UNIVERSITY OF LEEDS

Optimisation of cloud data centres and networking for big data applications

Sanaa Hamid Mohamed

University of Leeds

School of Electronic and Electrical Engineering

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

September, 2019

To the cloud computing, optical networking, and data centres research communities.

Intellectual Property Statement

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

The content in Chapters 2 and 3 has appeared or will potentially appear in publications as follows:

[1] S. H. Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "A Survey of Big Data Machine Learning Applications Optimization in Cloud Data Centres and Networks", *to be submitted to the IEEE Surveys and Tutorials.*

The candidate summarised and categorised a number of big data applications optimisation studies in terms of the performance and/or the energy efficiency and provided some research gaps and future directions.
Dr El-Gorashi reviewed the paper.
Prof. Elmirghani reviewed the paper and helped with the flow of content and suggested the review scope.

The work in Chapter 4 has appeared or will potentially appear in publications as follows:

[2] Ali Hammadi, S. H Mohamed, M. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Passive Optical Networks-based Data Centres for Fog and Cloud Computing", *to be submitted to IEEE Access.*

The candidate enhanced the modelling for a previously proposed data centre network design based on passive optical networks (PONs) technologies.
Dr Hammadi developed, optimised, and modelled five PON-based data centre designs and proposed heuristics to enhance workloads assignment.
Dr. Musa revised the initial model for the third design.
Dr El-Gorashi reviewed the designs and helped with the documentation.
Prof. Elmirghani is the originator of introducing PON technologies in data centre networks and he helped with the documentation.

The work in Chapter 5 has appeared or will potentially appear in publications as follows:

[3] S. H. Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "On the energy efficiency of MapReduce shuffling operations in data centers," *2017 19th International Conference on Transparent Optical Networks (ICTON), Girona, 2017, pp. 1-5.*

[4] S. H. Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Energy Efficiency of Server-Centric PON Data Center Architecture for Fog Computing," *2018 20th International Conference on Transparent Optical Networks (ICTON), Bucharest, 2018, pp. 1-4.*

[5] S. H. Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Optimizing Co-flows Scheduling and Routing in Data Centre Networks for Big Data Applications", *to be submitted to IEEE Transactions on Network and Service Management.*

The candidate initiated the optimisation of co-flows scheduling and routing of big data workloads in terms of the completion time and the energy efficiency in several state-of-the-art data centres in [3], additionally considered the fifth proposed PON-based data centre in [4], and enhanced the modelling of data centres, and considered the third design and other optical data centres in [5].

Dr El-Gorashi reviewed the papers and revised the models.

Prof. Elmirghani is the originator of the evaluation idea and reviewed the papers.

The work in Chapter 6 has appeared or will potentially appear in publications as follows:

[6] S. H. Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Impact of Link Failures on the Performance of MapReduce in Data Center Networks," *2018 20th International Conference on Transparent Optical Networks (ICTON), Bucharest, 2018, pp. 1-4.*

The candidate evaluated the performance of big data applications under links failure in different data centres in [6].

Dr El-Gorashi reviewed the papers.

Prof. Elmirghani proposed the evaluation and reviewed the papers.

The work in Chapter 7 has appeared or will potentially appear in publications as follows:

[7] M. B. Abdull Halim, S. Hamid Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Fog-Assisted Caching Employing Solar Renewable Energy for Delivering Video on Demand Service," *2019 21st International Conference on Transparent Optical Networks (ICTON), Angers, France, 2019, pp. 1-5.*

[8] M. B. A. Halim, S. H. Mohamed, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Fog-assisted Caching Employing Solar Renewable Energy and Energy Storage Devices for Delivering Video on Demand Service", *to be submitted to IEEE Access.*

The candidate developed the models and verified the results in [7] and [8].

A. Halim produced the results and wrote the initial draft of [7].

Dr El-Gorashi reviewed the papers.

Prof. Elmirghani proposed the study and reviewed the papers.

Acknowledgements

I would like to sincerely thank Prof. Elmirghani for all the opportunities he provided throughout different phases of pursuing my PhD degree. Not only I was awarded a prestigious scholarship based on his kind endorsement, but I also had great chances to attend local and abroad technical events. By joining his lab, I experienced both; a close access to well-established research and a space to explore new directions in trending topics. His remarkable supervision made the work in this thesis achievable. His encouragements and believing in my abilities when I doubted them are much appreciated.

I would like to thank my co-supervisor, Dr Elgorashi who ultimately supported me in sharpening my professional and research skills. On top of that, I am delighted she resembled a family for me in Leeds. I also thank Dr Iman, Dr Mamoun, Dr Moutaman, Dr Mohamed Hassan, Dr Elrefaie, Dr Landolsi, and Dr Al-Dweik for research experience prior to PhD.

I would like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC) and the University of Leeds for funding me through a Doctoral Training Grant Studentship and international fees waiver.

My appreciation is for my friends and colleagues in ICaPNet and previously I3S. I thank Anna and Louise for their kind assistance. I thank Dr Afzal for her inspirations to become a better researcher, Dr Samane, Dr Salim, Dr Howraa, Dr Zainab, Dr Quzweeni, Dr Nonde, Dr Nasralla, and Dr Hammadi for their guidance and Dr Haider for his help in ICTON19. I thank Hatem, Ida, Randa, and Barzan for their support and useful discussions during our PhD journey. I thank Amal, Fatemah, Opeyemi, Osama, Hadi, Abdullah, Azza, Mansourah, Kholoud, and specially Sarah for the precious times we had in the lab. Also, I thank Jenny for her kindness and morning greetings and treats.

My friends in Leeds complemented my experience and provided comfort and joy. I thank Gabrielė for our friendship that made it possible to cope and grow. My eternal gratitude is for my fiancé, Omer Mahdy for his

ultimate support. His generosity in reassuring and encouraging me is deeply appreciated. I am thankful to Dr Ayat for her sincere care, Razan for making me smile, Samah for her kindness, and Dr Sandra and Dr Salma for the amazing moments we had in Leeds and around.

My sincere gratefulness is for my parents for their everlasting love and prayers and their continues guidance and support. I cheerfully thank my siblings for recharging me during my breaks, being proud of me, always, and for making me proud of their achievements. I finally thank Dr Nassma, my beloved cousin, for always being there for me.

Sanaa Hamid Mohamed.

Abstract

Cloud data centre applications including big data applications and video content distribution services are rapidly evolving in terms of their data formats, frameworks, workloads, and data volumes which has resulted in a substantial increase in their storage and processing requirements. These applications and services usually utilise commodity clusters within geographically-distributed data centres. The increase in traffic between and within the data centres that migrate, store, and process big data and video content, is becoming a bottleneck that calls for enhanced infrastructures capable of reducing the congestion and power consumption. Moreover, it has resulted in the consideration of fog data centres in access networks in addition to the cloud data centres to reduce the delay and power consumption associated with delivering services. This thesis addresses the evaluation of the energy consumption and performance when deploying big data frameworks in intra data centre networks and the optimisation of delivering video demands through transport networks from cloud or fog data centres. First, two energy-efficient Passive Optical Network (PON)-based data centre designs which are a switch-centric design and a server-centric design are introduced. For big data applications, the first part of the work in this thesis focuses on the impact of data centre architectures on the performance, resilience, and energy efficiency of intra data centre networks. A time-slotted Mixed Integer Linear Programming (MILP) model that optimises the scheduling and routing of big data applications traffic was developed and utilised to compare the energy efficiency and the performance of four state-of-the-art data centres, in addition to the two PON-based data centres. Two different objectives are considered focusing on the minimisation of the completion time and the minimisation of the energy consumption. The comparisons are performed while considering a number of big data application-related parameters and data centre-related network architecture parameters. The results indicate that the topology has a significant impact on the perform-

ance, energy consumption, and the resilience of the data centre. Thus, investing in data centre designs is essential to meet future demands. For video-on-demand services, this thesis considers caching the content near the users in fog data centres to reduce the brown networking power consumption required to deliver the contents from cloud data centres. A MILP model is developed to optimse the delivery of content from cloud or fog data centres. The reduction in the brown power consumption is evaluated while considered several factors such as the core network design, the Power Usage Effectiveness (PUE) of the data centres, in addition to the use of renewable energy in the cloud data centres, and solar energy with Energy Storage Devices (ESDs) in the fog data centres.

Contents

List of Figures

List of Tables

Abbreviations

3R	Reamplification, Reshaping, and Retiming.	DCN	Data Centre Network.
AA	Application-specific Address.	DFS	Distributed File System.
AM	Application Master.	DOS	Data centre Optical Switch.
AMPL	Advanced Mathematical Programming Language.	DQPSK	Differential Quadrature Phase Shift Keying.
AWG	Arrayed Waveguide Grating.	DRAM	Dynamic Random Access Memory.
AWGR	Arrayed Waveguide Grating Router.	DSP	Digital Signal Processing.
AWS	Amazon Web Service.	DVFS	Dynamic Voltage and Frequency Scaling.
BAU	Business As Usual.	EC2	Elastic Compute Cloud.
BPON	Broadband Passive Optical Network.	ECMP	Equal Cost Multi Path.
BV-T	Bandwidth-Variable Transponder.	EDFA	Erbium-Doped Fiber Amplifier.
BV-WSS	Bandwidth-Variable Wavelength Selective Switch.	EEE	Energy Efficient Ethernet.
CAGR	Compound Annual Growth Rate.	EON	Elastic Optical Network.
CAPEX	Capital Expenditure.	EPON	Ethernet Passive Optical Network.
CD	Chromatic Dispersion.	EPS	Electronic Packet Switching.
CDC	Cloud Data Centre.	ESD	Energy Storage Device.
CDN	Content Delivery Network.	FCT	Flow Completion Time.
CMOS	Complementary Metal Oxide Semiconductor.	FDC	Fog Data Centre.
CO	Central Office.	FEC	Forward Error Correction.
CPU	Central Processing Unit.	FIFO	First-In First-Out.
CSP	Content Service Provider.	FPGA	Field-Programmable Gate Array.
DAC	Digital-to-Analogue Converter.	GbE	Gigabit Ethernet.

GFS	Google File System.	NIC	Network Interface Card.
GHG	Global Greenhouse Gas.	NM	Node Manager.
GPON	Gigabit-capable Passive Optical Network.	NSFNET	National Science Foundation Network.
HDFS	Hadoop Distributed File System.	OCS	Optical Circuit Switching.
HFS	Hadoop Fair Scheduler.	O/E/O	Optical-Electrical-Optical.
HTTP	HyperText Transfer Protocol.	OLT	Optical Line Terminal.
IaaS	Infrastructure-as-a-Service.	ONT	Optical Network Terminal.
IBGP	Internal Border Gateway Protocol.	ONU	Optical Network Unit.
ICN	Information-Centric Network.	O-OFDM	Optical Orthogonal Frequency Division Multiplexing.
I/O	Input/Output.	OPEX	Operational Expense.
IoT	Internet-of-Things.	OPS	Optical Packet Switching.
IP	Internet Protocol.	OSA	Optical Switching Architecture.
IPTV	Internet Protocol Television.	OSPF	Open Shortest Path First.
IR	Intermediate Results.	OXC	Optical Cross Connect.
ISP	Internet Service Provider.	P2MP	Point to Multi-Point.
ITU	International Telecommunication Union.	P2P	Point to Point.
JT	Job Tracker.	PUE	Power Usage Effectiveness.
JVM	Java Virtual Machine.	PaaS	Platform-as-a-Service.
LA	Location-specific Address.	QAM	Quadrature Amplitude Modulation.
LPI	Low Power Idle.	QoE	Quality-of-Experience.
MAC	Multiple Access Control.	QoS	Quality-of-Service.
MAN	Metropolitan Area Network.	RAID	Redundant Array of Independent Disks.
MCC	Mobile Cloud Computing.	RAM	Random Access Memory.
MCF	Multi Core Fiber.	RDD	Resilient Distributed Data sets.
MCF	Multi-Commodity Flow.	RM	Resource Manager.
MEMS	Micro-Electro-Mechanical System Switch.	ROADM	Reconfigurable Optical Add Drop Multiplexer.
MILP	Mixed Integer Linear Programming.	RTT	Round Trip Time.
MLR	Mixed Line Rate.	SaaS	Software-as-a-Service.
MMF	Multi Mode Fiber.	SDM	Space Division multiplexing.
MZI	Mach-Zehnder Interferometer.	SDN	Software Defined Networking.

SFP+	Enhanced Small Form Factor Pluggable.	TWC	Tuneable Wavelength Converter.
SIR	Shuffled Intermediate Results.	TWDM PON	Time Wavelength Division Multiplexing Passive Optical Network.
SLA	Service Level Agreement.	TT	Task Tracker.
SLR	Single Line Rate.	VLB	Valiant Load Balancing.
SMF	Single-Mode Fiber.	VM	Virtual Machine.
SOA	Semiconductor Optical Amplifier.	VoD	Vedio-on-Demand.
SPB	Shortest Path Bridging.	VNI	Visual Network Index.
SSD	Solid-State Drive.	VNE	Virtual Network Embedding.
STP	Spanning Tree Protocol.	WAN	Wide Area Network.
TCP	Transmission Control Protocol.	WC	Word Count.
TDM	Time Division Multiplexing.	WDM	Wavelength Division Multiplexing.
TMS	Traffic Matrix Scheduling.	WSS	Wavelength Selective Switch.
ToR	Top-of-Rack.	XG-PON	next-generation Passive Optical Network.
TRILL	Transparent Interconnection of Lots of Links.	YARN	Yet Another Resource Negotiator.

Chapter 1

Introduction

This chapter provides the motivation for optimising cloud data centres and networks for big data applications and content delivery services. The problem statement is presented and the tools and methodologies utilised are summarised. The thesis objectives and contributions are provided. Finally, the structure for the remainder of the thesis is provided.

1.1 Motivation

The evolving practice of big data is essential for critical advancements in data processing models and the underlying acquisition, transmission, and storage infrastructures [1–10]. Big data differs from traditional data in being potentially unstructured, rapidly generated, continuously changing, and massively produced typically by a large number of distributed users or devices. Typically, big data workloads are transferred into data centres containing sufficient storage and processing units for real-time or batch computations and analysis. A widely used characterisation for big data is the "5" notion which describes big data through its unique attributes of Volume, Velocity, Variety, Veracity, and Value [11]. In this notation, the volume refers to the vast amount of data produced which is usually measured in Exabytes (i.e. 2^{60} or 10^{18} bytes) or Zettabytes (i.e. 2^{70} or 10^{21} bytes), while the velocity reflects the high speed or rate of data generation and hence potentially the short lived useful lifetime of data. Variety indicates that big data can be composed of different types of data which can be categorised into structured and unstructured. The veracity measures the trustworthiness of the data as some generated portions could be erroneous or inaccurate, while the value measures the ability of the user or owner of the data to extract useful information from the data.

In 2020, the global data volume is predicted to be around 40,000 Exabytes which represents a 300 times growth factor compared to the global data volume in 2005 [10]. An estimate of the global data volume in 2010 is about 640 Exabytes [12], and in 2015 is about 2,700 Exabytes [13]. This huge growth in data volumes is the result of continuous developments in various applications that generate massive and rich content related to a wide range of human activities. For example, online business transactions are expected to have a rate of 450 Billion transactions per day by 2020 [13]. Social media such as Facebook, LinkedIn, and Twitter, which have between 300 Million and 2 Billion subscribers who access these social media platforms through web browsers in personal computers, or through applications installed in tablets and smart phones are enriching the content of the Internet with content in the range of several Terabytes (2^{40}

bytes) per day [14]. Analysing the thematic connections between the subscribers, for example by grouping people with similar interests, is opening remarkable opportunities for targeted marketing and e-commerce. Moreover, the subscribers' behaviours and preferences tracked by their activities, clickstreams, requests, and collected web log files can be analysed with big data mining tools for psychological, economical, business-oriented, and product improvement studies [15, 16].

To accelerate the delay-sensitive operations of web searching and indexing, distributed programming models for big data such as MapReduce were developed [17]. MapReduce is a powerful, reliable, and cost-effective programming model that performs parallel processing for large distributed datasets. These features have enabled the development of different distributed programming big data solutions and cloud computing applications. With the prevalence of mobile applications and services that have extensive computational and storage demands exceeding the capabilities of the current smart phones, emerging technologies such as Mobile Cloud Computing (MCC) were developed [18]. In MCC, the computational and storage demands of applications are outsourced to remote (or close as in mobile edge computing) powerful servers over the Internet. As a result, on-demand rich services such as video streaming, interactive video, and online gaming can be effectively delivered to the capacity and battery limited devices. Video content accounted for 51% of the total mobile data traffic in 2012 [18], and is predicted to account for 78% of an expected total volume of 49 Exabytes by 2021 [19]. Due to these huge demands, in addition to the large sizes of video files, big video data platforms are fronting several challenges related to video streaming, storage, and replication management, while needing to meet strict Quality-of-Experience (QoE) requirements [20].

In addition to mobile devices, the wide range of everyday physical objects that are increasingly interconnected for automated operations has formed what is known as the Internet-of-Things (IoT). In IoT systems, the underlying communication and networking infrastructures are typically integrated with big data computing systems for data collection, analysis, and decision-making. To process the big data generated by IoT

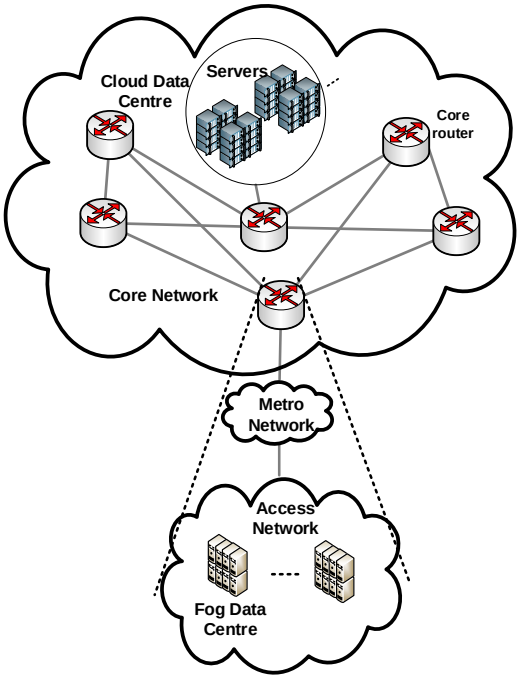


Figure 1.1: Cloud and fog data centres.

devices, different solutions such as cloud and fog computing were proposed [7, 21–36]. Cloud computing is a centralised solution to computing where a number of cloud data centers accessed through core networks are used, while fog computing is a decentralised solution where distributed fog data centres accessed through the access network are used. An illustration for cloud and fog data centres is provided in Figure 1.1. Existing cloud computing infrastructures could be utilised by aggregating and processing big data in powerful central data centres. Alternatively, data could be processed at the edge where fog computing units, typically with limited processing capacities compared to cloud, are utilised [37]. Fog computing reduces both the traffic in core networks and the latency by being closer to end devices. The connected devices could be sensors gathering different real-time measurements, or actuators performing automated control operations in industrial, agricultural, or smart building applications. IoT can support vehicle communication to realise smart transportation systems. IoT can also support

medical applications such as wearables and telecare applications for remote treatment, diagnosis and monitoring [38]. With this variety in IoT devices, the number of Internet-connected things is expected to exceed 50 Billion by 2020, and the services provided by IoT are expected to add $15 Trillion to the global gross domestic product in the next 20 years [39].

Achieving the full potential of big data requires a multidisciplinary collaboration between computer scientists, engineers, data scientists, as well as statisticians and other stakeholders [13]. One of the challenges of big data in enterprise and cloud infrastructures is the existence of various workloads and tenants with different Service Level Agreement (SLA) requirements that need to be hosted on the same set of clusters. An early solution to this challenge at the application level is to utilise a distributed file system to control the access and sharing of data within the clusters [40]. On the infrastructure level, solutions such as Virtual Machines (VMs) or Linux containers dedicated to each application or tenant were utilised to support the isolation between their assigned resources [10,41]. Big data systems are also challenged by security, privacy, and governance related concerns. However, addressing these challenges is out of the scope of the study presented in this thesis. As the increasing computational demands of the increasing data volumes are exceeding the capabilities of existing commodity infrastructures, future enhanced and energy efficient processing and networking infrastructure for big data have to be investigated and optimised.

1.2 Problem Statement

Typically, big data programming models and applications require extensive server-to-server communications to carry out their joint computations. As a result, the East-West data centre traffic (i.e. the traffic between the servers in a data centre) is currently becoming dominant over the South-North traffic (i.e. the traffic between the servers and the gateway router to and from the Internet) which was the focus in designing data centres in the past [10]. This bottleneck calls for enhanced data centre topologies

and technologies to cope with the evolving requirements of big data applications and emerging computing models. Moreover, as data volumes in big data applications are rapidly increasing beyond the capabilities of existing state-of-the-art infrastructures, a mix of both scale-up and scale-out designs are required. This study aims first to quantitatively assess the impact of different Data Centre Network (DCN) topologies on the performance of big data programming models such as MapReduce. The work is also performed under several typical failure scenarios in data centre environments including, link, switch, and server failures. This study considered the performance and resilience of several state-of-the-art data centre topologies, in addition to two recently introduced Passive Optical Network (PON)-based data centre topologies.

As the dominant inter cloud and cloud to user traffic is made up of video content delivered to users through access networks, this thesis also examines the reduction in brown power consumption of data centres and core networks when fog data centres are considered in the access network. The brown power can be defined as the power produced by non-renewable sources such as oil or coal that typically cause harmful Greenhouse Gas emission. These fog data centres can cache Video-on-Demand (VoD) contents and can then reduce the need for transporting the content through the backbone network. Compared to VoD caches that contain a number of streaming servers, a fog data centre can be considered as a larger cache containing higher number of servers that can be generic or specific to streaming. Several design factors that affect the power consumption at the cloud and fog layers are considered. Those include the Power Usage Effectiveness (PUE) of the data centres, the design of the backbone network, and the consideration of renewable sources for the cloud and fog data centres, in addition to Energy Storage Devices (ESDs) in the fog layer.

1.3 Tools and Methodologies

Understanding the complex characteristics of big data workloads is an essential step towards optimising the configurations for the frameworks used and identifying the sources

of bottlenecks in the underlying clusters. MapReduce and other big data frameworks are supported by several standard benchmarking suites such as GridMix, hibench, HcBench, and PUMA [1] which can be used in production environments for initial tuning, stress-testing, and debugging purposes. The workloads contained in these benchmarks are typically generated by a semantic that runs on previously collected or randomly generated datasets. Examples are text retrieval-based workloads, such as Word-Count (WC) and Sort, and web search-based workloads, such as Grep, Inverted Index, and Page Rank workloads. These workloads vary in being Input/Output (I/O), memory, or Central Processing Unit (CPU) intensive. For example, PageRank, Grep, and sort are I/O intensive, while WC, Page rank, and k-means are CPU intensive [17]. As the first part of the work in this thesis focuses on intra data centre networks, sort workloads were considered. Sort generates an alphabetically sorted output from input documents, hence generates final results with size equal to the input data. If performed through a distributed framework such as MapReduce, this allows testing the data centres when performing the required routing between the distributed data centre nodes (i.e. servers) to complete the job.

For the inter data centres traffic, consumer video traffic based on the Cisco Visual Network Index (VNI) forecast for 2020 was considered for the demands from the five cloud data centre locations to users in the The National Science Foundation Network (NSFNET) core network [42]. NSFNET is a topology for a core network covering 14 cities in the United States and is widely used in networking studies. The set of results and corresponding analysis presented in Chapters 4, 5, 6, and 7 of this Thesis were generated with the aid of the following tools and methodologies as summarised in Table 1.1.

1. Mixed Integer Linear Programming (MILP):

   MILP enables the modelling and optimisation of problems that can be described in a linear fashion. The optimisation can be carried out through algorithms such as the Simplex algorithm. The solution to a problem is presented as the optimum set of values for the variables that achieve the maximum or minimum value of a

defined objective function while meeting a number of constraints. The constraints reflect actual operational requirements of the modelled system. Solvers such as CPLEX can be used to solve MILP problems with up to few million variables. A brief tutorial for modelling networking problems through MILP is provided in Appendix A. The Advanced Mathematical Programming Language (AMPL) [43] was used to write the MILP codes and the associated data files.

2. MATLAB for workloads generation: To examine the data centres performance for MapReduce sort workloads when the data distribution is not uniform, uniform random number generation functions in MATLAB were used to generate and normalise the sizes of the flows to be routed in the data centre.

1.4 Thesis Objectives

The work reported in the thesis had the following objectives:

1. To model and evaluate the impact of choosing a given data centre topology on the performance of big data applications.

2. To evaluate the energy efficiency and resilience of different data centre network topologies.

3. To compare recent optical data centre topologies with current data centre topologies.

4. To optimise joint cloud-fog architectures in the presence of renewable energy and energy storage.

Table 1.1 provides a summary of the applications, tools and methodologies considered in addition to the workloads evaluated in this thesis.

Table 1.1: Summary of the content of the chapters and the workloads used.

Chapter	Applications	Intra DCN	Inter DCN	Tools and Methodologies	Workloads
4	-	✓		MILP model to optimise the connections and wavelength routing and assignment in the in the AWGR-based DCN design	-
5	MapReduce	✓		MILP Model to optimise the co-flows scheduling and routing for MapReduce traffic	Sort via MapReduce
6	MapReduce	✓		MILP Model to optimise the co-flows scheduling and routing with replication under server failure	Sort via MapReduce with replication
7	VoD services		✓	A MILP model to optimise the delivery of VoD contents from cloud data centres in the core network or fog data centres in the access network	Consumer video traffic based on Cisco Visual Network Index (VNI) forecast for 2020 [42]

1.5 Thesis Structure

Following Chapter 1, the remainder of this thesis is organised as follows: Chapter 2 provides a review of data centre applications including big data applications, cloud computing-based and content delivery services. Chapter 3 provides a review of cloud networking infrastructures and cloud data centres topologies, traffic characteristics, routing protocols, and energy efficiency. Chapter 4 introduces two PON-based DCN designs proposed for use in future cloud and fog data centres. Chapter 5 addresses the optimisation of the routing and scheduling in intra DCNs and compares the performance and energy efficiency of four state-of-the-art DCNs, in addition to the two proposed PON-based DCNs. Chapter 6 considers the resilience of DCNs for MapReduce workloads against link, switch, and server failures. Finally, Chapter 7 considers the optimisation of content delivery from cloud or fog data centres. The conclusions and future work are provided in Chapter 8.

Chapter 2

Data Centre Applications and Cloud-Based Services

This chapter reviews some of the programming models developed to provide parallel computation for big data including MapReduce. This chapter also reviews cloud computing services and how they are increasingly utilised for big data analytics, in addition to another class of data centre applications, namely content delivery, which contributes a high portion of the Internet traffic. The rest of this Chapter is organised as follows: Section 2.1 discusses MapReduce, Apache Hadoop and related applications, in addition to distributed in-memory big data applications, distributed graph processing, and big data stream processing applications. Section 2.2 describes cloud computing services and emphasises content delivery services.

2.1 Big Data Programming Models and Applications

2.1.1 The MapReduce Programming Model

The MapReduce programming model was introduced by Google in 2003 as a cost-effective solution for processing massive data sets. MapReduce utilises distributed computations in commodity clusters that run in two phases; *map* and *reduce* which are adopted from the Lisp functional programming language [17]. The MapReduce user defines the required functions of each phase by using a programming language such as C++, Java, or Python, and then submits the code as a single MapReduce job to process the data. The user also defines a set of parameters to configure the job. Each MapReduce job consists of a number of map and reduce tasks depending on the input data size and the configurations, respectively. Each map task is assigned to process a unique portion of the input data set, preferably available locally, and hence can run independently from other map tasks. The processing starts by transforming the input data into the key-value schema and applies it to the map function to compute another key-value pair known as the intermediate results. These results are then shuffled (i.e. transfered) to reduce tasks according to their keys where each reduce task is assigned to process intermediate results with a unique set of keys. Finally each reduce task generates a final output.

The internal operational details of MapReduce such as assigning the nodes within the cluster to map or reduce tasks, partitioning the input data, tasks scheduling, fault tolerance, and inter-machine communications are typically performed by the run-time system and are hidden from the users. The input and output data files are typically managed by a Distributed File System (DFS) that provides a unified view of the files and their details, and allows various modifications such as replication, read, and write operations. An example DFS is the Google File System (GFS) [44] which is a fault-tolerant, reliable, and scalable chunk-based distributed file system designed to support MapReduce in Google's commodity servers. The typical chunk size in GFS is 64 MB and each chunk is replicated in different nodes with a default value of replicating in

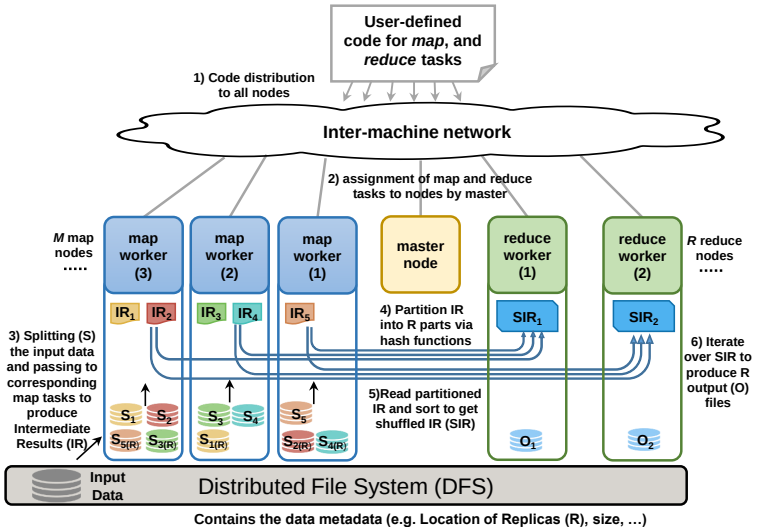


Figure 2.1: Google's MapReduce cluster components and the programming model implementation.

three nodes to support fault-tolerance.

The components of a typical MapReduce cluster and the implementation details of the MapReduce programming model are illustrated in Figure 2.1. One of the nodes in the cluster is set to be a *Master*, while the others are set to be *Workers* and are assigned to either a map or a reduce task by the master. Beside task assignments, the master is also responsible for monitoring the performance of the running tasks and for checking the statuses of the nodes within the cluster. The master also manages the information about the location of the running jobs, data replicas, and intermediate results. The detailed steps of implementing a MapReduce job are as follows [17]:

1. The MapReduce code is copied to all cluster nodes. In this code, the user defines the map and reduce functions and provides additional parameters such as input and output data types, names of the output files, and number of reduce workers.

2. The master assigns map and reduce tasks to available workers where typically, map workers are more than reduce workers and are assigned to several map tasks.

3. The input data, in the form of key-value pairs, is split into smaller partitions.

The splits (S) and their replicas (S_R) are distributed in the map workers local disks as illustrated in Figure 2.1. The splits are then processed concurrently in their assigned map workers according to their scheduling. Each map function produces intermediate results (IR) consisting of the intermediate key-value pairs. These results are then materialised (i.e. saved persistently) in the local disks of the map workers.

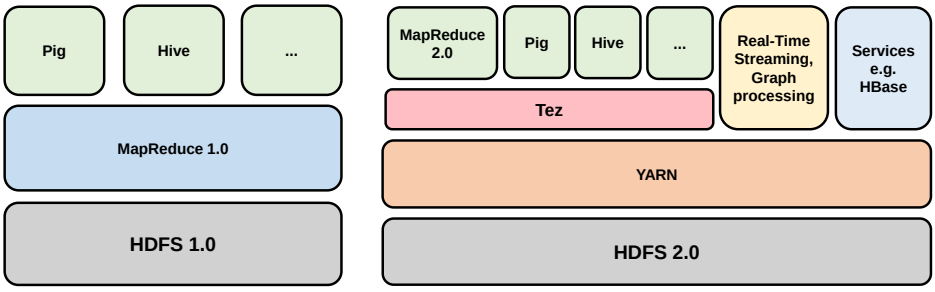
4. The intermediate results are divided into (R) parts to be processed by R reduce workers. Partitioning can be done through hash functions (e.g. hash(key) mod R) to ensure that each key is assigned to only one reduce worker. The locations of the hashed intermediate results and their file sizes are sent to the master node.

5. A reduce task is composed of shuffle, sort, and reduce phases. The shuffle phase can start when 5% of the map results are generated, however the last reduction phase cannot start unless all map tasks are completed. For shuffling, each reduce worker obtains the locations of the intermediate pairs with the keys assigned to it and fetches the corresponding results from the map workers local disks typically via the HyperText Transfer Protocol (HTTP).

6. Each reduce worker then sorts its intermediate results by the keys. The sorting is performed in the Random Access Memory (RAM) if the intermediate results can fit, otherwise, external sort-merge algorithms are used. The sorting groups all the occurrences of same key and forms the Shuffled Intermediate Results (SIR).

7. Each reduce worker applies the assigned user-defined reduce function on the shuffled data to generate the final key-value pairs output (O), the final output files are then saved in the distributed file system.

In MapReduce, fault-tolerance is achieved by re-executing the failed tasks. Failures can occur due to hardware causes such as disk failures, out of disk, out of memory, and socket time out. To improve MapReduce performance, speculative execution can be activated, where backup tasks are created to speed up the lacking in-progress tasks



(a) (b)

Figure 2.2: Framework components in Hadoop (a) Hadoop 1.x, and (b) Hadoop 2.x.

known as stragglers [17]. Developing efficient MapReduce applications requires advanced programming skills to fit the computations into the map and reduce pipeline, and deep knowledge of underlying infrastructures to properly configure and optimise a wide range of parameters [45, 46].

2.1.2 Apache Hadoop Architecture and Related Applications

Hadoop, which is currently under the auspices of the Apache Software Foundation, is an open source software framework written in Java for reliable and scalable distributed computing [47]. This framework was initiated by Doug Cutting who utilised the MapReduce programming model for indexing web crawls and was open-sourced by Yahoo in 2005. A wide range of applications and programming frameworks can run natively in Hadoop as depicted in Figure 2.2. Examples of these applications and frameworks are Pig, Tez, Hive, HBase, Storm, Giraph for graph processing, and Mahout for machine learning [1]. The basic components of the first versions of Hadoop; Hadoop 1.x are depicted in Figure 2.2(a). These versions contain a layer for the Hadoop Distributed File System (HDFS), a layer for the MapReduce 1.0 engine which resembles Google's MapReduce, and can have other applications on the top layer. The MapReduce 1.0 layer follows the master-slave architecture. The master is a single node containing a Job Tracker (JT), while each slave node contains a Task Tracker (TT). The JT handles jobs assignment and scheduling and maintains the data and metadata

of jobs, in addition to resources information. It also monitors the liveness of TTs and the availability of their resources by sending periodic heartbeat messages typically each 3 seconds. Each TT contains a predefined set of slots. Once it accepts a map or a reduce task, it launches a Java Virtual Machine (JVM) in one of its slots to perform the task, and periodically updates the JT with the task status [47]. The HDFS layer consists of a name node in the master and several data nodes in each slave node. The name node stores the details of the data nodes and the addresses of the data blocks and their replicas. It also checks the data nodes via heartbeat messages and manages load balancing.

Default tasks scheduling mechanisms in Hadoop are First-In First-Out (FIFO), capacity scheduler, and Hadoop Fair Scheduler (HFS). FIFO schedules the jobs according to their arrival time which leads to undesirable delays in environments with a mix of long batch jobs and small interactive jobs [48]. The Capacity scheduler developed at Yahoo reserves a pool containing minimum resources guarantees for each user, and hence suits systems with multiple users [49]. FIFO scheduling is then used for the jobs of the same user. The Fair scheduler developed at Facebook dynamically allocates the resources equally between jobs. It thus improves the response time of small jobs [50].

Hadoop 2.x, which is also depicted in Figure 2.2(b), introduced a resource management platform named YARN; Yet Another Resource Negotiator [51]. YARN decouples the resource management infrastructure from the processing components and enables the coexistence of different processing frameworks beside MapReduce which increases the flexibility in big data clusters. In YARN, the JT and TT are replaced with three components which are the Resource Manager (RM), the Node Manager (NM), and the Application Master (AM). The RM is a per-cluster global resources manager which runs as a daemon on a dedicated node. It contains a scheduler that dynamically leases the available cluster resources in the form of containers, which are considered as logical bundles (e.g. 2 GB RAM, 1 Central Processing Unit (CPU) core), among competing MapReduce jobs and other applications according to their demands and scheduling priorities. A NM is a per-server daemon that is responsible for monitoring the health

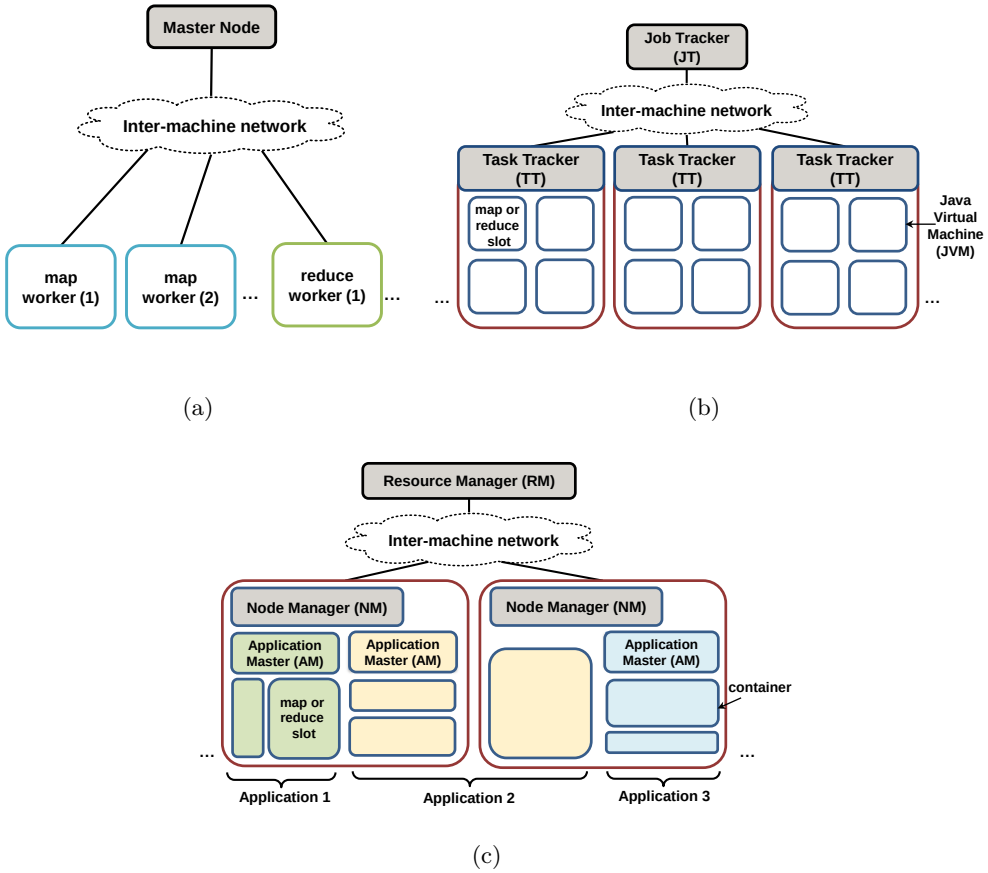


Figure 2.3: Comparison of clusters with (a) Google's MapReduce, (b) Hadoop 1.x, and (c) Hadoop 2.x with YARN.

of its physical node, tracking its containers assignments, and managing the containers lifecycle (i.e. starting and killing). An AM is a per-application container that manages the resources consumption, the jobs execution flow, and also handles the fault-tolerance tasks. The AM, which typically needs to harness resources from several nodes to finish its job, issues a resource request to the RM indicating the required number of containers, the required resources per a container, and the locality preferences.

Figure 2.3 illustrates the differences between Hadoop 1.x, and Hadoop 2.x with YARN. YARN increases the resource allocation flexibility in MapReduce 2 as it utilises flexible containers for resources allocations and hence eliminates the use of fixed

resources slots as in MapReduce 1. However, this advantage comes at the expenses of added system complexity and a slight increase in the power consumption when compared to MapReduce 1 [52]. Another difference is that the intermediate results shuffling operations in MapReduce 2 are performed via auxiliary services that preserve the output of a container's results before killing it.

2.1.3 Distributed In-memory Processing

A limitation in MapReduce that leads to un-utilised usage of disk I/O, CPU and network bandwidth resources is the need for outputs materialisation before being accessible for further computations, performed for fault-tolerance [53]. Hence, MapReduce is generally less suitable for interactive and iterative computations that require repetitive access to results. Instead, distributed in-memory processing systems are widely preferred. These use fast memory units such as Dynamic Random Access Memory (DRAM) or cache units to provide rapid access to data at run-time. However, as memory units are volatile, in-memory systems are required to adopt advanced mechanisms to guarantee fault-tolerance and data durability. To support in-memory interactive and iterative data processing, Spark, which is written in Scala, was introduced [54]. In Spark, the data is stored as Resilient Distributed Data sets (RDD) which are general purpose and fault-tolerant abstraction for data sharing in distributed computations. RDDs are created in the memory through course-grained deterministic transformations to datasets such as map, flatmap, filter, join, and GroupByKey.

2.1.4 Distributed Graph Processing

The partitioning and processing of graphs (i.e. a description of entities by vertices and their relationships by connecting edges) is considered a key class of big data applications especially for social networks that contain graphs with up to billions of entities and edges [55]. Most big data graph processing applications utilise in-memory systems due to the iterative nature of their algorithms. Examples are Pregel, Giraph, Trinity, GraphLab, and PowerGraph [1].

2.1.5 Big Data Streaming Applications

Applications that receive unbounded data, also known as *streams*, require stream processing applications with low latency and efficient processing to cope with the high arrival rate of the events within the streams. If they did not cope, the processing of some events is dropped leading to load shedding which is undesirable. An example is Storm which was developed at Twitter as a distributed and fault-tolerant platform for real-time processing of streams [56]. It utilised two primitives; *spout* and *bolts* to apply transformations to data streams, also named tuples, in a reliable and distributed manner. The spouts define the streams sources, while the bolts perform the required computations on the tuples, and emit the resultant modified tuples to other bolts. The computations in Storm are described as graphs where each node is either a spout or a bolt, and the vertices are the tuples routes. Storm relies on different grouping methods to specify the distribution of the tuples between the nodes. These include *shuffle* where streams are partitioned randomly, *field* where partitioning is performed according to a defined criteria, *all* where the streams are send to all bolts, and *global* where all streams are copied to a single bolt.

2.2 Cloud Computing Services

Cloud computing aims to enable seamless access for multi-users or tenants to a pool of computational, storage, and networking resources that typically reside within and between several geographically distributed data centres. Unlike traditional IT services that are limited by localised resources inaccessible by remote computing units, cloud computing services allow dynamic outsourcing to software and/or hardware resources. Hence, they can provide scalable and large-scale computational solutions while increasing the resources utilisation. Moreover, cloud computing considerably reduces both; the capital expenditure (CAPEX) and operational expense (OPEX) of software and hardware and increases the resilience. Consequently, it is continuing to encourage wide deployments of large-scale Internet-based services by various organisations and enter-

prises [57].

Cloud computing services can be categorised according to the outsourced resources and end-users privileges into Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS) [58]. SaaS provides on-demand Internet-based services and applications to end-users without providing the privilege of controlling or accessing the hardware, network, operating system, or the development platforms resources. Examples of SaaS are the services offered by Salesforce, Google Apps, and Microsoft Office 365. In the PaaS model, end-users have access privilege to the platform which enables them to develop, control, and upgrade their own cloud applications, but not to the underlying hardware. Thus, more flexibility is provided without the need for owning, operating and maintaining the hardware. Microsoft Azure and Amazon Web Services (AWS) provides PaaS offers. The IaaS model provides end-users with extra provisioning privileges that allow them to fully control the hardware, the operating systems, and the applications development platforms [59]. Examples of IaaS include Amazon Elastic Compute Cloud (EC2), Rackspace, and Google Compute Engine.

Although MapReduce and many other big data frameworks were originally provisioned for use in local clusters under controlled environments, there is an increasing number of cloud computing-based big data applications realisations and services to increase the profit and utilisation of cloud infrastructures despite the incurred overheads and challenges. Examples of such services are Amazon Elastic MapReduce, Microsoft's Azure HDInsight, VMWare Serengeti's project, Cloud MapReduce, and Resilin. Requirements and challenges of deploying big data applications in cloud computing infrastructures and the options for deploying big data application in geo-distributed clouds are discussed in [1].

2.2.1 Content Delivery Applications

Delivering different content such as Video and audio to Internet users is typically performed over a Content Distribution Network (CDN) [60] or an Information-Centric Network (ICN) [61]. CDNs are scalable networks that cache popular content throughout

Internet Service Provider (ISP) infrastructures, while ICNs support name-based routing to ease access to content. Content Service Providers (CSPs) place their workloads and content in geo-distributed data centres to improve the quality of their services, and rely on multiple ISPs or dedicated Wide Area Networks (WANs) [62] to connect these data centres. Advantages such as load balancing, increasing the capacity, availability and resilience against catastrophic failures, and reducing the latency by being close to the users are attained [63, 64]. A Quality of Service (QoS) metric that directly impacts the revenue of CSPs is the response time of interactive services. It was reported in [65] and [66] that a latency of 100 ms in search results caused 1% loss in the sales of Amazon, while a latency of 500 ms caused 20% sales drop, and a speed up of 5 seconds resulted in 10% sales increase in Google. QoE in video services is also impacted by latency. It was measured in [67] that with more than 2 seconds delay in content delivery, 60% of the users abandon the service.

2.3 Summary

This chapter reviews MapReduce, Apache Hadoop and related applications in addition to distributed in-memory big data applications, distributed graph processing, and big data stream processing applications. It shows the diversity in the available big data applications and frameworks and the extensive communication requirements that impose performance and energy consumption challenges in data centres. This chapter also reviews cloud computing services and emphasis on content delivery services which contributes a high portion of the Internet traffic and have strict QoE requirements. The following chapter reviews the networking and computing infrastructure utilised to deliver the applications and services summarised in this chapter.

Chapter 3

Cloud Networks and Data Centres

This chapter provides a brief review of cloud transport networks and cloud data centres. The remainder of this Chapter is organised as follows: Section 3.1 outlines the recent progress made in core IP over WDM networks and Passive Optical Network (PON)-based access networks while Section 3.2 discusses the implications of big data on the energy consumption of these networks. Finally, Section 3.3 provides a review of electronic switching, hybrid, and all optical data centre state-of-the-art topologies, summarises the characteristics of traffic inside data centres, summarises intra data centre routing protocols and traffic scheduling mechanisms, and discusses the energy efficiency of data centres.

3.1 Cloud Transport Networks

Cloud transport networks, as depicted in Figure 3.1, are composed of a core WAN, typically connected as a mesh to link cities in a country or across continents, Metropolitan Area Network (MAN) connected in a ring or star topologies, and access networks that typically use PON technologies and topologies and different mobile access networking technologies [68]. In the following Subsection, the advances in core networks and PON access networks are discussed.

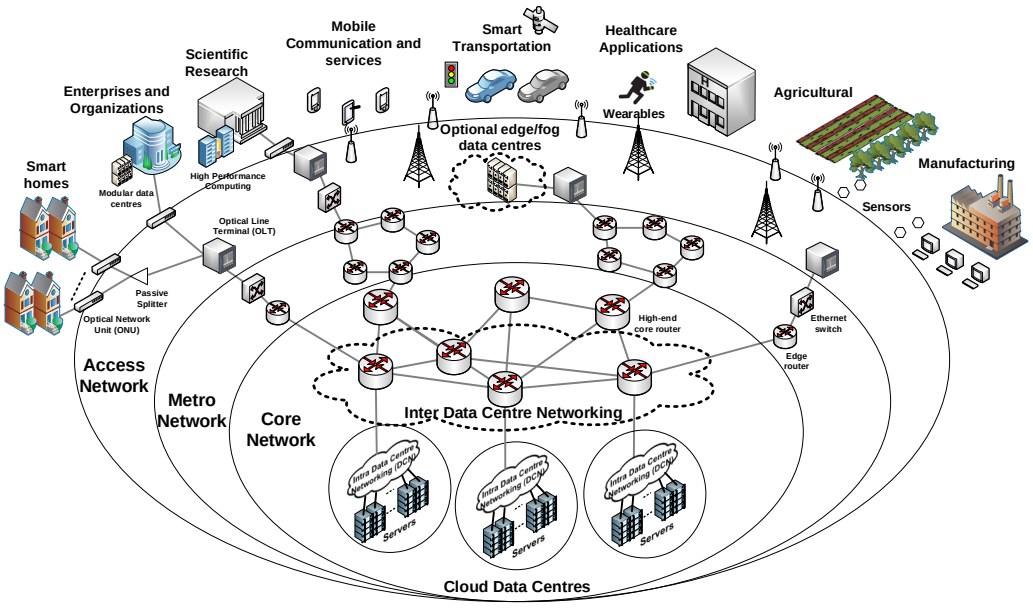


Figure 3.1: Cloud networking and computing infrastructure.

3.1.1 Core IP over WDM Networks

Core networks use fiber optic communication technologies due to their high capacity, reach, and energy efficiency. Internet Protocol (IP) routers in core networks are widely integrated with optical switches to form IP over Wavelength-Division-Multiplexing (WDM) architectures. The Internet protocol defines the basic rules of exchanging electrical data packets between devices over the Internet, while WDM technology mul-

tiplexes multiple optical signals with different wavelengths into optical fiber. In IP over WDM, the IP layer aggregates the traffic from metro and access networks and connects with the optical layer through short reach interfaces. The optical layer is composed of optical switching devices, mainly Reconfigurable Optical Add Drop Multiplexers (ROADM) realised by Wavelength Selective Switches (WSS), or Optical Cross Connect (OXC) switches to drop wavelengths of destined IP traffic, insert wavelengths for newly received IP traffic, and convert wavelengths, if necessary, to groom transit IP traffic. In addition, transponders for demodulation and modulation and Optical-Electrical-Optical (O/E/O) conversations are required [69]. Between the nodes, fiber links spanning up to thousands of kilometers are utilised. Due to physical impairments in fibers causing optical power losses and pulse dispersion, amplification is required mainly with Erbium-Doped Fiber Amplifiers (EDFAs) at fixed distances. Reamplification, Reshaping, and Retiming (3R) regenerators can also be installed at distances depending on the line rate. Such networks have sophisticated heterogeneous configurations and are typically configured to be static for long periods to reduce labor risk and costs. However, the increasing aggregated traffic due to Internet-based services with big data workloads makes legacy infrastructures incapable of serving future demands efficiently and recalling for improvements at different layers [70–73]. Legacy Coarse or Dense WDM systems, standardised by the International Telecommunication Union (ITU), utilised transponders based on On Off Keying (i.e. intensity) modulation and direct detection with carriers in the Conventional band (C-band) (1530-1565 nm) with 50 GHz spacing between channels and with up to 96 channels [63]. Typically, Single-Mode Fiber (SMF) links with a Single Line Rate (SLR) at 10 Gbps or 40 Gbps per fiber are utilised.

To increase the capacity with existing fiber plants and to improve the spectral efficiency, the use of Mixed-Line Rate (MLR) and multiple modulation formats with more than 1 bit per symbol such as duobinary and Differential Quadrature Phase Shift Keying (DQPSK) were proposed [74]. The rates in MLR systems can be adjusted to transport traffic with short/long reach with high/low rates to reduce regeneration re-

quirements. With the advances in digital coherent detection enabled with high sampling rate Digital-to-Analogue Converters (DACs), Digital Signal Processors (DSP), pre and post digital Chromatic Dispersion (CD) compensation, and Forward Error Correction (FEC) [63, 75, 76], the use of coherent higher order modulation formats such as QPSK, Quadrature Amplitude Modulation (QAM), in addition to Polarisation Multiplexing were also proposed to enhance the spectral efficiency. DSP can realise matched-filters for detection and hence, enables transmitting at the Nyquist limit for Inter Symbol Interference free transmission and relaxes the 50 GHz guard bands requirements between adjacent WDM channels allowing more channels in the C-band [63]. The use of the Long-band (L-band) (1565-1625 nm), and the Short-band (S-band) (1460-1530 nm) were also proposed to accommodate more channels, however, costly and complex filters such as Raman amplifiers are required in addition to careful impairments and non-linearities compensations approaches [63, 77]. Space Division multiplexing (SDM)-based solutions that enable the wavelengths reuse in the fiber are also considered to increase links' capacity. However, these solutions call for the replacement of SMFs with Multi Mode Fibers (MMFs) (i.e. with large core diameter to accommodate several lightpaths propagating at different modes), or Multi Core Fiber (MCFs) (i.e. with several cores within the fiber) to spatially separate lightpaths that use the same carrier [77].

To further improve the spectral efficiency and allow dynamic bandwidth allocation, the concept of superchannels in Elastic Optical Networks (EONs) is also introduced [78–81]. A superchannel is composed of bundled spatial or spectral channels with variable bandwidths at the granularity of 12.5 GHz, as defined by the ITU-T SG15/G.694.1 FlexGrid. These channels can be transmitted as a single entity with guard bands only between superchannels. Such channels can be constructed with Nyquist WDM or with coherent Optical Orthogonal Frequency Division Multiplexing (O-OFDM) which overlaps adjacent subcarriers [82–84]. Such flexibility in bandwidth assignments in FlexGrids requires programmable and adaptive networking equipment such as Bandwidth-Variable Transponders (BV-Ts), Bandwidth-Variable Wavelength Selective Switches (BV-WSSs), and Contentionless, Directionless, and Col-

orless ROADMs as detailed in [71, 72], and [77]. The modulation format, line rate, and bandwidth of each channel can then be dynamically configured with Software-Defined Networking (SDN) control [72] to provide programmable and agile Software-Defined Optical Networks with fine-grained control at the optical components and IP layer levels [85, 86].

3.1.2 Passive Optical Network (PON) Access Networks

Passive Optical Networks (PONs) are optical distribution networks composed of fiber optic links in addition to a subset of passive components selected according to the design requirements [68]. The passive devices include splitters, star couplers, star reflector, fibre Bragg Grating, Arrayed Waveguide Gratings (AWGs), and Arrayed Waveguide Grating Routers (AWGRs). PONs are widely used to connect networking equipment in the access network due to their advantages that include the reduction in costs while being data-rate agnostic, low maintenance costs, and the reduction in the power consumption in the last mile by eliminating the need for additional batteries or powering costs for active devices between the Central Office (CO) and the premises [87]. An example is the connection between a CO equipment that includes the Optical Line Terminals (OLTs) and the premises such as residential houses through an Optical Network Unit (ONU). Several standards can be used in PONs to control the upload and download rates such as Point to Multi-Point (P2MP), Point to Point (P2P), Broadband Passive Optical Network (BPON), Ethernet PON (EPON), Gigabit-capable PON (GPON), 10G EPON, next-generation PON (XG-PON), and Time Wavelength Division Multiplexing PON (TWDM PON) [88].

3.2 Big Data Implications on the Energy Consumption of Cloud Networking Infrastructures

Driven by the economic, environmental, and social impacts of the increased CAPEX, OPEX, Global Greenhouse Gas (GHG) emission, and carbon footprints as a result

of the expanding demands for Internet-based services, tremendous efforts have been devoted by industry and academia to reduce the power consumption and increase the energy efficiency of transport networks [89, 90]. These services empowered by fog, edge, and cloud computing, and various big data frameworks, incur huge traffic loads on networking infrastructures and computational loads on hosting data centres which in turn, increase the power consumption and hence, the carbon footprints of these infrastructures [10]. In 2018, the global energy demand for data centres was about 198 TWh which is equivalent to 1% of the total demand, while for transport networks was about 260 TWh which is equivalent to 1.1% of the total demand. Under an assumption that data centres infrastructures will be more efficient, an estimation for the data centres demand in 2021 is 191 TWh despite expected 80% increase in traffic and 50% increase in workloads. In 2021, transport networks are expected to consume 280 TWh if moderately improved [91]. Thus, improving both infrastructures is essential to reduce or maintain their demand under increasing traffic in the future. Optimising core networks plays an important role in improving the energy efficiency of the cloud networking infrastructures challenged by big data. The reductions of energy consumption and carbon footprint in core networks have been widely considered in the literature by optimising the design of their systems, devices, and/or routing protocols [42,68,92–108], utilising renewable energy sources [109–115], and by optimising the resource assignment and content placement in different Internet-based applications [60, 61, 116–128].

The early positioning study in [92] to green the Internet addressed the impact of coordinated and uncoordinated sleeps (for line cards, crossbars, and main processors within switches) on the switching protocols such as Open Shortest Path First (OSPF) and Internal Border Gateway Protocol (IBGP). Factors such as how, when, and where to cause devices to sleep, and the overheads of redirecting the traffic and awakening the devices were addressed. The study pointed out that energy savings are feasible but are challenging due to the modification required in devices and protocols. In [93], several energy minimisation approaches were proposed such as dynamic voltage scaling and dynamic frequency scaling at the circuit level, and efficient routing based on equipment

with efficient energy profiles at the network level. The consideration of switching off idle nodes and rate adaptation have also been reported in [94]. through Mixed Integer Linear Programming (MILP) and heuristic methods. The non-bypass approach requires O/E/O conversation to lightpaths (i.e. traffic carried optically in fiber links and optical devices) in all intermediate nodes, to be processed electronically in the IP layer and routed to following lightpaths. On the other hand, the bypass approach omits the need for O/E/O conversation in intermediate nodes, and hence reduces the number of IP router ports needed, and achieves power consumption savings between 25% and 45% compared to the non-bypass approach. In [96], a joint optimisation for the physical topology of core IP over WDM networks, the energy consumption and average propagation delay is considered under bypass or non-bypass virtual topologies for symmetric and asymmetric traffic profiles. Additional 10% saving was achieved compared to the work in [95].

Traffic-focused optimisations for IP over WDM networks were also considered, for example in [97–102]. Optimising static and dynamic traffic scheduling and grooming were considered in [97–100] in normal and post-disaster situations to reduce the energy consumption and demands blocking ratio. Techniques such as utilising excess capacity, traffic filtering, protection path utilisation, and services differentiation were examined. To achieve lossless reduction for transmitted traffic, the use of Network Coding in non-bypass IP over WDM was proposed in [101, 102]. Network-coded ports encode bidirectional traffic flows via XOR operations, and hence reduce the number of router ports required compared to un-coded ports. Utilising renewable energy resources such as solar and wind to reduce non-renewable energy usage in IP over WDM networks with data centres was proposed in [109], and [110]. Factors such as renewable energy average availability and their transmission losses, regular and inter data centre traffic, and the network topology were considered to optimise the locations of the data centres and an average reduction by 73% in non-renewable energy usage was achieved. The work in [111] considered periodical reconfiguration to virtual topologies in IP over WDM networks based on a “follow the sun, follow the wind” operational strategy. Renewable

energy was also considered in IP over WDM networks for cloud computing to green their traffic routing [112], content distribution [113], services migration [114], and for Virtual Network Embedding (VNE) assignments [115].

The placements of data centres and their contents in IP over WDM core nodes were addressed in [117] while considering the energy consumption, propagation delay, and users upload and download traffic. An early effort to green the Internet [118] suggested distributing Nano Data Centres (NaDa) next to home gateways to provide various caching services. In [119], the energy efficiency of Video-on-Demand (VoD) services was examined by evaluating five strategic caching locations in core, metro, and access networks. The work in [120–122] addressed the energy efficiency of Internet Protocol Television (IPTV) services by optimising video content caching in IP over WDM networks while considering the size and power consumption of the caches and the popularity of the contents. To maximise cache hit rates, the dynamics of TV viewing behaviors throughout the day were explored. Several optimised content replacement strategies were proposed and up to 89% power consumption reduction was achieved compared to networks with no caching. The energy efficiency of peer-to-peer protocol-based CDNs in IP over WDM networks was examined in [123] while considering the network topology, content replications, and the behaviours of users. In [124], the energy efficiency and performance of various cloud computing services over non-bypass IP over WDM networks under centralised and distributed computing modes were considered. Energy-aware MILP models were developed to optimise the number, location, capacity and contents of the clouds for three cloud services namely; content delivery, Storage-as-a-Service, and VM-based applications. An energy efficient cloud content delivery heuristic (DEER-CD) and a real-time VM placement heuristic (DEERVM) were developed to minimise the power consumption of these services. The results showed that replicating popular contents and services in several clouds yielded 43% power saving compared to centralised placements. The placement of VMs in IP over WDM networks for cloud computing was optimised in [125] while considering their workloads, intra-VM traffic, number of users, and replicas distribution and an energy saving of 23% was

achieved compared to one location placements. The computing and networking energy efficiency of cloud services realised with VMs and VNs in scenarios using a server, a data centre, or multiple geo-distributed data centres were considered in [116]. A Real-time heuristics for Energy Optimised Virtual Network Embedding (REOViNE) that considered the delay, clients locations, load distribution, and efficient energy profiles for data centres was proposed and up to 60% power savings were achieved compared to bandwidth cost optimised VNE.

To bridge the gap between traffic growth and networking energy efficiency in wired access, mobile, and core networks, GreenTouch [1], which is a leading Information and Communication Technology (ICT) research consortium composed of fifty industrial and academic contributors, was formed in 2010 to provide architectures and specifications targeting energy efficiency improvements by a factor of 1000 in 2020 compared to 2010. As part of the GreenTouch recommendations, and to provide a road map to ISP operators for energy efficient design for cloud networks, the work in [42, 107] proposed a combined considerations for IP over WDM design approaches, and the cloud networking-focused approaches in [116], and [124]. The design approaches jointly consider optical bypass, sleep mode for components, efficient protection, MLR, optimised topology and routing, in addition to improvements in hardware where two scenarios; Business-As-Usual (BAU), and BAU with GreenTouch improvements are examined. Evaluations on AT&T core network with realistic 7 data centres locations and 2020 projected traffic, based on Cisco Visual Network Index (VNI) forecast and a population-based gravity model, indicated energy efficiency improvements of 315x compared to 2010 core networks. Focusing on big data and its applications, the work in [127–130] addressed improving the energy efficiency of transport networks while considering different "5V" characteristics of big data and suggested progressive processing in intermediate nodes as the data traverse from source to central data centres.

[1] Available at the time of writing at: www.greentouch.org

3.3 Cloud Data Centres

Data centres can be defined as a collection of servers, switches, and storage devices to provide various data processing and retrieval services [131]. Intra Data Centre Networking (DCN), defined by the topology (i.e. the connections between the servers and switches), links capacity, and the switching technologies utilised, and routing protocols, is an important design aspect that impacts the performance, power consumption, scalability, resilience, and cost of data centres. The rest of this section is organised as follows: Subsection 3.3.1 reviews electronic switching-based data centres, while Subsection 3.3.2 reviews proposed and demonstrated hybrid electronic/optical and optical switching-based data centres. Subsection 3.3.3 presents traffic characteristics in cloud data centres, while Subsection 3.3.4 reviews intra DCN routing protocol and scheduling mechanisms. Finally, Subsection 3.3.5 addresses the energy efficiency in data centres.

3.3.1 Electronic Switching Data Centres

Servers in data centres are typically organised in "racks" where each rack typically accommodates between 16-32 servers. A Top-of-Rack (ToR) switch (also known as access or edge switch) is used to provide direct connections between the rack's servers and indirect connections with other racks via higher layer/layers switches according to the DCN topology. Most of legacy DCNs have a multi-rooted tree structure where the ToR layer is connected either to an upper core layer (two-tiers) or upper aggregation and core layers (three-tiers) [131]. For various improvement purposes, alternative designs based on Clos networks, flattened connections with high-radix switches, unstructured connections, and wireless transceivers were also considered. These architectures can be classified as switch-centric as the servers are only connected to ToR switches and the routing functionalities are exclusive to the switches. Another class of DCNs, known as server-centric, utilises the servers/set of servers with multiport NIC and software-based routing to aid the process of traffic forwarding. A brief description for some electronic switching DCNs while emphasising on their suitability for big data and cloud

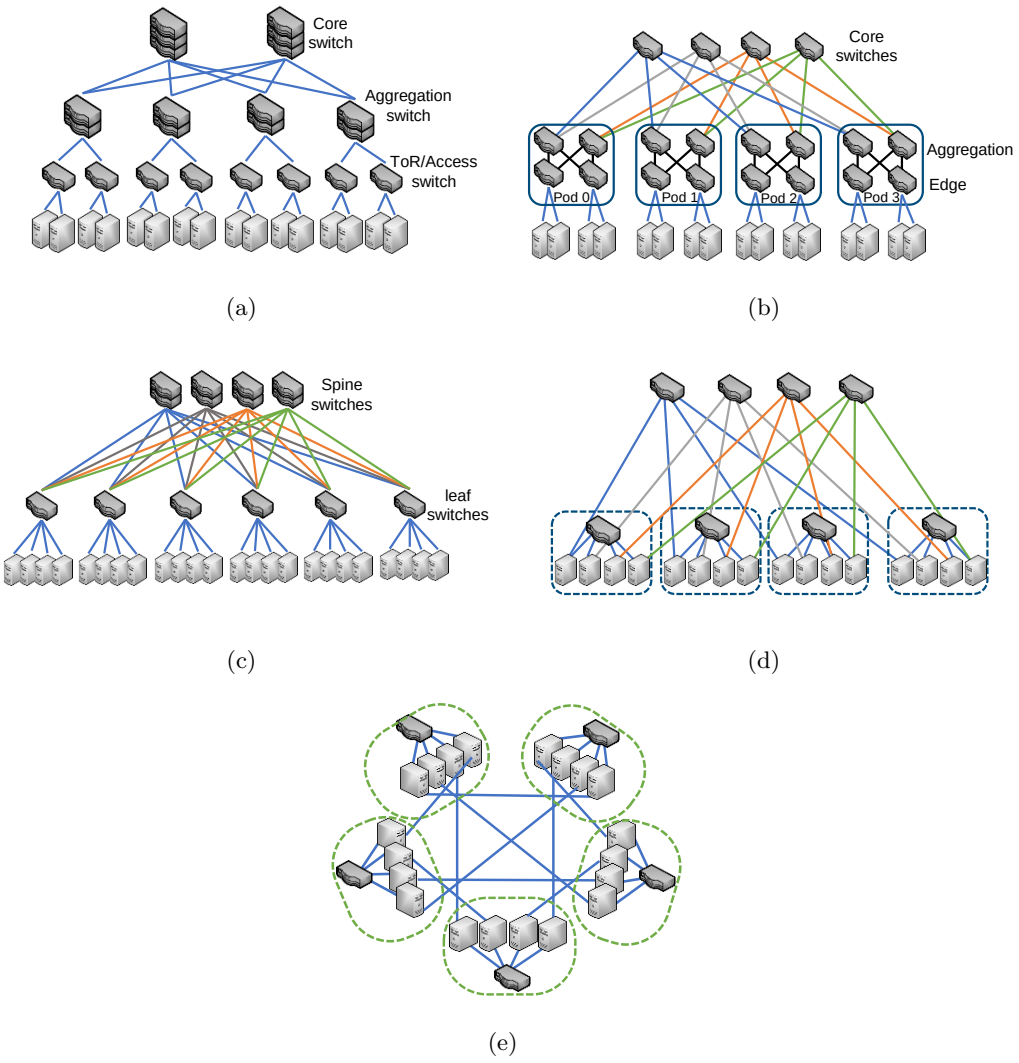


Figure 3.2: Examples of electronic switching DCNs (a) Three-tier, (b) Fat-tree, (c) Spine-leaf, (d) BCube, and (e) DCell.

applications [1] is provided below and some examples of small topologies are illustrated in Figure 3.2 showing the architecture in each case

- **Three-tier data centres** [131]: Three-tier designs have access, aggregation, and core layers (Figure 3.2(a)). Different subsets of ToR/access switches are connected to aggregation switches which connect to core switches with higher capacity to ensure all-to-all racks connectivity. This increases the over-subscription ratio

as the bisection bandwidth between different layers varies due to links sharing. Supported by firewall, load balancing, and security features in their expensive switches, three-tier data centres were sufficient for legacy Internet-based services with dominant south-north traffic and were widely adopted in production data centres.

- **k-ary Fat-tree** [132]: Fat-tree was proposed to provide 1:1 oversubscription and multiple equal-cost paths between servers in a cost-effective manner by utilising commodity switches with the same number of ports (k) at all layers. Fat-tree organises sets of equal edge and aggregation switches in *pods* and connects each pod as a complete bipartite graph. Each edge switch is connected to a fixed number of servers and each pod is connected to all core switches forming a folded-Clos network (Figure 3.2(b)). The Fat-tree architecture is widely considered in industry and research [133] indicating its efficiency with various workloads. However, its wiring complexities increase massively with scaling.

- **Spine-leaf** (e.g. [134, 135]): Spine-leaf DCNs are folded Clos-based architectures that gained widespread adoption by industry as they utilise commercially-available high-capacity and high-radix switches. Spine-leaf allows flexibility in the number of spine, leaf, and servers per leaf and links capacities at all layers (e.g. in Figure 3.2(c)). Hence, controllable oversubscription according to cost-performance trade-offs can be attained. Their commercial usage indicates acceptable performance with big data and cloud applications. However, wiring complexities are still high.

- **BCube** [136]: BCube is a generalised hyper cube-based architecture that targets modular data centres with scales that fit in shipping containers. The scaling in BCube is recursive where the first building block "BCube0" is composed of n servers and an n-port commodity switch and the k^{th} level (i.e. BCube$_k$) is composed of n BCube$_{k-1}$ and n^k n-port switches. Figure 3.2(d) shows a BCube$_1$ with n=4. For its multipath routing and to provide low latency and high bisec-

tion bandwidth and fault-tolerance, BCube utilises switches and servers equipped with multiple ports to connect with switches at different levels. BCube is hence, suitable for several traffic patterns such as 1-1, 1-many, one-all, and all-all which arise in big data workloads. However, with large scales, lower level to higher level bottlenecks increase and address space are to be overwritten.

- **DCell** [137]: DCell_k is a recursively-scaled data centre that utilises a commodity switch per DCell_0 pod to connect its servers, and the remaining of the $(k+1)$ ports in the servers for direct connections with servers in other pods of same level and in higher levels pods. Figure 3.2(e) shows a DCell_1 with 4 servers per pod. DCell provides high bandwidth, scalability, and fault-tolerance at low costs. In addition, under all-all, many-1, and 1-many traffic patterns, DCell achieves balanced routing, which ensures high performance for big data applications. However, as it scales, longer paths between servers in different levels are required.

3.3.2 Hybrid Electronic/Optical and All Optical Switching Data Centres

Optical switching technologies have been proposed for full or partial use in DCNs as solutions to overcome the bandwidth limitations of electronic switching, reduce costs, and to improve their performance and energy efficiency [138–143]. Such technologies eliminate the need for O/E/O conversion at intermediate hops and make the interconnections data-rate agnostic. Hybrid architectures add Optical Circuit Switching (OCS), typically realised with Micro-Electro-Mechanical System Switches (MEMSs) or free-space links, to enhance the capacity of an existing Electronic Packet Switching (EPS) network. To benefit from both technologies, bursty traffic (i.e. for mice flows) is offloaded to EPS while bulky traffic (i.e. for elephant flows) is offloaded to the OCS. MEMS-based OCS requires reconfiguration time in the scale of ms or μs before setting paths between pairs of ToR switches, and because packet headers are not processed, external control is needed for the reconfigurations. Another shortcoming of MEMS is their limited port count. WDM technology can increase the capacity of ports without huge increase in the power consumption [144], resolve wavelength contention, and reduce

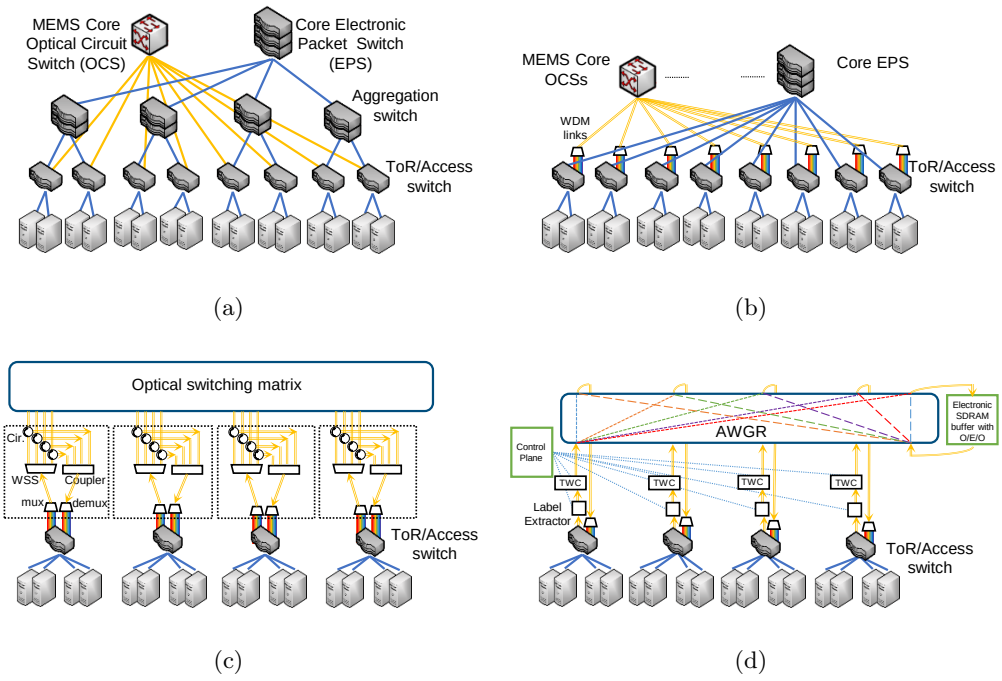


Figure 3.3: Examples of hybrid/all optical switching DCNs (a) c-Through, (b) Helios, (c) OSA/Proteus, and (d) DOS.

wiring complexities at the cost of additional devices for multiplexing, de-multiplexing, and fast tuning lasers and tuneable transceivers at ToRs or servers.

In hybrid and all optical DCNs, both active and passive components were considered. The passive components including fibers, waveguides, splitters, couplers, AWGs, and AWGRs, do not consume power but have insertion, crosstalk, and attenuation losses. Active components include Wavelength Selective Switches (WSSs), that can be configured to route different sets of wavelengths out of a total of M wavelengths in an input port to different N output ports (i.e. $1\times N$ switch), MEMSs, Semiconductor Optical Amplifiers (SOAs) that can provide switching time in the range of ns, Tuneable Wavelength Converters (TWCs), and Mach-Zehnder Interferometer (MZI) which are external modulators based on controllable phase shifts in split optical signals. In addition to OCS, Optical Packet Switching (OPS) [145–148] was also considered with or without intermediate electronic buffering. Examples of hybrid electrical/optical and all

optical switching DCNs are summarised below, and some are illustrated in Figure 3.3:

- **c-Through** [149]: In c-Through, electronic ToR switches are connected to a two-tier EPSs network and a MEMS-based OCS as depicted in Figure 3.3(a). The EPS maintains persistent but low bandwidth connections between all ToRs and handles mice flows, while the OCS must be configured to provide high bandwidth links between pairs of ToRs at a time to handle elephant flows. As the used MEMS had ms switching time, c-Through was only proven to improve the performance of workloads with slowly varying traffic.

- **Helios** [150]: In Helios, electronic ToR switches are connected to a single tier containing arbitrary number of EPSs and MEMS-based OCSs as in Figure 3.3(b). Helios performs WDM multiplexing in the OCS links and hence requires WDM transceivers in the ToRs. Due to its complex control, Helios was demonstrated to improve the performance of applications with second-scale traffic stability.

- **Optical Switching Architecture (OSA)** [151] / **Proteus** [152]: OSA and Proteus utilise a single MEMS-based optical switching matrix to dynamically change the physical topology of electronic ToRs connections. Each ToR is connected to the MEMS via an optical module that contains multiplexers/demultiplexers for WDM, a WSS, circulators, and couplers as depicted in Figure 3.3(c). This flexible design allows multiple connections per ToR to handle elephant flows and eliminates blocking for mice flows by enabling multi-hop connections via relaying ToRs. OSA was examined with bulk transfers and mice flows and minimal overheads were reported while achieving 60%-100% non-blocking bisection bandwidth.

- **Data centre Optical Switch (DOS)** [153]: DOS utilises an N+1-ports AWGR to connect N ToR electronic switches through OPS with the aid of optical label extractors as shown in Figure 3.3(d). Each ToR is connected via a TWC to the AWGR to enable it to connect to one other ToR at a time. At the same time, each ToR can receive from multiple ToR simultaneously. The last ports of the AWGR are connected to an electronic buffer to resolve contention for transmitting ToRs.

DOS suits applications with bursty traffic patterns, however, its disadvantages include the limited scalability of AWGRs and the power hungry buffering.

3.3.3 Characteristics of Traffic inside Data Centres

Traffic characteristics within enterprise, cloud, social networking, and university campus data centres have been reported and analysed in [154–158] to provide several insights about traffic patterns, volume variations, and congestion, in addition to various flows statistics such as their duration, arrivals, and inter-arrival times. Intra data centre traffic is mainly composed of different mixes of data centre applications including retrieval services and big data analytics, and provisioning operations such as data transfers, replication and backups. The first three pioneering studies by Microsoft Research [154–156] pointed that traffic monitoring tools used by ISPs in WANs do not suit data centres environments as their traffic characteristics are not statistically similar. The authors in [154] utilised low-overhead socket-level instrumentation at 1500 servers to collect application-aware traffic information. Two traffic patterns were defined; *Work-Seeks-Bandwidth*: for engineered applications with high locality, and *Scatter-Gather*: for applications that require servers to push or pull data from several others. It was found that 90% of the traffic stays inside the racks and that 80% of the flows last less than 10s, while less than 0.1% last more than 200s. In [155], an empirical study for traffic patterns in 19 tree-based two-tier and three-tier data centres for web-based services was carried based on coarse-grained measurements for links utilisation and packets loss rate. Average link loads were found the highest at the core switches and the highest packets losses were found at edge (i.e. ToR) switches. Additionally, fine-grained packet-level statistics at five edge switches in a smaller cluster showed a clear ON-OFF intensity and log-normal inter-arrivals during ON intervals.

SNMP logs were also utilised in [156] to study the traffic empirically while considering broader data centres usages and topologies. Those included 10 data centres with two-tier, three-tier, star-like, and Middle-of-Rack switches-based (i.e. connecting servers in several racks) topologies for university campuses, private enterprises, in ad-

dition to web-based services and big data analytics cloud data centres. It was found that 80% of the traffic stayed inside the racks in cloud data centres, while 40-90% left the racks in universities and enterprises data centres. Fine-grained packet traces from selected switches in 4 DCNs indicated that 80% of the flows are small (i.e. $\leq$ 10 KB), active flows were less than 10,000 per second per rack, and Inter-arrivals were less than 10 μs for 2-13% of the flows. The recent studies [157], and [158] presented the traffic characteristics inside Facebook's 4-tier Clos-based data centre that hosts hundreds of thousands of 10 Gbps servers. In [157], wide monitoring tools and per-host packet-level traces were utilised to characterise the traffic while focusing on its locality, stability, and predictability. The practice in this architecture recommends assigning each machine to one role (e.g. cache, web server, Hadoop), localising each type of workloads in certain racks, and varying the level of oversubscription according to the workload needs. Traffic was found to be neither fully rack-local nor all-all, and without ON-OFF behaviour. Also, it was found that servers communicated with up to 100s of servers concurrently, most of the flows were long-lived, and non-Hadoop packets were <200 Bytes. To capture fine-grained network behaviours such as μbursts (i.e. high utilisation events lasting <1 ms), high-resolution measurements at the rack-level with granularity of tens/hundreds of μs were utilised in [158] for the same data centre and application sets in the previous study. The measurements were based on a developed counter collection framework that poll packet counters and buffer utilisation statistics every 25μs, and 50μs, respectively. It was found that high utilisation events were short-lived as more than 70% of bursts lasted at most for tens of μs, and that load is very unbalanced as web and Hadoop racks had hot downlink ports while Cache had uplink hot ports. 90% of bursts lasted <200μs for all application types, and <50μs for Web racks, and the highest tail was recorded for Hadoop racks at 0.5 ms. It was noticed that the packets included in μbursts are larger than in the outside, μbursts were caused by application behavioural changes, and that the arrival rate of μbursts was not Poisson with 40% of inter-arrivals being> 100 μs for Cache and Web racks. Regarding the impact of μbursts on shared buffers, Hadoop racks had ports buffers at

>50% utilisation, while web and cache racks had a maximum of 71% and 64% of their ports buffers at high utilisation. Latency and packet loss measurements between VMs in different public clouds were presented and performed in [159] through a developed tool, PTPmesh, that aided cloud users in monitoring network conditions. The results for one way messaging delay between data centres in the same and different clouds were shown to range between μs and ms values. Specific traffic measurements for big data applications were presented in [160] and three traffic patterns; Single peak, repeated fixed-width peaks, varying heights and widths peaks were reported.

3.3.4 Intra Data Centres Routing Protocols and Traffic Scheduling Mechanisms

Routing protocols, which define the rules for choosing the paths for flows or flowlets between source and destination servers, were extensively surveyed in [161–166]. Routing in DCNs can be static or adaptive where paths assignments can be dynamic according to criteria measured by a feedback mechanism. Adaptive routing or traffic scheduling can be centralised where a single controller is required to gather network-wide information and to distribute routing and rate decisions to switches and servers, or distributed where the decisions are taken independently by the switches or servers according to local decision based on partial view of the network. Centralised mechanisms provide optimal decisions but have limited scalability while distributed mechanisms are scalable but not always optimal.

Tree-based data centres such as three-tier designs typically utilise VLAN with Spanning Tree Protocol (STP), which is a simple Layer2 protocol that eliminates loops by disabling redundant links and enforcing the traffic to route through core switches. Spine-leaf DCNs typically use improved protocols such as Transparent Interconnection of Lots of Links (TRILL) or Shortest Path Bridging (SPB) that enables the utilisation of all available links while ensuring loop-free routing. CONGA [167] was proposed as distributed flowlets routing mechanism for spine-leaf data centres that achieves load balancing by utilising leaf-leaf congestion feedback. Improved tree-based DCNs such as

Fat-tree and server-centric DCNs require designing their routing protocols closely with their topological properties to fully exploit the topology. For example, Fat-tree requires specific routing with two-level forwarding tables for servers with fixed pre-defined addressed [132]. For agility, VL2 [168] uses two addresses for servers; a Location-specific Address (LA), and an Application-specific Address (AA). For packets forwarding, VL2 employs Equal Cost Multi-Path (ECMP), which is a static layer3 routing protocol that distributes flows to paths by hashing, and Valiant Load Balancing (VLB), that randomly selects intermediate nodes between a source and destination. BCube employs a Source Routing protocol [136], DCell adopts a distributed routing protocol [137], and JellyFish [169] uses a k-shortest paths algorithm. c-Through uses the Edmond's algorithm to obtain the MEMSs configurations from traffic matrix, then the ToR switches traffic is sent via VLAN-based routing into the OSC or the EPS [149]. Helios has a complex control scheme of three modules; Topology Manager, circuit switch manager, and pod switch manager [150]. Mordia utilises a Traffic Matrix Scheduling (TMS) algorithm that obtain effective short-lived circuits schedules, based on predicted demands, that can be applied to configure the MEMS and WSSs sequentially [170]. OSA, and Proteus use the maximum-weight b-matching problem to enable the connection of multiple ToR switches, configure the WSS to match capacities and then use shortest path-based routing [151].

Using the Transmission Control Protocol (TCP), used in the Internet, in data centre environments has been proved to be inefficient due to the difference in the nature of traffic, the higher sensitivity to incast in data centres, and the key requirement for data centre applications of minimised Flow Completion Time (FCT) [163]. Thus, different transport protocols were proposed for DCNs. DCTCP [171] provides similar or better throughput than TCP and guarantees low Round Trip Time (RTT) by active control to queue lengths in the switches. MPTCP [172] splits flows to sub-flows and balances the load across several paths via linked congestion control. However, it might perform excessive splitting which requires extensive CPU and memory resources at end hosts. D2TCP [173] is a Deadline-aware Data centre TCP protocol that considers single

path for flows and performs load balancing. For FCT reduction, D3 proposed in [174] uses flows deadline information to control the transmission rate. pFabric [175] and PDQ [176] enable the prioritisation of the flows close to completion, and DeTail [177] splits flows, and performs adaptive load balancing based on queues occupancy to reduce the highest FCT. Alternatively, the centralised schedulers; Orchestra [178], Varys [179], and Baraat [180] target reducing the completion time of coflows which are sets of flows with applications-related semantic such as intermediate data shuffling in MapReduce.

SDN has been widely considered for data centres as its flexibility and agility can improve the load balancing, congestion detection and mitigation [181–183]. To allow users to make bandwidth reservation in data centres for their VM-to-VM communications, a centralised controller is used to determine the rate and path for each user's flow. SecondNet was proposed in [184] and is such a controller. Hedera in [185], detects elephant flows and maximises their throughput via a centralised SDN-based controller. ElasticTree in [186] improves the energy efficiency of Fat-tree DCNs by dynamically switching-off sets of links and switches while meeting demands and maintaining fault-tolerance.

3.3.5 Energy Efficiency in Data Centres

The energy consumption in data centres is attributed to servers and storage, networking devices, in addition to cooling, powering, and lighting facilities with percentages of 26%, 10%, 50%, 11%, and 3%, respectively of the total energy consumption [187]. As the energy consumption of servers is becoming proportional to loads, hence their energy efficiency is improving faster, the portion of the networking is expected to increase [188]. In [187], techniques for modeling the data centres energy consumption were comprehensively surveyed. In [189], green metrics including the Power Usage Effectiveness (PUE) (defined as the total facility power over the IT equipment power), and measurement tools that can characterise emissions were surveyed to aid in sustaining distributed data centres.

Several studies considered reducing the energy consumption and costs in data

centres at different levels [190–194]. For the hardware, dynamically switching off the idle components, proposing efficient hardware with inherent higher efficiency components, Dynamic Voltage and Frequency Scaling (DVFS), and utilising optical networking elements were considered. For example, to improve the energy proportionality of Ethernet switches, the Energy Efficient Ethernet (EEE) standard [194] was developed. EEE enables three states for interfaces which are active, idle with no transmission, and low power idle (i.e. deep sleep). Although EEE has gained industrial adoption, its activation is not advised due to uncertainty with its impact on applications performance [195]. Placement of workloads and VMs into fewer servers, and scheduling tasks to shave peak power usage were also proposed to balance the power consumption and utilisation in data centres.

3.4 Summary

This chapter reviews cloud transport networks while focusing on core IP over WDM networks and PON-based access networks and reviews data centre topologies, traffic characteristics, routing protocols and traffic scheduling. It also discusses the implications of big data on the energy consumption of these infrastructures and the need for improving their energy efficiency. The work in the reminder of this thesis considers only the physical layer of the technologies presented in this chapter. The following chapter introduces two PON-based data centre topologies that improve the performance and energy efficiency of data centres by utilising passive optical technologies. The work in chapters 5 and 6 addresses the performance, energy efficiency and resilience of different data centre topologies when running MapReduce workloads. Chapter 7 evaluates the energy consumption of transport networks when optimising the delivery of VoD traffic from cloud of fog data centres.

Chapter 4

Passive Optical Networks-based Data Centre Architectures

This chapter introduces two PON-based DCN designs that target both cloud and fog data centre environments. Both designs utilise ports in OLT line cards for inter and possibly intra data centre networking in addition to passive interconnects for the intra data centre networking between different PON groups (i.e. racks) within a PON cell (i.e. number of PON groups connected to a single OLT port). The first design is a switch-centric design that uses two AWGRs and the second is a server-centric design. The remainder of this chapter is organised as follows: Section 4.1 provides some related studies and introduces the designs, while Section 4.2 briefly describes the technologies utilised and requirements in both designs. Section 4.3 provides a MILP model to optimise the connections and wavelength routing and assignment in the AWGR-based design, while Section 4.4 provides the results. Finally, Section 4.5 provides the chapter summary.

4.1 Introduction

The limitations of current Data Centres Networks (DCNs) in terms of capacity, cost, and energy efficiency have triggered the need for new architectures capable of efficiently meeting the growing demands of cloud and fog computing distributed applications [4]. The proven high-performance and cost-effectiveness of Passive Optical Networks (PONs) in access networks has motivated the use of their technologies in designing energy efficient, low cost, scalable, and elastic future cloud and fog DCNs. The integration of optical line terminals in access networks and data centres or additional processing devices for extended fog computing and caching capacities was suggested in [196–199].

The benefits of using PONs in data centre networks include low equipment cost, low power consumption, data rate agnostic operation, and high scalability of PONs compared to EPS. Different PON technologies were considered for data centre networks, mainly while maintaining electronic ToR switches, including OFDM, WDM PON, and AWGRs [200–204]. To improve the scalability and reliability of large-scale data centres, the work in [205] proposed a passive optical cross-connect (PONX) with an efficient distributed Multiple Access Control (MAC) protocol that can also support fairness and QoS. The cross connect divides the signal from each input port to all other output ports equally which eliminates the need for reconfiguration, but reduces the spectral efficiency as spatial wavelengths reuse is not possible. Chen et. al. focused on the access tier in data centres and proposed a passive optical ToR interconnect architecture, POTORI [206], to be used with EPS-based or optical core and aggregation switches in data centres. A centralised cyclic-based MAC that supports WDM was proposed for the control and data planes.

In [207], five novel PON-based designs were introduced to provide scalable, low cost, energy-efficient and high capacity intra and inter rack interconnections for future DCNs. These designs can replace typical access, aggregation, and/or core switches in current DCNs with OLTs and different passive intra-rack (i.e. between servers) and

inter-rack (i.e. between racks) interconnections. The first two designs are TDM and TDM/WDM-based analogues to PONs in last-mile access networks where servers are organised in several racks and form cells that can communicate only through the OLT. Various desired oversubscription ratios and MAC protocols can be adopted. To improve intra rack communication, three passive technologies were suggested and to improve inter rack communication within each cell, three designs where proposed. The third design, which is further discussed in [208] and [209], utilises AWGRs to provide high-performance interconnections between racks within each cell, and an array of photo detectors and tuneable lasers at each server for wavelength detection and transmission. An optimisation study for the wavelengths assignment for inter-rack communication was presented in [208]. The energy efficiency of the design was assessed and compared to Fat-Tree and BCube, and energy savings of 45% and 80% were achieved, respectively. In [209], an SDN-based framework was suggested for the AWGR-based PON DCN, to improve the energy efficiency of routing and resource provisioning for inter-cells communication and improvement by up to 90% were obtained with no blocking. Further resources provision optimisations were carried in [210] while considering the delay-performance trade-offs for different applications. The results show that the delay can be decreased by 62% for delay-sensitive applications and the power consumption can be decreased by 22% for batch applications. For partial reduction of the server-attached tuneable lasers costs, the fourth design introduced the use of special servers to perform wavelength conversion for inter-rack communication [211, 212]. The work in [213] introduced the fifth design which is a cost-effective server-centric PON DCN that does not require tuneable lasers and instead, it utilises Network Interface Cards (NIC) with non-tunable optical transceivers in the relaying servers for inter-rack communication. Experimental evaluations were provided in [214, 215]. Benchmark studies were conducted against electronic, hybrid, and optical DCNs through evaluating the completion time of sort operations performed on big data workloads in [4], while resilience benchmark evaluations are reported in [6].

4.2 PON Technologies and Design Requirements

4.2.1 PON in Access Networks and for DCNs

In what follows, we discuss and compare the differences between using PONs in access networks and in data centre environments. In access networks, PONs provide high speed broadband voice, data, and video streaming (i.e. triple play) services through efficient and flexible protocols to end user in premises. A single strand of fiber connected to an OLT port is passively split via splitters or Arrayed Waveguide Gratings (AWGs) to provision services to 128-256 end locations equipped with ONU or Optical Network Terminal (ONT) at distances of up to 60 km from the central office [216]. Thus, no active components are required between the OLT and the end users' ONUs and ONTs. The split ratio and design of the passive optical distribution network in the middle depends on the area requirement and the PON protocols and standards utilised. Typically, this design can be a tree-based or a point to point design. The OLT switches in the central office are then responsible for channel access arbitration and upload and download bandwidth allocation. Figure 4.1 shows two different setups for PONs in access networks which are a Time Division Multiplexing (TDM)-based PON access network (Figure 4.1(a)) and hybrid TDM-Wavelength Division Multiplexing (WDM)-based PON access network (Figure 4.1(b)).

In access networks, ONU to ONU communication is not a concern as the traffic is mostly transmitted from OLT to users (download) or from the users to OLT (upload). Hence, the tree based topology design and the much lower upload bandwidth compared to the download bandwidth are suitable to meet the requirements in residential applications [9, 207].

A TDM PON and a hybrid TDM WDM PON were proposed for the use in data centres in [207] by connecting a number of racks containing servers (i.e. organised in cells, each contains a number of racks) passively to OLT ports with flexible superscription ratios. These designs replace electronic access and aggregation switches by passive connections and core switches by the OLT. Figure 4.2 depicts the use of a future-proof

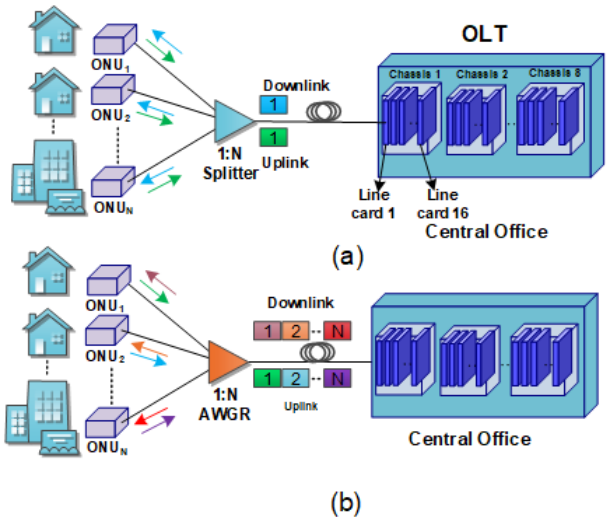


Figure 4.1: PON in access networks [9] (a) TDM PON, and (b) WDM PON.

OLT chassis such as [217] for PON cells (i.e. number of racks) interconnections. Typically, such an OLT chassis contains 16 service cards (i.e. line cards) where each card can provide up to 16 ports depending on the PON protocol. Each PON cell is then connected to a port in one of the line cards.

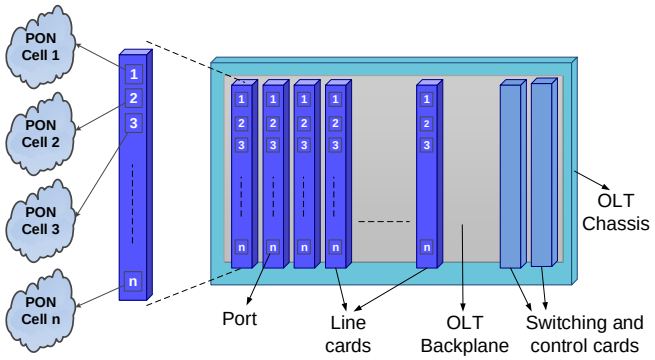


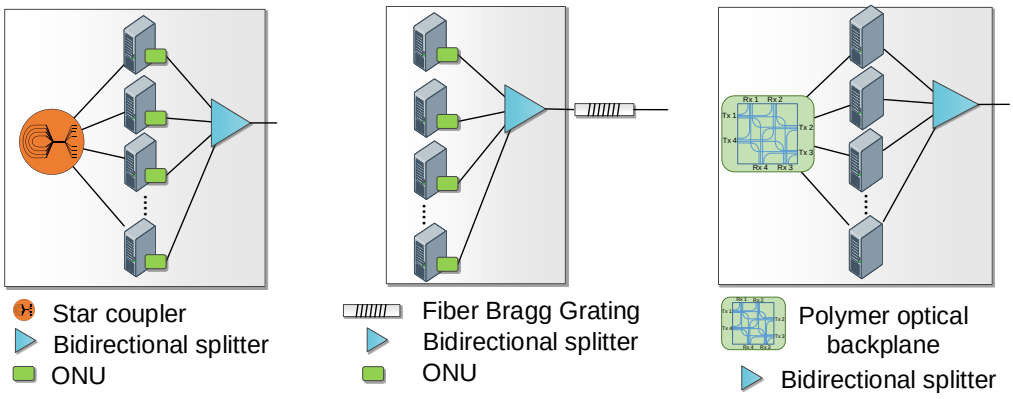
Figure 4.2: Architecture of an OLT chassis with the interconnections of PON-based DCN cells [9].

In the TDM PON and hybrid TDM WDM PON based designs, the servers utilise OLT ports to communicate with other servers in the same cell or in other cells. This

design limits the per server share of an OLT port upload and download bandwidth. To improve the performance and increase the per server bandwidth, several challenges are to be addressed. Typical PON setups are not sufficient for server to server communications as user to user traffic was not a concern in the access network. Using the OLT ports for all intra (including intra and inter rack communication) and inter cell communication incurs high overheads and provides limited bandwidth, thus, improving intra rack and intra cell communication is required. A number of designs utilising different passive technologies were also proposed in [9, 207].

4.2.2 Passive Technologies to Improve Intra-rack Communication

To improve intra rack communication within an individual rack in a cell, three passive technologies were proposed in [9,207]. Those are a star coupler, a Fiber Bragg Grating, and a Polymer optical backplane that was proposed in [218]. This backplane can provide non-blocking full mesh connectivity with a total of 1 Tbps capacity with multimode polymer waveguides each used at 10 Gbps rate. Figure 4.3 illustrates the connections when using these technologies for intra rack communication.



(a) (b) (c)

Figure 4.3: Passive technologies to improve intra-rack communication [9] (a) Star reflector, (b) Fiber Bragg Grating, and (c) Polymer optical backplane.

This will reduce the need to use the OLT port for intra rack traffic. Then, the

servers within the rack are interconnected with the remaining of the cell and with the OLT port through bidirectional splitters. Two of the designs proposed in [9, 207] for intra cell connections, which are a switch-centric and a server-centric designs, are discussed in the following two Subsections. Additional inter-cell designs, where any of the proposed five designs are used for the cells, can be used to attach servers in OLTs to increase the processing and storage capabilities at access network and the utilisation of OLTs at reduced data centre networking power consumption and cost.

4.2.3 The AWGR-centric Design (PON3 [1])

For intra cell connections (i.e. the connections between the racks and the OLT port), the use of two AWGRs was proposed for the AWGR-centric design in [9, 207–209]. An AWGR contains equal number of input and output ports (i.e. N × N AWGR) and provides passive N × N links between the input and output ports. Each input port routes different wavelengths to different output ports and each output port must receive a different wavelength from an output port. This can be realised with cyclic and acyclic designs for the wavelengths routing [69]. The number of AWGR ports required is a function of the number of racks within the cell and number of OLT ports the cell is connected to. An example of this design with four racks and connection to a single OLT port is depicted in Figure 4.4.

Each rack is to be connected to a single AWGR port and a single output port. The OLT port can be connected to both AWGRs through a single input and output port in each AWGR. If intra rack communication is to be realised only through one of the solutions proposed in the previous subsection, then a total of M-1 wavelengths are required to realise connections between different racks and between each rack and the OLT port where M is the total number of racks and OLT ports communicating. Section 4.3 provides a MILP model to optimise the connections and the wavelength assignment in this design. This design requires equipping the servers with tuneable transceivers. In addition to offloading intra rack traffic, this design also reduces the

[1] The third data centre design in [207].

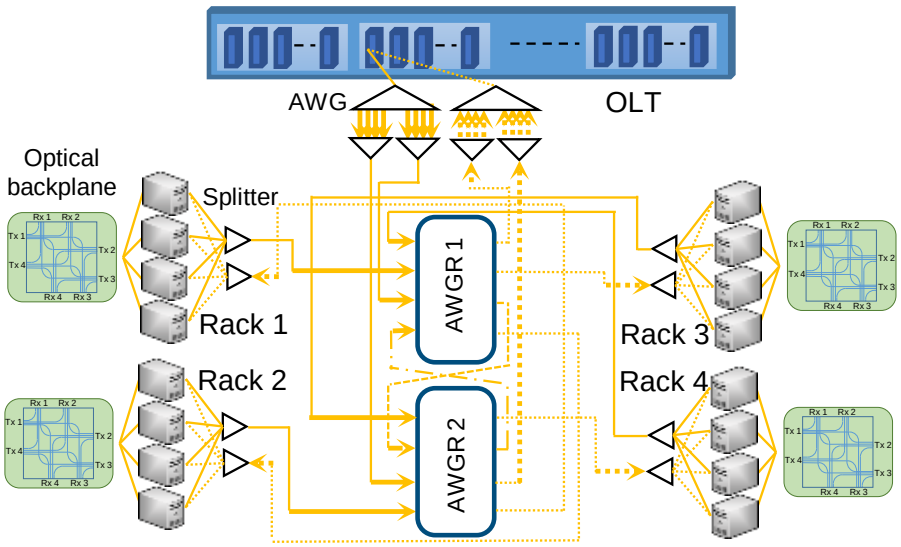


Figure 4.4: An example of the AWGR-centric design.

need to use the OLT port for intra cell traffic.

4.2.4 The Server-centric Design (PON5 [1])

The server-centric design is depicted in Figure 4.5. This design utilises NICs in servers to forward intra cell traffic between different racks and connects each rack with the OLT port through a single server in that rack. If a single wavelength is to be used, a star coupler can be used to connect the OLT port with the racks and all the servers share the bandwidth of that port through TDM only. If WDM is to be used, an AWG is used. This design provides multiple paths between servers in different racks at reduced costs compared to PON3.

4.3 MILP Model for Optimising the AWGR-centric Design

This section provides a MILP model to optimise the connections and wavelength routing and assignment in PON3. This model takes an initial topology where all input and output ports of racks and AWGRs are connected to all output and input ports of the

[1] The fifth data centre design in [207].

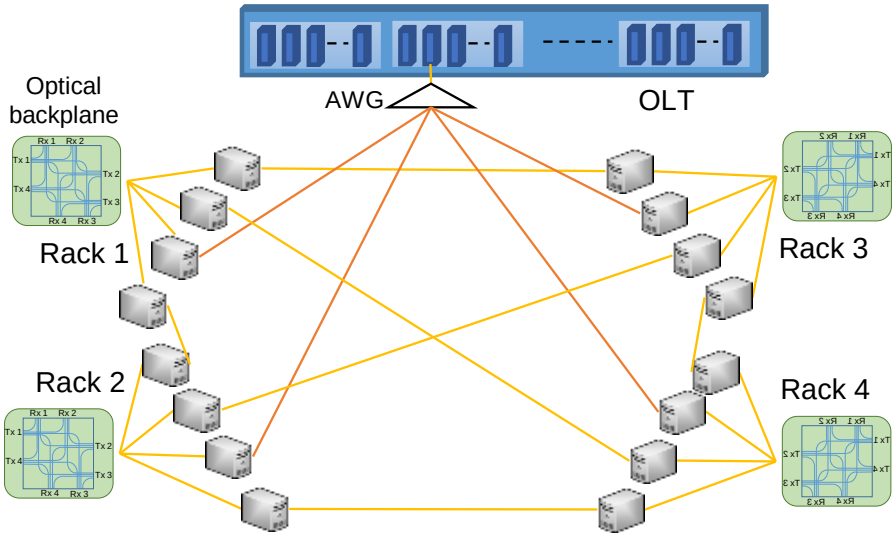


Figure 4.5: An example of the server-centric design.

AWGRs, respectively and maximises the number of achieved connections between the racks and the OLT port by assigning wavelengths to these connections while selecting the unique ports connections and maintaining correct routing and AWGR usage. The sets, parameters, and variables used in the model are provided below:

The objective is to maximise the connections between vertices $s, d \in \mathbb{P}, s \neq d$ which can be expressed as:

$$\max \sum_{\substack{j \in \mathbb{W}, s, d \in \mathbb{P} \\ s \neq d}} \mu_j^{sd} \tag{4.1}$$

Subject to the following constraints:

1. Flow conservation: The allocation of the links and wavelengths to connections

Sets and parameters:

G	Number of communicating vertices including OLT ports and PON groups
$\mathbb{W}$	Set of wavelengths used (count to $G-1$)
$\mathbb{K}$	Set of AWGRs
M	The size of an AWGR (i.e. $M \times M$) which is equal to the number of wavelengths needed (i.e. $G-1$)
$\mathbb{T}$	Set of OLT ports (initially one port is needed per PON cell)
$\mathbb{R}$	Set of PON groups (i.e. set of racks)
$\mathbb{P}$	$= \mathbb{T} \cup \mathbb{R}$, Set of all communicating vertices
$\mathbb{I}_k$	Set of input ports of AWGR $k; k \in \mathbb{K}$
$\mathbb{O}_k$	Set of output ports of AWGR $k; k \in \mathbb{K}$
$\mathbb{N}$	Set of all vertices (i.e. OLT ports, PON groups, and AWGRs ports) in a cell
$\mathbb{N}_m$	Set of potential neighbours of vertex m; $m \in \mathbb{N}$

Variables:

β_{mn}	Binary variable which is equal to one if vertex m is chosen to be connected with vertex n and is equal to zero otherwise; $m \in \mathbb{N}, n \in \mathbb{N}_m$
χ^{sd}_{jmn}	Binary variable which is equal to one if wavelength j is used in link (m,n) if it is chosen to connect vertex s and vertex d and is equal to zero otherwise; $j \in \mathbb{W}, m \in \mathbb{N}, n \in \mathbb{N}_m, s, d \in \mathbb{P}, s \neq d$
μ^{sd}_j	Binary variable which is equal to one if wavelength j is chosen to connect vertex s and vertex d and is equal to zero otherwise; $j \in \mathbb{W}, s, d \in \mathbb{P}, s \neq d$

follows the flow conservation law [9].

$$\sum_{n \in \mathbb{N}_m} \chi^{sd}_{jmn} - \sum_{n \in \mathbb{N}_m} \chi^{sd}_{jnm} = \begin{cases} \mu^{sd}_j & m = s \\ -\mu^{sd}_j & m = d \\ 0 & otherwise, \end{cases}$$

$$\forall s, d \in \mathbb{P}, s \neq d, m \in \mathbb{N}, j \in \mathbb{W} \tag{4.2}$$

2. Wavelength allocation: Constraint (4.3) ensures that a single wavelength is selected per a communicating pair. Constraint (4.4) ensures that each destination receives from different sources through different wavelengths. Constraint (4.5) ensures that each source transmits to different destinations through different wavelengths [9].

$$\sum_{j\in\mathbb{W}} \mu_j^{sd} \leq 1, \forall s, d \in \mathbb{P}, s \neq d \tag{4.3}$$

$$\sum_{s\in\mathbb{P}, s\neq d} \mu_j^{sd} \leq 1, \forall d \in \mathbb{P}, j \in \mathbb{W} \tag{4.4}$$

$$\sum_{d\in\mathbb{P}, s\neq d} \mu_j^{sd} \leq 1, \forall s \in \mathbb{P}, j \in \mathbb{W} \tag{4.5}$$

3. Routing constraints: Constraint (4.6) ensures that the flow for a connection between any pair is not relayed by any other vertex in $\mathbb{P}$. Constraints (4.7) and (4.8) are for routing within the AWGRs, the first ensures that flows are only directed from input to output ports and the second ensures that each input port in an AWGR sends a different and single wavelength to each output port [9].

$$\sum_{\substack{s,d\in\mathbb{P}, s\neq d \\ n\in\mathbb{N}_i, j\in\mathbb{W}}} \chi_{jin}^{sd} - \sum_{\substack{d\in\mathbb{P}, d\neq i \\ j\in\mathbb{W}}} \mu_j^{id} \leq 0, \forall i \in \mathbb{P} \tag{4.6}$$

$$\sum_{n\in\mathbb{I}_k} \chi_{jmn}^{sd} \leq 0, \forall s, d \in \mathbb{P}, s \neq d, k \in \mathbb{K}, m \in \mathbb{O}_k, j \in \mathbb{W} \tag{4.7}$$

$$\sum_{\substack{s,d\in\mathbb{P}, s\neq d \\ j\in\mathbb{W}}} \chi_{jmn}^{sd} \leq 1, \forall k \in \mathbb{K}, n \in \mathbb{O}_k, m \in \mathbb{I}_k, \tag{4.8}$$

4. Constraint to ensure that flows are routed only between connected communicating vertices selected according to constraints (4.10)-(4.18). Constraint (4.9) is to ensure that the sum of traffic in link (m, n) (i.e. $\sum_{s,d\in\mathbb{P}, s\neq d} \chi_{jmn}^{sd}$), which can maximally equal to one according to constraint (4.8), is equal to zero if β_{mn} is equal zero and allows the sum to equal to one if β_{mn} is equal to one.

$$\sum_{s\in\mathbb{P}, d\in\mathbb{P}, s\neq d} \chi_{jmn}^{sd} \leq \beta_{mn}, \forall m \in \mathbb{N}, n \in \mathbb{N}_m, j \in \mathbb{W} \tag{4.9}$$

5. Constraints to determine the connections of the input and output ports of each AWGR with the OLT ports, PON groups and input and output ports of the other AWGR. Constraint (4.10) is to ensure that each PON group is connected to a single AWGR input port, while Constraint (4.11) is to ensure that each PON group is connected to a single AWGR output port. Constraint (4.12) is to assign a single input port in each AWGR to the connection with the OLT port, while Constraint (4.13) is to assign a single output port in each AWGR to the connection with the OLT port. Constraint (4.14) is to ensure that each *input* port in an AWGR has a unique connection with either a PON group, OLT port, or an *output* port in the other AWGR. Constraint (4.15) is to ensure that each *output* port in an AWGR have a unique connection with either a PON group, OLT port, or an *input* port in the other AWGR. Constraint (4.16) is to ensure that all input and output ports of an AWGRs are internally connected. Constraint (4.17) is to ensure that remaining output ports of each AWGR are connected to the remaining input ports of the other AWGR. Constraint (4.18) is to ensure mutual neighboring between connected vertices.

$$\sum_{k\in\mathbb{K},n\in\mathbb{I}_k} \beta_{mn} \leq 1, \forall m \in \mathbb{R} \tag{4.10}$$

$$\sum_{k\in\mathbb{K},n\in\mathbb{O}_k} \beta_{mn} \leq 1, \forall m \in \mathbb{R} \tag{4.11}$$

$$\sum_{n\in\mathbb{I}_k} \beta_{mn} \leq 1, \forall k \in \mathbb{K}, m \in \mathbb{T} \tag{4.12}$$

$$\sum_{n\in\mathbb{O}_k} \beta_{mn} \leq 1, \forall k \in \mathbb{K}, m \in \mathbb{T} \tag{4.13}$$

$$\sum_{m\in\mathbb{P}\cup\mathbb{O}_q} \beta_{mn} \leq 1, \forall k \in \mathbb{K}, q \in \mathbb{K}, k \neq q, n \in \mathbb{I}_k \tag{4.14}$$

$$\sum_{m\in\mathbb{P}\cup\mathbb{I}_q} \beta_{mn} \leq 1, \forall k \in \mathbb{K}, q \in \mathbb{K}, k \neq q, n \in \mathbb{O}_k \tag{4.15}$$

$$\beta_{mn} = 1, \forall k \in \mathbb{K}, m \in \mathbb{I}_k, n \in \mathbb{O}_k \tag{4.16}$$

$$\sum_{m\in\mathbb{O}_k,n\in\mathbb{I}_q} \beta_{mn} \leq \frac{M}{2} - 1, \forall k \in \mathbb{K}, q \in \mathbb{K}, k \neq q \tag{4.17}$$

$$\beta_{mn} = \beta_{nm}, \forall m \in \mathbb{N}, n \in \mathbb{N}_m \tag{4.18}$$

4.4 Connections and Wavelength Routing and Assignment Results

For four racks, single OLT port and when two 4×4 AWGRs are used, the MILP results for the connections and the wavelength assignment for communication between the racks and the OLT port are provided in Figures 4.6 and 4.7. Figure 4.6 shows the detailed routing of each wavelength, while Figure 4.7 shows only the wavelength continuity from the source to the destination. The assignments are also summarised in Table 4.1.

Table 4.1: MILP obtained results for the wavelengths assignment to OLT ports and PON groups communications in the AWGR-based PON DCN.

	OLT port 1 $AWGR_1, \mathbb{O}_1 = 1$ $AWGR_2, \mathbb{O}_2 = 3$	PON group 1 $AWGR_2, \mathbb{O}_2 = 4$	PON group 2 $AWGR_1, \mathbb{O}_1 = 4$	PON group 3 $AWGR_1, \mathbb{O}_1 = 2$	PON group 4 $AWGR_2, \mathbb{O}_2 = 1$
OLT port 1 $AWGR_1, \mathbb{I}_1 = 3$ $AWGR_2, \mathbb{I}_2 = 3$	-	λ_3 1 hop	λ_2 1 hop	λ_1 1 hop	λ_4 1 hop
PON group 1 $AWGR_1, \mathbb{I}_1 = 2$	λ_2 1 hop	-	λ_3 1 hop	λ_4 1 hop	λ_1 2 hops
PON group 2 $AWGR_2, \mathbb{I}_2 = 4$	λ_1 1 hop	λ_4 1 hop	-	λ_2 2 hops	λ_3 1 hop
PON group 3 $AWGR_2, \mathbb{I}_2 = 1$	λ_3 1 hop	λ_1 1 hop	λ_4 2 hops	-	λ_2 1 hop
PON group 4 $AWGR_1, \mathbb{I}_1 = 1$	λ_4 1 hop	λ_2 2 hops	λ_1 1 hop	λ_3 1 hop	-

4.5 Summary

This chapter introduces two PON-based data centre designs; an AWGR-centric design and a server-centric design that were proposed in [9, 207]. Both designs utilise ports with OLT line cards for inter and possibly intra data centre networking in addition to passive interconnects for the intra data centre networking between different PON groups (i.e. racks) within a PON cell (i.e. number of PON groups connected to a single OLT port). The AWGR-centric design example presented in this chapter allows up to 20 simultaneous connections between different racks and the OLT. If 10 Gbps tuneable transceivers are used in servers, a bisection bandwidth of 200 Gbps is achieved while using 4×4 AWGRs. The server-centric data centre provides a design with better resilient and cost effectiveness compared to the AWGR-centric design. The performance and resilience of these two designs are examined in the work in Chapter 5, and Chapter 6.

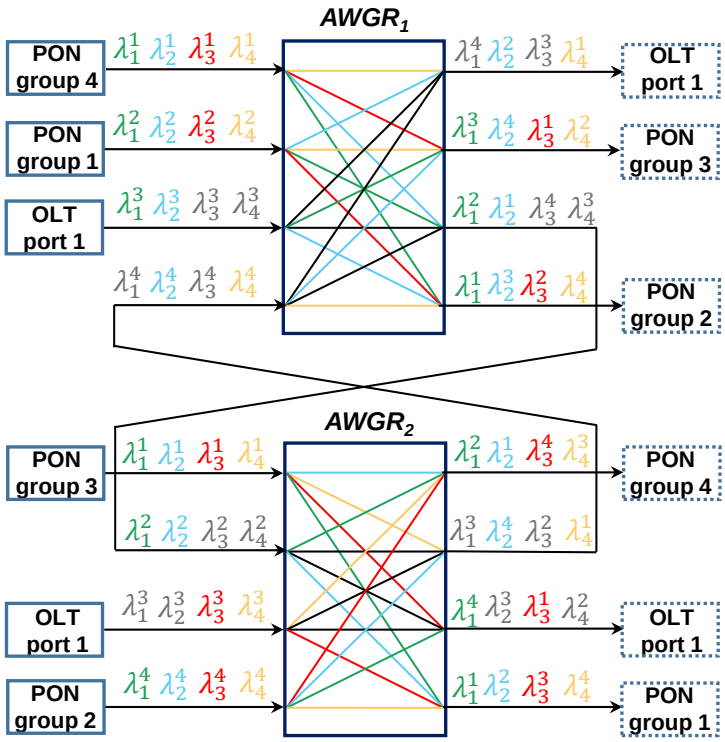


Figure 4.6: MILP results for the wavelength assignments and the connections between the PON groups and OLT port and the ports of the two AWGRs. Rectangles represent input ports of PON groups (i.e. racks) and dashed-line rectangles represent output ports of PON groups.

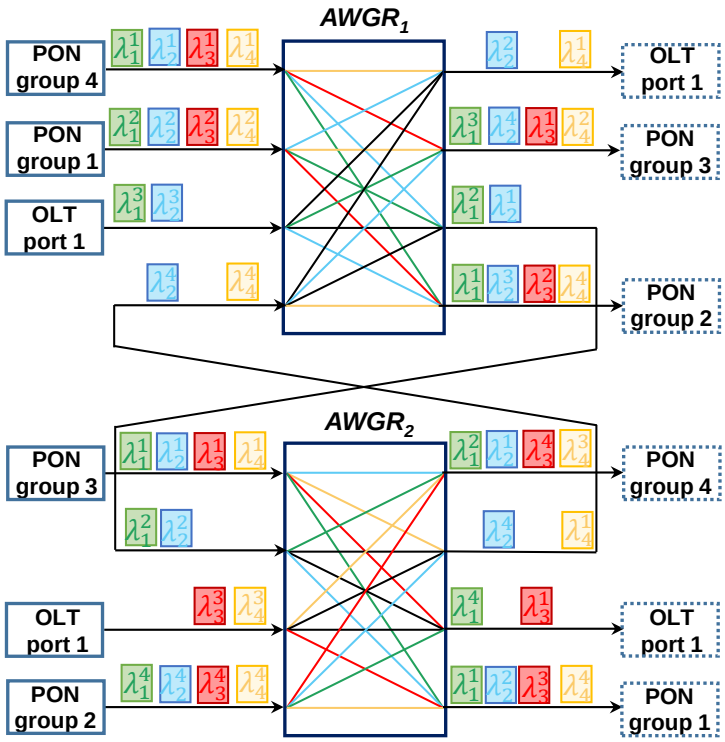


Figure 4.7: MILP results for the wavelength assignments and the connections showing wavelength continuity. Rectangles represent input ports of PON groups (i.e. racks) and dashed-line rectangles represent output ports of PON groups.

Chapter 5

Optimisation of Co-flows Scheduling and Routing in Intra Data Centre Networks for MapReduce

This chapter aims to investigate the optimisation of the scheduling and routing of the co-flows of MapReduce shuffling phase with the objective of minimising either the total energy consumption or the completion time through a MILP model. The study is carried out while considering different DCN architectures and different traffic patterns defined according to the distribution of map and reduce tasks in the servers and data skewness. The remainder of this chapter is organised as follows: Section 5.1 provides some data centre-related optimisation studies for big data applications. Section 5.2 describes the system models used for the considered data centres and the characteristics of the workloads examined. Section 5.3 presents the MILP model, while Section 5.4 provides the results and discussions. Finally, Section 5.5 provides a summary.

5.1 Related Studies

What follows provide some data centre-related optimisation studies to improve the performance or energy efficiency of MapReduce and related big data applications.

5.1.1 Data Centre Topology

Evaluating the performance and energy efficiency of big data applications in different data centres topologies was considered in [3, 219–223]. The authors in [219] modelled Hadoop clusters with up to four ToR switches and a core switch to measure the influence of the network on the performance. A simulator, MRPerf, was utilised in [220] to study the effect on Hadoop performance of several parameters related to clusters (e.g. CPU, RAM, and disk resources), configurations (e.g. chunk size, number of map and reduce slots), and framework (e.g. data placement and task scheduling). DCell was compared to double-rack clusters with 72 nodes under the assumptions of 1 replica and no speculative execution and was found to improve sorting by 99%. The authors in [221] estimated the completion time of jobs in different topologies with different workload distributions compared to a hypothetically optimal topology for MapReduce with a dedicated link for each intermediate data shuffling flow. Different levels of intermediate data skew were also examined and worse performance was reported for all topologies. In [3], we examined the effects of the topology on the performance and energy efficiency of MapReduce shuffling for sort workloads in different electronic, hybrid and all-optical switching data centres and different rate-per-server values. The results indicated that optical switching technologies achieved an average power consumption reduction by 54% compared to electronic switching data centres with comparable performance. In [222], the Network Power Effectiveness (NPE) defined as the ratio of the aggregate throughput to the power consumption was evaluated for six electronic switching data centre topologies under regular and energy-aware routing. The power consumption of the switches, the server's NIC ports and CPU cores used to process and forward packets in server-centric topologies were considered. Design choices such

as link speeds, oversubscription ratio, and buffer sizes in spine and leaf architectures with realistic web search queries were examined by simulations in [223].

Several big data frameworks utilised the topology in the computations as in [224, 225]. Camdoop in [224] is a MapReduce-like system that run in CamCube and exploits its topology by aggregating the intermediate data along the path to reduce workers. In [225], the utilisation of existing or attached networking hardware was proposed to improve the performance of query applications. The topologies of data centres were also considered in optimising VM assignments as in [226, 227]. A Traffic-aware VM Placement Problem (TVMPP) and its solution were proposed in [226] to improve the scalability of data centres. TVMPP follows two-tier approximating algorithm that leverages knowledge of traffic demands and the data centre topology to co-allocate VMs with heavy traffic in nearby hosts. The work in [227] proposed *Oktopus* to tackle intra data centre network performance variability in multi-tenant data centres. The results showed that allocating VMs while accounting for the oversubscription ratio improved the completion time and reduced tenant costs by up to 75% while maintaining the revenue.

The performance of big data applications in SDN-controlled electronic and hybrid electronic/optical switching data centres was considered in [228–230]. To evaluate the impact of networking configurations on the performance of big data applications in SDN-controlled data centres with multi-racks before deployments, a Flow Optimised Route Configuration Engine (FORCE) was proposed in [228]. To address big data applications need for frequent reconfigurations, the work in [229] examined a ToR-level SDN-based topology modification in a hybrid data centre with core MEMS switch and electrical Ethernet-based switches at run-time. The work in [230] experimentally examined the performance of MapReduce in two hybrid electronic/optical switching data centres namely c-Through and Helios. An "observe-analyse-act" control framework was utilised for the configurations of the OCS and the packet networks. The authors discussed the challenges and emphasised the need for near real-time analysis of application requirements to optimally obtain hybrid switching scheduling decisions.

5.1.2 Data Centre Routing

In [231], a reduce tasks placement problem was analysed in multi-rack environments to decrease cross-rack traffic based on two greedy approaches and up to 32% speedup in completion time was achieved. A scalable DCN-aware load balancing technique for key distribution and routing in the shuffling phase of MapReduce was proposed in [232] while considering bandwidth constraints and addressing data skewness. To improve shuffling under varying data sizes and data reduction ratios, a joint intermediate data partitioning and aggregation scheme was proposed in [233]. A decomposition-based distributed online algorithm was proposed to dynamically adjust data partitioning by assigning keys with larger data sizes to reduce tasks closer to map tasks. To effectively use the bandwidth in BCube data centres, the work in [234] proposed and optimised two schemes for in-network aggregation at the servers and switches.

The energy efficiency of routing big data applications traffic was considered in [235, 236]. In [235], preemptive flow scheduling and energy efficient routing were combined to improve the utilisation in Fat-tree data centres. To improve the energy efficiency of MapReduce-like systems, the work in [236] examined combining VM assignments with traffic engineering and total average savings by 60% in Fat-tree, and 30% in BCube data centres were achieved. SDN-based solutions that optimise the routing of big data applications traffic were discussed in [237–239]. To improve the routing of shuffling traffic in Fat-tree, an application-aware SDN routing scheme was proposed in [237]. The results indicated a reduction in the shuffling time by 20% and 10% compared to Round Robin-based ECMP under no skew, and with skew, respectively. To enhance the shuffling between map and reduce VMs under background traffic, the work in [238] suggested dynamic flows assignment to queues with different rates. The results showed that prioritising Hadoop traffic and providing more bandwidth to straggler reduce tasks improved the completion time by 42% compared to solely using a 50 Mbps queue. XPath was proposed in [239] to allow applications to explicitly route their flows without the overheads of dynamic establishment of paths in routing tables. For MapReduce shuffling, XPath achieved 3× completion time reduction compared to ECMP.

5.1.3 Scheduling of Flows, Coflows, and Jobs in Data Centres

Scheduling big data traffic at the flow level was addressed in [178], and at the co-flow level which is more applications-aware in [179, 180, 240, 241]. *Orchestra* was proposed in [178] as a task-aware centralised cluster manager to reduce the average completion time for batch, iterative, and interactive workloads. *Varys* was proposed in [179] as a coordinated inter-coflow scheduler in data centres targeting predictable performance. A greedy co-flow scheduler and a per-flow rate allocator were utilised and trace-driven simulations indicated that Varys achieved 3.66×, 5.53×, and 5.65× improvements compared to fair sharing, per-flow scheduling, and FIFO, respectively. The authors in [180] proposed *Baraat* which is a decentralised task-aware scheduling mechanism for co-flows to reduce their tail completion times. Compared to pFabric [175] and *Orchestra*, the completion time of 95% of MapReduce workloads was reduced by 43% and 93%, respectively. A decentralised coflow-aware scheduling system that dynamically sets the priorities of flows according to the maximum load was proposed in [240] and outperformed Baraat by 1.4 and 4 times for homogeneous and heterogeneous workloads, respectively. *Rapier* in [241] integrated routing and scheduling at the coflow level in DCNs with commodity switches. Compared to Varys with ECMP and optimised routing only, about 80% and 60% improvement in coflow completion time was achieved.

Scheduling traffic in DCNs with SDN environments was addressed in [242, 243]. *Pythia* in [242] focused on improving the network performance under skewed workloads. A run-time intermediate data size prediction and a centralised controller were utilised and up to 46% improvement in completion time was obtained compared to ECMP. The authors in [243], proposed and experimentally demonstrated a heuristic for Bandwidth-Aware Scheduling with SDN (BASS) to reduce the minimum job completion time in Hadoop clusters. The heuristic prioritises scheduling the tasks locally but considers remote assignment. The works in [144, 244, 245] focused on scheduling traffic in optical and hybrid DCNs. In [144], the gaps between evaluating OCS interconnects performance and latency-sensitive applications requirements were addressed. A centralised Static Circuit Flexible Topology (SCFT) algorithm and a distributed

Flexible Circuit Flexible Topology (FCFT) algorithm were proposed and up to 2.44× improvement was achieved over Mordia [170]. Resource allocation in NEPHELE was addressed in [244] while accounting for the SDN controller delay and random allocation of iterative MapReduce tasks. Compared to Mordia, NEPHELE uses multiple WDM rings and introduces an application-aware and feedback-based synchronous slotted scheduling algorithms. Effective traffic scheduling for a proposed Packet-Switched Optical Network (PSON) with space switches and layers of AWGRs was examined in [245]. Scheduling algorithms that consider priority of flows and occupancy of buffers were proposed and reduction in packet loss ratio and average delay compared to Round Robin were reported.

The energy efficiency of data centres through workloads and traffic scheduling was considered in [195, 246, 247]. The work in [195] examined the performance-energy tradeoffs when using the Low Power Idle (LPI) link sleep mode of the EEE standard [194] with MapReduce workloads. The timing parameters of entering and leaving the LPI mode in 10GbE links were optimised while utilising packet coalescing (i.e delaying outgoing packets during the quite mode and aggregating them for transmission in the following active mode). Depending on the superscription ratio and workloads, EEE achieved power saving between 5 and 8 times compared to legacy Ethernet. *Willow* in [246] aimed to reduce switches energy consumption in Fat-tree through SDN-based dynamic scheduling. Compared to ECMP and simulated annealing and particle swarm optimisation-based heuristics, up to 60% savings were achieved. The authors in [247] proposed *JouleMR* as a green-aware and cost-effective tasks and jobs scheduling framework for MapReduce workloads that maximised renewable energy usage while accounting for brown energy dynamic pricing. Compared to Hadoop with YARN, a reduction by 35% in non-renewable energy usage and by 21% in overall energy consumption was obtained.

5.2 System Model and Parameters

To quantitatively assess the impact of the data centre topology and workloads characteristics on the completion time and the energy efficiency of MapReduce shuffling operations, a MILP model that minimises either the completion time or the total energy consumption while optimising the scheduling and routing of co-flows was developed. The problem of optimising the routing of co-flows with known sources and destinations through a capacitated network that performs shuffling operations can be categorised as a Multi-Commodity Flow (MCF) problem which is NP-complete, but can be solved with solvers as CPLEX. The MILP model developed contains additional constraints to model the routing requirements of each data centre topology. Also, the model can be considered time-slotted as a discrete time dimension is introduced in the variables to account for the scheduling of flows or the remainder of flows in the following scheduling time slots at the granularity of a second or less. Moreover, as the workloads characteristics are highly coupled with the generated traffic in the data centre, the impact of the intermediate data skewness on the performance and energy efficiency of the routing and scheduling of the shuffling co-flows is also examined. The following Subsections provide the data centre models and the workloads modelling, in addition to the parameters and assumptions considered.

In this work, we optimise the routing and scheduling for pre-allocated map and reduce tasks. Although optimising the tasks placement to improve the data locality at different stages of MapReduce can improve the performance and energy efficiency in lightly loaded data centres, with larger data sizes and larger data centre scales, it becomes harder to maintain locality for all tasks. Hence, we present the evaluation for the data centre topology impact under random tasks allocation which also complies with the random allocation in native unmodified frameworks such as Hadoop [47]. Also, the comparison between the data centres was not performed under similar bisection bandwidth or network diameter (i.e. number of hops between servers) as the work in [222]. Instead, we compared the performance and energy consumption required to

shuffle the same amount of data when placing the map and reduce tasks in a fixed number of server (i.e. 16 servers interconnected with different architectures). The data centres are compared based on available technologies such as unifying the maximum data rate per transponder per wavelength to 10 Gbps while using the most suitable commodity hardware required for each architecture as will be detailed in the following Section.

5.2.1 Data Centre Models

Six DCN topologies, depicted in Figures 5.1, and 5.2 are considered. Those are Fat-tree, Spine-leaf, BCube, and DCell as electronic switching DCNs, in addition to the two PON-based DCNs introduced in Chapter 4. Each data centre is modelled as a graph with vertices in the set $\mathbb{G}$ including the servers and the switches. The topology of each data centre is defined by a neighbouring set ($\mathbb{G}_u$), where u is a vertex in $\mathbb{G}$. Depending on these two sets, the edges of the graph are defined where each edge, denoted as (u, v), represents a link between vertex u and v, where $u, v \in \mathbb{G}$. The capacity of each edge per a wavelength is selected to be 10 Gbps for all topologies. The power consumption and details of the electronic and optical equipment used in each topology are summarised in Table 5.1. The number of servers needed to accommodate map or reduce tasks is set to 16 to enable comparison between the different data centre architectures. To accommodate 16 servers, a Fat-tree network [132] with $k = 4$ is sufficient. Such a Fat-tree requires a total of $3/4k^3 = 48$ links as modelled in Figure 5.1(a). For BCube [136], 16 servers can be connected in a $k = 1, n = 4$ configuration where a BCube_0 is composed of a 4-port switch and 4 servers and the BCube_1 is composed of 4 BCube_0 units and additional four 4-port switches as depicted in Figure 5.1(c). As a DCell_k must be constructed recursively from $n - 1$ DCell_{k-1} units, where n is the number of servers in each DCell_0 [137], the best configuration to connect 16 servers is a DCell_1 with five DCell_0 each with 4 servers. To obtain results comparable to other topologies, the additional four servers are not assigned any additional tasks but can be used for the routing. The remaining topologies provide more flexibility with the number of servers

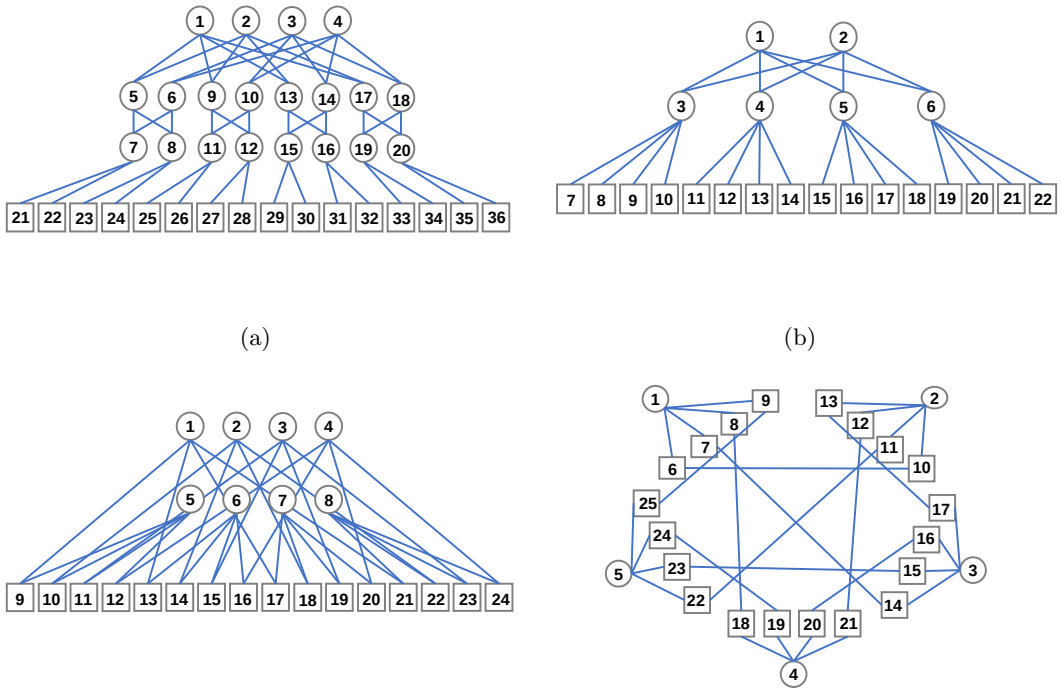
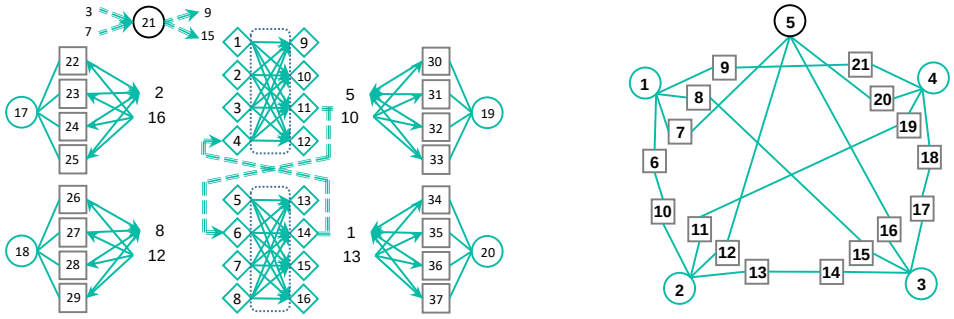


Figure 5.1: Electronic DCNs graph representation. Squares and circles represent servers and switches, respectively. Blue edges represent EPS bidirectional links. (a) Fat-tree, (b) Spine-leaf, (c) BCube, and (d) DCell.

in a rack and were configured as in Figure 5.1.

A 1 rack unit form-factor Cisco switch, model 3524X, was used as the switch in the spine-leaf DCN. It is a 10 Gbps Ethernet switch that has 24 ports providing 480 Gbps switching capacity and a total of 18 MB memory for the packet buffers [248]. For the electronic Top-of-Rack (ToR) switches in the remaining data centres, SG500XG-8F8T, with eight 10 fiber-based Gbps ports and eight copper-based 10 Gbps ports, was used [249]. The SG500XG-8F8T switch has a total of 16 MB memory for the packet buffers and provides a total of 320 Gbps switching capacity. In the servers of switch-centric topologies and in the ports of electronic switches, 10 Gbps Enhanced Small Form Factor Pluggable (SFP+) transceivers with power consumption of 1 Watt



(a) (b)

Figure 5.2: PON-based DCNs graph representation. Squares and circles represent servers, and switches, respectively. Cyan edges represent bidirectional links, while triple-dashed edges represent WDM links. Diamonds represent AWGR ports and rounded dashed rectangles represent AWGRs (a) PON3, and (b) PON5.

are considered [250]. In the servers of server-centric topologies (i.e. BCube, DCell, and PON5), PE10G2T-SR which is a commodity 10 Gbps Network Interface Card (NIC) from Broadcom was considered [251]. PE10G2T-SR is based on short range Fiber (IEEE standard 802.3ae) connections and contains two 10 Gbps ports that can maximally provide a total of 18.7 Gbps capacity per port due to host and protocol overheads. We considered that a server consumes 0.07 W to process 1 Gbps of traffic (i.e. minimal overhead). However, the power consumption can reach a value of up to 14 W to process the same amount of data [251].

Two rates (ρ) are considered for the transmission rate from each server. These rates are coupled with the data read speed from disks, memory, or caches to the transceiver or the NIC of the server. The 2.8 Gbps rate matches the read speed of the MapReduce framework. It is used to read the results of map workers from Solid-State Drives (SSD) of the server before sending to the network. A rate of 8 Gbps can be considered if the framework uses memory, optimises the use of Redundant Array of Independent Disks (RAID), or if it has caching capabilities in NICs. In this work, the completion time (M) calculation is done by only considering the transmission delay that results from the optimum routing and scheduling of the flows. Generally, there are four types of delay

in communication networks which are the propagation delays, transmission delays, processing delays, and queuing delays [252]. The first is mostly related to the speed of light in fiber links and can be ignored in DCN environments as the length of cables is typically small (i.e. less than 10 km) and the delay is estimated as 5μs per km [206]. The second is related to the capacity of links and the size of the transmitted packets and is considered dominant for elephant flows. The processing delay is related to the control overheads in the switches and NICs and depends on their CPU and RAM resources. The last delay (i.e. queuing delay) is determined by the limited buffer sizes in switches, the processing speed per packet, and the rate of packets arrival to each switch and can be estimated according to queuing theory which is complex in multi-path, and multi-hop connections in DCNs. The authors in [223] studied the performance of applications in Spine-leaf DCNs while assuming that leaf switches are ideal non-blocking switches and that the spine switch is a one large output-queued switch with infinite capacity. The authors in [223] assumed that the switches have unlimited buffer space and modelled them as shared-memory output queued switches that do not drop packets. Based on measurements, the processing latency in leaf switches was found to be 700 ns, while at spine switches 2 μs. The host networking stack added a 10 μs delay resulting in a total RTT of 50 μs. In studies that optimise the routing and scheduling of large flows, the processing, propagation, and queuing delays can be considered negligible compared to the transmission delay [253].

For the AWGR-based data centre (i.e. PON3), we considered the design presented in Figure 4.6. The corresponding system model is depicted in figure 6.1(b). PON group (i.e. rack) 1 contains servers 22, 23, 24, and 25, while the forth group contains servers 34, 35, 36, and 37. The OLT WDM port is in node 17, while nodes 1 to 16 are for the ports of the two AWGRs. The power consumption of the OLT port is estimated by considering a single Ethernet card in the OLT in [217]. The power consumption required to operate the OLT was estimated to be 187 W by considering the maximum power consumption values for the power cards, common interface card, and the switching and control cards. The maximum power consumption of a single Ethernet interface card,

Table 5.1: Data centre-related parameters

Topology	No of servers	No of Switches	No of links	Wavelengths use $\mathbb{W}$	Networking Devices Characteristics		
					Equipment	No	$O_{i(max)}$ Watts
Fat-tree [132]	16	20	48	Grey (colorless)	SG500XG-8F8T [249]	20	94.33
Spine-leaf [134]	16	6	24	Grey (colorless)	Nexus 3524X [248]	6	193
BCube [136]	16	8	32	Grey (colorless)	SG500XG-8F8T [249]	8	94.33
					PE10G2T-SR† [251]	16	14
DCell [137]	20*	5	30	Grey (colorless)	SG500XG-8F8T [249]	5	94.33
					PE10G2T-SR† [251]	20	14
PON3 [207, 208]	16	7	64*‡	WDM	4×4 Polymer back-plane	4	12
					OLT with one card [217]	1	217
					4×4 AWGR	2	0
PON5 [207, 213]	16	5	23	Grey (colorless)	4×4 Polymer back-plane	4	12
					OLT with one card [217]	1	217
					PE10G2T-SR† [251]	16	14

*The number of servers is kept 20 as it is a design scale requirement but workloads are allocated only in 16 servers, * Excluding internal AWGRs links, ‡ directional. ** 0.24 Watts per port [150]

which has 10G optical modules is 30 W. A tuneable transceiver per server is required for the connections with the AWGRs and OLT port through the AWGRs. We considered SFP-10GDWZR-TC [254] which is a dual fiber 10 Gbps Tunable DWDM transceivers. SFP-10GDWZR-TC consumes a maximum of 2 Watts and has span of 80 km which is more than sufficient in data centre environments. A Tuneable transceiver can only transmit at a single wavelength at a time, but can receive at multiple wavelengths if a wide band filter and appropriate receiver design and network interface are used. We considered a Field-Programmable Gate Array (FPGA)-based Network Interface card which has a power consumption between 11-12.3 W [255]. For the design presented in Figure 4.6, the servers can communicate with other servers in the rack only through an optical backplane. For the optical backplane connections we considered additional grey transceivers in the servers with total power consumption of 12 W per rack.

The server-centric PON-based design, PON5, is assumed to have a TDM connection with the OLT, hence the 4 connected servers are to share this link using a single grey wavelength. One cell of the server-centric PON-based DCN design was considered to have 4 servers per rack and a total of 4 racks in a single cell [6]. We considered

PE10G2T-SR [251] for the NICs in the servers, in addition to the OLT and optical backplane equipment as in PON3.

5.2.2 MapReduce Shuffling Traffic Modeling

In this evaluation, we considered a scenario where ten servers are dedicated for map tasks and six different servers are dedicated for reduce tasks. This configuration resembles a typical tasks ratio in the original Google's MapReduce [17]. The placement of map and reduce workers was randomly generated for all the topologies. To effectively examine network bottlenecks, sort workloads are considered. Sorting via MapReduce utilises identity map functions to generate $< word, 1 >$ pairs from large text files. The entire intermediate data is to be shuffled according to words (i.e. keys) to reduce workers in order to be sorted and finally saved. Hence, input, intermediate, and output data are all equal in size. The volume of total data to be sorted is varied from 1 Gbits to either 60 Gbits or 120 Gbits. A total of equivalent data is to be shuffled and transferred from map tasks to reduce tasks. We omit the details of assigning the data to individual tasks in each server and considered the traffic to be shuffled from a server containing several map tasks to one of the servers containing several reduce tasks as a single flow. This generates a total of 60 flows in the data centre network. Beside the shuffling traffic, DFS data transfers and control messages (i.e. heartbeats) are also required for the MapReduce framework. Also, as input data placement is not deterministic in most of MapReduce-based frameworks and as the locality for map input data cannot be always ensured, a step before starting the map phase may include DFS input data transfers. Also, HDFS final output data write can be assigned in servers different than the ones that are assigned to reduce tasks which will require additional transmission at the end of the MapReduce job. In this work we only considered the shuffling traffic and for simplicity, we assume that all map tasks finish at the same time and hence, all data is ready for transmission at the beginning of the shuffling phase. Such configuration can be realised by modifying the slow start option whose default configuration in Hadoop enables the shuffling to start when 3% of the map task output is ready [47].

We considered two cases for the flows sizes distribution. In the first case, the results of all map tasks are of equal size, and hence, they generate equal size flows. Such workload is an equivalent to the Indy GraySort benchmark which has uniform intermediate key distributions due to balanced words count [256]. The second case considers uneven map task output sizes, which is equivalent to the Daytona GraySort benchmark [257]. The map output file sizes were generated randomly through a uniform distribution-based random generator with values that range between zero to the size of the total shuffling data volume. To ensure that the total sum of the randomly generated flow sizes (i.e. map output sizes) is maintained, proper scaling was performed. Figure 5.3 shows the range of the skewed flow sizes as an error bar at each value of the total shuffling data volume.

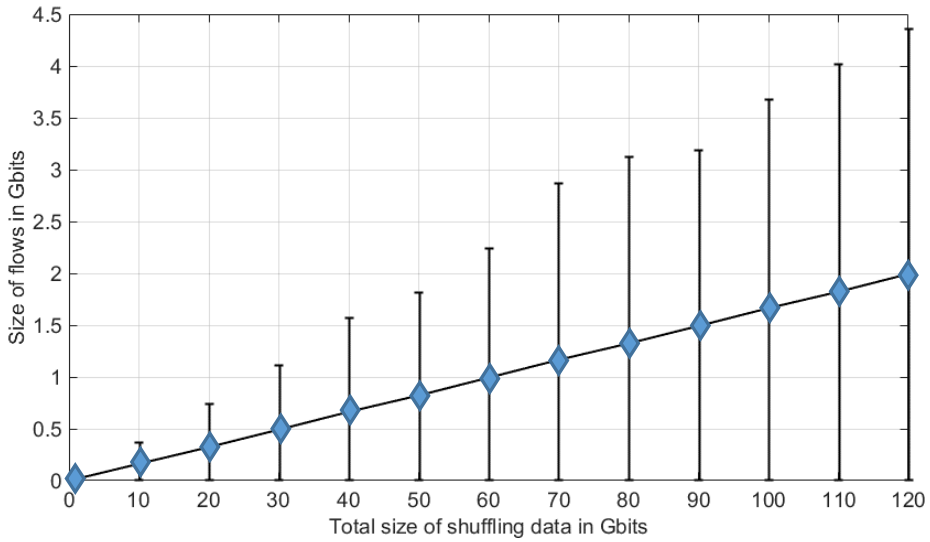


Figure 5.3: Size of shuffling flows in Gbps. Blue diamonds represent the uniform flow size with no data skew.

5.3 MILP Model for Optimising the Co-flows Scheduling and Routing for MapReduce Traffic

Ideally, in a MCF problem and based on the source, destination, and file size information of the flows in a co-flow, a MILP model is to determine the path and the data rate for each source. Then, the largest ratio of data to be sent divided by the data rate value will determine the largest completion time required to transmit the total co-flow. In the following model, we alternatively considered a time-slotted approach, which allow the scheduling of the flows fully or partially in one of the time slots and in one of the routes between the servers containing the map and reduce tasks. In this case, the most congested link in the last used time slot will determine the maximum transmission time of its flow, and hence will determine the completion time of the co-flow.

By finding the most congested link, we have identified the link with the largest ratio of data to be send over that link divided by the data rate of that link. Hence this is equivalent to replacing the multiple sources that send over this link by one source that has data to be send equal to the sum of the data to be send by each individual source, hence identifying the largest ratio. Based on the determined schedule and routing, the model also calculates the energy consumption based on the power consumption values of the networking equipment and the duration of using each element in the network while assuming that it will be on during the time slots it is utilised in.

In what follows, a developed MILP model for this optimisation problem is described. This model takes the topology of the data centre (i.e. the connections and the capacity of links), the power consumption for all equipment, and the total shuffling traffic as input and provides the schedule (i.e. time slot and amount of assigned traffic for each link), the completion time, and the total energy consumption. This is obtained while considering the architectural and routing constraints under one of two objectives which are to reduce the energy consumption or to reduce the completion time. The parameters and variables of the MILP model are provided below. All variables are set to be less than or equal to zero. The sets are represented as double-lined letters while the small

letters in the subscripts and postscripts indicate the indices of parameters or variables.

Sets and parameters:

$\mathbb{G}$	Set of all vertices (servers and switches) in the data centre
$\mathbb{G}_u$	Set of neighbors of vertex; $u \in \mathbb{G}$
$\mathbb{R}$	Set of servers in the data centre ($\mathbb{R} \subset \mathbb{G}$)
$\mathbb{S}$	Set of switches in the data centre ($\mathbb{S} \subset \mathbb{G}, \mathbb{R} \cap \mathbb{S} = \varnothing$)
$\mathbb{W}$	Set of wavelengths
$\mathbb{T}$	Set of time slots
D	The duration of a time slot (in seconds)
Δ_{sd}	The total shuffling traffic to be transmitter from server s to server d; $s, d \in R$ (in Gbits)
C_{uvw}	Capacity of link (u,v); $u, v \in G$, at wavelength $w \in W$ (in Gbps)
$P_{i(max)}$	The maximum power consumption of a transceiver in server i; $i \in \mathbb{R}$ or switch i; $i \in \mathbb{S}$ or an NIC in server i; $i \in \mathbb{R}$ (in Watts)
ϵ	The server performance per Watt in servers with an NIC (in Watt per Gbps)
$O_{i(max)}$	The maximum power consumption of switch i; $i \in \mathbb{S}$ (in Watts)
ρ	The maximum rate per server (in Gbps)
σ	The maximum rate per switch (in Gbps)
L	A very large number
Q	A weighting factor

The power consumption of the transceiver in server i at wavelength w and time slot t with an ON/OFF power profile is equal to:

$$\theta_{iwt} = B_{iwt}\, P_{i(max)}, \qquad \forall i \in \mathbb{R},\, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.1}$$

The power consumption of an NIC in server i at wavelength w and time slot t is equal to:

$$\theta_{iwt} = B_{iwt}\, P_{i(max)} + \epsilon\, \beta_{iwt}, \qquad \forall i \in \mathbb{R},\, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.2}$$

Variables:

M	Completion time which is equal to the time when the last transmission ends
E	The total energy consumption
B_{iwt}	Binary variable which is equal to one if the transceiver/NIC of server i is used at wavelength w and time slot t and is equal to zero otherwise; $i \in \mathbb{R}, w \in \mathbb{W}, t \in \mathbb{T}$
A_{iwt}	Binary variable which is equal to one if switch i is used at wavelength w and time slot t and is equal to zero otherwise; $i \in \mathbb{S}, w \in \mathbb{W}, t \in \mathbb{T}$
Γ_{uvwt}	Binary variable which is equal to one if link (u, v) is used at wavelength w and time slot t and is equal to zero otherwise; $u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}$
Z_{uvwt}	Binary variable which is equal to one at the link (u, v), wavelength w, and time slot t where M occurs (i.e. the last used link) and is equal to zero otherwise; $u \in \mathbb{G}, v \in \mathbb{G}_u$, $w \in \mathbb{W}, t \in \mathbb{T}$
χ^{sd}_{uvwt}	Traffic in link (u, v) that contributes to the shuffling data flow to be transmitted from server s to server d at wavelength w and time slot t; $s, d \in \mathbb{R}, s \neq d, u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}$
ψ_{uvwt}	Total traffic in link (u, v) at wavelength w and time slot t; $u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}$
β_{iwt}	The total ingress and egress traffic of the transceiver/NIC of server i at wavelength w and time slot t; $i \in \mathbb{R}, w \in \mathbb{W}, t \in \mathbb{T}$
α_{iwt}	The total ingress and egress traffic of switch i at wavelength w and time slot t; $i \in \mathbb{S}, w \in \mathbb{W}$, $t \in \mathbb{T}$
δ_{sdt}	Traffic for shuffling data flows from server s to server d selected to be transmitted at time slot t
θ_{iwt}	Power consumption of the transciever of server i at wavelength w and time slot t; $i \in \mathbb{R}, w \in \mathbb{W}, t \in \mathbb{T}$
ϕ_{iwt}	Power consumption of switch i at wavelength w and time slot t; $i \in \mathbb{S}, w \in \mathbb{W}, t \in \mathbb{T}$
Ω_{uvwt}	The earliest possible completion time of flow ψ_{uvwt} that starts at time slot t and is routed over link (u, v) at wavelength w; $u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}$, $t \in \mathbb{T}$
τ_{uvwt}	Variable that has the same definition as Ω_{uvwt} with the exception that it takes a value of zero if the link (u, v) is inactive; $u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}$, $t \in \mathbb{T}$

The power consumption of switch i at wavelength w and time slot t is equal to:

$$\phi_{iwt} = A_{iwt}\, O_{i(max)}, \qquad (5.3)$$
$$\forall i \in \mathbb{S},\ w \in \mathbb{W}, t \in \mathbb{T}.$$

The total energy consumption is equal to:

$$E = D \left[\sum_{i \in \mathbb{R}, w \in \mathbb{W}, t \in \mathbb{T}} \theta_{iwt} + \sum_{i \in \mathbb{S}, w \in \mathbb{W}, t \in \mathbb{T}} \phi_{iwt} \right]. \tag{5.4}$$

The model can have one of the following two objectives. The first objective is to minimise, E, the total energy consumption which is expressed as:

$$\min \left[E + Q \sum_{s,d \in \mathbb{R}, t \in \mathbb{T}, s \neq d} (t\, \delta_{sdt}) \right], \tag{5.5}$$

and the second objective is to minimise, M, the latest completion time of shuffling which is expressed as:

$$\min \left[M + Q \sum_{s,d \in \mathbb{R}, t \in \mathbb{T}, s \neq d} (t\, \delta_{sdt}) \right]. \tag{5.6}$$

The second term in both objectives is to schedule the flows in the earliest time slots as possible (i.e. encourage the use of first slots) and therefore the optimisation is not skewed by large files which takes the longest to be transmitted, hence causing the model to possibly send small files late (i.e. near completion time of the largest file). This term in effect improves the fairness for small files when large files are also present. Both objectives are to be calculated under the following capacity and architectural constraints in data centres:

1. Flow conservation in the data centre: The allocation of the links to the flows follows the flow conservation law at each time slot t and wavelength w:

$$\sum_{v \in \mathbb{G}_u} \chi^{sd}_{uvwt} - \sum_{v \in \mathbb{G}_u} \chi^{sd}_{vuwt} = \begin{cases} \delta_{sdt} & u = s \\ -\delta_{sdt} & u = d \\ 0 & otherwise, \end{cases} \tag{5.7}$$

$$\forall s, d \in \mathbb{R}, s \neq d, u \in \mathbb{G}, w \in \mathbb{W}, t \in \mathbb{T}.$$

2. Constraint to ensure that the total egress traffic from a server does not exceed the maximum rate per server at each time slot t:

$$\sum_{v \in \mathbb{G}i, w \in \mathbb{W}} \psi_{ivwt} \leq \rho; \forall i \in \mathbb{R}, t \in \mathbb{T}. \tag{5.8}$$

3. Constraint to ensure that the total ingress traffic of a switch does not exceed the maximum allowed rate per switch at each time slot t:

$$\sum_{u \in \mathbb{G}_i, w \in \mathbb{W}} \psi_{uiwt} \leq \sigma; \forall i \in \mathbb{S}, t \in \mathbb{T}. \tag{5.9}$$

4. Constraint to ensure that the total traffic for shuffling data flows in link (u, v) at wavelength w and time slot t does not exceed its capacity:

$$\psi_{uvwt} \leq D\, C_{uvw}; \forall u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.10}$$

5. Constraint to calculate ψ_{uvwt} by summing the traffic for all shuffling data flows between all servers that pass through link (u, v) at wavelength w and time slot t:

$$\psi_{uvwt} = \sum_{s,d \in \mathbb{R}, s \neq d} \chi^{sd}_{uvwt}; \forall u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.11}$$

6. Constraint to ensure that the sum of shuffling data flow sizes to be send from server s to server d in all time slots is equal to the total flow size:

$$\sum_{t \in \mathbb{T}} \delta_{sdt} = \Delta_{sd}; \forall s, d \in \mathbb{R}, s \neq d \tag{5.12}$$

7. Constraints to find which transceivers/NICs are used (i.e. B_{iwt} is equals to one only if $\beta_{iwt} >$ zero and is equal to zero otherwise):

$$\beta_{iwt} = \sum_{v \in \mathbb{G}_u} \psi_{ivwt} + \sum_{u \in \mathbb{G}_u} \psi_{uiwt}, \quad \text{and} \tag{5.13}$$

$$L\, \beta_{iwt} \geq B_{iwt}, \quad \text{and} \tag{5.14}$$

$$\beta_{iwt} \leq L\, B_{iwt}; \forall i \in \mathbb{R}, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.15}$$

8. Constraints to find which switches are used (i.e. A_{iwt} is equal to one only if $\alpha_{iwt} >$ zero and is equal to zero otherwise):

$$\alpha_{iwt} = \sum_{v \in \mathbb{G}_u} \psi_{ivwt} + \sum_{u \in \mathbb{G}_u} \psi_{uiwt}, \quad \text{and} \tag{5.16}$$

$$L\,\alpha_{iwt} \geq A_{iwt}, \quad \text{and} \tag{5.17}$$

$$\alpha_{iwt} \leq L\,A_{iwt}; \forall i \in \mathbb{S}, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.18}$$

9. Constraints to find if link (u, v) is active (i.e. $\Gamma_{uvwt} =$ one only if $\psi_{uvwt} >$ zero and is equal to zero otherwise):

$$L\,\psi_{uvwt} \geq \Gamma_{uvwt}, \quad \text{and} \tag{5.19}$$

$$\psi_{uvwt} \leq L\,\Gamma_{uvwt}; \forall u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.20}$$

10. Constraints to find the transmission time in link (u, v) at wavelength w if it is used up to time slot t and is active at it:

$$\Omega_{uvwt} = D\,(t-1) + \frac{\psi_{uvwt}}{C_{uvw}}, \quad \text{and} \tag{5.21}$$

$$\tau_{uvwt} \leq L\,\Gamma_{uvwt}, \quad \text{and} \tag{5.22}$$

$$\tau_{uvwt} \leq \Omega_{uvwt}, \quad \text{and} \tag{5.23}$$

$$\tau_{uvwt} \geq \Omega_{uvwt} - L\,(1 - \Gamma_{uvwt}), \quad \text{and} \tag{5.24}$$

$$\forall u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}.$$

11. Constraints to calculate M, which is the completion time determined by the calculated transmission time at the last used link:

$$M \geq \tau_{uvwt}, \quad \text{and} \tag{5.25}$$

$$M \leq \tau_{uvwt} + L\,[1 - Z_{uvwt}], \text{and} \tag{5.26}$$

$$\forall u \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T},$$

$$\sum_{i \in \mathbb{G}, v \in \mathbb{G}_u, w \in \mathbb{W}, t \in \mathbb{T}} Z_{uvwt} = 1. \tag{5.27}$$

For PON3, the following additional set and constraints are required:

Sets and parameters:

$\mathbb{I}$ Set of input ports of the two AWGRs

12. Constraint to ensure that servers do not forward the traffic of other servers:

$$\sum_{u\in\mathbb{R},v\in\mathbb{G}_u,u\neq s} \chi_{uvwt}^{sd} \leq 0, \tag{5.28}$$

$$\forall s \in \mathbb{R}, d \in \mathbb{R}, w \in \mathbb{W}, t \in \mathbb{T}, s \neq d.$$

13. Constraint to ensure that each server transmits only at one wavelength w in a given time slot t:

$$\sum_{w\in\mathbb{W}} \Gamma_{uvwt} \leq 1, \tag{5.29}$$

$$\forall u \in \mathbb{R}, v \in \mathbb{G}_u \cap \mathbb{I}, t \in \mathbb{T}.$$

For PON5, the following additional set and constraints are required:

Sets and parameters:

$\mathbb{O}$ Set of OLT ports (One port initially)

12. Constraints to ensure TDM (i.e. share of link) between the servers connected to the OLT port and the OLT port:

$$\sum_{v\in\mathbb{G}_u} \psi_{uvwt} \leq C \tag{5.30}$$

$$\forall u \in \mathbb{O}, w \in \mathbb{W}, t \in \mathbb{T}. \tag{5.31}$$

$$\sum_{v\in\mathbb{G}_u} \psi_{vuwt} \leq C$$

$$\forall u \in \mathbb{O}, w \in \mathbb{W}, t \in \mathbb{T}.$$

5.4 Results and Discussion

This Subsection provides the total energy consumption calculated by Equation 5.4 and the completion time estimated by Equations 5.25, 5.26, and 5.27 when optimising the

Table 5.2: Parameters related to the MILP model for optimising the co-flows scheduling and routing of MapReduce traffic

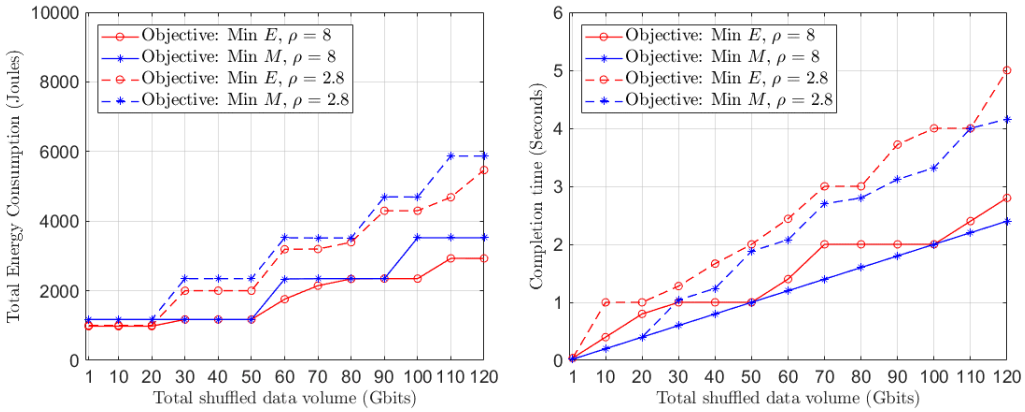
Parameter		Values
$\mathbb{G}, \mathbb{G}_u, \mathbb{R}, \mathbb{S}$		Check Figures 5.1, and 5.2
$\mathbb{T}, D$	Fattree, Spineleaf, BCube, DCell, PON5	Up to 6 slots, 1 second
	PON3	Up to 6 slots, 0.25 seconds
$\sum_{s,d \in \mathbb{R}, s \neq d} \Delta_{sd}$		1-120 Gbits without skew and with skew
C		10 Gbps
$P_{i(max)}$	Transceiver	1 Watt
	NIC	14 Watts
ϵ		0.07 Watt/Gbps
$O_{i(max)}$		Check table 5.1
ρ		8 Gbps, 2.8 Gbps
σ		The maximum switching capacity of the switch
L		5000 - 50000
Q		100

routing and scheduling of shuffling traffic under the objective of minimising the total energy consumption (i.e Equation 5.5) or the completion time (i.e. Equation 5.6). The results are generated for several scenarios while considering the parameters in Table 5.2 for the MILP model in Section 5.3

5.4.1 Electronic DCNs

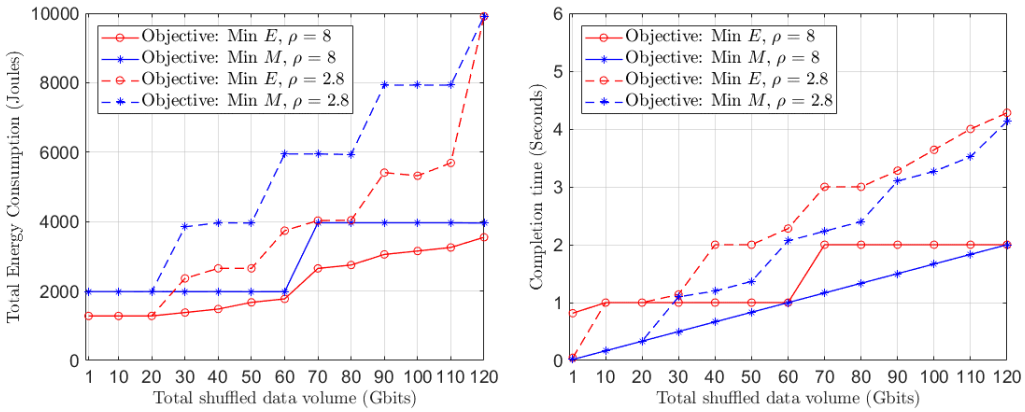
5.4.1.1 Energy Consumption and Completion Time with ON/OFF power profile and no intermediate data skew

Figures 5.4, 5.5, 5.6, and 5.7 show the results based on the MILP-obtained optimum routing and scheduling for shuffling traffic in Spine-leaf, Fat-tree, BCube, and DCell DCNs, respectively. The shuffling data is assumed to have no skew, ranging from 1 Gbits to 120 Gbits and the ON/OFF power profile was considered for the networking equipment with the specifications detailed in Table 5.1. Two rates per server values (ρ) were considered which are 2.8 Gbps and 8 Gbps. As the power profile is ON/OFF for the switches, transceivers, and NICs consume the same amount of power if their traffic is high or low. Hence, higher server rates will lead to lower energy use for the same amount of data to be sent due to higher utilisation for shorted duration.



(a) (b)

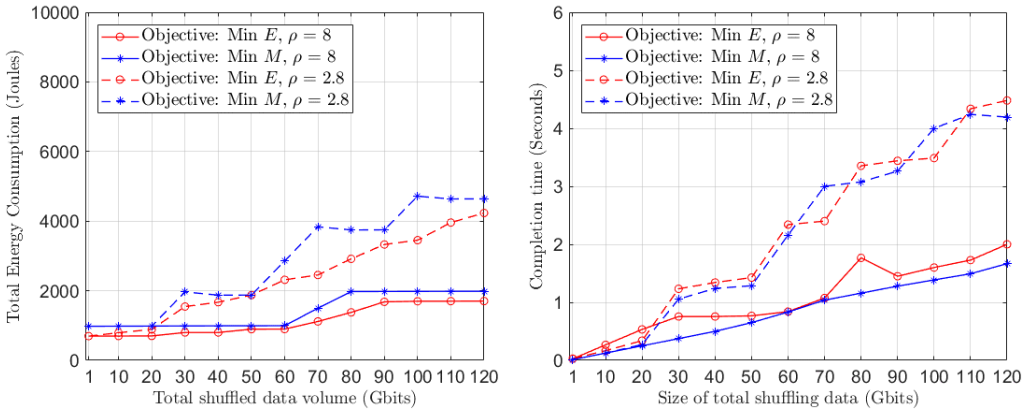
Figure 5.4: Energy consumption and completion time for Spine-leaf DCN with no intermediate data skew.



(a) (b)

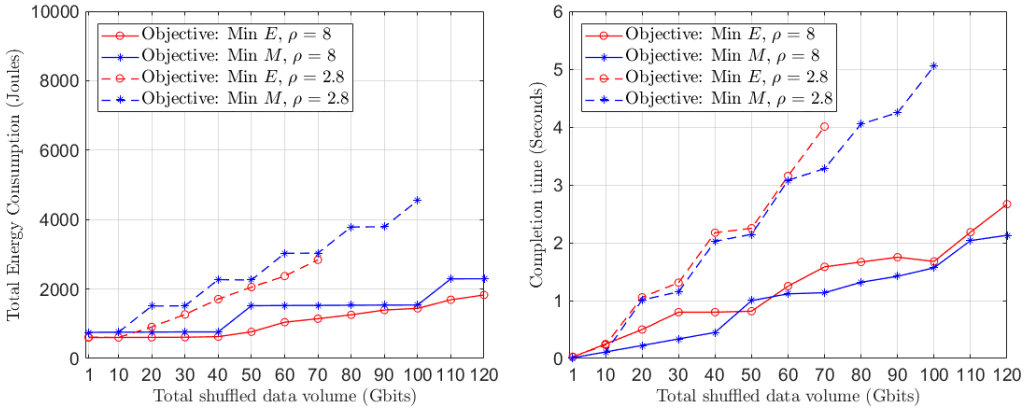
Figure 5.5: Energy consumption and completion time for Fat-tree DCN with no intermediate data skew.

For the electronic DCNs, D, which is the time slot used for scheduling, was set to be 1 second. As the time is discrete and increases in integer multiples of D, a higher server rate does not necessarily provide an advantage if the file size is small. Namely if the file is small, the lower server rate may complete the transmission for example at 0.9 D and a higher rate server may complete the transmission in 0.2 D. Both systems will



(a) (b)

Figure 5.6: Energy consumption and completion time for BCube DCN with no intermediate data skew.



(a) (b)

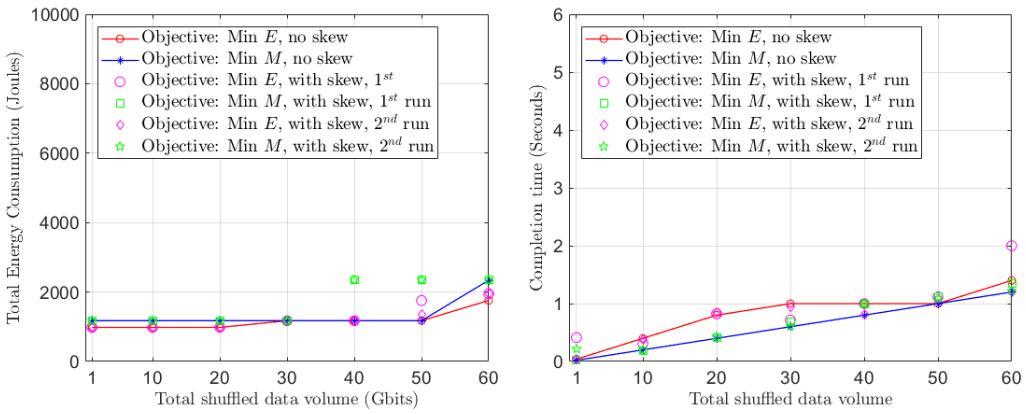
Figure 5.7: Energy consumption and completion time for DCell DCN with no intermediate data skew.

need a full time slot, hence the advantage of a higher data rate in term of the power efficiency is small at small file sizes. If the time is continuous (i.e. not discrete), then a higher data rate per server will mean that the data is transmitted in shorter time and hence the equipment can be switched off sooner leading to higher energy efficiency for higher rates per server. The best strategy to minimise the energy consumption

with ON/OFF power profile is to transmit at the maximum rate in fewer devices while switching off the remaining devices.

5.4.1.2 Energy Consumption and Completion Time with ON/OFF power profile and intermediate data skew

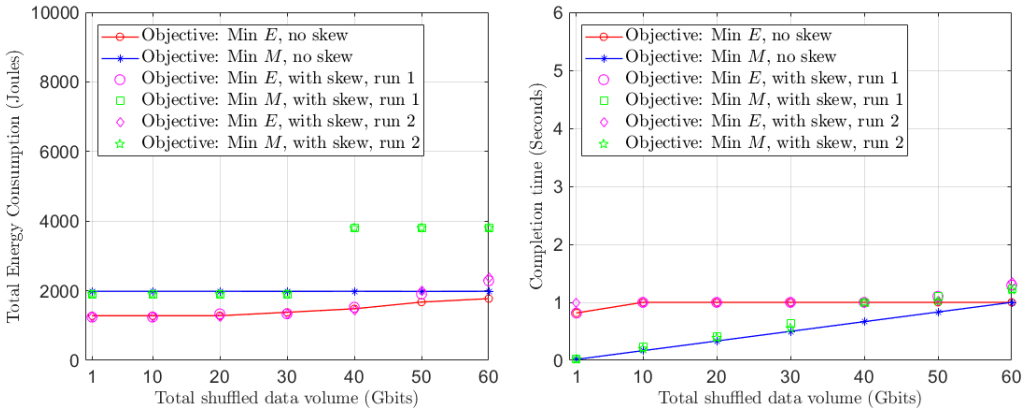
The results for the energy consumption and completion time when the data is skewed were generated for optimising the scheduling and routing for total shuffling data sizes of up to 60 Gbits with a rate per server (ρ) of 8 Gbps. For each total shuffling data size and each data centre, two runs for the flow sizes, randomly generated, were utilised. Figures 5.8, 5.9, 5.10, and 5.11 illustrate the results for Spine-leaf, Fat-tree, BCube, and DCell DCNs, respectively.



(a) (b)

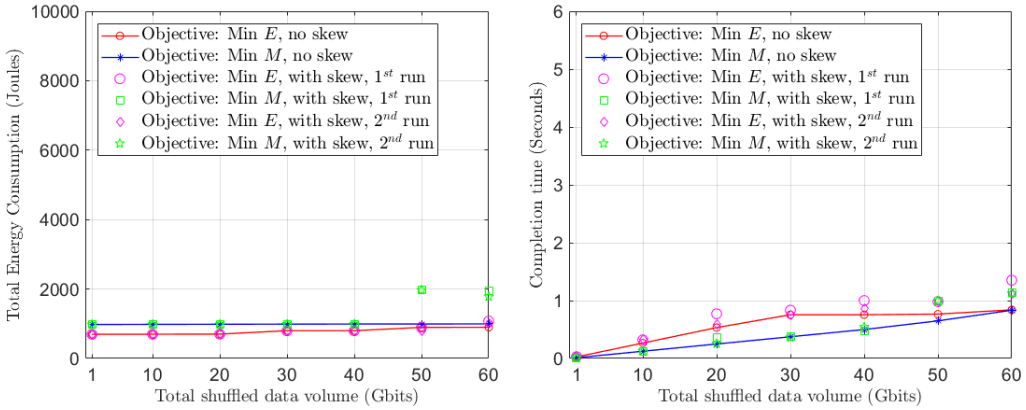
Figure 5.8: Energy consumption and completion time for Spine-leaf DCN with intermediate data skew.

The results show that under different objectives (i.e. minimising the energy consumption or the completion time), the abilities of different data centres to overcome the overheads of data skew are different. For Fat-tree data centre, the completion time results when minimising the power consumption or the completion time for skewed data indicated negligible impact on either objective. However, to achieve this for the objective of minimising the completion time for total data sizes larger than 30 Gbits,



(a) (b)

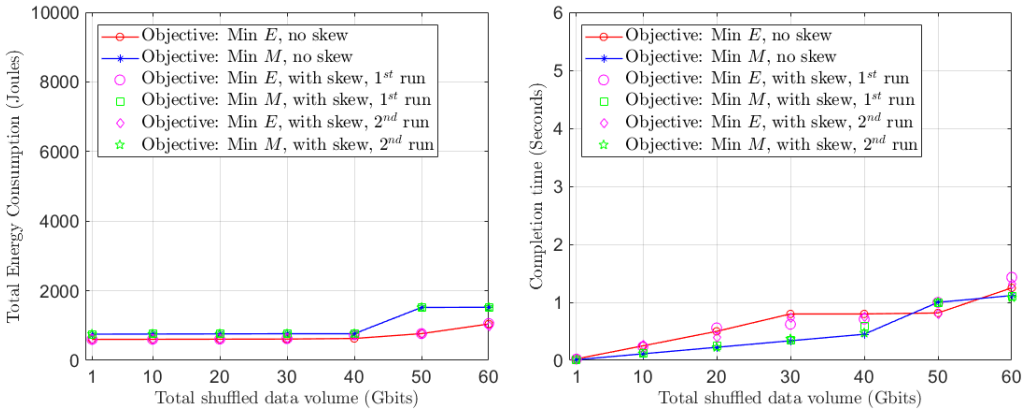
Figure 5.9: Energy consumption and completion time for Fat-tree DCN with intermediate data skew.



(a) (b)

Figure 5.10: Energy consumption and completion time for BCube DCN with intermediate data skew.

an increase by about 200% in the energy consumption is required. Similar trends are found for spine-leaf and BCube as depicted in Figure 5.8, and Figure 5.10. In contrast, the DCell results depicted in Figure 5.11 indicated that optimising the scheduling and routing with any of the two objectives resulted in balancing the impact of intermediate data skew.



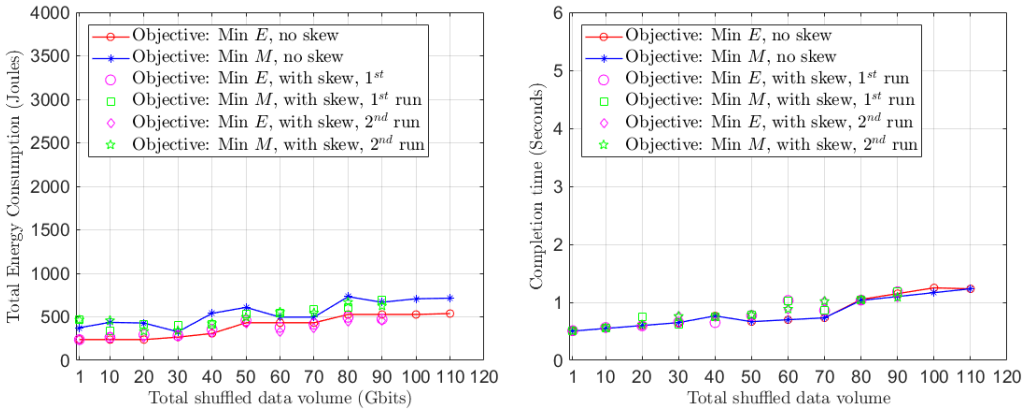
(a) (b)

Figure 5.11: Energy consumption and completion time for DCell DCN with intermediate data skew.

5.4.2 PON-based Optical DCNs

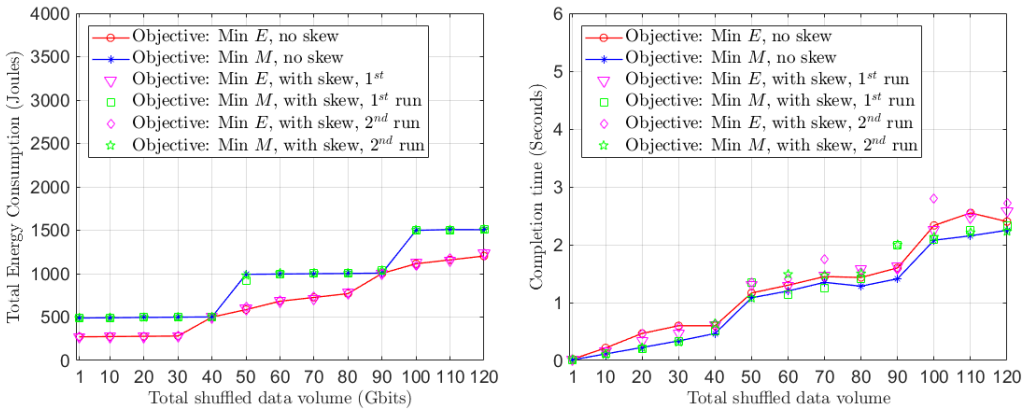
5.4.2.1 Energy Consumption and Completion Time without and with Intermediate Data Skew

We considered a rate per server (ρ) of 8 Gbps for the evaluation of the completion time and energy consumption in the PON-based DCNs. For PON3, and as a tuneable transceiver allows each server to transmit at a single wavelength in a given time slot, D was reduced to 0.25 seconds and an adequate number of slots was considered (i.e. up to 6 slots). This allows each map server to communicate with the six reduce servers in different time slots through the AWGR in case they are located in different racks. The results in Figure 5.12 indicate that the completion time is reduced by about 50% compared to other data centres. This reduction is attributed to the many server to server routes at different wavelengths achieved by the PON design that allow better utilisation of links. The results in Figure 5.13 indicate that PON5 achieves similar completion time to that of DCell while having lower power consumption. Data skew was also considered and the results show that PON5 has less sensitivity to skew compared to PON3.



(a) (b)

Figure 5.12: Energy consumption and completion time for PON3 DCN without and with intermediate data skew.



(a) (b)

Figure 5.13: Energy consumption and completion time for PON5 DCN without and with intermediate data skew.

5.5 Summary

This chapter provides a time-slotted MILP model to optimise the scheduling and routing of the flows in the shuffling phase of MapReduce. Two objectives are considered, which are minimising the total energy consumption of the data centre and minimising the completion time of the shuffling phase. This model is used to compare the energy

consumption and the performance of four electronic switching data centres and two PON-based data centres under several scenarios. A general observation is that when the objective is to minimise the completion time, higher energy is required indicating the need to activate more devices in the data centre. When the objective is to minimise the energy consumption, higher completion time results as redundant equipment is preferentially switched off. It is worth noting that for both objectives, an additional aim is to try to utilise earlier time slots which also helps in reducing the completion time. Thus, the completion time reduction objective aims to purely reduce completion time without any consideration for the energy consumption, while the energy minimisation objective also targets reducing the completion time but as a lower priority. The smallest completion time was obtained for PON3 due to the use of WDM in the architecture which offers higher capacities per route. Fat-tree was found to have the highest energy consumption when minimising the completion time due to the use of several commodity switches in the architecture in the path of each flow. With the objective of minimizing the total energy consumption and for a rate per server value of 8 Gbps and a total shuffling data volume of 60 Gbits, the two PON-based data centres achieve an average energy consumption reduction of about 83% compared to the electronic data centres. With the objective of minimizing the completion time, the average reduction in the energy consumption for the PON-based data centres is about 56% compared to the electronic data centres. The impact of data skew and the ability of data centres to avoid its overhead in terms of the power consumption or the completion time was also examined. The least sensitive DCNs to intermediate data skew are DCell, and PON5 due to the availability of a large number of routes between the servers. In the following chapter, the energy consumption and the completion time results are evaluated when considering link, switch and server failures.

Chapter 6

Resilience of Data Centre Networks and the Performance of Big Data Applications

This Chapter addresses the impact of link, switch, and server failure on MapReduce shuffling performance and energy efficiency in three data centre topologies. The MILP model in the previous Chapter is utilised while considering different non-fatal switch and link failure scenarios to evaluate the completion time and total energy consumption. A modified MILP model is also developed to examine the trade-offs in terms of the performance and energy efficiency when considering a replication factor of two and different allocation schemes for the input data to map tasks in the servers. The remainder of this chapter is organised as follows: Section 6.1 summarises some related studies and Section 6.2 discusses the system model and the scenarios considered and parameters. Section 6.3 provides the MILP model used to optimise the routing of shuffling traffic when considering the replication factor and server failures, while Section 6.4 discusses the results. Finally, Section 6.5 provides the summary.

6.1 Related Studies

Providing high reliability and continuity is a key requirement for cloud computing services as any disruption or disconnection in the infrastructure typically leads to revenue losses and customer departures. Failures in cloud infrastructure can occur in the transporting networks, data centres, or in the applications [258]. Overcoming the severe impacts of these failures requires the adoption of resilient designs and restoration plans. The resilience and energy efficiency of different IP over WDM network topologies were examined in [103] under fiber cuts or a core node failure. An energy efficient NC-based 1+1 protection scheme was proposed in [104], and [105] where the encoding of multiple flows sharing protection paths in non-bypass IP over WDM networks was optimised. MILP, heuristics, in addition to closed form expressions were used to obtain results for the networking power consumption as a function of the hop count, network size, and demands. The results indicated saving by up to 37% compared to conventional 1+1 protection. The authors in [106] optimised the traffic grooming and the assignment of router ports to protection or working links under different protection schemes while considering the sleep mode for protection ports and cards. Up to 40% saving in the power consumption was achieved. The work in [99] and [100] considered static and dynamic adaptation to traffic surges resulting from re-routing demands after links failures. The authors in [259] addressed the resilience of geo-distributed transport networks that link cloud data centres by jointly considering the content placement, and anycast routing under failures.

At the application level, solutions such as maintaining several replicas of data and software components are considered. For example, MapReduce, which is a widely used framework for big data parallel computations, has a data replication mechanism with default value of 3, and considers speculative execution for straggling tasks that cause completion time imbalance [17]. The authors in [260] proposed improved speculative execution strategies where Locally Weighted Regression (LWR) is used to identify straggler tasks and a cost-benefit model is used to optimise the selection of the backup

nodes to re-run them. To improve the performance of Hadoop under failures, a modified MapReduce work flow with fine-grained fault-tolerance mechanism called BEneath the Task Level (BeTL) was proposed in [261]. BeTL allows generating more files during the shuffling to create more checkpoints. It improved the performance of Hadoop under no failures by 6.6% and under failures by up to 51%. The work in [262] proposed four multi-queue size-based scheduling policies to reduce jobs slowdown variability which is defined as the idle time to wait for resources or I/O operations. Several factors such as parameters sensitivity, load unbalance, heavy-traffic, and fairness were considered. The work in [263] optimised the number of reduce tasks, their configurations, and memory allocations based on profiling the intermediate results size. The results indicated a complete disregard for job failures due to insufficient memory and a reduction in the completion time by up to 88.79% compared to legacy memory allocation approaches.

To improve the performance of MapReduce in memory-constrained systems, *Mammoth* was proposed in [264] to provide global memory management. In Mammoth, related map and reduce tasks were launched in a single JVM as threads that share the memory at run time. Mammoth actively pushes intermediate results to reducers unlike Hadoop that passively pulls from disks. A rule-based heuristic was used to prioritise memory allocations and revocations among map, shuffle, and reduce operations. Mammoth was found to be 5.19 times faster than Hadoop 1 and it outperformed Spark for interactive and iterative jobs when the memory is insufficient [264]. An automatic skew mitigation approach; *SkewTune* was proposed and optimised in [265]. SkewTune detects different types of skew, and effectively re-partitions the unprocessed data of the stragglers to process them in idle nodes. The results indicated a reduction by a factor of 4 in the completion time for workloads with skew and minimal overhead for workloads without skew.

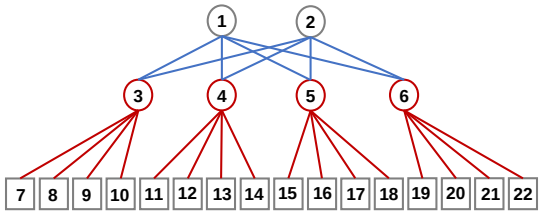
Several studies have considered the resilience of DCNs (e.g. [266–269]) by addressing the sources of links, switches, and servers' software and hardware failures. The resilience of a data centre is highly coupled with its power consumption and cost as it requires doubling the resources and links that can be underutilised. Fat-tree [132]

tackles the fault tolerance by allowing multi-paths between paired servers and by having aggregation and access switches connected as a biperate graph. BCube [136] provides uniform multi-path between servers that utilise switches at different levels and other servers for forwarding the traffic. DCell [137] uses a decentralised Fault-tolerant Routing Protocol to effectively handle hardware and software failures. The authors in [270] measured the resilience of data centres by the number of cuts required to fully disconnect a logical link. XPath in [239] considered the assignment of backup paths to flows in its application-aware SDN routing scheme. In [271], the authors considered the capacity and reliability of links between VMs and defined the traffic stress as the product of the traffic rates, route lengths, and the inverse of the links reliability. Their objective was to find service nodes placement that minimise the traffic stress.

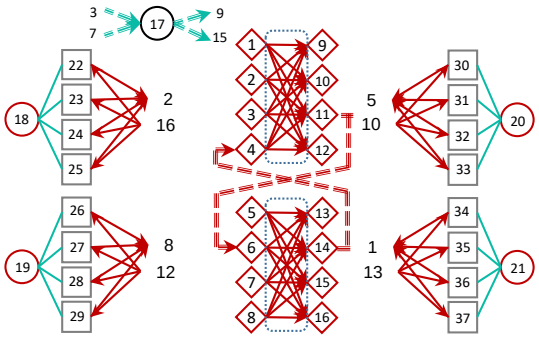
6.2 System Model and Parameters

We considered the optimisation of the scheduling and routing of shuffling MapReduce flows under non-fatal link or port failure in three data centres with the objective of minimising the energy consumption or the completion time. The data centres considered are the Spine-leaf, PON3, and PON5. Figure 6.1 shows the graph representation of the data centres while outlining the fatal links and switches in red. Any failure in one of those links or switches will require replicating the data in other servers to be accessible. Non-fatal failures in links or switches reduce the performance of the DCN due to the loss of redundant paths and the need to transmit data using reduced DCN capacity.

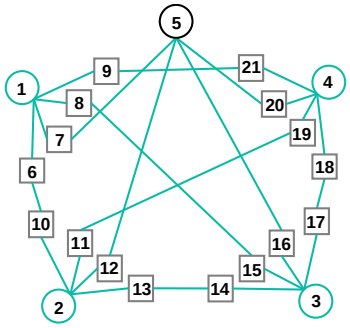
We considered several cases for switch, link, and OLT port failure under no data replication. The rate per server, ρ, was set to 8 Gbps and no data skew was considered. To examine the impact of fatal link and switches failure in addition to server failure on the energy efficiency and completion time, a modified MILP model is also developed to examine the trade-offs in the performance and energy efficiency when considering a replication factor of two for input data and different allocation schemes for the input data to map tasks in the servers.



(a)



(b)



(c)

Figure 6.1: Non-fatal and fatal link and switch failures in DCNs. Red outlines edges (i.e. links) and switches whose failure is fatal to MapReduce shuffling. Squares and circles represent servers and switches, respectively. (a) Spine-leaf, (b) PON3, and (c) PON5.

6.3 MILP Model for Optimising the Routing of MapReduce Shuffling Traffic under Server Failures

This MILP model takes the topology of the data centre (i.e. the connections and the capacity of links), the locations of data copies, the failed servers, the power consumption for all equipment, and the total shuffling traffic as input and provides the schedule (i.e. time slot and amount of assigned traffic for each link), the completion time, and the total energy consumption. This is obtained under one of two objectives which are to reduce the energy consumption or to reduce the completion time and while considering the architectural and routing constraints. In addition to the parameters in the MILP model presented in Chapter 5, the following are additionally defined:

Sets and parameters:

$\mathbb{R}_{\mathbb{M}}$ Set of servers containing map tasks

Φ_{zs} Binary parameter which is equal to one if a copy for the original data assigned to map tasks in server s is also available in server z and is equal to zero otherwise; $s, z \in \mathbb{R}_{\mathbb{M}}$

Λ_s Binary parameter which is equal to one if server s is working and is equal to zero if there is a failure in it; $s \in \mathbb{R}_{\mathbb{M}}$

The variable δ_{sdt} is replaced by the following variable to allow shuffling flows corresponding to map results from a server containing a copy to servers with reduce tasks.

Variables:

Ξ_{szdt} Variable to indicate the amount of map data, which is originally to be shuffled from server s to reduce tasks in server d, that is shuffled instead from server z that contain a copy at time slot t; $s, z \in \mathbb{R}_{\mathbb{M}}, d \in \mathbb{R}, t \in \mathbb{T}$.

Accordingly, the objectives in Equations 5.5, and 5.6 are replaced by what follows:

The first objective to minimise, E, the total energy consumption which is expressed as:

$$\min \left[E + Q \sum_{s \in \mathbb{R}_{\mathbb{M}}, d \in \mathbb{R}, t \in \mathbb{T}, s \neq d} \left(t \sum_{z \in \mathbb{R}_{\mathbb{M}}} \Xi_{zsdt} \right) \right], \tag{6.1}$$

and the second objective to minimise, M, the latest completion time of shuffling is expressed as:

$$\min \left[M + Q \sum_{s, d \in \mathbb{R}, t \in \mathbb{T}, s \neq d} \left(t \sum_{z \in \mathbb{R}_{\mathbb{M}}} \Xi_{zsdt} \right) \right]. \tag{6.2}$$

Also, the constraints in Equations 5.7, and 5.12 in the MILP model presented in Chapter 5 are replaced by Equations 6.3, and 6.4, respectively:

1. Flow conservation in the data centre: The allocation of the links to the flows follows the flow conservation law at each time slot t and wavelength w:

$$\sum_{v \in \mathbb{G}_u} \chi^{sd}_{uvwt} - \sum_{v \in \mathbb{G}_u} \chi^{sd}_{vuwt} = \begin{cases} \sum_{z \in \mathbb{R}_{\mathbb{M}}} \Xi_{zsdt} & u = s \\ -\sum_{z \in \mathbb{R}_{\mathbb{M}}} \Xi_{zsdt} & u = d \\ 0 & otherwise \end{cases} \tag{6.3}$$

$$\forall s \in \mathbb{R}_{\mathbb{M}}, d \in \mathbb{R}, s \neq d, u \in \mathbb{G}, w \in \mathbb{W}, t \in \mathbb{T}.$$

6. Constraint to ensure that the sum of shuffling data flow originally assigned between server s and server d served in all time slots is equal to the total flow size while allowing serving that flow fully or partially from another server z if it contains a copy and is working:

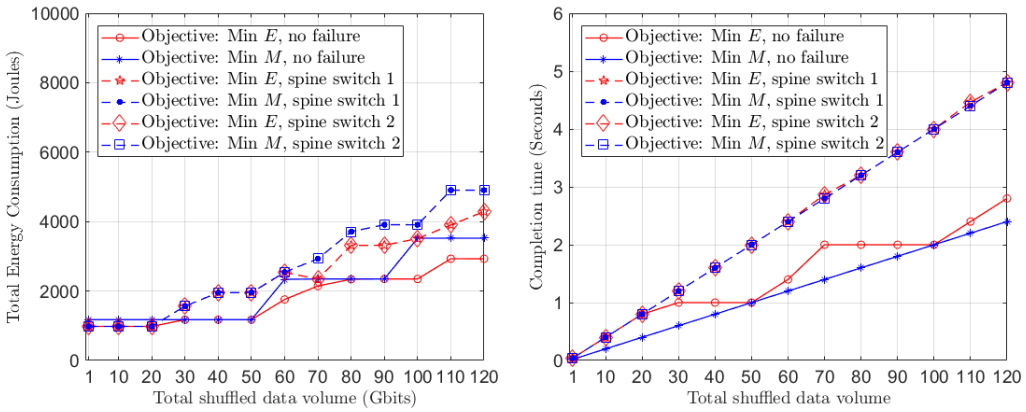
$$\sum_{t \in \mathbb{T}, z \in \mathbb{R}_{\mathbb{M}}} \Xi_{szdt} \wedge_z \Phi_{zs} = \Delta_{sd}; \forall s \in \mathbb{R}_{\mathbb{M}}, d \in \mathbb{R}, s \neq d \tag{6.4}$$

6.4 Results and Discussion

Link and Switch Non-Fatal Failures

Figure 6.2 shows the energy consumption and the completion time results when considering a failure in one of the spine switches. The result indicate that the completion time almost doubles for all objectives and that their impact is equivalent. With the

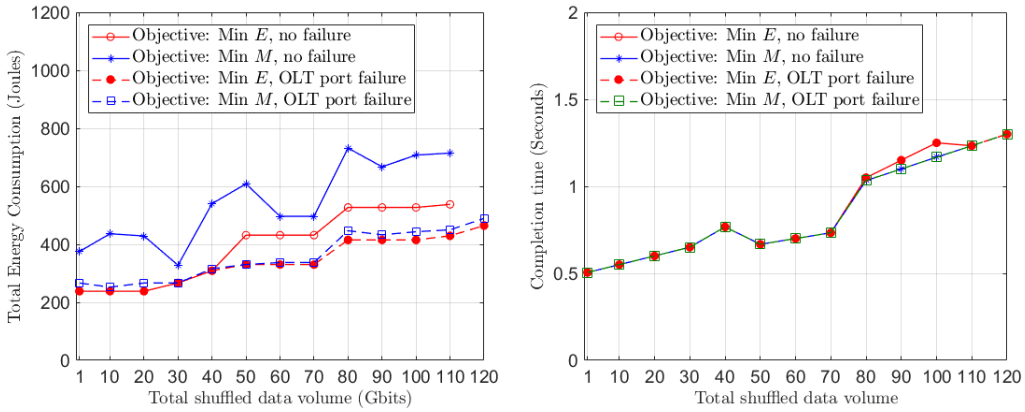
objective of minimising the energy consumption, the model will try to schedule the flows so that the usage time of the leaf switches is minimised throughout all time slots required. On the other hand, with the objective of minimising the completion time, all leaf switches will be ON most of the time, however, both objectives achieved similar completion time.



(a) (b)

Figure 6.2: Energy consumption and completion time for Spine-leaf DCN under link and switch non-fatal failure.

For PON3, we examined the impact of the OLT port failure. Using the OLT port in this topology can help in creating paths between more servers in a given time slot. However, the OLT is equipped with the highest energy consumption. The results in Figure 6.3 show the completion time when the OLT port is not used. The reduction in the energy consumption for any of the two objectives is due to not considering the OLT ports in offloading any traffic. The high power consumption for the objective of minimising the completion time under no OLT port failure is due to excessive use of the OLT port to achieve lower completion time. However, the completion time results shows that the benefit is minimal for the considered workloads. Finally, Figure 6.4 presents the impact of OLT port, single link, and two links failure in PON5. A failure in the OLT port causes a significant increase in the completion time results. That is due to the limited capacity in PON5 compared to PON3. Two link failures resulted



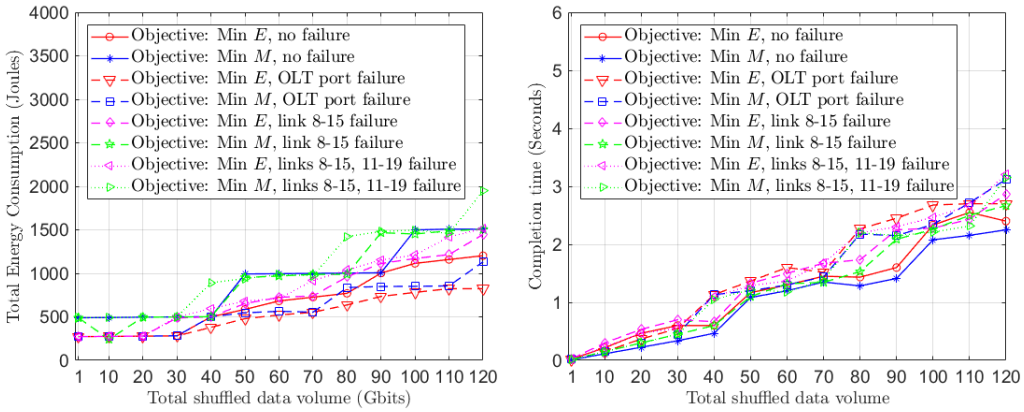
(a) (b)

Figure 6.3: Energy consumption and completion time for PON3 DCN under OLT port failure.

in increasing the completion time by 33% for the objective of minimising the energy consumption and resulted in increased power consumption by 50% for the objective of minimising the completion time. It was also observed that different combinations of single and two links failure resulted comparable results to those presented in Figure 6.4 as the load is generally balanced among the links. Thus, only a single case was presented for each of single and two links failure.

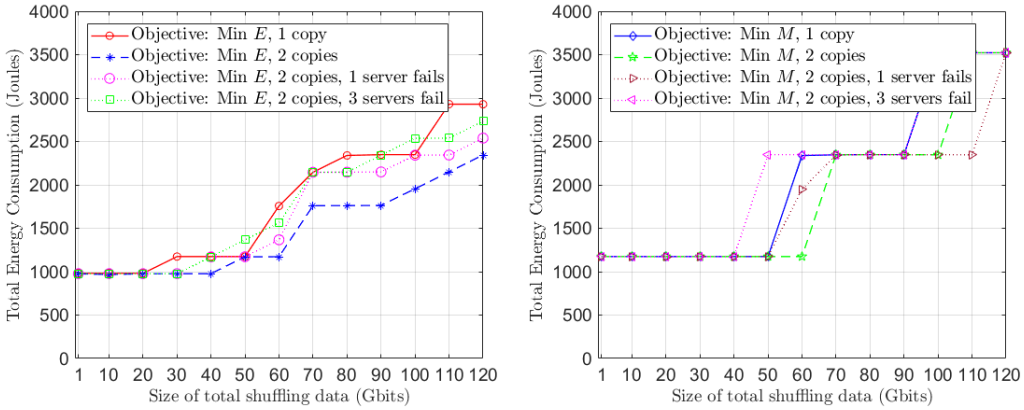
Server Failures

Figures 6.5, and 6.6, show the energy consumption and completion time results, respectively for Spine-leaf DCNs when considering a replication factor of two and one or three server failures. The results for the case of two servers failures is omitted due to their marginal difference compared to single server failure results. First, the energy consumption and completion time results are obtained under no server failure. The reduction in the completion time under the objective of minimising completion time and the improvement in the energy efficiency under the objective of minimising energy consumption are attributed to the possibility of improving the locality by selecting the optimal copies that can improve each objective. Single and three server failures reduce



(a) (b)

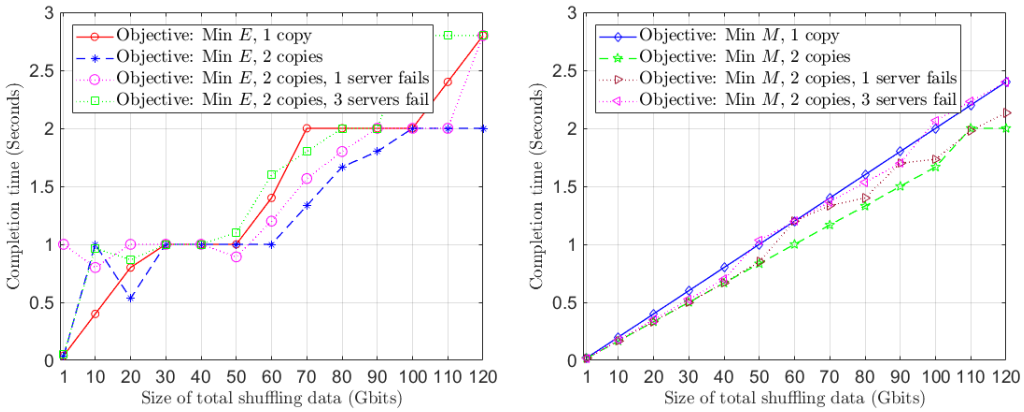
Figure 6.4: Energy consumption and completion time for PON5 DCN under OLT port and link failure.



(a) (b)

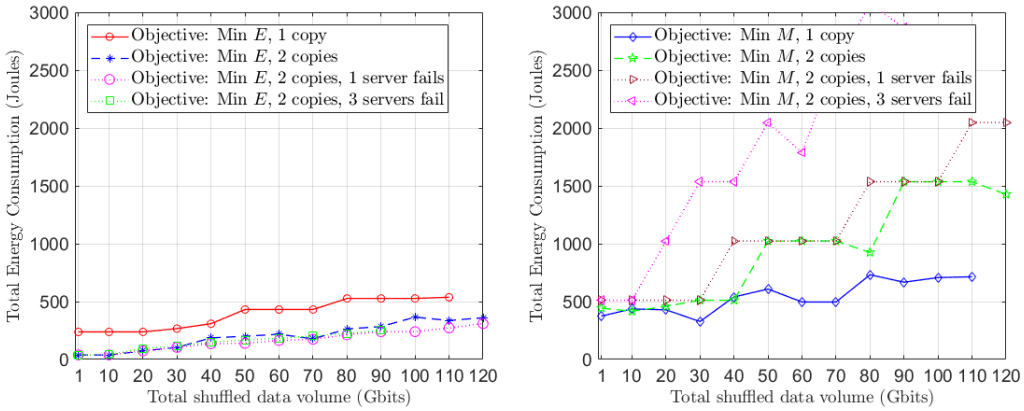
Figure 6.5: Energy consumption for Spine-leaf DCN with data replication and server failure, (a) minimise energy consumption objective, and (b) minimise completion time objective.

this advantage as the copy options needed to perform the shuffling are reduced. Figures 6.7 - 6.10 show the results for PON3 and PON5 when considering replication and server failures. As in Spine-leaf, considering a replication factor of two under no server failure improves the results. In PON3, it enables the utilisation of fewer time slots,



(a) (b)

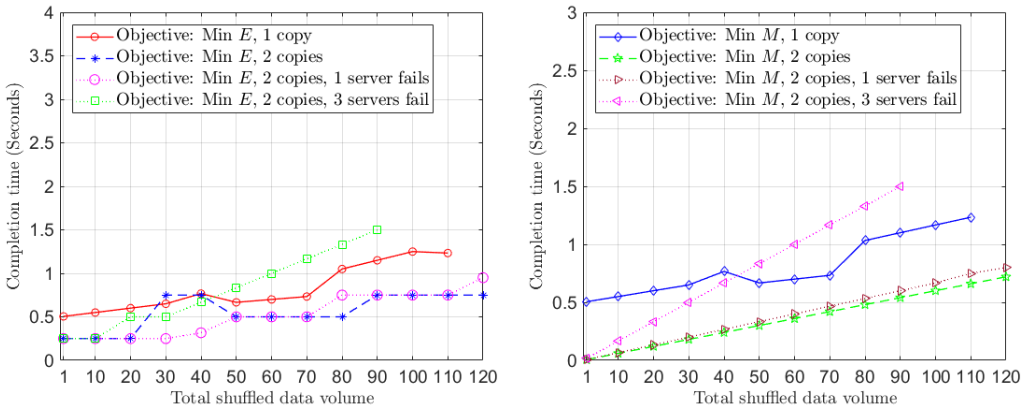
Figure 6.6: Completion time for Spine-leaf DCN with data replication and server failure, (a) minimise energy consumption objective, and (b) minimise completion time objective.



(a) (b)

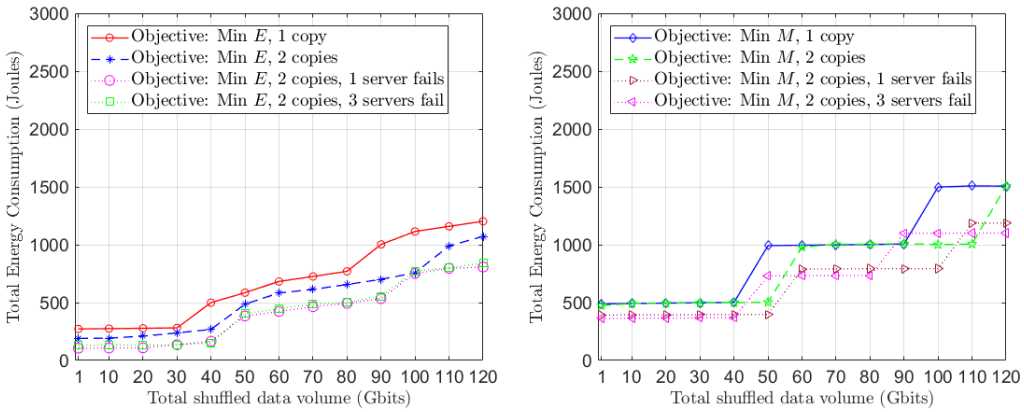
Figure 6.7: Energy consumption for PON3 DCN with data replication and server failure, (a) minimise energy consumption objective, and (b) minimise completion time objective.

and in PON5, it improves the locality and hence, the routing and scheduling. In all results, the energy consumption of the failed servers is not considered as they will also be unavailable for routing traffic.



(a) (b)

Figure 6.8: Completion time for PON3 DCN with data replication and server failure, (a) minimise energy consumption objective, and (b) minimise completion time objective.

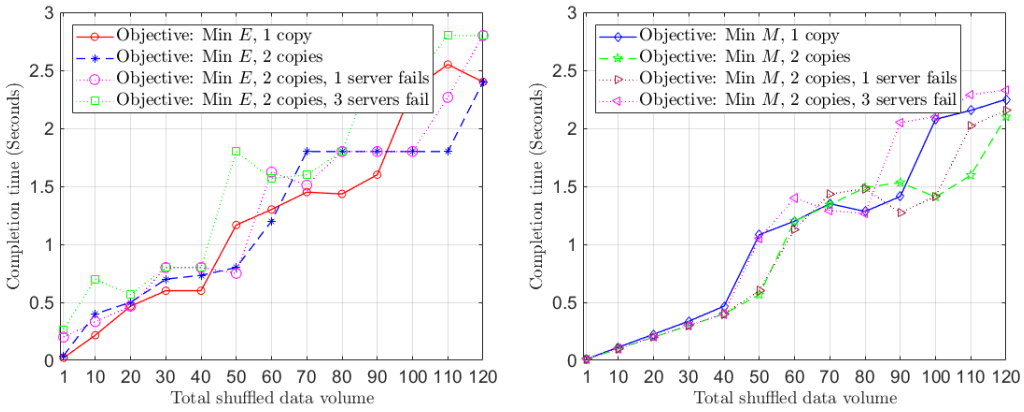


(a) (b)

Figure 6.9: Energy consumption for PON5 DCN with data replication and server failure, (a) minimise energy consumption objective, and (b) minimise completion time objective.

6.5 Summary

In this Chapter, we examine the impact of non-fatal and fatal failures in three data centres. The non-fatal failures included switch, link, and OLT port failures. For fatal failures, replicating the data is required to finish the shuffling operations. In this case,



(a) (b)

Figure 6.10: Completion time for PON5 DCN with data replication and server failure, (a) minimise energy consumption, and (b) minimise completion time.

a modified MILP model was utilised to calculate the energy consumption required and completion time achieved when optimising the scheduling and routing while having a replication factor of two. The results show that different DCNs experience different degradation in the energy efficiency or the completion time based on the links redundancy, and links utilisation. For spine-leaf data centres with two spine switches, the results for a spine switch failure indicate that the completion time almost doubles for all objectives with equivalent impact. Any leaf switch failure is considered fatal as the access to any of the servers connected to it is not possible. For PON3, the only non-fatal failure is the OLT port failure. Under OLT port failure, the two objective achieves comparable energy efficiency and completion time results as the additional paths that can improve both metrics are not available. Any link failure in this architecture is considered fatal, however failures are less likely to occur in such passive links compared to electronic switching-based links. For PON5, and as available capacities are less than PON3, OLT port failure causes a significant increase in the completion time results. Two link failures in PON5 resulted in increasing the completion time by 33% for the objective of minimising the energy consumption and resulted in increased power consumption by 50% for the objective of minimising the completion time. For servers failure results, it

was observed that allowing the use of any of the two data copies improves the locality and also the results (i.e. reduces the energy consumption for the energy minimisation objective and reduces the completion time for the completion minimisation objective). Under single and three server failures, this advantage is reduced.

Chapter 7

Optimisation of Content Delivery in Geo-distributed Cloud and Fog Data Centres

This chapter examines the reduction in the brown power consumption of transport networks including core, metro and access layers that can be achieved by caching Video-on-Demand (VoD) content in solar-powered fog data centres with Energy Storage Devices (ESDs). The effects of considering optical bypass routing, and MLR in the core network, the availability of solar renewable energy in the access network, and optimising the use of ESDs were addressed. A MILP model that considers the above-mentioned factors was developed to optimise the delivery of VoD content from cloud data centres in the core network or fog data centres in the access network. The remaining of this chapter is organised as follows: Section 7.1 provides some related studies and Section 7.2 shows the considered system model and parameters. Section 7.3 shows the MILP model for efficient content delivery while Section 7.4 presents the results and discussions. Finally, Section 7.5 provides the summary.

7.1 Related Studies

Video traffic is estimated to have a Compound Annual Growth Rate (CAGR) of 54% from 2016 to 2021 [19]. As a result, the power consumption of transport networks that link cloud data centres containing video workloads and end users in the access network is expected to massively increase. As these systems are typically powered by brown energy (i.e. non-renewable energy), this would also leads to steep rise in CO_2 gas emission and operational costs due to high utilisation and cooling requirements against thermal dissipation [95]. To overcome both issues, several greening approaches were considered in the last decade such as improving the hardware, optimising the routing and workload scheduling algorithms, in addition to considering renewable power sources [68]. The authors of [95] considered lightpath bypassing in IP over WDM core networks to reduce the power consumption and achieved energy savings in the range of 25% to 45%. As part of the efforts in GreenTouch, the work in [42, 107] investigated a combination of greening approaches for IP over WDM core networks through a comprehensive MILP model and heuristics. Those included the consideration of optical bypassing, optimising the core network topologies, employing MLR, utilising efficient protection and sleep modes, in addition to considering two improvements schemes for hardware which are the Business-As-Usual (BAU) improvement in equipment due to advances in Complementary Metal Oxide Semiconductor (CMOS) technology, and BAU with further GreenTouch improvements. The former indicated 4.23× energy efficiency improvements compared to 2010 networks while the later indicated 20× improvements.

The energy efficiency in Content Delivery Networks (CDNs) and Information-centric Networks (ICNs) was extensively surveyed in [60], and [61]. Optimising the workloads and content placement to reduce the traffic and hence the power consumption was also considered to green core networks as in [117, 119–122, 124]. In [117], the authors focused on data centre and popular content placement strategies and found that placing the data centres at the centre of the network and replicating the contents on multiple data centres according to their popularity minimised the power consumption by 28%.

In [119], the energy efficiency of Video-on-Demand (VoD) services was examined by numerically evaluating five strategic locations for caching in core, metro, and access networks. In [120–122], the caching of VoD contents was optimised to reduce its storage and transport energy consumption while considering the sizes of the caches, the contents popularity at different hours and dynamic cache contents replacement. To reduce the energy consumption of various cloud services, the work in [124] optimised the distribution of contents and services in cloud data centres and found that the optimised replications reduced the power consumption by 43% compared to centralised placements.

To reduce the CO_2 emission coupled with the rise in brown power consumption, the use of renewable resources has been considered to power different networking and data centre elements. The authors in [109] suggested using renewable energy sources in IP over WDM core nodes and optimised the routing to maximise the renewable energy usage which resulted reductions in CO_2 emission between 47% and 52%. The dynamics of solar power availability and workloads was considered in [115] while optimising the use of solar energy for cloud data centres and IP over WDM equipment and reductions by up to 32% in CO_2 emission was obtained. In [113], wind energy was considered for cloud services while considering the cloud locations, contents replications and the renewable energy transmission losses.

Different implementations such as Mobile Edge Computing (MEC), Fog Computing, and cloudlet Computing were recently emphasised on to reduce the latency of delivering various services through cloud computing to users in access networks [37, 189, 272]. Such implementations are also capable of reducing the energy consumption of core networks [273]. Nano Data Centres (NaDa) were introduced in the early work in [118] as a peer-to-peer computing and storage infrastructure at gateways and energy consumption reduction by at least 20-30% was obtained. The use of fog data centres for smart city applications was discussed in [274] to reduce core networks power consumption and maintain the QoS. The authors in [275] suggesting edge caching for Device-to-Device communications to improve the performance and reduce the consumption of back-haul

networks. The performance and power consumption tradeoffs of using of different data centre topologies for big data computations in fog environments were discussed in [4]. In [276], the concept of integrating micro data centre (Micro-DC) into OLTs of PONs was discussed to partially reduce core networks traffic. The authors in [199] proposed Fog Co-located OLT architecture which integrates fog computing at Central Offices PONs to improve telecommunication services. In [119], the power consumption and delay tradeoffs of caching VoD contents from different layers including core routers, metro routers and switches, and at the OLT and ONU were analysed.

To enhance the use of interrupted renewable sources such as solar power for data centres, the use of Energy Storage Devices (ESDs) is suggested. In [277], ESDs were utilised to store surplus renewable energy and discharge it during high workload peaks or when the brown energy price is expensive. The authors in [278] optimised the use of ESDs when the renewable energy is not available or during peak workloads and proposed an opportunistic scheduling algorithm to delay batch workloads until the renewable energy is available. The work in [279] reduced the energy cost for cloud data centres by implementing ESDs and energy trading for different renewable sources and optimising the consumption, storage and trading with power grids while addressing the inefficiencies with charging and discharging the batteries.

The work in this chapter utilises a MILP model to examine the reduction in the non-renewable power consumption of transport networks when delivering VoD traffic by maximising the use of solar renewable energy in fog data centres with ESDs in the access network as detailed in the following sections.

7.2 System Model and Parameters

7.2.1 Transport Network

In this work, the IP over WDM architecture with the by-pass approach was utilised for the core network. The network is modelled as a weighted unidirectional graph $G = (N, L)$, where N is the set of core nodes, and L is the set of physical links between

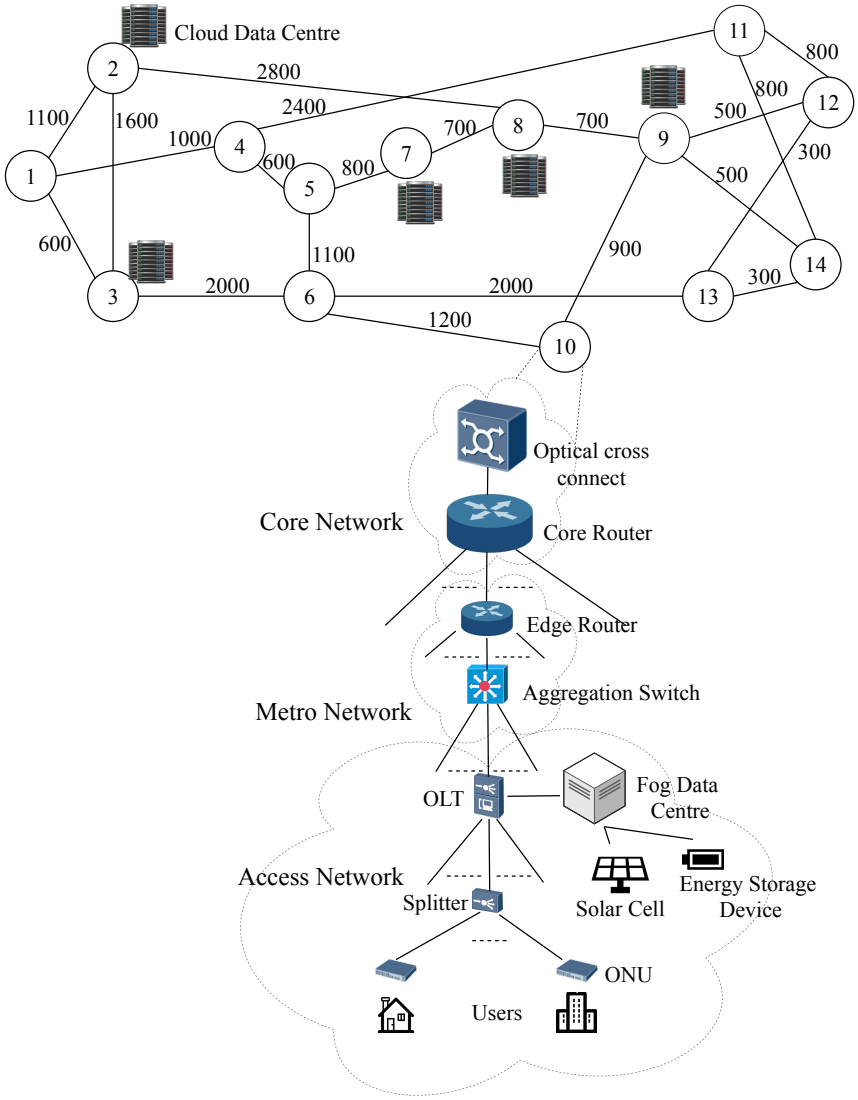


Figure 7.1: Fog data centre caching model to assist cloud VoD service.

the nodes. The NSFNET network topology was considered. In NSFNET, 14 nodes are connected through 21 bidirectional links with distances in km, $D_{(m,n)}$, as provided in Figure 7.1. Each core node is equipped with adequate IP router and transponder ports in addition to an optical switch. For the links, a number of Erbium-Doped Fiber Amplifiers (EDFAs) and regenerators are considered according to the link budget requirements at different line rates. The cloud data centres (CDCs) are pre-located in

nodes 2, 3, 7, 8, and 9 [42]. In each core node, a metro network, composed of edge routers and Ethernet switches, is utilised to provide connection with access networks associated with that node. All the aggregation ports in the core and metro routers are assumed to be 40 Gbps as it is the state-of-the-art technology at the time of writing. For metro Ethernet, C9500-32QC switches [280] are utilised. The access network is composed of OLTs connecting the metro network with Fog Data Centres (FDCs) to assist the five CDCs in delivering VoD services to users, in addition to splitters, ONUs, and end users. The capacities and power consumption values of the OLT chassis are obtained from [217]. The 40 GE (4 ports) are utilised for metro network connections, and the 10 GE (12 ports) of the Ethernet uplink and the switching and control cards are utilised for the connections with the fog data centre. Additional 2 Ethernet interface service cards providing a total of 4 10 GE ports are also utilised for the FDC. This provides up to 160 Gbps capacity between the OLT and the metro network and also 160 Gbps capacity between the OLT and the FDC.

7.2.2 Cloud and Fog Data Centres

For both CDCs and FDCs, the networking equipment power consumption is assumed to be 30% of the servers' power consumption [187]. The content server in [113] which has a maximum streaming capacity of 1.8 Gbps was considered. This allows the FDC to maximally provide 160 Gbps via about 88 servers. Each FDC is powered by brown sources, directly by solar cells with areas between 50 and 250 m^2, or by stored solar energy in an ESD with a capacity of 100 kWh [281]. Power Usage Effectiveness (PUE) values between 1.25 and 1.1 for FDC and of 1.1 for CDC were considered. Table 7.1 summarises the networking equipment and data centre parameters considered. In this work, consumer video traffic based on the Cisco Visual Network Index (VNI) forecast for 2020 [42] was considered for the demands from the five CDCs to users in the 14 NSFNET nodes. Figure 7.2 shows the total volumes at different times of the day in Tbps.

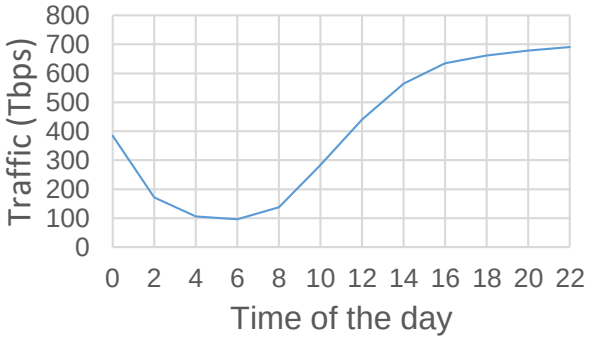


Figure 7.2: 2020 consumer video traffic.

7.2.3 Renewable Energy Sources

We considered solar renewable energy for its suitability for installment in fog environments within cities. The solar irradiance values in all 14 nodes of NSFNET were collected from [282] in W/m^2 which is based on SOLPOS calculator-based predictions and were averaged over two hour windows. Considering several factors that affect solar cells efficiency and based on [283], an efficiency of 26.3% was considered. Accordingly, Table 7.2 summaries the solar power availability in W/m^2.

7.3 MILP Model for Efficient Content Delivery

Below, we list the parameters, the variables, objective, and constraints of the MILP model. Small letters in superscripts and subscripts indicate indices while double-lined letters indicate sets.

The power consumption in IP over WDM networks at t; $t \in \mathbb{T}$, $P_{t(IPoverWDM)}$, is composed of the following:

1. IP Router ports under optical bypass:

$$P_{t(IP)} = \sum_{i \in \mathbb{N}} PR_M(CCQ_{it} + CMQ_{it}) + \sum_{j \in \mathbb{N}: i \neq j} \sum_{r \in \mathbb{WR}} PR_r\, C_{ijrt}. \tag{7.1}$$

Sets and parameters:

$\mathbb{N}$	Set of IP over WDM nodes
$\mathbb{N}_m$	Set of neighbours of node m; $m \in \mathbb{N}$
$\mathbb{T}$	Set of hours in the day
$\mathbb{WR}$	Set of wavelength rates
$\mathbb{CDC}$	Set of cloud data centres (CDCs) ($\mathbb{CDC} \subset \mathbb{N}$)
$\mathbb{S}$	Set of servers in a fog data centre
S	The duration between consecutive hours in $\mathbb{T}$
W	Number of wavelength in a fibre
B_r	Line rate at wavelength rate r; $r \in \mathbb{WR}$ (In Gbps)
B_M	Line rate of an aggregation port (In Gbps)
D_{mn}	Length of the physical link (m, n); $m \in \mathbb{N}, n \in \mathbb{N}_m$ (In km)
DA	Span between neighbouring EDFAs (In km)
A_{mn}	$= \left\lfloor \frac{D_{mn}}{DA} - 1 \right\rfloor + 2$, number of EDFAs in the physical link (m, n); $m \in \mathbb{N}, n \in \mathbb{N}_m$
R_r	Reach of regenerators at wavelength rate r; $r \in \mathbb{WR}$ (In km)
G_{mnr}	$= \left\lfloor \frac{D_{mn}}{R_r} - 1 \right\rfloor$, number of regenerators in the physical link (m, n) at wavelength rate r; $m \in \mathbb{N}, n \in \mathbb{N}_m, r \in \mathbb{WR}$
PR_r	Power consumption of a router port at wavelength rate r; $r \in \mathbb{WR}$
PR_M	Power consumption of an aggregation router port
PT_r	Power consumption of a transponder at wavelength rate r; $r \in \mathbb{WR}$
PO_m	Power consumption of an optical switch at core node m; $m \in \mathbb{N}$
PE	Power consumption of an EDFA
PS	Power consumption of a metro Ethernet switch port
P_{CS}	Power consumption of a content server per Gbps
C_S	Capacity of a content server
PUE_C	PUE of cloud data centres
PUE_F	PUE of fog data centres
PUE_N	PUE of core, metro, and access networking equipment
Z_{CDC}	Ratio to account for networking equipment power consumption in cloud data centres
Z_{FDC}	Ratio to account for networking equipment power consumption in fog data centres
P_{OLT}	Power consumption of an OLT
C_{OLT}	Capacity between an OLT and metro network
C_{FDC}	Capacity between an OLT and fog data centre
PG_r	Power consumption of a regenerator at wavelength rate r; $r \in \mathbb{WR}$
PS	Power consumption of a metro Ethernet switch port
VoD_{cdt}	Demands from CDC c to node d at time t; $c \in \mathbb{CDC}, d \in \mathbb{N}, t \in \mathbb{T}$ (In Gbps)
SP_{dt}	Available solar power per m^2 in node d at time t; $d \in \mathbb{N}, t \in \mathbb{T}$ (In Watts)
SSC	Size of a solar cell
M	A very large number
E_{MAX}	Battery maximum capacity
α	Charging percentage per hour
β	Discharging percentage per hour

Variables:

λ_{ijt}^{cd}	Traffic between node pair (c,d) passing through virtual link (i,j) at time t; $c \in \mathbb{CDC}, d \in \mathbb{N}, i,j \in \mathbb{N}, t \in \mathbb{T}, c \neq d$
C_{ijrt}	Number of wavelengths at rate r on the virtual link (i,j) at time t; $i,j \in \mathbb{N}, r \in \mathbb{WR}, t \in \mathbb{T}, i \neq j$
ω_{ijrt}^{mn}	Number of wavelengths at rate r of the virtual link (i,j) in the physical link (m,n) at time t; $i,j,m,n \in \mathbb{N}, r \in \mathbb{WR}, t \in \mathbb{T}, i \neq j$
F_{mnt}	Number of fibres used on the link (m,n) at time t; $m \in \mathbb{N}, n \in \mathbb{N}_m, t \in \mathbb{T}$
W_{mnrt}	Total number of wavelengths at rate r in the physical link (m,n) at time t; $m \in \mathbb{N}, n \in \mathbb{N}_m, r \in \mathbb{WR}, t \in \mathbb{T}$
CCQ_{ct}	Number of aggregation ports required to connect core node c with the CDC in c at time t; $c \in \mathbb{CDC}, t \in \mathbb{T}$
CMQ_{dt}	Number of aggregation ports required to connect core node d with the metro network in d at time t; $d \in \mathbb{N}, t \in \mathbb{T}$
MCQ_{ct}	Number of aggregation ports required to connect the metro network in c with the cloud data centre c at time t; $c \in \mathbb{CDC}, t \in \mathbb{T}$
MAQ_{dt}	Number of aggregation ports required to connect the metro network in d with the access network in d at time t; $d \in \mathbb{N}, t \in \mathbb{T}$
$VoDC_{cdt}$	Demands by users in node d that are served by CDC c at time t; $c \in \mathbb{CDC}, d \in \mathbb{N}, t \in \mathbb{T}$ (In Gbps)
$VoDF_{cdt}$	Demands by users in node d from cloud data centre c that is instead served by the FDC in d at time t; $c \in \mathbb{CDC}, d \in \mathbb{N}, t \in \mathbb{T}$ (In Gbps)
OLT_{dt}	Number of OLTs required in node d to accommodate VoD demands at time t; $d \in \mathbb{N}, t \in \mathbb{T}$
$VoDFS_{dst}$	Demands served in FDC d by server s powered by solar at time t; $d \in \mathbb{N}, s \in \mathbb{S}, t \in \mathbb{T}$ (In Gbps)
$VoDFB_{dst}$	Demands served in FDC d by server s powered by brown sources at time t; $d \in \mathbb{N}, s \in \mathbb{S}, t \in \mathbb{T}$ (In Gbps)
$VoDFE_{dst}$	Demands served in FDC d by server s powered by stored solar power at time t; $d \in \mathbb{N}, s \in \mathbb{S}, t \in \mathbb{T}$ (In Gbps)
E_{dt}	Energy stored in the battery at FDC d at time t; $i \in \mathbb{N}, t \in \mathbb{T}$
RS_{dt}	Energy to be charged in the battery from the surplus renewable energy at FDC d at time t; $i \in \mathbb{N}, t \in \mathbb{T}$
ED_{dt}	Energy to be discharged from battery to the FDC d at time t; $d \in \mathbb{N}, t \in \mathbb{T}$

2. Transponders:

$$P_{t(T)} = \sum_{m \in \mathbb{N}} \sum_{n \in \mathbb{N}_m} \sum_{r \in \mathbb{WR}} PT_r \, W_{mnrt}. \tag{7.2}$$

3. Optical switches:

$$P_{t(O)} = \sum_{m \in \mathbb{N}} PO. \tag{7.3}$$

Table 7.1: Parameters for the cloud transport network and fog and cloud data centres.

Set of wavelength rates ($\mathbb{WR}$)	40,100,400,1000 Gbps		
Span between two neighbouring EDFAs (DA)	80 km		
Number of wavelengths in a fibre (W) [116]	32		
The duration between consecutive hours in $\mathbb{T}$ (S)	2 hours		
Reach of regenerator (R_r) and power consumption of a regenerator (PG_r) at a wavelength rate r; [42] $r \in \mathbb{WR}$	r (Gbps)	R_r (km)	PG_r (W)
	40	2500	71.4
	100	1200	221.8
	400	400	857.4
	1000	350	2065.2
Power consumption of a router port (PR_r) at wavelength rate r; $r \in \mathbb{WR}$ [42]	r (Gbps)		PR_r (W)
	40		178.2
	100		309.3
	400		367.8
	1000		425.1
Power consumption of a transponder (PT_r) at wavelength rate r; $r \in \mathbb{WR}$ [42]	r (Gbps)		PT_r (W)
	40		35.7
	100		110.9
	400		428
	1000		1032.6
Power consumption of an optical switch (PO_m) at core node m; $m \in \mathbb{N}$ [116]	85 W		
Power consumption of an EDFA (PE) [42]	15.3 W		
Power consumption of a metro Ethernet switch port at a rate of 40 Gbps (PS) [280]	50 W		
Power consumption of a content server per Gbps (P_{CS}) [124]	211.1 W/Gbps		
Capacity of a content server (C_S) [124]	1.8 Gbps		
PUE of cloud data centres (PUE_C)	1.1		
PUE of fog data centres (PUE_F)	1.1-1.3		
PUE of core, metro, and access networking equipment (PUE_N) [42]	1.5		
Ratio to account for networking equipment power consumption in fog data centres (Z_{FDC}) [187]	1-1.3		
Ratio to account for networking equipment power consumption in cloud data centres (Z_{CDC}) [187]	1.3		
Power consumption of an OLT (P_{OLT}) [217]	904 W		
Total capacity of links between OLT and metro network (C_{OLT})	160 Gbps		
Total capacity of links between OLT and fog data centre (C_{FDC})	160 Gbps		
Size of a solar cell per OLT (SSC)	50, 100, 150, 200, 250 m^2		
A very large number (M)	10000000000		
Battery maximum capacity (E_{MAX}) [281]	100 kWh		
Charging efficiency during S (α) [279]	72.25%		
Discharging efficiency during S (β) [279]	90.25%		

Table 7.2: Solar power availability per m^2 in Watts recorded in February 2018 from different cities in the NSFNET network.

$\mathbb{N}$ \ $\mathbb{T}$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	Seattle, WA	***Palo Alto, CA***	***San Diego, CA***	***Salt Lake, UT***	***Boulder, CO***	***Houston, TX***	***Lincoln, NE***	***Champaign, IL***	***Pittsburgh, PA***	***Atlanta, GA***	***Ann Arbor, MI***	***Ithaca, NY***	***College Park, MD***	***Princeton, NJ***
00:00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
02:00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
04:00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
06:00	12.13	22.8	45.1	9.66	24.9	21.5	9.94	30.5	13.31	9.60	5.93	18.90	22.3	26.110
08:00	105.46	141.00	176.00	112.00	139.00	151.00	113.00	146.00	120.60	122.74	100.00	126.00	137.00	140.48
10:00	178.53	227.00	255.00	203.00	218.00	252.00	203.00	220.00	208.29	227.45	192.00	204.00	220.00	217.69
12:00	191.34	241.00	257.00	228.00	227.00	277.00	228.00	225.00	229.22	261.87	221.00	217.00	233.00	225.59
14:00	140.41	181.00	181.00	181.00	165.00	219.00	180.00	158.00	177.71	216.68	179.00	159.00	173.00	162.06
16:00	41.70	62.60	50.60	75.30	49.80	95.00	74.30	42.20	68.06	104.17	78.90	49.50	57.10	46.88
18:00	0.00	0.00	0.00	0.40	0.00	1.16	0.33	0.00	0.03	3.65	0.98	0.00	0.00	0.00
20:00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22:00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

4. EDFAs:

$$P_{t(E)} = PE \sum_{m \in \mathbb{N}} \sum_{n \in \mathbb{N}_m} A_{mn} F_{mnt}. \tag{7.4}$$

5. Regenerators:

$$P_{t(R)} = \sum_{m \in \mathbb{N}} \sum_{n \in \mathbb{N}_m} PG_r G_{mnr} W_{mnrt}. \tag{7.5}$$

Then, $P_{t(IPoverWDM)} = PUE_N \left(P_{t(IP)} + P_{t(T)} + P_{t(O)} + P_{t(E)} + P_{t(R)} \right)$.

The power consumption of metro and access networks, FDCs and CDCs at time t is composed of the following:

1. Metro router and Ethernet switch ports:

$$P_{t(Metro)} = PUE_N \left(\sum_{i \in \mathbb{N}} PS \ (CMQ_{it} + MCQ_{it} + MAQ_{it}) \right). \tag{7.6}$$

2. OLTs in Access Network:

$$P_{t(Access)} = PUE_N \sum_{d \in \mathbb{N}} P_{OLT} OLT_{dt}. \tag{7.7}$$

3. FDCs and CDCs:

$$P_{t(FDC)} = P_{CS} PUE_F Z_{FDC} \sum_{c \in \mathbb{CDC}} \sum_{d \in \mathbb{N}} VoDF_{cdt}. \tag{7.8}$$

$$P_{t(CDC)} = P_{CS} PUE_C Z_{CDC} \sum_{c \in \mathbb{CDC}} \sum_{d \in \mathbb{N}} VoDC_{cdt}. \tag{7.9}$$

Objective: Minimise the total brown energy consumption, (PC_b), subject to the following constraints:

1. Flow conservation in IP layer: The allocation of virtual links to the demands follows the flow conservation law.

$$\sum_{j \in \mathbb{N}, i \neq j} \lambda_{ijt}^{cd} - \sum_{j \in \mathbb{N}, i \neq j} \lambda_{jit}^{cd} = \begin{cases} VoDC_{cdt} & i = c \\ -VoDC_{cdt} & i = d \\ 0 & otherwise, \end{cases} \tag{7.10}$$

$$\forall c \in \mathbb{CDC}, d \in \mathbb{N}, i \in \mathbb{N}, t \in \mathbb{T}, c \neq d.$$

2. Flow conservation in optical layer: Allocation of wavelengths to virtual demands follows flow conservation law.

$$\sum_{n\in\mathbb{N}_m} \omega_{ijrt}^{mn} - \sum_{n\in\mathbb{N}_m} \omega_{jirt}^{mn} = \begin{cases} C_{ijrt} & m = i \\ -C_{ijrt} & m = j \\ 0 & otherwise, \end{cases} \tag{7.11}$$

$$\forall i, j, m \in \mathbb{N}, r \in \mathbb{WR}, t \in \mathbb{T}, i \neq j.$$

3. Virtual IP link capacity constraint: To ensure that traffic flows through a virtual link do not exceed its capacity.

$$\sum_{c\in\mathbb{CDC}} \sum_{d\in\mathbb{N},c\neq d} \lambda_{jit}^{cd} \leq \sum_{r\in\mathbb{WR}} C_{ijrt}\, B_r, \tag{7.12}$$

$$\forall i, j \in \mathbb{N}, t \in \mathbb{T}, i \neq j.$$

4. Capacity constraints: Constraint 7.13 ensures that wavelengths in the physical link do not exceed the maximum capacity of the fibres. Constraint 7.14 calculates W_{mnrt}.

$$\sum_{i\in\mathbb{N}} \sum_{j\in\mathbb{N},i\neq j} \sum_{r\in\mathbb{WR}} \omega_{ijrt}^{mn} \leq W\, F_{mnt}, \forall m \in \mathbb{N}, n \in \mathbb{N}_m, t \in \mathbb{T}. \tag{7.13}$$

$$\sum_{i\in\mathbb{N}} \sum_{j\in\mathbb{N},i\neq j} \omega_{ijrt}^{mn} = W_{mnrt}, \tag{7.14}$$

$$\forall m \in \mathbb{N}, n \in \mathbb{N}_m, r \in \mathbb{WR}, t \in \mathbb{T}, m \neq n.$$

5. Aggregation ports constraints: Constraints 7.15, and 7.16 determine the aggregation ports required in the core node. Constraint 7.15 determines the number of aggregation ports required to connect core node c; $c \in \mathbb{CDC}$ with the cloud data centre c located nearby, while Constraint 7.16 specifies the number of aggregation ports that are required to connect core node d with the networking equipment of the metro network at d; $d \in \mathbb{N}$. This is required to deliver VoD demands from CDCs in other nodes. Constraints 7.17, and 7.18 calculates the remaining aggregation ports in the metro network. Constraint 7.17 calculates the number

of aggregation ports required to connect the cloud data centre c with the metro node in c; $c \in \mathbb{CDC}$. Constraint 7.18 calculates the number of aggregation ports required to connect metro networking equipment at node d with the access network located in node d; $d \in \mathbb{N}$. This is required to deliver the total VoD demands to users in node d.

$$B_M \, CCQ_{ct} = \sum_{d \in \mathbb{N},\, c \neq d} VoDC_{cdt}, \forall c \in \mathbb{CDC}, t \in \mathbb{T}. \tag{7.15}$$

$$B_M \, CMQ_{dt} = \sum_{c \in \mathbb{CDC}, c \neq d} VoDC_{cdt}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.16}$$

$$B_M \, MCQ_{ct} = \sum_{d \in \mathbb{N}, c = d} VoDC_{cdt}, \forall c \in \mathbb{CDC}, t \in \mathbb{T}. \tag{7.17}$$

$$B_M \, MAQ_{dt} = \sum_{c \in \mathbb{CDC}} VoDC_{cdt}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.18}$$

6. Number of OLTs in the access network: Constraint 7.19 determines the number of OLTs required.

$$OLT_{dt} = \sum_{c \in \mathbb{CDC}} VoD_{cdt}/C_{OLT}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.19}$$

7. Demands distribution: Constraint 7.20 ensures that the sum of the demands served by CDCs and the demands served by FDCs is equal to the total demands.

$$VoDC_{cdt} + VoDF_{cdt} \, OLT_{dt} = VoD_{cdt}, \forall c \in \mathbb{CDC}, d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.20}$$

8. OLT capacity: Constraint 7.21 ensures that FDC demands do not exceed the capacity of its links to OLT.

$$\sum_{c \in \mathbb{CDC}} VoDF_{cdt} \leq C_{FDC}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.21}$$

9. Servers in FDCs: Constraint 7.22 ensures that demands per server do not exceed its capacity. Constraint 7.23 equates all servers demands with the total FDC

demands.

$$VoDFS_{dst} + VoDFB_{dst} + VoDFE_{dst} \leq C_S, \forall d \in \mathbb{N}, s \in \mathbb{S}, t \in \mathbb{T}. \tag{7.22}$$

$$\sum_{s \in \mathbb{S}} (VoDFS_{dst} + VoDFB_{dst} + VoDFE_{dst}) = \sum_{c \in \mathbb{CDC}} VoDF_{cdt}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.23}$$

10. Solar power: Constraint 7.24 ensures that nodes do not exceed available solar power.

$$\sum_{s \in \mathbb{S}} VoDFS_{dst}\, P_{CS}\, PUE_F\, Z_{FDC} \leq SP_{dt}\, SSC, \quad \forall d \in \mathbb{N}, s \in \mathbb{S}, t \in \mathbb{T}. \tag{7.24}$$

11. Discharge limit: Constraint 7.25 ensures that the energy discharge does not exceed the amount stored in the ESD. ED_{dt} at $t = 0, \forall d \in \mathbb{N}$ is assumed to be zero.

$$ED_{dt} \leq E_{dt}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.25}$$

12. Charge limit: Constraint 7.26 ensures that the stored energy is within the remaining capacity of the ESD.

$$RS_{dt} \leq E_{MAX} - E_{dt}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.26}$$

13. Energy storage constraints: Constraint 7.27 relates the energy stored in ESDs at t; $t \in \mathbb{T}$ with the energy stored at $t - S; t \in \mathbb{T}$. Constraint 7.28 ensures that the energy stored in the ESD is within the maximum capacity [279].

$$E_{dt} = \begin{cases} [E_{d\,(t-S)} - ED_{d\,(t-S)} + \alpha\, RS_{d\,(t-S)}] & t \neq 0 \\ 0 & t = 0, \end{cases} \quad \forall i \in \mathbb{N}, t \in \mathbb{T}. \tag{7.27}$$

$$E_{dt} \leq E_{MAX}, \forall d \in \mathbb{N}, t \in \mathbb{T}. \tag{7.28}$$

14. Energy discharge: Constraint 7.29 ensures that the stored energy used does not exceed the available battery energy.

$$S \sum_{s \in \mathbb{S}} VoDFE_{dst}\, P_{CS}\, PUE_F\, Z_{FDC} \leq \beta\, ED_{dt}, \tag{7.29}$$
$$\forall d \in \mathbb{N}, t \in \mathbb{T}$$

15. Surplus renewable energy: Constraint 7.30 specifies the surplus renewable energy to be stored into the battery.

$$S \times SSC \times SP_{dt} = RS_{dt} + S \sum_{s \in \mathbb{S}} VoDFS_{dst}\, P_{CS} \cdot PUE_F\, Z_{FDC}, \tag{7.30}$$
$$\forall d \in \mathbb{N}, t \in \mathbb{T}$$

7.4 Results and Discussions

7.4.1 Power Consumption with Brown-powered CDCs and FDCs

We start by evaluating the brown power consumption (PC_B) at time t required to optimally deliver VoD demands in terms of power consumption efficiency from brown-powered cloud and fog data centres for different values of PUE_F. In this case:

$$PC_{Bt} = \left(P_{t(IPoverWDM)} + P_{t(Metro)} + P_{t(Access)} + P_{t(CDC)} + P_{t(FDC)}\right), \tag{7.31}$$

and only Constraints 7.10 to 7.23 are considered while setting the variables $VoDFS_{dst}$, and $VoDFE_{dst}$ equal to zero; $\forall d \in \mathbb{N}, s \in \mathbb{S}, t \in \mathbb{T}$. For all the cases in this evaluation, PUE_C is set to 1.1, and Z_{CDC} and Z_{FDC} are set to 1.3. Figures 7.3, 7.4, 7.5, and 7.6 show the PC_{Bt} at the time of the day when considering PUE_F values of 1.25, 1.2, 1.15, and 1.1, respectively. The results show that for PUE_F of 1.25, delivering fully from CDCs is the most efficient approach. As PUE_F improves, it becomes more efficient to deliver partially from FDCs. When PUE_F is equivalent to PUE_C, it becomes more efficient to fully stream from FDCs as $P_{(FDC)}$ and $P_{(CDC)}$ required to deliver the same amount of traffic will be equivalent, and the power consumption of the transport network will be the factor that determines the differences in PC_{Bt}.

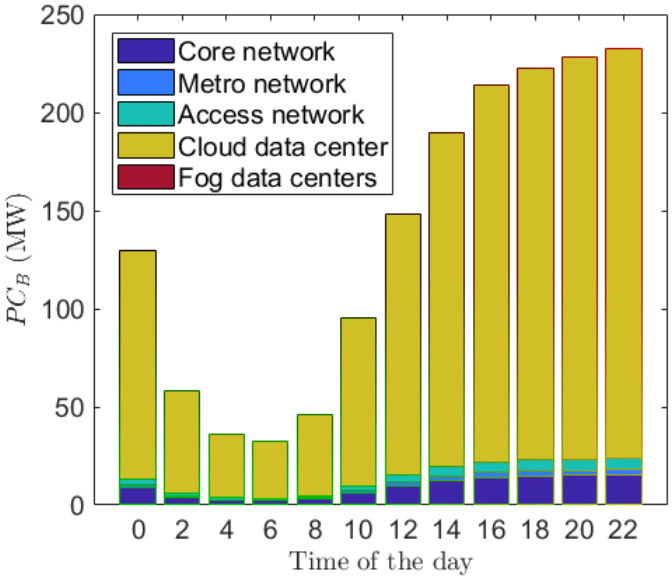


Figure 7.3: Brown power consumption (PC_{Bt}) for a PUE_F of 1.25.

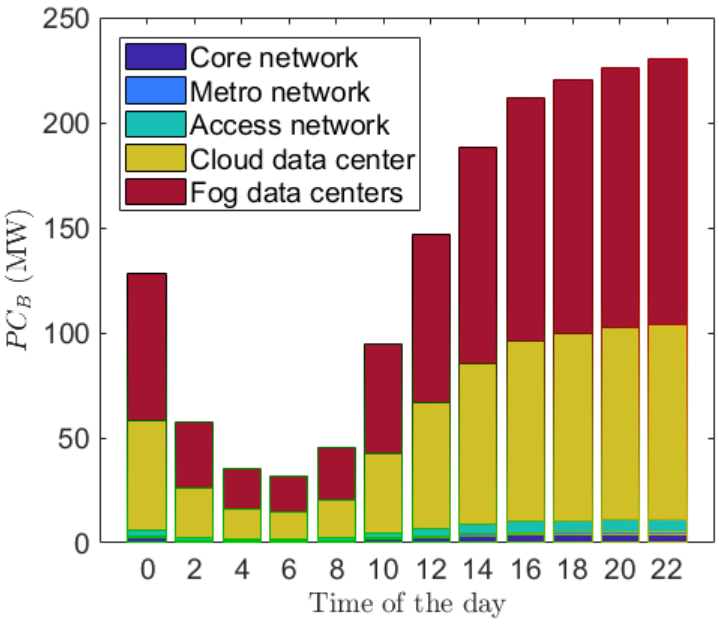


Figure 7.4: Brown power consumption (PC_{Bt}) for a PUE_F of 1.2.

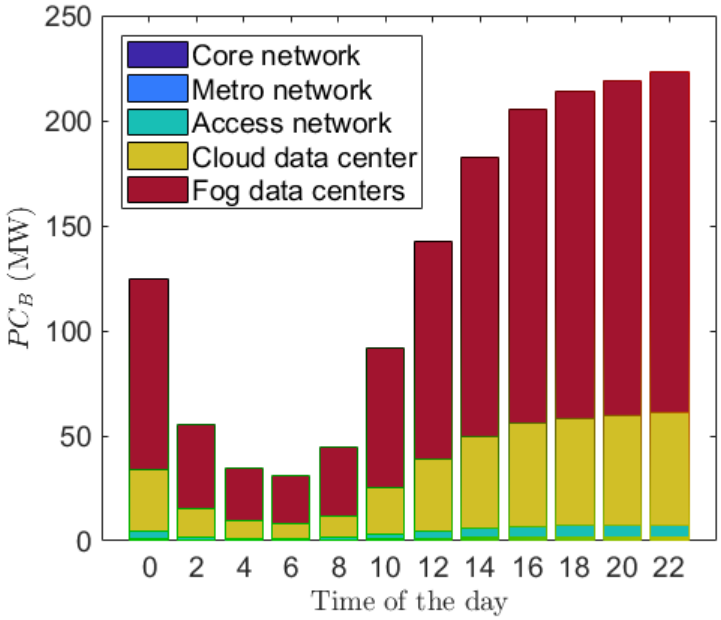


Figure 7.5: Brown power consumption (PC_{Bt}) for a PUE_F of 1.15.

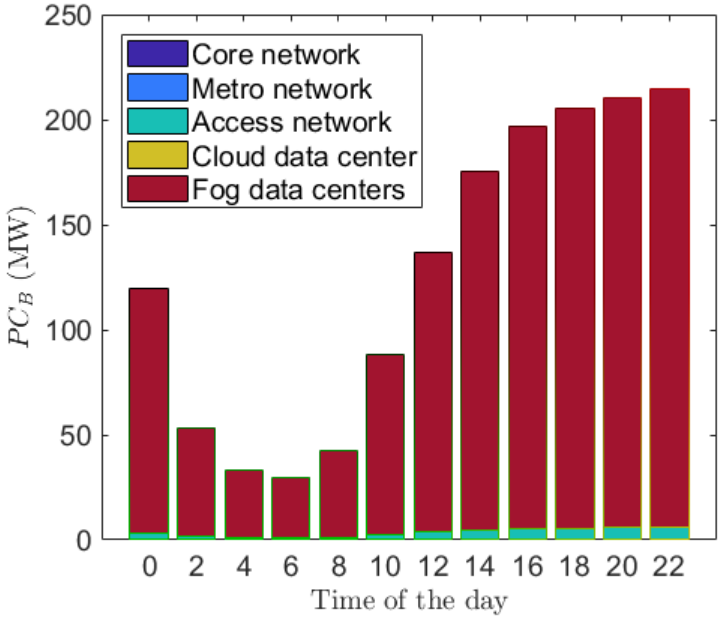


Figure 7.6: Brown power consumption (PC_{Bt}) for a PUE_F of 1.1.

Figures 7.7 - 7.10 show the total amount of traffic served from the cloud data centres (i.e. $\sum_{c\in\mathbb{CDC},t\in\mathbb{T}} VoDC_{cdt}$) and from the fog data centres (i.e. $\sum_{c\in\mathbb{CDC},t\in\mathbb{T}} VoDF_{cdt}$) at different nodes d in the NSFNET network. Figure 7.7 shows that when PUE_F is as high as 1.25, the fog data centres are not selected to serve the traffic at any node. In this case, the total brown networking power consumption (PC_N) which can be expressed as:

$$PC_N = \sum_{t\in\mathbb{T}} \left(P_{t(IPoverWDM)} + P_{t(Metro)} + P_{t(Access)} \right), \tag{7.32}$$

was found to be about 167.545 MW. Figure 7.8 shows that at PUE_F of 1.2, about half of the traffic is served from the fog data centres. The savings in the total brown networking power consumption was found to be 53% compared to the case where the optimal delivery is from the cloud data centres only. Figure 7.9 shows that when PUE_F is further reduced to 1.15, the majority of the traffic is served from the fog data centre (i.e. about 75% of the VoD traffic). Also, it shows that the cloud data centres are selected to serve the nodes that contain them (i.e. at nodes 2, 3, 7, 8, and 9) as serving at these locations will have less brown networking power consumption compared to serving at the remaining nodes in the NSFNET network. In this case, the savings in the total brown networking power consumption was found to be 67% compared to the case of serving fully from the cloud data centres. Finally, Figure 7.10 shows that the traffic is fully delivered from the fog data centres. The savings in the total brown networking power consumption in this case was found to be 75% compared to the same base case of delivering the traffic fully from the cloud data centres.

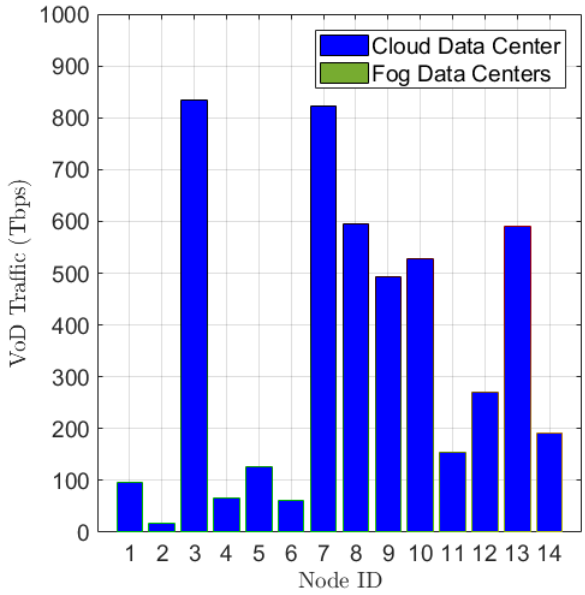


Figure 7.7: Volumes of cloud-served and fog-served VoD traffic for PUE_F of 1.25.

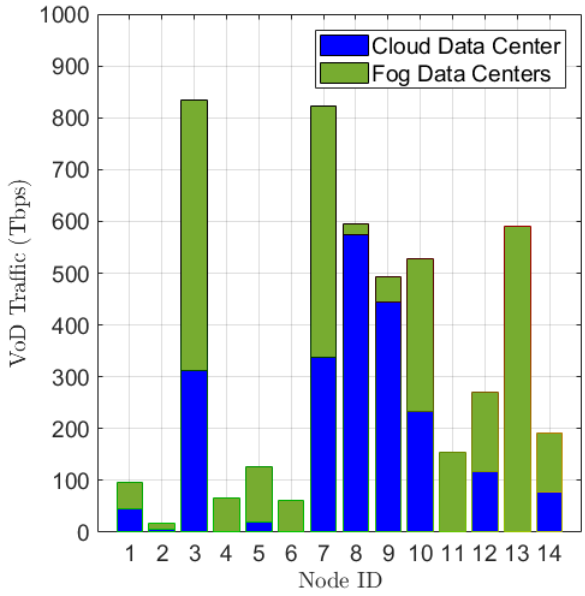


Figure 7.8: Volumes of cloud-served and fog-served VoD traffic for PUE_F of 1.2.

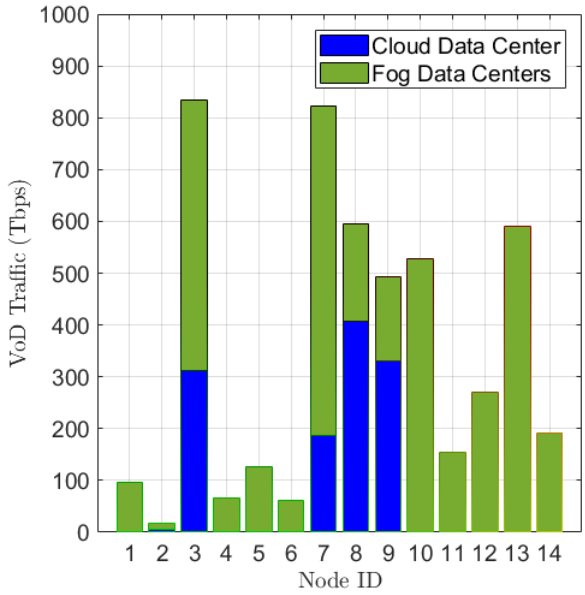


Figure 7.9: Volumes of cloud-served and fog-served VoD traffic for PUE_F of 1.15.

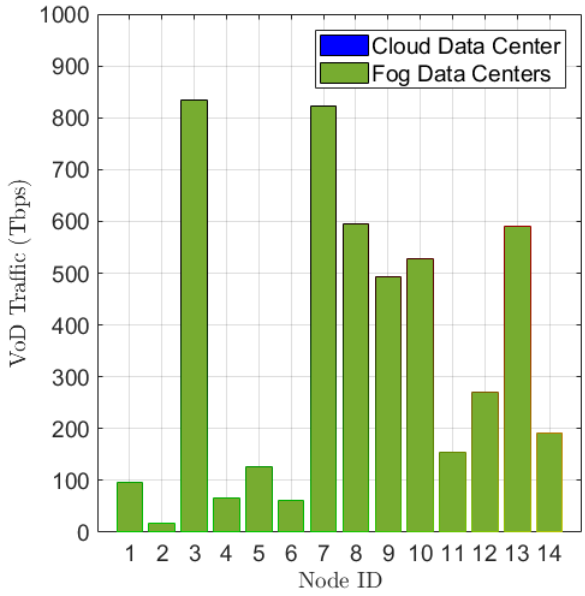


Figure 7.10: Volumes of cloud-served and fog-served VoD traffic for PUE_F of 1.1.

7.4.2 Power Consumption with Fully Renewable-powered CDCs and Solar-powered FDCs

We now consider fully renewable-powered CDCs and solar-powered FDCs with PUE_F equal to 1.1 and solar cells of different capacities. In this case:

$$PC_{Bt} = P_{t(IPoverWDM)} + P_{t(Metro)} + P_{t(Access)} + P_{CS}\, PUE_F\, Z_{FDC} \sum_{s\in\mathbb{S},d\in\mathbb{N}} VoDFB_{dst}\, OLT_{dt}, \tag{7.33}$$

and only constraints 7.10 to 7.24 are considered while setting $VoDFE_{dst} = 0, \forall d \in \mathbb{N}, s \in \mathbb{S}$. Figures 7.11 - 7.12 show the total brown power consumption (i.e. PC_{Bt}) at the time of the day when considering different sizes for the solar cells (i.e. SSC). The results when SSC is equal to 50 m^2 (i.e. in Figure 7.11) indicate a total reduction in the brown power consumption by 15% compared to the case of fully delivering from the cloud data centres. When SSC is equal to 150 m^2 (i.e. the results in Figure 7.12), the total saving in the brown power consumption was found to be 26%. For SSC of 250 m^2 (i.e. the results in Figure 7.13), the reduction in the brown power consumption was found to be 33%. It can be noticed from Figure 7.12, and 7.13 that the brown power consumption is no longer proportional to the total traffic in Figure 7.2. The reduction in the brown power consumption is achieved only between 6:00 and 18:00 during the availability hours of solar power as presented in Table 7.2. For SSC of 250 m^2, the high availability of the solar power between 10:00 and 12:00 enabled almost complete delivery from fog data centres resulting in negligible brown power consumption in the metro and core networks. Figures 7.14 - 7.16 show the total amount of traffic served from the cloud data centres and from the fog data centres while being powered by brown sources (i.e. $\sum_{s\in\mathbb{S},d\in\mathbb{N}} VoDFB_{dst}\, OLT_{dt}$) or solar cells with different capacities (i.e. $\sum_{s\in\mathbb{S},d\in\mathbb{N}} VoDFS_{dst}\, OLT_{dt}$). It was observed that utilising brown power for the fog data centres is not optimal for the objective of reducing the total brown power consumption when using the above mentioned parameters, in addition to the parameters in Table 7.1.

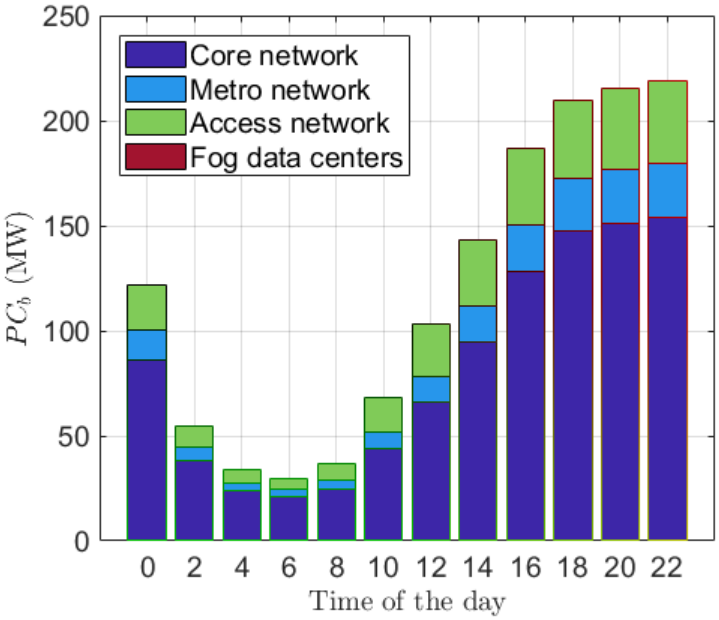


Figure 7.11: Brown power consumption (PC_{Bt}) for a SSC of 50 m^2.

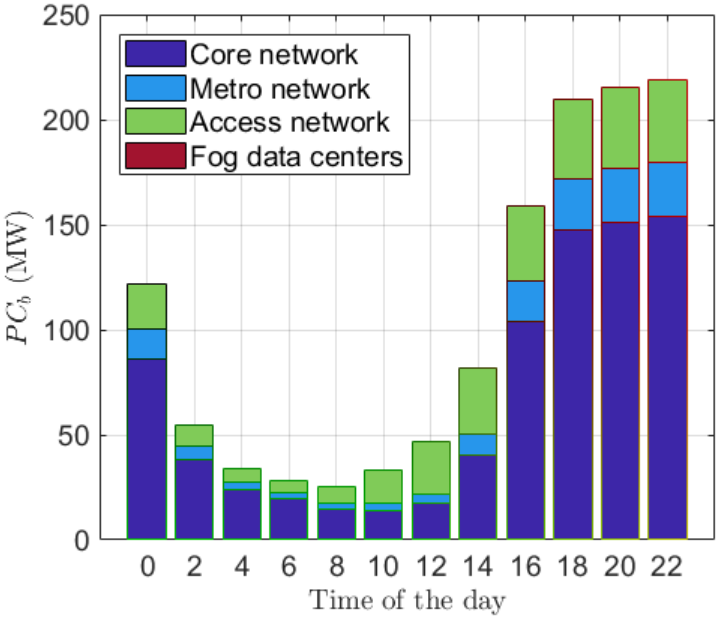


Figure 7.12: Brown power consumption (PC_{Bt}) for a SSC of 150 m^2.

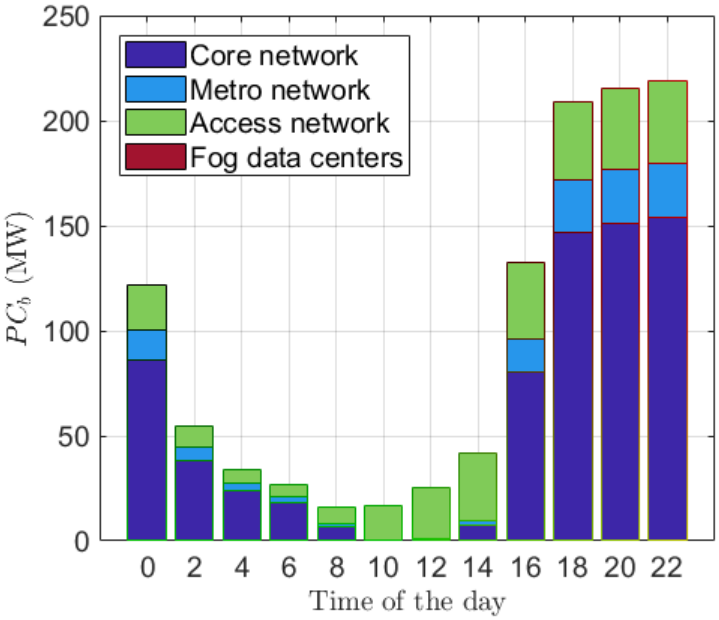


Figure 7.13: Brown power consumption (PC_{Bt}) for a SSC of 250 m^2.

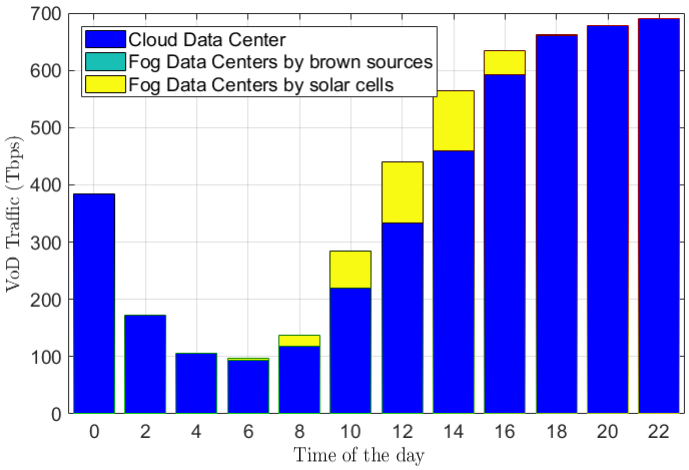


Figure 7.14: Volumes of cloud-served and fog-served VoD traffic for a SSC of 50 m^2.

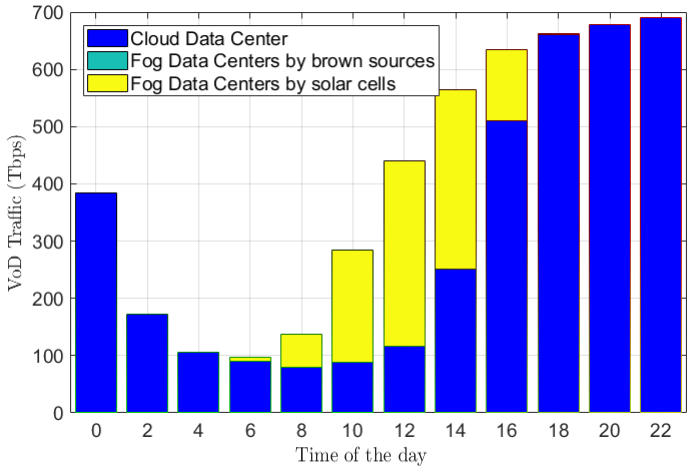


Figure 7.15: Volumes of cloud-served and fog-served VoD traffic for a SSC of 150 m^2.

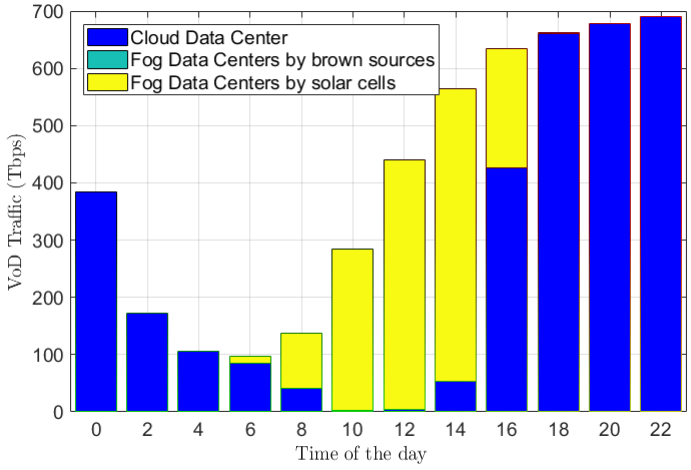


Figure 7.16: Volumes of cloud-served and fog-served VoD traffic for a SSC of 250 m^2.

7.4.3 Power Consumption with Fully Renewable-powered CDCs and Solar-powered FDCs with ESDs

In the case of renewable-powered cloud data centres and fog data centres with solar cells and energy storage devices (ESDs), we consider Constraints 7.10 to 7.30. The delivery of VoD demands is optimised so that it is from the cloud data centres or from the fog data centres with PUE_F of 1.1 and SSC of 250 m^2 while considering

the usage of an ESD with a capacity of 100 kWh. Figure 7.17 shows the total brown power consumption (i.e. PC_{Bt}) at the time of the day. In this case, the reduction in the total brown power consumption compared to the case of fully streaming from cloud data centres in 43%. The additional reduction in the brown power is due to optimising the direct use of solar power in the fog data centres and charging the ESD for use when the solar power is not available. Figure 7.18 shows the total amount of traffic served from the cloud data centres and from the fog data centres while being powered by brown sources (i.e. $\sum_{s\in\mathbb{S},d\in\mathbb{N}} VoDFB_{dst}\, OLT_{dt}$), directly by the solar cells (i.e. $\sum_{s\in\mathbb{S},d\in\mathbb{N}} VoDFS_{dst}\, OLT_{dt}$), or with ESD (i.e. $\sum_{s\in\mathbb{S},d\in\mathbb{N}} VoDFE_{dst}\, OLT_{dt}$). It shows that the optimal use of ESDs is between 14:00 and 22:00.

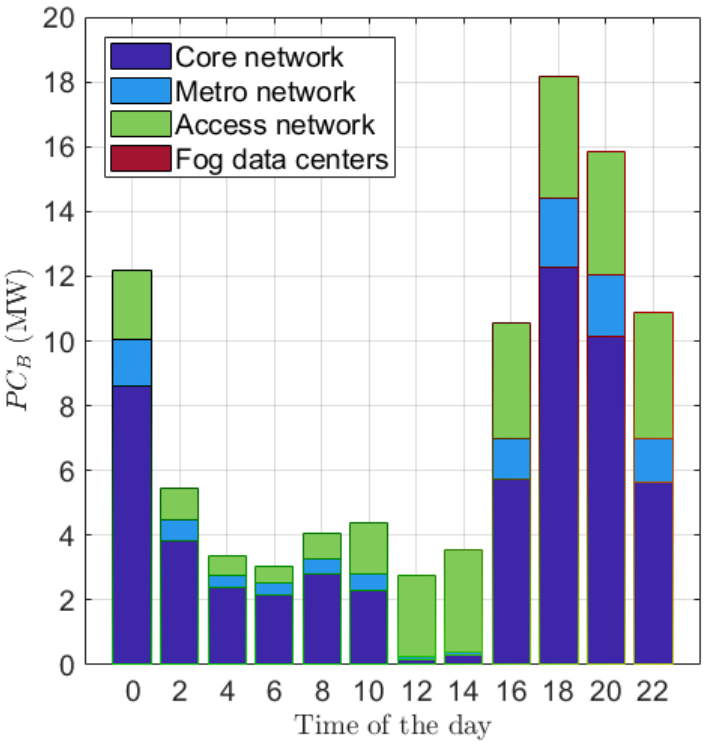


Figure 7.17: Brown power consumption (PC_{Bt}) for a SSC of 250 m^2 and E_{MAX} of 100 kWh.

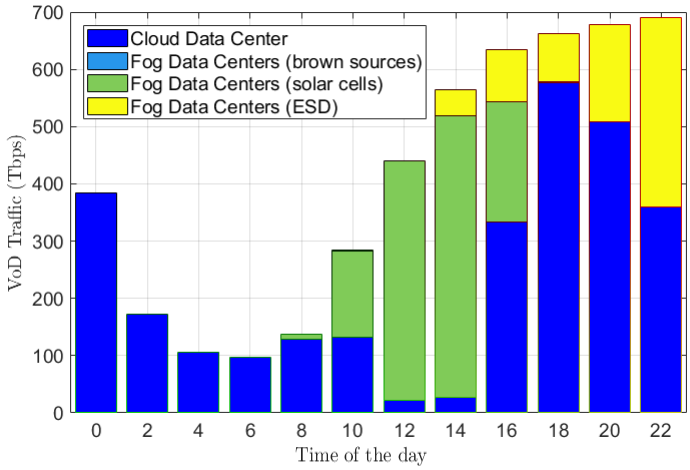


Figure 7.18: Volumes of cloud-served and fog-served VoD traffic for a SSC of 250 m^2 and E_{MAX} of 100 kWh.

7.5 Summary

This Chapter presents a comprehensive optimisation model for delivering VoD services from cloud data centres or distributed fog data centres in the access network with solar cells and ESDs. The model considers reducing the total brown power consumption which includes the brown power used by the brown-powered data centres in addition to the transport network with PON access network, Ethernet metro network, and IP over WDM core network. For the IP over WDM network, 2020 equipment power consumption was assumed, optical bypassing and MLR were considered to examine the reduction in the brown power consumption while considering efficient future networks. Different scenarios were considered for powering the cloud and fog data centres. For the first scenario (i.e. brown powered cloud and fog data centres), the results show that as the PUE_F reduces, it becomes more energy efficient to deliver from fog data centres. WhenPUE_F is equivalent to PUE_C, it is more efficient to deliver fully from fog data centres. In this case, the reduction in the brown power consumption is 75% compared to the case of delivering fully from the cloud data centres. As many cloud providers are utilising renewable power for their data centres, we also examined the optimisation

when the cloud data centres are fully powered by renewable sources and the fog data centres are solar-powered. The results indicated that savings by up to 33% can be achieved when considering 250 m^2 solar cells (which is considered to be of a suitable size for central offices in the access network) for the fog data centres. Additional saving of about 10% can be achieved when also considering ESDs with capacity of 100 kWh to store surplus solar energy.

The results presented in this chapter reflect the need of joint optmisation for the routing in transport networks and the usage of cloud and fog data centres when the objective is to reduce the non-renewable power consumption. A first main finding is that building fog data centres with worst power usage efficiency compared to existing cloud data centres is not optimal, thus investing in these distributed data centres is required to improve the energy efficiency of delivering VoD traffic. When considering renewable solar power, more savings can be achieved with larger solar cells and energy storage devices.

Chapter 8

Conclusions and Future Work

This chapter provides the conclusions and contributions for the work presented in this thesis and also provides some future research directions

8.1 Conclusions

This thesis addressed some of the challenges associated with deploying big data applications in intra data centre networks and in delivering the increasing video traffic volumes in transport networks. The performance of big data applications is determined by the underlying hardware specifications such as CPU, memory, and disk, cluster networking topology, selected programming framework, in addition to the configurations selected and mechanisms for data and job placements and tasks scheduling. The energy efficiency of a data centre is highly related to the topology and the equipment utilised. Chapter 4 introduced two energy efficient PON-based designs proposed in [9, 207] for use in fog and cloud data centres. The AWGR-centric data centre design was optimised through a MILP model that determined the topology, and the wavelength routing and assignment for communication between the racks in a single PON cell. For big data frameworks, this thesis considered evaluating the completion time and the energy efficiency of the shuffling phase in MapReduce while considering four state-of-the-art data centres in addition to the two PON-based designs. In Chapter 5 and Chapter 6, we considered big data application parameters such as the allocation of the map and reduce tasks, the replication factor, the data volume, and the data skewness. Concurrently, we considered data centre parameters such as the rate per server, the topology, the power consumption of the equipment used, and the resilience of links, switches and servers. The results presented reveal that the topology of the architecture has a significant impact on the performance and energy efficiency of MapReduce. Using the PON-based data centres improves the energy efficiency by about 83% compared to the electronic state-of-the-art data centres. The best performance and energy efficiency are achieved by the AWGR-based data centre design as it uses passive devices and utilise WDM technology. When failures in links, switches, and servers occur, server-centric data centres have less impact on performance and energy efficiency as optimising the routing and scheduling under different failures is aided by the availability of more routes between different servers. However, routing and scheduling in such data centres is more

complex. One suggestion to improve the resilience of the design of widely used data centres such as spine-leaf is to have dual links to and from each server, which can increase the over-subscription ratio. The results when considering a replication factor of two indicate that integrating the optimisation of the routing and scheduling in data centres and the application parameters can further improve the performance of the aplication and the energy efficiency of the data centre. For the video-on- demand, the work in Chapter 7 addressed the optimisation of the delivery of content from cloud data centres or distributed fog data centres in the access networks. The work presented in this Chapter considered several factors such as the design of the core network, the PUE of the cloud and fog data centres, the consideration of renewable energy in cloud data centres, and solar energy with ESDs in the fog data centres. The results presented in this chapter reflect that for non-renewable power consumption reduction when optimising services delivery, joint optmisation for the routing in transport networks and the usage of cloud and fog data centres is to be considered. Using fog data centres is more efficient only if they have similar power usage effectiveness to existing cloud data centres, thus investing in the infrastructure of these distributed data centres is required. This is to be widely considered as the use of fog computing is increasingly suggested due to reduced latency. When considering renewable solar power, more savings can be achieved with larger solar cells and energy storage devices.

8.2 Thesis Contributions

This thesis makes the following contributions to the knowledge:

1. It develops several MILP models that can be used to design the topologies and architectures, routing, and scheduling protocols of data centres. Chapter 4 provides a MILP model that can be used to obtain the optimum connections and wavelength routing and assignment in a proposed AWGR-centric optical data centre architecture. The AWGR-centric PON-based data centre uses passive devices such splitters and AWGRs with tunable optical sources and broadband

optical receivers able to receive a wide range of wavelengths.

2. It designs and optimises in Chapter 5 a time-slotted data centre architecture. A MILP model was developed and used to optimise the routing and scheduling of flows in the data centre network. The MILP optimisation minimises the completion time of tasks (an essential performance metric) and the total energy required to achieve this performance. This model is used to compare the performance and energy efficiency of six data centre networks when considering sort workloads and different objectives.

3. It evaluates through MILP optimisation the resilience of a range of data centre architectures when used to support big data applications. Chapter 6 provides a resilience study based on the MILP model and examines the impact of link and switch failure on the performance and energy efficiency of three data centre architectures. This Chapter also modifies the MILP model developed in Chapter 5 by introducing the added dimension of data replication to provide additional resilience and to examine the impact of server failures on the three data centre architectures.

4. Finally, in Chapter 7 the thesis reports a comprehensive MILP model developed to optimise the delivery of VoD traffic to users in the access network from cloud data centres or fog data centres. This chapter provides results for the reduction of the total brown energy consumption required to deliver the VoD traffic when considering fog data centres.

8.3 Future Research Directions

In what follows, we list some future research directions and their relation to the work presented in this thesis:

- **PON-based data centre designs:** The work in Chapter 4 introduced two PON-based data centre designs. One of future research directions is to enhance

and/or propose new data centre architectures to improve the performance and the energy efficiency of big data applications with increased traffic. Developing a routing protocol for a proposed data centre design and comparing the performance results with MILP-based routing will aid in adopting the design by the industry.

- **Data centre scales and big data volumes:** The work in Chapter 5 and Chapter 6 considered data centres with 16 servers and big data volumes of up to 120 Gbits due to the complexity of modelling larger scale data centres and more time slots and the memory requirement of running the MILP optimisation using AMPL. Also, the results were generated using a limited set of predefined allocation for the map and reduce tasks. Related future work can include averaging the results of several allocations and the development of heuristics to perform the same optimisation and the automation of constructing larger scale data centres and workloads.

- **Workload characteristics and their modelling:** The work in Chapter 5 and Chapter 6 considered a single type of big data workloads, which is sort batch workloads, to examine the resultant networking bottlenecks related to the data centre topology and the impact on the completion time and energy efficiency. Related future work can include the consideration of other MapReduce workloads and other streaming applications where the latency is more critical than the bandwidth.

- **Queueing delay in data centre networks:** The work in Chapter 5 and Chapter 6 considered the transmission delay and assumed that the queue and processing delays are much smaller than the transmission delay. For workloads other than batch workloads, using the models presented there, the consideration of these additional delays is important.

- **VoD content popularity and battery life time:** The work in Chapter 7 can be improved by considering the VoD content popularity which can result in higher reduction in the brown power consumption if these contents are cached

in the fog data centres. Also, the lifetime of ESDs defined by charge/discharge cycles should be considered in the optimisation.

Appendix A

Mixed Integer Linear Programming (MILP) modeling

This appendix explains the general form of linear programming problems and provides an example of a network design problem.

A.1 General Form of a Linear Programming Problem

The general form of a linear programming problem contains a set of decision variables which are to be the outcome values of the optimisation problem, an objective linear function that defines the solution and should be minimised or maximised, and a number of constraints in the form of linear functions that defines the requirements of the system to be optimised [284]. Each constraint is to be equal to, less than or equal, or large than or equal a parameter. The decision variables should typically have non-negative values. Nonlinear equations (e.g. multiplication, maximisation, and absolute values for

the variables) are not allowed in the constraints or the objective equations. Dealing with such equations requires applying several linearisation techniques. Mixed Integer Linear Programming (MILP) models contain integer in addition to binary variables that have values of either zero or one.

Let $x_1, ..., x_n$ be the variables, $c_j, \ \forall \ j = 1, 2, ..., n$ be the objective parameters (i.e. cost or revenue), and $a_{ij}, \ \forall \ i, j = 1, 2, ..., n$ and $b_1, ..., b_m$ be the parameters that define the available resources or requirements in the system to be optimised. Then, the general form of a linear programming problem can be written as [284]:

$$\min \ or \ \max \quad c_1x_1 + c_2x_2 + ... + c_nx_n \tag{A.1}$$

$$subject\,to$$

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \ (\leq, =, or \geq)b_1 \tag{A.2}$$

$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \ (\leq, =, or \geq)b_2 \tag{A.3}$$

$$...$$

$$a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \ (\leq, =, or \geq)b_m \tag{A.4}$$

$$x_j \geq 0 \ \forall j = 1, 2, ..., n \tag{A.5}$$

A.2 Formulation of a Network Design Problem

Figure A.1 shows a graph for a simple network that has three nodes, each is connected to the others forming a topology of a triangle shape [285]. The nodes can represent routers, switches or any equipment. Let Δ_{sd} be the demand value that can represent the traffic to be send from a source node s to a destination node d. In this network, a demand can be routed over two paths fully or partially. For example if Δ_{12} is 2 Gbps, this amount can be routed directly from node 1 to node 2 or from node 1 to node 2 through node 3. Partial amount of traffic can also be routed in these two available paths (e.g. 1 Gbps can be routed directly from node 1 to node 2 and 1 Gbps from node 1 to node 2 through node 3).

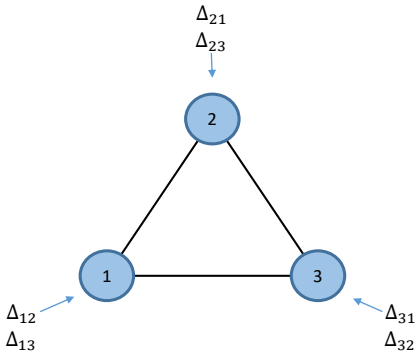


Figure A.1: Three nodes network example.

In a network design problem, determining the distribution of traffic occurs according to the optimisation problem and its requirements. There are two ways of formulating such problems which are the link-path formulation and the node-link formulation [285]. In this thesis and in what follows, the node-link formulation is used. In the node-link formulation, the variables for the traffic distribution are to be defined by the links and by the source and destination nodes. For example we can use χ_{uv}^{sd} to represent the traffic in link (u, v) that contributes to the traffic from node (vertex) s to node d. For example, χ_{12}^{13} is the portion of traffic to be routed from node 1 to node 3 that is assigned to link (1,2). The total traffic that can be routed in each link in the network is limited by its physical capacity ψ_{uv}, where u, and v are the nodes at the link edges (e.g. this network can have $\psi_{12}, \psi_{21}, \psi_{23}$, and $\psi_{32} = 10$ Gbps, and ψ_{13}, and $\psi_{31} = 20$ Gbps). Then for example, for link (1,2), the following should apply:

$$\chi_{12}^{12} + \chi_{12}^{21} + \chi_{12}^{13} + \chi_{12}^{31} + \chi_{12}^{23} + \chi_{12}^{32} \leq 10; \tag{A.6}$$

The node-link formulation requires following the flow conservation law which states that total traffic entering a node should be equal to the total outgoing traffic from that node. If the node is a source of a demand, then the total outgoing traffic for that demand minus the total incoming traffic should equal the demand. If the node is a destination of a demand, then the total incoming traffic for that demand minus the total outgoing traffic should equal the demand. For example for Δ_{13}, the following

equations should be applied:

$$\chi_{12}^{13} + \chi_{13}^{13} - \chi_{21}^{13} - \chi_{31}^{13} = \Delta_{13}\, at\, node\, 1\, (source\, node) \tag{A.7}$$

$$\chi_{21}^{13} + \chi_{23}^{13} - \chi_{12}^{13} - \chi_{32}^{13} = 0\, at\, node\, 2\, (transit\, node) \tag{A.8}$$

$$\chi_{31}^{13} + \chi_{32}^{13} - \chi_{13}^{13} - \chi_{23}^{13} = -\Delta_{13}\, at\, node\, 3\, (destination\, node) \tag{A.9}$$

An optimisation problem for designing a network should have a meaningful objective such as minimising the total routing cost. Such problems can be categorised as a multi-commodity network flow problem where multiple demands (i.e. commodities) need to be routed simultaneously while competing for link capacities [285]. A more generalised way to represent this problem, which is followed in this thesis, is as the following where all sets (e.g. nodes), parameters (e.g. capacities), and variables are first defined. Then, the objective function is presented followed by the constraints.

Sets and parameters:

$\mathbb{G}$	Set of all vertices (i.e. $\mathbb{G}$=[1,2,3])
$\mathbb{G}_u$	Set of neighbors of vertex; $u \in \mathbb{G}$ (i.e. $G_1 = [2,3], G_2 = [1,3], G_3 = [1,2]$)
Δ_{sd}	Traffic to be send from vertex s to vertex d; $s, d \in \mathbb{G}$
C_{uv}	Capacity of link (u,v); $u, v \in \mathbb{G}$

Variables:

χ_{uv}^{sd}	Traffic in link (u, v) that contributes to the traffic from vertex s to vertex d; $s, d \in \mathbb{G}, s \neq d, u \in \mathbb{G}, v \in \mathbb{G}_u$, $\chi_{uv}^{sd} \geq 0$

The objective is to minimise the routing cost:

$$\min \sum_{s,d,u,v \in \mathbb{G}, s \neq d} \left(\chi_{uv}^{sd}\right) \tag{A.10}$$

The objective is to be calculated under the following capacity and flow conservation constraints:

1. Flow conservation: The allocation of the links to the flows follows the flow con-

servation law:

$$\sum_{v \in \mathbb{G}_u} \chi_{uv}^{sd} - \sum_{v \in \mathbb{G}_u} \chi_{vu}^{sd} = \begin{cases} \Delta_{sd} & u = s \\ -\Delta_{sd} & u = d \\ 0 & otherwise, \end{cases} \tag{A.11}$$

$$\forall s, d \in \mathbb{G}, s \neq d, u \in \mathbb{G}.$$

2. Constraint to ensure that the total traffic in link (u, v) does not exceed its capacity:

$$\sum_{s,d \in \mathbb{G}, s \neq d} \chi_{uv}^{sd} \leq C_{uv}; \forall u \in \mathbb{G}, v \in \mathbb{G}_u. \tag{A.12}$$

A numerical example that can be solved through the previous model and can be solved manually is the following:

If there only one demand $\Delta_{12} = 8$ Gbps, minimising the routing cost will result that only the variable χ_{12}^{12} will equal 8 Gbps making the objective function equal to 8 Gbps. If the objective is to maximise the routing cost, then each of the variables χ_{13}^{12}, and χ_{32}^{12} will equal 8 Gbps making the objective function equal to 16 Gbps. The first solution indicates that the traffic is routed directly from node 1 to node 2, while the second solution indicates that the traffic is routed from node 1 to node 2 through node 3 which increases the routing cost.

A.3 Formulation of the Thesis MILP Models

Several extension and modifications are considered to the above example model to address the objectives of this thesis. For example for the models in Chapters 5, and 6, nodes were categorised into servers, AWGR nodes and switches. Additional binary variables were introduced to check for example if links or switches have traffic passing through them. Linearisation techniques such as those explained in [286] were used. Additional parameters to identify the power consumption of each equipment (located in a node) are used to aid in calculating the overall power consumption at a time.

Time dimension was added to all variable to aid in optimising the scheduling and to calculate the total energy consumption. The graph of each data centre (i.e. topology) was manually defined in the AMPL code. For the model in Chapter 7, the NSFNET topology is used and additional constraints for the routing in IP over WDM and usage of different power sources are introduced.

References

[1] S. H. Mohamed, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "A Survey of Big Data Machine Learning Applications Optimization in Cloud Data Centers and Networks," *to be submitted to the IEEE Surveys and Tutorials*, 2019.

[2] A. Hammadi, S. H. Mohamed, M. I. Musa, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Passive Optical Networks-based Data Centers for Fog and Cloud Computing," *to be submitted to IEEE Access*, 2019.

[3] S. H. Mohamed, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "On the energy efficiency of MapReduce shuffling operations in data centers," in *2017 19th International Conference on Transparent Optical Networks (ICTON)*, July 2017, pp. 1–5.

[4] S. H. Mohamed and T. E. H. El-Gorashi and J. M. H. Elmirghani, "Energy Efficiency of Server-Centric PON Data Center Architecture for Fog Computing," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, July 2018, pp. 1–4.

[5] S. H. Mohamed and T. E. H. El-Gorashi and J. M. H. Elmirghani, "Optimizing Co-flows Scheduling and Routing in Data Center Networks for Big Data Applications," *to be submitted to IEEE Transactions on Network and Service Management*, 2019.

[6] S. H. Mohamed, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Impact of Link Failures on the Performance of MapReduce in Data Center Networks," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, July 2018, pp. 1–4.

[7] M. B. Abdull Halim, S. Hamid Mohamed, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Fog-Assisted Caching Employing Solar Renewable Energy for Delivering Video on Demand Service," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, July 2019, pp. 1–5.

[8] S. H. Mohamed and Mohamad Bin Abdull Halim and T. E. H. El-Gorashi and J. M. H. Elmirghani, "Fog-assisted Caching Employing Solar Renewable Energy and Energy Storage Devices for Delivering Video on Demand Service," *to be submitted to IEEE Access*, 2019.

[9] A. A. Hammadi, "Future PON Data Centre Networks," Ph.D. dissertation, University of Leeds, School of Electronic and Electrical Engineering, Aug. 2016.

[10] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *Access, IEEE*, vol. 2, pp. 652–687, 2014.

[11] Y. Demchenko, C. de Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem," in *Collaboration Technologies and Systems (CTS), 2014 International Conference on*, May 2014, pp. 104–112.

[12] X. Yi, F. Liu, J. Liu, and H. Jin, "Building a network highway for big data: architecture and challenges," *Network, IEEE*, vol. 28, no. 4, pp. 5–13, July 2014.

[13] H. Fang, Z. Zhang, C. J. Wang, M. Daneshmand, C. Wang, and H. Wang, "A survey of big data research," *Network, IEEE*, vol. 29, no. 5, pp. 6–9, September 2015.

[14] W. Tan, M. Blake, I. Saleh, and S. Dustdar, "Social-Network-Sourced Big Data Analytics," *Internet Computing, IEEE*, vol. 17, no. 5, pp. 62–69, Sept 2013.

[15] C. Fang, J. Liu, and Z. Lei, "Parallelized user clicks recognition from massive HTTP data based on dependency graph model," *Communications, China*, vol. 11, no. 12, pp. 13–25, Dec 2014.

[16] H. Hu, Y. Wen, Y. Gao, T.-S. Chua, and X. Li, "Toward an SDN-enabled big data platform for social TV analytics," *Network, IEEE*, vol. 29, no. 5, pp. 43–49, September 2015.

[17] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[18] Y. Xu and S. Mao, "A survey of mobile cloud computing for rich media applications," *Wireless Communications, IEEE*, vol. 20, no. 3, pp. 46–53, June 2013.

[19] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021," White Paper, Cisco, March 2017.

[20] X. He, K. Wang, H. Huang, and B. Liu, "QoE-Driven Big Data Architecture for Smart City," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 88–93, Feb 2018.

[21] E. Marín-Tordera, X. Masip-Bruin, J. G. Almiñana, A. Jukan, G. Ren, J. Zhu, and J. Farre, "What is a Fog Node A Tutorial on Current Concepts towards a Common Definition," *CoRR*, vol. abs/1611.09193, 2016.

[22] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1628–1656, thirdquarter 2017.

[23] I. Stojmenovic, "Fog computing: A cloud to the ground support for smart things and machine-to-machine networks," in *2014 Australasian Telecommunication Networks and Applications Conference (ATNAC)*, Nov 2014, pp. 117–122.

[24] S. Wang, X. Wang, J. Huang, R. Bie, and X. Cheng, "Analyzing the potential of mobile opportunistic networks for big data applications," *IEEE Network*, vol. 29, no. 5, pp. 57–63, September 2015.

[25] Z. T. Al-Azez, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient IoT virtualization framework with passive optical access networks," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, July 2016, pp. 1–4.

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

[27] A. A. Alahmadi, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Distributed processing in vehicular cloud networks," in *2017 8th International Conference on the Network of the Future (NOF)*, Nov 2017, pp. 22–26.

[28] H. Q. Al-Shammari, A. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient service embedding in IoT networks," in *2018 27th Wireless and Optical Communication Conference (WOCC)*, April 2018, pp. 1–5.

[29] B. Yosuf, M. Musa, T. Elgorashi, A. Q. Lawey, and J. M. H. Elmirghani, "Energy Efficient Service Distribution in Internet of Things," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, July 2018, pp. 1–4.

[30] I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi, A. Q. Lawey, and J. M. H. Elmirghani, "Energy Efficiency of Fog Computing Health Monitoring Applications," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, July 2018, pp. 1–5.

[31] F. S. Behbehani, M. Musa, T. Elgorashi, and J. M. H. Elmirghani, "Energy Efficient Distributed Processing in Vehicular Cloud Architecture," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, July 2019, pp. 1–4.

[32] B. A. Yosuf, M. Musa, T. Elgorashi, and J. M. H. Elmirghani, "Impact of Distributed Processing on Power Consumption for IoT Based Surveillance Applications," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, July 2019, pp. 1–5.

[33] I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy Efficient and Resilient Infrastructure for Fog Computing Health Monitoring Applications," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, July 2019, pp. 1–5.

[34] R. Ma, A. A. Alahmadi, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy Efficient Software Matching in Vehicular Fog," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, July 2019, pp. 1–4.

[35] Z. T. Al-Azez, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy Efficient IoT Virtualization Framework with Peer to Peer Networking and Processing," *IEEE Access*, pp. 1–1, 2019.

[36] M. S. H. Graduate, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Patient-Centric Cellular Networks Optimization using Big Data Analytics," *IEEE Access*, pp. 1–1, 2019.

[37] K. Dolui and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," in *2017 Global Internet of Things Summit (GIoTS)*, June 2017, pp. 1–6.

[38] J. Andreu-Perez, C. Poon, R. Merrifield, S. Wong, and G.-Z. Yang, "Big Data for Health," *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 4, pp. 1193–1208, July 2015.

[39] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *Communications Surveys Tutorials, IEEE*, vol. 17, no. 4, pp. 2347–2376, Fourthquarter 2015.

[40] S. Mazumdar and S. Dhar, "Hadoop as Big Data Operating System – The Emerging Approach for Managing Challenges of Enterprise Big Data Platform," in *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*, March 2015, pp. 499–505.

[41] X. Xu, Q. Sheng, L.-J. Zhang, Y. Fan, and S. Dustdar, "From Big Data to Big Service," *Computer*, vol. 48, no. 7, pp. 80–83, July 2015.

[42] J. M. H. Elmirghani, T. Klein, K. Hinton, L. Nonde, A. Q. Lawey, T. E. H. El-Gorashi, M. O. I. Musa, and X. Dong, "GreenTouch GreenMeter core network energy-efficiency improvement measures and optimization," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A250–A269, Feb 2018.

[43] R. Fourer, D. Gay, and B. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*, ser. Scientific Press series. Thomson/Brooks/Cole, 2003. [Online]. Available: https://books.google.co.uk/books?id=Ij8ZAQAAIAAJ

[44] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google File System," *SIGOPS Oper. Syst. Rev.*, vol. 37, no. 5, pp. 29–43, Oct. 2003.

[45] S. Babu, "Towards Automatic Optimization of MapReduce Programs," in *Proceedings of the 1st ACM Symposium on Cloud Computing*, ser. SoCC '10. New York, NY, USA: ACM, 2010, pp. 137–142.

[46] D. Cheng, J. Rao, Y. Guo, C. Jiang, and X. Zhou, "Improving Performance of Heterogeneous MapReduce Clusters with Adaptive Task Tuning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 3, pp. 774–786, March 2017.

[47] T. White, *Hadoop: The Definitive Guide*, 1st ed. O'Reilly Media, Inc., 2009.

[48] G. Rumi, C. Colella, and D. Ardagna, "Optimization Techniques within the Hadoop Eco-system: A Survey," in *Symbolic and Numeric Algorithms for Scientific*

Computing (SYNASC), 2014 16th International Symposium on, Sept 2014, pp. 437–444.

[49] B. T. Rao and L. S. S. Reddy, "Survey on Improved Scheduling in Hadoop MapReduce in Cloud Environments," *CoRR*, vol. abs/1207.0780, 2012.

[50] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, "Delay Scheduling: A Simple Technique for Achieving Locality and Fairness in Cluster Scheduling," in *Proceedings of the 5th European Conference on Computer Systems*, ser. EuroSys '10. New York, NY, USA: ACM, 2010, pp. 265–278.

[51] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, and E. Baldeschwieler, "Apache Hadoop YARN: Yet Another Resource Negotiator," in *Proceedings of the 4th Annual Symposium on Cloud Computing*, ser. SOCC '13. New York, NY, USA: ACM, 2013, pp. 5:1–5:16.

[52] I. Polato, D. Barbosa, A. Hindle, and F. Kon, "Hadoop branching: Architectural impacts on energy and performance," in *Green Computing Conference and Sustainable Computing Conference (IGSC), 2015 Sixth International*, Dec 2015, pp. 1–4.

[53] Y. Zhang, T. Cao, S. Li, X. Tian, L. Yuan, H. Jia, and A. V. Vasilakos, "Parallel Processing Systems for Big Data: A Survey," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2114–2136, Nov 2016.

[54] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster Computing with Working Sets," in *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 10–10.

[55] A. Ching, S. Edunov, M. Kabiljo, D. Logothetis, and S. Muthukrishnan, "One Trillion Edges: Graph Processing at Facebook-scale," *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 1804–1815, Aug. 2015.

[56] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham, N. Bhagat, S. Mittal, and D. Ryaboy, "Storm@twitter," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: ACM, 2014, pp. 147–156.

[57] M. Sadiku, S. Musa, and O. Momoh, "Cloud Computing: Opportunities and Challenges," *Potentials, IEEE*, vol. 33, no. 1, pp. 34–36, Jan 2014.

[58] B. Biocic, D. Tomic, and D. Ogrizovic, "Economics of the cloud computing," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 1438–1442.

[59] N. da Fonseca and R. Boutaba, *Cloud Architectures, Networks, Services, and Management*. Wiley-IEEE Press, 2015, p. 432.

[60] C. Ge, Z. Sun, and N. Wang, "A Survey of Power-Saving Techniques on Data Centers and Content Delivery Networks," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1334–1354, Third 2013.

[61] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "A Survey of Green Information-Centric Networking: Research Issues and Challenges," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1455–1472, thirdquarter 2015.

[62] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Hölzle, S. Stuart, and A. Vahdat, "B4: Experience with a Globally-deployed Software Defined Wan," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 3–14, Aug. 2013.

[63] X. Zhao, V. Vusirikala, B. Koley, V. Kamalov, and T. Hofmeister, "The prospect of inter-data-center optical networks," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 32–38, September 2013.

[64] M. H. Ghahramani, M. Zhou, and C. T. Hon, "Toward cloud computing QoS architecture: analysis of cloud systems and cloud services," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 1, pp. 6–18, Jan 2017.

[65] Latency is Everywhere and it Costs you Sales - How to Crush it. (Cited on 2017, Dec). [Online]. Available: http://highscalability.com/latency-everywhere-and-it-costs-you-sales-how-crush-it

[66] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical Guide to Controlled Experiments on the Web: Listen to Your Customers Not to the Hippo," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 959–967.

[67] S. S. Krishnan and R. K. Sitaraman, "Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-experimental Designs," in *Proceedings of the 2012 Internet Measurement Conference*, ser. IMC '12. New York, NY, USA: ACM, 2012, pp. 211–224.

[68] Y. Zhang, P. Chowdhury, M. Tornatore, and B. Mukherjee, "Energy Efficiency in Telecom Optical Networks," *Communications Surveys Tutorials, IEEE*, vol. 12, no. 4, pp. 441–458, Fourth 2010.

[69] R. Ramaswami, K. N. Sivarajan, and G. H. Sasaki, *Optical Networks: A Practical Prespective*, 3rd ed. Morgan Kaufmann, 2010.

[70] H. Yin, Y. Jiang, C. Lin, Y. Luo, and Y. Liu, "Big data: transforming the design philosophy of future internet," *Network, IEEE*, vol. 28, no. 4, pp. 14–19, July 2014.

[71] K.-I. Kitayama, A. Hiramatsu, M. Fukui, T. Tsuritani, N. Yamanaka, S. Okamoto, M. Jinno, and M. Koga, "Photonic Network Vision 2020 – Toward Smart Photonic Cloud," *Lightwave Technology, Journal of*, vol. 32, no. 16, pp. 2760–2770, Aug 2014.

[72] A. S. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein, and W. Kellerer, "Software Defined Optical Networks (SDONs): A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2738–2786, Fourthquarter 2016.

[73] Y. Yin, L. Liu, R. Proietti, and S. J. B. Yoo, "Software Defined Elastic Optical Networks for Cloud Computing," *IEEE Network*, vol. 31, no. 1, pp. 4–10, January 2017.

[74] A. Nag, M. Tornatore, and B. Mukherjee, "Optical Network Design With Mixed Line Rates and Multiple Modulation Formats," *Journal of Lightwave Technology*, vol. 28, no. 4, pp. 466–475, Feb 2010.

[75] Y. Ji, J. Zhang, Y. Zhao, H. Li, Q. Yang, C. Ge, Q. Xiong, D. Xue, J. Yu, and S. Qiu, "All Optical Switching Networks With Energy-Efficient Technologies From Components Level to Network Level," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 8, pp. 1600–1614, Aug 2014.

[76] G. Tzimpragos, C. Kachris, I. B. Djordjevic, M. Cvijetic, D. Soudris, and I. Tomkos, "A Survey on FEC Codes for 100 G and Beyond Optical Networks," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 209–221, Firstquarter 2016.

[77] D. M. Marom, P. D. Colbourne, A. Dâ€™errico, N. K. Fontaine, Y. Ikuma, R. Proietti, L. Zong, J. M. Rivas-Moscoso, and I. Tomkos, "Survey of photonic switching architectures and technologies in support of spatially and spectrally flexible optical networking [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 1, pp. 1–26, Jan 2017.

[78] X. Yu, M. Tornatore, M. Xia, J. Wang, J. Zhang, Y. Zhao, J. Zhang, and B. Mukherjee, "Migration from fixed grid to flexible grid in optical networks," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 34–43, Feb 2015.

[79] M. Jinno, H. Takara, B. Kozicki, Y. Tsukishima, Y. Sone, and S. Matsuoka, "Spectrum-efficient and scalable elastic optical path network: architecture, benefits, and enabling technologies," *IEEE Communications Magazine*, vol. 47, no. 11, pp. 66–73, November 2009.

[80] O. Gerstel, M. Jinno, A. Lord, and S. J. B. Yoo, "Elastic optical networking: a new dawn for the optical layer?" *IEEE Communications Magazine*, vol. 50, no. 2, pp. s12–s20, February 2012.

[81] B. C. Chatterjee, N. Sarma, and E. Oki, "Routing and Spectrum Allocation in Elastic Optical Networks: A Tutorial," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1776–1800, thirdquarter 2015.

[82] G. Zhang, M. D. Leenheer, A. Morea, and B. Mukherjee, "A Survey on OFDM-Based Elastic Core Optical Networking," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 65–87, First 2013.

[83] A. Klekamp, U. Gebhard, and F. Ilchmann, "Energy and Cost Efficiency of Adaptive and Mixed-Line-Rate IP Over DWDM Networks," *Journal of Lightwave Technology*, vol. 30, no. 2, pp. 215–221, Jan 2012.

[84] T. E. El-Gorashi, X. Dong, and J. M. Elmirghani, "Green optical orthogonal frequency-division multiplexing networks," *IET Optoelectronics*, vol. 8, pp. 137–148(11), June 2014.

[85] S. Das, Y. Yiakoumis, G. Parulkar, N. McKeown, P. Singh, D. Getachew, and P. D. Desai, "Application-aware aggregation and traffic engineering in a converged packet-circuit network," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, March 2011, pp. 1–3.

[86] H. Harai, H. Furukawa, K. Fujikawa, T. Miyazawa, and N. Wada, "Optical Packet and Circuit Integrated Networks and Software Defined Networking Extension," *Journal of Lightwave Technology*, vol. 32, no. 16, pp. 2751–2759, Aug 2014.

[87] L. Valcarenghi, D. P. Van, P. G. Raponi, P. Castoldi, D. R. Campelo, S. Wong, S. Yen, L. G. Kazovsky, and S. Yamashita, "Energy efficiency in passive optical networks: where, when, and how?" *IEEE Network*, vol. 26, no. 6, pp. 61–68, November 2012.

[88] S. B. Weinstein, Y. Luo, and T. Wang, *PON Architecture and Components*. IEEE, 2012.

[89] F. Idzikowski, L. Chiaraviglio, A. Cianfrani, J. L. VizcaÃno, M. Polverini, and Y. Ye, "A Survey on Energy-Aware Design and Operation of Core Networks," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1453–1499, Secondquarter 2016.

[90] W. V. Heddeghem, B. Lannoo, D. Colle, M. Pickavet, and P. Demeester, "A Quantitative Survey of the Power Saving Potential in IP-Over-WDM Backbone Networks," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 706–731, Firstquarter 2016.

[91] Data centres and data transmission networks. (Cited on 2020, April). [Online]. Available: https://www.iea.org/reports/tracking-buildings/data-centres-and-data-transmission-networks

[92] M. Gupta and S. Singh, "Greening of the Internet," in *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '03. New York, NY, USA: ACM, 2003, pp. 19–26.

[93] J. C. C. Restrepo, C. G. Gruber, and C. M. Machuca, "Energy Profile Aware Routing," in *2009 IEEE International Conference on Communications Workshops*, June 2009, pp. 1–5.

[94] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, "Reducing Network Energy Consumption via Sleeping and Rate-adaptation," in *Pro-*

ceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, ser. NSDI'08. Berkeley, CA, USA: USENIX Association, 2008, pp. 323–336.

[95] G. Shen and R. Tucker, "Energy-Minimized Design for IP Over WDM Networks," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 1, no. 1, pp. 176–186, June 2009.

[96] 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, no. 12, pp. 1931–1942, June 2012.

[97] S. Zhang, D. Shen, and C. K. Chan, "Energy-Efficient Traffic Grooming in WDM Networks With Scheduled Time Traffic," *Journal of Lightwave Technology*, vol. 29, no. 17, pp. 2577–2584, Sept 2011.

[98] 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 *2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies*, Sept 2015, pp. 161–164.

[99] Z. H. Nasralla and T. E. H. El-Gorashi and M. O. I. Musa and J. M. H. Elmirghani, "Routing post-disaster traffic floods in optical core networks," in *2016 International Conference on Optical Network Design and Modeling (ONDM)*, May 2016, pp. 1–5.

[100] Z. H. Nasralla, M. O. I. Musa, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Routing post-disaster traffic floods heuristics," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, July 2016, pp. 1–4.

[101] M. O. I. Musa, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Network coding for energy efficiency in bypass IP/WDM networks," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, July 2016, pp. 1–3.

[102] M. O. I. Musa and 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)*, July 2015, pp. 1–4.

[103] T. E. H. El-Gorashi, X. Dong, A. Lawey, and J. M. H. Elmirghani, "Core network physical topology design for energy efficiency and resilience," in *2013 15th International Conference on Transparent Optical Networks (ICTON)*, June 2013, pp. 1–7.

[104] M. Musa, T. Elgorashi, and J. Elmirghani, "Energy efficient survivable IP-over-WDM networks with network coding," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 3, pp. 207–217, March 2017.

[105] M. Musa and T. Elgorashi and J. Elmirghani, "Bounds for energy-efficient survivable IP over WDMnetworks with network coding," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 5, pp. 471–481, May 2018.

[106] Y. Li, L. Zhu, S. K. Bose, and G. Shen, "Energy-Saving in IP Over WDM Networks by Putting Protection Router Cards to Sleep," *Journal of Lightwave Technology*, vol. 36, no. 14, pp. 3003–3017, July 2018.

[107] J. M. H. Elmirghani, L. Nonde, A. Q. Lawey, T. E. H. El-Gorashi, M. O. I. Musa, X. Dong, K. Hinton, and T. Klein, "Energy efficiency measures for future core networks," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–3.

[108] M. O. I. Musa, T. El-Gorashi, and J. M. H. Elmirghani, "Bounds on GreenTouch GreenMeter Network Energy Efficiency," *Journal of Lightwave Technology*, pp. 1–1, 2018.

[109] X. Dong, T. El-Gorashi, and J. Elmirghani, "IP Over WDM Networks Employing Renewable Energy Sources," *Lightwave Technology, Journal of*, vol. 29, no. 1, pp. 3–14, Jan 2011.

[110] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Green IP over WDM Networks: Solar and Wind Renewable Sources and Data Centres," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, Dec 2011, pp. 1–6.

[111] G. Shen, Y. Lui, and S. K. Bose, ""Follow the Sun, Follow the Wind" Lightpath Virtual Topology Reconfiguration in IP Over WDM Network," *Journal of Lightwave Technology*, vol. 32, no. 11, pp. 2094–2105, June 2014.

[112] M. Gattulli, M. Tornatore, R. Fiandra, and A. Pattavina, "Low-Emissions Routing for Cloud Computing in IP-over-WDM Networks with Data Centers," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 1, pp. 28–38, January 2014.

[113] 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)*, June 2015, pp. 128–134.

[114] S. K. Dey and A. Adhya, "Delay-aware green service migration schemes for data center traffic," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 8, no. 12, pp. 962–975, December 2016.

[115] L. Nonde, T. E. H. Elgorashi, and J. M. H. Elmirgahni, "Virtual Network Embedding Employing Renewable Energy Sources," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[116] L. Nonde, T. El-Gorashi, and J. Elmirghani, "Energy Efficient Virtual Network Embedding for Cloud Networks," *Lightwave Technology, Journal of*, vol. 33, no. 9, pp. 1828–1849, May 2015.

[117] X. Dong, T. El-Gorashi, and J. Elmirghani, "Green IP Over WDM Networks With Data Centers," *Lightwave Technology, Journal of*, vol. 29, no. 12, pp. 1861–1880, June 2011.

[118] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez, "Greening the Internet with Nano Data Centers," in *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '09. New York, NY, USA: ACM, 2009, pp. 37–48.

[119] C. Jayasundara, A. Nirmalathas, E. Wong, and C. Chan, "Improving Energy Efficiency of Video on Demand Services," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 3, no. 11, pp. 870–880, November 2011.

[120] 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*, May 2011, pp. 1–5.

[121] N. I. Osman and T. El-Gorashi and J. M. H. Elmirghani, "The impact of content popularity distribution on energy efficient caching," in *2013 15th International Conference on Transparent Optical Networks (ICTON)*, June 2013, pp. 1–6.

[122] 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, no. 13, pp. 2364–2381, July 2014.

[123] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "BitTorrent Content Distribution in Optical Networks," *Journal of Lightwave Technology*, vol. 32, no. 21, pp. 4209–4225, Nov 2014.

[124] A. Lawey, T. El-Gorashi, and J. Elmirghani, "Distributed Energy Efficient Clouds Over Core Networks," *Lightwave Technology, Journal of*, vol. 32, no. 7, pp. 1261–1281, April 2014.

[125] H. A. Alharbi, T. E. H. El-Gorashi, A. Q. Lawey, and J. M. H. Elmirghani, "Energy efficient virtual machines placement in IP over WDM networks," in *2017 19th International Conference on Transparent Optical Networks (ICTON)*, July 2017, pp. 1–4.

[126] U. Wajid, c. cappiello, P. Plebani, B. Pernici, N. Mehandjiev, M. Vitali, M. Gienger, K. Kavoussanakis, D. Margery, D. Perez, and P. Sampaio, "On Achieving Energy Efficiency and Reducing CO2 Footprint in Cloud Computing," *Cloud Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[127] A. Al-Salim, A. Lawey, T. El-Gorashi, and J. Elmirghani, "Energy Efficient Tapered Data Networks for Big Data processing in IP/WDM networks," in *Transparent Optical Networks (ICTON), 2015 17th International Conference on*, July 2015, pp. 1–5.

[128] A. M. Al-Salim, H. M. M. Ali, A. Q. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "Greening big data networks: Volume impact," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, July 2016, pp. 1–6.

[129] A. M. Al-Salim, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy Efficient Big Data Networks: Impact of Volume and Variety," *IEEE Transactions on Network and Service Management*, vol. PP, no. 99, pp. 1–1, 2017.

[130] A. M. Al-Salim, T. E. El-Gorashi, A. Q. Lawey, and J. M. Elmirghani, "Greening big data networks: velocity impact," *IET Optoelectronics*, November 2017.

[131] L. A. Barroso and U. Hoelzle, *The Datacenter As a Computer: An Introduction to the Design of Warehouse-Scale Machines*, 1st ed. Morgan and Claypool Publishers, 2009.

[132] M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, Aug. 2008.

[133] T. Wang, Z. Su, Y. Xia, and M. Hamdi, "Rethinking the Data Center Networking: Architecture, Network Protocols, and Resource Sharing," *Access, IEEE*, vol. 2, pp. 1481–1496, 2014.

[134] Pall Beck, Peter Clemens, Santiago Freitas, Jeff Gatz, Michele Girola, Jason Gmitter, Holger Mueller, Ray O'Hanlon, Veerendra Para, Joe Robinson, Andy Sholomon, Jason Walker, and Jon Tate, *IBM and Cisco: Together for a World Class Data Center*. IBM Redbooks, 2013.

[135] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, and A. Vahdat, "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," in *Sigcomm '15*, 2015.

[136] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 63–74, Aug. 2009.

[137] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers," in *SIGCOMM08*. Association for Computing Machinery, Inc., August 2008.

[138] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," *Communications Surveys Tutorials, IEEE*, vol. 14, no. 4, pp. 1021–1036, Fourth 2012.

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

[140] L. Schares, D. M. Kuchta, and A. F. Benner, "Optics in Future Data Center Networks," in *2010 18th IEEE Symposium on High Performance Interconnects*, Aug 2010, pp. 104–108.

[141] H. Ballani, P. Costa, I. Haller, K. Jozwik, K. Shi, B. Thomsen, and H. Williams, "Bridging the Last Mile for Optical Switching in Data Centers," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[142] G. Papen, "Optical components for datacenters," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–53.

[143] L. Chen, E. Hall, L. Theogarajan, and J. Bowers, "Photonic Switching for Data Center Applications," *IEEE Photonics Journal*, vol. 3, no. 5, pp. 834–844, Oct 2011.

[144] H. Rodrigues, R. Strong, A. Akyurek, and T. Rosing, "Dynamic optical switching for latency sensitive applications," in *Architectures for Networking and Communications Systems (ANCS), 2015 ACM/IEEE Symposium on*, May 2015, pp. 75–86.

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

[146] H. J. S. Dorren, S. Di Lucente, J. Luo, O. Raz, and N. Calabretta, "Scaling photonic packet switches to a large number of ports [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, no. 9, pp. A82–A89, Sep. 2012.

[147] F. Yan, W. Miao, O. Raz, and N. Calabretta, "Opsquare: A flat DCN architecture based on flow-controlled optical packet switches," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 4, pp. 291–303, April 2017.

[148] X. Yu, H. Gu, K. Wang, and S. Ma, "Petascale: A Scalable Buffer-Less All-Optical Network for Cloud Computing Data Center," *IEEE Access*, vol. 7, pp. 42 596–42 608, 2019.

[149] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan, "c-Through: part-time optics in data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, Aug. 2010.

[150] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, Aug. 2010.

[151] A. Singla, A. Singh, and Y. Chen, "OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility," in *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. San Jose, CA: USENIX, 2012, pp. 239–252.

[152] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: A Topology Malleable Data Center Network," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, ser. Hotnets-IX. New York, NY, USA: ACM, 2010, pp. 8:1–8:6.

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

[154] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The Nature of Data Center Traffic: Measurements & Analysis," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '09. New York, NY, USA: ACM, 2009, pp. 202–208.

[155] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding Data Center Traffic Characteristics," *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 1, pp. 92–99, Jan. 2010.

[156] 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*, ser. IMC '10. New York, NY, USA: ACM, 2010, pp. 267–280.

[157] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the Social Network's (Datacenter) Network," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 5, pp. 123–137, Aug. 2015.

[158] Q. Zhang, V. Liu, H. Zeng, and A. Krishnamurthy, "High-resolution Measurement of Data Center Microbursts," in *Proceedings of the 2017 Internet Measurement Conference*, ser. IMC '17. New York, NY, USA: ACM, 2017, pp. 78–85.

[159] D. A. Popescu and A. W. Moore, "A First Look at Data Center Network Condition Through The Eyes of PTPmesh," in *2018 Network Traffic Measurement and Analysis Conference (TMA)*, June 2018, pp. 1–8.

[160] D. Xie, N. Ding, Y. C. Hu, and R. Kompella, "The Only Constant is Change: Incorporating Time-varying Network Reservations in Data Centers," in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '12. New York, NY, USA: ACM, 2012, pp. 199–210.

[161] M. Noormohammadpour and C. S. Raghavendra, "Datacenter Traffic Control: Understanding Techniques and Tradeoffs," *IEEE Communications Surveys Tutorials*, vol. 20, no. 2, pp. 1492–1525, Secondquarter 2018.

[162] K. Chen, C. Hu, X. Zhang, K. Zheng, Y. Chen, and A. V. Vasilakos, "Survey on routing in data centers: insights and future directions," *IEEE Network*, vol. 25, no. 4, pp. 6–10, July 2011.

[163] R. Rojas-Cessa, Y. Kaymak, and Z. Dong, "Schemes for Fast Transmission of Flows in Data Center Networks," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1391–1422, thirdquarter 2015.

[164] J. Qadir, A. Ali, K. A. Yau, A. Sathiaseelan, and J. Crowcroft, "Exploiting the Power of Multiplicity: A Holistic Survey of Network-Layer Multipath," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2176–2213, Fourthquarter 2015.

[165] J. Zhang, F. R. Yu, S. Wang, T. Huang, Z. Liu, and Y. Liu, "Load Balancing in Data Center Networks: A Survey," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2018.

[166] Y. Zhang and N. Ansari, "On Architecture Design, Congestion Notification, TCP Incast and Power Consumption in Data Centers," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 39–64, First 2013.

[167] M. Alizadeh, T. Edsall, S. Dharmapurikar, R. Vaidyanathan, K. Chu, A. Fingerhut, V. T. Lam, F. Matus, R. Pan, N. Yadav, and G. Varghese, "CONGA: Distributed Congestion-aware Load Balancing for Datacenters," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 503–514, Aug. 2014.

[168] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: A Scalable and Flexible Data Center Network," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 51–62, Aug. 2009.

[169] A. Singla, C. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking Data Centers Randomly," *CoRR*, vol. abs/1110.1687, 2011.

[170] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat, "Integrating Microsecond Circuit Switching into the Data Center," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 447–458, Aug. 2013.

[171] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, "Data center TCP (DCTCP)," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. –, Aug. 2010.

[172] C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley, "Improving Datacenter Performance and Robustness with Multipath TCP," in *Proceedings of the ACM SIGCOMM 2011 Conference*, ser. SIGCOMM '11. New York, NY, USA: ACM, 2011, pp. 266–277.

[173] B. Vamanan, J. Hasan, and T. Vijaykumar, "Deadline-aware Datacenter TCP (D2TCP)," in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '12. New York, NY, USA: ACM, 2012, pp. 115–126.

[174] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron, "Better never than late: Meeting deadlines in datacenter networks," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 50–61, Aug. 2011.

[175] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, and S. Shenker, "pFabric: Minimal Near-optimal Datacenter Transport," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 435–446, Aug. 2013.

[176] C.-Y. Hong, M. Caesar, and P. B. Godfrey, "Finishing Flows Quickly with Preemptive Scheduling," in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '12. New York, NY, USA: ACM, 2012, pp. 127–138.

[177] D. Zats, T. Das, P. Mohan, D. Borthakur, and R. Katz, "DeTail: Reducing the Flow Completion Time Tail in Datacenter Networks," in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '12. New York, NY, USA: ACM, 2012, pp. 139–150.

[178] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica, "Managing Data Transfers in Computer Clusters with Orchestra," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 98–109, Aug. 2011.

[179] M. Chowdhury, Y. Zhong, and I. Stoica, "Efficient Coflow Scheduling with Varys," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 443–454, Aug. 2014.

[180] F. R. Dogar, T. Karagiannis, H. Ballani, and A. Rowstron, "Decentralized Task-aware Scheduling for Data Center Networks," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 431–442, Aug. 2014.

[181] S. Raghul, T. Subashri, and K. R. Vimal, "Literature survey on traffic-based server load balancing using SDN and open flow," in *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*, March 2017, pp. 1–6.

[182] T. Hafeez, N. Ahmed, B. Ahmed, and A. W. Malik, "Detection and Mitigation of Congestion in SDN Enabled Data Center Networks: A Survey," *IEEE Access*, vol. 6, pp. 1730–1740, 2018.

[183] A. Mendiola, J. Astorga, E. Jacob, and M. Higuero, "A Survey on the Contributions of Software-Defined Networking to Traffic Engineering," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 918–953, Secondquarter 2017.

[184] C. Guo, G. Lu, H. J. Wang, S. Yang, C. Kong, P. Sun, W. Wu, and Y. Zhang, "SecondNet: A Data Center Network Virtualization Architecture with Bandwidth Guarantees," in *Proceedings of the 6th International COnference*, ser. Co-NEXT '10. New York, NY, USA: ACM, 2010, pp. 15:1–15:12.

[185] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic Flow Scheduling for Data Center Networks," in *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 19–19.

[186] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "ElasticTree: Saving Energy in Data Center Networks," in *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 17–17.

[187] M. Dayarathna, Y. Wen, and R. Fan, "Data Center Energy Consumption Modeling: A Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 732–794, Firstquarter 2016.

[188] D. Ã‡avdar and F. Alagoz, "A survey of research on greening data centers," in *2012 IEEE Global Communications Conference (GLOBECOM)*, Dec 2012, pp. 3237–3242.

[189] A. C. Riekstin, B. B. Rodrigues, K. K. Nguyen, T. C. M. de Brito Carvalho, C. Meirosu, B. Stiller, and M. Cheriet, "A Survey on Metrics and Measurement Tools for Sustainable Distributed Cloud Networks," *IEEE Communications Surveys Tutorials*, vol. 20, no. 2, pp. 1244–1270, Secondquarter 2018.

[190] 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, no. 1, pp. 68–73, Dec. 2008.

[191] W. Zhang, Y. Wen, Y. W. Wong, K. C. Toh, and C. H. Chen, "Towards Joint Optimization Over ICT and Cooling Systems in Data Centre: A Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1596–1616, thirdquarter 2016.

[192] K. Bilal, S. U. R. Malik, S. U. Khan, and A. Y. Zomaya, "Trends and challenges in cloud datacenters," *IEEE Cloud Computing*, vol. 1, no. 1, pp. 10–20, May 2014.

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

[194] K. Christensen, P. Reviriego, B. Nordman, M. Bennett, M. Mostowfi, and J. Maestro, "IEEE 802.3az: the road to energy efficient ethernet," *Communications Magazine, IEEE*, vol. 48, no. 11, pp. 50–56, November 2010.

[195] R. F. e Silva and P. M. Carpenter, "Energy Efficient Ethernet on MapReduce Clusters: Packet Coalescing To Improve 10GbE Links," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2731–2742, Oct 2017.

[196] Y. Luo, F. Effenberger, and M. Sui, "Cloud computing provisioning over Passive Optical Networks," in *2012 1st IEEE International Conference on Communications in China (ICCC)*, Aug 2012, pp. 255–259.

[197] M. Taheri and N. Ansari, "A feasible solution to provide cloud computing over optical networks," *IEEE Network*, vol. 27, no. 6, pp. 31–35, November 2013.

[198] A. H. Helmy and A. Nayak, "Integrating Fog With Long-Reach PONs From a Dynamic Bandwidth Allocation Perspective," *Journal of Lightwave Technology*, vol. 36, no. 22, pp. 5276–5284, Nov 2018.

[199] S. H. S. Newaz, W. S. binti Haji Suhaili, G. M. Lee, M. R. Uddin, A. F. Y. Mohammed, and J. K. Choi, "Towards realizing the importance of placing fog computing facilities at the central office of a PON," in *2017 19th International Conference on Advanced Communication Technology (ICACT)*, Feb 2017, pp. 152–157.

[200] H. Liu, F. Lu, A. Forencich, R. Kapoor, M. Tewari, G. M. Voelker, G. Papen, A. C. Snoeren, and G. Porter, "Circuit Switching Under the Radar with REACToR," in *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*. Seattle, WA: USENIX Association, 2014, pp. 1–15.

[201] C. Kachris and I. Tomkos, "Power consumption evaluation of hybrid WDM PON networks for data centers," in *2011 16th European Conference on Networks and Optical Communications*, July 2011, pp. 118–121.

[202] P. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos, "Design and Evaluation of a Flexible-Bandwidth OFDM-Based Intra-Data Center Interconnect," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 19, no. 2, pp. 3 700 310–3 700 310, March 2013.

[203] K. Wang, L. Zhao, H. Gu, X. Yu, G. Wu, and J. Cai, "ADON: a scalable AWG-based topology for datacenter optical network," *Optical and Quantum Electronics*, vol. 47, no. 8, pp. 2541–2554, Aug 2015.

[204] P. N. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos, "Design and Evaluation of a Flexible-Bandwidth OFDM-Based Intra-Data Center Interconnect," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 2, pp. 3 700 310–3 700 310, March 2013.

[205] W. Ni, C. Huang, Y. L. Liu, W. Li, K. Leong, and J. Wu, "POXN: A New Passive Optical Cross-Connection Network for Low-Cost Power-Efficient Datacenters," *Journal of Lightwave Technology*, vol. 32, no. 8, pp. 1482–1500, April 2014.

[206] Y. Cheng, M. Fiorani, R. Lin, L. Wosinska, and J. Chen, "POTORI: a passive optical top-of-rack interconnect architecture for data centers," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 5, pp. 401–411, May 2017.

[207] J. Elmirghani, T. EL-GORASHI, and A. HAMMADI, "Passive optical-based data center networks," 2016, wO Patent App. PCT/GB2015/053,604. [Online]. Available: http://google.com/patents/WO2016083812A1?cl=und

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

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

[210] 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)*, June 2016, pp. 178–182.

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

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

[213] 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)*, July 2016, pp. 1–4.

[214] A. E. A. Eltraify, M. O. I. Musa, A. Al-Quzweeni, and J. M. H. Elmirghani, "Experimental Evaluation of Passive Optical Network Based Data Centre Architecture," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, July 2018, pp. 1–4.

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

[216] M. I. Olmedo, L. Suhr, K. Prince, R. Rodes, C. Mikkelsen, E. Hviid, C. Neumeyr, G. Vollrath, E. Goobar, P. Öhlén, and I. T. Monroy, "Gigabit Access Passive Optical Network Using Wavelength Division Multiplexing - GigaWaM," *Journal of Lightwave Technology*, vol. 32, no. 22, pp. 4285–4293, Nov 2014.

[217] ZXA10 C300:The Industry's First Future-proof Optical Access platform. (Cited on 2018, Apr). [Online]. Available: http://www.zte.com.cn/global/products/access/xpon/PON-OLT/424194

[218] J. Beals, N. Bamiedakis, A. Wonfor, R. V. Penty, I. H. White, J. V. DeGroot, K. Hueston, T. V. Clapp, and M. Glick, "A terabit capacity passive polymer optical backplane based on a novel meshed waveguide architecture," *Applied Physics A*, vol. 95, no. 4, pp. 983–988, Jun 2009.

[219] J. Han, M. Ishii, and H. Makino, "A Hadoop performance model for multi-rack clusters," in *Computer Science and Information Technology (CSIT), 2013 5th International Conference on*, March 2013, pp. 265–274.

[220] G. Wang, A. R. Butt, P. Pandey, and K. Gupta, "A simulation approach to evaluating design decisions in MapReduce setups," in *Modeling, Analysis Simulation of Computer and Telecommunication Systems, 2009. MASCOTS '09. IEEE International Symposium on*, Sept 2009, pp. 1–11.

[221] Z. Kouba, O. Tomanek, and L. Kencl, "Evaluation of Datacenter Network Topology Influence on Hadoop MapReduce Performance," in *2016 5th IEEE International Conference on Cloud Networking (Cloudnet)*, Oct 2016, pp. 95–100.

[222] Y. Shang, D. Li, J. Zhu, and M. Xu, "On the Network Power Effectiveness of Data Center Architectures," *Computers, IEEE Transactions on*, vol. 64, no. 11, pp. 3237–3248, Nov 2015.

[223] M. Alizadeh and T. Edsall, "On the Data Path Performance of Leaf-Spine Datacenter Fabrics," in *High-Performance Interconnects (HOTI), 2013 IEEE 21st Annual Symposium on*, Aug 2013, pp. 71–74.

[224] P. Costa, A. Donnelly, A. Rowstron, and G. O'Shea, "Camdoop: Exploiting In-network Aggregation for Big Data Applications," in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 3–3.

[225] L. Rupprecht, "Exploiting In-network Processing for Big Data Management," in *Proceedings of the 2013 SIGMOD/PODS Ph.D. Symposium*, ser. SIGMOD'13 PhD Symposium. New York, NY, USA: ACM, 2013, pp. 1–6.

[226] X. Meng, V. Pappas, and L. Zhang, "Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement," in *Proceedings of the 29th Conference on Information Communications*, ser. INFOCOM'10. Piscataway, NJ, USA: IEEE Press, 2010, pp. 1154–1162.

[227] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "Towards predictable datacenter networks," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 242–253, Aug. 2011.

[228] W. C. Moody, J. Anderson, K.-C. Wange, and A. Apon, "Reconfigurable Network Testbed for Evaluation of Datacenter Topologies," in *Proceedings of the Sixth International Workshop on Data Intensive Distributed Computing*, ser. DIDC '14. New York, NY, USA: ACM, 2014, pp. 11–20.

[229] G. Wang, T. E. Ng, and A. Shaikh, "Programming Your Network at Run-time for Big Data Applications," in *Proceedings of the First Workshop on Hot Topics in Software Defined Networks*, ser. HotSDN '12. New York, NY, USA: ACM, 2012, pp. 103–108.

[230] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. E. Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, and A. Vahdat, "Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks," in *Proceedings of the 2Nd ACM Symposium on Cloud Computing*, ser. SOCC '11. New York, NY, USA: ACM, 2011, pp. 30:1–30:8.

[231] L. Y. Ho, J. J. Wu, and P. Liu, "Optimal Algorithms for Cross-Rack Communication Optimization in MapReduce Framework," in *2011 IEEE 4th International Conference on Cloud Computing*, July 2011, pp. 420–427.

[232] Y. Le, F. Wang, J. Liu, and F. Ergün, "On Datacenter-Network-Aware Load Balancing in MapReduce," in *Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on*, June 2015, pp. 485–492.

[233] H. Ke, P. Li, S. Guo, and M. Guo, "On Traffic-Aware Partition and Aggregation in MapReduce for Big Data Applications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 3, pp. 818–828, March 2016.

[234] D. Guo, J. Xie, X. Zhou, X. Zhu, W. Wei, and X. Luo, "Exploiting Efficient and Scalable Shuffle Transfers in Future Data Center Networks," *Parallel and*

Distributed Systems, IEEE Transactions on, vol. 26, no. 4, pp. 997–1009, April 2015.

[235] Y. Shang, D. Li, and M. Xu, "Greening data center networks with flow preemption and energy-aware routing," in *Local Metropolitan Area Networks (LANMAN), 2013 19th IEEE Workshop on*, April 2013, pp. 1–6.

[236] L. Wang, F. Zhang, and Z. Liu, "Improving the Network Energy Efficiency in MapReduce Systems," in *Computer Communications and Networks (ICCCN), 2013 22nd International Conference on*, July 2013, pp. 1–7.

[237] L. W. Cheng and S. Y. Wang, "Application-Aware SDN Routing for Big Data Networking," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2014, pp. 1–6.

[238] S. Narayan, S. Bailey, and A. Daga, "Hadoop Acceleration in an OpenFlow-Based Cluster," in *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, Nov 2012, pp. 535–538.

[239] S. Hu, K. Chen, H. Wu, W. Bai, C. Lan, H. Wang, H. Zhao, and C. Guo, "Explicit path control in commodity data centers: Design and applications," *Networking, IEEE/ACM Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[240] S. Luo, H. Yu, Y. Zhao, S. Wang, S. Yu, and L. Li, "Towards Practical and Near-optimal Coflow Scheduling for Data Center Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[241] Y. Zhao, K. Chen, W. Bai, M. Yu, C. Tian, Y. Geng, Y. Zhang, D. Li, and S. Wang, "Rapier: Integrating routing and scheduling for coflow-aware data center networks," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 424–432.

[242] M. V. Neves, C. A. F. D. Rose, K. Katrinis, and H. Franke, "Pythia: Faster Big Data in Motion through Predictive Software-Defined Network Optimization at

Runtime," in *2014 IEEE 28th International Parallel and Distributed Processing Symposium*, May 2014, pp. 82–90.

[243] P. Qin, B. Dai, B. Huang, and G. Xu, "Bandwidth-Aware Scheduling With SDN in Hadoop: A New Trend for Big Data," *Systems Journal, IEEE*, vol. PP, no. 99, pp. 1–8, 2015.

[244] K. Kontodimas, K. Christodoulopoulos, E. Zahavi, and E. Varvarigos, "Resource allocation in slotted optical data center networks," in *2018 International Conference on Optical Network Design and Modeling (ONDM)*, May 2018, pp. 248–253.

[245] L. Wang, X. Wang, M. Tornatore, K. J. Kim, S. M. Kim, D. Kim, K. Han, and B. Mukherjee, "Scheduling with machine-learning-based flow detection for packet-switched optical data center networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 4, pp. 365–375, April 2018.

[246] D. Li, Y. Yu, W. He, K. Zheng, and B. He, "Willow: Saving Data Center Network Energy for Network-Limited Flows," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 26, no. 9, pp. 2610–2620, Sept 2015.

[247] Z. Niu, B. He, and F. Liu, "JouleMR: Towards Cost-Effective and Green-Aware Data Processing Frameworks," *IEEE Transactions on Big Data*, vol. 4, no. 2, pp. 258–272, June 2018.

[248] Cisco Nexus 3548-X, 3524-X, 3548-XL, and 3524-XL Switches Data Sheet. (Cited on 2019, Sept). [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-3548-switch/data_sheet_c78-707001.pdf

[249] Cisco 500 Series Stackable Managed Switches Data Sheet. (Cited on 2019, Sept). [Online]. Available: http://www.cisco.com/c/en/us/products/collateral/switches/small-business-500-series-stackable-managed-switches/c78-695646_data_sheet.html

[250] Cisco 10GBASE SFP+ Modules Data Sheet. (Cited on 2019, Sept). [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/transceiver-modules/data_sheet_c78-455693.htm

[251] R. Sohan, A. Rice, W. M. Andrew, and K. Mansley, "Characterizing 10 Gbps network interface energy consumption," in *IEEE Local Computer Network Conference*, Oct 2010, pp. 268–271.

[252] P. X. Gao, A. Narayan, S. Karandikar, J. Carreira, S. Han, R. Agarwal, S. Ratnasamy, and S. Shenker, "Network Requirements for Resource Disaggregation," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. Savannah, GA: USENIX Association, 2016, pp. 249–264.

[253] R. Xie and X. Jia, "Data Transfer Scheduling for Maximizing Throughput of Big-Data Computing in Cloud Systems," *Cloud Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[254] MRV Pluggable Transceivers. (Cited on 2019, Sept). [Online]. Available: http://s1.dtsheet.com/store/data/001288813.pdf?key=b765b7adbd723bf63be5c121991bab76&r=1

[255] F. Guo, O. Ormond, L. Fialho de queiroz, M. Collier, and X. Wang, "Power consumption analysis of a netfpga based router," *Journal of China Universities of Posts and Telecommunications*, vol. 19, no. SUPPL. 1, pp. 94–99, 6 2012.

[256] A. Rasmussen, G. Porter, M. Conley, H. V. Madhyastha, R. N. Mysore, A. Pucher, and A. Vahdat, "TritonSort: A Balanced and Energy-Efficient Large-Scale Sorting System," *ACM Trans. Comput. Syst.*, vol. 31, no. 1, pp. 3:1–3:28, Feb. 2013.

[257] Sort Benchmark Home Page. (Cited on 2019, Sept). [Online]. Available: http://sortbenchmark.org/

[258] C. Colman-Meixner, C. Develder, M. Tornatore, and B. Mukherjee, "A Survey on Resiliency Techniques in Cloud Computing Infrastructures and Applications," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2244–2281, thirdquarter 2016.

[259] M. F. Habib, M. Tornatore, M. D. Leenheer, F. Dikbiyik, and B. Mukherjee, "Design of disaster-resilient optical datacenter networks," *Journal of Lightwave Technology*, vol. 30, no. 16, pp. 2563–2573, Aug 2012.

[260] X. Liu and Q. Liu, "An Optimized Speculative Execution Strategy Based on Local Data Prediction in a Heterogeneous Hadoop Environment," in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, vol. 2, July 2017, pp. 128–131.

[261] H. Wang, H. Chen, Z. Du, and F. Hu, "BeTL: MapReduce Checkpoint Tactics Beneath the Task Level," *Services Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[262] B. Ghit and D. Epema, "Reducing Job Slowdown Variability for Data-Intensive Workloads," in *Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), 2015 IEEE 23rd International Symposium on*, Oct 2015, pp. 61–70.

[263] S. M. Nabavinejad, M. Goudarzi, and S. Mozaffari, "The Memory Challenge in Reduce Phase of MapReduce Applications," *IEEE Transactions on Big Data*, vol. PP, no. 99, pp. 1–1, 2016.

[264] X. Shi, M. Chen, L. He, X. Xie, L. Lu, H. Jin, Y. Chen, and S. Wu, "Mammoth: Gearing Hadoop Towards Memory-Intensive MapReduce Applications," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 26, no. 8, pp. 2300–2315, Aug 2015.

[265] Y. Kwon, M. Balazinska, B. Howe, and J. Rolia, "SkewTune: Mitigating Skew in Mapreduce Applications," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '12. New York, NY, USA: ACM, 2012, pp. 25–36.

[266] R. Potharaju and N. Jain, "When the network crumbles: An empirical study of cloud network failures and their impact on services," in *Proceedings of the 4th Annual Symposium on Cloud Computing*, ser. SOCC '13. New York, NY, USA: ACM, 2013, pp. 15:1–15:17.

[267] Potharaju, Rahul and Jain, Navendu, "An empirical analysis of intra- and inter-datacenter network failures for geo-distributed services," *SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 1, pp. 335–336, Jun. 2013.

[268] P. Gill, N. Jain, and N. Nagappan, "Understanding network failures in data centers: Measurement, analysis, and implications," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 350–361, Aug. 2011.

[269] P. Bodík, I. Menache, M. Chowdhury, P. Mani, D. A. Maltz, and I. Stoica, "Surviving failures in bandwidth-constrained datacenters," in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '12. New York, NY, USA: ACM, 2012, pp. 431–442.

[270] K. C. Webb, A. C. Snoeren, and K. Yocum, "Topology Switching for Data Center Networks," in *Proceedings of the 11th USENIX Conference on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services*, ser. HotICE'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 14–14.

[271] Z. Wu, Y. Zhang, V. Singh, G. Jiang, and H. Wang, "Automating Cloud Network Optimization and Evolution," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 12, pp. 2620–2631, December 2013.

[272] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A Comprehensive Survey on Fog Computing: State-of-the-Art and Research Challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 1, pp. 416–464, Firstquarter 2018.

[273] F. Jalali, K. Hinton, R. Ayre, T. Alpcan, and R. S. Tucker, "Fog computing may help to save energy in cloud computing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1728–1739, May 2016.

[274] S. Igder, S. Bhattacharya, and J. M. H. Elmirghani, "Energy Efficient Fog Servers for Internet of Things Information Piece Delivery (IoTIPD) in a Smart City Vehicular Environment," in *2016 10th International Conference on Next Generation Mobile Applications, Security and Technologies (NGMAST)*, Aug 2016, pp. 99–104.

[275] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, Aug 2014.

[276] B. Yang, Z. Zhang, K. Zhang, and W. Hu, "Integration of micro data center with optical line terminal in passive optical network," in *2016 21st OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS)*, July 2016, pp. 1–3.

[277] C. Gu, H. Huang, and X. Jia, "Green scheduling for cloud data centers using ESDs to store renewable energy," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–7.

[278] Y. Li, A. Orgerie, and J. Menaud, "Balancing the Use of Batteries and Opportunistic Scheduling Policies for Maximizing Renewable Energy Consumption in a Cloud Data Center," in *2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, March 2017, pp. 408–415.

[279] C. Gu, K. Hu, Z. Li, Q. Yuan, H. Huang, and X. Jia, "Lowering Down the Cost for Green Cloud Data Centers by Using ESDs and Energy Trading," in *2016 IEEE Trustcom/BigDataSE/ISPA*, Aug 2016, pp. 1508–1515.

[280] Cisco Catalyst 9500 series switches data sheet. (Cited on 2018, Apr). [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/switches/catalyst-9500-series-switches/datasheet-c78-738978.html

[281] H. Chen, T. N. Cong, W. Yang, C. Tan, Y. Li, and Y. Ding, "Progress in electrical energy storage system: A critical review," *Progress in Natural Science*, vol. 19, no. 3, pp. 291 – 312, 2009.

[282] NREL: MIDC/NREL Solar Radiation Research Laboratory (BMS). (Cited on 2018, Feb). [Online]. Available: https://midcdmz.nrel.gov/apps/go2url.pl?site=BMS

[283] K. Yoshikawa, H. Kawasaki, W. Yoshida, T. Irie, K. Konishi, K. Nakano, T. Uto, D. Adachi, M. Kanematsu, H. Uzu, and K. Yamamoto, "Silicon heterojunction solar cell with interdigitated back contacts for a photoconversion efficiency over 26%," *Nature Energy*, vol. 2, p. 17032, 2017.

[284] J. C. Smith and Z. C. Taskin, "A Tutorial Guide to Mixed Integer Programming Models and Solution Techniques," 2007.

[285] M. Pióro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2004.

[286] Linearization of the product of two variables. (Cited on 2020, April). [Online]. Available: https://www.leandro-coelho.com/linearization-product-variables/