

Future PON Data Centre Networks

Ali Abdullah Hammadi

*Submitted in accordance with the requirements for the degree of
Doctor of Philosophy*

The University of Leeds

School of Electronic and Electrical Engineering

August 2016

The candidate confirms that the work submitted is his 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 2 of the thesis has appeared in publications as follows:

[1] A. Hammadi and L. Mhamdi, "Review: A survey on architectures and energy efficiency in Data Centre Networks," Computer Communication. vol. 40, pp. 1-21, 2014".

My contribution: studied and surveyed most data centre architectures designed and implemented to date with special emphasis on energy efficiency techniques for the design of green data centres.

Dr. Lotfi Mhamdi: Helped with the preparation of paper.

This paper has received high citations in the last two years (57 citations) and is one of the most downloaded papers in the Elsevier computer communication journal for the last two years.

The work in Chapter 3 and 4 of the thesis has appeared in publications as follows:

[2] J. M. H. Elmirghani, Hammadi, A. and El-Gorashi, T.E., "Passive optical based Data Centre Networks", Patent filed on 26 November 2014 and published on 2nd of June 2016.

My contribution: presented and compared five novel data centre architectures based on optical passive devices, carried a benchmark study for cost and power consumption to compare with designs such as Fat- Tree and BCube and prepared documentation describing the designs.

Professor Elmirghani: Originator of the idea of introducing PONs in data centre architecture, proposed the first Tree-based PON data centre design and helped with the preparation of documentation.

Dr. Taisir: Reviewed the designs and helped with the preparation of the documentation.

[3] Hammadi, A. and El-Gorashi, T.E., J. M. H. Elmirghani, "Future PON Data Centre Networks", (to be submitted to IEEE Green Communications Magazine)

My contribution: literature review, described and compared the five PON designs, proposed techniques for intra and inter racks communications, demonstrated results for benchmark study for cost and power with option3 design with Fat-Tree and BCube designs.

Professor Elmirghani: originator of the idea and helped with the preparation of paper.

Dr. Taisir: Helped with the preparation of paper.

The work in Chapter 5 of the thesis has appeared in publications as follows:

[4] Hammadi, T. E. H. El-Gorashi, and J.M.H. Elmirghani, "High Performance AWGR PONs in Data Centre Networks," IEEE International Conference on Transparent Optical Networks, Hungary, 2015.

My contribution: described the architecture of PON data centre with tuneable lasers, developed MILP model to optimise the interconnection fabric in terms of wavelength routing and assignment within the PON cell. Demonstrated different PON cell sizes with 4 and 8 racks and shown the model results for wavelength routing and assignment.

Professor Elmirghani: originator of the idea and helped with the preparation of paper.

Dr. Taisir: reviewed the model, results and helped with the preparation of paper.

The work in Chapter 6 of the thesis has appeared in publications as follows:

[5] Hammadi, T. E. H. El-Gorashi, and J.M.H. Elmirghani, "Energy-Efficient Software-Defined AWGR-Based PON Data Centre Network" IEEE International Conference on Transparent Optical Network, Italy, 2016.

My contribution: described the idea of introducing software defined networking in PON data centres, developed and described a MILP model for energy-efficient routing among multiple PON cells examining different rates.

Professor Elmirghani: originator of the idea and helped with the preparation of paper.

Dr. Taisir: reviewed the MILP model and results and also helped with the preparation of paper.

The work in Chapter 7 of the thesis has appeared in publications as follows:

[6] Hammadi, Musa Mohammad, T. E. H. El-Gorashi, and J.M.H. Elmirghani, "Resource Provisioning for AWGR-Based PON Cloud Data Centre Network" IEEE 21st European Conference on Network and Optical Communications (NOC), Portugal, 2016.

My contribution: developed a MILP model for resource provisioning in cloud PON data centre to optimise power consumption and delay for different applications that can be hosted in data centre.

Professor Elmirghani: originator of the idea and helped with the preparation of paper.

Dr. Taisir: Reviewed the MILP model and helped with the preparation of paper.

Musa Mohammad: Reviewed and improved the MILP model.

[7] Hammadi, A., Musa Mohammad, El-Gorashi, T.E., and J. M. H. Elmirghani, "Green AWGR-Based PON Data Centre Networks", (to be submitted to IEEE Journal of Lightwave Technology)

My contribution: developed a MILP model for resource provisioning in cloud PON data centres to minimise power consumption and delay (individually and jointly) for different applications that can be hosted in a data centre. Developed an algorithm for minimisation of both delay and power

consumption, compared the algorithm developed with random placement and Best Fit Decreasing algorithms.

Professor Elmirghani: originator of the idea, reviewed and enhanced MILP and helped with the preparation of paper.

Dr. Taisir: Reviewed and enhanced the MILP model and algorithm and helped with the preparation of paper.

Musa Mohammad: reviewed and improved the MILP model.

The work in Chapter 8 and 9 of the thesis has appeared in publications as follows:

[8] Hammadi, T. E. H. El-Gorashi, Musa Mohammad and J.M.H. Elmirghani, "Server-Centric PON Data Centre Network" IEEE International Conference on Transparent Optical Network, Italy, 2016.

My contribution: described the design of server centric PON data centre. Presented and discussed the MILP results for energy efficient routing with resource provisioning (VM placement) in server centric PON data centres.

Professor Elmirghani: originator of the idea and helped with the preparation of paper.

Dr. Taisir: Reviewed and enhanced the MILP model, and helped with the preparation of paper.

Musa Mohammad: Reviewed the work and helped with paper preparation.

[9] Hammadi, T. E. H. El-Gorashi and J.M.H. Elmirghani, "On the Architecture of Energy-Efficient Server-Centric PON Data Centre Networks",

(to be submitted to IEEE Journal of Optical Communications and Networking)

My contribution: literature review, described the design of a server centric PON data centre and presented a benchmark study to demonstrate power savings of the proposed design compared to the three tier conventional data centre architecture. Developed two MILP models one for energy efficient routing and the second for resource provisioning with energy-efficient routing. Developed two algorithms to mimic the behaviour of the two developed MILP models.

Professor Elmirghani: originator of the idea, reviewed and enhanced the MILP and helped with the preparation of paper.

Dr. Taisir: Reviewed and enhanced the MILP models and the two algorithms and helped with the preparation of paper.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Acknowledgements

First and foremost, I would like to sincerely acknowledge my supervisor, Professor Jaafar Elmirghani for his guidance, patience, support and assistance through my PhD journey.

Also I would like to acknowledge Dr. Taisir Elgorashi, my academic co-supervisor for her effort throughout the journey of my PhD, for her useful discussion and advice.

I am very grateful and thankful to my beloved family back home, my mother, my father, my wife and my children. I don't have enough words to thank them for supporting me in all possible ways. I hope I made them proud.

My gratitude goes to my country Kuwait for fully funding my PhD. I was very fortunate to be awarded this scholarship; my dream of a PhD would have never been possible.

Finally, Many thanks to my colleagues in the I3S group in the School of Electrical and Electronic Engineering. I was very fortunate to have the opportunity to meet and collaborate with many outstanding students.

Ali Hammadi

Abstract

Significant research efforts have been devoted over the last decade to design efficient data centre networks. However, major concerns are still raised about the power consumption of data centres and its impact on global warming in the first place and on the electricity bill of data centres in the second place. Passive Optical Network (PON) technology with its proven performance in residential access networks can provide energy efficient, high capacity, low cost, scalable, and highly elastic solutions to support connectivity inside modern data centres.

Here, we focus on introducing PONs in the architecture of data centres to resolve many issues in current data centre designs such as high cost and high power consumption resulting from the large number of access and aggregation switches needed to interconnect hundreds of thousands of servers. PONs can also overcome the problems of switch oversubscription and unbalanced traffic in data centres where PON architectures and protocols have historically been optimised to deal with these problems and handle bursty traffic efficiently.

In this thesis, five novel PON data centre designs are proposed and compared to facilitate intra and inter rack communications. In addition to maximising the use of only passive optical devices, other challenges have to be addressed by these designs including off-loading the inter-rack traffic from the Optical Line Terminal (OLT) switch to avoid undesired power consumption and delays, facilitating multi-path routing, and reducing or eliminating the need for expensive tuneable lasers. The Scalability of the

proposed architectures in terms of efficiently accommodating hundreds of thousands of servers is discussed. CAPEX and energy consumption of the proposed architectures are also investigated and savings compared to conventional architectures, such as the Fat-Tree and BCube, are demonstrated. The Routing and Wavelength Assignment (RWA) in intra and inter rack communication and the resource provisioning needed to cater for different applications that can be hosted in data centre are optimised using Mixed Integer Linear Programming (MILP) models to minimise the PON designs power consumption. Furthermore, real-time energy-efficient routing and resource provisioning algorithms are developed. In addition to optimising the power consumption, delay is also considered for the delay sensitive applications that can be hosted in the proposed data centre architectures. To further reduce power consumption and overcome issues related to link oversubscription and multi-path routing, Software Defined Network (SDN) based design is proposed.

Table of Contents

Acknowledgements	i
Abstract	ii
Table of Contents	iv
List of Figures.....	x
List of Tables.....	xvi
List of Abbreviation.....	xviii
1 Introduction	1
1.1 Research objectives.....	5
1.2 Original contributions	7
1.3 Related publications.....	9
1.4 Organisation of the thesis	10
2 Review on data centre architectures.....	14
2.1 Introduction	14
2.2 Conventional Data Centre Architecture and Challenges	14
2.2.1 Conventional data centre design	14
2.2.2 Conventional data centre challenges.....	15
2.3 Data centre architectural evolution.....	17
2.3.1 Switch centric data centre architectures	18
2.3.2 Server-centric data centres.....	25

2.3.3	Optical data centres	29
2.4	Comparison and Discussion of DCN Architectures	36
2.5	Techniques for Energy Efficient Data Centres	38
2.5.1	Virtualization	39
2.5.2	Energy-Aware Routing.....	41
2.5.3	Dynamic Voltage and Frequency Scaling (DVFS)	42
2.5.4	Rate adaptation in networks	43
2.5.5	Dynamic Power Management (DPM)	43
2.5.6	Energy Aware Scheduling.....	44
2.6	Summary.....	45
3	Review of Passive Optical Network (PON) in access network (FTTx) ...	47
3.1	Introduction	47
3.2	General overview of PON devices	47
3.2.1	Arrayed Wave guides grating (AWG).....	48
3.2.2	Fibre Brag Grating	50
3.2.3	Splitters/couplers	51
3.2.4	Star reflector	54
3.3	PON deployment in access networks.....	54
3.4	Categorisation of FTTx PONs based on Media Access Control (MAC) protocol	58
3.5	Summary.....	59
4	Proposed PON architectures for data centre networks	60

4.1	Introduction	60
4.2	PON Emergence in future data centre architectures.....	61
4.2.1	The need for PONs in the data centre interconnection design....	63
4.2.2	Related work.....	65
4.3	PON capability study for data centres	66
4.4	Study of traffic patterns in FTTx and data centres.....	69
4.5	Proposed PON architectures for future data centres.....	70
4.5.1	Design Options 1 and 2: PON designs for data centres adopted from Fttx deployments	71
4.5.2	Design Option 3: PON designs for data centres with servers equipped with tuneable lasers	75
4.5.3	Design Option 4: PON designs for data centres with few tuneable lasers	77
4.5.4	Design Option 5: PON designs for data centres without tuneable lasers	79
4.6	Comparison and discussion	80
4.7	Summary.....	82
5	Energy and cost efficient AWGR-based PON data centre architecture (Design-Option 3).....	84
5.1	Introduction	84
5.2	Architecture of proposed AWGR-based PON Cell	84
5.3	Network optimisation through Mixed Integer Linear Programming (MILP)	87

5.4	MILP model for wavelength routing and assignment within a PON cell	90
5.5	The wavelength routing and assignment results within a PON cell..	95
5.6	Power consumption benchmarking of PON data centre design	104
5.6.1	Fat-Tree data centre architecture	104
5.6.2	BCube data centre architecture	106
5.6.3	PON data centre architecture	107
5.6.4	Comparison and discussion.....	108
5.7	Cost-based comparison between PON DC with Fat-tree and BCube architectures	110
5.8	Summary.....	112
6	Reduced Inter cell oversubscription through energy efficient software defined AWGR PON based Network	114
6.1	Introduction	114
6.2	Modified architecture for reduced over subscription in Inter Cell Communication	115
6.3	Inter-cell wavelengths routing and assignment	117
6.3.1	MILP model for inter cell wavelength routing and assignment..	118
6.3.2	Results and discussion	122
6.4	Energy efficient software defined AWGR-based PON data centre network	129
6.4.1	MILP model.....	130

6.4.2 Results and discussion	135
6.5 Summary.....	138
7 Resource provisioning for AWGR PON based cloud data centre	139
7.1 Introduction	139
7.2 MILP model for Energy Aware Resource Provisioning in PON data centre	140
7.3 Results and discussions.....	146
7.4 Clus-BF Greedy algorithm for VM placement in PON data centre .	151
7.5 Evaluation of the CLus-BF algorithm.....	153
7.6 Summary.....	158
8 Energy efficient server centric PON data centre network.....	160
8.1 Introduction	160
8.2 Traffic locality study for PON deployment in data centre networks	161
8.3 The architecture of the server-centric PON cloud data centre	162
8.4 Power consumption benchmark study of server-centric PON design against the 3-tier conventional DCN.....	165
8.5 MILP model for energy aware routing in PON data centre.....	169
8.6 Results and discussions.....	180
8.7 Energy Aware Routing Heuristic (EAR) for Server Centric PON Data Centre Architecture	189
8.8 Summary.....	191

9	Energy efficient routing with virtual machine placement in server- centric PON data centre	193
9.1	Introduction	193
9.2	MILP model for Energy Aware Routing and VM placement in Server- centric PON data centre design	193
9.3	Results and discussions.....	198
9.4	Energy Aware VM Placement in PON Data Centre Heuristic.....	204
9.5	Summary.....	209
10	Conclusions and future directions	210
10.1	Conclusions	210
10.2	Future work	214
	References.....	217

List of Figures

Figure 2.1 Conventional data centre with traffic flow classification [37].....	15
Figure 2.2. Fat-Tree Topology with $k=4$ [27].....	19
Figure 2.3. VL2 Topology [34].....	20
Figure 2.4. Portland Topology [28].....	22
Figure 2.5. Juniper one tier Q-fabric [39].....	24
Figure 2.6. BCube [30].....	26
Figure 2.7. DCell Topology [29].....	27
Figure 2.8. Ficonn Topology [31].....	28
Figure 2.9. Helios [33].....	30
Figure 2.10. C-through [35].....	32
Figure 2.11. Hybrid electro-WDM PON [43].....	34
Figure 2.12. Petabit 4 x 4 switch fabric [32].....	35
Figure 2.13. Summary of well-known data centre architectures.....	38
Figure 3.1. 4x4 Arrayed waveguide grating [64].....	48
Figure 3.2. Applications of AWG devices. (a) Mux/Demux (b) drop/add multiplexer (c) full interconnection [64].....	50
Figure 3.3. Fibre Bragg grating [64].....	50
Figure 3.4. Total internal reflection (TIR) [66].....	52
Figure 3.5. Optical tunnelling or frustrated TIR [66].....	52
Figure 3.6. Prism splitter [66].....	53
Figure 3.7. Cascaded 1x2 couplers/splitters to produce 1x8 coupler [36] ...	54
Figure 3.8. Star reflector [36].....	54
Figure 3.9. (a) TDM PON architecture (b) WDM PON architecture.....	56

Figure 4.1. (a) OLT card with 8 ports to connect 8 PON Cells (each PON Cell hosts 128 servers), (b) OLT Chassis with 16 cards to connect 128 PON Cells, (c) OLT switch with 8 chassis, (d) Five OLT switches	67
Figure 4.2. Total Processing capability and memory capacity for (a) PON Cell, (b) PON OLT Card, (c) PON OLT Switch Chassis, and (d) PON OLT Switch with 8 Chassis for data centre hosting Intel core 980x servers, with 8 GB RAM memory and 147,600 MIPS processing.....	68
Figure 4.3. 3-Tier data centre traffic flow classification.....	70
Figure 4.4. (a) Option design 1: TDM-PON DCN architecture, (b) Option design 2: hybrid WDM-TDM PON DCN with multi-carrier generator	71
Figure 4.5. Schematic of 10x10 passive polymer optical backplane [44]	74
Figure 4.6. Option-3 PON Data centre interconnection for PON Cell employing servers equipped with tuneable lasers and utilizing AWGR and star coupler/splitters for fabric interconnection	75
Figure 4.7. Design option 4: PON DCN with few servers equipped with tuneable lasers	77
Figure 4.8. Design option 5: PON DCN with no tuneable lasers	79
Figure 5.1. Architecture of proposed AWGR based PON cell	86
Figure 5.2. Architecture of the Optical line terminal (OLT) and tuneable ONU [36]	86
Figure 5.3. (a) Architecture of proposed PON data centre with tuneable lasers (b) Obtained MILP configuration for 4x4 AWGRs interconnection for wavelength routing (c) MILP obtained wavelength assignment for PON to PON and PONs to OLT communication	94

Figure 5.4. Architecture of proposed PON data centre with tunable lasers for 8 PONs groups.....	99
Figure 5.5. Obtained MILP configuration for 8x8 AWGRs interconnection for wavelength routing and assignment for the 8 PON groups design.	100
Figure 5.6. MILP obtained wavelength assignment for PON to PON and PONs to OLT communication for the 8 PON groups design.....	100
Figure 5.7. Worse-case server share of resources against different sizes of PON cell for 4,8,12, and 12 PON groups.....	101
Figure 5.8. Number of wavelengths needed with respect to the number of PON groups in a PON cell.....	101
Figure 5.9. architecture of PON data centre with tuneable lasers for intra and inter rack communication through passive AWGRs.....	102
Figure 5.10. MILP obtained wavelength interconnection assignment for inter/intra communication	103
Figure 5.11. Obtained MILP configuration for 8x8 AWGRs interconnection for wavelength routing and assignment to provision inter and intra rack communication.....	103
Figure 5.12. Fat-Tree data centre topology with $k=4$	105
Figure 5.13. BCube data centre topology (BCube1) with $n=4$ and $k=1$	106
Figure 5.14. Power consumption saving for PON architecture against BCube and Fat-Tree architectures	110
Figure 5.15. CAPEX saving for PON architecture against BCube and Fat-Tree.....	112
Figure 6.1. Upper level connectivity for the decentralized design option 3 architecture.....	116
Figure 6.2. Upper-level connectivity for SDN based option-3 architecture	117

Figure 6.3. 4x4 AWGR input/output numbering diagram.....	123
Figure 6.4. Energy-Efficient Bandwidth Allocation through reconfiguration and grouping.....	130
Figure 6.5. Power consumption evaluation for SDN over decentralized designs	137
Figure 6.6. Blocking percentages of SDN minimised energy model against the decentralized design.....	138
Figure 7.1. The average delay provisioning different sets of VMs; under the three models and the developed algorithm Clus_BF	148
Figure 7.2. Total number of used servers examining three sets of VMs; 20, 40, and 60 for the three objective functions and the developed algorithm Clus_BF	148
Figure 7.3. Average servers' utilization examining three sets of VMs; 20, 40, and 60 for the three objective functions and the developed algorithm Clus_BF.....	149
Figure 7.4. The total power consumption different sets of VMs; under the three models and the developed algorithm Clus_BF	149
Figure 7.5. Clus_BF Algorithm for Multi-constraint resource provisioning in AWG PON Cell.....	153
Figure 7.6. Total number of used servers for the random, best fit decreasing, best effort best fit, and Cluster best fit algorithms.....	155
Figure 7.7. Total power consumption of traffic flow reduction of the Cluster- BF, Random and Best Fit Decreasing algorithms.....	155
Figure 7.8. The percentage of traffic flow reduction of the Cluster-BF, Random and Best Fit Decreasing algorithms	156

Figure 7.9. Average servers' utilization of traffic flow reduction of the Cluster-BF, Random and Best Fit Decreasing algorithms.....	156
Figure 7.10. Average delay of traffic flow reduction of the Cluster-BF, Random and Best Fit Decreasing algorithms	157
Figure 8.1. Proposed PON Cell design with no tuneable lasers.....	163
Figure 8.2. Architecture of an ONU in access network [94].....	167
Figure 8.3. Proposed architecture of an ONU for PON data centre. GbE switch is passive and optional for the case where multiple servers to be connected by one ONU	168
Figure 8.4. Total power consumption for cases where no threshold for OLT are enforced and server's threshold is varied for the different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack.....	183
Figure 8.5. Total power consumption for a PON cell for cases where no threshold for servers are enforced and OLT threshold is varied for the different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack.....	184
Figure 8.6. Average path delay for cases where no threshold for OLT are enforced and server's threshold is varied for the different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack.....	188
Figure 8.7. Average path delay for cases where no threshold for servers are enforced and OLT threshold is varied for different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack	189

Figure 8.8. Heuristic flow chart for Energy Minimised Routing (EAR-PON) for a PON cell	191
Figure 9.1. The modelled server centric PON data centre architecture....	199
Figure 9.2. PON cell power consumption for different sizes of received clients' requests for the two objectives cases of energy aware (EA) and non-energy aware (NEA) of VMs routing and placement	201
Figure 9.3. Servers' utilisation showing processing and relay utilisation for 20 VMs for the EA objective	201
Figure 9.4. Servers' utilisation showing processing and relay utilisation for 30 VMs for the EA objective	202
Figure 9.5. Servers' utilisation showing processing and relay utilisation for 40 VMs for the EA objective	202
Figure 9.6. Servers' utilisation showing processing and relay utilisation for 50 VMs for the EA objective	203
Figure 9.7. Location and relay utilisation of servers selected for routing request for the non-energy aware objective (random placement)...	203
Figure 9.8. Average servers' utilization for different sizes of received clients' requests for the two objectives cases of energy aware (EA) and non-energy aware (NEA) of VMs routing and placement.....	204
Figure 9.9. Number of selected servers for the different cases of received requests for the two objectives cases of energy aware (EA) and non-energy aware (NEA) of VMs routing and placement.....	204
Figure 9.10. The flow chart for the Modified Best-Fit Decreasing (MBFD) algorithm for energy aware VM placement in a PON Cell	208
Figure 9.11. Power consumption results for MILP modelled network and MBFD and MBF algorithms	209

List of Tables

Table 3.1 PON access network classifications.....	56
Table 4.1. Comparison between the proposed technologies for Intra-rack communication	72
Table 4.2. Comparison of the five PON cell design options	81
Table 5.1. Cost Breakdown for the networking devices used in DC architectures.....	111
Table 6.1. MILP obtained wavelength assignments for uplink connections from PONs to OLTs	123
Table 6.2. MILP obtained wavelength assignment for down link connections from OLTs to PONs	124
Table 6.3. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-1 to OLT switches	125
Table 6.4. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-2 to OLT switches	126
Table 6.5. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-3 to OLT switches.	127
Table 6.6. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-4 to OLT switches	128

Table 6.7. Input parameters used in the model	137
Table 7.1. Input data for the model	146
Table 8.1. Breakdown of the power consumption of networking equipment of 3-tier data centre architecture provisioning connectivity for 5120 servers.....	166
Table 8.2. Power consumption for main components of 10G ONU for 20km reach [94]	168
Table 8.3. Breakdown analysis for power consumption for the proposed data centre architecture to support 10G and for provisioning 5120 servers	168
Table 8.4. Input data for the model	181
Table 9.1. Input data for the model	199

List of Abbreviation

ACPI	Advanced Configuration and Power Interface
AWGR	Arrayed Waveguide Grating Router
BFD-BP	Best Fit Decreasing Bin Packing
BSR	Basic Source Routing
BSR	Basic Source Routing
DCN	Data Centre Network
DENS	Data centre Energy efficient Network aware Scheduling
DPM	Dynamic power management
DVFS	Dynamic Voltage and Frequency Scaling
EA	Energy Aware
EAR	Energy Aware Routing
ECMP	Equal Cost Multipath routing protocol
EPA	Environmental Protection Agency
FBT	Fused Bi-conical Taper
FTTx	Fibre to the Premises
IaaS	Infrastructure as a Service
ITU	International Telecommunication Union
MAC	Media Access Control
MBFD-BP	Modified Best Decreasing Bin Packing
MEMS	Micro Electro Mechanical Switches
MILP	Mixed Integer Linear Programming
MoCA	Multimedia Over Coax
NEA	Not Energy Aware

NIC	Network Interface Card
OLT	Optical Line Terminal
ONU	Optical Network Unit
OSPF	Open Shortest Path First
PaaS	Platform as a Service
PCIe	Peripheral Component Interconnect Express
PON	Passive Optical Network
PUE	Power Usage Efficiency
QoS	Quality of Service
SaaS	Software as a Service
SDN	Software Defined Network
SDN	Software Defined Network
SLIC	Subscriber Line Interface
SOA	Semiconductor Optical Amplifier
STP	Spanning Tree Protocol
TAR	Traffic Aware Routing algorithm
TOR	Top of Rack
TWC	Tuneable Wavelength Converters
VLB	Valiant Load Balancing
VM	Virtual Machine
WoA	Wake Up on Arrival

1 Introduction

Recent years have witnessed an unprecedented growth in services and applications housed in modern data centres, such as web-search, scientific computations, social networks, file storage and distributed files systems. Today's data centres host hundreds of thousands of servers, interconnected via switches, routers and high-speed links, making the choice of networking architecture within data centre of premium importance as it impacts data centre scalability, cost, fault-tolerance, agility and power consumption.

Significant research efforts have been devoted over the last decade to design energy-efficient core networks [1-22]. However, major concerns are still raised about the power consumption of data centres and its impact on global warming in the first place and on the electricity bill of data centres in the second place. The US Environmental Protection Agency (EPA) has reported that power usage of data centres in the US has doubled between 2000 and 2006 to nearly 61 billion kilowatt-hours, accounting for 1.5% of the US total electricity demand [23].

Given the steadily increasing number of servers and the exponentially growing traffic inside data centres, conventional data centre networking architectures suffer from performance limitations such as links oversubscription and inefficient load balancing [24]. Also conventional data centre architectures are based on expensive and power hungry devices such as access switches, aggregation switches and core switches,

accounting for 20% of the total power consumption of a data centre [24]. These limitations have stimulated the search for new low cost, scalable, energy-efficient architectures to efficiently serve the increasing demands [24].

The choice of a DCN solution to address one challenge impacts and often limits the alternatives available to address other issues. Furthermore, DCNs are deployed for various applications, and the solutions differ such as between enterprise DCNs and cloud-service DCNs [25]. Irrespective of the DCN type, various common challenges for the design of DCNs have been observed at different levels, including: i) the architecture of the DCN and its topology, ii) minimising energy and keeping the DCN power budget manageable while providing virtualisation, network load management and scheduling, iii) congestion handling in DCNs including congestion notification and avoidance. A typical challenge is the problem of TCP incast, iv) routing in DCNs with the provision of efficient and cost-effective routing mechanisms such as multipath routing.

The architecture of a DCN, or its topology, directly reflects on its scalability, cost, fault-tolerance, agility and power consumption [26]. Conventional DCNs have been designed using a tree-like topology. A typical example of this topology is the three-tier topology proposed by [26] where the tree's leaves (servers) are connected to Top-of-Rack (ToR) switches and these (ToR) switches are connected to aggregation switches which are in turn connected to core routers at the root of the tree. This topology suffered numerous drawbacks of scale, capacity, reliability, utilisation and power budget [24]. As a result, efforts have been dedicated to address some of the

above problems encountered in the tree-based DCN topology and various DCN architectures have appeared as a result [27-34]. These architectures can be classified as server-centric and switch-centric and/or based on their infrastructure technologies such as electronic versus optical DCNs. Energy efficiency is a central issue in modern data centres. DCNs are typically high-capacity networks that are tuned for maximum performance, making them extremely power hungry.

Optical switching and networking for data centres have recently been proposed to establish high speed server-to-server connections in modern data centres. However, these solutions still suffer from a number of problems that can be overcome by PONs. The PON solutions we are proposing are an all optical data centre solution as they enable wavelengths to be allocated directly to end-to-end server connections not only within the same rack, but also between different racks in the data centre.

PON solutions are scalable: This is readily proven in the combination of core and access networks that are able to connect easily tens of millions of homes. PONs achieve scalability due to their cellular architecture. A PON cell may connect say 256 servers and many cells can then provide coverage of a small data centre or a large data centre with say 1 million servers. Scalability is assured at very low cost compared to optical switches, in a similar fashion to wireless cellular being scalable. Optical switches are not naturally built in large scale and cascading them can suffer from failures and can involve long paths. The cellular design also allows wavelength reuse without the need for expensive tuneable wavelength convertors.

PON solutions enable efficient bandwidth utilisation naturally: PONs can assign a wavelength to large “elephant” flows between servers. They can also allocate a time slot in their TDM-WDM structure to accommodate “mice” flows. This enables better wavelength / bandwidth utilisation compared to optical switches that are typically only able to allocate a whole wavelength to a flow, so optical switches deal well with “elephant flows” using optical circuits in a fashion similar to PONs, but can waste a wavelength on a “mouse flow”. Optical packet switching can use fast expensive optical switches with nanosecond switching time to accommodate mice flows. Optical switches have a trade-off between scale and switching speed typically.

PON solutions provide better resilience: Compared to optical switching in data centres, PONs have a cellular structure and a failure in a cell does not affect other cells and may only affect part of the cell. Failure in cascaded optical switches can be disruptive to the whole data centre.

PONs are more energy efficient than other all optical networking solutions: They rely on passive optical networks, and have lower control compared to optical switching / optical networking solutions. Such optical networking solutions are Helios [33], c-through [35], and Petabit [32].

Our version of PONs has many additional advantages, for example our proposed solutions i) enable direct server to server communication within the rack through reflected wavelengths or by deploying technologies such as the terabit capacity polymer passive optical backplane [36] , ii) enable server to server communication direct through our cellular architectures without going

to the “central office” or OLT switch. Note that standard PONs are used in residential access and do not provide a simple route between its nodes as home-to-home communication is not typical, iii) achieve load balancing by providing more than a single route between servers, where typically home to home communication has a single path through the OLT switch, and iv) overcome link oversubscription by allowing PON cells to access multiple OLT ports.

PONs have both advantages and disadvantages. One of the disadvantages of PONs is the high signal power loss resulted from splitting and coupling. This accumulated loss coming from the high split ratio can result in the need of high power lasers which negates the objective of the design. However, our design overcome this issue by introducing more PON groups with small number of servers connected to the couplers. Another issue is the broadcast nature of PONs where all servers connected to the same coupler receives all the traffic destined to any server connected to the same coupler. This issue is also reduced by reducing the number of servers connected to the same coupler.

1.1 Research objectives

The primary research objectives are as follows:

- To study the multi-tier conventional data centre architecture and evaluate the main challenges facing conventional DCNs which led to the development of other architectures. Survey the most recent advances in DCN architectures introduced to overcome the limitations

of existing conventional DCN as classified into many categories; electronic switch-centric, electronic server-centric, hybrid electro-optical, full optical, and hybrid Electro with WDM Passive Optical Network (PON) technology.

- To study different energy saving approaches and industry adopted techniques for energy efficient data centres.
- To study PON deployment in access networks to make use of its attractive proven performance in residential access networks to overcome the main limitations in current designs of data centres.
- To propose a number of scalable, energy-efficient, low cost, and high capacity designs for future PON data centre architectures relying mostly on PON devices to (i) manage different types of traffic that can co-exist within data centres for inter-rack and intra-rack communication among servers, and (ii) facilitate multipath routing.
- To investigate the use of a centralised Software Defined Network (SDN) control and management system to coordinate and arbitrate the channel access for communication through the OLT links with PONs via wavelength reconfiguration and energy-efficient grooming.
- To study the impact of SDN based architecture design against the decentralised design with respect to energy saving and blocking.
- To optimise the routing and wavelength assignments for inter and intra PON cell communications, and optimise resource provisioning for cloud applications in the proposed data centre architectures.

- To develop real-time heuristics for energy-efficient routing and virtual machine (VM) resource provisioning for the proposed data centre architectures.

1.2 Original contributions

The main contributions are summarised as follows:

1. Published a survey paper on the architectures and energy efficiency techniques of data centre networks. This can be used as a guide to academia and industry to understand and assess the evolution of data centre architecture designs and the main adopted energy efficiency improvement techniques in current data centres.
2. Filed and published a patent on the five proposed energy efficient novel architecture designs for future data centre interconnections using mostly low cost and low energy readily available passive optical devices to manage intra and inter rack communication. Two of these designs are based on current FTTx deployment technologies, however inter-rack communication has to be processed through the OLT switch. To facilitate high speed interconnection among racks within a PON cell, we proposed an architecture where each server is equipped with an array of photo detectors and tuneable lasers for wavelength detection and selection. Another design is proposed to reduce the need for expensive tuneable lasers by deploying special servers to manage and perform the wavelength conversion needed to support inter-rack traffic connections. The final proposed design is the

most cost efficient as it eliminates the need for tuneable lasers and facilitates high speed interconnection among racks.

3. Proposed three novel techniques and technologies to manage intra-rack communication using fibre Bragg grating, star reflectors, and optical backplane to replace high cost and power electronic access switches.
4. Demonstrated that Design option-3 which is based on AWGR does not suffer the limitations of the normal PON in access network as high per server rate, and multi-path routing can be achieved. Developed a mathematical model for wavelength routing and assignment within the AWGRs PON fabric to facilitate intra and inter cell communication.
5. Achieved low oversubscription ratio for intra PON cell and showed the worse-case per server rate for intra-cell communication can approach rates up to 5 Gb/s.
6. Improved the AWGR based design for more reduction in the deployment cost by managing intra-rack communication through the existing two intermediate AWGRs and the usage of additional hardware such as FBGs, optical backplanes, or star reflectors can be avoided. Developed MILP models and algorithms for energy efficient resource provisioning and routing within a PON data centre for the AWGR based design.
7. Introduced and designed energy-efficient SDN control for AWG PON based data centre architecture to achieve low oversubscription and provide inter cell multi-path routing.
8. Designed a server centric PON data centre architecture to eliminate the need for costly and power hungry devices such as tuneable

lasers. Within the cell of racks, the design give full wavelength connectivity and hence server to server communication within the cell can appear as point to point. The design also employs servers to take part in routing and traffic forwarding. Presented a power consumption benchmark study to compare the conventional 3-tier data centre architecture with the proposed server-centric design and shown that 69% of savings can be achieved.

9. Developed a mathematical optimisation model along with a heuristic for the PON server-centric design for energy efficient routing, examining different inter/intra ratios of traffic flows within the PON cell.
10. Developed a mathematical optimisation model along with a heuristic for the PON server-centric design for energy efficient VM placement.

1.3 Related publications

1. A. Hammadi and L. Mhamdi, "Review: A survey on architectures and energy efficiency in Data Centre Networks," *Computer Communication*. vol. 40, pp. 1-21, 2014".
2. J. M. H. Elmirghani, Hammadi, A. and El-Gorashi, T.E., "Passive Optical Based Data Centre Networks", UK patent, Filed on 26 November 2014, Published 02 June, 2016.

3. A. Hammadi, T. E. H. El-Gorashi, and J.M.H. Elmirghani, "High Performance AWGR PONs in Data Centre Networks," IEEE ICTON, Hungary, 2015.
4. Hammadi, A., El-Gorashi, T.E.H., Musa, M.O.I. and Elmirghani, J.M.H., "Server-Centric PON Data Centre Architecture," Proc IEEE 18th International Conference on Transparent Optical Networks (ICTON), Trento, Italy, July 10-14, 2016
5. Hammadi, A., El-Gorashi, T.E.H., Musa, M.O.I. and Elmirghani, J.M.H., "Energy-Efficient Software Defined AWGR Based Data Centre Network," Proc IEEE 18th International Conference on Transparent Optical Networks (ICTON), Trento, Italy, July 10-14, 2016
6. Hammadi, A., El-Gorashi, T.E.H., Musa, M.O.I. and Elmirghani, J.M.H., "Resource Provisioning for Cloud PON AWGR-Based Data Centre Architecture", 21st European Conference on Network and Optical Communications (NOC), Lisbon, Portugal, June 1-3, 2016.
7. IEEE communication magazine journal titled as "PONs in future data centres". (To be submitted to IEEE Green Communication Magazine)
8. "Energy-Efficient Server-centric PON Data Centre" (to be submitted to IEEE Journal of Light wave Technologies)
9. "Green PON AWGR PON Data Centre Networks" (to be submitted to IEEE Journal of Light wave Technologies)

1.4 Organisation of the thesis

Following the introduction in Chapter 1, the remaining parts of this thesis are organised as follows:

Chapter 2 reviews the evolution of data centre designs and the main shortcomings that appeared in the last decade. Special attention is given to energy saving approaches and the techniques most implemented by industry for the design of green data centres.

In Chapter 3, a review on the Passive Optical Networks (PON) in access networks is presented. The review covers the functionality, construction and main applications of well-known PON devices used in access networks such as arrayed waveguide routers, star couplers, star reflectors, and fibre Bragg gratings. PON deployed architectures, standards, MAC categorisation in Fibre to the Premises (FTTx) access networks are discussed and presented as well.

Chapter 4 discusses the PON capabilities needed to provision connectivity in modern data centres. We compare and classify the traffic patterns between FTTx access network and data centres based on flow and application workload. We introduce three different designs based mostly on optical passive devices to manage intra-rack communication without the need to reach the OLT switch. Then, 5 novel PON designs are presented to furnish connectivity for modern data centres applications. We conclude this chapter with a qualitative comparison between the proposed designs.

Chapter 5 presents a MILP model to optimise the fabric interconnection for AWGR based PON data centre architecture for wavelength routing and assignment. The chapter will discuss the issue of per server rate and the maximum rates that can be achieved for different sizes of PON cells. The work in this chapter also presents a method to make use of existing intermediate AWGRs to manage intra rack communication and avoids the

use of additional hardware. The chapter presents a benchmark study for cost and power consumption to compare the proposed AWGR PON design proposed with two well-known architectures such as the Fat-Tree and BCube.

Chapter 6 tackles the issue of link oversubscription in the AWGR PON based design and provides a solution to provide multi path routing and also reduce oversubscription along with energy through SDN enabled design. Two Mathematical MILP models for energy-efficient SDN enabled architecture and another for inter-cell wavelength routing and assignment are presented in this chapter.

Chapter 7 further investigates the AWGR PON based architecture for cloud application in data centres. A mathematical optimisation model along with algorithms is presented for resource provisioning considering minimisation of power consumption and delay.

In Chapter 8, the server-centric PON design is described in detail. The work in this chapter includes a benchmark study to compare the 3-tier architecture with the proposed server-centric design with respect to power consumption. A developed mathematical optimisation model along with a heuristic for energy efficient routing is presented for the described architecture.

Chapter 9 further investigates the server centric design to develop a mathematical optimisation model along with an algorithm for energy efficient resources provisioning for cloud applications that can be hosted in the described architecture. For optimum energy savings, the MILP model along

with the developed resource provisioning greedy algorithm attempts to optimise the selection of hosting servers, routing paths and relay servers to achieve efficient resource utilisation.

Finally, Chapter 10 concludes the thesis with a summary and tentative future plan.

2 Review on data centre architectures

2.1 Introduction

This chapter provides a detailed survey that covers the most recent advances in DCNs with a special emphasis on the architectures and energy efficiency in DCNs. Data centre architectures are classified as switch-centric and server-centric topologies with underlying electronics and optical technologies. At the end of the data centre architectures section, a qualitative comparison and discussion of the surveyed DCN architectures is provided. The end of this chapter presents a study on different methods for energy saving approaches and industry adopted techniques for energy efficient data centres. Virtualisation, dynamic frequency and voltage scaling, dynamic network management, efficient green routing, green schedulers, and rate adaptation are examples of such energy saving techniques.

2.2 Conventional Data Centre Architecture and Challenges

2.2.1 Conventional data centre design

The classic data centre design architecture [26] consists of switches and routers in two or three tier hierarchal structures as shown in Figure 2.1. The hierarchy in the case of three tiers consists of layer-3 with border routers, layer-2 with aggregation switches, and layer-1 with Top of Rack (ToR) access switches. A ToR switch usually connects 20-40 servers placed in a

rack with 1Gbps links, and for redundancy each ToR is connected with two aggregation switches which in turn connect with the core layers through multiple high speed 10 Gbps links. The aggregation layer provides and manages many functions and services such as spanning tree processing, server to server traffic flow, load balancing, firewall and more. Core routers/switches, running 10 Gbps high-speed links, are at the top of the hierarchy. These are used for traffic going in and out of the data centre. The Core routers/switches also run well-known routing algorithms such as Open Shortest Path First (OSPF) and can load balance traffic between core and aggregation layers [26]. Unfortunately, the hierarchal three tiers DCN structure suffers various issues as will be discussed next.

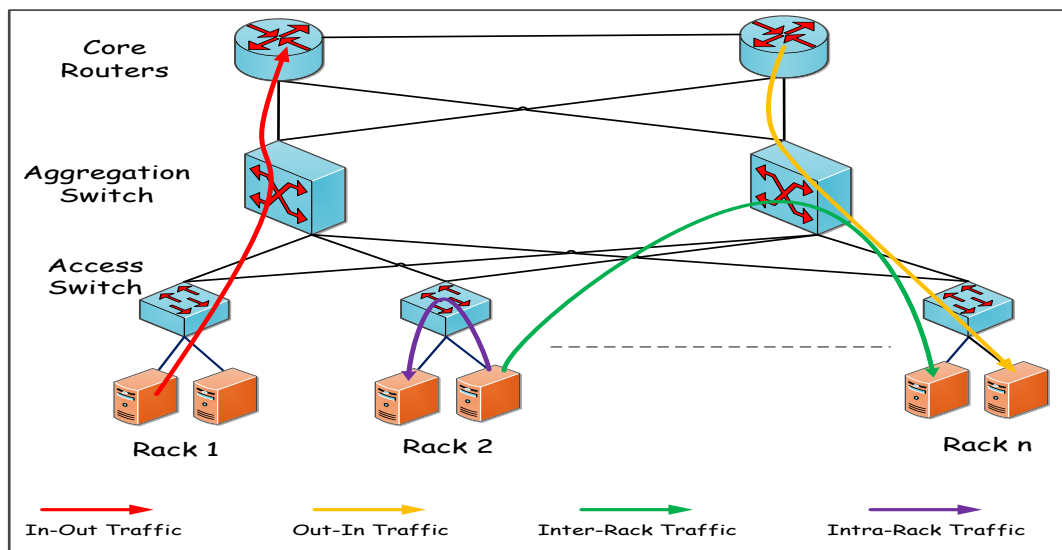


Figure 2.1 Conventional data centre with traffic flow classification [37]

2.2.2 Conventional data centre challenges

Several challenges [24] and issues have appeared with conventional data centres which have led many researches to explore and intensively study

alternative designs and approaches to provide scalable, fault tolerant and efficient data centres. One of the most important performance handicaps that could lead to congestion is oversubscription. Oversubscription is the ratio between the servers' bandwidth and the total uplink bandwidth at the access layer. Moving up to aggregation and core layer, the number of servers sharing the uplinks increases and hence the oversubscription ratio also increases and results in bottlenecks. Oversubscription limits the server to server capacity where the ratio should be 1:1 so hosts can communicate with their full network interface bandwidth. On the other hand, congestion resulting from oversubscription can also lead to overloading switch buffers which will in turn start dropping packets. Hence, another issue arises because of the lack of a mechanism to avoid packet drops at congested switches. Moreover congestion can also occur at switches where simultaneous transmission of packets from multiple senders arrive at the same time. The switch gets overloaded and starts dropping packets leading to TCP timeout and hence a collapse in TCP throughput, known as TCP incast [24].

Other challenges introduced with classical data centre networks include the lack of fault tolerance especially at the upper levels of the tree due to the low physical connectivity. Hardware failures in the core or aggregation layers result in sharp degradation of the overall network performance. Additionally, poor utilisation of resources can occur because of the fact that within the layer-2 domain, the Spanning Tree Protocol (STP) only uses one path even though multiple paths exist. In addition another issue with load balancing

arises since traffic cannot be evenly distributed over paths within core and aggregation layers.

The fast growth of DCNs has focused attention on the issue of power consumption due to the high number of power hungry devices used and cooling systems. Most of these devices are underutilised, as statistics have shown that a typical utilisation of a data centre is only 30% [38]. Hence, dynamic reassignment of resources among servers running on the data centre is an optimal solution to consolidate most jobs on 30% of the servers while being able to shut down the other unused servers and hence save power. The ability to assign any server to any service without considering topology is called Agility.

Many barriers like VLANs, access lists (ACLs), broadcast domains, and Load Balancers (LB) were obstacles that prevented researchers and industry from immediate implementation of VM migration (agility) on conventional data centres. The static network assignment between servers and services in conventional data centres prevent idle servers from being assigned for overloaded services thus resulting in underutilisation of resources. VL2 and Portland data centres as will be explained in subsequent sections where we will address this issue and present methods to overcome it.

2.3 Data centre architectural evolution

Numerous problems in conventional data centres have driven researchers to propose and design various data centre architectures to solve these

issues. Data centres can be categorised mainly in two classes, switch-centric and the server-centric. In switch-centric data centres, switches are the dominant components for interconnection and routing whereas in server-centric data centres, servers with multiple Network Interface Cards (NIC) exist and take part in routing and packet forwarding decisions.

The conventional data centre is a switch-centric design. Other examples of switch-centric include VL2 [34], Portland [28] and Fat-tree [27]. The server-centric topology also has attracted great interest and many designs have been proposed such as Dcell, Bcube, and FiConn. These topologies and designs are based on packet-switched electronic networks, however; hybrid electro-optical packet switches along with full optical solutions were also proposed and implemented for low power consumption and high bandwidth.

2.3.1 Switch centric data centre architectures

In this section, the most well-known switch centric data centre designs such as Fat-tree, Portland, VL2, and one-tier Qfabric [39] are covered. Such designs rely on switches for interconnection and traffic routing. The different design choices in the switch centric class resolve many issues that existed with the conventional data centre. These issues, as shall be explained in subsequent sections, are oversubscription, agility, load balancing and high power consumption.

2.3.1.1 Fat-Tree

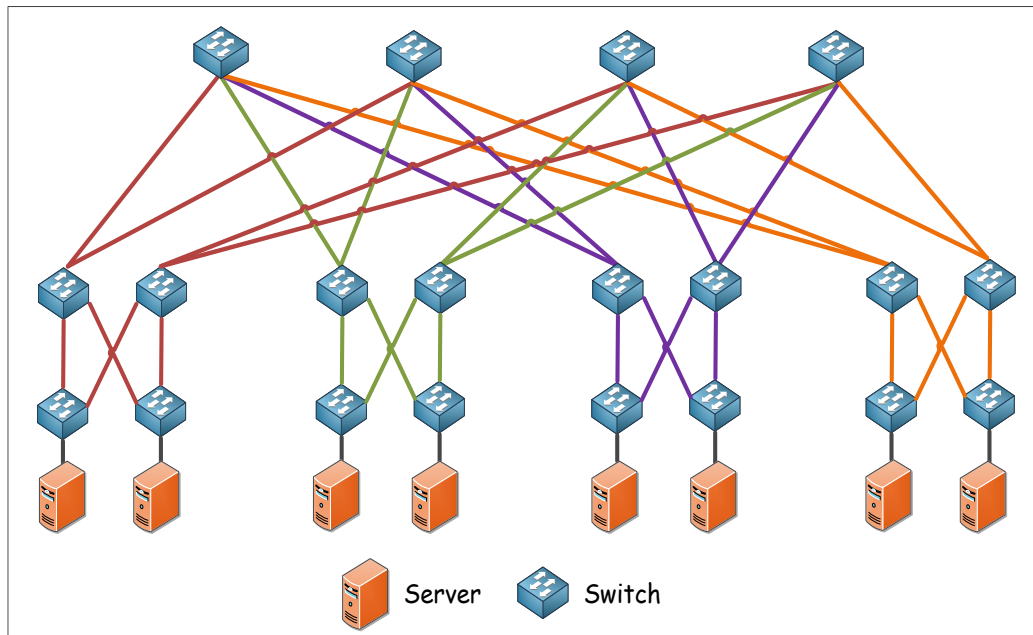


Figure 2.2. Fat-Tree Topology with k=4 [27]

The fat-tree topology, depicted in Figure 2.2, consists of k pods, each of which consist of $k/2$ edge switches and $k/2$ aggregation switches. Edge and aggregation switches are connected as a Clos topology and form a complete bipartite in each pod. Also each pod is connected to all core switches forming another bipartite graph. The Fat-Tree is built with k -port identical switches in all layers of the topology each of which supports $k^3/4$ hosts. Fat-Tree IP addresses are in the form 10:pod:subnet:host. The Fat-Tree topology resolves the issues with oversubscription, costly aggregation and core switches, fault tolerance, and scalability. Fat-Tree established a solid topology for researchers to work on to solve other important issues such as agility through virtualization.

2.3.1.2 VL2

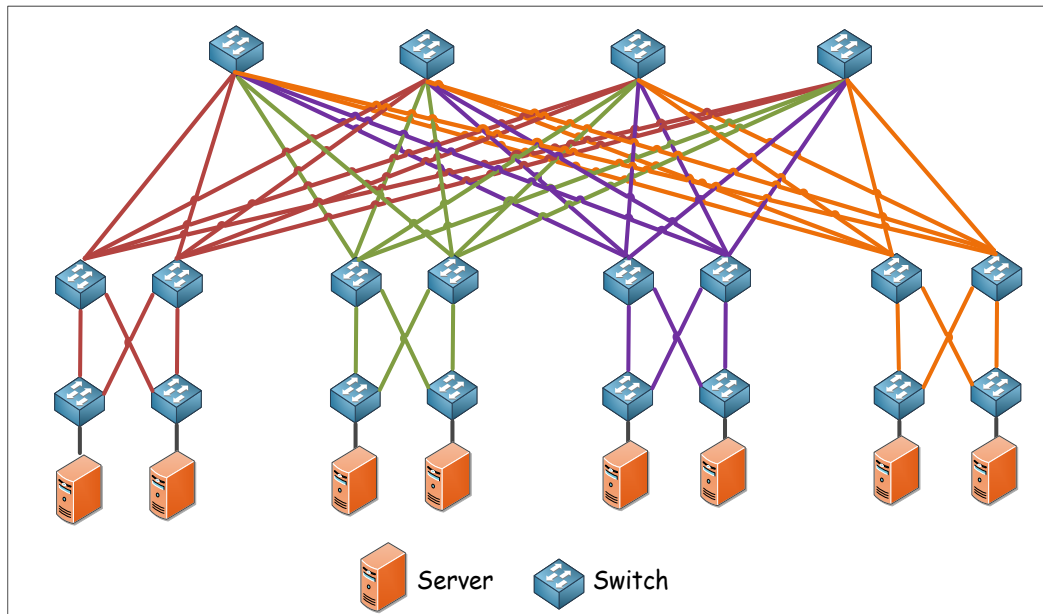


Figure 2.3. VL2 Topology [34]

VL2 was proposed in [34] and is considered as a solution to overcome some of the critical issues in conventional data centres such as oversubscription, agility and fault tolerance. VL2, shown in Figure 2.3, exploits a uniform high capacity fabric from server to server, supports VM migration from server to server without breaking the TCP connection and while keeping the same address. It is very similar to the three-tier architecture DCN proposed by Cisco, except that it implements a Clos topology between the core and aggregation layers to provide multipath and rich connectivity between the two top tiers. The architecture design of the VL2 topology enhances the availability and reliability of the network, especially in the presence of link or hardware failures. VL2 employs Valiant Load Balancing (VLB) to evenly load balance traffic flows over the paths using Equal Cost Multi Path (ECMP). VL2 also employs TCP for end to end congestion control. As additional advantage, VL2 can be easily implemented on low cost existing commodity switches since it uses already existing ECMP for packet forwarding and link state routing for topology updates.

In addition to the fact that VL2 can be implemented on existing hardware and can provide high load balancing, VL2 can support agility among servers. VL2 uses a special flat addressing scheme that separates server names (AA) from their locations (LA), then mapping between the AA and LA can be managed and handled by a directory system. LAs are addresses assigned to switches and interfaces (network infrastructure) while applications are assigned with permanent AAs. AAs remain unchanged no matter how servers' location changes because of the VM migration. Each AA is associated with LA which is the IP of the ToR switch to which the application server is connected. The sender server, before sending, must encapsulate the packets in the outer header with the LA of the destination AA. Once packets arrive at the LA (ToR), the ToR switch encapsulates the packets and sends them to the destination AA. All servers believe that they all belong to the same subnet, hence when any application sends a packet to AA for the first time; the servers' network stack broadcasts an ARP request. The VL2 agent intercepts the request and sends a unicast query message to the directory server which replies with the LA of the ToR switch where packets should be tunnelled.

Virtualisation has been given a great attention by researchers and has become the most widely adopted technique for data centres power saving. Virtualisation is a method to enable services to be moved between servers that have multiple VMs Machines which can serve different applications multiplexed to share one server. Knowing that idle servers consume about 66% of their peak power usage [38] and having in mind that data centre resources are underutilised since the average traffic load accounts for about

30% [38] of its resources; agility can achieve servers statistical multiplexing and give the illusion to services that they are all connected to the same switch. Hence, servers can be placed anywhere within the network and can be assigned to any service. The migration of virtual machines to consolidate workloads on a set of servers and then shutting down underutilised servers can lead to a great power saving in data centres.

2.3.1.3 Portland

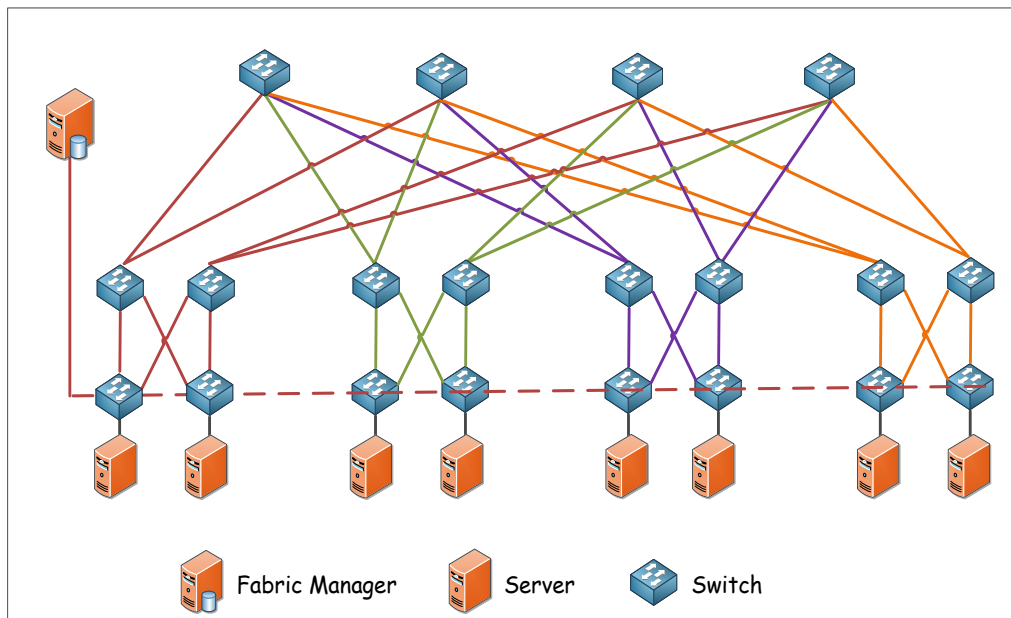


Figure 2.4. Portland Topology [28]

The Portland DCN topology, proposed in [28], is similar to VL2 in that both are based on a Fat-tree network topology. Portland, depicted in Figure 2.4, consists of three layers: edge, aggregation and core. It is built out of low cost commodity switches. Portland and VL2 differ in the way of associating and separating names from locators but both at the end aim at providing agility among services running on multiple machines. Both reduce broadcast by intercepting Address Resolution Protocol (ARP) requests and employ a unicast query through a centralised lookup service. Portland imposes additional

requirements on the switch software and hardware unlike VL2 where implementation only takes place in the servers' network stack. Portland has proposed another way to solve the agility issue in data centres. Portland, just like VL2 separates names from locators and reduces broadcast by intercepting ARP requests and employs a unicast query through a centralised lookup service.

Portland assigns Pseudo MAC (PMAC) to all end hosts to encode their positions within the topology. This is changed whenever the location of the host changes. The Portland fabric manager is used for centralised lookup services, it is used to reduce broadcast overhead from the network and it works in the following manner: The switches intercept the ARP requests for IP to MAC mapping and forward a unicast query to the fabric manager which then provides the requested information to the switch. The switch then forwards it to the requesting end host. In the case where the mapping details are not available, the fabric manager broadcasts to the core/aggregation/edge/hosts, host which will reply with its AMAC which will be rewritten by the egress switch to the appropriate PMAC before forwarding to the requesting host and the fabric manager.

For load balancing, Portland and VL2 employ ECMP; except that VL2 employs VLB which before forwarding a packet randomly selects an intermediate switch. This was found to be impractical in the case where two hosts, connected to the same edge switch, want to communicate.

2.3.1.4 One-tier fabric architecture

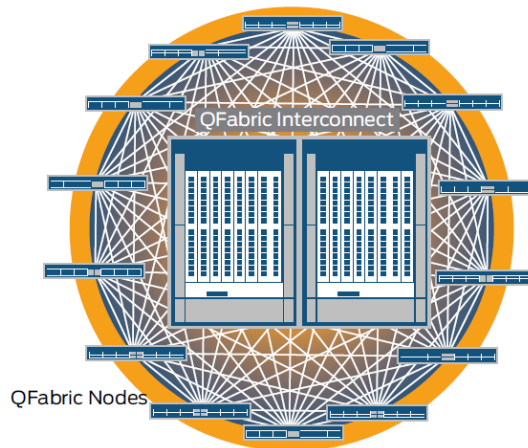


Figure 2.5. Juniper one tier Q-fabric [39]

Flattening the three-tier tree structure to a one tier fabric is an existing solution proposed for a modern data centre architecture as introduced by Juniper [39]. Juniper Qfabric architecture as shown in Figure 2.5 has flattened the data centre network and simplified the management of the data centre by reducing the number of switches. Furthermore, since there is no tree structure, there is no need for multiple hops traversing between any communicating nodes within the network. The location of hosts is not any more an issue since the network diameter and the shortest path between any two communicating nodes is always equal to one, no more oversubscription or congestion issues arise and all nodes can benefit from their all line card bandwidth.

The Qfabric single switch has an added value to the DCN since it reduces the complexity, operational cost, cooling cost, occupied floor space and power consumption. The Qfabric supports high speed server to server connectivity with low latency which makes it an attractive structure for modern data centres hosting delay sensitive applications. It also smoothes the process of virtualisation among servers within the data centre leading to

great energy savings. Qfabric can provide a power saving of about 77% if the reduced number of switches, links, cooling systems are considered along with applying other energy saving techniques such as virtualisation among data centre resources [39]. Consequently, Qfabric is considered to be a green data centre architecture that can contribute to reducing carbon footprint in the environment.

2.3.2 Server-centric data centres

Unlike switch centric designs, server centric designs were introduced to use servers as relay nodes to other servers, thus servers participate in the traffic forwarding. Server centric schemes such as BCube [30], Dcell [29], and Ficonn [31] can provide low diameter compared to switch centric schemes, can provide high capacity and support all types of traffic, especially for intensive computing applications with very low delays. In this section, an overview of BCube, Dcell, and Ficonn server centric schemes is provided along with their properties.

2.3.2.1 BCube

BCube is an example of a server-centric DCN structure which consists of servers equipped with multiple network ports connecting multiple low cost mini switches. In BCube, servers are not only hosts but they also act as relay nodes for each other and take part in traffic forwarding through multiple parallel short paths between any pair of servers. The design is driven by demands for intensive computing and higher bandwidth requirements to support applications for different traffic patterns such as one to one, one to

many, one to all and all to all. BCube supports and accelerates all types of traffic patterns and provides high network capacity due to its low diameter. The benefits of BCube design are that it can provide fault tolerance and load balancing and while requiring lower cooling and manufacturing cost.

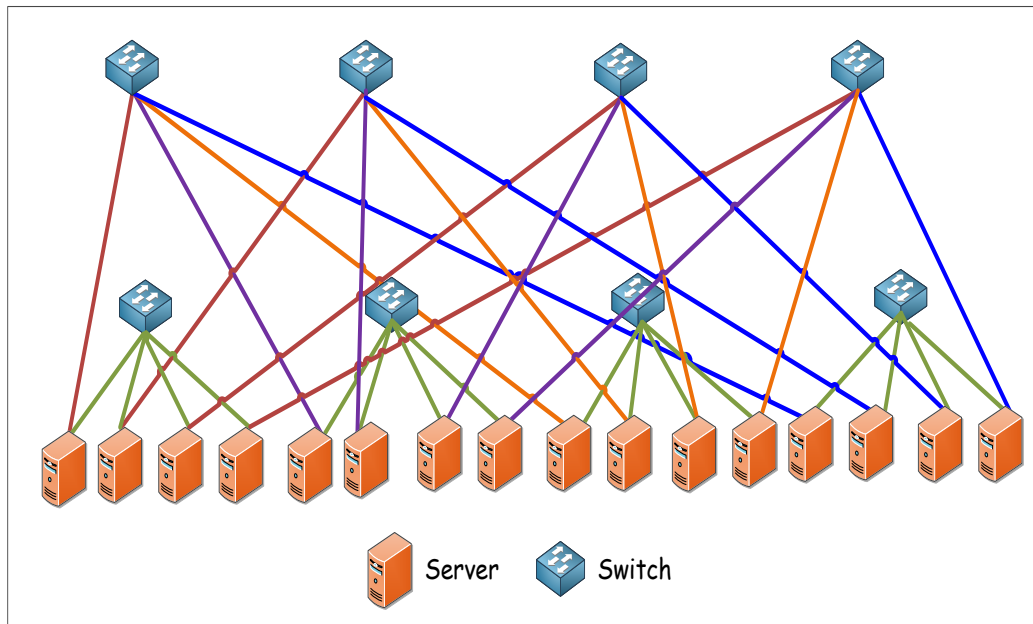


Figure 2.6. BCube [30]

BCube, as shown in Figure 2.6, can be constructed in a recursive manner starting at BCube0 as its basic building block, which is built around n -servers connected to n -port switch. Then, BCube1 is built out of n -BCube0 each of which has n -servers. BCube employs source routing protocol (BSR) when existing routing protocol such as OSPF cannot scale to thousands of servers. BSR can utilise high multipath capacity and also load balance the traffic automatically. With BSR, the source server controls the selection of the path without coordination with intermediate servers which is only responsible for forwarding received packets based on information obtained from the header. BSR probes the network to select the best path which

eliminates the need for frequent link state broadcasting which is not scalable since the network consists of 1000s of servers.

2.3.2.2 DCell

DCell is another server-centric structure for data centres that can provide desirable properties to overcome issues with scalability, fault tolerance and network capacity. As illustrated in Figure 2.7, DCell is a structure with rich physical connectivity among servers and switches and replaces expensive core and aggregation switches with mini low cost switches. However, an additional cost is introduced because of the additional and lengthy wiring communication links between switches and servers.

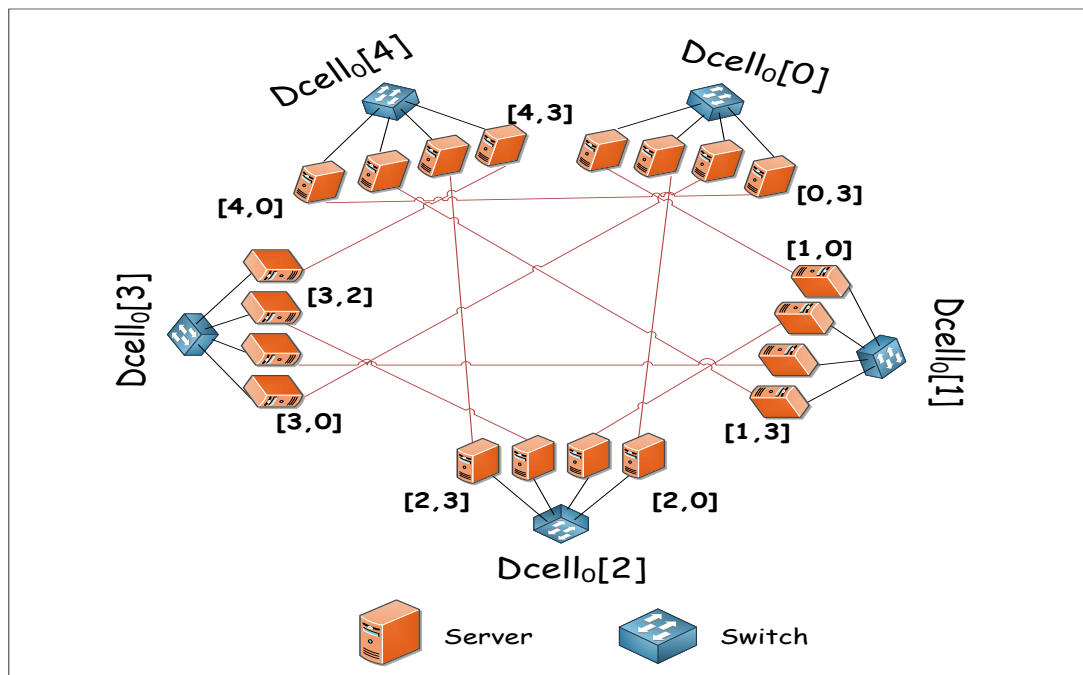


Figure 2.7. DCell Topology [29]

Similar to BCube, large DCells are recursively constructed from smaller DCells, with $DCell_0$ as the initial building block. A $DCell_0$ is constructed by connecting n servers to one low cost mini-switch with small port count. A $DCell_1$ consists of $(n + 1)$ $DCell_0$, where every $DCell_0$ is connected to every

other DCell0 in full mesh fashion as depicted in Figure 2.7. Servers in a generalized DCell topology have two interfaces each, one connects to its mini-switch and the other interface is connected to another server in a neighbouring DCell0. Any two servers with 2-tuples $[i, j - 1]$ and $[j, i]$ are connected with a link to every i and every $j > i$. As an example, in Figure 2.7, server with tuple $[4, 1]$ is connected to $[1, 3]$.

2.3.2.3 Ficonn

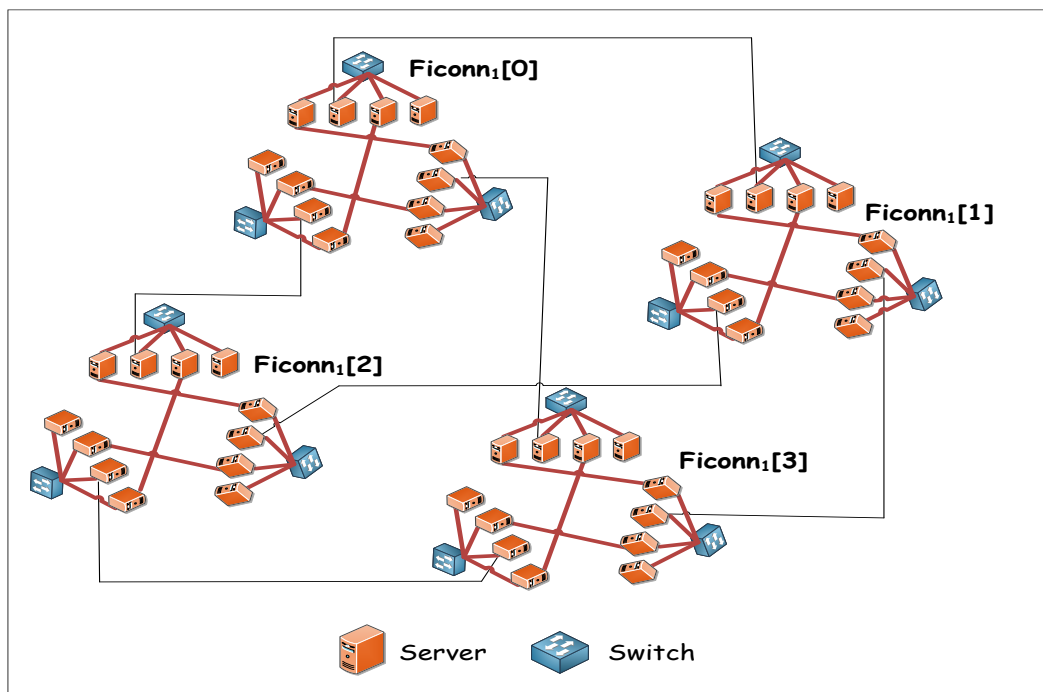


Figure 2.8. Ficonn Topology [31]

Ficonn employs an interconnection structure using commodity servers and switches to establish a scalable data centre network. It differs from BCube and DCell by making use of the two built-in Ethernet ports in the servers to establish connections and load balance traffic on two outgoing links through a Traffic Aware Routing algorithm (TAR). The throughput and routing path length can be severely affected by changing network condition. The TAR in Ficonn has the capability to monitor link capacities and, based

on information obtained on the status of network condition, it adapts accordingly.

The construction of the Ficonn interconnection can be explained by reference to Figure 2.8 where the Ficonn physical topology consists of FiConn2 with $n = 4$. FiConn2 is composed of 4 FiConn1, and each FiConn1 is composed of 3 FiConn0. There are three different level-links to constitute the interconnection within the topology, level 0 link connects each server with its switch within the same Ficonn0, level 1 or level 2 links connect the second port of the server to either another server within the same Ficonn1 or a server in another Ficonn1 within Ficonn2.

Ficonn is found to be scalable since its number of servers can be scaled up and increase exponentially with the increase of levels. Ficonn has a relative small diameter which makes the structure suitable for real time applications. Most attractively, the Ficonn's cost is much lower than other topologies since it employs a smaller number of switches and mostly relies on servers and efficient routing algorithms for switching and packet forwarding.

2.3.3 Optical data centres

A 2009 US Department of Energy vision and roadmap report estimated that a 75% energy saving can be obtained if data centre infrastructure moved toward full optical networking [40]. Optical interconnect schemes in data centres mainly rely on a mixture of active and passive optical devices to provide switching, routing, and interconnection. Such devices include

Tuneable Wavelength Converters (TWC), Optical Amplifiers, Arrayed-Waveguide Gratings (AWG), Micro-Electro-Mechanical Systems Switches (MEMS), Couplers, and Splitters. Optical interconnect schemes are mainly classified into two categories, the hybrid scheme where optical along with electrical switches are considered in the design to constitute the interconnection fabric, and the full optical network where only optical devices are employed. An insight into each scheme is presented in this section with a discussion of the architecture and main properties of the most well-known schemes such as Helios, C-through and Petabit.

2.3.3.1 Hybrid Electro-Optical data centres

2.3.3.1.1 Helios

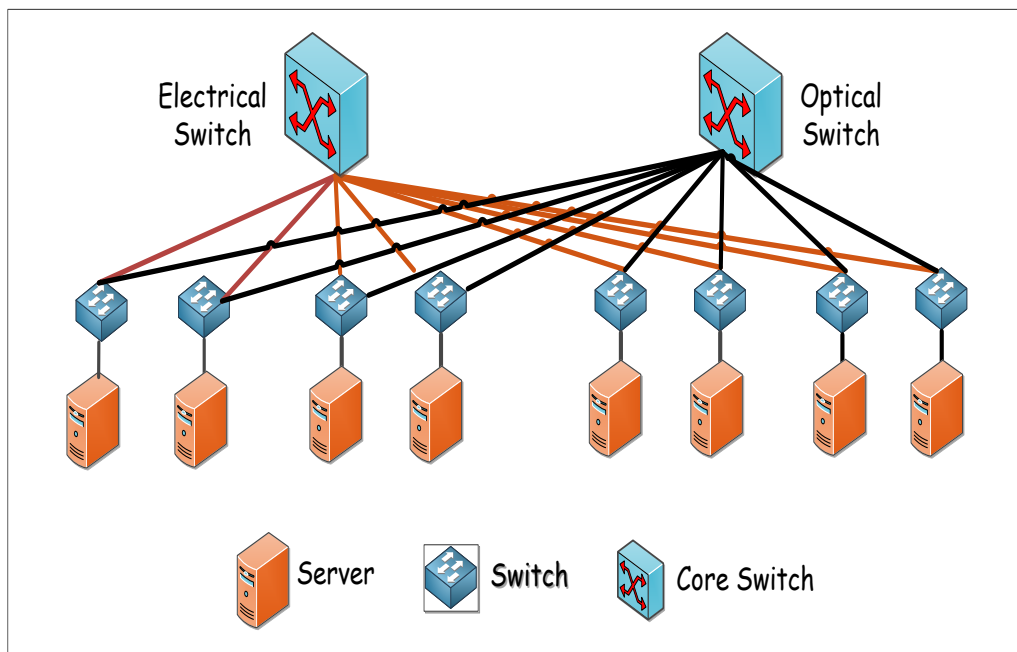


Figure 2.9. Helios [33]

Helios is a hybrid Electronic/Optical data centre architecture proposed by [33] as a design to reduce the number of switches, number of cables, cost and power consumption while maintaining full bisectional bandwidth at

minimum oversubscription ratio. Helios is a two-tier network consisting of ToR and core switches. The ToR switches are electronic packet switches while the core switches are a combination of optical and electronic switches. The electronic switches are used for all-to-all communication among pods, while the optical ones are used for long lived high bandwidth communication. Each ToR switch has two types of transceivers: 10G colourless for connecting pods to electronic core switches and $W \times 10G$ (where W can be from 1 to 32 and is the number of wavelengths multiplexed) for connecting pods to optical core switches.

The optical circuit switching in Helios relies on MEMS [41] technology. MEMS consist of crossbar fabric made of mirrors which can direct light beams from inputs to outputs without decoding or processing packets. Employing MEMS eliminates the need for signal conversion from optical to electronic. This results in high performance and lower delays. Furthermore, MEMS consume less power compared to electronic switches (240mW vs. 12.5W per port). However, MEMS have an issue with the reconfiguration time (few ms) which is seen to be long. A simplified Helios topology consisted of 64 pods, with 1024 hosts and two core switches; one for optical circuit switching and the other for packet switching. Depending on communication patterns, traffic shift and assignment are done statically between core switches through control software

The Helios design as depicted in Figure 2.9 was based on three main modules for its control software: Topology Manager (TM), Circuit Switch Manager (CSM) and Pod Switch Manager (PSM). Each module has a distinct role. The TM is responsible for monitoring and estimating pods traffic

demands between servers. Then, it computes a new topology with optical switch configuration to sustain high network throughput all the time. The CSM is responsible for configuring the MEMS after receiving the traffic connection graph. The PSM module resides in the pod switches and has a connection interfacing with the topology manager. The PSM maintains statistical details about traffic sent out from its pods. Based on calculations made by the TM for traffic routing decisions, the PSM gets the information and routes traffic accordingly either through the colourless transceivers or the WDM transceivers [41].

2.3.3.1.2 C-Through

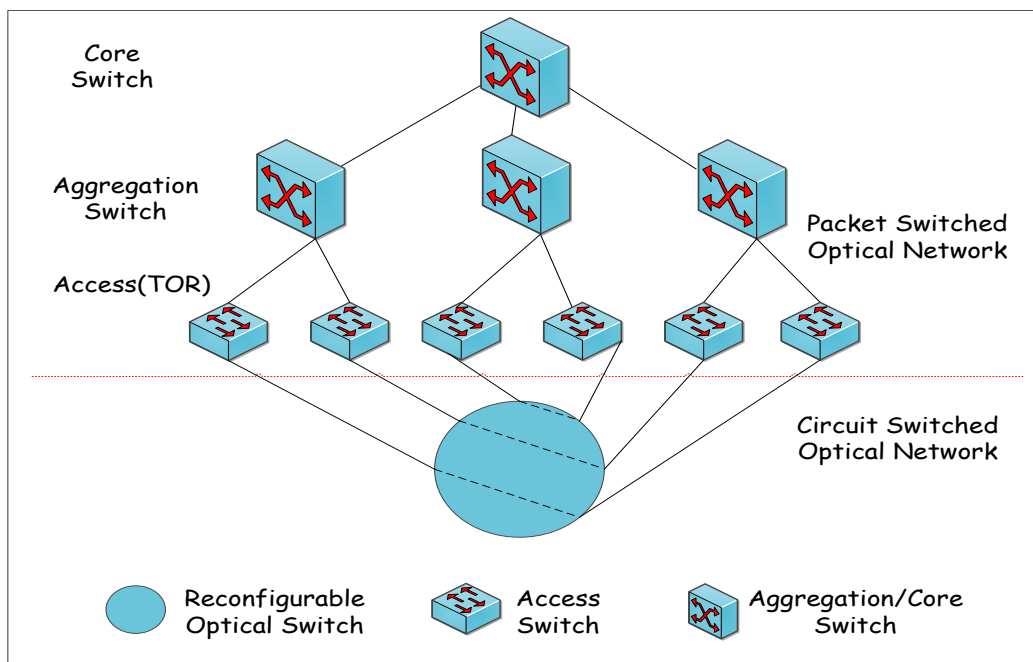


Figure 2.10. C-through [35]

C-Through as depicted in Figure 2.10 is a hybrid packet and circuit switched data centre network architecture (HyPaC) introduced in [35]. The design aims at supplying high bandwidth to data intensive applications through high speed optical circuit switched network that interconnects the

DCN's ToR switches. The HyPaC configuration, as can be seen in Figure 2.10, consists of traditional packet switched DCN tree hierarchy with access, aggregation and core switches in the top part and in the lower part, optical circuit switched network is used for rack to rack high-speed connectivity. Each rack can have one circuit switched connection at a time to communicate with any other rack in the network. For changing traffic demands over time, the optical switch can be reconfigured (few milliseconds) to establish new matching between different pairs of racks.

The traffic demands are analysed and hence links are formulated by Edmond's algorithm [42] for best maximum weight matching to satisfy dynamic intensive traffic requests among racks. The design relies on optical configuration manager that collects traffic information from the traffic monitoring systems placed on each host. Based on collected information, the configuration manager establishes circuit switched optical links among racks with respect to the bandwidth requirement among every pair of racks. Once the optical switch is configured, the ToR switches are informed about the new set up to route traffic via a special preconfigured VLAN that is dedicated to serve only optical circuits.

2.3.3.1.3 Hybrid electro-WDM PON data centres

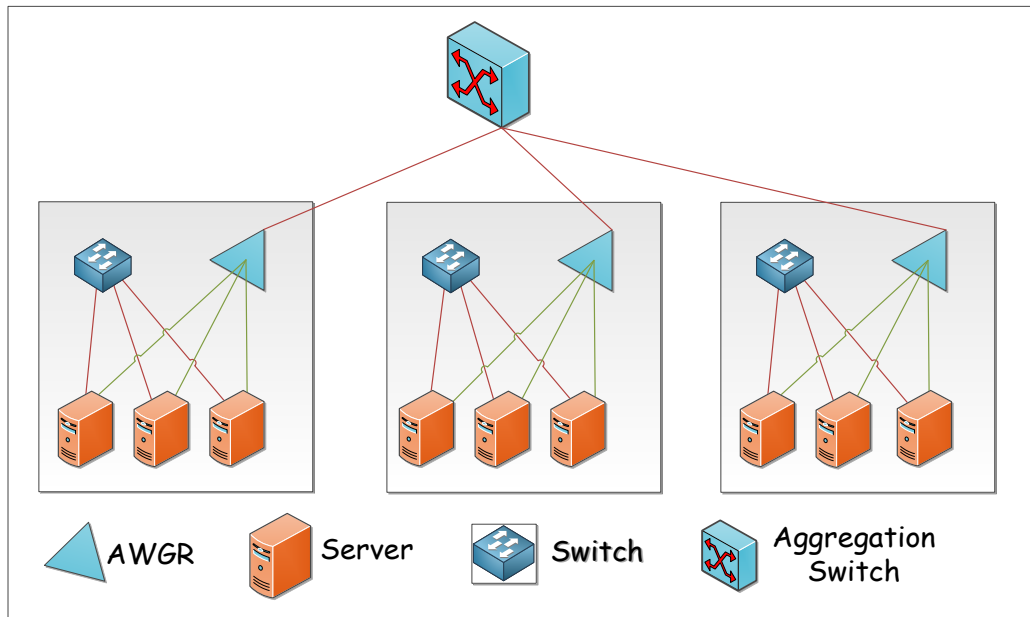


Figure 2.11. Hybrid electro-WDM PON [43]

Kachris and Tomkos in [43] proposed a novel design that introduces passive optical network devices (PON) such as Arrayed Wave Guide Routers (AWGR) in data centres. The design scheme as shown in Figure 2.11 consist of Ethernet ToR electronic switches that are used for intra rack communication and WDM PON devices (AWGR) for inter rack communication. Each server is equipped with Ethernet and optical WDM transceivers. WDM PON participates in offloading inter-rack traffic and eliminating additional processing on ToR switches, hence the power dissipated by TOR switches is reduced and high throughputs between racks are achieved with low delays. The authors reported a 10% power saving through simulation using three different traffic ratios for inter-rack and intra-rack flows.

2.3.3.2 Full optical data centres

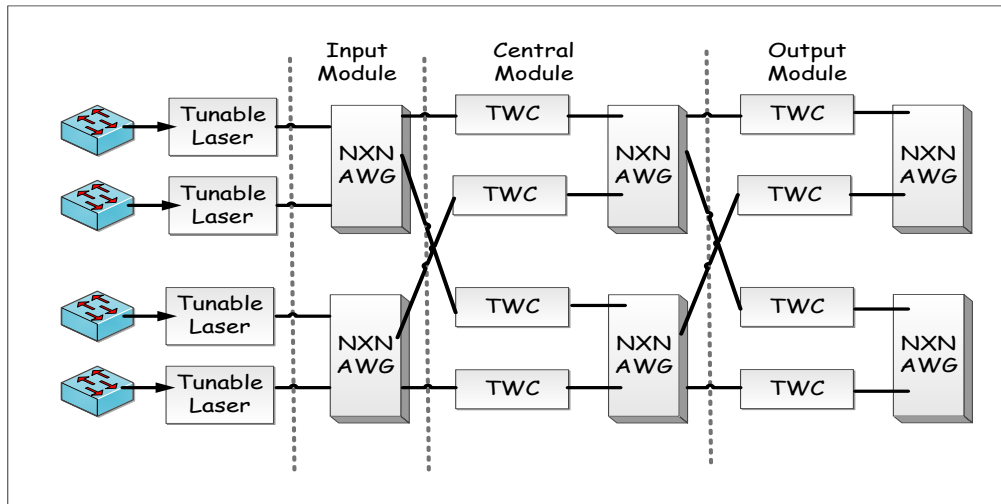


Figure 2.12. Petabit 4 x 4 switch fabric [32]

Petabit [32, 44] is a full optical switching solution for data centre networks based on a bufferless optical switch fabric using commercially available Array Waveguide Grating Router (AWGR) and Tuneable Wavelength Converters (TWC). The Petabit design objective is to overcome the issues with oversubscription, bottlenecks, latency, wiring complexity and high power consumption. The Petabit switch flattened the network by designing one switch that is capable of connecting all racks within the data centre. The design is targeting 10,000 of 100 Gbps ports by using one optical switch that is capable of delivering Petabit per second capacity. The structure of the Petabit switch as shown in Figure 2.12 is composed of a three-stage Clos network fabric with Input Modules (IMs), Central Modules (CMs) and Output Modules (OMs), where each module has an AWGR [44]. Multiple AWGRs are required for the Petabit switch since each AWGR can support few ports (128x128). Although the AWGR is passive and not configurable, the routing path from an input to an output and reconfiguration of the switch fabric are managed by TWCs which take care of wavelength conversion and hence

traffic can be routed from any input to any output. To overcome the switch fabric reconfiguration time delay when dealing with small packets, Petabit assembles packets in frames of 200 ns duration to allow sufficient time for fabric reconfiguration. In addition, the Petabit switch employs an iterative frame scheduling algorithm to coordinate input output traffic assignment. The performance of Petabit was shown to improve with the employment of three iterations and speed up of 1.6. The scheduling algorithm achieved 100% throughput, a detailed description of the scheduling algorithm is presented in [44].

Numerous other full optical designs for DCN interconnection have been presented to provide viable solutions for future data centres, allowing for high bandwidth interconnection for especially video streaming and cloud computing applications with acceptable reduced latency. Such full optical solutions include DOS [45], Proteus [46] , OSMOSIS [47], Space-WL [48] , E-RAPID [49], IRIS [50], and Data vortex [51].

2.4 Comparison and Discussion of DCN Architectures

Over the past few years, the emergence of bandwidth intensive applications with power consumption concerns have driven the evolution of data centre architectural designs. Figure 2.13 depicts a classification of the most well-known DCN architectures and their categorisations. DCNs are mainly classified into two classes: the electronic switch centric and server centric designs and the optical DCN designs. The efforts in the design of electronic data centres have succeeded to mitigate many dilemmas and obstacles in providing switch centric architectures that can support fault tolerance, load

balancing, agility and also overcome high oversubscription ratios. Server centric data centres then came next to use servers as relay nodes to each other and provide an infrastructure with low diameter and high capacity in order to support different traffic types for applications with intensive computing requirements. However, in server centric designs, additional wiring cost and complexity are a result of having servers equipped with more than one port.

Advances in the optical networking technologies in providing optical transceivers, arrayed wave guide routers, wave division multiplexing, tuneable lasers and passive optical devices have attracted great attention by researchers in academia and industry to adopt these technologies to overcome many existing issues in the design of electronic switches and server centric data centres. The driving force for the redesign of data centres to include optical switching along with electronic switches has become an attractive option because of the advancement of optical technology which has brought the prices of optical switches and transceivers down and also due to the fact that optical switching can provide high bandwidth, low power consumption and less complexity as compared to the designs which only include electronic switching technology.

Hybrid schemes such as Helios and C-through are based on readily commercially available optical components and can be implemented by upgrading current data centres. Helios and C-through are quite similar in the design except that C-through uses WDM links. The main drawback of hybrid schemes is that MEMS take few milliseconds to be reconfigured, however, MEMS are a solution that can replace high power consuming electronic

switches, where MEMs consume 0.24 Watts and electronic switches consume 12.5 Watts per port. On the other hand, most of the full optical data centre schemes are based on Semiconductor Optical Amplifiers (SOA) switches which can replace MEMS and sustain negligible reconfiguration time. Unlike hybrid schemes, full optical schemes require a complete change of current data centre in order to be implemented. Therefore, optical data centre schemes seem to be promising solutions to gradually replace electronic data centre schemes as they tend to provide low power consumption and high bandwidth with low latency.

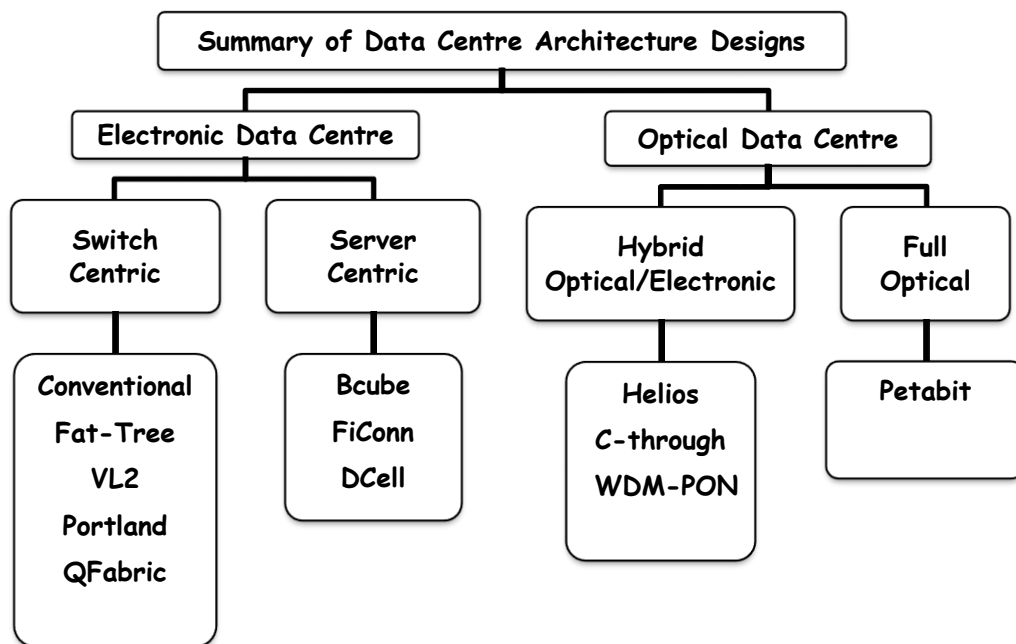


Figure 2.13. Summary of well-known data centre architectures

2.5 Techniques for Energy Efficient Data Centres

The United States (US) Environmental Protection Agency (EPA) has reported in 2007 that data centre power usage in the US doubled between

2000 and 2006 to nearly 61 billion kilowatt-hours, which represented about 1.5% of all US electricity consumption [23]. The increase in power consumption will definitely result in large carbon foot print and more emission of greenhouse gases which are the main contributors to global warming. The IT equipment is the most power hungry components in data centres, represented by servers, switches, routers and power distribution infrastructure [38]. A performance metric, Power Usage Efficiency (PUE), is used to measure how efficient a data centre is in using its power and can be calculated by dividing the total facility power by the IT equipment power consumption. The value of the PUE is typically in the range 1.2 to 2, where a PUE value of 1.2 indicates a highly energy efficient data centre [52].

In the following sections, we will present: an overview of the different methods for energy saving in data centres and industry adopted techniques for energy efficient data centres such as virtualisation, dynamic frequency and voltage scaling, dynamic network management, efficient green routing, green schedulers, scheduling schemes, and rate adaptation.

2.5.1 Virtualization

Virtualisation has received great attention by researchers and is the most adopted technique for data centres power saving [53] . Virtualisation is a method to enable services to be moved between servers and support multiple VMs which can serve different applications multiplexed to share one server. Knowing that idle servers consume about 66% of their peak power usage and having in mind that data centre resources are underutilised since the average traffic load accounts for about 30% of its resources[54], agility

can achieve servers statistical multiplexing and can give the illusion to services that they are all connected to the same switch. Hence, servers can be placed anywhere within the network and can be assigned to any service. The migration of virtual machines to consolidate workloads on a set of servers and then shutting down underutilised servers can lead to a great power saving in data centres. Researchers [55] have proposed a method to optimise data centre resources through dynamic consolidation of VMs on few servers while putting the rest on sleep state hence bringing substantial energy savings while providing the required Quality of Services (QoS).

The migration of VMs is optimised by selecting the VMs locations based on heuristics that employ utilisation thresholds. By setting up predefined threshold values and through continuous monitoring of the servers' resources utilisation, a decision of migrating VMs can be taken if these thresholds are exceeded. This results in better performance for servers and also in lower power consumption because of the overheating and the cooling system. On the other hand, VMs migration will also take place if servers' resources utilisation is below certain predefined threshold values, which allows these servers to be shut off, and save the power consumed by idle or underutilised servers.

A typical system structure consists of a dispatcher, global and local managers. The local manager role in the structure is to monitor the thermal status and resources utilisation of the network devices [56]. Based on the local manager observations, it sends to the global managers the collected information about the utilisation of resources and the VMs that have to be migrated, when the global manager becomes responsible for issuing

commands for live migration of VMs, resizing the network and hence eliminating servers by switching them off.

SecondNet [57] is a virtual data centre network architecture that can be built on top of many existing data centre network topologies such as Fat-Tree, VL2, and BCube. In SecondNet, a central Virtual Data Centre (VDC) manages the VM requests and controls virtual to physical mapping with guaranteed bandwidth reservation. Neighbouring servers are grouped into clusters, so that when VM requests are received, VDC allocation requires a search in specific clusters instead of searching in the whole network, which reduces the time complexity. In addition, grouping servers into clusters can place communicating VMs in the same cluster or within a close distance which is in fact more bandwidth efficient. The VDC manager uses spanning tree for signalling. Devices also use a the spanning tree to deliver failure messages to the manager VDC which in turn changes the routing paths and reallocate VMs if required. Path reallocation can be done in seconds where VM migration takes tens of seconds.

2.5.2 Energy-Aware Routing

The objective of energy aware routing is to save power consumption via putting idle devices in sleep mode or by shutting them down and using a few network devices to provide routing with no sacrifice in network performance. Network devices consume 20%-30% of the energy of the whole data centre [58]. The objective is to find a routing scheme for a specific topology where the total number of switches involved in the routing can sustain a network throughput that is equal to or higher than a predefined threshold. In [59] a

heuristic routing algorithm were proposed. The algorithm made of three modules: Route Generation (RG), Throughput Computation (TC), and Switch Elimination (SE). Basically, the algorithm first computes the network throughput through basic routing. Then, it gradually removes switches until the network throughput approaches the predefined performance threshold. Finally, it powers off or puts in sleep mode the switches that are not involved in the final routing. The output of the heuristic consists of an (R,G) tuple where R is energy-aware routing chosen for the traffic matrix, and G is a final topology with SE.

2.5.3 Dynamic Voltage and Frequency Scaling (DVFS)

Frequency and voltage scaling represent another method to reduce servers' power consumption, where there is a relation between voltage/frequency and the power consumed as described by: $P = V^2 * f$, (f is the frequency, V is the voltage and P is the power). The servers' memory, bus, I/O resources and disks power consumptions are not affected since they do not rely on the CPU frequency. Still, a significant saving can be achieved by reducing power via reducing frequency or voltage supplied to the processing chips [38]. In order to implement the DVFS technique on computing devices such as servers, hardware support for Advanced Configuration and Power Interface (ACPI), power management is required. The ACPI has four modes of power states: G0 for power-on, G1 for partial sleeping that is subdivided into four states; G2 is for soft-off except with having the Power Supply Unit (PSU) still supplying power and G3 for power-off state [60].

2.5.4 Rate adaptation in networks

Similar to the servers, DVS can be applied to links and switches to reduce power consumption. With respect to traffic patterns and link utilizations, data rate can be reduced by applying DVS on transceivers and ports. The energy consumed by a switch can be defined as [38]:

$$P_{switch} = P_{chassis} + n_{line\ cards} \times P_{line\ card} + \sum_{i=0}^R n_{ports} \times P_r \quad (2.1)$$

where P_r is the power consumed with respect to rate, $P_{chassis}$ is the power consumed by the chassis, $n_{line\ cards}$ is the number of line cards, $P_{line\ card}$ is the power consumed by the line card, and n_{ports} is the total number of ports.

An Ethernet link dissipates 2–4W when operating at 100 Mbps–1 Gbps and can dissipate 10–20W when operating at 10 Gbps. Hence, lowering the operating data rate could have a dramatic effect on power saving in data centres [61]. However, attention has to be paid when reducing the rate to keep the overall performance of the network intact where data rate reduction can cause link congestion.

2.5.5 Dynamic Power Management (DPM)

Dynamic power management (DPM) is a method used in data centres to reduce power consumptions of some IT infrastructure components by switching them off or by lowering the power state when inactive. Such components can be the NICs, access switches, aggregation switches, and servers as well. Putting network elements to sleep is not a new idea; it has

been already implemented for microprocessors and smart phones. The idea is to put line cards in sleep mode one by one, then to put route processor and switch fabric to sleep if all line cards are on sleep [61].

Measures and considerations for modelling a network sleep state should take care of power draw of sleep state over idle state, transition time in and out of a sleep mode, and the method to enter and exit a sleep state [61]. The Wake Up on Arrival (WOA) method was proposed in [62] for green Internet is another example deployed for data centre routers. The routers have a sensing circuit that is left powered on during the sleep mode, and it senses traffic arrival and hence wakes up routers to forward and then return to sleep if no more packets are arriving. An issue of lost bits which arrive first to wake up the router which takes time to transit from sleep to active mode was also solved by having dummy packets. In [62], an issue of frequent transitions due to small packet sizes was discussed and a solution was proposed to overcome this issue by shaping traffic into bursts, the routers arrange and maintain packets destined to the same egress into bursts and then forward them. This approach is called Buffer and Burst (B&B) and allows routers to sleep for longer time and hence save more power.

2.5.6 Energy Aware Scheduling

Different traffic scheduling approaches [38] in data centres were studied and proposed to either consolidate workloads on a few set of servers or to fairly distribute workload on the servers. A trade-off is always present between energy saving and performance, hence the scheduling should

always consider delay bounds, rate threshold and buffers occupancy in order to avoid performance degradation while achieving a considerable saving in power consumption in data centres.

In [63], the authors proposed Data centre Energy efficient Network aware Scheduling (DENS) with a main objective to balance the energy consumption of a data centre with performance, QoS and traffic demands. DENS achieves this objective via the implementation of feedback channels between network switches for workload consolidation distribution amendments to avoid any congestion or hot spots occurrences within the network which can affect the overall performance. Congestion notification signal by overloaded switches can prevent congestion which may lead to packet losses and result in high data centre network utilisation.

On the other hand, the green scheduler [63] performs workload consolidation on minimum possible set of links, switches, and servers and then uses DPM to switch off unused servers and switches. Finally, round robin scheduler can be implemented for uniform distribution of workload over all servers, which results in underutilisation of resources of data centres. The higher power consumption in DENS scheduler compared to the green scheduler can be justified by the necessity of involving extra number of servers and resources to guarantee the desired quality of service and avoid congestions.

2.6 Summary

This chapter provided a detailed survey of the most recent advances in DCNs with a special emphasis on the architectures and energy efficiency in DCNs. The survey has described the conventional tree-based DCN architecture and discussed the challenges inherited from this architecture. The architectural evolution in switch-centric and server-centric DCNs in the last decade was described and discussed. The chapter has also covered the underlying technologies used to build the electronic, optical and hybrid electro-optical DCN architectures. The switch-centric architectures surveyed include the Fat-Tree, VL2 and Portland. The server centric architectures surveyed include BCube, DCell and FiConn. The study was enriched by presenting a quantitative comparison and detailed discussion of the described DCN architectures. In parallel to the architectural evolution in DCNs, a detailed survey of recent advances in energy efficiency has been conducted. Techniques such as virtualisation, energy-aware routing in DCNs, dynamic voltage/frequency scaling, rate adaptation, dynamic power management (DPM), energy-aware scheduling methods and dynamic adjustment of active network elements in DCNs were described.

3 Review of Passive Optical Network (PON) in access network (FTTx)

3.1 Introduction

In this chapter, Passive Optical Network (PON) devices in access networks are reviewed. PON devices such as arrayed waveguide routers, star couplers, star reflectors, and fibre Bragg grating are described along with the construction and main applications of each PON device. A study on the PON deployment in Fibre to the Premises (FTTx) access networks is also described to evaluate standards, protocols and classifications of PONs in access network.

3.2 General overview of PON devices

This section briefly introduces the main passive optical network devices and their applications. Description of PON devices' physical characteristics, advantages and their main deployment in FTTx access network will be described. Typical PON devices include the Arrayed wave guides grating, couplers, and splitters fibre brag grating and star reflectors.

3.2.1 Arrayed Wave guides grating (AWG)

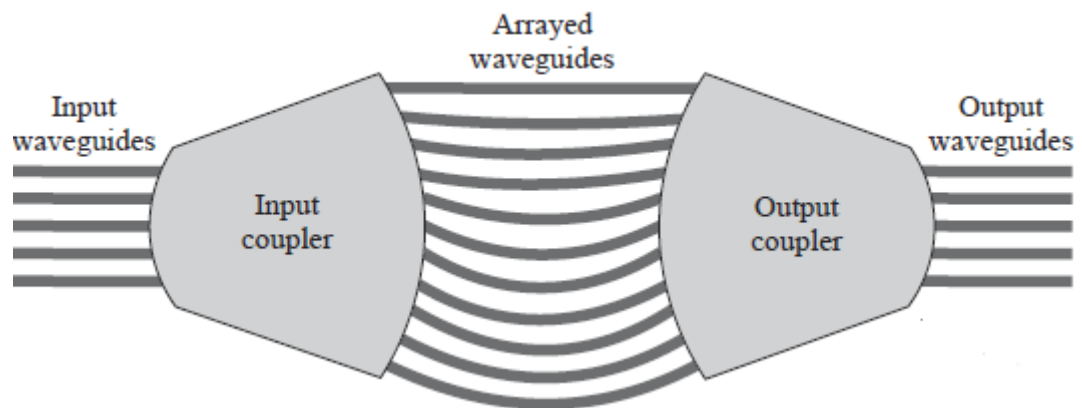


Figure 3.1. 4x4 Arrayed waveguide grating [64]

3.2.1.1 Components of the AWGR

An AWGR consists of Two $N \times M$ star couplers, one at the input and another one at the output, they are also called free propagation regions (FPR) and M Arrayed waveguides with different lengths that connect the two couplers. The difference in length of every two adjacent arrays is constant and equal to ΔL .

3.2.1.2 The operation of the AWGR

When a multi wavelength signal beam arrives at the device at any port through the input star coupler, the signal spreads out and its power is equally divided among the M waveguides which in turn, propagate the signals to the output coupler. Because of the different length of the array waveguides, the phase of each of the propagating wavelengths is changed and shifted differently. At the output of the array, M waves with different phases enter the output coupler. The wavelength with the phase that interferes constructively at an output fibre will exit the system otherwise it will destructively interfere.

3.2.1.3 Applications of AWG

A. Multiplexer AWG

An example multiplexing AWG is depicted in Figure 3.2(a), four different wavelengths are interfaced with the four input ports of the Nx1 AWG. These are multiplexed and directed to the output port.

B. Demultiplexer AWG

Figure 3.2(b) shows a 1xN AWG as a de-multiplexer where an input signal consisting of four different wavelengths λ_1 , λ_2 , λ_3 and λ_4 distinguished and directed to ports 1, 2, 3, and 4, respectively.

C. Add-drop multiplexer AWG

In an add-drop multiplexer, data contained in a light beam of wavelength λ_2 , for example, is dropped and different new data received from different systems are added to replace the dropped data exiting at output port 1 as shown in Figure 3.2(c).

D. Full interconnection AWG

In a full interconnect, a multi wavelength signal arriving at input port 1, for example, is distributed to the output ports according to the signal wavelengths. In Figure 3.2(d), a signal of wavelength λ_1 is routed to output port 1, where wavelength λ_2 is routed to output port 2. This application will be covered in more detail in later sections.

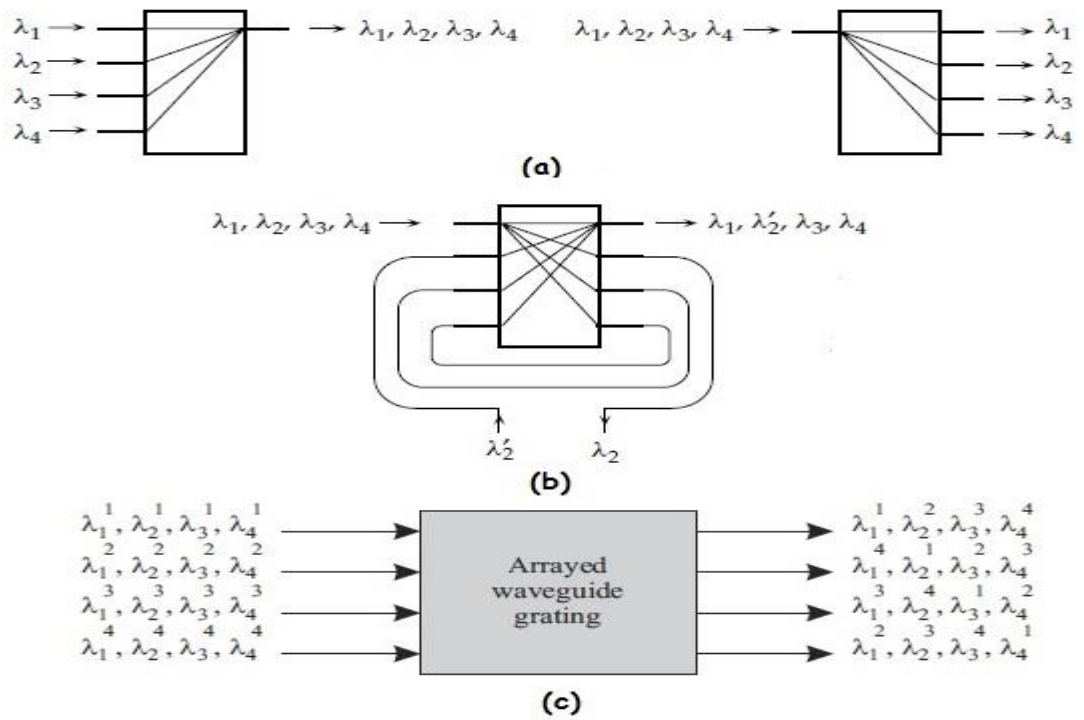


Figure 3.2. Applications of AWG devices. (a) Mux/Demux (b) drop/add multiplexer (c) full interconnection [64]

3.2.2 Fibre Bragg Grating

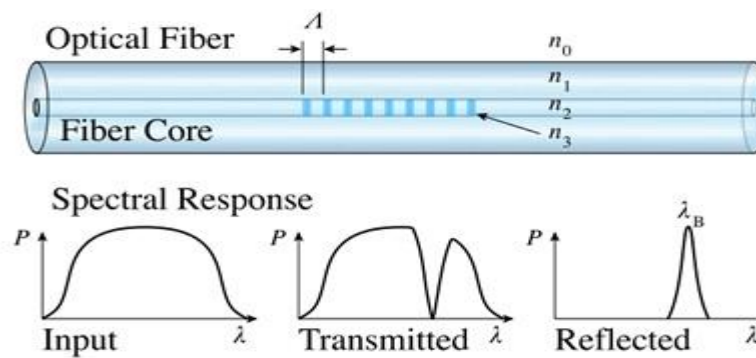


Figure 3.3. Fibre Bragg grating [64]

A fibre grating as shown in Figure 3.3 is a periodic variation (Λ) in small section of the core refractive index (n) of an optical fibre for the purpose of

reflecting a certain wavelength. The reflected wavelength obeys Bragg's law and can be expressed as [36]:

$$\Lambda = \frac{\lambda}{2n} \quad (3.1)$$

Other wavelengths that don't obey Bragg's law are transmitted through and will not be affected by the periodical variation of the core. Applications of the Brag grating include, strain and temperature sensors, WDM filters, and tuneable filters and many others [36, 65].

3.2.3 Splitters/couplers

The concept behind beam splitting in passive optical devices is explained in this section. Snell's law leads to a method to split an optical beam in to two beams with equally divided intensity. Snell's law basically relates the angles of incidence and refraction to the medium refractive indices as presented in the following equation [66]:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \quad (3.2)$$

Figure 3.4 shows a light wave that travels in a denser medium and strikes a dielectric material with a lower index of refraction, a portion of the light is transmitted and the other part is reflected. If the angle of incidence is increased and exceeds the critical angle (transmitted angle = 90 degree), there will be no transmitted light and the light will be reflected back into the dense medium. This phenomenon is called total internal reflection (TIR) where the transmitted angle equal to 90 degree ($\sin 90 = 1$) and the angle of incidence will be the critical angle [66].

$$\theta_c = \arcsin \frac{n_2}{n_1} \quad (3.3)$$

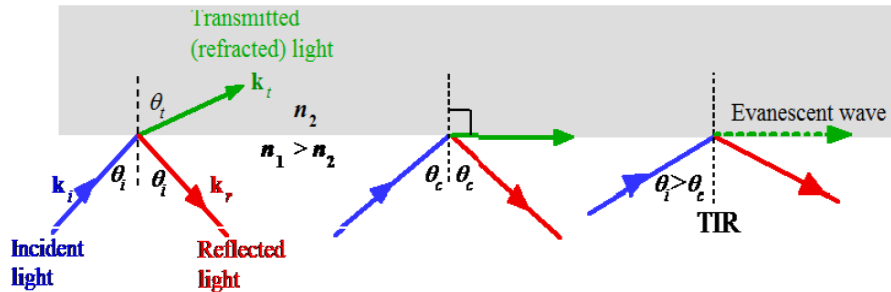


Figure 3.4. Total internal reflection (TIR) [66]

Total internal reflection is a result of the propagation of a light beam from a dense medium to a less dense medium where light is incident at an angle greater than or equal to the critical angle. Looking at Figure 3.5, another medium (medium C) with a refractive index equals to n_1 is introduced and by having medium B sandwiched in the middle between A and C. If medium C is close enough to medium A, the decaying evanescent wave will penetrate into medium B reaching the interface BC. This phenomenon is called Optical Tunnelling or frustrated total internal reflection resulting from the proximity of C which frustrates the TIR and reduces the intensity of the reflected light.

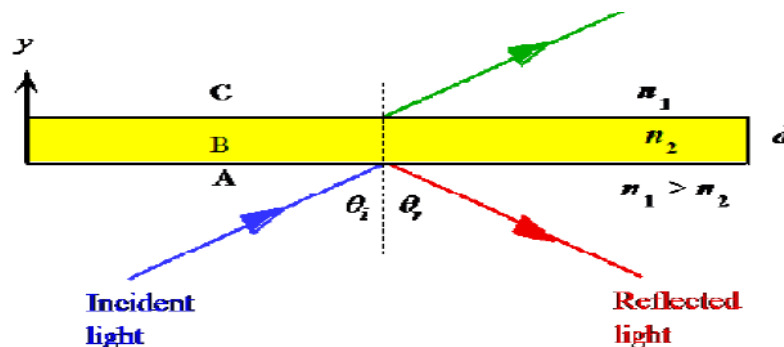


Figure 3.5. Optical tunnelling or frustrated TIR [66]

Frustration of total internal reflection was found to be one of the ideas that can be used to split a light beam using a beam splitter. The idea is very simple; a prism cube can be made by using two prisms where the prisms are separated by a low refractive index film as showing in Figure 3.6. When a light beam is incident at prism A at an angle greater than the critical angle, total internal reflection can be achieved and half of the light intensity is tunnelled through the thin film and transmitted into C.

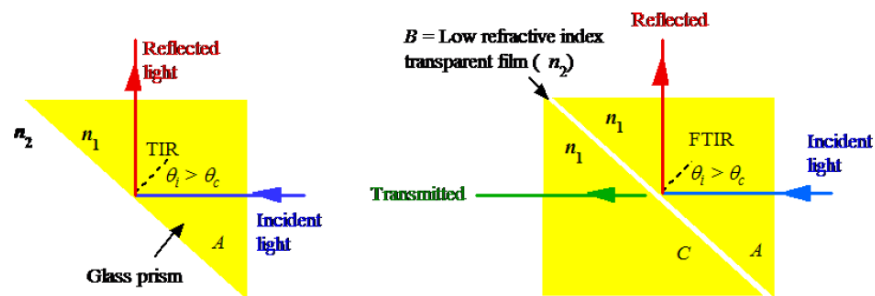


Figure 3.6. Prism splitter [66]

Splitters are attractive solution and are implemented in many applications. Passive splitting/coupling has been widely implemented in fibre to the premises access networks. A signal can be passively split by cascading multiple of 1x2 splitters as shown in Figure 3.7. Splitters can be formed by branching the waveguides in the form of a Y junction. Figure 3.7 presents cascaded 1x2 couplers used to produce a 1x8 coupler. A basic technology for building passive optical network splitters is the Fused Biconical Taper (FBT). An FBT splitter is made by wrapping two fibre cores together, putting tension on the optical fibres, and then heating the junction to make the two fibres tapered and fused together [65].

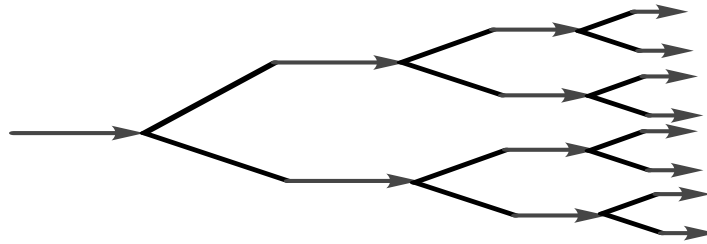


Figure 3.7. Cascaded 1x2 couplers/splitters to produce 1x8 coupler [36]

3.2.4 Star reflector

In brief, with star reflector, light incident to any port is coupled to all other ports. Figure 3.8 shows a star reflector passive device with 8 inputs. An incident light signal to any input port will result in broadcast to all ports.

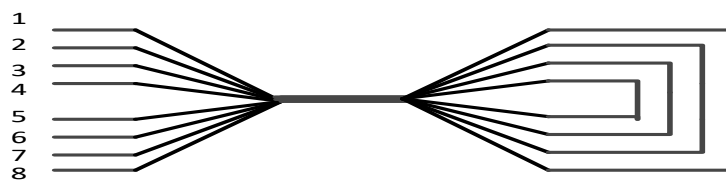


Figure 3.8. Star reflector [36]

The remaining sections of this chapter present PON deployment in Fibre to the Premises (FTTx) access network. The main standards developed over the years and their different categorisations based on the media access control protocols used are covered.

3.3 PON deployment in access networks

The high capacity links to end users, cost efficiency, and low power consumption of PON technology have motivated its mass-deployment in access networks. PONs are currently mainly deployed in fibre to the home, curb, or premises (FTTx) access networks to provide high speed broadband

triple play services of voice, data, and video over a single strand of fibre to be shared by multiple users in residential areas. PONs can be extended to reach users who are 20-60 km away from the Central Office (CO) with a 128-256 maximum split ratio which is equivalent to the number of subscribers that can be provisioned with triple play services from one port.

Figure 3.9, depicts a typical FTTx PON deployment. An Optical Line Terminal (OLT) switch is located at the CO. This switch is responsible for managing traffic flows among different PONs, controlling resource allocation to fulfil subscribers' traffic demands, and coordinating the arbitration of channel access to avoid collision. On the other side, Optical Network Units (ONUs) are installed at the subscribers' premises to deliver services such as telephone, Ethernet data, and IPTV. The uplink traffic flow from the ONUs to the OLT is a Point to Point (P2P) connection. The downstream traffic flow is in the form of Point to Multi Points (P2MP), originating from the OLT and broadcasted to the ONUs connected within the same PON. The P2P and P2MP nature of traffic forwarding in PONs is a result of the directionality and photonic functionality of the passive optical splitters/couplers that can aggregate and separate the optical signal passively. The bandwidth and resource allocation to the ONUs are managed by the OLT switch and can be performed statically or dynamically.

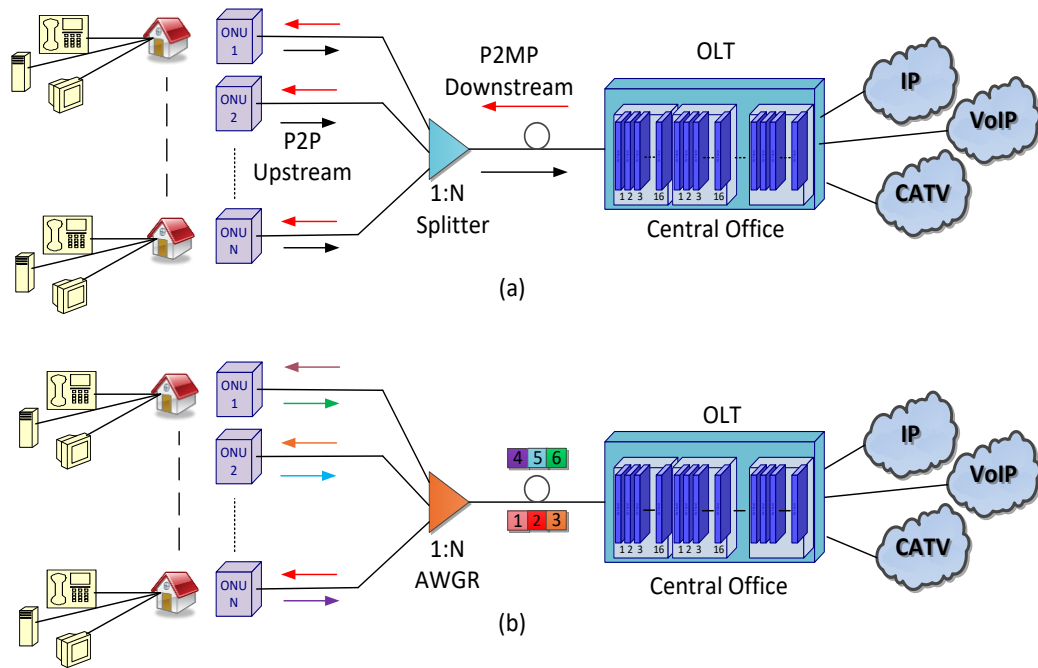


Figure 3.9. (a) TDM PON architecture (b) WDM PON architecture

PONs technology has evolved over the years. The ITU along with IEEE have standardised several designs to meet a range of requirements in terms of uplink and downlink transmission rates, number of splits, reach, and transmission protocol. The three main PON standards are Broadband PON (BPON), Ethernet PON (EPON), and Gigabit PON (GPON/XGPON). The characteristics of each PON are described in Table 1 to compare transmission capability, number of allowed splits, operating wavelengths, distance and transmission protocol of each scheme [67, 68].

Table 3.1 PON access network classifications

	BPON	EPON	GPON	XG-PON
Standard	ITU-T G.983	IEEE 802.3ah	ITU-T G.984	ITU-T G.987

Transmission Protocol	ATM	Ethernet	Ethernet, ATM, TDM	Ethernet, ATM, TDM
Transmission Speed	Up to 622 Mb/s downstream and up to 155 Mbit/s upstream	Symmetric up to 1.25 Gbit/s	Up to 2.5 Gb/s downstream and up to 1.25 Gbit/s upstream	Up to 10 Gb/s downstream and up to 2.5 Gbit/s upstream
Upstream wavelength	1310 nm	1310 nm	1310 nm	1270 nm
Downstream wavelength	1490 nm and 1550 nm	1550 nm	1490 nm and 1550 nm	1578 nm
Distance	20 km	10 km	Up to 60 km	Up to 60 km
Maximum Allowed splits	32	32	32,64, and 128	32,64, and 128

Three Classes of optical transceivers were standardised in order to define the permitted ranges of attenuation caused mainly by split ratio and distance between the transmitter and receiver. Transceivers' Classes with respect to power budgets in PON access networks are as follows[69]:

[1] Class A: 5-20 dB

[2] Class B: 10-25 dB

[3] Class C: 15-30 dB

3.4 Categorisation of FTTx PONs based on Media Access Control (MAC) protocol

Based on the Media Access Control (MAC) protocol implemented, PONs can be classified to two main categories; TDM-PON and WDM-PON. In TDM PON as shown in Figure 3.9(a), two wavelengths are used; one for uplink stream and the other for downlink stream. In the uplink direction ONUs compete to access the shared transmission channel. Many bandwidth allocation algorithms were proposed to mitigate and enhance media access in PON [70-73]. In static bandwidth allocation algorithms, ONUs are assigned with a predefined bandwidth whether there is a need to use it or not while in dynamic bandwidth allocation algorithms, bandwidth is dynamically allocated based on demand, quality of service requirements, and resources availability.

On the other hand, WDM-PONs as shown in Figure 3.9(b) avoid resource sharing among ONUs through the use of AWGRs instead of the star splitters/couplers. In WDM-PON, each ONU is assigned a pair of wavelengths dedicated for its upstream and downstream transmission and hence there is no need for bandwidth allocation algorithms. However, WDM-PONs suffers from a number of limitations in terms of scalability of wavelengths, number of costly tuneable lasers and photo detectors to be deployed, and bandwidth utilisation.

Hybrid WDM-TDM architectures were proposed to overcome these limitations. In hybrid architectures, wavelengths can be dynamically assigned and shared by multiple ONUs located at different PONs. The ability to dynamically tune to different wavelengths, allows ONUs to join other TDM-PONs which enhances the bandwidth utilisation at low loads and also avoids congestions at high loads. The number of laser diodes at the OLT can also be reduced by using a multicarrier generator [74] that is capable of supplying hundreds of carriers with ONUs that have no light source and directly modulate the optical carrier received by the OLT for uplink transmission.

3.5 Summary

This chapter has focused on FTTx technology implementation for residential access network and studied its deployments, architectures, hardware, characteristics, classifications, implemented protocols, and standards. The goal is to make use of PON technology with its attractive proven performance in residential access networks to provide energy efficient, high capacity, low cost, low latency, scalable, and highly elastic solutions to support connectivity inside modern data centres and to overcome some of the limitations in current data centres architectures. The next chapter of the thesis investigates the feasibility of designing new data centre architectures relying mainly on PONs to facilitate inter and intra rack communication.

4 Proposed PON architectures for data centre networks

4.1 Introduction

As PON technology performance has been proven in access networks and has shown its capability in provisioning low cost, high capacity, low latency, scalable, and energy efficient networks, it has become more attractive to be adopted to provide interconnection fabric in modern data centres. The use of Passive Optical Networking (PON) technology in data centres and the useful functionalities provided by devices like Arrayed Waveguide Grating Routers (AWGR), Fibre Bragg gratings (FBG), and star couplers/splitters have attracted much attention from the research community in the last few years.

In this chapter, the emergence of PONs in the design of interconnection fabric in data centres is reviewed and new designs are introduced. A review of previous works that considered partial implementation of PONs in the data centre is provided. Five novel designs for PON implementation in data centres are proposed. Different solutions using mostly passive optical devices to manage inter-rack and intra-rack communication among servers are presented. At the end of this chapter, a qualitative comparison between the five proposed designs is given to outline the advantages and

disadvantages of each design when compared to other designs. Detailed evaluation of the key proposed architectures is then given in Chapters 5 to 9.

4.2 PON Emergence in future data centre architectures

Recent progress in data centres has led to increased data rate requirements per server, where servers now have 1 Gb/s cards as standard and higher rate cards are being considered. This coupled with the large number of servers per data centre potentially reaching and exceeding 1 million servers per data centre, has meant that the use of optical networking approaches in data centres has become essential to provide the aggregate data rates required [75].

In addition, studies have shown that network connectivity is becoming a bottleneck in the data centre. More importantly improving the per server data rate can significantly reduce the number of servers needed and hence the overall data centre power consumption and cost. For example [75] illustrates through a sorting example that increasing the per server data rate by a factor of 100 from 2.8 MB/s to 293 MB/s reduces the number of servers needed by a factor of 66 from 3452 to 52 servers resulting in significant cost and power savings, all achieved through improved networking. In addition to sorting, data centres perform a range of other unrelated functions such as hosting and streaming content to users, where the impact of improved networking may not be as dramatic, but is still important. Server networking resembles in a number of ways access networks that interconnect homes, with the server replacing the home. Therefore PONs developed for residential networking are a natural contender, being both an optical networking

approach and a proven scalable solution that can support hundreds to millions of homes in a country.

However, if PONs are to be deployed in data centres, then three major challenges have to be tackled which are specific to the data centre environment. Firstly traditional PON architectures were developed to cater for home (i.e. Optical Network Units (ONUs) in the home) to the telecommunications office (OLT) traffic, for example for Internet access. Here home to home traffic is not a major strand. In data centres, server to server traffic is essential and therefore new PON architectures that support server to server traffic are essential.

Data centre traffic strands are considered to propose a number of new PON architecture options that support both forms of server to server traffic: inter-rack and intra-rack communications. Secondly, residential PONs typically have one route from ONU to ONU, and this is realised via the OLT switch. This loads the OLT switch if server to server communication is large and with a single path, load balancing becomes difficult. Therefore, designing new PON architectures that provide multiple routes between servers within the PON 'cell' and/or through the rest of the architecture is needed. Thirdly PONs are designed typically for asymmetric traffic where home users typically download from the Internet more than they upload. Server to server communication can however be symmetric and can exhibit different degrees of traffic asymmetry. Therefore our proposed designs introduce architectures with more than a single route (direct in several cases) between servers, thus providing symmetry. Moreover, MAC protocols can be re-designed to provide the required symmetry and variable degrees

of asymmetry. While possible solutions are proposed here, one of the goals is to outline the open research issues and the important problems to be tackled in realising efficient PON data centre networks.

As discussed, significant research efforts have been devoted over the last decade to the design of efficient data centre networks to mitigate the limitations of conventional data centre architectures. However, major concerns are still raised about the power consumption of data centres and its impact on global warming in the first place and on the electricity bill of data centres in the second place.

Given the steadily increasing number of servers and the exponentially growing traffic inside data centres, the limitations of conventional data centre networking architectures such as in respect of link oversubscription and inefficient load balancing have become even more critical. Also conventional data centre architectures are based on expensive and power hungry devices such as access switches, aggregation switches and core switches.

The use of PONs can overcome the problems of switch oversubscription and unbalanced traffic in data centres where PONs architectures and protocols have historically been optimised to deal with these problems as well as handling bursty traffic efficiently through flexible protocols.

4.2.1 The need for PONs in the data centre interconnection design.

The limitations of current proposed data centre networking infrastructure in terms of capacity, cost and energy efficiency have triggered the need for new architectures to meet efficiently the growing demands of modern data centres. The choice of the architecture of a DCN is of premium importance

as it impacts the overall efficiency of the DCN. The architecture of a DCN, or its topology, directly reflects on its scalability, cost, fault-tolerance, agility and power consumption. DCNs continue to evolve and considerable research efforts are being made to address the various challenges observed. The choice of a DCN solution to address one challenge impacts and often limits the alternatives and how to address other issues.

Attention has recently been directed, in respect of DCNs [76] [43], to Fibre to the Premises FTTx technology for residential access networks to study its deployments, architectures, hardware, characteristics, classifications, implemented protocols, and standards to design new data centre architectures that mostly rely on passive optical networks. The goal is to make use of PON technology with its attractive proven performance in residential access networks to provide energy efficient, high capacity, low cost, low latency, scalable, and highly elastic solutions to support connectivity inside modern data centres and to overcome most of limitations in current data centre architectures.

A number of PON designs will be presented, discussed, and compared to replace the high power consuming access and aggregation switches in current data centres infrastructures. In subsequent sections of this chapter, the deployment of PONs in future data centres are tackled by re-designing the current paradigm of PONs (used traditionally in residential access networks) to furnish scalable, low cost, energy-efficient, and high capacity interconnections infrastructure to accommodate the different traffic patterns in data centres.

In addition to the use of mostly passive optical devices, other challenges addressed by these designs include off-loading the inter-rack traffic from the OLT switch to avoid undesired power consumption and delays and reducing or eliminating the need for expensive tuneable lasers.

4.2.2 Related work

Previous work as discussed in Chapter 2 Section 2.3.3 has considered partial implementation of PONs in data centres to take advantage of the different merits provided by the PON architectures. Huawei in [76] presented a study on the PON technology and its capability in terms of storage, processing and interconnection speed to furnish all the requirements for cloud computing. In [43] the authors introduced Array Waveguide Grating Router (AWGR) based PONs in addition to Top of Rack (TOR) and aggregation switches to off-load the burden of managing the inter-racks communication traffic from access switches to PONs. Therefore, the power consumed by access switches was reduced by 10% and low delays with high throughputs were obtained for rack to rack communications. Another recent partial PON implementation for data centre interconnection is presented in [77] where aggregation switches are replaced by passive optical AWGRs supported with Orthogonal Frequency Division Multiplexing (OFDM) modulation. The proposed architecture in [77] has demonstrated low delays and high throughput with flexibility of bandwidth allocation through efficient subcarrier assignments using access TOR switches and AWGR for interconnection fabric.

No designs have considered full passive interconnection for intra and inter racks communication within data centres. The objective of the proposed

designs is to tackle the deployment of PONs in future data centres by re-designing the current paradigm of PONs used for access networks to furnish a scalable, low cost, energy-efficient, and high capacity interconnections infrastructure to accommodate the different traffic patterns in data centres. The subsequent section will provide a general overview of PON devices and its applications along with its deployment in access networks.

In this chapter, different PON architectures are proposed to furnish a scalable, high speed, and energy efficient data centre interconnection. First, PON capability to provide infrastructure for cloud computing in data centres is explained. Followed by a Study on traffic patterns and locality in data centres and access networks is reviewed. Then, five proposed PON designs are presented. The chapter concludes with a qualitative comparison between the proposed designs to summarise the main differences along with advantages and disadvantages of the different proposed architectures.

4.3 PON capability study for data centres

In this section PON deployment in data centres is introduced. Figure 4.1 shows the connectivity created by a typical PON deployment. The OLT switch consists, typically of 8 chassis, hosting up to 16 cards each. Each OLT card has the capacity to connect 8 ports, each of which provides a transmission rate up to 10 Gb/s. With a splitting ratio of 128, a single card port can connect 128 subscribers, and therefore one card can connect 1024 subscribers, and one chassis can provision connection to 16,384 subscribers.

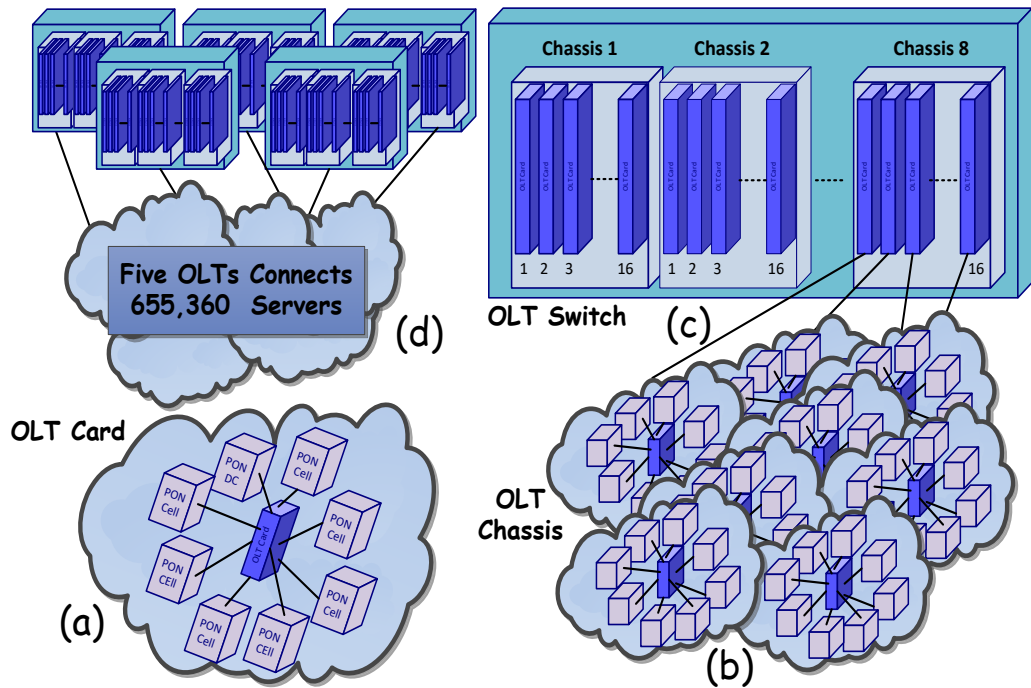


Figure 4.1. (a) OLT card with 8 ports to connect 8 PON Cells (each PON Cell hosts 128 servers), (b) OLT Chassis with 16 cards to connect 128 PON Cells, (c) OLT switch with 8 chassis, (d) Five OLT switches

This large connectivity furnishes a great infrastructure for cloud computing in data centres to provide services to clients that have different applications such as Infrastructure as a service (IaaS), Platform as a service (PaaS), and Software as a service (SaaS). The fact that data centre interconnection fabric covers distances of only hundreds of meters significantly reduces the attenuation incurred by signals and therefore more splitting can be provisioned and higher number of servers can be connected to the PON. With such connectivity and connecting the ONUs to Intel core 980x servers, with 8 GB RAM memory and 147,600 MIPS processing capability, a total processing capacity of 2,418,278 GIPS and memory of 131,072 GB can be obtained per an OLT switch chassis. Figure 4.2 illustrates the memory and processing capability of PON network for data centres.

The deployment of PON to provide connectivity inside data centres eliminates the need for access and aggregation switches used in current data centre connectivity (see Figure 4.1), and therefore reduces the power consumption while guaranteeing excellent performance in terms of resource allocation and speed of interconnection between servers.

Given the potential of PONs to support connectivity within data centres, the challenge will be to re-design the current interconnection relying only on passive optical devices to serve the different traffic patterns in data centres. In the following Section the traffic patterns in data centres are discussed.

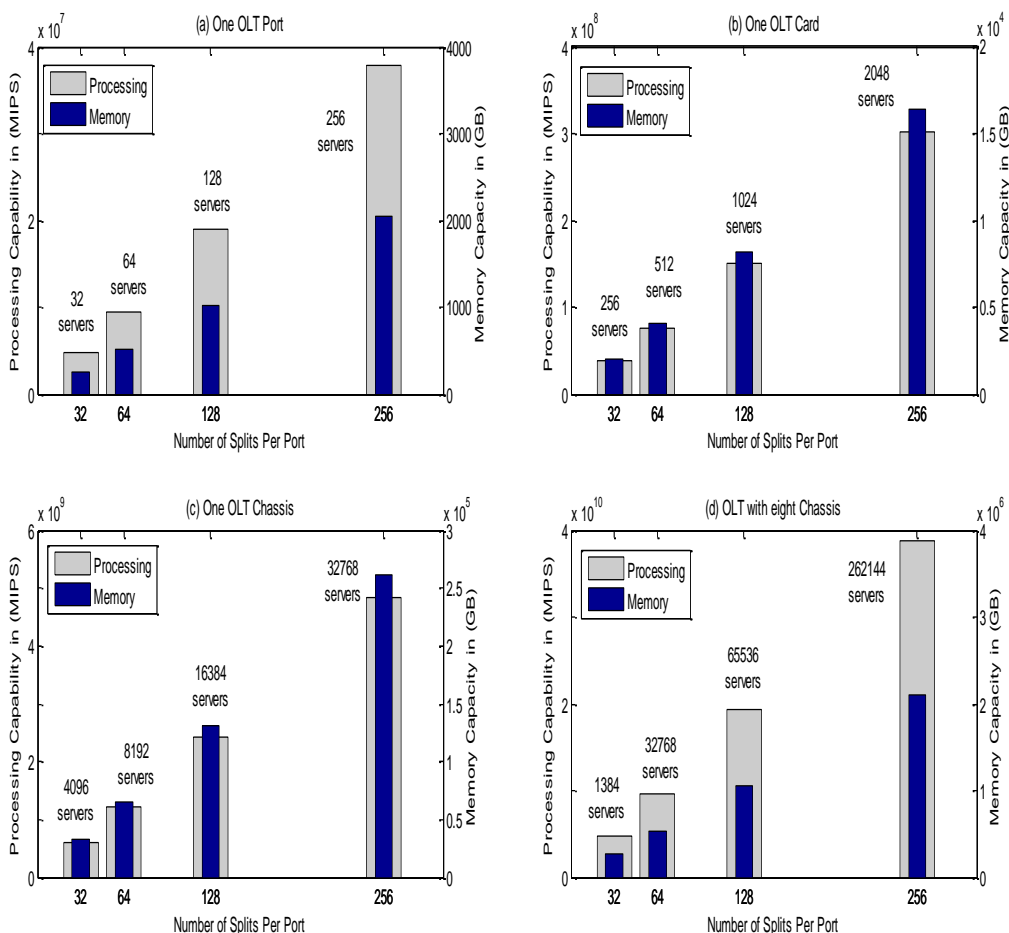


Figure 4.2. Total Processing capability and memory capacity for (a) PON Cell, (b) PON OLT Card, (c) PON OLT Switch Chassis, and (d) PON OLT Switch with 8 Chassis for data centre hosting Intel core 980x servers, with 8 GB RAM memory and 147,600 MIPS processing

4.4 Study of traffic patterns in FTTx and data centres

Traffic patterns in FTTx PON access networks consist of a mixture of voice, data and video. The tree topology nature of PON along with its directionality provides a suitable architecture for the interconnection between the OLT and the ONUs where traffic is either destined from the ONUs to the OLT (uplink) or from the OLT to the ONUs (Downlink). In FTTx, ONU to ONU traffic is not a major concern for residential areas since it is only common for voice traffic. In such a case, a request has to be forwarded to the OLT switch which in turn facilitates a connection between the two ONUs whether they are neighbours connected to the same PON or apart in two different PONs. However, the delay introduced by establishing a connection between neighbouring ONUs through the OLT switch is negligible for voice applications.

However, ONU to ONU traffic needs to be addressed more efficiently if PONs are deployed to provide interconnection fabric in data centres. In data centres, traffic patterns differ based on the application provided. Traffic within a data centre, as shown in Figure 4.3, can be classified into four main categories; (i) In-Out traffic destined out of the data centre through access switches, aggregation switches and core routers, (ii) Intra-rack traffic between servers located within the same rack through the access switches, (iii) Inter-rack traffic between servers located in different racks through the access switches and the aggregation switch linking the two racks, (iv) Out-In traffic entering the data centre through core routers, aggregation switches, and access switches.

The percentage of inter-rack and intra-rack traffic within a data centre varies between 20-80% depending on the type of data centre and the running applications [78]. Therefore, eliminating the access and aggregation switches and replacing them with directional PON splitters/couplers results in over-loading the OLT switch making it the bottleneck for all types of traffic. Despite the fact that the OLT switch backplane can provide non-blocking hundreds of gigabit interconnection among its cards, offloading the burden on the OLT switch will avoid undesired delays and power consumption resulting from O/E/O conversions, queuing, buffering and processing. In the subsequent sections, different design approaches to mitigate the challenges facing the implementing of PONs to support connectivity inside data centres are presented.

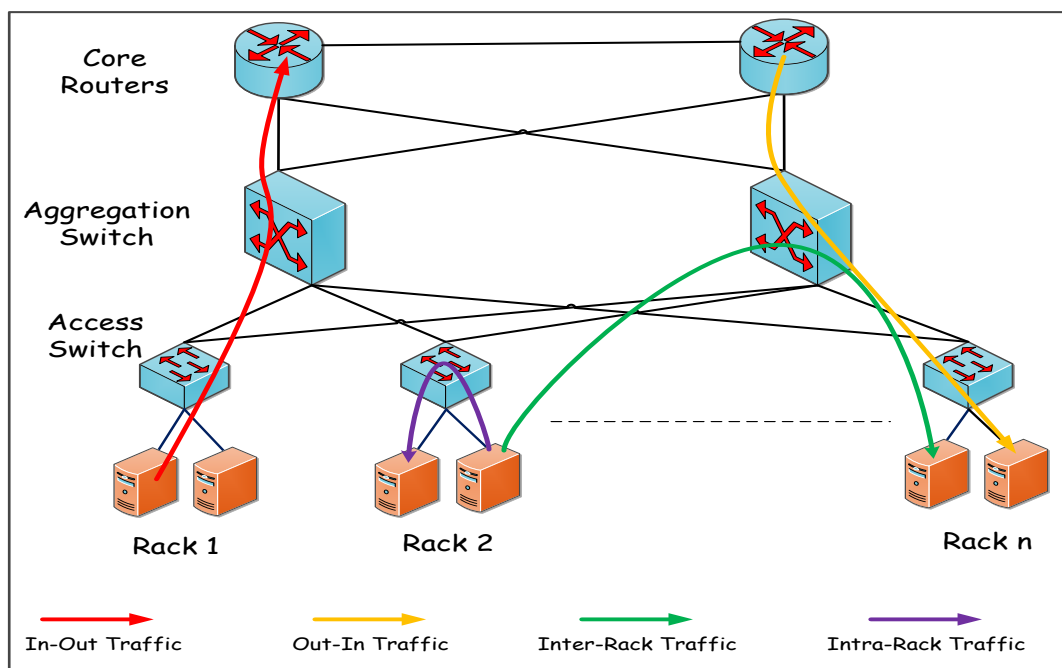


Figure 4.3. 3-Tier data centre traffic flow classification

4.5 Proposed PON architectures for future data centres

In this section, a number of novel PON designs to support connectivity inside data centres are described. In addition to the use of only passive optical devices, other challenges facing such designs include i) off-loading the inter-rack traffic from the OLT switch to avoid undesired power consumption and delays and ii) reducing or eliminating the need for expensive tuneable lasers.

4.5.1 Design Options 1 and 2: PON designs for data centres adopted from Fttx deployments

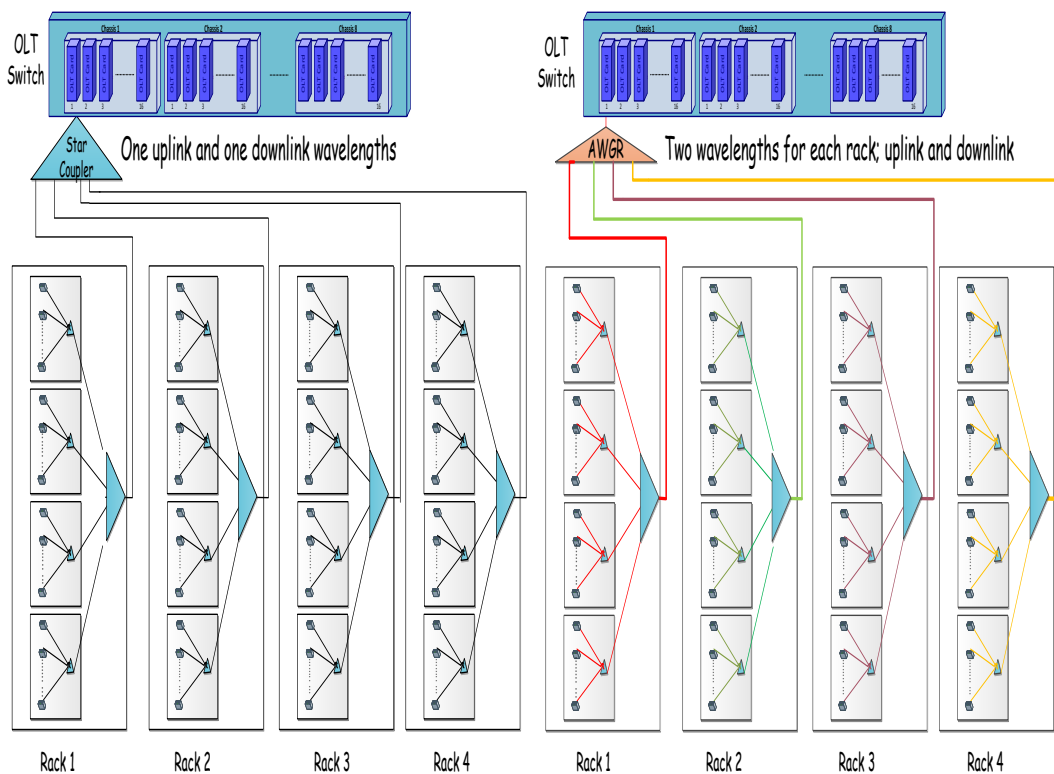


Figure 4.4. (a) Option design 1: TDM-PON DCN architecture, (b) Option design 2: hybrid WDM-TDM PON DCN with multi-carrier generator

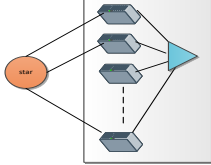
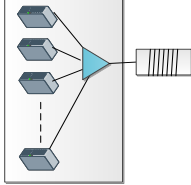
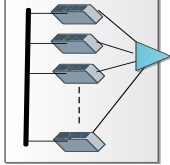
Figure 4.4 illustrates two designs of the PON interconnection fabric inside data centres adopted from the FTTx deployments. For simplicity, the architecture of a single PON cell is presented, i.e. the connection

provisioned by a single OLT card port to connect 128 servers distributed in four racks. The designs depicted in Figure 4.4 eliminate the need for access and aggregation switches. Figure 4.4(a) is a demonstration of a TDM-PON data centre where only passive star splitters/couplers are used to provide interconnection and a pair of wavelengths for upstream and downstream flows is shared among the 128 servers distributed in the four racks. The data centre connectivity can also be based on a hybrid TDM-WDM PON with a combination of star splitters/couplers and AWGR as shown in Figure 4.4(b). The hybrid TDM-WDM PON reduces congestion and facilitates more bandwidth for each rack. To reduce the cost of laser diodes at the OLT and avoid them at the servers, the TDM-WDM PON can be designed with multicarrier generator. At the server end, low cost multimode transceivers can be employed to directly modulate the carrier signal received from the OLT for the upstream transmission.

As discussed in Section 4.4, the nature of application and traffic locality within data centres requires servers to communicate and exchange information. In Table 4.1 three different designs based on optical passive devices are introduced to manage intra-rack communication without the need to reach the OLT switch.

Table 4.1. Comparison between the proposed technologies for Intra-rack communication

Design	Passive Star Reflector [36]	Passive Fibre Brag Grating [36]	Passive Polymer Optical Backplane [79]

Interconnection model			
Additional Requirements	Additional Multi-Wavelength (MW) transceiver and additional wiring	Additional MW transceiver or OFDM transceiver	Rack with passive polymer optical backplane
Drawbacks	Cost, wiring complexity, contention on MAC	Contention on MAC for first option and costly transceivers for OFDM option	Loss and need for regeneration for large racks
MAC Mechanism	Unidirectional TDMA	TDMA for the additional transceiver and OFDM for the second option	Terabit capacity Non-blocking Full mesh interconnectivity and no MAC is required

The first proposed design uses a passive star reflector to connect servers within a rack allowing each server to broadcast to other servers using an additional transceiver. The main limitation of such a design is the complexity of the MAC protocol needed to coordinate and arbitrate channel access.

Another solution to support intra-rack connectivity is to deploy a Fibre Brag Grating (FBG) after the star coupler connecting the servers in the rack to reflect a dedicated wavelength assigned for intra-rack communication

traffic. To facilitate the use of the FBG for the intra-rack communication, each server can be equipped with a second multimode transceiver. OFDM technology can be used to allow a single transceiver to generate multiple carriers, one for intra-rack communication and another for connections to the OLT or other racks. However, the expensive OFDM transceivers will increase the deployment cost of the PON design.

A third alternative which is found more practical for intra-rack communication is the Passive Polymer Backplane developed in [79]. This technology employs a passive backplane with multimode polymer waveguides and can provide non-blocking full mesh connectivity with 10 Gb/s rates per waveguide, exhibiting a total capacity of a 1 Tb/s. Figure 4.5 presents a 10 cards backplane layout with 100 waveguides.

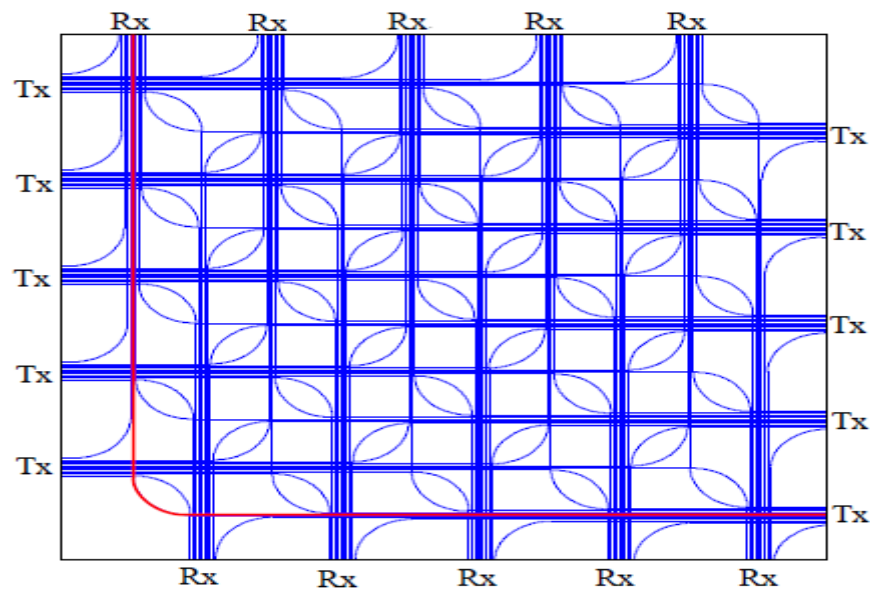


Figure 4.5. Schematic of 10x10 passive polymer optical backplane [44]

The main limitation of the designs depicted in Figure 4.4 is the need to forward all inter-rack traffic for racks within the same PON cell and rack in

different cell to the OLT switch, which buffers, processes, and reroutes traffic to the destination servers. Forwarding through OLT introduces delays and undesired power consumption. Therefore, new designs are proposed for PON cells to reduce or avoid the forwarding of the inter-rack traffic to the OLT for communication within a PON Cell.

4.5.2 Design Option 3: PON designs for data centres with servers equipped with tuneable lasers

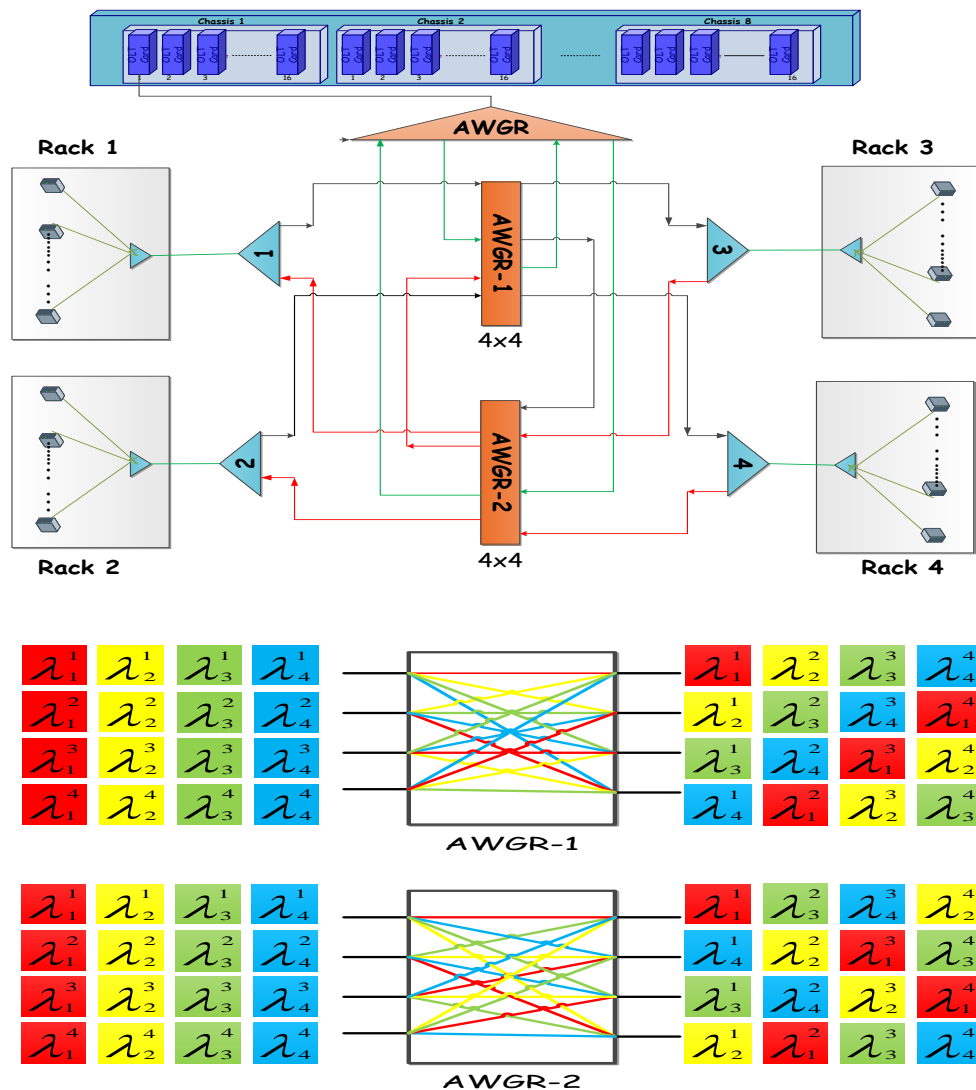


Figure 4.6. Option-3 PON Data centre interconnection for PON Cell employing servers equipped with tuneable lasers and utilizing AWGR and star coupler/splitters for fabric interconnection

In this design, as depicted in Figure 4.6, the PON Cell is equipped with two AWGRs to provision full interconnection between the 4 racks, each of 32 servers, using 4 wavelengths. The connection between the racks and the OLT is established via a 1: N AWGR.

The network interface card of each server is equipped with an array of photo detectors and tuneable lasers for wavelength detection and selection. Inter-rack communication can be provisioned either via the OLT switch or directly through the AWGR where a wavelength is selected for transmission based on the location of the destination server. Alternative routes facilitate multi-path routing and load balancing at high traffic load, however; as mentioned previously the OLT switch traffic forwarding should be avoided if possible to reduce delay and power consumption. A server can reach servers in other racks in the same cell via tuning its transceiver to the proper wavelength that matches with the AWGR wavelength routing interconnection map.

AWGR interconnection configuration for wavelength routing is shown in Figure 6b. For a server in rack 1 to communicate with a server in rack 2, its transceiver has to tune to wavelength 2 which will be input to AWGR-1 input-1 (λ_2^1) and forwarded to output 2 to be transmitted to AWGR-2 at input 1 and then forwarded to output 4. To establish a connection through the OLT switch, servers in rack 1 should tune to wavelength 3 (λ_3^1) that will be routed to output 3 of AWGR-1. This design is similar to the cellular network, in the essence that wavelengths can be reused to connect other racks connected to different OLT ports. Intra-rack communication can be provisioned using one of the previously described techniques in Table 4.1.

The main drawback of the design is its high deployment cost as all servers are equipped with tuneable transceivers. The following two designs are presented for PON cells to reduce or eliminate the need for tuneable lasers.

4.5.3 Design Option 4: PON designs for data centres with few tuneable lasers

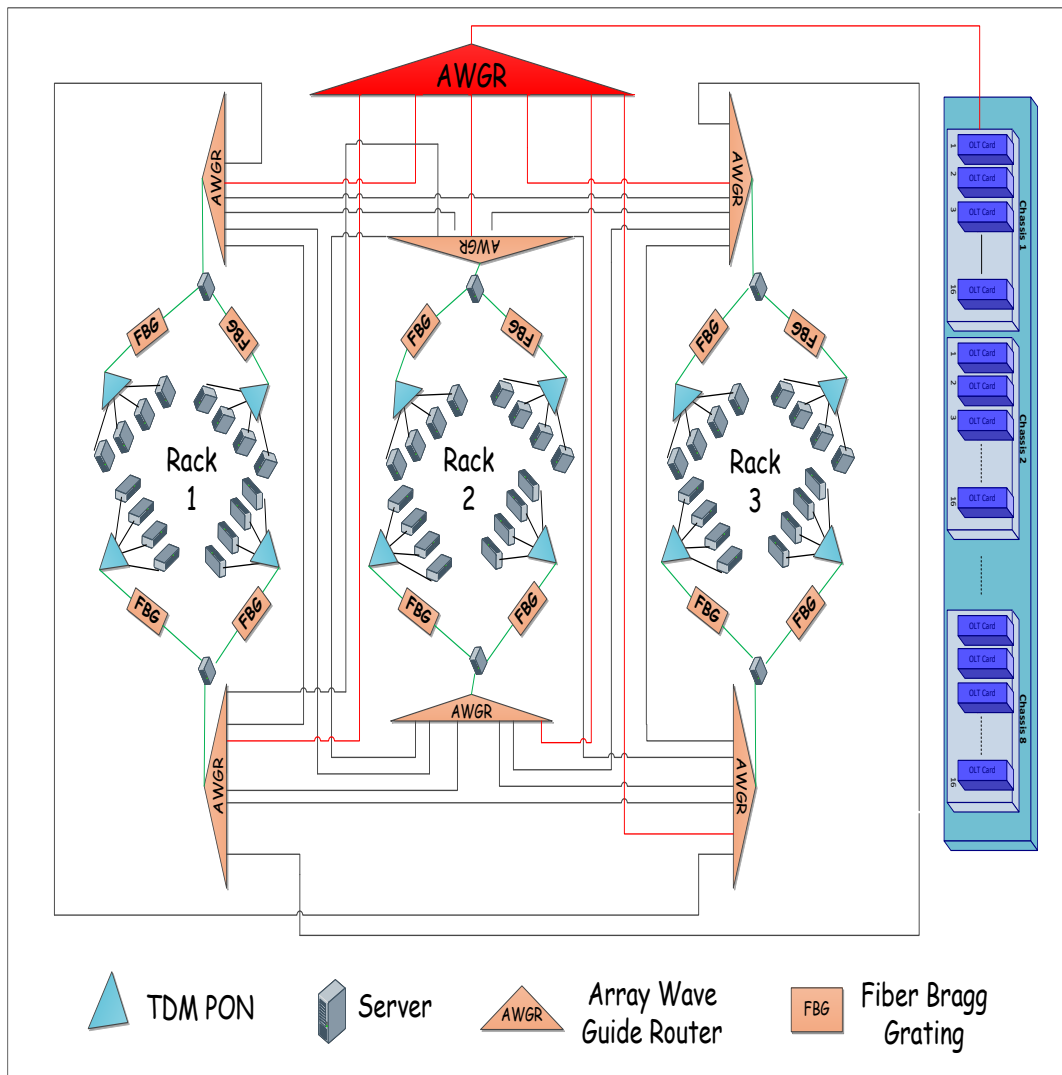


Figure 4.7. Design option 4: PON DCN with few servers equipped with tuneable lasers

This design aims to reduce the number of tuneable lasers needed to provide inter-rack interconnections by dividing racks into groups and

employing a special server to manage inter-group communication within the same rack and in different racks. In this design, as shown in Figure 4.7, each rack hosts 32 servers distributed in multiple groups. A group can host N servers where N is the splittings ratio for each TDM PON. For a rack of 32 servers, the design can have four groups each of 8 servers.

The connectivity within a group is maintained by reflecting the wavelength selected for intra-group communication using a FBG. As the reflected wavelength does not propagate out of the rack or into other groups within the same rack, the same wavelength can be used for all groups in different racks. This will simplify and unify the design of the transceivers for all servers. The same wavelength can be used for transmitting and receiving as it is a one way communication for the reflected wavelengths. To avoid collision and to manage contention on channel access, servers have to send a control message to the special server located after the FBG to gain permission to the intra-group communication wavelength.

Each group is assigned two wavelengths, one for uplink link and one for downlink transmissions. The specialised servers maintain a database of servers' addresses in the groups and wavelengths assigned to each group and can perform wavelength conversion to facilitate inter-rack communication which can take place either through the AWGRs as full mesh connectivity exists, or by approaching the OLT card. Specialised servers periodically exchange updates and status of their connectivity to update their databases. These updates can be exchanged between the servers directly or through the OLT.

The number of groups per rack can be reduced to two groups each of 16 servers to reduce the number of AWGRs, TDM-PONs, and the wiring complexity. However this will increase the load on the specialised servers which are the focal point of the design.

4.5.4 Design Option 5: PON designs for data centres without tuneable lasers

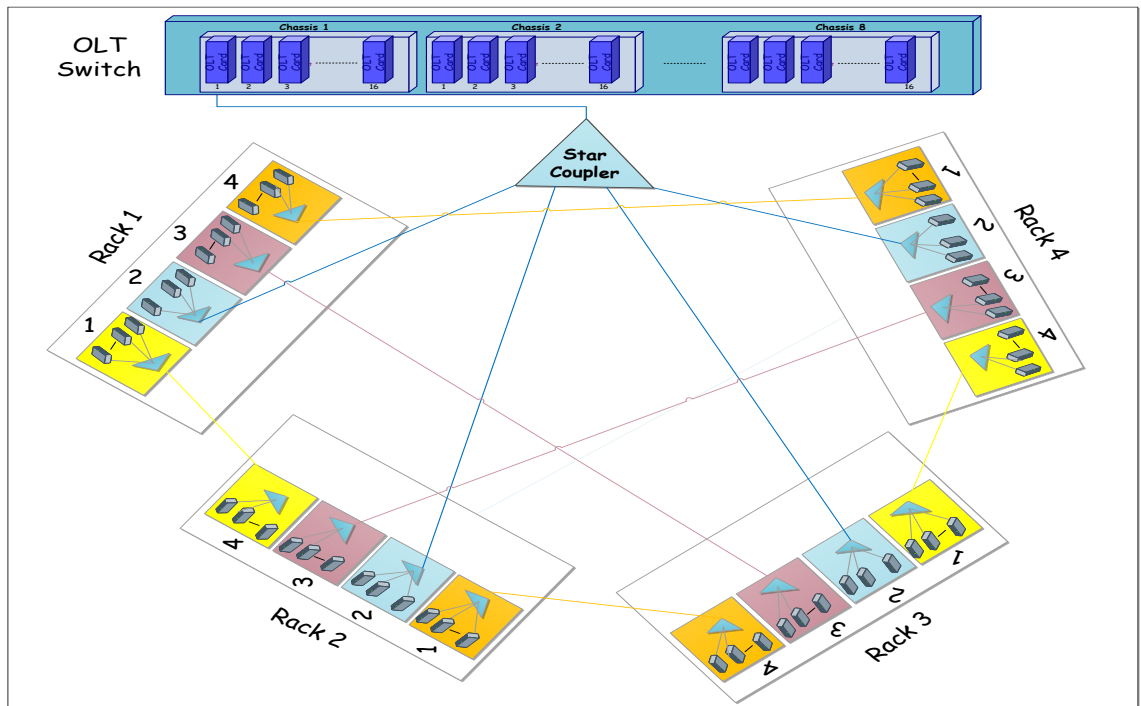


Figure 4.8. Design option 5: PON DCN with no tuneable lasers

The design, depicted in Figure 4.8, eliminates the need for tuneable lasers and facilitates high speed interconnection among racks within a PON cell by dividing each rack into 4 groups each of 8 servers connected by a TDM star coupler. Servers of different groups in the same rack are connected via Terabit capacity passive polymer optical backplane. Three of the 4 groups are connected to the other three racks, and one group connects the rack with the OLT. Inter-rack communication between servers

that do not belong to groups with a direct connection is established by using one of the servers of the group with a direct connection with the rack of the destination server. Relay server selection can be based on servers' utilisation or traffic load within the group. Similarly servers can establish connection with the OLT switch. Connections between the OLT switch and racks can be provisioned either through the use of a star coupler (TDM) as depicted in Figure 4.8 or via an AWGR (WDM) to facilitate more bandwidth.

4.6 Comparison and discussion

Table 4.2 summarises the key differences and similarities among the new PON architectures for future data centres proposed in this article. In terms of processing and memory capabilities, all the proposed PON designs can deliver an infrastructure suitable for modern applications such as cloud distributed computing with transparent and efficient resource assignment as per clients' needs. In addition, all the designs are scalable in terms of wavelength requirements to facilitate high speed interconnection (up to 10 Gb/s and beyond) among servers within a PON cell, where the same wavelengths can be reused in other PON cells, thus ensuring scalability in a fashion similar to cellular wireless and residential PONs. For control and management mechanism, all the proposed designs can make use of the MAC methods used in access networks (TDM, WDM and Hybrid TDM-WDM). The single exception here is that with option 4, a centralised server that is responsible for routing and coordination of channel access assignment and arbitration is introduced.

Design options 1 and 2 are simple architectures based on current FTTX deployment technologies, however; inter-rack communication has to be processed through the OLT switch creating a bottleneck as conventional PONs and OLT switches were not designed to handle large peer-to-peer (server-to-server) traffic between ONUs. As for inter-rack flows, the deployment of AWGRs in design option 3 is a premium solution to provision high speed interconnections that provide multi-path routing for inter-rack traffic. However, the drawback of such a design is its high cost incurred as a result of the expensive tuneable lasers required for each server. Design option 4 reduces the number of tuneable lasers required to support inter-rack communication by deploying special servers. Design option 5 is the most attractive solution as no tuneable lasers are deployed and multi-path routing for server to server communication is supported.

Table 4.2. Comparison of the five PON cell design options

	Option 1	Option 2	Option 3	Option 4	Option 5
Tuneable lasers	None	None	All servers	Few (4 servers)	None
Inter-rack traffic forwarding	Via OLT	Via OLT	Via AWGR or OLT	Via AWGR or OLT	Via relay servers or OLT
Intra-rack traffic forwarding	FBG /Star reflector/ Optical backplane	FBG /Star reflector/ Optical backplane	FBG /Star reflector/ Optical backplane	FBG / /Star reflector /Optical backplane	FBG / Star reflector /Optical backplane

Fabric interconnection	PONs	PONs	PONs	PONs and servers	PONs and servers
Multi-routes and Load balancing	No	No	Yes	Yes	Yes
Number of wavelengths/ Scalability	2, one uplink and one downlink	8, a pair for each rack	4	6 for inter-rack + 1 FBG + 2 for each rack	2 for connection between group pairs + 2 for TDM PON to OLT, or 8 for WDM PON to OLT
Wiring complexity	Low	Low	Moderate	High	Moderate
Management and control mechanism	Via OLT	Via OLT	Via OLT/ control system	Via OLT and centralized servers	Via OLT

4.7 Summary

In this chapter the deployment of PON based architectures to provide energy efficient, high capacity, low cost, low latency, scalable, and highly elastic networking infrastructures to sustain the applications and services hosted by modern data centres is discussed. Five designs for PON deployment in data centres are proposed. Two of these designs are based on current FTTx deployment technologies, however inter-rack communication has to be processed through the OLT switch. To facilitate high speed interconnection among racks within a PON cell, we proposed an architecture where each server is equipped with an array of photo detectors and tuneable lasers for wavelength detection and selection. Another design is proposed to reduce the need for expensive tuneable lasers by deploying special servers to manage and perform the wavelength conversion needed to support inter-rack traffic connections. The final proposed design is the most cost efficient as it eliminates the need for tuneable lasers and facilitates high speed interconnection among racks within a PON cell. We have also proposed different methods to facilitate intra-rack communication using star reflectors, FBGs, and a passive optical backplane.

5 Energy and cost efficient AWGR-based PON data centre architecture (Design-Option 3)

5.1 Introduction

In this chapter, a detailed description of the energy and cost efficient AWGR based PON data centre architecture (Design-Option 3) is given. A mathematical optimisation model for the routing and wavelength assignment problem within the AWGR PON cell fabric is presented. The issue of link oversubscription and the method to improve it within the PON cell are discussed. Further reduction in the deployment cost by making use of existing intermediate AWGRs and avoiding the use of additional optical backplanes, FBGs or star reflectors will be described. Finally, a benchmark study of the cost and power consumption of is presented for the AWGR based PON design compared to the most well-known architectures, the Fat-Tree and BCube.

5.2 Architecture of proposed AWGR-based PON Cell

In this Section, the PON cell of the AWGR based PON data centre architecture is described in detail. In the design depicted in Figure 5.1, the PON Cell is equipped with two intermediate AWGRs to provide full interconnection among the PON cell racks. Each AWGR has 4 inputs and 4 outputs. One of the outputs of AWGR-1 connects to AWGR-2. One output of

AWGR-1 connects with the OLT. Two outputs of the AWGR-1 connect to 2 of the 4 racks. Similar connections for outputs of AWGR-2 and for the inputs of AWGR-1 and AWGR-2 are followed. This pre-choice of connections is followed by the wavelength routing optimisation as will be demonstrated in Section 5.3. The connection between the PON groups and the OLT is established via the two AWGR. For simplicity, directions from/to the AWGRs connecting to the OLT will be drawn together in one AWGR like in Figure 5.3.

Each rack can host 32 servers divided into a number of PON groups. In the design shown in Figure 5.1 the whole rack is placed in one group. The traffic among servers in the same group is referred to as intra-group traffic while the traffic among servers in different groups is referred to as inter-group server traffic. The number of wavelengths used for inter-rack communications in a PON cell is proportional to the number of PON groups.

Servers can be either connected to a tuneable ONU as shown in Figure 5.2 or equipped with a network interface card PCI-e that is equipped with an array of fixed tuned receivers and a tuneable laser for wavelength detection and selection. Each ONU receives traffic destined to all ONUs connected to same splitter/coupler (broadcast). The ONU accepts traffic intended for the server it is connected to and discards other packets intended for other ONUs. Idle servers are switched off and ONUs connected to idle servers change from active to sleep state for further energy savings.

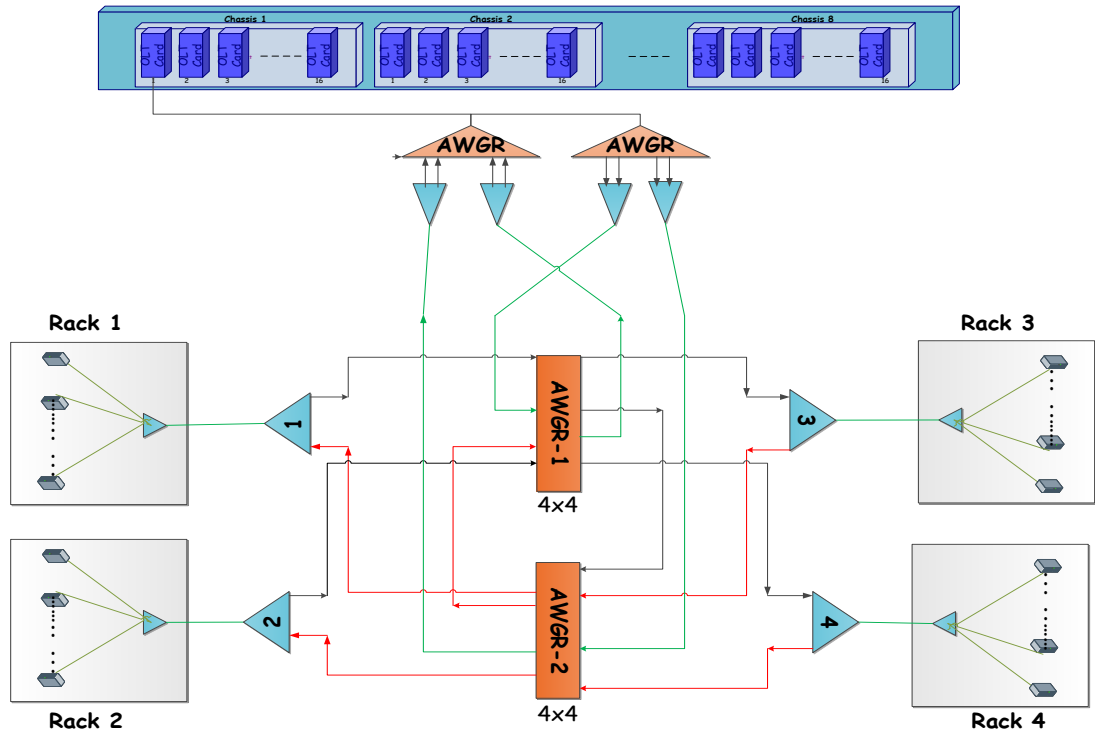


Figure 5.1. Architecture of proposed AWGR based PON cell

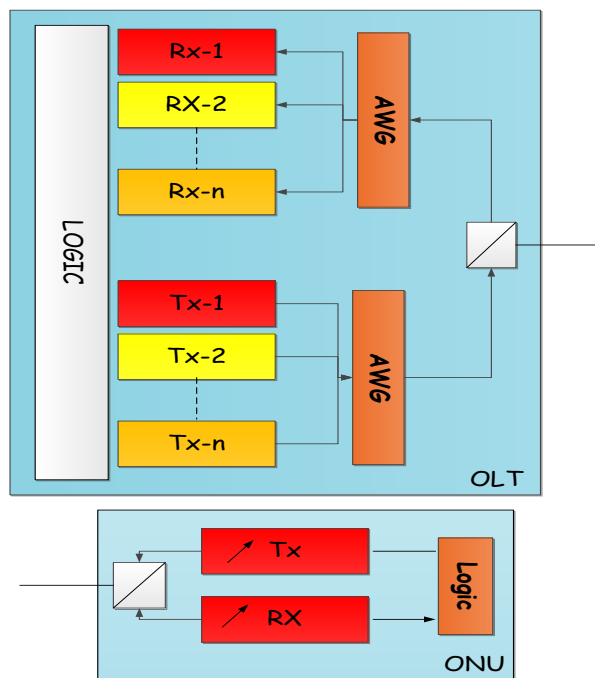


Figure 5.2. Architecture of the Optical line terminal (OLT) and tuneable ONU [36]

Inter-group communication within the PON cell can be provisioned either via the OLT switch or directly through the intermediate AWGR where a wavelength is selected for transmission based on the location of the

destination server. Alternative routes facilitate multi-path routing and load balancing at high traffic load, however, the OLT switch traffic forwarding should be avoided if possible to reduce delay as the OLT switch serves many PON cells at the same time and manages traffic coming in and leaving the data centre. A centralised scheduler in the OLT is responsible for scheduling inter group communication. Servers with demands send a request control message containing destination address and resources requirements to the scheduler. If the centralised scheduler grants the request, it replies to the source and the destination ONUs connecting the pair of servers with information about the assignment of time slots in the designated wavelength which both servers' ONUs need to tune to. Idle servers by default should be tuned to wavelengths connecting them with the OLT.

For intra group communication, one of the three different designs based on passive devices (backplanes, FBG, and star reflectors) proposed in Chapter 4 (Table 4.1) can be used. Later in this chapter we propose an improvement to the design that reduces the deployment cost by avoiding the use of these passive devices for intra group communication. The improved design demonstrates that intra group communications can be provisioned through the same intermediate AWGRs used for routing the inter group traffic by only employing one extra wavelength.

5.3 Network optimisation through Mixed Integer Linear Programming (MILP)

Linear programming is a mathematical modelling approach used for optimisation. Optimum solution with best outcome for an objective to either maximise a profit or minimise a cost can be achieved through linear programming by setting an objective function that can be subject to equality and inequality constraints. These constraints will further restrict the solution within the feasible region where a minimum or a maximum solution can be obtained through the intersections of the linear expressions represented by the equality and the inequality constraints [80]. The standard form of an objective function can be represented as follows:

Maximize or Minimize

$$F = C_1 * X_1 + C_2 * X_2 \dots \dots \dots C_n * X_n \quad (5.1)$$

where C_n is the coefficient and X_n is the variable. Usually the variables are continuous; however in some cases variables can be discrete and can be constrained to integers. In such cases the problem is called Mix Integer Linear Programming (MILP).

In order to restrict the values of the variables, we need to use constraints. Constraint expressions can take the following form:

$$a_{11} * X_1 + a_{12} * X_2 \dots \dots \dots a_{1n} * X_n \leq b_1 \quad (5.2)$$

$$a_{21} * X_1 + a_{22} * X_2 \dots \dots \dots a_{2n} * X_n \leq b_2 \quad (5.3)$$

Non Negative Constraints:

$$X_1 \geq 0, \quad X_2 \geq 0, \quad X_n \geq 0 \quad (5.4)$$

The idea behind using linear mathematical programming in designing a network is to represent the basic requirements of the network in the form of a cost function that relates to a quantity to be optimised (for example power minimisation) and constraints, where the latter constitute a feasible region through mathematical linear expressions. The solver program (CPLEX [81] running on a 2.5 GHz PC with 16GB RAM is used in this thesis) searches within the feasible region and selects the optimum values of the variables that will maximise or minimise a defined objective function. Optimisation problems in communications are mostly used to tackle issues such as delay, congestion, and energy consumption to better design networks in terms of interconnection topology or flow routing. A network can be presented as a graph that consists of nodes and links. A node is the entity that is capable of switching or routing flows (intermediate node) and at the same time can be a transmitting source (source node) or a receiving destination (destination node) where flows need to be terminated. Links are the communication channels that interconnect the nodes to constitute the graph or the topology and links can be either directed or undirected. Links have a capacity that should not be exceeded by accumulated flows. The link capacity is expressed in Gb/s.

For every network design problem, constraints such as capacity constraints, flow conservation constraints, and demand satisfaction constraints are essential as they form the basic requirements and functions of the network. The capacity constraints state that the total flows passing

through each link should not exceed its capacity. Where the flow conservation constraints ensure that the total traffic going into a node is equal to the total traffic leaving the same node for the case the node is neither a source nor a destination. If the node is a source, then the traffic out of the node minus the traffic entering the node is equal to the demand size that originates in that node. If the node is a destination, the net traffic in minus the traffic out of the node is equal to the amount of traffic destined to that node.

Tools

CPLEX along with AMPL are used to solve the mixed integer linear programming problem. CPLEX was developed and is maintained by IBM. AMPL was developed in Bell laboratories and is a powerful algebraic modelling language for optimisation problems. We used CPLEX version 12.5 that offers a 64 bit architecture support with unlimited memory use. Then we use AMPL to interact with CPLEX where the model and the data files are written in AMPL language and then fed to the CPLEX solver. AMPL is used again to analyse the results once CPLEX solves the problem.

5.4 MILP model for wavelength routing and assignment within a PON cell

In this section, a MILP model is developed to optimise the interconnection fabric within the cell and provide the configuration for wavelength routing to facilitate the inter rack communication among all servers located in different PON groups within the PON cell.

Parameters and variables used in the model:

Parameters:

- N Set of nodes (AWGR's ports, PON groups and the OLT)
- P Set of PON groups and OLT
- W Set of wavelengths
- A_k Set of output ports of AWGR k
- B_k Set of input ports of AWGR k
- N_m Set of neighbours of node m
- (s, d) Denotes source and destination of a connection, $s, d \in P$
- (m, n) Denotes end points of a physical link where $m \in N$ and $n \in N$

Variables:

- φ_{sd}^{jmn} Defined as $\varphi_{sd}^{jmn} = 1$ if wavelength j on link (m, n) is used for a connection (s, d) , otherwise $\varphi_{sd}^{jmn} = 0$
- μ_{sd}^j Defined as $\mu_{sd}^j = 1$ if wavelength j is used for the connection (s, d) , otherwise $\mu_{sd}^j = 0$

The model is defined as follows:

Objective:

Maximise:

$$\sum_{s \in P} \sum_{\substack{d \in P \\ s \neq d}} \sum_{j \in W} \mu_{sd}^j \quad (5.5)$$

Equation (5.5) gives the model objective which is to maximise the total number of connections (wavelengths) among the PON groups and between PON groups and OLT through the intermediate AWGRs.

Subject to:

$$\sum_{j \in W} \mu_{sd}^j \leq 1 \quad (5.6)$$

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

Constraint (5.6) ensures that a single wavelength is selected for communication among the PON groups and between PON groups and OLT.

$$\sum_{\substack{s \in P \\ s \neq d}} \mu_{sd}^j \leq 1 \quad (5.7)$$

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

Constraint (5.7) ensures that each destination receives a different wavelength from each transmitting source.

$$\sum_{\substack{d \in P \\ s \neq d}} \mu_{sd}^j \leq 1 \quad (5.8)$$

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

Constraint (5.8) ensures that each source transmits to different destinations on a different wavelength.

$$\sum_{\substack{n \in N_m \\ m \neq n}} \varphi_{sd}^{jmn} - \sum_{\substack{n \in N_m \\ m \neq n}} \varphi_{sd}^{jnm} \quad (5.9)$$

$$= \left\{ \begin{array}{ll} \mu_{sd}^j & m = s \\ -\mu_{sd}^j & m = d \\ 0 & \text{otherwise} \end{array} \right\}$$

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

Constraint (5.9) is the wavelength continuity flow conservation constraint. It ensures that the flow going into a node in a certain wavelength leaves the node on the same wavelength for all nodes except the source and destination.

$$\sum_{s \in P} \sum_{d \in P} \varphi_{sd}^{jnm} \leq 1 \quad (5.10)$$

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

Constraint (5.10) ensures that a wavelength j is not used more than once on link (m, n) to connect (s, d) .

$$\sum_{s \in P} \sum_{d \in P} \sum_{n \in N_i} \sum_{j \in W} \varphi_{sd}^{jin} - \sum_{\substack{d \in P \\ d \neq i}} \sum_{j \in W} \mu_{id}^j \leq 0 \quad (5.11)$$

$$\forall i \in P$$

Constraint (5.11) ensures that the flow between a certain PON group pair is not relayed by any of the other PON groups.

$$\sum_{\substack{n \in B_k \\ s \neq d}} \varphi_{sd}^{jmn} \leq 0 \quad (5.12)$$

$$\forall s, d \in P, \forall m \in A_k \text{ and } j \in W$$

$$\sum_{s \in P} \sum_{\substack{d \in P \\ s \neq d}} \sum_{j \in W} \varphi_{sd}^{jmn} \leq 1 \quad (5.13)$$

$$\forall m \in B_k \text{ and } \forall n \in A_k$$

Constraints (5.12) and (5.13) are for routing within the AWGRs. Constraint (5.12) ensures that flows are only directed from input to output ports of the AWGR. Constraint (5.13) is for routing within the AWGRs. The constraint ensures that each input port of the arrayed waveguide must send only one wavelength for an output port of the same AWGR for every (s, d) .

$$\sum_{s \in P} \sum_{d \in P} \sum_{j \in W} \mu_{sd}^j = (X - 1)X \quad (5.14)$$

Constraint (5.14) ensures all demands are met. However, it can be removed and given that equation (5.6) has less than or equal to, therefore demand rejection maybe allowed depending on network resources availability.

The term $(X - 1)X$ is the total number of wavelengths needed to satisfy all demands where X is the total number of PON groups and $(X - 1)$ is the number of PON groups that need to be connected with X .

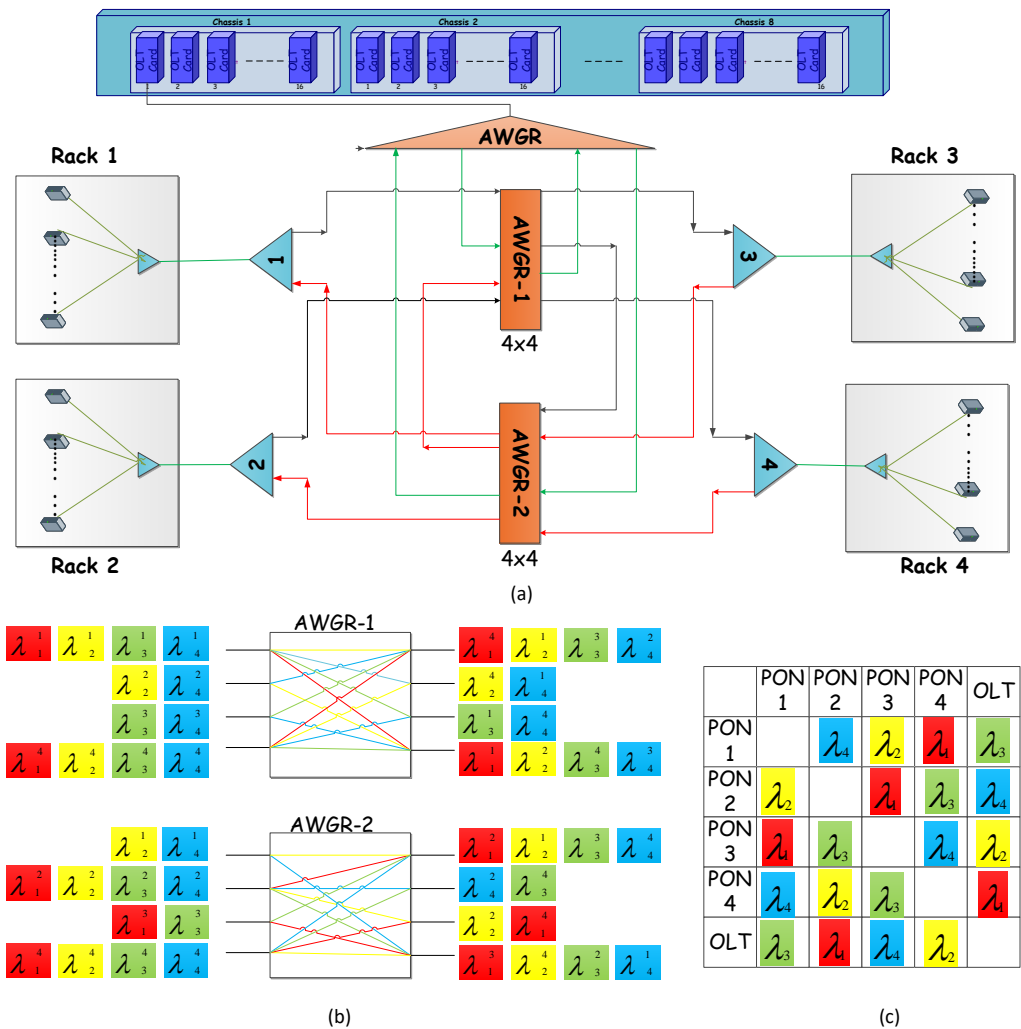


Figure 5.3. (a) Architecture of proposed PON data centre with tuneable lasers (b) Obtained MILP configuration for 4x4 AWGRs interconnection for wavelength routing (c) MILP obtained wavelength assignment for PON to PON and PONs to OLT communication

5.5 The wavelength routing and assignment results within a PON cell

Figures 5.3(b) and 5.3(c) present the configuration of the AWGRs fabric and wavelength routing table for the inter rack communications within the PON cell obtained through the MILP model for the architecture depicted in Figure 5.3(a).

This design is a Wavelength Routing Network (WRN) with $N+1$ entities (N racks and the OLT) to communicate with each other. In a WRN for $N + 1$ entities to communicate with N entities, we require either N fibres with N^2 wavelengths or N^2 fibres with N wavelengths. In our design, we select N wavelengths and employ the 2 AWGRs to represent the N^2 fibres (connections). For the architecture depicted in Figure 5.4 with $N=4$, 4 wavelengths are needed.

According to the wavelength routing table shown in Figure 5.3(c), if a demand exists between servers A and B located in groups 1 and 4, respectively. A request control message is sent to the OLT switch using wavelength 3 routed through AWGR-1 input port 1 to output port 3. If the OLT decides to grant the request, the OLT then replies with control messages to the two servers A and B using wavelengths 3 and 2 for racks 1 and 4 respectively. The control messages contain information about the wavelengths both servers need to tune to and assigned resources. Upon reception of the control information from the OLT switch, servers A and B tune their transceivers to wavelength 1. Idle servers by default should be tuned to wavelengths connecting them with the OLT.

The architectures depicted in Figure 5.4 and Figure 5.9 are for the AWGR based PON cell but with different number of PON groups. The rack is divided into a number of PON groups instead of having all servers within the same rack to be connected to the same coupler. The objective of increasing the number of PON groups is to reduce the number of servers connected to a single PON coupler. The impact of this reduction is to reduce the number of servers competing for channel access and resources, and also to reduce the broadcast traffic to be received by each ONU connected to the same PON coupler. There is no major impact on the cost if the number of PON groups is increased as these passive couplers are inexpensive.

The developed model is capable of designing AWGRs interconnection for as many PONs as required as long as ONUs are equipped with tuneable lasers that can be tuned to a sufficient number of wavelengths. In the architectures depicted in Figure 5.3 and Figure 5.4, the number of wavelengths required is 4 and 8, respectively, equal to the number of PON groups in a cell.

In Figure 5.4 the PON cell design has 8 PON groups with 8 servers each instead of 4 PON groups with 16 servers each.

For the different number of servers in the PON cell, as the number of PON groups increases the oversubscription rate decreases and the per server share of resources increases. For the case where the number of groups in a PON cell increases from 4 to 8, the oversubscription ratio is reduced by 50% and the per server share increases by 100%. For example, Figure 5.4 shows that for a PON cell with 64 servers, if the number of PON

groups increases from 8 to 16, the per server share of wavelength increases from 1.25Gb/s to 2.5Gb/s. As the number of PON groups increases, the number of wavelengths has to be increased to sustain all to all communication among all the PON groups. Figure 5.8 presents the number of wavelengths needed as the number of groups increases for the different number of servers that can be considered to be hosted in the PON cells.

Figures 5.3(b) and 5.3(c) present the resultant MILP configuration of the AWGRs fabric and wavelength routing table for the inter-rack communications within the PON cell for the architecture depicted in Section 4.5.2.

The architectures depicted in Figure 5.3 and 5.4 are for the same design but with different number of PON groups. The rack is divided into a number of PON groups instead of having all servers within the same rack to be connected to the same coupler. The objective of increasing the number of PON groups is to reduce the number of servers connected to a single PON coupler. The impact of this reduction is to reduce the number of servers competing for channel access and resources, and also to reduce the broadcast traffic to be received by each ONU connected to the same PON coupler. The developed model is flexible to design AWGRs interconnection for as many PONs as required as long as ONUs are equipped with tuneable lasers that can be tuned to a sufficient number of wavelengths. In the architectures depicted in Figure 5.3 and 5.4, the number of wavelengths required is 4 and 8 respectively, equal to the number of PON groups in a cell. There is no major impact on the cost if the number of PON groups is increased as these passive couplers are inexpensive.

The PON cell design can have 8 PON groups with 8 servers each instead of 4 PON groups with 16 servers each. This increase in PON groups reduces the number of servers competing for resources in each group hence reduces the per wavelength oversubscription for the point to point intra cell communication. This has led to resource assignment improvement for the point to point links by 100%.

For the different number of servers in the PON cell, as the number of PON groups increases the oversubscription rate decreases and the per server share of resources increases. For the case where the number of groups in a PON cell increases from 4 to 8, the oversubscription ratio reduces by 50% and the per server share increases by 100%. For example, Figure 5.7 shows that for a PON cell with 64 servers, if the number of PON groups increases from 8 to 16, the per server share of wavelength increases from 1.25Gb/s to 2.5Gb/s. An observation is that of increasing the number of PON groups, the number of wavelengths has to be increased to sustain all to all communication among all the PON groups. Figure 5.8 presents the number of wavelengths needed as the number of groups increases for the different number of servers that can be considered to be hosted in the PON cells.

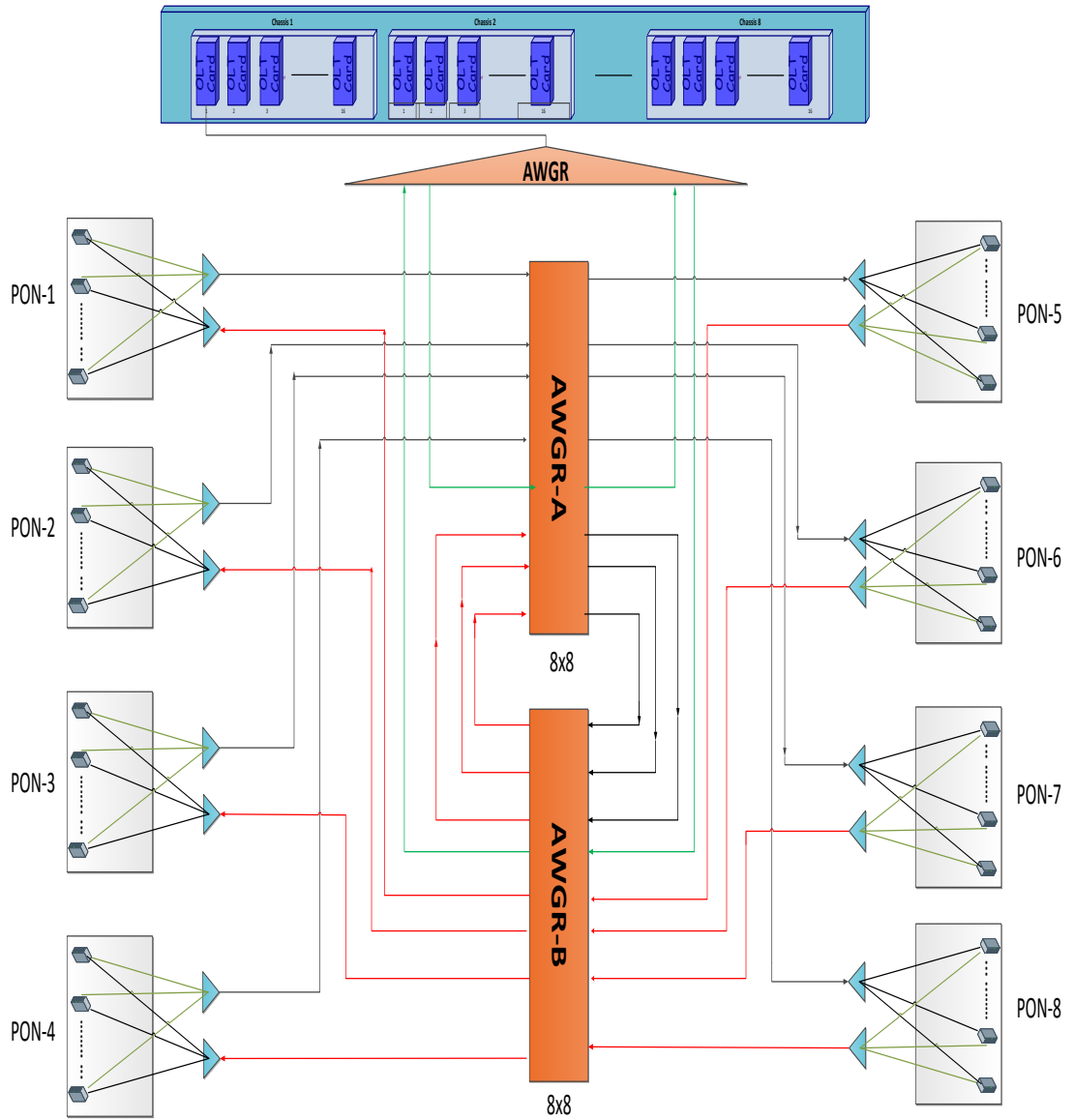


Figure 5.4. Architecture of proposed PON data centre with tunable lasers for 8 PONs groups

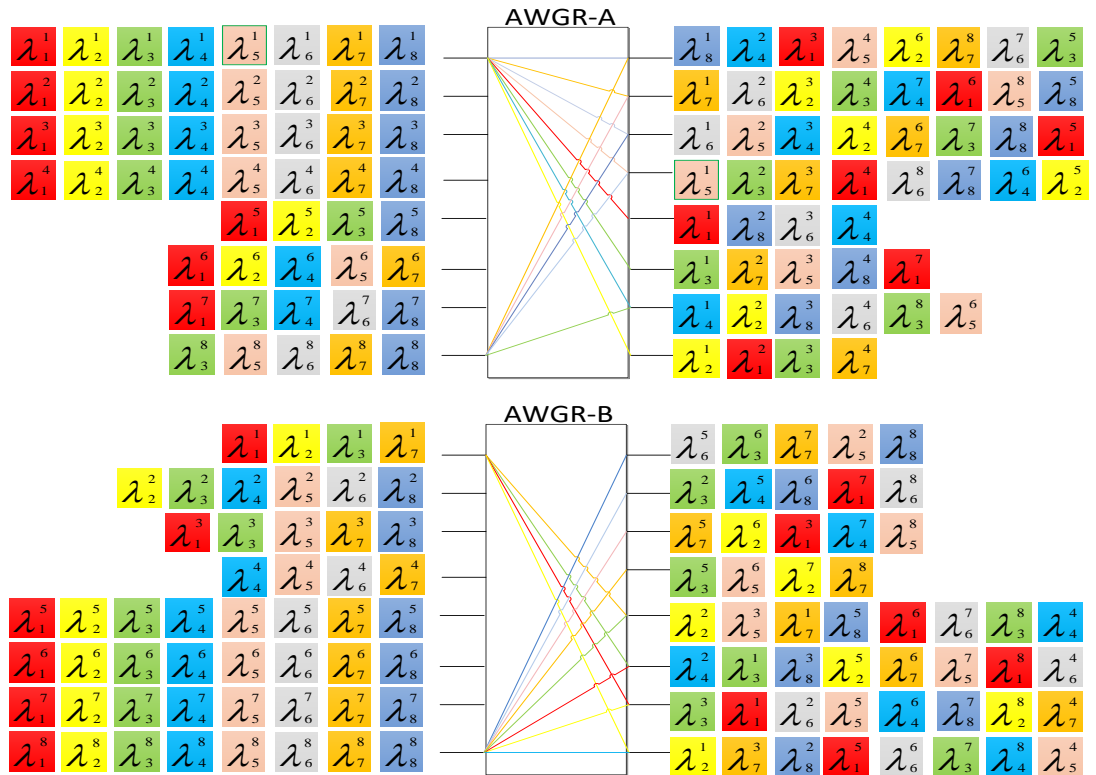


Figure 5.5. Obtained MILP configuration for 8x8 AWGRs interconnection for wavelength routing and assignment for the 8 PON groups design.

D \ S	PON 1	PON 2	PON 3	PON 4	PON 5	PON 6	PON 7	PON 8	OLT
PON 1		λ_4^A	λ_3^A	λ_2^A	λ_8^A	λ_7^A	λ_6^A	λ_5^A	λ_1^A
PON 2	λ_2^A		λ_1^A	λ_7^A	λ_4^A	λ_6^A	λ_5^A	λ_3^A	λ_8^A
PON 3	λ_5^A	λ_3^A		λ_8^A	λ_1^A	λ_2^A	λ_4^A	λ_7^A	λ_6^A
PON 4	λ_7^A	λ_8^A	λ_6^A		λ_5^A	λ_3^A	λ_2^A	λ_1^A	λ_4^A
PON 5	λ_8^B	λ_2^B	λ_5^B	λ_1^B		λ_4^B	λ_7^B	λ_6^B	λ_3^B
PON 6	λ_1^B	λ_7^B	λ_4^B	λ_6^B	λ_2^B		λ_3^B	λ_8^B	λ_5^B
PON 7	λ_6^B	λ_5^B	λ_8^B	λ_3^B	λ_7^B	λ_1^B		λ_4^B	λ_2^B
PON 8	λ_3^B	λ_1^B	λ_2^B	λ_4^B	λ_6^B	λ_5^B	λ_8^B		λ_7^B
OLT	λ_4^B	λ_6^B	λ_7^B	λ_5^B	λ_3^A	λ_8^A	λ_1^A	λ_2^A	

Figure 5.6. MILP obtained wavelength assignment for PON to PON and PONs to OLT communication for the 8 PON groups design

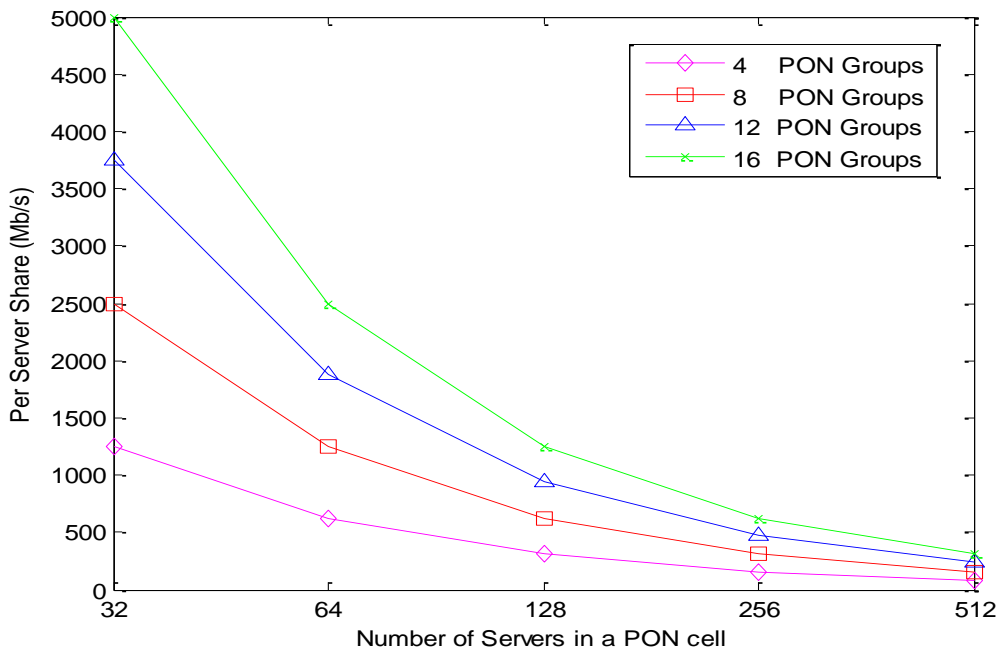


Figure 5.7. Worse-case server share of resources against different sizes of PON cell for 4,8,12, and 12 PON groups.

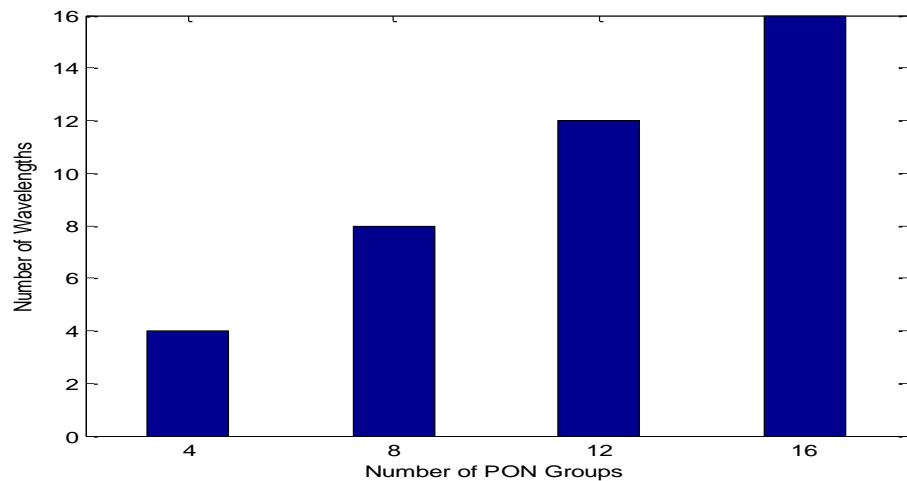


Figure 5.8. Number of wavelengths needed with respect to the number of PON groups in a PON cell

Figure 5.9 shows an architecture where intra-rack communication along with inter-rack can be provisioned via the intermediate AWGRs, which results in reduction in the implementation cost by avoiding the deployment of optical passive backplane or FBGs. The number of fixed tuned receivers for

each server in a PON cell is governed by the number of racks and number of OLT ports connecting the PON cell. In this design, N+1 entities (N PON groups and the OLT) need to communicate with each other and with themselves except one entity (OLT) does not need to talk to itself. Therefore, we need $(N+1)^2-1$ fibres / physical paths and N+1 wavelengths. For the architecture depicted in Figure 5.9 with N=6, 48 connections and 7 wavelengths are needed.

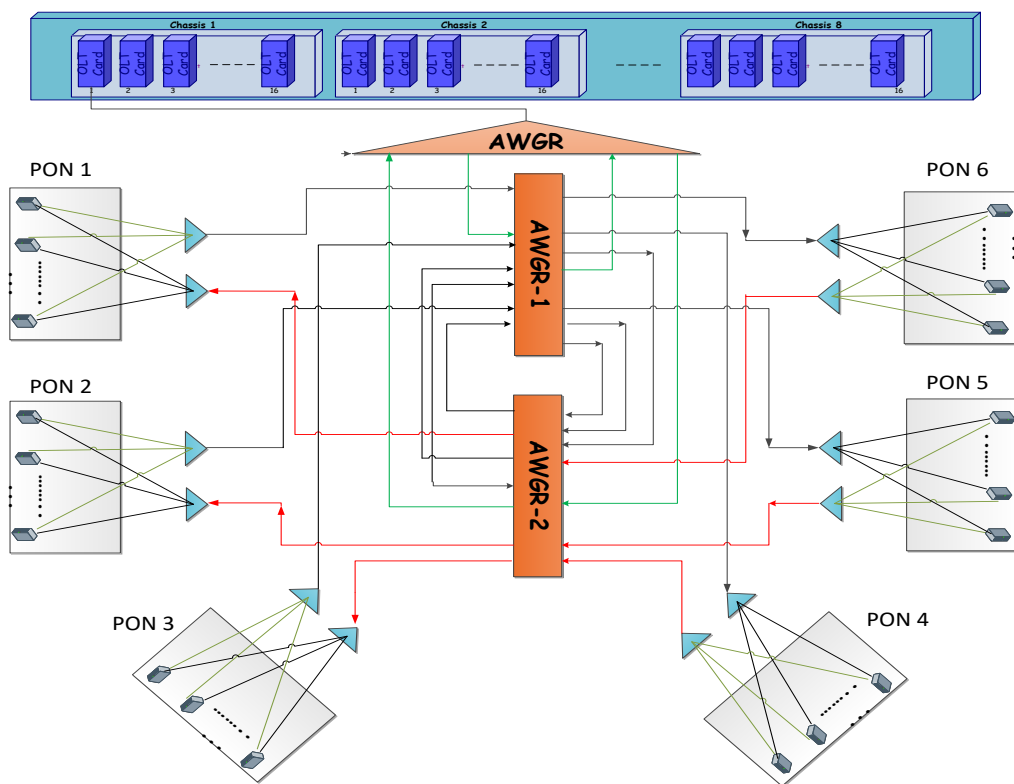


Figure 5.9. architecture of PON data centre with tuneable lasers for intra and inter rack communication through passive AWGRs

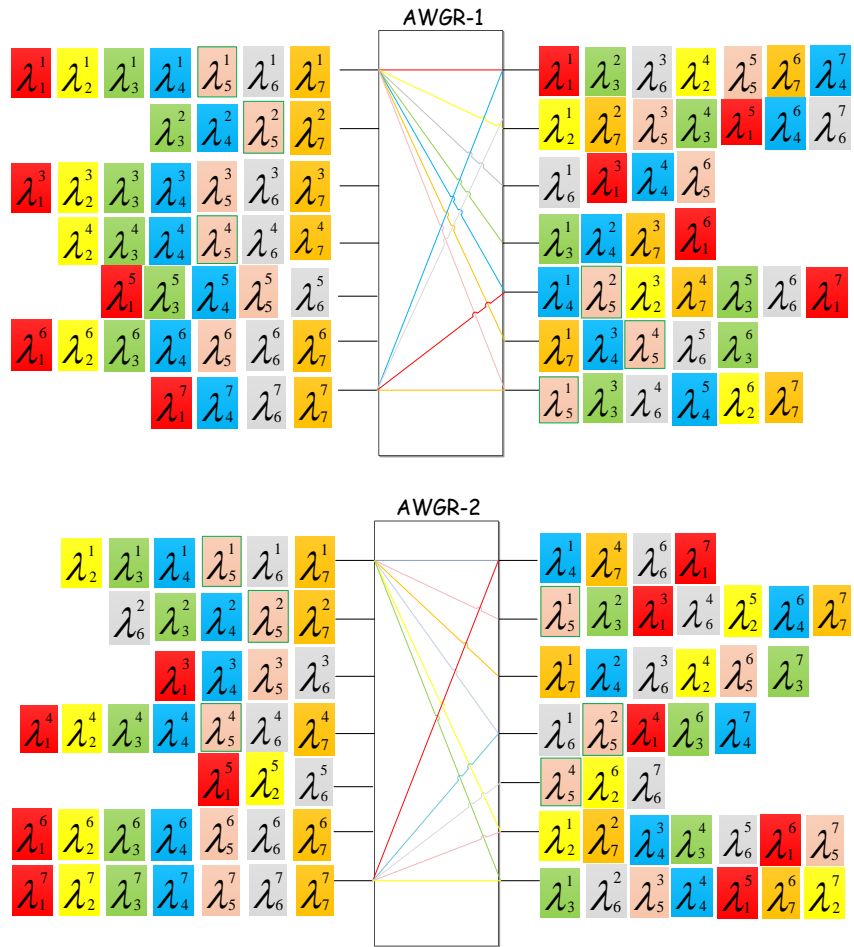


Figure 5.10. MILP obtained wavelength interconnection assignment for inter/intra communication

	PON 1	PON 2	PON 3	PON 4	PON 5	PON 6	OLT
PON 1	λ_5	λ_7	λ_6	λ_2	λ_4	λ_1	λ_3
PON 2	λ_3	λ_2	λ_5	λ_4	λ_6	λ_7	λ_1
PON 3	λ_1	λ_4	λ_3	λ_5	λ_2	λ_6	λ_7
PON 4	λ_7	λ_5	λ_2	λ_3	λ_1	λ_4	λ_6
PON 5	λ_4	λ_1	λ_7	λ_6	λ_3	λ_5	λ_2
PON 6	λ_6	λ_3	λ_4	λ_1	λ_7	λ_2	λ_5
OLT	λ_2	λ_6	λ_1	λ_7	λ_5	λ_3	λ_4

Figure 5.11. Obtained MILP configuration for 8x8 AWGRs interconnection for wavelength routing and assignment to provision inter and intra rack communication.

For intra-rack communications through AWGRs, assume a server in PON-1 wants to communicate with a server in the same group for the architecture depicted in Figure 3. A control message is sent to the OLT using wavelength 3 routed through AWGR-1 input port 1 to output port 4 that connects with the OLT switch. The OLT replies with a control message to PON-1 with wavelengths 2. The control message has information about communication wavelength and resources assigned for the communication. The source and destination servers in PON-1 tune their transceivers to wavelength 5. Traffic is transmitted through wavelength 5 traversing the link connected to AWGR-1 at input 1 exiting at output 8 of the same AWGR propagating into the second AWGR at input 1 and exiting at output 2.

5.6 Power consumption benchmarking of PON data centre design

In this section, a benchmarking study that compares the power consumption of our proposed PON data centre to the most common data centre architectures; the Fat-tree [27] and BCube [37] is presented.

5.6.1 Fat-Tree data centre architecture

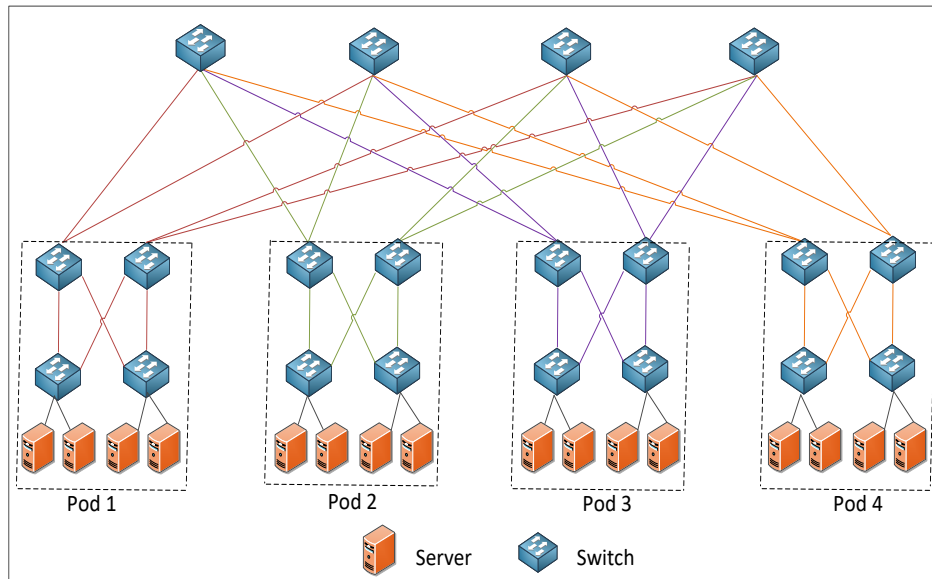


Figure 5.12. Fat-Tree data centre topology with $k=4$

The Fat-Tree data centre topology [27], depicted in Figure 5.12, is build using identical k -port low cost and low power commodity switches in access, aggregation and core layers. The Fat-Tree topology consists of k pods, each of which consists of $\frac{k}{2}$ access switches and $\frac{k}{2}$ aggregation switches, and hosts k servers. In each pod, interconnection between ToR and aggregation switches forms a complete bipartite. Similarly, another bipartite interconnection is formed between aggregation switches in each pod with all core switches. Fat-Tree topology with k -pods consists of $\frac{k^2}{4}$ core switches and can support $\frac{k^3}{4}$ servers.

The power consumption evaluation is carried out for different configurations of Fat-Tree switching fabrics. We evaluated power consumption for topologies with 24 and 48 pods using 24 and 48 ports commodity switches respectively. We considered Cisco commodity switches of 24 [82] and 48 [83] ports consuming 27W and 39W respectively. The

power consumption of a server's transceiver is 3W [84]. The network power consumption of Fat-Tree topology can be calculated using equation (5.14):

$$PC_k = PC_{pr} \left(\frac{k^3}{4} \right) + PC_{sw} \left[k^2 + \left(\frac{k}{2} \right)^2 \right] \quad (5.15)$$

where PC_{ser} is the server's power consumption, PC_{pr} is power consumption of server's port, PC_{sw} is the power consumption of the switch and k is the number of pods.

5.6.2 BCube data centre architecture

BCube data centre topology [30] is categorised as a server-centric structure as it employs multi ports servers to act as relay nodes to take part in routing and traffic forwarding decisions.

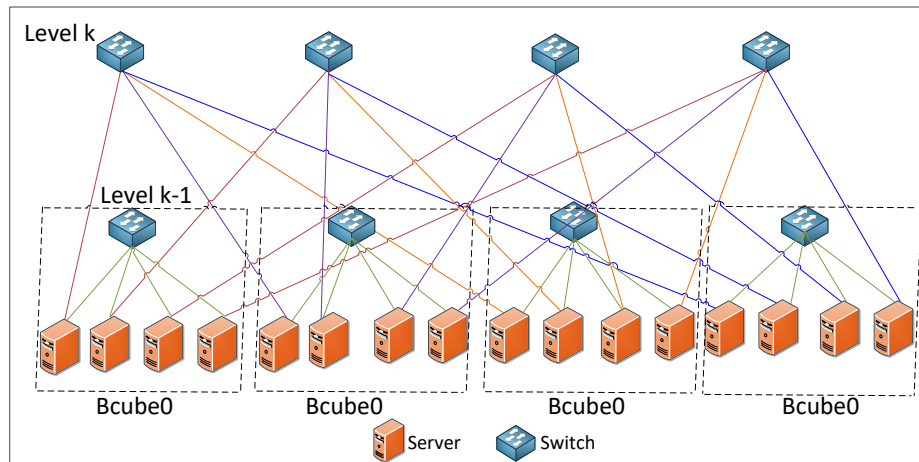


Figure 5.13. BCube data centre topology (BCube1) with $n=4$ and $k=1$

BCube as shown in Figure 5.13 is constructed in a recursive manner starting at BCube0 as its basic building block. A BCube0 consists of an n -port commodity switch connecting n -servers. A BCube1 is then constructed from n -BCube0 with a total of n , n -port commodity switches. The architecture in general is denoted as BCubek where $k+1$ defines the number of levels. For $k \geq 1$, the BCube topology is constructed from n BCube($k-1$)s, n^k n -port

switches and servers with $k + 1$ ports. In BCubek, the total number of servers is n^{k+1} and total number of levels is $k+1$ with each level consisting of n n -port switches. Figure (5.13) presents an example of BCubek structure with $k=1$ and $n=4$.

The network power consumption evaluation for BCube topology is evaluated for different fabric configurations for different values of k and n . The evaluated network power consumption for topologies with $k=2, 3$ and 4 for $n=8$ to provision connectivity for 512, 4096 and 32768 servers respectively is considered. Cisco commodity switches of 8 ports consuming 12W [85] is used for the evaluation and comparison study. The network power consumption of a BCube topology can be calculated using equation (5.15).

$$PC = PC_{sw} \left(\frac{(k + 1)n^{k+1}}{n} \right) + PC_{pr} \left((k + 1)n^{k+1} \right) \quad (5.16)$$

Note that the power consumption calculation of the B-Cube data centre architecture does not take into account the power consumption of the servers (CPU utilisation) taking part in the routing of inter rack traffic.

5.6.3 PON data centre architecture

The networking equipment of FTTx access networks is designed for long reach to connect subscribers (ONUs) located up to 20km away from the central office where the OLT is located. For PON data centre interconnections distances between racks and OLT will not exceed 100m typically. In addition, current ONUs are designed to support triple play services for video, audio and data while for the data centre only the data

service is required. Therefore the hardware requirements of ONUs and the OLT switch for a PON data centre are simpler than the FTTx, resulting lower power consumption. The power consumption of the ONUs and OLT is also a function of the transmission rate. 10 Gb/s tuneable ONU's TRX power consumption was reported to consume 2.5 W in [86].

To the best of our knowledge, no studies or vendors have provided power consumption specification for OLT PON transceivers that supports rates of 10G. GPON OLT (NEC CM7700S OLT) [65] supports 1G data rate for typical distance of PON system (20km) and consume 12.5W per port. For the proposed PON data centre, a linear profile for power consumption for 10G is assumed, one OLT port thus consumes 125W of. As a conservative estimate the same power consumption for a 10Gb/s port supporting a transmission distance of 100m is assumed.

5.6.4 Comparison and discussion

The evaluated PON architecture is depicted in Figure 5.6 and consists of 8 PON cells connected to an OLT card port where each card consists of 8 ports each of which can support a transmission rate of 10Gb/s. A PON cell with 64 servers, made up of 8 groups each of which hosting 8 servers, is considered. Intra group communication is maintained by an FBG and the same wavelength is used for all groups.

Networking equipment energy savings of the proposed PON architecture compared to Fat-Tree and BCube architectures are shown in Figure 5.14. The high energy consumption of BCube and Fat-Tree architectures is mainly due to the high number of switches used for the interconnections. As

discussed above, these switches are eliminated from the PON design and replaced by passive optical devices. Therefore the proposed PON architecture has reduced the power consumption by 45% and 80% compared to the Fat-tree and BCube architectures for 3,456 and 32,768 servers, respectively. The BCube architecture has the highest power consumption as it is a server centric architecture where servers are equipped with multiple transceivers needed to establish connectivity with all the levels. As the levels increase, the architecture can be scaled up to host more servers and the number of transceivers increases as each server needs to have connections with a switch in every level, hence the power consumption increases. The Fat-Tree architecture is a switch centric architecture and has lower power consumption as it is designed to have servers with single transceivers to connect to the ToR switch. The savings achieved by the proposed PON architecture compared to the Fat-Tree architecture decrease as of the number of server increases. This because the power consumption of the switches used to build the 24 pods and 48 pods Fat-Tree architectures does not increase linearly as the number of pods increases. On the other hand the power consumption savings achieved compared to the BCube architecture increases as the number of servers increases as a result of the increase in the number of transceivers needed as the architecture is scaled up to host a higher number of servers.

Note that the power consumption calculation of the B-Cube data centre architecture does not take into account the power consumption of the servers (CPU utilisation) taking part in the routing of inters rack traffic.

Considering cooling power draw for the huge number of switches in Fat-Tree and BCube architectures along with CPU utilisation for traffic forwarding in BCube, more savings can be demonstrated in the comparison. Moreover; if higher number of servers of 128 is introduced in the design for each PON cell, more savings can be obtained as double the number of servers can be served by each OLT port.

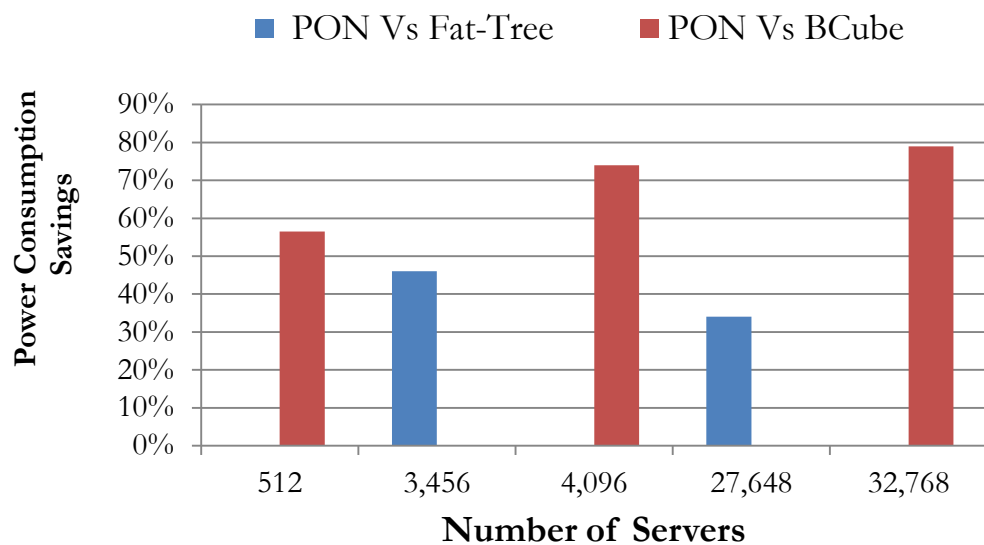


Figure 5.14. Power consumption saving for PON architecture against BCube and Fat-Tree architectures

In the next section, a study on CAPEX deployment cost for the PON architecture with benchmarking comparison with Fat-Tree and BCube architecture designs is presented.

5.7 Cost-based comparison between PON DC with Fat-tree and BCube architectures

An important consideration when deploying PONs in data centre interconnection is the CAPEX deployment cost involved. In this section, a

study on hardware cost of the proposed architecture is presented. A benchmarking study of the proposed architecture with well known architectures such as Fat-tree and BCube is presented. Table 5.1 presents the cost of the main components used for the PON and conventional data centre architectures. Equation 5.16 is used to compute the total cost of the PON data centre architecture.

Table 5.1. Cost Breakdown for the networking devices used in DC architectures

Equipment	Cost in Dollars
OLT chassis (C_{ch})	5000 [87]
System Controller Module (C_{SCM})	10,000 [87]
Switching Module (C_{SM})	25,000 [87]
Access Module (C_{AM})	15,000 [87]
10 Gb/s Burst-mode with tuneable TRX ONU (C_{ONU})	175 [86]
1 Gb/s Ethernet Transceiver (C_{eth})	74 [79] [88]
8 ports commodity switch (C_{SW8})	895 [89]
24 ports commodity switch (C_{SW24})	1525 [89]
48 ports commodity switch (C_{SW48})	2850 [89]

$$PON_{DC} COST = \left(\frac{S_{tot}}{S_{ch}}\right) (C_{ch} + C_{SM} + C_{SCM}) + \left(\frac{S_{tot}}{S_{AM}}\right) (C_{AM}) + (S_{tot})(C_{ONU}) \quad (5.17)$$

where S_{tot} presents the total number of servers, S_{ch} is the number of servers that can be connected to one OLT chassis. The architecture and prices of 10G OLT equipment is described in [87]. An OLT chassis consist of 8 Access Modules (AM), each of which has four 10Gbps ports. In addition to the AM modules, the OLT switch is equipped with a System Controller Module (SCM) and a Switching Module (SM) for management and traffic

switching. We assumed 64 servers are hosted in a PON cell and to be connected by one OLT port. CAPEX savings between the PON architecture and the conventional architectures; Fat-Tree and BCube are shown in Figure 5.15.

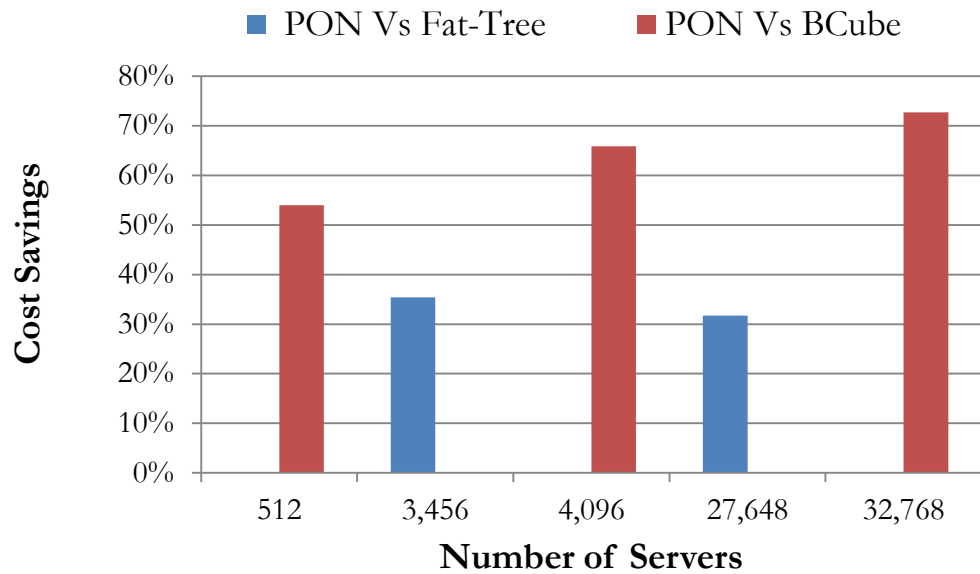


Figure 5.15. CAPEX saving for PON architecture against BCube and Fat-Tree

5.8 Summary

This chapter has discussed the deployment of the AWGR PON based architectures to provide energy efficient, high capacity, low cost, scalable, and highly elastic networking infrastructures to sustain the applications and services hosted by modern data centres. AWGR cellular based architecture does not suffer the limitations of the normal PON in access network as high per server rate, and multi-path routing can be achieved. A mathematical model was developed for wavelength routing and assignment within the

AWGRs PON cell fabric. Low oversubscription ratios are achieved for intra-PON cell as the worst-case per server rate for intra-cell communication can approach rates up to 5 Gb/s. The chapter also presented an improvement in the design for more reduction in the deployment cost as intra-rack communication can be managed through the existing two intermediate AWGRs and the use of additional hardware such as FBGs, optical backplanes, or star reflectors can be avoided. A benchmarking study has shown that our proposed PON architecture energy efficient design can reduce the network power consumption of data centres by 45% and 80% compared to Fat-tree and BCube architectures, respectively. Similarly, The proposed design has also shown CAPEX savings up to 40% and 76% compared to the Fat-Tree and BCube architectures, respectively.

6 Reduced Inter cell oversubscription through energy efficient software defined AWGR PON based Network

6.1 Introduction

In this chapter, the oversubscription issue in the inter-cell communication in the Arrayed Waveguide PON based data centre architecture is tackled. The improvement in the design is achieved by Introducing 2-tiers of AWGRs to facilitate multipath routing and energy-efficient utilisation of resources for inter-cell communication. This improved architecture will be presented and discussed. This chapter introduces a centralised SDN control and management system to coordinate and arbitrate the channel access for communication through the OLT links with PONs via wavelength reconfiguration and energy efficient grouping. MILP model along with detailed results on wavelength routing and assignment for the inter-cell communication is described. A benchmarking study that compares the proposed SDN architecture against the decentralised design is presented to show that with the SDN enabled architecture, the power consumption can be decreased by up to 90% for typical average data rates while maintaining zero blocking.

6.2 Modified architecture for reduced over subscription in Inter Cell Communication

Figure 6.1 shows a schematic of the upper level PON cell to PON cell connectivity in the AWGR based PON data centre design using only passive devices. In this design, each PON cell is only connected to a single OLT switch which has a maximum capacity because of the technology e.g. (16 ports per card and 16 cards per chassis). Demand pairs for servers located in two different PON cells cannot be grouped to a single OLT and have to be forwarded to core routers which increases the overall power consumption as more core routers ports and OLTs uplink ports are required. This increase in the traffic flow between OLTs and core routers will result in oversubscription for resources especially in the case of high data rates.

Oversubscription in the AWGR based PON data centre architecture can be resolved by introducing 2-tiers of AWGRs in the upper level connectivity to connect each PON cell to multiple OLT switches as seen in Figure 6.2. The main advantage of this design is its flexibility as it allows servers to join different OLT ports which enhance the performance in terms of resources provisioning and energy efficiency. An SDN enabled control and management system can be used to coordinate and arbitrate the channel access for inter-cell communication. Depending on the activity ratio of servers in different PON groups, the SDN enabled control and management system can, through reconfiguration, distribute the data centre traffic flows among different OLTs for load balancing or alternatively consolidate the traffic flows within fewer OLTs to save power through power shedding.

For simplicity, in Figure 6.2 only uplink connections from PON cells to the optical switches are shown.

The SDN controller allows any PON group to access any of the OLT switches through the introduced 2-tier of AWGRs by assigning the proper wavelength to the server with demand. All demand requests are forwarded to the SDN controller through the OLTs via a commodity switch (SW) for the assignment of transmission wavelengths and duration of transmission. Decisions made by the SDN controller are based on its global knowledge of the network parameters such as utilisations of uplink/downlink wavelengths between PONs and OLTs, active OLTs and core routers, utilisation of ports in OLTs and core switches.

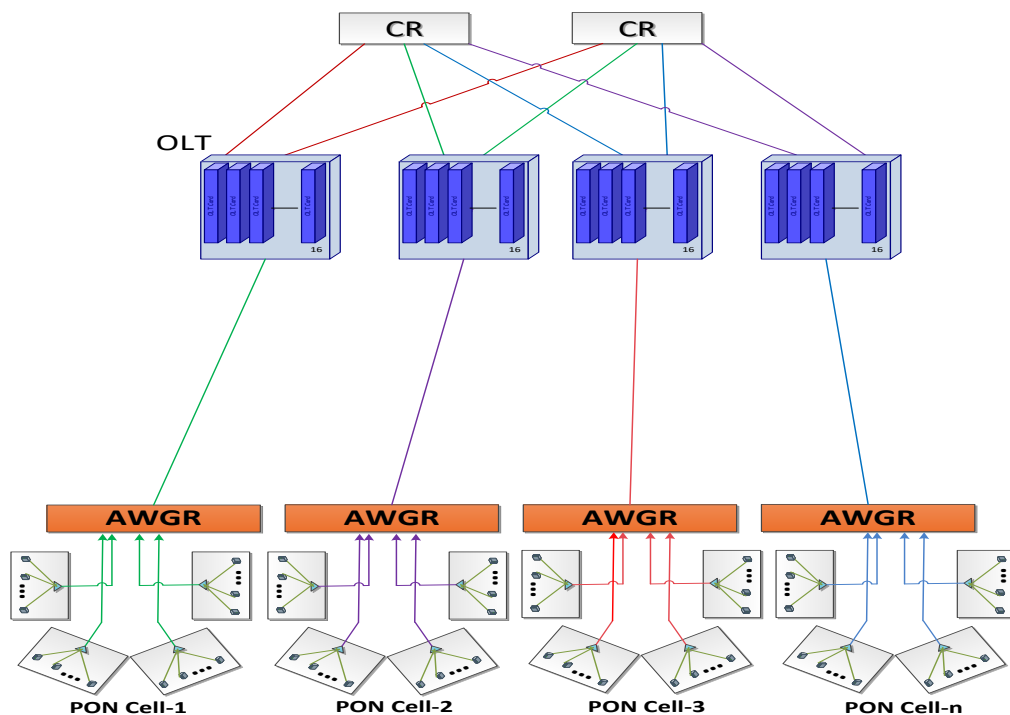


Figure 6.1. Upper level connectivity for the decentralized design option 3 architecture

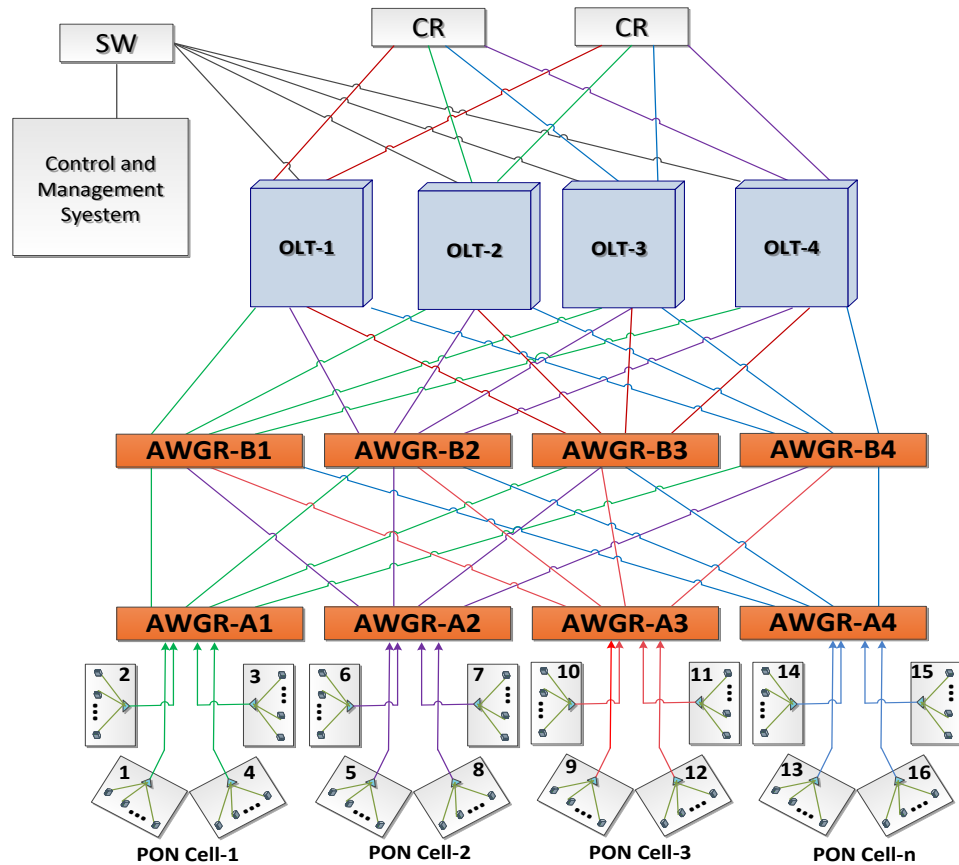


Figure 6.2. Upper-level connectivity for SDN based option-3 architecture

6.3 Inter-cell wavelengths routing and assignment

Earlier in Section 5.2 in Figure 5.6, a PON cell of 6 PON groups connected to an OLT port is demonstrated. In that design 7 different wavelengths were needed, one for intra-rack and six for inter-rack and OLT port connectivity. Now, after introducing a number of OLT switches. All PON groups in each PON cell are required to have the capability to connect with all OLTs. Therefore, a new set of wavelengths should be introduced for uplink and downlink flows between OLTs and PON Cells and tuneable lasers need to be capable of tuning to additional wavelengths equals to the total number of OLTs.

For intra-cell communication, we have shown that the number of wavelengths needed for N racks is $N-1$ wavelengths if intra rack communication is not considered, and N wavelengths are needed if intra rack communication is required. For inter-cell communication through OLT switches, the number of wavelengths needed by each rack is equal to S OLT switches. Therefore; the total number of wavelengths needed for the design depicted in Figure 6.2 is $N+S$ wavelengths.

Assuming that we have 4 OLT switches, 4 PON cells with 4 racks within each PON cell, where each rack needs to connect with all OLT switches and have full rack to rack connectivity including intra rack communication within the cell, the total number of different wavelengths is $4+4=8$ wavelengths. Each tuneable transmitter is capable of tuning to 8 different wavelengths. The same wavelengths can be reused in all PON cells.

6.3.1 MILP model for inter cell wavelength routing and assignment

In this section, a MILP model is developed to optimise wavelength routing and assignment of the upper level connectivity of the SDN enabled option-3 design to facilitate multi-path uplink and downlink communications between OLTs and PON groups within the PON cells.

Note that the model given below optimises the wavelength routing for uplink traffic. Downlink traffic routing is optimised using a similar model.

The parameters and variables used in the model are as follows:

Parameters:

N	Set of nodes (AWGR's ports, PON groups and OLTs)
P	Set of PON groups.
C	Set of PON cells.
P_c	Set of PON groups within PON cell c .
OS	Set of OLT switches
OSP	Set of all OLTs and PON groups
W	Set of wavelengths
A_k	Set of output ports of AWGR k
B_k	Set of input ports of AWGR k
N_m	Set of neighbours of node $m \in N$
(s, d)	Denotes source and destination of a connection, $s \in P$ and $d \in OS$
(m, n)	Denotes end points of a physical link, $m, n \in N$

Variables:

φ_{sd}^{jmn} Defined as $\varphi_{sd}^{jmn} = 1$ if wavelength j on link (m, n) is used for a connection (s, d) , otherwise $\varphi_{sd}^{jmn} = 0$

μ_{sd}^j Defined as $\mu_{sd}^j = 1$ if wavelength j is used for the connection (s, d) , otherwise $\mu_{sd}^j = 0$

Objective:

Maximise:

$$\sum_{s \in P} \sum_{d \in OS} \sum_{j \in W} \mu_{sd}^j \quad (6.1)$$

Equation (6.1) gives the model objective which is to maximise the total number of connections (wavelengths) among PON groups and OLT switches through the intermediate 2-tiers of the AWGRs.

Subject to:

$$\sum_{j \in W} \mu_{sd}^j \leq 1 \quad (6.2)$$

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

Constraint (6.2) ensures that a single wavelength is selected for communication between each PON group and the OLT switch.

$$\sum_{s \in P[i]} \mu_{sd}^j \leq 1 \quad (6.3)$$

$$\forall i \in C, \forall d \in OS, \forall j \in W$$

Constraint (6.3) ensures that PON groups within every PON cell are assigned different wavelengths to connect to every OLT switch.

$$\sum_{d \in OS} \mu_{sd}^j = 1 \quad (6.4)$$

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

Constraint (6.4) ensures that a given source s uses the wavelength j once only among all the links it has to all destinations

$$\sum_{\substack{n \in N_m \\ m \neq n}} \varphi_{sd}^{jmn} - \sum_{\substack{n \in N_m \\ m \neq n}} \varphi_{sd}^{jnm} = \begin{cases} \mu_{sd}^j & m = s \\ -\mu_{sd}^j & m = d \\ 0 & \text{otherwise} \end{cases} \quad (6.5)$$

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

Constraint (6.5) is the wavelength continuity flow conservation constraint. It ensures that the flow going into a node in a certain wavelength leaves the node on the same wavelength for all nodes except the source and destination nodes.

$$\sum_{s \in P} \sum_{d \in OS} \varphi_{sdj}^{nm} \leq 1 \quad (6.6)$$

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

Constraint (6.6) ensures that a wavelength in a certain link is used to connect only one source destination pair.

$$\sum_{s \in P} \sum_{d \in OS} \sum_{n \in N_i} \sum_{j \in W} \varphi_{sdj}^{in} - \sum_{\substack{d \in OS \\ d \neq i}} \sum_{j \in W} \mu_{id}^j \leq 0 \quad (6.7)$$

$$\forall i \in OSP$$

Constraint (6.7) ensures that the PON groups and OLTs do not act as wavelength relays

$$\sum_{s \in P} \sum_{\substack{d \in OS \\ s \neq d}} \sum_{j \in W} \varphi_{sd}^{jmn} \leq 1 \quad (6.8)$$

$$\forall m \in B_k \text{ and } \forall n \in A_k$$

$$\sum_{n \in B_k} \varphi_{sd}^{jmn} \leq 0 \quad (6.9)$$

$$\forall s \in P, \forall d \in OS, \forall m \in A_k \text{ and } j \in W$$

Constraints (6.8) and (6.9) are for the routing within the upper level connectivity AWGRs. Constraint (6.8) ensures that each input port of the each AWGR sends only one wavelength to an output port of the same AWGR. Constraint (6.9) ensures that flows are only directed from input to output ports of the AWGR.

6.3.2 Results and discussion

Tables 6.1 and 6.2 present the wavelength routing and assignment for uplink and downlink flows between the OLTs to the PON groups for the network depicted in Figure 6.2. Connecting 4 PON cells each of 4 PON groups to 4 OLT switches requires each PON group in each PON cell to use 4 different wavelengths in order to reach the four OLT switches through the 2-tier AWGRs CLOS topology. The same wavelengths can be reused for all PON groups as each PON group is connected to a different port of the lower AWGR at tier-1 connected to the designated PON cell. The 4 different wavelengths from each PON group interfaced at the input of the lower AWGR at tier-1 are directed to the 4 different output ports of the same AWGR. Each of the 4 wavelengths then reaches a different AWGR at tier-2 where it will be directed to a different OLT switch. Similarly other PON groups at the 4 different PON cells route their wavelengths obeying the AWGRs property of wavelength routing. As a result, all PON groups in each PON cell can reach all OLT switches using the four different wavelengths obeying the directionality and wavelength routing rules of all the AWGRs. The total number of wavelengths needed for uplink is 4 as the total number of OLT switches to be connected to is 4. The model described in Section

6.3.1 is used to obtain the results for the uplink connections. The same model is used to obtain the results for the downlink connections by only changing the source and destination e.g. OLTs will be the source nodes and the PON groups are the destination nodes. For verification of obtained results shown in Tables 6.1 and 6.2; rows for each PON group to OLT, shows that unique wavelength is assigned for communication. Similarly, columns in each PON cell shows unique wavelength is assigned between PON groups within the cell to the OLT. The results shown may appear intuitive, but MILP allows dealing with larger designs and also allows for different constraints to be used.

Detailed routing paths for connections from PON cells to OLT switches as obtained from MILP are shown in Tables 6.3, 6.4, 6.5 and 6.7. Figure 6.3 show the numbering of the inputs and outputs ports for the AWGRs used in the upper 2-tier AWGRs layers. The numbering sequence of the ports is followed in the results of detailed routing tables.



Figure 6.3. 4x4 AWGR input/output numbering diagram

Table 6.1. MILP obtained wavelength assignments for uplink connections from PONs to OLTs

		OLT-1	OLT-2	OLT-3	OLT-4
PON Cell-1	PON-1	λ 4	λ 3	λ 2	λ 1
	PON-2	λ 1	λ 2	λ 3	λ 4
	PON-3	λ 3	λ 4	λ 1	λ 2
	PON-4	λ 2	λ 1	λ 4	λ 3

PON Cell-2	PON-5	λ 1	λ 3	λ 2	λ 4
	PON-6	λ 3	λ 1	λ 4	λ 2
	PON-7	λ 4	λ 2	λ 3	λ 1
	PON-8	λ 2	λ 4	λ 1	λ 3
PON Cell-3	PON-9	λ 4	λ 1	λ 2	λ 3
	PON-10	λ 2	λ 3	λ 4	λ 1
	PON-11	λ 1	λ 4	λ 3	λ 2
	PON-12	λ 3	λ 2	λ 1	λ 4
PON Cell-4	PON-13	λ 4	λ 2	λ 1	λ 3
	PON-14	λ 3	λ 1	λ 2	λ 4
	PON-15	λ 1	λ 3	λ 4	λ 2
	PON-16	λ 2	λ 4	λ 3	λ 1

Table 6.2. MILP obtained wavelength assignment for down link connections from OLTs to PONs

		OLT-1	OLT-2	OLT-3	OLT-4
PON Cell-1	PON-1	λ 3	λ 1	λ 4	λ 2
	PON-2	λ 2	λ 4	λ 1	λ 3
	PON-3	λ 1	λ 3	λ 2	λ 4
	PON-4	λ 4	λ 2	λ 3	λ 1
PON Cell-2	PON-5	λ 2	λ 1	λ 4	λ 3
	PON-6	λ 1	λ 2	λ 3	λ 4
	PON-7	λ 3	λ 4	λ 1	λ 2
	PON-8	λ 4	λ 3	λ 2	λ 1
PON Cell-3	PON-9	λ 3	λ 2	λ 4	λ 1
	PON-10	λ 4	λ 1	λ 3	λ 2
	PON-11	λ 2	λ 3	λ 1	λ 4
	PON-12	λ 1	λ 4	λ 2	λ 3
PON Cell-4	PON-13	λ 3	λ 4	λ 2	λ 1
	PON-14	λ 1	λ 2	λ 4	λ 3
	PON-14	λ 2	λ 1	λ 3	λ 4
	PON-16	λ 4	λ 3	λ 1	λ 2

Table 6.3. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-1 to OLT switches

		AWGR(A)	AWGR(B)	OLT	λ
PON Cell-1	PON-1	AWGR-A1 I(1)-O(8)	AWGR-B4 I(1)-O(5)	OLT-1 I(4)	λ 4
	PON-1	AWGR-A1 I(1)-O(6)	AWGR-B2 I(1)-O(6)	OLT-2 I(2)	λ 3
	PON-1	AWGR-A1 I(1)-O(5)	AWGR-B1 I(1)-O(7)	OLT-3 I(1)	λ 2
	PON-1	AWGR-A1 I(1)-O(7)	AWGR-B3 I(1)-O(8)	OLT-4 I(3)	λ 1
	PON-2	AWGR-A1 I(2)-O(5)	AWGR-B1 I(1)-O(5)	OLT-1 I(1)	λ 1
	PON-2	AWGR-A1 I(2)-O(7)	AWGR-B3 I(1)-O(6)	OLT-2 I(3)	λ 2
	PON-2	AWGR-A1 I(2)-O(8)	AWGR-B4 I(1)-O(7)	OLT-3 I(4)	λ 3
	PON-2	AWGR-A1 I(2)-O(6)	AWGR-B2 I(1)-O(8)	OLT-4 I(2)	λ 4
	PON-3	AWGR-A1 I(3)-O(7)	AWGR-B3 I(1)-O(5)	OLT-1 I(3)	λ 3
	PON-3	AWGR-A1 I(3)-O(5)	AWGR-B1 I(1)-O(6)	OLT-2 I(1)	λ 4
	PON-3	AWGR-A1 I(3)-O(6)	AWGR-B2 I(1)-O(7)	OLT-3 I(2)	λ 1
	PON-3	AWGR-A1 I(3)-O(8)	AWGR-B4 I(1)-O(8)	OLT-4 I(4)	λ 2
	PON-4	AWGR-A1 I(4)-O(6)	AWGR-B2 I(1)-O(5)	OLT-1 I(2)	λ 2
	PON-4	AWGR-A1 I(4)-O(8)	AWGR-B4 I(1)-O(6)	OLT-2 I(4)	λ 1
	PON-4	AWGR-A1 I(4)-O(7)	AWGR-B3 I(1)-O(7)	OLT-3 I(3)	λ 4
	PON-4	AWGR-A1 I(4)-O(5)	AWGR-B1 I(1)-O(8)	OLT-4 I(1)	λ 3

Table 6.4. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-2 to OLT switches

		AWGR(A)	AWGR(B)	OLT	λ
PON Cell-2	PON-5	AWGR-A2 I(1)-O(8)	AWGR-B4 I(2)-O(5)	OLT-1 I(4)	$\lambda 1$
	PON-5	AWGR-A2 I(1)-O(7)	AWGR-B3 I(2)-O(6)	OLT-2 I(3)	$\lambda 3$
	PON-5	AWGR-A2 I(1)-O(6)	AWGR-B2 I(2)-O(7)	OLT-3 I(2)	$\lambda 2$
	PON-5	AWGR-A2 I(1)-O(5)	AWGR-B1 I(2)-O(8)	OLT-4 I(1)	$\lambda 4$
	PON-6	AWGR-A2 I(2)-O(6)	AWGR-B2 I(2)-O(5)	OLT-1 I(2)	$\lambda 3$
	PON-6	AWGR-A2 I(2)-O(5)	AWGR-B1 I(2)-O(6)	OLT-2 I(1)	$\lambda 1$
	PON-6	AWGR-A2 I(2)-O(8)	AWGR-B4 I(2)-O(7)	OLT-3 I(4)	$\lambda 4$
	PON-6	AWGR-A2 I(2)-O(7)	AWGR-B3 I(2)-O(8)	OLT-4 I(3)	$\lambda 2$
	PON-7	AWGR-A2 I(3)-O(7)	AWGR-B3 I(2)-O(5)	OLT-1 I(3)	$\lambda 4$
	PON-7	AWGR-A2 I(3)-O(8)	AWGR-B4 I(2)-O(6)	OLT-2 I(4)	$\lambda 2$
	PON-7	AWGR-A2 I(3)-O(5)	AWGR-B1 I(2)-O(7)	OLT-3 I(1)	$\lambda 3$
	PON-7	AWGR-A2 I(3)-O(6)	AWGR-B2 I(2)-O(8)	OLT-4 I(2)	$\lambda 1$
	PON-8	AWGR-A2 I(4)-O(5)	AWGR-B1 I(2)-O(5)	OLT-1 I(1)	$\lambda 2$
	PON-8	AWGR-A2 I(4)-O(6)	AWGR-B2 I(2)-O(6)	OLT-2 I(2)	$\lambda 4$
	PON-8	AWGR-A2 I(4)-O(7)	AWGR-B3 I(2)-O(7)	OLT-3 I(3)	$\lambda 1$
	PON-8	AWGR-A2 I(4)-O(8)	AWGR-B4 I(2)-O(8)	OLT-4 I(4)	$\lambda 3$

Table 6.5. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-3 to OLT switches.

		AWGR(A)	AWGR(B)	OLT	λ
PON Cell-3	PON-9	AWGR-A3 I(1)-O(5)	AWGR-B1 I(3)-O(5)	OLT-1 I(1)	$\lambda 4$
	PON-9	AWGR-A3 I(1)-O(7)	AWGR-B3 I(3)-O(6)	OLT-2 I(3)	$\lambda 1$
	PON-9	AWGR-A3 I(1)- O(8)	AWGR-B4 I(3)-O(7)	OLT-3 I(4)	$\lambda 2$
	PON-9	AWGR-A3 I(1)-O(6)	AWGR-B2 I(3)-O(8)	OLT-4 I(2)	$\lambda 3$
	PON-10	AWGR-A3 I(2)-O(7)	AWGR-B3 I(3)-O(5)	OLT-1 I(3)	$\lambda 2$
	PON-10	AWGR-A3 I(2)-O(5)	AWGR-B1 I(3)-O(6)	OLT-2 I(1)	$\lambda 3$
	PON-10	AWGR-A3 I(2)-O(6)	AWGR-B2 I(3)-O(7)	OLT-3 I(2)	$\lambda 4$
	PON-10	AWGR-A3 I(2)-O(8)	AWGR-B4 I(3)-O(8)	OLT-4 I(4)	$\lambda 1$
	PON-11	AWGR-A3 I(3)-O(6)	AWGR-B2 I(3)-O(5)	OLT-1 I(2)	$\lambda 1$
	PON-11	AWGR-A3 I(3)-O(8)	AWGR-B4 I(3)-O(6)	OLT-2 I(4)	$\lambda 4$
	PON-11	AWGR-A3 I(3)-O(7)	AWGR-B3 I(3)-O(7)	OLT-3 I(3)	$\lambda 3$
	PON-11	AWGR-A3 I(3)-O(5)	AWGR-B1 I(3)-O(8)	OLT-4 I(1)	$\lambda 2$
	PON-12	AWGR-A3 I(4)-O(8)	AWGR-B4 I(3)-O(5)	OLT-1 I(4)	$\lambda 3$
	PON-12	AWGR-A3 I(4)-O(6)	AWGR-B2 I(3)-O(6)	OLT-2 I(2)	$\lambda 2$
	PON-12	AWGR-A3 I(4)-O(5)	AWGR-B1 I(3)-O(7)	OLT-3 I(1)	$\lambda 1$
	PON-12	AWGR-A3 I(4)-O(7)	AWGR-B3 I(3)-O(8)	OLT-4 I(3)	$\lambda 4$

Table 6.6. MILP obtained detailed routing paths with wavelength assignments for uplink connections between PON groups in PON cell-4 to OLT switches

		AWGR(A)	AWGR(B)	OLT	λ
PON Cell-4	PON-13	AWGR-A4 I(1)-O(6)	AWGR-B2 I(4)-O(5)	OLT-1 I(2)	λ 4
	PON-13	AWGR-A4 I(1)-O(5)	AWGR-B1 I(4)-O(6)	OLT-2 I(1)	λ 2
	PON-13	AWGR-A4 I(1)-O(8)	AWGR-B4 I(4)-O(7)	OLT-3 I(4)	λ 1
	PON-13	AWGR-A4 I(1)-O(7)	AWGR-B3 I(4)-O(8)	OLT-4 I(3)	λ 3
	PON-14	AWGR-A4 I(2)-O(5)	AWGR-B1 I(4)-O(5)	OLT-1 I(1)	λ 3
	PON-14	AWGR-A4 I(2)-O(6)	AWGR-B2 I(4)-O(6)	OLT-2 I(2)	λ 1
	PON-14	AWGR-A4 I(2)-O(7)	AWGR-B3 I(4)-O(7)	OLT-3 I(3)	λ 2
	PON-14	AWGR-A4 I(2)-O(8)	AWGR-B4 I(4)-O(8)	OLT-4 I(4)	λ 4
	PON-15	AWGR-A4 I(3)-O(7)	AWGR-B3 I(4)-O(5)	OLT-1 I(3)	λ 1
	PON-15	AWGR-A4 I(3)-O(8)	AWGR-B4 I(4)-O(6)	OLT-2 I(4)	λ 3
	PON-15	AWGR-A4 I(3)-O(5)	AWGR-B1 I(4)-O(7)	OLT-3 I(1)	λ 4
	PON-15	AWGR-A4 I(3)-O(6)	AWGR-B2 I(4)-O(8)	OLT-4 I(2)	λ 2
	PON-16	AWGR-A4 I(4)-O(8)	AWGR-B4 I(4)-O(5)	OLT-1 I(4)	λ 2
	PON-16	AWGR-A4 I(4)-O(7)	AWGR-B3 I(4)-O(6)	OLT-2 I(3)	λ 4
	PON-16	AWGR-A4 I(4)-O(6)	AWGR-B2 I(4)-O(7)	OLT-3 I(2)	λ 3
	PON-16	AWGR-A4 I(4)-O(5)	AWGR-B1 I(4)-O(8)	OLT-4 I(1)	λ 1

6.4 Energy efficient software defined AWGR-based PON data centre network

Figure 6.4 shows an example of a simplified schematic describing the role of the SDN controller in assigning paths and resources for PON groups with demands. Assume the SDN controller has received requests to assign a route and resources for demands between the following pair of servers: (A, D), (B, F), and (C, E). Assume server A is located in PON Cell 1 in PON Group 1 and server D is located in PON Cell 2 in PON Group 5. In Figure 6.4(a) we have shown that servers are randomly assigned to OLTs and traffic is traversing multiple OLTs and CR.

In Figure 6.4(b) we show that demanding servers can be grouped to a common OLT switch. Based on routing details and wavelengths assignment obtained from MILP model and presented in Tables 6.1 and 6.2, the scheduler can assign wavelength 2 for server A to join OLT-3 and wavelength 4 for OLT 3 for downlink transmission to server D. Therefore; servers A and D are grouped to join OLT-3 to avoid routing the demand through the CR. Demands for (B, F) and (C, E) can be managed similarly. For further efficient use of the resources, Figure 6.4(b) shows that demanding servers can be grouped to join a common OLT switch so unused OLTs can be switched off.

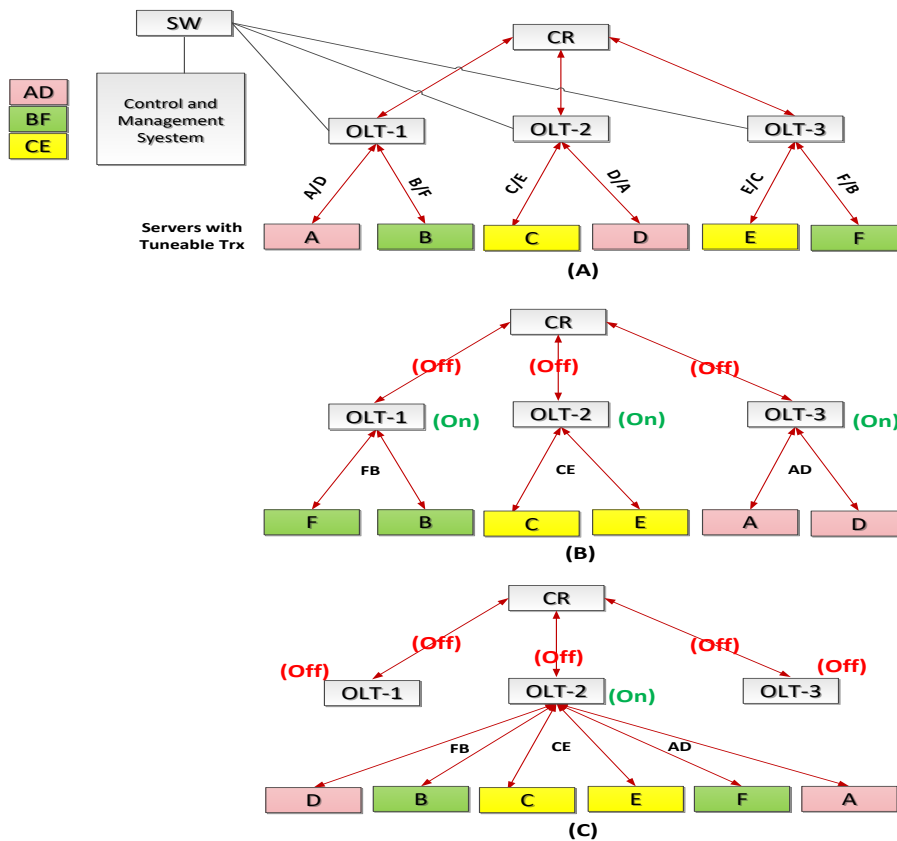


Figure 6.4. Energy-Efficient Bandwidth Allocation through reconfiguration and grouping

In this section a MILP model is developed for energy-efficient bandwidth allocation in the upper level connectivity to minimise the upper level connectivity power consumption. We evaluate the power savings achieved by the SDN enabled architecture compared to the decentralised one.

6.4.1 MILP model

Before introducing the model, the parameters and variables used in the model are listed:

Parameters:

OS	Set of OLT switches
CR	Set of core routers
H	Set of AM OLT ports, uplink OLT ports, CR ports and PON Groups
AM	Set of Access Module OLT ports
OU	Set of uplink OLT ports
CRP	Set of CR ports
P	Set of PON groups
N_i	Set of nodes with connectivity with node $i, i \in N$
(s, d)	Denotes source and destination PON groups of a traffic demand
L^{sd}	Traffic demand between PON group pair (s, d) in b/s
Cu	Uplink capacity between a PON group and an OLT in b/s
Cd	Downlink capacity between an OLT and a PON group in b/s
OS_{up}	Capacity of uplink between OLT switch and core routers in b/s
α	weight coefficient (unit less)
β	weight coefficient with a unit of $1/W$
OS_{ch}	Power consumption of an OLT chassis
OS_{sc}	Power consumption of a switch controller card of an OLT switch
OS_{pr}	Power consumption of an OLT Access Module port.
OS_{pc_up}	Power consumption of an OLT uplink port connecting with CRs
CR_{pr}	Power consumption of CR port
CRU	Core router port capacity in b/s
NPC	Total number of ports in core router
M	Large Multiplication number

Variables:

L_{ij}^{sd} Portion of traffic demand (s, d) traversing link (i, j)

$Cdown^{ij}$ Downlink traffic flow from node i to node j , $i, j \in N$

Cup^{ij} Uplink traffic flow from node i to node j , $i, j \in N$

ϕ_j Defined as $\phi_j = 1$ to indicate that OLT switch j is on, otherwise $\phi_j = 0$

bL^{sd} Defined as $bL^{sd} = 1$ if demand L^{sd} is not served, otherwise $bL^{sd} = 0$

The power consumption of the upper level connectivity consists of:

1) The power consumption of OLT switches

$$\begin{aligned} \sum_{i \in OS} \phi_i (OS_{ch} + OS_{sc}) + \sum_{i \in AM} \sum_{j \in P} \left(\frac{Cdown^{ij} + Cup^{ji}}{Cd + Cu} \right) OS_{pr} \\ + \sum_{i \in OU} \sum_{j \in CRP} \left(\frac{Cdown^{ji} + Cup^{ij}}{OS_{up}} \right) OS_{pc_up} \end{aligned} \quad (6.10)$$

Note that the second and third terms in (6.10) assume rate adaptation in AM and uplink ports in OLT switches.

2) The power consumption of CRs

$$\sum_{i \in CRP} \sum_{j \in OU} \left(\frac{Cdown^{ij} + Cup^{ji}}{CRU} \right) CR_{pr} \quad (6.11)$$

Objective:

Minimise:

$$\begin{aligned}
& \beta \left[\sum_{i \in OS} \phi_i (OS_{ch} + OS_{sc}) + \sum_{i \in AM} \sum_{j \in P} \left(\frac{Cdown^{ij} + Cup^{ji}}{Cd + Cu} \right) OS_{pr} \right. \\
& \quad + \sum_{i \in OU} \sum_{j \in CRP} \left(\frac{Cdown^{ij} + Cup^{ji}}{OS_{up}} \right) OS_{pc_up} \\
& \quad \left. + \sum_{i \in CRP} \sum_{j \in OU} \left(\frac{Cdown^{ij} + Cup^{ji}}{CRU} \right) CR_{pr} \right] + \alpha \sum_{s \in P} \sum_{d \in P} bL^{sd}
\end{aligned} \tag{6.12}$$

Equation (6.12) gives the model objective which is to minimise the total power consumption of the network and the blocking.

Subject to:

Flow Conservation constraint

$$\sum_{\substack{j \in N_i \\ i \neq j}} L_{ij}^{sd} - \sum_{\substack{j \in N_i \\ i \neq j}} L_{ji}^{sd} = \begin{cases} L^{sd} (1 - bL^{sd}) & i = s \\ -L^{sd} (1 - bL^{sd}) & i = d \\ 0 & otherwise \end{cases} \tag{6.13}$$

$$\forall s, d \in P : s \neq d \text{ and } \forall i \in H$$

Constraint (6.13) is the flow conservation constraint. It ensures that the total traffic going into a node is equal to the total traffic leaving it for all nodes except the source and destination of a demand.

$$Cup^{ij} = \sum_{s \in P} \sum_{\substack{d \in P \\ s \neq d}} L_{ij}^{sd} \tag{6.14}$$

$$\forall i \in P \text{ and } \forall j \in AM$$

Equation (6.14) calculates total traffic for uplink wavelengths between PON group i and OLT switches j .

$$C_{down}^{ij} = \sum_{s \in P} \sum_{\substack{d \in P \\ s \neq d}} L_{ij}^{sd} \quad (6.15)$$

$$\forall i \in AM \text{ and } \forall j \in P$$

Equation (6.15) calculates total traffic for downlink wavelengths between OLT switch i and PON group j .

Link capacity constraints

$$C_{up}^{ij} \leq C_u$$

$$\forall i \in P \text{ and } \forall j \in AM \quad (6.16)$$

Constraint (6.16) ensures that the uplink traffic traversing from PON groups to OLTs does not exceed its capacity.

$$C_{down}^{ij} \leq C_d$$

$$\forall i \in AM \text{ and } \forall j \in P \quad (6.17)$$

Constraint (6.17) ensures that the downlink traffic from an OLT AM port to a PON group does not exceed the downlink capacity.

$$\sum_{i \in OU} \sum_{j \in CR} C_{up}^{ij} \leq CRU \ NPC$$

$$(6.18)$$

Constraint (6.18) ensures that the uplink traffic from an OLT uplink port to a CRs does not exceed the ports capacity.

Constraints to switch off idle OLTs

$$\sum_{s \in P} \sum_{\substack{d \in P \\ :s \neq d}} \sum_{n \in N_i} L_{ni}^{sd} \geq \phi_i \quad (6.19)$$

$$\forall i \in OS$$

$$\sum_{s \in P} \sum_{\substack{d \in P \\ :s \neq d}} \sum_{n \in N_i} L_{ni}^{sd} \leq M \phi_i \quad (6.20)$$

$$\forall i \in OS$$

Constraints (6.20) and (6.21) ensure that OLT switches not used are switched off.

6.4.2 Results and discussion

The decentralised network is depicted in Figure 6.1 where each PON group connects to a single OLT switch and the second with a centralised control and management SDN is depicted in Figure 6.2.

Table 6.7 shows the input parameters used in the model. The MILP results as presented in Figure 6.5 show that with the SDN enabled architecture, the power consumption of the upper level connectivity can be decreased by 65% to 90% compared to the decentralised architecture for average data rate of 250 Mb/s and to 2500 Mb/s, respectively. This decrease is due to the efficient utilisation of uplink and downlink resources between OLTs and PON groups in the SDN enabled architecture which reduces the number of OLT switches and access ports needed. The introduction of the 2-tiers of AWGRS layers with full interconnection also allows SDN to assign resources to servers to join any of the OLTs which

further increases the energy saving as routing through core routers is avoided.

On the other hand in the conventional decentralised design, demand pairs for servers located in two different PON cells cannot be grouped to a single OLT and have to be forwarded to core routers which increases the overall power consumption as more core router ports and OLT uplink ports are required. As the flow rate increases, the flow of traffic between core routers and OLTs increases in the decentralized design. Hence, more high power consuming core router and OLT uplink ports will be needed to satisfy demands. While in the SDN enabled technology no traffic is routed between OLTs and core routers and through the introduced two layers of the AWGRs and efficient grouping mechanism employed, all traffic is directly routed between PON groups and the OLTs through the up/down links of the AM ports. This explain the increase of power saving as the rate of flow increases between demanding servers in different PON groups.

This increase in the traffic flow between OLTs and core routers, results in resources oversubscription especially in the high data rates case. Figure 6.5 shows that with the conventional decentralised architecture, the blocking probability increases from 13% to 45% as data rates increase from 1Gb/s to 2.5Gb/s whereas in the SDN enabled network zero blocking is achieved for the evaluated data rates. The conventional design at rate of 750Mb/s fully utilises all resources between the OLT uplink and core routers ports, hence the power consumption on this rate and onward rates appear flat as shown in Figure 6.5 and number of blocked requests increases as rate increases and more demands cannot be satisfied as in Figure 6.6.

Table 6.7. Input parameters used in the model

Power consumption of Cisco ME4600 OLT Chassis	60 W [90]
Power consumption of Cisco ME4600 XCO Switch Fabric Card	180 W [90]
Power consumption of Cisco ME4600 AMX Access Card with 16 x GPON	90 W [90]
Cisco ME4600 UMX Uplink Card with 4 x 10GE	40 W [90]
Total GPON ports per OLT	256 [90]
GPON port uplink capacity	10Gb/s [90]
GPON port downlink capacity	10Gb/s [90]
Core router port capacity	10Gb/s
Power consumption of core router port	300 W [38]
Average demand rate following uniform distribution	250-2500 Mb/s

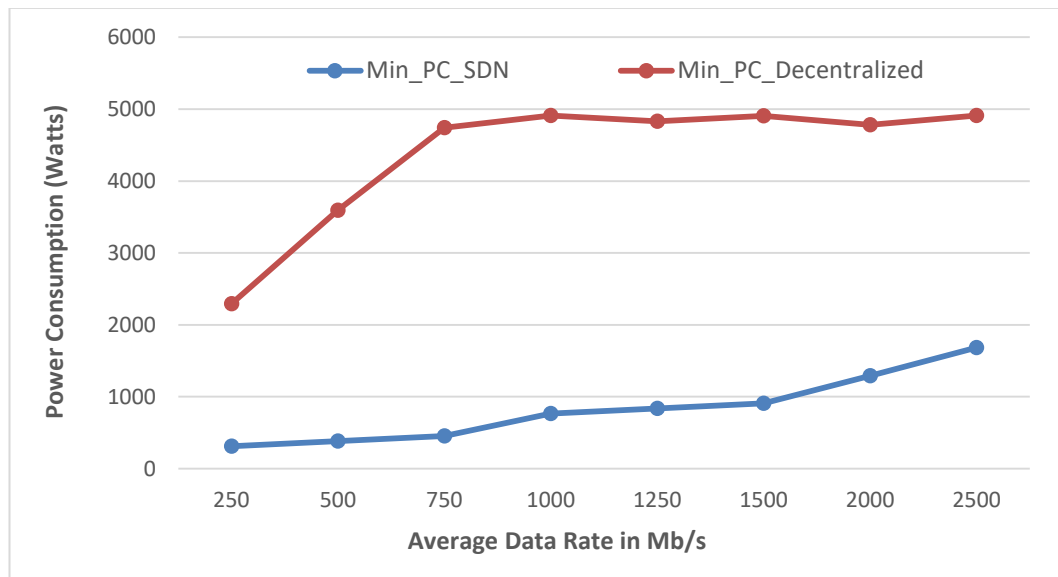


Figure 6.5. Power consumption evaluation for SDN over decentralized designs

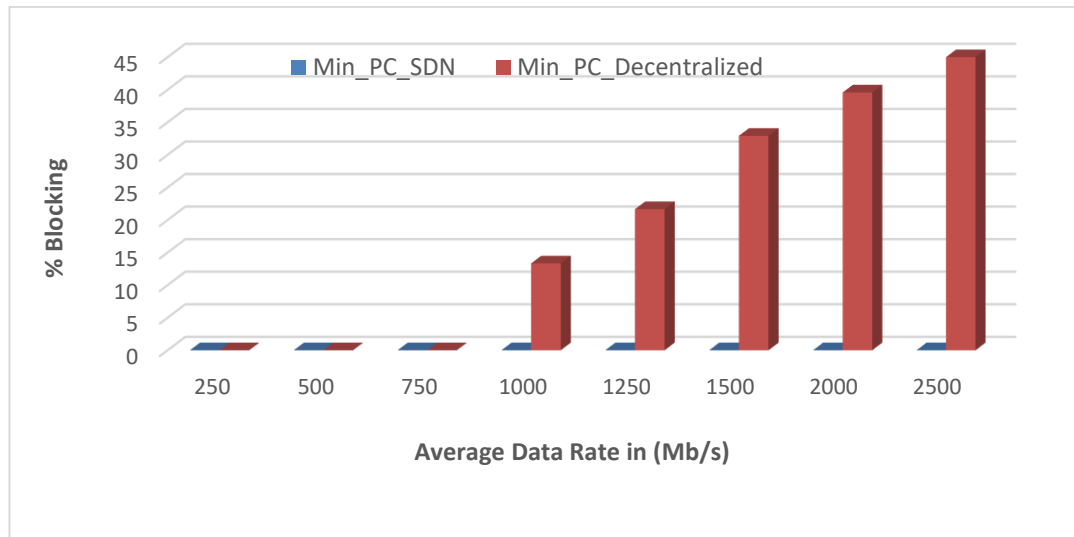


Figure 6.6. Blocking percentages of SDN minimised energy model against the decentralized design

6.5 Summary

This chapter has proposed an SDN based AWGR PON data centre interconnection design to provide energy efficient and highly elastic networking infrastructure to sustain the applications and services hosted by modern data centres. The design has shown that SDN can facilitate dynamic networks with wavelength configurability for efficient utilisation and allocation of bandwidth resources. The proposed inter-cell interconnection fabric through the 2-tiers of AWGRs improved the design by reducing oversubscription ratios and provisioning multi-path routing. A benchmarking study between the proposed SDN architecture against the decentralised conventional design shows that with the SDN enabled architecture, the power consumption can be decreased by up to 90% for typical average data rates while maintaining zero blocking.

7 Resource provisioning for AWGR PON based cloud data centre

7.1 Introduction

Consolidation along with virtualisation can result in substantial savings in power consumption if efficient algorithms for resources provisioning are designed. As most of current data centres' resources are underutilised, studies have reported that the average utilisation of data centre computing capacity is around 30% [38]. This is mainly a result of non-energy efficient resource provisioning based on random or round robin algorithms to map the VMs with servers. By improving resource utilisation through efficient mapping between VMs and servers, the number of active servers along with cooling power draw will be significantly reduced.

This chapter tackles the problem of resource provisioning optimisation for cloud applications in AWGR based PON data centre architecture. A MILP model is developed to minimise the power consumption and delay for different cloud applications by optimising resource provisioning. A trade-off between power consumption and delay with respect to the applications that can co-exist in data centres is considered. A novel greedy algorithm along with the MILP mathematical modelling for resource provisioning optimisation

are developed considering the physical constraints of servers (CPU and memory) and virtual machines communication traffic. The developed algorithm is compared with random and Best Fit Decreasing Bin-Packing conventional algorithms.

7.2 MILP model for Energy Aware Resource Provisioning in PON data centre

In this section, a MILP optimisation model is introduced to minimise the power consumption of the AWGR based PON data center architecture through efficient resource provisioning. We consider the power consumption of the servers hosting the VMs and the ONUs connected to them.

Parameters and variables used in the model are as follows:

Parameters:

S	Set of servers
P	Set of PON groups
G_p	Set of servers connected to PON group P
V	Set of VM requests
PO	Idle power consumption of a server
PM	Maximum power consumption of a server
C_j	CPU capacity of server j
M_j	Memory capacity (RAM) of server j
K	Number of servers allowed to serve a VM request
ρ_i	VM request i requirement of CPU for processing

m_i	VM request i requirements of RAM
A^{if}	Traffic demand between VMs i and f
PU	ONU power consumption
CU	ONU data rate
CW	Wavelength capacity for TDM
M	Large positive number

Variables:

δ_j	Defined as $\delta_j = 1$ if server j is activated, otherwise $\delta_j = 0$
ρ_j^i	processing resources of server j in GHz assigned for request i
ω_j^i	Defined as $\omega_j^i = 1$ if request i is served by server j , otherwise $\omega_j^i = 0$
β^{if}	Defined as $\beta^{if} = 1$ if VMs i and f are assigned to the same server, otherwise $\beta^{if} = 0$
U_j	Uplink traffic for server j
v_j	Number of VMs placed in server j
ε_{ij}^{sd}	ε_{ij}^{sd} is the AND of ω_j^i and ω_k^f , i.e. $\varepsilon_{ij}^{sd} = \omega_j^i + \omega_k^f$
θ_j^{if}	θ_j^{if} is the AND of ω_j^i and ω_j^f , i.e. $\theta_j^{if} = \omega_j^i + \omega_j^f$
T^{jk}	Traffic between server j and server k resulting from VMs placement

The power consumption of the PON cell is composed of:

The power consumption of a physical machine (server) j is given as:

$$\sum_{j \in \mathcal{S}} \left(PO \delta_j + (PM - PO) \sum_{i \in \mathcal{V}} \frac{\rho_j^i}{C_j} \right) \quad (7.1)$$

Power consumption of ONUs connected to servers

$$\frac{PU}{CU} \sum_{j \in S} U_j \quad (7.2)$$

Objective:

Minimise:

$$\sum_{j \in S} \left(PO \delta_j + (PM - PO) \sum_{i \in V} \frac{\rho_j^i}{C_j} \right) + \frac{PU}{CU} \sum_{j \in S} U_j \quad (7.3)$$

Equation (7.3) gives the model objective which is to minimise the total servers' and network power consumption. This is achieved through optimising the servers selected to provision VMs.

Subject to:

$$U_j = \sum_{i \in V} \sum_{f \in V} A^{if} (1 - \beta^{if}) \quad (7.4)$$

$$\forall j \in S$$

Constraint (7.4) calculates the total uplink traffic from server j , where β^{if} is given by constraint (7.5):

$$\beta^{if} = \sum_{j \in S} \omega_j^i \omega_j^f \quad (7.5)$$

$$\forall i, f \in V$$

Constraint (7.5) contains multiplication of two binary variables which makes the model nonlinear; we replace it by the following three constraints to maintain the linearity of the model (constraints (7.6)-(7.9)).

$$\beta^{if} = \sum_{j \in S} \theta_j^{if} \quad (7.6)$$

$$\forall i, f \in V$$

$$\theta_j^{if} \leq \omega_j^i \quad (7.7)$$

$$\forall i, f \in V \text{ and } j \in S$$

$$\theta_j^{if} \leq \omega_j^f \quad (7.8)$$

$$\forall i, f \in V \text{ and } j \in S$$

$$\theta_j^{if} \geq \omega_j^i + \omega_j^f - 1 \quad (7.9)$$

$$\forall i, f \in V \text{ and } j \in S$$

Note that equations (7.7) and (7.8) ensures that θ_j^{if} can only be equal to 1 if both ω_j^i and ω_j^f are 1. However, equation (7.7) and (7.8) can be satisfied with $\theta_j^{if} = 0$ while are ω_j^i and ω_j^f are both equal 1. This last issue is resolved through (7.9)

$$\sum_{i \in V} m_i \omega_j^i \leq M_j \quad (7.10)$$

$$\forall j \in S$$

$$\sum_{i \in V} \rho_i \omega_j^i \leq C_j \quad (7.11)$$

$$\forall j \in S$$

$$U_j \leq CU \quad (7.12)$$

$$\forall j \in S$$

Constraints (7.10)-(7.12) are the memory, processing and ONU link capacity constraints, respectively

$$\sum_{j \in S} \omega_j^i = K \quad (7.13)$$

$$\forall i \in V$$

Constraint (7.13) limits the number of servers that can be used to serve a VM.

$$v_j = \sum_{i \in V} \omega_j^i \quad (7.14)$$

$$\forall j \in S$$

Constraint (7.14) calculates the total number of VMs served by server j .

$$\delta_j \leq v_j \quad (7.15)$$

$$\forall j \in S$$

$$M \delta_j \geq v_j \quad (7.16)$$

$$\forall j \in S$$

Constraints (7.15) and (7.16) are used to relate the binary variable δ_j , to the non-binary variable v_j where M is a large enough number, the value used in the model is 1000.

$$T^{jk} = \sum_{i \in V} \sum_{f \in V} A^{if} (\omega_j^i \omega_k^f) \quad (7.17)$$

$$\forall j, k \in S$$

Constraint (7.17) is used to calculate the traffic between server pairs after assigning the VMs to the servers. Multiplication of two binary variables $(\omega_j^i \omega_k^f)$ makes the model nonlinear. It is replaced by constraints (7.18) – (7.21) to maintain the linearity of the model.

$$T^{jk} = \sum_{i \in V} \sum_{f \in V} A^{if} \varepsilon_{jk}^{if} \quad (7.18)$$

$$\forall j, k \in S$$

$$(7.19)$$

$$\varepsilon_{jk}^{if} \leq \omega_j^i$$

$$\forall j, k \in S \ \& \ \forall i, f \in V$$

$$\varepsilon_{jk}^{if} \leq \omega_k^f \quad (7.20)$$

$$\forall j, k \in S \ \& \ \forall i, f \in V$$

$$\varepsilon_{jk}^{if} \geq \omega_j^i + \omega_k^f - 1 \quad (7.21)$$

$$\forall j, k \in S \ \& \ \forall i, f \in V$$

$$\sum_{j \in P_p} \sum_{k \in P_o} T^{jk} \leq CW \quad (7.22)$$

$$\forall p, o \in P$$

Constraint (7.22) ensures that servers in each PON group does not exceed the shared wavelength capacity while communicating to other servers in different PON groups.

7.3 Results and discussions

In this section, an evaluation of the power consumption of resource provisioning in the AWGR based PON data center architecture using the model developed in Section (7.2) is presented. The AWGR based PON architecture depicted in Figure 5.4 is modelled. Different number of VM requests (20, 40, and 60) are examined. The CPU, memory, and communication traffic requirements of VMs are randomly distributed within the values in Table 7.1. Traffic between VMs pairs is generated so each VM communicates with a number of VMs randomly distributed between 1-3. This traffic is randomly (uniform distribution) assigned values between 40Mb/s and 200Mb/s as given in Table 7.1. Table 7.1 also presents the other input parameters used for the model.

Table 7.1. Input data for the model

Link capacity (CW)	10Gbps
Power consumption for idle servers PO [38]	201 W
Maximum power consumption for servers PM [38]	301 W
Clients processing requirements in M CPU cycles (ρ_i)	500-2000
Clients memory requirements in MB (M_i)	500-2000
Server's processing capacity (C_j) in GHz	2.5
Server's memory (RAM) (M_j) in GB	8

ONU power consumption (PU) in	2.5W
VMs traffic (A^{if})	40-200 Mb/s

We compare the power consumption of the model developed in Section 7.2 with the objective of minimising both the power consumption of physical machines (PM), the power consumption of the network, and the minimisation of both the physical machines power consumption and the network power consumption (Min Both). The results compare the minimisations of both, the minimisation of physical machines power consumption only (Min PMs PC), the minimisation of network power consumption only (Min Net PC) described in equations 7.1, 7.2, and 7.3, respectively.

Figures 7.4 presents the total power consumption for the three examined models. The minimisation of the servers' power consumption model allocates CPU and RAM resources requested by clients to the minimum possible number of servers by means of consolidation.

Minimisation of the servers' power consumption objective does not look after the communication demands among the VMs to decide on the location of placement for each VM.

While minimisation of network power objective aims at minimisation of traffic flow on communication links to reduce ONUs transmit output power and doesn't take into accounts the number of servers used to provision resources to the clients. The multi objective of minimisation of both considers the minimisation of the total power which results from servers and network.

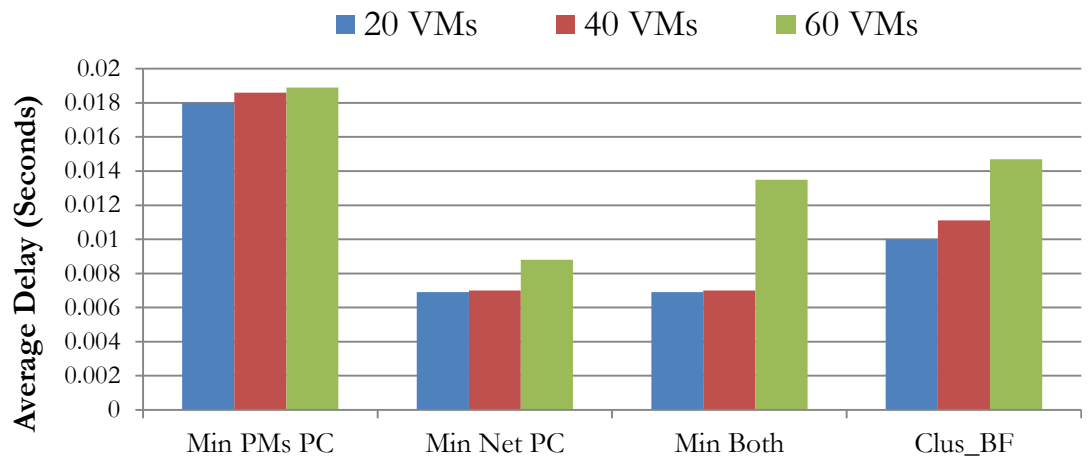


Figure 7.1. The average delay provisioning different sets of VMs; under the three models and the developed algorithm Clus_BF

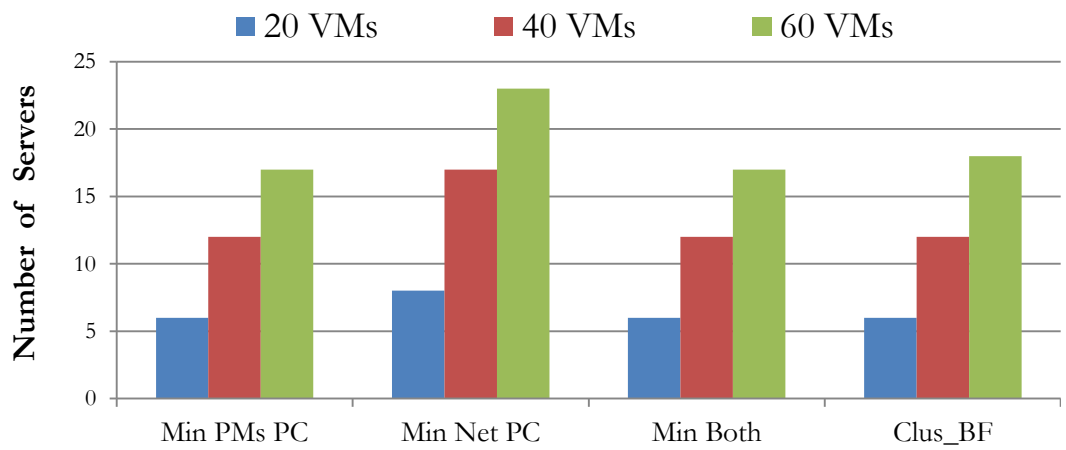


Figure 7.2. Total number of used servers examining three sets of VMs; 20, 40, and 60 for the three objective functions and the developed algorithm Clus_BF

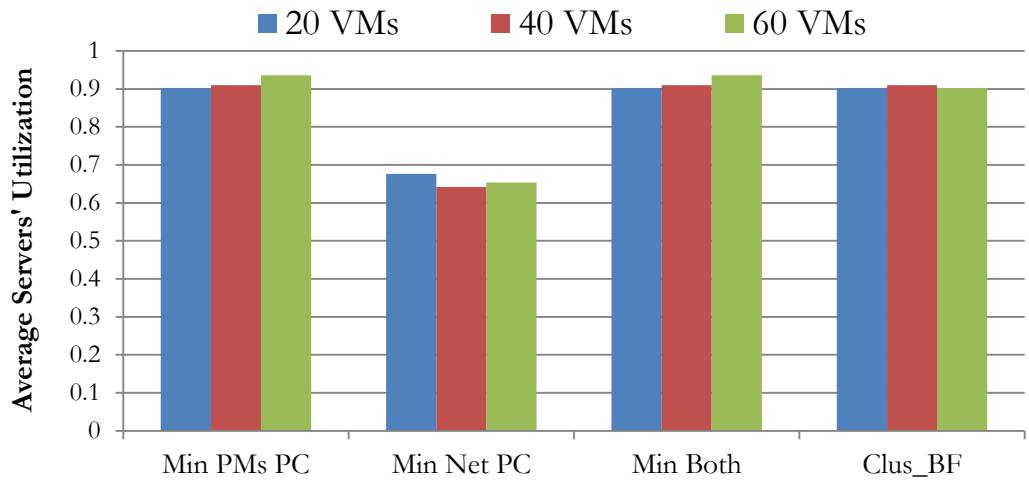


Figure 7.3. Average servers' utilization examining three sets of VMs; 20, 40, and 60 for the three objective functions and the developed algorithm Clus_BF

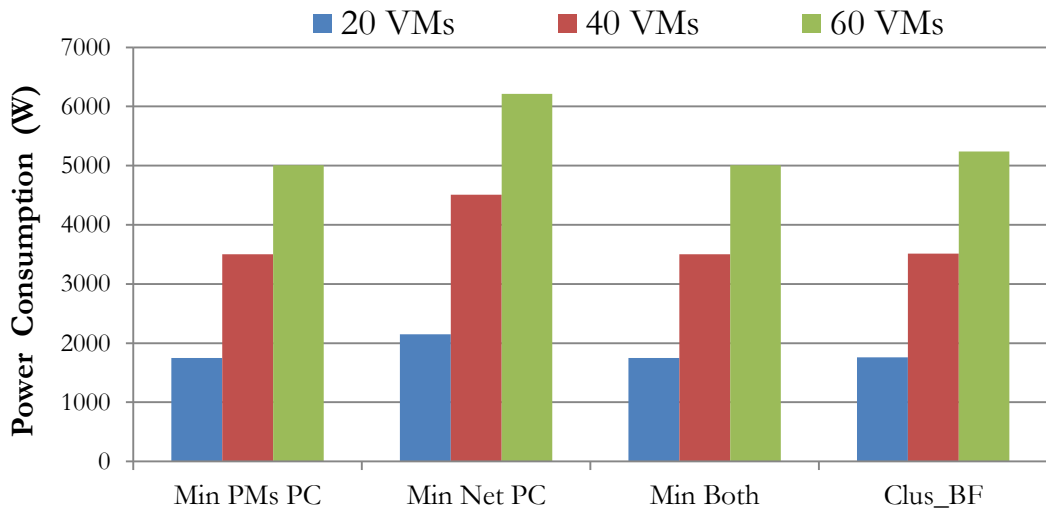


Figure 7.4. The total power consumption different sets of VMs; under the three models and the developed algorithm Clus_BF

The network power minimisation model tries to allocate communicating VMs in the same servers as much as possible. As a result, traffic flow among servers, ONU power consumption, and average transmission delay

are reduced. However; the total power consumption is increased compared to the minimisation of servers' power consumption model as more servers are switched on to accommodate groups of communicating VMs.

For the examined VMs (20, 40, and 60 VMs), the power consumption under the Min Net PC model is reduced by 20% compared to the Min PMs PC model.

Figures 7.2 and 7.3 present the average servers' utilisation and number of activated servers for the different models for 20, 40, and 60 VMs. Min Net PC also results in lower utilisation of servers' resources compared to the Min PMs PC model as more servers are switched on to serve the same number of VMs. The network power consumption minimisation model results in average utilisation of about 65%, while under the servers' power consumption minimisation model; servers' utilisation is more efficient and approaches 90%.

Figure 7.1 presents the average delay for the different objectives. Minimisation of network power reduces the total traffic flow among servers, it therefore results in reduction in the average transmission delay. Minimisation of network power results in a decrease in the average delay by 60% compared with the minimisation of servers' power consumption model.

The results of the multi-objective model as shown in Figures 7.2, 7.3, and 7.4 achieves similar power consumption, number of used servers, and servers' average utilisation as the minimizing the servers' power consumption model for 20, 40, and 60 VMs. Average delay as shown in Figure 7.1 has shown similar results with the minimised network power for

20 and 40 VMs. For 60 VMs, the average delay is lower than the approach that minimised the servers' power consumption. The average delay decreased by 29%.

Next, the greedy algorithm developed is described to mimic the behaviour of the multi resources constraints placement MILP and to act as verification for results obtained from the MILP model. Also for comparison purposes, an implementation of conventional placement algorithms; Best Fit Decreasing Bin-Packing (BFD) and random (Rnd) is reported and the results are compared with our developed greedy algorithm.

7.4 Clus-BF Greedy algorithm for VM placement in PON data centre

The multi-objective optimisation problem is an NP-hard problem. Well-known optimisation algorithms derived from ant colony or genetic algorithms can be implemented for optimum solution. Such optimisation algorithms are avoided as the main concern in a practical implementation is the time needed by the algorithm, in addition to the reduction in energy consumption achieved as these algorithms need to be implemented in hardware.

In order to solve the multi-objective problem of minimising the total power consumption of physical machines and network, an algorithm that combines VMs clustering and best fit placement with respect to CPU, RAM, and link resources is proposed. The input parameters are the network topology(G), traffic matrix for VMs (A^{if}), VM requirements of CPU , VM requirement of RAM, servers' CPU capacity (P_j), servers' RAM capacity (M_j), and link capacity (C_j).

The algorithm as described in the pseudo-code depicted in Figure 7.5 consists of two main steps. The first step is to cluster VMs in groups by visiting each VM and creating a group for that VM to include other VMs with mutual traffic only. Then these groups are sorted in a decreasing order starting with groups that have the highest intra traffic flows as described in lines 1 and 2.

Secondly, the algorithm in line 3 visits a sorted list of groups one by one and filters the groups to generate a queue of clusters of VMs with total resources that do not exceed a server resources in terms of CPU, memory and link capacity. Then in line 4, the algorithm searches for a server with the minimum remaining resources that are still sufficient and that best fit the VMs group with the highest mutual intra traffic and place such VMs in that server. In lines 7, 8 and 9, the placed VMs are removed from the list of VMs, and the remaining resources of the server selected are updated. Line 13 is used to search for all remaining not assigned VMs and sort them in a non-decreasing order for best fit assignment. In line 15 and 16, a list of servers with enough resources is searched and selection is based on the server with existing VMs that has the highest mutual bandwidth. Lines 17-19 are used to update resources of the server after the placement of VMs. In line 22 the algorithm ends. The output of the algorithm is the mapping between VMs and PMs, average server utilisation, total power consumption, mapping between VMs and servers (δ_j^i), traffic flow in all links (U^{ij}), average delay (d), number of servers used (\overline{N}), total power consumption (PC), and average servers utilization (\overline{u}).

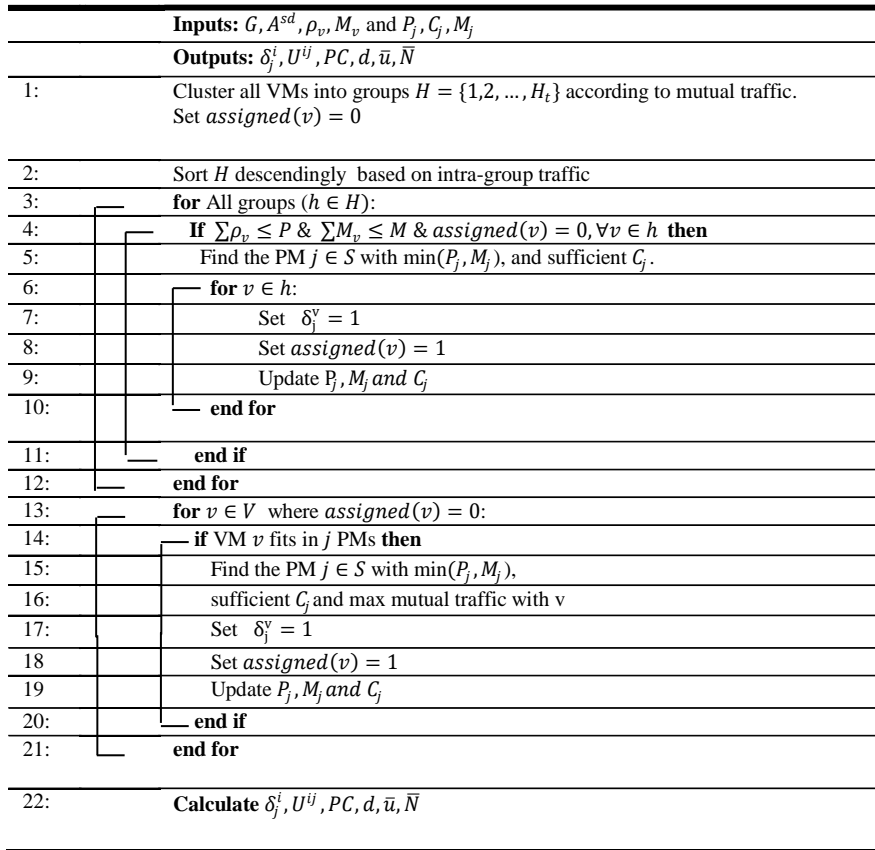


Figure 7.5. Clus_BF Algorithm for Multi-constraint resource provisioning in AWG PON Cell

7.5 Evaluation of the CLus-BF algorithm

The Clus-BF greedy algorithm as shown in Figure 7.5 has produced results similar to those obtained from the multi-objective model in terms of total power consumption; number of servers used, and average servers' utilisation. Conventional Best Fit Bin Packing (BFD-BP) and random (Rnd) placement algorithms are implemented so that the results can also be compared and evaluated with the developed algorithm in PON data centre. Comparison will be in terms of power consumption, average server's

utilisation, inter-flow traffic reduction, average delay, and number of servers used.

Compared with Random placement algorithm, the BFD-BP based scheduler results in efficient utilisation of computing resources of servers and hence reduces the number of servers used and the total power consumption. Consolidation of workloads as single server's resources can be sliced to be shared efficiently by multiple VMs results in efficient utilisation of servers' resources. Figure 7.8 shows that the average server utilisation is 90% for almost all the sets of different number of VMs for BFD, while the random placement results in utilisation values between 30%-40% of PMs' resources.

The underutilised resources are a result of the behaviour of the random algorithm as random placement scheduler tends to distribute VMs randomly rather than consolidating workloads on minimum set of machines.

The BFD-BP based scheduler as shown in Figure 7.6 and 7.7 achieves 47% reduction in overall power consumption as the number of used servers is reduced to 55% compared to the case with random scheduler.

The BFD-BP algorithm achieves excellent results for efficient utilisation and power savings; however traffic flow among communicating VMs hosted by different PMs is not part of the objective to be minimised. The developed Clus-BF algorithm achieves power consumption minimisation while addressing the multi constraints placed on resources as described in the MILP model in terms of traffic flow on communicational links.

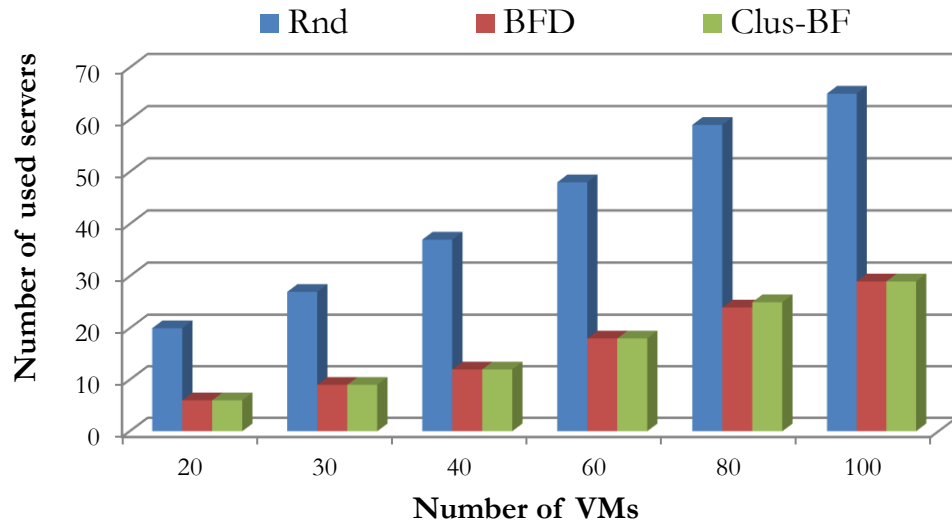


Figure 7.6. Total number of used servers for the random, best fit decreasing, best effort best fit, and Cluster best fit algorithms

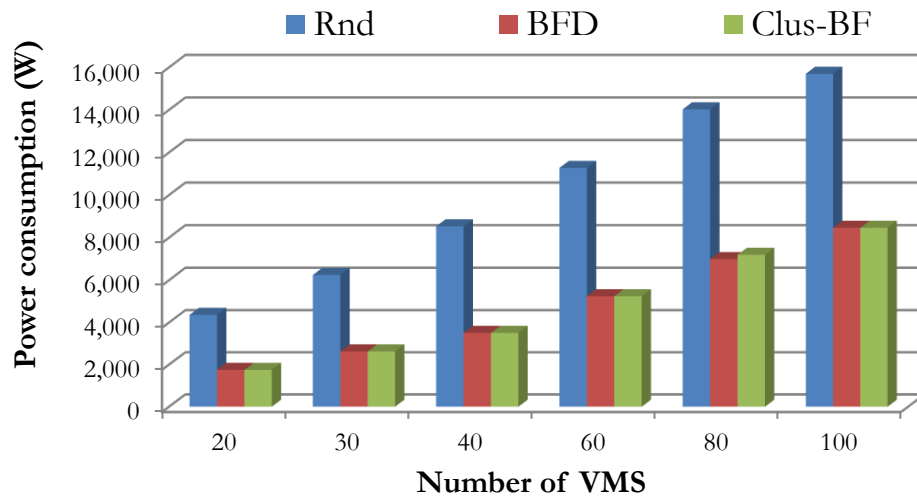


Figure 7.7. Total power consumption of traffic flow reduction of the Cluster-BF, Random and Best Fit Decreasing algorithms

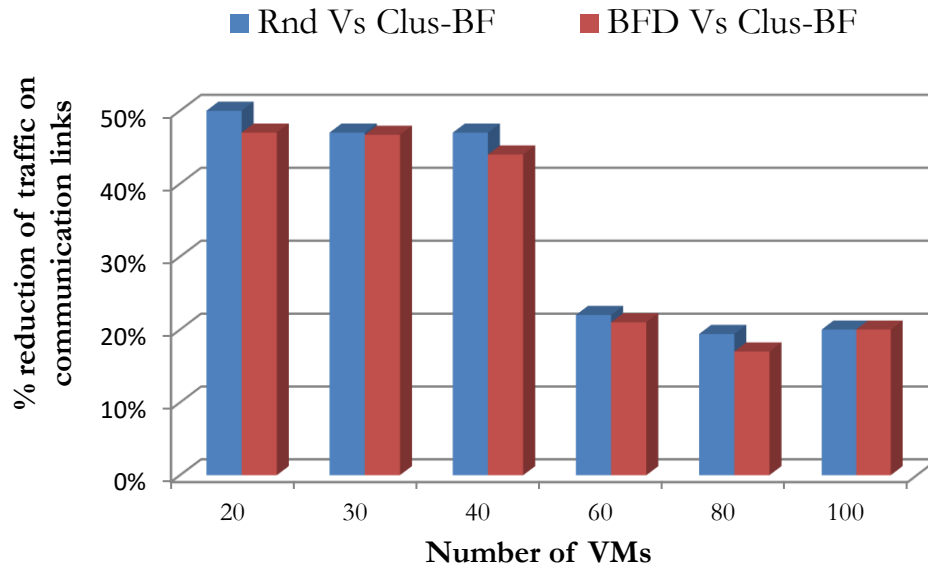


Figure 7.8. The percentage of traffic flow reduction of the Cluster-BF, Random and Best Fit Decreasing algorithms

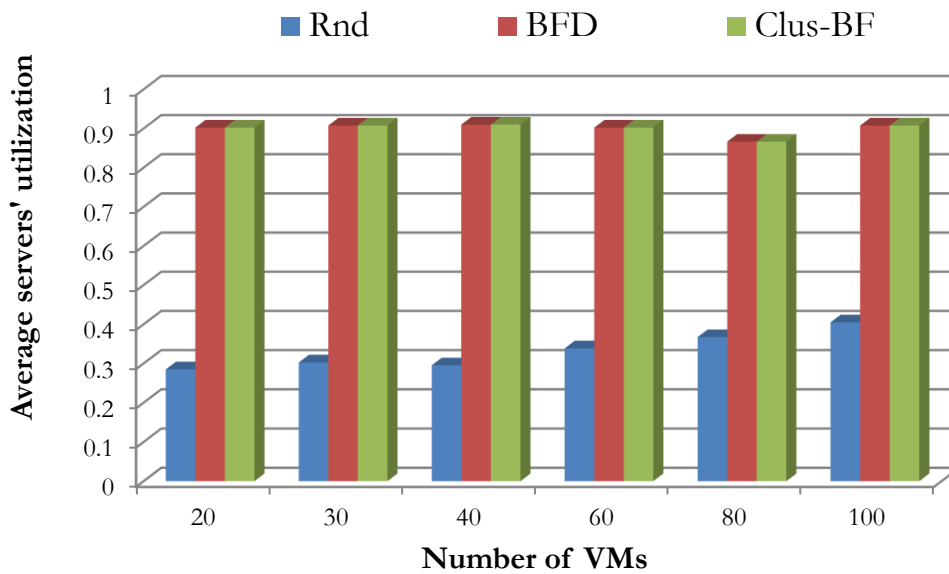


Figure 7.9. Average servers' utilization of traffic flow reduction of the Cluster-BF, Random and Best Fit Decreasing algorithms

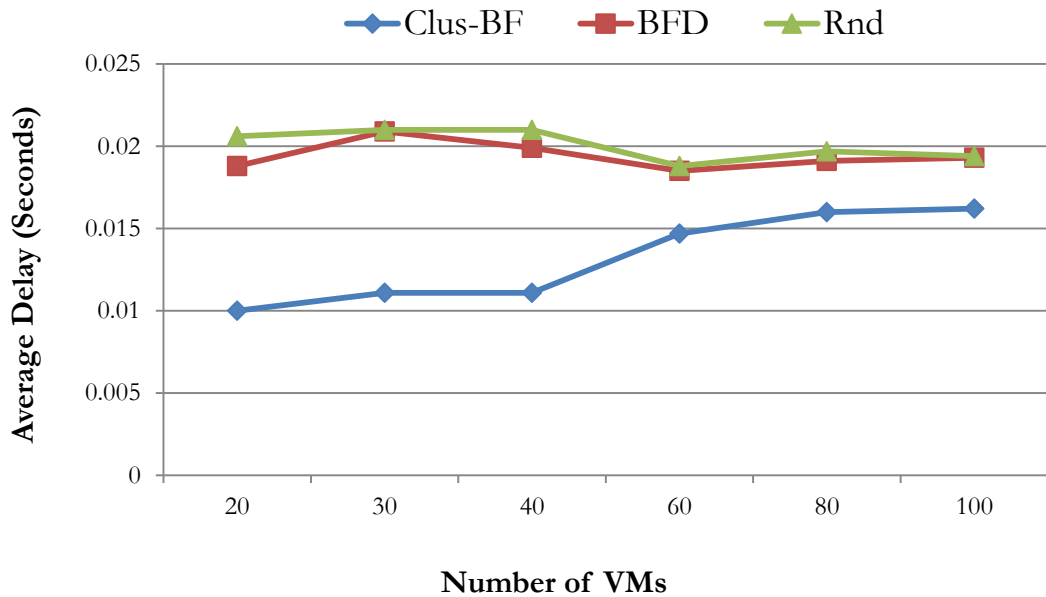


Figure 7.10. Average delay of traffic flow reduction of the Cluster-BF, Random and Best Fit Decreasing algorithms

In comparison with BFD-BP and random algorithms, the Clus-BF algorithm shows good performance in terms of power consumption, and reduction of traffic flow for the different sets of VMs. Clus-BF algorithm as shown in Figure 7.7 produces results very similar to BFD-PB for power consumption, number of servers used and average server utilisation. Clus-BF as expected demonstrates an average of 47% savings in power consumption with more efficient PM resources utilisation reaching (90%) if compared with random algorithm.

On the other hand, traffic flow reduction is shown in Figure 7.8 to compare the Clus-BF with the other examined algorithms. For the different sets of VMs, Clus-BF always produces a reduction in the inter-flow traffic on communicational links between servers. Clus-BF achieves 20%-50% and 17%-47% reduction compared with random and BFD algorithms respectively. The decrease in the percent of the reduction of the traffic flow

is a result of the increase of the number of VMs with communication requirements as the number of servers are fixed for all sets of examined VMs to be provisioned. Obviously with constrained physical resources as the number of VMs increases, the size of flow increases as well and as a result the transmission delay increases. This is clearly happening in our modelled architecture when the size of VMs increases beyond 40 VMs. This is also shown in the results of Figure 7.10. Therefore the percent of reduction shown in Figure 7.8 decreases.

7.6 Summary

In this chapter, a further investigation of the proposed AWGR PON based architecture for cloud application is studied. A mathematical optimisation model was developed for resource provisioning considering minimisation of power consumption, delay, and both. The results show that delay can be decreased by 62% for delay-sensitive applications and power consumption can be decreased by 22% for non-delay sensitive applications. Multiple resource provisioning for servers' physical resources and communication traffic is tackled as well. A mathematical optimisation model was developed along with a placement algorithm to solve the multiple resource optimisation problem. The results show good agreement between the results from the MILP models and the introduced greedy algorithm. Further study that compares the proposed algorithm with algorithms like random and BFD-BP are also presented and evaluated for delay, power consumption, average servers' utilisation and number of servers used. Our algorithm demonstrated an average of 47% savings in power consumption with more efficient PM

resources utilisation reaching (90%) when compared with the random algorithm. The proposed algorithm has also shown a maximum of 50% and 47% reduction in delay when compared with random and BFD algorithms respectively.

8 Energy efficient server centric PON data centre network

8.1 Introduction

The AWGR-based PON architecture has achieved energy savings of 45% and 80% compared to the Fat-Tree and BCube architectures, respectively. The main drawback of this design, however, is its high deployment cost as all servers are equipped with tuneable transceivers (or a transceiver with several fixed tuned wavelengths). In this chapter, a new server centric PON architecture is introduced. The different merits of this form of PONs and servers in the design of energy efficient, high capacity, low cost, scalable, and highly elastic networking infrastructures that support the applications and services hosted by modern data centres is considered. The proposed architecture eliminates the need for costly and power hungry devices such as tuneable lasers and electronic switches. The work in this chapter includes a benchmark study to compare the 3-tier architecture with the proposed server-centric design with respect to power consumption. A mathematical optimisation model is developed along with a heuristic for energy efficient routing and presented for the described architecture.

8.2 Traffic locality study for PON deployment in data centre networks

In data centres, the nature of traffic differs from PONs' traffic. The traffic within a data centre, as shown and discussed in Chapter 4, in Figure 4.3, can be categorised into four main types; (i) In-Out traffic destined out of the data centre through access switches, aggregation switches and core routers, (ii) Intra-rack traffic between servers located in the same rack through the top of rack access switches, (iii) Inter-rack traffic between servers located in different racks through the top of rack access switches and the layer-2 aggregation switch linking the two racks, (iv) Out-In traffic entering the data centre through core routers, aggregation switches, and top of rack access switches.

A number of studies have analysed the characteristics of the traffic generated within data centres [43], [78], [91]. Based on the application, the majority of data centre traffic may remain inside the racks or span several racks. The studies have shown that 50% of the enterprise and university data centres' traffic is inter-rack flows while for cloud data centre type, the inter-rack flows are under 25%. According to [43], [78], [91] the intra rack and inter rack traffic flow percentages can take a range of ratios; 20%-80%, 50%-50%, and 80%-20%.

Therefore, eliminating the access and aggregation switches and replacing them with directional PON splitters/couplers, will result in over-loading the OLT switch making it the bottleneck for all types of traffic. Despite the fact that the OLT switch backplane can provide non-blocking hundreds of Gb/s interconnection among its cards, offloading the burden on the OLT switch will avoid undesired delays and power consumption resulting from O/E/O conversions, queuing, buffering and processing. Hence; the proposed PON architecture design addresses all the challenges to furnish an interconnection fabric to sustain all types of traffic patterns for the different applications that can be hosted within the data centre.

In the subsequent section, a detailed description of the server centric architecture relying mostly on PONs by re-designing the current paradigm of PONs used for access networks is presented.

8.3 The architecture of the server-centric PON cloud data centre

Unlike the proposed design of the AWG based PON data centre presented in Chapter 5, the design depicted in Figure 8.1, eliminates the need for costly tuneable lasers and facilitates high speed interconnection among racks within a PON cell by dividing each rack into 4 groups each of 8 servers connected by a TDM star coupler. Three of the 4 groups are connected to the other three racks, and one group connects the rack with the OLT at 10 Gb/s rate. Therefore, OLT port receives 4 x 10 Gb/s. Inter-rack communication between servers that do not belong to groups with a direct connection is established by using one of the

servers of the group with a direct connection with the rack of the destination server. Relay server selection can be based on servers' utilisation or traffic load within the PON group. Similarly servers can establish connection with the OLT switch. Connections between the OLT switch and racks can be provisioned either through the use of a star coupler (TDM) as depicted in Figure 3 or via an AWGR (WDM) to provide more bandwidth.

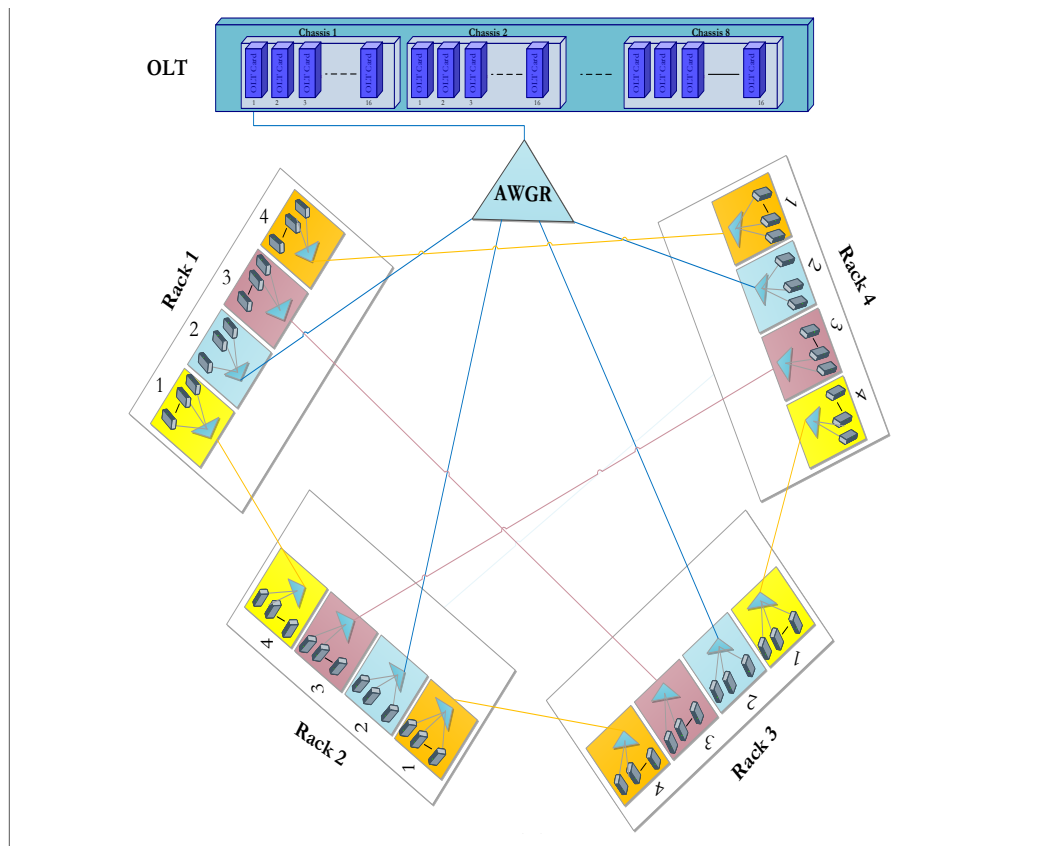


Figure 8.1. Proposed PON Cell design with no tuneable lasers

Servers in different groups in the same rack are connected via a backplane, for example the terabit capacity passive polymer optical backplane in [79]. This technology as depicted earlier in Chapter 4 in Figure 4.5 employs a passive

backplane with multimode polymer waveguides and can provide non-blocking full mesh connectivity with 10 Gb/s rates per waveguide, exhibiting a total capacity of 1 Tb/s. Other options for intra-rack interconnection are described in Section 4.5. These proposed technologies manage intra-rack communication among servers within the rack without the need to reach the OLT switch.

An observation in this architecture is that the 10 Gb/s port rate is shared by the 128 servers in a PON cell. If each server needs a sustained 1 Gb/s rate the capacity available to a PON cell may not be enough. If a high server activity ratio (proportion of total number of servers that are simultaneously active and working at the full 1 Gb/s) is expected in the data centre, then the architecture can be modified so that the 10 Gb/s is shared by a smaller number of servers which calls for a larger number of OLT cards. At lower activity ratios the PON protocol can do elastic bandwidth allocation to cater for traffic bursts. Furthermore the data centre load can be distributed among different PON cells for load balancing or alternatively consolidated in fewer PON cells to save power through power shedding in response to long term daily load variation, followed by sleep in response to shorter inactivity periods within the hour / minutes. Power saving and architecture and protocol optimisation for performance are interesting topics for future research.

In PON residential access networks, ONU to ONU communication is not a major concern as traffic is either transmitted from ONUs to the OLT or from OLT to the ONUs. In addition current PONs in access network as studied and presented in Section 8.2 do not perform well for rates above 1Gb/s where point

to point fibre links become more energy efficient than a PON. However, the cellular architecture depicted in the proposed design (See architectures in Chapters 5-7) does not suffer the limitations of the normal PON as (i) within the rack, the optical backplane can provide full wavelength rate, and (ii) within the cell of racks, the AWGRs give full wavelength connectivity and hence server to server communication within the cell can appear as point to point.

8.4 Power consumption benchmark study of server-centric PON design against the 3-tier conventional DCN

In this section, a benchmarking study that compares the power consumption of the proposed server centric PON data centre architecture to the most commonly implemented data centre architecture nowadays, the 3-tier conventional data centre is presented.

A 3-tier conventional data centre architecture is depicted in Figure 4.3. Core switches at tier 3 are connected by 10GE links to aggregation switches at tier 2. The number of core switches is typically limited to 8 as the Equal Cost Multipath routing protocol (ECMP) can only support 8 paths [38]. Aggregation switches are typically double the number of core switches. Top of Rack (TOR) access switches located at tier-1 can connect 20-40 servers with 1GE links [38]. Table 8.1 breaks down the power consumption of the networking equipment of a 3-tier data centre architecture provisioning connectivity for 5120 servers.

Table 8.1. Breakdown of the power consumption of networking equipment of 3-tier data centre architecture provisioning connectivity for 5120 servers

	Access switch	Aggregation switch	Core switch
Model	Cisco nexus 2148T [92]	Cisco nexus 5020 [93]	Cisco nexus 7000 F2 [94]
Capacity	32-48 ports (1Gb/s+ 4SFP- 10Gb/s)	40 ports (each 10 Gb/s)	48 ports (each 10 Gb/s)
Power consumption	200W	750W	7.5W per port
Number required	160	8	4 core switches (8 ports each)
Total power consumption	32kW	6kW	240

The power consumption of PON equipment operating at 10Gb/s was discussed in Section 5.6.3.

Figure 8.2 depicts a typical schematic architecture of an ONU used for FTTx access network [95] where a subscriber line interface (SLIC) is used for telephony voice service, the Multimedia Over Coax (MoCA) chipset is used for TV service provisioning. in Section 5.6.3, the PON data centre ONU is not required to provide video and audio services, hence, the SLIC and MoCa which are reported in [60] to consume 150mW and 1262mW, respectively, are not required in the PON data centre ONU design as seen in Figure 8.3. The PON data centre ONU can have a passive GbE switch to support multiple servers'

connection. In small data centres with low traffic, a single ONU can be used to connect multiple servers to reduce the cost and power consumption.

Table 8.2 gives the active and idle power consumption of the major components of a 10Gb/s ONU designed for telecom access telecom networking covering 20km based on a study presented in [96]. From these values, the ONU in active mode consumes 6.2W, whereas in the sleep mode it consumes 0.42W. As mentioned in Section 5.6.3, PON data centre interconnection will not exceed 100 meters. Assuming linear profile to estimate the transceiver power consumption for the proposed PON data centre ONU with interconnections not exceeding 100 meters, ONU's transceiver power consumption is given as 17.5mW for active mode. Total power consumption of an ONU is 2.72W. For idle (sleep) mode, ONU will consumes 0.42W.

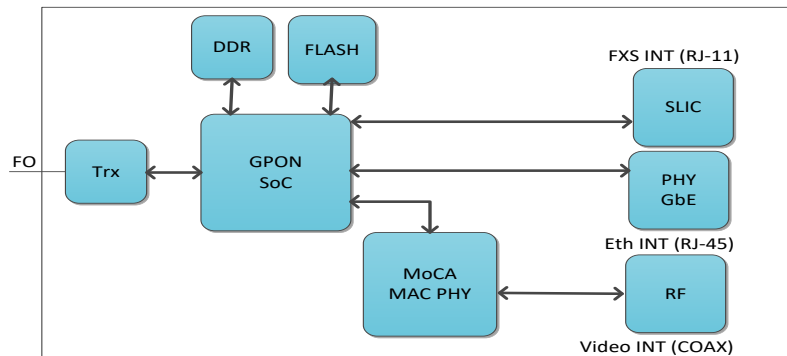


Figure 8.2. Architecture of an ONU in access network [96]

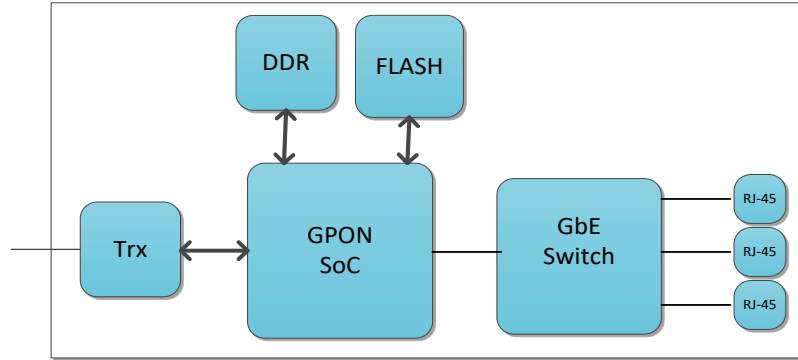


Figure 8.3. Proposed architecture of an ONU for PON data centre. GbE switch is passive and optional for the case where multiple servers to be connected by one ONU

Table 8.2. Power consumption for main components of 10G ONU for 20km reach [96]

	Active mode	Sleep mode
Transceiver	3.5W	0 W
System on chip (SoC)	2W	0.4W
DDR (RAM)	0.7W	0.02W

Table 8.3 calculates the power consumption of PON data centre design supporting 10 Gb/s with typical rack to rack interconnection distances. We consider two designs; one uses an ONU to connect one server and the other uses an ONU to connect to two servers.

Table 8.3. Breakdown analysis for power consumption for the proposed data centre architecture to support 10G and for provisioning 5120 servers

	Capacity and interfaces	number of servers per port	Number required	Power consumption	Total power consumption
OLT	8 ports per card (10Gb/s)	128	40 ports (5 cards)	1 kW per card 125W/port	5 kW
		64	80 ports (10 cards)	1 kW per card 125W/port	10 kW
ONU	One Ethernet interface (10Gb/s)	-	5120 ONUs	2.72W	13.926 kW
	Two Ethernet interface (10Gb/s)	-	2560 ONUs	2.72W	6.963 kW

For a split ratio of 128, the total power consumption of a PON data centre architecture supporting 5120 servers is 18,926 and 11,963W for ONUs with one interface and two interfaces respectively. Networking equipment energy savings of the PON architecture with ONUs of a single interface compared to the conventional architectures is 50.5% and for ONU with two interfaces is 69%. For a splitting ratio of 64, the total power consumption for 5120 servers is 23,926 and 16,963W for ONU with one interface and two interfaces, respectively. Energy savings compared to the conventional architectures will be 38% for the ONU option with one interface and 56% for the ONU option with two interfaces. Note that the power consumption calculation of the PON data centre architecture does not take into account the power consumption of the servers taking part in the routing of inter rack traffic.

In the next section, an optimisation model for energy aware routing within the server centric design is presented.

8.5 MILP model for energy aware routing in PON data centre

As discussed in Section 8.3 servers can take part in the routing of inter rack traffic to off-load the burden on the OLT switch and make efficient use of the servers underutilised processing capabilities. In this section, an optimisation MILP model is developed to minimise the power consumption of the server centric PON cell design by optimising the routing of inter rack traffic within the PON cell.

The parameters and variables used in the model:

Parameters

L	Set of all nodes
A	Set of active nodes (servers and OLT ports)
P	Set of passive nodes (PON)
N_i	Set of neighbours of node $i \in L$
s and d	Denotes source and destination, $s, d \in A$
m and n	Denotes end points of a physical link, $m, n \in L$
C_{mn}	Capacity of link $m, n \in L$
PO	Idle power consumption of a server
PI_{OLT}	Idle power consumption of an OLT port
PM	Maximum power consumption of a server
PM_{OLT}	Maximum power consumption of an OLT port
R_s	Data rate per request in b/s
γ_{src}	Portion of a server processing capacity used for transmitting one request (server acts as a transmitter)
γ_{dest}	Portion of a server processing capacity used for processing one received request (server acts as a destination)
γ_{fwd}	Portion of a server processing capacity used for forwarding one request (server acts as a relay)

β_{src}	Portion of the OLT processing capacity used for transmitting one request
β_{dest}	Portion of the OLT processing capacity used for processing a single received request
β_{fwd}	Portion of the OLT processing capacity used for forwarding one request
T_{scr_i}	Total traffic originated and transmitted by node i
T_{dest_i}	Total traffic destined to node i
Th_{olt}	Threshold on maximum allowed CPU utilization which determines the maximum number of requests that can be allowed to be forwarded by the OLT switch
Th_{server}	Threshold on maximum allowed CPU utilization which determines the maximum number of requests that can be allowed to be forwarded by a server
Uti_{Max}	The threshold on maximum allowed server utilization
T	Processing delay per request
α^{sd}	Traffic demand from source node s to destination node d

Variables:

χ_{mn}^{sd}	Portion of traffic demand (s, d) traversing physical link (m, n) in b/s
Tin_i	Total traffic entering node i in b/s

$Tout_i$	Total traffic leaving node i in b/s
$Tfwd_i$	Total traffic forwarded by node i in b/s
USr_i	Utilization of server i
$UOLT_i$	Utilization of OLT port i
PS_i	Power consumption of server i
PO_i	Power consumption of OLT i
δ_i	Defined as $\delta_i = 1$ if server node i is activated, otherwise $\delta_i = 0$
δOLT_i	Defined as $\delta OLT_i = 1$ if OLT port i is activated, otherwise $\delta OLT_i = 0$
Tr_{mn}	The total traffic traversing physical link (m, n)
ψ_{mn}^{sd}	Defined as the binary equivalent of χ_{mn}^{sd} , $\psi_{mn}^{sd} = 1$ if traffic demand (s, d) traversing physical link (m, n) , otherwise $\psi_{mn}^{sd} = 0$
D_i	Average delay experienced by requests queued to be served by server i
Y_{im}^{sd}	Defined as the forwarding delay experienced by traffic demand (s, d) at node i to be transmitted to node m

The power consumption of a cell in the server centric PON data centre architecture is composed of the power consumption of servers, optical network units (ONUs), and the optical line terminal (OLT) port. In our model we will only consider the power consumed by servers and the OLT as the ONU can be assumed for full load to have power consumption of 2.72W.

Before introducing the model, we define the power consumption of the PON cell and other quantities we will need in the model and for performance evaluation:

Power consumption of server i

$$PS_i = (PO \delta_i) + USr_i(PM - PO)$$

$$\forall i \in A \text{ and } i \neq OLT \quad (8.1)$$

Power consumption of OLT port i

$$PO_i = PI_{OLT} \delta_{OLT_i} + UOLT_i (PM_{OLT} - PI_{OLT})$$

$$\forall i \in A \text{ and } i = OLT \quad (8.2)$$

Total traffic entering active node i

$$Tin_i = \sum_{s \in A} \sum_{d \in A: s \neq d} \sum_{n \in N_i} \chi_{ni}^{sd}$$

$$\forall i \in A \quad (8.3)$$

Total traffic originating and transmitting by active node i

$$Tscr_i = \sum_{d \in A: d \neq i} \alpha^{id}$$

$$\forall i \in A \quad (8.4)$$

Total traffic destined to active node i

$$T_{dest_i} = \sum_{s \in A: s \neq i} \alpha^{si} \quad (8.5)$$

$$\forall i \in A$$

Total traffic transmitted out of an active node i

$$T_{out_i} = \sum_{s \in A} \sum_{d \in A: s \neq d} \sum_{n \in N_i} \chi_{in}^{sd} \quad (8.6)$$

$$\forall i \in A$$

Total traffic forwarded by active node i

$$T_{fwd_i} = \left(\sum_{s \in A} \sum_{d \in A: s \neq d} \sum_{n \in N_i} \chi_{ni}^{sd} \right) - \left(\sum_{s \in A: s \neq i} \alpha^{si} \right) \quad (8.7)$$

$$\forall i \in A$$

Utilisation of server i

$$USr_i = \left(\left(\frac{T_{scr_i}}{R_s} \right) \gamma_{src} \right) + \left(\left(\frac{T_{dest_i}}{R_s} \right) \gamma_{dest} \right) + \left(\left(\frac{T_{fwd_i}}{R_s} \right) \gamma_{fwd} \right) \quad (8.8)$$

$$\forall i \in A \text{ and } i \neq OLT$$

Utilisation of OLT port i

$$UOLT_i = \left(\left(\frac{T_{scr_i}}{R_s} \right) \beta_{src} \right) + \left(\left(\frac{T_{dest_i}}{R_s} \right) \beta_{dest} \right) + \quad (8.9)$$

$$\left(\left(\frac{T_{fwd_i}}{R_s} \right) \beta_{fwd} \right)$$

$$\forall i \in A \text{ and } i = OLT$$

Average server queuing delay (D_i)

The delay experienced by a request waiting to be processed for transmission is a function of the number of requests in the queue. The first request arriving to the queue will experience no queuing delay only processing delay T . The second request in the queue will experience a total delay of $2T$ and the third request will wait $3T$...etc. So the average delay experienced by requests relayed by server i with a queue of R_i requests is calculated as:

$$D_i = \frac{T \sum_{j=1}^{R_i} (j)}{R_i} = \frac{T R_i (R_i + 1)}{R_i \cdot 2} = \frac{T (R_i + 1)}{2} \quad (8.10)$$

where T is the processing delay of a request, and R_i is the total number of requests to be forwarded by a server including requests originated and relayed by the server.

Server's average queuing delay then can be calculated using:

$$D_i = \frac{T}{2} \left(\frac{T_{out_i}}{R_s} + 1 \right) \quad (8.11)$$

$$\forall i \in A$$

Average path queuing delay:

$$\chi_{mn}^{sd} \geq \psi_{mn}^{sd} \quad (8.12)$$

$$\forall s, d, m \in A \text{ \& } \forall n \in N_m$$

$$\chi_{mn}^{sd} \leq M \psi_{mn}^{sd}$$

$$\forall s, d, m \in A \ \& \ \forall n \in N_m \quad (8.13)$$

$$Y_{im}^{sd} \leq M \psi_{im}^{sd}$$

$$\forall s, d, i \in A \ \& \ \forall m \in N_i \quad (8.14)$$

$$Y_{im}^{sd} \leq D_i$$

$$\forall s, d, i \in A \ \& \ \forall m \in N_i \quad (8.15)$$

$$Y_{im}^{sd} \geq D_i - M(1 - \psi_{im}^{sd})$$

$$\forall s, d, i \in A \ \& \ \forall m \in N_i \quad (8.16)$$

Equations (8.12) and (8.13) relate variable χ_{mn}^{sd} to its binary equivalent ψ_{mn}^{sd} . Equations (8.14)-(8.16) are used to calculate the Y_{im}^{sd} variable (forwarding delay experienced by traffic demand (s, d) at node i to be sent to node m). The queuing delay experienced by a request is the total forwarding delay experienced at all node along the path from the source to the destination and is given as:

$$\sum_{s \in A} \sum_{d \in A: s \neq d} \sum_{i \in A} \sum_{m \in N_i} \left(\frac{\alpha^{sd}}{R_s} \right) Y_{im}^{sd} \quad (8.17)$$

The model objective function is defined as follows:

Minimise:

$$\begin{aligned} & \sum_{i \in A: i \neq OLT} (PO \delta_i) + USr_i (PM - PO) \\ & + \sum_{i \in A: i = OLT} (PI_{OLT} \delta_{OLT_i}) + UOLT_i (PM_{OLT} - PI_{OLT}) \end{aligned} \quad (8.18)$$

Equation (8.18) gives the model objective which is to minimise the total power consumption within a PON cell considering the servers and OLT ports power consumption.

Subject to:

Flow conservation constraint

$$\sum_{\substack{n \in N_m \\ m \neq n}} \chi_{mn}^{sd} - \sum_{\substack{n \in N_m \\ m \neq n}} \chi_{nm}^{sd} = \begin{cases} \alpha^{sd} & m = s \\ -\alpha^{sd} & m = d \\ 0 & \text{otherwise} \end{cases} \quad (8.19)$$

$$\forall s, d, m \in L : s \neq d$$

Constraint (8.19) is the flow conservation constraint. It ensures that the total traffic going into a node is equal to the total traffic leaving it for all nodes except the source and destination of a demand.

Link capacity constraint

$$\sum_{s \in N} \sum_{d \in N: s \neq d} \chi_{mn}^{sd} \leq C_{mn} \quad (8.20)$$

$$\forall m \in L \text{ and } \forall n \in N_m$$

Constraint (8.20) ensures that the traffic traversing any physical link does not exceed its capacity.

Server utilisation constraint

$$USr_i \leq Uti_{Max} \quad (8.21)$$

$$\forall i \in A, \text{ and } i \neq OLT$$

Constraint (8.21) assures that the utilisation of a server as calculated in Equation (8.8) does not exceed the predefined threshold on servers' utilisation to avoid performance degradation and overheating.

OLT power consumption constraint

$$PO_i \leq PM_{OLT} \quad (8.22)$$

$$\forall i \in A, \text{ and } i = OLT$$

Constrain (8.22) ensures that the total power consumption of the OLT does not exceed the maximum power as it is governed by the OLT utilisation variable($UOLT$).

Constraints on forwarding requests by the OLT and servers

$$\left(\left(\frac{T_{fwd_i}}{R_s} \right) \beta_{fwd} \right) \leq Th_{olt} \quad (8.23)$$

$$\forall i \in A \text{ and } i = OLT$$

$$\left(\left(\frac{T_{fwd_i}}{R_s} \right) \gamma_{fwd} \right) \leq Th_{server} \quad (8.24)$$

$$\forall i \in A \text{ and } i \neq OLT$$

Constraints (8.23) and (8.24) set the threshold on the maximum number of requests allowed to be forwarded through the OLT switch and servers, respectively. The threshold on the OLT switch ensures offloading some of the burden on it and the threshold on the servers will balance the load on the active servers to avoid undesired delays resulting from queuing requests.

Constraints to switch off idle servers

$$M USr_i \geq \delta_i$$

$$\forall i \in A \text{ and } i \neq OLT \quad (8.25)$$

$$USr_i \leq M \delta_i \quad (8.26)$$

$$\forall i \in A \text{ and } i \neq OLT$$

Constraints (8.25) and (8.26) ensure that servers with zero utilization are switched off.

Constraints to switch off idle OLT port

$$\begin{aligned}
 M UOLT_i &\geq \delta OLT_i \\
 \forall i \in A \text{ and } i = OLT
 \end{aligned}
 \tag{8.27}$$

$$\begin{aligned}
 UOLT_i &\leq M \delta OLT_i \\
 \forall i \in A \text{ and } i = OLT
 \end{aligned}
 \tag{8.28}$$

Constraints (8.27) and (8.28) ensure that OLT ports with zero utilisation are switched off.

Constraint on Traffic bifurcation

$$\begin{aligned}
 \sum_{n \in N_m} \psi_{mn}^{sd} &\leq 1 \\
 \forall s, d, m \in L
 \end{aligned}
 \tag{8.29}$$

Constraint (8.28) guarantees that the request is routed through a single path.

8.6 Results and discussions

In this section, the power consumption of the proposed PON data centre architecture and the delay are evaluated when optimising the routing of inter rack traffic using the model introduced above.

A PON cell is modelled consisting of three racks each hosting 12 servers divided into 3 groups each of 4 servers. Table 8.4 summarises the input parameters of the model. In our evaluation we consider OLT equipment capable of supporting up to 8 GPON cells each of 10 Gb/s. The NEC CM7700S OLT of

8 ports each working at 1Gb/s is reported in [65] to consume 100 W. Assuming a linear power profile as a conservative estimate (usually equipment power consumption grows under the linear profile as the data rate increases), 1 port working at 10 Gb/s will consume 125W. An idle port is assumed to consume 70% of the full utilisation power consumption, i.e. 88W. The 88 W idle power accounts only if OLT chassis is fully loaded, if for example one port only is used the 88 W is an under estimation of idle power.

The servers considered in the modelled architecture are the same servers used in Chapter 6. The servers are equipped with Intel Xeon processors. The servers consume 301W at full CPU utilisation; of which 130 W is consumed by CPU and 171W is consumed by server's memory, motherboard, fan and other peripheral device [38]. Idle servers are also assumed to consume 70% of the full utilisation power consumption, i.e. 201W [38].

Table 8.4. Input data for the model

Link capacity (C)	10 Gb/s
Power consumption for idle servers (PI_{server})	201 W [38]
Maximum power consumption for servers (PM_{server})	301 W [38]
Power consumption for idle OLT port (PI_{OLT})	88 W
Maximum power consumption for OLT port (PM_{OLT})	125 W
Portion of a server processing capacity used for transmitting one request (server acts as transmitter) (γ_{src})	0.3%
Portion of a server processing capacity used for processing one received request (execution of a job) (γ_{dest})	2%

Portion of a server processing capacity used for relaying one request (server acts as a router) (γ_{fwd})	1.5%
Large constant (M)	1000
Portion of the OLT processing capacity used for transmitting one request (β_{src})	0.5%
Portion of the OLT processing capacity used for processing a single received request (β_{dest})	0.5%
Portion of the OLT processing capacity reserved for forwarding one request (β_{fwd})	1.5%
Total number of requests (Rt)	354
Forwarding processing delay per request (T)	0.4 ms
Maximum utilization allowed for a server (Uti_{Max})	90%
Data rate per request (Rs)	200 Mb/s

The sending, forwarding and processing of a request is assumed to consume 0.3%, 1.5%, 2% of the total server processing capacity, respectively. The server utilisation is constrained not to exceed 90%. The request size Rs is assumed to be equal to 200 Mb/s and the forwarding delay per request is assumed to be equal to 0.4 ms.

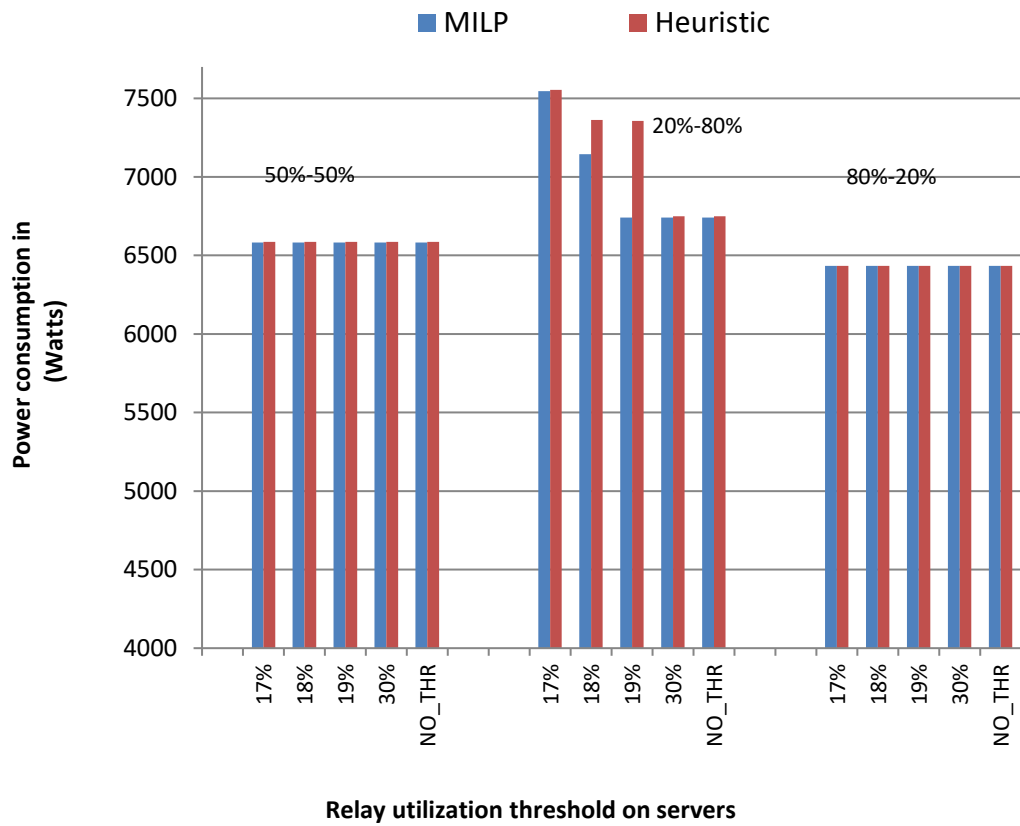


Figure 8.4. Total power consumption for cases where no threshold for OLT are enforced and server's threshold is varied for the different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack

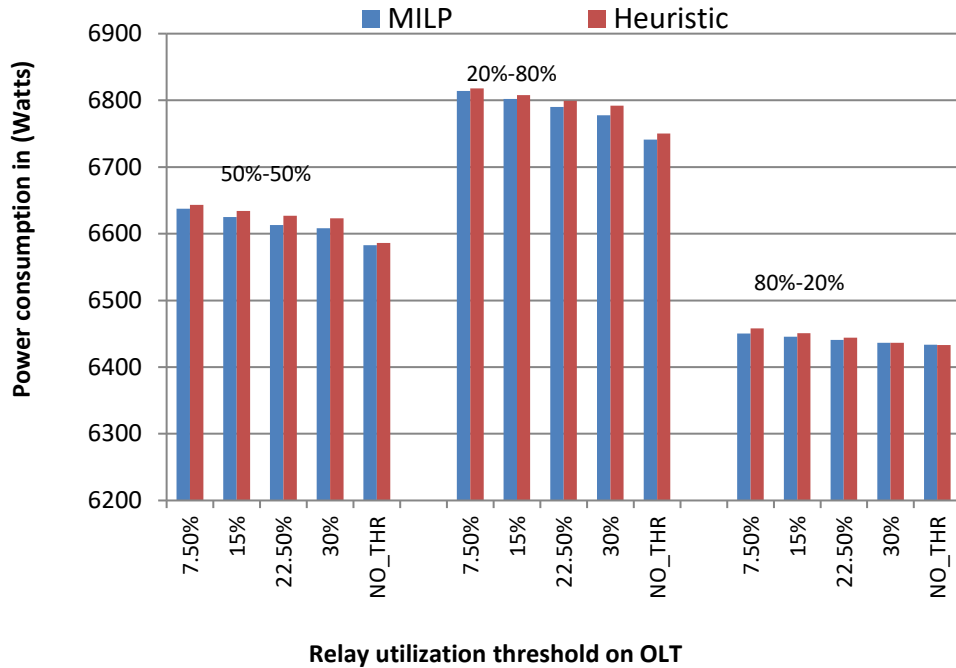


Figure 8.5. Total power consumption for a PON cell for cases where no threshold for servers are enforced and OLT threshold is varied for the different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack

The evaluated traffic scenarios consist of 354 requests; where 15% of the traffic (54 requests) is in/out traffic flows between the OLT and servers while the remaining traffic (300 requests) is inter rack and intra rack flows within the PON cell. A number of studies have analysed the characteristics of the traffic generated within data centres. As discussed above, depending on the nature of the application, traffic within the data centre may remain inside the racks or span several racks. The studies have shown that 50% of the enterprise and university data centres' traffic is inter-rack flows while for cloud data centre type, the inter-rack flows are under 25%. In our analysis we study the performance of the data centre PON cell under the following traffic scenarios which represent

the important cases and the extreme ends: the intra rack traffic/inter rack traffic percentages are: 20%/80%, 50% /50%, and 80%/20%.

This work evaluates the power consumption of the PON cell and the average queuing delay experienced by requests versus varying thresholds on the number of requests forwarded by the OLT switch and servers for the different intra rack and inter rack traffic percentages. As discussed above, varying the threshold on the number of requests forwarded by the OLT switch can be used to study the effect of off-loading some of the burden on the OLT switch. In contrast the threshold on the number of requests forwarded by the servers, balances the load of requests to be forwarded among servers and therefore reduces the average queuing delay experienced by requests.

Figure 8.4 shows the power consumption of the PON cell versus varying thresholds on the number of requests forwarded by servers while allowing the OLT to be fully utilised. Figure 8.5 shows the power consumption of the PON cell versus varying thresholds on the number of requests forwarded by the OLT while allowing servers to be fully utilised. Setting no thresholds on the number of requests forwarded by the servers and the OLT switch resulted in the minimum power consumption for the different traffic scenarios. By setting no threshold on the forwarding by the OLT, the model is given the freedom to favour routing the inter rack traffic through the OLT switch which consumes less power compared to routing through servers. Setting no forwarding thresholds on the servers allows the MILP model to efficiently utilise active servers before activating idle servers.

The results in Figure 8.4 examined thresholds of 17%, 18%, 19%, and 30% on the servers' utilisation.

The 20%-80% intra rack and inter rack traffic scenario has the highest power consumption as most of traffic is routed through servers and/or the OLT switch. The intra rack traffic will always be routed through the passive optical backplane without traversing intermediate nodes.

Setting thresholds on the servers utilisation has not increased the power consumption of the PON cell under the traffic scenarios (80%-20%) and (50%-50%) for the different applied forwarding thresholds as a high proportion of the traffic will be routed within the same rack through the passive optical backplane. However; with the 20%-80% traffic flow scenario, setting thresholds of 17% and 18% on servers has increased the power consumption by 12% and 6%, respectively compared to the case with no threshold. Limiting the number of forwarded requests by servers increases the power consumption as some idle servers need to be switched on to meet the threshold constraint. For the thresholds 19% - 30% and no threshold no major difference in the power consumption is observed.

With a server forwarding threshold of 17%, the PON cell under the 20%-80% traffic scenario as shown in Figure 8.4, consumes 13% and 15% more power than the 50%-50% and 80%-20%, respectively.

Figure 8.5 shows that varying the forwarding threshold on the OLT has slightly increased the PON cell power consumption under the different traffic

scenarios. For the high inter rack traffic flow scenario (20%-80%) a threshold of 7.5% on OLT has increased the power consumption by 1% compared to the case with no threshold. As a result, introducing a threshold on OLT has no major impact on the PON cell power consumption and can off-load the burden on the switch as it serves a high number of PON cells.

Figures 8.6 and 8.7 present the average queuing delay experienced by requests versus varying thresholds on the number of requests forwarded by the OLT switch and servers for the different intra rack and inter rack traffic percentages.

Figure 8.6 shows the average queuing delay of the PON cell versus varying thresholds on the number of requests forwarded by servers while allowing the OLT switch to be fully utilised. The 20%-80% intra rack and inter rack traffic scenario has the highest queuing delay as most of the traffic is routed through servers and/or the OLT switch. With a server forwarding threshold of 17%, the PON cell under the 20%-80% traffic scenario results in 17% and 50% increase in queuing delay compared to the 50%-50% and 80%-20%, respectively.

For the 20%-80% intra rack and inter rack traffic scenario a threshold of 17% and 18% on servers has decreased the average queuing delay by 14% and 8.5%, respectively compared to the case with no threshold. On the other hand, varying the forwarding threshold for the 50%-50% and 80%-20% has no significant impact on the delay as a high proportion of the traffic will be routed within the same rack through the passive optical backplane.

Figure 8.7 shows the average queuing delay of the PON cell versus varying thresholds on the number of requests forwarded by the OLT while allowing servers to be fully utilised. For the high inter rack traffic flow scenario (20%-80%) a threshold of 7.5% on the OLT forwarding utilisation has increased the average queuing delay by 13% compared to the case with no threshold. However, decreasing the OLT threshold will not necessarily increase the servers queuing delay as the increased load on servers can be reduced by activating more server nodes and therefore the average queuing delay is reduced by decreasing the threshold on servers as seen with forwarding threshold cases with 17% and 18%.

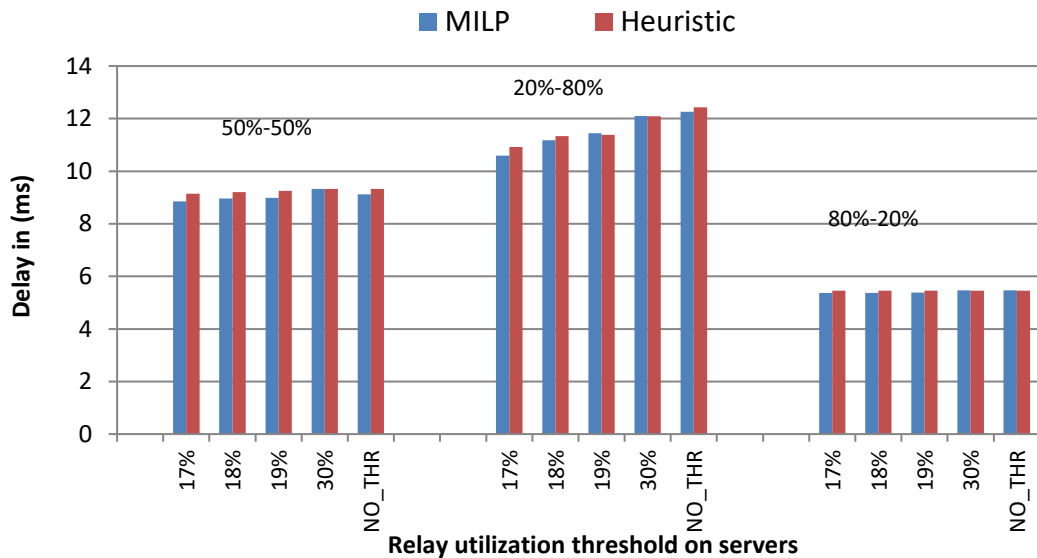


Figure 8.6. Average path delay for cases where no threshold for OLT are enforced and server's threshold is varied for the different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack

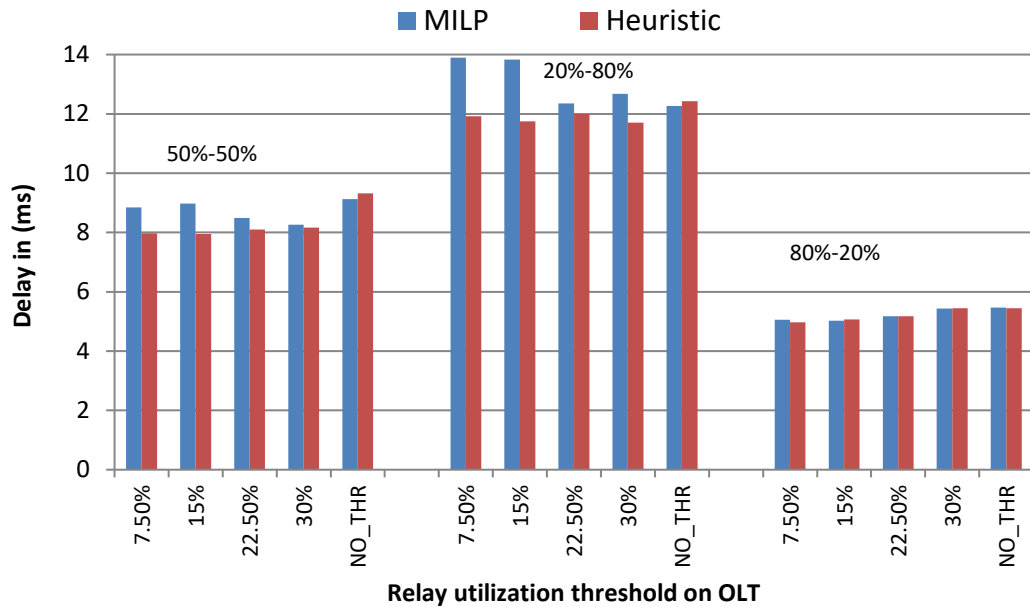


Figure 8.7. Average path delay for cases where no threshold for servers are enforced and OLT threshold is varied for different traffic flows where for e.g. 20%-80% of traffic refer to 20% intra rack and 80% inter rack

8.7 Energy Aware Routing Heuristic (EAR) for Server Centric PON Data Centre Architecture

For real time implementation of the energy aware routing approach in the proposed PON architecture we developed a heuristic that mimics the behaviour of the MILP model. The heuristic relies on a central approach. Every server with a request to be sent to another server either within the same PON cell or in another PON cell has to send a control message to a central management entity in the OLT switch requesting a connection to the destination server. This central entity has global knowledge of the data centre network in terms of links and servers utilisation in the different PON cells. The flow chart of the proposed heuristic is shown in Figure 8.8.

The incoming control messages from a PON cell are queued to be served by the central management entity at the OLT switch. The heuristic starts by creating a virtual topology of the OLT switch and active servers in the PON cell. Each link in the PON cell is assigned a cost value to distinguish between the different nodes, favouring the OLT node as it consumes less power for processing forwarded requests.

The first control message in the queue is retrieved and the associated request is assigned the minimum cost path given that the threshold on the OLT utilisation is not exceeded. If any of the servers on the minimum cost path exceed their utilisation threshold, the heuristic tries to replace it with another active server in its group that has not reached its threshold. If no active servers are available, an idle server within the group can be activated. The request is blocked if all servers within the group cannot serve it and there are no alternative routes. After assigning paths to all the incoming requests, the total power consumption and average queuing delay are computed based on the equations presented in Section 8.5.

The heuristic results are based on one run. Note that the maximum difference between MILP and the heuristic as shown in Figure 8.4 for the 20%-80% traffic is about 500W which corresponds to only 7%. Noting that the heuristic results are based on one run and the approach used in designing the algorithm is based on greedy algorithm. The heuristic has achieved power consumption and delay levels approaching those of the MILP model under the different thresholds as seen in Figures 8.4, 8.5, 8.6, and 8.7.

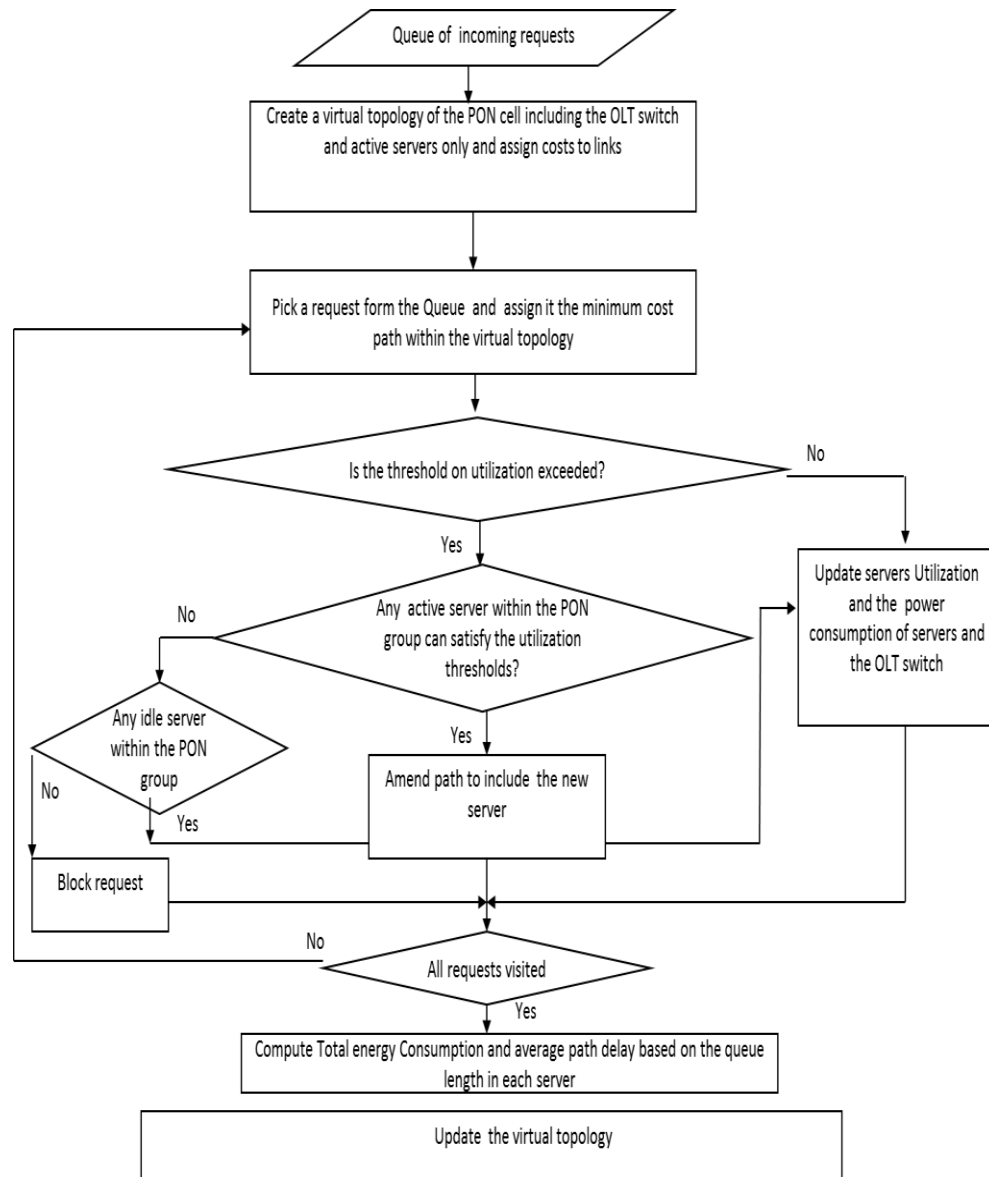


Figure 8.8. Heuristic flow chart for Energy Minimised Routing (EAR-PON) for a PON cell

8.8 Summary

In this chapter a novel scalable, high capacity, low cost, energy efficient server-centric PON data centre architecture was presented. A power consumption benchmarking study of the proposed architecture with the most

implemented 3-tier data centre architecture was reported and results have shown that power saving up to 69% can be achieved. A MILP model along with a heuristic were developed for energy efficient routing in the server-centric PON data centre architecture examining different inter/intra ratios of traffic flows. The selection of routing paths and relay servers is optimised to achieve optimum power savings while maintaining acceptable performance. The results have also shown a trade off between delay and power consumption when the servers forwarding thresholds are varied especially in the case of high inter rack flows (20%-80%). Limiting the number of forwarded requests by servers increases the power consumption as some idle servers need to be switched on to meet the threshold constraint. On the other hand, limiting the number of forwarded requests on servers reduces delay as more servers are involved in the routing.

9 Energy efficient routing with virtual machine placement in server- centric PON data centre

9.1 Introduction

This chapter is a further study on the server-centric PON design to develop a MILP mathematical optimisation model along with an algorithm for energy efficient resources provisioning for cloud applications. For optimum energy savings, the MILP model along with a resource provisioning greedy algorithm attempt to optimise the selection of hosting servers, routing paths and relay servers to achieve efficient resource utilisation.

9.2 MILP model for Energy Aware Routing and VM placement in Server-centric PON data centre design

In addition to the parameters and variable defined in Chapter 8, the following variables are defined:

Parameters:

V Set of requests for processing and memory

S Set of servers

OS OLT switch

M Large positive number

C_j	CPU capacity of server j
M_j	Memory capacity (RAM) of server j
ρ_i	CPU requirements of Request i
m_i	Memory requirements of Request i

Variables:

$Tdest_j$	Total traffic destined to server j
ϕR_j	CPU capacity of server j utilised for relaying requests
ϕP_j	CPU capacity of server j utilised for processing requests
ϕM_j	RAM capacity of server j utilised for requests
$\forall M_{ij}$	Fraction of memory resources of server j assigned for request i
$\forall P_{ij}$	Fraction of processing resources of server j assigned for request i
P_{ij}	Defined as $P_{ij} = 1$ if request i processing requirements are served by server j , otherwise $P_{ij} = 0$
M_{ij}	Defined as $M_{ij} = 1$ if request i memory requirements are served by server j , otherwise $M_{ij} = 0$
η_{nj}^{sd}	Number of requests from node s to node d arriving at node j from neighbouring node n
$NRfwd_j$	Number of requests not served by server j , hence forwarded to other servers

The model is defined as follows:

Objective:

We examine two objective functions in (9.1) and in (9.2) which result in power minimization and VM mapping maximization, respectively.

Minimise:

$$\sum_{j \in S} (PO_j) \delta_j + (\phi R_j + \phi P_j)(PM_j - PO_j) \quad (9.1)$$

Equation (9.1) gives the model objective function 1 which is to minimise the total power consumption of servers by optimising the servers selected to route and provision resources to VMs within the PON cell.

Maximise:

$$\sum_{j \in S} \sum_{i \in V} P_{ij} \quad (9.2)$$

Equation (9.2) gives objective function 2 which is to maximise the total mapping of VMs with servers. This objective function aims to serve VMs with the required processing and RAM resources without energy saving consideration.

Constraints (8.19) and (8.20) introduced in Chapter 8 apply to the resource provisioning model. The following additional constraints are introduced:

$$NRfwd_j = \left(\sum_{s \in OS} \sum_{d \in S} \sum_{n \in N_j} \eta_{nj}^{sd} \right) - \sum_{i \in V} P_{ij} \quad (9.3)$$

$$\forall j \in S$$

Constraint (9.3) calculates the number requests rerouted by a server as the difference between the incoming traffic and the traffic destined to the server.

$$\sum_{j \in S} C_j \quad \forall P_{ij} = \rho_i \quad (9.4)$$

$$\forall i \in V$$

Constraint (9.4) ensures that the processing requirements of all VMs are met.

$$\phi P_j = \sum_{i \in V} \nabla P_{ij} \quad (9.5)$$

$$\forall j \in S$$

Constraint (9.5) computes the CPU capacity of server j utilised for processing requests

$$\sum_{j \in S} M_j \quad \forall M_{ij} = m_i \quad (9.6)$$

$$\forall i \in V$$

Constraint (9.6) ensures that the memory requirements of all VMs are met.

$$\phi M_j = \sum_{i \in V} \nabla M_{ij} \quad (9.7)$$

$$\forall j \in S$$

Constraint (9.7) computes RAM capacity of server j utilised for serving requests

$$\sum_{j \in S} P_{ij} = 1 \quad (9.8)$$

$$\forall i \in V$$

Constraint (9.8) limits the number of servers that can be used to serve a VM to one server.

$$M_{ij} = P_{ij} \quad (9.9)$$

$$\forall i \in V, \forall j \in S$$

Constraint (9.9) ensures that the processing and memory requirements for a client are assigned in the same machine to reduce communication overhead.

$$(\phi R_j + \phi P_j) \leq 1 \quad (9.10)$$

$$\forall j \in S$$

Constraint (9.10) ensures that the processing and relaying load on each server does not exceed its processing capacity.

$$\phi M_j \leq 1 \quad (9.11)$$

$$\forall j \in S$$

Constraints (9.11) ensure that the memory requirements of VMs served by a server do not exceed the RAM capacity of servers.

$$\forall P_{ij} \leq P_{ij} \quad (9.12)$$

$$\forall i \in V, \forall j \in S$$

$$M \forall P_{ij} \geq P_{ij} \quad (9.13)$$

$$\forall i \in V, \forall j \in S$$

Constraints (9.12) and (9.13) relate $\forall P_{ij}$ to its binary equivalent P_{ij} .

$$\begin{aligned} \forall M_{ij} &\leq M_{ij} \\ \forall i \in V, \forall j \in S \end{aligned} \tag{9.14}$$

$$\begin{aligned} M \forall M_{ij} &\geq M_{ij} \\ \forall i \in V, \forall j \in S \end{aligned} \tag{9.15}$$

Constraints (9.14) and (9.15) relate $\forall M_{ij}$ to its binary equivalent M_{ij}

$$M (\phi R_j + \phi P_j) \geq \delta_j \tag{9.16}$$

$$(\phi R_j + \phi P_j) \leq \delta_j \tag{9.17}$$

$$\forall j \in S$$

Constraints (9.16) and (9.17) ensure that servers with zero utilization are switched off.

9.3 Results and discussions

This section evaluates the power consumption and servers' utilization resulting from relaying and serving VMs in the PON cell depicted in Figure 9.1 when optimising the routing and resource provisioning using the model introduced Section 9.2. The modelled PON cell consists of three racks each hosting 6 servers divided into 3 groups each of 2 servers. In addition to the input parameters defined in Section 8.6, additional parameters are defined in Table 9.1 for the modelled network.

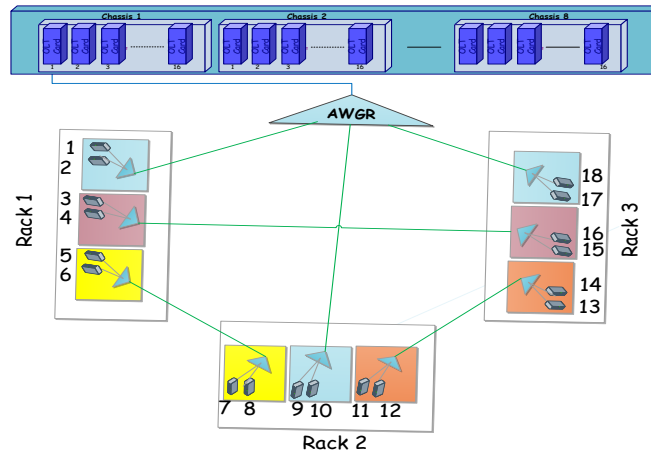


Figure 9.1. The modelled server centric PON data centre architecture

Table 9.1. Input data for the model

Server's utilisation for processing a relayed request (γfwd_j)	5%
Clients processing requirements	500-2000 M CPU cycles
Clients memory requirements	500-2000 MB
Server's processing capacity (C_j)	2500 MHz
Server's memory (RAM) (M_j)	8 GB

As discussed, in the proposed design, servers are not only used to store and process data but they can also participate in traffic forwarding and act as relay nodes. For resource provisioning in the proposed design, the selection of host servers is of premium importance as the scheduler attempts not only to slice servers' resources to be shared by multiple of VMs but also to reduce the number of servers relaying requests as much as possible to allow switching them off for maximum energy saving.

For comparison purposes and to evaluate the behaviour of the MILP model in minimising the power consumption, two objectives are evaluated as described in Section 9.2; one for energy minimisation (EA) and one for only provisioning all VMs without targeting energy saving (NEA).

The power consumption of the EA and NEA models are shown in Figure 9.2 for serving different numbers of VMs. The EA model achieves significant power savings ranging between 9% for the 50 VMs and 59% for 20 VMs. These savings are attributed to the model selection of host servers.

Figures 9.3, 9.4, 9.5, and 9.6 give a detailed overview of total utilisation of servers resources resulting from forwarding and processing requests under the different numbers of VMs. For 20 VMs (Figure 9.3) relaying requests through intermediate servers is avoided as the required processing resources and memory can be sustained within the gateway servers (servers 1,2,9,10,17 and 18 with direct connection to the OLT switch). As the number of VMs increases as in Figures 9.4, 9.5, and 9.6, gateway servers become highly exploited as relays. The model selects at most one gateway server to forward the request to the assigned host server. It can also be seen that an activated server is fully utilised before activating another one.

In NEA, the model randomly selects host servers and paths for all VMs which is reflected in the high number of activated underutilised servers to relay traffic as seen in Figure 9.7. The random selection of hosts along with the multiple intermediate servers resulted in underutilisation of servers' resources and therefore high power consumption.

Figures 9.8 and 9.9 show the number of activated servers and the average utilisation of servers for the different sets of examined VMs. The EA model reduced the number of activated servers by 66% while improving their average utilisation by 200% compared to the NEA model.

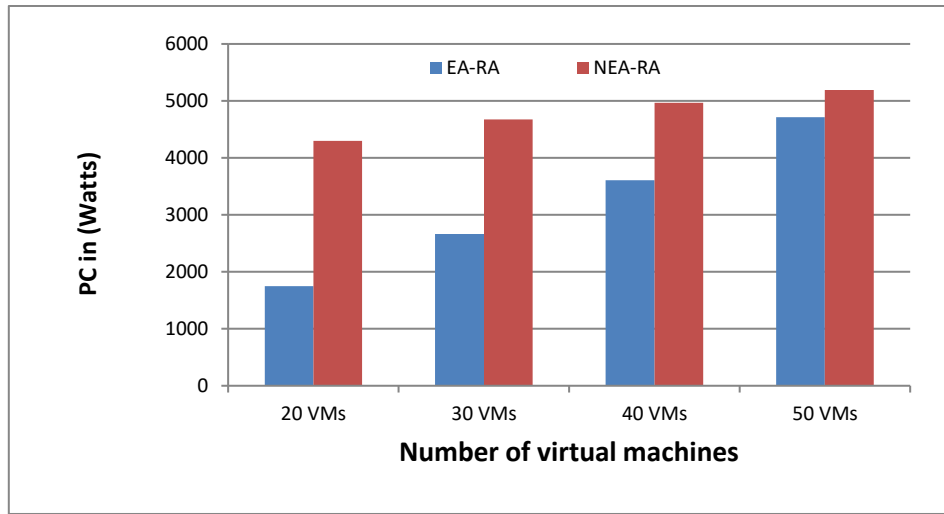


Figure 9.2. PON cell power consumption for different sizes of received clients' requests for the two objectives cases of energy aware (EA) and non-energy aware (NEA) of VMs routing and placement

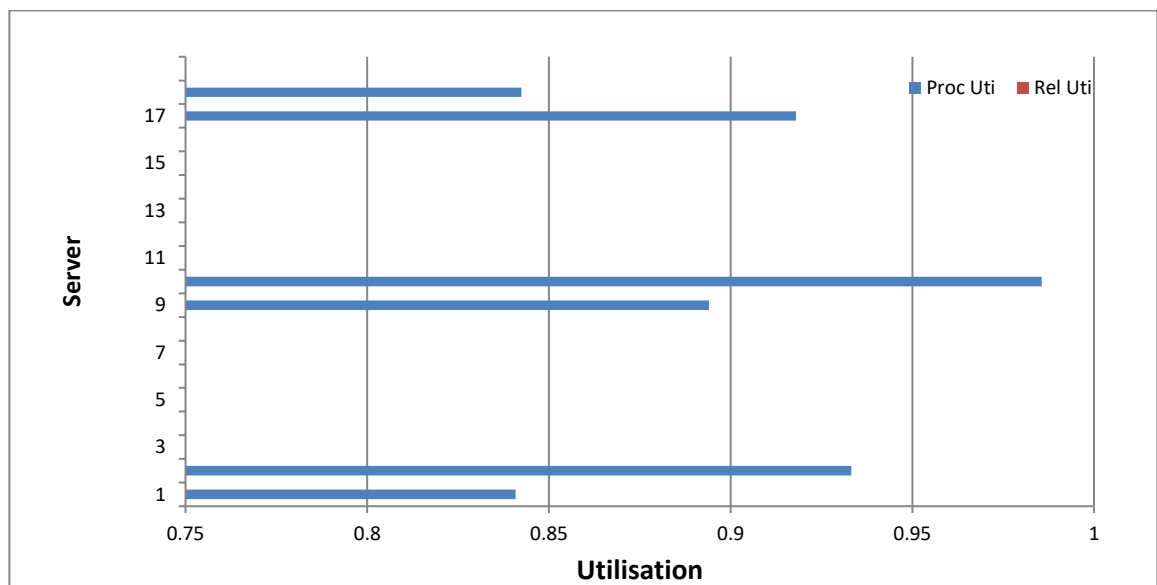


Figure 9.3. Servers' utilisation showing processing and relay utilisation for 20 VMs for the EA objective

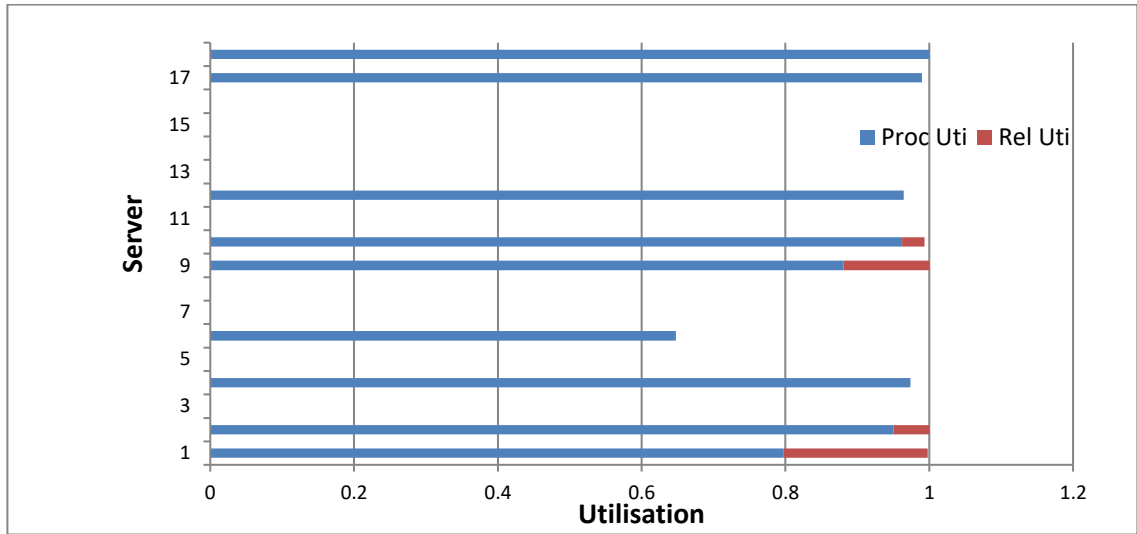


Figure 9.4. Servers' utilisation showing processing and relay utilisation for 30 VMs for the EA objective

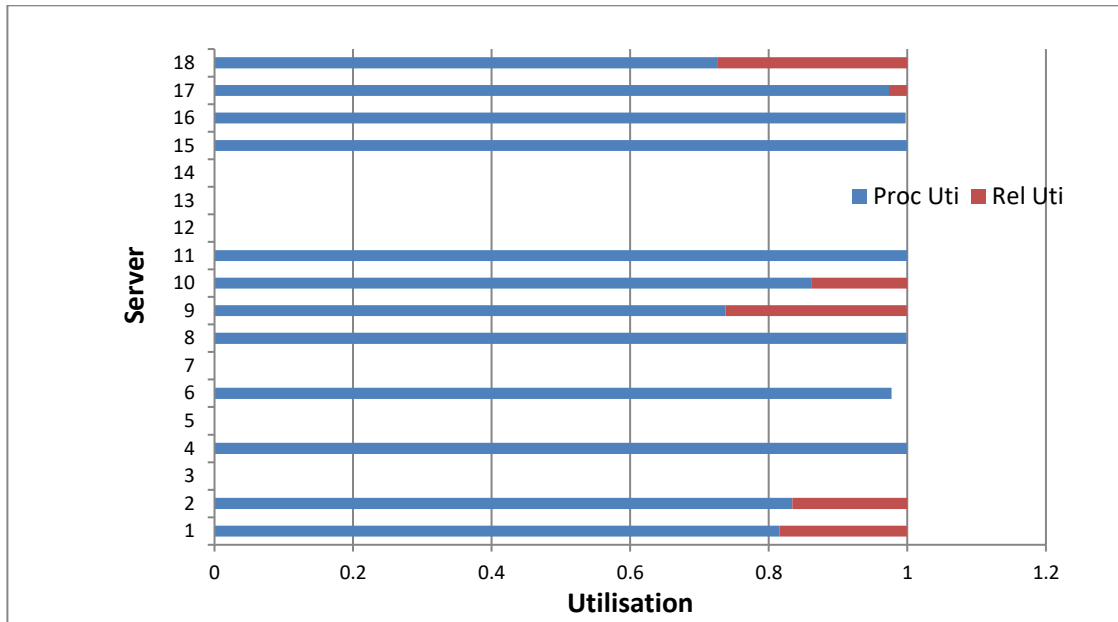


Figure 9.5. Servers' utilisation showing processing and relay utilisation for 40 VMs for the EA objective

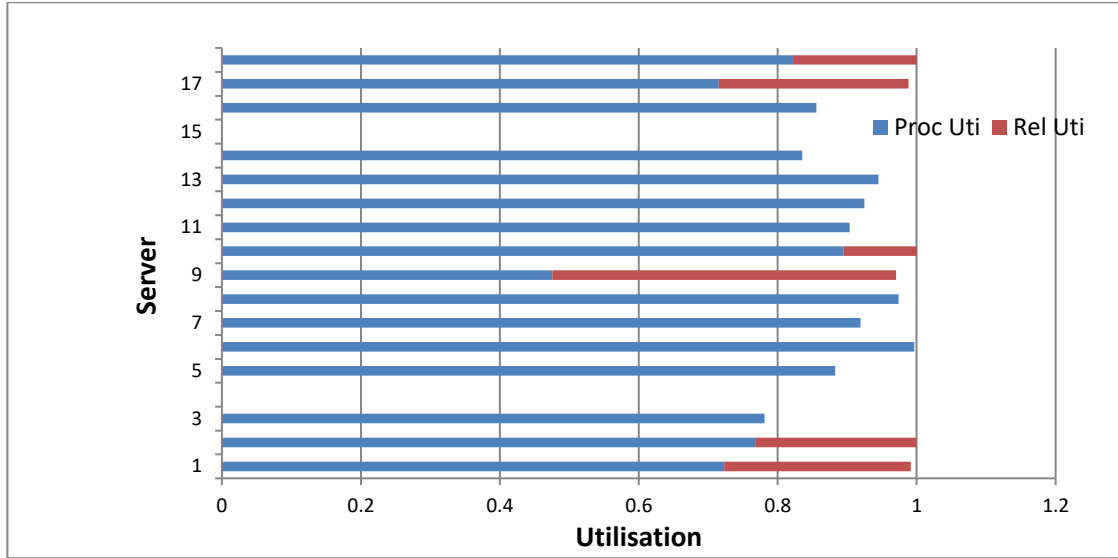


Figure 9.6. Servers' utilisation showing processing and relay utilisation for 50 VMs for the EA objective

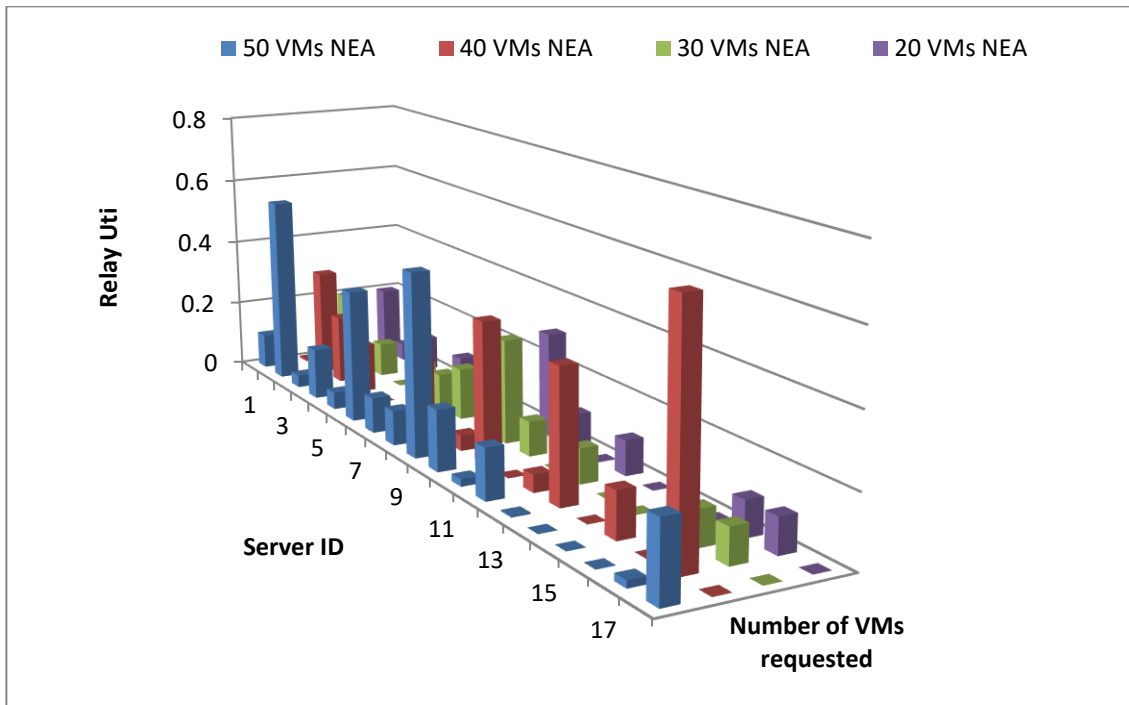


Figure 9.7. Location and relay utilisation of servers selected for routing request for the non-energy aware objective (random placement)

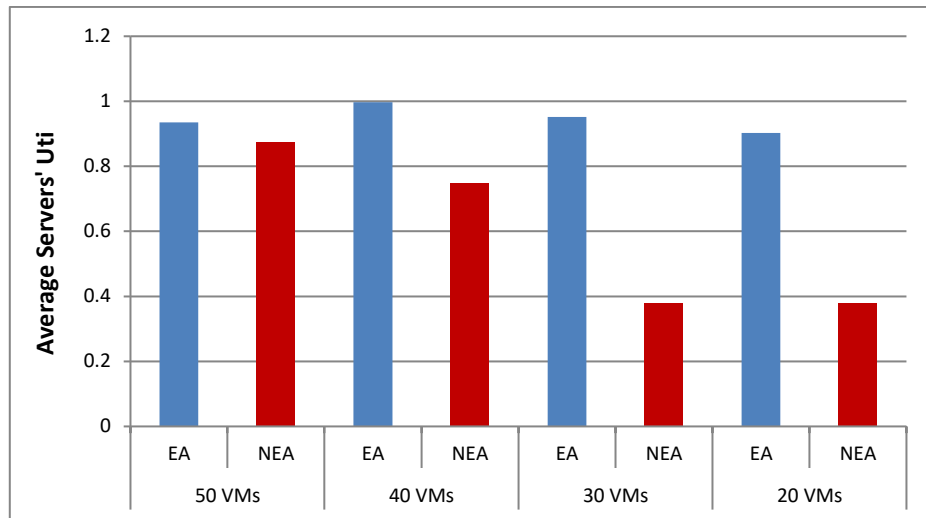


Figure 9.8. Average servers' utilization for different sizes of received clients' requests for the two objectives cases of energy aware (EA) and non-energy aware (NEA) of VMs routing and placement

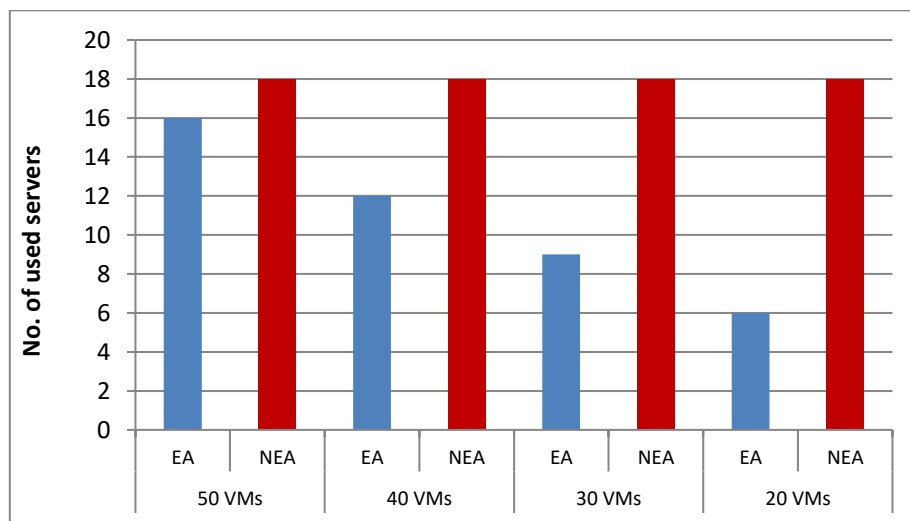


Figure 9.9. Number of selected servers for the different cases of received requests for the two objectives cases of energy aware (EA) and non-energy aware (NEA) of VMs routing and placement

9.4 Energy Aware VM Placement in PON Data Centre Heuristic

For real time implementation of the energy efficient routing and resource provisioning approach that mimics the behaviour of the MILP in host server selection, we proposed a heuristic to be run by a centralised scheduler that receives requests from clients and provisions them the required processing and memory resources.

The input parameters to the heuristic are: the VM requests CPU and memory requirement, servers' resources capacities, and network topology. The scheduler initially computes the Lower Bound (LB) for number of servers needed to serve the VM requests. Two values are calculated; one for memory LB_M given in equation (9.18) and one for computational process LB_P , given in equation (9.19). The higher of LB_M and LB_P is considered as the LB .

$$LB_P = \frac{\sum_{i \in V} \rho_i}{CP} \quad (9.18)$$

$$LB_M = \frac{\sum_{i \in V} m_i}{CM} \quad (9.19)$$

where CP and CM are the CPU and Memory capacity of the server, respectively assuming all servers are of the same capacity.

For LB values less than the total number of gateway servers in the architecture, the scheduler assigns resources for VMs by means of consolidation in gateway servers to avoid routing through intermediate servers and achieves minimum power consumption. The placement is based on Best Fit Decreasing Bin Packing (BFD-BP). In BFD-BP the CPU and memory resources of a server are sliced and shared among multiple of VMs and the minimum number of servers with sufficient resources are selected to serve a

group of VMs. A queue of VM requests is generated by the scheduler in a non-decreasing order with respect to the size of resources required by the VM. The placement of VMs is carried out in best fit manner where one of the servers with minimum remaining sufficient capacity is always selected to host the VM.

If the minimum number of servers required to host the VMs (LB) is higher than the number of gateway servers, the scheduler needs to decide in which rack to place the VM and whether to place the VM in a gateway server or in a neighbouring server of a gateway server. The scheduler attempts to efficiently utilise the servers' resources by efficient selection of host and gateway servers. We developed a greedy algorithm referred to as a Modified BFD-BP to solve this specific energy aware placement problem. The flow chart of the algorithm is shown in Figure 9.10.

The algorithm starts by sorting the VMs in decreasing order based on CPU requirements. Then the first VM in the queue is retrieved and a search is carried out to find a server with minimum remaining capacity that satisfies the process and memory requirements of the VM. The non-gateway servers (NGS) are given higher priorities for selection as the lower bound value is higher than the number of gateway servers within the rack. Then, the scheduler identifies the location of the selected server within the rack to find if the server is a gateway server or not. If the selected server is not a gateway server, the scheduler searches for a gateway server within the same rack with sufficient resources to route the request. Then, the resources of the selected host and gateway servers are updated before the next VM is served. If no gateway server was

found to relay the request or relaying resources of gateways (RPC) within the rack is insufficient, a new rack is opened and a new lower bound is calculated to estimate the minimum number of servers needed to host the remaining queued VMs.

If the new lower bound value is less than the number of gateway servers within the rack, priority for assigning the remaining VMs will be given to the gateway servers, which will host the VMs in this case. If the lower bound is higher than the number of gateway servers, priority assignment starts with a non-gateway server. The process is repeated until all VMs are visited and mapped to servers.

Figure 9.11 shows that the heuristic has achieved power consumption levels approaching those of the MILP model for the different sets of examined VMs. A case where no sorting of VMs is examined and the results have shown that with a small number of VMs, for example 30VMs, better results can be achieved.

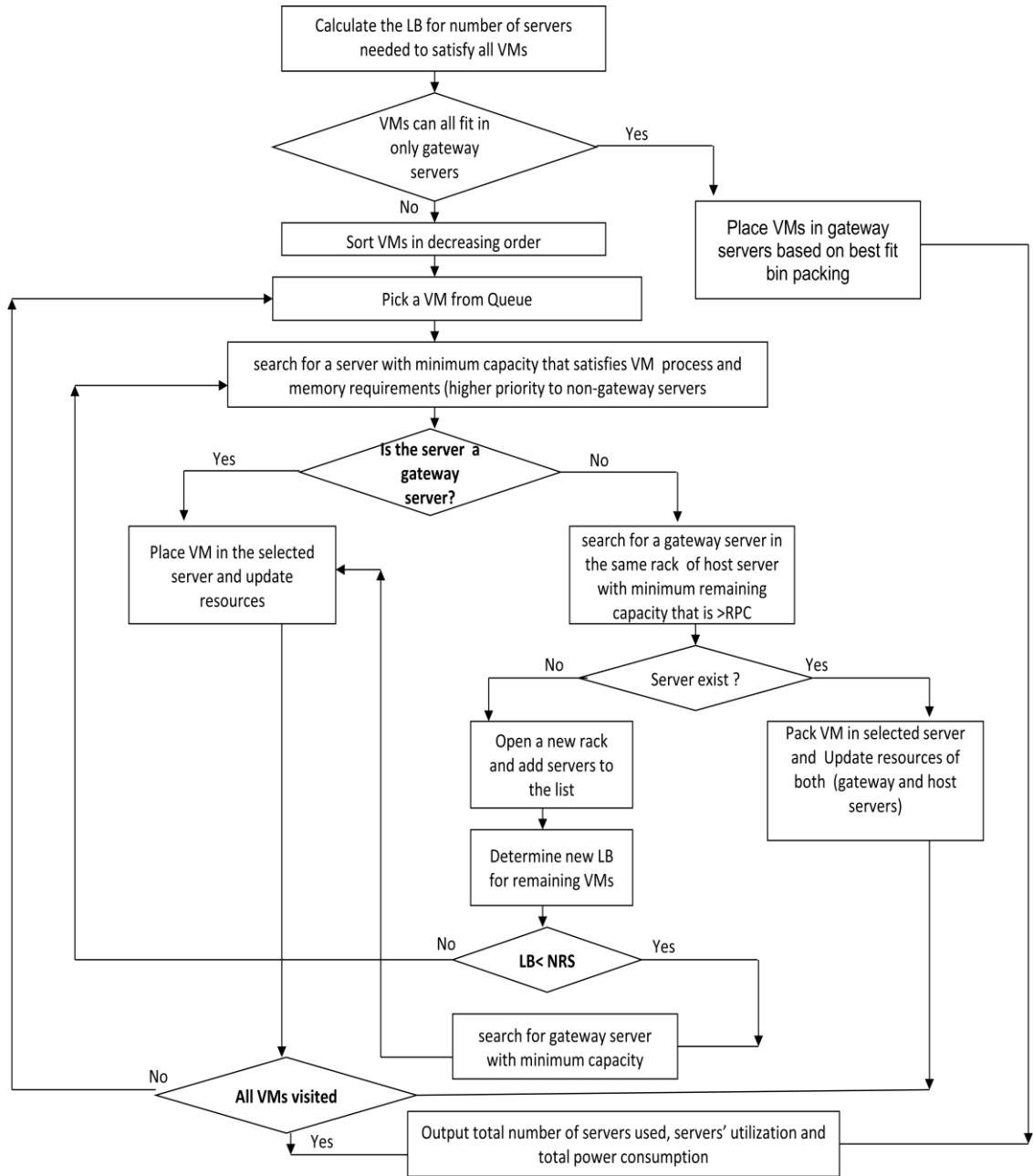


Figure 9.10. The flow chart for the Modified Best-Fit Decreasing (MBFD) algorithm for energy aware VM placement in a PON Cell

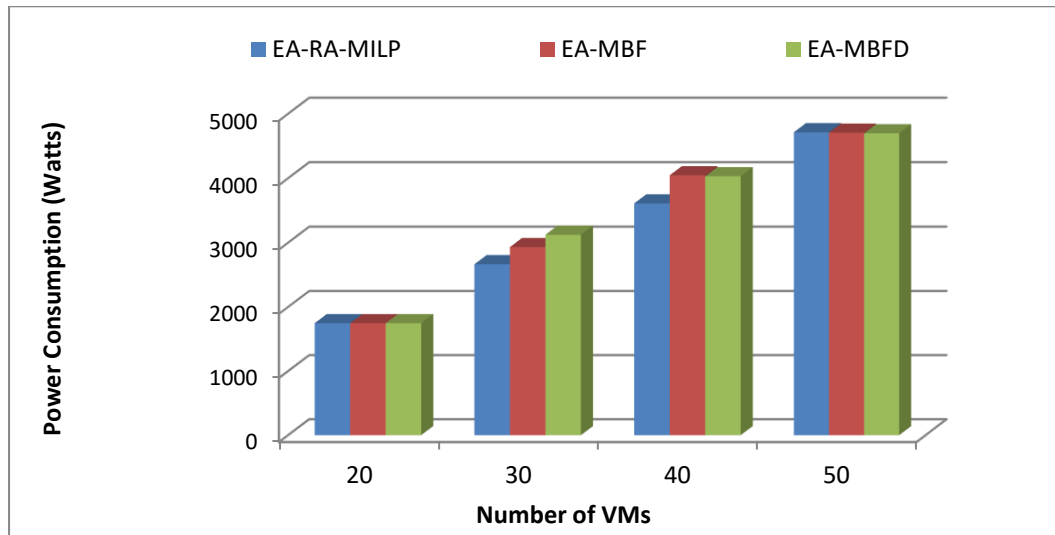


Figure 9.11. Power consumption results for MILP modelled network and MBFD and MBF algorithms

9.5 Summary

In this chapter, further investigation of the server-centric PON data centre architecture has been carried out to optimise the resource provisioning of VM requests in the architecture. A MILP model for energy efficient VM placement and routing has been developed. Significant power savings up to 59% are achieved by optimising the selection of host servers. In addition to the developed energy efficient MILP optimisation model, an algorithm was developed to mimic the behaviour of the MILP for real time implementation. The algorithm results in savings similar to those obtained from the MILP optimisation model.

10 Conclusions and future directions

This chapter summarises the main contributions for the work described in this thesis. Furthermore, it states the main conclusions and highlights directions for future research.

10.1 Conclusions

The architecture of a DCN directly reflects on its scalability, cost, fault tolerance and power consumption. DCNs are continuing to evolve and considerable research efforts by academia and industry are being devoted to address the various challenges observed. Chapters 2 reviewed different architecture designs proposed and implemented for data centre interconnections fabrics. These designs grew to overcome many challenges that appeared in the conventional design. Power consumption is one of the main concern as it has a great impact on global warming in the first place and on the electricity bill of data centres in the second place. This chapter provided a detailed survey on the most recent advances in DCNs with a special emphasis on the architectures and energy efficiency in DCNs [37].

Chapter 3 focused on PON deployment in FTTx access networks. The chapter covered PON technologies, architectures, classifications, standards, and implemented protocols. The objective was to understand PONs with their proven performance in access networks to provide energy efficient, high

capacity, low cost, scalable, and highly elastic solutions to help adapt PON architectures to support connectivity inside modern data centres.

Chapter 4 proposed five energy-efficient novel designs for PON implementation in data centres. Different solutions and technologies to manage intra/inter rack communications are proposed using mostly passive optical devices such as FBGs, optical backplanes, star reflectors, passive couplers/splitters, and AWGRs. A qualitative comparison for the proposed designs was given to determine the advantages and disadvantages of each design when compared to other designs [97].

Chapter 5 investigated the AWGR PON based design in detail to further study its architecture, interconnection topology, wavelength routing and assignment. A mathematical model was developed to optimise the routing and wavelength assignment of inter rack communication through intermediate AWGRs. The per server rate was taken into consideration and the design has shown its capability in providing 5 Gb/s rate. It has been shown that by introducing PONs in the design of data centres, the power consumption can typically be reduced by 45% and 80% compared to Fat-tree and BCube architectures, respectively. Similarly, the proposed design has also shown CAPEX savings up to 40% and 76% compared to the Fat-Tree and BCube architectures, respectively [11].

In Chapter 6, the oversubscription issue for the inter-cell communication in the AWGR PON based design was investigated. A solution was provided to reduce oversubscription and provision multipath routing through a centralised SDN. A benchmarking study that compared the proposed SDN architecture against the decentralised design was presented. The results showed that that in the SDN enabled architecture, the power consumption can be decreased by up to 90% for average data rates while maintaining zero blocking [98].

Chapter 7 further investigated the AWGR PON based design for cloud applications. The chapter studied optimisation of the resource provisioning for delay sensitive and non-delay sensitive applications, to cater for different applications that can be hosted in a PON cloud data centre. A mathematical optimisation model was developed for resource provisioning considering the minimisation of power consumption, delay, and both. The results have shown the trade-off between minimisation of power consumption and minimisation of delay objectives. The results have shown that delay can be decreased by 62% for delay-sensitive applications and power consumption can be decreased by 22% for non-delay sensitive applications [99]. A real time energy efficient resources provisioning greedy algorithm is also developed. The results revealed good agreement between the MILP model and the described greedy algorithm. The proposed algorithm was compared to other algorithms like random provisioning and BFD-BP. Our algorithm demonstrated an average of 47% savings in power consumption with more efficient PM resources utilisation

reaching (90%) when compared with the random algorithm. The proposed algorithm has also shown a maximum of 50% and 47% reduction in delay when compared with random and BFD algorithms respectively.

In Chapter 8, a detailed description of the server-centric PON design was presented. A power consumption benchmarking study of the proposed architecture of the server centric PON against the most implemented 3-tier data centre architecture was presented and the results have shown that a maximum saving of 69% can be achieved [100]. A mathematical optimisation model was developed along with a heuristic for energy efficient routing examining different inter/intra ratios of traffic flows. This work evaluated the power consumption of the PON cell and the average queuing delay experienced by requests versus varying thresholds on the number of requests forwarded by the OLT switch and servers for the different intra rack and inter rack traffic percentages. The results have shown that limiting the number of forwarded requests by servers increases the power consumption as some idle servers need to be switched on to meet the threshold constraint. On the other hand, limiting the number of forwarded requests per server reduces delay as more servers are involved in the routing.

Chapter 9 further investigated the Server Centric PON architecture for cloud applications. The work presented a mathematical optimisation model along with an algorithm for energy efficient routing and resources provisioning for cloud applications in the server centric PON data centre taking into

consideration the physical constraints of servers' and communication links resources. The MILP model achieved efficient resource utilisation reaching 95% and optimum saving in energy consumption reaching 59% saving. The algorithm results have shown similar results of those obtained from the MILP optimisation model.

10.2 Future work

The topic of energy-efficient data centre design is a hot research topic with several areas to be investigated. The energy efficient PON based architectures, energy efficient routing heuristics, and energy efficient VM placement algorithms for cloud applications proposed and demonstrated along with results in this thesis motivate the investigation of further issues in PON data centres.

The work presented in this thesis has been limited to mathematical models and computer simulations. However, experimental demonstration is another method to validate and verify the results obtained from the MILP mathematical models and the developed heuristics and simulators.

An important aspect to examine in order to improve the PON design and further reduce oversubscription is to consider the coexistence of hybrid optical technologies. In addition to OLT switches, MEMS and/or Semiconductor Optical Switches (SOA) optical switches can be considered to allow for flow

classification. Classification of the inter-cell rack to rack flows by assigning paths along with resources based on the flow size and duration can result in significant efficiency gains. For mice flows, the communicating servers' resources are assigned and grouped to join the same OLT port/switch through efficient grooming. While for elephant flows such as the case where a server needs to use a full wavelength, OLT traffic forwarding can be avoided and a circuit through the optical switch can be established between communicating entities in the different PON cells.

The implementation of network coding in our PON data centre designs is an interesting research topic to investigate. Recent work in wireless networks has shown the potential of network coding in reducing the energy consumption in such networks. In wireless networks, network coding demonstrated enhancement in the throughput and hence further reduction in power consumption. It is worth investigating the influence of introducing network coding at the OLTs and ONUs in our proposed PON data centre architectures to study its impact on power consumption and delay.

Another interesting topic is to consider disaggregated data centre design to replace the servers based architecture with CPU racks, memory racks, I/O cards racks and storage racks for maximum efficiency in provisioning resources such as memory, processing, communications and storage. PON based networking for such implementation will not only improve the resources usage efficiency but can also further reduce power consumption. However delay will

be a challenge as the distance between processors and memory in the disaggregated design will increase. The increase in the distance impacts the communication speed. The choice of the interconnection design is of premium importance as it impacts on the overall efficiency of the disaggregated data centre design. Therefore photonic technology along with racks interconnection fabric design have to be taken into careful consideration to not sacrifice performance when considering disaggregated data centre design. This is a hot research topic as substantial improvement in current designs is required.

References

- [1] Z. T. Al-Azez, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Virtualization framework for energy efficient IoT networks," in *Cloud Networking (CloudNet), 2015 IEEE 4th International Conference on*, 2015, pp. 74-77.
- [2] H. M. M. Ali, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient disaggregated servers for future data centers," in *Networks and Optical Communications - (NOC), 2015 20th European Conference on*, 2015, pp. 1-6.
- [3] A. M. Al-Salim, A. Q. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "Energy Efficient Tapered Data Networks for Big Data processing in IP/WDM networks," in *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 2015, pp. 1-5.
- [4] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Green IP over WDM Networks: Solar and Wind Renewable Sources and Data Centres," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, 2011, pp. 1-6.
- [5] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Energy-efficient IP over WDM networks with data centres," in *2011 13th International Conference on Transparent Optical Networks*, 2011, pp. 1-8.
- [6] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Use of renewable energy in an IP over WDM network with data centres," *IET Optoelectronics*, vol. 6, pp. 155-164, 2012.
- [7] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Green IP Over WDM Networks With Data Centers," *Journal of Lightwave Technology*, vol. 29, pp. 1861-1880, 2011.
- [8] X. Dong, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Joint optimization of power, electricity cost and delay in IP over WDM networks," in *2013 IEEE International Conference on Communications (ICC)*, 2013, pp. 2370-2375.
- [9] X. Dong, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "On the Energy Efficiency of Physical Topology Design for IP Over WDM Networks," *Journal of Lightwave Technology*, vol. 30, pp. 1931-1942, 2012.
- [10] J. M. H. Elmirghani, T. Klein, K. Hinton, T. e. h. El-Gorashi, A. Q. Lawey, and X. Dong, "GreenTouch GreenMeter core network power consumption models and results," in *Green Communications (OnlineGreencomm), 2014 IEEE Online Conference on*, 2014, pp. 1-8.
- [11] A. Hammadi, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "High performance AWGR PONs in data centre networks," in *Transparent Optical Networks (ICTON), 2015 17th International Conference on*, 2015, pp. 1-5.
- [12] A. Q. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "Energy-efficient peer selection mechanism for BitTorrent content distribution," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, 2012, pp. 1562-1567.

- [13] A. Q. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient cloud content delivery in core networks," in *2013 IEEE Globecom Workshops (GC Wkshps)*, 2013, pp. 420-426.
- [14] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Renewable energy in distributed energy efficient content delivery clouds," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 128-134.
- [15] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Distributed Energy Efficient Clouds Over Core Networks," *Journal of Lightwave Technology*, vol. 32, pp. 1261-1281, 2014.
- [16] M. O. I. Musa, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient core networks using network coding," in *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 2015, pp. 1-4.
- [17] Z. H. Nasralla, T. E. H. El-Gorashi, M. O. I. Musa, and J. M. H. Elmirghani, "Energy-Efficient Traffic Scheduling in IP over WDM Networks," in *Next Generation Mobile Applications, Services and Technologies, 2015 9th International Conference on*, 2015, pp. 161-164.
- [18] L. Nonde, T. E. H. Elgorashi, and J. M. H. Elmirghani, "Cloud Virtual Network Embedding: Profit, Power and Acceptance," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1-6.
- [19] L. Nonde, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy Efficient Virtual Network Embedding for Cloud Networks," *Journal of Lightwave Technology*, vol. 33, pp. 1828-1849, 2015.
- [20] N. I. Osman, T. El-Gorashi, and J. M. H. Elmirghani, "Reduction of energy consumption of Video-on-Demand services using cache size optimization," in *2011 Eighth International Conference on Wireless and Optical Communications Networks*, 2011, pp. 1-5.
- [21] N. I. Osman, T. El-Gorashi, L. Krug, and J. M. H. Elmirghani, "Energy-Efficient Future High-Definition TV," *Journal of Lightwave Technology*, vol. 32, pp. 2364-2381, 2014.
- [22] D. Xiaowen, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "On the Energy Efficiency of Physical Topology Design for IP Over WDM Networks," *Lightwave Technology, Journal of*, vol. 30, pp. 1931-1942, 2012.
- [23] "B. R., Report to congress on server and data center energy efficiency public law 109-431, Tech. rep. (2007).".
- [24] Z. Yan and N. Ansari, "On Architecture Design, Congestion Notification, TCP Incast and Power Consumption in Data Centers," *Communications Surveys & Tutorials, IEEE*, vol. 15, pp. 39-64, 2013.
- [25] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *SIGCOMM Comput. Commun. Rev.*, vol. 39, pp. 68-73, 2008.
- [26] Cisco.
https://www.cisco.com/application/pdf/en/us/guest/netso/ns107/c649/cc_migration_09186a008073377d.pdf.

- [27] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," presented at the Proceedings of the ACM SIGCOMM 2008 conference on Data communication, Seattle, WA, USA, 2008.
- [28] R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, *et al.*, "PortLand: a scalable fault-tolerant layer 2 data center network fabric," *SIGCOMM Comput. Commun. Rev.*, vol. 39, pp. 39-50, 2009.
- [29] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "Dcell: a scalable and fault-tolerant network structure for data centers," presented at the Proceedings of the ACM SIGCOMM 2008 conference on Data communication, Seattle, WA, USA, 2008.
- [30] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, *et al.*, "BCube: a high performance, server-centric network architecture for modular data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 39, pp. 63-74, 2009.
- [31] L. Dan, G. Chuanxiong, W. Haitao, K. Tan, Z. Yongguang, and L. Songwu, "FiConn: Using Backup Port for Server Interconnection in Data Centers," in *INFOCOM 2009, IEEE*, 2009, pp. 2276-2285.
- [32] H. J. Chao, D. Kung-Li, and Z. Jing, "PetaStar: a petabit photonic packet switch," *Selected Areas in Communications, IEEE Journal on*, vol. 21, pp. 1096-1112, 2003.
- [33] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, *et al.*, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 40, pp. 339-350, 2010.
- [34] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, *et al.*, "VL2: a scalable and flexible data center network," *SIGCOMM Comput. Commun. Rev.*, vol. 39, pp. 51-62, 2009.
- [35] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, *et al.*, "c-Through: part-time optics in data centers," presented at the Proceedings of the ACM SIGCOMM 2010 conference, New Delhi, India, 2010.
- [36] J. C. Palais, *Fiber optic communications*: Prentice Hall, 1988.
- [37] A. Hammadi and L. Mhamdi, "Review: A survey on architectures and energy efficiency in Data Center Networks," *Comput. Commun.*, vol. 40, pp. 1-21, 2014.
- [38] D. Kliazovich, P. Bouvry, and S. U. Khan, "GreenCloud: a packet-level simulator of energy-aware cloud computing data centers," *The Journal of Supercomputing*, vol. 62, pp. 1263-1283, 2012.
- [39] "Revolutonizing Network Design Flattening the Data Center Network with the QFabric Architecture. ."
- [40] "Vision and Roadmap: Routing Telecom and Data Centers Toward Efficient Energy Use. Vision and Roadmap Workshop on Routing Telecom and Data Centers, 2009.."

- [41] C. Kachris and I. Tomkos, "A Survey on Optical Interconnects for Data Centers," *Communications Surveys & Tutorials, IEEE*, vol. 14, pp. 1021-1036, 2012.
- [42] J. Edmonds and R. M. Karp, "Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems," *J. ACM*, vol. 19, pp. 248-264, 1972.
- [43] C. Kachris and I. Tomkos, "Power consumption evaluation of hybrid WDM PON networks for data centers," in *Networks and Optical Communications (NOC), 2011 16th European Conference on*, 2011, pp. 118-121.
- [44] "M. Y. K. Xia, Y.-H. Kaob, H. J. Chao, Petabit optical switch for data center networks, Tech. rep., Polytechnic Institute of NYU (2010).".
- [45] Y. Xiaohui, P. Mejia, Y. Yawei, R. Proietti, S. J. B. Yoo, and V. Akella, "DOS - A scalable optical switch for datacenters," in *Architectures for Networking and Communications Systems (ANCS), 2010 ACM/IEEE Symposium on*, 2010, pp. 1-12.
- [46] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," presented at the Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, Monterey, California, 2010.
- [47] R. Luijten, W. E. Denzel, R. R. Grzybowski, and R. Hemenway, "Optical interconnection networks: The OSMOSIS project," in *Lasers and Electro-Optics Society, 2004. LEOS 2004. The 17th Annual Meeting of the IEEE*, 2004, pp. 563-564 Vol.2.
- [48] O. Liboiron-Ladouceur, P. G. Raponi, N. Andriolli, I. Cerutti, M. S. Hai, and P. Castoldi, "A Scalable Space-Time Multi-plane Optical Interconnection Network Using Energy-Efficient Enabling Technologies [Invited]," *Journal of Optical Communications and Networking*, vol. 3, pp. A1-A11, 2011/08/01 2011.
- [49] A. K. Kodi and A. Louri, "Energy-Efficient and Bandwidth-Reconfigurable Photonic Networks for High-Performance Computing (HPC) Systems," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, pp. 384-395, 2011.
- [50] J. Gripp, J. E. Simsarian, J. D. LeGrange, P. Bernasconi, and D. T. Neilson, "Photonic terabit routers: The IRIS project," in *Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC)*, 2010, pp. 1-3.
- [51] O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, *et al.*, "The Data Vortex Optical Packet Switched Interconnection Network," *Journal of Lightwave Technology*, vol. 26, pp. 1777-1789, 2008.
- [52] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," presented at the Proceedings of the 10th ACM

- SIGCOMM conference on Internet measurement, Melbourne, Australia, 2010.
- [53] T. S. G. t. D. C. E. E. D. Chernicoff, Realtime Publisher, Newyork, 2009. Available: <http://nexus.realtimepublishers.com/sgdcee.php?ref=gbooks>
- [54] L. Jie, Z. Feng, L. Xue, and H. Wenbo, "Challenges Towards Elastic Power Management in Internet Data Centers," in *Distributed Computing Systems Workshops, 2009. ICDCS Workshops '09. 29th IEEE International Conference on*, 2009, pp. 65-72.
- [55] A. Beloglazov and R. Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers," in *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*, 2010, pp. 826-831.
- [56] J. Shuja, S. Madani, K. Bilal, K. Hayat, S. Khan, and S. Sarwar, "Energy-efficient data centers," *Computing*, vol. 94, pp. 973-994, 2012/12/01 2012.
- [57] C. Guo, G. Lu, H. J. Wang, S. Yang, C. Kong, P. Sun, *et al.*, "SecondNet: a data center network virtualization architecture with bandwidth guarantees," presented at the Proceedings of the 6th International COnference, Philadelphia, Pennsylvania, 2010.
- [58] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, *et al.*, "ElasticTree: saving energy in data center networks," presented at the Proceedings of the 7th USENIX conference on Networked systems design and implementation, San Jose, California, 2010.
- [59] Y. Shang, D. Li, and M. Xu, "Energy-aware routing in data center network," presented at the Proceedings of the first ACM SIGCOMM workshop on Green networking, New Delhi, India, 2010.
- [60] D. K.-M. Dave Hood – Ericsson, Frank Effenberger - Huawei, Dan Parsons – BroadLight, Eli Elmoalem - BroadLight, "ONT Power Saving Proposal," ITU Q2/SG15, Stockholm, June 2008.
- [61] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, "Reducing network energy consumption via sleeping and rate-adaptation," presented at the Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, San Francisco, California, 2008.
- [62] M. Gupta and S. Singh, "Greening of the internet," presented at the Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications, Karlsruhe, Germany, 2003.
- [63] D. Kliazovich, P. Bouvry, and S. U. Khan, "DENS: Data Center Energy-Efficient Network-Aware Scheduling," in *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*, 2010, pp. 69-75.

- [64] Available:
<http://www.morganclaypool.com/doi/pdf/10.2200/s00193ed1v01y200905cac006>
- [65] J. Baliga, R. W. A. Ayre, W. V. Sorin, K. Hinton, and R. Tucker, "Energy Consumption in Access Networks," in *Optical Fiber communication/National Fiber Optic Engineers Conference, 2008. OFC/NFOEC 2008. Conference on*, 2008, pp. 1-3.
- [66] S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*: Prentice Hall: Englewood Cliffs, NJ, 2000.
- [67] "http://www.itu.int/dms_pub/itu-t/oth/0B/04/T0B040000382C01PDFE.pdf."
- [68] Cisco,
["http://www.cisco.com/web/HR/expo08/pdf/Thomas_Martin_Fiber_To_The_Home.pdf"](http://www.cisco.com/web/HR/expo08/pdf/Thomas_Martin_Fiber_To_The_Home.pdf), 2008.
- [69] G. Keiser, *FTTX concepts and applications* vol. 91: John Wiley & Sons, 2006.
- [70] G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT a dynamic protocol for an Ethernet PON (EPON)," *Communications Magazine, IEEE*, vol. 40, pp. 74-80, 2002.
- [71] C. M. Assi, Y. Ye, S. Dixit, and M. A. Ali, "Dynamic bandwidth allocation for quality-of-service over Ethernet PONs," *Selected Areas in Communications, IEEE Journal on*, vol. 21, pp. 1467-1477, 2003.
- [72] M. McGarry, M. Reisslein, and M. Maier, "Ethernet passive optical network architectures and dynamic bandwidth allocation algorithms," *Communications Surveys & Tutorials, IEEE*, vol. 10, pp. 46-60, 2008.
- [73] J. Zheng and H. T. Mouftah, "A survey of dynamic bandwidth allocation algorithms for Ethernet Passive Optical Networks," *Optical Switching and Networking*, vol. 6, pp. 151-162, 2009.
- [74] D. J. Shin, D. K. Jung, H. S. Shin, J. W. Kwon, S. Hwang, Y. Oh, *et al.*, "Hybrid WDM/TDM-PON with wavelength-selection-free transmitters," *Journal of lightwave technology*, vol. 23, p. 187, 2005.
- [75] A. Vahdat, H. Liu, Z. Xiaoxue, and C. Johnson, "The emerging optical data center," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, 2011, pp. 1-3.
- [76] L. Yuanqiu, F. Effenberger, and S. Meng, "Cloud computing provisioning over Passive Optical Networks," in *Communications in China (ICCC), 2012 1st IEEE International Conference on*, 2012, pp. 255-259.
- [77] P. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos, "Design and evaluation of a flexible-bandwidth OFDM-based intra data center interconnect," 2012.
- [78] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 267-280.

- [79] J. Beals IV, N. Bamiedakis, A. Wonfor, R. Penty, I. White, J. DeGroot Jr, *et al.*, "A terabit capacity passive polymer optical backplane based on a novel meshed waveguide architecture," *Applied Physics A*, vol. 95, pp. 983-988, 2009.
- [80] R. Fourer, D. Gay, and B. Kernighan, *AMPL* vol. 117: Boyd & Fraser Danvers, MA, 1993.
- [81] I. I. CPLEX, "V12. 1: User's Manual for CPLEX," *International Business Machines Corporation*, vol. 46, p. 157, 2009.
- [82] "Cisco-2960-24TC-L DataSheet."
- [83] "Cisco-2960-48TC-L DataSheet."
- [84] Gyarmati and T. A. Trinh, "How can architecture help to reduce energy consumption in data center networking?," presented at the Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking, Passau, Germany, 2010.
- [85] "Cisco-2960-8TC-L DataSheet."
- [86] K. Grobe, M. Roppelt, A. Autenrieth, J. P. Elbers, and M. Eiselt, "Cost and energy consumption analysis of advanced WDM-PONs," *Communications Magazine, IEEE*, vol. 49, pp. s25-s32, 2011.
- [87] K. R. C. Bhagat, R. Shetye and A. Vaity, "Technological and cost-based comparison of next generation PON technologies: 10GPON and WDM PON," in *A capstone paper submitted as partial fulfillment of the requirements for degree of Masters in Interdisciplinary Telecommunications at the University of Colorado, Boulder, University of Colorado, USA, 2012.*
- [88] "Intel PRO/1000 PT Server Adapter DataSheet,[Online]. Available: <http://ark.intel.com/products/50497/Intel-PRO1000-PT-Server-Adapter..>"
- [89] [Online] Available <http://www.3anetwork.com/>.
- [90] "Cisco: Data sheet of Cisco-ME 4600 Series Optical Line Terminal DataSheet,[Online].Available : <http://www.cisco.com/c/en/us/products/collateral/switches/me-4600-series-multiservice-optical-access-platform/datasheet-c78-730445.pdf>."
- [91] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: measurements & analysis," presented at the Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, Chicago, Illinois, USA, 2009.
- [92] "Cisco Nexus 2148T Fabric Extender DataSheet."
- [93] "Cisco Nexus 5020 Switch DataSheet."
- [94] "Cisco Nexus 7000 F2-Series 48-Port 1 and 10 Gigabit Ethernet Module Data Sheet."
- [95] "PAS740x GPON ONT SoC DataSheet."
- [96] B. Mukherjee, "Energy Savings in Telecom Networks," *Tutorial: Brazilian Symposium on Networks and Distributed Systems (SBRC) 2011, Campo Grande, Brazil.*

- [97] J. Elmirghani, T. EL-GORASHI, and A. HAMMADI, "Passive optical-based data center networks," ed: Google Patents, 2016.
- [98] J. Elmirghani, T. EL-GORASHI, and A. HAMMADI, "Passive optical-based data center networks," ed: Patent Number WO2016083812 A1, 2016.
- [99] A. Hammadi, T. E. H. El-Gorashi, M. O. I. Musa, and J. M. H. Elmirghani, "Server-centric PON data center architecture," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, 2016, pp. 1-4.
- [100] A. Hammadi, M. Musa, T. E. H. 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, pp. 178-182.