Evaluation of Data Centre Networks and Future Directions

Sohini Basu

B.Tech., M.Sc.(Eng)

A dissertation submitted in fulfilment of the

requirements for the award of

Master of Engineering by Research

School of Electronic Engineering

Faculty of Engineering and Computing

Dublin City University

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

September 2017

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Engineering by Research 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.: 14211373

Date:

Acknowledgement

I would like to thank my supervisors, Prof. Liam P. Barry and Dr. Conor McArdle for giving me this opportunity to pursue my research under their tutelage and expert guidance. Both of them have been instrumental in shaping my research direction and I would like to express my sincere gratitude to them for their insightful advice, constant encouragement and unwavering support over the last two and half years, during the course of my research completion.

I gratefully acknowledge the funding received from the Daniel O'Hare Research Scholarship Scheme and I am indebted for being selected as a recipient of this generous scholarship.

My appreciation extends towards my colleagues in the Radio and Optical Communications Laboratory in DCU for all their help and support and I would specially like to thank Dr. Jingyan Wang for guiding me and helping me to settle down during my initial days.

All this would not have been possible without the constant support and inspiration of my family and I would specially like to thank my parents Soma and Asish and my little cousin Rounak for their patience, love and constant belief in me throughout the whole course of my studies.

Lastly, my special regards goes to my friend Lokesh for always being there with me even during the toughest of times.

Contents

List of Acronyms

AoD ..... Architecture on Demand

AQM ..... Active Queue Management

ASIC ..... Application Specific Integrated Circuit

AWG ..... Arrayed Waveguide Grating

B&S ..... Broadcast-and-Select

BDP ..... Bandwidth Delay Product

CAGR ..... Compound Annual Growth Rate

CAPEX ..... Capital Expenditure

CE ..... Congestion Encountered

CM ..... Central Module

COTS ..... Commercial of the shelf

CSN ..... Compute Storage Node

DARD ..... Distributed Adaptive Routing for Datacentres

DCN ..... Data Centre Network

DCTCP ..... Datacentre TCP

DEMUX ..... Demultiplexer

DLB ..... Distributed Loopback Buffer

E/O ..... Electro-Optic

ECMP ..... Equal Cost Multipath

ECN ..... Explicit Congestion Notification

EON ..... Elastic Optical Network

EPS ..... Electronic Packet Switch

FCT ..... Flow Completion Time

FDL ..... Fibre Delay Line

FIFO ..... First In First Out

FWC ..... Fixed Wavelength Converter

HOL ..... Head of the Line

IM ..... Input Module

IPP ..... Interrupted Poisson Process

IQ ..... Input Queued

LE ..... Label Extractor

MEMS ..... Micro-Electro-Mechanical System

MILP ..... Mixed Integer Linear Programming

MLB ..... Mixed Loopback Buffer

MSS ..... Maximum Segment Size

MTU ..... Maximum Transmission Unit

MUX ..... Multiplexer

O/E ..... Optic-Electro

O/E/O ..... Optical-Electrical-Optical

OCS ..... Optical Circuit Switching

OI ..... Optical Interface

OLE ..... Optical Label Extractor

OLT ..... Optical Link Terminator

OM ..... Output Module

ONU ..... Optical Network Unit

OPEX ..... Operational Expenditure

OPS ..... Optical Packet Switching

OQ ..... Output Queued

OXC ..... Optical Cross-Connect

PASTA ..... Poisson Arrivals See Time Averages

PON ..... Passive Optical Network

PreRes ..... Pre-Reservation

PUE ..... Power Usage Effectiveness

PWM ..... Passive Wavelength stripped Mapping

QoS ..... Quality-of-Service

QSFP ..... Quad Small Form-factor Pluggable

RAM ..... Random Access Memory

ROADM ..... Reconfigurable Optical add-drop Multiplexer

RTT ..... Round-Trip Time

SFP ..... Small Form-factor Pluggable

SLB ..... Shared Loopback Buffer

SOA ..... Semiconductor Optical Amplifier

SRAM ..... Static Random Access Memory

TCP ..... Transmission Control Protocol

TDMA ..... Time Division Multiple Access

ToR ..... Top-of-Rack

TWC ..... Tunable Wavelength Converter

VLAN ..... Virtual LAN

VLB ..... Valiant Load Balancing

VM ..... Virtual Machine

VOQ ..... Virtual Output Queue

WC ..... Wavelength Converter

WDM ..... Wavelength Division Multiplexing

WS ..... Wavelength Selector

WSS ..... Wavelength Selective Switch

List of Tables

List of Figures

Evaluation of Data Centre Networks and Future Directions

Sohini Basu

Abstract

Traffic forecasts predict a more than threefold increase in the global datacentre workload in coming years, caused by the increasing adoption of cloud and data-intensive applications. Consequently, there has been an unprecedented need for ultra-high throughput and minimal latency. Currently deployed hierarchical architectures using electronic packet switching technologies are costly and energy-inefficient. Very high capacity switches are required to satisfy the enormous bandwidth requirements of cloud datacentres and this limits the overall network scalability. With the maturity of photonic components, turning to optical switching in datacentres is a viable option to accommodate greater bandwidth and network flexibility while potentially minimising the latency, cost and power consumption.

Various DCN architectures have been proposed to date and this thesis includes a comparative analysis of such electronic and optical topologies to judge their suitability based on network performance parameters and cost/energy effectiveness, while identifying the challenges faced by recent DCN infrastructures. An analytical Layer 2 switching model is introduced that can alleviate the simulation scalability problem and evaluate the performance of the underlying DCN architecture. This model is also used to judge the variation in traffic arrival/offloading at the intermediate queueing stages and the findings are used to derive closed form expressions for traffic arrival rates and delay. The results from the simulated network demonstrate the impact of buffering and oversubscription and reveal the potential bottlenecks and network design tradeoffs. TCP traffic forms the bulk of current DCN workload and so the designed network is further modified to include TCP flows generated from a realistic traffic generator for assessing the impact of Layer 4 congestion control on the DCN performance with standard TCP and datacentre specific TCP protocols (DCTCP). Optical DCN architectures mostly concentrate on core-tier switching. However, substantial energy saving is possible by introducing optics in the edge tiers. Hence, a new approach to optical switching is introduced using Optical ToR switches which can offer better delay performance than commodity switches of similiar size, while having far less power dissipation. An all-optical topology has been further outlined for the efficient implementation of the optical switch meeting the future scalability demands.

Chapter 1

Introduction

In recent times, when computing and storage facilities are being transferred from desktops to the cloud, data centre networks (DCN) are gaining more and more prominence for extensive exchange of a huge volume of traffic between interconnected servers. The emergence of the world wide web in the 1990s led to the concept of DCNs when companies started to look for their permanent existence on the Internet. Gradually the number of websites started to grow from a few thousands to several millions [1] and data centres started to open up for hosting these websites and web-services. Since then the evolution have been constant and from hosting a few hundred servers, DCNs have now grown into massive cloud establishments capable of hosting hundreds of thousands of servers. Today, cloud services are generating an enormous demand for bandwidth which is almost doubling every year. With such rapidly increasing volumes of traffic, the number of servers in data centres is expanding at an exponential rate which is creating a series of changes in the DCN topology that interconnects the computing and storage nodes within the data centres.

Compared to traditional communication networks such as LAN or WAN, data centre architectural principles are quite different and based on the unique challenges and requirements of the sector for which the DCN is purpose built. Irrespective of the type of data centre there exists several challenges that can hinder it's successful architecture design such as (i) Scalability to meet the growing demand for storage and sufficient network capacity to support bandwidth-hungry cloud services which challenge network design in terms of interconnection and cost. (ii) Need for optimal congestion control, load management and scheduling techniques to overcome typical DCN traffic bottlenecks such as TCP Incast (iii) Dynamic traffic characteristics rising from a diverse range of DCN services and applications which hinders cost-effective routing mechanisms. (iv) The constant demand for service

from the servers hosted by a DCN that requires a high degree of robustness and agile failure recovery. (v) Energy efficiency while keeping the DCN power and CAPEX/OPEX budget under control.

Current DCNs are mostly deployed with a multi-tier hierarchical switching architecture that relies on commodity Ethernet switches for processing the data streams. The electronic packets are then transformed to optical form for facilitating Wavelength Division Multiplexing (WDM) across very high capacity point to point links that interconnect the network servers and switches. However, such architectures are not scalable enough to meet the demands of modern data centres because of their layered topology which causes a potential bottleneck at the upper network tiers due to the aggregation of traffic from the lower layers. This, in turn, creates an increased demand for bandwidth and network capacity across the oversubscribed network tiers. Additionally, the ever increasing traffic volume requires the DCN switches to be upgraded to very high data rates, but the opaque electronic switching is failing to deliver such flexibility and moreover, the capital/operational expenses coupled with the power budget is continuously on the rise in recent DCN infrastructures. So, the current scenario has shifted the attention of industry and academia towards enhancement of DCN infrastructure through designing flexible, scalable, greener and cost efficient architectures.

Compared to the electronic domain which uses only space and time dimensions for switching, optical domain introduces wavelength switching where enhanced WDM technologies enable tens of multiplexed signals to be carried over distinct wavelengths and provide a very large capacity. Consequently, optical switching has emerged as a potential solution for addressing the DCN performance requisites of scalability, network capacity, reduced latency and low footprint requirements since it can exploit the potential of WDM technology in terms of very high bandwidth, low latency, data rate transparent switching and low power consumption. Optical switching can be mainly categorised into Optical Circuit Switching (OCS) and Optical Packet Switching (OPS). OCS based switches use WDM technology so that every link can have several available wavelengths and every wavelength can be simultaneously used in different lightpaths that are not sharing any common link. These switches can generally provide sufficiently high bandwidth capacity with minimal power consumption but has reconfiguration speed in the order of milliseconds and so is not fit for handling the bursty DCN traffic flows. OPS which is the other alternative has a much finer switching granularity and provides on demand bandwidth with switching speed equivalent to their electronic counterparts. OPS technology shows promise of a very low latency and high bandwidth solution. However, the OPS bottleneck lies in contention resolution due to lack of optical Random Access Memory (RAM) options.

Even though, research has been focussed towards finding innovative optical switching solutions for increasing the bandwidth capacity while simultaneously reducing energy consumption, most of the recent architectures, concentrate on developing optical interconnects solely for the core/aggregation tier that deals with inter rack switching. Because of very heavy thermal dissipation in DCNs, the current acceptable rate of energy consumption is much lower than the growing rate of network capacity [2] and hence having energy efficient switching in every network tier is essential for sustainable development in large scale data centres. However, electronic switching is still used in the edge tier for interconnecting the servers within a rack and this consumes up to 90% of the total switching power in DCNs due to the presence of a huge number of ToR (Top of the Rack) switches [3].

Thesis Contribution

In this thesis, a threefold approach has been taken for a thorough evaluation of DCN infrastructure and standards that are currently prevalent and accordingly future research directions are suggested. Firstly, the data centre background history is covered with an extensive study of DCN characteristics and evolving architectures. An in depth survey of data centre topologies currently in use and those that have been proposed are then discussed. The rationale behind using optical solutions in data centres is evaluated and an extensive list of optical DCN architectures with varying underlying switching technologies are reviewed to judge the suitability of such solutions in the context of modern data centre performance requirements. The focus then shifts towards judging the performance of present day data centre environment with varying traffic pattern. (i) Layer 2 switching traffic is used for analytical modelling based on multi-stage tandem queue networks for identification of potential bottlenecks in traffic arrival/offloading and deriving relevant closed form delay expressions. Further performance evaluation helps to realize the impact of buffering and oversubscription and reveals the design tradeoffs in oversubscribed networks (ii) Variations of TCP traffic is used to assess the impact of Layer 4 congestion control on DCN performance and judge the suitability of the used protocols. The novelty of the designed network lies in the fact that it can use traffic flows generated internally by customised modules as well as have provisions for flows generated externally by a realistic traffic generator. Finally, in keeping with the reality that the future of DCN architecture perhaps lies in Optical Switching, an optical switch for the power hungry edge layer has been proposed with an outline of a highly scalable all-optical architecture for its implementation. Simulation results revealed that

the optical solution can have better delay performance and improved power efficiency compared to it's electronic switching counterpart.

Layout of the Thesis

The remainder of the thesis has been structured as follows:

- In Chapter 2, the dominant trends of recent data centres are identified and how the current data centre infrastructures have evolved to cope with the current demands are then described. Over the years a variety of DCN architectures have been proposed and the chapter provides an extensive overview of such topologies and architectures, highlighting their pros and cons. The impending challenges that are still faced by recent data centres are also discussed.
- An overview of optical switching paradigms is provided in Chapter 3, along with a background for the rise of optical interconnects in DCNs. The current optical data centre proposals are categorised and compared based on their underlying switching technology, to show how the architectures are evolving to tackle the diverse DCN traffic and performance requirements.
- In Chapter 4, an electronic Leaf-Spine network is evaluated, based on different network configurations in order to judge the performance variations arising from various combinations of buffer placement and varying oversubscription parameters. An analytical modelling is also done for Layer 2 switching to understand the difference in traffic loading in different switching stages and identify the potential bottlenecks.
- The challenges and impact of Layer 4 switching in data centres are studied in Chapter 5, as congestion control is an important aspect of DCN traffic management. Since TCP is the most common transport layer protocol used in data centres, a Leaf-Spine network was designed to evaluate the performance of standard TCP compared to a DCN specific TCP variant. The designed network mostly used TCP traffic flows generated externally by a realistic traffic generator (DCT^{2}Gen) that is based on distributions generated from data centre traffic traces.
- In Chapter 6, an optical interconnect architecture is presented, which utilises the cyclic routing capability of AWGRs to develop a fast optical ToR (OToR) switch that minimises power consumption by using passive components (like AWGR) and effectively handles the intense intra-rack traffic with proper contention resolution through FDLs. An all optical architecture is

also proposed where the OToR can be utilised for efficient DCN communication.

- The overall observations of the thesis are summed up in Chapter 7 and some relevant topics and areas where future research direction can be focused upon are pointed out.

Part I

Background

Chapter 2

Data Centre Networks - Trends, Standards and Challenges

Traffic within data centres is growing at an exponential rate due to the rapid emergence of cloud computing services (Microsoft Azure, Amazon EC2, Chrome, Google App Engine etc.), social networking (Facebook, Twitter), video streaming (YouTube, NetFlix), web search (Bing, Google), data intensive applications (like GFS, Map Reduce, Hadoop) and High Performance Computing [4]. Data centre traffic is estimated to be 1 million times more than that of external IP traffic and so for every byte of data sent across the Internet at least 1 MB of data gets transferred within a data centre [5]. This rise in CaaS (Computing as a Service) based cloud applications and server-side computing needs very large scale storage infrastructures with high performance guarantees [6]. As a result, many data centre networks (DCNs) are evolving from enterprise networks to ware house scale mega data centres comprising of 100,000 or more servers to keep up with the massive traffic growth [7].

2.1 Current Data Centre Trends

2.1.1 Traffic Growth Pattern

Data centre traffic will continue to dominate Internet traffic for the foreseeable future, but the nature of data centre traffic is undergoing a fundamental transformation. Ciscos Global Cloud index report forecasts that global data centre traffic will be more than tripled from a mere 4.7 ZB in 2015 as shown in Figure 2.1 to reach 15.3 ZB (zettabytes = 10^{21} bytes) annually by the end of 2020 , representing a 27% CAGR [8] and this enormous amount of traffic can be actually equivalent to the traffic generated

by 213 trillion hours or 38 months of continuous music streaming for every individual of the world's population in 2020 [9]. Additionally, global cloud data centre traffic, the fastest-growing component of data centre traffic is projected to have a 30% CAGR and by 2020, 92% of the traffic will be handled by cloud data centres while only 8% will be handled by traditional data centres [8, 9].

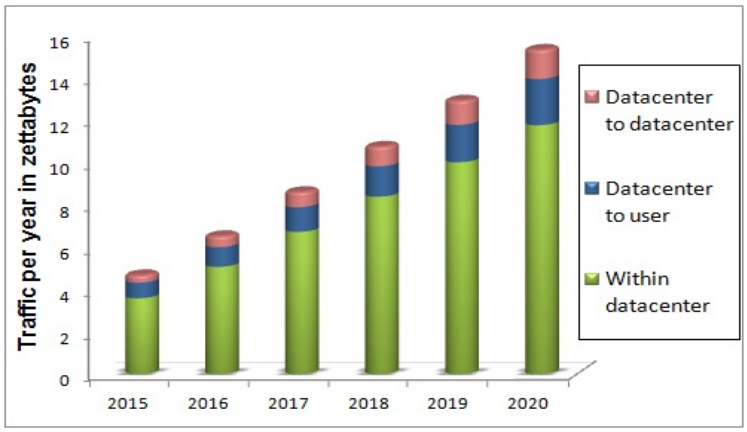

Figure 2.1: Net Global Datacentre Traffic [8]

Figure 2.2 shows that an average of 77% traffic from 2015 through 2020 is projected to remain inside the data centres [8,9]. This represents the traffic between racks/clusters contributed by authentication, back-end queries and storage databases, which causes the applications to scan through all servers generating replication, read/write and backup per user query [9]. However inter DCN traffic though much smaller in proportion than intra DCN traffic, is growing at a faster rate and by 2020, it will account for almost 9% of the total DCN traffic, up from 7% in 2015 [8]. This increase can be attributed to the growth of locally distributed data centres pertaining to dedicated services, data replication, big data analysis, inter-database searching and the need for exchanging data between clouds [9].

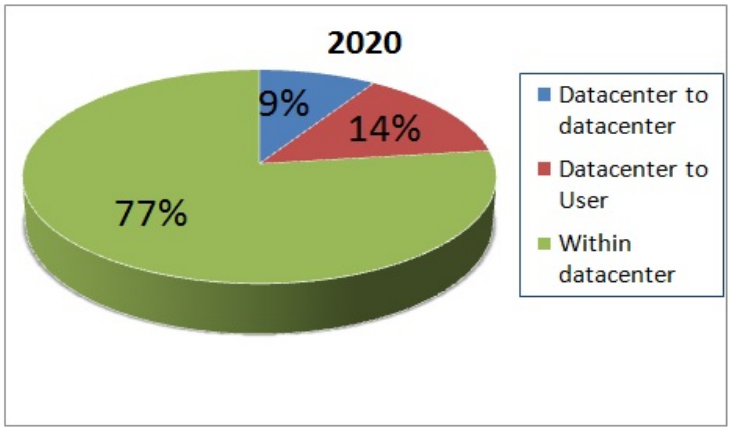

Figure 2.2: Global Datacentre Traffic by Destination [8]

2.1.2 Traffic Characteristics

DCN traffic can be broadly divided into two categories namely; North-South traffic which corresponds to the communication between data centre and the end users (such as Video streaming to a mobile device or PC, Web search etc.) and East-West traffic which refers to the communication within the data centre (such as Big Data/MapReduce applications which requires coordination among its VMs) [10]. As per recent traffic growth patterns traditional north and southbound traffic is being replaced by east-west or horizontally oriented traffic patterns [11] across virtualized clouds of data and currently, more than 75% of the DCN traffic is east-west in nature [8, 9]. Major factors [12–14] contributing to the rapid rise in east-west/intra data centre traffic are as follows:

- Convergence which occurs in storage traffic often sharing the same physical network as the application traffic. Storage traffic occurs between hosts and arrays that are in the same network segment, logically right next to each other.
- Growing tendency to virtualize physical hosts into virtual machines (VM); the ability to move workloads easily, has become a mainstream, normative function. VMs moving from physical host to physical host within a network segment leads to a rise in intra DCN traffic.
- Parallel processing divides tasks and sends them to multiple servers, contributing to the increase in internal DCN traffic.

Inherent traffic patterns within DCNs differs from that of the other communication networks and so it is crucial to understand the basic characteristics to analyse the performance of any proposed data centre architecture. The information obtained from traffic engineering studies such as [15–19] has been aggregated to characterise cloud DCN traffic as mentioned in the following subsection:

2.1.2.1 Rack Awareness

Most of the traffic generated by servers (average of 75% - 80%) stays within the rack in a cloud DCN [15, 16]. There is only 11% probability of intra rack transfer and 0.5% probability of inter rack transfer between server pairs [10]. Due to the rack-awareness feature, modern DCNs running applications such as Hadoop, MapReduce etc. try to keep the frequently communicating servers within the same rack to achieve higher throughput and lower latency. Servers tend to communicate with either almost all other servers inside a rack or with less than 25% of the other servers in the same

rack. In case of inter rack traffic, servers tend to communicate with less than 10% of servers in other racks or do not communicate with any server outside its rack. [16].

2.1.2.2 Flow Distribution

Traffic flows within DCNs can be of two types: (i) Long lived, throughput bound communications or elephant flows that are large in volume and can be generated by data backup, virtual machine migrations, MapReduce functions etc. (ii) Latency sensitive communications or mice flows that are bursty by nature and can be generated by transactional traffic, web browsing or database search queries [20, 21]. Mice flow dominate data centres and constitute around 90% of the total number of flows. They are less than 10 KB in volume and last for only a few hundreds of milliseconds [10, 19]. However, elephant flows even though far less common in DCNs accounting for only 10% of the total number of flows, carry more than 80% of the traffic volume [10, 15]. They last for several seconds transferring tens of megabytes across the DCN causing the mice flows to suffer from undue latency. Highly persistent TCP flows tend to fill up network buffers which add significant queuing delay to the mouse flows that share these buffers and this leads to long flow completion time (FCT). Since TCP traffic in DCNs is very dynamic, at any time there may exist around 10,000 active flows per second within a rack. Hence for a large scale network of 100,000 servers, it is possible to have 25,000,000 or more active flows [10] at any point in time.

2.1.2.3 Hotspot Formation

The unbalanced and skewed pattern of bursty DCN traffic causes congestion and formation of localised hot-spots across particular links. Hotspots are actually switches with high buffer occupancy rates, where the arrival of extra packets causes the buffer to overflow and affect the DCN performance by causing packet loss. Due to the presence of TCP flow control the lost packets are retransmitted which in turn makes the switches more unstable. Heavy elephant flows potentially create hot nodes with transient congestion. Studies suggest that hot-spots are usually formed across links having a utilisation of 70% or more but overall less than 25% of the links experience hot-spot traffic [10,16].

2.1.3 Flattening of Networks

With the rapid rise in cloud computing applications, multi-tiered DCN topologies can no longer sustain the dense horizontally oriented (east-west) traffic which requires collaboration between servers

for handling a single request with several compute and store stages [22]. In a legacy/enterprise data centre up to 50% of the switch ports can be used just for interconnecting network tiers rather than connecting servers [23]. East-west traffic between servers needs to move vertically and switches at each level need to do multiple iterations of packet processing. As a result the traffic can get unnecessarily delayed. Removing the upper switching tiers and directly connecting the servers can reduce network diameter and create a low latency and non-blocking Clos type fabric [24] which can be vital for time sensitive system applications such as web search, financial trading, video streaming etc.

A flattened DCN fabric represents a network of servers and interconnected switches with tightly woven connections between all nodes [22, 24]. It can create one-to-one connections between servers allowing them to interconnect with a fixed latency by traversing a set number of switches and can subsequently ensure virtual machine migration, simplified traffic flow, improved full cross sectional bandwidth, reduced management overheads, power and cost savings and ability to assign any application to any server. Due to the need for reducing switching tiers, the most common DCN topology used today are the two-layered folded Clos networks or the Leaf-Spine network [25]. Very high radix spine or tier 2 switches are required to scale up even though it allows load balancing across several equal cost paths. Such switches with very high radix are far more expensive than regular top-of-rack (ToR) switches and can consume more energy [26]. So, current DCN research is being focused towards finding alternative flattened networks that can connect the ToR switches directly.

2.1.4 Hyperscale Datacentre

Today	Hyperscale Datacenter Trends and Figures	2020
21%	Share of global data center servers	47%
39%	Share of global data center processing power	68%
49%	Share of global stored data in data centers	57%
34%	Share of global data center traffic	53%

Figure 2.3: Scale of Hyperscale Datacentres [8]

Growing demand for cloud computing applications and their increasing complexity have led to the rise in the number of Hyperscale data centres which are designed to provide a massively scalable public cloud DCN facility with custom hardware, software, and architecture [27]. They are typically made up of several thousands of individual nodes for providing networking, compute and storage facilities.

Cisco has predefined the eligible criteria for cloud operators to be classified as hyperscale data centres and based on that 24 operators have been identified currently [8] such as Amazon, Facebook, Google, Microsoft, Apple, eBay, Alibaba, Yahoo etc. Due to the rapid demand for consumer and business oriented cloud services and big data, such massive data centres are going to rise to 485 in number by 2020 from 259 in 2015, representing 47% of all data centre servers globally by 2020 (as shown in Figure 2.3). This shows the tremendous growth of hyperscale DCNs in the Public Cloud sector since almost 83% of all public cloud servers will be installed in hyperscale data centres [8].

2.1.5 Changing Data Rates

Generally, the traffic within data centres is several times more than the output volume and a very large amount of traffic is constantly being transferred between several thousand nodes that may cause congestion and bottlenecks. Several applications such as online transactions (Financial Trading, Web search etc.) and cloud services require very fast connectivity to avoid untimely delays in getting updates and streaming. Additionally, modern DCN require minimal hardware with very high performance capability to reduce complexity. All these factors constantly lead to the rising demand for data rate upgradations and increasing bandwidth capacity to minimise latency and operational costs.

Currently, the observed trend in data centres is to move towards 25G/100G from 10G/40G [28]. Intra rack connections have evolved from 1G to 10G and are moving towards 25G and 50G in near future. Whereas inter rack communication has evolved from 10G to 40G and is soon moving towards 100G and beyond. Inter DCN connections have already started using 100G and 400G is being standardised now [29]. According to statistics in [30, 31] at least 50% of all DCN switch shipments will be either 25G or 100G by 2020. In order to keep up with this growing demand, several hyperscale operators have already introduced 100G technology in 2016 and replaced their off-the-shelf switches with custom built 100G port switches such as the 100G Pigeon switch introduced by LinkedIn [32] or the 100G Wedge 100 [33] and Backpack switches [34] introduced by Facebook.

2.2 Traffic Routing

Routing protocols used in enterprise networks cannot be used in data centres due to their specific architecture and traffic pattern. DCNs require multiple paths between communicating hosts to scale

horizontally while handling the arbitrary traffic through path diversity. Communication within DCNs covers shorter distance links and cannot spread out widely. Moreover, Cloud based applications also demand very high intracluster bandwidth requirements [35]. Since it may not be possible to accommodate all components of a particular application within same rack/cluster, high bisection bandwidth will be required to prevent hotspot formation [36]. So, data centres need highly optimised routing and load balancing techniques for providing the required bandwidth, avoid path complexity and congestion while still maintaining the desired level of network performance and latency requirements.

For full utilisation of network resources, modern DCNs makes use of commodity switches to connect multiple equal cost paths in the tree based topologies (such as fat-tree, leaf-spine) where a single flow might pass through multiple paths to avoid congestion. However, standard transport protocols such as TCP lack the ability to select routes based on network load. While fixed DCN architectures like FatTree [37], BCube [38] etc employ a single routing strategy, flexible architectures such as CamCube [39] uses a two-stage routing where a link state routing protocol can take care of the multipath diversity while the servers can also reroute packets as and when required. In order to maximise throughput and resource utilisation, DCN networks either make use of a Centralised Dynamic Routing or a Distributed Traffic-Oblivious Routing approach. But the two approaches are not entirely exclusive since many distributed approaches will gather localised information from end nodes in a particular area to implement a centralised routing strategy for that area whereas there may exist distributed location discovery mechanisms in centralised protocols such as PortLand [40].

2.2.1 Centralised Routing

This type of routing scheme employs a central scheduler that is aware of every link-state for distributing the elephant flows dynamically across multiple paths. Hedera [36] is the most common centralised routing scheme which uses edge switches to detect elephant flows and subsequently notifies a central controller so that it can reroute the traffic through a new path and dynamically update the switch routing tables. Since the controller has a global view of the entire network and its corresponding links so it can adaptively control congestion and formation of bottlenecks. Even though this type of routing is beneficial in hierarchical networks for utilising the full bisection bandwidth, the centralised controller can in turn limit scalability and be a single point of failure that can disrupt the whole network. On widening the size of the network, more and more traffic messages being sent to and from the cen-

tral controller tends to congest the connecting links and hence parallelism [15] may be required for handling bursty traffic patterns and enhancing route computations.

Advantages of this routing approach [41] can be summarised as below:

- Ensures better use of system hardware since it has a complete view of the whole system which leads to optimal routing decisions.
- Fault handling is more efficient because the centralised controller can gather information about updates and system failures from the whole data centre,
- As centralised routing protocols are deployed on selected nodes, so reconfiguration becomes easier while changing protocols.

2.2.2 Distributed Routing

This routing approach is adaptive in nature and allows end hosts to route traffic on their own from congested paths to the less traffic loaded paths without any centralised controller. Distributed routing approach can scale efficiently and can be used for the modern DCNs, but the lack of knowledge of workloads in each path can create hotspots. ECMP (Equal Cost Multi-Path routing) [42] and VLB (Valiant Load Balancing routing) [19, 43] are examples of traffic oblivious load balancing approach; a variation of Distributed Routing. ECMP is employed at the flow level by switches and uses simple path assignment technique that uses more than one path for each connection and divides traffic across the multiple paths based on hash values of each path without requiring packet reordering at the final destination. VLB routing uses path selection at the edge switches for randomly selecting a core switch to forward the traffic to the required destination node. This approach tries to utilise the full bisection bandwidth under all traffic conditions and is well suited for hierarchical data centres due to the ease of scalability. Both these approaches assign traffic paths regardless of the workload and utilisation, which can create elephant flows that might collide and lead to congestion.

A more optimised routing approach is used in DARD (Distributed Adaptive Routing for Data centre networks) [35] which differs from ECMP and VLB in two major aspects. Firstly the traffic path is selected dynamically depending on the workload so that very large flows can be detected and distributed accordingly. Secondly, instead of being implemented only at the switch level like its peers, it can be deployed in server-centric DCN topologies where servers can monitor the traffic paths and equally distribute the workload across all available paths.

Advantages of this routing [41] approach can be summarised as below:

- As the routing information is distributed across all servers, these protocols can operate for the entire duration when data centre remains connected.
- Since routing decision is based on the information obtained from the local nodes; substantial time is saved for gathering information about the entire data centre.
- It creates less overhead since the servers do not need to exchange routing related information with a large number of nodes.

2.3 Data Centre Architecture Topologies

Data centres can house several thousands of servers that can dynamically share resources such as memory, processors and I/O devices in order to support data intensive communications. The servers are arranged into racks with the help of Top of Rack (ToR) switches. Racks of servers can be further grouped together to form pods or clusters depending on network requirements. ToR switches normally connect the server racks using higher layer COTS (commercial of the shelf) switches. Based on the underlying topology which defines how the different network devices (routers, servers and switches) are interconnected, DCNs can be classified into different types and their performance heavily depends on the type of networking infrastructure being used.

2.3.1 Commercially Used Architectures

Commercially used DCN topologies are generally based on tree-type hierarchical networks with high-density hardware. While a three tier tree network was the choice of topology in the past, DCN operators have now started using a relatively flatter, folded-clos topology for better network performance. Additionally, hyperscale data centres are using specially customised architectures particularly suited to their network characteristics and traffic pattern which is usually more dense than the enterprise data centres, for example, Facebook DCN using the fabric architecture [44].

2.3.1.1 Multi-Tier Tree

Traditionally used DCN models use a tree type topology where switches are arranged in multiple layers with varying link capacity. The Access layer is formed by ToR switches which provide server-

to-server interconnections. Aggregation layer switches collect the traffic from the ToRs and forward them to the core switches at the top layer for inter-rack communication. High-capacity core switches interconnect the switches in the aggregation layer and form the backbone of the network. Redundancy is provided at core and aggregation layers to enhance bandwidth capacity and network reliability. Each ToR switch is connected to multiple aggregate switches which in turn are linked with multiple core switches. This over-provisioning of resources aims at enhancing network reliability and bandwidth capacity [37]. While the aggregation layer handles the job of load balancing, the core layer maintains the flow of traffic in the north-south direction by interconnecting the lower layers.

Multi-tiered tree networks (see Figure 2.4) tend to have high oversubscription ratios where servers within the same rack have significantly more provisioned bandwidth between one another than with servers in other racks [45]. Heavy utilisation of the upper layer links imposes added stress on the core switches and introduces buffer congestion affecting the overall latency. Due to lack of sufficient redundancy, the core switches are prone to failure which leads to the loss of connectivity among the ToR switches and congestion across certain links whereas other paths may be free. Scalability is limited by the core switch radix which makes it difficult to accommodate very large number of servers. Cost and power consumption figures rise at a significant rate with increased demand for the high-performance processing capacity of the core network.

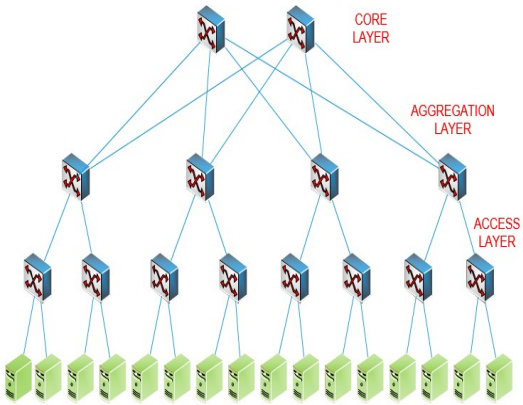

Figure 2.4: Multi Tier Tree Topology

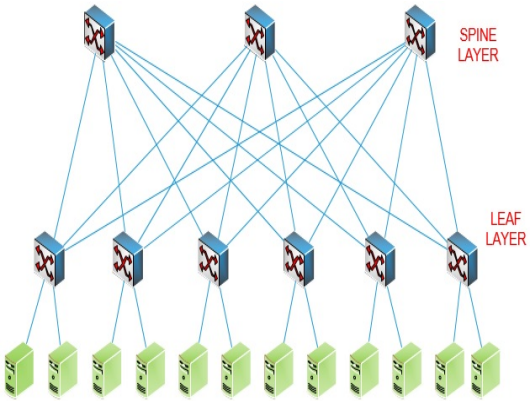

Figure 2.5: Leaf Spine Topology

2.3.1.2 Leaf-Spine Architecture

Leaf-Spine is a fully connected multipath topology wherein every node is equidistant from all its neighbours [25]. This folded clos network (see Figure 2.5) is currently the de-facto standard for DCN

topology amongst commercial vendors due to its predictable performance, ease of scaling and high bandwidth capacity. The Leaf switches connect the racks of servers and are fully meshed to a series of non-blocking Spine switches at the upper layer to form any-to-any connectivity between all servers. Leaf-Spine networks offers better network performance and lower oversubscription ratio compared to the previous multi-tier networks.

Since all nodes are equidistant from each other, applications running across the network behave in a predictable manner. The presence of multipath redundancy ensures network reliability and fault tolerance. It allows flexible incremental expansion up to a certain extent since new server racks can be introduced by using additional Leaf switches. However, further scalability can be restricted by the limit on the Spine radix and CAPEX/OPEX costs will also increase with the growing size of Spine switches. Leaf-Spine has a high wiring complexity due to the fully connected nature of the topology which tends to increase the number of interconnections with an increase in Spine radix.

2.3.1.3 New Hyperscale Architectures

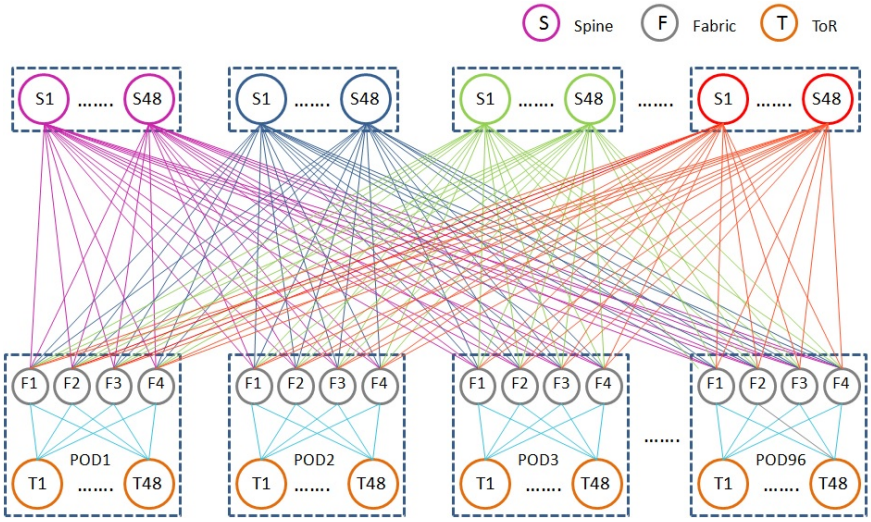

Figure 2.6: Facebook Fabric Datacentre Topology

Using the traditional tree-based architectures, hyperscale cloud operators are finding it hard to keep up with the growing demand for network bandwidth capacity and hence are deploying customised architectures enabling rapid network deployment, performance scalability and very high bisection bandwidth required for agile cloud applications. One such prominent example is Facebook who has recently moved towards a modular, robust and scalable fabric-based architecture [44], see Figure 2.6.

This modular network is divided into smaller units called pods, containing 48 ToR switches and 4 Fabric switches. For ensuring very wide connectivity, the network is divided into four independent Spine planes, with up to 48 independent devices in each plane. Each of the 48 ToR switches use 4X40Gbps links to connect to the four fabric switches which are in turn connected to 48 Spine switches in each Spine plane via 48X40Gbps links with 1:1 oversubscription. Thus the 4 fabric switches in each pod are linked with the 4 different Spine planes at the top layer.

With this modular approach, the top layer no longer needs very high radix Spine switches which overcomes the switch size and capacity limitations. So, network capacity, incremental scalability and network reliability is improved considerably. By using smaller switches the CAPEX/OPEX costs are also lowered. However, such an intricate design leads to high routing and cabling complexity.

2.3.2 Taxonomy of Electronic Data Centre Architectures

According to Dally et al. [47], interconnected networks can be classified as direct network where the backbone network directly connects with the node endpoints or indirect network where the node endpoints only connect to the edge of the network devices. In DCN terms, in a direct topology the routers will be directly connected to the hosts, e.g. the Recursive architectures. Whereas in an indirect topology the routers may not have a direct connection with the hosts, e.g. the Clos type architectures [48]. Popa et al. [49] further grouped the DCN topologies from the hardware perspective, as Switch centric or Server Centric or Hybrid in nature. Switch centric are those where routing is handled solely by the switches such as the Clos-based networks. Server Centric are those where routing is explicitly handled by the servers and they forward packets as well as run regular applications, e.g. the 3D torus topologies like Camcube where servers are directly connected to each other through server to server links. Lastly, Hybrid networks are the ones where both switches and servers are used for forwarding packets such as the recursive networks having servers connected across different layers.

According to Yang Liu et al. [41] data centres can be classified as either fixed or flexible networks based on the reconfigurability of the deployed architecture. If the deployed network supports incremental growth without disturbing the current network or replacing existing switches then that particular topology can be considered to be flexible such as the Random architectures. Otherwise, if the scaling up process is not configurable and networks require systematic upgradations due to their deterministic structure then the topology can be termed as fixed such as the Clos-based architectures.

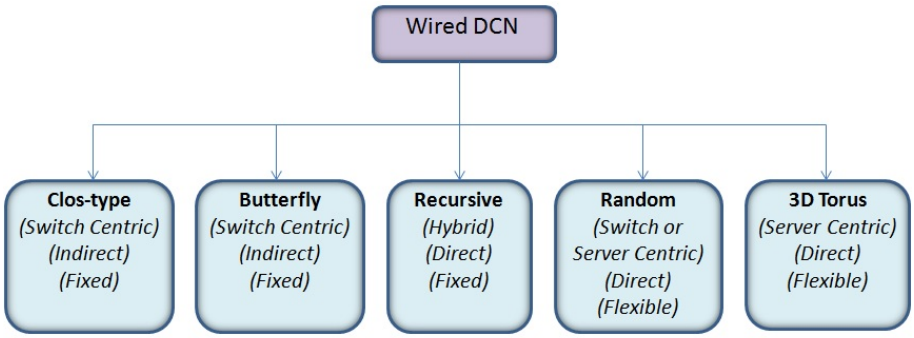

Figure 2.7: Classification of DCN Topology

Over the years a series of architectures have been proposed for DCNs, each trying to improve upon the shortcomings of its prior solutions. Based on their physical layout and network characteristics DCN topologies can be divided into five main groups as shown in Figure 2.7.

2.3.2.1 Clos based Architectures

The inefficiency of traditional multi-tiered networks generated new interest in the Clos [50] networks (initially proposed for telecom networks in 1953) in the last decade. This type of topology consists of a non-blocking multistage architecture where each core switch connects to all switches in the lower layer in order to reduce the network diameter and latency by minimising the number of intersecting nodes. Fat-tree, VL2, Elastic Tree, Portland are some of the typical Clos-based architectures.

2.3.2.1.1 Fat Tree

To relieve the core switching layer from the bottleneck of oversubscription and high failure rate, Al-Fares et al. [37] introduced the fat-tree networks (as shown in Figure 2.8) based on a folded-Clos topology. Similar to the multi-tier tree network, fat-tree also has three layers but is composed of similar switches and links all throughout, resulting in a homogeneous topology with perfectly balanced oversubscription ratio. The ToR and aggregation switches are arranged in blocks called pods which are interconnected through the Core layer.

Building the entire network using similar switches improves the cost efficiency compared to multi-tier networks. Fat-tree network is rearrangeably non-blocking and the presence of an equal number of links in all layers relieves the bandwidth bottleneck. All devices can transmit at line speed if packets are distributed uniformly along the available paths and the network can operate at its full bisection

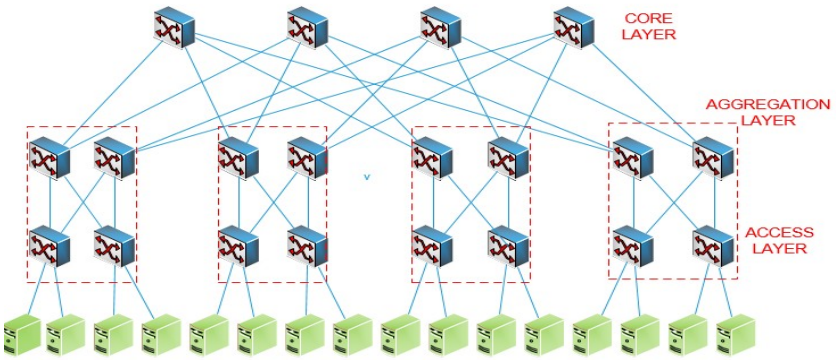

Figure 2.8: Fat Tree Topology

bandwidth because of the same number of uplinks and downlinks in each pod. However, scalability is restricted by the switch radix and it cannot scale out to a massive size with COTS switches. Reliability can be affected by the lower layer switch failures [5] which in turn hampers the overall performance. Wiring complexity of O(n^3) is also substantially high for designing large networks.

2.3.2.1.2 VL2

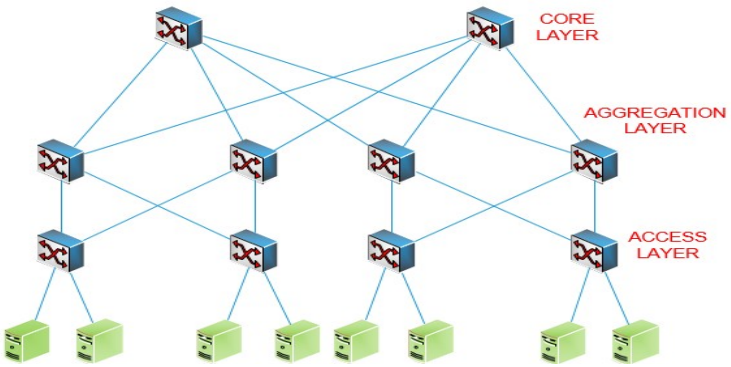

Figure 2.9: VL2 Topology

Similar to the fat-tree topology, Greenberg et al. [19] proposed the VL2 topology (as shown in Figure 2.9) to design a DCN with dynamic resource allocation. It uses a folded-Clos topology to introduce redundancy and load balancing is achieved through VLB routing technique which was introduced by Valiant [51] to handle the volatile traffic mix in interconnection networks. VLB routing can result in a hot spot free DCN by segregating traffic flows uniformly into several paths. Flat addressing is used to eliminate fragmentation of resources and ensure dynamic resource allocation across all servers. It is possible to increase cost efficiency by increasing oversubscription ratio and minimising VM cost. However, scalability is restricted by switch radix and it may not be suitable for accommodating cloud

DCNs. Similar to fat-tree, it also has a high wiring complexity of the order of O(n^3). Priority based traffic handling is not possible in VL2 networks since the VLB approach treats all flows uniformly. So bursty or latency sensitive applications might suffer.

2.3.2.1.3 Elastic Tree

Heller et al. [52] proposed the Elastic tree architecture (see Figure 2.10) to tackle the randomness and volatile nature of DCN traffic and improve the energy efficiency in data centres. All server racks need not be fully connected at all times and so depending on the traffic pattern this network dynamically adjusts the number of active links and switches. The basic model consists of an optimizer which calculates the subnet with minimum power requirement while maintaining system performance and fault tolerance, the power control module which controls the power state of all network elements and the routing module that schedules the routing of packets in the optimised network.

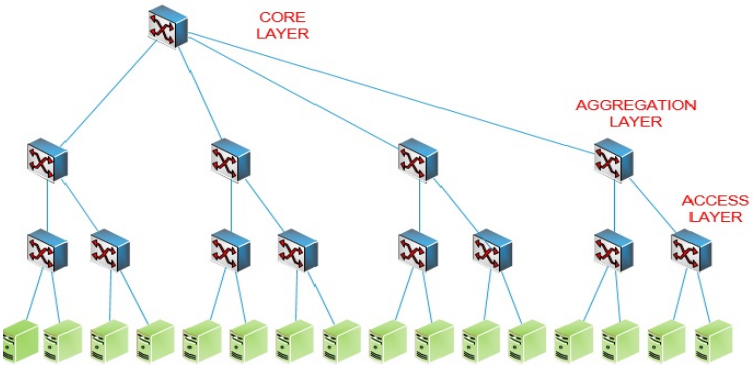

Figure 2.10: Elastic Tree Topology

Substantial amount of energy can be saved in this network by turning off switching elements as per traffic demand. As mentioned in [52] up to 50% of energy savings is possible and the expense on cooling of network devices is also proportionately minimised. This subsequently results in lowering maintenance costs as well. But since this network is based on the fat tree model it inherently faces similar limitations in terms of scalability and network reliability.

2.3.2.2 Butterfly based Architectures

This topology was proposed in 1982 by BBN technologies [53] to solve the issues with parallel computing devices. A k-ary n-fly network has k terminal ports per router and n intermediate stages in between. Butterfly networks try to combine connecting paths between multi-stage networks in

order to achieve path diversity and improve the performance and fault tolerance. Flattened Butterfly (FBFly) and Dragonfly are some of the most prominent butterfly based DCN topologies.

2.3.2.2.1 Flattened Butterfly (FBFLY)

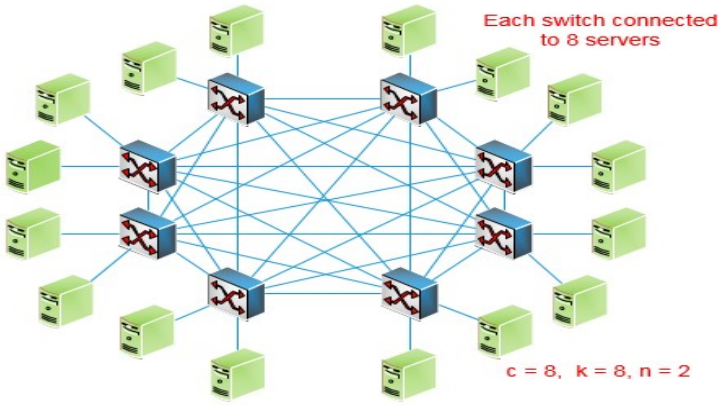

Figure 2.11: Flattened Butterfly Topology

John Kim et al. [54] proposed this topology for creating a flat network using high radix switches to improve network capacity and cost efficiency. The multistage routers in the original butterfly network are combined into a single router in each row to flatten the network and introduce path diversity. In a traditional butterfly network, a single path exists between any two nodes. But FBFLY (see Figure 2.11) introduces path diversity to improve system reliability and performance [54]. The flattened nature of the network creates a small network diameter which minimises latency and improves network throughput. The maximum distance between any two nodes is usually two hops and the network cost is lesser than an equivalent sized Clos based networks due to fewer routers being used. Butterfly network was originally introduced for on-chip interconnection networks. So, scalability can be restricted in case of cloud DCNs due to the high switch radix requirement. As the physical channel count per dimension increases quadratically with the increase in the number of nodes, wiring complexity also rises and longer global wires are required for interconnection of routers. The high degree of dedicated point to point links being used, further affects the overall link utilisation.

2.3.2.3 Recursive Architectures

The bandwidth bottleneck and single point of failure constraint in hierarchical tree based architectures can be overcome by using recursive topologies which provide a scalable network capacity by means

of a well-connected interconnection network, where a single unit is replicated several times to scale out to a very large capacity, spread out across multiple levels. Recursive architectures are hybrid [49] in nature where the switching functions are handled by both servers and routers. DCell, FiConn, BCube, MDCube etc. are some of the major recursive architectures proposed to date.

2.3.2.3.1 DCell

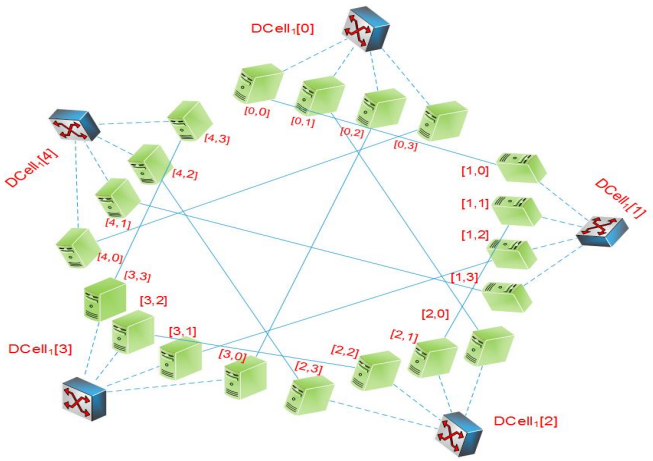

Figure 2.12: DCell Topology

DCell, proposed by Guo et al. [55] is a homogeneous topology which uses similar small radix commodity switches and links across all levels and recursively builds up the network from a basic unit called DCell_0 consisting of n servers connected to a mini-switch with n ports. DCell_{k+1} is built out of m+1 DCell_k units where m is the number of the server in DCell_k and k is the dimension of the network. It can scale up doubly exponentially with the rise in node degree and hence can support a large number of servers with a fairly small number of switches. As shown in Figure 2.12, each server connects to a mini-switch and another server in each level to provide a high network capacity and large bisection bandwidth for supporting bandwidth consuming applications and the presence of multipath ensures fault tolerance of network links. A distributed fault-tolerant routing algorithm used for packet forwarding can transfer packets even in case of severe link or node failures.

Most of the switching operations are handled by the lower level servers and hence load balancing can be critical in case of heavy all to all communications. Since the network is spread out across several layers, so the servers need multiple network ports to scale up and it can be costly to install multiple NICs per server. Due to the recursive structure, the network diameter keeps on increasing with the

addition of every new layer which causes the network latency to rise. The number of interconnections and wiring complexity gradually increases due to the massive link redundancy. Even though the network can offer good scalability, it may not be a practical solution for massive cloud data centres requiring incremental expansion.

2.3.2.3.2 FiConn

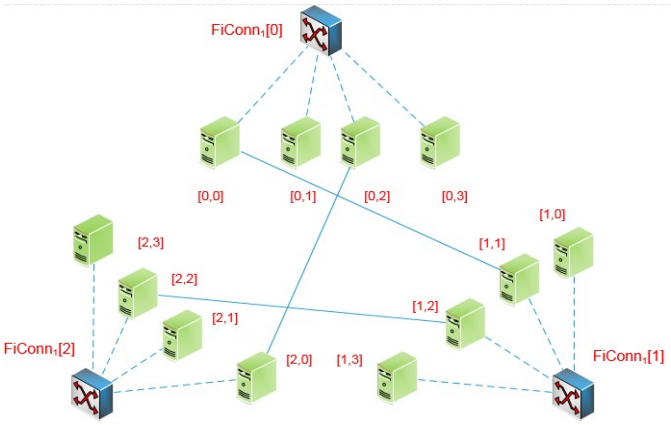

Figure 2.13: FiConn Topology

In order to simplify the need for having servers with multiple network ports in DCell, Li et al. [56] proposed a simpler network which makes use of only two network ports per server, one for packet forwarding and reception and the other for redundancy. Unlike DCell, FiConn (see Figure 2.13) does not create fully connected paths among all virtual nodes at the same level and only the idle ports are used for connecting with other devices. Hence network overload due to link redundancy is reduced by quite some extent at the cost of performance. By using two NICs per server, extra hardware related cost is minimised and wiring complexity is reduced by decreasing the number of interconnecting links and server ports. Traffic aware routing in FiConn can efficiently use available link capacity according to the traffic level ensuring a balanced link utilisation at all levels. However, the reduction in path redundancy lowers the network capacity to some extent and due to the absence of enough redundant links between connecting servers, the average path length becomes quite long.

2.3.2.3.3 BCube

The BCube model based on hypercube [57] topology was proposed by Guo et al. [38], for modular data centres. Being recursive and homogeneous in nature, BCube (see Figure 2.14) uses similar links and switches at all levels and builds the network from the basic unit called $BCube_0$ consisting of n

servers connected to a n-port commodity mini-switch. BCube$_k$ is constructed from n BCube$_{k-1}$ units connecting to n^k n-port switches. Unlike DCell, the servers connect to a separate switch in each layer and are not connected with other servers. Every server has k+1 parallel paths of varying length.

BCube can support bandwidth intensive applications by speeding up one to one, one to many and one to all traffic patterns depending on the number of available network adapters. It demonstrates efficient network capacity for all to all traffic patterns and multiple server ports are available to ensure selective routing and high fault tolerance. Unlike DCell, it supports effective load balancing in the low-level servers. The number of switches required is approximately k times that of DCell to support a certain number of servers. So, BCube cannot be cost effective and scalable enough for cloud DCNS. Due to multiple parallel paths available at each node, the wiring complexity is relatively high. Similar to DCell, servers need multiple NICs (typically 4 to 5) to scale up which adds up to CAPEX budget. Because of the recursive structure, the network diameter keeps on increasing with the addition of every new layer and it in turn increases the network latency.

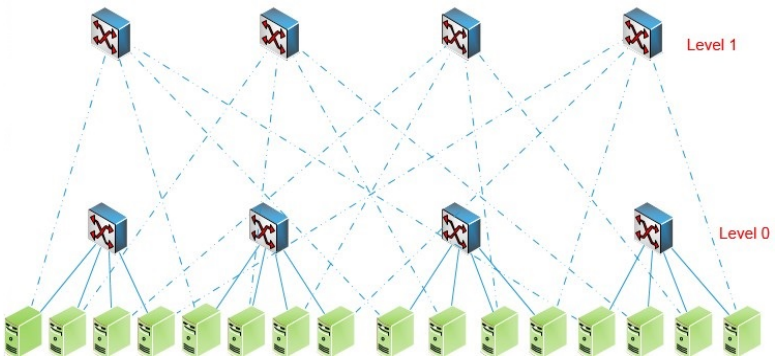

Figure 2.14: BCube Topology

2.3.2.4 Random Architectures

A randomly connected topology consists of servers and switches connected arbitrarily in no particular order. Hierarchical network models suffer from scalability and bandwidth bottlenecks but random architectures can offer greater flexibility in terms of scaling, by moving away from a symmetric and homogeneous design. Scafida and Jellyfish are prominent examples of random architectures.

2.3.2.4.1 Scafida

Gyarmati and Trinh [58] proposed Scafida based on the scale-free network generation algorithm of Barabsi and Albert [59] where the topology puts a constraint on the degree of a node in order to meet

switch radix limitations. Scafida networks (see Figure 2.15) exhibit randomness in terms of scaling up and can be expanded to any size to ensure flexibility in network design. Average path length is less than comparable symmetric networks like fat-tree leading to greater capacity and lower latency. Due to the scale-free nature of the network, Scafida can provide a high degree of fault tolerance. As a result of the random nature, network diameter is slightly more than an equal cost fat-tree network which leads to lesser bisection bandwidth. The need for correlation among the edges makes this topology not particularly suited for supporting incremental expansion.

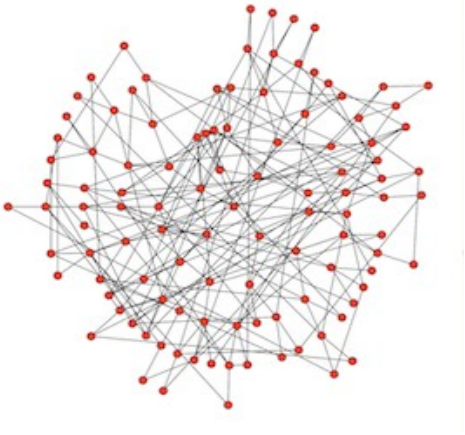

Figure 2.15: Representative Scafida Topology

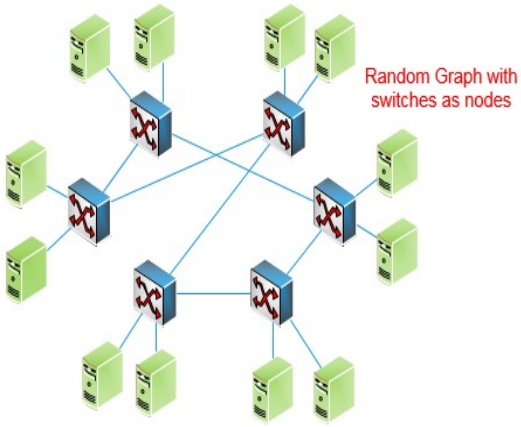

Figure 2.16: Jellyfish Topology

2.3.2.4.2 Jellyfish

To flatten the network to a single layer mesh-like topology, Singla et al. [60] introduced Jellyfish (see Figure 2.16) which allows incremental expansion of DCNs. A random graph created at the ToR switch level can connect r_i ports of a ToR switch (having k_i ports) to other ToR switches and remaining k_i-r_i ports are connected to servers and considering the network to be built with N similiar ToR switches, N(k-r) servers can be supported in total. Jellyfish is a flexible network that can scale out easily to support a larger number of servers than an equal cost fat-tree at the full bisection bandwidth. This leads to cost efficiency for building networks with high port count switches. Reliability of this network is higher than the hierarchical networks due to the availability of substantial path diversity. Even though the network diameter is similar to that of a fat-tree, the average path length is shorter which results in reduced overall latency and increased network capacity. Being based on a random regular graph, the wiring complexity is very high and it constrains the possible locations of the ToRs and due to the random nature of the network, designing an optimal routing path is challenging and debugging can be highly tricky.

2.3.2.5 3D Torus based Architectures

3D Torus networks are devoid of any network switching devices and fully rely on servers (equipped with extra NICs) for the purpose of routing and packet forwarding. Hence they can be classified as server centric or direct networks. These kinds of architectures depend on grid interconnections and aim to achieve a very high bisection bandwidth with low system latency and improve the overall failure resiliency. CamCube [39] is an example of such type of networks.

2.3.2.5.1 CamCube

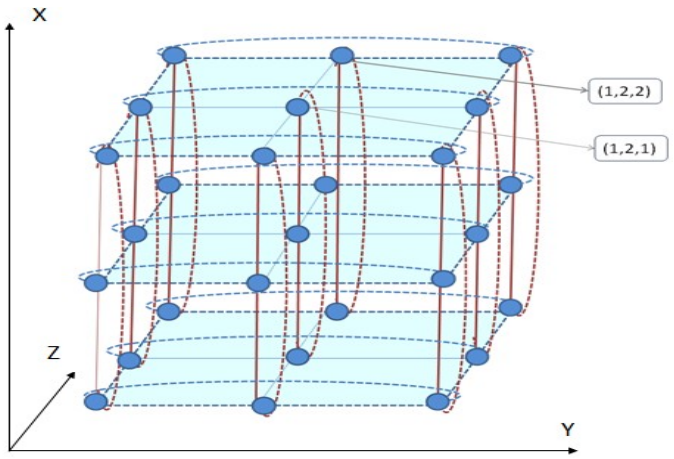

Figure 2.17: 3D torus based CamCube Topology

Abu. Libdeh et al. [39] proposed the CamCube topology, where each server is connected to six other servers (two servers in each direction) in a 3D torus topology for providing a modular structure with high link redundancy. As shown in Figure 2.17, a particular 3D (x,y,z) coordinate is assigned to each server to indicate its position for sending or receiving packets to/from one hop neighbours. All the servers are equipped with multiple NIC ports for running network applications as well as forwarding packets. The modular size of the torus topology is targeted at shipping container sized DCNs. Being devoid of network switching components, Camcube overcomes the oversubscription and bandwidth bottleneck and has lower cost of deployment and power consumption (related to cooling) while offering better fault tolerance. Architectural symmetry of the torus topology lowers wiring complexity and allows ease of incremental expansion. But, long routing paths of the order of O($N^{\frac{1}{3}}$) hop counts (N = no. of servers) increases latency and leads to poor routing efficiency. Large network diameter requires multiple network adapters in each server and hinders scaling up to very large sizes.

2.3.3 Summary of Wired DCN Topologies

All the topologies analysed in Section 1.3 have been summarised in tablular form in Figure 2.18 for a comparative analysis of the architectures based on standard network performance parameters.

Architecture	Scalability	Bisection Bandwidth	Cabling Complexity	Fault Tolerance	Power Efficiency	Cost
Multi-tier Tree	Medium	Low	Medium	Low	Low	High
Leaf-Spine	High	High	Medium	Medium	Medium	High
Fat Tree	Medium	High	Medium	Medium	High	Low
VL2	High	High	Medium	Medium	High	Medium
Elastic Tree	Medium	High	Medium	Medium	High	Medium
FBFly	Medium	High	High	Low	High	High
Dcell	Quite High	Medium	Quite high	Medium	Low	High
FiConn	Quite High	Medium	High	Medium	Low	High
Bcube	Medium	High	High	Medium	Medium	Medium
Scafida	Medium	High	Very high	Medium-High[2]	Energy Proportional[3]	Low[4]
Jellyfish	Incremental Expansion[1]	High	Very high	High	High	Low[4]
CamCube	Low	High	High	High	Medium	High
	1. Degree of scalability is dependant on the Cabling and Routing Strategy			2. Not good in case of failure of highly connected nodes	3. Power consumption of the network is proportional to the number of servers.	4. Depends on the availability of network equipments.

Figure 2.18: Comparative Analysis of Wired DCN Topology

2.4 Challenges for Optimal DCN Design

The rapid surge in traffic which is inherently dynamic by nature forces the datacentre architecture to evolve continuously. The key challenges faced by current DCNs for ensuring the desired performance are listed in the below subsections.

2.4.1 Scalability

In order to accommodate the growing traffic, DCN operators constantly need to expand their infrastructures. For example Microsoft, is always growing their DCNs in size and currently hosts over 1 million servers across more than 100 data centres around the world [62]. Another DCN giant Facebook is also constantly enlarging their existing networks over the last few years [17]. Traditionally data centres expand by replacing existing switching components with higher end ones that can support a greater number of servers. In order to effectively satisfy the ever increasing demand for data storage, modern data centres must have the option of scaling out through the addition of individual

components without requiring to upgrade their existing architecture to ensure QoS. This kind of expansion is always crucial for the DCN operator's business and hence it must not create any bottlenecks or fragmentation within the network while allowing maximum resource utilization.

2.4.2 Fault Tolerance and Reliability

Network components are prone to failures and a single point of failure can lead to subsequent performance degradation by blocking the usage of many servers. Traditional multi-tiered DCN architectures exhibit poor fault tolerance due to lack of redundant interconnections. Hierarchical architectures such as VL2, Fat-tree, FBFly allows path redundancy but cannot degrade smoothly with increased failure rate due to their fixed topology. Recursive architectures can offer better fault tolerance by virtue of several node-disjointed paths, but it leads to increased wiring complexity and longer average path length which hampers network performance by increasing latency. Ideally, the DCN system performance should degrade gracefully with rapid fault detection and minimal recovery time [63]. Sufficient redundant paths are necessary to distribute the load evenly and ensure that network resources can be deployed as and when required for enhancing server utilisation and reliability.

2.4.3 Oversubscription and Resource Fragmentation

Oversubscription ratio is the ratio of the downstream bandwidth capacity to the total upstream provisioned bandwidth at any level of a network [37]. Servers should ideally have 1:1 oversubscription to other servers inside a rack for communicating at their full bisection bandwidth. But it is usually not the case as one moves up in the network, as oversubscription is introduced by DCN operators for reducing design and hardware costs. The ToR uplinks are usually oversubscribed from 3:1 to 20:1 [19] and the top switching layers can be up to 240:1 oversubscribed [64] in extreme cases. So, due to the hierarchical nature of DCN topologies, resources get fragmented which limits server capacity. Hence, servers are clustered close to each other in the hierarchy as the distance in hierarchy affects the system performance. However, when an application gradually expands and needs more storage, resource fragmentation can constrain it from using idle servers within other applications and hinder the overall network performance. The rapid surge in the server to server traffic has increased the bandwidth requirements for the top switching tiers which now requires a large number of switches and links leading to substantial hardware expenses.

2.4.4 Network Capacity and Performance Guarantee

Data centres need ample capacity to satisfy the bandwidth hungry cloud applications. While web search applications are latency sensitive, the online infrastructure services like GFS (Google File System), MapReduce, Hadoop requires a reliable connection with a very high throughput. Oversubscription in DCNs limits the server to server capacity and restricts the availability of bandwidth in upper layer switches. Previously conducted studies have revealed [15, 65, 66] that bandwidth availability for VMs can vary by a factor of five over a period of time within a cloud platform. So suitable mechanisms need to be implemented to ensure inter host communication at the full available bandwidth within a heterogeneous environment. Currently deployed DCN interconnection links can be upgraded to up to 100Gbps capacity but will need to push further in near future.

2.4.5 Congestion Control and Load Balancing

During TCP incast [67] phenomenon, certain network links become overloaded and congestion occurs leading to packet drops once the buffers get filled up. Additionally, elephant flows can block a series of mice flows in congested links [68] causing hot-spots and packet drops from the smaller flows. Congestion may also occur if the traffic does not get evenly spread across the network due to lack of path redundancy and the volatile nature of cloud workload makes prefix routing inefficient. In these scenarios, when TCP retransmission takes place it can substantially degrade the application throughput. So, load balancing is of utmost importance in cloud based DCNs for even distribution of packets across multiple paths in the network for guaranteeing large bisection bandwidth.

2.4.6 Latency

In case of tree based or multi-dimensional hierarchical topologies, switches/servers at each level need to perform multiple iterations of packet and frame processing and hence the traffic experiences much longer delay to reach its destination than if the sender and receiver nodes were directly connected across the network. This undesired processing delay degrades system performance for latency sensitive applications such as web search, video streaming or financial transactions. Hence, traffic aware adaptive routing schemes need to be used and optimal DCN topologies need to be flattened to minimise the delay incurred at each hop.

2.4.7 Cost of Deployment

Hierarchical architectures can incur a greater cost for switching elements than recursive networks that do most of their switching through servers. Heterogeneous networks require advanced routers for the upper layer operations resulting in increased CAPEX cost. On the other hand, recursive homogeneous DCNs using low cost commodity switches incur additional expenses in the form of extra NIC and CPU cores for packet processing. Availability of multipath also increases cost related to wiring and requirement of O-E-O transceivers. Lack of incremental scalability in certain architectures hampers easy upgrading and increases OPEX cost. Ideally, DCNs must be able to enhance bandwidth capacity and network performance at a reasonable cost. Another major component of DCN OPEX budget is the cost related to power consumption. Even for a standard DCN, the energy bill can be as much as that of 25,000 households [69,70] and alarmingly such costs are doubling every five years [71].

2.4.8 Energy efficiency

Global demand in DCN energy consumptions is on the rise and is projected to reach 140 billion kilowatt hours by 2020 [72, 73], equivalent to the consumption of 50 power plants and totaling about $13 billion annually. By 2030 DCNs may consume as much as 13% of the global energy consumption [74]. Over-provisioning of server resources causes rapid energy drain and even when idle, servers consume up to 50% of its full power [75]. A lot of effort has been devoted towards making servers more energy efficient since they consume the highest percentage of DCN energy. However, this increases the percentage of power consumed by switches/routers which are also associated with a substantial amount of cooling cost. Consumption in case of recursive architectures gets further accelerated since they require servers to be always kept switched on for packet forwarding even when no application is running on it thus resulting in a poor return on investment (ROI). Hence modern DCNs should be using greener architectures where power consumption can be made to be proportional to server utilisation.

2.4.9 Modularity and Compatibility

Modular data centres (MDC) are packed inside shipping size containers which are flexible enough to run by themselves or facilitate the expansion of DCNs through the interconnection of containers containing 1k-4k servers each [76]. The plug and play nature of MDC ensures rapid and easy de-

ployment and can improve cost and power efficiency through means of easy wiring and lower cooling costs. It can support around six times more servers than a regular enterprise data centre using the same space. However, the heterogeneous nature of data centres and server compatibility, limits such an approach. Data centres tend to grow over time and as a result, new network elements need to be periodically introduced to accommodate the expansion. This leads to a mismatch since containers built over different periods of time can have varying hardware configurations. So heterogeneity poses a serious challenge to the widespread use of modular DCNs.

2.5 Conclusion

Data centres form a critical part of modern storage and communication systems. However, the rapidly evolving data centre traffic patterns and the vast application diversity make the optimal DCN infrastructure design highly challenging. DCN performance criteria are heavily dependent on the underlying network architecture with relation to throughput, response time, scalability, cost and power efficiency. A flexible network with high bandwidth and fault tolerance is required for communication intensive traffic like social networking and web searches while data intensive traffic like grid computing rely on the programmability of servers. The traffic in data centres being significantly different from traditional communication networks introduces various infrastructure and design challenges for efficient intra-DCN communication. After analysing the different DCN topologies mentioned in this chapter, it can be deduced that each solution tries to improve upon the shortcomings of its prior solutions, but still, the opaque electronic network architectures face significant problems related to scalability, network performance and energy and cost efficiency. So, in accordance with the type of workload and underlying business applications, DCN architectures need to be optimised to ensure that they meet their desired performance goals in a cost-efficient way. However, with the growing rate of expansion, electronically switched networks can tend to form a bottleneck and become unsustainable in near future.

Hence advanced switching solutions are required to overcome the shortcomings of current DCN networks and use of optical switching can prove to be the most viable solution with regard to network capacity, latency and power consumption requirements. So, Chapter 3 of this thesis will deal with the optical switching paradigm and give a wide overview of the evolution of optical DCN networks.

Chapter 3

The Role of Optics in Data Centres

Rapid rise in DCN traffic has led to the expansion of enterprise-scale data centres into warehouse scale with several hundred thousand servers which needs to be connected through a low latency and high capacity network. However, current DCNs using electronic packet switches and point-to-point interconnects suffer from limited throughput, high latency, low flexibility and poor network performance. Additionally, the overall network scalability is restricted due to the limit on the ever widening switch radix which in turn is increasing the network cost and cabling complexity. Commodity switches will not be able to sustain the projected enormous traffic growth and associated bandwidth demand without a substantial increase in power consumption especially when wider Core/Spine switches with more than 1000 ports are being required. In such a scenario, electronic switches are soon going to reach their limit and become unsustainable for further expansion.

With the maturity of photonic components, optical switching, that makes use of wavelength division multiplexing (WDM), has been considered as a viable solution for providing an ultra high capacity, very low latency and high-performance transparent switching medium. Compared to their electronic counterparts, optical interconnects can prove to be more cost-effective and power efficient by performing switching in the optical domain and provide increased bandwidth capacity ($\sim$Tbps in each fiber with WDM routing) irrespective of distance with smaller interference, and crosstalk [77]. Owing to the availability of WDM routing facility, optical switches can achieve better scalability and have all-to-all contention-free interconnection between servers due to multi-wavelength routing. Due to the absence of repeaters and the non-store and forward mechanism of optical switching, substantial improvement in latency is achieved in comparison to commodity switching.

Transparent optical networks are on-demand connection-oriented networks that are highly flexible in nature and can be reconfigured during network operation to be best suited for the highly volatile DCN traffic. Very large radix optical switches cause less heat dissipation than electrical switches and so cooling costs are reduced. Moreover, to accommodate the growing demand of storage, optical switching can facilitate the smooth upgrading of link bandwidths to 100 Gbps and beyond. This chapter aims to evaluate the rationale behind using optical solutions in data centres against the backdrop of different contributing factors. Thereafter major optical switching topologies are reviewed and future challenges in this domain are identified.

3.1 Emergence of Optical Switching in Data Centres

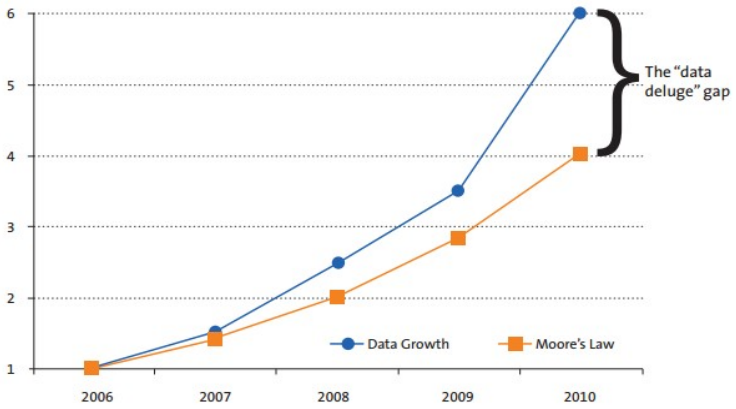

Figure 3.1: Projected Data growth in comparison to Moore's Law [78]

The growing popularity of cloud and web-based applications have considerably increased the DCN workload and this growth pattern has been even outgrowing Moores Law. Figure 3.1 depicts the data deluge gap (availability of more data than the ability to process it efficiently) with relation to transistor scaling, observed since the last decade [78]. DCNs host servers or compute-storage nodes, each of which can contain one or more virtual machines (VM) with multiple network interfaces [41] and it is very vital for the DCN architecture to be efficiently designed for easy deployment coupled with convenient and budget friendly maintenance. Some of the major factors responsible for causing a paradigm shift in DCN switching technology are explained in the following subsections.

3.1.1 Servers and Network Capacity

For tackling the scalability bottleneck and the increase in traffic, DCN operators are resorting to very powerful servers based on advanced processors. Previously, increase in processor performance was

solely dependent on increasing the number of transistors on-chip, operating at higher clock frequencies. However, since the last decade, due to lower energy budget and heat dissipation constraints, on-chip frequencies are unable to follow this trend. So, multi-core processors operating at lower clock frequencies are being used to overcome the performance scaling problem and consequently, the DCN interconnects linking the servers together are becoming increasingly important [77, 79]. However, as a consequence of multi-core processors, very high bandwidth interfaces are required for communicating with servers situated in separate racks [2].

Figure 3.2: Projected Bandwidth Requirement in Datacentres

Alternatively, DCN operators are adding more servers to increase the processing power. Some hyperscale data centres have already scaled up to more than 100,000 servers. Network connectivity between such massive number of servers and virtual machines play a vital role in deciding each job completion time [41]. This exponential growth in size is posing significant challenges for the interconnection networks which require very high bandwidth interconnects for ensuring high throughput and minimal packet delay. In traditional DCNs, server racks are connected with ToR switches while a higher radix aggregate switch interconnects racks to form clusters. Figure 3.2 shows the projected bandwidth requirement for the switching ports in DCNs based on the capacity of the server port being used (considering 48 servers in each rack and 32 racks in each pod/cluster) [80]. With data rates rising from 10G to 25G and from 40G to 100G in the near future, the bandwidth requirements for the switching layers will be much higher than the currently available switches.

Another possible way of scaling up the capacity is to interconnect multiple small or mid-sized DCNs (closely located to the end users) to form a single unit with faster switching and improved scalability. But, inter DCN connectivity in such a scenario can be difficult to maintain due to the significant rise in

the required bandwidth of the network interconnects linking the distributed DCNs and subsequently, very high multi-Tbps capacity will be required [77].

3.1.2 Switching and Network Architectures

The need for high radix and high throughput COTS switches in massive sized DCNs limits network scalability. These switches can be power inefficient and have limited throughput and normally store and forward packets across multiple switching layers, increasing the latency, cabling cost and complexity. Since every incoming packet is processed and then dispatched individually to designated output ports, a packet traversing the network has to suffer from unnecessary queuing and processing delay. Even though low-latency cut-through switches are currently available (from Arista and Mellanox [81]), they have a smaller radix compared to commodity store-and-forward switches. Thus, DCNs either need to employ high latency COTS switches in the core layer or add an extra layer on top, thereby increasing the network diameter and chances of congestion. Even when using cut-through switches, latency spikes can occur due to the congestion focal points at the upper switching layers. Hence, latency-sensitive applications need to use dedicated networks to meet their performance demands in current DCNs. Additionally, commodity switches tend to consume power at a rate that is proportional to their packet handling capacity. So, power consumption naturally increases on increasing the switch capacity [82].

In order to build a DCN with several thousand servers ($\geq$10,000) each operating at very high data rates, massive commodity switches are required for avoiding hierarchical multi-layer architectures that can lead to bandwidth bottlenecks, high server-to-server latency, and poor cost efficiency [83]. However, the limited I/O bandwidth of the switch application specific integrated circuit (ASIC) caused by the limited ball grid array (BGA) density limits the number of ports that can be supported in each switch with the desired link rates [84]. Multi-tiered architectures lead to throughput bottleneck across the oversubscribed network paths particularly higher up in the network and hence, the throughput across racks becomes limited with respect to that within a rack which makes inter-rack communication more difficult. However, the prevalent east-west traffic pattern requires efficient connectivity between server racks. So, alternative architectures are obviously required for better DCN performance and to overcome the limit on network scalability.

3.1.3 Power Consumption

Rising energy consumption and the associated operational cost (OPEX) is a major concern for the evolving DCN infrastructure. The ratio between the energy consumed by the IT equipment and overall energy consumption forms the PUE (power usage effectiveness) which denotes the rate of energy consumption in DCNs. Modern data centres tend to exhibit a very low PUE, such as Google whose most power efficient DCNs have a PUE of just 1.12 which means that for every Watt of energy spent on IT equipment, only 0.12 Watt is spent for cooling and power distribution [85]. Recently, Facebook has been choosing Nordic locations for its newer data centres to utilise the icy Arctic climates and further bring down the PUE to around 1.07 [86]. Such low PUE values clearly indicate that major energy savings are possible by limiting the energy usage of the IT components. IT energy expenditure can be subdivided into the consumption of servers, storage devices and switching network. As per the GreenData Project survey, network components consume at least 23% of the total IT energy consumption in DCNs and up to approximately 15% of the total DCN energy and this is expected to rise sharply in near future [87]. For operating the network equipment, a large amount of power is also wasted in cooling and studies indicate that for estimating actual power budget, power consumption of the network equipment should be doubled to take account of the cooling costs [88].

While the bandwidth requirement in DCNs is growing at a rate of 20 times every 4 years resulting in a CAGR of 111% from 2012 to 2020 (1 Pb/s in 2012 ->20 Pb/s in 2016 ->400 Pb/s in 2020), the net rise in network power consumption in DCNs is only allowed to grow at a mere rate of 2 times every 4 years, with a CAGR of 19% (0.5 MW in 2012, 1 MW in 2016, and 2 MW in 2020) [89]. So it is evident that the recent network infrastructure will be unable to sustain the data centre traffic in the near future when the energy efficiency will need to drop to less than a milliwatt/Gbps [90].

Currently, commercial DCNs use electronic switches and optics have only been introduced as optical fibers and transceivers for point-to-point communication. Very high data rates are difficult to achieve in electrical interconnects due to crosstalk and attenuation, so a large number of copper interconnects are required for high capacity networks leading to high power consumption and limited scalability. Optical interconnects, in turn, can provide larger data rates over a longer transmission distance and are more scalable. But optical interconnects have to convert packets to optical signals at the transmitter side and then back to electrical signals at the receiver side since packet switching is still done in the electronic domain. This extra E/O and O/E conversion wastes a lot of power and adds to the net packet

latency which is even more aggravated when electronic buffering is required to resolve contention. The power consumption problem is further worsened as the network equipment tends to consume similiar power when they are idle as compared to working at full load capacity [91]. Using optical interconnects in place of the currently used 10GBase-T cabling (copper interconnect) can generate savings worth approximately 150 million dollars of electrical power [92]. As per estimation by the US Department of Energy, 75% of total IT energy can be saved if all optical switching networks are introduced in DCNS [93]. So, data centre architecture definitely needs to be redesigned to meet the future performance requirements and still remain within the power consumption limits.

3.2 Taxonomy of Optical Switching Architectures

To overcome the limitations of wired electronically switched data centres, optical interconnection networks have been proposed that can exploit the merits of the optical domain to design a very high capacity, cost effective, green network. Optical DCNs can be mainly classified into two categories; hybrid networks where optical switching solutions are coupled with existing commodity switches or all-optical networks which can be further subdivided into different switching architectures based on the type of optical switching architecture being used as shown in Figure 3.3.

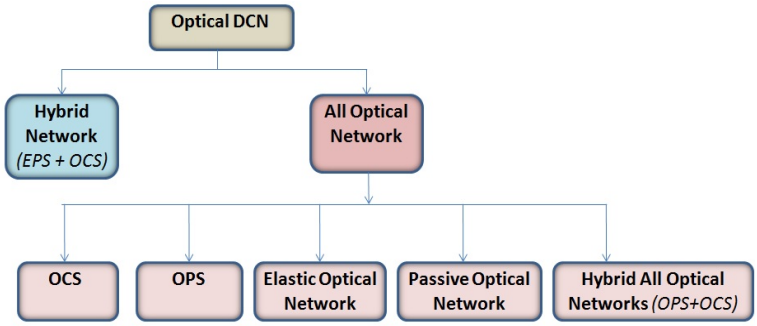

Figure 3.3: Taxonomy of Optical Switching Architectures

3.2.1 Hybrid Switching Schemes

A hybrid approach can have optical interconnects working in parallel with existing legacy switches to offer the advantage of an easy incremental upgrade for operating DCNs. The ToR switches can be expanded by adding optical modules that will increase bandwidth and reduce latency while at the same time the legacy network can be used for all-to-all communication as before. Since OCS (such

as 3D MEMS switches) are currently available off the shelf whereas OPS have not been launched on a mass scale most of the hybrid optical solutions have used OCS to alleviate the problem of elephant flows while the electronic switches handle the packet level switching granularity. Brief reviews of such solutions have been provided in the next section of this chapter.

3.2.1.1 c-Through

Based on the traditional multi-tier tree topology, c-Through [94] connects the ToR switches to an auxiliary optical network (see Figure 3.4) for facilitating high-speed and high-capacity communication. Depending on the varying traffic load, the control plane dynamically reconfigures the optical circuit switch to set up new connection paths between server racks within milliseconds. Electrical and optical networks are mutually isolated through a data plane which demultiplexes traffic from the ToR switches onto a circuit or packet based path giving higher priority to the optical path when a circuit between two racks is available. Every server has two virtual VLANs for mapping packets to either the electrical or optical network. The c-through solution can be feasibly incorporated into an existing network, based on inter-rack traffic analysis. But scalability is limited and configuration complexity goes up when several optical managers are required for handling the growing number of servers in current DCNs. Since fixed-bandwidth links are maintained within each ToR, a bottleneck is reached when two ToR switches try to use the full bandwidth connection simultaneously.

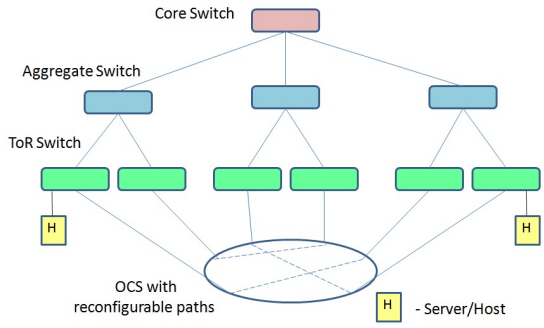

Figure 3.4: c-Through Architecture

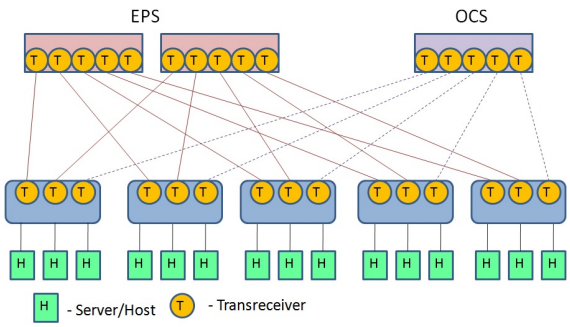

Figure 3.5: Helios Architecture

3.2.1.2 Helios

Helios [95] is a two-tier architecture particularly suited for modular data centres and is similar to the two tiered Leaf-Spine topology (see Figure 3.5). While the ToR switches are traditional commodity switches, the core switch can be electronic (packet switch) or optical (MEMS-based circuit switch)

where the packet and circuit switch paths remain mutually complementary. Unlike c-through which has no switching differentiation based on traffic type, the electronic core switches in Helios handle the short-lived/bursty flows while the optical circuit switches can handle the long-lived inter-pod flows using WDM via high capacity optical links. Inter-pod traffic demands are estimated by a central topology manager based on real time observation and accordingly circuit switches and uplinks are reconfigured. ToR switches use optical transceivers to connect with optical core switches through a passive optical multiplexer, forming super links capable of fully flexible bisection bandwidth assignment. The OCS network uses MEMS optical switches which are easily available, consume very little power and don't require any end-host modification. However switching speed is not adequate for bursty traffic since MEMS may require several milliseconds for reconfiguration.

3.2.1.3 Mordia

Mordia [96] is an extension of the Helios [95] model with parallel electrical and optical networks where a control plane manages the OCS reconfiguration in the microsecond range. Logically it resembles a 24 node fully connected mesh but actually, it has 6 connecting stations placed in an unidirectional optical ring with one-by-four port at each station (see Figure 3.6). Hosts have dual 10G ports, one connected to the commodity switch and the other linked to an OCS for routing wavelengths on different ports through a WSS (Wavelength Selective Switch). Four wavelengths are multiplexed at each station and the combined signal is sent to the connected device, where wavelengths are dropped or passed on to the next station. Though, Mordia has an estimated OCS reconfiguration time of 11.5μs (faster than MEMS) it is still not fully compatible with the current Ethernet packet granularity. With multiple parallel rings, it is possible to scale further than the Helios model, but this ROADM approach cannot scale up sufficiently to meet the current cloud DCN requirements due to the limit on the number of wavelengths that can be multiplexed within an optical ring. Additionally, in case of a link failure, connectivity for the entire ring may be lost.

3.2.1.4 REACToR

The REACToR [97] prototype uses hybrid ToR switches that can support up to 100 Gbps downlinks and connects to legacy 10 Gbps electronic network as well as to 100 Gbps optical network links (see Figure 3.7). For controlling contention in the OCS network, REACToR uses low-cost DRAM memory in end hosts to buffer the packet bursts until the circuits are explicitly established. A control

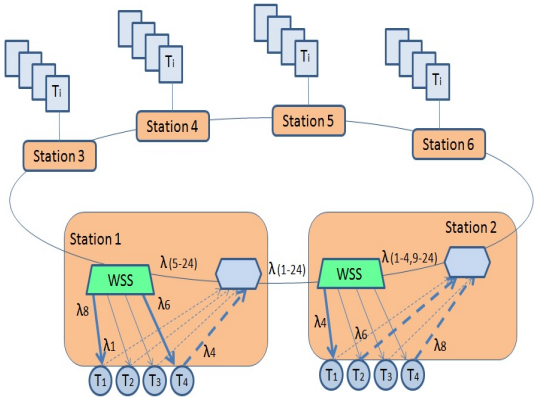

Figure 3.6: Mordia Architecture [96]

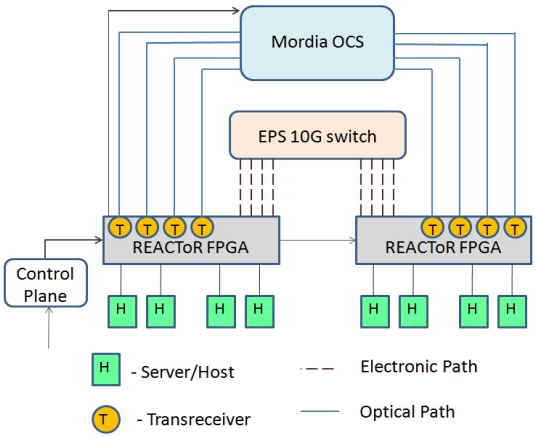

Figure 3.7: REACToR Prototype

plane routes the latency sensitive traffic across the EPS path and a synchronous signaling protocol ensures that the switch configuration matches the workload. Performance degradation of packet-based protocols at the end-hosts, due to flow-level TDMA is balanced by the fast OCS switching speed. Built on top of Mordia's circuit switch, REACToR offers packet-switch like performance with adequate bandwidth utilisation and as OCS technology is further improved this hybrid model might blur the distinction between packets and circuits. But it can be difficult to interconnect the base network amongst a very large number of REACToRs while still maintaining the buffer synchronisation.

3.2.1.5 SSX (Single-stage Shuffle eXchange)

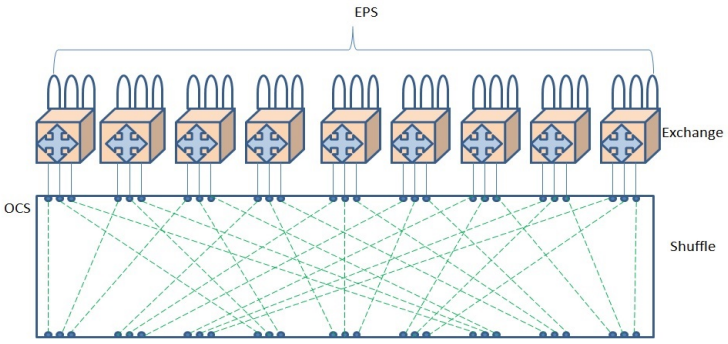

Figure 3.8: SSX Switching Architecture

Unlike c-Through [94] and Helios [95], SSX [98] employs electronic and optical networks combined together to build the interconnection network. The OCS acts as the central switching fabric (see Figure 3.8) and links individual EPS, each of which has multiple I/O ports connected to an OCS port through unidirectional fibers leading to multiple available connections between a pair of edge EPS. The optical signal fed to the EPS is switched to the desired output port which in turn connects with an OCS input

port and the OCS acts as a space switch for delivering the signal to the destination EPS. With recent development in OCS port count, the SSX architecture might be scalable for larger DCNs but the coarse-grained switching granularity and high reconfiguration time could create a bottleneck. In order to avoid multiple reconfigurations, SSX uses a multi-hop strategy to allow communication between a pair of EPS not linked by the OCS. But this leads to multiple packet processing and O/E/O conversion before reaching the destination EPS which increases latency and power consumption.

3.2.2 All Optical Switching

Wavelength Division Multiplexing (WDM) based optical solutions can be an option for the high bandwidth, low latency, scalable and energy efficient needs of modern Data Centres. Moreover, the flexible multiplexing and super-high switching capacity of all-optical switching technology can provide the possibility of flattening the DCN architecture. Depending on the switching granularity and the type of components being used, all optical switching techniques can be divided into five categories as explained in the following subsections.

3.2.2.1 OCS (Optical Circuit Switching)

OCS is a relatively mature technology that can provide a guaranteed amount of bandwidth allocated for the exclusive use of an active connection, ensuring Quality-of-Service (QoS). By using WDM technology every network link can have several available wavelengths and a wavelength can be simultaneously used in different lightpaths if not sharing any common link. However as OCS supports the deployment of static routing with pre-established lightpaths, it cannot handle bursty DCN traffic, which leads to congestion in overloaded links. Being a coarse-grained switching technology operating at the level of a wavelength channel, it exhibits low flexibility, poor spectral efficiency and inefficiency in switching fine-grained DCN traffic. OCS technology is mainly available through optical 3D MEMS that can support quite a high number of ports, but the high switch reconfiguration time makes it suitable for handling only the long-lived elephant flows. Brief reviews of some recent OCS solutions have been provided in the following subsections.

3.2.2.1.1 OSA (Optical Switching Architecture)

MEMS acts as the OCS switching matrix in OSA [99]; an extension of the Proteus [100] architecture for interconnecting the server racks. As shown in Figure 3.9, N transceivers in the ToRs do the

E/O conversion and passes the signal onto a MUX which combines them before sending to a WSS where the signal is mapped into K subsets and directed to one of the K ports of the MEMS as per the transmitted wavelength. At the receiving end, a $K \times 1$ coupler collects the wavelengths from the OCS and combines them into a WDM signal which is then divided into N wavelengths by a DEMUX before transferring to appropriate ToR switch ports. OSA allows multipath between the ToR and OCS matrix by virtue of high bisection bandwidth and link capacities. A central controller adjusts the MEMS configuration for localising heavy traffic flows and assigning appropriate outgoing links to each ToR. Circulators used to connect the WSS with the couplers, simplifies OCS interconnection and allows bidirectional optical transmissions. With an arbitrary topology of K-regular connection graphs, OSA allows flexible link capacities and minimises ToR hotspot formation within the rack, but it is not enough to support the needs of dynamic DCN traffic owing to inherent OCS switching limitations. Similar to the SSX model, OSA has a multi-hop routing policy for supporting bursty communications but this, in turn, increases routing complexity, latency and cost of O/E/O conversions.

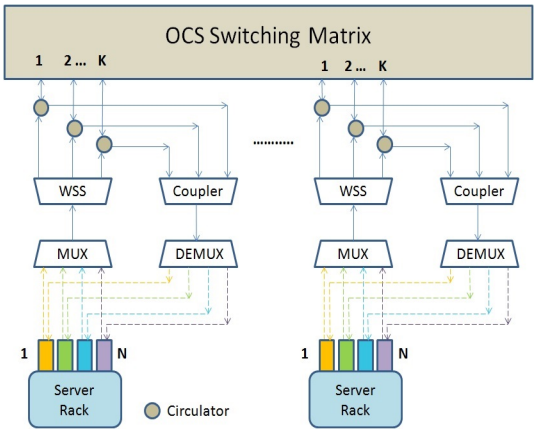

Figure 3.9: OSA Network Architecture

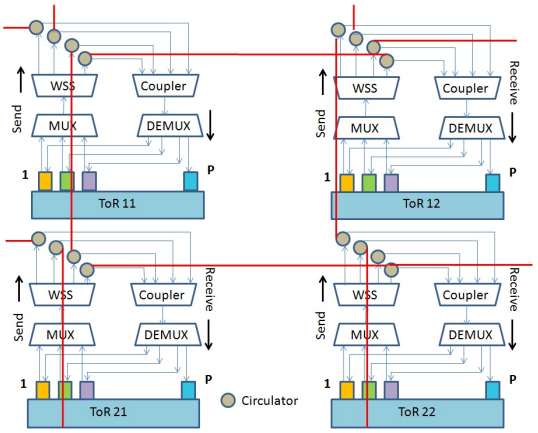

Figure 3.10: WaveCube Architecture

3.2.2.1.2 WaveCube

WaveCube [101] does away with the use of MEMS to overcome the bottleneck of scalability and single point of failure. Every pair of ToRs has multiple node-disjoint paths for ensuring high performance through dynamic bandwidth assignment. It is a flattened single layer architecture where a central controller manages the wavelength distribution and fault-tolerant routing. As shown in Figure 3.10, a n-dimensional cube can have 2n node-disjoint paths between connecting ToRs and dynamic link bandwidth is used over these 2n multi-paths through WSS where MUX multiplexes all wavelengths from each ToR and divides them into different groups. The number of links between ToRs is

equivalent to the number of wavelengths in each group. On the receiver end, wavelengths are directed to the DEMUX via a 3-way circulator and coupler to reach the destination ToR. WaveCube shows graceful degradation in case of link failures and can be scalable to quite an extent. By removing most of the active components it is less expensive, consumes less power than contemporary architectures and has a low wiring complexity. However, the dynamic bandwidth adjustment algorithm is not optimal and in the case of very bursty DCN traffic the network performance can be compromised.

3.2.2.2 OPS (Optical Packet Switching)

OPS offers fine-grained and adaptive switching capability at packet level but is still technologically immature. It can provide virtual optical paths for every connection that emulates the behaviour of a lightpath while still provisioning bandwidths less than a full wavelength. Thus it allows more flexibility and better bandwidth utilisation than OCS networks. OPS can support both statistical multiplexing and sub-wavelength granularity which minimises energy consumption, delay and allows easy capacity upgrade and is suitable for the dynamic DCN traffic. But the lack of all-optical buffers hinders contention resolution. Wavelength conversion using spectral efficiency that transfers data from congested wavelengths onto a free wavelength can be a solution. Alternatively, Fiber Delay Lines (FDLs) can offer fixed length delays until contention is resolved even though it introduces signal attenuation. Packet synchronisation at the end host can also be a solution but is impractical to use in DCNs with a large number of links. OPS can be categorised as follows and brief reviews of some recent architecture in each subgroup have been provided in the following subsections.

- **AWGR based networks** use passive cyclic AWGR where input ports can connect with any of the N output ports by using N different wavelengths based on a static routing table. Additionally, it exhibits wavelength parallelism and so wavelengths coming from different input ports can be simultaneously received at a particular output port. When coupled with TWCs (Tunable Wavelength Converters) AWGR forms a fast, reconfigurable switching matrix that allows all to all connectivity by transferring an optical signal from any input port to any output port.
- **Broadcast and Select (B&S) network** can be a single or multi-stage which involves input signals being broadcasted to all the output ports and receivers can tune to receive the desired signals based on the inherent multicasting property of the network. This type of network uses a passive star coupler/splitter type device connected to multiple nodes.

3.2.2.2.1 Petabit

Petabit [102, 103] uses interconnected AWGR and TWCs to form a bufferless OPS fabric based on the Clos network model and consists of input modules (IMs), central modules (CMs), and output modules (OMs) (see Figure 3.11). By appropriately tuning the TWC a high-speed connection between the source and destination nodes is laid out. Being a bufferless design, Petabit makes use of the linecards (containing buffers) to resolve contention and the line card buffers store the contending packets awaiting transmission. For avoiding the head-of-line blocking (HOL) problem, packets are grouped into virtual output queues (VOQ) maintained on a per-OM basis, according to their destination address. Though HOL blocking is not fully eliminated, this does reduce the queue numbers and by using an efficient scheduling algorithm coupled with a large speedup, the switch fabric is estimated to achieve a near 100% throughput. It is possible to scale up to a fairly large switch size, using multiple AWGRs with the multi-stage design but this increases the scheduling complexity. Since a large number of TWCs are required with the growing number of AWGRs, it increases power consumption and deployment cost which makes it a not so practical choice for large scale DCNs.

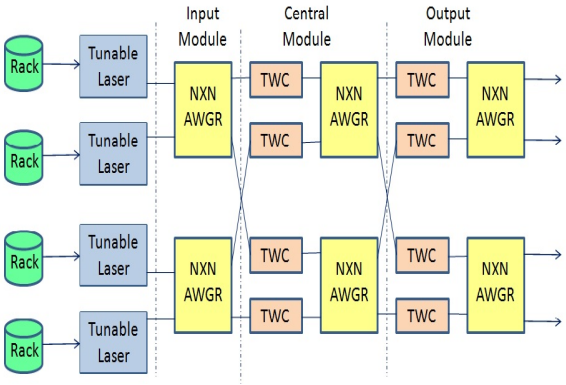

Figure 3.11: Petabit Network Architecture

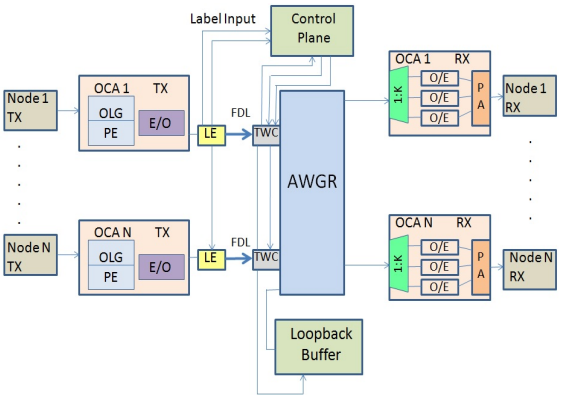

Figure 3.12: LIONS Architecture [105]

3.2.2.2.2 LIONS

As shown in Figure 3.12, the basic LIONS [105] model derived from the DOS [104] architecture uses an array of TWCs, a high radix AWGR and Label Extractors (LEs), to form an OPS fabric where the LEs receive packets from the ToR and detach the optical labels from the payload before sending them to the control plane for processing. As per the information contained in the packet header, the control plane configures the TWC and assigns the required wavelength to the AWGR inputs so that packets can reach their desired. The limit on the number of receivers available per output port leads

to contention and LIONS uses a forward-store strategy where the contended packets that fail to get grants from the arbiter are stored temporarily in a shared loopback buffer.

Replacing the single SDRAM used in DOS, LIONS uses three different loopback buffer designs namely; the shared loopback buffer (SLB), the distributed loopback buffer (DLB), and the mixed loopback buffer (MLB). SLB receives delayed packets from different inputs concurrently on different wavelengths and stores the contended packets in parallel shared memory. On receiving the grant for a particular output, it serially sends all delayed packets. DLB uses N separate memory blocks, one for each input port and to avoid HOL blocking, multiple transmitters per queue are deployed to simultaneously transmit multiple packets to different outputs on different wavelengths. MLB combines the advantage of both SLB and DLB and uses multiple AWGR inputs to support multiple separate queues so that each queue is specified for every single AWGR input. Unlike the DLB, in which each queue serves only one node, each queue in the MLB serves multiple nodes. DLB can achieve lower latency than SLB, but it uses N ports from AWGR while SLB uses only one to support the queues for N hosts. The number of transmitters required to alleviate HOL blocking is more in DLB than SLB but lesser than MLB. DLB and MLB together can give better performance than using SLB alone.

The LIONS architecture offers a low latency flattened network suited for dynamic DCN traffic and is designed with high radix passive AWGRs that have high throughput and low power dissipation. However, scalability can be limited by AWGR size and the use of a high-capacity, high-speed shared buffers and TWCs increases power consumption and latency due to additional O/E and E/O conversions. So, to overcome the switching bottlenecks, LIONS further replaces the shared buffering with a negative acknowledgement scheme which can promptly notify the end nodes whenever one of their packets cannot reach the desired output due to contention.

There has been another recent innovation on the basic DOS model, proposed by J. Wang et al. [106]. While DOS requires a very high-capacity shared electronic buffer having a port count and port speed equivalent to that of the overall switch for resolving contention and results in substantial cost and power-consumption; the new architecture utilises a hybrid-buffered AWG architecture and packet scheduling scheme for dealing with the contended packets. The OPS switch uses a number of multi-wavelength FDLs coupled with a small electronic buffer for integrating the FDL wavelengths into the AWG switch. This results in a lossless switch with a good latency performance and lower power consumption, as compared to the standalone use of either of the two buffering strategies.

3.2.2.2.3 OSMOSIS (Optical Shared Memory Supercomputer Interconnect System)

OSMOSIS is a two-stage low-latency, B&S architecture based on WDM and space division multiplexing. In the first stage, wavelengths are multiplexed in a WDM line and a coupler broadcasts the WDM signal to all the second stage modules (see Figure 3.13) where SOAs (Semiconductor Optical Amplifiers) are used as fiber-selector gates for choosing desired wavelengths. OSMOSIS [107] model has a 64 node interconnect scheme with 8 wavelengths each on 8 fibers and an OPS switching matrix is formed with a 8:1 fiber-selection stage followed by a 8:1 wavelength-selection stage at each output port. Using a separate scheduler that synchronises with every packet arrival, a central arbitration unit reconfigures the switch matrix and enables high-efficiency and high throughput switching. Scalability can be enhanced by deploying several switching blocks in a fat tree topology but the design involves a very complex control plane and using SOA devices increases cost and power consumption.

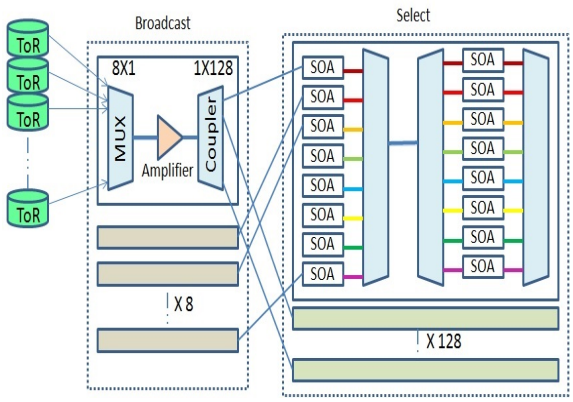

Figure 3.13: OSMOSIS Network Architecture

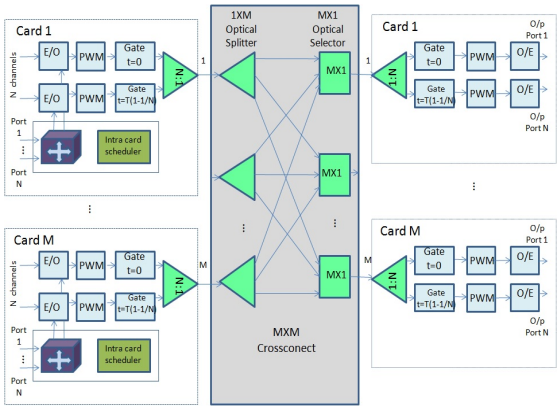

Figure 3.14: STIA Architecture

3.2.2.2.4 STIA

STIA [108] is a B&S network combining wavelength, time and space domain switching. As shown in Figure 3.14, the OPS switching matrix contains M line cards with N ports linked by an M x M space switch. Packets get switched between linecards in the space domain, they are switched between the i/p and o/p ports in the time domain and WDM encodes packets on multiple wavelengths to increase the throughput in the wavelength domain. WDM packets are used to comb an array of N optical channels with one modulator. Then via a PWM (passive wavelength stripped mapping) device packets are delayed from other modulated channels and gated in time to create WDM packets which are transmitted and time-multiplexed by an N:1 coupler for forwarding to the MxM space switch. At the receiving end, the opposite process is simply applied by each output port, whereby a 1:N splitter transmits the

multiplexed WDM packets back to the linecard and through SOA gates, selected WDM packets are reversely delayed by the PWMs and sent to the respective ports. STIA gives a low power consuming, high throughput architecture with a low wiring complexity. But as only a small portion of power is transmitted to the receiving node, expensive amplifiers have to compensate for the lost power. This, in turn, increases operational cost and hinders scalability and the deployment of wavelength-based time compression technique increases the data processing time.

3.2.2.3 EON (Elastic Optical Networks)

EONs enhance the flexibility of bandwidth provisioning by adaptively assigning bandwidth resources to applications, based on traffic demand. Switching granularity is realised through flexible spectrum allocation via bandwidth variable transponders (BVTs) and variable bandwidth crossconnects. Multi-carrier techniques using low-rate subcarriers generated by low-speed modulators, WSS supporting optical switching of arbitrary spectrum slices or AWGR can be used as the variable bandwidth cross-connect. A brief review of a recent EON architecture is provided in the next subsection.

3.2.2.3.1 EODCN

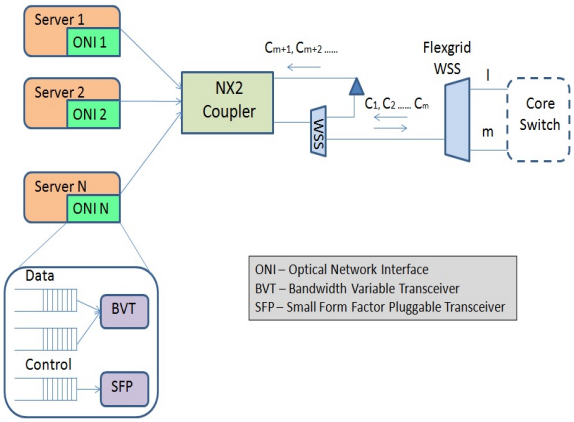

Figure 3.15: EODCN switching at the ToR [109]

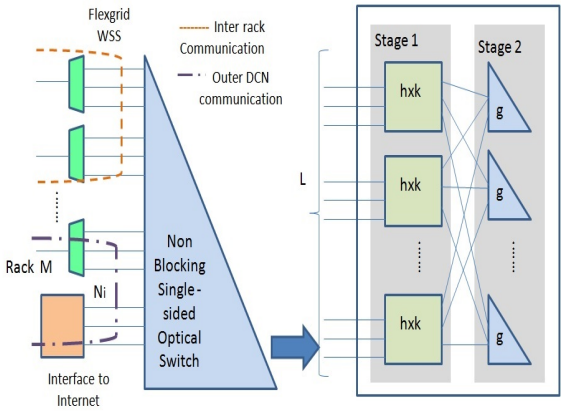

Figure 3.16: Elastic Optical Core Switch [109]

WDM, OCS and B&S switching are combined together in EODCN [109]. It uses optical ToRs made up of Nx2 couplers, connected to a WSS that sends selected wavelengths back for intra-rack communication. The core switching matrix resembles a non-blocking Clos network comprised of MEMS and a flexi-grid WSS that enables inter-rack communication. An intra-rack and inter-rack scheduler is used to compute the transmission time and duration for the packets, which are buffered at the compute/ storage node (CSN). BVT is used to tune to an appropriate wavelength for supporting multiple

data rates. When packets are transmitted from the CSN's transceiver the remaining circuit is already preconfigured and after the propagation delay, packets are received by the corresponding receiver. As core networks are bufferless, the CSN buffers should be large enough to hold packets until the network path is setup. EODCN is scalable and the optical ToRs can be significantly power efficient. But, the transceiver at the CSN can be power consuming. Since, inter-rack communication relies on MEMS, it can slow down the network due to high reconfiguration time. Optical losses can further limit the number of CSNs supported per rack by the optical ToR which in turn affects scalability.

3.2.2.4 Hybrid All Optical Networks

DCN traffic is very dynamic. While certain applications in DCNs tend to have bulky and long-lived traffic flows which can be efficiently offloaded through OCS paths with point to point bandwidth provisioning, the other applications tend to have bursty and short-lived flows that can be best served by the fine-grained OPS paths. So hybrid all optical networks try to combine the merits of both OPS and OCS to deliver enhanced resource utilisation, network flexibility and QoS. A brief review of a recent hybrid architecture is provided in the next subsection.

3.2.2.4.1 LIGHTNESS

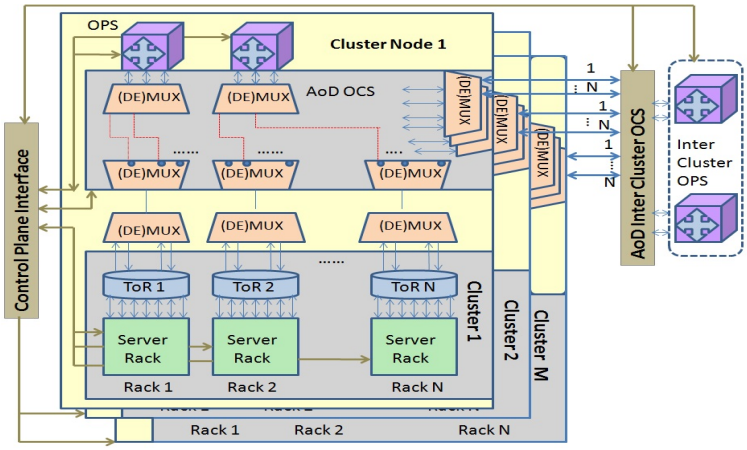

Figure 3.17: LIGHTNESS Hybrid Architecture [110, 111]

LIGHTNESS [110, 111] is a flattened two-layer transparent network utilising OPS and OCS switching technologies. An Architecture on Demand (AoD) OCS switch node provides mesh connectivity between the OPS and OCS paths in the core layer to enable inter-rack communication within a cluster. MEMS based OCS switching handles the elephant flows having high bit rates with bandwidth

guarantee while OPS switching based on the B&S architecture offloads the bursty traffic flows with low bit-rate demands. ToRs perform flow control and classifies the traffic type for co-ordination between both switching networks and the inter-cluster AoD switch provides connectivity between the clusters. LIGHTNESS provides a high bandwidth, low latency, flexible architecture for supporting the variation in DCN traffic, but it needs an effective traffic classification technique for fast scheduling prior to optical switching which in turn increases deployment cost due to its complex hardware and software requirements. Partitioning the network into clusters of ToR switches improves scalability but the B&S OPS switch facilitating highly distributed control needs concurrent interconnectivity and the power lost through splitting can also prove to be a bottleneck for scaling out any further.

3.2.2.5 PON (Passive Optical Networks)

While OCS has a slow reconfiguration time, OPS switches suffer from a buffering bottleneck and are not yet commercially available. So, the limitations in active optical switching solutions have led to the development of low cost, scalable, energy-efficient architectures using passive optical components. PONS which are already in use in access networks has been introduced to replace the power hungry COTS switches across conventional DCNs to design a reliable and fast switching network. A brief review of some PON based DCN architecture is provided in the next subsection.

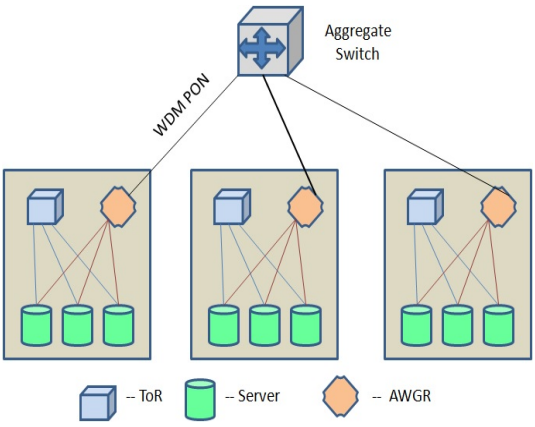

Figure 3.18: WDM Passive Optical Network Model

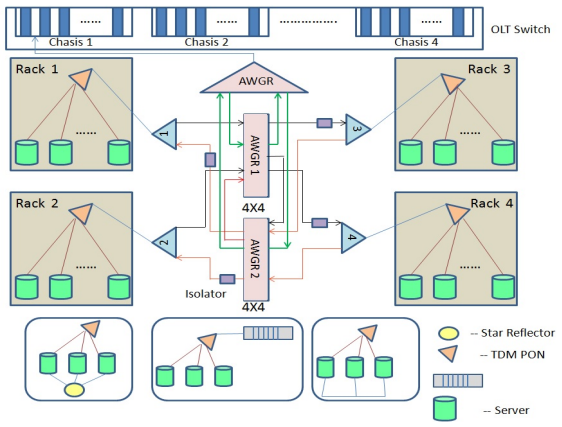

Figure 3.19: AWGR based PON DCN model [112]

3.2.2.5.1 WDM-Passive Optical Network (PON)

As shown in Figure 3.18, WDM-PON architecture uses passive AWGR and WDM-PON devices for communication between commodity racks using multi-mode fibers while commodity switches can

communicate within the rack [113]. Each server has two transceivers, one optical (WDM) and one electronic one connecting to the ToRs. Analogous to telecom networks, ToR switches serve as optical network units (ONU) and the Aggregate switches serve as optical link terminators (OLT). The WDM layer can offload the heavy inter-rack traffic and eliminate additional processing on ToR switches. Thus power dissipation at ToR level is reduced and high throughputs between racks are achieved with lower latency. At least a 10% reduction in power consumption has been estimated with this model with no additional latency gain and the architecture can be further extended to higher switching layers. But such WDM-PON architectures tend to waste bandwidth due to lack of flexibility.

3.2.2.5.2 AWGR based PON Data Centre

In this architecture each PON Cell has two AWGRs for connecting racks and the OLT switch contains 8 chassis, each with up to 16 optical line-cards. All servers in a group share the same wavelength for inter-rack and OLT communication and these wavelengths can be reused in other PON cells [112]. Servers either connect to a tuneable ONU or use NIC and fixed-tuned receivers and a tuneable laser for wavelength detection and selection. Both the OLT switch and direct AWGR connection can facilitate inter-rack communication while any of the designs shown in Figure 3.19 can be used for intra-rack communication. OLT controls all communications, and servers send a request message containing the destination address and resources requirements. If the request is granted, the OLT sends a gate message to the source and the destination ONUs to inform the servers about the wavelength assignment. A mixed Integer Linear Programming (MILP) model optimises the AWGR interconnection for wavelength routing. Availability of multi-path routing ensures load balancing with heavy traffic. However, forwarding traffic through the OLT can introduce extra delay and increase power consumption due to O/E/O conversions, queuing, and processing. The model is scalable enough to support the needs of dynamic cloud DCNs but has a high deployment cost as all servers are equipped with tuneable transceivers.

3.2.3 Summary of Optical DCN Topologies

All the topologies analysed in Section 2.2 have been summarised in a tabular form in Figure 3.20 for a comparative analysis of the architectures based on standard network performance parameters.

Architecture	Technology	Connectivity	Topology	Scalability	Capacity	Throughput	Latency	Power Efficiency	Cost
c-Through	Hybrid	EPS + OCS	Tree	Low	Medium	Medium	Medium	Low	Medium
Helios	Hybrid	EPS + OCS	Tree	Low	Medium	Medium	Medium	Low	Medium
Mordia	Hybrid	EPS + OCS	Ring	Medium	Medium	Low	Low	Medium	High
REACToR	Hybrid	EPS + OCS	Tree	Low	Medium	Medium	Medium	Low	Medium
SSX	Hybrid	EPS + OCS	Tree	Medium	Medium	Medium	High	Low	Medium
OSA	All Optical	OCS	Tree	High	High	High	Low	High	Low
WaveCube	All Optical	OCS	Cube	High	High	Medium	Low	High	High
Petabit	All Optical	OPS	Tree	High	High	High	Medium	High	Medium
LIONS	All Optical	OPS	Tree	Medium	Medium	High	Medium	Medium	Low
OSMOSIS	All Optical	OPS	Tree	Medium	Medium	High	Medium	Medium	Medium
STIA	All Optical	OPS	Tree	High	Medium	Medium	High	High	High
EODCN	Elastic Optical	OCS	Tree	Medium	Medium	High	Medium	High	Medium
LIGHTNESS	Hybrid All Optical	OPS + EPS	Tree	High	High	High	Low	Medium	High
WDM-PON	Passive Optical	OPS	Tree	Medium	Medium	Medium	Low	Medium	Low
AWGR-PON	Passive Optical	OPS	Tree	High	Medium	High	Medium	Medium	High

Figure 3.20: Comparative Analysis of Optical DCN Topology

3.3 Conclusion

The massive increase in traffic is constantly pushing the data centres to scale up beyond their limits and has reinforced the need for flexible and high performance interconnects. In such a scenario, optical switching can fulfil the requirements for ultra-high capacity, low latency and power efficiency with reduced CAPEX and OPEX budgets. The OCS approach with a coarse bandwidth granularity is more suited for heavy traffic between two end-to-end nodes where the communication pattern requires a full-wavelength capacity. On the other hand, the OPS approach can offer a finer bandwidth granularity and is suited for networks with short and bursty flows between end nodes where all-to-all connectivity is required. However, the slow reconfiguration time in OCS solutions and lack of efficient optical buffering schemes in OPS solutions are potential bottlenecks. Commercial availability of optical switches is an issue and to date, no all-optical DCN has been fully in operation. However, the concept of hybrid networks using both optical and electrical switching in parallel is getting accepted due to its lower OPEX costs and easier upgrade path, even though it cannot provide a long term solution. OPS switches are not readily available while OCS (3D based MEMS switches) are available off the shelf. This has led to combinations of EPS and OCS in commercial switching solutions such as in [114, 115] with OpenFlow standard based SDN infrastructures.

As observed from the architectures analysed in this chapter, most of the research in the optical DCN

paradigm to date has been concentrated towards the aggregation/core layers and switching at the ToRs is still handled by the costly and power hungry commodity switches. Hence for making the edge layer more energy efficient, using passive optical component based ToR switches can be profitable in the long run and since they don't have switching/buffering delays, so latency can also be minimised. So, it is evident that, optical solutions are necessary for complying with the changing DCN performance demands, but they need to be enriched with improved architectures that could take advantage of the inherent merits of optical switching and integrate it with newer and existing technologies.

After an in depth background study in Part I about the existing DCN infrastructure and how optical switching is evolving to meet the challenges of electronically switched architectures, Part II tries to look into the current scenario and analyses the performance of a typically used data centre topology in order to judge its potential bottlenecks in terms of Layer 2 and TCP switching.

Part II

Current Data Centre Network Scenario

Chapter 4

Evaluation of Electronic Data Centre Network

4.1 Introduction

As mentioned in Chapter 2, due to the increasing virtualisation of servers/applications, current DCN traffic needs flattened topologies to sustain the dense horizontally oriented traffic requiring collaboration between servers for handling a single request with several compute and store stages. There has been a shift in the traffic pattern from the former North-South (User to Application) to East-West (Application to Application) communication. As a result, 3-Tiered topologies developed to handle the vertical traffic in client-server applications can no longer be the best fit since they cannot provide a shortcut for application to application communication. Thus Leaf-Spine topology has emerged as the new successor to the multi-tier model for use in modern data centres [25]. The flattened nature of Leaf-Spine architecture boosts the network agility and allows host nodes to interconnect with lower latency than traditional tree networks. This also leads to improvement in the network bandwidth (any-to-any) due to the availability of multiple equal cost paths between any two endpoints.

Leaf-Spine topology has been adopted from the Clos networks developed in 1952 and comprises of a flat, non-blocking switching architecture where the core is divided into a number of modular spine switches. The Leaf-Spine architectures do away with the giant chassis-based switches that dominated the previous DCN architectures and instead uses simpler and leaner switches that are less expensive but more flexible to deploy. The major networking vendors like Cisco, Juniper and Arista [118] have already adopted the Leaf-Spine architecture and are actively marketing it to the enterprise data centres.

The hyperscale data centre operators are also using a modified version of the flattened Clos networks (the precursor of Leaf-Spine networks) to allow massive inter-pod connectivity (such as FaceBook's DCN design). So, it can be anticipated that the Leaf-Spine architecture will likely dominate data centres of all size in the coming years. Hence, the Leaf-Spine topology was selected for the electronic DCN implementation in this chapter and the next chapter as well since it has become the de facto reference architecture [119] used in modern data centres and the aim was to evaluate a network very similar to what is being commercially used in recent times.

Layer 2 network performance is predominantly based on queueing delay and it is possible to get a reasonably accurate representation of the underlying switching behaviour through simulations. However simulating very large networks can run into a scalability problem due to computational speed degradation and high memory consumption. So, an analytical Layer 2 model based on multi-stage tandem queued network is introduced in this chapter, that can be used for evaluating the performance of any Leaf-Spine DCN architecture. Comparison of the theoretical results with the simulation results validates the analytical design which can provide a basis for further network optimization and can be used for developing simulation tools for performance evaluation beyond the scope of theoretical modelling. The designed Leaf-Spine network is analysed to have a fair estimation of the current standards of scalability, cost and power consumption in DCNs. Detailed analytical modelling of Layer 2 traffic helps to judge the variation in traffic arrival/offloading at the intermediate queueing stages and based on the findings, closed form expressions for traffic arrival rates and delay are derived. Further performance evaluation helps to realize the impact of buffering and oversubscription and revealed the potential bottlenecks created at the Leaf uplink queues which experiences significant traffic overloading in oversubscribed networks. With a fairly simple Layer 2 switching model, the results from the simulated network demonstrated that having large buffers solely in the Spine switches will be unable to control traffic bursts in the Leaf layer and can result in substantial packet loss. This finding validates a similiar trend shown in [25] which used a more complex TCP traffic model but with similar buffer placement ratios.

4.2 Overview of the Leaf Spine Topology

The Leaf-Spine topology was chosen for the electronic DCN implementation. It is a flattened two layer network which is fully connected and every switch in the lower layer connects with all the

switches in the higher layer and vice-versa [25] and as shown in Figure 4.1, all the Leaf switches are meshed to the high port density Spine switches. This type of network needs high bandwidth links and high-speed switches with minimal processing delay to interconnect the storage servers in a grid-like fashion where every server has a direct connection to every other server. As a result, the entire network can be approximated to a single logical unit like a very large Output Queued switch [25, 120, 121] as shown in Figure 4.2 where packets entering from one port or transmitting from one server can be directly sent to another port or received at any server, thus resulting in a full bi-partite graph.

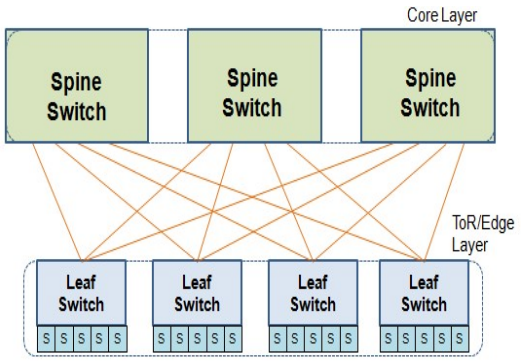

Figure 4.1: Leaf-Spine Network Topology

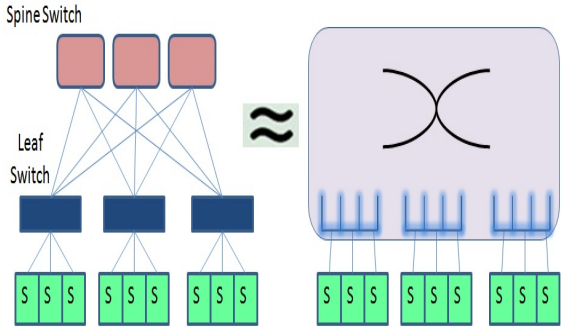

Figure 4.2: Leaf Spine network compared with an ideal Output-Queued Switch

Such networks exhibit high bandwidth availability and reliability and have good scope for scalability except for in very large sized DCNs where the Spine switch radix can run out of ports to interconnect more Leaf switches (after extending to a certain point), or when the oversubscription rate becomes unacceptable. Currently, such electronic networks can connect approximately 100K-110K servers (considering each rack with 48 servers) at 10Gbps, with very few commercially available Spine switches having $>$2000 ports [122, 123] such as Arista 7300X series (2048), Juniper QFX 10000 (2304). Such large Spine switches can be quite expensive and power hungry and are unavailable with link rates $>$10 Gbps. Due to the fully connected nature of the topology, the wiring complexity is high and building very large DCNs can only be possible by adding a third switching layer called the Super Spine layer [124] at the expense of additional cost, power consumption and latency.

4.3 Topological Parameters

Leaf-Spine network parameters are discussed in this section, followed by a summary in Table 4.1.

- **Degree of Servers** is the number of ports required per server for interconnection within the

network. Since servers are not used for switching in this network, it has a server degree of *1*.

- **Network Diameter** is the maximum hop count of a packet traveling between any two servers. In intra-rack communication, the network diameter is only *2* since there is no need to travel up to the Spine layer. However, for inter-rack communication, the network diameter is always *4*.

- **Network scalability**

  Considering **S** the number of ports in each Spine and **L** the number of uplinks from each Leaf switch, the **total number of Spines = L** and **total number of Leafs = S**. With an oversubscription ratio of $1 : A$ where Leaf-Server connectivity is R_1 Gbps (where R_1 is usually 10 Gbps as per current standards) and Leaf-Spine connectivity is R_2 Gbps links (where R_2 is usually 10 or 40 Gbps); the total number of servers in the network (**N**) is given by eq. 4.1. Here the radix of the Spine (**S**) is mainly responsible for scaling up and accommodating more servers since the number of Spines (**L**) is usually fixed to accommodate standardised ECMP routing [125].

$$No.\ of\ servers\ (N) = K * S = \left(\frac{A * L * R_2}{R_1}\right) * S, \quad \text{K = no. of servers in a rack.} \tag{4.1}$$

  In order to determine the total number of required 10Gbps ports (given by eq. 4.2) we need to consider the number of Leaf downlink ports connecting the Leaf to the Servers and the number of Leaf uplink ports connecting the Leaf to the Spine as well as the number of Spine downlink ports connecting the Spine to the Leaf.

$$\begin{aligned} Total\ no.\ of\ 10G\ ports\ (P) &= \left(\frac{A * L * R_2}{R_1}\right) * S + \left(\frac{L * R_2}{10}\right) * S + \left(\frac{S * R_2}{10}\right) * L \\ &= R_2 * L * S\left(\frac{A}{R_1} + \frac{1}{5}\right) \end{aligned} \tag{4.2}$$

- **Switch Radix variation** – Variations in DCN switch latencies and cost as mentioned in [126] based on data obtained from Cisco and Arista datasheets [127–129], show that the cost and latency in electronic packet switches increase slowly until a limiting switch radix and thereafter tends to increase rapidly and Figure 4.3 depicts such a trend. Normally an ASIC can accommodate up to 128 ports and so switches with larger radix requires multiple ASICs. For example a 256 port switch (see Figure 4.4) is formed by connecting six ASIC in a Clos topology on a single main board [127]. However, due to space constraint and power dissipation issues, only

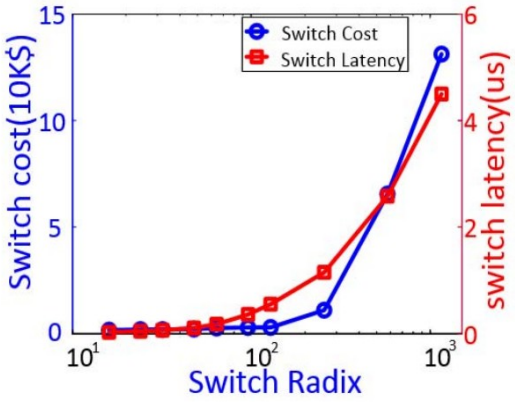

Figure 4.3: Effect of switch radix on CAPEX and Switch Latency [126]

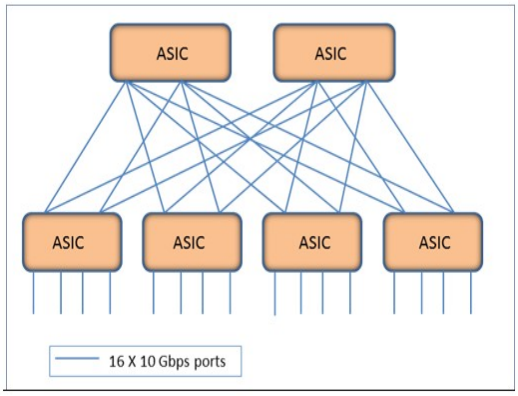

Figure 4.4: Formation of a 256 port switch from 128 port ASICs connected on a single mainboard

a few ASIC can be connected within a main board. Thus when increasing the radix, multiple mainboards may be required, such as in a 512 radix switch, six 256 port main boards are interconnected. So, naturally the cost increases with a rise in the number of components required for building larger switches. Moreover, a packet also has to pass through multiple component stages in a large switch which adds on extra processing delay to the overall switch latency.

- **Interconnecting Links** – Due to the fabric path nature of the Leaf-Spine topology, the network has a high path redundancy which leads to a large number of interconnections between both the switching layers. If the number of Leaf and Spine switches are considered as **L** and **S** respectively, then the total number of available interconnections will be **L*S**.

Degree of Servers	**Network Diameter**	**No. of Servers in a Rack (K)**	**Total No. of Servers**	**Total No. of 10G ports**	**Total No. of Interconnections**
1	4	$\left(\frac{ALR_2}{R_1}\right)$	KS	$R_2LS\left(\frac{A}{R_1}+\frac{1}{5}\right)$	LS

Table 4.1: Summary of Leaf-Spine Topology Parameters

4.4 Network Size, Cost and Power Consumption Approximation

With 2048 port Spine switches (e.g. Arista 7316) for building a Leaf-Spine network, the approximate Scalability, CAPEX and Power Consumption figures are estimated in the below subsections.

4.4.1 Scalability

Using the parameters mentioned in Subsection 4.3 and Table 4.1, an estimation about the maximum possible network size has been made and the findings are summarised in Table 4.2.

Network Size	**No. of Racks**	**Rack Size**	**No. of Spine**	**Spine Radix**	**No. of Leaf**	**Leaf Radix**	**No. of Interconnections between two layers**
98304 servers	2048	48 servers	16	2048, 10G ports	2048	64, 10G ports	32768

Table 4.2: Summary of estimated network topology parameters

- The no. of Leaf switches (*S*) required for a network with **2048** port Spine switch is equal to **2048** which is also the same as the number of server racks present in the network.
- Considering the no. of servers in each rack *(K)* to be 48 and the Oversubscription ratio *(1:A)* as **1:3**, the no. of Leaf uplinks (*L*) required are KR_1/AR_2 = **(48/3) = 16** when both Leaf-Server (R_1) as well as Leaf-Spine (R_2) link rates are fixed at 10 Gbps.
- Total no. of servers that can be supported by the network is equal to *KS* = **(2048 X 48) = 98304**.
- So, the no. of Spine switches required (*L*) is **16** and the total no. of ports in each Leaf switch is equal to **(48 + 16) = 64**.
- Finally, the no. of interconnecting links between the two layers is *(LS)* = **16 X 2048 = 32768**.

4.4.2 Power Consumption and CAPEX Budget

Power Consumption per Leaf Switch Port (P_{leaf})	**Power Consumption per Spine Switch Port (P_{spine})**	**Power Consumption per Transceiver (P_{tr})**	**No. of racks/ Spine Radix (S)**	**No. of Servers in a rack (K)**	**No. of Spine (L)**
3.5W	6W	1.0W	2048	48	16

Table 4.3: Power Consumption parameters

Power consumption figures for a network of 98304 servers (obtained from [123, 128, 131–133]), are listed in Table 6.6 along with other network parameters and using those parameters an approximation about the net power requirement of DCN switching components has been given below.

Net power consumption in the ToR or Leaf Switching layer is given by eq. 4.3

$$P_{ToR} = S * (L + K) * (P_{leaf} + P_{tr}) \tag{4.3}$$

Net power consumption in the Spine Switching layer is given by eq. 4.4

$$P_{core} = L * (S) * (P_{spine} + P_{tr}) \tag{4.4}$$

So, the net power consumption due to the network switching components is given by eq. 4.5

$$P_{tot} = P_{ToR} + P_{core} + N * P_{tr} \approx 916KW \tag{4.5}$$

Cost per Leaf Switch Port (C_{leaf})	**Cost per Spine Switch Port** (C_{spine})	**Cost per Transceiver** (C_{tr})	**Cost per fiber used for switch inter-connection** (C_{fiber1})	**Cost of intra-rack connec-tion fiber** (C_{fiber2})	**No. of Leaf-Spine Links (LXS)**	**No. of Leaf-Server Links (KXS)**
250$	500$	200$	40$	25$	32768	98304

Table 4.4: Cost Estimation parameters

Component cost figures (obtained from [134–137]), have been listed in Table 4.4 and based on that, an approximation of the CAPEX budget for DCN switching and cabling is given below.

Net CAPEX cost of the ToR or Leaf Switching layer is given by eq. 4.6

$$C_{ToR} = S * (L + K) * (C_{leaf} + C_{tr}) \tag{4.6}$$

Net CAPEX cost of the Spine Switching layer is given by eq. 4.7

$$C_{core} = L * (S) * (C_{spine} + C_{tr}) \tag{4.7}$$

Net CAPEX cost that can be attributed to the network switching components is given by eq. 4.8

$$C_{tot} = C_{ToR} + C_{core} + N * C_{tr} \tag{4.8}$$

Net CAPEX cost that can be attributed to the network interconnections is given by eq. 4.9

$$C_{link} = C_{fiber1} * (L * S) + C_{fiber2} * (K * S) \tag{4.9}$$

So, the net network switching and cabling cost can be summed up to be C_{tot} + $C_{link} \approx$ 105M \$.

The estimated figures are fairly close to the normalised cost given in [126] and is lower than similiar sized tree-based networks. However, the pricing is based on component costs available in online catalogues and may vary slightly from actual market price. A major component of the network cost can be attributed to the need for transceivers and the requirement of very large Spine switches with external deep buffers. However, this approximation does not take into account the cost and power consumed in the server stacks which actually occupies a significant amount of DCN expenditure.

4.5 Details of the Simulated Network

4.5.1 Simulation Setup

The network and its components were designed using the OMNeT++ [138] Network Simulation Framework, which is an open-source, event-driven C++ Discrete Event simulation tool supporting Object-Oriented Programming. OMNeT++ comes with a large number of inbuilt tools and utilities and it is also possible to customise component modules with user defined modelling functions where the individual modules are subsequently connected and assembled in an OMNeT++ simulation network to represent network devices, links, topologies and applications. OMNeT++ is a flexible, repeatable, scalable and reliable simulation framework with a user-friendly interface and can be used for simulating any network with discrete event approach, and then conveniently mapped into entities that communicate by exchanging messages. These attributes make OMNeT++ a very useful simulation tool for designing and testing communication networks, and hence it was chosen for network modelling for this chapter and also for simulation modelling in the remaining chapters as well.

4.5.2 Electronic Switch Performance

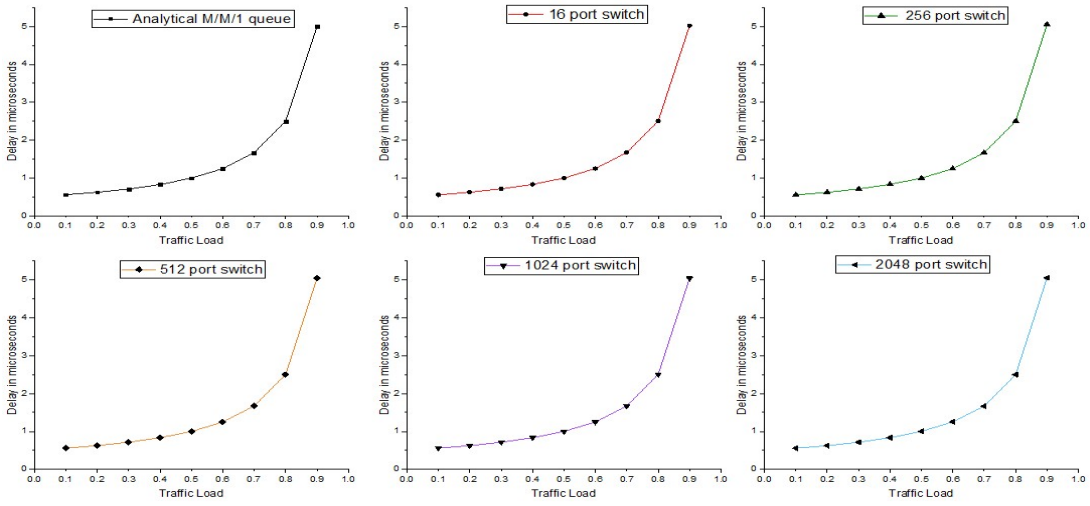

Figure 4.5: Delay Performance of Output Queued Switch Model with infinite buffer

Electronic packet switches (EPS) can be classified as per their queue arrangement which gives them a varying degree of speedup and contention resolution capability. Speedup of S denotes that up to S packets can be delivered to each output port within a time slot, which is the time gap between subsequent packet arrivals at a port. Usually, EPS can be either (i) Output queued switch (OQ) which can

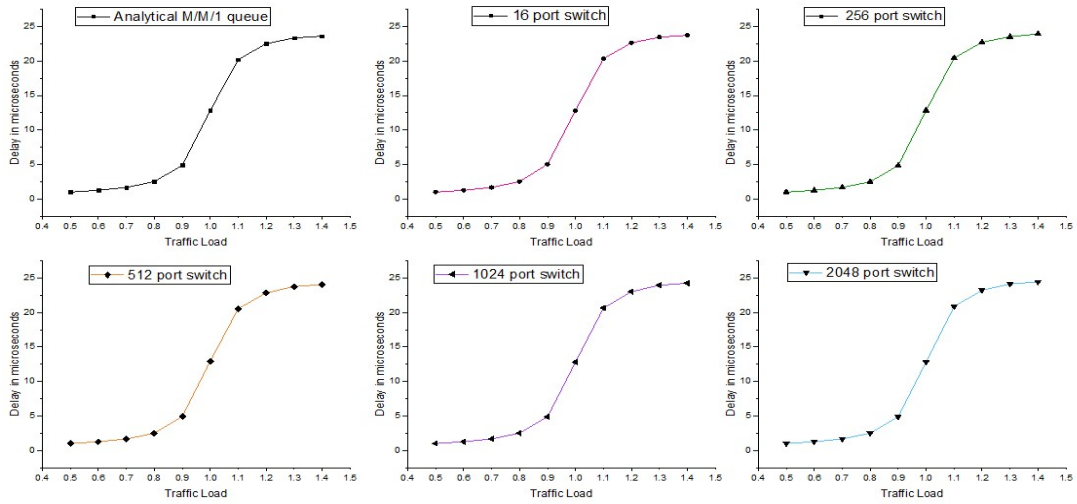

Figure 4.6: Delay Performance of Output Queued Switch Model with finite buffer

achieve a full 100% throughput and have a speedup of N where each output port can ideally receive packets from N input ports (N = no. of switch ports) simultaneously which boosts up the switching speed and gives minimal latency. (ii) Input queued switch (IQ) is the other commonly used solution which has queues at the input ports, but it can lead to the Head of Line blocking problem where due to a contention with the first packet in the queue, even if the next packet is without contention, it cannot be transmitted without breaking the FIFO queue order. Having virtual output queues at the input can solve this problem and provide considerable throughput performance with a proper scheduling algorithm. But IQ/VOQ switches have a speedup of *1* which is not practically suited for low latency DCNs and moreover, the number of buffers required are also more (N^2) than the OQ switches which needs only N buffers and no scheduler to handle packet contention. So to have a minimal latency in line with the requirements of recent DCN architectures, the network simulation was built with OQ switches which require fewer queues than VOQ switches and so will also be faster in simulating larger networks due to less number of network modules.

The switch model was compared with analytical models for M/M/1 and M/M/1/K [139] queues for varying switch radix. Poisson traffic [140] with exponential inter-arrival time and exponential service time distribution was used and the link rate was considered to be 10 Gbps and the average packet length as 625 bytes. The delay performance of the switch for various port counts and varying traffic load is given in Figure 4.5 which relates to a M/M/1 model and Figure 4.6 which relates to a M/M/1/K model. The plotted graphs show that the delay performance from the various switch configurations

closely match each other and also coincide with the analytical queuing model approximations.

4.5.3 Network Architecture

The output queued switch was used as the main building block as already mentioned and apart from that the other network parameters are described in this subsection:

Oversubscription should be ideally 1:1, but that requires lots of switches and cabling involving extra cost and often such high bandwidth remains under-utilised. So, DCNs try to reduce the number of switches depending on traffic since all devices do not communicate simultaneously. Legacy DCNs using low radix switches can have oversubscription ratio as high as 1:20 but currently, oversubscription ratio of 1:3 is being used commercially [130] to have an acceptable number of Spine switches.

Link rates – With the introduction of advanced processor cores like Intel's Romley server platform, growth in cloud computing and virtualization have increased server bandwidth demands [141], leading to higher data rate migrations. However, Leaf-Spine data rate is dependent on the Spine switch type and for very large networks, high radix Spines with only 10 Gbps data rates are currently available [122, 123]. So 10 Gbps is the chosen all around data rate in the simulated network.

Buffer Size – Generally, two types of buffers are used in the DCN switches. The first type uses on-chip shared SRAM with approx. 12 MB of memory per 128 10G ports in an ASIC which translates to about 100 KB buffering memory per 10G port [142]. The other type uses external DRAM to enable deep buffering with approx. 100 MB buffer per 10G port, thus multiplying the capacity by almost 1K compared to the previous type. However simulating a large network with very big buffers is time-consuming and requires very high memory capacity. So buffers with approx 100 KB per port have been used for the majority of the simulations. Larger buffer sizes have been used in certain scenarios to judge the effect of buffer size variations in both Leaf and Spine switches.

Traffic – Uniform Poisson traffic with exponential inter-arrival and service time distribution is used for analysing the performance of the network where packets generated by the servers in a rack are sent to any server in any rack with uniform probability. As shown in Figure 4.7, servers act as a source for generating packets and sends the traffic to the Leaf switch. For intra-rack switching, traffic is then routed through the Leaf switch and for inter-rack switching, traffic is routed through the Spine switch chosen randomly to implement Valiant Load Balancing type routing [19, 47]. Since servers can transmit packets with similar probability to all racks, a randomised routing helps to spread traffic

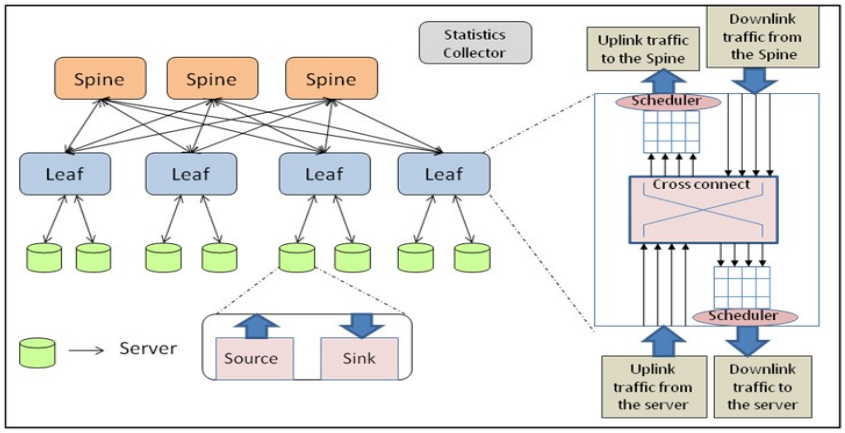

Figure 4.7: Simulation model for Leaf Spine Architecture

uniformly independent of the destination and prevents any sort of hot spot formation in the Spine layer while balancing the load across all available paths. From the Spine switch packets are transferred to the Leaf switches in the next hop as per the destination server.

4.6 Analytical Modelling of the Designed Network

An analytical M/M/1/K model equivalent to a Leaf-Spine network was implemented in Java. Performance analysis was targeted towards the more time consuming inter-rack communication where packets generated by a server are sent to any server but in a separate rack with uniform probability to ensure that the intra-rack packet departures at the Leaf switch do not affect the packet arrival time for the proceeding layers. The Leaf-Spine analytical network can be modelled as a stack of interconnected queues where packets pass through a series of nodes to reach the destination. Each switching node can be considered to be composed of a number of queues whose output can be further connected to multiple queues in the proceeding switching stage and packet arrivals can be exogenous (external traffic flow) or endogenous (relay traffic transmitted between nodes) in nature. So, in inter-rack communication, the uplink traffic arriving at the Leaf from the servers will be the only exogenous traffic entering the network while all traffic arrivals in the following stages will be internal relay traffic. Given the structure of the Leaf-Spine network, the rate of traffic arrival in the uplink and downlink paths will be dependent on the oversubscription introduced in the Leaf layer and the uplink traffic rate is bound be more intensive for inter-rack traffic arrival in an oversubscribed network.

4.6.1 Analysis of Traffic Arrival Rate and Delay

In a tandem queuing network, λ is the cumulative exogenous traffic arrival rate for layer1 queues. On receiving service at node i, a packet can proceed to node j with a deterministic probability r_{ij}, where both the relay and external flows are aggregated. Considering that each layer1 queue with arrival rate λ, sends traffic to 'x' layer2 queues with equal probability, $y(\lambda/x)$ will be the cumulative endogenous traffic arrival rate in each layer2 queue receiving traffic from $'y'$ queues of the preceding layer.

When queues are arranged in tandem, arrival time of packets in a queue is correlated to the departure time in the preceding queues which makes it difficult to make an analytical approximation. However, Kleinrock's Independence Approximation [143] considers that merging several packet streams on a single transmission line cancels out the interdependence of inter-arrival and service times, making them independent of each other. So, a M/M/1 model can be used to analyse the transmission pattern across each link. Furthermore, Burkes Theorem [144] states that for a M/M/1 queue in the steady state, if Poisson traffic arrives at a node with an arrival rate λ then it also departs from the node with a departure rate λ while still retaining it's Poisson nature and at any intermediate time t, the number of packets in the queue is independent of the departure of the preceding packets prior to t.

So, the Leaf-Spine data centre network with Poisson traffic can be considered to be analytically equivalent to a tandem M/M/n/K (when n = 1 and K is infinite) queue model which is also in accordance to the open Jackson queuing network [145] or an open network of n M/M/n state-independent queuing system with unlimited queue capacity at each node and having the following features.

1. There is only one class of packets arriving to the system.
2. Exogenous packets arrive at node j according to a Poisson process with rate $\lambda_j \geq 0$.
3. The service times of the packets at j^{th} queue are exponentially distributed with mean $1/\mu_j$.

Considering *r* to be the nxn probability matrix (where n is the no. of nodes) describing the routing of packets within an open queuing system [145], γ = (γ_1, γ_2,..., γ_n) can be termed as the mean arrival rates of the relayed packets, and λ = (λ_1, λ_2,...., λ_n) can be the mean arrival rates of the exogenous packets. Using the flow conservation principle while assuming that the network reaches equilibrium, the cumulative arrival rates entering the system will be equal to the total departure rate under the steady-state conditions. So, the traffic equation can be written as,

$$\gamma_j = \lambda_j + \sum_{i}^{n} \gamma_i r_{ij}, \qquad j = 1, 2, ...n. \tag{4.10}$$

In steady state, considering a stable network, the aggregate input traffic arrival rate γ_j into the node j will be equal to the combined output rate from the node i in the preceding stage, where $i = 1, 2, .., n$. Thus it can give a system of n equations and n unknowns in the matrix form as shown below,

$$\vec{\gamma} = \vec{\lambda} + \vec{\gamma}\mathbf{R} \tag{4.11}$$

Once the aggregate arrival rate into an individual node is available, the network is simplified into a chain of tandem queues where every node represents an independent queuing system with Poisson input and follows the average delay pattern of the corresponding M/M/1 queue given in eq. 4.12 [139]. Assuming the exogenous traffic flows to be independent Poisson processes, the service times are assumed to be mutually independent of the arrival process at that queue.

$$\overline{T} = \frac{\overline{N}}{\lambda} = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\lambda} = \frac{1}{\mu - \lambda} \tag{4.12}$$

where $\overline{T}$ is the average time spent per packet in a node or the node latency, $\overline{N}$ is the average number of packets in a node and $\rho = \frac{\lambda}{\mu}$ is the traffic load intensity.

However, on introducing oversubscription in the network, the queues can get overloaded and the buffer length needs to be restricted to control the latency from being infinite (when the offered load $\geq$ 100%). So, the previous M/M/1 model is no longer analytically equivalent to the network where the possibility of packet loss, makes the departure process from queues to be non Poisson in nature. The departure rate from each queue is no longer the same as the arrival rate λ. A packet arriving at node i will not necessarily join the queue and may be dropped if the queue is full. So, a type of reduced load approximation [139] is used here to estimate the effective arrival rate for the intermediate stages using the eq. 4.13,

$$\lambda_{eff} = \lambda(1 - p_K) = \lambda\left(1 - \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}\right) \tag{4.13}$$

where λ is the arrival rate into a node i which is dependent on the departure rate of the preceding queues connected to node i, p_K is the blocking probability and $K-1$ is the buffer length or K is the net system storage capacity. λ_{eff} is the effective arrival rate to a particular queuing stage for the packets that successfully join the queue. The departure rate from the node i is thus equivalent to the effective arrival rate (λ_{eff}) considering the packet loss. So, the overloaded network can now be considered to be analytically equivalent to a tandem queue model with reduced load approximation, where every node represents an independent M/M/1/K queuing system. It is to be noted that the departure process can still be assumed to be approximately Poisson in nature, which should be a reasonably accurate assumption when the loss rate is low. The model thus follows the average delay

pattern as given by eq. 4.14. λ_{eff} gives a measure of the throughput of node i since all packets that enter the queue are processed by the node i. So, the effective load or the system utilisation is given by λ_{eff}/μ.

$$\overline{T} = \frac{\overline{N}}{\lambda_{eff}} = \frac{\overline{N}}{\lambda(\frac{1-\rho^K}{1-\rho^{K+1}})} \tag{4.14}$$

4.6.2 Leaf Spine Analytical Queueing Model

A Leaf Spine network (as in Figure 4.8) was analysed using uniform inter-rack Poisson traffic with varying oversubscription ratios and the used network parameters are mentioned in Table 4.5.

No. of Spine	No. of servers per rack	Oversubscription	Link Rate	Average Packet Length	Queue scheduling type
Sp	N	Sp/N	10 Gbps	625 bytes	FIFO

Table 4.5: Summary of Analytical Model Parameters

Due to the nature of the topology, uplink connections from the Leaf tend to suffer from heavy traffic and difference in oversubscription introduces variation in the level of traffic offloading between Server-Leaf and Leaf-Spine uplinks affecting the arrival rate of packets in the intermediate switching nodes. When oversubscription is introduced, the arrival rate in the uplink queues increase and the system starts getting overloaded, leading to increased delay and after a certain point latency tends to infinity. Finite buffer length can resolve this problem but introduces packet loss. So, two different scenarios have been tested where the network is perfectly balanced or over-subscribed to compare the variation in traffic arrival rates for the analytical tandem network model.

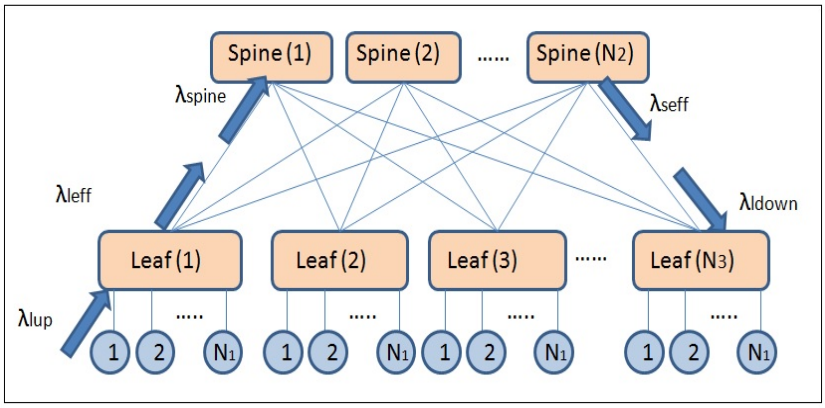

Figure 4.8: DCN model for analytical comparison

4.6.2.1 Balanced Network

An optimal oversubscription ratio of 1:1 is used in case of the balanced network. Hence traffic from the Server-Leaf uplinks are evenly distributed across the Leaf-Spine uplinks without any contention and so the queues in the different switching stages behave uniformly and exhibit equivalent latency figures. Due to lack of oversubscription, this network behave logically equivalent to a tandem M/M/1 queue network.

Arrival Rate Calculation:

Based on Figure 4.8, λ_s is the arrival rate for Server-Leaf traffic streams, where each Leaf switch receives traffic from N_1 servers and has N_2 output queues. So, net arrival rate in each Leaf queue is (N_1*λ_s)/N_2. Each Leaf queue output is directed towards a particular Spine which can receive traffic from N_3 separate Leaf switches. But the net traffic in each Spine also gets divided into N_3 output queues. So, the net arrival rate for each of the downlink traffic queue from the Spine to the Leaf is N_3*($\frac{N_1\lambda_s}{N_2}$)/N_3 or $\frac{N_1\lambda_s}{N_2}$. Each Spine switch output queue is directed towards only one Leaf switch which in return receives traffic stream from N_2 separate Spine switches. But the net traffic in each Leaf gets divided into N_1 separate output queues towards N_1 separate servers. So, the net arrival rate for each downlink traffic queue connected from the Leaf to the Server is N_2*($\frac{N_1\lambda_s}{N_2}$)/N_3 or λ_s. Based on these calculations, the traffic arrival rates for the simulated network are listed in Table 4.6.

Leaf Uplink Arrival Rate (λ_{lup})	**Leaf Uplink Departure Rate** (λ_{leff})	**Spine Arrival Rate** (λ_{spine})	**Spine Departure Rate** (λ_{seff})	**Leaf Downlink Arrival Rate** (λ_{ldown})
$\frac{N_1\lambda_s}{N_2} = \lambda_s$	$\lambda_{lup} = \lambda_s$	$\frac{N_1\lambda_s}{N_2} = \lambda_s$	$\lambda_{spine} = \lambda_s$	λ_s

Table 4.6: Summary of Traffic arrival rates for individual queueing stages in Balanced Network (1:1 Oversubscription) where $N_1 = N_2$

Analytical results for the Leaf Switch uplink queue, Spine switch queue and Leaf switch downlink queue are calculated based on the M/M/1 model with corresponding inter-arrival rates as mentioned in Table 4.6 and the simulation results are plotted in Figure 4.9. It is observed that the simulation results for all the three individual queuing stage matches with the analytical results. As the network is balanced without any traffic overloading across any link, all queueing stages shows uniform latency as expected due to equivalent traffic flows.

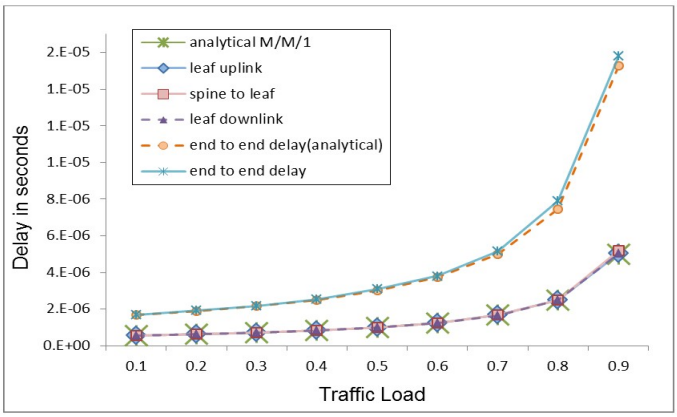

Figure 4.9: Delay performance analysis for Balanced Network

4.6.2.2 Oversubscribed Network

An oversubscription ratio of 1:A is introduced in the network with buffer lengths of approx 100 KB per switching port as used by switches with standard SRAM buffers [142]. So, there are less Leaf-Spine links than there are Server-Leaf links connected to the Leaf switch and all queues in the Leaf uplink switching stage experiences contention due to greater traffic load, leading to increased latency. Since the arrival rate in the Spine layer is dependent on the departure rate from the Leaf layer so the queuing stage in the Spine layer does not behave identically to its preceding stage because of the introduction of buffers in the network which leads to packet dropping. So, a reduced load approximation is used to estimate the effective arrival rate for the Spine and Leaf downlink queue based on the arrival rate of the preceding stages.

Arrival Rate Calculation:

Based on Figure 4.8, λ_s is the arrival rate for Server-Leaf traffic streams, where each Leaf switch receives traffic from N_1 servers and has N_2 output queues. So, the net arrival rate in each Leaf switch queue is $\lambda_{lup} = (N_1\lambda_s)/N_2$. The departure rate from each Leaf uplink queue is $\lambda_{leff} = \lambda_{lup}(1\text{-}p_{Kleaf})$. Each Leaf queue output is directed towards a particular Spine switch which in return receives traffic stream from N_3 separate Leaf switches. But the net traffic in each Spine also gets divided into N_3 separate output queues. So, the net arrival rate for each Spine-Leaf downlink traffic queue is $\lambda_{spine} = N_3*(\lambda_{leff})/N_3 = \lambda_{leff}$. The departure rate from each Spine queue is $\lambda_{seff} = \lambda_{spine}(1\text{-}p_{Kspine})$. Each Spine switch output queue is directed towards only one Leaf switch which in return receives

traffic streams from N_2 separate Spine switches. But the net traffic in each Leaf gets divided into N_1 output queues towards the server. So, the net arrival rate for each downlink traffic queue from Leaf to Server should be λ_{ldown} = (N_2*(λ_{seff}))/N_1. Based on these calculations, the traffic arrival rates for the simulated network are listed in Table 4.7.

Leaf Uplink Arrival Rate (λ_{lup})	**Leaf Uplink Departure Rate** (λ_{leff})	**Spine Arrival Rate** (λ_{spine})	**Spine Departure Rate** (λ_{seff})	**Leaf Downlink Arrival Rate** (λ_{ldown})
$\frac{N_1\lambda_s}{N_2} = 3\lambda_s$	λ_{lup}(1-p_{kleaf})	$\frac{N_3\lambda_{leff}}{N_3} = \lambda_{leff}$	λ_{spine}(1-p_{kspine})	$\frac{N_2\lambda_{seff}}{N_1} = \lambda_{seff}/3$

Table 4.7: Summary of Traffic arrival rates for individual queueing stages in Oversubscribed Network (1:3 Oversubscription), where $N_1 = 3N_2$

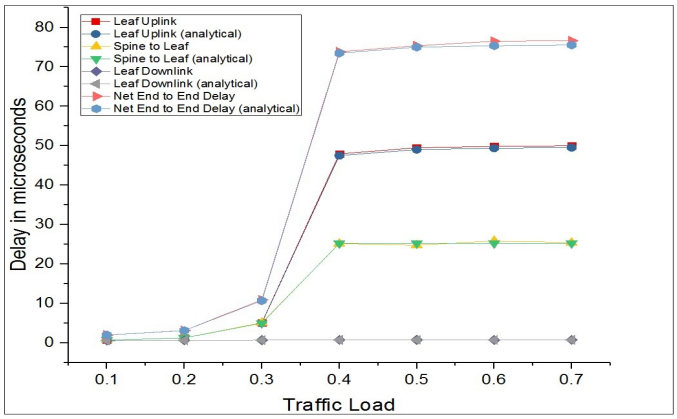

Figure 4.10: Delay performance analysis for Oversubscribed Network

Analytical results for Leaf Switch uplink queue, Spine switch queue and Leaf switch downlink queue are calculated based on the M/M/1/K tandem network model with inter-arrival rates as mentioned in Table 4.7 and the simulation results are plotted in Figure 4.10. Due to the oversubscription introduced, the network tends to become overloaded (arrival rate $\geq$ service rate) as the server traffic load increases (particularly as $\rho \geq 0.4$). Thereafter the delay surges to quite an extent as the arrival rate in the network tends to be more than the service rate. Arrival rates in Leaf uplink queues and the Spine queues no longer remain the same due to limited buffer length and as observed from the traffic rates in Table 4.7 the Leaf uplink stage experiences the maximum traffic offloading which leads to contention within the buffers. So, the delay experienced by the Leaf uplink queue rises with the increase in buffer occupancy due to the oversubscription introduced in the Leaf switch layer. The effective departure rate (λ_{eff}) at each stage has been calculated to estimate the arrival process for the next stage. The plotted delay for the different stages (Figure 4.10) show that all the queues behave similarly to the

equivalent analytical estimation. Mixing of multiple traffic streams in the endogenous flow arriving at the second and third switching stages leads to a close match with analytical results and reduces dependencies amongst the queues as per Kleinrock Independence Assumption [143].

4.6.3 Estimation of Analytical Parameters for Leaf-Spine Queueing Network

Service Rate	Inter-Arrival Rate	Server Load	Oversubscription	Mean No. of packets in the system	Blocking Probability
μ	λ	$\rho = \lambda/\mu$	1:S	N_q	P_{kq}

Table 4.8: Basic Network Parameters (where $q = 1, 2, 3$ denoting the queueing stage)

As observed from the analytical modeling, the arrival rates for any intermediate stage becomes dependent on the departure rate of the previous stage when oversubscription is introduced in the network, which leads to overloading. So, calculating the arrival and departure rates of traffic streams is essential for determining different network performance parameters. Considering a three staged Leaf-Spine queueing model with Poisson inter-rack traffic, shortened formulae for calculating the traffic flow rates and network parameters for all the queueing stages have been presented in this subsection. Servers act as the source for initiating the traffic flows in the multi-staged tandem network and the basic server and network parameters used in all further calculations are listed in Table 4.8.

Leaf Uplink Queue

Traffic flows from multiple servers, aggregate at the Leaf switch queues to be further directed towards the Spine. However, all the packets reaching the Leaf are not forwarded to the Spine due to packet loss in the Leaf queues. Formulae derived for this queueing stage are given in eq. 4.15–4.18.

$$\boldsymbol{Traffic\ Arrival\ Rate = \lambda_1 = S\lambda} \tag{4.15}$$

$$\boldsymbol{Traffic\ Load = \rho_1 = (S * \lambda/\mu) = S\rho} \tag{4.16}$$

$$\boldsymbol{Traffic\ Departure\ rate = \lambda_{d1} = \lambda_1(1 - P_{k1}) = S\lambda(1 - P_{k1})} \tag{4.17}$$

$$\boldsymbol{Delay = N_1/\lambda_{d1}} \tag{4.18}$$

Spine Queue

Upward traffic flows from multiple Leaf switches, queue up in the Spine to be directed downwards to various Leaf switches. However, all packets reaching the Spine are not forwarded further due to

packet dropping in overloaded buffers. Formulae derived for this queueing stage are given in eq. 4.19–4.22.

$$\boldsymbol{Traffic\ Arrival\ Rate = \lambda_2 = \lambda_{d1} = S\lambda(1 - P_{k1})} \tag{4.19}$$

$$\boldsymbol{Traffic\ Load = \rho_2 = (S\lambda(1 - P_{k1})/\mu) = S\rho(1 - P_{k2})} \tag{4.20}$$

$$\boldsymbol{Traffic\ Departure\ rate = \lambda_{d2} = \lambda_2(1 - P_{k2}) = S\lambda(1 - P_{k1})(1 - P_{k2})} \tag{4.21}$$

$$\boldsymbol{Delay = N_2/\lambda_{d2}} \tag{4.22}$$

Leaf Downlink Queue

Before reaching the destination servers, the downward traffic flows from multiple Spines are merged into the Leaf queues where certain packets are dropped depending on the buffering capacity. Formulae derived for this queueing stage are given in eq. 4.23–4.26.

$$\boldsymbol{Traffic\ Arrival\ Rate = \lambda_3 = \lambda_{d2}/S = \lambda(1 - P_{k1})(1 - P_{k2})} \tag{4.23}$$

$$\boldsymbol{Traffic\ Load = \rho_3 = (\lambda(1 - P_{k1})(1 - P_{k2}))/\mu) = \rho(1 - P_{k1})(1 - P_{k2})} \tag{4.24}$$

$$\boldsymbol{Traffic\ Departure\ rate = \lambda_{d3} = \lambda 3(1 - P_{k3}) = \lambda(1 - P_{k1})(1 - P_{k2})(1 - P_{k3})} \tag{4.25}$$

$$\boldsymbol{Delay = N_3/\lambda_{d3}} \tag{4.26}$$

4.7 Performance Analysis of DCN

While analysing the performance of the Leaf-Spine network modelled with OMNeT++, only the delay within the switching network is taken into account and the additional delay that might be experienced due to processing in the host stack is not considered. Further, the minimal propagation delay (only a few ns per 1m of fiber) incurred across the fiber has also been ignored.

4.7.1 Buffer size variation

Placement and size of buffers can be crucial for desired network performance. 1:3 oversubscription with 10 Gbps link rates has been used for this evaluation. Two sets of buffer capacity have been analysed in four different combinations of the Leaf and Spine switches. Buffer sizes are either 20 KB

or 1000 KB approx. per 10 Gbps port and the network is simulated with uniform Poisson traffic to determine the effectiveness of buffer size as per its placement in the edge or core layer. As evident from Figure 4.11 the rate of packet loss differs with variation in the placement and size of buffers. Using large buffers in both switches result in minimum packet loss and using small buffers in both switches result in maximum packet loss as expected. However using smaller buffers in the Leaf in contrast to the Spine degrades the performance to some extent and results in more packet loss as compared to having a bigger buffer in the Leaf and a small one in the Spine switch. Larger buffer length can ease packet drops in the switch but at the same time increases queuing delay which leads to increase in network latency as evident in Figure 4.12.

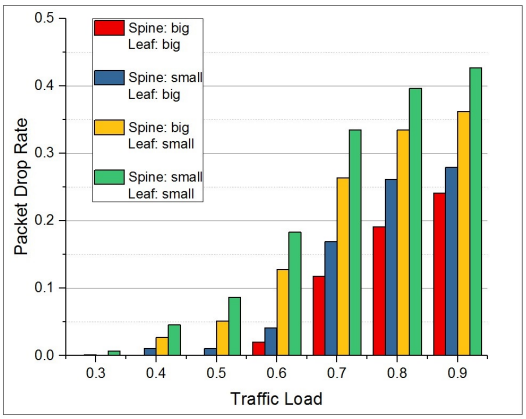

Figure 4.11: Variation of packet drop rate with different buffer combinations

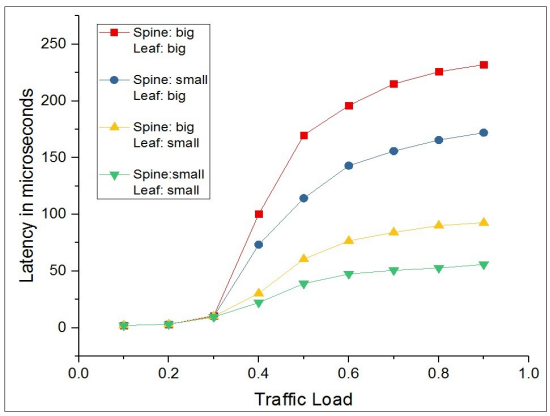

Figure 4.12: Latency distribution with variation in buffer size

This variation in latency and packet loss happens due to contention occurring in both the fabric ports (uplinks) as well as the server ports (downlinks) of the Leaf switch as pointed out in [25, 146]. Spine switches have contention only across the downlinks which operate at the same link speed as that of the Leaf uplinks but the later has additional contention across the Leaf-Server downlinks as well. The major traffic overload is due to the oversubscribed Leaf switching layer which increases the load in the Leaf uplinks directed towards the Spine when it tries to distribute the traffic across a small number of Spine switches leading to packet drops. As seen in the analytical model mentioned in Subsection 4.6.2.2, latency across the Leaf-Spine uplink is more than the Spine-Leaf downlink in oversubscribed networks due to the greater buffer occupancy in the uplink queues of the edge layer. So for optimal performance, it is necessary to have big buffers in both the core and edge layers instead of concentrating on the core only. If bigger buffers are solely allocated for the core layer then it might be difficult to control the traffic bursts in the Leaf layers resulting in substantial packet drops. Very

large buffers are especially useful in case of heavy traffic loads for handling traffic bursts that may occur due to change in link rates or due to oversubscription. But usually (around 95% of the time), traffic in DCNs does not exceed 10% of the maximum network capacity [15] and for more than 99% of the time, traffic load (ρ) is ≤ 0.3 [147]. So, when the load ratio (ρ) is ≤ 0.3, having deep buffers in both layers can eliminate packet loss and assure latency within few μs (as evident in fig 4.12) which is acceptable as per DCN standards and even if in the remaining 1% of time the traffic load is > 0.3, the results confirm that, the network can handle the data centre traffic and limit packet loss if deep buffers are used in both the switching layers or at least in the Leaf layers.

4.7.2 Variation of Oversubscription

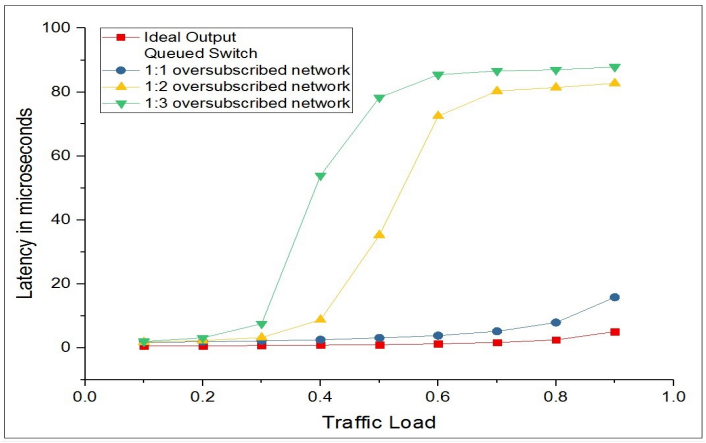

Figure 4.13: Variation in network latency with changing oversubscription factor

For evaluating the performance with varying degree of oversubscription, the network was simulated with 10 Gbps links and compared with an ideal non-blocking uniform switch network. As already discussed in Section 4.2, Leaf-Spine networks can be broadly compared with very large Output-Queued switches having the same number of ports as the number of servers in the network. Minimum latency is observed in case of the ideal OQ switch with infinite buffer capacity, but the non-oversubscribed network also delivers a similar performance when compared to the non-blocking OQ network. Figure 4.13, shows that when the network capacity is not bandwidth constrained, decreasing the oversubscription helps in improving load balancing across the network resulting in better performance. Lowering the number of Leaf-Spine uplinks compared to the Leaf-Server downlinks leads to contention and greater buffer occupancy in the edge layer and so network latency increases on introducing oversubscription in the network. However, at low traffic intensity, particularly at less than 30% traffic load

it can be observed that the oversubscribed network performs at par with the non-oversubscribed ones because below that load, arrival rates remain comparatively less than service rates.

As shown in the analytical models mentioned in Subsection 4.6.2, in a non-oversubscribed system, all three queueing stages behave almost identically, equivalent to a regular M/M/1 queue and hence latency remains quite minimal. On introducing an oversubscription factor of 1:x, the arrival rates in the uplink queues becomes x times that of the downlink Leaf queues and so, latency increases to a great extent causing a huge gap in performance between the oversubscribed and non-oversubscribed networks. As the traffic load increases beyond 30% for 1:3 oversubscribed network and beyond 40% for 1:2 oversubscribed network the system tends to become unstable leading to a sharp rise in latency caused by contention and heavy buffer occupancy in the constrained queues.

Ideally, non-oversubscribed networks can provide the best performance due to minimal buffer occupancy but it is not realistic to use a large number of Spine switches especially when the usual traffic load (approx 30% as mentioned in [147]) can be low enough to not to cause any buffer contention. It can be pointed out that in an oversubscribed network, increasing the link speed between the two switching layers can upgrade the overall performance. As the link speed rises between the two switching layers, it leads to a speed jump for packets moving towards the Spine from the Leaf switch layer and this can naturally lead to a drop in latency even in an oversubscribed network. So, to achieve enhanced network performance it is necessary to increase the link speed across the Leaf and Spine fabric network and minimise oversubscription when the traffic intensity tends to be high.

4.8 Conclusion

This chapter is focused towards the evaluation and analysis of an electronic DCN, and Leaf-Spine architecture was chosen as per the current trend for topological analysis of the network to have a fair estimation of the recent standards of scalability, cost and power consumption figures. Leaf-Spine networks can be related to multi stage tandem queued networks with individual nodes/switches further considered to be composed of a series of queues. Based on this concept, a detailed analytical modelling with Kleinrock Independance Assumption and Reduced Load Approximation was introduced for evaluating the performance of any generic Leaf-Spine DCN architecture and the theoretical values shows near perfect match with the network simulation results. The designed network and the analytical model were both tested against different levels of blocking in the network to judge the variation in

traffic arrival and offloading at all the intermediate queueing stages and based on the findings, closed form expressions for traffic arrival and departure rates as well as delay from individual switching nodes were derived.

DCN network was investigated with different variations of buffer placements in Leaf and Spine switches and with different oversubscription figures while comparing it to a uniform non-blocking switch. From the simulation results it is observed that when networks are oversubscribed, significant traffic overload is created in the Leaf-Spine uplinks, initiating blocking. Results also confirm that having larger buffers in the Leaf and smaller in the Spine can perform better than having smaller buffer in the Leaf and larger buffer in the Spine switches. Non-oversubscribed network showed similiar delay performance when compared to an ideal non-blocking OQ network and the oversubscribed networks showed matching results until traffic overloading was introduced. From this analysis it can be inferred that increasing the uplink data rates from the traffic burdened Leaf layer, can significantly help to improve network performance by initiating faster traffic offloading. All simulations in this chapter dealt with Layer 2 switching, but for a better understanding of the data centre environment, the next chapter will deal with TCP traffic which is the prevalent communication protocol in DCNs.

Chapter 5

Modelling and Evaluation of TCP Traffic

5.1 Introduction

The bulk of DCN traffic emanates from the transport layer comprising of both TCP and UDP flows. As congestion control is critical for most data centre applications, such as distributed filesystems, cluster computing, parallel databases and disaster recovery, the majority of the packets transferred in DCNs are from TCP flows (82% of packets, according to one study in [148]). Hence, in this chapter, TCP traffic is chosen for studying the transport layer behaviour in modern DCNs.

The inherent nature of TCP whereby it supports reliable and ordered bidirectional transmission and provides congestion control and fairness among different flows is relevant in WANs and LANs. However, the traffic characteristics in DCNs being different from other communication networks require newer versions of TCP to cater for their specific performance requirements. In this chapter, the TCP bottlenecks are studied and a Leaf-Spine network is designed to judge the impact of Layer 4 switching in data centre environment, with a standard TCP variant and a DCN specific TCP variant (DCTCP [149]). Through network simulation it is validated that DCTCP can give a better performance in comparison to standard TCP due to their inbuilt ECN marking scheme. However, TCP can give a similiar performance to DCTCP in smaller networks having less concurrent traffic flows. Network performance analysis further reveals that the use of a smaller congestion window, in DCTCP can lead to frequent C_{wnd} resizing, which lowers the throughput than the link bandwidth. The novelty

of the designed network lies in the fact that it can use traffic flows generated internally by customised Layer 4 OMNeT++/INET [150] modules as well as have provisions for flows generated externally by a realistic traffic generator (DCT^{2}Gen [151]). The developed network model can provide further scope for facilitating TCP traffic implementation in other network topologies by changing the layout of the network hosts/routers that interfaces with the INET custom modules.

5.2 TCP in Data Centres

The diverse mix of DCN traffic can generate both small request/response type flows for supporting interactive applications as well as large flows with bulky data replication and data shuffling traffic [152]. Hence it is essential to maintain a high burst tolerance and minimal delay for the latency sensitive short flows and high throughput utilisation for the long data flows [16, 19]. Commercially deployed topologies such as Fat Tree and Leaf-Spine networks are characterised by the availability of very high-speed links, regularity in topology [153] and minimal propagation delay. This leads to a reduced round trip time (RTT) in DCNs which is fairly lower than the TCP RTT in wide area networks. Subsequently, the bandwidth delay product in DCNs is smaller than regular TCP traffic [16] and commodity switches used in data centres usually have shallow buffers for minimising latency.

5.2.1 Problems related to TCP traffic in Data Centres

Implementing standard TCP protocol in data centres does not ensure the desired network performance due to the inherent difference in DCN workload from other IP network traffic patterns and the main challenges faced by such TCP implementation in data centres are summarised below.

5.2.1.1 Minimising Latency

Connecting nodes inside the DCN can be placed only a few hundred meters apart generating a negligible propagation delay [155] and TCP segments (as per standard Ethernet MTU of 1500 bytes) experiences only 1.2μs serialisation delay when transmitted through 10 Gbps links. So, queuing delay in switches can be considered as the main reason behind the undue latency in DCNs which can adversely affect the user experience [154]. Elephant flows deliver the majority of traffic, so when both mice and elephant flows enter the same bottleneck switch, the mice flow ends up queuing behind several packets of the elephant flow [152] and can be dropped frequently. A smaller queue length can alleviate

the latency problem. However, traditional TCP congestion control is based on packet loss. So, the nodes transmitting elephant flows continue to send TCP segments until the buffers overflow and loose packets leading to packet retransmission and queuing delay for the mice flows [152, 153].

5.2.1.2 TCP Incast

In a network characterised by low latency and high throughput, when multiple nodes simultaneously try to communicate with a single receiving node, the cumulative traffic overflows the buffer of the switch connected to the receiving server leading to the TCP Incast problem which causes severe congestion and performance degradation. This type of many to one traffic pattern is quite prevalent in DCN applications such as MapReduce (where several mappers send data to a single reducer), distributed storage systems (where several storage nodes communicate with one request node), web search queries etc. [152, 153]. The loss of packets causes TCP retransmission timeouts and triggers TCP Incast scenario whereby packets are dropped at the Ethernet switches, blocking the reception of the duplicate ACKs at the sender. As a result, the fast retransmit and recovery phase is not triggered and the network waits for retransmission timeouts at the servers adding up substantial delay to the network. This whole phenomenon wastes network resources, causes link underutilization and can lower the goodput (throughput at the receiving server application side) by up to 90% [153].

5.2.1.3 TCP Outcast

When using TCP in DCNs, the throughput of flows with small RTT might turn out to be much lower than the throughput of flows with longer RTT, leading to the Outcast problem. The main reason behind it is the Port Blackout phenomenon [153] which occurs when several packets enter a switch from different input ports to be routed to the same output port. In commodity switches with shallow buffers, some of the contending packets succeed in entering the buffer while the rest are discarded once the buffer starts to overflow. If multiple servers transmit flows to a single receiver at the same time, flows with a shorter RTT arrive ahead of the flows with a longer RTT and the buffer tends to be filled up with such short flows [148]. However, once a single buffer slot becomes available and a packet from the flow with a longer RTT manages to get inside, any packets arriving from any of the shorter flows will be subsequently dropped until the switch finishes transmitting the longer RTT flow. As a result, the throughput of the shorter flows is degraded and it leads to severe unfairness.

5.2.2 DCTCP (Data Centre TCP) Congestion Control

Data Centre TCP (DCTCP) is one of the earliest TCP protocols that was designed to meet the performance demands of DCN applications [149]. It is an ECN (Explicit Congestion Notification) and AQM (Active Queue Management) congestion control technique based on the TCP Reno [156] variant. Commercial availability of ECN enabled switches makes DCTCP implementation feasible and it has been considered as a baseline protocol based on which newer protocols are being continuously proposed for DCNs. DCTCP aims to lower queueing delay and provide high burst tolerance for the bursty mice flows and high throughput for the bulky elephant flows [152, 153]. Congestion control in standard TCP is based on packet loss and congestion notification comes after a severe queue has already built up, but ECN based marking in DCTCP notifies the sender when the queues are filled up beyond a predefined limit by setting the CE codepoint in the packet header. Subsequently, the receiving servers notify the sender of the building congestion by setting the ECN ECHO flag in the header and the senders react by controlling the congestion window as per the extent of congestion and not just by halving it as done in standard TCP. Thus, DCTCP reacts quickly to growing congestion levels and operates with 90% less buffer capacity [148, 149]. ECN features in DCTCP can be related to RED [157] whereby both the minimum and maximum buffer threshold are prefixed at the same value. With a shorter congestion window DCTCP can reduce sender transmission speed which can effectively alleviate queuing congestion and subsequent overflows and TCP retransmissions caused by the Incast problem. Minimal queue lengths in DCTCP provides fairness to both short and large flows and can alleviate the Outcast problem by not letting the queues to become full [16].

5.3 TCP Network Implementation

In order to evaluate the performance of TCP in DCNs, a Leaf-Spine network was designed with OMNeT++ and added with INET [150] framework modules with additional provisions for passing external flow files from a realistic traffic generator to generate TCP traffic as per requirement.

5.3.1 Network Modelling

Real world traffic originating from various DCN applications is dynamic in nature and requires efficient congestion control which leads to the introduction of switching with L4 (Layer 4) TCP protocol in the designed network. Congestion window (C_{wnd}) regulation at the sender side is the driving force

behind TCP congestion control and it controls the number of packets that can be in transit at any given point of time with its growth spread over a number of stages. TCP Reno model is a popular TCP variant supporting all the standard C_{wnd} phases such as Slow Start, Congestion Avoidance as well as Fast Retransmit and Recovery Phases and hence was chosen for TCP implementation. Even DCTCP [149] can be considered as a modification of the Reno variant with added AQM features.

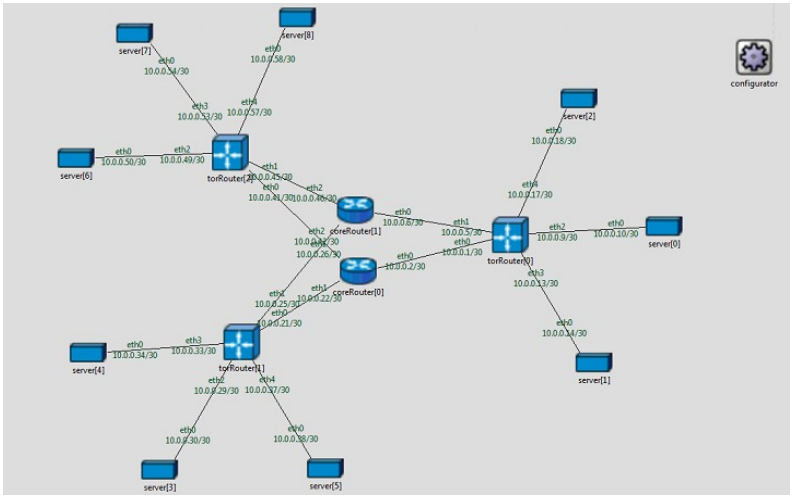

Figure 5.1: Basic Network Model

The OMNeT++/INET framework has an inbuilt TCP stack implementation and by using customised L4 Traffic Generator applications (explained in Subsection 5.3.2), TCP traffic transmission was made possible. The L4 standard hosts in INET were used to inherit the network nodes with added functionality to drive the custom traffic generator. L3 routers were modified for functioning as the Leaf and Spine switches respectively. Both the switches and servers had additional L3 and L4 features to drive the packet marking and acknowledgement facilities for DCTCP Congestion Window Control (as explained in Subsection 5.3.3.1). The number of servers, Leaf and Spine switches (depending on over subscription ratio), buffer lengths as well as the link rates all remained reconfigurable.

The routing protocol inherently used in INET modules is based on the Djikstra's algorithm [158] with shortest weighted routing pattern but since all inter-rack routes are equal hops in length in Leaf Spine topology, the L3 routing had to be redefined by changing the inbuilt routing protocol and routing tables. VLB Routing Protocol was used and the Spine switches were randomly chosen for packet transmission so as to spread traffic uniformly independent of the destination. A *vlbRoutingTable* module was linked with the Layer 3 interface of the switches to suppress the inbuilt routing mechanism and randomise the next destination depending on the switching stage. Random selection of the

intermediate switching helps to minimise the hot spot formation in the Spine layer and balances the load across all available paths. From the Spine switch, packets are transferred to the Leaf switches in the next hop as per the destination server.

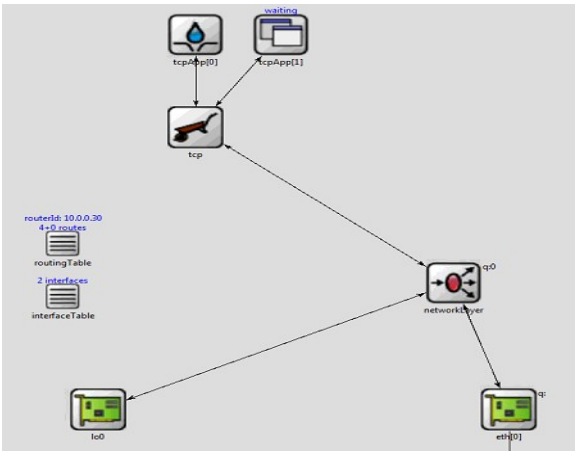

Figure 5.2: Server Model

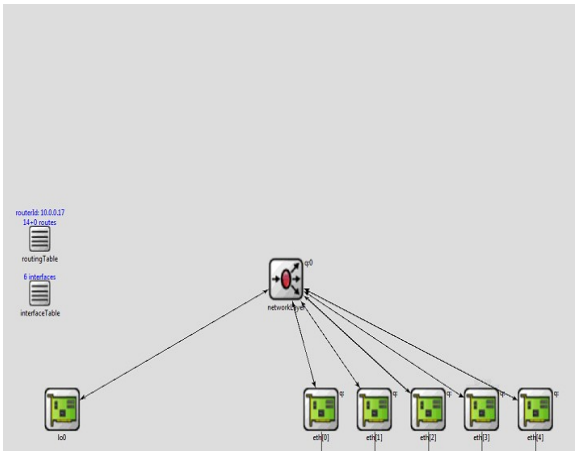

Figure 5.3: Router Model

Figure 5.1 shows the layout of the network model with all servers and routers having designated IP addresses for L3 packet forwarding and packet marking as required. The server used in the network model is shown in Figure 5.2 which demonstrates the layered TCP/IP stack. The application layer contains two TCP traffic generators serving as the source and sink respectively and driving the underlying transport layer protocol. The *tcpApp* in the application layer has been customised to generate traffic with a custom flow schedule as well as through random uniform allocation and the TCP interface in the transport layer has been modified to add an additional TCP variant (DCTCP). Figure 5.3 shows the basic router model. It can be seen that the network layer in both the servers and routers, help in VLB routing through customised routing tables while the Ethernet layer interconnects the different modules through reconfigurable Ethernet links.

5.3.2 TCP Traffic Generation

TCP traffic is generated in flows of varying size which are further divided into fixed size packets as per the MSS (Maximum Segment Size) of the network. A L4 *TCPTrafficGenerator* application module has been created which interfaces with the inbuilt TCP interface and accordingly drives the TCP flows into the *tcpApp* provided by INET. Traffic generation is made possible by two methods; by using the customised INET traffic generators or by using realistic traffic flow files generated externally whereby the custom built L4 Traffic Generator App can parse the .csv files produced by DCT^{2}Gen [151] and

accordingly feed the traffic flows into the TCP interface of the network.

5.3.2.1 DCT^{2}Gen Traffic Generator

The Data Centre TCP Traffic Generator (DCT^{2}Gen) uses a set of L2 traffic distributions to generate L4payload traffic. If the generated L4 traffic is transported using TCP, the resulting L2 traffic pattern is expected to comply with the observed L2 traffic from which the distributions are derived [151]. This generator uses the rack-awareness feature of modern data centres running applications such as Hadoop, MapReduce etc., which keeps as much traffic in the same rack as possible to achieve a higher throughput and lower latency. TCP can be bidirectional in nature but for simplicity in this generator, it has been considered as non-interactive/asymmetric in nature, where the payload is transported only in a single direction because available DCN traces for interactive TCP connections does not provide information about the relationship between the amount of sent and received data per host.

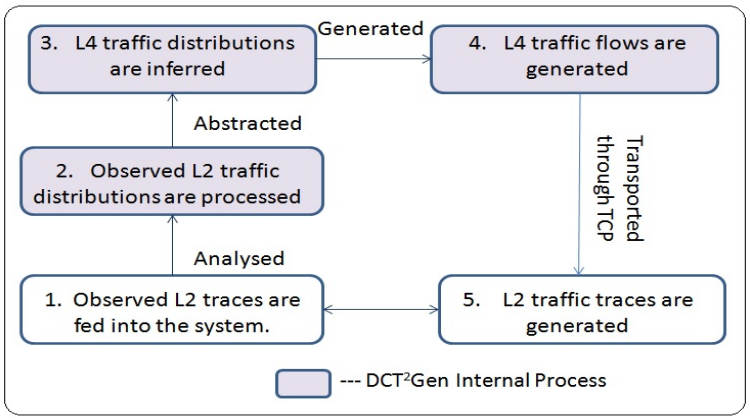

Figure 5.4: DCT^{2}Gen Traffic Generation Process

Figure 5.4 shows the flow diagram of this traffic generator which uses L2 traffic distributions derived from the realistic data centre traffic patterns reported in [15, 16] and accordingly abstracts the L4 payload distribution. The used traffic distributions are listed below:

- Number of observed inter and intra rack communication partners per node.
- Observed amount of bytes exchanged between a pair of nodes in inter rack and intra rack communication.
- Observed flow size and observed flow inter-arrival times.

After subtracting the required ACK from the L4 traffic, the desired traffic matrix (tm) is generated

and then flows are scheduled for each source-destination pair, which when transmitted through TCP results in the observed distribution in Layer 2. DCT^{2}Gen works at the flow level and 5 different parameters are required to be passed in order to generate a tm. Each *tm* is generated for a fixed duration and the no. of racks, no. of servers in each rack, no. of *tm* matrix required (according to used simulation duration), traffic scale and a .csv filepath for storing the flow distributions needs to be fed to the generator as input.

5.3.2.2 Output of DCT^{2}Gen

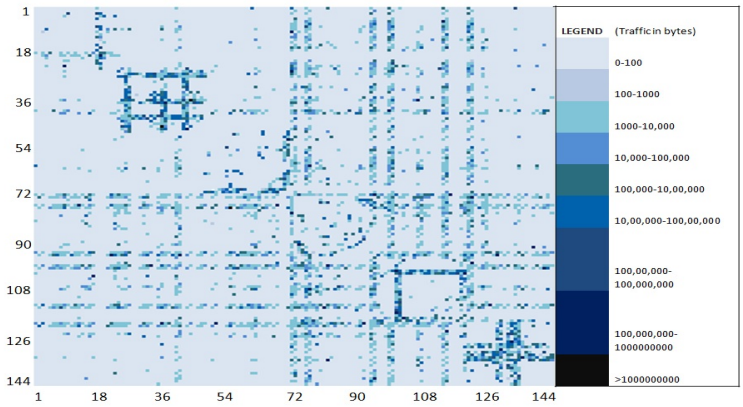

Figure 5.5: Sample Traffic Matrix Heatmap generated by DCT^{2}Gen

After processing the traffic distribution files using R, a $tm.csv$ file is generated denoting the total amount of traffic to be transferred between a server pair within the given duration. Additionally, a traffic schedule is generated ($flows.csv$) with 4 output parameters; Source, Destination, Start time and Flow length which contains flows of varying sizes between different server pairs that can be either intra/inter rack. A sample traffic heatmap (with 144 servers) generated from the *flows.csv* file, given in Figure 5.5 shows that the generated pattern matches closely with the DCN tendency of forming traffic hotspots where certain servers that may be distributed across multiple racks or within the same rack, exchange large volume of traffic amongst each other. This *flow.csv* file is passed on to the custom TCP traffic generator module in the designed network, for generating the TCP traffic flows.

5.3.3 DCTCP Model

DCTCP implementation was done by adding a new TCP class with modifications for packet marking in IP/TCP headers to reflect AQM and window control characteristics. The DCTCP variant is inherited from the TCP Reno variant within the L4 INET TCP module and additional methods are added

to replicate the DCTCP behaviour. For connection setup and Congestion window control, new state variables are included into the TCP interface module and the DCTCP algorithm is extended into all TCP classes. Details about the algorithm have been provided in the next section.

5.3.3.1 DCTCP Algorithm Implementation

DCTCP reacts to congestion in proportion to it's the extent and uses a marking scheme to make an estimation about the network congestion as soon as the buffer occupancy exceeds a fixed threshold, and accordingly scales down the congestion window. Congestion control processes used in standard TCP variants, such as Slow Start, congestion Avoidance, Fast retransmit and Recovery remain the same in DCTCP and the additional stages used in DCTCP implementation are described as follows:

1. **Marking Congestion at the Switch:**

   All switches in the network path are configured with a Congestion Threshold (K) and when the switch's instantaneous queue length is greater than K during a packet arrival then the switch sets the *CE* codepoint in the packet IP header to indicate congestion. An *ECN* flag is also introduced in the IP header to indicate to the receiving server that the packet was marked.

2. **Echoing Congestion Information:**

   An *ECN_ECHO* flag is set in the TCP header of the ACK by the receiver to indicate to the sender (when it receives the ACK from the receiver) that the packet being acknowledged had *CE* code point marking. Based on the principle of Delayed ACK, a boolean state variable, (*DCTCP.CE*) is initialized to false. When sending an *ACK*, the *ECN_ECHO* flag is set only if *DCTCP.CE* is true and while receiving packets, the *CE* codepoint is processed as follows:

   (a) If *DCTCP.CE* is false, but the *CE* codepoint is set, an *ACK* is sent for any previously unacknowledged packets and DCTCP.CE is set to true.

   (b) If *DCTCP.CE* is true, but the *CE* codepoint is not set, an *ACK* is sent for any previously unacknowledged packets and *DCTCP.CE* is set to false.

   (c) Otherwise, the *CE* codepoint is ignored and *DCTCP.CE* remains unchanged.

3. **Processing Congestion Indications:**

   Sending nodes check the fraction of sent bytes encountering congestion and stores the estimate in a state variable; *DCTCP.Alpha*, which is initialized to 1 and updated as given in eq. 5.1.

$$DCTCP.Alpha = DCTCP.Alpha * (1 - g) + g * M \tag{5.1}$$

where g is the estimated gain varying between 0 and 1. M denotes the fraction of sent bytes encountering congestion during the previous observation window. Whenever congestion estimate is updated, the sender updates the TCP Congestion Window (C_{wnd}) as given in eq. 5.2.

$$Cwnd = Cwnd * (1 - DCTCP.Alpha/2) \tag{5.2}$$

Parameters K and g are reconfigurable and based on information provided in [149], have been set to 40 and $1/16$ respectively and $DCTCP.Alpha$ and M are calculated within the code.

4. **DCTCP . ALPHA Calculation:**

For updating $DCTCP.Alpha$, three TCP state variables are used, which are explained next.

- ***DCTCP.WindowEnd*** – It denotes the TCP sequence number threshold for beginning a new observation window and is initialized to *SND.UNA* which represents the sequence number of first outstanding byte of data that has been sent but not yet acknowledged.
- ***DCTCP.BytesSent*** – It denotes the number of bytes sent during the current window and is initialized to zero.
- ***DCTCP.BytesMarked*** – It denotes the number of bytes sent during the current window that encountered congestion and is initialized to zero.

The congestion estimation process on the sender side processes the received *ACK* as follows:

(a) **Computation of acknowledged bytes:**

$$BytesAcked = SEG.ACK - SND.UNA \tag{5.3}$$

where $SEG.ACK$ is the ACK from the receiver indicating the next sequence number.

(b) **Updating the number of bytes sent:**

$$DCTCP.BytesSent+ = BytesAcked \tag{5.4}$$

(c) **Updating the number of marked bytes:**

If the *ECN_ECHO* flag is set, then the counter for marked bytes is updated as follows.

$$DCTCP.BytesMarked = BytesAcked + 1 \tag{5.5}$$

(d) If sequence number $\leq$ *DCTCP.WindowEnd*, then processing is stopped, else it indicates the observation window end and the congestion estimate is updated as follows.

- **Computation of current window congestion:**

$$M = DCTCP.BytesMarked/DCTCP.BytesSent \tag{5.6}$$

- **Updating the congestion estimate:**

$$DCTCP.Alpha = DCTCP.Alpha * (1 - g) + g * M \tag{5.7}$$

- **Setting the size of the new window:**

$$DCTCP.WindowEnd = SND.NXT \tag{5.8}$$

where $SND.NXT$ indicates the sequence no. of the next byte of data to be sent.

- **Reseting the byte counters:**

$$DCTCP.BytesSent = DCTCP.BytesMarked = 0 \tag{5.9}$$

5.4 TCP Performance Analysis

The Layer 4 switching performance of the Leaf-Spine network was adjudged with two distinct congestion control protocols; TCP and DCTCP to understand the suitability of both in current DCN environments. Traffic flow schedules were generated externally using a realistic traffic generator (DCT^{2}Gen) which were then passed on to the network to initiate the flows.

5.4.1 Performance across a single link

A single bottleneck link is tested with up to 10 concurrent flows at a time with varying congestion threshold (*K*) values. Each of the source TCP applications generated packets based on a Poisson arrival process with exponential inter-arrival and service time distributions with a link rate of 10 Gbps to emulate a backlogged sender. Though the application generates this amount of data, the amount of data that is forwarded is limited by TCP's data sending rate. The ports can have a maximum buffer capacity of 1.5 MB in this scenario. Figure 5.6 shows the average delay figures and the delay provided by TCP Reno is observed to be higher compared to DCTCP. This is due to the larger congestion window size used by TCP Reno which leads to greater queue length. So, even though it can have higher throughput values than DCTCP, the increased delay is a significant disadvantage. Delay provided by

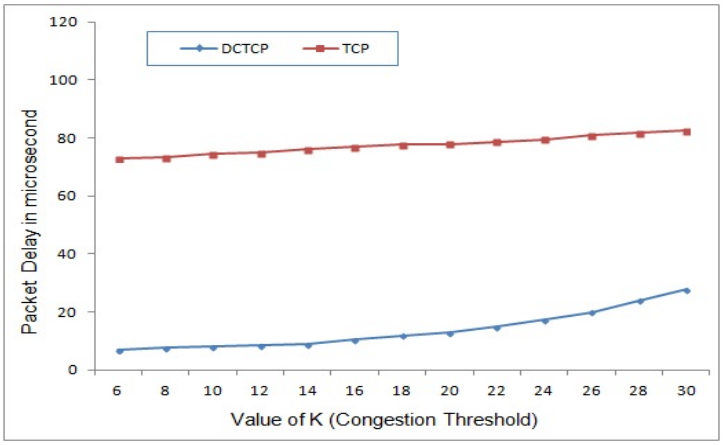

Figure 5.6: Average Delay for single bottleneck link

DCTCP is lower than standard TCP because of the lower queue length and the smaller congestion window size in the router. The packet marking scheme alerts the sender about impending congestion and accordingly the contention window size decreases early on to lower system latency.

5.4.2 Incast Bottleneck Scenario

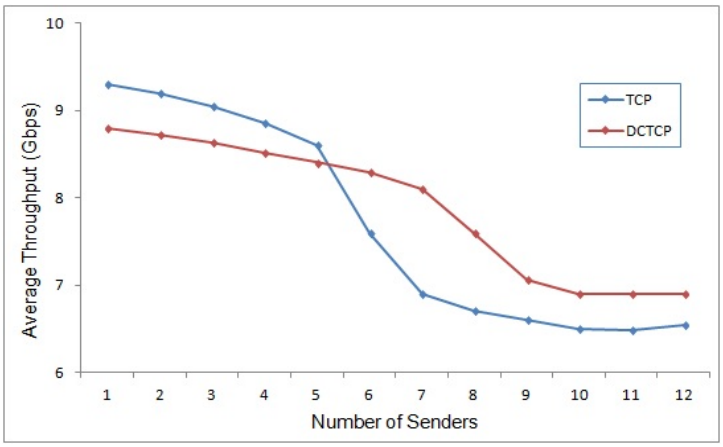

Figure 5.7: Incast Congestion Scenario

Incast congestion is the primary challenge for TCP implementation in DCNs which occurs in traffic with many to one communication pattern. In this case, Incast congestion is created by simulating many to one traffic flows from multiple host nodes scattered across a small network to a single destination and so, the link between the Leaf switch and the receiving node forms the main network bottleneck. As observed from Figure 5.7, Incast congestion brings down the network throughput to a lower value than the maximum link capacity in both cases. However, the Incast scenario sets in ear-

lier in TCP and throughput degradation is more compared to DCTCP. This performance degradation is mainly caused by timeouts experienced by several servers due to buffer overloading and loss of packets at the switching queues. So, even if the remaining servers complete their transfers they fail to receive the next request until a timeout occurs which keeps the bottleneck link idle and underutilised for an extended period of time resulting in throughput degradation. Results show that DCTCP has a better performance compared to TCP since it can adjust its window size early on, in accordance to network congestion, but it is not immune to Incast congestion specially when the number of flows in the network increases substantially dropping the throughput close to standard TCP values.

5.4.3 Average Network Delay

The parameters used for average packet round trip delay calculation are mentioned as follows.

1. Packet or TCP segment size - 1460 bytes, as per the standard TCP MSS value which results in 1500 bytes Ethernet frames.
2. Flow pattern of traffic - Realistic traffic flow files generated by DCT^{2}Gen.
3. Link Speed - 10 Gbps
4. Buffer size - 1.5 MB for 10 Gbps switches.

Figure 5.8 shows the average delay experienced by packets when TCP (Reno) is implemented in the network and it is observed that the delay increases almost linearly as the queue builds up. This is because a greater number of servers in the network leads to more concurrent flows to the routers. Since congestion control in TCP is only triggered after a packet is lost or retransmission timer expires so the router queues continue to grow to add up an extra queuing delay that increases the overall latency. Standard TCP congestion-control algorithms such as TCP Reno can perform well in large bandwidth-delay product (BDP) networks but they might not give the best delay performance in data centres since the RTT in DCN is much smaller than in traditional IP networks.

As observed in Figure 5.9, the average delay graph remains reasonably smooth when DCTCP is introduced in the network. By using ECN marking for congestion detection, DCTCP can significantly lower the latency for traffic flows and improve the network burst tolerance while using upto 90% less buffer space [153]. The improvement in performance is due to the fact that DCTCP can prejudge the level of congestion in the network and accordingly control its congestion window. As a result,

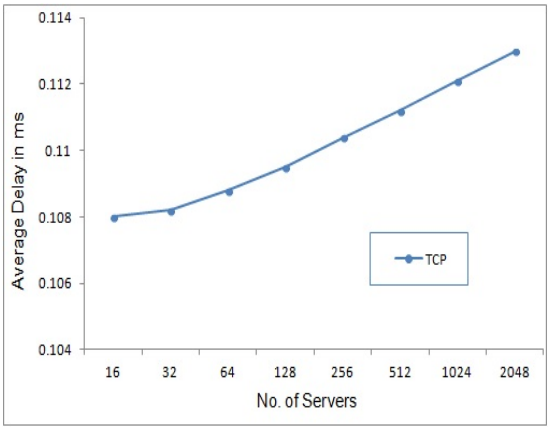

Figure 5.8: Average Delay with TCP Reno Implementation

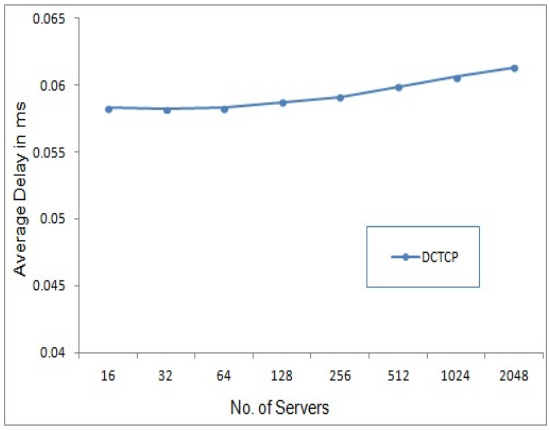

Figure 5.9: Average Delay with DCTCP Implementation

the system is not overwhelmed with too many concurrent flows at a time and hence the buffers are not overloaded leading to lower queuing time and lower overall latency. It can be observed that even with rise in network size, DCTCP can outperform TCP with an average delay of approx. 45% less as compared to TCP due to the efficient use of congestion window which leads to its suitability in networks with lower bandwidth delay product.

5.4.4 Average Network Throughput

Using the same parameters mentioned in the previous subsection, the network was analysed to judge the effect of TCP (Reno) on the network throughput as shown in Figure 5.10. It can be observed that the average TCP throughput decreases gradually with an increase in the number of servers which increases the traffic flows generated in the network leading to queuing delay in switching nodes. TCP is not ideally optimised to control congestion from an early stage based on the extent of congestion in DCNs and C_{wnd} resizing is only triggered with a packet loss when a considerable amount of congestion has already built up in the network. This contention in the switching queues lead to packet drop and blocking which affects reception of ACKs at the sender side causing subsequent timeout and throughput degradation. However having a larger C_{wnd} compared to DCTCP helps to have better throughput performance than DCTCP in small networks where the network congestion may not be too heavy to overload the buffers and cause packet drops. The observed TCP throughput figures in the network are relatively better than expected figures in multi-tiered networks, because of the presence of enough redundant network paths in Leaf-Spine topology.

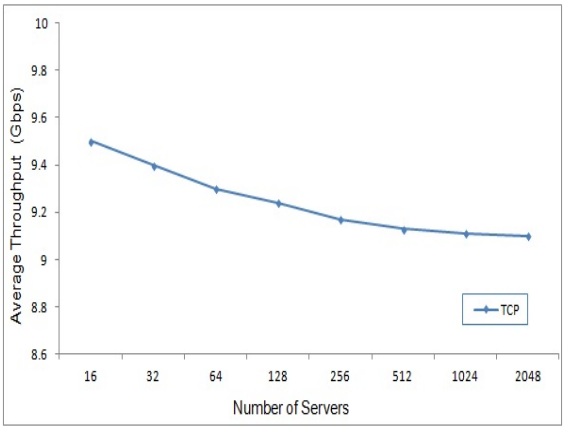

Figure 5.10: Average Throughput with TCP Reno Implementation

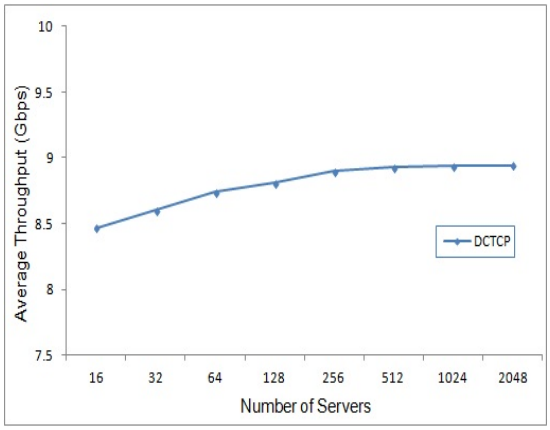

Figure 5.11: Average Throughput with DCTCP Implementation

Figure 5.11 shows that when DCTCP is used for congestion control, the throughput figures do not show signs of gradual degradation with an increase in network size as was observed in the case of TCP implementation. This is due to the inherent nature of DCTCP which has a high burst tolerance and subsequently frees up buffer space, leading to late queue build up. Since DCTCP can detect congestion from an early stage through ECN marking, the C_{wnd} resizing starts fairly soon which prevents buffer overloading in the network even in presence of a considerable number of flows. As a result, no unnecessary queueing delay and packet loss takes place in the switching nodes and the latency figures remains relatively lower as shown in Figure 5.9 leading to better throughput performance. The richly connected topology of Leaf-Spine networks also helps to alleviate the congestion bottlenecks and the throughput starts to flatten gradually only when the network size grows up sufficiently. In order to have a stable network state, DCTCP throughput remains around 90% of the link bandwidth since the C_{wnd} reacts early on to prevent packet loss and overcome Incast bottlenecks.

5.5 Conclusion

In order to lower network latency by avoiding congestion buildup and utilising the available bandwidth effectively, DCN operators resort to using Layer 4 congestion control protocols. In such a scenario, TCP provides a mature technology for ordered and reliable bidirectional delivery of data and hence TCP-based traffic forms the major share of DCN traffic flows in recent times. However, TCP has a large bandwidth delay product and may not provide the optimum performance required in DCN environments where the traffic flows typically have shorter RTT and can result in increased latency

and throughput degradation. This Chapter analyses Layer 4 traffic performance with TCP implementation using INET framework in a Leaf-Spine network model. The bottlenecks of TCP in data centres were studied and one of the baseline DCN specific transport protocol (DCTCP) was also implemented for a comparative analysis with standard TCP. Provisions were made in the designed network to generate TCP traffic flows internally by customising Layer 4 INET traffic generators or by using external traffic flow schedules generated by a realistic DCN specific traffic generator (DCT^2Gen).

DCTCP is believed to lower queueing delay and provide high burst tolerance for the mice flows and high throughput for the elephant flows [152, 153] with lower buffer occupancy and so are suited for use in shallow buffered DCN switches. Based on the performance analysis, this chapter can be concluded with the following findings: (a) DCTCP gives better performance than TCP (where congestion notification comes only after packet loss and timeout) due to having an inbuilt ECN marking scheme and shorter congestion window which subsequently reduces sender transmission speed and alleviates queuing delay, overflows and TCP retransmissions as the network size widens. (b) In smaller networks without a large number of concurrent traffic flows overloading the buffers and causing packet loss, TCP can give a similiar performance to DCTCP. (c) To achieve a stable network state DCTCP tries to reduce delay considerably with a smaller congestion window but may result in frequent C_{wnd} resizing and the throughput remains lower than the full link bandwidth since the C_{wnd} reacts early on to prevent packet loss and overcomes Incast bottlenecks. The developed Layer 4 network model used in this chapter provides further scope for facilitating TCP traffic implementation in other network topologies by changing the layout of the electronic hosts and routers that interfaces with the INET custom nodes. By changing the Layer 4 TCP interfacing modules, the developed model can be further extended to include different data centre specific TCP variants other than DCTCP.

The future of DCN topologies lies in optical switching and hence the last part of the thesis proposes an optical interconnect suited for the emerging DCN environment and gives an outline of a network where the proposed interconnect can be implemented. Some recommendations for future research in the DCN switching paradigm are then provided in the concluding chapter.

Part III

Looking into the Future

Chapter 6

Optical Interconnects of the Future

6.1 Introduction

A rapid surge in bandwidth hungry cloud-based applications, virtualisation and data intensive computations is continuously pushing the data centre capacity and storage demands leading to a steady annual rise in DCN traffic by almost 25% [8]. To cope with the growing traffic, the data rate is gradually increasing from 10 Gbps to 40 Gbps and possibly to 100 Gbps or higher in the near future. Additionally, power consumption efficiency is a major concern for sustainable development in very large scale DCN infrastructures [116] since DCN network equipments tend to consume approximately 20% of the total IT energy and this value is expected to rise higher [82]. As a whole, the electronic interconnects currently in use in DCNs are proving to be energy-inefficient, hindering link capacity upgrading and are not massively scalable. Thus, innovation in switching is important in modern data centres where optical switching technologies can provide high data rates for switching high aggregation bandwidths while showing an immediate improvement in energy efficiency compared to their electrical counterparts since it eliminates the O-E-O conversions at the switches and can be future proof due to the data rate transparency of the system.

Depending on the varying user applications in DCNs the heavy intra-rack traffic flows can result in very high bandwidth demands. However, for increasing the bandwidth capacity while simultaneously reducing energy consumption, most of the DCN architectures, concentrate on developing switching architectures for the upper network tiers (core/aggregation layer) which mainly deal with switching between server racks. Conventional commodity ToR switches are still used at the edge tiers in such

switching schemes and it should be noted that the electronic edge tier switching which is responsible for interconnecting the servers within a rack can consume up to 90% of the total power consumed by DCN switching equipments due to the presence of a huge number of ToR switches [3]. Hence, reducing power consumption at the edge tier is highly essential for improving the overall DCN power efficiency and optical ToR switches can be a viable solution in that direction.

This chapter introduces an Optical Top of Rack (OToR) switch that can minimise energy consumption in the power hungry edge tier. Utilising the fast optical switching property of AWGRs, the OToR can effectively handle the intra-rack traffic flows with contention resolution provided by FDLs and it is observed that they offer better delay performance (average latency in *ns* range) than similiar sized commodity switches (with latency in μs range) while causing far less power dissipation (approx. 33% to 55% lower). An all optical architecture is thereafter proposed where the OToR can be utilised for efficient DCN communication and can be interconnected through a mega optical space switch (OSS), based on a non-blocking Clos network, which can satisfy future scalability demands. This network architecture can connect a fairly large number of servers while providing full bisection bandwidth through a combination of optical cross-connects and WDM rings.

6.2 Optical Top of Rack Switch

OToRs are a new concept in data centre switching solutions which can provide enhanced intra-rack bandwidth availability, and by mostly using passive components they can minimise energy consumption in the edge tier as compared to the commodity switches that are being used at present. OToRs can also eliminate the need for extra transceivers required in O-E-O conversions while enabling easy upgrading of network links. AWGR based optical switches have been used for handling core layer switching in previously proposed architectures such as in [104–106] since fast optical switching of the order of nanoseconds is possible using AWGRs coupled with tunable lasers at the input ports. In order to utilise the advantage of AWGRs for switching in edge layer, they have been used as the main switching element in the OToRs proposed in the next subsection because of its high bandwidth switching capability, architectural simplicity, fairly high port count, low cost and high power efficiency due to being a passive optical device. Additionally, AWGRs have the benefit of fixed cyclic wavelength routing property which allows the selection of output port based on the wavelength of the transmitted signal. Hence, AWGR based switches in the edge layer can form a reconfigurable wavelength-routed optical cross-connect for routing any signal to any output port as required and effectively handle the

intra-rack traffic that is generally bursty by nature.

6.2.1 Switch Architecture

Figure 6.1 shows the proposed OToR architecture that utilises the cyclic routing property of the AWGR to establish a non-overlapping one-to-one connection between any two servers inside a rack. Each server is equipped with an optical interface (OI) which is connected to the AWGR switching fabric. The OI consists of a wavelength tunable transmitter (WTT) and receiver and can send/receive optical signals to/from other servers by tuning the optical carrier to the desired wavelength. N wavelengths are required for connecting N servers across each port of the AWGR and by appropriately tuning the lasers, incoming optical packets on any single wavelength channel with 10Gbps link rate can be routed to any one of the N output ports.

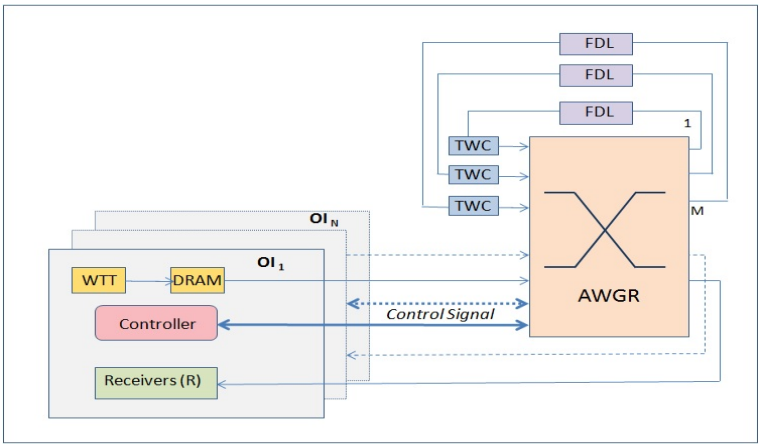

Figure 6.1: Optical Top of Rack Switch Architecture

OToRs have been recently proposed in a few other DCN architectures, such as in [116, 160] which are also energy efficient but do not include physical provisions for resolving traffic contention within the rack and instead resort to complex bandwidth allocation strategies to avoid simultaneous traffic conflict at a particular destination. Moreover they use broadcast-and-select architectures with optical combiners (such as splitters/couplers) to aggregate signals and provide multipoint connections. Broadcasting traffic to all available servers can lead to substantial optical insertion loss if the number of servers in a rack is high and hence limits further scalability. Since each OI has a limited number of receivers (R), traffic contention might arise if several flows within a rack are concurrently destined towards the same server across different wavelengths. Hence buffering needs to be provisioned within the switching fabric, but instead of including costly buffers within the ToR, cheap DRAMs

are included in the OI in the proposed architecture to buffer the traffic until the desired connection path is available. In order to fully eliminate contention in the traffic flows within the OToR, taking cue from [106], a number of FDLs are used in the switch that can hold the contended traffic for a fixed number of intervals until the desired connection becomes free. As shown in Figure 6.1, each OI transmitter has a simple FIFO queue buffer which can hold the transmitted packets and if the channel connecting the server and the OToR is free, then the controller sends the first packet in the queue to the desired output port unless the switch control signal indicates that all the FDLs are currently busy which means that there is a potential chance of the packet getting dropped before reaching the destination. This single bit signal sent from the OToR after the controller detects the packet destination port is the initial feedback required between the OI and the switch and then the contention resolution is entirely handled by the FDLs present in the switch.

If the destination port is free then the packets are sent out on dedicated wavelengths but in case the destination output port is busy, the incoming packets are directed to a wavelength channel in any of the *M* FDLs, and re-circulated a fixed number of times until an output port channel becomes free. Every FDL is fitted with a small number of TWCs *(W)* so that it can carry up to *W* wavelength channels at a time and the TWCs tune the wavelength as required for connecting to the destination server. FDL inputs are interleaved with the incoming fiber connections from the servers to the AWGR and the type of FDLs used are degenerate in nature where delay times increase uniformly as a multiple of the base delay. The number of FDLs (M) required in the OToR will depend on the traffic conditions and the number of incoming links coming to the OToR when placed within a DCN architecture.

6.2.2 Performance Analysis

The proposed OToR switch was modelled using OMNeT++ and subsequent simulations were performed to judge the packet loss performance and delay in comparison to standard commodity ToR switches. Servers generated packets with uniform Poisson traffic distribution whereby packet inter arrival times and service times are exponentially distributed and the input traffic is evenly sent across all destination ports or any server can send packets to any of the other *N* servers with equal probability. The simulated switch is considered to be of size 64 ports in accordance with the 48 server racks and 16 port connectivity to the Tier 2/Spine layer used in commercial DCNs deployed with folded Clos topology. Mean data frame size is set to be 625 bytes which give the service rate (μ) as 10^{10}/5000 packets per second or the mean packet duration (T) to be 500 ns. The FDL base delay D is set to be

0.1T, which is chosen in order to balance the trade-off between latency and packet loss.

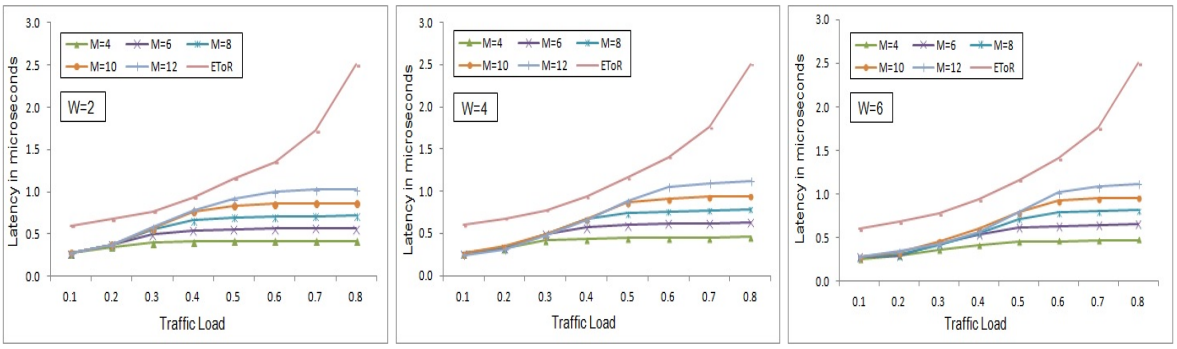

Figure 6.2: Latency comparison with single receiver (*R=1*)

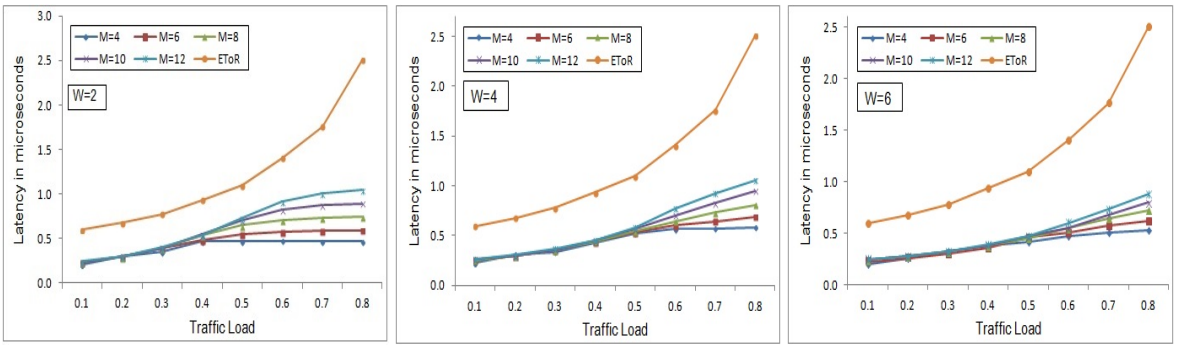

Figure 6.3: Latency comparison with two receivers (*R=2*)

Figure 6.2 and 6.3 shows the switching performance with respect to average end to end latency for packets transferred through the proposed OToR with possible server and FDL buffering. The delay values are plotted for varying traffic load with variations in the number of FDLs (M), number of TWCs (W) used per FDL and the number of receivers (R) used in each server OI. As expected the latency increases gradually with increase in load but it also rises with the increase in the number of FDLs for specific combinations of W and R, e.g. at 50% load the delay increases from 0.45 μs (M=*4*) to 0.85 μs (M=*12*) when using single receiver (R) with W=*4*. This is because the mean FDL delay grows with the rise in the FDL count in the different scenarios. Since the FDLs are degenerate in nature, so they have a fixed base delay which is used for the shortest FDL and then gradually the maximum FDL delay rises with an increase in M. It is also observed that instead of using only a single TWC per FDL, if the value of W is increased the delay can be lowered due to using multiple wavelength channels in each FDL. Generally every server OI may have a single receiver to reduce cost but if multiple receivers are introduced in every OI then the latency can be further reduced since it will allow reception of multiple packets at once at any server. Even though recirculating packets through the FDL might

incur some additional delay to the otherwise fast AWGR switching, the simulation results shows that the overall latency (within *ns* range) still remains lower compared to a similiar sized electronic ToR (EToR) switch (similiar to the one as used in chapter 4) with latency in μs range.

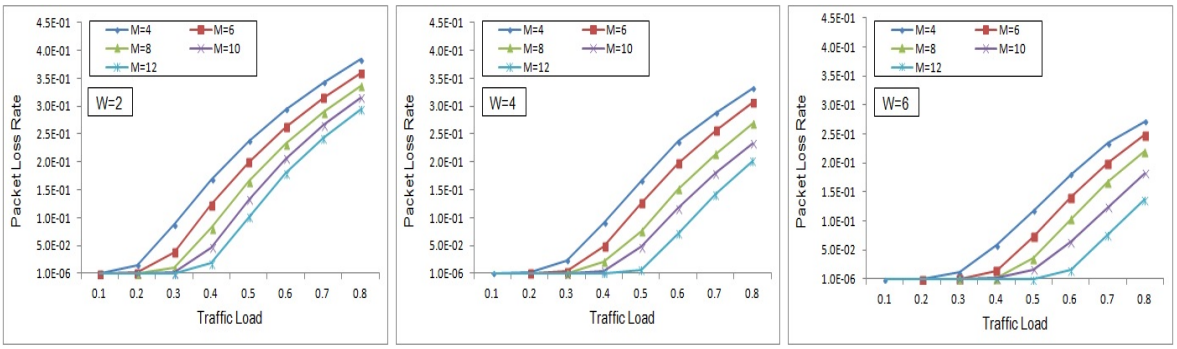

Figure 6.4: Variations in Packet Loss with single receiver (*R=1*)

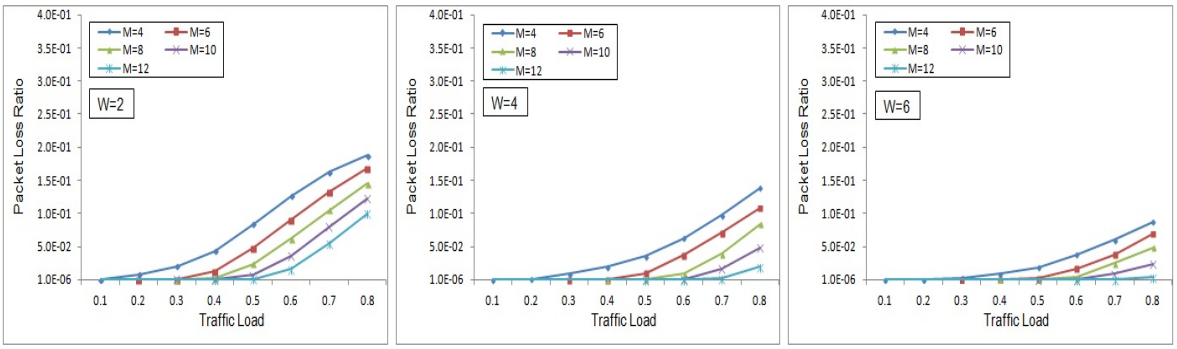

Figure 6.5: Variations in Packet Loss with two receivers (*R=2*)

Figure 6.4 and 6.5 plot the packet loss ratio as a function of the traffic load with varying switch configurations. With increasing traffic load, the switch tends to be overloaded with a large number of input traffic streams which try to get transmitted through a small number of output transmission channels and as a result the packet loss increases gradually as expected. However with the increase in the number of available FDLs in the switch, the availability of buffering option also increases and as a result the packet loss tends to decrease considerably, e.g. at 50% load the packet loss rate decreases from 0.15 (*M=4*) to 0.007 (*M=12*) when using single receiver (*R*) with *W=4*. Similar results are observed with the increase in the number of available TWC which introduces the availability of multi-wavelength channels in each FDL and boosts buffering possibility. Additionally the introduction of multiple receivers in the server OI can further reduce packet loss. Packet loss in a network is related to the level of desired QoS (Quality of Service) and is dependent on the underlying applications responsible for driving the switching traffic. As per [161, 162], an approx. 1% -2.5% packet loss ratio

is considered acceptable for bursty traffic such as audio/video streaming and for majority of the time, the traffic load (ρ) is ≤ 0.3 [147, 162]. Simulation results show that the switch gives such acceptable performance under standard network conditions (specially with increase in number of FDLs) and this performance can always be significantly improved by using larger buffers and multiple receivers in the server OI.

6.2.3 Power Consumption Analysis

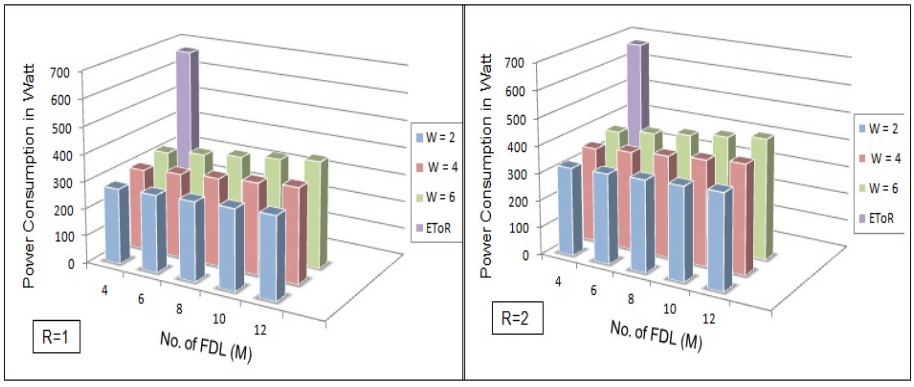

Figure 6.6: Power Consumption Figures for Proposed OToR Switch

Optical switching in the edge tier should ideally lead to substantial improvement in terms of energy efficiency and for computing the gross power consumption figures of the proposed OToR, the major active switching components are considered and a graphical analysis compared to electronic ToR (EToR) switches is presented in Figure 6.6. However, it is to be noted that the power consumption estimations do not consider the power dissipation due to the controller action in the server optical interface and only include the major energy consumers like TWCs, wavelength tuneable transmitters, tuneable receivers and electronic buffer used in the OI. As per figures obtained from [163] every individual TWC consumes 1.5W, each tuneable transmitter uses approx. 1.5W [116] and each receiver uses 0.8W [164] of power. According to the data provided in [122, 165], electronic buffers can be considered to use approx. 2.5W of power for every port. So, by calculating the required number of active components of each type, the energy figures of the proposed architecture is plotted as a function of the number of FDLs (M) for different combinations of TWCs (W) and receivers (R) considering a standard 64 port ToR switch. For comparison with a similiar size commodity ToR switch, the per port power consumption is assumed to be 9.4W [116] and the power required by each network interface card (NIC) is considered as 1W [116].

As observed from the plotted graphs in Figure 6.6, the proposed switch model proves to be more energy efficient than the standard commodity ToR switches primarily due to using passive components as the main switching fabric and buffering substitute. It can be seen that the power consumption per port increases linearly with the increase in the number of FDLs for all scenarios. The results suggest that the proposed AWGR based architecture can be highly scalable with regard to power consumption and can show an improvement of approx. 33% to 55% (under the two given configurations) over the net power dissipation in commodity switches. Additionally, it can also be inferred that with an increase in switch size the energy efficiency will be further improved.

6.3 Proposed All Optical Data Centre Architecture

Using the optical ToR switch presented in Section 6.2, the outline for an all optical combined Intra and Inter-rack interconnection network [168] is proposed in this section which can be highly scalable and power efficient while providing sufficient bandwidth capacity. A mega space switch that can enable connectivity among the different server racks by using a combination of fast OCS and optical rings providing dedicated non-blocking circuits form the core switching layer of the proposed DCN architecture. The OCS switching architecture is inspired by the Clos [50] network which provides connectivity through smaller switching blocks. Previous studies confirmed that in cloud DCNs, around 80% of the generated traffic stay within the same rack [15] and so high speed AWGR based optical ToRs can handle that traffic requiring fine switching granularity while the OCS can transmit the bulky DCN hotspot traffic that is usually delay-insensitive and slowly varying by nature.

6.3.1 Outline of Network Architecture

Recent DCNs are mostly deployed using the Leaf-Spine (see Figure 6.7) topology which connects servers with Core layer (Spine) through ToRs (Leaf) and can provide good path redundancy for interconnecting servers at the full bisection bandwidth. However, this results in a poor speed/cost ratio where a large number of switches and long cables are required and with an increase in DCN size, the size of the Spine also widens and puts a limit on the network scalability. On top of that, the growing power consumption figures is also a critical concern. So, to overcome the electronic switching bottlenecks, the proposed network shown in Figure 6.8 introduces an all optical architecture to minimise power dissipation while providing improved network performance. Thus, OToRs have been used in

place of the Leaf switches and OCS based switching replaces the Spine layer. Instead of multiple very large Spine switches, a single unified module composed of optical cross-connects (OXCs) and WDM rings can act as the inter-rack switching element which is logically equivalent to a very high radix optical space switch (OSS). The intra DCN OSS can be considered to be a fully operational point-to-point circuit switch with a reconfigurable connectivity substrate which can provide an agile network-wide connectivity across the server racks. In order to emulate the multipath nature of the Leaf-Spine network, each OToR can connect to the mega OCS switching node through a number of optical links, depending on the network size and traffic conditions.

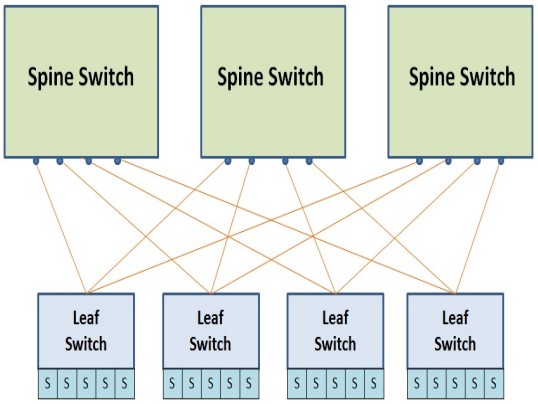

Figure 6.7: Current Data Centre Architecture

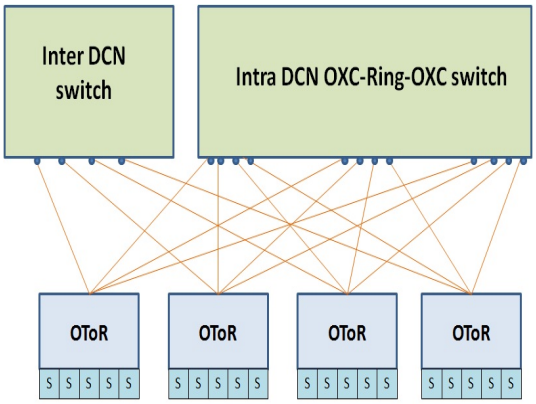

Figure 6.8: Proposed Data Centre Architecture

OCS switching is a relatively inexpensive and mature technology that can be easily deployed, though it can incur switching delay due to high reconfiguration times. However, recent works have demonstrated faster circuit switches which can reconfigure within a few μs [96, 166]. Around 95% of the DCN traffic tends to be transferred by the top 10% of the elephant flows [15]. Under such conditions OCS can provide a stable pathway for off-loading the bulky flows in complete isolation from other traffic; e.g. in network virtualisation, virtual machine migrations. The OCS based OXC-Ring-OXC switch handles the inter rack traffic and once the circuits are setup it will not result in any packet loss or incur additional delay due to buffering like packet switches and flows can be transferred in Tbps speed in each fiber with WDM routing [95]. To improve fault tolerance, the network is not based on a single centralised switch and the provisions for multiple paths between the edge tier and OCS based core switching tier ensures path redundancy. The flattened two layer nature of the network does not compromise on scalability (as in electronic DCNs) as there is no dependence on very high-radix switches and instead the switching function is distributed across a number of switching nodes.

6.3.2 Intra Rack Connectivity

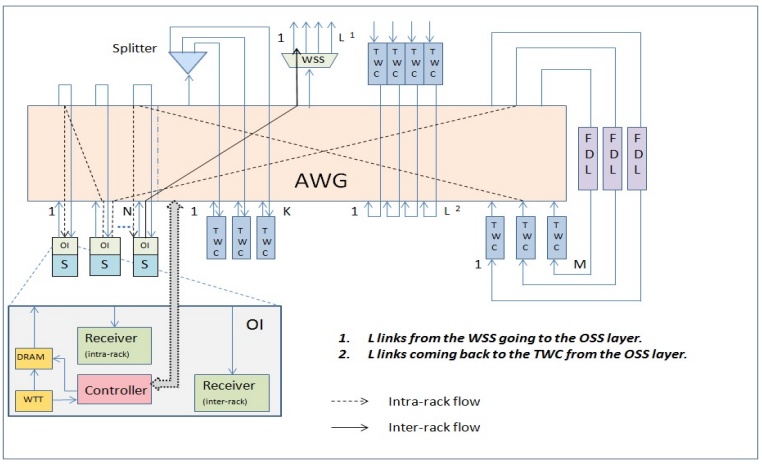

Figure 6.9: Proposed Intra Rack Switch

The OToR switch described in Subsection 6.2.1 can be further modified to include additional functionalities to fit in with the proposed all optical DCN architecture and the upgraded switch architecture is shown in Figure 6.9. In order to facilitate the transfer of traffic to other server racks, each OToR needs to be fitted with a WSS that can aggregate and send the traffic across L links (maintaining the OToR's connection to the OSS), and the incoming traffic to the OToR in each of the L links will first need to pass through a TWC to eliminate any wavelength contention before being transferred to the input of the AWGR for routing to the destination ports. The servers should have dedicated receivers for Intra and Inter rack traffic to allow simultaneous reception of traffic coming from within/outside the rack. Multicast traffic is a frequent phenomenon in DCNs with big data and parallel computing applications where a server simultaneously tries to transmit to several other servers (around 10 [15]). In the proposed Intra rack switch, a dedicated optical splitter can be used for multicasting and such traffic can pass through the AWGR to the splitter. The splitter can then multicast the traffic to up to K servers through the AWGR by appropriately tuning the wavelength so that it is sent to the desired servers who can filter out the designated wavelength using a wavelength tuneable filter.

6.3.3 Inter rack Connectivity (Optical Space Switch)

In warehouse scale DCNs, inter-rack traffic consists of huge flows that are slowly varying by nature and hence require a massively scalable, stable, high capacity, non-blocking switching network for its transmission. Previous works have proposed OCS solutions that can provide very high bandwidth

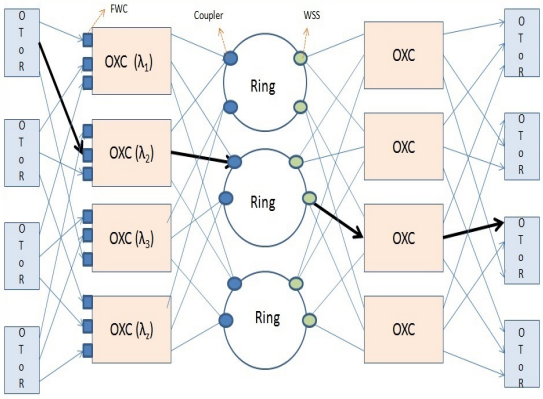

Figure 6.10: Inter-Rack Switching Network

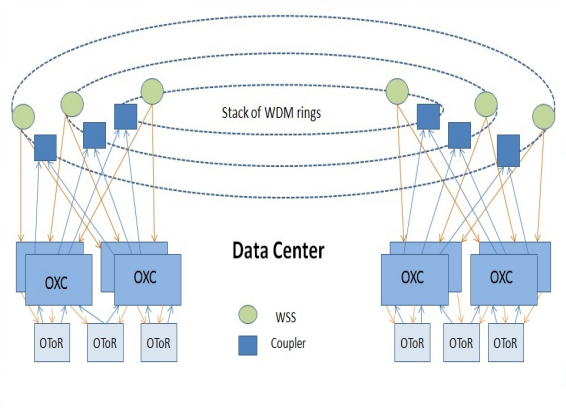

Figure 6.11: Implementation of OSS Network

paths between connecting nodes but are not highly scalable. MEMS used for inter-rack connectivity in [94, 95] are restricted by the unavailability of sufficient number of server ports since commercially available MEMS can have a maximum of a few hundred ports which cannot fully interconnect a warehouse scale DCN. Mordia [96] suggested using a ROADM ring to directly connect ToRs by adding/dropping wavelengths through couplers/WSSs. This approach also fails to scale up sufficiently due to the limit on the number of wavelengths that can be multiplexed within an optical ring and in the case of a link failure, connectivity of entire network may be compromised due to its dependence on the central ring. More recent OCS networks like Wavecube [101] can offer better scalability but cannot provide full bisection bandwidth and in the case of very bursty DCN traffic the network performance can be affected. In networks such as OSA [99] the multi-hop routing policy increases the latency and routing complexity and since the connections are non transparent, the O-E-O conversions required at every intermediate rack increases energy consumption and limits easy capacity upgrade.

This proposed core switching layer as shown in Figure 6.10 uses optical WDM rings along with a series of OXC switches at the input/output stages and is based on the concept of Clos networks utilising smaller switching units for connecting large number of server racks. Hierarchical switching pattern has been avoided to minimise cabling complexity and latency build-up. Instead, direct ToR to ToR non-blocking paths through the OSS layer are made feasible. The WSS present in each OToR can aggregate the inter-rack traffic across L links and direct each to individual input stages of the OSS layer in appropriate paths as per availability and traffic demands. The mega OSS is a three point structure with Z $(a \times b)$ and Z $(b \times a)$ MEMS/fast circuit switches at the entry and exit stages with a being the number of wavelengths that are multiplexed within the WDM rings and b being the number of WDM rings in the center that can multiplex a given set of wavelengths $(\lambda_1, \ldots \lambda_Z)$.

When incoming flows from the OToRs enter the OSS node they are converted to specific wavelengths assigned to each input node by a set of fixed wavelength converters (FWC). A $(1 \times n)$ coupler and $(1 \times n)$ WSS (value of n can be varied depending upon the number of required wavelengths) acts as the ROADM pair and adds/drops the wavelengths to/from the optical ring. By having a dedicated wavelength for every flow, the ring essentially acts as a point-to-point cross connect setting up a path between the source and destination server racks.

For minimising contention and ensuring one to one correspondence between the wavelengths, the non-blocking criteria of this Clos-inspired network [50] demands that b is $\geq a$ and no two connections coming from the same input module pass through the same central ring and similarly no two connections going to the same output module comes from the same central ring. Moreover, only a single wavelength can be added/dropped by each link of the coupler/WSS. This can avoid wavelength contention and additionally multiple links from/to each OToR take care of the situation where several flows are directed towards a single rack. So this proposed network can be considered as an improvement upon the network demonstrated in [167] which uses a single connection between the ToR switch and OSS switching layer and hence lacks the path redundancy required to handle the contention that might arise specifically in the case of an Incast scenario (a common occurrence in TCP traffic as illustrated in Chapter 5) where a large number of flows from within the network can be targeted towards a single rack. Additionally [167] also does not exemplify the ToR level connectivity of the network which has been well described in this case and this architecture also takes into consideration the use of optical ToR switches in contrast to the use of commodity switches at the edge tier.

As shown in Figure 6.11, the input and output OXC nodes can be collocated with the OToRs for practical implementation and the stack of WDM rings can be extended as required to cover areas where groups of server racks may be distributed locally. This will ease the cabling complexity by doing away with the need for very long cables compared to Clos networks and the server racks can be connected to the input and output stages of the Intra DCN OSS through shorter fibers. Traffic within data centres is usually selective by nature and so grouping the frequently communicating servers within the same rack can further ease the inter-rack traffic load by limiting the OCS reconfigurations.

6.4 Scalability and Power Consumption Efficiency

The proposed network can scale up linearly depending on the type of OXC that is being used and the wavelength multiplexing capability of the WDM ring. As of now an optical ring can carry upto 160

wavelengths [169] by utilising the C and L bands of the optical fiber and considering commercially available optical circuit switches such as Calient S320 with 320 ports [170] and Polatis 7000n with 384 ports [171] it is possible to connect to over 3K server racks within the proposed network. However, from the scalability figures given in Table 6.1, it can be observed that even by using the largest Spine switches available commercially such as Arista 7316 [122] and Juniper QFX10000 [123] a two-tiered Leaf-Spine network cannot even scale up to 2.5K server racks without requiring a third switching tier that incurs additional latency and cost. DCNs usually have 48 servers in each rack and in such a scenario the proposed network can accommodate over 150K servers with 16 links between the core and edge switching tiers in equivalence to the standard 16-way ECMP in electronic networks; while Leaf-Spine networks can cater to approx. 110K servers only at 10 Gbps link speed. Since OCS technology is fairly mature, a lot of advanced switch architectures are constantly being proposed with much larger port counts than currently available commodity switches which pave the way for even greater scalability in the proposed network. For example, considering recently proposed optical crossconnects in [172] with 1536 ports, it can be possible to connect at least 15K server racks.

Optical Switch Type	**No. of Racks in Proposed Network**	**Net Server Count**	**Electronic Switch Type**	**No. of Racks in Leaf-Spine Network**	**Net Server Count**
Calient S320 (320 ports)	51200 (L=1)	2457600	Arista 7316	2048	98304
	6400 (L=8)	307200			
	3200 (L=16)	153600			
Polatis 7000n (384 ports)	61440 (L=1)	2949120	Juniper QFX10000	2304	110592
	7680 (L=8)	368640			
	3840 (L=16)	184320			

Table 6.1: Scalability Comparison (where L = no. of interconnections between OToR and OSS layer)

Electronic switches can be highly power consuming with increased port count and are not widely available with very large link rates. However, OCS switching is not data rate dependent and since the proposed architecture mostly uses passive switching components like AWGRs, splitters, couplers, optical rings, FDLs etc, the network can be easily upgraded to suit higher data rates without any additional expenditure. Use of optical switching in all network tiers does away with the requirement for additional transceivers to connect the edge tier to the OSS layer as opposed to 2-4 transceivers required in the electronic tree-based networks for O-E-O conversion and this can lead to considerable energy savings. The only active components used in the OToR switch are the set of TWCs and a WSS and their share of power consumption can be minimised by lowering the number of TWC required for

multicasting/FDL-buffering depending on the traffic intensity. Due to the dynamic nature of the intra-rack traffic, the OI components can be kept in low power mode to further enhance power efficiency. Optical cross connects used in the OSS core switching layer can consume a minimal amount of power if using OCS variants such as MEMS and the other required components will be the set of FWCs at the input of the OSS and a WSS used in each optical ring. Considering a 320 port MEMS as the OXC component with each ring multiplexing about 160 distinct wavelengths as already discussed; the net power consumed in the OSS layer can be represented by eq. 6.1.

$$P_{OSS} = 160 * 320 * P_{FWC} + 160 * P_{MEMS} + 320 * P_{WSS} \sim 58.6KW \tag{6.1}$$

where each FWC consumes 0.8W [164], each MEMS consumes 0.25W per port [109] and each WSS consumes 15W [116] of power. So, the net optical core layer (OSS layer) power consumption (while supporting at least 1.5K servers) proves to be much lower than the Spine layer power consumption for a similiar sized Leaf-Spine network, as already deduced in Subsection 3.4.2.

6.5 Conclusion

Optically switched networks have the potential for high scalability, low power consumption and increased bandwidth capacity and are considered as a promising solution for overcoming the bottlenecks present in traditional DCNs powered by electronic switches. However, no fully-fledged optical DCNs are in operation yet and designing an optical network that can interconnect a warehouse scale data centre is still not fully feasible. Even though the major share of power consumption occurs in the edge tier, most of the proposals for future DCNs have been concentrated on optical switching in the core layer. In this chapter, an optical Top of Rack switch was proposed that utilises the cyclic routing capability of AWGRs to enable fast optical switching at the edge tier and substantially minimise energy consumption by using passive components (like AWGR). Effective handling of the intra-rack traffic is made possible with contention resolution through FDLs and inexpensive DRAMS in the server OI. It is observed that the proposed OToR causes far less power dissipation (up to 55%) than similiar sized commodity ToR switches while offering better delay performance in *ns* range than the commodity switches which operates in μs range.

A network has been further outlined where the OToR can be interconnected within the DCN through a mega optical space switch (OSS), based on a non-blocking Clos network. This architecture can

connect servers through a combination of optical cross-connects and WDM rings and establish optical connections between servers in the same or different racks. Since the network is based on OCS, it is comparatively inexpensive, easily deployable and can deliver very high capacity links for handling the bulky inter-rack traffic. Optical rings which form the core of the OSS layer are easily extendable, resulting in a reduction of overall cabling complexity. As per recent availability of optical circuit switches, the proposed network can satisfy the scalability demands of future DCNs and accommodate far more servers than the currently operational two-tiered Leaf-Spine data centres. Moreover the OSS switching layer can prove to be way more energy efficient than the Spine layer for an equivalent sized network.

However, it needs to be mentioned that although optical interconnects can be highly power efficient, provide greater bandwidth capacity and have better network performance, all optical components are not easily available (due to lack of mass production) as compared to commodity switches, which is why optically switched DCN networks might not prove to be highly cost efficient to start with, but with the advancement in optical integration technology and greater commercial availability the CAPEX figures are expected to come down substantially in the near future.

The next chapter concludes the findings and observations gathered over the course of the entire thesis and sheds some light on viable future research directions in the field of DCN architecture design.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

The trends and standards of current DCNs have been analysed in this thesis for gaining a better understanding of the prevalent infrastructure. Further assessment of the performance parameters (for both Layer2 and TCP traffic) of a current topology was done to judge the future directions in this field. Studying the needs of modern data centres, particularly in relation to energy efficiency, an optical switching solution is then proposed for the edge layer to address the demands for optimal performance and low power consumption. Thereafter a topology has been proposed whereby such an optical solution can be implemented in an all optical architecture for effective Intra DCN communication.

Evolution of DCN Infrastructure

In Chapter 2, the dominant trends of recent data centres and how their infrastructures have evolved with time were identified. It is observed that commercial DCN designs based on multi-tiered tree networks using commodity switches suffer from the bandwidth limitations introduced by oversubscription, inefficient network scalability, low network resource utilisation, complex cabling and high CAPEX and OPEX budgets. By studying the various DCN architecture proposals, the design challenges that can give rise to a series of performance-based dilemmas (as described below), impacting the underlying applications and services can be identified.

Network Performance or Cost Efficiency – Path redundancy introduced in a network can ensure minimal latency and maximum bisection bandwidth. Availability of a large number of network devices in hierarchical DCNs like Fat-Tree and Leaf-Spine networks improves the throughput and reliability but adversely affects the cost and energy expenditure. Recursive architectures like DCell and BCube require less power-hungry networking elements since they can use servers for packet forwarding, but they cannot provide full bisection bandwidth. Moreover, expanding the recursive network across multiple layers lengthens the network diameter and affects the system latency as well.

Switch or Server based architectures – While switches experience less processing delay than servers, the later has better programmability and enables a network to be built using smaller commodity switches. This can be economical in terms of cost (by using cheaper and less number of switches) and energy (does away with high-grade switches and associated cooling cost). Nonetheless, the servers are kept switched on for packet forwarding even when not undergoing any computing operation which proves to be expensive. Usually, server-based architectures can scale out efficiently as they are not restricted by switch size limitations, but installing multiple NICs per server cannot be a practical approach for large DCNs.

Scalability or Incremental expansion – Modern cloud DCNs need to expand their network without fully upgrading the existing architecture, in order to meet their storage demand. Even though homogeneous topologies can be scaled up to a fairly large size, they are not flexible enough to allow incremental expansion without replacing existing switches. Full bisection bandwidth is only possible with a specific number of servers, for a given switch radix and so introducing servers on demand can affect network performance. Asymmetric architectures can be arbitrarily expanded as required, but being based on random graphs they have high wiring complexity and require difficult path routing that can make scaling up very problematic.

Rise of Optics in Data Centres

An overview of optical switching paradigm and the rise of optical interconnects in data centres is presented in Chapter 3. The current optical DCN proposals are analysed based on their underlying switching technology to show how the architectures are evolving to tackle the diverse DCN traffic. It is observed that over the last few years, several research approaches have proved that optical interconnects can provide the desired performance requisites of high bandwidth, reduced latency and low power consumption but optical DCN architectures need to attain a very degree of flexibility and

reconfigurability while addressing the existing challenges of buffering capability, scalability and resilience.

Since, the setup/release of lightpaths and the switch reconfiguration timings in OCS is usually non-negligible, it results in higher latency and creates a possibility of limited network connectivity. However, it is observed that fast OCS solutions such as those proposed in [96, 166] can possibly reconfigure within a few microseconds and exhibit a greater resource utilisation. Large port count MEMS switches are now available and such switching solutions being energy efficient and data rate agnostic can give orders of magnitude lower per-bit energy consumption than commodity switches. Recent research in the field of integrated optics have demonstrated the possibility of compact and very high port-count AWGs and coupled with advanced TWCs, it is the preferred choice for OPS solutions with the best possible optical switching granularity. OPS interconnects are continuously evolving to fulfil the demands of modern DCNs but such solutions still lack the desired buffering capacity required during header processing. This is due to the unavailability of optical Random Access Memory (RAM) which can limit switch scalability and introduce complex contention resolution mechanisms.

Chapter 3 findings reveal that, for overcoming the switching bottlenecks, the recent trend is to combine the benefits of both OCS and OPS technologies instead of relying on a particular one. Such an approach has been adopted in some current optical switching architectures [111, 173] which have a dedicated routing mechanism based on traffic type and offloads the bursty traffic flows through the OPS while OCS takes care of the long-lived elephant flows.

Evaluation of DCN Network Performance

In order to judge the performance variations arising from various combinations of buffer placement and varying oversubscription parameters, an electronic DCN is evaluated in Chapter 4. A representative Leaf-Spine network was designed to have a fair estimation of the current standards of scalability, cost and power consumption in DCNs. Detailed analytical modelling of Layer 2 switching analogous to a multi-stage tandem queued network quantified the variation in traffic arrival and offloading at the intermediate queueing stages and based on the findings, relevant closed form expressions were derived. Analytical estimations and further performance evaluation revealed the potential bottlenecks created at the Leaf uplink queues directed towards the Spine which experiences heavy contention and significant traffic overloading in oversubscribed networks. While having large buffers in both the Spine and Leaf switches can give the optimal performance, simulation results confirmed that having

large buffers solely in the Spine switches will be unable to control traffic bursts in the Leaf layer and can result in substantial packet loss. Thus with a simpler Layer 2 switching model, the results from the simulated network validated the trend represented in [25] (which used TCP traffic modelling with similar buffer placement ratios as that used in the simulated network) that having larger buffers in the Leaf and smaller in the Spine can perform better than having smaller buffer in the Leaf and larger buffer in the Spine switches.

The performance of a DCN network is analysed in Chapter 5 in the presence of TCP traffic which is currently the most significant contributor of traffic in data centres. Various bottlenecks of TCP Congestion Control in data centres were studied and a DCN specific TCP variant (DCTCP) was implemented for a comparative analysis with standard TCP in the designed DCN environment. Provisions were made to generate TCP traffic flows internally by customising INET traffic generators or by using traffic flow schedules generated externally by DCT^{2}Gen, a DCN specific traffic generator. DCTCP with an ECN marking scheme can react to congestion from an early stage in proportion to its extent and gives better performance than standard TCP with lower buffer occupancy. Results confirm the suitability of DCTCP particularly for larger DCNs with smaller bandwidth delay products where it shows approx. 45% lower latency and better throughput figures than standard TCP. However, in smaller networks without a large number of concurrent traffic flows overloading the buffers and causing packet loss, TCP can give a similiar performance to DCTCP. It was observed that having smaller buffers can reduce queueing delay and make DCTCP suitable for latency sensitive applications but this may result in frequent resizing of window size due to timeouts particularly in Incast scenarios resulting in throughput degradation. So, the ECN marking threshold needs to be chosen with precision for optimal performance.

New Innovations for Future Data Centres

Energy efficiency is a major concern for sustainable development in current DCNs and introducing optical interconnects in the edge tiers is essential for improving the overall DCN power efficiency. An optical interconnect is proposed in Chapter 6, which can minimise power consumption by using passive components (AWGR) as the main switching fabric and utilises the cyclic routing capability of AWGRs to develop a fast optical edge tier switch for handling the intra-rack traffic with proper contention resolution through FDLs. Simulation results shows that the mean latency in the OToR remains lower (within *ns* range) than a similiar sized commodity ToR switch (with latency in μs

range). An all optical architecture is further proposed where the OToR can be interconnected within the DCN through a optical space switch (OSS), based on a non-blocking Clos network, satisfying the scalability demands for future data centres. Power consumption estimations show that the proposed OToR can consume 33% to 55% less power than commodity ToR switches and the OSS Switching layer can also be more energy efficient than and the Spine/Core switching layer. However, currently such optically switched interconnects and DCN networks may not be very cost efficient due to a lack of market availability of optical components, but with the growing development in optical integration technology, the costs are expected to come down.

7.2 Future Work

Apart from the conventional use of optical switching in the upper network tiers, introducing optical interconnects at the top of the rack offers newer advantages in DCN design and to date very few such solutions have been proposed. By using passive components as the main building block, it is possible to increase energy efficiency compared to the electronic counterpart. However since this is a new paradigm in data centre architecture, a lot more remains to be explored. As of now, all-optical networks, even though showing great potential for future DCNs, have only been explored through research and no commercial implementation has yet been possible. Hybrid networks are a more viable option for faster commercialisation of optical interconnects by introducing it along with the existing electronic DCN infrastructure rather than having a complete makeover and so, in that respect, hybrid solutions with Optical ToRs should be looked into.

Core layer switches often suffer from scalability limitations due to their huge radix size but ToR/Leaf switches are normally far smaller in size and so the main bottleneck in OToRs would be proper contention resolution schemes which lead to the possibility of newer options to be investigated. An all-optical switching network has been proposed in Chapter 6 where the OToR can be implemented and this network could be further explored by introducing better traffic scheduling algorithms for faster OCS reconfiguration for routing bursty inter-rack flows. Additionally, considering OToRs as standalone switching components, they might be introduced in both all optical and hybrid networks to replace existing commodity ToR switches. Another prospective area for further investigation is the use of Architecture on Demand (AoD) nodes within all optical/hybrid DCN architectures since they can offer an effective way of offloading traffic onto desired optical paths as per network conditions

instead of having prefixed topologies, and can also introduce multicast/incast traffic handling options through dedicated optical crossconnects.

As evident from the qualitative comparison in Chapter 3, most of the proposed optical architectures follow a tree topology for linking the optical interconnects. Though these networks provide a very high capacity, they require very large switches for handling the core layer traffic. Recursive networks, on the other hand, make use of small or moderate radix switches to scale up gradually in layers while still providing a rich fault-tolerant connectivity. Recently a few optical architectures have ventured in this direction, where the designed networks are logically equivalent to recursive topologies such as DCell (as in [174]) and BCube (as in [162]) and possibly more attention should be paid to this direction in future. Another issue that should not be ignored in future networks is the SDN compliance of the switching architectures. DCNs experience a very diverse mix of traffic and so networks should be able to differentiate between elephant and mice flows with appropriate SDN control and accordingly use different routing and scheduling strategies to prevent the buildup of network congestion.

TCP traffic has been used to judge the network performance in Chapter 5 through simulation which leaves the scope for proposing analytical frameworks for modeling basic TCP behaviour in data centres. Normalised fluid models have been previously used in [149] to do a steady state analysis but using queueing theory models can give a better perspective in this respect with a comprehensive understanding of the network bottlenecks which can result in better resource utilisation. In particular, Incast behaviour which is the main TCP traffic bottleneck needs to be analysed through detailed Markovian models to capture the true impact of many to one workload on the system throughput in the presence of duplicate ACKs and retransmission timeouts. TCP implementations for DCNs are being constantly upgraded for more efficient data transfer and so a general analytical framework needs to be developed which can be customised to include the TCP version-specific features.

Though TCP is the prevalent communication pattern in modern data centres, most of the proposed optical switching architectures do not tend to judge network performance with TCP traffic which possibly leads to an unfair estimation of their efficiency. Hence, TCP traffic should be included in the network performance analysis to gain a detailed understanding of the suitability of any network. However, generating TCP traffic flows is quite tedious and hardly any traffic generator exists (apart from DCT^{2}Gen) that is suited for DCN traffic. Thus, more options in designing simplified TCP traffic generators specifically suited for DCN communication patterns with customisation of different bottleneck scenarios such as Incast and Outcast, should be explored. The developed network model used

for performance analysis of transport layer traffic in electronic DCNs in this thesis can be used to accommodate TCP traffic in optical networks as well if the electronic hosts and routers are replaced with optical modules that can interface with the INET custom nodes. While DCTCP was the chosen congestion control protocol used to judge the DCN performance, there has been a few more DCN specific L4 routing protocols (such as D^2TCP [175], TDCTCP [176], ICTCP [177], MTCP [178] etc.) with enhanced performance guarantee. So a thorough comparative analysis of all suggestive topologies is essential to predict the optimal choice. By changing the Layer 4 TCP interfacing modules, the developed model can be further extended to include such newer data centre specific TCP variants.

Bibliography

[1] Netcraft, Netcraft web server survey, 2009. [Online], Available: http://news.netcraft.com/archives/2016/07/19/july-2016-web-server-survey.html , [Last Accessed: 24^{th} February 2017].

[2] C. Kachris and I. Tomkos, "Optical interconnection networks for data centers," in *Proceedings of the ONDM*, pp. 19-22, 2013.

[3] R. Pries, M. Jarschel, D. Schlosser and M. Klopf, "Power Consumption Analysis of Data Center Architectures," in *Green Communication and Networking*, 2012.

[4] S. Saha, "Data center design: Architecture and energy consumption," PhD. Dissertation, University of Nebraska Lincoln, 2012.

[5] R. Ho, H. Schwetman, M. O. McCracken, P. Koka, J. Lexau, J. E. Cunningham, X. Zheng and A. V. Krishnamoorthy, "Optical systems for data centers," in *OFC/NFOEC* 2011.

[6] IBM Academy of Technology, "Cloud Computing Simplified: The Thoughts on Cloud Way," in *IBM redbooks*, [Online], Available: http://www.redbooks.ibm.com/, 2015.

[7] A. Greenberg, J. Hamilton, D. A. Maltz and P. Patel, "The cost of a cloud: Research problems in data center networks," in *Proceedings of ACM SIGCOMM*, 2008.

[8] Cisco Systems, "Cisco Global Cloud Index: Forecast and Methodology," White Paper, 2015-2020, November 2016.

[9] Cisco Systems, "Growth in the cloud," [Online]. Available: http://www.cisco.com/c/dam/assets/sol/sp/network_infrastructure/growth_cloud_infographic.html , 2016, [Last Accessed: 10^{th} April 2017].

[10] D. S. Marcon, R.R. Olieveira, L.P. Gaspary, M.P. Barcellos, "Datacenter Networks and Rel-

evant Standards," in *Cloud Services, Networking and Management.*, 1st Edition, IEEE-Wiley Series on Network Management, 2015.

[11] R Recio, "The coming decade of data center networking discontinuities," in *ICNC*, Keynote Speaker, August 2012.

[12] D. Tardent and N. Farrington, "Building a data-center network with optimal performance and economy," in *Lightwave Online*, Vol. 31, Issue 4, 10^{th} July, 2014.

[13] D. Kilper, K. Bergman, V. Chan, I. Monga, G. Porter, and K. Roauschenbach, "Optical networks come of age," in *Optics and Photonics News*, Vol 25, Issue 9, pp. 50-57, 2014.

[14] B. Wang, Z. Qi, R. Ma, H. Guan, Athanasios V. Vasilakos, "A survey on data center networking for cloud computing," in *Computer Networks*, Vol. 91, pp. 528-547, 2015.

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

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

[17] A. Roy, H. Zengy, J. Baggay, G. Porter, and A. C. Snoeren, "Inside the social networks (Datacenter) network," in *ACM SIGCOMM*, August 2015.

[18] P. Bodk, I. Menache, M. Chowdhury, P. Mani, David A. Maltz, and I. Stoica, "Surviving failures in bandwidth-constrained datacenters," in *ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, August 2012.

[19] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel and S. Sengupta, "VL2: a scalable and flexible data center network," in *ACM SIGCOMM Computer Communication Review*, Vol. 39. Issue. 4, ACM, 2009.

[20] D. Abts and B. Felderman, "A guided tour of data-center networking," in *Communication ACM*, Vol. 55, Issue 6, pp. 44-51, June 2012.

[21] L. Guo and I. Matta, "The War between Mice and Elephants," in *Proceedings of the Ninth International Conference on Network Protocols*, November 2001.

[22] J. Thomson-Melanson, "Learn about Data Center Fabric Fundamentals," Juniper Networks, [Online], Available: `http://www.juniper.net/techpubs/en_US/learn-`

about/LA_DCFabricsFundamentals.pdf , [Last Accessed: 8^{th} March 2017].

[23] D. Lin, Y. Liu, M. Hamdi, and J. Muppala, "Flatnet: Towards a flatter data center network," in *IEEE Global Communications Conference (GLOBECOM)*, 2012.

[24] T. Telgin, N. Pleros, R. Pitwon and A. Hakansson, "Data Center Architectures," in *Optical Interconnects for Data Centers*, Woodhead Publishing, 1st Edition, November 2016.

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

[26] F. Durr, "A Flat and Scalable Data Center Network Topology Based on De Bruijn Graphs," Networking and Internet Architecture, arXiv:1610.03245, October 2016.

[27] Ericsson, "Hyperscale Cloud-Reimagining Data Centers From Hardware To Applications," White Paper, May 2016.

[28] S. Kipp, B. Smith, C. Cole, M. Nowell, "The State of Ethernet Optics," in *Ethernet Alliance, OFC*, March 2016.

[29] C. Urricariet, "Latest Trends in Optical Interconnects," in *NANOG (North American Network Operators Group)*, February 2016.

[30] Crehan Research Inc., "25GbE and 100GbE to Comprise Over Half of all Data Center Ethernet Switch," Press Release, [Online]. Available: http://www.crehanresearch.com/wp-content/uploads/2017/01/CREHAN-Data-Center-Networking-PR-January-2017.pdf , [Last Accessed: 14^{th} March 2017].

[31] M. Machowinski, "Networking Ports: 1G, 2.5G, 40G, 100G Market Tracker," *IHS.com*, Nov. 2016, [Online], Available: https://technology.ihs.com/573062/networking-ports-1g-25g-10g-25g-40g-100g-market-tracker-regional-h2-2016 , [Last Accessed: 5^{th} March 2017].

[32] Z. Ali. Kahn, "Project Falco: Decoupling Switching Hardware and Software," in Engineering.linkedin.com, [Online]. Available: https://engineering.linkedin.com/blog/2016/02/falco-decoupling-switching-hardware-and-software-pigeon , [Last Accessed: 6^{th} March 2017].

[33] A. Eckert, L. M. Garcia, R. Niazmand, X. Wang, "Wedge 100: More open

and versatile than ever," in Facebook Code, 2017. [Online]. Available: https://code.facebook.com/posts/1802489260027439/wedge-100-more-open-and-versatile-than-ever/ , [Last Accessed: 6^{th} March 2017].

[34] Z. Yao, J. Bagga, H. Morsy, "Introducing Backpack: Our second-generation modular open switch," in Facebook Code, 2017. [Online]. Available: https://code.facebook.com/posts/864213503715814/introducing-backpack-our-second-generation-modular-open-switch/, [Last Accessed: 6^{th} March 2017].

[35] X. Wu and X. Yang, "Dard: Distributed adaptive routing for datacenter networks," in *IEEE Distributed Computing Systems (ICDCS)*, June 2012.

[36] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang and A. Vahdat, "Hedera: Dynamic flow scheduling for data center networks," in *Proceedings of the 7th USENIX conference on Networked systems design and implementation*, 2010.

[37] M. Al-Fares, A. Loukissas, and A. Vahdat. "A scalable, commodity data center network architecture," in *ACM SIGCOMM Computer Communication Review*, Vol. 38, Issue. 4, 2008.

[38] 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 Computer Communication Review*, Vol. 39, Issue.4, pp. 63-74, 2009.

[39] H. Abu-Libdeh, P. Costa, A. Rowstron, G. O'Shea and A. Donnelly, "Symbiotic routing in future data centers," in *ACM SIGCOMM Computer Communication Review*, Vol. 40, Issue. 4, pp. 51-62, 2010.

[40] R. Niranjan Mysore, et al., "PortLand: a scalable fault-tolerant layer 2 data center network fabric," in *SIGCOMM Computer Communication Review*, Vol. 39, Issue. 4, pp. 39-50, 2009.

[41] Y. Liu, J. K. Muppala, M. Veeraraghavan, D. Lin and M. Hamdi, "Data center networks: Topologies, architectures and fault-tolerance characteristics," in *Springer Briefs in Computer Science*, 2013.

[42] C. Hopps, "Analysis of an Equal-Cost Multi-Path Algorithm," RFC 2992, 2000.

[43] A. Greenberg, P. Lahiri, David, A. Maltz, P. Patel, and S. Sengupta, "Towards a next generation data center architecture: scalability and commoditization," in *ACM PRESTO*, 2008.

[44] A. Andreyev, "Introducing datacenter fabric, the next-generation Facebook datacenter network," [Online]. Available: https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network , [Last Accessed: 6^{th} March 2017].

[45] Dennis Abts and Bob Felderman, "A guided tour of data-center networking," in *ACM Queue*, Vol. 55, Issue. 6, pp. 44-51, June 2012.

[46] Arista Networks, "Cloud Networking: Scaling Out Data Center Networks, White Paper, 2016. [Online]. Available: https://www.arista.com/assets/data/pdf/Whitepapers/Cloud_Networking__Scaling_Out_Data_Center_Networks.pdf , [Last Accessed: 6^{th} March 2017].

[47] W. J. Dally, B. Towles, "Principles and practices of interconnection networks," Morgan Kaufmann Publishers, 2004.

[48] D. Abts, J. Kim, "High performance datacenter networks: Architectures algorithms and opportunity," in *Synthesis Lectures on Computer Architecture*, Morgan and Claypool Publishers, March 2011.

[49] L. Popa, S. Ratnasamy, G. Iannaccone, A. Krishnamurthy, and I. Stoica, "A cost comparison of datacenter network architectures," in *ACM CoNEXT*, 2010.

[50] C. Clos, "A study of non-blocking switching networks," in *Bell System Technical Journal*, Vol. 32, Issue. 2, pp. 406-424, March 1953.

[51] L. G. Valiant, "A scheme for fast parallel communication," in *SIAM Journal on Computing*, Vol. 11, Issue. 2, pp. 350-361, 1982.

[52] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown. "Elastictree: Saving energy in data center networks," in *USENIX NSDI*, April 2010.

[53] BBN Laboratories, "Butterfly Parallel Processor Overview," in *BBN Report #6148*, Version I, Cambridge, MA, March 1986.

[54] J. Kim, J. Balfour, and W. Dally. "Flattened butterfly topology for on-chip networks," in *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*, IEEE Computer Society, 2007.

[55] C. Guo, H. Wu and K. Tan, "Dcell: a scalable and fault-tolerant network structure for data centers," in *ACM SIGCOMM Conference on Data Communication*, 2008.

[56] D. Li, C. Guo and H. Wu, "Ficonn: using backup port for server interconnection in data centers," in *IEEE INFOCOM*, pp. 2276-2285. 2009.

[57] L. Bhuyan and D. Agrawal, "Generalized Hypercube and Hyperbus Structures for a Computer Network," in *IEEE Transactions on Computers*, April 1984.

[58] L. Gyarmati and T. A. Trinh, "Scafida: A scale-free network inspired data center architecture," in *SIGCOMM Computer Communications Review*, Vol. 40, Issue. 5, October 2010.

[59] A. L. Barabsi and R. Albert. "Emergence of Scaling in Random Networks," in *Science*, Vol.286, Issue.5439, pp. 509-512, 1999.

[60] A. Singla, C. Y. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking data centers randomly," in *USENIX NSDI*, 2012.

[61] C. Wu and R. Buyya, "Cloud Data Centers and Cost Modeling," Morgan Kaufmann Publishers, 2015.

[62] Microsoft, "Microsofts Cloud Infrastructure: Datacenters and Network Fact Sheet," [Online]. Available: `http://download.microsoft.com/download/8/2/9/8297F7C7-AE81-4E99-B1DB-D65A01F7A8EF/Microsoft_Cloud_Infrastructure_Datacenter_and_Network_Fact_Sheet.pdf`, [Last Accessed: 8^{th} March 2017].

[63] Y. Zhang, N. Ansari, "On architecture design congestion notification tcp incast and power consumption in data centers," in *IEEE Communication Surveys & Tutorial*, Vol. 15, 2013.

[64] K. Nagaraj, H. Khandelwal, C. Killian, and R. R. Kompella, "Hierarchy-aware distributed overlays in data centers using DC2," in *Fourth International Conference on Communication Systems and Networks (COMSNETS)*, pp. 1-10, January 2012.

[65] H. Ballani, K. Jang, T. Karagiannis, C. Kim, D. Gunawardena, and G. OShea, "Chatty Tenants and the Cloud Network Sharing Problem," in *USENIX NSDI*, 2013.

[66] R. Shea, F. Wang, H. Wang, and J. Liu, "A Deep Investigation Into Network Performance in Virtual Machine Based Cloud Environment," in *IEEE INFOCOM*, 2014.

[67] Y. Chen, R. Griffith, J. Liu, et al. "Understanding TCP incast throughput collapse in datacenter

networks," in *Proceedings of 1st ACM Workshop on Research on Enterprise Networking*, pp. 73-82, 2009.

[68] Alan Shieh, Srikanth Kandula, Albert Greenberg, Changhoon Kim, and Bikas Saha. "Sharing the data center Network," in *USENIX NSDI*, 2011.

[69] M. Poess and R. O. Nambiar, "Energy cost, the key challenge of todays data centers: A power consumption analysis of TPC-C results," in *Proceedings of VLDB Endowment*, Vol. 1, Issue. 2, pp. 1229-1240, August 2008.

[70] M. Dayarathna, Y. Wen, R. Fan, "Data Center Energy Consumption Modeling: A Survey," in *IEEE Communication Surveys and Tutorials*, Vol. 18, Issue. 1, pp. 732-794, 2016.

[71] R. Buyya, C. Vecchiola, and S. Selvi, "Mastering Cloud Computing: Foundations and Applications Programming," Morgan Kaufmann Publishing (imprint of Elsevier), 2013.

[72] "Make IT Green: Cloud Computing and its Contribution to Climate Change," Greenpeace International, 2010.

[73] P. Delforge, "America's Data Centers Consuming and Wasting Growing Amounts of Energy," National Resources Defense Counsel, [Online], Available: `https://www.nrdc.org/resources/americas-data-centers-consuming-and-wasting-growing-amounts-energy` , [Last Accessed: 12^{th} March 2017].

[74] A. S. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030," in *Challenges 2015*, Vol. 6, Issue. 1, pp. 117-157, 2015.

[75] L. A. Barroso and U. H. Olzle, "The Case for Energy Proportional Computing," in *IEEE Computer*, Vol. 40, Issue. 12, pp. 33-37, December 2007.

[76] D. Li, M. Xu, H. Zhao, X. Fu, "Building mega data center from heterogeneous containers," in *Proceedings of IEEE ICNP*, Vancouver, Canada, 2011.

[77] E. Agrell, et al., "Roadmap of optical communications," in *Journal of Optics*, Vol. 18, 2016.

[78] "High Performance and Embedded Architecture and Compilation", *HiPEAC Vision 2015 Report*.

[79] D. A. B. Miller, "The role of optics in computing," in *Nature Photonics*, Vol. 4, 2010.

[80] R. D. Williams, T. Sze, D. Huang, S. Pannala and C. Fang, "Server memory roadmap," in

JEDEC Report, 2012.

[81] Arista 7100 Series. December 2013. [Online], Available: http://bit.ly/VBo1Ll.

[82] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," in *IEEE Communication Survey Tutorials*, Vol. 14, Issue. 4, pp. 1021-1036, 2012.

[83] A. Mohammad, A. Kabbani, T. Edsall, B. Prabhakar, A. Vahdat, and M. Yasuda, "Less is more: Trading a little bandwidth for ultra-low latency in the data center," in *9th USENIX Conference on Networked Systems Design and Implementation*, San Jose, CA, 2012.

[84] A. Ghiasi, "Large data centers interconnect bottlenecks," in *Optics Express*, Vol. 23, Issue. 3, pp. 2085-2090, 2015.

[85] Google Data Centers, [Online], Available: https://www.google.com/about/datacenters/efficiency/internal/, [Last Accessed: 12^{th} March 2017].

[86] J. Park, "Designing a Very Efficient Data Center," [Online], Available: https://www.facebook.com/notes/facebook-engineering/designing-a-very-efficient-data-center/10150148003778920/ , [Last Accessed: 12^{th} March 2017].

[87] A. Hurson and H. Azad, "Energy Efficiency in Data Centers and Clouds," Academic Press, 1st edition, 2016.

[88] A. Tzanakaki, et al., "Energy efficiency in integrated IT and optical network infrastructures: The geysers approach," in *Computer Communications Workshops, IEEE INFOCOM*, pp. 343-348, April, 2011.

[89] C. Kachris, K. Kanonakis, and I. Tomkos, "Optical Interconnection Networks in Data Centers: Recent Trends and Future Challenges," in *IEEE Communications Magazine*, Vol. 14, pp. 39-45, September 2013.

[90] P. Pepeljugoski et al., "Low Power and High Density Optical Interconnects for Future Supercomputers," in *OFC*, 2010.

[91] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright, "Power awareness in network design and routing," in *27th IEEE INFOCOM Conference on Computer Communications*, 2008.

[92] A. Benner, "Optical interconnect opportunities in supercomputers and high end computing" in *Optical Fiber Communication Conference*, pp. 1-60, 2012.

[93] US Department of Energy, "Vision and roadmap: Routing telecom and data centers toward efficient energy use," in *Vision and Roadmap Workshop on Routing Telecom and Data Centers*, pp. 1-27, 2009.

[94] 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*, October 2010.

[95] 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*, October 2010.

[96] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen and A. Vahdat, "Integrating microsecond circuit switching into the data center," in *ACM SIGCOMM*, pp. 447-458, 2013.

[97] H. Liu, F. Lu, A. Forencich, R. Kapoor, M. Tewari, G.M. Voelker, G. Papen, A.C. Snoeren and G. Porter, "Circuit switching under the radar with reactor," in *USENIX NSDI*, pp. 1-15, 2014.

[98] 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.

[99] 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" in *USENIX NSDI*, pp. 1-14, 2012.

[100] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," in *ACM SIGCOMM HotNets*, pp. 1-6, 2010.

[101] K. Chen, X. Wen, X. Ma, Y. Chen, Y. Xia, C. Hu, and Q. Dong, "WaveCube: A scalable, fault-tolerant, high-performance optical data center architecture", in *IEEE INFOCOM*, pp. 1-9, 2015.

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

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

[104] Xiaohui Ye, Yawei Yin, S. J. B. Yoo, Paul Mejia, Roberto Proietti, and Venkatesh Akella, "DOS: A scalable optical switch for datacenters," in *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, pp. 1-12, 2010.

[105] 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," in *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 19, Issue. 2, March 2013.

[106] 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, October 2015.

[107] Ronald P. Luijten, Richard R. Grzybowski, and Roe Hemenway, "Optical interconnection networks:The OSMOSIS project," in *17th Annual Meeting of the IEEE Lasers and Electro-Optics Society*, 2004.

[108] 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," in *Journal of Optical Communications and Networking*, Vol. 3, Issue. 8, pp. 1-11, August 2011.

[109] M. Fiorani,S. Aleksic, M. Casoni, L. Wosinska, and J. Chen, "Energy-efficient elastic optical interconnect architecture for data centers," in *IEEE Communication Letters*, Vol. 18, pp. 1531-1534, 2014.

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

[111] 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)*, September 2014.

[112] A. Hammadi, T. E. H. El-Gorashi, and J.M.H. Elmirghani, "High performance AWGR PONs in data centre networks," in *Proceedings of ICTON*, Hungary, 2015.

[113] C. Kachris and I. Tomkos, "Power consumption evaluation of hybrid wdm pon networks for

data centers," in *2011 16th European Conference on Networks and Optical Communications*, pp. 118-121, 2011.

[114] Calient Technologies, "The Software Defined Hybrid Packet Optical Datacenter Network," in *OFC*, 2013.

[115] SFlow, "SDN control of hybrid packet / optical leaf and spine network," [Online], Available: http://blog.sflow.com/2014/09/sdn-control-of-hybrid-packet-optical.html , [Last Accessed: 18^{th} March 2017].

[116] J. Chen, Y. Gong, M. Fiorani and S. Alexsic, "Optical interconnects at top of the rack for energy-efficient datacenters," in *IEEE Commun. Mag.*, Vol. 53, Issue. 8, August 2015.

[117] R. M. Indre, J. Pesic, and J. Roberts, "POPI: A passive optical Pod interconnect for high performance data centers," in *Proceedings of IEEE ONDM*, pp. 84-89, May 2014.

[118] SdxCentral, " 2017 Next Gen Data Center Networking Report," Market Report, 2017. [Online]. Available: http://www.communicationstoday.co.in/images/reports/20170419-sdxcentral-next-gen-data-center-networking-report.pdf , [Last Accessed: 6^{th} September 2017].

[119] Cisco Systems, " Software Defined Networking (SDN) and its role in automating the provision and scaling of cloud native architectures in the data center through network programmability," White Paper, [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/enterprise/cisco-on-cisco/cs-en-08022017-sdn-thesis.pdf , 2017, [Last Accessed: 6^{th} September 2017].

[120] C. Hintz and C. Obediente, "Next Generation Data Center - Spine Leaf Fabric Design," [Online], Available: http://dcnextgen.blogspot.ie/2014/07/next-generation-data-center-spine-leaf.html , [Last Accessed: 18^{th} March 2017].

[121] Ethernet Storage Forum, "Controlling Congestion in New Storage Architectures" in *SNIA Ethernet Storage Forum*, [Online], Available: https://www.snia.org/sites/default/files/ESF/DCC_SNIA_ESF_Final.pdf , [Last Accessed: 18^{th} March 2017].

[122] Arista Networks, "Arista 7300 Series Tech. Specifications," [Online], Available: https://www.arista.com/en/products/7300-series, [Last Accessed: 18^{th} March 2017].

[123] Juniper Networks, "QFX10000 Modular Ethernet Switches," Data Sheet, [Online], Available: http://www.juniper.net/assets/us/en/local/pdf/datasheets/1000529-en.pdf, [Last Accessed: 19^{th} March 2017].

[124] Cisco Networks, "Ciscos Massively Scalable Data Center," [Online], Available: http://www.cisco.com/c/dam/en/us/td/docs/solutions/Enterprise/Data_Center/MSDC/1-0/MSDC_AAG_1.pdf, [Last Accessed: 19^{th} March 2017].

[125] Arista Networks, "Cloud Networking: Scaling Out Data Center Networks," White Paper, [Online], Available: https://www.arista.com/assets/data/pdf/Whitepapers/Cloud_Networking__Scaling_Out_Data_Center_Networks.pdf , [Last Accessed: 19^{th} March 2017].

[126] F. Yan, W. Miao, H. Dorren and N. Calabretta, "On the cost, latency, and bandwidth of LIGHTNESS data center network architecture," in *IEEE Photonics in Switching*, pp. 130-132, 2015.

[127] Arista Networks, "Arista datasheet for 7500 Series Data Center Switch," [Online], Available: http://www.aristanetworks.com/media/system/pdf/Datasheets/7500_Datasheet.pdf, [Last Accessed: 26^{th} March 2017].

[128] Arista Networks, "Arista datasheet for 7050QX Series Data Center switch, 2016," [Online], Available: https://www.arista.com/assets/data/pdf/Datasheets/7050S_Datasheet.pdf , [Last Accessed: 26^{th} March 2017].

[129] Cisco, "Cisco Nexus 7000 Series Switches, 2013," [Online], Available: http://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-10-slot-switch/Data_Sheet_C78-437762.html , [Last Accessed: 26^{th} March 2017].

[130] Arista Networks, "Arista 7500 Scale-Out Cloud Network Designs," White Paper, [Online], Available: http://www.arista.com/assets/data/pdf/Whitepaper/Arista_7500_Scale_Out_Designs.pdf , [Last Accessed: 28^{th} March 2017].

[131] Arista Networks, "Arista datasheet for 7300 Series Data Center Switch," [Online], Available: https://www.arista.com/assets/data/pdf/Datasheets/7300X_DS.pdf , [Last Accessed: 29^{th} March 2017].

[132] Cisco, "Cisco Nexus 9500 Platform Common Equipment Data Sheet" [Online], Avail-

able: http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-729404.html , [Last Accessed: 29th March 2017].

[133] Finisar, "Multi-Rate SFP+ Transceiver, Product Specification" [Online], Available: https://www.finisar.com/sites/default/files/downloads/finisar_ftlx1672d3btl_10gbase-er_40km_sfp_optical_transceiver_product_specb1.pdf , [Last Accessed: 29th March 2017]

[134] Colfaxdirect.com, "Online Catalogue," [Online], Available: http://www.colfaxdirect.com/store/pc/viewCategories.asp?idCategory=7 , [Last Accessed: 2nd April 2017].

[135] Convergedigest.com, [Online], Available: http://www.convergedigest.com/2013/11/arista-introduces-switches-for-cloud.html , [Last Accessed: 2nd April 2017].

[136] Vion.com, "DataCenter Price Catalogue," [Online], Available: http://www.vion.com/assets/site_18/files/vion%20dir-tso-2710%20price%20list%2019nov14.pdf , [Last Accessed: 2nd April 2017].

[137] QSFPS.com, "Online Catalogues," [Online], Available: http://www.qsfps.com , [Last Accessed: 2nd April 2017].

[138] Discrete Event Simulator, OMNeT++. [Online], Available: http://www.omnetpp.org

[139] D. Gross, C.M. Harris, "Fundamentals of queueing theory", Wiley, Chichester, 1985.

[140] V. S. Frost and B. Melamed, "Traffic Modeling for Telecommunications Networks," in *IEEE Communications*, March, 1994.

[141] OPEN Compute Project, "Data Center - Mechanical Specifications," version 1.0.

[142] A. Bechtolsheim, L. Dale, H. Holbrook, and A. Li, "Why Big Data Needs Big Buffer Switches," White Paper, [Online], Available: http://www.arista.com/assets/data/pdf/Whitepapers/BigDataBigBuffers-WP.pdf.

[143] L. Kleinrock, "Queuing Systems" in *Volume I: Theory*, John Wiley, Inc., 1975.

[144] P. J. Burke, "The Output of a Queuing System," in *Operations Research*, Vol. 4, Issue. 6, pp.

699-704, 1956.

[145] J. R. Jackson. "Networks of waiting lines," in *Operations Research*, Vol. 5, August 1957.

[146] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R.Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Holzle, S. Stuart and A. Vahdat "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenters." in *SIGCOMM*, 2015.

[147] N. Calabretta, R. Centelles, S. Di Lucente, and H. Dorren, "On the Performance of a Large-Scale Optical Packet Switch Under Realistic Data Center Traffic" in *Journal of Optical Communications and Networking*, Vol. 5, Issue. 6, pp. 565-573, June 2013.

[148] Y. Qin, W. Yang , Y. Ye and Y. Shi, "Analysis for TCP in data center networks: Outcast and Incast" in *Journal of Network and Computer Applications*, pp. 140-150, June 2016.

[149] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, "Data center TCP (DCTCP)" in *ACM SIGCOMM Computer Communication Review*, pp. 63-74, 2010.

[150] INET Framework, [Online], Available: `https://inet.omnetpp.org/`.

[151] P. Wette and H. Karl, "DCT^{2}Gen: A traffic generator for data centers" in *Computer Communications*, pp. 45-58, April, 2016.

[152] 5. S. Liu, H. Xu and Z. Cai, "Low latency datacenter networking: A short survey", [Online] Available: `http://arxiv.org/abs/1312.3455`, 2013.

[153] P. Sreekumari and J. Jung, "Transport protocols for data center networks: a survey of issues, solutions and challenges," in *Photonic Network Communications*, pp. 1-17, 2015.

[154] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron, "Better Never than Late: Meeting Deadlines in Datacenter Networks" in *Proceedings of the ACM SIGCOMM conference*, 2011.

[155] C. Lee, K. Jang and S. Moon, "Reviving Delay-based TCP for Data Centers" in *Proceedings of the ACM SIGCOMM conference*, 2012.

[156] J. Kurose and K. Ross, "Computer Networking A Top-Down Approach", 4th edition, Addison Wesley, 2008.

[157] S. Floyd and V. Jacobson, "Random Early Detection gateways for congestion avoidance," in

IEEE/ACM Transactions on Networking, Vol. 1, Issue. 4, August 1997.

[158] T. Cormen, C. Leiserson, L. Rivest, C. Stein, "Introduction to Algorithms", Section 24.3: Dijkstra's algorithm, MIT Press and McGrawHill. Second edition, pp. 595-601, 2001.

[159] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron, "Better Never than Late: Meeting Deadlines in Datacenter Networks" in *Proceedings of the ACM SIGCOMM conference*, 2011.

[160] Y. Gong, X. Hong, Y. Lu, S. He and J. Chen, "Passive optical interconnects at top of the rack: offering high energy efficiency for datacenters," in *Optics Express*, Vol. 23, Issue. 6, pp. 7957-7970, 2015.

[161] L. Cottrell, W. Matthews, and C. Logg, "Tutorial on Internet Monitoring & PingER at SLAC," [Online], Available: `http://www.slac.stanford.edu/comp/net/wan-mon/tutorial.html`, [Last Accessed: 6^{th} May 2017].

[162] F. Yan, W. Miao, O. Raz and N. Calabretta, "Opsquare: A flat DCN architecture based on flow-controlled optical packet switches," in *IEEE/OSA Journal of Optical Communications and Networking*, Vol. 9, Issue. 4, pp. 291-303, 2017.

[163] C. Kachris and I. Tomkos, "Power consumption evaluation of all optical data center networks," in *Cluster Computing*, Vol. 16, Issue. 3, pp. 611-623, 2013.

[164] R. S. Tucker, "The role of optics and electronics in high-capacity routers," in *Journal of Lightwave Technology*, Vol. 24, Issue. 12, pp. 4655-4673, 2006.

[165] Arista Networks, "7250QX-64 Switch Datasheet," [Online], Available: `https://www.arista.com/assets/data/pdf/Datasheets/7250QX-64_Datasheet_S.pdf`, [Last Accessed: 12^{th} May 2017].

[166] S. Han, T. J. Seok, N. Quack, B. Yoo and M. C. Wu ,"Monolithic 50x50 mems silicon photonic switches with microsecond response time," in *OFC*, 2014.

[167] L. Zhao, T. Ye, and W. Hu, "Nonblocking Clos networks of multiple ROADM rings for mega data centers," in *Optics Express*, Vol. 23, October 2015.

[168] S. Basu, C. McArdle and L. P. Barry, "Scalable OCS-based Intra/Inter Data Center Network with Optical ToR Switches" in *ICTON*, July 2016.

[169] Z. Hua and S. Zhong, "Scalable and topology adaptive intra-data center networking enabled

by wavelength selective switching," in *OFC*, 2014.

[170] Calient Technologies, "S320 Photonic Switch Hardware User Manual," [Online], Available: http://www.calient.net/wp-content/uploads/downloads/2013/04/CALIENT-S-Series-Photonic-Switch-Hardware-User-Manual-Rev-A-460xxx-00-v10.pdf, [Last Accessed: 19^{th} May 2017].

[171] Polatis, "Series 7000n Network Optical Matrix Switch datasheet," [Online], Available: https://www.phoenixdatacom.com/wp-content/uploads/2017/02/Polatis_384x384_7000n_Network_Optical_Switch_pdl.pdf , [Last Accessed: 26^{th} May 2017]

[172] K. Ueda, Y. Mori, H. Hasegawa, K. Sato and T. Watanabe, "Large-Scale and Simple-Configuration Optical Switch Enabled by Asymmetric-Port-Count Subswitches," in *IEEE Photonics Journal*, Vol. 8, Issue. 2, 2016.

[173] M. Fiorani, S. Aleksic, and M. Casoni, "Hybrid Optical Switching for Data Center Networks, in *Journal of Electrical and Computer Engineering*, Vol. 2014, Article ID 139213, 2014.

[174] X. Hong, Y. Yang, Y. Gong and J. Chen, "Passive optical interconnects based on cascading wavelength routing devices for datacenters: A cross-layer perspective," *in IEEE/OSA Journal of Optical Communications and Networking*, Vol. 9, Issue. 4, pp. 45-53, 2017.

[175] B. Vamanan, J. Hasan and T. N. Vijaykumar, "Deadline-Aware Datacenter TCP (D2TCP)," in *Proceedings of the ACM SIGCOMM conference (SIGCOMM 12)*, August 2012.

[176] T. Das and K. M. Sivalingam, "TCP improvements for data center networks," in *Proceedings of 5th International Conference on Communication Systems & Networks (COMSNETS)*, 2013.

[177] W. Haitao, Z. Feng, C. Guo, Y. Zhang, "ICTCP: Incast congestion control for TCP in data-center networks" in *IEEE/ACM Transactions on Networking*, Vol. 21, pp. 345-358, 2013.

[178] S. Fang, C. H. Foh and K. M. M Aung, "Prompt congestion reaction scheme for data center network using multiple congestion points," in *IEEE International Conference on Communications (ICC)*, pp. 2679-2683, 2012.

List of Publications

1. S. Basu, C. McArdle and L. P. Barry, ”Scalable OCS-based Intra/Inter Data Center Network with Optical ToR Switches” in *18th International Conference on Transparent Optical Networks (ICTON)*, Trento, Italy, July 2016.

2. 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.

3. S. Basu, C. McArdle and L. P. Barry, ”Optical Switching in Datacenter Networks: Outlook” in *Photonics Ireland*, Cork, Sep. 2015.