Arrayed Waveguide Grating Router and server-based Passive Optical Network data centres

Randa Abduljabbar Thabit

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Electronic and Electrical Engineering

March, 2020

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapter 4 of the thesis has appeared in the following publication:

Alani, R., Hammadi, A., El-Gorashi, T.E. and Elmirghani, J.M., 2017, July. PON data centre design with AWGR and server-based routing. In 2017 19th International Conference on Transparent Optical Networks (ICTON) (pp. 1-4). IEEE.

The candidate modelled the problem and provided solutions for it, also gathered data from literature to motivate the work.

Ali Hammadi's work was used as the basis for the extensions.

Dr Taisir held regular meetings to confirm and validate the findings.

Prof. Jaafar Elmirghani came up with the concept and also provided guidance throughout the development of the work.

The work in Chapter 5 of the thesis has appeared in the following publication:

Alani, R.A., El-Gorashi, T.E. and Elmirghani, J.M., 2019, July. Virtual Machines Embedding for Cloud PON AWGR and Server Based Data Centres. In *2019 21st International Conference on Transparent Optical Networks (ICTON)* (pp. 1-5). IEEE.

The candidate modelled the problem and provided solutions for it, also gathered data from literature to motivate the work.

Dr Taisir held regular meetings to confirm and validate the findings.

Prof. Jaafar Elmirghani came up with the concept and also provided guidance throughout the development of the work.

The work in Chapter 6 of the thesis has appeared in the following publication:

Thabit, R.A., El-Gorashi, T.E. and Elmirghani, J.M., 2020. Resilient AWGR and server based PON data centre architecture. In *2020 22nd International Conference on Transparent Optical Networks (ICTON)*. IEEE*.*

The candidate modelled the problem and provided solutions for it, also gathered data from literature to motivate the work.

Dr Taisir held regular meetings to confirm and validate the findings.

Prof. Jaafar Elmirghani came up with the concept and also provided guidance throughout the development of the work.

The work in Chapters 4 and 5 of the thesis has appeared in the following publication:

Thabit, R.A., El-Gorashi, T.E. and Elmirghani, J.M., Resource allocation in AWGR and server based PON data centre architecture, to be submitted to IEEE Access.

The candidate modelled the problem and provided solutions for it, also gathered data from literature to motivate the work.

Dr Taisir held regular meetings to confirm and validate the findings.

Prof. Jaafar Elmirghani came up with the concept and also provided guidance throughout the development of the work.

The work in Chapter 6 of the thesis has appeared in the following publication:

Thabit, R.A., S. H. Mohamed, El-Gorashi, T.E. and Elmirghani, J.M., Resilient PON-based data centre architectures, to be submitted to IEEE Access.

The candidate modelled the problem and provided solutions for it, also gathered data from literature to motivate the work for three PON based data centre architectures.

Sanaa M. modelled the problem and provided solutions for it, also gathered data from literature to motivate the work for two PON based data centre architectures

Dr Taisir held regular meetings to confirm and validate the findings.

Prof. Jaafar Elmirghani came up with the concept and also provided guidance throughout the development of the work.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Professor Jaafar Elmirghani and Dr Taisir Elgorashi for accepting to supervise me and for their continuous support of my PhD study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better supervisors and mentors for my PhD study.

I would like to acknowledge the Iraqi Ministry of Higher Education and Scientific Research (MOHE) and AL-Nahrain University for funding my PhD.

From all of my heart I would like to say a special thank you for my husband Mustafa for his help, support, advice and endless love. Without him I would never be able to complete my thesis. I also want to express my gratitude to my special daughter Basma for her patience, understanding and support and to my lovely twins Sama and Mazin for the joy they brought to my life especially during my PhD journey. Also, I wish to express my love and gratitude to all my family, especially to the memory of my father, Abdul-Jabbar for his wisdom and patience that always inspired me, to my mother, Batool, for her care and support, my kind brother Muhened and my sisters Hind, Hala and especially Nadia for her endless support during my PhD.

Also, I want to express my gratitude to Victoria Master and Liz Oxley for their valuable support that made a huge difference in my life. Finally, I would like to extend the acknowledgement to all my friends especially Baraa and all my colleagues at the University of Leeds for their friendship, support and valuable discussions.

Abstract

The continuous growth in Internet-connected devices results in significant increase in data centres' traffic which in turn necessitates the development of scalable, high bandwidth, low power consumption data centre architectures. Passive Optical Networks (PONs) with their proven performance in access networks can provide efficient solutions to support connectivity inside modern data centres. Recently, different novel PON architectures were proposed to manage the inter-rack and intra-rack communication in a data centre.

In this thesis, we further studied one of the PON designs proposed earlier by our group where routing is performed by Arrayed Waveguide Grating Routers (AWGRs) and servers. This work is the first to analyse the AWGR and server-based PON data centre architecture. A Mixed Integer Linear Programming (MILP) model was developed, where the routing and wavelength assignment involving the AWGRs of the PON cell is optimised to support inter group interconnections. Also, the power consumption of this design was compared to a traditional server-based data centre architecture, BCube. Our study showed that the AWGR and server-based data centre architecture reduced the power consumption by 83% as compared to the standard BCube architecture. In addition, we further investigated this AWGR and server-based PON data centre architecture in cloud applications. We developed a MILP model along with a heuristic that minimise power consumption by optimising the embedding of virtual machine requests which is achieved by optimising the servers selected to host VMs. Our study showed that the AWGR and server-based PON data centre architecture reduced the power consumption by up to 34%, compared to the non-optimised embedding model of VMs which tries to fulfil all the requests by maximising the number of VM requests served. Furthermore, a third MILP model was developed to evaluate the resilience of the AWGR and server-based PON data centre architecture modified designs. We also studied the impact of failure of the main components in the proposed PON data centre architecture. In particular, we evaluated the impact of different failure modes on the power consumption and delay of the proposed AWGR and server-based PON data centre architecture. Our study showed that duplicating the special servers reduces power consumption and delay compared to the option where the servers in each rack share a star coupler / backplane. However, choosing between these two resilient designs is a compromise between cost and performance.

Table of Contents

List of Figures

List of Tables

List of Abbreviation

API	Application Programming Interface
ATM	Asynchronous Transfer Mode
AWG	Arrayed Waveguide Grating
AWGR	Arrayed Waveguide Grating Router
BPON	Broadband Passive Optical Network
BPPRD	Backplane per Rack Design
CO	Central Office
CPU	Central Processing Unit
DCeP	Data Centre energy Productivity
DCiE	Data Centre infrastructure Efficiency
DCN	Data Centre Network
DSL	Digital Subscriber Line
DVFS	Dynamic Voltage and Frequency Scaling
E/O	Electrical-to-Optical
EPON	Ethernet Passive Optical Network
EPS	Electronic Packet Switching
FBG	Fibre Bragg Grating
FSAN	Full-Service Access Network
FTTx	Fibre to the Premises
GbE	Gigabit Ethernet
GPON	Gigabit Passive Optical Network
ICT	Information and Communication Technology
IEEE	Institute of Electrical and Electronics Engineers
IT	Information Technology

ITU-T	International Telecommunication Union-Telecommunication
LAN	Local Area Network
MEMS	Micro Electro Mechanical Switches
MILP	Mixed Integer Linear Programming
MoCA	Multimedia Over Coax
MTTF	Mean Time to Failure
MTTR	Mean Time to Repair
NG-PON	Next Generation Passive Optical Network
NIC	Network Interface Card
ODN	Optical Distribution Network
O/E	Optical-to-Electrical
OFDM	Orthogonal Frequency Division Multiplexing
OSC	Optical Circuit Switches
OLT	Optical Line Terminal
ONT	Optical Network Terminal
ONU	Optical Network Unit
P2MP	Point-to-MultiPoint
PCB	Printed Circuit Board
PON	Passive Optical Network
PUE	Power Usage Efficiency
SDN	Software Defined Network
SLIC	Subscriber Line Interface
SSBD	Special Server Duplication Based Design
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access

ToR	Top of Rack
TWC	Tuneable Wavelength Converters
VDSL	Very high-speed Digital Subscriber Line
VLB	Valiant Load Balancing
VM	Virtual Machine
WDM	Wavelength Division Multiplexing
WSS	Wavelength Selective Switch

Chapter 1 Introduction

1.1 General overview

Information and communication technology (ICT) has considerably changed our everyday life. Its ability to transcend time and space has made it possible to access, store, transmit, share, and manipulate data anytime and anywhere [1]. This has promoted a sharp increase in the number of applications and services that depend on ICT, which in turn is expected to cause exponential growth in Internet traffic in the coming years [2]. According to [3], in 2018, the global Internet traffic was three times higher than its value in 2015 and is likely to continue to increase to 4.2 Zetta Bytes per year by 2022. In addition, the number of total Internet users in 2018 was 3.9 billion. This number is expected to become 5.3 billion by 2023 [4]. According to these trends, demands have rapidly increased for data centre which is a facility where computing and networking equipment is housed to collect, store, process, distribute, or allow access to large amounts of data. Companies and organizations rely heavily on the applications, services and data contained within a data centre, making it a crucial point for everyday operations.

To respond to this dramatic increase in the internet traffic, further effort has to be dedicated to improving network architectures and technologies. This involves adding network devices, which will result in higher power consumption. The global energy usage of Internet data transmission networks in 2018 was approximately 260 TWh, which was around 1.1% of the total global electricity demand [3]. Moreover, the data centres' global electricity demand was approximately 198 TWh in 2018, which represented 1% of the total global electricity demand. A slight reduction in the global data centre energy demands to 191 TWh is expected in 2021 in spite of the probable increase of 80% in the data centre traffic and the expected increase of 50% in the data centre workload during the coming three years. This is attributed to the current developments in the efficiency of data centre infrastructure and hardware [3].

Over the last few decades, there has been an increasing interest [5-35] in designing energy-efficient core and access communication networks where the access network represents the last mile of telecommunication networks and the core network provides services to customers who are connected by the access network.

Moreover, a significant number of energy-efficient solutions are concentrating on data centre architectures [36-43]. These efforts tend to enhance the legacy data centre network (DCN) architecture [44] by optimising the design and operations of the data centre network. This legacy data centre network is based on tiered topologies that use equipment (switches and routers) to manage intra-data centre communications and inter-data centre communications. Although this equipment provides high performance, they are power-hungry and use high-priced devices that increase the deployment and maintenance costs and contribute to global warming. Hence, to satisfy the data centres' demands in terms of scaling up, network designers have proposed new architectures with less cost and power consumption. Mainly, the research has progressed in three directions, namely the use of commodity switches, all-optical data centres, or hybrids [45].

With respect to the use of commodity switches where their cost and power consumption is much lower than enterprise-level switches, architectures are classified as switch-centric or server-centric [46]. The factors to consider while designing these architectures include bandwidth capacity, latency, complexity, scalability, resilience, and cost. The all-optical data centre architectures are considered one of the most active areas in data centre architecture research nowadays. This is because of the high data rates, low power consumption and flexibility of photonic systems [47]. In contrast, an electronic data centre has many limitations although it has been widely used. These limitations include low throughput, high cost, high latency, management complexity, and limited scalability [47]. Therefore, the optical approach is seen as a solution to overcome the electrical data centre limitations. In particular, the use of a passive optical network (PON) [48] for data centres has been one of the major interesting research subjects because of the low cost, low latency, and low power consumption as well as the high capacity and scalability of PONs [49].

On the other hand, reducing data centres power consumption can be also achieved by using different energy efficient approaches which reduce data centre resources power consumption, maximise data centre utilisation, and balance data centre thermal load [52]. Reducing data centres power consumption can be achieved by using energy efficiency techniques such as dynamic speed scaling and traffic management. Also, there are different approaches used to maximise data centre utilisation such as energy aware virtual machines embedding.

In this thesis we investigate the performance of a passive Arrayed Waveguide Grating Router (AWGR) and server-based PON data centre design. This work is the first to analyse the AWGR and server-based PON data centre architecture and, hence, is concerned with designing and optimising its routing and wavelength assignment also optimising the embedding of the virtual machine requests in the servers. In addition, this work studies the resilience ability of the AWGR and server-based PON data centre architecture in the face of different kinds of failures.

1.2 Thesis objectives

The main research objectives of the work presented in this thesis are as follows:

1. To study the traditional data centre architectures and review the main challenges facing traditional DCNs which motivates the introduction of new architecture designs;
2. To study the PON-based data centre architectures proposed in [49], focusing on the AWGR and server-based PON data centre architecture;
3. To optimise the routing and wavelength assignments within the AWGRs of a PON cell in the AWGR and server-based PON data centre architecture;
4. To optimise virtual machine embedding for cloud applications in the AWGR and server-based PON data centre architecture to minimise power consumption;
5. To evaluate the resilience of AWGR and server-based PON data centre architecture.

1.3 Thesis contributions

The main contributions of this thesis are summarised as follows:

1. A Mixed Integer Linear Programming (MILP) model is developed for the routing and wavelength assignments within the AWGR and server-based PON data centre architecture.
2. A MILP model along with a heuristic are developed to investigate virtual machine placement for cloud applications in the AWGR and server-based PON data centre architecture.
3. Improved designs of the AWGR and server-based PON data centre architecture are proposed for resilience against link failures. A MILP model is developed to optimise traffic routing over the proposed designs under failures scenarios.

1.4 Publications

The work in this thesis has resulted in the following publications:

1. Alani, R., Hammadi, A., El-Gorashi, T.E. and Elmirghani, J.M., 2017, July. PON data centre design with AWGR and server based routing. In 2017 19th International Conference on Transparent Optical Networks (ICTON) (pp. 1-4). IEEE.
2. Alani, R.A., El-Gorashi, T.E. and Elmirghani, J.M., 2019, July. Virtual Machines Embedding for Cloud PON AWGR and Server Based Data Centres. In *2019 21st International Conference on Transparent Optical Networks (ICTON)* (pp. 1-5). IEEE.
3. Thabit, R.A., El-Gorashi, T.E. and Elmirghani, J.M.,2020. Resilient AWGR and server based PON data centre architecture. In *2020 22nd International Conference on Transparent Optical Networks (ICTON).* IEEE.

4. Thabit, R.A., El-Gorashi, T.E. and Elmirghani, J.M., Resource allocation in AWGR and server based PON data centre architecture, to be submitted to IEEE Access.
5. Thabit, R.A., S. H. Mohamed, El-Gorashi, T.E. and Elmirghani, J.M., Resilient PON-based data centre architectures, to be submitted to IEEE Access.

1.5 Thesis structure

The reminder of the thesis is organised as follows:

Chapter 2 reviews the current data centre architectures, including electrical, optical, and hybrid architectures along with their main advantages and limitations.

Chapter 3 presents a review of the PON technology in access networks, including architectures, topologies, and components. Also, five PON based data centre architectures including an AWGR and server-based PON data centre architecture are described.

Chapter 4 discusses the optimisation of the routing and wavelength assignments of the AWGR and server-based PON data centre architecture. In addition, a benchmark study to compare the power consumption of the AWGR and server-based PON data centre architecture against a traditional data centre architecture, BCube, is presented.

Chapter 5 addresses the optimisation of virtual machine embedding in the servers of the AWGR and server-based PON data centre architecture. Moreover, a heuristic is developed to verify and validate the results of the MILP model.

Chapter 6 evaluates the resilience of the AWGR and server-based PON data centre architecture against link failure and introduces a MILP model to optimise traffic routing under link failures scenarios.

Finally, Chapter 7 presents the conclusions and some future work directions.

Chapter 2 Review of data centre architectures

2.1 Overview

In this chapter, a review of the current data centre architectures, including electronic, optical, and hybrid architectures, is presented along with their main advantages and limitations. Electronic data centre architectures are classified into switch-centric and server-centric architectures, while optical data centre architectures are classified as hybrid or fully optical.

2.1 Traditional data centre

The dramatic growth of Internet services, such as social networks and cloud computing, has led to the need for cost-effective data centre infrastructures with scalable computing and storage resources [50]. A data centre consists of a large number of servers that are grouped into racks where each rack typically hosts 20 to 40 servers [51]. The servers in a rack are connected by a top-of-rack (ToR) switch, which is responsible for routing inter-rack and intra-rack traffic. These ToR switches (also known as access switches) are interconnected by a network known as the data centre network (DCN), as shown in Figure 2.1 [52]. An effective and efficient DCN can help in reducing the cost of a data centre's deployment and maintenance. Moreover, the DCN needs to be scalable, reliable, and energy efficient in order to serve the various applications and services in a data centre [53].

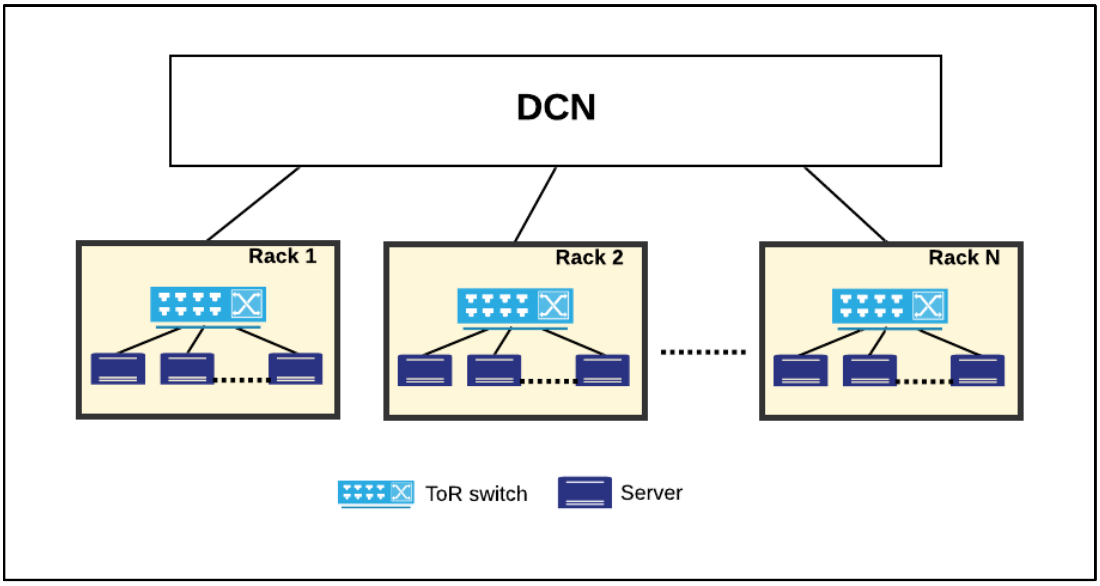


Figure 2.1: Typical data centre architecture [52]

2.2 DCN infrastructure

The DCN is one of the main factors that specify the characteristics of a data centre, such as resilience, latency and scalability. There are three main types of DCNs, namely electronic, optical, and wireless DCNs [45, 54]. Electronic DCNs use twisted pair cables, which provide transmission rates of 10 Mbps, 100 Mbps, 1 Gbps, and 10 Gbps. Optical DCNs are attracting increasing attention because of their high bandwidth and low power consumption. Optical fibres used in optical DCNs are light weight and small as compared to electrical cables and can support data rates up to 100 Gbps and beyond. Wireless DCNs mainly use the unlicensed 60-GHz frequency band, which offers a multi-gigabit data rate that makes it as a good alternative electronic DCN [44]. Our discussion in this Chapter is limited to electronic and optical data centres.

2.2 Electronic DCN

DCNs based on commodity Electronic switches are classified as either switch-centric or server-centric [46]. The factors to consider while designing these architectures are bandwidth capacity, latency, complexity, scalability, resilience, and cost.

2.2.1 Switch-centric architectures

These architectures depend on switches to provide connectivity and forward packets. The switches in these architectures are organised in a hierarchal topology with two-tier architecture for moderate data centres and three-tier architectures for large data centres [1]. The two-tier architecture is composed of edge-tier switches connected to core-tier switches, while the three-tier architecture adds aggregation-tier switches in the middle of the previous two, as shown in Figure 2.2 [45]. The following sections discuss some examples of switch-centric architectures.

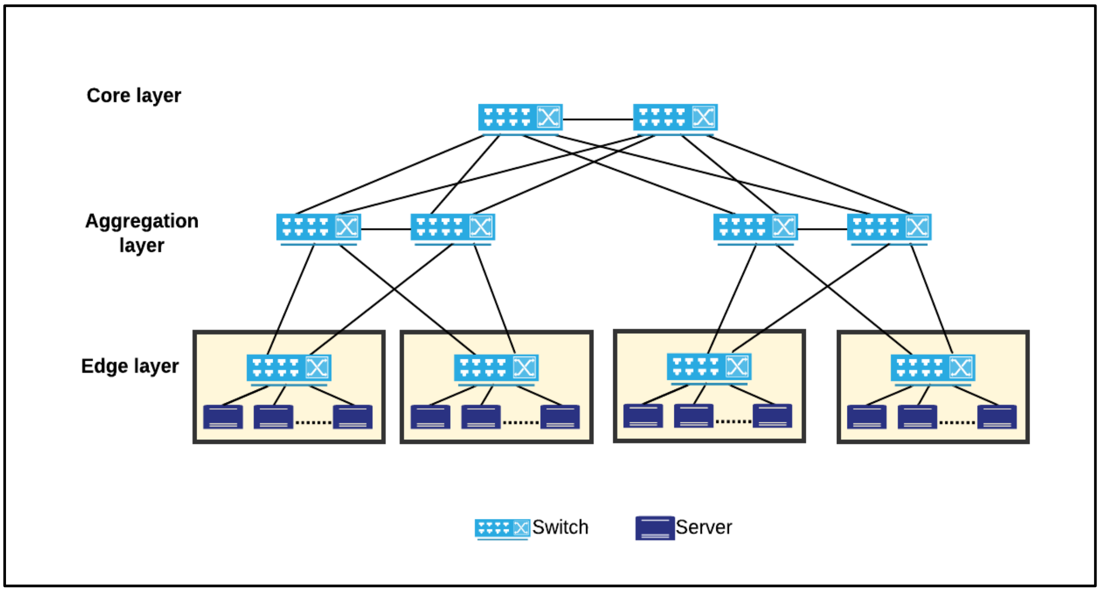


Figure 2.2: Three-tier data centre architecture [45]

2.2.1.1 Fat tree

This architecture is depicted in Figure 2.3 [54]. It replaces the high-end switches used in traditional data centres with identical commodity Ethernet switches, which reduces the cost and power consumption. Moreover, it increases the number of core and aggregation switches, which in turn decreases the oversubscription ratio. Furthermore, increasing the number of core and aggregation switches increases the number of available paths between servers. Scalability is another advantage of this architecture because of its ability to host $K^3/4$ servers, where K represents the number of switch ports [54].

However, in this architecture, the cabling complexity and cost are higher than in traditional data centres. Moreover, this architecture requires modification in the IP addressing where manual location-based addressing is required along with a modification for the routing table structure of switches.

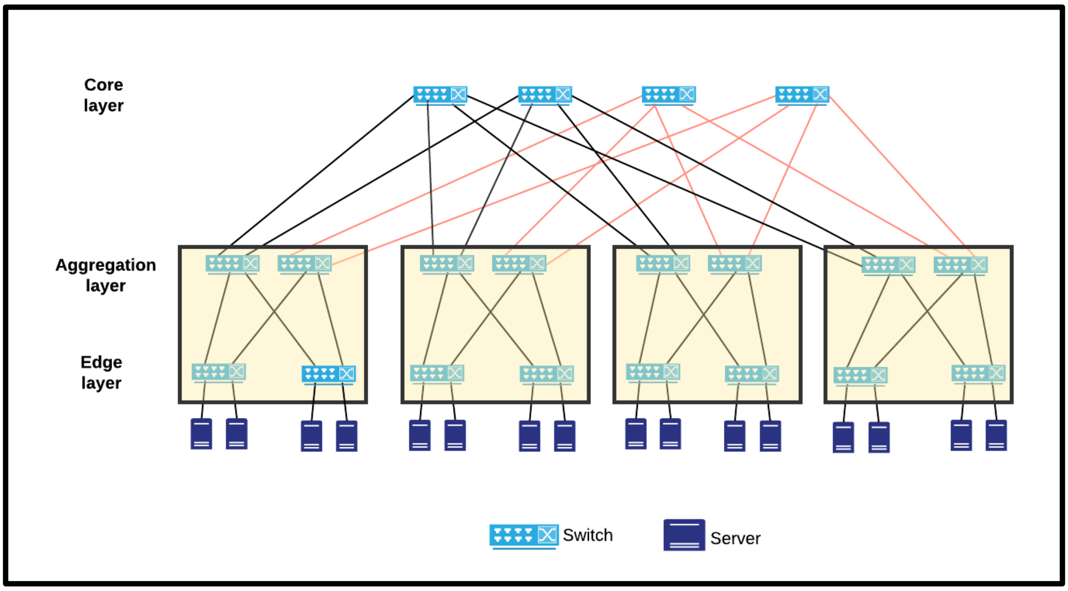


Figure 2.3: Fat tree topology [54]

2.2.1.2 VL2

The Virtual Layer 2 (VL2) architecture, shown in Figure 2.4 [55], is similar to a fat tree in terms of the use of commodity switches but differs by assigning higher capacity to the switch-to-switch links than the server-to-switch links. Another difference is the reduction in the number of cables needed to connect the core layer to the aggregation layer. Moreover, flat addressing is used in this architecture to allow the allocation of services anywhere in the DCN without the need for manual addressing configuration. The VL2 architecture performs traffic flow load balancing for a high cross-section bandwidth. However, in this architecture, the directory services may become the bottleneck in case of heavy traffic load along with virtual relay and centralised management [55].

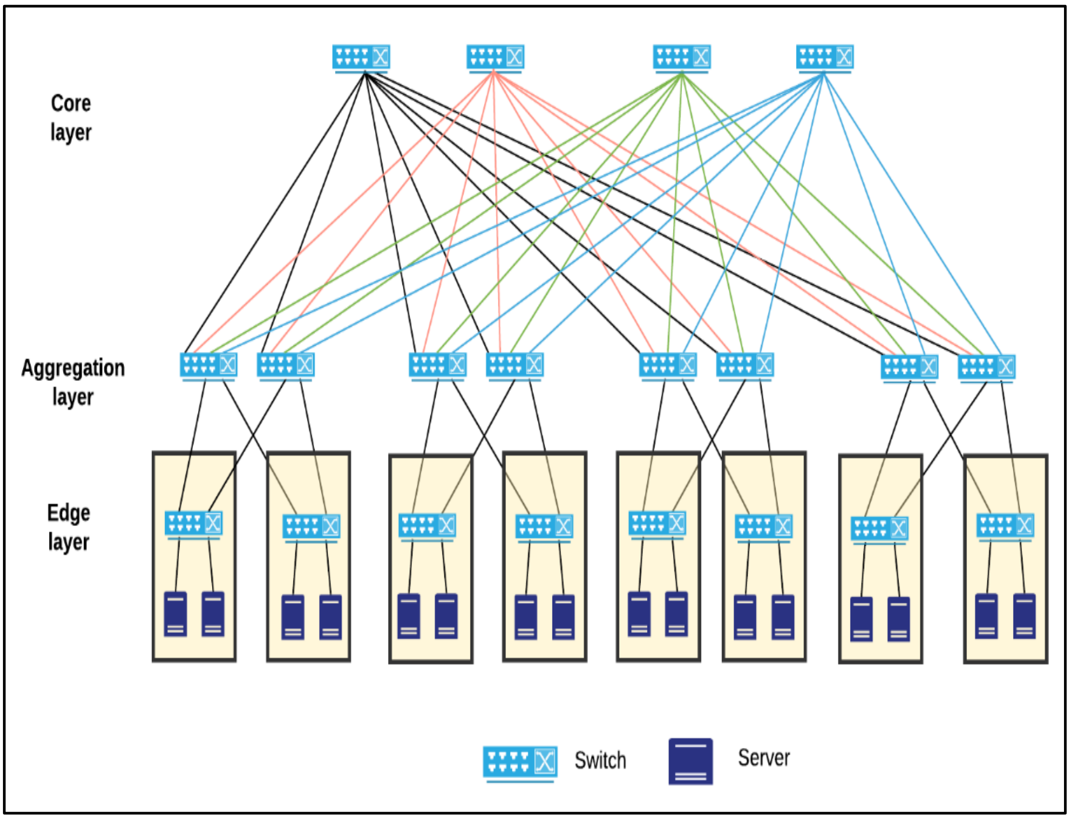


Figure 2.4: VL2 topology [55]

2.2.1.3 Jellyfish

This architecture connectivity is based on a random graph topology which allows incremental and heterogeneous growth [56]. The number of servers connected to ToR switches is random where these switches are directly connected in a random manner, as shown in Figure 2.5 [57]. Jellyfish is a scalable architecture that provides high bandwidth, random interconnectivity, and resilience. Compared with a fat tree, it offers shorter path lengths and less cost for the same number of servers. In contrast, Jellyfish requires more work on the routing schemes since no routing approach is provided by the developers[56].

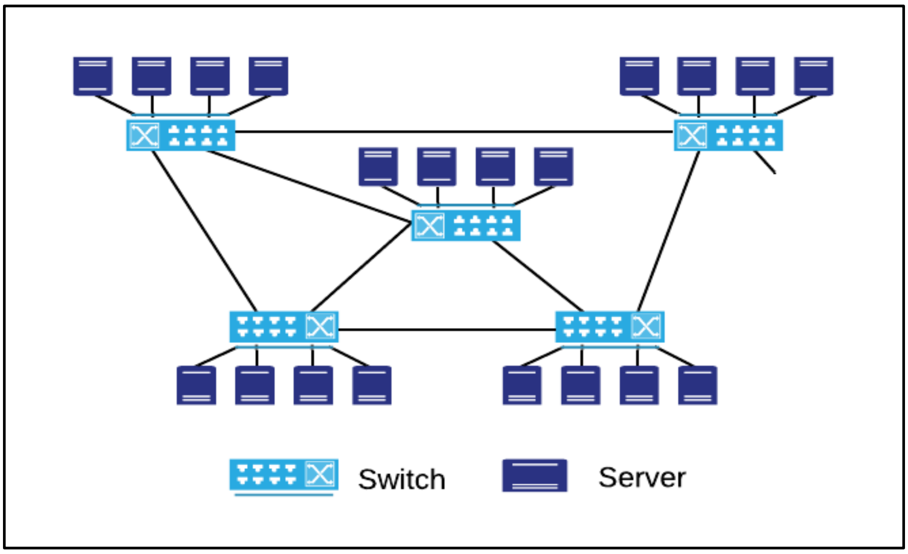


Figure 2.5: Jellyfish topology [57]

2.2.2 Server-centric architectures

Here, servers are assigned a role in data forwarding benefiting from the server hardware advancement in central processing unit (CPU) and network interface cards (NIC) ports [44]. Some of the server-centric architectures are described in the following sections.

2.2.2.1 GRIN

GRIN, as shown in Figure 2.6 [58], is the most basic solution that uses servers as relay nodes for traffic forwarding. It can simply be used over any topology by connecting the servers in the same or neighbouring racks to each other using their free ports. This will enable servers to benefit from the free bandwidth of the neighbouring servers connected to them. In this case, two paths are available for the server traffic: one provided by the topology itself, and the other through the servers connected to it. This is a cheap and easily deployed solution [58].

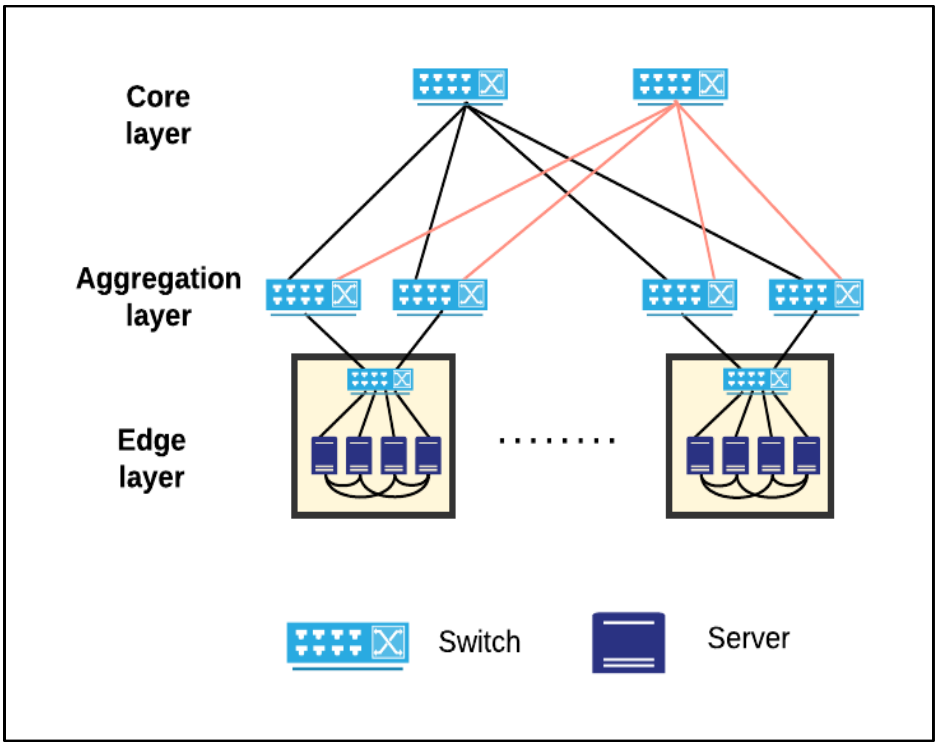


Figure 2.6: GRIN topology [58]

2.2.2.2 DCell

This is a hierarchical architecture that is based on a basic unit, referred to as DCell, built by connecting n servers to a switch [59]. A higher level of the network can be obtained by adding $n + 1$ DCell units. In this architecture, each server in a DCell unit is connected to a server in a different DCell unit, resulting in connecting the DCell unit to all the other DCell units, as depicted in Figure 2.7 [59]. Although it is a scalable, resilient, and cost-effective architecture, it increases the oversubscription ratio and the path length and decreases the cross-sectional capacity.

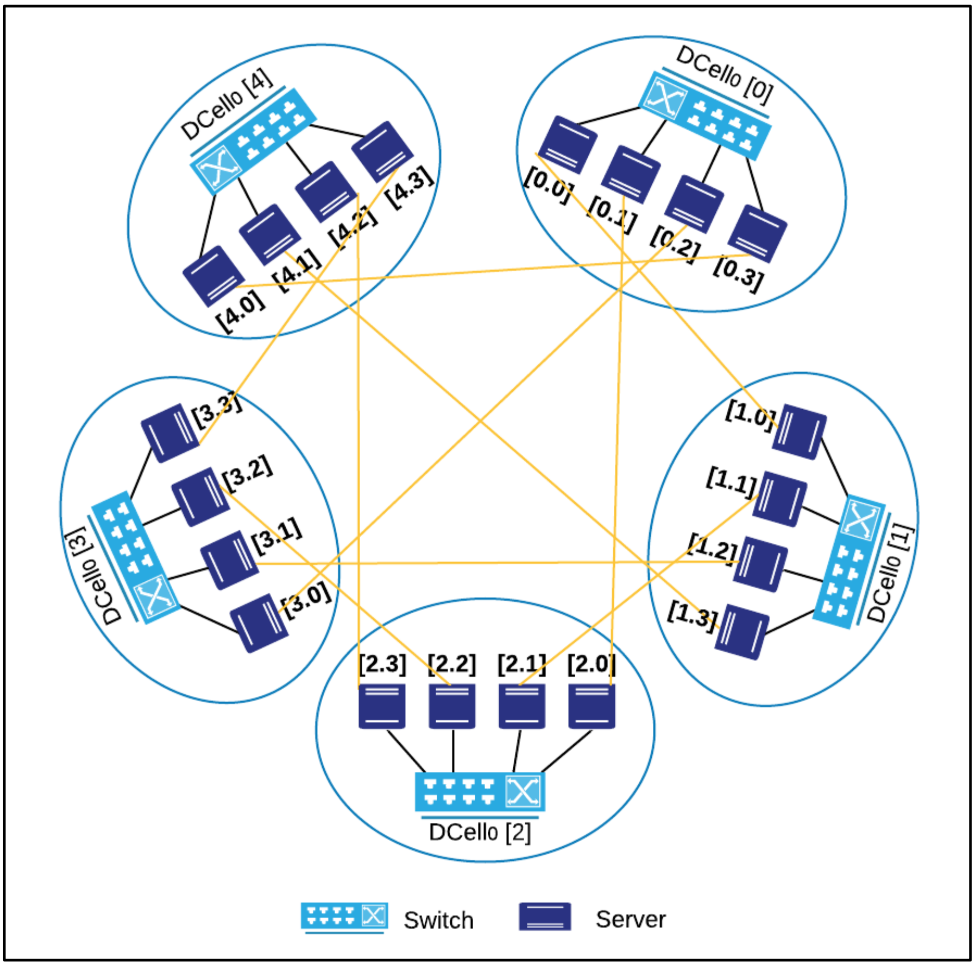


Figure 2.7: DCell topology [59]

2.2.2.3 BCube

Similar to the concept of the DCell topology, BCube repeats a basic BCube unit to scale up the network [60]. Here, the servers need to have multiple network ports to connect to multiple commodity switches. Intercell traffic is handled by switches instead of servers, and this is different from the DCell topology, as shown in Figure 2.8 [60]. The advantages of the BCube topology include high performance, resilience, and efficient support to one-to-one, one-to-several, and one-to-all communication. On the downside, BCube lacks scalability, as it is designed for container data centres and its cabling cost is high.

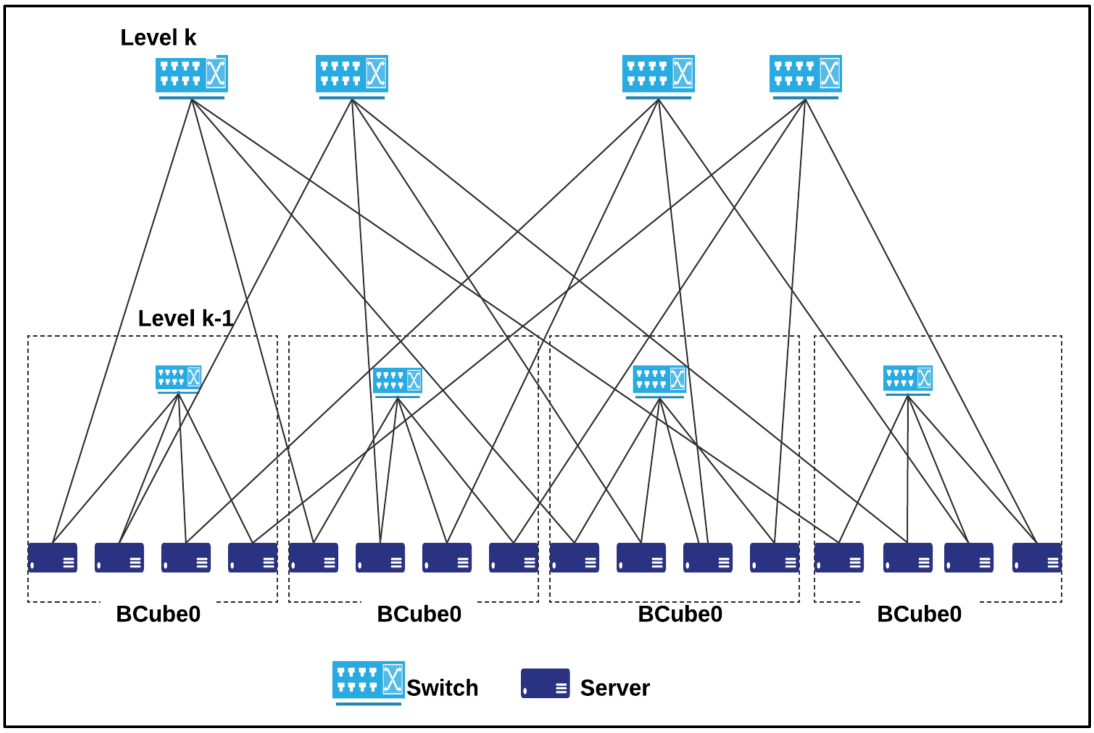


Figure 2.8: BCube topology [60]

2.3 Optical DCN

Optical data centres are one of the most active areas in data centre architecture research today. This is because of its photonic systems that provide high data rates, low power consumption, and high flexibility [47]. In contrast, an electronic data centre architecture has many limitations, although it has been widely used in data centres. These limitations include throughput, high cost, latency, management complexity, and scalability. Therefore, optical approaches have been proposed as a solution to overcome the limitations of an electronic data centre [61]. In the following subsections we discuss the two main categories of optical data centres; hybrid data centres and all optical data centres.

2.3.1 Hybrid DCN

Hybrid data centre architectures made up of electronic and optical components were developed first by researchers who wished to develop all optical data centre architectures, but whose efforts were limited by the lack of optical memories that can be integrated in all optical switches in optical data centre architectures. They tended to combine electronic switches and cables with optical circuit switches (OCSs) and optical fibres [62]. This resulted in a significant reduction in power consumption and latency since OCSs do not need to perform packet processing. Moreover, this approach reduced cable complexity and enabled considerably higher data rates, as optical fibres can provide communication rates of up to hundreds of gigabits per second. The following subsection discusses some examples of hybrid architectures.

2.3.1.1 HyPaC

This is a hybrid packet and circuit-switched data centre architecture that is based on a tree topology [63]. It was proposed to enhance the bandwidth for data centre applications by providing alternative optical connections between data centre racks in addition to the electronic packet switching between them, as shown in Figure 2.9 [63]. This is accomplished by connecting ToR electronic switches via a reconfigurable circuit-switched optical network. A rack can only be connected to one other rack at a time and uses the temporary high-capacity optical links.

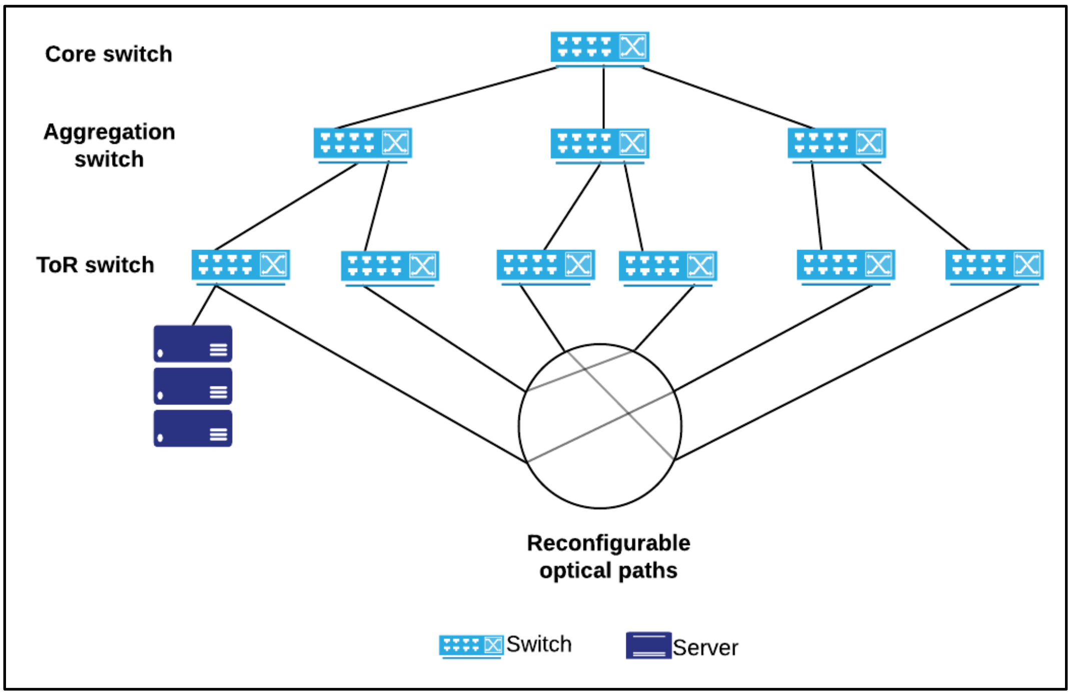


Figure 2.9: HyPaC topology [63]

2.3.1.2 Helios

Helios is a two-level multi-rooted tree architecture made up of electronic and optical switches [64]. The first level consists of pods with thousands of servers where each pod is connected to an electronic switch via short copper cables, as shown in Figure 2.10 [64]. In contrast, the second level, which is the core level, uses a mixture of electronic and optical switches. All the links between the pod switches and the core switches are optical, which requires the switches to have optical transceivers.

To achieve a higher bandwidth, this architecture utilises two optical technologies, namely Micro Electro Mechanical Switches MEMS-based optical switches and WDM transceivers. Moreover, for enhanced performance, Helios selects the traffic path according to its size. Accordingly, large flows are sent to the optical core switches and small flows through the electronic core switches. In addition, this architecture reduces cost, complexity, and power consumption as compared to a non-blocking electronic switch topology.

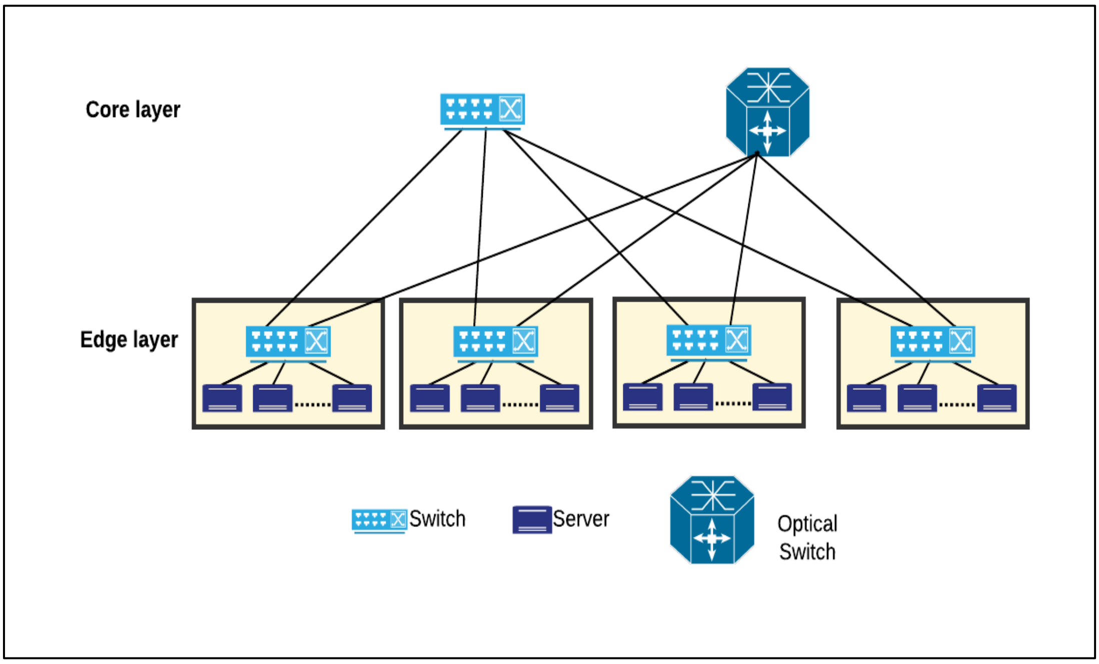


Figure 2.10: Helios topology [64]

2.3.2 Fully optical DCN

Although hybrid electronic/optical architectures have brought many advances to the data centre, they still consume considerable power because of the use of the power-hungry electronic-to-optical (E/O) and optical-to-electronic (O/E) transceivers. Hence, researchers have begun to further study the feasibility of applying fully optical DCNs to reduce power consumption and latency and increase bandwidth [45].

Typically, an all-optical DCN is obtained by directly connecting racks to an optical interconnection network. The optical interconnection network is a term which refers to any network topology that uses only optical switching. In these topologies, the top of the rack switches can be either optical switches or electronic switches [47].

Compared with the previously mentioned DCN infrastructures, optical interconnection networks reduce power consumption and latency and increase bandwidth. Nevertheless, the replacement of all the commodity switches is considerably costly; hence, to reduce the cost, some optical interconnects use either OCSs or packet-based optical switches.

2.3.2.1 OSA

The optical switching architecture (OSA) is an optical switching architecture for ToR switches that can modify its topology and link capacity dynamically and thus, introduces an exceptional flexibility to adjust to dynamically changing traffic patterns [65]. This is accomplished by benefiting from the reconfigurability of the optical devices used, such as MEMSs, wavelength selective switches (WSSs), optical circulators, and optical transceivers.

In this architecture, each electronic ToR switch communicates with k ports of MEMSs, where *k* specifies the number of other ToR switches it is directly connected to simultaneously, as shown in Figure 2.11 [65]. Accordingly, MEMS choose which sets of ToR switches are connected together on the basis of their configurations. Hence, to connect a set of ToR switches to the remaining ToR switches that are not directly connected to them, multi-hop routing is used. At each hop, the packet is converted from optical to electrical to read the packet's header to determine its destination and then convert it back to optical and send it.

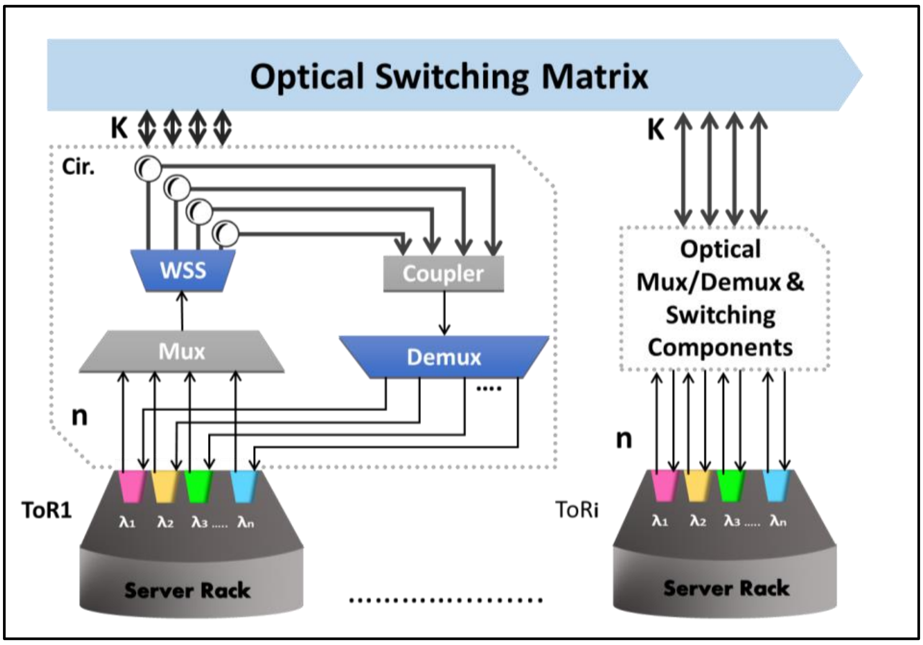


Figure 2.11: OSA topology [65]

In contrast, link capacity flexibility is achieved through the use of the WDM technology and wavelength selective switch (WSS). The WDM technology enables the transmission of multiple wavelengths over a single fibre simultaneously, while WSS provides dynamic reconfigurability. The MEMS and WSS configurations are controlled by an OSA manager.

Note that the value of k affects the network size inversely and the performance proportionally. Moreover, the slow switching speed of MEMS (9 ms) and WSS switches (12 ms) might affect the time-sensitive mice flows.

2.3.2.2 DOS

This is a datacentre optical switch (DOS) that uses an array waveguide grating router (AWGR), tuneable wavelength converters (TWC), and a loopback-shared SDRAM buffer as its basic elements [66]. Moreover, a control plane is implemented to guarantee that each packet passes through the appropriate AWGR output port. This is done by setting the TWC wavelength based on the destination address of the packet that is obtained using O/E/O converters. Although the delay in DOS is not dependent on the input load, it uses the power hungry O/E and E/O converters that increase the packet delay and the power consumption.

2.4 Energy efficiency in data centres

The sharp increase in the demand for applications and services, that depend on data centres, resulted in the exponential growth in data centres number and size. Accordingly, the data centre power consumption, cost and environmental impact are also increasing [53]. The main components of these large scaled data centres are servers, DCN and cooling system equipment which need to be addressed when considering energy efficiency. It was reported in [67] that 26% of data centres power consumption is caused by servers and storage equipment while cooling system consumes 50%.

In order to achieve efficient data centre power consumption, researchers adopted different approaches. These approaches attempt to reduce data centre resources power consumption, maximise data centre utilisation, balance data centre thermal load, and develop performance metrics [53].

Reducing data centres power consumption can be achieved by using energy efficiency techniques such as dynamic speed scaling and traffic management. Dynamic speed scaling (also known as dynamic voltage and frequency scaling DVFS) is implemented by reducing the devices speed in order to save energy since power consumption is proportional to the device speed or the supply voltage as deployed in [68, 69]. As for traffic management [45, 50], it is achieved by determining the devices and links needed to satisfy the traffic demands and then by trying to reroute the traffic through these devices and links and switch off the underutilised equipment as presented in [70]. Also, there are different approaches used to maximise data centre utilisation such as energy aware virtual machines embedding. Many studies [29, 71, 72] deployed VM embedding, which attempts to place VMs efficiently in servers in a way that uses as few servers as possible while satisfying the demands. In addition, the use of renewable energy sources such as geothermal and solar energy is a new trend in data centres in order to replace the traditional fossil fuel-based energy sources [53]. In addition, recent research investigates replacing the commonly used air conditioning units by other types of cooling systems such as heat exchanger pipes under data centre racks to maintain room temperature as in [73].

In order to evaluate the energy efficiency of a data centre, different metrics are used such as Power Usage Effectiveness (PUE), Data Centre infrastructure Efficiency (DCiE), and Data Centre energy Productivity (DCeP). PUE [74, 75] is the ratio of the total facility power consumption to the IT equipment power consumption. The lower the PUE, the more efficient is the data centre, which is achieved by enhancing the efficiency of the cooling and power distribution systems. DCiE [74] is the ratio of the IT equipment power usage to the total facility power. DCeP [53, 75] is used to identify the data centre computing efficiency and it is the ratio of useful work produced to total energy consumed to perform that work. Numerous studies have attempted to reduce the data centre's power consumption as in [76-79].

2.5 Summary

This chapter provided a detailed review of the recent data centre architectures, including electronic, optical, and hybrid architectures. The electronic data centre architectures were classified into switch-centric and server-centric architectures. We reviewed switch-centric data centre architectures including fat tree, VL2, and Jellyfish and server-centric data centre architectures including GRIN, DCell, and BCube. The optical data centre architectures were classified as hybrid data centres and all optical data centre architectures. HyPac and Helios were studied as examples of hybrid data centre architectures whereas OSA and DOS represented all optical data centre architectures. In addition, approaches adopted for energy efficiency in data centres which attempt to reduce resources power consumption, maximise utilisation, balance thermal load, and develop performance metrics were illustrated.

Chapter 3 PON in data centre

3.1 Overview

This chapter reviews PONs in access networks, including architecture, topologies, technologies, and components. The goal is to provide sufficient background to motivate the use of PONs in optical data centre architectures where the traffic patterns and connectivity are different from those of the access networks. Detailed descriptions of five fully passive optical data centre architectures are presented explaining how inter-rack connectivity and intra-rack connectivity are provided.

3.2 Review of PON in access network

The performance of Passive Optical Network (PON) technologies in terms of cost, energy consumption, and bandwidth is proven in access networks. The following subsections review PONs in access networks, including evolution, architecture, topologies, technologies, and components.

3.2.1 PON evolution

The access network represents the last mile of telecommunication networks. It connects the subscribers to a particular service provider. The bandwidth limitations in the deployed access solutions, such as data cable and Digital Subscriber Line (DSL) create bottlenecks. To overcome these bottlenecks, the largest carriers around the world started to invest in Fibre To The Premises (FTTP) solutions. One of the low-cost and high-bandwidth solutions is PON. In contrast to other wired solutions, such as Very High-Speed Digital Subscriber Line (VDSL) and active optical networks, PON utilises passive components, which reduces the network's deployment and maintenance costs [80].

The literature shows that the first attempt to develop optical access networks was carried out by the major international carriers in the 1980s in small experimental scales. However, far too little attention has been paid to optical solutions at that time because of the high cost associated with deploying them by digging the streets and due to the low traffic demand. In the following two decades, the rapid development in the Internet urged the call for effective broadband solutions. This led, in 1995, to the initiation of the Full-Service Access Network (FSAN) consortium, which was responsible for defining the optical access system's requirements. Later on, FSAN/ International Telecommunication Union-Telecommunication (ITU-T) and Institute of Electrical and Electronics Engineers (IEEE) developed a series of standards to increase upstream and downstream bandwidths, as depicted in Figure 3.1 [80, 81].

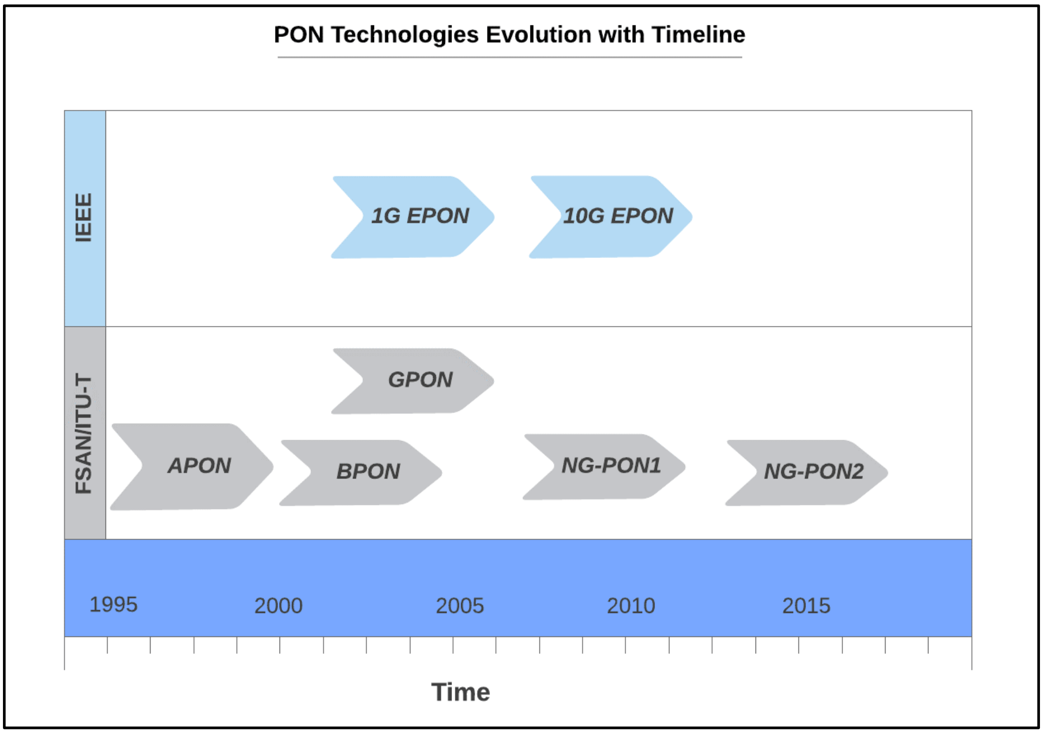


Figure 3.1: PON evolution in the access network [81]

3.2.2 PONs architecture

Typically, a PON is composed of a single Optical Line Terminal (OLT), typically deployed in the service provider's Central Office (CO), a number of Optical Network Units (ONUs) located near subscribers, and the Optical Distribution Network (ODN). The OLT connects the PON to the backbone network, while the ONUs provide the network interface to the subscribers. In order to connect the OLT to the ONUs, the ODN uses optical fibre and passive optical splitters, as shown in Figure 3.2 [82].

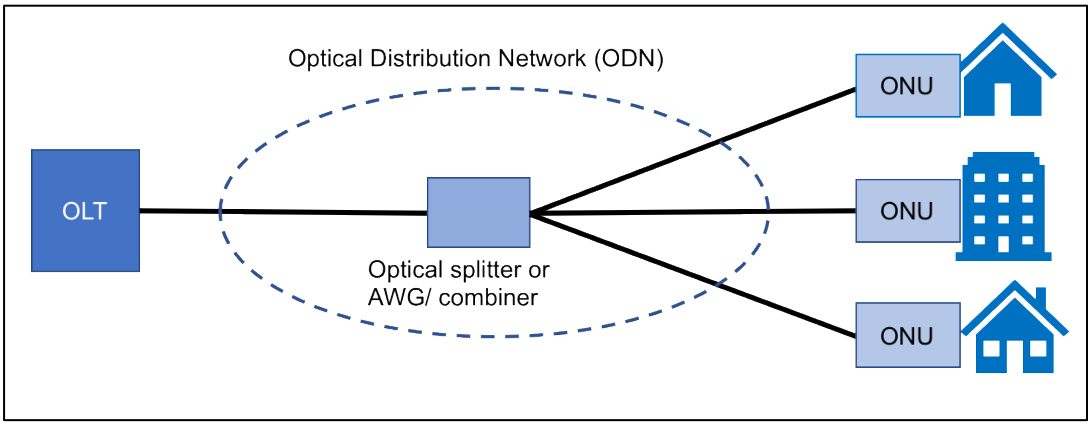


Figure 3.2: General PON architecture in the access networks [82]

A PON performs point-to-multipoint (P2MP) communication, where the downstream signals from the OLT are split by a 1:N splitter into n single fibres each connected to an ONU, whereas the upstream signals from ONUs are coupled onto one upstream fibre to the OLT.

3.2.3 PON topologies

There are three main types of topologies for PONs to choose from, namely tree, bus, and ring. The tree topology, shown in Figure 3.3(a), is preferred when the ONUs are close to the OLT and to each other, such as in urban areas. In contrast, the bus topology, shown in Figure 3.3(b), is used when ONUs are located far from the OLT and each other such as in rural areas. For failure protection, a ring topology is chosen where two fibres are used to back each other up, as shown in Figure 3.3(c) [80].

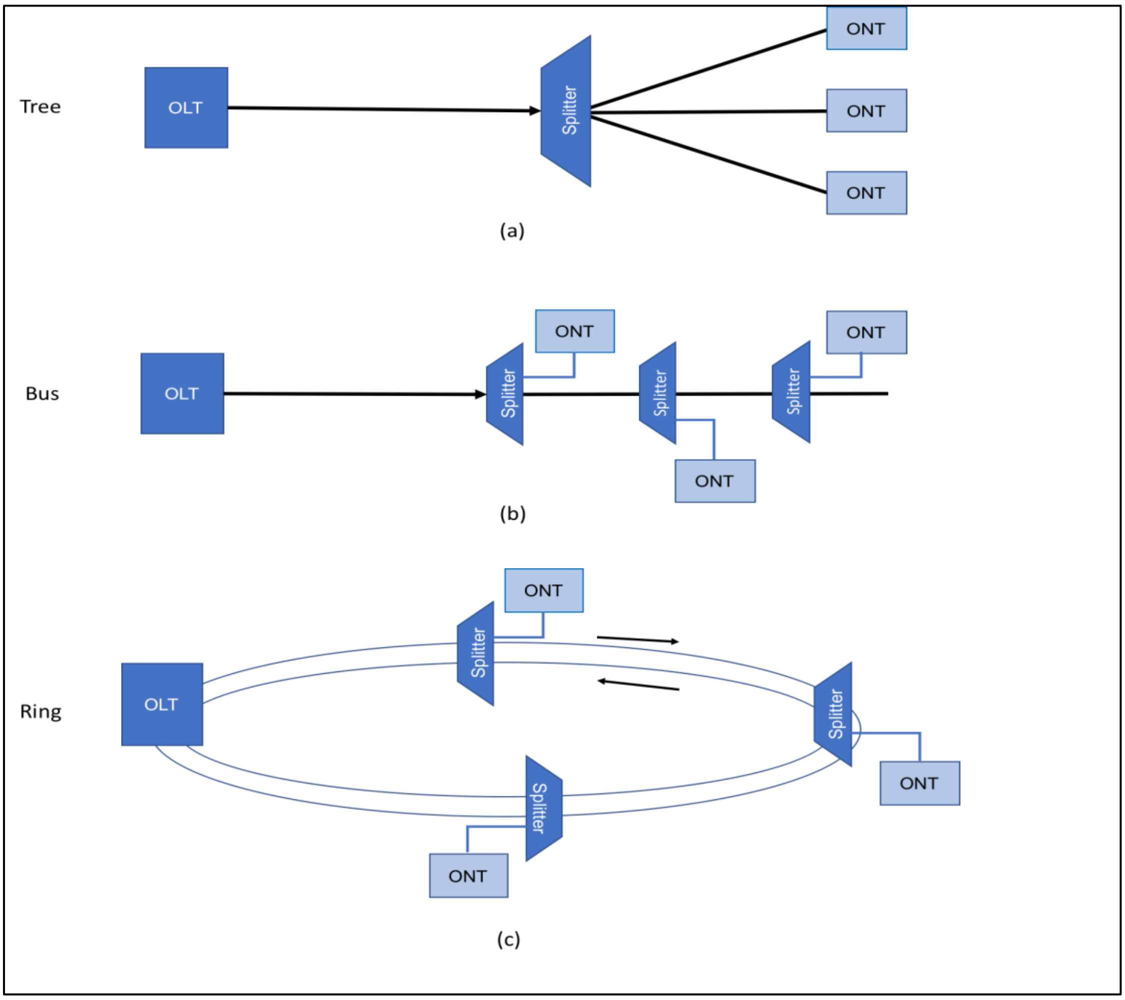


Figure 3.3: Basic PON topologies [80]

3.2.4 Technologies

There are several technologies that are used in PONs as briefly described below [80, 82].

3.2.4.1 APON/BPON

The first standardised technology deployed in PONs is APON or ATM-PON by FSAN, which uses the asynchronous transfer mode (ATM) switching technique. BPON is an enhancement of APON with higher performance. The APON/BPON transmission speed for the downstream is 155 Mbps or 622 Mbps, while that of the upstream is 155 Mbps.

3.2.4.2 EPON/10G-EPON

EPON was initiated by IEEE in order for the Ethernet to be adopted in access networks in addition to LAN networks. The downstream/upstream transmission speeds are symmetric at a rate of 1 Gbps over a transmission distance of 10–20 km. EPON is based on the P2MP tree topology that supports 16 ONUs per OLT. With a view to achieve a higher bandwidth, IEEE developed 10G-EPON with 10 Gbps downstream and a choice of 10 Gbps or 1 Gbps for the upstream data rate.

3.2.4.3 GPON

The reason behind introducing GPON, by FSAN/ITU-T, was to increase the network's capacity, reach, split ratio, and flexibility. It provides various transmission speeds: symmetric downstream and upstream of 622 Mbps, symmetric downstream and upstream of 1.244 Gbps, and downstream of 2.488 Gbps with upstream of 1.244 Gbps. It provides transmission over 60 km with 128 ONUs per OLT and supports ATM, Ethernet, voice, TDM, and wireless extensions.

3.2.4.4 NG-PON

Aiming for a higher bandwidth, FSAN/ITU-T started investigating the next generation of PON (NG-PON). This involves two categories—NG-PON1 and NG-PON2—where NG-PON1 is compatible with GPON standards and considered as a mid-term upgrade. In contrast, NG-PON2 is a long-term solution, which is independent of GPON standards and can be used over new ODNs.

3.2.5 Components

The optical network components are either active or passive devices. Active components require electronic control to operate whereas passive components does not [83]. Since active components are not used in a PON network, this section only describes the passive components. The main passive components used in PONs are described below.

3.2.5.1 Splitter

It is a 1 × N passive device that receives optical signals coming from a single input fibre and splits it into N output fibres. The number of output fibres affects the amount of loss in the splitter. A splitter with a desired splitting ratio of 1 : N can be obtained directly by using a 1 × N splitter [84].

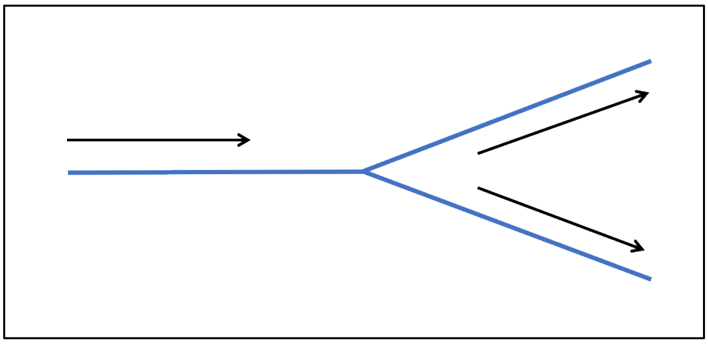


Figure 3.4: Basic operation of a splitter [84]

3.2.5.2 Coupler

It is an N × N passive device which combines optical signals from N incoming fibres and sends them through N outgoing fibres, as shown in Figure 3.5. There are two types of couplers: wavelength-independent and wavelength-selective. The former’s coupling ratio is independent of the wavelength, while in the latter, it is. In PONs, usually, the splitter and the coupler are considered one device that splits downstream signals and combines the upstream signals with a similar loss for both directions [84].

3.2.5.3 Star coupler

It is a multi-way coupler in which every input signal is received by every output fibre, as shown in Figure 3.5(a) [85]. Another type of a star coupler is a reflection star coupler in which the input signal can be on any fibre, whereas the output signal is split among all the fibres, as shown in Figure 3.5(b).

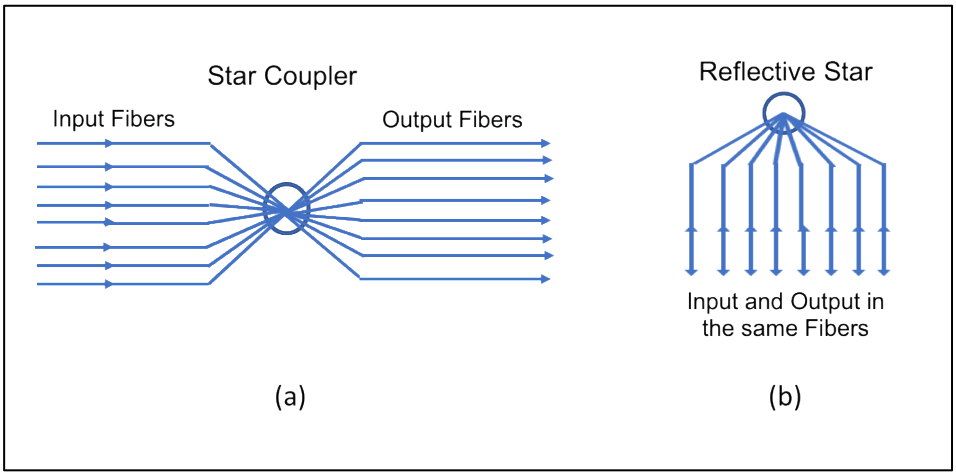


Figure 3.5: Basic operation of star coupler (a) Generic star coupler (b) Reflective star [85]

3.2.5.4 AWG

It is a 1 × N passive device that multiplexes signals from N inputs into one output fibre and demultiplexes signals from a single fibre into N outputs, as shown in Figure 3.6 [80]. As shown in the figure, when AWG operates as a demultiplexer, the signals entering from the single fibre pass through different stages. First, signals traverse a splitter and then enter a number of waveguides. Then, the signals outgoing the waveguides travel across a combiner and leave the AWG through the output waveguides. Accordingly, each output port delivers only one wavelength. In contrast, the reverse steps are applied when the AWG operates as a multiplexer.

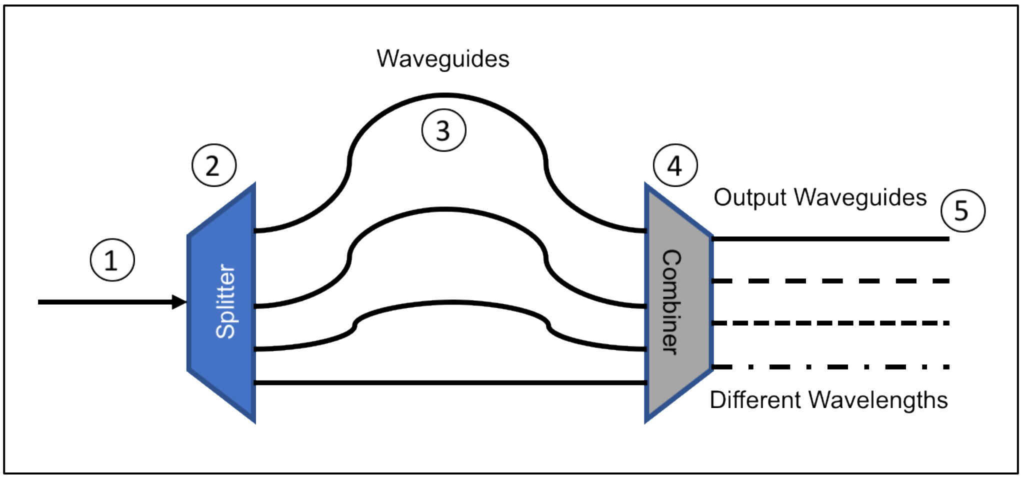


Figure 3.6: Basic operation of an AWG [80]

3.2.5.5 FBG

A fibre Bragg grating, as shown in Figure 3.7 [86], is a passive device that operates as a reflector for a specific wavelength (λ_b), while other wavelengths passing through FBG are not affected.

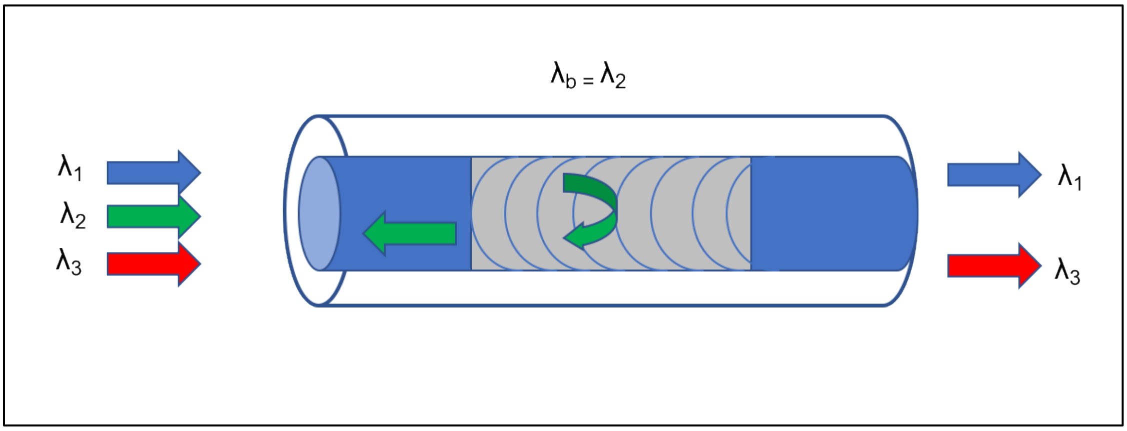


Figure 3.7: Basic operation of FBG [86]

3.2.5.6 Passive polymer optical backplane

It is a meshed polymer waveguide architecture on Printed Circuit Board (PCB) as shown in Figure 3.8 [87]. Its links represent a mesh of non-blocking connections that connect line cards equipped with optical transmitter and receiver arrays. Each card can communicate with any other card at the transmission rate of 10 Gbps or even loop back with a low link loss and remarkable crosstalk performance.

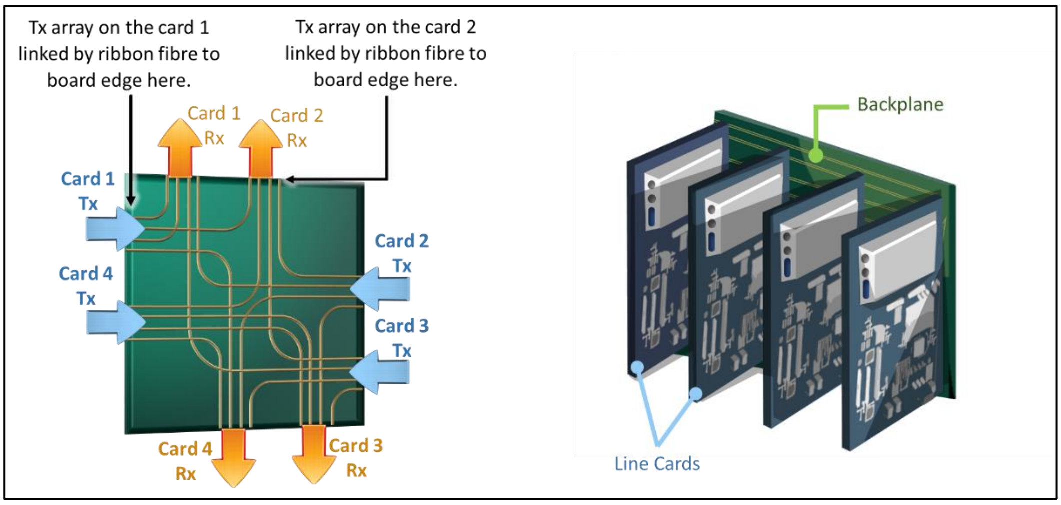

Tx array on the card 1 linked by ribbon fibre to board edge here.
Tx array on the card 2 linked by ribbon fibre to board edge here.
Card 1 Rx
Card 2 Rx
Card 1 Tx
Card 4 Tx
Card 2 Tx
Card 3 Tx
Card 4 Rx
Card 3 Rx
Backplane
Line Cards

Figure 3.8: Passive polymer backplane: (a) planar polymer waveguide routing and (b) backplane architecture [87]

3.3 PON in data centre networks

The deployment of PON technologies in a data centre architecture needs to take into account the differences in terms of the traffic patterns and connectivity between the data centres and access networks, where PON technology was first deployed. In access networks, traffic is mainly a north–south flow between subscribers and the Optical Line Terminal (OLT) placed in the telecom office. Accordingly, no direct connection exists between subscribers; instead, they communicate through the OLT if needed. In contrast, in a data centre, the east–west traffic between servers is as important as the north–south traffic. Burdening the OLT with the intra-rack traffic and the inter-rack traffic will result in increased power consumption and increased delay. Therefore, it is necessary to redesign PON connections so that multiple routes are available between servers in a data centre [49].

Studies related to the design of fully passive interconnections for a data centre have been relatively scanty as compared to hybrid data centre architectures that use passive and active components. A hybrid approach is proposed in [88], where the inter-rack communication is served by WDM PON, while Ethernet switches are still used for intra-rack connectivity. Another solution, in [89], focuses only on the ToR switch which is passive optical irrespective of the aggregation and core switches, which could be optical or electronic packet switches (EPS) switches.

Recent studies proposed data centre architectures that are based on fully passive intra-rack and inter-rack interconnections [49, 90-94], which are described fully in the upcoming sections.

3.3.1 PON-based data centre architectures

In [49, 90-94], five novel fully passive optical architectures were proposed, which eliminate the need for aggregation and core switches. Consider using an OLT switch (adapted from the access network) of eight chasses, as depicted in Figure 3.9 [49], used to connect servers in a data centre. Each chassis hosts up to 16 cards, and each card has eight ports each supports a transmission rate of 10Gb/s and 128 subscribers [49]. Accordingly, each card can provide connectivity to 1024 servers, while each chassis can connect 16,384 servers.

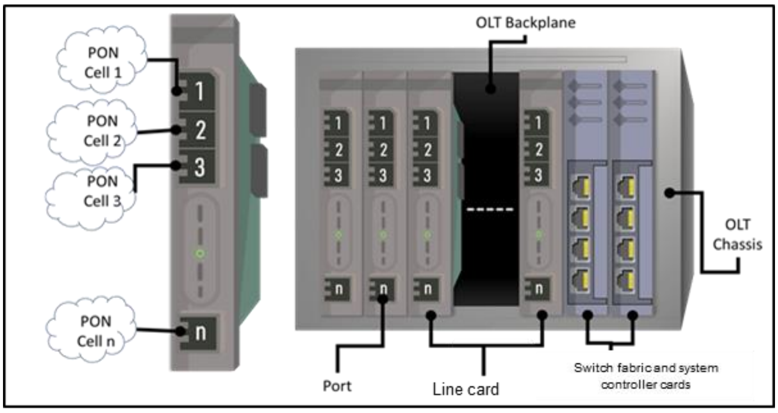


Figure 3.9: OLT chassis architecture [49]

The intra-rack and inter-rack connectivity in a PON data centre is redesigned so that other routes in addition to the route through the OLT are available. The intra-rack connectivity of the five architectures in [90-94] is fully passive and has three possible options: star coupler, fibre Bragg grating (FBG), or passive polymer backplane. The inter-rack connectivity is what distinguishes each design. Therefore, in the following subsections we review the three intra-rack connectivity technologies that can be used in the five PON designs and the inter-rack connectivity for each design.

3.3.2 Intra-rack connectivity options

3.3.2.1 Reflection star coupler connectivity

In this connectivity, servers in a rack are connected to each other using a reflection star coupler, as shown in Figure 3.10. It is a broadcast-based connection in which the power input into any port is divided equally to all the ports. Accordingly, the received power at each port is inversely proportional to the number of servers that are connected to the reflection star coupler. Every fibre connected to the reflection star coupler carries both the transmitted and the received data, which creates the need for a directional coupler or circulator to separate the two signals at each terminal. With the use of a reflection star coupler, the intra-rack traffic is separated from the inter-rack traffic and uses separate transceivers (although this is an architecture choice, but it is simpler to use different transceivers to couple to forward (inter-rack) or reverse (intra-rack) directions. This requires extra wiring and multi-wavelength transceivers, and any fault in the links leading to the reflection star coupler or in the reflection star coupler itself affects only the intra-rack communication [49].

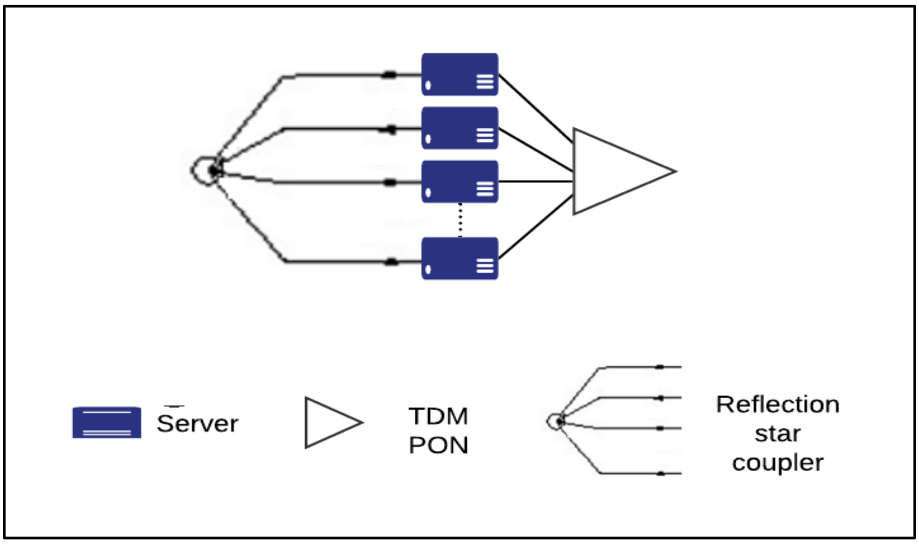


Figure 3.10: Reflection star coupler for intra-rack communication [49]

3.3.2.2 FBG connectivity

In this connectivity, the FBG is located after the star coupler that connects the servers in the rack to the rest of the network, as illustrated in Figure 3.11. Accordingly, the intra-rack and inter-rack traffic use a shared link in which the intra-rack traffic is reflected by the FBG back to the rack, while the inter-rack traffic continues out of the rack. This is accomplished by using one wavelength for intra-rack communication and another for communicating with the rest of the network. Therefore, an additional transceiver is required in each server, or a single high-cost OFDM transceiver can be used to produce the two wavelengths by generating multiple carriers; However, this increases the deployment cost [28, 32].The use of a shared link for the intra-rack and the inter-rack traffic reduces the wiring needed, although any fault in links or FBG will affect both traffics [49].

Note that the reflected signal is divided equally among all the ports; therefore, the received power at each port is inversely proportional to the number of servers in the rack.

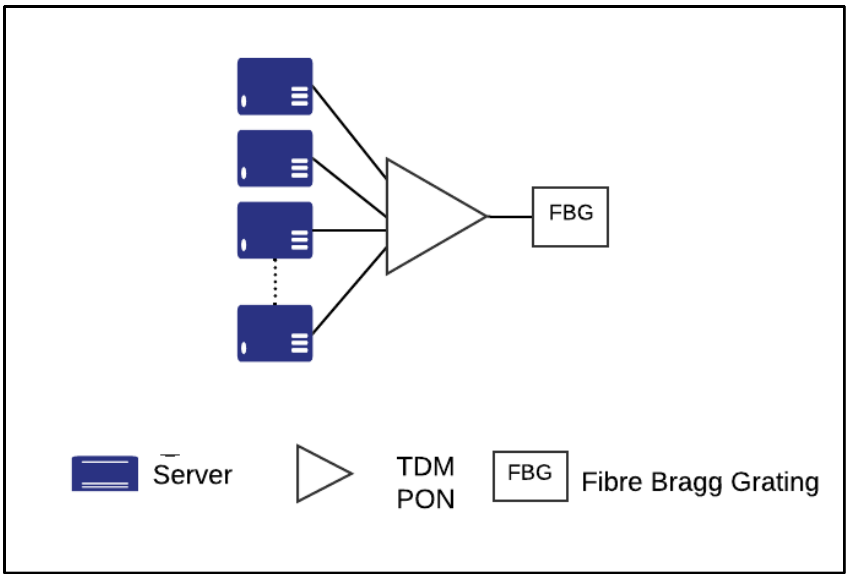


Figure 3.11: Fibre Bragg grating for intra-rack communication [49]

3.3.2.3 Passive polymer optical backplane

Similar to a reflection star coupler, a passive polymer optical backplane connects the servers in a rack in such a way that the intra-rack traffic is separated from the inter-rack traffic, as depicted in Figure 3.12. It offers a mesh of non-blocking connections that can connect typically up to ten or more server line cards [87], which means that the power input into any port is not divided among all the ports, as it is not broadcasted to all ports but transmitted to a specific port. Therefore, the number of servers in a rack does not affect the received power [49].

With the use of this backplane, the intra-rack traffic is separated from the inter-rack traffic, where any fault in the backplane affects only the intra-rack communication.

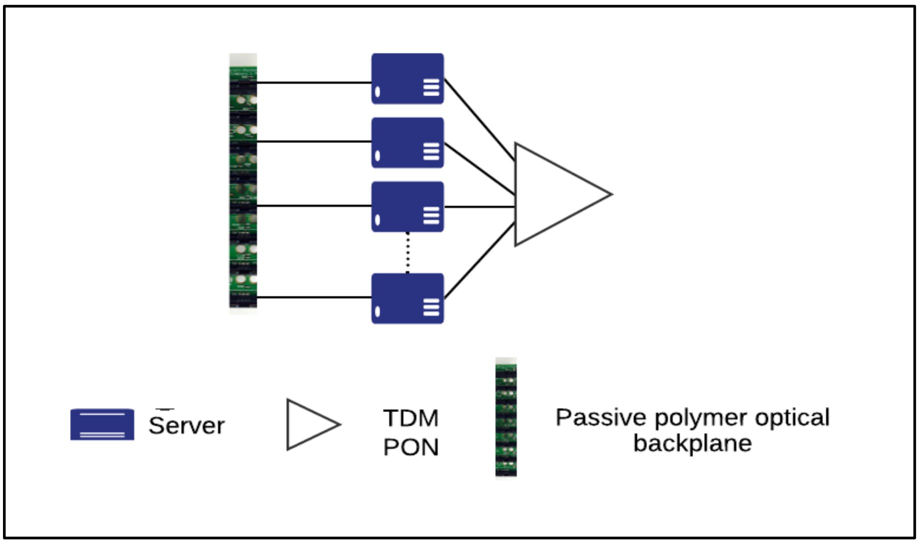


Figure 3.12: Passive polymer optical backplane for intra-rack communication [49]

3.3.3 Inter-rack connectivity

All the five architectures eliminate the need for aggregation and core switches. Moreover, the servers in each rack are divided into groups, where in each group, servers are connected to a coupler. The distribution of servers in groups within a rack is implemented to minimise the oversubscription rate and maximise the server share of resources as compared to connecting all the servers in a rack with one coupler [49].

3.3.3.1 Cellular PON data centre architecture

The data centre is divided into cells where each cell can have for example 16 to 256 servers connected in a cell architecture. The work in [49] has proposed 5 such cell architectures. A large data centre that has thousands or millions of servers is then constructed by interconnecting many cells in an optical hierarchy using AWGRs for example [49] as shown in Figure 3.13. Here, the hierarchy of AWGRs leads to the core routers (CRs) that connect the data centre to the outside world. The AWGRs fan-out the traffic to several OLTs with the load being balanced through a control and management system that links to the OLTs via a switch (SW). Figure 3.13 shows the overall proposed data centre architecture [49]. The focus of this thesis is on the cell architecture that fits in the “clouds” at the bottom of Figure 3.13.

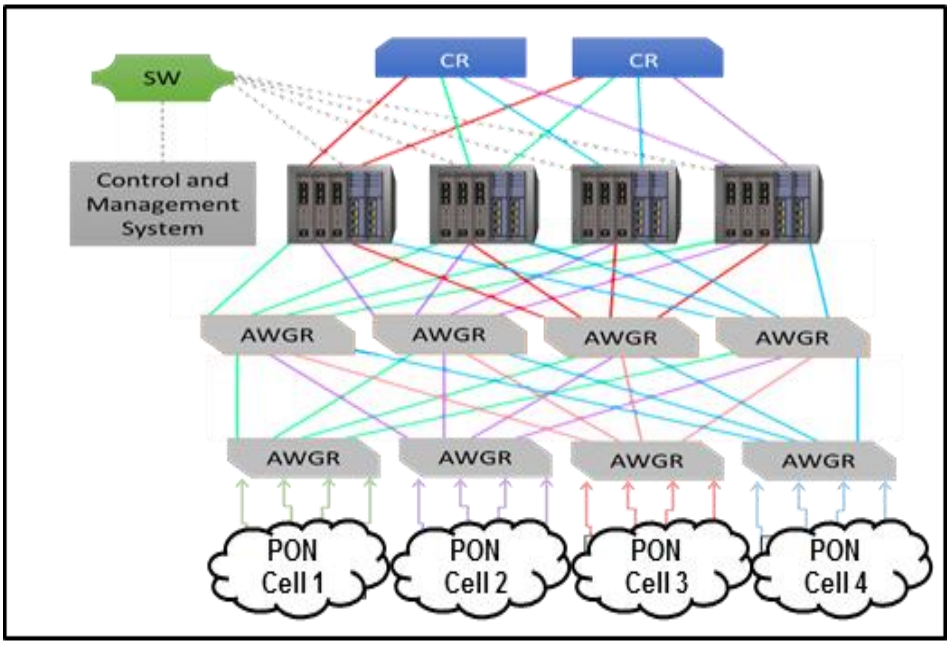


Figure 3.13: Cellular PON data centre Architecture [49]

The cellular architecture is very attractive as it enables scalability by allowing wavelength (and time slot) resources to be reused in different cells in a fashion similar to wireless cellular communications where a limited spectrum is reused in cells to cover a nation that may have tens to hundreds of millions of cellular phones (servers in our case). The work in [49] investigated in detail two of the five proposed architectures, named PON 3 and PON 5 architectures. Although [49] proposed the PON 4 architecture which reduces the number of tuneable lasers needed, it did not investigate PON 4. This thesis investigates PON 4 in detail. The five PON cell architectures are described next in detail for completeness. Each of the cell architectures described next replaces the cells structure at the bottom layer of Figure 3.13.

3.3.3.2 TDM-based PON cell data centre architecture (PON 1)

This architecture includes a TDM-based PON that is composed of only passive splitter and coupler devices that connect the cell’s racks to the OLT, as illustrated in Figure 3.14 [49]. It uses two wavelengths only: one for upstream and the other for downstream. This architecture is very similar to that of PON in an access network, where all the inter-rack traffic has to pass through the OLT.

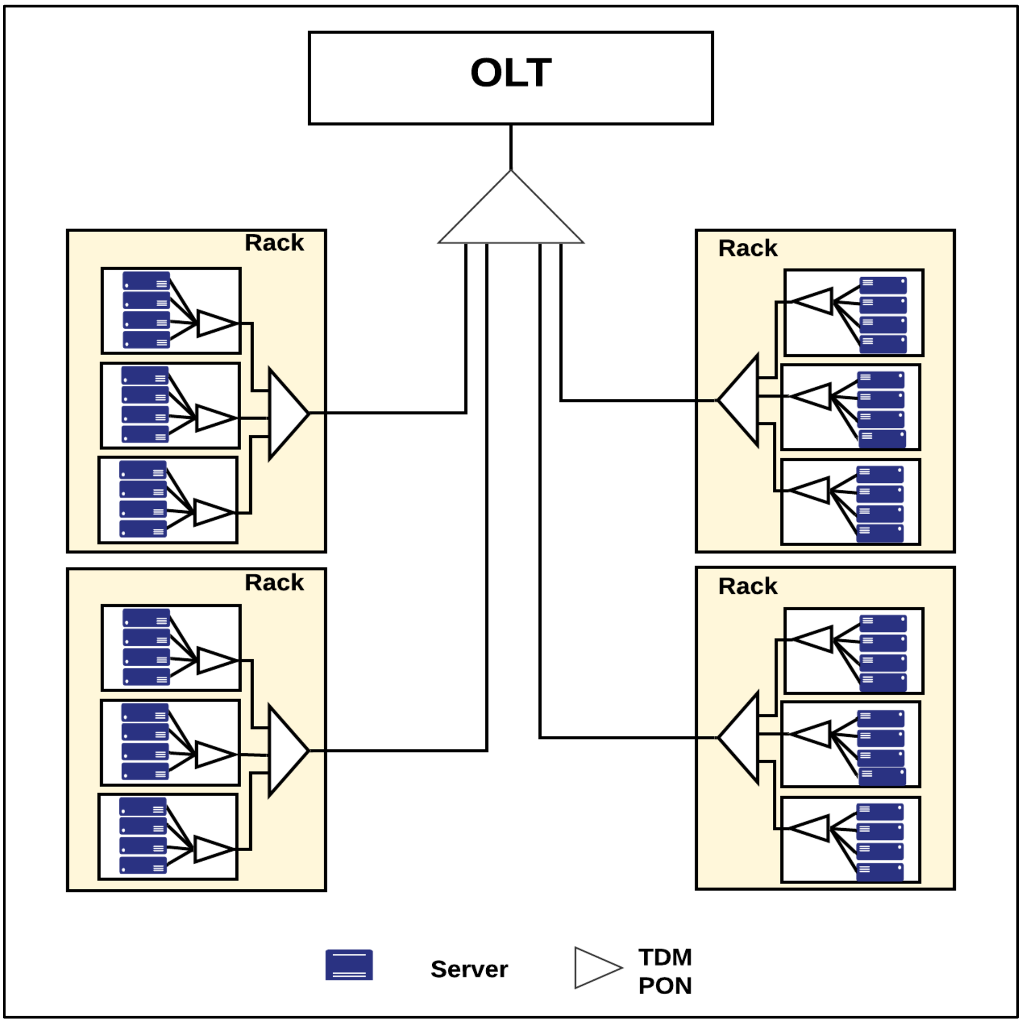


Figure 3.14: PON 1 cell architecture [49]

3.3.3.3 TDM/WDM-based PON cell data centre architecture (PON 2)

This architecture uses both TDM and WDM PON techniques, where only passive splitters, couplers and AWGR devices are utilised, as shown in Figure 3.15 [49]. Here, a specific wavelength for each rack’s uplink and downlink is used, which minimises congestion. Accordingly, high-cost laser diodes are required in the OLT and servers, although this can be avoided by using a multicarrier laser generator at the OLT, low cost multimode fibres and multimode transceivers at the servers [49]. This architecture is also very similar to that of PON in an access network, where all the inter-rack traffic has to pass through the OLT.

For TDM-based PON and TDM/WDM PON data centre architectures, having all the inter rack traffic passing through the OLT, where it needs to be buffered, processed, and rerouted to the desired rack, will result in additional unwanted delay and power consumption. The next three architectures were proposed in [49] to avoid this unnecessary action.

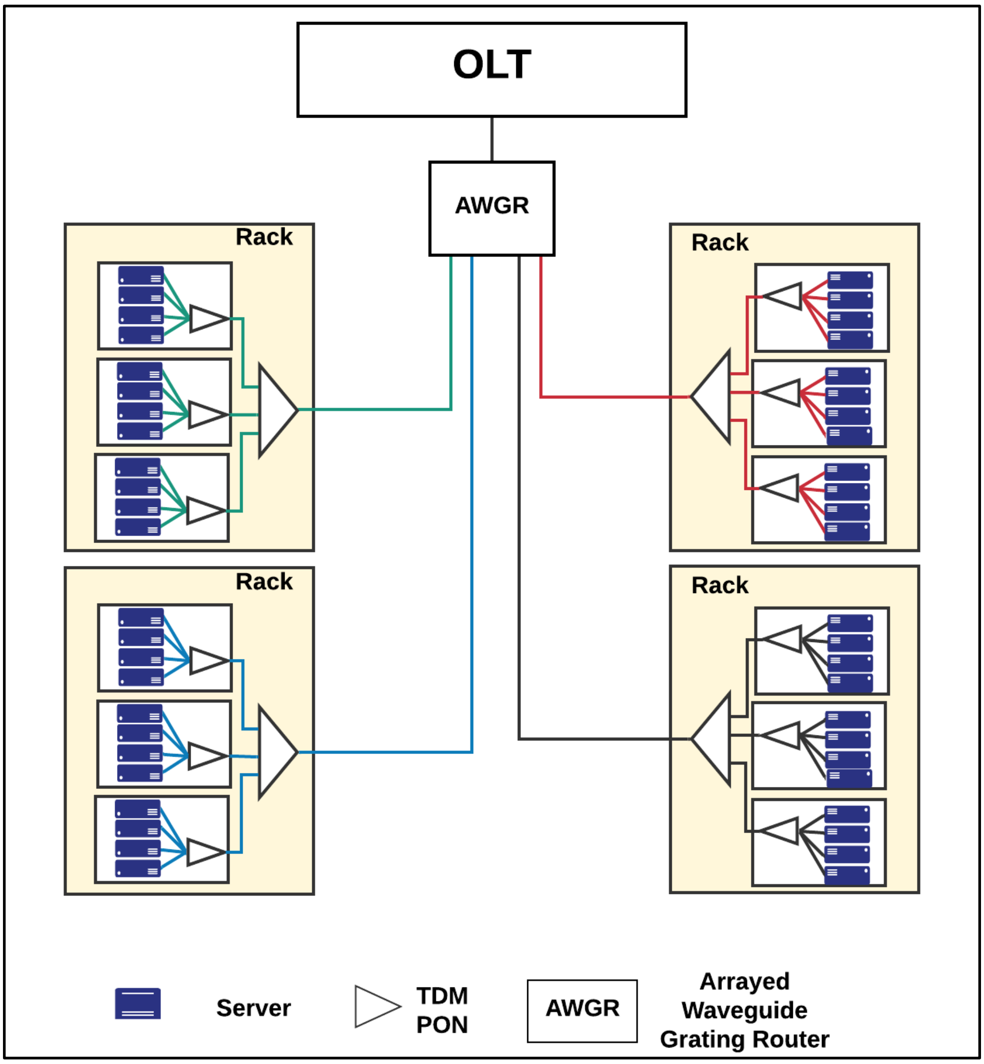


Figure 3.15: PON 2 cell architecture [49]

3.3.3.4 AWGR-based PON cell data centre architecture (PON 3)

In this architecture, the servers in each rack are either divided into groups or placed in one group [29]. Therefore, there are two types of traffic in this architecture: intra-group traffic and inter-group traffic. With respect to the intra-group traffic, one of the previously described intra-rack connections is deployed. In contrast, the inter-group communication is achieved by connecting these groups via two N × N AWGRs, as shown in Figure 3.16, where N is the number of groups. Thus, the number of wavelengths needed for the inter-group communication is equal to the number of groups in the PON cell. In order for each server to reach the servers in the other groups, a tuneable PON optical network unit (ONU) or network interface card (NIC) has to be fitted in the server. Regarding the NIC, it needs to contain tuneable lasers for wavelength detection and selection, along with an array of fixed tuned receivers.

The routing of the inter-group traffic is accomplished either by the OLT, which adds unnecessary delay and power consumption, or via the two N × N AWGRs, where a direct connection is possible because of the use of ONUs. The ONUs allow the servers to use a specific wavelength to communicate with a server in a specific group.

This architecture reduces the power consumption and delay, as it only uses passive components for connectivity. However, it increases the deployment cost because of the use of an ONU for every server.

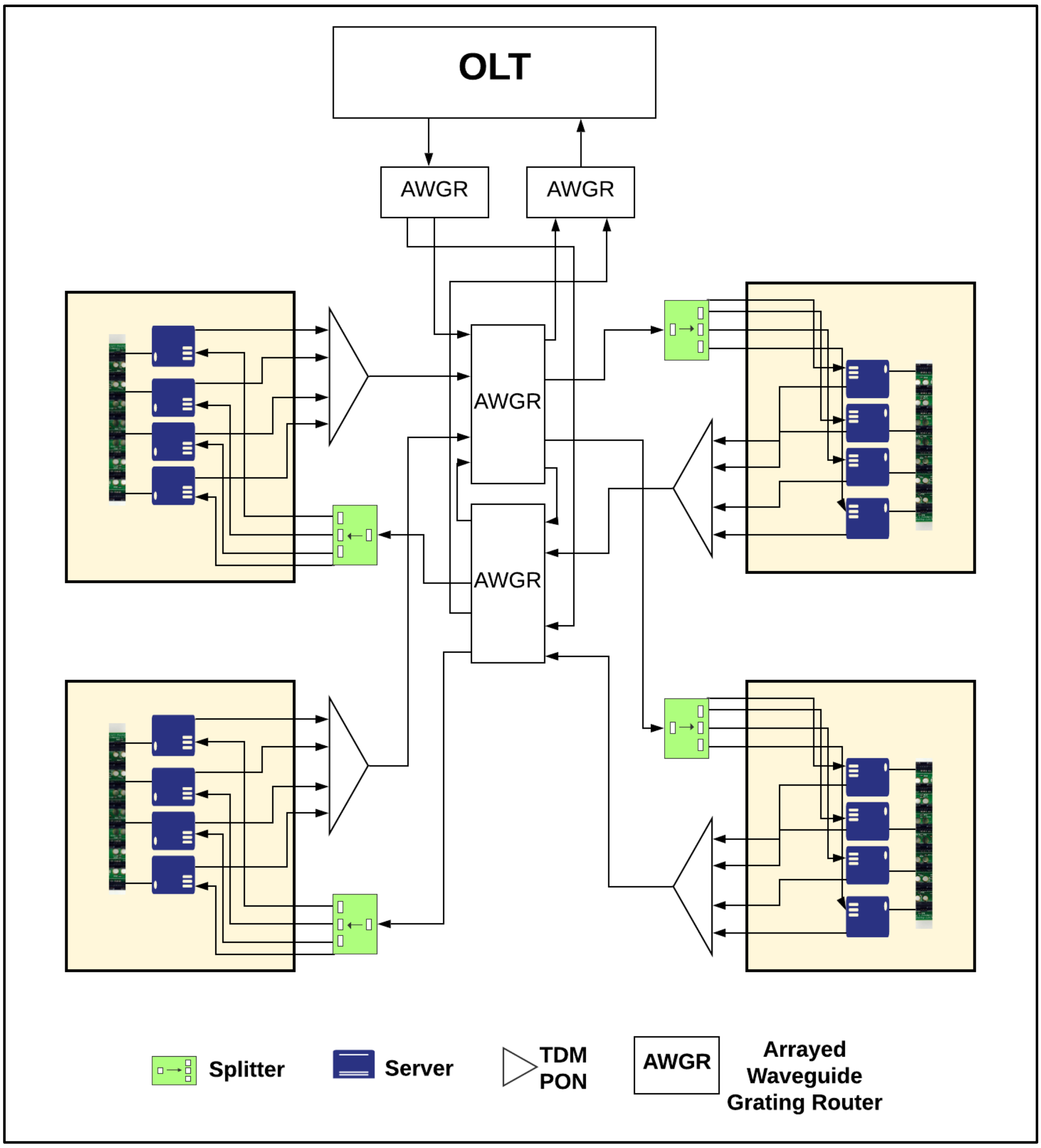


Figure 3.16: PON 3 cell architecture [29]

3.3.3.5 AWGR and server-based PON cell data centre architecture (PON 4)

The aim of this architecture is to reduce the number of tuneable lasers used, thereby reducing the deployment cost. Here, as shown in Figure 3.17 [49], the servers in each rack are either divided into groups or placed in one group. A group is made of multiple subgroups, and the number of servers hosted by a subgroup depends on the splitting ratio of the TDM PON connected to it. For example, a group can host 16 servers placed into two subgroups of eight servers each.

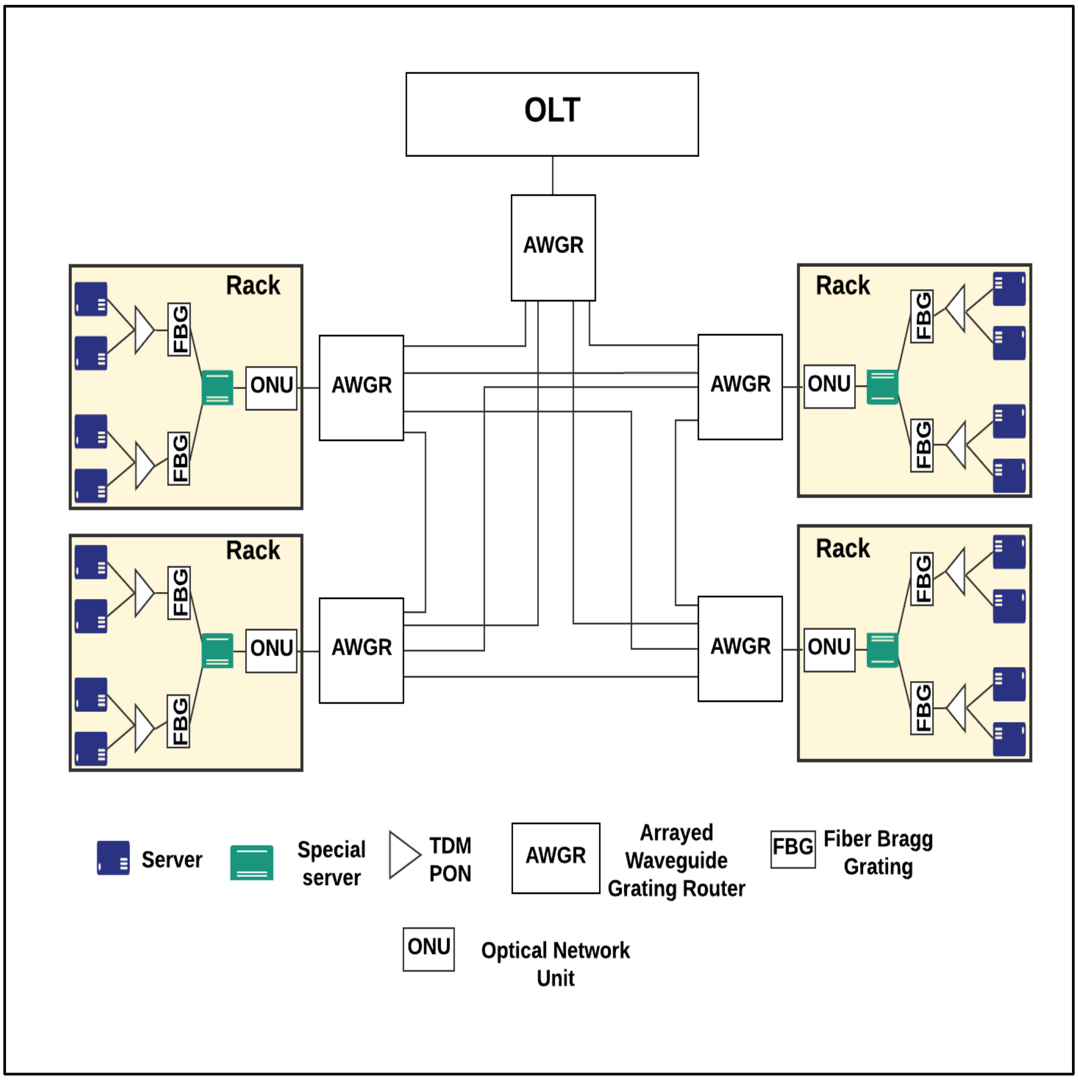


Figure 3.17: PON 4 cell architecture [49]

This architecture involves two types of communication: intra-group and inter-group communication. The intra-group communication can be either between the servers in a subgroup (intra-subgroup communication) or between the servers in the different subgroups in the group (inter-subgroup communication). The subgroups in each group are connected to a special server whose task is to maintain the inter-subgroup and inter-group communication. These special servers are equipped with tuneable lasers in the ONUs, which make it possible for the special servers to use a specific wavelength to direct the traffic from a server in its group to a server in another group. The rest of the servers do not have tuneable lasers and this significantly reduces the number of tuneable lasers needed in the architecture as the number of special servers (gateway servers) is small, with one such server per group.

The intra-subgroup communication is achieved using a fibre Bragg grating (FBG), which reflects only the wavelength assigned for the intra-subgroup communication. All the subgroups can use the same wavelength for the intra-subgroup communication because of the fact that this wavelength does not pass through the FBG to the other groups. Accordingly, the design of the servers' transceivers will be simplified and unified as the same wavelength is used for transmitting and receiving. The servers should get permission from the special server (that lies after the FBG) for the intra-subgroup communication wavelength. This will help eliminate collision and control the channel access contention.

Each subgroup uses only two wavelengths for the inter-group and the inter-subgroup communication: one for uplink communication and the other for downlink communication. For the inter-subgroup communication, the wavelength passes the FBG to the special server that sends it back to the same group. As for the inter-group communication, each special server is connected to two AWGRs: one for sending to other groups and the other for receiving from them. The number of wavelengths needed for the inter-group communication is equal to the number of groups in the PON cell. In order for special servers to perform effectively, they keep a database which contains the server addresses in the groups and the wavelength assigned to each group. The special server manages the inter-group communication by applying wavelength conversion and connects to other AWGRs or the OLT card. In addition, these special servers share information with each other, which is done directly or via the OLT, in order to update their databases. This architecture is a scalable design that can be expanded to support hundreds of thousands of servers.

3.3.3.6 Server-based PON cell data centre architecture (PON 5)

This architecture [37] further reduces the deployment cost by completely eliminating the need for tuneable ONUs. It however relies on server forwarding and is therefore slower in terms of throughput. It depends on the data centre's servers themselves to implement the routing inside the PON cell. In this architecture, the servers in each rack must be divided into groups. The number of groups in each rack is equal to the number of racks in the PON cell.

Regarding the intra-rack traffic, one of the previously described intra-rack connections can be deployed. For the inter-rack traffic, each group in a rack is responsible for connecting the rack to one of the other racks, as shown in Figure 3.18. Moreover, one of the groups in a rack should only be responsible for connecting the rack to the OLT. The OLT is connected to the racks either via a star coupler or by using an AWGR. For example, if a server in Rack 1 wants to communicate with a server in Rack 2, the traffic is routed through a server in Group 2 of Rack 1 and is then delivered to a server in Group 2 of Rack 2 and then passed to the destined server.

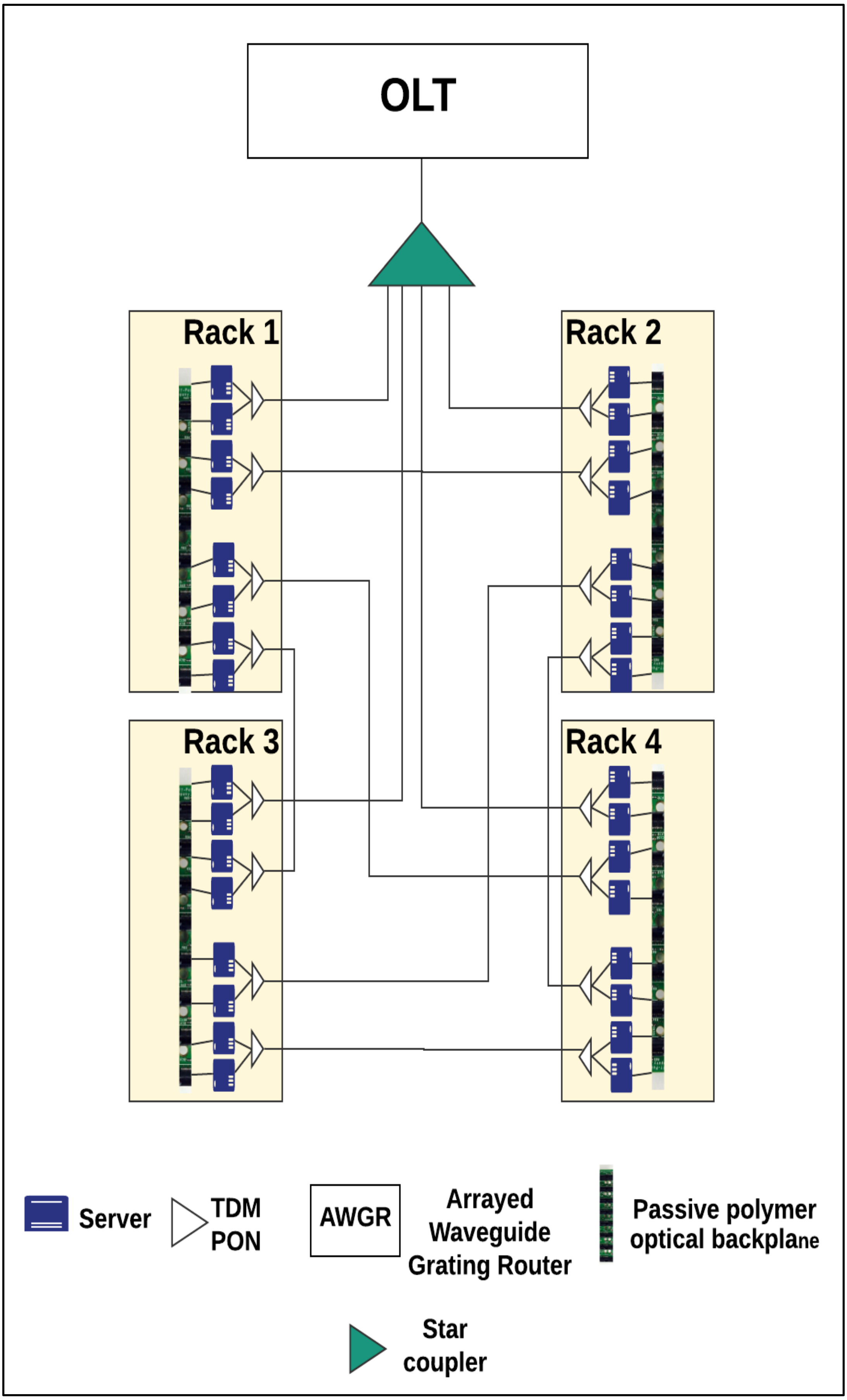


Figure 3.18: PON 5 cell architecture [37]

3.3.4 Significance of the AWGR and server-based PON cell data centre architecture (PON 4)

To sum up, the AWGR-based PON data centre (PON 3) architecture has a high deployment cost because of the use of an ONU that has a tuneable laser in every server. It has high throughput though as it avoids routing via servers. The server-based PON data centre architecture (PON 5) reduces the deployment cost, as it eliminates the use of ONUs and AWGRs. However, it increases the delay, as it requires hopping between servers for routing traffic. On the other hand, the AWGR and server-based PON data centre architecture (PON 4) achieves a trade-off between cost and delay by equipping only the special servers with ONUs that have tuneable lasers and assigns the routing responsibilities to the special servers which reduces the delay. In this thesis, the AWGR and server-based PON data centre architecture is considered for further thorough investigation.

3.4 Summary

This chapter discussed the use of PONs in access networks starting by giving a historical overview and then describing the architecture, topologies, technologies, and components. The PON topologies studied included tree, bus, and ring and the technologies studied included APON, BPON, EPON, 10G-EPON, GPON, and NG-PON. The components of PONs reviewed were the splitter, coupler, AWG, FBG, and the passive polymer optical backplane. This chapter also discussed the deployment of PON in data centre architectures. A detailed description of five recently proposed fully passive optical architecture options was presented. The intra-rack connectivity in these designs could be one of three types, namely reflection star coupler, FBG, and passive polymer optical backplane. The inter-rack connectivity proposed in these designs include TDM-based PON cells, TDM-WDM-based PON cells, AWGR-based PON cells, server-based PON cells, and AWGR and server-based PON cell data centre architectures. The AWGR and server-based PON data centre architecture is chosen for further thorough investigation, as it gives a trade-off between the features of the AWGR-based PON and server-based PON data centre architectures.

Chapter 4 Wavelength Assignment in PON Data Centre Design with AWGR and Server-Based Routing

4.1 Overview

In this chapter, the wavelength routing and assignment within the AWGRs of the cell of the AWGR and server-based PON data centre architecture, described in Section 3.3.3.5, are optimised using an MILP model. Furthermore, a benchmark study is presented to compare the power consumption of the AWGR and server-based PON data centre architecture against a traditional server-centric data centre architecture, BCube.

4.2 Communications in the AWGR and server-based PON data centre cell

When a server in the AWGR and server-based PON data centre cell, depicted in Figure 4.1, needs to talk to another server in the same subgroup, it transmits the data to the connected FBG. Then, the FBG reflects the data back to the subgroup to be received by the destination server. On the contrary, if the server needs to talk to another server in another subgroup, it first talks to the special server (gateway) to ask for a grant. Accordingly, the special server replies to the requesting server to offer the grant and tunes its transceiver to the destination wavelength. In order for the special server to communicate with the servers, TDMA over WDM or any other orthogonal technique can be used to route data to a given destination server.

This work is the first to analyse the AWGR and server-based PON data centre architecture and, hence, is concerned with the lowest layer, i.e., the WDM layer. Accordingly, this work designs and optimises the routing from a group to another group.

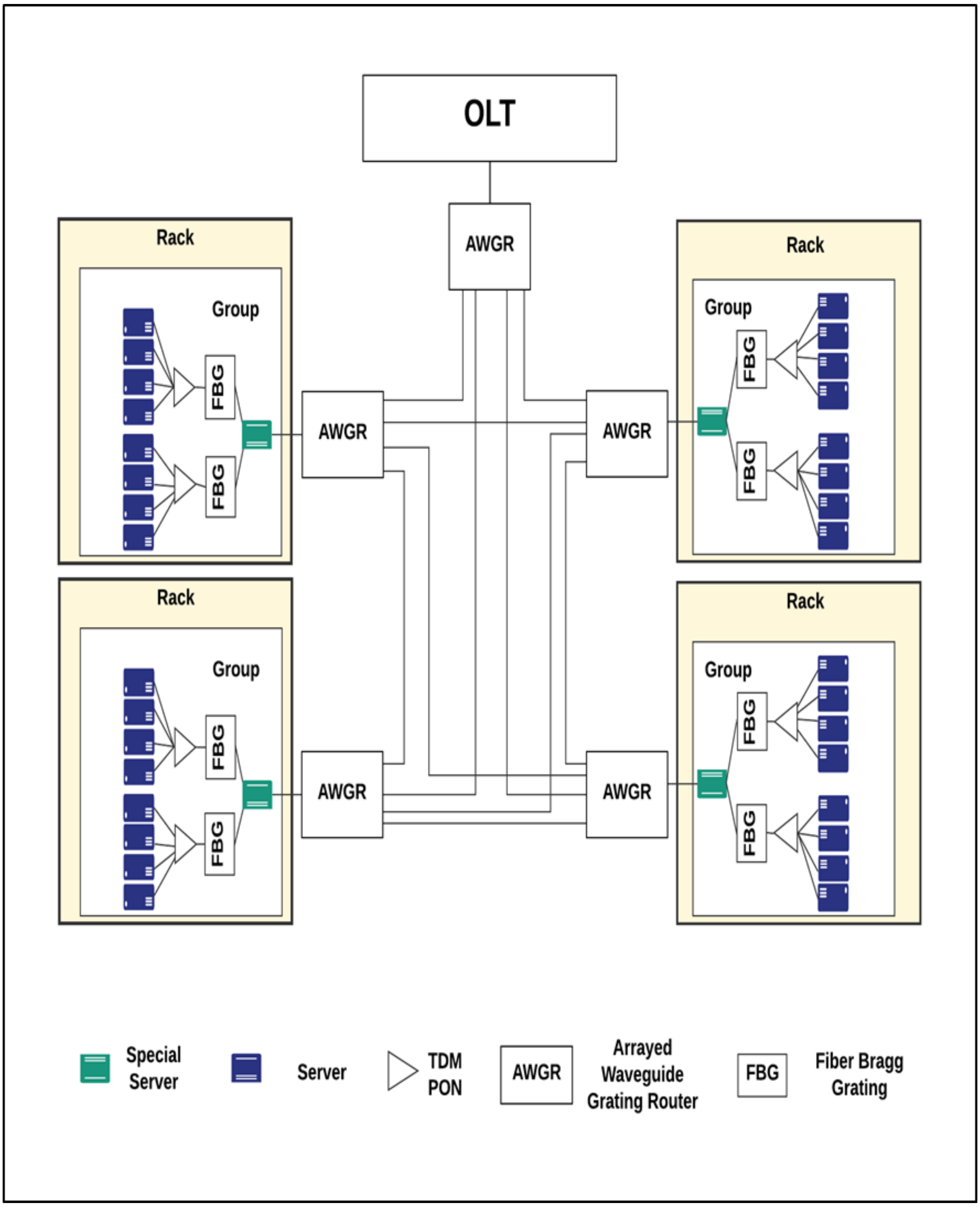


Figure 4.1: Simplified AWGR and server-based PON data centre architecture

4.3 MILP model description for wavelength routing and assignment

This work is the first to analyse the AWGR and server-based PON data centre architecture and, hence, is concerned with designing and optimising its routing and wavelength assignment. A Mixed Integer Linear Programming (MILP) model is developed to optimise the routing and wavelength assignment using the AWGRs of the PON cell shown in Figure 4.1 to support inter-group connections and connections between the OLT and the groups. The sets, parameters, and variables used in this model are as follows:

Sets:

N Set of nodes (AWGRs' ports, PON groups, and the OLT)

P Set of PON groups and OLT, where $P \subset N$

W Set of wavelengths

N_m Set of neighbouring nodes of node $m \in N$ that receive from node m

NB_m Set of neighbouring nodes of node $m \in N$ that transmit to node m

Variables:

μ_{sd}^{j} μ_{sd}^{j} = 1 if wavelength j is used for the connection (s, d); otherwise, μ_{sd}^{j} = 0, where $j \in W$ and $s, d \in P$.

φ_{sd}^{jmn} φ_{sd}^{jmn} = 1 if wavelength j on link (m, n) is used for a connection (s, d); otherwise, φ_{sd}^{jmn}= 0, where $j \in W, s, d\ P$ and $m \in N\ and\ n \in N_m$.

The model is defined as follows:

Objective:

Maximise:

$$\sum_{s\in P}\sum_{\substack{d\in P\\ s\neq d}}\sum_{j\in W} \mu_{sd}^{j} \tag{4.1}$$

Equation (4.1) states the model objective, which is to maximise the number of connections supported by the AWGRs. This equation works under two conditions — if the physical topology has a physical link or possible physical route from every source to every destination and if there are enough wavelengths—both of which are satisfied in our model.

Subject to:

$$\sum_{j\in W} \mu_{sd}^{j} \leq 1 \tag{4.2}$$

$$\forall\, s,d \in P, s \neq d$$

Constraint (4.2) guarantees that only a single wavelength is used for a connection among the PON groups and between PON groups and OLT.

$$\sum_{\substack{s\in P\\ s\neq d}} \mu_{sd}^{j} \leq 1 \tag{4.3}$$

$$\forall\, d \in P, \forall\, j \in W$$

Constraint (4.3) ensures that different source nodes use different wavelengths to communicate with a destination node. Hence, each destination receives a different wavelength from each transmitting source.

In addition, we introduce:

$$\sum_{\substack{d \in P \\ s \neq d}} \mu_{sd}^{j} \quad \leq \quad 1 \tag{4.4}$$

$$\forall\, s \ \in P, \forall\, j \in W$$

Constraint (4.4) ensures that each source node uses different wavelengths to communicate with different destination nodes. Hence, each source transmits to different destinations on a different wavelength.

$$\sum_{\substack{n \in N_m \\ m \neq n}} \varphi_{sd}^{jmn} \ - \sum_{\substack{n \in NB_m \\ m \neq n}} \varphi_{sd}^{jnm} \ = \begin{cases} \mu_{sd}^{j}, & m = s \\ -\mu_{sd}^{j}, & m = d \\ 0 & otherwise \end{cases} \tag{4.5}$$

$$\forall\, s, d \in P\,, \forall\, m \in N\,, \forall\, j \in W, s \neq d$$

Constraint (4.5) represents the wavelength continuity flow conservation constraint following the flow conservation law [95]. It guarantees that the flow entering a node at a certain wavelength departs the node at the same wavelength for all the nodes apart from the source and the destination.

$$\sum_{s \in P} \sum_{\substack{d \in P \\ s \neq d}} \varphi_{sd}^{jmn} \quad \leq \quad 1 \tag{4.6}$$

$$\forall\, m \ \in N\,, \forall n \in N_m\,, \forall\, j \in W$$

Constraint (4.6) ensures that a wavelength is only used once on a link.

$$\sum_{s \in P} \sum_{d \in P} \sum_{n \in N_i} \sum_{j \in W} \varphi_{sd}^{jin} \ - \sum_{\substack{d \in P \\ d \neq i}} \sum_{j \in W} \mu_{id}^{j} \ \leq 0 \tag{4.7}$$

$$\forall\, i \ \in P$$

Constraint (4.7) prevents relaying flows by PON groups by ensuring that a PON group relays only the traffic generated by itself. The first term gives the traffic traversing output links from a PON group, and the second term gives the traffic originating from the PON group. The difference between the two terms should always be zero.

4.4 Wavelength routing and assignment results

We consider a PON cell composed of 96 servers divided into six groups of 16 servers. Each group is divided into two subgroups, as shown in Figure 4.2. Table 4.1 shows the resulting wavelength assignment for the inter-group communication and the communication between PON groups and the OLT, which is also shown in Figure 4.2.

Table 4.1: Optimised wavelength assignment for the proposed architecture obtained from the MILP

Destination / Source	PON Group 1	PON Group 2	PON Group 3	PON Group 4	PON Group 5	PON Group 6	OLT
PON Group 1		λ_5	λ_3	λ_2	λ_1	λ_4	λ_6
PON Group 2	λ_2		λ_1	λ_4	λ_6	λ_5	λ_3
PON Group 3	λ_5	λ_1		λ_3	λ_2	λ_6	λ_4
PON Group 4	λ_6	λ_4	λ_5		λ_3	λ_2	λ_1
PON Group 5	λ_4	λ_2	λ_6	λ_1		λ_3	λ_5
PON Group 6	λ_1	λ_3	λ_4	λ_6	λ_5		λ_2
OLT	λ_3	λ_6	λ_2	λ_5	λ_4	λ_1	

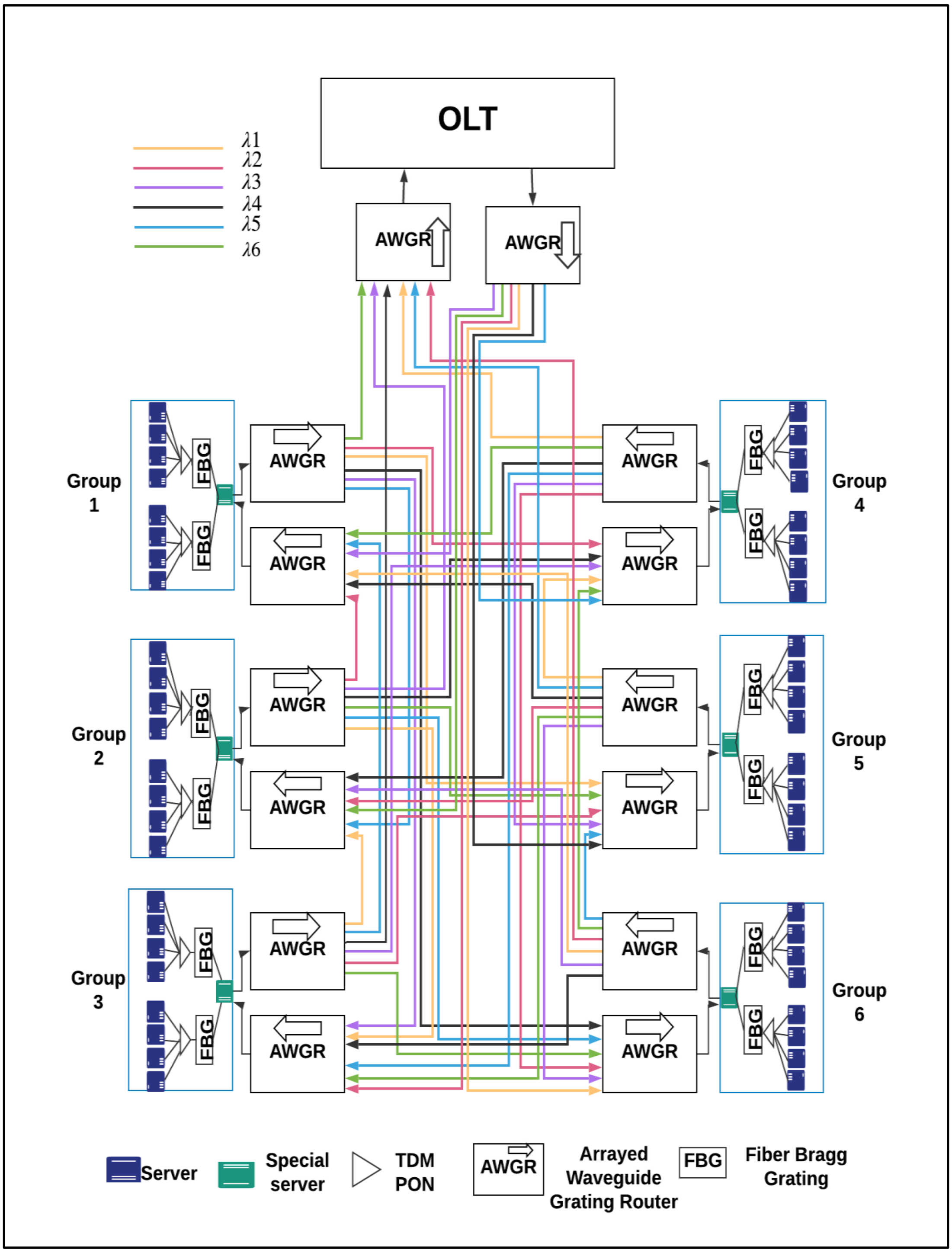


Figure 4.2: Optimised wavelength assignment for the proposed architecture

4.5 Power consumption benchmark

A benchmark study which compares the power consumption of our design to a server-centric traditional data centre architecture (BCube) is presented in this section.

4.5.1 BCube architecture

The BCube architecture is characterised by the use of servers for routing and forwarding decisions along with the switches. As shown in Figure 4.3, this recursive architecture's elementary unit is called BCube0. Accordingly, the number of servers, n, hosted by BCube0 matches the number of ports of the single commodity switch connecting them. Furthermore, combining $n\ BCube0$s is achieved by connecting them via n commodity switches which forms $BCube1$, where the switches are only connected to the servers and not allowed to connect to another switch. In general, a BCube data centre architecture is referred to as $BCubek$, which is composed of $n\ BCubek - 1$s, $(k+1)n^k$ switches, and n^{k+1} servers, where k is the level number [60].

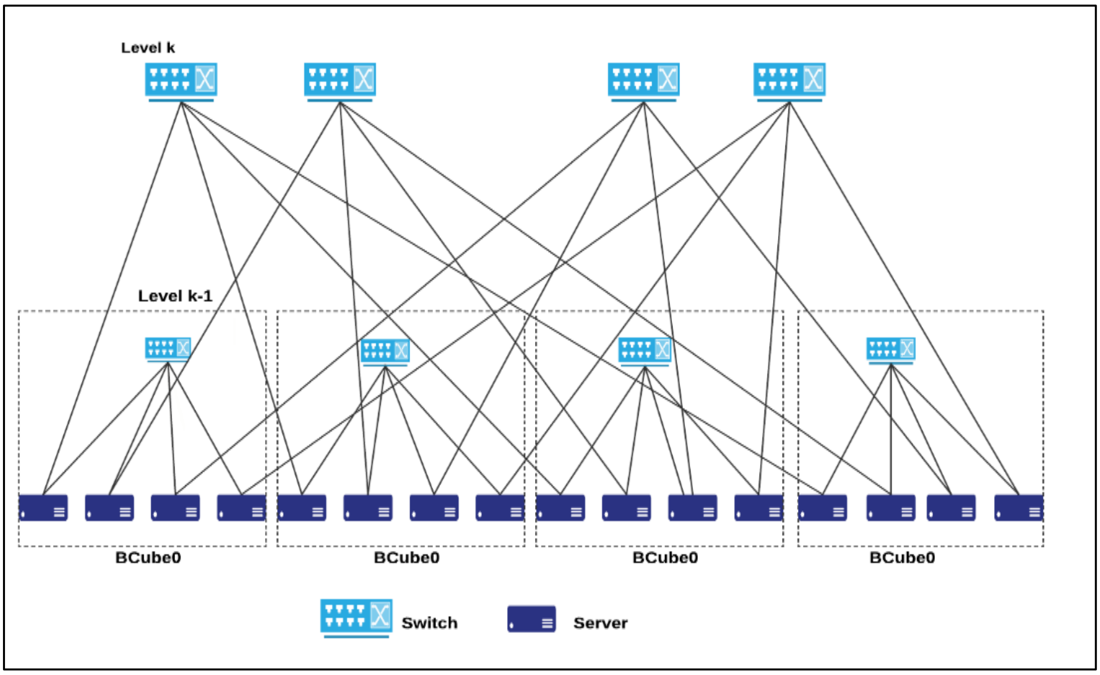


Figure 4.3: BCube architecture when n = 4, k = 1 [60].

The power consumption of the BCube architecture (P_{BCube}) is calculated assuming that all the servers participate in the routing, (as many routes cannot be spanned by using switches only), as expressed in the following equation [96]:

$$P_{BCube} = \; PT \;\, N_s \;\, L \; + \; \rho \, PS_s \, N_s + \;\; PW \;\, N_w \tag{4.8}$$

where PT, PS_s, and PW represent the server's transceiver power consumption, server's maximum power consumption, and the switch power consumption, respectively. Furthermore, L is the number of levels, and ρ is the fraction of the server power used for communication which is a dynamically changing variable. In addition, N_s and N_w are the total number of servers and the total number of switches used in the BCube architecture, respectively, and are calculated as follows:

$$L = k + 1 \tag{4.9}$$

$$N_s = n^L \tag{4.10}$$

$$N_w = L \, n^k \tag{4.11}$$

4.5.2 The AWGR and server-based PON data centre architecture

The power consumption of this architecture (P) is calculated using the following equation:

$$P = \; N_c \;\, N_r \;\, (\, \beta \, PR_r \; + PO \,) \;\; + \;\;\; N_P \;\;\, OMP + \;\, N_c \;\, N_r N_g \;\;\, PT \tag{4.12}$$

where PR_r and PO represent the power consumption of special servers and the power consumption of a tuneable ONU device. Moreover, OMP is the power consumption of an OLT port, while PT is the server's transceiver power consumption. N_c represents the number of cells, and N_r is the number of special servers needed in a cell. Furthermore, N_p is the number of OLT ports needed for connectivity, whereas N_g is the number of servers per group; β is used to ensure that the special server power used for communication in the AWGR and server-based PON data centre architecture is comparable to the server power used for communication in the BCube architecture and is calculated as follows:

$$\beta = \rho \frac{N_g}{n-1} \tag{4.13}$$

where n is the number of servers hosted in BCube0. $N_c, N_r, and\ N_P$ are determined as follows:

$$N_c = \frac{N_a}{N_e} \tag{4.14}$$

where N_a is the total number of servers used and N_e is the number of servers per cell.

$$N_r = \frac{N_e}{N_g} \tag{4.15}$$

According to the design, each PON cell is connected to an OLT port. Therefore,

$$N_p = N_c \tag{4.16}$$

4.5.3 Benchmark results

Here, we compare the power needed to provide connectivity for 512, 4096, and 32768 servers using the BCube and the AWGR and server-based PON data centre architecture. The BCube architectures are based on n = 8 and k = 2, 3, and 4. The values used for ρ are 0.05, 0.1, 0.2, and 0.3. The AWGR and server-based PON data centre architectures are considered using different PON cell sizes of 64, 96, and 128 servers, where 128 is the maximum number of servers supported by an OLT port, to evaluate the impact of cell size on power consumption. Each PON cell is composed of a number of groups, where each group hosts 16 servers.

The power consumption of the devices of the AWGR and server-based PON data centre architecture used in this benchmark study are shown in Table 4.2. Furthermore, the power consumption of the devices of the BCube architecture used in this benchmark study are shown in Table 4.3.

Table 4.2: the AWGR and server-based PON data centre architecture component power consumption

Device	Power Consumption
10-Gb/s tuneable ONU	2.5 W [97]
Special server's full power	457 W [98]
10-Gb/s OLT port	14.3 W [99]
Server's transceiver power	3 W [96]

Table 4.3: : BCube architecture component power consumption

Device	Power Consumption
Server's transceiver	3 W [96]
Cisco 2960-8TC-L	12 W [100]
Server's full power	457 W [98]

Regarding the OLT power consumption [99], the OLT considered in this model has 16 XG-PON cards, where each card has 8 ports and the OLT port power consumption is calculated on the basis of the maximum power consumption of the OLT switching, power cards, and the fan card.

As for the ONU power consumption [49], note that the power consumption of the ONU used in the PON data centres differs from that of the ONU used in access networks (FTTx) because of the differences in the architecture and the services provided. Regarding architecture differences, ONUs (subscribers) in FTTx access networks can be positioned up to 20 km away from the OLT, while in PON data centres, the distance between ONUs and the OLT does not exceed 1 km typically, even for a large data centre, and is usually under 100m. The distance between ONUs and the OLT has a direct effect on the ONU’s transceiver power consumption. As the ONU’s transceiver power consumption used in FTTx access networks is 3.5 W, and when a linear profile is assumed, the ONU’s transceiver power consumption used in the PON data centre is 17.5 mW.

As for the services provided, FTTx access networks support video, audio, and data services, while in PON data centres, only data services are required, which eliminates the need for a subscriber line interface (SLIC) and the Multimedia Over Coax (MoCA) used for audio and video services, respectively. In addition, a passive GbE switch can replace the GbE switch to reduce the power consumption. Accordingly, the total power consumption of the ONU used in a PON data centre is 2.72 W considering all the ONU components needed. Figure 4.4 shows the architecture of the ONU used in an FTTx access network, while Figure 4.5 shows the architecture of the ONU used in a PON data centre. Furthermore, Table 4.4 presents the power consumption of the main components of an ONU used in an FTTx access network.

Table 4.4: 10-G ONU components' power consumption [49]

Component	Power Consumption
Transceiver	3.5 W
DDR	0.7 W
System-on-chip (SoC)	2 W

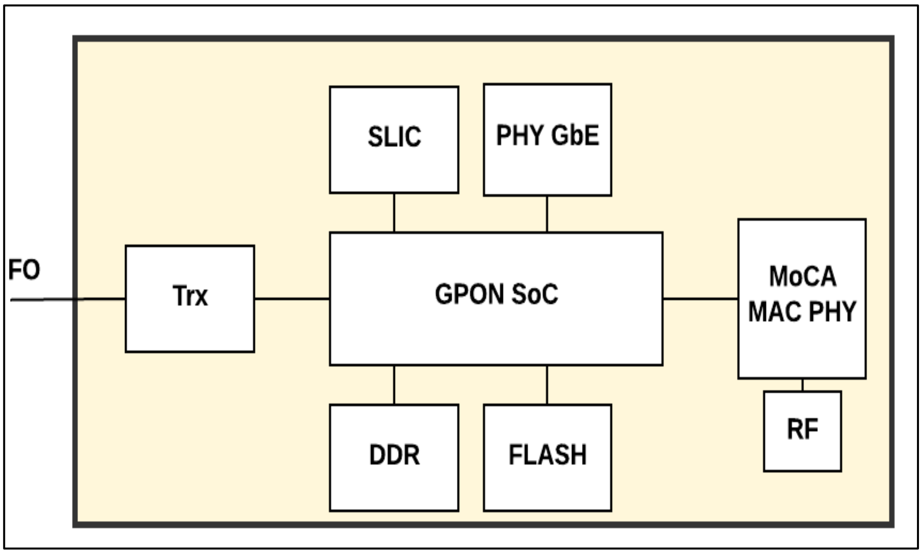


Figure 4.4: ONU architecture in FTTx access network [49, 101]

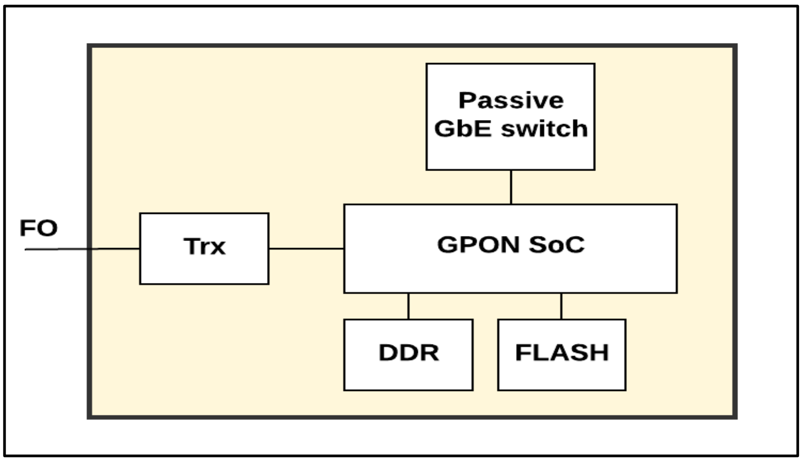


Figure 4.5: ONU architecture in PON data centre [49]

Figure 4.6 shows the power savings of the AWGR and server-based PON data centre architecture over those of the BCube data centre considering the different cell sizes of 64, 96, and 128 servers for $\rho = 0.05$ case to study the impact of cell size at a given value of ρ. This figure reveals several points.

Firstly, it shows that as the number of servers increases, the AWGR and server-based PON data centre architecture power savings compared to BCube's increase. This is attributed to the linear increase in the AWGR and server-based PON data centre architecture's power with an increase in the number of servers, while BCube's power consumption increases exponentially. This is due to the fact that the BCube architecture depends on servers for routing in addition to the commodity switches. This results in the servers having multiple transceivers to handle connectivity with all the levels. Therefore, as the number of levels increases in the BCube topology, the power consumption increases, because more transceivers are needed to establish connections to switches at every level. Hence, the power savings of our proposed design as compared to those of BCube increase with an increase in the number of BCube's levels.

Secondly, the AWGR and server-based PON data centre cell size impact on the power consumption savings is limited as compared to the BCube data centre. This is because the main contributor to the power consumption of the AWGR and server-based PON data centre architecture compared to BCube architecture is the special servers. The number of special servers is a function of the number of groups (a group is composed of 16 servers for the different cell sizes). The only impact of the cell size is on the OLT port power consumption, which is considerably lower than the special server's power consumption. Accordingly, for $\rho = 0.05$, the AWGR and server-based PON data centre architecture succeeds in minimising the power consumption by 80% compared to a BCube topology of 512 servers, by 82% as compared to a BCube topology of 4096 servers, and by 83% as compared to a BCube architecture of 32768 servers.

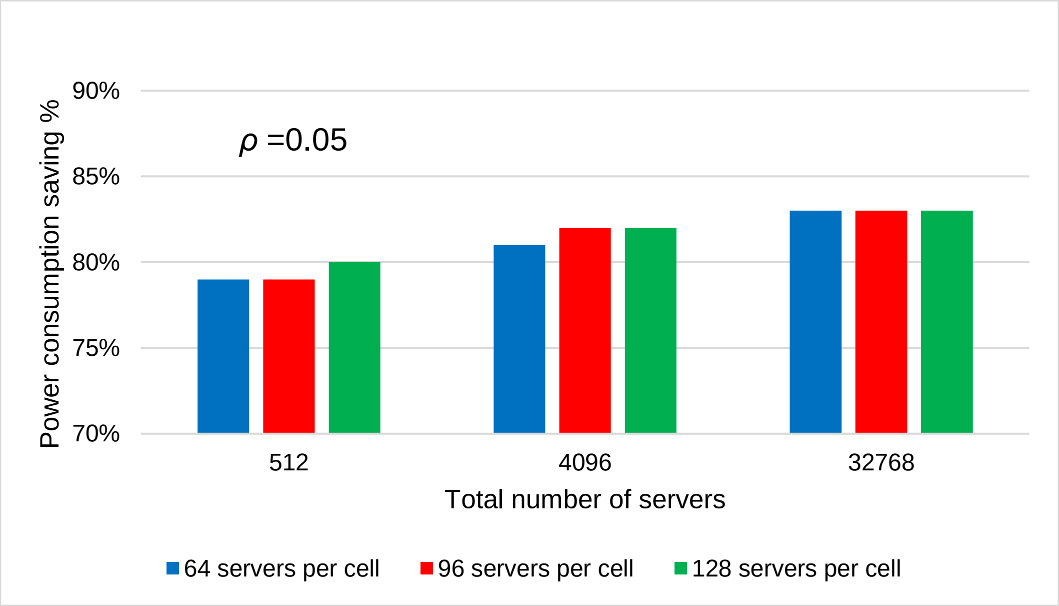


Figure 4.6: Power consumption benchmarking study comparing the AWGR and server-based PON data centre architecture to BCube architecture using $\rho = 0.05$

Note that the larger the number of servers in a PON cell is, the smaller is the resource share that they can have. Moreover, as the cell size effect on the power consumption savings as compared to the BCube architecture's is limited, we considered the cell size of 96 servers. Figure 4.7 presents a comparison of the power savings of the AWGR and server-based PON data centre architecture and the BCube data centre considering different values for ρ, which represent the fraction of the server power used for communication in the BCube architecture and affect the special server power used for communication in the AWGR and server-based PON data centre architecture, as shown in equation (4.16). In this figure, the results show that the different values of ρ have a limited effect on the power savings of the AWGR and server-based PON data centre architecture as compared to those of the BCube data centre. The reason behind this is that ρ effects only the power consumption of the special servers in the AWGR and server-based PON data centre architecture, while it has effect on the power consumption of all the servers in BCube data centre.

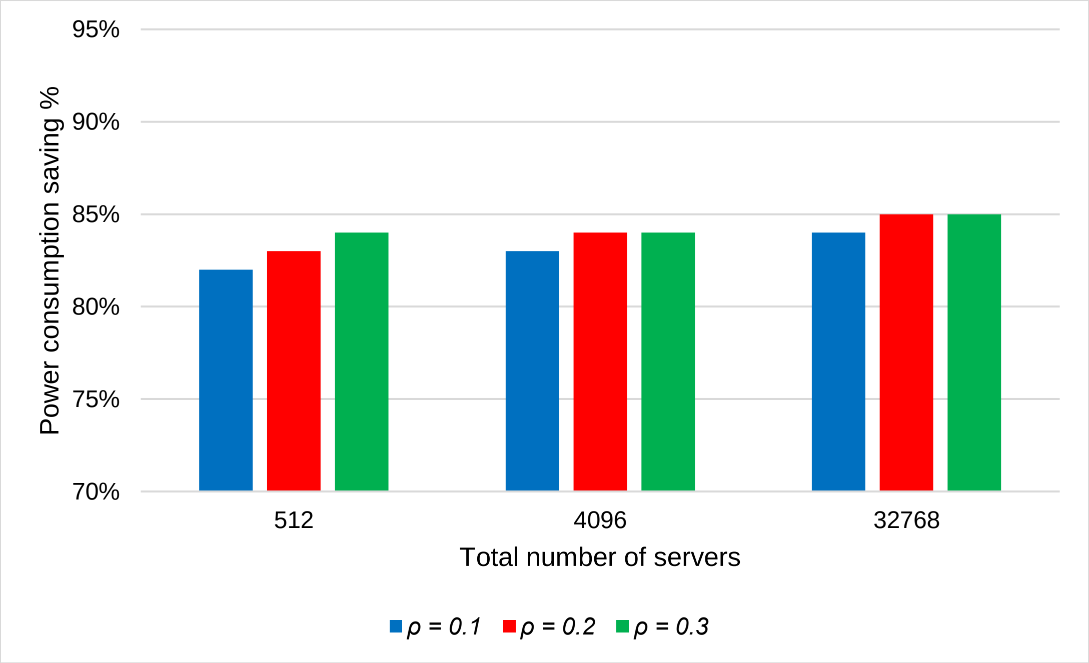


Figure 4.7: Power consumption benchmarking study comparing the AWGR and server-based PON data centre architecture to BCube architecture using different values of ρ

4.6 Summary

In this chapter, the development of an MILP model to optimise the wavelength routing and assignment using the AWGRs of the cell of the AWGR and server-based PON data centre architecture was discussed. In addition, a benchmark study was conducted to compare the power consumption of the AWGR and server-based PON data centre architecture versus a traditional server-centric data centre architecture, BCube. The results revealed that the proposed architecture reduced the power consumption by 80% as compared to a BCube topology of 512 servers, by 82% as compared to a BCube topology of 4096 servers, and by 83% as compared to a BCube architecture of 32768 servers using $\rho = 0.05$. Moreover, the use of different values of ρ resulted in a slight difference in the power savings of the AWGR and server-based PON data centre architecture as compared to those of the BCube data centre.

Chapter 5 Optimal Virtual Machine Placement in PON AWGR and Server-Based Data Centres

5.1 Overview

As discussed in Chapter 2, virtualisation techniques are used to offer better control and improved resource utilisation. In this chapter, we further investigate the AWGR and server-based PON data centre architecture by developing an MILP model along with a heuristic to minimise the power consumption of the AWGR and server-based PON data centre architecture by optimising the embedding of the virtual machine requests in the servers of the AWGR and server-based PON data centre architecture.

5.2 Power optimisation of virtual machine embedding in PON data centre

An MILP model is developed to minimise the power consumption of the AWGR and server-based PON data centre cell by optimising the embedding of virtual machine requests in servers. The sets, parameters, and variables used in this model are as follows:

Sets:

N	Set of all nodes (servers and special servers)
S	Set of servers, where $S \subset N$
SS	Set of special servers, where $SS \subset N$
NB_m	Set of neighbouring nodes of node m in the topology, where $m \in N$

Parameters:

TD_{vw} Traffic demand between VMs v and w, where v , $w \in V$

L Capacity of a physical link

M Sufficiently large number

PSI Idle power consumption of a server

PSM Maximum power consumption of a server

PT Transceiver power consumption

OP Maximum ONU power consumption

CV_v Processing capacity requirement of request v, where $v \in V$

CS_s Maximum processing capacity of server s, where $s \in S$

MV_v Memory requirement of request v, where $v \in V$

MS_s Memory capacity of server s, where $s \in S$

OR Maximum ONU data rate

DR Maximum data rate that server $s \in S$ can support

γ_f Fraction of a special server processing capacity used for forwarding one request

SN Maximum number of servers for embedding a VM request

K Traffic bifurcation degree

Variables:

T_{sd} Traffic demand of server pair (s, d), aggregated from all VMs placed in these servers, where $s, d \in S$

X_{mn}^{sd} Traffic demand of server pair (s,d) passing through link (m,n) in the topology, where $s, d \in S, m \in N$ and $n \in NB_m$

η_{mn}^{sd} Binary equivalent of X_{mn}^{sd}, η_{mn}^{sd} = 1 if there is a traffic demand between node pair (s,d) passing through link (m,n); otherwise, η_{mn}^{sd} = 0, where $s, d \in S, m \in N$, and $n \in NB_m$

TR_r Total traffic forwarded (relayed) by special server r, where $r \in SS$

TS_s Total traffic transmitted by server s, where $s \in S$

NRf_r Number of requests forwarded by special server r , where $r \in SS$

α_s α_s = 1 if server s is activated; otherwise, α_s = 0, where $s \in S$

αR_r αR_r = 1 if special server r is activated; otherwise, αR_r = 0, where $r \in SS$

ω_s^v ω_s^v = 1 if request v is served by server s; otherwise, ω_s^v = 0, where $v \in V$ and $s \in S$

ε_{sd}^{vw} ε_{sd}^{vw} is the ANDing of ω_s^v and ω_d^w, ε_{sd}^{vw} **=** 1 if VMs v and w are embedded in different servers (s and d) ; otherwise, ε_{sd}^{vw} = 0, where $v, w \in V$ *and* $s, d \in S$

VN_s Number of VMs placed in server s, where $s \in S$

PS_s Total power consumed by a server $s \in S$

PF_r Power consumed by the CPU of a special server $r \in SS$ for request forwarding

OP_r Power consumption of an ONU attached to special server $r \in SS$

PR_r Power consumed by the special server $r \in SS$

The total power consumption is composed of the power consumed by the servers hosting VMs and the power consumed by the special servers that route the traffic and the ONUs attached to the special servers. The power consumed by the server (PS_s) is composed of (i) the server's idle power; (ii) the proportional power, which is a function of the CPU utilisation due to embedding VMs; and (iii) the power consumed by the server's transceiver for communication. It is calculated as follows:

$$PS_s = \alpha_s \, PSI + (PSM - PSI) \sum_{v \in V} \omega_s^v \; \frac{CV_v}{CS_s} + \; \alpha_s \; PT \tag{5.1}$$

The power consumed by the special server (PR_r) consists of the special server's idle power; the proportional power, which is a function of the CPU utilisation resulting from forwarding requests; and the power consumed by the ONU for communication. The power consumed by the CPU of a special server (PF_r) for request forwarding is calculated as follows:

$$PF_r = \; NRf_r \;\; \gamma_f \;\; (PSM - PSI) \tag{5.2}$$

The power consumed by an ONU (OP_r) for communication is assumed to follow a linear profile and is expressed as follows:

$$OP_r = \; \frac{OP}{OR} \;\; TR_r \tag{5.3}$$

PR_r is determined as follows:

$$PR_r = \;\; \alpha R_r \, PSI \; + \; PF_r + \; OP_r \tag{5.4}$$

On the basis of the types of communication mentioned in Section 3.3.3.5 and the VM placement, there will be three levels of traffic:

1. Traffic between VMs: This is the requested traffic demand between VMs and is expressed as TD_{vw}.

2. Traffic between servers T_{sd}: This results from aggregating the traffic between the VMs hosted in these servers.

3. Traffic between servers on different subgroups that needs to be forwarded by special servers TR_r.

The model is defined as follows:

Objective:

Minimise:

$$\sum_{s \in S} PS_s \; + \; \sum_{r \in SS} PR_r \tag{5.5}$$

where equation (5.5) states the model objective, which is to minimise the total power consumed by the servers and the special servers. This is achieved by optimising the servers selected to host VMs.

Subject to:

1) VM constraints:

$$\sum_{v \in V} MV_v \quad \omega_s^v \; \leq \; MS_s \tag{5.6}$$

$$\forall \, s \; \in S$$

Constraint (5.6) ensures that the memory requirements of the demands hosted by a server do not exceed the server's memory capacity.

$$\sum_{v \in V} CV_v \quad \omega_s^v \; \leq \; CS_s \tag{5.7}$$

$$\forall \, s \; \in S$$

Constraint (5.7) ensures that the CPU requirements of the demands hosted by a server do not exceed the server's CPU capacity.

$$VN_s \; = \sum_{v \in V} \; \omega_s^v \tag{5.8}$$

$$\forall \, s \; \in S$$

Constraint (5.8) determines the number of VMs embedded in a server.

$$\sum_{s \in S} \omega_s^v = SN \tag{5.9}$$

$$\forall\, v \in V$$

Constraint (5.9) controls the number of servers that can be used to host a VM. In this work, we set SN = 1.

2) Traffic constraints:

$$T_{sd} = \sum_{v \in V} \sum_{\substack{w \in V \\ w \neq v}} TD_{vw} \;\; \varepsilon_{sd}^{vw} \tag{5.10}$$

$$\forall\, s, d \in S\colon\; s \neq d$$

Constraint (5.10) gives the traffic between a pair of servers where ε_{sd}^{vw} is expressed as follows:

$$\varepsilon_{sd}^{vw} = \omega_s^v \;\, \omega_d^w \tag{5.11}$$

$$\forall\, v, w \in V\colon w \neq v, \qquad \forall\, s, d \in S\colon s \neq d$$

Equation (5.11) contains the multiplication of two binary variables, which makes the model nonlinear. In order to maintain the linearity of the model, it is replaced by the following three constraints (5.12) – (5.14).

$$\varepsilon_{sd}^{vw} \leq \omega_s^v \tag{5.12}$$

$$\forall\, v, w \in V\colon w \neq v, \qquad \forall\, s, d \in S\colon s \neq d$$

$$\varepsilon_{sd}^{vw} \leq \omega_d^w \tag{5.13}$$

$$\forall\, v, w \in V\colon w \neq v, \qquad \forall\, s, d \in S\colon s \neq d$$

$$\varepsilon_{sd}^{vw} \geq \omega_s^v + \omega_d^w - 1 \tag{5.14}$$

$$\forall\, v, w \in V: w \neq v, \qquad \forall\, s, d \in S: s \neq d$$

$$\sum_{\substack{n \in NB_m \\ n \neq m}} X_{mn}^{sd} - \sum_{\substack{n \in NB_m \\ n \neq m}} X_{nm}^{sd} = \begin{cases} T_{sd} & if\ m = s \\ -T_{sd} & if\ m = d \\ 0 & Otherwise \end{cases} \tag{5.15}$$

$$\forall\, s, d \in S : s \neq d, \forall\, m \in N$$

Constraint (5.15) represents the flow conservation constraint for the traffic flows in the network following the flow conservation law [95].

$$X_{mn}^{sd} \geq \eta_{\ mn}^{\ sd} \tag{5.16}$$

$$\forall\, s, d \in S: s \neq d, \qquad \forall\, m \in N, \qquad \forall\, n \in NB_m: m \neq n$$

$$X_{mn}^{sd} \leq \eta_{\ mn}^{\ sd} \quad M \tag{5.17}$$

$$\forall\, s, d \in S: s \neq d, \qquad \forall\, m \in N, \qquad \forall\, n \in NB_m: m \neq n$$

Constraints (5.16) and (5.17) relate X_{mn}^{sd} to its binary equivalent $\eta_{\ mn}^{\ sd}$, where M is set to 1000.

$$\sum_{\substack{n \in NB_m \\ m \neq n}} \eta_{\ mn}^{\ sd} \leq K \tag{5.18}$$

$$\forall\, s, d \in S: s \neq d, \qquad \forall\, m \in N$$

Constraint (5.18) specifies the number of routes a traffic demand can be bifurcated through. In this work, we have set $K = 1$, so a single path is used; i.e., traffic bifurcation is not allowed; however, the model is general.

$$TS_s = \sum_{\substack{d \in S \\ s \neq d}} T_{sd} \tag{5.19}$$

$$\forall\, s \in S$$

Constraint (6.19) computes the total traffic transmitted by a server.

$$TS_s \leq DR \tag{5.20}$$

$$\forall\, s \in S$$

Constraint (5.20) ensures that the total traffic of a server is within its data rate.

$$TR_r = \sum_{s \in S} \sum_{\substack{d \in S \\ s \neq d}} \sum_{n \in NB_r} X_{n\,r}^{sd} \tag{5.21}$$

$$\forall\, r \in SS$$

Constraint (5.21) calculates the total traffic forwarded by a special server.

$$NRf_r \;\; \gamma_f \;\; \leq \;\; 1 \tag{5.22}$$

$$\forall\, r \in SS$$

Constraint (5.22) guarantees that the special server's processing used for request forwarding does not exceed the special server's capacity.

$$NRf_r = \sum_{s \in S} \sum_{\substack{d \in S \\ s \neq d}} \sum_{n \in NB_r} \eta_{\;n\,r}^{\;sd} \tag{5.23}$$

$$\forall\, r \in SS$$

Constraint (5.23) computes the number of requests forwarded by a special server.

$$\sum_{s \in S} \sum_{\substack{d \in S \\ s \neq d}} X_{mn}^{sd} \;\; \leq \;\; L \tag{5.24}$$

$$\forall\, m \; \in N \,, \forall\, n \; \in \; NB_m$$

Constraint (5.24) ensures that the total traffic passing through a link does not exceed the capacity of the link.

$$TR_r \; \leq \; OR \tag{5.25}$$

$$\forall\, r \; \in SS$$

Constraint (5.25) ensures that the traffic forwarded by a special server does not exceed its ONU data rate.

3) Server constraints:

$$\alpha_s \leq VN_s \quad \forall s \in S \tag{5.26}$$

$$M\,\alpha_s \geq VN_s \quad \forall s \in S \tag{5.27}$$

Constraints (5.26) and (5.27) are used to relate the binary variable α_s to the non-binary variable VN_s.

4) Special server constraints:

$$\alpha R_r \leq NRf_r \quad \forall r \in SS \tag{5.28}$$

$$M\,\alpha R_r \geq NRf_r \quad \forall r \in SS \tag{5.29}$$

Constraints (5.28) and (5.29) are used to relate the binary variable αR_r to the non-binary variable NRf_r.

5.3 Results and discussions

In this subsection, the network settings, parameter values, and the results of the MILP model are presented and discussed in detail. The network considered for this model is composed of 16 servers, where the intra-group and the inter-group communication between these servers are covered by this reduced architecture. Here, we focus on the traffic within the cell, East–West traffic, which is typically 76% of the total traffic [102]. The extension to the North–South traffic is a straightforward extension. The server's data rate is 1Gbps, while the special servers provide a data rate of up to 10Gbps. A number of VM requests (5, 10, 15 and 20) are considered, and each VM can communicate with 1–3 other VMs. The processing, memory, and traffic demands for VMs are randomly and uniformly distributed. Table 5.1 shows the input parameters used in this model. Two scenarios are studied; (i) the first investigates the effect of the number of servers in a subgroup on the power consumption of the AWGR and server-based PON data centre architecture; (ii) the second compares the AWGR and server-based PON data centre architecture optimised model to a non-power-optimised model (whose goal is to serve all the demands) for embedding VMs.

Table 5.1: Input data for the model

Parameter	Value
Traffic demand between VMs (TD_{vm})	200–500 Mb/s, random and uniformly distributed
Capacity of physical link (L)	10 Gbps
Idle power consumption of a server or special server (PI)	301.6 W [98]
Maximum power consumption of a server or a special server (PM)	457 W [96]
Processing capacity requested by a client in CPU cycles (CV_v)	500–3000 MHz, random and uniformly distributed
Maximum processing capacity of server (CS_s)	3.9 GHz [103]
Portion of a server or special server's processing capacity used for forwarding one request (γ_f)	5%
Total ONU power consumption	2.5 W [97]
ONU data rate	10 Gbps
VM request requirements of RAM	500–3000 MB, random and uniformly distributed
Memory capacity (RAM) of server	50 GB [103]
Server’s data rate	1 Gbps

5.3.1 Effect of number of servers per subgroup on power consumption

First, we will examine the effect of the number of servers per subgroup on the total power consumption, the number of activated special servers, the special servers' utilisation, and the ONU's utilisation. We consider placing the 16 servers of the studied architecture into 2, 4, and 8 subgroups, as depicted in Figures 5.1–5.3.

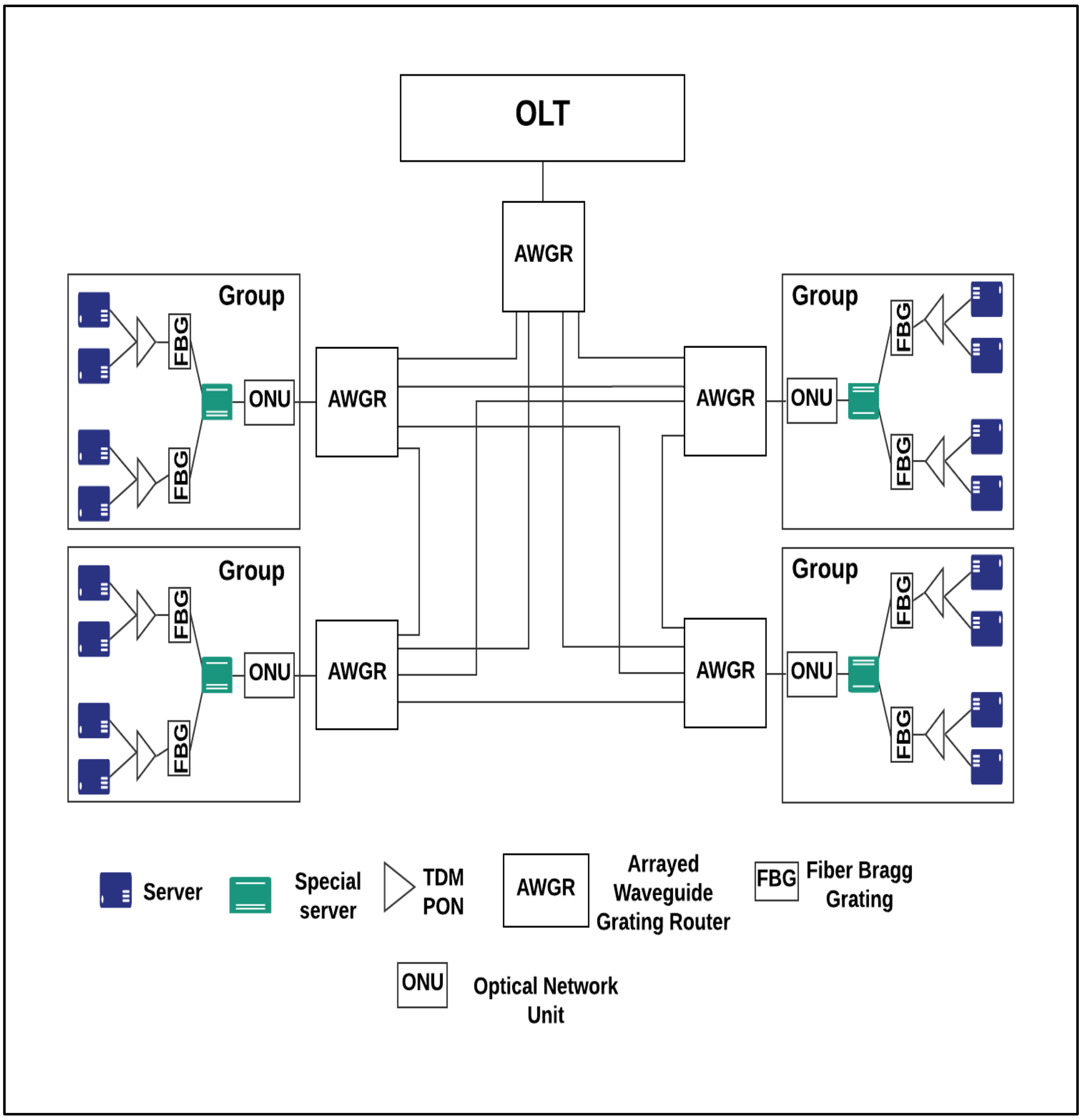


Figure 5.1: The AWGR and server-based PON data centre architecture with two servers in each subgroup

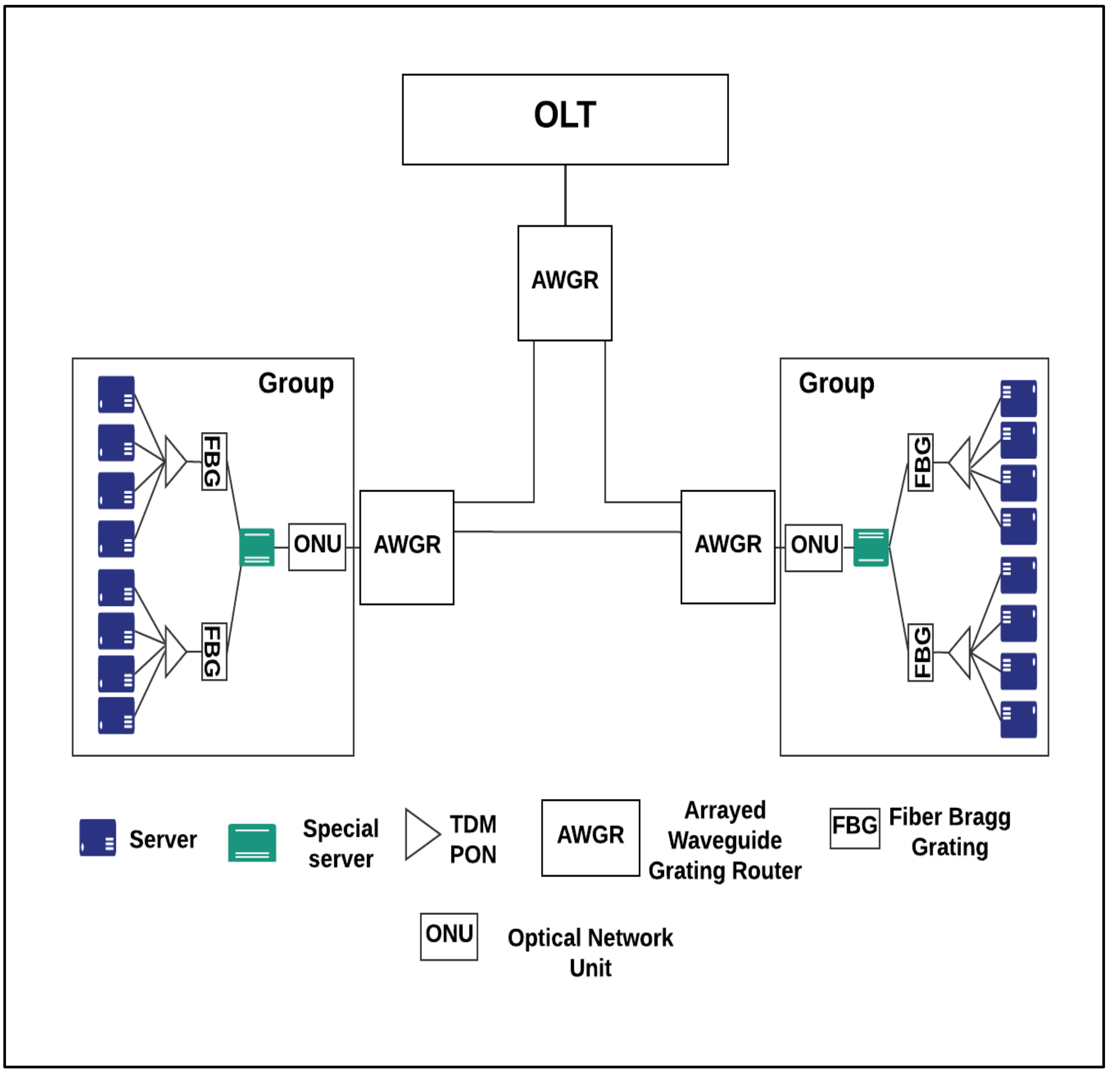


Figure 5.2: The AWGR and server-based PON data centre architecture with four servers in each subgroup

The power consumption associated with hosting VMs in the AWGR and server-based PON data centre architecture is presented in Figure 5.4. As shown in this figure, the power consumption of placing VMs is inversely proportional to the number of servers in each subgroup. The idea behind this is that whenever VMs having traffic between them are placed in one subgroup, there will be no need for activating the special servers connected to the subgroup. This is more likely to happen when the number of servers in a subgroup is large. Moreover, even if some VMs are distributed over more than one subgroup, the model tries to allocate the VMs with a high traffic in the same subgroup as far as possible. Accordingly, a small amount of traffic traverses the special servers connecting these VMs.

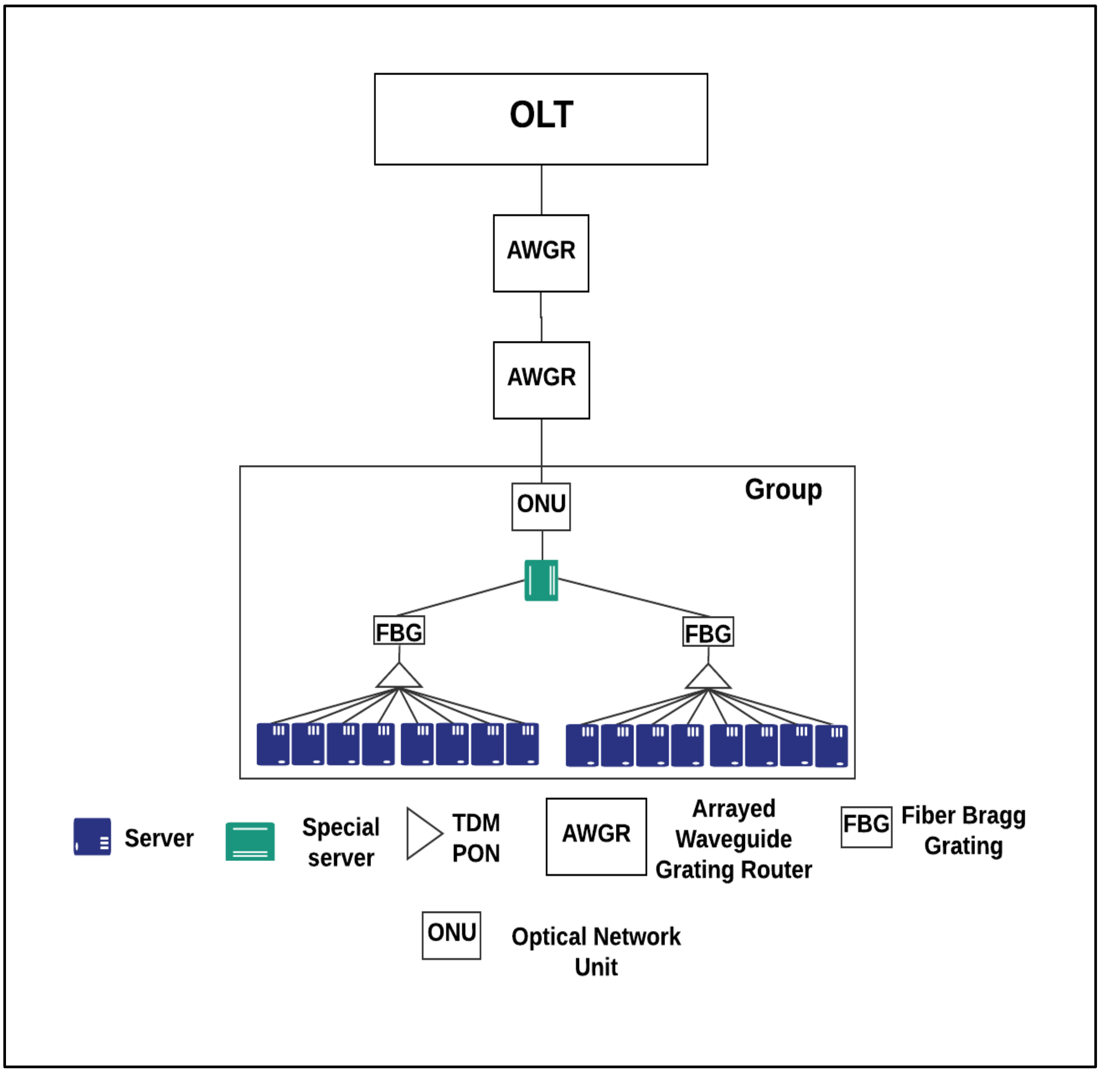


Figure 5.3: The AWGR and server-based PON data centre architecture with eight servers in each subgroup

This is important because the amount of power consumed by special servers is affected by the amount of traffic traversing them. In other words, with a small number of servers in a subgroup, the communicating VMs will need to be distributed over more than one subgroup, which activates the special servers and consumes more power. Note that the increase in the number of servers in a subgroup is limited by the splitting ratio of the TDM

PON connected to it. Moreover, the oversubscription rate and the server's share of resources are affected by the number of servers in the subgroups.

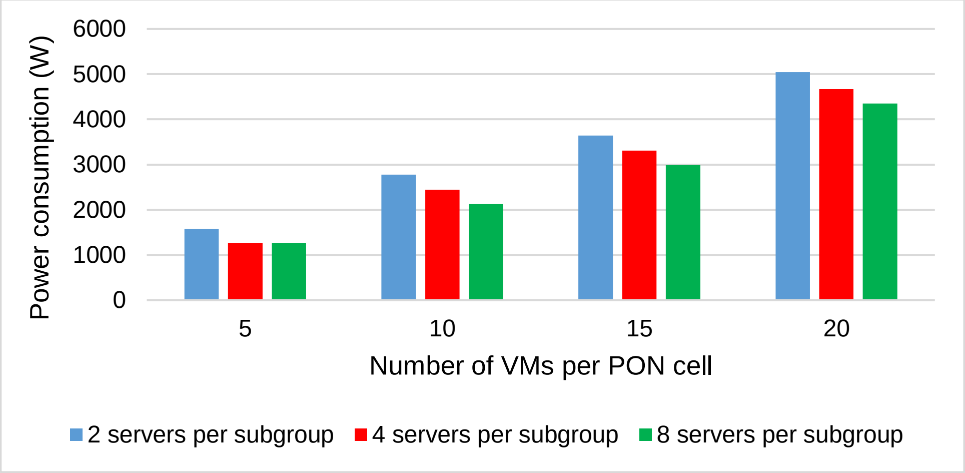


Figure 5.4: VM power consumption considering different number of servers per subgroup; with 16 servers in total per PON cell

To get a detailed view of the impact of changing the number of servers in each subgroup on each term of the objective function and, hence, the total power consumption, their values are presented in Figures 5.5–5.8. It is clear from Figure 5.5 that the number of activated servers for the VM placement is affected by the number of VMs allocated irrespective of the number of servers in each subgroup. On the contrary, the number of activated 'special servers' (Figure 5.6) is inversely proportional to the number of servers in each subgroup and directly proportional to the number of VMs placed.

The special server's CPU utilisation is affected by the number of requests traversing through it irrespective of their traffic (Figure 5.7). Reducing the number of servers in each subgroup increases the likelihood of having more requests that need to pass through the special servers. Having a small number of servers per subgroup along with a large number of VMs increases the special server's CPU utilisation, as shown in Figure 5.7.

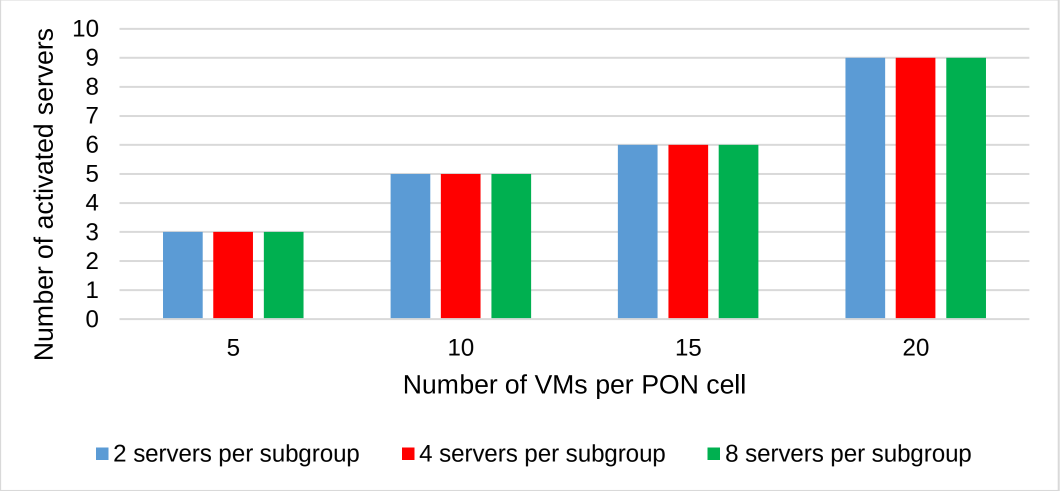


Figure 5.5: Number of activated servers (α_s) based on the number of servers in each subgroup, with 16 servers in total per PON cell

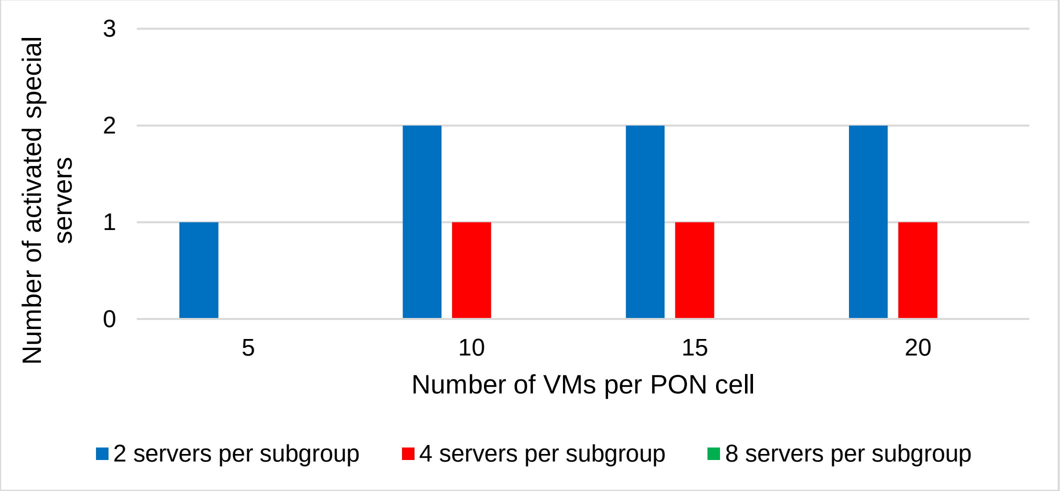


Figure 5.6: Number of activated special Servers (αR_r) based on the number of servers in each subgroup

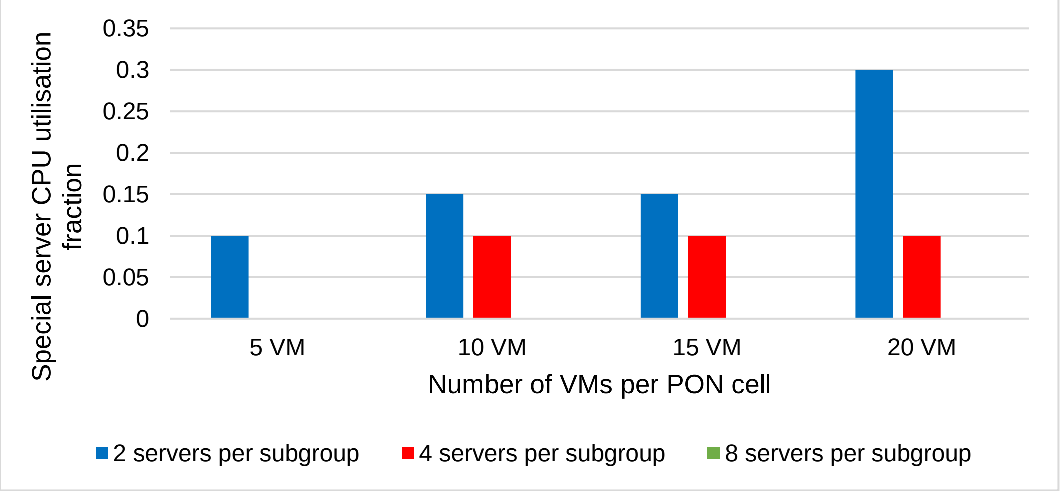


Figure 5.7: Special servers CPU utilisation $(NRF_r \; \gamma_f)$ based on the number of servers in each subgroup, with 16 servers in total per PON cell

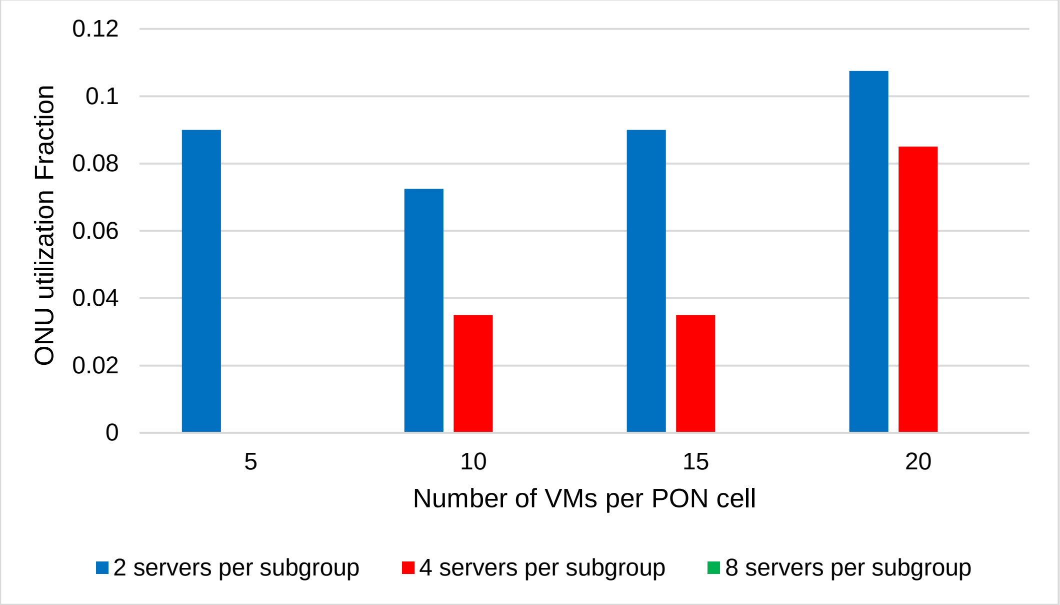


Figure 5.8: ONU utilisation (TR_r/OR) based on the number of servers in each subgroup, with 16 servers in total per PON cell

Regarding Figure 5.8, note that in general, increasing the number of VMs means that more communication may be needed between the VMs in the different subgroups passing through the special server (gateway). Moreover, with few servers per subgroup, the use of the special sever increases. More importantly, minimising the total relaying power consumption is about minimising the sum of the special server's power consumption and the ONU's power consumption, as shown by equation 5.4.

Regarding the power consumption of the special server, it is dictated by the number of requests relayed through the relay server, as expressed by equation (5.2), where the number of relayed requests (jobs) determines the CPU power consumption. As for the ONU power consumption, it is dictated by the data rate, as shown in equation (5.3).

As the power consumption of the special server dominates that of the ONU, the optimisation minimises the number of requests (most of the time) and not the data rate of the requests relayed. As such, at 5 VMs in Figure 5.8, the data rate of the requests may be high, and hence, the ONU may consume more power than the 10-VM case shown in Figure 5.8. In all the cases, the sum of the power consumption is minimised, but this may sometimes involve a higher ONU power consumption.

5.3.2 Optimised and non-optimised VM embedding

Here, the number of servers in each subgroup is set to 4. The power consumption resulting from the optimised VMs embedding is compared to that resulting from non-optimised VM embedding. The non-optimised VM embedding model tries to fulfil all the requests using the following cost function whose objective is to maximise the number of VM requests served:

$$\sum_{v \in V} \sum_{s \in S} \omega_s^v \tag{5.30}$$

Figure 5.9 shows that the optimised VM embedding reduces the power consumption by 34%, 21%, 34%, and 18% compared to the non-optimised VM embedding of 5, 10, 15, and 20 VMs, respectively. It is worthy to mention that the non-optimised VM embedding maximises the number of VM requests served regardless of the hosting servers' number or location in the network. This behaviour explains the unstable variation in their power consumption while increasing the number of VMs.

The main reasons for the power savings are the minimised number of servers and the use of the special server for hosting VMs and providing communication, as shown in Figures 5.10 and 5.11. Furthermore, minimising the number of requests traversing the special servers reduces their CPU utilisation, as shown in Figure 5.12. Moreover, minimising the traffic passing through the special servers reduces their operational power and ONU power consumption, as depicted in Figure 5.13. This is done by allocating VMs that communicate with each other considering the following priority order: same server, subgroup, or group as much as possible.

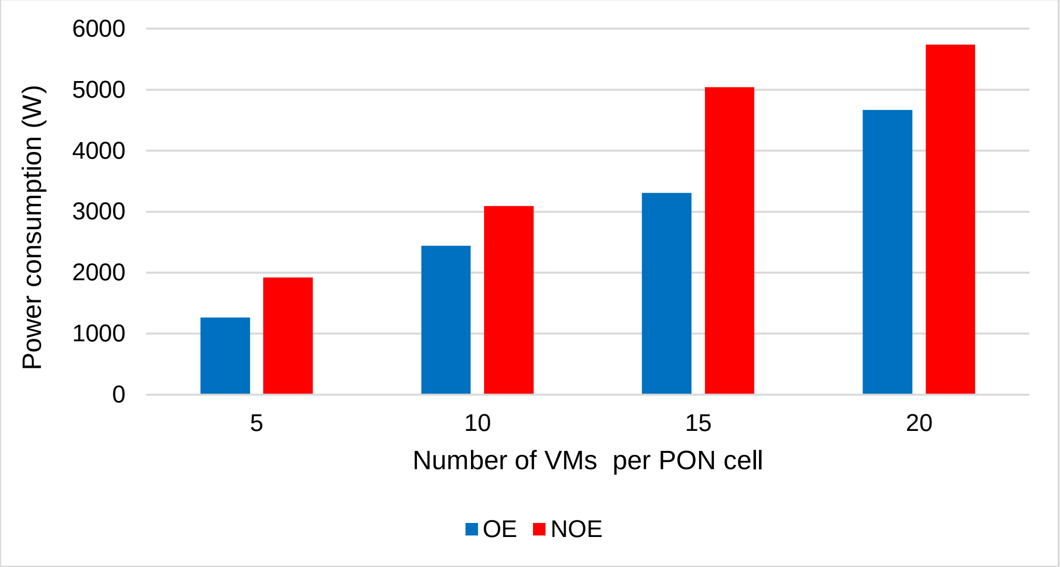


Figure 5.9: VM power consumption for optimised embedding and non-optimised embedding, with 16 servers in total per PON cell

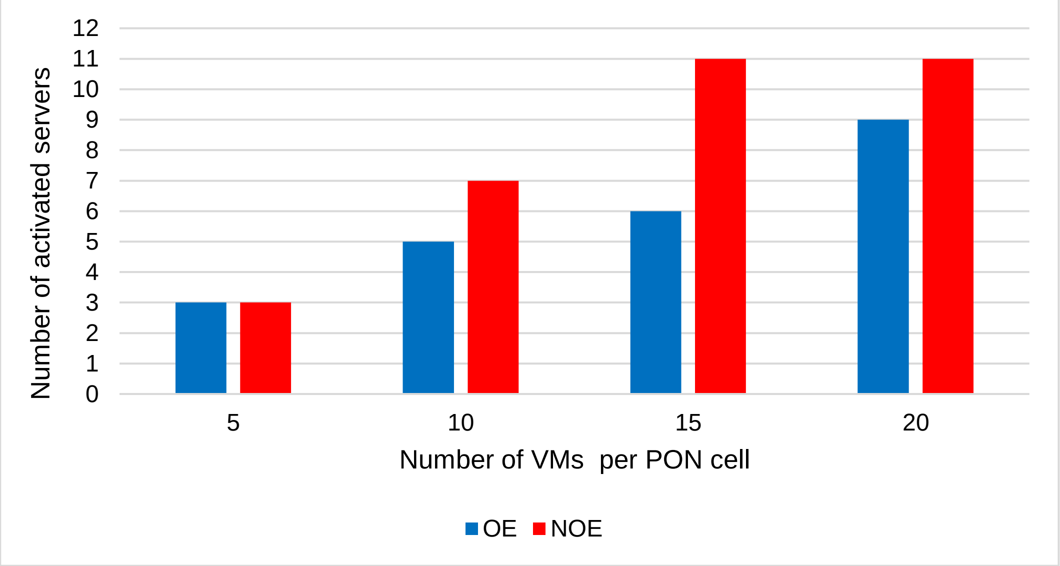


Figure 5.10: Number of activated servers (α_s) for optimised embedding and non-optimised embedding, with 16 servers in total per PON cell

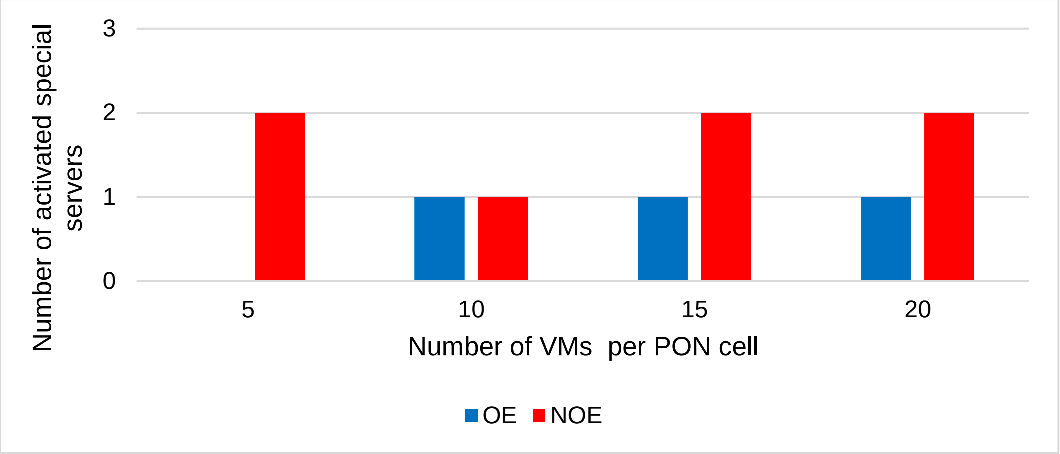


Figure 5.11: Number of activated special servers (αR_r) for optimised embedding and non-optimised VM embedding, with 16 servers in total per PON cell

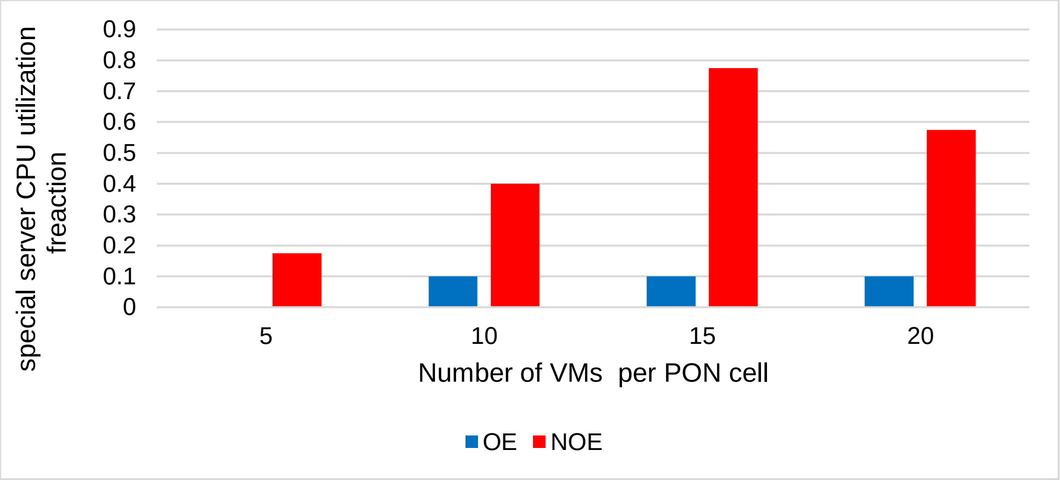


Figure 5.12: Special servers’ CPU utilisation ($NRF_r.\ \gamma_f$) for optimised embedding and non-optimised embedding, with 16 servers in total per PON cell

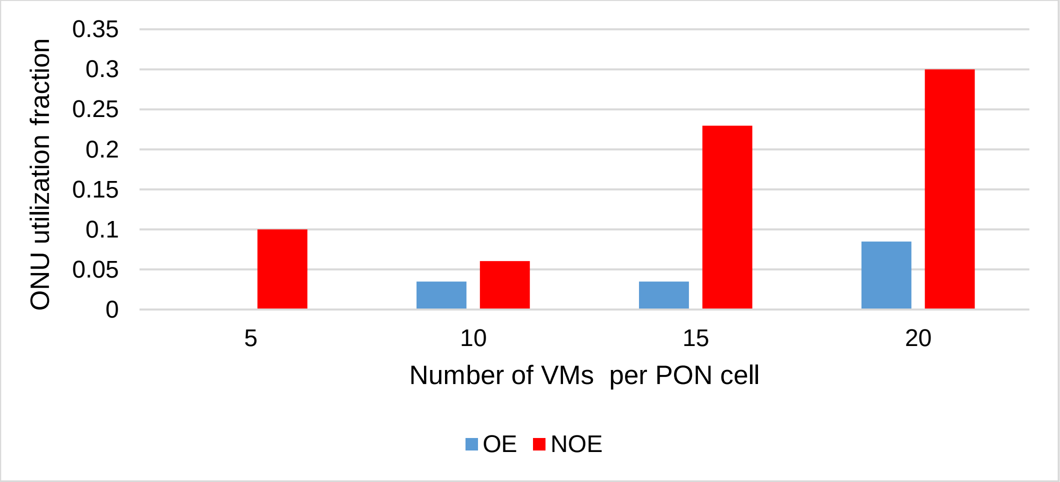


Figure 5.13: ONU utilisation (TR_r/OR) for optimised embedding and non-optimised embedding, with 16 servers in total per PON cell

5.4 VM placement heuristic

A heuristic is developed to verify and validate the VM placement MILP model. Using the heuristic, results for larger problems (networks having a large number of servers and a large number of VMs) can be obtained. The flowchart presented in Figure 5.14 shows the steps that have been adopted in this heuristic.

The inputs to the heuristic are the data centre network topology; the servers processing (CS_s), memory (MS_s), and link capacity (L); and the special servers processing (CR_r) and memory (MR_r) capacities. Furthermore, the VM requirements of processing capacity CV_v , memory capacity MV_v, and traffic TD_{vw} are provided to the heuristic.

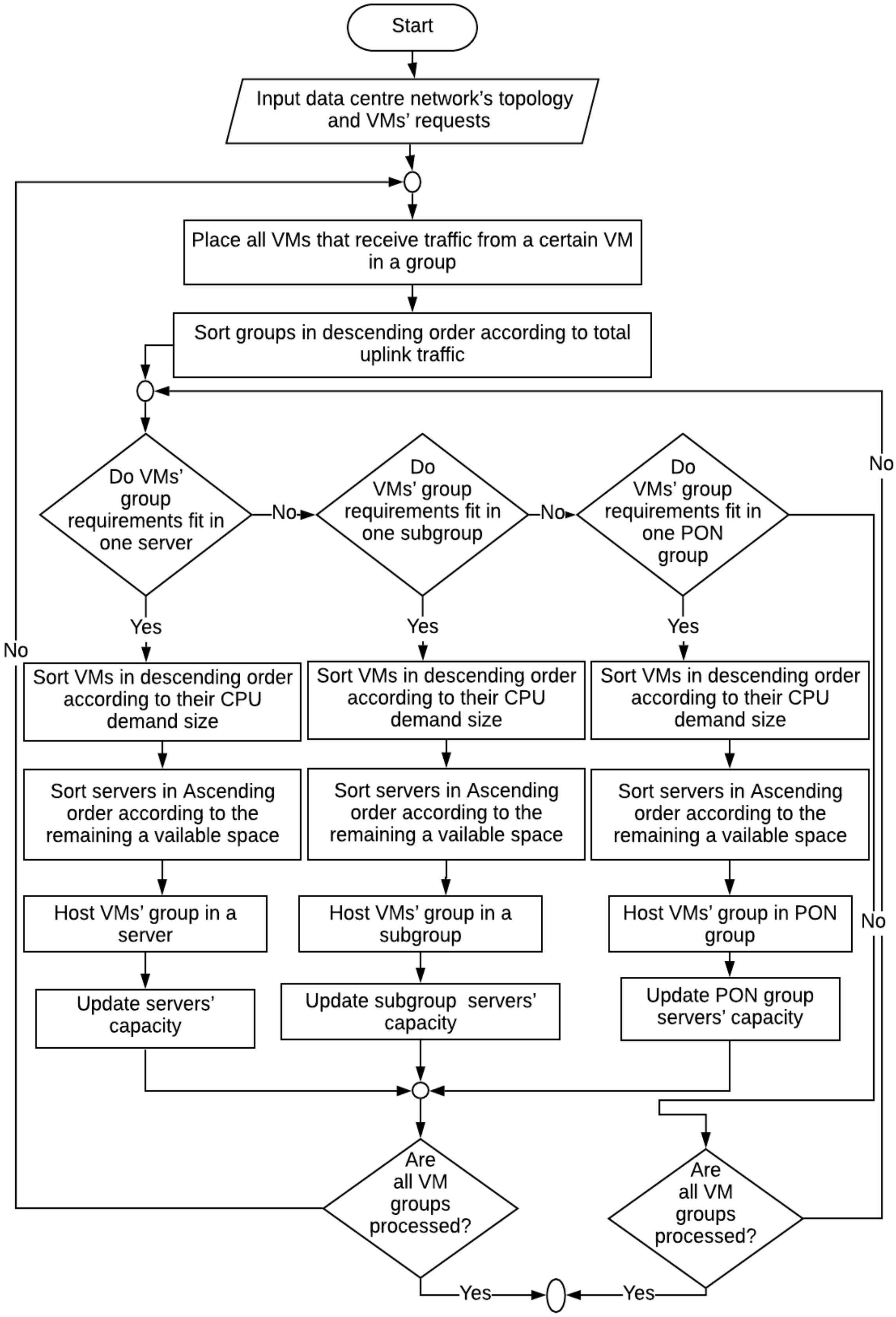
The processing of these input data begins with creating groups by placing all the VMs that receive traffic from a certain VM in a group; i.e., the number of groups is equal to the number of VMs sending traffic and each VM receiving traffic can exist in multiple groups depending on the number of VMs that it receives traffic from. This grouping process is essential to attempt hosting VM with inter-VM traffic in the same server if possible, if not, then the VMs should be hosted in the same subgroup, or finally in the same group if the former two options are not possible. This reduces the power consumed by the servers and the special servers. The groups are sorted in descending order according to the total traffic aggregated from each VM group to other VMs.

The heuristic then arranges the VMs in the group according to their CPU demand size and arranges the servers in the subgroup or the group according to the remaining available space. It then picks the largest VM and tries to fit it in the smallest remaining available space that can accommodate it. The heuristic does this by checking whether the VM's group requirements of memory, CPU, and uplink data rate can fit in any of the servers with available CPU and memory capacity. If a server that can host the VM group is found, the server's capacity is updated and the VMs of the group are removed from the other VM groups. Then, the servers are ordered in an ascending order on the basis of their remaining capacity.

If the VM group requirements of memory, CPU, and uplink do not fit into a server, the heuristic checks whether the VM group requirements can fit into one subgroup of servers. If a subgroup with sufficient capacity is found, then the VM group is placed in it and the subgroup servers' capacity is updated and the VMs are removed from the other VM groups.

If the VM group requirements of memory, CPU, and uplink do not fit into a server subgroup, the heuristic checks whether the VM group requirements can fit into a group of servers. If a group of servers with sufficient capacity is found, then the VM group is placed into it and the group servers' capacity is updated and the VMs are removed from the other VM groups.

If the VM group does not fit into any group of servers, the heuristic moves to place the next VM group. After attempting to place all the VM groups, the heuristic places any remaining VMs individually, i.e. not as groups. This is done by ordering the remaining VMs according to their CPU demand size and then trying to fit the largest first into the smallest remaining space in a server. Finally, the total power consumption of placing VMs is calculated.


Start
Input data centre network's topology and VMs' requests
Place all VMs that receive traffic from a certain VM in a group
Sort groups in descending order according to total uplink traffic
Do VMs' group requirements fit in one server
No
Do VMs' group requirements fit in one subgroup
No
Do VMs' group requirements fit in one PON group
No
Yes
Yes
Yes
No
Sort VMs in descending order according to their CPU demand size
Sort VMs in descending order according to their CPU demand size
Sort VMs in descending order according to their CPU demand size
Sort servers in Ascending order according to the remaining a vailable space
Sort servers in Ascending order according to the remaining a vailable space
Sort servers in Ascending order according to the remaining a vailable space
Host VMs' group in a server
Host VMs' group in a subgroup
Host VMs' group in PON group
No
Update servers' capacity
Update subgroup servers' capacity
Update PON group servers' capacity
Are all VM groups processed?
Are all VM groups processed?
Yes
Yes

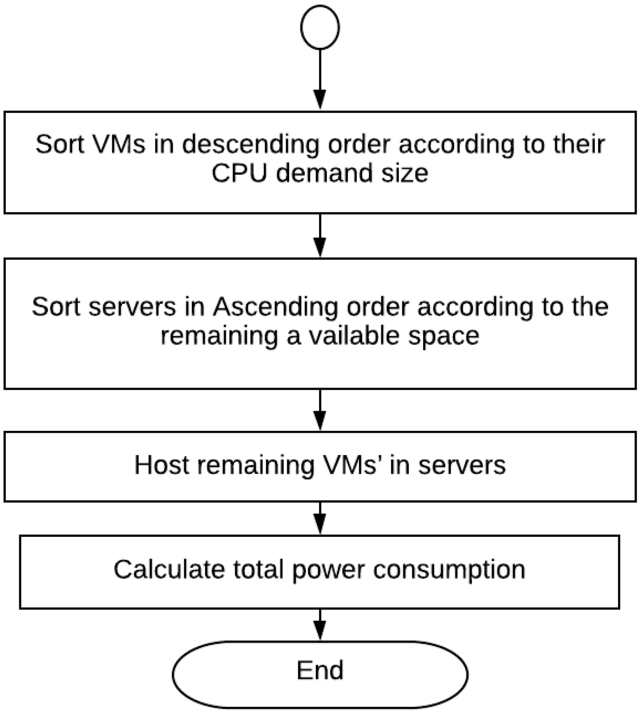

Sort VMs in descending order according to their CPU demand size
Sort servers in Ascending order according to the remaining a vailable space
Host remaining VMs' in servers
Calculate total power consumption
End

Figure 5.14: Flowchart of the heuristic for VM placement in the AWGR and server-based PON data centre architecture

Figures 5.15 shows the results obtained from the heuristic for different numbers of VMs (10, 20, 40, 60, 80, and 100). The heuristic results are compared to the MILP results in terms of the power consumption for 5, 10, 15, and 20 VMs for verification. As the heuristic results provide acceptable agreement with the MILP results (the gap in performance is limited to 20% maximum). Note that the difference between the heuristic and MILP results is due to the sequential nature of the heuristic which takes sequential decisions that cannot be reversed, while the MILP carries out a global optimisation and hence provides the optimum results. The heuristic is used to study the power consumption of the AWGR and server-based PON data centre architecture, as the network size and VM number expand. The expanded network is composed of 32 servers placed into four groups each of eight servers, while the numbers of VMs used are 40, 60, 80, and 100.

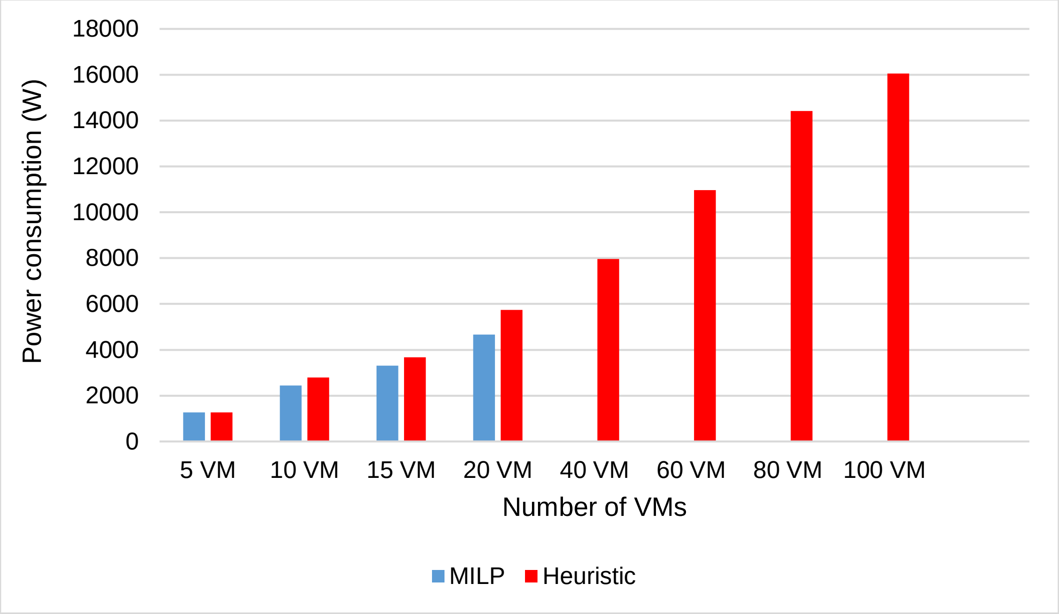


Figure 5.15: Power consumption of placing VMs using the heuristic

5.5 Summary

This chapter investigated the placement of VMs in the AWGR and server-based PON data centre architecture. We optimised the power consumption of virtual machines allocation by using an MILP model and presented a range of results. The results showed that the power consumption was affected by the number of servers in each subgroup because of the change in the number of activated special servers. Our study showed that the proposed model reduced the power consumption by 34%, 21%, 34%, and 18% compared to the non-optimised embedding model of 5, 10, 15, and 20 VMs, respectively. In addition, a heuristic was developed to place the VM requests in the AWGR and server-based PON data centre architecture taking into account the power consumption for the expanded network size, VM number and VM requirements.

Chapter 6 Resilient AWGR and server-based PON data centre architecture

6.1 Overview

In this chapter, we investigate the resilience of the AWGR and server-based PON data centre architecture against different link failures scenarios and propose modified designs for improved resilience. A MILP model is developed to optimise traffic routing under different failure scenarios. The performance is evaluated in terms of delay and power consumption.

6.2 Resilience approach

Resilience in general represents the ability of a system to function during and after a disruption. Many disciplines such as ecology, health, and engineering use the concept of resilience yet in different contexts to evaluate and enhance their systems under disruption [104, 105].

For engineering, namely communication networks, resilience stands for the ability to deliver a service of sufficient quality under failure circumstances [106]. The main sources of failures in communication networks are link cuts, wear out, overload, malicious attacks and environmental disasters [107, 108]. To achieve enhanced network resilience, different techniques are used [30, 107, 109]. One of these techniques is prevention where measures are introduced to stop failures from happening by placing the network components in secured locations and also by providing backup power supplies. Traffic restoration techniques can be used to improve network resilience by rerouting demands and reducing the effect of a failure. These techniques are used to maximise the proportion of traffic carried under single or multiple failures.

In addition, an important performance metric regarding network resilience is availability which is the likelihood that a network component is available when needed. It is calculated as follows [110]:

$$Availability = \frac{MTTF}{MTTF + MTTR} \qquad (6.1)$$

where $MTTF$ represents mean time to failure and $MTTR$ is the mean time to repair [107, 110]. This availability metric makes it possible to assess and enhance the network designs in terms of resilience. Availability is suitable for dynamic operation but, since our focus is on long term design, this method is not used here.

Another technique is resilient network design where diversity and redundancy are the focus. For example, connecting the network nodes to multiple network interfaces. This technique is used when the probability of multiple simultaneous link and node failures is very low. Here, redundancy is introduced so that each link and each node has protection and the routes selected are disjoint [107].

In our PON cell, the loss of a link in a 4-group PON cell can cause 75% loss in traffic. In this chapter we consider varying the design of the AWGR and server-based PON data centre to provide resilience, then a MILP model is used to route traffic over these designs.

6.3 Methodology

The PON data centre architectures in Chapter 4 and in Chapter 5 are not resilient and can suffer from single points of failure. These failures can happen in two broad areas; active components and passive components. Regarding passive components such as AWGRs and fibres, they have very high reliability and are housed mostly within the rack and exclusively within the data centre, therefore accidental fibre cuts are not very likely; unlike fibres installed outdoors. Therefore, in this work, no additional protection/redundancy will be provided for AWGRs, but we examine the impact of link disruptions as transceivers can fail and even when fibres are not cut in the data centre, links can thus fail.

As for active components, in the AWGR and server-based PON data centre architecture, they include the special relay servers, the backplane, the optical transceivers in normal servers and the normal servers. We do not consider the failure of normal servers and the failure of transceivers in such servers because failing servers in data centres are very common, and are usually not repaired but left in place until it is time to replace them in the normal replacement cycle every year. Therefore, this work considers the failure of the special servers, backplanes (which contain active components) and fibre link disruptions (due to transceiver failures) in the AWGR and server-based PON data centre architecture. Our approach to these three forms of failure are described below.

To deal with the possible failure of the special server we modified the architecture and added special server redundancy in the form of additional special servers. Regarding backplane failure, here we consider the backplane to provide different levels of connectivity and we study the impact of its failure on the PON data centre architecture in terms of increased power consumption and delay. As for link failures, we consider these fibre link failures in a fashion similar to backplane failure and consider their impact in terms of increased power consumption and increased delay.

6.4 The impact of different intra rack connectivity techniques on the resilience of the AWGR and server-based PON data centre architecture

Different intra-rack physical connections are investigated to find out which one will provide resilient connectivity for the AWGR and server-based PON data centre architecture. The intra-rack connections studied are the ones described in Section 3.3.2.

6.4.1 FBG connectivity

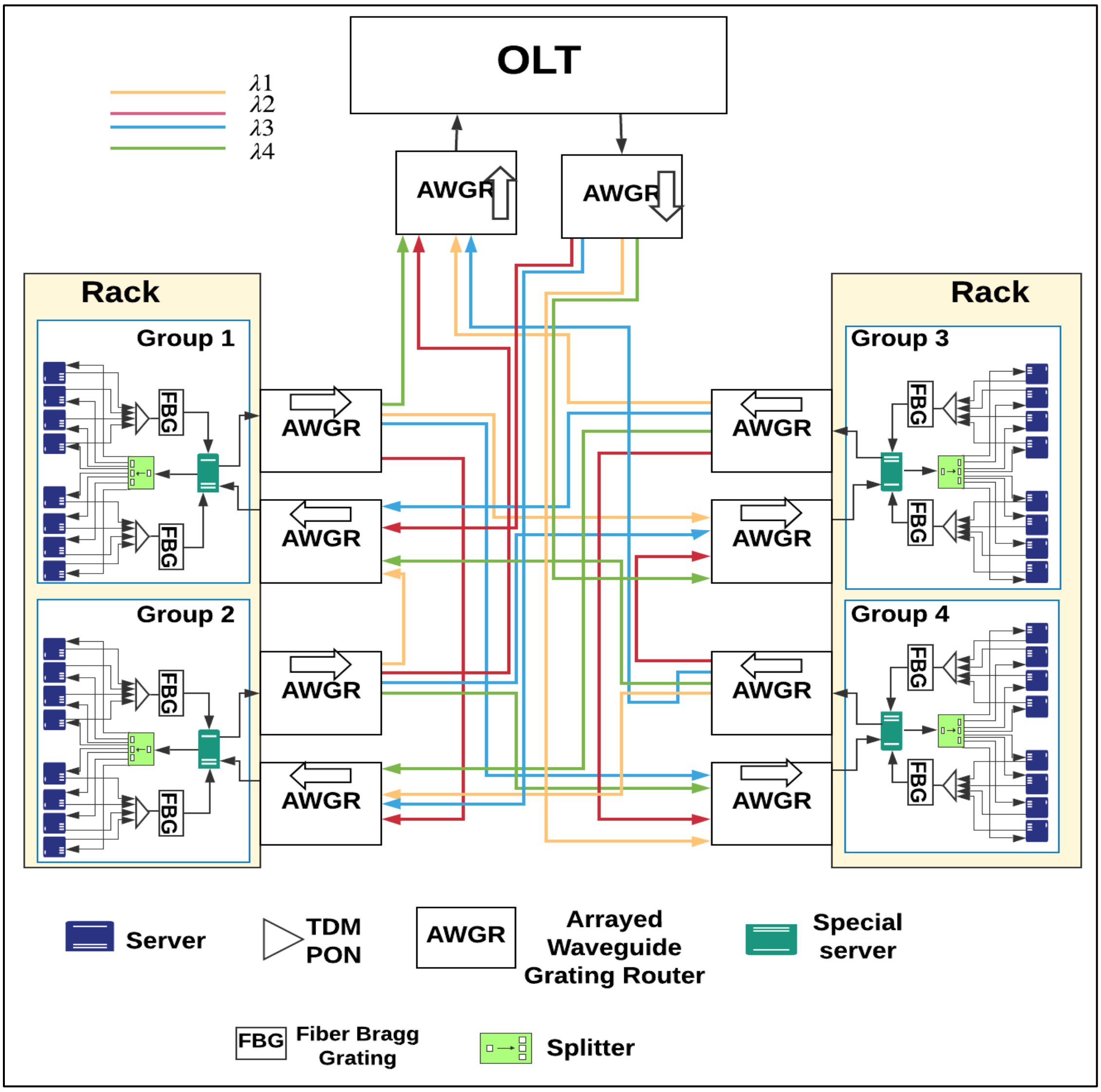


Figure 6.1: FBG connectivity for subgroup in AWGR and server-based PON data centre architecture

Regarding FBG connectivity, servers in each subgroup are connected by an FBG as shown in Figure 6.1. The intra and inter-group traffic use a shared link in which intra subgroup traffic is reflected by the FBG back to the rack while inter-subgroup and inter group traffic continues to the special server. Regarding intra group and inter-group traffic, any fault in wiring or FBG will affect both inter-group and intra-group traffic. This architecture can survive only two out of eight possible types of link failures as illustrated in Table 6.1. Accordingly, for failure scenario 8 (S8) where the failure is between any two AWGRs, traffic will use the other PON groups as a relay to reach its destination. Also, the architecture can survive S5, where the failure is in the link connecting the splitter to the normal server, by using another normal server in the same subgroup as a relay then the traffic continues to the special server and back to its destination.

All other kinds of link failure will affect the inter or intra-subgroup communication or both. For example, if the link connecting a server to TDM PON coupler fails (S2), then there is no other way for that server to send traffic for intra and inter-subgroup servers. In addition, if the link connecting a special server to a splitter fails (S4), then all servers in the related group will not be able to receive inter subgroup / group traffic. Moreover, if the link connecting a special server to AWGR or the opposite fails (S6 and S7), then the inter-group traffic is lost. Also, if the link connecting a TDM PON coupler to FBG fails (S9), then there is no other way for all servers connected to that TDM PON coupler to send traffic for intra and inter-subgroup servers which is also the case when the link connecting FBG to a special server fails (S10).

Table 6.1: Resilience of a connection against link failure

Link failure Intra-rack connection	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
	Backplane link	Server to TDM PON link	Link between TDMPON and special server	Link from special server to Splitter	Splitter link to server	Link from special server to AWGR	Link from AWGR to special server	Link between AWGRs	Link between TDMPON and FBG	Link between FBG and special server
Duplicating the special servers	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	NA	NA
FBG connecting a subgroup	NA	No	NA	No	Yes	No	No	Yes	No	No
Star coupler/backplane connecting a subgroup	Yes	Yes	No	No	Yes	No	No	Yes	NA	NA
Star coupler/backplane connecting a group	Yes	Yes	Yes	No	Yes	No	No	Yes	NA	NA
Star coupler/backplane connecting a rack	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	NA	NA

6.4.2 Reflection Star coupler and passive polymer optical backplane

Architectures using a reflection star coupler or a passive polymer optical backplane to provide connectivity are affected by failures in the same way. Therefore, they are discussed together. In the following, we show different variations of the AWGR and server-based PON data centre architecture where a reflection star coupler or a passive polymer optical backplane is used to connect a subgroup, group, or rack.

6.4.3 Connecting a subgroup

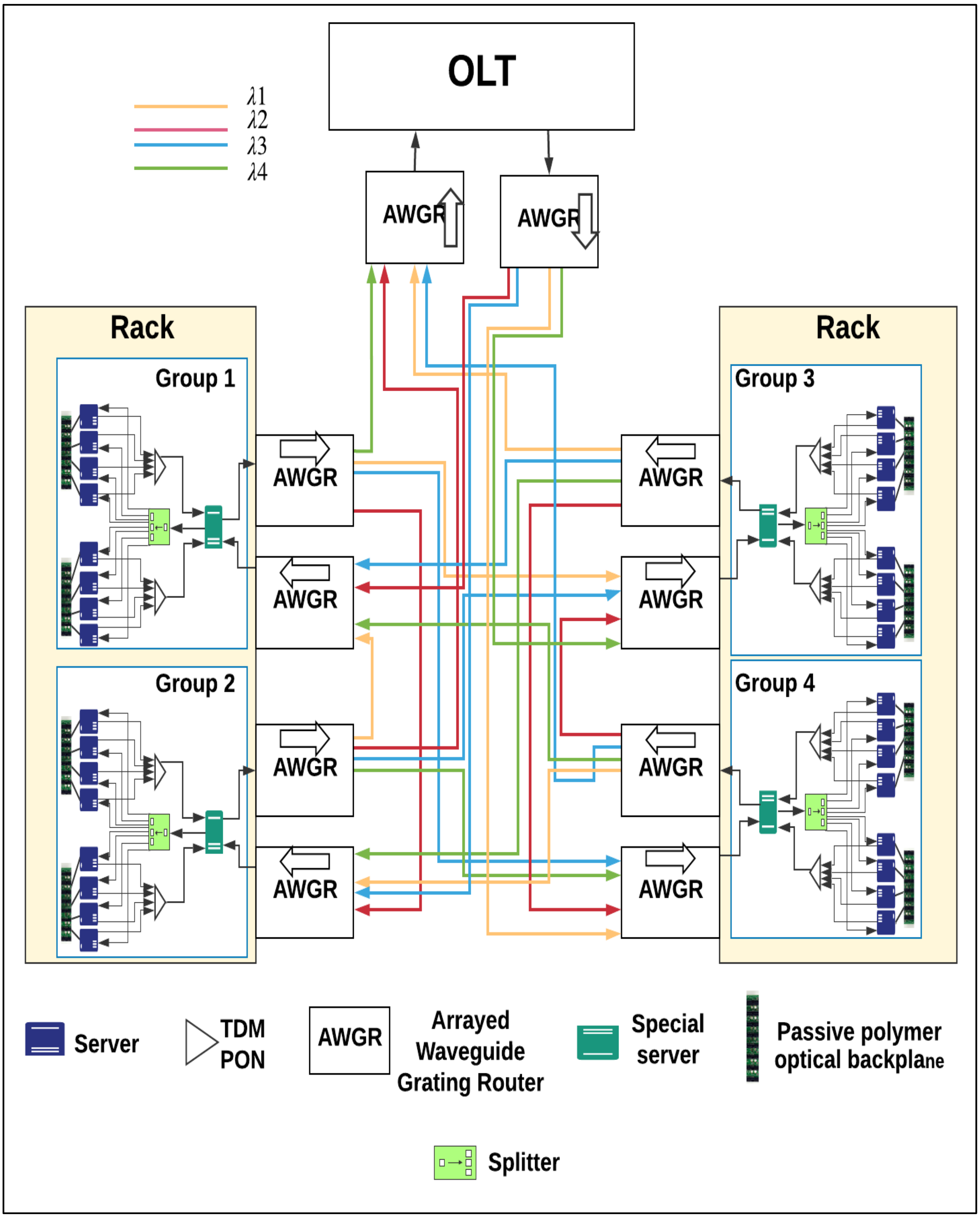


Figure 6.2: Backplane connectivity for a subgroup in AWGR and server-based PON data centre architecture

Here, servers in each subgroup are connected by a star coupler/ backplane as shown in Figure 6.2 (the figure shows connections using a backplane). In this connection, intra-subgroup is separated from inter subgroup traffic where intra-subgroup traffic passes through the backplane and inter-subgroup traffic traverses through the special server. This choice is resilient against only four out of eight possible types of link failures as illustrated in Table 6.1. Accordingly, if the link connecting a server to the backplane fails (S1), the architecture can still allow communication between the affected server and servers in the same subgroup by using multi-hop routing through the special server. Also, if the link connecting the server to TDM PON coupler fails (S2), then the server affected can reach another server in the subgroup through the backplane and can use it as a relay to reach the TDM PON coupler for inter-subgroup communication. As for S5, where the failure is in the link connecting a splitter to a server, the splitter can overcome it by sending the traffic to a server in the same subgroup as a relay, then the traffic will pass through the backplane to reach its destination. Regarding S8, where the failure is between any two AWGRs, the traffic will use the other PON groups as a relay to reach its destination.

It is worth mentioning that all the modified architectures we designed will survive (S1), (S2), (S5) and (S8) in a similar way. These will be discussed in the next sections.

All other kinds of link failure will affect the inter-subgroup communication. For example, if the link connecting a TDM PON to a special server fails (S3), then all servers in the related subgroup will not be able to send inter subgroup traffic. Also, if the link between a special server to splitter fails (S4) then all servers in the related group will not be able to receive inter-subgroup/group traffic. In addition, if the link connecting a special server to AWGR or the opposite fails (S6 and S7), then the inter-group traffic is lost.

6.4.4 Connecting a group

As for this choice, servers in each group are connected by a star coupler/ backplane as shown in Figure 6.3. In this connection, intra-group is separated from inter-group traffic where intra group traffic passes through the backplane and inter-group traffic traverses through the special server. This choice can survive only five out of eight possible types of link failure as illustrated in Table 6.1. Accordingly, S1, S2, S5 and S8 are survived as mentioned in Section 6.4.3. Also, if the link connecting a TDM PON coupler to special server fails (S3), then the server affected can reach another server in the group through the backplane and use it as a relay to reach the other TDM PON coupler in that group for inter-group communication. All other kinds of link failure will affect the inter group communication. For example, if the link connecting a special server to a splitter fails, then all servers in the relevant group will not be able to receive inter-group traffic. Also, if the link connecting a special server to AWGR or the opposite fails (S6 and S7), then all servers in the relevant group will not be able to send or receive inter group traffic.

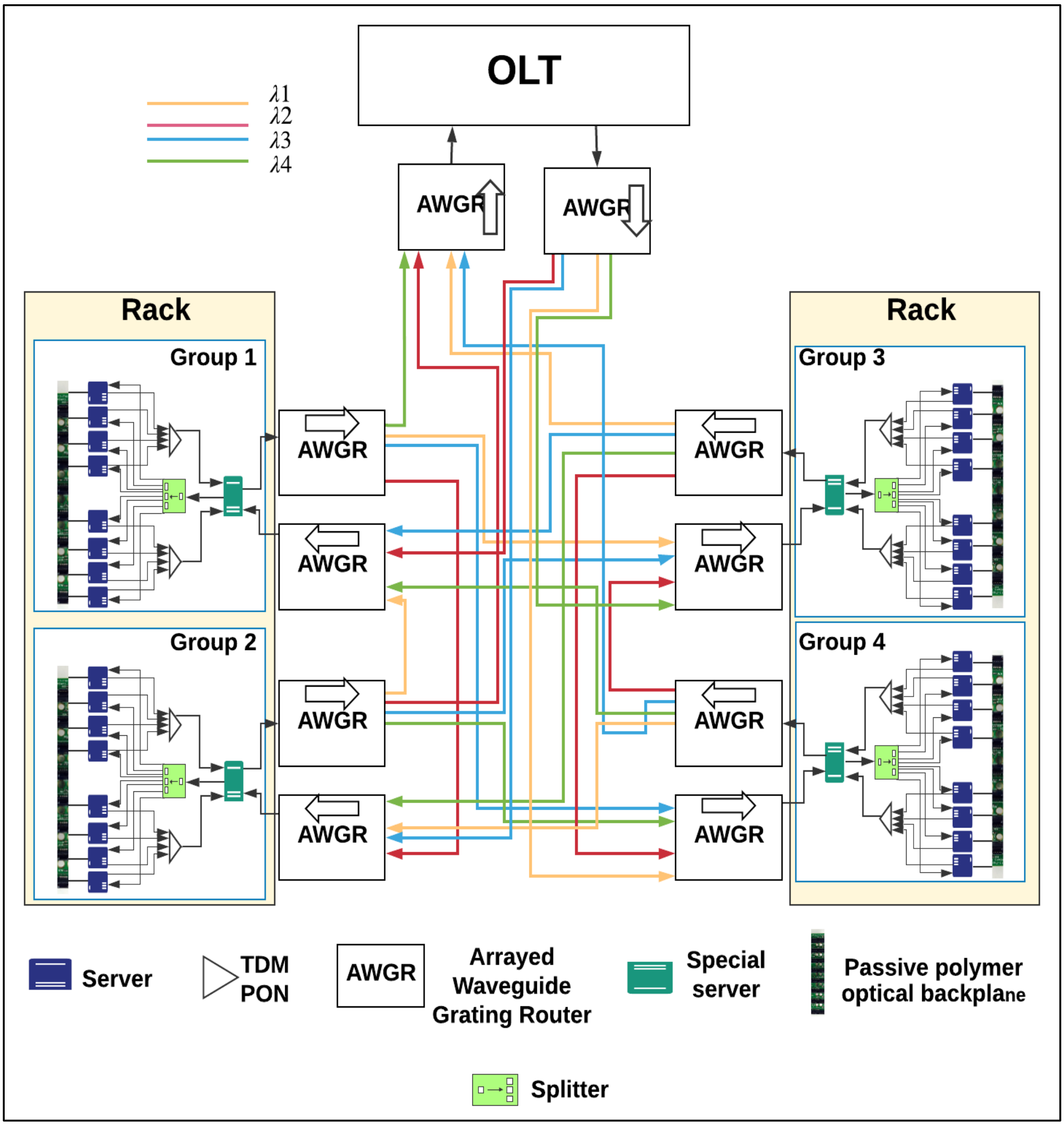


Figure 6.3: Backplane connectivity for a group in AWGR and server-based PON data centre architecture

6.4.5 Connecting a rack

Considering this choice, servers in each rack are connected by a star coupler/ backplane as shown in Figure 6.4. In this connection, intra-rack traffic is separated from inter-rack traffic where intra rack traffic passes through the backplane and inter-rack traffic traverses through the special server. This connection makes AWGR and server-based PON data centre architecture resilient since it survives all the 8 possible types of link failure as illustrated in

Table 6.1. Accordingly, (S1), (S2), (S5) and (S8) are survived as mentioned in Section 6.4.3, whereas (S3) is survived as mentioned in Section 6.4.4.

Furthermore, if a link between a special server and AWGR or the opposite fails (S6 and S7), then the traffic will use the other special server in that rack as a relay to reach its destination. Also, if the failure is in the link connecting a special server to a splitter (S4), the traffic will use the other special server in that rack as a relay to reach its destination.

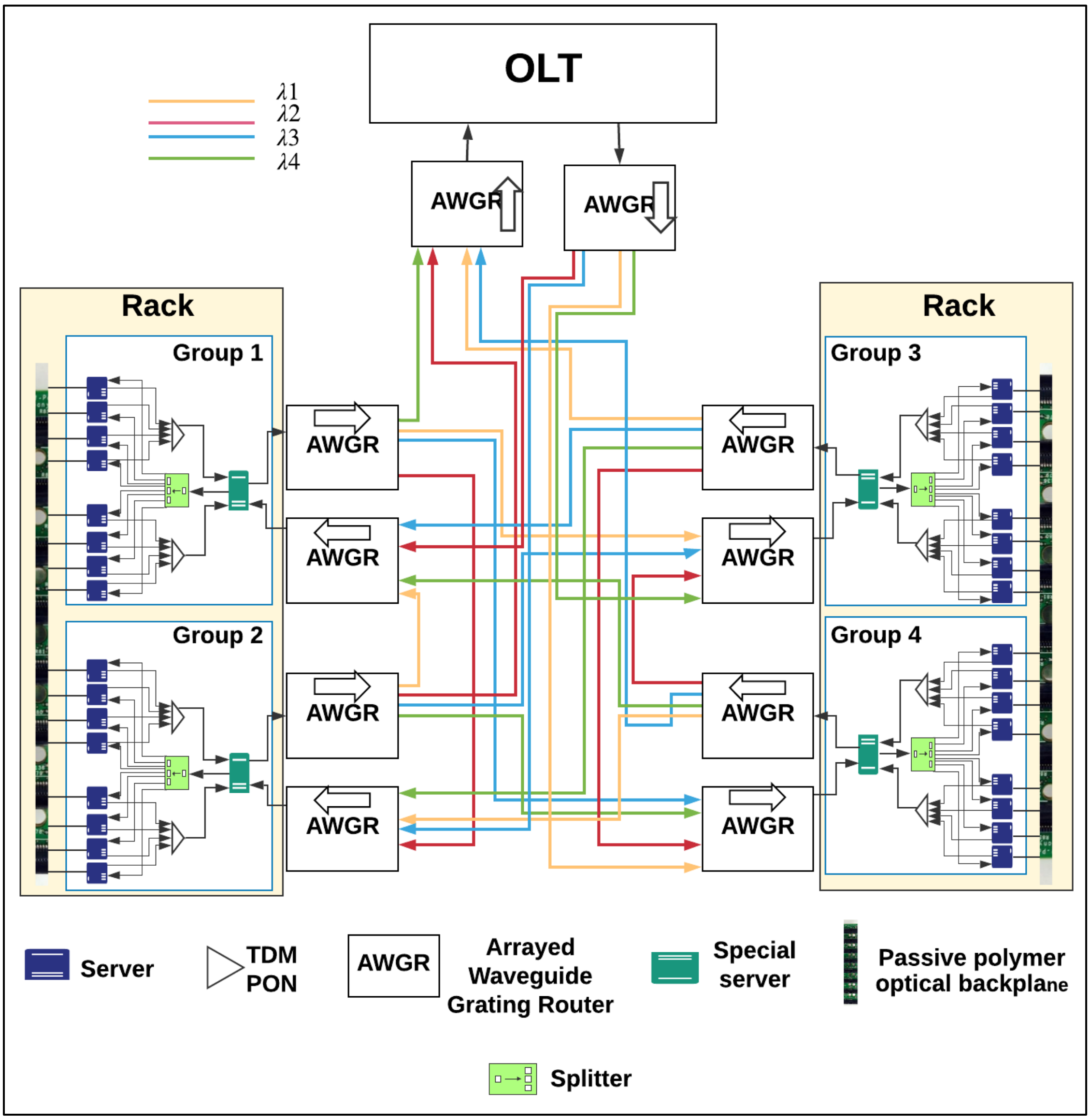


Figure 6.4: Backplane connectivity for a rack in AWGR and server-based PON data centre architecture

6.5 The impact of duplicating the special servers on the resilience of the AWGR and server-based PON data centre architecture

The concern about the design given in Figure 6.4 is that the server’s share of resources decreases when a link failure occurs. For example, when link failure occurs, it may not be possible to reach a special server in a rack. As a result, there will be only one special server to serve all servers in the affected rack instead of two. Based on this concern, we considered duplicating the special servers and the link connecting each special server to the coupler/splitter and AWGR to avoid this limitation. Here, the servers in each subgroup are connected by a backplane and servers in a group are connected to two special servers instead of one. One of these two special servers serves the group in the normal state, and the other is a backup in case the first one fails. However, this option will increase the deployment cost since more special servers will be added to the architecture. Therefore, the choice between duplicating the servers and using a backplane per rack is a compromise between cost and performance since both of them can survive link failures. The architecture with duplicated special servers is depicted in Figure 6.5.

In this connection, the intra-subgroup traffic is separated from inter-subgroup traffic where intra-subgroup traffic passes through the backplane and inter-subgroup traffic traverses through the special server. This choice is resilient against all the 8 possible types of link failure as illustrated in Table 6.1. Accordingly, (S1), (S2), (S5) and (S8) are survived the same way as mentioned in Section 6.4.3. For link failure between TDM PON coupler and special server (S3), the TDM PON coupler will send traffic through the backup special server. When the failure is in the link connecting a special server to AWGR or the opposite (S6 and S7), then the backup special server is used instead to deliver traffic which is also the case if the failure is in the link connecting a special server to a splitter (S4).

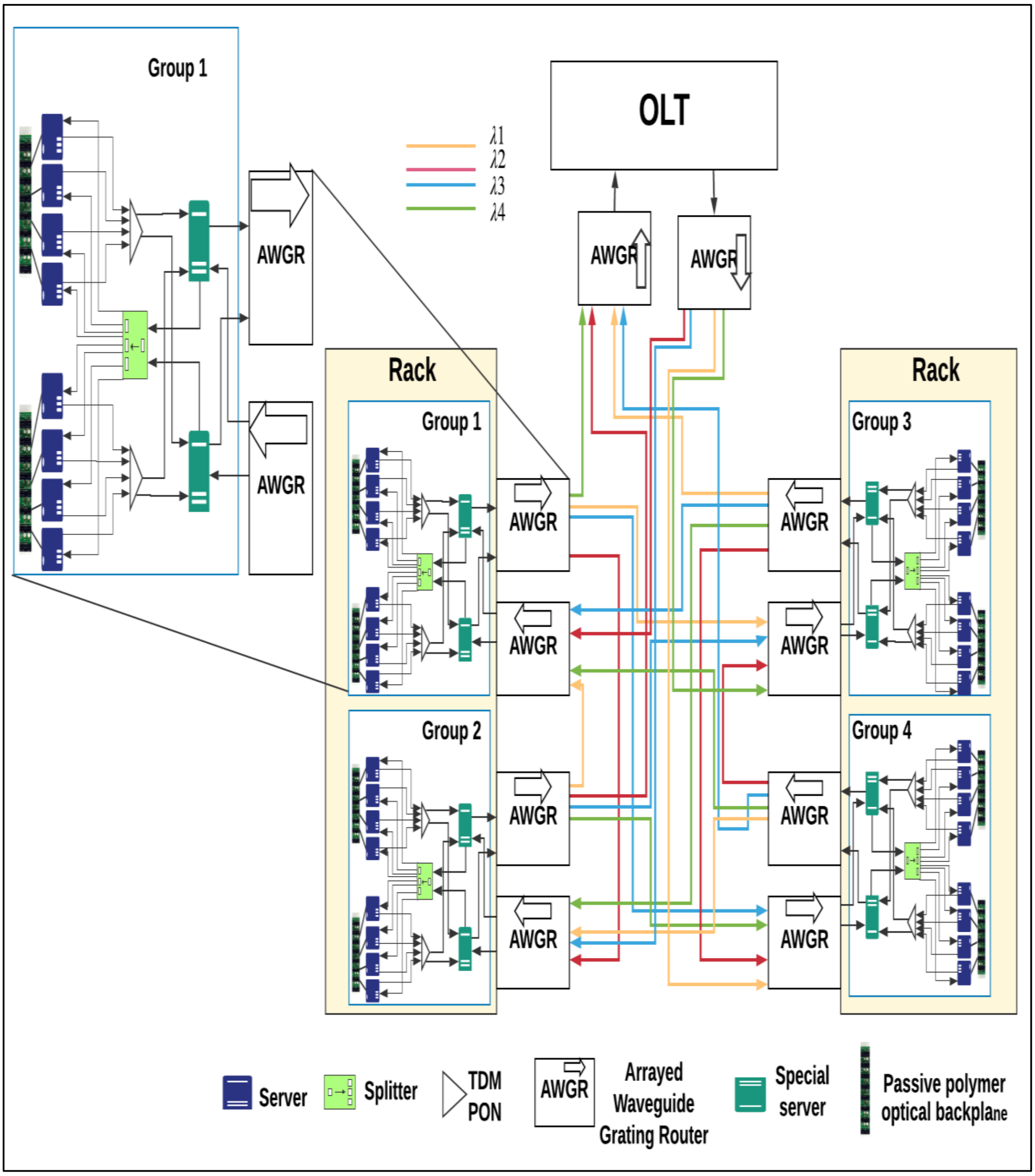


Figure 6.5: Duplicating the special servers in AWGR and server-based PON data centre architecture

6.6 MILP model for evaluating the resilient modified AWGR and server-based PON data centre architectures

Based on the study in the previous sections, connecting the servers in each rack via a star coupler/ backplane or duplicating the special servers are the most resilient options. A MILP model is developed to optimise routing over these designs considering different failure scenarios. The sets, parameters and variables used in this model are:

Sets:

N Set of nodes (servers, special servers, AWGR, splitters, couplers, backplanes and the OLT).

S Set of servers, where $S \subset N$.

SS Set of Special servers, where $SS \subset N$.

OLT Set of OLT's ports, where $OLT \subset N$.

T Set of servers' transceivers connected to the backplane ports, where $B \subset N$.

SP Set of splitters, where $SP \subset N$.

C Set of couplers, where $C \subset N$.

$SOLT$ Set of $S \cup OLT$.

Y Set of traffic demands for servers.

YO Set of traffic demands for OLT ports.

N_m Set of neighbouring nodes of node $m \in N$ that receive from node m.

NB_m Set of neighbouring nodes of node $m \in N$ that transmit to node m.

Parameters:

T_{sd} Traffic demand between servers s and $d \in SOLT$.

L_{mn} Capacity of physical link between m and $n \in N$.

M Large enough number.

PSI Idle power consumption of a server.

PSM Maximum power consumption of a server.

OIP Idle power consumption of an OLT port.

OMP Maximum power consumption of an OLT port.

PT Transceiver power consumption.

OP Maximum ONU power consumption.

DR Maximum data rate that server $s \in S$ can support.

OR Maximum ONU data rate.

$OLTR$ Maximum data rate of an OLT port.

γf Fraction of a server or special server processing capacity used for forwarding one traffic demand.

γs Fraction of a server's processing capacity used for transmitting one traffic demand.

γd Fraction of a server's processing capacity used for receiving one traffic demand.

K Traffic bifurcation degree.

UZ Allowed utilization fraction for server’s CPU.

$THOi$ OLT port forwarding threshold; i.e. upper bound on OLT port capacity dedicated for forwarding.

$THSs$ Server’s forwarding threshold; i.e. upper bound on server capacity dedicated for forwarding.

Q_{λ} Server’s Delay for each traffic load.

QO_{λ} Special server or OLT’s Delay for each traffic load.

σ Fraction of server’s total power used for relaying.

τ Fraction of special server’s total power used for relaying.

Variables:

PS_s Total power consumed by server $s \in S$.

PF_r Total power consumed by the CPU of special server $r \in SS$ for requests forwarding.

PO_i Total power consumed by OLT port $i \in OLT$ for requests forwarding.

PR_r Power consumed by the special server $r \in SS$.

X_{mn}^{sd} Amount of traffic demand between node pair s and $d \in SOLT$ passing through link m and $n \in N$ in the substrate network.

η_{mn}^{sd} Binary equivalent of X_{mn}^{sd} .

TR_r Total traffic forwarded (relayed) by special relay server $r \in SS$.

TS_s Total traffic forwarded and received by node $s \in SOLT$.

Ns_s Number of traffic requests originated and transmitted by server $s \in SOLT$.

Nd_s Number of traffic requests destined to server $s \in SOLT$.

Nf_s Number of traffic requests forwarded by server $s \in SOLT$.

NRf_r Number of traffic requests forwarded by special server $r \in SS$.

α_s α_s = 1 if server s is activated, otherwise, α_s = 0, where $s \in S$.

αO_i αO_i = 1 if OLT’s port i is activated, otherwise, αO_i = 0, where $i \in OLT$.

αR_r αR_r = 1 if special server r is activated, otherwise, αR_r = 0, where $r \in SS$.

αB_b αB_b = 1 if a transceiver b is activated for intra rack (in BPPRD) or intra subgroup (in SSBD) communication, otherwise, αB_b = 0, where $b \in T$.

αT_s αT_s = 1 if a transceiver in server s is activated for inter rack (in BPPRD) or inter subgroup (in SSBD) communication, otherwise, αT_s = 0, where $s \in S$.

Nb_b Total number of active transceivers connecting servers to the backplane, i.e. servers' transceivers used for intra rack (in BPPRD) or intra subgroup (in SSBD) communication, where $b \in T$.

NC_s Total number of requests sent by server s to the couplers, i.e. requests sent for inter rack (in BPPRD) or inter subgroup (in SSBD) communication, where $s \in S$.

NP_s Total number of requests received by server s from the splitters, i.e. requests received from inter rack (in BPPRD) or inter subgroup (in SSBD) communication, where $s \in S$.

NT_s Total number of requests using server's transceiver for sending and receiving inter rack (in BPPRD) or inter subgroup (in SSBD) communication, where $s \in S$.

UO_i Utilization of OLT's port $i \in OLT$.

US_s Utilization of server $\in S$.

UR_r Utilization of special server $r \in SS$.

DS_s Delay for each server $s \in S$.

DR_r Delay for each special server $r \in R$.

DO_i Delay for each OLT port $i \in OLT$.

$I_{s\lambda}$ Traffic load indicator: $I_{s\lambda}$= 1 when the traffic load of server s is equal to λ , otherwise, $I_{s\lambda}$ = 0, where $s \in S$ and $\lambda \in Y$.

$IR_{r\lambda}$ Traffic load indicator: $IR_{r\lambda}$=1 when the traffic load of special server r is equal to λ , otherwise, $IR_{r\lambda}$ = 0, where $r \in SS$ and $\lambda \in Y$.

$IO_{i\lambda}$ Traffic load indicator: $IO_{i\lambda}$= 1 when the traffic load of OLT port i is equal to λ , otherwise, $IO_{i\lambda}$ = 0, where $i \in OLT$ and $\lambda \in Y$.

DS_{xn}^{sd} Delay experienced at server x by demand between node pair s and $d \in SOLT$ passing through link x and n where $x \in S$ and $n \in N_x$ in the substrate network.

DR_{xn}^{sd} Delay experienced at special server x by demand between node pair s and $d \in SOLT$ passing through link x and n where $x \in SS$ and $n \in N_x$ in the substrate network.

DO_{xn}^{sd} Delay experienced at OLT port x by demand between node pair s and $d \in SOLT$ passing through link x and n where $x \in OLT$ and $n \in N_x$ in the substrate network.

D_{sd} Delay for each demand between node pair s and $d \in SOLT$.

D Total delay.

NR Total number of demands.

The total power consumption is composed of the power consumed by the servers, special servers, and OLT ports. It is worth mentioning that ONUs are only attached to the special servers. Regarding the power consumed by a server (PS_s), it involves the server's idle power, and the power consumed by the CPU utilisation due to transmitting, processing and forwarding traffic. It is calculated as follows:

$$PS_s = [\,\alpha_s\ PSI + (PSM - PSI)\ US_s]\,\sigma\ \ + PT\ \alpha T_s \qquad (6.2)$$

Here, σ is applied to the idle power and the operational power of the server because these general servers are there primarily to do processing and only σ of their total power is used for relaying. The power consumed by OLT's port for traffic forwarding (PO_i) is given as:

$$PO_i = \alpha O_i \text{ OIP} + (OMP - OIP)\, UO_i \tag{6.3}$$

As for the power consumed by the special servers (PR_r), it includes the special servers' idle power, the power consumed by the CPU utilisation for traffic forwarding and the power consumed by ONUs for communication which is assumed as a linear profile. It is given as:

$$PR_r = [\, \alpha R_r \; PSI + (PSM - PSI) \; UR_r \,]\, \tau + \frac{OP}{OR} \; TR_r \tag{6.4}$$

Accordingly, the model is defined as follows:

Objective:

Minimise:

$$\sum_{s \in S} PS_s + \sum_{r \in SS} PR_r + \sum_{i \in OLT} PO_i \; + PT \sum_{b \,\epsilon\, T} \alpha B_b \tag{6.5}$$

Equation (6.5) gives the model's objective which is to minimise the power consumed by servers, special servers, and OLT ports.

Subject to:

1) Traffic constraints:

$$\sum_{\substack{n \in N_m \\ n \neq m}} X_{mn}^{sd} - \sum_{\substack{n \in NB_m \\ n \neq m}} X_{nm}^{sd} = \begin{cases} T_{sd} & if\ m = s \\ -T_{sd} & if\ m = d \\ 0 & Otherwise \end{cases} \tag{6.6}$$

$$\forall\, s, d \in SOLT : s \neq d, \;\; \forall m \in N$$

Constraint (6.6) represents the flow conservation constraint for the traffic flows in the network following the flow conservation law [95].

$$X_{mn}^{sd} \geq \eta_{\;mn}^{\;sd} \tag{6.7}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, m \in N, \qquad \forall\, n \in N_m: m \neq n$$

$$X_{mn}^{sd} \leq \eta_{\;mn}^{\;sd} \quad M \tag{6.8}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, m \in N, \qquad \forall\, n \in N_m: m \neq n$$

Constraints (6.7) and (6.8) relate X_{mn}^{sd} to its binary equivalent $\eta\,_{mn}^{sd}$, the value used for M in the model is 100000.

$$\sum_{s \in SOLT} \sum_{\substack{d \in SOLT \\ s \neq d}} X_{mn}^{sd} \;\leq\; L_{mn} \qquad (6.9)$$

$$\forall m \;\in N\,, \forall n \;\in\; N_m$$

Constraint (6.9) ensures that the total traffic traversing link m and n does not exceed the links capacity.

$$\sum_{\substack{n \in N_m \\ m \neq n}} \eta\,_{mn}^{sd} \;\leq K \qquad (6.10)$$

$$\forall\, s, d \in SOLT\colon s \neq d, \qquad \forall\, m \in N$$

Constraint (6.10) specifies the number of routes a traffic demand can be bifurcated through. In this work, we have set $K = 1$, so a single path is used, i.e., traffic bifurcation is not allowed however the model is general.

$$Ns_s = \sum_{\substack{d \in SOLT \\ s \neq d}} \sum_{n \in N_s} \eta\,_{sn}^{sd} \qquad (6.11)$$

$$\forall\, s \in SOLT$$

Constraint (6.11) calculates the total number of traffic requests originated by node s.

$$Nd_s = \sum_{\substack{d \in SOLT \\ s \neq d}} \sum_{n \in NB_s} \eta\,_{ns}^{ds} \qquad (6.12)$$

$$\forall\, s \in SOLT$$

Constraint (6.12) calculates the total number of traffic requests destined to node s.

$$Nf_s = \sum_{m \in SOLT} \sum_{\substack{d \in SOLT \\ m \neq d}} \sum_{n \in NB_s} \eta\,_{ns}^{md} - \sum_{m \in SOLT} \sum_{n \in NB_s} \eta\,_{ns}^{ms} \qquad (6.13)$$

$$\forall\, s \in SOLT$$

Constraint (6.13) calculates the total number of traffic requests forwarded by node s.

$$NRf_r = \sum_{s \in SOLT} \sum_{\substack{d \in SOLT \\ s \neq d}} \sum_{n \in NB_r} \eta\,_{nr}^{sd} \tag{6.14}$$

$$\forall\, r \in SS$$

Constraint (6.14) calculates the total number of traffic requests forwarded by special server r.

$$Nf_s \ \gamma f \ \leq THSs \tag{6.15}$$

$$\forall\, s \in S$$

Constraint (6.15) controls the total CPU portion used for forwarding traffic by a server which should not exceed the $THSs$ value. $THSs$ value used in the model is 1.

$$UO_i \leq THOi \tag{6.16}$$

$$\forall\, i \in OLT$$

Constraint (6.16) controls the total link capacity used for forwarding traffic by OLT port which should not exceed the $THOi$ value. $THOi$ value used in the model is 1.

$$TS_s = \sum_{m \in SOLT} \sum_{\substack{d \in SOLT \\ m \neq d}} \sum_{n \in N_s} X_{sn}^{md} + \sum_{m \in SOLT} \sum_{\substack{d \in SOLT \\ m \neq d}} \sum_{n \in NB_s} X_{ns}^{md} \tag{6.17}$$

$$\forall\, s \in SOLT$$

Constraint (6.17) calculates the total traffic of a node s whether it is originated, destined or forwarded by node s.

$$TS_S \leq DR \tag{6.18}$$

$$\forall\, s \ \in S$$

Constraint (6.18) ensures that the total traffic of a server s does not exceed its data rate.

$$TS_S \leq \; OLTR \qquad (6.19)$$
$$\forall \, s \; \in OLT$$

Constraint (6.19) ensures that the total traffic of an OLT port s does not exceed its data rate.

$$TR_r = \sum_{s \in SOLT} \sum_{\substack{d \in SOLT \\ s \neq d}} \sum_{n \in NB_r} X_{nr}^{sd} \qquad (6.20)$$
$$\forall \, r \in SS$$

$$TR_r \leq ONUR \qquad (6.21)$$
$$\forall \, r \; \in SS$$

Constraint (6.20) calculates the total traffic outgoing of a special server r while constraint (6.21) ensures that the total traffic of a special server r does not exceed its fitted ONU data rate.

<u>Server constraints</u>

$$US_s = \; Ns_s \;\; \gamma s + Nd_s \;\; \gamma d + \; Nf_s \;\; \gamma f \qquad (6.22)$$
$$\forall \, s \; \in S$$

$$US_s \leq UZ \qquad (6.23)$$
$$\forall \, s \; \in S$$

Constraint (6.22) calculates the CPU utilisation of a server which results from sending, receiving, and forwarding traffic requests, while constraint (6.23) controls the total portion of CPU utilisation of a server which should not exceed the UZ value. The UZ value used in the model is 0.9 [49].

$$\alpha_s \; \leq \; US_s \qquad (6.24)$$
$$\forall \, s \; \in S$$

$$M\ \alpha_s \ \geq \ US_s \tag{6.25}$$

$$\forall\, s\ \in S$$

Constraints (6.24) and (6.25) are used to relate the binary variable α_s, to the non-binary variable US_s.

$$Nb_b = \sum_{m \in SOLT} \sum_{\substack{d \in SOLT \\ m \neq d}} \sum_{n \in N_b} \eta_{\,bn}^{\,md} \tag{6.26}$$

$$\forall\, b\ \in T$$

Constraint (6.26) is used to calculate the total number of requests passing through a transceiver b.

$$\alpha B_b \ \leq \ Nb_b \tag{6.27}$$

$$\forall\, b\ \in T$$

$$M\ \alpha B_b \ \geq \ Nb_b \tag{6.28}$$

$$\forall\, b\ \in T$$

Constraints (6.27) and (6.28) are used to relate the binary variable αB_b, to the non-binary variable Nb_b.

$$NC_n = \sum_{m \in SOLT} \sum_{\substack{d \in SOLT \\ m \neq d}} \eta_{\,nc}^{\,md} \tag{6.29}$$

$$\forall\, c\ \in C, \forall\, n \in NB_c$$

Constraint (6.29) is used to calculate the total number of requests sent by a server n to a coupler c.

$$NP_n = \sum_{m \in SOLT} \sum_{\substack{d \in SOLT \\ m \neq d}} \eta_{\,sp\,n}^{\,md} \tag{6.30}$$

$$\forall sp\ \in SP, \forall\, n \in N_{sp}$$

Constraint (6.30) is used to calculate the total number of requests sent by a splitter sp to one of its neighbouring servers n.

$$NT_s = NC_s + NP_s \tag{6.31}$$
$$\forall s \in S$$

Constraint (6.31) is used to calculate the total number of requests using the server's transceiver for sending and receiving inter rack (in BPPRD) or inter subgroup (in SSBD) communication.

$$\alpha T_s \leq NT_s \tag{6.32}$$
$$\forall s \in S$$

$$M \ \alpha T_s \geq NT_s \tag{6.33}$$
$$\forall s \in S$$

Constraints (6.32) and (6.33) are used to relate the binary variable αT_s, to the non-binary variable NT_s.

Special server constraints

$$UR_r = NRf_r \quad \gamma f \tag{6.34}$$
$$\forall r \in SS$$

$$UR_r \leq 1 \tag{6.35}$$

$$\forall r \in SS$$

Constraint (6.34) calculates the CPU utilisation of a special server which results from forwarding traffic requests, while constraint (6.35) controls the total portion of CPU utilisation of a special server.

$$\alpha R_r \leq NRf_r \tag{6.36}$$
$$\forall r \in SS$$

$$M \, \alpha R_r \geq NRf_r$$
$$\forall r \in SS \tag{6.37}$$

Constraints (6.36) and (6.37) are used to relate the binary variable αR_r, to the non-binary variable NRf_r.

OLT constraints

$$UO_i = \frac{TS_i}{OLTR} \qquad (6.38)$$

$$i \ \in OLT$$

Constraint (6.38) calculates the utilisation of an OLT port which results from sending, receiving, and forwarding traffic requests.

$$\alpha O_i \ \leq \ UO_i \, M \qquad (6.39)$$

$$\forall \, i \ \in OLT$$

$$M \, \alpha O_i \ \geq \ UO_i \qquad (6.40)$$

$$\forall \, i \ \in OLT$$

Constraints (6.39) and (6.40) are used to relate the binary variable αO_i, to the non-binary variable UO_i.

Delay constraints

Delay in these architectures is in four main parts which are queueing delay, propagation delay, transmission delay (time to transmit a packet) and reception delay (time to receive a packet) [111]. We ignored propagation delay as the distances between racks are in metres and we ignored the transmission delay and reception delay as these are common to all data centre architectures. We have used the Ethernet maximum packet size of 1500 bytes (12,000 bits) as the packet size, since this is the most popular packet size in the Internet.

The queueing delay was modelled based on queuing theory where the mean delay through an G/G/1 queuing system [112] is the mean time a packet takes to pass through the queue plus the node as in equation (6.41).

$$Mean\ system\ delay\ (Q) = \frac{1}{\mu - \lambda} \qquad (6.41)$$

where μ is the server processing rate in packet/s and λ is the mean arrival rate in packet/s; and the delay Q is in second per packet.

$$\sum_{\lambda \in Y} \lambda \quad I_{s\lambda} = TS_s \tag{6.42}$$

$$\forall\, s \in S$$

Constraint (6.42) is used to relate the server traffic load TS_s with the corresponding λ.

$$\sum_{\lambda \in Y} I_{s\lambda} \leq 1 \tag{6.43}$$

$$\forall\, s \in S$$

Constraint (7.43) ensures that only one value of λ is corresponding to server s.

$$DS_s = \sum_{\lambda \in Y} Q_\lambda \quad I_{s\lambda} \tag{6.44}$$

$$\forall\, s \in S$$

Constraint (7.44) is used to determine the delay of server s.

$$DS_{xn}^{sd} = \eta_{\,xn}^{\,sd} \; DS_x \tag{6.45}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, x \in S, \qquad \forall\, n \in N_x$$

Constraint (6.45) is used to relate the delay of node x to the traffic between nodes s and d passing through node x. Constraint (6.45) represent a non-linear equation that is replaced by constraints (6.46) - (6.48).

$$DS_{xn}^{sd} \leq \eta_{\,xn}^{\,sd} \; M \tag{6.46}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, x \in S, \qquad \forall\, n \in N_x$$

$$DS_{xn}^{sd} \leq DS_x \tag{6.47}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, x \in S, \qquad \forall\, n \in N_x$$

$$DS_{xn}^{sd} \geq \eta_{\,xn}^{\,sd} \; M + \; DS_x \; - M \tag{6.48}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, x \in S, \qquad \forall\, n \in N_x$$

$$\sum_{\lambda \in Y} \lambda \quad IR_{r\lambda} = TR_r \tag{6.49}$$

$$\forall\, r \in SS$$

Constraint (6.49) is used to relate the special server traffic load TR_r with the corresponding λ.

$$\sum_{\lambda \in Y} IR_{r\lambda} \leq 1 \tag{6.50}$$

$$\forall\, r \in SS$$

Constraint (6.50) ensures that only one value of λ is corresponding to special server r.

$$DR_r = \sum_{\lambda \in Y} QO_\lambda \quad IR_{r\lambda} \tag{6.51}$$

$$\forall\, r \in SS$$

Constraint (6.51) is used to determine the delay of special server r.

$$DR_{xn}^{sd} = \eta_{\ xn}^{\ sd} \quad DR_x \tag{6.52}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, x \in SS, \qquad \forall\, n \in N_x$$

Constraint (6.52) is used to relate the special server x delay to the traffic between nodes s and d passing through special server x. Constraint (6.52) represent a non-linear equation that is replaced by constraints (6.53) - (6.55).

$$DR_{xn}^{sd} \leq \eta_{\ xn}^{\ sd} \ M \tag{6.53}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, x \in SS, \qquad \forall\, n \in N_x$$

$$DR_{xn}^{sd} \leq DR_x \tag{6.54}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\, x \in SS, \qquad \forall\, n \in N_x$$

$$DR_{xn}^{sd} \geq \eta_{\;xn}^{\;sd} \; M + \; DR_{x} \; - M \tag{6.55}$$

$$\forall \, s, d \in SOLT: s \neq d, \qquad \forall \, x \in SS, \qquad \forall \, n \in N_{x}$$

$$\sum_{\lambda \in YOLT} \lambda \quad IO_{i\lambda} = \; TS_{i} \tag{6.56}$$

$$\forall \, i \; \in OLT$$

Constraint (6.56) is used to relate the OLT traffic load TS_i with the corresponding λ.

$$\sum_{\lambda \in YOLT} IO_{i\lambda} \leq \; 1 \tag{6.57}$$

$$\forall \, i \; \in OLT$$

Constraint (6.57) ensures that only one value of λ corresponds to OLT port i.

$$DO_{i} = \sum_{\lambda \in YOLT} QO_{\lambda} \quad IO_{i\lambda} \tag{6.58}$$

$$\forall \, i \; \in OLT$$

Constraint (6.58) is used to determine the delay of OLT port i.

$$DO_{xn}^{sd} \; = \; \eta_{\;xn}^{\;sd} \quad DO_{i} \tag{6.59}$$

$$\forall \, s, d \in S: s \neq d, \qquad \forall \, x \; \in OLT, \qquad \forall \, n \in N_{x}$$

Constraint (6.59) is used to relate the OLT port x delay to the traffic between nodes s and d passing through OLT port x. Constraint (6.59) represent a non-linear equation that is replaced by constraints (6.60) - (6.62).

$$DO_{xn}^{sd} \; \leq \; \eta_{\;xn}^{\;sd} \; M \tag{6.60}$$

$$\forall \, s, d \in SOLT: s \neq d, \qquad \forall \; x \; \in OLT, \qquad \forall \, n \in N_{x}$$

$$DO_{xn}^{sd} \; \leq DO_{x} \tag{6.61}$$

$$\forall \, s, d \in SOLT: s \neq d, \qquad \forall \; x \; \in OLT, \qquad \forall \, n \in N_{x}$$

$$DO_{xn}^{sd} \; \geq \; \eta_{\;xn}^{\;sd} \; M + \; DO_{x} \; - M \tag{6.62}$$

$$\forall\, s, d \in SOLT: s \neq d, \qquad \forall\ x\ \in OLT, \qquad \forall\, n \in N_x$$

$$D_{sd} = \sum_{x \in S} \sum_{n \in N_x} DS_{xn}^{sd} + \sum_{x \in OLT} \sum_{n \in N_x} DO_{xn}^{sd} + \sum_{x \in SS} \sum_{n \in N_x} DR_{xn}^{sd} \qquad (6.63)$$

$$\forall\, s, d \in SOLT: s \neq d$$

Constraint (6.63) calculates the total delay per request.

$$D = \sum_{s \in SOLT} \sum_{\substack{d \in SOLT \\ s \neq d}} D_{sd} \qquad (6.64)$$

Constraint (6.64) calculates the total delay for all the traffic requests.

$$NR = \sum_{s \in SOLT} Ns_s \qquad (6.65)$$

Constraint (6.65) calculates the total number of traffic requests.

Then, the average delay per request is obtained as:

$$Average\ Delay = \frac{D}{NR} \qquad (6.66)$$

6.7 Results and discussion:

In this subsection, the network settings, parameters' values along with the results of the MILP model are presented and discussed in detail. The network considered for this model consists of 16 servers placed into two racks each hosting 8 servers. Each rack contains two groups which are further divided into subgroups.

Table 6.2 presents the input parameters used for the model while Table 6.3 provides the values of Q_λ and λ used in queuing delay calculation for servers. Also, Table 6.4 provides the values of QO_λ and λ used for queuing delay calculations for special servers and OLT ports.

Table 6.2: Input data for the model

Parameter	Value
Traffic demand between servers TD_{sd}	200-800 Mbps random and uniformly distributed
Capacity of physical link L_{mn}	10 Gbps
Large enough number M	100000
Idle power consumption of a server PSI	301.6 W [98]
Maximum power consumption of a server PSM	457 W [98]
Idle power consumption of an OLT port OIP.	2 W [99]
Maximum operational power consumption of an OLT port.	14.3 [99]
Fraction of a server's processing capacity used for forwarding one traffic demand.	1.5%
Fraction of a server's processing capacity used for transmitting one traffic demand	0.3%
Fraction of a server's processing capacity used for receiving one traffic demand	0.2%
Server’s Data rate	1 Gbps
OLT’s port data rate	10 Gbps
Total ONU power consumption	2.5 W [97]
ONU data rate	10 Gbps
Fraction of server’s total power used for relaying (σ)	0.05
Fraction of special server’s total power used for relaying (τ)	0.15

Table 6.3: Traffic load and corresponding delay for a server

λ (Mb/s)	*Q (μsec/packet)*
200	15
400	20
600	30.1
800	60.2

Table 6.4: A sample of Traffic load and corresponding delay for a special server or OLT

λ (Mb/s)	*Q (μsec/packet)*
200	1.22
400	1.25
600	1.27
800	1.30
1000	1.33
1200	1.36
1400	1.39

The power consumption and delay in AWGR and server-based PON data centre architecture are reported as performance metrics that can be used to evaluate the two resilient modified AWGR and server-based PON data centre architecture designs. The performance is evaluated in the presence of different types of link failures. These two resilient modified designs connect the servers in each rack via a star coupler / backplane and duplicate the special servers.

For backplane per-rack design (BPPRD), in the normal state, the path for any inter-rack demand starts from the source server then traverses through two special servers (one connected to source and one connected to destination) and finally reaches its destination. In contrast, for intra-rack communication there are two path choices for any demand; (i) traversing through the backplane or (ii) acting similar to inter-rack traffic and reaching the destination through the special servers. This path choice is based on reducing power consumption, however it has to be noted that, traversing through a backplane, switches on the source and destination transceivers that are connected to the backplane.

As for the special server duplication-based design (SSBD), when running in the normal state, the path for any inter-subgroup demand starts from the source then traverses through two special servers (one connected to the source server and the other connected to the destination server), and finally reaches its destination. On the contrary, for intra-subgroup communication there will be two path choices for any demand: (i) traversing through the backplane; or (ii) acting similar to inter-subgroup and reaching the destination through special servers. This path choice will depend on which path consumes less power, noting that traversing through the backplane, switches on the source and destination transceivers that are connected to the backplane.

Accordingly, power consumption in the normal state, Figure 6.6 shows that the power consumption of the SSBD is higher than the backplane-based design since more demands need to be forwarded through two special servers. Here, half of the servers send to the other half, hence all the servers are active and all the special servers are active. Therefore, the power consumption difference between the two designs results from the power consumption of special servers dedicated for forwarding demands which is considerably small. From the aforementioned reasons, the difference in power consumption between the two designs is too small.

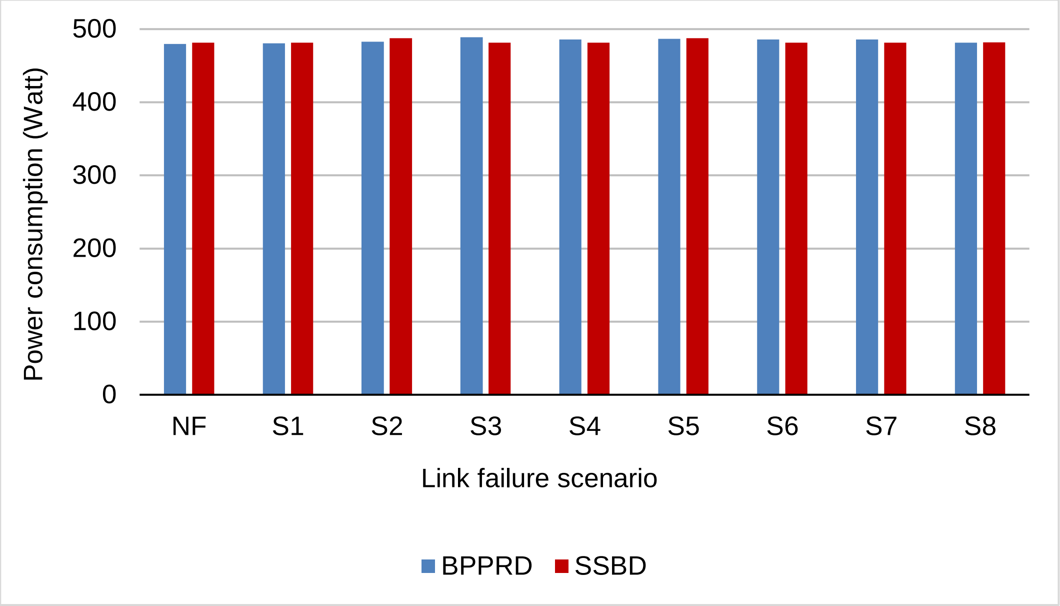


Figure 6.6: Power consumption for the two modified architectures (BPPRD and SSBD) under different failure scenarios

Regarding link failure S1, for both designs, the demands of the affected server need to traverse through the corresponding special server and back to their destination. This will add special server's power consumption for forwarding demands which is considerably small. This results in a slight difference in power consumption compared to the normal state.

Link failures S2 and S5 have similar effect on both designs causing the demands affected to relay through one more server. This increases power consumption due to adding the server forwarding power consumption which is small. This explains the slight difference in power consumption compared to the normal state of each design.

Link failures S3, S4, S6 and S7 cause the SSBD to only use the backup special server which results in no added power consumption to the normal case. On the contrary, these link failures cause the affected demands in the backplane-based design to relay through only one additional server. Again, this relaying (forwarding) power consumption is too small which causes slight difference in power consumption between the two designs. Regarding link failure S8, both designs react in the same way by using a relay PON group to recover from this failure which causes them to consume almost similar amount of power.

It is worthy to mention that the slight difference of both designs' power consumption under failure compared to the normal state (NF) represent a strength point of the improved designs. This is due to the ability of these improved designs to resist link failures with a very low power consumption cost.

Figure 6.7 shows the queuing delay for the two modified architectures. Here, the queuing delay of the SSBD is higher than the queueing delay of BPPRD. This is because in SSBD more demands will need to be forwarded through two more special servers compared to BPPRD. Since the queuing delay in a special server is too small compared to servers' queuing delay (as shown in Tables 6.3 and 6.4), therefore, there is a slight difference in queuing delay between the two designs.

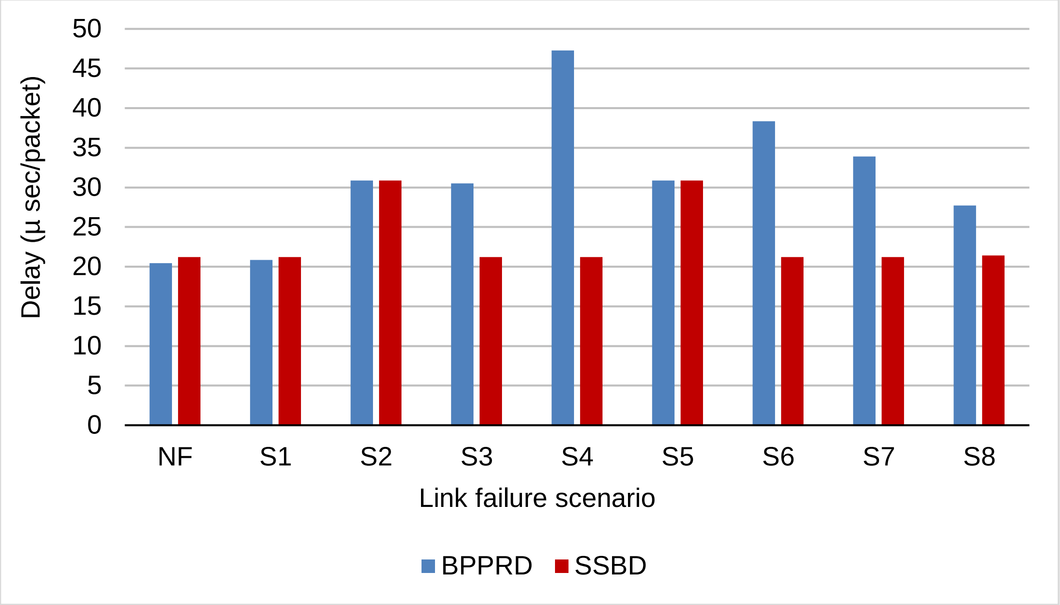


Figure 6.7: Delay for the two modified architectures (BPPRD and SSBD) under different failure scenarios

For link failure S1, in both designs, only the affected server demands need to relay through one more special server and back to their destination. This causes slight queuing delay difference compared to the normal state of each design.

As for link failures S2 and S5, in both designs, only the affected server demands need to be relayed through one more server which causes extra queuing delay compared to the normal state of each design. Here, the difference between the two designs is because the demand in backplane-based architecture has more relaying server options to choose from since all servers in a rack are connected by the backplane, so they can relay through a server with less queuing delay while in the SSBD this choice is limited to the servers in the corresponding subgroup.

Regarding link failures S3, S4, S6 and S7, the special server duplication-based design experiences no difference in queuing delay compared to the normal state since the effected demands use the backup special server. On the contrary, the backplane-based design reacts differently to these failures.

For S3, this influences the demands of the affected subgroup which need to traverse through the backplane, and has to be relayed by other servers in the same rack. This increases the queuing delay compared to the normal state. As for S4, S6, and S7, these failures influence all the inter-group demands transmitted and destined to the affected group servers, which causes them to be relayed by the other special server and servers in the corresponding rack to reach their destination. This, in turn, causes considerable difference in queuing delay compared to the normal state.

As for link failure S8, the affected demands in both designs; SSBD and BPPRD, need to be relayed through only one extra special server. The relaying special server choice depends on which one causes lower added power consumption. Here, for BPPRD, the special server with lower power consumption which is used to relay the traffic is the one in the same rack of the affected special server. The relaying continues through relay server(s) using the backplane to reach the destination. This is why there is considerably higher queuing delay compared to the normal state and the SSBD.

with regards to the performance, we considered the following metrics:

a) The ratio of the total traffic served by servers under failure scenario i to total traffic served in normal state which is calculated as

$$S_i = \frac{\sum_{s \in S} TS_s | c = i}{\sum_{s \in S} TS_s | c = 0} \qquad (6.67)$$

b) The ratio of total traffic served by special servers under failure scenario i to total traffic served in normal state which is calculated as:

$$R_i = \frac{\sum_{r \in SS} TR_r | c = i}{\sum_{r \in SS} TR_r | c = 0} \qquad (6.68)$$

c) The ratio of total traffic served by OLT ports under failure scenario i to total traffic served in normal state which is calculated as

$$O_i = \frac{\sum_{i \in OLT} TS_i | c = i}{\sum_{i \in OLT} TS_i | c = 0} \tag{6.69}$$

d) The ratio of total traffic served under failure scenario i to total traffic served in normal state which is calculated as

$$A_i = \frac{TA_a | c = i}{TA_a | c = 0} \tag{6.70}$$

where TA_a is calculated as

$$TA_a = \sum_{s \in SOLT} TS_s + \sum_{r \in SS} TR_r \tag{6.71}$$

The results for these performance metrics are shown in Figures 6.8 - 6.11. Regarding metrics S_i and R_i , it is clear that under failures and under BPPRD, servers carry more traffic than special servers as shown in Figures 6.8 and 6.9. Also, Figure 6.10 shows that OLT ports traffic is not affected under failures in both designs. This means that the OLT did not relay any affected traffic under failures in both designs, and the affected traffic was rerouted within the PON cell. The metric that is most sensitive to failures is A_i , which reflects the fact that under failures and under BPPRD, the backplane carries more traffic as shown in Figure 6.11.

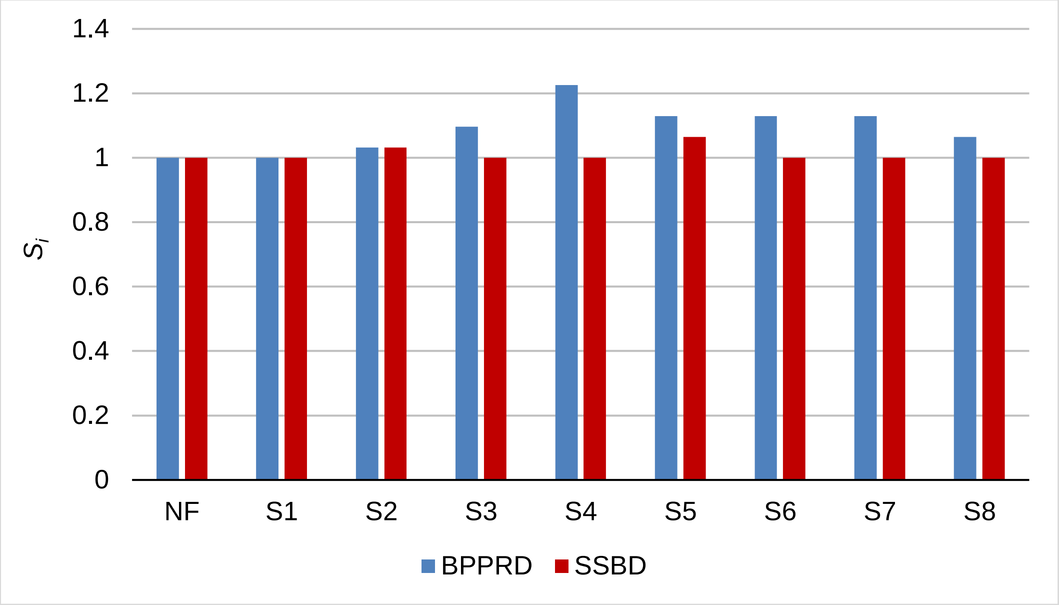


Figure 6.8: **The ratio of the total traffic served by servers under failure scenario i to total traffic served in normal state for the two modified architectures (BPPRD and SSBD)**

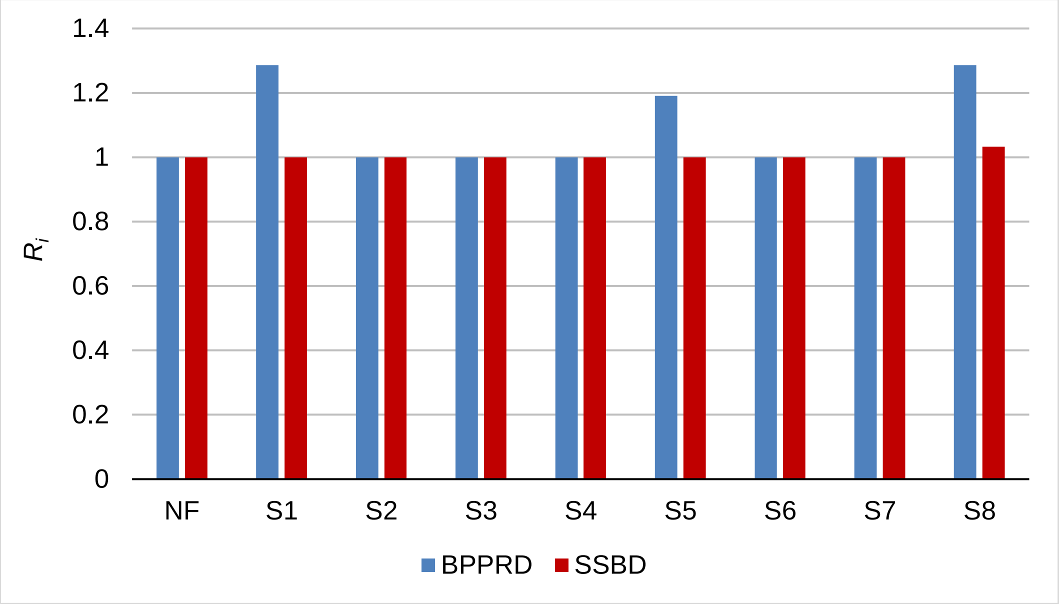


Figure 6.9: The ratio of total traffic served by special servers under failure scenario i to total traffic served in normal state for the two modified architectures (BPPRD and SSBD)

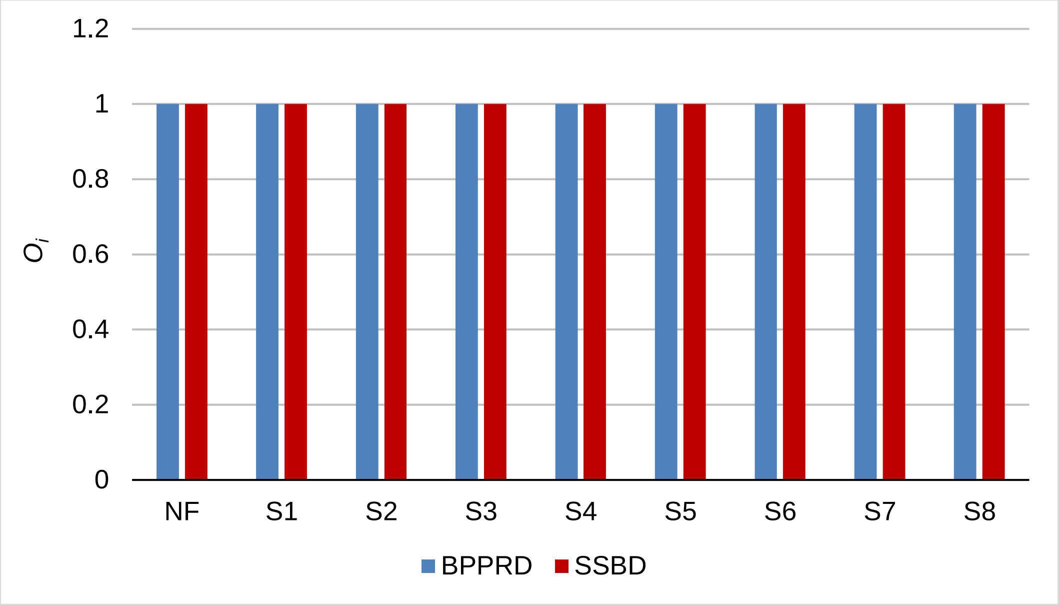


Figure 6.10: **The ratio of total traffic served by OLT port under failure scenario i to total traffic served in normal state for the two modified architectures (BPPRD and SSBD)**

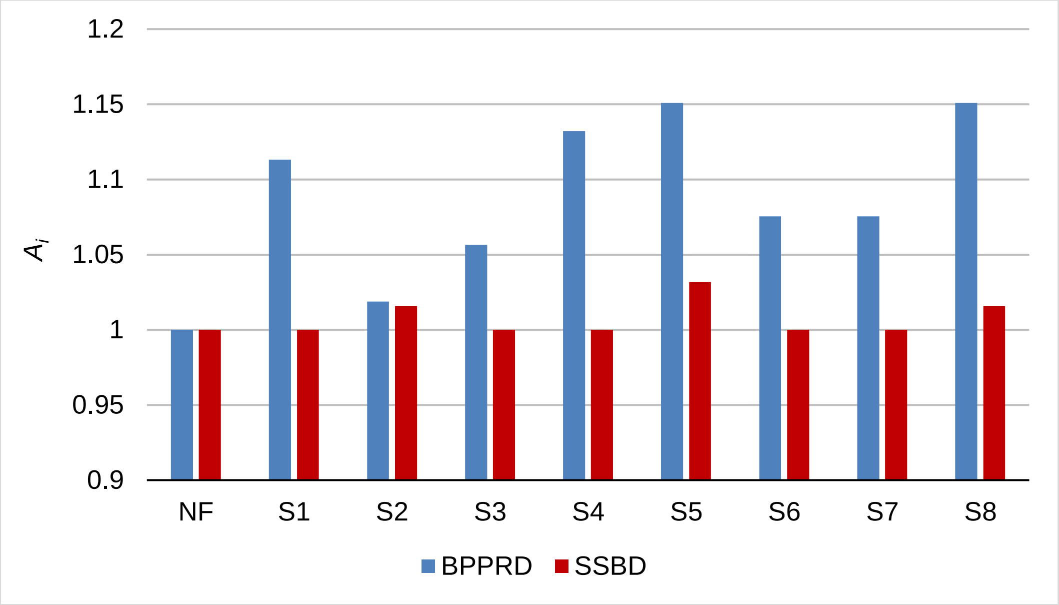


Figure 6.11: **The ratio of total traffic served under failure scenario i to total traffic served in normal state for the two modified architectures (BPPRD and SSBD)**

6.8 Summary

This chapter introduced several modified designs for the AWGR and server-based PON data centre architecture to study their resilience ability in the face of different kinds of failures. It was found that connecting the servers in each rack via a star coupler / backplane or duplicating the special servers are the most resilient options.

A MILP model was developed to evaluate the performance of these two modified designs considering different failure scenarios. The results show that duplicating the special servers reduces power consumption and delay compared to the servers in each rack sharing a star coupler / backplane. However, choosing between these two resilient designs is a compromise between cost and performance.

Chapter 7 Conclusion and future work

This chapter provides the conclusions for the work presented in this thesis and highlights some aspects which are suggested as possible future research directions.

7.1 Conclusions

The DCN infrastructure is one of the main factors that specify the properties of a data centre, such as its resilience, latency, and expansion complexity. Chapter 2 introduced the traditional electronically-switched data centre architectures, described the all optical and hybrid data centre architectures, and presented a review of their main advantages and limitations. Chapter 3 presented an overview of PON in access networks, with respect to its architecture, topologies, technologies, and components. This was followed by a detailed description of five recently proposed fully passive optical architecture options. Important conclusions can be drawn regarding these five architectures. First, for the TDM-based PON and TDM-WDM-based PON data centre architectures (i.e., PON 1 and PON 2), all the inter-rack traffic passes through the OLT, resulting in additional unwanted delay and power consumption because of the buffering, processing, and rerouting that has to be carried out by the OLT. Second, the AWGR-based PON data centre architecture (PON 3) minimises the power consumption and delay, as it provides direct high data rate routes for inter-rack traffic using tuneable lasers in ONUs. However, it has a high deployment cost, as it uses an ONU with a tuneable laser for each server. Third, the server-based PON data centre architecture (PON 5) has a low deployment cost, as it eliminates the use of ONUs that have tuneable lasers and AWGRs: However, the delay increases as traffic needs to hop between servers for routing traffic. Fourth, the AWGR and server-based PON data centre architecture (PON 4) is a trade-off between cost and delay. Compared with the AWGR-based PON data centre architecture, it reduces the number of ONUs by equipping only the special servers with ONUs that have tuneable lasers which reduces cost. On the contrary, compared with the server-based PON data centre architecture, no hopping is required, as the special servers are responsible for routing which reduces delay. Therefore, the AWGR and server-based PON data centre architecture was chosen for a further thorough investigation.

In Chapter 4, the AWGR and server-based PON data centre architecture wavelength routing and assignment for the inter-group communication was optimised using a MILP model. A benchmark study showed that the AWGR and server-based PON data centre architecture reduced the power consumption by 80% compared to a BCube topology of 512 servers, by 82% compared to a BCube topology of 4096 servers, and by 83% compared to a BCube architecture of 32768 servers.

Chapter 5 presented an investigation of the embedding of VMs in the AWGR and server-based PON data centre architecture servers. We minimised the power consumption of cloud applications allocation by using a MILP model and presented a range of results. The results showed that the power consumption is affected by the number of servers in each subgroup because of the change in the number of activated special servers. Our study showed that the proposed model reduced the power consumption by 34%, 21%, 34%, and 18%, compared to the non-optimised embedding model of 5, 10, 15, and 20 VMs, respectively.

In Chapter 6, a MILP model was developed to evaluate the resilience of the modified PON data centre architecture design based on the AWGR and server-based PON data centre architecture. The impact of the main active components' failure is analysed in terms of the power consumption and the delay. Consideration was given to link failure (due to active i.e., electronics and/or photonics transceiver components failure) and the failure of the special server (which is an active component with active components).

7.2 Contributions to knowledge

The main contributions to knowledge from this study help to investigate the performance of the AWGR and server-based PON data centre design. The contributions of the current work are summarised as follows:

1. It is the first to analyse the AWGR and server-based PON data centre architecture and, hence, is concerned with designing and optimising its routing and wavelength assignment.
2. It optimises virtual machine placement for cloud applications in the AWGR and server-based PON data centre architecture.
3. It studies the resilience ability of the AWGR and server-based PON data centre architecture in the face of different kinds of failures.

7.3 Future research directions

1. **Experimental demonstration**: The work in Chapters 4, 5, and 6 is based on mathematical models and computer simulations. A possible future direction can be to develop experimental demonstrations to validate and verify the results obtained from the MILP models and the developed heuristic.

2. **Deployment cost reduction**: The architectures of the server-based PON and the AWGR and server-based PON data centre architectures were proposed in [48] to reduce or eliminate the need for ONUs that have tuneable lasers, which in turn reduced the power consumption and cost. A possible further research direction is to reduce the deployment cost by reducing or eliminating the AWGRs (which reduces the number of fibre-polymer complex links) used in the AWGR and server-based PON data centre architecture. The proposed architectures in Figures 7.1 and 7.2 have the potential to address the above-suggested future work.

   Further, the proposed architecture in Figure 7.1 (proposed PON 6) is derived from the AWGR-based PON and the AWGR and server-based PON data centre architectures, where it combines the rack structure of the AWGR and server-based PON data centre architecture with the connection between the racks of the AWGR-based PON data centre architecture. The servers in a rack are distributed in a manner similar to that used in the AWGR and server-based PON data centre architecture, where a special server is used for the inter-rack communication, while the connection between racks is performed by using only two AWGRs. Note that this architecture (PON 6) uses fewer AWGRs compared to PON 4, however these fewer AWGRs have a larger number of ports similar to PON 3.

   The proposed architecture in Figure 7.2 (proposed PON 7) eliminates the use of AWGRs in the AWGR and server-based PON data centre architecture. Here, the servers in a rack are distributed in a manner similar to that used in the AWGR and server-based PON data centre architecture, where a special server is used for the inter-rack communication, while the connection between racks is performed by using a passive polymer backplane.

These two architectures need to be studied in terms of their impact on scalability, per-server share of wavelength, resilience, and latency.

3. **Software Defined Network (SDN) adoption:** In the AWGR and server-based PON data centre architecture, the special servers are responsible for routing the inter-group traffic based on a database of server addresses and the assigned wavelength of each group. In addition, these special servers share information with one another, which is done directly or via the OLT, in order to update their databases. Therefore, these special serves can adopt and benefit from SDN features such as load balancing. These special servers can include application programming interfaces (API) which can then interface with an SDN controller that can compose cells, disintegrate cells and assign tasks (load) to cells to balance the processing and networking loads.

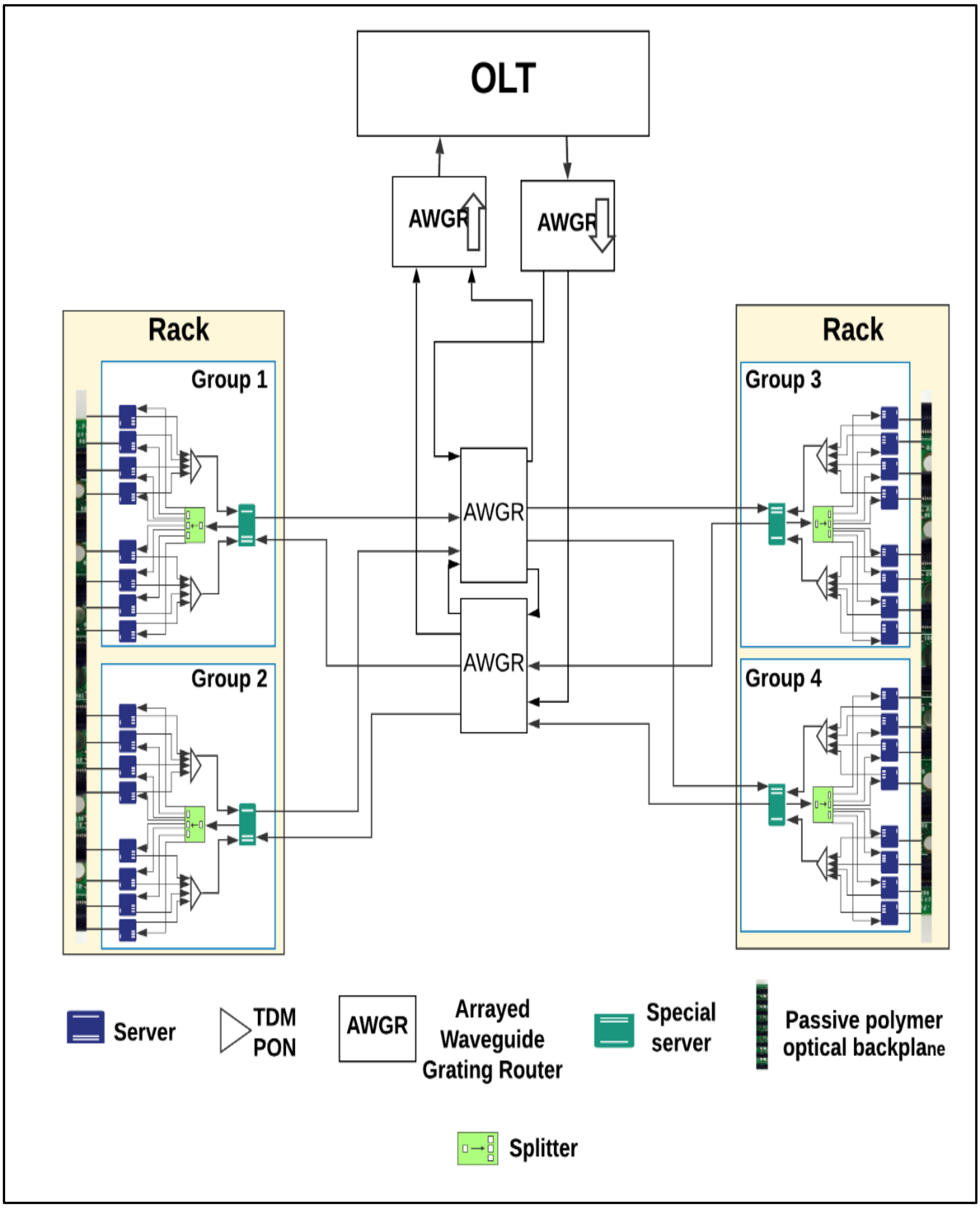


Figure 7.1: Proposed PON 6 architecture

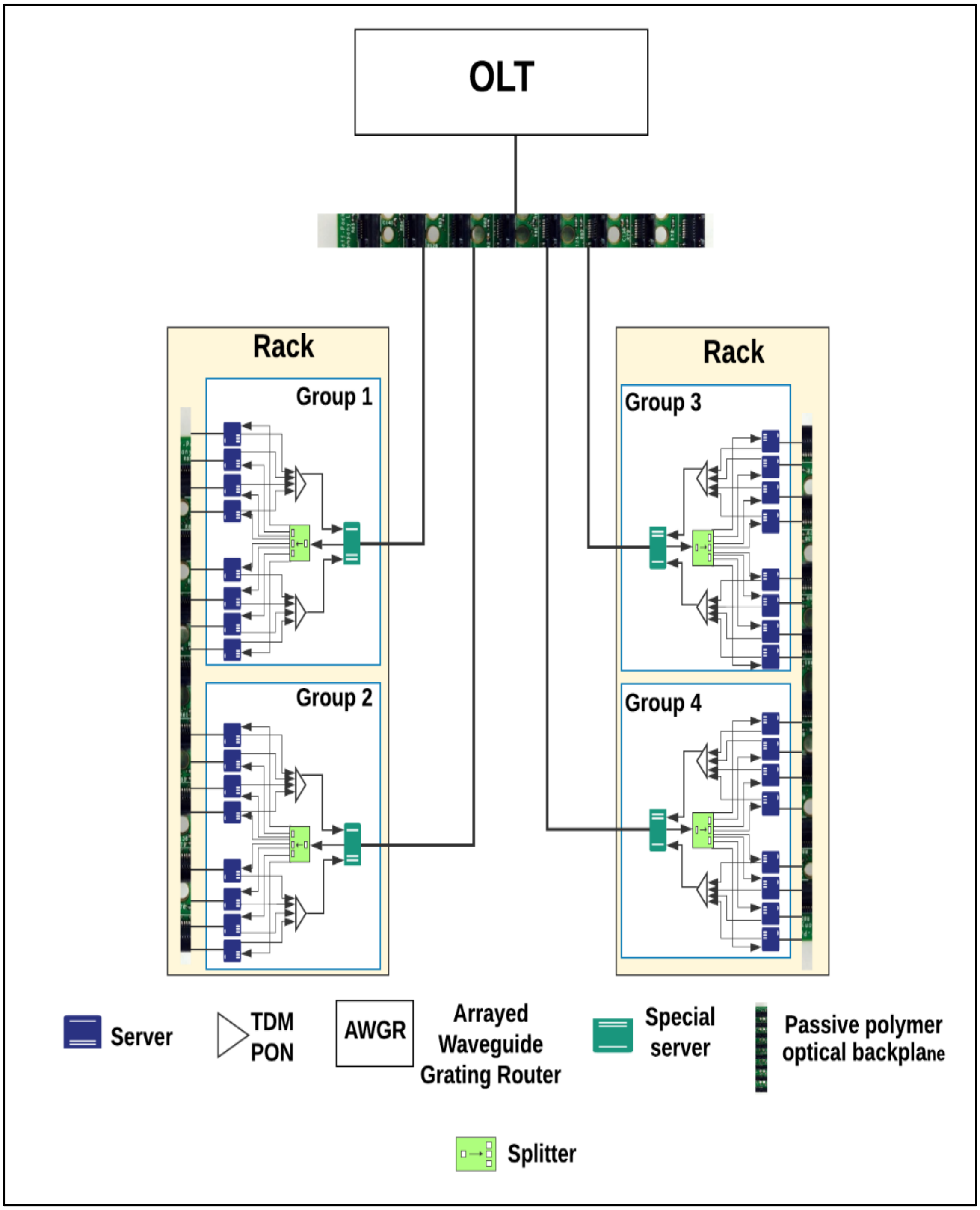


Figure 7.2: Proposed PON 7 architecture

List of References

[1] J. Shuja *et al.*, "Survey of techniques and architectures for designing energy-efficient data centers," vol. 10, no. 2, pp. 507-519, 2014.

[2] J. Elmirghani *et al.*, "GreenTouch GreenMeter core network energy-efficiency improvement measures and optimization," vol. 10, no. 2, pp. A250-A269, 2018.

[3] IEA. "Data centres and data transmission networks." https://www.iea.org/reports/tracking-buildings.

[4] Cisco. "Cisco Annual Internet Report (2018–2023) White Paper." https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[5] Z. T. Al-Azez, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Virtualization framework for energy efficient IoT networks," in *2015 IEEE 4th International Conference on Cloud Networking (CloudNet)*, 2015: IEEE, pp. 74-77.

[6] H. M. M. Ali, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Energy efficient disaggregated servers for future data centers," in *2015 20th European Conference on Networks and Optical Communications-(NOC)*, 2015: IEEE, pp. 1-6.

[7] X. Dong, T. El-Gorashi, and J. M. Elmirghani, "Green IP over WDM networks: Solar and wind renewable sources and data centres," in *2011 IEEE Global Telecommunications Conference-GLOBECOM 2011*, 2011: IEEE, pp. 1-6.

[8] X. Dong, T. El-Gorashi, and J. M. Elmirghani, "Energy-efficient IP over WDM networks with data centres," in *2011 13th International Conference on Transparent Optical Networks*, 2011: IEEE, pp. 1-8.

[9] X. Dong, T. El-Gorashi, and J. J. I. o. Elmirghani, "Use of renewable energy in an IP over WDM network with data centres," vol. 6, no. 4, pp. 155-164, 2012.

[10] X. Dong, T. El-Gorashi, and J. M. J. J. o. L. T. Elmirghani, "Green IP over WDM networks with data centers," vol. 29, no. 12, pp. 1861-1880, 2011.

[11] X. Dong, T. E. El-Gorashi, and J. M. Elmirghani, "Joint optimization of power, electricity cost and delay in IP over WDM networks," in *2013 IEEE International Conference on Communications (ICC)*, 2013: IEEE, pp. 2370-2375.

[12] X. Dong, T. E. El-Gorashi, and J. M. J. J. o. L. T. Elmirghani, "On the energy efficiency of physical topology design for IP over WDM networks," vol. 30, no. 12, pp. 1931-1942, 2012.

[13] A. Q. Lawey, T. E. El-Gorashi, and J. M. J. J. o. L. T. Elmirghani, "Distributed energy efficient clouds over core networks," vol. 32, no. 7, pp. 1261-1281, 2014.

[14] M. O. Musa, T. E. El-Gorashi, and J. M. Elmirghani, "Energy efficient core networks using network coding," in *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 2015: IEEE, pp. 1-4.

[15] M. O. Musa, T. E. El-Gorashi, and J. M. J. J. o. L. T. Elmirghani, "Bounds on GreenTouch GreenMeter network energy efficiency," vol. 36, no. 23, pp. 5395-5405, 2018.

[16] L. Nonde, T. E. El-Gorashi, and J. M. J. J. o. L. T. Elmirghani, "Energy efficient virtual network embedding for cloud networks," vol. 33, no. 9, pp. 1828-1849, 2014.

[17] A. T. Hussein, M. T. Alresheedi, and J. M. J. J. o. l. t. Elmirghani, "20 Gb/s mobile indoor visible light communication system employing beam steering and computer generated holograms," vol. 33, no. 24, pp. 5242-5260, 2015.

[18] A. T. Hussein and J. M. J. J. o. L. T. Elmirghani, "Mobile multi-gigabit visible light communication system in realistic indoor environment," vol. 33, no. 15, pp. 3293-3307, 2015.

[19] A. Q. Lawey, T. E. El-Gorashi, and J. M. J. J. o. l. t. Elmirghani, "BitTorrent content distribution in optical networks," vol. 32, no. 21, pp. 4209-4225, 2014.

[20] A. T. Hussein, J. M. J. J. o. O. C. Elmirghani, and Networking, "10 Gbps mobile visible light communication system employing angle diversity, imaging receivers, and relay nodes," vol. 7, no. 8, pp. 718-735, 2015.

[21] M. T. Alresheedi, A. T. Hussein, and J. M. J. I. C. Elmirghani, "Uplink design in VLC systems with IR sources and beam steering," vol. 11, no. 3, pp. 311-317, 2017.

[22] H. M. M. Ali, T. E. El-Gorashi, A. Q. Lawey, and J. M. J. J. o. L. T. Elmirghani, "Future energy efficient data centers with disaggregated servers," vol. 35, no. 24, pp. 5361-5380, 2017.

[23] A. M. Al-Salim, A. Q. Lawey, T. E. El-Gorashi, J. M. J. I. T. o. N. Elmirghani, and S. Management, "Energy efficient big data networks: Impact of volume and variety," vol. 15, no. 1, pp. 458-474, 2017.

[24] M. Musa, T. Elgorashi, J. J. J. o. O. C. Elmirghani, and Networking, "Energy efficient survivable IP-over-WDM networks with network coding," vol. 9, no. 3, pp. 207-217, 2017.

[25] M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. J. C. N. Elmirghani, "Big data analytics for wireless and wired network design: A survey," vol. 132, pp. 180-199, 2018.

[26] M. Musa, T. Elgorashi, J. J. J. o. O. C. Elmirghani, and Networking, "Bounds for energy-efficient survivable IP over WDM networks with network coding," vol. 10, no. 5, pp. 471-481, 2018.

[27] A. N. Al-Quzweeni, A. Q. Lawey, T. E. Elgorashi, and J. M. J. I. A. Elmirghani, "Optimized energy aware 5G network function virtualization," vol. 7, pp. 44939-44958, 2019.

[28] T. E. El-Gorashi, X. Dong, and J. M. J. I. O. Elmirghani, "Green optical orthogonal frequency-division multiplexing networks," vol. 8, no. 3, pp. 137-148, 2014.

[29] A. Hammadi, M. Musa, T. E. El-Gorashi, and J. H. Elmirghani, "Resource provisioning for cloud PON AWGR-based data center architecture," in *2016 21st European Conference on Networks and Optical Communications (NOC)*, 2016: IEEE, pp. 178-182.

[30] Y. Ye *et al.*, "Energy-efficient resilient optical networks: Challenges and trade-offs," vol. 53, no. 2, pp. 144-150, 2015.

[31] A. M. Al-Salim, T. E. El-Gorashi, A. Q. Lawey, and J. M. J. I. O. Elmirghani, "Greening big data networks: Velocity impact," vol. 12, no. 3, pp. 126-135, 2017.

[32] X. Dong, T. E. El-Gorashi, and J. M. Elmirghani, "Energy efficiency of optical OFDM-based networks," in *2013 IEEE International Conference on Communications (ICC)*, 2013: IEEE, pp. 4131-4136.

[33] N. I. Osman, T. El-Gorashi, and J. M. J. J. o. H. S. N. Elmirghani, "Caching in green IP over WDM networks," vol. 19, no. 1, pp. 33-53, 2013.

[34] A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Renewable energy in distributed energy efficient content delivery clouds," in *2015 IEEE International Conference on Communications (ICC)*, 2015: IEEE, pp. 128-134.

[35] J. Elmirghani *et al.*, "Energy efficiency measures for future core networks," in *Optical Fiber Communication Conference*, 2017: Optical Society of America, p. Th1I. 4.

[36] A. Hammadi, T. E. El-Gorashi, and J. M. Elmirghani, "High performance AWGR PONs in data centre networks," in *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 2015: IEEE, pp. 1-5.

[37] A. Hammadi, T. E. El-Gorashi, M. O. Musa, and J. M. Elmirghani, "Server-centric PON data center architecture," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, 2016: IEEE, pp. 1-4.

[38] O. Z. Alsulami, M. O. Musa, M. T. Alresheedi, and J. M. Elmirghani, "Visible light optical data centre links," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, 2019: IEEE, pp. 1-5.

[39] A. E. Eltraify, M. O. Musa, A. Al-Quzweeni, and J. M. Elmirghani, "Experimental evaluation of passive optical network based data centre architecture," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, 2018: IEEE, pp. 1-4.

[40] A. E. Eltraify, M. O. Musa, A. Al-Quzweeni, and J. M. Elmirghani, "Experimental Evaluation of Server Centric Passive Optical Network Based Data Centre Architecture," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, 2019: IEEE, pp. 1-5.

[41] A. E. Eltraify, M. O. Musa, and J. M. Elmirghani, "TDM/WDM over AWGR Based Passive Optical Network Data Centre Architecture," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, 2019: IEEE, pp. 1-5.

[42] J. Elmirghani, T. El-Gorashi, and A. Hammadi, "Passive optical-based data center networks," ed: Leeds, 2015.

[43] S. H. Mohamed, T. E. El-Gorashi, and J. M. Elmirghani, "Energy efficiency of server-centric PON data center architecture for fog computing," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, 2018: IEEE, pp. 1-4.

[44] W. Xia, P. Zhao, Y. Wen, H. J. I. c. s. Xie, and tutorials, "A survey on data center networking (DCN): Infrastructure and operations," vol. 19, no. 1, pp. 640-656, 2016.

[45] K. Bilal *et al.*, "A taxonomy and survey on green data center networks," vol. 36, pp. 189-208, 2014.

[46] L. Popa, S. Ratnasamy, G. Iannaccone, A. Krishnamurthy, and I. Stoica, "A cost comparison of datacenter network architectures," in *Proceedings of the 6th International COnference*, 2010, pp. 1-12.

[47] C. Kachris, I. J. I. C. S. Tomkos, and Tutorials, "A survey on optical interconnects for data centers," vol. 14, no. 4, pp. 1021-1036, 2012.

[48] C. F. Lam, *Passive optical networks: principles and practice*. Elsevier, 2011.

[49] A. A. Hammadi, "Future PON data centre networks," University of Leeds, 2016.

[50] M. Gupta and S. Singh, "Greening of the Internet," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, 2003, pp. 19-26.

[51] F. P. Tso, S. Jouet, and D. P. Pezaros, "Network and server resource management strategies for data centre infrastructures: A survey," *Computer Networks,* vol. 106, pp. 209-225, 2016.

[52] M. Noormohammadpour, C. S. J. I. C. S. Raghavendra, and Tutorials, "Datacenter traffic control: Understanding techniques and tradeoffs," vol. 20, no. 2, pp. 1492-1525, 2017.

[53] *Green data centers: A survey, perspectives, and future directions,* a. p. arXiv: , 2016.

[54] M. Al-Fares, A. Loukissas, and A. J. A. S. c. c. r. Vahdat, "A scalable, commodity data center network architecture," vol. 38, no. 4, pp. 63-74, 2008.

[55] A. Greenberg *et al.*, "VL2: a scalable and flexible data center network," in *Proceedings of the ACM SIGCOMM 2009 conference on Data communication*, 2009, pp. 51-62.

[56] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking data centers randomly," in *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, 2012, pp. 225-238.

[57] X. Yuan, S. Mahapatra, W. Nienaber, S. Pakin, and M. Lang, "A new routing scheme for Jellyfish and its performance with HPC workloads," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2013, pp. 1-11.

[58] A. Agache, R. Deaconescu, and C. Raiciu, "Increasing Datacenter Network Utilisation with {GRIN}," in *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, 2015, pp. 29-42.

[59] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "Dcell: a scalable and fault-tolerant network structure for data centers," in *Proceedings of the ACM SIGCOMM 2008 conference on Data communication*, 2008, pp. 75-86.

[60] C. Guo *et al.*, "BCube: a high performance, server-centric network architecture for modular data centers," in *Proceedings of the ACM SIGCOMM 2009 conference on Data communication*, 2009, pp. 63-74.

[61] S. B. Yoo, Y. Yin, and K. Wen, "Intra and inter datacenter networking: The role of optical packet switching and flexible bandwidth optical networking," in *2012 16th International Conference on Optical Network Design and Modelling (ONDM)*, 2012: IEEE, pp. 1-6.

[62] O. Liboiron-Ladouceur, "Optical interconnection networks for data centers," in *2014 IEEE Photonics Conference*, 2014: IEEE, pp. 67-68.

[63] G. Wang *et al.*, "c-Through: Part-time optics in data centers," in *Proceedings of the ACM SIGCOMM 2010 conference*, 2010, pp. 327-338.

[64] N. Farrington *et al.*, "Helios: a hybrid electrical/optical switch architecture for modular data centers," in *Proceedings of the ACM SIGCOMM 2010 conference*, 2010, pp. 339-350.

[65] K. Chen *et al.*, "OSA: An optical switching architecture for data center networks with unprecedented flexibility," vol. 22, no. 2, pp. 498-511, 2013.

[66] X. Ye, Y. Yin, S. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: A scalable optical switch for datacenters," in *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, 2010, pp. 1-12.

[67] M. Dayarathna, Y. Wen, R. J. I. C. S. Fan, and Tutorials, "Data center energy consumption modeling: A survey," vol. 18, no. 1, pp. 732-794, 2015.

[68] M. Li, A. C. Yao, and F. F. J. P. o. t. N. A. o. S. Yao, "Discrete and continuous min-energy schedules for variable voltage processors," vol. 103, no. 11, pp. 3983-3987, 2006.

[69] X. Jin, F. Zhang, Y. Song, L. Fan, and Z. Liu, "Energy-efficient scheduling with time and processors eligibility restrictions," in *European Conference on Parallel Processing*, 2013: Springer, pp. 66-77.

[70] S. Irani, S. Shukla, and R. J. A. T. o. E. C. S. Gupta, "Online strategies for dynamic power management in systems with multiple power-saving states," vol. 2, no. 3, pp. 325-346, 2003.

[71] H. M. M. Ali, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Energy efficient resource provisioning in disaggregated data centres," in *Asia Communications and Photonics Conference*, 2015: Optical Society of America, p. AM1H. 1.

[72] H. M. M. Ali, A. M. Al-Salim, A. Q. Lawey, T. El-Gorashi, and J. M. Elmirghani, "Energy efficient resource provisioning with VM migration heuristic for disaggregated server design," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, 2016: IEEE, pp. 1-5.

[73] Y. T. S. Al-Anii, "Holistic and Energy-Efficient Management of Datacentres," University of Leeds, 2017.

[74] G. G. DATA, "Center power efficiency metrics: Pue and dcie," 2008.

[75] C. Yu and L. Lai, "Study on metrics model for energy efficiency in data centers," in *2016 International Conference on Sensor Network and Computer Engineering*, 2016: Atlantis Press.

[76] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," ed: ACM New York, NY, USA, 2008.

[77] W. Zhang, Y. Wen, Y. W. Wong, K. C. Toh, C.-H. J. I. C. S. Chen, and Tutorials, "Towards joint optimization over ICT and cooling systems in data centre: A survey," vol. 18, no. 3, pp. 1596-1616, 2016.

[78] K. Bilal, S. U. R. Malik, S. U. Khan, and A. Y. J. I. c. c. Zomaya, "Trends and challenges in cloud datacenters," vol. 1, no. 1, pp. 10-20, 2014.

[79] K. Christensen, P. Reviriego, B. Nordman, M. Bennett, M. Mostowfi, and J. A. J. I. C. M. Maestro, "IEEE 802.3 az: the road to energy efficient ethernet," vol. 48, no. 11, pp. 50-56, 2010.

[80] S. B. Weinstein, Y. Luo, and T. Wang, *The ComSoc guide to passive optical networks: Enhancing the last mile access*. John Wiley & Sons, 2012.

[81] N. Ansari and J. Zhang, *Media access control and resource allocation: For next generation passive optical networks*. Springer Science & Business Media, 2013.

[82] S. Lallukka and P. Raatikainen, *Passive optical networks: transport concepts*. VTT Technical Research Centre of Finland, 2006.

[83] G. Keiser, *FTTX concepts and applications*. John Wiley & Sons, 2006.

[84] J. M. Senior and M. Y. Jamro, *Optical fiber communications: principles and practice*. Pearson Education, 2009.

[85] H. J. Dutton, *Understanding optical communications*. Prentice Hall PTR Durham, North Carolina, USA, 1998.

[86] M. M. Saleh, R. Ani, and I. K. J. P. C. J. Onees, "Simulation of uniform and apodized fiber Bragg grating," vol. 9, pp. 239-46, 2014.

[87] J. Beals *et al.*, "Terabit capacity passive polymer optical backplane," in *2008 Conference on Lasers and Electro-Optics and 2008 Conference on Quantum Electronics and Laser Science*, 2008: IEEE, pp. 1-2.

[88] 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*, 2011: IEEE, pp. 118-121.

[89] Y. Cheng, M. Fiorani, R. Lin, L. Wosinska, J. J. J. o. O. C. Chen, and Networking, "POTORI: a passive optical top-of-rack interconnect architecture for data centers," vol. 9, no. 5, pp. 401-411, 2017.

[90] A. Hammadi, T. E. El-Gorashi, and J. M. Elmirghani, "High performance AWGR PONs in data centre networks," in *Transparent Optical Networks (ICTON), 2015 17th International Conference on*, 2015: IEEE, pp. 1-5.

[91] A. Hammadi, T. E. El-Gorashi, M. O. Musa, and J. M. Elmirghani, "Server-centric PON data center architecture," in *Transparent Optical Networks (ICTON), 2016 18th International Conference on*, 2016: IEEE, pp. 1-4.

[92] A. Hammadi, T. E. El-Gorashi, and J. M. Elmirghani, "Energy-efficient software-defined AWGR-based PON data center network," in *Transparent Optical Networks (ICTON), 2016 18th International Conference on*, 2016: IEEE, pp. 1-5.

[93] A. Hammadi, M. Musa, T. E. El-Gorashi, and J. H. Elmirghani, "Resource provisioning for cloud PON AWGR-based data center architecture," in *Networks and Optical Communications (NOC), 2016 21st European Conference on*, 2016: IEEE, pp. 178-182.

[94] J. Elmirghani, T. EL-GORASHI, and A. HAMMADI, "Passive optical-based data center networks," ed: Google Patents, 2016.

[95] J. W. Chinneck, "Practical optimization: a gentle introduction," *Systems and Computer Engineering), Carleton University, Ottawa. http://www.sce.carleton.ca/faculty/chinneck/po.html,* p. 11, 2006.

[96] L. Gyarmati and T. A. Trinh, "How can architecture help to reduce energy consumption in data center networking?," in *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*, 2010, pp. 183-186.

[97] K. Grobe, M. Roppelt, A. Autenrieth, J.-P. Elbers, and M. J. I. C. M. Eiselt, "Cost and energy consumption analysis of advanced WDM-PONs," vol. 49, no. 2, pp. s25-s32, 2011.

[98] S. P. E. Corporation. "Dell Inc. PowerEdge R740." https://www.spec.org/power_ssj2008/results/res2017q3/power_ssj2008-20170829-00780.html.

[99] Z. CORPORATION, "ZXA10 C300 Optical Access Convergence Equipment Hardware Description" 2013.

[100] Cisco. "Cisco Catalyst 2960-S and 2960 Series Switches with LAN Lite Software Data Sheet." https://www.cisco.com/c/en/us/products/collateral/switches/catalyst-2960-series-switches/product_data_sheet0900aecd806b0bd8.html.

[101] PMC. "PAS740x GPON ONT SoC DataSheet." https://www.datasheetarchive.com/PAS740x-datasheet.html.

[102] C. G. V. Networking and C. Index, "Forecast and Methodology, 2011-2016," ed.

[103] Intel. "Intel® Xeon® Gold 6210U Processor." https://ark.intel.com/content/www/us/en/ark/products/192452/intel-xeon-gold-6210u-processor-27-5m-cache-2-50-ghz.html.

[104] R. Bhamra, S. Dani, and K. J. I. J. o. P. R. Burnard, "Resilience: the concept, a literature review and future directions," vol. 49, no. 18, pp. 5375-5393, 2011.

[105] P. Martin-Breen and J. M. Anderies, "Resilience: A literature review," 2011.

[106] S. SECTOR and O. ITU, "ITU-Tfg-DR&NRR," 2014.

[107] D. J. T. S. Tipper, "Resilient network design: challenges and future directions," vol. 56, no. 1, pp. 5-16, 2014.

[108] H. Zhang, J. Zhang, W. Bai, K. Chen, and M. Chowdhury, "Resilient datacenter load balancing in the wild," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, 2017, pp. 253-266.

[109] J.-P. Vasseur, M. Pickavet, and P. Demeester, *Network recovery: Protection and Restoration of Optical, SONET-SDH, IP, and MPLS*. Elsevier, 2004.

[110] W. Ahmad, O. Hasan, U. Pervez, J. J. J. o. N. Qadir, and C. Applications, "Reliability modeling and analysis of communication networks," vol. 78, pp. 191-215, 2017.

[111] D. Serpanos and H. Meleis, "Communication systems for high-speed networks," in *EFOC LAN*, 1992: EFOC & N, pp. 278-281.

[112] L. Lipsky, *Queueing Theory: A linear algebraic approach*. Springer Science & Business Media, 2008.