Performance Analysis of High-speed Optical Packet Switching in High Performance Computing and Datacentre Networks

Jingyan Wang

B.Eng., M.Sc.

A dissertation submitted in fulfilment of the
requirements for the award of
Doctor of Philosophy (Ph.D.)

School of Electronic Engineering

Faculty of Engineering and Computing

Dublin City University

Supervisors: Prof. Liam P. Barry, Dr. Conor J. McArdle

June 2016

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No.:

Date:

Acknowledgements

I would like to thank my supervisor Prof. Liam P. Barry for giving me the opportunity to join his research group and for his expert, sincere and valuable guidance and constant support throughout this PhD. My deepest appreciation and gratitude goes to my co-supervisor Dr. Conor McArdle under whose inspiration and guidance this work has been completed. I sincerely acknowledge his guidance, encouragement, suggestions, understanding and very constructive criticism throughout my research.

My appreciation also extends to my colleagues past and present in the Radio and Optical Communications Laboratory, for the enormous help and support they have provided throughout my time in DCU.

Most of all, I am fully indebted to all my friends and family for their unceasing support and encouragement. I would like to thank my parents and my sister for their endless love and support. Thank you for always believing in me.

Contents

List of Acronyms

AWG	Arrayed Waveguide Grating
B&S	Broadcast-and-Select
CAGR	Compound Annual Growth Rate
CWDM	Coarse Wavelength Division Multiplexing
DEMUX	Demultiplexer
DWDM	Dense Wavelength Division Multiplexing
E/O	Electro-Optic
FDL	Fibre Delay Line
FWC	Fixed Wavelength Converter
HPC	High Performance Computing
IID	Independent and Identically Distributed
IPP	Interrupted Poisson Process
LAUC-VF	Latest Available Unused Channel with Void Filling
MEMS	Micro-Electro-Mechanical System
MMPP	Markov-modulated Poisson Process
MUX	Multiplexer
NAK	Negative Acknowledgement
O/E	Optic-Electro
O/E/O	Optical-Electrical-Optical
OCS	Optical Circuit Switching
OLE	Optical Label Extractor
OPS	Optical Packet Switching
OXC	Optical Cross-Connect
PreRes	Pre-Reservation

QoS	Quality-of-Service
RAM	Random Access Memory
RTT	Round-Trip Time
SOA	Semiconductor Optical Amplifier
ToR	Top-of-Rack
TWC	Tunable Wavelength Converter
WC	Wavelength Converter
WDM	Wavelength Division Multiplexing
WS	Wavelength Selector
WSS	Wavelength Selective Switch

List of Tables

List of Figures

Performance Analysis of High-speed Optical Packet Switching in High Performance Computing and Datacentre Networks

Jingyan Wang

Abstract

The massive growth in datacentre traffic, due to a huge increase in the deployment of data-intensive applications, is forcing datacentre infrastructure to migrate away from conventional electronic packet-switched networks, where capacity scaling imposes significant financial and technical constraints, and to evolve towards more advanced architectures. Motivated by this, new optical switching technologies and networking architectures, capable of providing very large bandwidth capacity, high scalability, high switching speed and high energy efficiency, are being targeted for building next-generation high-performance datacentre and High Performance Computing (HPC) networks. Optical packet switching technology is considered to be a long-term solution to meet these design challenges, as it can exploit fully the enormous potential capacity enabled by optics and support high switching flexibility at packet level.

In this thesis, new wavelength-routed optical packet switching architectures, which exploit the functionalities of key optical switching components such as Tunable Wavelength Converters (TWCs), Arrayed Waveguide Gratings (AWGs) and Wavelength Selective Switches (WSSs), to realise all-optical switching, are proposed for use as next-generation datacentre and HPC networks. To maximise the efficiency of the proposed optical switching architecture, a dynamic bandwidth provisioning algorithm, which allocates switch resources to traffic demands based on application requirements, is developed to further enhance network flexibility, resource utilisation and network performance. Moreover, based on the proposed switch architectures, a large-scale high-performance datacentre network with flexible central control is modelled, with a view of determining the optimal network topology and traffic scheduling methods. This flexible network architecture employs a modular design and combines transparent optical packet switches, based on Arrayed Waveguide Grating (AWG) routers, and a hybrid congestion control scheme using recirculating Fibre Delay Lines (FDLs) along with novel packet retransmission schemes. The work carried out in this thesis indicates that the proposed network structure not only provides high scalability, capable of hosting hundreds of thousands of severs, but also delivers high bandwidth utilisation and network provisioning flexibility. The network offers a promising and viable networking solution to address current and future application needs in datacentre and HPC environments.

Introduction

Cloud computing applications and data-intensive Internet services are motivating the installation of warehouse-scale datacentres - huge facilities with hundreds of thousands of servers that are required to exchange very large volumes of data at high speeds and low latencies. Coinciding with the explosion of data traffic volumes exchanged between servers within datacentres is an escalation of the required communication capacity which presents significant technological challenges for the datacentre networking infrastructure. Current network architectures follow a three-tier structure comprising access (L1), aggregation (L2) and core (L3) layers. The basic network elements are switches and routers which process data streams electronically and then convert them to optical form for Wavelength Division Multiplexing (WDM) transmission on high-capacity point-to-point optical links between the switches and routers. A problem associated with this type of architecture is the capacity scaling issue, stemming from the multi-layer nature of the network topology. More precisely, the aggregated traffic demands at higher levels of the system hierarchy potentially require progressively larger bandwidth capacities, and consequently put the statically-provisioned and oversubscribed upper tiers under strain. Additionally, as datacentres continue to expand, the traffic demands will increase massively. To keep up with this rapid growth, the electronic switches and routers need to scale with regard to data rates, connectivity and bandwidth capacity, which is technically challenging and requires excessive amounts of power. Further, the installation of the required Optical-Electrical-Optical (O/E/O) converters for WDM transmission dramatically increases the investment in equipment and equipment power consumption. These drawbacks are stressing the existing datacentre networks and motivating the network to evolve towards advanced all-optical switching and routing.

Optical switching is a strong candidate for building next-generation large-scale networks, as it allows the exploitation of the unprecedented potential provided by WDM technology in terms of super-high bandwidth, fast switching speed, low latency, wavelength-based switching/routing, data

rate/format transparency and low power consumption per bit transmitted. Depending on the switching granularity, optically switched solutions can be classified into two categories: Optical Circuit Switching (OCS) and Optical Packet Switching (OPS). OCS is usually built based on Micro-Electro-Mechanical System (MEMS), which promises high bandwidth capacity, high switch port count and low power consumption. However, a MEMS-based OCS exhibits switching speeds in the order of milliseconds and is therefore unable to handle fast changing bursty traffic in high performance computing and datacentre environments. Alternatively, OCS can be used to setup semi-permanent dedicated lightpaths between the source and destination, maintained for the entire transmission session, which ensures the data traffic traverses the network in optical format with guaranteed bandwidth. A drawback with this technique is the underutilisation of the available bandwidth on average, as the established paths cannot be used for other communications even when their links are idle. These restrictions of OCS with respect to switching speed and bandwidth utilisation make it challenging to fulfill the performance requirements of the applications, especially when the amount of data traffic involved is so massive and with unpredictable and dynamic traffic patterns. For that reason, a flexible switching paradigm, which supports all-optical switching at packet granularity, is preferable, and this can be provided by OPS. Optical packet switching has the potential of delivering on-demand bandwidth and fine-grained granularity with fast switching speed, making it a suitable substitution for large electronic packet switches and a solution for future interconnection networks.

Optical packet switch design is currently being greatly facilitated by the increased availability of a diverse range of active and passive optical components, including the optical Multiplexer (MUX) and Demultiplexer (DEMUX), the Tunable Wavelength Converter (TWC), the Wavelength Selective Switch (WSS) and the Arrayed Waveguide Grating (AWG) which is a low-loss passive optical switching device where the reconfiguration is performed by tuning the input signal on the appropriate wavelength using fast TWCs. Progress in advanced optical components and switching technologies offers the possibility of constructing a high-speed, high-bandwidth and energy-efficient optical packet switch (OPS), with simple routing control, suitable for interconnecting many servers.

Despite these advantages, the issue of packet contention resolution in OPS is problematic, due to the lack of viable optical Random Access Memory (RAM). As such, congestion management significantly impacts application performance, and eventually determines the future role of OPS in High Performance Computing (HPC) and datacentre networks, thus the research and development of contention avoidance solutions are extremely important. In response to this, optical buffers emu-

lated with Fibre Delay Lines (FDLs) have been proposed to delay the contending packets for a fixed time, depending on fibre length. Other techniques to address the need for packet buffering include adding electronic buffers, deflection routing schemes and wavelength conversions. Alternatively, data recovery mechanisms have also been exploited to eliminate the traffic loss in optical network. Further, additional packet scheduling/routing flexibility in the network could be a tremendous benefit, where network capacities are dynamically allocated to meet the dynamically changing traffic patterns, thus alleviating packet contention conditions. Moreover, hybrid network design (optical and electronic switching) is another effective approach, as it exploits the advantages of different switch paradigms and improves the overall effectiveness of the network, which in turn reduces the required buffering capacity. This thesis focuses on the development of scalable, high-performance OPS networks for datacentres and HPC with efficient contention solutions.

Main Thesis Contributions

The main contributions of this thesis are:

- A novel high-port-count optical packet-switched architecture, where AWG and TWCs are employed to support high-speed switching, is developed for HPC and datacentres. To resolve packet contention, the network is configured with share-per-node optical buffers. Extensive simulations are performed to analyse the impact of various FDL parameters on network performance. With proper dimensioning of optical buffers, significant performance improvements are obtained. In addition, to realise a non-blocking switching scheme, the inclusion of small-scale electronic buffering functionality is investigated, with the aim of minimising the number of the switch components required. A follow-up investigation on power consumption and packet scheduling computation complexity further confirms the feasibility of the AWG switch in large-scale networks.
- Simulation-based dimensioning turns out to be complex and computationally-intensive, thus it is essential to provide an analytical performance model, which accurately captures the behaviours of the presented AWG switch with optical buffers and accelerates performance evaluation studies. A mathematical queuing model is developed, which allows a detailed analysis of the overall performance against various network design parameters. Complemented with a heuristic procedure, the model is extended to resolve the dimensioning problem which is formulated as a constrained optimisation problem with the objective of allocating the appro-

priate network parameter configuration which potentially achieves the most power savings, simultaneously meeting network performance constraints in terms of packet loss probability and overall latency.

- An enhanced optical packet switch design, based on combining AWG and WSSs, is developed, which allows more flexible and dynamic end-to-end bandwidth re-provisioning. The prominent attribute of the WSS is the elastic switching granularity, which enables the dynamic allocation of capacity to traffic demands. This flexibility enhances the overall resource utilisation as well as the network's ability to adapt to changing communication patterns. The system performance is examined under realistic non-uniform traffic, and compared with that of a static bandwidth allocation strategy. The performance improvements suggest that the proposed scheme is an effective solution for more efficiently provisioning applications in HPC and datacentres.
- A scalable datacentre interconnection network of combined optical and electronic packet switches is proposed, together with a centralised dynamic reconfiguration algorithm. The network possesses a number of advantages. First, the network platform is highly scalable, and capable of interconnecting a large number of servers, due to the layered topology. Additionally, the dynamically reconfigurable OPS is implemented in conjunction with mature electronic switching technology, making possible the exploitation of the benefits of high-capacity optical switching and fine-grained electronic switching. Furthermore, the flexible global routing scheme, which supports dynamic network reconfigurations, significantly enhances the overall performance. Another important benefit is the resolution of traffic loss, which is accomplished through the deployment of a combined contention control scheme employing recirculating optical buffering (FDLs) and a novel packet retransmission mechanism.

Thesis Structure

The remainder of this thesis is structured as follows:

- Chapter 1 reports the forecasted growth of datacentre traffic and expected datacentre traffic characteristics, along with the networking requirements for datacentres arising from these trends. Fundamental networking structures and key enabling technologies used for building large-scale datacentre interconnection networks are discussed. The remainder of this chapter

provides an overview of recent developments in switching technologies, operation strategies and architectural designs for HPC and datacentre networks. Several recently proposed networking technologies and proposals are reviewed in detail. The benefits and design issues of these proposals are also discussed.

- Chapter 2 proposes a novel high-port count, high-speed optical packet switch which employs an Arrayed Waveguide Grating (AWG) component for cross-connection, a hybrid optical/ electronic buffering scheme and a method for efficiently integrating Fibre Delay Line (FDL) buffer capacity into the AWG switch. A detailed description of the simulation model of the proposed optical network is presented in this chapter. Extensive simulations have been carried out to investigate the impacts of various network parameters on network performance and power consumption. The overall goal is to resolve the network dimensioning problem where the objective is to optimally dimension the optical and electronic buffer resources in an energy-efficient manner. Further, the computational complexity of the scheduling algorithm in the hybrid-buffered switch is investigated.

- Chapter 3 is devoted to developing a Markov-modulated mathematical model for the proposed optical packet switch with recirculating Fibre Delay Line buffers. The analytical framework numerically models the behaviours of the optical switch, and allows a detailed performance analysis regarding contention probability and communication latency. The correctness and accuracy of the proposed analytical model is confirmed by the comparison with the simulation results under both smooth and bursty traffic types. Further, on the basis of the developed analytical model, the dimensioning of the switch's component parts is resolved as an optimisation problem with the objective of addressing the performance requirements and minimising the deployment and power costs. A heuristic optimisation algorithm is implemented to resolve the constrained optimisation problem.

- Chapter 4 introduces a flexible optical packet-switched architecture for HPC systems and datacentre networks using passive optical components of the AWG and Wavelength Selective Switches (WSSs). A novel aspect of the architecture is the use of a flexible resource allocation algorithm, which is developed in this chapter. To facilitate the flexible wavelength assignment in FDLs, an analytical model of the optical switch which accurately approximates the overflow traffic loads from optical switch to FDL buffers, is developed. A non-uniform, realistic traffic model is presented, and the network performance and energy consumption of the flexible optical network is investigated under both uniform and non-uniform traffic

patterns. The network dimensioning problem is also resolved with the aim of satisfying the performance constraints and minimising the power consumption.

- Chapter 5 introduces the network architecture of the proposed large-scale hybrid electronic/optical datacentre interconnection network, followed by a detailed description of the dynamic global traffic scheduling procedure. Further, a novel congestion control method, combining the recirculating optical buffers and the packet retransmission mechanism, is introduced. Through extensive simulations, the system performance of the proposed datacentre network is evaluated under both all-to-all traffic and a more realistic non-uniform datacentre workload, highlighting its benefits regarding network throughput, scalability and flexibility.
- Chapter 6 presents the concluding remarks of the thesis and proposes some potential future research topics related to the subject.

Chapter 1

Networking Technologies and Architectures for HPC and Datacentres

A datacentre is a massive warehouse-scale computer system with hundreds or thousands of servers [1]. The datacentres support a high-level computational capacity for running big data and information-intensive applications efficiently, quickly and reliably. High Performance Computing (HPC) is similar to datacentres, the difference being that HPC incorporates massively parallel processing such that the processors can work together to resolve an enormously complex computational problem. In both datacentres and HPC systems, it is of crucial importance to design an efficient interconnection network which supports high-speed interactions between servers. Various interconnects for datacentres and HPC systems have been proposed and prototyped in recent years, such as electronic networks, hybrid electronic/optical systems and all-optical schemes. This chapter provides a qualitative categorisation and comparison of some newly proposed networking technologies and designs for HPC and intra-datacentre networks. Fundamental networking structures and key enabling technologies used for building these interconnection networks are also reported. As motivation and background, the forecasted growth of datacentre traffic and expected datacentre traffic characteristics are first reviewed and the technological challenges arising from these trends are discussed.

1.1 Background

1.1.1 Datacentre Traffic Growth

The network platforms in HPC and datacentres are constantly scaling and evolving to sustain the unprecedented growth in application data traffic. Figure 1.1 illustrates the forecasted growth of global datacentre traffic, based on recent statistics from Cisco [2, 3]. The annual global datacentre traffic is projected to grow more than threefold, from 2.6 zettabytes [1] in 2012 to 8.6 zettabytes by 2018, with a 22 percent Compound Annual Growth Rate (CAGR). It is estimated that the forecasted traffic volume of 8.6 zettabytes in 2018 is equivalent to the amount of traffic generated by three hours of high-defination (HD) video streaming for the entire population (7.6 billion people) [4].

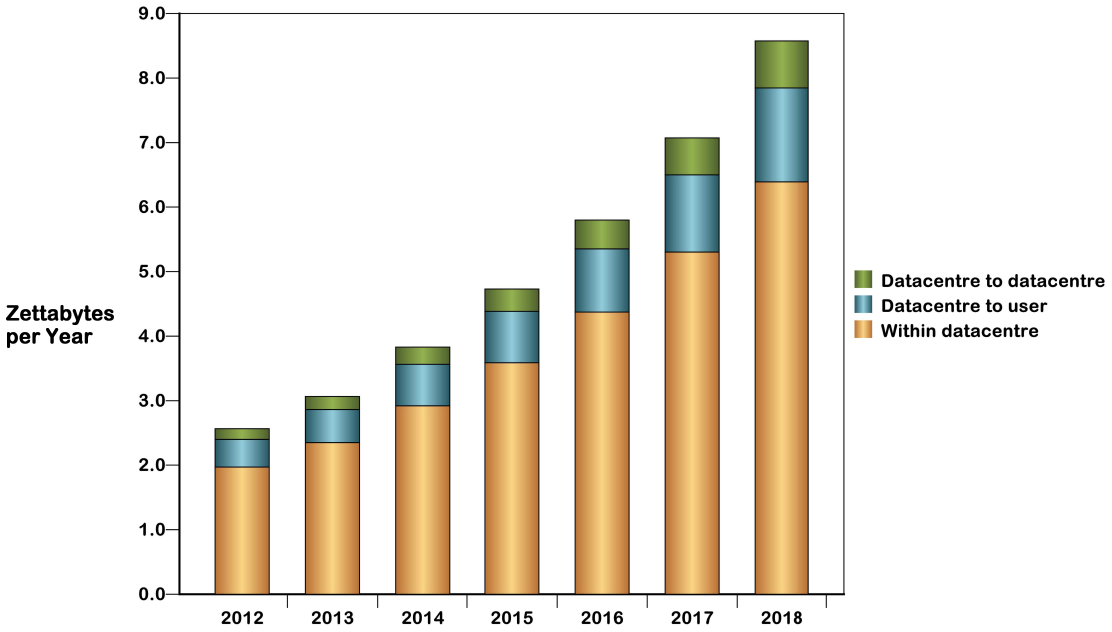


Figure 1.1: Current trends and future projections of overall datacentre traffic. Global datacentre traffic is estimated to grow at a CAGR of 22%, and intra-datacentre traffic will grow at a 21% CAGR.

Apparently, the main contributor to this growth is the traffic exchanged within datacentres, making up roughly 75 percent of the overall datacentre traffic for the period 2012-2018. By 2018, the intra-datacentre traffic is expected to reach 6.4 zettabytes annually from 2.0 zettabytes in 2012, which amounts to a CAGR of 21%. This explosive growth is largely due to the deployment of cloud technology, as illustrated by Figure 1.2, which depicts the predicated growth of cloud datacentre traffic [2, 3]. Cloud datacentre traffic is growing rapidly at a CAGR of ~32%, expected to reach

[1] A zettabyte is 10^{21} bytes

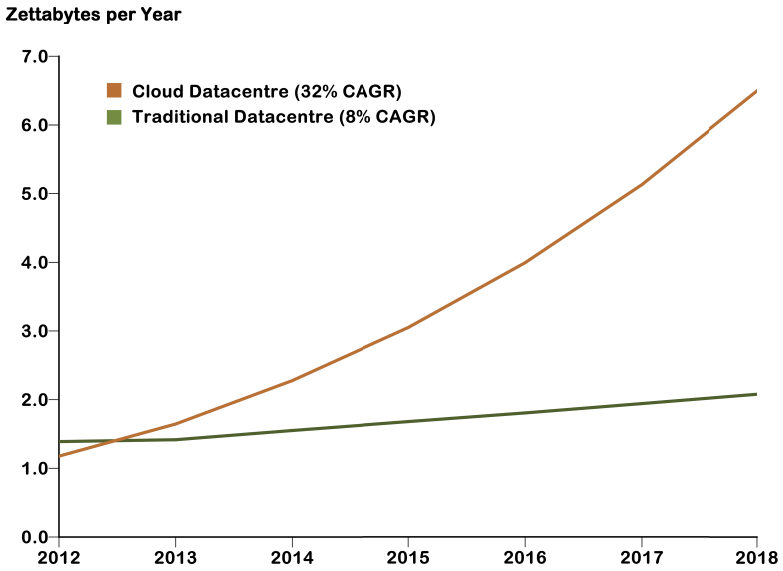


Figure 1.2: Cloud datacentre traffic is predicted to grow with a CAGR of 32%.

6.5 zettabytes by the end of 2018. Approximately 76 percent of data centre traffic will come from cloud computing applications by 2018, up significantly from 46 percent in 2012, thus cloud traffic is dominating the datacentre [4]. This continuing trend is changing the nature of datacentre traffic, as discussed in the next section.

1.1.2 Datacentre Traffic Characteristics

Intra-datacentre traffic is growing rapidly, mainly as a consequence of hosting a variety of cloud applications and services. This ever-increasing trend highlights the fact that the design of a highly reconfigurable network infrastructure which efficiently supports the heterogeneous traffic profiles of a wide range of applications, is of particular importance. Analysing the traffic characteristics facilitates the design of switch technology, network structure and resource management protocols. For that reason, significant research efforts [5–11] have been undertaken to explore the behaviours of datacentre applications. Even through the traffic demands evolve rapidly and constantly in the datacentre environment, and it is challenging to accurately model the communication patterns, these studies manage to statistically characterise various aspects of datacentre applications.

According to [6, 7], datacentre traffic is non-uniform and exhibits rack locality. More specifically, the majority of traffic generated by a rack, up to 80%, stays inside the rack, and only a small portion of traffic, referred to as inter-rack communication, traverses across the network. Analysis in [5, 7–9] suggests that there exists a number of hot-spots in network traffic, and as such the communication

matrix is extremely sparse and skewed, because an individual host tends to communicate with a small subset of remote racks at any time [5]. Further, [10] reveals that the traffic pattern exhibits low concurrency, that is, a server generates a relatively small number of concurrent connections compared to the network-wide scale. These findings provide important guidances for the design of datacentre networks.

Traffic characterisation is a crucial part of datacentre networking, as the dynamics in traffic patterns have significant impacts on service quality of datacentre applications. However, most recently proposed datacentre architectures are designed for the worst-cast uniform traffic pattern where traffic demands are uniformly distributed across all servers. As a consequence, these networks can exhibit low flexibility, low resource utilisation, and are unable to support high-demand hot-spot traffic patterns and adapt to changing traffic demands. Therefore, when designing a datacentre interconnection network, it is of importance to consider traffic variations and evaluate the network architecture under various types of communication patterns so as to ensure the robustness and reliability of the network architecture to future traffic trends in datacentres.

1.1.3 Requirements for Future Datacentre/HPC Networks

The continuing growth of datacentre traffic, coupled with its dynamic nature, imposes significant technical and operational challenges on current datacentre networks. To address these challenges, the network infrastructure will need to evolve towards more advanced network architectures, which provide larger bandwidth capacity, more network flexibility and faster communication speed, so as to adapt to new dynamic, data-intensive cloud-driven applications. Although much effort has been devoted to designing innovative electronic-switched interconnection networks, these architectures are based on electronic switching and routing, and will ultimately be constrained with respect to bandwidth capacity, network connectivity and power consumption. This has led to the emergence of optical switching technology in datacentres which exploits the ultra-high bandwidth capacity offered by optical fibres. Due to its high data-rate transmission, optical transparency, high capacity and flexibility and low energy consumption, optical switching technology has been considered as a long-term solution to meet the tremendous capacity needs highlighted in Section 1.1.1. Several different high-performance datacentre architectures, deploying photonic technologies have been developed in the research literature. These pioneering proposals have successfully demonstrated the enormous potential of optical switching technology in resolving the application needs in large-scale datacentres.

Apart from switching technologies, various switch topologies have been proposed for use in building large-scale networks. These fundamental network structures remain mostly the same, despite the evolution in the networking technologies used to implement the topologies. The next section explores the fundamental switching topology principles which are commonly used in datacentre networking.

1.2 Fundamental Network Theory

In datacentre and HPC networks, a large-scale network is required to interconnect hundreds of thousands of servers. Such a network can be constructed using a single-stage crossconnect with very high port count or a multi-stage network architecture consisting of multiple levels of smaller switch fabrics. The single-stage cross-connect provides direct connections between any input/output, but the device fabrication of a hundred-thousand port-count switch is technologically challenging and economically not feasible. An effective solution to this problem is through multi-stage switching which makes use of low-radix switch elements to build very large scalable cross-connects. The fundamental structures of multi-stage network designs include Crossbar, Fat-Tree, Clos and Spanke topologies. A multi-stage architecture is mainly characterised by three aspects: the number of basic switch modules required, the non-blocking pattern and the loss uniformity.

Number of Switch Modules: In a multi-stage architecture, the large-scale switch fabric is constructed by interconnecting several basic switching blocks in a multi-stage topology. The number of switch modules needed hinges on the implementation details of the network structure. This property has a significant impact on deployment costs, blocking characteristics and network control complexity, as discussed later.

Non-blocking Pattern: The blocking characteristic is an important factor in the design of large-scale switching networks. A switch fabric is either *blocking*, in which case some input/output connections cannot be realised [12], or *nonblocking*, in which case any input/output configuration is possible at any time. For most services and applications, nonblocking switch network is demanded in order to ensure communication quality. There are three types of non-blocking pattern: rearrangeable nonblocking, wide-sense nonblocking and strict-sense nonblocking.

- Rearrangeable nonblocking. A rearrangeable nonblocking switch architecture is able to interconnect any idle input/output pair by allowing the existing active connections to be interrupted. Although it ensures the nonblocking property, the connections in progress may be

disturbed. Problems of this type can significantly impair the application performance. Furthermore, a high level of control complexity is required, especially in large-scale switching networks, so as to execute an appropriate routing path for a new connection.

- Wide-sense nonblocking. In a wide-sense nonblocking network, any idle input/output connections can be provisioned without interfering with current active connections. However, this nonblocking path provisioning can only be realised with the implementation of an intelligent routing algorithm which considers not only current connections but also future connections. Compared to rearrangeable nonblocking, this scheme improves the network performance and requires a relatively simpler control complexity. Nonetheless, this imposes the requirement of installing more connection links and hardware devices.

- Strict-sense nonblocking. Strict-sense nonblocking switch can always provide nonblocking connectivity between any free input/output configurations, without rearrangement of current connections and implementation of complex routing algorithm. This simplicity is a result of deploying more connection links and a greater number of switch elements. Intuitively, there is a trade-off between network control complexity and the number of cascaded switch elements required.

Loss Uniformity:. As a multi-stage network is composed of multiple levels of cascaded switches, data traffic from an input port to an output port needs to traverse through multiple switching stages. It is possible that different input/output connecting paths contain different numbers of switching stages. Consequently, the quality of the transmitted signals depends on the chosen input/output paths. This path-dependent feature is undesirable in the networking of communication systems, as it potentially leads to performance variation in terms of signal losses and delays. Typically, loss uniformity is measured in terms of the shortest path length and the longest path length.

In what follows, the fundamental topologies which are commonly used in the design of large-scale switching networks, are presented and discussed. Table 1.1 shows a comparison of the main characteristics of these architectures.

Architectures	No. Switch Elements	Nonblocking Pattern	Loss Uniformity	
			Shortest Path	Longest Path
Crossbar	N^2	Wide sense	1	$2N-1$
Fat-Tree	$\frac{5}{4}\sqrt[3]{16N^2}$	Rearrangeable	3	5
Clos	$2\sqrt{2N}$	Rearrangeable	3	3
	$4\sqrt{N-1}$	Strict sense		
Spanke	$2N$	Strict sense	2	2

Table 1.1: Comparison of some fundamental network structures. N denotes the switch size.

1.2.1 Crossbar

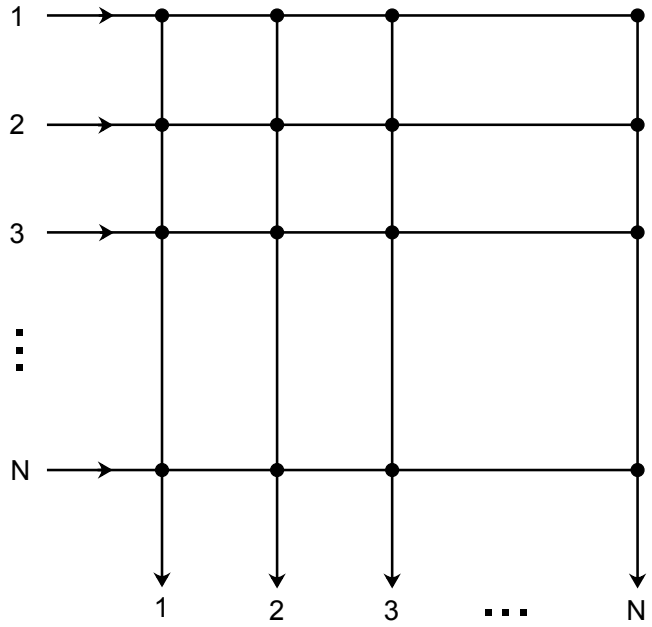
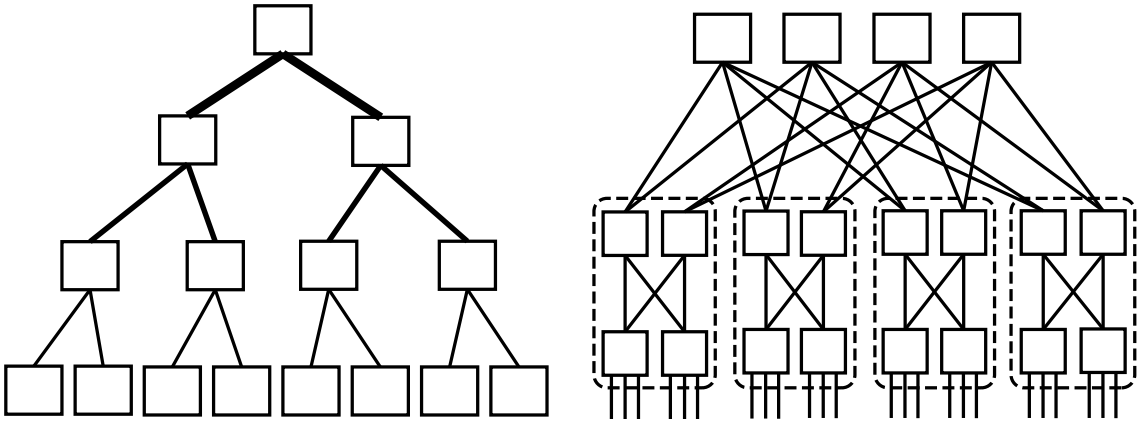


Figure 1.3: A $N \times N$ Crossbar switch

Crossbar is one of the simplest techniques to fabricate a large-scale switch. As illustrated in Figure 1.3, a $N \times N$ Crossbar architecture contains N inputs, N outputs and N^2 cross-point switching devices at the junctions. The interconnections between N inputs and N outputs are established by configuring the cross-points [13]. Note that the crossbar architecture supports wide-sense nonblocking connectivity. For an input/output pair (i, j), a connection path from input i to output j can be established by connecting row i to column j. Using this specific configuration algorithm, the switch architecture is strict-sense nonblocking. In the Crossbar switch, the shortest path has a length of 1, and the longest path includes $(2N - 1)$ hops [12]. Evidently, as switch size N scales, the worst-case path length increases linearly. This is an important issue, because it can lead to significantly unbalanced loss between different connections. Another challenge associated with this structure is the large number of cross-points required, N^2, which inevitably limits the scalability of the switch fabric, due to very high fabrication cost and reconfiguration complexity. Thus, the crossbar switch is often used as the basic building block in the design of multi-stage switch architectures.



(a) Single-root Fat-Tree (b) Multi-root Fat-Tree

Figure 1.4: Tree-based structures

1.2.2 Fat-Tree

A tree topology is a hierarchical structure which includes multiple layers of switch elements. Each switch element is configured to connect its branches to a single parent node, and as such the switch has multiple downlinks, but only one uplink [14]. Consequently, this network can suffer from severe capacity limitations at higher levels of the topology. A Fat-Tree is an enhanced version of the Tree structure, where the bandwidth capacity increases when moving up towards the root of the tree. The simplest Fat-Tree is single-root Fat-Tree architecture, as illustrated in Figure 1.4(a). Although this tree structure alleviates the capacity issue, it is not robust and still capacity constrained, as only one connection path exists between any input/output pair and the network performance relies completely on the single common root node. For that reason, the Fat-Tree has evolved from a single common root to multiple roots, see Figure 1.4(b). In a multi-rooted Tree, more switch elements and connection links are deployed in the higher levels of the tree, thereby achieving improved network reliability, greater scalability and higher bandwidth capacity.

The multi-rooted Fat-Tree is a rearrangeable nonblocking network provided that full bisection bandwidth [2] is supported at each network level [16]. Suppose that the switch module used has n ports, with $n/2$ ports connecting to $n/2$ upper-layer switches and $n/2$ ports connecting to lower-layer modules, the three-level tree network shown in Figure 1.4(b) can support up to $n^3/4$ port counts [14]. To construct a $N \times N$ switching network, n is computed to be $\sqrt[3]{4N}$, thus the total number of switch modules required is $\frac{5}{4}\sqrt[3]{16N^2}$. In this network, the longest path occurs when

[2] If a network of N nodes is partitioned into two subsets of $N/2$ nodes so that the available bandwidth between these two equal halves is minimum, then this worst-cast bandwidth is defined as the bisection bandwidth [15].

data streams need to travel from edge switches, up the network to the core and back down again, leading to a path length of 5. This methodology offers the possibility of building a high port-count nonblocking switch network. This is one of the major advantages of Fat-Tree and is the main reason behind its dominant use in today's datacentre networking. Nonetheless, there are trade-offs to be considered. The pros and cons of this solution are detailed in the next session.

1.2.3 Clos

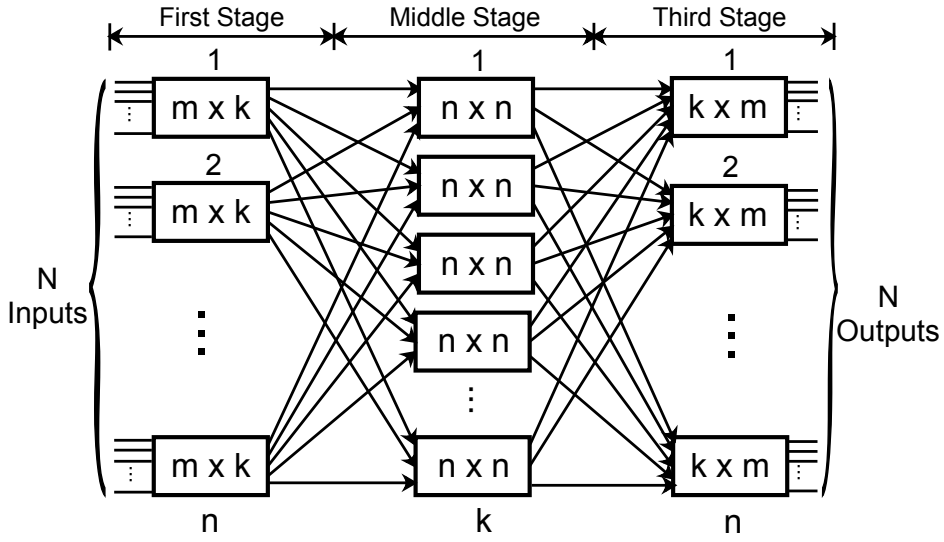


Figure 1.5: Three-stage Clos

The Clos architecture is a multi-stage networking model which was first proposed by Charles Clos in 1953 for the telephone switching network [17]. It has been an attractive structural model for many years, and now is a key networking topology in datacentres. Figure 1.5 shows a $N \times N$ symmetric three-stage Clos switch, which is composed of the ingress stage, the middle stage and the egress stage. The ingress stage consists of n $m \times k$ crossbar switches, the middle stage contains k $n \times n$ crossbar switches, and the egress stage includes n $k \times m$ crossbar switches, where $N = nm$. Every ingress switch is connected to each of the k middle-stage switches, which in turn is connected to all of the n egress switches. A three-stage Clos network can be folded into a two-stage folded-Clos architecture by combining the corresponding ingress and egress switches into one switch.

The Clos structure provides high scalability and loss uniformity with a path length of 3. In this type of switch, the non-blocking property is determined by the number of the crossbar switches used in the middle stage, k. More specifically, if $k \geq m$, the Clos switch is an rearrangeable nonblocking architecture which comprises $2n + m$ switch modules [18], where $m = N/n$. Apparently, the re-

quired number of switches is determined by parameter n. To minimise the total number of switches involved, n is calculated to be $n = \sqrt{\frac{N}{2}}$, and in this case the minimum amount of switch elements required in the network is $2\sqrt{2N}$. Alternatively, to form a strictly non-blocking switch matrix, the condition $k \geq 2m - 1$ has to be satisfied [12, 14], which means that at least $2m - 1$ middle-stage switches need to be installed. Consider $k = 2m - 1$, the total number of required crossbar switches is $2n + 2m - 1$. In such a case, the optimised network configuration is achieved when $n = \sqrt{N}$, and as a result, the minimum amount of hardware devices required is $4\sqrt{N} - 1$. This trade-off between the number of switches required and the nonblocking characteristics is presented in Table 1.1.

1.2.4 Spanke

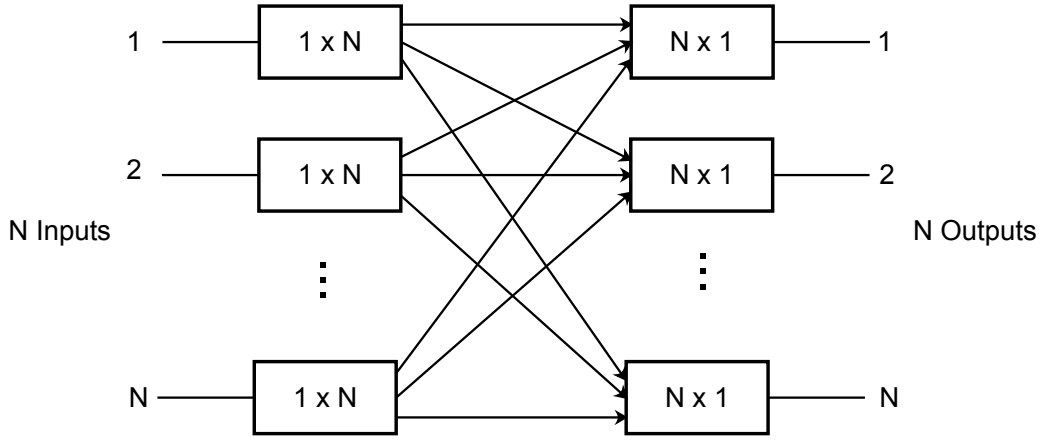


Figure 1.6: Spanke architecture

The Spanke architecture is another commonly used structural model for building large-scale interconnection networks [12, 18]. The general structure of a $N \times N$ Spanke switch is shown in Figure 1.6. It consists of N $1 \times N$ input switches and N $N \times 1$ output switches. Each of the N input switches is connected to all the output switches, thus the Spanke structure is a strict non-blocking switch fabric. In addition, this architecture utilises $2N$ switch elements in total, thereby the deployment cost exhibits a linear growth with the switch size N. Other attractive features of the Spanke architecture include its simplicity and the uniform loss property. However, in this architecture the port count of the basic switch element scales with switch size N, which consequently constrains the scalability of the switch architecture. This is one of the main drawbacks with this architecture. A typical application of the Spanke structure is the Broadcast-and-Select (B&S) architecture which will be discussed later.

1.2.5 Conclusions for Large-Scale Switch Architectures

Other fundamental networking structures include Omega, Baseline, Beneš, Spanke-Beneš and Butterfly [19]. The main problems for the use of these structures in building a high-radix switching network relate to the large number of stages required and the high implementation complexity. To alleviate these challenges, flattened multi-stage designs with simplified network connections have been suggested. A typical example of the flattened networks is Flattened-Butterfly [20] which significantly reduces the number of stages involved and reduces routing complexity. All these fundamental structures described are representatives of large-scale multi-stage switching architectures. Based on these structures, various application platforms have been proposed to support different network services. In the following sections, several recently proposed datacentre/HPC interconnection networks that exploit recent advances in switching technologies, switch control strategies and topology designs, are described in detail.

1.3 Electronic Switching in Datacentre Networks

Today's datacentres support the communications of hundreds of thousands of servers by leveraging commodity electronic packet switches (EPSs). These commodity electronic switches and routers are typically interconnected in a multi-layer hierarchical topology [16] so as to provide all-to-all connectivity. The design of multi-tier architectures is governed by three important network properties: oversubscription rate, scalability and Quality-of-Service (QoS). These features are discussed as follows.

Oversubscription– Oversubscription is described as the ratio between downstream bandwidth capacity and upstream provisioned bandwidth, which is calculated by the formula: Oversubscription Ratio = (downlink ports * bandwidth) / (uplink ports * bandwidth). Evidently, the oversubscription level depends on the network implementation. In a multi-tier datacentre network, due to the explosive growth of server-to-server traffic fuelled by the data-intensive applications requiring interactions among hundreds or thousands of servers, the bandwidth requirements become progressively higher when moving up to the top level of the hierarchy. This introduces difficulty as, in order to meet this resultant bandwidth requirements, a large number of links and switches will need to be provisioned at higher layers, leading to significant financial and technical challenges. Because of this, datacentre operators are forced to choose a compromised solution with the network being heavily oversubscribed.

Scalability– With the incremental deployment of cloud-based applications and on-demand services which demand high-performance computation power, today's datacentres are integrating more physical and virtual servers into networks [21, 22]. To keep up with this rapid expansion, the overlaying network infrastructure needs to scale out to interconnect hundreds of thousands of servers, and scale up regarding the bandwidth capacity so as to sustain the greatly increased communication demands in the network. To that end, the individual switching and routing components need to scale in terms of port intensity, switching capacity and power consumption. Scalability is an important feature that allows the expansion of datacentre framework to accommodate future application needs. However, in the existing hierarchical design, poor scalability is one of the major concerns.

QoS– Quality-of-Service (QoS) represents a collection of performance parameters involving bandwidth, throughput, latency and data loss [23]. It measures the quality of the communication services. In the multi-tier network model, the traffic stream from a source server may need to flow up through access, aggregation, and core layers, and then back down again to reach its destination server, which inevitably induces latency and bandwidth issues. Furthermore, if the network is already oversubscribed, the congestion condition would be exacerbated, and consequently degrading the network performance and system throughput. Thus, the multi-tier datacentres may not meet the QoS requirements of the applications.

1.3.1 Three-tier Datacentre Architecture

The structure of a typical three-tier tree-based datacentre network is illustrated in Figure 1.7. The bottom of the network is the access layer (L1), which consists of a collection of Top-of-Rack (ToR) switches. The ToR switch provides direct server-to-server connections within a rack which typically hosts 20 to 40 servers [24]. To support inter-rack communications, the ToR switches are attached to the aggregate switches at the aggregation layer (L2) via high-capacity uplinks. The aggregate switches collect the data streams from the ToR switches and forward them to the core switches at the top layer (L3). The high-capacity core switches form the backbone of the datacentre network, delivering mesh connectivity between aggregate switches. Redundancy is provided at L2 and L3, that is, a ToR switch is connected to multiple aggregate switches which in turn attach to multiple core switches. This over-provisioning of resources aims at enhancing network reliability and bandwidth flexibility and capacity [16].

By way of example, one of the most technologically advanced installed three-tier systems is Face-

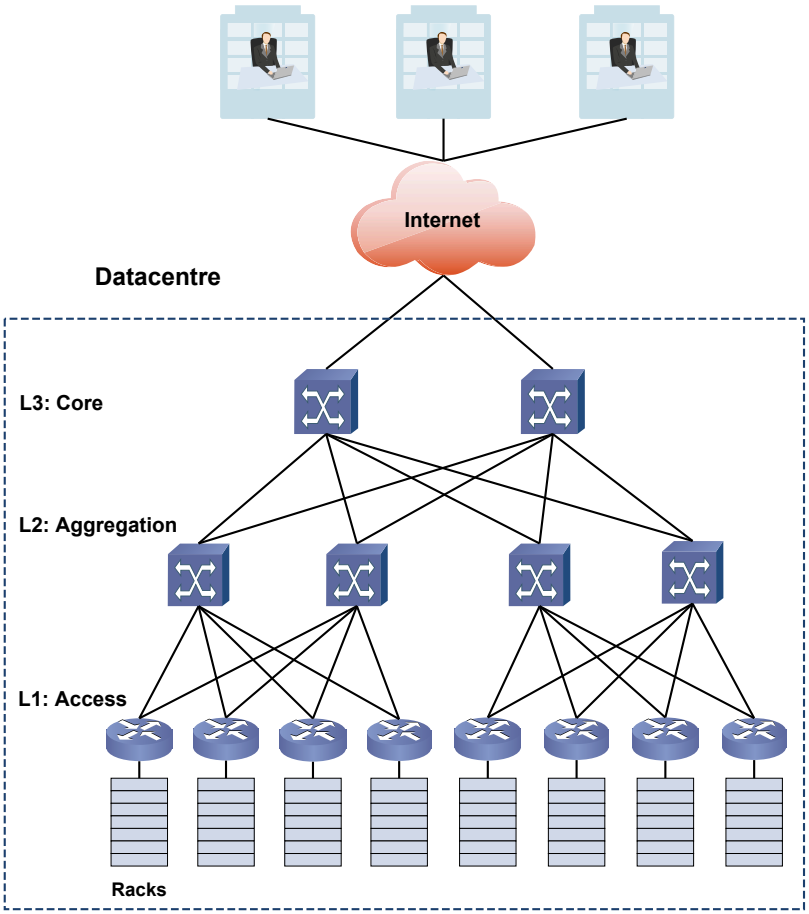


Figure 1.7: Three-tier datacentre network

book's Datacentre. Figure 1.8 illustrates the network architecture of current Facebook datacentre [25]. The network platform follows a hierarchical tree-like networking model composed of three layers of electronic switches and routers. The network infrastructure is constructed in a modular structure using clusters. A cluster contains a large number of ToR switches and four high-radix Cluster Electronic Switches (CESs). Each ToR switch is connected to the four high-radix cluster switches (CESs) via 4×10Gbps uplinks, forming a four-post networking fabric. The four-post fabric has the advantage of supporting uninterrupted services in the case of a link or switch failure, because multiple communication connections exist between ToR switches. Note that the four CESs are interconnected in a ring via 80Gbps links. Then, by using 4×40Gbps uplinks, the cluster switch (CES) is attached to all four Core Switches (CSWs) at the core layer which provides redundant interconnectivity between clusters. Also, the four CSW are interconnected in a ring via high-capacity, multi-wavelength channels up to 160Gbps.

According to Facebook [26], the current datacentre framework is reaching its limits. On one

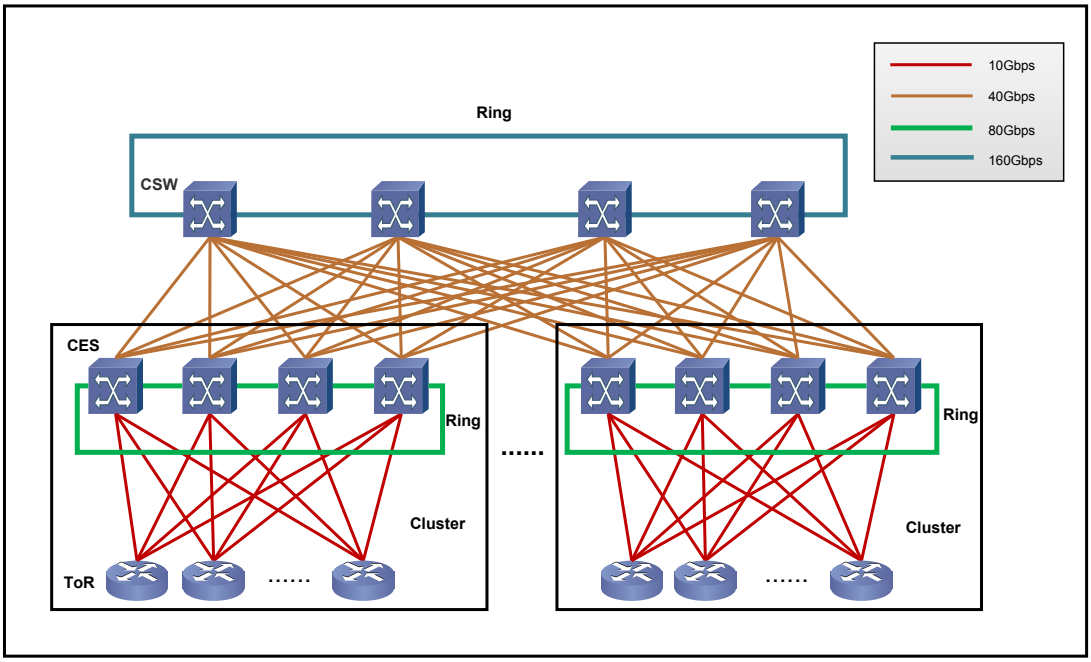


Figure 1.8: Current Facebook Datacentre Network. ToR: Top-of-Rack Switch. CES: Cluster Electronic Switch. CSW: Core Switch.

hand, it is technologically challenging to fabricate high-capacity, high-port count Cluster Electronic Switches (CESs). On the other hand, the Core Switch (CSW) at the core layer is heavily oversubscribed, and consequently unable to sustain the rapid growth of inter-cluster data traffic. The mentioned issues and challenges have motivated Facebook to upgrade its network infrastructure towards a fabric-based architecture, see Figure 1.9.

The next-generation Facebook datacentre still maintains a hierarchy in the interconnection network. Also, it employs a modular design, except that the modular unit, termed a "pod", is much smaller, containing only 48 ToR switches and four Fabric Switches (FSs). In a pod, to ensure reliability, the four-post architecture is preserved, that is, each ToR switch is connected to four FSs via 4×40Gbps uplinks. Each FS can reach down to 48 ToR switches at the access layer, and also connects to 48 Spine Switches (SPSs) in a Spine Plane at the core layer via 48×40Gbps uplinks, resulting in an ideal over-subscription of $1:1$. In this way, the four FSs in a pod are directly connected to four separate Spine Planes at the core layer. A prominent aspect of the new architecture is that the cluster size shrinks, and the core layer is composed of a network of low-radix Spine Switches (SPSs). This arrangement significantly alleviates the capacity and port intensity limitations in the aggregation and core layers, thus improving the network scalability, reliability and bandwidth capacity, but at the cost of high cabling and routing complexity.

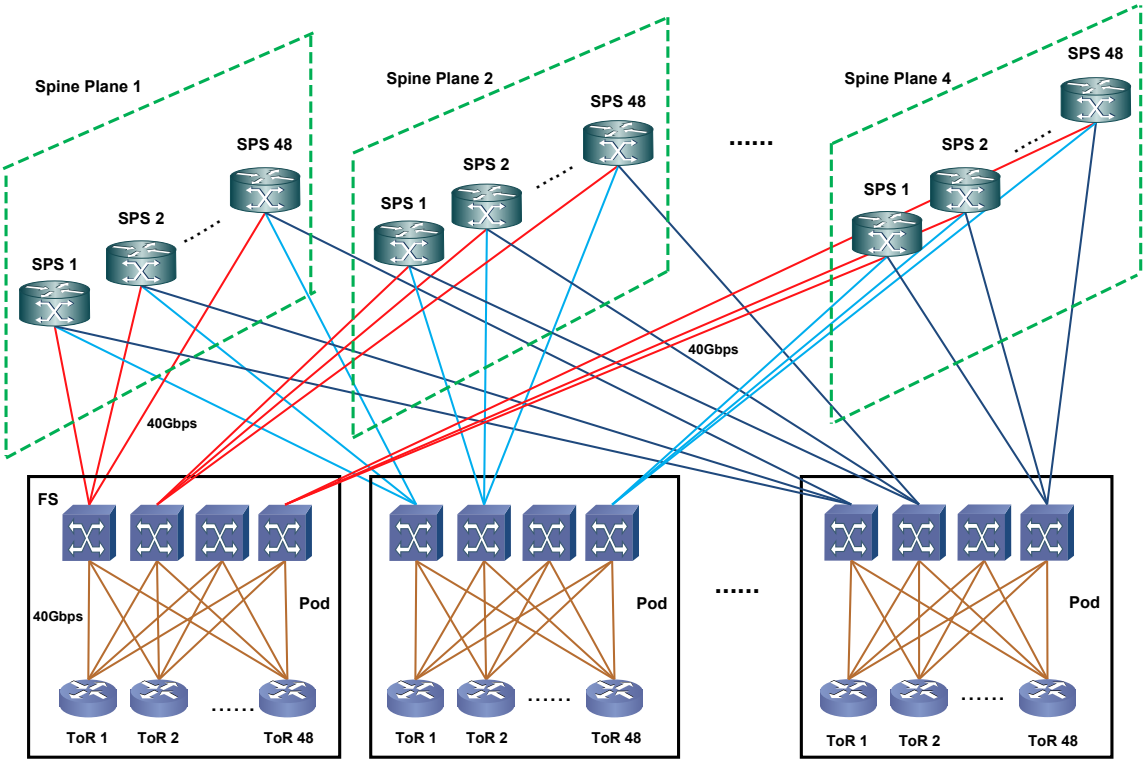


Figure 1.9: Next-generation Facebook Datacentre Network. ToR: Top-of-Rack Switch. FS: Fabric Switch. SPS: Spine Switch.

Obviously, Facebook's new networking model follows a three-tier leaf-spine architecture where the pod is viewed as a leaf, and the core switch is considered as the spine. Facebook's attempt to use leaf-spine architecture to flatten its network overlay indicates that current datacentre infrastructures are migrating away from the legacy three-tier network, and evolving towards two-level networking design.

1.3.2 Leaf-Spine Datacentre Architecture

In modern datacentres, a two-tier network platform consisting of leaf and spine layers, has emerged as an alternative to the traditional three-tier network topology [27]. The leaf-spine architecture is a folded-Clos network. Figure 1.10 demonstrates the basic network structure of a leaf-spine network. The leaf (access) layer comprises a network of leaf (ToR) switches which are fully meshed to a series of non-blocking spine switches at the spine layer, thereby realising any-to-any connectivity. At the spine layer, traffic is distributed among spine switches. The leaf-spine architecture is considered an improved version of the classic three-tier network structure by combining L2 and L3. It offers the attributes of simplified network cabling and operational management, decreased over-subscription

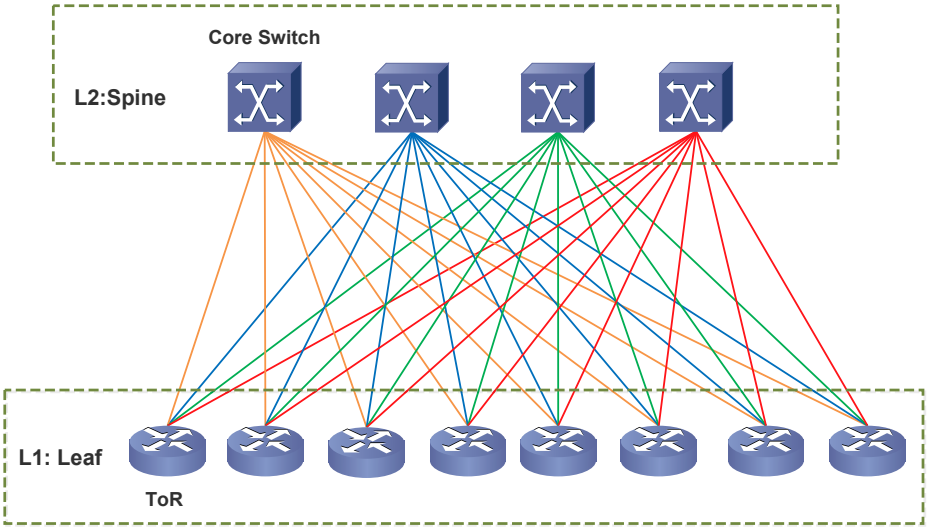


Figure 1.10: A leaf-spine network.

level and improved network performance. On the downside, the horizontal scaling of the leaf-spine structure is largely determined by the design of the high-port count spine switches. That is, the addition of leaf switches will require the scaling of spine switches. High-radix spine switches supporting 2048×10Gbps have been brought to market to facilitate the construction of leaf-spine networks [28]. Nevertheless, ultimately, the spine switch will run into a switch port limit due to high CAPEX/OPEX costs, and become unsustainable given predictions for future datacentre traffic growth [3]. Motivated by that, more advanced optical technologies are being investigated by vendors and organisations to facilitate the building of next-generation large-scale datacentres.

1.3.3 Other Electronically-Switched Datacentre Architectures

The previously discussed multi-tier networks are switch-centric architectures which integrate routing and switching intelligence in commodity electronic switches. In contrast to the switch-centric structure, a new networking concept - server-centric architecture, suggests placing the interconnection and routing operations into servers. Most of the server-centric interconnects, such as DCell [29], BCube [30] and FiConn [31, 32], follow a modular structure which is constructed by recursively interconnecting multi-port servers and low-radix electronic switches.

The scale-out of these modular datacentres can be realised by adding more ports to servers and using more electronic switches, which gives rise to a problem, as servers need to scale with the expansion of the datacentre infrastructure. In this type of network, the main drawback lies in the

fact that servers are not the proper devices to perform switching and routing operations. In addition, complex cabling structure, oversubscription, as well as Quality-of-Service (QoS) are also important design issues.

1.3.4 Discussion

Current, opaque network architectures relying on commodity electronic switches face significant challenges regarding, among other things, capacity scaling, network performance and heat dissipation. Undoubtedly, datacentre networks will continue to expand, and there will come a point at which the electronic switching technology reaches its limits and becomes unsustainable. This calls for investigating more advanced technologies to devise high-capacity, high-speed, power-efficient solutions for intra-datacentre networks. In response to this, all-optical interconnects have been envisioned as a promising candidate, which address the capacity, port intensity, speed and power consumption challenges by employing all-optical switching technologies. The evolution towards transparent optical networks largely depends on the functionalities of various optical components and enabling technologies being exploited to enable all-optical switching, which will be discussed in the following section.

1.4 Enabling Technologies: Optical Switch Components

This section briefly discusses the characteristics of optical components and enabling technologies for optical switching. The basic building elements that may be used in an optical datacentre interconnection network include the

- Micro-Electro-Mechanical System (MEMS) Optical Switch
- Optical Multiplexer/Demultiplexer (MUX/DEMUX)
- Optical Coupler/Splitter
- Semiconductor Optical Amplifier (SOA)
- Wavelength Converter (WC)
- Arrayed Waveguide Grating (AWG)
- Wavelength Selective Switch (WSS).

1.4.1 Micro-Electro-Mechanical System (MEMS)

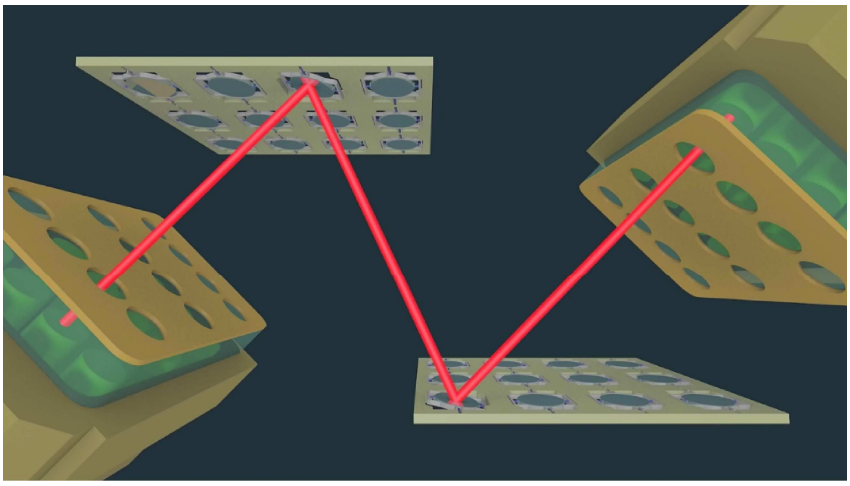


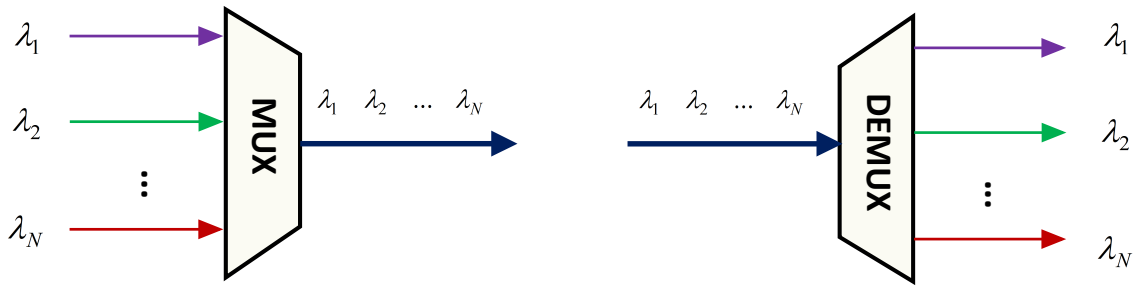
Figure 1.11: A three-dimensional optical MEMS switch.

The optical micro-electro-mechanical system (MEMS) is an optical component integrating mechanical, optical and electronic technologies. Generally, there are two classes of free-space optical MEMS devices: two-dimensional (2D) and three-dimensional (3D) optical MEMS switches [18]. In particular, the free-space 3D optical MEMS, where optical elements are positioned in three axes, is recognised as a key architectural module for building large-scale Optical Cross-Connects (OXCs), due to its high scalability enabled by spatial parallelism. Figure 1.11 shows the basic schematic of a free-space 3D optical MEMS switch [33]. The primary functional components include an input fibre array, output fibre array, lens array, input mirror array and output mirror array. In MEMS, the arrays of mirrors are critical elements to ensure the precise switching operations of optical signals [34], as the lightpath between input/output fibres is controlled by tilt angles of the MEMS mirrors. As illustrated in Figure 1.11, by rotating the input mirror array, the light beam from an input fibre is redirected to the output MEMS mirror array which adjusts its angle to reflect the light to the destined outgoing fibre. Once the connection is established, it is maintained for the whole transmission, which guarantees the Quality-of-Service (QoS) for applications. Thus MEMS is a coarse-grained switching technology which performs switching operations at the granularity of the full capacity of an optical fibre. Also, the MEMS switch is fully transparent to line rate, modulation format and protocol, since no signal processing is applied to optical signals in the MEMS.

The potential benefits of optical MEMS, coupled with the rapid development of MEMS technology, motivates the integration of MEMS into large-scale optical switches. To date, commercially

available MEMS-based optical circuit switch can support up to 320 ports [35]. However, owing to the utilisation of mechanical operations in mirror motion, optical MEMS switches exhibit relatively high reconfiguration times which are typically in the order of tens to hundreds of milliseconds, and therefore lead to a prohibitively large switching overhead for use in switching at packet granularity. Alternatively, the application of MEMS technology also extends to the realisation of tunable micro-optical devices such as tunable waveguides, couplers, filters, and other optical structures, due to its microfabrication technology and system integration and operation characteristics [36, 37].

1.4.2 Multiplexer/Demultiplexer (MUX/DEMUX)



(a) A $N \times 1$ MUX (b) A $1 \times N$ DEMUX

Figure 1.12: MUX/DEMUX

Figure 1.12 illustrates the functional block diagrams of optical MUX/DEMUX devices. In a Multiplexer (MUX), optical signals of different wavelengths are multiplexed into a composite WDM signal so that they can be transmitted on a single optical fibre. Conversely, an optical Demultiplexer (DEMUX) module splits the input WDM signal into its wavelength constituents, with each wavelength channel sent out from a different output port. The MUX/DEMUX allows the exploitation of the enormous transmission capacity in optical fibres, making it a fundamental key device in communication networks.

There are two types of optical MUX/DEMUX devices: Coarse Wavelength Division Multiplexing (CWDM) and Dense Wavelength Division Multiplexing (DWDM). Although the two technologies perform the same functionality of multiplexing several wavelengths on a single optical fibre, they differ in channel spacing and the number of wavelengths. More specifically, the CWDM MUX/DEMUX carries a small number of wavelength channels, typically eight with a channel separation of 20nm from 1470nm to 1610nm, thus it requires a relatively simple design architecture. In contrast, the DWDM MUX/DEMUX is designed to combine a large number of closely spaced optical signals,

up to 160 wavelengths with a dense channel spacing of 0.2nm (25GHz) [38]. However, because of the very narrow frequency window, the DWDM MUX/DEMUX requires a precise control of optical signals, leading to greatly increased design complexity. Generally, the optical MUX/DEMUX is built based on technologies such as the Arrayed Waveguide Grating (AWG), thin-film interference filters and fibre Bragg gratings [12].

1.4.3 Coupler/Splitter

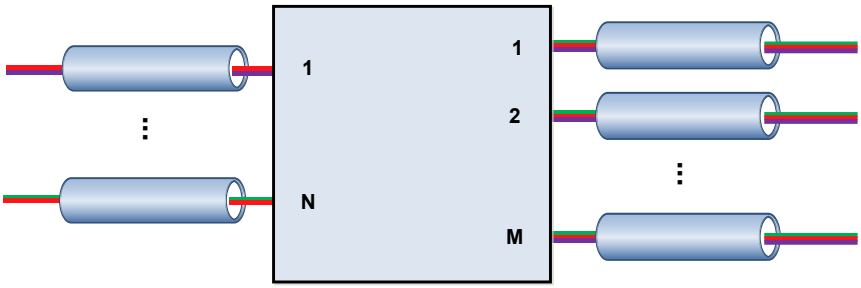


Figure 1.13: A $N \times M$ optical coupler.

An optical coupler is a passive device that combines one or more incoming light channels into one set which is subsequently distributed in power to different outputs. Generally, the output signal is the combined optical signal from all input ports. As illustrated in Figure 1.13, for a $N \times M$ coupler, optical signals coming from N input ports are combined, regardless of the signal wavelengths, and then distributed to M output ports. A $N \times 1$ coupler is referred to as a combiner.

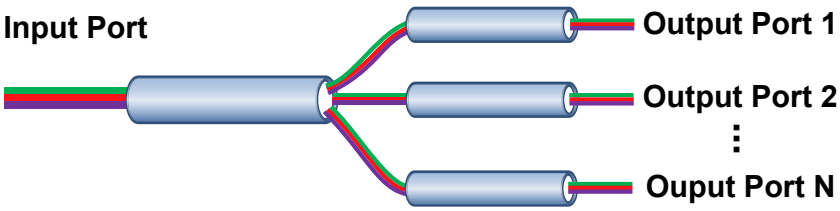


Figure 1.14: A $1 \times N$ optical splitter.

A $1 \times N$ optical splitter is a simplified coupler with 1 input port and N output ports, see Figure 1.14. The splitting is applied to the signal power, rather than to wavelengths as in an optical DEMUX. In an optical splitter, the optical power of the signal is divided into N paths, each passed into one output fibre [39].

1.4.4 Semiconductor Optical Amplifier (SOA)

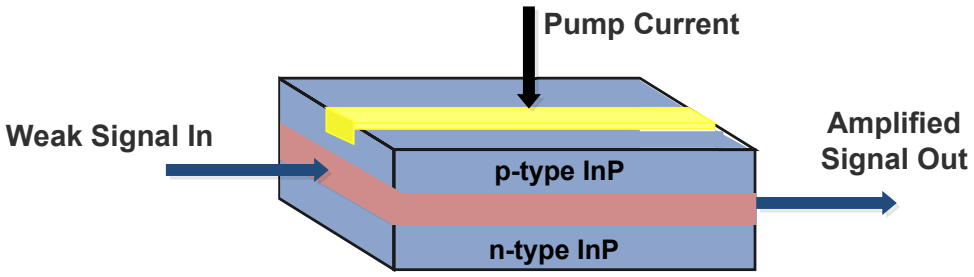


Figure 1.15: A Semiconductor Optical Amplifier (SOA).

An SOA is an active optical component which amplifies optical signals through the stimulated emission process [40]. Figure 1.15 shows the generic structure of the SOA, where the active medium is a *p-n* junction pumped with an injection current. In the SOA, the amplification gain is mainly determined by two processes: stimulated absorption and emission. By injecting sufficient current into the active region, the stimulated emission exceeds the absorption, that is, population inversion is triggered and, as a result, an optical gain is achieved.

The SOA offers the advantage of supporting all-optical amplification of the incident signals, which makes it particularly useful in high-speed optical networking. The main applications of SOA involve optical switching and high-speed wavelength conversion. In all-optical switching, the SOA is usually adopted as the gate component which either blocks or amplifies the injected optical signals depending on the controlling signals [18]. Alternatively, by exploiting fibre non-linearity, the optical SOA provides a practical technique for realising ultra-fast optical Tunable Wavelength Converter (TWC) which is transparent to data rates and modulation formats [41].

1.4.5 Wavelength Converter (WC)

A WC is designed to transfer the data information carried on one wavelength to a different wavelength channel. There are two general classes of wavelength converters: the Fixed Wavelength Converter (FWC) and the Tunable Wavelength Converter (TWC). A FWC is the simplest WC device which converts the injected wavelength to a specific outgoing wavelength [42]. In contrast to the FWC, a TWC can convert the injected wavelength to an arbitrary wavelength in a waveband. Various techniques have been proposed to realise WCs. One method is known as *opto-electronic wavelength conversion*, which makes use of receiver and transmitter devices. This particular scheme converters the incoming optical signal back to electrical format and subsequently retransmits it on

an appropriate wavelength. This scheme is indeed widely used today. Another approach is transparent conversion of the optical signal, which is made possible by using SOA components [41, 43, 44]. In optical networks, all-optical conversion is a desirable functionality, as it avoids Optic-Electro (O/E) and Electro-Optic (E/O) conversions.

The introduction of wavelength conversion technology into optical networking can greatly boost system performance and network flexibility, as it enables wavelength re-use and helps to alleviate the link congestion issue. Also important is that it can be deployed with other optical components like the AWG to build advanced, reconfigurable all-optical switching devices. However, the wavelength conversion of an optical signal with data rates of 10Gbps and beyond poses significant challenges, involving extinction ratio, signal-to-noise ratio (SNR), power penalty and conversion efficiency. Another key issue related to the TWC is difficulty in achieving a very large tuning range. To address these problems and make wavelength conversion feasible, significant advances in optical conversion technologies have been achieved. A tunable optical wavelength converter with a tuning range of 100nm in S and C bands is demonstrated in [45], which supports a positive conversion efficiency with low optical input power. [46] further demonstrates a broadband tunable wavelength conversion scheme that enables a 180nm span covering the S, C and L wavelength bands.

1.4.6 Arrayed Waveguide Grating (AWG)

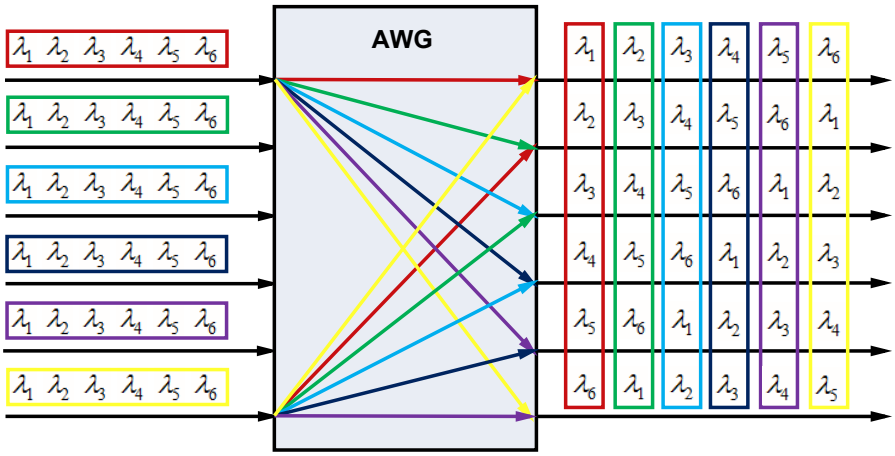


Figure 1.16: The cyclic wavelength routing pattern of a 6×6 AWG switch.

The AWG is a passive multiple-input-multiple-output optical component with a cyclic wavelength routing property. As illustrated in Figure 1.16, an AWG has a fixed switching matrix which determines the routing path of an incoming signal based on its wavelength. So an optical signal carried

by a certain wavelength can only be forwarded to a specific output port. This unique feature of wavelength routing enables simple yet fast optical switching. It is evident from Figure 1.16 that for a $N \times N$ AWG, an input port can communicate with N output ports by using different wavelength channels. Additionally, due to the wavelength parallelism at switch outputs, signals of different wavelengths entering from different input ports can be delivered to the same output port simultaneously, on different wavelengths. Thus, the AWG has the potential of supporting non-blocking all-to-all communications. However, due to the static routing, the AWG itself is a non-reconfigurable router. To facilitate a reconfigurable switching fabric, fast Tunable Wavelength Converters (TWCs) are added at AWG input ports. By configuring the TWCs to output the appropriate wavelength, the AWG can route an optical signal from an input port to any output ports, thereby enabling all-to-all connectivity. Suppose that each AWG output port is equipped with N receivers, a non-blocking reconfigurable $N \times N$ all-optical packet switch is then obtained. In such a configuration, a total of N^2 receivers are required. However, for large port count AWG switches, this installation is infeasible, thus fewer receivers are usually employed at switch outputs, consequently resulting in collisions. To address congestion, contention resolution strategies in AWG-based switches are necessary.

The optical AWG switch, coupled with wideband Tunable Wavelength Converters (TWCs), can potentially achieve high throughput, high-speed switching and low power consumption. Progress in integrated photonic technologies suggest that the AWG is also highly scalable, supporting up to 512 ports or more [47]. Hence the all-optical AWG switch provides a viable basis for building next-generation large-scale optical communication networks.

1.4.7 Wavelength Selective Switch (WSS)

A WSS is a switch module with one common input port and N output ports. The input port receives a light signal composed of several wavelengths. Then, the waveband of the input light is partitioned into N subsets, each including an arbitrary number of wavelengths, and subsequently the N subsets are routed to N separate output ports. As an example, Figure 1.17 illustrates the switching functionality of a 1×3 WSS which accepts a light signal consisting of 7 wavelength channels at its input port, and distributes them to 3 output ports. The WSS is a flexible switching technology where any wavelength of the input optical signal can be delivered to any one of the N output ports, allowing for high flexibility with regard to the allocation of wavelength resources. For that, the WSS technology has been extensively exploited in all-optical networking to realise various functions such as wavelength demultiplexing/multiplexing, optical switching and routing, dropping

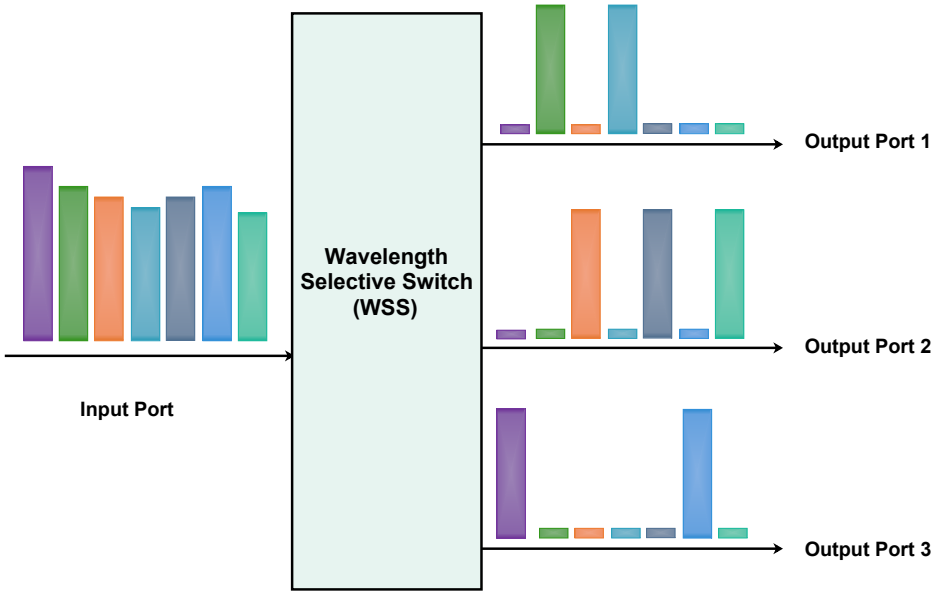


Figure 1.17: The switching functionality of a 1×3 WSS. The input port receives an optical signal containing 7 wavelength channels

and adding of wavelength channels at nodes and traffic balancing. There are a number of technical approaches to the wavelength selective switch design, including MEMS switching technology [48] and liquid crystal on silicon (LCoS) technology [49]. These manufacturing technologies lead to a relatively slow switching speed of WSS.

Having introduced the key optical switching components which are currently being deployed in optical networks, the remainder of this chapter reviews the proposed networking technologies and architectural designs for current and future datacentres.

1.5 Optical Switching in Datacentre Networks

Wavelength Division Multiplexing (WDM) is an optical technology enabling the statistical multiplexing of multiple wavelength channels into one fibre, which dramatically increases system bandwidth capacity, and therefore holds the potential to support future application needs in datacentres and HPC systems. All-optical WDM networks, which support ultra-high capacity, optical transparency and low power consumption, undoubtedly, represent a strong candidate for deployment in next-generation datacentres. Depending on the switching granularity, optical switching technologies can be generally classified into two categories: Optical Circuit Switching (OCS) and Optical Packet Switching (OPS).

Circuit vs. Packet Switching

- OCS. Optical Circuit Switching (OCS) is a coarse-grained switching technique operating at the granularity of a wavelength channel [50]. It provides point-to-point optical connections between input/output fibres with guaranteed bandwidth, ensuring Quality-of-Service (QoS). OCS is mainly built based on optical 3D MEMS technology with commercially available MEMS-based OCS supporting up to 320 ports [35]. This port intensity facilitates the wide deployment of OCS as optical cross-connects (OXCs) in large-scale networks. The main problem with the MEMS-based OCS is the high switch reconfiguration times which are in the orders of milliseconds. As such, OCS exhibits low flexibility and inefficient resource utilisation when it comes to bursty, dynamic traffic. Thus, optical circuit switching (OCS) is mainly used to handle long-lived bulky data transfers in optical networks.

- OPS. Optical Packet Switching (OPS) is a fine-grained, adaptive switching paradigm which supports fast optical switching at packet level. This technique allows more efficient utilisation of available bandwidth resources, because of the high switching flexibility enabled by the bandwidth provisioning on a per-packet basis. In OPS, the traffic is served with high efficiency, regardless of whether it is bursty in nature. For that, the OPS switching paradigm is well suited to the unpredictable nature of datacentre applications, and thus considered to be the future technology of datacentre networks, but it also faces some design challenges. The main impediment to employing all-optical packet switching is contention resolution, owing to the lack of technologically viable optical Random Access Memory (RAM). For the purpose of contention mitigation, several resolution strategies have been proposed for use. Wavelength conversion (WC), which transfers the data carried by a congested wavelength onto a free wavelength channel, is an effective contention management method, as it exploits spectral capacity to resolve congestions. Another technique is deflection routing which addresses collisions by directing optical signals onto an alternative path. Also important is optical buffering using Fibre Delay Lines (FDLs). This technique resolves contention by transmitting the contending packet on optical fibre lines, thereby offering a set of fixed length delays. Differently from FDLs, an electronic buffer can store a contending packet for any arbitrary time period, but it requires power-intensive O/E/O conversions and compromises the signal transparency. These methods are referred to as contention resolution strategies which attempt to avoid the occurrence of contention before it happens. An alternative to contention resolution is loss recovery mechanism which recovers the lost data after the contention happens.

Particularly, packet retransmission is an important data recovery technique, which allows the senders to retransmit the lost traffic, and thus achieving loss minimisation.

In recent years, several research efforts have been devoted to developing novel optical networking technologies and interconnection topologies for datacentre and HPC networks. This section introduces several recently proposed optical switching interconnection networks in the research literature. The benefits and design issues of these schemes are discussed and compared.

1.5.1 Optical Circuit Switching Networks

In this section, several optical circuit-switched datacentre architectures are presented and discussed. There are broadly two categories of OCS networks: hybrid EPS/OCS datacentre networks and transparent OCS-based datacentre networks. The former combines the merits of mature electronic packet switching (EPS) and advanced optical circuit switching (OCS), while the latter employs a high-radix OCS as the central switching fabric.

1.5.1.1 Hybrid EPS/OCS Datacentre Networks

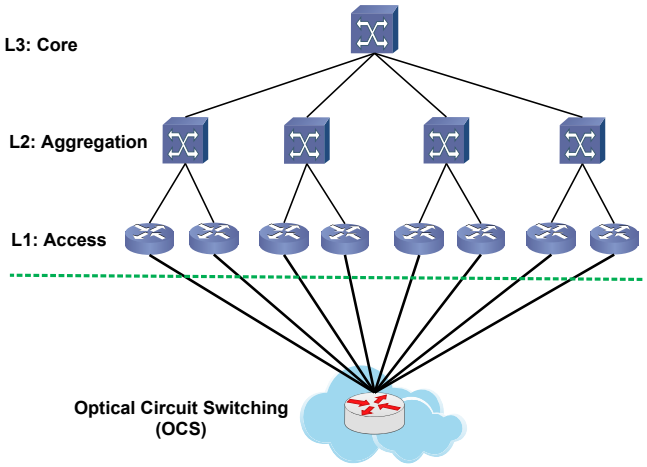


Figure 1.18: Hybrid c-Through network

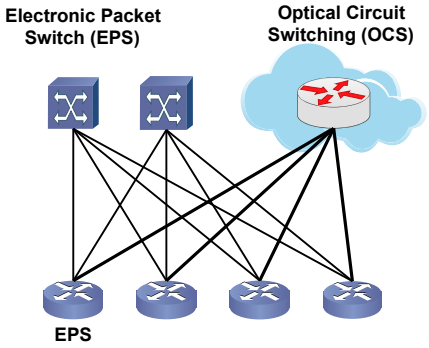


Figure 1.19: Hybrid Helios network

c-Through [51] and Helios [52] are two examples of hybrid electronic/optical intra-datacentre networks, which integrate advanced OCS technology into existing well-established electronic packet-switched network technologies, aiming to achieve the best of the two technologies. The structures of the c-Through and Helios networks are illustrated in Figures 1.18 and 1.19, respectively. Both architectures consist of two separate interconnection networks: a traditional electronic packet switch-

ing (EPS) network and an optical circuit switching (OCS) network. In c-Through, the electronic network follows a three-tier spanning-tree topology, whereas Helios deploys a two-level networking structure. The optical circuit-switched network uses a slow, high-capacity MEMS-based OCS which provides direct interconnections between access switches. In both c-Through and Helios, the electronic and optical systems are operating in parallel, with EPS targeting bursty traffic flows and OCS supporting large data transfers with guaranteed bandwidth, thereby resulting in greatly enhanced network capacity and flexibility. The hybrid designs, to some extent, alleviate the oversubscription and bandwidth capacity issues in current datacentres without requiring a complete replacement of hardware equipments. A key problem of the two schemes is that, due to the inherent drawbacks of multi-tier electronic networks and OCS technology, the EPS/OCS architecture is unable to host diverse datacentre applications with QoS guarantees. Although the EPS/OCS network performs reasonably well with applications with steady traffic requirements [53], for applications that exhibit dynamic, fast shifting traffic patterns, the network performance largely relies on the electronic network, instead of high-capacity OCS because of its low reconfigurability.

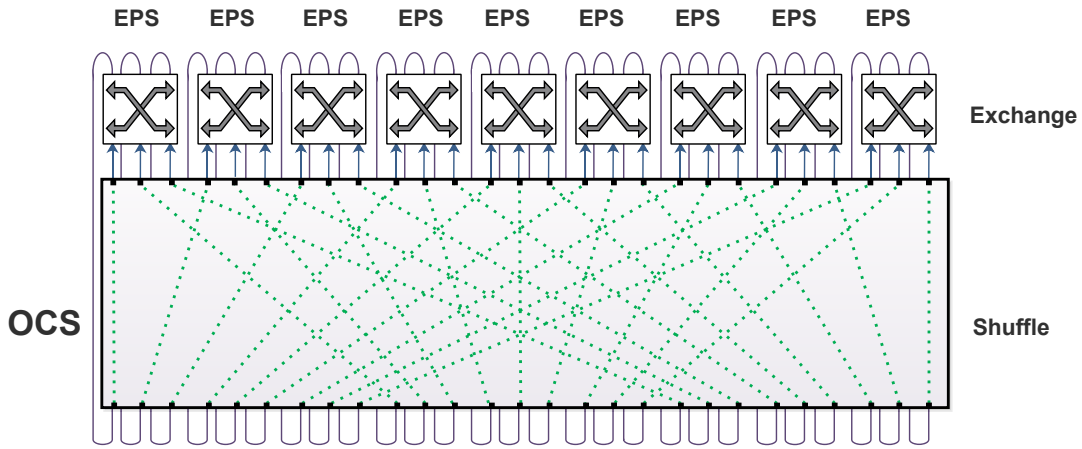


Figure 1.20: Hybrid SSX architecture.

The Single-stage Shuffle eXchange (SSX) architecture [54] is another hybrid EPS/OCS intra-datacentre architecture. Unlike c-Through and Helios, SSX employs EPSs and OCS together to build one interconnection network in which the OCS functions as the central switching fabric that interconnects EPSs, as shown in Figure 1.20. Each EPS has several input/output ports, each of which is attached to a port of the OCS via a uni-directional fibre, so there exists multiple connections between a pair of edge EPSs. In such a configuration, flow control is necessary, which is performed in the edge EPSs. This implies that the SSX architecture operates on the principle that

the *exchange* is performed first, and then the *shuffling*. More precisely, an optical signal injected to an EPS is first switched to the appropriate output port of the EPS which directly leads to an input port of the OCS, then through the OCS the optical signal is delivered to the destination EPS. In doing so, fine-grained EPSs perform efficient routing and flow control functionalities.

With the development of high-radix OCS, the SSX architecture now has the port densities for large-scale datacentres. As mentioned earlier, the OCS has a coarse-grained switching granularity and high reconfiguration times, which offset the benefits of the high-capacity optical switching. To avoid constant reconfiguration and expand the connectivity of the OCS, SSX employs hop-by-hop routing strategy which allows communications between two EPSs which are not currently being interconnected by the OCS. However, to do this, the input signal needs to travel through multiple EPSs and undergoes multiple O/E/O conversions before reaching its destined EPS, which consequently imposes increased communication latency and power consumption. Hence, a trade-off needs to be considered when using the SSX architecture.

1.5.1.2 Optical Circuit-Switched Datacentre Networks

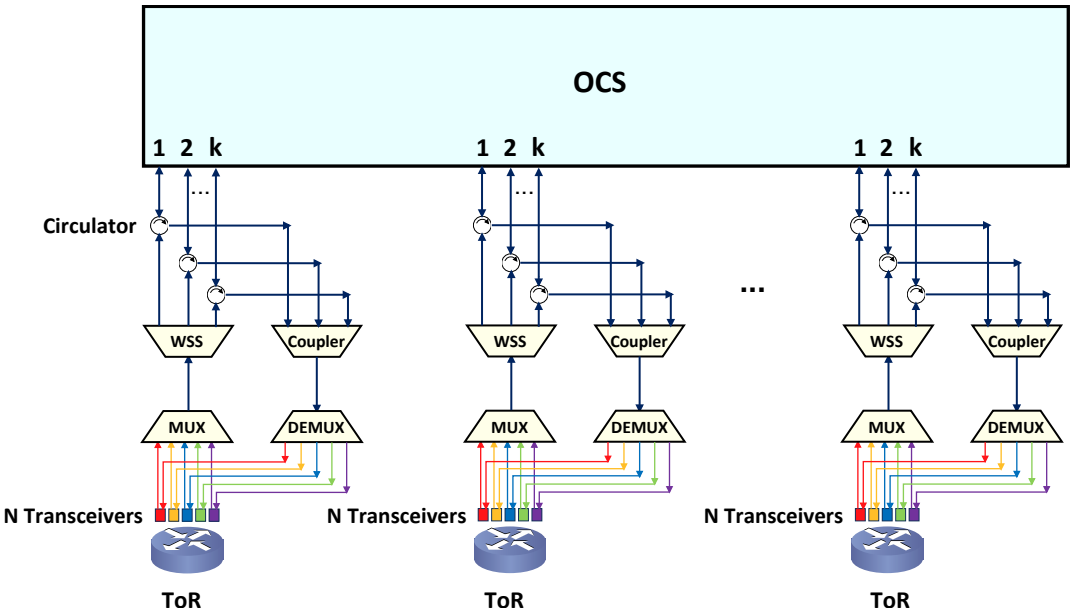


Figure 1.21: Optical circuit-switched OSA architecture.

OSA [55, 56], also referred to as Proteus, is a transparent optical circuit-switched architecture where the OCS functions as the core switching matrix. As illustrated in Figure 1.21, OSA is composed of optical MUX/DEMUX modules, WSSs, couplers and a high-port count OCS backplane. The

ToR switch is equipped with N transceivers transmitting and receiving optical signals on different wavelengths. Each ToR switch is connected to k ports of OCS through a MUX/DEMUX, a $1 \times k$ WSS and a $k \times 1$ coupler. N optical wavelengths coming from a ToR switch are combined into a composite WDM signal by the MUX which is followed by a $1 \times k$ WSS. The WSS module partitions the optical WDM signal into k subsets, each passed to an appropriate output port which is attached to an input port of the MEMS-based OCS. At the receiving side, the $k \times 1$ coupler collects the light signals coming from k output ports of OCS, and combines them into one WDM signal. The WDM signal is then split into N wavelengths by the DEMUX, and each wavelength is directed to the associated receiver in the ToR switch.

In OSA, a ToR switch can communicate with up to k ToR switches using N wavelength channels. Through the use of reconfigurable WSS, these wavelengths are adaptively assigned to the communications based on their traffic demands, thus introducing another degree of network flexibility, and also improving the quality of the communications.

Despite the use of flexible bandwidth assignment, the OSA scheme is still not flexible enough to efficiently support the rapidly changing traffic patterns in datacentre networks, owing to the inherent limitations of the OCS. Similar to the SSX architecture, OSA employs a multi-hop routing policy so as to alleviate these disadvantages, but at the cost of increased routing complexity, communication latency and O/E/O conversions. Therefore, to meet the traffic dynamics in datacentre applications, more flexible optical switching technology, which supports ultra-fast switching at packet level, high capacity scaling and high port density, is desirable. In response to this, OPS technology has been proposed for datacentre networking.

1.5.2 Optical Packet Switching Networks

There are mainly three classes of optical packet switching interconnects: Broadcast-and-Select (B&S) interconnection networks, wavelength-routed interconnection networks, and elastic interconnection networks. The B&S architecture is a space switch including two switching stages: broadcast and select. An optical signal from an ingress node is first divided by a power splitter, and broadcast to all edge nodes. Then the transmitted signal is selected and received by the designated egress node. This architecture is based on the Spanke structure demonstrated in Section 1.2.4. In wavelength-routed optical networks, the overlay network employs wavelength-routed optical components (AWGs) and wavelength converters (TWCs) to support optical packet-switched communi-

cations. Both networking concepts have been extensively exploited by a number of research projects including KEOPS [57], OPERA [58] and WASPNET [59]. Elastic networking is a flexible scheme which dynamically allocates the bandwidth resources to communications based on the application demands. In this type of architecture, the switching granularity ranges from sub-wavelength to super-wavelength, thereby achieving significantly enhanced spectral efficiency.

1.5.2.1 Broadcast-and-Select (B&S) OPS Networks

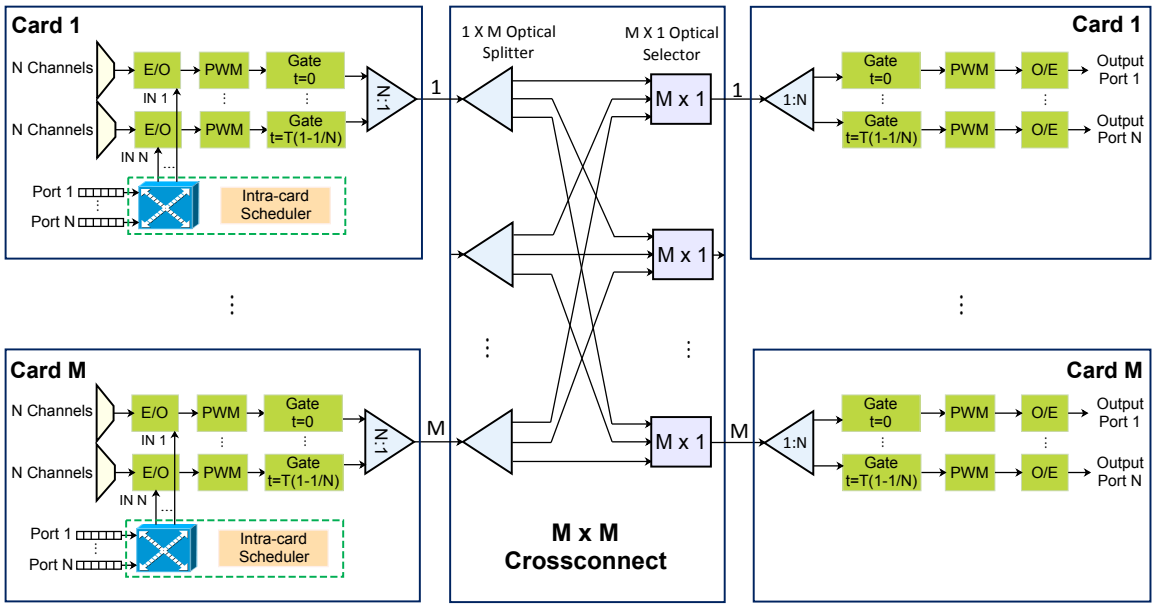


Figure 1.22: Optical packet-switched STIA architecture.

Space-Time Interconnection Architecture (STIA) [60] is a novel B&S optical packet-switched network which exploits three switching domains: wavelength domain, time domain and space domain. The structure of STIA is shown in Figure 1.22. The STIA architecture contains M switching cards and a $M \times M$ B&S Optical Cross-Connect (OXC). The $M \times M$ space OXC consisting of M $1 \times M$ power splitters and M $M \times 1$ Wavelength Selectors (WSs), serves as the core switching matrix and manages the inter-card communications. In the switching card, a packet compression technique exploiting wavelength domain and time domain, is adopted. This technique compresses an optical packet with a duration of T into a WDM packet of length T/N by modulating the packet on N wavelengths. In this way, N WDM packets coming from N input ports in a switching card can be accommodated into one time-frame T and transmitted in parallel, each occupying T/N time. Subsequently, the compressed optical packet is directed to the B&S OXC where the optical power of the packet is divided into M parts, each delivered to one Wavelength Selector (WS). Eventually,

a WS will accept the optical packet, and send it to the designated switching card.

The main networking challenge in STIA is that, due to signal splitting in the B&S OXC, only a small fraction of the transmitted power is received by the destined node, and consequently, costly, power-intensive optical amplifiers are required to compensate for the power loss. This requirement not only raises the capital and operational investments, but also places a constraint on network scalability. Alternatively, the deployment of a wavelength-based time compression technique, which demands a complex hardware and software implementation, greatly increases the design complexity and data processing time.

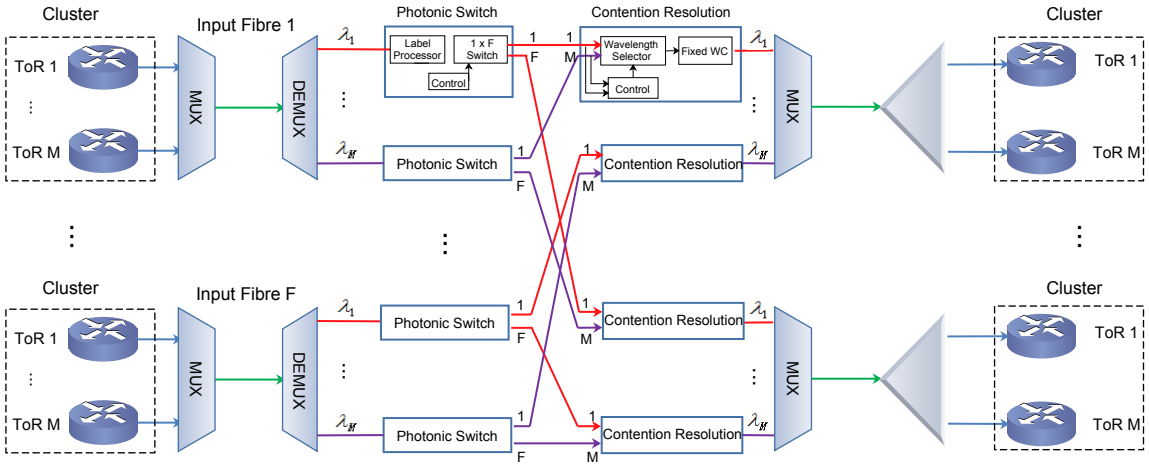


Figure 1.23: Optical packet-switched Spanke-type architecture.

An alternative design of B&S networks is the Spanke-type architecture with highly distributed control [61, 62]. As depicted in Figure 1.23, a $N \times N$ Spanke architecture has F input/output fibres, each input/output fibre connecting to M ToR switches inside a cluster by using M distinct wavelength channels, where $N = F \times M$. The input fibre is attached to a $1 \times M$ Demultiplexer (DEMUX) which separates the WDM signal into M component wavelengths, each of which is delivered to a $1 \times F$ photonic switch. The photonic switch, controlled by a local controller, broadcasts the received wavelength to F output fibres. At the output fibre, the selection operation is performed by using F $M \times 1$ Wavelength Selectors (WSs) which also work independently, thus enabling highly distributed control. The $M \times 1$ WS accepts M input wavelengths, but only allows one wavelength pass through, and blocks the rest of the wavelengths. In this case, contention arises among the M input wavelengths in each WS. To avoid performance impairments due to contentions, a retransmission mechanism is envisaged in combination with electronic buffers installed in the ToR switches.

The Spanke architecture supports highly distributed control, which can be of great benefit in dat-

acentre networking, since in this case the controlling complexity and switching latency is made independent of the switch size. The main disadvantage of the Spanke architecture is the high power splitting loss, because it follows a B&S architecture. Also, it exhibits a low degree of concurrent connectivity. More specifically, one out of M ToR switches in a cluster is allowed to communicate with a ToR switch in another cluster at any given time. This is problematic, as data-intensive applications may require massive communications across the interconnection network.

1.5.2.2 Wavelength-Routed OPS Networks

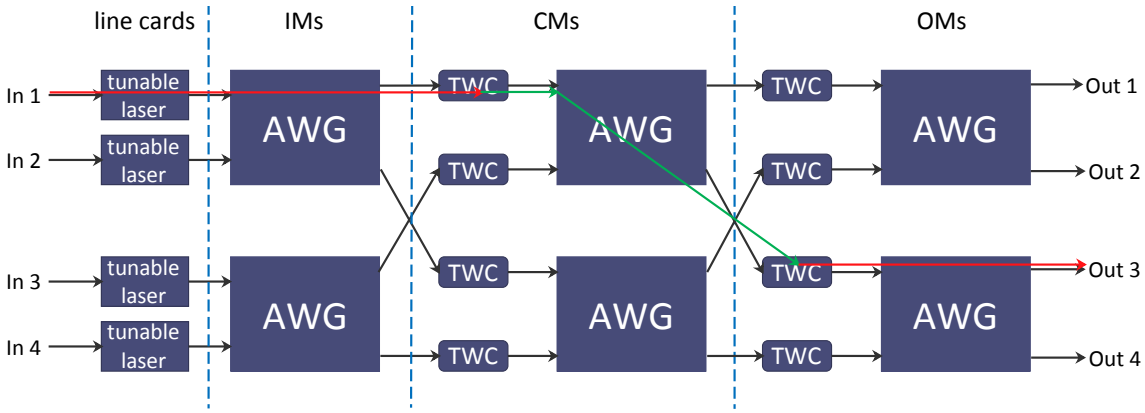


Figure 1.24: Optical packet-switched Petabit architecture.

Petabit [63, 64] is a transparent OPS architecture which supports fast optical switching by exploiting the wavelength routing property of the AWG. As depicted in Figure 1.24, the data switching plane is a three-stage Clos network consisting of input modules (IMs), central modules (CMs), and output modules (OMs). In each module, a passive $M \times M$ AWG router is deployed in conjunction with M Tunable Wavelength Converters (TWCs) to provide high-speed wavelength-routed switching. A connection between source and destination nodes can be established by setting up TWCs along the path on the appropriate wavelengths, based on the static wavelength routing table of the AWG. Petabit is a bufferless design, that is, the optical network does not include any buffers or fibre delay lines (FDLs). Instead, contention resolution is performed in linecards where electronic buffers are augmented. Once congestion occurs, the contending packet is converted back to the electronic domain and stored in an electronic buffer until a free path is available for data transmission. To avoid the head-of-line blocking problem, the electronic buffer is arranged into virtual output queues (VOQs) which are maintained on a per-OM basis.

The multi-stage design of Petabit allows the network to scale out easily by using more low-radix

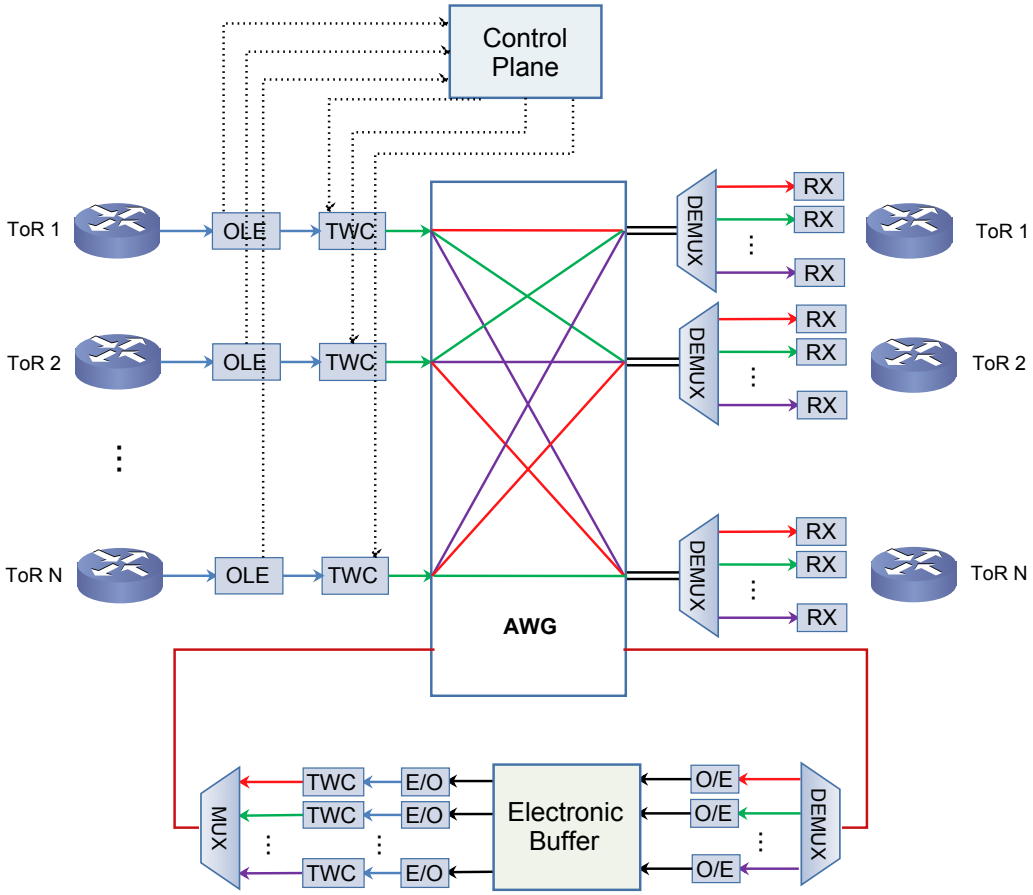


Figure 1.25: AWG-based LIONS architecture.

AWGs, but adds significant complexity to the controlling algorithm, as the routing operation involves multiple optical components. A major concern in Petabit is the very high deployment cost and energy consumption associated with the large number of TWCs required for wavelength routing and contention resolution, making this architecture particularly unsuitable for use in large-scale datacentres.

To mitigate the disadvantages of a multi-stage design, the LIONS (DOS) switch [65, 66] suggests using a single high-radix AWG switch to optically interconnect a large number of ToR switches, see Figure 1.25. In LIONS, the data switching plane is composed of four main functional modules: Optical Label Extractors (OLEs), TWCs, an AWG router and a feedback electronic buffer. The passive AWG functions as the core switching matrix of the OPS, which routes wavelengths from N input ports to N output ports. Each input port of the AWG is equipped with a dedicated OLE and TWC. The OLE component extracts the header packet from the data packet, and delivers it to the controller plane for processing. The controller identifies the routing information stored in

the header packet, and sets up the TWC accordingly, prior to the arrival of the corresponding data packet. After that, the data packet travels through the configured TWC and enters into the AWG switch. Based on its wavelength routing table, the AWG switch delivers the wavelength-adjusted data packet to the destined output port.

The LIONS architecture supports non-blocking all-to-all communications, provided that each output port has N receivers, one for each input port. This introduces difficulty, especially in large-scale LIONS switches, owing to high capital/operational investments. Therefore, at each output port, k receivers are utilised, where $k < N$. With this implementation, at most k out of N wavelengths at an output are allowed to travel through the optical DEMUX concurrently [66], which consequently leads to packet collisions. To mitigate the effects of contentions, LIONS employs a feed-back electronic buffer to store the contending packet until the required receiver at designated output port becomes free.

The LIONS architecture is designed based on a high-radix AWG device, allowing flattening the existing hierarchical datacentre networks. Also, it offers the attributes of fast all-optical switching, high throughput and low power dissipation. However, this scheme adopts electronic buffers for contention resolution, and a high-capacity, high-speed electronic buffer is expected to consume a large amount of power, which can be a significant handicap for enabling a cost-and energy-efficient interconnect.

1.5.2.3 Elastic OPS Networks

To enhance bandwidth provisioning flexibility in datacentre interconnects, [67, 68] propose a new networking concept–the Elastic Networking paradigm, which adaptively assigns the bandwidth resources to communications based on the traffic demands. A novel aspect of the elastic networking solution is its arbitrary switching granularity ranging from sub-wavelength to super-wavelength, which is realised through the use of a flexible spectrum-sliced grid. The flexible spectrum allocation involves two essential enabling technologies: variable bandwidth transponders (VBTs) and variable bandwidth cross-connects. Multicarrier techniques, which form a broadband waveform using many low-rate subcarriers generated by low-speed modulators, have been proposed as possible transponder implementations for elastic optical networks [68]. Alternatively, the variable bandwidth cross-connect is often designed based on WSSs which support optical switching of arbitrary spectrum slices. Also, the AWG switch can be envisaged as a variable bandwidth cross-connect [69].

In elastic networking, dynamic adaptation of line rate and bandwidth capacity to traffic demands is considered to be the most important attribute, which results in finer-grained arbitrary switching granularity and very high spectral efficiency, in comparison with conventional WDM networks. Thus, the elastic optical network represents an attractive solution for matching bandwidth capacity to application needs, but there are some problems to be addressed. One is spectral fragmentation, where the scattered and fragmented spectrum slots remain unused, attributed to a waste of spectral resources. Also challenging is the high design complexity of an efficient routing and spectrum allocation (RSS) algorithm in large-scale data-intensive datacentres.

1.5.2.4 Hybird OCS/OPS Datacentre Networks

In datacentres, some applications host relatively long-lived bulky data transfers, which would benefit from low-speed circuit switching technology, due to its point-to-point bandwidth provisioning. Conversely, some other applications are comprised of rapid changing, short-lived traffic flows which can be delivered by high-speed, fine-grained optical packet switching. Motivated by this, hybrid OCS/OPS architectures exploiting the relative merits of both circuit switching and packet switching technologies are proposed to effectively support various datacentre applications. The hybrid design promises significantly improved resource utilisation, network flexibility and QoS.

A prominent example of hybrid OCS/OPS interconnects is the LIGHTNESS architecture proposed by the EC FP7 LIGHTNESS project [70, 71]. Figure 1.26 shows a high-level overview of the hybrid OCS/OPS LIGHTNESS framework. The network deploys a flat two-level structure of core and access layers. The core layer includes transparent OCS and OPS networks, both providing fully meshed interconnectivity among ToR switches. The OCS, which is designed based on MEMS technology, is responsible for handling large traffic streams with bandwidth guarantee. In contrast, the envisioned OPS network, which is built based on the B&S Spanke-type switch described in Section 1.5.2.1, targets short-lived traffic flows. In order to coordinate between OCS and OPS, traffic classification and flow control is performed in the ToR switches.

The hybrid OCS/OPS technology fits better to datacentre networks, as it holds the flexibility to support the heterogeneity in datacentre applications. However, this comes with the requirement for an efficient traffic classification technique which enables fast scheduling prior to optical switching. To achieve this, complex hardware and software equipments need to be installed, resulting in significantly increased deployment costs. Alternatively, in the LIGHTNESS architecture, the deployment of the B&S Spanke-type OPS switch facilitates highly distributed control, but it also suffers from

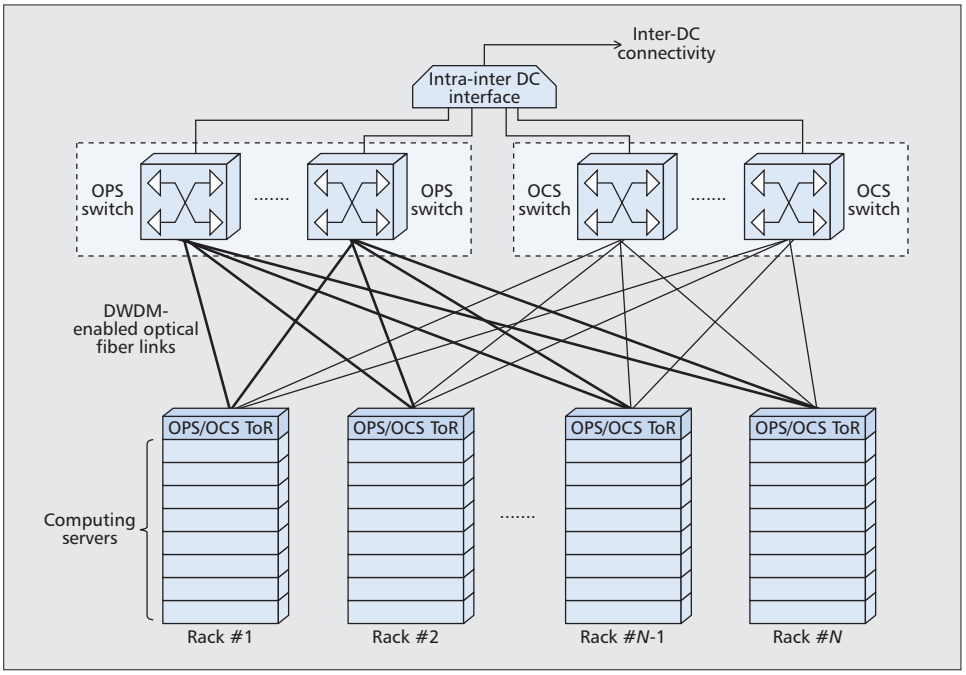


Figure 1.26: Hybrid OCS/OPS LIGHTNESS architecture [70, 71].

issues relating to concurrent interconnectivity and power splitting loss.

Another hybrid OCS/OPS interconnect is proposed in [72]. Differently from the LIGHTNESS switch, this hybrid architecture makes use of one Optical Cross-Connect (OXC) for both circuit and packet switching. Figure 1.27 depicts the structure of a $2NM \times 2NM$ OXC, which is a space wavelength-routed switch containing $2NM$ $1 \times M$ optical splitters, NM $2M \times 1$ optical combiners, M $N \times N$ AWG routers, and NM 1×2 couplers. An optical signal entering from an input port reaches a $1 \times M$ optical splitter which broadcasts the signal to M AWG routers through M $2M \times 1$ combiners. The function of the $2M \times 1$ combiner is to collect the $2M$ optical signals injected to its input ports and combine them into a composite WDM signal. The combined light is then delivered to an input port of the AWG router. Together with TWCs, the AWG forwards the input light to different output ports. At each output port, a 1×2 coupler is utilised to divide the output WDM signal into two groups, each passed to a broadband receiver.

This hybrid architecture utilises a shared optical switching fabric for both circuit and packet switching. Depending on the traffic identification process performed in the ToR switches, the OXC is set up to perform circuit or packet switching. Importantly, due to the fact that circuit-switched traffic contains large volumes of data, the circuit switching has higher priority than packet switching. A consequence of this implementation is the severely degraded system performance, because if the

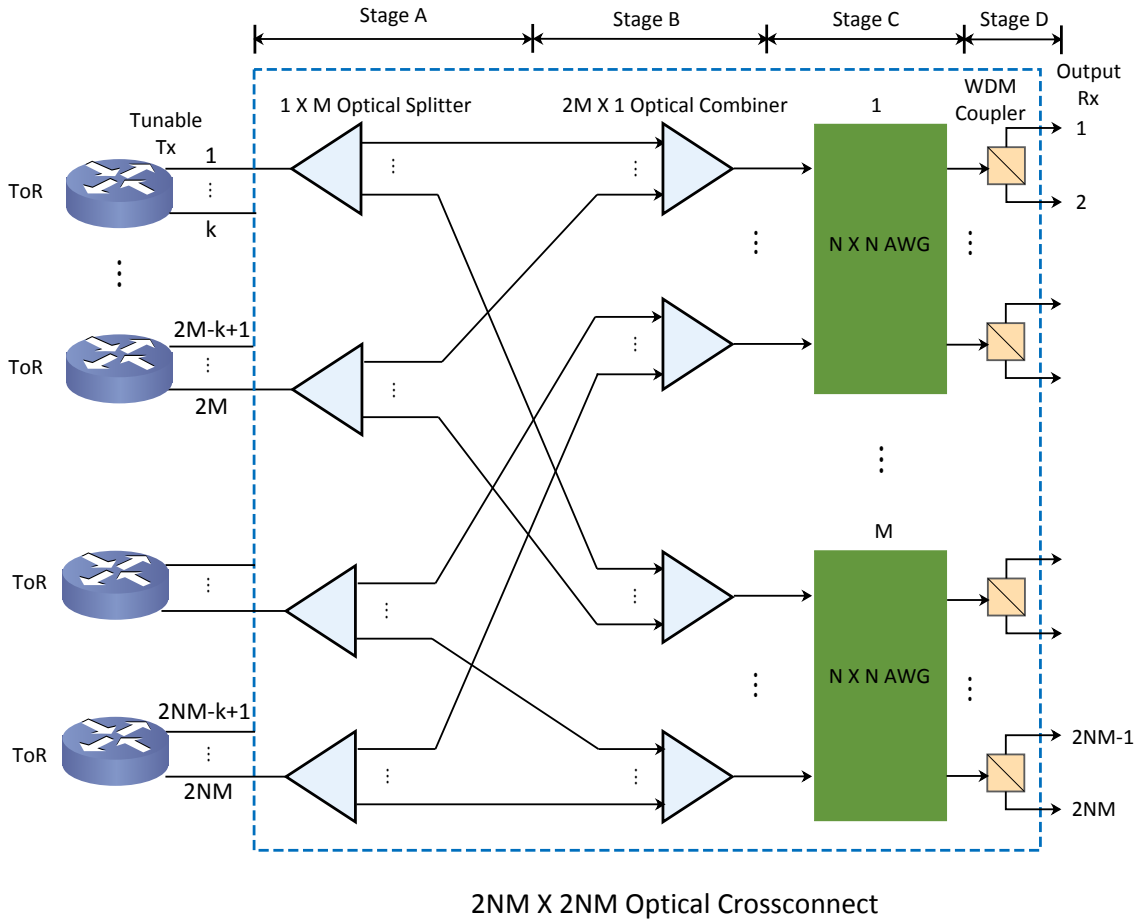


Figure 1.27: The space wavelength-routed optical cross-connect (OXC).

shared OXC is in use for circuit switching, all packet-switched traffic will need to be stored in electronic virtual output queues (VOQs) inside ToR switches. Another drawback of this architecture is the adoption of complex traffic classification and a huge number of transceivers and TWCs, which contributes considerately to the overall cost and power consumption.

1.6 Conclusions

The rapid growth of datacentre traffic due to cloud-computing applications and on-demand services motivates the deployment of high-performance datcentre interconnection networks. Optical networking concepts and technologies present challenges and opportunities for satisfying current and future application needs in datacentres, owing to ultra-high bandwidth capacity and transparency to datarates, modulation formats and protocols. Of utmost importance is high-speed Optical Packet

Switching (OPS) which promises high network flexibility and bandwidth capacity to support service heterogeneity in datacentres. Remarkable technological advances in integrated photonic technologies have been achieved, especially in AWG switching technology. It is indicated that AWG can scale up to 512 ports [47], highlighting the opportunity to fabricate high-radix optical AWG switches. The potentially wide deployment of high-radix AWG switches makes possible the idea of flattening current hierarchical datacentre networks.

The AWG-based OPS holds some attractive features including ultra-high bandwidth, high-speed switching, low power consumption and operational flexibility, thereby it offers an excellent solution to overcome the challenges in large-scale datacentres, and deserves continued attention. Nonetheless, contention resolution in this type of switch is problematic, due to the absence of viable optical RAM technology. In the remainder of this thesis, various implementations of high-radix AWG switches employing advanced enabling optical technologies and contention resolution techniques are proposed, with the purpose of building large-scale flattened datacentre networks with improved energy efficiency and network performance. The performance evaluation and power estimation of the proposed optical packet switches will be a major focus of the work.

Bibliography

[1] Dennis Abts and John Kim, "High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities," Morgan & Claypool Publishers, Mar. 2011.

[2] Cisco Systems, "Cisco global cloud index: forecast and methodology 2012-2017," White Paper, Oct. 2013. [Online]. Available: http://knoema.com/jlbtvff/cisco-global-cloud-index-2012-2017.

[3] Cisco Systems, "Cisco global cloud index: forecast and methodology 2013-2018," White Paper, Nov. 2014. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.pdf.

[4] Cisco Systems, "Growth in the cloud," 2014. [Online]. Available: http://www.cisco.com/c/dam/assets/sol/sp/network_infrastructure/growth_cloud_infographic.html.

[5] W. E. Denzel, J. Li, P. Walker, and Y. Jin, "A framework for end-to-end simulation of high-performance computing systems". in *First International Conference on Simulation Tools and Techniques for Communications, Networks and Systems (SIMUTools 2008)*, Mar. 2008.

[6] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken,"Nature of datacenter traffic: measurements and analysis," in *ACM SIGCOMM*, 2009.

[7] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *ACM SIGCOMM*, 2010.

[8] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: dynamic flow scheduling for data center networks," in *Networked Systems Design and Implementation (NSDI) Symposium*, Apr. 2010.

[9] P. Bodk, I. Menache, and M, Chowdhury, " Surviving failures in bandwidth-constrained datacenters," in *ACM SIGCOMM conference on Applications, technologies, architectures, and*

protocols for computer communication, Aug. 2012.

[10] FP Tso and D. Pezaros, "Improving data center network utilization using near-optimal traffic engineering," *IEEE Transactions on Parallel & Distributed Systems*, vol. 24, no. 6, pp. 1139-1148, June 2013.

[11] A. Roy, H. Zengy, J. Baggay, G. Porter, and A. C. Snoeren, "Inside the social networks (Datacenter) network", in *ACM SIGCOMM*, pp. 123-137, Aug. 2015.

[12] R. Ramaswami, K. N. Sivarajan, and G. H. Sasaki, "Optical Networks: A Practical Perspective," Elsevier, USA, 2010.

[13] Andre DeHon, "Rent's rule based switching requirements," in *International Workshop on System-level Interconnect Prediction*, pp. 197-204, 2001.

[14] Tarek S. El-Bawab, "Optical Switching," Springer, 2006.

[15] J. L. Hennessy and D. A. Patterson, "Computer Architecture: A Quantitative Approach," Morgan Kaufmann Publishers, 2003.

[16] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *ACM SIGCOMM*, Aug. 2008.

[17] Charles Clos, "A study of non-blocking switching networks," *Bell System Technical Journal*, vol. 32, no. 2, pp. 406-424, Mar. 1953.

[18] B. Mukherjee,"Optical WDM Networks," Springer, 2006.

[19] M. Moazez, F. Safaei, and M. Rezazadeh, "Design and implementation of multistage interconnection networks for SoC networks," *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, vol. 2, no. 5, Oct. 2012.

[20] J. Kim, W. J. Dally, and D. Abts, "Flattened Butterfly: a cost-efficient topology for high-radix networks," *International Symposium on Computer Architecture (ISCA)* , pp. 126-137, June 2007.

[21] Hewlett-Packard Development Company, " Effects of virtualization and cloud computing on data center networks," Oct. 2011. [Online]. Available: http://h20565.www2.hp.com/hpsc/doc/public/display?docId=emr_na-c03042885.

[22] R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "PortLand: a scalable fault-tolerant layer 2 data center network fabric,"

in *ACM SIGCOMM*, pp. 39-50, Aug. 2009.

[23] A. A. Kist, and L. Brodie, "Quality of service, quality of experience and online learning," *Frontiers in Education Conference (FIE)*, pp. 1-6, Oct. 2012.

[24] Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, and Changhoon Kim, "VL2: a scalable and flexible data center network," in *SIGCOMM*, Aug. 2009.

[25] N. Farrington and A. Andreyev, "Facebooks data center network architecture," in *IEEE Optical Interconnects Conference (OI)*, 2013.

[26] A. Andreyev, "Introducing data center fabric, the next-generation Facebook data center network," Nov. 2014. [Online]. Available: https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network.

[27] M. Alizadeh, and T. Edsall, "On the data path performance of Leaf-Spine datacenter fabrics," *High-Performance Interconnects (HOTI)*, pp. 71-74, Aug. 2013.

[28] Arista, "High Performance, Feature Rich 10/40GbE Systems: 40Tbps & 512×40GbE, 2048×10GbE, 768×10GBASE-T," Nov. 2014. [Online]. Available: https://www.arista.com/en/products/7300-series.

[29] C. Guo, H. Wu, K. Tan, L. Shiy, Y. Zhang, and S. Lu,"DCell: a scalable and fault-tolerant network structure for data centers," in *ACM SIGCOMM*, Aug. 2008.

[30] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: a high performance, server-centric network architecture for modular data centers," in *ACM SIGCOMM*, Aug. 2009.

[31] D. Li, C. Guo, H. Wu, K. Tan, Y. Zhang, and S. Lu, "FiConn: using backup port for server interconnection in data centers," in *IEEE International Conference on Computer Communications*, pp. 2276-2285, Apr. 2009.

[32] D. Li, C. Guo, H. Wu, K. Tan, Y. Zhang, S. Lu, and J. Wu, "Scalable and cost-effective interconnection of data-center servers using dual server ports," *IEEE/ACM Transactions on Networking*, vol. 19, no. 1, pp. 102-114, Feb. 2011.

[33] CALIENT, "3D MEMS Optical Circuit Switching Demonstration," Oct. 2013. [Online]. Available: http://www.calient.net/2013/10/calient-3d-mems-optical-circuit-switching-demonstration.

[34] A. Neukermans, and R. Ramaswami, "MEMS technology for optical networking applications," *IEEE Communications Magazine*, vol. 39, no. 1, pp. 62-69, Jan. 2001.

[35] CALIENT, "S320 Photonic Switching", 2013. [Online]. Available: http://www.enterprisemanagement360.com/wp-content/files_mf/1377248645CalientS320DataSheetMarch12013.pdf.

[36] O. Solgaard, A. A. Godil, R. T. Howe, L. P. Lee, Y. A. Peter and H. Zappe, "Optical MEMS: from micromirrors to complex systems," *Journal of Microelectromechanical Systems*, vol. 23, no. 3, pp. 517-538, June 2014.

[37] H. Du, F. S. Chau, and G. Zhou, "Mechanically-tunable photonic devices with on-chip integrated MEMS/NEMS actuators," *Micromachines*, vol. 7, no. 4, pp. 1-24, Apr. 2016.

[38] International Telecommunications Union,"G.694.1 spectral grids for WDM applications: DWDM frequency grid" Series G: Transmission Systems and Media, Digital Systems and Networks, 2012.

[39] Alan Rogers, "Understanding Optical Fiber Communications," Artech House, USA, 2001.

[40] P. Urquhart, "Advances in Optical Amplifiers," InTech, Rijeka, Croatia, 2011.

[41] A. K. Jaiswal and S. Verma, "Performance on SOA-based wavelength conversion at 10 GB/s by dual pump four wave mixing over a 50 nm," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 5, May 2013.

[42] Martin Maier, "Optical Switching Networks," Cambridge University Press, Feb. 2008.

[43] J. M. H. Elmirghani, and H. T. Mouftah, "All-optical wavelength conversion: technologies and applications in DWDM networks," *IEEE Communications Magazine*, vol. 38, no. 3, pp. 86-92, Mar. 2000.

[44] F. Bontempi, N. Andriolli, S. Faralli, J. Klamkin, E. Kleijn, T. de Vries, and G. Contestabile, "An InP monolithically integrated multi-frequency wavelength converter," in *2014 Optical Fiber Communication Conference (OFC)*, Mar. 2014.

[45] G. Contestabile, Y. Yoshida, A. Maruta, and K. Kitayama, "Ultra-broad band, low power, highly efficient coherent wavelength conversion in quantum dot SOA," *Optics Express*, vol. 20, no. 25, pp. 27902-27907, Dec. 2012.

[46] N. A. Awang, A. A. Latif, M. Z. Zulkifli, Z. A. Ghani, S. W. Harun, and H. Ahmad, "S + C + L Band tunable wavelength conversion using FWM dual-wavelength fiber laser in a highly non-

linear fiber," *Microwave and Optical Technology Letters*, vol. 55, no. 2, pp. 379-382, Feb. 2013.

[47] S. Cheung, T. Su, K. Okamoto, and S. J. B. Yoo, "Ultra-compact silicon photonic 512×512 25-GHz Arrayed Waveguide Grating Router," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 4, pp. 310-316, Aug. 2014.

[48] D. M. Marom et al., "Wavelength-selective 1×4 switch for 128 WDM channels at 50 GHz spacing," in *Optical Fiber Communication Conference (OFC)*, pp. FB7-1-FB7-3, Mar. 2002.

[49] R. Ryf et al., "Wavelength-selective switch for few-mode fiber transmission," in *39th European Conference on Optical Communications (ECOC)*, 2013.

[50] M. Chen, H. Jin, Y. Wen, and V. C. M. Leung, "Enabling technologies for future data center networking: a primer," *IEEE Network*, vol. 27, no. 4, pp. 8-15, Aug. 2013.

[51] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. Eugene Ng, M. Kozuch, and M. Ryan, "c-Through: part-time optics in data centers," in *ACM SIGCOMM*, Oct. 2010.

[52] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," in *ACM SIGCOMM*, Oct. 2010.

[53] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. Eugene Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, and A. Vahdat, "Switching the optical divide: fundamental challenges for hybrid electrical/optical datacenter networks," in *ACM Symposium on Cloud Computing (SOCC)*, pp. 1-8, Oct. 2011.

[54] D. Lugones, K. Katrinis, and M. Collier, "A reconfigurable optical/electrical interconnect architecture for large-scale clusters and datacenters," in *ACM International Conference on Computing Frontiers*, pp. 13-22, May 2012.

[55] L. Xu, W. Zhang, H. L. R. Lira, M. Lipson, and K. Bergman, "A hybrid optical packet and wavelength selective switching platform for high-performance data center networks," *Optical Express*, vol. 19, no. 24, Nov. 2011.

[56] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, "OSA: an optical switching architecture for data center networks with unprecedented flexibility," *Northwestern University ,Tech Rep*, 2012.

[57] C. Guillemot et al., "Transparent optical packet switching: the European ACTS KEOPS

project approach," *Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2117-2134, Dec. 1998.

[58] A. Carena, M. D. Vaughn, R. Gaudino, M Shell, and D. J. Blumenthal, "OPERA: an optical packet experimental routing architecture with label swapping capability", *Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2135-2145, Dec. 1998.

[59] D. K. Hunter et al., "WASPNET: a wavelength switched packet network," *Communications Magazine*, vol. 37, no. 3, pp. 120-129, Mar. 1999.

[60] O. Liboiron-Ladouceur, P. G. Raponi, N. Andriolli, I. Cerutti, M. S. Hai, and P. Castoldi, "A scalable space-time multi-plane optical interconnection network using energy-efficient enabling technologies," *Journal of Optical Communications and Networking*, vol. 3, no. 8, pp. 1-11, Aug. 2011.

[61] S. D. Lucente, N. Calabretta, J. A. C. Resing, and H. J. S. Dorren, "Scaling low-latency optical packet switches to a thousand Ports," *Journal of Optical Communications and Networking*, vol. 4, no. 9, pp. 17-28, Sept. 2012.

[62] N. Calabretta, R. P. Centelles, S. D. Lucente, and H. J. S. Dorren, "On the performance of a large-scale optical packet switch under realistic data center traffic," *Journal of Optical Communications and Networking*, vol. 5, no. 6, pp. 565-573, June 2013.

[63] K. Xia, Y.H. Kaob, M. Yangb, and H. J. Chao, "Petabit optical switch for data center networks," *Technical report, Polytechnic Institute of NYU*, 2010.

[64] H. J. Chao and K. Xi, "Bufferless optical Clos switches for data centers," in *OSA/OFC/NFOEC*, 2011.

[65] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS–a scalable optical switch for datacenters," *Architectures for Networking and Communications Systems (ANCS)*, pp. 1-12, Oct. 2010.

[66] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella and S. J. B. Yoo, "LIONS: an AWGR-based low latency optical switch for high-performance computing and data centers," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 2, Mar. 2013.

[67] B. Kozicki, H. Takara, and M. Jinno, "Enabling technologies for adaptive resource allocation in elastic optical path network (SLICE)," in *Communications and Photonics Conference and Exhibition (ACP)*, pp. 23-24, Dec. 2010.

[68] O. Gerstel, M. Jinno, A. Lord, and S.J.B. Yoo, "Elastic optical networking: a new dawn for the optical layer?," *Communications Magazine*, vol. 50, no. 2, pp. 12-20, Feb. 2012.

[69] S. J. B. Yoo, Y. Yawei, and W. Ke, "Intra and inter datacenter networking: the role of optical packet switching and flexible bandwidth optical networking," in *Optical Network Design and Modeling (ONDM)*, pp. 1-6, Apr. 2012.

[70] N. Calabretta, "Scalable and low latency optical packet switching architectures for high performance data center networks," *Photonics in Switching (PS)*, July 2014.

[71] S. Peng, R. Nejabati, B. Guo, Y. Shu, G. Zervas, S. Spadaro, A. Pages, and D. Simeonidou, "Enabling multi-tenancy in hybrid optical packet/circuit switched data center networks," in *40th European Conference and Exhibition on Optical Communication (ECOC)*, Sept. 2014.

[72] Q. Huang, Y. K. Yeo, and L. Zhou, "Combining circuit and packet switching using a large port-count optical cross-connect for data center networks," *Optics Communications*, vol. 285, no. 21-22, pp. 4268-4274, Oct. 2012.

Chapter 2

A Novel Energy-Efficient High-Speed Optical Packet Switching Datacentre Network

2.1 Introduction

As stated in the previous chapter, high-capacity optical packet switching has the potential to support the communication requirements between ever increasing numbers of highly interdependent servers in datacentres. An attractive candidate solution for optical packet switching is the wavelength-switched architecture, which adopts Tunable Wavelength Converters (TWCs) and an Arrayed Waveguide Grating (AWG) router for cross-connection. A key aspect of the AWG is that it is a passive optical device and can potentially scale to very high port counts [1]. Importantly, due to its fixed cyclic wavelength routing property, the AWG allows the selection of output ports based on the wavelength of the input optical signal [2]. Thus, by tuning the TWCs installed at input ports, an AWG can passively route the optical signals to any output ports, forming a high-speed, reconfigurable wavelength-routed switching fabric. Moreover, at switch output ports, the AWG allows multiple packets on different wavelengths to be detected concurrently, if multiple parallel receivers are located there. In particular, a contention-free $N \times N$ switch is possible provided that each AWG output has a $1 : N$ Demultiplexer (DEMUX) and N separate receivers [3]. However, it is generally not scalable to implement N receivers at each of N outputs, especially for high-port count AWGs, hence many fewer receivers are typically used and as such, the resulting contention needs to be

handled by some form of packet buffering. This makes contention resolution an important issue for the realisation of AWG-based optical switches.

Previous research, e.g. the Datacenter Optical Switch (DOS) [3] resolves conflicts by buffering contending packets in a single shared electronic buffering. However, due to considerable contention at output ports, especially under high load conditions, this design requires a high-capacity electronic buffer with the same port count and same port speed as the overall switch. Essentially, such an electronic buffer amounts to a high port-count, high bandwidth switch, yielding a potentially high-cost, high power-consumption architecture overall. [4] makes extensive comparisons between electronic buffers and single-wavelength Fibre Delay Lines (FDLs) in the context of AWG switches. The authors conclude that the electronic buffer outperforms FDLs with respect to latency and throughput, though at the expense of higher power consumption. In related work, [5] and [6] studied the enhanced DOS architectures (a.k.a. LIONS) with three different loop-back electronic buffering designs, but did not further consider FDLs. In [7–10], AWG-based packet switches with N feedback optical buffers are proposed. However, these architectures face network scalability and overall system complexity issues, due to the large number of FDL buffers utilised.

In comparison to these previous findings, the main novelty and advancement of the work presented in this chapter is that this work solves the contention resolution challenge in AWG-based packet switches by designing an efficient buffering hardware and packet scheduling scheme. More precisely, a novel hybrid-buffered AWG architecture, which employs a relatively small number of multi-wavelength Fibre Delay Lines (FDLs), in combination with a very small electronic buffer, using an optimal method for integrating FDL wavelengths into the AWG switch, is proposed for deployment in building scalable, energy-efficient, high-performance interconnection networks. The motivations behind and the advantages of using such a hybrid-buffered AWG architecture are examined by simulation work. The network dimensioning problem is also investigated.

This chapter is organised as follows. Section 2.2 presents details of the new architecture together with its contention resolution scheme. The format of the optical packet and the packet scheduling algorithm are also described in this section. Section 2.3 details the simulation setup and the experimental work carried out to examine the network performance and power consumption of the proposed optical switch. Also investigated is the execution complexity of the packet scheduling scheme. In Section 2.4, some important conclusions are outlined.

2.2 Switch Node Architecture

2.2.1 Switch Architecture

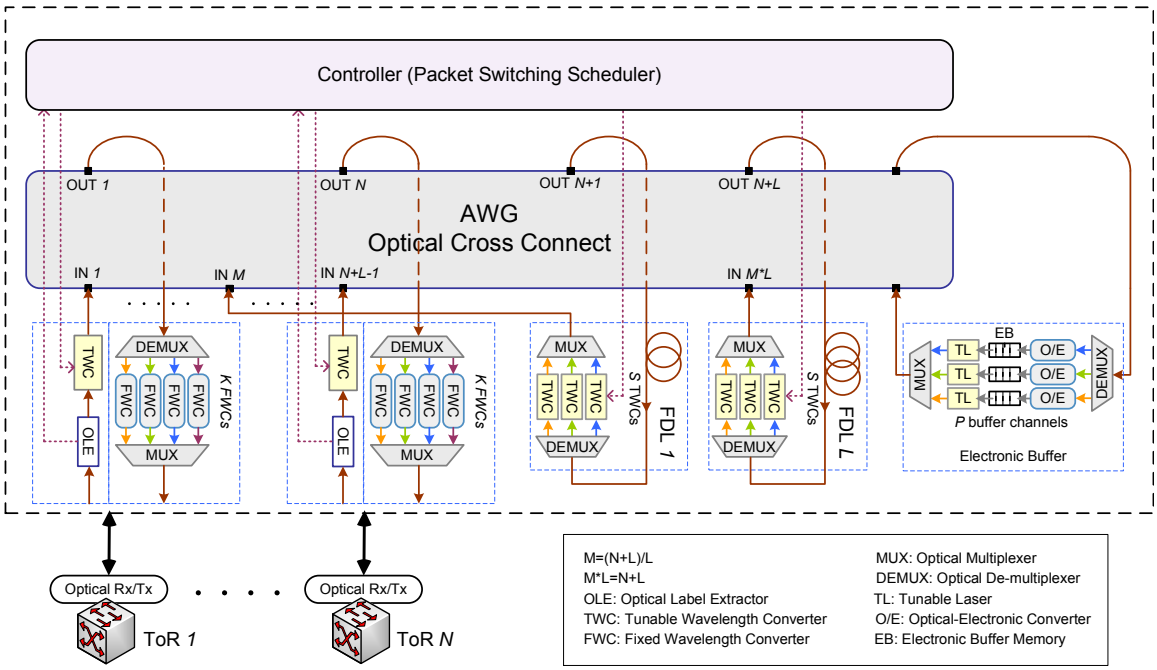


Figure 2.1: Proposed hybrid buffer $N \times N$ AWG-based packet switch architecture

Figure 2.1 shows the proposed hybrid-buffered $N \times N$ AWG switch. The switching plane is composed of N linecards, an all-optical switching matrix, a bank of shared optical buffers and a small-scale electronic buffer. Of particular note in this architecture (and discussed later) is that the re-circulating FDL fibres are returned to port numbers that are interleaved with incoming fibre ports.

Each linecard in the switch connects to a bidirectional optical link and includes three main functional modules: an Optical Label Extractor (OLE), an input Tunable Wavelength Converter (TWC) and an output port optical Demultiplexer (DEMUX) with Fixed Wavelength Converters (FWCs). The OLE separates the incoming optical label from the packet and sends it to the controller. The controller identifies the routing information stored in the label, and then configures the TWC prior to the arrival of the data packet. Upon arrival, the data packet is directed to the destined output port through the switching matrix, by appropriate tuning of the TWC. The switching matrix is an Arrayed Waveguide Grating (AWG) router, which allows multiple different wavelengths from multiple inputs to be switched to an output simultaneously, and thus itself is a non-blocking wavelength

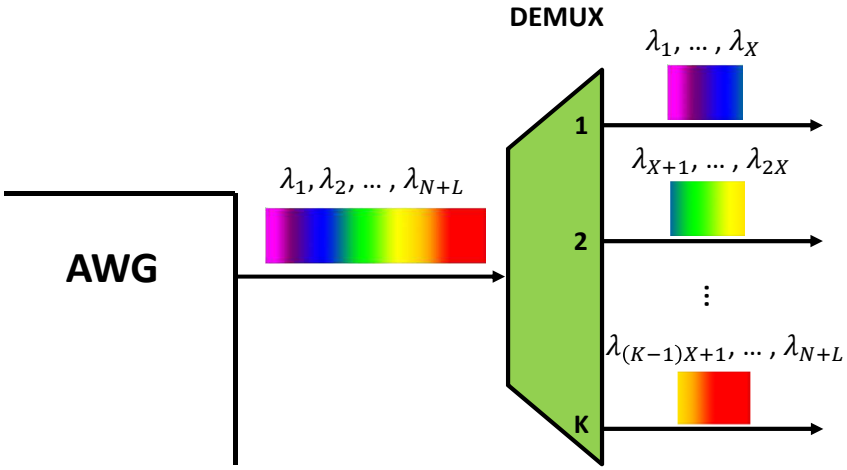
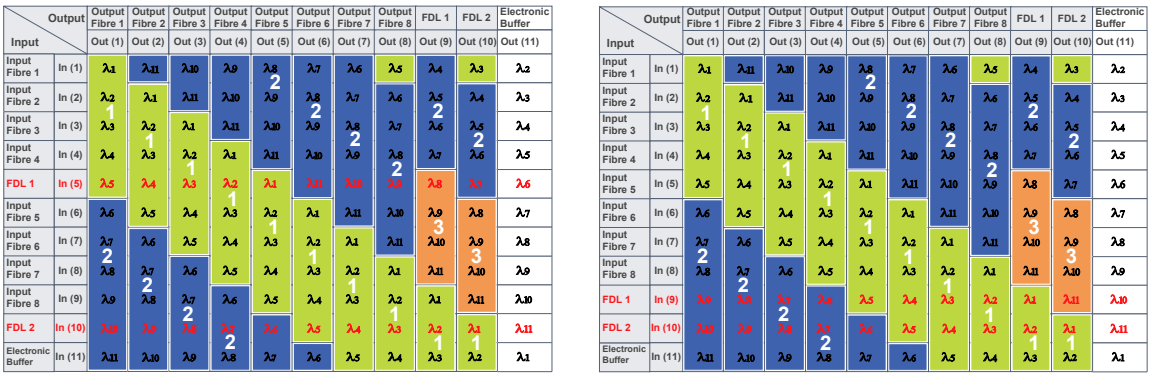


Figure 2.2: Waveband partitioning. The WDM signal at an AWG output port consists of $(N + L)$ wavelengths, each coming from an AWG input. The optical DEMUX divides the WDM signal into K continuous wavebands, each of which contains $X = (N + L)/K$ wavelengths.

router. At the output side, the wavelength division multiplexed (WDM) channel from the AWG output port is directed to the optical DEMUX in the linecard, which partitions the WDM channel into K continuous wavebands, as illustrated in Figure 2.2. Each waveband contains $(N + L)/K$ wavelengths, and is delivered to a Fixed Wavelength Converter (FWC) which converts the input wavelengths to a fixed output wavelength for transmission, see Figure 2.1. This can potentially cause collisions between wavelengths in the waveband, as only one wavelength can be carried by the FWC at any one time. To resolve conflicts, a contention resolution policy is required to buffer and separate contending packet transmissions in time. In regards to the basic AWG architecture and wavelength routing scheme as described here, this is identical to the previously proposed DOS architecture [3]. What is significantly different is the architecture and operation of the proposed hybrid optical/electronic buffer, as detailed next. Optical FDLs are implemented in a feed-back configuration with each of L FDLs having S TWCs in its return port. Similarly to an output port, each FDL buffer is equipped with a $1 : S$ optical DEMUX which divides the whole waveband into S subsets. Each waveband is composed of $(N + L)/S$ wavelengths, and fed into a dedicated TWC. The optical buffer return ports to the AWG are evenly interleaved with the input port fibre connections, that is, the AWG input port numbers connected to FDLs, denoted by $M_k, k \in \{1, 2, \ldots, L\}$, are a multiple of M, where $M = (N + L)/L$, so that $M_k = kM$.

To explain the benefits of the above described *spread* FDL connection, the waveband partition scheme of a 8×8 AWG switch with 2 FDLs is shown in Figure 2.3(a). The parameters K and S are taken as $K = 2$ and $S = 3$. This means that the waveband at the switch output is divided into



(a) OPS with spread FDLs

(b) OPS with non-spread FDLs

Figure 2.3: Waveband assignment of a 8×8 AWG switch with 2 FDLs. That is, $N = 8$ and $L = 2$. K and S are set to be $K = 2$ and $S = 3$. (a) OPS with spread FDLs. (b) OPS with non-spread FDLs.

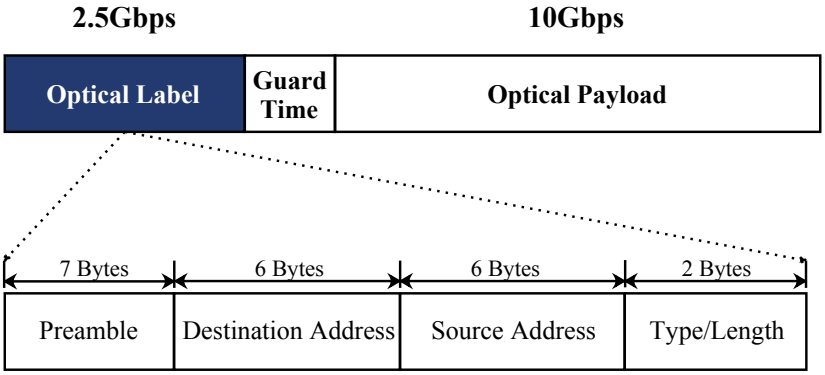
two subsets and the waveband in the FDL is divided into three subsets. For comparison, Fig. 2.3(b) depicts the waveband partition of the *non-spread* connection scheme where FDLs are connected to a contiguous set of AWG inputs. It can be seen, comparing the figures, that with the proposed spread connection, see Figure 2.3(a), optical signals coming from different FDLs to an output port are grouped into different wavebands, which in turn allows these signals to be scheduled concurrently, as they request different transmission channels (FWCs). This, on one hand, facilitates the resolution of contentions, on the other hand, balances the retrial traffic among K output FWCs, thus enhancing the resource utilisation. In experimental section of this chapter, the exact gain of interleaving FDL buffers in this way is reported in detail. Numerical results show that the interleaved optical buffering is significantly more effective than the non-spread FDLs.

Fibre delay lines may carry multiple channels, but have a limited number of associated TWCs (S) to redirect delay packets back to the intended output port, thus packet contention can additionally occur in the FDLs. To achieve a fully contention-free switch, the switch is configured with a low-port count electronic buffer which sustains the rest of the contention traffic. In the presented architecture, the optical buffer is preferred to resolve contention and handles the majority of the overflow traffic.

The logical structure of the required electronic buffer is also shown in Figure 2.1. There are N buffer inputs, fed from a $1:N$ DEMUX, so that packets from all switch inputs can be detected concurrently. Electronically delayed packets are stored in Random Access Memory (RAM), having P virtual input queues with each input queue offering buffering to N/P wavelengths. As will be shown through simulations, due to the primary buffering function being performed by the FDLs, the

required throughput of the electronic buffer is reduced by several orders of magnitude compared to an electronic-only buffer, which allows P to be small compared to N and the required bitrates of the P switch inputs to be relatively low. This makes the electronic buffer in the hybrid buffering scheme easily realisable as a low power-consumption component of the architecture. The estimations of the comparative power consumptions are presented later in this chapter.

2.2.2 Label Processing



Header Fields	Functions
Preamble	1. synchronization 2. indicates the start of an optical packet
Destination Address	identification of the destination node
Source Address	identification of the source node
Type/Length	1. protocol type 2. payload length

Figure 2.4: The Ethernet-based optical packet format

The format of an optical packet is illustrated in Figure 2.4, which consists of header bits and a payload. The header contains scheduling and routing information, including preamble, source address, destination address, protocol type and payload length, while the payload delivers the data information. The optical label is transmitted at a data rate of 2.5Gbps and the payload data rate is 10Gbps. The operating principle here is to use a low-speed label signal to carry the header, while maintaining high-speed optical packet switching, so that optical labels can be extracted and processed with

lower speed electronics. There are two main techniques to optical label coding: bit-serial labelling and subcarrier multiplexing [11, 12]. In bit-serial labelling, the optical label and the payload are encoded on the same wavelength and transmitted serially. A guard time is inserted between the bit-serial label and the data packet so as to separate these two components, see Figure 2.4. Suppose that the average payload size is 625 Bytes, then the control signal overhead is estimated to be around 12%. In subcarrier multiplexing, the optical label is modulated on a separate wavelength, thereby the label and the payload can be multiplexed into one optical signal [12]. This multiplexing allows the label information and the user data to be transmitted in parallel. Using this scheme, all-optical extraction of the control packet from the data packet can be realised by employing an optical passive filter [13, 14].

Figure 2.5 indicates that in the proposed switch architecture, after label extraction, the header packet is subsequently routed to the control plane which performs both header processing and control signal generation. Once the control unit decodes the label packet and schedules data transmission, control signals are generated to configure the switch components. Meanwhile, the data packet is traversing through an input fibre of fixed length. In doing so, an extra delay is generated to account for the label processing time and the switch setup time. In the TWC, the wavelength of the data packet is tuned to an alternative wavelength for optical switching.

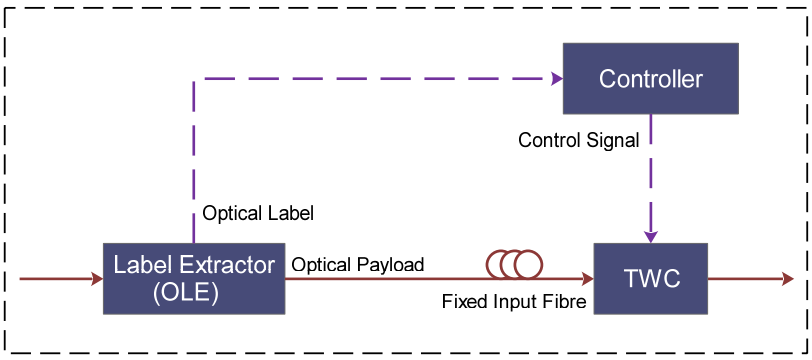


Figure 2.5: Label separation and processing

2.2.3 Scheduling Algorithm

In the proposed optical switch, the scheduling algorithm assigns output port and FDL resources in a Pre-Reservation (PreRes) manner. That is, the output FWCs and FDL channels are both reserved prior to the arrival of the optical data packet. Figure 2.6 describes the main steps of the scheduling algorithm implemented here. An arriving packet to the switch finding its designated output FWC free will occupy the FWC. Otherwise, it is blocked from the output port and requires a delay to

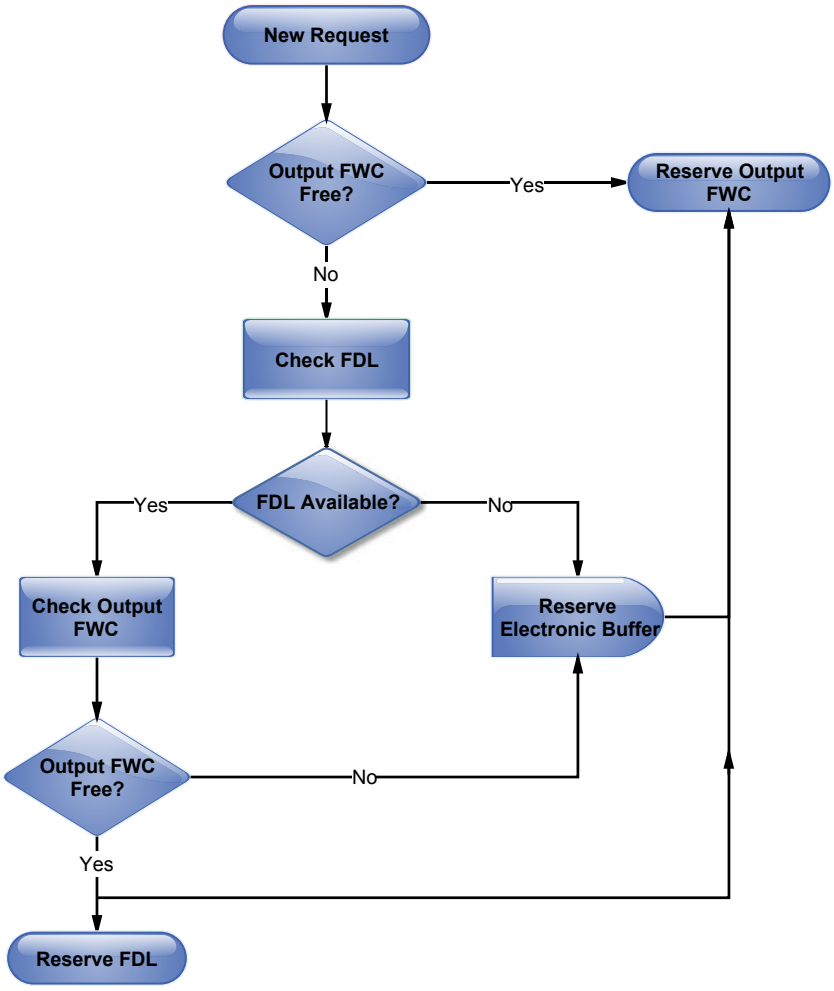


Figure 2.6: Scheduling/reservation procedure

address the contention. The scheduler firstly determines if an FDL buffer can resolve the contention. FDLs of different lengths are available and each FDL can carry multiple wavelengths accessible by appropriate tuning control of input port TWCs. A packet may circulate only once in the FDL block. If there exists an FDL delay that allows the contending packet to be scheduled in a future conflict-free timeslot of the destined output FWC, then the FDL channel (TWC) and the future timeslot of the output FWC are reserved for the incoming optical packet. However, if FDL and output FWC reservation fails, the packet is instead stored in the electronic buffer. This scheduling technique significantly enhances the resource utilisation, because the reservation task is performed before the packet enters the switch so that it does not occupy additional resources. Alternatively, as the FWCs at switch outputs operate independently due to waveband partitioning, the reservation process of each FWC is executed independently, resulting in reduced complexity of the switch architecture. In Section 2.3.2.7, the complexity of the scheduling procedure just described is investigated.

2.3 Simulation and Numerical Results

In this section, the main objective is to evaluate and analyse the influence of various network parameters on the performance of the AWG switch so as to allow optimal dimensioning of the switch hardware components. This is followed by a detailed investigation on power efficiency of the considered architecture. Finally, the complexity of channel scheduling is examined.

2.3.1 Simulation Setup

The network is simulated using a model developed with the Discrete Event Simulator OMNeT++ [15]. OMNeT++ is an event-driven C++ simulation tool which supports Object-Oriented Programming (OOP). That is, it allows users to define and create special-purpose modelling functions for the component modules. These fundamental component modules are subsequently connected and assembled into an OMNeT++ simulation network which accurately models network devices, link connectivity, network topology, operating systems and applications. According to [16], OMNeT++ is a flexible, repeatable, scalable and reliable simulation framework. These attributes make OMNeT++ a suitable simulation tool for the purpose of designing, building, and testing network architectures, and hence it is used in this chapter to model the proposed optical switch shown in Figure 2.1.

The simulation model used in this section is illustrated in Figure 2.7, where an $N \times N$ Optical Packet Switch (OPS) interconnects N Top-of-Rack (ToR) switches via high-capacity optical links. Figure 2.8 shows the structural models of the ToR switch and the OPS fabric. The developed OMNeT++ model simulates both the traffic generation and detection behaviours, and the OPS switching and buffering functions. More specifically, the framework characterises the following optical packet node, network and scheduling features:

- user-defined optical packet format containing routing details such as source, destination and packet duration,
- traffic generation process configurable for each ToR switch,
- up to N edge nodes (ToR switches) connected to the OPS via N input/output WDM fibres,
- configurable number of incoming/outgoing transmission channels per fibre,
- an incoming packet aggregator,

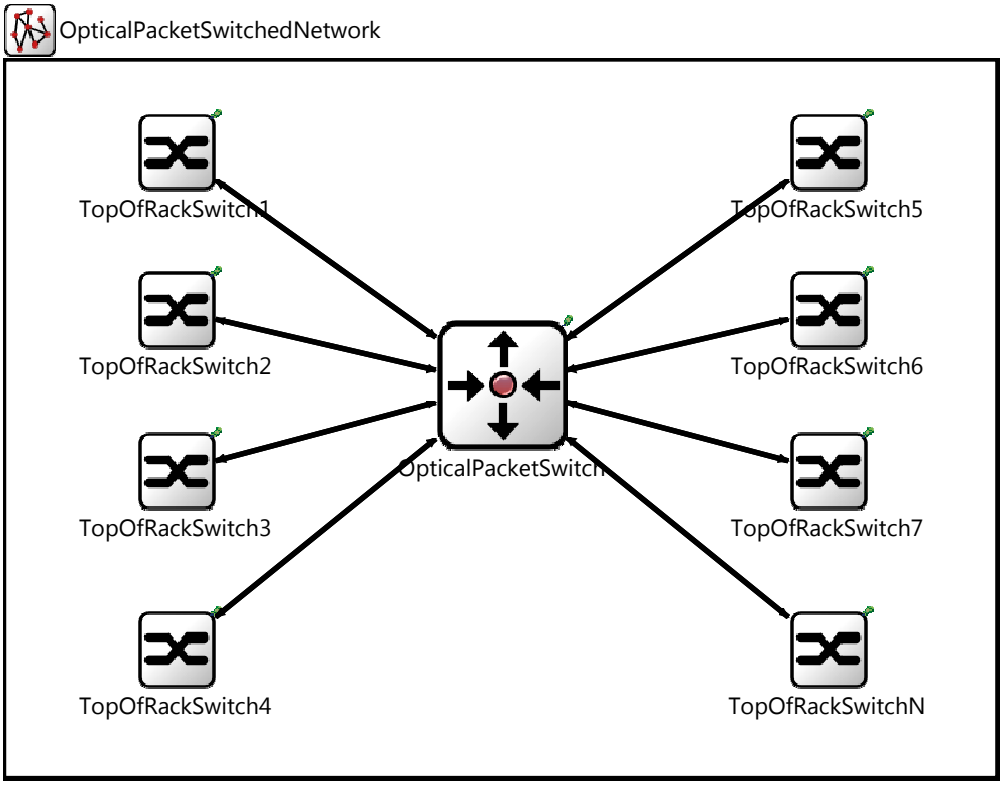
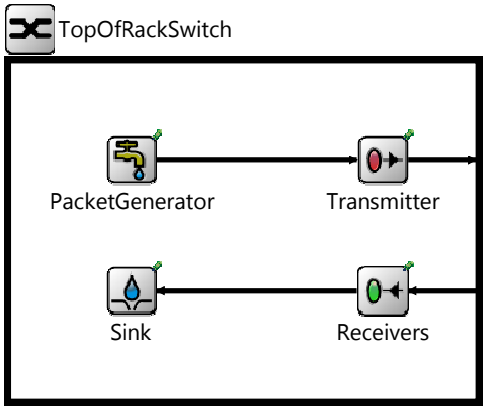


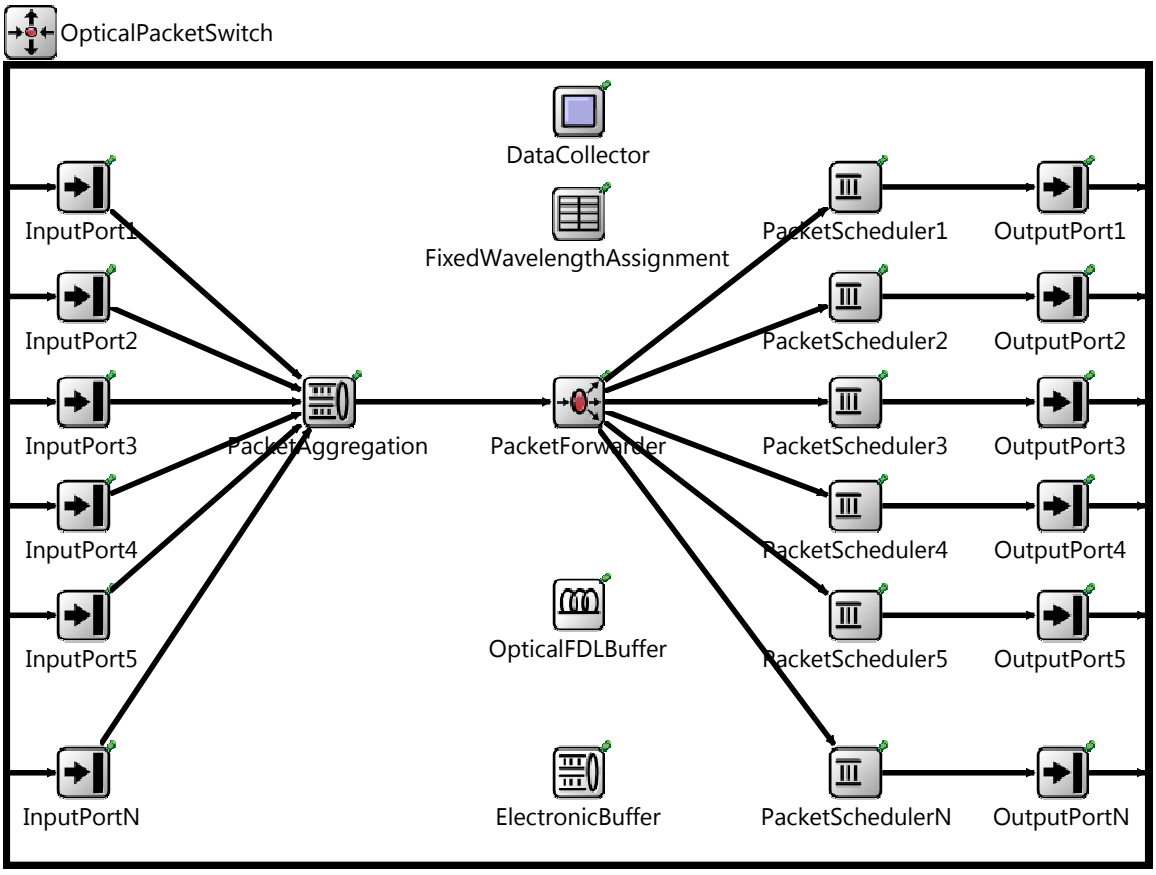
Figure 2.7: Simulation setup for an $N \times N$ optical packet switching network

- a forwarding engine, easily configured from user-specified routing table,
- configurable traffic load matrix which allows load-balancing routes,
- the Latest Available Unused Channel with Void Filling (LAUC-VF) scheduling technique,
- configurable share-per-node feedback optical Fibre Delay Line (FDL) buffers,
- configurable electronic buffers,
- data collector which records numerical results and statistical properties.

As shown in Figure 2.8(a), the Packet Generator module generates the traffic flows and populates the header fields of the optical packets. The traffic generation process is defined by users. That is, the packet transmission time is randomly selected according to the user-defined packet length distribution and packet length mean. Alternatively, the inter-arrival times between the packets are determined by the provided inter-arrival time distribution and the mean inter-arrival time. The destination node of a new packet is randomly chosen from all destinations following a uniform distribution. This means that a packet from this source node is equally likely to be addressed to any



(a) Simulation model of the Top-of-Rack switch



(b) Simulation model of the Optical Packet Switch

Figure 2.8: Simulation models of the ToR switch and the core OPS switch

of the N edge nodes. Subsequently, the generated packet is transmitted towards the core OPS by the Transmitter which is connected to the bi-directional fibre. The Receiver block detects and reads the incoming optical packets, and forwards them to the Sink module which destroys the packets.

In the OPS simulation model, see Figure 2.8(b), the simulator receives the traffic flows coming from the ToR switches, and directs them to the Packet Aggregator block. The Packet Aggregator collects the incoming packets and sends them to the Packet Forwarder Module. By looking up its forwarding table for each packet destination, the Packet Forwarder routes the packet to the Packet Scheduler module associated with its destined switch output port. As illustrated in Figure 2.8(b), each switch output is equipped with a separate Packet Scheduler which keeps track of all previously scheduled packets on all outgoing transmission channels in the output port. The Packet Scheduler block is essentially an OMNeT++ queue with one sub-queue assigned to each transmission channel. The scheduled data is stored (inserted) as user-defined schedule packets, each containing the start and end time of a packet transmission on the outgoing channel. To conserve memory in the simulator, an interrupt, scheduled for the end of the packet transmission time, causes the expired scheduling packet to be removed from the sub-queue. Sub-queues are kept sorted (prioritised) in order of packet arrival time so that removal of the correct scheduling packet is simple. When a new packet arrives to the Packet Scheduler, the start and end transmission time for the packet is calculated from the current simulation time and the packet duration value contained in the header packet. The scheduler then determines the required output transmission channel of the packet, which is assigned based on the fixed bandwidth partitioning results stored in the Fixed Wavelength Assignment module, and attempts to reserve this transmission channel. This process reads the current state of the required transmission channel in the output port (as stored in the queue of scheduling packets) and, if possible, schedules the arriving packet as specified by the field values in the packet header. The scheduling algorithm implemented by the Packet Scheduler is the Latest Available Channel with Void Filling (LAUC-VF) algorithm. The scheduler stores all voids, and choose between available voids i.e. the void which minimises the time between start of packet and end of previously scheduled packet is chosen. If the output channel reservation fails, an attempt to resolve the contention using an FDL is made (if switched on). In this case, the scheduler attempts to find the shortest available FDL that can accommodate the packet and then attempts to schedule the output transmission channel in order to accommodate the delayed packet. If no suitable output channel is found at this stage, then the next shortest FDL is examined, etc. The multi-wavelength FDL buffers are also modulated as an OMNeT++ queue with multiple sub-queues. The scheduler uses the sub-queues to reserve the corresponding FDL wavelength channels. This is done in a similar manner described

for the output channel scheduling processes above by inserting/removing scheduled packets to keep track of the occupancy of its delay line input port. If the FDL buffering is unable to resolve the contention, the scheduler then reserves the electronic buffer and the corresponding output transmission channel. This scheduling procedure is referred to as the Pre-Reservation (PreRes) void-filling scheduling scheme, as described in Section 2.2.3. Finally, the resultant statistical characteristics and numerical results are recorded by the Data Collector module.

Table 2.1: Table of Notation

N	Number of incoming/outgoing fibres in OPS. It also represents the number of the ToR switches in the network.
K	Number of transmission channels (FWCs) per output
L	Number of Fibre Delay Lines (FDLs)
S	Number of channels (TWCs) per FDL
D	Unit delay in FDL buffers. The discrete set of delays generated by FDLs are integer multiples of D, i.e., D, $2D$, ..., LD
λ	Packet arrival rate of the Poisson process
T	Mean packet transmission time
μ	Packet service rate, defined by $1/T$
ρ	Mean input traffic load generated by the ToR switch. The value is given by λ/μ
r	Coefficient of burstiness of the IPP
C	Coefficient of variation of the IPP, which is expressed as the ratio between the variance and the squared mean of the inter-arrival times.
P_{loss}	Packet contention probability in the optical network
D_{avg}	Average end-to-end latency in the optical network
D_{fdl}	Average delay imposed by FDLs in the optical network
D_e	Average electronic buffering delay in the optical network
T_e	Throughput of the electronic buffer

The detailed notation relating to the model is presented in Table 2.1. N represents the number of incoming/outgoing fibres connected to the AWG-based OPS. Each input fibre carries one wavelength channel operating at a data rate of 10Gbps, while the outgoing fibre accommodates K 10Gbps transmission channels (FWCs). L is the number of FDLs utilised, and each FDL has S TWCs in its return port so that it may simultaneously carry S different channels. Note that the degenerate FDLs are employed, that is, the available delay times are integer multiples of a base delay time D. In addition, the mean input traffic load generated by the packet generator inside a ToR switch is denoted by ρ.

In the experiments, two different input traffic processes are applied: the Poisson Process and the Interrupted Poisson Process (IPP). The Poisson process is considered to be of smooth type, whereas IPP represents a bursty traffic process. In the Poisson process, inter-arrival times of optical packets

are exponentially distributed with an average of $1/\lambda$, λ being the mean arrival rate. The mean packet duration is set to be T. In IPP, the traffic process alternates between ON states and OFF states. During the on-period, the packets arrive according to a Poisson process, but upon entering the off-period, the flow is interrupted and no arrivals occur. Both the on-periods and off-periods are exponentially distributed, and they are mutually independent. The burstiness of an IPP traffic is determined by two parameters: *coefficient of burstiness*, denoted by r and *coefficient of variation*, denoted by C. Parameter r determines the relative on-period of the IPP, and C represents the variation of the inter-arrival times, which is calculated as the ratio between the variance and the squared mean of the inter-arrival times. Note that when $r = 1$, the IPP becomes exactly a Poisson process.

In this section, the packet loss rate, P_{loss}, is computed as the ratio between the number of dropped packets and the total number of generated optical packets during a given time period. Average end-to-end delay, denoted D_{avg}, is defined as the average latency over all optical packets, whether they require buffering or not. FDL delay, denoted D_{fdl}, is the average delay imposed by FDLs. Similarly, electronic delay, D_e, is calculated as the average delay through the electronic buffer. It should be pointed out that the constant propagation delay within the switch is not considered here. Additionally, the electronic buffer throughput, T_e, is described as the average amount of traffic, in bits per second, that traverses the electronic buffer.

2.3.2 Simulation Results and Discussions

Experimental results presented in this section are all attained by using the simulation model described in Section 2.3.1. The subsequent simulations are performed on switch scenarios with $N = \{64, 256, 512\}$, so as to draw conclusions on how switch performance scales. The analysis starts with a preliminary investigation of the network performance of the AWG switch with only FDL buffers. In this case, the main performance metric is the switch blocking probability P_{loss}. As a benchmark, the contention performance of the switch with no buffering is first evaluated, as shown in Figure 2.9. Simulation results indicate that the bufferless switch suffers from severe packet loss, which is around 10%-20% under high loads. This is expected given the large over-subscription from N input traffic streams competing for a small number of output transmission channels K.

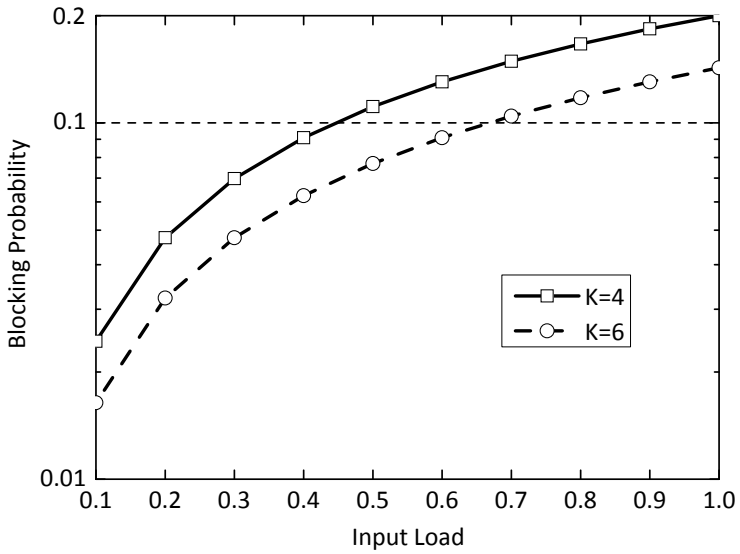
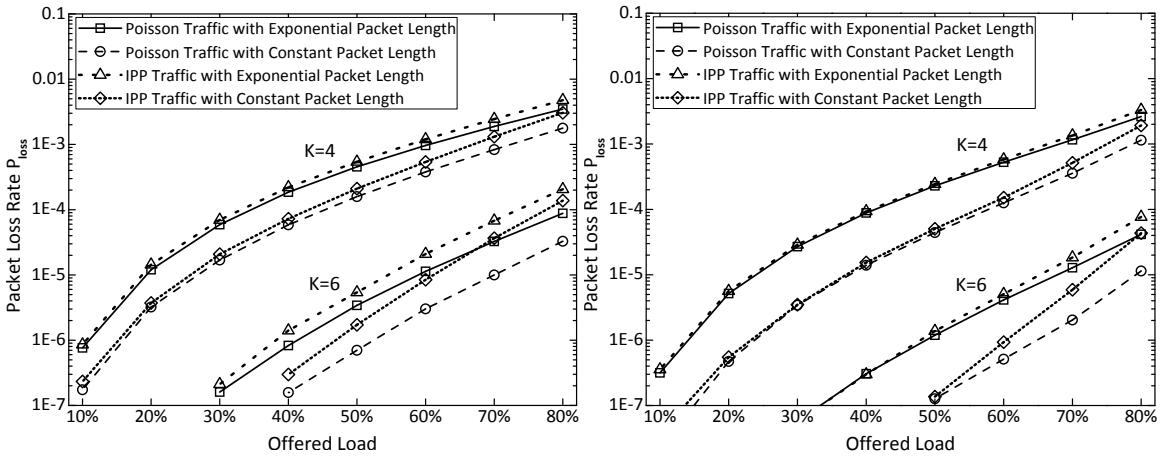


Figure 2.9: Contention performance of the bufferless switch.

2.3.2.1 Impact of Input Traffic Process

To investigate the impact of the packet arrival process on network performance, extensive simulations have been carried out to measure the performance behaviours of the proposed optical-buffered switch with both Poisson arrivals and IPP arrivals under varying load conditions. For the IPP traffic model, the parameter r is set to be 0.5, and C to be 6. Figure 2.10 plots the blocking probability of the switch architecture with four different traffic arrival processes: (1) Poisson traffic with exponential packet length, (2) Poisson traffic with constant packet length, (3) IPP traffic with exponential packet length, and (4) IPP traffic with constant packet length. Evidently, for all traffic types, the introduction of FDLs reduces the packet loss rate substantially, compared to the bufferless switch scenarios shown in Figure 2.9, even under high loads, from approximately 10% of the switch offered load to less than 0.01%. This indicates that FDLs can efficiently resolve the contentions and carry the majority of the contention traffic. Additionally, increasing K lowers the contention probability significantly. This is expected as increasing the number of transmission channels at the output would alleviate the congestion conditions in the switch. Furthermore, it is noticed that the IPP traffic causes higher blocking probability than Poisson traffic, suggesting that the contention performance degrades with the increase of the burstiness of the input traffic. Another important point to note from Figure 2.10 is that the overall performance behaviours of the four traffic arrival processes exhibit a similar pattern. Owing to this property, in the rest of this chapter, only Poisson traffic with exponential packet length is considered, which is a classical traffic model commonly used in performance analysis.



(a) $N = 64$, $L = 8$, $S = 6$

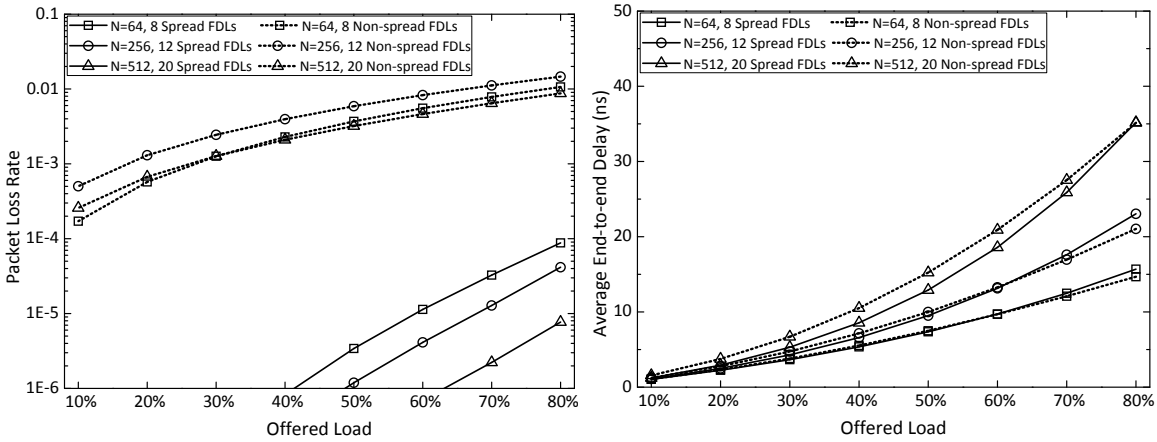
(b) $N = 256$, $L = 12$, $S = 10$

(c) $N = 512$, $L = 20$, $S = 10$

Figure 2.10: Packet drop ratio P_{loss} versus input load for two values of K under four different input traffic processes: (1) Poisson traffic with exponential packet length, (2) Poisson traffic with constant packet length, (3) IPP traffic with exponential packet length, and (4) IPP traffic with constant packet length.

2.3.2.2 FDL Port Interleaving

In this section, the emphasis is on determining the benefits of the proposed spread FDL connection where FDL returning ports are evenly interleaved with input fibres, as illustrated in Figure 2.3(a). This is achieved by analysing the performance quantities of the switch, measured in terms of the contention probability and the average end-to-end delay. Extensive comparisons have been conducted between spread FDL connection and non-spread FDL connection where FDLs are connected to a set of successive inputs, see Figure 2.3(b). In this set of simulations, the values of K and D are fixed to be $K = 6$ and $D = 0.1T$. Figure 2.11 plots the network performance under a wide range of load conditions in different switch configurations. Obviously, having FDLs evenly spread yields



(a) Packet loss-rate comparison - spread versus non-spread FDLs.

(b) Latency performance comparison - spread versus non-spread FDLs.

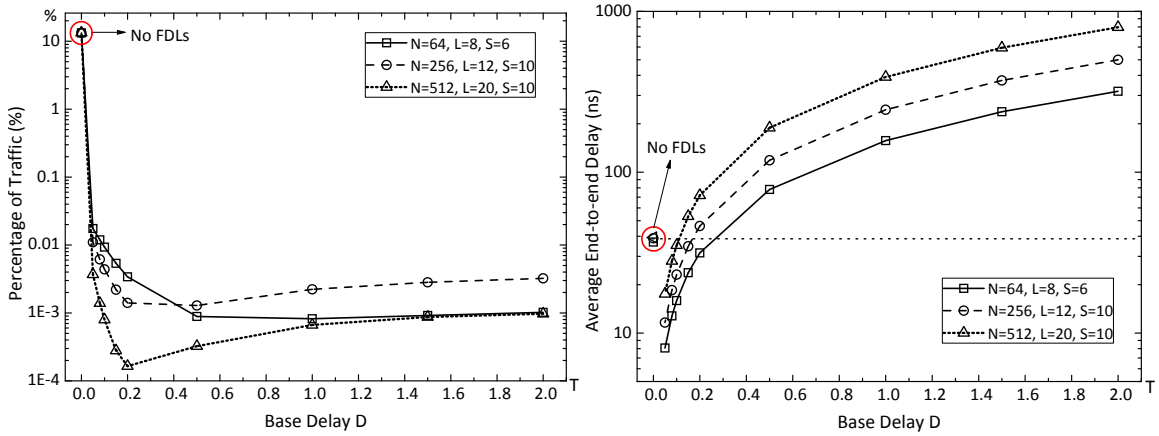
Figure 2.11: Performance comparison between spread and non-spread FDL schemes as a function of the offered load for scenarios with fixed parameter values of $K = 6$ and $D = 0.1T$.

a significant reduction in packet loss rate, by more than two orders of magnitude, without inducing additional delays. This is due to the fact that with the spread connection, the contended packets returning from different FDLs are scattered into different output port wavebands and have access to multiple FWCs. It verifies that the proposed technique makes more efficient use of available buffering resources.

2.3.2.3 FDL Base Delay D

For the remainder of the simulation results, the inclusion of an electronic buffer is considered, giving the full hybrid switch architecture of Figure 2.1. As stated previously, in the hybrid electronic/FDL buffer, FDLs are utilised in conjunction with the electronic buffer, thereby the performance of FDL buffers has a significant impact on the requirements for electronic buffer, as well as the overall network performance. One important parameter that affects the performance of optical buffers is the FDL base delay D. To explore the sensitivity of the network performance to D, the following set of simulations is carried out, with parameter K being 6 and input load ρ being 80%.

The plots of Figure 2.12(a) show the % of offered traffic that overflows from FDLs. As previously described, the overflown traffic from FDLs will be handled by electronic buffer, thus the numerical results presented in Figure 2.12(a) also represent the electronic buffer throughput. It can be seen from Figure 2.12(a) that for a given N, L, and S, the overflows from FDLs first reduce with the increase in D, but beyond a certain value of D, they tend to increase as D continues to increase. This is not typical of a shared FDL architecture (with void-filling scheduling, as used here) but is



(a) Input traffic to electronic buffer versus base delay D, presented as the percentage (%) of switch total offered load.

(b) Average system delay D_{avg} versus base delay D.

Figure 2.12: Performance measurements of AWG switch with hybrid buffer as a function of FDL base delay D for three values of N. Parameter K is 6, and input load ρ is 80% of maximum load. The 'No FDLs' point, when $D = 0$, means that only electronic buffer is employed for buffering.

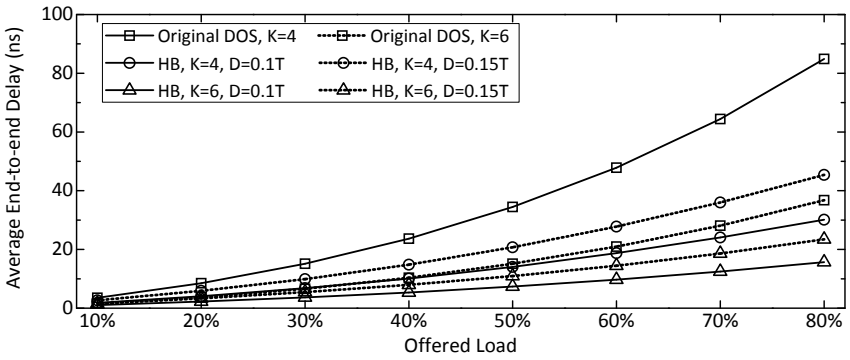
due to the output port and FDL waveband partitions and the interaction between electronic and FDL buffering in this case.

Figure 2.12(b) depicts the average system delay as a function of base delay parameter D for the three different switch sizes. As expected with the growth of D, the packet delay increases beyond a few packet lengths. For all three scenarios tested, in order to maintain a mean latency less than or close to the value that is attained for electronic-only buffer (no FDLs), the base delay parameter should be less than $0.15T$. Additionally, to keep the overflow traffic (equal to electronic buffer throughput) at a low level (referring to Figure 2.12(a)), D is chosen as $0.1T$ or $0.15T$. In the following sections, the two values will be used, in certain results, to examine sensitivity of the system's performance to a small change in D around its nominal design value.

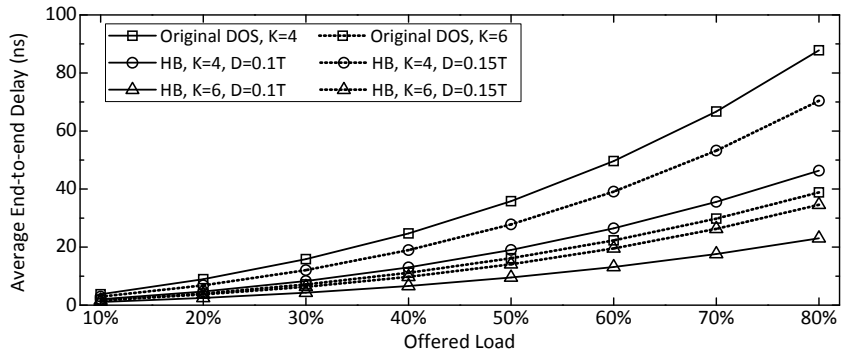
2.3.2.4 Hybrid Switch Latency

Figures 2.13(a)-2.13(c) show packet latency properties of the hybrid buffer versus electronic-only buffer for three different values of switch size N, under a wide range of system loads. The three figures exhibit almost the same pattern, except that, in the case of $N = 64$ and $N = 256$, the electronic buffer yields the worst performance, whereas, for $N = 512$, the hybrid buffer (when $D = 0.15T$) exhibits higher latency. In addition, the best delay performance is gained by the hybrid switches with $D = 0.1T$. Considering the top three curves at load 80% in Figure 2.13(c), it is observed that

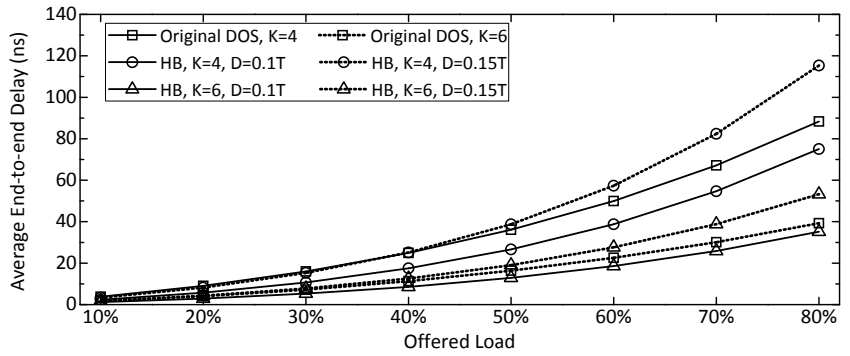
when moving from $D = 0.15T$ to $0.1T$, the end-to-end delay varies from 115ns to 75ns, which is 15% lower than the measured delay for the electronic-only buffer. However, as presented in Figure 2.12, larger values of base delay allow FDLs to grasp more contention traffic, which reduces requirements of the electronic buffer. This is a trade-off one needs to consider in dimensioning. It is fair to conclude that the hybrid buffering strategy can yield very good performance, when the base delay time D is appropriately chosen.



(a) $N = 64, L = 8, S = 6$



(b) $N = 256, L = 12, S = 10$



(c) $N = 512, L = 20, S = 10$

Figure 2.13: System delay versus offered load for three different configurations of switch size N and FDL size L and S, with two values of K and base delay parameter D in each configuration.

2.3.2.5 Dimensioning the FDLs

A final, but primary, factor determining performance of the hybrid buffered switch is the number of FDLs installed. A case study is conducted on three switch scenarios: (a) $N = 64, S = 6$, (b) $N = 256, S = 10$ and (c) $N = 512, S = 10$, fixing parameter $K = 6$, $D = 0.1T$ and $\rho = 80\%$. The impact of L on latency and electronic buffer throughput is examined in detail. It is additionally shown that, with all things considered, the hybrid buffer can considerably lower energy consumption.

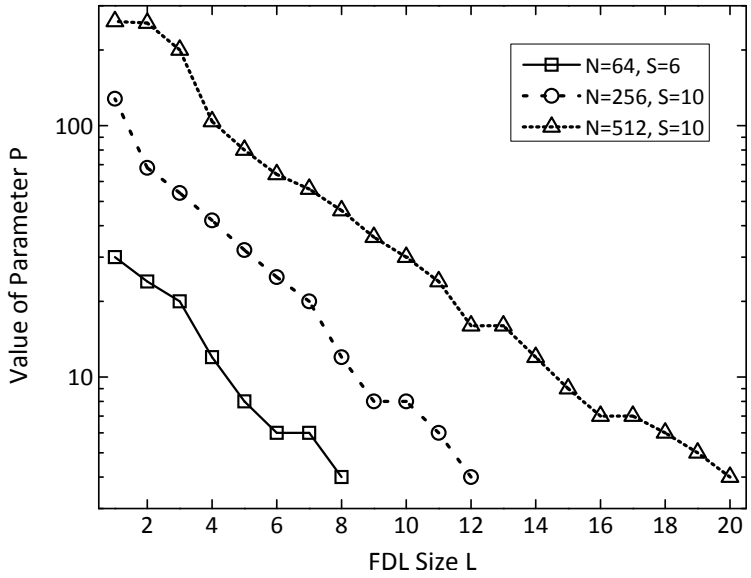


Figure 2.14: Appropriate values of P for different numbers of FDLs L, for different switch sizes N. The value of P for a given L corresponds to the maximum number of simultaneously busy electronic buffer outputs at 80% load, $K = 6$, $D = 0.1T$.

In deciding L, it is important to determine an appropriate dimension for the electronic buffer size P, as they together determine mean buffering latency. To this end, P is first set to be N, and then, by simulations, the maximum number of output TWCs in the electronic buffer that are simultaneously busy, are estimated as a function of L. Figure 2.14 shows the results for the three different switch sizes. Thus, these plots give an estimate of the minimum number of TWCs required (P) for a given L. If P is set less than the value indicated, electronic buffer latency will increase. If P is set greater than necessary, switch energy consumption will increase, without a large expected improvement in latency.

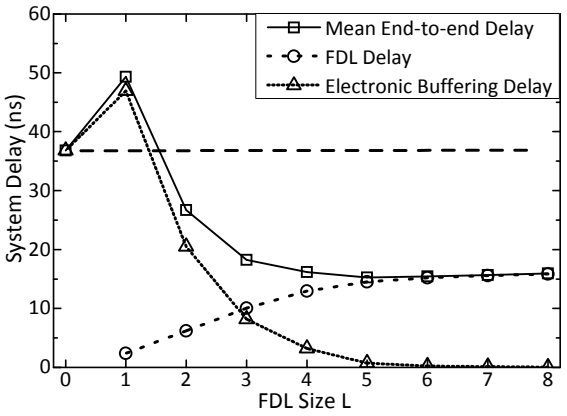
Note that this method of setting P is approximate as it only takes into account how latency changes as the electronic buffer size P is scaled, but not the effect of the necessary change to the buffering

structure feeding the cross-connect (c.f. Figure 2.1). When P is set to some given value less than N in a design, this determines the value of M_u of the $M_u : 1$ multiplexers which in turn determines the amount of contention for each of the P virtual queues. This contention is not reflected in Figure 2.14 where $P = N$ and there is no multiplexing contention. Nonetheless, Figure 2.14 gives a viable and simple way to estimate a suitable value of P given a design point for L and throughout the remainder of the results, the value of P is set based on the estimations in Figure 2.14. The effect of the approximation appears in Figure 2.15 and is discussed below.

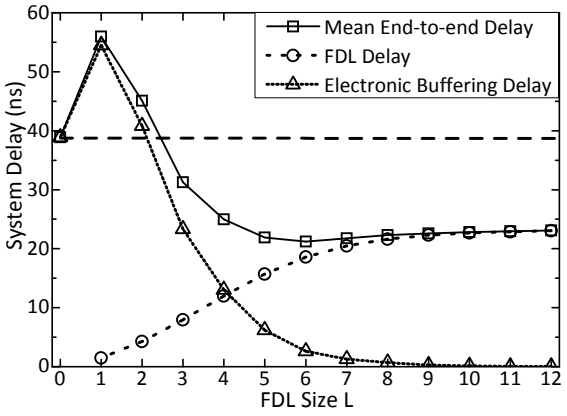
In Figure 2.15, mean latencies are shown versus the number of FDLs L, for different values of N. Initially, when the first FDL is added to the switch, latency is seen to increase. This is due to the previously mentioned effect of multiplexing multiple contending packets from different inputs into the same virtual queue in the electronic buffer, which consequently imposes additional waiting time. When $L = 0$, each of N cross-connect inputs has a dedicated uncontended input. When $L = 1$, there is only one FDL and each queue is fed by (on average) two competing switch inputs. This causes longer queues in the electronic buffer as the dominant source of latency, as indicated by the individual latency plots for electronic and FDL buffers in Figure 2.15, supported by the throughput plots of Figure 2.17.

When more than one FDL is added ($L > 1$), the benefits of FDL buffering begin to dominate and latency reduces (Figure 2.15). Perhaps counter to intuition, as L increases the latency actually reduces below that of the $N \times N$ electronic-only buffer. Intuitively, mean system latency should grow as the number of FDLs L grows, given that the FDL base delay (delay of shortest FDL) is fixed and the longest available FDL delay thus increases with L, in addition to the fact that generally FDLs make less efficient buffers due to their fixed delay times. Overall, the latency reduces with increasing L, at least up to a point. This situation can be explained by considering the relative percentages of contended traffic resolved by the electronic buffer versus the FDL buffers in Figure 2.16. When there are few FDLs, there is considerable load on the electronic buffer which dominates system latency with long queues. As the number of FDLs is increased, this relative overload of the electronic buffer is alleviated as more overflow load is carried by the adequate FDL buffering capacity. Eventually, as the number of FDLs increases further, latency begins to increase as is more normally expected for FDLs. In all, what is interesting is that FDLs, when combined with an electronic buffer, can in fact reduce system latency in an AWG-type architecture, which reveals the benefits of the proposed hybrid buffering technique, compared to electronic buffers.

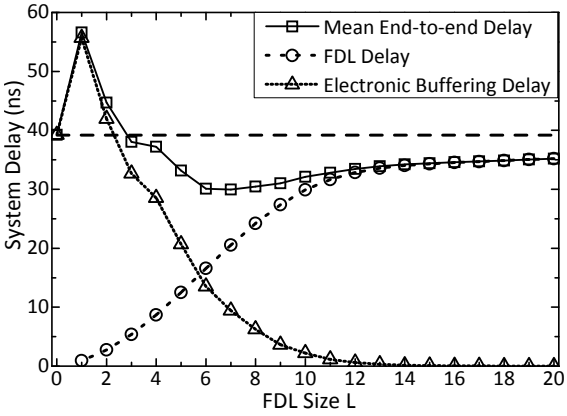
Figure 2.17 importantly indicates how varying the number of FDLs affects the electronic buffer



(a) N=64, S=6

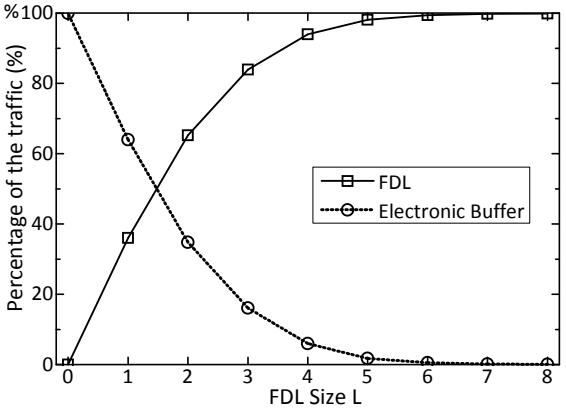


(b) N=256, S=10

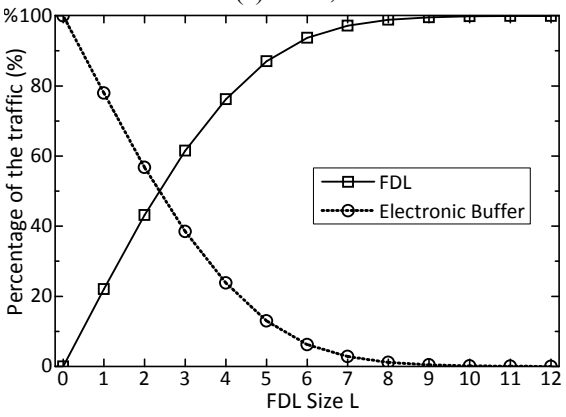


(c) N=512, S=10

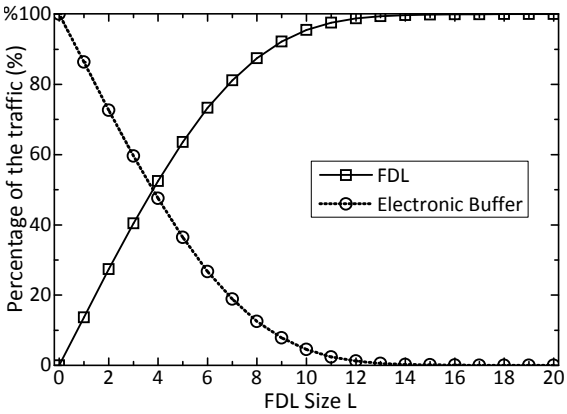
Figure 2.15: Average packet delays versus FDL size L. P is set for each different value of L according to Figure 2.14. The dashed straight lines represent the results for switch architecture with only electronic buffer.



(a) N=64, S=6



(b) N=256, S=10



(c) N=512, S=10

Figure 2.16: The percentage (%) of contention traffic carried by two different buffers. Parameters $K = 6$, $D = 0.1T$, $\rho = 80\%$. P is set for each different value of L according to Figure 2.14.

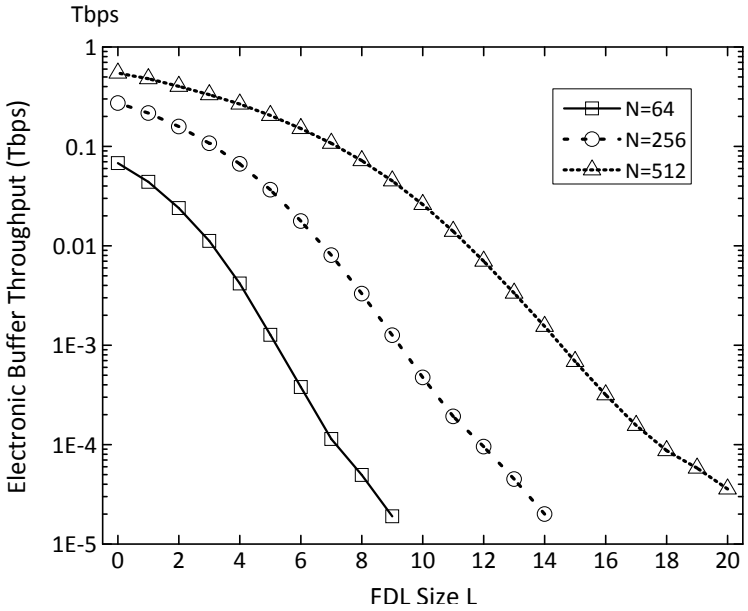


Figure 2.17: The electronic buffer throughput versus FDL size L in three switch configurations: (1) $N = 64$, (2) $N = 256$, and (3) $N = 512$. Parameters $K = 6$, $D = 0.1T$, $\rho = 80\%$.

throughput. When there are no FDLs, or a very small number, the throughput of the electronic buffer is extremely high, up to around 0.5Tbps in the case of $N = 512$. To support such traffic requirements, a sophisticated electronic buffer with high capacity, high datarate, heavy parallelization and low processing delay is required, which is costly and has relatively high energy consumption relative to the rest of the switch architecture. However, in the proposed architecture with just a modest number of FDLs, a much smaller electronic buffer is sufficient to support the remaining buffered traffic, reducing to less than 100Mbps total throughput. This downsizing of the electronic buffer contributes to a significant improvement in overall energy efficiency of the switch, as discussed in the next section.

2.3.2.6 Switch Power Consumption

Having investigated performance characteristics, this section focuses on the assessment of energy consumption of the switch architecture under study. In Section 2.3.2.5, the analysis shows that in the hybrid-buffered architecture, even for modest numbers of deployed FDLs, the required number of electronic buffer TWCs can be reduced substantially, compared to the number that needs to be deployed in an all-electronic architecture (always N TWCs). Overall, the reduction in the required electronic buffering gives a reduction in switch power consumption, as estimated as follows.

It is important to point out that the estimations in this section do not include the power consumption of the control unit, and to simplify the calculations, only the major consumers of the power, including TWCs, FWCs, tunable lasers and the electronic buffer/switch, are considered here. [17] indicates that a 10GBASE SFP+ transceiver consumes 1W power. Additionally, the supply power of the TWC is set to be 1.5W [18], FWC to be 0.6W [19, 20] and O/E/O to be 0.8W [20]. Further, according to the data in [21–23], it is assumed that the electronic buffer is essentially a forwarding engining plus buffer (router) with a 2.5W/port power consumption.

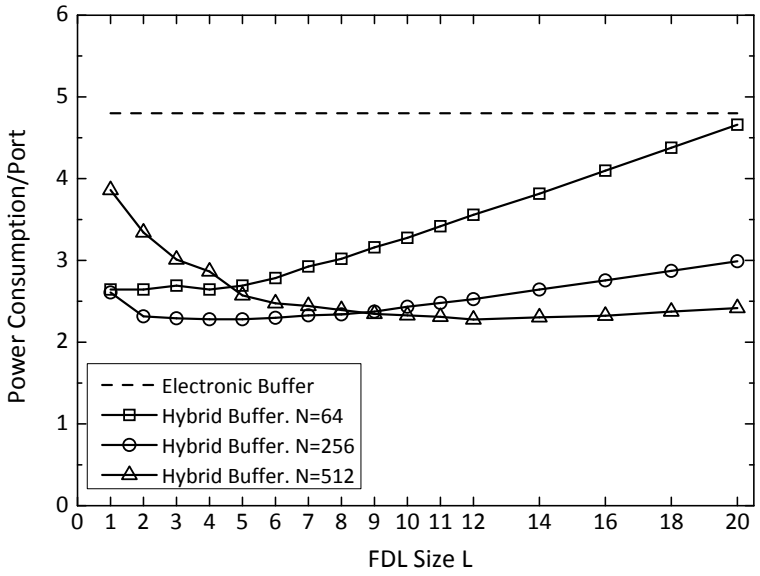


Figure 2.18: The power efficiency is estimated as a function of parameter L in three different scenarios : (a) $N = 64, S = 6$ (b) $N = 256, S = 10$ (c) $N = 512, S = 10$.

With these numbers, the energy efficiency of the studied architecture is computed as a function of FDL size parameter L for different values of N and plotted in Figure 2.18. Notably, the proposed architecture tends to be considerably more energy efficient than the all-electronic buffering scheme, primarily due to the reduced power supply to the electronic buffer. It is clear to see from the figure that, for a given N, beyond a certain value of L, denoting it L_{const}, there will be no further improvements in energy efficiency. According to Figure 2.18, L_{const} is taken as 6 for $N = 64$, 8 for $N = 256$, and 12 for $N = 512$. With L set to be L_{const}, Table 2.2 presents a comparison between electronically-buffered and hybrid-buffered architectures in the three different scenarios. It is observed that the necessary number of the various devices in the electronic buffer case grows linearly with N. For the hybrid buffer case, the required number of tunable transmitters does not change much with switch size. Additionally, the number of input-tunable-output-tunable TWCs deployed in the FDLs increases sub-linearly with switch size N. Furthermore, the buffer power per

port is reduced substantially, and so is the total power consumption per port, by as much as 46%. These findings suggest that the proposed buffering scheme makes the AWG-based switch highly scalable in terms of power dissipation.

	$N = 64$		$N = 256$		$N = 512$	
	Electronic Buffer	Hybrid Buffer	Electronic Buffer	Hybrid Buffer	Electronic Buffer	Hybrid Buffer
FDL Size L	-	6	-	8	-	12
FDL TWC ($L \times S$)	-	36	-	80	-	120
Tunable Tx ($= P$)	-	6	-	12	-	16
E/O Tx	64	-	256	-	512	-
O/E Rx	64	64	256	256	512	512
Buffer Power (W)	211	82.2	844.8	214.8	1689.6	357.6
Buffer Power/Port	3.30	1.24	3.30	0.82	3.30	0.68
Power/Linecard	2.5	2.5	2.5	2.5	2.5	2.5
Total Power/Port	5.80	3.74	5.80	3.32	5.8	3.18

Table 2.2: The power consumption comparison between different buffers.

2.3.2.7 Scheduling Complexity

This section presents a detailed analysis of computational complexity of the scheduling/reservation algorithm in the proposed architecture. Two types of reservation scheme are examined: horizon and void filling. The horizon scheduler does not utilise gaps between previous reservations on output port or optical buffer channels, thus the scheduling is simple and fast. In contrast, void filling (VF) keeps tracks of all void intervals and schedules the incoming payload on the idle gaps [24]. The experiments are performed on the network implementation of a 256×256 switch with 8 FDLs and a 12×12 electronic router ($P = 12$) as presented in Table 2.2. The input traffic load ρ is set to 80%. It is worth mentioning that the computational times are obtained with reference to profiling of the execution of the actual scheduling algorithms running in the simulator, on a single thread of an Intel i7-2600 processor.

For comparison, the scheduling times in both the hybrid buffer scenarios and the electronic-only buffer case are examined. Figure 2.19 plots the packet header loss rates due to late scheduling with respect to different computation time limits (equal to input FDL delay) of the scheduler, when the controller processor has insufficient processing power. Further, it is estimated that the maximum computation time for LAUC-VF algorithm is around 2.4μs, for horizon algorithm is 1.8μs, and for the electronic-buffer only case is 1.5μs. As expected, the computational power required to perform LAUC-VF scheduling is significantly more than for others, as intensive void checks are performed

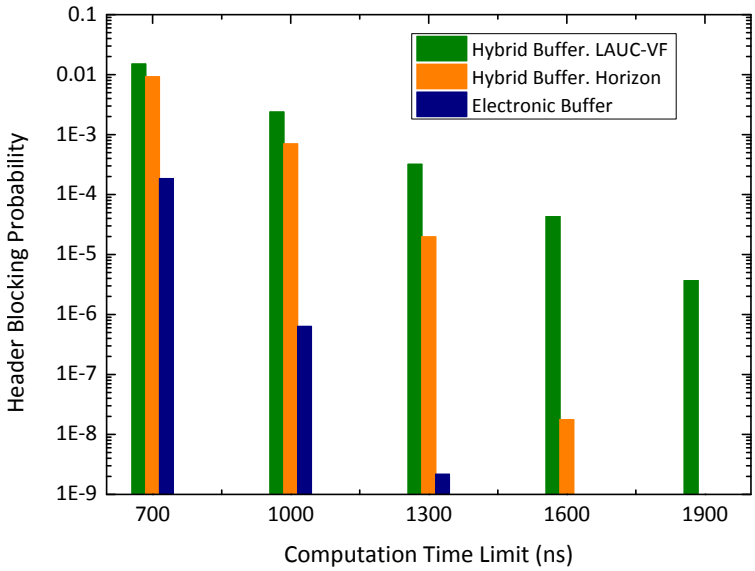


Figure 2.19: The packet header blocking probability due to late scheduling versus computation time limit. In hybrid buffer scenarios, the worst-case computation time for LAUC-VF algorithm is around 2.4μs and the computation time for horizon algorithm is 1.8μs. The maximum computation time for electronic-buffer only scenario is 1.5μs

on delay fibres and output ports. In contrast, the electronic-only buffer requires the least processing time, despite that all electrically buffered packets having the same target output port will always request the same output FWC, and consequently, the controller performs intensive gap checks on one FWC, at the cost of higher power consumption in the data switching plane.

The computational analysis shows that the horizon algorithm requires less computation time than LAUC-VF algorithm in the proposed hybrid buffer. The question then arises as to whether the performance differences between these two algorithms are significant. From Figures 2.20 and 2.21, it is observed that, provided there are sufficient optical buffers, there is little difference in optical buffering capacity and packet latency performance between void filling and the simpler horizon for this particular switch architecture, due to the traffic balancing action of the optical buffer arrangement. Note that the packet latency of either algorithm is significantly lower than an all-electronic buffer which has a mean buffering latency of 292ns. This prompts the conclusion that overall the proposed hybrid-buffered switch implementing horizon scheduling algorithm is preferred.

It is worth mentioning that FPGA or GPU implementations, as opposed to serial code considered here, could substantially enhance the processing speed, and thus reduce the computation time of the scheduling algorithm significantly. This is because a large degree of parallelism is achievable as each output port can be scheduled independently.

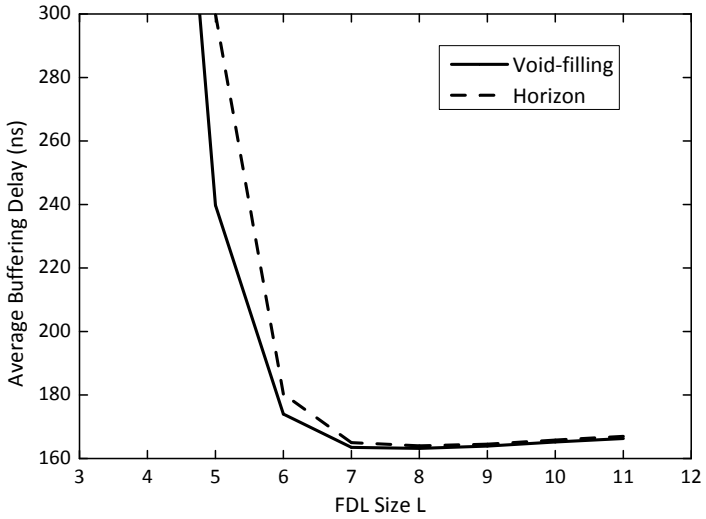


Figure 2.20: The average latency versus FDL size L when P is fixed as 12.

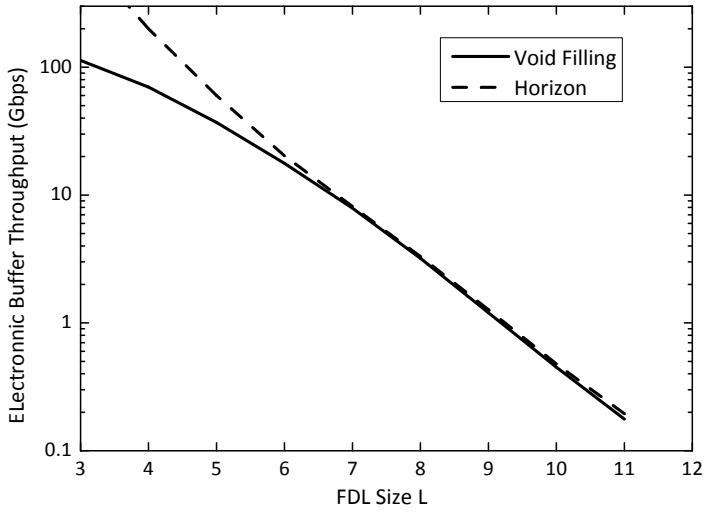


Figure 2.21: The electronic buffer throughput versus FDL size L when P is fixed as 12.

In support of the above code profiling results, it is necessary to analyse the theoretic worst-case computational complexity of scheduling, which occurs when N inputs send optical packets to the same output. As each output has K transmission channels working independently, each transmission channel receives N/K packets. In scheduling, only one packet will reserve the link successfully, and the other $N/K-1$ packets need to be buffered. For the horizon algorithm, the worst-case scenario is that the $N/K-1$ packets check L FDLs and output channels, resulting in a complexity of O(NL/K). For the void-filling (VF) algorithm, the 1st packet is scheduled immediately, the 2nd packet needs buffering and checks at least 2 voids, the (N/K)-th packet checks at least N/K voids. $(1+2+3+\ldots+N/K) = \text{O}((N/K+1)*N/K/2) = \text{O}(((N/K)^2)/2+N/K/2)$, thus the approximate complexity is O$((N/K)^2)$. It should be noted that the complexity of the

VF controlling algorithm is known to be O (log$a$), where a is the number of the void checks on a channel [25].

2.4 Conclusions

The subject of this chapter is to investigate and simulate AWG-type switches with different implementations of contention resolution: electronic buffering, FDL buffering and a hybrid electronic/FDL buffering scheme. The performance analysis shows that the proposed method of interleaving FDLs into the AWGR input ports can provide low packet loss rates. Further, combining FDLs and electronic buffering together yields a lossless switch with good latency performance and power consumption, in comparison to standalone use of either of the two buffering strategies. Alternatively, additional experiments were carried out to evaluate the computational complexity of void filling (VF) and horizon algorithms in the presented switching architecture, and the results indicate that the simpler horizon algorithm outperforms the VF algorithm in terms of processing power requirements, throughput and latency, thereby the horizon scheduling is more suitable in the proposed switch. These findings indicate the feasibility of realising a high-port count, high-capacity optical switch for data centre or HPC deployments.

Although a line bitrate of 10Gbps has been assumed in the simulations, with the advent of 100Gbps optical channels, the results presented are expected to remain valid, assuming packet sizes will scale accordingly (c.f. 9000 byte 'Jumbo' frame sizes of Gigabit Ethernet) and so packet transmission time (and appropriate FDL lengths) remains approximately constant. With this increase in bit rate, a reduction of the already high throughput of the electronic buffer, by means of hybrid buffering, becomes even more important.

In the analysis, the values of parameters S and K are decided through simulation experimentation, so as to produce illustrative results, but these values are not optimised. This is due to the fact that solving the dimensioning problem of the proposed switch architecture through simulations would require an impractically large number of long-duration simulation runs. All dimensioning parameters (N, K, L, S and P) are inter-dependent and have a complex relationship with switch latency and power consumption and an optimisation would need to consider all parameters at once. This is a non-convex, multi-objective optimisation problem, which will be resolved using mathematical modelling in the next chapter.

Bibliography

[1] S. Cheung, T. Su, K. Okamoto, and S. J. B. Yoo, "Ultra-compact silicon photonic 512×512 25 GHz Arrayed Waveguide Grating Router," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 4, pp. 310-316, July 2014.

[2] B. Mukherjee, "Optical WDM Networks", Springer, pp. 96-98, 2006.

[3] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS - a scalable optical switch for datacenters," in *ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, pp. 1-12, Oct. 2010.

[4] X. Ye, V. Akella, and S. J. B. Yoo, "Comparative studies of all-optical vs. electrical vs. hybrid switches in datacom and in telecom networks," in *Optical Fiber Communication Conference and Exposition, and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, pp. 1-3, Mar. 2011.

[5] X. Ye, R. Proietti, Y. Yin, S. J. B. Yoo, and V. Akella, "Buffering and flow control in optical switches for high performance computing," *Journal of Optical Communications and Networking*, vol. 3, no. 8, pp. 59-72, July 2011.

[6] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. J. B. Yoo, "LIONS: an AWGR-based low-latency optical switch for high-performance computing and data centers," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 2, Apr. 2013.

[7] S. Pallavi and M. Lakshmi, "AWG based optical packet switch architecture," *MECS*, pp. 30-39, Mar. 2013.

[8] S. Pallavi and M. Lakshmi, "An AWG based optical router," in *International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 245-248, Feb. 2014.

[9] R. Srivastava and Y. N. Singh, "Feedback Fiber Delay Lines and AWG based optical packet

switch architecture," *Journal of Optical Switching and Networking*, vol. 7, no. 2, pp. 75-84, Apr. 2010.

[10] VaibhavShukla and R. Srivastava, "WDM Fiber Delay Lines and AWG based optical packet switch architecture," in *National Conference on Innovative Trends in Computer Science Engineering (ITCSE)*, pp. 47-49, Apr. 2015.

[11] E. Udvary and T. Berceli, "Optical Subcarrier Label Swapping by Semiconductor Optical Amplifiers," Journal of Lightwave Technology, vol. 21, no. 12, pp. 3221-3225, Dec. 2003.

[12] K. M. Sivalingam and S. Subramaniam, "Emerging Optical Network Technologies: Architectures, Protocols and Performance," Springer US, 2005.

[13] M. T. Hill, A. Srivatsa, N. Calabretta, Y. Liu, H. D. Waardt, G. D. Khoe, and H. J. S. Dorren, "1×2 optical packet switch using all-optical header processing," *Electronics Letters*, vol. 37, no. 12, pp. 774-775, 2001.

[14] N. Calabretta, H.-D. Jung, E. Tangdiongga, and H. Dorren, "All-optical packet switching and label rewriting for data packets beyond 160 Gb/s," *IEEE Photonics*, vol. 2, no. 2, pp. 113-129, Apr. 2010.

[15] Discrete Event Simulator, OMNeT++. http://www.omnetpp.org.

[16] OMNeT++ User Manual. [Online]. Available: https://omnetpp.org/doc/omnetpp/Manual.pdf.

[17] Cisco 10GBASE SFP+ Modules, Datasheet, Cisco Inc., 2010.

[18] C. Kachris and I. Tomkos,"Power consumption evaluation of all-optical data center networks," *Cluster Computing*, vol. 16, no. 3, pp. 611-623, Sept. 2013.

[19] N. Bikash, J. Sangirov, M. R. Uddin, and Y. H. Won, "Efficient approach towards energy minimized optical network". [Online]. Available: https://www.itu.int/dms_pub/itu-t/oth/06/2F/T062F0010000004PDFE.pdf.

[20] R. S. Tucker, "The role of optics and electronics in high-capacity routers," *Journal of Lightwave Technology*, vol. 24, no. 12, pp. 4655-4673, Dec. 2006.

[21] Arista 7050S-64 Switch, Datasheet, Arista Networks, Inc., 2011.

[22] Arista 7250QX-64 Switch, Datasheet, Arista Networks, Inc., 2013.

[23] Arista 7300 Switch, Datasheet, Arista Networks, Inc., 2013.

[24] P. K. Chandra, “Survey on optical burst switching in WDM networks,” in *4th International Conference on Industrial and Information Systems (ICIIS)*, pp. 83-88, Dec. 2009.

[25] V. M. Vokkarane, G. P. V. Thodime, V. U. B. Challagulla, and J. P. Jue, “Channel scheduling algorithms using burst segmentation and FDLs for optical burst-switched networks,” in *IEEE International Conference on Communications (ICC)*, pp. 1443-1447, 2003.

Chapter 3

Analytic Modelling and Dimensioning of a Datacentre Optical Packet Switch with Recirculating Optical Buffers

3.1 Introduction

Wavelength-routed optical packet switching, which exploits the functionalities of advanced optical switching components such as high port-count Arrayed Waveguide Gratings (AWGs) and Tunable Wavelength Converters (TWCs), represents an attractive candidate to address current and future application needs in large-scale datacentre interconnection networks, due to advantages of bandwidth capacity, switching speed and operational flexibility [1–3]. A major challenge associated with this networking concept is the effective handling of contention resolution. In the previous chapter, a hybrid electronic/optical buffer, where multi-wavelength Fibre Delay Line (FDL) buffers are deployed in combination with an electronic buffer, is developed to resolve congestion in the wavelength-routed optical switches. Differently, this chapter proposes an efficient AWG switch architecture with no electronic buffering, that uses recirculating optical buffers of variable lengths. This feature of the architecture allows the contending packets to be recirculated in optical buffers multiple times until the requested switch outputs become free. Nevertheless, there is no guarantee the contention will be resolved and some level of packet blocking (loss) occurs, as the FDL buffering provides limited buffering capacity and a limited choice of delay times. Thus, it is of paramount importance that this type of optical switch is efficiently dimensioned such that a wide variety of

applications and services can be supported with the required Quality-of-Service (QoS).

A popular strategy for addressing the network design challenge is through mathematical network performance modelling [4, 5]. The principal purpose is to determine the required minimum network resources to fulfil the traffic demands without overloading the network infrastructure and degrading the quality of the service provisioning. In this sense, network dimensioning can be formulated as a constrained optimisation problem which is quantified as a function of several network design parameters, with the design/operating cost being the objective function. The ultimate goal is to allocate an optimal network configuration. The concept of optimality of a network configuration is essentially a balance between the power-related operating cost and the network performance. To enable this balance, a proper assessment of the trade-off between energy usage and network QoS performance is required. Understanding the network and investigating the impact of network parameters on QoS measurements like congestion and communication latency is thus clearly necessary. Motivated by this, analytical modelling is developed, which is a powerful tool for performance analysis, as it numerically characterises the basic behavioural processes in the network, such as the switching, buffering and scheduling behaviours etc., and allows the derivation of mathematical expressions for the performance characteristics, which in turn allows a detailed theoretical analysis of the network performance against various network design parameters. A major requirement for the analytical modelling is that the mathematical framework provides accurate estimations with reasonable computation complexity, so that it can be extended to resolve the network dimensioning problem.

3.1.1 Related Work

The performance analysis methods applied to optical packet switching are generally based around models relating to the G/G/N/N blocking system, a system with N channels (switch output ports) with no queueing and/or the G/G/N queueing system, where an unlimited queue length approximates a limited number of FDLs, or some variation such a re-trial queueing model approximating a combined switch and FDL buffer. The simplest approximation techniques model the switch and FDL buffers as a single lumped queueing model and with simple assumptions about the traffic characteristics. For example, in [6] the performance of an Optical Burst Switch (OBS) with buffers is modelled using the M/M/N model, i.e. Poisson arrivals and exponential packet lengths assumed. A similar approximation is made in [7] for OBS with FDLs. Other examples are [8] and [9], where approximate models of optical packet switches with FDLs are developed. In [10, 11], the

authors consider models combining multple queueing models with more generally distributed arrivals, that is, Gamma-distributed inter-arrival times, in the queueing analysis, and the numerical results suggest that the variance of the traffic has a significant influence on the network performance and must be considered. The authors further propose an approximate two-moment traffic model in [12] to estimate the contention performance of OBS switches with feed-back FDLs. In other authors' work, to capture the correlation and burstiness in the arrival traffic, [13] and [14] develop a multi-server queueing model based on the Markov-modulated Poisson Process (MMPP), where a Markov-modulated mathematical framework is developed for modelling the traffic overflow process. Along similar lines, in [15], a Markov chain model is developed to analyse the performance of optical switch architectures with shared optical buffers. In [16], MMPP-based analytical modelling is used to analyse an optical packet switch with recirculation-enabled FDL buffers. The arrival process is represented by a simple ON-OFF MMPP process, also referred to as an Interrupted Poisson Process (IPP). Based on [13] and [14], mentioned above, the authors in [17] study an optical packet switch with recirculating FDL buffers where arrivals are Poisson. In the queueing analysis, the recirculating traffic flows from the optical buffers are also assumed to be of Poisson type and the blocking probability performance in different network scenarios is evaluated.

This chapter focuses on developing an effective modelling method based on the MMPP to analyse the performance of the proposed optical packet switch with recirculating FDLs. This allows practical AWG-based switches (where the number of required transceivers is limited by cost constraints) to be dimensioned. In comparison to previous modelling approaches, the generality of the input process in the proposed model is extended, allowing it to be any MMPP process, and also the re-trial flows are modelled as MMPPs so as to enhance accuracy (rather than simplifying to the Poisson process). Additionally, a method to model different FDL lengths is incorporated in the performance modelling, which later allows the optimisation of the FDLs. Overall, the modelling methodology used is based on the fundamental methods of [13] and [14], but its application significantly extends the application of the approach in [17]. The work in this chapter also relates to the previous work in [18], which investigated a similar type of switch through simulation studies alone. Comparing analytic results to simulations, the contention probability, the overall end-to-end delay and the maximum buffering latency can be accurately estimated under different input traffic types. Furthermore, extending previous work in the literature on optical packet switch modelling, using the proposed analytic model, efficient heuristic solutions for the network dimensioning problem are developed, with a view to selecting the optimal value of network resource parameters, which ensures QoS metrics while minimising power consumption of the network infrastructure.

3.1.2 Overview of Modelling Approach

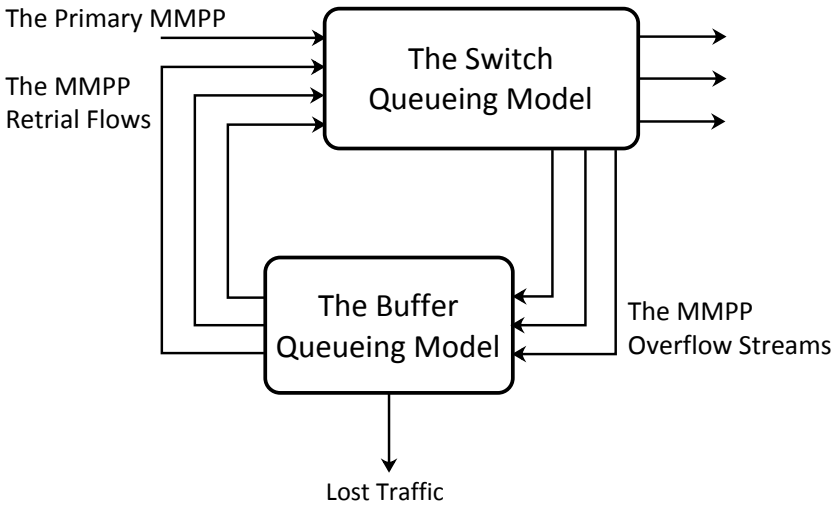


Figure 3.1: An overview of the analytical queueing model.

In overview, the switch cross-connect and optical buffers are each modelled as distinct queueing systems operating in parallel, as illustrated in Figure 3.1. The switch is approximated by a Markov-modulated finite-server queueing system with the arriving process composed of a primary MMPP traffic process and multiple MMPP-based retrial flows, from the recirculating buffers. The traffic streams that overflow from the switch queueing model (blocked traffic) are merged and characterised by MMPPs, which form the input traffic to the shared FDL buffering resource. The behaviour of the optical buffers is represented by an additional Markov-modulated finite-server queueing system. The departing processes from the FDL buffers, again modelled by MMPPs, then loop back to the switch queueing system, whereas the overflowing traffic from the optical buffers is abandoned, leading to packet loss in the system. It is important to note that the optical buffering includes multiple fibre lines, hence it can be viewed as a queueing system with multiple servers or as a cascade of multiple single-server queueing systems. For that reason, in this work, two distinct models for the analysis of the multi-channel optical buffering with L fibres are developed. More specifically, a single MMPP/M/L queueing system is suggested to approximate the behaviours of FDLs with small unit delay, whereas in the scenarios of large delay granularity, a chain of L MMPP/M/1 loss systems is proposed. The main advantage of this arrangement is that, based on the two models, numerical estimations of the performance measurements can be derived as a function of the unit delay, allowing a detailed investigation of the trade-off between the lengths of the buffers and the overall switch performance.

From Figure 3.1, it is clear that the analysis will involve an iteration around the switch+buffer loop system. Further, as FDL recirculations are allowed in the network, meaning that the contention traffic can travel through the loop system multiple times, there exists multiple retrial traffic processes, i.e. the i-th retrial flow represents the traffic that has traversed through the loop system i times. Thus, it is critical to model this network dependence and recirculation feature in order to obtain accurate approximations, which consequently adds more complexity to the modelling. The solution to the problem is to apply a recursive method. To simplify the analysis, traffic independence assumptions are made, that is, the retrial traffic processes are assumed to be independent. In this case, the overall arrival traffic to the switch model is composed of a primary MMPP arrival process and multiple independent MMPP retrial flows returning from FDL buffers. In addition, the overflowing traffic streams seen by these input processes, which are offered to the optical buffers as arrival process, are also independent of each other. Based on these premises, an iterative procedure is deployed to derive the resulting mathematical formulation, which provides a detailed analysis of the impact of the incoming traffic and switch dimensions on performance characteristics in terms of the contention probability and the communication delay.

The remainder of this chapter is structured as follows. Section 3.2 details the proposed optical datacentre network architecture. Section 3.3 reports the definition and some properties of the Markov-modulated Poisson Process (MMPP), followed by a detailed description of the proposed analytical framework for the optical switch under study. In Section 3.4, an extensive numerical analysis indicates that the developed analytical network can accurately approximate the performance measurements in terms of the contention probability and the overall delay for different input traffic types. In Section 3.5, an optimisation algorithm is developed for the network dimensioning problem, using the proposed analytical modelling technique. Some optimal solutions are presented. Finally, Section 3.6 summarises the work in this chapter.

3.2 Optical Packet Switch Architecture

Figure 3.2 details the proposed datacentre switch architecture and its buffering functionality. The switching plane consists of three main functional modules: N linecards, an optical cross-connect and a bank of shared-per-node optical buffers. Each linecard is composed of an Optical Label Extractor (OLE), a Tunable Wavelength Converter (TWC), an optical Demultiplexer (DEMUX), multiple Fixed Wavelength Converters (FWCs) and an optical Multiplexer (MUX). The OLE ex-

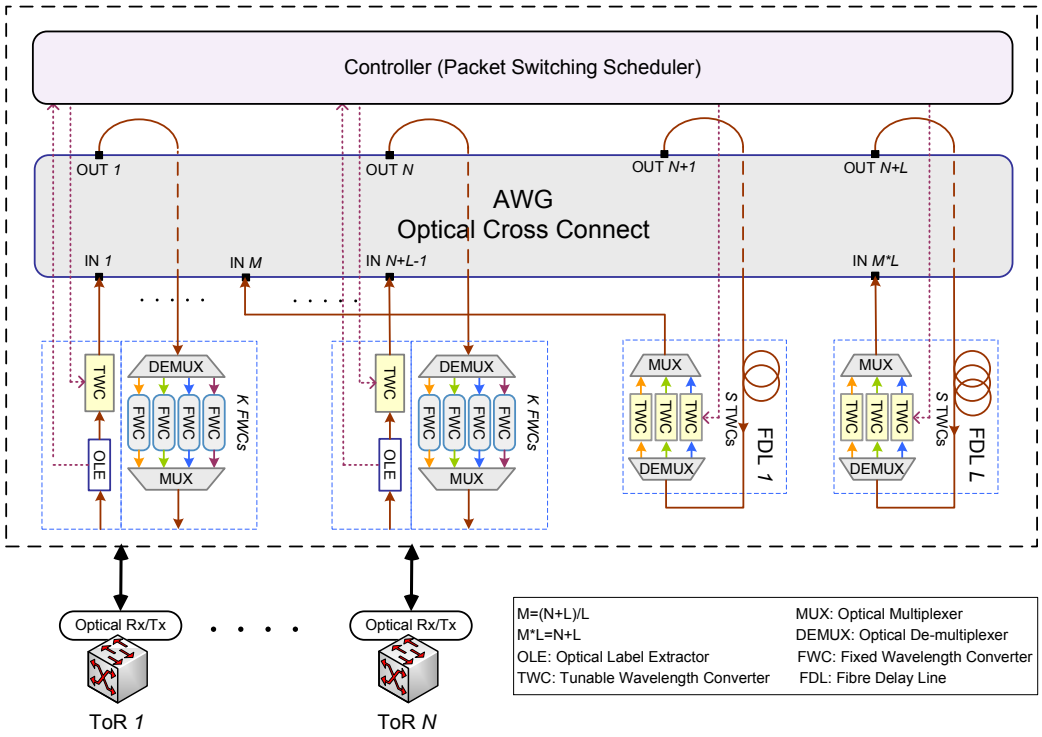


Figure 3.2: The AWG-based optical packet switch with optical FDL buffers, providing high bandwidth connectivity between N different Top-of-Rack (ToR) datacentre switches.

tracts the optical label from the arriving packet and directs it to the controller, which decodes the label and then sends a control signal to configure the TWC. By appropriate tuning of the TWC, the data packet is switched to the designated output port through the all-optical core cross-connect. The core cross-connect of the switch is an $N \times N$ wavelength-routed optical Arrayed Waveguide Grating (AWG) component, which allows multiple wavelengths from multiple different inputs to be switched to an output simultaneously. At AWG output port, the wavelength division multiplexed (WDM) channel, which comprises N wavelengths coming from N input ports, is directed to an optical DEMUX which partitions the input WDM signal into K ($<< N$) continuous wavebands. Each waveband, containing N/K wavelengths injected from N/K different switch input ports, is fed to a dedicated FWC which converts the input wavelengths to a fixed transmission wavelength. In that regard, the $N \times K$ FWCs at switch outputs operate independently with each other. Subsequently, the K converted transmission channels are multiplexed by the optical MUX for transmission. As each FWC can carry only one wavelength (one packet) in the waveband at any time, there is potential for packet collisions. At the same time, as relatively few FWCs are used ($K << N$), the architecture is efficient with FWC being well utilised. To resolve the conflicts between packets, the switch is equipped with a bank of L feed-back optical Fibre Delay Lines (FDLs), each having

S Tunable Wavelength Converters (TWCs) in its return port. Thus each FDL can carry up to S packets simultaneously. As illustrated in Figure 3.2, the returning ports of the optical FDL buffers are evenly interleaved with the incoming fibres, that is, the AWG input port numbers connected to FDLs, denoted by $M_i, i \in \{1, 2, \ldots, L\}$, are a multiple of M, where $M = (N + L)/L$, so that $M_i = iM$. This spread connection of FDLs balances the retrial traffic among K transmission channels (FWCs) at a switch output, since optical signals coming from different FDLs are grouped into different wavebands, and as such they request different FWCs (transmission channels). This arrangement significantly enhances the resource utilisation and facilitates the resolution of contentions. Delay times of FDLs are integer multiples of a base delay D. Similarly to switch output ports, in the FDL buffers the waveband is divided into S subsets, each fed into a dedicated TWC, thereby the S TWCs in an FDL also operate independently. Consequently, a packet experiencing contention at an output port only has access to one fixed TWC in an FDL, instead of S TWCs.

In the optical switch, the scheduling/reservation of output port transmission channels (FWCs) and buffering channels follows a Pre-Reservation (PreRes) strategy, which is executed by the packet scheduler in the controller. An arriving packet to the switch finding its designated output port FWC free will occupy the FWC and the corresponding transmission channel. Otherwise, it is blocked from the output port and directed to the shared bank of FDLs. In the FDL bank, the contending packet is recirculated possibly multiple times and subsequently returns to its destined output port when it has become free. However, as the FDL bank has limited buffering capacity and limited delay time, it is possible that no feasible schedule can be made for the contending packet, and consequently, the packet is lost from the network. More specifically, a contending packet can be lost due to either of two causes: (i) it finds all FDL channels busy, and (ii) it fails to reserve the destination switch output port, despite having an idle buffer channel available. Work in [18] reported a detailed evaluation of these two packet loss causes, that is, loss due to buffer congestion versus loss due to switch output congestion. In the next section, an approximate analytical model is developed to mathematically characterise the switch architecture and scheduling behaviour just described.

3.3 The Analytical Queueing Model

In the proposed optical packet switch with recirculating optical buffers of Figure 3.2, the overflowing traffic (contending packets) at the switch output ports is directed to the optical buffers, rather

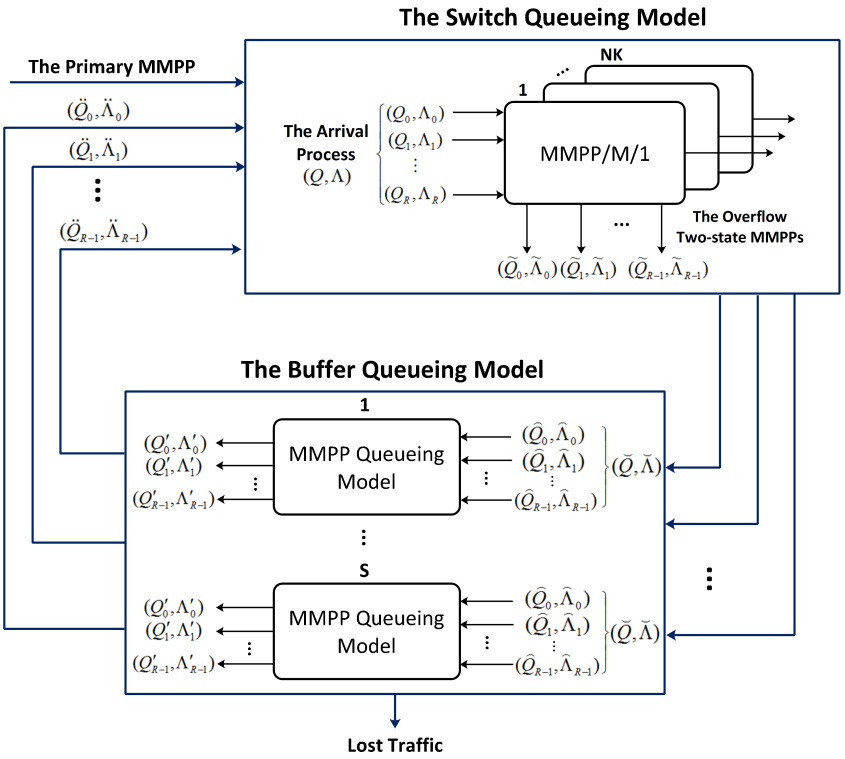


Figure 3.3: The analytical model of the proposed optical switch. The switch queueing model consists of $N \times K$ independent and identical MMPP/M/1 queues. The arrival process to the MMPP/M/1 queue is composed of a primary MMPP (Q_0, Λ_0), and R independent recirculating traffic flows (Q_i, Λ_i), $i = 1, \ldots, R$. The optical buffer queueing model is composed of S independent and identical MMPP-based queueing models with the MMPP arrival process which includes R independent traffic flows.

than being immediately discarded, as illustrated in the traffic flow representation of the switch in Figure 3.3. The traffic carried by the optical buffers is re-offered to the optical switch output ports as a retrial flow. This retrial traffic process may overflow from the switch output ports and then be directed to the optical buffers again. Thus the switch node and FDL buffers form a traffic feedback (re-trial) loop and so must be modelled in conjunction with each other to resolve overall packet blocking probability and delay. An accurate traffic modelling technique, that can take account of correlations between the various traffic streams in the loop, is required to quantify the behaviours of the traffic overflowing from the switch to the optical buffers, the traffic carried by FDLs and the overflowing traffic from FDLs. To this end, a Markov-modulated modelling approach, which can well capture the correlations in the traffic streams while still maintaining analytical tractability, is applied [13, 14, 19].

Figure 3.3 shows that the proposed analytical framework is logically partitioned into two parts.

One captures the behaviours of the optical switch, and the other focuses on approximating the performance of the proposed contention resolution scheme. The two models are dependent on each other, requiring a recursive resolution to the queueing equations. The switch queueing system is modelled as a collection of independent and identical MMPP/M/1 loss systems with MMPP arrivals, each representing an outgoing channel (FWC) at switch outputs. This decomposition is attributed to the implementation of the waveband partitioning strategy, which divides the entire waveband at each switch output into K subsets, each allocated a dedicated FWC, thus all FWCs operate independently of each other, as stated previously in Section 3.2. The traffic streams overflowing from the switch loss systems are combined and approximated by the MMPPs, and then directed to the optical buffer queueing system as input traffic processes. Note that in the optical buffers, waveband partitioning is also employed, which divides the waveband in the FDL into S subsets, and each subset is assigned a dedicated TWC, thus the S TWCs in an FDL work independently. For that reason, the optical buffer queueing system can be decomposed into S independent and identical Markov-modulated queueing models, as shown in Figure 3.3. The departing processes from the S identical buffer queueing models are merged and approximated by the MMPPs, and subsequently return to the switch queueing system. In what follows, the switch queueing model, the buffer queueing model, and the associated traffic arrival process, traffic departure process and traffic overflow process are explained in detail. Before proceeding to the developed analytical model, the fundamentals of the Markov-modulated Poisson Process (MMPP) are first reviewed, which is based on [13, 14] and reviewed here for the readers convenience.

3.3.1 Preliminaries

3.3.1.1 Markov-modulated Poisson Process (MMPP)

The Markov-modulated Poisson process (MMPP) is a point process whose arrival rate is governed by an underlying Markov chain [19]. Considering an MMPP with k states, during a time period in which the process is in state i, arrivals occur according to a Poisson process with rate λ_i, $i \in \{1, \ldots, k\}$. A k-state MMPP is characterised by the two parameters Q and Λ. Q is the *state transition rate matrix* of the underlying Markov chain, which determines the time intervals of the process in each state, and Λ is the diagonal rate matrix, defined as $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_k)$, specifying the arrival rates for each state. In this work, the focus is mainly on the two-state MMPP (Q, Λ)

which takes the specific form

$$Q = \begin{bmatrix} -\theta_1 & \theta_1 \\ \theta_2 & -\theta_2 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{3.1}$$

The process is thus defined by just four scalar parameters $\theta_1, \theta_2, \lambda_1$ and λ_2. If the two-state MMPP is in state 1, the traffic arrives according to a Poisson process with intensity λ_1, otherwise, the traffic arrival rate is λ_2. The intervals of time when the MMPP is in state 1 or 2 are both exponentially distributed with average intensity respectively defined by θ_1 and θ_2.

The superposition of multiple mutually independent Markov-Modulated Poisson Processes (MMPPs) is also an MMPP, but with a larger state space than its constituents. For m independent two-state MMPPs (Q_i, Λ_i), $i = 1, \ldots, m$, the superimposed MMPP process $(Q^{(m)}, \Lambda^{(m)})$ is expressed as,

$$Q^{(m)} = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_m,$$

$$\Lambda^{(m)} = \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_m.$$

where $\oplus$ represents the Kronecker sum, explained as follows. Given A and B, square matrices of respective orders p and q, the Kronecker sum of A and B is given by [13],

$$A \oplus B = (A \otimes I_q) + (I_p \otimes B) \tag{3.2}$$

where I_q is the identity matrix of order q, and $\otimes$ denotes the Kronecker product. The Kronecker product, $A \otimes I_q$ (or $I_p \otimes B$), is a square matrix of order pq with block elements $\{A_{ij} I_q\}$ (respectively $\{I_{ij} B\}$) [20]. Thus, the resulting Kronecker sum of A and B is a square matrix of size pq. Thus the superposed MMPP process $(Q^{(m)}, \Lambda^{(m)})$ of m two-state MMPPs has a state space of size 2^m. It should be noted that, in the case that the m two-state MMPPs (Q, Λ) are identical, the aggregated multi-state MMPP $(Q^{(m)}, \Lambda^{(m)})$ can be simplified to a $(m+1)$-state process, which is expressed as,

$$Q^{(m)} = \begin{bmatrix} -m\theta_1 & m\theta_1 & & & \\ \theta_2 & -((m-1)\theta_1 + \theta_2) & (m-1)\theta_1 & & \\ \ddots & \ddots & \ddots & \ddots & \\ & (m-1)\theta_2 & -(\theta_1 + (m-1)\theta_2) & \theta_1 & \\ & & m\theta_2 & -m\theta_2 & \end{bmatrix} \tag{3.3}$$

$$\Lambda^{(m)} = \begin{bmatrix} m\lambda_1 & & & \\ & ((m-1)\lambda_1 + \lambda_2) & & \\ & & \ddots & \\ & & & m\lambda_2 \end{bmatrix}. \tag{3.4}$$

In a MMPP queueing analysis problem, multi-state MMPPs may lead to the state space explosion issue, that is, the high dimensionality of the state space causes an exponential computational complexity rise, consequently making the queueing system not amenable to analysis [21]. To avoid state space explosion, the multi-state MMPPs arising in the queueing analysis will be approximated as simpler two-state MMPP models by using the moment matching method, following [13, 14, 19] and summarised below.

The Two-state MMPP Approximation

The stationary probability vector of the multi-state MMPP $(Q^{(m)}, \Lambda^{(m)})$ is denoted as $\pi^{(m)}$, which represents the probability distributions of the states in the MMPP. It can be obtained by solving the system equilibrium equations,

$$\pi^{(m)} Q^{(m)} = \mathbf{0}, \quad \pi^{(m)} \mathbf{e} = 1 \tag{3.5}$$

where $\mathbf{e}$ is a column vector of ones, i.e. $\mathbf{e} = [1, \ldots, 1]$. Let u_j represent the j-th non-central moment of the MMPP, and v be variance, so that $v = u_2 - u_1^2$. In the matching procedure, four statistics of the multi-state MMPP $(Q^{(m)}, \Lambda^{(m)})$ are chosen to be matched with the characteristics of a two-state MMPP (Q, Λ): (i) the mean arrival rate μ_1, (ii) the variance v, (iii) the third non-central moment μ_3, and (iv) the time constant τ_c. The first three moments can be calculated by

$$\mu_1 = \pi^{(m)} \Lambda^{(m)} \mathbf{e}, \quad \mu_2 = \pi^{(m)} (\Lambda^{(m)})^2 \mathbf{e}, \quad \mu_3 = \pi^{(m)} (\Lambda^{(m)})^3 \mathbf{e} \tag{3.6}$$

and the time constant τ_c is computed as the integral of the arrival rate covariance function $r(t)$ [13], given by

$$\begin{aligned} \tau_c &= \tfrac{1}{v} \textstyle\int_0^\infty r(t) dt \\ &= \tfrac{1}{v} [\pi^{(m)} \Lambda^{(m)} (\mathbf{e}\pi^{(m)} - Q^{(m)})^{-1} \Lambda^{(m)} \mathbf{e} - u_1^2]. \end{aligned} \tag{3.7}$$

Based on [19], the parameters of the two-state MMPP (Q, Λ) are calculated as

$$\theta_1 = \frac{1}{\tau_c(1+a)}, \quad \theta_2 = \frac{a}{\tau_c(1+a)}, \tag{3.8}$$

$$\lambda_1 = u_1 + \sqrt{v/a}, \quad \lambda_2 = u_1 - \sqrt{va} \tag{3.9}$$

where

$$b = \frac{u_3 - 3u_1 v - u_1^3}{v^{3/2}}, a = 1 + \frac{b}{2}(b - \sqrt{4+b^2}) \tag{3.10}$$

In this way, the multi-state arrival process has been reduced to a two-state MMPP (Q, Λ), which is a simpler approximate traffic model, allowing fast queueing analysis.

3.3.1.2 Interrupted Poisson Process (IPP)

The Interrupted Poisson Process (IPP) is a special case of the two-state MMPP, where one of the arrival rates, say λ_2, equals zero. An IPP is thus defined by

$$Q_{IPP} = \begin{bmatrix} -\theta_1 & \theta_1 \\ \theta_2 & -\theta_2 \end{bmatrix}, \quad \Lambda_{IPP} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.11}$$

The process alternates between two states, namely, an ON state and an OFF state. The on-periods are exponentially distributed with mean $1/\theta_1$, and the off-periods are exponentially distributed with mean $1/\theta_2$. The on-periods and off-periods are considered to be mutually independent. During the on-period, the packets arrive according to a Poisson process with rate λ_1, but upon entering the off-period, the flow is interrupted and no arrivals occur. Thus, IPP represents a bursty traffic process. The IPP is uniquely characterised by parameters θ_1, θ_2 and λ_1. The probability that the process is in the ON state, denoted by r, is expressed as

$$r = \frac{1/\theta_1}{1/\theta_1 + 1/\theta_2} = \frac{\theta_2}{\theta_1 + \theta_2}. \tag{3.12}$$

r is referred to as the *coefficient of burstiness*. Generally, $r < 1$, and when $r = 1$, the IPP becomes exactly a Poisson process. Given λ_1 and r, the average packet arrival rate in an IPP is defined by

$$\lambda_{avg} = r\lambda_1. \tag{3.13}$$

Then the mean inter-arrival time is $\bar{m} = \frac{1}{\lambda_{avg}}$. Let $\bar{v}$ represent the variance of the inter-arrival times. Then the *coefficient of variation*, denoted by C, is computed as,

$$C = \frac{\bar{v}}{\bar{m}^2} \tag{3.14}$$

The burstiness of an IPP traffic is determined by parameters r and C.

3.3.2 The Switch Queueing System

Figure 3.3 shows the developed mathematical modelling framework of the proposed optical switch with recirculating optical buffers. The model consists of the switch queueing system and the optical buffer queueing system. Table 3.1 presents the parameters related to the model. In this section, the main objective is to introduce the switch queueing system, which characterises the behaviours of the all-optical switching plane. Recall that the proposed $N \times N$ optical switch node can be decomposed into a network of isolated queues, more precisely, NK identical MMPP/M/1 queueing models, each approximating the behaviour of one FWC. The primary input traffic to the queueing model is modelled by a two-state MMPP (Q_0, Λ_0) formed by N/K Independent and Identically Distributed (IID) MMPPs. Further, due to FDL re-circulations, the contention traffic which traverses through the optical buffers and then returns to the switch output, also contributes to the arrival process. Thus, the arrival process to the switch queue consists of a primary MMPP input traffic and multiple two-state MMPP retrial flows, as shown in Figure 3.3. The i-th retrial MMPP flow represents the traffic stream that has travelled through optical buffers i times. Note that the number of allowable FDL recirculations is limited to R, thereby there exists R retrial flows, denoted by (Q_i, Λ_i), $1 \leq i \leq R$. The primary two-state MMPP arrival and the two-state MMPP retrial traffic flows are expressed as,

$$Q_i = \begin{bmatrix} -\theta_{i1} & \theta_{i1} \\ \theta_{i2} & -\theta_{i2} \end{bmatrix}, \quad \Lambda_i = \begin{bmatrix} \lambda_{i1} & 0 \\ 0 & \lambda_{i2} \end{bmatrix}. \tag{3.15}$$

It is further assumed that the input traffic flows are independent. The arrival process (Q, Λ) is calculated as,

$$Q = Q_0 \oplus Q_1 \oplus Q_2 \oplus \cdots \oplus Q_R, \tag{3.16}$$

$$\Lambda = \Lambda_0 \oplus \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_R. \tag{3.17}$$

Table 3.1: Table of Notation

N	Number of incoming/outgoing fibres connected to the switch architecture
K	Number of transmission links (FWCs) per output
L	Number of Fibre Delay Lines (FDLs)
S	Number of channels (TWCs) per FDL
D	Unit delay in FDL buffers. The delay times of FDLs are integer multiples of D, i.e., D, $2D$,..., LD
D_{G}	Generalized unit delay, defined by D/t
R	Number of allowable FDL re-circulations
λ_{avg}	Average arrival rate of the input process
t	Mean packet transmission time
μ	Packet service rate, defined by $1/t$
ρ	Input link load, which is given by λ_{avg}/μ
r	Coefficient of burstiness of the IPP
C	Coefficient of variation of the IPP, which is expressed as the ratio between the variance and the squared mean of the inter-arrival times.
P	Packet contention probability in the optical network
T	Mean end-to-end latency in the optical network
E	Power consumption of the switch

The state space of the aggregated MMPP has a size of $2^{(R+1)}$. The contribution of the j-th ($0 \leq j \leq R$) input flow, denoted by $\Lambda(j)$, is represented by,

$$\begin{aligned} \Lambda(j) = & \quad 0 \oplus 0 \oplus \cdots 0 \oplus \Lambda_j \oplus 0 \cdots \oplus 0 \\ = & \quad I_2 \otimes I_2 \otimes \cdots I_2 \otimes \Lambda_j \otimes I_2 \cdots \otimes I_2 \end{aligned} \tag{3.18}$$

where I_2 is an 2×2 identity matrix. Note that $\Lambda = \sum_{j=0}^{j=R} \Lambda(j)$.

As mentioned earlier, the proposed switch node is decomposed into NK identical MMPP/M/1 models working in isolation. The MMPP/M/1 is a single-server queueing system with exponentially distributed service times and an MMPP arrival process composed of a primary MMPP and R independent two-state MMPP retrial flows. The behaviour of the MMPP/M/1 model can be analysed as a Markov process with state space $\{(p, q), 0 \leq p \leq 1, 1 \leq q \leq 2^{(R+1)}\}$. The infinitesimal generator $\overline{Q}$ of this process is defined as,

$$\overline{Q} = \begin{array}{c} \\ \mathbf{0} \\ \mathbf{1} \end{array} \begin{array}{c} \begin{array}{cc} \quad & \quad \end{array} \\ \begin{bmatrix} Q - \Lambda & \Lambda \\ I & Q - I \end{bmatrix} \end{array}. \tag{3.19}$$

When the Markov process $\overline{Q}$ is in the state $\mathbf{0}$, which means the server is free, the traffic is served and then departs from the system. In this case, there is no overflow traffic. Conversely, if the process $\overline{Q}$

is in the state $\mathbf{1}$, the arriving traffic overflows from the queue. In that regard, the rate matrix of the traffic overflowing from the j-th input traffic flow, denoted by $\overline{\Lambda}(j)$, is expressed as,

$$\overline{\Lambda}(j) = \begin{matrix} \mathbf{0} \\ \mathbf{1} \end{matrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda(j) \end{bmatrix}. \tag{3.20}$$

Thus, the overflow process seen by the j-th input traffic flow is characterised by an MMPP $(\overline{Q}, \overline{\Lambda}(j))$. The overflow traffic streams are approximated on a per-input-flow basis. It is important to point out that in the switch, the PreRes scheduling scheme is employed, which reserves the switch output resources in advance for the incoming traffic, thus the traffic recirculating in the optical buffers is guaranteed to be served. To capture this feature, the overflow traffic seen by the R-th input flow is added to the MMPP $(\overline{Q}, \overline{\Lambda}(0))$, and consider that the R-th input flow results in no overflow traffic. Consequently, there exists R overflow traffic streams $(\overline{Q}, \overline{\Lambda}(j))$, $j = 0, \ldots, R-1$, caused by R independent input MMPPs (Q_j, Λ_j), $j = 0, \ldots, R-1$. These traffic streams are offered to the buffering system as arrival process. In order to simplify the queueing analysis and reduce the computation complexity, the multi-state MMPPs $(\overline{Q}, \overline{\Lambda}(j))$ are approximated as two-state MMPPs by matching their statistical properties. To do that, it is necessary to calculate the stationary vector $\overline{\pi}$, which satisfies $\overline{\pi}\overline{Q} = \mathbf{0}, \overline{\pi}\mathbf{e} = 1$. Based on the approximation procedure described in Section 3.3.1.1, R two-state MMPP overflow streams $(\widetilde{Q}_j, \widetilde{\Lambda}_j)$, $j = 0, \ldots, R-1$, are obtained, which are defined as,

$$\widetilde{Q}_j = \begin{bmatrix} -\widetilde{\theta}_{j1} & \widetilde{\theta}_{j1} \\ \widetilde{\theta}_{j2} & -\widetilde{\theta}_{j2} \end{bmatrix}, \quad \widetilde{\Lambda}_j = \begin{bmatrix} \widetilde{\lambda}_{j1} & 0 \\ 0 & \widetilde{\lambda}_{j2} \end{bmatrix}.$$

It should be noted that the two-state MMPP $(\widetilde{Q}_j, \widetilde{\Lambda}_j)$ represents the overflow traffic from one MMPP/M/1 queueing model, the overall overflow traffic from the switch node is the superposition of NK independent and identical MMPP $(\widetilde{Q}_j, \widetilde{\Lambda}_j)$ flows, which forms the input traffic to the optical buffers of increasing length.

Recall that the optical buffering network can be decomposed into S identical Markov-modulated finite-server queueing models with the input process being the superposition of NK/S IID MMPP $(\widetilde{Q}_j, \widetilde{\Lambda}_j)$ flows. Before introducing the buffer queueing model, the superposition of NK/S IID MMPPs $(\widetilde{Q}_j, \widetilde{\Lambda}_j)$ is first parameterised. Consider that $X = NK/S$. As mentioned in Section 3.3.1.1, the superposition of X independent and identical two-state MMPPs $(\widetilde{Q}_j, \widetilde{\Lambda}_j)$ can be

approximated by a $(X+1)$-state MMPP $(\overline{\overline{Q}}_j, \overline{\overline{\Lambda}}_j)$, which is given by,

$$\overline{\overline{Q}}_j = \begin{bmatrix} -X\widetilde{\theta}_{j1} & X\widetilde{\theta}_{j1} & & & \\ \widetilde{\theta}_{j2} & -((X-1)\widetilde{\theta}_{j1}+\widetilde{\theta}_{j2}) & (X-1)\widetilde{\theta}_{j1} & & \\ \ddots & & \ddots & \ddots & \\ & (X-1)\widetilde{\theta}_{j2} & -(\widetilde{\theta}_{j1}+(X-1)\widetilde{\theta}_{j2}) & \widetilde{\theta}_{j1} & \\ & & X\widetilde{\theta}_{j2} & -X\widetilde{\theta}_{j2} & \end{bmatrix} \tag{3.21}$$

$$\overline{\overline{\Lambda}}_j = \begin{bmatrix} X\widetilde{\lambda}_{j1} & & & \\ & ((X-1)\widetilde{\lambda}_{j1}+\widetilde{\lambda}_{j2}) & & \\ & & \ddots & \\ & & & X\widetilde{\lambda}_{j2} \end{bmatrix}. \tag{3.22}$$

where $j = 0, \ldots, R-1$. To avoid the state-space explosion issue, the multi-state MMPPs $(\overline{\overline{Q}}_j, \overline{\overline{\Lambda}}_j)$ are approximated as two-state MMPPs $(\widehat{Q}_j, \widehat{\Lambda}_j)$, $j = 0, \ldots, R-1$, based on the approximation procedure in Section 3.3.1.1. Having obtained the representation of the arrival process to the buffer queue, in the following section, the optical buffer queueing model is detailed.

3.3.3 The Optical Buffer Queueing System

This section introduces the developed optical buffer queueing model which approximates the behaviour of the contention resolution scheme. Consider that the optical buffers consist of L Fibre Delay Lines (FDLs), and each fibre contains S outgoing channels. As indicated in the previous section, the optical buffering network is decomposed into S parallel Markov-modulated finite-server queues, and the input process to the queue is represented by R two-state MMPPs $(\widehat{Q}_j, \widehat{\Lambda}_j)$, $j = 0, \ldots, R-1$, see Figure 3.3. Note that the resource scheduling mechanism in the proposed switch relies on the PreRes scheme which reserves the switch and FDL resources before the packets enter into the switch. This strategy determines that the two-state MMPP traffic $(\widehat{Q}_j, \widehat{\Lambda}_j)$, for $1 \leq j \leq R-1$, which is about to traverse through the optical buffer for the $(j+1)$-th time, goes though the buffering queue without any traffic loss, and turns into the $(j+1)$-th input retrial flow directly. That means, in the buffer queueing analysis, the R MMPP traffic streams $(\widehat{Q}_j, \widehat{\Lambda}_j)$, $j = 0, \ldots, R-1$, will compete and occupy the resources, but the MMPPs $(\widehat{Q}_j, \widehat{\Lambda}_j)$, $j = 1, \ldots, R-1$, have higher priority, and only the MMPP $(\widehat{Q}_0, \widehat{\Lambda}_0)$ leads to traffic loss in the system.

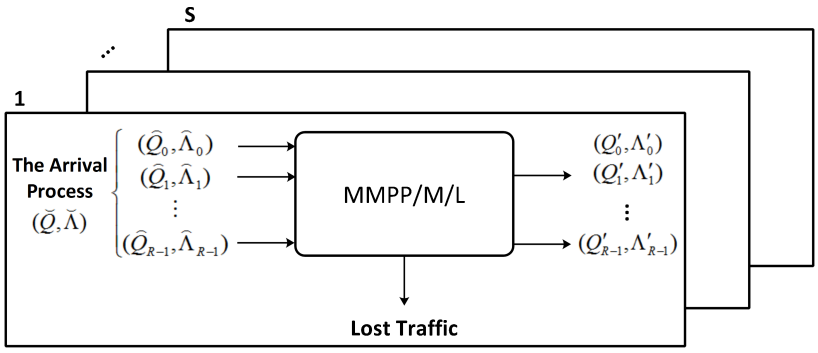
The optical buffers utilised are assumed to be of degenerate type, that is, the discrete-time delays introduced by FDLs are integer multiples of the delay granularity D, i.e. $D, 2D, \ldots, LD$. It is obvious that for small values of D, the time resolution of the FDL buffers is low, thus the behaviours of the optical buffers can be well modelled by a queueing system with L servers operating in parallel [12, 17]. When D is large, the time resolution of the FDL buffers is high, and in this case each fibre is more likely to function as an individual queue and carry the incoming traffic independently, so the buffers can be modelled as a cascaded chain of L distinct queues [22]. Depending on the FDL delay granularity, two different Markov-modulated finite-server queueing models are developed: (a) an MMPP/M/L queue, as shown in Figure 3.4(a), and (b) a chain of L MMPP/M/1 queues with cascading overflows, as shown in Figure 3.4(b). The MMPP/M/L queue is expected to approximate the performance of the FDLs with small D, whereas the cascaded model of L MMPP/M/1 queues is expected to quantify the optical buffers with large D. As shown in Figure 3.4(a), the MMPP/M/L model is a L-server queue receiving MMPP-based arrival process which includes R two-state MMPPs $(\widehat{Q}_j, \widehat{\Lambda}_j)$, $j = 0, \ldots, R-1$. The arrival process $(\breve{Q}, \breve{\Lambda})$ is computed as,

$$\breve{Q} = \widehat{Q}_0 \oplus \widehat{Q}_1 \oplus \cdots \oplus \widehat{Q}_{R-1}, \quad \breve{\Lambda} = \widehat{\Lambda}_0 \oplus \widehat{\Lambda}_1 \oplus \cdots \oplus \widehat{\Lambda}_{R-1}. \tag{3.23}$$

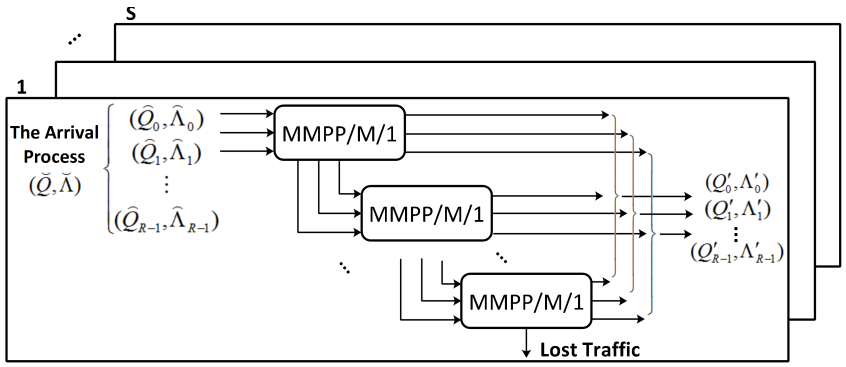
The MMPP/M/L queueing model can be approximated by a $L+1$-state Markov process where the infinitesimal generator $\check{\breve{Q}}$ is expressed as,

$$\check{\breve{Q}} = \begin{array}{c} \\ \mathbf{0} \\ \mathbf{1} \\ \vdots \\ \mathbf{L} \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{0} & \mathbf{1} & & & & \end{array} \\ \begin{bmatrix} \breve{Q}-\breve{\Lambda} & \breve{\Lambda} & & & \\ I & \breve{Q}-\breve{\Lambda}-I & \breve{\Lambda} & & \\ & \ddots & \ddots & \ddots & \\ & & LI & \breve{Q}-LI \end{bmatrix} \end{array}. \tag{3.24}$$

When the Markov process is in the state $\mathbf{i}$, $\mathbf{i} \in \{\mathbf{0}, \ldots, \mathbf{L}-\mathbf{1}\}$, the arrival traffic is processed and departs from the system, but if the process is in the state $\mathbf{L}$, which means all of the L servers are busy, the arrival traffic overflows from the queue. The MMPP departing processes, which are quantified by two-state MMPPs $(\acute{Q_j}, \acute{\Lambda_j})$, $j = 0, \ldots, R-1$, will loop back to the switch queueing system as part of the arrival process. Conversely, the traffic overflowing from the queue will be abandoned and considered lost from the network. Differently, in the cascaded model of L MMPP/M/1 queues, the arrival process to the first MMPP/M/1 is MMPP $(\breve{Q}, \breve{\Lambda})$. The overflowing process, modelled as MMPP, forms the arrival process to the next MMPP/M/1, and so on down the chain, as shown in Figure 3.4(b). In the L-th MMPP/M/1 queue, the overflow traffic from the queue is discarded.



(a) An MMPP/M/L queue



(b) A cascaded model of L MMPP/M/1 queues

Figure 3.4: The buffer queueing model consists of S parallel Markov-modulated finite-server queues. Two different representations are developed to characterise the Markov-modulated finite-server queue: (a) a single MMPP/M/L model and (b) a cascaded model of L MMPP/M/1 queues.

Obviously, the departing process from the cascaded model is the superposition of L individual MMPP departing streams. Note that these individual streams are not identical. Further, the multi-state departing processes are approximated by two-state MMPPs $(\acute{Q_j}, \acute{\Lambda_j})$, $j = 0, \ldots, R-1$, so as to simplify the analysis. It is important to point out that in both buffer models, for $j = 1, \ldots, R-1$, the MMPP $(\acute{Q_j}, \acute{\Lambda_j})$ and the MMPP $(\widehat{Q}_j, \widehat{\Lambda}_j)$ are the same process, due to the PreRes scheme. Therefore, the overall departing process from the buffer queueing model is composed of S identical two-state MMPPs $(\acute{Q_j}, \acute{\Lambda_j})$, which is then reduced to a two-state MMPP $(\ddot{Q}_j, \ddot{\Lambda}_j)$, $j = 0, \ldots, R-1$. The MMPP $(\ddot{Q}_j, \ddot{\Lambda}_j)$ is uniformly distributed to NK switch models as arrival process (Q_{j+1}, Λ_{j+1}), $j = 0, \ldots, R-1$, which is MMPP (Q_i, Λ_i), $i = 1, \ldots, R$.

An important feature of the cascaded model is that it keeps track of the individual traffic flows, from and to, each FDL, thus allowing for the estimation of the mean occupancy of each FDL. Given the

occupancy measures $\mathbf{M} = \{M(k)|1 \leq k \leq L\}$, the overall end-to-end delay T is defined by

$$T = D \sum_{k=1}^{L} kM(k). \tag{3.25}$$

The equation reveals that a linear dependency exists between the mean communication delay and the FDL base delay D.

To validate the performance of the analytical model, a detailed evaluation of the proposed analytical framework has been performed in the next section, and extensive comparisons against simulation results have been carried out.

3.4 Performance Analysis

The subject of this section is to analyse the performance characteristics of the proposed optical packet-switched network by means of both analytical modelling and network simulations. The mathematical modelling is implemented in MATLAB [23], and the network simulation is performed by using the discrete-event simulator OMNeT++ [24]. Extensive comparisons between numerical approximations and simulation results have been carried out so as to validate the correctness and accuracy of the proposed modelling method.

In the experiments, a uniform routing policy is employed, that is, the output destinations of the input traffic are selected with equal probability. The contention traffic is allowed to traverse through the optical buffers multiple times until the destined output becomes free. To control latency performance and signal degradation issues, the maximum number of allowable FDL recirculations is limited to R. The detailed notation used in this section is presented in Table 3.1. There are many parameters that can affect the network performance: the input traffic characteristics, the number of transmission channels (FWCs) per switch output port K, the number of FDLs L, the number of channels (TWCs) per FDL S, the FDL delay granularity D or D_G, and the number of allowable FDL recirculations R. In the exploration of the accuracy of the proposed model, the impact of these network parameters in different configurations of the switch architecture will be investigated in detail. Performance metrics of interest include the contention probability, the FDL occupancy and the mean end-to-end delay. The analysis starts with the performance evaluation of the analytical model with Poisson primary arrival process.

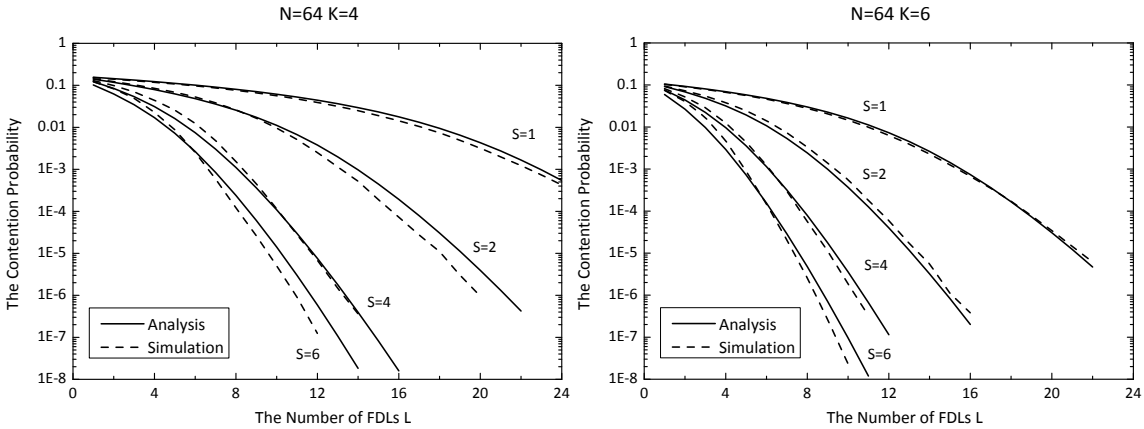
3.4.1 Poisson Arrival Traffic

In this section, the emphasis is on the performance analysis of the proposed Markov-modulated analytical framework with Poisson arrival process. The performance quantities of the optical switch, measured in terms of the contention probability and the average end-to-end delay, are evaluated and discussed. To assess the quality of the approximations, extensive comparisons between analytical estimations and simulation results have been performed. Noting that for all experiments, two different switch size configurations are considered: $N = 64$ and $N = 256$. And the maximum number of allowable FDL recirculations is fixed to be $R = 3$.

3.4.1.1 Approximations of the Contention Probability

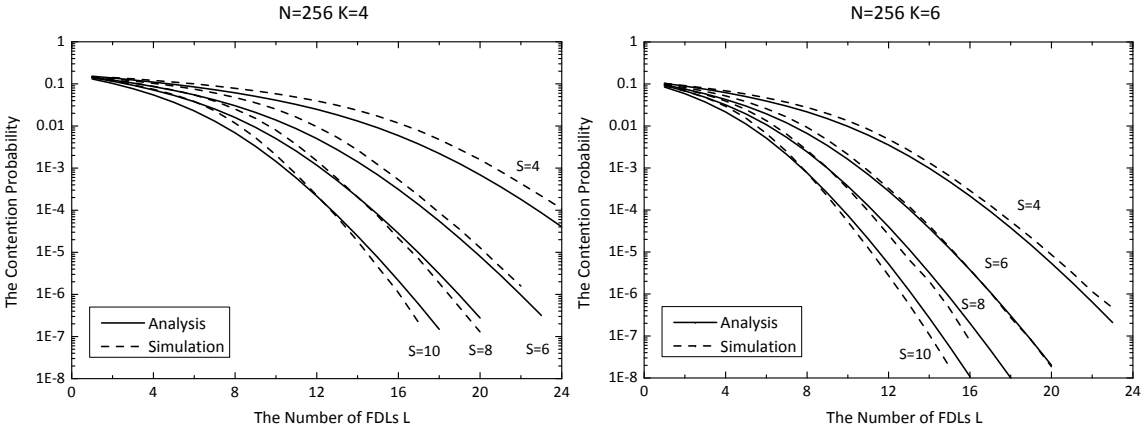
As mentioned in Section 3.3, due to the unique feature of waveband partitioning at outputs, the switch queueing model consists of a network of independent and identical MMPP/M/1 loss nodes, each representing an outgoing channel. Further, in order to study the behaviours of the optical buffers in depth, two different models are developed. One refers to a single MMPP/M/L queueing system, which represents a L-server loss node with MMPP arrivals. This model is suggested for the FDL implementations with small unit delay D_G, due to the fact that the time resolution of the optical buffers is low, and the L buffers can be viewed as a system having L servers operating in parallel. To investigate the accuracy of the model, analytical estimations of the contention probability are compared with simulation results under high load condition, as depicted in Figures 3.5 - 3.6. In the simulations, the FDL delay granularity D_G is set to be 0.1. The blocking probability is plotted as a function of the number of optical buffers L in different network configurations. It is shown that the numerical results obtained from the analytical model are in good agreement with simulations results for a wide range of working conditions, e.g. different switch sizes N, different channels per switch output K,various FDL sizes L, and different values of parameter S. As expected, increasing K yields a significant reduction in packet loss rate, due to the fact that increasing the number of FWCs, K, alleviates the contention conditions at switch outputs, thus resulting in less overflow traffic. It is also observed that increasing the number of FDLs, L, lowers the average drop rate significantly. Additionally, while moving to higher values of the S parameter, the blocking probability becomes smaller. Thus, higher buffering capacity can effectively improve contention performance.

The other model refers to a chain of L MMPP/M/1 queues with cascading overflows. That is to say, the i-th MMPP/M/1 queueing system generates overflows which are fed to $(i + 1)$-th MMPP/M/1



(a) $N = 64$ and $K = 4$ (b) $N = 64$ and $K = 6$

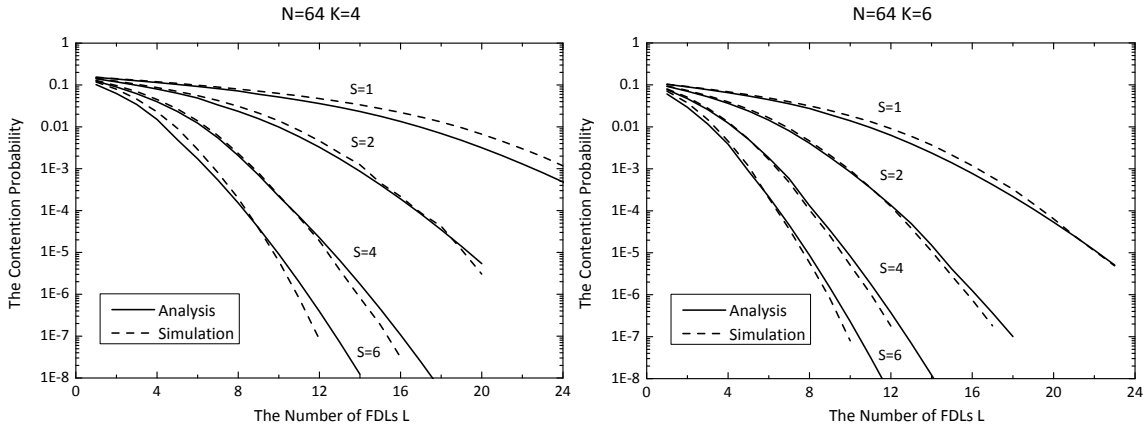
Figure 3.5: $N = 64$, $\rho = 0.8$, $R = 3$ and $D_G = 0.1$. The observations are obtained for two different scenarios: $K = 4$ and $K = 6$.



(a) $N = 256$ and $K = 4$ (b) $N = 256$ and $K = 6$

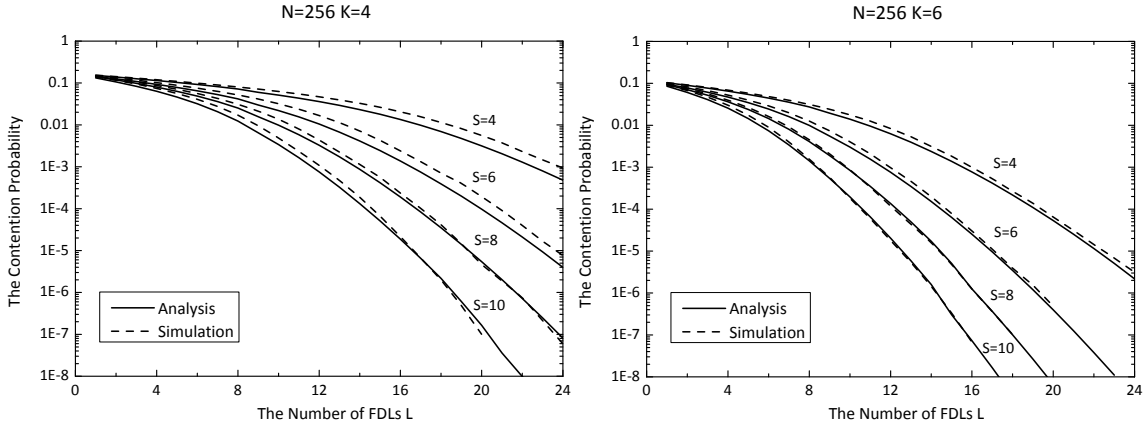
Figure 3.6: $N = 256$, $\rho = 0.8$, $R = 3$ and $D_G = 0.1$. The observations are obtained for two different scenarios: $K = 4$ and $K = 6$.

queue. This model targets network configurations with large FDL delay granularity D_G. In such a switch, the delay values provided by FDLs are significantly different, it is therefore anticipated that the fibres are more likely to handle the incoming traffic independently. To measure the accuracy of the queueing analysis, numerical results are plotted in Figures 3.7 - 3.8, which provide convincing support for the developed analytical model. In the simulations, the base delay is taken as $D_G = 1$. Evidently, the observations show a similar behaviour as for the previous example. When comparing the results from two sets of experiments, an important observation is that decreasing the delay granularity D_G lowers the blocking probability considerately. This beneficial feature of the proposed switch allows the performance to be improved with reduced length of the optical fibres.



(a) $N = 64$ and $K = 4$ (b) $N = 64$ and $K = 6$

Figure 3.7: $N = 64$, $\rho = 0.8$, $R = 3$ and $D_G = 1$. The observations are obtained for two different scenarios: $K = 4$ and $K = 6$.



(a) $N = 256$ and $K = 4$ (b) $N = 256$ and $K = 6$

Figure 3.8: $N = 256$, $\rho = 0.8$, $R = 3$ and $D_G = 1$. The observations are obtained for two different scenarios: $K = 4$ and $K = 6$.

The comparisons presented above verified the accuracy of the proposed queueing models. The behaviours of the systems having different FDL delay granularities are quite different, which is reflected in the performance plots. Nonetheless, exact impact of the FDL delay granularity on the performance is still unclear, which is fundamentally important for the analysis. In that regard, explicit expression of the contention performance as a function of the delay unit D_G is thus required to better understand the effects of such parameters. To resolve this problem, a linear mathematical expression of the blocking probability P is proposed based on the proposed two models, which is expressed as

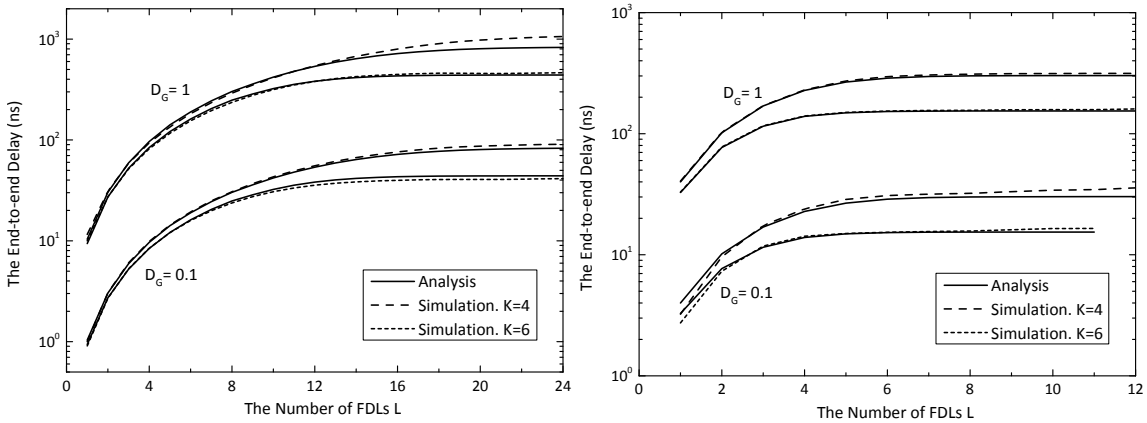
$$P(D_G) = P_2 + (D_G - 1)\frac{P_2 - P_1}{0.9} \tag{3.26}$$

where $0.1 \leq D_G \leq 1$. P_1 denotes the analytical estimation of the packet loss rate obtained from the former model targeting $D_G = 0.1$, and P_2 represents the approximation of the packet loss rate obtained from the latter model targeting $D_G = 1$. Assessment of the above equation comes from Figure 3.11, which illustrates a detailed analysis on how contention performance changes when varying the delay unit D_G. The experiments are carried out for two different scenarios: (i) $N = 64$ and (ii) $N = 256$. Clearly, in both scenarios, the contention performance improves significantly with the decrease of the delay unit D_G. Numerical comparison between analysis and simulation shows that the mathematical estimations yield a high level of accuracy, suggesting the applicability of our linear approximation.

Extensive experiments have been carried out in this section to illustrate how various network parameters affect the contention performance. The consistency between the analysis and the simulation results verifies the accuracy of the analytical representation. In the next subsection, a discussion on the latency performance of the proposed network will be presented.

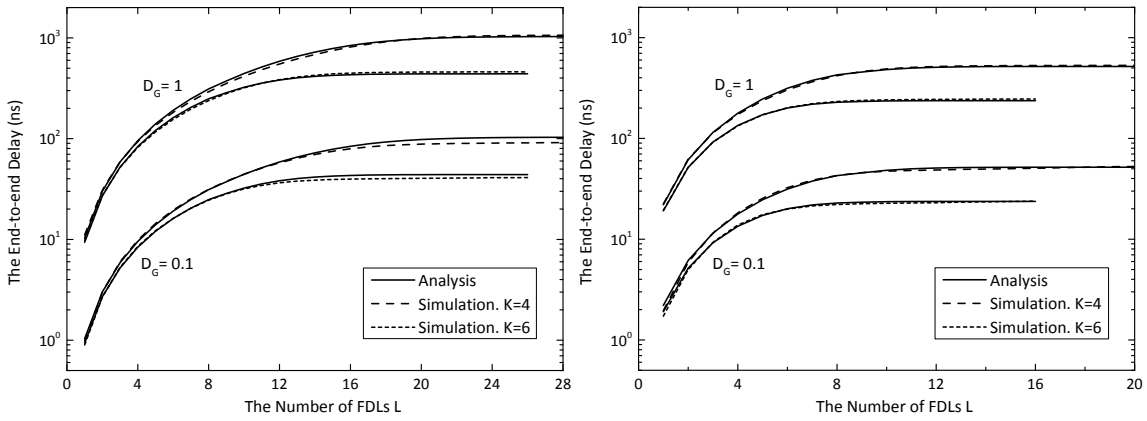
3.4.1.2 Approximations of Packet Latency

For the purpose of formulating an expression for the communication delay, it is necessary to resolve the occupancy in each of the L optical buffers, as indicated in Eq. 3.25. The cascaded model presented in Figure 3.4(b) characterises the input traffic, the carried traffic and the overflow traffic in each FDL queue. This feature allows for the estimation of the mean utilisation of the individual FDL buffers, thus making the theoretical approximation of the mean end-to-end delays possible. In this case, the mean communication delay depends heavily on the base delay D_G. This important property of the model will be illustrated later. To validate the proposed method, the numerical estimations are compared against the simulation results under a high load condition, as plotted in Figures 3.9 - 3.10. Evidently, the model captures the delay characteristics accurately, thus it is fair to say that the proposed model provides a simple and accurate tool for the approximations of the overall network delay. In the figures, a number of different network parameters are at play here, involving channels per switch output K, number of FDLs L, channels per FDL S and base delay D_G. As explained in the previous subsection, larger K leads to less traffic contentions at switch outputs, thereby reducing the amount of contention traffic which needs optical buffering. As a result, the end-to-end delay is reduced considerately. Longer base delay D_G implies that the FDLs can delay the packet for a longer time, which explains the increase in the mean delay when moving from $D_G = 0.1$ to $D_G = 1$. Further, for increasing values of the FDL size L, the overall delay



(a) $N = 64$ and $S = 1$ (b) $N = 64$ and $S = 6$

Figure 3.9: $N = 64$, $\rho = 0.8$ and $R = 3$. The overall system delays are plotted versus the number of FDLs L for different values of parameter D_G: $D_G = 0.1$ and $D_G = 1$



(a) $N = 256$ and $S = 4$ (b) $N = 256$ and $S = 10$

Figure 3.10: $N = 256$, $\rho = 0.8$ and $R = 3$. The overall system delays are plotted versus the number of FDLs L for different values of parameter D_G: $D_G = 0.1$ and $D_G = 1$

increases and then becomes stable.

As described in Section 3.3, an important implication of the proposed approximation is that the predicted overall latency is proportional to the base delay D_G. Convincing support is provided by the extensive simulation results, as illustrated in Figure 3.12. In Figure 3.12, analytical approximations and simulation results are depicted versus base delay D in two different scenarios: (i) $N = 64$ and (ii) $N = 256$. The evidence of results confirms the linear dependence of the overall latency on the FDL unit delay D_G. Furthermore, in both scenarios, the curves are very close together, verifying that the proposed modelling method exhibits a high degree of accuracy.

Observing the results in Figures 3.5 - 3.10, it is interesting to note that the smaller is the delay unit

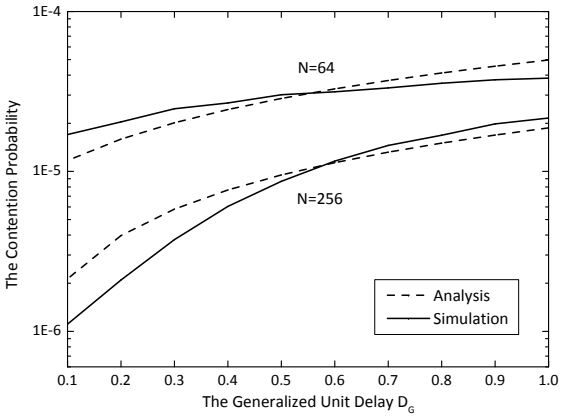


Figure 3.11: The blocking probability is plotted versus FDL unit delay D_G in two different scenarios: (i) $N = 64$, $K = 6$, $S = 2$ and $L = 13$ and (ii) $N = 256$, $K = 4$, $S = 10$ and $L = 16$.

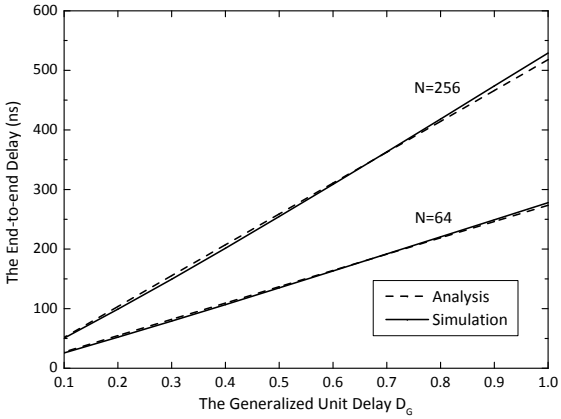
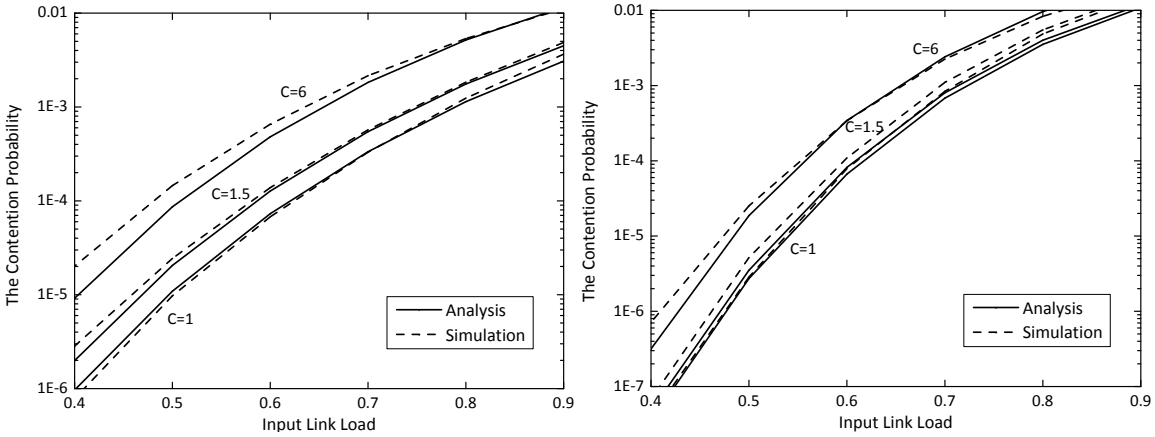


Figure 3.12: The overall delay is plotted versus FDL unit delay D_G in two different scenarios: (i) $N = 64$, $K = 6$, $S = 2$ and $L = 13$ and (ii) $N = 256$, $K = 4$, $S = 10$ and $L = 16$.

D_G, the lower is the blocking probability, and so is the end-to-end delay. Owing to this unique property, small base delay D_G is preferable when designing the optical switch. The above analysis of some important performance quantities suggests that the performance of the system is to a large extent dependent on switch parameter settings. In that regard, with the proper dimensioning of the optical buffers, a target performance can be achieved. Noting that it is concluded for Poisson arrivals, which is a relatively smooth traffic process. In the next section, the queueing performance is further investigated by employing bursty input traffic, which is characterised by an Interrupted Poisson Process (IPP).

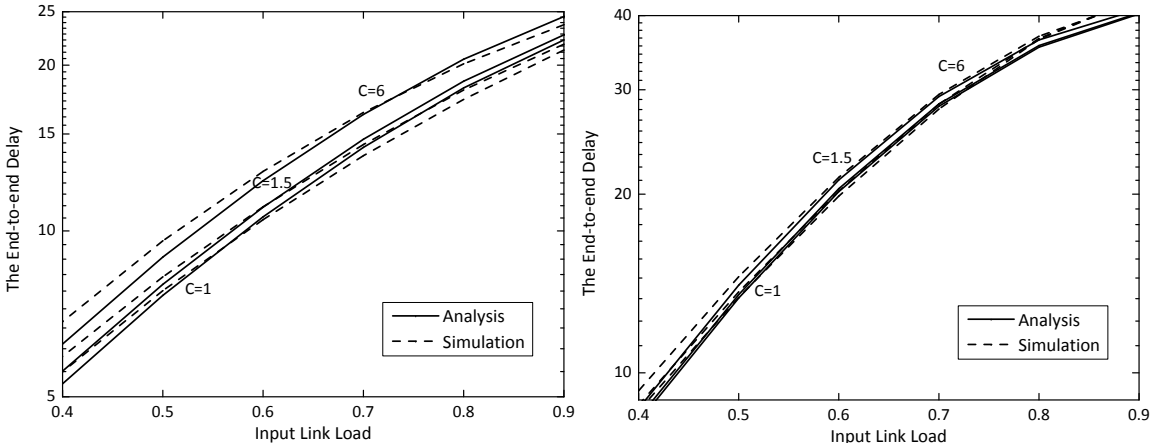
3.4.2 Interrupted Poisson Arrivals (IPP)

In this section, the performance analysis is extended to models in which the arrival processes are modelled as Interrupted Poisson Process (IPP), rather than Poisson. IPP is a traffic model that exhibits a type of burstiness. The burstiness of an IPP is determined by the coefficient of burstiness, denoted by r and the coefficient of variation, denoted by C. Ratio r defines the relative on-time of the IPP traffic, and C represents the variation of the inter-arrival times. To facilitate a more insightful analysis, by varying the burstiness of the IPP process, the influence of ratio C on the performance of the queueing network is demonstrated. Note that when $C = 1$, the input process is a 'smooth' Poisson process. Conversely, a large C indicates high variation in the traffic flow. The experiments are carried out in two different switch configurations: (i) $N = 64$, $K = 6$, $L = 6$, and $S = 4$ and (ii) $N = 256$, $K = 6$, $L = 12$, and $S = 4$. In the experiments, ratio r is set to be 0.3.



(a) $N = 64, K = 6, L = 6$, and $S = 4$. (b) $N = 256, K = 6, L = 11$, and $S = 4$.

Figure 3.13: The contention probability is plotted versus the link load for different input IPP traffics.



(a) $N = 64, K = 6, L = 6$, and $S = 4$. (b) $N = 256, K = 6, L = 12$, and $S = 4$.

Figure 3.14: The overall delay is plotted versus the link load for different input IPP traffics.

The resulting packet loss probability and overall latency obtained by analysis and simulation is depicted under varying load conditions in Figures 3.13 and 3.14, which clearly demonstrate the agreement between analysis and simulation. In Figure 3.13, the average packet drop rate is illustrated as a function of the mean input load ρ, and in Figure 3.14, the average end-to-end delay is depicted versus the mean input load ρ. As seen from the figures, for small values of C, the predicted contention performance gets fairly close to the simulated results, whereas in the case of larger values of C, that is, the IPP traffic is more bursty, the model becomes less accurate. Overall, the model provides quite good approximations. As expected, the packet drop rate and communication latency increases dramatically with the average input load ρ, because more traffic competing for the network resources will result in more frequent congestion. Additionally, it is observed that

the contention and latency performance degrades significantly with the increase of the burstiness parameter C. This is also expected as higher variation in the input traffic stream causes higher contention. Notice, however, that this degradation trend of the performance, especially the overall latency performance, becomes less obvious, when moving from $N = 64$ to $N = 256$. This can be explained by the fact that increasing switch size N leads to more Independent and Identically Distributed (IDD) IPP streams arriving to the same output, and consequently, the behaviours of the resulting traffic formed by these streams is getting closer to that of a Poisson process.

Validation with respect to simulation for the IPP arrivals shows that the predicted results closely meet the desired performance. This indicates that the model is fairly robust, and thus general enough to be applied to different traffic classes. For the next step, by exploiting the proposed model evaluation technique, the network design and planning problem is formulated and resolved.

3.5 Network Dimensioning

Having validated that the proposed analytical modelling method is accurate and both mathematically and numerically simple to apply, it is then exploited to resolving the network planning and dimensioning issue.

3.5.1 Problem Formulation

In this context, network dimensioning is formulated as an optimisation problem, with the power consumption cost being the objective function. The primary task is to obtain the optimal network parameter settings which support service provisioning with a target QoS level, while also minimising the overall power usage. In this regard, the optimisation algorithm needs to address two major issues: constraints satisfaction and optimisation. Thereby, the optimisation formula is defined as follows,

- Given : the switch size N; the input link load ρ; the mean packet length t; the FDL base delay D_G; the number of allowable FDL re-circulations R.
- Optimisation Variables : K, S and L. K is the number of transmission channels (FWCs) per switch output port, S is the number of TWCs per FDL, and L is the number of FDLs.
- Performance Constrains : the blocking probability P_{limit}, and the overall latency T_{limit}.

- Objective function : minimise the overall power consumption of the switch, denoted by $E(K, S, L)$.

The above defines the optimisation problem. The optimisation variables are K, S and L, hence the state space, denoted by Δ, is defined as $\Delta = \{\overline{\mathbf{m}} | \overline{\mathbf{m}} = (K, S, L)\}$, a three-dimensional search space. The objective function is the overall power consumption $E(K, S, L)$ which is formulated as a function of the variables K, S and L. The calculation of $E(K, S, L)$ includes the major consumers of the power, including TWCs at switch inputs, FWCs at switch ouputs and TWCs in FDLs. [25] indicates that a 10GBASE SFP+ transceiver consumes 1W power. The supply power of the TWC is set to be 1.5W [26], and the power of the FWC is set to be 0.6W [27]. The network performance measured in terms of contention probability $P(K, S, L)$ and overall delay $T(K, S, L)$, which are defined in Section 3.3, is taken as constraints. The overall aim is to find a state (K, S, L) which minimises the power consumption $E(K, S, L)$, and also satisfies the performance criterion. As this optimisation search space is most likely complex and non-convex, given the nature of the performance equations, a heuristic search procedure is implemented to address the optimisation problem.

3.5.2 Optimisation Algorithm

The detailed steps of the heuristic optimisation algorithm used here are summarised in Algorithm 1. The first step is to identify an initial point $\overline{\mathbf{m}}_0 = (K_0, S_0, L_0)$ where the network performance quantities, predicted by the analytical model, meet the conditions. Suppose that the step size of the search algorithm is taken as 1, then the neighborhood state space is described by $\mathbf{Z} = \{(K, S, L) | K_0 - 1 \leq K \leq K_0 + 1, S_0 - 1 \leq S \leq S_0 + 1, L_0 - 1 \leq L \leq L_0 + 1\}$. Next, the algorithm explores the neighborhood region $\mathbf{Z}$ to check for minimum. This way, the algorithm allocates the optimal point in the neighborhood area, which achieves the minimum power usage and also ensures the guaranteed performance. The least power consumption E_{min} is then updated. Subsequently, the algorithm proceeds to search the neighborhood region of the current optimal point obtained in the previous search. The procedure comes out of the loop until there is no further improvement for the power consumption. Obviously, the proposed algorithm proceeds in a descent direction, and eventually converges to an optimum. This provides a practical and computationally efficient way to estimate the network parameters. However, the heuristic has a main drawback that the outcome may be a local optima. Additionally, different initial starting points $\overline{\mathbf{m}}_0$ may result in different outcomes, which is confirmed by the numerical results in Tables 3.3 and 3.4. In order to obtain reasonably accurate dimensioning, multiple runs with different starting points are made, and

Algorithm 1 Min. Energy Heuristic Search Algorithm

Initialization : $\overline{\mathbf{m}}_0 \leftarrow (K_0, S_0, L_0)$
if $P(\overline{\mathbf{m}}_0) < P_{limit}$ and $T(\overline{\mathbf{m}}_0) < T_{limit}$, **then**
  $i \leftarrow 0$
  goto *loop*
else
  goto *Initialization*
end if
loop:
current state $\overline{\mathbf{m}}_i = (K_i, S_i, L_i)$
neighborhood state space $\mathbf{Z}$:
$\mathbf{Z} = \{(K, S, L) | K_i - 1 \leq K \leq K_i + 1, S_i - 1 \leq S \leq S_i + 1, L_i - 1 \leq L \leq L_i + 1\}$
check the neighborhood state space, and obtain the neighborhood feasible region $\overline{\mathbf{Z}}$:
$\overline{\mathbf{Z}} = \{\overline{\mathbf{m}} | \overline{\mathbf{m}} \in \mathbf{Z}, P(\overline{\mathbf{m}}) < P_{limit}, T(\overline{\mathbf{m}}) < T_{limit}\}$
search the feasible state space $\overline{\mathbf{Z}}$
find the state point $\overline{\mathbf{m}}_j$, which satisfies
$E(\overline{\mathbf{m}}_j) \leq E(\overline{\mathbf{m}})$, for all $\overline{\mathbf{m}} \in \overline{\mathbf{Z}}$
if $\overline{\mathbf{m}}_i \neq \overline{\mathbf{m}}_j$, **then**
  $i \leftarrow j$ ($\overline{\mathbf{m}}_i \leftarrow \overline{\mathbf{m}}_j$)
  goto *loop*.
else
  return $\overline{\mathbf{m}}_i$;
end if

the best solution is chosen as the final solution.

3.5.3 Optimal Solutions

Table 3.2: A collection of conditions for the optimisation

Given	N	64	256
	ρ	0.8	0.8
	D_G	0.1	0.1
Constraints	$P(K, S, L)$	$< 10^{-8}$	$< 10^{-8}$
	$T(K, S, L)$	$< 60ns$	$< 80ns$
Search Limits	K_{max}	6	6
	S_{max}	10	12
	L_{max}	24	32

The heuristic procedure described above is now deployed to seek the optimal networking parameters for two different network configurations, using the proposed analytical modelling solution method. Table 3.2 lists the performance constraints with respect to the contention probability and the overall delay. To facilitate the analysis, it is assumed that the maxima K_{max}, S_{max} and L_{max} are imposed on the variables K, S and L. As a result, the state space involving K, S and L becomes

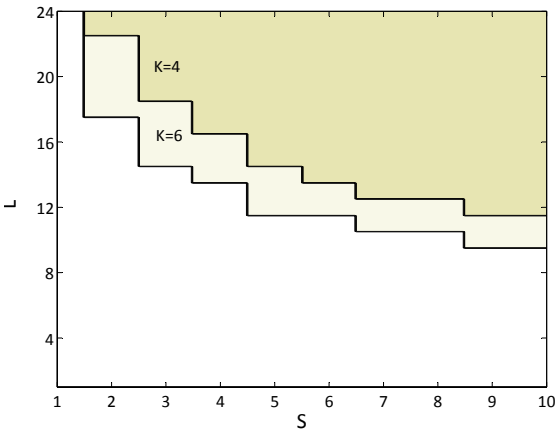


Figure 3.15: $N = 64$. The feasible state space of S and L in the scenarios: $K = 4$ and $K = 6$. Note that the blank space represents the area in which the state point breaks the constraints.

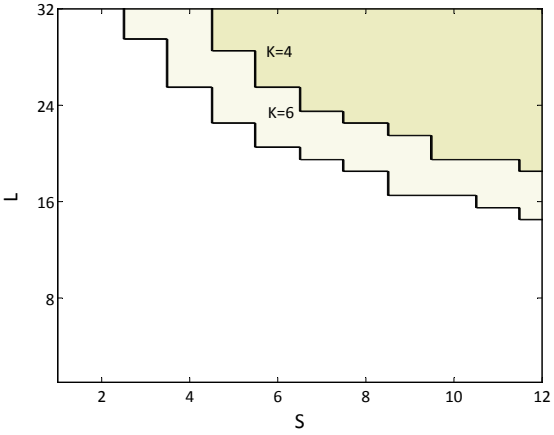


Figure 3.16: $N = 256$. The feasible state space of S and L in the scenarios: $K = 4$ and $K = 6$. Note that the blank space represents the area in which the state point breaks the constraints.

a constrained three-dimensional area. The heuristic procedure searches for the optimal solution in the feasible region of the state space which is determined by the analytical framework. As an example, in Figures 3.15 and 3.16, the feasible state space of variables S and L is plotted in the case of $K = 4$ and $K = 6$ for two different switch sizes: (i) $N = 64$ and (ii) $N = 256$. The graphs show that increasing K expands the state space for variable S and L, and the same is expected to apply to variables S and L. Note that sufficiently many replications have been conducted such that different local optima are obtained. During the experiments, it is noticed that the proposed optimisation algorithm converges quickly. Experimental results are shown in Tables 3.3 and 3.4. Observing that the resulting optimal solutions collected from the experiments are quite different, due to the fact that the heuristic procedure starting with a different initial state point may converge to a different local optimum. For $N = 64$, three sets of parameters are obtained, and the parameter setting $(4, 2, 23)$ is chosen as the final optimal solution. Alternatively, in the case of $N = 256$, four distinct optimal points are collected, and the optimal configuration $(4, 5, 29)$, which achieves the

Table 3.3: Optimal solutions for $N = 64$. $E_{min}(K, S, L)$ represents the minimum power consumption (W). $\overline{E}_{min}(K, S, L)$ represents the minimum power consumption per switch port, defined by $\overline{E}(K, S, L) = \frac{E(K,S,L)}{N}$

	N=64				
Optimal Solutions	K	S	L	$E_{min}(K, S, L)$	$\overline{E}_{min}(K, S, L)$
Solution 1	4	2	23	222.6	3.478
Solution 2	4	3	19	239.1	3.736
Solution 3	3	5	17	242.7	3.792
Global Optimum	4	2	23	222.6	3.478

Table 3.4: Optimal solutions for $N = 256$. $E_{min}(K, S, L)$ represents the minimum power consumption (W). $\overline{E}_{min}(K, S, L)$ represents the minimum power consumption per switch port, defined by $\overline{E}(K, S, L) = \frac{E(K,S,L)}{N}$

	N=256				
Optimal Solutions	K	S	L	$E_{min}(K, S, L)$	$\overline{E}_{min}(K, S, L)$
Solution 1	4	5	29	831.9	3.250
Solution 2	4	6	26	848.4	3.314
Solution 3	4	7	24	866.4	3.384
Solution 4	4	10	20	914.4	3.572
Global Optimum	4	5	29	831.9	3.250

least power consumption, is selected.

Finally, to measure the performance of the proposed optimisation algorithm, a comparison has been performed between the optimal solutions obtained in Tables 3.3 - 3.4 and the worst-case solutions computed using the same heuristic algorithm except that the objective is to maximise the power usage of the network. The worst-case solutions also fulfill the target performance requirements. In Table 3.5, the numerical comparison shows that the obtained optimal solution not only enhances the energy efficiency, but also lowers the deployment cost significantly, as the number of required optical components in the network is greatly reduced. Specifically, in the scenario of $N = 64$, the power usage in the network is reduced by 42% in comparison with the power requirement of the worse-case solution. For $N = 256$, the optimal solution results in a reduction of 30% in power consumption. This indicates that the proposed optimisation procedure is successful in finding a good optimum, given the simplicity of the formulation concerning computational time and resources.

Table 3.5: Power consumption comparison. The worst-case solutions are obtained by exploring the feasible state space for the maximum, and the best-case solutions are the global optimal solutions in Tables 3.3 - 3.4

	$N = 64$		$N = 256$	
	Worst-case Solution	Best-case Solution	Worst-case Solution	Best-case Solution
K	6	4	6	4
S	10	2	12	5
L	10	23	15	29
$E(K, S, L)$	380.4	222.6	1191.6	831.9
$\overline{E}(K, S, L)$	5.944	3.478	4.655	3.250

3.6 Conclusion

The network infrastructure design problem requires the network parameters to be optimally set such that the power-related operating cost is minimised without compromising the overall performance, and modelling based on queueing network analysis provides a powerful tool for this purpose. In this chapter, a mathematical framework has been developed for the proposed transparent optical packet switch with recirculating optical buffers, based on Markov-modulated finite-server queueing models which were shown to be simple and analytically tractable. The framework follows a recursive structure where the overflowing MMPP traffic from the switch queue forms the arrival process to the buffering queue, and the departing MMPP traffic from the buffering queue contributes to the input process of the switch queue. Using the model, the behaviour of the switch has been numerically analysed, more particularly, to assess the impact of various network parameters on the contention probability and the overall latency. Convincing support provided by extensive simulations shows that the proposed mathematical performance model exhibits a good degree of accuracy for Poisson primary input process. Also, further experiments have been conducted with bursty primary input traffic, which is modelled by Interrupted Poisson Process (IPP), and the accuracy of the evaluation queueing model is validated by a comparative simulation study. All these findings indicate that the proposed analytical framework is sufficiently accurate, and applicable for different traffic classes, making it a practical and cost-effective evaluation tool to study the network performance. Finally, the mathematical framework is exploited for network dimensioning. The task of the dimensioning is to allocate the optimal network parameter setting which ensures network performance with the least power-related cost. In this work, the optimisation problem is addressed by adopting a heuristic search procedure, which provides fast convergence. To evaluate the performance of the proposed optimisation technique, additional experiments have been carried out. The numerical evidence shows that the obtained optimal solution can potentially improve the energy efficiency by 42% in the scenario of $N = 64$, in comparison with the power requirement of the worse-case solution, and for $N = 256$, the optimal solution results in a reduction of 30% in power consumption. Therefore, the algorithm provides a reliable solution for designing and dimensioning this type of optical switch.

Bibliography

[1] D. K. Hunter et al., "WASPNET: a wavelength switched packet network," *IEEE Commun. Mag.*, vol. 37, no. 3, pp. 120-129, 1999.

[2] R. Srivastava and Y. N. Singh, "Feedback fiber delay lines and AWG based optical packet switch architecture," *Opt. Switch. Network.*, vol. 7, no. 2, pp. 75-84, 2010.

[3] S. Pallavi and M. Lakshmi, "An AWG based optical router," in: *International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 245-248, 2014.

[4] C. M. Lee, C. C. R. Hui, F. Tong, and P. Yum, "Network dimensioning in WDM-based all-optical networks," in *Global Communications Conference (GLOBECOM'98)*, pp. 328-333, 1998.

[5] P. Lassila, A. Penttinen, and J. Virtamo, "Dimensioning methods for data networks with flow-level QoS requirements," in *IEEE Conference on Local Computer Networks*, pp. 353-359, Oct. 2007.

[6] J. Turner, "Terabit burst switching," *Journal of High Speed Networks* vol. 8, pp. 3-16, 1999.

[7] M. Yoo, C. Qiao, and S. Dixit, "QoS performance of optical burst switching in IP over WDM networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, pp. 2062-2071, Oct. 2000.

[8] J. Yan and G. Kuo, "Performance analysis of WDM optical buffers for asynchronous variable length packets," in *High Performance Switching and Routing (HPSR'05)*, pp. 302-305, 2005.

[9] J. Liu, H. Liu, and T. T. Lee, "Performance modeling of optical buffers supporting variable length packets," in *High Performance Switching and Routing (HPSR'05)*, pp. 535-538, 2005.

[10] C. McArdle, D. Tafani, and L. P. Barry, "Analysis of a buffered optical switch with general interarrival times," *Journal of Networks*, vol. 6, no. 4, pp. 536-548, Apr. 2011.

[11] C. McArdle, D. Tafani, T. Curran, A. Holohan, and L. P. Barry, "Renewal model of a buffered optical burst switch," *IEEE Communications Letters*, vol. 15, no. 1, pp. 91-93, Jan. 2011.

[12] D. Tafani, C. McArdle, and L. P. Barry, "A two-moment performance analysis of optical burst switched networks with shared fibre delay lines in a feedback configuration," *Journal of Optical Switching and Networking*, vol. 9, no. 4, pp. 323-335, Nov. 2012.

[13] K. S. Meier-Hellstern, "The analysis of a queue arising in overflow models," *IEEE Transactions on Communications*, vol. 37, no. 4, pp. 367-372, Apr. 1989.

[14] W. Fischer and K. S. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Performance Evaluation*, vol, 18, no. 12, pp. 149-171, Sept. 1993.

[15] T. Zhang, K. Lu, and J. P. Jue, "An analytical model for shared fiber-delay line buffers in asynchronous optical packet and burst switches," in *IEEE International Conference (ICC'05)*, pp. 1636-1640, 2005.

[16] X. Shao, L. Zhang, and P. Shum, "Modeling and performance analysis of realizable optical queue with service differentiation capability," *Journal of Computer Communications*, vol. 31, no. 15, pp. 3654-3661, Sept. 2008.

[17] N. Akar and Y. Gunalay, "Dimensioning shared-per-node recirculating Fiber Delay Line buffers in an Optical Packet Switch," *Performance Evaluation*, vol. 70, no. 12, pp. 1059-1071, Dec. 2013.

[18] J. Wang, C. McArdle, and L. P. Barry, "Energy-efficient optical packet switch with recirculating fiber delay line buffers for data center interconnects," in *International Conference on Transparent Optical Networks (ICTON)*, 2014.

[19] H. Heffes," A class of data traffic processes-covariance function characterization and related queuing results," *Bell System Technical Journal*, vol. 59, no. 6, pp. 897-929, July-Aug. 1980.

[20] R. Bellman, "Introduction to Matrix Analysis," Society for Industrial and Applied Mathematics, USA, 1997.

[21] M. Yu, "A new model reduction method for traffic described by Markov Modulated Poisson Processes," in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1-6, 2008.

[22] C. McArdle, D. Tafani, L. P. Barry, A. Holohan, and T. Curran, "Simplified overflow analysis of an optical burst switch with fibre delay lines," in *Sixth International Conference on Broadband Communications, Networks, and Systems (BROADNETS)*, pp. 1-8, 2009.

[23] MATLAB, URL: http://uk.mathworks.com/products.

[24] Discrete Event Simulator, OMNeT++, URL: http://www.omnetpp.org.

[25] Cisco 10GBASE SFP+ Modules, Datasheet, Cisco Inc. (2010).

[26] C. Kachris and I. Tomkos, "Power consumption evaluation of all-optical data center networks," *Cluster Computing*, vol. 16, no. 3, pp. 611-623, Aug. 2013.

[27] R. S. Tucker, "Scalability and energy consumption of optical and electronic packet switching," *Journal of Lightwave Technology*, vol. 29, no. 16, pp. 2410-2421, Aug. 2011.

Chapter 4

Flexible Optical Packet Switching for HPC and Datacentres

4.1 Introduction

As discussed in Section 1.1.2, the traffic requirements in datacentres evolve rapidly and constantly, and exhibit non-uniform communication patterns. To support the dynamics of traffic patterns and service differentiation of datacentre applications, a more flexible, more scalable and more bandwidth efficient datacentre network infrastructure with on-demand high-capacity service provisioning needs to be developed. This motivates the design of flexible optical switching solutions which allow network capacities to be dynamically assigned to traffic demands depending on the application needs. High flexibility is an important requirement in datacentre networking, which introduces tremendous benefits with regard to network reconfigurability, resource utilisation, Quality-of-Service (QoS) and scalability. Nevertheless, it is challenging to incorporate these features into optical networks due to additional network control and scheduling complexity.

To overcome this key challenge, this chapter introduces a novel energy-efficient and flexible optical packet-switched architecture for high-performance computing (HPC) systems and datacentre networks (DCNs) using passive optical components of the Arrayed Waveguide Grating (AWG) and Wavelength Selective Switches (WSSs). The AWG device, as mentioned previously, is a passive optical wavelength-routing component, which directs an incoming optical packet to a chosen output port depending on the wavelength of light used to carry the signal at the input. The AWG can facilitate building N-input$\times N$-output high-speed optical packet switches. Nonetheless, the inherent

drawback of the AWG is the large numbers of output receiver components required, deployment costs and energy consumption. The work carried out in the previous chapters has shown the potential of optical Fibre Delay Line (FDL) buffers in resolving contentions in AWG-based switches and so greatly reducing the numbers of output receivers required. Based on these findings, this chapter proposes to further enhance the contention performance and energy efficiency of the AWG-based architectures by incorporating a dynamic resource allocation algorithm into the switch. This adaptive feature is made possible through the deployment of a Wavelength Selective Switch (WSS) at each switch output, which is a flexible switching technology supporting finer-grained arbitrary switching granularity. The proposed dynamic resource allocation scheme supports the dynamic allocation of capacity to traffic demands, thus greatly improving network flexibility, making more efficient use of the shared network resources and considerably alleviating the contention conditions, which in turn reduces the required optical switching and buffering capacities.

The main contribution of this chapter is an extension of the architectures presented previously. The proposed novel switch network employs an AWG, WSSs and a novel method to allocate flexible bandwidth resources at switch outputs, so that the architecture can be re-configured to more efficiently switch non-uniform, non-constant hotspot datacentre traffic. It is common in datacentres, due to the nature of the deployed applications, that a Top-of-Rack (ToR) switch only communicates with relatively few other ToRs so that patterns in the source-to-destination network traffic typically show hotspots [1]. Similarly, in high-performance computing applications, traffic patterns are commonly observed to be degree-bounded [2]. Previous proposals for optical datacentre networks have not been evaluated under such traffic patterns, where uniform traffic is assumed in simulation performance studies. This chapter introduces a method for generating synthetic hotspot traffic for simulation studies, and then derives a load-sensitive algorithm which dynamically allocates bandwidth capacity resources based on the traffic demands in the network. Further, a tractable analytic model, which accurately approximates the contention performance of the switch node, is developed to facilitate the implementation of the proposed flexible resource allocation scheme. Evidence from the simulation results indicates that uniformly dimensioned output transmission capacity performs poorly under hotspot traffic and that the flexible allocation scheme considerably improves network performance, throughput and energy efficiency. Moreover, the work in this chapter also reports a method for choosing dimensions of the main switch components that meet packet loss targets whilst minimising the architecture's energy consumption.

The chapter is organised as follows. Section 4.2 presents the proposed switch architecture. Section

4.3 introduces the hotspot traffic model. Section 4.4 develops the flexible bandwidth assignment algorithm at switch output ports and in the FDLs. In Section 4.5, the network performance and power consumption of the proposed switch architecture are evaluated and compared with the static bandwidth allocation. Finally, Section 4.6 outlines some important conclusions that can be drawn from the experimental results.

4.2 OPS Architecture

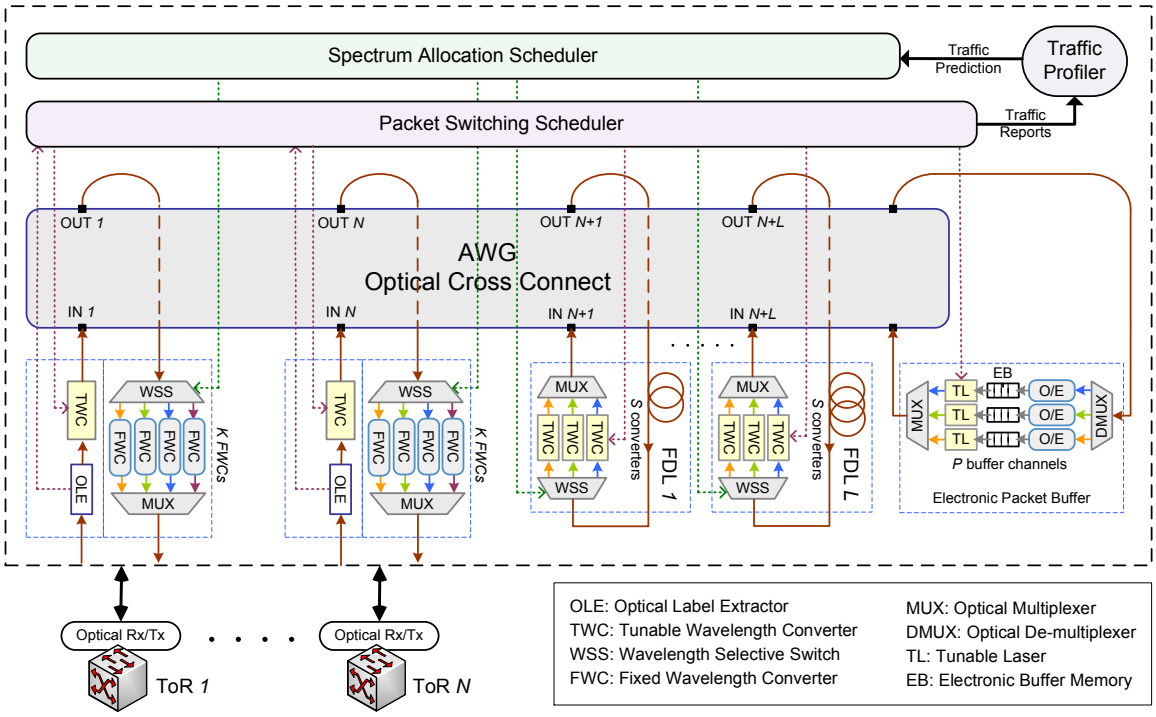


Figure 4.1: Architecture of proposed flexible-bandwidth datacentre optical packet switch.

The details of the proposed all-optical packet switch are shown in Figure 4.1. Multiple Top-of-Rack (ToR) switches are directly interconnected by the Optical Packet Switch (OPS) via high-bandwidth links, 10 or 40 Gbit/s optical links. Using a single optical switch, which could be scaled to the order of 512 ports [3], and assuming each ToR is connected to 48 servers, a medium-scale network of 24,576 servers is realisable. Figure 4.2 indicates how larger networks of hundreds of thousands of servers can be realised by additional layers in the switching hierarchy. The design of a large-scale datacentre network using the proposed OPS architecture will be explored in the next chapter, and the focus in this chapter is particularly on how a flexible bandwidth provisioning scheme can be realised in an optical architecture and the advantages this brings in terms of switch throughput and energy

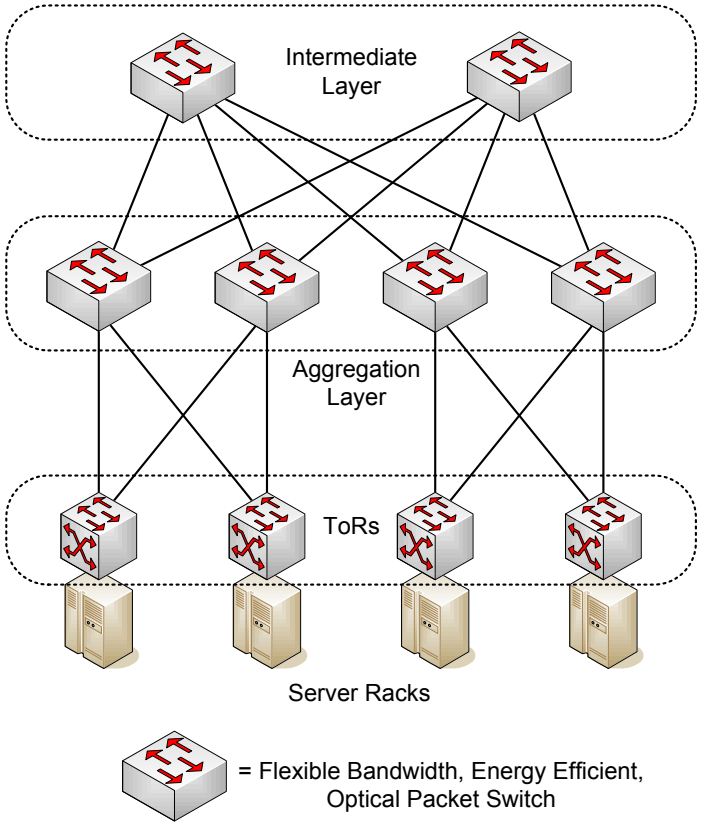


Figure 4.2: An example network architecture (Clos Network) using multiple identical optical switches to build out a large datacentre network.

efficiency. The study in this chapter focuses on the single switch scenario, but note that flexible bandwidth provisioning in a large interconnection network has similar advantages, allowing over-subscription of link bandwidth to be reduced, compared to statically provisioned networks.

The core of the proposed switch is the arrayed waveguide grating (AWG) router, which has $N+L+1$ input and output ports. The input port of the switch carries a single, fixed optical wavelength channel. The wavelength of an incoming optical packet is tuned at switch input by the Tunable Wavelength Converter (TWC), so that the AWG directs the packet on the new wavelength to the required output port. An AWG can carry multiple different packets on different wavelengths simultaneously and is itself a non-blocking wavelength router. These multiplexed optical channels from the AWG output port are directed to the wavelength selective switch (WSS) which can be configured to split arbitrary sub-sets of its incoming optical channels to multiple (K) outputs. Each WSS output is directed to a Fixed Wavelength Converter (FWC) which converts the input wavelengths to a fixed transmission wavelength, and then all the converted signals are optically multiplexed by the Multiplexer (MUX) and transmitted from the switch output port back towards the ToR.

As there are fewer transmission channels (K) at output port than there are input wavelength channels (N), contention arises at switch output as each transmission channel (FWC) can carry only one wavelength at any time. To deal with the potential packet loss due to collisions, contending packets are directed to different length Fibre Delay Line (FDL) buffers, where they are delayed until a transmission channel (FWC) in the required output port becomes free. Fibre delay lines may carry multiple channels, but have a limited number of associated TWCs (S) to redirect delay packets back to the intended output port, thus packet contention can additionally occur in the FDLs. None-the-less the majority of buffered traffic can be accommodated with FDLs and the remainder directed to a small-scale electronic buffer, having P input channels, with each input channel offering buffering to $(N + L)/P$ wavelengths. The packet scheduler receives packet headers (via the OLEs) and controls all signaling for optical packet routing and scheduling.

The prominent feature of the architecture is that flexible wavelength allocation at output ports is made possible by the WSS, where the incoming wavelength channels can be arbitrarily divided and assigned to outgoing channels (FWCs) in accordance with expected source-destination traffic flow volumes between switch ports (between the connected ToRs). Additionally, to achieve better wavelength channel utilisation in FDL buffers, the flexible wavelength allocation is also applied to FDLs. This flexibility provides a finer match between the required and provided bandwidths, and thus significantly enhances resource utilisation under non-uniform and changing traffic conditions. Given the increased efficiency, the dimensions of the main power consuming components in the architecture can be reduced, avoiding over-dimensioning of an architecture with fixed bandwidth allocation.

The performance of optical switch architectures are usually evaluated assuming uniform loading, where all input ports send equal traffic load to all output ports. To evaluate the reconfigurable switch architecture of this paper, a more realistic non-uniform traffic scenario is considered, which will be reported in the next section.

4.3 Hot-spot Traffic Model

As stated previously in Section 1.1.2, due to the fact that a ToR switch prefers to communicate with a small subset of ToR switches close by, the communication pattern of datacentre applications is extremely sparse and skewed, and includes a number of hot-spots, thereby the traffic pattern is a sparse matrix with relatively high diagonal values. Based on these traffic characteristics, a hot-spot

traffic matrix representing realistic traffic demands in datacentre/HPC applications is estimated. It is assumed that, due to the implementation of the application task placement technique, the overall switch load is balanced, that is, each ToR sends and receives the same aggregate traffic volume, however, individual source-destination traffic volumes are not balanced, i.e. a ToR may send different traffic loads to different destination ToRs. This switch loading scenario is intended to reflect a deployment of different applications distributed across the set of datacentre servers, where each application has different communications requirements between its allocated servers, but under the premise that the allocation of servers to application modules has been controlled to achieve an overall balanced mean traffic load between ToRs, with a view to reducing congestion conditions.

The mean traffic loads between N ToRs are described using the $N \times N$ traffic matrix $\mathbf{D}$, where $\mathbf{D}(i,j)$ is the mean traffic load emanating from ToR i (connected to input port i of the switch) which is destined for ToR j (connected to output port j of the switch). As the aggregate load is required to be balanced, $\mathbf{D}$ must be a doubly stochastic matrix up to a scalar constant, that is, all column and row sums are equal to $\overline{M}, 0 \leq \overline{M} \leq 1$, the aggregate input (or output) load per switch port, that is

$$\sum_{\forall i} \mathbf{D}(i,j) = \overline{M},\ \forall j \quad \text{and} \quad \sum_{\forall j} \mathbf{D}(i,j) = \overline{M},\ \forall i.$$

Note that a ToR does not use its optical input/output switch port to route traffic between its directly connected servers, so the diagonal entries of $\mathbf{D}$ must be zero. It is assumed that traffic patterns are dominated by *hotspot* traffic, groups of servers distributed across multiple ToRs that exchange large volumes of traffic. To represent multiple different hotspots, $\mathbf{D_{blk}}$ is constructed with a block-diagonal form, where

$$\mathbf{D_{blk}} = \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{S}_l \end{bmatrix}$$

with l blocks (hotspots) having dimensions $b_1 \times b_1, b_2 \times b_2, \ldots, b_l \times b_l$, where $\sum_{k=1}^{l} b_k = N$. In order to control the proportions of small and large flows between ToRs in a hotspot and also have some randomization in the traffic matrix, the following strategy is adopted to construct each block. To construct each $\mathbf{S}_k, k = 1, 2, \ldots, l$, a *seed vector* $\mathbf{s}_k = [0, s_1, s_2, \ldots, s_{b_k-1}]$, with $\sum_{i=1}^{b_k-1} s_i = \overline{M}$, is chosen as the first row of the block. The second row of the block is formed by right-shifting and cycling the entries of $\mathbf{s}_k$, and similarly for subsequent rows, to yield a circulant matrix with

zero diagonal

$$\mathbf{S'}_k = \begin{bmatrix} 0 & s_1 & s_2 & \cdots & s_{b_k-2} & s_{b_k-1} \\ s_{b_k-1} & 0 & s_1 & s_2 & \cdots & s_{b_k-2} \\ \vdots & & & & & \vdots \\ s_2 & \cdots & s_{b_k-2} & s_{b_k-1} & 0 & s_1 \\ s_1 & s_2 & \cdots & s_{b_k-2} & s_{b_k-1} & 0 \end{bmatrix}.$$

The columns of $\mathbf{S'}_k$ are randomly permuted and then the rows are exchanged to restore the zero diagonal, yielding the matrix $\mathbf{S}_k$, the k^{th} block of $\mathbf{D_{blk}}$. As each matrix $\mathbf{S'}_k/\overline{M}$ is doubly stochastic, and permuting columns and exchanging rows preserves column/row sums, then it follows that $\mathbf{D_{blk}}/\overline{M}$ constructed from the $\mathbf{S}_k$ blocks is also doubly stochastic. To include traffic other than hotspot traffic in the final matrix $\mathbf{D}$, using the same method as for $\mathbf{S}_k$, a random matrix $\mathbf{D_{rnd}}$ of size $N \times N$ is formed, starting with a random seed vector of length N, and subsequently a scalar multiple of this is added to $\mathbf{D_{blk}}$ and the formulate is then re-scaled to yield the resultant traffic matrix

$$\mathbf{D} = \frac{\mathbf{D_{blk}} + \alpha \cdot \mathbf{D_{rnd}}}{1 + \alpha},$$

which has mean aggregate loads, at each input and each output port, equal to $\overline{M}$. A 128×128 traffic matrix generated by this method, and used later for simulations, is illustrated in Figure 4.3.

There are two 8×8 hotspots and three hotspots of dimension 16, 32 and 64, respectively. $\mathbf{D_{rnd}}$ has been constructed to be sparse and the scalar parameter α is set to be $\alpha = 1/3$, so the block (hotspot) form of $\mathbf{D}$ dominates the traffic pattern, whilst each group of hotspot ToRs is not completely isolated from others.

4.4 Elastic Wavelength Allocation

A novel aspect of the proposed switch is the use of adaptive resource allocation, which is performed by the spectrum allocation scheduler in the switch. The scheduler is responsible for optimally configuring which input wavelengths (input port traffic streams) are directed to which output receivers, in each output port. This section details the algorithm implemented by this scheduler that aims to minimise output port congestion for any given non-uniform traffic load matrix.

Given a traffic matrix $\mathbf{D}$, an output port j will receive a set of loads $\mathbf{A} = \{\mathbf{D}(i,j), i = 1, 2, \ldots, N\}$ from the input ports. Traffic from each input port arrives on a separate wavelength channel. These

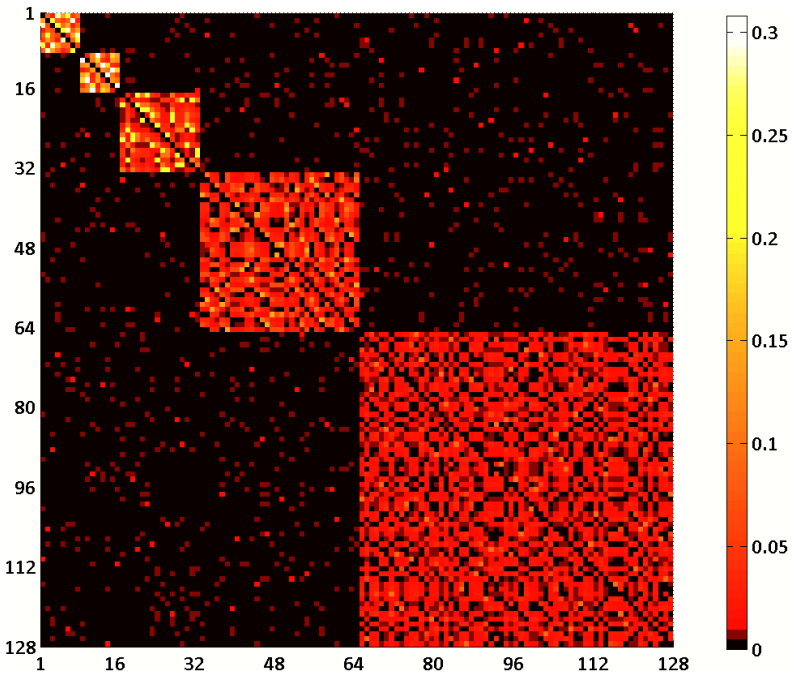


Figure 4.3: Heatmap of 128×128 hotspot traffic matrix, generated using the random circulant block-matrix scheme. The matrix has full load ($\overline{M} = 1$), with individual source-destination loads ranging between 0% and 30%.

incoming wavelengths are to be partitioned into groups, by appropriate configuration of the WSS, and each group is assigned to one of K FWCs in the output port, with the objective of balancing load across the outgoing channels. Each output port is treated separately, in Stage I of the algorithm detailed below. Additionally, some of the direct input load is initially blocked and must be buffered by FDLs, which require a similar load-balancing treatment of the loads directed to each TWC in each FDL. The traffics returning from FDLs must also be re-assigned to their destination output port channels, whilst preserving load balance. Stage II of the algorithm handles this FDL allocation procedure.

4.4.1 Stage I: Direct Input Traffic Assignment

Let the set $\{B_k, k = 1, \ldots, K\}$ be some partition of $\mathbf{A}$, the set of loads directed to a given output port j. Each member of the partition corresponds to an assignment of a group of the input port loads (group of wavelengths) to a particular transmission channel (FWC) in the output port in question. Let m_k denote the sum of the load values in B_k. Thus, the average load over all transmission

channels in the output port is

$$\overline{m} = \frac{1}{K} \sum_{k=1}^{K} m_k.$$

Defining the variance function $V(m_k) = (m_k - \overline{m})^2$, the load-balancing optimisation objective is to choose a partitioning $\{B_k, k = 1, \ldots, K\}$ that minimises the mean variance of loads over all transmission channels (FWCs) in the output port, formally

$$\underset{B_k}{\text{minimize}} \quad \sum_{k=1}^{K} V(m_k).$$

An iterative algorithm that implements this objective is listed as Algorithm 2. In summary, the algorithm firstly assigns the K largest loads in $\mathbf{A}$ to K groups. Then, one by one and starting with the smallest, each of the remaining $(N - K)$ loads are added to these K groups, such that each assignment gives a reduction in $V(m_k)$ of the target group, i.e. brings the mean load of the group closer to $\overline{m}$. Thus an initial partition of $\mathbf{A}$ is formed. The remainder of the algorithm then iteratively moves loads from one group to another, seeking the largest improvement in the objective function with each move, until convergence, when no further improvement between iterations is achieved. The procedure is repeated separately for each output port.

4.4.2 Stage II: FDL Traffic Assignment

In order to load-balance the FDL wavelength group allocations, it is crucial to estimate the input load to FDLs. This load is composed of the overflow (blocked) traffics at output ports. Each output port FWC (output WSS wavelength group) receives a portion of the total traffic emanating from each input port, where the load values are determined by the optimal load partitioning $\{B_k\}$ decided in Stage I of the algorithm. As an input port splits its load to several output ports, it is necessary to consider blocking of all input traffic streams to a FWC (transmission channel) simultaneously in the calculation, as follows.

Let $\rho_{i,j,k}$ be the load from input port i intended for FWC k in output port j, where $\rho_{i,j,k} = \mathbf{D}(i, j)$ if this input wavelength i has been assigned to FWC k by the partitioning $\{B_k\}$ for output port j. Otherwise $\rho_{i,j,k} = 0$. This particular FWC is denoted with label $R(j, k)$. At each input port, packets are assumed to have random independent identically distributed (iid) transmission times. Additionally, idle times between packet transmission periods are also iid. Each FWC in an output port can be modelled as a single server with no queue. Under these traffic assumptions, the call

Algorithm 2 Load Balancing Algorithm

Given $\mathbf{A}$ (vector of loads to a given output port) and K (number of transmission channels (FWCs) in output port) *% $\mathbf{A}$ consists of N elements*

procedure 1 (Initialisation)

$\hat{\mathbf{A}} \leftarrow \mathbf{A}$ sorted in descending order *% the numbers in $\mathbf{A}$ are arranged in descending order*

for $k = 1 \ldots K$ **do** *% take the first (largest) K elements in $\hat{\mathbf{A}}$ and put them into K different groups*

$\mathbf{B}_k \leftarrow \hat{\mathbf{A}}(k)$

end for

$k \leftarrow 1$

$i \leftarrow N$ *% for the remaining $(N-K)$ numbers in $\hat{\mathbf{A}}$, start with the last element, each number is added to a group*

while $k \leq K \& i > K$ **do**

loop :

$\hat{\mathbf{B}}_k \leftarrow \mathbf{B}_k \cup \hat{\mathbf{A}}(i)$ *% if the mean variance becomes smaller when putting the element $\hat{\mathbf{A}}(i)$ into group k*

if $V(\hat{m}_k) < V(m_k)$ **then**

$\mathbf{B}_k \leftarrow \hat{\mathbf{B}}_k$ *% then the element $\hat{\mathbf{A}}(i)$ is added to group k*

$i \leftarrow i - 1$ *% move on to the next element*

if $i > K$ **then**

goto *loop*

end if

end if

$k \leftarrow k + 1$ *% otherwise, move on to the next group*

end while

end procedure

procedure 2 (Iterative Load Reassignments)

while not converged **do** *% repeat until no further improvement*

for $j = 1$ to N **do** *% start with the first element in $\mathbf{A}$*

remove $\mathbf{A}(j)$ from its original group

for $k = 1$ to K **do** *% group index $k = 1, 2, \ldots, K$*

$\overline{\mathbf{B}_k} \leftarrow \mathbf{B}_k \cup \mathbf{A}(j)$ *% put the element $\mathbf{A}(j)$ into group k*

$\overline{\Delta}(k) \leftarrow \Delta$ for set $\{\mathbf{B}_1 \ldots \overline{\mathbf{B}_k} \ldots \mathbf{B}_K\}$ *% a mean variance $\overline{\Delta}(k)$ is obtained*

end for *% In total, K mean variances are gained*

$l \leftarrow \overset{\arg\min}{l} \overline{\Delta}$, subject to: $l \in \{1 \ldots K\}$ *% choose the smallest mean variance $\overline{\Delta}(l)$*

$\mathbf{B}_l \leftarrow \overline{\mathbf{B}_l}$ *% assign the element $\mathbf{A}(j)$ to group l*

end for

end while

end procedure

congestion (aka blocking probability) at this server (FWC) is the probability that an arbitrary arriving packet finds the server busy. With respect to an arriving packet generated at input port i and directed to output FWC $R(j,k)$, the blocking probability is equal to the carried load at the server due to all *other* packets from inputs other than input i (by way of the *arrival theorem* for finite-source traffic). This observation yields a system of N linear independent equations in the blocking probability ($B_{i,j,k}$) of each input load offered to receiver $R(j,k)$:

$$B_{i,j,k} = \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \rho_{i',j,k} \cdot (1 - B_{i',j,k}), \quad i = 1, 2, \ldots, N.$$

It is important to note that this system of linear equations is easily transformed to have a coefficient matrix of Toeplitz form (by way of pre-multiplying the system by $[[1]^{N \times N} - 2 \cdot \mathbf{I}^{N \times N}]$) and so has an efficient and numerically stable solution method. By comparison with simulation results, it has been verified that this analytic model is exact, under the iid traffic assumptions given above.

Having solved for all $B_{i,j,k}$, the mean overflow traffic at FWC $R(j,k)$ is then $\hat{\rho}_{i,j,k} = B_{i,j,k} \cdot \rho_{i,j,k}$. The aggregate load that is lost from the load at input port i is then

$$\hat{\rho}_i = \sum_{\forall j, \forall k} \hat{\rho}_{i,j,k} \quad i = 1, \ldots, N.$$

The overflowing loads $\hat{\rho}_i$ form the input loads to FDL 1. Some of the traffic in FDL 1 is blocked, which is offered to FDL 2, and so on down the chain of L FDLs. Thus, in this case each FDL is treated as a blocking system generating its own overflowing loads. For allocating input wavelength groups to the TWCs in FDLs, each FDL is dimensioned using the same algorithm as Stage I but with number of transmission channels in each FDL being S, rather than K.

Finally, the traffic that is carried by the FDLs and subsequently re-offered to the designated switch output ports, is re-allocated to output port wavelength groups as follows. Traffic delayed in the shortest FDL is assigned to the group with the current lowest load allocation. Load returning from the next shortest FDL is allocated similarly, and so on in a round-robin fashion until all FDL return loads are allocated to their target output ports. It is assumed in this that all load offered to the FDLs is carried, which is approximately the case as the analysis is targeting very low overall switch blocking probabilities ($< 10^{-6}$).

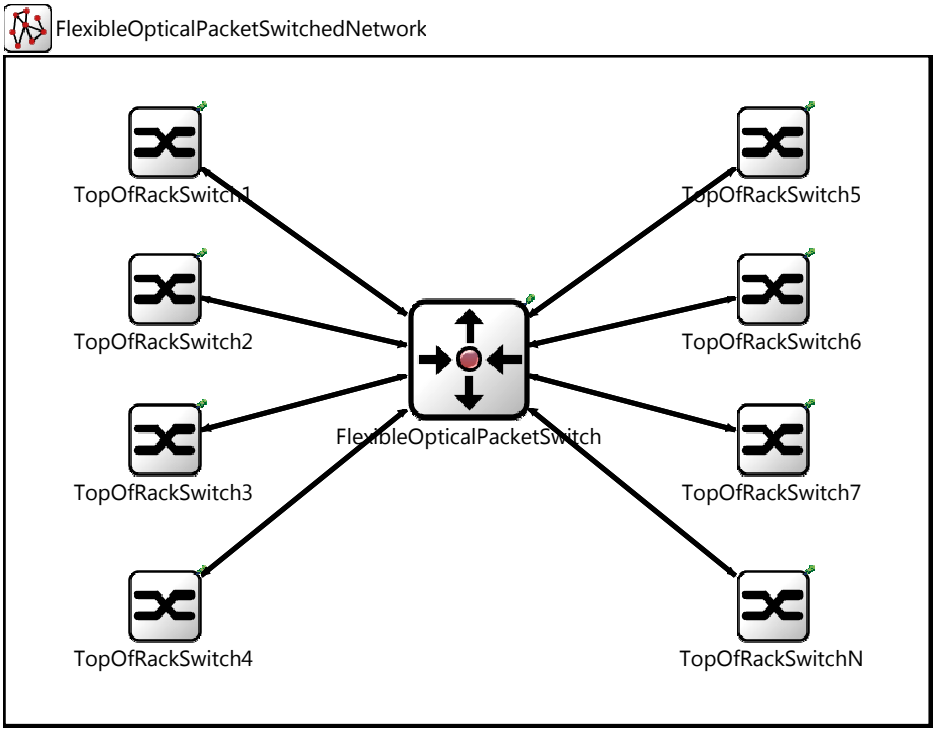
4.5 Performance Evaluation

In the proposed optical packet-based network, not only is the switch architecture being considered, but also the wavelength assignment control. The experiments and results described in this section indicate the efficacy of the flexible Wavelength Allocation (WA) scheme, in regard to packet blocking performance and energy efficiency of the proposed switch under both uniform and non-uniform traffic load. The previously presented hotspot traffic matrix (Section 4.3) is used for the input load parameters in a discrete-event simulation model of the presented optical network. Through simulations, analyses of the impacts introduced by various network parameters on network performance have been conducted, so as to resolve the network optimisation problem where the objective is to minimise the total network hardware components whilst achieving the target packet loss performance. The optical system performance of various switch sizes is investigated.

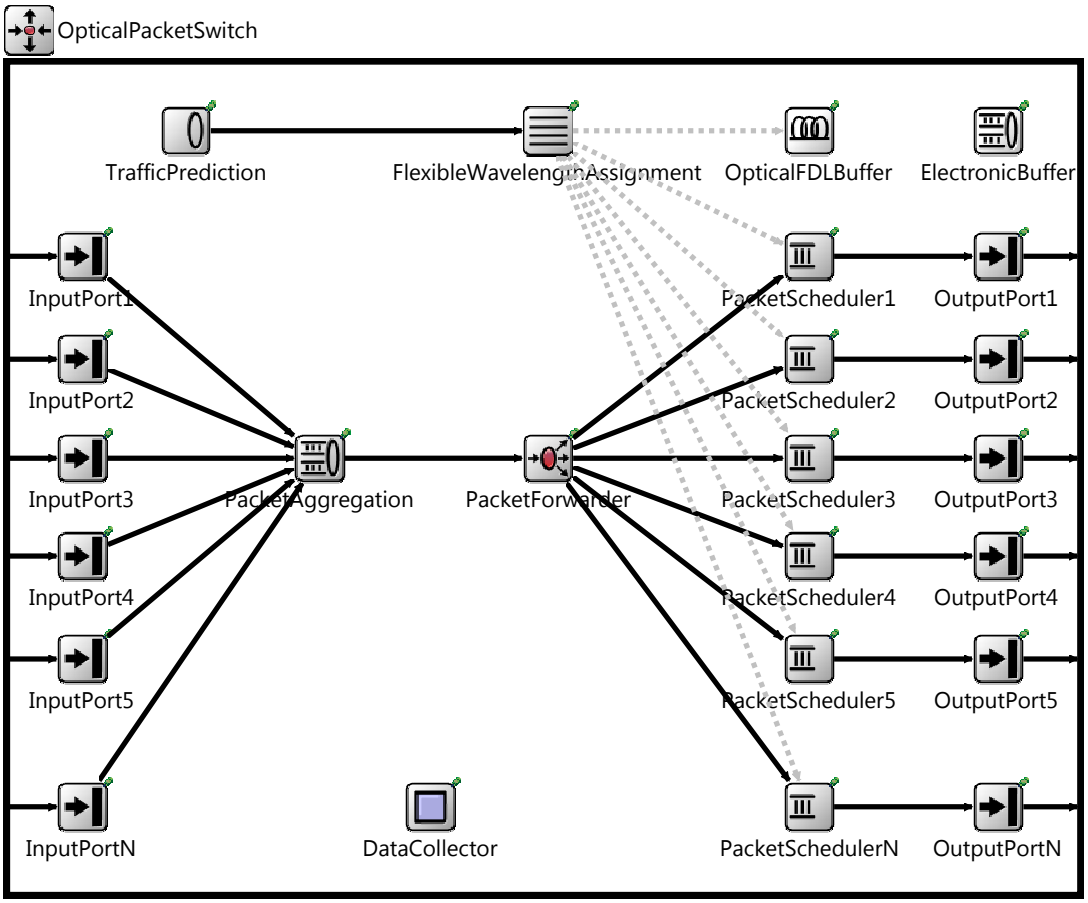
4.5.1 Simulation Setup

The simulation model of the optical switching network is implemented in OMNeT++ [4] which is an Object-Oriented Programming (OOP) discrete-event simulation tool as discussed in Section 2.3.1. The experimental setup is shown in Figure 4.4. Table 4.1 tabulates the network parameters and the performance measurements used in the experiments. The simulator is an enhanced version of the simulation model detailed in Section 2.3.1, which models the traffic generation and detection processes, the flexible wavelength assignment scheme based on a load-balancing optimisation algorithm, optical packet switching, optical FDL buffering and electronic buffering.

Similar to the previously described model in Section 2.3.1, traffic is generated by the ToR switches and transmitted towards the flexible OPS node via optical fibres. Note that the traffic load matrix in the network follows the traffic pattern reported in Section 4.3. In OPS, see Figure 4.4(b), the traffic streams are aggregated by the Packet Aggregation block and then directed to the Packet Forwarder module. Based on the destination node of the packet, the Packet Forwarder sends the packet to the appropriate Packet Scheduler which is associated with the packet destination. The Packet Scheduler module performs the packet scheduling process in a similar manner as described in Section 2.3.1. The difference is that the flexible wavelength assignment, as introduced in Section 4.4, is implemented in the switch output ports and FDLs, with the objective to adaptively balance the traffic loads among different transmission channels in each output port, and among different buffering channels in each FDL. This optimisation algorithm is implemented by the Flexible Wavelength



(a) The simulation model of the flexible optical packet-switched network



(b) The simulation model of the flexible OPS node

Figure 4.4: Simulation setup of the flexible optical network.

Assignment module, based on the changing traffic loads recorded in the Traffic Prediction module. Importantly, in order to balance the loads among different buffering channels in the FDL, a mathematical model of the switch node is developed so as to estimate the traffic loads overflowing from the switch outputs to the FDL buffers. This approximation procedure is also conducted in the Flexible Wavelength Assignment module. Given the wavelength assignment results and the source and destination nodes of the packet, the Packet Scheduler determines the assigned output transmission channel for the arriving packet, and then attempts to reserve the allocated output channel. If this reservation is unsuccessful, the scheduler checks the availability of the assigned FDL buffering channels in the degenerate FDLs. In the case that an FDL exists which resolves the contention, the scheduler reserves the required FDL channel and the corresponding output link. However, if the FDL/outgoing channel reservation fails, the scheduler reserves the electronic buffer (if switched on) and the corresponding new assigned outgoing channel.

Table 4.1: Table of Notation

N	Number of incoming/outgoing fibres in OPS. It also represents the number of the ToR switches in the network.
K	Number of transmission channels (FWCs) per switch output
L	Number of Fibre Delay Lines (FDLs)
S	Number of channels (TWCs) per FDL
D	Unit delay in FDL buffers. The discrete set of delays generated by FDLs are integer multiples of D, i.e., D, $2D$, ..., LD
τ	Mean packet transmission time
$\overline{M}$	Mean traffic load generated by the ToR switch

In the simulations, the traffic generation process follows an Engset traffic model which has exponentially distributed packet transmission times and exponentially distributed idle times between packet transmission periods. The mean traffic load $\overline{M}$ generated by a ToR switch is fixed to be 80%. All scenarios considered focus on using OPS with $N = 128$ ports, each port connected to a ToR switch via input/output fibre links. The input fibre of the OPS carries a single transmission channel, and the number of outgoing channels at each output port of the OPS is K, which is equivalent to the number of FWCs installed at each switch output. L is the number of FDLs utilised, and each FDL has S buffering channels (TWCs). The base delay of FDLs, denoted by D, is fixed to be 0.2τ, where τ is the mean packet duration. It should be noted that degenerate FDLs are adopted here, that is, the delays provided by FDLs are integer multiples of D, i.e., D, $2D$, ..., LD.

In the next section, the experimental work is carried out and the numerical results are presented and discussed. The analysis starts with the performance evaluation of the network scenario where there

is no additional electronic buffer in the architecture. The purpose is to illustrate buffering of the majority of the traffic in FDLs, and then, with the addition of a small scale electronic buffer, show how blocking probability can be further reduced to very low levels.

4.5.2 Results and Discussions

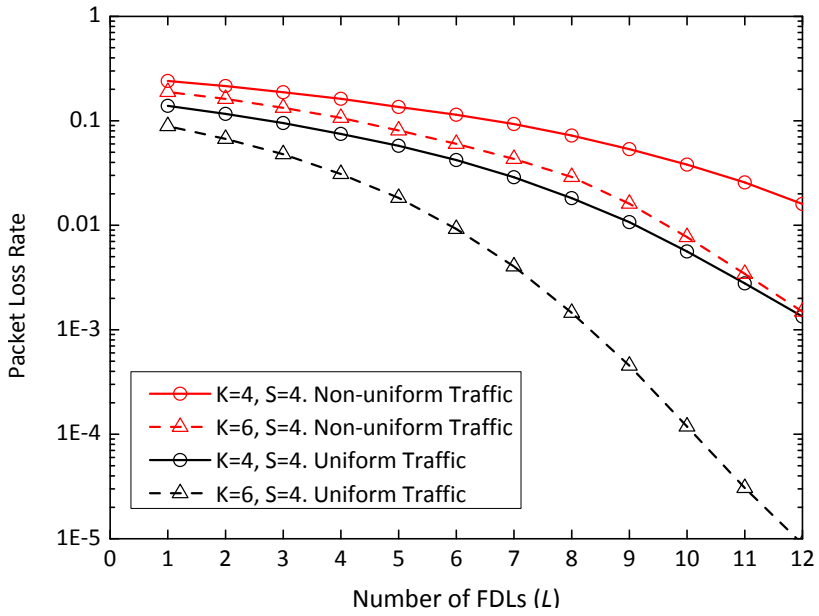
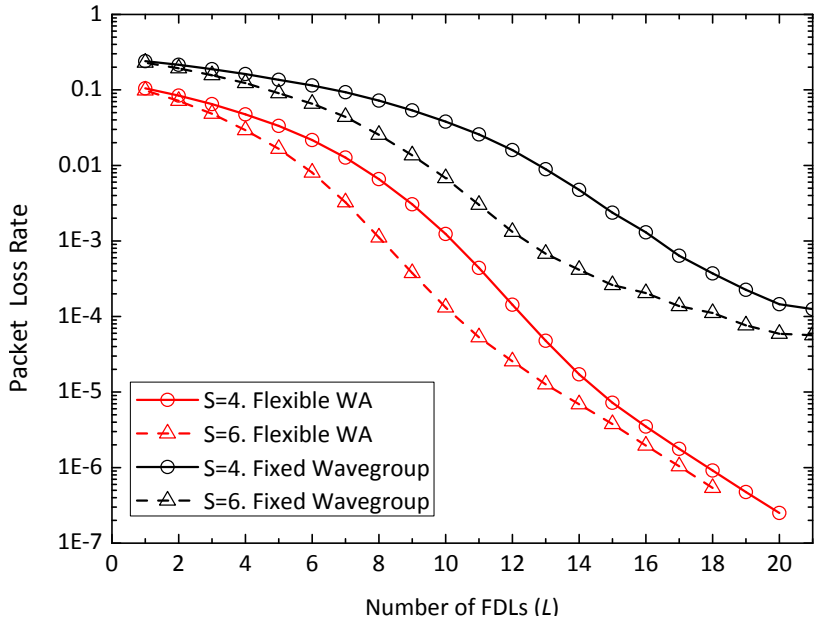
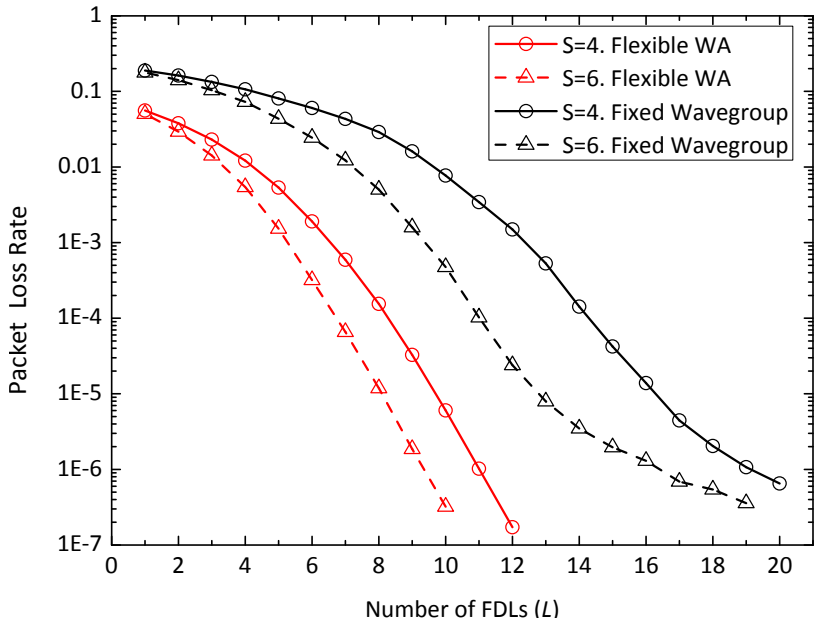


Figure 4.5: Packet loss rates in switch with fixed Wavelength Allocation (WA) scheme under uniform and non-uniform (hotspot) traffic, for varying switch dimensions: number of FDLs L, number of transmission channels (FWCs) per output port K, number of TWCs per FDL S.

The motivation for the use of flexible resource allocation in the switch architecture under study is first illustrated by evaluating the impact of hotspot traffic on a switch architecture where the output bandwidth capacity have a fixed dimensioning. In the fixed allocation scheme, each of K FWCs in an output port is assigned a block of N/K contiguous wavelengths (input source streams), regardless of the traffic loads. That is, all FWCs (transmission channels) are equally dimensioned and the allocation scheme does not take account of or change with the traffic demands. A fixed assignment of wavelengths to TWCs in the FDLs is similarly made. Figure 4.5 indicates that, with proper dimensioning of K and S, this fixed configuration can effectively carry uniform traffic (all source-destination loads equal) but that, under hotspot traffic, performance deteriorates dramatically. It is clear to see from the figure that under non-uniform traffic the optical switch with fixed resource assignment suffers from severer contention issue. This consequently highlights the necessity of deploying flexible resource allocation scheme in the presented optical packet switch.



(a) Number of transmission channels (FWCs) per switch output port $K = 4$



(b) Number of transmission channels (FWCs) per switch output port $K = 6$

Figure 4.6: Comparison of packet loss rates for fixed and flexible allocation schemes under hotspot traffic load.

To reveal the benefits of the proposed flexible WA, Figure 4.6 compares packet loss performance of fixed uniformly-allocated switch network to the proposed flexible wavegroup allocation technique, when both are under hotspot traffic. Shown is the packet loss rate versus the number of FDLs L for $K = 4$ and $K = 6$. In all scenarios shown, the flexible scheme delivers a significantly lower packet loss rate than the fixed scheme, by almost four orders of magnitude in the case of $K = 6$ and this difference is expected to be further increased as L increases. This is due to the fact that the flexible wavelength allocation makes better use of the shared network resources, and thus yielding lower packet contention probability.

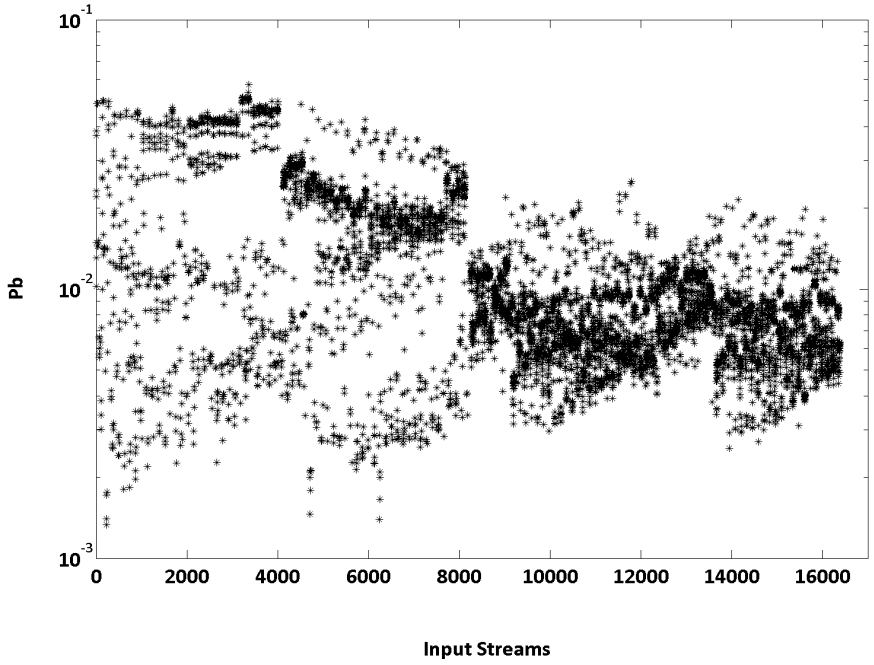


Figure 4.7: The packet loss difference between fixed resource allocation and flexible resource allocation, $Pb = P_{\text{fixed}} - P_{\text{flexible}}$, for each source-destination traffic stream in the scenario $K = 4$, $S = 4$, $L = 12$

Although the aggregate packet loss rates for the flexible scheme can be seen to be favorable, it is also important to examine the fairness of the resource allocation across individual source-destination traffic streams. In response to this, packet loss probabilities of both fixed and flexible resource allocation schemes for each source-destination traffic stream, denoted by P_{fixed} and P_{flexible} respectively, are evaluated. Figure 4.7 indicates the packet loss difference, Pb, between the two wavegroup allocation schemes, where $Pb = P_{\text{fixed}} - P_{\text{flexible}}$. Intuitively, there is an improvement in packet loss performance for all source-destination pairs, for the flexible allocation scheme. Figure 4.8 further shows this improvement on a per input stream basis. In this case, the contention performance difference between the two schemes is measured for each input port which corresponds to a source

ToR switch.

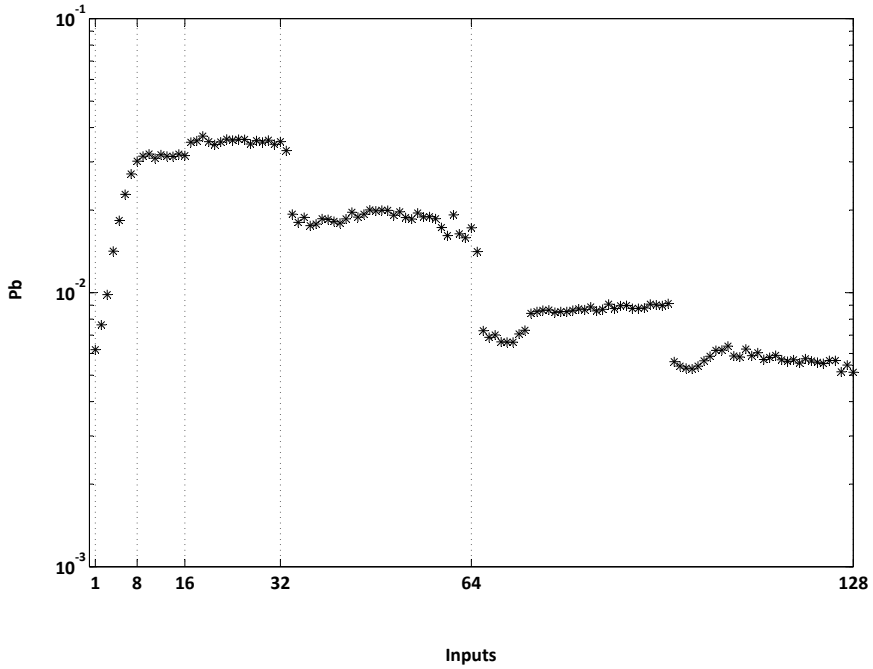


Figure 4.8: The packet loss difference between fixed resource allocation and flexible resource allocation, per input port.

4.5.3 Power Consumption Minimisation

The flexible allocation scheme allows component counts (output port FWCs, FDLs, FDL TWCs) to be reduced, thus resulting in a considerable reduction of energy consumption, compared to the fixed scheme, when dimensioning the switch to achieve a given low packet loss rate. In this section, the major subject is to evaluate the power saving possibilities when the flexible allocation scheme is utilised.

It can be appreciated from Figure 4.6 that there are multiple switch configurations (values of L, K and S) that could achieve the same packet blocking rate. However, each configuration will have different component counts and a different overall switch power consumption. By using discrete event simulation, multiple solutions (L, K and S values) that achieve a loss probability of $\leq 10^{-6}$, are attained under the constraint that the number of FDLs is limited to less than 22, in order to constrain the maximum latency. Having found candidate solutions, the power consumption for each solution is estimated based on the corresponding component count and estimated power consumption per component [5–9].

Table 4.2 presents the solutions found for both flexible and fixed allocation. It is noticed that flexible WA has four solutions, while fixed WA only has two solutions. Comparing the minimum power dissipation, a significant power reduction, up to 25%, is obtained by flexible WA, and it is anticipated that more energy savings can be achieved by performing a full-area searching, which indicates that the flexible scheme can significantly enhance network scalability in terms of both energy efficiency and system performance.

Table 4.2: Power consumption for a target packet loss rate of 10^{-6}

Variables	Solutions					
	Flexible Wavelength Assignment				Fixed Wavelength Assignment	
	1	2	3	4	1	2
K	4	4	6	6	6	6
S	4	6	4	6	4	6
L	18	17	11	9	19	17
Power (W)	261.6	306.6	296.4	311.4	344.4	383.4
Minimum Power	261.6				344.4	
Power per Port	2.04				2.70	

4.5.4 Hybrid Buffering

Although FDLs can buffer the majority of traffic in the proposed architecture, when lower blocking probabilities than 10^{-6} are required (which would generally be the case in a datacentre network), adding the required numbers of FDLs would become inefficient. The remaining traffic to be buffered becomes more correlated and an electronic buffer is then more efficient than fixed length FDLs. For that, a small-size electronic buffer with P input/output ports is included into the optical-buffered switch network, resulting in a hybrid-buffered optical network, as illustrated in Figure 4.1. In this subsection, a performance analysis is conducted for the proposed flexible switch architecture with hybrid buffer. The optimal network configuration $(4, 4, 18)$ obtained in the previous section is utilised here. The simulation results in Table 4.3 below show the impact of adding a small ($P \times P$) electronic loop-back buffer to the optical switch on packet loss performance. As expected, increasing the electronic buffer size P yields substantial reduction on packet loss rate. With the addition

Table 4.3: Blocking probability versus electronic buffer size P (P input/output ports)

Electronic Buffer Size P	1	2
Packet Loss Rate	3.12e-008	6.86e-009
90% Confidence Interval	$\pm 2.81\%$	$\pm 8.23\%$

of a 2×2 electronic buffer, the packet loss rate is reduced from 10^{-6} to the order of 10^{-9}. On the basis of this, it can be concluded that even a small dimension buffer could reduce overall packet loss to a sufficiently low level (10^{-9}) with little impact on switch power consumption. Alternatively, it is important to point out that in the hybrid buffer, for a target packet loss rate, increasing the electronic buffer size P allows the required FDL buffering capacity to be scaled, and accordingly the cost/power associated with the TWCs in FDLs is lowered. Therefore, when dimensioning the hybrid buffer, there exists a trade-off between the required FDL buffering capacity and the electronic buffer size P.

4.6 Conclusion

High performance computing systems and datacentres already deploy fibre carrying 10/40/100 Gigabit Ethernet to interconnect ToRs and aggregate/spine switches, however, the electronic-optical-electronic (E/O/E) conversion process at each end of the links is power hungry. By avoiding this conversion, optical switching can alleviate a substantial portion of this problem. This chapter proposes a novel optical packet-switched architecture, which is designed based on Arrayed Waveguide Grating (AWG) device and Wavelength Selective Switches (WSSs). A key feature of the switch is its use of dynamic wavelength assignment technique to enhance the network flexibility and resource utilisation. A major challenge for the proposed optical packet switching scheme is the lack of an optical equivalent of the electronic buffer. Although previous schemes suggest using electronics to buffer all traffic, the buffer itself becomes a dominant power consumer. The architecture presented in this chapter shows the energy efficiency of FDL buffering and further shows that by combining fibre and electronic buffers, and by allowing network resources to be dynamically allocated to the expected datacentre workload, that power-efficient switching can be achieved. Specifically, the simulation results indicate that in the proposed flexible optical network with hybrid buffer, the packet loss probability can be reduced to a sufficiently low level, which in this case is of the order of 10^{-9}. Further, the comparison of the minimum power dissipation between the proposed flexible network scheduling and the commonly used fixed scheduling reveals that a significant power reduction, up to 25%, is obtained by the flexible scheme, and it is anticipated that more energy savings can be achieved by performing a full-area searching. In the next chapter, the topic of research is to investigate the feasibility and reliability of large-scale networks composed of a number of the proposed flexible optical switches, with a view to determining the appropriate network topology and control management technique.

Bibliography

[1] S. Kandula, J. Padhye, and P. Bahl, "Flyways to de-congest data center networks," in *ACM HotNets*, Oct. 2009.

[2] K. Barker, "On the feasibility of optical circuit switching for high performance computing systems," in *ACM/IEEE SC 2005 Conference*, pp. 1-22, Nov. 2005.

[3] S. Cheung, T. Su, K. Okamoto, and S. J. B. Yoo, "Ultra-compact silicon photonic 512×512 25-GHz Arrayed Waveguide Grating Router," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 4, pp. 310-316, Aug. 2014.

[4] OMNeT++ discrete event simulator, URL: http://www.omnetpp.org/.

[5] Cisco 10GBASE SFP+ Modules, Datasheet, Cisco Inc., 2010.

[6] Nakarmi Bikash, J. Sangirov, M. Rakib Uddin, Y. H. Won, "Efficient approach towards energy minimized optical network," in *ITU-T Workshop-ICTs: Building the green city of the future*, May 2010.

[7] R. S. Tucker, "The role of optics and electronics in high-capacity routers," *Journal of Lightwave Technology*, vol. 24, no. 12, pp. 4655-4673, Dec. 2006.

[8] R. S. Tucker, "Scalability and energy consumption of optical and electronic packet switching," *Journal of Lightwave Technology*, vol. 29, no. 16, pp. 2410-2421, Aug. 2011.

[9] Christoforos Kachris, Ioannis Tomkos, "Power consumption evaluation of all-optical data center networks,", *Cluster Computing*, pp. 1-13, Aug. 2012.

Chapter 5

Large-scale Optical Packet Switching Datacentre Networks Using Combined Optical Buffering and Packet Retransmission

5.1 Introduction

By exploiting the high-radix, flexible optical packet switch architecture developed previously, this chapter proposes a large-scale modular datacentre architecture where a network of optically transparent switches forms the backbone of the interconnection network. An important feature of the network is the integration of a flexible global resource scheduling algorithm, which dynamically allocates network capacities to dynamically changing traffic patterns, thus significantly enhancing network flexibility and resource utilisation, and alleviating contention conditions. The proposed novel optical packet switch and resource scheduling method enables the design of a highly reconfigurable network infrastructure, which allows dynamically reconfiguring the bandwidth of switch output ports in conjunction with load balancing on input links to accommodate changes in datacentre traffic reconfigurations without over-subscription penalties on heavily loaded links. Furthermore, in order to resolve the contention problem, a hybrid scheme combining recirculating optical Fibre Delay Line (FDL) buffering and a packet retransmission mechanism is employed. The principle is to reduce the packet contentions to a sufficiently low level by using recirculating optical

buffering, such that only a small portion of the contention traffic overflows from the optical buffers and requires packet retransmission. With this arrangement, the optical network under study has a relatively low retransmission rate, as the FDL buffering carries the majority of the contention traffic, thus the proposed hybrid scheme can efficiently manage the traffic conflicts in the network without compromising the latency performance.

Packet retransmission is an important loss recovery mechanism, especially in short-distance local area networks like datacentres and High Performance Computing (HPC) systems [1]. This technique eliminates traffic loss by allowing the lost data to be retransmitted. More precisely, in our proposed scheme, when a packet loss occurs, a Negative Acknowledgement (NAK) control message is generated by the core OPS and then sent back to the source node to request packet retransmission. Once the sender detects this message, it retransmits the optical packet. In this work, two different retransmission strategies are investigated: Random Retransmission (RR) and a novel scheme called Pre-Reservation Retransmission (PRR). The RR policy is a commonly used retransmission strategy which allows the senders to retransmit the lost data as new packets [2]. In this case, to achieve a lossless transmission scheme, it is theoretically possible that a packet needs to be retransmitted an infinite number of times. Conversely, in PRR, an abandoned packet only requires one retransmission, as a future timeslot on the destined channel is reserved for the lost packet when the contention happens, and this timeslot information is stored in the NAK message so as to notify the source node when to resend the packet. Although per-reserving channels can involve quite complex resource scheduling algorithms, in this case reservation is only necessary for a very small proportion of packets and so, overall, the scheme is efficient.

In this chapter, the emphasis is on the performance analysis and network dimensioning of the proposed optical packet switching network. Extensive simulation experiments have been carried out to investigate the performance of the proposed optical network architecture and its flexible traffic scheduling strategy under a worst-case (in terms of connection degree) all-to-all traffic pattern and a more realistic non-uniform datacentre load (hotspot traffic) and we show how performance is relatively immune to large changes in workload characteristics. Additionally, the behaviour of the recirculating FDL buffering, the packet retransmission mechanism and a coordinated congestion management technique, combining the recirculating FDLs and the packet retransmission scheme, have also been explored in detail in the proposed flexible optical network. The simulation results reveal the benefits of the combined recirculating FDL/retransmission scheme in resolving contentions. The remainder of this chapter is organised as follows. Section 5.2 introduces the envisioned net-

work architecture and technologies of the proposed datacentre interconnection network, followed by a detailed description of the global traffic scheduling procedure in Section 5.3. In Section 5.4, the packet retransmission strategies are explained. Section 5.5 presents a detailed performance analysis of the proposed network through simulations, highlighting its benefits regarding network throughput, scalability and flexibility. Finally, Section 5.7 outlines the main conclusions.

5.2 Network Architecture

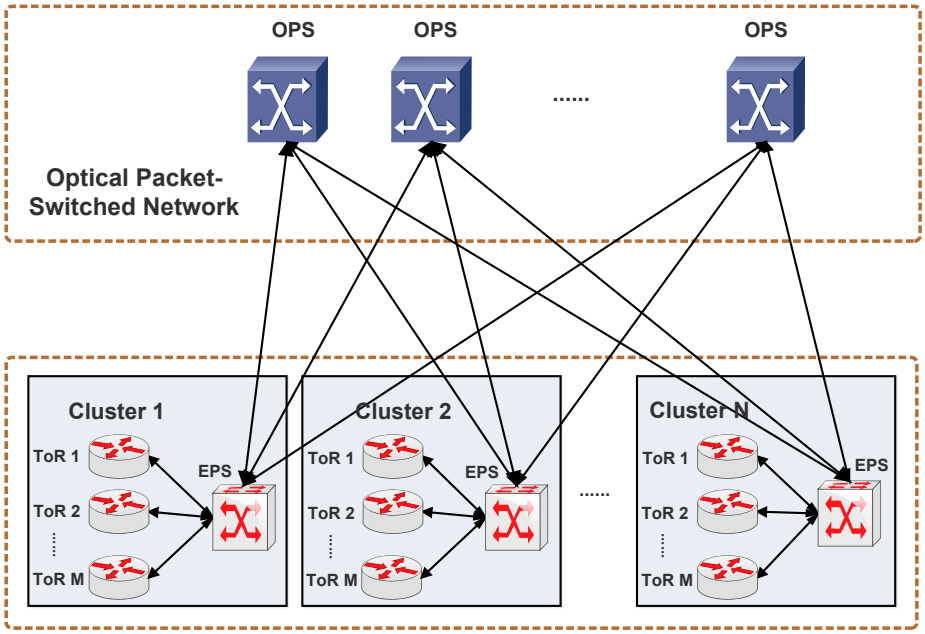


Figure 5.1: The proposed modular optical packet-switched datacentre network (DCN).

The proposed optical datacentre network employs a hierarchical modular structure, as shown in Figure 5.1. In the multi-layer network, the servers are arranged into racks and connected to Top-of-Rack (ToR) switches which manage the inter-server communications inside racks. Several ToR switches are fully interconnected by a low-radix, high-capacity cluster electronic packet switch (EPS), forming together a small-size cluster network. This modular design not only alleviates the capacity limitations on the aggregation switches, but also makes the network scale out easily. The cluster switch, on one hand, handles the communications among ToR switches, and on the other hand, provides inter-cluster connectivity. Traffic from ToR switches is aggregated at cluster switches and transported over an all-optical core switching network, which is designed based on the optical packet switching (OPS) paradigm. Large port-count optical packet switches (OPSs)

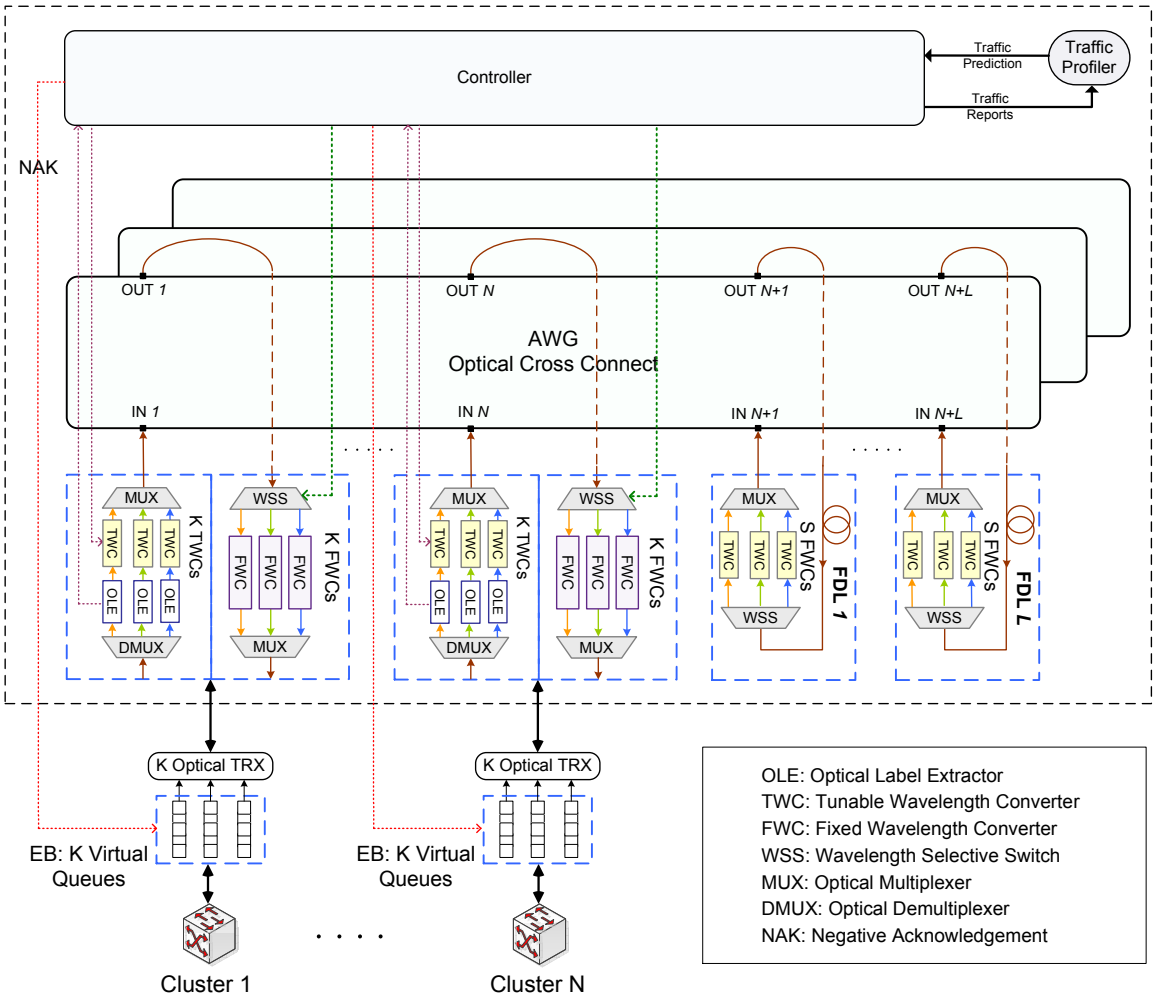


Figure 5.2: The proposed AWG-based optical packet switching network.

with high switching speed make up the backbone of the interconnection network, which can efficiently support the massive traffic demands between clusters by exploiting the enormous bandwidth capacity provided by optical technology. Assuming that each cluster includes 16 ToR switches, and OPSs have a port count of 128×128, and with each rack having 48 servers, a large-scale datacentre network totaling 98,304 servers is then obtained.

As shown in Figure 5.1, the upper layer of the proposed datacentre network is composed of a collection of optical packet switches. The envisioned node architecture and technologies for all-optical packet switches are described as follows. An overview of the proposed optical packet switch is shown in Figure 5.2. The switch connects N input fibres to N output fibres, each fibre carrying K wavelength channels. The OPS node consists of N linecards, an all-optical AWG switching fabric, and a pool of shared optical buffers. The linecard includes three main functional modules, namely,

Optical Label Extractors (OLEs), input Tunable Wavelength Converters (TWCs) and an output port optical Wavelength Selective Switch (WSS) with Fixed Wavelength Converters (FWCs). Each input channel is equipped with a dedicated OLE and TWC. The OLE separates the optical label from the user data. The label is then sent to the controller while the data packet is forwarded to the switching plane. The controller decodes and reads the label information, and then accordingly configures the input TWC. After traversing through a TWC for wavelength shifting, the data packet enters the core switching fabric and is directed to the destined output port. The core switching fabric is an Arrayed Waveguide Grating (AWG) router, which allows multiple different wavelengths from multiple inputs to be switched to an output simultaneously. However, the AWG is not a non-blocking wavelength router, as it can only establish one lightpath between an input-output pair. That is, if multiple channels from the same input request the same output, the input channels would be tuned to the same wavelength and so contention occurs. This contention is referred to as *input contention*. After the AWG, the multiplexed optical channel from the switch output port is directed to the WSS in the linecard, which partitions the WDM channel into K sets of wavebands, where each can be configured to have an arbitrary number of wavelengths. Subsequently, the current wavelength channel on each of the K wavebands is converted to a fixed transmission wavelength by a dedicated Fixed Wavelength Converter (FWC). This can potentially cause congestion between wavelengths within the same waveband (referred to as *output contention*), as only one wavelength can be converted by the FWC and transmitted on the outgoing channel at any one time.

In the proposed OPS-based modular network, a source-destination cluster pair is interconnected by multiple intermediate optical packet switches (OPS) to overcome the limitations in link capacity and switch port density. The performance of the datacentre network crucially depends on the flow scheduling strategy which adaptively performs traffic distribution among all the core OPS switches, based on regular updates of current traffic demands in the network. This scheduling strategy exploits the path diversity provided by topological redundancy, and thus optimises the use of the scarce OPS resources. This access scheduling algorithm is integrated as part of the global routing and scheduling procedure which is detailed in Section 5.3.

Based on the access scheduling decisions, the inter-cluster traffic is routed to the appropriate OPS for transmission. A prominent feature of the proposed OPS architecture is that flexible bandwidth partitioning at output ports is made possible by the WSS, where the incoming wavelength channels can be arbitrarily divided and converted to the transmission wavelengths by FWCs in accordance with expected source-destination traffic flow volumes between switch ports. This flexibility pro-

vides a finer match between expected traffic demands and provisioned bandwidths, and thus significantly enhances resource utilisation under non-uniform and dynamic traffic conditions. In this context, the bandwidth partition optimisation problem is addressed as part of the central routing and scheduling procedure.

Given the increased efficiency provided by an adaptive global routing mechanism, the dimensions of the main optical components in the OPS architecture can be reduced, avoiding over-dimensioning of an architecture with fixed bandwidth partitioning.

5.2.1 Overview of Contention Resolution Methods

There are two essential issues to be addressed in the proposed OPS network: (i) input contention and (ii) output contention. To avoid packet loss due to collisions, different congestion control techniques are adopted in the network.

As input collisions only happen among input channels coming from the same cluster, this type of contention can be eliminated by deploying electronic buffering in the cluster electronic switches, which has the advantage (over FDL buffering) of storing the contended packets for an arbitrary amount of time. Each cluster switch maintains K virtual queues, one for each transmission wavelength. That is, every virtual queue is fed into a dedicated optical transmitter which serves as an interface between the electronic cluster switch and the optical switching network. The queue management details and the queueing delay analysis will be included in the simulations.

To address output contention, the switch deploys different length optical Fibre Delay Line (FDL) buffers. The optical buffers are configured in a feedback manner with each of L FDLs having S TWCs in its return port. Note that FDL recirculation is allowed in the switch, that is, a contending packet is allowed to traverse through optical buffers and OPS multiple times until its destination switch output becomes free. To avoid excessive delay and optical signal degradation issues, a limit is usually placed on the number of retransmission attempts. It is important to point out that in the proposed optical network, a PreRes scheduling strategy is employed, that is, the scheduling/reservation of a wavelength channel at the switch output port or an optical buffering channel in the FDL is made prior to the optical packet entering the optical switch. It can be of great benefit, as the abandoned optical packets only occupy the transmission channels at switch inputs, without consuming any network resources at switch output ports and any FDL buffering resources.

Fibre delay lines may carry multiple channels, but only support a discrete set of delays and limited

buffering capacity, thus packet contention still exists in the optical network and cannot be eliminated, which consequently leads to traffic loss. This is one of the main drawbacks of FDL buffering. To recover the lost data, novel packet retransmission policies are incorporated into the network so that, overall, a non-blocking optical packet-switched network is realised. Specifically, if a packet contention in an OPS cannot be resolved by the recirculating FDLs, the packet will necessarily be discarded, and then a Negative Acknowledgement (NAK) message is generated by the core OPS and sent back to the source cluster electronic switch. Once receiving the congestion notification NAK, the source switch retransmits the data traffic identified in the NAK message. Otherwise, after a given timeout, the packet is removed from the retransmission buffer in the source cluster switch, and assumed to have successfully reached its destination. The timeout is estimated based on the Round-Trip Time (RTT) which represents the propagation delay between a source/destination cluster pair [3]. Note once again that the scheduling/reservation of the network resources follows the PreRes strategy, where channel scheduling is performed prior to the arrival of the optical packets, hence the unsuccessful transmissions in packet retransmission only consume input wavelength channels and do not occupy any output wavelengths.

As mentioned above, to enable packet retransmission, retransmission buffers are deployed in the cluster switches so as to store the copies of the packets for future possible retransmission. An important problem associated with this is the requirement for large retransmission buffers, especially under high traffic load conditions where frequent packet retransmissions happen. However, in practical implementations, the buffer size is necessarily limited, and this could potentially give rise to traffic loss, as the traffic may overflow from the limited retransmission buffer. To avoid packet loss due to traffic overflowing, FDL buffering is efficiently dimensioned to carry the majority of the contention traffic such that only a small fraction of traffic needs retransmission, and thus the retransmission buffer size requirements are controlled.

5.3 Flexible Network Load Scheduling Algorithm

Given regular updates of current traffic demands in the datacentre network, the controller adaptively performs network-wide routing and scheduling. The purpose of the global load-adaptive scheduling is to improve resource utilisation and network throughput by optimising the load balancing with dynamic re-configuration. It consists of two phases: an access control scheduling phase and an OPS resource allocation phase, as demonstrated in Figure 5.3. Both are realised by utilising a

least-loading algorithm.

The least-loading algorithm is a simple but efficient heuristic algorithm. It realises an overall balanced traffic by using mean traffic load as heuristic information, with a view to reducing congestion conditions in the network. Given a set of traffic loads, the algorithm firstly arranges the traffic loads in descending order. Then, one by one and starting with the largest, each of the remaining loads are assigned to the least-load group. The heuristic least-loading algorithm provides a fast solution for the load-balancing optimisation problem.

Before detailing the global scheduling algorithm, some important notations are introduced. Consider a datacentre network with P optical packet switches (OPSs). Each OPS has N input/output ports, each connecting to an optical input/output fibre carrying K incoming/outgoing transmission channels. Every input/output fibre directly attaches to one of N clusters, which interconnects M ToR switches. Thus, in total, the network includes MN ToR switches. The overall traffic demand in the network, which is represented by a block matrix $\mathbf{D}$ of size $MN \times MN$, can be expressed as,

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{1,1} & \mathbf{D}_{1,2} & \dots & \mathbf{D}_{1,N} \\ \mathbf{D}_{2,1} & \mathbf{D}_{2,2} & \dots & \mathbf{D}_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_{N,1} & \mathbf{D}_{N,2} & \dots & \mathbf{D}_{N,N} \end{bmatrix} \tag{5.1}$$

where $M \times M$ matrix $\mathbf{D}_{i,j}$ contains the traffic loads from the M ToR switches in Cluster i to the M ToR switches in Cluster j, e.g. $\mathbf{D}_{i,j}(p, q)$ represents the traffic load sent from ToR p residing in Cluster i to ToR q residing in Cluster j, where $1 \leq i \leq N, 1 \leq j \leq N$. When $i = j$, $\mathbf{D}_{ij}$ represents the intra-cluster traffic matrix. Otherwise, $\mathbf{D}_{i,j}$ represents the inter-cluster traffic matrix. In the traffic matrix $\mathbf{D}$, each individual element defines the traffic demand between a source/destination ToR pair.

The access scheduling phase performs the assignment of traffic flows across a bank of OPSs operating in parallel, see Figure 5.3. As mentioned above, each element in the traffic matrix $\mathbf{D}$ corresponds to a traffic flow between a source/destination ToR pair, thereby the forwarding and traffic splitting decisions are made on a per-flow basis, which also makes sure that the packet arriving order is maintained. In this phase, the objective is to select the most suitable OPS for the traffic flows to ensure the resource utilisation and transmission efficiency. To do this, the algorithm uniformly distributes the M^2N traffic loads originating from each source cluster independently into P groups by using the least-loading procedure. More specifically, for a source cluster i, M^2N traffic loads, which

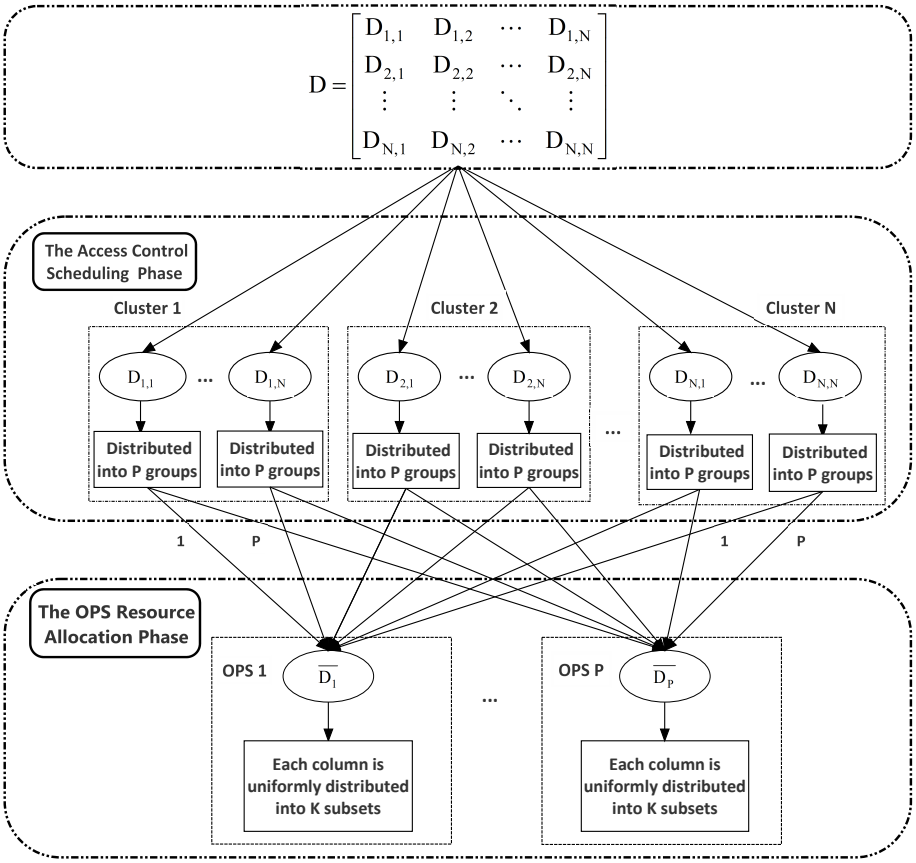


Figure 5.3: The flexible network scheduling algorithm.

are accommodated in block matrices $\mathbf{D}_{i,j}$ $(j = 1, \ldots, N)$, are transferred from this cluster to all other clusters, and by using the least-loading algorithm, these traffic loads are evenly divided into P groups. In this way, the traffic matrix $\mathbf{D}_{ij}$ is evenly divided into P constituents $\mathbf{D}_{ij,k}$ $(1 \leq k \leq P)$ such that $\mathbf{D}_{ij} = \sum_{k=1}^{P} \mathbf{D}_{ij,k}$. Consequently, the overall traffic matrix $\mathbf{D}$ is uniformly divided into P matrices of size $MN \times MN$, i.e. $\overline{\mathbf{D}}_1, \overline{\mathbf{D}}_2 \ldots \overline{\mathbf{D}}_P$, and $\mathbf{D} = \overline{\mathbf{D}}_1 + \overline{\mathbf{D}}_2 + \ldots + \overline{\mathbf{D}}_P$. The matrix $\overline{\mathbf{D}}_i$ represents the traffic loads assigned to the i-th OPS. An important aspect of the algorithm is that the traffic distribution procedure of the traffic matrices $\mathbf{D}_{ij}$ can be performed independently, which significantly reduces the computation time and allows for a fast scheduling.

The access scheduling strategy resolves the traffic flow distribution optimisation problem and decides the amount of traffic demands, $\overline{\mathbf{D}}_i$, routed to i-th OPS. This is then followed by the OPS resource allocation procedure, which performs local load balancing among K outgoing links at each output of the optical packet switches by employing the least-loading algorithm, based on the traffic demands $\overline{\mathbf{D}}_i$. More precisely, as illustrated in Figure 5.3, the traffic flows requested for the same output port in traffic matrix $\overline{\mathbf{D}}_i$ are uniformly divided into K subsets by employing the least-loading

algorithm, and each subset is assigned to an outgoing channel. This flexible resource allocation process optimally configures which input wavelengths (input port traffic streams) are directed to which output outgoing channel/FWC, in each output port of the OPS, partitioning incoming wavelengths to an output into groups by appropriate configuration of the WSS, with each group assigned to one of K outgoing channels/FWCs in the output port, with the objective of balancing loads across the K outgoing links. Accordingly, this procedure off-loads the traffic flows from a highly utilised link to an underutilised outgoing channel, thus improving resource utilisation and alleviating link congestion.

The proposed network global scheduling strategy balances the traffic loads and best exploits the OPS resources. Additionally, the benefit of the proposed algorithm with respect to control complexity is also evident. As noted before, in the access scheduling phase, the traffic distribution can be processed in parallel by independent local controllers. Furthermore, the OPS resource allocation procedure in the optical switches is performed on a per-output-port basis. Thus, the controller complexity and scheduling time is drastically reduced, provided that parallel computation is implemented, enabling fast switching in the presented network. In Section 5.5.2, the proposed flexible network scheduling algorithm is applied to the presented optical network, and the system performance is examined in detail under both uniform and non-uniform traffic patterns.

5.4 Retransmission Mechanism

As alluded to previously, in the proposed optical network (see Figure 5.2), when an ingress cluster node sends a data packet to an egress cluster node, a copy of the data packet is stored in the electronic retransmission buffer in the ingress switch for possible future packet retransmission. When the optical packet arrives at the core OPS, the packet scheduler in the controller attempts to schedule the arriving optical packet on the designated transmission channel by using the PreRes strategy. If the scheduling/reservation fails, the packet is dropped immediately and considered lost, and subsequently a control message called Negative Acknowledgement (NAK) is sent by the OPS controller back to the source cluster switch as a request to retransmit the packet. Once the source node receives the NAK message, it retransmits the data packet. If no NAK has been received from the core OPS after a given timeout, the copy of the packet is removed from the retransmission buffer, implying a successful data transfer. Different policies for packet retransmission have been proposed. In this work, two retransmission strategies are considered, which are explained as follows.

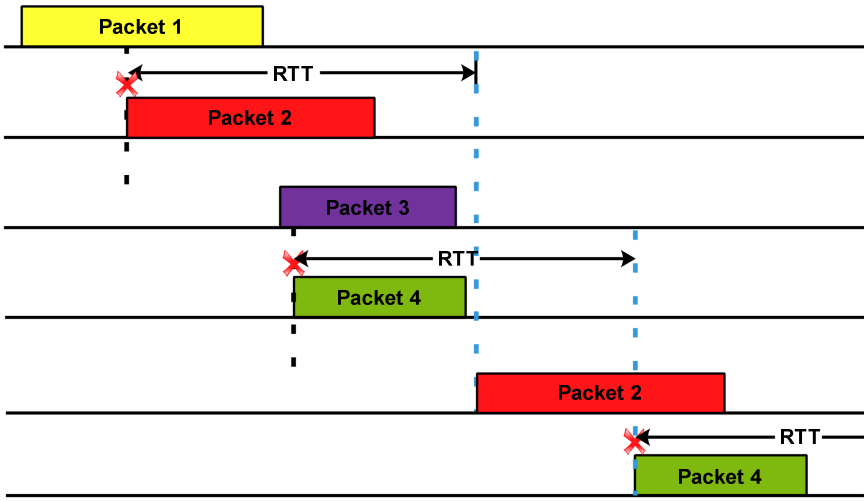


Figure 5.4: Random Retransmission (RR)

- Random Retransmission (RR). Random Retransmission is the simplest retransmission strategy. Under RR, the source ingress node retransmits the lost traffic as new packets [2]. However, this potentially leads to the situation where the retransmitted optical packet is blocked again in the core OPS, and requires another retransmission attempt, as demonstrated in Figure 5.4. Figure 5.4 shows that a contention occurs between packet 1 and packet 2, then packet 2 is discarded immediately without occupying any wavelength resources, which allows for the successful scheduling of packet 3, and subsequently packet 2 is retransmitted as a new packet. In the RR scheme, the number of retransmissions experienced by a packet is random, and it is theoretically possible that a packet may need to be retransmitted an infinite number of times. This is problematic, as multiple retransmissions result in greatly increased communication delay, low resource utilisation and potential overflow of a limited transmission buffer, especially under heavy loads. For that reason, limits are placed on the number of allowable retransmissions.

- Pre-Reservation Retransmission (PRR). We propose a novel scheme where the NAK message is utilised not only to notify the source node of the packet retransmission, but also to carry retransmission timing information. When a packet is discarded at the OPS, the OPS scheduler searches its scheduling time-line and pre-reserves a timeslot on the designated output channel in advance of the future arrival of the retransmitted packet, as illustrated in Figure 5.5. This scheduling timing, is stored in the NAK message sent back to the source cluster switch. The time window between the current time and the future arrival time is set as the sum of the path RTT and an additional waiting time. The additional time is determined by two factors: (i) the

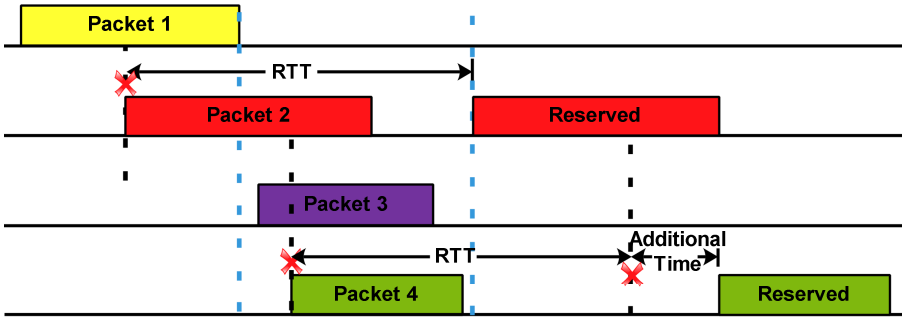


Figure 5.5: Pre-Reservation Retransmission (PRR)

time of availability of the designated OPS output channel, as shown in Figure 5.5, and (ii) the availability of the input transmission channels to the OPS (c.f. *input contention*). More specifically, as each incoming fibre of the OPS comprises only K transmission channels, if there are more than K optical packets for retransmission, retransmissions will compete for access to the channels. Hence, a packet may need to be stored in the buffer for an additional time before it is retransmitted. The main benefit of the PRR scheme is that it guarantees successful transmission with at most one retransmission, though requires a more complex scheduling scheme.

In both RR and PRR schemes, under heavy load conditions, frequent packet retransmissions would occur, which greatly increase the network load, consume significantly transmission bandwidth, require large retransmission buffers and induce additional delays. To avoid frequent retransmissions and mitigate these potential problems, the retransmission strategies have been used in conjunction with optical buffering, with optical buffers dimensioned to handle the majority of the contention traffic. In what follows, in order to study the impact of the recirculating optical buffering on contention and latency performance, the performance of the optical network with recirculating FDLs is first investigated for different numbers of FDL buffering channels and different numbers of allowable FDL circulations. Then, the advantages and limitations of the two packet retransmission schemes, when employed in the proposed optical network, are extensively exploited. A detailed performance analysis is conducted for these two schemes so as to examine their efficacy in recovering data loss. Furthermore, the system performance of the proposed datacentre network with the combined optical buffering/packet retransmission scheme is evaluated and analysed in terms of the packet loss, the overall communication latency and the packet retransmission rate under both uniform and non-uniform traffic patterns.

5.5 System Performance Evaluation

5.5.1 Simulation Setup

5.5.1.1 General Parameter Settings

The network is simulated using the event-driven simulator OMNeT++. Table 5.1 tabulates some important network parameters utilised in simulations. In the model, the number of ToR switches in a cluster, M, is set to be 16 and the number of clusters, N, which is also the OPS switch size, is fixed to be 128, thereby the resultant datacentre network provides connectivity to 2048 ToR switches via 10 or 40 Gbps links. Each ToR switch is modelled as a traffic source/sink, which generates and sends traffic flows to other ToR switches. Note that the traffic generation process follows an Engset traffic model where the packet transmission times are exponentially distributed and the idle times between packet transmission periods are also exponentially distributed. The number of OPSs in the optical network is represented by P, and each input/output fibre of the OPS carries K transmission channels. This implies that each input port of the OPS is equipped with K TWCs, and each output port has K FWCs. The number of FDLs is denoted L, and each FDL has S TWCs. The normalised base delay of FDLs, which is computed as the ratio between the FDL base delay and the mean packet transmission time, is taken as 0.1, based on the FDL base delay analysis in Chapter 2. The maximum number of allowable FDL circulations is denoted by R. In packet retransmission, the maximum number of allowable packet retransmissions is defined by B. If $B = 0$, no packet retransmission is allowed in the network, whereas if $B = \infty$, the number of allowable packet retransmissions is unlimited.

The amount of (normalised) traffic transferred from a rack (ToR) to all other racks (ToRs), which is referred to as inter-rack traffic, is represented by a. Particularly, if $a = 0$, there is no inter-rack traffic, whereas when $a = 1$, all traffic generated by a ToR switch is destined for other ToR switches (either in the same local cluster or other remote cluster). According to previous datacenter traffic studies [4], a is typically observed to be around 0.2 (20%), but in this section the performance analysis is performed for inter-rack traffic loads far beyond 0.2, to test how the optical network will scale under much heavier load. Given a, the total traffic load generated inside a cluster is $a \times M$, which can be classified into two types of traffic: intra-cluster traffic and inter-cluster traffic. The intra-cluster traffic is exchanged between the ToR switches within the same cluster, thereby it is handled by the cluster electronic switches without travelling through the core optical network.

Table 5.1: Table of Notation

M	Number of ToR switches in a cluster.
N	OPS input/output port count. It also defines the number of clusters in the network
P	Number of Optical Packet Switches (OPSs) in the network
K	Number of transmission channels per incoming/outgoing fibre of the OPS
L	Number of Fibre Delay Lines (FDLs)
S	Number of channels (TWCs) per FDL
R	Maximum number of allowable FDL circulations
B	Maximum number of allowable packet retransmissions
a	Inter-rack traffic load, the (normalised) load sent out from a ToR, that is destined for other ToRs (in the same or in different clusters)
β	Average proportion of the traffic load (a) generated by each of the 2048 ToR switches that needs to traverse through the optical network. In the uniform traffic pattern, $\beta \approx 1$, whereas in the generated non-uniform traffic pattern, $\beta = 0.835$
ρ	Optical network load, the proportion of load offered to the optical network, relative to its full load capacity

Conversely, the inter-cluster traffic, which represents the data traffic exchanged between two ToR switches residing in different clusters, needs to travel through the core optical network. The ratio between the inter-cluster traffic load and the total cluster traffic load ($a \times M$) is expressed as β. This ratio defines the average proportion of the traffic load (a) generated by each of the 2048 ToR switches that needs to traverse the proposed optical network. Note that β is determined by the particular traffic matrix used. In the case of a uniform traffic pattern where the traffic load from a ToR switch is uniformly distributed to all ToR switches, β is computed as, $\beta = \frac{MN-M}{MN-1} \approx 1$, that is, almost all traffic load generated by each of the ToR switches needs to traverse the optical network. In the case of our chosen hot-spot traffic matrix, β is still maintained at a high value, which is calculated as 0.835.

Based on β, the total traffic load sent from a cluster to the optical network is computed as $\beta \times a \times M$. Since the proposed network is composed of N (= 128) clusters, the total requested optical network capacity is $\beta \times a \times M \times N$. As the total optical network capacity is $P \times K \times N$, the offered network load on the core optical network, denoted by ρ, is defined as $\rho = \frac{\beta \times a \times M \times N}{P \times K \times N} = \frac{\beta a M}{PK}$, which is the ratio of the requested optical network capacity to the total optical network capacity. Thus, ρ is equivalent to the average traffic load on each transmission link of the OPS. We remark that as each input/output port of the OPS carries K transmission channels, the offered traffic load per switch port is ρK. Note that $\frac{M}{PK}$ determines the over-subscription ratio of the core optical network. For the given dimensions of M, P and K, the network is oversubscribed by the ratio $4:3$ in the uniform traffic pattern ($\beta = 1$). For non-uniform traffic (given the lower value of β) the over-subscription is

approximately $4\beta : 3$. In either case, the optical network understudy is oversubscribed, so that by scaling a, its performance can be fully stress-tested in the simulation studies that follow.

5.5.1.2 Performance Measurements

The packet loss rate, also referred to as the blocking probability, is calculated as the ratio between the lost traffic from the optical network and the total traffic offered to the optical network. The input queueing (buffering) delay represents the average amount of time a packet needs to be stored in the cluster switch before the switch can forward it to the optical network without input contention. The overall communication latency is defined as the average end-to-end delay experienced by the optical packets. The resulting network workload is computed as the sum of the offered network load ρ and the retransmission traffic load, this also being referred to as the effective network load. The packet retransmission rate is measured as the ratio of the total number of retransmission requests to the total number of packets offered to the optical network. The network throughput is quantified as the amount of traffic carried by the optical network. Additionally, the retransmission buffer size is calculated as the queue length required in the cluster electronic switch in order to store the copies of the packets for possible retransmission.

5.5.1.3 Traffic Patterns

In the datacentre interconnection network, an important requirement is to support heterogeneous traffic profiles of datacentre applications with a specified Quality-of-Service (QoS). For that reason, the network performance of the proposed datacentre interconnect is studied under two types of traffic patterns: all-to-all, worst-case communication pattern and a more representative heterogeneous hot-spot traffic pattern. The traffic demands in the network are described by ToR-to-ToR traffic matrix (TM), as expressed in Equation 5.1. Note that the traffic matrix has zero entries on its diagonal, that is, all block matrices $\mathbf{D}_{i,i}$ are zero matrices, due to the fact that the intra-cluster traffic is handled by the cluster electronic packet switches (EPSs) and thus will not traverse through the optical network. In the all-to-all traffic pattern, each ToR switch sends traffic to all remote ToR switches with uniform probability, resulting in a large number of concurrent flows in the network, which is worst-case in terms of the required network connectivity. For hot-spot datacentre communication patterns, due to the nature of datacentre applications, the traffic demands evolve rapidly and constantly, thus it is challenging to accurately estimate the volumes of traffic flows. Some re-

cent studies [4–8] have explored the traffic characteristics of datacentre applications and services. As introduced in Section 1.1.2, servers that communicate often with each other are arranged in the same rack, and only 0.5% of server pairs belonging to different racks exchange data [5]. A server generates about 10 concurrent flows more than 50% of the time, and at least 5% of the time, it generates more than 80 flows [6, 7], so the traffic matrix is a sparse matrix with hotspots, as most racks communicate with a small subset of remote racks at any time [8]. Importantly, [4] reveals that the amount of traffic destined outside the rack is less than 25% of the total rack traffic. Assume that each rack includes 48 servers, thus the proposed network consists of $98,304$ servers. Following the communication pattern described in Table 5.2, a non-uniform traffic matrix of size 2048×2048 is generated, which is very similar to the traffic matrix in [5]. This allows for the performance analysis of the proposed datacentre network under non-uniform, hot-spot traffic. It should be noted that in the non-uniform traffic matrix, the ratio β equals to 0.835, which means that on average 83.5% of the traffic load generated by each of the 2048 ToR switches will travel through the optical network.

Table 5.2: Traffic generation pattern

0.5% server pairs	non-zero traffic load
Intra-rack traffic	traffic load equals zero
50% of ToRs	generate $0-10$ concurrent traffic flows, one to a ToR switch
45% of ToRs	generate $10-80$ concurrent flows, one to a ToR switch
5% of ToRs	generate $80-2048$ current traffic flows, one to a ToR switch
50% of ToRs	receive $0-10$ concurrent traffic flows, one from a ToR switch
45% of ToRs	receive $10-80$ concurrent traffic flows, one from a ToR switch
5% of ToRs	receive $80-2048$ concurrent traffic flows, one from a ToR switch
Normalised traffic load generated by an active ToR	1
Normalised traffic load received by a ToR	1

In simulation experiments we will produce results for different offered network loads ρ. It is important to note that at very high loads, some of the traffic demands in hotspots could be larger than 1, even though the average offered network load ρ is less than 1. Thus, in order to devise a hotspot traffic matrix that would not unfairly overload switch input/output ports, compared to uniform traffic, it is necessary to cap individual traffic loads as the average network load is scaled-up in experiments. Specifically, when a traffic load in the traffic matrix reaches 0.8, it is then fixed to be 0.8. Clearly, the resulting traffic matrix is indeed a truncated version of the original hot-spot traffic matrix for high load values. Figure 5.6 shows the relationship between the original mean traffic

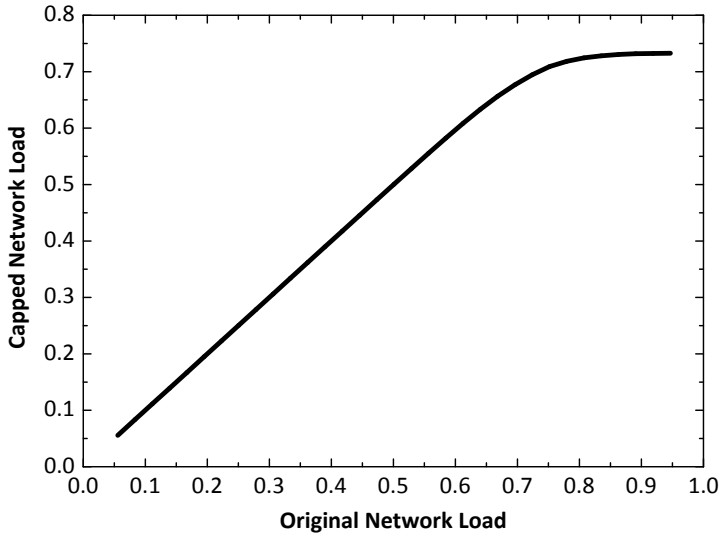


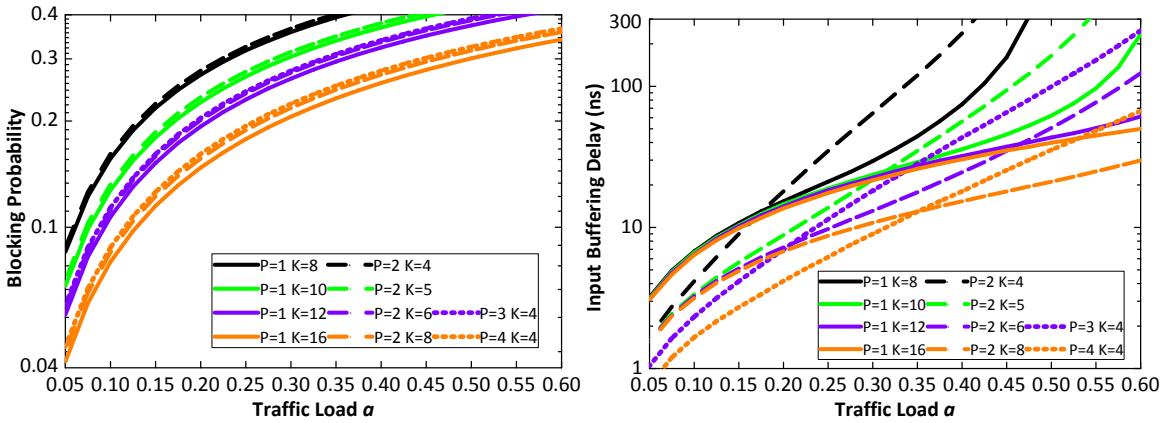
Figure 5.6: The relationship between the original traffic load and the capped traffic load in hot-spot traffic pattern.

load and the modified, scaled mean traffic load. Only at very high loads, the network load is capped and thus reduced. This scaling method ensures a fair evaluation of the proposed optical network while maintaining the essential characteristics of the hot-spot traffic pattern.

Under the above mentioned premises, a detailed performance analysis of the proposed flexible optical network architecture is carried out under both uniform and hot-spot traffic patterns in the following sections. Note that under hot-spot communication pattern, the offered network load ρ is indeed referred to as the capped network load.

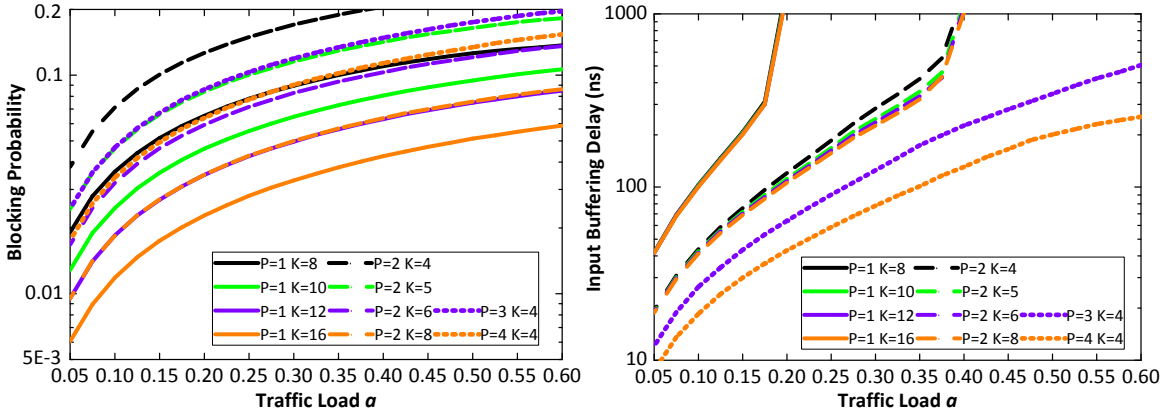
5.5.2 Flexible Network Load Scheduling Analysis

In order to determine baseline performance, we initially assume there are no optical buffers and no packet retransmission in the network, meaning that no contention management techniques are applied to resolve output contention. The network performance, measured in terms of the packet loss rate and the average input buffering delay, is evaluated under worst-case uniform traffic pattern and hotspot non-uniform traffic pattern for different inter-rack traffic loads a. There are two major objectives in this section. One is to examine the benefits of the proposed flexible network scheduling technique in the developed optical network, by comparing the performance of the flexible scheduling with that of fixed resource scheduling. Note that in the fixed scheduling, the resource scheduling result is obtained based on the uniform traffic matrix and will not change with the traffic demands in the network. For that reason, under uniform traffic pattern, the fixed and flexible routing algorithms



(a) Blocking probability (b) Input buffering latency

Figure 5.7: Performance evaluation of the flexible resource scheduling under uniform traffic pattern.



(a) Blocking probability (b) Input buffering latency

Figure 5.8: Performance evaluation of the flexible resource scheduling under hot-spot traffic pattern.

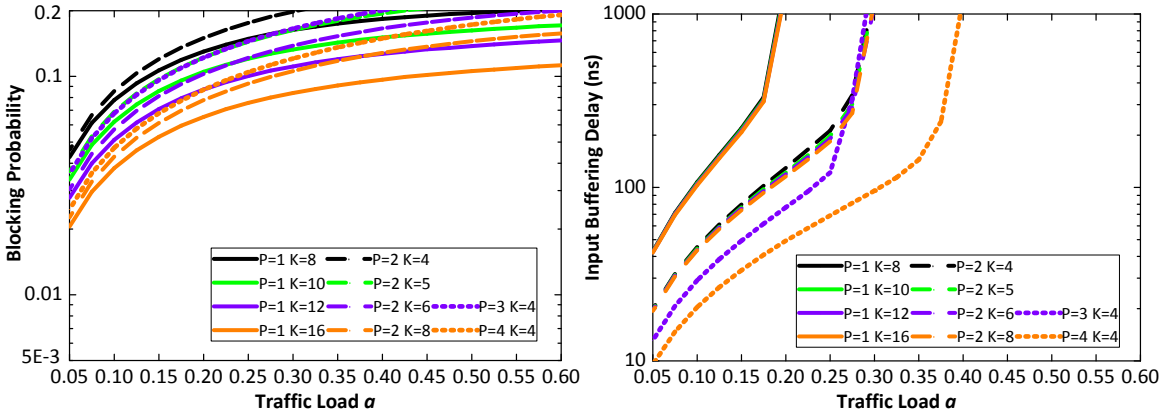
achieve the same network performance. Thus, the performance comparison between the fixed and flexible routing algorithms is only performed for the non-uniform traffic pattern. The other aim is to decide the proper values for the number of OPSs in the optical network, P, and the number of transmission channels per input/output port of the OPS, K, by illustrating how packet loss rate and input buffering delay scale with $P \times K$ under a wide range of inter-rack traffic loads a.

Figures 5.7 and 5.8 plot the performance measurements of the proposed optical network with flexible global resource scheduling as a function of the inter-rack traffic load a under, respectively, all-to-all communication pattern and non-uniform hotspot traffic pattern. For the purpose of comparison, Figure 5.9 depicts the system performance of the optical network with fixed resource scheduling under hot-spot traffic pattern. From Figures 5.7 and 5.8, it can be seen that in all scenarios, with the increase of the inter-rack traffic load a, the network performance degrades significantly. This can

be explained by the fact that as the amount of inter-rack communication increases, the volume of datacentre traffic that needs to traverse through the optical network grows progressively, and as a consequence, more traffic competes for the network resources, resulting in more packet contentions and larger buffering delay. It is further shown that the observations in both figures exhibit the same trend, that is, with the increase of $P \times K$, the packet loss rate and the input buffering delay decrease substantially. This is expected as increasing the effective network switching capacity always yields an improvement in network performance. Notably, in Figure 5.7(a), for a set of implementations where the product of P and K are the same, the packet loss rates have very similar values, due to the uniform distribution feature of the all-to-all communication pattern. In contrast, Figure 5.8 shows that under non-uniform traffic, when $P \times K$ is the same, the higher is the value of P, the higher is the packet loss rate, but the lower is the input buffering delay. This is because the proposed access scheduling alleviates the input congestion by spreading the traffic among P input fibres of P OPSs, but it potentially imbalances the traffic loads among output fibres and consequently increases the output congestion. Thus there exists a trade-off between the packet loss rate and the input buffering delay. Furthermore, the latency results show that, for a certain value of P, as K increases, the delay first reduces drastically. This is due to the fact that a larger number of transmission channels give less input congestion, and thus the waiting time in the queue is reduced. But beyond a certain value of K, the input buffering delay is sufficiently low and the latency improvement is less significant. Based on these measurements, it can be concluded that a compromise needs to be made between the packet loss rate and the latency performance when dimensioning the optical network.

As previously stated, in datacentre applications, only a small fraction of the rack traffic, less than 25%, is destined towards remote racks, thereby the network configuration with $P \times K = 12$ is chosen, leading to a ratio of $4 : 3$ between the downlink capacity and the uplink capacity in the cluster switch. Consider the above mentioned trade-off between packet loss and input queueing delay, P is dimensioned to be 3 and K to be 4. The experiments in the following sections will be performed in this network implementation. Given the configuration settings, the relationship between the offered network load ρ and the inter-rack traffic load a can be quantified. As explained in Section 5.5.1.1, in the uniform traffic pattern, ρ is calculated as $\rho = \frac{aM}{PK} = \frac{4}{3}a \approx 1.33a$, whereas in the non-uniform traffic pattern, $\rho = \frac{\beta aM}{PK} = \frac{0.835*16}{12}a \approx 1.11a$. Apparently, the network is under-dimensioned, that is to say, if $a = 1$, the offered traffic load to the optical network is significantly higher than 1.

More importantly, the results in Figures 5.8 and 5.9 indicate that the proposed adaptive global scheduling algorithm greatly enhances optical network utilisation, packet loss performance and



(a) Blocking probability (b) Input buffering latency

Figure 5.9: Performance evaluation for the fixed scheduling under hot-spot traffic pattern.

communication latency, compared to commonly used fixed scheduling. The figures show that using the flexible scheduling method, the packet loss rate varies from 0.006 to 0.2, whereas for the fixed scheduling, the loss rate ranges within the interval $[0.02, 0.2]$, thus the traffic is less likely to overflow from the optical network in flexible scheduling. Note also that the overall network throughput is $MNa\,(=2048a)$, so even a small reduction in loss rate results in a considerable decrease in the amount of blocked traffic. The reason behind the loss improvement is that the access scheduling phase balances the traffic demands among P OPSs and then the OPS resource allocation procedure further balances the traffic loads among different outgoing links at each output port, which significantly alleviates the congestion conditions in the network and enhances resource utilisation, as discussed in Section 5.3. Further, it is observed that using the flexible scheduling method, the input buffering delay is greatly reduced from $>1\mu$s to less than 300ns, as the adaptive access scheduling enables load-balancing and thus alleviates the input congestion, which in turn reduces the required queueing time in the buffer queue. The performance improvements highlight the fact that the proposed flexible scheduling algorithm significantly outperforms the fixed scheduling scheme, because it allocates the network resources to the transmission requests based on the traffic demands.

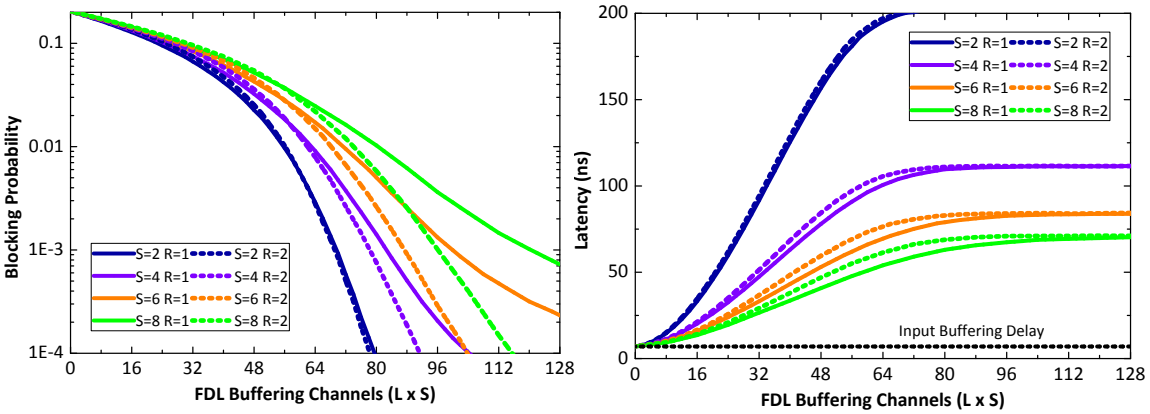
Figures 5.7 and 5.8 indicate that a significant fraction of the total traffic overflows from the network, even when the switching capacity is over-provisioned. This is the main inherent drawback of optical packet switching. In order to mitigate the contention problem in the proposed optical packet-switched datacentre network, different congestion control techniques are investigated, including optical buffering, packet retransmission and hybrid FDL/retransmission scheme. In what follows, extensive experiments are carried out to study the performance of the proposed contention control strategies in the presented flexible optical packet-switched network.

5.5.3 Contention Control Analysis

In this section, the emphasis is on analysing the system performance of the proposed optical packet-switched network architecture employing different congestion control techniques. Based on the network dimensioning in the previous section, P is set to be 3 and K to be 4, thus the network configuration under study is composed of 3 Optical Packet Switches (OPSs), and each input/output port of the OPS carries 4 transmission channels. In this network configuration, three different contention management techniques are exploited under various offered network loads: recirculating Fibre Delay Lines (FDLs), packet retransmission mechanism and the combined FDLs/retransmission technique. Note that the offered network load in the uniform traffic pattern, as stated previously, is defined as $\rho \approx 1.33a$, whereas in the non-uniform traffic pattern, $\rho \approx 1.11a$.

5.5.3.1 Recirculating Optical Buffering Analysis

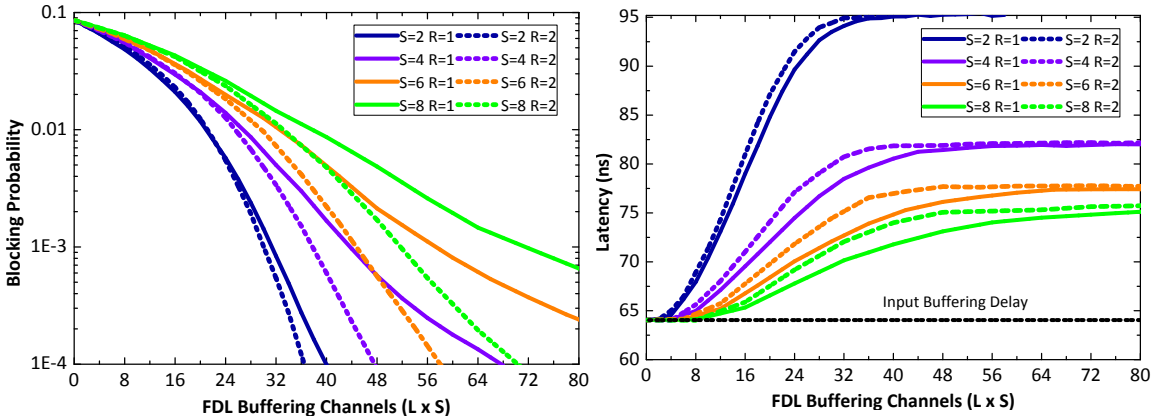
The performance evaluation of the recirculating optical buffers is analysed in this section with the objective of establishing the benefits of the recirculating FDLs in resolving packet contentions. Consider that the inter-rack traffic load a is set to 0.2, that is, 20% of traffic is generated and destined outside the rack (ToR switch), which is a typical value in cloud applications. This corresponds to an offered network load of $\rho = 0.267$ in the uniform traffic pattern and a network load of $\rho = 0.223$ in the non-uniform traffic pattern. The performance results of the analysed network with all-optical buffering are obtained for both the all-to-all communication pattern and the hot-spot traffic pattern. These results provide some important insights into the impact of the number of allowable FDL circulations, R, and the number of buffer channels per FDL, S, on contention and latency performance. Figures 5.10 - 5.11 show that under both uniform and non-uniform traffic patterns, an increase in FDL buffering capacity results in an evident reduction in the amount of traffic overflowing from the optical switches, with higher FDL buffering delays. This confirms that the deployment of optical buffers results in substantial improvements in packet loss and greatly enhances the resource utilisation of optical packet switches, but also induces additional delay. It is further demonstrated that the higher the number of TWCs S per FDL, the higher the packet loss rate and the lower the delays induced by FDL buffering. This is because the larger is the value of S, the smaller is the number of FDLs L, and as a consequence, the set of delay times provided by FDLs is reduced. In this case, despite the fact that the total number of optical buffering channels ($L \times S$) is the same, the effectiveness of FDLs in resolving congestion becomes lower, but the delay performance improves. Hence,



(a) Packet loss rate

(b) Overall communication latency

Figure 5.10: All-to-all communication pattern



(a) Packet loss rate

(b) Overall communication latency

Figure 5.11: Hot-spot traffic pattern

a trade-off between the packet loss rate and the average end-to-end latency needs to be considered when dimensioning the parameter S (the number of TWCs per FDL). Another important observation is that when the number of available FDL buffering channels is small, increasing the number of allowable FDL circulations, R, results in no major differences in contention performance, however the situation is substantially different for higher numbers of available FDL buffering channels, in which case increasing R yields a significant improvement on packet loss performance, and this improvement becomes increasingly more evident as the buffering capacity increases. Nonetheless, increasing R also results in an increase of communication latency.

The performance measurements illustrated in Figures 5.10 - 5.11 indicate that under both all-to-all communication pattern and hot-spot datacentre traffic pattern, with proper dimensioning of optical buffers, the vast majority of traffic can be switched all optically, while still achieving low communication latency in the datacentre network, verifying the efficacy of the optical buffering to diverse

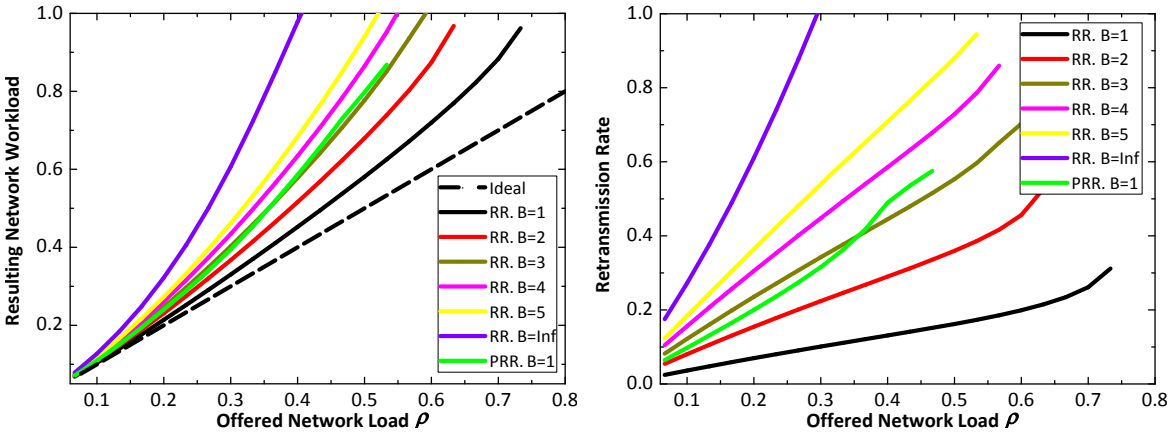
traffic patterns. Therefore, the system performance of the proposed optical network is to a large extent dependent on switch parameter settings, meaning that a target performance can be achieved, with proper network dimensioning. In the following experiments, the number of buffer channels per FDL is set to be 8 so as to guarantee the latency performance, and the number of FDLs is taken as 16. With this configuration, each switch input of the OPS, comprising K(=4) transmission channels, is assigned a virtual optical buffering channel.

Although increasing optical buffering capacity and the number of allowable FDL circulations results in a drastic reduction in blocking probability, packet contention still exists and cannot be eliminated, thus incurring traffic loss in the optical network. This is one of the main drawbacks of FDL buffering. Motivated by this, the packet retransmission mechanism is introduced to the optical network so as to recover the lost data, thereby achieving loss minimisation. The advantages and limitations of the packet retransmission scheme, when employed in the proposed optical network, will be studied and discussed in detail in the next section.

5.5.3.2 Packet Retransmission Analysis (without FDLs)

To investigate the effectiveness of the packet retransmission mechanism in recovering packet losses in the proposed flexible optical network, the performance quantities, including the resulting network workload, the packet retransmission rate, the system throughput, the retransmission buffer size, the packet loss rate and the overall communication latency, are evaluated under both uniform and non-uniform traffic patterns. Two different retransmission policies are considered: Random Retransmission (RR) and Pre-Reservation Retransmission (PRR), which were explained in Section 5.4. To explore the influence of the two retransmission techniques on network performance, extensive simulations are carried out in network scenarios with different numbers of allowable retransmissions B. In the experiments, the parameter B is set to be 0 (no retransmission),$1, 2, 3, 4, 5$ and ∞ (an infinite number of retransmissions). Note that in the PRR scheme, the maximum number of packet retransmissions required is 1, implying that the maximum value of B in this case is 1. Additionally, the link distance between the clusters and the core OPSs is set to be 25m, resulting in a propagation delay of 125ns. We examine the impact of increasing the link length in a later section.

The performance measurements of the optical network employing the two packet retransmission mechanisms are first evaluated and analysed under uniform traffic pattern. Shown in Figures 5.12 - 5.13 are the resulting network workload, the packet retransmission rate, the network throughput

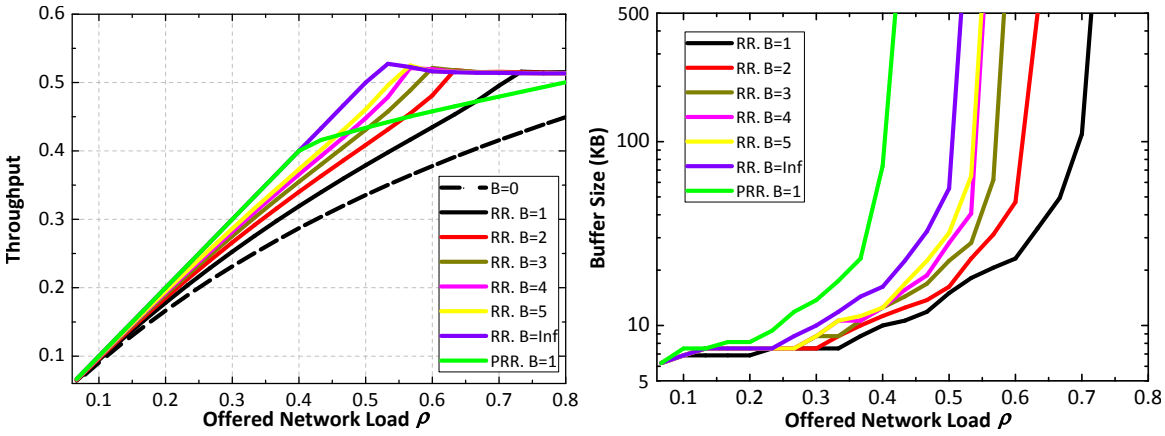


(a) Resulting network load versus offered network load ρ (b) Packet retransmission rate versus offered network load ρ

Figure 5.12: Network workload and packet retransmission rate under uniform traffic pattern.

and the required retransmission buffer capacity. Numerical results are plotted as a function of the offered network load ρ for different values of B. These graphs give an insight of the effectiveness of the two retransmission strategies in resolving traffic loss under various network traffic loads.

Figure 5.12 plots the resulting network workload and the packet retransmission rate versus the offered network load ρ for different retransmission scheme configurations. It is intuitive that in the RR scheme, increasing the number of allowable retransmissions B results in an evident increase in the resultant network workload. This is mainly due to the fact that increasing B leads to more frequent packet retransmissions, as shown in Figure 5.12(b), and as packet retransmissions occupy additional network resources, which in turn causes more frequent packet contentions, the amount of traffic in the network is greatly increased. It is further shown that as the input traffic load grows, the increase of the network workload becomes progressively more significant, and consequently, system overloading occurs. This overloading is indeed the overfilling of the input transmission channels of the optical network rather than the output tranmission channels, due to the PreRes scheduling strategy detailed previously. Evidently, the larger is the value of B, the faster is the network infrastructure overloaded. It should be noted that in the scenario of $B = \infty$, the network infrastructure is overloaded more rapidly than the other scenarios illustrated. More precisely, the system overloading happens when the network load is only around 40%. Furthermore, it is noticed that the PRR scheme appears to result in a smaller increment of the overall network workload and the packet retransmission rate than the RR scheme with $B \geq 4$, as it limits the number of required retransmissions to 1, but the PRR policy actually results in a more rapidly overloaded system than the RR scheme, owing to the low resource utilisation, which is attributed to the pre-reservation



(a) Network throughput versus offered network load ρ (b) Buffer size requirement versus offered network load ρ

Figure 5.13: Network throughput and buffer size requirement under uniform traffic pattern.

feature. This drawback of the PRR scheme is confirmed by the network throughput performance and the required buffer size plotted in Figure 5.13. However, also due to the pre-reservation feature, the PRR policy holds an important advantage over the RR scheme, that is, it guarantees the contention performance, since the maximum number of packet retransmissions required in PRR is only 1. This aspect of the PRR will be investigated in detail later.

Figure 5.13(a) illustrates the normalised network throughput under a wide range of offered network loads. The results show that the introduction of the packet retransmission mechanism has greatly enhanced the throughput performance, as it facilitates the data recovery in the optical network. From the graph, it is observed that the throughput improvement attained by the RR scheme largely depends on the number of retransmissions B. This is expected as increasing the number of retransmissions B allows for a greater number of lost packets to be recovered, thus improving the network throughput. Nevertheless, the relationship between the offered network load ρ and the resultant throughput exhibits a non-linear pattern, due to the existence of traffic loss in the optical network, as indicated in Figure 5.15. It is further shown in Figure 5.13(a) that at low network loads, the PRR scheme outperforms the RR scheme, but beyond a load of 0.4, the PRR scheme yields less throughput improvement. This mainly depends on the fact that the PRR scheme reserves the transmission channels for the lost packets prior to the packet retransmissions, which ensures the contention performance, but the downside is that the PRR scheme has low resource utilisation when the packet retransmissions are frequent.

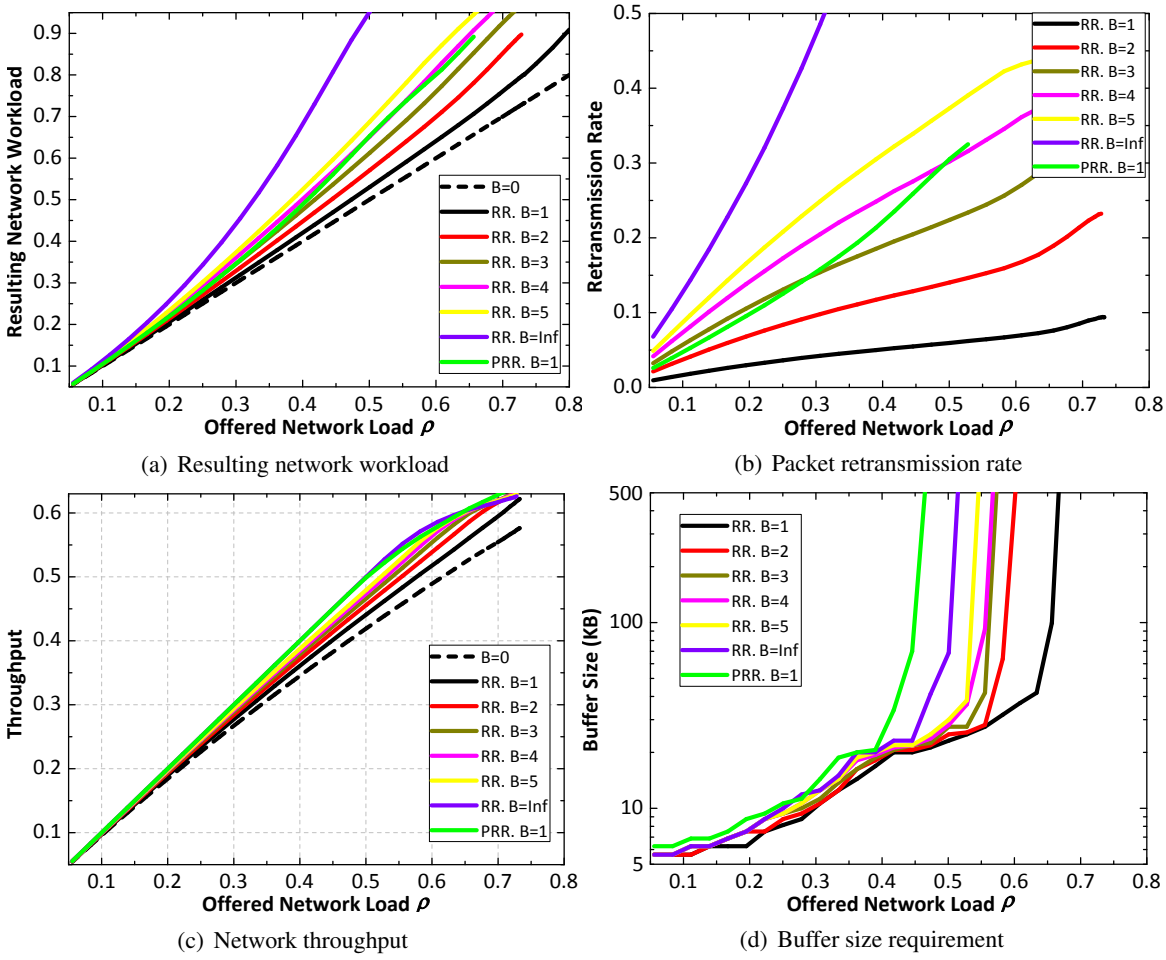
Another important observation is that in all scenarios tested, as the network load continues to grow, the throughput performance degrades significantly. Particularly, the larger is the value of B, the

earlier this degradation occurs. The reason behind this situation is the presence of frequent packet retransmissions, which greatly limits the network capacity due to significant network resources being wasted by failed transmission attempts. It should be pointed out that at the point where the system overloading just starts to occur, a small increment in network throughput is observed. As stated previously in Section 5.2, each input transmission channel of the OPSs is associated with a virtual electronic buffer queue in the cluster switch. When the system overloading just starts to occur, the virtual queue stores the new arriving packets and then transmits them towards the optical network where these packets may be successfully scheduled on the output channels of the OPSs. In this case, the overloading issue is slightly alleviated and thus a performance improvement is achieved. However, as the network load is further increased, the buffer queue is quickly overwhelmed and new arrivals are being blocked from the optical network, and consequently, no substantial improvement on network throughput can be achieved. It is noticed that the network throughput levels off at about 0.52, which corresponds to a normalised throughput of ~ 2.1 per switch port, as each switch port carries $K(=4)$ wavelength channels.

Aside from the impact on network throughput performance, the capacity requirement of the retransmission buffer also increases dramatically as the network load increases, see Figure 5.13(b). Importantly, a further increase in buffer size is observed when increasing the number of allowable packet retransmissions B. Once again, it is because frequent packet retransmissions considerably increase the resultant traffic load in the network, and accordingly a larger buffer size is required to store the copies of the increased amount of traffic.

Moreover, to illustrate the above mentioned performance characteristics of the packet retransmission schemes for the case of non-uniform traffic, another set of simulations is conducted under hot-spot traffic pattern. The measurements are presented in Figure 5.14. Clearly, the observations show a similar behaviour as the simulation results plotted in Figures 5.12 - 5.13. The difference is that although the average offered network load ρ is still low, the transmission channels at some switch ports are already overloaded, which is attributed to two main factors: (i) the existence of hotspots in the non-uniform traffic matrix and (ii) the increased resulting network workload incurred by frequent packet retransmissions.

From Figures 5.12 - 5.14, it is observed that in all scenarios under study, once system overloading occurs, the throughput performance degrades rapidly and the required buffer capacity increases drastically. This is an undesirable situation in datacentre networking. A common way of overcoming this issue is to limit the retransmission buffer and the number of allowable retransmissions so



(a) Resulting network workload

(b) Packet retransmission rate

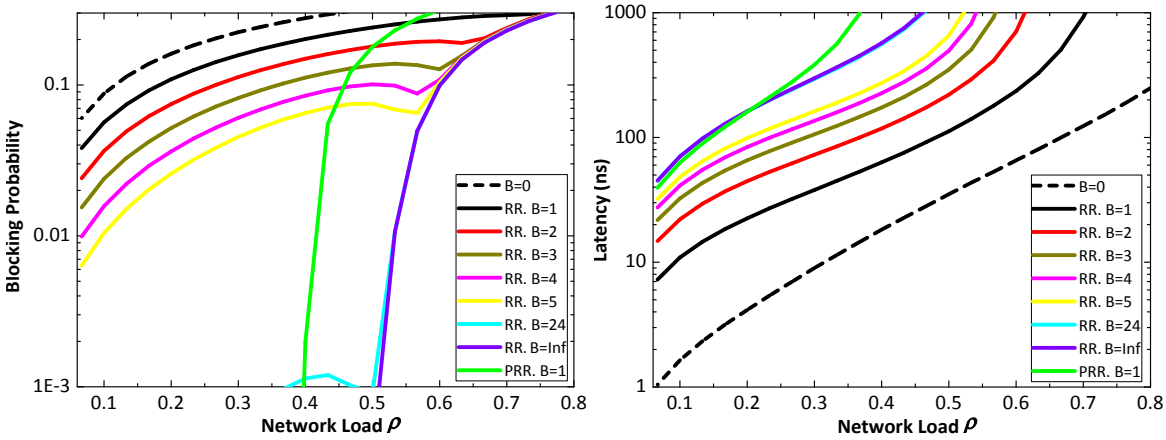
(c) Network throughput

(d) Buffer size requirement

Figure 5.14: (a) resulting network workload, (b) packet retransmission rate, (c) network throughput and (d) buffer size requirement are plotted as a function of the offered network load ρ under non-uniform traffic pattern.

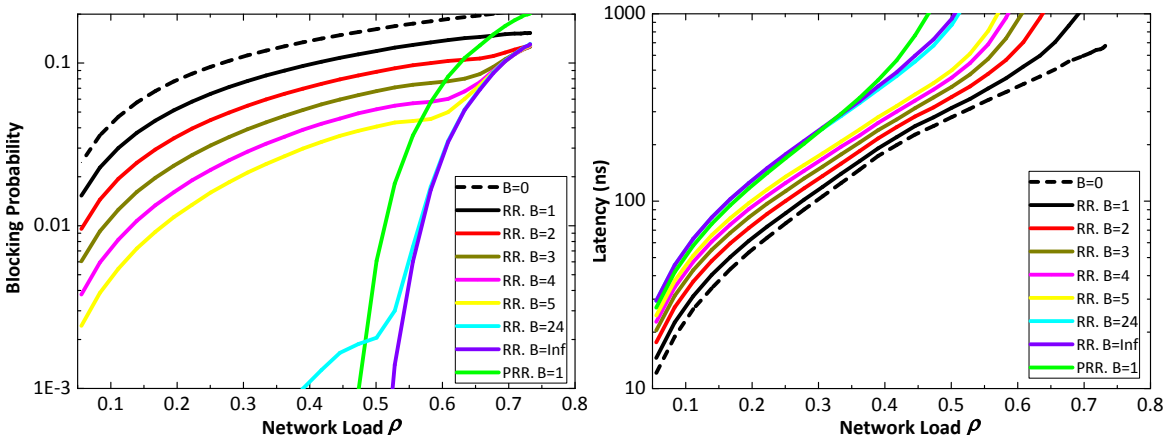
as to avoid the network capacity being wasted by unsuccessful retransmissions. Based on the buffer size analysis illustrated in Figures 5.13(b) and 5.14(d), the retransmission buffer length is limited to 25KB, which allows the storage of 40 packets. Nonetheless, limiting the buffer size leads to high packet loss, especially under heavy load conditions, due to traffic overflowing from the queue. This effect will be illustrated in the performance analysis regarding the packet loss rate and the communication latency in the following set of experiments. The results are taken for both the uniform and non-uniform traffic patterns.

Figures 5.15 and 5.16 show the performance characteristics with respect to packet loss rate and mean end-to-end communication delay under both uniform and hot-spot traffic patterns. The measurements indicate that in the RR scheme, as the number of allowable retransmissions B increases,



(a) Blocking probability (b) Overall communication latency

Figure 5.15: Network performance versus offered network load ρ under uniform traffic pattern.



(a) Blocking probability (b) Overall communication latency

Figure 5.16: Network performance versus offered network load ρ under hot-spot traffic pattern.

the contention performance improves considerably. Of course, there is a trade-off to be considered as frequent packet retransmissions introduce additional input queueing delay and retransmission delay. Evidently, when B increases to a sufficiently high value, say $B = 24$, there is no major difference in performance between the scenarios of $B = 24$ and $B = \infty$. Note that in the case of $B = \infty$, although the number of allowable packet retransmissions is unlimited, traffic loss still occurs under very high load conditions due to retransmission buffer overflow, as a result of the retransmission buffer size being limited. It has also been found that PRR exhibits a very similar performance pattern as RR with large B. The observations show that the PRR scheme generally achieves significantly better loss performance than the RR strategy, but it leads to a higher communication latency. This is attributed to the fact that PRR guarantees the successful retransmission of the lost packets,

but to achieve this, it may need to store the retransmitted packets in the virtual buffer queue for a longer time so as to avoid input contentions between the retransmitted packets at switch inputs. The figures further illustrate that as the network load ρ increases, the packet loss rate and the communication latency first rise smoothly, but at a certain point the performance impairment becomes more evident, because the optical network has reached its capacity limitation as a consequence of frequent packet retransmissions. At this point, the packet loss, arising from traffic overflowing from the limited buffer queue, starts to dominant the overall contention performance. It should be pointed out that, in Figure 5.15, before reaching this particular point, the performance degradation appears to slow down. This is attributed to the effective network load being reduced, since part of the offered network load is blocked from the optical network due to limited buffer depth, and thus the contention problem is slightly improved, yielding a small reduction in the average packet loss rate. Beyond this point, there are no major differences in loss performance between different scenarios, but a more significant increase of the communication delay is observed.

The above performance analysis suggests that the packet retransmission mechanism can substantially enhance the contention and throughput performance of the proposed flexible optical network, with good latency performance under low traffic loads, especially the PRR scheme. Nevertheless, at high traffic loads, the network performance degrades drastically, arising from the system overloading issue due to frequent packet retransmissions. Thus, the retransmission mechanism is only suitable when the amount of traffic that needs retransmission is low. For that reason, the packet retransmission scheme is utilised in conjunction with optical buffering which has proven to be highly effective in lowering packet loss, and is capable of carrying the majority of the contention traffic with proper dimensioning. This hybrid congestion control scheme will be the subject of investigation in the following section.

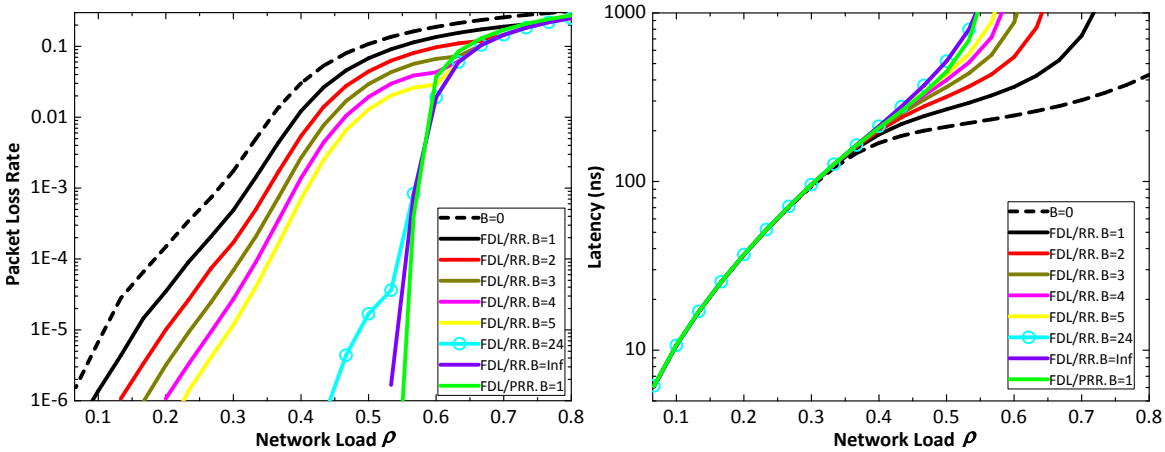
5.5.3.3 Hybrid Recirculating FDL/Retransmission Analysis

In the hybrid optical buffering/retransmission scheme, the recirculating optical buffering (FDLs) is utilised as the primary contention resolution technique, and the traffic that overflows from FDL buffers is dropped and then recovered by the packet retransmission mechanism. However, frequent retransmissions in the network, as previously detailed, would lead to system overloading issue. To overcome this challenge, it is necessary to provide sufficient optical buffering capacity so that the amount of colliding traffic that needs packet retransmission is reduced to a sufficiently low level. With this arrangement, the recirculating FDL buffering and the packet retransmission mechanism

compensate for each other's disadvantageous effects. In this section, the main goal is to measure the Quality-of-Service (QoS) of different hybrid FDL/retransmission congestion control schemes under both uniform and non-uniform traffic patterns in the proposed flexible network architecture. The QoS measurement is quantified by performance metrics such as the packet loss rate, the communication latency and the packet retransmission rate. Simulation results indicate that by combining the recirculating FDLs and the packet retransmission mechanisms, packet collisions in the proposed datacentre network can be efficiently resolved without impairing the latency performance. Additionally, the robustness of the flexible network architecture with hybrid optical buffering/retransmission scheme to different communication patterns is also verified.

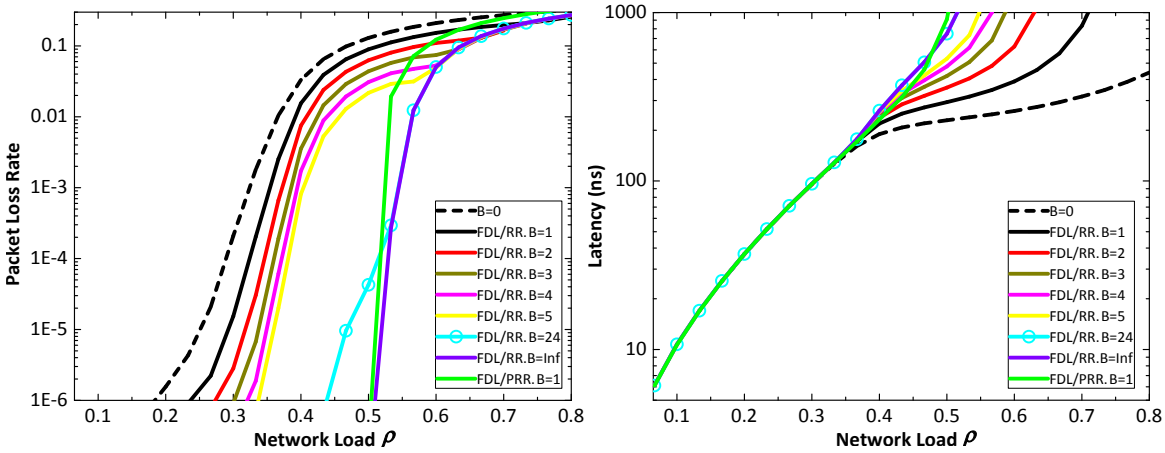
Before proceeding to the performance analysis, the dimensioning of the FDL buffers is first resolved, based on the performance evaluation of the FDL buffering conducted in Section 5.5.3.1. Consider the trade-off between the packet loss rate and the communication latency, the parameter S is set to 8. Further, to make sure that each switch input of the OPS, which comprises K (= 4) transmission channels, is assigned a virtual buffering channel, L is dimensioned as 16. Alternatively, on the basis of the packet retransmission analysis in Section 5.5.3.2, the retransmission buffer size is set to 25KB, which allows the storage of 40 packets. Given these measurements, the system performance of the proposed flexible optical network employing the coordinated recirculating FDLs/retransmission contention scheme is investigated through extensive simulations.

Figures 5.17 - 5.19 plot the network performance, measured in terms of packet loss, latency and retransmission rate, as a function of the offered network load ρ for two different cases of $R = 1$ and 2 under all-to-all traffic pattern. Figures 5.20 - 5.22 illustrate the performance characteristics of the proposed optical network under hot-spot communication pattern. For the purpose of comparison, the packet loss and the communication latency of the optical network using FDLs only (no retransmission $B = 0$) are also presented in figures (black dashed curves). Numerical results show that the combined FDLs/retransmission approach greatly enhances the contention and latency performance, when comparing to the standalone uses of the FDL buffers and the packet retransmission schemes, which were investigated in the previous sections. Also shown in figures is how the contention probability can be reduced by up to several orders of magnitude by increasing the number of FDL circulations, R, and the number of packet retransmissions, B. Note that in the scenario of FDL/RR, $B = \infty$, despite that an unlimited number of retransmissions is allowed, traffic loss still exists in the optical network. The same situation occurs in the scenario of FDL/PRR, $B = 1$. This performance degradation results from the overflow traffic from the limited retransmission buffer. Note that as



(a) Packet loss rate versus offered network load ρ (b) Communication latency versus offered network load ρ

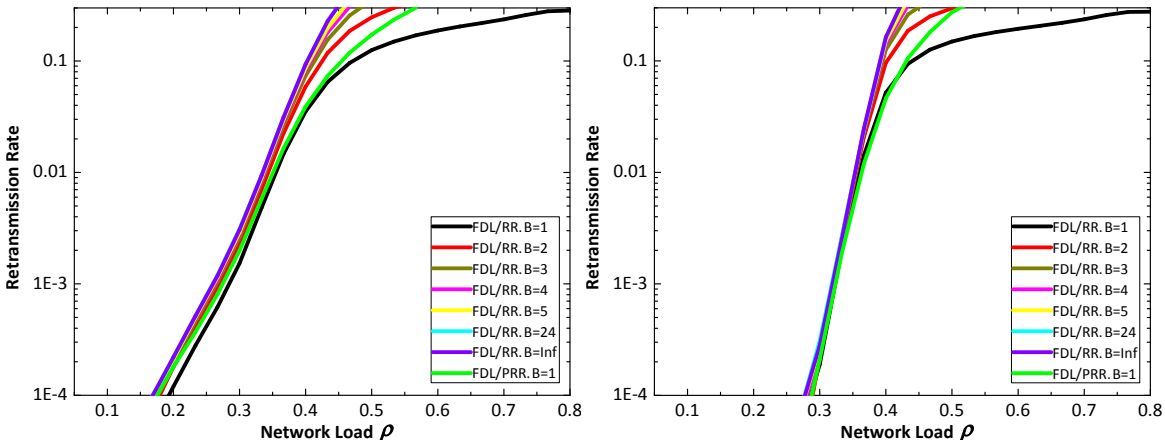
Figure 5.17: Network performance of the combined contention control scheme with $R = 1$ under uniform traffic pattern.



(a) Packet loss rate versus offered network load ρ (b) Communication latency versus offered network load ρ

Figure 5.18: Network performance of the combined contention control scheme with $R = 2$ under uniform traffic pattern.

the offered network load ρ increases, this effect becomes more significant, due to the fact that packet contentions at higher network loads are more frequent, leading to more packets overflowing from the buffer. It is also noticed that increasing the number of allowable retransmissions B in the hybrid FDL/RR scheme yields an evident reduction in packet loss rate, without significantly increasing the communication latency. Nonetheless, increasing B to a certain value in the hybrid FDL/RR scheme, the performance gain appears to be less obvious. Indeed, there is a small difference in performance between the scenarios of $B = 24$ and $B = \infty$. The simulation results further show that the hybrid FDL/PRR strategy achieves considerably better network performance than the hybrid FDL/RR scheme under low and medium load conditions. Conversely, under very high load conditions, the FDL/RR strategy, especially in the case of $B = \infty$, outperforms the FDL/PRR scheme.



(a) Number of FDL circulations R is 1

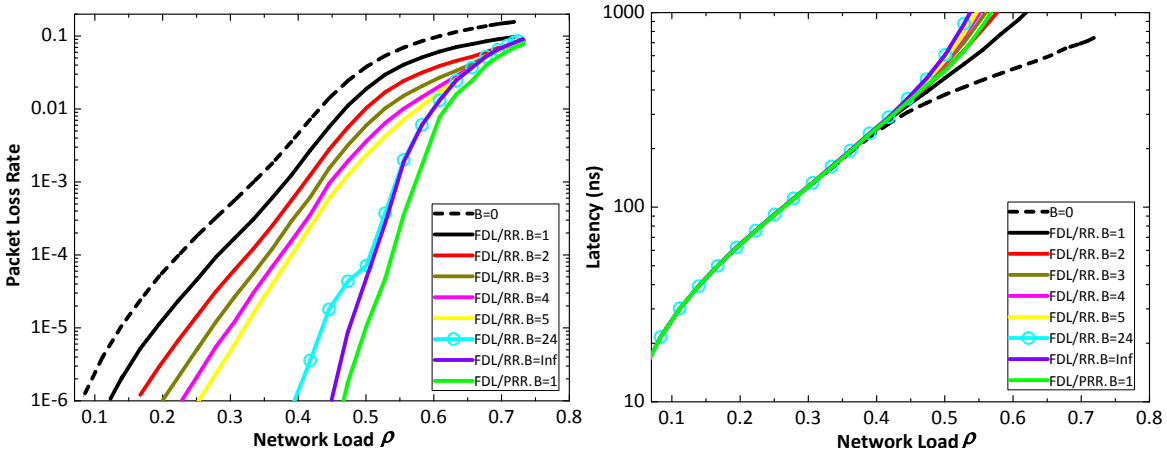
(b) Number of FDL circulations R is 2

Figure 5.19: Packet retransmission rate in the combined recirculating FDL/retransmission scheme under uniform traffic pattern.

This performance difference is attributed to the low resource utilisation of the PRR policy under heavy traffic conditions, because of the pre-reservation feature.

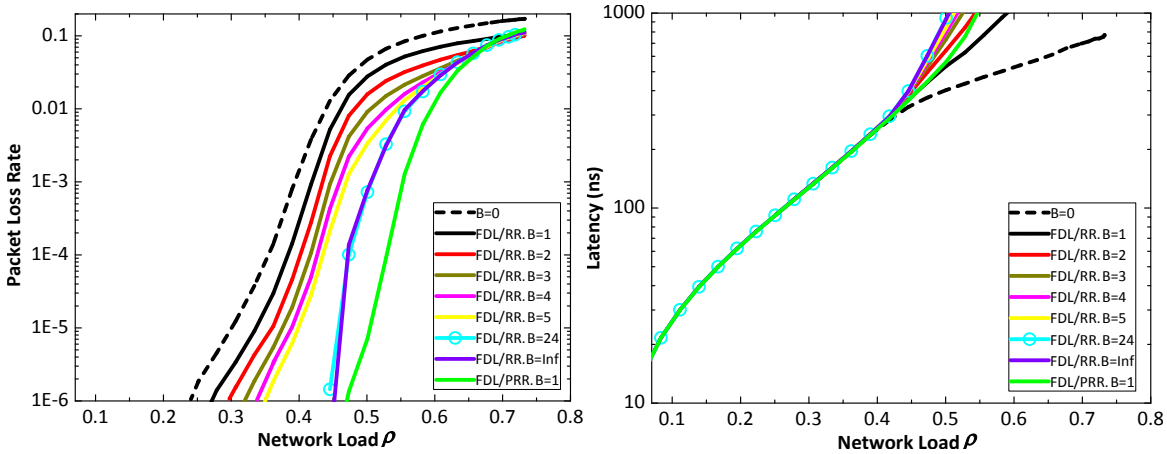
Under both uniform and non-uniform traffic patterns, an evident improvement in contention performance is obtained by increasing the number of FDL circulations R from 1 to 2. This is expected as increasing R results in a more efficient use of the shared buffering resources, thus allowing a greater number of the contending packets to be carried by the optical buffering. This in turn largely reduces the packet retransmission rate in the network, which will be illustrated later. Alternatively, it is observed that for a target packet loss rate equal or lower than 10^{-6}, the overall network utilisation is also greatly improved. For example, in Figures 5.17(a) and 5.18(a), the analysis shows that in the scenario of hybrid FDL/RR scheme with $B = 5$, by increasing R from 1 to 2, the network utilisation is enhanced by more than 10%, approximately up to 34%. In addition, a network utilisation close to 45% is supported by the hybrid FDL/RR scheme with $B = 24$. Furthermore, the hybrid FDL/PRR scheme achieves a network utilisation of 55%, which corresponds to a mean resource utilisation of 220% per input/output port of the OPS, since each switch port contains K $(= 4)$ transmission channels, as described in Section 5.5.1.1. Nevertheless, this improvement comes at the cost of a significantly increased end-to-end communication latency.

As mentioned above, in the hybrid scheme, the inclusion of the optical buffering can efficiently address the packet collisions in the optical network, thus the packet retransmission rate is significantly decreased. This aspect is illustrated in detail in Figures 5.19 and 5.22, where the packet retransmission rate is plotted as a function of the offered network load ρ. Comparing to the results shown in Figures 5.12(b) and 5.14(b), the packet retransmission rate in the network is dramatically low-



(a) Packet loss rate versus offered network load ρ (b) Latency versus offered network load ρ

Figure 5.20: Network performance of the combined contention control scheme under hot-spot traffic pattern. Note that $R = 1$.

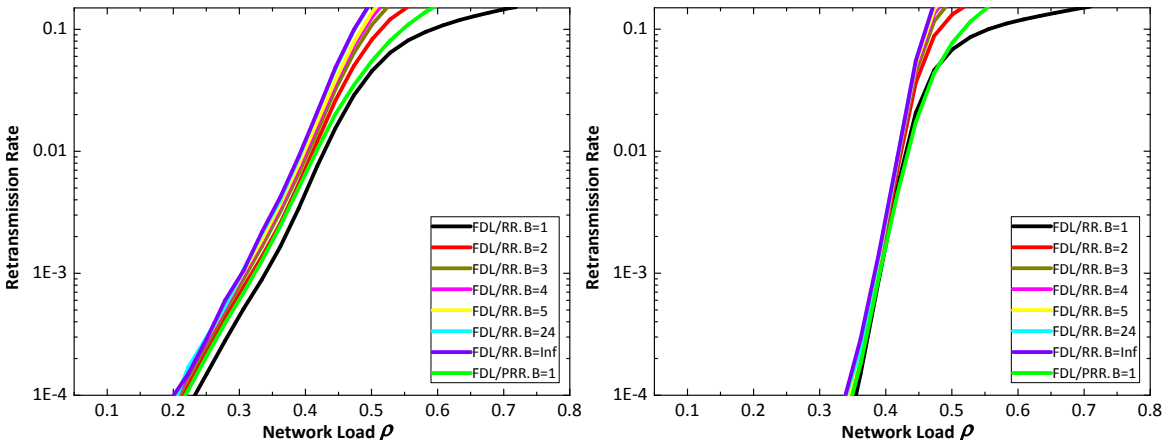


(a) Packet loss rate versus offered network load ρ (b) Latency versus offered network load ρ

Figure 5.21: Network performance of the combined contention control scheme under hot-spot traffic pattern. Note that $R = 2$.

ered, by several orders of magnitude. Moreover, a further reduction in packet retransmission rate is gained by increasing the number of allowable FDL circulations R from 1 to 2. It should be noted that at high network loads, the recirculating FDLs become increasingly less effective at resolving contentions, and therefore the amount of traffic that needs retransmissions increases greatly. As a consequence, the packet loss due to traffic overflowing from the retransmission queue dominates the overall system performance.

Simulation studies presented in this section give an indication as to the level of performance enhancement that can be attained by combining the recirculating optical buffers and the packet retransmission techniques in the proposed flexible datacentre network. It is indicated that a high net-



(a) Number of FDL circulations R is 1

(b) Number of FDL circulations R is 2

Figure 5.22: Packet retransmission rate in the combined contention control scheme under non-uniform traffic pattern.

work link utilisation, approximately up to 55%, can be achieved by using the hybrid recirculating FDL/PRR scheme, which corresponds to an average resource utilisation of 220% per input/output port of the OPS, thus highlighting the efficiency of the proposed optical network structure.

5.6 Impact of Geographical Distance on Network Performance

The performance analysis in the previous section was performed under the assumption that the link distance between the cluster and the core OPS is set to 25m. In this section, additional experiments are conducted under link lengths ranging from 25m to 1000m so as to indicate the robustness of the proposed datacentre network structure and the packet retransmission schemes to geographical distance. In simulations, the number of allowable FDL circulations is fixed to be 2, and the network load is taken as $\rho = 0.35$ ($a \simeq 0.25$). Note that only worst-cast uniform traffic pattern is considered. Figure 5.23 plots the contention performance of the optical network with hybrid recirculating FDLs/retransmission scheme as a function of the link length in different network scenarios. In all scenarios illustrated, the packet loss rate first reduces greatly with the increase of the link length. The main reason is that the increased link length leads to significantly increased Round-Trip Time (RTT), which means that the time window between the current time and the future arriving time of the retransmitted packet is enlarged, as illustrated in Figures 5.4 and 5.5, and therefore it is more likely for the congestion to be resolved between two colliding packets. Overall, the proposed network achieves excellent contention performance for link length up to 800m. Nevertheless, as the link distance continues to increase, the contention performance degrades significantly, due to the

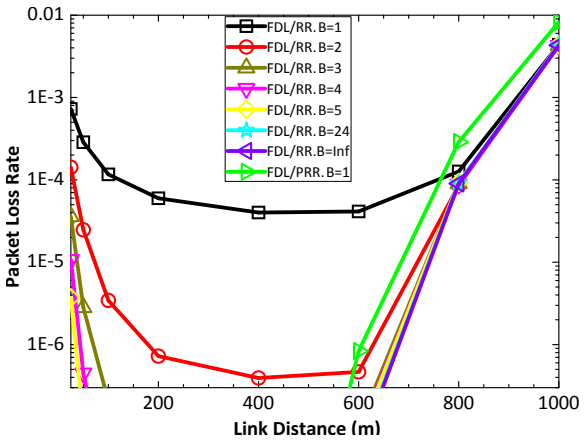


Figure 5.23: Contention performance

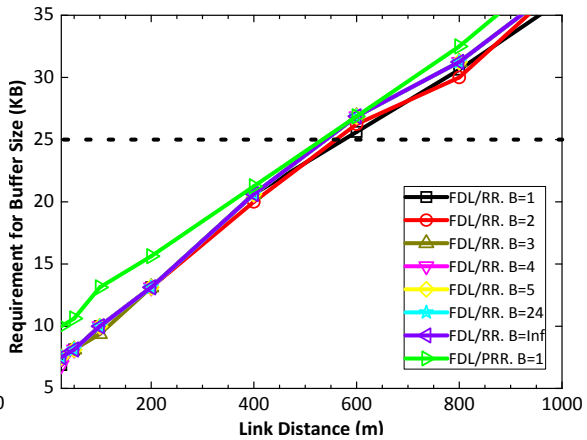
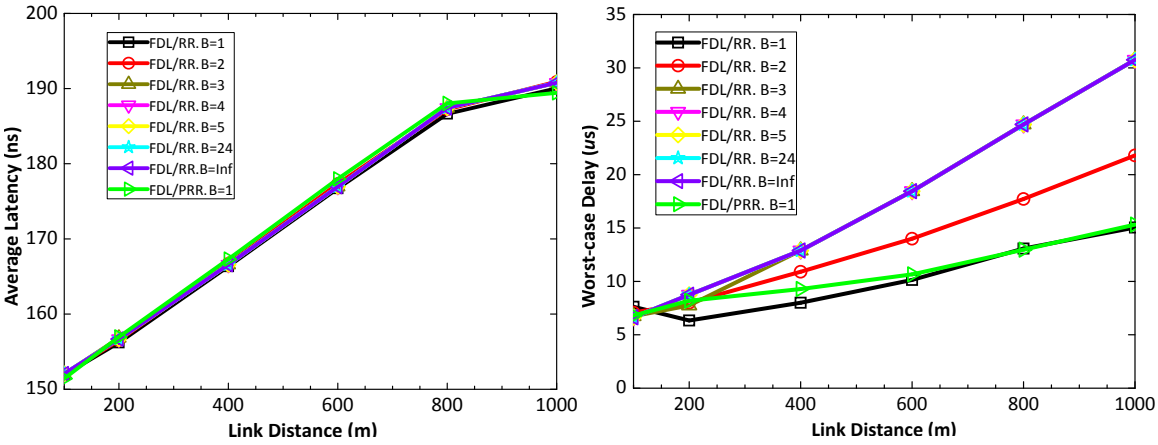


Figure 5.24: Requirement for buffer size

fact that the retransmission buffer is overfilled and the problem of traffic overflow occurs. This is confirmed by the simulation data in Figure 5.24, which plots the requirement for buffer size as a function of the link distance. It is evident from the figure that if the link distance increases, the required retransmission buffer size will also increase. The reason behind this situation, as noted previously, is that increasing the link distance results in a significantly increased RTT, and accordingly the timeout value is also greatly increased, meaning that the copy of a transmitted packet needs to be stored in the retransmission buffer for a longer time, thereby a larger buffer size is clearly required. However, the available buffer size in the simulations is fixed to be 25KB, and as a result, the traffic overflowing issue arises when the link length increases to a sufficiently large value and the required buffer depth is larger than 25KB. The solution to this particular problem is to increase the retransmission buffer size, as illustrated in Figure 5.24.

On the other hand, observing the latency performance plotted in Figure 5.25(a), it can be seen that the link distance has a small impact on the average end-to-end delay. Specifically, when the link length increases from 25m to 1000m, the mean communication latency only increases by approximately 25%. In addition, there is almost no relative difference in latency between different scenarios. This is due to the fact that the vast majority of traffic is carried by the optical network, and only a very small fraction of traffic needs retransmissions, thus the time delay introduced by the packet retransmission mechanism contributes slightly to the overall communication latency. Differently, Figure 5.25(b) illustrates that the worst-cast communication latency grows progressively with the increase of the link distance. Furthermore, the results show that in the FDL/RR scheme, increasing the number of allowable packet retransmissions, B, leads to an evident increase in the worst-cast communication latency, but no substantial increase is observed when B reaches 3. This



(a) Average end-to-end latency

(b) Worst-case latency

Figure 5.25: Latency performance as a function of the link length.

can be explained by the fact that under this load condition, when the link distance is larger than 400m, the maximum number of required packet retransmissions is 3, since in this case the traffic loss rate has already been reduced to a sufficiently low level, see Figure 5.23 (note that the degradation is caused by the traffic overflowing). Another important observation is that although the combined FDL/PRR scheme results in a slightly higher average latency, it achieves significantly lower worst-case communication latency.

5.7 Conclusions

In this chapter, a hybrid modular interconnection network, partially based on mature electronic switching, is proposed for future large-scale datacentres or HPC systems. The proposed OPS/EPS network employs a modular design which not only addresses the capacity limitations on the aggregation switches, but also makes the network highly scalable, thus providing a migration path from existing networks. The introduction of transparent optical packet switching technologies to the core network of the intra-DC platform promises high bandwidth capacity and high switching speed. Importantly, in order to further enhance the network flexibility and throughput, a flexible global routing algorithm based on a least-loading heuristic algorithm is developed to exploit the path diversity in the network. Moreover, to resolve the contention problem in the proposed flexible optical network, a congestion control technique combining the recirculating Fibre Delay Lines (FDLs) and the packet retransmission mechanism has been employed. By simulations, it has been found that the combined recirculating FDL/retransmission contention resolution technique has a large impact

with regard to reducing the traffic loss in the optical network, while still achieving good latency performance. A significant enhancement on network link utilisation, up to around 55% for a target contention performance of 10^{-6}, can be achieved. This corresponds to an average resource utilisation of 220% per switch input/output port, as each switch port carries $K = 4$ transmission channels. It has been further illustrated that the proposed network structure can provide excellent performance even for very large link lengths. Thus the proposed optical network architecture employing the hybrid contention management strategy provides robust performance under different traffic patterns for a wide range of geographical distances, and therefore has the potential to meet the requirements of future datacentre networks.

Bibliography

[1] R. Proietti, Y. Yin, R. Yu, X. Ye, C. Nitta, V. Akella, and S. J. B. Yoo, "All-optical physical layer NACK in AWGR-based optical interconnects," *Photonics Technology Letters, IEEE*, vol. 24, no. 5, pp. 410-412, Mar. 2012.

[2] Akbar G. Rahbar, "Quality of Service in Optical Packet Switched Networks," Wiley-IEEE Press, Apr. 2015.

[3] S. Di Lucente, R. P. Centelles, H. J. S. Dorren, and N. Calabretta, "Study of the performance of an optical packet switch architecture with highly distributed control in a data center environment," in *16th International Conference on Optical Network Design and Modeling (ONDM)*, pp.1-6, Apr. 2012.

[4] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *ACM SIGCOMM*, 2010.

[5] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken,"Nature of datacenter traffic: measurements and analysis," in *ACM SIGCOMM*, 2009.

[6] A. Greenberg, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. Maltz, P. Patel, and S. Sengupta, "VL2: A scalable and flexible data center network," in *ACM SIGCOMM*, Aug. 2009.

[7] FP Tso and D. Pezaros, "Improving data center network utilization using near-optimal traffic engineering," *IEEE Transactions on Parallel & Distributed Systems*, vol. 24, no. 6, pp. 1139-1148, June 2013.

[8] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: dynamic flow scheduling for data center networks," in *Networked Systems Design and Implementation (NSDI) Symposium*, Apr. 2010.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The continuous growth of datacentre traffic imposed by existing and new applications motivates the deployment of high-capacity, large-scale data networks. To construct these next-generation datacentre and High Performance Computing (HPC) networks, advanced optical networking concepts and technologies and novel optical components are being exploited to build high-capacity datacentre networking architectures, which are capable of providing high scalability, super-high bandwidth capacity, fast switching speed, low latency, wavelength-based switching/routing, data rate/format transparency and low power consumption per bit transmitted. Further, a high level of flexibility and reconfigurability is also required to be incorporated into the future optical switching infrastructure so as to achieve high utilisation of the shared optical network resources and ensure Quality-of-Service (QoS).

To overcome these challenges, remarkable technological advances in integrated photonic technologies, especially in Arrayed Waveguide Grating (AWG) and Tunable Wavelength Converter (TWC) switching technology, have been realised. The AWG-based Optical Packet Switch (OPS) holds some attractive features including ultra-high bandwidth, high-speed switching, low power consumption, operational flexibility and high scalability, thereby it offers a promising and viable networking solution to address current and future application needs in datacentres, and deserves continued attention. Nonetheless, contention resolution in this type of switch is problematic, due to the absence of viable optical Random Access Memory (RAM) technology. In this thesis, various wavelength-routed optical network architecture employing advanced enabling optical technologies

and contention resolution techniques are proposed, with the purpose of building large-scale flattened datacentre networks with improved energy efficiency and network performance. The performance evaluation, power estimation, bandwidth provisioning and load scheduling of the proposed optical packet switches is a major focus of this work.

In Chapter 2, a novel high-port-count optical packet-switched architecture, where optical components AWG and TWCs are employed to support high-speed switching, is developed for use in datacentre and HPC networks. The wavelength-routed switch architecture is configured with a hybrid electronic/optical buffering scheme combining multi-wavelength Fibre Delay Line (FDL) buffering and a small-size electronic buffer. The shared optical buffering is utilised as the primary contention resolution technique, and the electronic buffer is used to sustain the overflow traffic from FDLs, thus together they facilitates the realisation of a non-blocking switching scheme. Numerical results show that with proper dimensioning, the hybrid-buffered AWG switch achieves significantly increased overall energy efficiency, by as much as 46%, compared to electronic only buffering, while maintaining low latency and non-blocking performance. A further investigation on the computational complexity of the required scheduling algorithm in the hybrid-buffered switch is performed, which in turn allows the estimation of the required processing power of the switch controller. All these findings confirm the feasibility of the AWG switch in large-scale networks.

An alternative method to resolve contention in wavelength-routed AWG switches is to employ recirculating FDL buffers of variable lengths, which is the topic of research in Chapter 3. In this scheme, a contending packet is allowed to traverse through optical buffers and the optical switch multiple times until the destined switch transmission channel becomes free. A significant challenge of using the recirculating FDLs is the resource optimisation of the proposed optical-buffered OPS architecture. To address this, an analytical performance model, which accurately captures the behaviours of the proposed OPS network with optical buffers, is derived, allowing a detailed analysis of the impact of various network design parameters and traffic characteristics on network performance with regard to the blocking probability and the communication latency. Complemented with a heuristic procedure, the mathematical model is extended to resolve the dimensioning problem which is formulated as a constrained optimisation problem with the objective of allocating the appropriate network parameter configuration which potentially achieves the most power savings, simultaneously meeting network performance constrains in terms of contention probability and overall latency. It is shown that the obtained optimal solution can potentially improve the energy efficiency by 42% in the scenario of $N = 64$, in comparison with the power requirement of the worse-case solution,

and for $N = 256$, the optimal solution results in a reduction of 30% in power consumption. Therefore, the proposed queueing model and the optimisation algorithm provides a reliable solution for designing and dimensioning this type of optical switch.

Further, to enhance network flexibility and reconfigurability so as to adapt to the dynamic, fast shifting traffic patterns in datacentres, Chapter 4 proposes a dynamically reconfigurable optical packet switching design, based on the AWG and the WSSs, by incorporating a flexible resource allocation algorithm into the switch. The proposed switch architecture supports the dynamic allocation of network capacity to traffic demands based on application requirements. The key to enabling this functionality is the deployment of a Wavelength Selective Switch (WSS) at each switch output, which is a flexible switching technology supporting finer-grained arbitrary bandwidth granularity. The network performance of the proposed flexible switch architecture is evaluated and analysed under both a uniform communication pattern and a more realistic datacentre traffic pattern. The simulation results show that the packet loss probability can be reduced to a sufficiently low level, which is of the order of 10^{-9} under the hotspot traffic pattern. Further, the comparison of the minimum power dissipation between the proposed flexible network scheduling and the commonly used fixed scheduling reveals that a significant power reduction, up to 25%, is obtained by the flexible scheme, and it is anticipated that more energy savings can be achieved by performing a full-area searching. Therefore, the proposed flexible optical switching architecture makes more efficient use of the shared network resources, considerably alleviating the contention conditions, and greatly improves the network's ability to adapt to changing communication patterns, which in turn reduces the required optical switching and buffering capacities, thus facilitating the realisation of power-efficient switching.

In Chapter 5, a large-scale high-performance datacentre network is developed based on the optical switch architecture proposed in Chapter 4. The datacentre network employs a modular design, which makes it highly scalable, capable of hosting hundreds of thousands of servers. The core network of the intra-datacentre platform is composed of multiple transparent optical packet switches based on Arrayed Waveguide Grating (AWG) routers and Wavelength Selective Switches (WSSs), thus supporting ultra-high bandwidth capacity. A key advancement of this network structure is the introduction of a flexible network scheduling technique, which greatly enhances the network flexibility and reconfigurability. Another aspect of this network is the deployment of a combined congestion control scheme using recirculating Fibre Delay Lines (FDLs) along with novel packet retransmission schemes. By simulations, it has been found that the combined recirculating FDL/re-

transmission scheme has a large impact with regard to reducing the traffic loss in the optical network, while still maintaining good latency performance. A significant enhancement on network utilisation, up to around 55% for a target contention performance of 10^{-6}, can be achieved, which corresponds to a resource utilisation of 220% per switch port of the OPS. A further study of the impact of the geographical distances on network performance indicates that the proposed network structure achieves good performance even for large link distances. Thus the proposed optical network architecture employing the coordinated contention management technique provides robust performance under different traffic patterns for a wide range of geographical distances, and therefore has the potential to meet the requirements of future datacentre networks, where longer inter-switch links become necessary.

6.2 Future Work

The work in this thesis reveals the enormous potential of wavelength-routed optical packet switches in meeting current and future datacentre application requirements, and provides contributions towards the construction of future large-scale high-capacity datacentre/HPC network infrastructure. Nonetheless, significant effort still needs to be devoted to further enhancing the optical switching and networking technologies related to this work, so as to realise the wide deployment of commercially viable optical packet-switched datacentre networks.

The design of large-scale optical networks using optical packet-switched WDM networks has been well investigated in this work, and the author believes that further improvements could be achieved by subscribing to the advantages of the hybrid architecture which is designed based on Optical Circuit Switching (OCS) and Optical Packet Switching (OPS) paradigms. In the joint implementation of OPS and OCS, large port-count optical packet switches with fast switching speed, are responsible for accommodating short-lived bursty traffic, due to its packet-level granularity, whereas slow, high port-count optical circuit switches target larger data transfers. This parallel mode of hybrid operation could provide finer switching granularity, greatly improved network flexibility and higher data rates/improved performance, and also expand the network connectivity. However, an important requirement in this hybrid architecture is to map the applications to the transport network that best suits the traffic demands. To overcome this, the development of a sophisticated network scheduling algorithm, which supports efficient traffic distribution, and thus maximises the merits of the two optical switching paradigms, is clearly necessary. The investigation of the control complexity and

the operation time in the hybrid network is another important subject.

Further exploration also includes the study of the dynamics of the flexible network scheduling procedure when the traffic demands evolve over time in the proposed optical datacentre networks. The work in this thesis only dealt with the performance analysis of the flexible network scheduling algorithm under one hot-spot traffic pattern, and it would be of great interest to investigate how the flexible network scheduling reacts to the constantly changing traffic patterns, particularly, to study how the network capacities are re-assigned based on the predicted traffic demands of the applications, and how the optical network is reconfigured. In doing so, the impact of the traffic heterogeneity on network performance can be examined and analysed, which gives an indication as to the effectiveness, robustness and suitability of the proposed flexible optical network structures in supporting traffic heterogeneity in datacentre/HPC applications.

In Chapter 4, a simple analytic framework was proposed to model the system performance of the proposed flexible OPS architecture, but it only characterises the switching function and neglects the behaviours of the optical buffering. To gain a more comprehensive understanding of the optical-buffered flexible optical packet switch, a more complete mathematical model, which accurately captures the functionalities of the flexible scheduling and switching and all-optical buffering, will need to be developed. The new mathematical queueing model is expected to take into account all the important network design parameters, thus allowing for the resolution of the resource optimisation problem of the optical-buffered OPS architecture. Further, with the development of such a mathematical model, dimensioning at the network level is also made possible. In particular, using the new queueing model, the resource dimensioning of the large-scale datacentre network architecture proposed in Chapter 5 could potentially be addressed, thereby facilitating the prediction of the required networking resources in datacentre networks so as to guarantee the QoS in a cost and energy efficient manner.

Appendix A

List of Publications

1. **J. Wang**, C. McArdle, and L. P. Barry, "Modelling and Dimensioning of a High-Radix Datacentre Optical Packet Switch with Recirculating Optical Buffers," *Elsevier Optical Switching and Networking*, pending review.
2. **J. Wang**, C. McArdle, and L. P. Barry, "Retransmission Schemes for Lossless Transparent Optical Packet Switching in Large-scale Datacentre Networks," in *IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, Vienna, Austria, Aug. 2016.
3. **J. Wang**, C. McArdle, and L. P. Barry, "Large-scale Optical Datacentre Networks using Hybrid Fibre Delay Line Buffers and Packet Retransmission," in *18th International Conference on Transparent Optical Networks (ICTON)*, Trento, Italy, July 2016.
4. **J. Wang**, C. McArdle, and L. P. Barry, "Optical Packet Switch with Energy-Efficient Hybrid Optical/Electronic Buffering for Data Center and HPC Networks," *Springer Photonic Network Communication*, DOI 10.1007/s11107-015-0578-z, pp. 1-15, Nov. 2015.
5. **J. Wang**, S. Basu, C. McArdle, and L. P. Barry, "Large-scale Hybrid Electronic/Optical Switching Networks for Datacenters and HPC Systems," in *IEEE 4th International Conference on Cloud Networking (CloudNet)*, Niagara Falls, Canada, Oct. 2015.
6. **J. Wang**, C. McArdle, and L. P. Barry, "Energy-Efficient Optical HPC and Datacenter Networks using Optimized Wavelength Channel Allocation," in *Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems, (SPECTS), Best Paper Award,* Chicago, USA, July 2015.
7. **J. Wang**, C. McArdle, and L. P. Barry, "Energy-Efficient Optical Packet Switch with Recirculating Fiber Delay Line Buffers for Data Center Interconnects," in *16th International Conference on Transparent Optical Networks (ICTON)*, Graz, Austria, July 2014.