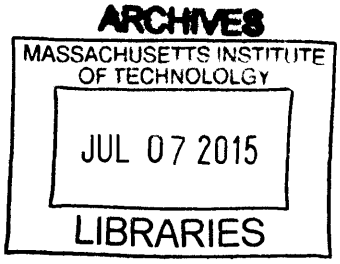
Metropolitan Area Network Architecture Design for Optical Flow Switching



by

Xijia Zheng

B.Eng., The University of Hong Kong (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author . . . Signature redacted .....
Department of Electrical Engineering and Computer Science
May 18, 2015

Certified by . . Signature redacted .....
Vincent W. S. Chan
Joan and Irwin Jacobs Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . Signature redacted .....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Metropolitan Area Network Architecture Design for Optical Flow Switching

by

Xijia Zheng

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Optical Flow switching (OFS) is a key enabler of future scalable all-optical networks for the large traffic flows. In this thesis, we provide design concepts of efficient physical topology and routing architectures for an all-optical Metropolitan Area Network (MAN) that supports OFS.

We use all-to-one stochastic flows to model inter-MAN traffic demands and adopt Moore Graphs and Generalized Moore Graphs as the physical topology. We found good MAN architectures are coupled intimately with media access control protocol designs and must be optimized jointly. Two routing architectures that represent extreme cases were proposed and examined: Quasi-Static Architecture (QSA) and Dynamic Per Flow Routing Architecture (DPFRA). The performance and costs are compared to provide an economical architecture building strategy.

We find for the MAN, DPFRA always has the lower queueing delay and lower blocking probability than that of QSA at the expense of more complexity in scheduling, switching, and network management and control. Our analysis based on Moore Graphs and Generalized Moore Graphs indicates that QSA becomes cheaper when the product of the average offered load per node and the normalized delay are equal to or larger than ~ 2 units of wavelengths, with both architectures essentially meeting the same delay or blocking probability requirements. Also, the cost boundary shows that DPFRA with shortest-queue node first routing strategy (sq-first strategy) is preferred only when the delay requirement is stringent and the offered load is low, while QSA is much more suitable for the all-optical MAN to accommodate modest to heavy network traffic. Since OFS is only going to be used in heavy load situations brought on by elephants in the traffic, QSA is the preferred architecture. We have shown the hybrid architecture of QSA and DPFRA is impractical and thus it should be avoided.

Thesis Supervisor: Vincent W. S. Chan
Title: Joan and Irwin Jacobs Professor of Electrical Engineering and Computer Science

Acknowledgments

First and foremost, I would like to show my ultimate gratitude to my research advisor, Professor Vincent W. S. Chan, for his expertise, support, and patience. Without his guidance and encouragement, this thesis would not have been possible, let alone finished one year in advance. His emphasis on practical oriented design, creative thinking and effective communication skills continues to be a great inspiration to me. Moreover, his dedication and invaluable life advice have been and will continue to help me find my direction in life.

I would like to sincerely thank my undergraduate research supervisor Professor Victor O. K. Li at the University of Hong Kong, for introducing me into the field of research and encouraging me to pursue my graduate study at MIT.

A special thanks to Lei Zhang and Henna Huang for their help and support on my research. I would like to thank Donna Beaudry for her care and support for the whole group. I would like to thank my fellow labmates: Matthew Carey, Antonia Feffer, Jessica Weaver, John Metzger, Esther Jang, and Andrew Song. All of them have made my time in "Chan's Clan" colorful and enjoyable.

Thank you to my friends at MIT: Zi Wang, Hsin-Yu Lai, Yunming Zhang, David Qiu, Yewen Pu, Tianren Liu, Chengtao Li, Guowei Zhang, and Jiajun Wu. It was really wonderful to spend time with you all, playing badminton, playing bridge, and exploring Boston.

Last but not the least, I would like to thank my parents for their unconditional love, guidance, and support all the time. I feel very grateful to be the child of these two excellent engineers, who opened the door to the field of engineering for me when I was young. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid developments in technology nowadays have led to the explosively growing volume of data and information, which demands the synchronous advancements in data transmission. For example, the increasing high data rate products such as high-definition (HD) video streaming, video call, cloud computing, and big data analytics require accesses to the high capacity of fiber network to accommodate the increasing number of users with satisfactory performances. One of the bottlenecks constraining (mostly due to cost) the high data rate performance over the entire network is the use of electronically switched schemes in Metropolitan Area Network (MAN) and access network [4][5]. Though the existing electronically switched network is excellent in handling small transactions, the costs of electronic components and electrical switching schemes grow with increasing data rates and become unaffordable for traffic that goes through disruptive increases such as in heavy-tail (elephants) transactions. All-optical networking provides relief in the cost structure for large transactions and thus transport such as Optical Flow Switching (OFS), can reduce the cost per bit of elephant traffic.

OFS is a scheduled, all-optical, end-to-end user service to accommodate "elephant flows" of the dramatic increase in data volume demands of emerging applications [4]. Users can directly get the access to vast backbone network bandwidth without the compromise of the cost and delay of data processing in the MAN and access network. To achieve excellent performance and high cost efficiency, an all-optical MAN that

can support OFS efficiently should be properly designed from the Physical to the Transport Layers. For high network utilization operating at an acceptable performance (e.g. delay), the right physical and media access control (MAC) architecture must also be used.

Optical flows can be done over quasi-statically provisioned tunnels, but the lack of agility (including per flow switching) may compromise the efficiency of the architecture and thus the cost of the network is not minimized. However, agile per flow switching incurs additional cost for the network in cost for network management and control and scheduling and in some cases faster optical switches are more expensive. In this thesis, we focus on the architecture design of the MAN for OFS. Both the physical topology and the routing architecture of the MAN based on the MAN traffic model will be discussed in detail. [14]

1.1 Introduction to Optical Flow Switching

OFS is an agile all-optical network service for users with large traffic flows [4]. Upon the users' traffic flow transmission requests, the schedulers at both the ingress MAN and the egress MAN coordinate to set up the user-to-user connections for each request prior to traffic flow transmissions. When the dedicated paths are set up, traffic flow transmissions start immediately and no collision will happen [4] [11] [12].

Figure 1-1 is an example showing the scheduling process of OFS across MANs and Wide Area Network (WAN). In this scenario, both source users S1 and S2 request to send their own traffic flows to the corresponding destination D1 and D2, respectively. When both S1 and S2 request transmissions, the schedulers in the source MAN where S1 and S2 reside, located in San Francisco, will reserve resources and allocate paths for each transmission. Then the scheduler in the destination MAN where D1 and D2 reside, located in Boston, will synchronize with the scheduler in the source MAN, reserving resources and allocating paths in the destination MAN. Finally, the dedicated end-to-end paths across all MANs and WAN are properly allocated for both users. Both S1 and S2 can therefore transmit the traffic flows individually without collision.

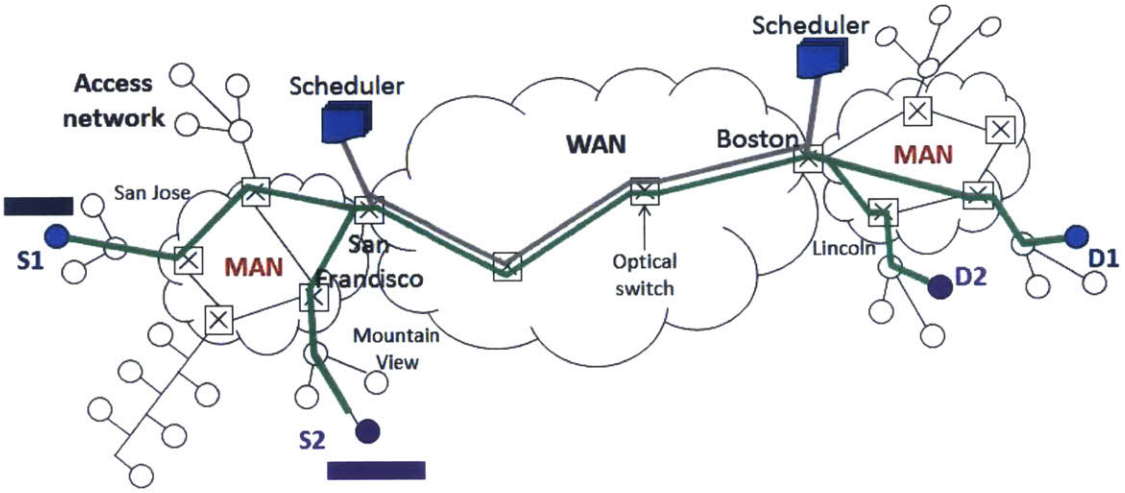


Figure 1-1: Optical Flow Switching Architecture with WAN/MAN/LAN. Reproduced from [4]

We assume in this thesis that over the WAN wavelength tunnels are set up between the MANs [13] to slow down the control plane and the complexity of cross traffic scheduling in the WAN. The remaining question is whether in the MAN should per flow switching be used or should there be quasi-static wavelength tunnels as well.

1.2 MAN Design For Optical Flow Switching

A MAN in this context serves as the inter-connection between Local Area Network (LAN)s where end-users reside and the WAN (and subsequent egress MAN and LAN). The Hierarchy of the network is shown in Figure 1-2. The primary function of MANs is to manage the transmission of traffic generated from the end users in the LANs to the destination. There are two important constructs to the MAN architecture design for OFS: 1. Physical layer topology (links and switches); 2. Scheduling, routing, control, and management.

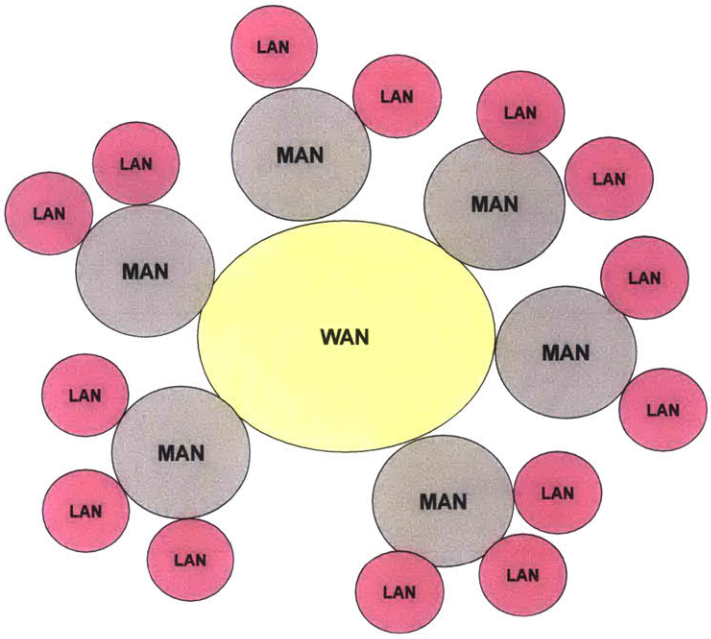


Figure 1-2: Hierarchy of the network.

1.2.1 Previous Work on MAN for Optical Flow Switching

Our work focus on the MAN architecture design based on the two important MAN architecture constructs mentioned above. Parametric cost models for the MAN involving both constructs are then used to optimize the architecture. Previous research has mainly focused only on the first construct, the physical topology design [7][10]. Guan has performed a detailed analysis on the MAN physical topology selection on regular topologies. He also provides the parametric, first-order, and homogeneous network cost model for MAN including the transceiver cost, fiber connection cost, and the switching cost. Based on his work, Generalized Moore Graphs are the good candidates for the MAN physical topologies, since Generalized Moore Graphs achieve near minimum cost (minimum in the case of Moore Graphs) with the least number of wavelengths. Lin has shown Generalized Moore Graphs with node degrees between $0.05N$ and $0.08N$, where N is the number of nodes, are both power and cost minimal for an all-optical network. She has also proved that shortest path and minimum hop routing is power optimal in most topologies and traffic load balanced routing should

be avoided.

It should be noted that those works only deal with the design of the MAN physical topology, while further research on the architecture constructs, cost and efficiency involving network management and control and dynamic resource allocation are needed. Thus the cost model for MAN also need modifications to include the cost incurred by the network management and control. Besides, both Guan and Lin adopted a deterministic traffic model in the analysis, which does not include stochastic traffic model for the MAN architecture. Moreover, there were no considerations of the management and control and scheduling of bursty traffic in those works. In this work, we provide a MAN architecture design for OFS including both constructs (physical topology and network management and control) on the foundation of the previous work.

1.2.2 Key Objectives and Contributions

In this work, we first provide a MAN traffic model to describe the bursty traffic flow characteristics. Then we focus on the architecture involving the dynamic management and control. We propose and examine two extreme architectures: 1. Quasi-Static Architectures (QSA) that only changes with traffic trends; 2. Dynamic Per Flow Routing Architectures (DPFRA) that allows per flow routing and switching. For QSA, the network management and control algorithms (for spatial and wavelength switches) do not have to reconfigure the network topology very quickly. Intuitively, it does not have to make routing decisions at high per flow speeds and only adjust quasi-statically. On the other hand, DPFRA, though certainly more efficient, (using optical switching) needs fast network management and control which will be a challenge to the design and implementation of a large scale network. For DPFRA, we design two rerouting strategies, shortest-queue node first routing strategy (sq-first strategy) and traffic-receiving node first routing strategy (tr-first strategy).

The objective of this work is to determine which of the two architectures (or a hybrid) is preferred and recommend an appropriate physical architecture and MAC protocol for the access network and MAN. We will identify cost efficient MAN physical

and MAC protocol network architectures that are easy to implement. In particular we are looking for efficient and simple to implement architectures that can avoid per flow switching (which is a big burden for network management and control) if at all possible.

1.3 Architecture Evaluating Metrics

The differences in the architectures require a clear comparison using different metrics to evaluate the cost and performance. Here, relevant metrics include delay, blocking probability and cost of implementing the network. There are two kinds of delay, the transmission delay and the queueing delay. The transmission delay is the time a traffic flow spends on the transmission on the wavelength. The queueing delay is the time a traffic flow waits to be transmitted, the time interval between the arrival and the transmission. The blocking probability is the probability that the transmission request of a flow is rejected. Costs are measured via throughput or number of wavelengths, switches and fibers used and network management and control efforts.

There are trade-offs between the cost and the performance metrics including the delay and the blocking probability. To keep a lower delay or a lower blocking probability of the MAN, we need more network capacity, expenditures and more network control effort. Our aim is to find an economical building strategy which meet the operating requirements in terms of the delay and the blocking probability.

It is not hard to determine that dynamic architecture which allow per flow routing has the shorter delay with a lower blocking probability, while the quasi-static architecture needs to use more wavelengths to meet the same operating requirement. However, if we take costs into consideration, the preferred architecture may change. Therefore, a comprehensive evaluation is required in the design.

1.4 Thesis Organization

The rest of the thesis is organized as follows.

In Chapter 2, we provide the MAN traffic model for the further MAN architecture design. The MAN traffic model is based on the analysis of the characteristics of the bursty MAN traffic.

In Chapter 3, we introduce the MAN physical topology based on Guan's work [7] and the MAN traffic model generated in Chapter 2. We discuss the regularity of the MAN topology and thus regular graphs are the good candidates. Moore Graph and Generalized Moore Graph are discussed and used as the physical topology of MAN in this work.

In Chapter 4, we propose the MAN routing architecture. First, we provide the basic queueing model for each node in the MAN to analyze the network performance in terms of the delay and the blocking probability. The schedule holder size design is discussed to provide a guideline on the routing architecture design based on different operating requirements. Then we proposed two extreme routing architectures, one extreme where rerouting is not allowed, and the other extreme where rerouting is enabled for each flow. We provide the detailed rerouting strategies for the two cases and model them into different queueing systems based on the basic queueing model. The comparisons of the performances based on the analytical results are provided to evaluate different rerouting strategies and thus the different routing architectures. The idea of hybrid architecture is proposed for the further analysis.

In Chapter 5, we generate the parametric cost models for the different architectures. The parametric cost models are based on the traffic model of the MAN, the physical topology of the MAN, and the cost model for each network component, including both the capital expenditures and the operating expenditures. We compare the cost of different architectures based on the different operating conditions. Finally, we summarize this chapter by comparing different architecture designs to provide a suitable building strategy of MAN for OFS.

In Chapter 6, we conclude the thesis with a summary of our contributions. Besides,

we discuss the future work for the continuing research in this area.

Chapter 2

MAN Traffic Model

One of the functionalities of a MAN is to transfer traffic between LANs and the WAN. From Section 1.2, we know that a MAN can be considered as an interface between LANs and the WAN. One node in the MAN serves as the hub and is the gateway to the WAN. The hub is accessible by all the other MAN nodes. The MAN nodes also interconnect LANs, transferring traffic to/from LANs. In this structure, we distinguish two kinds of traffic in the MAN: intra-MAN traffic and inter-MAN traffic. Intra-MAN traffic describes data transfer between LANs within the same MAN. Inter-MAN traffic describes data transfer between MANs, where traffic travels through the WAN. In [7] [10], Guan and Lin have studied intra-MAN traffic. In this work, we mainly focus on inter-MAN traffic.

The traffic generated in a geographic area can be approximated to be proportional to the population of that area. For example, an area with a dense population generates more traffic than an area with a sparse population. To reduce network cost, it is likely that several sparsely populated areas will share the same end node to the MAN. Conversely, a densely populated area will have an entire single node served by the MAN. We assume the average traffic volume generated between each MAN node and the hub is identical. This is accomplished by partitioning a MAN by population distribution. Thus, MAN nodes either connect a single LAN or multiple LANs, depending on the population distribution of each LAN. Also we assume the traffic transmission of every node to the hub is independent of each other, since they aggregate the traffic

from different areas which are considered statistically independent.

Previous analyses tend to assume a static traffic demand between nodes. However, this assumption does not capture the bursty nature of data traffic. One of the main contributions of this work is to model the time varying stochastic nature in MANs. In this chapter, we build a traffic model for the MAN based on the characteristics of the MAN traffic. In the following chapters, we design a network architecture to satisfy the traffic demand based on this traffic model presented in this chapter.

2.1 All-to-one Stochastic Traffic Model

In our traffic model, we make three assumptions about MAN traffic originating from the MAN node destined for the hub. First, we assume the traffic is all-to-one, which best describes the inter-MAN traffic characteristics. The other commonly used all-to-all uniform traffic model is not addressed here, since that model mainly describes intra-MAN traffic among dense areas with uniform and well-balanced traffic. Second, we assume the traffic is stochastic. This is due to the transmission of bursty data traffic from the end nodes of the MAN to the WAN. Third, we assume the distributions of the arrival process in all the end nodes are independent and identically distributed (I.I.D.) and the sizes of flows are I.I.D. random variables. The I.I.D. arrival process is due to the evenly balanced traffic of all the nodes. The I.I.D. flow size is used to simplify analyses.

The all-to-one traffic, addresses the traffic transmission constituting a tree logical topology, which is the transmission between the end nodes and the hub [7]. One example of all-to-one traffic is illustrated in Figure 2-1. A graph G models the physical MAN topology with the total number of nodes as $V(G) = N$, the total number of edges as $E(G) = M$, node degree as Δ, and the graph diameter as D. Node V_0 is designated as the hub from the MAN to the WAN. Meanwhile, the other $N-1$ nodes are end nodes. The bold lines in Figure 2-1 represent the traffic directions. Denote $\mathbf{T} = [T_{i,j}]$ as the traffic matrix for MAN. Each end node V_i sends $T_{i,0}$ amount of traffic per second to the hub V_0 and receives $T_{0,i}$ traffic from the hub. The traffic

transmission between each end node is ignored for now and should be addressed in a different work. By the all-to-one traffic pattern, we have

$$T_{i,j} = 0, \qquad i = j, \text{or } i, j \neq 0. \tag{2.1}$$

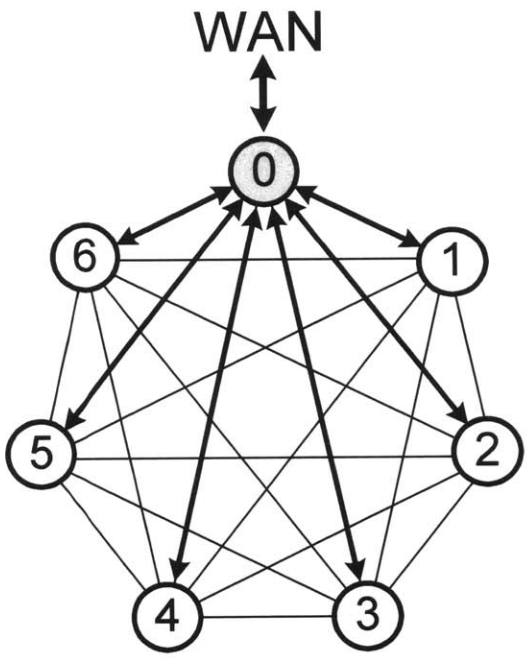


Figure 2-1: Schematic diagram showing the fully connected network topology with all-to-one traffic among nodes. Node V_0 is the hub. All other nodes are the end nodes.

Since sending and receiving are assumed to be symmetric processes, only the direction of sending traffic from end nodes to the hub is considered in the following analysis. Also, to simplify the notation, the end node is referred to the node, which is distinguished from the hub. The end node, the node, and the MAN node are used interchangeably.

The arrival process of the stochastic traffic of each node is assumed to be a Poisson process with the average arrival rate of λ. Denote the arrival rate at epoch t as $\lambda(t)$. In addition, we assume the sizes of all flows are also I.I.D. random variables to reflect

the variability associated with the traffic. It is assumed the size of each traffic flow L is exponentially distributed with mean $\overline{L}$. There is another time scale of interest and that is the rate at which the average traffic load changes. This can occur in hours, minutes or as short as seconds. So the average traffic function for the the nodes to the hub

$$T_{i,0} = \lambda(t)L, \qquad i \neq 0 \tag{2.2}$$

The average traffic $\overline{T}_{i,0}$ per node is

$$\overline{T}_{i,0} = E[\lambda(t)L] = \lambda\overline{L} \tag{2.3}$$

2.2 Traffic Load

We employ the traffic load to further evaluate the traffic situation in the MAN. With k wavelengths to the MAN with the hub V_0 in total, we evenly divide k wavelengths to all the $N-1$ nodes. So the number of wavelengths assigned to each node is defined as

$$x = \frac{k}{N-1} \tag{2.4}$$

Namely, we can regard it as that we evenly divide the traffic into x wavelengths of total traffic per node. To ensure stability and/or no excessive blocking of the network, we have the constraint as

$$k \geq \frac{\sum E[T_{i,0}]}{R} \tag{2.5}$$

where R is the transmission rate of a wavelength channel.

Therefore, the load ρ is defined as

$$\rho = \frac{\sum E[T_{i,0}]}{kR} = \frac{\overline{T}_{i,0}}{xR} < 1 \tag{2.6}$$

To reflect the loading situation in terms of the number of lighpaths, we define another variable $\overline{F}$ as the offered load per node in unit of wavelength

$$\overline{F} = x\rho \tag{2.7}$$

Chapter 3

Topology Design for MAN

In this chapter, we describe the MAN physical topology. Most of the observations in this chapter can be found in Guan's work on OFS physical network topology [7][8].

In designing the network topology, we consider the physical characteristics of the network. For example, the WAN spans a wide geographic area over thousands of kilometers and connects many MANs. Its network topology is usually sparse and arbitrary. Thus, it is difficult to categorize the WAN physical topology into a specific class of graphical models. In contrast, the MAN spans a relatively smaller geographic area and has a higher concentration of nodes than the WAN. Thus, regular topologies with symmetry and well-defined connectivity pattern are good candidate structures to model the MAN physical topology. In this work, we focus on these regular toplogies. More specifically, we adopt the definition of regular topologies stated in [7]. It says a topology is regular with node degree Δ and diameter D if it satisfies the following conditions:

- There are Δ outgoing edges from and Δ incoming edges to each node.
- The topology has nodal symmetry such that each node links to Δ other nodes following predefined connectivity rules.
- The topology has Δ-connectedness such that the number of nodes that are i hops away from a node (define as $n(i)$) via minimum hop routing for i less than the diameter of the network is at lease Δ. That is $n(i) \geq \Delta, 1 \leq i \leq D-1$.

- Average minimum hop distance H_{min} between node pairs of the topology can be expressed as

$$H_{min} = \frac{1}{N-1} \sum_{i=1}^{D} in(i) \tag{3.1}$$

There are plenty of regular graphs candidates, including Moore Graphs, Generalized Moore Graphs, Δ-Nearest Neighbors Graphs, Symmetric Hamilton Graphs, ShuffleNet, Hypercube, etc. [7]. From the results in [7][8][10][11], Moore Graphs are the optimal candidates and Generalized Moore Graphs are the nearly optimal candidates for OFS in terms of both cost and power consumption. Also, Guan has shown that Moore Graphs and Generalized Moore Graphs have the highest robustness (in cost) to traffic demand uncertainties [7], which shows they are good physical topologies to accommodate the stochastic traffic that we deal with in this work. In our subsequent analysis, we only consider Moore Graphs and Generalized Moore Graphs to represent the MAN physical topology.

3.1 Moore Graphs

Moore Graphs are a class of ideal regular topology satisfying the Moore bound. From the literature of graph theory [3], the Moore bound is an upper bound on the number of nodes in a graph with given the diameter D and the maximum node degree Δ. For a directed graph, the Moore bound is

$$\begin{aligned} N_{max}(\Delta_{max}, D) &\leq 1 + \sum_{i=1}^{D} (\Delta_{max})^i \\ &= \frac{\Delta_{max}^{D+1} - 1}{\Delta_{max} - 1} \end{aligned} \tag{3.2}$$

For an undirected graph, the Moore bound is

$$N_{max}(\Delta_{max}, D) \leq 1 + \sum_{i=0}^{D-1}(\Delta_{max} - 1)^i \tag{3.3}$$
$$= 1 + \Delta_{max}\frac{(\Delta_{max} - 1)^D - 1}{\Delta_{max} - 2}$$

Here, notice that $\Delta_{max} = \Delta$ for regular graphs.

Since Moore Graphs achieve this upper bound, each node in a Moore graph can reach every other node in a fully populated Δ-ary minimum hop routing spanning tree, where a spanning tree is a connected subgraph including all of the nodes of the original graph and has no cycles. The path between each node pair along this Δ-ary minimum hop routing spanning tree is unique. This characteristic of Moore Graphs is very important to the network topology design, since it achieves the lower bound on the average minimum hop distance H_{min} among all regular topologies with the same node number and node degree. So the minimum cost with the least number of wavelengths is achieved by Moore Graphs.

From [7], for a directed Moore Graph with node degree Δ and diameter D, the average minimum hop distance is

$$H_{min_{dir.Moore}} = \frac{D\Delta^D}{\Delta^D - 1} - \frac{1}{\Delta - 1} \tag{3.4}$$

For an undirected Moore Graph, the average minimum hop distance is

$$H_{min_{undir.Moore}} = \frac{D(\Delta - 1)^D}{(\Delta - 1)^D - 1} - \frac{1}{\Delta - 2} \tag{3.5}$$

One example of Moore Graph is the Petersen Graph with $N = 10$, $M = 15$, $\Delta = 3$, and $D = 2$ is shown in Figure 3-1. Referring to the all-to-one traffic model in Chapter 2, the Petersen Graph can be transformed into a spanning tree with the top most node denoted as V_0, which is the hub. All the other nine nodes at different levels will send their traffic to the hub. The solid lines in (b) represent the fiber connection in the embedded tree topology in (a). Dashed lines are fiber connections not included

in the spanning tree.

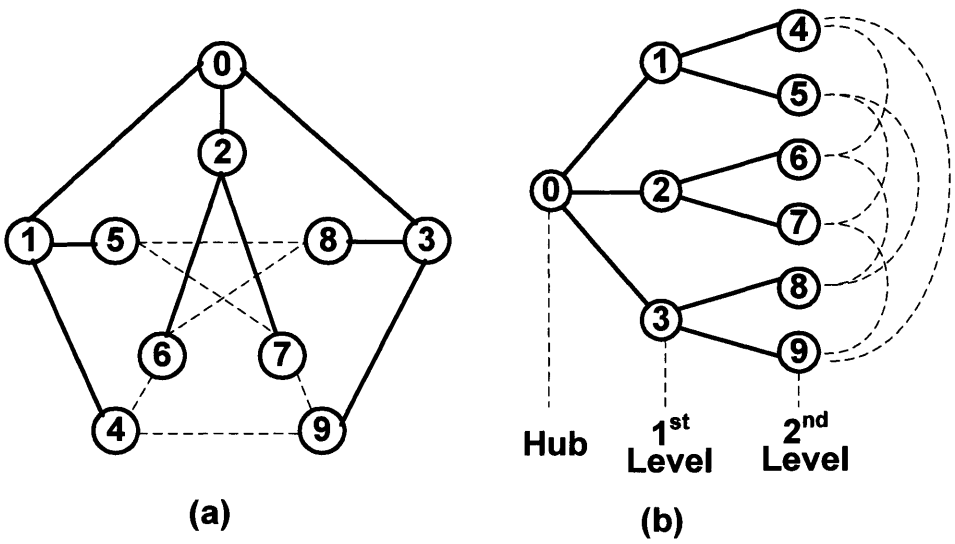


Figure 3-1: (a) The MAN topology with Petersen Graph, $N = 10, M = 15, \Delta = 3$, and $D = 2$; (b) Routing spanning tree embedded in physical architecture with node V_0 as the hub. Reproduced from [7], Figure 4-3.

3.2 Generalized Moore Graphs

Generalized Moore Graphs are a class of ideal regular topology which do not achieve the Moore bound, but achieve the lower bound on the average minimum hop distance. A Generalized Moore Graph has a Δ-ary minimum hop routing spanning tree with all levels fully filled, except possibly the last level, level D. Therefore, Generalized Moore Graphs achieve near minimum cost with the least number of wavelengths. If the last level is full, it is Moore Graph. In other words, Moore graphs are a special class of Generalized Moore Graph [7][10].

For a directed Generalized Moore Graph with node size N and node degree Δ, the average minimum hop distance is

$$H_{min_{dir.G.Moore}} = \frac{\Delta - \Delta^{D+1} + ND(\Delta-1)^2 + D(\Delta-1)}{(N-1)(\Delta-1)^2} \tag{3.6}$$

As $N \to \infty$, $H_{min_{dir.G.Moore}} \to \log_{\Delta} N$.

For an undirected Generalized Moore Graph, the average minimum hop distance is

$$H_{min_{dir.G.Moore}} = \frac{\Delta[1-(\Delta-1)^D] + ND(\Delta-2)^2 + 2D(\Delta-2)}{(N-1)(\Delta-2)^2} \tag{3.7}$$

As $N \to \infty$, $H_{min_{undir.G.Moore}} \to \log_{\Delta-1} N$.

One example of a Generalized Moore Graph is the Heawood Graph with $N = 14, M = 21, \Delta = 3$, and $D = 3$ as shown in Figure 3-2. Here, node V_0 is the hub while all the other 13 nodes at different levels will send their traffic to the hub. The solid lines in (b) represent the fibers connections in an embedded tree topology in (a). Dashed lines are fiber connections not included in the spanning tree.

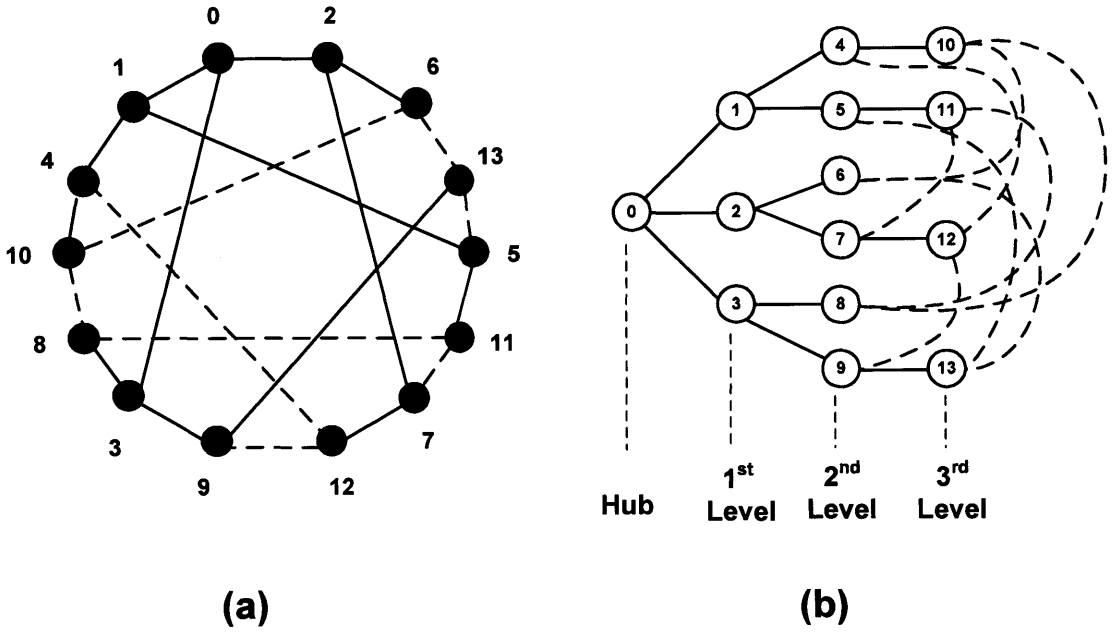


Figure 3-2: (a) The MAN topology with Heawood Graph, $N = 14, M = 21, \Delta = 3$, and $D = 3$; (b) Routing spanning tree embedded in physical architecture with node V_0 as the hub. Reproduced from [7], Figure 4-6.

Chapter 4

MAN Routing Architecture Design

Given the traffic model described in Chapter 2 and the physical network topology model described in Chapter 3, we move on to the design of network scheduling, control, and management. In this work, we focus on the OFS routing architecture design for the MAN. Routing is a critical process in a network architecture because it is the decision process of selecting the transmission paths for the offered traffic. With a given physical network topology, the strategies of how to allocate resources and select paths can vary greatly. We need to figure out what OFS MAN routing strategy gives the best performance while keeping the architecture cost-efficient.

In this work, we focus on routing decisions inside the MAN, where traffic originates from the MAN nodes and is destined for the hub. Ultimately, the traffic will be transmitted from the hub and through the WAN. In Section 2.1, it is assumed that the traffic generated from the LANs arrives at each MAN node as a Poisson Process of rate λ. The routing of traffic between node pairs within the same MAN will not be considered here, since in this work we only focus on inter-MAN traffic transmission. All traffic is first attempted to be routed along nodes via shortest path routing unless blocking necessitates rerouting through other intermediary nodes. The routing of traffic through an intermediate node that is not on the original attempted path is called rerouting, which is illustrated in Figure 4-1. The schedule holder at a node is considered as marker to the corresponding wavelength channels at this node specifying the future assignment [4][13]. In this work, the schedule holder at a node only holds

transmission requests from this node to the hub. In the example of Figure 4-1, the schedule holder at node V_i only holds requests from node V_i to the hub. Once an incoming traffic flow is received at node V_i, there are two routing options:

- The incoming traffic flow waits in the schedule holder of node V_i. The incoming request will be transmitted from node V_i to the hub according to the schedule holder policy. In this work, we assume a first-come first-serve queueing policy.

- The incoming traffic flow is rerouted to node V_j because V_j has the shortest queue at the moment and node j can be reached through fiber connection(s) from node V_i. The traffic is then transmitted from node V_j to the hub on a first-come-first-serve basis. This rerouting strategy will be discussed in detail in the following sections.

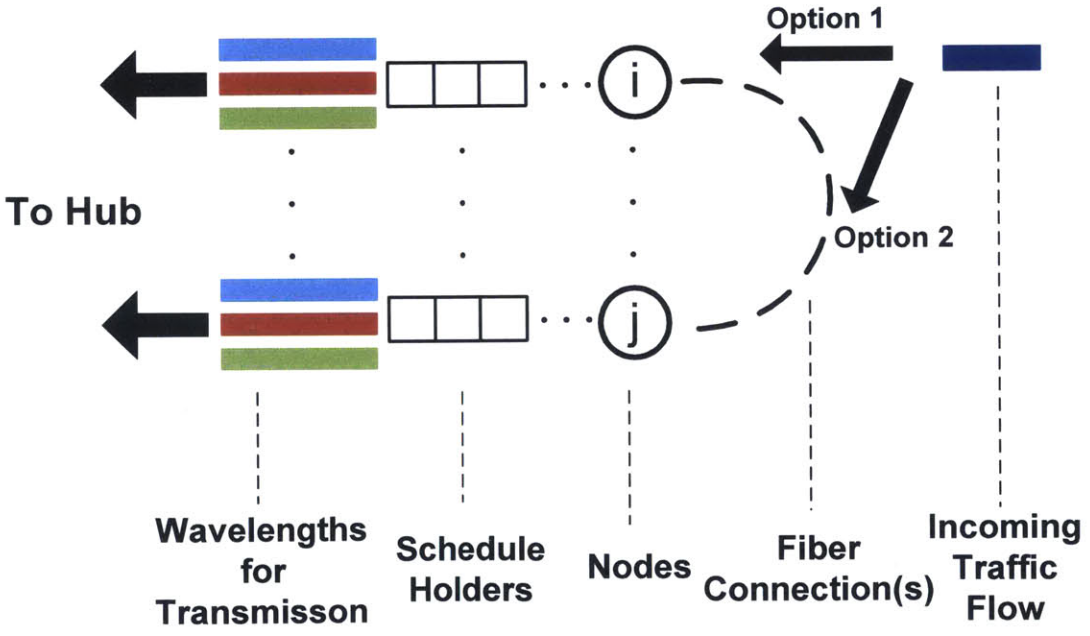


Figure 4-1: Routing options for the incoming traffic flow received at node V_i.

4.1 Queueing Model

We can evaluate the performance of different routing architectures analytically by modeling the nodes in the MAN with a queueing system. In the MAN, flows waiting to be transmitted at each node can be modeled as a finite queue. The traffic volume is in unit of flows. Flows are generated by users in the LANs, and arrive at each node with rate λ. The arriving traffic will be transmitted to the hub on the assigned wavelength(s). The transmission rate is determined by the number of wavelengths, where the number of wavelengths is modeled as the number of servers in the queueing system. If the amount of traffic arriving exceeds the capacity (in terms of the number of wavelengths) of the node and the finite number of schedule holders [13], the arriving traffic will be blocked due to overflow at the schedule holders. The MAN therefore can be modeled as a queueing system with finite queues. In this section, we provide queueing models for the all to one stochastic MAN traffic and the design of schedule holder.

4.1.1 Schedule Holder Size Design

One important assumption in this work is that the size of the schedule-holders are finite for each node. The idea of virtual schedule holders[4][13] is used for schedulers to determine how to schedule the transmission of the flows. Consider the situation when the arrival rate exceeds the service rate of the system; it leads to congestion which may never be quenched by the system. To guarantee reasonable delay performance, we do not allow a single flow to wait at a node for an infinitely long time. The traffic is blocked and will be dropped out of the queueing system. When an arbitrary flow is blocked the first time it first arrives, the user will re-enter the system and request retransmission after a random time delay. A reasonable blocking probability for OFS is 10^{-2}. In this case, the probability for one particular flow to be blocked three times is less than $(10^{-2})^3 = 10^{-6}$. Such blocking probability can be considered negligible in our analysis. If we allow the limit of retransmissons to be greater than three, the blocking of flows will tend to zero. So we must determine the size of the schedule

holders to satisfy the target blocking probability. Here, we can use Little's Theorem [9] to determine the average size of the schedule holders.

Theorem 1 (Little's Theorem). *The average number of customers in a queueing system (with possibly infinite holding time) is equal to the average arrival rate of customers to that system, times the average time spent in the system.*

Since the blocking probability is small we can use Little's Theorem to find a good approximation of the necessary size of the holder. The number of flows in the queueing system m includes the number of flows being transmitted by the wavelengths and the number of flows waiting in the schedule holders. So the size of schedule holders of one node should at least be (plus some margin to be determined later) the average number of flows in the queueing system of this node minus the number of transmission wavelengths of this node. If the average queueing time of a flow is τ_Q and the average transmission time of a flow is τ_{TR}, then the average total delay of a flow in the queueing system is $\tau_Q + \tau_{TR}$. With the average arrival rate of the flow at each node as λ, we have the average number of flows in the queueing system of each node N_S as

$$N_S \sim \lambda(\tau_Q + \tau_{TR}) \tag{4.1}$$

Notice this is an approximation to the average number of flows in the systems guided by Little's Theorem. The equality in Little's Theorem is replaced by $\sim$ because of blocking. For the design of the schedule holder size, we need to also consider the fluctuation of the number of flows in the systems which sometimes can exceed the mean. So we should design the size of the schedule holder to be greater than the mean number of flows in the system minus the number of wavelengths. The fluctuations in the flow arrival process can be characterized by the moments of the arrival process. Here, we consider only the second moment, or variance σ^2. We determine the number of flows in the queueing system just before overflow to be expressed as $N_S + \alpha\sigma$, where $\alpha\sigma$ is a design margin over the mean. Since $N_S + \alpha\sigma$ may not be an integer, we round it up to give a integer value to the number of flows in the queueing system

$$m = \lceil N_S + \alpha\sigma \rceil \tag{4.2}$$

where $\lceil x \rceil$ is the ceiling function and the smallest integer not less than x.

The size of the schedule holder is then $m - x = N_S + \alpha\sigma - x$, where x is the number of wavelengths assigned to each node as defined in Section 2.2.

For the Poisson arrival process with λ arrival rate assumed in this work, the number of arrivals in time interval $(t, t + \tau_Q + \tau_{TR}]$ follows a Poisson distribution with associated parameter $\lambda(\tau_Q + \tau_{TR})$. The variance is $\lambda(\tau_Q + \tau_{TR})$ and the standard deviation is

$$\sigma = \sqrt{\lambda(\tau_Q + \tau_{TR})} = \sqrt{N_S} \tag{4.3}$$

So the number of flows in the queueing system for the Poisson arrival process is

$$m = \left\lceil N_S + \alpha\sqrt{N_S} \right\rceil \tag{4.4}$$

There are many ways to determine the proper α depending on the specific network requirements. For example, one common requirement is the blocking probability requirement that we use in this work. Assume that the blocking probability is expected to be below a certain threshold $P_{b_{thr.}}$. So by Chebyshev's Inequality [2], we have

$$\Pr\{m - N_S \geq \alpha\sigma\} \leq \Pr\{|m - N_S| \geq \alpha\sigma\} \tag{4.5}$$

$$\leq \frac{1}{\alpha^2} = P_{b_{thr.}} \tag{4.6}$$

Finally, we have

$$\alpha = \frac{1}{\sqrt{P_{b_{thr.}}}} \tag{4.7}$$

By Equation 4.7, we can get the desired α for a given blocking probability constraint. Without loss of generality, we round α up to give an integer marginal coeffi-

cient to simplify our analysis. For example, $\alpha = 10$ when $P_{b_{thr.}} = 10^{-2}$. $\alpha = 32$ when $P_{b_{thr.}} = 10^{-3}$.

Notice that Chebyshev's Inequality only provides a general upper bound on α for all distributions with the same σ. If we know the arrival statistics we can further tighten the bound. Assume an arrival process with number of arrivals X. Let m^+ be the maximum number of flows that can be accommodated in the queueing system before overflow. To satisfy the blocking probability constraint, we must satisfy the constraint

$$Pr\{X \geq m^+\} \leq P_{b_{thr.}} \tag{4.8}$$

By Equation 4.2 and Equation 4.8, we have

$$\alpha = \left\lceil \frac{m^+ - \overline{m}}{\sigma} \right\rceil \tag{4.9}$$

where $\overline{m}$ is the mean number of flows in the system and σ is the standard deviation.

For the Poisson arrival process assumed in this work, we get this tighter bound $\alpha = 4$ for $P_{b_{thr.}} = 0.01$. Details are shown in Appendix A.1.

4.1.2 Single Node Queue Model

From 4.1.1, we know each node in the MAN can be modeled as a finite queue. As assumed, the queue at each node can accommodate m flows, including the flow(s) being transmitted and the flow(s) waiting in the schedule holders. Since in Section 2.1 we take the flow size to be exponentially distributed with mean $\overline{L}$, the service time is also exponentially distributed. With transmission rate R, the average transmission delay (the service time) is

$$\tau_{TR} = \frac{\overline{L}}{R} \tag{4.10}$$

The average service rate is

$$\mu = \frac{1}{\tau_{TR}} = \frac{R}{\overline{L}} \tag{4.11}$$

Each node can be therefore modeled as an $M/M/x/m$ queue. The first M means the arrival of the traffic is a Poisson arrival process and the second M means the transmission time of each flow is exponentially distributed. x represents the number of servers, which is number of wavelengths assigned to this node. m is the number of flows can be accommodated in this queueing system, which is the sum of the number of servers and the number of schedule holders at this node.

Further combining Equation 2.3 and 2.6, we have

$$\lambda = \frac{kR\rho}{(N-1)\overline{L}} \tag{4.12}$$

Since $x = \frac{k}{N-1}$, we get the load as

$$\rho = \lambda \frac{N-1}{k} \frac{\overline{L}}{R} = \frac{\lambda}{x\mu} < 1 \tag{4.13}$$

A detailed derivation of blocking probability and queueing delay for an M/M/x/m queueing model is given in Appendix A.2. The blocking probability of an $M/M/x/m$ queue is

$$P_B = \frac{1}{x^{m-x}x!}(x\rho)^m p_0 \tag{4.14}$$

where p_0 is the probability that the system is empty. For p_0, we have

$$p_0 = [\sum_{n=0}^{x} \frac{(x\rho)^n}{n!} + \frac{(x\rho)^{x+1}}{x \cdot x!} \cdot \frac{1-\rho^{m-x}}{1-\rho}]^{-1} \tag{4.15}$$

The average queueing delay τ_Q of $M/M/x/m$ queue is

$$\tau_Q = \frac{\sum_{n=x}^{m}(n-x)\frac{x^x}{x!}\rho^n p_0}{\lambda(1-P_B)} \tag{4.16}$$

4.1.3 Merged Node Queue Model

For the rerouting case, we need to generalize the single node queueing model in Section 4.1.2 to a merged node queueing model. In the single node queueing model, each MAN node is considered as a separate queueing system. In the merged node queueing model, a subset of i nodes in the MAN is considered as one merged node and modeled by a single shared queueing system. Such a merged node has arrival rate $i\lambda$, number of servers ix, and number of users in the system im. We can get the queueing model of this merged node with certain modifications to the model of an $M/M/x/m$ queue. It can thus be modeled into an $M/M/ix/im$ queue with an arrival rate of $i\lambda$. Notice that the load stays the same, which is

$$\rho = \frac{i\lambda}{ix\mu} = \frac{\lambda}{x\mu} < 1 \tag{4.17}$$

For the merged queue of of i I.I.D. nodes, the blocking probability P_{B_i} is

$$P_{B_i} = \frac{1}{(ix)^{im-ix}(ix)!}(ix\rho)^{im}p_{0_i} \tag{4.18}$$

where p_{0_i} is the probability that the merged queueing system is empty. For p_{0_i}, we have

$$p_{0_i} = [\sum_{n=0}^{ix} \frac{(ix\rho)^n}{n!} + \frac{(ix\rho)^{ix+1}}{ix \cdot (ix)!} \cdot \frac{1-\rho^{im-ix}}{1-\rho}]^{-1} \tag{4.19}$$

The average queueing delay τ_{Q_i} of the $M/M/ix/im$ queue with an arrival rate of $i\lambda$ is

$$\tau_{Q_i} = \frac{\sum_{n=ix}^{im}(n-ix)\frac{(ix)^{ix}}{(ix)!}\rho^n p_{0_i}}{i\lambda(1-P_{B_i})} \tag{4.20}$$

4.2 Quasi-static Architecture (QSA)

The key idea of QSA is that no rerouting is allowed. In this scheme, all traffic generated in a LAN is routed to the hub via the corresponding MAN node that connects this LAN to the MAN. If the queue of this MAN node is full, then the traffic will be blocked. Since we assume all the MAN nodes are I.I.D., traffic transmission on one node is independent from traffic transmission on any other node. Thus, each node in QSA can be modeled as an $M/M/x/m$ queue. The average delay and the blocking probability of QSA can be acquired by the $M/M/x/m$ queueing model. With the queueing model derived in Section 4.1, the blocking probability of QSA is

$$P_{qsa} = P_{B_1} \tag{4.21}$$

The average queueing delay of QSA is

$$\tau_{qsa} = \tau_{Q_1} \tag{4.22}$$

4.3 Dynamic Per Flow Routing Architecture (DPFRA)

The key idea of DPFRA is that rerouting is enabled for every flow. In this scheme, all traffic generated in a LAN can be rerouted to another MAN node with free wavelengths and then sent from the new node to the hub. Notice that the new MAN node is not the same MAN node that connects this LAN to the MAN and the traffic will be blocked only if the capacity of the MAN as a whole is exceeded.

With rerouting enabled in DPFRA, the next design question is how to perform rerouting. In flow transmission, the rerouting strategy depends on both the wavelength occupancy level and the queue occupancy level. If one or more wavelengths are available when a flow arrives from the LAN to a MAN node, it will receive service immediately. However, if all wavelengths are currently serving other users, the incoming flow will need to wait in the queue and incur a waiting delay penalty. Finally, if the queue is full, the flow will be blocked. Thus, there are two situations where

rerouting may be performed. The first is to reroute whenever all wavelengths are occupied in the transmitting MAN node, regardless of whether or not the queue is full. The second is to only perform rerouting when the queue of the MAN node is full. We propose two rerouting strategies according to these two situations:

1. **shortest-queue node first routing strategy (sq-first strategy):** In sq-first strategy, the arriving traffic always finds the node with the shortest queue, minimizing its time to wait for data transmission.

2. **traffic-receiving node first routing strategy (tr-first strategy):** In tr-first strategy, the arriving traffic always waits in the queue of the MAN node that receives the traffic, unless the queue is full. Only when the queue is full, the flow be rerouted to the MAN node with the shortest queue length.

4.3.1 Shortest-queue Node First Routing Strategy (Sq-first Strategy)

In DPFRA with sq-first strategy, all the $N-1$ nodes in the MAN can be regarded as an overall single node for queueing delay calculation and all wavelengths will be shared as a single group. So this is the merged queueing model described in Section 4.1.3 with $N-1$ I.I.D. nodes. It can thus be modeled as an $M/M/(N-1)x/(N-1)m$ queueing system with arrival rate $(N-1)\lambda$. So the blocking probability of DPFRA with sq-first strategy is

$$P_{sq} = P_{B_{N-1}} \tag{4.23}$$

The average queueing delay of DPFRA with sq-first strategy is

$$\tau_{sq} = \tau_{Q_{N-1}} \tag{4.24}$$

4.3.2 Traffic-receiving Node First Routing Strategy (Tr-first Strategy)

DPFRA with tr-first strategy can be regarded as a two-stage queueing system. The first stage is the queueing system at the traffic-receiving node. The queue of the traffic-receiving node is an $M/M/x/m$ queue. The second stage is the queueing system of all the other $N-2$ nodes. Since all other $N-2$ nodes will share all their wavelengths together as a single group, they can be considered as a merged queue of $N-2$ I.I.D. nodes. The queue at the second state can be modeled as an $M/M/(N-2)x/(N-2)m$ queue with arrival rate $(N-2)\lambda$.

To get the blocking probability of DPFRA with tr-first strategy, denote event B_1 as the event that one node (traffic-receiving node) are blocked and event B_{N-2} as the event that $N-2$ nodes (all non-traffic-receiving nodes) are blocked. The blocking probability of DPFRA with tr-first strategy is therefore as

$$P_{tr} = P[B_{N-2}|B_1]P[B_1] \tag{4.25}$$

To obtain the closed-form analytic result for P_{tr} is very difficult, since the conditional probability $P[B_{N-2}|B_1]$ is hard to calculate analytically. However, we can prove the following bound:

The blocking probability of DPFRA with tr-first strategy is not less than the blocking probability of DPFRA with sq-first strategy with the same network resources in terms of wavelengths and schedule holder size. That is

$$P_{tr} \geq P_{sq} \tag{4.26}$$

This is because DPFRA with sq-first strategy will use the earliest available time and wavelength slot to the destination whereas DPFRA with tr-first strategy will not place the sessions in the slots with the shortest wait. A more detailed discussion is given in Appendix A.3.

QSA always has a larger blocking probability than that of DPFRA with tr-first

strategy, since enabling rerouting decreases the blocking probability. That is

$$P_{tr} \leq P_{qsa} \tag{4.27}$$

The average queueing delay of DPFRA with tr-first strategy consists of two parts, the delay τ_{Q_1} of the traffic-receiving node and the delay $\tau_{Q_{N-2}}$ of the other nodes with the shortest queue. Though it is hard to get the accurate conditional blocking probability $P[B_{N-2}|B_1]$, we can use $P_{B_{N-2}}$ as an approximation when we calculate the delay and then do the normalization. The flow waits in the traffic-receiving node with probability $1 - P_{B_1}$ and waits in the shortest-queue node with probability $P_{B_1}(1 - P_{B_{N-2}})$, where P_{B_1} is the blocking probability of the traffic-receiving node and $P_{B_{N-2}}$ is the blocking probability of all the other $N - 2$ nodes. Therefore, the normalized queueing delay of DPFRA with tr-first strategy is approximately

$$\tau_{tr} \approx \frac{(1 - P_{B_1})\tau_{Q_1} + P_{B_1}(1 - P_{B_{N-2}})\tau_{Q_{N-2}}}{(1 - P_{B_1}) + P_{B_1}(1 - P_{B_{N-2}})} \tag{4.28}$$

4.4 Routing Architecture Performance Comparisons

To further evaluate the performance of the different routing architectures described in this chapter, we compare the analytical results of the different architectures graphically. In this section, the comparisons of schedule holder size, normalized delay, blocking probability, and load will be shown. Without loss of generality, in our analysis we assume a Petersen Graph as the underlying static network topology.

4.4.1 Queue Size Comparisons

Figure 4-2 shows the comparisons of schedule holder size with different wavelengths assigned to each node versus load with the requirement that the blocking probability is within 10^{-2}. From it, we can see when the load tends to 1, the size of schedule holder tends to infinity. It agrees with the fact the blocking probability increases dramatically when the load tends to 1 for the system with the fixed schedule holder

size. So the size of the schedule holder needs to increase at the same pace if we want to satisfy the blocking probability requirement. Besides, with the same load, the schedule holder size decreases when the number of wavelengths assigned to each node increases. For $x = 1$, the schedule holder is always required to meet the requirement that the blocking probability requirement is less than 10^{-2}. For $x = 2$, the schedule holder is required when $\rho > 0.1$. When $\rho \approx 0.2$, each wavelengths requires 1 holder, and the total schedule holder size of a node is 2. For $x = 10$, the schedule holder is required when $\rho > 0.45$. When $\rho \approx 0.8$, the schedule holder size of a node is 10, meaning that 1 holder per wavelength is sufficient.

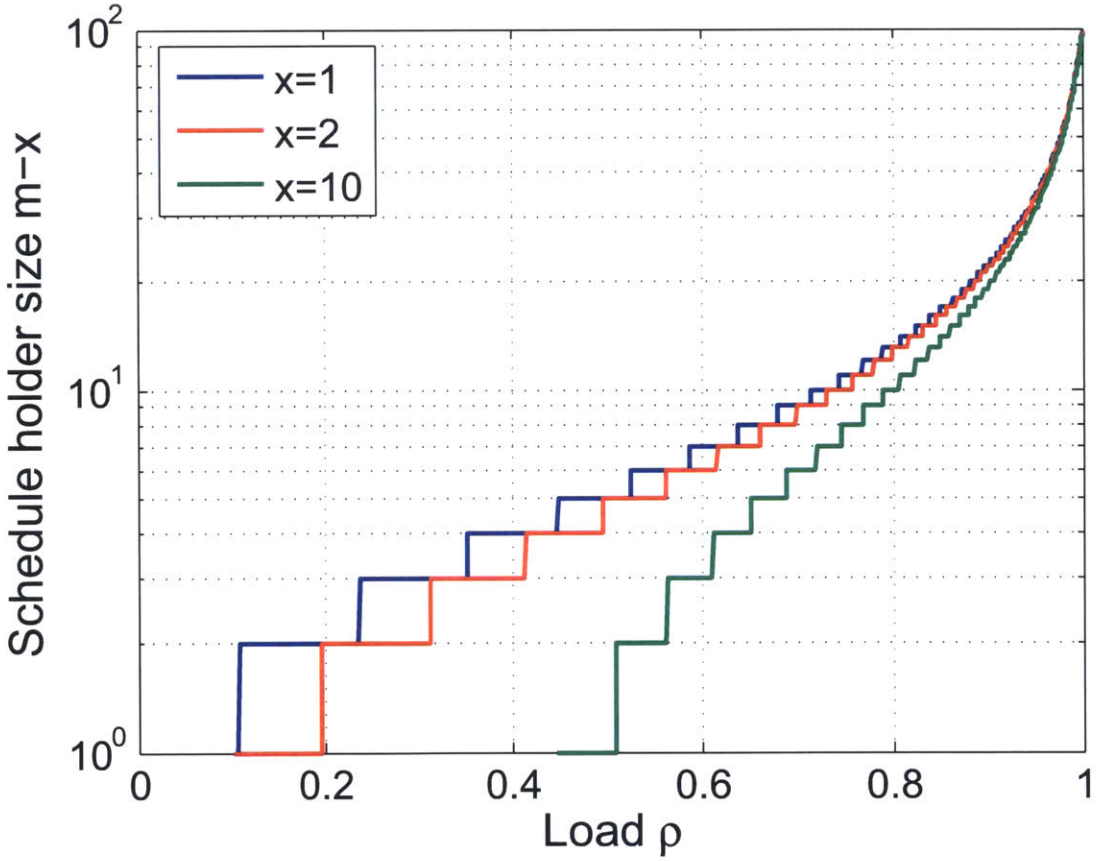


Figure 4-2: The comparisons of the schedule holder size per node $m - x$ versus load ρ. Blocking probability is 0.01.

4.4.2 Normalized Delay Comparisons

Figure 4-3 shows the comparisons of the normalized delays with different k wavelengths of traffic versus $\overline{F}$ of QSA with primary no-rerouting strategy and DPFRA with two rerouting strategies, sq-first strategy and tr-first strategy. To make the comparison, here we fix the schedule holder size. We set the number of flows in the queueing system to be five times of the number of wavelengths at each node, including the flow(s) being transmitted. The normalized delay τ consists of both the transmission delay and the queueing delay. τ is the normalized delay in terms of the time of transmission of one session.

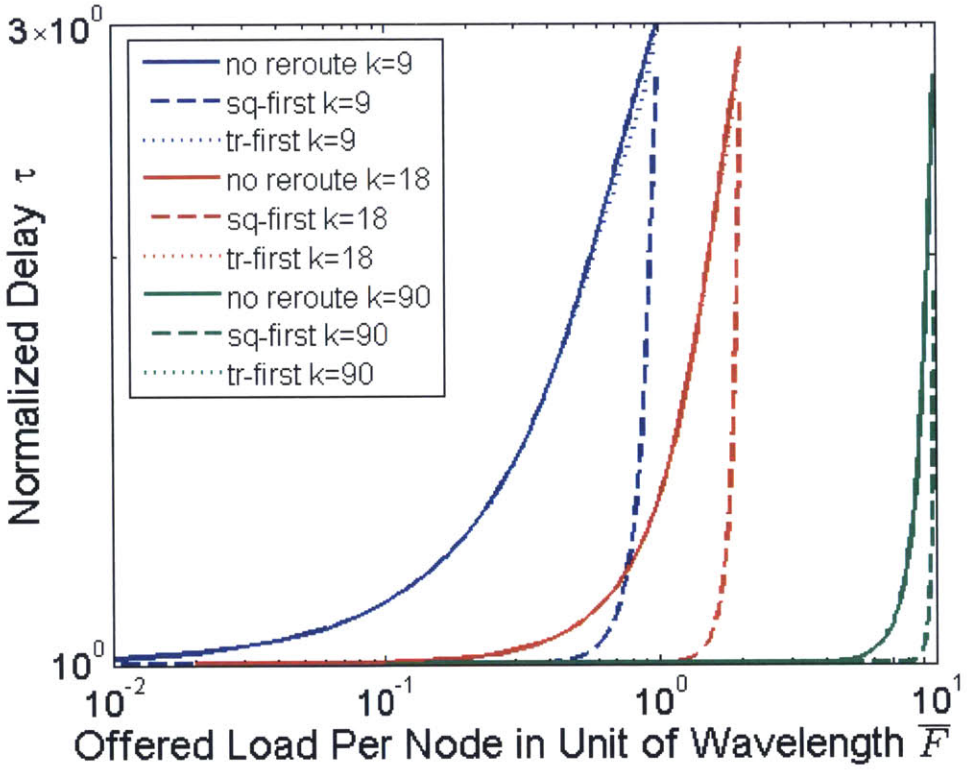


Figure 4-3: The comparisons of normalized delays τ with different total number of wavelengths assigned to the MAN k (traffic in number of wavelengths) versus offered load per node in unit of wavelength $\overline{F}$. τ is the normalized delay in terms of the time of transmission of one session. The number of flows in the queueing system is five times of the number of wavelengths at each node, including the flow(s) being transmitted in the wavelength channel(s).

In Figure 4-3, the normalized delay grows exponentially with the increase of $\overline{F}$. DPFRA with sq-first strategy has the shortest delay compared to QSA(with primary no-rerouting strategy) and DPFRA with tr-first strategy. With increasing traffic (and thus total assigned wavelength k), the differences between QSA and DPFRA with tr-first strategy are insignificantly small. Therefore, we can conclude that DPFRA with tr-first strategy has only a slight enhancement in reducing the delay. In contrast, DPFRA with sq-first strategy can decrease the queueing delay more substantially.

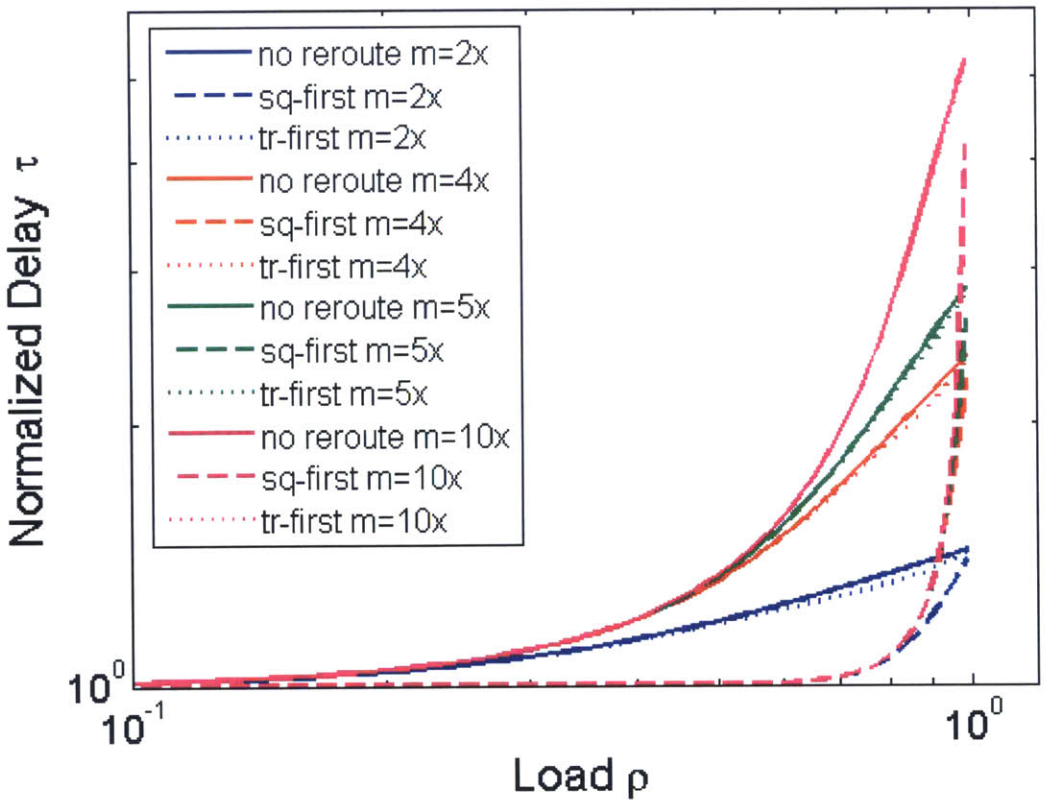


Figure 4-4: The comparisons of the normalized delay τ with different m (the total number of flows in the schedule holders and the wavelength channels of each node) versus offered load per node in unit of wavelength $\overline{F}$. τ is the normalized delay in terms of the time of transmission of one session. $x = 2$ wavelengths are assigned to each node.

Figure 4-4 shows the comparisons of the normalized delay with different schedule holder size plus the number of wavelengths m versus load ρ of QSA and DPFRA with two rerouting strategies. To make the comparison, here we fix the total number

of wavelengths assigned to the network k. We assign two wavelengths to each node. With the same load, the normalized delay increases with the increasing schedule holder size as expected. Comparing different strategies, we still have the same results that DPFRA with sq-first strategy can decrease the queueing delay substantially compared to the QSA (no-rerouting one), while the enhancement in reducing the delay of DPFRA with tr-first strategy is insignificantly small. Notice that the limits of the normalized delay is different for different schedule holder sizes. The smaller the of the schedule holder is, the smaller the limit of the normalized delay is. However, the delay due to re-entry into the system is not included in this result. The overall delay if that is counted will actually be worse.

4.4.3 Blocking Probability Comparisons

Figure 4-5 shows the comparisons of the blocking probability with different k wavelengths of traffic versus load ρ of QSA and DPFRA with two rerouting strategies. Again, we fix the number of flows in the queueing system to be five times of the number of wavelengths at each node, including the flow being transmitted. From it, the blocking probability decreases with increasing k. When the load ρ increases, the blocking probability increases as expected. When the load ρ tends to 1, all the blocking probabilities tend to 0.1, which is too large to be acceptable. Compared to QSA (the no-rerouting situation), DPFRA with sq-first strategy can greatly reduce the blocking probability. When the load tends to 1, the blocking probability of DPFRA with sq-first strategy tends to 0.01, which is 10% of that of QSA. For larger k, the decrease in the blocking probability for the same load ρ becomes obvious. This is due to statistical multiplexing making the normalized spread around the mean smaller resulting in a smaller blocking probability. Now the blocking probability of DPFRA with sq-first strategy is small enough to be within the tolerance. Thus for heavy traffic which is the case of interest for OFS, we need not consider the blocking probability as an evaluation criterion, provided we are mindful in keeping it low.

Figure 4-6 shows the comparisons of blocking probabilities with different schedule holder size plus the number of wavelengths m versus load ρ of different routing

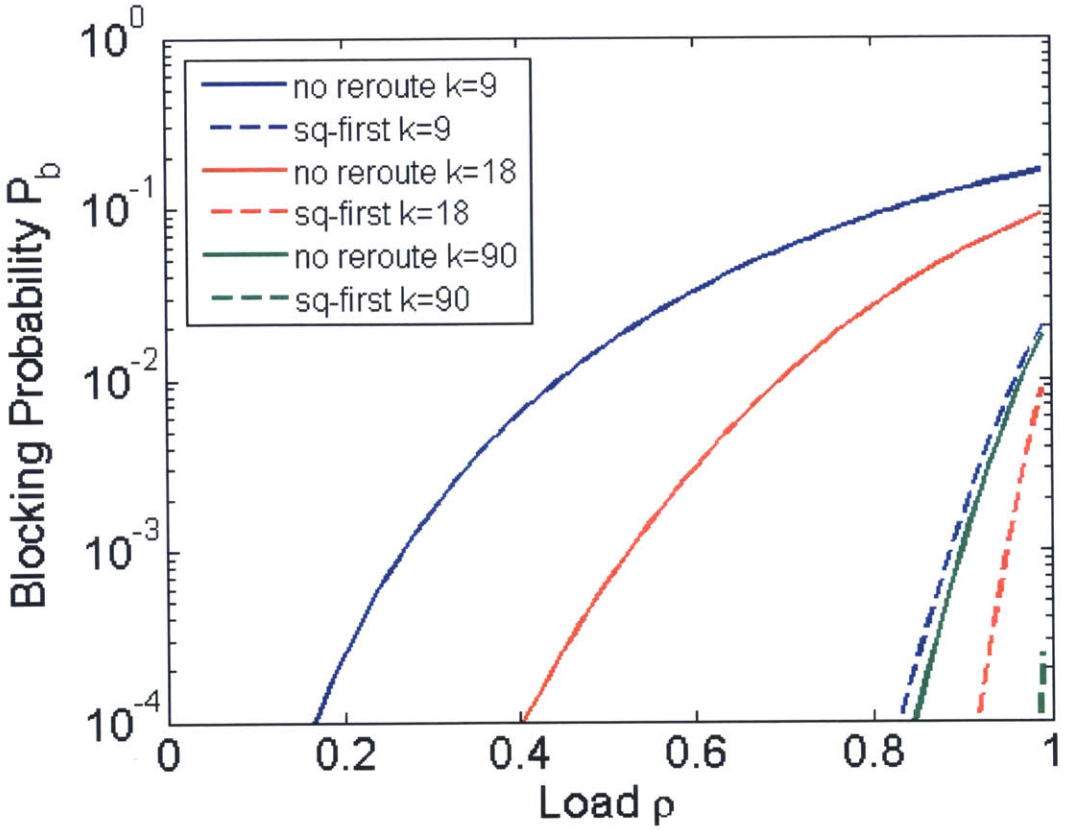


Figure 4-5: The comparisons of blocking probabilities P_b with different total number of wavelengths assigned to the MAN k (traffic in number of wavelengths) versus load ρ. The number of flows in the queueing system is five times of the number of wavelengths at each node, including the flow(s) being transmitted in the wavelength channel(s).

architectures. Again, here we fix the total number of wavelengths assigned to the network k, assigning two wavelengths to each node. We can see with the increase of the size of schedule holders, the blocking probability decreases. It agrees with the design principle of schedule holder size and the results in Figure 4-2. Also, it still holds that DPFRA with sq-first strategy can greatly reduce the blocking probability. Notice that the system with smaller number of schedule holder size has a higher blocking probability. It agrees with the analysis of Figure 4-4 that the overall delay will actually be worse if we count the delay due to the re-entry into the system, since the blocking probability of the first entry becomes higher.

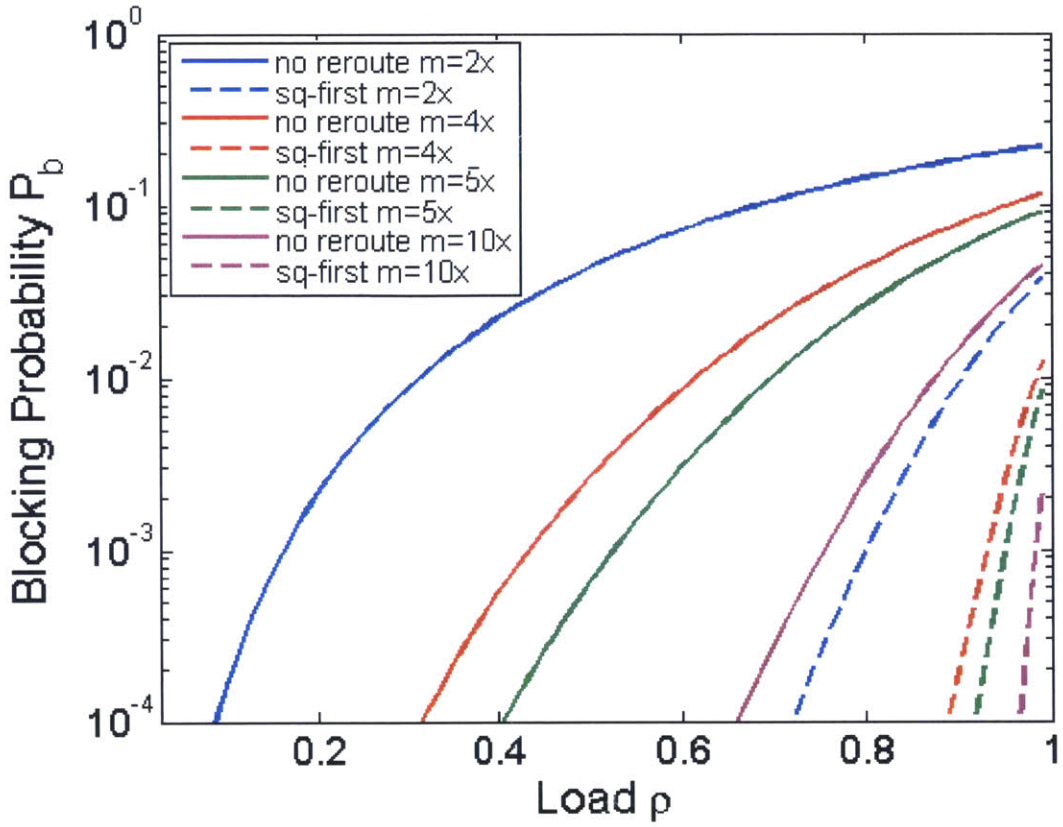


Figure 4-6: The comparisons of blocking probabilities P_b with different m (the total number of flows in the schedule holders and the wavelength channels of each node) versus load ρ. $x = 2$ wavelengths are assigned to each node.

4.4.4 Load Comparisons

Figure 4-7 shows the maximum load of both QSA and DPFRA with sq-first strategy versus the number of wavelength traffic per node with a given blocking probability. From Figure 4-7, we can see that rerouting guarantees higher load compared to no-rerouting for the same blocking probability. Also, for blocking probability ~ 0.01, the network tends to be nearly fully loaded in DPFRA when the number of wavelength per node x is larger than 2. Therefore, DPFRA with sq-first strategy has the greater efficiency which should be intuitively obvious.

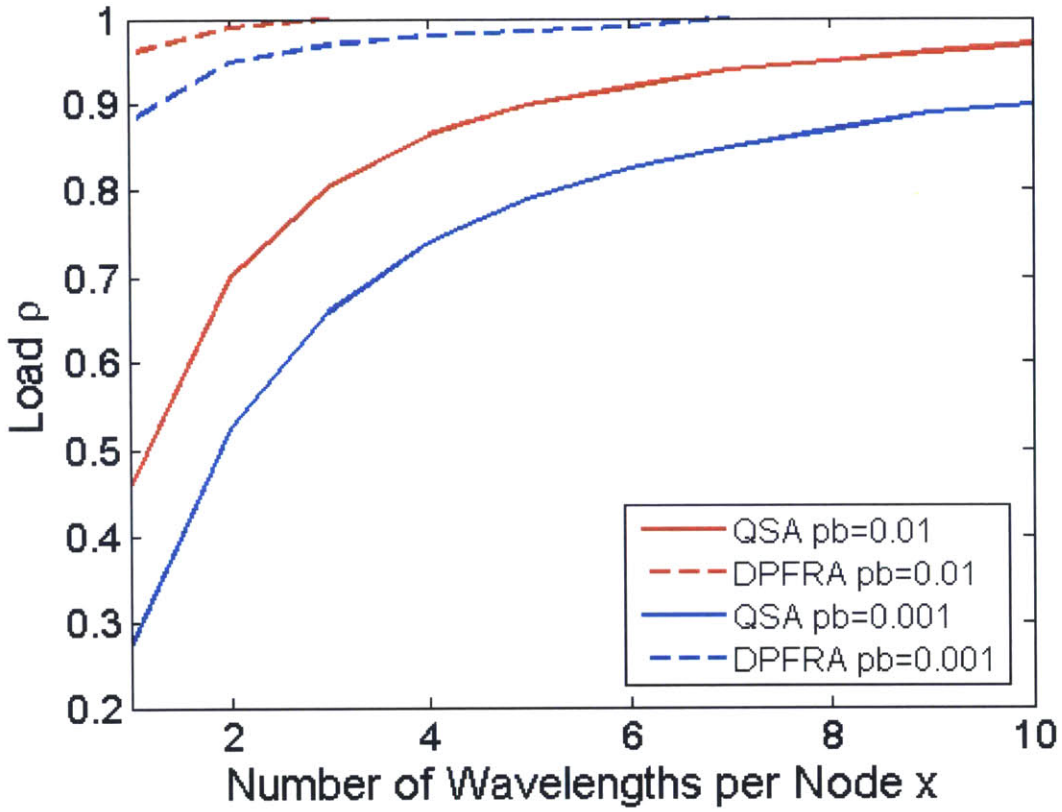


Figure 4-7: Load comparisons between QSA and DPFRA with sq-first strategy versus number of wavelengths per node x with same blocking probability requirements. The blocking probability requirement is 0.01 or 0.001.

4.5 Hybrid Architecture

The hybrid architecture lies between the two extremes of the QSA and DPFRA architectures. In QSA, rerouting is never allowed. In DPFRA, per flow rerouting is always allowed. In the hybrid architecture, rerouting is allowed for a fraction of users or time periods. The decision on whether or not to allow rerouting for each flow is based on the real-time network scenarios and requirements. For example, rerouting may be allowed only when the queue lengths pass a pre-determined threshold. The aim of the hybrid architecture is explore the space between the two extreme architectures presented in this chapter.

From the analytical results in Section 4.4, it is clear that with the same network resources, DPFRA which enables per flow rerouting can decrease both the delay and

the blocking probability of flows compared to QSA. So one example of the hybrid architecture can be as follows: when the traffic amount is high, we enable per flow rerouting to make sure the blocking probability is within a tolerable level; when the traffic amount is low, we switch back to QSA to simplify the network control and configuration and thus reduce operating costs. In this way, the hybrid architecture tries to take advantages of both extreme architectures. We will discuss the performance of the hybrid architecture after we generate a cost model for the network and gain a better understanding of the two extreme architectures, QSA and DPFRA.

4.6 Summary of MAN Routing Architecture

In this section, we summarize the design of the MAN routing architecture, including the schedule holder size design and the routing strategy design. We generate the queueing models for the different routing architectures, mainly the two extreme cases, QSA and DPFRA. For DPFRA, we design two rerouting strategies, sq-first strategy and tr-first strategy. The summary of the modeling results is shown in Table 4.1 and Table 4.2. Also, the idea of the hybrid architecture is proposed, which will be discussed in the following chapter.

From the analytical results in Section 4.4, we find DPFRA has a lower delay and lower blocking probability compared to QSA. However, rerouting in DPFRA requires higher computational complexity due to a more complex routing decision at each nodes. Furthermore, sending additional routing decisions to MAN nodes will result in a higher volume of control traffic.

Table 4.1: Summary of Queueing Model of Different Routing Architectures

Routing Architecture	Queueing Model
QSA	$M/M/x/m$
DPFRA(sq-first)	$M/M/(N-1)x/(N-1)m$
DPFRA(tr-first)	$M/M/x/m + M/M/(N-2)x/(N-2)m$

Table 4.2: Summary of Average Queueing Delay and Blocking Probability of Different Routing Architectures

Routing Architecture	Average Queueing Delay	Blocking Probability
QSA	$\tau_{qsa} = \tau_{Q_1}$	$P_{qsa} = P_{B_1}$
DPFRA(sq-first)	$\tau_{sq} = \tau_{Q_{N-1}}$	$P_{sq} = P_{B_{N-1}}$
DPFRA(tr-first)	$\tau_{tr} = \frac{(1-P_{B_1})\tau_{Q_1}+P_{B_1}(1-P_{B_{N-2}})\tau_{Q_{N-2}}}{(1-P_{B_1})+P_{B_1}(1-P_{B_{N-2}})}$	$P_{tr} = P[B_{N-2}\|B_1]P[B_1]$

Comparing the two rerouting strategies in DPFRA, we see DPFRA with tr-first strategy only makes slight improvements compared to QSA in the normalized delay comparison. As the number of wavelengths or the size of the schedule holder increases, such enhancements are insignificant and negligible. Also, bound 4.26 shows that the blocking probability of DPFRA with tr-first strategy is always higher than that of DPFRA with sq-first strategy. So in the following chapters, we will no longer consider DPFRA with tr-first strategy.

Chapter 5

Parametric cost model for MAN

In this chapter, we build the parametric cost model used to evaluate the costs for architectures with different topologies and different routing architectures. Network cost can be divided into two categories: capital expenditure and operating expenditure. The capital expenditure includes the cost to construct the physical network and the cost of the network components. The operating expenditure includes the cost of power consumption and cooling of network components, and the cost of network management/control and maintenance [7][10]. Another component of network cost is what is associated with dynamic network management and control. The dynamic control architecture involves routing decisions that is not considered in the cost of the physical topology. In [7], Guan provides the parametric network cost model of the static network costs for MAN based on uniform all-to-all deterministic traffic demand. It only includes the transceiver cost, fiber connection cost, and the switching cost. In this work, we will consider both the static network costs and the dynamic control costs based on the all-to-one stochastic traffic model. In this work we do not consider ALL the operating costs, particularly those that involve human.

Based on the above assumptions, the total cost of the MAN mainly consists of five components:

1. transceiver cost

2. fiber connection cost

3. switch cost

4. control traffic cost

5. computational complexity cost

The transceiver cost, the fiber connection cost, and the switch cost are in the category of static network costs, while the control traffic cost and the computational complexity cost are in the category of the dynamic control costs. We assume the average life of the switches is five years. So here we perform the cost comparison in terms of one life cycle of the switches, which is five years. The total costs in five years of QSA and DPFRA are denoted as C_{qs} and C_{dr}, respectively.

The network architecture topology, the traffic demands, and the cost coefficients are the three factors driving the cost model. The cost coefficients are derived from the marginal costs of network components [7]. Therefore, the network costs are modeled as functions of the number of nodes N, the node degree Δ, the number of wavelengths assigned to each node x, the offered load per node in unit of wavelengths $\overline{F}$, the load ρ, and the different cost coefficients.

5.1 Cost Model for Quasi-static Architecture

The cost for QSA contains five parts: transceiver cost C_{qtr}, fiber connection cost C_{qfb}, switch cost C_{qsw}, control traffic cost C_{qct}, and computational complexity cost C_{qcx}.

5.1.1 Transceiver Cost

For all-to-one traffic, each node sends traffic to the hub and receives traffic from the hub. So a total of $(N-1)$ transceivers are needed for all the $(N-1)$ nodes. For the hub, we assume no transceiver is needed, since the traffic will be directly transmitted to the switches in the WAN through this gateway. Therefore, the cost of transceiver is [7]

$$C_{qtr} = \alpha_{tr}(N-1) \tag{5.1}$$

where α_{tr} the cost per transceivers.

5.1.2 Fiber Connection Cost

From [8], since the distances in the MAN are much shorter than that in the WAN, amplifiers or regenerators are not used. Besides, the fiber connection cost only includes the capital expenditure of the fibers and no operation expenditure is included, since fibers are not active network components. For a single mode fiber, it can support hundreds of wavelengths. So we assume that the changes of the number of wavelengths will not affect the fiber connection cost. Therefore, the cost of fiber connection for a certain network topology is [7]

$$C_{qfb} = \alpha_f \Delta N \tag{5.2}$$

where α_f is the marginal cost of a new fiber connection.

5.1.3 Switch Cost

As stated in [8], the cost of a switch is a linear function of the number of switch ports (for at least low to modest number port count switches), which is determined by the amount of traffic going through the switch. This is a close enough approximation for the purpose of this thesis though for large port counts (> 1000) the cost becomes nonlinear. Each node can generate at most x wavelengths of traffic and each wavelength of traffic is sent from the node to the hub via H_{min} hops, where H_{min} is the average minimum hop distance of the MAN topology [8]. When the traffic reaches one node in the path, it occupies a port pair in that node. Since there are a total of $(H_{min}+1)$ nodes, the number of switch port pairs consumed in one lightpath is $x(H_{min}+1)$. The total number of switch port pairs for $(N-1)$ identical nodes is $(N-1)x(H_{min}+1)$. Therefore, the switch cost is

$$C_{qsw} = \alpha_{s_1}(N-1)(H_{min}+1)x + \alpha_{s_2}(N-1)(H_{min}+1)t\frac{R}{L}\overline{F}$$
$$= (N-1)(H_{min}+1)(\alpha_{s_1}x + \alpha_{s_2}t\frac{R}{L}\overline{F}) \quad (5.3)$$

where α_{s_1} is the cost coefficient of the capital expenditure of switching, α_{s_2} is the cost coefficient of the operating expenditure of switching, and t is the operating time interval.

5.1.4 Control Traffic Cost

The control traffic is the information that scheduler sends to the nodes for preparation and reconfiguration. It consumes resources when the scheduler has to send the control traffic to the source node to announce the schedule and reconfigure that node. Since there are a total of $(N-1)$ source nodes, the control traffic cost is

$$C_{qct} = \alpha_{ct}(N-1)t\frac{R}{L}\overline{F} \quad (5.4)$$

where α_{ct} is the cost of control traffic for one flow.

5.1.5 Computational Complexity Cost

Computational complexity is a part of the operating cost. Once the routing table is established, the scheduler uses table look-up and no rerouting decision is needed. So the complexity in the QSA is $O(1)$. The computational complexity cost is

$$C_{qcx} = \alpha_{cx}O(1)(N-1) \quad (5.5)$$

where α_{cx} is the cost coefficient of complexity.

5.1.6 Total Network Cost

The total network cost for QSA is:

$$
\begin{aligned}
C_{qs} &= C_{qtr} + C_{qfb} + C_{qsw} + C_{qct} + C_{qcx} \\
&= \alpha_{tr}(N-1) + \alpha_f \Delta N + (N-1)(H_{min}+1)(\alpha_{s_1} x + \alpha_{s_2} t \frac{R}{L} \overline{F}) \\
&+ \alpha_{ct}(N-1) t \frac{R}{L} \overline{F} + \alpha_{cx} O(1)(N-1)
\end{aligned} \tag{5.6}
$$

In realistic situations, the sum of C_{qtr},C_{qfb} and C_{qsw} takes up approximately 90% of the cost. C_{qct} and C_{qcx} are relatively insignificant and can be ignored. This is because the tunneled architecture for OFS reduces control plane traffic and processing complexity by orders of magnitude [13]. So the formulation can be simplified as

$$
\begin{aligned}
C_{qs} &\approx C_{qtr} + C_{qfb} + C_{qsw} \\
&\approx \alpha_{tr}(N-1) + \alpha_f \Delta N + (N-1)(H_{min}+1)(\alpha_{s_1} x + \alpha_{s_2} t \frac{R}{L} \overline{F})
\end{aligned} \tag{5.7}
$$

5.2 Cost Model for Dynamic Per Flow Routing Architecture

The cost for DPFRA contains five parts: transceiver cost C_{dtr}, fiber connection cost C_{dfb}, switch cost C_{dsw}, control traffic cost C_{dct}, and computational complexity cost C_{dcx}.

5.2.1 Transceiver Cost

The transceiver cost in DPFRA are the same as that in QSA due to the unchanged network topology. Therefore, the transceiver cost is [7]

$$
C_{dtr} = \alpha_{tr}(N-1) \tag{5.8}
$$

5.2.2 Fiber Connection Cost

Similarly, the fiber connection cost in DPFRA are the same as that in QSA due to the unchanged network topology. The cost of fiber connection is [7]

$$C_{dfb} = \alpha_f \Delta N \tag{5.9}$$

5.2.3 Switch Cost

Rerouting in DPFRA requires additional cost over QSA cost which is determined by the number of rerouting hops. The difference of the average minimum number of rerouting hops h_{min} with the average minimum hops distance H_{min} [7] is that the hub is not included in the rerouting topology. In other words, no traffic will be sent from or rerouted to the hub when rerouting is performed. Therefore, we cannot simply use H_{min} to compute the average minimum number of hops.

For each node, β of average traffic $\overline{F}$ has to be rerouted from the traffic-receiving node to the shortest-queue node passing through h_{min} hops. Since it is equiprobable for each node to be the node with the shortest queue, we have

$$h_{min} = \frac{1}{N-2}[(N-1)H_{min} - \frac{\Delta}{N-1}\sum_{i=1}^{D-1} i(\Delta-1)^{i-1} - (1 - \frac{\Delta}{N-1}\frac{1-(\Delta-1)^{D-1}}{2-\Delta})D] \tag{5.10}$$

The detailed derivation is shown in Appendix B.1.

The total number of extra switch port pairs used for rerouting is $(N-1)\beta x h_{min}$, since additional β amount of traffic requires extra $\beta x h_{min}$ switch port pairs. So the total switch cost in the dynamic per flow switching is

$$\begin{aligned} C_{dsw} &= \alpha_{S_1}(N-1)(H_{min} + \beta h_{min} + 1)x + \alpha_{s_2}(N-1)(H_{min} + \beta h_{min} + 1)t\frac{R}{L}\overline{F} \\ &= (N-1)(H_{min} + \beta h_{min} + 1)(\alpha_{s_1}x + \alpha_{s_2}t\frac{R}{L}\overline{F}) \end{aligned} \tag{5.11}$$

β is the average fraction of the average total traffic needed to be rerouted in the rerouting strategy. Since all the nodes are symmetric, for large N, we have

$$\beta = \frac{N-2}{N-1} \approx 1 \tag{5.12}$$

5.2.4 Control Traffic Cost

Compared to QSA, DPFRA requires more control traffic to manage the extra per flow rerouting traffic. For one node, such extra control traffic to arrange the rerouting path for βx wavelengths of traffic has to be sent to all the nodes between the traffic-receiving node and the newly selected shortest-queue node. We assume the control traffic amount is proportional to the amount of data traffic. Since the number of hops it passes through is h_{min}, so the number of nodes on the rerouting path is $(h_{min}+1)$. For all $(h_{min}+1)$ nodes, the total amount of extra control traffic is $(N-1)\beta x(h_{min}+1)$. Therefore, the control traffic cost is

$$\begin{aligned} C_{dct} &= \alpha_{ct}(N-1)t\frac{R}{L}\overline{F} + \alpha_{ct}(N-1)\beta(h_{min}+1)t\frac{R}{L}\overline{F} \\ &= \alpha_{ct}(N-1)[1+\beta(h_{min}+1)]t\frac{R}{L}\overline{F} \end{aligned} \tag{5.13}$$

5.2.5 Computational Complexity Cost

The computational complexity of DPFRA are mainly generated in three steps: 1. finding the shortest queue; 2. assigning the wavelength; 3. looking up the routing table. Both step 1 and step 2 require extra computation compared to QSA. Since the shortest path algorithm is adopted, the complexity of the three steps are $O(N)$, $O(1)$, and $O(1)$, respectively. Therefore, the computational complexity cost is

$$C_{dcx} = \alpha_{cx}O(N)(N-1) \tag{5.14}$$

5.2.6 Total Network Cost

The total network cost for DPFRA is:

$$\begin{aligned} C_{dr} &= C_{dtr} + C_{dfb} + C_{dsw} + C_{dct} + C_{dcx} \\ &= \alpha_{tr}(N-1) + \alpha_f \Delta N + (N-1)(H_{min} + \beta h_{min} + 1)(\alpha_{s_1} x + \alpha_{s_2} t \frac{R}{L} \overline{F}) \\ &+ \alpha_{ct}(N-1)[1 + \beta(h_{min} + 1)] t \frac{R}{L} \overline{F} + \alpha_{cx} O(N)(N-1) \end{aligned} \tag{5.15}$$

Similarly, the cost of DPFRA can be simplified as the sum of the transceiver cost, fiber connection cost, and the switch cost. So

$$\begin{aligned} C_{dr} &\approx C_{dtr} + C_{dfb} + C_{dsw} \\ &= \alpha_{tr}(N-1) + \alpha_f \Delta N + (N-1)(H_{min} + \beta h_{min} + 1)(\alpha_{s_1} x + \alpha_{s_2} t \frac{R}{L} \overline{F}) \end{aligned} \tag{5.16}$$

5.3 Overall MAN Architecture Cost Comparisons

In this section, we compare the overall MAN architecture cost between QSA and DPFRA with sq-first strategy (DPFRA with tr-first strategy is eliminated based on the performances in terms of delay and blocking probability). For consistency, we keep employing Petersen Graph to demonstrate the results. Since the costs of transceivers and fiber connections are the same for the two architectures with the same traffic, we ignore this part of the cost and only compare the costs of switches.

For fiber connect cost, from [10], the marginal cost of a new fiber connection is in the range of $\$2k/km$ to $\$25k/km$. Since a typical fiber is $5km$ to $20km$, the estimated value of α_f lies between $\$10,000/fiber$ and $\$500,000/fiber$.

For switch cost, the capital expenditure of an 8×8 OXC switch is $\$10,000/port\ pair$ for five years, so $\alpha_{s_1} = \$10,000/port\ pair$. The operating expenditure is $\$3,700/port\ pair$ for five years. We assume each flow takes $1s$ per port pair, then $\alpha_{s_2} = \$2.35 \times 10^{-5}/(flow \cdot port\ pair)$. The total life cycle of the hardware is assumed to be

$t = 5$ *years*.

5.3.1 Cost Comparisons with Same Delay Requirement

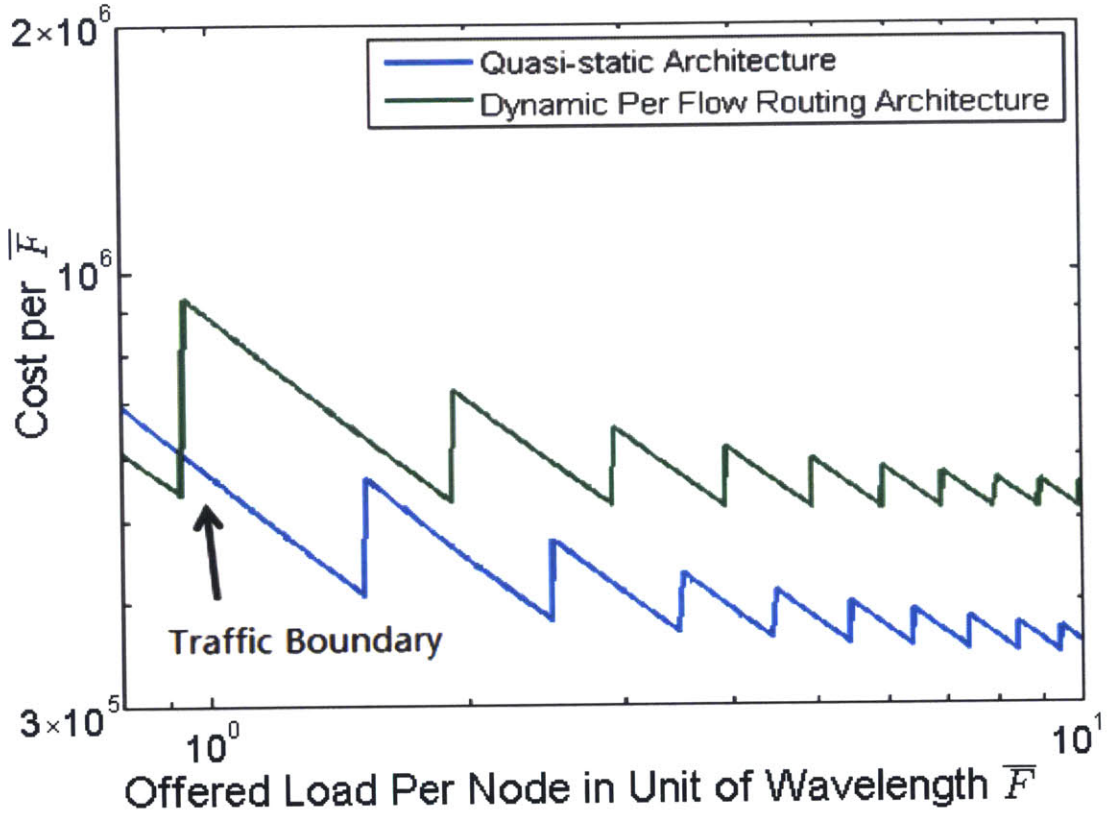


Figure 5-1: Cost comparisons between QSA and DPFRA with sq-first strategy versus offered load per node in unit of wavelength $\overline{F}$ with same delay requirement. The average queueing delay requirement of each flow is the transmission time of one flow. Parameter Assumptions: $\alpha_{s_1} = \$10,000/port\ pair$, $\alpha_{s_2} = \$2.35\times10^{-5}/(flow{\cdot}port\ pair)$, $t = 5$ *years*.

For our comparison we keep queueing delays requirement of both QSA and DPFRA with sq-first strategy to be equal to the transmission times of one flow. The comparison between the two architectures is shown in Figure 5-1. The zigzag shapes of the costs are due to the integer property of the number of the wavelengths and the number of switch ports. The lines of cost per unit traffic for the two architectures crosses approximately at $\overline{F} = 0.8$. This intersection is denoted as the traffic boundary and it will move depending on the delay. The intersection shows the transition between

the two preferred architectures. Note that when the number of wavelengths of offered traffic is less than the traffic boundary, the cost of DPFRA with sq-first strategy is cheaper. When it goes beyond the traffic boundary, QSA is cheaper. When $\overline{F}$ tends to 10, QSA tends to save cost by 35.7%.

5.3.2 Cost Comparisons with Same Blocking Probability Requirement

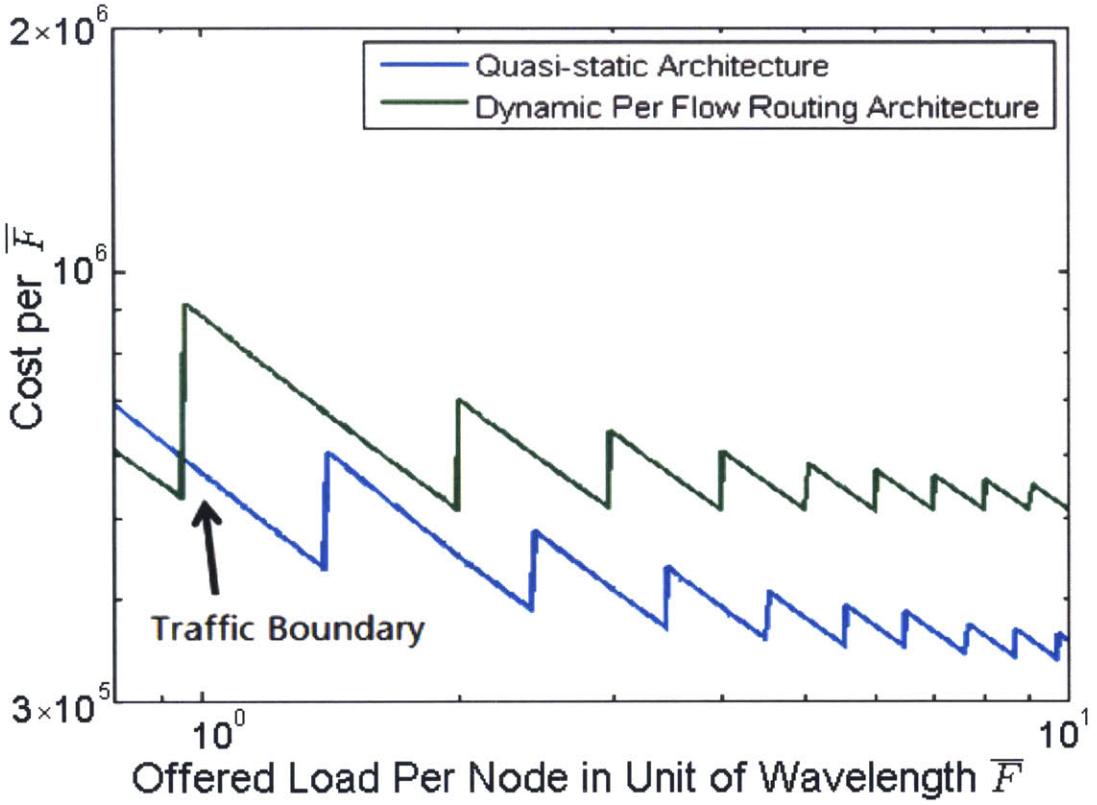


Figure 5-2: Cost comparisons between QSA and DPFRA with sq-first strategy versus offered load per node in unit of wavelength $\overline{F}$ with same blocking probability requirement. The blocking probability requirement is 0.01. Parameter Assumptions: $\alpha_{s_1} = \$10,000/port\ pair$, $\alpha_{s_2} = \$2.35 \times 10^{-5}/(flow \cdot port\ pair)$, $t = 5\ years$.

As an example we keep both the blocking probabilities in QSA and DPFRA with sq-first strategy to be 0.01, the plot of cost comparisons between the two architectures is shown in Figure 5-2. Similar to the result in the cost comparisons with the same delay, the cost curves of the two architectures intersect at the traffic boundary, which denotes the preference of the architectures in term of cost. Here, the crossover is approximately at $\overline{F} = 1$. When the offered traffic is light, DPFRA with sq-first strategy is cheaper; when the traffic is heavy, QSA is cheaper. QSA is cheaper by 31.4% when $\overline{F}$ tends to 10. Note that the blocking probability is below 0.1% in the DPFRA with sq-first strategy when the number of wavelengths assigned to each node is greater than 8.

5.3.3 Cost Boundaries

Figure 5-3 shows the cost boundary of QSA and DPFRA with sq-first strategy versus different delay requirements and different $\overline{F}$. For any point on the plot, one of the two architectures has the lower cost. In Figure 5-3, the blocking probability is set to 0.01. DPFRA with sq-first strategy is preferred when the delay requirement is stringent and $\overline{F}$ is low. With either the relaxation of delay requirement or increase of $\overline{F}$, QSA is cheaper to implement than DPFRA with sq-first strategy. When the product of $\overline{F}$ and the normalized delay τ is approximately larger than 2, QSA is the only choice, indicating DPFRA with sq-first strategy always costs more when the traffic volume is modest or high. This is mostly due to the fact that rerouting uses more resources in larger hop number nad more switch ports, adding to the cost of the architecture. Note that with the increase of $\overline{F}$, the difference of the cost between the two extreme architectures becomes larger, showing that QSA is always preferred in the high offered load situation.

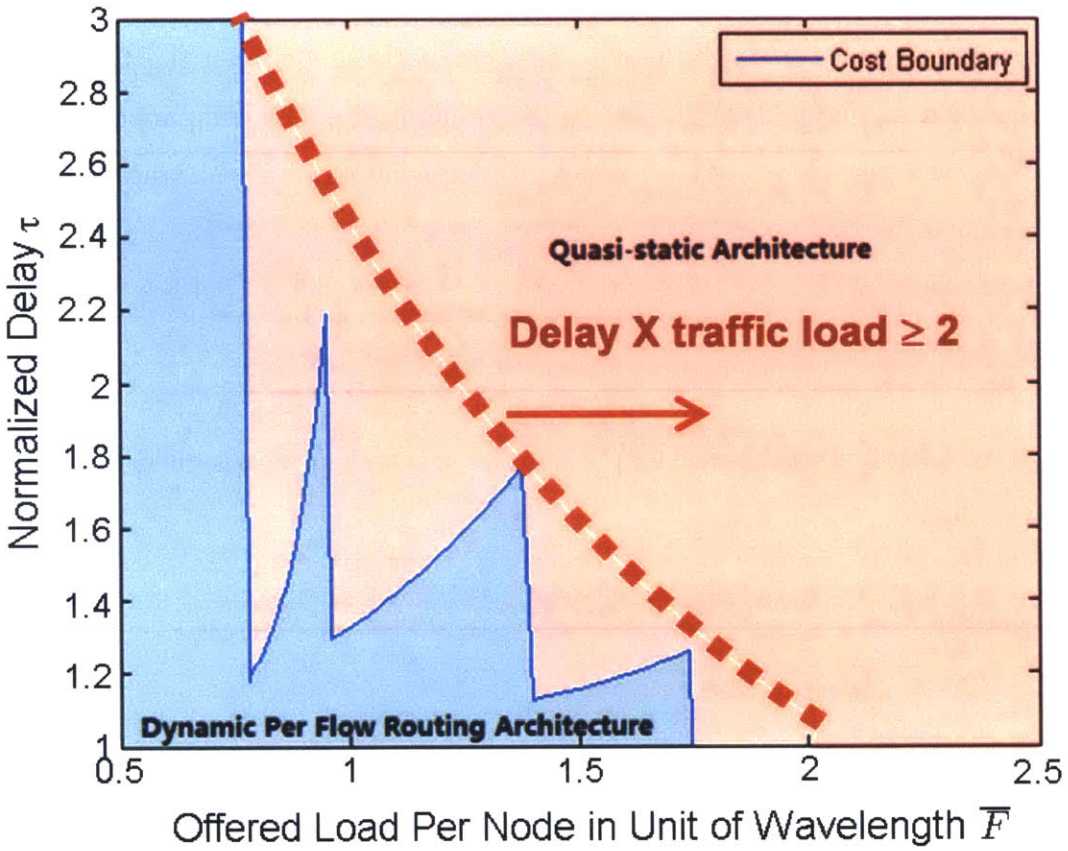


Figure 5-3: Cost boundary between QSA and DPFRA with sq-first strategy versus different delay requirements and different offered load per node in unit of wavelengths $\overline{F}$ with the same blocking probability requirement. The blocking probability requirement is 0.01. Parameter Assumptions: $\alpha_{s_1} = \$10,000/port\ pair$, $\alpha_{s_2} = \$2.35 \times 10^{-5}/(flow \cdot port\ pair)$, $t = 5\ years$.

5.4 Hybrid Architecture Discussion

The idea of the hybrid architecture is to dynamically choose the routing strategy to optimize the network performance. However, with the results above, we find such intermediate architecture is not practical, and is not recommended. The two main reasons are as follows.

First, such change makes the architecture more complicated. It is less scalable and manageable when the size of the network increases.

Second, such change of routing strategies makes no significant improvement to the architecture in terms of the cost. For a given network with certain numbers of

wavelengths, QSA tends to use all the wavelengths at a lower loading, while DPFRA with sq-first strategy tends to fully utilize each wavelength to make the total number of wavelengths used as small as possible. So it is only possible to change QSA to DPFRA with sq-first strategy, since we cannot build new fiber connections with switches immediately for the hybrid architecture. With the rerouting enabled on the old QSA architecture, the load of wavelengths will decrease. However, the cost increases rather than decreases, since the operating cost of switches, the control traffic cost, and the computational complexity cost grow, while other costs keep the same. If the loads of several wavelengths are low enough, we may be able to turn off these wavelengths to reduce the cost. However, the reduced cost is insignificantly small if we only turn off a small number of wavelengths, like one or two. If a large amount of wavelengths can be turned off, like more than a half of the wavelengths, we should no longer use OFS, since the traffic amount is light, and the existing electrical switching is more cost effective. Therefore, such change in architecture cannot help to reduce the cost.

Based on the two reasons above, the idea of hybrid architecture of QSA and DPFRA is no longer considered.

5.5 Summary of Optimized Architecture

We propose two extreme architectures for the MAN, QSA and DPFRA, which has been shown that these two options are enough to handle different network scenarios. The important metric to distinguish the two options are delay, throughput and ultimately cost for the same quality of service. Figures in Section 4.4 shows the expected performance that routing to the shortest queue is always more efficient in terms of blocking probability and delay, with the no-routing architecture being only $\sim$ 80% as efficient as the routed case for a blocking probability of 0.01. However, if cost of components such as switches and network management and control are factored in the decision, as illustrated in Figure 5-3, any time when the product of the required normalized delay and the traffic load ≥ 2, it is cheaper to pick the no-routing architecture albeit using more wavelengths. Notice here the extra switch port cost dominates

the extra costs. The extra cost of network management and control is insignificant. Thus, we conclude the best architecture for the MAN is a quasi-static topology without per flow switching. This would make the MAC protocol much simpler and the hardware switching speeds (part of MAC) are quasi-static (can be done $> 100mS$) as opposed to $\sim 10mS$.

Chapter 6

Conclusion

In this thesis, we have focused on the MAN architecture design in both physical topology and routing architecture. We have provided the MAN architecture building strategies based on the MAN traffic demand, performance requirements, and the parametric cost model.

6.1 Summary of Contributions

Based on the characteristics of the MAN traffic, we have shown the all-to-one stochastic traffic best addresses the bursty traffic transmission between the nodes and the hub. Based on this traffic model and the designed-in regularity of the MAN topology, we adopt Moore Graphs and Generalized Moore Graphs as the MAN physical topology.

In the routing architecture of the MAN, we first showed the $M/M/x/m$ queue describes the queueing system of each node in the MAN. We considered two extreme routing architectures, Quasi-static Architecture (QSA) where the rerouting is not performed, and Dynamic Per Flow Routing Architecture (DPFRA) where the rerouting is possible for each flow. For DPFRA, we designed two rerouting strategies, shortest-queue node first routing strategy (sq-first strategy) and traffic-receiving node first routing strategy (tr-first strategy). The performance comparisons showed the advantages of DPFRA with tr-first strategy in terms of the normalized delay compared to

that of QSA or the blocking probability compared to that of DPFRA with sq-first strategy can be ignored. So DPFRA with tr-first strategy was eliminated from the further analysis. Also, we have shown the hybrid architecture, the intermediate state of QSA and DPFRA, is impractical. Therefore, we only consider the choice of either QSA or DPFRA with sq-first strategy based on the required operating conditions.

From the results in the performance comparisons of QSA and DPFRA with sq-first strategy in terms of the normalized delay, blocking probability, load, and cost, we find for the MAN, DPFRA with sq-first strategy always has the lower queueing delay and lower blocking probability than that of a QSA at the expense of more complexity in scheduling, switching and network management and control. The network configuration and management of QSA are much simpler so that it reduces the operating expenditure and MAC complexity. Our analysis based on Moore Graphs and Generalized Moore Graphs indicates that QSA becomes cheaper when the product of the average offered load per node and the normalized delay are equal to or larger than ~ 2 units of wavelengths, with both architectures meeting the same delay or blocking probability requirements. Also, the cost boundary shows that DPFRA with sq-first strategy is preferred only when the delay requirement is stringent and the offered load is low, while QSA is much more suitable for the all-optical MAN to accommodate modest to heavy network traffic. Since each node of the MAN is generally tied to an access network which can be a tree or a bus, the quasi-static physical architecture for the MAN is favorable if the amount of traffic per access network to the same destination is ~ 2 wavelengths or above. This result is a great relief since per flow switching is complicated and the complexity of management and control (and MAC) may prevent OFS from being deployed in the near future. In essence the quasi-static architecture uses wavelength channels a little more inefficiently in exchange for using less network resources, simpler network management and control without higher complexity, and cost of fast per flow switching.

6.2 Future Work

The possible directions for future work are listed as below.

1. In our modeling of the physical network topology of the MAN, we assume the MAN topology is a regular graph and thus use the optimal candidate Moore Graphs and Generalized Moore Graphs. In reality, the physical topology for some MANs may not be regular graphs. So one can study the cases where the MAN physical topologies are irregular graphs to provide a more generalized understanding to the MAN architecture performances and the cost.

2. In our modeling of the routing architecture, we model different routing architectures into different queueing models based on the basic queueing model, $M/M/x/m$ queue. One can generalize the basic queueing model to $G/G/x/m$ queue to study the performance of the MAN with the different arrival rates and the service rates.

Appendix A

Discussions and Derivations for Chapter 4

A.1 Coefficient α for Poisson Process

For a Poisson process with mean $\overline{m}$, the probability mass function for its arrival counting process $\{N(t); t > 0\}$ is [6]

$$Pr\{N(t) = n\} = \frac{\overline{m}^n e^{\overline{m}}}{n!} \tag{A.1}$$

So by Equation 4.8, we have

$$\sum_{n=0}^{m^+-1} \frac{\overline{m}^n e^{\overline{m}}}{n!} \geq 1 - P_{b_{thr.}} \tag{A.2}$$

From Equation A.2, we get m^+ for any given $\overline{m}$. Since the standard deviation for this Poisson process is $\sqrt{\overline{m}}$, further by 4.9, we have

$$\alpha = \left\lceil \frac{m^+ - \overline{m}}{\sqrt{\overline{m}}} \right\rceil \tag{A.3}$$

From Figure A-1, we can find α is bounded by 4 when the blocking probability requirement is 0.01.

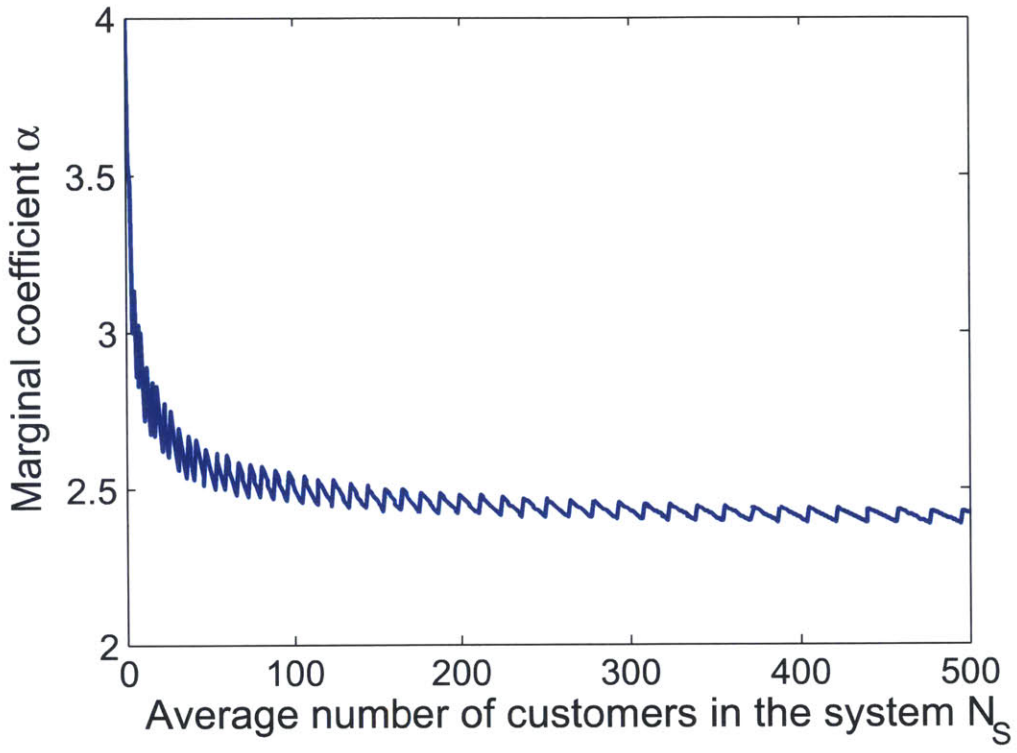


Figure A-1: Coefficient α of Poisson Distribution with different means (Average number of customers in the system N_S) when the blocking probability requirement is less than 0.01.

A.2 Derivation of $M/M/x/m$ queue model

In Section 4.1, we model the transmission process of each node in the MAN into an $M/M/x/m$ queue with arrival rate λ and service rate μ. x is the number of servers, which is the number of wavelengths a node has the access to. m is the number of the flows that the system can accommodate, including flows in the schedule holder and the flows in transmission. The state transition diagram is shown in Figure A-2. Similar to the derivation of $M/M/m/m$ queue model shown in [1], we derive the $M/M/x/m$ queue model as follows.

Assume the steady-state probability of state i is p_i. By the global balance equations for the steady-sate probability, we obtain the equations as follows

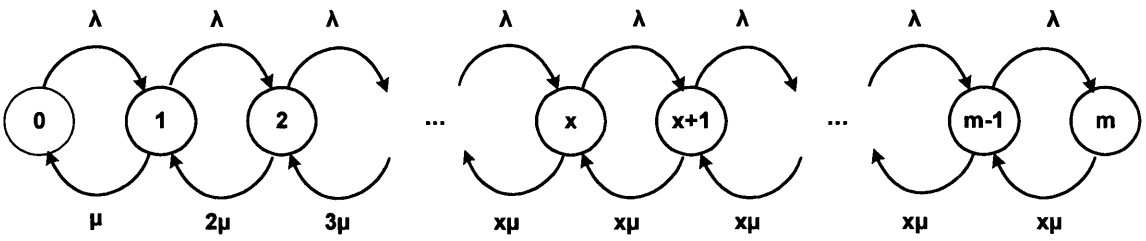


Figure A-2: State transition diagram of $M/M/x/m$ queue

$$\lambda p_{n-1} = n\mu p_n, \quad 0 < n \leq x \tag{A.4}$$

$$\lambda p_{n-1} = x\mu p_n, \quad x < n \leq m \tag{A.5}$$

From these equations we obtain

$$p_n = \begin{cases} \frac{1}{n!}(\frac{\lambda}{\mu})^n p_0 & 0 < n \leq x \\ \frac{x^x}{x!}(\frac{\lambda}{x\mu})^n p_0 & x < n \leq m \end{cases} \tag{A.6}$$

We can get p_0 using the above equations and the condition $\sum_{n=0}^{\infty} p_n = 1$. We obtain

$$\begin{aligned} p_0 &= [\sum_{n=0}^{x} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^{x+1}}{x \cdot x!} \cdot \frac{1 - (\frac{\lambda}{x\mu})^{m-x}}{1 - \frac{\lambda}{x\mu}}]^{-1} \\ &= [\sum_{n=0}^{x} \frac{(x\rho)^n}{n!} + \frac{(x\rho)^{x+1}}{x \cdot x!} \cdot \frac{1 - \rho^{m-x}}{1 - \rho}]^{-1} \end{aligned} \tag{A.7}$$

where ρ is given by

$$\rho = \frac{\lambda}{x\mu} < 1$$

So the probability for a flow to be in the queue is

$$
\begin{aligned}
P_Q &= \sum_{n=x}^{m-1} p_n \\
&= \frac{x^x}{x!} p_0 \sum_{n=x}^{m-1} (\frac{\lambda}{x\mu})^n \\
&= \frac{\rho^x - \rho^m}{1-\rho} \cdot \frac{x^x}{x!} p_0
\end{aligned} \tag{A.8}
$$

The probability for a flow to be blocked from the system is the probability that the system is in state m, where the schedule holder with size $m - x$ are full. So the blocking probability is

$$
P_B = p_m = \frac{1}{x^{m-x} x!} (x\rho)^m p_0 \tag{A.9}
$$

So the average number of flows waiting in the schedule holder is

$$
\begin{aligned}
N_Q &= \sum_{n=x}^{m} (n-x) p_n \\
&= \sum_{n=x}^{m} (n-x) \frac{x^x}{x!} \rho^n p_0
\end{aligned} \tag{A.10}
$$

By Little's Theorem, we have the average waiting time for a flow to be transmitted as

$$
W_Q = \frac{N_Q}{\lambda(1 - P_B)} \tag{A.11}
$$

A.3 Discussion of Bound 4.26

Bound 4.26 is restated here:

The blocking probability of DPFRA with tr-first strategy is not less than the blocking probability of DPFRA with sq-first strategy with the same queueing system settings. That is

$$P_{tr} \geq P_{sq} \tag{A.12}$$

Denote the queueing system with tr-first strategy as S_{tr} and the queueing system with sq-first strategy as S_{sq}. Assume both S_{tr} and S_{sq} have the same settings except the rerouting strategy. They have the same network topology with $n-1$ nodes and 1 hub, the same number of wavelengths assigned to each node as x, the same schedule holder size as $m-x$. The traffic generated from all the nodes to the hub are independent and identically distributed.

To clearly show Inequality 4.26 holds, let's consider an extreme case: Assume both S_{tr} and S_{sq} are empty in the initial state. At epoch $t = t_1$, which can be considered as a very short time interval, m traffic flows come to an arbitrary node V_i and no traffic flows comes to the rest nodes. For S_{tr}, all the m traffic flows stay in V_i with x flows being transmitted and $m-x$ in the schedule holder. There is no flows in the rest nodes. For S_{sq}, there are two cases:

1. If $m-x$ is less or equal than $(n-2)x$, all the flows have the wavelengths to be transmitted and there is no flows in the schedule holder of any node;

2. If $m-x$ is greater than $(n-2)x$, there must be $m-(n-1)x$ in the schedule holder(s) of one or several arbitrary node(s).

At epoch $t = t_1 + \tau_{TR}$, the transmission of all the flows not in the schedule holders are finished. For S_{tr}, there are still $m-x$ flows remaining. For S_{sq}, there are no flows remained in case 1 and $m-(n-1)x$ flows remained in case 2. At this epoch, if $m(n-1)$ flows in total comes to both S_{tr} and S_{sq}, $m-x$ of newly-coming flows will be blocked in S_{tr}. In contrast, no newly-coming flows will be blocked in case 1 of S_{sq}.

Therefore, $P_{tr} > P_{sq}$ in case 1. In case 2 of S_{sq}, $m-(n-1)x$ of newly-coming flows will be blocked. Notice that we have $n \geq 2$, for the topology needs to have one hub and at least one node. So if $m-(n-1)x < m-x$, namely, $n > 2$, we have $P_{tr} > P_{sq}$; If $m-(n-1)x = m-x$, namely, $n = 2$, we can have $P_{tr} = P_{sq}$. Besides, another case that the equality can be achieved is that both the number of flows that S_{tr} and S_{sq} can accommodate are large enough. The coming flows can be quenched in time. So it leads to the equal blocking probability of S_{tr} and S_{sq}.

In general, since S_{sq} can always achieve the full utilization of all the wavelengths and the schedule holders of all the nodes while S_{tr} cannot, the situation that more flows are remained in S_{tr} than S_{sq} is impossible and thus P_{tr} is not less than P_{sq}.

Appendix B

Derivations for Chapter 5

B.1 Derivation of h_{min} in Eq. (5.10)

In section 5.2.3, we propose the idea of the average minimum number of rerouting hops h_{min}. The detailed derivation is shown as follows.

To calculate h_{min}, we should consider the nodes in different level. To facilitate analysis, by definition in [7], $n(i)$ denotes the number of nodes that are i hops away from a node via minimum hop routing; D denotes the diameter of a topology; Δ denotes the degree of a topology.

The number of levels is Δ. According to the property of Generalized Moore Graph, there are $\Delta(\Delta-1)^{k-1}$ level k nodes, where $1 \leq k \leq N-1$. Since the last level of routing spanning tree for Generalized Moore Graph is not full, so the number of level D nodes is $N-1-\frac{\Delta[(\Delta-1)^{D-1}-1]}{\Delta-2}$. Assume the selected node is the node where the traffic will reroute to. Therefore, the probability that the selected node is level k node is:

$$P_k = \begin{cases} \frac{\Delta(\Delta-1)^{k-1}}{N-1}, & 1 \leq k \leq D-1 \\ 1-\frac{\Delta}{N-1}\frac{(\Delta-1)^{D-1}-1}{\Delta-2}, & k = D \end{cases} \tag{B.1}$$

The average minimum number of hops the traffic goes through from the other node to the selected node varies from different level node to node, since we have

to consider that the nodes do not include the hub. Therefore, for level k nodes, the average minimum number of hops h_k should be $\frac{1}{N-2}[\sum_{i=1}^{D} in(i) - k]$, where $1 \leq k \leq N$.

The average minimum number of rerouting hops h_{min} is

$$\begin{aligned} h_{min} &= \sum P_k h_k \\ &= \frac{1}{N-2}[\sum_{i=1}^{D} in(i) - \frac{\Delta}{N-1}\sum_{i=1}^{D-1} i(\Delta-1)^{i-1} - (1 - \frac{\Delta}{N-1}\frac{1-(\Delta-1)^{D-1}}{2-\Delta})D] \end{aligned} \tag{B.2}$$

Since the average minimum hop distance H_{min} is defined as [7]

$$H_{min} = \frac{1}{N-1}\sum_{i=1}^{D} in(i) \tag{B.3}$$

so the average minimum rerouting hop distance is

$$h_{min} = \frac{1}{N-2}[(N-1)H_{min} - \frac{\Delta}{N-1}\sum_{i=1}^{D-1} i(\Delta-1)^{i-1} - (1 - \frac{\Delta}{N-1}\frac{1-(\Delta-1)^{D-1}}{2-\Delta})D] \tag{B.4}$$

Bibliography

[1] D.P. Bertsekas and R.G. Gallager. *Data Networks (2Nd Ed.).* Prentice-Hall, Inc., Upper Saddle River, NJ, 1992.

[2] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability.* Athena Scientific, Belmont, MA, 2008.

[3] B. Bollobas. *Modern Graph Theory.* Springer-Verlag Now York Inc., New York, NY, 1998.

[4] V.W.S. Chan. Optical flow switching networks. *Proceedings of the IEEE*, 100(5):1079–1091, May 2012.

[5] V.W.S. Chan, Lei Zheng, H. Huang, G. Weichenberg, and A. Ganguly. Optical flow switching: An end-to-end "ultraflow " architecture. In *Transparent Optical Networks (ICTON), 2013 15th International Conference on*, pages 1–4, June 2013.

[6] R.G. Gallager. *Stochastic Processes: Theory for Applications.* Cambridge University Press, New York, NY, 2014.

[7] K.C. Guan. *Cost-effective optical network architecture: a joint optimization of topology, switching, routing and wavelength assignment.* PhD thesis, Massachusetts Institute of Technology, 2007.

[8] K.C. Guan and V.W.S. Chan. Cost-efficient fiber connection topology design for metropolitan area wdm networks. *Optical Communications and Networking, IEEE/OSA Journal of*, 1(1):158–175, June 2009.

[9] L. Kleinrock. *Queueing Systems*, volume I: Theory. Wiley Interscience, 1975. (Published in Russian, 1979. Published in Japanese, 1979. Published in Hungarian, 1979. Published in Italian 1992.).

[10] K.X. Lin. Green optical network design: Power optimization of wide area and metropolitan area networks. Master's thesis, Massachusetts Institute of Technology, 2011.

[11] G. Weichenberg, V.W.S. Chan, and M. Medard. Design and analysis of optical flow-switched networks. *Optical Communications and Networking, IEEE/OSA Journal of*, 1(3):B81–B97, August 2009.

[12] K.C. Weichenberg. *Design and Analysis of Optical Flow Switched Networks.* PhD thesis, Massachusetts Institute of Technology, 2009.

[13] L. Zhang. *Network Management and Control of Flow-Switched Optical Networks: Joint Architecture Design and Analysis of Control Plane and Data Plane with Physical-Layer Impairments.* PhD thesis, Massachusetts Institute of Technology, 2014.

[14] A.X. Zheng, L. Zhang, and V.W.S. Chan. Metropolitan and access network architecture design for optical flow switching. In *Global Communications Conference (GLOBECOM), 2014 IEEE*, pages 2036–2041, Dec 2014.