IMPROVING FAIRNESS, THROUGHPUT AND BLOCKING PERFORMANCE FOR LONG HAUL AND SHORT REACH OPTICAL NETWORKS

by

SANA TARIQ
B.E. National University of Science and Technology (NUST) Pakistan, 2005
M.S. Rochester Institute of Technology NY, 2009

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2015

Major Professor: Mostafa Bassiouni

ABSTRACT

Innovations in optical communication are expected to transform the landscape of global communications, internet and datacenter networks. This dissertation investigates several important issues in optical communication such as fairness, throughput, blocking probability and differentiated quality of service (QoS). Novel algorithms and new approaches have been presented to improve the performance of optical circuit switching (OCS) and optical burst switching (OBS) for long haul, and datacenter networks. Extensive simulations tests have been conducted to evaluate the effectiveness of the proposed algorithms. These simulation tests were performed over a number of network topologies such as ring, mesh and U.S. Long-Haul, some high processing computing (HPC) topologies such as 2D and 6D mesh torus topologies and modern datacenter topologies such as FatTree and BCube.

Two new schemes are proposed for long haul networks to improve throughput and hop count fairness in OBS networks. The idea is motivated by the observation that providing a slightly more priority to longer bursts over short bursts can significantly improve the throughput of the OBS networks without adversely affecting hop-count fairness. The results of extensive performance tests have shown that proposed schemes improve the throughput of optical OBS networks and enhance the hop-count fairness.

Another contribution of this dissertation is the research work on developing routing and wavelength assignment schemes in multimode fiber networks. Two additional schemes for long haul networks are presented and evaluated over multimode fiber networks. First for alleviating the fairness problem in OBS networks using wavelength-division multiplexing as well as mode-

division multiplexing while the second scheme for achieving higher throughput without sacrificing hop count fairness.

We have also shown the significant benefits of using both mode division multiplexing and wavelength division multiplexing in real-life short-distance optical networks such as the optical circuit switching networks used in the hybrid electronic-optical switching architectures for datacenters. We evaluated four mode and wavelength assignment heuristics and compared their throughput performance. We also included preliminary results of impact of the cascaded mode conversion constraint on network throughput.

Datacenter and high performance computing networks share a number of common performance goals. Another highly efficient adaptive mode wavelength- routing algorithm is presented over OBS networks to improve throughput of these networks. The effectiveness of the proposed model has been validated by extensive simulation results.

In order to optimize bandwidth and maximize throughput of datacenters, an extension of TCP called multipath-TCP (MPTCP) has been evaluated over an OBS network using dense interconnect datacenter topologies. We have proposed a service differentiation scheme using MPTCP over OBS for datacenter traffic. The scheme is evaluated over mixed workload traffic model of datacenters and is shown to provide tangible service differentiation between flows of different priority levels.

An adaptive QoS differentiation architecture is proposed for software defined optical datacenter networks using MPTCP over OBS. This scheme prioritizes flows based on current network state.

ACKNOWLEDGMENTS

I want to express heartfelt gratitude to all those whose unconditional support helped me complete this work.

First of all I want to thank God Almighty for blessing me with the opportunity for studying at doctoral level under Fulbright scholarship program and sending people along the way who made it possible to get through this journey with success.

Dr. Mostafa Bassiouni would be the first and foremost whose help, guidance and constant encouragement helped complete my research from initial stages to its final form. I owe him an immense magnitude of gratitude and respect in eternity.

I also want to extend special thanks to my external committee member Dr. Guifang Li for his innovative ideas that gave a new direction to our research. I also want to extend thanks to Dr. Zou Cliff and Dr. Damla Turgut for serving in my doctoral committee and providing valued comments and continued support.

I attribute a big share of my success in completing this work to my husband Saad Arif whose support helped me stay focused. I owe special thanks to my loving and supporting parents who always took pride in my accomplishments and to my beautiful daughters Irsa and Aysel who keeps my spirits high. I would also like extend special thanks go to my best friend and colleague at department of EECS, Amerah Alabrah who served as a constant source energy throughout my research at UCF.

In the last, I want to extend a special thanks to my manager Kathleen Foster at T-Mobile whose support and encouragement helped me complete the final stages of my research and dissertation.

To each of the above, I extend my deepest appreciation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. CHAPTER ONE: INTRODUCTION

In this chapter, we review optical networks using wavelength division multiplexing (WDM) and mode-wavelength division multiplexing (MWDM). We also review optical routing and wavelength assignment problem in the context of multi mode fiber networks. A brief over view of optical switching architectures is discussed. This section will provide an introduction to our contributions [1-5] discussed throughput this dissertation. In the end we provide the organization of the dissertation.

1.1 Wavelength division multiplexing (WDM)

Wavelength division multiplexed (WDM) optical networks [6-9] have been envisioned as technology of choice for next-generation Internet architectures. WDM optical networks have been successfully deployed in the backbone of commercial telecommunication networks. These networks have been known to provide exceptional bandwidth, low processing cost, protocol transparency, and have been well accepted by industry and academia. However, over the last decade, the exponential growth of traffic due to the surge of new generation bandwidth hungry applications have led to the point where wavelength division multiplexing (WDM) networks are approaching their fundamental Shannon limit for transmission capacity [10].

1.2 Mode-division multiplexing (MDM)

The demand for high-bandwidth communications in optical networks is continuously increasing. Mode-division multiplexing can offer an additional degree of freedom to alleviate bandwidth bottleneck in optical networks and have recently received considerable attention [6, 11-16]. In wavelength division multiplexing (WDM) networks the available bandwidth is

divided into a number of wavelengths each of which can carry information over a separate channel within a single fiber.

1.3 Optical communication architectures

In order to efficiently utilize the available bandwidth in optical network, design efficient routing and switching protocols have to be developed. There are three optical communication switching schemes namely, optical packet switching (OPS), optical burst switching (OBS) and optical circuit switching (OCS). These optical communication paradigms differ in the size of data-unit at switching level and switching mechanism. Switching techniques primarily differ based on whether data will use 'switch cut-through' or 'store and forward'. OPS have the smallest data unit and provides excellent granularity for bursty traffic and provides high bandwidth utilization but incurs high cost and power consumption. Furthermore the lack buffering technology and strict synchronization and control issues makes OPS architecture unfeasible with currently existing technology. OPS will not be favorable in the foreseeable future until optical buffers gains maturity. OCS has the largest data unit and needs an end-to-end connection setup before communication by two way reservation scheme. Once the connection is established, all relevant resources are exclusively reserved until the data transfer is complete. Eventually the connection is broken down to free the resources. The connection setup and tear down is time consuming hence, OCS is only suitable for static and stable traffic patterns. The last switching scheme in optical communication is optical burst switching (OBS) that combines the advantages of OPS and OCS networks. OBS follows one way reservation scheme in which the data burst follows a corresponding control packet without waiting for an acknowledgment. One way reservation reduces the burst switching time as compared to OCS technique. OBS is so far

the most acceptable technology that achieves the best compromise between coarse-grained circuit switching and the fine-grained packet switching by consolidating the currently available techniques.

1.4 Routing and mode wavelength assignment (RMWA)

The end to end connection is in an optical network called a lightpath. In the absence of wavelength converters, a lightpath must occupy the same wavelength on all the fiber links through which it traverses [8]. This property is known as the wavelength continuity constraint. Given a set of connections, the problem of setting up lightpaths by routing and assigning a wavelength to each connection is called the routing and wavelength assignment (RWA) problem. In Wavelength and mode routed optical networks, mode converters also needs to be present to perform mode conversion in case the same wavelength, mode pair is not available over the link to next hop. Setting up light path in case of WDM and MDM is called routing mode wavelength assignment (RMWA) problem. The wavelength and mode continuity constraint implies that when the RMWA protocol is unable to find a path and allocate the same wavelength to all links along the path, the connection request will be blocked.

1.5 Beat down unfairness problem in OBS

In WDM OBS networks, the data is dynamically switched at sub-wavelength level by combining electronics and optics. The unit of data is a collection of packets called a *burst*. The control information is sent over a reserved optical channel, called the *control channel*, ahead of the data burst in order to reserve the wavelengths across all OXCs. The control information is electronically processed at each optical router while the payload is transmitted all-optically with full transparency through the lightpath. Hence, control packets would have to experience O/E/O

conversion for resource reservation at each intermediate optical node. The source node waits for a pre determined time called an offset time between the transmission of the control packet and the data burst. During this offset time, the data burst is buffered electronically at the source node while the control packet propagates forward to reconfigure each OXC along the lightpath. At the end of the offset time, the data burst is transmitted and is switching all-optically through the network. In an OBS network there is no connection establishment requirement the control packet may or may not have reserved a channel when the data starts in transmission. This results in blocking. It has been observed in OBS networks that longer the lightpaths of OBS connection suffer greater blocking than shorter lightpath connections. This is called "beat down" unfairness problem of OBS networks. When the number of hops *m* in the light path increases, the probability that the burst will be delivered successfully to the destination decreases. Single-hop light paths (i.e., $m=1$) have the highest probability of successful burst delivery. In this dissertation we address beat down unfairness issue and present schemes to alleviate this unfairness in the context of single and multi mode WDM networks.

1.6 Quality of service differentiation (QoS)

Quality of traffic is determined in terms of throughput, blocking and latency etc. Quality of service (QoS) is the overall network performance as seen by the end users of the network. Various protocols supporting QoS requirements have been studied for OBS networks. In an OBS network supporting a diverse range of applications, the data bursts may belong to different priority classes. Higher priority bursts needs a preferential treatment in order to reduce their drop probability and end-to-end delay. Service differentiation based on traffic of various priority levels and requirements has been an important topic in internet architectures however; very little work has been done to address

the service differentiation issue in datacenter networks. In this dissertation we address the need for service differentiation in datacenter networks and present one possible solution.

1.7 Dissertation Organization

The dissertation is organized as follows. In Chapter 1 we survey and discuss the relevant literature. In Chapter 2 we present the throughput and fairness improving schemes in long haul WDM networks. In Chapter 3, two new schemes are discussed in multi mode fiber networks, one for improving fairness and the second for concurrent improvement of throughput and fairness. Chapter 5 presents routing and mode wavelength assignment problem in OCS networks for datacenters. Chapter 6 presents performance evaluation of newly emerging transport protocol Multi path-TCP over OBS networks in datacenters and propose QoS differentiation scheme for MPTCP over OBS networks in datacenters. In Chapter 7 we present performance evaluation of MPTCP and compare it with standard TCP in cloud based datacenter optical networks using OBS. In Chapter 8, we present adaptive QoS scheme for software defined optical networks in shared datacenters using MPTCP over OBS. In Chapter 9 we present an adaptive mode wavelength routing algorithm for short-reach MDM optical networks. Chapter 10 concludes the dissertation presents possible future work.

2. CHAPTER TWO: REVIEW OF THE LITERATURE

In this chapter, a review of the relevant literature is surveyed to analyze the course of research in this area, and visualize the status of our current research within the larger paradigm of optical communication.

2.1 Introduction

Last decade has seen an exponential growth in internet, could computing and high performance computing applications. In this dissertation, we investigate fairness and throughput in OBS networks for long haul using wavelength division multiplexing (WDM) and mode and wavelength division multiplexing (MWDM). Cloud computing is becoming the heart of computational world over the past few years. We examine the problem of mode-wavelength routing and assignment in datacenter optical networks and demonstrate the viability and significant benefits of using both wavelength and mode division multiplexing. We also addressed issues of wavelength mode cascaded conversion constraint. A newly emerging transport protocol, Multi-path TCP (MPTCP) has been evaluated for OBS networks in datacenters and a service differentiation scheme for MPTCP over OBS is proposed for datacenter networks. Software defined networks have gained significant attention in research as well as industry in recent years. An adaptive QoS scheme is also presented for software defined optical networks.

Below, we provide an extensive survey on the background of each of these issues, present review of related work and elaborate on motivations that led to our proposed contributions.

2.2 Improving fairness and throughput in OBS networks

In our first contribution [1] we modified basic Just in time (JIT) [17-19] reservation protocol for OBS networks to improve throughput. The signaling protocol plays a crucial role in the burst

transmission. Various wavelength reservation schemes have been proposed and studied. Just in time (JIT) [17-19] is a reservation protocol that received attention for its simplicity. In this contribution, we will illustrate our proposed protocols for OBS networks using JIT. In JIT, the source node delays the transmission of a burst by certain amount of time after the control packet is sent. This delay is called the *offset* time. The offset time should be large enough to allow all the OXCs in the lightpath of the burst to process the control packet and configure their input/output ports. The switch configuration time, also known as the *cut-through* time, should be taken into consideration because a burst may get dropped if it arrives at an OXC before the OXC has completed the configuration of its ports. If the lightpath of a burst consists of *m* hops, the offset time t_d used in JIT can be defined as:

$$t_d \geq m t_p + t_\delta \tag{2.1}$$

Where t_p is the control packet processing time in each OXC including the time required for O/E/O conversions and t_δ is an additional delay required to complete the cut through time at the last OXC. The standard JIT scheme does not take the fairness issue into account which is generally an important area of research in OBS networks. Improving fairness in OBS networks has been addressed in [20] [21-27]. In [20]the authors presented two different approaches, balanced JIT (BJIT) and prioritized random early discard (PRED), to alleviate the unfairness problem. The hop-count unfairness of OBS networks is due to the fact that the dropping probability for bursts travelling through paths with longer hop count tends to be greater than for bursts with shorter paths. BJIT was extended from standard JIT while PRED was a proactive discarding scheme that probabilistically discards bursts at the network access station (NAS) of OBS networks. Throughput of OBS networks is another important issue that needs to be

addressed to make WDM networks a winning choice for future applications. Maximizing throughput in OBS networks is addressed in [28-33]. Our two proposed schemes in this contribution fundamentally differ from our previous approach described in [29] in that they do not use burst preemption and are easier to implement. The two schemes enhance the throughput of the OBS network through a simple but elegant approach of considering burst length and giving a selective priority to longer bursts over shorter bursts. Improving network utilization by giving preference for the transmission of larger data units over smaller data units is a well established concept in IP electronic networks. This concept was utilized in the Gigabit Ethernet Standard (IEEE 802.3z- 1998) which introduced an extension called *frame bursting* that allows the sending station to combine several data frames into one transmission frame. It has been found that frame bursting substantially increases the throughput of Ethernet networks and that the larger the size of the assembled bursts the better the network utilization. An optical burst is similar in nature to a Gigabit Ethernet burst. In OBS, packets are aggregated into *data bursts* at the source access station to form the optical data payload in the same way that data frames are aggregated into *frame bursts* at the sending Gigabit Ethernet switch (hub) to form the electronic data payload.

OBS networks can provide a fine granularity for bandwidth and improve the utilization of WDM networks [33-35]. In this section, we review previous work to improve the hop-count fairness of OBS networks [24-26] or improve the throughput of OBS networks [30-33, 36], then we give more detailed review of the BJIT and PRED protocols proposed in [20].

The Fair Prioritized Preemption algorithm (FPP) proposed in [24] computes a dynamic priority based on several characteristics of the burst. When contention occurs, the scheme picks

the burst with the highest priority and drops the other bursts. The simulation results showed that this scheme could improve fairness at the expense of little deterioration in the burst dropping performance. In [25], the authors present a scheme to improve fairness in OBS networks using random scheduling for hop-based burst-cluster transmission. When a burst-cluster is generated, the scheduler at the ingress node computes a random actual waiting time after which the control packet is transmitted. One drawback is that the transmission delay may become large. The scalability of the scheme for large hop counts was not adequately demonstrated as all tests reported in [25] were limited to a network topology whose maximum number of hops is 3.

The link scheduling state based offset selection (LSOS) scheme proposed in [26] is based on collecting link scheduling state information and using it to determine the offset times for routes with different hop lengths. The scheme has signaling overhead associated with the periodic link state collection; as in [25], the scalability of the LSOS scheme for large hop counts was not adequately demonstrated. A different type of unfairness, called the *burst length priority effect* (BLPE), is addressed in [27]. The BLPE fairness occurs when void-filling scheduling algorithms, such as the *latest available unused channel-void filling* (LAUC-VF), are used. The void-filling algorithms tend to favor shorter bursts as they fit more easily into voids. As a result, longer bursts will have a higher drop probability compared to shorter bursts. The BLPE unfairness is not practically important and the majorities of research assessments and reports on LAUC-VF has disregarded the BLPE unfairness and have viewed LAUC-VF positively as a desirable extension to improve the performance of OBS networks. In this contribution, we only focus on hop-count fairness.

The scheme proposed in [30] focuses on improving the performance of TCP over OBS networks by using a drop policy in the core optical nodes based on the hop count. When contention occurs in core nodes, the bursts with the larger total hop count are given higher priority. The goal of the scheme is to avoid retransmission of bursts with large hop counts because these bursts are the ones who introduce extra traffic load, and are the key factor for the increased network delay. Further, when a burst is dropped in a core node, a NAK is sent to the edge node for the possible retransmission of this burst. Depending on some thresholds, the edge node may retransmit this burst to avoid the delay associated with letting the upper TCP layer handle the retransmission of the dropped burst. The BATCHOPT algorithm proposed in [31] improves the throughput of OBS networks by grouping the largest possible number of control packets and processing them together rather than the greedy processing to individual control packets. The edge node gathers reservation requests during a certain time interval, and then schedules them as a batch of requests. Exact knowledge of the reservation requests in the batch allows the authors to develop optimal solution for the job scheduling problem and theoretically analyze the computational complexity of the solution.

The extensive study reported in [32] clearly shows that for both types of scheduling algorithms (non-void-filling and void-filling) and different TCP versions (Reno, New Reno and SACK), the goodput of TCP connections increases as the number of burst assemblers per egress node is increased for an OBS network employing timer-based assembly algorithm. The *source ordering* scheme for improving TCP performance in OBS networks [36] is only applicable to configurations that uses load balance routing with static route calculation and dynamic route-selection. The ingress OBS node dynamically selects the least-congested path (among the two

static link disjoint minimum hop paths) to all destination nodes using the cumulative congestion information of all the links along the two pre-calculated paths. In source ordering, the ingress node pre-calculates the path delay-differential δ between the primary minimum-hop path and the alternate second minimum-hop path. Every time a long-to-short path-switch occurs, the node electronically buffers the bursts for δ seconds before transmitting it on the shorter path. This prevents bursts transmitted on the primary path from overtaking the previously transmitted bursts on the longer alternate path before reaching the destination. The study published in [33]proposes the following three methods to improve OBS performance: 1) addition of simple fiber delay lines (FDLs) to delay the incoming data burst and avoid contention, 2) random extra offset time, and 3) window-based channel scheduling (WBS) which delays the channel/routing assignment for a specific additional duration after reading the information of a control packet. This delay provides better accuracy in estimating the impact of the channel requests (control packets) arriving in the future and leads to better channel/routing assignment decisions.

2.3 Feasibility of Mode division multiplexing

In order to meet the data centers' and High Performance Computing centers' growth forecasts in terms of traffic, computational ability, latency below microseconds and communication at the rate tens of gigabits/sec, optical communication using wavelength division multiplxing would need to be complemneted with alternative ways for increasing bandwidth capacity. Multi mode fiber networks are expected to be one of the next big breakthroughs in the field of optical networks. While the demand for high-bandwidth communications in optical networks is continuously increasing, the technology of wavelength division multiplexing (WDM) is approaching the fundamental Shannon limit for transmission

capacity [10]. Mode-division multiplexing (MDM) has recently received considerable attention as an alternative way to increase the optical fiber capacity and a number of successful experiments have demonstrated the feasibility of mode-division multiplexed WDM transmission for distances in the range 40 km to 177 km. The authors in [11] successfully transported data at 100 Gb/s over 40 km using MDM with five modes. We have successfully demonstrated mode-division multiplexed WDM transmission over 50-km [15] and adaptive frequency-domain equalization for MDM transmission [13]. The authors in [37] transmitted 32 WDM channels over 12 spatial and polarization modes of 177 km few-mode fiber at a record spectral efficiency of 32 bit/s/Hz. The authors in [12] demonstrated wavelength- and mode-division multiplexed transmission over a fiber re-circulating loop comprising 50-km of low-DMGD few-mode fiber and an optimized few-mode EDFA. The schemes in [12, 37] use Mode Mux/DeMux that work for WDM channels. There is a strong optimism in the optical community that MDM will become a feasible transport technology that can be used in conjunction with WDM.

2.4 Optics in datacenters

Many internet applications today are powered by data centers. The traffic generated by these bandwidth intensive applications grew exponentially over the last few years resulting in a massive increase in the computational, storage and scalability requirements of data centers. Meeting performance goals for data centers networks (DCN) became very challenging under these conditions. The research on improving communication efficiency in data center networks received considerable attention in recent years [7, 38-47]. With the rise of a newer generation of applications, such as web search, retail advertising and social media, optical networking holds a huge promise in cloud computing and datacenters [7, 40, 44, 45, 48]. The research in the area of

optical networking for data centers is rich and still growing. There are three switching paradigms that are considered for WDM optical communications. In optical circuit switching (OCS), that two way reservation protocol needs to establish a reserved path for communication and incurs high latency. The switching environment in datacenters is very fast hence OCS, circuit set-up latency makes it misfit for ON-OFF [43] bursty traffic. OCS has been deployed in hybrid OCS-EPS (electronic packet switching) networks in datacenters [49, 50], however, due to its slower switching times they are currently serving only the larger flows in datacenter. It is assumed that OCS is only a short term solution to achieve optics in datacenter [44]. It will not be able to provide all optical switching in future datacenter networks. The second switching technique, optical packet switching OPS is still not feasible due to lack of optical buffering technology. The high data rates of optical networks also imply that switching times in OPS networks must be in the order of a few nanoseconds [51] making it even harder for the technology to catch up. This leaves us with third switching technology optical burst switching OBS that combines the advantages of OCS and OPS, providing burst size granularity, bandwidth flexibility, a one-way reservation scheme, variable burst length, and no optical buffer. In one way reservation, OBS incurs much lower latency than OCS. OBS is the only technology that can achieve all-optical networking in future datacenters once the technology gets matured [44]. OBS has been proposed for a number of datacenter architectures e.g., Sowailem et al. in [41] proposed the use of optical burst switching in data center networks and provided hardware level design modifications to establish its feasibility in the DCN environment. Similarly, Li et al. in [45] suggested the use of OBS in inter-pod data communication to avoid bottleneck in this layer of DCN.

2.5 Multi-path TCP (MPTCP)

Another approach that can significantly improve bandwidth efficiency in datacenters is utilizing Multipath-TCP (MPTCP) [38, 42, 52-54]. Modern datacenter topologies already contain redundant multiple paths such as FatTree, BCube [38] and VL2 [47] etc and single path TCP is not capable of using multiple paths simultaneously hence the available capacity is under used.

TCP has been considered as the most dominant transport layer protocol in internet. It was natural that researchers studied the performance of TCP over OBS [55-57] for future optical networks. They discussed a number of issues e.g., performance evaluation of TCP over OBS [58] TCP unfairness over OBS [59], OBS random burst loses and effect on congestion window [56] and dissertations to improve TCP performance over OBS networks [51].

Multipath-TCP, an extension of TCP has recently gained attention and studied over electronic packet switching in datacenters[38, 42]. Due to densely interconnected topologies of modern datacenters [38, 42, 47], MPTCP has shown to improve the performance and efficiency through multiple path utilization. However none of research studies have evaluated the performance of multi-path TCP over Optical burst switching networks. In this dissertation, this is the first research that evaluates the performance of MPTCP over OBS in datacenter networks. We combine the advantages of Multipath-TCP with optical networking and present an evaluation of MPTCP over OBS for data center network. We compare the performance of standard TCP with MPTCP under different network loads and topologies.

The research on improving communication efficiency in data center networks received considerable attention in recent years [7, 38-47]. One popular approach to achieve bandwidth optimization is introduction of optical networking in datacenter [7, 40, 44, 45, 48]. Optical burst

switching (OBS) has been considered as the best compromise between optical circuit switching (OCS) and optical packet switching (OPS) due to its granularity and bandwidth flexibility, and would be suitable for datacenters eventually as optical switching technology gets mature [44].

Another approach that can significantly improve bandwidth efficiency in datacenters is utilizing Multipath-TCP (MPTCP) [38, 42, 52-54]. Modern datacenter topologies already contain redundant multiple paths such as FatTree, BCube[38] and VL2[47] etc and single path TCP is not capable of using multiple paths simultaneously hence the available capacity is under used.
Moving towards optical networks and implementing multiple access transport protocols such as Multi-path TCP can successfully address bandwidth constraints. TCP has been considered as the most dominant transport layer protocol in internet. The researchers studied the performance of TCP over OBS [48, 53, 54] for future optical networks and discussed a number of issues e.g., performance evaluation of TCP over OBS [55], TCP unfairness over OBS [56], OBS random burst loses and effect on congestion window [53] and dissertations to improve TCP performance over OBS networks [48]. However none of research studies have evaluated the performance of multi-path TCP over Optical burst switching networks.
Data centers are becoming the heart of the computational world over the past few years. The emergence of cloud computing and the growth of data-intensive applications have driven the need for finding alternative ways to improve communication efficiency in data center networks.

With the rise of a newer generation of applications, such as web search, retail advertising and social media, optical networking holds a huge promise in cloud computing and datacenters. The research in the area of optical networking for data centers is rich and still growing. There are three switching paradigms that are considered for WDM optical communications. In optical

circuit switching (OCS), that two way reservation protocol needs to establish a reserved path for communication and incurs high latency. The switching environment in datacenters is very fast hence OCS, circuit set-up latency makes it misfit for ON-OFF [43] bursty traffic. OCS has been deployed in hybrid OCS-EPS (electronic packet switching) networks in datacenters [49, 50], however, due to its slower switching times they are currently serving only the larger flows in datacenter. It is assumed that OCS is only a short term solution to achieve optics in datacenter [44]. It will not be able to provide all optical switching in future datacenter networks. The second switching technique, optical packet switching OPS is still not feasible due to lack of optical buffering technology. The high data rates of optical networks also imply that switching times in OPS networks must be in the order of a few nanoseconds [51] making it even harder for the technology to catch up. This leaves us with third switching technology optical burst switching OBS that combines the advantages of OCS and OPS, providing burst size granularity, bandwidth flexibility, a one-way reservation scheme, variable burst length, and no optical buffer. In one way reservation, OBS incurs much lower latency than OCS. OBS is very feasible technology that can achieve all-optical networking in future datacenters once the technology gets matured[44]. OBS has been proposed for a number of datacenter architectures e.g., Sowailem et al. in [41] proposed the use of optical burst switching in data center networks and provided hardware level design modifications to establish its feasibility in the DCN environment. Similarly, Li et al. in [45] suggested the use of OBS in inter-pod data communication to avoid bottleneck in this layer of DCN.

TCP has been considered as the most dominant transport layer protocol in internet. It was natural that researchers studied the performance of TCP over OBS [55-57] for future optical

networks. They discussed a number of issues e.g., performance evaluation of TCP over OBS[58] TCP unfairness over OBS [59], OBS random burst loses and effect on congestion window[56] and proposals to improve TCP performance over OBS networks [51].

2.6 Service differentiation in datacenter network

There is a growing interest in introducing QoS differentiation in datacenters, motivated by the need to improve the quality of service for time sensitive datacenter applications and to provide clients with a range of service-quality levels at different prices. Over the past decade, considerable attention has been given to different areas of cloud computing e.g., efficient sharing of computational resources, virtualization, scalability and security. However, less attention has been paid to network management and QoS (Quality-of-Service) provisioning in datacenters. The inability of today's cloud technologies to provide dependable and predictable services is a major showstopper for the widespread adoption of the cloud paradigm [60].

The type of applications hosted by datacenters are diverse in nature ranging from back-end services such as search indexing, data replication, MapReduce jobs to front end services triggered by clients such as web search, online gaming and live video streaming [39]. The background traffic contains longer flows and is throughput sensitive while the interactive front end traffic is composed of shorter messages and is delay sensitive. The traffic belonging to the same class can also have differences in relative priority levels and performance objectives [61].

Future data center consumers will require quality of service QoS as a fundamental feature. There have been some recent research studies on traffic modeling, network resource management and QoS provisioning in data centers [39, 43, 62]. Ranjan, et. al., studied the problem of QoS guarantees in data-center environments in [62]. However, this work is not

suitable for highly loaded shared data-centers with computationally intensive applications due to the two sided nature of communication. Song Ying et al. in [63] proposed a resource scheduling scheme which automatically provides on-demand capacities to the hosted services, preferentially ensuring performance of some critical services while degrading others when resource competition arises. However research studies on QoS provisioning in data centers did not employ optical networks nor did they use multi-access transport protocols such as MPTCP.

There is rich research on QoS schemes in optical burst switching for wide area networks [64-66]. In this dissertation we present and evaluate a QoS provisioning algorithm called QAMO, 'QoS aware MPTCP over OBS' that provides QoS provisioning algorithm for service differentiation using MPTCP over OBS in datacenters.

2.7 Software Defined networks

Cloud services are expanding and organizations are shrinking their datacenters to virtualization technologies in order to take advantage of the predictability, continuity, and quality of service. Future data center consumers will require quality of service QoS as a fundamental feature. Software defined networks have received significant attention in industry and research community in recent years [67-72]. There have been some recent research studies on traffic modeling, network resource management and QoS provisioning in data centers [39, 43, 62]. Ranjan, et. al., studied the problem of QoS guarantees in data-center environments in [62]. However, this work is not suitable for highly loaded shared data-centers with computationally intensive applications due to the two sided nature of communication. Song Ying et al. in [63] proposed a resource scheduling scheme which automatically provides on-demand capacities to the hosted services, preferentially ensuring performance of some critical services while

degrading others when resource competition arises. Hong et al. in [73] proposed a flow scheduling protocol called Preemptive Distributed Quick (PDQ), designed to complete flows quickly by emulating a shortest job first algorithm and giving priority to the short flows. Similarly authors in [74] propose taxonomy to categorize existing works based on three main techniques, reducing queue length, prioritizing mice flows, and exploiting multi-path. Zats et al. in [75] proposed DeTail, which designed a cross-layer network stack aiming to improve the tail of completion time for delay-sensitive flows. Wilson et al. [76] presented a deadline-aware control protocol, named D3, which controlled the transmission rate of network flows according to their deadline requirements. D3 gave priority to mice flows and improved the transmission capacity of data center networks.

Previously, research studies on QoS provisioning in data centers did not employ optical networks nor did they use multi-access transport protocols such as MPTCP. Extending our previous work on MPTCP over OBS in datacenters [77], and QoS provision scheme QAMO, allows us to create a network model and QoS provisioning scheme for software defined datacenters called QAMO-SDN.

There is rich research on QoS schemes in optical burst switching for wide area networks [64-66]. OBS has been considered as the best compromise between optical circuit switching (OCS)and optical packet switching (OPS) due to its granularity and bandwidth flexibility, and would be suitable for data centers eventually as optical switching technology gets mature [44]. TCP is the most dominant transport layer protocol in internet and TCP over OBS has been extensively studied [55-57]. Multipath-TCP (MPTCP) has been shown to provide significant

improvement in throughput and reliability in electronic packet switched networks in data centers [38, 42]. However, MPTCP has not been studied in the context of OBS networks before.

3. CHAPTER THREE: IMPROVING FAIRNESS AND THROUGHPUT IN WDM OBS NETWORKS

3.1 Introduction

In this chapter, we propose and evaluate two schemes for improving bandwidth utilization in optical burst-switched (OBS) networks employing timer-based burst assembly routines. The first scheme adjusts the size of the search space for a free wavelength in optical switches using a balancing formula that promotes throughput and hop count fairness. The formula achieves controllable increase in the size of the search space either when the size of the burst increases or when the hop count of the traveling burst increases. The second scheme proactively discards bursts at the source network access station using a dropping probability matrix that satisfies certain horizontal and vertical constraints. The matrix assigns smaller dropping probabilities to bursts with larger sizes and longer lightpaths. The results of extensive performance tests to evaluate the two schemes and compare them with previous schemes for improving fairness and throughput are presented and discussed. The results show that the two schemes improve the throughput of optical OBS networks and enhance the hop-count fairness.

Rest of the chapter is organized as follows. Section 3.2 highlights the motivation for the idea. Section 3.3 describes the proposed idea and Section 3.4 provides a detailed analysis of simulation and performance results and section 3.5 provides a high level summary of the chapter.

3.2 Motivation for the proposed idea

The motivation for this work came from two previously proposed schemes in [20] called BJIT and PRED to alleviate hop-count unfairness problem. When the number of hops *m* in the light path increases, the probability that the burst will be delivered successfully to the destination

decreases. Single-hop light paths (i.e., *m=1*) have the highest probability of successful burst delivery. The first scheme BJIT proposed in [20] deals with the fairness problem by adjusting the size of the search space for a free wavelength based on the number of hops traveled by the burst. As the burst moves from one hop to the next, BJIT increases the number of wavelengths that can be used to switch this burst at each OXC. As the burst travels from source to destination, a larger fraction of the available spectrum of W wavelengths is searched for a free channel in each next hop.

The performance results presented in [20] showed that BJIT can alleviate the unfairness problem with a very small impact on the overall throughput of the system. In BJIT, a burst is allowed to select a free wavelength from n_i wavelengths at its i^{th} hop, where $n_i \le n_j \forall i < j$. The value of n_i is determined according to the following equation

$$n_i = \lceil (1-g) \times W + g \times i \times W / D \rceil \qquad 0 \le g \le 1 \tag{3.1}$$

The ceiling function is used to yield an integer number of wavelengths which should be at least equal to 1. The symbol W is the number of wavelengths in a fiber link and the diameter D is the maximum hop count of any lightpath in the network topology. The parameter g controls the degree of effectiveness of resolving fairness. Generally speaking, the larger we assign a value to g, the better fairness we can obtain but at the expense of slightly dropping some bursts which have shorter hops to destination. Two simple policies can be employed to choose the subset of n_i wavelengths from the entire W wavelength set $[1, W]$ at the burst's i^{th} hop — choose the first n_i applicable wavelengths sequentially, i.e., search the subset $[1, n_i]$; or select a total of n_i wavelengths randomly as suggested in [22].

PRED is the second scheme proposed in [20] which alleviates hop-count unfairness in OBS networks by proactively discarding bursts with lesser hops to destination at source node NAS. The PRED scheme adapts the concept of random early discard (RED) to the OBS environment and prioritizes the levels of discarding based on the length of the lightpath. The discarding probability decreases as the length of the lightpath increases. Specifically, all proactive discarding in PRED is done in the network access station (NAS) of the source node that generated the burst. Unlike the scheme proposed in [29], PRED does not preempt any burst after the lightpath has been established and the burst has been accepted. This has the advantage that the discarded bursts will not waste any bandwidth resources in the core of the optical network.

Let α_i be the probability used by PRED at the source NAS to discard a newly incoming burst whose light path has a length of i hops. To alleviate the hop-count unfairness problem, the values of the discarding probabilities must satisfy the following constraint.

$$\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_i \geq \alpha_{i+1} \geq \ldots \geq \alpha_{D-1} \geq \alpha_D \tag{3.2}$$

The proactive discarding in PRED should not take place if the load on the OXC is light. This is because unfairness is not noticeable at light loads and most bursts are expected to reach their destination successfully. The mechanism adopted by the authors in [20] is to disable/enable proactive discarding in PRED based on the current load on the network. In OBS, each NAS uses a buffer to hold assembled bursts until they are sent to the local OXC. Bursts arriving when this buffer is empty do not get discarded by PRED. When the buffer is not empty, new bursts are subjected to the probabilistic discarding of PRED.

3.3 Proposed idea

The schemes proposed in this section are motivated by the observation that providing a slightly more priority to longer bursts over short bursts can significantly improve the throughput of the OBS networks without adversely affecting hop-count fairness[1]. We will now describe the two new schemes BJIT-S and PRED-S that take the burst size S into consideration.

3.3.1 BJIT-S

In the BJIT scheme [20], the control message searches for a free wavelength from a subset of maximum available wavelengths at each OXC and the size of the search subset gradually increases as the burst moves towards the destination node. The largest fraction of wavelengths is searched to reserve free wavelengths closer to the destination. If at a particular node, the control packet is unable to find a free wavelength from the designated subset, the packet is considered blocked and gets dropped, even though there may be wavelengths available at that node outside the subset. This approach was only sufficient for improving fairness in OBS networks. Typically in an OBS network, the arriving bursts are of different sizes and a bandwidth reservation technique can simply look into the burst size in order to enhance the overall throughput of the system. The new scheme presented here, BJIT-S, will incorporate the burst size to positively enhance throughput. We will introduce a new term for BJIT-S, the *size factor* η where η is node-dependent and is equal to the ratio of the size of the current burst S to the maximum allowed burst length S_{max}. The variable S_{max} can be conveyed in the control packet of the burst so that the computation is based on the burst size distribution of the node that originated this burst.

$$\eta = S / S_{\max} \tag{3.3}$$

In BJIT-S, we tune the search subset by adding the size factor for each burst. Specifically, the number of wavelengths n_i that are searched in the i^{th} hop of a burst is given by

$$n_i = \lceil (1-g) \times W \times \eta + g \times i \times W \times \eta / D \rceil \qquad 0 \le g \le 1 \tag{3.4}$$

Where,

n_i = number of wavelengths searched at the i^{th} hop

W = maximum number of wavelengths

D = diameter of the network
η = size factor
i = current hop

The ceiling function is used to yield an integer number of wavelengths which should be at least equal to 1. The *size factor* η in Eq. 3.4 adjusts the wavelength search subset based on the size of the current burst, and allows a larger number of wavelengths to be searched for larger bursts. Consequently, for two bursts of different sizes but with the same hop count, BJIT-S will allow a larger wavelength search space to the burst that is of larger size. Notice that the BJIT-S scheme is not biased against an ingress node which produces bursts of sizes smaller than bursts generated by other nodes. This is because Eq. 3.4 gives a search size n_i that is based on the burst size distribution of the source node that originated the burst. Unfairness against a particular node because of its burst size distribution does not exist in the BITJ-S scheme.

The first term of Eq. 3.4 assigns a fixed number of wavelengths for bursts of particular sizes, whereas the second term is adjustable and determines the wavelength subset size based on the hop count and the burst size. The value of the constant g in this equation varies from 0 to 1

inclusive and decides the fraction of wavelengths available for the fixed and adjustable part. For $g = 0$, each burst is assigned a subset irrespective of its hop count, and for $g = 1$, only the adjustable part comes into play. As we will discuss in Section 4, we have tested our scheme using different values of g, keeping all other factors constant.

3.3.2 PRED-S

The PRED scheme [20] employs probabilistic discarding of bursts at the source NAS when the network is heavily congested. PRED only takes into account the hop distance and does not distinguish between bursts of different sizes. In PRED-S, we bring the burst size into consideration when deciding random early discard. Under heavy network loads when the NAS buffer is not empty and a new burst arrives, the burst is dropped with a certain probability based on its lightpath length and its size. Like PRED, PRED-S only becomes active once the network is congested.

In PRED-S, a dropping probability matrix α is generated at each source NAS. The matrix α is an $m \times n$ matrix in which α_{ij} signifies the probability of proactively dropping a newly arriving burst at the source node if the light-path has a length of i hops and the integer value j is given by

$$j = \left\lfloor S / S_{\min} \right\rfloor \tag{3.5}$$

where S is the size of the current burst and $S_{\min}$ is the minimum allowed burst size. PRED-S must satisfy two constraints when generating α **:**

$$\alpha_{1,j} \geq \alpha_{2,j} \geq \ldots \geq \alpha_{i,j} \geq \alpha_{i+1,j} \geq \ldots \geq \alpha_{D-1,j} \geq \alpha_{D,j} \tag{3.6}$$

$$\alpha_{i,1} \geq \alpha_{i,2} \geq \ldots \geq \alpha_{i,j} \geq \alpha_{i,j+1} \geq \ldots \geq \alpha_{i,M-1} \geq \alpha_{i,M} \tag{3.7}$$

where D is the network diameter and M is given by

$$M = \left\lfloor S_{\max} / S_{\min} \right\rfloor \tag{3.8}$$

Here, $S_{\max}$ is maximum allowed burst length. The PRED-S' dropping probability matrix is shown in Eq. 3.9. The element in the first row and first column $\alpha_{1,1}$ will have the largest probability and the values of the discarding probability decreases as we go down or right in the matrix.

$$\alpha = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,M} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,M} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{D,1} & \alpha_{D,2} & \cdots & \alpha_{D,M} \end{bmatrix} \tag{3.9}$$

Equations 3.8 and 3.9 are computed by each ingress (access) node based on its own parameters. The parameter S_{max} in a node represents the maximum burst length for this node, which could be smaller than S_{max} in another node. Each node constructs the dropping probability matrix that is most suitable for its own burst size distribution. Bursts having the maximum size in a node will have the best preferential treatment even if this maximum size is smaller than the maximum burst size in other nodes. Unfairness against a particular node because of its burst size distribution does not exist in the PRED-S scheme as will be shown in Section 3.3.3.

The constrained given in Eq. 3.6 shows that as the number of hops increases, the dropping probability decreases thereby alleviating the hop-count unfairness problem. Eq. 3.7 shows that bursts of larger sizes will have lower dropping probabilities compared to bursts of smaller sizes, thereby maximizing OBS network throughput. Algorithm in Figure 3-1 shows the procedure of generating the matrix α that will be used by each NAS node to implement PRED-S. The terms δ_h and δ_s are step functions for the hop count and burst size, respectively. The

algorithm starts by checking whether the transmission buffer of the source node is not empty, and if so it enables proactive discarding. We start by filling in the dropping probabilities in α from the lower right corner of the matrix given in Eq. 3.9. This element $\alpha_{D,M}$ signifies the burst with the largest hop count and maximum burst size, therefore is given 0 dropping probability. The algorithm then works its way up by filling the last column. Each element in the column is assigned a probability greater than the probability of the element below it by adding the hop step function δ_h. Then each row of the matrix is filled from column $(M-1)$ to the first column, with each element assigned a probability greater than the probability of the element towards its right by adding the burst size step function δ_s.

Table 3.1 shows a 10×4 α matrix used by the PRED-S scheme in our simulation tests.

Algorithm 1 Generate-Dropping-Probability-Matrix

```
Input: D, M, δh, δs
Initialization: Allocate matrix α(D×M)
If (NAS transmission buffer is not empty) then
    { α_{D,M} := 0;
      For i = (D-1) to 1 step -1 do
            { α_{i,M} := α_{i+1,M} + δ_h }
      end for;
      For i = 1 to D do
            For j = M -1 to 1 step -1 do
                  { α_{i,j} := α_{i,j+1} + δ_s }
            end for;
      end for;
    }
end if;
return
```

Figure 3-1: Algorithm for generating the matrix α in PRED-S

Table 3.1:10×4 α matrix used by PRED-S in simulation

Dropping probability α matrix $\delta_h = 0.02\ \delta_s = 0.015$				
Burst size (Mbits)/ hops	250	500	750	1,000
1	0.225	0.21	0.195	0.18
2	0.205	0.19	0.175	0.16
3	0.185	0.17	0.155	0.14
4	0.165	0.15	0.135	0.12
5	0.145	0.13	0.115	0.1
6	0.125	0.11	0.095	0.08
7	0.105	0.09	0.075	0.06
8	0.085	0.07	0.055	0.04
9	0.065	0.05	0.035	0.02
10	0.045	0.03	0.015	0

3.4 Performance Evaluation

3.4.1 Simulation detail

Our proposed schemes have been extensively tested using a simulation testbed. The simulation assumes that assembled bursts arrive at the network with Poisson distribution. The arrival rate λ is controllable and both schemes BJIT-S and PRED-S, along with the scheme BJIT and PRED proposed in [20], are tested using various network loads and burst sizes. A source-destination pair is randomly chosen for each arriving burst. To establish the static lightpath, the simulation calculates the shortest path between these nodes using Dijkstra's algorithm as was done in [20, 78-80]. The control message originates from the source node, requesting for a free wavelength at each hop until it either reaches the destination node or gets blocked due to the unavailability of free wavelength at any hop along the path. Our simulation assumes that each node which is an OXC is equipped with full wavelength conversion capability. The source node

waits for a predetermined time depending on the hop distance to the destination before transmitting the optical burst message.

The simulation clock is divided into time units, where each simulation time unit corresponds to 1 millisecond. The optical node and network parameters are similar to those typically used in the literature. Each node has a control packet processing time of 10 milliseconds and its cut through time is set to 1 millisecond. A burst length of 100 units corresponds to 500 Mbits bursts at 5 Gbits/sec transmission rate. In order to evaluate the performance of our proposed schemes, we have used variable burst sizes between S_{min}=250 Mbits to S_{max}= 1000 Mbits and in some of our tests we increased the upper limit to S_{max}=2500 or 3000 Mbits. Each node can have a certain maximum number W of allowed wavelengths and for every run, all the schemes are tested using the same value of W. We used W = 4, 8, 16, 32, 64 and 128 maximum wavelengths in our tests. Each of the performance graphs in this contribution was generated by averaging 7-10 test runs using different randomly generated seeds. We used multiple batch means at 95% confidence interval, and each simulation was run for sufficiently long time (up to 100 million simulated units of time) to obtain stable statistics.

Figure 3-2 and Figure 3-3 show the two network topologies used in our simulation (US Long Haul with 28 nodes and 5X5 Mesh torus with 25 nodes). For the analysis of BJIT-S and PRED-S on various network sizes, we have used mesh topologies of different sizes.

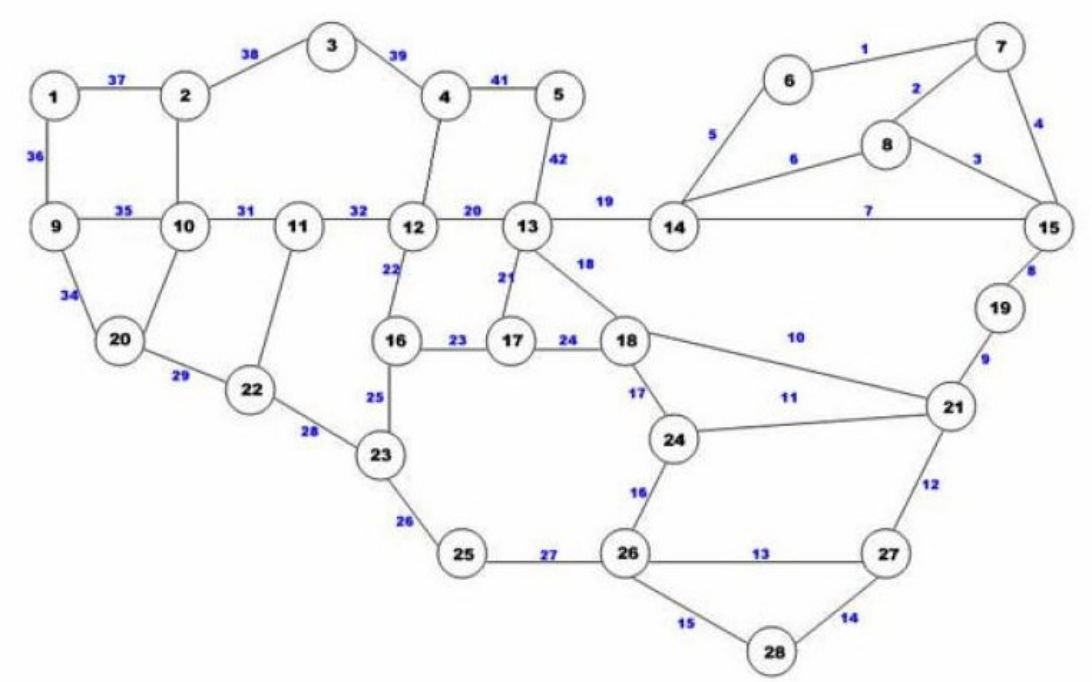

Figure 3-2: US Long Haul Network topology

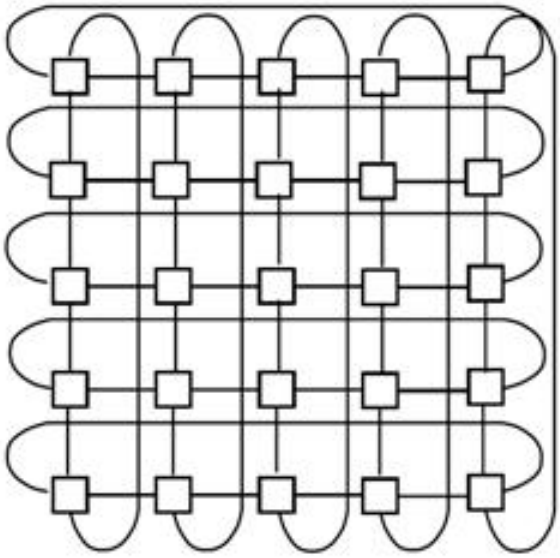

Figure 3-3: 5x5 Mesh Torus topology

Note that the longest lightpath in the US-Long Haul network has the diameter of 8 hops while that of the 5x5 mesh torus network is 4 hops. The traffic used in our simulation is

uniformly distributed among nodes as was done in [20, 79] which means that any node can be a source or a destination. The mesh-torus network has more links than the Long Haul network and it often has multiple shortest-path routes connecting the same source-destination pair. The mesh-torus network therefore requires higher total load than the Long Haul network to induce a certain level of congestion on the individual links.

Our tests have been done under the assumption that the OBS network is employing timer-based burst assembly routines and that the size of the assembled burst in any access node does not depend on the destination of this burst. This is an acceptable assumption that reflects the traffic patterns of practical applications. Under this assumption, the hop count fairness is not affected by the distribution of burst size. This is simply because the hop count of the burst and the size of the burst are two independent variables.

3.4.2 Throughput analysis

Figure 3-4 and Figure 3-5 show the throughput comparison of various schemes including JIT, BJIT, BJIT-S, PRED and PRED-S under different network loads on the US Long Haul and 5x5 Mesh torus networks, respectively. The horizontal axis gives the value of the arrival rate λ as 2, 4, 6, 8, 10 and 12 bursts/unit time and the Y-axis shows the throughput of the network under various schemes as gigabits/time unit, where a time unit (tu) corresponds to 1 millisecond. The value of the parameter g used by BJIT and BJIT-S is g = 0.5. Our simulation tests used randomly generated bursts for each scheme with burst sizes between 250 Mbits to 1000 Mbits. It can be observed that the throughput of BJIT-S and PRED-S is higher than the throughput of the other schemes.

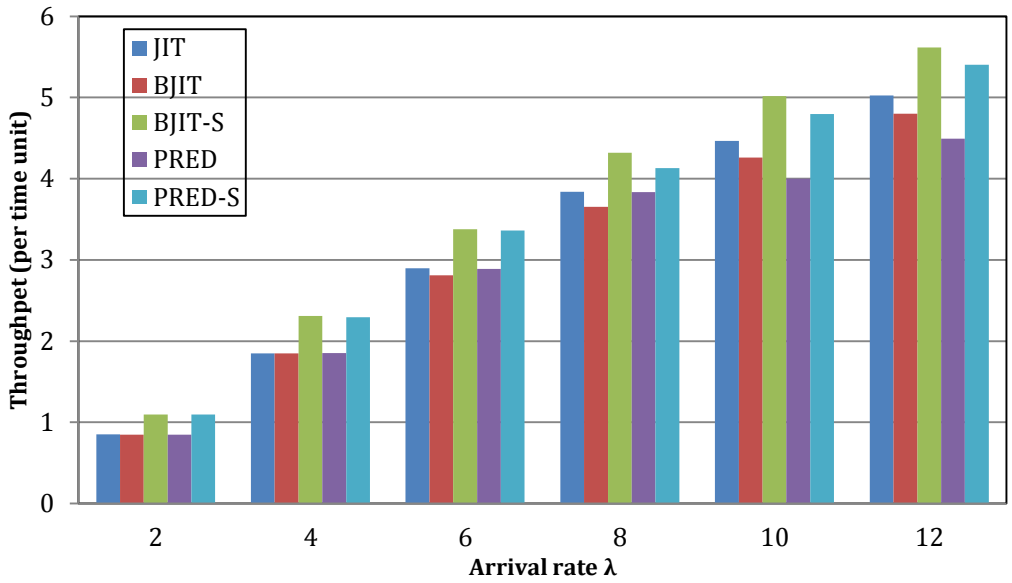

Figure 3-4: Throughput comparison of various schemes on US Long Haul at different loads, W = 64

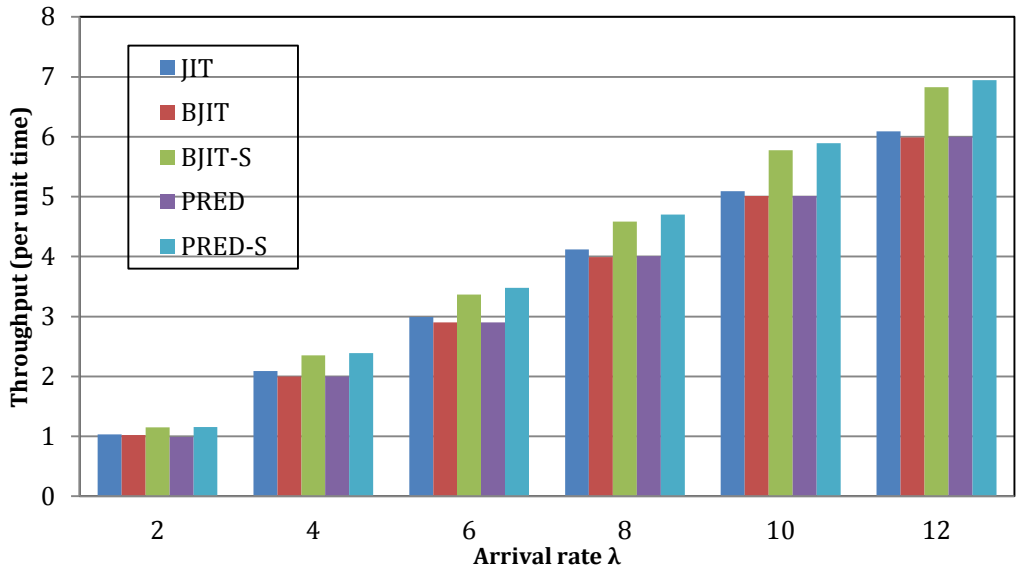

Figure 3-5 Throughput comparison of various schemes on 5x5 Mesh Torus network, W = 64

Figure 3-6 shows a comparison of dropping probabilities of nodes in the US Long Haul network with small and large burst size distributions. For this simulation, we have used burst

sizes between 250 MBits and 1500 MBits. Out of all the nodes, the simulation randomly chooses 25% nodes that are allowed to have small burst size distribution with burst size range between 250 MBits and 750 MBits, whereas, the remaining 75% nodes have the full (large) burst size distribution with burst size range between 250 MBits and 1500 MBits. Since each node calculates its own PRED-S dropping probability matrix based on its minimum and maximum burst sizes, therefore, none of the nodes suffers from the logic of the PRED-S scheme. If a node is only transmitting smaller bursts, this node is not adversely affected by the random early drop strategy of PRED-S as shown in Figure 3-6. Nodes with both small and large burst size distributions have approximately the same dropping probabilities under all network loads.

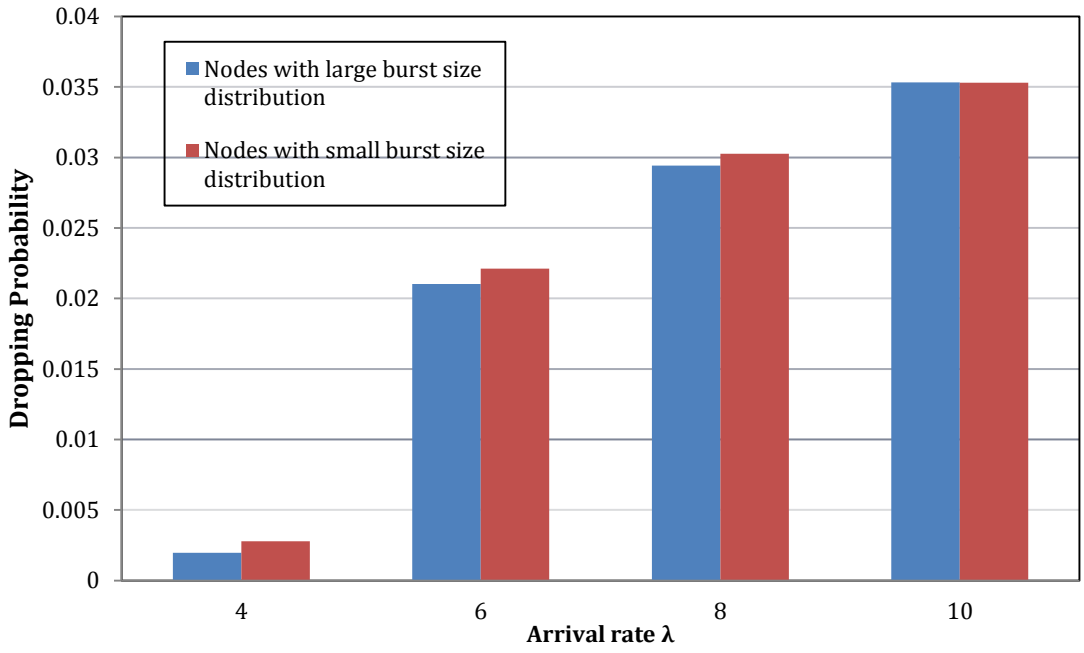

Figure 3-6 Dropping probabilities in PRED-S of nodes with small and large burst size distributions, US Long Haul, W = 64

3.4.3 Fairness analysis

Figure 3-7 and Figure 3-8 show the dropping probabilities for various hop counts under different network loads. The value of the parameter g used by BJIT and BJIT-S is g = 0.5. The figures clearly show that BJIT-S and PRED-S improve the throughput compared to the other schemes without negatively impacting the fairness of the network.

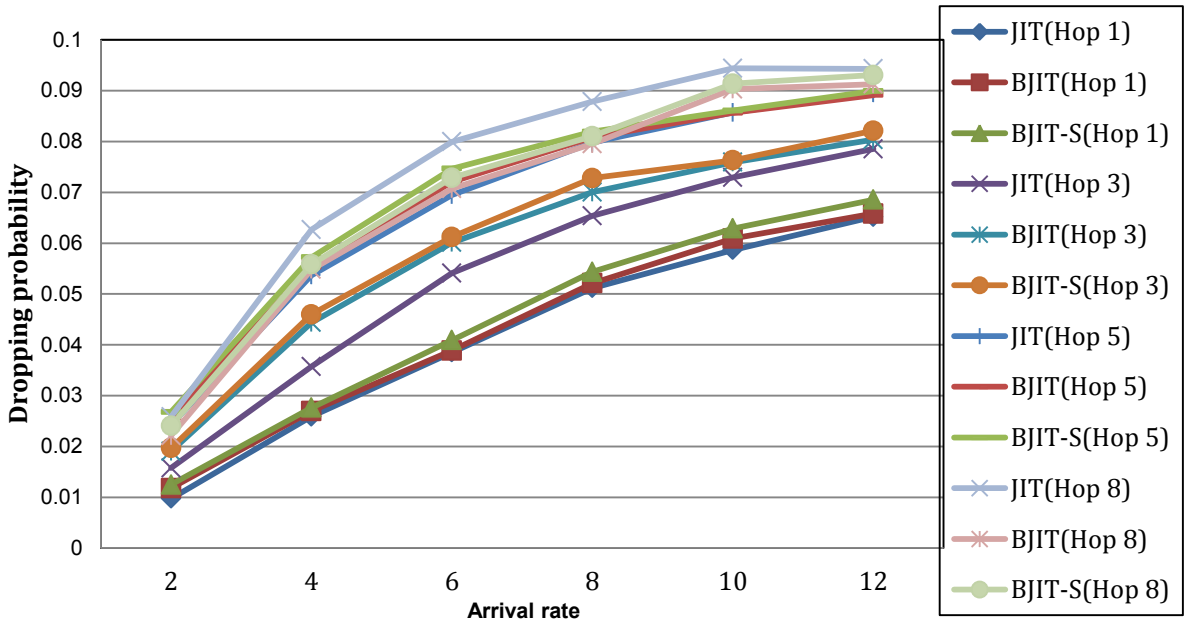

Figure 3-7 Drop Probabilities of JIT vs BJIT vs BJIT-S on US Long Haul with different loads, W = 64

Figure 3-7 shows that the per hop dropping probabilities of the BJIT-S scheme are close to those of the BJIT scheme. Both schemes, BJIT and BJIT-S, have better fairness than the JIT scheme as the dropping probabilities for bursts with larger hop counts in BJIT and BJIT-S are closer to the dropping probabilities for bursts with smaller hop counts than in the case of JIT. Figure 3-8 shows that the per-hop dropping probabilities of PRED-S and PRED are approximately the same even though PRED-S is superior in throughput.

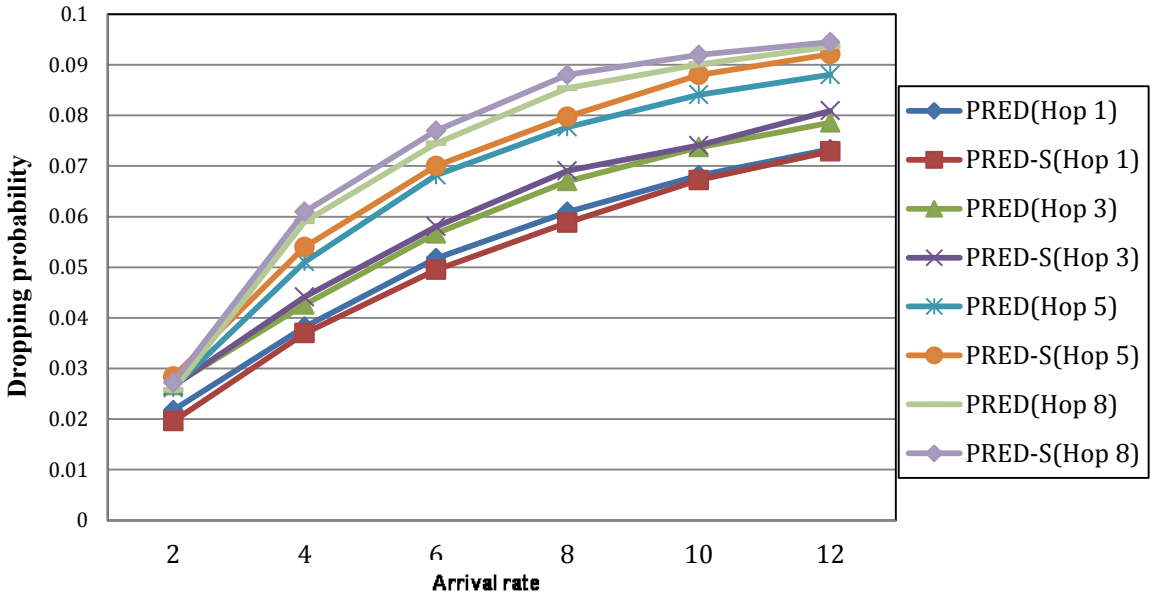

Figure 3-8 Drop Probabilities of PRED vs PRED-S on US Long Haul with different loads, W = 64

The fairness of the three schemes JIT, BJIT and BJIT-S is further compared in Figure 3-9 using the coefficient of variation (standard deviation over mean) of the individual average blocking probabilities for bursts with different hop counts. We call this metric the Unfairness Coefficient (UC); the smaller the value of the unfairness coefficient the better the fairness of the scheme. Figure 3-9 show that our proposed BJIT-S scheme gives more or less the same level of fairness as the BJIT scheme even though it is superior in throughput.

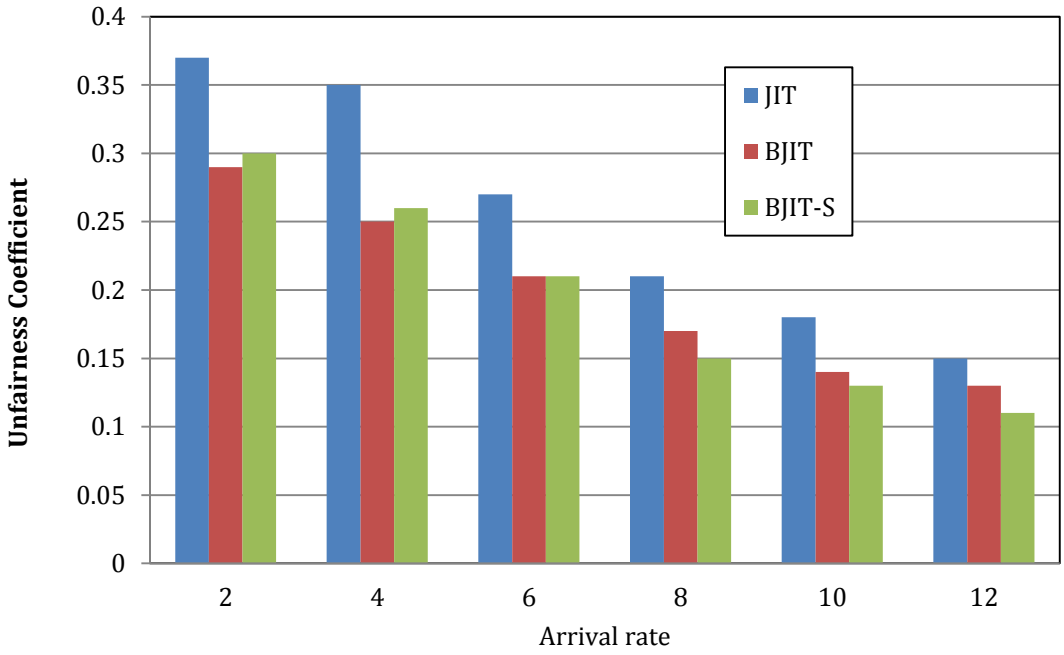

Figure 3-9 Unfairness Coefficient for the US Long haul at g = 0.5, W = 64

3.4.4 Variation of the parameter g in BJIT and BJIT-S

The value of g used in Figures 3-8 was g = 0.5. The authors in [20] suggested that the best compromise between the level of hop-count unfairness and the dropping probabilities was achieved at g = 0.5. The throughput of BJIT was degraded for values of g approaching 1. We tested BJIT-S for different values of g and found that although the trend of decreasing throughput as the value of g increases applies to both BJIT and BJIT-S, the BJIT-S scheme performs slightly better than BJIT as g approaches the value of 1. Our tests used randomly generated bursts with burst sizes between 250 Mbits to 1000 Mbits. Figure 3-10 shows the throughput analysis of BJIT and BJIT-S using various values of g in the US Long Haul network with W = 64 and $\lambda = 4$.

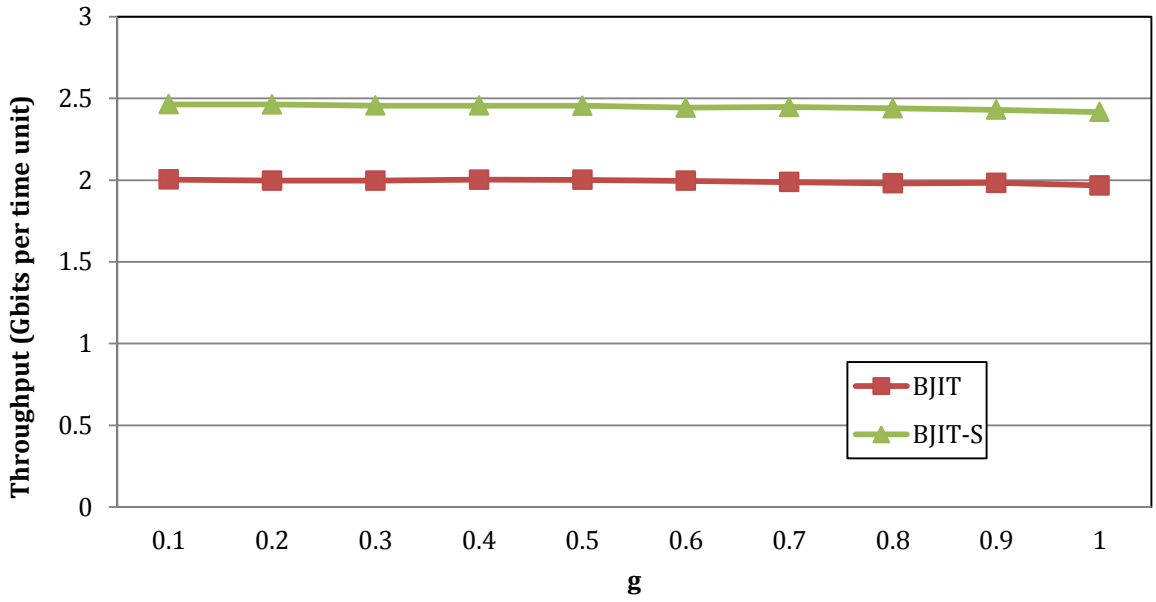

Figure 3-10 Throughput comparison of BJIT-S and BJIT for Long Haul at different values of g, W = 64, λ = 4

Figure 3-11 shows the throughput analysis of BJIT and BJIT-S for various values of *g* in the 5x5 Mesh Torus network with W = 64 and λ = 6. We again observe that there is a decreasing trend of throughput values as g increases but the BJIT-S scheme performs better than BJIT as the value of *g* approaches 1. This means that if fairness is of absolute importance to the network designer and the value of *g* has to be selected in the range 0.5-0.7 to improve fairness, then BJIT-S can outperform BJIT in terms of throughput while still achieve the same level of fairness as was shown in Figure 3-11.

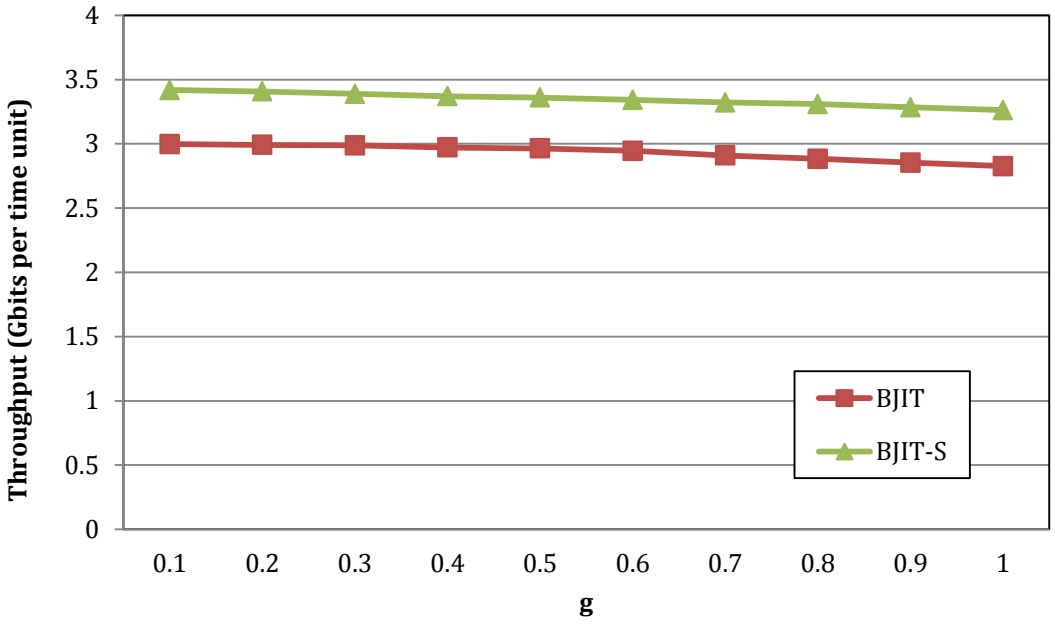

Figure 3-11: Throughput comparison of BJIT-S and BJIT for Mesh Torus using different g values, W = 64, λ = 6

3.4.5 Analysis with mesh grids of different sizes

An L x M two dimensional mesh topology is a grid with length L and width M. The total number of nodes in this mesh grid is N = L x M and its diameter is equal to *(L-1) + (M-1)*. Figure 11 shows the throughput of the different schemes using mesh networks with increasing number of nodes along the X-axis. The size of the bursts used in these test ranged from 750 Mbits to 3000 Mbits. The size of the mesh ranged from a small 3x 3 mesh of 9 nodes to a large 16 x 16 mesh of 256 nodes. The number of wavelengths available at each OXC was W = 64 and the total arrival rate to the entire network was λ = 13 bursts per unit of time (13000 bursts per second corresponding to an average offered load of 24.375 Gbps). For small mesh sizes, the arrival rate exceeds the capacity of the mesh and causes link congestion and burst dropping. As

the size of the mesh increases, the number of links increases and the load on each link decreases causing smaller burst dropping and higher throughput.

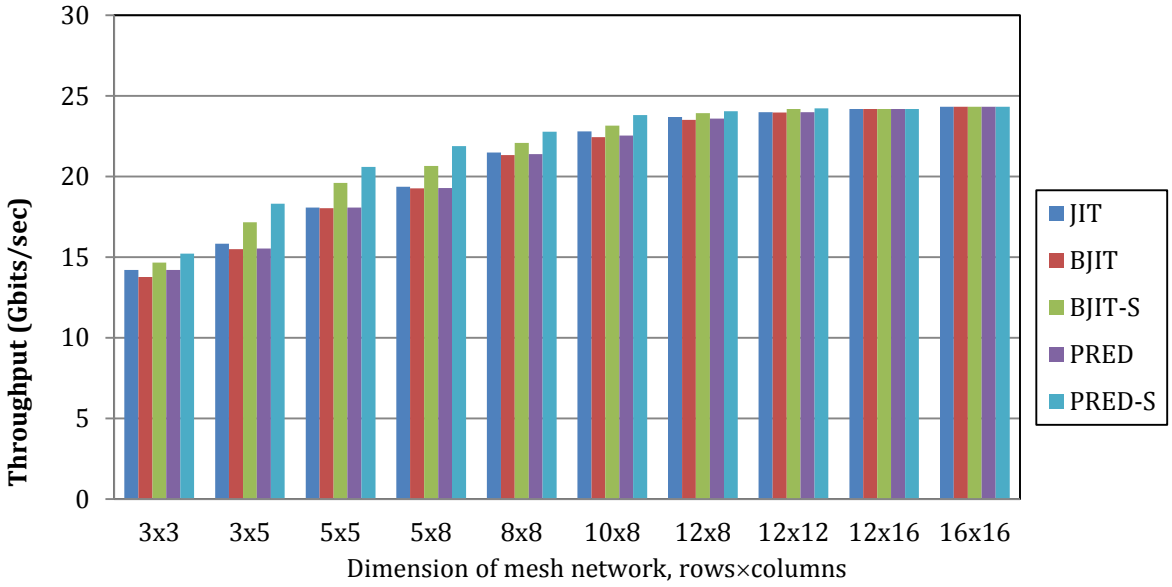

Figure 3-12 Throughput of mesh networks with increasing number of nodes at g = 0.5, W = 64, λ = 13

The arrival rate in Figure 3-12 is λ = 13 for all the mesh configurations. We can observe two phases of behavior when the size of the mesh network is increased.

Phase I: For small mesh networks, e.g., 3×3= 9 nodes, the load of λ = 13 is too high and causes too much dropping. The improvement of PRED-S and BJIT-S over JIT is not significant. When the mesh network gets larger, congestion becomes less and the throughput improves. The improvement of PRED-S and BJIT-S over JIT becomes larger and larger until it reaches maximum improvement at mesh network of size 5×8 = 40 nodes; in this case the throughput of PRED-S is smaller than the average input load of 24.375 Gbps due to some proactive dropping and routing dropping.

Phase II: Exceeding the mesh size above 5x8 at the same load causes the congestion to start disappearing and the throughput of all schemes start approaching the input load because there are no losses. The improvement of PRED-S over JIT will start decreasing until it becomes zero.

3.4.6 Analysis of varying number of wavelengths at OXCs

Figure 3-13 shows the effect of increasing the number of wavelengths W on the throughput of the various schemes in the US Long Haul network with $g = 0.5$ and $\lambda = 4$. We used randomly generated bursts with burst sizes between 250 Mbits to 1000 Mbits. As expected, all the schemes perform better as the network resources, i.e., the number of wavelengths at each OXC is increased. At W=4, the network is quite congested and all schemes have low throughput due to burst dropping; PRED-S has the highest throughput. At higher values of W, e.g., W=16 and W=32, there are more free wavelengths to establish successful lightpaths allowing greater number of successful burst transmissions. The greatest improvement on throughput values is observed in BJIT-S and PRED-S as compared to the other schemes. When W increases to 128, there is almost no congestion in the network and the throughput of all schemes approaches the offered load of 2.5 Gbps.

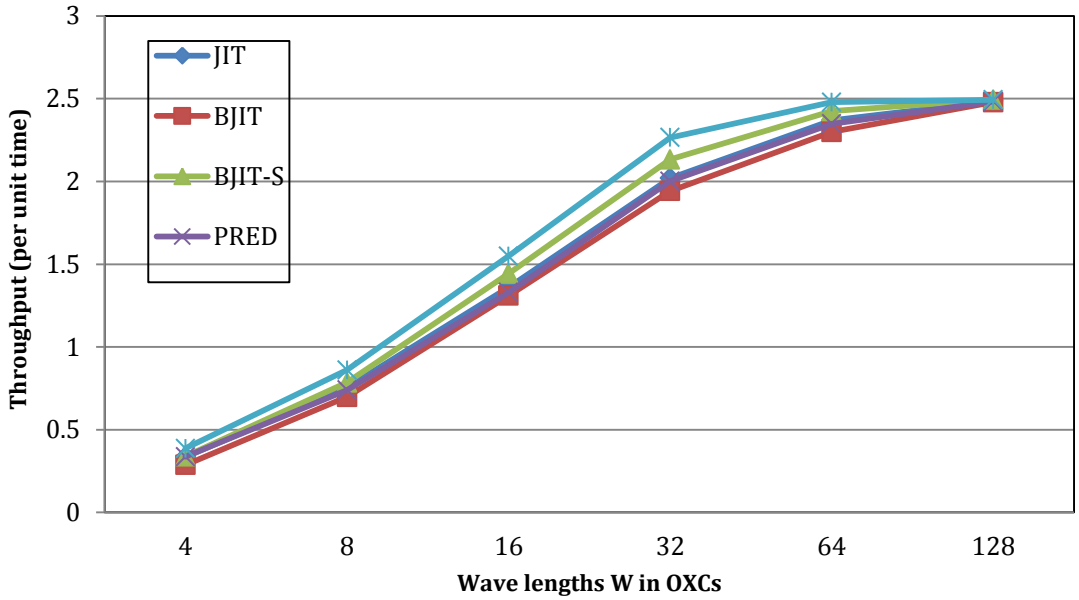

Figure 3-13 Throughput comparison with increasing W at OXCs in US Long Haul, g = 0.5, λ = 4

Figure 3-14 shows a similar trend for the effect of increasing the number of wavelengths W on the throughput of the various schemes in the 5x5 mesh torus network with $g = 0.5$ and $\lambda = 6$.

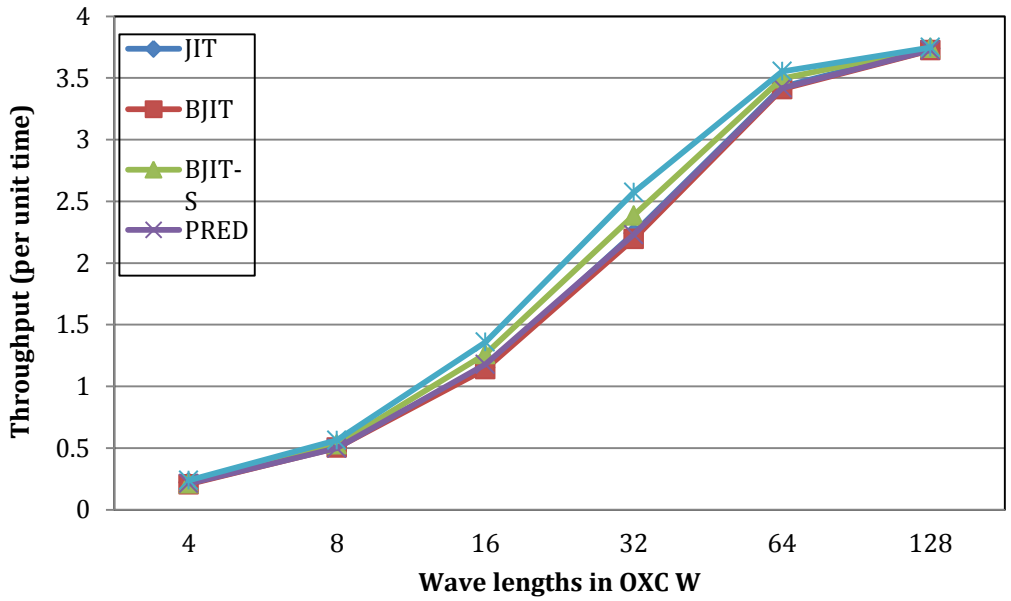

Figure 3-14 Throughput comparison with increasing W at OXCs in 5x5 mesh torus, g = 0.5, λ = 6

3.4.7 Analysis with variation of burst sizes

In this section, we investigate the performance of our proposed schemes using different sets of bursts of increasing range of burst sizes. The different sets used in our tests are given in Table 3.2 Sets of burst sizes.

Table 3.2 Sets of burst sizes

Set	Range of Burst size (Mbits)
Set 1	250-500
Set 2	250-1000
Set 3	250-1500
Set 4	250-2000
Set 5	250-2500

Figure 3-15 shows the throughput for the different sets in the US Long Haul network using arrival rate $\lambda = 6$, $W = 64$ and $g = 0.5$. It can be observed that BJIT-S and PRED-S perform better than the other schemes as the average burst size increases at larger ranges.

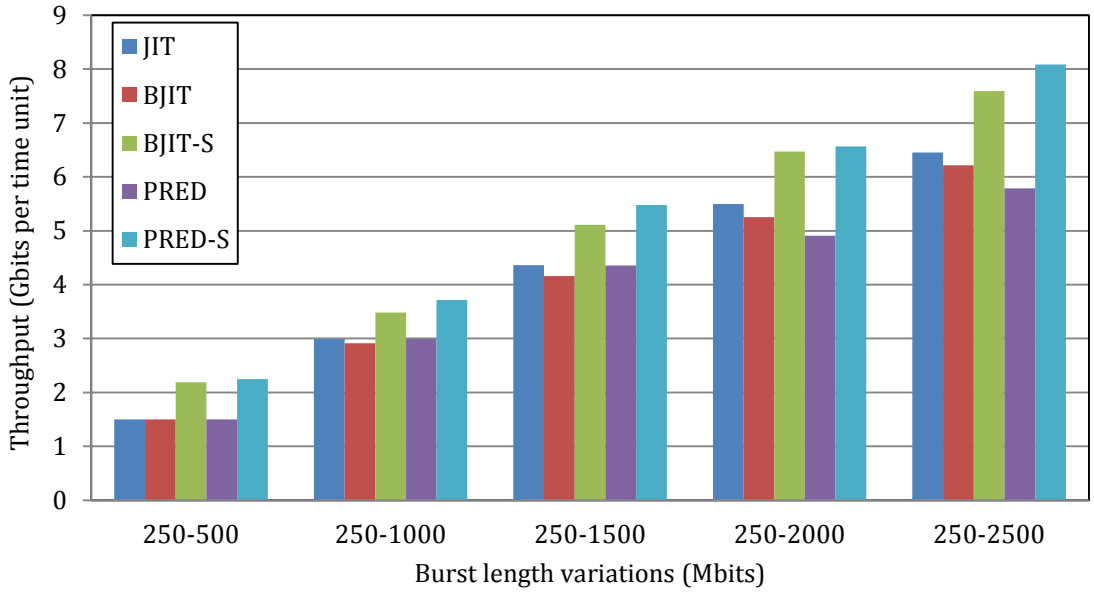

Figure 3-15 Throughput analysis of different burst sizes, US Long Haul, $W = 64$, $g = 0.5$, $\lambda = 6$

3.4.8 Analysis of PRED-S step functions δ_h and δ_s

Algorithm in Figure 3-1 showed how to generate the PRED-S' dropping probability matrix α of Eq. 9 using the two parameters: the hop step function δ_h and the burst size step function δ_s.

Table 3.1 gave the value of α using constant values of $\delta_h = 0.02$ and $\delta_s = 0.015$. In this section, we examine the throughput and unfairness coefficient for different values of δ_h and δ_s in the US Long Haul network and we show how we arrived at the values of δ_h and δ_s that give the best compromise between throughput and fairness.

Figure 3-16 shows the throughput with variable δ_h while keeping δ_s constant at 0.015. It can be observed that the best throughput is achieved when value of δ_h is at minimum. The parameter δ_h deals with hop count fairness and we will investigate the value of δ_h in next three figures to see when the best fairness is achieved.

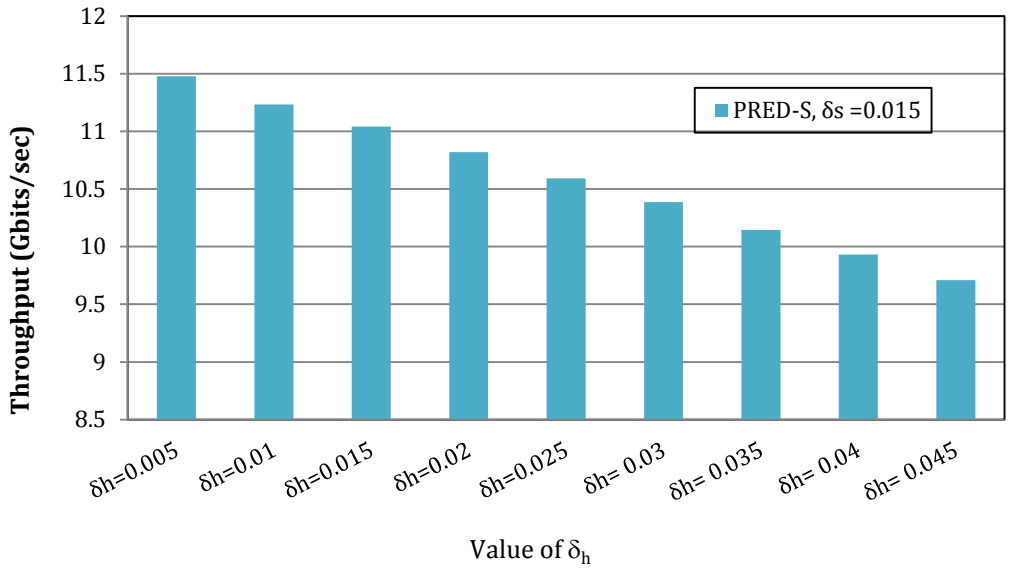

Figure 3-16 Throughput analysis, variable δ_h with δ_s =0.015, US Long Haul, W = 64, g = 0.5, λ = 7

Figure 3-17 shows the throughput with variable δ_s while keeping δ_h constant at 0.02. Again the best throughput is achieved when the value of δ_s is at minimum.

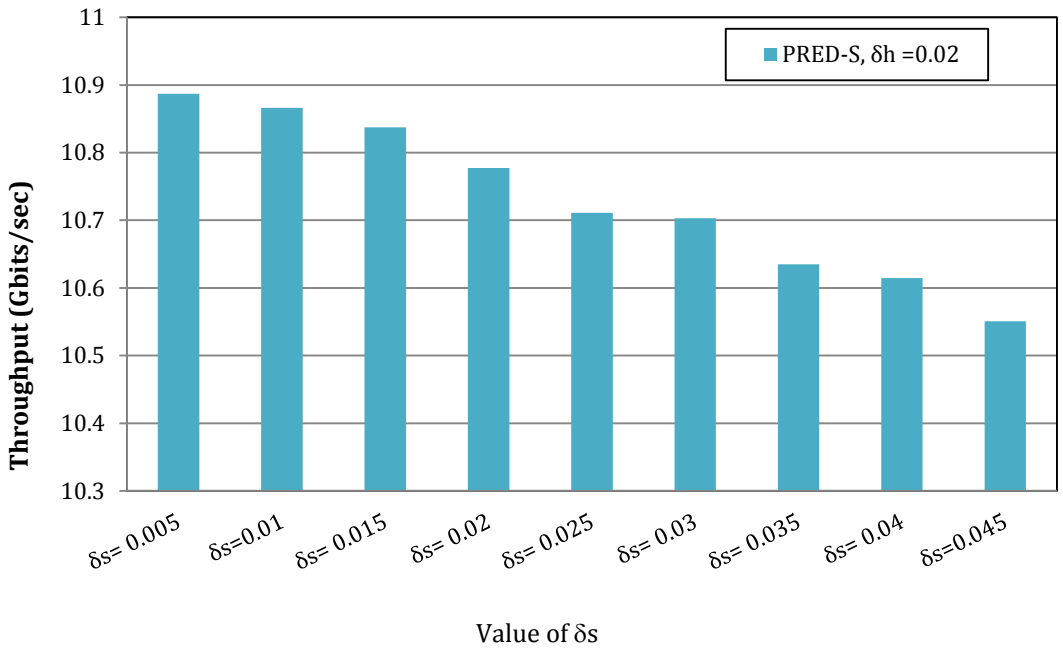

Figure 3-17 Throughput analysis, variable δ_S with δ_h=0.02, US Long Haul, W = 64, g = 0.5, λ = 7

Figure 3-18 shows the unfairness coefficient with variable δ_h while keeping δ_S constant at 0.015. The best (smallest) value of the unfairness coefficient is achieved at δ_h =0.02.

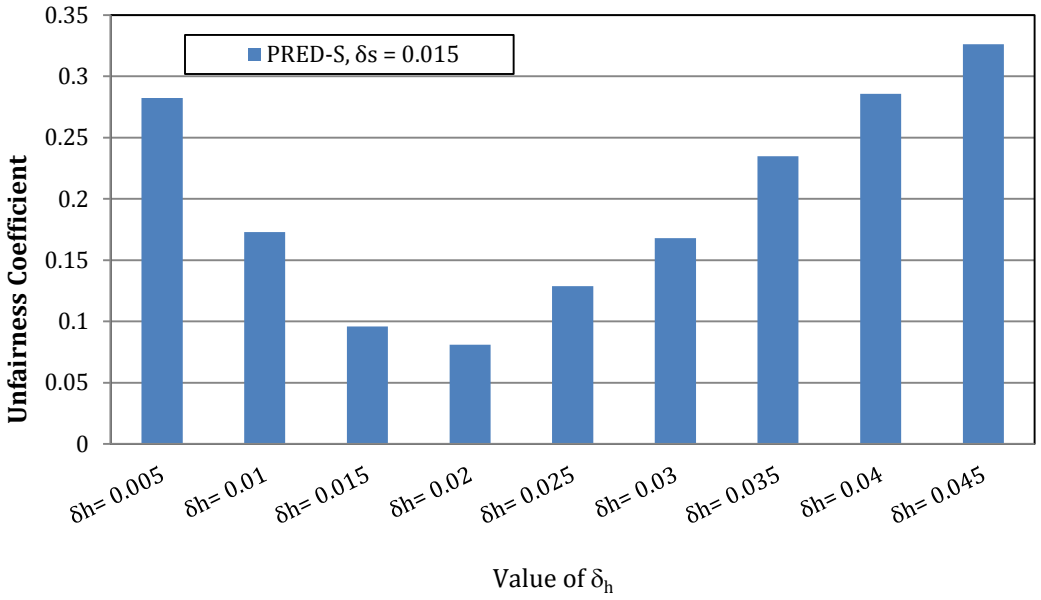

Figure 3-18 Unfairness Coefficient, variable δ_h with δ_s =0.015, US Long Haul, W = 64, g = 0.5, λ = 7

Figure 3-19 shows that best fairness is achieved at δ_h=0.02 and δ_s=0.015. In Fig. 18, the Unfairness Coefficient for the two values δ_s=0.015 and δ_s=0.005 are compared at different values of δ_h for the US Long Haul network with W = 64, g = 0.5, and λ = 7. The value δ_h=0.02 gives the best Unfairness Coefficient and at this value of δ_h=0.02, the value δ_s=0.015 gives better Unfairness Coefficient than the value δ_s=0.005. Therefore the best compromise for throughput and fairness seems to be at the values δ_h=0.02 and δ_s=0.015, which are the values used in Table 1 and in all the performance tests presented in the previous sections.

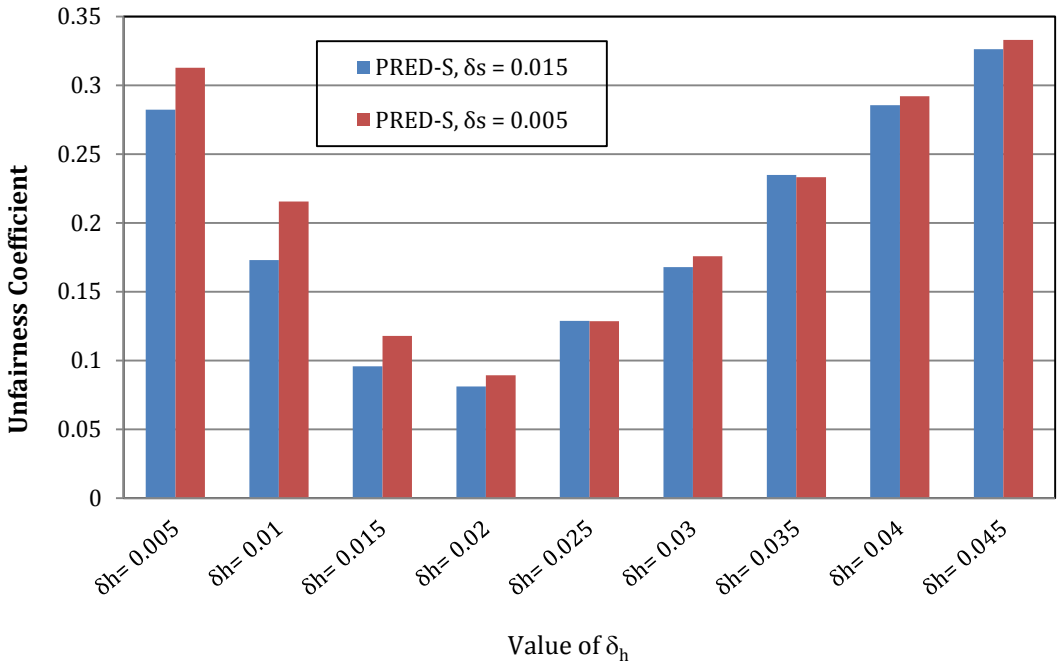

Figure 3-19 Unfairness Coefficient comparisons for δ_h and δ_s, US Long Haul, W = 64, g = 0.5, λ = 7

3.5 Summary

In this chapter we presented two new schemes, BJIT-S and PRED-S, that considered the burst size to maximize throughput without affecting fairness in OBS networks. We evaluated the effectiveness of these schemes in maximizing throughput of the OBS networks with simulations. Our schemes have proven to be effective in maximizing throughput in the US Long Haul and Mesh networks. These networks were extensively tested with variable network loads, various values of factor g, and the number of wavelengths W at OXCs. Under all test conditions, both PRED-S and BJIT-S have shown to perform better compared to JIT, BJIT and PRED.

4. CHAPTER FOUR: IMPROVING FAIRNESS OF OBS IN MULTIMODE FIBER NETWORKS

4.1 Introduction

In this chapter, we propose and evaluate two new schemes for alleviating the fairness problem in optical burst switching networks that use mode-division multiplexing as well as wavelength-division multiplexing. The two schemes use formulas that adjust the size of the search space for a free mode or a free wavelength based on the distance of the current hop of the burst from the source node. Additionally the second scheme uses a formula to adjust the size of the search based on the size of the burst, thereby attaining higher throughput without sacrificing hop count fairness. Extensive performance tests are presented to evaluate the two schemes and analyze their effectiveness in improving fairness either without negatively impacting network throughput or with an improved throughput for the second scheme.

Rest of the chapter is organized as follows. Section 4.2 highlights the motivation for the idea. Section 4.3 describes the proposed idea and Section 4.4 provides a detailed analysis of simulation and performance results. Finally section 4.5 summarizes the chapter.

4.2 Motivation for the proposed idea

OBS networks experience a hop count fairness problem. The optical bursts traveling through longer lightpaths with larger hop counts tend to have higher dropping probabilities than bursts with lighpaths having smaller number of hops. Previously, the hop count fairness problem in OBS networks has been investigated in the context of single mode fibers. In [81], an OBS reservation scheme was proposed using parallel backward reservation paradigm in OBS networks operated under the wavelength-continuity constraint. The fairness was achieved by

classifying bursts into several groups according to their total hop counts and then limiting the number of wavelengths dedicated to the group with shorter-hop bursts. Two schemes were proposed in [20] to alleviate the fairness problem in OBS networks. In the first scheme, the size of the search space for a free wavelength is adjusted based on the number of hops traveled by the burst. The second scheme uses the concept of random early discard (RED); the scheme applies proactive discarding of bursts at the network access station (NAS) using discarding probabilities computed based on the hop count of the lightpath of the burst. In [23], our group proposed fairness-aware hop by hop adaptive routing schemes using metrics based on forward channel reservation or link connectivity. All previous schemes have addressed the fairness problem in the context of single mode fibers. In this contribution, we propose and evaluate schemes for solving the fairness problem in OBS networks that use mode-division multiplexing[82].

4.3 Proposed idea

4.3.1 Fairness formula based Optical Routing (FFOR)

This scheme is proposed in order to address the hop count fairness problem that exists in the standard shortest path first (SPF) algorithm. In the FFOR scheme, at any node, the control packet will try to use the same wavelength and same mode that were used in the previous hop. If this is not possible, it will attempt mode conversion first. It is assumed that mode converters as well as wavelength converters are present in the switching/routing component throughout the network. Using Eq. 4.1 below, the search is conducted for a free mode on the same wavelength used in the previous hop suing a subset of the total set of available modes.

$$\text{Mode search size} = \lceil i * M / D \rceil \tag{4.1}$$

Where, M is the maximum number of modes, D is diameter of the network (the largest lightpath in the network) and i is the current hop. The factor i/D determines the size of the subset of modes to be searched among the total modes M and it increases with the number of hops travelled by the burst; when $i=D$, all M modes are searched. The ceiling function is used to yield an integer number of modes subset which should be at least equal to 1. For example if the network diameter is $D=10$, the number of modes searched is $0.1*M$ at the first hop, $0.2*M$ at the second hop, and so on.

If no mode is free, FFOR then attempts wavelength conversion. Eq. 4.2 determines the subset of wavelengths that can be searched, keeping the same mode as the previous hop.

$$n_i = (1-g)*W_M + g*i*W_M / D, \quad 0 \le g \le 1 \tag{4.2}$$

Where, n_i is number of wavelengths searched at the i^{th} hop, W_M is the maximum number of wavelengths per mode and g is a constant between 0 and 1 inclusive. Because the value of M is practically much smaller than the value of W_m, we have used a different equation for the wavelength search that always includes a subset with a base size. The parameter g divides the search spectrum in each OXC into two parts: a base part and an adjustable part. The base part has a fixed size of $(1-g)*W_M$ wavelengths regardless of the hop count of the lightpath. The adjustable part gives higher priority (larger wavelength subset) to the burst having travelled larger distance and can reach a maximum size of $g*W$ wavelengths. For example if the network diameter D is 10, the size of the adjustable part is $0.1*g*W$ at the first hop, $0.2*g*W$ at the second hop, etc.

The parameter g controls the degree of effectiveness of resolving fairness. Generally speaking, the larger we assign a value to g, the better fairness we can obtain but at the expense of

slightly dropping some bursts which have shorter hops to destination. Best value of g has been found to be around 0.5.

If none of these wavelengths is free, FFOR will start searching in the entire subset with both mode and wavelength conversion using Eq. 4.3.

$$\textit{Wavelength-mode search size} = (1-g)*W_M*M + g*i*M*W_M/D, \quad 0 \le g \le 1 \qquad (4.3)$$

It can be seen in each of the above three equations that the subset for either modes or wavelengths or both depends on the current hop distance of the control packet from the source OXC. When the burst is closer to the destination, a larger subset of wavelengths or modes is searched to find and reserve a free wavelength & mode. If at a particular node, the control packet is unable to find a free wavelength from the designated subset on all available modes, the packet is considered blocked and gets dropped.

4.3.2 Fairness Throughput Formula based Optical Routing (FTFOR)

FFOR is sufficient for improving fairness in OBS networks but does not attempt to improve throughput further. Typically in an OBS network, the arriving bursts are of different sizes and a bandwidth reservation technique can simply look into the burst size in order to enhance the overall throughput of the system. FTFOR, the new scheme presented here, will incorporate the burst size to positively enhance throughput. We will introduce a new term, the *size factor* η which is the ratio of current burst's size S to the maximum allowed burst length S_{max}. The search equations for FTFOR are given below.

$$\textit{Mode search size} = \lceil i*M*\eta/D \rceil \qquad (4.4)$$

where,

$\eta = S / S_{\max}$

S: burst length

S_{max}*:* max allowed burst length

Eq. 4.4 yields a larger subset of modes for larger bursts and longer lightpaths thereby improving fairness and throughput. The role of the factor *i/D* and the ceiling function being the same as already mentioned in the FFOR scheme.

If the mode search fails, we try wavelength conversion. The wavelength search subset is given by:

$$n_i = (1-g)*W_M*\eta + g*i*W_M*\eta/D, \quad 0 \le g \le 1 \tag{4.5}$$

In Eq. 4.5, we have introduced η = size factor. With the presence of the size factor, a bigger subset of wavelengths is searched for larger bursts thereby giving them a higher probability to reach the destination successfully. If the above search fails, we need to change both wavelength and mode using Eq. 4.6.

$$\text{Wavelength-mode search size} = (1-g)*W_M*M*\eta + g*i*M*W_M*\eta/D, \quad 0 \le g \le 1 \tag{4.6}$$

The size factor η adjusts the wavelength search subset based on the size of the current burst, and allows a larger number of wavelengths to be searched for larger bursts. Consequently, for two bursts of different sizes but with the same hop count, FTFOR will allow a larger wavelength search space to the burst that is of larger size. Because the hop count of the burst is independent of its size, FTFOR tends to have the same level of fairness as FFOR but achieves higher throughput.

4.4 Performance Results

Our proposed schemes have been extensively tested using a simulation testbed written in C++. The simulation assumes that assembled bursts arrive at the network with Poisson distribution. The arrival rate λ is controllable and both schemes FFOR and FTFOR are tested and compared with SPF using various network loads and burst sizes. A source-destination pair is randomly chosen for each arriving burst. To establish the static lightpath, the simulation calculates the shortest path between these nodes using Dijkstra's algorithm. The network nodes are assumed to be equipped with mode as well as wavelength converters. The control packet which originates from the source node acquires an initial free wavelength & mode then travels to the destination using the Just-in-Time signaling protocol [19]. When blocked at the next hop, the control packet searches for the same wavelength on all available modes. If the same wavelength is not available then it tries wavelength conversion and if not successful it tries both mode and wavelength conversion. The process continues until the control packet either reaches the destination node or gets blocked due to the unavailability of free wavelength on all modes at any hop along the path. The source node waits for a predetermined time depending on the hop distance to the destination before transmitting the optical burst message.

The simulation clock is divided into time units, where each simulation time unit corresponds to 1 millisecond. The optical node and network parameters are similar to those typically used in the literature. Each node has a control packet processing time of 10 milliseconds and its cut through time is set to 1 millisecond. In order to evaluate the performance of our proposed schemes, we have used variable burst sizes between S_{min}=250 Mbits to S_{max}= 1000 Mbits. Each node can have a certain maximum number W of allowed wavelengths and all

the schemes are tested using the same value of W. Each of the performance graphs in this contribution was generated by averaging 7-10 tests where each test was run for a sufficient large number of time units to produce stable results.

The topologies used in our simulation tests are the US Long Haul Network with 28 nodes and a 5x5 Mesh torus Network. The longest lightpath in the US Long Haul network has the diameter of 8 hops while that of the 5x5 mesh torus network is 4 hops. The traffic used in our simulation is uniformly distributed, i.e., any node can be a source or a destination. The mesh-torus network has more links than the Long Haul network and it often has multiple shortest-path routes connecting the same source-destination pair. The mesh-torus network therefore requires higher total load than the Long Haul network to induce a certain level of congestion on the individual links.

Solutions to remedy fairness usually have the side effect of decreasing the overall throughput of the network. Before examining the fairness performance, we will show that the proposed schemes do not negatively impact the throughput of the network and that FTFOR actually improves the throughput.

Figure 4-1 shows the throughput of the US Long Haul network for the schemes SPF, FFOR and FTFOR under various available numbers of modes with burst arrival rate of 35 bursts/s. It can be observed that the throughput of FFOR is roughly the same as that of SPF or very slightly smaller while the throughput of FTFOR is generally higher. It is also interesting to note that the gain in throughput for increasing the number of modes is multiplicative, e.g., the throughput with three modes is triple the throughput with a single mode.

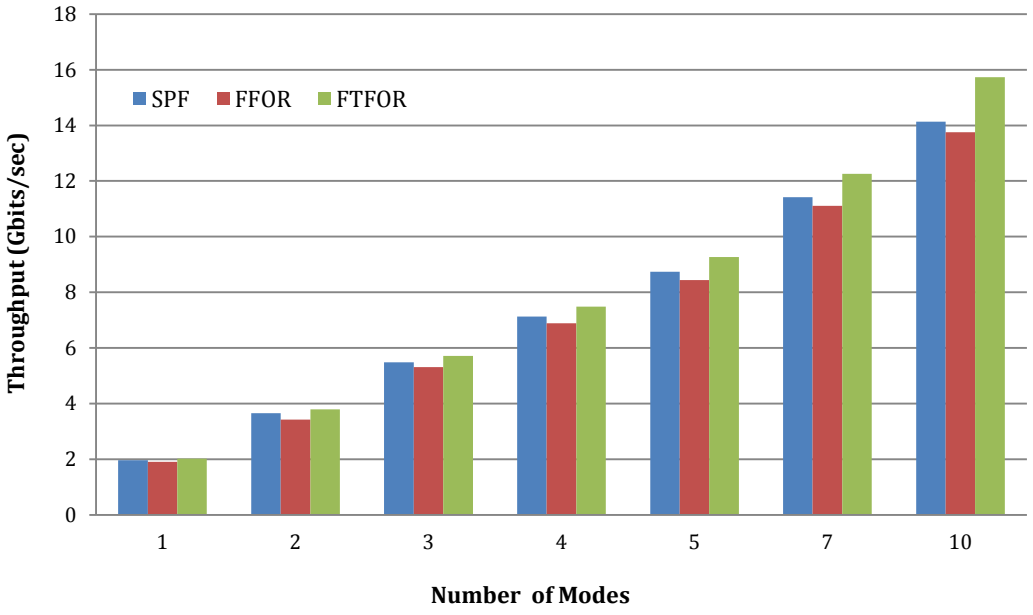

Figure 4-1Throughput comparison in US Long Haul network, Max wavelengths W=20, arrival rate=35/s, g=0.5

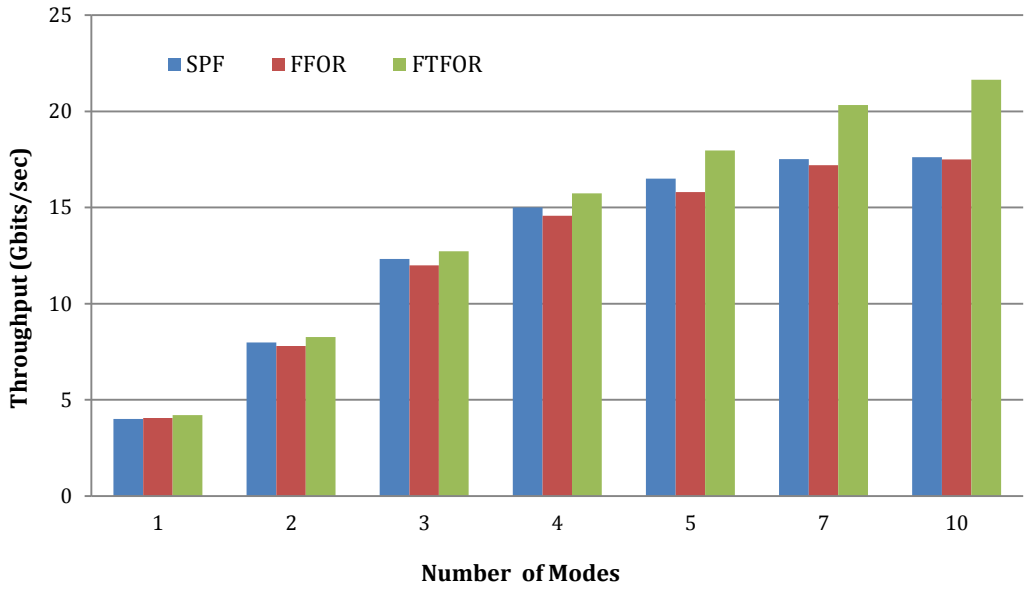

Figure 4-2: Throughput comparison in 5x5 Mesh Torus network. Max wavelengths W=20, arrival rate=35/s, g=0.5

Figure 4-2 shows the throughput for the 5x5 mesh torus network with a burst arrival rate of 35/s. Again the schemes SPF, FFOR and FTFOR show the multiplicative trend of increasing throughput when the number of modes is increased. FTFOR performs best while the throughput of FFOR is very slightly smaller; this is a very small penalty for achieving better fairness.

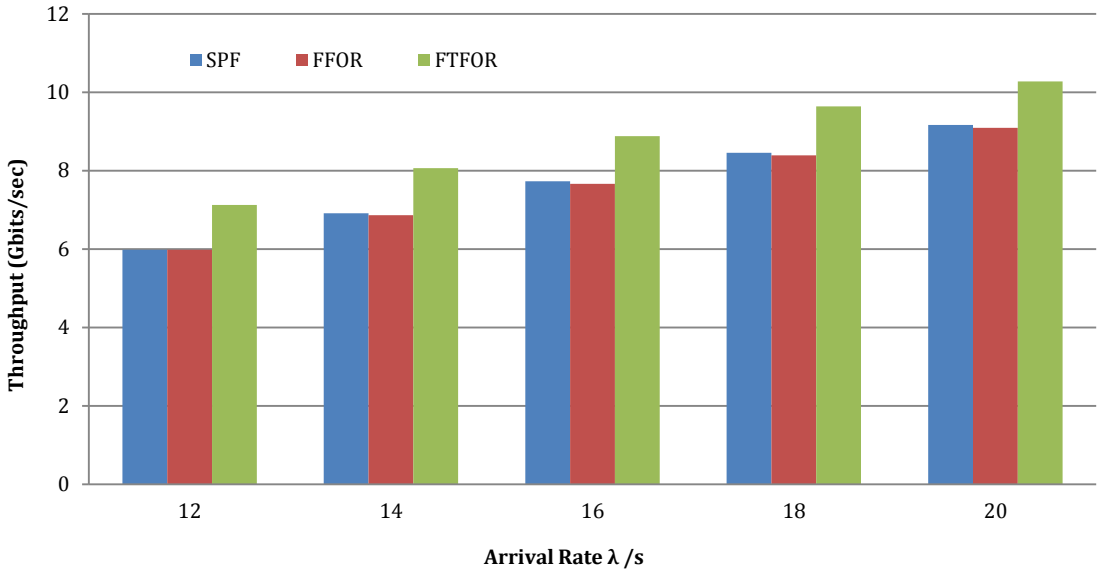

Figure 4-3: Throughput comparison in 5x5 Mesh Torus network. Max wavelengths W=20, g=0.5, modes=3

Figure 4-3 shows the throughput of the three schemes with three modes and different arrival rates for the 5x5 mesh topology; similar results have been obtained for the US long Haul topology.

We next investigate the fairness performance by examining the per-hop dropping probabilities.

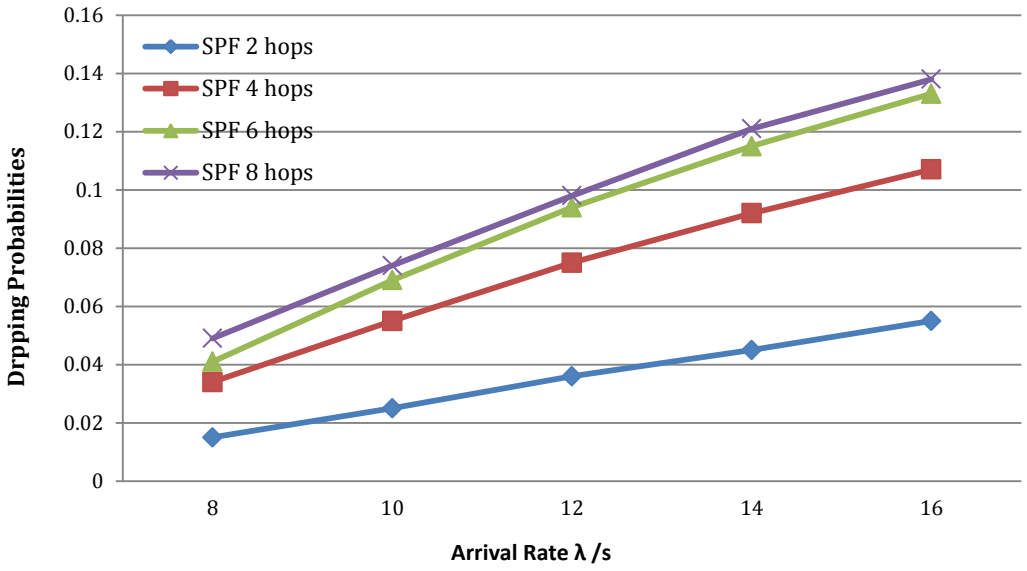

Figure 4-4: Per hop dropping probabilities in US Long Haul network -SPF. Max wavelengths W=20, g=0.5, modes=3

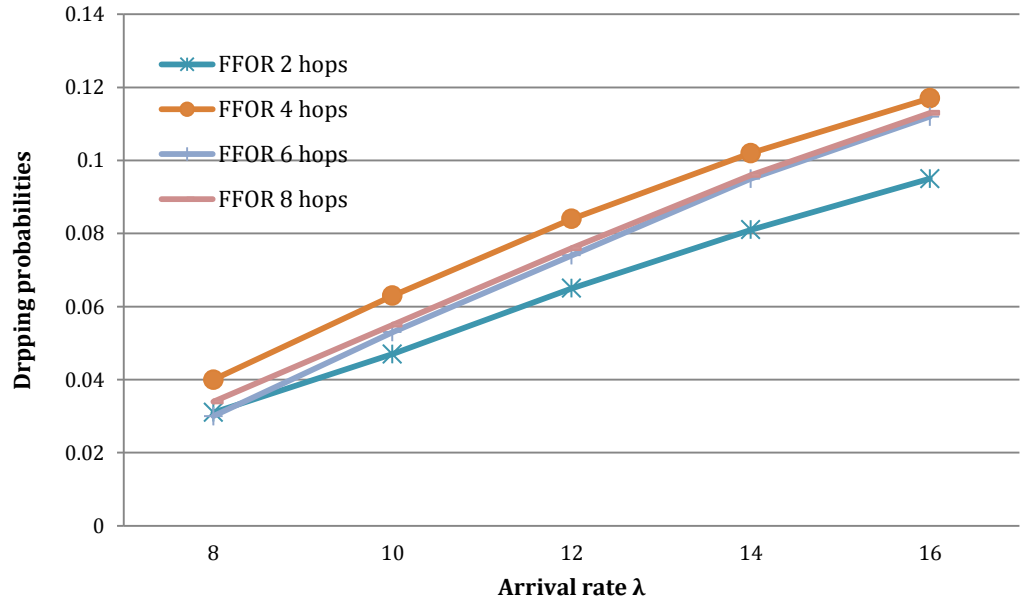

Figure 4-5: Per hop dropping probabilities in US Long Haul network-FFOR. Max wavelengths W=20, g=0.5, modes=3

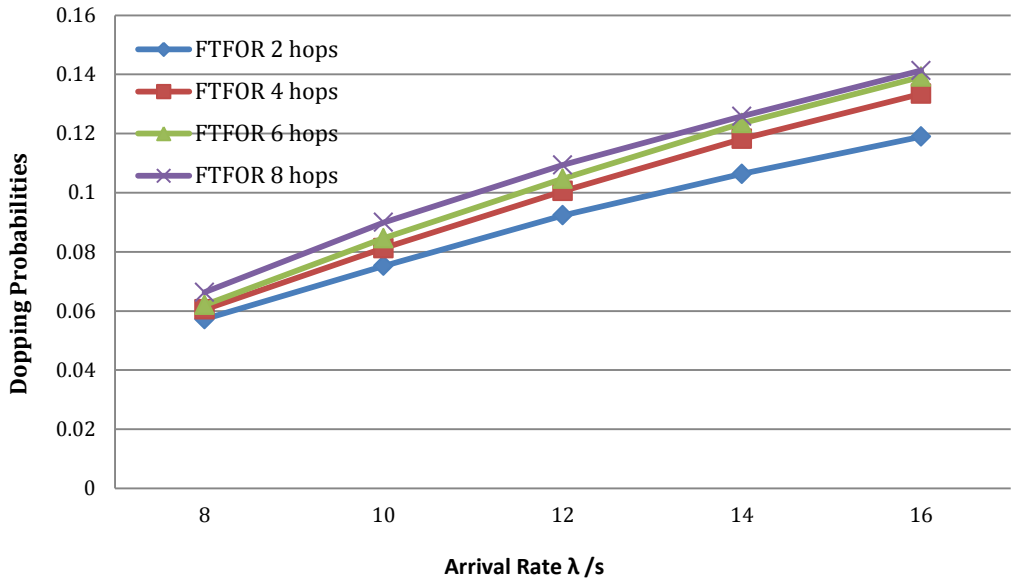

Figure 4-6 Per hop dropping probabilities in US Long Haul network FTFOR. Max wavelengths W=20, g=0.5, modes=3

Figure 4-4 shows the per hop dropping probabilities in the US Long Haul for SPF while Figure 4-5 and Figure 4-6 show the corresponding per hop dropping probabilities for FFOR and FTFOR, respectively. It can be observed from these figures that the dropping probabilities in SPF for smaller hop counts (e.g. 2 hops or 4 hops) are much less than the dropping probabilities for larger hop counts (e.g., 6 hops or 8 hops). This is the expected behavior of all optical routing schemes that do not have fairness-improving mechanisms. Under SPF and similar routing protocols, the delivery of bursts between two nodes far away from each other is much less reliable and has lesser throughput than the delivery of bursts between two nodes that are near each other. Figure 4-5 and Figure 4-6 clearly show that the bias against bursts with longer lightpaths has substantially decreased when FFOR or FTFOR are used. Compared to Fig. 4, small and large hop counts in Figure 4-5 and Figure 4-6 exhibit small differences in the blocking

probabilities at all arrival rates. It is no longer the case that a connection between two distant (far away) nodes will have significantly less throughput than a connection between two nearby nodes.

To evaluate the fairness of the proposed schemes FFOR and FTFOR, we calculate in Table 4.1, the coefficient of variation (standard deviation over mean) of the individual average blocking probabilities for bursts with different hop counts. We call this metric the Unfairness Coefficient: the smaller the value of the unfairness coefficient the better the fairness of the scheme. Table 4.1 shows the improved fairness of FFOR and FTFOR over SPF for all the arrival rates. The coefficient of unfairness decreases with increasing arrival rate λ. It can be clearly observed that both new schemes FFOR and FTFOR have much better fairness in multimode fiber networks than the standard SPF routing protocol.

Table 4.1 Unfairness Coefficient for U.S. Long Haul

Arrival Rate λ	SPF	FFOR	FTFOR
8	0.60	0.39	0.39
10	0.56	0.38	0.37
12	0.54	0.36	0.36
14	0.53	0.35	0.35
16	0.51	0.34	0.34

4.5 Summary

In this chapter, we have proposed a new scheme FFOR that has proved to improve fairness in OBS networks for multimode fiber networks. An additional scheme FTFOR is also introduced that attempts to maximize throughput while maintaining the fairness of FFOR by

selectively giving priority to larger bursts over smaller bursts. Multi mode fiber networks is expected to be one of the next big breakthroughs in the field of optical networks and the schemes proposed in this contribution represent a first attempt to solve the fairness problem in multimode OBS networks. Extensive simulation tests have shown the effectiveness of the proposed schemes.

5. CHAPTER FIVE: ROUTING AND MODE-WAVELENGTH ASSIGNMENT IN MULTIMODE FIBER NETWORKS

5.1 Introduction

In this chapter, we motivate the use of mode-division multiplexing as an additional degree of freedom to enhance the performance of optical networks. We show the significant benefits of using both mode division multiplexing and wavelength division multiplexing in real-life short-distance optical networks such as the optical circuit switching networks used in the hybrid electronic-optical switching architectures for datacenters. We next evaluate four mode and wavelength assignment heuristics and compare their throughput performance. To our knowledge, this is the first research work that evaluates mode division multiplexing and presents results on mode-wavelength assignment for wavelength-mode-routed optical networks. We conclude this chapter by evaluating the impact of the cascaded mode conversion constraint on network throughput.

Rest of the chapter is organized as follows. In section 5.2, we demonstrate the viability and significant benefits of using both wavelength and mode division multiplexing. We also explain the concept of mode-wavelength switching over an arbitrary network. In section 5.3, we show the benefits of using mode-wavelength division multiplexing (MWDM) in datecenter network topologies. We evaluate four mode and wavelength assignment heuristics in section 5.4. In section 5.5, we present preliminary results of the impact of the cascaded mode conversion contstraint on network throughput.

5.2 Utilization of mode-wavelength switching

In addition to the crucially important research on the hardware and device level implementation, research efforts on higher levels of MDM networking is needed to prepare for the successful deployment of MDM in short-distance optical networks. Current optical networks are enabled by wavelength-division multiplexed (WDM) transmission systems serving as point-to-point links between routers. WDM transport also presented the possibility of high-throughput routing and switching in the optical domain in which bits/packets/bursts of information carried on an entire wavelength are switched and routed completely in the optical domain using devices such as (reconfigurable) add/drop multiplexers [(R)OADM] and optical cross connects (OXC). Optical networks based on WDM transport, ROADMs and OXCs are called wavelength-routed optical networks (WRONs), shown schematically in Figure 5-1.

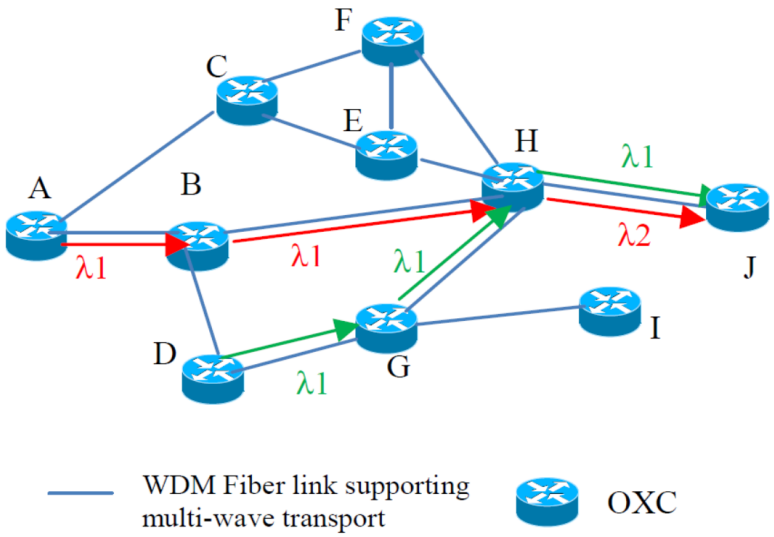

Figure 5-1 Schematic of WRON

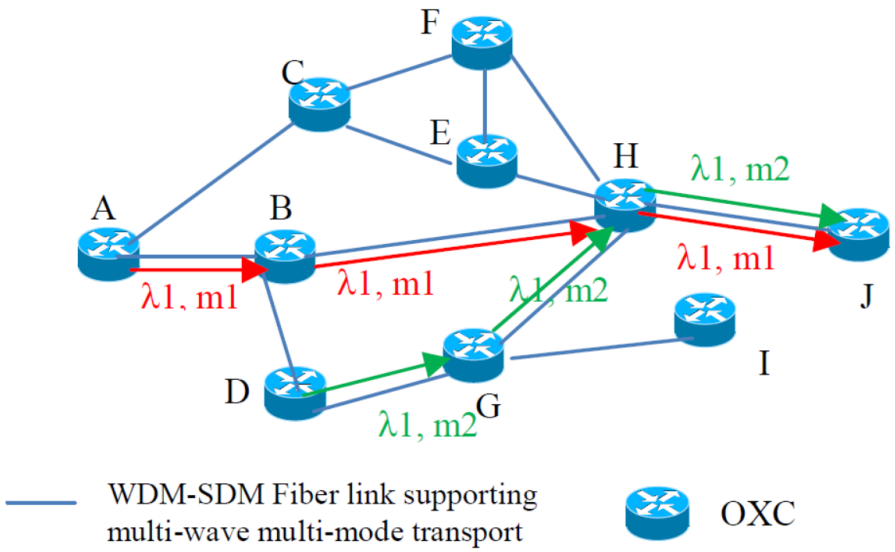

Figure 5-2 Schematic of WMRON

Figure 5-1 depicts two connections with overlapping lightpaths. The lightpath of the first connection originating from node A to node J is identified by the red color and the lightpath of the second connection originating from node D to node J is identified by the green color. Notice that the connection from node A to node J uses two different wavelengths identified by the labels λ_1 and λ_2. In particular, the wavelength for the connection from node A to node J is converted at node H. This wavelength conversion is needed because otherwise the two connections will have conflict in the link from node H to node J. In WRON, each connection must be identified by a unique wavelength. The number of connections is limited by the number of wavelengths supported by the WDM transport system.

In this chapter, we investigate a (spatial) mode-routed optical network, in combination with wavelength routing, i.e., wavelength- and mode-routed optical network (WMRON), as shown in Figure 5-2. The purpose of WMRON is to reduce the blocking probability and increase the throughput of future optical networks. Each optical transport link in WMRON supports not

only multiple wavelengths (same as WDM) but also multiple spatial modes for each wavelength. As a result, the two connections in the WMRON shown in Figure 5-2, corresponding to those in the WRON shown in Figure 5-1, can be carried on the same wavelength but using two different modes. We assume that the wavelength-mode routed OXC (WMROXC) can route any mode (group) on any specific wavelength on any particular input fiber to a specific mode on a specific wavelength on the corresponding particular output fiber.

5.3 Performance Evaluation of Mode-wavelength division multiplexing (MWDM)

Hybrid network architectures supporting both optical circuit switching (OCS) and electronic packet switching (EPS) have received considerable attention as promising networking architectures for datacenters [49, 50, 83-88]. In these hybrid architectures, optical circuit switching (OCS) is used for transmitting longer (stable) flows and electronic packet switching (EPS) is used for transmitting smaller (bursty) flows. Some common examples of hybrid OCS and EPS architectures are Helios [49], HyPaC [85] and Proteus [50]. In this section, we demonstrate the significant benefits of using mode wavelength multiplexing in the OCS component of hybrid packet-optical circuit switched data center networks. We assume that the optical switch within the hybrid architecture is equipped with both mode division multiplexing (MDM) and wavelength division multiplexing (WDM) capabilities.

Currently, the WDM optical circuit switches in the datacenter hybrid architectures are used to provide bandwidth support by routing a limited number of paths in a reasonable period of time. The electrical packet switches in these hybrid architectures are used to provide fast communication among thousands of rack servers and other database servers. Our main goal in this contribution is to motivate and evaluate the concept of combined mode-wavelength division

multiplexing (MWDM) and demonstrate the benefits of integrating MWDM into the optical components of the various hybrid electronic-optical switching architectures proposed in literature. Our contribution is not intended to provide ranking or performance comparison among these hybrid architectures nor modify or improve their fundamental algorithms.

For the purpose of illustrating the benefits of mode wavelength multiplexing in this section, we select the standard shortest path first (SPF) algorithm and we use the notation SPF_WDM to denote routing using WDM only and SPF_MWDM to denote routing using both MDM and WDM. For mode and wavelength assignment, we select the First-Fit heuristic. The performance of other mode and wavelength assignment heuristics will be examined in section V.

5.3.1 Network Topologies

Our OCS simulation testbed assumes that the flows arrive at the network following a lognormal distribution with an ON-OFF pattern. Several research studies of datacenter traffic characteristics have reported that the lognormal distribution with an ON-OFF pattern accurately represents traffic behavior in data centers [43, 46] and recent research on hybrid OCS for datacenters used the lognormal distribution in their performance tests [89]. All the tests reported in this contribution use the lognormal distribution but we have included, as a benchmark, one test with a Poisson arrival since the Poisson distribution has also been used in the literature to model traffic in data centers [90, 91]. The load in the lognormal arrival pattern is controlled by two variables: the mean μ and the standard deviation σ. In the case of the Poisson distribution, the load is controlled by the arrival rate λ. A source-destination pair is randomly chosen for each arriving flow. The two schemes SPF_WDM and SPF_MWDM are tested using various network loads and flow sizes. To establish the static lightpath for a source-destination pair, the simulation

software calculates the shortest path between the source and destination using Dijkstra's algorithm. The optical network nodes, i.e., the optical switches within the hybrid electronic-optical switching architecture, are assumed to be equipped with mode as well as wavelength converters. The simulation clock is divided into time units, where each simulation time unit corresponds to 1 microsecond. The mean circuit setup time of optical circuit switching is assumed to be of the order of 11 microseconds, which is the same setup time reported in the Mordia (Microsecond Optical Research Datacenter Interconnect Architecture) [92, 93].

The bandwidth-intensive long-living flows (also called elephant flows) constitute the background traffic in datacenters [94]. The OCS network is mostly used for serving these long background flows. In our simulation tests, we have used three ranges for the sizes of the background flows: small, medium and large. Within each range, the sizes of the flows are uniformly distributed. The three ranges are as follows:

Small size range: from $S_{min=}$25Mb to $S_{max =}$ 250 Mb

Medium size range: from $S_{min=}$200Mb to $S_{max =}$ 800 Mb

Large size range: from $S_{min=}$500 Mb to S_{max} = 1250 Mb

We adopt the same uniform traffic model used in [23], i.e., any host node can be a source or a destination. Because current hardware implementations for MDM are suitable for short distances, our simulation tests used modern datacenters topologies such as FatTree and BCube shown in

Figure 5-3.3 and Figure 5-4, respectively. We also used the OCS subset of hybrid networks such as Helios [49] and Mordia [92, 93]. The hybrid network of Helios is shown in

Figure 5-5. The links connecting a Core node to Pods make up an optical subset that has the star topology shown in

Figure 5-6. The Mordia hybrid datacenter network is shown in Figure 5-7 and its optical component uses the ring network shown in Figure 5-8.

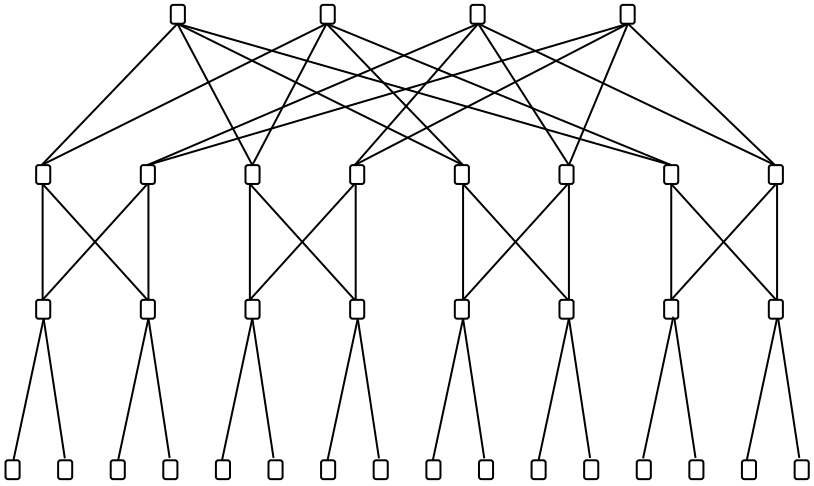

Figure 5-3 Fat Tree

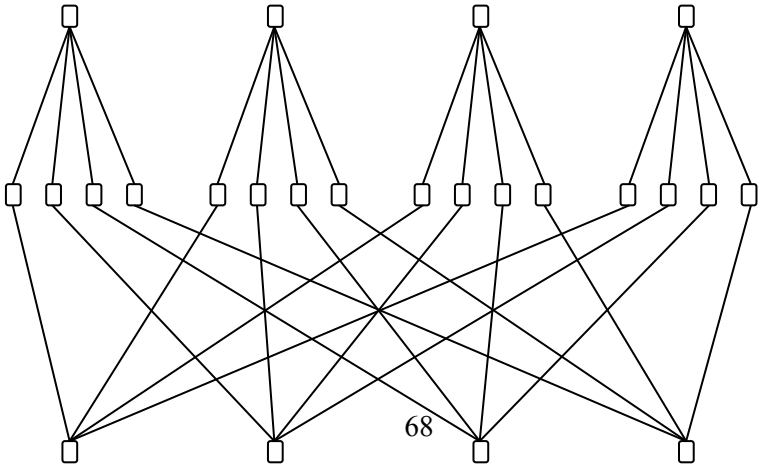

Figure 5-4 BCube

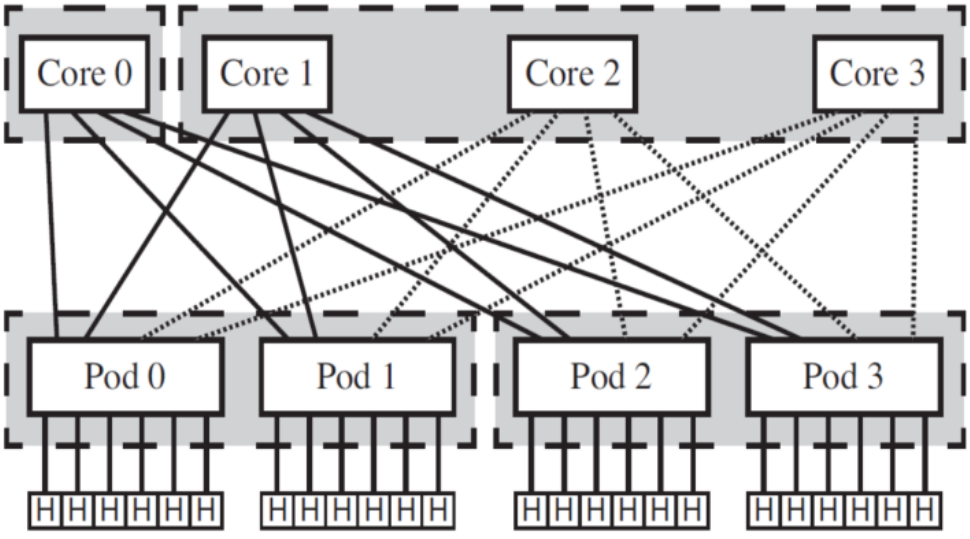

Figure 5-5 Helios network [49]

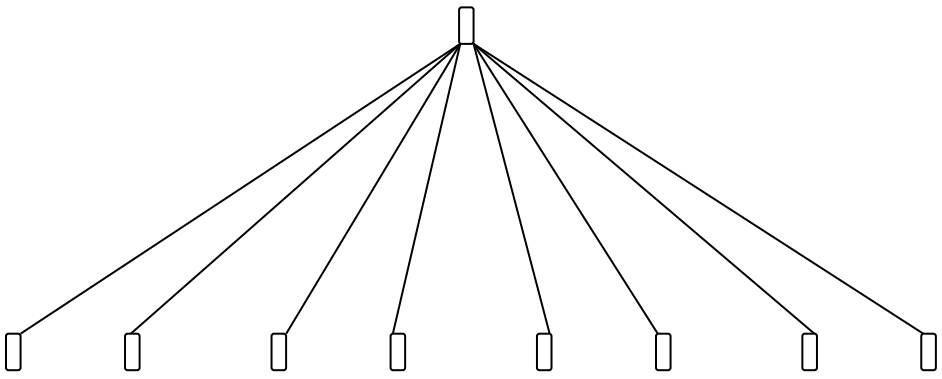

Figure 5-6 Optical subset of the Helios network

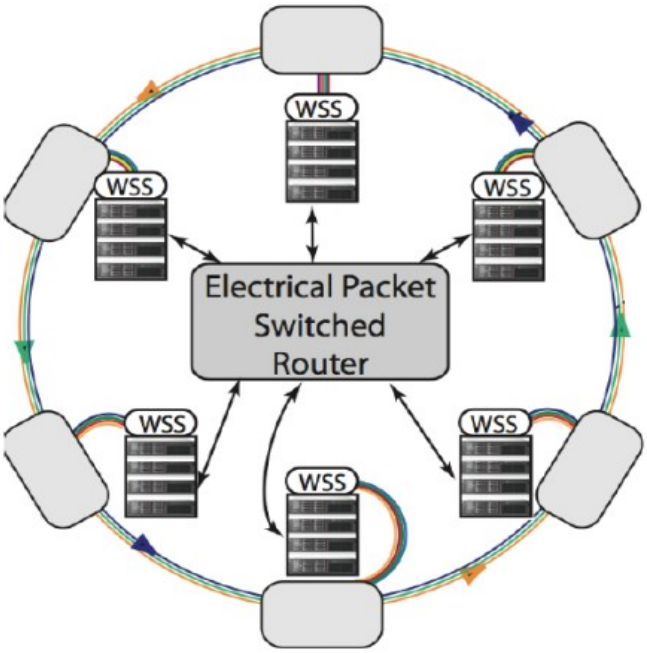

Figure 5-7 Mordia hybrid datacenter network [93]

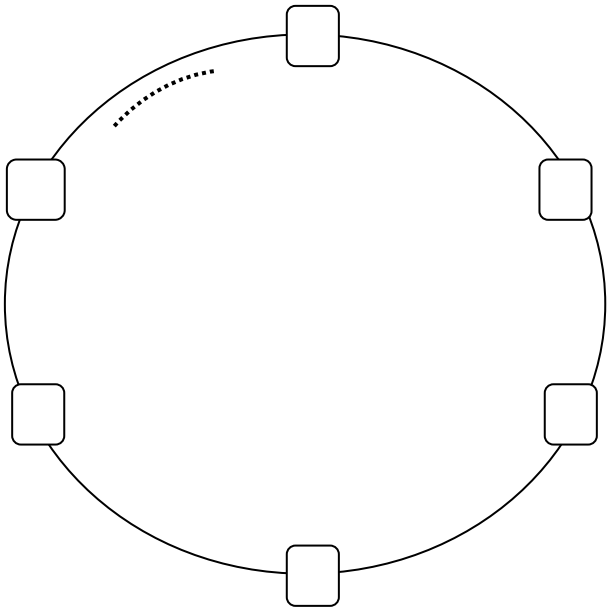

Figure 5-8 Optical ring of Mordia hybrid datacenter network

5.3.2 Performance Results

We have performed extensive performance tests using a wide range of parameter values for flow arrival rates, flow sizes, number of wavelengths W, number of modes M, number of nodes in the Helios star topology, and number of nodes in the Mordia ring topology. In this section, we present the results of a selection of these tests.

Figure 5-9 shows the throughput of the 26 node BCube network with increasing number of modes. SPF_WDM is the shortest path first scheme with only one mode per fiber while SPF_MWDM is the scheme that supports multiple modes and multiple wavelengths with mode conversion and wavelength conversion capability using the First-Fit heuristic for mode and wavelength assignment. Figure 5-9 is tested using the lognormal distribution with mean μ=2.75, standard deviation σ=1 and flow sizes in the small size range of 25 -250 Mb. Because of the ON-OFF feature of traffic in datacenters, the average arrival rate in Figure 5-9 is smaller than the arrival rate of a continuous lognormal process having the same mean and standard deviation. The input load is much higher than the saturation load of 0.524 Mbits/second for the network operating under SPF_WDM. Although the scheme SPF-WDM uses only a single mode (M=1), its saturation throughput is plotted in higher values of modes in Figure 5-9 for the purpose of convenience of comparison with SPF-MWDM. As the number of modes is increased,Figure 5-9 shows that the improvement in throughput has a multiplicative trend in general.

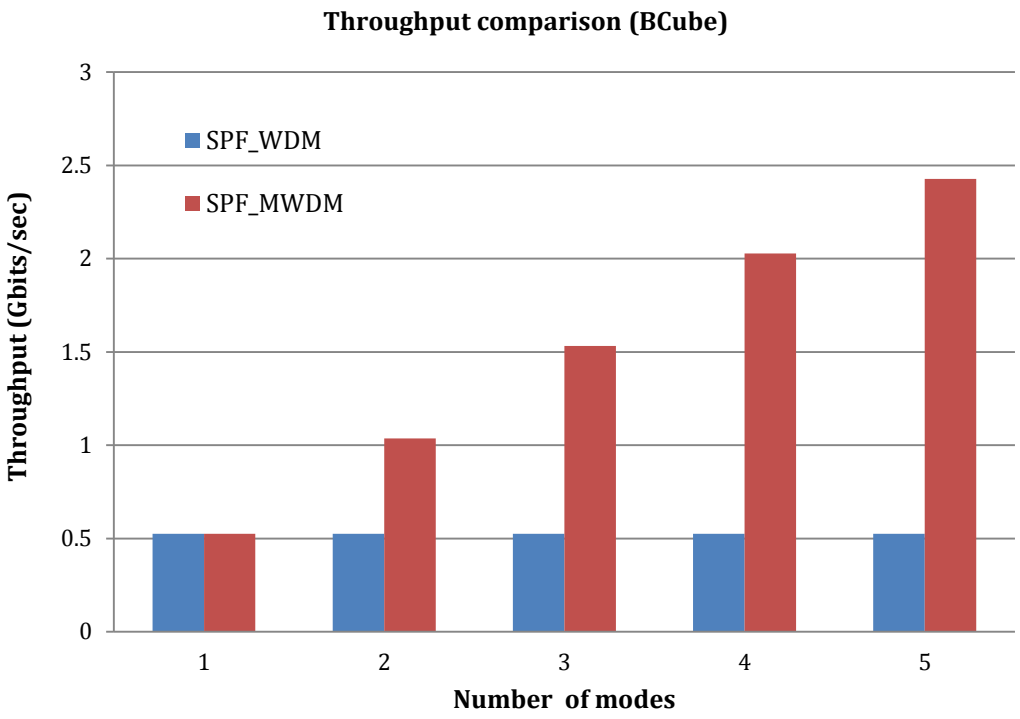

Figure 5-9 Throughput comparison BCube, Max Wavelengths=20, Arrival rate= 18.1 flows/s

Figure 5-10 shows the corresponding throughput of the 36-node FatTree network with increasing number of modes. The flow sizes used in this test are in the large size range of 500 - 1250 Mb. The traffic pattern for flow arrivals has the lognormal distribution with ON-OFF behavior, characteristic of datacenters, with mean μ=3.625 and standard deviation σ=1. It can be observed again that with the introduction of multiple modes in fibers, the throughput can be improved multiplicatively as the number of modes increases.

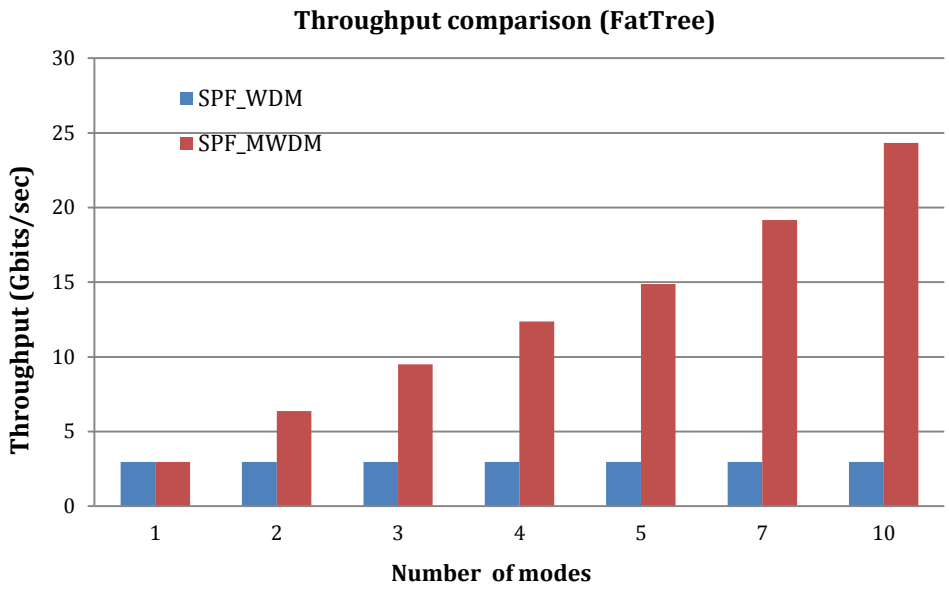

Figure 5-10 Throughput comparison FatTree, Max Wavelengths=20, Arrival rate 44.5 flows/s

Figure 5-111 shows the corresponding throughput of the 24-node Mordia ring with increasing number of modes. The flow sizes used in this test are in the small size range of 25 - 250 Mb. The traffic pattern for arrivals has the lognormal distribution with ON-OFF behavior, with mean μ=2.075 and standard deviation σ=1. It can be observed again that with the introduction of multiple modes in fibers, the throughput can be improved multiplicatively as the number of modes increases.

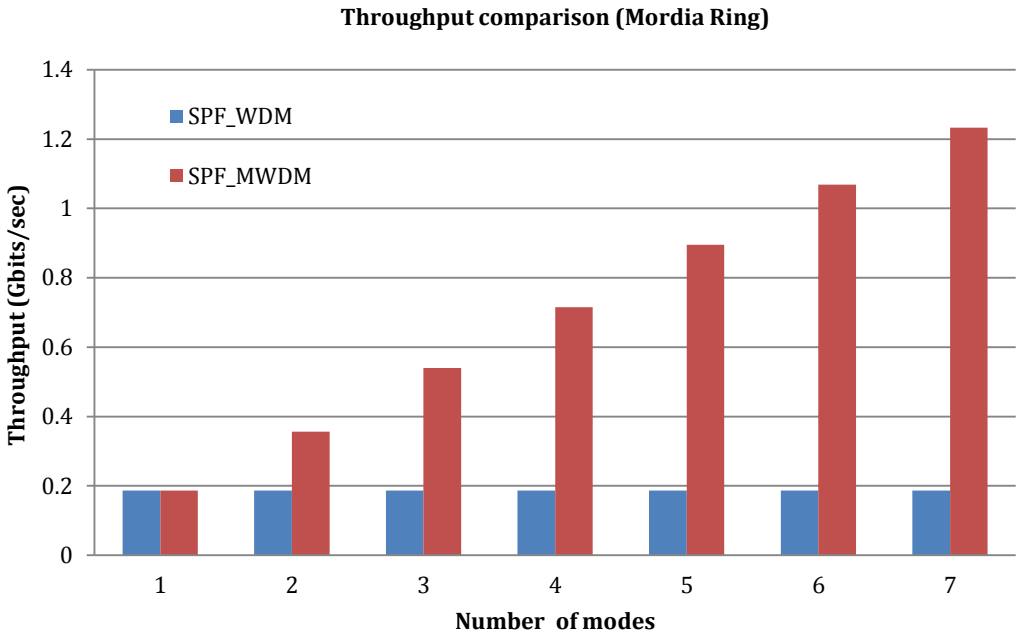

Figure 5-11 Throughput comparison Mordia Ring, Max Wavelengths=24, Arrival rate 9.37 flows/s

Figure 5-12 shows the corresponding throughput of the 16 node Helios star with increasing number of modes. The flow sizes used in this test are in the small size range of 25 - 250 Mb. The traffic pattern for flow arrivals has the lognormal distribution with ON-OFF behavior, with mean $\mu=3$ and standard deviation $\sigma=1$. As before, it can be observed that with the introduction of multiple modes in fibers, the throughput can be improved multiplicatively as the number of modes increases.

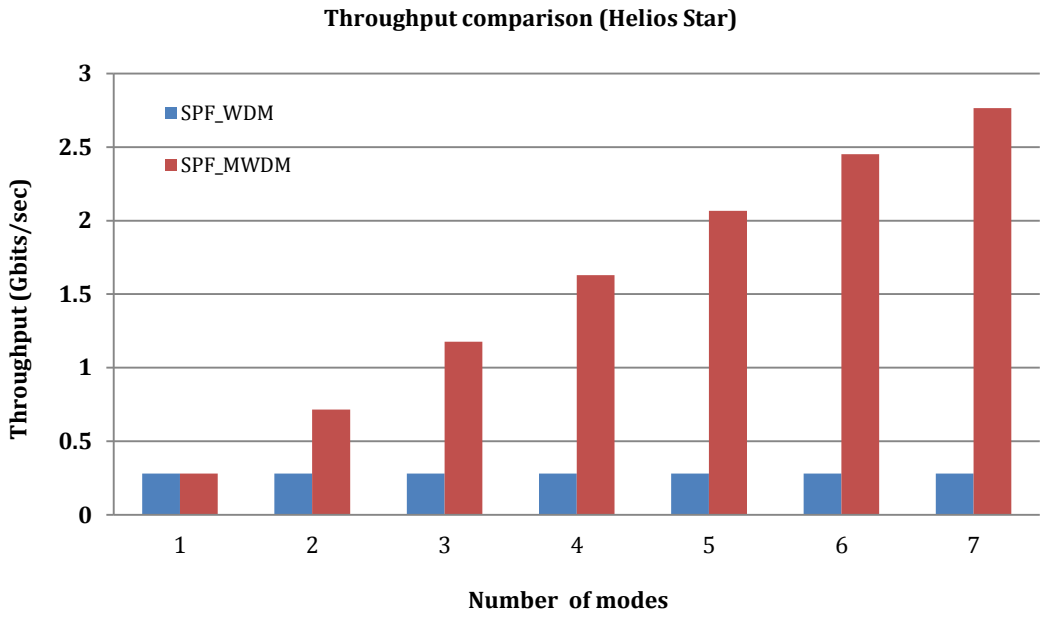

Figure 5-12 Throughput comparison Helios Star, Max Wavelengths=20, Arrival rate 23.87 flows/s

Figure 5-13 shows the performance comparisons of three routing schemes in the BCube network. The arrival rate has the lognormal distribution with mean μ=2.125 and standard deviation σ=1, and flow sizes are in the small range of 25-250 Mb. The first scheme is SPF_WDM with W wavelengths and one mode per fiber. The second scheme is also a single mode SPF_WDM but with M*W wavelengths. The third scheme is the multimode scheme SPF_MWDM with M modes and W wavelengths. For example, if W=8 and M=2, the first SPF routing scheme uses WDM with 8 wavelengths, the second SPF scheme uses WDM with 16 wavelengths, and the third scheme uses both MDM with 2 modes and WDM with 8 wavelengths.

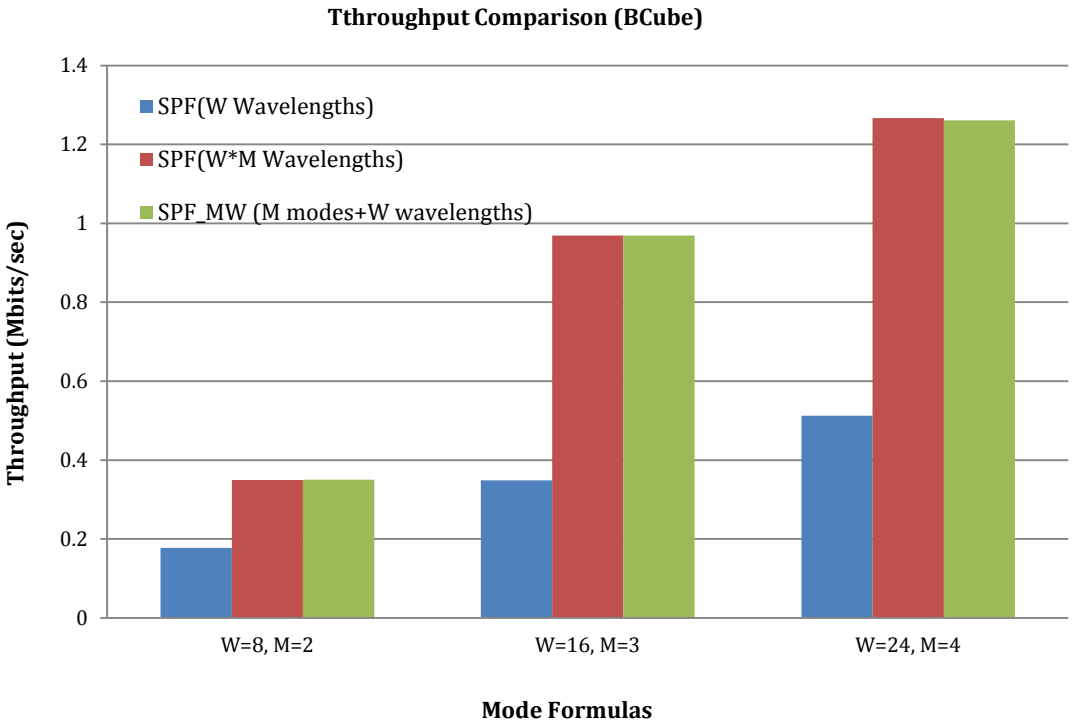

Figure 5-13 Throughput Comparison BCube, Arrival rate = 9.7 flows/s

It can be seen from Figure 5-13 that increasing the number of modes by 1 can produce throughput gain similar to the gain obtained in WDM by increasing the number of wavelengths by W.

Figure 5-14 shows the throughput in the BCube network with increasing input load; the curve for SPF_WDM uses W=20 and M=1 while the curve for SPF_MWDM uses W=20 and M=4. The flow arrival pattern follows the lognormal distribution with ON-OFF behavior using mean from $\mu=1$ to $\mu=2$ and standard deviation $\sigma=1$. Flow sizes used in this test are in the medium range of 200-800 Mb.

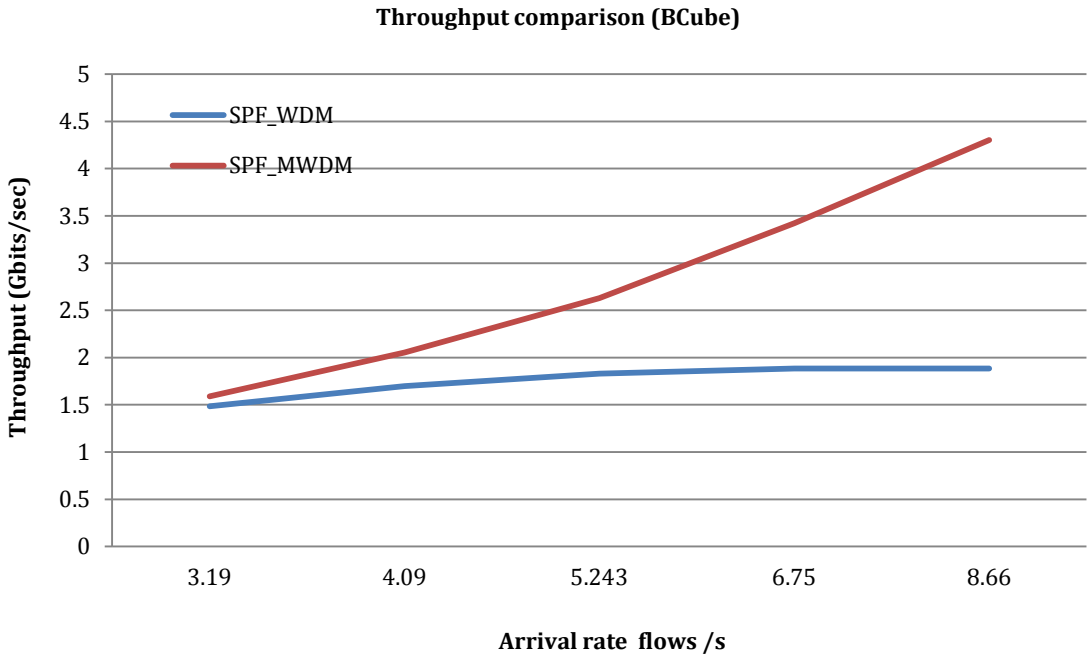

Figure 5-14 Throughput comparison BCube, Max Wavelengths=20, Modes =4

Figure 5-15 shows the capacity increase ratio (CIR) obtained by adding mode-division multiplexing to WDM in the 36-node Fat Tree network. CIR is defined as the ratio between the throughput obtained by SPF_MWDM and the throughput obtained by SPF_WDM. Flow arrivals follow the lognormal distribution with mean μ=3.65, standard deviation σ=1 and ON-OFF nature. Flow sizes used in this test are in the large range of 500 -1250 Mb.

Figure 5-16 shows that the increase in throughput is largely multiplicative in nature when the number of modes is increased.

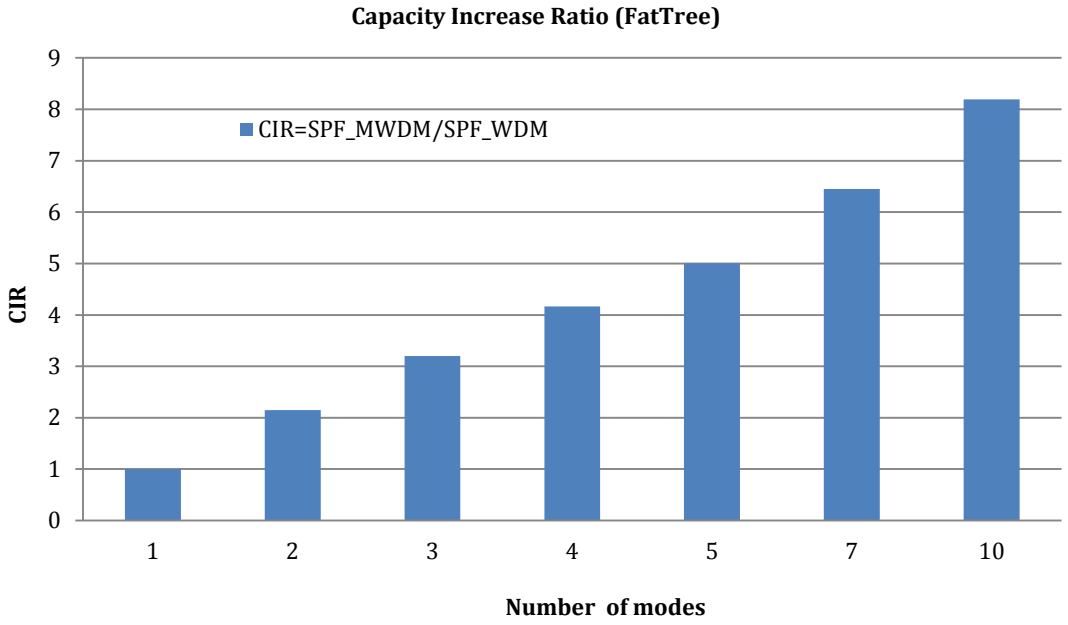

Figure 5-15 Capacity Increase Ratio FatTree, Max Wavelengths=20, Arrival rate=44.5 flows/s

Figure 5-16 shows the corresponding CIR test using the 24 node Mordia ring. Flow arrivals follow the lognormal distribution with mean μ=2.07, standard deviation σ =1 and ON-OFF nature. Flow sizes used in this test are in the small range of 25-250 Mb. Figure 5-16 shows that the increase in throughput is largely multiplicative in nature when the number of modes is increased.

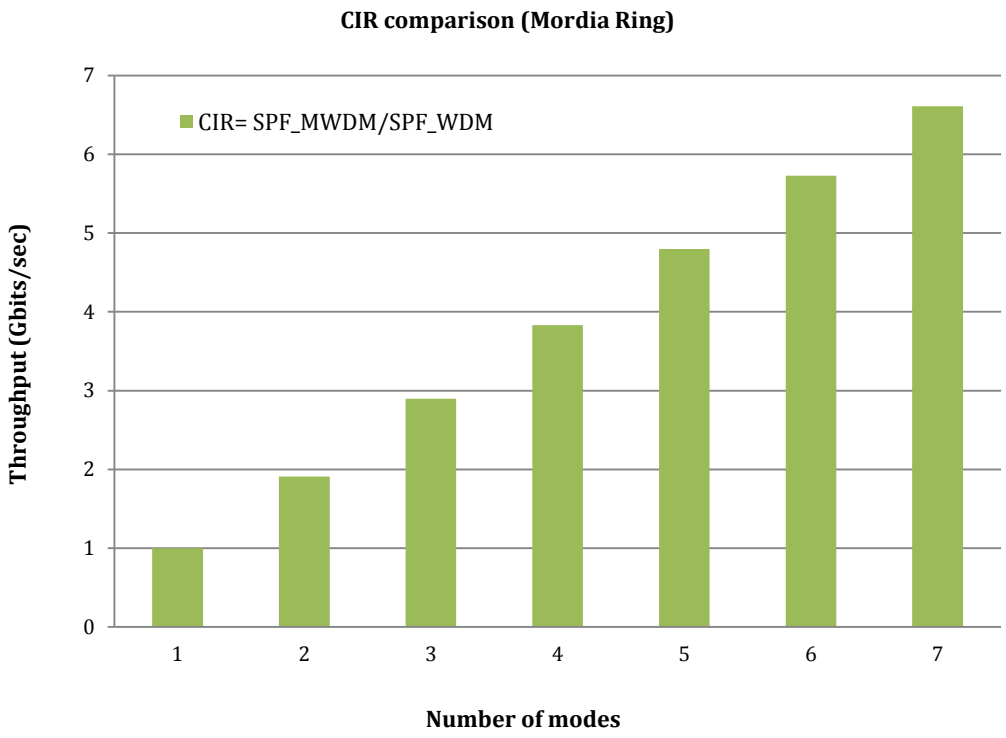

Figure 5-16 CIR comparison Mordia Ring, Max Wavelengths=24, Arrival rate=9.37 flows/s

Figure 5-17 shows the throughput in the 24 node Mordia Ring with increasing input load; the curve for SPF_WDM uses W=24 and M=1 while the curve for SPF_MWDM uses W=24 and M=3. The flow arrival process is lognormal with ON-OFF pattern using mean from μ= 0.5 to μ=1.75 and standard deviation σ=1. Flow sizes used in this test are in the small range of 25-250 Mb.

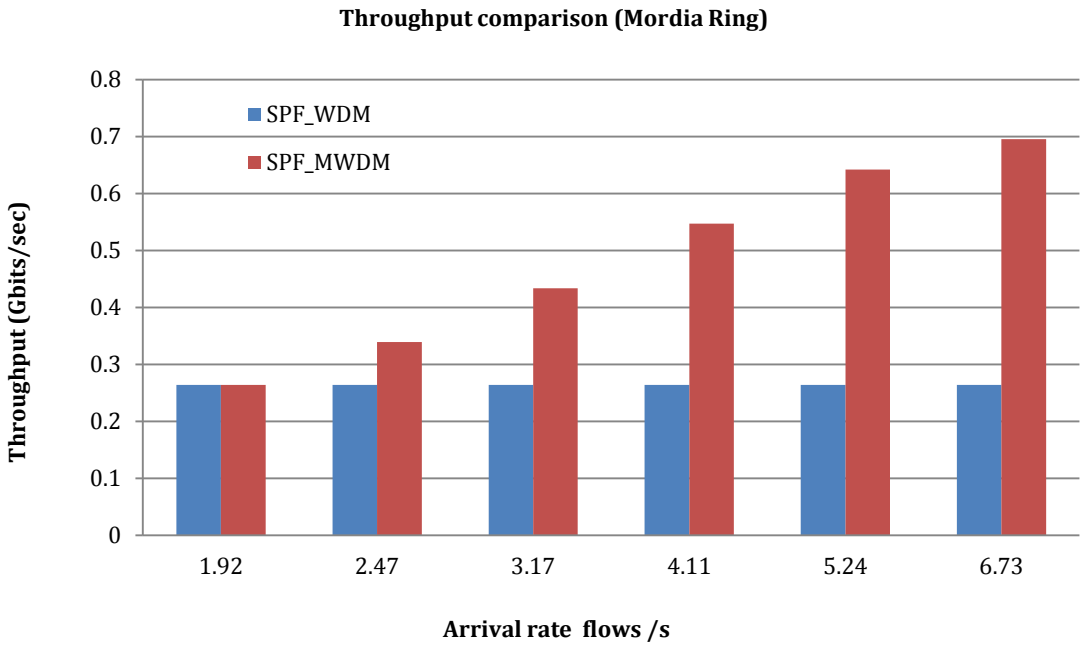

Figure 5-17 Throughput comparison Mordia Ring, Max Wavelengths=24, Modes =3

5.4 Routing mode wavelength assignment (RMWA) heuristics

In this section we evaluate different mode and wavelength assignment heuristics. The shortest path first (SPF) algorithm is used for routing. Four different heuristics are employed for mode and wavelength assignment, namely, most used (MU), first fit (FF), least used (LU) and random assignment. In order to properly evaluate the mode-wavelength assignment heuristics, we have configured the network to operate under the wavelength and mode continuity constraint, i.e., the cross-connects in these tests do not possess mode or wavelength conversion capability.

The most used heuristic, MU, selects the free wavelength (mode) used on most links and assigns that wavelength (mode) to the new lightpath request. The first fit heuristic, FF, selects the free lowest-index wavelength (mode) and assigns it to the new request. The random heuristic,

Random, selects the new wavelength (mode) randomly while the least used heuristic, LU, is just the opposite of MU, i.e., it selects the free wavelength (mode) least used so far on network links

Figure 5-18 compares the throughput obtained by the different mode-wavelength assignment heuristics for the Fat Tree network using the SPF_MWDM routing algorithm. The tests in this figure used W=24, lognormal arrival with ON-OFF traffic pattern using mean μ=2.8 and standard deviation σ=1. Figure 5-19 is similar to Figure 5-18 but uses the Poisson distribution for the arrival process with λ = 35/s and W=20. Both figures used number of modes ranging from M=1 to M=10. Figure 5-18 and Figure 5-19 show that Random and LU are the best heuristics while MU is the worst heuristic.

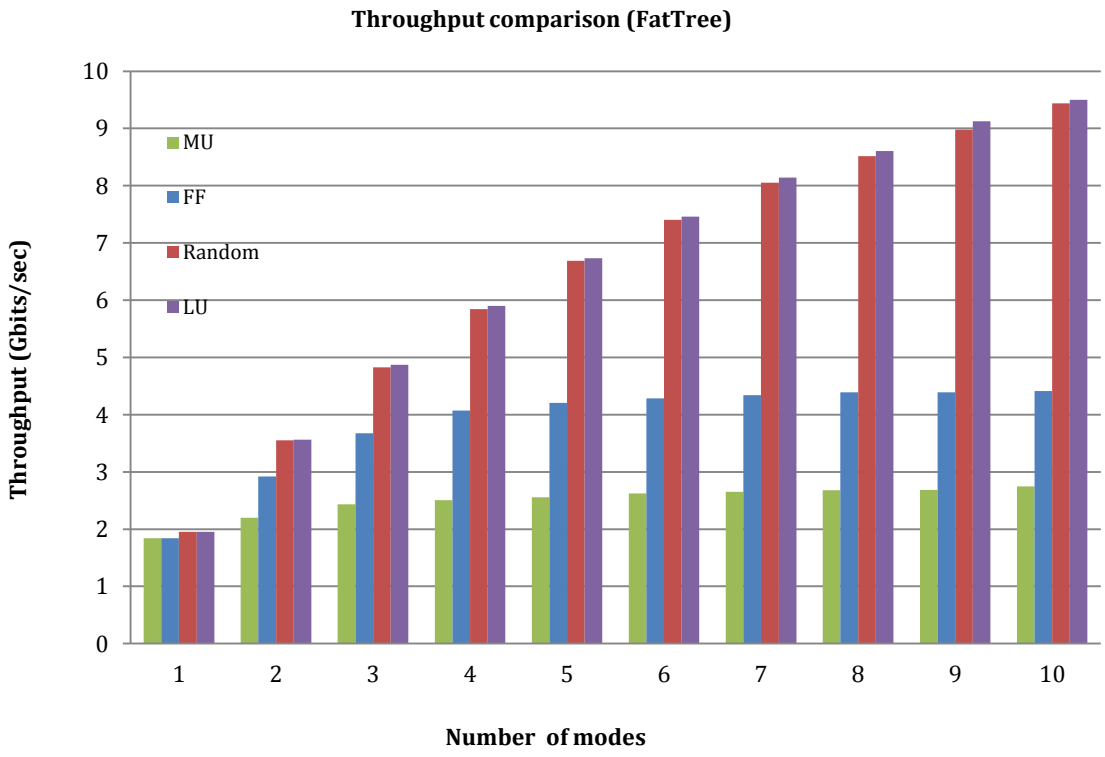

Figure 5-18 Throughput comparison FatTree, Max Wavelengths=24, Lognormal Arrival rate= 19.37 flows/s

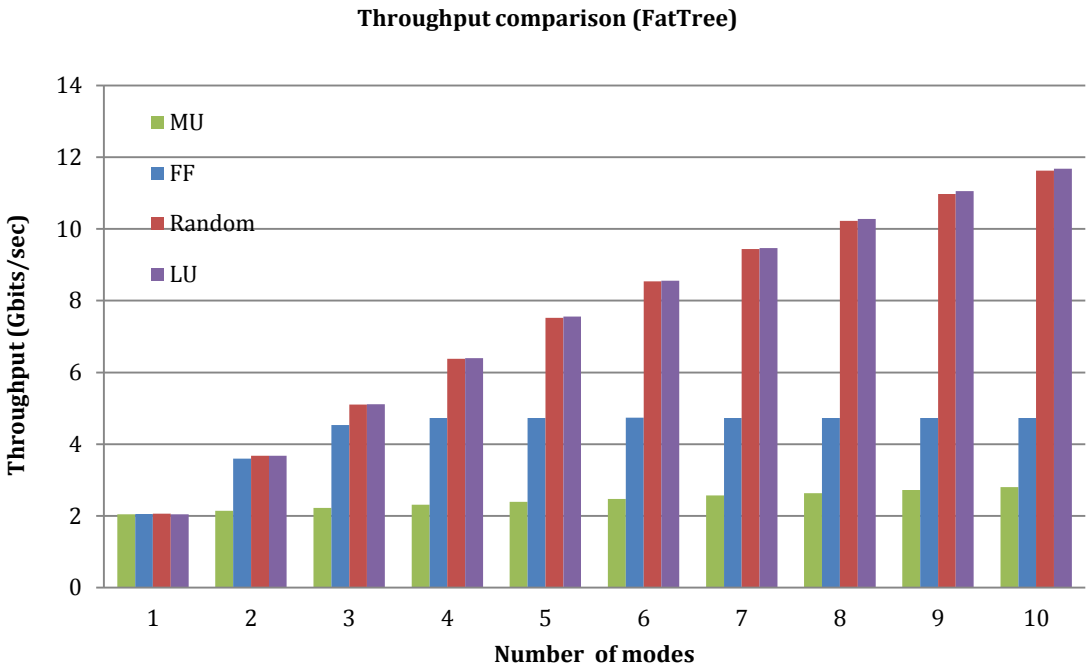

Figure 5-19 Throughput comparison FatTree, Wavelengths=20, Poisson Arrival Rate=35 flows/s

Figure 5-20 compares the throughput obtained by the different mode-wavelength assignment heuristics for the Fat Tree network using SPF_MWDM routing with 4 modes, 18 wavelengths and different input loads. The flow arrival process is lognormal with ON-OFF pattern using mean from μ=1.5 to μ=3.5 and standard deviation σ=1. In Figure 5-20, the performance of MU is still the worst. The performance of FF improves with increasing network load while Random and LU are consistently the best heuristics.

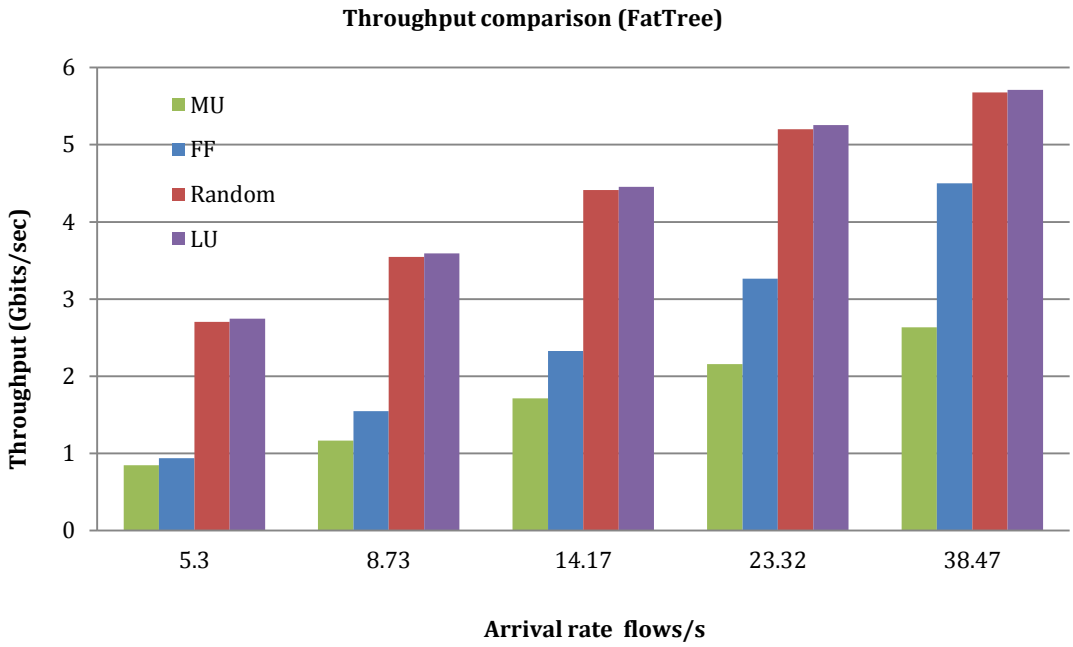

Figure 5-20 Throughput comparison FatTree, Max Wavelengths=18, Modes=4

In the tests presented in, Figure 5-18,Figure 5-19 and Figure 5-20 above, the same heuristic is used for both mode assignment and wavelength assignment. In Figure 5-21, different heuristics are used for mode assignment and wavelength assignment. All Figures in this section used flow sizes in the large range of 500 -1250 Mb.

Figure 5-211 shows the throughput using different heuristics for mode assignment and wavelength assignment. This test used 4 modes and lognormal arrival process with ON-OFF pattern. The value of mean μ=2.5 and standard deviation σ=1 for the lognormal distribution. We tested the following configurations for mode-wavelength assignment: LU-FF, FF-LU, Random-FF, FF-Random, LU-Random, Random-LU, FF-MU, MU-FF, LU-MU and MU-LU.

It is clear from Figure 5-21 that when a better performing heuristic is applied to wavelength conversion the overall throughput gets better because of the larger number of

wavelengths compared to the number of available modes. Hence we can conclude that the wavelength assignment heuristic is more significant than the mode assignment heuristic in deciding the overall throughput.

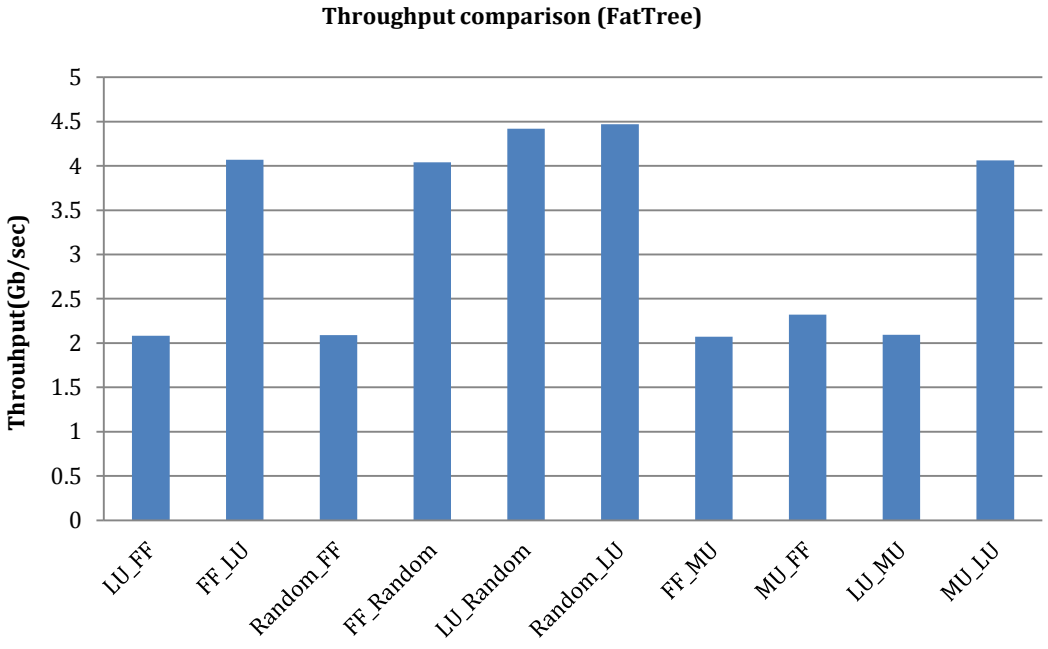

Figure 5-21 Throughput comparison FatTree, Max Wavelengths=20, Modes=4, Arrival rate =14.27 flows/s

The topic of mode-wavelength assignment heuristics is obviously diverse and complex. We have presented preliminary results on some mode-wavelength assignment heuristics and hope that our work will encourage further research on this subject.

5.5 Evaluation of mode cascaded conversion constraint

It is well understood that wavelength conversion degrades the quality of the signal and reduces the signal to noise ratio. Cascading wavelength conversion further aggravates this problem and it is important to realize that a signal can undergo only a certain number of

wavelength conversions to maintain its quality [95 , 96]. Likewise, we cannot perform mode conversions for an individual connection unlimitedly as each mode conversion would deteriorate the quality of the data being transmitted, resulting in the data being rendered useless at the end nodes. Although the rich literature on the performance evaluation of WDM optical routing has largely ignored the problem of cascaded wavelength conversions, there have been a number of studies that investigated this problem. For example, we have previously evaluated the level of deterioration of the blocking performance of all-optical routing due to a constraint on the maximum number of allowed wavelength conversions within the lightpath of circuit-switched optical connections and we developed conversion cascading constraint-aware adaptive routing for WDM optical networks [97-99]. While the issue of wavelength cascaded conversion has been studied in the literature; there has not been any work to evaluate the impact of mode conversions on the blocking performance of optical routing algorithms. In this section, we explore this issue and present preliminary results on the throughput deterioration of optical networks when a mode conversion constraint is applied on the number of allowed mode conversions within the lightpath of an optical flow. For simplicity and to present preliminary results, we will only consider mode conversion constraints without any wavelength conversion constraints.

Figure 5-22 shows the throughput performance of the SPF routing algorithm in the FatTree under different mode cascaded conversion constraint values. The curve labeled SPF_MWDM_NC gives the SPF performance when the negative impact of mode conversion is ignored (NC= no constraint). The curve labeled SPF_MWDM_MCCk gives the SPF performance when the negative impact of mode conversion is such that at most k mode conversions are allowed in the lightpath of a flow, where k=1,2 or 3 . If more than k mode

conversions are needed for a flow routed by SPF_MWDM_MCCk, the flow is dropped to avoid allocating any further resources to a severely degraded optical signal.

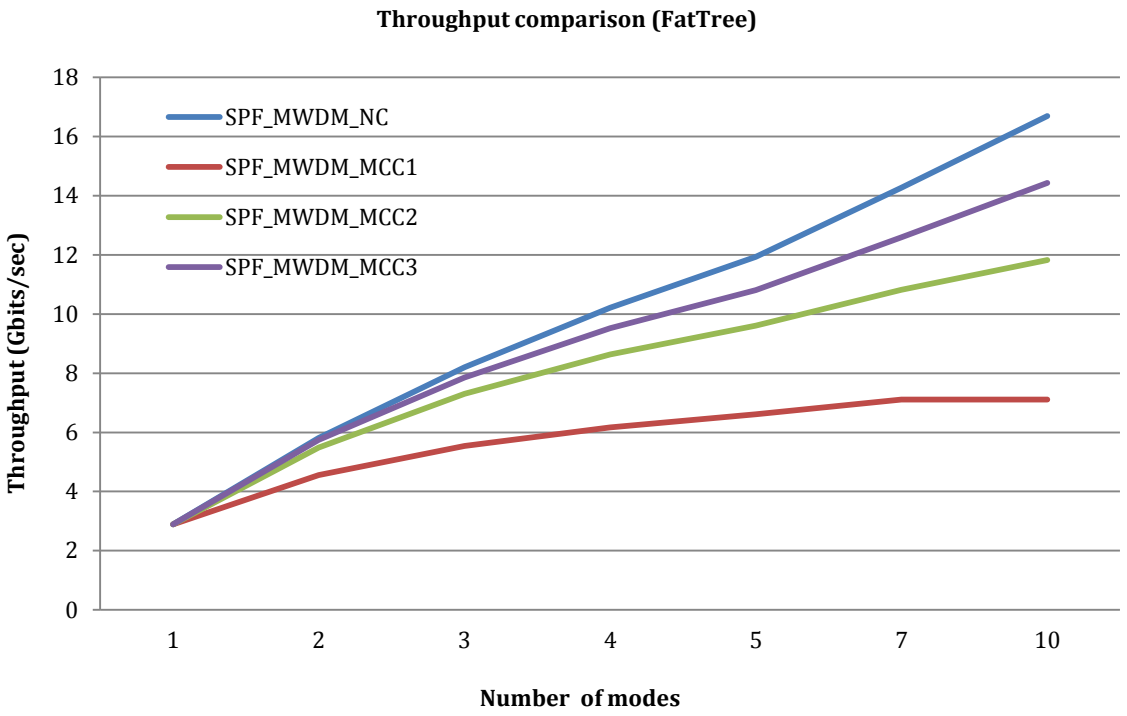

Figure 5-22: Throughput comparison FatTree, Max Wavelengths=20, Modes=4, Arrival rate= 23.66 flows/s.

The figure shows the throughput comparison of SPF_MWDM with no constraint and one, two or three mode conversion constraints on the Fat Tree topology. The flow arrival follows the lognormal distribution with ON-OFF pattern, with mean $\mu=3$ and standard deviation $\sigma=1$. Flow sizes used in this test are in the large range of 500-1250Mb. It is clear from the figure that cascaded mode conversions can practically lower the throughput of the routing algorithm. As mentioned earlier, our investigation uses a simplified environment and is preliminary. Future research should investigate the combined impact of wavelength cascaded conversions and mode cascaded conversions, and should provide more rigorous evaluation of the maximum number of

hops (number of conversions) in the connection lightpath and the maximum number of nodes that can be analyzed.

5.6 Summary

In this chapter, we investigated mode division multiplexing as an additional dimension to enhance network bandwidth in the OCS subset of hybrid electronic-optical datacenter networks. We demonstrated the feasibility of implementing mode-wavelength switching in the hardware domain and presented a possible mode-wavelength switching architecture. Our test results highlighted the benefits of mode wavelength division multiplexing. The chapter evaluated four heuristic algorithms for the mode and wavelength assignment problem in MWDM networks. We have so far presented encouraging preliminary results on mode division multiplexing and mode-wavelength assignment heuristics and we hope that our work will encourage further research on mode wavelength division multiplexing. We have also shown the effect of cascaded mode and wavelength conversion on burst loss probability and throughput of the system.

6. CHAPTER SIX: QUALITY OF SERVICE (QOS) USING MPTCP OVER OPTICAL BURST SWITCHING IN DATA CENTERS

6.1 Introduction

The rapid advancement in cloud computing is leading to a promising future for shared data centers hosting diverse applications. These applications constitute a complex mix of workloads from multiple organizations. Some workloads require small predictable latency while others require large sustained throughput. Such shared data-centers are expected to provide potential service differentiation to client's individual flows. Multipath-TCP (MPTCP) protocol provides improved bandwidth utilization over an OBS network in dense interconnect datacenter networks. In this chapter we will present a simple and efficient service differentiation algorithm called 'QoS aware MPTCP over OBS' (QAMO) in datacenters. Our experimental results show that QAMO algorithm achieves tangible service differentiation without impacting the throughput of the system.

Rest of the chapter is organized as follows. In section 0, we discuss the motivations for the proposed idea. In section 6.3 we discuss briefly our network model. Section 6.4 presents service differentiation scheme for datacenter networks called 'QoS-aware MPTCP over OBS', QAMO. Simulation details and performance analysis is discussed in section 6.5. Finally we summarize the chapter in section 6.6.

6.2 Motivation for the proposed work

Future data center consumers will require quality of service QoS as a fundamental feature. There have been some recent research studies on traffic modeling, network resource management and QoS provisioning in data centers [39, 43, 62]. Ranjan, et. al., studied the

problem of QoS guarantees in data-center environments in [62]. However, this work is not suitable for highly loaded shared data-centers with computationally intensive applications due to the two sided nature of communication. Song Ying et al. in [63] proposed a resource scheduling scheme which automatically provides on-demand capacities to the hosted services, preferentially ensuring performance of some critical services while degrading others when resource competition arises. However research studies on QoS provisioning in data centers did not employ optical networks nor did they use multi-access transport protocols such as MPTCP. MPTCP over OBS provide significant improvement in throughput and reliability and fairness in datacenters. It also makes the data center network more fault tolerant by providing alternative routes in situations of link/node failures. In order to utilize the available bandwidth and network resources more efficiently in previously proposed architecture of MPTCP over OBS we develop a QoS provisioning scheme for data center networks. We evaluate QoS scheme's performance under a realistic datacenter traffic model that will be discussed in detail in this chapter.

The type of applications hosted by datacenters are diverse in nature ranging from back-end services such as search indexing, data replication, MapReduce jobs to front end services triggered by clients such as web search, online gaming and live video streaming [39]. The background traffic contains longer flows and is throughput sensitive while the interactive front end traffic is composed of shorter messages and is delay sensitive. The traffic belonging to the same class can also have differences in relative priority levels and performance objectives [61].

6.3 Proposed Network model

With the popularity of new data center topologies such as Fat Tree and VL2 and the multitude of available network paths, it becomes natural to switch to multi path transport

protocol such as MPTCP to seek performance gains. MPTCP provides significant improvement in bandwidth, throughput and fairness. We have used MPTCP over OBS in our proposed network architecture. In an OBS network, the control information is sent over a reserved optical channel, called the control channel, ahead of the data burst in order to reserve the wavelengths across all OXCs. The control information is electronically processed at each optical router while the payload is transmitted all-optically with full transparency through the lightpath. The wavelength reservation protocol plays a crucial role in the burst transmission and we have used just-in-time (JIT) [19] for its simplicity. The necessary hardware level modifications of optical switches for supporting OBS in data centers have been discussed in [41], and will not be repeated in this contribution.

6.4 QoS aware MPTCP over OBS algorithm

Our proposed algorithm QoS aware MPTCP over OBS called QAMO combines the multiple paths of MPTCP and resource reservation in OBS to develop an adaptive and efficient QoS-aware mechanism. Data centers handle a diverse range of traffic generated from different applications. The traffic generated from real time applications e.g., web search, retail advertising, and recommendation systems consists of shorter flows and requires faster response. These shorter flows (foreground traffic) are coupled with bandwidth intensive longer flow (background traffic) carrying out bulk transfers. The bottleneck created by heavy background traffic impacts the performance of latency sensitive foreground traffic. It is extremely important to provide a preferential treatment to time sensitive shorter flows to achieve an expected performance for data center applications. QoS technologies should be able to prioritize traffic belonging to more critical applications. Our proposed algorithm provides priority to latency-sensitive flows at two

levels, i) MPTCP path selection stage and ii) OBS wavelength reservation stage. We propose that larger bandwidth be dynamically allocated to high priority flows, in order to minimize latency and reduce their drop probability. QAMO algorithm just does that.

Let W be the maximum number of wavelengths per fiber, and K be the number of paths that exist between a given souce-destination pair. We will introduce a new term, the *priority factor P* for a burst priority defined as the ratio of P_{curr} (priority level of the current burst) to P_{max} (maximum priority levels) i.e., $P = P_{curr}/P_{max}$. Priorities of individual bursts are represented in ascending order as P_1, P_2, $P_{3...}$ P_{max} while P_{max} is the highest priority level in the bursts. We next define the number of allocated paths k_{curr} for the burst of a particular priority level as follows.

$$k_{curr} = \lceil K \times P \rceil \tag{6.1}$$

At path allocation stage a larger number of paths is allocated for a high priority burst thus reducing its latency. For example, if $P_{curr}=P_{max}$, then P = 1. This will result in k_{curr} = K paths whereas if P_{curr} = 0.5*P_{max}, then P=0.5 and the number of allocated paths is reduced to half the set of *K paths*. This will give the low priority burst, half the number of paths. We now define the size of the wavelength search space controlled by the following equation.

$$\textit{Wavelength search size} = \lceil W \times P \rceil \tag{6.2}$$

At wavelength reservation stage in OBS, equation 2 allocates a larger subset of wavelength search space for a burst with higher priority level thereby allowing it a greater chance to get through and reduce its blocking probability.

QAMO (QoS Aware MPTCP over OBS) Algorithm

Input:

$P = P_{cur}/P_{max}$
K = maximum number of paths
W = maximum number of wavelengths
w_{cur} = current wavelength reserved for current burst
N_k = vector of all nodes on path k
k_{cur} = paths allocated to the current burst
$burst_{cur}$ = current burst

Algorithm:

```
for each k in ⌈K × P⌉:
  concurrency::parallel_invoke: lightpath(k)

function lightpath(path k)
  Initialize w_cur
  for each n in N_k:
    if n = N_k [length(N_k) – 1]          // destination node
      break;
    if n = N_k                            // source node
      for each w in ⌈W × P⌉:
        if w is free
          reserve w for burst_cur at n
          w_cur = w
          break;
    else
      if w_cur is free at n
        reserve w_cur for burst_cur at n
        continue;
      for each w in ⌈W × P⌉:
        if w is free
          reserve w for burst_cur at node n
          w_cur = w
          break;

    if no free wavelength at n
      return (error);            // search failed at node n
  return(success);
```

Figure 6-1 QoS-aware MPTCP over OBS, QAMO Algorithm

In the above algorithm, the priority factor P is used to adjust the number of allocated paths for concurrent transmission and the size of the wavelength search space based on the priority level of the burst. For high priority bursts, more concurrent MPTCP paths result in larger bandwidth, and more OBS network wavelengths reduce dropping probability. The parameter P_{max} can be flexible to accommodate changes in network statistics over time as bursts of different priority levels are encountered.

We assume that QAMO algorithm has access to available information about QoS requirements of different bursts to process them correctly. AT MPTCP layer this capability may be implemented using a specific interface such as the Implicit Packet Meta Header (IPMH) promoted in [100]. It is possible to assign priority levels for different flows in MPTCP at IPMH interface [101, 102]. Because of IPMH interface, it is also possible to gather priority information for each type of flow at a particular end host. This information can be passed on to the OBS network during burst segmentation process from MPTCP layer. At OBS network, the current burst priority P_{curr}, *or the ratio* $P = P_{curr}/P_{max}$, can be easily passed from one OXC to the next via the control packet and does not demand any significant resources in the OXC's. Implementing the reduced (adjustable) search as in the case of QAMO, to find a free wavelength requires minor modification to the standard JIT channel allocation scheme. The adjustable search in a smaller space of $\lceil W \times P \rceil$ for wavelengths actually leads to a smaller average search time.

The QAMO scheme has been extensively tested on the simulation testbed using data center network topologies FatTree and BCube to provide tangible QoS differentiation without negatively impacting the overall throughput of the system.

6.5 Performance Evaluation

In this section we will discuss simulation detail, network topologies used in our tests, traffic model and present performance results with their analysis.

6.5.1 Simulation Detail

The simulation testbed has been developed using C++. A source-destination pair amongst host nodes is randomly chosen for each originated burst. For TCP, to establish the static lightpath, simulation calculates the shortest path between these nodes using Dijkstra's algorithm. In case of MPTCP, it uses K shortest paths algorithm (derived from Dijkstra's algorithm) to find K paths between the source-destination pair. The wavelength assignment heuristic is first-fit as done in [1, 2]. Recent research studies on traffic characteristics of data centers have shown that the traffic in data centers follows the lognormal distribution with ON-OFF pattern [43, 46]. The lognormal distribution is also considered to be the most fitted distribution for modeling various categories of internet traffic including TCP [103]. We have used lognormal arrival with an ON-OFF behavior in our simulation. The network nodes are assumed to be equipped with wavelength converters. We assume that MPTCP is running at end hosts. Based on the priority of the burst, K control packets originate from the source node to establish K lightpaths. Each control packet acquires an initial free wavelength at the source node, then travels to the destination node and reserves wavelengths following QAMO algorithm. If at any node, the same wavelength as the one reserved on the previous node is not available then it tries wavelength conversion. The process continues until the control packet either reaches the destination node or gets blocked due to the unavailability of free wavelength at any hop along the path. Thus, number of lightpaths established = K – number of control packets blocked. The source node

waits for a predetermined time depending on the hop distance to the destination called offset time before transmitting the optical burst message. The traffic used in our simulation is uniformly distributed, i.e., any host node can be a source or a destination [1, 23].

The simulation clock is divided into time units (tu), where each simulation time unit corresponds to 1 microsecond. Each node has a control packet processing time of 20 microseconds and a cut through time of 1 microsecond as proposed for OBS networks in data centers [40]. Each node can have a certain maximum number W of allowed wavelengths. Arrival rate/tu denotes the average arrival rate of the lognormal ON-OFF traffic.

In data center environment a complex mix of short and long flows is generated. The shorter flows are usually latency-critical and represent the largest proportion of flows in data centers [43]. The medium sized and longer flows constitute background traffic and may belong to different priority levels [94]. To represent these scenarios of data center mixed traffic, we have used variable burst sizes in different ranges with uniform distribution within each range [94].

Short burst sizes: S_{min}=5 Kbits to S_{max}= 20 KB

Medium burst sizes: S_{min}=200 Kbits to S_{max}= 1 MB

Long burst sizes: S_{min}=20 Mbits to S_{max}= 100 Mbits.

Our traffic model is based on the findings on data center traffic characteristics in [39, 43, 46, 94]. To model our traffic we assume that 70-80% of bursts generated are short burst belonging to latency sensitive applications, 10-15% have medium burst sizes while 5-10% of bursts belongs to large burst size range. In order to assign the priorities, 95% of short burst messages have the randomly assigned priorities from the highest priority range [P5-P6]; the remaining 5% can have any priority level. Similarly, 95% of medium and large burst sizes are

randomly assigned priorities from sets [P3 – P4] and [P1 – P2] respectively. The remaining 5% from these ranges are assigned random priorities from set [P1 –P6].

6.5.2 Performance results and discussion

The topologies used in our simulation tests are FatTree with 36 nodes and BCube with 24 nodes as shown earlier in Figure 5-3 and Figure 5-4 respectively. All the figures in this section are tested following lognormal distribution. Because of the ON-OFF pattern of traffic the average arrival rate is smaller than the arrival rate of a continuous lognormal process having the same mean and standard deviation. The tests are conducted over burst distribution of our proposed traffic model discussed in section 6.5.

Figure 6-2 motivates the use of MPTCP in data center networks for improving throughput. Figure 6-2 is tested using the lognormal distribution with mean μ=1.8 and standard deviation σ=1, corresponding to an arrival rate of 7.12/tu in BCube topology. Figure 6-2 shows the throughput comparison between TCP (K = 1) and MPTCP (K = 2, 3, 4), where K is the number of paths (i.e., number of subflows) used by each MPTCP connection. It can be observed that, MPTCP gives much higher throughput as compared to single path TCP. It can also be observed that MPTCP performs better with increasing number of paths. Similar results were achieved for FatTree topology.

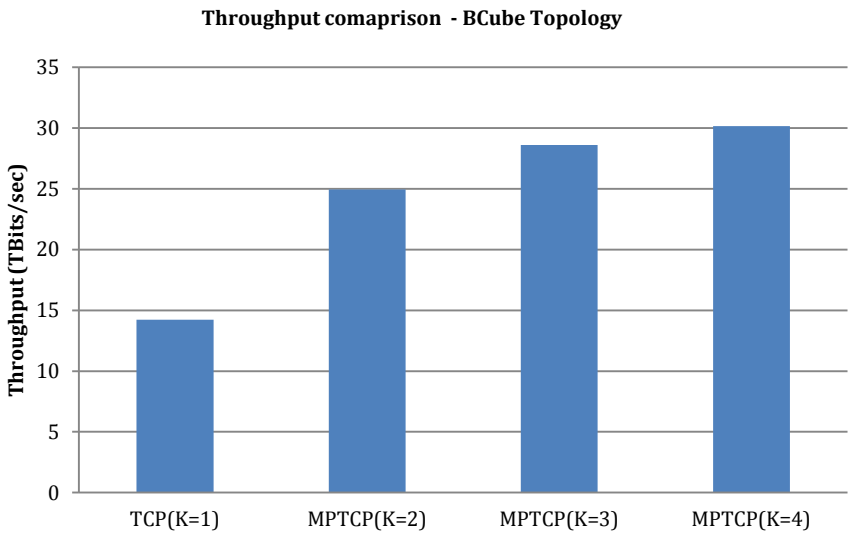

Figure 6-2 Throughput comparison, BCube, Arrival Rate /tu = 7.12, W=64

Figure 6-3 Dropping Probability – FatTree, Variable arrival rate, W=64shows the ability of QAMO algorithm to achieve QoS differentiation when tested for bursts of various sizes and priority levels as proposed in our traffic model. The dropping probability comparison for six priority levels is shown with increasing load in a FatTree topology. For lognormal traffic, the mean values used in this test are from $\mu=1$ to $\mu=3$ and standard deviation $\sigma=1$. It can be observed that the algorithm achieves substantial QoS differentiation for all priority levels. For example, P6 being the highest priority level, experiences the least dropping at all values of input load. Similar results were achieved for BCube topology.

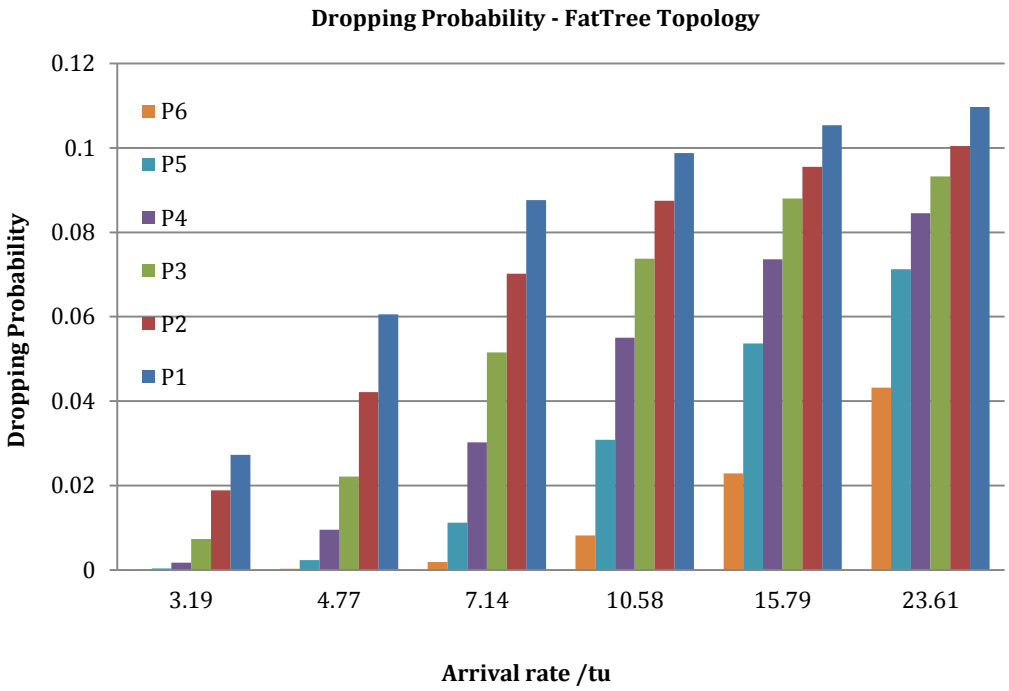

Figure 6-3 Dropping Probability – FatTree, Variable arrival rate, W=64

Figure 6-4 shows the average throughput comparison of TCP, MPTCP (K=4) and QAMO. The lognormal mean values used in this test are from μ=0.5 to μ=1.75 and standard deviation σ=1. It can be observed that QAMO and MPTCP (K=4) both performs much better than standard TCP. The throughput of QAMO is slightly less than MPTCP (K=4) at small values of input load while the difference in throughput becomes less at higher loads. The reason for QAMO's degraded throughput is its preferential treatment for higher priority bursts, which are mostly very small in size.

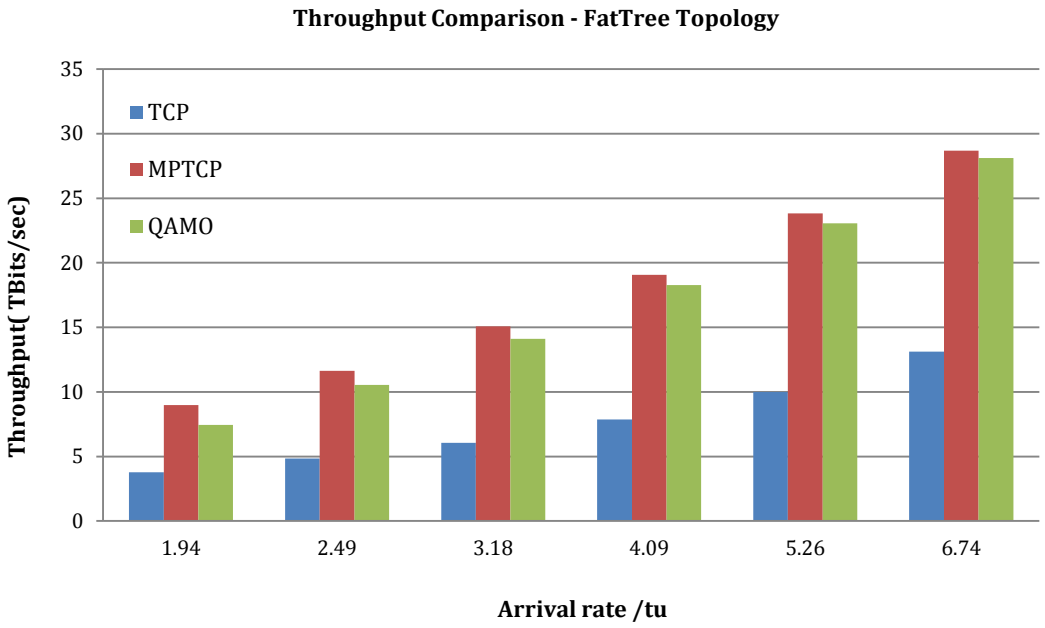

Figure 6-4 Throughput Comparison, FatTree, Variable arrival rate, W=64

Figure 6-5 provides deeper analysis of throughput breakdown in terms of burst priorities at one of the loads from Figure 5, specifically at arrival rate = 2.49 bursts/tu. The lognormal mean in Figure 6 is μ=0.75 and standard deviation σ=1. It can be observed that in TCP and MPTCP the greatest share of throughput is achieved by low priority background traffic, giving less importance to the time sensitive foreground flows in the absence of QoS provisioning. The throughput of QAMO is well distributed between high priority (foreground) and low priority (background) traffic. Hence, the slight degradation of QAMO throughput compared to MPTCP is acceptable for achieving better share of network resources for more critical traffic in data centers.

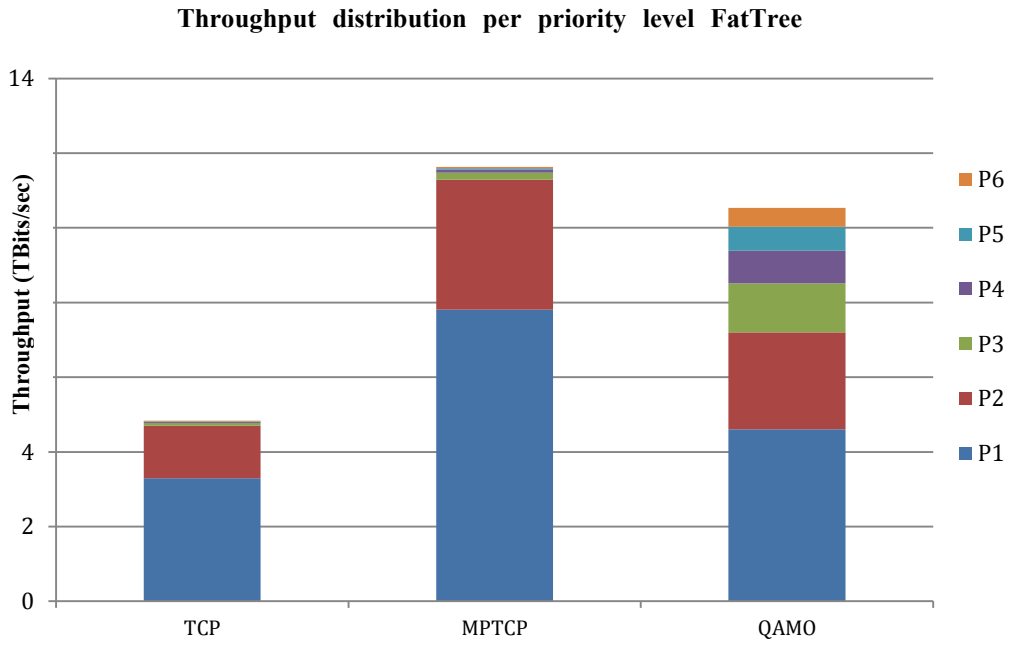

Figure 6-5 Throughput distribution per priority level – FatTree, Arrival Rate /tu = 2.49, W=64

6.6 Summary

In this chapter we have presented and evaluated QoS-aware MPTCP over OBS (QAMO) scheme to provide service differentiation in data center traffic. QAMO algorithm provides tangible QoS differentiation to bursts of various classes without impacting the throughput of the system. We presented the results of extensive performance tests to evaluate the effectiveness of the proposed scheme.

7. CHAPTER SEVEN: MULTIPATH-TCP (MPTCP) IN CLOUD BASED OPTICAL DATACENTERS

7.1 Introduction

Data centers have become the heart of the computational world over the past few years. The emergence of cloud computing and the growth of data-intensive applications have driven the need for finding alternative ways to improve communication efficiency in data center networks. In this paper, we combine the advantages of Multipath-TCP with optical networking to maximize bandwidth in datacenters and present an evaluation of MPTCP over optical burst switching (OBS) for data center network. We compare the performance of standard TCP with MPTCP under different network loads and topologies using realistic data center traffic models. Our simulation tests have established that Multipath-TCP over OBS provides significant performance advantage in terms of improving throughput, reliability and fairness for data center networks.

The rest of the chapter is organized as follows. In section 7.2, we provide motivation for the proposed work. In section 7.3 we discuss our proposed work, networking model that uses MPTCP protocol over an optical burst switching network for data centers. In section 7.4 simulation details are discussed and the performance analysis and experimental results are given in section 7.4.2. We give the conclusion and summarize the work in Section 7.5.

7.2 Motivation for the Proposed Idea

The emergence of cloud computing and the growth of data-intensive applications have driven the need for finding alternative ways to improve communication efficiency in data center networks. Many internet applications today are powered by data centers. The traffic generated by these bandwidth intensive applications grew exponentially over the last few years resulting in a

massive increase in the computational, storage and scalability requirements of data centers. Meeting performance goals for data centers networks (DCN) became very challenging under these conditions.

7.3 Proposed Idea

In this chapter, we propose and evaluate the potential benefits of implementing the newly emerging transport protocol, Multipath TCP, over an optical OBS network in data centers. Our proposed data center networking strategy is evaluated over the FatTree and BCube topologies and our tests have established that Multipath TCP over OBS provides huge performance advantage in terms of improving throughput, reliability and robustness of data center networks.

7.3.1 Network model

A. Understanding data center traffic pattern

The datacenters traffic model is very complex in nature. Datacenters handle a diverse range of traffic generated from different applications. The traffic generated from real time applications e.g., web search, retail advertising, and recommendation systems consists of shorter flows. These shorter flows (foreground traffic) are coupled with bandwidth intensive longer flows (background traffic) carrying out bulk transfers. The bottleneck created by heavy background traffic impacts the performance of latency sensitive “foreground traffic hence it becomes extremely important to bring communication efficiency in datacenter networks. Recent research studies on traffic characteristics of datacenters have shown that the arrival pattern of datacenter traffic follows lognormal distribution with an ON-OFF pattern [43, 46].

B. Datacenter topologies and MPTCP

Traditionally data centers have been built using hierarchical topologies. Such topologies are suitable if there a good mix of inter and intra-data center traffic flows. It has been observed, however, that most of the traffic generated is intra-data center [42]. Therefore, traditional data center topologies cannot provide sufficient bandwidth and serious communication bottlenecks exist between hosts-edge and edge-core layers. For this reason, the FatTree, VL2 and BCube topologies with dense interconnects have started to be deployed in modern DCNs [42, 47]. We have used these modern datacenter topologies in our proposed network model as shown in Figure 5-3 and Figure 5-4.

With the popularity of new data center topologies and multitude of available network paths, it becomes natural to seek performance gains through adopting multi path transport protocols such as MPTCP. Multipath TCP can provide numerous benefits over single path TCP[38, 42]. MPTCP provides significant improvement in bandwidth, throughput and fairness in these modern DCN topologies over electronic packet switched networks

C. Optical Burst switching in DCN

This section provides only a brief introduction of optical burst switching. A burst is a collection of packets and is the basic data unit of OBS. There can be different burst size variation in a typical OBS network. Larger bursts are suitable for a large session scenario (as in OCS) and smaller bursts can serve the granular and bursty traffic scenarios (as in OPS). This feature makes OBS a promising solution to achieve the all-optical switching in intra- data center communication, since it can handle both continuous and bursty traffic[48]. In OBS, the control information is sent over a reserved optical channel, called the control channel, ahead of the data burst in order to reserve the wavelengths across all OXCs. The control information is

electronically processed at each optical router while the payload is transmitted all-optically with full transparency through the lightpath. Hence, control packets would have to experience O/E/O conversion for resource reservation at each intermediate optical node. The resource reservation scheme and wavelength assignment heuristics play an important role in OBS network performance. For the purpose of our simulations, we have used just-in-time (JIT) [19, 104] reservation scheme for its simplicity. The wavelength assignment heuristic is first-fit as done in [1, 2]. The necessary hardware level modifications of optical switches for supporting OBS in data centers have been discussed in [48], and will not be repeated in this chapter.

D. *Fairness in MPTCP over OBS*

It is important for a networking protocol to provide fair share of system resources to all the nodes/applications. In order to evaluate our proposed network architecture for fairness, we need to investigate two types of possible unfairness. i) The TCP unfairness caused in multi-rooted tree topology for flows with smaller and larger number of hops competing for a common output port [59] and ii) the beat down unfairness problem naturally present in OBS networks [2]. We will show in our results section that due to availability of multiple paths as in case of MPTCP both types of unfairness are reduced.

7.4 Performance Evaluation

7.4.1 Simulation Detail

MPTCP over OBS has been extensively tested using a simulation testbed written in C++. A source-destination pair amongst host nodes is randomly chosen for each originated burst. The traffic used in our simulation is uniformly distributed, i.e., any host node can be a source or a destination as was done in [1] . For TCP, to establish the static lightpath, simulation calculates

the shortest path between these nodes using Dijkstra's algorithm. In case of MPTCP, it uses K shortest paths algorithm (derived from Dijkstra's algorithm) to find k paths between the source-destination pair. TCP over OBS carries an inherent problem of TCP congestion control mechanism [55-57]. In an OBS network, TCP is implemented at a higher layer and each assembled burst contains packets from several TCP flows. Random burst losses may be mistakenly interpreted by TCP layer to reduce congestion window un-necessarily even when congestion in the network is low. In our simulation, a lost burst is re-transmitted internally in OBS network as proposed in [55] to avoid false propagation of burst loss to TCP layer. Recent research studies on traffic characteristics of datacenters have shown that lognormal distribution with ON-OFF pattern accurately represents traffic behavior in datacenters [43, 46]. The Lognormal distribution is also considered to be the most fitted distribution for modeling various categories of internet traffic including TCP [103]. Hence we have used lognormal arrival with an ON-OFF behavior in our simulation. The network nodes are assumed to be equipped with wavelength converters. The control packet which originates from the source node acquires an initial free wavelength then travels to the destination using the Just-in-Time signaling protocol [19, 104] and reserves wavelengths along K paths. If the same wavelength is not available at any hop in the along the path then it tries wavelength conversion. The process continues until the control packet either reaches the destination node or gets blocked due to the unavailability of free wavelength at any hop along the path. In that case we have light paths = K – number of blocked control packets. The source node waits for a predetermined time depending on the hop distance to the destination called offset time before transmitting the optical burst message.

The simulation clock is divided into time units, where each simulation time unit corresponds to 1 microsecond. Each node has a control packet processing time of 20 microseconds and a cut through time of 1 microsecond, as proposed for OBS networks in datacenters [40]. Cloud applications produce bursts of various size ranges [94]. Amongst these ranges, burst sizes are distributed uniformly (from S_{min} to S_{max}) as follows:

Small burst sizes: S_{min}=100 Kbits to S_{max}= 1 MB

Long burst sizes: S_{min}=10 Mbits to S_{max}= 100 Mbits.

Each node can have a certain maximum number of allowed wavelengths W. Arrival rate (A/tu) denotes the average arrival rate of the lognormal ON-OFF traffic per time unit (μsec). Each of the performance graphs in this chapter was generated by running the simulation for more than 10 million iterations to produce stable results i.e., within 95% confidence interval range.

7.4.2 Results and Discussion

The topologies used in our simulation tests are FatTree with 36 nodes and BCube with 24 nodes as shown in Figure 5-3 Fat Tree and Figure 5-4 BCube. In FatTree and BCube the root level nodes are called high level aggregators (HLAs), the next layer of nodes are medium level aggregators (MLAs).

Figure 7-1 shows the throughput comparison between TCP (K = 1) and MPTCP (K = 2, 3, 4), where K is the number of paths (i.e., number of subflows) used by each connection. Burst sizes used in this test are in the range of 100 Kb- 1Mb. It can be observed that for an arrival rate of 2 connection requests/μs, TCP gives much lower throughput as compared to MPTCP. It can

also be observed that MPTCP performs better with increasing number of paths. Similar results were achieved for BCube.

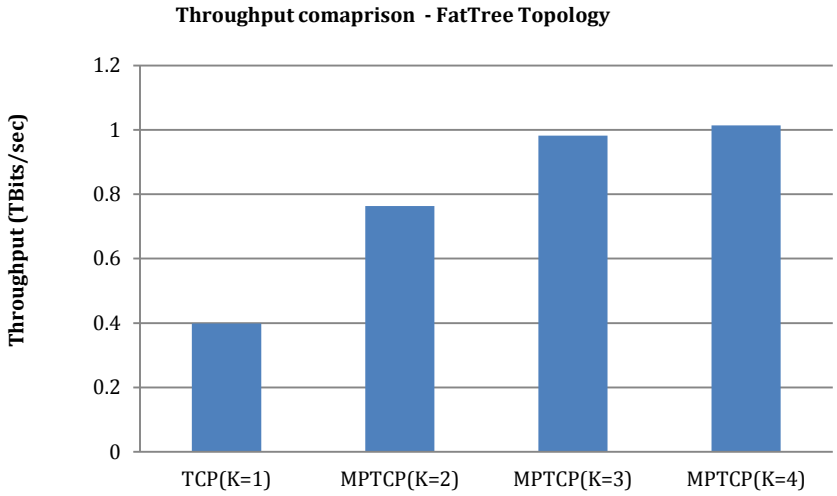

Figure 7-1: Arrival rate = 2 bursts/µs, W=64

Figure 7-2 shows the throughput of MPTCP with K=3 and TCP over increasing network loads. Burst sizes used in this test are in the range of 10-100 Mb. MPTCP (K=3) throughput values are better than regular TCP even after network becomes congested.

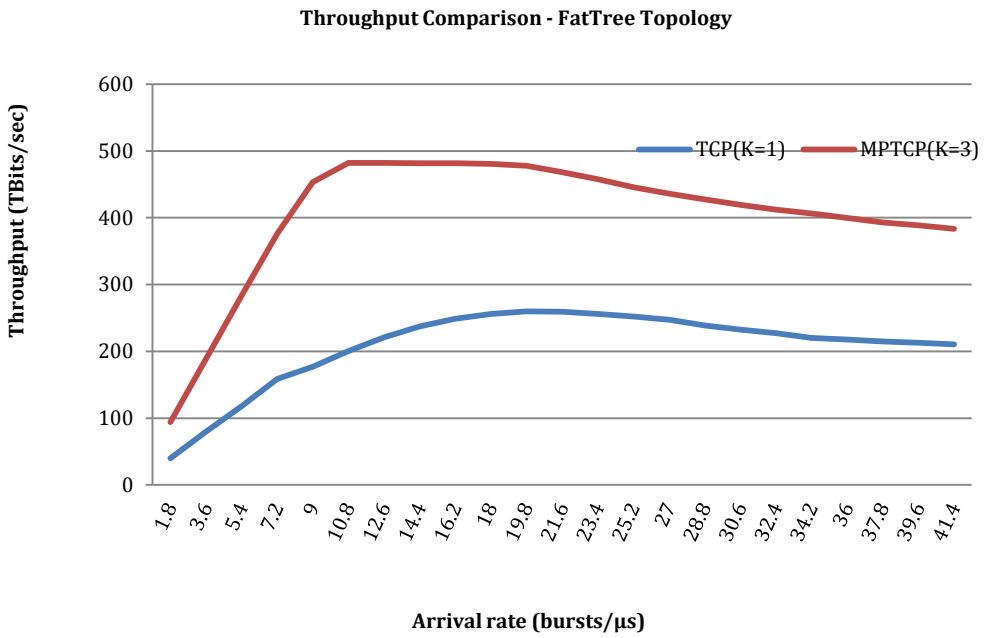

Figure 7-2: Variable arrival rate, W = 80

Figure 7-3 shows a comparison between the two protocols in terms of the time taken to transmit a certain amount of data. Burst sizes used in this test are in the range of 100Kb - 1Mb. The simulation generated same amount of data (2500 GBits) for both protocols and we can see in Figure 6 that MPTCP takes less time to transmit the data under similar network load. MPTCP consisting of multiple TCP subflows uses the alternative paths in the network more efficiently and transmits that data in much shorter time. We have observed that MPTCP with K=2 is using approximately 50% less time, than that of regular TCP. In case of K=3 and K=4, this improvement becomes approximately 60%.

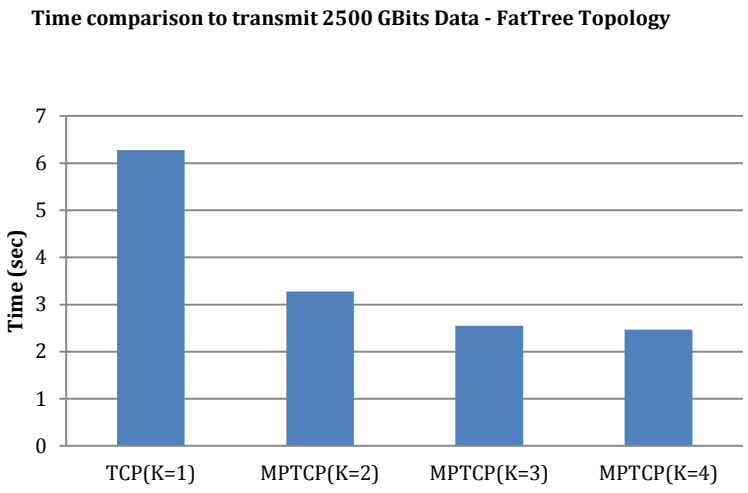

Figure 7-3: Arrival rate = 2 bursts/µs, W=64

Figure 7-4 shows similar comparison between the two protocols for BCube topology. Burst sizes used in this test are in the range of 10- 100Mb. The time improvement of MPTCP with K=2 is approximately 40% less than that of regular TCP. In case of K=3 and K=4, this improvement becomes approximately 45%. BCube topology has much longer alternative paths (no. of hops) hence; percentage improvement in time of MPTCP over TCP is relatively smaller than FatTree topology.

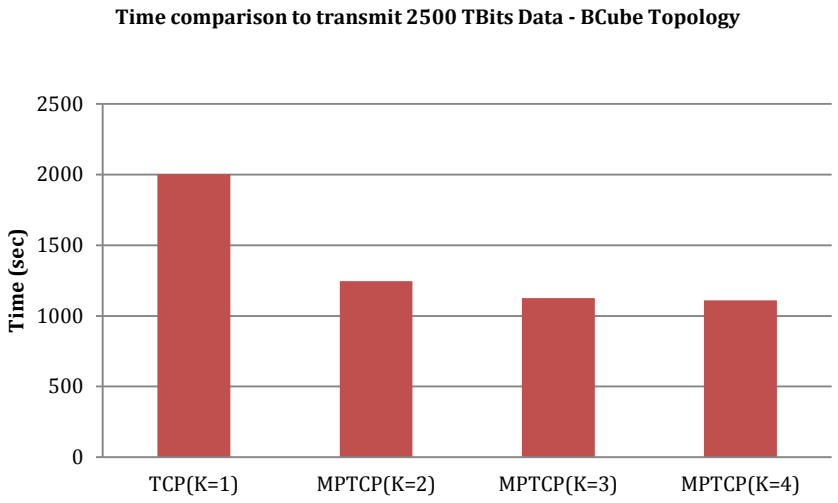

Figure 7-4: Arrival rate = 1.8 bursts/µs, W=64

Figure 7-5 shows the various stages of data transmission over different points of time (seconds). At every time instant MPTCP has transmitted more data than TCP. It is also noticeable that as we increase the number of paths K (i.e., number of subflows per connection), the volume of data transmitted continues to increase. Similar results were achieved for FatTree topology.

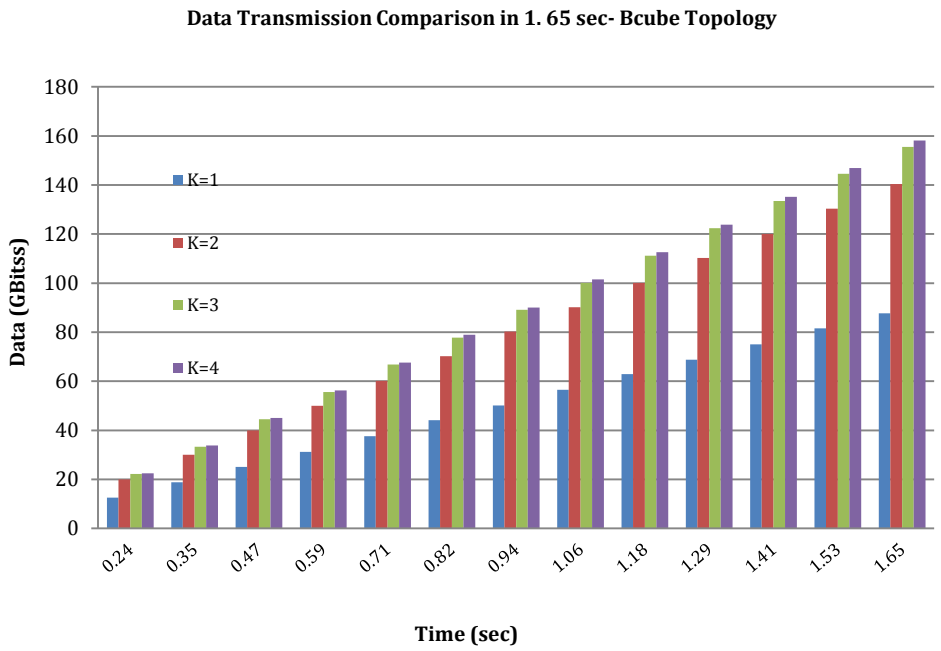

Figure 7-5: Arrival rate =1.8 bursts/µs, W=64

The robustness of MPTCP could be well established through its ability to use alternative routes in the network in case of link failures. In Figure 7-6, we have simulated medium level aggregator (MLA) and high level aggregator (HLA) link failures of the FatTree topology. We used the burst sizes in the range of 100 Kb-1 Mb. Any two random links were failed with HLA or MLA nodes. MPTCP performs better under the situations of link failures than regular TCP and uses possible alternative routes.

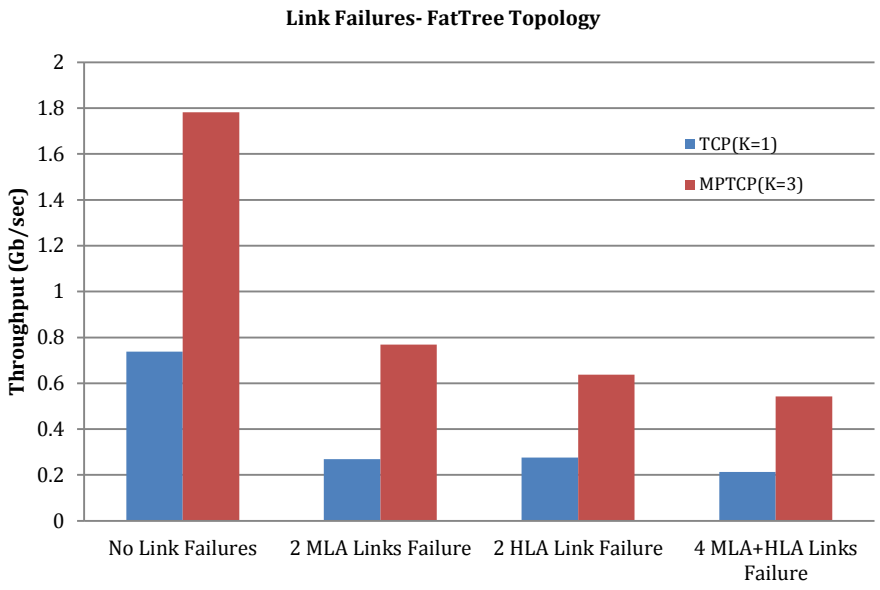

Figure 7-6: Arrival rate = 3.3 bursts/µs, W=64

In order to evaluate fairness of TCP vs. MPTCP over OBS, independent throughput of nodes is shown in Figure 7-7. Burst sizes used in this test are in the range of 100 Kb-1 Mb. The throughput was recorded at randomly selected five nodes when those nodes were acting as source and trying to send data to a common destination node. The source nodes had different hop counts to destination nodes (hop-count unfairness problem of OBS) and near and far flows were competing for common output ports to reach the common destination node (TCP unfairness problem in datacenters). Per node throughput varies in case of TCP under congestion while it stays uniform in case of MPTCP due to usage of multiple paths.

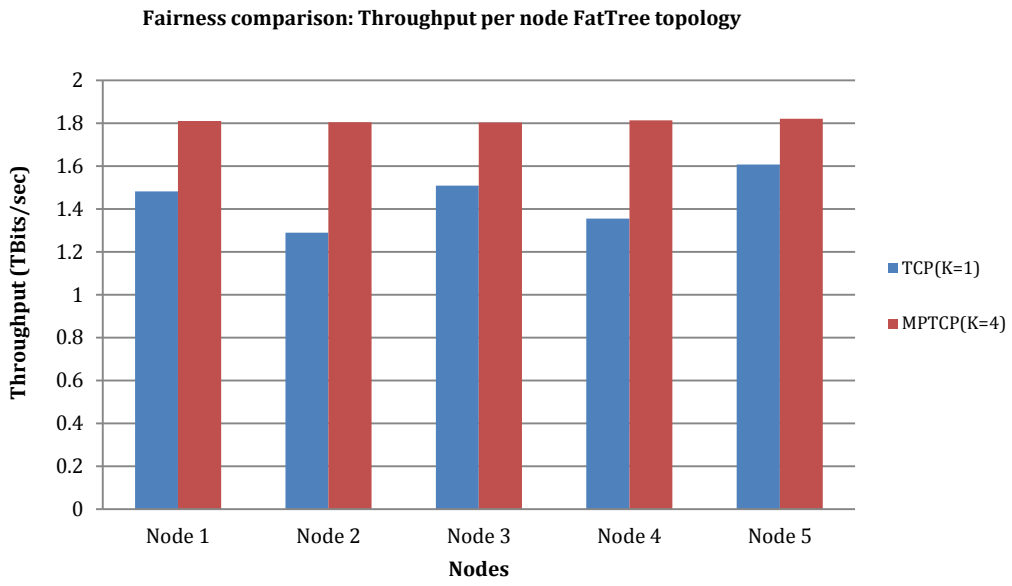

Figure 7-7: Arrival rate = 6.5 bursts/µs, W=64 (Large size)

7.5 Summary

Densely interconnected topologies such as Fat Tree and BCube leverage multiple parallel paths to offer high bandwidth between end hosts for datacenter. MPTCP over OBS is shown to efficiently utilize available multiple paths and provide improvement in throughput, fairness and robustness. It also makes the data center network more fault tolerant by providing alternative routes in situations of link/node failures. The benefit of MPTCP over OBS increases at higher traffic levels where multiple paths serve to alleviate severe bottlenecks and allows more efficient usage of network resources. In the future we plan to investigate schemes for dynamic load balancing on multiple paths of MPTCP and develop QoS algorithms using MPTCP over OBS in datacenter networks.

8. CHAPTER EIGHT: QOS IN SOFTWARE DEFINED OPTICAL NETWORKS

8.1 Introduction

Cloud based datacenters will be most suitable candidates for future software defined networking. The QoS requirements for shared data centers, hosting diverse applications could be successfully achieved through SDN architecture. This chapter provides an extension of our previously proposed scheme QAMO discussed in Chapter 6 that was aimed at achieving tangible QoS in datacenters through controlling bandwidth reservation in Multipath TCP and OBS layer while maintaining throughput efficiency. However, QAMO was designed for traditional networks and does not have the capability to adapt to current network status as expected from future software defined networks. The chapter presents an enhanced algorithm called QAMO-SDN that introduces a controller layer in previously proposed architecture and achieves adaptive QoS differentiation based on current network feedback. QAMO-SDN inherits the architecture of QAMO, using Multipath TCP over OBS networks. We evaluate the performance of QAMO-SDN under different network loads and topologies using realistic data center traffic models and present the results of our detailed simulation tests.

Rest of the chapter is organized as follows. In section 8.2, we review motivation for the proposed work. In section 8.3, we describe the proposed idea, our networking model that uses MPTCP protocol over an optical burst switching network for data centers, and 'QoS Aware Multipath TCP for Software Defined Optical Networks' algorithm. Simulation details are discussed performance analysis is given in Section 8.4. We conclude the chapter in Section 8.5.

8.2 Motivation for the proposed Idea

There is a growing interest in introducing QoS (Quality-of-Service) differentiation in datacenters, motivated by the need to improve the quality of service for time sensitive datacenter applications and to provide clients with a range of service-quality levels at different prices. There is also a growing trend towards software defined networks in datacenters and QoS schemes should adapt to SDN based cloud architectures. Software defined networks could be well understood by a simple analogy of sending a package through a courier who sets off his way in the network to deliver it. Traditionally he would ask different people and change his route multiple times to find an optimal path. With a Software defined network, assume that the courier has a GPS system with an up to date data of all the routes, traffic conditions, packet size and its requirements to find the best route for it dynamically before it is launched into the network. In the traditional network the routers contain the rules and logic for controlling the flow and modifications of packets. In traditional network there no centrally controlled mechanism to route the traffic. Software defined networks decouples the control plane from the data plane so the packet traverse the network with a pre-defined knowledge of the route it will take and the control of the traffic lies within the software defined network controller. Achieving QoS through current network feedback as expected in software defined networks will improve the performance and efficiency of QoS schemes.

8.3 Proposed Idea

The type of applications hosted by datacenters are diverse in nature ranging from back-end services such as search indexing, data replication, MapReduce jobs to front end services triggered by clients such as web search, online gaming and live video streaming [39]. The

background traffic contains longer flows and is throughput sensitive while the interactive front end traffic is composed of shorter messages and is delay sensitive. The traffic belonging to the same class can also have differences in relative priority levels and performance objectives [61].

In this chapter, we employ MPTCP over OBS for datacenter networks for efficiency and robustness as was done in our previous work [77] and present and evaluate a QoS provisioning algorithm in software defined networks called QAMO-SDN, 'QoS aware MPTCP over software defined optical network'. To our knowledge, this is the first research report that provides QoS provisioning algorithm for service differentiation using MPTCP over OBS in software defined datacenter.

8.3.1 Network model

Since SDN separates the control plane and data-forwarding plane, the entity that implements the control-plane functionalities is called the SDN controller. Software defined network has SDN capable devices hence software defined optical network will have SDN enabled optical cross connects that can communicate with upper layers[67, 105]. With an SDN architecture the controller layer has a lower level network view that enables the QoS schemes to perform prioritization of flows based on actual bandwidth on the links and network state. Figure 8-1 below shows the high level diagram of software defined network architecture.

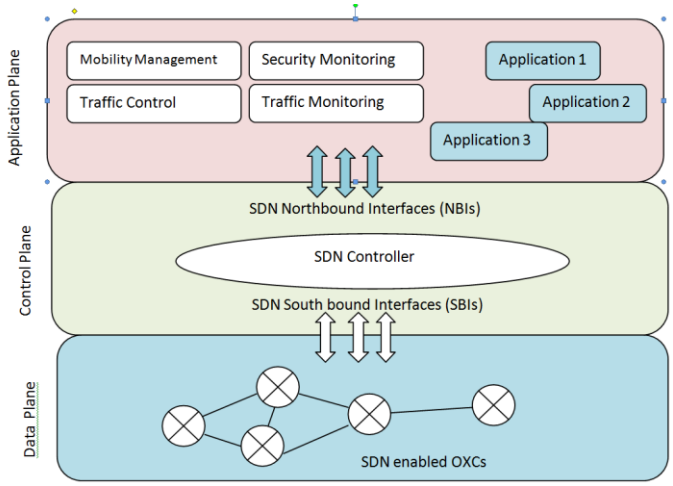

Figure 8-1: High level SDN architecture

With the popularity of new data center topologies such as Fat Tree and VL2 and the multitude of available network paths, it becomes natural to switch to multi path transport protocol such as MPTCP to seek performance gains. MPTCP provides significant improvement in bandwidth, throughput and fairness. We have used MPTCP over OBS in our proposed network architecture. In an OBS network, the control information is sent over a reserved optical channel, called the control channel, ahead of the data burst in order to reserve the wavelengths across all OXCs (Optical cross connects).In our SDN architecture we assume that our optical cross connects will be SDN enabled and will have the functionality to communicate the available wavelengths with upper layers [106]. The wavelength reservation protocol plays a crucial role in the burst transmission and we have used just-in-time (JIT) [19] for its simplicity. The necessary hardware level modifications of optical switches for supporting OBS in data centers have been discussed in [41], and will not be repeated in this chapter.

8.3.2 QoS aware MPTCP over OBS algorithm

Our proposed algorithm QAMO-SDN combines the multiple paths of MPTCP and resource reservation in OBS to develop an adaptive and efficient QoS-aware mechanism. Data centers handle a diverse range of traffic generated from different applications. The traffic generated from real time applications e.g., web search, retail advertising, and recommendation systems consists of shorter flows and requires faster response. These shorter flows (foreground traffic) are coupled with bandwidth intensive longer flow (background traffic) carrying out bulk transfers. The bottleneck created by heavy background traffic impacts the performance of latency sensitive foreground traffic. It is extremely important to provide a preferential treatment to time sensitive shorter flows to achieve an expected performance for data center applications. QoS technologies should be able to prioritize traffic belonging to more critical applications. Our proposed algorithm provides priority to latency-sensitive flows at two levels, i) MPTCP path selection stage and ii) OBS wavelength reservation stage. We propose that larger bandwidth be dynamically allocated to high priority flows, in order to minimize latency and reduce their drop probability. Datacenter networks are continuously changing and the concept of software defined networking is becoming increasingly popular. QoS algorithm should adapt to current network and dynamically change routing decisions that achieves service differentiation for current network state. QAMO-SDN algorithm just does that.

Let W be the maximum number of wavelengths per fiber, and K be the number of paths that exist between a given souce-destination pair. We will introduce a new term, the *priority factor P* for a burst priority defined as the ratio of P_{curr} (priority level of the current burst) to P_{max} (maximum priority levels) i.e., $P = P_{curr}/P_{max}$. Priorities of individual bursts are represented in

ascending order as P_1, P_2, P_3... P_{max} while P_{max} is the highest priority level in the bursts. As discussed before, the number of allocated paths for the burst of a particular priority level as follows.

$$\text{max\_paths} = \lceil K \times P \rceil \tag{8.1}$$

We also define a new vector $Path_{i,j}$ which is a collection of all paths that exist between nodes *i* and *j*. The number of paths this vector must store can be limited based on set value of *K* to reduce overhead. Another matrix is introduced in the algorithm, *L*, link state matrix. Each element $L_{i,j}$ of the matrix shows the state of the link between the nodes *i* and *j* as below:

$$L_{i,j} = \textit{number of available wavelengths between nodes i and j / W} \tag{8.2}$$

$L_{i,j}$ is initialized to 1 as all wavelengths are available and as the networks becomes congested, the matrix gets updated as shown in equation (2). We then sort the vector of paths in descending order of available bandwidths along the path. So the path having higher availability will be put on top of the path having lower number of paths. Since the number of paths in path vector can be limited, only the shortest paths will be chosen and then arranged accordingly. Lightpath creation is done only on the subset of paths from this vector. This subset is chosen from the top, so the path having higher wavelength availability will be preferred over others. At path allocation stage a larger number of paths is allocated for a high priority burst thus reducing its latency. For example, if P_{curr}=P_{max}, then P = 1. This will result in k_{curr} = K paths whereas if P_{curr} = 0.5*P_{max}, then *P* = 0.5 and the number of allocated paths is reduced to half the set of *K paths*. This will give the low priority burst, half the number of paths. We now define the size of the wavelength search space controlled by the following equation.

$$\textit{Wavelength search size} = \lceil W \times P \rceil \tag{8.3}$$

At wavelength reservation stage in OBS, equation 3 allocates a larger subset of wavelength search space for a burst with higher priority level thereby allowing it a greater chance to get through and reduce its blocking probability.

After reserving all the lightpaths for current source/destination pair, matrix L is updated according to equation (2) and made available to be used during other reservations.

QAMO (QoS Aware MPTCP over OBS) Algorithm

Input:
$P = P_{cur}/P_{max}$
K = maximum number of paths
W = maximum number of wavelengths
w_{cur} = current wavelength reserved for current burst
N_k = vector of all nodes on path k
$Paths_{i,j}$ = vector of all paths between node i and node j
$burst_{cur}$ = current burst
L = link state matrix
$L_{i,j}$ = state of link between node i and node j
N = vector of all nodes in the network

Algorithm:
```
if L is not initialized
  Initialize matrix L: set L_{i,j} = 1
arrange_paths(i, j)
max_paths = ⌈K × P⌉
for each k in max_paths:
  path_curr = Paths_{i, j}[k]
  lightpath(path_curr)
update_link_state_matrix()

function lightpath(path k)
Initialize w_cur
for each n in N_k:
  if n = N_k [length(N_k) – 1]              // destination node
    break;
  if n = N_k                          // source node
    for each w in ⌈W × P⌉:
      if w is free
        reserve w for burst_cur at n
        w_cur = w
```

```
            break;
        else
          if w_cur is free at n
            reserve w_cur for burst_cur at n
            continue;
          for each w in ⌈W × P⌉:
            if w is free
              reserve w for burst_cur at node n
              w_cur = w
              break;

          if no free wavelength at n
            return (error);                  // search failed at node n
        return(success);

        function arrange_paths(node i, node j)
          sort all paths in Paths_i, j in descending order of average bandwidth availability on all
links along the path

        function update_link_state_matrix()
          for each node i in N
            for each node j in N
              if i = j
                continue;
              L_i, j = number of available wavelengths between i and j / W
```

Figure 8-2: Algorithm QAMO-SDN

In the above algorithm, the priority factor P is used to adjust the number of allocated paths for concurrent transmission and the size of the wavelength search space based on the priority level of the burst. The chosen paths will always have higher availability compared to others. For high priority bursts, more concurrent MPTCP paths result in larger bandwidth, and more OBS network wavelengths reduce dropping probability. The parameter P_{max} can be flexible to accommodate changes in network statistics over time as bursts of different priority levels are encountered.

Figure 8-3 shows the cross layer design on QAMO-SDN algorithm. We assume that QAMO-SDN algorithm has access to available information about QoS requirements of different bursts and network conditions to process them correctly. The Controller layer receives feedback from lower layers and establishes an inner view of underlying network topology and stat in terms of link/node congestion. This layer provides feed back to QAMO-SDN layer to calculate the best path for new burst based on its priority level and current situation of wavelengths at OXS along various possible light paths.

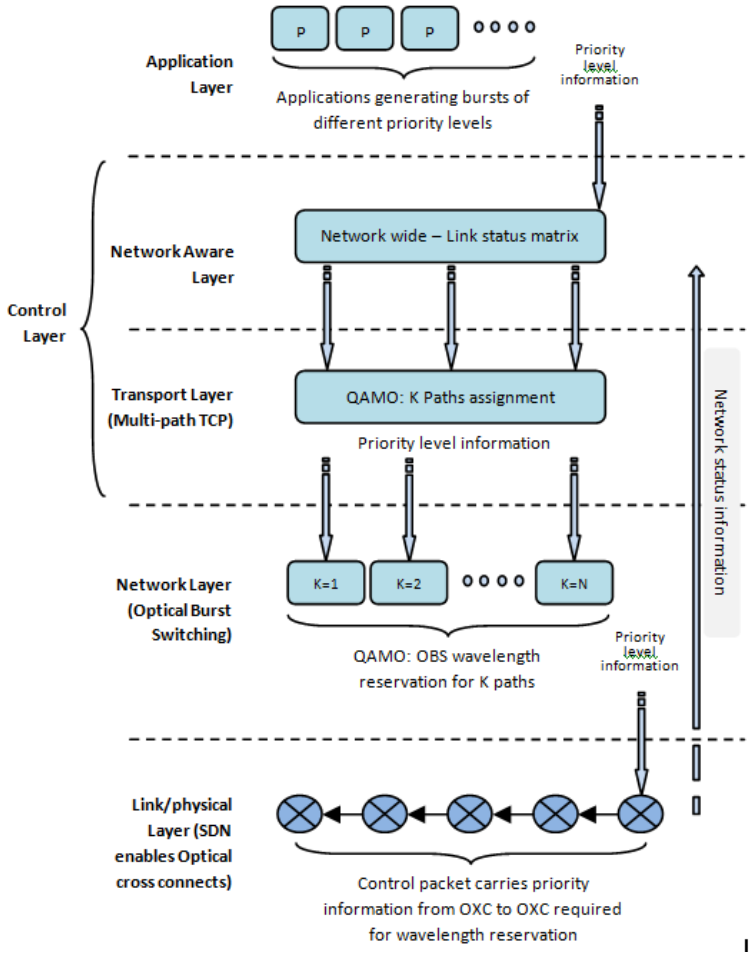

Figure 8-3: QAMO-SDN's cross-layer design: Changes to the Protocol stack and the burst priority level information flow.

We have assumed that priority level information will flow from application to MPTCP layer. This capability may be implemented using a specific interface such as the Implicit Packet Meta Header (IPMH) promoted in [100]. It is possible to assign priority levels for different flows in MPTCP at IPMH interface [101, 102]. Because of IPMH interface, it is also possible to gather priority information for each type of flow at a particular end host. This information can be passed on to the OBS network during burst segmentation process from MPTCP layer. At OBS network, the current burst priority P_{curr}, *or the ratio* $P = P_{curr}/P_{max}$, can be easily passed from one SDN capable OXC to the next and upper layers via the control packet and does not demand any significant resources in the OXC's. Implementing the reduced (adjustable) search as in the case of QAMO-SDN, to find a free wavelength requires minor modification to the standard JIT channel allocation scheme. The adjustable search in a smaller space of $\lceil W \times P \rceil$ for wavelengths actually leads to a smaller average search time.

The QAMO-SDN scheme has been extensively tested on the simulation testbed using data center network topologies FatTree and BCube and is shown to provide tangible QoS differentiation without negatively impacting the overall throughput of the system. It is also observed that QAMO-SDN utlises available capacity better than basic QAMO scheme due to SDN architecture.

8.4 Performance Evaluation

8.4.1 Simulation Details

The simulation testbed has been developed using C++. A source-destination pair amongst host nodes is randomly chosen for each originated burst. For TCP, to establish the static lightpath, simulation calculates the shortest path between these nodes using Dijkstra's algorithm.

In case of MPTCP, it uses K shortest paths algorithm (derived from Dijkstra's algorithm) to find K paths between the source-destination pair. The wavelength assignment heuristic is first-fit as done in [1, 2]. Recent research studies on traffic characteristics of data centers have shown that the traffic in data centers follows the lognormal distribution with ON-OFF pattern [43, 46]. The lognormal distribution is also considered to be the most fitted distribution for modeling various categories of internet traffic including TCP [103]. We have used lognormal arrival with an ON-OFF behavior in our simulation. The network nodes are assumed to be equipped with wavelength converters. We assume that MPTCP is running at end hosts. Based on the priority of the burst, K control packets originate from the source node to establish K lightpaths. Each control packet acquires an initial free wavelength at the source node, then travels to the destination node and reserves wavelengths following QAMO algorithm. If at any node, the same wavelength as the one reserved on the previous node is not available then it tries wavelength conversion. The process continues until the control packet either reaches the destination node or gets blocked due to the unavailability of free wavelength at any hop along the path. Thus, number of lightpaths established = K – number of control packets blocked. The source node waits for a predetermined time depending on the hop distance to the destination called offset time before transmitting the optical burst message. The traffic used in our simulation is uniformly distributed, i.e., any host node can be a source or a destination [1, 23].

The simulation clock is divided into time units, where each simulation time unit corresponds to 1 microsecond (μs). Each node has a control packet processing time of 20 microseconds and a cut through time of 1 microsecond as proposed for OBS networks in data

centers [40]. Each node can have a certain maximum number W of allowed wavelengths. Arrival rate/tu denotes the average arrival rate of the lognormal ON-OFF traffic.

In data center environment a complex mix of short and long flows is generated. The shorter flows are usually latency-critical and represent the largest proportion of flows in data centers [43]. The medium sized and longer flows constitute background traffic and may belong to different priority levels [94]. To represent these scenarios of data center mixed traffic, we have used variable burst sizes in different ranges with uniform distribution within each range [94].

Short burst sizes: S_{min}=5 Kbits to S_{max}= 20 KB

Medium burst sizes: S_{min}=200 Kbits to S_{max}= 1 MB

Long burst sizes: S_{min}=20 Mbits to S_{max}= 100 Mbits

Our traffic model is based on the findings on data center traffic characteristics in [39, 43, 46, 94]. To model our traffic we assume dynamically changing traffic with an average of 70-80% of bursts generated in short burst range belonging to latency sensitive applications, 10-15% in medium burst sizes while 5-10% of bursts belongs to large burst size range. In order to assign the priorities we use dynamically changing priority levels and relative percentages of various priority classes with an average of 95% short burst messages having the randomly assigned priorities from the highest priority range [P5-P6]; the remaining 5% can have any priority level. Similarly, 95% of medium and large burst sizes are randomly assigned priorities from sets [P3 – P4] and [P1 – P2] respectively. The remaining 5% from these ranges are assigned random priorities from set [P1 – P6]

8.4.2 Results and Discussion

The topologies used in our simulation tests are FatTree with 36 nodes and BCube with 24 nodes. All the figures in this section are tested following lognormal distribution. Because of the ON-OFF pattern of traffic the average arrival rate is smaller than the arrival rate of a continuous lognormal process having the same mean and standard deviation. The tests are conducted over burst distribution of our proposed traffic model.

Figure 8-4 motivates the use of MPTCP in data center networks for improving throughput. Figure 3 is tested using the lognormal distribution with mean μ=1.8 and standard deviation σ=1, corresponding to an arrival rate of 7.12/tu in BCube topology. Figure 3 shows the throughput comparison between TCP (K = 1) and MPTCP (K = 2, 3, 4), where K is the number of paths (i.e., number of subflows) used by each MPTCP connection. It can be observed that, MPTCP gives much higher throughput as compared to single path TCP. It can also be observed that MPTCP performs better with increasing number of paths. Similar results were achieved for FatTree topology.

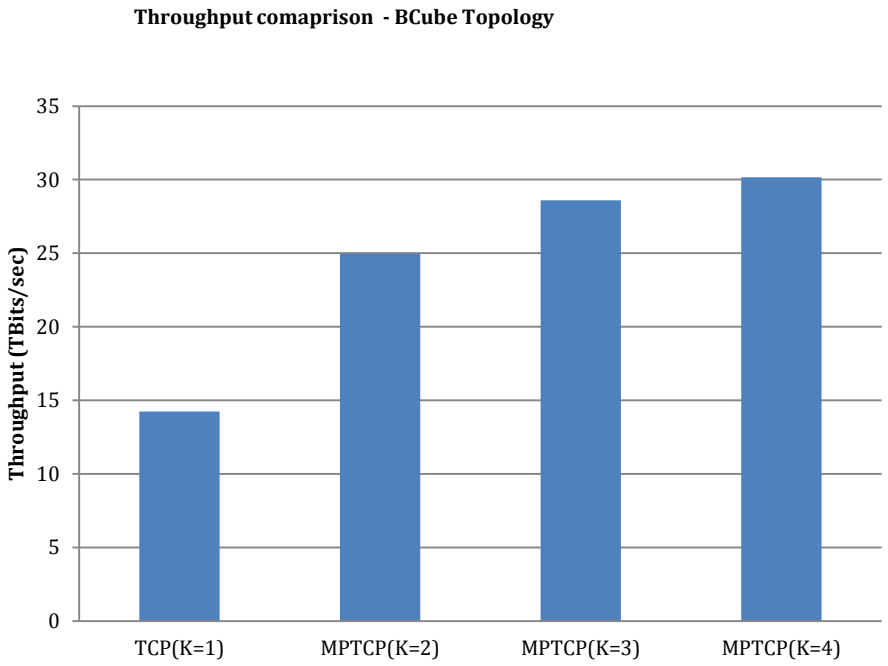

Figure 8-4: Arrival Rate /μs = 7.12, W=64

Figure 8-5 shows the ability of QAMO-SDN algorithm to achieve QoS differentiation when tested for bursts of various sizes and priority levels as proposed in our traffic model. The dropping probability comparison for six priority levels is shown with increasing load in a FatTree topology. For lognormal traffic, the mean values used in this test are from μ=1 to μ=3 and standard deviation σ=1. It can be observed that the algorithm achieves substantial QoS differentiation for all priority levels. For example, P6 being the highest priority level, experiences the least dropping at all values of input load. Similar results were achieved for BCube topology.

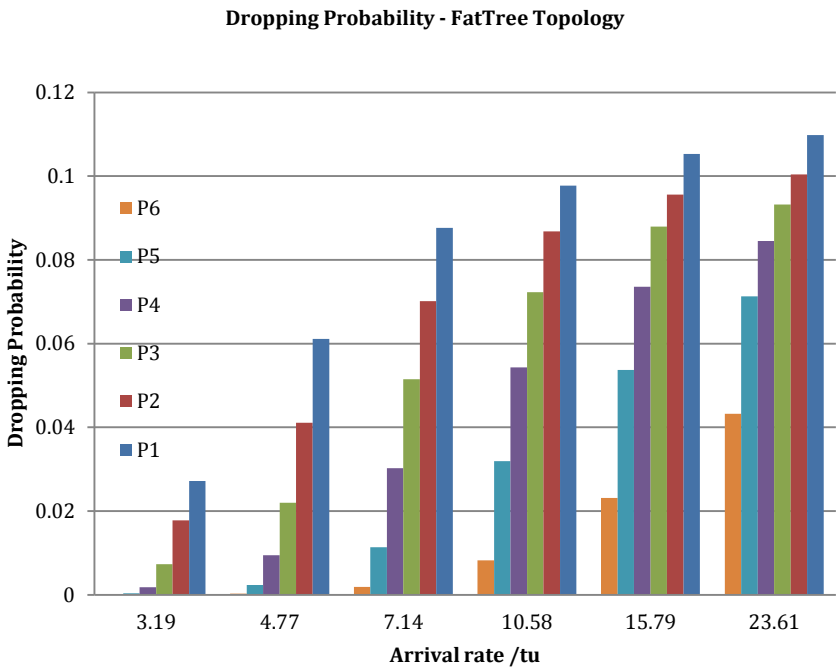

Figure 8-5: Variable arrival rate, W=64

Figure 8-6 shows the average throughput comparison of TCP, MPTCP (K=4), QAMO and QAMO-SDN. The lognormal mean values used in this test are from μ=0.5 to μ=1.75 and standard deviation σ=1. It can be observed that QAMO and MPTCP (K=4) both performs much better than standard TCP. The throughput of QAMO is slightly less than MPTCP (K=4) at small values of input load while the difference in throughput becomes less at higher loads. QAMO-SDN utilizes the available bandwidth better hence there is an improvement in QAMO-SDN compared to QAMO. The reason for QAMO's and QAMO-SDNs' degraded throughput is its preferential treatment for higher priority bursts, which are mostly very small in size.

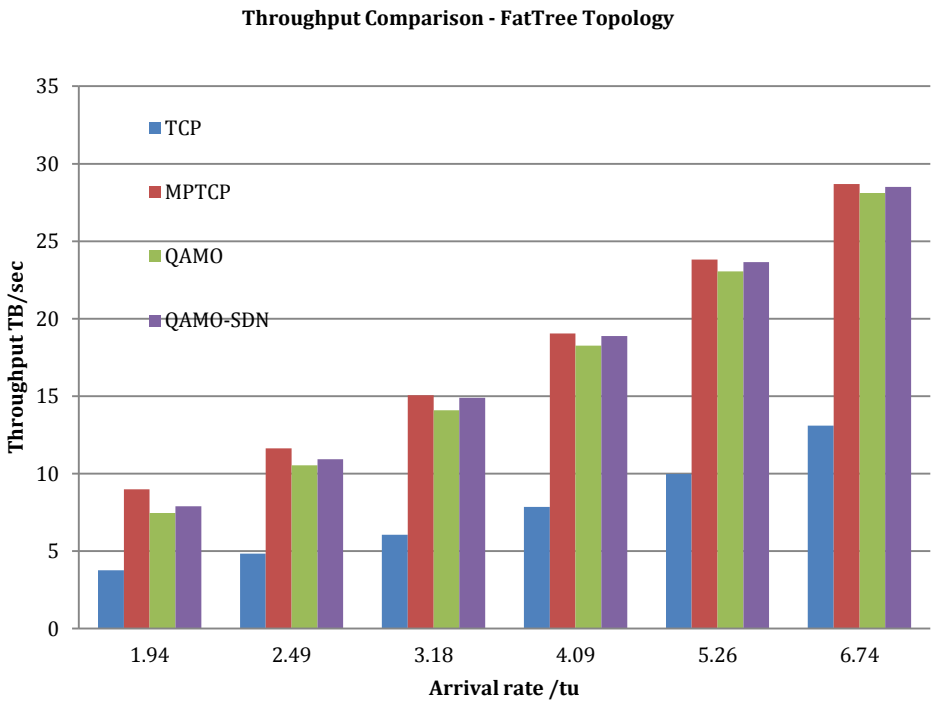

Figure 8-6: Variable arrival rate, W=64

Figure 8-7 provides deeper analysis of throughput breakdown in terms of burst priorities at one of the loads from Figure 5, specifically at arrival rate = 2.49 bursts/tu. The lognormal mean in Figure 6 is μ=0.75 and standard deviation σ=1. It can be observed that in TCP and MPTCP the greatest share of throughput is achieved by low priority background traffic, giving less importance to the time sensitive foreground flows in the absence of QoS provisioning. The throughput of QAMO and QAMO-SDN is well distributed between high priority (foreground) and low priority (background) traffic. Hence, the slight degradation of throughput compared to MPTCP is acceptable for achieving better share of network resources for more critical traffic in data centers. QAMO-SDN achieves better throughput than QAMO due to SDN architecture.

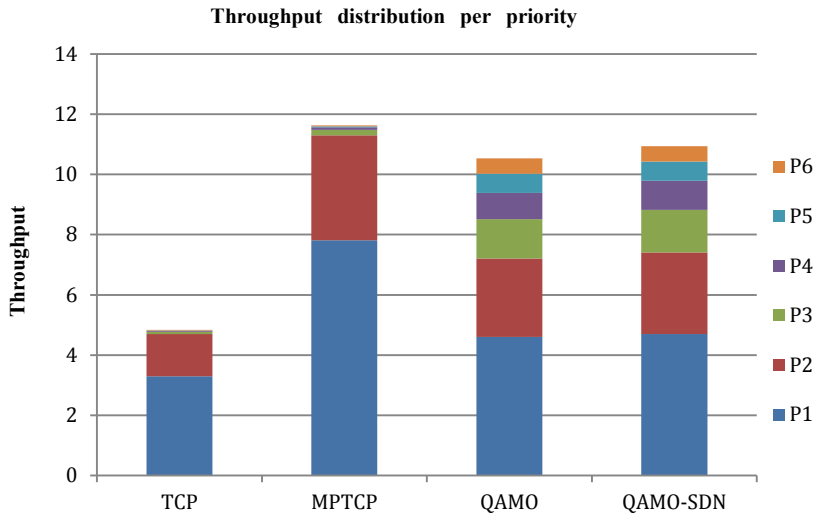

Figure 8-7: Arrival Rate /μs = 2.49, W=64

8.5 Summary

In this chapter we have shown a possible architecture of the Software defined optical network employing newly emerging transport protocol MPTCP over OBS networks and extended QoS provisioning algorithm for SDN in cloud datacenter. We have seen that MPTCP improves the throughput and reliability in data center networks by parallel transmission on multiple paths. We have presented and evaluated QoS-aware MPTCP over OBS for software defined optical networks (QAMO-SDN) scheme to provide service differentiation in data center traffic. QAMO-SDN provides tangible QoS differentiation to bursts of various classes without impacting the throughput of the system. QAMO-SDN is also an adaptive and self configurable scheme that changes its dynamics based on current network feedback making it applicable in software defined networks (SDN) for future datacenters. QAMO-SDN performs better than our previously proposed QAMO. It must also be noted that the slight improvement in throughput is not the only benefit of SDN architecture over standard QAMO. The motivation to use software

defined architecture lies in its simplicity, predictability, ease of network management through a central control and scaling and will continue to grow in the future.

9. CHAPTER NINE: IMPROVING THROUGHPUT FOR DATA CENTER NETWORKS AND HIGH PERFORMANCE COMPUTING

9.1 Introduction

Data centers and High performance computing networks share a number of common performance goals such as computational ability, exponentially growng demands for throughput, latency below microseconds and communication at the rate tens of gigabits/sec. As we have seen in Chapter 4 and Chapter 5, Mode-division multiplexing can offer an additional degree of freedom to enhance the bandwidth and throughput of optical networks. In this chapter, we present a highly efficient adaptive mode-wavelength-routing algorithm to improve the throughput of data centers and high performance computing networks. We will also present extensive simulation results to evaluate the proposed scheme.

Rest of the chapter is organized as follows. In section 9.2, we provide the motivations for the proposed idea and demostrate mode-wavelength division multiplexing (MWDM) with an example. In Section 9.3 we present the idea which is an adaptive mode-wavlength routing (AMWR) algorithm. We present the results of adaptive mode-wavelength-routing (AMWR) algorithm in section 9.4.1 and summarzie the chapter in section 9.5.

9.2 Motivation for the proposed Idea

As discussed in chapter 5, Wavelength-division multiplexed (WDM) carries bits/packets/bursts of information on an entire wavelength which is switched and routed completely in the optical domain using devices such as (reconfigurable) add/drop multiplexers [(R)OADM] and optical cross connects (OXC). In this chapter we provide an additional example of lightpath establishment in Optical networks based on WDM called wavelength-routed optical

networks (WRONs), shown schematically in Figure 9-1 and Optical networks based on MWDM called wavelength- and mode-routed optical network (WMRON), as shown in Figure 9-2.

Figure 9-1 shows wavelength routed optical OXCs having a single mode and two wavelengths λ_1 and λ_2. The figure depicts three connections with overlapping lightpaths. The lightpath of the first connection originating from node E to node H is identified by the red color and the lightpath of the second connection originating from node I to node H is identified by the blue color. While the third lightpath connection originating from node J to node H is shown in green color. The first connection from node E to node H uses the first available wavelength identified by the label λ_1 over the entire lightpath. The second connection from node I to node H requires a wavelength conversion from λ_1 to λ_2 at node F, as λ_1 is already in use by the first connection. This wavelength conversion is needed because otherwise the two connections will have conflict in the path from node F to node H. The third lightpath from node J to node H gets blocked at node G as both the possible wavelengths λ_1 and λ_2 are busy. In WRON, each connection must be identified by a unique wavelength. The number of connections is limited by the number of wavelengths supported by the WDM transport system.

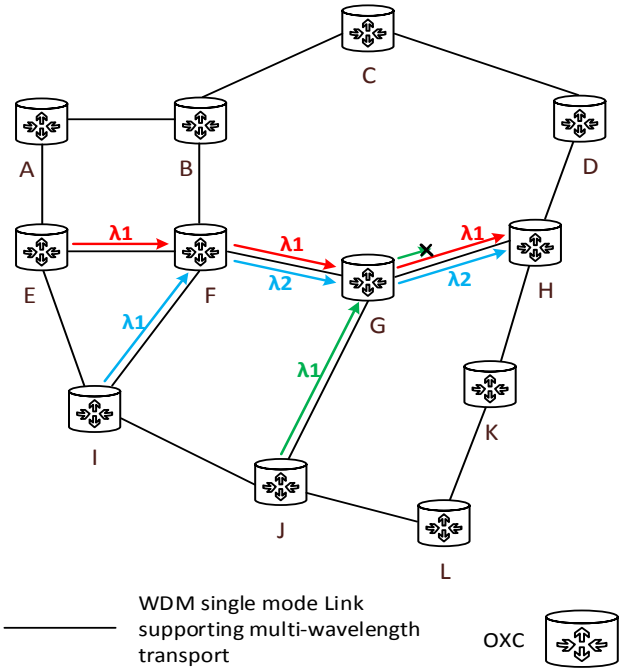

Figure 9-1: Wavelength Routed Optical Network

.

In Figure 9-2 Wavelength and mode routed optical network, we employ (spatial) mode-routing, in combination with wavelength routing, with wavelength-mode routed OXC (WMROXC) having two modes m1 and m2 and two wavelengths λ_1 and λ_2. The purpose of WMRON is to reduce the blocking probability and increase the throughput of future optical networks. Each optical transport link in WMRON supports not only multiple wavelengths (same as WDM) but also multiple spatial modes for each wavelength. As a result, the first two connections in the WMRON shown in Figure 9-2 Wavelength and mode routed optical network, corresponding to those in the WRON shown in **Error! Reference source not found.**, can be arried on the same wavelength λ_1 but using two different modes m_1 and m_2. The third connection can now avoid blocking by using the free wavelength λ_2 with mode m_1.

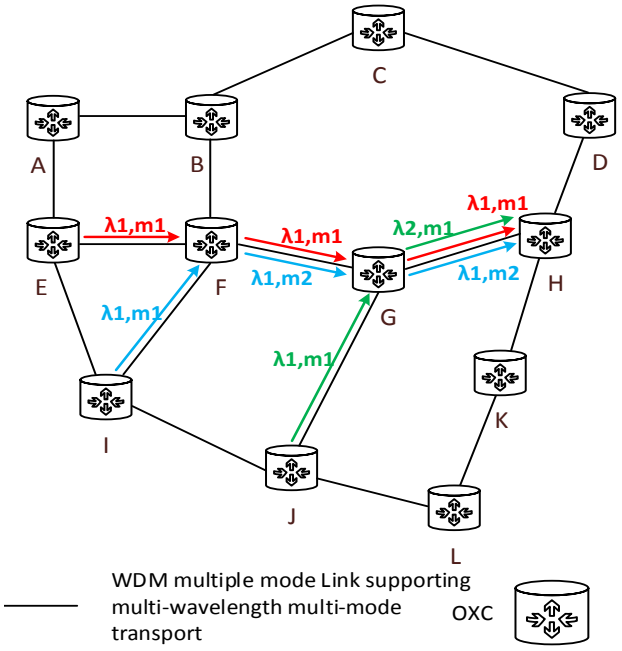

Figure 9-2 Wavelength and mode routed optical network

The network traffic of datacenters and HPCs will continue to increase in future driving a greater need for bandwidth and communication efficiency. In order to meet the throughput demands of future optical networks, we propose a simple and highly efficient routing algorithm for OBS networks called the Adaptive Mode-Wavelength-Routing scheme (AMWR), employing mode division multiplexing. OBS has been considered as the best compromise between OCS and OPS due to its granularity and bandwidth flexibility, and would be suitable for datacenters eventually as optical switching technology gets mature [44].

The proposed scheme uses a formula-based approach similar to the approach we used in the Hop-LC scheme [23] but with two major differences: i) Hop-LC is a single-mode fiber routing scheme whereas AMWR is a multi-mode fiber routing scheme that uses formulas for both mode-multiplexing and wavelength multiplexing; the formulas used in AMWR are entirely different formulas from those used in [23] and ii) the goal of Hop-LC is to enhance the hop-

count fairness of optical burst switched networks without negatively impacting the throughput of the network, whereas the goal of AMWR is to improve the throughput of the optical network without negatively impacting the hop-count fairness of the network. Hop-count fairness here means that bursts with longer lightpaths do not suffer excessively higher blocking probabilities than bursts with shorter lightpaths.

9.3 Proposed idea

Typically in an OBS network, the arriving bursts are of different sizes and a bandwidth reservation technique can use the burst size in making decisions that enhance the overall throughput of the system. AMWR does just that. Let M and W be the maximum number of modes and maximum number of wavelengths per fiber, respectively. We will introduce a new term, the *size factor* η for a burst defined as the ratio of the size S of that burst to the maximum allowed burst size S_{max}, i.e., $\eta = S/S_{max}$. We next define the size of the mode search space for that burst as follows.

$$Mode\ search\ size = \lceil M \times \eta \rceil \tag{9. 1}$$

For example, if $S=S_{max}$, then $\eta=1$ and the search size is the set of all M modes whereas if $S=0.5*S_{max}$, then $\eta=0.5$ and the search size is half the set of M modes, thereby giving the smaller burst, half the chance of finding a free mode. During the first mode search step, we will assume wavelength conversion cannot be done, i.e., we will search for available modes within the current wavelength only. We have chosen to perform mode search first before wavelength search because the value of M is typically smaller than W and because mode conversion can potentially be accomplished using only linear optics. If the first mode search step fails, we then try a wavelength search keeping the same mode, i.e., during the wavelength search step, we will

assume mode conversion cannot be done. The size of the wavelength search space is controlled by the following equation.

$$Wavelength\ search\ size = \lceil W \times \eta \rceil \quad (9.2)$$

If the wavelength search step also fails, we perform both mode and wavelength conversion simultaneously. The algorithm below shows the steps for this process.

Algorithm 2- Algorithm to find free Mode/Wavelength (AMWR)

Input:
$\eta = S/S_{max}$
M = maximum number of modes
W = maximum number of wavelengths
m_{curr} = current mode
w_{curr} = current wavelength
The pair [m_{curr}, w_{curr}] is not free on the output fiber

Output:
Algorithm:
for each m in $\lceil M \times \eta \rceil$:
    if [m, w_{curr}] is free
        m_{new} = m
        return [m_{new}, w_{curr}]; exit()
for each w in $\lceil W \times \eta \rceil$:
    if [m_{curr}, w] is free
        w_{new} = w
        return [m_{curr}, w_{new}]; exit()
for each m in $\lceil M \times \eta \rceil$:
for each w in $\lceil W \times \eta \rceil$:
        if [m, w] is free
            w_{new} = w
    m_{new} = m
        return [m_{new}, w_{new}], exit()
Return(Error) // search has failed

Figure 9-3 Algorithm (AMWR) find free Mode/Wavelength

In the above algorithm, the size factor η is used to adjust the size of the mode search subset and the size of the wavelength search subset based on the size of the current burst, thereby allowing a larger number of modes and wavelengths to be searched for larger bursts. For example if we have two bursts of different sizes, the AMWR scheme will provide a larger wavelength search space to the burst that has larger size hence this larger burst will contribute in obtaining higher throughput. It should be mentioned that the AMWR scheme does not affect the hop-count fairness of the optical network, positively or negatively. The AMWR scheme and the SPF scheme have essentially the same level of hop-count fairness. This is because the size of the burst and the length of its lightpath can be considered independent variables, and generally speaking, small bursts and large bursts have equal likelihood to be destined to near or far destination nodes. In order to appreciate the role of the size factor $\eta = S/S_{max}$ and the flexibility of the AMWR scheme in setting the maximum burst size in each node, it is important to note that equations 6.1 and 6.2 of the AMWR scheme are executed independently by each node (wavelength-mode routed OXC) when it receives a new burst. The parameter S_{max} in a node represents the maximum burst length for this node, which could be smaller or larger than S_{max} in other nodes. The parameter S_{max} in each node can flexibly change over time as this node encounters different size distributions of bursts.

Our proposed AMWR scheme is easy to implement and does not demand any significant resources in the OXC's. The current burst size *S, or the ratio* $\eta = S/S_{max}$, could be easily passed from one OXC to the next via the control packet. Implementing the reduced (adjustable) search as in the case of AMWR, to find a free wavelength requires minor modification to the standard

SPF channel allocation scheme. The adjustable search in a smaller space of size $\lceil M \times \eta \rceil$ for modes or size $\lceil W \times \eta \rceil$ for wavelengths actually leads to a smaller average search time.

9.4 Performance Evaluation

In this section we will discuss simulation detail, network topologies used in our tests and present performance results with their analysis.

9.4.1 Simulation detail

Our OBS simulation testbed assumes that assembled bursts arrive at the network with lognormal distribution with an ON-OFF pattern as found in datacenters [43, 46]. We have also used the lognormal arrival process to simulate arrivals of new requests in High Processing Computing systems as done in [107, 108]. A source-destination pair is randomly chosen for each arriving burst. The load in the lognoraml arrival pattern is controlled by two variables: Mean μ and Standard Deviation σ. The two schemes SPF_MWDM and AMWR are tested using various network loads and burst sizes. To establish the static lightpath for source-destination pair, the simulation software calculates the shortest path between these nodes using Dijkstra's algorithm. The network nodes are assumed to be equipped with mode as well as wavelength converters. The simulation clock is divided into time units, where each simulation time unit corresponds to 1 microsecond. Each node has a control packet processing time of 20 microseconds [40] and a cut through time of 1 microsecond.

Many cloud applications could produce bursts of large sizes and some HPC applications produce bursts of medium sizes. We have used uniformly distributed variable burst sizes in three ranges (from S_{min} to S_{max}) as follows:

- Fat Tree and BCube (datacenter topologies): S_{min} = 100 Mb to S_{max} = 400 Mb.
- Fat Tree and BCube (datacenter topologies): S_{min} = 250 Mb to S_{max} = 1000 Mb.
- Mesh torus and 6D Mesh torus (supercomputers/HPC topologies): S_{min} = 128 Kb to S_{max} = 512 Kb.

9.4.2 Network Topologies

Because current hardware implementations for MDM, Mode division multiplexed networks are suitable for short distances, our simulation tests used topologies suitable for short transmission distances such as those found in the optical interconnects for datacenters and supercomputers. We used FatTree topology with 36 nodes (Figure 5-3) and the BCube topology with 24 nodes (Figure 5-4); both topologies are used in modern datacenters [109] and discussed in chapter 5. We also used two more network topologies in our simulation tests: the 5x5 Mesh torus with 25 nodes (Figure 3-3) and the 3x3 6D Mesh Torus with 27 nodes. The 3x3 6D mesh topology as shown in Figure 9-4 is used in the Fujitsu K next generation supercomputer project [110]. We adopt the same traffic model used in [23], which is uniformly distributed such that any node can be a source or a destination.

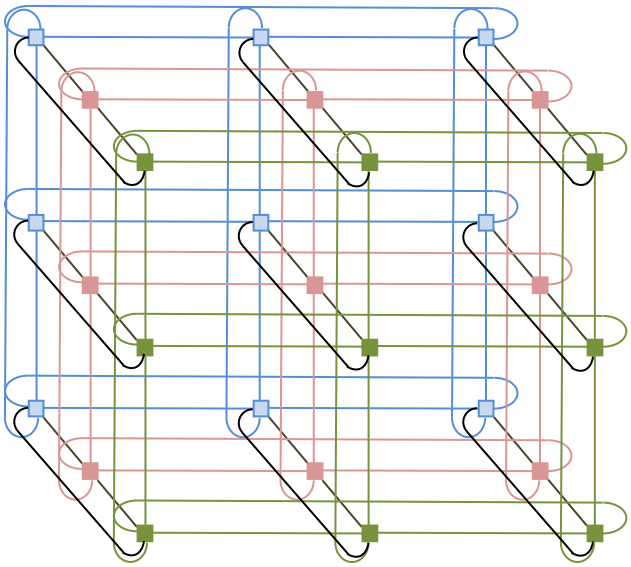

Figure 9-4 3x3 6D Mesh Torus

9.4.3 Performance results and discussion

In this section, we present results on AMWR scheme over optical burst switched (OBS) networks equipped with both mode division multiplexing (MDM) and wavelength division multiplexing (WDM). For routing, we select the standard shortest path first (SPF) algorithm and we use the notation SPF_WDM to denote routing using WDM only and SPF_MWDM to denote routing using both MDM and WDM. For mode and wavelength assignment, we select the First-Fit heuristic.

The AMWR scheme has been extensively tested using the simulation testbed and the four short-distance network topologies discussed in section 9.4.2. Figure 9-5 gives a comparison between the throughput of SPF_MWDM and the throughput of the AMWR scheme. The traffic has lognormal distribution with ON-OFF behavior. Mean for lognormal is $\mu=3$ and standard deviation $\sigma=1$. Burst sizes used in this test are the range of 250 -1000 Mb. Figure 9-5 shows that

the throughput of AMWR is consistently better than that of SPF_MWDM for all numbers of modes.

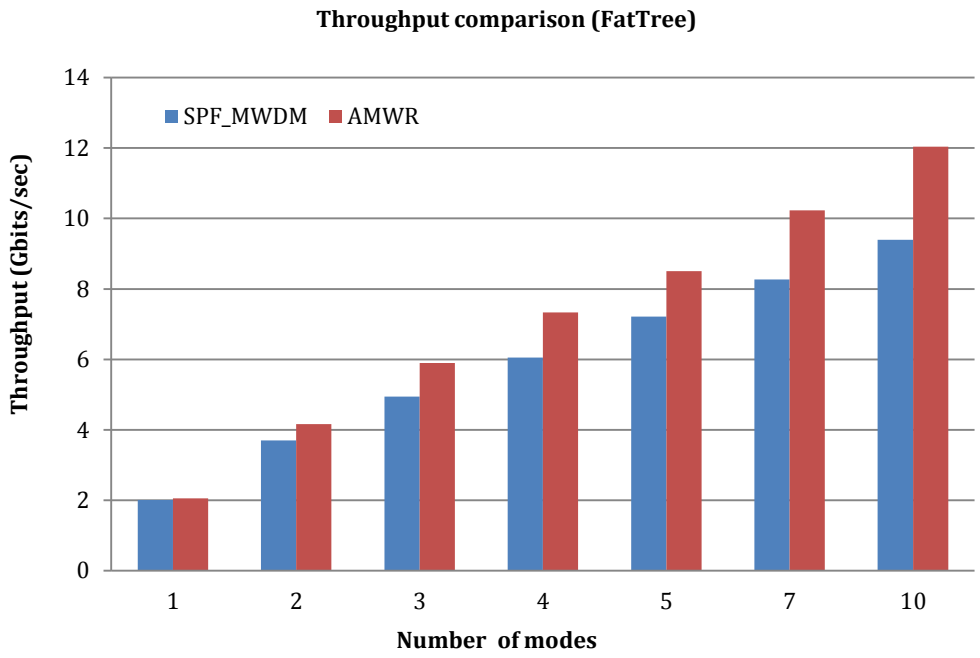

Figure 9-5 Throughput comparison FatTree, Max Wavelengths=16, Arrival rate=23.6/s

Figure 9-6 gives a comparison between the throughput of SPF_MWDM and the throughput of AMWR for the BCube topology when the number of modes is equal to M=4. Traffic arrival is lognormal with ON-OFF pattern. Burst sizes used in this test are in the range of 100-400 Mb. The throughput of AMWR is better than that of SPF_MWDM for all values of the input load. Similar results were obtained for the other network topologies.

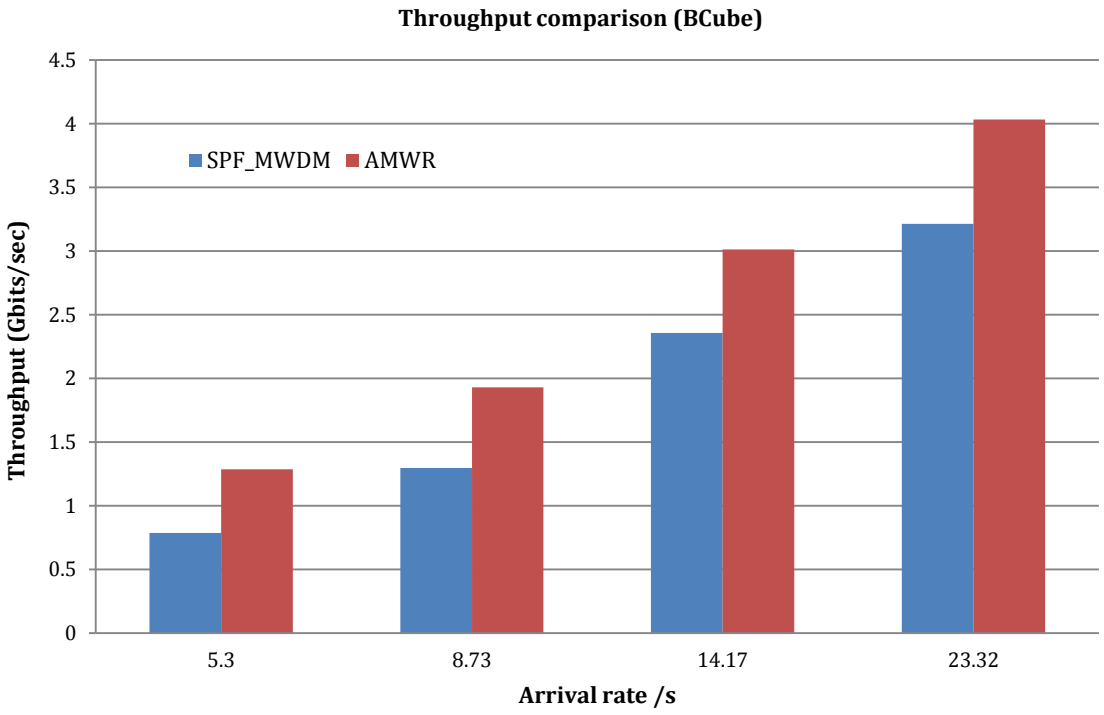

Figure 9-6 Throughput comparison BCube, Max Wavelengths=20, Modes= 4

Figure 9-7 shows the throughput of the SPF_MWDM scheme and the throughput of the AMWR scheme for the 3x3 6D mesh torus network using three modes (M=3) and different ranges of burst sizes. New bursts arrive with lognormal distribution having a mean μ=3.0 and standard deviation σ=1. The range of burst sizes S_{min} ─ S_{max} is increased gradually from 128─256 Kb to 128─1280 Kb. It can be observed that with a larger range (i.e., greater size difference between largest and smallest bursts), the AMWR scheme outperforms SPF_MWDM with an increasing margin. As the range [S_{min} ─S_{max}] increases, the blocking of bursts in the network also increases due to longer bursts holding the resources for longer time. Under these circumstances, favoring larger bursts slightly as done in AMWR enhances throughput over SPF_MWDM.

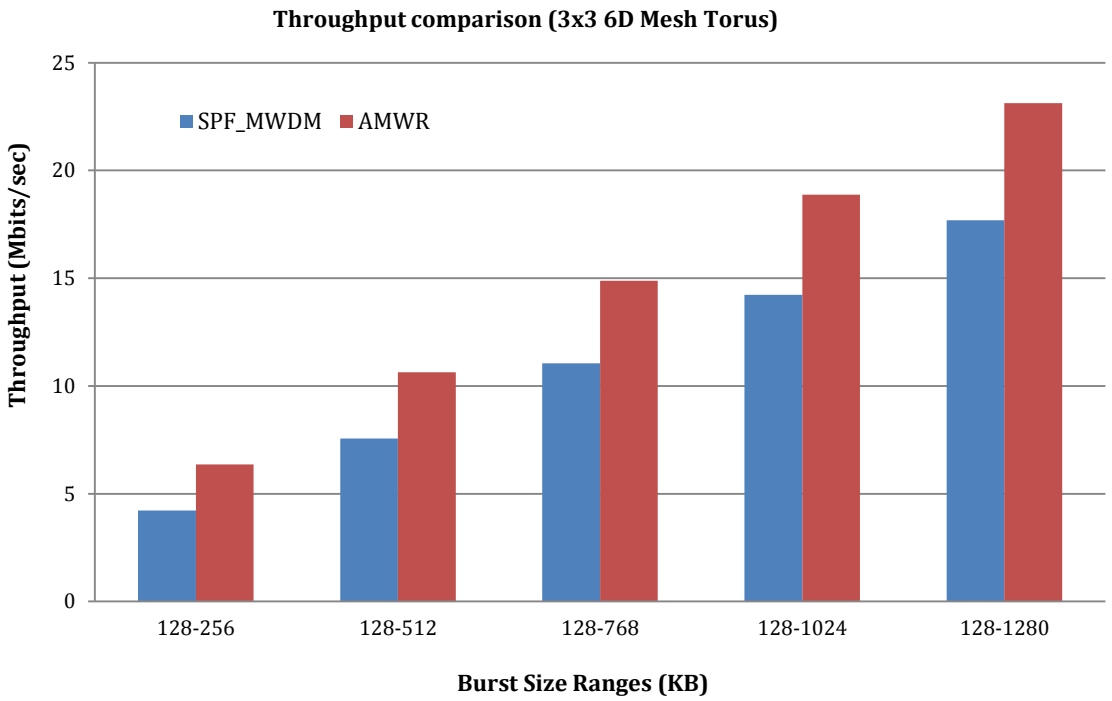

Figure 9-7 Throughput comparison 3x3 mesh torus, Max Wavelengths=20, Modes=3, Arrival rate =33/s

9.5 Summary

In this chapter we have discussed mode division multiplexing as a new dimension in increasing network bandwidth. We have proposed a simple and highly efficient routing algorithm called adaptive mode wavelength routing, AMWR employing combined wavelength and mode division multiplexing. We have tested the performance of our proposed algorithm using various data center and HPC topologies and have shown good improvement in all scenarios.

10.CHAPTER TEN: CONCLUSION AND FUTURE WORK

Below, we present a summary of our contributions in different areas of optical communication networks improving fairness, throughput and blocking performance in single and multimode fiber networks. We have done work in both Wide area networks (WAN) for long haul optical communication and short reach optical networks of datacenter and HPCs. In this chapter we will briefly discuss our main findings and contributions. We will also provide possible extension of our various contributions and planned future work.

10.1 Summary of Contributions

In first contribution [1], we presented two new schemes, BJIT-S and PRED-S, that considered the burst size to maximize throughput without affecting fairness in OBS networks. We evaluated the effectiveness of these schemes in maximizing throughput of the OBS networks with simulations. Our schemes have proven to be effective in maximizing throughput in the US Long Haul and Mesh networks. These networks were extensively tested with variable network loads, various values of factor g, and the number of wavelengths W at OXCs. Under all test conditions, both PRED-S and BJIT-S have shown to perform better than JIT, BJIT and PRED. Both schemes do not preempt any burst after the burst has been accepted and the lightpath has been established. Blocked bursts will not waste any bandwidth resources in the core of the optical network. The two schemes can be used to efficiently improve the throughput of optical OBS networks and enhance the hop-count fairness.

In our next contribution [2] we proposed two new schemes using Wavelength-Division Multiplexing (WDM) as well as mode division multiplexing (MDM) in optical fiber networks. We have seen that multi-mode fibers can serve as a promising technology in all areas of optical

networks. The availability of multiple modes over the same fiber can multiplicatively increase the available wavelengths and adds a new dimension to enhancing capacity of the network. We first proposed scheme FFOR was proved to improve fairness in OBS for multimode fiber networks while our second scheme FTFOR was shown to maximize throughput while maintaining the fairness of FFOR by selectively giving priority to larger bursts over smaller bursts. Multi-mode fiber networks is expected to be one of the next big breakthroughs in the field of optical networks and the schemes proposed in this contribution represent a first attempt to solve the fairness problem in multimode OBS networks.

Data centers have become the heart of the computational world over the past few years. The emergence of cloud computing and the growth of data-intensive applications have driven the need for finding alternative ways to improve communication efficiency in data center networks. The following contributions addressed datacenter's performance objectives and proposed various approaches to improve bandwidth, throughput, reliability and quality of service differentiation.

In Chapter 5, we discussed mode division multiplexing in a greater detail describing its current state and feasibility in future. We presented the opportunities possible for future optical communication networks using Mode division multiplexing. We showed the significant benefits of using both mode division multiplexing and wavelength division multiplexing in real-life short-distance optical networks such as the optical circuit switching networks used in the hybrid electronic-optical switching architectures for datacenters. We next evaluated four mode and wavelength assignment heuristics and compared their throughput performance. To our knowledge, this was the first research work that evaluated mode division multiplexing and presented results on mode-wavelength assignment for wavelength-mode-routed optical networks.

We concluded the chapter by evaluating the impact of the cascaded mode conversion constraint on network throughput.

In chapter 7 we propose and evaluate the potential benefits of implementing the newly emerging transport protocol, Multipath TCP, over an optical OBS network in data centers [77]. Our proposed data center networking strategy is evaluated over the FatTree and BCube topologies and our tests have established that Multipath TCP over OBS provides huge performance advantage in terms of improving throughput, reliability and robustness of data center networks.

Shared datacenter networks supporting diverse range of applications constitute a complex mix of workloads from multiple organizations. Some workloads require small predictable latency while others require large sustained throughput. Such shared data-centers are expected to provide potential service differentiation to client's individual flows. We presented a simple and efficient service differentiation scheme called 'QoS aware MPTCP over OBS' (QAMO) in datacenters. Our extensive experimental results showed that QAMO algorithm achieves tangible service differentiation without impacting the throughput of the system.

In our next contribution, an extension of QAMO called QAMO-SDN has been proposed that presents architecture for software defined networks using MPTCP over OBS and proposes and evaluates service differentiation scheme QAMO-SDN for software defined optical datacenter networks.

In Chapter 9, a new scheme is presented for improving throughput in datacenter and High Performance Computing networks. Typically in an OBS network, the arriving bursts are of different sizes and a bandwidth reservation technique can use the burst size in making decisions that enhance the overall throughput of the system. Adaptive mode-wavelength-routing (AMWR)

scheme just does that. Extensive simulation results on different network topologies showed that the scheme significantly improves the throughput of data centers and high performance computing networks. The proposed scheme is simple, efficient and easy to implement. This research work presented routing results for wavelength-mode-routed optical networks.

10.2 Proposed future work

A number of schemes and approaches presented in this dissertation can be extended in a variety of ways. The schemes for hop count fairness problem discussed in Chapter 3 and Chapter 4 can be extended for software defined cloud based optical datacenter networks. The proposed schemes are suitable for long haul optical networks and have considered regular optical cross-connects. Software defined networks assume using network devices that are programmable [105].

Similarly, the hybrid electro-optical datacenter networks that utilized mode-division multiplexing discussed in Chapter 5, can be extended in different ways. The Mode division multiplexing technology will continue to mature and will open new possibilities for designing robust routing algorithms and wavelength assignment techniques to exploit the multiplicative increase in data rates offered by them. The enhanced data rates can find significant attention for Big Data network architectures.

The work shown on MPTCP over OBS in Chapter 6, 7 and 8 can be extended in multiple ways. We have evaluated the performances of MPTCP over OBS networks using Just-in-time (JIT) wavelength reservation technique; other wavelength assignment heuristics can be evaluated to compare the relative performance in this architecture. Multi-Path TCP can be evaluated for all possible network architectures where there are multiple paths and regular TCP is currently

utilized and its performance should be evaluated considering optical networks in those scenarios. The performance of MPTCP can compared against other multi-path transport protocols such as Stream Control Transmission Control (SCTP) in optical datacenter networks.

Software defined networking will open the possibilities for a number of new research areas in optical communication networks. Wavelength reservation schemes and routing algorithms will need to keep the SDN architecture in design perspective to ensure that data plane and control plane are separated in assumed architecture.

Another interesting area of research could be developing an adaptive routing algorithm that considers cascaded mode-wavelength conversion constraint in software defined networks. The controller layer will not only have the information of current network in terms of links utilization but also the number of conversions that happened on each path and will govern the routing decisions in light of this knowledge.

LIST OF REFERENCES

[1] S. Tariq and M. A. Bassiouni, "Hop-count fairness-aware protocols for improved bandwidth utilization in WDM burst-switched networks," *Photonic Network Communications,* vol. 25, pp. 35-46, 2013.

[2] S. Tariq, M. Bassiouni, and G. Li, "Improving Fairness of OBS Routing Protocols in Multimode Fiber Networks," In the Proceedings of IEEE International Conference on Computing, Networking and Communications (ICNC), 2013.

[3] S. Tariq and M. Bassiouni, "Performance evaluation of MPTCP over optical burst switching in data centers," In the Proceedings of IEEE International *Telecommunications Symposium (ITS), 2014 International*, 2014, pp. 1-5.

[4] S. Tariq and M. bassiouni, "QAMO: QoS Aware Multipath-TCP Over Optical Burst Switching in Data Centers," In the Proceedings of International Conference on Optical Communication Systems (OPTICS 2014) Vienna, Austria, 2014.

[5] S. Tariq and M. Bassiouni, "QAMO-SDN: QoS Aware Multipath TCP for Software Defined Optical Networks," In the Proceedings of 12th annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vagas, NV, 2015.

[6] S. Xu, L. Li, and S. Wang, "Dynamic routing and assignment of wavelength algorithms in multifiber wavelength division multiplexing networks," *Selected Areas in Communications, IEEE Journal on,* vol. 18, pp. 2130-2137, 2000.

[7] W. Zhang, D. Liu, H. Wang, and K. Bergman, "Experimental demonstration of wavelength-reconfigurable optical packet-and circuit-switched platform for data center networks," *Proc. 1st IEEE Opt. Int. Con.(OI),* 2012.

[8] H. Zang, J. P. Jue, and B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks," *Optical Networks Magazine,* vol. 1, pp. 47-60, 2000.

[9] B. Ramamurthy and B. Mukherjee, "Wavelength conversion in WDM networking," 1998.

[10] R. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *Lightwave Technology, Journal of,* vol. 28, pp. 662-701, 2010.

[11] C. Koebele, M. Salsi, L. Milord, R. Ryf, C. A. Bolle, P. Sillard*, et al.*, "40km transmission of five mode division multiplexed data streams at 100Gb/s with low MIMO-DSP complexity," in *European Conference and Exposition on Optical Communications*, 2011, p. Th. 13. C. 3.

[12] E. Ip, M.-J. Li, Y.-K. Huang, A. Tanaka, E. Mateo, W. Wood*, et al.*, "146λx6x19-Gbaud Wavelength-and Mode-Division Multiplexed Transmission over 10x50-km Spans of Few-Mode Fiber with a Gain-Equalized Few-Mode EDFA," in *Optical Fiber Communication Conference*, 2013, p. PDP5A. 2.

[13] N. Bai and G. Li, "Adaptive frequency-domain equalization for mode-division multiplexed transmission," *Photonics Technology Letters, IEEE,* vol. 24, pp. 1918-1921, 2012.

[14] N. Riesen, J. Love, and J. Arkwright, "Few-core spatial-mode multiplexers/demultiplexers based on evanescent coupling," 2013.

[15] N. Bai, E. Ip, Y.-K. Huang, E. Mateo, F. Yaman, M.-J. Li*, et al.*, "Mode-division multiplexed transmission with inline few-mode fiber amplifier," *Optics express,* vol. 20, pp. 2668-2680, 2012.

[16] B. Huang, C. Xia, G. Matz, N. Bai, and G. Li, "Structured directional coupler pair for multiplexing of degenerate modes," in *National Fiber Optic Engineers Conference*, 2013, p. JW2A. 25.

[17] J. J. Rodrigues, M. M. Freire, N. M. Garcia, and P. P. Monteiro, "Enhanced Just-in-Time: A New Resource Reservation Protocol for Optical Burst Switching Networks," in *ISCC*, 2007, pp. 121-126.

[18] J. Y. Wei, J. L. Pastor, R. S. Ramamurthy, and Y. Tsai, "Just-in-time optical burst switching for multiwavelength networks," in *Broadband communications*, 2000, pp. 339-352.

[19] J. Y. Wei and R. I. McFarland, "Just-in-time signaling for WDM optical burst switching networks," *Journal of lightwave technology,* vol. 18, p. 2019, 2000.

[20] B. Zhou, M. Bassiouni, and G. Li, "Improving fairness in optical-burst-switching networks," *Journal of Optical Networking,* vol. 3, pp. 214-228, 2004.

[21] T. Orawiwattanakul, Y. Ji, and N. Sonehara, "Fair bandwidth allocation with distance fairness provisioning in optical burst switching networks," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, 2010, pp. 1-5.

[22] X. Gao and M. A. Bassiouni, "Fairness-improving adaptive routing in optical burst switching mesh networks," in *Communications, 2008. ICC'08. IEEE International Conference on*, 2008, pp. 5209-5213.

[23] X. Gao and M. A. Bassiouni, "Improving fairness with novel adaptive routing in optical burst-switched networks," *Journal of Lightwave Technology,* vol. 27, pp. 4480-4492, 2009.

[24] C.-F. Hsu and L.-C. Yang, "On the fairness improvement of channel scheduling in optical burst-switched networks," *Photonic Network Communications,* vol. 15, pp. 51-66, 2008.

[25] B. O. Nassar, T. Tachibana, and K. Sugimoto, "Random scheduling based on transmission delay and buffer size for hop-based burst-cluster transmission in OBS networks," *Photonic Network Communications,* vol. 19, pp. 292-300, 2010.

[26] S. K. Tan, G. Mohan, and K. C. Chua, "Link scheduling state information based offset management for fairness improvement in WDM optical burst switching networks," *Computer Networks,* vol. 45, pp. 819-834, 2004.

[27] H. Li, M. W. L. Tan, and I. L.-J. Thng, "Fairness issue and monitor-based algorithm in optical burst switching networks," *Computer Networks,* vol. 50, pp. 1384-1405, 2006.

[28] J. Li, C. Qiao, J. Xu, and D. Xu, "Maximizing throughput for optical burst switching networks," *IEEE/ACM Transactions on Networking (TON),* vol. 15, pp. 1163-1176, 2007.

[29] B. Zhou and M. A. Bassiouni, "Threshold-based preemption scheme for improving throughput in OBS networks," *Photonic Network Communications,* vol. 24, pp. 12-21, 2012.

[30] S. Peng, Z. Li, Z. Zhang, Y. He, and A. Xu, "Drop policies and multiple edge thresholds to enhance the performance of TCP over OBS networks with retransmission," *Photonic Network Communications,* vol. 17, pp. 183-190, 2009.

[31] G. B. Figueiredo, E. Candido Xavier, and N. L. Da Fonseca, "Optimal algorithms for the batch scheduling problem in OBS networks," *Computer Networks,* vol. 56, pp. 3274-3286, 2012.

[32] G. Gurel and E. Karasan, "Using multiple per egress burstifiers for enhanced TCP performance in OBS networks," *Photonic Network Communications,* vol. 17, pp. 105-117, 2009.

[33] C. Y. Li, P. Wai, and V. O.-K. Li, "Performance improvement methods for burst-switched networks," *Journal of Optical Communications and Networking,* vol. 3, pp. 104-116, 2011.

[34] N. Barakat and E. H. Sargent, "Dual-header optical burst switching: A new architecture for WDM burst-switched networks," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, 2005, pp. 685-693.

[35] M. Yoo and C. Qiao, "New optical burst-switching protocol for supporting quality of service," in *Photonics East (ISAM, VVDC, IEMB)*, 1998, pp. 396-405.

[36] B. Komatireddy, N. Charbonneau, and V. M. Vokkarane, "Source-ordering for improved TCP performance over load-balanced optical burst-switched (OBS) networks," *Photonic Network Communications,* vol. 19, pp. 1-8, 2010.

[37] R. Ryf, S. Randel, N. K. Fontaine, M. Montoliu, E. Burrows, S. Chandrasekhar*, et al.*, "32-bit/s/Hz spectral efficiency WDM transmission over 177-km few-mode fiber," in *Optical Fiber Communication Conference*, 2013, p. PDP5A. 1.

[38] C. Raiciu, C. Pluntke, S. Barre, A. Greenhalgh, D. Wischik, and M. Handley, "Data center networking with multipath TCP," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, 2010, p. 10.

[39] Y. Chen, S. Jain, V. K. Adhikari, Z.-L. Zhang, and K. Xu, "A first look at inter-data center traffic characteristics via yahoo! datasets," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 1620-1628.

[40] S. Saha, J. S. Deogun, and L. Xu, "Hyscaleii: A high performance hybrid optical network architecture for data centers," in *Sarnoff Symposium (SARNOFF), 2012 35th IEEE*, 2012, pp. 1-5.

[41] M. Y. Sowailem, D. V. Plant, and O. Liboiron-Ladouceur, "Implementation of optical burst switching in data centers," in *Photonics Conference (PHO), 2011 IEEE*, 2011, pp. 445-446.

[42] C. P. C. Raiciu Sebastien Barre, Adam Greenhalgh, Damon Wischik, and Mark Handley, "Improving Data center Performance and Robustness with Multipath TCP," presented at the ACM SIGCOMM 2011, Toronto, Canada, 2011.

[43] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 267-280.

[44] L. Peng, C.-H. Youn, W. Tang, and C. Qiao, "A novel approach to optical switching for intradatacenter networking," *Lightwave Technology, Journal of,* vol. 30, pp. 252-266, 2012.

[45] C. Li, N. Deng, M. Li, Q. Xue, and P. Wai, "Performance analysis and experimental demonstration of a novel network architecture using optical burst rings for interpod communications in data centers," *Selected Topics in Quantum Electronics, IEEE Journal of,* vol. 19, pp. 3700508-3700508, 2013.

[46] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," *ACM SIGCOMM Computer Communication Review,* vol. 40, pp. 92-99, 2010.

[47] A. Greenberg et al., "VL2: A Scalable and Flexible Data Center Network," *Communications of the ACM,* vol. 54, March 2011.

[48] M. Y. Sowailem, D. V. Plant, and O. Liboiron-Ladouceur, "Implementation of optical burst switching in data centers," *IEEE PHO, WG4,* 2011.

[49] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, *et al.*, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *ACM SIGCOMM Computer Communication Review,* vol. 41, pp. 339-350, 2011.

[50] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, 2010, p. 8.

[51] J. Sullivan, N. Charbonneau, and V. M. Vokkarane, "Performance Evaluation of TCP over Optical Burst Switched (OBS) Networks Using Coordinated Burst Cloning and Forward-Segment Redundancy," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, 2010, pp. 1-6.

[52] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, "Architectural guidelines for multipath TCP development," *RFC6182 (March 2011), www. ietf. ort/rfc/6182,* 2011.

[53] C. Raiciu, M. Handley, and D. Wischik, "Coupled congestion control for multipath transport protocols," *Engineering Task Force- RFC 6356 draft-ietf-mptcp-congestion-01 (work in progress),* 2011.

[54] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "TCP extensions for multipath operation with multiple addresses," *IETF MPTCP RFC 6824 l-http://tools. ietf. org/id/draft-ford-mptcp-multiaddressed-03. txt,* 2011.

[55] Q. Zhang, V. M. Vokkarane, Y. Wang, and J. P. Jue, "Analysis of TCP over optical burst-switched networks with burst retransmission," in *Global Telecommunications Conference, 2005. GLOBECOM'05. IEEE*, 2005, pp. 6 pp.-1983.

[56] A. Lazzez, N. Boudriga, and M. S. Obaidat, "Improving TCP QoS over OBS networks: A scheme based on optical segment retransmission," in *Performance Evaluation of Computer and Telecommunication Systems, 2008. SPECTS 2008. International Symposium on*, 2008, pp. 233-240.

[57] B. Shihada, P.-H. Ho, and Q. Zhang, "A novel congestion detection scheme in TCP over OBS networks," *Journal of Lightwave Technology,* vol. 27, pp. 386-395, 2009.

[58] S. Gowda, R. K. Shenai, K. M. Sivalingam, and H. C. Cankaya, "Performance evaluation of TCP over optical burst-switched (OBS) WDM networks," in *Communications, 2003. ICC'03. IEEE International Conference on*, 2003, pp. 1433-1437.

[59] P. Prakash, A. Dixit, Y. C. Hu, and R. Kompella, "The TCP outcast problem: Exposing unfairness in data center networks," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, 2012, pp. 30-30.

[60] P. Rygielski and S. Kounev, "Network Virtualization for QoS-Aware Resource Management in Cloud Data Centers: A Survey," *Praxis der Informationsverarbeitung und Kommunikation,* vol. 36, pp. 55-64, 2013.

[61] A. Ghosh, S. Ha, E. Crabbe, M. Chiang, and J. Rexford, "Scalable Multi-Class Traffic Management in Data Center Backbone Networks," *under submission*.

[62] S. Ranjan, J. Rolia, H. Fu, and E. Knightly, "Qos-driven server migration for internet data centers," in *Quality of Service, 2002. Tenth IEEE International Workshop on*, 2002, pp. 3-12.

[63] Y. Song, H. Wang, Y. Li, B. Feng, and Y. Sun, "Multi-tiered on-demand resource scheduling for VM-based data center," in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2009, pp. 148-155.

[64] Y. Chen, M. Hamdi, and D. H. Tsang, "Proportional QoS over OBS networks," in *Global Telecommunications Conference, 2001. GLOBECOM'01. IEEE*, 2001, pp. 1510-1514.

[65] Q. Zhang, V. M. Vokkarane, B. Chen, and J. P. Jue, "Early drop scheme for providing absolute QoS differentiation in optical burst-switched networks," in *High Performance Switching and Routing, 2003, HPSR. Workshop on*, 2003, pp. 153-157.

[66] M. Yoo, C. Qiao, and S. Dixit, "QoS performance of optical burst switching in IP-over-WDM networks," *Selected Areas in Communications, IEEE Journal on,* vol. 18, pp. 2062-2071, 2000.

[67] M. Channegowda, R. Nejabati, and D. Simeonidou, "Software-defined optical networks technology and infrastructure: enabling software-defined optical network operations [Invited]," *Optical Communications and Networking, IEEE/OSA Journal of,* vol. 5, pp. A274-A282, 2013.

[68] D. Li, Y. Shang, and C. Chen, "Software Defined Green Data Center Network with Exclusive Routing."

[69] N. McKeown, "Software-defined networking," *INFOCOM keynote talk,* 2009.

[70] C. Monsanto, J. Reich, N. Foster, J. Rexford, and D. Walker, "Composing Software Defined Networks," in *NSDI*, 2013, pp. 1-13.

[71] S. Sezer, S. Scott-Hayward, P.-K. Chouhan, B. Fraser, D. Lake, J. Finnegan*, et al.*, "Are we ready for SDN? Implementation challenges for software-defined networks," *Communications Magazine, IEEE,* vol. 51, 2013.

[72] A. Tootoonchian, S. Gorbunov, Y. Ganjali, M. Casado, and R. Sherwood, "On controller performance in software-defined networks," in *USENIX Workshop on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services (Hot-ICE)*, 2012.

[73] C.-Y. Hong, M. Caesar, and P. Godfrey, "Finishing flows quickly with preemptive scheduling," *ACM SIGCOMM Computer Communication Review,* vol. 42, pp. 127-138, 2012.

[74] S. Liu, H. Xu, and Z. Cai, "Low Latency Datacenter Networking: A Short Survey," *arXiv preprint arXiv:1312.3455,* 2013.

[75] D. Zats, T. Das, P. Mohan, D. Borthakur, and R. Katz, "DeTail: reducing the flow completion time tail in datacenter networks," *ACM SIGCOMM Computer Communication Review,* vol. 42, pp. 139-150, 2012.

[76] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron, "Better never than late: Meeting deadlines in datacenter networks," in *ACM SIGCOMM Computer Communication Review*, 2011, pp. 50-61.

[77] S. T. a. M. Bassiouni, "Performance Evaluation of MPTCP over Optical Burst Switching in Data Centers," presented at the ITS Brazil 2014, São Paulo, Brazil, 2014.

[78] M. El Houmaidi, M. Bassiouni, and G. Li, "Dominating set algorithms for sparse placement of full and limited wavelength converters in WDM optical networks," *Journal of Optical Networking,* vol. 2, pp. 162-177, 2003.

[79] J. A. White, R. S. Tucker, and K. Long, "Merit-based scheduling algorithm for optical burst switching," in *Proceedings of the international conference on optical Internet*, 2002, pp. 75-77.

[80] M. El Houmaidi, G. Li, and M. A. Bassiouni, "Architecture and sparse placement of limited-wavelength converters for optical networks," *Optical Engineering,* vol. 43, pp. 137-147, 2004.

[81] I. Ogushi, S. i. Arakawa, M. Murata, and K.-i. Kitayama, "Parallel reservation protocols for achieving fairness in optical burst switching," in *High Performance Switching and Routing, 2001 IEEE Workshop on*, 2001, pp. 213-217.

[82] S. Tariq, M. Bassiouni, and G. Li, "Improving fairness of OBS routing protocols in multimode fiber networks," in *Proceedings of the 2013 International Conference on Computing, Networking and Communications (ICNC)*, 2013, pp. 1146-1150.

[83] R. Ho, H. Schwetman, M. O. McCracken, P. Koka, J. Lexau, J. Cunningham*, et al.*, "Optical systems for data centers," in *Optical Fiber Communication Conference*, 2011, p. OTuH1.

[84] K. J. Barker, A. Benner, R. Hoare, A. Hoisie, A. K. Jones, D. K. Kerbyson*, et al.*, "On the feasibility of optical circuit switching for high performance computing systems," in *Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, 2005, p. 16.

[85] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. Ng, M. Kozuch, *et al.*, "c-Through: Part-time optics in data centers," in *ACM SIGCOMM Computer Communication Review*, 2010, pp. 327-338.

[86] M. Glick, "Optical switching and routing for the data center," in *Photonics Society Winter Topicals Meeting Series (WTM)*, 2010.

[87] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, "The emerging optical data center," in *Optical Fiber Communication Conference*, 2011, p. OTuH2.

[88] L. Schares, D. M. Kuchta, and A. F. Benner, "Optics in Future Data Center Networks," in *Hot Interconnects*, 2010, pp. 104-108.

[89] M. Fiorani, M. Casoni, and S. Aleksic, "Large data center interconnects employing hybrid optical switching," in *Network and Optical Communications (NOC), 2013 18th European Conference on and Optical Cabling and Infrastructure (OC&i), 2013 8th Conference on*, 2013, pp. 61-68.

[90] M. Alizadeh, A. Kabbani, T. Edsall, B. Prabhakar, A. Vahdat, and M. Yasuda, "Less is more: trading a little bandwidth for ultra-low latency in the data center," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, 2012, pp. 19-19.

[91] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "A performance analysis of EC2 cloud computing services for scientific computing," in *Cloud Computing*, ed: Springer, 2010, pp. 115-131.

[92] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, *et al.*, "Integrating microsecond circuit switching into the data center," in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, 2013, pp. 447-458.

[93] N. Farrington, A. Forencich, P.-C. Sun, S. Fainman, J. Ford, A. Vahdat, *et al.*, "A 10 us Hybrid Optical-Circuit/Electrical-Packet Network for Datacenters," in *Optical Fiber Communication Conference*, 2013, p. OW3H. 3.

[94] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, *et al.*, "Data center tcp (dctcp)," *ACM SIGCOMM Computer Communication Review,* vol. 40, pp. 63-74, 2010.

[95] R. B. Lee, D. F. Geraghty, M. Verdiell, M. Ziari, A. Mathur, and K. J. Vahala, "Cascaded wavelength conversion by four-wave mixing in a strained semiconductor optical amplifier at 10 Gb/s," *Photonics Technology Letters, IEEE,* vol. 9, pp. 752-754, 1997.

[96] S. B. Yoo, "Wavelength conversion technologies for WDM network applications," *Lightwave Technology, Journal of,* vol. 14, pp. 955-966, 1996.

[97] X. Gao, M. A. Bassiouni, and G. Li, "Addressing conversion cascading constraint in OBS networks through proactive routing," *Photonic Network Communications,* vol. 18, pp. 90-104, 2009.

[98] X. Gao, M. A. Bassiouni, and G. Li, "Conversion cascading constraint-aware adaptive routing for WDM optical networks," *Journal of Optical Networking,* vol. 6, pp. 278-294, 2007.

[99] X. Gao, G. Li, and M. A. Bassiouni, "Effective preemptive scheduling scheme for optical burst-switched networks with cascaded wavelength conversion consideration," *Optical Engineering,* vol. 49, pp. 035004-035004-10, 2010.

[100] E. Exposito, M. Gineste, L. Dairaine, and C. Chassot, "Building self-optimized communication systems based on applicative cross-layer information," *Computer Standards & Interfaces,* vol. 31, pp. 354-361, 2009.

[101] C. Diop, G. Dugué, C. Chassot, E. Exposito, and J. Gomez, "QoS-aware and autonomic-oriented multi-path TCP extensions for mobile and multimedia applications," *International Journal of Pervasive Computing and Communications,* vol. 8, pp. 306-328, 2012.

[102] C. Diop, G. Dugué, C. Chassot, and E. Exposito, "QoS-aware multipath-TCP extensions for mobile and multimedia applications," in *Proceedings of the 9th International Conference on Advances in Mobile Computing and Multimedia*, 2011, pp. 139-146.

[103] M. Pustisek, I. Humar, and J. Bester, "Empirical analysis and modeling of peer-to-peer traffic flows," in *Electrotechnical Conference, 2008. MELECON 2008. The 14th IEEE Mediterranean*, 2008, pp. 169-175.

[104] J. J. Rodrigues, M. M. Freire, N. M. Garcia, and P. M. Monteiro, "Enhanced just-in-time: a new resource reservation protocol for optical burst switching networks," in *Computers and Communications, 2007. ISCC 2007. 12th IEEE Symposium on*, 2007, pp. 121-126.

[105] S. Gringeri, N. Bitar, and T. J. Xia, "Extending software defined network principles to include optical transport," *Communications Magazine, IEEE,* vol. 51, pp. 32-40, 2013.

[106] N. Cvijetic, "OFDM for next-generation optical access networks," *Lightwave Technology, Journal of,* vol. 30, pp. 384-398, 2012.

[107] H. Li, D. Groep, and L. Wolters, "Workload characteristics of a multi-cluster supercomputer," in *Job Scheduling Strategies for Parallel Processing*, 2005, pp. 176-193.

[108] H. Li, L. Wolters, and D. Groep, "Workload Characteristics of the DAS-2 Supercomputer."

[109] C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley, "Improving datacenter performance and robustness with multipath tcp," in *ACM SIGCOMM Computer Communication Review*, 2011, pp. 266-277.

[110] M. Taubenblatt, "Space division multiplexing in data communications and high performance computing," presented at the IEEE Summer Topical 2012.