Design and Analysis of Optical Packet-Switched Networks

By

Shun Yao
B.E., Electronic Engineering (Tsinghua University, China) 1997

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical and Computer Engineering

in the
OFFICE OF GRADUATE STUDIES
of the
UNIVERSITY OF CALIFORNIA
DAVIS

Approved by the Committee in Charge:

Dr. Biswanath Mukherjee
Committee Chair and
Professor of Computer Science

Dr. S. J. Ben Yoo
Associate Professor of
Electrical and Computer Engineering

Dr. Jonathan P. Heritage
Professor of
Electrical and Computer Engineering

2001
Design and Analysis of Optical Packet-Switched Networks

Abstract

Over the past few years, the Internet traffic has grown rapidly and the optical transport bandwidth has been continuously increasing. These changes are stimulating the evolution of data networks. In such a dynamic environment, a network architecture that accommodates multiple data formats, supports high throughput, and is capable of flexible bandwidth provisioning is the key to the design and development of the next-generation Internet. The past evolution of Internet has proved that packet switching is a scalable, robust technology that is capable of adapting to various traffic patterns. Packet switching in the optical domain provides data-rate and data-format transparency, and it scales well to match the growing bandwidth of fiber optics. Furthermore, optical packet switching can remove the optical-electrical-optical (OEO) conversion in the network and, therefore, reduce the footprint of the switches.

This dissertation examines the performance and design issues of optical packet-switched networks. It first presents a survey of the current state of the art in optical packet switching. Then, it investigates one of the most important design issues: contention resolution in time, wavelength, and space domains. Following this comprehensive study, it develops an analytical modeling technique, called PLATO, for evaluation of the network performance. The PLATO technique is an accurate approach applicable to various network topologies and switch architectures. After the PLATO model is a unified and extended study on contention-resolution schemes. This part of the dissertation introduces priority-based routing, which leads to the implementation of Class of Service (CoS). It also proposes an optical-electrical hybrid contention-resolution scheme that demonstrates excellent performance with a cost-effective switch architecture. Finally this dissertation explores the possibility of applying optical packet switching to metropolitan-area networks.
Contents

1 Introduction .............................................. 1
  1.1 Motivation .......................................... 1
  1.2 Contributions ....................................... 3
  1.3 Dissertation Outline ................................. 3

2 Advances in Optical Packet Switching: An Overview ...... 4
  2.1 Introduction .......................................... 4
    2.1.1 Delay Variation Between Nodes ................. 6
    2.1.2 Delay Variations Inside the Nodes ............ 6
  2.2 Synchronization of Slotted Networks .................. 6
    2.2.1 Impact of Packet Format ....................... 6
    2.2.2 Bit-Level Synchronization ...................... 8
    2.2.3 Packet-Synchronization Schemes ............... 9
  2.3 Unslotted Networks .................................... 10
  2.4 Contention Resolution ............................... 10
    2.4.1 Optical Buffering ............................... 11
    2.4.2 Deflection Routing .............................. 14
    2.4.3 Wavelength Conversion ......................... 17
  2.5 Header and Packet Format ................................ 18
  2.6 Advances Toward Optical Sub-Wavelength Switching ... 21
    2.6.1 Optical Label Switching ....................... 21
    2.6.2 Photonic Slot Routing ........................... 21
    2.6.3 Optical Burst Switching ....................... 22
  2.7 Summary ............................................. 25

3 Contention-Resolution Schemes in Optical Packet-Switched Networks ...... 26
  3.1 Introduction ......................................... 26
  3.2 Contention-Resolution Schemes ....................... 28
3.2.1 Optical Buffering .................................................. 29
3.2.2 Wavelength Conversion ........................................... 30
3.2.3 Space Deflection .................................................. 31
3.2.4 Combinational Schemes .......................................... 32
3.3 Numerical Results and Performance Comparison .................. 33
   3.3.1 Simulation Experiments and Configuration ................... 33
   3.3.2 Comparison of the Basic Contention-Resolution Schemes ...... 38
   3.3.3 Comparison of Combinational Schemes ....................... 42
   3.3.4 Limited Wavelength Conversion ............................... 44
   3.3.5 Comparison Study of Slotted and Unslotted Networks ......... 45
3.4 Conclusion .......................................................... 46

4 PLATO: A Generic Modeling Technique for Optical Packet-Switched Networks ................................................. 48
   4.1 Introduction ....................................................... 48
   4.2 Basic Algorithms of PLATO ....................................... 49
      4.2.1 Traffic-Flow Analysis ....................................... 51
      4.2.2 Local Switch Modeling ...................................... 53
      4.2.3 Constructing the Network Model ............................ 55
   4.3 Delay-Line Model ................................................ 57
   4.4 Numerical Results and Discussion .................................. 59
   4.5 Conclusion ........................................................ 62

5 Unification and Extension of Contention-Resolution Studies in Optical Packet-Switched Networks ............................. 65
   5.1 Introduction ....................................................... 65
   5.2 Priority-Based Routing in Optical Packet-Switched Networks .... 65
      5.2.1 Node Architecture and Routing Policies ...................... 66
      5.2.2 Illustrative Results and Discussion ............................ 68
      5.2.3 Summary ...................................................... 74
   5.3 Comparison of Slotted and Unslotted Networks with Priority-Based Routing .................................................. 74
      5.3.1 Node Architecture and Routing Policies ...................... 74
      5.3.2 Illustrative Results and Discussion ............................ 76
      5.3.3 Summary ...................................................... 76
   5.4 A Hybrid Contention-Resolution Scheme ........................... 79
      5.4.1 Network Architecture .......................................... 79
      5.4.2 Simulation Configuration ...................................... 81
      5.4.3 Illustrative Results and Discussion ............................ 82
6 Optical Packet Switching for Metropolitan-Area Networks: Opportunities and Challenges

6.1 Introduction .......................................................... 89
6.2 Optical Packet Switching for MAN ................................. 89
  6.2.1 Network Requirements .......................................... 89
  6.2.2 Promises of Optical Packet Switching ......................... 91
  6.2.3 Inter-Network Interfaces ....................................... 93
6.3 Enabling Technologies ............................................... 94
6.4 Summary ............................................................... 96

7 Conclusion .............................................................. 97
  7.1 Advances in Optical Packet Switching: An Overview .......... 97
  7.2 Contention-Resolution Schemes in Optical Packet-Switched Networks ............................................. 98
  7.3 PLATO: A Generic Modeling Technique for Optical Packet-Switched Networks ...................................... 99
  7.4 Unification and Extension of Contention-Resolution Studies .......................................................... 99
  7.5 Optical Packet Switching for MAN ............................... 101
List of Figures

1.1 The evolution of the data-network protocol stack. ........................................ 2
2.1 A generic node architecture of the slotted network. .................................. 5
2.2 Functional block diagram of synchronization of the packets. ...................... 7
2.3 Two possible packet formats that determine the synchronization schemes. .... 8
2.4 A scheme for input synchronization stage in a node. .................................. 9
2.5 A generic node architecture of the unslotted network. .............................. 10
2.6 Broadcast-and-select switch proposed in the KEOPS project. ..................... 12
2.7 The shared-memory optical packet switch. .................................................. 13
2.8 Fiber-loop-memory switch from ATMOS. ................................................... 13
2.9 A 2x2 switching element containing optical buffer. ................................. 14
2.10 The Manhattan Street Network and ShuffleNet. ...................................... 15
2.11 Example of node architecture for hot-potato routing or deflection with limited
    buffer routing in MSN and ShuffleNet. ...................................................... 15
2.12 Power spectrum of the laser modulation current. ....................................... 19
2.13 Header retrieval in SCM. .......................................................................... 19
2.14 PSR node architecture. ............................................................................. 22
2.15 OBS node architecture. ........................................................................... 23
2.16 Offset time and delayed reservation in the JET protocol. ......................... 24

3.1 Node architectures for different contention-resolution schemes: (a) buf; (b)
    bufwdm; (c) wc; (d) wc+bufwdm. ............................................................. 29
3.2 Parametric wavelength conversion. ............................................................. 31
3.3 An example mesh network topology. .......................................................... 33
3.4 Probability distribution function (PDF) of IP packet sizes. ....................... 36
3.5 Performance comparison of the basic schemes – baseline and wavelength con-
    version. (a) Throughput comparison; (b) packet-loss rate comparison; (c)
    average end-to-end delay comparison; (d) average hop distance comparison. 38
3.6 Performance comparison of the basic schemes – baseline, optical buffering, and deflection. (a) Throughput comparison; (b) packet-loss rate comparison; (c) average end-to-end delay comparison; (d) average hop distance comparison.

3.7 Performance comparison of the combinational schemes. (a) Throughput comparison; (b) packet-loss rate comparison; (c) average end-to-end delay comparison; (d) average hop distance comparison.

3.8 Performance comparison of limited wavelength conversion. (a) Throughput comparison; (b) packet-loss rate comparison.

3.9 Packet-loss rate comparison between slotted and unslotted networks.

4.1 Building blocks of PLATO.

4.2 An example six-node network.

4.3 Possible contentions a $t_{14}$ packet can encounter.

4.4 Contention for $t_{14}$ at node 2.

4.5 A baseline optical packet switch.

4.6 Vulnerable period of the shaded packet.

4.7 Switch architecture with one dedicated delay line for each output port.

4.8 Three streams contending for one output port, with one dedicated delay line.

4.9 The NSF network topology with link cost.

4.10 PLR with baseline switch architecture for the 6-node topology and the NSF topology.

4.11 A zoom-in view of the light-load region of Fig. 4.10.

4.12 PLR with buffering switch architecture for the NSF topology.

5.1 Network topology under study.

5.2 (a) Architecture for node 1, 4; (b) architecture for node 2, 3, 5, 6.

5.3 Comparison of end-to-end delay (10%-20%-60% priority distribution).

5.4 Comparison of average hop distance (10%-20%-60% priority distribution).

5.5 Comparison of packet-loss rate (10%-20%-60% priority distribution).

5.6 Comparison of percentage of packets lost at receiver (10%-20%-60% priority distribution).

5.7 Comparison of end-to-end delay (30%-30%-40% priority distribution).

5.8 Comparison of average hop distance (30%-30%-40% priority distribution).

5.9 Comparison of packet-loss rate (30%-30%-40% priority distribution).

5.10 Comparison of percentage of packets lost at receiver (30%-30%-40% priority distribution).

5.11 (a) Network topology and (b) node architecture under study.

5.12 Overall packet-loss rate.
5.13 Class 1 packet-loss rate. .................................................. 77
5.14 Class 2 packet-loss rate. .................................................. 78
5.15 Class 3 packet-loss rate. .................................................. 78
5.16 Client interface of optical packet-switched networks. ............... 80
5.17 Architecture of the proposed hybrid contention-resolution scheme. 80
5.18 Packet-loss rate of different simulation scenarios. ..................... 83
5.19 Average end-to-end delay of different simulation scenarios. ............. 83
5.20 Proposed node architecture. ................................................ 85
5.21 Network topology for TCP experiment. .................................. 86
5.22 Comparison of $T_{FTP}$ for different TCP window sizes. ............... 87
5.23 Comparison of $T_{FTP}$ for different aggregation schemes. ............. 88
5.24 Comparison of packet-loss rate. ........................................... 88

6.1 The SONET ring-based network. .......................................... 91
6.2 Generic block diagram of an optical packet switch. ....................... 92
6.3 Client interface between optical packet-switched MAN and access networks. 94
6.4 Two options to inter-connect the optical packet-switched MAN with optical-crossconnected backbone. (a) Using electrical equipment at the interface. (b) Using label stacking and end-to-end signaling. ............... 95
ACKNOWLEDGMENTS

I thank Professor Biswanath Mukherjee, my research advisor and committee chair, for his guidance and encouragement throughout the course of my graduate study. When I came to UC Davis three years ago, it was Dr. Mukherjee who discovered my interest in optical networks and recognized my research ability. His boundless enthusiasm, dedication to excellence, careful attention to detail, and infinite patience will always be inspiring to me.

I am grateful to Professor S. J. Ben Yoo for guiding my research and giving me valuable advice. I have benefited greatly from his expertise in the field of optical communications.

I thank Professor Jonathan Heritage for serving on my PhD dissertation committee and PhD qualifying exam committee.

I thank Professor Dipak Ghosal for chairing my PhD qualifying exam.

I thank Dr. Sudhir Dixit at Nokia Research Center, for giving me valuable guidance and advice during my summer internships.

I gratefully acknowledge the support from Nokia Research Center in funding this research.

I thank the Electrical Engineering and Computer Science faculty at the University of California, Davis, for their superb teaching and guidance. I thank the office staff and the systems support staff for all their assistance.

I thank all of the members of the UC Davis Networks Research Lab for creating such a dynamic culture for research and study. Many thanks to Toshit Antani, Yash Bansal, Matthew Caesar, Dr. Wonhong Cho, Dr. Debasish Datta, Aysegul Gencata, Yurong Huang, Dr. Jason Jue, Dr. Young-Chon Kim, Keith Kong, Glen Kramer, Canhui Ou, Vijoy Pandey, Kama Potherlanka, Dr. Laxman Sahasrabuddhe, Narendra Singhal, Jian Wang, Dr. Wushao Wen, Dr. Xiaoxin Wu, Dr. Hui Zang, Jing Zhang, Hongyue Zhu, and Keyao Zhu for their technical expertise and general camaraderie.

Finally, and most importantly, I thank my parents and my fiancee Florence for their constant support and encouragement.
Chapter 1

Introduction

1.1 Motivation

Over the past few years, wavelength-division multiplexing (WDM) technology has brought dramatic and fundamental changes to the network design [1, 2]. Wavelength-routed networks, in which lightpaths are setup on specific wavelengths, have been the focus of extensive studies [3, 4]. Over a short period of a few years, these networks have evolved from textbook subjects to real-life products. The current ongoing efforts to automate and expedite wavelength and bandwidth provisioning in the optical layer indicate the inevitable trends that lead to more intelligent optical networks [5, 6, 7]. Migration of certain switching functionality from electronics to optics will remove the incumbent layers that impose unnecessary optical-electrical-optical conversions and unnecessary signal processing. The pre-optical Open System Interconnection (OSI) protocol stack, defined decades ago, are no longer applicable in the new networking environment (Fig. 1.1).

Until recently, Internet Protocol (IP) routers were inter-connected by virtual circuits provided by Asynchronous Transfer Mode (ATM) cell switches. The ATM network is built with links provided by Synchronous Optical Networks (SONET) in the form of time-division multiplexed (TDM) circuits. As the networks evolved, the ATM was gradually replaced by IP, and IP routers are directly interfaced with SONET equipment. The recently proposed Multiprotocol Label Switching (MPLS) protocols further enrich the functionalities of IP. The increase in an IP router’s port data rate and aggregate capacity is leading the IP routers to be directly linked with wavelengths in the WDM layer. As the networks continue to evolve and new technologies keep emerging, an optical packet-switched (OPS) layer can carry multiple-protocol client networks and provide a seamless integration with the WDM layer. Such a flexible WDM optical network is desired by service providers to meet their versatile traffic demands. As a result, a number of research groups have proposed various
optical (sub-wavelength) switching paradigms [8, 9].

From the demand side, the tremendous increase in the transport networks’ bandwidth is stimulating the high volume of gigabit multimedia services. A robust network supporting various kinds of traffic is the cornerstone for the next-generation Internet. The past evolution of Internet has proved that packet switching, regardless of the underlying transport technology, is such a flexible and robust technology that is scalable and is able to adapt to new traffic patterns. The key to the success of packet-switched networks lies in the deployment of cost-effective packet switching systems that provide simple access to the large bandwidth in the optical networks.

Optical packet switching has been increasingly receiving attention because it can seamlessly interface with the WDM transport layer while offering a number of layer-two and layer-three functions [10]. It is capable of dynamically allocating network resources with fine granularity, and of supporting incremental scaling of the network. Optical packet switching appears to be a candidate transport solution to close the gap between the electrical (IP/MPLS) layer and the optical (WDM) layer [11, 12]. Various research groups have proposed several node architectures for optical packet switching. To contribute to the body of knowledge on optical packet switching, this dissertation conducts a comprehensive study of optical packet switching in a network context, and aims to fully understand the design issues in building such a network. Moreover, it is also our goal to find a cost-effective network architecture based on currently available and emerging technologies.

![Protocol stack evolution]

Figure 1.1: The evolution of the data-network protocol stack.
1.2 Contributions

This dissertation makes five important contributions to the body of knowledge on optical packet-switched networks. First, it provides an overview of some of the most important design issues in optical packet switching \([13]\): synchronization and contention resolutions. It reviews the previously proposed architectures and contention-resolution algorithms, and pinpoints the direction for further investigation. Second, it conducts a comprehensive, network-based study of contention-resolution schemes in time, wavelength, and space domains \([14]\). Third, it develops a generic analytical modeling technique, called PLATO, that can accurately evaluate the network’s performance with an arbitrary topology and switch architecture. Fourth, it unifies and extends the study on contention-resolution schemes, and investigates priority-based routing \([15]\). It also proposes an optical-electrical contention-resolution scheme that gives excellent packet-loss performance with a simple architecture \([16]\). Fifth, it investigates the application of optical packet switching in metropolitan-area networks and proposes two interface models \([12]\).

1.3 Dissertation Outline

The dissertation is organized as follows. Chapter 2 discusses the main design issues in optical packet switching and provides an overview of the current state of the art. Chapter 3 presents a comprehensive study of contention-resolution schemes in time, wavelength, and space domains, based on an arbitrary mesh network topology. Chapter 4 develops PLATO, a generic analytical modeling technique to evaluate packet-loss performance in optical packet-switched networks. Chapter 5 presents the study on priority-based routing and comparison between slotted and unslotted networks. It also proposes an optical-electrical hybrid contention-resolution scheme that performs well under self-similar IP traffic, followed by a TCP-layer performance study. Chapter 6 investigates the application of optical packet switching in metropolitan-area networks. Chapter 7 concludes the dissertation with a summary of its results.
Chapter 2

Advances in Optical Packet Switching: An Overview

2.1 Introduction

As telecommunications and computer communications continue to converge, the data traffic is gradually exceeding the telephony traffic. This means that many of the existing connection-oriented, circuit-switched networks need to support packet-switched data traffic. The WDM technology has provided us an opportunity to multiply the network capacity. Current optical-switching technologies allow us to rapidly deliver the enormous bandwidth of WDM networks. Among all the switching technologies, all-optical packet switching appears to be a strong candidate because of the high speed, data rate/format transparency, and configurability it offers. The goal of this chapter is to examine some of the critical issues involved in the design and implementation of optical packet-switched networks. It first discusses the synchronization issues, then the contention resolution and switching strategies, followed by the header and packet format. The chapter concludes by describing some of the proposed optical sub-wavelength switching techniques.

In general, there are two categories of optical packet-switched networks: slotted (synchronous) and unslotted (asynchronous) networks. When individual photonic switching systems form a network, at the input ports of each node, packets can arrive at different times. Since switch fabric can change its state incrementally (set up one input-output connection at a time) or jointly (set up multiple input-output connections together at the same time), it is possible to switch multiple aligned packets together or to switch each packet individually ‘on the fly’. In both cases, bit-level synchronization and fast clock recovery are necessary for packet-header recognition and packet delineation.

In a slotted network, all the packets have the same size. A fixed-size time slot contains
both the payload and the header. The time slot has longer duration than the whole packet to provide guard time. There are a number of studies on slotted networks \cite{17, 18, 19, 20, 21, 22}, since they help to make efficient use of optical fiber delay lines, which are the first choice for buffering in the store-and-forward type contention resolution. Optical buffering consists of fiber loops or delay lines that have a fixed propagation delay equal to a few multiples of the time-slot duration. This requires that all the input packets arriving at the input ports have the same size and to be aligned in phase with one another (see Fig. 2.1).

![Figure 2.1: A generic node architecture of the slotted network.](image)

In an unslotted network, the packets may or may not have the same size. Packets arrive and enter the switch without being aligned. Therefore, the packet-by-packet switching action could take place at any point in time. Obviously, in unslotted networks, the chance for contention is larger because the behavior of the packets is more unpredictable (similar to the contention in the slotted and unslotted ALOHA networks \cite{23}). On the other hand, unslotted networks are more flexible compared with slotted networks, since they are better at accommodating packets with variable sizes.

Before we delve into the details of synchronization schemes and architectures, it would be insightful to first examine the source for delay variation of packets within the network \cite{24}. 
2.1.1 Delay Variation Between Nodes

The time for a packet to travel through a certain distance of the fiber depends on the fiber length and the optical signal’s group velocity. This time varies with the chromatic dispersion and the temperature. When WDM is used, the effect of chromatic dispersion needs to be considered. Chromatic dispersion results in different propagation speed for packets transmitted on different wavelengths; therefore, different propagation delays occur. For example, with a typical fiber dispersion of 20 ps/nm/km (where ps is the time unit for delay variation, nm the unit for wavelength difference, and km the unit for propagation distance), a wavelength variation of 30 nm (consistent with the typical Erbium-Doped Fiber Amplifier’s 1530 - 1560 nm window) and a propagation distance of 100 km, the propagation delay variation would be about 60 ns. If dispersion compensation fibers are used, the above delay variation can be reduced by one order of magnitude.

The packet propagation speed also varies with temperature, with a typical figure of 40 ps/°C/km. 100 km of fiber under temperature variation range of 0 - 25 °C means a delay variation of 100 ns.

The delay variations mentioned above are relatively slow with respect of time; the compensation can be static instead of dynamic (on a packet-by-packet basis).

2.1.2 Delay Variations Inside the Nodes

How much delay each packet experiences inside a node depends on the switch fabric and contention-resolution scheme. In a slotted network which uses fiber delay lines as optical buffers, a packet can take different paths with unequal lengths within the switch fabric. All the considerations given to delay variations in the inter-node links apply here. It is worth mentioning that the fast time jitter (as compared with the slow delay variation mentioned previously) induced by dispersion between different wavelengths and unequal optical paths varies from packet to packet at the output of the switch; therefore, a fast output synchronization interface might be required, depending on the position of the payload and header. Thermal effects are smaller here because the temperature varies more slowly and can be easily controlled within the node.

2.2 Synchronization of Slotted Networks

2.2.1 Impact of Packet Format

Figure 2.2 shows a functional block diagram of the node architecture in a slotted network. A tap splits a small amount of power from the incoming packets for the header reading. If header swapping is necessary, a device such as a filter (in the case of subcarrier multiplexed
header) will extract the header completely. The header-processing circuits recognize a preamble at the beginning of the packet and then read the header information. The circuits also pass the timing information of the incoming packet to the control unit to configure the synchronization stages and the switch fabric. The input synchronization stage aligns packets before they enter the switch fabric. The output synchronization stage, which is not shown in Fig. 2.1, is to further compensate the fast time jitter that occurs inside the node. It may or may not be necessary, depending on the actual packet format and node architecture.

In general, the required resolution of synchronization (how fine we want to tune the position of each incoming packet) depends on the actual packet format, i.e., the size and position of the header, payload, and guard time. The longer the packet is, the more guard times we could put in there without significantly sacrificing link utilization; and longer guard time means less strict requirement for alignment.

With regard to the position of payload, header, and guard times, there are two cases as shown in Fig. 2.3 [24]:

(a) Headers define the beginning of time slots. In this case, since the switch only needs to read the information contained in the header, the position of the whole slot would vary slowly with different propagation delay and local clock frequency drift. The time jitter before and after the payload does not have a strong impact on the synchronization.

(b) A guard time is placed between the header and the beginning of the slot, as well as between the header and the payload. In this case, the consecutive packets coming in from the same link could have different mis-alignments and the fast clock recovery for header reading is necessary on a packet-by-packet basis.
Chapter 2: Advances in Optical Packet Switching: An Overview

8

Figure 2.3: Two possible packet formats that determine the synchronization schemes.

In case (a), the header is precisely at the beginning of the time slot. Therefore, the consecutive packet headers arrive at a node with a constant time interval with respect to the preceding node. Since the header appears exactly at the beginning of each time slot and the control unit only needs to read the header to switch the packets correctly, the packet delineation and electronic control of the input synchronizer are relatively simple to implement. The input synchronizer stage only has to deal with slow delay variations. However, every effort should be made to keep the header well aligned with the slot boundary. In the diagram in Fig. 2.2, the output synchronizer stage should be a fast and high-resolution solution to compensate for the jitter of header and payload occurring in the different optical paths inside the switch fabric on a packet-by-packet basis. In case (b), since there is guard time between the header and time-slot boundary, and header jitter is allowed, the header-reading electronics has to deal with fast clock recovery of jittered header on a packet-by-packet basis. In other words, the switching node cannot predict precisely the exact arrival time of the header, since it only has the knowledge of the time when the slot begins, which is in the middle of a guard time. A fast and high-resolution output synchronization stage becomes optional because the header jitter is taken care of by the header recognizing and reading electronics at the following node’s input synchronization stage.

2.2.2 Bit-Level Synchronization

Packet delineation is essential for both slotted and unslotted networks. During packet delineation, the incoming bits are locked in phase with the clock in order for the node to read the header information. Traditional phase-locking loop approach is not applicable because it requires too many bits. Burst-mode receiver have demonstrated to achieve bit-
level synchronization within nanosecond range \cite{25, 26}. Reference \cite{27} describes a burst-mode receiver setup in which the transmitter frequency multiplexes its bit-clock with the baseband data and modulates the optical carrier with the composite signal. The data and clock travel along the fiber with negligible dispersion. At the receiver end, a photodiode detects the optical signal. Its RF output is amplified and split. The data and clock are first separated by a low-pass filter that cuts off the baseband and then by a narrow bandpass filter centered at the clock frequency. The retrieved clock is fed into an analog receiver. If the delay from the output of the photodiode to the input of the receiver is matched, the clock and data will be in bit-synchronization for all incoming packets.

### 2.2.3 Packet-Synchronization Schemes

Since packets are entering a node from different links, for all the previously stated reasons, they could arrive totally out of phase with one another. Figure 2.4 \cite{28} shows a typical synchronization stage consisting of a series of switches and delay lines, as they appear in the input synchronization stage of a node.

![Figure 2.4: A scheme for input synchronization stage in a node.](image)

Once the header processor recognizes the bit pattern in the header and performs packet delineation, it identifies the packet start time, and the control unit will calculate the necessary delay and configure the correct path through these switched delay lines. The length of the delay lines are in an exponential sequence between the 2x2 switches, i.e., the first delay line is equal to 1/2 time-slot duration, the second delay is equal to 1/4 time-slot duration, etc. The resolution of this scheme is 1/2^n of the time-slot duration, where n is the number of delay lines. This type of synchronization scheme is suitable for both static (slow) and dynamic (fast) synchronization. At the system initialization, the synchronization is setup to compensate for delay variations between different inputs and to keep this configuration throughout the system operation time (static). For the packet-based (dynamic) synchronization, much faster switches are necessary to operate during the guard time.

From a physical point of view, such a packet-synchronization scheme introduces insertion loss and crosstalk due to the switches used. Cascading the switches will inevitably require
optical amplification, which will result in degraded signal-to-noise ratio. Meanwhile, the
crosstalk accumulated through the switches will also increase the bit-error rate. In a multi-
node network, the power penalty brought by all the synchronization stages may significantly
impair the system performance.

2.3 Unslotted Networks

Figure 2.5 shows the general node architecture and packet behavior for unslotted networks.
Note the absence of synchronization stages and packet alignment. The fixed-length fiber
delay lines hold the packet when header processing and switch reconfiguration are taking
place. There is no packet-alignment stage and all the packets go through the same amount of
delay with the same relative position in which they arrived, provided there is no contention.
The unslotted networks circumvent the requirement of synchronization stages. However,
given the same traffic load, the network throughput in an unslotted network is a bit lower
than that in the slotted network because contention is more likely to occur.

![Figure 2.5: A generic node architecture of the unslotted network.](image)

2.4 Contention Resolution

In a packet-switched network, each packet often has to go through a number of switching
nodes to reach its destination. When the packets are being switched, contention occurs
whenever two or more packets are trying to leave the switch from the same output port,
on the same wavelength, at the same time. How the contention is resolved may have a
large effect on the network performance. Here, we will examine three types of contention resolutions: optical buffering, deflection routing, and wavelength conversion.

### 2.4.1 Optical Buffering

In electronic switches, contention is usually resolved by the store-and-forward technique, which means that the switch will first store contending packets in a queue and send them out at a later time when the output port becomes available. This is possible because of the availability of electronic random-access memory (RAM). In an optical switch, there is no ready-to-use random-access optical memory. An optical buffer can only consist of optical delay lines. The main difference between electronic RAM and an optical buffer is that optical buffers typically utilize fiber delay lines and provide sequential access. Once a packet has entered the fiber, it must emerge from the other end after a fixed amount of time; there is no way to retrieve the packet any time earlier (except for re-circulation fiber loops, which will be discussed later).

There are various designs of node architectures applying optical buffer and there are different ways to categorize them. One way is to compare them to the buffering in electronic switches (input, output, shared, and re-circulating buffering). There is also a simpler, more direct way: single-stage buffering or multi-stage buffering, forward buffering or feedback buffering. (A stage is a single continuous piece of delay line.)

The first example, proposed in the European ACTS KEOPS (Keys to Optical Packet Switching project) \[21\], is a broadcast-and-select space switch using a single-stage forward-buffering scheme for contention resolution (Fig. 2.6). The wavelength converters encode the packet streams entering each input; therefore, the packets on each input are distinguished by a separate wavelength. The streams are then combined by a wavelength multiplexer and distributed to \(k\) groups of delay lines of different lengths, which give the packets necessary delays to resolve contention. With the use of semiconductor optical amplifier (SOA) gates and passive couplers, each output port is able to select the packets with proper delays. At the final stage, the wavelength demultiplexer, SOA gates, and multiplexer can select one packet from a specific input port. In this architecture, there is only one stage of buffer and each delay line feeds forward to the next part of the switch. Since each packet is broadcast to all delay lines and every output port, it is possible to offer multicast operation and also implement packet priorities. The drawback of this architecture is the use of a large number of components and controls, which considerably increases the cost. For example, one switch needs \(n\) wavelength converters, \(n \cdot k + n^2\) SOA gates, and \(2n + 1\) wavelength multiplexers/demultiplexers; where \(n\) is the number of input ports.
The second example is the Shared-Memory Optical Packet (SMOP) Switch [18], which belongs to the single-stage feedback category. Figure 2.7 shows its principle of function. The lengths of the delay lines could be 1, 2, 3, ..., m packet duration. The \((n + m) \times (n + m)\) space switch can switch a packet either directly to an output port or to one of the delay lines, according to how much delay the packet needs. Delay lines of length greater than one packet duration significantly reduce the number of re-circulation loops needed, resulting in the reduced need for amplifiers and in less noise. This scheme also allows prioritized packet switching since a lower-priority packet may be preempted by being sent to another circulation. Since the number of re-circulations a packet will take is unpredictable, some packets could suffer more power loss than others, making optical amplification necessary. This will inevitably introduce additional signal-to-noise ratio (SNR) degradation to the re-circulating packets.

Another case of single-stage feedback optical buffering is the Fiber Loop Memory Switch introduced in the Asynchronous Transfer Mode Optical Switching (ATMOS) project [20] of Research and development in Advanced Communications in Europe (RACE), as shown in Fig. 2.8. The buffer is a fiber-loop delay line containing multiple wavelength channels. When contention occurs, the input packet is converted to one of the available wavelengths in the loop and kept circulating by activating the corresponding passive fixed filter (i.e., by turning on the related SOA gate). At the input of the loop, half of the power enters...
the loop, and the other half goes toward the outputs through the passive coupler. When
the contention is resolved, the packet travels to the destination link with the proper tuning
of the corresponding output tunable filter. At the same time, the passive filter in the
loop is turned off to erase the packet in the buffer. It is possible for incoming packets
to preempt those that are already waiting; hence, this type of switch can also implement
packet priorities.

Figure 2.7: The shared-memory optical packet switch.

Figure 2.8: Fiber-loop-memory switch from ATMOS.
For multi-stage feed-forward buffering examples, there are several node architectures applying cascaded 2x2 switching elements containing optical buffer [19] (Fig. 2.9). Each of these switching elements provides buffering delay of one or more packet duration in case of contention. A larger switch fabric can be constructed by cascading a number of these 2x2 elements in, for example, a Banyan configuration.

![Figure 2.9: A 2x2 switching element containing optical buffer.](image)

There are various designs of optical buffering, such as the staggering switch [17], Switched Fiber Delay Lines (SDL) (CORD: Contention Resolution by Delay Lines) [29], Switch with Large Optical Buffers (SLOB) [30], etc. Packet-loss rate, network latency, hardware cost, control-circuit complexity, and packet reordering are among the many important issues to be considered in the design, which depends on the network specification, i.e., network dimension, topology, traffic load and pattern, etc.

### 2.4.2 Deflection Routing

Optical buffering is, to a great extent, an imitation of their electronic network counterparts. In an electronic network, the link bandwidth is much less than today’s optical fiber’s capacity, and increasing the link utilization is the main goal of electronic buffering. In a network deploying optical buffers, each packet will normally arrive at its destination along the shortest path, and the expected number of hops is minimized. Implementing optical buffers involves a significant amount of hardware and complex electronic controls. Another issue that arises with optical buffer is that the optical signal suffers from power loss in the switch elements (optical buffering causes the packet to be switched more times), and optical amplifiers may be necessary. The accumulated noise from the cascaded amplifiers can limit the network size at very high bit rates, unless signal regeneration is applied. In deflection routing, as the name implies, contention is resolved as follows: if two or more packets need to use the same output link to achieve minimum distance routing, then only one will be routed along the desired link, while others are forwarded on paths which may lead to a
non-shortest-path route. Hence, for each source-destination pair, the number of hops taken by a packet is no longer fixed. Deflection routing does not necessarily exclude the use of optical buffers. The simplest form of deflection routing is hot-potato routing [31], which is a special case where buffers are not provided at all.

Most studies on deflection focus on regular network topologies with uniform traffic load. These logical topologies can be setup over different physical topologies such as ring, star, or mesh. Figure 2.10 shows two typical logical topologies used for network performance simulation: the Manhattan Street Network (MSN) [32] and ShuffleNet [33]. Each node in these two topologies has two input ports and two output ports. Figure 2.11 shows example node architectures for MSN and ShuffleNet.

A comparison between store-and-forward and hot-potato routing with the ShuffleNet topology shows that the average number of hops for each packet is larger for hot-potato routing, because not all the packets take the shortest route toward their destinations [31].

---

Figure 2.10: The Manhattan Street Network and ShuffleNet.

Figure 2.11: Example of node architecture for hot-potato routing or deflection with limited buffer routing in MSN and ShuffleNet.
As the number of users (or number of nodes) increases, both the average number of hops and aggregate capacities increase for both routing strategies. In multi-hop networks, where information from a source node to a destination node may be routed through intermediate nodes\cite{1}, only a portion of the network capacity is used for newly-generated traffic. A certain amount of network capacity is taken up by “bypassing traffic” as packets hop from one node to another to reach their destinations. The overall capacity of the network is inversely proportional to the average number of hops, and proportional to the number of nodes and the capacity of each link between two nodes. Store-and-forward routing can improve the network utilization as the number of nodes increases. It has also been shown that, even for networks containing several thousand nodes, the aggregate capacity of hot-potato routing is not worse than 25% of that for store-and-forward routing\cite{34}. A more intuitive explanation is that, in hot-potato routing, the nodes use the entire network as a big buffer and route the packet in contention to the rest of the network. This type of routing trades off network utilization for simpler hardware implementation.

In MSN and ShuffleNet, there are three characteristics that determine the performance of the network with deflection routing:

- Diameter. It is the maximum distance in number of hops between any node pair in the network. The diameter is a good indicator of how compact a network is.

- Deflection Cost. This is the maximum increase in path length in number of hops due to a single deflection.

- Don’t care nodes. For a given destination, any node that has both its output links as part of the shortest path is a don’t care node.

A high percentage of don’t care nodes helps keep the number of deflections to a low level even at high loads. The performance of ShuffleNet is better because the initial advantage in the diameter of ShuffleNet over MSN is preserved under heavy traffic by the high percentage of don’t care nodes. The expected number of hops depends on the routing algorithm. For store-and-forward routing with infinite buffers, the average number of hops is a minimum, since the packets always take the shortest path to their destinations. However, the queuing delay could diverge to infinity when the network approaches saturation, i.e., when probability of packet generation in each time slot approaches 1. For deflection routing, the average number of hops becomes an increasing function of link load and the throughput is, therefore, lower than that with store-and-forward routing.

So far, the routing strategies described above have assumed slotted (synchronous) network operation, which involves complex and expensive packet-alignment schemes. As for the unslotted networks, they suffer from severe congestion as the offered load increases,
and its throughput would collapse when the load exceeds a certain threshold. The reason is that, with increasing congestion, there are more and more deflected packets "wandering" around in the network and they further lower the network capacity to process newly-generated packets; meanwhile, more packets are being generated. The whole scenario forms a positive-feed-back cycle and as a result, the network throughput will collapse. To avoid such saturation, the number of hops a packet traverses has to be monitored and kept under a maximum value.

One way to improve the network throughput and eliminate congestion is to provide optical buffering to such unslotted deflection networks. The corresponding performance improvement is very encouraging in the high-load region, as shown in later chapters. Also the congestion is highly reduced with one re-circulating loop. One practical concern of the asynchronous deflection routing with limited number of optical buffer is the number of times a packet is allowed to re-circulate in the loop. Optical amplification imposes noise to the signal. Network latency also increases with the number of circulations. Therefore, it is necessary to establish an upper bound on the number of re-circulation for the packets.

Monitoring the number of hops a packet has taken is essential to mitigate looping problems, which can be caused by too many packets "wandering" in the network. Having a \textit{time-to-live} (TTL) field in the packet header (like in IP packets) is one way to keep track of the hop count of a packet. An alternative is to have the source node put a time stamp on each packet and the intermediate nodes compare it with a global clock when the packet is in transit.

Deflection routing plays a prominent role in many optical network architectures, since it can be implemented with none or modest amount of optical buffering. Asynchronous (unslotted) deflection routing combined with limited buffering can help avoid complex synchronization schemes and provide good performance with careful design.

\subsection{Wavelength Conversion}

Optical buffering and deflection routing are both deflection in general, one in the time domain and the other in the space domain. With today’s enabling WDM technology, the wavelength domain presents us one more dimension to solve the contention problem. Both buffering and deflection have their advantages and disadvantages: buffering offers a better network throughput but involves more hardware and controls, deflection is easier to implement but can offer only a limited improvement in network performance. When combined with wavelength conversion, their disadvantages could be overcome or minimized.

In a switching node applying wavelength conversion and buffering, the input stage de-multiplexes wavelength channels and the wavelength converters locate available wavelengths for certain output ports. The space switch selects the appropriate output port.
Chapter 2: Advances in Optical Packet Switching: An Overview

An asynchronous network can incorporate wavelength conversion, optical buffering, and deflection. An example of an optical packet switch without packet alignment is described in [35]. It is very similar to the broadcast-and-select architecture proposed in the KEOPS project, except that it is for WDM operation with wavelength conversion. Since the switch is made of optical gates and it is fully non-blocking, it can be configured incrementally, making the architecture ready for asynchronous operation.

Wavelength conversion can reduce the number of optical buffers and the packet-loss probability. When nodes are provided with a number of optical receivers/transmitters equal to the number of wavelengths, hot-potato routing in conjunction with wavelength conversion becomes an interesting option for mesh topologies such as Manhattan Street network and ShuffleNet [36].

Another advantage of wavelength conversion is that it provides noise suppression and signal re-shaping. In a network with only a small number of wavelengths, buffering might be more desirable. In a network with a large number of wavelengths and full wavelength conversion, buffers may not be necessary.

There are several possible combinations of optical buffering, wavelength conversion, and deflection routing. The following chapters will present the research work to understand which scheme offers a low end-to-end delay, low packet-loss rate, high network throughput, etc.

2.5 Header and Packet Format

In an electronic network, the transmitter transmits the packet header serially with payload data at the same data rate, e.g., IP packets and ATM cells. Electronic routers or switches process the header information at the same data rate as the payload. In an optical network, the channel bandwidth is much higher. A typical wavelength channel has the line speed of 10 Gbps (OC-192). Although there are various techniques to detect and recognize packet headers at Gbps speed, either electronically or optically, it is costly to implement electronic header processors operating at such high speeds [37].

Among several proposed solutions, packet switching with subcarrier-multiplexed (SCM) header is attracting an increasing amount of interest [38]. In this approach, the header and payload data are multiplexed on the same wavelength (optical carrier). Figure 2.12 shows that, in the current that modulates the laser transmitter, payload data is encoded at the baseband frequency, while header bits are encoded on a properly-chosen subcarrier frequency at a lower bit rate. A conventional photodetector can detect header information on different wavelengths by detecting a small fraction of the light in the fiber, without any type of optical filtering. Figure 2.13 illustrates that, in the output current of the
photodetector, various data streams from different wavelengths jam at baseband, but the subcarrier remains distinct and the header can be retrieved by an electrical filter.

![Power spectrum of laser modulation current](image1.png)  
**Figure 2.12:** Power spectrum of the laser modulation current.

![Header retrieval in SCM](image2.png)  
**Figure 2.13:** Header retrieval in SCM.

A nice feature of subcarrier-multiplexed header is that the header can propagate in parallel with the payload data and it can take up the whole payload transmission time. However, subcarrier multiplexing also has its limit. Since today’s optical transmitters can easily operate at 10 Gbps and beyond, the upper-bound frequency limit of the available
RF components imposes a constraint to the number of subcarrier channels. Because the minimum subcarrier spacing cannot be less than twice the header bit rate, it is desirable to have a lower header bit rate so that there can be more subcarrier channels. On the other hand, if the header bit rate is too low, it would take longer to transmit the header and the header bandwidth overhead may be non-negligible.

In many of the routing and switching protocols, headers have to be updated at each node. There are several approaches proposed on optical same-wavelength header replacement for headers transmitted serially with the payload data stream [39, 40]. The switching node can perform header swapping by blocking the old header with a fast optical switch and inserting the new header, generated locally by another laser, at the proper time. If the network is WDM-based, it is important that the new header is precisely at the same wavelength as the payload data.

An alternative approach is to update the header by transmitting the payload and the header on separate wavelengths, demultiplex the header for optoelectronic conversion, and electronically process and re-transmit on the header wavelength. This approach suffers from fiber dispersion, which separates the header and the payload as the packet propagates through the network. Subcarrier-multiplexed headers have less dispersion problems since they are very close to the baseband frequency. SCM header could be removed by narrow-band optical filters, but it would be very sensitive to wavelength drift. Previous practical SCM header-replacement schemes are limited to full optoelectronic conversion of the entire packet followed by electronic filtering, re-modulation, and re-transmission on a new laser. Reference [41] proposed a technique to update the SCM header with simultaneous wavelength conversion of baseband payload using SOAs. It involves a two-stage process: first, simultaneous SCM header suppression and wavelength conversion of the baseband payload is achieved due to the low-pass frequency response of cross-gain modulation in the SOA’s; then, header replacement is achieved by optically re-modulating the wavelength-converted signal with a new header at the original subcarrier frequency.

Packet length is another issue of concern for network designers. A short packet might not give good throughput because a larger percentage of the bandwidth is now given to the header or guard time between the header and the payload. On the other hand, a very long packet would need longer optical buffers, and it may not provide a granularity that is fine enough for IP traffic. Also, re-transmission of a long packet will take much longer. From a physical point of view, balancing the packet-error rate (PER) between payload and header is important [42]. PER is different from bit-error rate (BER); PER is the probability of the entire packet received in error. PER increases with BER and the number of bits contained in the packet. For efficient network operation, the PER for payload and header should be approximately the same, in order to deliver the packets as successfully as possible [43].
Chapter 2: Advances in Optical Packet Switching: An Overview

Payload usually contains many more bits than the header. If the header is updated at every traversed node, the bits in the payload will suffer more physical impairment than the bits in the header. Another fact is that, with SCM, the header is usually transmitted at a lower bit rate than the payload data. All these facts lead to a big advantage of lower BER for header bits over the payload bits. Therefore, it is necessary to optimize the amount of power distributed between the header and the payload to achieve a balanced PER for payload and header at the destination node.

2.6 Advances Toward Optical Sub-Wavelength Switching

2.6.1 Optical Label Switching

Reference [44] documented the first demonstration of computer-to-computer communication using an optical packet-switching mechanism, namely Optical Label Switching, implemented in a testbed at Telcordia. This approach adopts the unslotted operation and uses SCM for optical header (optical label) encoding. The demonstrated network utilizes deflection routing, wavelength conversion, and a preemption scheme based on packet priority. It also proposes an optical single sideband (OSSB) subcarrier header erasing and replacement technique [45]. References [46, 47] also documented successful SCM-header extraction and replacement. An optical label switching technique based on fast tunable wavelength converter and Array Waveguide Grating (AWG) was demonstrated in [48].

2.6.2 Photonic Slot Routing

Another recently proposed approach to optical packet switching is Photonic Slot Routing (PSR) [49, 50, 51]. In PSR, time is divided into “photonic” slots. All the wavelength are slotted synchronously. Each photonic slot, across all the wavelengths, has a preassigned destination; therefore, all the packets transmitted in one photonic slot on different wavelengths are routed together through the same path. By requiring the packets to have the same destination, the photonic slot is routed as a single entity, without the need for demultiplexing individual wavelengths at intermediate nodes. Thus, each node only needs wavelength-insensitive components, resulting in less complexity, faster routing, and lower network cost [49, 52]. PSR reduces the optical hardware complexity by shifting the burden to the electrical-buffer management; therefore, most of the intelligence of PSR resides in the design of the access-control protocols.

In a PSR mesh network, each node consists of a wavelength-insensitive optical packet switch, optical delay lines, and electronic buffers (for ingress packet buffering). Each node (while acting as a source) maintains a separate queue for each destination node. A diagram
of the node architecture is shown in Fig. 2.14.

A photonic slot spans across all of the WDM channels wavelengths in the network, and each wavelength can carry a single packet. At the input fiber, a wavelength-selective drop device (such as a filter) extracts the header of each photonic slot, which may be carried on a separate wavelength. The header contains information such as the destination of the slot, and which wavelengths in the slot are occupied by packets. The destination information is used to determine/configure the switch setting so that the slot can be appropriately routed toward its destination according to a standard routing algorithm. The slot-occupancy information determines which wavelengths are still available in the current slot. Please see [49, 50, 51] for more details on PSR and its performance characteristics.

![Figure 2.14: PSR node architecture.](image)

### 2.6.3 Optical Burst Switching

Optical burst switching (OBS), like PSR, is another approach that attempts to shift the computation and control complexity from the optical domain to the electrical domain (at the edge nodes). OBS has a switching granularity between a circuit and a packet. Burst switching was first proposed for voice and data communications in the early 1980s [53]. It was introduced to optical networks in the mid 90’s [54, 55, 56]. In an IP-over-WDM
context, a burst formed at the network edge contains a number of IP packets and can be of
tens of kilobytes to a few megabytes long. In OBS, a control packet, separate from the data
burst that carries the payload, is sent first. The switch control units along the path will
configure each switch (set up and reserve certain input-output ports connection) according
to the information carried in the control packet as it propagates along. Shortly after the
control packet leaves the source node, the optical burst leaves the source node (without any
acknowledgment about the reservation along the switches on the path) and will go through
the configured switches. Compared with optical packet switching, OBS eliminates the need
for optical buffering of payload while the header is processed by pushing the buffering
function to the edge nodes where electrical buffer is available.

Figure 2.15 shows a typically OBS node architecture. The control packet which precedes
the data burst resides on a separate control wavelength. The switch control unit controls
the burst injection and establishes an internal all-optical path inside the switch according to
the control packet. Once the switch is configured, the data burst will arrive and go through
the switch without any extra latency.

Similar to PSR, the intelligence of OBS is mostly implemented in the burst-transmission
algorithm and protocol rather than the switch architecture itself. Two important aspects in
the design of OBS protocol are (1) the relation between the control packet and the payload data burst, and (2) the duration or size of the burst and how the configured switch state is released. When bandwidth is reserved in the switch, the control packet may or may not indicate the duration of the burst. If the burst duration is not known to the switch, the reservation will not be released until the switch is explicitly told to do so. This can be done by sending another release control packet, or by detecting the end of a burst by the switch itself. The importance of the offset between control packet and data burst is best demonstrated in a proposed protocol called Just-Enough-Time (JET) [55]. In JET, the source node is assumed to have knowledge of the explicit route for the data burst, and the control-packet processing time $\delta$ at each switch (Fig. 2.16). The offset between data burst and control packet should be no less than the total processing delay of the control packet for all the switches along the routing path. This ensures that any switch along the path will have been configured when the burst arrives. Also, since the burst does not arrive immediately after the control packet is processed, the switch does not need to establish the required input-output connection until the burst actually arrives. This delayed reservation makes more efficient use of the bandwidth at the expense of delayed setup time.

Figure 2.16: Offset time and delayed reservation in the JET protocol.
2.7 Summary

It is impossible to cover every aspect of optical packet switching in one chapter. This topic involves routing, synchronization, contention resolution, header format/updating, switch fabrics, physical impairment of the transmission, network control, protocol, etc. Only a handful of the important aspects were covered here. Optical packet switching is promising because it offers a much higher capacity and data transparency, despite the fact that it lacks sophisticated queuing mechanism. It is our belief that optical packet switches will be deployed in the future networks through technical breakthroughs, clever network design, and making optimal use of optics and electronics wherever they fit best.
Chapter 3

Contention-Resolution Schemes in Optical Packet-Switched Networks

3.1 Introduction

In recent years, WDM technology has brought fundamental changes to the network design [1, 2]. Previous studies have investigated wavelength-routed networks, in which lightpaths are set up with the bandwidth of full wavelength channels [3, 4]. The current ongoing efforts to automate and expedite wavelength and bandwidth provisioning in the optical layer indicate the inevitable trends that lead to more intelligent optical networks [7]. Migration of the switching functionality from electronics to optics will eventually remove the incumbent layers that impose the optical-electrical-optical (O-E-O) conversion and additional signal processing requirements. Rapid increases in the data traffic suggest that an optical WDM networking technology capable of switching at sub-wavelength granularity is attractive for meeting diverse traffic demands of the next-generation networks.

A number of research groups have reported various optical sub-wavelength switching approaches [8, 9]. Among these approaches, optical packet switching is attracting increasing attention. It can seamlessly interface with the WDM transport layer and potentially offer a number of layer-two functions (e.g., similar to the virtual circuit in ATM networks, or label-switched path in Multiprotocol Label Switching (MPLS) networks) and layer-three functions (e.g., similar to packet-by-packet routing in IP networks) via packet-level switching in the optical domain [10]. Packet switching is capable of dynamically allocating network resources with fine granularity while it offers excellent scalability. Optical packet switching is a candidate transport solution to close the gap between the electrical (IP/MPLS) layer and the optical (WDM) layer [11, 12, 57, 58, 59].

In an optical packet-switched network, contention occurs at a switching node whenever
two or more packets try to leave the switch fabric on the same output port, on the same wavelength, at the same time. In electrical packet-switched networks, contention is resolved with the store-and-forward technique, which requires the packets losing the contention to be stored in a memory bank, and to be sent out at a later time when the desired output port becomes available. This is possible because of the availability of electronic random-access memory (RAM). There is no equivalent optical RAM technology; therefore, the optical packet switches have to adopt different approaches for contention resolution. Meanwhile, WDM networks provide one new additional dimension, namely wavelength, for contention resolution. This chapter explores all three dimensions of contention-resolution schemes: wavelength, time, and space. Specifically, this chapter proposes a combinational scheme based on an integrated optical packet router architecture, which demonstrates excellent network performance. The contention-resolution mechanism of the three dimensions are outlined below.

- Wavelength conversion offers effective contention resolution without relying on buffer memory [35, 60, 61, 62, 63]. Wavelength converters can convert wavelengths of packets which are contending for the same wavelength of the same output port. It is a powerful and the most preferred contention-resolution scheme (as this study will demonstrate) that does not cause extra packet latency, jitter, and re-sequencing problems.

- Optical delay line (which provides sequential buffering) is a close imitation of the RAM in electrical routers, although it offers fixed and finite amount of delay. Many previously proposed architectures employ optical delay lines to resolve the contention [17, 21, 22, 29, 30, 64, 65, 66]. However, since optical delay lines rely on the propagation delay of the optical signal in silica to buffer the packet in time, i.e., due to their sequential access, they have more limitations than the electrical RAM. To implement large buffer capacity, the switches need to include a large number of delay lines.

- The space deflection approach [31, 34, 67, 68, 69] is a multiple-path routing technique. Packets that lose the contention are routed to nodes other than their preferred next-hop nodes, with the expectation that they will eventually be routed to their destinations. The effectiveness of deflection routing depends heavily on the network topology and the offered traffic pattern.

Both wavelength conversion and optical buffering require extra hardware (wavelength converters and lasers for wavelength conversion; fibers and additional switch ports for optical buffering) and control software. Deflection routing can be implemented with extra control software only.
Chapter 3: Contention-Resolution Schemes in Optical Packet-Switched Networks

With an orthogonal classification, optical packet-switched networks can be divided into two categories: time-slotted networks with fixed-length packets [24] and unslotted networks with fixed-size or variable-size packets [35, 65]. In a slotted network, packets of fixed size are placed in time slots. When they arrive at a node, they are aligned before being switched jointly [21]. In an unslotted network, the nodes do not align the packets and switch them one by one ‘on the fly’; therefore, they do not need synchronization stages. Because of such unslotted operation, they can switch variable-length packets. However, unslotted networks have lower overall throughput than slotted networks, because of the increased packet-contention probability [68]. Similar contrast exists between the unslotted and slotted version of the ALOHA network [23]. Due to the lack of viable optical RAM technologies, all-optical networks find it difficult to provide packet-level synchronization, which is required in slotted networks. In addition, it is preferred that a network can accommodate natural IP packets with variable lengths. Therefore, this chapter primarily focuses on optical contention-resolution schemes in unslotted, asynchronous networks.

Previous studies have concentrated on one or two-dimensional contention-resolution schemes in optical packet-switched networks. There has been no report on unified contention resolution across wavelength, time, and space domains for irregular mesh networks. The goal of this study is to design an effective contention-resolution scheme that incorporates all three dimensions, and to examine the performance of various schemes. It is shown that, with careful planning of optical resources according to network parameters (e.g., topology, traffic pattern, etc.), one can design an optical packet-switching node architecture that offers excellent network performance.

Section 3.2 describes the node architectures for a number of contention-resolution schemes. Section 3.3 presents simulation experiments and performance results of these schemes, including a comparison of slotted and unslotted networks, and it explains the network behavior observed under these schemes. It also proposes setting the number of optical buffers according to the node connectivity, and this method provides higher performance without imposing a large number of optical buffers for every node. Section 3.4 concludes the chapter.

3.2 Contention-Resolution Schemes

Contention-resolution schemes have different effects in slotted and unslotted networks. In a slotted network, all packets have the same length. Each slot contains a guard time and the fixed-size packet which consists of header and payload. Most optical-buffering architectures utilize delay lines of fixed propagation delay equal to some multiple of the time-slot duration. Such architectures in the context of slotted networks will require synchronization of packets at the input of the switch fabric.
Unslotted networks utilize packets of variable lengths, and require no synchronization of packets at the input of the switching fabric. All packets experience the same amount of delay with the same relative position in which they arrived before they enter the switch.

Figure 3.1 shows the generic node architectures utilizing contention-resolution schemes in time, space, and wavelength domains. Every node has a number of add/drop ports, and this number will vary depending on the nodal degree. Each add/drop fiber ports will correspond to multiple client interfaces reflecting multiple wavelengths on each fiber. Each client interface-input (-output) will be connected to a local transmitter (receiver). Different contention-resolution schemes give rise to different architecture.

![Node architectures for different contention-resolution schemes](image)

Figure 3.1: Node architectures for different contention-resolution schemes: (a) buf; (b) bufwdm; (c) wc; (d) wc+bufwdm.

### 3.2.1 Optical Buffering

Optical buffering utilizes one or more optical fiber delay line looping the signal from the output back to the input of the switch fabric. Figures 3.1(a) and 3.1(b) illustrate opti-
Chapter 3: Contention-Resolution Schemes in Optical Packet-Switched Networks

3.2.2 Wavelength Conversion

Figures 3.1(c) and 3.1(d) show contention resolution utilizing wavelength conversion, where the signal on each wavelength from the input fiber is first demultiplexed and sent into the switch, which is capable of recognizing the contention and selecting a suitable wavelength converter leading to the desired output fiber. The wavelength converters can operate with a full degree (i.e., they can convert any incoming wavelength to a fixed desired wavelength) or with a limited range (i.e., they can convert one or several pre-determined incoming wavelengths to a fixed desired wavelength).

The majority of wavelength-conversion techniques demonstrated to date are for one single wavelength channel. Parametric wavelength conversion is a promising technique offering multi-channel wavelength conversion without measurable crosstalk [70]. Furthermore, this conversion mechanism can scale well without a large number of wavelength converters by virtue of limited multi-channel wavelength conversion [71]. The nonlinear interaction in the device results in generation of the converted signal at a wavelength corresponding to the frequency difference between the pump and signal waves \( f_2 = f_P - f_1 \). In Fig. 3.2(a), a pump laser at frequency \( f_P \) converts signals between \( f_1 \) and \( f_4 \), centered around the frequency mirror of \( f_P \) (namely \( f_P/2 \), which corresponds to half of the pump’s frequency) located between \( f_1 \) and \( f_4 \). The same pump can also convert between \( f_2 \) and \( f_3 \). Figure
3.2(b) shows that two pump lasers, \( f_{P1-2} \) and \( f_{P1-3} \), convert signals between \( f_1/f_2 \) and \( f_1/f_3 \).

![Diagram](image-url)

3.2.3 Space Deflection

Space deflection relies on another neighboring node to route the packet when contention occurs. As a result, the node itself can adopt any node architecture in Fig. 3.1. Space deflection resolves contention at the expense of the network capacity and the switching capacity of another node. Obviously, this is not the first choice among contention-resolution schemes in three dimensions. As the later sections will reveal, the node will seek wavelength-domain contention resolution first, time-domain contention resolution second, and space-domain contention resolution third. In practice, the contention resolution will often employ a combination of time-, space-, and wavelength-domain contention resolutions. For instance, the contention resolution may involve buffering in time and wavelength, in which case wavelength conversion may precede time buffering. When contention resolution in all three dimensions fail, the packet will be dropped.

In most deflection networks, certain mechanisms have to be implemented to prevent looping (a packet being sent back to a node it has visited before), such as setting a maximum
hop count and discarding all the packets that have passed more hops than this number. This is similar to the time-to-live (TTL) mechanism for routing IP packets.

### 3.2.4 Combinational Schemes

By mixing the basic contention-resolution schemes discussed so far, we propose combinations of these three approaches. For example, Fig. 3.1(d) shows the node architecture for wavelength conversion combined with multi-wavelength buffering. Note that a packet can be dropped at any node under all of these schemes due to (1) unavailability of a free wavelength at the output port, (2) unavailability of a free buffer, and/or (3) the fact that the packet may have reached its maximum hop count.

We define the notations for these schemes as follows:

- **baseline**: No contention resolution. Packet in contention is dropped immediately. The *baseline* case is being studied here for purposes of performance comparison.

- **$N_{buf}$, $N_{bufwdm}$**: Buffering. The node has $N$ delay lines, and each delay line can take one or multiple wavelengths at a time.

- **def**: Deflection.

- **wc**: Wavelength conversion with full conversion range.

- **wclimC**: Limited wavelength conversion. One wavelength can be converted to other $C$ fixed wavelengths. When parametric wavelength conversion is used, the chosen pump frequencies can operate between multiple wavelength pairs. For example, when $C = 2$, we can set a pump frequency whose half value lies in the center between the frequencies of $\lambda_1$ and $\lambda_{16}$ (therefore the same pump laser can be used to convert between $\lambda_1$ and $\lambda_{16}$, $\lambda_2$ and $\lambda_{15}$, $\lambda_3$ and $\lambda_{14}$, and so on, provided the wavelengths are equally spaced). A second pump frequency can be placed such that its half value lies in the middle between the frequencies of $\lambda_1$ and $\lambda_8$, and it can also convert signals between $\lambda_2$ and $\lambda_7$.

In the following section, these notations indicate the different approaches and their priorities. For example, $4wav_{wc}+16buf+def$ means a combination of full-range wavelength conversion, single-wavelength buffering with 16 delay lines, and deflection in a 4-wavelength system. The order of contention resolution is as follows: A packet that loses the contention (it fails to occupy the preferred output port) will first seek a vacant wavelength on the preferred output port. If no such wavelength exists, it will seek a vacant delay line. If no delay line is available, it will seek a vacant wavelength on the deflection output port. When all of
the above options fail, the packet will be dropped. This contention resolution order provides the best performance in terms of the packet-loss rate and the packet delay.

3.3 Numerical Results and Performance Comparison

3.3.1 Simulation Experiments and Configuration

Network Topology

For purposes of illustration, the network topology under study is a part of a telco’s optical mesh network, as shown in Fig. 3.3. Each link $i$ is $L_i$ km long and consists of two fibers to form a bi-directional link. Every fiber contains $W$ wavelengths, each of which carries a data stream at data rate $R$. Each node is equipped with an array of $W$ transmitters and an array of $W$ receivers that operate on one of the $W$ wavelengths independently at data rate $R$. By default, all packets are routed statically via the shortest path.

![Figure 3.3: An example mesh network topology.](image)

Deflection Routing

It is worth mentioning that, because some nodes have many neighbors, the deflection scheme should avoid ‘blind deflections’. In the example network, deflection is carried out only in hub nodes, which have more than two incoming/outgoing links and serve as major routing nodes. Nodes 1, 6, 7, and 13 in Fig. 3.3 are such hub nodes. Most of the routing is done by these four hub nodes, and it is possible to have a deflection policy, with which a packet
would only be deflected to a node that can eventually lead to the packet’s original ‘next-hop’ node with no more than two extra hops in this example network. If no such node exists, the packet will be dropped. Deflection routing requires the node to have a second entry (or perhaps a third entry as well) in the routing table that contains preferred deflection ports for each destination.

For example, in Fig. 3.3, if a packet from node 3 is destined for node 9 via node 1, and there is contention at the output port in node 1 leading to node 7, the packet will be deflected to the port leading to node 11; if the port leading to 11 is also busy, the packet will be dropped immediately instead of being deflected to node 13 or node 6. Table 3.1 shows an example of the deflection entries in the routing table at node 7.

Table 3.1: Deflection table of node 7. Note that packets whose next hop is node 9 will be dropped instead of deflected to node 1, because it would take more than two extra hops to reach node 9.

<table>
<thead>
<tr>
<th>Next-hop node</th>
<th>Deflect to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>Drop</td>
</tr>
</tbody>
</table>

**Packet Generation**

One of the important requirements for the simulation experiment is to model the network traffic as close to reality as possible. One main characteristic of Internet traffic is its burstiness, or self-similarity. It has been shown in the literature that self-similar traffic can be generated by multiplexing multiple sources of Pareto-distributed ON/OFF periods. In the context of a packet-switched network, the ON periods correspond to packet trains, and OFF periods are the periods of silence between packet trains [72].

The probability density function (pdf) and probability distribution function (PDF) of the Pareto distribution are:

\[
p(x) = \frac{\alpha \cdot b^\alpha}{x^{\alpha+1}} \tag{3.1}
\]

\[
P(x) = \int_b^x \frac{\alpha \cdot b^\alpha}{x^{\alpha+1}} \, dx = 1 - \frac{b^\alpha}{x^\alpha} \tag{3.2}
\]

where \( \alpha \) is the shape parameter (tail index), and \( b \) is minimum value of \( x \). When \( \alpha \leq 2 \), the variance of the distribution is infinite. When \( \alpha \geq 2 \), the mean value is infinite as well.
For self-similar traffic, $\alpha$ should be between 1 and 2. The Hurst parameter $H$ is given as $H = (3 - \alpha)/2$.

Since $0 < P(x) \leq 1$, the value of $x$ can be generated from a random number $\mathcal{RND}$ with the range $(0, 1)$:

$$\frac{b^{\alpha}}{x^{\alpha}} = \mathcal{RND}$$

$$x = b \cdot \left( \frac{1}{\mathcal{RND}} \right)^{\frac{1}{\alpha}}$$

(3.3)  (3.4)

The mean value of Pareto distribution is given by:

$$E(x) = \frac{\alpha \cdot b}{\alpha - 1}$$

(3.5)

Once $\alpha_{on}, \alpha_{off}, b_{on}, b_{off}$ are given, the distribution of the ON/OFF periods are determined. $b_{on}$ is the minimum ON period length, equal to the smallest packet size divided by the line rate. The average load of each ON/OFF source, $L$, is:

$$L = \frac{E_{on}}{E_{on} + E_{off}}$$

where $E_{on}$ and $E_{off}$ are the mean value of ON and OFF period. Therefore,

$$b_{off} = (\frac{1}{L} - 1) \frac{\alpha_{on} \cdot (\alpha_{off} - 1)}{(\alpha_{on} - 1) \cdot \alpha_{off} \cdot b_{on}}$$

(3.6)

We assign $\alpha_{on} = 1.4$, $\alpha_{off} = 1.2$ in the simulation. During the ON period of the ON/OFF source, packets are sent back-to-back.

In our simulation experiments, the network traffic is assumed to consist of IP packets. The nature of IP packets is known to be hard to capture [73]. Statistical data indicates a predominance of small packets, with peaks at the common sizes of 44, 552, 576, and 1500 bytes. The small packets of 40-44 bytes in length include TCP acknowledgment segments, TCP control segments, and telnet packets carrying single characters. Many TCP implementations that do not implement Path Maximum Transmission Unit (MTU) Discovery use either 512 or 536 bytes as the default Maximum Segment Size (MSS) for non-local IP destinations, yielding a 552-byte or 576-byte packet size. An MTU size of 1500 bytes is the characteristic of Ethernet-attached hosts. The cumulative distribution of packet sizes in Fig. 3.4 (statistical-data plot from [73]) shows that almost 75% of the packets are smaller than the typical TCP MSS of 552 bytes. Nearly half of the packets are 40 to 44 bytes in length. On the other hand, over half of the total traffic is carried in packets of size 1500 bytes or larger. This irregular packet-size distribution is difficult to express with a closed-form expression. We adopt a truncated 19-order polynomial, fitted from the statistical data, to faithfully reproduce the IP packet-size distribution (as shown by the polynomial-fit plot in
Fig. 3.4. The number of orders, 19, is the smallest number that can reproduce a visually close match with the steep turns in the statistical data. We set the maximum packet size to be 1500 bytes, since the percentage of packets larger than 1500 bytes is negligibly small.

![Probability distribution function (PDF) of IP packet sizes.](image)

Figure 3.4: Probability distribution function (PDF) of IP packet sizes.

**Maximum Hop Count**

With deflection routing, a loop-mitigation mechanism is necessary. A maximum hop count $H$ is set to limit how many hops a packet can travel (each time the packet passes through a delay line or is transmitted from one node to another, it is counted as one hop). This is similar to the TTL field in the IP packet header, which prevents an IP packet from wandering endlessly in the network. Nevertheless, the network topology of Fig. 3.3 and the aforementioned deflection policy can automatically eliminate looping (since the shortest path between any source-destination pair in Fig. 3.3 involves no more than two hub nodes, and we can ensure that the deflection table will not cause any looping). The purpose of setting the maximum hop count in this study is to limit physical impairments (signal-to-
noise-ratio degradation, accumulated crosstalk, accumulated insertion loss, etc.) of packets, which is mainly introduced by optical buffering.

**Performance Metrics**

We have chosen four metrics to evaluate the network performance with different contention-resolution schemes: network throughput, packet-loss rate, average end-to-end delay, and average hop distance. They indicate the network utilization, reliability, latency, and physical impairments to the signal. The packet-loss rate is the total number of lost packets divided by the total number of packets generated. The network throughput is defined as:

\[
\text{Network throughput} = \frac{\text{total number of bits successfully delivered}}{\text{network transmission capacity} \times \text{simulation time} \times \text{ideal average hop distance}}
\]

\[
\text{Network transmission capacity} = (\text{total } # \text{ of links}) \times (# \text{ of wavelengths}) \times (\text{data rate})
\]

Network throughput is the fraction of the network resource that successfully delivers data. Because packets can be dropped, a part of the network capacity is wasted in transporting the bits that are dropped. In an ideal situation where no packets are dropped and there is no idle time on any link, the network will be fully utilized and the throughput will reach unity. Average hop distance is the hop distance a packet can travel, averaged over all the possible source-destination pairs in the network. The ideal average hop distance (i.e., no packet dropping) of the network in Fig. 3.3 is 2.42.

Table 3.2 shows the values of the parameters used in the simulation experiments.

<table>
<thead>
<tr>
<th>Link length $L$</th>
<th>Data rate $R$</th>
<th># of wavelengths $W$</th>
<th>Max hop distance $H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 km</td>
<td>2.5 Gbps</td>
<td>4, 8, 16</td>
<td>8</td>
</tr>
</tbody>
</table>

All the results are plotted against average offered transmitter load, i.e., the total number of bits offered per unit time divided by the line speed. (For example, if the source is generating 0.5 Gbits of data per second and the transmitter/line capacity is 2.5 Gbps, the transmitter load would be 0.2.) With a given average offered transmitter load and a uniform traffic matrix, the average offered link load per wavelength is:

\[
\text{Ave. offered link load} = \frac{\text{ave. offered TX load} \times \text{total } # \text{ of TX in the network} \times \text{ave. hop distance} \times \text{# of wavelengths} \times \text{total number of uni-directional links in the network}}{	ext{# of wavelengths}}
\]
3.3.2 Comparison of the Basic Contention-Resolution Schemes

Wavelength Conversion

Figure 3.5: Performance comparison of the basic schemes – baseline and wavelength conversion. (a) Throughput comparison; (b) packet-loss rate comparison; (c) average end-to-end delay comparison; (d) average hop distance comparison.

Figure 3.5(a) compares the network throughput of the schemes incorporating wavelength
conversion with different number of wavelengths. For reference, it also shows the throughput of baseline. We simulate four wavelengths in baseline, although the number of wavelengths does not affect the results of baseline since each wavelength plane operates independently. We find that more wavelengths provide better throughput performance for the wavelength-conversion scheme. Meanwhile, the margin of improvement in throughput decreases when the wavelength number increases; with 16 wavelengths, the network throughput is nearly linear to transmitter load.

Figure 3.5(b) compares the packet-loss rate (represented in fraction) of these schemes. It is a good complement to Fig. 3.5(a) because the network’s throughput is directly related to the packet-loss rate, while packet-loss rate can reveal more subtle performance differences, especially under light load. To estimate the upper-bound requirement of packet-loss rate for TCP traffic, we adopt the following criteria: the product of packet-loss rate and the square of the throughput-delay product should be less than unity [74], where the throughput is measured in packets/second on a per-TCP-connection basis. Let us first evaluate the upper bound of packet-loss rate for a nation-wide network, whose round-trip delay is approximately 50 ms. For a 1 Mbps TCP connection, the pipe throughput is approximately 100 packets/second. This gives an estimated upper bound of packet-loss rate of 0.01. With a smaller-sized optical packet-switched network, the round-trip delay is less; thus, the upper bound of the packet-loss rate can be higher than 0.01 for TCP applications. The transmitter load corresponding to 0.01 packet-loss rate for 4wav, 8wav, and 16wav are approximately 0.16, 0.31, and 0.44. This clearly indicates the advantage of using more wavelengths.

Figure 3.5(c) compares the average end-to-end packet delays. We can see the general trend of decreasing delay with higher load, for all values of wavelengths. This is because, when the load increases, the packet-loss rate also increases. Packets with closer destinations are more likely to survive, while packets that have to travel longer distances are more likely to be dropped by the network. The overall effect is that, when we consider only survived packets in the performance statistics, the delay decreases. This effect is very prominent with baseline, because it has the highest packet-loss rate. The same reasoning can be applied to explain the lower delay of 4wav.

Figure 3.5(d) shows the average hop-distance comparison. Since neither baseline nor wc involves any buffering, the average hop distance is proportional to the average end-to-end delay.
Optical Buffering and Deflection

Figure 3.6: Performance comparison of the basic schemes — baseline, optical buffering, and deflection. (a) Throughput comparison; (b) packet-loss rate comparison; (c) average end-to-end delay comparison; (d) average hop distance comparison.

In the schemes incorporating optical buffering, all the optical delay lines are of length 1 km (with propagation delay of 5 µs), enough to hold a packet with maximum length (12000 bits). We simulate four wavelengths in the network, with different buffering and deflection
settings; this is because other number of wavelengths renders similar results, and larger number of wavelengths requires much longer time to simulate. Figure 3.6(a) compares the throughput of different optical-buffer schemes and the deflection scheme. For the buffering schemes, we observe the difference in performance between different numbers of optical delay lines. Network throughput first increases with the number of delay lines, but it saturates after the number of delay lines reaches 16: the throughput curve for 4wav_16buf, 4wav_32buf, and 4wav_32bufwdm are almost indistinguishable. This effect is due to the maximum hop count, which is set to 8. To validate this explanation, we also plot another curve, 4wav_32bufwdm_hopinf, without the maximum-hop-count constraint. It is clear that the utilization of optical buffering increases considerably if packets can be buffered for an unlimited number of times. For deflection, the throughput of 4wav_def is only marginally higher than that of baseline, indicating that deflection, by itself, is not a very effective contention-resolution technique in this network. For the given topology, only about one quarter of the nodes can perform deflection. Furthermore, any source node can reach any destination node through at most two hub nodes (nodes that have degree higher than two); therefore, the chance of a packet being deflected is relatively low. Space deflection can be a good approach in a network with high-connectivity topology, such as ShuffleNet ([34, 68]), but is less effective with a low-connectivity topology.

Figure 3.6(b) shows the packet-loss rates. Deflection alone, in this network, cannot meet the required packet-loss rate of 0.01. The threshold transmitter loads for all the buffering schemes with maximum hop count are approximately 0.2. The packet-loss rate for 4wav_32bufwdm_hopinf is very low and the upper-bound transmitter load is 0.63.

The plots of end-to-end delay and average hop count in Figs. 3.6(c) and 3.6(d) reveal more detail on the effect of setting the maximum hop count. The hop count for all the optical-buffering schemes increases with load, because packets are more likely to be buffered with higher load. However, the end-to-end delay decreases (except for 4wav_32bufwdm_hopinf) with load. This is because (1) packets destined to closer nodes are more likely to survive; and (2) the buffer delay introduced by optical delay lines is small compared with link propagation delay. Without the maximum-hop-count constraint, both end-to-end delay and average hop count rise quickly with load, indicating that unrestricted buffering can effectively resolve the contention. The amount of extra delay caused by unrestricted buffering is less than 0.5 ms with transmitter load of 0.63, and this extra delay may be acceptable for most applications. The disadvantage of unrestricted buffering is the physical impairment of the signal (e.g., attenuation, cross-talk, etc.), since packets must traverse the switch fabric and delay lines many more times. In this case, optical regeneration may become necessary. We also notice that, when the load is light, the end-to-end delay increases for 4wav_def; this is because, under light load, deflection is resolving the contention effectively by deflecting
packets to nodes outside the shortest path, thus introducing extra propagation delay. For average end-to-end delay, the effect of the extra propagation delay due to deflections is more prominent than the effect of having more survived packets with closer destinations. This explains the initial increase of delay in the deflection scheme.

Table 3.3 lists the upper bound of average offered transmitter load with acceptable packet-loss rate set at 0.01. 4wav, 32buf, wdm, hopinf offers the best packet-loss performance, but it may be more expensive to implement due to the large number of optical delay lines and switch ports, and it may also require optical amplification/regeneration. Wavelength conversion is a very effective approach to resolve contention, and its effectiveness depends on the number of wavelengths in the system. Deflection is the least effective approach in the example network, but its benefit can be achieved when we combine it with other schemes.

Table 3.3: Comparison of upper-bound average offered transmitter load with packet-loss rate of 0.01. (For 4wav_def, we did not obtain a packet-loss rate less than 0.01 because the lowest simulated average offered transmitter load is 0.05; therefore, we assume that its upper-bound transmitter should be less than 0.05.)

<table>
<thead>
<tr>
<th>Scheme</th>
<th>4wav_wc</th>
<th>16wav_wc</th>
<th>4wav_4buf</th>
<th>4wav_16buf</th>
<th>4wav_32buf_wdm_hopinf</th>
<th>4wav_def</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. load</td>
<td>0.16</td>
<td>0.44</td>
<td>0.18</td>
<td>0.2</td>
<td>0.63</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>

### 3.3.3 Comparison of Combinational Schemes

This study chose four scenarios for different combinations of contention-resolution schemes: 16wav_wc+16buf, 16wav_wc+def, 16wav_16buf+def, and 16wav_wc+16buf+def.

Figure 3.7(a) shows the throughput comparison of these schemes. The scheme that incorporates all three dimensions (wavelength, time, and space) for contention resolution offers the best throughput. One can also observe the benefit of using wavelength converters: the schemes involving wavelength conversion perform better under heavy load.

Figure 3.7(b) compares the packet-loss rates. Although the throughput for all these schemes are quite close, their packet-loss rates have large differences. The best performer is 16wav_wc+16buf+def, followed by 16wav_wc+16buf, 16wav_wc+def, and 16wav_wc+16buf+def. The benefit of wavelength conversion and buffering appears to be dominant. The upper-bound transmitter load for a packet-loss rate less than 0.01 is as follows: 0.2 for 16wav_16buf+def, 0.45 for 16wav_wc+def, 0.61 for 16wav_wc+16buf, and 0.65 for 16wav_wc+16buf+def.

Figures 3.7(c) and 3.7(d) show the average end-to-end delay and average hop distance of these schemes. The end-to-end delay presents a general trend of decrease for all four schemes, due to the dropping of packets with far-away destinations with increasing traf-
fic load. $16wav_{-}16buf+def$ has a more prominent trend of decrease, indicating buffering and deflection alone can make the network prefer packets with closer destinations. $16wav_{-}16buf+def$ also introduces more extra hops due to the high utilization of optical delay lines.

Figure 3.7: Performance comparison of the combinational schemes. (a) Throughput comparison; (b) packet-loss rate comparison; (c) average end-to-end delay comparison; (d) average hop distance comparison.
3.3.4 Limited Wavelength Conversion

Both wavelength conversion and optical buffering can effectively resolve contention. However, they require extra hardware which may increase the system cost. Full-range wavelength conversion requires a fast tunable laser as the pump laser for every wavelength converter, and optical buffering requires extra ports on the switch fabric. Here we consider the case of limited wavelength conversion, which can be realized through simultaneous multiple-channel conversion based on parametric wavelength conversion. We investigate different degrees of limited wavelength conversion. A degree-$C$ wavelength conversion means that a given wavelength can be converted to $C$ pre-designated wavelengths. With parametric wavelength converters, one can save the cost on pump lasers. Table 3.4 shows an example of the possible wavelengths that $\lambda_1$ can be converted to in a 16-wavelength system, when $C$ is set to 1, 2, and 4.

Table 3.4: Target conversion wavelengths for $\lambda_1$ in a 16-wavelength system.

<table>
<thead>
<tr>
<th>Conversion degree $C$</th>
<th>1</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target $\lambda$’s</td>
<td>$\lambda_{16}$</td>
<td>$\lambda_{16}, \lambda_8$</td>
<td>$\lambda_{16}, \lambda_8, \lambda_4, \lambda_2$</td>
</tr>
</tbody>
</table>

![Figure 3.8](image)

Figure 3.8: Performance comparison of limited wavelength conversion. (a) Throughput comparison; (b) packet-loss rate comparison.

Figure 3.8 shows the throughput and packet-loss rate comparison of the three limited wavelength conversion cases—$16\text{wav}_wclim1$, $16\text{wav}_wclim2$, and $16\text{wav}_wclim4$. For refer-
ence, it also shows $4_{wav,wc}$. The performance of these schemes ranks in the following order (starting with the best): $16_{wav,wclim4}$, $4_{wav,wc}$, $16_{wav,wclim2}$, $16_{wav,wclim1}$; their respective upper-bound transmitter load with packet-loss rate of 0.01 are 0.31, 0.17, 0.13 and 0.07. $16_{wav,wclim4}$ can accommodate nearly twice the load relative to that for $4_{wav,wc}$, indicating that limited wavelength conversion in a system with more wavelengths is better than full-range wavelength conversion in a system with fewer wavelengths.

**3.3.5 Comparison Study of Slotted and Unslotted Networks**

Since many previously-proposed architectures have employed the slotted-network approach, this chapter also presents a quantitative comparison study between the slotted and the unslotted networks. In the slotted-network scenario, packets have fixed size of 4096 bits (corresponding to 512 bytes, the average IP packet size) and they are placed in a 1.5 $\mu$s time slot (for a data rate of 2.5 Gbps). Before the packets enter each switching node, they must be aligned; therefore, the switching occurs only at the slot boundaries. Synchronization stages are necessary in slotted networks for alignment of packets. In unslotted networks, packets are of variable size, with the size distribution of IP packets. There are no synchronization stages needed and packets are switched ‘on the fly’. It is assumed that the packet arrival processes in both cases are self-similar. Each node is equipped with wavelength converters, multi-wavelength fiber delay line(s), and a secondary routing-table for deflection. In the case of the slotted network, the delay line’s length is equal to the slot size ($1.5 \mu$s, or 300 m).

In the case of the unslotted network, since we observe that a delay line size smaller than the maximum packet size can cause degradation of network performance, the delay line’s length is set to be equal to the maximum packet size ($5 \mu$s, or 1 km). The slotted network node is equipped with full-range wavelength converters and one multi-wavelength fiber delay line, operating in a network with four wavelengths (denoted as slotted$_{4_{wav,wc}+1_{bufwdm}}$). For the unslotted network, since the contention mostly occurs at the nodes with higher degrees, the number of optical delay lines is set according to the nodal degree, and we denote such a scheme as unslotted$_{W_{wav,wc}+(bufwdm=deg)+def}$. This approach can greatly reduce the cost of nodes that have only two incoming and outgoing fibers.

Figure 3.9 compares the packet-loss rates for slotted and unslotted networks. There is a large improvement in the unslotted network with more wavelengths. Under light load, both unslotted$_{8_{wav,wc}+(bufwdm=deg)+def}$ and unslotted$_{16_{wav,wc}+(bufwdm=deg)+def}$ outperform the slotted network. However, the packet-loss rate of unslotted networks increases faster with load than that of the slotted network, because both buffering and deflection are very sensitive to a certain load threshold. This is especially true under uniform traffic. During the simulation experiment, we observed that all the hub nodes become congested nearly simultaneously, when the traffic load reaches the threshold value. Once this load
threshold is reached, neither deflection in time domain (optical buffering), nor deflection in space domain seems to be effective, since the chance of contention is the same everywhere in time and space. The upper-bound transmitter load with packet-loss rate of 0.01 for the slotted network and the unslotted network (with 4, 8, and 16 wavelengths) are 0.62, 0.51, 0.58, and 0.63, respectively. Although, by nature, the slotted network performs better than the unslotted network, it requires complex synchronization stages at the switching nodes. The unslotted network can accommodate packets in its native size and does not require the synchronization stages. From the performance results, we can tell that, with more wavelengths and careful allocation of buffering capacity, the unslotted network can provide comparable or better performance than the slotted network. For a more detailed comparison study of slotted and unslotted networks, the reader is referred to [15].

![Figure 3.9: Packet-loss rate comparison between slotted and unslotted networks.](image)

### 3.4 Conclusion

This chapter presented a unified study on various contention-resolution schemes in an optical packet-switched network with irregular mesh topology. It compared the advantages and limitations of contention-resolution schemes in wavelength, time, and space domains using an integrated optical packet-switching architecture. Limited wavelength conversion and selective deflection were also investigated. Among all the schemes, wavelength conversion, combined with optical buffering (allocated according to topology) and selective
deflection, appears to be the most efficient scheme. By applying wavelength conversion in a 16-wavelength network, and by setting the number of optical delay lines according to nodal degree, we are able to demonstrate a packet-loss rate of less than 0.01 with average transmitter load up to 0.63, with typical IP packet-size distribution and bursty packet arrivals.

It is possible to design a simple optical packet-switching node architecture with excellent network performance by incorporating contention-resolution schemes within all three dimensions, and by allocating network resources (e.g., the number of delay lines, the degree of wavelength conversion, deflection-table entries, etc.) according to the specific network topology and traffic demand. Chapter 5 will explore different contention-resolution algorithms in both optical and electrical layers.
Chapter 4

PLATO: A Generic Modeling Technique for Optical Packet-Switched Networks

4.1 Introduction

One of the objectives in designing the optical packet-switched networks is low packet-loss rate (PLR, sometimes also denoted as probability of packet loss, PPL). Packets may be dropped in contentions, when there are more packets traveling to the output port of a switch than the port can actually accommodate. Various node architectures have been proposed in the past [17, 21, 29, 60, 63, 75], as well as contention-resolution algorithms in time, space, and wavelength domains [14, 22, 36, 61]. Although some analytical studies have been presented for different node architectures, simulation is still the dominant tool for studying the network-wide performance of large, practical networks.

Today’s optical links have a line speed as high as 10 Gbps (OC-192) and many optical packet formats have sizes based on IP packets, which is no more than tens of thousand of bits (with optical burst switching [8] as an exception). Therefore, to simulate the performance of a network, millions to billions of packets need to be generated. The burden of computation becomes much heavier when the network size increases, with multiple wavelengths and when the PLR is low. In many cases, such large simulations become impractical.

This chapter proposes and develops PLATO (Packet-Loss Analysis with Traffic flOw), a generic analytical modeling technique. It is suitable for a large variety of network topologies and node architectures. This technique combines numeric method with traffic and probability analysis. The comparison between simulation and analysis results demonstrates that PLATO is both computationally tractable and accurate.
Chapter 4: PLATO: A Generic Modeling Technique

The analysis of optical packet-switching architectures has received some attention in the literature. However, most past work has concentrated on modeling one single switch, or a network with one specific node architecture. A generic approach, independent of topology and node architecture, has been an unchartered problem, until now. Our proposed technique is an attempt to contribute to this subject. It is aimed at providing the network designer an accurate and fast method for evaluating network PLR without running computation-intensive simulations.

After presenting the basic building blocks of the technique (Sections 4.2 and 4.3), the chapter shows its application to two example networks, with comparison results (Section 4.4). Section 4.5 concludes the chapter.

4.2 Basic Algorithms of PLATO

![Diagram of PLATO building blocks]

Figure 4.1: Building blocks of PLATO.

![Diagram of an example six-node network]

Figure 4.2: An example six-node network.
Table 4.1: Routing table of the six-node network.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>*</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>*</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>*</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>*</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>*</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>*</td>
</tr>
</tbody>
</table>

Figure 4.1 shows the basic approach of PLATO. The model consists of two parts: traffic-flow analysis and local node modeling. In traffic-flow analysis, based on the topology, traffic matrix, and routing algorithm, the traffic intensity and its composition at each input port of every switch is computed. This information is then fed into the PLR model for each node along the path of every source-destination pair, to calculate the corresponding PLR. Because the final model is a multiple-variable system, an iterative approach is used to obtain the actual value of the PLR.

Figure 4.2 shows a simple network topology. Suppose a packet is going from node 1 to node 4, through node 2 and node 3. This packet could be dropped in any of these nodes, due to contention or lack of sufficient amount buffering. Here we define $\text{plr}_{ij}$ as the end-to-end packet-loss probability of any packet from node $i$ to node $j$, and $p_{ijk}$ as the packet-loss probability at intermediate node $k$ for packets belonging to the $i \rightarrow j$ source-destination pair. $t_{ij}$ is defined as the offered traffic intensity (offered load) between source node $i$ and destination node $j$, and $\text{PLR}$ is defined as the network-wide packet-loss probability. The packet-loss probability equals the ratio of the number of lost packets to the number of total offered packet; therefore, “packet-loss rate” and “packet-loss probability” are used interchangeably throughout this chapter. The end-to-end packet-loss probability can be obtained by:

$$\text{plr}_{ij} = 1 - \prod_{k \in i \rightarrow j} (1 - p_{ijk})$$  \hspace{1cm} (4.1)$$

where $k$ denotes the index of nodes along the path from $i$ to $j$ (including $i$ and $j$.)

To compute the network packet-loss probability, all the traffic load from every possible source-destination pair should be taken into consideration. We have
Chapter 4: PLATO: A Generic Modeling Technique

51

To compute $PLR$, we need to find out $plr_{ij}$, and therefore $p_{ijk}$ for any $i \to j$ pair. This requires a careful inspection of the route that every $t_{ij}$ takes and computation of the packet-loss probability at every intermediate node. Consider $t_{14}$ for example in Fig. 4.2. According to the routing information in Table 4.1, the nodes that a $t_{14}$ packet traverses include nodes 1, 2, 3 and 4. At each node, the packet might be dropped, depending on the contentions it encounters. Figure 4.3 illustrates the possible contentions at each switch, where the gray arrows represent the contending packet streams. The input/output ports are indexed by the node number they are leading to. In order to compute the packet-loss probability at each switch (in this case, $p_{141}$, $p_{142}$, $p_{143}$, $p_{144}$), we need to find out (1) the contention traffic flows’ intensity and composition, and (2) the PLR model of each switch as a function of input traffic flows. The detailed algorithms are described in the following subsections.

4.2.1 Traffic-Flow Analysis

A switch, regardless of its internal architecture, is a system with fixed number of inputs and outputs. The packet loss probability of a packet stream from a given input port is a function of the input traffic flows only. The PLR of a packet can be determined as soon as all the contending traffic flows’ intensity is known, provided there is an accurate model for the specific switch architecture.

Let us take $p_{142}$ for example, as shown in Fig. 4.4. We follow the $t_{14}$ packets through node 2. The packets are arriving on port 1 and leaving from port 3. From the routing table in Table 4.1, there are three streams contending for output port 3. Each stream is composed of traffic from different source-destination pairs. The top stream consists of packets from

\[
PLR = \frac{\sum_{i,j,i \neq j} (t_{ij} \cdot plr_{ij})}{\sum_{i,j,i \neq j} t_{ij}} \quad (4.2)
\]

Figure 4.3: Possible contentions a $t_{14}$ packet can encounter.
Chapter 4: PLATO: A Generic Modeling Technique

\(t_{13}\) and \(t_{14}\), the middle one from \(t_{63}\), and the bottom one from \(t_{23}, t_{24}\) and \(t_{25}\). Since, before these tributary flows reach node 2, they have already suffered from possible packet loss from the previous nodes along their respective paths, we need to take into account the individual packet-loss rate they have suffered. Therefore, the load of the three streams contending for output port 3 at the input ports of node 2 are:

\[
t_{13} \cdot (1 - p_{131}) + t_{14} \cdot (1 - p_{141}) + t_{63} \cdot (1 - p_{636}) + t_{23} + t_{24} + t_{25}
\]

According to the switch model specific to the node architecture, with the above values of traffic intensity, we are able to compute the packet-loss probability of \(p_{142}\). In this case, \(p_{132} = p_{142}\) because \(t_{13}\) and \(t_{14}\) belong to the same tributary stream contending for output port 3.

![Figure 4.4: Contention for \(t_{14}\) at node 2.](image)

It is likely that an optical link has multiple wavelengths, and a node has more than one add/drop port. Nevertheless, the contending traffic flows at a certain node are determined with given topology, traffic matrix \(t_{ij}\), and static routing.

We have demonstrated how to analyze the contending traffic for computing \(p_{142}\). Our goal is to find out \(plr_{ij}\), which is given by Eqn. (4.1). Therefore, we have to perform the same analysis for each \(i \to j\) pair at every switch on its path. This can be performed easily with modern computing facilities. The pseudo-code for the traffic-flow analysis algorithm will follow the completion of local switch models.

Once the traffic analysis is complete, it can be combined with the local model of each switch to render a networked model to compute the global PLR.
4.2.2 Local Switch Modeling

There have been several proposals for optical packet-switch architectures in the literature. Many of them are provided with analytical models. One main characteristic of an optical packet switch is that it does not have optical random-access memory (RAM), unlike the electronic peers which can take advantage of inexpensive RAM for packet buffering. This has resulted in various alternative approaches to resolve contentions, namely buffering (based on fiber delay lines), wavelength conversion, deflection routing, and over-provisioning (offer more capacity to the network).

Since this work does not aim to discuss in depth the modeling of specific switch architectures, but rather a generic approach for network modeling, we will first develop a simple but representative example switch model for verification of the PLATO technique. The example switch we adopt is a baseline switch architecture, which is simply an \( N \times N \) space switch with no buffers and operates on a single wavelength, as shown in Fig. 4.5.

For the moment, we consider the case when there are three packet streams contending for one output port. Whenever there is a collision, the later-arriving packets will be dropped. For this example, we assume packets are of fixed size and arrive according to Poisson pro-
cess. Our results will show that the model is also suitable for variable packet sizes with negative exponential distribution. The traffic load of the streams are $l_1$, $l_2$, and $l_3$, while the packet-loss probabilities are $p_1$, $p_2$, and $p_3$, respectively. Let us first study the packets arriving from stream 1. When a packet arrives, it will not suffer a collision if no other frames are sent within one packet time of its start, as shown in Fig. 4.6. The colliding packets can only be from stream 2 or stream 3, since stream 1 packets will never contend among themselves. Let $t$ be the time required to send a packet. If stream 2 and/or stream 3 has generated a packet between time $t_0$ and $t_0 + t$ and it has successfully occupied the output port, the end of that packet will collide with the beginning of the shaded one from stream 1. The probability that $k$ packets are generated during one packet time is given by the Poisson distribution: $Pr[k] = (G^k e^{-G})/k!$, where $G$ is the average load on the output port, contributed by stream 2 and stream 3: $G = l_2(1 - p_2) + l_3(1 - p_3)$. Therefore, the packet-loss probability for stream 1 is:

$$p_1 = 1 - Pr[0] = 1 - e^{-[l_2(1-p_2)+l_3(1-p_3)]}$$  \(4.3\)

Similarly, we have:

$$p_2 = 1 - e^{-[l_1(1-p_1)+l_3(1-p_3)]}$$  \(4.4\)

$$p_3 = 1 - e^{-[l_1(1-p_1)+l_2(1-p_2)]}$$  \(4.5\)

Equations (4.3)-(4.5) form a system of multiple-variable transcendental equations. It is difficult to obtain the analytical solutions for these equations. PLATO adopts a modified iteration algorithm to obtain the numerical solutions. Iteration is useful for solving single-variable nonlinear equations. First, give a set of initial values (for example, zero) to $p_1$, $p_2$, $p_3$. Then plug them into the right-hand side of Eqns. (4.3)-(4.5), which will give a new set of values for $p_1$, $p_2$, $p_3$. Repeat the substitution process until the values of $p_1$, $p_2$, $p_3$ converge. (The proof of the convergence of this algorithm is provided in the Appendix.) When this system of equations is applied in the network model, we write a function of $(l_1, l_2, l_3)$, which returns the value of $p_1$. The pseudo-code is as follows:

```plaintext
plr_3stream(l1, l2, l3)
1   p1_new, p2_new, p3_new ← 0
2   p1_old ← 1
3   while Distance(p1_new, p1_old) > e
4       p1_old ← p1_new
5       p2_old ← p2_new
6       p3_old ← p3_new
7       p1_new ← 1-exp(-(12(1-p2_old)+13(1-p3_old)))
```
Chapter 4: PLATO: A Generic Modeling Technique

8 \text{p2\_new} \leftarrow 1 - \exp(-(1\text{-}p1\_old)+13\text{(1}\text{-}p3\_old)))
9 \text{p3\_new} \leftarrow 1 - \exp(-(1\text{-}p1\_old)+12\text{(1}\text{-}p2\_old)))
10 \text{return (p1\_new)}

Similar functions can be written (e.g., \text{plr\_2stream()}, \text{plr\_4stream()} \ldots) for any number of streams contending for one output port.

4.2.3 Constructing the Network Model

Now we have the complete information on traffic intensity and its composition at every input port of every switch, as well as the model to compute packet-loss probability at any intermediate node of any source-destination pair. We are ready to compute \text{plr}ij and PLR.

Let us look back at the example of \text{plr}14\_14. From Eqn. (4.1):

\[
\text{plr}14 = 1 - (1 - p141)(1 - p142)(1 - p143)(1 - p144)
\]

where

\[
p141 = 0
\]

\[
p142 = \text{plr\_3stream}\left(t13(1-p131) + t14(1-p141),
\quad t63(1-p636), \quad t23 + t24 + t25\right)
\]

\[
p143 = \text{plr\_2stream}\left(t14(1-p141)(1-p142) +
\quad t24(1-p242), \quad t34\right)
\]

\[
p144 = \text{plr\_2stream}\left(t14(1-p141)(1-p142)\cdot
\quad (1-p143) + t24(1-p242)(1-p243)
\quad +t34(1-p343), \quad t54(1-p545) + t64(1-p645)\right)
\]

With flow analysis as described earlier, it is possible to construct similar equations for every \text{p}ijk, \forall i, j, k; \ i \neq j; \ k \in i \rightarrow j. This leads to a system of nonlinear equations with all \text{p}ijk as the variables. To solve them, we apply the numeric iteration method again, as previously described in the pseudo-code. This algorithm will not only give the network \text{PLR}, but also give the packet-loss rate at every switch of any source-destination pair, \text{plr}ijk. Thus, it enables us to investigate the “hot spots” in the network.

Here is the pseudo-code of the algorithm:

\text{Main(T, e)}
1 Init(q)
2
3 \text{PLR\_old} \leftarrow 0
4  PLR_new ← 1
5
6  while Distance(PLR_new, PLR_old)>e
7      do PLR_old ← PLR_new
8          PLR_new ← Step(T, q)
9
10 return PLR_new

where T is the network topology and e is the numeric iteration error limit. Init(q) initializes every element of array q as 0. Distance(a, b) returns |a - b|/a and Step(T, q) returns the network PLR.

Step(T, q)
1  for i ← 1 to Node_Number(T)
2      do for j ← 1 to Node_Number(T)
3          do tmp_p ← 1
4              for each vertex k in Route[i][j]
5                  do Compute(p[i][j][k], q)
6                      tmp_p ← tmp_p * (1 - p[i][j][k])
7                      plr[i][j] ← 1 - tmp_p
8                      PLR_new ← \sum_{i,j,i\neq j} plr_{ij} \cdot t_{ij} / \sum_{i,j,i\neq j} t_{ij}
9
10 for i ← 1 to Node_Number(T)
11      do for j ← 1 to Node_Number(T)
12          do for k ← 1 to Node_Number(T)
13              do q[i][j][k] ← p[i][j][k]
14
15 return PLR_new

Compute(p[i][j][k], q)
1  b ← the input port index when packet flow from i to j goes through node k
2  find all the other packet flows and their loads, which go through node k using output port b
3 merge the flows from the same input port to output port b in node k
4 compute $p[i][j][k]$ according to the number of conflicting flows by fitting them into suitable local switch models

where $Node\_Number(T)$ returns the total number of nodes in network T, and $Route[i][j]$ contains the route from node i to node j.

### 4.3 Delay-Line Model

To verify the accuracy of the PLATO technique with a different node architecture, this study also presents a sample switch model that implements fiber delay lines as buffers for the packets. Figure 4.7 shows the switch architecture. There is one fiber delay line dedicated to each output port, including the local drop port. Although this is not a very economical way of utilizing buffers, it simplifies the modeling process and can still capture sufficiently well the effect of buffering. To further reduce the model complexity, we impose that each packet is allowed to enter the delay line only once during contention. Packets could be dropped due to a busy delay line or a busy output port when it emerges from the delay line. Each delay line is 5-km long in the numerical example here to ensure that (1) it can accommodate a full packet and (2) when the packet emerges from the end of the delay line, the previous packet occupying the preferred output port has finished transmission. It is necessary to develop new functions as local switch models, namely $bufplr\_nstream(l_1, l_2, \ldots, l_n)$, where $n$ is the number of contending streams, to take into account the effect of buffering.

![Figure 4.7: Switch architecture with one dedicated delay line for each output port.](image)

Figure 4.7 shows an example of 3 packet streams contending for one output port. Let us follow the packets arriving from stream 1 and investigate the packet-loss probability they suffer. Packets from stream 1 are contending with packets from stream 2, stream 3, and the delay line. We denote $l_k$ as the load on stream $k$, $l_b$ as the load on the delay line, $b_k$
as the probability that a packet from stream $k$ needs to be buffered due to contention on the output port, and $bb_k$ as the probability that a packet from stream $k$ will find the delay line occupied. We will also use the functions previously written for the baseline switch, \texttt{plr\_nstream($l_1, l_2, \ldots, l_n$)}, to compute $b_1$. There are two scenarios where the packet from stream 1 is dropped: (1) it finds the output port and the delay line busy at the same time, and (2) it is successfully buffered but finds the output port busy after emerging from the delay line. Therefore, the packet-loss probability for stream 1 packets, $p_1$, can be written as:

$$p_1 = b_1 \cdot (bb_1 + (1 - bb_1) \cdot \text{plr\_4stream($l_b, l_1, l_2, l_3$)})$$  \hspace{1cm} (4.6)

where

$$bb_1 = \text{plr\_3stream($l_1b_1, l_2b_2, l_3b_3$)}$$  \hspace{1cm} (4.7)

$$bb_2 = \text{plr\_3stream($l_2b_2, l_1b_1, l_3b_3$)}$$  \hspace{1cm} (4.8)

$$bb_3 = \text{plr\_3stream($l_3b_3, l_1b_1, l_2b_2$)}$$  \hspace{1cm} (4.9)

$$b_1 = \text{plr\_4stream($l_1, l_2, l_3, l_b$)}$$  \hspace{1cm} (4.10)

$$b_2 = \text{plr\_4stream($l_2, l_1, l_3, l_b$)}$$  \hspace{1cm} (4.11)

$$b_3 = \text{plr\_4stream($l_3, l_1, l_2, l_b$)}$$  \hspace{1cm} (4.12)

$$l_b = l_1b_1(1 - bb_1) + l_2b_2(1 - bb_2) + l_3b_3(1 - bb_3)$$  \hspace{1cm} (4.13)

From Eqns. (4.6)-(4.13), together with the functions written for the baseline switch architecture and the iteration method, we can construct function \texttt{bufplr\_3stream($l_1, l_2, l_3$)} to compute the packet-loss probability $p_1$ with one dedicated delay line. Similarly, \texttt{bufplr\_nstream($l_1, l_2, \ldots, l_n$)} can be written and plugged into the traffic-flow model, and the network PLR model can be constructed.

The pseudo-code is as follows:

\begin{verbatim}
bufplr_3stream(l1, l2, l3)
1  lb ← lb_3stream(l1, l2, l3)
\end{verbatim}
Chapter 4: PLATO: A Generic Modeling Technique

4.4 Numerical Results and Discussion

Using the above algorithms, we are now able to model the network’s performance (packet-loss rate) with a given topology and a static routing algorithm. We apply the algorithms to two topologies: the six-node topology in Fig. 4.2 and the NSF network topology in Fig. 4.9. Shortest-path routing is used in both topologies. Our first example switch architecture is the baseline switch with no contention resolutions, as shown in Fig. 4.5. Each switch has one local add port and one local drop port. Offered traffic is injected to the network from the transmitter at the local add port. Each transmitter is operating at OC-48 line speed. Packets are of size 12000 bits.

Figure 4.10 shows a comparison between simulation and analytical results. PLR is plotted against average offered transmitter load, which is the average fraction of busy time.
over total simulated time for the transmitter at the local add port of every switch. The traffic injected from a transmitter is uniformly distributed among all the rest of the nodes as destinations. The model provides excellent matching with the simulation results. Figure 4.11 also gives a zoom-in view of the light-load region, since the low PLR region is more important. Both figures show that the PLATO technique offers accurate results of PLR with both light and medium load, with different topologies.

Figure 4.12 shows the PLR results with buffering for the NSF topology. The analytical
Figure 4.11: A zoom-in view of the light-load region of Fig. 4.10.

Figure 4.12: PLR with buffering switch architecture for the NSF topology.
model demonstrates a very close match with the simulation results, indicating that the PLATO technique has great accuracy with different switching architectures.

We also carried out simulations with variable packet size. The packet-size distribution is negative exponential, with mean of 12000 bits. The results show no noticeable difference from the fixed-size case.

4.5 Conclusion

This chapter presented PLATO, an analytical modeling technique for evaluating packet-loss probability for optical packet-switched networks. This technique can accurately model different topologies and node architectures.

The traffic-analysis part provides an algorithm to decompose all the traffic flows and to construct a system of equations that represent the contentions at every output port of every node. The switch-modeling part models each switch in such a way that the packet-loss probability of any input packet stream can be computed as a function of the intensities of all contending streams. These functions representing local packet-loss probability are called by the system of equations from the traffic-flow analysis. Through numerical iterations, the values of the packet-loss rate of all the flows finally converge and give the network-wide packet-loss rate.

Comparison between analytical and simulation results shows excellent accuracy of the PLATO technique. It is also demonstrated that PLATO performs well with different topologies and node architectures; therefore, it is a good candidate as a generic modeling technique for optical packet-switched networks, and it should be customized for each specific application.

The work presented here can be pursued in several directions. On the local-switch modeling side, the study assumes Poisson arrival. In practice, packet arrival is most likely self-similar. Packet sizes are closely related with IP traffic, in which the packet-size distribution has a hard-to-capture bipolar nature. Moreover, we only considered two simple node architectures, while there exist a large number of complex architectures incorporating wavelength conversion and sophisticated buffering schemes. On the network-traffic-analysis side, the PLATO technique is based on static routing. Dynamic routing, such as deflection routing, has been proposed as one of the important contention-resolution techniques. The traffic-flow-analysis algorithm could be extended to include dynamic routing.

A significant contribution of this investigation is the proposed PLATO technique. It combines (1) network-traffic-flow analysis, (2) switch modeling, and (3) numeric method to accurately compute the packet-loss probability of the network and each flow. It can be applied to an arbitrary network topology and node architecture. We believe that our
present work can serve as an excellent foundation for further exciting research along the
lines outlined in the previous paragraphs.

Appendix A

Iteration is a common numeric method to solve certain nonlinear equations \[76\]. Let us
begin with the single-variable case. Suppose the equation to solve is \( f(x) = 0 \). It can be
re-written as \( x = \varphi(x) \). We start with an estimate of the root \( x_0 \), and plug it into the
equation to obtain \( x_1 = \varphi(x_0) \). Generally, \( x_1 \neq x_0 \). Then, plug \( x_1 \) back into the equation
and obtain \( x_2 = \varphi(x_1) \). By repeating this substitution, we have a series of approximate
solutions: \( x_0, x_1, x_2, \ldots, x_n, \ldots \). If we have a pre-defined tolerable error limit \( \varepsilon \), then for
certain functions \( \varphi(x) \), this process is limited. The condition to end the process is
\( |x_{n+1} - x_n| < \varepsilon \). \( x_{n+1} \) is the approximation of the root with error less than \( \varepsilon \). The sufficient
condition for the iteration process to converge is \( |\varphi'(\xi)| < 1, \forall \xi \).

For a system of nonlinear equations of \( k \) variables

\[
\begin{align*}
x_1 &= \varphi_1(x_1, x_2, \ldots, x_n) \\
x_2 &= \varphi_2(x_1, x_2, \ldots, x_n) \\
&\vdots \\
x_k &= \varphi_k(x_1, x_2, \ldots, x_k)
\end{align*}
\]

the initial estimate of solutions are:

\( x_{1,0}, x_{2,0}, \ldots, x_{k,0} \)

and the condition of convergence is:

\[
\left| \frac{\partial \varphi_i(x_1, x_2, \ldots, x_k)}{\partial x_j} \right| < 1 \quad i, j = 1, 2, \ldots, k
\]

(4.14)

In the equation systems developed from both traffic-flow analysis and local switch mod-
els, the variables are packet-loss probabilities, which are between 0 and 1. All the equations
can satisfy Eqn. (4.14). Take \texttt{plr\_3stream}(l_1, l_2, l_3) in Fig. 4.5, for example. This function
performs iteration to solve

\[
\begin{align*}
p_1 &= 1 - e^{-[l_2(1-p_2)+l_3(1-p_3)]} \\
p_2 &= 1 - e^{-[l_1(1-p_1)+l_3(1-p_3)]} \\
p_3 &= 1 - e^{-[l_1(1-p_1)+l_2(1-p_2)]}
\end{align*}
\]

The partial differentials of \( p_1 \) are:

\[
\begin{align*}
\frac{\partial p_1}{\partial p_2} &= -l_2 \cdot e^{-[l_2(1-p_2)+l_3(1-p_3)]} \quad (4.15) \\
\frac{\partial p_1}{\partial p_3} &= -l_3 \cdot e^{-[l_2(1-p_2)+l_3(1-p_3)]} \quad (4.16)
\end{align*}
\]
Since \( l_i, p_i < 1 \), we have:

\[
\left| \frac{\partial p_1}{\partial p_2} \right| < 1 \quad (4.17)
\]

\[
\left| \frac{\partial p_2}{\partial p_3} \right| < 1 \quad (4.18)
\]

Therefore, they are all converging.
Chapter 5

Unification and Extension of Contention-Resolution Studies in Optical Packet-Switched Networks

5.1 Introduction

So far we have investigated different contention-resolution schemes based on wavelength conversion, optical buffering, and deflection routing. This chapter will extend the studies of contention resolutions toward more practical issues, such as implementing packet priorities and interfacing with electronic client networks. Section 5.2 studies the effect of combining different contention-resolution policies with packet priority classes and explains the reasons for certain observed network behaviors. Section 5.3 presents a comparison study of the packet-loss rate between slotted and unslotted networks with priority-based routing. Section 5.4 proposes a hybrid electrical-optical contention-resolution scheme, which utilizes electrical buffering at the edge of the networks. This hybrid scheme offers a packet-loss rate of less than 0.01 for transmitter load up to 0.65, with self-similar IP traffic. Section 5.5 conducts a study on the TCP performance of optical packet switching with electrical ingress buffering and aggregation. Section 5.6 concludes the chapter.

5.2 Priority-Based Routing in Optical Packet-Switched Networks

A network built on optical packet switching will have to not only provide IP-like connectionless services, but also higher-quality connection-oriented services. This will require the network to have differentiated service qualities on the packet level. Although, at this stage,
it seems early to delve into the realm of Quality-of-Service (QoS) studies of such networks, since most of the optical packet switches are lab prototypes; nevertheless, it is worthwhile to study the network behavior with different packet priorities, which will lead us one step closer to the implementation of different Class of Service (CoS) in such networks.

5.2.1 Node Architecture and Routing Policies

Figure 5.1 shows the network topology under study. There are six nodes connected by 20 km long, bi-directional fiber links. Every fiber can accommodate four wavelengths, each of which is operating at 2.5 Gbps (OC-48).

![Network topology under study.](image)

The wavelength converters in the study are based on parametric wavelength conversion [70]. The advantage of doing so is that parametric wavelength converters can convert multiple wavelengths simultaneously. (The output frequencies are the mirror images of the input frequencies, where the ‘mirror’ is at half of pump frequency.) Considering the high cost associated with full-range wavelength conversion, parametric wavelength conversion appears to be a cost-effective solution. The four wavelengths are labeled as $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$, with the wavelength conversion pattern: $\lambda_1 \leftrightarrow \lambda_2$, and $\lambda_3 \leftrightarrow \lambda_4$.

The optical buffering at each switch is done by connecting a piece of fiber delay line, which is of one average packet size, between an input port and an output port of the switch. The delay line can accommodate four wavelengths at a time. This is a rather simple setup compared with many other optical buffering architectures [17, 29], but it performs well when combined with wavelength conversion and deflection routing. Each node keeps two routing tables: primary and secondary. Whenever deflection routing is required, the switch control will route the packet according to the secondary routing table. The primary routing table applies shortest-path routing. Because of the specific topology of this network, when
only shortest-path routing is concerned, it can be easily observed that most of the packet forwarding happens at nodes 2, 3, 5 and 6; while nodes 1 and 4 only generate and receive packets. Therefore, for nodes 1 and 4, the contention mainly takes place at the local drop ports. For this reason, we place the wavelength converters only at the local drop ports, but not at the other output ports of the switch fabric, in order to more effectively resolve contentions. At the other nodes (2, 3, 5, and 6), the wavelength converters are placed at the output ports leading to other nodes, but not the local drop ports. The characteristics of each contention-resolution scheme are described in much more detail in [14].

The packets are of average size of 12,000 bits, with negative exponential distribution. The arrival of the packets is assumed to be a Poisson process. There are three packet priority classes: class 3, 2, and 1, with class 3 being the highest priority.

In case of contention, a packet will first seek an alternative vacant wavelength on the desired output port; if none is available, it will seek a vacant wavelength in the buffer. If there are no suitable wavelengths in the buffer, it will compare its priority class with the packets currently occupying the preferred port. If the new packet has higher priority, it can preempt a lower priority packet and be transmitted successfully (either on its original wavelength or on a converted wavelength). If the two packets in contention are of the same priority, the one arriving later is switched according to the secondary next hop (from the secondary routing table) and the above process will be repeated. If the secondary routing entry fails as well, the packet will be dropped. These routing policies will conform to the actual node architecture. For example, at node 3, when a packet is destined to node 3
and contends on the same wavelength with another packet which is also destined to node 3, its wavelength cannot be converted because there will not be any wavelength converters available for ports leading to the local receivers.

### 5.2.2 Illustrative Results and Discussion

All the simulation results are plotted against transmitter load, with is the mean fraction of time when the transmitter is busy. The packet-class distribution used in the first set of simulations is: 10% class 3, 30% class 2, and 60% class 1.

Figure 5.3 shows the end-to-end delay comparison of the three classes. Because the propagation delay between two nodes is much larger than the delay introduced in the fiber delay line, a prominent increase in the end-to-end delay indicates that deflection routing is being applied to more packets. Class 1 packets experience the most increase of end-to-end delay as the load increases. This is due to the fact that they are not able to preempt any other class of packets, and have to resort to deflection more often. Class 2 and class 3 packets have similar end-to-end delays, which is because that both of these classes can preempt class 1 packets and can maintain a lower probability of using deflection routing to resolve contention.

![Figure 5.3: Comparison of end-to-end delay (10%-20%-60% priority distribution).](image)

Figure 5.4 compares the average hop distance. The average hop distance is a measurement of the physical impairment which packets suffer inside the network. When a packet
travels through fibers, switch fabrics, multiplexers/demultiplexers, wavelength converters, or when its power is tapped for header reading at each node, the signal experiences various physical impairments such as power loss, cross talk, dispersion, nonlinear effects, etc. In this study, every time a packet travels through a switch, it is counted as one hop (including the case when a packet is buffered and re-enters the same switch). There is a hop-distance counter for every packet to keep track of how many times it has traversed the switches, similar to the TTL field in the IP header. Moreover, since deflection routing is used, we set a maximum hop count equal to 8 to mitigate the looping effect. From the figure, it can be observed that all three classes have nearly the same average hop distances, which increase steadily with load. This is due to the effect of optical buffering. During contention, optical buffering is attempted before preemption or deflection takes place. All three classes are utilizing buffering with similar probability.

Figure 5.5 shows another important metric related to the network performance: packet-loss rate. Not surprisingly, class 1 packets have the highest packet-loss rate, which also increases rapidly with the increasing load. Class 2 packets have reasonably good packet-loss rate, which remains below 0.01 until the transmitter load reaches 0.2. The increase of packet-loss rate for class 3 packets are barely noticeable, and for the whole simulated range (transmitter load < 0.4), the packet-loss rate stays well below 0.01. To estimate the requirement of packet-loss rate for TCP traffic (which makes more than 90% of IP traffic), we adopt the criteria derived in [74], which states that the product of packet-loss probability and
the square of the throughput-delay product should be less than one. Here, the throughput is measured in packets/second on a per-TCP-connection basis. The nation-wide round-trip propagation delay is about 50 ms. Since, in an optical packet-switched network, there is no equivalent to a complex electrical queuing system, the delay occurred at each switching node is negligibly small. For a 1 Mbps TCP connection, the pipe throughput is approximately 100 packets/second (with average IP packet size of 12000 bits). This gives an estimated (order of magnitude) packet-loss rate of 0.01, or 1%. The simulation results show that, for a load lower than 0.2, both class 3 and class 2 packets can maintain the packet-loss rate lower than 0.01. The low packet-loss rate achieved by higher priority packets is at the cost of dropping more lower priority packets.

![Comparison of packet-loss rate (10%-20%-60% priority distribution).](image)

To better understand the source of contention, we also measured the fraction of packets dropped due to contention at the receiver, as shown in Fig. 5.6. All the lost class 3 packets are due to receiver contentions, implying that they can well survive the middle hops but eventually have to compete among themselves for the receivers. This is especially true under the case of uniform traffic, where each node in a N-node network is receiving packets from the remaining N-1 nodes. For both class 2, and class 1 packets, the fraction of packets lost at the receivers decrease gradually with the load, implying the middle-hop contentions intensify faster than the contentions at the receivers.

In order to further examine the effect of priority distribution on the network performance, we have simulated a different priority distribution: 30% class 3, 30% class 2, and
Figure 5.6: Comparison of percentage of packets lost at receiver (10%-20%-60% priority distribution).

40% class 1.

The end-to-end delay comparison in Fig. 5.7 and average hop-distance comparison in Fig. 5.8 look very similar to those in Fig. 5.3 and Fig. 5.4, indicating that, although packet-priority distribution has changed, the probability of packets being buffered or deflected did not change much.

In the packet-loss rate comparison shown in Fig. 5.9, class 2 packets experience two times more packet loss than in the previous simulation. The cause is that there are two times more class 3 packets, which will preempt both class 2 and class 1 packets; and fewer class 1 packets for class 2 packets to preempt. Figure 5.10 shows the fraction of lost packets at the receivers of all the lost packets. Most of the dropped packets in class 3 are still at the receiver side, implying that class 3 packets only compete among themselves and the contentions mostly happen at the receiver. For class 2 and class 1, it appears that more fraction of the lost packets are due to contention at the transit switches, because of the increased number of class 3 packets. For class 3 packet-loss rate to stay below 0.01, the transmitter load must be less than 0.25; and for class 2 and class 1, the corresponding transmitter load thresholds are 0.15 and 0.1, respectively.

For the three different classes of priority simulated, class 3 has the best performance in terms of latency, signal quality, and packet-loss rate. It will be suitable to carry various mission-critical, real-time traffic, and could be further explored to establish virtual
circuit-like connections to accommodate stringent user demands. Class 2, which has higher delay and higher packet-loss rate, appears to be a good candidate to carry medium-quality connection-oriented traffic (such as TCP or voice) or connection-less data traffic (such as
Figure 5.9: Comparison of packet-loss rate (30%-30%-40% priority distribution).

Figure 5.10: Comparison of percentage of packets lost at receiver (30%-30%-40% priority distribution).

UDP. Class 1 has the highest latency and highest loss rate. It can be used for applications and data services that do not require real-time connection and can recover from frequent
packet losses. It is important to note that the performance of each class is not only related with the specific routing policies and the network topology, but also how the bandwidth is divided between the classes. The priority distribution is crucial to achieve the desired packet-loss performance for each class. In general, the high quality of the higher priority service is obtained at the cost of lower priority services; therefore, the portion of bandwidth assigned to high priority classes has to be kept small. In other words, in such optical packet-switched networks, we are trading off bandwidth utilization for better flexibility and service quality.

5.2.3 Summary

For the first time, we investigated the possibility of implementing priority-based routing in an optical packet-switched network. With the use of different combinations of contention-resolution schemes and preemption according to the packet priorities, our simulation indicates that differentiated and reasonable performance can be obtained. It is shown that the more difference of load there is between different priority classes, the better the performance will be for the higher priority traffic. The priority distribution plays an important role in meeting the targeted service quality for different types of traffic. The contention resolutions used in our example is not as complex and sophisticated as the electrical queuing mechanism or some other optical packet switch prototypes [21]. For this type of network to be competitive in the future market, its cost should be low. Introducing a great amount of buffering, as suggested by many earlier literatures, to resolve contention, will introduce more switch elements and lead to higher cost. With the simple architecture in our study, we are able to obtain decent performance while keeping the node cost low.

5.3 Comparison of Slotted and Unslotted Networks with Priority-Based Routing

5.3.1 Node Architecture and Routing Policies

Figure 5.11(a) shows the network topology under study. Each link is a bi-directional fiber link, containing four wavelengths and transmitting at 2.5 Gbps. The link length is 20 km. In the slotted network, packets are of fixed size (12,000 bits) and placed in a 5 µs time slot. Before the packets enter each switching node, they must be aligned; therefore, the switching happens only at the slot boundaries. Synchronization stages are required for this type of network for the alignment of packets. In the unslotted network, packets are of variable size. The packet-size distribution in this study is set to be negative simulation exponential distribution, with the average of 12,000 bits. There are no synchronization stages needed.
and packets are switched ‘on the fly’. The packet arrival in both cases is assumed to be Poisson process. The chance of contention is greater in an unslotted network because of its asynchronous nature.

Figure 5.11(b) shows the node architecture. Each node is equipped with parametric wavelength converters, multi-wavelength fiber delay line(s) and a secondary routing table for deflection. The parametric wavelength converters can convert multiple wavelengths simultaneously, with the wavelength conversion pattern $\lambda_1 \leftrightarrow \lambda_4$, and $\lambda_2 \leftrightarrow \lambda_3$. The fiber delay line is of one average packet size. The secondary routing table is used for deflection routing. Detailed discussion of the characteristics of different contention resolutions was presented earlier Chapter 3. In our study, the slotted network node is equipped with one fiber delay line (noted as slotted$_{1buf}$), while the unslotted network node is equipped with 1, 2, 4, 8, or 16 fiber delay lines (noted as unslotted$_{1,2,4,8,16buf}$). The traffic generated between any source-destination pair is assumed to be equal, i.e., uniform traffic matrix. The packet maximum hop count is set to 8; each time a packet is buffered it is also counted as a hop.

![Network topology and node architecture under study.](image)

There are three packet priority classes: class 3, class 2, and class 1, with class 3 being the highest priority. The traffic distribution among these classes for a representative simulation example is 10% class 3, 30% class 2, and 60% class 1. The priority-based routing policies are the same as described in the previous section.
5.3.2 Illustrative Results and Discussion

Figure 5.12 shows the overall (including all three classes) packet-loss rate comparison. There is a large improvement between unslotted\textsubscript{1buf} and unslotted\textsubscript{2buf}. The amount of improvement decreases quickly with increasing number of buffers. There is no significant improvement between unslotted\textsubscript{8buf} and unslotted\textsubscript{16buf}. The packet-loss rate of slotted\textsubscript{1buf} is between unslotted\textsubscript{4buf} and unslotted\textsubscript{8buf} and stays well below 0.01 for transmitter load under 0.3. Although, by nature, the slotted network performs much better than unslotted network, it involves complicated synchronization stages at the switching nodes. The unslotted network can accommodate packets in its native size and does not require the synchronization stages. From the simulation results, we can tell that, with more buffering capacity, the unslotted network can provide reasonable performance. Having more delay lines means increasing the port count of the switch fabric, which could be a concern because it will add to the cost of the network.

Figures 5.13 and 5.14 show the packet-loss rate comparison for class 1 and class 2 traffic. Both these figures have similar trend as Fig. 5.12. For class 1 packets, unslotted\textsubscript{4buf} can match the packet-loss rate of slotted\textsubscript{1buf}, but for class 2 packets, slotted\textsubscript{1buf} outperforms all the unslotted configurations, regardless of the number of buffers. For class 1 packet to have packet-loss rate less than 0.01 with transmitter load less than 0.3, there needs to be four fiber delay lines, while for class 2 with four fiber delay lines they can achieve packet-loss rate less than 0.01 under transmitter load less than 0.35.

Figure 5.15 shows the packet-loss rate comparison for class 3 packets. Interestingly, under heavy load, the packet-loss rate in the unslotted network increases with the number of buffers. This is against the intuition that more buffering will bring down the packet-loss rate. This phenomenon is caused by the large amount of buffering and the maximum hop-count mechanism. In the case of contention, a higher-priority packet is more likely to be successfully buffered than to preempt lower-priority packets. Therefore, more packets are dropped because of reaching the maximum hop count. For all the unslotted network configurations, class 3 packet-loss rate stays well below 0.001 with transmitter load less than 0.2 and below 0.01 with transmitter less than 0.35.

5.3.3 Summary

This section presented a comparison between a slotted and an unslotted network. By increasing the number of fiber delay lines in the unslotted network node, we are able to reach the performance of a slotted network with one fiber delay line. It is shown that, for the topology in Fig. 5.11(a), four fiber delay lines per node will provide packet-loss rate less than 0.01 for all three priority classes with transmitter load less than 0.3. The results
Figure 5.12: Overall packet-loss rate.

Figure 5.13: Class 1 packet-loss rate.
Figure 5.14: Class 2 packet-loss rate.

Figure 5.15: Class 3 packet-loss rate.
indicate that it is possible to avoid the complicated packet fragmentation/re-assembly and synchronization stages required by the slotted network and accommodate variable packet size without sacrificing network performance.

5.4 A Hybrid Contention-Resolution Scheme

There exist a number of studies on various node architectures and contention-resolution algorithms. Most of these studies focus on the optical domain of one optical packet switch, or a network inter-connected by such a switch. An optical packet-switched network, by its name, should perform packet-based switching optically. However, it does have to interface with other types of networks to provide end-to-end connectivity. The other networks, especially the client networks, are often electrical. This means that there should be an interface at the edge of the optical packet-switched network. In this section, we propose to take advantage of the availability of electrical buffers at the edge, to resolve contentions, improve performance, and lower network cost. Our simulation experiments indicate that this method offers excellent network performance.

5.4.1 Network Architecture

When packets arrive from client networks, which are mostly electrical, they need to be converted into optical format before being sent to the optical packet-switched network. This conversion is performed at the client interface of the network (see Fig. 5.16). The optical packet switch performs two types of packet forwarding: the forwarding of transit packets from other optical packet switches, and the forwarding of local packets received from the client interface. A transit packet has to cope with possible contention from the local packets as well as other transit packets. In most proposed architectures, the contention resolution usually requires a large amount of optical resources, such as wavelength converters and delay lines. In our proposed contention-resolution scheme, the local packets are first queued in the electrical buffers, which can be easily implemented in the electrical part of the client interface. These packets enter the optical switch only when there is no transit packet occupying the preferred wavelength/output port. This buffering mechanism ensures that all the wavelength converters and delay lines are only used for transit packets. Since the switching is still carried out by the optical components and there is no O-E-O conversion in the network, the use of electrical buffers at ingress nodes for the client packets does not compromise the all-optical nature of the core network.
Figure 5.16: Client interface of optical packet-switched networks.

Figure 5.17: Architecture of the proposed hybrid contention-resolution scheme.

**Node Architecture**

Figure 5.17 shows the node architecture that implements the proposed hybrid contention-resolution scheme. In the optical portion of the switch, both optical delay lines and wave-
length converters are used to resolve contention. During contention, a packet will first seek an available wavelength on the preferred output port. If no vacant wavelength is available, it will seek a vacant wavelength in the optical delay line for buffering. If no buffering is available, deflection is used to route the packet to a second preferred port.

In the electrical-optical interface portion of the switch, FIFO queues are included. All the packets from client networks will be queued first. A scheduler observes the state of every wavelength/output port of the switch from the optical portion. When the packet’s output port clears, the transmitter will convert the packet to optical format and send it to the optical switch fabric.

5.4.2 Simulation Configuration

Figure 5.11(a) shows the network topology. It consists of 15 nodes, with the node architecture similar to that of Fig. 5.17. On each link of the network, there are two fibers to form a bi-directional link. Each link is 20 km long. The fibers contain \( W \) wavelengths, with the line rate of 2.5 Gbps. Every node is equipped with \( W \) transmitters, with corresponding FIFO queues. A FIFO queue is fed with a self-similar IP packet generator consisting of 12 ON/OFF sources (see Section 3.3) The traffic is uniformly distributed. The scheduler constantly monitors the state of the wavelengths at the output ports. Whenever the FIFO queue is not empty and there is a vacant wavelength on the preferred output port, the scheduler will retrieve the packet at the head of the queue and send it to the optical switch fabric. The switch includes a number of optical delay lines, which are of length of 12500 bits (500 bits more than the maximum packet size of 12000 bits). In the polynomial fitting of packet sizes, we set the maximum packet size to be 1500 bytes (12000 bits), since the probability of having IP packets larger than 1500 bytes is negligibly small in real data traffic. This ensures that a delay line can fully accommodate any IP packet.

As shown in Fig. 5.17, there are wavelength converters at each output port of the switch fabric. In our simulation experiments, the number of wavelength, \( W \), is set to 4, 8, 16, and 32. Since nodes with more incoming/outgoing links perform more routing than others, the number of optical delay lines at each node is set according to the nodal degree. After experimenting with various numbers of delay lines at each node, we choose the number of delay lines to be \((\text{nodal degree} - 1)\) to produce good results. Deflection is the third contention resolution. For more discussion on the different contention-resolution schemes, please see Chapter 3.
5.4.3 Illustrative Results and Discussion

Figure 5.18 shows the packet-loss rates plotted against the average offered transmitter load. For the 4-wavelength scenario, the packet-loss rate is kept below 0.01 when the offered transmitter load is less than 0.5. For the 32-wavelength case, the acceptable transmitter load is close to 0.65. The light-load portion of the plot is not shown on the figure, because the simulation did not encounter any dropped packets during the simulated time.

Figure 5.19 shows the average end-to-end delay. The ideal average hop distance of this network topology is 2.42, corresponding to 0.242 ms propagation delay with each link 20 km long. With average transmitter lower than 0.6, the end-to-end delay is dominated by propagation delay. The delay introduced by optical delay lines is much lower than the link propagation delay. Moreover, the electrical queuing delay in the access FIFO queues are also negligibly small. This is because the queued packet does not have to wait for more than a few packets’ transmission delay to find a vacant wavelength. With 2.5 Gbps line rate and native IP packet sizes, the transmission delay of packets are on the order of microseconds. Since the network uses static routing, there will be a most-congested link in the network. In this topology, the most-congested link is the one between node 1 and 11. There is traffic from 21 source-destination pairs passing this link. This link becomes congested when the average transmitter load reaches 0.66. When the link is approaching congestion, electrical queuing delay and deflection delay start dominating. This explains the significant increase in delay in the figure.

5.4.4 Summary

This section presented an optical-electrical hybrid contention-resolution scheme for optical packet-switched networks. This scheme uses the electrical buffer available at the ingress of the network to buffer packets before they enter the optical domain. Without introducing any noticeable extra latency, this mechanism largely improves the efficiency of optical contention-resolution resources, such as wavelength converters and optical delay lines. The simulation experiments adopted self-similar IP packet traffic. With the given topology and 32 wavelengths, the proposed scheme is able to achieve a packet-loss rate less than 0.01 with average transmitter load up to 0.6.
Figure 5.18: Packet-loss rate of different simulation scenarios.

Figure 5.19: Average end-to-end delay of different simulation scenarios.
5.5 TCP Performance with Electrical Ingress Buffering and Packet Aggregation

The previous section proposes to use inexpensive electrical buffers at the ingress interfaces to reduce packet loss. Since our goal is not only to lower packet-loss rate, but also to provide a better transport for the TCP/IP traffic, this section extends the investigation to the TCP performance of optical packet-switched networks. To improve the TCP performance, the work in [77] proposed optical flow-routing, a type of aggregation that is supposed to reduce packet reordering. Our work proposes a packet-aggregation mechanism that allows many packets to be grouped together into a larger entity that can be more efficiently transported through the network. Such a mechanism can help reduce the traffic burstiness and improve the system performance.

5.5.1 Node Architecture

Figure 5.20 shows the proposed node architecture. The packet aggregator assembles client packets into larger entities (referred to as aggregation packets) in a FIFO manner. It directly interfaces with the client-network elements (typically IP routers), and it consists of a number of FIFO sub-queues. Each sub-queue buffers packets going to the same destination. A sub-queue transmits all the buffered packets in an aggregation packet after a certain period of time, $t_a$. To avoid unnecessary delay, we set a packet-count threshold $C$ such that, when the number of buffered packets reaches $C$, the sub-queue will transmit the aggregation packet even if the time from last transmission is less than $t_a$. (Alternatively, $C$ can base on the total number of received bits, instead of packets [78].) This aggregation mechanism can be compared to a bus system (as in public transportation): At any time, there is one bus with one or more empty seats waiting for passengers for each destination. A bus has a maximum capacity of $C$ passengers and it leaves every $t_a$ seconds. If the bus is full before its scheduled departure time, it will leave early and the next empty bus will pull into the station. The aggregator not only preserves the order of packets, but also shapes the traffic by injecting more evenly-sized aggregation packets at more regular time intervals.

The ingress buffer controls when an aggregation packet can be injected into the optical switch in order to avoid contention. By using an ingress buffer, we can reduce its contention with transit packets by allowing local aggregation packets to be injected only when there is no transit packet occupying the preferred output port. The local aggregation packets are first stored in the ingress buffer electrically, and they are converted to optical format and then injected into the optical switch. A scheduler is used to constantly monitor the state of the switch fabric, and to control the transmission of the ingress buffer.
5.5.2 Simulation Configuration and Illustrative Results

Figure 5.21 shows the network topology for the simulation. The experiment uses a file transfer protocol (FTP) session to measure the TCP performance. The main performance metric is the transfer time of a large file (assumed to be 1.6 Mbytes in this example). It is reasonable to assume that both hosts have Ethernet interfaces; therefore, the maximum transfer unit (MTU) is 1500 bytes. For the network scenario to be realistic, each link also carries some background traffic. The self-similar traffic generator consists of 100 ON/OFF packet trains. The packet-size distribution is based on IP packet sizes, as described previously. Each node is equipped with four transmitters, fed by four traffic generators independently. The intensity of the background traffic is controlled by the average offered transmitter load (TX load).

One of the main factors that affect TCP performance is the receiver window size, whose typical values are 8, 32, or 64 Kbyte. The aggregation threshold $C$ can also impact the TCP performance. With $C$ values, the aggregation timer value $t_a$ and the delay-line size should be adjusted accordingly. In the experiments, both $t_a$ and the delay-line size are set to be equal to the transmission delay of $C$ packets with maximum length (1500 bytes each).
The simulation experiments were done with the **OPNET** simulation tool, and the running time for each data point varies between 4 and 75 hours on a 500-MHz Pentium III machine, depending on the TX load. The maximum TX load was 0.5 because larger values made the simulation time prohibitively long.

![Network topology for TCP experiment.](image)

**Figure 5.21:** Network topology for TCP experiment.

**Figure 5.22** compares the file transfer time $T_{FTP}$ for different TCP window sizes and different values of $C$. For reference, it also shows the $T_{FTP}$ without any background traffic for a client-server pair directly connected through a 100 Mbps link with the same propagation delay, i.e., a link length of 60 km. Without aggregation, a window size of 32 Kbyte provides the best result because the measured TCP round-trip time (RTT) is approximately 3 ms, and the TCP connection’s data rate is 100 Mbps. (Note that the optimal window size should be the product of RTT and the data rate.)

Next, in **Fig. 5.23**, one can see the effect of the aggregation threshold with different $C$ values of 10, 30, and 100 packets, while window size is equal to 8 Kbyte. For a TX load less than 0.2, the aggregation threshold does not have any effect on the system performance. As the TX load increases, the 10-packet aggregation scheme has the lowest $T_{FTP}$, followed by the 30-packet and the 100-packet schemes. The 10-packet scheme also performs better than the one without aggregation, indicating that aggregation improves TCP performance. However, with more packets aggregated, the performance deteriorates because more queuing delay is introduced in the packet aggregator and the ingress buffer. Intuitively, one would imagine that the ideal aggregation packet should contain all the TCP segments sent within one window size. Unfortunately, the aggregator has to hold the first segment for at least the whole transmission delay of all the segments in that window. This defeats the purpose of pipelining in the TCP sliding window. Therefore, aggregation does not appear to directly improve the TCP performance. However, aggregation reduces the traffic burstiness and lowers the packet-loss rate.

**Figure 5.24** shows the packet-loss rates of aggregation schemes with different values of
C. The 10-packet and 30-packet schemes offer the lowest packet-loss rate, followed by the 100-packet scheme and the no-aggregation scheme. Large aggregation appears to be less effective because the aggregation threshold \( C \) is based on the number of packets instead of bits. Since a large portion of IP traffic consists of very small packets, the size of the aggregation packet can vary significantly when \( C \) is large, and the inter-departure time of the aggregation packets can still be quite bursty. Hence using bit count, instead of packet count, might be a better approach for aggregation.

### 5.6 Conclusion

This chapter presented the unifications and extensions of the contention-resolution schemes. The simulation results show that priority-based routing is one possible solution to implement Class of Service (CoS). Although unslotted networks do not perform as well as slotted networks, they appear to be easier to implement and offer satisfactory performance. With the use of electrical buffer at the ingress nodes, an optical-electrical hybrid contention resolution significantly improves the network performance. The last section of the chapter investigates the TCP performance with ingress buffering and packet aggregation.

![Figure 5.22: Comparison of \( T_{FTP} \) for different TCP window sizes.](image-url)
Chapter 5: Unification and Extension of Contention-Resolution Studies in...

Figure 5.23: Comparison of $T_{FTP}$ for different aggregation schemes.

Figure 5.24: Comparison of packet-loss rate.
Chapter 6

Optical Packet Switching for Metropolitan-Area Networks: Opportunities and Challenges

6.1 Introduction

So far, this dissertation has discussed various design and implementation issues. Nevertheless, the application of optical packet switching is equally important. This chapter will present a high-level view of how such networks can be integrated with other network segments and provide users end-to-end connectivity with performance and simplicity.

6.2 Optical Packet Switching for MAN

6.2.1 Network Requirements

In an end-to-end connectivity picture, the current networks consist of three major segments: the access network, Metropolitan-Area Network (MAN), and the backbone Wide-Area Network (WAN). The access network is responsible for collecting end-user traffic and is usually less than a few tens of miles in its extent. Examples of access networks include intra-building Ethernet and networks operated by local Internet service providers (ISP’s). The MAN is responsible for transporting traffic between different access networks and route part of the traffic onto the backbone WAN. The MAN usually does not exceed a few hundreds of miles in size. The backbone WAN interconnects MAN’s that are typically hundreds to thousands of miles away. Optics plays a key role in transmission in the backbone WAN. These three segments of networks complete the end-to-end delivery path for user data, and each of them
has different characteristics. The access network does not need to offer high bandwidth as MAN or WAN, since it deals primarily with low-end users. The main consideration for building an access network is to keep the cost low. Therefore, most access networks consist of electrical media (copper wire) or passive optics (low-cost fiber systems without regeneration) with a simple medium-access control (MAC) protocol. In MAN, the clients are different access networks and high-end users (such as financial institutions and large ISP’s, who require reliable, high-bandwidth connectivity nationwide). The MAN needs to provide a large variety of service qualities, such as best-effort-based, connection-less datagram delivery, and/or QoS-based connection-oriented virtual circuits. It should also be scalable to accommodate the rapidly growing number of access networks. Another issue a MAN has to cope with is the time-dependency of traffic patterns. The heavy load of the network may move from the business district in downtown during working hours to the residence area in the suburbs in the evening. Part of the MAN traffic will travel inside the same MAN, while the rest have to reach another MAN through the backbone WAN. The WAN have varying extent of link lengths, typically long (>100 miles) but some times shorter in densely-populated areas where a number of a MAN’s reside close to one another. The traffic on the backbone WAN, compared with that of MAN, is often aggregated, groomed, relatively more static, and predictable. High-bandwidth lightpaths are the main logic links to build the connectivity for WAN’s.

Numerous companies target MAN as their primary market because it can have flexible network architectures and its market is rapidly growing. Future MAN’s must provide users with high bandwidth, together with different granularities of bandwidth. Meanwhile, the emergence of various Internet applications and the ever-growing number of users demand a future-proof MAN with excellent scalability, multi-protocol support, and rapid-provisioning capabilities. Typical MAN’s of today are based on ring topology and SONET technology (Fig. 6.1). Connectivity is established by using SONET add/drop multiplexers (ADM’s), and digital crossconnects (DXC’s) in the inter-connected-ring case. SONET equipment provides Time-Division Multiplexing (TDM) sub-channel switching within a wavelength. SONET terminals, ADM’s, and DXC’s all require OEO conversion and time multiplexing/demultiplexing. Lately, a few companies started manufacturing optical crossconnect-based products, which are capable of performing wavelength switching without time demultiplexing the data carried by the wavelength. The recent Multi-Protocol Lambda Switching (MPAS) and Generalized Multi-Protocol Label Switching (GMPLS) activities in IETF further suggest using MPLS protocols to expedite wavelength provisioning (and therefore end-to-end connectivity when combined with IP routers). Although optical crossconnects (OXC’s) and optical add-drop multiplexers (OADM’s) can circumvent processing the bits and maneuvering them in the time domain, the granularity cannot be finer than a single
wavelength. Therefore, the flexibility of such MAN’s constructed with OXC’s and OADM’s is limited.

![SONET ring-based network](image)

**Figure 6.1: The SONET ring-based network.**

### 6.2.2 Promises of Optical Packet Switching

Optical packet switching, while combining high throughput and packet-level switching, appears to be a good candidate for MAN applications. Although most of the past investigation took place in research institutions and there are no standards regarding the architectures to implement optical packet-switching, there are two common features that an optical packet switch should have:

1. The switching takes place in the optical domain without OEO conversion. Optical switching ensures higher throughput and less power consumption.

2. The switch should be able to perform packet-level or, more generally, sub-wavelength switching. This is what differentiates an optical packet switch from an OXC. Theoretically, the switch fabric should also be able to switch with more coarse granularities such as wavelengths.

Figure 6.2 shows a generalized illustration of an optical packet-switching node.

The input stage is responsible for pre-amplification when necessary, reading the packet headers, and packet alignment if so desired. Delay lines are used between the input stage and the switch unit to give enough time for the header-processing unit to configure the switch before the packets enter the switch unit. The header-processing unit reads the header and set up the switch accordingly. The switch unit, besides switching the packet to the desired port, also carries out contention resolutions. The output stage might perform
header re-writing and power amplification if necessary.

The ubiquity of IP routers today is an example that packet switching is more capable of adapting to the fast changing networks, because of its high flexibility. An IP network can carry connection-oriented (TCP) traffic as well as connection-less (UDP) traffic, and be built upon various lower-layer technologies (ATM or SONET). Similarly, optical packet-switched networks will be able to carry traffic from different upper layers. The limitations of IP include: (a) the simple destination-based routing imposes limitations on the routing functionality and traffic engineering and (b) switching in the electronic domain is limited by the speed and the power consumption of the electronics. MPLS successfully solves the first issue by using exact-match forwarding algorithm and constraint-based routing. Optical packet switching will further remove the OEO conversion bottleneck and offer a combination of high throughput, considerable flexibility, and a rich set of routing functionality similar to that of MPLS.

The future MAN will be a very dynamic network, consisting of a large variety of users. Different users will have different service requirements. For example, the corporate clients are likely to have traffic mainly generated by IP, and therefore require high-bandwidth, IP-based connectivity between major offices. High-end clients, such as financial institutions, are likely to request highly-reliable, connection-oriented connectivity. Campus networks and local ISP’s are likely to generate IP traffic mainly consisting of web browsing; therefore, they are expected to request best-effort IP-based connectivity. Optical packet switching will be able to provide a large variety of services to meet these different needs because of its capability to switch with different bandwidth granularities. An optical packet switch, since
it does not read the actual bits in the payload data, can stay in a given switching state for an arbitrary amount of time. This unique feature enables it to provide a large range of services. For example, for campus networks or low-end ISP’s, it can provide best-effort service; for corporate clients, it can provide reliable virtual-circuit (similar to ATM) service; and for high-end clients, it can provide a whole protected lightpath [79].

Another advantage of using optical packet switching for MAN is that it can use the existing protocol suites for routing and signaling. The on-going work at IETF is extending the MPLS protocol suite to build a seamlessly integrated network of IP routers running with OXC’s [80]. It is likely that, in the future, the generalized MPLS will be running on both label-switched routers (LSR’s) and OXC’s. Since an optical packet switch can function both like an LSR and an OXC, it will be easy to run the same protocols and integrate them with the existing equipment.

### 6.2.3 Inter-Network Interfaces

The aforementioned characteristics of optical packet switching make it an excellent candidate for a MAN. But to complete the end-to-end connectivity, we also need to consider how the optical packet-switched MAN interfaces with the rest of the networks. On the client side are the access networks and high-end users, whose networks most likely consist of IP routers or ATM switches. On the other side, the MAN will have one or more egress nodes connected to the high-speed backbone WAN.

Figure 6.3 shows the client side of an optical packet-switching node. The client interface should be able to process layer-three/layer-two traffic, or connection requests when a wavelength is requested, based on different protocols such as IP, ATM, and SONET. In the case of IP packets, the client interface will first look at the IP header, assign an optical header to it according to the routing table, and transmit the packet to the next optical packet-switching node. To fully exploit the capacity of optical packet switching, the client interface should be a high performance edge label-switched router-like device with optical interfaces in order to aggregate user traffic and switch it at optical line speed.

There are two options as to how to send packets from the optical packet-switched MAN to the backbone WAN. The first option is to have the egress node perform OEO conversion, read the layer-three or layer-two headers, and then transmit the packets onto the established lightpaths through the backbone (Fig. 6.4(a)). This is similar to the overlay model in MPλS, where the inside of the optical-crossconnected network is opaque to the outside. The other option, as shown in Fig. 6.4(b), is to utilize label stacking to deliver the packets to the desired optical packet-switched node or LSR on the other side of the backbone. This will require the label-binding information from several network segments to be propagated through the whole delivery path during the label-switched path (LSP) setup period. In the
example in Fig. 6.4(b), a packet has to go from node X.1 to node Z.3. Networks X and Z are optical packet-switched MANs, while network Y is the optical-crossconnected backbone. Network Y is assumed to be running MPλS and label-binding information is implicit in the wavelength index b. During the LSP setup phase, both network Z and network Y will send the label-binding information to node X.1. (In the context of optical packet switching, we assume network segment-wide unique label because label swapping in optical domain is a nontrivial task, unlike in the electronic domain.) Node X.1 will stack the labels from all three networks (label a from network Z, label/wavelength index b from network Y, and label c from network X), form an optical header and assign it to the packet. When the packet is leaving network X, the egress node will pop one label from the stack and the ingress node at network Y will use the next-level label (b) to assign the wavelength. At the egress node of network Y, the same label-pop process will take place and the packet will eventually reach the desired egress node Z.3. If optical label swapping is applicable, it will not be necessary for the label-binding information to be passed to the source node. The labels will be of local significance only.

6.3 Enabling Technologies

The core of an optical packet-switching node is the switch fabric. To be able to switch on packet level at multi-Gbps speed, the switch must have nanosecond switching time. Micro electrical-mechanical system (MEMS) technology is not applicable here because the MEMS switches can only perform millisecond switching. Potential candidates include LiNbO₃-based switch elements and semiconductor optical amplifier (SOA) gate switch elements. An alternative is to use ultra-fast tunable lasers, wavelength converters, and static wavelength routing devices, such as Arrayed Waveguide Gratings (AWG’s), to construct a fast, large-port-count switch fabric. This approach can avoid concatenating multiple stages of 2x2
Figure 6.4: Two options to inter-connect the optical packet-switched MAN with optical-crossconnected backbone. (a) Using electrical equipment at the interface. (b) Using label stacking and end-to-end signaling.

Wavelength conversion remains one key technology in contention resolutions. Although the current stage of wavelength converters is not as mature as other devices, it is only a matter of time to have off-the-shelf wavelength converters. For example, Alcatel already has a packaged optical, active Mach-Zehnder structure-based wavelength converter operating at 10 Gbps [81]. Parametric wavelength conversion may be of special interest in optical packet switching because of its capability of converting multiple wavelengths simultaneously [70].

The current research in optical RAM is still in an early stage, while there are some
promising discoveries in laboratory level (e.g., the chiropticene molecular switch [82] and molecular transistor [83]). Before optical RAM becomes available, we still need to rely on fiber delay lines as optical buffers.

On the software side, the optical packet-switched network can utilize the existing MPLS and MPλS protocols with necessary extension for signaling, and IP routing protocols for routing information distribution. There is no need to re-invent another protocol layer exclusively dedicated to optical packet switching.

6.4 Summary

As the data traffic continues to grow rapidly, the MAN will have to manage a large variety of clients and meet different client requirements in terms of bandwidth provisioning and service qualities. Optical packet switching appears to be a suitable candidate to build a flexible, high-throughput, and scalable MAN. Powered by the generalized MPLS protocols, such optical packet-switched networks can provide different switching granularities, such as packets, bursts, and circuits, and support different user data formats. To compete with high-speed IP routers and LSR’s, the optical packet switch should operate beyond 10 Gbps. Optical regeneration will be necessary as the bit rate approaches 40 Gbps.

With the current switching and buffering technology, we still have to make a trade-off between packet-loss rate, network utilization, and node complexity. Fast, large-port-count switch fabrics and wavelength converters remain the key enabling technologies before there is a breakthrough in optical RAM.
Chapter 7

Conclusion

This dissertation makes five important contributions to the body of knowledge on the design and analysis of optical packet-switched networks. The architectures, algorithms, and modeling techniques developed in this dissertation can be used in the design and implementation of the next-generation optical Internet. This chapter summarizes the main results and contributions in this dissertation.

7.1 Advances in Optical Packet Switching: An Overview

As telecommunications and computer communications continue to converge, the data traffic is gradually exceeding the telephony traffic. This means that many of the existing connection-oriented, circuit-switched networks will need to be upgraded to support packet-switched data traffic. The concept of WDM has provided us an opportunity to multiply the network capacity. Current optical-switching technologies allow us to rapidly deliver the enormous bandwidth of WDM networks. Among all the switching schemes, optical packet switching appears to be a strong candidate because of the high speed, data rate/format transparency, and configurability it offers. Chapter 2 discussed some of the critical issues involved in designing and implementing optical packet-switched networks.

In general, optical packet-switched networks can be divided into two categories: slotted (synchronous) and unslotted (asynchronous) networks. When individual photonic switches are combined to form a network, at the input ports of each node, packets can arrive at different times. Since switch fabric can change its state incrementally (set up one input-output connection at a time) or jointly (set up multiple input-output connections together at the same time), it is crucial for the network designer to decide whether to have all the packets aligned before entering the switch fabric.

In a packet-switched network, each packet has to go through a number of switches to
reach its destination. When the packets are being switched, contention occurs whenever two or more packets are trying to leave the switch from the same output port. How the contention is resolved can have a large impact on the network performance. Chapter 2 briefly discussed three types of contention resolutions: optical buffering, deflection routing, and wavelength conversion. Various existing node architectures were studied.

7.2 Contention-Resolution Schemes in Optical Packet-Switched Networks

In electrical packet networks, contention is usually resolved with the store-and-forward technique, which requires the packets in contention to be stored in a memory bank and sent out at a later time when the desired output port clears. This is possible because of the availability of electronic RAM. There is no equivalent applicable optical RAM technology; therefore, different approaches have to be used. Meanwhile, WDM networks provide one new additional dimension, namely wavelength, for contention resolution. Chapter 3 explored all three dimensions for contention-resolution schemes: wavelength, time, and space in an irregular mesh network.

Many previous works on optical packet networks employ optical delay lines to resolve contentions. The deflection approach to resolve contentions has been studied for electronic networks, while the wavelength-conversion approach has also received attention recently. Chapter 3 presented a comprehensive study of the advantages and trade-offs of the three approaches and investigated other factors, such as network topology, that can affect the effectiveness of these approaches.

A fair amount of work has been conducted on slotted networks with fixed-length packets, in terms of network architecture and performance. In an unslotted network, the packets are not aligned and they are switched one by one ‘on the fly’. While unslotted networks generally have lower cost (no expensive synchronization stages), higher flexibility (variable packet length), and robustness (simpler control), they may have slightly lower overall throughput than synchronous networks, because of the increased contention probability. Since today’s WDM networks offer high aggregate bandwidth, we are interested in the performance of unslotted networks with different contention-resolution schemes. Chapter 3 focused on an unslotted irregular mesh network since most of the published work has been carried out on slotted networks or unslotted networks with regular topology.

The simulation experiments were based on a large network with a number of contention-resolution schemes, such as different combinations of wavelength conversion, optical buffering, and space deflection. This chapter also investigated the impact of selective deflection and limited wavelength conversion.
7.3 PLATO: A Generic Modeling Technique for Optical Packet-Switched Networks

One of the objectives in designing an optical packet-switched network is high performance, or low packet-loss rate. Packets in contention may be dropped when there are more packets traveling to the output port of a switch than the port can actually accommodate. Various node architectures have been proposed in the past, as well as contention-resolution algorithms in time, space, and wavelength domains. Although some analytical studies have been presented for different node architectures, simulation is still the dominant tool for studying the performance of large, practical networks.

Chapter 4 presented PLATO, a generic modeling technique for evaluating packet-loss probability for optical packet-switched networks. This technique can accurately model different topologies and node architectures.

In the traffic-analysis part of PLATO, an algorithm is provided to decompose all the traffic flows and to construct a system of equations that represents the contentions at every output port of every node. In the switch-modeling part, each switch is modeled in such a way that the packet-loss probability of any input packet stream can be computed as a function of the intensities of all contending streams. These functions representing local packet-loss probability are called by the system of equations from the traffic-flow analysis. Through numerical iterations, the values of the packet-loss rate of all the flows will converge and give the network-wide packet-loss rate. Comparison between analytical and simulation results shows excellent accuracy of the PLATO technique. It is also demonstrated that PLATO performs well with different topologies and node architectures; therefore, it is a good generic modeling technique for optical packet-switched networks.

The main contribution of the PLATO technique is that it combines network-traffic-flow analysis, switch modeling, and a numerical method to accurately compute both the network-wide and local per-flow-based packet-loss probability, regardless of the network topology and node architecture. We believe that our present work can serve as an excellent foundation for further exciting research on this subject.

7.4 Unification and Extension of Contention-Resolution Studies

Chapter 5 presented a unified and extended study of contention-resolutions schemes: priority-based routing, comparison of slotted and unslotted networks, proposal of an optical-electrical hybrid contention-resolution scheme, and a study of TCP performance with ingress electrical buffering and packet aggregation.
A network built on optical packet switching will have to provide not only IP-like connection-less services, but also higher-quality, connection-oriented services. This will require the network to have differentiated service qualities at the packet level. It is worthwhile to study the network behavior with different packet priorities, which will lead us one step closer to the implementation of Class of Service (CoS). For the first time, our study investigated the possibility of implementing priority-based routing in an optical packet-switched network. With the use of different combinations of contention-resolution schemes and preemption according to the packet priorities, our simulation results indicated that differentiated and reasonable performance can be obtained. It was shown that the more difference of load there is between different priority classes, the better the performance will be for the higher priority traffic. The priority distribution plays an important role in meeting the targeted service quality for different classes of traffic.

In the comparison between a slotted and an unslotted network, the results indicate that the unslotted network can have the same performance of a slotted network when there are more fiber delay lines. It was shown that, for a representative network topology, four fiber delay lines per node will provide packet-loss rate less than 0.01 for all three priority classes with transmitter load less than 0.3. It is possible to avoid the complicated packet fragmentation/re-assembly and synchronization stages required by the slotted network and accommodate variable packet sizes without sacrificing network performance.

Most of the proposed contention-resolution schemes and optical packet switch architectures focus on the optical domain. An optical packet-switched network, by its name, should perform packet-based switching optically. However, it does have to interface with other types of networks to provide end-to-end connectivity. The other networks, especially the clients networks, are often electrical. This means that there has to be an interface at the edge of the optical packet-switched network. It is possible to take advantage of the availability of electrical buffers at the edge to resolve contentions, to improve performance, and to lower the network cost. Such a hybrid contention-resolution scheme can offer satisfactory network performance. For a given topology and 32 wavelengths, it achieved a packet-loss rate less than 0.01 with average transmitter load below 0.6.

Since our goal is not only to lower packet-loss rate, but also to provide a better transport for the TCP/IP traffic, we extended the investigation to the TCP performance of optical packet-switched networks. Our research proposed a packet-aggregation mechanism that allows many packets to be grouped together into a larger entity that can be more efficiently transported through the network. Such a mechanism can help reduce the traffic burstiness and improve the system performance.
7.5 Optical Packet Switching for MAN

In an end-to-end connectivity picture, the current networks consist of three major segments: the access network, MAN, and the backbone WAN. The access network is responsible for collecting end-user traffic and is usually less than a few tens of miles in its extent. The MAN is responsible for transporting traffic between different access networks and route part of the traffic onto the backbone WAN. The backbone WAN interconnects MAN’s that are typically hundreds to thousands of miles away. These three segments of networks complete the end-to-end delivery path for user data, and each of them has different characteristics.

The MAN needs to provide a large variety of service qualities, such as best-effort-based, connection-less datagram delivery and/or QoS-based connection-oriented virtual circuits. It should also offer scalability to accommodate the rapidly-growing number of access networks. Another issue a MAN has to cope with is the time dependency of traffic patterns.

The future MAN will be a very dynamic network, consisting of a large variety of users. An optical packet switch, since it does not read the actual bits in the payload data, can stay in a given switching state for an arbitrary amount of time. This unique feature enables it to provide a large range of services. Another advantage of using optical packet switching for a MAN is that we can use the existing protocol suites for routing and signaling. The on-going work at IETF is extending the MPLS protocol suite to build a seamlessly integrated network of IP routers running with OXC’s. It is possible that, in the future, the generalized MPLS will be running on both LSR’s and OXC’s. Since an optical packet switch can function both like LSR and OXC, it will be easy to run the same protocols and integrate them with the existing equipment.

As the data traffic continues to grow rapidly, the MAN will have to manage a large variety of clients and meet different requirements in terms of bandwidth provisioning and service qualities. Optical packet switching appears to be a suitable candidate to build a flexible, high-throughput, and scalable MAN. Powered by the generalized MPLS protocols, such optical packet-switched networks can provide different switching granularities such as packets, bursts, and circuits, and support different user data formats. To be able to compete with high-speed IP routers and LSR’s, the optical packet switch should operate beyond 10 Gbps. Optical regeneration will be necessary as the bit rate approaches 40 Gbps.
Bibliography


detection and swapping technique incorporating a fiber Bragg grating filter,” IEEE

switching and routing by rapidly tunable wavelength conversion and uniform loss cyclic
frequency array-waveguide grating,” Proc. OFC ’01, vol. 3, pp. wdd49–1 – wdd49–3,

architecture based on photonic slot routing,” IEEE/ACM Transactions on Networking,

transparent scalable WDM wide area networks,” Photonic Network Communications,

[51] H. Zang, J. P. Jue, and B. Mukherjee, “Capacity allocation and contention resolution
in a photonic slot routing all-optical WDM mesh network,” IEEE/OSA Journal of

based on photonic slot routing and switched delay lines,” Proc. IEEE INFOCOM ’97,

vol. 21, pp. 36–42, Nov. 1983.

[54] G. C. Hudek and D. J. Muder, “Signaling analysis for a multi-switch all-optical net-
work,” Proc. IEEE International Conference on Communications (ICC ’95), pp. 1206–
1210, June 1995.

in optical networks,” Proc. IEEE/LEOS Technologies for a Global Information Infra-


length conversion by difference-frequency generation in AlGaAs waveguides with peri-
2611, 1996.

[71] S. J. B. Yoo, “Reduced parametric wavelength-interchanging crossconnect architectures 

[72] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, “Self-similarity through 
high-variability: statistical analysis of Ethernet LAN traffic at the source level,” 


bandwidth-delay products and random loss,” *IEEE/ACM Transactions on Networking*, 

a wavelength-interchanging crossconnect utilizing parametric wavelength converters,” 


[78] F. Xue, S. Yao, and S. J. B. Yoo, “The performance improvement in optical packet-
switched networks by traffic shaping of self-similar traffic,” *Submitted to OFC2002*, 


