


2007

# Fiber optic networks: fairness, access controls and prototyping

Nathan Allan VanderHorn  
*Iowa State University*

Follow this and additional works at: <https://lib.dr.iastate.edu/rtd>

 Part of the [Computer Sciences Commons](#), and the [Electrical and Electronics Commons](#)

## Recommended Citation

VanderHorn, Nathan Allan, "Fiber optic networks: fairness, access controls and prototyping" (2007). *Retrospective Theses and Dissertations*. 15510.  
<https://lib.dr.iastate.edu/rtd/15510>

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact [digirep@iastate.edu](mailto:digirep@iastate.edu).

**Fiber optic networks: fairness, access controls and prototyping**

by

Nathan Allan VanderHorn

A dissertation submitted to the graduate faculty  
in partial fulfillment of the requirements for the degree of  
**DOCTOR OF PHILOSOPHY**

Major: Computer Engineering

Program of Study Committee:  
Arun K. Somani, Major Professor  
Mani Mina  
Ahmed Kamal  
Robert Weber  
Cliff Bergman

Iowa State University

Ames, Iowa

2007

Copyright © Nathan Allan VanderHorn, 2007. All rights reserved.

UMI Number: 3259452

UMI<sup>®</sup>

---

UMI Microform 3259452

Copyright 2007 by ProQuest Information and Learning Company.  
All rights reserved. This microform edition is protected against  
unauthorized copying under Title 17, United States Code.

---

ProQuest Information and Learning Company  
300 North Zeeb Road  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

## DEDICATION

To me...



## TABLE OF CONTENTS

|  |      |
|--|------|
| <b>LIST OF TABLES</b> . . . . .                                | viii |
| <b>LIST OF FIGURES</b> . . . . .                               | ix   |
| <b>ACKNOWLEDGEMENTS</b> . . . . .                              | xii  |
| <b>ABSTRACT</b> . . . . .                                      | xiv  |
| <b>CHAPTER 1. Optical Communication Networks</b> . . . . .     | 1    |
| 1.1 History of Communication Networks . . . . .                | 1    |
| 1.1.1 Optical Telegraph . . . . .                              | 2    |
| 1.1.2 Electrical Telegraph . . . . .                           | 4    |
| 1.1.3 Modern Optical Communication . . . . .                   | 6    |
| 1.2 Modern Communication Networks . . . . .                    | 8    |
| 1.2.1 Circuit-switched Telecommunications Networking . . . . . | 8    |
| 1.2.2 Packet-switched Data Networking . . . . .                | 9    |
| 1.2.3 Merging Voice and Data . . . . .                         | 10   |
| 1.3 Switching and Multiplexing . . . . .                       | 11   |
| 1.3.1 Wavelength Division Multiplexing (WDM) . . . . .         | 12   |
| 1.3.2 Grooming . . . . .                                       | 13   |
| 1.3.3 Optical Switching Techniques . . . . .                   | 14   |
| 1.3.4 Optical Burst Switching (OBS) . . . . .                  | 15   |
| 1.3.5 Multiprotocol Label Switching (MPLS) . . . . .           | 15   |
| 1.3.6 Optical Packet Switching (OPS) . . . . .                 | 17   |
| 1.4 Modern Communication Networks . . . . .                    | 18   |

|   |  |           |
|---|--|-----------|
| 1.4.1   | Access Networks . . . . .  | 18        |
| 1.4.2   | Wide Area Networks (WAN) . . . . .                                   | 21        |
| 1.4.3   | Metropolitan Area Networks (MAN) . . . . .                           | 21        |
| 1.5   | Contributions . . . . .  | 24        |
| <b>CHAPTER 2. Fairness and Access Control Protocols . . . . .</b> |  | <b>26</b> |
| 2.1   | Introduction . . . . .   | 26        |
| 2.2   | Fairness Methods . . . . .   | 27        |
| 2.2.1   | Free Access With Backpressure . . . . .                              | 28        |
| 2.2.2   | Bandwidth Reservation . . . . .                                      | 28        |
| 2.2.3   | Station Polling . . . . .  | 28        |
| 2.3   | Fairness Models . . . . .  | 29        |
| 2.3.1   | Flow Based Max-min Fairness . . . . .                                | 30        |
| 2.3.2   | Node Ingress Traffic Fairness . . . . .                              | 31        |
| 2.3.3   | Utility and Proportional Fairness . . . . .                          | 32        |
| 2.4   | Feedback Based Resilient Packet Rings . . . . .                      | 35        |
| 2.4.1   | Ring Ingress Aggregated with Spatial Reuse (RIAS) Fairness . . . . . | 35        |
| 2.4.2   | RPR Node Architecture . . . . .                                      | 36        |
| 2.4.3   | RPR Fairness Algorithm . . . . .                                     | 36        |
| 2.4.4   | RPR Oscillations . . . . .   | 37        |
| 2.5   | Reservation Mechanisms for Slotted Bus Networks . . . . .            | 39        |
| 2.5.1   | Distributed Queue Dual Bus . . . . .                                 | 39        |
| 2.5.2   | Pi-persistent Protocol . . . . .                                     | 41        |
| 2.6   | Token Polling Protocols . . . . .                                    | 44        |
| 2.6.1   | Fiber Distributed Data Interface (FDDI) Protocol . . . . .           | 44        |
| 2.7   | Combination Protocols . . . . .                                      | 45        |
| 2.7.1   | Cyclic Reservation Multiple Access (CRMA) . . . . .                  | 46        |
| 2.8   | Bandwidth Budget Fairness Model . . . . .                            | 48        |
| 2.9   | Summary . . . . .  | 49        |

|   |           |
|---|-----------|
| <b>CHAPTER 3. Robust, Dynamic and Fair Network . . . . .</b>        | <b>50</b> |
| 3.1 Introduction . . . . .  | 50        |
| 3.2 Metropolitan Area Network Structure . . . . .                   | 51        |
| 3.2.1 RPR Revisited . . . . .                                       | 52        |
| 3.3 RDFN Network Architecture . . . . .                             | 53        |
| 3.3.1 Node Structure . . . . .                                      | 53        |
| 3.3.2 Controller Structure . . . . .                                | 54        |
| 3.4 Data Transport Method . . . . .                                 | 55        |
| 3.4.1 Network Operation . . . . .                                   | 57        |
| 3.4.2 Network Operation Example . . . . .                           | 58        |
| 3.4.3 Multiple Controller Operation . . . . .                       | 59        |
| 3.4.4 Additional Considerations . . . . .                           | 59        |
| 3.5 RDFN Fairness Scheme . . . . .                                  | 60        |
| 3.5.1 Additional Considerations . . . . .                           | 61        |
| 3.6 Performance Evaluation . . . . .                                | 62        |
| 3.6.1 Request Generation . . . . .                                  | 62        |
| 3.6.2 Central Controller Slot Queue . . . . .                       | 63        |
| 3.6.3 Central Controller Slot Issue . . . . .                       | 64        |
| 3.6.4 Node Processing . . . . .                                     | 64        |
| 3.7 Simulation Results . . . . .                                    | 65        |
| 3.7.1 Random Wavelength Selection Scheme . . . . .                  | 65        |
| 3.7.2 Fixed Wavelength Selection Scheme . . . . .                   | 67        |
| 3.7.3 Wait Time per Packet Element . . . . .                        | 68        |
| 3.8 Summary . . . . .   | 69        |
| <b>CHAPTER 4. Light-Trails: Architecture and Fairness . . . . .</b> | <b>72</b> |
| 4.1 Introduction . . . . .  | 74        |
| 4.2 Light-Trail Architecture . . . . .                              | 74        |
| 4.2.1 Light-trail Literature . . . . .                              | 77        |

|  |   |            |
|--|---|------------|
| 4.2.2  | Light-trail Network . . . . .                                       | 78         |
| 4.2.3  | Light-trail Metro Architectures . . . . .                           | 79         |
| 4.3  | Light-Trail Medium Access Controls . . . . .                        | 80         |
| 4.3.1  | Light-trail MAC . . . . .   | 80         |
| 4.3.2  | Light-bus MAC . . . . .   | 81         |
| 4.3.3  | Greedy Pi-persistent Light-trail MAC . . . . .                      | 82         |
| 4.3.4  | Performance Evaluation - Light-Trail and Light-Bus . . . . .        | 83         |
| 4.3.5  | Performance Evaluation - Greedy Pi-persistent Light-trail . . . . . | 85         |
| 4.4  | Token LT and Light-Trail Fair Access (LT-FA) MAC . . . . .          | 87         |
| 4.4.1  | Bandwidth Budget Advertisement . . . . .                            | 87         |
| 4.4.2  | Token LT - Extra Capacity Distribution . . . . .                    | 89         |
| 4.4.3  | LT-FA - Extra Capacity Distribution . . . . .                       | 89         |
| 4.4.4  | Round Interval Calculation for LT-FA . . . . .                      | 90         |
| 4.4.5  | Normal Operation . . . . .  | 92         |
| 4.5  | Token LT and LT-FA Performance Evaluation . . . . .                 | 92         |
| 4.5.1  | Traffic Generation . . . . .  | 92         |
| 4.5.2  | Token LT and LT-FA Operation . . . . .                              | 94         |
| 4.5.3  | Performance Analysis . . . . .                                      | 94         |
| 4.6  | Summary . . . . .   | 104        |
| <b>CHAPTER 5. Rapid Prototyping Platform and Light-Trail TestBed . . . .</b> |   | <b>106</b> |
| 5.1  | Introduction . . . . .  | 106        |
| 5.2  | Reconfigurable Rapid Prototyping Platform . . . . .                 | 107        |
| 5.2.1  | FPGA History . . . . .  | 108        |
| 5.2.2  | Xilinx Virtex II Pro Development Boards . . . . .                   | 109        |
| 5.2.3  | Real-Time Radon Transform Prototype . . . . .                       | 112        |
| 5.2.4  | Griffin Parallel Computing Platform . . . . .                       | 113        |
| 5.3  | Light-Trail Test Bed Design . . . . .                               | 117        |
| 5.3.1  | High Level Testbed Description . . . . .                            | 118        |

|  |  |            |
|--|--|------------|
| 5.3.2                                  | Complete Testbed Functional Description . . . . .  | 119        |
| 5.3.3                                  | FPGA System Design . . . . .                       | 120        |
| 5.3.4                                  | Optical System Components . . . . .                | 126        |
| 5.3.5                                  | Testbed Physical Limitations . . . . .             | 127        |
| 5.4                                    | Basic Testbed Operation and MAC . . . . .          | 129        |
| 5.4.1                                  | Basic Operation . . . . .                          | 129        |
| 5.4.2                                  | Testbed MAC protocol . . . . .                     | 130        |
| 5.5                                    | Light-trail Streaming Media Application . . . . .  | 132        |
| 5.5.1                                  | Client Media Application . . . . .                 | 133        |
| 5.5.2                                  | Sender Client Interface . . . . .                  | 134        |
| 5.5.3                                  | Light-trail Streaming Media Transmission . . . . . | 134        |
| 5.5.4                                  | Receiver Client Interface . . . . .                | 135        |
| 5.5.5                                  | Experimental Results . . . . .                     | 135        |
| 5.6                                    | Summary . . . . .                                  | 136        |
| <b>CHAPTER 6. Conclusion . . . . .</b> |  | <b>137</b> |
| <b>BIBLIOGRAPHY . . . . .</b>          |  | <b>139</b> |

## LIST OF TABLES

|           |   |     |
|-----------|---|-----|
| Table 3.1 | Average wait time with 10% connection oriented traffic . . . . .                        | 66  |
| Table 3.2 | Average wait time with 20% connection oriented traffic . . . . .                        | 66  |
| Table 3.3 | Average wait time with 30% connection oriented traffic . . . . .                        | 67  |
| Table 3.4 | Average wait time for all wavelengths with 20% connection oriented<br>traffic . . . . . | 68  |
| Table 4.1 | Example Utilization Rates . . . . .   | 88  |
| Table 5.1 | Measured receive power . . . . .  | 127 |

## LIST OF FIGURES

|            |  |    |
|------------|--|----|
| Figure 1.1 | Chappe telegraph station near Nalbach, Germany [85] . . . . .  | 2  |
| Figure 1.2 | 1846 Chappe telegraph network in Europe [25] . . . . .   | 3  |
| Figure 1.3 | Map of electrical telegraph stations in the United States, Canadas and<br>Nova Scotia (1853), [10] . . . . .                                     | 5  |
| Figure 1.4 | Map of the Level 3 optical backbone network . . . . .  | 7  |
| Figure 1.5 | MPLS enabled network illustrating a LSP from nodes 1 to 5 . . . . .  | 17 |
| Figure 1.6 | Illustration of the communications network geographical hierarchy in-<br>cluding access, metro edge, metro core and long haul networks . . . . . | 19 |
| Figure 2.1 | Parking lot scenario . . . . .   | 30 |
| Figure 2.2 | Two exit parking lot . . . . .   | 31 |
| Figure 2.3 | Proportional fairness network example . . . . .  | 33 |
| Figure 2.4 | Parallel parking lot . . . . .   | 36 |
| Figure 2.5 | 3 node network portion to illustrate the RPR oscillation scenarios . . . . .   | 37 |
| Figure 3.1 | RDFN ring architecture illustrating regular and controller nodes . . . . .   | 53 |
| Figure 3.2 | Upstream Time Slot . . . . .   | 56 |
| Figure 3.3 | Active upstream time slot (invalid/generic unique ID) . . . . .  | 57 |
| Figure 3.4 | Downstream time slot . . . . .   | 58 |
| Figure 3.5 | Network operation example . . . . .  | 58 |
| Figure 3.6 | Average wait times of all nodes for Wavelength 0 . . . . .   | 67 |
| Figure 3.7 | Average wait times for all wavelengths . . . . .   | 69 |

|             |   |    |
|-------------|---|----|
| Figure 3.8  | Comparison of wait times for the random and fixed wavelength selection schemes . . . . .  | 70 |
| Figure 3.9  | Wait time per packet element (20% voice) . . . . .  | 71 |
| Figure 4.1  | Multiple wavelength light-trail node featuring multiple Light-trail Access Units (LAU) . . . . .  | 75 |
| Figure 4.2  | Four node light-trail - optical connectivity path is displayed in blue . .  | 75 |
| Figure 4.3  | The L-Bone network is an example of how light-trails can be configured as a complete network solution. The node color indicates which light-trail a node is active on. A gray shade indicates that a node is not active on the trail. . . . . | 78 |
| Figure 4.4  | A metro network employing WDM light-trails . . . . .  | 80 |
| Figure 4.5  | Average queuing delay Vs load for a 5 node LT using the light-trail MAC   | 84 |
| Figure 4.6  | Average queuing delay Vs load for a 5 node light-bus using the light-bus MAC . . . . .  | 85 |
| Figure 4.7  | Average queuing delay Vs load for a 5 node light-trail using the greedy Pi-persistent protocol . . . . .  | 86 |
| Figure 4.8  | Maximum queuing delay Vs load for a 5 node light-trail using the greedy Pi-persistent protocol . . . . .  | 87 |
| Figure 4.9  | Fairness Control Message . . . . .  | 88 |
| Figure 4.10 | Average queuing delay Vs load for a 6 node light-trail using the Token LT protocol . . . . .  | 95 |
| Figure 4.11 | Maximum access delay Vs load for a 6 node light-trail using the Token LT protocol . . . . .   | 96 |
| Figure 4.12 | Average queuing delay Vs load for a 6 node light-trail using the LT-FA protocol . . . . .   | 98 |
| Figure 4.13 | Maximum access time Vs load for a 6 node light-trail using the LT-FA protocol . . . . .   | 99 |



|             |   |     |
|-------------|---|-----|
| Figure 4.14 | Average queuing delay Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 19 slot buffer . . . . .  | 100 |
| Figure 4.15 | Average queuing delay Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 9 slot buffer . . . . .   | 102 |
| Figure 4.16 | Average queuing delay Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 4 slot buffer . . . . .   | 103 |
| Figure 4.17 | Maximum access time Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 4 slot buffer . . . . .   | 104 |
| Figure 5.1  | Chart showing the growth and cost reduction of the Xilinx XCV4000 family FPGA from 1991 to 1998 . . . . .   | 108 |
| Figure 5.2  | Memec development board containing the Xilinx Virtex II Pro FPGA  | 110 |
| Figure 5.3  | Digilent development board containing the Xilinx Virtex II Pro FPGA   | 111 |
| Figure 5.4  | Griffin Cluster containing 16 Dell computers connected to 16 Digilent FPGA boards . . . . .   | 116 |
| Figure 5.5  | High level diagram of our 4 four node testbed showing a uni-directional LT from node $N_1$ to $N_4$ . . . . .   | 118 |
| Figure 5.6  | Complete functional block diagram of the 4 node LT testbed illustrating the FPGA components and optical devices used in the design . . . . .                            | 119 |
| Figure 5.7  | Single Light-trail node illustrating the hardware and software components used. . . . .   | 120 |
| Figure 5.8  | Light-trail testbed segment illustrating the optical component organization and coupling ratios . . . . .   | 126 |
| Figure 5.9  | Our four node light-trail along with the eye diagram showing a healthy signal at the end of the fourth node . . . . .   | 130 |
| Figure 5.10 | SHOUTcast streaming media operation. DJ clients connect to the SHOUTcast server where data is buffered awaiting a connection request from SHOUTcast listeners . . . . . | 133 |

## ACKNOWLEDGEMENTS

First and foremost I would like to thank my parents. They instilled in me the value of a strong work ethic and encouraged me to strive for the best. They raised me to understand the importance of ambition and enthusiasm without which I would not have been able to complete this research work. They also provided me with financial and emotional support throughout my tenure at ISU.

Special thanks to my major professor, Dr. Arun Somani, for his guidance over the past 5 years in the work leading up to the successful completion of both my Masters and Ph.D. degrees. Thanks not only for his professional advice but also for his motivation when I was discouraged. His technical expertise and lifelong experience in the academic community has indeed improved the quality of my research. He has imparted me with his wisdom which I will take with me in my future endeavors. Without his leadership this work would not have been possible.

I would like to thank my committee members, Dr. Mani Mina, Dr. Robert Weber, Dr. Ahmed Kamal and Dr. Cliff Bergman for their valuable suggestions to improve the quality of this dissertation. Special thanks to Dr. Mina, whose personal conversations were inspiring and often comical. I really appreciate his encouragement and genuine care for my success.

I would like to thank all my friends and coaches at the cheer gym where I could always go to burn off a little steam and take my mind off research for a while. Thanks to my fellow labmates, Ramon Mercado, Rashmi Bahuguna, Joe Schneider and Mike Frederick whose afternoon political discussions were always a good break from day-to-day activities. Special thanks to Mike, my roommate and fellow researcher, for his understanding of the challenges of Ph.D. research and providing motivation to keep going.

Thanks to my girlfriend, Amy Sutterer, who was always there to lend an ear when the stress of late night dissertation writing was getting the best of me.

Lastly, I would like to thank the administrative support staff from the graduate office, especially Pam Myers, for guiding me through the paperwork and making the graduation process go as smoothly as possible.

## ABSTRACT

Fiber optic technologies enabling high-speed, high-capacity digital information transport have only been around for about 3 decades but in their short life have completely revolutionized global communications. To keep pace with the growing demand for digital communications and entertainment, fiber optic networks and technologies continue to grow and mature. As new applications in telecommunications, computer networking and entertainment emerge, reliability, scalability, and high Quality of Service (QoS) requirements are increasing the complexity of optical transport networks.

This dissertation is devoted to providing a discussion of existing and emerging technologies in modern optical communications networks. To this end, we first outline traditional telecommunication and data networks that enable high speed, long distance information transport. We examine various network architectures including mesh, ring and bus topologies of modern Local, Metropolitan and Wide area networks. We present some of the most successful technologies used in todays communications networks, outline their shortcomings and introduce promising new technologies to meet the demands of future transport networks.

The capacity of a single wavelength optical signal is 10 Gbps today and is likely to increase to over 100 Gbps as demonstrated in laboratory settings. In addition, Wavelength Division Multiplexing (WDM) techniques, able to support over 160 wavelengths on a single optical fiber, have effectively increased the capacity of a single optical fiber to well over 1 Tbps. However, user requirements are often of a sub-wavelength order. This mis-match between individual user requirements and single wavelength offerings necessitates bandwidth sharing mechanisms to efficiently multiplex multiple low rate streams on to high rate wavelength channels, called traffic grooming.

This dissertation examines traffic grooming in the context of circuit, packet, burst and trail switching paradigms. Of primary interest are the Media Access Control (MAC) protocols used to provide QoS and fairness in optical networks. We present a comprehensive discussion of the most recognized fairness models and MACs for ring and bus networks which lay the groundwork for the development of the Robust, Dynamic and Fair Network (RDFN) protocol for ring networks. The RDFN protocol is a novel solution to fairly share ring bandwidth for bursty asynchronous data traffic while providing bandwidth and delay guarantees for synchronous voice traffic.

We explain the light-trail (LT) architecture and technology introduced in [37] as a solution to providing high network resource utilization, seamless scalability and network transparency for metropolitan area networks. The goal of light-trails is to eliminate Optical Electronic Optical (O-E-O) conversion, minimize active switching, maximize wavelength utilization, and offer protocol and bit-rate transparency to address the growing demands placed on WDM networks. Light-trail technology is a physical layer architecture that combines commercially available optical components to allow multiple nodes along a lightpath to participate in time multiplexed communication without the need for burst or packet level switch reconfiguration. We present three medium access control protocols for light-trails that provide collision protection but do not consider fair network access. As an improvement to these light-trail MAC protocols we introduce the Token LT and light-trail Fair Access (LT-FA) MAC protocols and evaluate their performance. We illustrate how fairness is achieved and access delay guarantees are made to satisfy the bandwidth budget fairness model. The goal of light-trails and our access control solution is to combine commercially available components with emerging network technologies to provide a transparent, reliable and highly scalable communication network.

The second area of discussion in this dissertation deals with the rapid prototyping platform. We discuss how the reconfigurable rapid prototyping platform (RRPP) is being utilized to bridge the gap between academic research, education and industry. We provide details of the Real-time Radon transform and the Griffin parallel computing platform implemented using the RRPP. We discuss how the RRPP provides additional visibility to academic research initiatives

and facilitates understanding of system level designs. As a proof of concept, we introduce the light-trail testbed developed at the High Speed Systems Engineering lab. We discuss how a light-trail test bed has been developed using the RRPP to provide additional insight on the real-world limitations of light-trail technology. We provide details on its operation and discuss the steps required to and decisions made to realize test-bed operation. Two applications are presented to illustrate the use of the LT-FA MAC in the test-bed and demonstrate streaming media over light-trails.

As a whole, this dissertation aims to provide a comprehensive discussion of current and future technologies and trends for optical communication networks. In addition, we provide media access control solutions for ring and bus networks to address fair resource sharing and access delay guarantees. The light-trail testbed demonstrates proof of concept and outlines system level design challenges for future optical networks.

## CHAPTER 1. Optical Communication Networks

In this, the first chapter of my dissertation, we trace the evolution of long distance “optical” communication networks from the Roman smoke signal network through to the modern fiber optic transport networks of today. We outline the technological developments in switching, routing and multiplexing that enable today's high speed, high capacity optical networks. Furthermore, we present the common network hierarchy with a discussion of local, metropolitan and wide area network architectures.

### 1.1 History of Communication Networks

The first documented use of “optical” communication dates back to around 150 A.D. with the Roman smoke signal telegraph. As the Roman empire continued to expand throughout the first century, coordinating military forces through the use of hand delivered messages became a daunting and time consuming task. In order to overcome this communication dilemma the Romans developed a highly sophisticated network of towers within visible range of each other where smoke signals were used to relay military messages up to a distance of over 4500 kilometers. The use of such “optical” signals and repeaters marked the introduction of the optical communication network. Although the devices and methods used in optical communication have changed dramatically over the past 20 centuries the primary motivation for high-speed, long distance communication has remained the same. The following sections discuss various engineering achievements over the past 20 centuries that have brought us to the modern optical networks as we know them today.



Figure 1.1 Chappé telegraph station near Nalbach, Germany [85]

### 1.1.1 Optical Telegraph

Fast forward now to 1790. In the wake of the French revolution, and again realizing the importance of high-speed, long distance communication, Claude Chappé designed and built the first “optical telegraph” literally meaning “visible distance writing”. Covering a distance of about 120 miles from Paris to Lillie, France, Chappé’s original fixed “optical” network consisted of a series of 15 towers placed 10 to 20 miles apart. Each tower is equipped with telescopes and large articulating arms sitting atop like those of Figure 1.1. Messages are sent from station to station by configuring the semaphore arms to represent specific code characters. These characters are viewed through a telescope by an operator at another tower and relayed to the final destination. Depending upon weather conditions and operator competency, messages can be sent at a speed of nearly 3000 Km/h. Due to its initial success, the optical telegraph



became a noticeable part of the European landscape for almost half a century. New lines were constructed until 1846 and the network grew to 556 stations covering over 3000 miles connecting cities like Amsterdam, Brussels, Mainz, Milan, Turin and Venice. Figure 1.2 illustrates the vast coverage of the Chappe telegraph network.

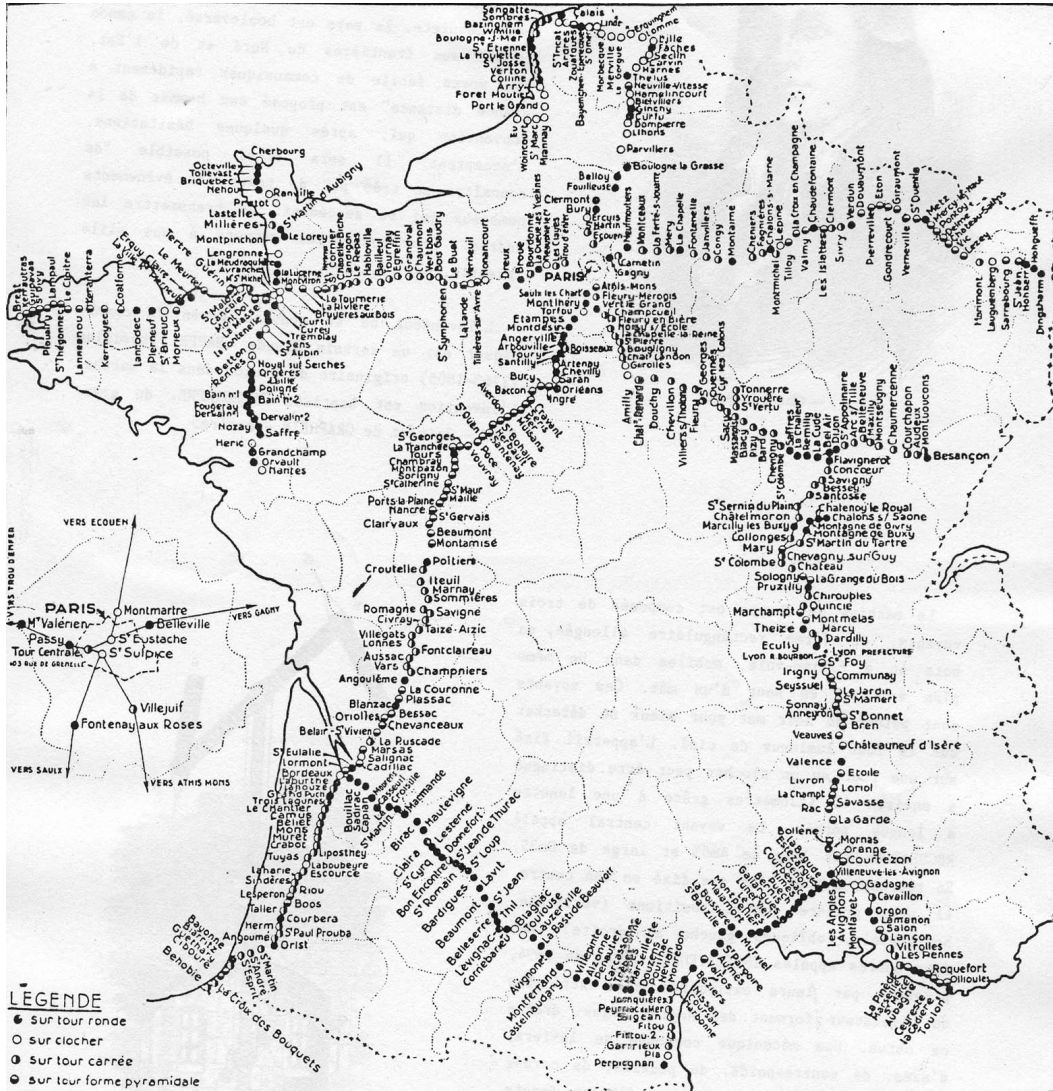


Figure 1.2 1846 Chappe telegraph network in Europe [25]

Tom Standage wrote in “The Victorian Internet” that the Chappe telegraph was “the mother of all networks” and that, at the end, the lines formed “sort of a mechanical Internet”. The optical telegraph was a highly sophisticated communication network which paved the way for future long distance communication networks. Some of the very same principles used in

today's networks such as source coding, error detection, synchronization and flow control were originally developed through the use of the optical telegraph [25].

Although the optical telegraph was in use for over 60 years its success was limited. Due to its complexity, high cost of operation and inability to operate in fog, mist or darkness the optical telegraph ultimately gave way to an electronic version beginning in the early 1830's.

### 1.1.2 Electrical Telegraph

The invention of the electromagnet by William Sturgeon in 1825 paved the way for the first electronic communications. The electronic telegraph operates on a very straightforward principle. The transmitter simply opens and closes an electric circuit which can then be interpreted by a receiver. This principle of "on-off keying" is still in use today in modern optical communications.

Although many inventors of the time pursued electronic communications, Samuel Morse and his partner Alfred Vail popularized its use with the introduction of the Morse code signaling alphabet in 1837. Morse sent his first electronic message in 1838 over two miles of wiring. However, the U.S economic disaster known as the Panic of 1837 delayed the acceptance and deployment of his technology until 1844 when Morse sent a message from Washington D.C to Baltimore MD; a distance of over 40 miles. The telegraph continued to gain acceptance in the U.S and grew rapidly throughout the 19<sup>th</sup> century. Just 8 years after the first line from Baltimore to N.Y was completed, telegraph lines connected most of the Central and Eastern United States as shown in Figure 1.3. Then in 1861 the first transcontinental telegraph line extending from California to New York was completed.

A similar transformation was taking place in Europe and Asia and in 1866 the first robust transatlantic telegraph cable was completed<sup>1</sup>. After its inception, it is said that the worldwide telecommunications system spread more rapidly than the internet of today. For the next 150 years, companies such as Western Union would use this technology to send telegram messages

<sup>1</sup>Earlier transatlantic cables were installed in 1857 but failed after only a few weeks due to the lack of knowledge in transmission line theory.





Figure 1.3 Map of electrical telegraph stations in the United States, Canadas and Nova Scotia (1853), [10]

across the country<sup>2</sup>.

The electronic telegraph was a booming success, however, the cost to send a message over the first telegraph lines was extremely high due to the infrastructure costs and the need for skilled telegraph operators. Thus, usage was typically limited to government officials, railroad operators and emergency workers. Efforts were made to reduce the cost per message by optimizing the message code (similar to video compression of today). Telegraphese was one such language that omitted unimportant words and increased the use of abbreviation. In addition, the telephone began to make its mark in public telecommunications further eliminating the

<sup>2</sup>Western Union publicly announced that they have discontinued telegram service as of January 2006

need for trained telegraph operators. Although, these optimizations significantly reduced the cost of sending electronic messages, a major cost reduction came in 1935 when automatic message routing was introduced. Since 1935, technological advancements in message switching, routing, and multiple access would shape the global communications network as we know it today.

For nearly the past 2 centuries electronic communication has continued to grow in popularity. New technologies have allowed inventors and innovators to significantly enhance electronic communication networks. Modern communication networks have adopted many of the concepts from these early incarnations, however increased demands for high-speed, high capacity and improved reliability paved the way for the fiber optic revolution of today.

### 1.1.3 Modern Optical Communication

As the telecommunications industry continued to flourish into the mid 1900's researchers began looking into ways to increase the speed and capacity of contemporary copper wire transmission medium. Taking heed of the success of the first "optical" networks, researchers began looking for ways to once again carry information via light. The first breakthrough came with the Light Emitting Diode (LED) made from Gallium Arsenide Phosphide in the late 50's which would provide the light source for the first modern fiber optic communications. LED's would eventually be replaced by the Light Amplification by Stimulated Emission of Radiation (LASER) as the light source, however, the LED can still be found today in some optical communication networks.

Although the invention of the LED was promising for communication via light, a fundamental problem still remained; how to "guide" light over long distances without significant signal degradation. It wasn't until 1970 when the Corning Glass Works company introduced a titanium doped silica glass fiber, with an attenuation of only 17dB per kilometer that fiber optic communication began to take off. After continuing research in light emitting devices and low attenuation glass waveguides the first fiber optic transmission was carried out in Long Beach, California in 1977. The data rate at that time was a whopping 6 Mbit/s.

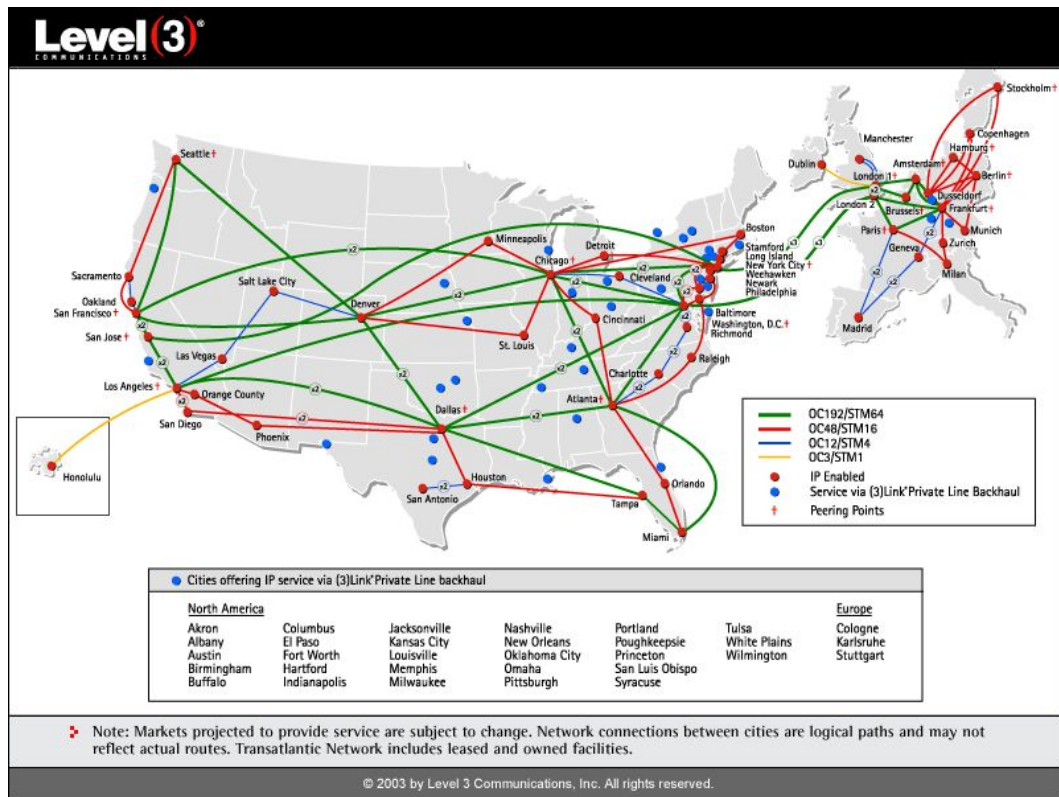


Figure 1.4 Map of the Level 3 optical backbone network

As fiber optic technology continued to blossom and gain wide acceptance as the next generation communication medium, fiber optic lines began popping up all across the world to compliment or replace existing copper wire transmission lines. And in 1988 the first transatlantic fiber, TAT-8, carrying 40,000 telephone circuits, was installed between the U.S., England and France. Figure 1.4 illustrates one of the largest global optical networks owned and operated by Level3 communications. However, this is only a fraction of the mass expanse of optical fiber deployments around the world. In fact, fiber optic cable is continuously being deployed and “lit” every day and may one day reach every home.

Over the past 30 years fiber optic communication has seen many technological advancements that have increased the bit rate to over 1 terabit/s on a single fiber. However, creating a fiber optic network to connect millions of end stations requires a sophisticated architecture and organization. In order to make efficient use of fiber optic capabilities networks must be able to intelligently switch, route and multiplex messages over the shared infrastructure. The re-

mainder of this dissertation discusses modern optical network architectures and protocols and proposes new architectures and protocols to enable the future of fiber optic communication networks.

## 1.2 Modern Communication Networks

A communications network is defined as conglomeration of physical links arranged such that messages may be passed from a location on one part of the network to a destination at another using one or multiple links. Just as the devices and mechanisms utilized for long distance communications have changed dramatically over the centuries from the smoke signals and towers of the Romans to the LASER's and fiber optic waveguides of modern optical communications the network architecture used to carry these messages have also changed. As networks continue to grow in geographical size and complexity new technologies have emerged to support switching, routing and multiplexing of information as discussed in the following sections.

### 1.2.1 Circuit-switched Telecommunications Networking

As telecommunications networks began to grow to connect millions of people across the world, the networks architectures, components and technologies evolved. Switchboard operators were replaced with electronic switching machines, unidirectional lines were upgraded to full duplex and analog signals became digital. This switch from analog to digital signals was the first revolutionary change to effect the telecommunications network. This digital network came to be known as the Public Switched Telephone Network (PSTN). The PSTN connected millions of end stations through copper wires and enabled what is commonly referred to as Plain Old Telephone Service (POTS).

The PSTN was the first incarnation of circuit switched technology where a connection is established through the use of telephone exchanges that are configured to provide dedicated end-to-end circuit for the duration of the communication. This circuit-switched concept provides sufficient quality of service required for typical telephone communication.



In order to group multiple phone circuits together for transport over the long distance links circuit-switched networks rely on Time Division Multiplexing (TDM). Using TDM, time-slots are pre-allocated to the each circuit so that a number of calls can be grouped together for transport across the shared network. As the capacity of the shared communication medium continued to grow, telephone and telegraph providers needed a standardized method for bundling multiple TDM channels together which led to the development of the Plesiochronous Digital Hierarchy (PDH) in the early 1980's.

PDH was the first method used to create a hierarchy for multiplexing higher order frames. PDH effectively increased the network capacity by multiplexing standard TDM frames. The concept introduced with PDH worked for some time, however, due to some of its drawbacks it eventually gave way to the Synchronous Digital Hierarchy (SDH) and Synchronous Optical Network (SONET) in the late 1980's. SONET/SDH continues to be used in today's PSTN's and is expected to continue operation for many years to come due to its large install base and world-wide interoperability. However, as we will see in later sections new features and capabilities are continually being added to SONET/SDH to make it more accommodating to the packet-switching nature of data networks as discussed in the next section.

### 1.2.2 Packet-switched Data Networking

At the same time telecommunications networks were emerging to bring people together through voice communications, computer data networks were emerging as a way to connect computing machines. Although the goal of long distance data communication is similar that of telecommunications the fundamental method in which a connection is established is much different. Instead of the circuit-switching paradigm of telecommunications networks, data networks investigated the use of packet switching. Packet switching attempts to make data transport more efficient by placing less emphasis on dedicated service provisioning and more emphasis on high throughput.

The world's first operational packet switched network was developed by United States Department of Defense (DoD) Advanced Research Projects Agency or ARPA and was thus ap-

appropriately named ARPANET. The initial ARPANET connected 4 machines from UCLA, Stanford, UC Santa Barbara and University of Utah and was operational on Dec 5, 1969. Over the next decade the ARPANET grew rapidly to over 200 hosts including links to Hawaii, Norway and London. ARPANET laid the foundation of the packet switching paradigm of the internet and eventually gave rise to many other data networks such as the military network MILNET and the Defense Data Network (DDN). Then, after the introduction of TCP/IP in the late 70's ARPANET was eventually replaced by the National Science Foundations NSFNet and finally discontinued operation in 1989.

After the introduction of the NSFNet, data communications began to grow at an exponential rate and the NSFNet became the first 45Mbit/s network in 1991. In addition, the growth of the personal computer and the launch of the World Wide Web continued to increase the demand for network communication for commercial and personal use and in 1998 the NSF relinquished its direct role in the Internet to the commercial sector. This privatization of data networking and the deregulation of the national telephone carriers led to huge competition in the private sector. Thus, new technologies began to emerge as a way to provide value-added services and integrate voice and data communications. The next section looks at the challenges of combining voice and data for transport over a shared network.

### 1.2.3 Merging Voice and Data

The past decade has seen a networking paradigm shift from connection oriented communication to high bandwidth IP-centric packet switched data traffic. The explosive growth of the Internet, video conferencing, Storage Area Networks (SAN), Virtual Private Network's (VPN) and other related high bandwidth services combined with the deregulation and privatization of telephone carriers and data networks facilitated a change in the traditional network operation. As public interest in the Internet and other data services grew, telecommunications carriers began offering packet switched services over their existing circuit switched architecture. These new communication networks not only needed to provide high quality of service for voice traffic in terms of latency, jitter and protection but were also required to provide high bandwidth



connectionless services for data traffic. These new requirements coupled with the increased capacity of fiber optics led to new switching and multiplexing solutions to suit both types of traffic.

The following section outlines a few of these emerging switching, routing and grooming techniques that are making their way into today's communication networks. We follow this discussion with existing network architectures that lay the groundwork for the next-generation communication networks.

### 1.3 Switching and Multiplexing

Although fiber optics continue to replace copper wires as the primary transmission medium most switching, routing and multiplexing must still be done in the electronic domain. Switching, routing and multiplexing is relatively simple in the electrical domain because each of the intermediate stations can process the message, read the routing information and make the appropriate switching decision. In addition, SONET Add/Drop Multiplexers (ADM) allow ADM capable nodes to multiplex multiple low rate streams into a higher rate stream for transport over the outgoing port. Thus, to take advantage of the benefits of long distance communication using fiber optics the incoming optical signals must first be converted to the electronic domain and then converted back to the optical domain for transport across the next outgoing link. This optical-electronic-optical (O-E-O) conversion remains to be one of the most challenging aspects of future communication networks.

As single wavelength data rates continue to increase from 10 Gbps today to over 100 Gbps in the near future, electronic processing at every intermediate node adds to the overall complexity and cost of modern communication networks. Advances are being made to make electronic switches more scalable by adding additional ports to the switching fabric, however, as the optical data rate increases electronic switches will be hard-pressed to keep up.

Capacity upgrade in SONET is possible either through deploying new rings or through increasing TDM rates. The former requires new fiber routes while the latter necessitates equipment upgrades on all ring nodes both of which are expensive and time consuming. In

SONET, each transport path has a fixed bandwidth defined over a rigid rate hierarchy. This precludes the possibility of supporting a multitude of client data applications resulting in large bandwidth inefficient mappings. Besides, the burstiness in data traffic cannot be handled well since re-provisioning requires careful capacity planning and takes a long time. The network is not transparent, supports only constant bit rates and provides very little room for service differentiation. So, a need for a transparent, cost-effective architecture that can respond to dynamic traffic needs and allow for service differentiation while still offering support for legacy services is being increasingly realized. Thus, new technologies are being introduced to alleviate this optical-electronic bandwidth mismatch and are discussed below.

### 1.3.1 Wavelength Division Multiplexing (WDM)

WDM is an approach that exploits the huge bandwidth available in the fiber by splitting it up into multiple non-overlapping channels and allowing users to transmit on each channel at the peak electronic rate effectively increasing single-link bandwidth from 10 Gbps to over 1 Tbps. WDM is the sole technology that can support TDM, data, SAN, video traffic etc. independent of bit rates and protocol formats. Dense WDM (DWDM) technologies enable a single fiber to carry over 160 wavelengths. However, the benefits of such high capacity is somewhat overshadowed by the high precision component cost. Coarse WDM (CWDM) on the other hand does not place stringent requirements on optical equipment and can lead to significant performance increase with significant cost savings over DWDM. CWDM will allow operators to expand service offerings, support legacy services and prepare for future traffic growth.

One of the main features of WDM is that traffic can be carried from source to destination entirely in the optical domain using Optical Add Drop Multiplexers (OADM). An OADM is an all optical element that allows wavelengths to be added and dropped while allowing the other wavelengths to pass through without the need for O-E-O conversion. The alternative is to demultiplex all wavelengths, convert them to the electronic domain and use traditional add/drop methods. The use of OADMs can significantly reduce network cost by eliminating

the need for O-E-O components at each node. However, traditional OADMs are statically configured devices in which a determination of which wavelengths are to be dropped needs to be known a priori. This led to the development of reconfigurable OADMs (ROADM) which can be dynamically reconfigured to allow any wavelength to be added and dropped on the fly. ROADMs are rapidly becoming part of today's metropolitan and wide area networks [72].

One drawback of ROADMs is that they do not support wavelength routing as required in mesh networks. New devices such as wavelength selective cross-connects (WXC) and micro-electro-mechanical systems (MEMS) are being introduced to provide such support but are still immature and costly. The goal of designing transparent optical networks in which optical signals on an arriving wavelength can be switched to an output link without electronic conversion is slowly becoming a reality with technologies such as WDM, OADMs, and OXCs. However, in order to more fully utilize a single wavelength capacity traffic grooming is still required. The following section outlines grooming and switching technologies that are making an impact on future optical networks.

### 1.3.2 Grooming

One of the most pressing issues concerning current WDM implementations is that, once a lightpath is established, the entire wavelength is used exclusively by the connections source and destination node pair; no sub-wavelength sharing between nodes along the lightpath is allowed. And since sub rate SONET streams can be as little as Optical Carrier 1 (OC-1), the entire wavelength capacity may be underutilized unless the source and destination nodes efficiently aggregate traffic.

Grooming creates a unit of capacity smaller than an entire wavelength in a WDM network [78]. Instead of lower rate traffic monopolizing the use of an entire wavelength, multiple lower rate traffic streams can be multiplexed on the same wavelength, and the capacity more effectively utilized. As mentioned, sub-rate optical connections can vary from OC-1 to full wavelength capacity and, although WDM networks can utilize multiple wavelengths for each connection, a cost is associated with adding additional wavelengths due to the need for

additional transmitters and receivers. Thus, it is cost effective to combine sub-wavelength connections whenever possible to more efficiently utilize each wavelength in the form of traffic grooming.

WDM grooming networks can be classified into two categories [90]: dedicated-wavelength grooming (DWG) networks and shared-wavelength grooming (SWG) networks. In a DWG network, the source-destination node pairs are connected by lightpaths shared by connections between the pair. In a SWG network, the lightpath can be shared by connections from different s-d pairs. The performance of SWG networks depends on the efficient merging of fractional wavelength requirements into full or almost-full wavelength requirements. Grooming nodes are classified into various categories depending on the level of grooming capability they provide.

- ADM-constrained grooming node - The node can multiplex and demultiplex low-rate traffic only on dropped wavelengths using an add-drop multiplexer.
- Wavelength continuity constrained grooming node - The node can switch connections across different lightpaths, but cannot switch between different wavelengths.
- Full grooming node - The node can switch connections in any permutation from one wavelength to another.

The previous sections have discussed methods for packing multiple streams into a single channel for transport across the network. We also discussed electronic switching techniques using O-E-O conversion. The next section looks at some emerging technologies that attempt to provide switching and routing capabilities entirely in the optical domain.

### 1.3.3 Optical Switching Techniques

All-optical networks are able to transport data from source to destination entirely in the optical domain. This is a departure from current optical networks that rely on O-E-O conversion at each intermediate node to route data properly. The opacity inherent in traditional networks is costly in terms of limiting bandwidth and increasing switching complexity.

Multiprotocol Label Switching (MPLS), Optical Burst Switching (OBS) and Optical Packet Switching (OPS) have been proposed as solutions to realizing an all-optical network. MPLS and OBS have the advantages of creating all-optical end-to-end circuits but do not allow intermediate nodes access to the wavelength path. Additionally, optical switches are constantly being reconfigured to accommodate new connections. OPS, on the other hand can make switching decisions in the optical domain, but the technology is still immature. We discuss each of these technologies below.

#### 1.3.4 Optical Burst Switching (OBS)

OBS was one of the first methods suggested to create an all-optical path for information to travel from source to destination entirely in the optical domain [71, 35, 83]. In OBS, a control packet is sent ahead of the data to configure the switches along the path to establish the end-to-end connection. Data follows this packet, delayed by a guard time, never leaving in the optical domain. OBS is an attempt to offer burst level communication at the optical layer by pre-allocating resources for the duration of an IP traffic burst. However, in order to provide reasonable network utilization, current OBS solutions require very high-speed switches to lower the ratio between burst length and path set up time. Additionally, because resources must be reserved along the entire burst path, these resources are unusable to other connections which can cause the network to go underutilized.

Although much work on OBS has been presented in the literature the advantages it offers over traditional O-E-O circuit or packet switching paradigms discussed earlier are limited. Thus, it seems that OBS technology will probably never emerge as an effective transport protocol.

#### 1.3.5 Multiprotocol Label Switching (MPLS)

Although not a new technology, Multiprotocol Label Switching (MPLS) has been gaining attention as a way to offer QoS in next generation IP-based optical networks [84]. MPLS was originally designed to provide a way of grouping individual traffic flows together for transport

over a core MPLS-enabled network. MPLS is similar to OBS in that an all-optical path is established between two edge nodes for a given flow. The difference between OBS and MPLS lies in how the path is established.

The basic idea in MPLS is to add a routing label that MPLS enabled routers can easily understand and route, thus creating a label switched path (LSP) from source to destination in which intermediate routing decisions can be made more quickly based upon the label. The all-optical Wavelength Selective Cross-connect (WXC) is the first attempt to make all optical switching decisions where the label is actually the incoming wavelength (MPLambdaS). As an optical signal enters a switch, the wavelength is used to determine the correct output port. Two types of WXC are available in an all optical transport network [3].

- Lambda Switch Capable (LSC) - Forwarding occurs by switching a lightpath on the basis of its wavelength. The LSC WXC interface is not capable of receiving control messages in-band with the data, thus, data forwarding decisions are made based upon the incoming wavelength.
- Fiber Switch Capable (FSC) - Does not recognize bits or have any concept of wavelength. Out-of-band control information is used to configure each switch en route to the destination.

Figure 1.5 is an example of a network containing a Label Switched Path (LSP). The OADMs and WXC are configured such that an LSP is established from node 1 to 5. The data carried on this path travels from source to destination entirely in the optical domain. The LSP may exist for hours or days depending upon the connection requirements and the available network resources. Thus, the main difference from OBS is that an OBS path is only established for the duration of the information burst.

Although MPLS and OBS offer network designers the ability to utilize multiple wavelength routes without expensive OEO conversion, networks are still unable to provide sub-rate multiplexing or traffic grooming along the wavelength path. Optical packet switching, discussed in the next section, overcomes this shortcoming by attempting to provide dynamic switching

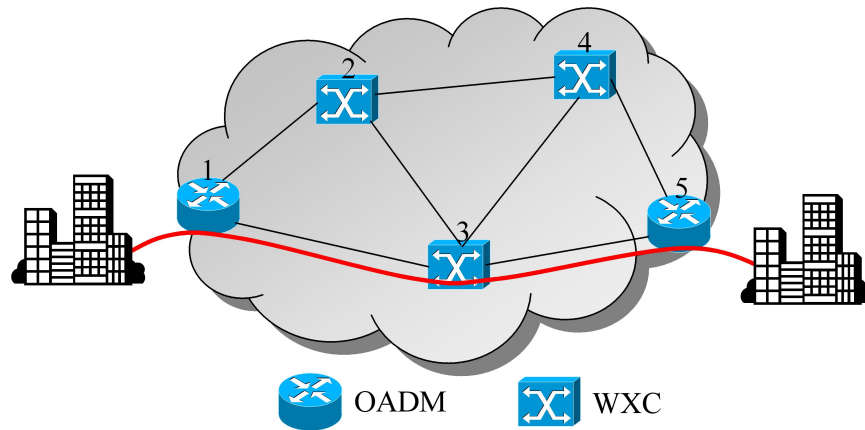


Figure 1.5 MPLS enabled network illustrating a LSP from nodes 1 to 5

on a packet by packet basis entirely in the optical domain.

### 1.3.6 Optical Packet Switching (OPS)

An OPS is the true optical equivalent of an electronic packet switch where a switching decision is made based upon the packet header information [67, 73]. The major advantage of OPS is its flexible and efficient bandwidth usage which enables the support of a variety of services. Pure OPS technology in which packet header recognition and control are performed in all-optical domain is still many years away, and may never become reality.

An OPS, with electronic header processing and control, is a more realistic near term solution. A practical OPS experiment has been performed under the European ACTS (Advanced Communications Technology and Services) KEOPS (KEYs to Optical Packet Switching) project [36, 89]. In KEOPS, the header is sent with the payload but at lower bit rate, and the header processing is still in the electrical domain. This potentially requires massive optical buffering at the input port to allow the header processing circuits to finish the job. Similarly, an optical buffer is required at the output side to avoid packet loss. At present, the buffering technology is not mature and has to overcome a number of technological constraints, including the large and varying size of optical buffering and high speed header processing.

The ultimate goal of switching, grooming and multiplexing technologies in optical networks is enable dynamic switching in the optical domain, allow traffic grooming at intermediate nodes

and utilize multiple wavelengths. The result will be a highly utilized bit-rate and protocol transparent network. Regardless of the components and technologies used in future communication networks an important feature of these networks is the physical architecture required to interconnect billions of end systems. Thus, the next section briefly outlines the current state of modern communication network architectures.

## 1.4 Modern Communication Networks

Up to this point we have discussed how signals are transported through the networks using various switching, routing and multiplexing techniques. However without an appropriate architecture on which to use these techniques a communications network is not complete. In this section we talk about current communications network architectures.

A common approach to classify networks is in relation to their geographic scope. Often times communication networks are roughly organized into three hierarchical categories: access, metro and long haul. Figure 1.6 is a graphical representation of this three tiered hierarchy and will be used in the following discussion.

On one side of the geographical spectrum is the access network. The access networks support a broad range of protocols and technologies to connect the subscriber (or end system) to the network. On the other end of the hierarchy is the long haul or core network which provides large tributary connectivity between regional and metropolitan area domains. Interfacing the access networks with the long haul network is the metro. Central Offices (COs) within the metro offer high speed media and application services to interconnect the access networks with the core. The following sections provide additional details for each network category respectively.

### 1.4.1 Access Networks

As the smallest of the geographical categories, access networks typically only span a distance of less than 1-2 kilometers and as such are often times referred to as Local Area Networks (LAN). The primary role of access networks or LANs is to connect end users or subscribers



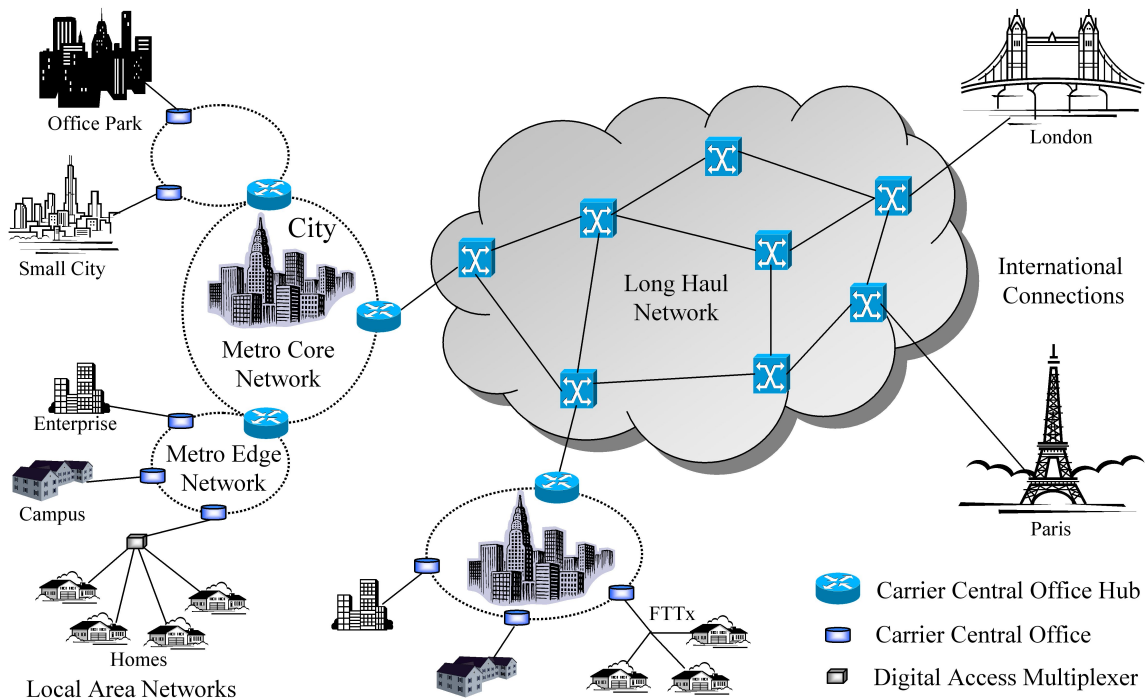


Figure 1.6 Illustration of the communications network geographical hierarchy including access, metro edge, metro core and long haul networks

to the edge switching elements of the larger transport network. Campus wide and enterprise networks, in which local communication may comprise a majority of the traffic, are often considered in the context of local area networks as well.

Over the past 30 years numerous transport protocols have been developed and utilized for use in the LAN arena, however, most installations have settled on Ethernet as the transport protocol of choice. Although a comprehensive discussion of all the access network technologies and protocols is out of the scope of this dissertation we provide a brief overview of some of the most common features and protocols of access networks.

For telecommunications operators the last mile or subscriber line connection is typically the most expensive to maintain. Thus, to reduce the amount of cable required to reach the subscriber premises carriers typically deploy remote multiplexing equipment such as digital subscriber line access multiplexers (DSLAM). These devices aggregate traffic from multiple end stations before being transmitted over higher speed lines to the local switching center or

central office where the traffic can be switched/routed to the PSTN or across the Internet. Such remote multiplexing units allow carriers to reduce the amount of cable required to reach each individual end station.

Typically the access network is the network bottleneck. Thus, efforts are underway by both telecommunications and cable TV service providers to push fiber optics closer to the user premises. Telecom providers are attempting to squeeze the maximum bandwidth out of the existing twisted pair copper wire by moving the remote multiplexing equipment, often connected to the CO with fiber, closer to the subscriber. The goal of this new infrastructure is to provide television, broadband internet and traditional telephony services to all subscribers; often referred to as the service “triple play”. Although the Digital Subscriber Line (DSL) has been very successful for telecom carriers the bandwidth limitation of legacy copper twisted pair cable is becoming a significant hurdle when trying to provide high definition video to subscribers. To remedy this problem telecommunication carriers are beginning to deploy fiber all the way to the home, thus, eliminating the access network bottleneck.

On the other hand, cable companies have spent billions of dollars to upgrade the existing coaxial infrastructure to provide the same “triple play” of services through their existing coaxial cable network. The cable network has been enhanced to provide upstream communication from the subscriber premises as well as the downstream broadcast capabilities for which it was designed. The major hurdle for cable providers is that the current cable access network architecture is a shared medium that may connect over 2000 subscribers to the same CO segment. This inherent drawback of cable networks limit the scalability of the cable infrastructure.

Cable and telecommunication companies will continue to compete for subscriber acceptance and it seems that cable is currently winning the race. However, with the billions of dollars being spent by the large telecommunication service providers this gap in beginning to narrow. With the widespread deployment of fiber to the home of near the home (I.e., Fiber to the “x” where x refers to home, curb, building, or premises) competition will continue to drive access network growth and technological advancements.

On the other end of the geographical spectrum is the long haul or core network. The follow-

ing section provides a brief discussion of their role in the modern communications architecture.

#### 1.4.2 Wide Area Networks (WAN)

Networks that cover a large geographical area are typically referred to as WAN's, Long Haul or backbone networks. The Level3 map of Figure 1.4 is one example of a modern WAN. As can be seen from Figure 1.6 the long haul network typically has a highly meshed architecture and provides the long distance interconnections between local and metropolitan area Central Offices, small communities and international gateways.

Connections across WAN's can be implemented through circuit or packet switching but are often times provisioned as a leased line service identified as T-carriers. Various protocols are used to carry traffic across the WAN such as Frame relay, Asynchronous Transfer Mode, SONET etc.

Of the three geographical categories long haul networks typically support the highest capacity which has historically provided the greatest source of revenue for the incumbent telecommunications carriers. As such, the newest optical technologies and components such as WDM, ROADM's and WXC's are typically first deployed in the long haul networks as the initial high cost can be mitigated. As the cost of such technology decreases and the bandwidth requirements of MANs increases these technologies typically find their way into the MAN market.

As WAN technologies continue to evolve to support ever increasing data traffic requirements their influence on metro networks will be profound. The remainder of this dissertation focuses on the Metropolitan Area Networks. Thus, we devote the next sections to discussing current technologies and challenges of MANs in greater detail.

#### 1.4.3 Metropolitan Area Networks (MAN)

Currently, most metro networks are based upon SONET/SDH ring architectures and are organized into a two-level hierarchy: metro edge and metro core. The metro edge refers to the space between subscriber access and central office location. Metro edge rings typically span about 10 to 40 kilometers, operate at OC-3/STM-1 or OC-12/STM-4 rates and employ

SONET add drop multiplexers that connect to digital loop carrier setups, enterprise networks, telephone public branch exchanges etc. Most edge traffic is usually outbound from the local ring and hence exhibit strongly hubbed traffic patterns with the central office as the hub [19]. This makes edge networks well suited for Unidirectional Path Switched Ring (UPSR) architectures.

The metro core refers to the rings that interconnect major central office hub locations and feed into long haul networks. Metro core rings typically span about 40 to 80 kilometers, operate at OC-48/STM-16 or OC-192/STM-64 rates and perform a higher level of aggregation than the corresponding edge rings. The traffic demands in metro core are much more meshed and improved bandwidth efficiency is obtained through Bidirectional Line Switched Ring (BLSR) architectures. Digital cross-connects are used to interconnect rings and to provide fine granular bandwidth management. The traditional ring architectures performed well when the dominant traffic was voice. However, there have been some emerging trends in design and deployment of optical networks that bring to the forefront the inherent deficiencies in existing architectures. The following two sections outline emerging solutions for metro area networks.

#### **1.4.3.1 Metro Core Solutions**

The requirements for metro core are different from that of the metro edge. In the metro core, the emphasis is on scalable bandwidth provisioning. With maturing optical technologies, ring- or mesh-based wavelength routed DWDM networks is an ideal fit here since it offers rapid provisioning, service transparency and low network costs (since they are amortized over a large user base). However, in the metro edge, the focus is on protocol heterogeneity, heavily sub-wavelength traffic and a price-sensitive limited user base. Hence the metro edge is seeing more diverse possibilities, ranging from improved SONET/SDH and Ethernet offerings, we discuss each of them in turn below.

### 1.4.3.2 Metro Edge Solutions

Recently, Next Generation SONET (NGS) has been introduced to improve SONET's capabilities for carrying packetized data traffic while still retaining its original protection and performance monitoring features [19]. This includes the Generic Framing procedure (GFP), Link Capacity Adjustment Scheme (LCAS) and Virtual Concatenation (VC) mechanisms. VC allows for concatenation of several payloads to provide flexible bandwidth provisioning and to minimize mismatch in data and port rates. GFP provides a simple framing technique to multiplex multiple client protocols and LCAS specifies a control mechanism to dynamically adjust the number of tributaries assigned to a connection [48]. Collectively, these features are the building blocks of the new data-aware NGS transport networks.

Despite the above enhancements, NGS is still an approach that attempts to bridge the packet and circuit switching paradigms of voice and data communications, both of which differ fundamentally in their philosophies. NGS systems process the signals electronically in all intermediate nodes thereby precluding transparency, reducing scalability and leading to increased equipment costs. Besides, NGS also has some framing requirements like STRATUM timing and pointer processing which can become expensive at high data rates like 40 Gbps. Thus, other approaches are being considered to enhance Metro Edge networks such as Next Generation Ethernet (NGE) as discussed below.

The features that are exclusive to SONET is its efficient support for survivability and performance monitoring. Ethernet services, on the other hand, are easily upgradeable and have the advantages of familiarity, simplicity and low cost. While Ethernet does not offer TDM-level guarantees for bandwidth and delay, SONET does not offer efficient data mappings. NGE is a ring based cost-effective and fault tolerant data transport solution that combines statistical multiplexing along with a fairness based access scheme called Resilient Packet Rings (RPR) which will be discussed in more detail in Chapter 2 [24].

There are some problems associated with the packet scheduling and rate adaptation approach followed by RPR. The scheduling stream gives priority to transit traffic over local traffic and hence delay seen by a node is dependent on upstream traffic patterns. In addition, if the

bandwidth requirement of a newly arriving traffic flow is the lowest among the contending flows all the upstream nodes are required to throttle their rate to this lowest rate, creating large oscillations in bandwidth allocation. Such a reactive approach in the presence of bursty traffic may result in large settling times to a fair rate. In general, packet rings have been designed based on enterprise requirements and consequently there is less support for TDM traffic. Since RPR terminates traffic on every node like NGS, their capacity scalability and cost-effectiveness is also questionable.

## 1.5 Contributions

The goals of this dissertation are two fold: addressing fairness in media access controls for metropolitan area networks and validating research claims through system development using the reconfigurable rapid prototyping platform. To this end, Chapter 2 begins our discussion of fairness with a presentation of three common methods used in bandwidth arbitration protocols, namely, backpressure, reservation and polling. We then delve further into fairness with an overview of some of the most recognized fairness models, such as max-min, node ingress and proportional fairness. Following that we examine fairness in existing media access control protocols for ring and bus networks. Chapter 2 is then concluded with an introduction of our bandwidth budget fairness model which will later be used as a basis for the light-trail fair access protocol for light-trail networks presented in Chapter 4.

Chapter 3 continues our discussion of fairness in ring networks with a presentation of the Robust, Dynamic and Fair Network (RDFN) protocol. The RDFN protocol is developed as a novel method to efficiently share ring network resources between traditional circuit switched voice traffic with highly variable and bursty connectionless data traffic. The protocol facilitates fairness through a centralized controller and a reservation based access control scheme. The reservation mechanism ensures that all stations competing for bandwidth have equal opportunity for service and no single node is allowed to monopolize the channel. In addition, isochronous traffic is supported for voice communications with bandwidth and delay guarantees. As we will see, the RDFN protocol does not suffer from the oscillations experienced with

the Resilient Packet Ring protocol nor is it susceptible to large network delays as in the Cyclic Reservation Multiple Access protocol.

In chapter 4 we give an overview of light-trail technology as introduced in [37] and discuss various access control protocols for light-trails. Light trail technology is a solution to the problem of all-optical network switching. The goal is to establish a connection between source and destination nodes as a unidirectional bus. This unidirectional bus also allows intermediate nodes to transmit data to any other node downstream. Light-trail technology avoids costly O-E-O switching at intermediate nodes and offers complete transparency to signal bit-rate, format, and protocol. We present three light-trail access control mechanisms for bandwidth arbitration in light-trails. We then introduce the Token LT and Light-Trail Fair Access (LT-FA) protocol that satisfies the bandwidth budget model presented in Chapter 2. The LT-FA protocol provides access delay bounds and distributes trail capacity to all competing stations in a fair manner. We also show that the LT-FA protocol is the best solution for light-trail media access control through performance comparisons with that of other unidirectional bus protocols suitable for adaptation to the light-trail architecture.

Chapter 5 introduces the reconfigurable rapid prototyping platform (RRPP). We provide details of the Real-time Radon transform and the Griffin parallel computing platform implemented using the RRPP. We discuss how the RRPP provides additional visibility to academic research initiatives and facilitates understanding of system level designs. As a proof of concept, we introduce the light-trail testbed developed at the High Speed Systems Engineering lab. We provide details on its operation and discuss two applications developed to enhance our understanding of light-trail technology. We show how the LT-FA MAC has been implemented on the testbed and demonstrate a streaming media application.

The final Chapter of provides concluding remarks on our work with fairness, media access controls and prototyping for metropolitan area networks.

## CHAPTER 2. Fairness and Access Control Protocols

As single wavelength data rates continue to increase to over 40 Gbps it is becoming less likely that single end to end connections will consume an entire wavelength capacity. Thus, it is of ever increasing importance for networks to effectively multiplex multiple streams over the shared medium architectures. Media Access Control (MAC) protocols play an important role in metro networks by providing collision protection and bandwidth arbitration to meet various quality of service requirements.

MAC protocols are designed and implemented to avoid collisions, provide acceptable delay and jitter characteristics, reduce blocking and most importantly provide fair access to available resources to competing flows. The following sections present an overview of various fairness definitions and existing access control techniques for ring, dual bus and unidirectional bus networks. We then propose a new fairness definition for metro networks based upon a “bandwidth budget” that suggests a method to provide acceptable use to all competing nodes while attempting to maximize network utilization.

### 2.1 Introduction

The explosive growth of the internet and IP-centric data communications is changing the face of networking. Traditional synchronous voice traffic with dedicated bandwidth requirements is being replaced with highly bursty asynchronous data traffic. This paradigm shift is affecting the way in which traffic is provisioned and how network resources are consumed. Typically, in traditional telecommunications networks, resources are provisioned using dedicated service channels. However, the overwhelming prevalence of bursty IP traffic causes dedicated channels to often times go underutilized or at other times be inefficient for handling bursty



data traffic. Thus, customers and network carriers are interested in more effective methods to arbitrate shared network capacity.

As mentioned in Chapter 1, WDM technology has increased single fiber capacity to over 1Tbps. However, WDM components such as tunable transceivers, wavelength converters, optical cross connects and high precision lasers are still relatively immature and too expensive for extensive deployment in metro networks. Until such devices become more available and affordable, network resource sharing will rely on optical or electronic wavelength grooming techniques.

The difficulty in designing bandwidth arbitration schemes that attempt to maximize network utilization while providing fairness and access delay guarantees is that these Quality of Service (QoS) metrics share inverse relationships. That is, a positive change in one metric will most likely have a negative effect on the others.

To understand these relationships, this chapter is devoted to providing a discussion of fairness, access delay and network utilization and their tradeoffs. We outline some of the most recognized MACs for ring, dual bus and unidirectional bus networks. We then provide a definition of fairness based on a “bandwidth budget” that attempts to maximize network utilization through residual capacity resource sharing while maintaining acceptable use for all stations. This “bandwidth budget” model will serve as the fairness criteria for the Light-trail media access control discussed in Chapter 4.

## 2.2 Fairness Methods

Media access control protocols provide two primary functions in shared medium networks: preventing/detecting collisions and regulating network access. Fairness in MAC protocols has been defined and implemented many different ways for various network topologies and QoS requirements. In general though, the three most common methods used to regulate fair medium access are backpressure, reservation and polling as described in the following sections.

### 2.2.1 Free Access With Backpressure

Protocols that provide free access to the shared network medium, such as the Resilient Packet Ring (RPR) protocol, rely on timely feedback of network status in the upstream direction to control downstream congestion. Because stations at the head of the ring or bus network can monopolize network resources, these protocols use backpressure messages to inform upstream stations of impending downstream congestion. Upon receiving these feedback messages, stations contributing to the congestion are forced to throttle their ingress traffic to some fair rate which will in turn alleviate congestion and allow for fair treatment of downstream nodes. However, backpressure access control protocols are only useful in networks that are bi-directional in nature and are capable of providing timely feedback in the opposite direction. As we will see in the the later sections this timely feedback has a direct effect on the performance of feedback based MAC protocols.

### 2.2.2 Bandwidth Reservation

Reservation based protocols facilitate resource sharing by requiring stations to make access reservations before transmission. Bandwidth is then provisioned based upon some knowledge of the outstanding network requests. Reservations can be queued in a distributed manner as in the Distributed Queue Dual Bus (DQDB) protocol or in a centralized control structure as in the Gigabit Passive Optical Network (GPON) protocol. Just as in backpressure protocols, reservation methods often require networks to support bi-directional communications.

### 2.2.3 Station Polling

Finally, protocols such as Token Bus and Fiber Distributed Data Interface (FDDI) use a polling mechanism to arbitrate network bandwidth. In these protocols a token is passed from station to station granting network access when needed. When a station acquires the token it is permitted to transmit for some fixed duration before passing the token to the next station. Token based protocols effectively share network resources by successively passing the token to each station in a round robin fashion. Polling protocols are the only category that does not

necessarily require full duplex operation for sufficient performance.

Each of these MAC approaches has its benefits and drawbacks. We will discuss each of these approaches in the context of ring and bus networks in the upcoming sections. We also take a look at access control protocols that use a combination of the three techniques such as the Cyclic Reservation Multiple Access (CRMA) protocol. However, to more fully understand the goals of fairness in MAC protocols we first define a few fairness metrics using some of the most common fairness models.

### 2.3 Fairness Models

Defining fairness is a relatively difficult task as there are many criteria that effect network quality of service. Depending upon the desired QoS requirements, fairness mechanisms have been proposed in the literature to provide equal resource utilization, equal performance metrics such as delay and throughput, and equal blocking probability [28]. Quantitative measures of fairness that represent the tradeoffs between competing performance measures is introduced in [54] with the concept of *power* which can be used to identify the operating point where a system delivers its best performance. In addition, game theoretic approaches to fairness have been presented in [60] using the Nash arbitration scheme where the product of individual station performance objectives is defined as the criterion for optimization. Nash equilibrium approaches are also presented in [43] to define fairness and optimal participation strategies in peer-to-peer networks with economic incentives.

It should be noted that fairness can be optimized in many contexts depending upon the requirements of the network architecture and protocol. Thus, fairness optimizations schemes must be developed and analyzed to suit the specific network and application QoS requirements. Due to the inverse relationships shared by network performance metrics, designing an all-inclusive fairness protocol is virtually impossible. Attempts have been made to enable both full network utilization and fairness as in [23] and [79]. However, physical limitations such as instantaneous feedback prohibits real-world operation. In addition, such schemes may introduce increased network overhead and call blocking probability. Hence, a complete understanding of

the system requirements and fairness models must be defined to guarantee satisfactory fairness performance.

As an exhaustive discussion of all the proposed fairness strategies is out of the scope of this dissertation, the following sections present a few of the most common fairness models as a framework for use in the various metropolitan area network media access control protocols discussed later. Thus, we begin our discussion with one of the most recognized fairness models: flow based max-min fairness.

### 2.3.1 Flow Based Max-min Fairness

The goal of flow based max-min fairness is to maximize the network use to the flows with the minimum allocation. That is, we attempt to maximize the allocation of flow  $i$  with the constraint that an increase in flow  $i$ 's allocation does not cause a decrease in some other flows allocation having the same or smaller rate than  $i$ . Under flow based max-min fairness, each flow is entitled to the same network usage as any other competing flow [14].

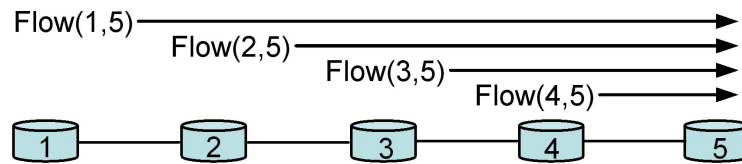


Figure 2.1 Parking lot scenario

To illustrate max-min fairness, consider the parking lot scenario depicted in Figure 2.1 in which 4 infinite demand flows share the most congested link from 4 to 5. In this example, flow based max-min fairness suggests that each flow will receive an equal share of the bandwidth over the most congested link. Thus, all flows will receive  $1/4$  of the bandwidth over link 4-5.

If it were possible to decompose traffic from all sources into bits, the most elemental building block, the easiest way to provide max-min fairness is to use a round robin service scheme similar to Generalized Processor Sharing (GPS). Under the GPS method, a server will sequentially visit each station servicing a single bit from each flow having a backlogged queue;

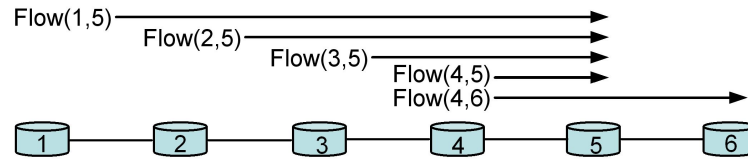


Figure 2.2 Two exit parking lot

in this manner each backlogged queue is guaranteed an equal service rate. In addition, the excess service is distributed equally among all backlogged flows.

One of the major advantages of GPS is that it is work conserving. That is, a work conserving protocol guarantees that the server is always busy if there are any backlogged queues in the system. A work conserving protocol ensures that no penalty on the total network delay is introduced by the access control scheme. Although such a GPS scheme can provide max-min fairness it can only be used as an optimal benchmark due to the following reasons.

- Traffic cannot typically be decomposed into bits
- Propagation time between stations must be taken into consideration
- Not all topologies are well supported for round robin service especially unidirectional bus architectures

Although flow based max-min fairness is one solution to regulate traffic traversing a congested link, in some situations it may still be unfair. The following sections discuss alternatives to flow based max-min fairness.

### 2.3.2 Node Ingress Traffic Fairness

Figure 2.2 shows another network flow scenario called the “two exit parking lot”. In this example, flow based max-min fairness suggests that all flows should receive an equal  $1/5$  share of the bandwidth over link 4-5. As can be seen, flow based max-min fairness rewards node 4 for spreading its traffic among multiple destinations. In addition, nodes 1 through 3 will be further penalized if more connections originating at node 4, such as (4,7) and (4,8) exist.

In this situation, it may be more desirable for fairness to be based upon node ingress traffic. Node ingress traffic fairness suggests that each station must split the available bandwidth over the most heavily congested link allowing each station to source the same amount of traffic. Assuming that each node has an equal weight, the node ingress fair rate is then 1/4 of the capacity of link 4-5. It is then the task of each individual node to determine how to distribute the available bandwidth among its competing flows [92, 33].

The per node granularity of node ingress fairness model is desirable in that MAC need not be concerned with the priority of each individual flow; rather the MAC will attempt to distribute the available capacity among all competing stations based upon some service level agreement. In addition, MACs that follow node ingress traffic fairness can accommodate weighted distributions relatively easily leading to proportional fairness as discussed in the next section.

### 2.3.3 Utility and Proportional Fairness

The former fairness models have illustrated situations in which fairness is based strictly upon actual network usage. That is, fairness is governed by the actual network resources given to each station over a bottleneck link. Such models are effective in demonstrating fairness in situations where external factors do not effect the perception of fairness. However, in many situations economical or environmental factors may play a role in determining the fair allocation of resources. To accommodate such factors a more general fairness concept called utility based fairness models are presented.

The utility based approach is defined as follows: every source  $i$  has a utility function  $U_i$  where  $U_i(x_i)$  defines the value (or utility) to source (or flow)  $i$  of having rate  $x_i$ . The goal of utility fairness is to then maximize the overall utility of the network under the constraint that the sum of the allocations,  $\sum x_i$ , does not exceed the individual link capacities [12, 58].

Assuming utilities are additive, e.g. the aggregate of all utilities is the sum of each sources' utility and a stations utility is weighted by a factor of  $w_i$  between 0 and 1, then, a utility fair allocation of rates is an allocation which satisfies the following equations.

$$\begin{aligned} & \text{Maximize } \sum_i w_i U_i(x_i) \\ & \text{Subject to } \sum_{i \in l} x_{i,l} \leq C_l \end{aligned} \quad (2.1)$$

where  $x_{i,l}$  is the allocation  $x_i$  over link  $l$  and  $C_l$  is the capacity of link  $l$ .

One of the most recognized utility fair models is that of proportional fairness presented in [53]. Under the assumption that internet traffic is elastic, e.g. traffic can be transferred at any rate up to the capacity limit imposed by the link or network, Kelly suggests a reasonable utility function for  $U()$  is  $\ln$ . Using this utility function and Lagrangian multiplier techniques Kelly found that the solution to the above equations must satisfy the following criteria

$$\sum_i w_i \frac{y_i - x_i}{x_i} \leq 0 \quad (2.2)$$

where  $x_i$  is the proportional fair rate,  $y_i$  is some other feasible allocation and  $w_i$  is again a weight factor.

To illustrate the conditions of proportional fairness as stated above we present the following commonly used example. Consider the network shown in Figure 2.3 which consists of  $n$  links of capacity 1 and  $(n+1)$  flows with weight  $w_i = 1$ . Flow 0 goes through all links; flow 1 through link 1; flow 2 through link 2 and so on. It is intuitive that all flows  $x_1$  through  $x_n$  must have an equal allocation say  $p$ . Hence, the allocation for flow  $x_0$  is  $(1-p)$ . The proportional fair utility function then becomes

$$\sum w_i U_i(x_i) = \ln(1-p) + n \ln(p) = f(p)$$

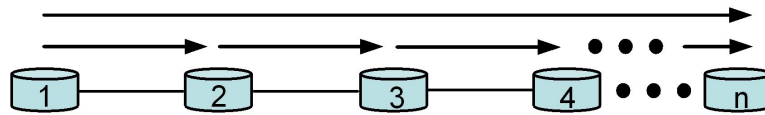


Figure 2.3 Proportional fairness network example

We can then determine the proportional fair rates by finding the maximum of  $f(p)$  which occurs when the derivative is equal to zero.

$$\frac{df}{dp} = -\frac{1}{1-p} + \frac{n}{p} = 0$$

which gives

$$p = \frac{n}{n+1}$$

As can be seen from this example, where the weight of each flow is equal, maximizing the proportional fair utility function provides greater overall network throughput than a max-min solution at the cost of penalizing long routes. The following examples illustrate the usefulness of the logarithmic proportional fair utility function in situations where economical factors may effect the weight of the utility functions.

Consider again the parking lot scenario of figure 2.1. In our first example we look at a situation in which the network is over-provisioned. Assume the weight factor of flow 1 and 2 is .4 and the weight of flow 3 and 4 is .2. Using eq. 2.3 we calculate the weighted proportional fair rates as 1/3 for flow 1 and 2 and 1/6 for flows 3 and 4. It is left to the reader to verify that these weighted proportional fair rates satisfy both eq. 2.1 and eq. 2.2.

$$x_i = \frac{w_i}{\sum w_i} \quad (2.3)$$

Our final example illustrates a weighted proportional fair allocation when the network is under-utilized. Again we look to figure 2.1 and assume a weight factor of .2 for flows 1 and 2 and .1 for flows 3 and 4. In this situation, proportional fairness suggests that we distribute the extra capacity proportionally among all competing flows such that flows 1 and 2 receive an allocation of 1/3 and flows 3 and 4 receive 1/6 just as in our previous example. In this manner each flow is proportionally entitled to use the additional capacity if their capacity requirements change. Again these allocations satisfy both eq. 2.1 and eq. 2.2

As mentioned earlier each fairness model has its strengths and weaknesses depending upon the network metric that is of the most importance for various QoS requirements. The following sections discuss fairness in a few common ring and bus MAC protocols.



## 2.4 Feedback Based Resilient Packet Rings

Due to the overwhelming prevalence of ring networks in the metropolitan area, efforts are underway to exploit the already existing ring infrastructure to more effectively accommodate asynchronous data traffic. The IEEE 802.17 RPR working group was formed in 2000 to develop a standard for bi-directional metropolitan packet rings that support both connection-oriented and dynamic bandwidth provisioning to suit both voice and data traffic. The objective is to overlay a bandwidth sharing protocol on existing ring infrastructures to support data traffic more efficiently.

One of the desirable properties of Resilient Packet Rings (RPR) is the support for destination packet removal. With destination packet stripping, a packet may not traverse all ring nodes, thus, total network utilization can be improved by taking advantage of spatial reuse. To maximize spatial reuse in packet ring networks the Ring Ingress Aggregated with Spatial Reuse (RIAS) fairness model is introduced.

### 2.4.1 Ring Ingress Aggregated with Spatial Reuse (RIAS) Fairness

RIAS fairness has two key components. The first component is based on the ingress fairness model discussed above to define the granularity for fairness determination. Thus, in the event of congestion, the RIAS reference model ensures that all nodes contributing to the congestion are entitled to an equal share of bandwidth over the congested link. The second component governs spatial reuse subject to the first constraint. That is, any node can reclaim unused bandwidth such that the bottleneck link is not affected [92].

To illustrate spatial reuse we present the parallel parking lot example in Figure 2.4. In this scenario, RIAS fairness suggests that each flow en route to station 5 is entitled to 1/4 of the link bandwidth of the most congested link, 4-5. In addition, to exploit spatial reuse, flow (1,2) is allowed to reclaim the excess capacity on link 1.

The key challenge to providing fairness with spatial reuse is the ability to design a bandwidth allocation scheme that can dynamically achieve such fair rates. The difficulty of dynamic fairness updates arises in that there must be timely coordination between nodes, i.e.

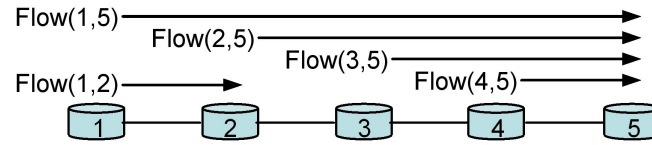


Figure 2.4 Parallel parking lot

upstream nodes must be aware of downstream nodes usage requirements to be sure they are not consuming more than their fair share. This coordination proves to be very difficult in that instantaneous feedback in the upstream direction is not practical as we will see with the oscillation phenomenon of the RPR protocol.

To understand the RPR protocol, we first take a look at the RPR node architecture and then discuss the properties of the RPR fairness mechanism.

#### 2.4.2 RPR Node Architecture

As traffic sourced from a node may or may not be destined for a congested link, each RPR node maintains multiple rate controllers for node ingress traffic. As traffic enters the First In First Out (FIFO) queues it passes through rate controllers that throttle traffic to the fair rate at a per destination granularity. This granularity allows a type of virtual output queue to avoid head of line blocking. I.e., traffic destined for a congested link may be throttled at a different rate than local traffic.

Another feature of the node architecture is that transit traffic has priority to enter the FIFO queue over transmit traffic. This ensures that once a packet is injected into the ring, it will not be dropped at a downstream node. However, since packets will not be dropped in transit to the destination, downstream nodes may not be allowed to source local data on to the network if they are overwhelmed by upstream traffic, hence the need for a fairness mechanism.

#### 2.4.3 RPR Fairness Algorithm

The RPR standard discusses two modes of operation: Aggressive Mode and Conservative Mode. The function of both modes is similar and described as follows. When a node experiences

congestion it will advertise to the upstream nodes, via the opposite ring, its congestion status and a suggested fair rate. Upon receiving a fairness control message, any station contributing to the congestion is required to decrease their send rate to the advertised value. Congestion is alleviated once all stations' rates are set to this fair rate. Once congestion clears, stations periodically increase their send rate in attempt to maximize throughput and spatial reuse [93, 24, 74].

Due to the propagation delay associated with backpressure media access control schemes, stations cannot converge to fairness instantaneously. Also, because congestion can come and go, aggressive congestion signals may not accurately reflect the immediate congestion status or true fair rate. Thus, in adverse situations nodes oscillate in search of the correct fair rates. The following sections outline these fairness drawbacks.

#### 2.4.4 RPR Oscillations

Under the aggressive mode of operation, stations increase their send rate until congestion occurs. Once congestion occurs, a backpressure message is sent upstream requiring all nodes contributing to the congested link to reduce their traffic flow to the minimum input rate. This advertised rate, however, may not be the true RIAS fair rate. Consequently, nodes may over-throttle their send rate below the RIAS rate. Under conditions where traffic from different nodes is severely unbalanced these oscillations will continue and lead to a throughput degradation. The following example illustrates this oscillation phenomenon.

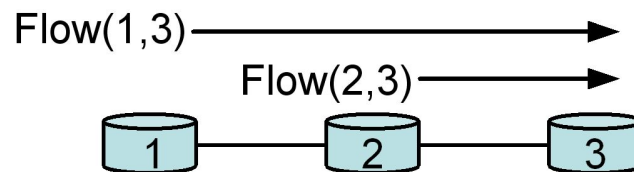


Figure 2.5 3 node network portion to illustrate the RPR oscillation scenarios

Consider the 3 node network portion of Figure 2.5. Assume flow (1,3) has demand for the full link capacity  $C$  and flow (2,3) requires some small capacity  $\epsilon$ . Since the aggregate

traffic exiting node 2 is  $C + \epsilon$ , the link will become congested. This congestion will perpetuate a backpressure message advertising the minimum input rate of  $\epsilon$ . Consequently, node 1 will throttle its flow from  $C$  to  $\epsilon$ . After congestion clears, node 1 will begin to increase its rate to  $C - \epsilon$  where congestion will occur again. This oscillation cycle will continue to repeat with the upstream flow permanently fluctuating between  $\epsilon$  and  $C - \epsilon$ .

A similar situation can be seen in the conservative mode of operation as well. Under conservative mode, congestion occurs when traffic exceeds some low-threshold; .8 of full link capacity by default. The fair rate is then determined by the number of active flows contributing to the congestion during a predefined period of time called the aging-interval. The fair rate is then given as,  $C/n_a$ , the full link capacity divided by the number of active stations contributing to the congested link.

For example, consider a similar two-flow situation as depicted in Figure 2.5. Assume flow (2,3) requires full capacity  $C$  and flow (1,3) has demand  $\epsilon$ . Since both nodes are active when congestion occurs, the advertised fair rate sent from node 2 is then  $C/2$ . Consequently, node 2 will throttle its ingress traffic to this rate during the next aging-interval. Congestion will clear and node 2 will periodically increase its flow until the low-threshold is reached again. Thus, node 2 oscillates between  $C/2$  and low-threshold -  $\epsilon$ .

Due to these bandwidth oscillations it has been shown in [33] through analysis and simulation that throughput loss compared to the RIAS fair rate can be as much as 26% using aggressive mode and up to 30% for conservative mode. This deviation from RIAS fairness and oscillation inherent in the RPR fairness algorithm has led to the development of the Distributed Virtual Time Scheduling In Rings (DVSR) protocol in [33]. Although a description of DVSR will not be presented here the authors show that the DVSR converges to the RIAS fair rates faster than either the aggressive or conservative mode RPR protocol in all situations. In addition, it is shown that the DVSR protocol does not suffer from permanent oscillations as significantly as the RPR MAC.

Although Ring networks are the most prevalent network topology in the metropolitan area it is important to discuss fairness in various other network topologies to arrive at a

complete network fairness solution. The following section presents a discussion of fairness using reservation mechanisms in high speed slotted bus networks.

## 2.5 Reservation Mechanisms for Slotted Bus Networks

Full utilization for a slotted bus network is defined in [23] as the property that a station with data segments to send never releases an idle slot unless it can be used by some further downstream node. This definition ensures that the work conserving property mentioned earlier is maintained. It is also shown that a protocol with full utilization provides the minimum average queuing delay.

A greedy access scheme in which any station with traffic to send fills each passing slot is an example of a fully utilized protocol. Although the greedy protocol provides full utilization it is not fair since upstream stations can monopolize the channel. In addition, there is no guaranteed bound on access delay since downstream stations may wait forever for an idle slot. Thus, we note that for an algorithm to provide fair access it must be non-greedy.

Non-greedy protocols that attempt to provide fair bandwidth allocation on slotted bus networks operate on the premises that stations at the head of the bus must be restricted in some manner. Often times however, blindly restricting upstream stations can have adverse affects on total network utilization.

One method for coordinating such access restrictions is to require all stations to make bandwidth reservations before transmitting. Based upon these reservations bandwidth can be allocated using a distributed schedule or through the use of a centralized controller. The Distributed Queue Dual Bus (DQDB) protocol attempts to construct a distributed FIFO queue to provide access and is discussed below.

### 2.5.1 Distributed Queue Dual Bus

DQDB was one of the first proposals to enable fairness on slotted bus networks. The fundamental access mechanism for DQDB has been around for over 20 years. Since then many enhancements have been proposed and implemented to overcome some of the inherent

drawbacks [20, 44, 68, 45, 50, 22]. In this paper we provide only a brief outline of its operation to gain a more complete understanding of reservation based fairness methods for metropolitan area network MAC protocols.

The central idea behind the DQDB protocol is to use bandwidth reservation to facilitate fair media access on shared bus networks. Each station in a DQDB network is connected to two unidirectional buses which carry traffic in opposite directions. Data is carried on slots of fixed size generated at the head of each bus. Each slot header contains a busy bit and three request bits, one for each level of priority supported by the protocol. The busy bit is used to indicate that a slot is full and the request bits are used to make reservations. For ease of discussion we assume only one priority and define one bus as the data bus and the opposite as the reservation bus. In general, when a station has a packet to send it must first make a request, by setting the request bit of a slot on the reservation channel. Subsequently, a non busy slot on the data channel will be used to carry the traffic to the intended destination.

Fairness is facilitated through the use of two counters maintained at each node, a request counter and a countdown counter. When a station is idle, the request counter is incremented for each request detected on the reservation channel and decremented for each empty slot passing on the data channel. Thus, each stations request counter forms a distributed queue indicating the number of unserved downstream reservations. When the station becomes active (a packet is awaiting service) it makes a request on the reservation bus and copies the value of the current request counter into the countdown counter and resets the request counter to zero. The countdown counter now reflects the number of reservations that are ahead of the local data segment. In addition, the request counter continues to count the number of reservations that are queued after the local segment. When the countdown counter reaches zero, the station is allowed to transmit its data segment in the next non busy slot.

As we saw in RPR, unfairness arises from the fact that backpressure messages are not instantaneous. Similarly, in DQDB, network propagation delay can effect the view each station has on the network status. That is, each stations distributed queue may not accurately reflect the actual number of outstanding requests. When propagation times are significantly longer

than the slot duration, unfairness in DQDB networks is accentuated. To illustrate a situation in which unfairness can occur in DQDB networks we consider the following scenario.

Assume two stations that require full capacity are separated by a large number of slot distances,  $D$ . If the upstream station begins transmitting before any requests are made from the downstream station it will consume all empty slots. Then, when the downstream station makes a request, it will take  $D$  slots before the upstream station notices the request and leaves an idle slot. It then takes an additional  $D$  slots before the downstream station can transmit its packet and issue another request. Hence, the upstream station will consume significantly more resources than an equally weighted downstream station and monopolize the bus. A similar situation occurs if the downstream station begins transmitting first. In this case the reservation bus will be saturated by requests from the downstream station forcing the upstream station to let the majority of slots pass unused.

Unfairness in DQDB networks has been studied extensively in the literature and various enhancements have been suggested [57, 80, 51, 17]. Most of these modifications attempt to promote additional fairness by placing additional restrictions on stations at the head of the bus. DQDB with bandwidth balancing is one such solution that requires upstream station to periodically pass unused slots even if they have not been reserved. It is shown that fairness characteristics can be improved with the use of bandwidth balancing under heavy network loads. However, this blind restriction of upstream station can have an adverse effect on network utilization.

DQDB is an example of how reservation mechanisms can work when a reservation bus exists in the reverse direction. In the case of unidirectional bus networks where such reservation is not available we need to look at different protocols to provide fairness. The next section presents the Pi-persistent protocol for unidirectional bus networks.

### 2.5.2 Pi-persistent Protocol

Much work has been presented in the literature pertaining to media access control for high speed unidirectional buses where feedback in the upstream direction is not available in a

timely manner. Without some sort of reservation or backpressure mechanism it is realized that stations near the end of the bus can be treated unfairly or starved of resources if upstream stations are not restricted in some manner. Realizing this, Mukherjee et al. presented the Pi-persistent protocol to govern bandwidth sharing on a unidirectional bus.

In this protocol, if station  $i$  has a packet to send, it will persist to transmit the packet in the next empty slot with a unique probability,  $p_i$ , until the transmission is successful. Various fairness criteria are modeled and simulated in [63, 52], including, equal mean packet delay, equal blocking probability (for a finite buffer implementation) and equal throughput for all stations. We will discuss the case of equal mean packet delay for all stations employing an infinite buffer as presented in [63].

### 2.5.2.1 Calculating the Persistence Factor

The queuing model for the Pi-persistent protocol for the infinite buffer case is relatively straight forward and is presented in more detail in [63, 52, 59, 61]. Here we only present a brief overview.

Under the Pi-persistent protocol, it is assumed that packets are of fixed length and have duration of one slot. In addition, the packet arrival process at station  $i$  is assumed to be Poisson with average arrival rate  $\lambda_i$  packet/slot. Thus, the average overall offered traffic to the network is also Poisson with parameter  $\lambda = \sum \lambda_i$ .

In order to determine the  $p_i$  that produces equal packet delay for each station, we define the packet service time,  $x_i$ , as the time taken by a packet after it has reached the head of the buffer to be successfully transmitted. Under the Bernoulli approximation,  $x_i$  has the following probability mass function.

$$Pr(x_i = k\text{slots}, k \geq 1) = (1 - r_i p_i)^{k-1} r_i p_i$$

Where  $r_i = \Pr(\text{slot arriving at station } i \text{ is empty at time } t)$  it then follows that the first moment of access delay at node  $i$  is given by



$$E[x_i] = \frac{1}{r_i p_i} \quad (2.4)$$

And

$$Var[x_i] = \frac{1 - r_i p_i}{(r_i p_i)^2} \quad (2.5)$$

We note that  $r_1 = 1$  because all slots arriving at station 1 are empty and  $p_n = 1$  since the last station should use an empty slot with probability 1 for maximum utilization. Furthermore, the probability that station  $i$  will successfully transmit on an empty slot is  $p_i q_i$  where  $q_i = \Pr(\text{station } i \text{ is busy at time } t)$ . Lastly, an empty slot arriving at station  $i-1$  will also be empty at station  $i$ , if station  $i-1$  does not use it for transmission. Thus, we have the recursive relationship governing  $r_i$ .

$$r_i = r_{i-1}(1 - p_{i-1}q_{i-1}) \quad (2.6)$$

Since station  $i$  can be approximated by a single queuing system, we apply the mean value analysis of the classical M/G/1 queue in combination with 2.4, 2.5 and 2.6 to arrive at the individual  $p_i$  probabilities given in 2.7.

$$p_i = \frac{2(1 - \lambda) + \lambda_i(1 + \lambda - \lambda_n)}{(2 - \lambda_n) \left( 1 - \sum_{j=0}^{i-1} \lambda_j \right)} \quad (2.7)$$

Although the Pi-persistent protocol performs well under light loads ( $\lambda \leq .5$ ) stations near the end of the bus showed significant degradation in the average queuing delay under heavy loads [52]. Thus, slight modifications have been proposed and are presented in the next section.

### 2.5.2.2 Pi-persistent Protocol Improvements

In order to maintain the work conserving nature of the network, improvements are proposed first in [59] and subsequently in [75]. Similar to both methods, slot preemption must take place for stations to regulate their average delay. For example; if station  $i$  has a packet to send, it

first attempts to transmit in an empty slot with probability  $p_i$ . A packet transmitted following this rule is marked as holding a transmission permit. If however, the station is not granted a permit via the original Pi-persistent model, the packet is sent without a transmission permit. Such packets sent (without permit) can be preempted in order to provide other stations the ability to transmit permit holding packets. In this fashion slots are fully utilized if any station has data to send on an empty slot. Although this mechanism enhances performance it requires each station to provide a store and forward buffer to assist slot preemption. This store and forward buffer concept is not a feasible in all optical networks without electronic buffering capabilities as in the unidirectional light trail network as we will see in Chapter 4.

As we have seen in the former section reservation based protocols can provide a certain level of fairness usually at the cost of network utilization. In the following section we look at the final common method of enabling fairness in MAC's: token polling protocols.

## 2.6 Token Polling Protocols

Many different flavors of token based protocols such as the token bus, token ring and FDDI have been standardized and are in use today [70, 69, 26, 27]. Fairness in token based protocols is achieved by successively polling each station competing for resources. Resources are then distributed on an as needed basis in a round robin fashion. Non-greedy token protocols provide fair resource sharing by restricting the duration for which a token can be held. Thus, no station is starved of resources. As with the other MAC protocols discussed within this Chapter many variations and enhancements have been suggested in the literature. Here we provide a overview of Fiber Distributed Data Interface (FDDI) protocol as it is one of the most widely accepted token protocols.

### 2.6.1 Fiber Distributed Data Interface (FDDI) Protocol

The FDDI MAC utilizes a timed token rotation method to control access to the medium. Under a timed token scheme, each station records the time elapsed since the token was last received and determines the fair amount of time available to hold the token for the current

cycle.

Before transmission begins, stations connected to the FDDI negotiate a Target Token Rotation Time (TTRT) which is used to set a bound on the maximum time for subsequent token accesses. By nature of the FDDI protocol, subsequent token access by any station is necessarily less than  $2 * TTRT$ . By bounding the token rotation time (TRT) FDDI supports both synchronous and asynchronous traffic.

Based upon the TTRT, each station is allocated an acceptable allocation to be used for high priority synchronous traffic, the aggregate of which is less than the TTRT. In the worst case each station is allowed to transmit this predetermined amount. If there is capacity remaining, stations are allowed to transmit bursty asynchronous traffic such that they do not overwhelm the transmission medium.

After initialization and TTRT negotiation, stations holding the token can begin transmission. When a station receives the token it loads the difference between the negotiated TTRT and the actual TRT into a Token Holding Timer (THT). If  $TTRT - TRT$  is positive, indicating that the token was early, the station is allowed to transmit asynchronous traffic for up to the THT. If the token arrived late, THT is negative, then the station must immediately pass the token after the guaranteed synchronous traffic has been transmitted.

Timed token protocols are generally considered to be fair in that the round robin service technique allows active stations equal access to the medium while providing equal average access delay bounds. The advantage of token protocols is that weights can be used to satisfy max-min fairness, node ingress traffic fairness and proportional fairness as discussed earlier. The one drawback of token protocols is the token propagation delay which effects total network utilization. Thus, token based protocols are not well suited for long distance networks.

## 2.7 Combination Protocols

To complete our discussion of common access control methods we note that not all protocols are limited to using only one of the methods described above. That is, protocols such as the Cyclic Reservation Multiple Access (CRMA) developed by IBM [65, 64] and the Light-Trial Fair

Access protocols use a combination of the aforementioned methods to provide network fairness. The CRMA protocol uses both reservation and backpressure mechanisms as presented below and the LT-FA protocol relies on reservation and token polling methods and will be discussed in Chapter 4.

### 2.7.1 Cyclic Reservation Multiple Access (CRMA)

CRMA is an access control scheme for high-speed slotted bus networks and consists of two primary access control features: cyclic reservation and reservation cancellation backpressure. The cyclic reservation mechanism is used to facilitate capacity allocation and the reservation cancellation mechanism is used to minimize access delay. In general, stations competing for network resources are allowed to make solicited reservations in a round robin fashion using the cyclic reservation features. If at any time the network becomes overloaded with reservations, these reservations can be canceled by using the cancellation features, which ultimately preclude any station from monopolizing the communication medium. CRMA works as follows.

CRMA is designed for use in a folded bus or dual bus network topology. In our discussion we assume a folded bus architecture in which the head end station provides the access control features and stations transmit on the outbound segment and receive on the inbound segment. In CRMA, stations access the bus according to cycles of slots. Stations reserve slots in a cycle and the head end generates cycles sufficiently long to satisfy these reservations. Two commands: *reserve* and *start*, generated from the head station, govern the reservation and generation of these cycles respectively.

To facilitate slot reservation, the head end periodically issues *reserve* commands, each with its own unique *cycle\_number* and initial *cycle\_length* set to zero. Stations can reserve slots in the cycle as it passes on the outbound bus segment by incrementing the *cycle\_length*. The number of reserved slots is equal to the number by which the *cycle\_length* is augmented. This *reserve\_number* along with the *cycle\_number* is stored in a local reservation queue awaiting future service. When the *reserve* command returns to the head end a reservation is entered into the global FIFO reservation queue reflecting the *reserve\_number* and *cycle\_length*.

Once the cycle is queued in the central queue the head end station issues a *start* command containing the next *cycle\_number* followed by the number of empty slots indicated by its corresponding *cycle\_length*. Upon noticing the *start* command, each station examines its local reservation queue for a matching *cycle\_number*. If the cycle numbers match, the station then transmits the number of slots previously reserved for the matching cycle.

The cyclic nature of access under CRMA provides fairness in that each station is allowed to make reservations on every cycle. In addition, a limit can be imposed on the number of slots each node may reserve in each cycle leading to a weighted or proportional fair solution. However, it is possible that, without restricting reservations, the head end queue can become overwhelmed and lead to network saturation and station starvation. To address this issue CRMA implements the reservation cancellation backpressure feature as described below.

The reservation cancellation mechanism requires two additional access commands, *confirm* and *reject*. In general, a threshold is set at the head node to limit the number of outstanding reserved slots. If the *cycle\_length* of a returning *reserve* command can be queued without exceeding the threshold, the head station sends a *confirm* message, with the corresponding *cycle\_number*, to notify the stations that the specific cycle reservations are accepted. If however, the threshold is overrun, a *reject* message is sent indicating that all outstanding reservations are not accepted.

The central control of the CRMA protocol and the cyclic reservation process is shown to provide high throughput efficiency independent of the network speed and distance or the number of active nodes [65]. Furthermore the CRMA access control mechanism provides a significant degree of control in allocating the bus capacity to meet various QoS requirements. One drawback of the CRMA protocol is that network access delay may suffer in that reservations can only be made when *reserve* messages are issued. Furthermore, stations must wait for the *reserve* messages to return to the head end and sit in queue before being granted which also leads to increased network delay. These drawbacks are addressed as we will see in the next Chapter with the introduction of the Robust Dynamic and Fair Network (RDFN) protocol.

## 2.8 Bandwidth Budget Fairness Model

Through a discussion of existing MAC protocols in the previous sections we see that fairness and network utilization typically share inverse relationships. Greedy algorithms can offer full utilization at the cost of being unfair and fair protocols must sacrifice network utilization. In response to this dichotomy we propose the bandwidth budget fairness model.

In many situations fairness from a stations point of view can be different than that of the network. That is, if a station  $i$  requires some allocation of say,  $x_i$ , then the station will perceive itself as being treated fairly if at anytime at least  $x_i$  is available, regardless of station  $j$ 's allocation. In addition, because internet traffic is bursty, a station may at times require some value  $\delta_i$  less than or more than  $x_i$  for short durations. Thus, it may be possible for the network to statistically share the extra capacity  $x_i - \delta_i$  to enhance some other station  $j$ 's perception of fairness without degrading station  $i$ 's fairness and vice-versa. The network on the other hand will only perceive this as a fair solution if each station  $j$  gets an equal or proportional share of the extra capacity  $x_i - \delta_i$ . However, regardless of which station  $j$  receives this extra capacity the overall utility of the network is improved.

Here is how bandwidth budgeting works. Each station  $i$  begins with a bandwidth budget  $x_i$  for which they pay some amount  $p_i$ . To ensure an acceptable QoS the network must guarantee that  $\sum_i x_i$  is less than the network capacity  $C$ . Now, if during some period, station  $i$  requires some  $\delta_i$  less than  $x_i$ , stations can compete for the  $x_i - \delta_i$  additional capacity to support short traffic bursts. The role of the network is to then keep track of these  $\delta_i$  deviations from the budgeted allocation  $x_i$ . Stations consuming less than their purchased budget will be compensated and those using more will incur additional costs similar to the way in which electricity usage is billed. Thus, no station can unwillingly be deprived of its acceptable fairness, however, stations wishing to be treated more "fairly" can do so with an associated cost.

The bandwidth budget model treats fairness as a commodity. That is, each station begins with an equal amount of "fairness" and must pay for any additional fairness that they receive. Thus, bandwidth budgeting maximizes the utility of fairness and network utilization.

## 2.9 Summary

As mentioned, the difficulty in designing bandwidth arbitration schemes that attempt to maximize network utilization while providing fairness and access delay guarantees is that Quality of Service (QoS) metrics share inverse relationships. That is, a positive change in one metric will most likely have a negative effect on the others.

In order to more fully understand fairness aspects in network protocols this chapter provides a discussion of fairness, access delay and network utilization and their tradeoffs. We presented various fairness models including max-min fairness, ingress traffic fairness and utility based fairness. We outlined three common methods typically used to facilitate fair bandwidth arbitration, namely, backpressure, reservation and token polling methods. We then outlined some of the most recognized MACs for ring, dual bus and unidirectional bus networks. Finally we presented a definition of fairness based on a “bandwidth budget” that attempts to maximize network utilization through residual capacity resource sharing while maintaining acceptable “budget” for all stations.

In Chapter 4 we discuss light-trail technology as a solution that supports all optical wavelength sharing in metro networks. We then present a media access control protocol for light-trails based upon the bandwidth budget fairness model as discussed. But, before we discuss light-trail technology we first present the Robust, Dynamic and Fair network protocol for metropolitan area ring networks. The RDFN protocol provides fair access to all stations on a ring network for bursty data traffic in addition to providing delay and throughput guarantees for isochronous voice traffic. Furthermore the RDFN protocol does not suffer from the oscillations experienced with the RPR protocol nor is it susceptible to large network delays as in the CRMA protocol.

## CHAPTER 3. Robust, Dynamic and Fair Network

An important trend in high-speed networking is the migration of packet-based technologies from Local Area Networks to Metropolitan Area Networks. Our goal in this research is to develop an efficient medium access control mechanism for high-speed metropolitan area networks. The protocol is aimed at fair bandwidth provisioning with varying service level agreements, optimization for both voice and data traffic, and fast recovery in case of node or link failures. The Robust, Dynamic and Fair (RDFN) protocol includes the advantages of both SONET/SDH and high-speed Ethernet technologies.

### 3.1 Introduction

Currently, the majority of traffic transmitted over Metropolitan Area Networks is transported on SONET based networks, which were originally designed for voice traffic. The explosive growth of the Internet, videoconferencing, Storage Area Network (SAN) applications, Virtual Private Network's (VPN) and related data services is causing metro networks to become overloaded and inefficient for bandwidth sharing.

Due to the high cost of overhauling existing metro area physical architectures it is imperative to take advantage of existing MAN infrastructures to provide cost effective solutions to accommodate the increase in data traffic. The already existing fiber ring structure of SONET provides a good physical transport medium to overlay our proposed medium access control.

Another major concern among carriers is how to provide service level agreements to customers. Traditional leased line service is becoming less attractive to customers due to the high variability of bursty data traffic. Although leased line provisioning provides high service guarantees, the customer may not want to pay for the entire channel if it is not fully utilized.



On the other hand, packet based services are attractive to customers in that bandwidth may not be limited to the capacity of a leased line, however, it is difficult for carriers to accurately charge for these services and quality of service guarantees are difficult to maintain.

Due to the increase in connectionless data traffic in the metropolitan area we seek a media access control that provides dedicated service to existing connection-oriented voice traffic and utilizes the remaining network bandwidth for use by connectionless data traffic. To minimize the implementation efforts of such a media access control we strive to make use of the existing fiber ring structure already deployed in the metropolitan area.

The following sections discuss a media access control mechanism for Metropolitan Area Networks that accommodates the increase in data traffic while maintaining a high quality of service for voice traffic. Also a solution for providing Service Level Agreements is given in which a carrier has the option to provision bandwidth by guaranteeing service or providing a “fair” bandwidth allocation service to all customers.

### 3.2 Metropolitan Area Network Structure

Metropolitan Area Networks (MANs) are typically designed as a logical ring, which connect a number of access nodes (typically between 10 and 30) throughout a metropolitan area. These nodes are in place to service geographically separated local area networks. The previous generations MANs employing optical technology are referred to as Synchronous Optical Networks. The limitation in these SONET rings is that they are designed for point-to-point, circuit-switched applications (e.g. voice traffic) where a connection is from end-to-end.

Some of the disadvantages in such rings are listed below.

- Fixed bandwidth allocations - Each node on the ring is given a fixed amount of available bandwidth, which is wasted if not used by that node, thus, dynamic bandwidth sharing is not allowed among nodes.
- Inefficient multicasting - Multicasting is a large application where data from one source is distributed to many nodes. The current SONET ring must create a separate circuit from the source to each multicast destination. A separate copy of the packet is then

sent to each destination resulting in bandwidth wastage and increased workload at the source.

- Wasted bandwidth for protection - Typically half of the bandwidth is reserved to provide protection to tolerate the effects of failures. Hence, the maximum usable bandwidth is only half of the available fiber bandwidth even when there are no failures.

### 3.2.1 RPR Revisited

SONET was designed to support connection-oriented service, a key requirement to guarantee quality for voice traffic. The most appealing element for SONET is the simplicity of the ring structure that provides fast recovery from link failures. More importantly, fiber rings are already present in the metropolitan network.

Ethernet was designed for data traffic where the underlying physical layer is a broadcast medium. Ethernet allows for efficient sharing of available bandwidth in a simpler and less expensive manner. However, the medium access control follows randomized procedures that are efficient for data traffic but do not guarantee quality of service for voice traffic.

Hence, efforts are underway to exploit the already existing ring infrastructure to accommodate data traffic. The IEEE working committee on 802.17 has recently published a standard for Resilient Packet Ring (RPR) architecture that supports both connection-oriented and dynamic bandwidth provisioning schemes to suit both voice and data traffic. The objective is to use existing infrastructure to support increasing data traffic efficiently. However, flexibility and fairness in terms of bandwidth allocation are the main difficulties. In addition, the RPR protocol suffers from large oscillations from the fair rate when stations contributing to congestion have unbalanced requirements as pointed out in Chapter 2.

Of great importance to the introduction of a new media access protocol is the method of providing fair access to fairness eligible bandwidth while maximizing total network throughput. The dilemma when addressing fairness in the context of maximum network utilization as mentioned in Chapter 2 is that under maximum utilization fairness may be compromised and likewise providing fair service may effect network utilization.

The motivation for creating our Robust Dynamic and Fair (RDF) Network is to take advantage of the current ring infrastructure and overlay it with a fair media access control that allows voice and data to exist on the same network. The idea is to make use of a bi-directional ring with one node acting as the controller for each wavelength. This node will be the origin for the scheduling and provisioning of all ring traffic, allowing voice and data traffic to share bandwidth fairly. The protocol will allow for multiple wavelengths scaling along with the ability to provide connection-oriented and connectionless service while offering fairness among such services.

### 3.3 RDFN Network Architecture

Our proposed architecture consists of a bi-directional ring architecture composed of two unidirectional rings similar to that employed in a SONET, but arbitrated by a centralized controller (one for each wavelength). Figure 3.1 shows the network architecture in which nodes are connected by fibers carrying multiple wavelengths.

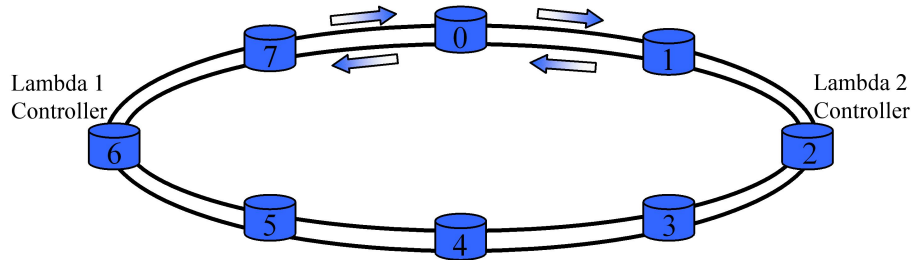


Figure 3.1 RDFN ring architecture illustrating regular and controller nodes

#### 3.3.1 Node Structure

##### 3.3.1.1 Receiver

Using the proposed multi-wavelength architecture each node must be able to receive data on any wavelength. Thus, each node employs a wavelength demultiplexer and multiple fixed receivers, one for each wavelength. Using a tap and continue mechanism, the WDM signal

arriving at the node it is demultiplexed and directed to the appropriate fixed receiver. Packets headers for each wavelength are read and processed at the node. If the data arriving on the specific wavelength is destined for that node the information is electronically stored and processed internally. The packet then continues through the network and, in the case of a multicast transmission, can be processed by each multicast destination without source retransmission. Once the packet arrives back at the controller that originated the slot it is stripped from the network. The node structure is designed to require minimal internal processing. Each node is only required to read the header of each packet and make the necessary decisions. It is not concerned with timing or bandwidth arbitration.

### 3.3.1.2 Transmitter

Two transmission schemes are developed in this Chapter. The first scheme, random wavelength selection, requires each node to possess a tunable transmitter allowing it to request and transmit on a randomly selected wavelength decided upon at packet arrival time. Under this scheme a request is sent to the respective controller on the selected wavelength. This request is granted on the same wavelength and data transmission takes place on the selected wavelength.

The other scheme, fixed wavelength selection, only requires each node to maintain a fixed transmitter. The transmission wavelength for each node is selected based on the proximity to the nearest controller and the transmitter is set to the corresponding wavelength. If the node lies equal distance between two controllers it will choose the downstream controller. For example, nodes 5, 6, 7 and 0 of Figure 3.1 transmit on wavelength 1 and nodes 1, 2, 3 and 4 transmit on wavelength 2. This scheme reduces the propagation time of the request to the controller, but may result in wavelength overload.

### 3.3.2 Controller Structure

The controller structure is somewhat more complex as timing and bandwidth arbitration must be handled at the controller nodes. A controller operates similar to the other nodes with some added functionality. The receiver structure of the control node is identical to that of

all other nodes as it must be able to receive data on all wavelengths requiring a wavelength demultiplexer and multiple fixed receivers.

As the control node may also require network bandwidth for data transmission a tunable transmitter is needed for operation with the random wavelength selection scheme and a fixed transmitter is required for the fixed wavelength selection scheme, similar to that of all other nodes.

In addition to the transmitter and receiver components, each controller queues and issues slots for use by it and the other nodes in a scheduled manner as discussed later; thus a queue must be maintained at the controller. The controller also has the capability to maintain various service level agreements along with the ability to issue connection-oriented services when needed.

The controllers are selected based on their location, number of nodes in the ring and the number of wavelengths used in the ring. For example: if the network has 24 nodes and 4 wavelengths,  $\lambda_0$  is controlled by node 0,  $\lambda_1$  by node 6 and so on. This scheme is used to evenly space control nodes throughout the network. It is possible however, to place control nodes at arbitrary locations on the ring to provide additional load balancing features but is not investigated in this work.

### 3.4 Data Transport Method

The goal for data transport is to develop an efficient bandwidth allocation scheme to achieve fairness and quality of service for both data and voice traffic. In order to utilize the maximum capacity of the ring, both inner and outer rings are used for data transmission.

If the upstream ring is required for data transport, the node sends a request to the control node via the downstream ring and vice versa. The sending node chooses a ring for data transmission based on the location of the destination node and controller. In the following explanation an upstream path is one in which a packet is sent in the clockwise direction. For example, referring to Figure 3.1, node 0 will transmit to nodes 1, 2, 3, 4, 5 and 6 using the upstream ring and will transmit to node 7 using the downstream ring. The controller allocates

and issues time slots containing the following sections:

- Unique ID - Each slot has a unique identifier consisting of an ID and COUNTER that corresponds to the request being filled. As a time slot travels through the network the unique identifier is examined by each node to determine which request is being serviced.
- Request - A request section is appended to each slot to allow for future requests. The request section contains an ID and COUNTER that serves as the unique identifier when the request is granted. Any node can make a request by incrementing the counter, this value is recorded by the node to match a future time slot for service. As time slots flow in both directions the requests are piggybacked upon all time slots.
- Data Slot - Each time slot also maintains a portion of bandwidth for data to be inserted. The proposed portion is 1500 bytes or approximately 1 full Ethernet frame.

As noted earlier, the control node is in charge of bandwidth provisioning. The control node issues time slots in both directions on the ring. The format of an upstream time slot is shown in Figure 3.2.



Figure 3.2 Upstream Time Slot

The upstream ID and upstream COUNTER (UP-ID and UP-CTR, respectively) are the unique slot identifiers for use by the requesting node. The data slot is the portion of bandwidth that is reserved for data packet information, which is inserted by the requesting node. The downstream ID (DN-ID) is incremented for each successive slot and is used to indicate a unique slot issued to the network. Finally the downstream COUNTER (DN-CTR) combines with the DN-ID to create a unique pair that serves as a request identifier for any node requesting a future time slot in the downstream direction.

### 3.4.1 Network Operation

Operation is explained for use with one wavelength and one control node. At the start of time slot issuance the network is empty with each node possibly having bandwidth requests. The control node issues a time slot in the upstream direction with the values shown in Figure 3.3. Each successive slot issued is similar with DN-ID being incremented until the first slot of the respective round arrives back the controller at which time the DN-ID is reset again to 1. Thus, only one specific DN-ID is present in the network at a given time.

|    |    |       |   |   |
|----|----|-------|---|---|
| -1 | -1 | EMPTY | 1 | 0 |
|----|----|-------|---|---|

Figure 3.3 Active upstream time slot (invalid/generic unique ID)

The values of UP-ID = -1 and UP-CTR = -1 signify that this particular slot is not intended to service any outstanding reservation, or that this slot is intended for connection-oriented service, and thus no node may use this slot for “data” transmission. The DN-ID value of 1 signifies this is the first slot issued in the upstream direction for the current round. During each round only one unique DN-ID may be present on the network.

As the time slot propagates through the ring, any node may request a future time slot by incrementing the DN-CTR by 1, as each node may only make 1 request per slot. This DN-ID.DN-CTR combination becomes the unique ID for the future time slot to be issued by the controller in downstream direction. For instance, if node 1 has a data awaiting service, the DN-CTR is incremented as the slot passes and the unique ID of 1.1 is remembered for future service. If node 2 has data awaiting service the DN-ID is again incremented and the unique ID 1.2 is remembered and so on. Once the slot returns to the control node it is stripped from the network. The DN-ID and DN-CTR are then examined and the reservations are queued in the central controller in a FIFO manner. That is, a slot with unique ID DN-ID.1 is placed into the queue followed by a slot with unique ID DN-ID.2 and so on until DN-ID.(max)DN-CTR is reached. Slots are then issued from the central queue in the opposite direction in format shown in Figure 3.4. If the central queue becomes empty, slots are issued with a generic ID.COUNTER pair as in Figure 3.3.



Figure 3.4 Downstream time slot

### 3.4.2 Network Operation Example

The following example will clarify the usage of the dual ring network operating with 1 controller and 4 ordinary nodes as seen in Figure 3.5.

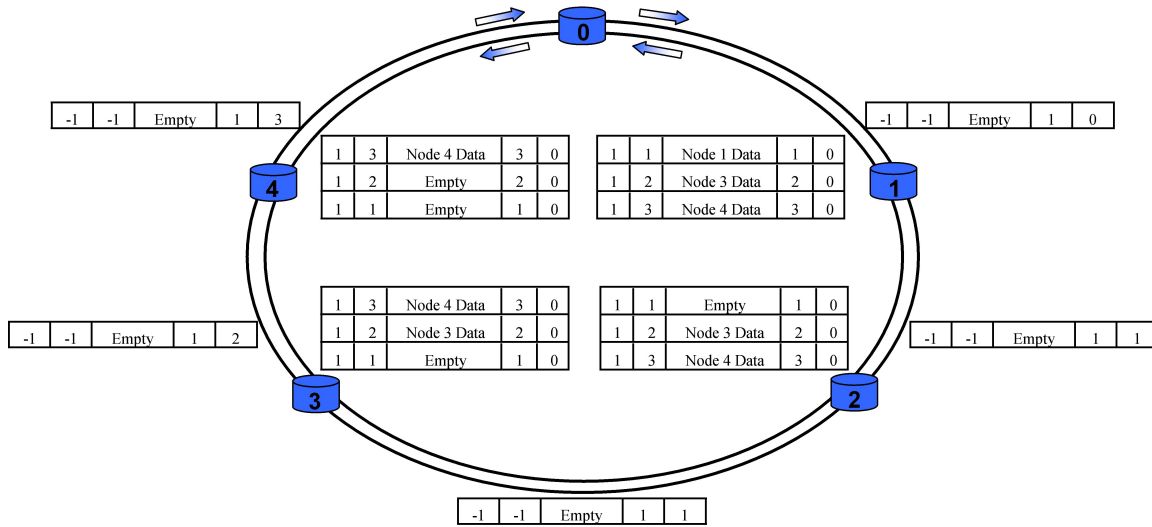


Figure 3.5 Network operation example

For instance, assume nodes 1, 3 and 4 have at least one request at the time the first slot (Figure 3.2) is issued in the upstream direction. Node 1 increments the counter and remembers the unique identifier 1.1. Node 2 allows the slot to pass unchanged as it is not in need of service. Node 3 increments the counter and remembers the identifier 1.2. Node 4 does likewise and stores the request identifier 1.3. Slot 1 with DN-ID = 1 is shown propagating on the outer ring of Figure 3.5.

Once the slot arrives back at the controller it is stripped from the network. The controller then examines the request portion of the received slot and determines the number of slots requested for downstream service and queues them in the central queue. The controller then begins to issue slots from the head of the queue in the downstream direction with the DN-ID and DN-CTR filled to satisfy the received requests.



For each slot issued on the inner ring it should be noted that the UP-ID (field 4) is incremented to identify a slot in the current round. Depending upon the size of the network the maximum UP-ID is variable. The inner ring shows a table of the three slots issued for service based on the received requests. Each node examines the header information to locate a matching unique identifier. When the time slot with a matching unique identifier is discovered the awaiting data is inserted.

In this example the first two slots are allowed to pass node 4 unchanged (unless a future upstream request is to be made). The third slot shown with UP-ID = 3 has a matching unique identifier of 1.3, thus, node 4 data is inserted. The three slots are shown with their respective changes after processing is done at each node. The slots continue through the network until they arrive again at the control node where the request portion of each slot is examined and processed.

### 3.4.3 Multiple Controller Operation

For scalability purposes the proposed network can be modified to accommodate multiple wavelengths by employing multiple controllers. In this case the controllers are spaced evenly throughout the ring each arbitrating bandwidth on a specific wavelength. By adding additional controllers and wavelengths the network is easily scalable to a wavelength division multiplexing solution.

The results given in Section 3.6 are for a network with 24 nodes employing 4 wavelengths. A similar request and service scheme, as explained above, is used with all controllers which divides network usage among all available wavelengths while maintaining fair bandwidth arbitration.

### 3.4.4 Additional Considerations

One of the advantages to stripping a packet from the network at the control node is seen with multicast traffic. The current SONET solution requires a new connection be established for each multicast destination. Our solution allows a packet to pass all nodes en route to the controller before being removed from the network. Thus, a packet need only be transmitted

twice, once in each direction, to reach all nodes on the ring.

Another improvement over existing solutions is the ability to provide fast recovery in the event of a failure without reserving bandwidth for protection as is done in SONET/SDH. Our solutions utilizes both inner and outer rings for data transport as opposed to reserving one for protection; in the event of a failure each node will make adjustments to only transmit requests on a chosen ring to avoid the failure.

### 3.5 RDFN Fairness Scheme

Fairness has been described in many different ways to suit different applications as discussed in Chapter 2. The fairness criterion for the RDF network is to provide equal access delay for all nodes while maintaining strict guarantees on delay and jitter to support isochronous voice traffic.

As mentioned earlier, the proposed access control mechanism makes efforts to allow connection oriented service along with a connectionless service for each node. Based upon various levels of dedicated service we show that available bandwidth can be shared among the remaining nodes without neglecting requests from any single node.

Because the connection-oriented service is of utmost importance to maintain quality of service, highest priority is given to this type of traffic. The bandwidth required for these services is not available for connectionless service. Each node has the ability to purchase such guaranteed service from the network carrier. Connection-oriented slots are provisioned by the controller node at predefined intervals and are not required to wait in the central queue. That is, if node 1 requires 5% capacity for dedicated service traffic, the controller issues 1 slot every 20 with a specific ID.COUNTER pair corresponding to node 1. Thus, reservations are not required for connection-oriented service, rather the central control node provisions this traffic regardless of the status of the central queue, similar to the way in which guaranteed service is provisioned under the FDDI protocol. After all dedicated service slots are issued the remaining bandwidth is provisioned based upon the reservations made to the central queue. This, additional bandwidth is available to all nodes, and is allocated in a fair manner.

The mechanism used for bandwidth provisioning is naturally controlled by the request scheme. As mentioned earlier, each time slot consists of a unique identifier, payload and request. Because each node may make only one request per time slot, all nodes have an equal opportunity to make a request before the slot arrives back at the controller and service slots are queued. When the control node receives the time slot and examines the request portion, time slots are queued in the central queue based on this information. By the nature of the reservation scheme each slot received by the central controller cannot contain more than one request from any single node and thus each nodes request is scheduled in the same time period.

In summary the request scheme guarantees that no node is allowed more than one successive time slot if any other node requires a portion of the available bandwidth. Furthermore, any node may be given successive slots if no other node requires bandwidth but is restricted from overloading the central queue.

This fairness scheme has been developed to allow all nodes to compete equally for available bandwidth while maintaining the ability to service connection-oriented requests. As shown in the results, the wait time per packet, measured in number of slots, is similar for all nodes on the network. Thus showing that available bandwidth is equally shared among all nodes independent of total network load.

### 3.5.1 Additional Considerations

The fairness scheme is aimed at providing maximum network throughput while sharing available bandwidth among all nodes. In order to maintain the proposed fairness scheme it is required that all nodes follow the bandwidth request rules. A problem may arise when a node acts without consideration of these rules. As stated, each node is only allowed to make one request per slot. This ensures that all nodes have equal priority to the available bandwidth. By incrementing the counter on any particular slot by more than one and storing both ID.COUNTER pairs a node can receive an unfair amount of bandwidth and go undetected. This approach though, if implemented correctly, can provide a weighted node scheme for fair bandwidth provisioning.

In such a case normal bandwidth provisioning is done in the manner described earlier with the possibility of any single node being granted an additional proportion of bandwidth based upon their individual load. For example, consider a node that services twice the amount of customers as all other nodes. This particular node is allowed to make two requests on each passing slot. This method effectively reduces the wait time experienced by the heavily burdened node in a proportional manner.

### 3.6 Performance Evaluation

In order to demonstrate the operation of the ring network described above, a network simulator is built using Visual C++. The following paragraphs describe the functions and methods used to build such a simulator. Tests are performed on the various functions to verify their functionality and are also briefly described.

#### 3.6.1 Request Generation

To simulate a large Metropolitan Area Network a request generation function is created to control the frequency of bandwidth requests. The duty of the request generator is to simulate a network in which a Poisson distribution governs the inter-arrival time of all network requests. For simulation purposes the network is assumed to have 100 slots per round. Two different types of bandwidth requests, connection-oriented (voice) and connectionless (data) and their generation characteristics are explained below.

To simulate connection oriented traffic a variable is used to specify the amount of connection oriented service to be provisioned. This variable is set at run time and can be changed to reflect an increase or decrease in voice traffic as is needed to provide varying service level agreements. The provisioning for voice traffic is done periodically throughout each round. I.e. if 20 percent of all network traffic is specified for voice traffic, two of every ten slots are issued for such traffic and thus not available for use by best effort data traffic requests. To signify slots for use by voice traffic the arbitrary unique identifier -1.-1 is used, thus no node can have a matching unique ID and the slots cannot be used for data transmission.

The connectionless traffic requests are generated in a somewhat different fashion. For simulation purposes each node maintains two request queues, one for both upstream and downstream requests. At the beginning of simulation a request generation function is called to fill each request queue. The request queues hold requests for each node and are generated using a randomized process.

We used a Poisson distribution to govern the arrival rate of the messages. Based upon total network load the average arrival rate is specified. Generic requests are generated for all nodes based on this arrival time. The requests are then associated with their respective node queue based on the source/destination pair and destined control node.

A second Poisson distribution is used to determine the message size (in number of slots) of each message based on an average packet size, which is set at run time. Each message follows the following generation characteristics.

- A negative exponential random number is used to determine the arrival time of the next message.
- A negative exponential random number is used to determine the size of the next message based upon the average message size.
- A uniform random number is used to determine the source and destination pair of the message.
- A uniform random number is generated to specify the destined wavelength for use with the random wavelength selection scheme.

Once the arrival time, size, source, destination and wavelength are determined the packet is queued in the respective node queue.

### 3.6.2 Central Controller Slot Queue

The simulation begins with a function that queues slots in the centralized queue for each respective wavelength. The slot queue function examines the request portion of the most recently received time slot. This information is used to queue the necessary slots for service

as described earlier. Each queued element is given the unique identifier associated with the request.

### 3.6.3 Central Controller Slot Issue

Once the requested slots are queued in the central queue a slot is issued onto the network in one of two ways.

1. If the next slot is to be taken from the pool of available bandwidth, for use by connection-less traffic, the first element of the central queue is removed and issued with the specific ID.COUNTER value. If the central queue is empty, a slot is issued with the ID -1.-1. In this case the slot goes unused and is wasted.
2. Otherwise the ID.COUNTER field is set to -1.-1 to suggest this slot is to be used for connection-oriented traffic, nodes are able to use this slot to request future slots but cannot transmit in this time slot.

As mentioned earlier connection-oriented slots are given the highest priority and are allowed to jump to the head of the queue. For a network with 100 slots per round and 20% connection-oriented service, 20 slots are issued each round (1 of every 5) for voice traffic. Thus, slots 1, 6, 11, 16 and so on are scheduled for voice traffic independent of the size of the central queue.

### 3.6.4 Node Processing

After a new slot is issued from the centralized controller, node processing begins. Node processing is done at each node as the slot passes through the network. Two functions handle the node processing section. The first deals with future requests and the second deals with slot usage.

To make a bandwidth reservation the head of the request queue for the current node is examined to determine if a packet is waiting. If so, a request is placed into the service queue awaiting a slot for service and the request counter of the current slot is incremented to reflect a slot has been requested.

The other function handled in the node processing section deals with slot usage. If the current slot is destined for the node doing the processing, the head of the service queue is examined for consistency and is removed from the queue, indicating the request has been serviced.

To accumulate the statistics of the simulation the node processing section also calculates network metrics. When a slot is requested and hence moved from the request queue to the service queue the request time is recorded. When the corresponding time slot arrives to service the request the service time is recorded. The difference between request time and service time is calculated to determine the wait time for each request. The slot then propagates through the network until it arrives again at the central controller to be stripped from the network and examined to begin the queue slots process in the opposite direction.

### 3.7 Simulation Results

#### 3.7.1 Random Wavelength Selection Scheme

The results displayed below are for the random wavelength selection scheme described previously. As noted, a uniform random number is used to choose the wavelength to transmit the corresponding request on. This method evenly distributes network load across all available wavelengths regardless of the geographical location of the requesting node.

One disadvantage of the random wavelength selection scheme is in the set up time. Set up time is described as the constant time required for a request to travel to the controller added to the time for the corresponding service slot to return to the requesting node. Although set up time is not examined extensively in this dissertation, consideration is given when comparing the random wavelength selection scheme to the fixed selection scheme.

As mentioned, the goal of the proposed media access control is to allow connection-oriented service and bursty data traffic to share bandwidth efficiently and fairly. In order to maintain a high quality of service for voice traffic, highest priority is given to connection-oriented requests with the remaining bandwidth being shared among all nodes. Thus, connection-oriented service requests are allowed to enter the network at any time while data requests must wait. This wait

is used to illustrate how fair service is achieved. The importance of this metric is to show that available bandwidth is indeed shared among all nodes fairly.

Samples from 5 million requests are used to calculate wait time averages for various network loads. In each case it is noted that wait times are similar for all nodes. In other words, a message from any node waits approximately the same amount of time before being fully serviced. The results displayed reflect the average wait times with varying connection oriented service requirements and network loads. The standard deviation and 95% confidence levels are listed to suggest only a small variance is exhibited between individual node wait times. Tables 3.1, 3.2 and 3.3 with corresponding Figure 3.6 shows wait times in number of slots for wavelength 0. Figure 3.6 shows the graphical depiction of Tables 3.1, 3.2 and 3.3; 95% confidence is signified with error bars shown on each data series.

Table 3.1 Average wait time with 10% connection oriented traffic

| Total Load     | 40%  | 50%  | 60%   | 70%   | 80%   |
|----------------|------|------|-------|-------|-------|
| Average Wait   | 5.74 | 8.62 | 13.30 | 20.85 | 36.44 |
| ST Dev         | 0.29 | 0.52 | 0.82  | 1.34  | 1.34  |
| 95% Confidence | 0.12 | 0.21 | 0.33  | 0.56  | 0.54  |

Table 3.2 Average wait time with 20% connection oriented traffic

| Total Load     | 40%  | 50%  | 60%   | 70%   | 80%   |
|----------------|------|------|-------|-------|-------|
| Average Wait   | 5.18 | 7.71 | 11.67 | 18.52 | 32.14 |
| ST Dev         | 0.22 | 0.33 | 0.67  | 0.98  | 1.22  |
| 95% Confidence | 0.09 | 0.13 | 0.27  | 0.39  | 0.49  |

For completeness and clarity Table 3.4 with the corresponding Figure 3.7 is shown to represent all wavelengths. It is noted that each wavelength showed similar wait times as can be seen from these data points. The data points shown are for a network with varying load and 20 percent connection-oriented service.



Table 3.3 Average wait time with 30% connection oriented traffic

| Total Load     | 40%  | 50%  | 60%   | 70%   | 80%   |
|----------------|------|------|-------|-------|-------|
| Average Wait   | 4.51 | 6.81 | 10.18 | 16.10 | 27.93 |
| ST Dev         | 0.23 | 0.26 | 0.39  | 0.79  | 1.19  |
| 95% Confidence | 0.09 | 0.10 | 0.16  | 0.32  | 0.48  |

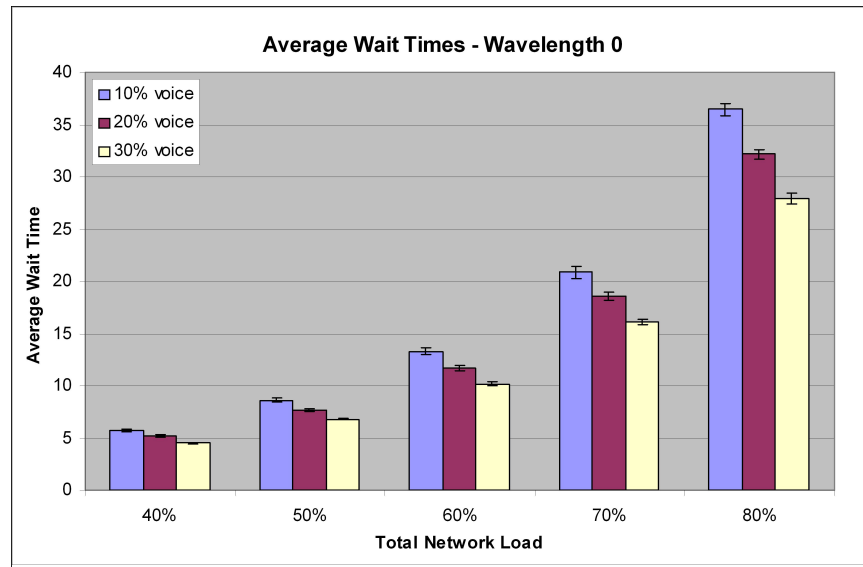


Figure 3.6 Average wait times of all nodes for Wavelength 0

### 3.7.2 Fixed Wavelength Selection Scheme

The results displayed below compare the fixed wavelength selection scheme described with the random selection scheme illustrated above. As noted, for the fixed wavelength selection scheme, wavelengths are selected on a fixed basis according to the physical location of each node in proximity to the nearest controller. Thus, all requests from any node are predetermined to be serviced by the nearest controller.

Experimental results obtained are similar to those of the random wavelength selection scheme with a slight difference pertaining to individual wavelength load due to the request generation process. With the fixed scheme, wavelength load show more variance between up and down requests on each wavelength. The average wait time combined for requests in

Table 3.4 Average wait time for all wavelengths with 20% connection oriented traffic

|                | Wavelength 0 | Wavelength 1 | Wavelength 2 | Wavelength 3 |
|----------------|--------------|--------------|--------------|--------------|
| 40% total load | 5.18         | 5.14         | 5.03         | 5.17         |
| 50% total load | 7.71         | 7.76         | 7.59         | 7.95         |
| 60% total load | 11.67        | 11.79        | 11.47        | 12.01        |
| 70% total load | 18.52        | 18.90        | 17.81        | 18.74        |
| 80% total load | 32.14        | 33.20        | 30.91        | 32.90        |

both directions for each node however remain similar with a high degree of confidence. This combined wait time is displayed in Figure 3.8 shown next to the wait times for similar network loads in the random selection scheme with voice traffic fixed at 20 percent.

In terms of set up time, the fixed wavelength selection method is preferable when multiple wavelengths are used. Since requests are sent to the nearest controller set up time is also fixed. As more wavelengths are added to the network this method becomes more and more appealing as fewer nodes are competing for the same available bandwidth and controllers become closer in proximity.

### 3.7.3 Wait Time per Packet Element

Of other interest is the total time taken to service requests of different size with 20 percent connection-oriented load and varying data loads. Shown in Figure 3.9 is the average wait time for each packet element. A packet element is considered as one request of any given message. This value is calculated by taking the average time to completely service a packet divided by the packet size. These values are averaged over all nodes and compared; the 95% confidence mark is also shown to emphasize the small variance between nodes.

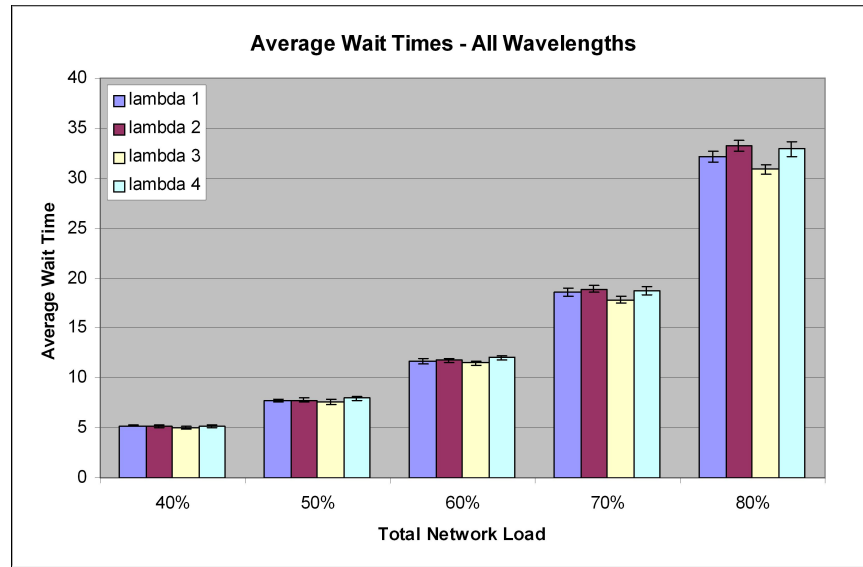


Figure 3.7 Average wait times for all wavelengths

### 3.8 Summary

As stated, the goal of the access control mechanism is to allow available bandwidth to be shared fairly for bursty data traffic among all nodes in a metropolitan area network along with maintaining a high level of service for connection-oriented traffic. To show this we examined queue wait times as experienced by all nodes in a simulated network. This wait time reflected the amount of time a single request must wait before a network time slot is issued for service. As shown in the data presented, all nodes experienced nearly the same wait time with a high level of confidence.

Another noteworthy aspect of this network is the influence of connection-oriented traffic on the network. As can be seen from Figure 3.6 an increase of voice traffic by 10 percent did not affect the request wait times as significantly as a 10 percent increase in data traffic. This indicates that connection-oriented or high-priority service can be added to such a network with less wait penalty than a similar increase in bursty data traffic.

As mentioned in the introduction, a major concern for network carriers is how to offer service level agreements to customers. As seen in the results, adding additional connection-oriented or guaranteed service has a lower wait time cost than adding an equal amount of

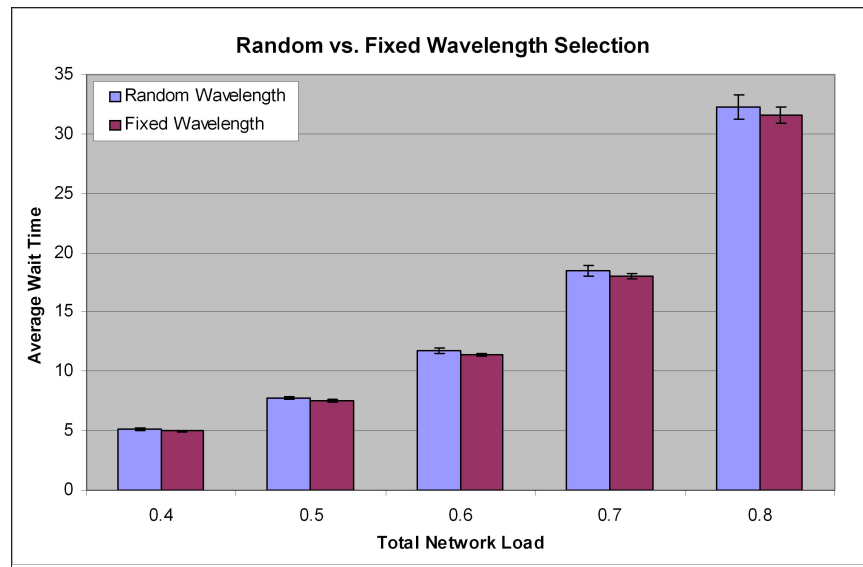


Figure 3.8 Comparison of wait times for the random and fixed wavelength selection schemes

bursty data traffic. Thus, with a slight modification to the control node, network carriers have the ability to provision guaranteed service at a fixed cost to customers with minimal effect to the service quality offered to all other nodes. Also, because of the nature of the request and service scheme the amount of traffic used by each node can be recorded allowing carriers to bill customers only for what is used.

Although the RDFN protocol resembles the CRMA protocol discussed in Chapter 2 the RDFN protocol exhibits a few differences and advantages. One key advantage is that the RDFN is easily scalable to a multiple wavelength solution as described above. Another major difference is that by allowing stations to make requests on every passing slot instead of waiting for a solicitation period as in CRMA total network delay is reduced. A final note that differentiates the RDFN protocol from CRMA is that stations can only make one reservation per slot whereas CRMA allows multiple slot reservations per cycle. By restricting stations to only one slot reservation no single node is allowed to monopolize the bus for duration of more than one slot if the network is congested which leads to improved fairness and lower access delay.

Through the use of the proposed protocol we are able to provide bandwidth for different traffic demands of voice and data. The MAC allows sharing available bandwidth in a fair

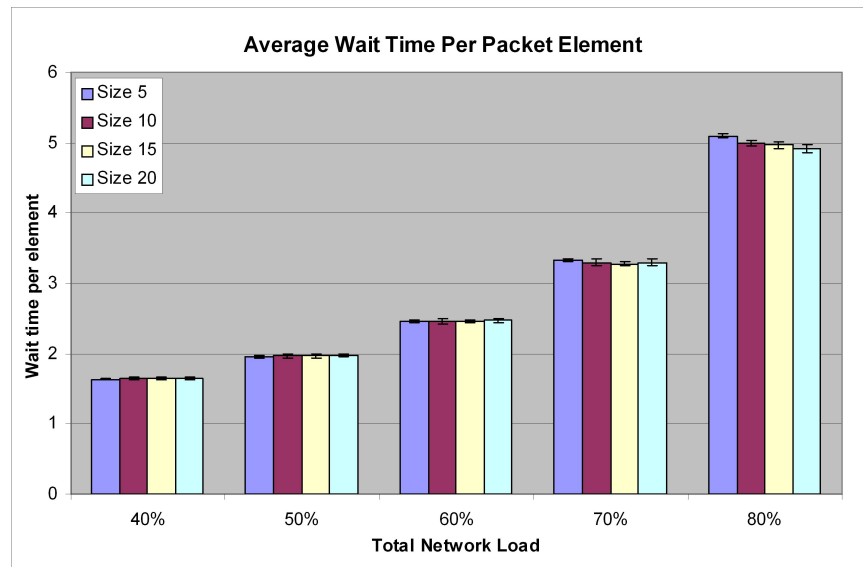


Figure 3.9 Wait time per packet element (20% voice)

manner among all nodes and provides for easy scalability. We also note that the request and service scheme used in the RDFN protocol does not rely on timely feedback coordination to enable fairness as in the RPR protocol. Thus, the bandwidth oscillations seen in the RPR protocol are avoided in the RDFN protocol. By utilizing the advantages of SONET and Ethernet we have created a metropolitan area network that meets traffic demands and achieves fairness, high quality of service, and survivability.

## CHAPTER 4. Light-Trails: Architecture and Fairness

In the previous chapters we presented various metropolitan area media access control protocols that rely on O-E-O conversion to provide traffic grooming. Although O-E-O conversion at intermediate nodes enables simple switching, routing and multiplexing of low rate data streams onto high rate optical links, features still unavailable in the optical domain, current electronic processing technology is very expensive at data rates above 10 Gbps. In addition, electronic processing at intermediate nodes throughout the network precludes network transparency. That is, each node in the network must be aware of the underlying transmission bit-rate and protocol in order to make the appropriate switching and routing decisions. This opacity limits the freedom that network engineers have in deploying various transport protocols such as Fiber Channel, Infiniband, RapidIO, etc.

As mentioned in Chapter 1, Wavelength Division Multiplexing technologies using Optical Add/Drop Multiplexers and Optical Crossconnects have been introduced to enable all optical transport networks. However, deployment of such technology in the metro arena remains limited due to the high cost and immaturity of the required components. In addition, traditional circuit switched WDM networks are provisioned for peak rate traffic due to lack of buffering capabilities in the optical domain and hence may be severely underutilized. Network utilization can be improved by equipping nodes with electronic grooming (e-grooming) capabilities that allow efficient packing of low rate streams onto high rate channels. However, grooming brings along with it concerns related to complexity, scalability, delay and transparency. Traffic engineering and statistical multiplexing gains are achievable in optical packet switched networks but high speed optical switches, scalable packet parsing mechanisms and fast and large random access units have not been realizable for large scale commercial deployment. Burst

switching provides a hybrid approach between circuit and packet switched paradigms, but the requirement of low switch reconfiguration times as compared with the burst duration leads to significant challenges in optical switch design.

As a solution to providing high network resource utilization, seamless scalability and network transparency, we discuss light-trail (LT) technology [37]. The goal of light-trails is to eliminate O-E-O conversion, minimize active switching, maximize wavelength utilization, and offer protocol and bit-rate transparency to address the growing demands placed on WDM networks. Light-trail technology is a physical layer architecture that combines commercially available optical components to allow multiple nodes along a lightpath to participate in time multiplexed communication without the need for burst or packet level switch reconfiguration.

In the following Chapter we develop light-trails as a novel and amenable control and management solution to address IP-centric communication at the optical layer. We first define the light-trail concept and outline existing research into light-trail technology. As a shared medium architecture light-trails must employ media access controls to avoid collision and facilitate bandwidth arbitration. To this end, we present existing light-trail and light-bus protocols [37, 37, 82, 9]. We note that these protocols suffer from unfairness and require packet fragmentation, in the case of the light-trail protocol. The greedy Pi-persistent LT protocol is introduced as a slotted TDM protocol to address the packet fragmentation issue, however does not consider fairness. As a complete MAC solution we introduce the Token LT and LT fair access protocols and evaluate their performance in comparison to a modified version of the Pi-persistent protocol as discussed in Chapter 2. It is shown that the LT-FA protocol satisfies the bandwidth budget fairness model and provides the most efficient mechanism for LT access control. The goal of light-trails and our access control solution is to combine commercially available components with emerging network technologies to provide a transparent, reliable and highly scalable communication network.

## 4.1 Introduction

As the primary motivation for light-trail technology is to avoid costly O-E-O conversion and electronic switching the LT solution is realized using passive optical components. The LT node architecture enables the dynamic opening of an optical path or trail between any chosen source and destination. In addition, TDM techniques allow all stations en-route from the trail head node to end node to access the trail without the need for switch reconfiguration or electronic conversion. This is contrary to point-to-point, lightpath or burst level path provisioning in conventional architectures where communication is restricted to source and destination and traffic grooming is only available at terminal stations. By allowing intermediate nodes access to an already existing lightpath, connections are not constantly being setup and torn down, but rather exist for as long as they are being used by any of the nodes along the trail.

## 4.2 Light-Trail Architecture

Light-trails are introduced in [37] as an architecture and protocol that allows the opening of an optical path between any chosen source and destination nodes while allowing optical access to all nodes en route to the destination. A light-trail is similar to lightpath in that it requires the establishment of a unidirectional optical circuit between the head and end nodes. The key difference is that intermediate nodes can also receive and transmit data on the same channel in a time multiplexed manner without electronic interference of transit traffic. The light-trail architecture combines inexpensive optical splitters, combiners and shutters to enable an all-optical trail from source to destination.

Figure 4.1 shows a typical multi-wavelength light-trail node. For each wavelength, a light-trail access unit (LAU) that consists of a splitter, shutter, combiner and an optional power compensator (typically, a semiconductor optical amplifier (SOA)) are provisioned. Although WDM can be supported, the remainder of this Chapter only considers the single wavelength solution.

Figure 4.2 shows a four node uni-directional light-trail (which is a slight variant of the system suggested in [39]). At each node, the signal passes through the LAU. A signal sourced



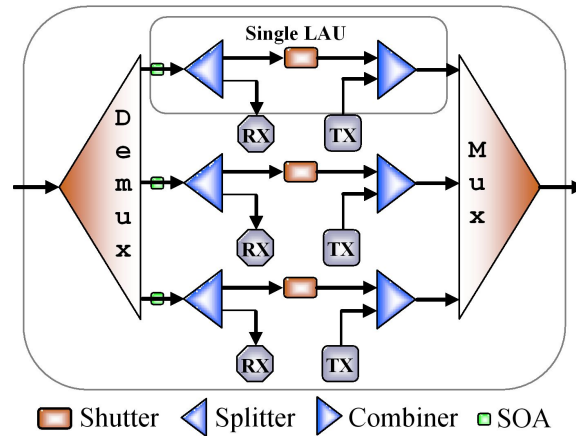


Figure 4.1 Multiple wavelength light-trail node featuring multiple Light-trail Access Units (LAU)

by a node traverses all nodes downstream to it on the trail. At each LAU splitter, a sufficient amount of optical power is tapped from the incoming signal for local processing. The local node may choose to ignore the packet or forward it to higher layer if the packet is destined for it. After the signal passes through the splitter, the remaining signal is sent through the optical shutter which is actuated through the use of an out of band communication channel to establish the light-trail. A simple optical attenuator such as the magneto-optic switch based on Faraday effect described in [6] can be configured to either block a selected wavelength or let the signal pass through.

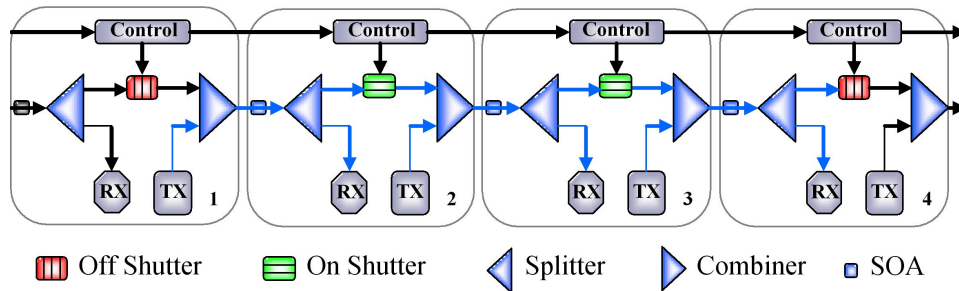


Figure 4.2 Four node light-trail - optical connectivity path is displayed in blue

If the node is the last or the first station on the trail, the shutter is configured to block the selected wavelength. For all intermediate nodes, the shutter lets the signal pass through.

This configuration isolates the particular wavelength from the rest of the network and enables spatial reuse of wavelengths. Lastly, if the signal is not blocked by the shutter, the transit signal is coupled with the local transmitter signal at the combiner. The coupler is used by the local transmitter to introduce its signal when channel access is required.

The node architecture in combination with the shutter configurations results in a single light-trail which is defined as a simple path in a network graph that can support multiple requests subject to the following constraints

- Containment Constraint: The light-trail can support any request  $(i,j)$  if  $i,j \in LT$  and  $j$  is downstream of  $i$  on  $LT$ .
- Capacity Constraint: The aggregate traffic supported by all connections in  $LT$  is at most the capacity of a single wavelength.

The bandwidth on demand feature of light-trail networks helps the network handle bursty and highly variable traffic in a more efficient way as opposed to conventional circuit switched networks. The key point to note in the architecture is that the optical shutters are not switched on a per packet basis but configured only on a longer time scale as opposed to burst or packet switched networks. This prevents light-trails from being constrained by optical switching technologies. Despite the absence of dynamic switching, by sharing the medium statistically, by expanding trails to meet new demands and by tearing down unused trails in a distributed manner, light-trails are able to provide the granularity required for data-centric communication.

In some sense, the  $LT$  concept may appear to be similar to the distributed queue dual bus (DQDB) architecture specified in IEEE 802.6 and discussed in Chapter 2. However, it is important to note two key differences between  $LT$ s and DQDB architectures. DQDB is bidirectional, whereas  $LT$  is not. The  $LT$ s unidirectional nature is best suited to meet the prevalent asymmetric traffic patterns of the Internet and to give the designer the flexibility to establish only those trails that optimally meet the traffic requirements. The second key difference is that the DQDB is a physical topology, whereas the collection of  $LT$ s defines a virtual topology that may be embedded over a mesh or ring physical network. Thus,  $LT$ s

lend themselves naturally to be a more general framework to cater to the needs of IP-centric applications.

#### 4.2.1 Light-trail Literature

Since the introduction of the light-trail concept in 2003 [37] much work has been presented in the literature to enhance the features and characteristics of light-trails. Much of this research is focused primarily on two areas of light-trail communication. The first deals with light-trail design, organization and fault tolerance as a overlaid virtual topology and the second focuses on communication within an established light-trail.

The question of how individual light-trails should be established and provisioned given a physical topology and traffic requirements has been discussed in [30] using a two step approach; preprocessing the traffic matrix and applying an ILP to optimize light-trail establishment. Heuristics for LT routing and wavelength assignment are given in [7]. LT efficiency is compared to OBS and lightpath routed networks in [31]. Other LT design optimization strategies are presented in [40, 8, 5, 91]. Suggestions for a mesh implementation of light-trail are found in [38] and bi-directional LT's are introduced in [55] and [41] in the context of Resilient Packet Rings. In addition, LT survivability, protection and restoration are discussed in [42] and [47].

As light-trails operate on a shared medium, media access controls must be considered to avoid collisions and address fairness and QoS. The remainder of this chapter is devoted to providing a comprehensive discussion of such bandwidth arbitration techniques for the unidirectional light-trail found in [37, 82, 9, 81]. Media Access control in bi-directional light-trail networks is considered in [41], however, due to the added complexity of the bi-directional solution they are not discussed in this Chapter.

One final note on the literature and work with light-trails is that although the "light-trail" architecture was introduced in 2003 similar architectures were considered in 2001 with the Dual Bus Optical Ring (DBORN) in [76, 13] and the RING Optical network (RINGO) in [34, 18]. Although the architectures suggested in these works were strictly devoted to ring networks, the node structure employing passive components to support tap and continue functionality is

very similar to that of light-trails.

#### 4.2.2 Light-trail Network

The L-Bone network is an example network that illustrates how LTs can be used as a complete mesh network solution. Light-trails in the L-bone network are established and torn down during the LT design phase and are not configured on a packet-by-packet basis. This is done through the use of an out-of-band control channel that is dropped and processed at each node. The signaling channel carries information pertaining to the setup, teardown, and dimensioning of LTs, and is responsible for provisioning “optical connections,” ranging in duration from IP bursts to virtual circuits.

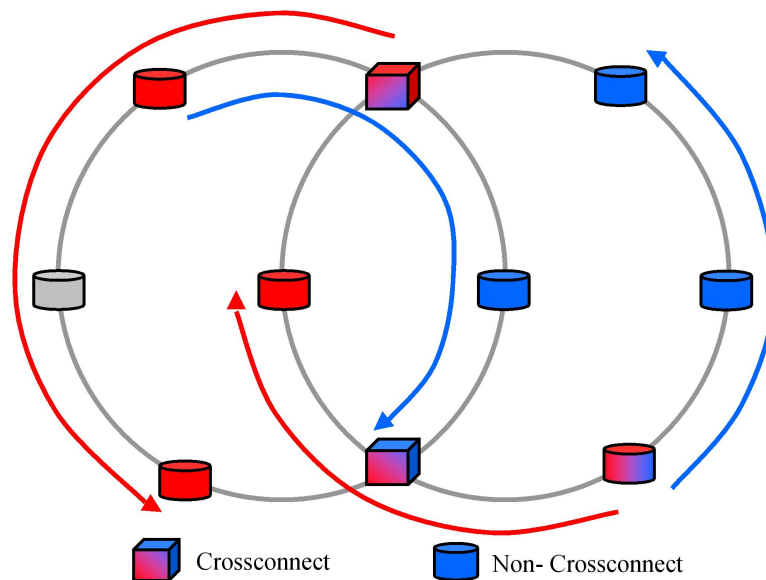


Figure 4.3 The L-Bone network is an example of how light-trails can be configured as a complete network solution. The node color indicates which light-trail a node is active on. A gray shade indicates that a node is not active on the trail.

Figure 4.3 shows an L-bone network. Note that the signal may not be tapped on every intermediate node of the trail. If the node is actively communicating on a trail, the incoming signal is demultiplexed, switched to the LAU, and reinserted back into an OXC to be multi-

plexed onto an output fiber. However, if the node is not involved with communication on this trail, the OXC lets the signal bypass the LAU and switches it directly onto the output fiber. For a detailed analysis of OXC architectures in light-trail networks, readers are referred to the study in [7].

The L-Bone network offers full optical connectivity that can share the wavelength among all nodes in the time domain, leading to dynamic sub-wavelength allocation and multicasting. It is important to note that despite the absence of dynamic switching within LTs, the granularity obtained is sufficient to provide IP-centric communication bursts in addition to full lightpath connections.

### 4.2.3 Light-trail Metro Architectures

Having discussed the requirement of WDM in metro networks and the rationale behind trail switching in WDM networks, we see how light-trails can be designed for metro networks. One such solution is shown in Figure 4.4. In this example we propose a light-trail ring or bus topology for metro edge networks. A variety of devices like GigE routers, ESCON main frames, Fiber Channel based SAN switches, ATM and telephony switches can connect to the subscriber access points on the LT ring/bus. In this example we consider two unidirectional trails being set up as shown in the figure. The downstream trail is used for the hub (central office) to transmit data to all the other nodes (access points) on the ring and the upstream trail is used for the access points to transmit data to the hub. While the downstream trail has only one source, the upstream trail has multiple sources and hence needs a medium access control for upstream communication.

The demands in metro core, however, are more meshed and voluminous and hence we propose either a ring or a mesh DWDM architecture. Figure 4.4 illustrates an example metro core network configured in the form of a mesh. Nodes in the core are connected to the long-haul network but is not shown in the figure. If node N1 in Figure 4.4 has data to be sent to node N2, the data is first sent on the upstream trail to the central office CO1 and then routed via the metro core which then reaches node N2 via the downstream trail originated by CO2.

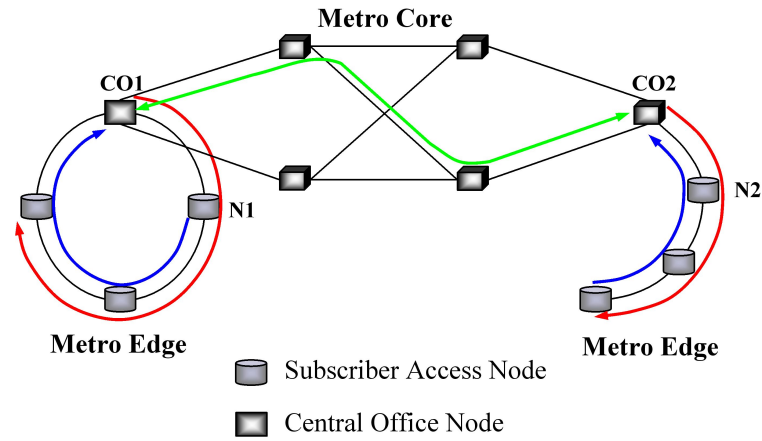


Figure 4.4 A metro network employing WDM light-trails

Because LTs operate on a shared medium, it is important that the design of a complete LT solution must include media access control to govern bandwidth arbitration. The next section discusses three LT MAC protocols designed to prevent collisions and provide bandwidth allocation policies.

### 4.3 Light-Trail Medium Access Controls

A number of priority based access protocols have been proposed for unidirectional bus topologies such as the Pi-persistent protocol as discussed in Chapter 2 and the Full Utilization and Fair (FUFA) in [23]. The following sections describe three medium access control methods specifically designed to address access control in light-trails.

#### 4.3.1 Light-trail MAC

A simple MAC protocol based on carrier sensing is proposed for light-trails in [37]. When a node, say X, wants to transmit data, it senses the carrier for upstream activity. If the channel is free, X sends a beacon signal downstream indicating that it has a packet to transmit. If the channel is busy, X waits until the channel is free to send the beacon signal. After some predetermined offset time, called the guard band, X transmits its data. This guard band

provides sufficient time for further downstream nodes to disable their transmitter so as not to interfere with X's packet transmission. Node X continues to sense the channel while its packet is being transmitted. During transmission, X may hear a beacon signal from an upstream node, say Y, which has data to transmit. Upon hearing this, X terminates its transmission and lets Y's packet pass through. X's truncated packet is discarded by the receiver(s) since it fails the link level error checks. Thus, by sending a beacon signal before transmitting a packet and by always giving priority to the upstream nodes, the protocol successfully resolves medium access contentions.

The key design parameter that affects protocol performance is the guard band gap. When the decision to stop transmission is made, the feedback control happens through a microcontroller and hence the delay is large due to the electronic processing overhead. In addition, the laser source must be completely deactivated which can take a significant amount of time depending upon the characteristics of the optical components used. The guard band should be set large enough to pull out of transmission after sensing the beacon signal and small enough to avoid significant overhead.

### 4.3.2 Light-bus MAC

A new architecture called the light-bus is proposed in [9]. The key difference is that the light-bus switch architecture includes a fiber delay loop at every node. The delay is statically set to the time needed to transmit a maximum size packet plus a sufficient guard band. When a node has data to transmit, it first checks the delay line for activity. If the delay line is free, the node transmits its packet. If the delay line is busy, the node waits until the delay line becomes free and then starts transmission. In the middle of a transmission, if an upstream packet reaches this node, it is buffered in the delay line and the current transmission is completed before the upstream packet exits the delay line.

The main difference between the light-trail and light-bus protocol is that in the transmitter need not be concerned with aborting and retransmitting packets. In addition, packet fragmentation is not required as a station will always be able to finish its current transmission before

being interrupted by upstream packets. Thus, the light-bus approach does not lead to wasted transmissions. The main drawback in the light-bus protocol is that the addition of fiber loops has an adverse effect on queuing and propagation delays as discussed in [9].

### 4.3.3 Greedy Pi-persistent Light-trail MAC

As noted earlier, the contention resolution solution in the original light-trail MAC is undesirable in that stations may be interrupted in the middle of packet necessitating retransmission. In order to reduce the probability that a downstream packet will be interrupted midstream, slotted TDM can be used. That is, a beacon signal can only be sent at the beginning of a time slot. Furthermore, a node may only begin transmission of a queued packet if it will finish before the end of the current time slot. Thus, packet interruption and retransmission is avoided at the cost of reduced network utilization due to a slot not being completely filled. Network utilization can be improved using traditional fragmentation, however, the increased overhead and added hardware complexity makes it prohibitive. Another way in which utilization can be improved is for each station to selectively choose packets from their queue to statistically fill a time slot regardless of the individual packet arrival time. [56] proposes such a scheme in the context of the Ethernet Passive Optical Network (EPON) protocol. However, the underlying bin-packing problem again creates additional overhead and may lead to undesirable packet delays and is not explored in this dissertation.

This modification to the original light-trail protocol can be likened to that of the Pi-persistent protocol, as discussed in Chapter 2, with each stations persistence value,  $p_i$ , equal to 1. This special case of the Pi-persistent protocol is considered to be a greedy solution. That is, stations are only restricted from transmission if a particular slot is occupied by data from upstream stations. Since no other restrictions are implemented, a station at the head of the trail can monopolize the bus as downstream stations must yield to upstream transmissions.

As we look at the performance evaluation of the light-trail, light-bus and greedy Pi-persistent MACs in the next section we will see that these protocols do indeed provide collision protection but do not address fairness. In fact, none of the fairness criterion such as equal



resource utilization, equal performance in delay and throughput, or equal blocking probability are satisfied. In addition, these basic light-trail access control techniques cannot be identified with any of the fairness models as discussed in Chapter 2. Thus, following a discussion of their performance we present the Token LT and the light-trail fair access (LT-FA) MAC to address fairness and quality of service in light-trails.

#### 4.3.4 Performance Evaluation - Light-Trail and Light-Bus

Performance analysis of the light-trail and light-bus protocols was carried out by Sirini Balasubramanian and presented in [82], the results are reiterated here for completeness. The simulations for various traffic loads is done using discrete event simulation techniques. Packet arrivals are based upon a Poisson distribution and uniform service times are assumed for the simulations on a 10 Gbps system. We considered five-node light-trails and light-buses. In the graphs, the x-axis plots the total offered load expressed as a fraction of the capacity of the wavelength and The y axis plots the log of the packet delay normalized to the maximum sized burst (and scaled by a factor of 1000). For the five node system, traffic density for nodes 1 through 4 are split in the ratio 4:3:2:1, respectively.

Our observation is that both the light-trail and the light-bus exhibit acceptable performance until approximately 70% of full capacity at which point average packet delay increases exponentially particularly due to increased interruptions, retransmissions and head-end trail monopolization. Of primary notability in the graphs of Figures 4.5 and 4.6 is the comparison of average delay encountered by a downstream nodes' packets compared to that of a packet from an upstream node. As expected, the light-trail and light-bus MAC, without fairness provisions, queuing delay increases considerably for further downstream stations. For results shown in Figures 4.5 and 4.6, the packet sizes are uniformly distributed between 500 and 1500 bytes. We notice that the performance of both the light-trial and light-bus are quite similar, however, it is noted that when packet sizes are small, the guard band required by the light-trail protocol is of the order of the packet transmission time and the light-bus protocol is able to outperform the light-trail protocol.

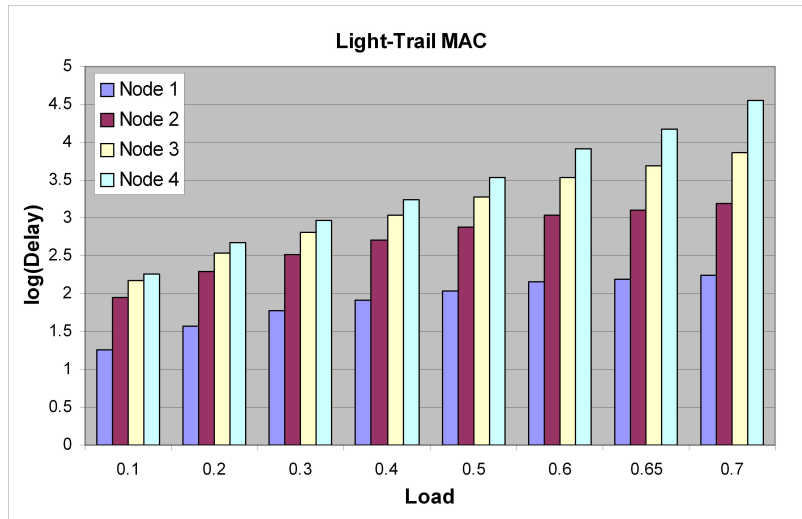


Figure 4.5 Average queuing delay Vs load for a 5 node LT using the light-trail MAC

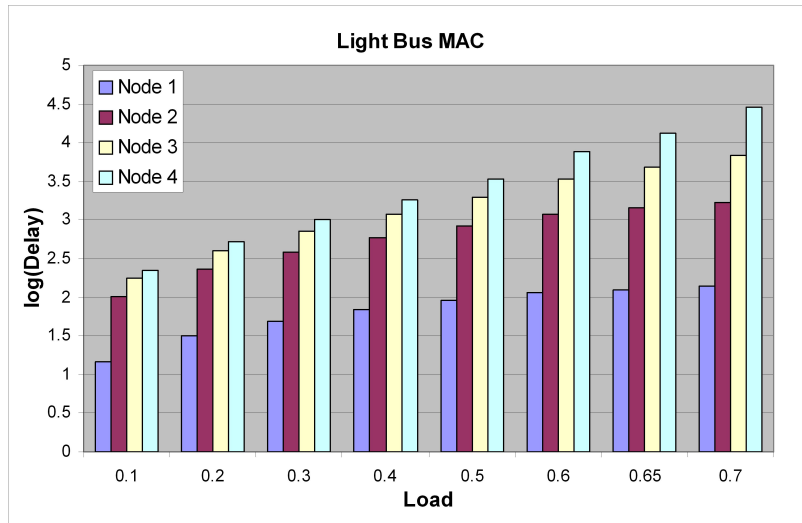


Figure 4.6 Average queuing delay Vs load for a 5 node light-bus using the light-bus MAC

#### 4.3.5 Performance Evaluation - Greedy Pi-persistent Light-trail

Performance results of the greedy Pi-persistent Light-trail protocol showed similar results as the aforementioned protocols. As before, discrete event simulations we used to arrive at the results shown in Figure 4.7. Inter arrival times were again governed using a Poisson distribution. For simplicity, packets are of static size and equal to the time slot duration. Thus, we did not take into consideration throughput degradation due to underutilized time slots. The results of Figure 4.7 illustrate the differences in average queue hold times (in number of slots) for the various stations of a 6 node light-trail. As can be seen from the charts the results showed similar trends as the former access control solutions as expected. Nodes at the tail end of the trail see higher delays than those at the head. In addition, Figure 4.8 depicts the maximum hold time seen at each node when the simulation is run for 2 million packets.

Although the three protocols discussed above are effective in preventing collisions, these schemes do not provide fairness among nodes and do not guarantee a bound on access delay, as can be seen in the performance analysis.

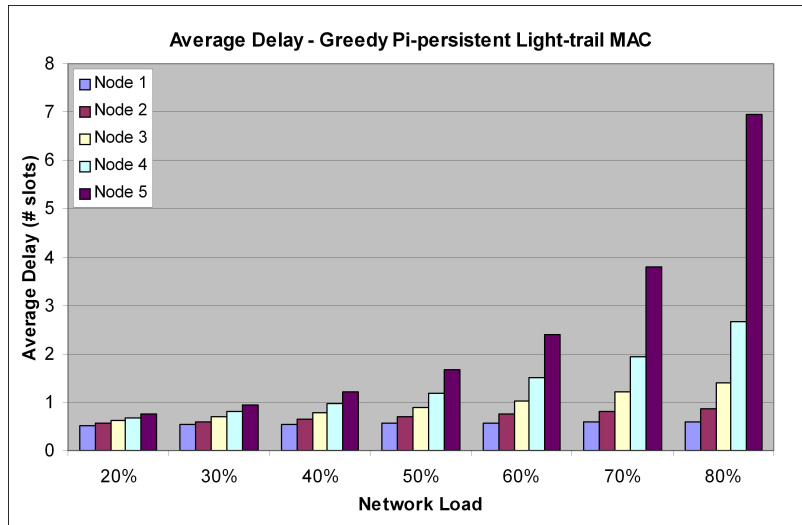


Figure 4.7 Average queuing delay Vs load for a 5 node light-trail using the greedy Pi-persistent protocol

Another factor that must be considered is the setup time required at each station in the light-trail before transmission can take place. Due to the shared optical medium characteristic of the light-trail architecture a requirement for collision free operation is that all stations must disable their transmitter when they are not transmitting and then enable the laser to participate in communication. As the laser source cannot be turned on and off instantaneously the corresponding on/off time must be taken into consideration which can be on the order of  $100\mu\text{s}$  for cleaved cavity lasers. If this on/off time is on the order of a slot or packet duration it is evident that significant overhead is realized when switching from station to station. For instance, assume that slot duration in the greedy Pi-persistent protocol is  $100\mu\text{s}$ , in this case 50% of the network capacity is wasted due to setup time which is clearly undesirable.

To solve this problem we present the Token LT MAC which reduces the effect of setup time by allowing a station to transmit for multiple slot durations before relinquishing control of the trail. The Token LT MAC will serve as a basis for the LT-FA MAC in which additional considerations will be made to reduce network delay while satisfying the bandwidth budget fairness model discussed in Chapter 2. The following sections present the Token LT and LT-FA media access controls followed by a performance evaluation.

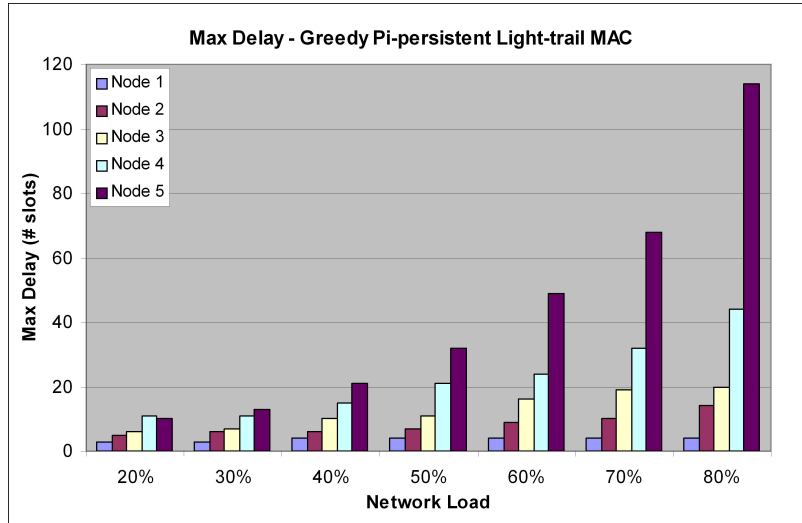


Figure 4.8 Maximum queuing delay Vs load for a 5 node light-trail using the greedy Pi-persistent protocol

#### 4.4 Token LT and Light-Trail Fair Access (LT-FA) MAC

In the following sections we present the Token LT and LT-FA protocols. These protocols are based upon the bandwidth budget fairness model discussed in Chapter 2 and use a combination of both reservation and token passing mechanisms. To facilitate reservation we utilize a key property of light-trails in that the end station is able to monitor all trail activity. Thus, all stations can communicate their bandwidth budget requirements to the end station where fair rates can be determined based upon global knowledge of the light-trail capacity requirements.

Both mechanism use a round robin scheduling technique with predetermined round intervals to bound access delay. Fairness is offered in terms of an acceptable use policy by negotiating an acceptable allocation estimate for each node; the aggregate of this allocation must be less than light-trail capacity.

##### 4.4.1 Bandwidth Budget Advertisement

At the start of each fairness interval (which consists of a number of round intervals as explained later) the head station sends a broadcast message containing information about its bandwidth budget. In addition, a maximum delay factor is sent to indicate the maximum

access delay requirements of each light trail station. The Fairness Control Message (FCM) is shown in Figure 4.9, where  $C_i$  is the connection oriented rate and  $U_i$  is the Poisson arrival rate of non deterministic traffic.

| Station # | Max Delay | $C_i$ rate | $U_i$ rate |
|-----------|-----------|------------|------------|
|-----------|-----------|------------|------------|

Figure 4.9 Fairness Control Message

When station 1 completes transmission of its FCM the next downstream station recognizes the channel being idle and begins transmission of its own FCM which also indicates its expected utilization and maximum delay. This process continues for all light trail stations.

For example, assume the usage characteristics of Table 4.1, sent as FCM's from the respective stations. Station 1 begins the fairness interval by sending the first FCM. Upon receiving this message, subsequent stations mark this round as a fairness interval and behave likewise.

Table 4.1 Example Utilization Rates

| Station # | Max Delay | $C_i$ Rate | $U_i$ Rate |
|-----------|-----------|------------|------------|
| 1         | 100       | .05        | .15        |
| 2         | 125       | .05        | .15        |
| 3         | 150       | .05        | .15        |
| 4         | 125       | .05        | .15        |

The job of the end station, after receiving all fairness messages is to determine if the aggregate of all bandwidth budgets is feasible. If so, the end station then sends a message, through the electronic control channel, to all stations accepting these rates and advertising the minimum delay (calculation for the LT-FA method below) and any spare capacity that can be used during normal operation. This spare capacity per round, identified by an *extraCounter*, is defined as the additional network capacity not requested in the FCM's and is used to aid burst mode traffic during normal operation.

Assuming the light-trail is not fully utilized, the *extraCounter* indicates the amount of

residual bandwidth remaining if all stations use only their bandwidth budget. Thus, this extra capacity can be provisioned during a single round without denying any station of their advertised bandwidth budget. The use of the *extraCounter* is explained for both the Token LT and LT-FA methods in more detail later.

It may occur in the light trail network that the advertised utilization rates for a particular fairness interval over-utilize the channel. In this case, once all fairness messages are received a weighted fair rate is calculated and reported to all stations in which 100% of the available capacity is proportionally split between all nodes and the *extraCounter* is set to zero. The current scheme is to use the proportional fairness method to reduce each stations budget. If over-utilization occurs in two consecutive fairness intervals a new light trail must be established to accommodate the additional traffic.

#### 4.4.2 Token LT - Extra Capacity Distribution

The primary difference between the Token LT and the LT-FA methods is the way in which the extra capacity is distributed. A seemingly fair method to distribute this extra capacity is based upon the proportional fairness model discussed in Chapter 2 which suggests that the extra capacity be divided proportionally among all competing stations based upon their expected utilizations. Thus, for the bandwidth budget advertisements given in Table 4.1 the proportional fairness model would allocate 5 additional slots to each of the four competing stations. This, is precisely the method used in the Token LT access control mechanism. As we will see in the performance evaluation this method does indeed facilitate fairness in terms of average queuing delay, however, the total network delay is greater than that seen with the LT-FA method, described in the next section.

#### 4.4.3 LT-FA - Extra Capacity Distribution

The LT-FA access control is best described with the use of an example. At the start of a new round station 1 begins transmission of packets from its local buffer. If additional capacity is needed, station 1 has the first opportunity to utilize the additional un-requested capacity

as indicated by the *extraCounter* (20 slots in our example). Each slot of extra capacity used by station 1 decrements the *extraCounter* by 1. If, however, station 1 does not use all of its advertised capacity, the *extraCounter* is incremented by the difference of its used and advertised rate in the hope that downstream nodes can take advantage of this additional capacity for the current round. Once station 1's transmission is complete, either because the transmission buffer is empty or transmission reaches the fair usage threshold and the *extraCounter* is expired, access is passed to the next downstream node by sending the *extraCounter* frame indicating the new *extraCounter* value. After station 1 completes transmission and station 2 receives the *extraCounter* frame, station 2 begins transmission and follows the same rules governing station 1 with regards to the extra capacity usage. Once the *extraCounter* frame reaches the last transmitting node, uninterrupted access is granted for the remainder of the round. As will be seen in the performance evaluation the LT-FA satisfies the bandwidth budget fairness model. In addition, it outperforms the Token LT in terms of total network delay.

One final consideration is how access delay bounds are guaranteed using the LT-FA access control mechanism. The following section outlines how access delay bounds are calculated.

#### 4.4.4 Round Interval Calculation for LT-FA

During a fairness interval, under normal operation, the head node begins a new round based upon the round interval as calculated and advertised from the end node. The round interval is determined from the maximum delay values sent in the FCMs. To compute the round interval, the end station examines the maximum delay factors and connection oriented utilization factors from each station and calculates a round interval that ensures no station must wait more than their maximum delay even in the worst case. This worst case access delay occurs if a heavily loaded head of trail round follows a lightly loaded head of trail round and applies only to stations 2 through N-1. A following example will explain this behavior in more detail.

The worst case access time is calculated by comparing the difference between the earliest possible access time  $T_{i_{early}}$  with the latest possible access time  $T_{i_{late}}$  for all stations. The



early round access time occurs when all previous stations with respect to station  $i$  do not have any traffic to transmit except their guaranteed service traffic. This time signifies the earliest time node  $i$  will gain access to the trail, in number of slots, with respect to the start of a new round. Calculation of  $T_{i_{early}}$  is given in eq. 4.1.

$$T_{i_{early}} = \sum_{j=0}^i C_j * T_{i_{round}} \quad (4.1)$$

The late round access time occurs when all previous stations with respect to station  $i$  transmit their fair share and the *extraCounter* has expired when station  $i$  begins transmission.  $T_{i_{late}}$  is calculated from 4.2.

$$T_{i_{late}} = \sum_{j=0}^i (U_j + C_j) * T_{i_{round}} + \text{extraCounter} \quad (4.2)$$

$T_{i_{round}}$  can then be calculated for station  $i$  by equating the advertised maximum delay for station  $i$  to the difference of  $T_{i_{early}}$  and  $T_{i_{late}}$  added to  $T_{i_{round}}$  and solving for  $T_{i_{round}}$ , as shown in 4.3.

$$\text{MaxDelay} = T_{i_{late}} - T_{i_{early}} + T_{i_{round}} \quad (4.3)$$

The two exceptions to this rule are for stations 1 and N-1,  $T_{i_{round}}$  for station 1 is simply  $\text{MaxDelay} + T_{i_{early}}$  and because station N-1 does not give up access to the trail to further downstream nodes, the last possible access time is the last slot of the current round. Thus,  $T_{i_{round}}$  for station N-1 is simply  $\text{MaxDelay} + T_{i_{late}}$ .

For example, assume a the allocation given by the FCMs in table 4.1. The calculated  $T_{i_{round}}$  for each station are as follows

- Station 1 -  $T_{i_{round}} = 105$
- Station 2 -  $T_{i_{round}} = 96.5$
- Station 3 -  $T_{i_{round}} = 103.45$
- Station 4 -  $T_{i_{round}} = 156.25$

In this example station 2 is the most highly restricted station and thus node 1 will begin a new round every 96 time slots from the start of the previous round. In this manner, operation is simplified because station 1 is the only station on the light trail that must keep a time reference.

#### 4.4.5 Normal Operation

Normal operation is similar for both the Token LT and the LT-FA access controls. As the light-trail master, the head node sends a beacon signal to interrupt current transmission and designate the start of a new round. During operation a nodes access is surrendered for the round when either the transmission buffer is empty or the fair access limit is reached. When access is surrendered the token is passed to the next subsequent station until the token arrives at the last transmitting station where access is granted until the start of a new round. The following section provides a performance evaluation of both access control mechanisms.

### 4.5 Token LT and LT-FA Performance Evaluation

In order to demonstrate the operation of our light trail media access control, a network simulator is developed using C++. The following sections describe the functions and methods used to develop such a simulator.

#### 4.5.1 Traffic Generation

To simulate the light trail operation a request generation function is created to control the frequency of bandwidth requests. Three traffic types are created for simulation purposes, guaranteed service, Poisson distributed messages of size one, and burst mode traffic streams with various burst sizes. For simplicity, the simulation assumes a round interval of 100 slots.

The provisioning for guaranteed service traffic is done at the beginning of each stations access period for each round, i.e. if station  $i$  requires 5% capacity for guaranteed service, 5 slots ( $5\% * 100\text{slots}/\text{round}$ ) are allocated at the start of station  $i$ 's channel access each round. These slots are not available for service of other traffic types.

The Poisson and burst mode traffic requests are generated in a somewhat different fashion. Each station maintains a single traffic queue to store all requests which are generated using a randomized process. We used a Poisson distribution to govern the arrival rate of all messages. Based upon the Poisson traffic load, the average arrival rate,  $\lambda p_i$ , is specified. In a similar fashion,  $\lambda b_i$ , the average burst rate is also assigned. Generic requests are then generated based on the governing inter-arrival time. The requests are then associated with their respective station queue based on the randomly identified source.

For burst mode traffic, the burst size (in number of slots) of each message is generated using an exponential distribution with average burst size,  $B_{size}$ . Each message follows the following generation characteristics.

- The arrival time of the next bandwidth request for both packet and burst traffic is governed by Poisson process.
- The burst size is exponentially distributed with a mean value of  $B_{size}$ .
- The source of the traffic is determined with a uniform random variable and the request is associated with the respective queue to fulfill the advertised distribution indicated by  $U_i$ .

Due to the connectionless property of the Light-trails a setup time of  $100\mu s$  is required to activate the laser source before transmission can begin. This setup time,  $S_i$ , is modeled similar to the method used for guaranteed service traffic. To simplify design, 1 slot time was set to  $100\mu s$  and thus setup time requires 1 slot each time a station begins transmission. Thus, the combined total of all three traffic types governs the total network utilization. The setup time is not included in the total network utilization and is considered overhead. In the simulation results presented below for a 6 node light-trail with 5 transmitting nodes and a round time of 100 slots the overhead due to setup time is 5% of the total network capacity, that is, 5 slots per 100 are wasted for setup time.

### 4.5.2 Token LT and LT-FA Operation

Simulation begins with a call to the request generation function where all requests are generated and assigned to their respective station queues. A token is established to indicate the current transmitting station. When a station begins transmission the first  $S_i$  and  $C_i$  slots are provisioned for setup time and guaranteed service traffic respectively. As network time progresses a station examines the head of its queue to determine if a request has arrived before the current network time. If a request has arrived by the start of each slot the packet is dequeued and network metrics are calculated.

Once all guaranteed service units are transmitted, operation continues to empty the station queue during each successive slot. If the station queue becomes empty during operation, or if  $S_i + C_i + U_i$  slots have been transmitted and the proportional allotment or the *extraCounter* has expired for the Token LT or LT-FA respectively, the token is advanced to the subsequent station where setup time and guaranteed service transmission begins for that station. When the token reaches station N-1, access is granted for the remainder of the current round.

### 4.5.3 Performance Analysis

All simulation results shown below operate using the same allocation and distribution of 10 million requests. In order to allow for proper network loading the first and last 10% of requests are not considered when calculating network metrics.

Metrics are collected for average queue hold time and maximum access delay for each station. Queue hold time is calculated when a request leaves the respective station queue by examining the current network time and comparing it to the time the request entered the station queue. Maximum access delay for station  $i$  is computed by noting the time between station  $i$ 's subsequent media accesses.

#### 4.5.3.1 Token LT Performance

Figure 4.10 shows the queuing delay experienced by each of the 5 transmitting nodes of the simulated network using the Token LT access method. In all simulations traffic is uniformly

distributed among all stations. The values for the 5 series in Figure 4.10 show the effect increasing Poisson traffic from 10 to 60 percent has on the average queue hold times while maintaining the burst mode and the guaranteed service traffic at a constant rate of 10 percent. The 5% allocated to setup time is not considered in the total network load, thus, network saturation is achieved at 95% utilization. The results shown are for an average burst size of 5 which gives a burst range between 1 and 85 (which varies slightly depending upon total network load).

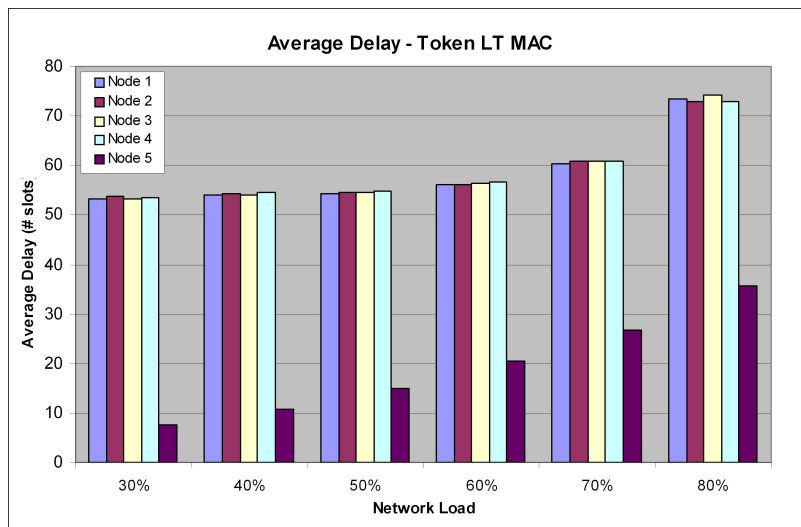


Figure 4.10 Average queuing delay Vs load for a 6 node light-trail using the Token LT protocol

As can be seen from the figure, each node exhibits similar packet access delay except for the last node on the trail. This is intuitive since the last station is allowed to transmit for the remainder of the round once the token is acquired. Although the last transmitting station has an advantage in packet access time from a network point of view this slight unfairness is acceptable because no other stations are penalized to provide this beneficial service. That is, the only reason that the last station received fairer treatment is because all upstream stations only relinquish control of the trail if either their fair share has been transmitted or their buffer is empty. As suggested by the bandwidth budget fairness model, no station is required to sacrifice their budgeted allocation for fairer treatment of downstream nodes. Thus, even

though the last station received better treatment in terms of access delay, all stations were treated fairly in terms of their bandwidth budget fairness. In addition, it is likely that the last station on the trail has less bandwidth requirements than all other stations as there is only one possible destination and thus the advantages of being the last station on the trail are reduced. For completeness, Figure 4.11 is included to illustrate that access delay bounds are guaranteed. Here again we note that the first and last stations see slight advantages with respect to maximum access delay. However, the only requirement here is that access delay is indeed bounded which is illustrated in the chart.

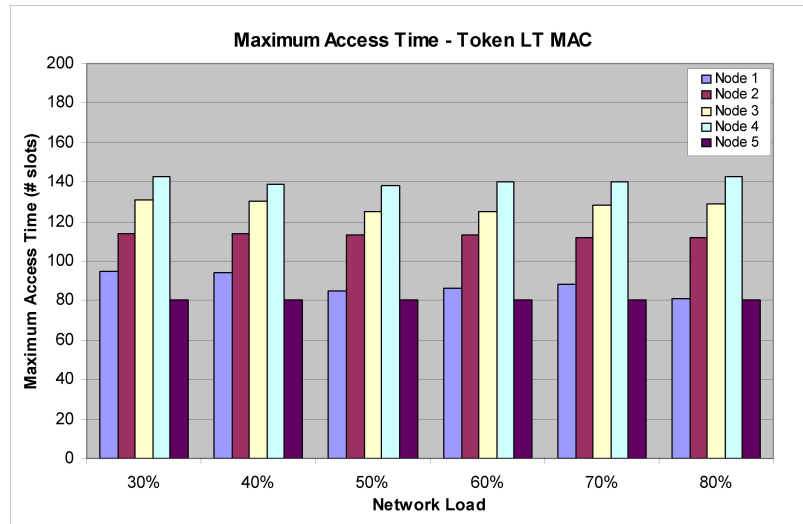


Figure 4.11 Maximum access delay Vs load for a 6 node light-trail using the Token LT protocol

Although the Token LT protocol satisfies the bandwidth budget fairness model it is noted that in the presence of highly bursty traffic, station queues may remain loaded for multiple rounds when large bursts arrive. This is because each station is only allowed a proportional fraction of the extra capacity not accounted for in the aggregate bandwidth budget. The next section presents the LT-FA protocol which, as will be seen, can more efficiently accommodate such large bursts while maintaining fairness and access delay guarantees. In addition, the LT-FA is able to improve total network delay.

#### 4.5.3.2 LT-FA Performance

Figure 4.12 shows the corresponding values for the LT-FA method using the same request generation characteristics. A comparison of the LT-FA MAC with the Token LT MAC indicates that total network delay is improved under the LT-FA scheme. That is, although a slight degradation in access delay is seen by further downstream nodes, all nodes experience slightly less average delay for identical traffic characteristics. As in the Token LT MAC the bandwidth budget fairness model is satisfied in that no station is denied at least their budgeted allocation each round if such requirements exist. Furthermore, guaranteed bounds for connection-oriented service traffic is supported while the extra capacity is shared among all nodes in a fair manner to support bursty data traffic. To illustrate this bounded service guarantees Figure 4.13 shows the maximum access delay experienced by each station.

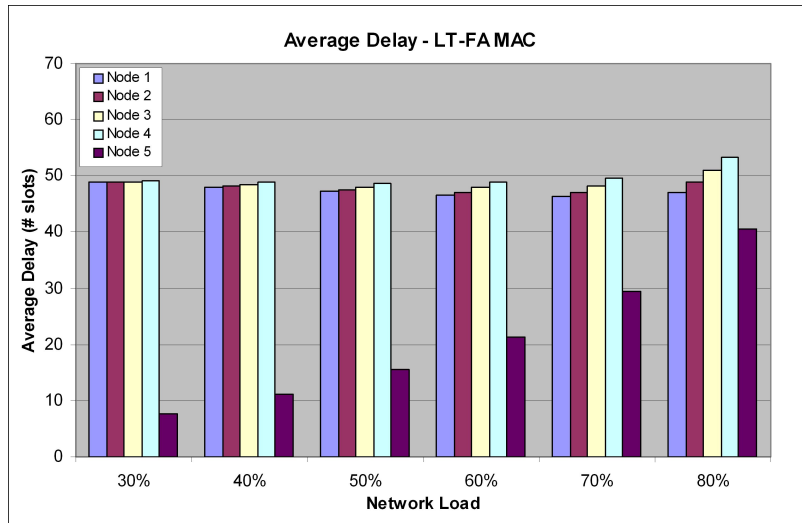


Figure 4.12 Average queuing delay Vs load for a 6 node light-trail using the LT-FA protocol



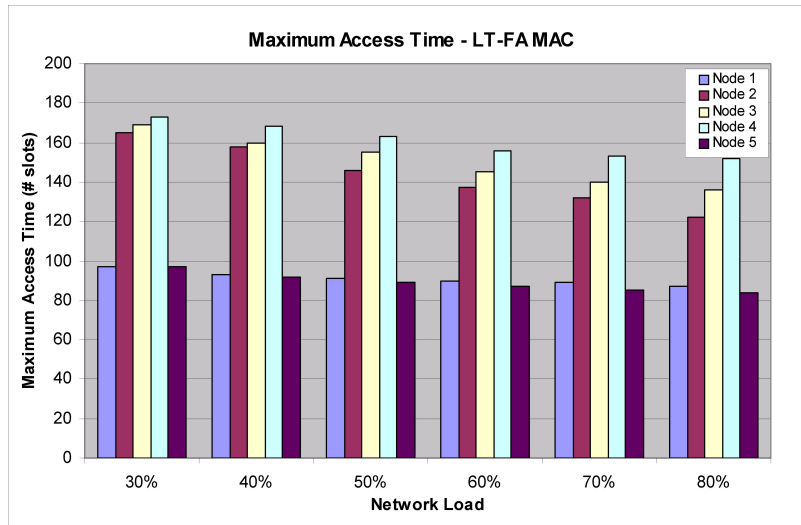


Figure 4.13 Maximum access time Vs load for a 6 node light-trail using the LT-FA protocol

We conclude that the LT-FA protocol not only satisfies the bandwidth budget fairness model but it also improves total network delay as compared to the Token LT protocol. As a final comparison we revisit the Pi-persistent protocol with a slight modification to accommodate for the large setup times inherent in the light-trail architecture.

#### 4.5.3.3 Modified Pi-persistent Protocol and its Performance

A simple modification to the Pi-persistent protocol is made to compare it with the Token LT and the LT-FA. As outlined in Chapter 2, the Pi-persistent protocol works as follows: if station  $i$  has a packet to send, it will persist to transmit the packet in the next empty slot with a unique probability,  $p_i$ , until the transmission is successful. However, as pointed out earlier, the setup time required on the light-trail network is prohibitive to the normal operation of the Pi-persistent protocol. That is, if the setup time is on the order of single slot time the overhead is near 50%. Thus, a slight modification is made.

In the above simulations of the Token LT and LT-FA protocols it is noted that the 5% of the total network capacity is devoted to setup time. To compare this with the pi-persistent protocol, we restrict a station from vieing for access to the trail until a sufficient number of packets arrive in the station queue. Then a station is allowed to compete for trail access on

regular boundaries based upon their persistence factor.

To simulate operation of the modified Pi-persistent protocol under similar traffic requirements a C++ network simulator is developed. Poisson and burst traffic is generated in the same fashion as above, however, guaranteed service and setup time is provisioned as follows. To model guaranteed service traffic, network requests are generated at regular intervals and placed at the head of the respective station queue. That is, if 2% load is specified for each of the 5 stations, a request is placed into each station queue every 50 slots. Setup time is accounted for by requiring the first slot to go unused when a station obtains access to the medium. Network metrics are collected and calculated in a similar fashion as described above.

In the first experiment we require a station to accumulate 19 packets in their queue before vying for trail access. Furthermore, a station is only allowed access the medium on a 20 slot boundary. That is, when a station queue accumulates 19 packets it waits for the next 20 slot boundary (i.e. slot 0, 20, 40 etc) and persists to transmit the packet based upon their unique probability. This represents a situation in which 5% of trail capacity is wasted for setup time, similar to that of the Token LT and LT-FA protocols. Figure 4.14 shows the average delay experienced by each node of a 6 node LT.

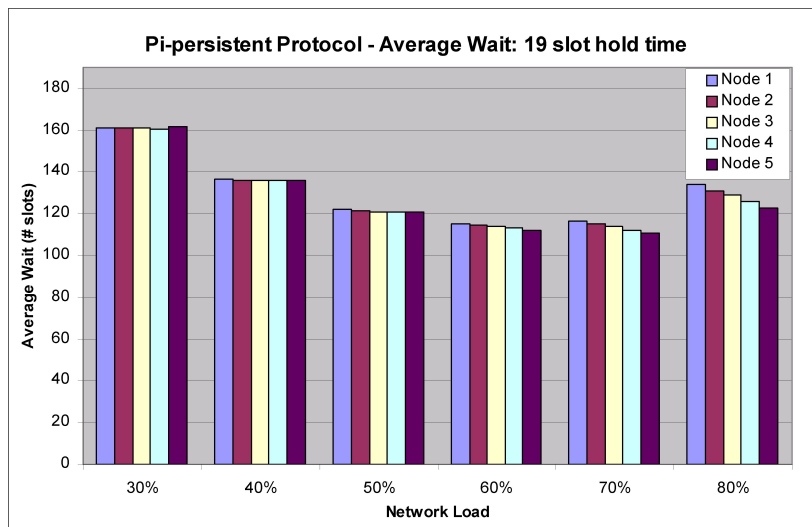


Figure 4.14 Average queuing delay Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 19 slot buffer

As can be seen from the figure, average delay experienced by each node on the trail is similar, however, the delay is significantly worse than what is seen using the LT-FA protocol. Moreover, under light network loads stations are significantly penalized as they are forced to wait until a sufficient burst arrives before transmission is allowed. In order to reduce this average delay the buffer capacity restriction is lowered to 9 slots and 4 slots. Figures 4.15 and 4.16 depict the average delay of these situations respectively.

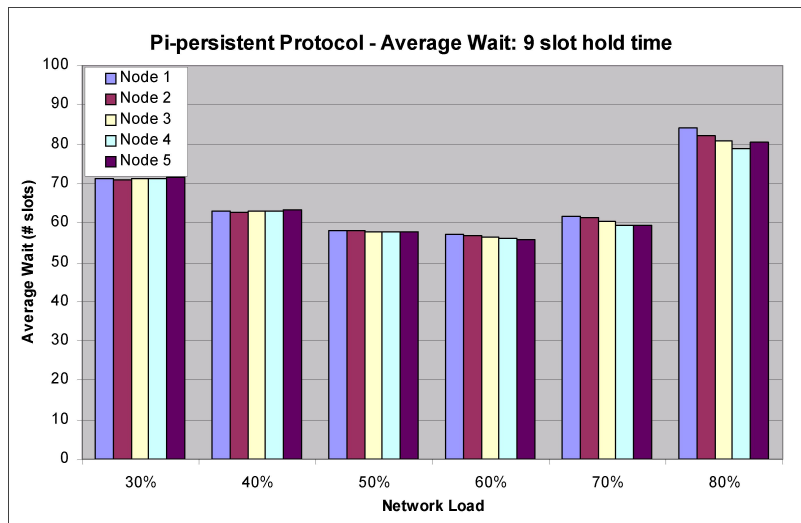


Figure 4.15 Average queuing delay Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 9 slot buffer

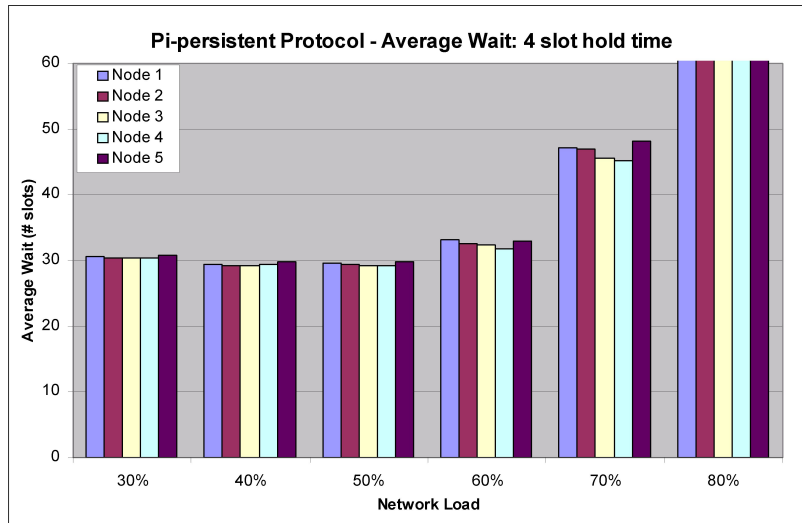


Figure 4.16 Average queuing delay Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 4 slot buffer

Again we note that each station experiences relatively similar average queuing delays in both situations. We point out that in the 9 slot buffer case no station sees improved performance over the LT-FA MAC for any network load. However, using a 4 slot wait time we notice that the modified Pi-persistent protocol is able to outperform the LT-FA in terms of average queuing delay for loads under .6C while still providing fair network service. However, at loads above .6C no improvement is seen and average delay begins to increase exponentially as the network nears saturation which is achieved at .8C due to the set up time requirement of 20%. Furthermore, we note that the probabilistic nature of the access control scheme of the Pi-persistent does not guarantee a bound on maximum access time, which is a requirement to supporting synchronous voice traffic. Figure 4.17 shows the maximum access time realized with 10 million requests under the Pi-persistent protocol with a 4 slot hold time. We note that the LT-FA protocol is able to outperform the modified Pi-persistent protocol in terms of maximum access time.

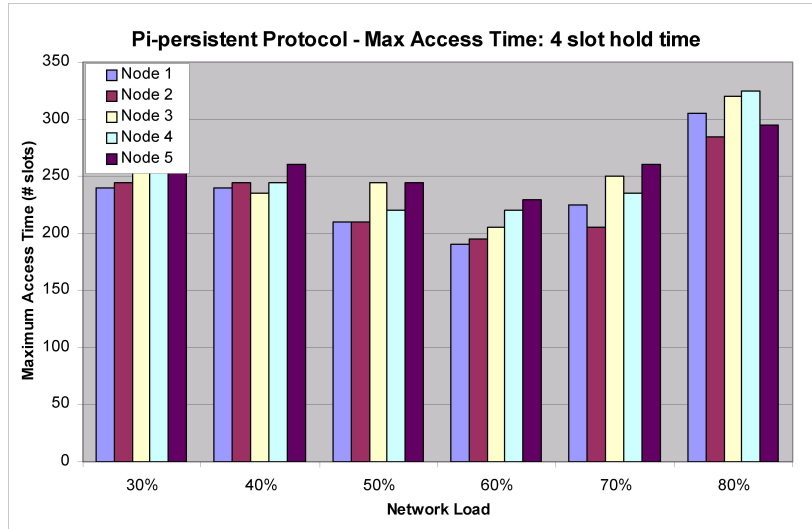


Figure 4.17 Maximum access time Vs load for a 6 node light-trail using the modified Pi-persistent protocol with a 4 slot buffer

## 4.6 Summary

In this chapter we introduced light-trail technology as a solution to wavelength grooming at the optical layer. Light-trail technology avoids costly OEO switching at intermediate nodes and offers complete transparency to signal bit-rate, format, and protocol. As a shared medium architecture light-trails must employ media access controls to avoid collision and facilitate bandwidth arbitration. To this end, we presented three medium access control protocols for light-trails that provide collision protection but do not consider fair network access. As an improvement to these light-trail MAC protocols we introduced the Token LT and light-trail Fair Access (LT-FA) MAC protocols and evaluate their performance. We illustrate how fairness is achieved and access delay guarantees are made to satisfy the bandwidth budget fairness model as introduced in Chapter 2. We also show that the LT-FA protocol is the best solution for light-trail media access control through performance comparisons with that of other unidirectional bus protocols suitable for adaptation to the light-trail architecture such as the Pi-persistent protocol.

We conclude through experimental simulation that the LT-FA protocol is able to bound maximum media access wait time while providing near equal queuing delay for all stations.

The results also show that the LT-FA protocol is able to accommodate busy traffic without significantly effecting queue hold times and maximum access delay.

## CHAPTER 5. Rapid Prototyping Platform and Light-Trail TestBed

Prototyping is defined as a process of quickly putting together a working model (a prototype) in order to test various aspects of a design, illustrate ideas or features and gather early feedback. The increasing availability and decreasing cost of prototyping platforms utilizing reconfigurable logic devices is facilitating a change in the engineering design process. Industry research and development centers are ever increasingly employing programmable devices to test new concepts at a fraction of the cost of full production development while gaining valuable insight into physical design and integration challenges. The use of reconfigurable rapid prototyping platform (RRPP) is proving to be indispensable in the engineering design process by providing a more comprehensive understanding of system level interactions.

This chapter introduces the RRPP in more detail and discusses its role in academic research and education and outlines a few of our recent projects utilizing the RRPP. We then introduce the Light-Trail Testbed built upon the RRPP. The testbed is developed to verify light-trail operation and enhance our understanding of light-trail technology. The complete testbed development is a unique implementation of light-trails and provides valuable insight into system level design challenges.

### 5.1 Introduction

As engineering solutions continue to become more complex requiring multidisciplinary approaches and complex hardware/software interactions the rapid prototyping platform has gained wide acceptance in both industry research and development and undergraduate engineering curriculum over the past decade. In addition, improved performance and enhanced capabilities combined with significant cost reduction is effecting somewhat of a paradigm shift



in engineering approaches and education which will continue to make the RRPP an invaluable engineering design and education tool. To this end, undergraduate courses at Iowa State [77], Georgia Tech [46] and many other universities [21, 15, 86] are being introduced to place emphasis on and examine system level design considerations and hardware/software co-design issues.

At the heart of this paradigm shift is the growth of system on a chip (SOC) designs and the proliferation of Programmable Logic Devices (PLDs) and Field Programmable Gate Array (FPGAs). SOC designs facilitate system level innovation by incorporating the speed of dedicated hardware with the flexibility of general purpose processors on a single chip. The reconfigurability and flexibility of FPGA's make such SOC designs easily attainable. Furthermore, rapid prototyping development boards that integrate PLD's with an array of peripheral devices are enabling the realization of complex engineering designs at a fraction of the cost of full production development.

In the following sections we provide a background on reprogrammable devices and the rapid prototyping platform and discuss recent projects utilizing the RRPP including the light-trail testbed. We suggest that the use of the RRPP in academic research is having a two fold effect. First by enhancing cooperation between research and education, undergraduate students are exposed to cutting edge academic research while at the same time gaining valuable experience using the same tools and design techniques incorporated by industry. And second, the rapid prototyping platform is making academic research projects more visible to the industrial community further fostering collaboration between industry and academia.

## 5.2 Reconfigurable Rapid Prototyping Platform

Today there are many flavors of reprogrammable logic devices on the market to suit various design requirements. A few of the most common are the Programmable Electrically Erasable Logic chips (PEELs) and Complex Programmable Logic Devices (CPLDs). However, the most flexible and advanced incarnation of the CPLDs is the Field Programmable Gate Array. And although PEELs and CPLDs continue to grow in size and functionality the remainder of this

chapter is focused on the characteristics and uses of the FPGA.

### 5.2.1 FPGA History

Since its first release to the market in the early 80's the FPGA has seen considerable advancements in size and speed along with significant cost reduction. Figure 5.1 illustrates the exponential growth and cost reduction of the Xilinx XCV4000 family FPGA. In the period from 1991-1998, the devices got 5 times faster, increased in complexity (gate count) by a factor of 20 and saw a 50x reduction in cost [1]. Today's FPGAs not only contain millions of programmable logic gates but can also be tailored for specific applications utilizing high-speed on-chip memory, dedicated multiplier blocks, gigabit per second communication modules and on-chip processors; all of which add to overall platform flexibility and speed.

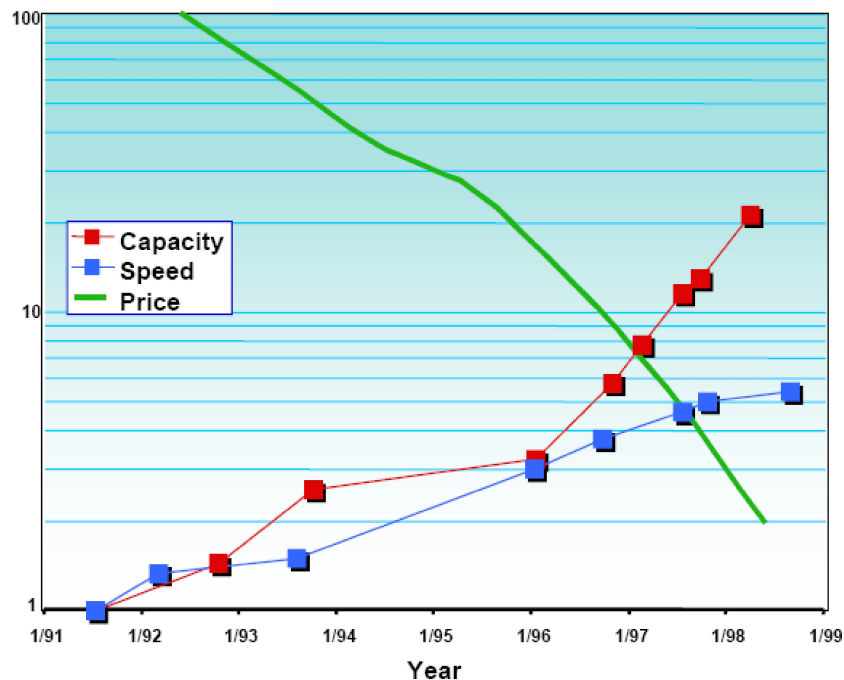


Figure 5.1 Chart showing the growth and cost reduction of the Xilinx XCV4000 family FPGA from 1991 to 1998

What makes the FPGA prototyping environment so useful is the combination of a powerful FPGA, a highly integrated development board and easy-to-use Computer Aided Design (CAD) tools. Today, many vendors provide FPGAs of various speeds, sizes and capabilities,

each employing slightly different proprietary technologies. In addition, board vendors provide many platforms that incorporate various peripheral devices that are easily integrated into the complete system design with the use of advanced CAD tools available from the FPGA vendor and simulation and modeling companies. These boards typically feature a host of common peripheral devices including, but not limited to, Universal Asynchronous Receiver Transmitters (UARTS), memory expansion modules, Universal Serial Bus (USB) interfaces, and audio/visual codecs. These board/chip combinations, associated CAD tools, and multitude of freely available Intellectual Property (IP) cores make complex hardware/software system designs easily attainable.

The following section delves further into the RRPP with a more detailed description of commercially available FPGA/development board combinations used in our research and education activities.

### 5.2.2 Xilinx Virtex II Pro Development Boards

Although many platforms are available to be tailored and optimized for specific design requirements, the primary goal of the rapid prototyping platform is to provide an easy-to-use, flexible and reprogrammable system design environment. Our research efforts at Iowa State University are based on Xilinx FPGA-based prototyping boards such as the Memec development system shown in Figure 5.2.

This board utilizes the Xilinx XC2VP30 FPGA which includes over 1 million programmable logic gates, 2 embedded microcontrollers, multigigabit serial transceivers and many other useful features such as dedicated multiplier blocks and over 200 KBytes of on-chip Block Random Access Memory (BRAM). The board also features many integrated components such as a 10/100Mbps Ethernet controller, SubMiniature version A (SMA) connectors for high-speed serial communications, peripheral memory and other useful media attachments. These features provide great flexibility and can be easily programmed using the Xilinx CAD tools and extensive library of free IP cores. Although the board used for in the original light-trail test bed are relatively expensive at \$1500, other boards with similar features such as the one manufactured

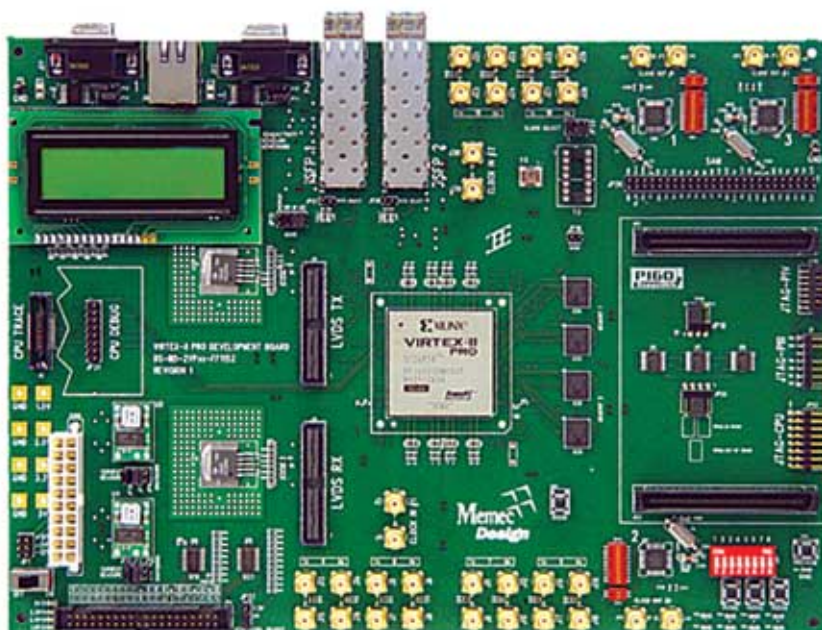


Figure 5.2 Memec development board containing the Xilinx Virtex II Pro FPGA

by Digilent Inc., shown in Figure 5.3, are available to universities at a cost of only \$300. This cost also includes a site license of Xilinx CAD tools and access to a variety of IP cores.

Over the past four years the RRPP has gained acceptance in our research efforts. In November of 2003 the Dependable Computing and Networking Lab (DCNL - Arun Somani) purchased the first Virtex II Pro development board at a cost of \$5000 to experiment with prototyping of high-speed serial communications and digital signal processing. Mike Frederick and myself began experimenting with the Xilinx CAD tools and hardware/software co-design projects. In January of 2004, we developed the High Speed Systems Engineering (HSSE - Mani Mina) Lab to incorporate characterization and measurements of high-speed communications utilizing the FPGA development boards. Shortly thereafter the Embedded Systems Lab (ESL - Dianne Rover) co-located with the HSSE to combine expertise on FPGA system level development. Since then the combined groups have collaborated on many research and senior design projects including the development of an undergraduate class now offered at ISU [11]. In addition, the members of the groups have provided consulting services to various other

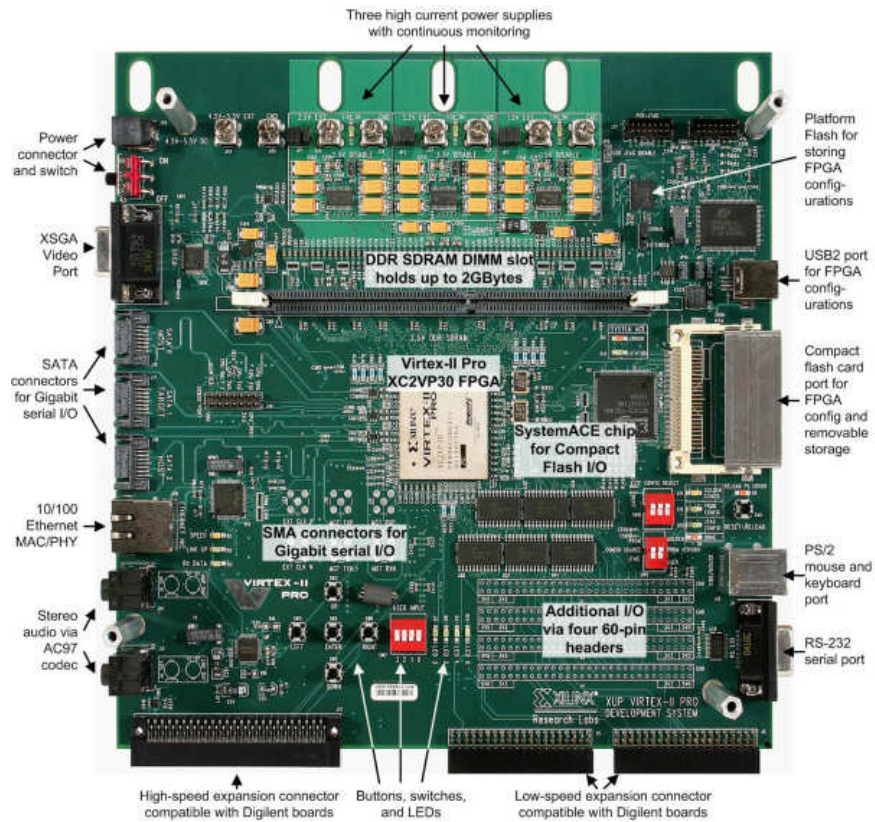


Figure 5.3 Digilent development board containing the Xilinx Virtex II Pro FPGA

research project areas in automation, parallel computing and computer architecture.

As our combined expertise in system level development grows, new projects and research directions continue to emerge. With a total of 4 Masters and 4 PhD students working with the RRPP these efforts have exposed over 20 undergraduate students to academic research initiatives as system developers over the past 3 years. Moreover, our combined labs have access to over 30 FPGA development boards, high speed electronic and optical measurement and characterization equipment and an expanding library of IP cores and CAD tools for system level development and simulation. We continue to strive to involve more and more undergraduate students in our research activities which is expanding the wealth of knowledge of system level development and exposure to academic research projects. In addition, we continue to build a multidisciplinary research environment with expertise in embedded systems, high-speed opto/electronics, parallel computing and much more.

The following sections detail two projects I have been directly involved with at Iowa State University that have been taken from a theoretical background and improved upon using the RRPP. We outline ongoing research projects that have complimented recent undergraduate engineering curriculum additions involving FPGAs and SOC design. We discuss how the FPGA is allowing researchers to promote innovations in their respective research areas while realizing the practical limitations of real world designs. The use of the FPGA platform is giving rise to more undergraduate opportunities to participate in academic research initiatives which in turn enhances project visibility to industry.

### 5.2.3 Real-Time Radon Transform Prototype

The rapid prototyping platform played an indispensable role in our work on a hardware implementation of the Radon transform as published in [32]. With the growth of interest in autonomous vehicle control, the use of digital signal processing for computer vision is gaining attention. The primary approach to enable computer vision is to extract certain desired features from images using common 2-dimensional transforms such as the Radon or Hough Transform.

The Radon transform (RT) is a widely studied algorithm used to perform image pattern extraction in fields such as computer graphics, medical imagery, and avionics. Real-time implementation of the Discrete RT (DRT) is extremely difficult due to its use of complex trigonometric functions and  $O(N^3)$  time complexity, where  $N$  is the number of pixels in the image, making its use in video applications difficult. The Approximate Discrete Radon Transform (ADRT) presented in [16] outlines a method to compute an approximation of the DRT in  $O(N^2 \lg N)$  time using up to  $N$  processors, reducing the time complexity to only  $O(\lg N)$ , and only requiring addition operations.

The goal of the project was to conduct research on how to optimize the ADRT for use in a dedicated hardware application to obtain a real-time solution i.e. 30 frames per second. Due to the high costs associated with Application Specific Integrated Circuit (ASIC) design and production, a complete ASIC solution is not feasible in an academic setting. Using a rapid



prototyping board, our research team, including 2 graduate students and 2 undergraduates, was able to produce a hardware/software design in a relatively short time with a limited budget. Although the original academic publication provided a conceptual solution for the ADRT, our implementation revealed other optimizations and properties of ADRT computation that could be exploited for improved performance.

We quickly realized that a complete system design had to not only consider theoretical runtime, but also how to optimize memory accesses and latencies. By making a small change to the original algorithm we were able to take advantage of spatial locality of the pixels stored in memory and yield a time-optimal pattern of memory accesses. Additionally, clever memory organization and manipulation allowed us to reduce memory usage resulting in nearly space-optimal memory requirements. Although the FPGA-based prototype fell slightly short of the designated goal of 30 frames per second, our researchers gained enough understanding of the system to propose how to achieve a more complete solution. In doing this project we obtained corporate visibility as well as gave two undergraduate students valuable experience used to secure industry jobs in the area of digital design and signal processing.

Although the original published ADRT algorithm identified and solved the major conceptual aspects it overlooked further optimizations that were identified in the complete embedded system design. Only through FPGA prototyping were we able to identify appropriate algorithm enhancements and interesting properties of the transform to exploit them for improved performance.

#### 5.2.4 Griffin Parallel Computing Platform

As we will see in this section, the RRPP is not only a valuable tool for rapid prototyping, but its flexibility lends itself nicely to applications in distributed computing.

With ever-increasing data processing and computing requirements, distributed computing has emerged as the focus of a growing field of research. The BlueGene experiment [49], by IBM, provides one approach to tackling the extensive computing requirements of computational biology using multiple general purpose processors. We are exploring the use of FPGAs in

distributed computing with our FPGA cluster, Griffin.

The classical approach to distributed computing is to divide and conquer a complex task using multiple general purpose processors. Although the general purpose processor approach provides a generalized and simple solution, it is often inefficient in that the processors themselves are not well suited for parallel problems that are highly regular in nature such as matrix multiplication. A potentially more efficient solution is to replace the general purpose processors with dedicated hardware modules designed for a specific task. The drawback to this approach is that development of dedicated hardware is expensive and time consuming compared to the programming of general purpose processors.

The Griffin project is a hybrid approach that combines the flexibility of general purpose processing with the speed of dedicated hardware to create a flexible, high-speed parallel processing platform. Figure 5.4 shows the 16 node FPGA cluster. Each FPGA development board is interconnected to one another via high-speed serial communication links and managed by its own computer. All computers are networked together and communicate through the Message Passing Interface (MPI) specification. Dedicated hardware modules are developed and loaded onto the FPGA boards to perform highly regular tasks and the computers connected to each board provide storage and coordination among all the FPGA boards. The Griffin hybrid approach optimizes the speed of parallel computing using a combination of both hardware and software components.

Designs developed for the Griffin platform are realized with two primary components. The first component is an application specific Intellectual Property (IP) processing element to be run on each FPGA cluster node. This processing element is implemented using a combination of both hardware logic gates and the embedded PPC microcontroller on each FPGA. The second component, called GriffinNET, provides communication capabilities between the processing IP's. Although the task of designing the processing element IP will vary for each task, the GriffinNET IP core will be used with any general design to provide high-speed communication between the processing nodes; much like MPI is used for the general purpose processing parallel platform. Designers are provided access to the GriffinNET communication system through the



use of library functions handled by the on board microcontroller.



Figure 5.4 Griffin Cluster containing 16 Dell computers connected to 16 Digilent FPGA boards

We are currently working on adapting the gene sequencing algorithm presented in [2] to the Griffin platform. The highly regular sequencing algorithm is a good fit for the dedicated hardware of the Griffin cluster. Thus, we expect to see significant performance enhancements compared to a software only approach.

The Griffin project has employed four undergraduate and two graduate students. The undergraduate students are used in the implementation of the cluster and are exposed to various disciplines such as high-speed networking, socket programming, distributed systems programming using MPI, and hardware development and simulation. Three undergraduates have been

recruited for internships by industry based specifically on their familiarity and experience with the networking, HDL development, and embedded systems programming aspects of the Griffin project.

The two projects described above illustrate the benefits to academic research that the RRPP can bring. The following sections discuss the light-trail testbed developed using the RRPP. As will be pointed out, the development of the testbed has increased our awareness of the system level design challenges of light-trail technology. As our early experiments suggest, we must consider many practical implementation issues such as, measurement and characterization of devices to guarantee system level compatibility, careful consideration of passive optical device placement to meet power budget constraints yet avoid receiver saturation and the necessity of burst mode lasers to increase network throughput and response time. In addition, various properties and design specifications of the constituent devices such as, couplers, connectors, shutters, laser on/off switches and fiber types must be considered.

The practical implementation dimension makes the light trail solution a more interesting, challenging, and rich problem. While the implementation aspects introduce limitations and added complexity, such practical problems are worth thinking, planning, and including in the comprehensive light-trail design.

### 5.3 Light-Trail Test Bed Design

As mentioned in the previous Chapters, telecommunication networks have rapidly added staggering amounts of capacity to their long haul networks at low costs per bit using WDM technologies. Concurrently, there has been a wave of new access technologies that are driving customers to demand high-speed, robust and customized data services. These dynamics have led to what is called the “metro gap” - the inability to leverage backbone capacity to create and distribute revenue generating services. This section focuses on our work to address the metro gap problem utilizing light-trail technology. As an initial step toward solving this problem, we present the light-trail testbed developed using the RRPP.

We begin our discussion of the light-trail testbed in the next section with a high level

overview of the complete testbed system which demonstrates end to end communication over a light trail backbone. Following the introduction of the system level design we present detailed component descriptions used to implement the 1.5 Gbps testbed. Included in this discussion we present design challenges encountered with the realization of our testbed.

### 5.3.1 High Level Testbed Description

Our primary goals in the development of the testbed are two fold; to demonstrate the feasibility of end to end communication over a light trail optical backbone and to address physical constraints and system level design challenges of LT technology. To this end we have developed a 4 node LT capable of multiplexing multiple client data streams over a single unidirectional optical channel entirely in the optical domain. Figure 5.5 depicts a high level view of our 4 node LT in which client workstations are connected to each LT node. Each client station is capable of communicating with their respective LT node via Ethernet, as discussed later. In addition, each LT node has the ability to transmit and receive LT frames using their respective LT access unit detailed in Chapter 4. As we will see in the next sections the development of the LT node required considerable planning and experimentation to verify the desired functionality.

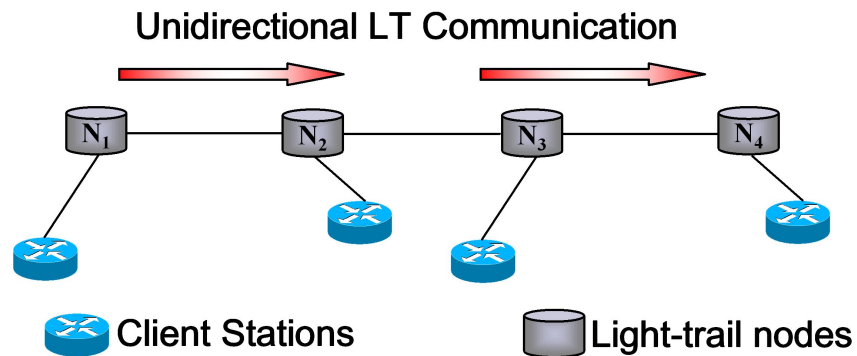


Figure 5.5 High level diagram of our 4 four node testbed showing a uni-directional LT from node  $N_1$  to  $N_4$

### 5.3.2 Complete Testbed Functional Description

Figure 5.6 illustrates a block diagram of the various components used in the design of our 4 node LT testbed. As seen in the figure, the 4 light-trail nodes are implemented on 2 Xilinx Virtex II Pro FPGA development boards. Detailed specification sheets corresponding to the FPGA and development board can be found in [88] and [4] respectively. Although the diagram illustrates that physical resources are shared between the co-located LT nodes, the hardware components are logically separated in the FPGA software application. That is, each logical node can perform all client communications and LT functions independent of one another.

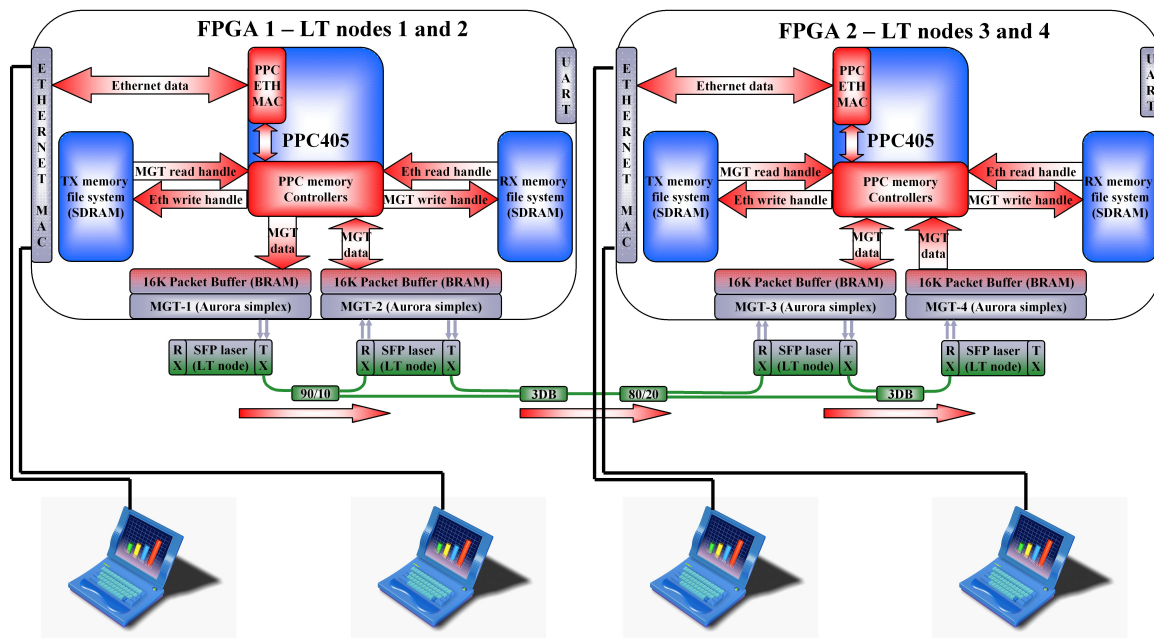


Figure 5.6 Complete functional block diagram of the 4 node LT testbed illustrating the FPGA components and optical devices used in the design

As shown in the figure, the complete testbed design combines electrical and optical components. The FPGA embedded system includes a PowerPC processing core that provides interfacing capabilities from/to the client stations with the LT backbone. In addition, the PPC performs memory management, synchronization and packet buffering capabilities to support end to end client communication. Also shown in the figure are the lasers and optical couplers used to establish the LT backbone. As we will discuss in the following sections the de-

sign of each embedded system component required careful planning and verification to ensure system level compatibility. In addition, careful measurements are performed to ensure reliable optical communication over the LT backbone.

### 5.3.3 FPGA System Design

Our 4 node testbed is realized with a combination of hardware Intellectual Property (IP) and a software application running on the embedded PPC 405 core. Figure 5.7 illustrates a conceptual block diagram of a single LT node. Custom hardware IP is coded in VHDL and C code is used to perform the software functions. A brief description of each FPGA component shown in the figure is given below. The forthcoming sections provide additional detail into the operation of the hardware IP modules and software application. We also present the challenges faced in the overall system level design.

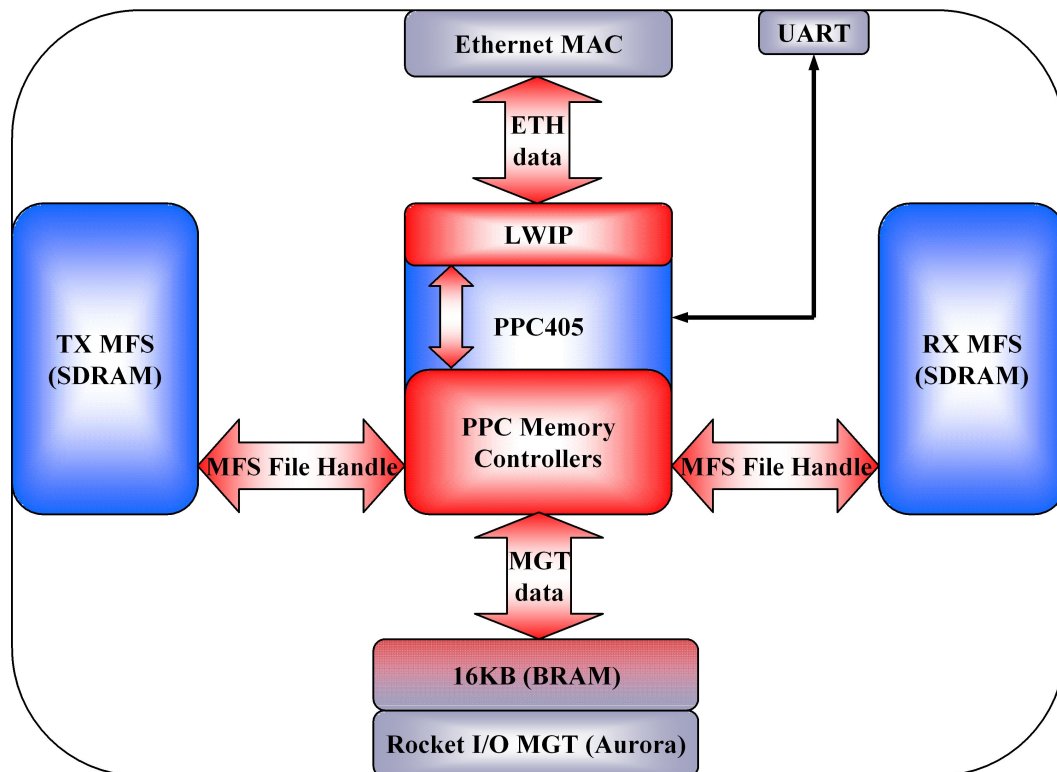


Figure 5.7 Single Light-trail node illustrating the hardware and software components used.



- **Power PC 405** - General purpose microcontroller running a software application which provides synchronization and bus communication between the various memory mapped peripherals.
- **Universal Asynchronous Receiver Transmitter (UART)** - Hardware interface to relay debug information and system status to the host computer.
- **Ethernet Media Access Controller (MAC)** - Hardware IP module enabling 10/100Mbps Ethernet communication to client stations.
- **Light Weight Internet Protocol (LWIP)** - Reduced TCP/IP software library enabling TCP socket communication with the client workstations.
- **Synchronous Dynamic RAM (SDRAM)** - 64 MB off-chip memory module to buffer and store client data.
- **Memory File System (MFS)** - Software library to interface the PPC with the SDRAM memory.
- **Block RAM** - High-speed hardware memory module with single cycle access latency to buffer LT frames.
- **ROCKET I/O MGT and Aurora Protocol** - Custom IP which provides the high-speed serial data stream to the LT backbone laser sources.

#### 5.3.3.1 PPC 405 Embedded Microcontroller

As a complete embedded system platform FPGA, the Xilinx Virtex II Pro device incorporates 2 PPC 405, 32-bit RISC processor cores for general purpose processing. Our testbed utilizes one of the available cores to provide interrupt handling, memory management, synchronization, and peripheral interface communication. The PPC operates at 300 MHz and uses a memory mapped I/O structure to communicate with the hardware peripherals via either the 64 bit Processor Local Bus (PLB) or the 32 bit On-Chip Peripheral Bus (OPB) both

operating at 100 MHz. A brief list of the processors responsibilities are given below. More detailed implementation and operation of these functions follow in subsequent sections.

- Provide TCP/IP communication capabilities to client stations using the LWIP software libraries in conjunction with the interrupt driven Ethernet MAC.
- Maintain a Memory File System to organize and store data for communication with the client workstations.
- Retrieve data from the SDRAM and load LT frames into BRAM for transmission over the LT backbone.
- Receive incoming LT frames from the BRAM and store them in the SDRAM in preparation for delivery to the end client station.
- Provide synchronization with the Multigigabit Transceivers through memory mapped handshake registers implemented within the Aurora protocol.
- Provide system status and debug information to the host computer via the UART.

### 5.3.3.2 Ethernet MAC

To support the sender/receiver client interface, the testbed utilizes an Ethernet MAC (EMAC) intellectual property hardware core provided by Xilinx [88]. The EMAC core is a fully functional 10/100 Mbps Ethernet MAC and is addressable from the PPC microcontroller via the OPB. The EMAC communicates with the PPC through an interrupt driven interface and delivers Ethernet frames to the software application. Incoming data from the EMAC, operating at 100Mbps, is retrieved from the OPB using the Light Weight Internet Protocol software functions. The combination of the Ethernet hardware core and the LWIP libraries give us the ability to communicate directly with the LT node from the client workstations using a standard socket interface.



### 5.3.3.3 Light Weight Internet Protocol (LWIP)

As mentioned, the testbed LWIP software Application Programming Interface (API) is used to interface the EMAC core with the PPC memory management application software. LWIP is an open source implementation of the TCP/IP protocol developed with the intention to reduce resource usage for embedded systems while still having full scale TCP capabilities [29]. Asynchronous network events (data received, data acknowledged, connection established etc.) are communicated to the PPC software application through interrupt callback functions. Socket connections are established and callback functions are registered to manage Ethernet communication with the clients. Incoming data from the client is presented to the software application layer using LWIP functions and stored in SDRAM memory awaiting transmission over the Light Trail. On the other hand, received LT frames, stored in the SDRAM, are transferred to the client station using LWIP functions upon request.

### 5.3.3.4 FPGA Memory Buffers

As shown in Figure 5.7 two memory structures are used in the LT node embedded system design, SDRAM and BRAM. Two memory structures are required to accommodate both high-speed and high-capacity data storage. The SDRAM provides high-capacity data buffering and storage to and from the client stations while the BRAM enables high-speed access to support LT communication. Due to capacity limitations of BRAM and speed limitations of the SDRAM both memory structures are used to compliment each other and provide complete system integration.

**SDRAM Memory File System** To provide sufficient data storage and buffering capability for Ethernet client data, each LT node maintains SDRAM memory data buffers. The SDRAM is an off chip memory peripheral module connected to the PPC's Processor Local Bus (PLB). Data is stored in the SDRAM through the use of the Memory File System (MFS) library functions which provide a simple interface to the SDRAM memory. Files created in the MFS can be dynamically created, accessed and destroyed through the use of one or more file

handles which are essentially address pointers into the associated file. The primary purpose for the use of the SDRAM is to provide additional buffer capacity not available with the high-speed on-chip BRAM. Each of our development boards has access to 64MB of SDRAM capacity, thus, it is partitioned into two 32MB blocks, one for each of the 2 LT nodes implemented on each board.

**BRAM Memory Structure** Due to the high-speed requirements of the Multigigabit Rocket I/O transceivers the off-chip SDRAM cannot supply data directly to the Aurora core due to high access latency over the PPC OPB. For this reason, each LT node maintains both a transmitter and receiver side Block RAM module which serve as temporary frame buffers for LT packets. BRAM is an on-chip memory structure capable of single cycle access latency. Each BRAM module is an asynchronous dual port memory used to provide a fast interface between the software application and the Aurora module. Due to limited BRAM resources, each LT node incorporates 2 16KB modules for LT frame transmission and reception temporary storage. Thus, as a limitation, a single LT frame can be a maximum of 16KB.

Prior to transmission over the optical medium the software application loads the BRAM with a packet via port A which is connected to the PPC OPB. Subsequently, the PPC notifies the Aurora protocol, through a software configuration register, that a packet is awaiting transmission. Upon notification, the Aurora TX state machine retrieves 32-bit words from the BRAM, every clock cycle, via port B. Similarly, on the receiver side, the Aurora RX state machine latches incoming 32 bit words into the receiver side BRAM via port B. Upon reception of a complete LT frame the Aurora module signals the PPC that a frame has arrived and can be accessed through port A.

#### 5.3.3.5 MultiGigabit Transceivers (MGT)

High-speed LT communication is enabled using Rocket I/O Multigigabit transceiver (MGT) modules embedded in the Virtex II Pro FPGA. The MGT modules provide the differential serial electronic data stream used to modulate the SFP laser source and receive the incoming electronic stream from the optical receivers. The MGT modules instantiated in the testbed

use a low jitter Low Voltage Differential Signal (LVDS) clock source operating at 75 MHz and multiplied by 20 with a Digital Clock Manager (DCM) incorporated into the MGT to enable the 1.5 Gbps data stream. Details of MGT operation and characterization including can be found in [88].

Due to the complexity of standalone MGT operation including initialization, synchronization, and encoding, Xilinx has provided the Aurora link layer protocol which provides a logical link interface between the sender and receiver MGT instantiations. In addition to providing frame encapsulation the Aurora protocol enables high-level clock synchronization and clock recovery functionality. That is, the Aurora protocol provides a much simpler user interface to the complex operation of the standalone MGTs. Each Aurora module is configured as a single lane, simplex channel and instantiates one hardware MGT for high-speed serial transmission. The MGTs, in turn, are driven by their own lane logic module which handles operations such as 8/10B symbol generation and decoding, transceiver initialization and error detection. Additional details of the Aurora protocol such as lane and channel initialization, clock recovery, clock correction and error and flow control can be found in [87].

To send data over the high-speed channel, the Aurora transmission interface module (state machine) retrieves data buffered in the 16KB TX side BRAM. LT packets are framed using a TX start of frame (TX\_SOF) flag sent at the beginning of frame transmission which signals the intended receiver(s) of the impending frame. Upon reception of the TX\_SOF flag, the receiver asserts the RX\_SOF flag notifying the Aurora RX interface module (state machine) to begin latching the incoming data into the receiver side BRAM data buffer. The end of frame is signaled by the transmitter via a TX\_EOF flag which in turn triggers the RX\_EOF flag at the receiver. Upon frame completion, the RX state machine notifies the PPC software application, via handshake registers, that a packet has arrived and is ready for processing.

In addition to the SOF and EOF flags used to frame LT messages each Aurora module provides addressing capability. Node addressing is handled using the first 32 bits of the data stream. That is, when a receiver enters the RX state machine, after the RX\_SOF is asserted, it examines the first 32 bits of the impending packet. If the address matches, the data is latched

into the local receive BRAM and the PPC application is notified of the packet reception. If the address fails to match, indicating the frame is intended for another destination the frame is simply dropped. The specific address of each node is controlled via software and can be modified by the application through software configuration registers. In addition to point-to-point addressing, a unique broadcast address is maintained to support such traffic.

The previous sections have detailed the FPGA components used in the design of the LT nodes. In the next section we discuss the optical components used in the testbed and discuss their functional characteristics.

### 5.3.4 Optical System Components

Figure 5.8 illustrates the optical components used to configure our 4 node LT testbed. To provide the optical signal we employ continuous wave Small Form Factor Pluggable (SFP) Fabry Perot lasers operating at 1310 nm and internally modulated at 1.5 Gbps using a differential serial signal provided from the FPGA MGT's<sup>1</sup>. Low insertion loss multimode optical splitters and combiners are used to configure a 4 node light-trail segment. As shown in the figure various splitting ratios are used on each link to divert sufficient optical power to each of the LT nodes.

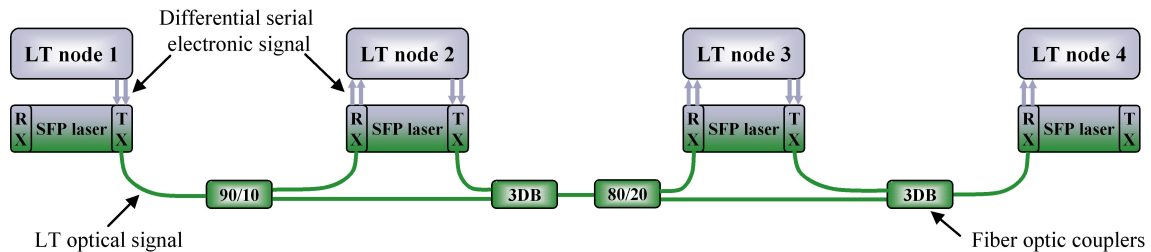


Figure 5.8 Light-trail testbed segment illustrating the optical component organization and coupling ratios

To ensure signal sufficient power is received at each node we performed optical power measurements at each of the downstream nodes. To verify that receiver power is sufficient to

<sup>1</sup>The testbed has been shown to operate successfully at 3.125 Gbps, however, transient faults are seen at this speed due to the maximum rated modulation speed of 2.5 Gbps for the lasers.

allow communication among all four stations of the light trail we monitored the power received at each station on the LT. Table 5.1 lists the power received at each LT node from the various transmitters. The specifications of our optical transceivers indicates a typical optical transmit power of -5 dBm and receiver sensitivity of -22 dBm which suggests an optical power budget of 17dBm. As indicated in the table, the power received at all LT nodes is adequate to support all possible LT connections.

Table 5.1 Measured receive power

|                          | Receive Power |            |            |
|--------------------------|---------------|------------|------------|
|                          | Receiver 2    | Receiver 3 | Receiver 4 |
| Transmitter 1 (-4.8 dBm) | -14.9 dBm     | -16 dBm    | -13.2 dBm  |
| Transmitter 2 (-4.6 dBm) | -             | -15.2 dBm  | -12.2 dBm  |
| Transmitter 3 (-4.5 dBm) | -             | -          | -7.7 dBm   |

Although not implemented in the test bed, a number of devices can be used as the optical shutter such as the magneto-optic switch based on Faraday effect described in [6]. As mentioned, we encountered many physical design limitations in the implementation of the 4 node testbed. The following section outlines two of the most prominent limitations; LT connection setup time and memory latency.

### 5.3.5 Testbed Physical Limitations

#### 5.3.5.1 Connection Setup Time

Because the LT trail architecture is a shared optical medium, individual connections must be established before LT communication is possible. The connection setup time consists of laser TX enable, MGT alignment and Aurora lane synchronization. To determine this connection set up time a simple application is developed as described below.

To accurately determine the connection setup time we developed a simple application operating on a single development board. The application consists of a single Aurora connection between two MGT instantiations, clocked with the same source, on the same FPGA. In this

application, the TX Aurora module begins continuously transmitting an electronic initialization sequence, as defined in Aurora reference manual [87], while the laser source is deactivated. The laser source is then activated and begins optical transmission of the initialization sequence. Upon obtaining MGT alignment and Aurora synchronization the receiver asserts that the channel is up and ready to begin receiving data. The time between these two events, laser enable and channel up, is recorded as the connection setup time.

In order to precisely determine the time at which the laser is activated, a single General Purpose I/O (GPIO) pin is connected from the FPGA to the TX enable pin of the laser source. When the GPIO pin is asserted, activating the laser source, a counter, initialized to zero, begins counting clock transitions until the receiver indicates that the channel is up. The number of clock periods accumulated is then observed as the connection setup time. Although the number of clock transitions varied between subsequent trials due to clock drift and jitter within the MGT modules, we found that a sufficient set up time is approximately 200  $\mu$ s.

As we will see in the subsequent sections this connection setup time is rather excessive and leads to inefficient LT operation. This setup time can be mitigated with the use of burst mode laser drivers such as the Mindspeed MO2090 [62] which suggest burst on/off time of as little as 2.5ns, however, at the time of development such devices were not available. In addition, the cost of such devices were prohibitive on an academic budget.

### 5.3.5.2 Memory Latencies

Another physical limitation of the LT testbed that became evident during development is memory access time. As mentioned in section 5.3.3.5 the Aurora protocol relies on dedicated BRAM to supply and store LT frames. Although the BRAM only requires a single cycle latency for hardware access, the software application relies on a PPC bus transaction to store or retrieve data from the BRAM. Thus, the application access time varies depending upon the PPC frequency and bus speed.

Two simple test designs were developed to determine memory access speed for the software application. The first design simply uses a *for* loop to load and retrieve data from the 16KB

BRAM. Software timing functions are used to determine the time required to fill the entire BRAM. Using a 300 MHz processor clock and a 100 MHz bus clock frequency the bandwidth of BRAM is found to be only about 50Mbps. Because the streaming media application, discussed later, transfers memory from the SDRAM to BRAM we also generated tests to determine the time required for 16KB block transfers between the two memory modules. With SDRAM cache enabled the bandwidth is found to be similar to that of direct BRAM access at about 50Mbps. It is noted that it may be possible to increase this transaction time using Direct Memory Access (DMA) or memory bursting, however, we did not investigate these optimizations in the testbed.

## 5.4 Basic Testbed Operation and MAC

### 5.4.1 Basic Operation

Now that we have defined the FPGA embedded system design and outlined the LT optical components we continue with a discussion of how we tested the operational characteristics of the testbed. In our experiments we designed applications to demonstrate both downstream (hubbed) and upstream communication as discussed Chapter 4. In the hubbed architecture, node 1 is the only transmitting node, thus, no precautions are taken to avoid collisions and node 1's laser source remains on for the duration of the experiment.

Our first experiment is designed to verify that all stations on the LT can receive information from a single transmitting station. For this experiment, 16KB packets are generated at node 1 and sent over the LT to all downstream nodes. The information sent in the LT frames is a simple sequence of numbers, e.g. 1, 2, 3, etc. such that each receiver can easily verify the packet integrity. Once the packet is received at the destination node(s) it is checked for correctness. Using this testbed setup we are able to verify node addressing and broadcast capabilities. After performing these simple tests it is noted that LT operation in the downstream (hubbed) configuration is functional, that is, all downstream nodes are able to receive packets sent from node 1. During the course of the experiment 16KB frames were sent every millisecond and no errors were seen. The oscilloscope shown in Figure 5.9 illustrates a healthy eye pattern



at the fourth node in our 4 node testbed. The test pattern was produced with a 1.5 Gbps data stream sent from node 1 to all downstream nodes of the light-trail.

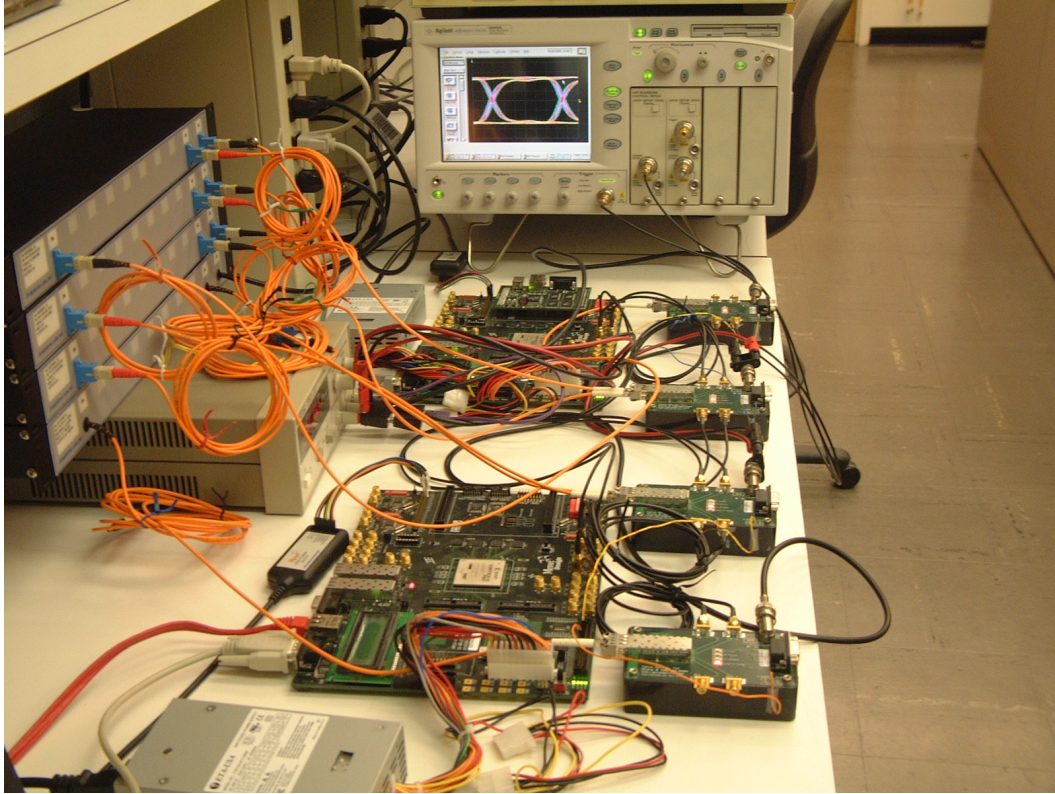


Figure 5.9 Our four node light-trail along with the eye diagram showing a healthy signal at the end of the fourth node

The upstream situation, where multiple connections exist is somewhat more complicated in that access restrictions are required to avoid trail collisions. Thus, a MAC protocol is required and is implemented based upon the LT-FA protocol described in Chapter 4. The following section describes the upstream application experiment.

#### 5.4.2 Testbed MAC protocol

To demonstrate bandwidth sharing on the light-trail, a simple application is designed to share bandwidth between nodes 1, 2 and 3. In our upstream demonstration, connections are established from nodes 1, 2 and 3 to node 4. To demonstrate media sharing a similar application to the one discussed in the previous section is designed. The difference is that in



this application, nodes 1, 2 and 3 take turns transmitting a 16KB frame.

Following the LT-FA MAC, node 1 begins transmission of its 16KB packet which is addressed to node 4. When transmission is complete, the hardware state machine de-asserts the laser TX enable pin, via a GPIO pin, to free the medium for downstream LT node access to the medium. The software application running on node 1 then waits  $1000\mu\text{s}$  before beginning a new round. A new round is begun by notifying the hardware state machine of node 1 to re-initialize, enable the laser source, send the Aurora initialization sequence and transmit the next packet awaiting transmission in the TX BRAM. This  $1000\mu\text{s}$  wait time is referred to as the round interval as described in the LT-FA MAC section of Chapter 4. This round interval is a sufficient amount of time to allow both nodes 2 and 3 to transmit a packet before beginning a new round.

Node 2 on the other hand, upon sensing that the medium is available, after being released by node 1, asserts its laser TX enable pin, waits  $200\mu\text{s}$  for connection setup time and proceeds to send a 16KB packet awaiting in the associated TX BRAM. When node 2's transmission is complete the TX enable pin is de-asserted to free the optical medium for node 3 to begin transmission. Node 3 begins transmission of its packet, addressed to node 4, in a similar fashion to that of node 2. When transmission is complete, node 3's laser is switched off which frees the medium for node 1 to begin a new round.

Although this is a crude implementation of the LT-FA MAC in that the packet size and transmission duration of each node is static, results show that in addition to protecting against collisions the testbed is able to support multiple simultaneous optical connections on the LT shared medium. The throughput of the testbed application, with packet sizes of 16KB, is approximately 384Mbps which is split evenly between nodes 1, 2 and 3. As it turns out, this throughput is still larger than the memory bandwidth limitation of the system as discussed earlier. Thus, packets sent from nodes 1, 2 and 3 are not modified in each round. In addition, node 4 cannot process all the information from a single packet before the next one arrives in the BRAM. Thus to verify packet integrity only the first and last word of each LT frame is examined at node 4. Because the contents of each packet is known a priori and do not change

in subsequent rounds node 4 is able to effectively verify the contents of each received packet. We note that no errors were detected at the link layer nor with the first or last word in each frame.

The previous experiments discussed are designed to verify the feasibility of LT technology; the next section describes a second application that uses the LT backbone to enable end-to-end client communication. We discuss the development and operation of a streaming media application that enables client computers to establish an end-to-end connection over the LT testbed.

### 5.5 Light-trail Streaming Media Application

To demonstrate more sophisticated end-to-end LT operation we designed a streaming media application to run over the testbed. The streaming media application operates over the single wavelength uni-directional light-trail as described earlier. The bus consists of a four node network ( $N_1, N_2, N_3, N_4$ ) as shown in Figure 5.5. Sender/Receiver client stations (host computers) connect to the LT nodes via Ethernet as explained earlier. The optical bus is configured as a static LT with node N1 as the head node and N4 as the end node just as in the aforementioned demonstrations.

A multimedia application runs on the sender client(s) which streams media content to the desired downstream clients(s) via the light-trail. The application implements two primary functions; the first is an interface from the sender and receiver client terminals to the LT nodes and the second allows LT nodes to communicate over the optical channel. In the demonstration, sender client stations stream media content to the FPGA boards via Ethernet socket communication. The source LT node(s) buffer the Ethernet data stream in the SDRAM MFS, encapsulate it into an LT frame and relay the information to downstream LT destination(s). At the receiving end, downstream LT node(s) examine the frame address, buffer the data in the SDRAM MFS, if they are the intended destination, and complete end to end communication to the receiver client via Ethernet.

The following sections describe the media application, client interface operation and LT

communication. We also discuss verification of end-to-end operation.

### 5.5.1 Client Media Application

Nullsoft's SHOUTcast [66] streaming media application is used to enable the end-to-end multimedia data stream. SHOUTcast is Nullsoft's free winamp-based distributed streaming audio system and is chosen because of its relatively simple client operation. SHOUTcast is designed to allow users to distribute audio or video content via the Internet. The simple client interface allows SHOUTcast DJ's to stream raw media content to a dedicated SHOUTcast server through socket based communications. Listeners then "tune in," via socket connection, to the desired server to retrieve the buffered media stream. Figure 5.10 illustrates the simple SHOUTcast streaming media interface. With the use of the freely available SHOUTcast client software and a SHOUTcast server, end-to-end client communication is established.

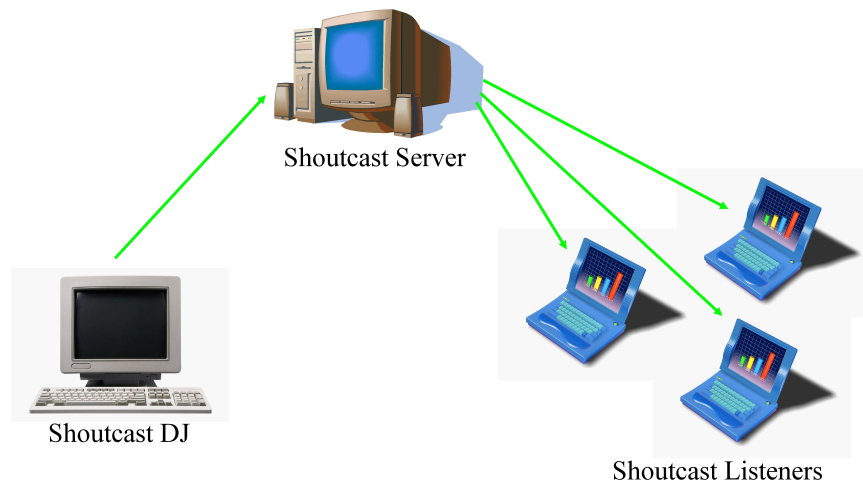


Figure 5.10 SHOUTcast streaming media operation. DJ clients connect to the SHOUTcast server where data is buffered awaiting a connection request from SHOUTcast listeners

In order to provide SHOUTcast streaming media via light-trail, client computers run the DJ and Listener software to connect to the desired LT node which emulates SHOUTcast server functionality. The use of such freely available client software allows us to provide efficient and easy to use communication between client computers and LT nodes. The following sections describe the operation of the streaming media application used in the testbed design.

### 5.5.2 Sender Client Interface

To emulate the DJ side server interface on the LT nodes the FPGA PPC application creates a TCP socket connection, binds the FPGA IP address with a well known port and listens for an impending connection from the sender (SHOUTcast DJ) client. When a sender client is ready to stream multimedia content, it initiates a TCP three way handshake on the listening FPGA port. The selected LT node acknowledges the three way handshake and registers the appropriate LWIP callback function to processes all subsequent packets arriving at the network interface. In addition, a 16MB file (EthDataFile) is created in the SDRAM memory buffer using the MFS to store the multimedia data stream as it is received. Prior to transmission over the LT, the incoming client data stream is written, using LWIP functions, to the EthDataFile from using an MFS file handle (EthWriteHandle) associated with the connection. When the SDRAM buffer becomes full, the file handle is reinitialized to the beginning of the data file and the old data is overwritten.

### 5.5.3 Light-trail Streaming Media Transmission

Data transmitted over the light-trail is supplied to the TX BRAM from the EthDataFile via the PPC memory management functions. The PPC uses a separate file handle, MgtReadHandle, which initially points to the beginning of the EthDataFile, to mark the location in the data file from which the next LT frame will begin. Prior to light-trail transmission, a 16KB data block is copied from the MFS into the TX BRAM. Subsequent to filling the BRAM, the PPC signals the Aurora protocol to proceed with clock synchronization. The first 16KB packet is then addressed to the appropriate destination, encapsulated in a LT frame and transmitted.

As the optical signal traverses the fiber, the LT tap couplers divert a fraction of the optical power to the receivers MGT while the remaining power continues on to all other downstream nodes. Upon detection of light-trail activity, the local MGT is synchronized with the transmission and decides whether or not to process the incoming data based upon the unique software controlled 32-bit address. As mentioned earlier, this address label in conjunction with the SOF flag generated by the Aurora protocol provides unique node addressability.

If the hardware on (say) node  $N_2$  detects that it is the intended destination, the impending data is latched into the receive BRAM and a notification signal is sent to the PPC application to indicate that a packet has been received. Following notification of packet reception, the PPC transfers the associated BRAM data into an separate MFS file (MgtDataFile) at the location of MgtWriteHandle. This file handle is created and initialized to the start of MgtDataFile upon reception of the first packet in the communication stream. The MGTdataFile has characteristics similar to the EthDataFile in that 16MB of space is allocated for the connection and is overwritten when full. That is, the SHOUTcast listener side FPGA server emulator uses the 16MB SDRAM data file to buffer the streaming media from the SHOUTcast DJ and awaits a connection from a listener client as explained in the next section.

#### 5.5.4 Receiver Client Interface

To complete end to end communication, receiver client(s) (SHOUTcast Listener) initiate an Ethernet session with the software application running on the downstream LT nodes. Similar to the sender connection initiation, the receiver opens a socket connection with the LT node which completes the tree way handshake. A new file handle, EthReadHandle, associated with the new connection, is created and initialized to the start of MgtDataFile. The MgtDataFile is used in a similar manner as the EthDataFile except that it is written from by the MGT with data received from the LT which is sent from the sender client. LWIP functions are associated with the connection and handle all subsequent communications. After connection setup has completed, packets are sent from the MgtDataFile at the location of EthReadHandle to the receiver client, thus completing end-to-end streaming connection.

#### 5.5.5 Experimental Results

To verify operation of the streaming media application TCP dump traces are obtained using the Ethereal network protocol analyzer. In this verification, traces of a SHOUTcast DJ client connected to node 1 are compared to traces obtained from a listener client connected to node 4 of testbed. The traces indicate that the streaming data, sent to LT node 1 from

the SHOUTcast DJ are successfully relayed through the LT and received by LT node 4. In addition, we are able to listen and view audio and video transmissions on both the sender and receiver client stations to verify correct operation.

## 5.6 Summary

The success of the RRPP in industry has had a strong impact on the research and education community. With such a universal platform, engineers, students and academic researchers are able to explore more innovative designs that can be rapidly developed, tested, and demonstrated while gaining valuable insight into complex implementation requirements. We believe research efforts can strongly benefit from the RRPP by enhancing project visibility which can ultimately lead to an increased cooperation between industry and academic research. In addition, the RRPP promotes the development of more robust research solutions at a fraction of the time and cost of full production development which is often not feasible in the academic setting.

In this Chapter we have discussed the valuable role that the RRPP has played in our research efforts in the DCNL, HSSE, and ESL. We presented the light-trail testbed and discussed how its development has enhanced our understanding of the system level design considerations of light-trail technology.

We conclude that the LT project has indeed gained visibility through the development of the RRPP prototype. In addition, we realize that a complete system design requires careful consideration of all system components, not just the underlying LT technology. That is, even though the LT is capable of supporting multiple connections at a relatively high data rate the limitations placed on the testbed by the system components is overwhelming. These factors are exactly what the LT prototype is designed to address. We have shown through testbed experiments that LT technology is feasible, however, we have exposed system level design issues such as the the need for burst mode laser drivers, efficient memory management and LT access control techniques must be considered to make LT's suitable for metro networking. It is these limitations that must be addressed in future work to make the LT prototype more efficient.

## CHAPTER 6. Conclusion

Fiber optic technologies enabling high-speed, high-capacity digital information transport have only been around for about 3 decades but in their short life have completely revolutionized global communications. To keep pace with the growing demand for digital communications and entertainment, fiber optic networks and technologies continue to grow and mature. As new applications in telecommunications, computer networking and entertainment emerge, reliability, scalability, and high Quality of Service (QoS) requirements are increasing the complexity of optical transport networks.

In this dissertation we discussed existing and emerging technologies in modern optical communications networks. We outlined traditional telecommunication and data networks that enable high speed, long distance information transport. We examined various network architectures including mesh, ring and bus topologies of modern Local, Metropolitan and Wide area networks. We presented some of the most successful technologies and network protocols used in today's communications networks, outlined their shortcomings and introduced promising new technologies to meet the demands of future transport networks.

We presented a comprehensive discussion of the most recognized fairness models and MACs for ring and bus networks which laid the groundwork for the introduction of the Robust, Dynamic and Fair Network (RDFN) protocol for ring networks as presented in Chapter 3. The RDFN protocol is a novel solution to fairly share ring bandwidth for bursty asynchronous data traffic while providing bandwidth and delay guarantees for synchronous voice traffic. It was shown that the RDFN network provided fairness in ring networks without suffering from the bandwidth oscillations inherent in the RPR protocol. Furthermore, the RDFN protocol is not susceptible to large network delays as in the Cyclic Reservation Multiple Access protocol.

We presented the light-trail architecture and technology in Chapter 4 as a solution to providing high network resource utilization, seamless scalability and network transparency for metropolitan area networks. We presented three medium access control protocols for light-trails that provide collision protection but do not consider fair network access. As an improvement to these light-trail MAC protocols we introduced the Token LT and light-trail Fair Access (LT-FA) MAC protocols and evaluate their performance. We illustrate how fairness is achieved and access delay guarantees are made to satisfy the bandwidth budget fairness model as introduced in Chapter 2. We also show that the LT-FA protocol is the best solution for light-trail media access control through performance comparisons with that of other unidirectional bus protocols suitable for adaptation to the light-trail architecture such as the Pi-persistent protocol.

The second area of discussion in this dissertation dealt with the rapid prototyping platform. We discussed how the reconfigurable rapid prototyping platform (RRPP) is being utilized to bridge the gap between academic research, education and industry. We presented of the Real-time Radon transform and the Griffin parallel computing platform implemented using the RRPP to illustrate how the RRPP has enhanced our academic research efforts and brought industry visibility to our research projects in addition to exposing undergraduate students to valuable research experiences. We presented the LT testbed and discussed how it is used to provide additional insight on the real-world limitations of light-trail technology. As a proof of concept, we introduce the light-trail testbed developed at the High Speed Systems Engineering lab. We provide details on its operation and discuss two applications developed to enhance our understanding of light-trail technology. We show how the LT-FA MAC has been implemented on the testbed and demonstrate a streaming media application.

As a whole, this dissertation provided a comprehensive discussion of current and future technologies and trends for optical communication networks. In addition, we provided media access control solutions for ring and bus networks to address fair resource sharing and access delay guarantees. Finally, we completed the work with a testbed developed using the RRPP which demonstrated proof of concept for light-trail technology and outlined system level design challenges for future optical networks.



## Bibliography

- [1] Peter Alfke. *The Future of Field-Programmable Gate Arrays*. 2003. Xilinx, Inc., San Jose, CA.
- [2] S. Aluru, N. Futamura, and K. Mehrotra. Parallel Biological Sequence Comparison Using Prefix Computations. In *International Parallel and Distributed Processing Symposium*, pages 653–659, April 1999.
- [3] J. Ash, Li Chung, K. D’Souza, Wai Sum Lai, H. Van der Linde, and Yung Yu. AT&T’s MPLS OAM Architecture, Experience, and Evolution. *IEEE Communications Magazine*, 40(10):100–111, October 2004.
- [4] Avnet. *Xilinx Viretex II Pro FF1152 Development Kit*. <http://www.em.avnet.com/>.
- [5] A.S Ayad, K.M. El Sayed, and S.H Ahmed. Efficient Solution of the Traffic Grooming Problem in Light-Trail Optical Networks. In *IEEE Symposium on Computers and Communications, ISCC '06*, pages 622–627, June 2006.
- [6] R. Bahuguna, M. Mina, and R.J. Weber. A novel all fiber magneto-optic on-off switch. In K. M. Iftekharuddin and A. A. S. Awwal, editors, *Photonic Devices and Algorithms for Computing VII*, pages 15–21, August 2005.
- [7] S. Balasubramanian, A.E. Kamal, and A.K. Somani. Network Design for IP-Centric Light Trail Networks. In *IEEE Conference on Local Computer Networks, LCN*, pages 174–181, November 2005.

- [8] S. Balasubramanian, A.E. Kamal, and A.K. Somani. Network Design for IP-Centric Light Trail Networks. In *International Conference on Broadband Networks*, pages 45–54, October 2005.
- [9] S. Balasubramanian, Ahmed Kamal, and A. K. Somani. Medium Access Control Protocols For Light-trail and Light-bus Networks. In *8th Working Conference on Optical Networks Design and Modelling, ONDM 2004*, February 2004.
- [10] Chas. B. Barr. *Telegraph Stations in the United States, the Canadas and Nova Scotia*. <http://commons.wikimedia.org/wiki/Image:OptischerTelegraf.jpg>.
- [11] Mikel Bezdek, Daniel Helvick, Ramon Mercado, Diane Rover, Akhilesh Tyagi, and Zhao Zhang. Developing and Teaching an Integrated Series of Courses in Embedded Computer Systems. In *Proceedings of the 36nd ASEE/IEEE Frontiers in Education Conference, FIE '02*, San Diego, California, October 2006.
- [12] T. Bonald, L. Massoulié, A. Proutiere, and J. Virtamo. A Queueing Analysis of Max-min Fairness, Proportional Fairness and Balanced Fairness. *Queueing Systems*, 53(1-2):65–84, June 2006.
- [13] N. Bouabdallah, A.L Beylot, E. Dotaro, and G. Pujolle. Resolving the Fairness Issues in Bus-Based Optical Access Networks. *IEEE Journal on Selected Areas in Communications*, 23(8):1444–1457, August 2005.
- [14] Jean-Yves Le Boudec. *Rate adaptation, Congestion Control and Fairness: A Tutorial*. November 2005.
- [15] D. Bouldin. Impacting Education Using FPGAs. In *International Parallel and Distributed Processing Symposium*, pages 142–147, April 2004.
- [16] Martin L. Brady. A Fast Discrete Approximation Algorithm for the Radon Transform. *Society for Industrial and Applied Mathematics*, 27(1):107–119, February 1998.

- [17] W.E. Burr, S. Wakid, Xiaomei Qian, and D. Vaman. A Comparison of FDDI Asynchronous Mode and DQDB Queue Arbitrated Mode Data Transmission for Metropolitan Area Network Applications. *IEEE Transactions on Communications*, 42(2/3/4):1758–1768, February 1994.
- [18] A. Carena, V. D. Feo, J. M. Finochietto, R. Gaudino, F. Neri, C. Piglione, and P. Poggiolini. RingO: An Experimental WDM Optical Packet Network for Metro Applications. *IEEE Journal on Selected Areas in Communications*, 22(8):1561–1571, October 2004.
- [19] D. Cavendish, K. Murakami, S.H. Yun, O. Matsuda, and M. Nishihara. New Transport Services For Next-generation SONET/SDH Systems. *IEEE Communications Magazine*, 40(5):80–87, May 2002.
- [20] G. Cena, L. Durante, R. Sisto, and A. Valenzano. Comparison of Adaptive Fairness Control Mechanisms for DQDB Metropolitan Area Networks. In *Proceedings of the 1995 IEEE Fourteenth Annual International Phoenix Conference on Computers and Communications*, pages 205–211, March 1995.
- [21] R. Chamberlain, J. Lockwood, S. Gayen, R. Hough, and P. Jones. Use of a Soft-core Processor in a Hardware/Software Codesign Laboratory. In *International Conference on Microelectronic Systems Education, MSE '05*, pages 97–98, June 2005.
- [22] Shun Yan Cheung. Controlled Request (DQDB): Achieving Fairness and Maximum Throughput in the DQDB Network. In *Proceedings of IEEE INFOCOM'92*, pages 180–189, 1992.
- [23] A.L. Chiu and R.G. Gallager. Full Utilization, Fairness and Bounded Access Delay on High Speed Bus Networks. In *International Conference on Network Protocols*, pages 154–161, Oct 1996.
- [24] F. Davik, M. Yilmaz, S. Gjessing, and N. Uzun. IEEE 802.17 Resilient Packet Ring Tutorial. *IEEE Communications Magazine*, 42(3):112–118, March 2004.

- [25] J-M. Dilhac. *The Telegraph of Claude Chappe - An Optical Telecommunication Network for the XVIII<sup>th</sup> Century*.
- [26] W. Dobosiewicz and P. Gburzynski. On Token Protocols for High-speed Multiple-ring Networks. In *International Conference on Network Protocols*, volume 24, pages 10–17, May 1986.
- [27] W. Dobosiewicz and P. Gburzynski. On the Use of Multiple Tokens on Ring Networks. In *Conference of the IEEE Computer and Communication Societies, INFOCOM 1990*, pages 15–22, June 1990.
- [28] C. Douligeris and L.N. Kumar. Access to a Network Channel: A Survey into the Unfairness Problem. In *IEEE International Conference on Communications, SUPERCOMM/ICC '92.*, pages 1184–1189, June 1992.
- [29] Adam Dunkels. *Light-Weight Internet Protocol*. <http://savannah.nongnu.org/projects/>.
- [30] Jing Fang, Wensheng He, and A.K. Somani. Optimal Light Trail Design in WDM Optical Networks. In *International Conference on Wireless and Optical Communications Networks*, volume 4, pages 1699–1703, June 2004.
- [31] M.T. Frederick, N.A. VanderHorn, and A.K Somani. Light Trails: A Sub-Wavelength Solution for Optical Networking. In *Workshop on High Performance Switching and Routing, HPSR 2004*, pages 175–179, March 2004.
- [32] M.T. Frederick, N.A. VanderHorn, and A.K. Somani. Real-time Hardware Implementation of the Approximate Discrete Radon Transform. In *International Conference on Application-Specific Systems, Architecture Processors, ASAP '05*, pages 399–404, July 2005.
- [33] Violeta Gambiroza, Ping Yuan, Laura Balzano, Yonghe Liu, Steve Sheafor, and Edward Knightly. Design, Analysis, and Implementation of DVSR: A Fair High-Performance Protocol for Packet Rings. *IEEE/ACM Transactions On Networking*, 12(1):85–102, February 2004.

- [34] R. Gaudino, A. Carena, V. Ferrero, A. Pozzi, V. De Feo, P. Gigante, F. Neri, and P. Poggiolini. RINGO: A WDM Ring Optical Packet Network Demonstrator. In *European Conference on Optical Communication, ECOC '01*, volume 4, pages 620–621, September 2001.
- [35] A. Ge, F. Callegati, and L.S. Tamil. On Optical Burst Switching and Self-similar Traffic. *IEEE Communications*, 4(3):90–100, 2000.
- [36] C. Guillemot, M. Henry, F. Clerot, A. Le Corre, J. Kervaree, A. Dupas, and P. Gravey. KEOPS Optical Packet Switch Demonstrator: Architecture and Test Bed Performance. In *Optical Fiber Communication Conference*, volume 3, pages 204–206, March 2000.
- [37] A. Gumaste and I. Chlamtac. Light-Trails: A Novel Conceptual Framework for Conducting Optical Communications. In *Workshop on High Performance Switching and Routing, HPSR 2003*, pages 24–27, June 2003.
- [38] A. Gumaste and I. Chlamtac. Mesh Implementation of Light-Trails: A Solution to IP Centric Communication. In *Conference on Computer Communications and Networks, ICCCN '03*, pages 178–183, October 2003.
- [39] A. Gumaste and I. Chlamtac. Light-trails: An Optical Solution for IP Transport. *Journal of Optical Networking*, 3(5):261–+, May 2004.
- [40] A. Gumaste, G. Kuper, and I. Chlamtac. Optimizing Light-Trail Assignment to WDM Networks for Dynamic IP Centric Traffic. In *IEEE Workshop on Local and Metropolitan Area, LANMAN '04*, pages 113–118, April 2004.
- [41] A. Gumaste and Si Qing Zheng. Optical Implementation of Resilient Packet Rings Using Light-Trails. In *21st Optical Fiber Conference/National Fiber Optic Engineers Conf, NFOEC/OFC '05*, March 2005.
- [42] A. Gumaste and Si Qing Zheng. Protection and Restoration Scheme for Light-Trail WDM Ring Networks. In *Optical Network Design and Modeling, ONDM '05*, pages 311–320, February 2005.

- [43] R. Gupta and A.K. Somani. Game Theory as a Tool to Strategize as Well as Predict Nodes' Behavior in Peer-to-Peer Networks. In *International Conference on Parallel and Distributed Systems*, volume 1, pages 244–249, July 2005.
- [44] E. Hahne, A. Choudhury, and N. Maxemchuk. DQDB Networks With and Without Bandwidth Balancing. *IEEE Transactions on Communications*, 40(7):1192–1204, July 1992.
- [45] E.L. Hahne, A.K Choudhury, and N.F Maxemchuk. Improving the Fairness of Distributed-Queue-Dual-Bus Networks. In *Proceedings of IEEE INFOCOM'90*, pages 175–184, June 1990.
- [46] T.S. Hall and J.O. Hamblen. System on a Programmable Chip Development Platforms in the Classroom. *IEEE Transactions on Education*, 47(4):502–507, November 2004.
- [47] Wensheng He, Jing Fang, and A. K. Somani. On Survivable Design in Light Trail Optical Networks. In *8th Working Conference on Optical Networks Design and Modelling, ONDM 2004*, February 2004.
- [48] E. Hernandez-Valencia, M. Scholten, and Zhenyu Zhu. The Generic Framing Procedure (GFP): An Overview. *IEEE Communications Magazine*, 40(5):63–71, May 2002.
- [49] IBM. *Blue Gene Project*. <http://www.research.ibm.com/bluegene/>.
- [50] A.E Kamal and A. Bissonauth. Priority Mechanism for the DQDB Network. *IEE Proceedings Communications*, 141(2):98–104, April 1994.
- [51] A.E. Kamal and H.S. Hassanein. Throughput Analysis of WDM-based Dual-Bus Local Area Networks. In *International Performance, Computing and Communications Conference*, pages 426–432, February 1999.
- [52] S. Kasemlonnappa and J.S. Meditch B. Mukherjee. A Delay-Throughput Performance Analysis of the Pi-persistent Protocol for Unidirectional Broadcast Bus Networks. In

- Conference of the IEEE Computer and Communications Societies, INFOCOM 1989*, pages 834–840, April 1989.
- [53] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability. *The Journal of the Operational Research Society*, 49(9):237–252, March 1998.
- [54] L. Kleinrock. Power and Deterministic Rules of Thumb for Probabilistic Problems in Computer Communications. In *International Conference on Communications, ICC '79*, volume 43, pages 1–10, June 1979.
- [55] D. Kliazovich, F. Granelli, H. Woesner, and I. Chlamtac. Bidirectional Light-Trails for Synchronous Communications in WDM Networks. In *IEEE Global Telecommunications Conference, GLOBECOM '05*, volume 4, page 5, November 2005.
- [56] G. Kramer, B. Mukherjee, and G. Pesavento. Ethernet PON (EPON): Design and Analysis of an Optical Access Network. *Photonic Network Communications*, 3(3), July 2001.
- [57] L.N Kumar and C. Douligeris. Demand and Service Matching at Heavy Loads: a Dynamic Bandwidth Control Mechanism for DQDB MANs. *IEEE Transactions on Communications*, 44(11):1485–1495, November 1996.
- [58] Richard J. La and Venkat Anantharam. Utility-Based Rate Control in the Internet for Elastic Traffic. *IEEE Transactions On Networking*, 10(2), April 2002.
- [59] D. Manjunath and M.L. Molle. The Effect of Bandwidth Allocation Policies on Delay in Unidirectional Bus Networks. *IEEE Journal on Selected Areas in Communications*, 13(7):1309–1323, September 1995.
- [60] Ravi Mazumdar, Lorne G. Mason, and C. Douligeris. Fairness in Network Optimal Control: Optimality of Product Flow Forms. *IEEE Transactions on communications*, 39(5), May 1991.

- [61] G.J. Miller and M. Paterakis. A Study of the Accuracy of the Bernoulli Approximation of the Slot Occupancy Process in High-speed Unidirectional-bus Networks. In *IEEE International Conference on Communications, SUPERCOMM/ICC 1994*, volume 2, pages 985–990, May 1994.
- [62] Mindspeed. *3.3V Burst-Mode Laser Driver and Integrated Limiting Amplifier with Eye-Minder(TM) Technology for Fiber-to-the-Premises*. <http://www.mindspeed.com/web/product/info.html?id=762>, 2006.
- [63] B. Mukherjee. On the Infinite Buffer Model and the Implementation Aspects of the Persistent Protocol for Unidirectional Broadcast Bus Networks. In *IEEE International Conference on Communications, ICC 1988*, pages 273–277, June 1988.
- [64] H.R. Muller, M.M. Nassehi, J.W. Wong, E. Zurfluh, W. Bux, and P. Zafiropulo. DQMA and CRMA: New Access Schemes for Gbit/s LANs and MANs. In *Ninth Annual Joint Conference of the IEEE Computer and Communication Societies, INFOCOM 1990*, volume 1, page 185.191, June 1990.
- [65] M.M. Nassehi. CRMA: An Access Scheme for High-speed LANs and MANs. In *IEEE International Conference on Communications, ICC 1990*, volume 4, pages 1697–1702, April 1990.
- [66] Nullsoft. *Nullsoft SHOUTcast Streaming Media Technology*. <http://www.shoutcast.com/>.
- [67] M.J. O’Mahony, D. Simeonidou, D.K. Hunter, and A. Tzanakaki. The Application of Optical Packet Switching in Future Communication networks. *IEEE Communications Magazine*, 39(3):128–135, March 2001.
- [68] S. O’Shea and J. Finucane. Generalized DQDB for Metropolitan Area Networks. In *International Conference on Broadband Services, Systems and Networks*, pages 73–77, November 1993.
- [69] J.W.M. Pang and F.A. Tobagi. FDDI - A Tutorial. *IEEE Communications Magazine*, 24(5):10–17, May 1986.



- [70] J.W.M. Pang and F.A. Tobagi. Throughput Analysis of a Timer Controlled Token Passing Protocol Under Heavy Load. *IEEE Transactions on Communications*, 37(7):694–702, July 1989.
- [71] C. Qiao and M. Yoo. Optical Burst Switching (OBS) - A New Paradigm for an Optical Internet. *Journal of High Speed Networks, JHSN 1999*, 8(1):69–84, 1999.
- [72] R. Ramaswami. Optical Networking Technologies: What Worked and What Didn't. *IEEE Communications Magazine*, 44(9):132–139, September 2006.
- [73] M. Renaud, F. Masetti, C. Guillemot, and B. Bostica. Network and System Concepts for Optical Packet Switching. *IEEE Communications Magazine*, 35(4):96–102, April 1997.
- [74] Y.F. Robichaud and Changcheng Huang. Improved Fairness Algorithm to Prevent Tail Node Induced Oscillations in RPR. In *IEEE International Conference on Communications, ICC 2005*, pages 402–406, May 2005.
- [75] N. Sarkar and K. Pawlikowski. A Delay-Throughput Performance Improvement to the Pi-persistent Protocol. In *IEEE Symposium on Computers and Communications, ISCC '01*, pages 615–620, July 2001.
- [76] N. Le Sauze, A. Dupas, E. Dotaro, L. Ciavaglia, N.H.M Nizam, A. Ge, and L. Dembeck. A Novel, Low Cost Optical Packet Metropolitan Ring Architecture. In *European Conference on Optical Communication, ECOC '01*, volume 23, pages 66–67, 2001.
- [77] J. Schneider, M. Bezdek, Z. Zhang, Z. Zhang, and D.T. Rover. A Platform FPGA-based Hardware-Software Undergraduate Laboratory. In *International Conference on Microelectronic Systems Education, MSE '05*, pages 53–54, June 2005.
- [78] R. Srinivasan and A.K. Somani. Blocking in Multiwavelength TDM Networks. In *International Conference on Telecommunication Systems, Modeling, and Analysis*, pages 535–541, March 1996.

- [79] R. Srinivasan and A.K. Somani. On Achieving Fairness and Efficiency in High-speed Shared Medium Access Networks. *IEEE/ACM Transactions on Networking*, 11(1):111–124, February 2003.
- [80] H.R. van As, J.W. Wong, and P. Zafiropulo. Fairness, Priority and Predictability of the DQDB MAC Protocol Under Heavy Load. In *International Zurich Seminar on Digital Communications, Electronic Circuits and Systems for Communications*, pages 410–417, March 1990.
- [81] N. VanderHorn, M. Mina, and A. K. Somani. Light Trails: A Passive Optical Networking Solution for Wavelength Sharing in the Metro. In *U.S. Pakistan International Conference on High Capacity Optical Networks and Enabling Technologies, HONET '04*, December 2004.
- [82] N.A VanderHorn, S. Balasubramanian, M. Mina, and A.K Somani. Light-trail Testbed for IP-centric Applications. *IEEE Communications Magazine*, 43(8):5–10, August 2005.
- [83] S. Verma, H. Chaskar, and R. Ravikanth. Optical Burst Switching: A Viable Solution for Terabit IP Backbone. *IEEE Network*, 14(6):48–53, Nov/Dec 2000.
- [84] A. Viswanathan, N. Feldman, Z. Wang, and R. Callon. Evolution of Multiprotocol Label Switching. *IEEE Communications Magazine*, 36(5):165–173, May 1998.
- [85] Wikipedia. *Wikipedia Commons*. The Library of Congress, Geography and Map Division, 1853.
- [86] M.J. Wirthlin. Senior-level Embedded System Design Project Using FPGAs. In *International Conference on Microelectronic Systems Education, MSE '05*, pages 91–92, June 2005.
- [87] Xilinx. *Aurora Link-Layer Protocol*. <http://www.xilinx.com/aurora>.
- [88] Xilinx. *Virtex II Pro User Guide*. <http://www.xilinx.com>.

- [89] Shun Yao, B. Mukherjee, and S. Dixit. Advances in Photonic Packet Switching: An Overview. *IEEE Communications Magazine*, 38(2):84–94, February 2000.
- [90] J. Yates, J. Lacey, and D. Everitt. Blocking in Multiwavelength TDM Networks. In *International Conference on Telecommunication Systems, Modeling, and Analysis*, pages 535–541, March 1996.
- [91] Yabin Ye, H. Woesner, R. Grasso, Tao Chen, and I. Chlamtac. Traffic Grooming in Light Trail Networks. In *Global Telecommunications Conference, GLOBECOM '05*, volume 4, page 6, November 2005.
- [92] Ping Yuan, V. Gambiroza, and E. Knightly. The IEEE 802.17 Media Access Protocol for High-speed Metropolitan-area Resilient Packet Rings. *IEEE/ACM Transactions On Networking*, 18(3):8–15, May-June 2004.
- [93] Xiaobo Zhou, Guowei Shi, Hongbo Fang, and Lieguang Zeng. Fairness Algorithm Analysis in Resilient Packet Ring. In *International Conference on Communication Technology, ICCT 2003*, pages 622–624, April 2003.