristics in terms of the data burst inter-arrival time and the data burst length, respectively. An analytical model and numerical results, which evaluate the burst delay at the edge node of the OBS-enabled WDM backbone, are presented.

2.1 Motivation

The OBS paradigm holds great promise for the IP over switched WDM networks because 1) it supports the bandwidth multiplexing within the individual wavelength, thus rendering finer granularity for traffic engineering in the optical domain, and 2) it implements most of the intelligence of the network at the IP layer, with simple and scalable control and management functionalities in the core routers, thus facilitating the better synergy of the mature electronic technologies and the advanced optical technologies[13, 20, 34].

One of the enabling technologies of the OBS-enabled system is the burst assembly/disassembly procedure, namely, the burstification/de-burstification process [13, 18], The burstification process is originally proposed to alleviate the optoelectronic capacity required at the core routers for the BHP processing—a potential bottleneck owing to the bandwidth mismatch of the transport capability of the optical channels and the electronic processing speed of the core routers. Transmitting and switching the traffic in the data burst granularity proves to enhance the system performance with reduced implementation complexity at the core routers [35].

The introduction of the burstification process implies the importance to identify the characteristics of the burst traffic forwarded into the core network, and to evaluate the performance of the system with the burst traffic as the input. The burstification mechanism, including the burstification interval determination, plays an important role in unleashing the potential of the OBS-enabled WDM network, and is among the critical issues in OBS networking that has received considerable attention.

This chapter investigates the burstification interval scaling problem when the timer-based burstification mechanisms, including the periodic and the non-periodic alternatives, are adopted. The burst traffic characteristics are studied in terms of the burst inter-arrival time and the burst length. Meanwhile, an analytical model is developped to evaluate the burst delay at the edge nodes of an WDM backbone, followed by the numerical and simulation results to validate the analysis.

2.2 System Model

This section describes the system scenario upon which the subsequent discussion is developed. Then, the investigation objectives addressed in this chapter will be formulated.

2.2.1 System Architecture

Figure 2.1 highlights the functional components of the ingress node under investigation, where each output port is associated with the individual burstification unit, which consists of an electric-optic converter, a scheduler with a waiting queue, and a group of assemblers.

The edge node implements the OBS MAC layer functionalities described in [20]. IP packets from the input port i ($i = \{1, ..., N\}$) are assembled and processed in the function unit of the output port j ($j = \{1, ..., M\}$) (which corresponds to the destination address j), and are launched into the WDM backbone via the data channels.

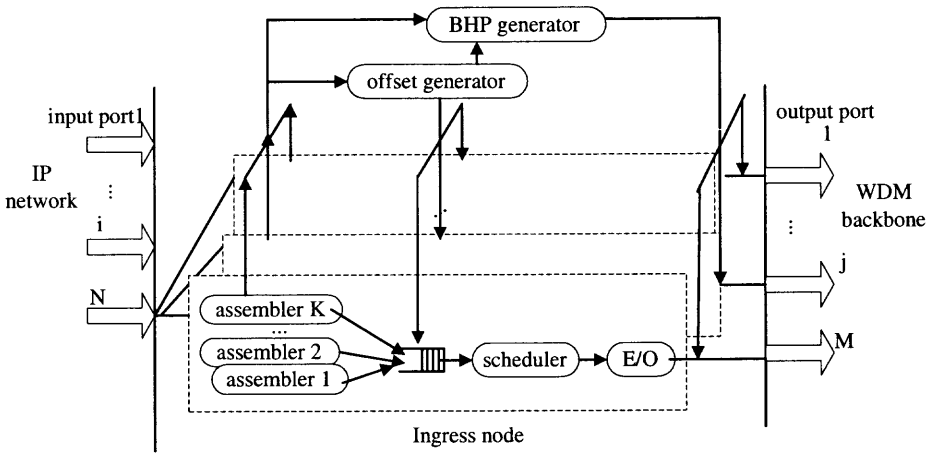

Figure 2.1 The functional components of the ingress node in the OBS system.

The event that a data burst is formed triggers the BHP generation, followed by the offset determination. The assembled data burst is either immediately assigned a data channel, or is inserted in the waiting queue for scheduling, depending on the availability of the data channels of the associated destination address (i.e., the output port).

The data burst is emitted into the core network when a data channel becomes available, with the corresponding BHP being transmitted in advance by an offset τ_o. τ_o is a system parameter to guarantee the transparent transmission of the data burst throughout the core network, and can be determined by a variety of algorithms [19, 21].

Figure 2.2 illustrates the intrinsic features of the timer-based burstification mechanisms, including the periodic and the non-periodic alternatives. Both mechanisms specify the burstification interval as a system parameter, denoted as τ_a. In the periodic mechanism, the burst is assembled back-to-back. That is, the new burstification process begins as soon as the previous data burst is assembled, and lasts until the pre-determined interval of τ_a elapses. A new data burst, however, is not actually generated if no packet arrives during the whole burstification interval. In the

non-periodic assembly mechanism, the next burstification process starts only when a new IP packet arrives. Both assembly mechanisms shape the traffic and change the traffic stream characteristics.

2.2.2 Objectives

The focus of this chapter is on the burstification interval scaling problem when either of the timer-based mechanisms is adopted. The investigation will be conducted from the following two perspectives:

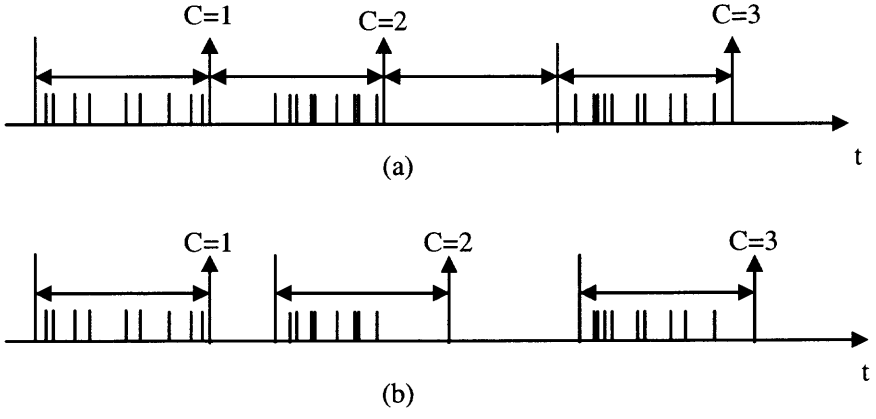

Figure 2.2 The timer-based burst assembly mechanisms (C: burst count): (a) the periodic mechanism; (b) the non-periodic mechanism.

- To identify the characteristics of the burst traffic which flows into the WDM backbone after the burstification process. The interested performance figures involve the burst inter-arrival time related ones and the burst length related ones. The mean values and the squared coefficients of variation of both categories are reported.

- To develop the analytical model for the evaluation of the delay that a data burst experiences at the edge node with the burst traffic as the input.

2.3 The Impact of the Burstification Interval

This section analyzes the impact of the burstification interval on the burst traffic characteristics, and utilizes the acquired traffic parameters to evaluate the burst delay at the ingress node.

2.3.1 τ_a on the Burst Traffic Characteristics

The following notations are defined for description simplicity:

- λ_a : The packet arrival rate of the IP stream flowing into the individual burst assembler.
- $f_T(t)$: The Probability Density Function (PDF) of the burst inter-arrival time when the non-periodic mechanism is adopted.
- E_T : The average data burst inter-arrival time.
- C_T^2 : The squared coefficient of variation of the data burst inter-arrival time.
- E_L : The average data burst length (in time).
- C_L^2 : The squared coefficient of variation of the data burst length (in time).

For computational tractability, the IP packet stream feeding into the assembler is assumed to be a Possion process. Given that the burstification interval is equal to τ_a, the PDF of the data burst inter-arrival time can be expressed as

$$f_T(t) = \begin{cases} (1 - e^{-\lambda_a \tau_a}) e^{-\lambda_a (t-\tau_a)} & t = k\tau_a, \text{ periodic,} \\ \lambda_a \cdot e^{-\lambda_a (t-\tau_a)} & t \geq \tau_a, \text{ non-periodic,} \end{cases} \tag{2.1}$$

where $k = 1, 2, ..., \infty$. Accordingly, the average burst inter-arrival time is

$$E_T = \begin{cases} \frac{\tau_a}{1-e^{-\lambda_a \tau_a}} & \text{periodic,} \\ \tau_a + \frac{1}{\lambda_a} & \text{non-periodic.} \end{cases} \tag{2.2}$$

Of more interest is the coefficient of variation of the burst inter-arrival time, defined by $C_T^2 = \frac{\sigma_T^2}{E_T^2}$ (where σ_T^2 is the variance of the burst inter-arrival time). C_T^2 is the parameter for the traffic burstiness measurement, and has been increasingly emphasized for the traffic characteristics description owing to its significant impact on the delay and loss of the network [36, 37]. Lowering C_T^2 will be greatly beneficial and important to improve the system performance.

In the predefined system scenario,

$$C_T^2 = \begin{cases} e^{-\lambda_a \tau_a} & \text{periodic,} \\ \frac{1}{(1+\lambda_a \tau_a)^2} & \text{non-periodic.} \end{cases} \tag{2.3}$$

Equation 2.3 indicates that, under the timer-based burstification process, the coefficient of variation of the data burst inter-arrival time is a function of the burstification interval (τ_a) and the input traffic density (λ_a). Given λ_a, C_T^2 decays exponentially with the burstification interval τ_a in the periodic assembly mechanism, while in the non-periodic mechanism, the decay is approximately proportional to the reciprocal of τ_a^2.

Equation 2.3 provides a guideline that, to limit the coefficient of variation of the data burst inter-arrival time to be lower than a certain level ϵ_1, the burstification interval should be constrained by

$$\tau_a \geq \begin{cases} -\frac{1}{\lambda_a} \cdot \ln \epsilon_1 & \text{periodic,} \\ \frac{1}{\lambda_a} \cdot (\sqrt{\frac{1}{\epsilon_1}} - 1) & \text{non-periodic.} \end{cases} \tag{2.4}$$

Also of interest are the parameters related to the burst size, which affects the throughput and the optoelectronic capacity requirement of the core routers [35].

Given the system scenario with the deterministic packet size, the average data burst length is

$$E_L = \begin{cases} \frac{\lambda_a \tau_a}{1-e^{-\lambda_a \tau_a}} L & \text{periodic,} \\ (\lambda_a \tau_a + 1) L & \text{non-periodic,} \end{cases} \tag{2.5}$$

where L is the average packet size. The data burst size coefficient of variation is the same as that for the burst inter-arrival time, i.e., $C_L^2 = C_T^2$ (see Equation 2.3).

2.3.2 The Delay Model at the Edge Node

This subsection discusses the burst delay in the scheduling queue, taking into account the burst traffic parameters obtained above. The analysis is based base on the following assumptions:

- The edge node is a wavelength constrained system, wherein the assembled data burst may need to be queued in the burstification unit for the next available data channel.
- The burstification unit contains a large amount of electronic buffer, thus rendering a lossless queuing system for the data burst to be scheduled.
- The maximum number of data channels which can simultaneously transport data burst to the same output port is m, and the speed-up factor of the edge node, i.e., the bandwidth ratio of the output data channel to the input one, is B_0.
- The EAC (Earliest Available Channel) algorithm is employed for data channel scheduling, and the offset value is equal for the individual data bursts.
- The IP traffic flowing into the WDM backbone presents a high input density, i.e., $\lambda_a \cdot \tau_a \gg 1$.

The edge node delay (D_e) which a data burst experiences consists of three components: the average burstification delay ($\frac{1}{2}\tau_a$), the offset delay (τ_o), and the queuing delay for the next available data channel (W). In the assumed system scenario,

$$D_e = \frac{\tau_a}{2} + W + \tau_o. \tag{2.6}$$

Given that τ_a and τ_o are system parameters pre-determined by the network management according to different protocols, while W is a performance figure of merit directly dependent on the burst traffic characteristics, this chapter merely targets on the burst delay in the scheduling queue owing to the data channel unavailability, and considers a preliminary system scenario, wherein the scheduling queue is associated with the single assembler.

After the burstification process, the burst traffic feeding to the scheduling queue resembles a general independent process, with the average burst arrival rate of $\lambda_q = 1/E_T$ (see Equation 2.2). Meanwhile, the channel holding time required by each burst (S) is proportional to the data burst length, leading to the equation of $\bar{S} = E_L/B_0$ (see Equation 2.5). This way, the assembler, the scheduler, and the associated queue can be modeled as the GI/G/m system.

Based on our model, the waiting time Cumulative Distribution Function (CDF) under heavy load traffic situation is [38]

$$F_W(w) \simeq \begin{cases} 1 - e^{-\delta \cdot \frac{1-e^{-\lambda_a \tau_a}}{\lambda_a \tau_a e^{-\lambda_a \tau_a}} \cdot w} & \text{periodic,} \\ 1 - e^{-\delta \cdot (\lambda_a \tau_a + 1) \cdot w} & \text{non-periodic,} \end{cases} \tag{2.7}$$

where δ is

$$\delta = \frac{2\lambda_a m B_0 (m B_0 - \lambda_a)}{m^2 B_0^2 + \lambda_a^2}. \tag{2.8}$$

Furthermore, based on the Allen-Cunneen formula, and taking into account of our burst traffic parameters, the average data burst waiting time in the scheduling queue can be approximated by

$$\bar{W} \simeq \begin{cases} P_m \cdot \frac{\lambda_q \tau_a e^{-\lambda_q \tau_a}}{(m B_0 - \lambda_q)(1 - e^{-\lambda_q \tau_a})} & \text{periodic,} \\ P_m \cdot \frac{1}{(m B_0 - \lambda_q)(\lambda_q \tau_a + 1)} & \text{non-periodic,} \end{cases} \tag{2.9}$$

where P_m is the probability that the data burst needs to be inserted into the scheduling queue until the next data channel becomes available. P_m can be calculated by [38]

$$P_m \simeq \begin{cases} \frac{(\frac{\lambda_q}{mB_0})^2+\frac{\lambda_q}{mB_0}}{2} & \frac{\lambda_q}{mB_0} > 0.7, \\ (\frac{\lambda_q}{mB_0})^{\frac{m+1}{2}} & \frac{\lambda_q}{mB_0} \leq 0.7\ . \end{cases} \tag{2.10}$$

2.4 Numerical and Simulation Results

In this section, we present the numerical and simulation results to justify our analysis. We focus on the impact of τ_a, and the sensitivity of such impact to λ_a. The data channel utilization of the individual output port, defined by $\rho = \lambda_a \cdot \bar{S}/m$, is set to be 0.2 and 0.5, respectively. The burst assembly duration (τ_a) and the queuing delay (W) are normalized with respect to the time to transmit one IP packet of 1500 bytes. The IP packets arrive at the input port according to a Poisson process.

Figure 2.3 presents the effect of the average data burst inter-arrival time under the periodic and the non-periodic burstification mechanisms. Both simulation and analytical results are presented. The non-periodic mechanism delivers longer burst inter-arrival time than the periodic mechanism does, leading to the following conclusion: to support the same network traffic, the non-periodic mechanism allows the core router process the BHP with a lower processing capability requirement, and better facilitates the benefits of the OBS-enabled WDM networks in terms of alleviated BHP processing requirement at the core routers.

Figure 2.4 illustrates the coefficient of variation of the burst inter-arrival time versus τ_a, indicating that in both burstification mechanisms, C_T^2 is reduced as τ_a increases, and that the effect of τ_a on C_T^2 is more considerable for the periodic mechanism than for the non-periodic mechanism. This is especially true when the input traffic load is relatively large (represented by the larger ρ), in which case the periodic mechanism delivers the slightly lower C_T^2 than the non-periodic counterpart does as τ_a gets large enough.

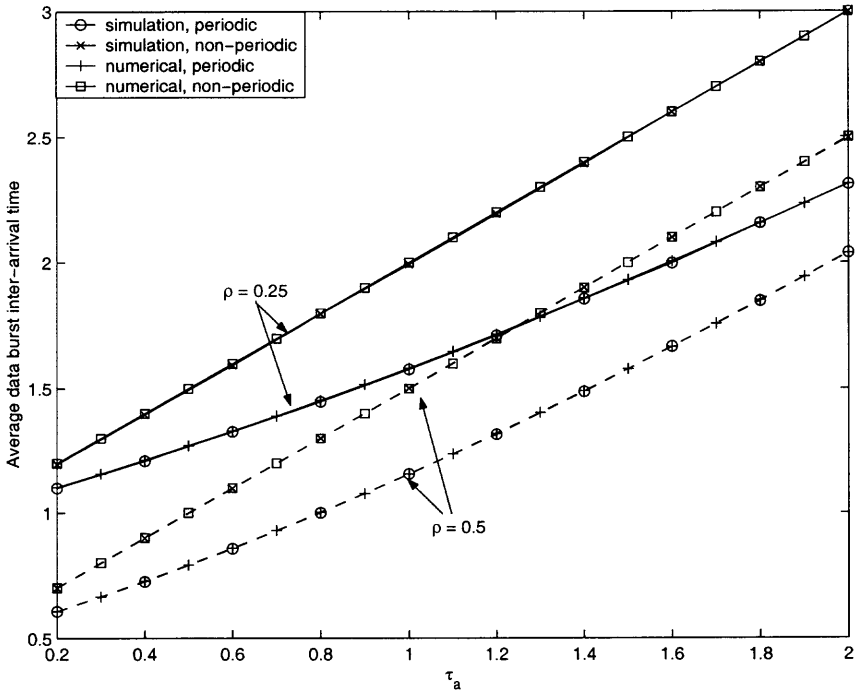

Figure 2.3 The average data burst inter-arrival time versus the burstification interval.

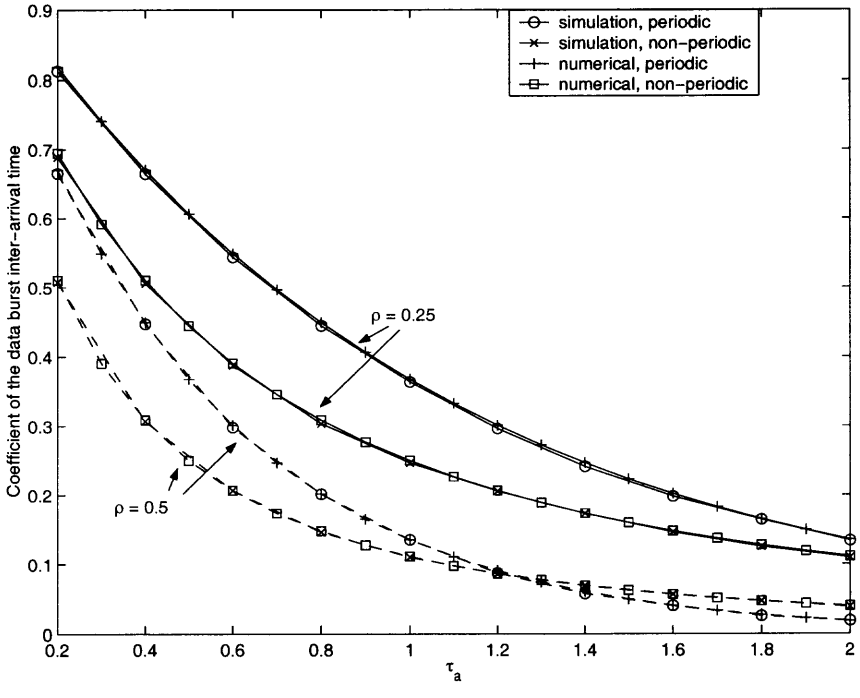

Figure 2.4 The squared coefficient of variation of the data burst inter-arrival time versus the burstification interval.

The simulation results of the waiting delay because of the unavailability of the output data channel are shown in Figure 2.5. It can be seen that the periodic mechanism results in larger waiting delay than the non-periodic mechanism does for both ρ values. Meanwhile, the average waiting delay goes downward as the burstification interval increases, and the effect of τ_a on the waiting delay is more significant when the traffic load increases.

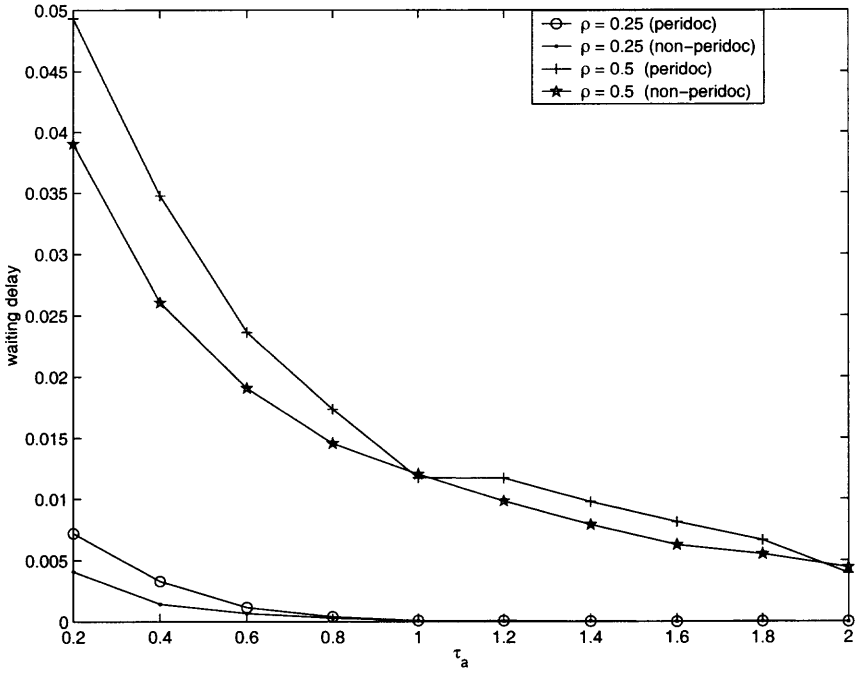

Figure 2.5 The simulation results of the average waiting time versus the burstification interval.

2.5 Summary

This chapter has studied the burstification interval scaling problem. The characteristics of the traffic after the burstification process is identified, and the burst delay due to the data channel unavailability is evaluated. Simulations results and the theoretical analysis indicate the following conclusions: 1) Given that τ_a is large enough (e.g., a few times $1/\lambda_a$), the non-periodic mechanism yields the larger average data burst size and the data burst inter-arrival time as compared with the periodic alternative, and

thus lowering signaling overhead in the backbone; 2) In both mechanisms, the coefficients of variation of the data burst inter-arrival time decays as τ_a increases within a reasonable range; 3) The lower average traffic delay is achievable by the non-periodic mechanism. In other words, the non-periodic mechanism improves performance of the OBS-enabled network with reduced buffer requirements at the ingress node.

CHAPTER 3

FORWARD RESOURCE RESERVATION FOR LATENCY REDUCTION IN OBS NETWORKS

In this chapter, a novel Forward Resource Reservation (FRR) transmission scheme is proposed for the OBS systems. The aim is to reduce end-to-end transport latency for delay-sensitive applications, and to facilitate QoS provisioning for different traffic classes. Following the motivations and system environments description, our discussion will cover the FRR scheme principle, its implementation and feasibility, the FRR extension for QoS provisioning, and the performance evaluation.

3.1 Motivations

Transporting and switching the traffic in the data burst granularity is one of the important features that alleviate the OBS system from the heavy burden for lightpath configuration. This advantage benefits from the particular burst assembly procedure (i.e., the burstification process) at the ingress nodes of an OBS system. A side effect imposed by such a burst-buildup process, however, is an artificial delay. The typical end-to-end delay of a data burst thus mainly consists of three components:

- Burst assembly delay at edge routers
- Path setup delay caused by control headers
- Propagation delay in the core network

To date, it has been widely recognized that, the bandwidth is no longer the transmission bottleneck in many core networks, but it is the latency that dominates the transmission time and is becoming of paramount importance [23, 39, 40]. Latency

reduction is an important consideration when building up any system for the next generation optical network [11, 41, 40].

There have been numerous proposals in the literature focusing on the latency reduction issue in OBS systems. For example, a typical OBS system features one-way reservation that lowers the round-trip delay for signaling transmission [22]. The Just-In-Time (JIT) protocol is proposed to reduce burst delay due to round-trip lightpath-setup [21, 23]. Xiong et al. [13] discussed the optimal switching architectures of the core routers to process control headers. All these strategies are focused on reducing the latency in the core network.

It is observed that the bandwidth at the core network (OC192 and beyond) is much higher than that in the edge network (OC3-OC48). The time for assembling a burst, which usually consists of hundreds of IP packets and is at the time scale of hundreds of microseconds, is comparable with the switching path setup time, which is also presumed to be in the range of hundred of microseconds [41, 41]. The burst delay at network edges is substantial and has a significant impact on the end-to-end transmission latency. This influence is especially detrimental to the real-time traffic, which has stringent delay constraints. Since the propagation time of a data burst, which is intrinsic, cannot be reduced, reducing burst delay at network ingresses will be greatly beneficial to latency reduction and QoS provisioning.

Therefore, an innovative transmission scheme, namely, the Forward Resource Reservation (FRR) scheme, is proposed for latency reduction at the edges of an OBS system. The FRR scheme is further extended for QoS differentiation on burst delay for different classes of traffic. Theoretical analysis and simulation results show that the proposed scheme substantially reduces the burst delay at the network edge without introducing deleterious system costs.

3.2 System Environment and Design Objective

This section describes the system environment in which the FRR scheme is applied. The objective of the FRR-embedded OBS system design is presented.

3.2.1 System Model

Figure 3.1 highlights the architecture of an OBS network under investigation. The timer-based LS-burst assembly mechanism is adopted for burstification. When a predefined threshold is reached (e.g., a timer expires), a new burst is generated and is ready to be sent into the core network.

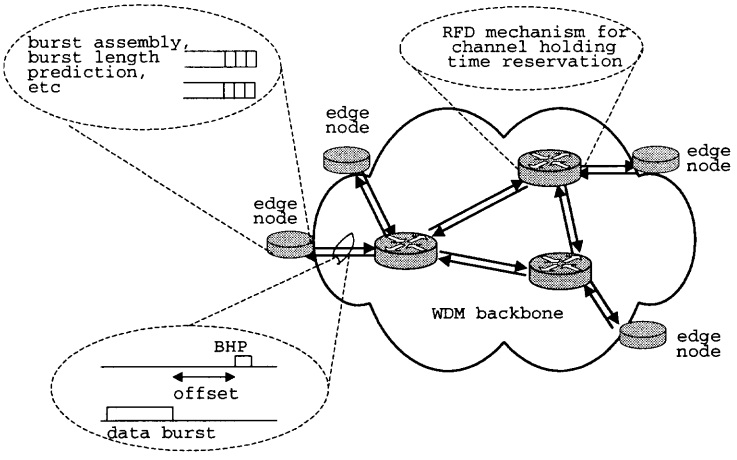

Figure 3.1 The system environment.

The lightpath is setup and reserved for a burst according to the Explicit Reservation Estimated Release approach [21], whereby a BHP requires the channel holding duration that corresponds to its burst length. Such an approach has been proved to deliver higher bandwidth utilization and lower burst loss probability at the core network [21].

In this scenario, a BHP has the knowledge of its payload, including the burst length. The BHP also specifies the offset time that its data payload lags behind. The offset value can adopt a pre-existing protocol that is most convenient in the system,

e.g., it may be the BHP processing time (end-to-end) in the core network [19], or it may be determined according to the JIT protocol [21, 23].

The application streams in the WDM network are specified by two parameters. One is the traffic load which characterizes the incoming traffic intensity. The other is the delay allowance which indicates the time constraint. According to this parameter, we partition the traffic into M QoS classes, with the class-j traffic being more delay-sensitive than the class-k one when $0 \leq j < k \leq M-1$. The QoS requirement considered in this chapter is the delay constraint.

3.2.2 Design Objectives

A brief summary of the FRR scheme design objective is as follows.

1. A BHP carries the information necessary for lightpath setup, including a reservation duration, which corresponds to the length of its data payload;

2. While preserving the all-optical lightpath advantage for its payload, a BHP should enable the data burst to be transmitted as early as possible, thus minimizing the latency at edge nodes;

3. The system can behave differently for different classes of traffic to achieve service differentiation in terms of the burst delay.

The proposed FRR scheme meets the first two requirements by a linear predictive filter (LPF)-based method, and is extended to facilitate the QoS capability required by the third one.

3.3 The FRR Scheme

In a typical DR mechanism, the transmission of a BHP depends on the burst assembly process. To acquire the necessary information of its data payload, including the data burst length, a BHP waits for the completion of the burst assembly before it is

transmitted for signaling and resource reservation. To allow enough time for switching nodes to process the BHP and to set up the switching matrix, the data payload should be further delayed at the ingress node for an offset time before being launched into the core network. The data burst delay at an edge node has to account for these two factors, both considerable sources of delay.

The intuitive idea on this observation is that, rather than performing the above two processes in sequence, the burst assembly procedure and the transmission of a BHP should be processed in parallel, and minimizing their impact on the total end-to-end burst delay.

3.3.1 The FRR Scheme Principle

The following notations are defined to simplify the description of the FRR scheme (i=0, 1, ..., $M-1$, where M is the number of traffic classes in the system):

- T_a^i : The time when a new burst of class-i traffic begins to assemble,
- T_h^i : The time when a class-i BHP is sent into the core network,
- T_d^i : The time when a class-i data burst is sent into the core network,
- τ_a^i : The duration to assemble a burst of class-i traffic,
- τ_o^i : The offset between a class-i BHP and its data payload.

In the rest of this chapter, for notational simplicity, the referencing of the class may be omitted when the behavior and performance of only the traffic class to which the FRR scheme applies is concerned.

An FRR scheme involves a three-step procedure as follows:

Phase 1: Prediction. As soon as the previous burstification is done and a new burst assembly begins at T_a^i, the BCU predicts the length of the next incoming

data burst. This estimation is based on a linear prediction method, as will be discussed in the next section.

Phase 2: Pre-transmission. Instead of waiting for the burst assembly to complete, a control header is constructed instantly upon the completion of the prediction. The BCU enters into the BHP the information necessary for path setup, including a resource reservation length which is determined with an aggressive reservation algorithm. The BHP is then launched into the core network at time $(T_h^i = max(T_a^i, T_a^i + \tau_a^i - \tau_o^i))$.

Phase 3: Examination. When the burst assembly is fully carried out, the actual burst length is compared with the reservation length in the pre-transmitted BHP to ensure the pre-reserved duration is enough for the actual burst length. There are two cases of interest to consider (Figure 3.2):

- If the actual burst length is less than or equal to the pre-reserved length, i.e., the BHP has reserved enough bandwidth for the data payload, the BHP pre-transmission is deemed a success. In this case, the data burst is sent into the core network at $T_d^i = T_h^i + \tau_o^i$.
- If the actual burst duration exceeds the reservation length, the BHP pre-transmission is deemed a failure. The BHP has to be re-transmitted for this burst at a later time of $T_a^i + \tau_o^i$, with the actual burst size, and the data payload lags behind by the offset τ_o^i.

Figure 3.3 presents the basic functional components of the FRR scheme. A variety of solutions can be employed to implement different functionalities. Note that the proposed FRR scheme does not introduce extra burst delay. A failed forward reservation causes the same latency with a transmission not using the FRR scheme.

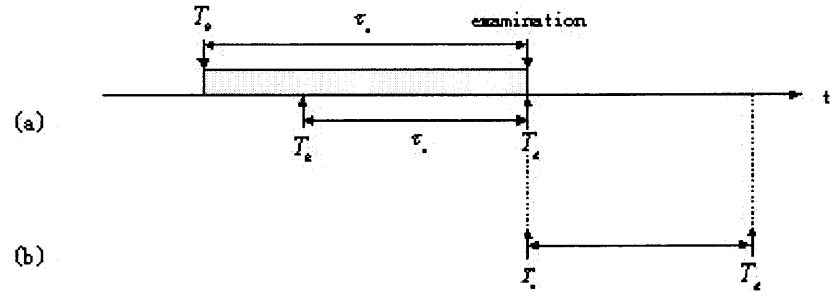

Figure 3.2 The FRR scheme principle. (a) The BHP pre-transmission succeeds; (b) The BHP pre-transmission fails.

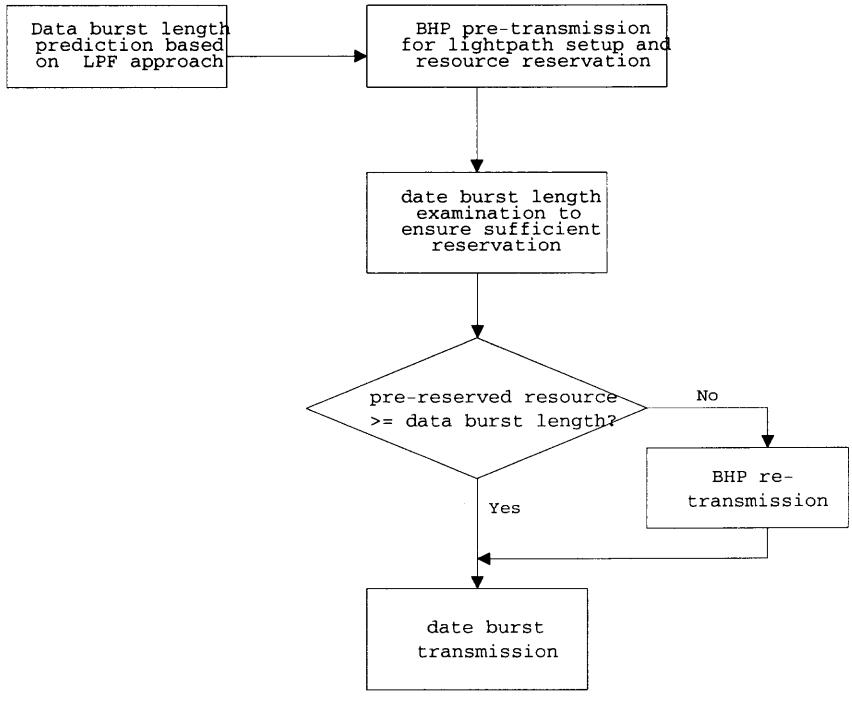

Figure 3.3 The basic functional components of the FRR system.

3.4 Basic Features of the FRR Scheme

The basic idea of the FRR scheme is to parallel the executions of multiple delay sources, thus reducing the end-to-end data burst delay. The proposed FRR scheme possesses the following salient features: the burst length prediction based on adaptive filters, the aggressive strategy for resource reservation, and the ease for QoS provisioning and QoS differentiation.

3.4.1 Traffic Prediction

The FRR scheme requires *a priori* knowledge of the burst length before it is fully assembled. This is made possible with an N-order LPF [42, 43]. Let $L_d(k)$ be the length (in the time scale) of the k-th burst, then the length of the next incoming burst is predicted according to the lengths of the previous N bursts by

$$\tilde{L_d}(k+1) = \sum_{i=1}^{N} h(i) \cdot L_d(k-i+1), \tag{3.1}$$

where $w(i)$, $i \in \{1,,N\}$, are the coefficients of the adaptive filter.

Two approaches are examined to obtain the predictive filter coefficients. One is based on the Yule-Walker method, whereby the predictive filter coefficients can be expressed as $\mathbf{Rh} = \mathbf{r}$, where $\mathbf{R}$ and $\mathbf{r}$ are the autocorrelation matrix and the autocorrelation vector of the data burst lengths, respectively, and $\mathbf{h}$ is the coefficient vector [44]. An alternative is the N-order normalized Least Mean Square (LMS)-based recursive LPF. The predictive filter coefficients are updated using an efficient algorithm [43], where the coefficients for the $(k+1)$-th prediction are defined as

$$\mathbf{h}(k+1) = \mathbf{h}(k) + \frac{\mu \cdot e(k) \cdot (\mathbf{L_d})^k}{\| \mathbf{L_d}^k \|^2}, \tag{3.2}$$

with μ being an adjustable parameter of the LPF, $e(k)$ the residual between the actual and the predicted length of the k-th data burst, and $(\mathbf{L_d})^k$ the vector of $L_d(j)(j \in (k-1) \cdot N + 1, ..., k \cdot N)$.

It is verified by simulations (the results of which are partly reported in the next section) that, the LMS-based method is more appropriate, within the context of burst length prediction, to forecast the length of the next data burst when the input IP traffic presents self-semilarity. The LMS-based method achieves satisfactory prediction performance without knowing the autocorrelation of the input traffic stream in advance, and thus can be used as an on-line algorithm for bandwidth forecast. Meanwhile, the LMS-based approach outperforms the other alternative in terms of computational simplicity. Its time complexity for the coefficient calculation is $O(N)$, which is much less than that of Yule-Walker equations ($O(N^2)$).

3.4.2 Aggressive Resource Reservation

A BHP makes an advance resource reservation according to the predicted value. The forward reservation length, denoted as $L_r(k+1)$, if optimal, should be equal to the actual burst length. Due to the imperfection of a predictor, however, an estimated length may turn out to be smaller or larger than the actual burst duration. Suppose the reservation length is set to be equal to the predicted length, a smaller prediction of burst length ($e(k+1) = L_d(k+1) - \tilde{L_d}(k+1) \geq 0$) will result in an insufficient reservation of path holding time for the data burst. This requires the BHP to be re-transmitted after the burst assembly finishes, thus degrading the FRR latency reduction performance.

This problem is compensated by an aggressive reservation method. Instead of making $L_r(k+1) = \tilde{L_d}(k+1)$, we define the reservation length as $L_r(k+1) = \tilde{L_d}(k+1)+\delta$, where δ is a small margin of correction. The value of δ has a significant impact on both the BHP pre-transmission success probability (and therefore the latency

reduction capability of the FRR scheme), and the system costs (e.g., the resource utilization and the signaling overhead). It should be carefully determined according to the tradeoff between these two performance metrics.

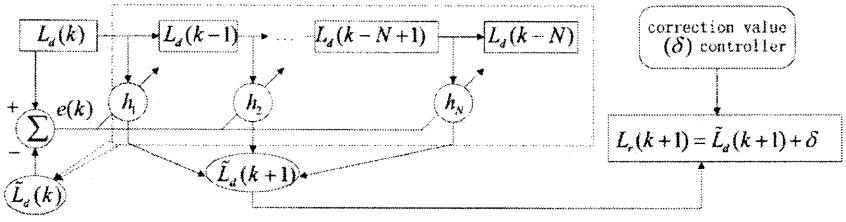

Figure 3.4 The burst length prediction and the resource reservation determination.

Figure 3.4 illustrates the principle of the burst length prediction and the aggressive reservation strategy.

3.4.3 FRR-based QoS Provisioning

As discussed above, real-time traffic has a higher traffic class and a more stringent constraint for burst delay. To achieve a flexible QoS differentiation for different classes of applications, the FRR scheme is extended for QoS provisioning.

The FRR-enabled OBS system facilitates the QoS provisioning by assigning each individual class-i traffic two system parameters: the interval (τ_p^i) to control when to launch the BHP into the core network prior to the burst assembly completion, and the real value α^i (defined in Figure 3.5) to achieve controllable BHP pre-transmission success probability. It is precisely the flexibility of (τ_p^i) and α^i that enables us to implement the scalable delay reduction and QoS differentiation degree [19, 25] between classes.

Figure 3.5 presents the discipline of our QoS strategy by illustrating the behaviors of BHPs belonging to two traffic classes (class-0: delay-sensitive; class-1: delay-tolerant) when $M = 2$. For simplicity, both classes are defined to have the same burst assembly time and offset time, denoted by τ_a and τ_o, respectively, and

that $\tau_p^0 < \tau_a$ and $\tau_p^1 = 0$. The advanced transmission of the class-0 data burst is achieved ($T_d^0 < T_d^1$). Accordingly, the average delay that the time-critical traffic experiences at the ingress node can be decreased, taking into account of $\alpha^0 \geq \alpha^1$ (as will be analyzed in the next section).

For a burst of class-1 traffic (i.e., non-real-time traffic), a simple resource reservation is executed, where a BHP is generated and is sent into the core network when the burst is fully assembled. The BHP carries the actual burst length (Figure 3.5).

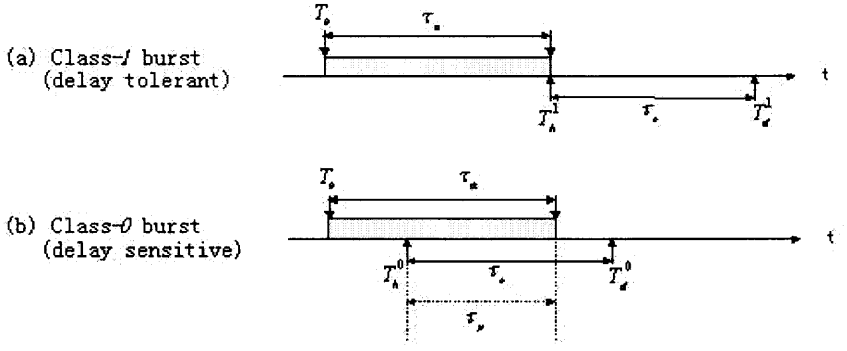

Figure 3.5 The FRR-based QoS provisioning. (a) For a delay-tolerant data burst, the BHP is not transmitted until the burstification is completed; (b) For a delay-sensitive data burst, the FRR scheme is adopted.

For a burst of class-0 traffic, however, an FRR-based process is triggered. A BHP is launched into the core network prior to the burst assembly completion by time τ_p (Figure 3.5). The delay of the time-critical traffic at the ingress node is thus decreased ($T_h^0 \leq T_h^1$). The advanced period τ_p is a system parameter and can be determined from a user or a system perspective. The user could specify the QoS constraint. Alternatively, the network operator can adapt the τ_p as a matter of policy, varying with the differentiation degree requirement between classes. In Figure 3.5, $\tau_a \geq \tau_p$.

3.5 Performance Analysis and Simulation Results

The system performance is evaluated via theoretical analysis and simulation results. Interesting performance metrics include: η - the latency reduction improvement of an FRR system, P_s - the BHP pre-transmission success probability, and γ - the bandwidth overhead. Hereafter, when we conduct the performance evaluation for the individual traffic class to which the FRR scheme applies, the referencing of traffic class will be omitted for notational simplicity. Of more interested is the performance of the FRR-based reservation scheme as compared with that of a simple reservation scheme (called NFRR for None Forward Resource Reservation). We also investigate the prediction performance of an LMS-based LPF under a variety of traffic parameters, and justify the predictability of the self-similar traffic. The order of LPF is 4, if not otherwise specified. To focus on the effect of the FRR scheme on latency reduction, we do not consider the queuing delay due to the edge node scheduling. The following notations will be used in the analysis:

- D_n : Average burst delay in an NFRR system
- D_a : Average burst delay in an FRR system
- D_f : Burst delay when the BHP pre-transmission fails
- D_s : Burst delay when the BHP pre-transmission succeeds
- P_s : The BHP pre-transmission success probability

3.5.1 Latency Reduction Improvement

This subsection is focused on the burst delay at the network edge with the NFRR or the FRR mode of transmission, and the latency improvement by the FRR scheme. The delay of a data burst is defined as the average delay of all the packets composed of this burst, due to the burst assembly and the basic offset time. Therefore, the burst delay due to burst assembly is $\frac{1}{2} \cdot \tau_a$.

1. Burst delay in an NFRR system

   In an NFRR system, the burst delay at an ingress node due to the burst assembly and the basic offset time is

$$D_n = \frac{1}{2} \cdot \tau_a + \tau_o. \tag{3.3}$$

2. Burst delay in an FRR system

   In a system with an FRR scheme, the burst delay at an ingress node differs according to the success or failure of the pre-transmission of a BHP. If fails, the delay is the same as D_n ($D_f = D_n$). Otherwise, the delay $D_s = \frac{1}{2} \cdot \tau_a$ when $\tau_a \geq \tau_o$, or $D_s = \frac{1}{2} \cdot \tau_a + \tau_o$ when $\tau_a \leq \tau_o$. Suppose the forward resource reservation succeeds with a probability of P_s, the average burst delay of a class-0 burst is therefore

$$D_a = P_s \cdot D_s + (1 - P_s) \cdot D_f, \tag{3.4}$$

   i.e.,

$$D_a = \begin{cases} \frac{1}{2} \cdot \tau_a + \tau_o - \tau_a \cdot P_s & \tau_a < \tau_o, \\ \frac{1}{2} \cdot \tau_a + \tau_o - \tau_o \cdot P_s & \tau_a \geq \tau_o. \end{cases} \tag{3.5}$$

Now that the burst delay depends on both τ_a and τ_o, we assume $\tau_o = \mu \cdot \tau_a$, where μ is a real value that represents the ratio of τ_o over τ_a . Hence, the latency improvement (η) of the FRR scheme over the NFRR scheme is given by

$$\eta = 1 - \frac{D_a}{D_n} = \begin{cases} \frac{2 \cdot \mu \cdot P_s}{1 + 2 \cdot \mu} & \mu < 1, \\ \frac{2 \cdot P_s}{1 + 2 \cdot \mu} & \mu \geq 1. \end{cases} \tag{3.6}$$

The systems performance improvement η depends on two parameters: the ratio of τ_o over τ_a (μ), and the probability that a forward reservation succeeds (P_s). Figure 3.6 presents the latency reduction percentage versus P_s, when μ varies. It

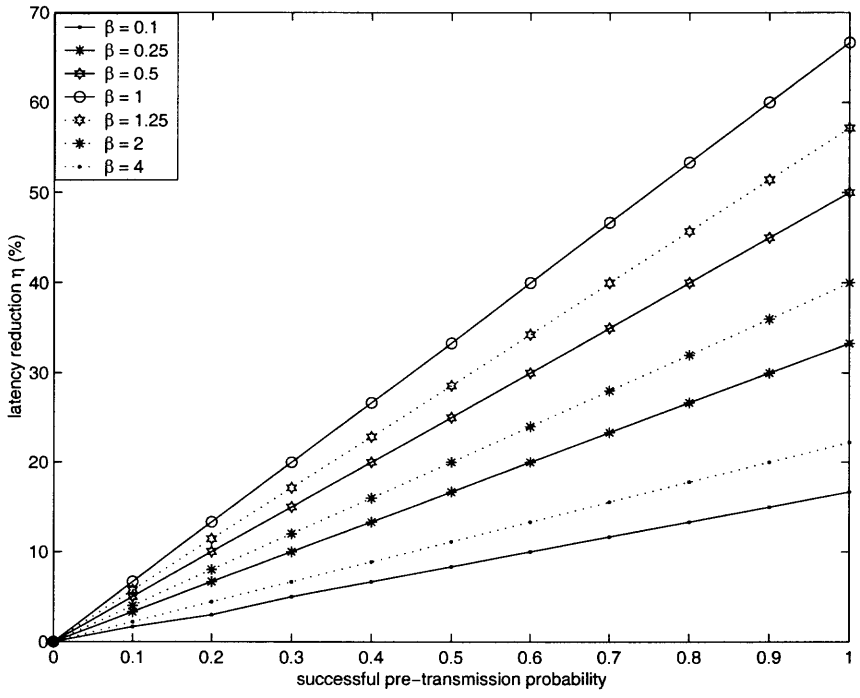

Figure 3.6 The latency reduction improvement.

shows that η increases as τ_o approaches τ_a, and reaches its maximum gain when the ratio is 1. Specifically, if the burst length can be predicted precisely such that the pre-transmission of the BHP succeeds with a high probability ($P_s \rightarrow 100\%$), our FRR scheme can reduce the latency for the high-class traffic by 66% when $\tau_a = \tau_o$. This observation can be further exploited when studying the design issues related to the burst assembly time and the offset values.

3.5.2 BHP Pre-transmission Success Probability

Equation 3.6 indicates that the probability that a BHP pre-transmission succeeds (P_s) has an important impact on the latency improvement η. Since P_s depends on, among others, the difference between the pre-reserved duration and the actual burst length ($L_r(k) - L_d(k) = \delta - e(k)$), it is important to study the effect of the correction margin (δ) on P_s.

Conceptually, P_s can be derived from

$$P_s = P(e(k) \leq \delta) = \int_{-\infty}^{\delta} f(e(k))de(k), \tag{3.7}$$

where $f(e(k))$ is the distribution of the prediction errors $(e(k))$. Assuming that $f(e(k))$ could be approximated by a zero-mean Gaussian function with variance equal to σ^2, P_s can be derived as

$$P_s = P(e(k)) = 1 - Q(\frac{\delta}{\sigma}), \tag{3.8}$$

where $Q(.)$ is the Q-function [44] defined as

$$Q(t) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{t}^{+\infty} e^{-\frac{t^2}{2}} dt. \tag{3.9}$$

The theoretical value of P_s, together with the simulation results of P_s under the real IP traffic and the video traffic, is plotted in Figure 3.7. It is shown that the BHP pre-transmission succeeds at a probability of more than 95%, if $\delta \geq 2 \cdot \sigma$, at which point the latency improvement is more than 60% (when $\tau_a = \tau_o$, as shown in Figure 3.6).

Figure 3.8 presents the simulation results by tracing the probability density function (PDF) of the number of bursts whose actual lengths differ with the pre-reserved length by a small region of reservation correction. The simulation platform is OPNET, with packets arriving according to a Poisson process.

For comparison, we also draw the PDF curve of a standard Gaussian distribution function (the dotted line in Figure 3.8). It shows that the PDF of the simulation results matches the theoretical curve very well. This also implies that controllable successful BHP pre-transmission probabilities are achievable as a function of the extra bandwidth reservation.

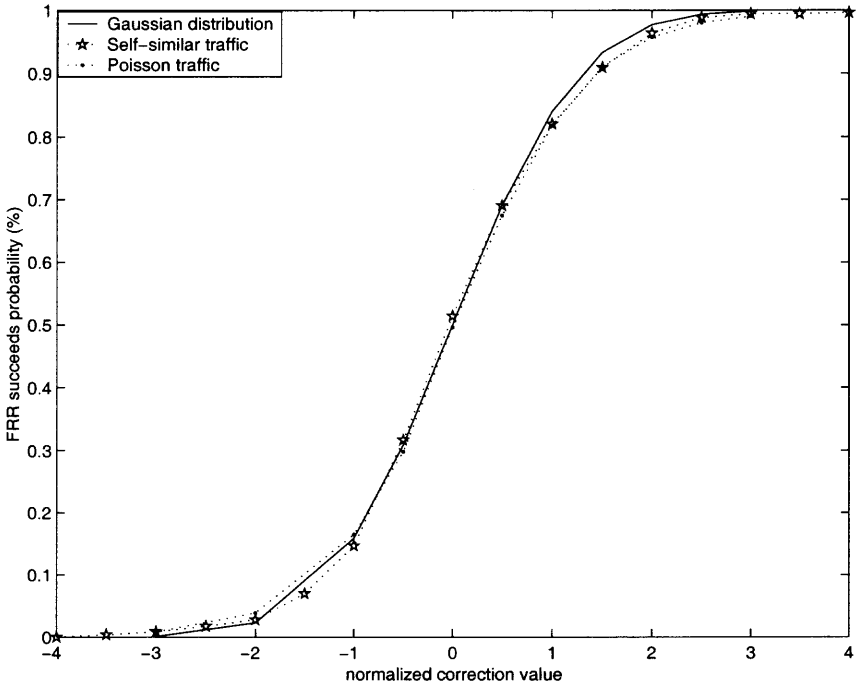

Figure 3.7 The BHP pre-transmission success probability versus δ.

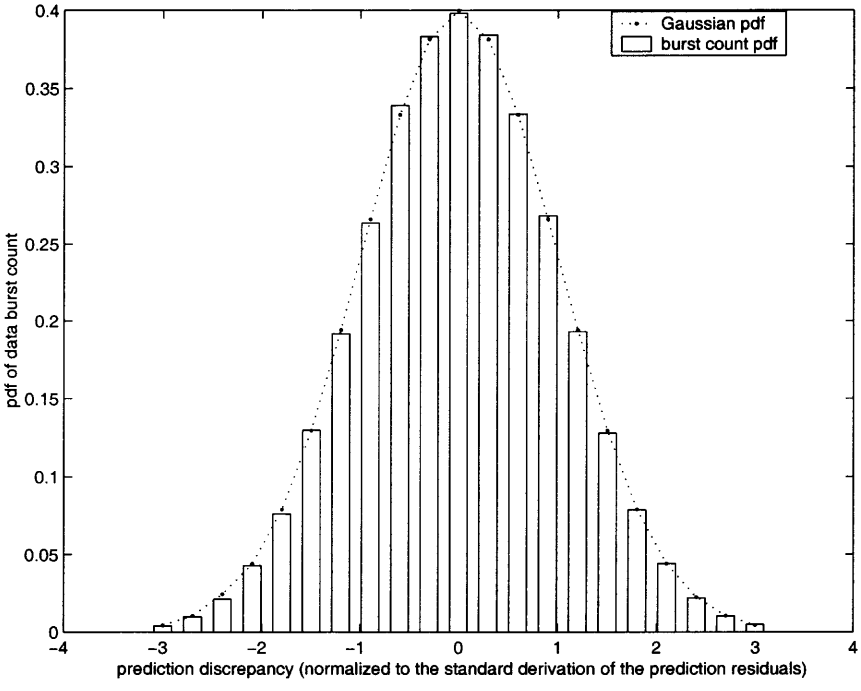

Figure 3.8 The PDF of burst numbers versus δ.

3.5.3 Bandwidth Overhead

The FRR strategy increases the BHP pre-transmission success probability and improves the latency reduction performance for the delay-sensitive traffic by means of an aggressive bandwidth reservation. For the class-i traffic to which the FRR scheme applies, let γ^i represent the ratio of the average extra reservation length to the average actual burst length. γ^i can be referred to as the bandwidth overhead of this traffic class. Now we consider the bandwidth overhead as a long-term system performance, and omit the index of the burst sequence number. This way, an advanced reservation length is simply denoted as L_r^i, which is equal to $\tilde{L}_d^i + \delta^i$, where $\tilde{L}_d^i$ is the estimated burst length and δ^i the correction margin. The actual burst length is referred to as L_d^i. Let ε^i and ζ^i represent the difference between L_d^i and $\tilde{L}_d^i$, and that between L_r^i and L_d^i , respectively. Then, the following relationships hold: $\varepsilon^i = L_d^i - \tilde{L}_d^i$, $\zeta^i = L_r^i - L_d^i$, and ${L_r}^i = \tilde{L}_d^i + \delta^i$.

The bandwidth overhead of the FRR scheme factors in both the successful and the unsuccessful pre-transmission probabilities of a BHP. A BHP pre-transmission succeeds when $\zeta^i > 0$, which implies $\varepsilon^i < \delta^i$. The average ε^i in this case, denoted as $\bar{\varepsilon}^i$, is given by

$$\bar{\varepsilon}^i = \int_{-\infty}^{\delta^i} \varepsilon^i \cdot f(\varepsilon^i) d\varepsilon^i, \tag{3.10}$$

where $f(\varepsilon^i)$ is the distribution function of ε^i.

The bandwidth overhead caused by a successful forward resource reservation is thus:

$$\gamma^i = \frac{\delta^i - \bar{\varepsilon}^i}{\tilde{L}_d^i + \bar{\varepsilon}^i} \cdot P_s^i. \tag{3.11}$$

Meanwhile, the bandwidth overhead caused by an unsuccessful pre-transmission of the BHP is 100%, i.e., $\gamma_f^i = P_f^i$.

Provided that the distribution of the residuals of our LPF is a zero-mean Gaussian function with variance σ^{i2}, and that we have $\delta^i = \alpha^i \cdot \sigma^i$, the bandwidth overhead of class-i traffic can thus be expressed as

$$\gamma^i = \gamma_s^i + \gamma_f^i = \frac{\delta^i + \frac{\sigma^i}{\sqrt{2\cdot\pi}} \cdot e^{-\frac{\delta^{i2}}{2\cdot\sigma^{i2}}}}{\tilde{L}_d - \frac{\sigma^i}{\sqrt{2\cdot\pi}} \cdot e^{-\frac{\delta^{i2}}{2\cdot\sigma^{i2}}}} \cdot (1 - Q(\frac{\delta^i}{\sigma^i})) + Q(\frac{\delta^i}{\sigma^i}). \tag{3.12}$$

Of more interest is the system bandwidth overhead, i.e., the bandwidth overhead of the whole system where multiple traffic classes exist, defined as

$$\bar{\gamma} = \sum_{i=theTraffic ClassIndex} \gamma^i \cdot \rho^i, \tag{3.13}$$

where ρ^i is the traffic load of class-i. For example, in a two-class QoS scenario where the FRR and the NFRR schemes are applied to the class-0 traffic and the class-1 traffic, respectively; suppose the traffic load distribution of the real-time traffic and the non-real-time traffic is $3:7$, then the system bandwidth overhead is $\bar{\gamma} = 0.3 \cdot \gamma^0$. Figure 3.9 illustrates $\bar{\gamma}$ as a function of α^0. Both theoretical values (Equation 3.12) and simulation results are presented.

It is interesting to see that by properly choosing a small margin of correction in addition to the predicted burst length, the aggressive resource reservation-enhanced FRR system actually reduces the bandwidth overhead as compared to a system with a zero-correction reservation algorithm. Provided $\alpha^0 \in [0, 3.0]$, the upper bound of the bandwidth overhead corresponds to the one with $\alpha^0 = 0$. The reason is that the correction value, which is much smaller than the length of a data burst, dramatically increases the BHP forward signaling success probability, and reduces the wasted resource reservation due to insufficient burst length prediction, which will otherwise contribute a greater bandwidth overhead. Correction values larger than some threshold (e.g., $\delta^0 \geq 2 \cdot \delta^0$), however, result in a slightly higher pre-transmission

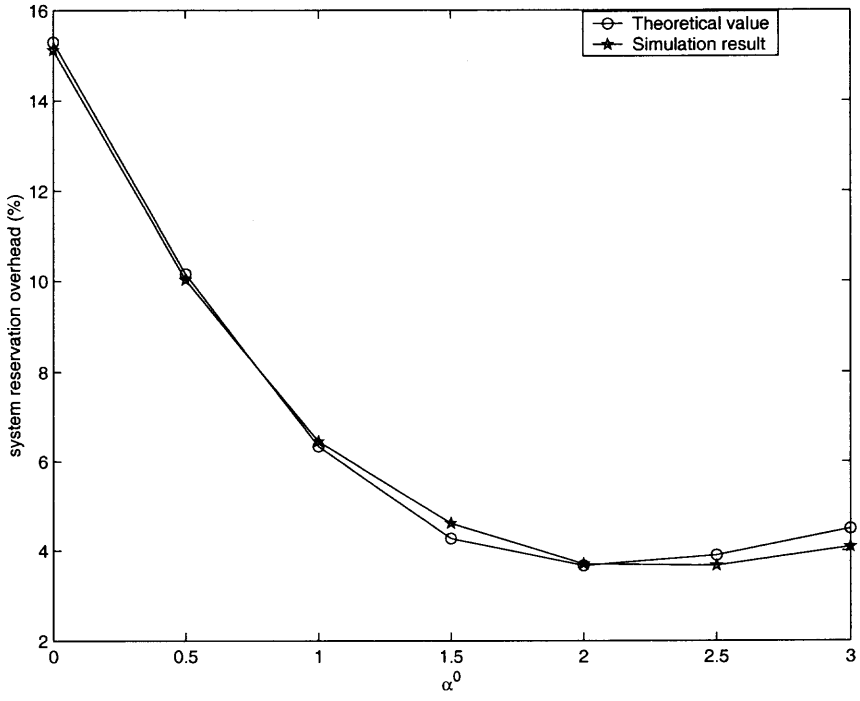

Figure 3.9 The bandwidth overhead versus δ.

success probability at the cost of a larger system bandwidth overhead (see Figures 3.8 and 3.9).

The FRR scheme gains a significant latency reduction at the cost of a very small system bandwidth overhead, as can be seen from Figures 3.6, 3.8 and 3.9, which reinforce the aforementioned conclusion that the FRR scheme should be applied in tandem with the aggressive reservation algorithm to achieve satisfactory performance figures of merits with minor operation overhead.

Although the aggressive reservation method results in a higher probability of successful BHP pre-transmissions at the cost of a bandwidth overhead, the benefit is more considerable, because bandwidth is no longer a limiting factor in the core network, and latency will be the major challenge to overcome in the future [45].

3.5.4 LPF Performance and Traffic Predictability

The accuracy of an LMS-based LPF is assessed by two parameters: $SNR^{-1} = \frac{\sum e^2(k)}{L^2(k)}$ which is the inverse of the Signal-to-Noise (SNR) ratio, and the autocorrelation of the residuals after the forecast. Special attention has been paid to the self-similar traffic scenario generated from the FFT-FGN model[46], if not otherwise specified.

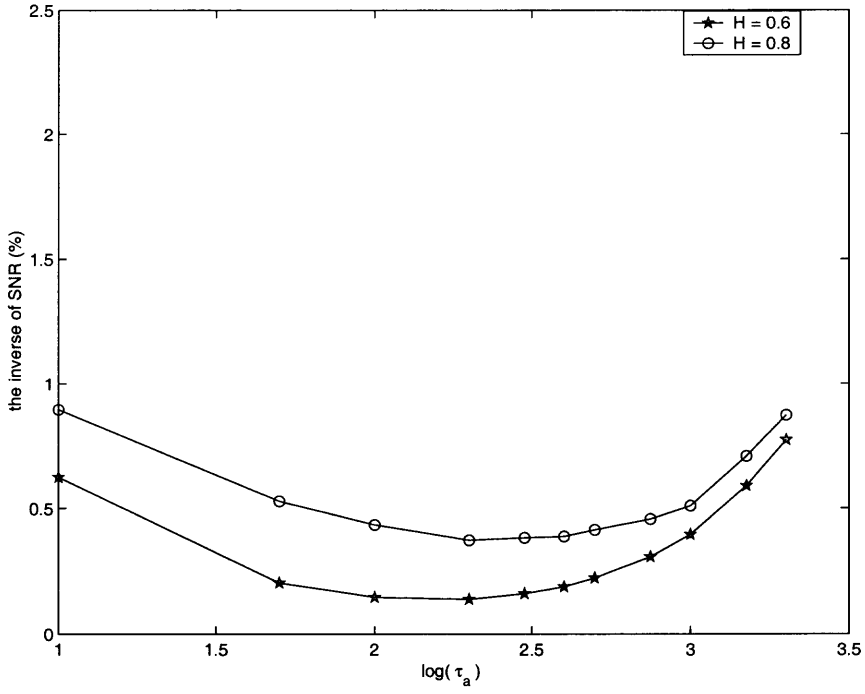

Figure 3.10 SNR^{-1} vs burst assembly interval τ_a. The mean and variance of the traffic flow are $2K$ and $100K$, respectively.

The first set of simulations are conducted by tracing the dependence of SNR^{-1} on the parameters of burst assembly duration τ_a, Hurst parameter H (the traffic bursty degree), and traffic load ρ, respectively (Figures 3.10, 3.11, and 3.12). The performance of an LPF is influenced by all the three variables. Figure 3.10 shows the effect of the burst assembly duration. While smaller SNR^{-1} values are achieved when the burst assembly time (τ_a) is between $100 - 1000\mu s$, a shorter or longer assembly time results in worse performance (i.e., larger SNR^{-1}). Meanwhile, it shows that the optimal τ_a, i.e., the burstification interval that delivers smaller SNR^{-1}, shifts

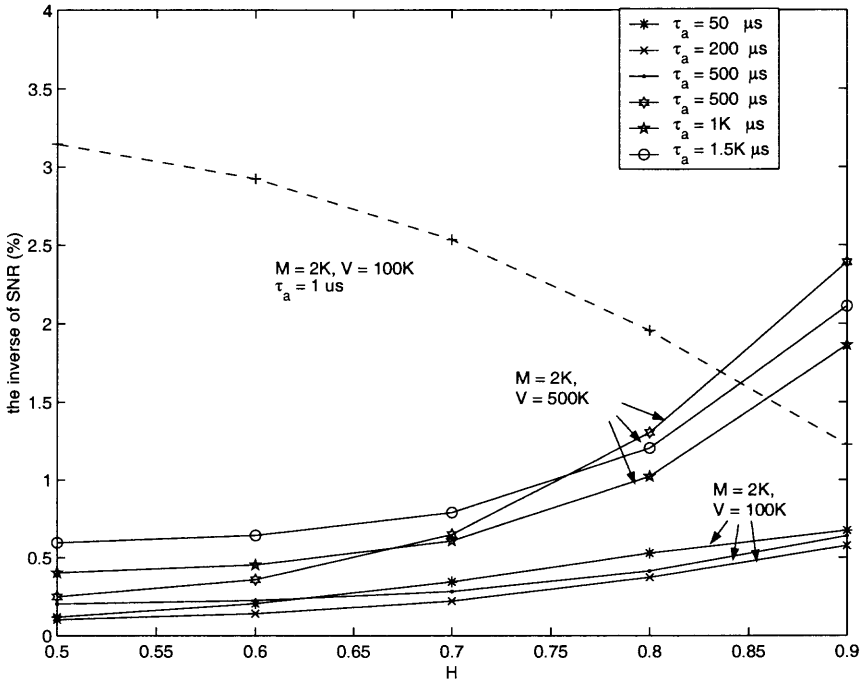

Figure 3.11 SNR^{-1} vs the Hurst parameter H. M and V represent the mean and variance of the input traffic flow, respectively.

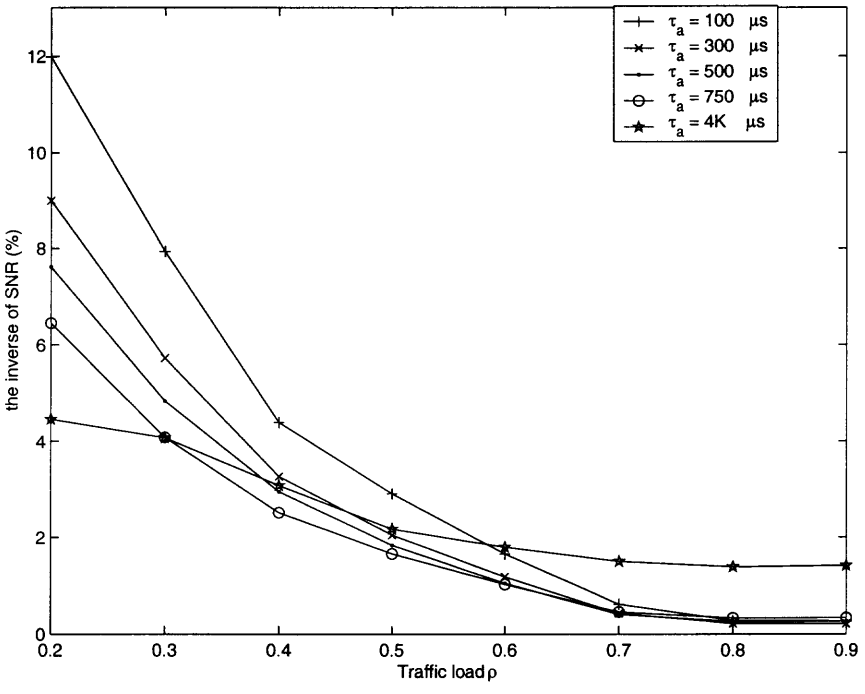

Figure 3.12 SNR^{-1} vs traffic load ρ. The input traffic is generated from 1024 ON-OFF sources. $H = 0.8$.

as the H value varies. This can also be seen from Figure 3.11, which shows that the prediction filter performance degrades slightly as the H value becomes larger. However, the LMS-based LPF presents acceptable prediction throughout the range from $H = 0.5$ to $H = 0.9$. For example, given that the mean value of the input traffic flow is 2000 $bytes/\mu s$ and variance 10^5, when the burst assembly time is $200\mu s$, the SNR^{-1} is 0.22% and 0.37% for H of 0.7 and 0.8, respectively. Note that the burstification interval changes the performance of an LPF on the self-similar traffic. For $\tau_a = 1\mu s$ (i.e., no further assembly on the input trace), the prediction performance is improved as H gets larger. This phenomenon is consistent with the conclusion given in [47, 48]. However, the effect of H on the prediction performance diminishes as the burst assembly interval grows. The traffic load also has substantial effect on the prediction performance (Figure 3.12). Given the same traffic bursty degree and burst assembly time, the performance of an LMS-based LPF increases dramatically as the traffic load increases.

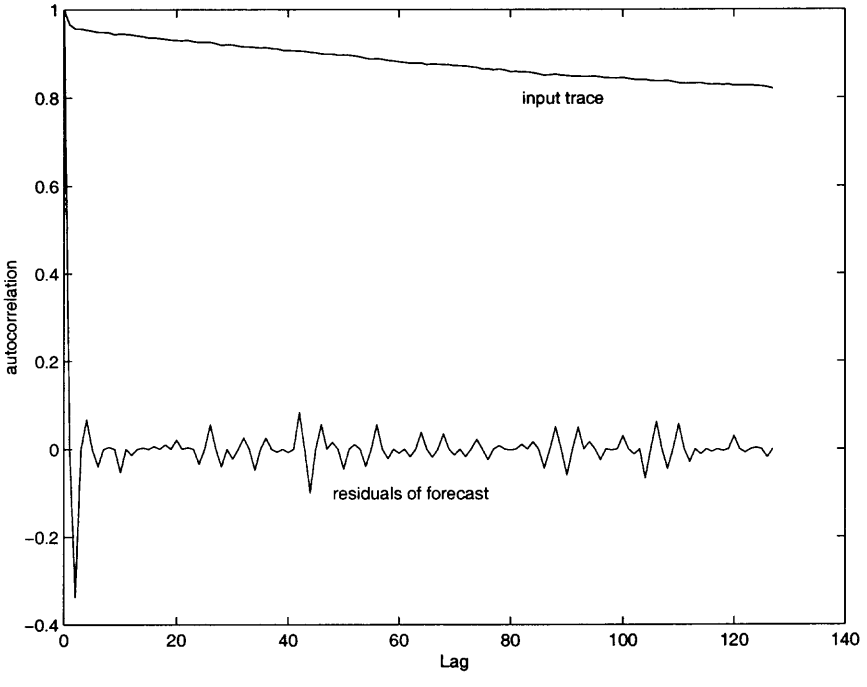

Figure 3.13 Autocorrelation of the input traffic flow and the residuals of forecast under an LMS-based LPF. $H = 0.8$.

Figure 3.13 shows the autocorrelation of the input traffic and the prediction errors. Although the input trace presents the long-range dependence, the residuals of the LMS-based forecast resemble white noise.

The simulation results on both performance metrics imply the following conclusions. First, the LMS approach can deliver satisfactory prediction for the self-similar traffic. The length of the next incoming burst can be forecasted very well. Since the real Internet traffic can be best modeled by self-similar processes, our conclusion strongly verifies the viability of our LPF-based FRR mechanism. Second, the residuals of the LMS-based forecast are approximately Gaussian distributed, justifying our previous derivations that are based on the white noise assumption. Third, in an LPF-based FRR system, a dynamic burst assembly interval is important to process the real-time traffic. The burst assembly time should be determined on-line, adaptive to the statistics derived from the previous traffic streams, i.e., $\tau_a = g(\rho, H)$. Meanwhile, we also propose that with the FRR mechanism, a burst assembly time should be no less than the burst offset time τ_o. The argument is that even though a burst assembly finishes earlier than the expiration of its offset time, the burst should wait for the end of the offset time and then be sent into the core network afterward. Algorithms to determine the optimal burst assembly duration τ_a, combining other constraints such as the number of data channels and control channels, are critically important and need further investigation.

3.6 Summary

In this chapter, a novel FRR scheme has been proposed and proved to be practical in reducing the data burst delay at network ingresses of an OBS system. The FRR scheme consists of three inherent features: a parallel execution of BHP signaling and burstification, an LMS-based LPF for burst-length prediction, and an aggressive

resource reservation. The FRR scheme has also been extended to facilitate QoS differentiation at network edges.

Theoretical analysis and simulations exhibit encouraging results. The FRR mechanism leads to a significant latency reduction based on simple algorithms and mature techniques. QoS differentiation is facilitated at network edges. Furthermore, it is shown that the FRR scheme, in tandem with the aggressive reservation strategy, results in less signaling re-transmissions and bandwidth overhead as compared to a zero-correction system. The LMS-based LPF delivers excellent forecasting performance for the self-similar traffic which best models the Internet traffic. Optimal performance of the LPF has been found to depend on a variety of traffic parameters, including the traffic load, self-similar degree, and prediction interval. Such dependence on prediction interval implies the importance to devise algorithms that dynamically determine the burstification duration.

The FRR scheme presented in this chapter is a skeleton based on which a variety of specific algorithms are possible to solve different practical problems. For example, the aggressive reservation strategy can be implemented to facilitate different determination algorithms of the correction values. Similarly, the LPF can be based on recursive algorithms or on block algorithms. The way in which the edge nodes and the intermediate nodes communicate to negotiate for parameter adjustments can also vary. In the following chapters, algorithms and solutions to some of these aspects will be proposed and investigated.

CHAPTER 4

AGGRESSIVE RESERVATION ALGORITHMS

One of the important features of our proposed FRR scheme is the aggressive reservation strategy. That is, the pre-transmitted BHP attempts to reserve resources at the intermediate nodes in an aggressive manner, rather than defining the reservation length to be exactly equal to the predicted data burst length. This chapter will investigate the aggressive reservation strategy for FRR-embedded OBS systems. Specifically, two algorithms, the success probability-driven (SPD) algorithm and a bandwidth usage-driven (BUD) algorithm, will be proposed to facilitate the FRR transmission scheme.

4.1 Motivation

According to the proceeding discussion, the salient characteristic of the aggressive reservation strategy can be expressed as

$$L_r = \tilde{L}_d + \delta, \tag{4.1}$$

where δ acts as the channel holding time adjustment to compensate for the imperfection of the underlying predictive filter. Aiming at increasing the forward reservation success probability, the aggressive reservation strategy improves the latency reduction capability of the FRR scheme.

Transmission latency and bandwidth utilization are two important performance figures of merits to be addressed in the next generation network. The correction value δ has a significant impact on both the BHP pre-transmission success probability (P_s)– therefore the latency reduction capability, and the resource usage efficiency (denoted as η). It should be determined based on the tradeoff consideration of these two

performance figures, whose relative importance to network management differs in different network applications.

In addition to the system scenario described in the last chapter, out-of-band signaling is assumed in this chapter, i.e., data payload and its control header are transmitted in separate channels and at different time domains. Signaling messages (e.g., BHPs) are queued and electronically processed by each intermediate node. Signaling channel is best effort link by link. Queuing losses are possible. However, signaling channel is presumed to possess a low bit error rate (BER), e.g., from 10^{-12} to 10^{-15} [21]. Taking into account the low BER and the heavy burden of maintaining the burst information, data burst re-transmission is not desirable in the WDM layer. The underlying predictive filer is assumed to be an N-order LMS-based LPF.

4.2 Aggressive Reservation Algorithms

In this section, we explain the proposed aggressive reservation algorithms, and assesses their performance in terms of the BHP pre-transmission success probability and the bandwidth usage efficiency.

4.2.1 Success Probability-driven (SPD) Algorithm

The intuitive motivation of this algorithm is to achieve explicit control on the BHP pre-transmission success probability (P_s), and thus achieve a deterministic latency reduction percentage (See Equation 3.6). For this purpose, the correction value δ is defined to be some multiple of the standard derivation (σ) of the prediction residuals resulted from the underlying LPF, i.e., $\delta = \alpha \cdot \sigma$, where $\alpha \geq 0$ is a real value. Given that the prediction residuals are approximately Gaussian distributed with mean zero and variance σ^2, based on Equation 3.7 and the analysis in Chapter 3, we get

$$P_s = P(e(k) \leq \delta) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot \int_{-\infty}^{\delta} e^{-\frac{t^2}{2 \cdot \sigma^2}} dt = 1 - Q(\alpha). \tag{4.2}$$

This way, to achieve some desired P_s, one can directly choose an α value that satisfies $\alpha = \frac{\delta}{\sigma} = Q^{-1}(1 - P_s)$.

The SPD algorithm features the advantage that P_s is deterministic to a specified α , and is independent of the burstification scheme, the traffic load, and the performance of the underlying LPF. This algorithm is more appropriate when the forward reservation success probability (and the latency reduction capability) is of essence to the network management.

From the implementation perspective, the correction value δ is defined to be some multiple of the sample Root-Mean-Square (RMS) of the forecast residuals of the underlying LPF, i.e.,

$$\delta = \alpha \cdot \sqrt{\frac{\sum_{i=0}^{N-1} e^2(i)}{N}}, \tag{4.3}$$

where $e^2(i), i \in 0 \ldots N-1$ represents the N latest prediction residuals.

Given a specific data burst with length L_d, the probability distribution function of the corresponding prediction length is

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{(x-L_d)^2}{2 \cdot \sigma^2}}. \tag{4.4}$$

Therefore, the bandwidth usage efficiency of the SPD algorithm can be expressed as

$$\begin{aligned} \omega \;=\; & 1 - [\frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \int_{L_d - \delta}^{\infty} (x + \delta - L_d) \cdot e^{-\frac{(x-L_d)^2}{2 \cdot \sigma^2}} dx] \cdot \frac{1}{L_d} - \\ & [\frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \int_{-\infty}^{L_d - \delta} (x + \delta) \cdot e^{-\frac{(x-L_d)^2}{2 \cdot \sigma^2}} dx] \cdot \frac{1}{L_d}, \end{aligned} \tag{4.5}$$

where the second and the third terms represent the bandwidth overhead incurred when the BHP pre-transmission succeeds and fails, respectively. From Equation 4.5,

$$\begin{aligned} \omega = 1 - & [\frac{\sigma}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{(\delta)^2}{2 \cdot \sigma^2}} + \delta \cdot (1 - Q(\frac{\delta}{\sigma}))] \cdot \frac{1}{L_d} - \\ & [\frac{-\sigma}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{(\delta)^2}{2 \cdot \sigma^2}} + (L_d + \delta) \cdot (1 - Q(\frac{\delta}{\sigma}))] \cdot \frac{1}{L_d}. \end{aligned} \tag{4.6}$$

Therefore,

$$\omega = 1 - Q(\frac{\delta}{\sigma}) - \frac{\delta}{L_d}. \tag{4.7}$$

Given that $\delta = \alpha \cdot \sigma$ in the SPD algorithm, Equation 4.7 can be expressed as

$$\omega = 1 - Q(\alpha) - \frac{\alpha \cdot \sigma}{L_d}. \tag{4.8}$$

Provided that the BHP pre-reservation succeeds at a high probability ($P_s \rightarrow 100\%$), the bandwidth usage efficiency is $\omega = 1 - \frac{\alpha \cdot \sigma}{L_d}$.

4.2.2 Bandwidth Usage-driven (BUD) Algorithm

An alternative algorithm is to define the compensation value from the bandwidth utilization viewpoint. In a system that the explicit control on the bandwidth usage efficiency is more concerned to the network management, the BUD algorithm is adopted, whereby δ is determined directly based on the bandwidth overhead allowance, expressed as $\delta = \rho \cdot \tilde{L}_d$, where $\rho \geq 0$ is a real-value constant. The pre-reservation length is thus defined as $L_r = \tilde{L}_d + \rho \cdot \tilde{L}_d$. The bandwidth usage efficiency of the BUD algorithm is thus

$$\omega = 1 - \rho \cdot P_s - (1 + \rho) \cdot (1 - P_s) = P_s - \rho. \tag{4.9}$$

Provided that the BHP pre-reservation succeeds at a high probability ($P_s \rightarrow 100\%$), the bandwidth usage efficiency due to the introduction of ρ is

$$\omega = 1 - \rho, \tag{4.10}$$

implying that the bandwidth usage efficiency caused by the BUD algorithm can be easily derived as long as ρ is determined, and is independent of the performance of the underlying LPF or the burst assembly scheme (e.g., the determination for τ_a).

In the BUD algorithm, the BHP pre-transmission succeeds when $L_r \geq L_d$, i.e., when $\tilde{L}_d + \rho \cdot \tilde{L}_d \geq L_d$. Based on Equation 3.7, P_s in this algorithm is

$$P_s = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot \int_{\frac{L_d}{1+\rho}}^{\infty} e^{-\frac{(x-L_d)^2}{2 \cdot \sigma^2}} dx, \tag{4.11}$$

which is further simplified to be

$$P_s = 1 - Q(\frac{\rho}{1+\rho} \cdot \frac{L_d}{\sigma}). \tag{4.12}$$

Equation 4.12 indicates that the BHP pre-transmission success probability of the BUD algorithm factors in not only the parameter of ρ, but also the standard derivation of the forecast residuals, and the average burst length which is affected by the average traffic load and the burst assembly time.

In both the SPD and the BUD algorithms, the channel holding time contained in a pre-transmitted BHP (L_r) is determined by both the predicted data burst length $\tilde{L}_d$ and the correction value δ. Hereafter, we define the compensation ratio to be the ratio of δ to $\tilde{L}_d$, and denote it as ϕ. The value of ϕ indicates the fraction of the aggressively reserved extra resource as compared to the actual burst length. Given a system with determined underlying LPF and the input traffic load, ϕ and τ_a are two parameters that the network management should specify to obtain the desired system performance.

4.3 Simulation Results and Observations

In this section, the two proposed algorithms are compared via simulations. Interesting performance metrics include the BHP pre-reservation success probability (P_s) and the bandwidth usage efficiency (ω), with an emphasis on their dependence on the compensation ratio (ϕ) and the burst assembly time (τ_a), given a determined underlying LPF and the input traffic load. It is expected that the SPD algorithm outperforms the BUD algorithm when the explicit control on P_s is of the essence, while the opposite

effect takes place if the control on ω is more important. In the following discussion, τ_a will be normalized with respect to the average time to transmit one IP packet of 1500 bytes, and will be referred to as τ.

Simulations are conducted based on the self-similarity traffic process, which best represents the traffic characteristic of the current Ethernet [49, 50]. The packet stream flowing into the edge node is produced by the FFT-FGN model [46], with the Hurst parameter of $H = 0.75$ and an average packet size of 2000 bytes. A 12-order LMS-based LPF is adopted for burst length prediction.

4.3.1 Performance of the BHP Pre-transmission Success Probability (P_s)

Figure 4.1 shows the relationship between P_s and ϕ. It can be seen that for the same compensation ratio, both algorithms present similar BHP pre-reservation success probability, and consequently similar latency reduction capability (see Equation 3.6). A high P_s is obtainable with small compensation ratio values. For example, when $\phi = 0.1$, both algorithms achieve $P_s \geq 95\%$ (for $\tau_a = 100$). Meanwhile, a larger compensation ratio contributes higher P_s. Since the latency reduction performance of an FRR scheme increases as P_s approaches to 1, the larger compensation ratio will deliver higher latency reduction improvement. In addition, the impact of ϕ on P_s is affected by the burst assembly time. A τ value larger than a single time unit leads to a more significant improvement on P_s.

Figure 4.2 demonstrates the advantage of the SPD algorithm, showing that with this algorithm, one can achieve explicit control on P_s by properly choosing α, and the resulted P_s is basically independent of τ . This is consistent with Equation 4.2, which indicates that P_s is only a function of α, and therefore it should remain constant as long as α is determined.

The independence of P_s to τ does not hold for the BUD algorithm, where P_s is considerably influenced by τ, and the behavior of P_s versus τ differs as ϕ changes.

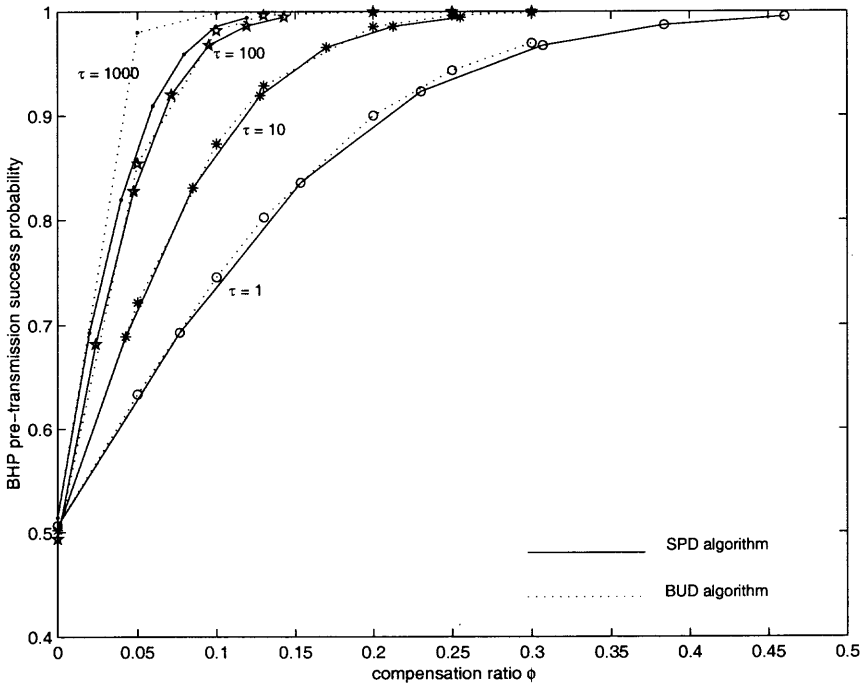

Figure 4.1 BHP pre-transmission success probability versus the compensation ratio.

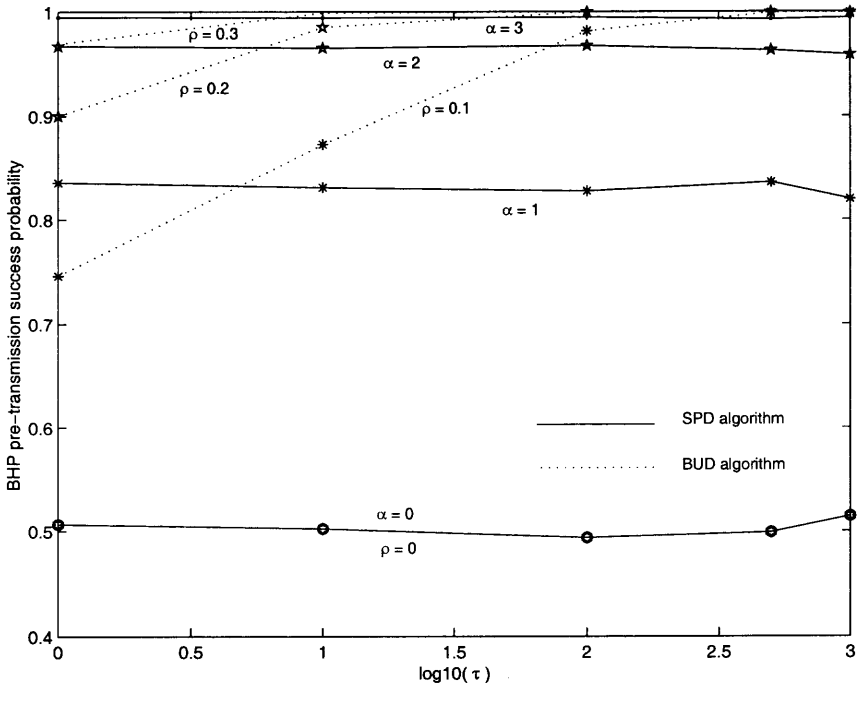

Figure 4.2 BHP pre-transmission success probability versus the burstification duration.

Equation 4.12 explains this phenomenon. Given that the average burst length is a function of both the average traffic load and the burst assembly interval, and that the average traffic load remains the same, a different τ yields a different P_s.

4.3.2 Performance of the Bandwidth Usage Efficiency (ω)

Figure 4.3 plots ω versus ϕ for several values of τ. As expected, given the determined burst assembly time, the two algorithms deliver similar bandwidth usage efficiency for the same compensation ratio. Meanwhile, it is shown that increasing the compensation ratio initially results in an improvement on ω, i.e., the aggressive reservation strategy actually improves the bandwidth utilization of an FRR system. This is because the correction value δ acts as a channel holding time adjustment that substantially increases the pre-reservation success probability. Since an unsuccessful forward resource reservation (due to the insufficient reservation duration) incurs the bandwidth wastage of roughly comparable to the burst size, the reduction of the unsuccessful BHP pre-transmission owing to the marginal reservation compensation will eventually result in less bandwidth overhead, consequently better bandwidth usage efficiency.

Another important implication of Figure 4.3 is that, for any given τ, there exists some optimal compensation ratio, i.e., a compensation ratio threshold that results in the highest bandwidth usage efficiency. Compensation ratios that exceed such a threshold will cause the bandwidth usage efficiency to degrade. This is because as ϕ gets larger, the bandwidth overhead due to the correction values becomes more significant, and consequently, the bandwidth utilization drops.

The impact of the compensation ratio on ω is also affected by τ. On one hand, for the same compensation ratio, a system with longer burst assembly time delivers better bandwidth utilization. On the other hand, the optimal compensation ratio

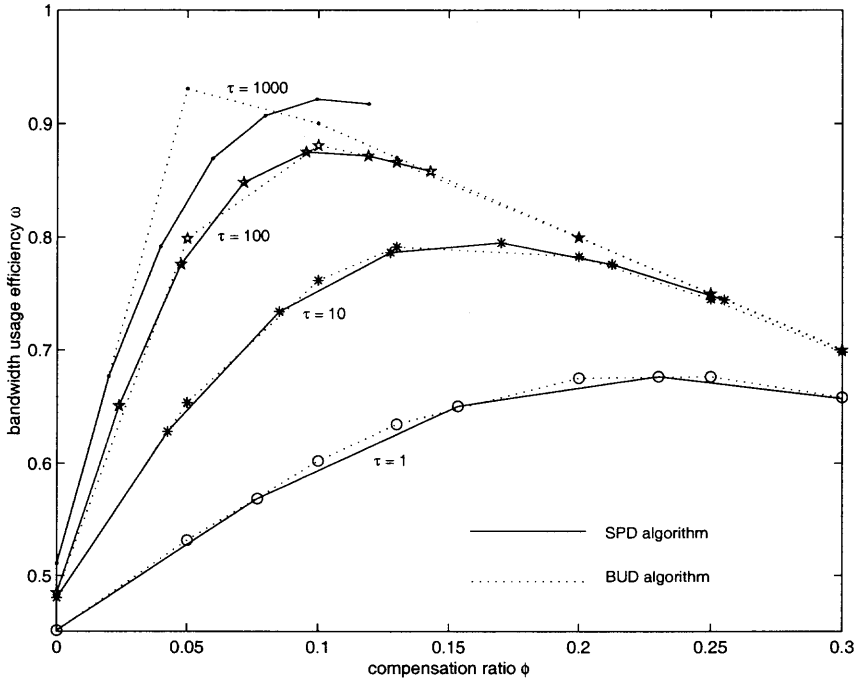

Figure 4.3 The bandwidth usage efficiency versus the compensation ratio.

becomes smaller as the burstification interval increases, while still resulting in better bandwidth usage efficiency.

Figure 4.4 exhibits the advantage of the BUD algorithm, in terms of the more stable and explicit control on the bandwidth usage efficiency as long as the compensation ratio is specified. As τ changes, the ω of the BUD algorithm remains constant. This is especially true for larger τ (e.g., $\tau \geq 10$ and accordingly $\log \tau \geq 1$), or larger ϕ (e.g., $\rho \geq 0.15$ which indicates that $\phi \geq 0.15$, whereby the BHP pre-transmission of both algorithms succeed at a high probability (i.e., $P_s \rightarrow 100\%$). Likewise, this conclusion is consistent with our expectation based on Equation 4.8 and Equation 4.10.

4.3.3 P_s Versus ω

The discussion in the previous sections leads to a conclusion that, the explicit control on P_s and ω, both important performance figures of an FRR system, can be achieved

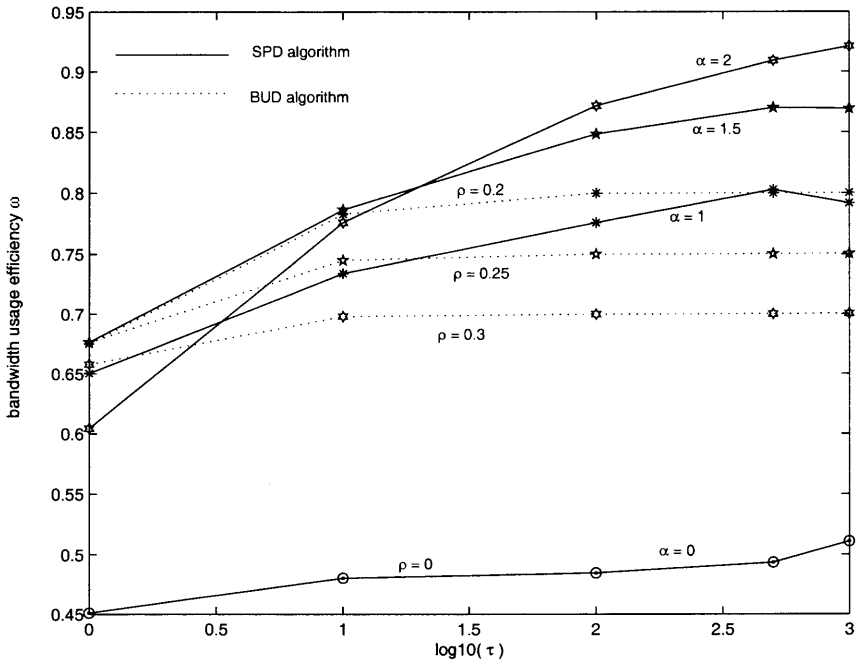

Figure 4.4 The bandwidth usage efficiency versus the burstification duration.

by selecting different implementation algorithms and deciding the respective system parameter (e.g., the compensation ratio). Since the FRR scheme employing the aggressive reservation strategy improves the latency reduction performance at the expense of the bandwidth overhead, an intuitive criterion to assess the quality of the system configuration is to decide how much the bandwidth usage efficiency is achieved for a determined P_s, or vice versa. An optimal system design should take into consideration of both performance figures, therefore balancing between the latency reduction capability (which is linear to P_s) and the bandwidth overhead.

The relationship between ω and P_s is directly plotted in Figure 4.5. As expected, both algorithms possess a similar relationship between these two performance figures. When P_s is low, ω grows as P_s increases, indicating that the compensation ratio within this range improves both ω and P_s. In this situation, increasing the compensation ratio will further benefit the system performance while still lowering the system cost. When P_s approaches a point (whereby the optimal compensation ratio is reached

or exceeded), ω goes downward. This reinforces the aforementioned conclusion that the compensation ratio larger than some threshold will result in reduced bandwidth utilization with limited improvement on P_s. In this case, the correction value should back off to reduce its negative impact on the system performance.

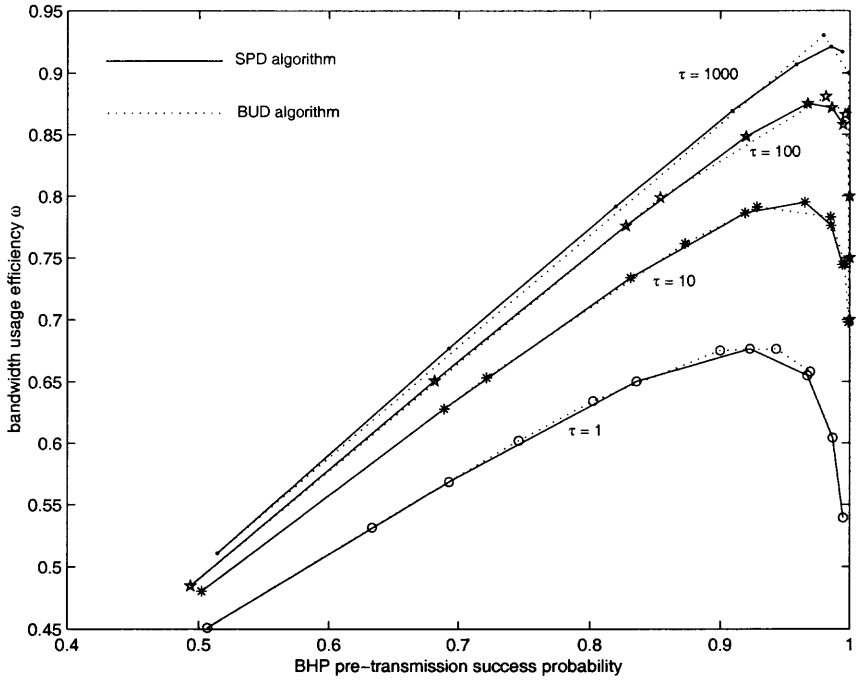

Figure 4.5 The bandwidth usage efficiency versus the BHP pre-transmission success probability.

4.4 Summary

This chapter has investigated the aggressive reservation strategy for an FRR-embedded OBS system. Two aggressive reservation algorithms, each of which is focused on the explicit control on the BHP pre-transmission success probability and the bandwidth usage efficiency, respectively, have been proposed. Theoretical analysis and simulation results demonstrated the respective advantages of the proposed algorithms.

The following observations can be concluded from the analysis and simulation results: 1) Given the same compensation ratio, both the SPD and the BUD algorithms

deliver similar performance in terms of the bandwidth usage efficiency and BHP pre-transmission success probability–therefore the latency reduction percentage of the FRR-embedded OBS system. 2) The SPD algorithm facilitates direct control on the BHP pre-transmission success probability. Deterministic latency reduction requirement can thus be achieved by directly selecting the compensation ratio. 3) In contrast, the BUD algorithm enables explicit control on the bandwidth usage efficiency, and is more appropriate when the bandwidth utilization is of more concern to the system management.

The analysis and simulation results presented in this chapter also imply that several issues remain to be addressed to further improve the performance of the FRR scheme. For example, in the BUD algorithm, ω diverges from the expected value defined in Equation 4.9 when the burstification duration is small (e.g., $\tau \leq 10$). Such performance discrepancy is primarily due to the inherent mechanism of the FRR scheme. That is, when the pre-reserved resource is insufficient to transport the data burst, the BHP is re-transmitted with a new reservation length equal to the actual burst length, and the pre-reserved resources are simply left unused. If the BHP pre-transmission fails at a high probability, the bandwidth usage efficiency degrades significantly. Therefore, innovative control schemes that can further reduce the bandwidth overhead caused by this source are highly desired.

CHAPTER 5

PERFORMANCE IMPROVEMENT FOR THE FRR SCHEME

This chapter is focused on the optimization issues of the FRR scheme. The aim is to enhance the performance of the FRR scheme by improving the bandwidth usage efficiency when the BHP pre-transmission fails or succeeds.

5.1 Motivation

The discussion in the previous chapters demonstrates that the bandwidth usage efficiency of the FRR scheme is affected by two factors: the compensation ratio ϕ which is introduced by the aggressive reservation strategy, and the bandwidth wastage due to insufficient forward resource reservations.

Accordingly, there are two disciplines to improve the bandwidth utilization of the FRR-embedded OBS network. The first one is a proper planning mechanism for the correction value δ, which serves as a channel holding time adjustment. A too large correction value (e.g., the compensation ratio is larger than the compensation threshold) induces considerable negative impact on the bandwidth utilization with only marginal improvement on the latency reduction. This problem can be overcomed by properly budgeting δ in order to balance the performance gains (e.g., the latency reduction capability) and the operation cost (e.g., the bandwidth overhead).

The other discipline is a control mechanism which reduces the bandwidth wastage due to unsuccessful forward resource reservations. In the FRR scheme, nothing is done to the pre-reserved resource which is insufficient to support the transmission of the actual data burst. This results in the bandwidth utilization discrepancy, especially when the BHP pre-transmission fails with a high probability. Mechanisms that make intelligent usage of such resources are highly desired.

In this chapter, the FRR scheme optimization issues will be investigated based on these two disciplines. First, the channel holding time adjustment is determined to optimize the system performance in terms of the minimum reservation overhead and the maximum net performance gain, respectively, thus providing a guideline for the network designer to configure the system parameters according to the desired performance figures of merits. Second, a bandwidth enhancement mechanism is explored to make better usage of the network resources. The aim is to render the bandwidth enhancement capability for the FRR scheme, thus improving the bandwidth utilization of the FRR-embedded OBS systems. Hereafter, for simplicity, an FRR scheme which adopts the proposed bandwidth enhancement mechanism will be referred to as the BEFRR (Bandwidth Enhanced FRR) scheme, and that without the bandwidth enhancement mechanism as the basic FRR scheme.

5.2 Determining the Channel Holding Time Adjustment

The channel holding time adjustment has a significant impact on the important performance figures of the FRR-enabled OBS networks. With a too small adjustment value (e.g., $\delta \rightarrow 0$), the potential benefits of the aggressive reservation strategy (in terms of the improved η and P_s, and the reduced γ) are not fully exploited, while a too large adjustment value induces adverse impact on the reservation overhead with only marginal improvement on the latency reduction. The intuitive criteria to justify the usage of a specific δ value are to assess how much the latency reduction it can contribute, how much the associated reservation overhead is, and how much the performance improvement outweighs the system cost, which is referred to in this chapter as the net performance gain.

In this section, the FRR-enabled system performance is improved by determining the channel holding time adjustment when the performance optimization criteria change.

5.2.1 Design Objectives and Assumptions

In this chapter, the system performance is optimized in terms of the following two perspectives:

- To minimize the reservation overhead γ. That is, to find the adjustment threshold δ^* which yields the minimum reservation overhead γ^*.
- To maximize the system net performance gain ψ, which is formulated as

$$\psi = \sum I - k \cdot \sum C. \tag{5.1}$$

In Equation 5.1, $\sum I$ represents the total performance improvement enabled by the aggressive reservation strategy, and $\sum C$ is the associated operation cost. k is the loss/gain ratio representing the relative importance of the system cost to that of the performance gain. The network management should specify its actual value according to the specific network conditions. ψ is thus a system characteristic to justify the optimality of the adjustment value.

Note that the interesting performance figures involved in both components of Equation 5.1 are optional. For example, the signaling message overhead may also be considered as a factor of the system cost in addition to the reservation overhead. In this chapter, however, a preliminary system optimization problem is considered, i.e., the system performance gain is merely targeted to be the latency reduction capability, and the reservation overhead the system cost.

The following assumptions are made for deriving the optimal channel holding time adjustment to meet different system optimization objectives.

The burstification interval (τ_a) is equal to the offset value between a BHP and the data payload (τ_o). This way, in Equation 3.6, $\mu = \tau_o/\tau_a = 1$.

Without loss of generality, the network situation is assumed to be $k = 1$, i.e., the bandwidth wastage and the latency improvement capability are of equal essence

to the network management. The proposed solution can be easily extended for other network scenarios with different loss/gain ratios.

The Success Probability Driven (SPD) algorithm is adopted to implement the aggressive strategy, wherein the channel holding time adjustment is set to be some multiple of the standard derivation of the prediction residuals, i.e., $\delta = \alpha \cdot \sigma$, where α is a real-value. This way, finding the optimal adjustment value can be simplified to determining the value of α, since σ is independent of the choice of the α value for the given underlying adaptive filter and the WDM network.

5.2.2 Optimize the Reservation Overhead γ

Based on the previous discussions, both γ_s and γ_f are functions of δ, and can be expressed as

$$\gamma_s = [\frac{1}{\sqrt{2\pi}\sigma} \int_{L_d-\delta}^{\infty} (x + \delta - L_d) \cdot e^{-\frac{(x-L_d)^2}{2\sigma^2}} dx] \cdot \frac{1}{L_d} \tag{5.2}$$

and

$$\gamma_f = [\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{L_d-\delta} (x + \delta) \cdot e^{-\frac{(x-L_d)^2}{2\sigma^2}} dx] \cdot \frac{1}{L_d} \tag{5.3}$$

respectively, where L_d is the average data burst length. Combining Equation 5.2 and Equation 5.3,

$$\gamma = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\infty} e^{-\frac{t^2}{2}} dt + \frac{\alpha \cdot \sigma}{L_d} \tag{5.4}$$

To achieve the optimal system performance in terms of the minimum reservation overhead, the α value should satisfy

$$\frac{d\gamma}{d\alpha} = \frac{\sigma}{L_d} - \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\alpha^2}{2}} = 0. \tag{5.5}$$

Therefore,

$$\alpha^* = (-2 \cdot ln(\sqrt{2\pi} \cdot \frac{\sigma}{L_d}))^{\frac{1}{2}} \tag{5.6}$$

Both σ and L_d are independent of the choice of the α value, and are a parameter of the burstification interval τ_a. Denote $\frac{\sigma}{L_d}$ as q_t. The channel holding time adjustment which delivers the minimum reservation overhead is thus

$$\delta^* = \sigma \cdot (-2 \cdot ln(\sqrt{2\pi} \cdot q_t))^{\frac{1}{2}}. \tag{5.7}$$

In this situation,

$$\gamma_s^* = \frac{q_t}{\sqrt{2\pi}} \cdot e^{-\frac{(\alpha^*)^2}{2}} + q_t \cdot \alpha^* Q(-\alpha^*), \tag{5.8}$$

and

$$\gamma_f^* = -\frac{q_t}{\sqrt{2\pi}} \cdot e^{-\frac{(\alpha^*)^2}{2}} + (1 + q_t \cdot \alpha^*) Q(\alpha^*), \tag{5.9}$$

respectively, wherein the $Q(\cdot)$ is the $Q-$ function [44]. Similarly, the minimum reservation overhead γ^* is

$$\gamma^* = q_t \cdot \alpha^* + Q(\alpha^*). \tag{5.10}$$

5.2.3 Optimize the Net Performance Gain ψ

In this subsection, the optimal adjustment is determinated according to its effect on the system net performance gain ψ, as defined in Equation 5.1. Based on the previous discussions, the parameter ψ becomes

$$\psi = \frac{2}{3} - \frac{5}{3} Q(\alpha) - \alpha \cdot q_t. \tag{5.11}$$

The α value which achieves the maximum value for ψ is thus

$$\alpha^{**} = (-2 \cdot ln(\frac{3\sqrt{2\pi}}{5} \cdot q_t)^{\frac{1}{2}}, \tag{5.12}$$

Therefore, the optimal adjustment value which optimizes the system in terms of the maximum ψ is

$$\delta^{**} = \sigma \cdot (-2 \cdot ln(\frac{3\sqrt{2\pi}}{5} \cdot q_t)^{\frac{1}{2}}. \tag{5.13}$$

Similarly, the latency reduction capability, the reservation overhead, and the system net performance gain in this situation can be expressed as

$$\eta^{**} = \frac{2}{3}(1 - Q(\alpha^{**})), \tag{5.14}$$

$$\gamma^{**} = Q(\alpha^{**}) + \alpha^{**} \cdot q_t, \tag{5.15}$$

and

$$\psi^{**} = \frac{2}{3} - \frac{5}{3} \cdot Q(\alpha^{**}) - \alpha^{**} \cdot q_t. \tag{5.16}$$

5.2.4 Numerical and Simulation Results

Simulations are conducted to justify the above adjustment value determination for different optimization objectives as the burstification interval changes. A 12-order LPF is utilized for data burst length prediction. The traffic flowing into the ingress node is assumed to be a self-similarity process, generated based on the FFT-FGN model [46], with the Hurst parameter of $H = 0.75$ and the average packet size of 2000 bytes. The burst assembly duration (τ_a) is normalized with respect to the time to transmit one IP packet of 1500 bytes.

Before verifying the optimality of the preferred α value for different design objectives, the effect of the burstification interval on q_t is first illustrated (Figure 5.1), as q_t is an important parameter of the optimal adjustment value. This figure, together with Equation 5.6 and Equation 5.12, implies that the optimal adjustment value should change dynamically as the burstification interval varies, so as to optimize the network resource utilization and the system performance.

Figure 5.2 represents the effect of the channel holding time adjustment on different performance figures of merits when the burstification interval is set to be 1000 and k (i.e., the loss/gain coefficient) is equal to 1. It can be seen that different optimal adjustment values exist for different system performance measurements.

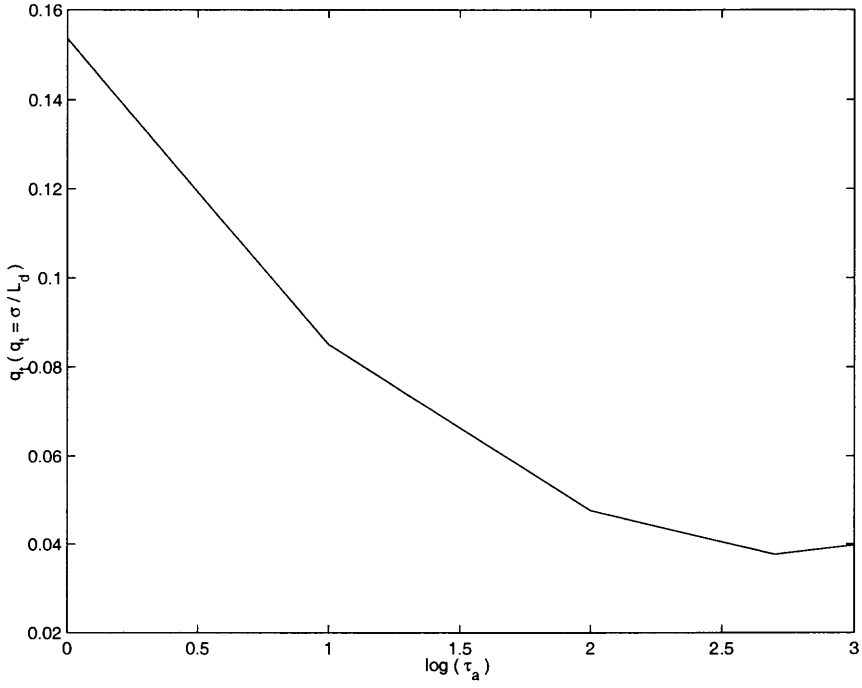

Figure 5.1 q_t versus the burstification interval τ_a.

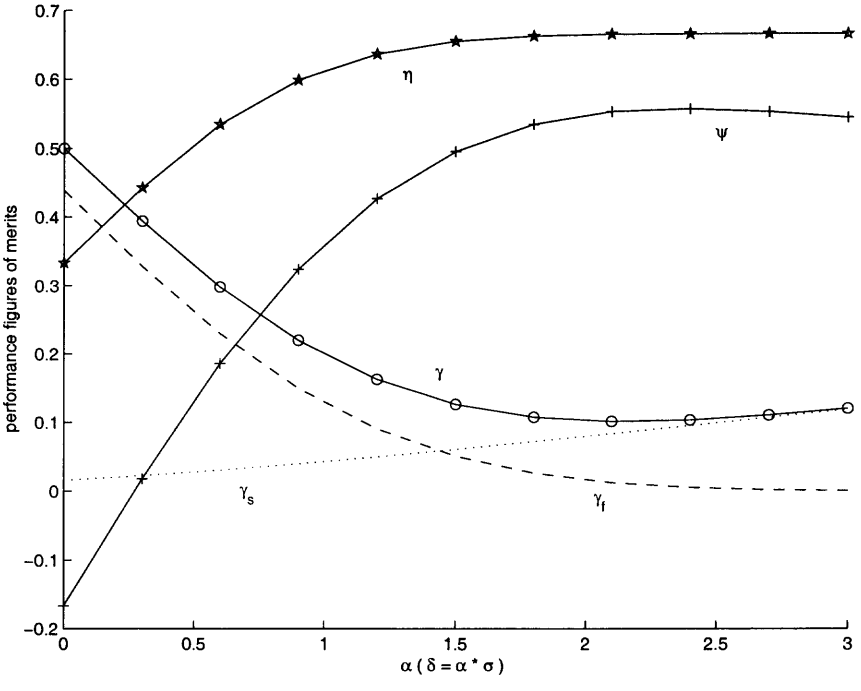

Figure 5.2 The system performance parameter versus the channel holding time adjustment.

Based on Equation 5.6, the α values which deliver the minimum γ and the maximum ψ are $\alpha^* = 2.15$ and $\alpha^{**} = 2.37$, respectively. The simulation results match very well with the analytical solutions.

It is also observed that both the reservation overhead and the net performance gain are improved when the α value initially grows larger than 0, and that their performance gets degraded when the α value grows larger than 3. This implies that the optimal adjustment value should be in between.

Figure 5.3 presents the performance of the reservation overhead delivered by different channel holding time adjustments as the burstification interval changes. It can be seen that the α value determined by Equation 5.6 always yields the minimum reservation overhead as the burstification interval changes. This justifies the optimality of the above adjustment value determination.

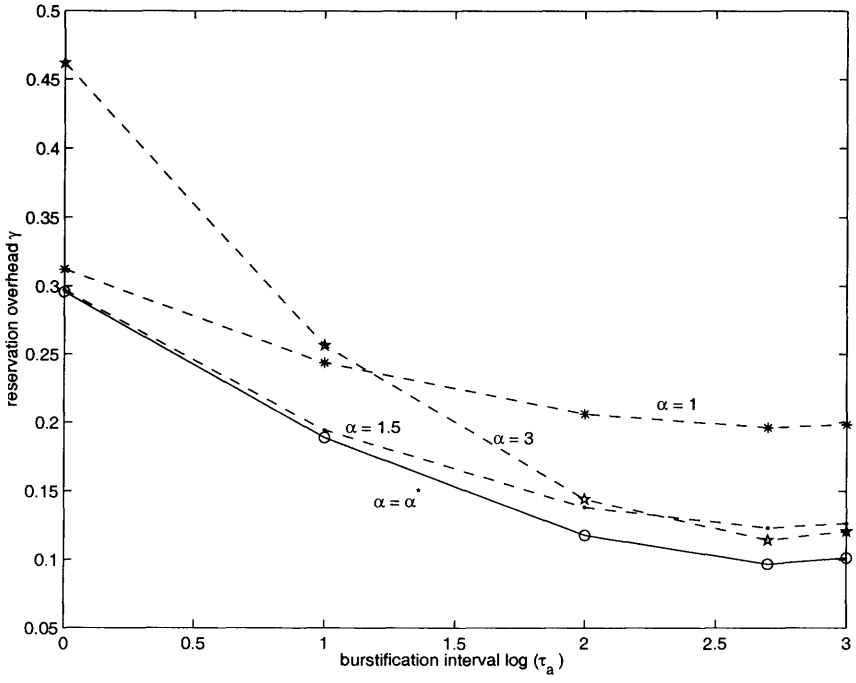

Figure 5.3 The reservation overhead versus the burstification interval.

When the system optimization objective is to maximize the net performance gain ψ, the α value concludes to be determined by Equation 5.12. Figure 5.4 reinforces

the derived solution, illustrating that the net performance gain delivered by δ^{**} (where $\delta^{**} = \alpha^{**} \cdot \sigma$) exceeds that delivered by other adjustment values.

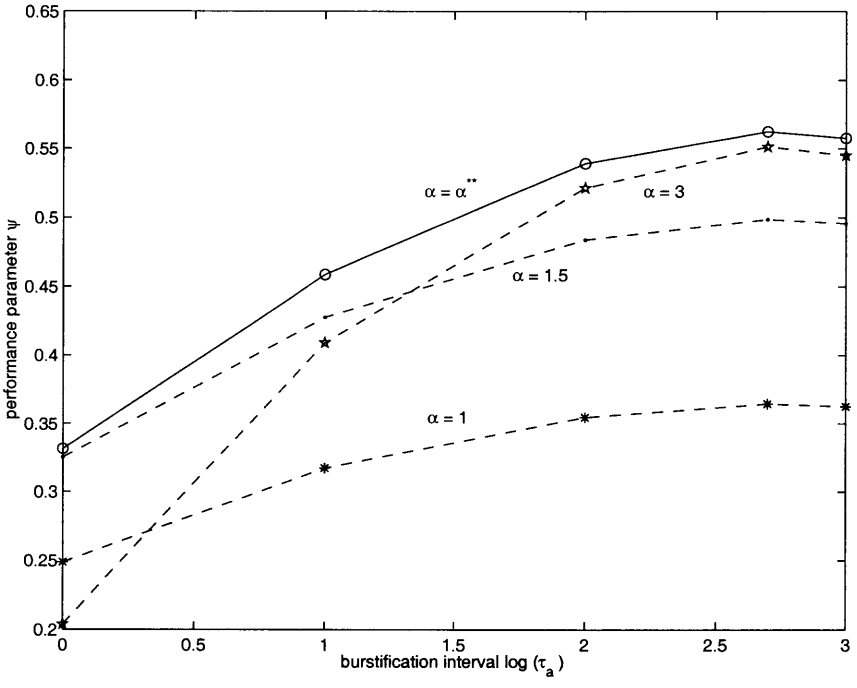

Figure 5.4 The performance versus the burstification interval.

5.3 Bandwidth Enhanced FRR Scheme

This section describes the system environment and the design objective of the BEFRR scheme, illustrates the BEFRR scheme principle, and assesses its performance in terms of the bandwidth savings and the signaling overhead.

5.3.1 System Model and Assumptions

The design for the BEFRR scheme is based on the same system environment as described in Chapter 3. Several functionalities of the edge nodes that are critical to our current study are described.

Both the ingresses and the intermediate nodes are equipped with timers to make sure an action is carried out within specific time constraints. For example, Timer A at the ingress node is set to be the burstification duration. A time-out on

Timer A indicates that the burstification is completed. Meanwhile, Timer B at the intermediate node is set to be the estimated amount of time that a data burst lags behind the receipt of the corresponding BHP. A time-out on Timer B indicates the transmission failure of a data burst. Likewise, a Timer C at the intermediate node monitors the channel holding time reserved for a data burst. In the assumed system scenario, the Switching Control Unit (SCU) at the intermediate node is responsible to release–when necessary–the bandwidth which has been reserved for a data burst. The bandwidth release operation is triggered by either a time-out event from Timer C, or a particular message which explicitly requires such an operation.

The OBS system under consideration adopts a void-filling (VF) strategy for data channel scheduling [51]. The basic principle of the VF strategy is that the interval between two previously scheduled periods of resources can be used to transmit the traffic which arrives later, thus filling the void. The void-filling method facilitates flexible utilization of the network resources.

Furthermore, each intermediate node of the core network is identified by an index i, where $i = 1, \ldots, n$, and n represents the total number of the intermediate nodes in the core network.

The following notations will be utilized in the BEFRR description ($i = 1, \ldots, n$):

- T_a : The time when a new burst traffic begins to assemble at the ingress node,
- $T_h(i)$: The time when a BHP is received at the i-th intermediate node. $T_h(0)$ represents the time when the BHP is transmitted into the the core network at the ingress node,
- $T_c(i)$: The time when the i-th intermediate node receives a signaling message requiring the release of the pre-reserved resources. $T_c(0)$ represents the time when the actually assembled data burst length exceeds the reservation value contained in a pre-transmitted BHP.

- $\vartheta(i)$: The time interval for the SCU of the i-th intermediate node to process a BHP. It is assumed that $\vartheta(i) = \vartheta$ for all $i \in \{1 \ldots n\}$;
- $\theta(i)$: The time interval for the switching matrix configuration at the i-th intermediate node to become stable. It is assumed that $\theta(i) = \theta$ for all $i \in \{1 \ldots n\}$;
- $\tau_o(i)$: The offset between a BHP and its data payload at the output port of the i-th intermediate node. $\tau_o(0)$ represents the initial offset between a BHP and its data payload at the ingress node. $\tau_o(i) = \tau_o(i-1) - \vartheta$;
- $T_s(i)$: The starting time when the resource at the i-th intermediate node is reserved for a data burst ($T_s(i) = T_h(i) + \vartheta(i) + \tau_o(i)$);
- $T_e(i)$: The ending time when the resource at the i-th intermediate node is reserved for a data burst ($T_e(i) = T_s(i) + L_r$).

The design of the BEFRR scheme is guided by the following considerations:

1. The bandwidth wastage of an FRR system, especially that due to the unsuccessful forward resource reservations, is minimized;
2. No extra end-to-end burst delay is induced;
3. The operation cost of the intermediate node, such as that for the lightpath tear-down and setup, and that for the switching matrix re-configurations, is maintained as low as possible.

The essence of the proposed bandwidth enhancement mechanism is to adopt a crank-back procedure at the intermediate nodes to release the pre-reserved resources which are insufficient to support the corresponding data burst. In order to maintain the BHP pre-transmission success probability, thus satisfying the latency reduction requirement, the BEFRR scheme still employs the aggressive strategy for resource reservation as the basic FRR scheme does. Likewise, the delayed-reservation is utilized to improve the network throughput.

5.3.2 The BEFRR Scheme Principle

The proposed bandwidth enhancement mechanism involves both the edge node behavior and the intermediate node behavior. To emphasize the bandwidth enhancement functionality, we present the BEFRR principle by describing its distinctive characteristics as compared to the basic FRR scheme:

1. Instead of comparing the data burst length with the reservation length carried in a pre-transmitted BHP until the burst assembly is completed, the BCU in the BEFRR scheme begins to monitor the actually assembled burst amount immediately after the BHP is sent out at $T_h(0)$.

   If by the time $T_a + \tau_a$, the actual burst length does not exceed the reservation value contained in the pre-transmitted BHP, the forward resource reservation succeeds and the data burst is transmitted without additional action to be taken.

   Otherwise, the following steps should be executed.

2. As soon as the actual burst length exceeds the pre-reservation length at some time $T_c(0)$, where $T_h(0) < T_c(0) \leq T_a + \tau_a$, the BCU issues a signaling message, namely, a CLEANUP message, to nullify the pre-reservation requirement (i.e., the pre-transmitted BHP). The CLEANUP message carries the identifier of the BHP which it attempts to invalidate.

3. At the i-th intermediate node, upon the reception of the CLEANUP message at $T_c(i)$, the SCU promptly triggers a crank-back procedure. That is, the SCU releases the pre-reserved resources for the corresponding data burst, and makes this period available to other burst transmissions. Simultaneously, the CLEANUP message is forwarded to the next intermediate node, until all nodes that have reserved resources for the corresponding data burst are notified.

   If by the time $T_c(i)$, the switching matrix has been configured for the corre-

sponding data burst, i.e., $T_s(i) - \theta \leq T_c(i) \leq T_s(i)$, the switching matrix should be released immediately.

Figure 5.5 illustrates the difference between the basic FRR scheme and the BEFRR scheme at the i-th intermediate node. For simplicity, only the circumstance when a BHP pre-transmission fails and the crank-back procedure occurs is presented.

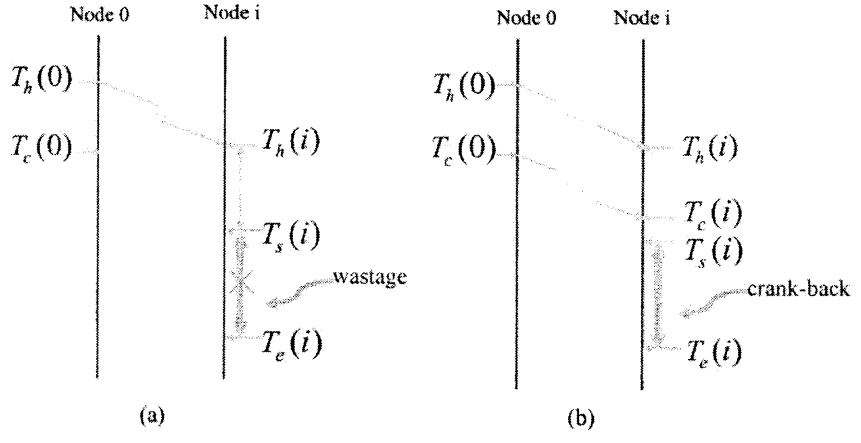

Figure 5.5 The comparison between the basic FRR scheme and the BEFRR scheme (a) in the basic FRR scheme, nothing is done with the insufficient pre-reserved resources; (b) in the BEFRR scheme, a crank-back procedure is employed at the intermediate node to release the pre-reserved resources.

Reservation clean-up is an essential feature of the BEFRR scheme to reduce the potential bandwidth wastage due to the insufficient pre-reserved resources. This procedure, in tandem with the VF-scheduling method, enables the intermediate node to make intelligent usage of the available network resources, and improves the system throughput.

Another important feature of the BEFRR scheme is that the delayed-reservation is adopted and the switching matrix is configured in a just-in-time manner, i.e., the lightpath at the intermediate node is not configured for a reservation until $T_s(i) - \theta$. This characteristic enables the CLEANUP message, which satisfies $T_h(i) + \vartheta \leq T_c(i) <$

$T_s(i) - \theta$, to not only reduce the bandwidth wastage, but also avoid the unnecessary operations for lightpath set-up and tear-down.

The benefits of the BEFRR scheme is facilitated by the message dialog between the ingress nodes and the intermediate nodes. The particular message, i.e., the CLEANUP message, is thus employed to make the intermediate nodes aware of the invalidity of the pre-transmitted reservation requirement.

Comparing with the basic FRR scheme, the system cost of the BEFRR scheme is induced by the extra signaling transmissions, and is equal to $O(m)$, where m is the number of the CLEANUP messages to be transmitted. Taking into account of the reduced switching matrix operation and the improved bandwidth usage efficiency, together with the fact that the BHP pre-transmission failure probability is typically small (e.g., less than 5% in a steady system), the benefits of the BEFRR scheme are more considerable.

5.3.3 Theoretical Analysis

The objective of the bandwidth enhancement mechanism is to reduce the potential bandwidth wastage caused by insufficient forward resource reservations in the basic FRR scheme. Therefore, the interesting performance figure of merit is the bandwidth savings for a given burst length L_d, and the associated signaling overhead, which is defined as the possibility to transmit the CLEANUP message. For simplicity, the effect of θ is considered to be negligible as compared to the length of the data burst.

The BEFRR scheme provides bandwidth savings when the pre-reserved resources are insufficient to support the actually assembled data burst. As empirically demonstrated in the previous chapters, the prediction residuals delivered by the underlying LPF is approximately Gaussian distributed with mean L_d and variance σ^2, and the probability distribution function of the prediction value ($\tilde{L}_d$) corresponding to a given L_d is expressed by Equation 4.4. Consequently, for a given burst length

L_d, the average bandwidth saving (L_s) is

$$\begin{aligned} L_s &= \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\int_{-\infty}^{L_d-\delta}(x+\delta)\cdot e^{-\frac{(x-L_d)^2}{2\cdot\sigma^2}}dx \\ &= -e^{-\frac{\delta^2}{2\cdot\sigma^2}}\cdot\frac{\sigma}{\sqrt{2\cdot\pi}}+(L_d+\delta)\cdot Q(\frac{\delta}{\sigma}). \end{aligned} \tag{5.17}$$

As a CLEANUP message is transmitted as soon as the actual burst length exceeds the reservation value contained in the pre-transmitted BHP, the associated signaling overhead, denoted as S_o, can be expressed as the probability that the forward resource reservation fails, i.e.,

$$\begin{aligned} S_o &= P(L_r=\tilde{L}_d+\delta<L_d) \\ &= \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\int_{-\infty}^{L_d-\delta}e^{-\frac{(x-L_d)^2}{2\cdot\sigma^2}}dx \\ &= Q(\frac{\delta}{\sigma}). \end{aligned} \tag{5.18}$$

Equation 5.17 and Equation 5.18 represent the upper bounds of the bandwidth savings and the signaling overhead associated with the BEFRR scheme, respectively.

Tighter bounds with respect to L_s and S_o are derivable for the BEFRR scheme, if a more practical and stricter situation is considered that the reservation requirement carried by a pre-transmitted BHP should be no less than 0. In this case, the average bandwidth wastage corresponding to a given data burst of length L_d is

$$\begin{aligned} L_s &= \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\int_{0}^{L_d-\delta}(x+\delta)\cdot e^{-\frac{(x-L_d)^2}{2\cdot\sigma^2}}dx \\ &= \frac{\sigma}{\sqrt{2\cdot\pi}}\cdot(e^{-\frac{L_d^2}{2\cdot\sigma^2}}-e^{-\frac{\delta^2}{2\cdot\sigma^2}})+ \\ &\quad (L_d+\delta)\cdot[Q(\frac{\delta}{\sigma})-Q(\frac{L_d}{\sigma})], \end{aligned} \tag{5.19}$$

and the potential signaling overhead is given by

$$\begin{aligned} S_o &= P(L_r=\tilde{L}_d+\delta<L_d) \\ &= \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\int_{0}^{L_d-\delta}e^{-\frac{(x-L_d)^2}{2\cdot\sigma^2}}dx \end{aligned}$$

$$= Q(\frac{\delta}{\sigma}) - Q(\frac{L_d}{\sigma}). \quad (5.20)$$

5.3.4 Simulation Results and Discussion

Simulations are conducted to examine the advantages of the BEFRR scheme as compared to the basic FRR scheme. Similar to the last section, a 12-order LPF is utilized for data burst length prediction, and the traffic flowing into the ingress node is assumed to be a self-similarity process, generated based on the FFT-FGN model, with the Hurst parameter of $H = 0.75$ and the average packet size of 2000 bytes. In the rest of the chapter, the burst assembly time (τ_a) will be normalized with respect to the average time to transmit one IP packet of 1500 bytes, and will simply be referred to as τ.

The proposed BEFRR scheme improves the bandwidth usage efficiency (ω) as compared to the basic FRR scheme, and the improvement is especially significant when the correction value (δ) is small, whereby the BHP pre-transmission fails at a higher probability (Figure 5.6). It is interesting to see that for any given burst assembly duration τ, the BEFRR scheme delivers similar bandwidth usage efficiency as that in the basic FRR scheme after the optimal correction value (i.e., the correction threshold that delivers the maximum ω for the basic FRR scheme) is reached. This implies that when δ is large, only a small fraction of the total bandwidth overhead is caused by the insufficient forward resource reservation, while the major part is induced by the aggressive reservation strategy. Therefore, the improvement of the BEFRR scheme is not significant in this case.

The above conclusions also hold in Figure 5.7, which plots the relationship between the bandwidth usage efficiency and the BHP pre-transmission success probability (P_s). As expected, both BEFRR scheme and the basic FRR scheme yield similar ω as P_s approaches 100%. Note that although Figure 5.6 and Figure 5.7 present only the simulation results where the SPD-based algorithm is utilized as the underlying

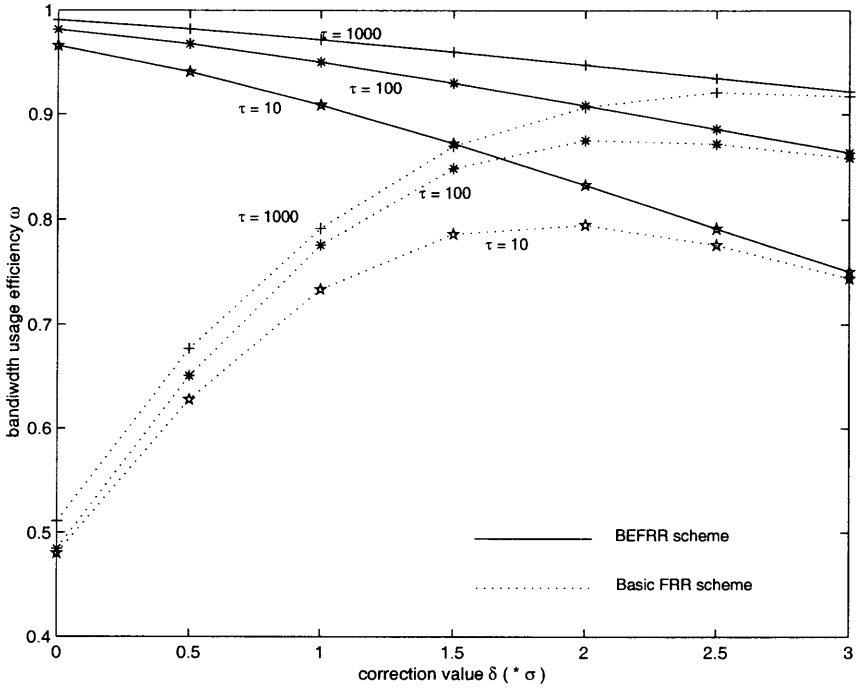

Figure 5.6 The bandwidth usage efficiency versus the correction value.

aggressive reservation strategy, the same conclusions hold for the system which adopts the BUD-based algorithm.

Another advantage of the BEFRR scheme is that it enables the aggressive strategy to perform more straightforward control on the bandwidth usage efficiency (Figure 5.8 and Figure 5.9). Figure 5.8 presents the bandwidth usage efficiency versus the burst assembly time when the BUD algorithm is utilized to implement the aggressive reservation strategy. As observed, as long as ρ ($\rho = \phi$) is specified, the BEFRR scheme achieves almost constant bandwidth usage efficiency as the expected value ($\omega = 1 - \rho$). That is, the BEFRR scheme enables the BUD algorithm to be more independent of τ. Such benefit is more significant for smaller ρ, with which the BHP pre-transmission fails at a higher probability. Although the independence of the bandwidth usage efficiency to the burst assembly time does not hold for the SPD-based aggressive reservation algorithm, the BEFRR scheme also benefits the SPD algorithm by making ω to be more proportional to the correction value δ (Figure

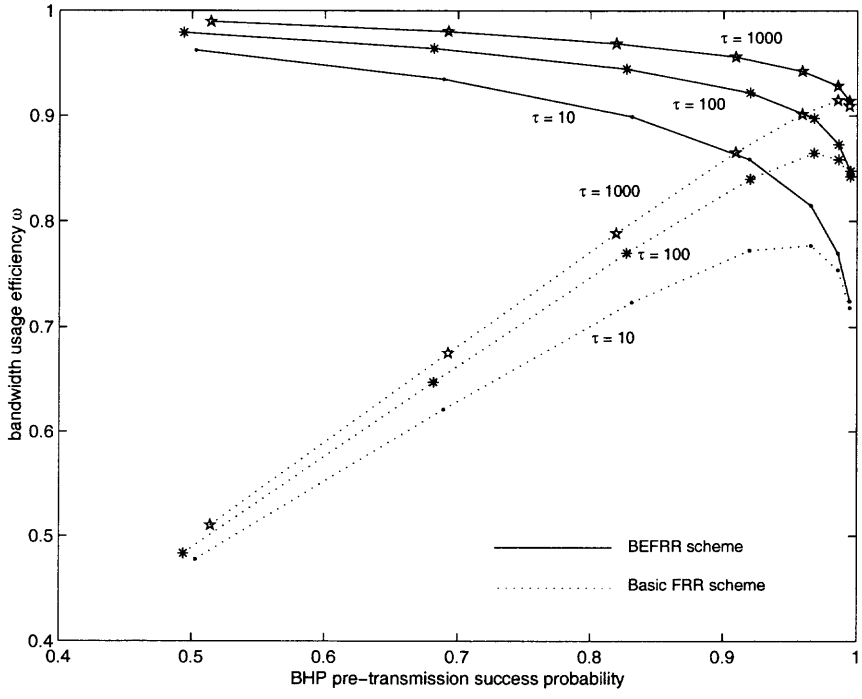

Figure 5.7 The bandwidth usage efficiency versus the BHP pre-transmission success probability.

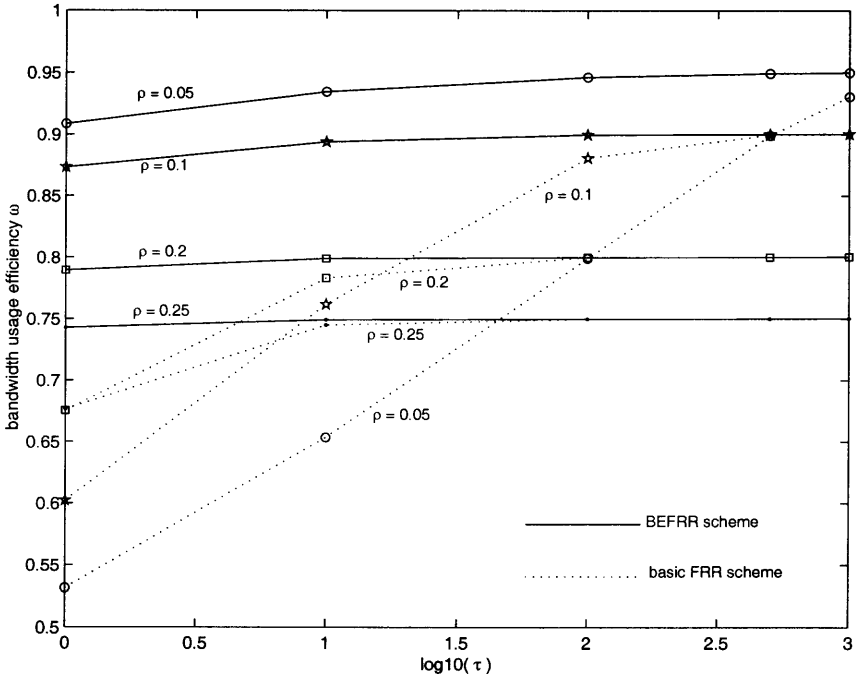

Figure 5.8 The bandwidth usage efficiency versus the burstification duration.

5.9). For example, in the basic FRR scheme, when $\tau = 100$, the bandwidth usage efficiency delivered by $\alpha = 2$ is higher than that both by $\alpha = 1.5$ and $\alpha = 3$ (Note that $\delta = \alpha \cdot \sigma$). On the contrary, in the BEFRR scheme, a larger α value (i.e., the larger correction value δ) results in the higher ω value. The curves in both figures reinforce

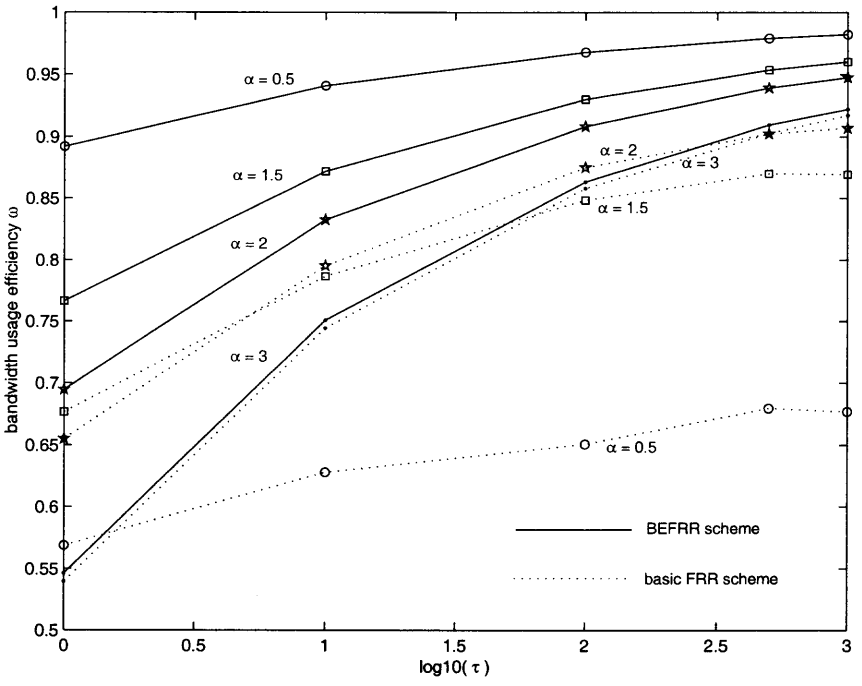

Figure 5.9 The bandwidth usage efficiency versus the burstification duration.

the aforementioned conclusion that the bandwidth enhancement capability of the BEFRR scheme is reduced as δ increases, and that the BEFRR scheme presents only marginal bandwidth enhancement capability when the compensation ratio is large enough, Therefore, when the correction value is relatively large, a proper planning mechanism for δ plays the more important role in the system bandwidth utilization.

5.4 Summary

This chapter studies the performance improvement issues of the FRR-enabled OBS system. Two solutions have been proposed, based on the channel holding time adjustment and bandwidth enhancement mechanism, respectively. Theoretical

analysis and simulations results demonstrate the advantages of the proposed solutions, and provide a guideline for the network designer to build up a network with the desired performance measurements.

The following conclusions can be drawn:

1. By adapting the channel holding time adjustment according to the changes of the burstification interval, the aggressive strategy-based FRR scheme can achieve the optimal performance of different objectives. Meanwhile, the effect of the channel holding time adjustment on the network performance is also influenced by the input traffic parameters.

2. By allowing a CLEANUP message to nullify the pre-reservation requirement and performing the switching matrix configuration just-in-time, the proposed BEFRR scheme yields better system performance. The BEFRR scheme reduces the bandwidth wastage due to the insufficient forward reservations at a low cost of signaling overhead, and enables the correction value to perform more straightforward control on the bandwidth usage efficiency. The benefit of the bandwidth enhancement mechanism is more significant when the correction value is small (whereby the BHP pre-transmission fails with a higher probability)

The performance improvement issue for the basic FRR scheme is an on-going research. There are still some work to be carried on to improve validate and implement the proposed algorithms and signaling scheme.

CHAPTER 6

CONTROL ARCHITECTURE AND ENABLING TECHNOLOGIES FOR ETHERNET-SUPPORTER IP OVER WDM MANS

This chapter is focused on the control architecture and the enabling technologies for Ethernet-supported IP-over-WDM metropolitan area networks. The general architecture of an access node in such networks will be presented, and solutions to facilitate the essential system functionalities will be proposed. The aim is to render the flexible and high capacity metropolitan network which provides service provisioning improvement and resource utilization efficiency for the data-dominated traffic. Specifically, an enhanced address resolution protocol is proposed to reduce the call setup latency and the signaling overhead associated with the address probing procedure; a burst-based transmission mechanism is adopted to improve the network throughput and resource utilization efficiency; and a hop-based wavelength allocation algorithm is investigated to provide flexible bandwidth multiplexing with fairness and high scalability.

6.1 Motivation

As stated in Chapter 1, the Ethernet-supported IP-over-WDM ring paradigm provides a feasible solution for the new generation metropolitan optical network, enabling a graceful migration from the current voice-oriented MAN prototype into a world optimized for packets. The unprecedented bandwidth supportability of WDM technology, in tandem with the packet-oriented Ethernet prototype, facilitates a common shared infrastructure, thus making a new generation of optical MAN optimized for scalable, survivable, and IP-dominated networks at gigabit speeds possible.

To unleash the potential of the packet-based WDM metropolitan ring network, efficient Media Access Control (MAC) protocols are needed to coordinate the system resources, in addition to the system infrastructure considerations. This is especially true when the data packet processing in the optical domain is still not yet mature. Ethernet, while a natural fit for data traffic, has evolved to support full duplex-switched infrastructures but lacks the flexible MAC mechanisms to manage the access across multiple users in the WDM ring prototype. The foregoing challenges require innovative protocols and algorithms in the metropolitan environment that retain all the advantages of the packet-based transport mechanism while rendering elastic bandwidth allocation and graded levels of services.

Intensive research endeavors have been devoted in the design and implementation of the metropolitan optical network ([30, 31, 52, 53, 54, 55, 56, 57, 58] and references therein). The efforts have been focused on either the network architectures and configurations, or the MAC protocols and the service provisioning mechanisms. For example, E. Hernandez- Valencia [31] presented a hybrid architecture consisting of Ethernet/TDM service solutions to enable storage networking and Ethernet transport over SONET/SDH networks. N. Madamopoulos et al. [53] investigated the impact of different add-drop modules implementations on the system performance. J. Cai et al. [55] proposed the MultiToken Interarrival Time (MTIT) protocol to provide the efficient bandwidth multiplexing for the WDM ring architecture. Alternative MAC protocols focusing on the fairness, scalability, and various service provisioning have also been reported.

This chapter designs the access nodes of the Ethernet-supported IP-over-WDM metropolitan network, and devises the enabling technologies combining the space, time, and channel domains to systematically facilitate the new network infrastructure. Specifically, a set of mechanisms comprised of the address resolution, the traffic engineering, and the wavelength allocation has been proposed to render

packet-optimized optical MAN with the transport performance betterment (e.g., reduced transport delay), the resource utilization efficiency (e.g., improved network throughput), and fairness and scalability for the network resource access control.

6.2 System Environment and Problem Statement

This section details the network environment upon which our investigation will be conducted. The general control architecture of an access node will be presented, followed by the design objectives of our proposed technologies and algorithms.

6.2.1 System Environment

In this chapter, a ring-shaped metropolitan network is considered, where N access nodes (ANs) are interconnected via counter-rotating dual fibers (i.e., the feeder ring) [59]. The fiber ring consists of the inner ringlet and the outer ringlet, each of which makes use of the full bandwidth of the fiber, i.e., the individual wavelength can be transported concurrently in both ringlets, assuming that each access node has adequate receiver capabilities (see Figure 6.1). Each fiber supports $W + 1$ wavelengths as parallel channels, of which W wavelengths (λ_1, ..., λ_W) are for data channels and one for the control channel (λ_0). If necessary, the network capacity can be gradually updated by parallel fibers. The aggregated bandwidth can scale to multi-terabits/second.

Vast research efforts have been focused on the bandwidth multiplexing and channel access control issues for the packet-supported WDM networks. The majority of the approaches centers on the WDM layer. The implementation complexity, cost, and performance have impact on the network design. Interested readers are referred to [54, 55, 56, 57, 58] for detailed discussions and to [13] for an overview. In this chapter, the efficient bandwidth sharing of the optical fiber is achieved from the perspectives of the signaling transmission and the space reuse of wavelengths.

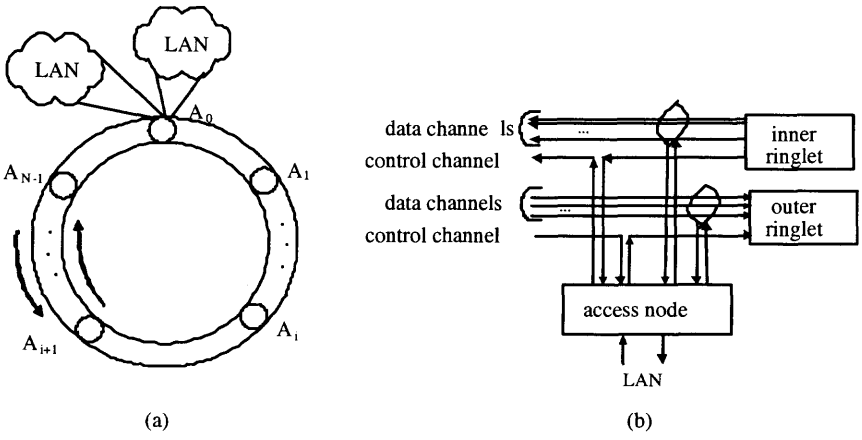

Figure 6.1 The prototype of a ring-based metropolitan optical network. (a) The dual-fiber ring; (b) The access node connecting the feeder ring and the LAN.

Access to the network resources (e.g., the data transmission channel) is typically based on two alternative schemes: pre-allocation based protocols and reservation based protocols. While the former technique assigns transmission rights to different nodes in a static and pre-determined manner, the reservation-based technique arbitrates the bandwidth access to the traffic demand in a real-time fashion, i.e., the resource reservation request is delivered throughout the ring layout when a data transmission is required. In the assumed system scenario, the reservation-based method is adopted, as it yields flexible bandwidth utilization and is a natural fit for the dynamic data traffic. The data payload is launched into the network after the corresponding reservation request is confirmed success.

Two or more data transmissions on the same wavelength along the same section of the fiber result in a collision. Depending on the approach the network resource contention is addressed, signaling protocols discussed in the literature can be classified into two main categories: 1) collision-free strategies, and 2) collision-and-retry strategies. These variants result in different complexity of network hardware requirements, and different optical bandwidth multiplexing efficiency. In this chapter,

the control packet transmission is decoupled from that of the data payload. While a control packet reserves the resources according to the collision-and-retry strategy, the data payload is guaranteed with a single-hop transmission without any delay or loss in the feeder ring.

In addition, the destination-stripping method is employed to extract the data traffic from the ring network. This method, together with the wavelength reuse of the data transmission (owing to the wavelength allocation algorithm as described in the next section), enables the concurrent data channel usage in disjoint source-destination pairs, yielding an improved degree of bandwidth multiplexing and resource utilization efficiency within the ring.

One or more Gigabit Ethernet (GbE) Local Area Network (LAN) is connected to the feeder ring via the access nodes. The LAN is typically composed of enterprises or high-speed end users. Hereafter, the $j-th$ router in the LAN attached to the access node A_i ($i \in \{ 0, ..., N-1 \}$) will be referred to as A_iR_j, assuming that $j \in \{0, ..., M-1\}$, where M is the total number of routers in the attached LANs. Besides the connectivity functionality, the AN also provides MAC solutions which are necessary to render efficient packet-oriented traffic processing and transmission at the WDM layer. The detailed architecture and functionality of the access node will be described in the next subsection. Throughout this chapter, the traffic flowing from an LAN to the metropolitan ring will be referred to as the upstream traffic, while that from the metropolitan ring to the LAN as the downstream traffic.

6.2.2 Access Node Architecture

Each access node has two interfaces: The gigabit Ethernet interface which is used for the packet processing and the interaction with the associated local access networks, and the optical link interface to access the WDM ring in the optical domain.

Figure 6.2 illustrates the functional architecture of an access node, which mainly consists of the transmitter logic, the receiver logic, and the control logic.

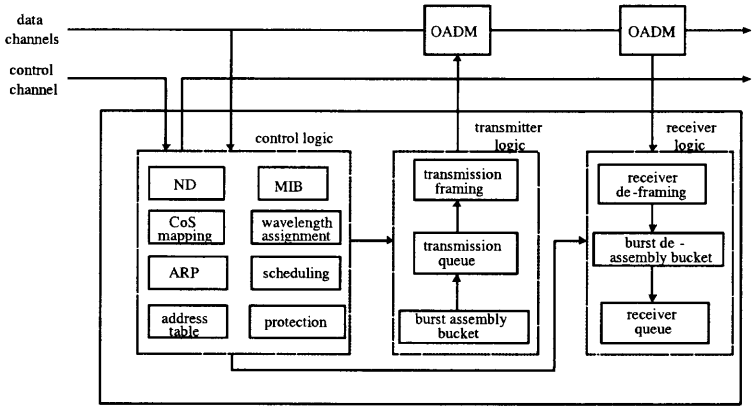

Figure 6.2 The functional architecture of an access node.

The main function of the transmitter logic is to adapt the low bit-rate tributaries (up to OC-12) of the local network to the transmission granularity suitable for WDM transport media, and to forward the traffic to the destination node. In the proposed system, the transmitter logic assembles the upstream data packets into the larger granularity, namely, a data burst. The traffic assembly mechanism is similar to that proposed for the backbone [15, 13], and is tailored for the Ethernet-supported WDM ring topology, as will be discussed in the next Section.

The transmitters emit the assembled traffic into the metropolitan network. A transmitter may be either fully tunable to all data channels, or it may be partially tunable to more than one but fewer than W wavelengths [57], depending on the preference of the network management. In the proposed system, the number of tunable transmitters is impacted by the wavelength allocation algorithm (as will be discussed in the next section). Multiple transmitters of an access node can transmit on different wavelengths concurrently with negligible tuning latency [60, 61], enabling the better exploit of the parallel transmission capability of the WDM technology. The

transmitter queues are provided to accommodate the traffic waiting for the access to the data channels.

The receiver logic is similar to the transmitter logic. Each node is equipped with tunable receivers, which can be fully tunable or partially tunable just like the transmitters. The downstream traffic is disassembled into packets before being transported to the individual host in the LAN. The receiver queue is provided to order the processing of the traffic according to their Class of Service (CoS).

The control logic coordinates the traffic transmission and processing, and facilitates the traffic engineering functionalities. Of particular interests in this chapter are the ARP module, the burst assembly control module, and the wavelength allocation module.

The ARP module is designed to support the address probing mechanism proposed in this paper. Each access node is embodied with two tables: the local address mapping (LAM) table registering the <protocol type,protocol address, physical address> triplet of all the routers in the subordinated LANs, and the remote address mapping (RAM) table recording the address translation results obtained from the executed address inquiry procedures.

The traffic assembly module controls the traffic assembly/disassembly procedure. To enable the connectivity services with graded levels of performance, the CoS mapping function should be incorporated in this module to map the incoming traffic flow into a specific transport class.

The wavelength allocation module, in conjunction with the scheduling module and the traffic selection module, enables the coordination between the control packet and the data payload, as well as resolves resource contentions. The implementation of such control functionalities requires the Management Information Base (MIB) and other functional components to keep track of the network status, such as the data

channels availability (full or empty) and the configuration status of the data channel connections.

Figure 6.2 also highlights some functional modules which are necessary for the multi-facet network infrastructure. For example, the node discoverer module maintains the topology information of all the nodes in the ring, and monitors the fiber cut or node failure events by periodically sending sub-wavelength signals to its neighboring nodes, both of which are essential to deploy the protection/restoration strategies. However, the implementation of such components is beyond the scope of this chapter.

The designed access node architecture has the following properties.

- It facilitates the Ethernet-supported IP-over-WDM integration to support the dynamic data traffic, thus overcoming the inefficiency and inflexibility of the current SONET-dominated infrastructure and its circuit-based provisioning model.

- It allows the deployment of mechanisms in both the Ethernet and WDM perspectives, thus enabling the better synergy of both mature electronic protocols and advanced optical technologies.

- It is flexible and cost effective. The individual functional component is a skeleton based on which a variety of algorithms and devices can be adopted and developed according to the preference of the network management. Meanwhile, the service provisioning requires no SONET overhead bytes and synchronization among access nodes, nor the dedicated protection bandwidth.

Theoretical analysis and simulation results will demonstrate that such an access node architecture, in tandem with the proposed enabling technologies, yields improved network performance with reduced system cost.

6.2.3 Problem Statement

Based on the aforementioned service requirements and network architecture, the design objectives of the proposed algorithms and technologies can be summarized as follows.

- Improve the service provisioning for the application streams in terms of the transport latency, including both the call setup delay and the transmission delay in the metropolitan network.
- Improve the system efficiency in terms of the resource utilization, the network throughput, and the signaling transmission requirements.
- Provide fairness and scalability control among the traffic between the access nodes, as well as differentiated classes of services for the multi-type applications.

In the proposed system, these requirements are implemented with mechanisms developed at both the data link layer and the medium layer, based on the presented architecture.

6.3 The Enabling Technologies

This section speculates the proposed enabling technologies and algorithms, and discusses their impacts on the Ethernet-supported IP-over-WDM metropolitan network. Notations are defined in Table 6.1 to simplify our description.

6.3.1 The Enhanced Address Resolution Protocol (E-ARP)

Typically, an Ethernet-supported network employs the address resolution protocol (ARP) [62] to translate the network layer address (i.e., the IP protocol address) into the link layer one (i.e., the hardware address). While very simple and well-suited to the LAN-hardened Ethernet (which is broadcast in nature), the original ARP cripples the address probing procedures in the metropolitan optical network.

Table 6.1 Notations for he proposed enabling technologies ($i \in 1, ..., N$, $l \in 1, ..., \infty$, $k \in 1, ..., H_{max}$)

Term	*Explanation*
pro	Protocol type of the data packet.
tha	Hardware address of the target of this packet.
tpa	Protocol address of the target of this packet.
sha	Hardware address of the sender of this packet.
spa	Protocol address of the sender of this packet.
t_d^a	The time when a data burst begins to assembly at the access node.
t_d^s	The time when an upstream data burst is sent into the metropolitan network.
$t_c^s(l)$	The time when a control packet is sent into the metropolitan network for the $l-th$ reservation attempt.
t_c^r	The time when an access node receives a control packet.
$t_r(l)$	The random time when a control packet is delayed at the source access node after the $l-th$ reservation attempt fails.
τ_a	The burst assembly duration.
T_w	The data burst delay owing to signaling transmissions.
T_c	The round-trip time for a control packet to be processed in the MAN.
R	The maximum number of resource reservation attempts before a data burst is dropped.
t_e	The time the specific data channel will be available at the destination access node.
h_{max}	The maximum number of hops a data burst is transported in the ring.
H	The total number of hops to support concurrent transmissions between any pair of access nodes.
h_k	The number of hops for a transmission to propagate from the source to the destination node.
C_t	The number of data channels required to support the concurrent transport between any pair of access nodes.
C	The total number of data channels available in the system.
S_k	The $k-th$ data channel subset shared by the traffic which traverses k hops.
$\|S_k\|$	The number of data channels in the data channel subset S_k.
ω_k^j	The $j-th$ data channel in the subset S_k.
$\omega_k^{pref}(A_i)$	The most preferred data channel for node A_i to transport the traffic requiring k hops.

Consider routers A_1R_1 and A_1R_2, both of which need to communicate to router A_2R_2. Using a conventional ARP, both senders execute individual call setup procedures on the overlapped route from A_1 to A_2, and A_2 to A_2R_2 (see Figure 6.3), resulting in the longer call setup delay and the higher consumption of network resources. A savvy address inquiry function is highly desirable for the new network scenario.

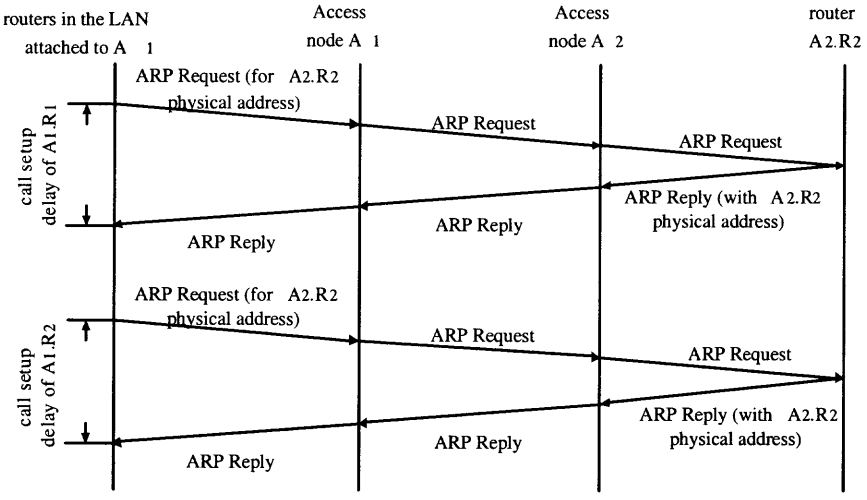

Figure 6.3 Packet flow in the ARP mechanism.

This chapter proposes an enhanced ARP, called E-ARP, to reduce the call setup latency and the gratuitous ARP packet transmissions. The basic idea is two-folded: the address translation is rendered as early as possible, and the address mapping obtained from previous ARP transmissions is retrievable for the subsequent inquiries. The distinctive characteristics of the E-ARP as compared to the conventional ARP can be described as follows:

- The access nodes incorporate the address translation function with the packet forwarding function. Upon the reception of an upstream ARP request (i.e., an ARP packet destined for a router in a remote LAN), the access node either directly replies it, or broadcasts it in the MAN, depending on whether or not the access node can find the required address mapping information in the RAM

table. Meanwhile, the access node updates - when necessary - the hardwire address field of the sender's entry in the LAM table with the information in the packet, or adds a new entry if the sender does not exist in the LAM. A downstream ARP request (i.e., an ARP packet broadcasted in the metropolitan ring) is received by all access nodes, and is replied only by the access node which connects the destination router to the MAN (i.e., the access node whose LAM table has the entry indexed by the destination protocol address).

- The access node re-edits the upstream ARP request before broadcasting it into the metropolitan ring, replacing the sender hardware address (*sha*) field with its own hardware address. The access node is also responsible to generate the uni-cast ARP reply packet. Note that the traffic is exchanged in the MAN ring according to the hardware address of the access node.

- The address mapping information obtained from the ARP reply packet is registered in the RAM table of the access node, and is retrievable by the subsequent address translation requests which are sent by other local routers. The redundant ARP transport on the metropolitan ring is thus avoided until the RAM table ages out the entry corresponding to the remote router.

Figure 6.4 depicts the ARP packet flow when the E-ARP is adopted, given the same address resolution requests as that in Figure 6.3. The detailed E-ARP procedure is shown in Figure 6.5. Note that the enhanced ARP functionality at access nodes also supports the basic address information management [62], e.g., the address update and address age-out. Such functionalities are omitted in Figure 6.5 for simplicity.

The proposed E-ARP mechanism features several advantages. First, the RAM table in the access node facilitates the information reuse of address inquiries, thus prompting the call setup procedure and reducing the unnecessary ARP packet transmissions in the metropolitan optical network. Second, the LAM table provides the

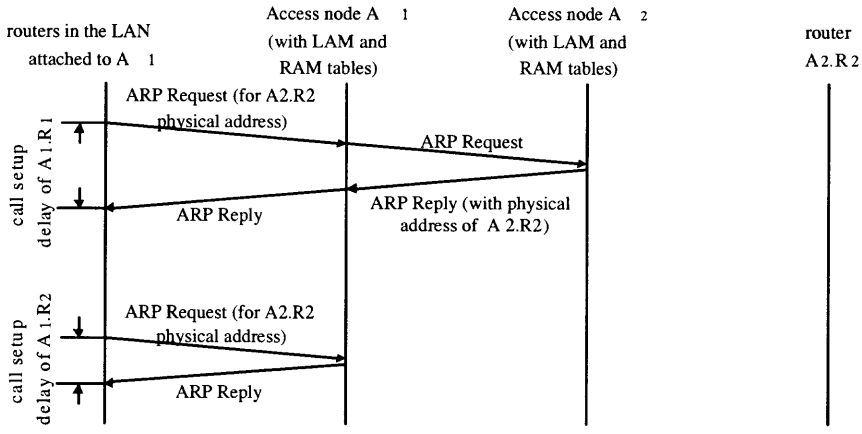

Figure 6.4 Packet flow in the E-ARP mechanism.

If the triplet < *pro, tpa, tha* > exists in the RAM table,
Returns the *tha* to the sender
else
{Re-writes the ARP request packet by replacing *spa* and *sha* fields
with the respective values of the current access node,
Forward the re-written ARP request packet on the WDM ring}

(a)

If the *tpa* carried in the ARP request packet is not in my LAM table,
Discard the ARP request packet
else
{Add the triplet < *pro, spa, sha* > to the RAM table
Swap *spa* and *sha* in the ARP request packet with its *tpa* and *tha*, respectively.
Putting the hardware address of the current access node in the *sha* field;
Mark the ARP packet with " *REPLY* ";
Send the ARP reply packet to the ring.
}

(b)

Figure 6.5 The pseudo-code for the E-ARP mechanism. (a) An access node receives an upstream ARP request packet. (b) An access node receives a downstream ARP request packet.

access node with the address information of its local routers. Therefore, an access node can reply an ARP request without going further into the LAN. Third, the E-ARP maps the protocol address of a router to the hardware address of its associated access node. This way, data packets are transported in the MAN according to the hardware address of the access node, and are delivered to the ultimate routers by the local access node. This solution ushers in the decoupling of the traffic transmission in the WDM domain from that in the Ethernet domain, which is consistent with the line of thought to operating in the individual network independently. Meanwhile, addressing the traffic according to the access node also benefits our system with reduced complexity for traffic management and traffic engineering (e.g., the traffic assembly procedure which will be explained in the next subsection). Another significant merit of our E-ARP is the performance improvement achieved without requiring new protocols. The address resolution enhancement is implemented by simply equipping the access node with the RAM and the LAM tables.

6.3.2 The Burst-based Transmission Mechanism

After the address translation procedure, the subsequent data traffic can be transported according to the obtained hardware address. The traffic transmission mechanism plays an important role in architecting the efficient WDM-enabled integration. Motivated by the mismatch between the transmission capacity of WDM fibers and the fine traffic granularity of data packets, the proposed system adopts a modified burst-based transmission mechanism to improve the system throughput and reduce the signaling overhead. The transmission mechanism is described in terms of the traffic assembly procedure, the signaling scheme, and the burstbased transmission benefits.

The burst assembly procedure is based on the one proposed in [13, 18], which has proved to be an efficient paradigm in the long-haul network. Different from their

proposal, however, in the proposed system, data packets are assembled at the access node according to the Ethernet address of the destination access node, a mechanism enabled by the proposed E-ARP. The individual access node is equipped with $N-1$ assembly units, each corresponding to one destination access node on the ring topology. To enable differentiated services for the incoming traffic, each unit has one or more assembly buckets corresponding to different CoS requirements. Broadcast service can be easily supported by the burst assembly mechanism in alternative ways: either add one specific assembly unit at each access node to assemble the broadcast packets so that a single copy of each packet is sent throughout the metropolitan network, or $N-1$ copies of the individual broadcast packet is replicated, each inserted in one of the $N-1$ assembly buckets.

The burst assembly interval (τ_a) is subject to the constraints imposed by the round-trip time of the control packet (i.e., the cycle latency for a control packet to travel throughout the WDM ring, denoted as T_c) and that by a pre-determined maximum burst length. In the metropolitan network adopting the reservation-based channel access mechanism, τ_a should be no less than T_c to upperbound the traffic generation rate by the traffic processing rate. By limiting the maximum burst length with a certain threshold, a simple fairness control is rendered to avoid one transmission occupying a certain channel for an excessive long time and potentially starving the other transmission requirements competing for the same channel or destination node.

The signaling protocol in the proposed burst-based transmission mechanism distinguishes from its counterpart in the backbone in the two-way signaling scheme. When the data burst is fully assembled, the associated control packet is generated and transported into the network. After being processed at each access node and attempting to reserve resources at the destination node, the control packet returns to the source node with a reservation acknowledgement. Depending on whether the corresponding reservation succeeds or fails, the data payload is either emitted into

the ring network, or is delayed at the transmission queue. Padding may be required if a minimum burst length is imposed [13]. The unsuccessful control packet will be re-transmitted after a random time at the source access node until the maximum retransmission attempts (R) is performed, or the reservation succeeds, with the random delay between two consecutive re-transmissions uniformly distributed from 0 to T_c. A data burst is discarded when its control packet fails to reserve the resource after R attempts. R should be properly engineered based on the tradeoff consideration between the traffic drop probability and the delay allowance.

When the fast service provisioning is of essence to the network management, our signaling protocol incorporates the parallel execution of the burst assembly procedure and the resource reservation, whereby the control packet transmission is triggered as soon as a burst begins assembling. The concurrent execution of both delay contributors reduces the inherent artificial delay of data traffic at the access node.

Maintaining the advantages of the typical burst-based transmission mechanism, the aforementioned two-way signaling protocol avoids the data traffic re-transmissions in the WDM domain. Once a data payload is launched into the ring topology, it is guaranteed to deliver without further delay or loss caused by resource contentions. Meanwhile, the signaling scheme accommodates the data payload with a single-hop transmission, which provides security and privacy transport services, and is convenient for CoS guarantees. Data payloads are propagated to the destination access node on the ringlet with the shorter hops. This way, the maximum number of hops a data burst is transported in the ring is $h_{max} = \lceil \frac{N-1}{2} \rceil$. Moreover, the reservation acknowledgement also carries the information of the network status which can be exploited for dynamic adjustment of the system parameters (e.g., the burst assembly interval τ_a, the delay period between consecutive signaling re-transmissions, and maximum resource reservation attempts R), thus enhancing the traffic engineering capability.

The proposed burst-based transmission mechanism is simple and efficient in that no complex determination for the offset time is involved, and that the value-added Delayed Reservation (DR) [25] mechanism can be easily implemented at the destination access node. Both system parameters (the offset time and the delayed reservation interval) are inherently equal to the cycle latency of the control packet. Such simplification is benefited from the ring topology of the metropolitan network. Meanwhile, the burst-based transmission mechanism enables us to engineering the data traffic at the access node according to the Ethernet address of the access node, the CoS requirements, and the multicast service. This solution complies with the *de facto* trend that only simple and straightforward processing is performed in the WDM layer, while most of the intelligence of the network is provided in the electronic domain. In other words, the burst-based transmission and the Ethernet-supported WDM metropolitan network are dual-benefited mechanisms.

6.3.3 Wavelength Allocation Algorithm

Besides the service provisioning and traffic engineering functionalities, the proposed network also features a hop-based wavelength allocation algorithm for efficient bandwidth utilization and contention resolution. The basic idea is to partition the bandwidth capacity of W data wavelengths into the disjoint subsets S_k ($k = 1, ..., h_{max}$), each containing a group of data channels, and being shared among the transmission demand with the same hop numbers. For example, the traffic sourced from access node A_i ($i \in \{0, ..., N-1\}$) and destined for A_{i+k} ($k \leq h_{max}$) share the same subset of data channels with that sourced from access node A_{i+1} and destined for A_{i+k+1}. Herein, the data channel consists of one wavelength or a portion of a wavelength. Assembling the data burst based on the hardware address of the destination access node enables the source access node to easily determine the number of hops that the traffic needs to be propagated.

By definition, there exists $S_k = \{ \omega_k^j \mid k = 1, ..., h_{max},\ j = 0, ..., S_k \}$. It is also observed that on the individual ringlet, the total number of sessions required for all pairs of access nodes to communicate is

$$H = \begin{cases} \sum_{k=1}^{h_{max}} k & N \text{ is an odd integer,} \\ \sum_{k=1}^{h_{max}} k - \lfloor \frac{h_{max}}{2} \rfloor & N \text{ is an even integer.} \end{cases} \tag{6.1}$$

Therefore, the minimum number of data channels required to support concurrent transmissions between all pairs of access nodes is $C_t = H$. The proposed wavelength allocation principle features the following characteristics:

1. The number of data channels allocated to the subset S_k is determined based on the associated transport distance (in hops) and the total available data channels, defined by

$$|S_k| = k \cdot \frac{C}{C_t} \tag{6.2}$$

   The contention-free wavelength allocation is theoretically achievable if $C \geq C_t$, i.e., $|S_k| \geq k$.

2. The data channel assignment can also take into consideration the reservation contention tolerance, the transport latency constraints, and the estimated traffic demand of the individual node, in order to facilitate the service differentiation.

3. Among the data channels of the subset S_k, the one chosen for the access node A_i to communicate with A_{i+k} can be determined via a variety of algorithms. For example, it can be selected randomly with a probability of $|S_k|^{-1}$. Alternatively, the data channel can be determined in a cyclic fashion. The most preferred channel for A_i to transport traffic requiring the hop number k is determined by (see Figure 6.6)

$$\omega_k{}^{pref}(A_i) = \omega_k{}^{imod|S_k|}. \tag{6.3}$$

```
Set picketChannel  = -1 ;
If preferredChannel  available
        set pickedChannel  = preferredChannel
else
        For ( l =0;  l < |Sk|; l++ ) {
                pickedChannel  = (preferredChannel  + l) mod  |Sk|;
                if pickedChannel  available
                        return  pickedChannel
                }

if pickedChannel  = -1
        return (No channel available).
```

Figure 6.6 The channel selection algorithm at the source access node.

The proposed algorithm provides a fair and scalable MAC mechanism. The transport demands requiring the same number of hops of propagation equally share the same group of data channels, regardless of the index of the source or the destination access nodes. Meanwhile, our subset-based wavelength allocation is advantageous in terms of computational simplicity. Its time complexity for data channel selection is $O(max(|S_k|))$, $k = 1, ..., h_{max}$.

6.4 Preformance Analysis

The system performance is evaluated via theoretical analysis and simulation results. Interesting performance metrics include the network throughput, the resource reservation blocking probability, the burst drop probability, and the transport latency. Theoretical analysis and simulations have been conducted to investigate their dependency on the number of access nodes, the maximum burst size, and the traffic intensity.

Simulations are based on the uniform traffic scenario, whereby the upstream traffic at each access node is destined for the other nodes with equal probabilities. The inter-arrival time of the data packets flowing into the access nodes are exponentially distributed. The metropolitan network consists of a regional size ring with

a circumference of $200km$. The cycle latency of a control packet is approximately $1000\mu s$. Table 6.2 summarizes the notations which will be used in the analysis.

Table 6.2 Notations for performance analysis ($i \in 1, ..., N$, $k \in 1, ..., H_{max}$) .

Term	*Explanation*
D	The transport latency owing to the call setup procedure.
B_0	The basic bandwidth capacity of one wavelength.
X_{max}	The maximum throughput achievable by the system.
X	The actually achievable network throughput.
P_f^k	The probability that a control packet fails to reserve resources for a transport requirement with hop number k.
P_f	The reservation blocking probability.
d_i	The depth of the LAN attached to the access node A_i.
t_l	The propagation time to transport a packet on one link of the LAN.
t_m	The propagation time to transport a packet on one link of the MAN.
D_c	The transport distance owing to the call setup procedure when the system adopts the E-ARP.
$D_c{}'$	The transport distance owing to the call setup procedure when the system adopts the traditional ARP.
α_i	latency reduction capability enabled by the E-ARP.
$\bar{T}_w^s$	The average signaling delay for a transmission requirement of k hops when the control packet and the burst assembly procedure are executed in sequence.
$\bar{T}_w^p$	The average signaling delay for a transmission requirement of k hops when the control packet and the burst assembly procedure are executed in parallel.
D_b^s	The average data burst delays induced by the burst-based transmission mechanism when the control packet and the burst assembly procedure are executed in sequence.
D_b^p	The average data burst delays induced by the burst-based transmission mechanism when the control packet and the burst assembly procedure are executed in parallel.

6.4.1 Network Throughput

Among many others, the property and efficiency of a system design are characterized by the network throughput, which is defined as the average data traffic (in bits) successfully transmitted by all the access nodes in a unit time. Of more interest are

the impact of the burst-based transmission mechanism and the hop-based wavelength allocation on the network throughput. In the presence of the balanced network traffic scenario, the transport request of different hop numbers is uniformly distributed.

Theoretically, the proposed system enables a throughput of

$$X_{max} = 2 \cdot \sum_{k=1}^{h_{max}} (B \cdot |S_k| \cdot \frac{N}{k}), \tag{6.4}$$

where B_c is the bandwidth capacity of a data channel, given by $B_c = \frac{W \cdot B_0}{H}$. The actually achievable network throughout is impacted by two factors: the number of the concurrently transported data burst limited by the signaling mechanism, and the signaling/data transmission ratio. The network throughput imposed by these constraints can be derived to be

$$X = N \cdot (N-1) \cdot \beta_1 \cdot \beta_2, \tag{6.5}$$

where β_1 and β_2 represent the ratio of the average reservation length ($\bar{L}$) over the cycle latency of the control packet on the ring (T_c), and that of the transported data bursts over the total reservation attempts, respectively. In a system with $W = H_t$, the network throughput normalized to the transport capacity of an MAN adopting SONET with the same configuration can be expressed as

$$X_0 = \begin{cases} 8 \cdot \frac{N}{N+1} \cdot \beta_1 \cdot \beta_2 & N \text{ is an odd integer,} \\ 8 \cdot \frac{N-1}{N} \cdot \beta_1 \cdot \beta_2 & N \text{ is an even integer,} \end{cases} \tag{6.6}$$

provided the total number of sessions in the network is $H_t = \frac{1}{8} \cdot (N^2 - 1)$ and $H_t = \frac{1}{8} \cdot (N^2)$ for even and odd integer of N, respectively.

Figure 6.7 shows the achievable network throughput versus the number of access nodes when the maximum burst size varies. The theoretical values are obtained based on Equation 6.6, assuming an ideal implementation (i.e., the resource reservation succeeds at the first signaling attempt). Simulations are based on two scenarios

whereby the data bursts are transported based on the loss-free fashion with the maximum burst size (L_{max}) being $2.5Mb$ and the loss-subjective fashion with $L_{max} = 1.5Mb$, respectively. It appears that increasing the maximum burst size improves the achievable network throughput (e.g., for $N = 15$, $X_0 = 1.13$ and 1.73 when $L_{max} = 1.5Mb$ and $2.5Mb$, respectively), and that while the experimental X_0 matches the theoretical values very well in the loss-subjective transport scenario, the two curves slightly diverge from each other in the loss-free alternative, revealing the impact of reservation blockings.

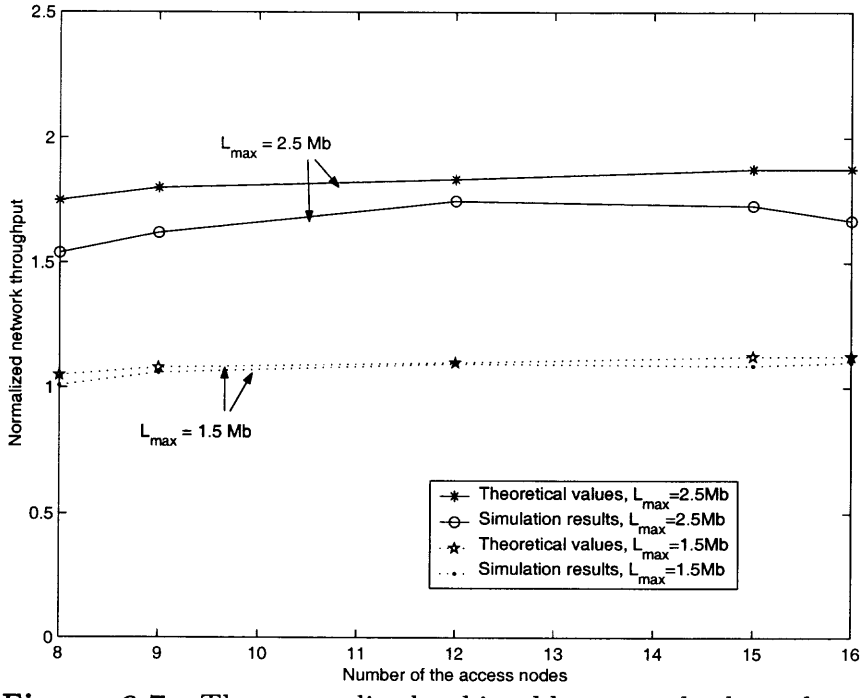

Figure 6.7 The normalized achievable network throughput versus the number of access nodes.

The simulation results in both scenarios imply the following conclusions. First, the proposed system delivers high throughput despite the large size of the ring or the large number of access nodes, indicating the efficient resource utilization capacity and high traffic volume supportability. Second, while incurring no burst loss by re-transmitting the control packets until the reservation succeeds, too many re-reservation attempts result in the large signaling response time at the access node

as well as the head of line (HOL) delay to the subsequent data bursts, both of which contribute to the lower network throughput. Dimensioning the maximum resource re-reservation attempts entails a tradeoff between the burst loss probability and the network throughput. Third, assembling the input packets into the larger transport granularity improves the throughput capacity with reduced signaling transmission requirement. This is consistent with the conclusion that the larger burst size within a certain range, the higher network throughput [63]. When the burst size keeps growing, however, the performance improvement slows down because the negative impact of the resource contention increases. This is also demonstrated in Figure 6.8, which illustrates the impact of L_{max} on the network throughput and the signaling reduction factor (i.e., the ratio of the signaling transmission requirements and the control packet processing complexity in a system without the assembly procedure over those in the proposed system), given that the maximum number of attempts for signaling re-transmission is limited. Figure 6.8 reinforces our previous conclusion that the network throughput is improved when L_{max} increases within a reasonable range.

6.4.2 Reservation Blocking Probability

The reservation blocking occurs when the data channel carried by a control packet is not available at the intermediate nodes between the source and destination access nodes. Assuming that the Probability Density Function (PDF) for the inter-arrival time of the reservation requests received at the access node is $f(t)$, and that the data channel is randomly selected from each subset with equal probability, the reservation blocking probability at an intermediate node is $P_c = \int_0^{L_{max}} f(t)dt$. Therefore, the resource reservation for a data burst propagating for k hops is blocked with the probability of

$$P_f^k = 1 - (1 - \frac{1}{|S_k|} \cdot P_c)^{k-1}. \tag{6.7}$$

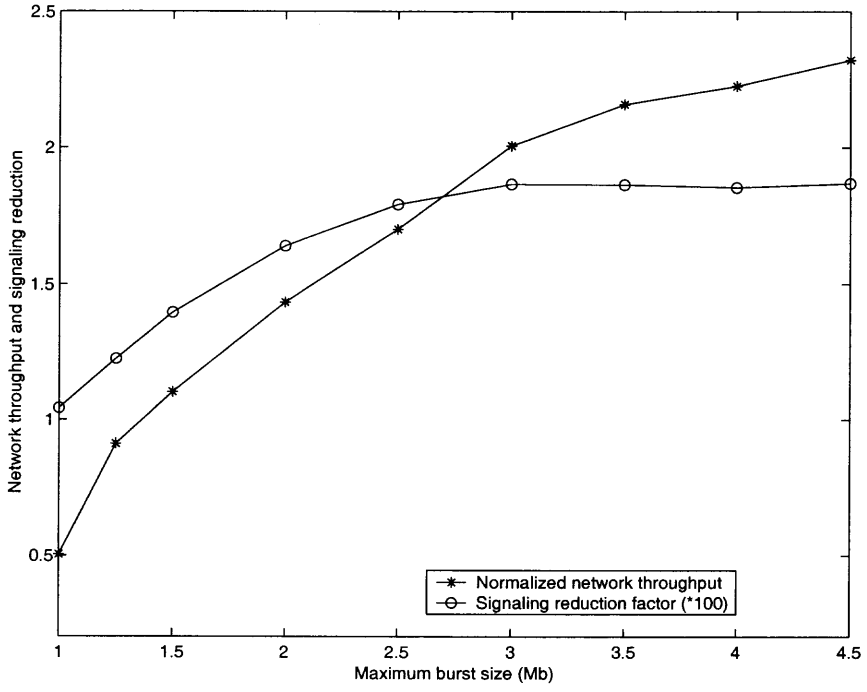

Figure 6.8 The impact of the maximum burst size (Mb) on the normalized achievable network throughput (X_0) and the signaling reduction factor (f) when $N = 1$ and $R = 3$.

By averaing the blocking probablity over all the possible distances, the average resource reservation blocking probability can be expressed as

$$P_f = 1 - \frac{\sum_{k=2}^{h_{max}} (1 - \frac{1}{|S_k|} \cdot P_c)^{k-1}}{h_{max} - 1}. \tag{6.8}$$

6.4.3 Transport Latency

The transport latency involves the delay experienced by the call setup procedure, and that by the data traffic at the source access node. To focus on the effect of the E-ARP and the burst-based transmission mechanism, the following analysis does not consider the delay due to the control packet generation and data channel determination.

The Call Setup Latency To analyze the latency reduction capability of the E-ARP, it is assumed that the LAN has a binary tree topology, wherein the root is the access node connecting the tree to the MAN and the leaf is typically a router from a corporate or campus network. Assuming $M_i \geq M_j$ and the distance (in hops)

between A_i and A_j is h_k, the total distance (in hops) required for call setup packets to build up the full communication between all the routers in the LAN A_i and those in A_j is

$$D_c = 2 \cdot [(d_i + h_k) + d_i \cdot (M_i - 1)] \cdot M_j. \tag{6.9}$$

When the conventional ARP is adopted, the distance for the same communication requirement is

$$D_c' = 2 \cdot [(d_i + h_k + d_j) \cdot M_i] \cdot M_j. \tag{6.10}$$

Without loss of generality, we assume each LAN has the same depth ($d_i = d_j$) and that the ARP packet propagation on one link in LAN is equal to that on one link in the ring (i.e., $t_l = t_m$). The average latency reduction delivered by E-ARP for node A_i to communicate with all other nodes is approximately

$$\alpha_i = \frac{\sum_{k=1}^{h_{max}} (1 - \frac{k + M_i \cdot \log_2 M_i}{(2 \cdot \log_2 M_i + k) \cdot M_i})}{h_{max}}. \tag{6.11}$$

The proposed E-ARP yields significant latency reduction for the call setup procedures as compared to the conventional ARP (see Figure 6.9). This is especially true when N is large, or when M is relatively small. By making the address mapping information retrievable for the subsequent inquiries at the access nodes, E-ARP obtains the larger α_i as N increases, and reduces the signaling overhead for address translations (i.e., the transmission requirements (in hops) of the call setup packets) at the WDM feeder ring by $\frac{M_i - 1}{M_i}$, as implied in Equation 6.10 and 6.11. In the assumed system scenario, the latency reduction percentage decreases as the size of the LAN gets larger, when $M_i \gg N$, $\alpha_i \to \frac{1}{2}$.

Data Burst Latency The proposed network architecture introduces two main sources that will cause the data burst delay at the access node: the delay caused the burst assembly procedure, which is given by $D_a = \frac{1}{2} \cdot \tau_a$, and the potential delay

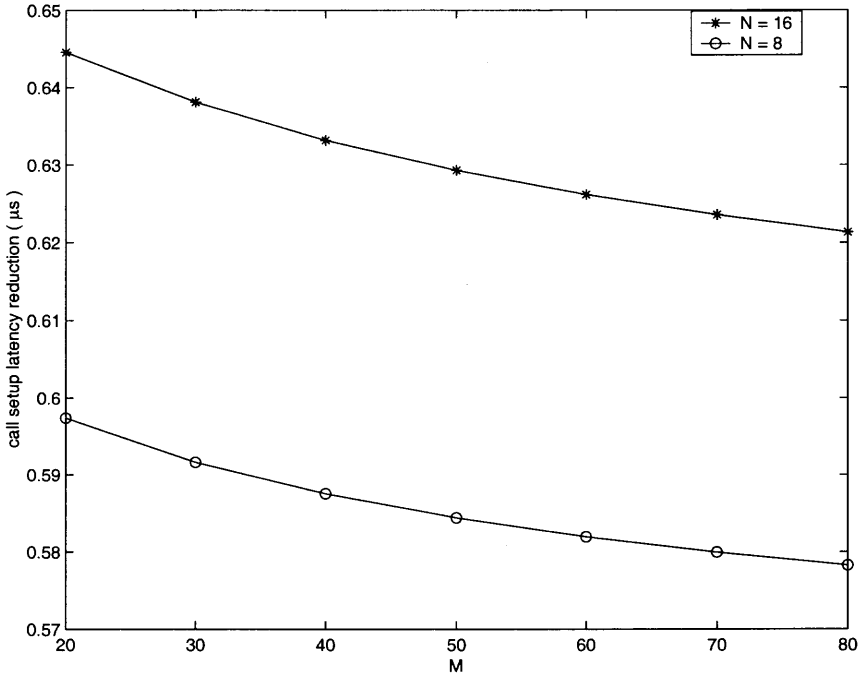

Figure 6.9 The latency reduction capability of the EARP with respect to the ARP.

because of unsuccessful resource reservations. Since the $i-th$ control packet retransmission induces a burst delay of $(T_c + t_r(i))$, the average signaling delay for a transmission requirement of k hops is

$$\bar{T_w}^s = T_c + \sum_{i=1}^{R} i \cdot (T_c + t_r(i)) P_k^f(i), \tag{6.12}$$

for a system wherein the control packet and the burst assembly procedure are executed in sequence, and

$$\bar{T_w}^p = \sum_{i=1}^{R} i \cdot (T_c + t_r(i)) P_k^f(i), \tag{6.13}$$

when the two procedures are performed in parallel. The average data burst delays are thus

$$D_b^s = \frac{1}{2} \cdot \tau_a + T_c + \frac{1}{h_{max} - 1} \cdot \sum_{k=2}^{h_{max}} \sum_{i=1}^{R} i \cdot (T_c + t_r(i)) P_k^f(i), \tag{6.14}$$

and

$$D_b^p = \frac{1}{2} \cdot \tau_a + max0, T_c - \tau_a + \frac{1}{h_{max} - 1} \cdot \sum_{k=2}^{h_{max}} \sum_{i=1}^{R} i \cdot (T_c + t_r(i)) P_k^f(i), \qquad (6.15)$$

respectively. In an ideal implementation, the average data burst delay is simply $D_b^s = D_a + T_c$ or $D_b^p = D_a$ when $\tau_a \geq T_c$.

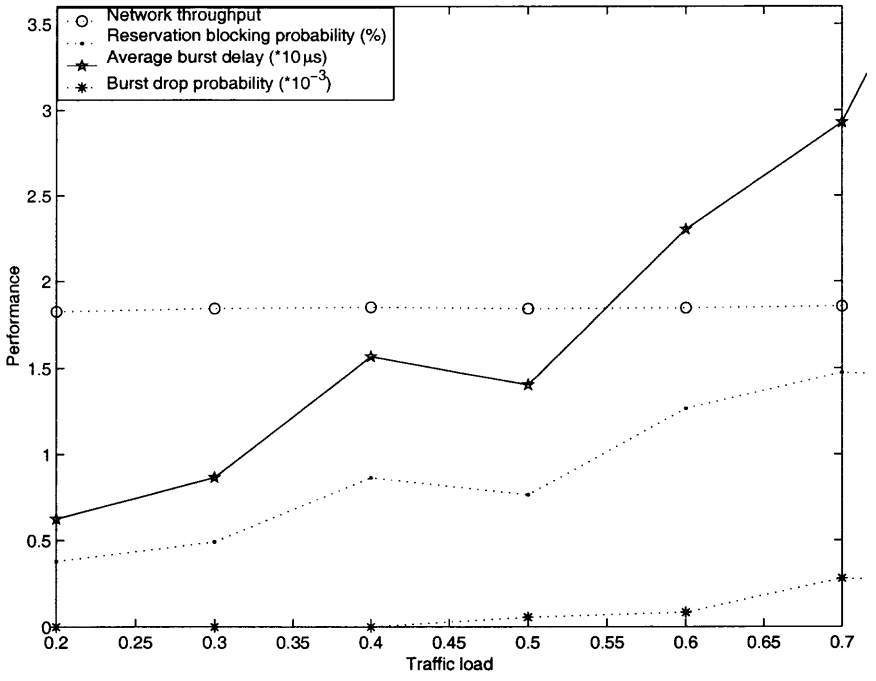

Figure 6.10 The preformance figures of merits versus the traffic load when the burst assembly and the control packet transmission are executed in parallel. The control packet reserves the resources according to the actual data burst length. $R = 3$, $N = 16$, and $L_{max} = 2.5$.

Given that in the sequential signaling scheme, the total signaling response time should also account for the cycle latency of the control packet which is approximately $1000\mu s$, the parallel signaling scheme is more favored for applications with the stringent time constraint. Figure 6.10 presents the burst delay performance respect to the input traffic intensity when the parallel transmission of the control packet and the data burst is examined. The reservation length is fixed to be the burst length. The proposed system yields satisfactory transport delay for today's voice and video

interactive applications with a stringent end-to-end latency bound requirement of a few hundred milliseconds. Besides the data burst latency at the access node, Figure 6.10 also presents the other performance metrics interested in this chapter.

6.5 Summary

This chapter addressed the control architecture and enabling technologies for the Ethernet-support IP-over- WDM metropolitan network. A set of mechanisms have been proposed by jointly considering both the Ethernet layer and the WDM layer that provide network service improvement to the real-time variable-length traffic.

Theoretical analysis and simulations exhibit encouraging results. The E-ARP protocol significantly reduces the transport latency and the transmission requirements for the call setup procedures. The burst-based traffic transmission is proved to be a feasible and effective paradigm for the metropolitan optical network. In addition, the novel hop-based wavelength allocation algorithm fairly arbitrates the channel access among all the access nodes. The enhancement of the Ethernet services, in tandem with the innovative mechanism in the WDM domain, facilitates a flexible and efficient prototype for the new generation metropolitan optical network dominated by the data-dominated traffic.

CHAPTER 7
CONCLUSIONS AND FUTURE RESEARCH

7.1 Conclusions

This dissertation research has been focused on the control architecture, enabling technologies, algorithm degisn, and performance analysis of the WDM burst-switched long haul and metropolitan area networks. Architectural frameworks for the IP over WDM networks that retain the advantages of the packet-oriented transmission mechanism while rendering elastic network resource utilization and graded levels of services have been proposed and investigated.

First and foremost, the burst-based transport mechanism has been investigated, with focus on the ingress traffic characteristics and its impact on the system performance. Theoretical analysis and simulation results show that, transmitting the traffic in the data burst granularity alleviates the signaling overhead at the intermediate nodes of IP over WDM networks, and is a feasible solution for the new generation Internet. Meanwhile, the coefficients of variation for the traffic inter-arrival time can be reduced as the traffic assembly interval increases at a reasonable range.

An innovative transport mechanism, namely, the Forward Resource Reservation (FRR) mechanism, has been proposed to reduce the end-to-end traffic delay for the burst-switched IP over WDM systems. The FRR mechanism explicitly adopts a linear predictive filter and incorporates an aggressive reservation strategy for data burst length prediction and resource reservation, respectively, and is subsequently extended for QoS differentiation at the network edges. The FRR scheme improves the realtime communication services for applications with time constraints without deleterious system costs.

One of the important enabling technologies of the proposed FRR-based WDM system is the aggressive strategy for the channel holding time reservation. It was

verified, via theoretical analysis and simulation results, that by properly choosing a small margin of correction for the resource reservation in addition to the predicted burst length, the proposed aggressive strategy-enhanced FRR scheme actually reduces the system overhead (in terms of operation cost and resource utilization) as compared to a system with zero-correction reservation algorithm. Specifically, two algorithms, the Success Probability-driven (SPD) and the Bandwidth Usage-driven (BUD) algorithms, were developed to facilitate the aggressive reservation strategy. These algorithms render explicit control on the latency reduction improvement and bandwidth usage efficiency, respectively, both of which are important figures of performance metrics.

The performance improvement issues for the FRR-enabled WDM system has been studied. Specifically, static and dynamic models targeting different desired system performance objectives (in terms of algorithm efficiency and system performance) have been created. Meanwhile, a "crank-back" based signaling scheme, termed as the Bandwidth Enhanced FRR (BEFRR) scheme, was devised to provide the bandwidth enhancement capability for the FRR-enabled WDM system. This mechanism, in tandem with a void-filling scheduling method, enables the intelligent usage of the network resources, and enhances the aggressive reservation algorithms with direct bandwidth utilization control.

The third major theme of the dissertation relates to the Ethernet-supported IP over WDM Metropolitan Area Network (MAN). The aim is to facilitate a common shared infrastructure enabling a graceful migration from the current voice-centralized infrastructure into a network dominated by data packets, thus making the new generation MAN optimized for scalable, survivable, and IP-dominated network at gigabit speed possible. An Enhanced Address Resolution Protocol (E-ARP) has been proposed to reduce the call setup latency and the signaling overhead associated with the address probing procedure. By re-using the retrievable information at the access

nodes, the proposed protocol decouples the traffic transport in the Ethernet domain from that in the WDM domain, and facilitates independent control operation in the individual network layers, thus benefiting the system with reduced management complexity and improved traffic engineering efficiency.

Meanwhile, the burst-based transmission mechanism, including traffic assembly mechanisms and signaling protocols, has been architected in the Ethernet-supported WDM infrastructure. It was pointed out, through a variety of concrete simulations, that the burst-based transmission mechanism addresses the traffic granularity problem in the MAN with a clean state, and that it complies with the *de facto* trend that only simple and straightforward control and management should be done in the WDM layer, while most of the intelligence of the network, such as traffic engineering and QoS provisioning, should be implemented at the edge nodes.

In addition, this dissertation studies the fair and scalable MAC protocol for network resource coordination and contention resolution. At the current stage, these problems are addressed by developing a novel hop-based wavelength allocation algorithm, whereby the data channels are grouped into subsets, each of which is shared by the transport demands requiring the same number of hops. The proposed algorithm is a skeleton based on which a variety of algorithms are possible for channel dedication and selection, with or without considerations of the traffic priorities.

7.2 Future Work

The directions of the future research efforts will be on two categories:

- Extention of the proposed architecture framework.
- Exploration of novel topics in other broadband networks.

First, as a natural outgrowth and maturity of the current research, efforts will be devoted to the applicability of the proposed protocols and techniques. For example, it will be interesting to examine the applicability and the impacts of the proposed

protocols and techniques (e.g., the FRR mechanism and the enhanced ARP) in the heterogeneous networks, such as IP Telephony and wireless IP further. Also of interest are the interoperability and contention resolution issues for the Ethernet-supported IP over WDM MAN.

Second, the future research will go beyond the current topics. It is important to apply the achieved experience and insight in an integrated and cohesive manner to the state of the art in other broadband networks. Specifically, such studies will cover i) performance evaluation and improvement, ii) the framework for end-to-end QoS provisioning, and iii) network reliability and protection/restoration.

REFERENCES

[1] A. Rodriguez-Moral, P. Bonenfant, S. Baroni, and R. Wu, "Optical Data Networking: Practical, Technologies, and Architectures for Next Generation Optical Transport Networks and Optical Internetworks," *J. Lightwave Technol.*, vol. 18, pp. 1855–1870, 2000.

[2] J. Wei, C. Liu, S. Park, K. Liu, R. Ramamurthy, H. Kim, and M. Meada, "Network Control and Management for the Next Generation Interne," *IEICE Trans. on Commun.*, vol. 18B, pp. 2191–2209, 2001.

[3] N. Ghani, S. Dixit, and T. Wang, "On IP-over-WDM Integration," *IEEE Commun. Mag.*, vol. 38, pp. 72–84, 2000.

[4] T. E. Stern and K. Bala, *Multiwavelength Optical Networks, A Layered Approach.* New Jersey: Addison-Wesley, 2000.

[5] W. Goralski, *SONET.* McGraw-Hill, 2000.

[6] I. White, R. Penty, M. Webster, C. Y.J., A. Wonfor, and S. Shahkooh, "Wavelength Switching Components for Future Photonic Networks," *IEEE Commun. Mag.*, vol. 40, no. 9, pp. 74–81, 2002.

[7] B. Rajagopalan, J. Luciani, D. Awduche, B. Cain, and B. Jamoussi, "Ip Over Optical Networks: A Framework," *IETF Document.*, vol. 40, no. 9, pp. 74–81, 2002.

[8] R. Xu, Q. Gong, and P. Ye, "A Novel IP With MPLS Over WDM-based Broadband Wavelength Switched IP Network," *J. Lightwave Technol.*, vol. 19, no. 5, pp. 596–602, 2001.

[9] T. EL-Bawab and J. Shin, "Optical Packet Switching in Core Networks: Between Vision and Reality," *IEEE Commun. Mag.*, vol. 40, no. 9, pp. 60–65, 2002.

[10] J. Diao and P. Chu, "Packet Rescheduling in WDM Star Networks With Real-time Service Differentiation," *J. Lightwave Technol.*, vol. 19, no. 12, pp. 1818–1828, 2001.

[11] B. Meagher, G. Chang, G. Ellinas, Y. Lin, W. Xin, T. Chen, X. Yang, A. Chowdhury, J. Young, S. J. Yoo, C. Lee, M. Z. Iqbal, T. Robe, H. Dai, Y. J. Chen, and W. I. Way, "Design and Implementation of Ultra-low Latency Optical Label Switching for Packet-switched WDM Networks," *J. Lightwave Technol.*, vol. 18, no. 12, pp. 1978–1987, 2000.

[12] B. Li, A. Ganz, and C. M. Krishna, "An In-band Signaling Protocol for Optical Packet Switching Networks," *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, pp. 1876–1884, 2000.

[13] Y. Xiong, M. Vandenhoute, and H. Cankaya, "Control Architecture in Optical Burst-switched WDM Networks," *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, pp. 1838–1851, 2000.

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

[15] S. Turner, "Wdm Burst Switching for Petabit Data Networks," *Optical Fiber Commun. Conf.*, vol. 2, pp. 47–49, 2000.

[16] J. Choi, M. Kang, G. Lee, J. Choi, Y. Cha, and W. Rhee, "Extension of GSMP for Optical Burst Switching," *IETF Document*, 2002.

[17] J. White, M. Zukerman, and H. Vu, "A Framework for Optical Burst Switching Network Design," *IEEE Commun. Lett.*, vol. 6, no. 6, pp. 268–270, 2002.

[18] A. Ge, F. Callegati, and L. Tamil, "On Optical Burst Switching and Self-similar Traffic," *IEEE Commun. Lett.*, vol. 4, no. 3, pp. 98–100, 2000.

[19] M. Yoo, C. Qiao, and S. Dixit, "Optical Burst Switching for Service Differentiation in the Next-generation Optical Internet," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 98–104, 2001.

[20] S. Verma, H. Chaskar, and R. Ravikanth, "Optical Burst Switching: A Viable Solution for Terabit IP Backbone," *IEEE Network*, vol. 14, no. 6, pp. 48–53, 2000.

[21] I. Baldine, G. Perros, and G. Stevenson, "JumpStart: A Just-In-Time Signaling Architecture for WDM Burst-switched Networks," *IEEE Commun. Mag.*, vol. 40, no. 2, pp. 82–89, 2002.

[22] M. Yoo and C. Qiao, "Just-Enough-Time (JET): A High Speed Protocol for Bursty Traffic in Optical Networks," *1997 Digest of the IEEE/LEOS Summer Topical Meeting*, pp. 26–27, 1997.

[23] J. Wei, R. McFarland, and Jr., "Just-In-Time Signaling for WDM Optical Burst Switching Networks," *J. Lightwave Technol.*, vol. 18, no. 12, pp. 2019–2037, 2000.

[24] M. Jeong, Y. Xiong, H. Cankaya, M. Vandehoute, and C. Qiao, "Efficient Multicast Schemes for Optical Burst-switched WDM Networks," *IEEE Intl. Conf.*, vol. 3, pp. 1289–1294, 2000.

[25] M. Yoo, C. Qiao, and S. Dixit, "Qos Performance of Optical Burst Switching in IP-over-WDM Networks," *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, pp. 2062–2071, 2000.

[26] F. Callegati, A. Cankaya, Y. Xiong, and M. Vandenhoute, "Design Issues of Optical IP Routers for Internet Backbone Applications," *IEEE Commun. Mag.*, vol. 37, no. 12, pp. 124–128, 1999.

[27] L. Liu, P. Wan, and O. Frieder, "Optical Burst Switching: the Next IP Revolution Worth Multiple Billions Dollars," *MILCOM 2000 21st Century Military Commun.*, vol. 2, pp. 881–885, 2000.

[28] M. Listanti, V. Eramo, and R. Sabella, "Architectural and Technological Issues for Future Optical Internet Networks," *IEEE Commun. Mag.*, vol. 38, no. 9, pp. 82–92, 2000.

[29] D. Comer, *Computer Networks and Internet with Internet Applications, Fourth Edition.* New Jersey: Prentice Hall, 2004.

[30] D. Stoll, P. Leisching, H. Bock, and A. Richter, "Metropolitan DWDM: A Dynamically Configurable Ring for the KomNet Field Trail in Berlin," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 106–113, 2001.

[31] E. Eernandez-Valencia, "Hybrid Transport Solutions for TDM/Data Networking Services," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 104–122, 2002.

[32] G. Kramer and G. Pesavento, "Ethernet Passive Optical Network (EPON): Building a Next-generation Optical Access Network," *IEEE Commun. Mag.*, vol. 40, no. 2, pp. 66–73, 2002.

[33] G. Chlruvolu, G. An, D. Elie-Dit-Cosaque, M. Ali., and J. Rouyer, "Issues and Approaches on Extending Ethernet Beyond LANs," *IEEE Commun. Mag.*, vol. 42, no. 3, pp. 80–86, 2004.

[34] J. Xu, H. Perros, and G. Rouskas., "Techniques for Optical Packet Switching and Optical Burst Switching," *IEEE Commun. Mag.*, vol. 39, no. 1, pp. 136–142, 2001.

[35] M. Duser and P. Bayvel, "Analysis of a Dynamically Wavelength-routed Optical Burst Switched Network Architecture," *J. Lightwave Technol.*, vol. 20, no. 4, pp. 574–585, 2002.

[36] S. Oh and M. Kang, "A Burst Assembly Algorithm in Optical Burst Switching Networks," *Optical Fiber Commun. Conf. and Exhibit*, pp. 17–22, March 2002.

[37] A. Detti and M. Listanti, "Impact of Segments Aggregation on TCP Reno Flows in Optical Burst Switching Networks," *IEEE INFOCOM*, pp. 1803–1812, June 2002.

[38] L. G. Alberto, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science and Applications.* New York: Wiley, 1998.

[39] Y. Yang and J. Wang, "Optimal All-to-all Personalized Exchange in a Class of Optical Multistage Networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, no. 3, pp. 261–274, 2000.

[40] L. Kleinrock, "The Latency/Bandwidth Tradeoff in Gigabit Networks," *IEEE Commun. Mag.*, vol. 30, no. 4, pp. 36–40, 1992.

[41] Y. Yang and J. Wang, "Optimal All-to-all Personalized Exchange in A Class of Optical Multistage Networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, no. 3, pp. 261–274, 2000.

[42] M. Honig and D. Messerschmitt, *Adaptive Filters: Structures, Algorithms, and Applications.* Boston: Kluwer Academic Publisher, 1984.

[43] J. Treichler, C. Johnson, Jr., and M. Larimore, *Theory and Design of Adaptive Filters.* New York: A Wiley-interscience Publication, 1987.

[44] L. Alberto, *Probability and Random Processes for Electrical Engineering.* Massachusetts: Addison-Wesley, 1994.

[45] K. Maney, "Many Fiber-optic Lines Unused Despite Rising Demand," *USA Today*, 2002.

[46] V. Paxson, "Fast Approximation of Self-similar Network," *Tech. Rep. LBL-36750*, 1995.

[47] I. Norros, "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks," *IEEE Select. Areas Commun.*, vol. 13, no. 6, pp. 953–962, 1995.

[48] J. Beran, *Statistics for Long-Memory Processes.* New York: Chapman & Hall, 1994.

[49] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the Self-similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[50] B. Tsybakov and N. Georganas, "Self-similar Processes in Communication Network," *IEEE Trans. Info. Theory*, vol. 44, no. 5, pp. 1713–1725, 1998.

[51] L. Tancevski, S. Yegnanarayananm, G. Castanon, and L. Tamil, "Optical Routing of Asynchronous, Variable Length Packets," *IEEE Select. Areas Commun.*, vol. 18, no. 10, pp. 2084–2093, 2000.

[52] M. Scholten, Z. Zhu, E. Hernandez-Valencia, and J. Hawkins, "Data Transport Applications Using GFP," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 96–103, 2002.

[53] N. Madamopoulos, D. C. Friedman, I. Tomkow, and A. Boskovic, "Study of the Performance of a Transparent and Recondifurable Metropolitan Area Network," *J. Lightwave Technol.*, vol. 20, no. 6, pp. 937–954, 2002.

[54] I. Chlamtac, V. Elek, A. Fumagalli, and C. Szabo, "Scalable WDM Access Network Architecture Based on Photonic Slot Routing," *IEEE/ACM Trans. Networking*, vol. 7, no. 1, pp. 1–9, 1999.

[55] J. Cai, A. Fumagalli, and I. Chlamtac, "The Multitoken Interarrival Time (MTIT) Access Protocol for Supporting Variable Size Packets Over WDM Ring Network," *IEEE Select. Areas Commun.*, vol. 18, no. 10, pp. 2094–2104, 2000.

[56] K. Bengi and H. R. Van As, "Efficient QoS Support in a Slotted Multihop WDM Metro Ring," *IEEE Select. Areas Commun.*, vol. 20, no. 1, pp. 216–227, 2002.

[57] A. C. Kam and K. Siu, "Supporting Bursty Traffic With Bandwidth Guarantee in WDM Distribution Networks," *IEEE Select. Areas Commun.*, vol. 18, no. 10, pp. 2029–2040, 2000.

[58] M. A. Marsan, A. Bianco, E. Keonardi, M. Meo, and F. Meri, "Mac Protocols and Fairness Control in WDM Multirings With Tunable Transmitters and Fixed Resceivers," *J. Lightwave Technol.*, vol. 14, no. 6, pp. 1230–1244, 1996.

[59] C. S. Jeiger and J. M. H. Elmirghani, "Photonic Packet WDM Ring Networks Architecture and Performance," *IEEE Commun. Mag.*, vol. 40, no. 11, pp. 110–115, 2002.

[60] J. P. Kamnisow, C. R. Doerr, and C. Dragone, "A Wideband All-optical WDM Network," *IEEE Select. Areas Commun.*, vol. 14, no. 6, pp. 780–799, 1996.

[61] K. Zhu and B. Mukherjee, "Traffic Grooming in An Optical WDM Mesh Network," *IEEE Select. Areas Commun.*, vol. 20, no. 1, pp. 122–133, 2002.

[62] D. C. Plummer, "Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48 Bit Ethernet Address for Transmission on Ethernet Hardware," *Request for Comments (Standard) RFC 826, IETF Document*, no. 11, 1982.

[63] J. Cai and A. Fumagalli, "An Analytical Framework for Performance Comparison of Bandwidth Reservation Schemes in WDM Rings," *IEEE INFOCOM*, pp. 41–47, June 2002.