Optical Technologies and Control Methods for Scalable Data Centre Networks

Hui Yuan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Electronic & Electrical Engineering
University College London

February 6, 2020

I, Hui Yuan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Attributing to the increasing adoption of cloud services, video services and associated machine learning applications, the traffic demand inside data centers is increasing exponentially, which necessitates an innovated networking infrastructure with high scalability and cost-efficiency. As a promising candidate to provide high capacity, low latency, cost-effective and scalable interconnections, optical technologies have been introduced to data center networks (DCNs) for approximately a decade. To further improve the DCN performance to meet the increasing traffic demand by using photonic technologies, two current trends are a) increasing the bandwidth density of the transmission links and b) maximizing IT and network resources utilization through disaggregated topologies and architectures. Therefore, this PhD thesis focuses on introducing and applying advanced and efficient technologies in these two fields to DCNs to improve their performance. On the one hand, at the link level, since the traditional single-mode fiber (SMF) solutions based on wavelength division multiplexing (WDM) over C+L band may fall short in satisfying the capacity, front panel density, power consumption, and cost requirements of high-performance DCNs, a space division multiplexing (SDM) based DCN using homogeneous multi-core fibers (MCFs) is proposed. With the exploited bi-directional model and proposed spectrum allocation algorithms, the proposed DCN shows great benefits over the SMF solution in terms of network capacity and spatial efficiency. In the meanwhile, it is found that the inter-core crosstalk (IC-XT) between the adjacent cores inside the MCF is dynamic rather than static, therefore, the behaviour of the IC-XT is experimentally investigated under different transmission conditions. On the other hand, an optically disaggregated DCN is developed and to ensure the performance of it, different architectures, topologies, resource routing and allocation algorithms are proposed and compared. Compared to the traditional server-based DCN, the resource utilization, scalability and the cost-efficiency are significantly improved.

Impact Statement

The rapid growth in Internet protocol traffic is pushing data centers into the "Zettabyte Era". Particularly, it is forecasted that at least 85% of the total traffic by 2021 will stay inside the data center to accommodate the requirements from the applications, such as cloud services, video services and associated machine learning applications, highlighting the importance of providing high capacity to the intra data center network. Moreover, the electricity demand from data centers at that time will be more than 1000 billion kWh and data centers are predicted to consume 3-13% of the global electricity usage for communication in 2030, necessitating the application of power efficient techniques in DCNs. Specifically, the servers in the data centers consume around 40%, storage consumes 37% and the network devices consume around 23% of the total IT power. In the past decade, optical technologies have been introduced to DCNs and significantly relieved the pressure on these two aspects. However, the current DCN configurations or architecture may be insufficient to support the future traffic requirements. Therefore, new interconnection schemes or architectures must be proposed to address the aforementioned challenges.

In this thesis, simulations and experiments were carried out to explore the effectiveness of using advanced fibers and architectures for future data center scaling. In particular, the investigation on the solution of using MCF-based SDM technique indicated that the network capacity and link spatial efficiency of DCN could be considerably increased while reducing the cabling complexity, front panel density and power consumption. Moreover, the main constraint of MCF, IC-XT, was comprehensively generalized and correlated to its root causes, such as signalling source, baud rate, modulation format, wavelength, data sequence pattern and temperature. Associated with the derived IC-XT equations, the processes of fiber selection, design and deployment in SDM-based DCN will be easier. Furthermore, the study on the IC-XT step distribution will contribute to the source signal classification, which may significantly increase the DCN security.

The introduction of the concept of disaggregation can also bring revolutionary benefits to DCNs. Compared to the traditional data center architecture, the

disaggregated data center architecture can fundamentally improve the flexibility, scalability of DCN while maximizing the resource utilization. Although extra latency and bandwidth are required, positive network gain can be achieved, attributing to the considerable processors and memory savings. Furthermore, the application of either the proposed parallel topologies or MCF-SMF hybrid architecture enables the disaggregated DCN to have lower latency and bandwidth requirements or power/space consumption, which makes the architecture perfectly satisfy the requirements for data center scaling.

To sum up, all the solutions that have been presented in this thesis have great potential to enable the data center to scale sustainably and efficiently in the future.

Acknowledgements

First and foremost, I wish to thank and appreciate my principal supervisor Dr. Georgios Zervas for his faith in my capability and giving me the opportunity to purse a PhD degree in the Optical Network Group (ONG) at UCL. Without his continuous support, professional guidance and advice throughout my PhD research, I could not have done it. Also, I would like to express my profound gratitude to my second supervisor Prof. Polina Bayvel, who founded the ONG and led it to be one of the best research groups in the optical transmission field. I am absolutely proud to be a member in this group.

Besides my supervisors, I would like to thank the entity of the ONG, lecturers, post-docs, fellow Phd students and visitors. Especially to Dr. Zhixin Liu and Dr. Lidia Galdino for supporting me on the multi-core fiber experiments and paper writing, to Prof. Hiroyuki Tsuda for guiding me for the experiment on the T-band routing system, to Dr. Arsalan Saljoghei, Dr. Yuta Wakayama, Mr. Eric Sillekens and Mr. Thomas Gerard for teaching me either the theory behind various experiments or the skills on coding and performing experiments, to Mr. Vaibhawa Mishra and Mr. Joshua Benjamin for teaching me how to use the FPGA and corresponding software. I would aslo like to thank Dr. Paris Andreades, Dr. Tianhua Xu, Miss Xun Mu and Miss Wenting Yi for discussions not limited to work and making my time in London enjoyable and fun.

Special thanks to my friends Miss Ke Zhu, Dr. Fanchao Meng, Dr. Adaranijo Peters and Mr. Mingfeng Zhu for their support, company, care and encouragement during my PhD life.

Finally, but most importantly, I would like to express my deepest gratitude to my parents and my sister for their constant supporting and deep love throughout the whole PhD program, without them this thesis would never have turned into reality.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Challenges for Data Centers

In the recent decades, due to the increasing application of cloud services, video services and associated machine learning applications, the internet protocol (IP) traffic in data centers is growing rapidly [1]. This tremendous growth in IP traffic is pushing data centers into the "Zettabyte Era" [2]. According to the white papers from Cisco and the data shown in Fig. 1.1, the annual global IP traffic for data center, including both internal ("east-west") and external ("north-south") traffic, will reach 20.6 zettabytes (increasing by 1.7 zettabytes per month) by the end of 2021, rising at 6.8 zettabytes per year (increasing by 568 exabytes per month) in 2016 [3, 4]. In other words, the global data center IP traffic will grow by 3-fold over five years. Particularly, it is predicted that at least 85% of the total traffic by 2021 will stay inside the data center to accommodate the applications' requirements, highlighting the importance of providing high capacity to the intra data center network (DCN).

Besides, power consumption is another essential prerequisite and key performance indicator (KPI) for the design and deployment of a DCN. Power consumption from the data center infrastructures [5] and the Internet network typologies [6] dominate the overall power consumption of cloud computing. Especially, the servers, storage and network devices consume around 40%, 37% and 23% of the total information technology (IT) power, respectively [7]. A report from Greenpeace [8] shows that the global electricity demand from data centers was nearly 330 billion kWh in 2007, whilst in 2020 this value will be more than 1000 billion kWh, which means the demand will be tripled. Moreover, it is forecasted that data centers will consume 3 to 13% of the global electricity usage for communication in 2030 [9]. Therefore, introducing power-efficient techniques to DCNs is urgent. It should be noted that along with the rapid growth of IT devices in data centers, the power required by the heating-ventilation and air-conditioning equipment, which is utilized to maintain the temperature of the

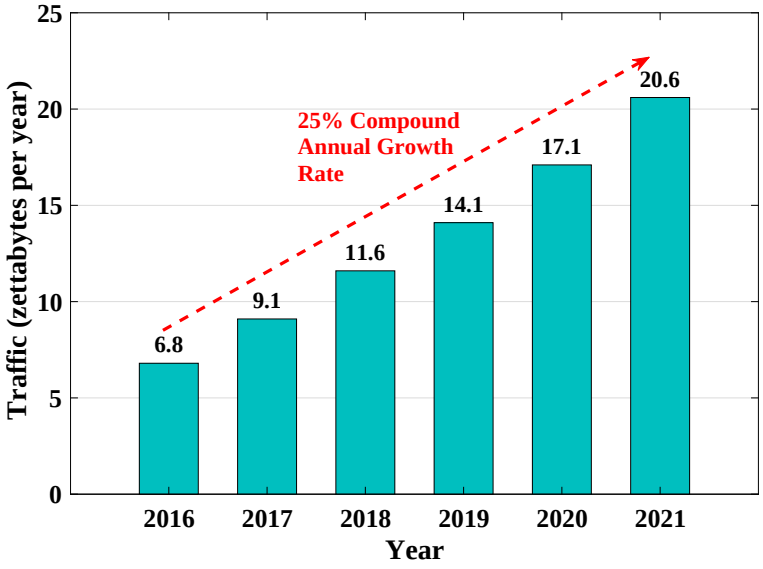


Figure 1.1: Global data center IP traffic growth

data center site, also increases. Thus, power consumption reduction from the network devices also has a considerable influence on the overall power consumption of the data center site.

Apart from the challenges in network capacity and power consumption, the wastage of IT (compute, memory, storage) resources in DCN should also be addressed. The conventional data centers are built based on the server-centric architecture, where each server tray acts as a functionally unit bundling compute, memory and acceleration units [10]. Also, it defines the physical boundaries of resource allocation, once and for all at design time. However, since the current DCNs need to support highly diverse workloads ranging up to 4-orders of magnitude on memory over compute demand [10], the mismatch between fixed proportionalities and diverse workloads can lead up to 60% IT resource wastage [11]. These unused resources can be translated into considerable cost and power wastage [12], indicating the importance of maximizing resource utilization in DCNs.

Additionally, low latency between end-to-end connections is necessary for supporting future generations of communication use cases and scenarios, which increasingly involve cloud facilities [13]. For some extreme examples, such as remote surgery and autonomous driving, even ultra-low latency is required. Other KPIs for data center deployment and scaling that need to be taken into consideration include front panel density, cabling complexity, spatial efficiency and cost. In order to address the aforementioned challenges coming with the increased communication in data centers, new interconnection schemes and

architectures should be proposed and developed.

1.2 Promising Solutions for Data Center Scaling

As a promising solution to provide high capacity, low latency, cost-effective and scalable interconnections to DCNs, optical technologies have been widely explored in research for approximately a decade [7, 14, 15]. Especially, optical links (fibers) and transceivers have been commercially adopted in DCNs. To further improve the DCN performance to meet the increasing traffic demand by using photonic technologies, two current trends [1] are a) increasing capacity or bandwidth density of transmission links/fibers and b) maximizing IT resources, i.e. central processing unit (CPU), memory and network, utilization through disaggregated topologies and architectures [16–18].

Based on the Shannon–Hartley theorem [19, 20], the total achievable summed up capacity (C) in the cable or fiber, including M parallel spatial channels (i.e. fibers, cores or modes) and two polarization, can be expressed as [21, 22]:

$$C = 2MBlog_2(1+SNR) \tag{1.1}$$

In which, B stands for the bandwidth of the channel and SNR denotes the received signal to noise ratio. The equation clearly indicates that, two approaches that can increase the fiber capacity are a) increasing the bandwidth of each optical channel and b) adding more spatial channels to the fiber or cable, which are regarded as space division multiplexing (SDM) techniques.

On the one hand, the most direct way to increase the channel bandwidth is the ultra-wide bandwidth transmission. That is to say, apart from the commonly used C+L band wavelengths, S-band [23–25], O-band [26–28] or even T-band [29–31] wavelengths can also be utilized for optical transmission systems. On the other hand, to date, the main approaches to realize SDM are using a) fiber ribbons/bundles, b) multi-core fibers (MCFs), which will be described in detail in Section 2.4.2, c) few/multi-mode fibers [32–34] and d) few-mode multi-core fibers [35–37]. Moreover, the bandwidth density can also be increased by increasing the spectral efficiency of transmission signals, such us using high-order modulated signals. Recently, a highly 16384-quadrature amplitude modulation (QAM) transmission system at 10 GBaud over 25-km has been experimentally demonstrated [38]. The aforementioned technologies have been successfully demonstrated in either long-haul or short-reach systems, showing their potential to scale the network capacity. However, the feasibility of applying them to the data centers for network scaling still needs to be further explored and evaluated.

As aforementioned the the previous section, the IT resource utilization in current DCN is low [10]. To improve resource efficiency, a new approach, resource disaggregation [17], has been proposed and explored aiming at increasing the data center resource utilization while reducing the cost and power consumption. A review on the history and current trend of this technique will be presented in Section 2.2.2.

1.3 Key Contributions

The key contributions from the work presented in this thesis are listed below:

- A comprehensive study on the relationship between the static and dynamic inter-core crosstalk (IC-XT) in MCF and the transmission parameters, such as signalling source, modulation format, baud rate, temperature, pseudo random binary sequence (PRBS) pattern, operating wavelength, and the number of excited cores are experimentally demonstrated in Section 3.4. The IC-XT behavior is characterized, which can be a reference for future MCF deployment in data centers. The results of this study were published in [39] and [40].

- For the first time, the impacts of the power meter averaging time and the observation time window on the accuracy of the observed IC-XT are evaluated and presented in Section 3.5. While in Section 3.6, a novel model for the distribution of IC-XT step is proposed and validated by the experimental results. These works can significantly benefit the studies on MCFs in the labs and in practical applications, which were published in [40] and [41].

- A SDM-based DCN using bi-directional MCFs is proposed and developed in Chapter 4, associating with several new resource allocation algorithms. Simulation results indicate that compared to the DCN based on wavelength division multiplexing (WDM) using uni-directional single-mode fiber (SMF), this new network architecture can considerably increase the network capacity, link spatial efficiency and front panel density. This comprehensive study was published in [42] and [43].

- New mathematical models to estimate wavelength-dependent IC-XT in bi-directional step-index/trench-assisted MCFs with euqal/unequal core pitches are derived in Section 4.2 and applied to the resource allocation process in a SDM-based DCN. Some of the results were published in [43] and [44].

- Three disaggregated data center (DDC) architectures and four resource routing and allocation algorithms for DDC are developed and compared with a Matlab simulator in Sections 5.2, 5.3 and 5.4. These new architectures can considerably save IT resources (i.e. CPU, memory) while presenting low blocking probability at the cost of extra bandwidth and latency. This comprehensive work was published in [45] and [46].

- A cost model is developed in Section 5.5 to compare the cost efficiency of the DDC and traditional data center (TDC). For most of the investigated situations, the DDC outperforms the TDC, which was presented in [46].

- Two ultra-low latency and scalable parallel topologies are proposed for DDCs in Section 6.2. Simulation results were published in [47], which indicate that they have the potential to save network resource while reducing the latency, cost and power consumption against the traditional 3-tier fat tree topology.

- A novel MCF-SMF hybrid architecture for DDC network is proposed based on the better parallel topology in Section 6.5. Compared to the pure SMF-based architecture, it significantly reduces power consumption at the switching layer and increases the space efficiency. This work was published in [48].

1.4 Structure of the Rest of this Thesis

The rest of this thesis is structured as follows:

Chapter 2 provides an overview of the existing network architectures, topologies, technologies as well as network resource routing and allocation algorithms that can contribute to future DCN deployment and scaling. In addition, some of them will be utilized in the following chapters.

Chapter 3 experimentally investigates the behavior and accuracy of IC-XT in trench-assisted MCFs (TA-MCFs). First of all, it characterizes the IC-XT behavior with the consideration of a variety of parameters. Secondly, this chapter evaluates the impact of experimental setup or observation time on the IC-XT accuracy. At last, a novel model for the IC-XT step distribution is presented and elaborated.

Chapter 4 focuses on a SDM-WDM hybrid DCN using bi-directional MCFs. It first derives multiple IC-XT equations for different types of bi-directional MCFs. Subsequently, several core and spectrum allocation algorithms are proposed for the new transmission model. A simulator is then developed in Matlab to compare the performance of DCN with different multiplexing techniques, fibers and topologies,

including blocking probability, network utilization, network capacity and link spatial efficiency.

Chapter 5 aims at improving the IT resource utilization of DCN by applying the concept of disaggregation. Several DDC architectures and resource (i.e. CPU, memory, storage and bandwidth) allocation algorithms are presented and compared in the simulation platform. This chapter also evaluates the network gain of the disaggregated data center versus the conventional one, by using a developed cost model.

Chapter 6 proposes two parallel topologies for DDC networks. The performance of them, in terms of blocking probability, resource utilization, round-trip latency, switch cost and power consumption, is compared to a widely used non-parallel topology. In addition, this chapter also designs and builds a MCF-SMF hybrid DDC architecture. The port utilization and in turn, the power/space efficiency of the new architecture is estimated.

Chapter 7 generally concludes and highlights the works that have been descried in the previous chapters. Suggestions for future work are also presented.

1.5 Publications Related to this Thesis

The following publications originating from the work described in this thesis are as follows, sorted by year.

Journal Papers

1. A. Saljoghei, **H. Yuan**, V. Mishra, M. Enrico, N. Parsons, C. Kochis, P. D. Dobbelaere, D. Theodoropoulos, D. Pnevmatikatos, D. Syrivelis, A. Reale, T. Hayashi, T. Nakanishi, G. Zervas, "MCF-SMF hybrid low-latency circuit switched optical network for disaggregated data centers," *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4017–4029, Aug 2019.

2. **H. Yuan**, M. Furdek, A. Muhammad, A. Saljoghei, L. Wosinska and G. Zervas, "Space-division multiplexing in data center networks: on multi-core fiber solutions and crosstalk-suppressed resource allocation," *Journal of Optical Communications and Networking (JOCN)*, vol. 10, no. 4, pp. 272-288, April 2018.

3. G. Zervas, **H. Yuan**, A. Saljoghei, Q. Chen, and V. Mishra, "Optically Disaggregated Data Centers With Minimal Remote Memory Latency: Technologies, Architectures, and Resource Allocation [Invited]," *Journal of Optical Communications and Networking (JOCN)*, vol. 10, no. 2, pp. A270-A285, Feb. 2018.

Conference Papers

4. **H. Yuan**, A. Saljoghei, T. Hayashi, T. Nakanishi, E. Sillekens, L. Galdino, P. Bayvel, Z. Liu, and G. Zervas, "Experimental Investigation of Static and Dynamic Crosstalk in Trench-Assisted Multi-Core Fiber," *Optical Fiber Communications Conference and Exhibition (OFC)*, San Diego, CA, USA, 2019, pp. 1-3.

5. **H. Yuan**, A. Saljoghei, A. Peters and G. Zervas,"Comparison of SDM-WDM based Data Center Networks with equal/unequal core pitch Multi-Core Fibers," *Optical Fiber Communications Conference and Exposition (OFC)*, San Diego, CA, 2018, pp. 1-3.

6. **H. Yuan**, A. Saljoghei, A. Peters and G. Zervas, "Disaggregated Optical Data Center in a Box Network using Parallel OCS Topologies," *Optical Fiber Communications Conference and Exposition (OFC)*, San Diego, CA, 2018, pp. 1-3.

7. G. Zervas, F. Jiang, Q. Chen, V. Mishra, **H. Yuan**, K. Katrinis, D. Syrivelis, A. Reale, D. Pnevmatikatos, M. Enrico, and N. Parsons, "Disaggregated Compute, Memory and Network Systems: A New Era for Optical Data Centre Architectures," *Optical Fiber Communications Conference and Exhibition (OFC)*, Los Angeles, CA, 2017, pp. 1-3.

8. Y. Liu, **H. Yuan**, A. Peters and G. Zervas, "Comparison of SDM and WDM on Direct and Indirect Optical Data Center Networks," *European Conference on Optical Communication (ECOC)*, Dusseldorf, Germany, 2016, pp. 1-3.

Book Chapter

9. N. Alachiotis, A. Andronikakis, O. Papadakis, D. Theodoropoulos, D. Pnevmatikatos, D. Syrivelis, A. Reale, K. Katrinis, G. Zervas, V. Mishra, **H. Yuan**, I. Syrigos, I. Igoumenos, T. Korakis, M. Torrents, and F. Zyulkyarov, *dReDBox: A Disaggregated Architectural Perspective for Data Centers*. Cham: Springer International Publishing, 2019, pp. 35–56.

Chapter 2

Literature Review

2.1 Introduction

To satisfy the rapidly increasing traffic requirements in DCNs, innovative network architectures, topologies, technologies and network resource routing and allocation policies should be proposed and developed. These new methods and technologies should aim to provide optimum levels of scalability, flexibility, cost and power efficiency, as well as network performance. In this chapter, an overview of existing techniques in the aforementioned aspects is provided. At the beginning, the current data center and DCN architectures are reviewed in Section 2.2. Next, the concept of disaggregation is introduced followed by the discussion on disaggregated data center (DDC) architecture. Moreover, existing research on memory disaggregation and DDCs are reviewed, including theoretical simulations and experimental demonstrations. At the end of Section 2.2, challenges that DDCs have to address in the future are presented. From then on, several typical data center topologies are described and compared in Section 2.3, including indirect and direct topologies. Subsequently, Section 2.4 first reviews the concepts and history of two widely used multiplexing techniques in optical networks, wavelength division multiplexing (WDM) and SDM, among which the latter one is one of the key points of this thesis. On the perspective of SDM technologies, the study focus on the SDM realization using MCFs. The constraint for this technique, i.e. inter-core crosstalk (IC-XT), and the related theoretical models for it are presented and analyzed. At last, Section 2.5 looks into several routing and allocation algorithms that will be used for network simulation in the following chapters. The algorithms include not only the general network routing and allocation algorithms considering the IT or bandwidth resources, but also the specific algorithms for spectrum and core resources allocations in WDM and/or SDM-based networks.

2.2 Data Center Architectures

Data center is regarded as a facility that consists of computer systems and corresponding components, including communication networks and storage systems [49]. IT companies rely on large scale data centers to store and process information created by tens of thousands users. International Data Corporation predictions present that 90% of IT industry growth in the next few years will be promoted by cloud services, big data and network technologies [50]. In order to efficiently manage the enormous traffic as well as realize high speed and highly quality services, the deployments for new data centers as well as the corresponding networks are imperative.

2.2.1 Traditional data center architecture

Traditional data center (TDC), based historically on a server-centric approach, is built using hundreds of servers with the ability for each to function independently. Each server can be regarded as a functionally integrated unit integrating a fixed number of CPU processors and directly attached memory resources within the boundary of a mainboard tray [10]. Figure. 2.1 depicts the aggregated architecture of the current DCN using this model. As shown, this hierarchical inter-networking model consists of three layers: core layer, aggregate layer and access layer. At the bottom layer, each rack hosts usually up to 48 servers (e.g. database, application or web servers) in the form of blades, while the servers are connected to the top of

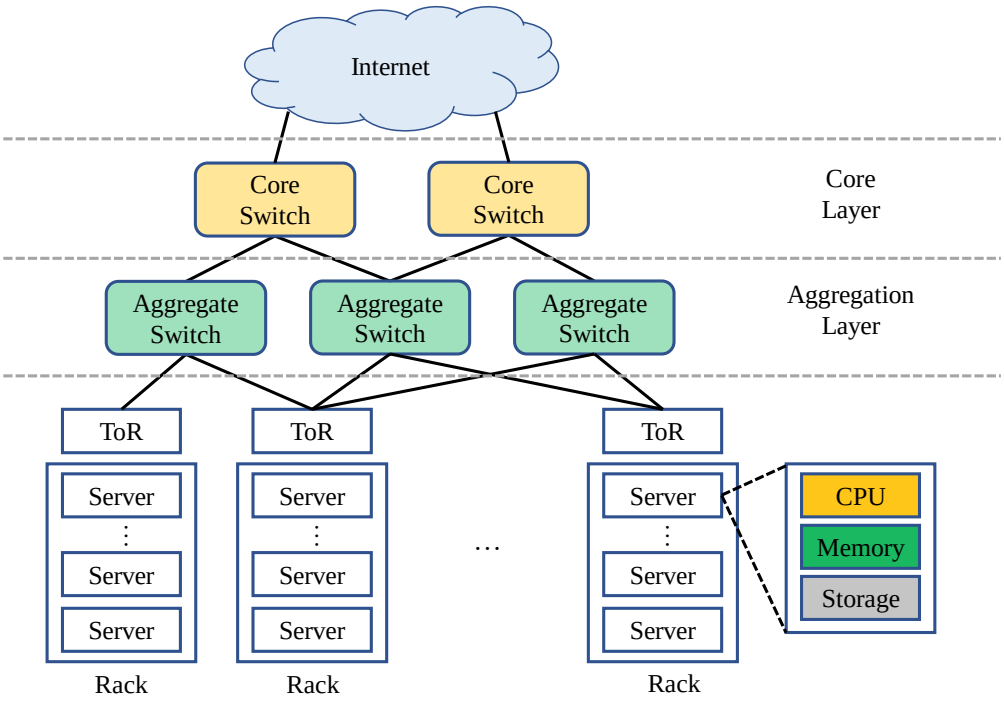


Figure 2.1: Architecture of traditional data center network

rack (ToR) switches through electrical cables or optical fibers. At the upper two layers, the ToRs are interconnected through the aggregate switches using optical links in a tree topology, where the aggregate switches are also interconnected through the core switches with a fat-tree topology. Therefore, a 3-tier network architecture is established, which enables the interconnection between all the servers and then offers access to data center clients through the Internet (i.e. wired and wireless networks).

With the increase of the traffic demand, conventional data centers using this architecture have to support highly diverse workloads ranging up to 4-orders of magnitude on memory/CPU demand to CPU usage [10]. However, it is found that this architecture presents shortcomings in flexibility and adaptability triggering that IT resources (i.e. CPU, memory and storage) cannot be fully utilized in each server. For instance, in compute dense applications, CPUs are largely required but only a minimal memory size is requested, which causes a considerable waste of memory; and in memory dense applications, an over-provision of CPUs usually occurs [51]. Authors in [52] illustrated that the memory occupancy is lower than 50% in a multi-purpose cluster. Additionally, workload trace from Google backed clusters indicates that the average memory utilization of the five clusters is only 42% [11]. These mismatches between the fixed proportionalities and a diverse set of workloads lead to substantially underutilized resources (often at only 40%), which account for 85% of the total data center cost [53]. Obviously, these issues make TDC unable to fulfill the demands of mutable application requirements. To address the issues, the concept of disaggregation is introduced.

2.2.2 Disaggregated data center architecture

Generally speaking, disaggregation indicates dividing an integrated object or system into several components based on their functionalities [54]. With the inherent benefit in modularity, it enables flexible customization and optimization of DCNs, also, it provides modularity for compute and memory resources. Unlike the traditional server-centric model that uses the mainboard as a functional unit, disaggregated architecture is based on the resource-centric model, which enables the creation of functional block unit. Specifically, in DDC, as shown in Fig. 2.2, IT resources are designed as modular resource pools, including CPU, memory, storage and accelerator pools [55]. An optical or opto-electronic network fabric is built to interconnect all the resource pools, which allows dynamic resource allocation according to different application demands [45]. This new architecture can realize increased IT resource utilization to allow high demands and variable workloads. Moreover, it can provide an immense level of scalability and flexibility

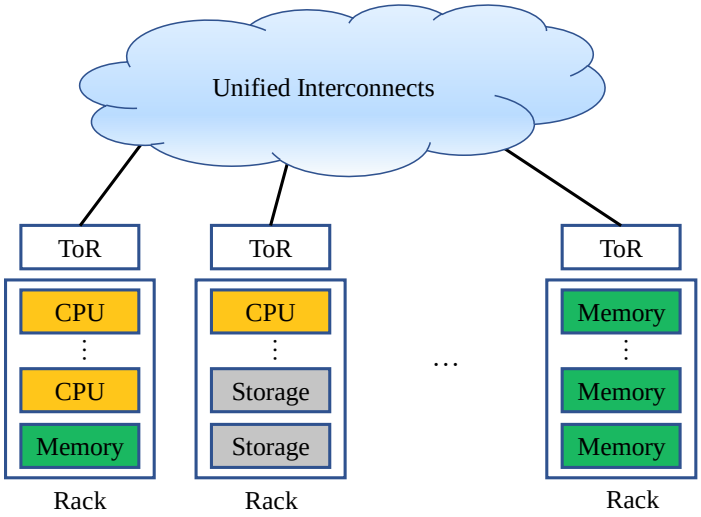


Figure 2.2: Architecture of disaggregated data center network

since each specific resource can be dynamically added or removed depending on various application requirements, which triggers the expansion on the capacity and optimized utilization of resources through co-operative sharing.

To date, research on memory disaggregation at the rack level has showcased approximately a 10-fold increase in performance over that of the server-centric architecture, while achieving up to 87% performance-per-dollar improvement [56]. Besides, a joint-virtual machine (VM) provisioning approach has realized efficient resource provisioning in compute clouds, which has achieved 45% reduction on resource requirements [57]. Moreover, investigation on a mixed-integer linear programming and heuristic model has shown that compared with the traditional server-centric data centers, DDCs can provide better power savings [58]. Apart from the theoretical analysis and simulations, multiple experiments also have been conducted to evaluate the concept of disaggregation and disaggregated data centers [17, 59–62]. According to [17], which concentrates on the effects of bandwidth and latency requirements on the application performance in DDC, for a certain application, under 10% (minimal) degradation in performance can be achieved with 20-40 Gb/s bandwidth for remote memory access. Moreover, communication between multi-processor system on chip hardware with multiple cores and remote access memory over 10 Gb/s lanes has been demonstrated in reference [61]. Based on the result, nearly 5 Gb/s bandwidth towards remote memory can be achieved by a single compute processor core, with a 45% penalty compared to that of the local access. In addition, an experimental demonstration on the memory disaggregation over a low latency optical network shows that 68% of memory bandwidth can be sustained by disaggregated memory access

compared to the local memory [62].

However, every coin has two sides. Several fundamental challenges have arisen on such resource-centric architectures that should be addressed before its application. First of all, compared with the traditional direct attached architecture, latency overhead results from the resource disaggregation need to be minimized. Secondly, as extra network resources are required by the IT resources communications, the new system has to provide specific performance and services based on different communication types, such as compute-to-memory and memory-to-storage communications, on the same substrate for the maximum flexibility. Furthermore, at the perspective of cost and power consumption, IT and network resources should be orchestrated aiming at maximizing the workload performance and the resource utilization at the minimum latency and cost. The new systems also need to support a substantially higher bandwidth and/or bandwidth density while saving cost and reducing power consumption.

2.3 Existing Data Center Topologies

As aforementioned in Section 2.2, in order to enable the data transmission or communication between different end nodes inside a data center or between data centers in a smooth manner, efficient optical network fabrics are required to interconnect all the servers, ToRs, upper layer switches and other involved devices. Data center topology is defined as the arrangement of these end nodes, intermediate nodes and fibers [63], which may significantly determine the key properties of DCNs. In order to design an efficient topology for data centers, several requirements should be taken into consideration: a) high bandwidth links should be provided to support even huge data transmission, such as applications requiring any to any communication [64], b) the topology should be cost-effective and power-effective, c) low end-to-end network latency should be guaranteed, d) it should be resilient to blocking or link failures and e) scalablity and flexibility are also indispensable .

By considering the DCN size, data center topologies can be categorized as rack scale topology, data center scale topology, data center cluster scale topology and multi-cluster scale topology. Many existing researches on data center topology have focused on the rack scale and data center scale topologies [63, 65–69] and based on the position of routing intelligence they can be classified as switch-centric topology and server-centric topology [70, 71]. To some extent, many of the data center topologies are similar to the widely used network topologies that link network components together, such as the star topology (as shown in Fig. 2.3), the linear topology and the ring topology. The main difference

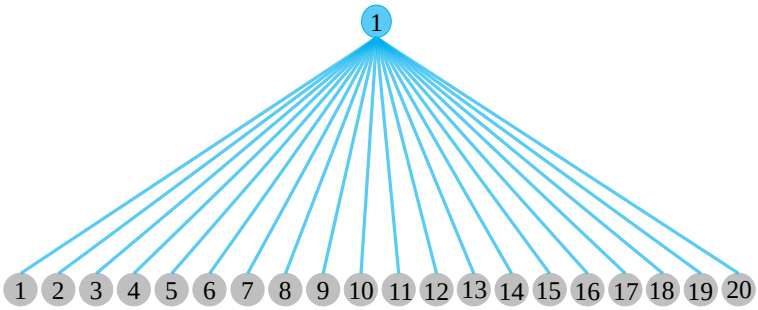


Figure 2.3: Star topology

between them is that data center topologies only serve the essential processes of the data center as a central place to keep all sorts of information. In the following sections, several popular data center topologies are reviewed and among which, some topologies can be used for both intra and inter DCNs.

2.3.1 Star and spine-leaf topologies

Star topology, as shown in Fig. 2.3, is the dominant physical topology for local areal networks [72]. In contrast, in traditional data centers it is mainly utilized for the connection between the servers inside a rack and the ToR switch [73]. In this topology, all the end nodes, e.g. servers or devices, are directly connected to a star node, which can be a switch or a hub, and there is no link connecting any two end nodes directly. That is to say, any end to end transmissions should through the star node. This topology is easy to implement and has great scalability, since only the star node need to be updated when a new end node is added or removed [74]. Moreover, it features good fault tolerance as any single cable/fiber broken or link failure will affect one node only, and provides low routing complexity as well as end-to-end latency attributing to the 1-hop transmission between any two end nodes. However, the star topology is not suitable for network with fast-changing and/or enormous traffic, e.g. inter-rack and inter data center networks, since all the pressure of system failure is put on the star node. In other words, if the star node fails, the whole system will break down.

To keep the benefits of star topology and make it suitable for higher level DCNs, the spine-leaf topology (also know as two-tier clos topology) is proposed. It is one of the most popular topologies for current DCNs, which consists of two layers of nodes, a leaf layer and a spine layer [75–77]. The leaf nodes can be servers, ToRs or higher level switches, while the spine nodes are usually the intermediate switches in DCNs. An example of this kind of topology consisting of 3 spine nodes and 20 leaf nodes is shown in Fig. 2.4. As seen, all the leaf nodes are connected to the spine nodes with a full-mesh topology, and in this way full

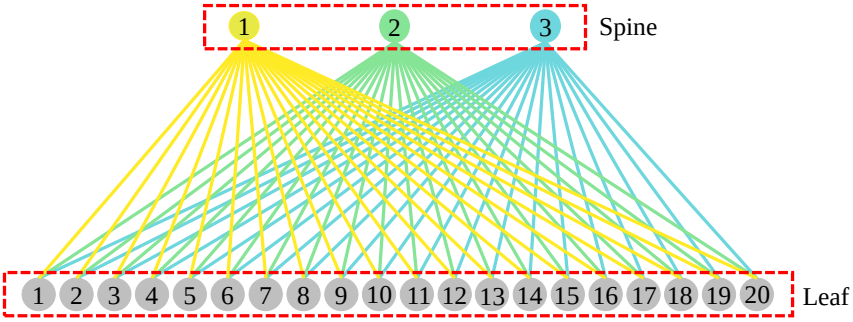


Figure 2.4: Spine-leaf topology

connectivity between all the leaf nodes is enabled. This topology keeps the advantages of star topology in terms of low latency and routing complexity, since any data from a leaf node only has to hop to a spine node and then another leaf node to reach its destination. Compared with the star topology, the sufficient number of links provide a much higher bandwidth for the DCN when the traffic is generated. Moreover, it has better fault tolerance, since if the performance of one spine node saturates or even the node fails, other spine nodes can takeover the responsibility and the performance throughout the DCN would be only slightly degraded [75]. In addition, more than one spines also feature better flexibility for the node selection process, which may lead to much lower blocking probability or request dropping rate (optimizing the east-west traffic), since alternative paths can be selected depending on the amount of spine nodes connected to every leaf node. Furthermore, even if port availability of a leaf node saturates and limits the network capacity, a new leaf node can be easily added by connecting it to every spine node and adding the network configuration to the spine node. This ease of expansion contributes to the optimization of the IT department's process of updating and/or scaling the DCN [78].

However, the spine-leaf topology also faces several challenges [76]. The leading concern is the cost and power consumption for the devices and links. With the rapidly increasing demand on network capacity, a considerable amount of fibers, optical switches and/or other network equipment are required to scale the DCN since every leaf node should be connected to every spine node. Particularly, more expensive optical switches with high port counts at the spine layer are necessitated. Besides, the number of leaf nodes that the topology can support is constrained by the port number of each spine switch (maximal leaf node number $=$ switch port number, assuming no upper links). Consequently, to scale the data center to support more racks, all the spine switches should be replaced by switches with higher port counts. Both the star and the spine-leaf topologies has only two

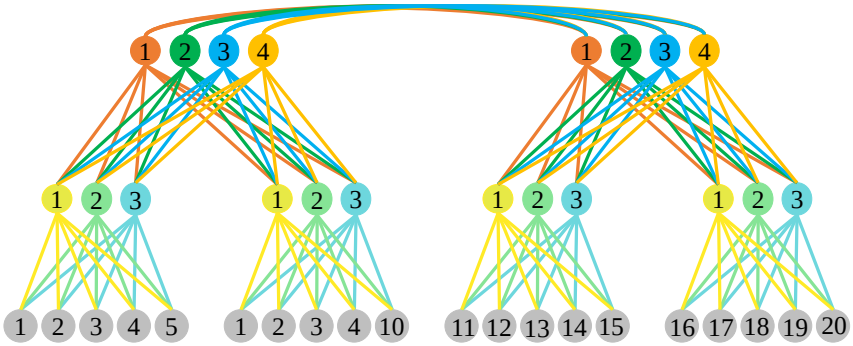


Figure 2.5: 3-tier fat tree toploogy

tiers, however, in practical DCNs 3-tier topologies are widely considered, which may consist of several spine-leaf topologies.

2.3.2 3-tier fat tree topology

Fat-tree topology is another widely investigated topology for both the high performance computing and data center networks [79–81]. As shown in Fig. 2.1, this kind of topology has served the conventional data center well for many years, which provides effective access to servers within the racks and communications between the servers. Figure 2.5 depicts an example of a 3-tier fat tree topology with 20 end nodes (e.g. servers or ToRs) and 20 intermediate nodes (i.e. switches), where the end nodes are equally divided into 4 groups. The 5 end nodes in each group are connected to 3 corresponding intermediate switches with a spine-leaf topology. At the second layer, a group of 6 switches are connected to a group of 4 higher level switches at the third layer with a spine-leaf topology. In the meanwhile, at the top layer, two groups of 4 nodes are fully connected and thus, the interconnection between any two end nodes is enabled. This topology provides great simplicity, as the network issues resulted from the end nodes are simplified because the node number in every group is limited. Moreover, the division on the end nodes also enables the isolation between each group, which increases the security, failure tolerance and scalablity of the network [76].

When the network scale is small, this topology requires only two layers of optical switches, which guarantees the cost and power consumption efficiency. Also, it greatly matches the server functions in traditional data centers that heavy east-west traffic within the rack or rack groups is required, while only limited north-south traffic from the rack groups to the core network is demanded. However, when the network moves to lager scales, lots of devices and fibers should be added corresponding to the layer increase. In addition, high latency attributes to the long-path (e.g. four hops) transmission between any two end nodes may be a

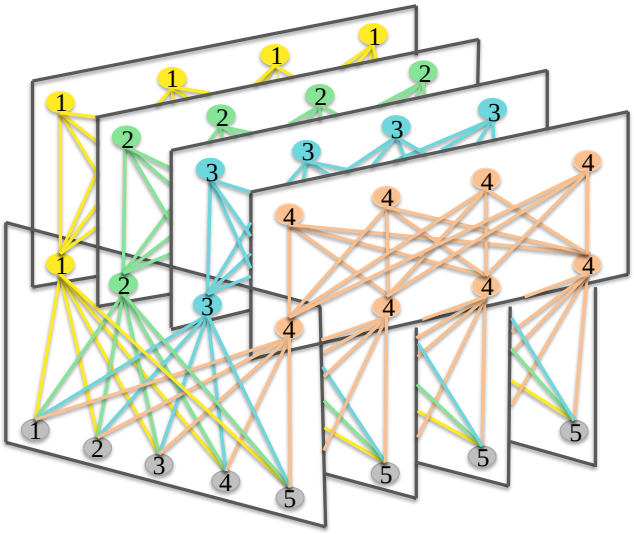


Figure 2.6: Facebook data center network topology

big issue. Oversubscription at the second and third layers should also be concerned since spine-leaf topology is adopted. A oversubscription ratio of < 5:1 between two layers is generally acceptable, however, the requirement is highly dependent on the traffic situation in practice [76].

2.3.3 Facebook data center topology

In order to replace the hierarchically oversubscribed topologies in traditional data centers and build a high performance network, the Facebook data center topology [82, 83] was proposed based on the "4-post" data center topology [84] and it is currently implemented in the Facebook data centers. Unlike the hierarchical tree topology, this topology consists of several new "units of network", which are configured in parallel. As shown in Fig. 2.6, each "unit of network" is a plane housing two layers of nodes, which are interconnected with a spine-leaf topology. In this example, two kinds of "unit of network" are considered for connections between the nodes in the first layer and the second layer, the second layer and the third layer, respectively, featuring a 3-tier micro-cluster structure overall. In order to maintain the same number of end nodes with the aforementioned topologies, each unit at the bottom includes only five end nodes, which can be up to 48 nodes in practical Facebook data center.

Attributing to the innovative disaggregated "unit of network", this topology can be easily extended by adding "unit of network" to satisfy the increasing capacity demand. Moreover, it is resilient to network element failures since individual devices and links are not that important. Compared with the 3-tier fat tree topology, it offers better connectivity for the nodes, which may significantly

reduce the blocking probability and improve the network performance. However, the increase of the number of intermediate nodes and links will considerably rise the total cost and power consumption.

2.3.4 Other data center topologies

All the topologies described previously are indirect topologies, which means that the end nodes in each topology are indirectly connected to each other through one or more intermediate nodes. On the contrary, in direct topology, some of the end nodes are directly connected to each other. Compared with the indirect topologies, the end-to-end latency in the DCN with this approach can be significantly reduced. Good examples for DCNs are 2D-Torus topology [85] and Torus-based topologies, e.g. NovaCube [86] and Mesh-of-Torus [87]. Apart from the aforementioned topologies, many other data center topologies have also been proposed aiming at efficiently interconnecting the servers or switches inside a data center [88, 89] and showing benefits in network capacity, scalability, flexibility, cost, power consumption or low latency, including Slim Fly [63], Jellyfish [66], Bcube [68], Flattened Butterfly [90], SprintNet [91] and so on. In this thesis, I only use three topologies presented in previous subsections, the network performance of the spine-leaf topology and the Facebook data center topology will only be evaluated in Chapter 4, while the 3-tier fat tree topology will be adopted and investigated in the proposed data center architectures in Chapter 4, Chapter 5 and Chapter 6.

2.4 Advanced Link Level Techniques for Maximizing Network Capacity

In order to satisfy the increasing needs of data centers in network capacity, at the link level, links with high bandwidth or bandwidth density are required. To achieve it, two of the most popular techniques are WDM using conventional single-mode fibers (SMFs) and SDM using MCFs.

2.4.1 SMF-based wavelength division multiplexing

WDM is one of the key technologies in optical transmission systems, which uses a single fiber or optical device to carry multiple individual channels [92]. The concept of WDM was proposed and first published in 1970 [93], however, the fundamental research on WDM was not actually started until the middle of 1977 [94]. As demonstrated in Fig. 2.7, in a basic WDM system, multiple signals in different wavelengths (λ) or frequency slots from various transmitters are multiplexed into a SMF through a multiplexer. Afterwards, the signals are transmitted simultaneously inside the fiber until being distinguished and separated

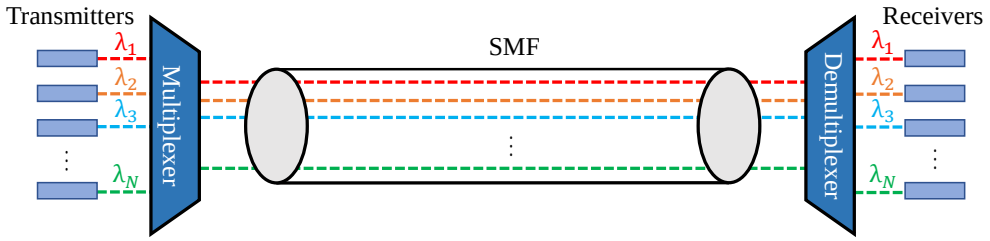


Figure 2.7: Basic WDM system

into independent channels by a demultiplexer before the receivers [95]. In this way, the capacity that a fiber can support is multiplied.

In the past decades, WDM technology has been developed rapidly in optical communication networks [94, 96, 97]. Based on the channel spacing (wavelength grid), the WDM systems can be categorized as the coarse wavelength division multiplexing (CWDM) and the dense wavelength division multiplexing (DWDM), where the former one supports a channel spacing of 20 nm [98] and the latter one can support various channel spacings ranging from 12.5 GHz to 100 GHz and/or wider [99]. However, due to the limitation of the WDM technologies, for example, arrayed-waveguide gratings (AWGs) and micro-electromechanical system (MEMS) devices, which only allow inherently fixed channel granularity set at the time of their manufacture rather than variable passband widths, in the conventional backbone optical networks, fixed grid (e.g. 50GHz or 100 GHz) WDM systems have been widely implemented [100]. This rigid framework may result in considerable wastage of bandwidth or network capacity and inefficient utilization of the spectrum since the traffic requirement may be less than the maximal capacity that one unit of this grid can support [101].

With the development of the WDM technology, e.g. wavelength selective switch (WSS), flexible grid WDM with 12.5 GHz (or even 6.25 GHz) granularity is enabled, which is pushing the optical network from fixed to flexible grid [101–104]. This kind of elastic optical networks [105] is a promising candidate for high-speed optical communications. Moreover, apart from the traditional WDM over wavelengths in C+L-band (12 THz in total), as shown in Fig. 2.8, where the optical fibers have quite low absorption loss, researchers have started to explore the network with O-band wavelengths for both long and short reach networks [26–28], and T-band [106–108] wavelengths for short-reach networks, aiming at realizing ultra-wide band WDM and in turn, increasing the network capacity. To date, WDM technology has played a key role in the optical communications as it has been widely applied to both academical researches and commercial products. However, attributing to the rapidly increasing capacity

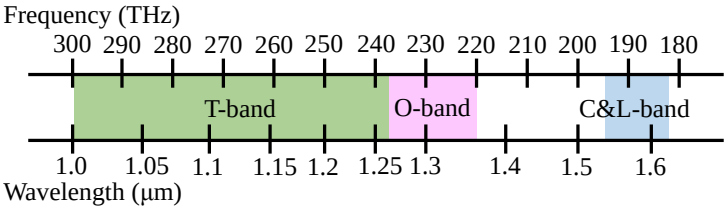


Figure 2.8: Frequency and wavelength bands in optical transmission

requirements in DCNs, even with the WDM technology, the capacity crunch in standard SMF may impose a limit in further scaling the optical transmission systems [109, 110].

2.4.2 MCF-based space division multiplexing

To overcome the crunch issue in SMF solutions, the concept of SDM is recalled, which refers to the multiplexing technology of using the controllable arrangement of optical fibers or channels in the spatial domain. Actually, the concept of using SDM technologies, in particular, combining multiple spatial channels (SMFs or cores) in a single cable or fiber, to increase the capacity can date back to the early 1980s [111, 112].

The initial attempts to deploy SDM were through the means of fiber ribbons or bundles, which could be obtained by packing tens to thousands of conventional SMFs together in parallel [113]. Fiber ribbons or bundles have a great potential for optical transmission systems or DCNs since they can provide high capacity and great compatibility with the WDM technologies. Moreover, compared with using many individual fibers, they can reduce the cabling complexity and increase the spatial efficiency. Nowadays, fiber ribbons or bundles have been commercially manufactured [114, 115] and commonly used in current data centers to connect the severs and racks for several years [116]. Furthermore, the combination of WDM and the aforementioned SDM techniques has been commercially used to further increase the data center capacity [117].

2.4.2.1 Multi-core fibers

However, with the increase of the data center traffic, numerous optical interconnects are required, the current systems using fiber ribbons or bundles will also be bulky and hard to manage. Moreover, the spatial efficiency and scalability will be challenged [118]. To deal with theses issue, the concept of placing multiple single-mode or few-mode cores in a single fiber has come back into the researchers' eyes, which was initially proposed in 1979 [111]. This kind of fiber is named as MCF and according to the coupling style, it can be categorized as uncoupled MCF [119], weakly coupled MCF [120, 121] and strongly coupled

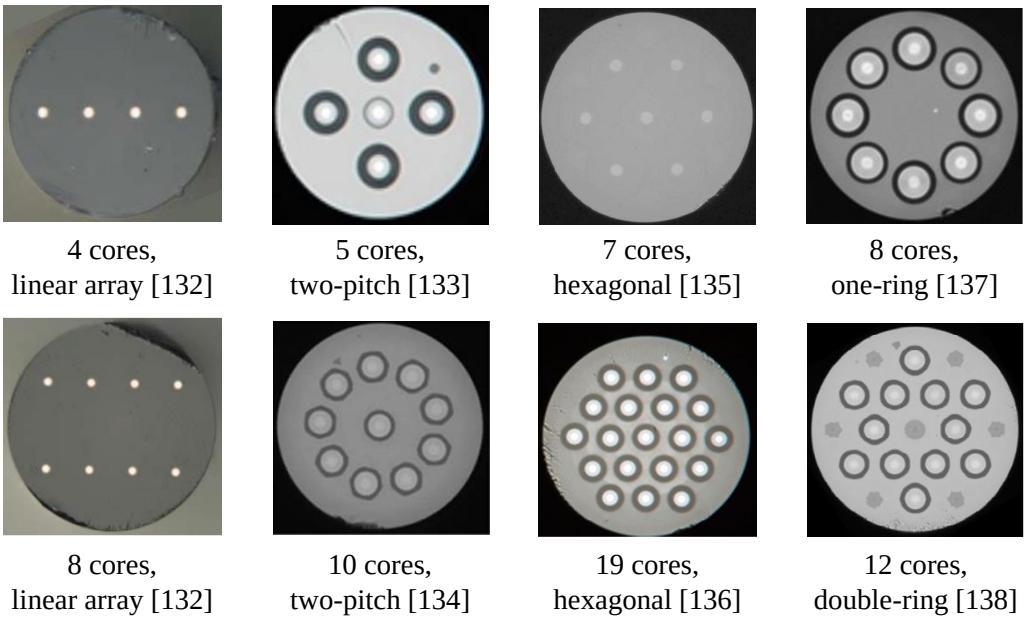


Figure 2.9: Cross-sectional area of different MCFs

MCF [122–124]. Uncoupled and weakly coupled MCFs have a relatively low IC-XT between the adjacent cores while the strongly coupled MCF enables high coupling between signals propagating in neighbouring cores at the cost of large mount of IC-XT even after few meters [125]. Note that, the IC-XT in MCF will be described and analyzed in depth in the next subsection. Due to this reason, the uncoupled and weakly coupled MCFs are more preferred in the existing researches or applications [113]. In fact, compared with SMFs, the principles used for MCF design and fabrication are completely distinctive. Researchers or manufacturers have been looking into various fundamental design aspects behind the MCFs, such as fiber diameter, core number, core arrangement (e.g. core pitch between the cores and the structure), bending radius and cladding thickness [120, 126–131]. According to these researches, the main challenge for placing as many cores as possible into a single MCF is how to avoid the large penalties from IC-XT.

To date, a variety of MCFs with different core numbers and layouts have been proposed and experimentally demonstrated, as shown in Fig. 2.9, including 4-core and 8-core MCFs with linear array structure [132], the latest 5-core [133] and 10-core [134] MCFs with two-pitch structure, the widely considered 7-core [135] and 19-core [130, 136] MCFs with hexagonal arrangement, the 8-core MCF with one-ring structure [137] and the 12-core MCF with double-ring arrangement [138]. These MCFs have various core pitches ranging from 31 to 50 μm and different cladding diameters between 125 μm, which is same as the diameter of a standard SMF, and 204.4 μm. To arrange more cores in a single MCF, such as 22

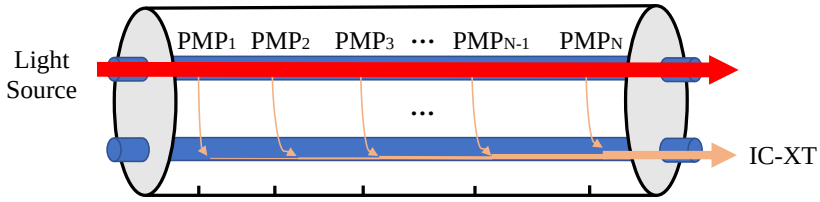


Figure 2.10: IC-XT generation process in MCF.

cores [139], 30 cores [140], 32 cores [141] and 37 cores [142], up to 260 μm cladding diameter is required to achieve an acceptable IC-XT. The state of art MCF is designed and fabricated for data center applications, which has 100 cores within a 320 μm cladding diameter, achieving a low IC-XT of < -25 dB for 1 km length [143].

2.4.2.2 Inter-core crosstalk

Ideally, the capacity that a MCF can offer is linearly proportional to the core number, however, as aforementioned that the core number and the performance of a MCF is limited by the inter-core interference or IC-XT between the adjacent cores in practice. IC-XT is an unwanted interference inherent to MCFs, which can be considered as the power leakage from the excited core to the target core. It occurs at stochastically distributed discrete points along the fiber where the principal and IC-XT signals match in phase [120]. These discrete points are called phase-matching points (PMPs), shown in Fig. 2.10, where the total IC-XT can be approximated as the sum of the IC-XT contributions overall.

The statistical mean IC-XT per meter of a homogeneous MCF, inside which, all the cores are made of the same material (i.e. same refractive index) and have equal radius, can be expressed as Eq. (2.1) [144, 145]:

$$h(C_P) = 2\frac{\kappa^2 R}{\beta C_P} \tag{2.1}$$

Where R, β and C_p stand for the bending radius, propagation constant and core pitch between two neighbouring cores, respectively. κ denotes the mode coupling coefficient, which is dependent on the MCF structures, i.e. normal step-index (SI) structure and trench-assisted (TA) structure. The profiles of the refractive index for SI-MCF is shown in the left of Fig. 2.11 and the mode coupling coefficient κ' between two adjacent cores can be expressed as [146, 147]:

$$\kappa' = \frac{\sqrt{\Delta_1}}{a}\frac{U_1^2}{V_1^3 K_1^2(W_1)}\sqrt{\frac{\pi a}{W_1 C_P}}e^{-\frac{W_1 C_P}{a}} \tag{2.2}$$

In this equation, a represents the core radius and Δ_1 is the refractive index

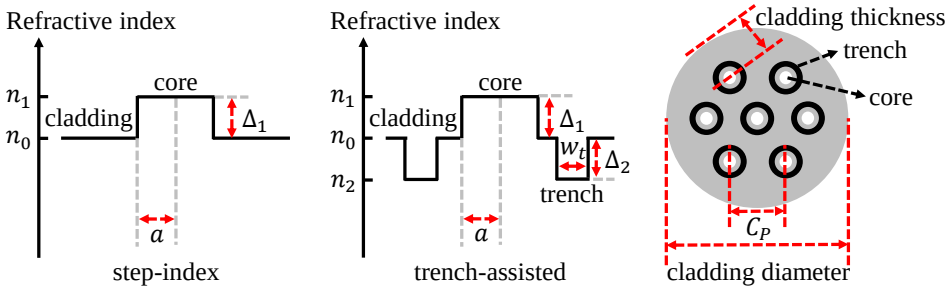


Figure 2.11: Profiles of refractive index and cross-sectional area of MCF

difference between the core (n_1) and cladding (n_0). $U_1 = a(k^2n_1^2 - \beta^2)^{1/2}$ and $V_1 = kan_1(2\Delta_1)^{1/2}$, where $k = 2\pi/\lambda$ stands for the wave number and λ is the wavelength of the light. $K_1(W_1)$ represents the modified Bessel function of the second kind with first order and $W_1 = a(\beta^2 - k^2n_0^2)^{1/2}$.

Compared with SI-MCF, trenches with low-index are added surrounding the cores to suppress the electric field distribution in each core. Therefore, the overlap of electric field between adjacent cores are smaller and thus, reduces IC-XT in MCF [148, 149]. The cross-sectional area and refractive index profile of a TA-MCF are also depicted in Fig. 2.11. As seen, two more parameters are considered, w_t, the trench width and Δ_2, the refractive index difference between trench (n_2) and cladding. According to [145, 150, 151], the mode coupling coefficient κ" in homogeneous TA-MCFs can be simplified as:

$$\kappa'' = \kappa' \sqrt{\Gamma} e^{-2(W_2 - W_1)\frac{w_t}{a}} \tag{2.3}$$

where $\Gamma = W_1/[W_1 + (W_2 - W_1)w_t/C_P]$, in which $W_2 = (V_2^2 + W_1^2)^{1/2}$ and $V_2 = kan_0(2|\Delta_2|)^{1/2}$.

Eq. (2.1) presents the mean IC-XT between any two cores, however, in a MCF the target core may have more than one adjacent cores. Based on [120, 127], the total crosstalk on any target core can be calculated by:

$$IC\text{-}XT_{total} = \frac{n_c - n_c e^{-(n_c+1)hL}}{1 + n_c e^{-(n_c+1)hL}} \tag{2.4}$$

In which, n_c is the number of the neighboring cores around the target core and L stands for the fiber length. The numerator stands for the total signal power of the adjacent cores, while the denominator represents the power of the target core, with the consideration of the power leakage among them. Note that, this equation is only for uni-directional homogeneous MCF with uniform core pitch. For uni-directional homogeneous MCF with unequal core pitch, bi-directional homogeneous MCF and/or heterogeneous MCF, inside which the cores may have

different refractive indexes or radius, new equations should be derived.

2.5 Routing and Resource Allocation Algorithms

The aforementioned data center architecture, topology and links (including the multiplexing techniques) can physically affect the DCN performance, while the routing and resource allocation algorithms can influence the DCN properties in the software aspect. Generally speaking, routing is the process of selecting a path or several paths for data transmission between any two nodes inside a network or across multiple networks [152], while resource allocation is the process of assigning and managing available resources for different network requirements. In optical DCNs, the nodes can be severs or optical switches while the path is the lightpath between them, which may consist of one or several optical links. Usually, the resource includes IT resources, for example, CPU, memory and storage resources, and network resource, such as bandwidth. Specifically, for DCNs using WDM and/or SDM technologies, the resources can be wavelengths (frequency slot) or cores [153–155]. Since developing a globally optimal solution for the routing and resource allocation problems is incredibly difficult, the adoption of heuristics that are locally optimal, and those that take a strategic approach to resource allocation may contribute to higher resource utilization and lower blocking probability. This is to say, if the system is physically configured, the routing and allocation algorithms can dominantly determine how efficient the network is. Several commonly utilized algorithms for both standard networks and specific WDM and/or SDM-based networks are reviewed in this section.

2.5.1 Dijkstra's algorithm and K-shortest path algorithm

Lots of routing algorithms have been developed for selecting the path between any two nodes, for example, shortest path algorithm, flooding algorithm, distance vector routing algorithm, link state routing algorithm, hierarchical routing algorithm and multicast routing algorithm [156]. For the research in this thesis, I mainly use the famous K-shortest path algorithm [157].

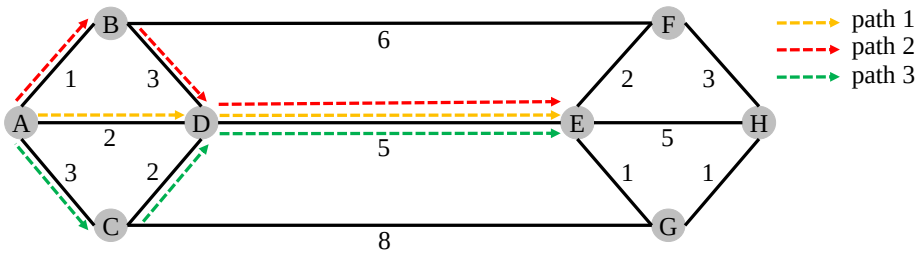


Figure 2.12: Routing with K-shortest path algorithm

Before describing the K-shortest path algorithm, Dijkstra's algorithm [158] should be introduced first, which is one of the shortest path algorithms. As shown in Fig. 2.12, the distances between any two adjacent nodes are provided, where the distance can also be replaced by weights considering bandwidth and/or latency according to the network requirements. Note that, the distances or weights should be non-negative. At the beginning, no paths are known and all nodes are labeled with infinity. Subsequently, the algorithm will go through nodes that have not been checked and calculate the total distances or weights from the nodes to the source node. In the meanwhile, the label on each node will keep updating until the smallest value is found. Once it is found that the label shows the shortest path between the source node and that node, the label will be made permanent and never be checked or changed again. Step by step, every node will be labeled with the smallest distance or weight to the source node as well as the corresponding intermediate node information. It can be seen that, with this approach, although the shortest path can be found, it may consume long time and this time will exponentially increase with the network size increase [156].

K-shortest path algorithm is a generalization and an extension to the Dijkstra's algorithm, which can offer K alternative shortest paths. Figure. 2.12 shows an example of using 3-shortest algorithm for routing between nodes A and E. As seen, three paths: A-D-E, A-C-D-E and A-B-D-E are found when K=3, while with Dijkstra's algorithm, only the path A-D-E will be found. With this approach, the blocking probability or traffic congestion of the network can be notably reduced since part of the traffic pressure can be undertaken by the other two paths. Moreover, since the algorithm will also follow the rule of the shortest path the first, there will be no extra resource wastage compared to the routing process with the Dijkstra's algorithm.

2.5.2 First-fit algorithm

First-fit algorithm is one of the simplest resource allocation algorithms, in which the resources are allocated following a first-fit basis [159–161]. In other words, each request will be allocated to the first available resource(s) during the resource identification procedure. Since it is simple and easy to implement, it has been widely used for memory allocation [162] or wavelength assignment in optical networks [163]. However, if the network is large or the request number is huge, this algorithm can hardly find the optimal allocation solutions since the algorithm does not take the availability of network resources into consideration. That is to say, there may be no links can provide sufficient bandwidth between the two selected end nodes in DCNs, or the nodes founds are far away from each other. As

a consequence, this may lead to high blocking probability (or request dropping rate) or high transmission latency. In terms of the spectrum allocation in WDM-based network, it may also result in severe fragmentation issues and in turn degrading the network performance.

2.5.3 Spectrum and core allocation algorithms

In WDM and/or SDM-based networks, spectrum and cores are also resources need to be considered in the allocation process. To date, numerous algorithms has been developed for SMF-based long-haul backbone networks and recently, a few studies have proposed resource allocation algorithms that take the MCF-based SDM technique [155, 164, 165] into account. The introduction of MCF brings additional flexibility in spatial domain for the network, however, it also increases the complexity of the allocation process [153, 166], since more constraints such as IC-XT need to be taken into consideration. To address these constraints, different mechanisms for resource allocation has been adopted, including core priority, core switching and slot split schemes.

Core priority scheme was proposed as a mechanism to reduce the IC-XT between neighbouring cores in MCF, which could pre-define the sequence of core usage for uni-directional transmission [164, 167]. The employment of core switching scheme aims at alleviating the effect of spectrum continuity constraint, which indicates that the same spectrum or wavelength should be allocated at all links along the whole transmission path. This mechanism allows the connections to use different cores on each link along the path while still using the same wavelength. With this approach, the freedom of the spectrum allocation can be increased compared to the case that each connection is limited to always use the same core [155].

Spectrum	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Core1															
Core2															
Core3															

(a) Before the 3-slot request arrives

Spectrum	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Core1															
Core2											Low	Low	High		
Core3															

(b) 3-slot request allocation limited by the spectrum contiguity constraint

Spectrum	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Core1															
Core2				Low							Low	Low			
Core3															

(c) 3-slot request allocation without spectrum contiguity constraint

Figure 2.13: Effect of spectrum contiguity constraint on IC-XT

The spectrum contiguity constraint enforces the network to assign contiguous spectrum slots to each individual request [166], as a consequence, the flexibility of spectrum-slot selection process can be reduced. Moreover, it can potentially lead to spectrum allocations with higher IC-XT. Figure. 2.13 shows an example to demonstrate the effect of the spectrum contiguity constraint on the IC-XT. As seen, Fig. 2.13 (a) depicts the initial status of three cores with 15 spectrum slots at a random time slot during the allocation process, where the occupied slots are shown in red and the unused ones in white. When a 3-slot request arrives, due to the spectrum contiguity constraint enforced in Fig. 2.13 (b), the only possible choice is Slots 11-13 in core 2. Such assignment results with a high level of IC-XT for Slot 13 because it has already been used in cores 1 and 3. If the contiguity constraint is relaxed, as shown in Fig. 2.13 (c), the request can be split and assigned Slots 4, 11 and 12 of core 2, which will result with minimal IC-XT from neighbouring cores, since they have not been used in any other cores.

These IC-XT increases can have a significant effect on the requests using higher-order modulation formats due to their sensitivity to distortion and noise. To alleviate this constraint, the slot split scheme has been developed, which can divide a request demanding a wide spectral bandwidth into several requests with smaller spectral bandwidth requirements, where the bandwidth of each split request can be as small as the minimum slot size accepted in the system [168].

2.6 Summary

This chapter has reviewed the fundamental concepts for understanding the research presented in the following chapters. Theory described in Section 2.2 will be used in Chapter 5 and Chapter 6, while the topologies and algorithms introduced in Section 2.3 and Section 2.5 will be used in Chapter 4, Chapter 5 and Chapter 6. Theory presented in Section 2.4 will be mainly used in Chapter 3 and Chapter 4.

Chapter 3

Variations and Accuracy of IC-XT in Trench-assisted MCFs

3.1 Introduction

As previously described in Section 2.4.2, SDM using MCF is a promising solution to cope with the capacity crunch in standard SMF-based optical communication systems. Nevertheless, the achievable capacity of the MCF is limited by the inherent IC-XT between the adjacent cores, since it can potentially reduce the optical signal-to-noise ratio (OSNR), and subsequently, affects the performance and power budget of the system [169]. Heretofore, lots of studies on MCF-based networks have considered IC-XT as a static value that is able to be calculated from the existing mathematical models [44, 155, 167, 170–172]. However, several recent researches [173–176] have shown that the number and/or distribution of the phase matching points (PMPs), which are the discrete points where IC-XT is from, can vary randomly with each core's characteristics, fiber structure and the conditions under which the fiber is placed. Especially, the longitudinally varying perturbations along the MCF, including macro bending, twists and structural fluctuations, that lead to the phase offset variance, can make the IC-XT dynamic. For instance, for a continuous-wave (CW) source light, the induced IC-XT can vary by up to more than 15 dB over a 1-hour time window [174, 176]. As a consequence, these IC-XT fluctuations may enforce the optical system to set a much higher OSNR margin to ensure sufficient system performance for a long run. To date, many analytical models have been developed to analyze the IC-XT characteristics, including the fluctuations in homogeneous MCFs [173, 175, 177, 178]. Especially, a model that was proposed based on a Brownian motion for the time dependent IC-XT showed an great fitness to the experimental results [178]. In addition, some of the studies have theoretically shown that IC-XT and the IC-XT fluctuations are dependent on the

wavelength [147] and skew between the neighboring cores [176].

In addition, the properties of the launching optical signals, e.g. symbol rate and modulation format, can also impact the IC-XT variations. Several experimental results have shown that by either increasing the bandwidth of intensity modulated signals or adopting the phase modulation techniques, the variations can be suppressed accordingly [176]. Referred to [179], this suppression can be attributed to the reduction or elimination of the power of the residual optical carrier. However, to the best of the author's knowledge, there is no research on comparing the performance of IC-XT induced by different intensity modulated formats, such as on-off keying (OOK) and pulse amplitude modulation (PAM). In contrast, in terms of the phase modulated formats, theoretically and experimentally evaluations on the impacts of quadrature phase shift keying (QPSK) and binary phase shift keying (BPSK) with a variety of symbol rates on IC-XT variations have been done [176], respectively. However, considering the higher order phase modulated formats, only the dual-polarization 16-quadrature amplitude modulation (DP-16QAM) that was operated at a single wavelength of 1550 nm and a single baud rate of 24.5 GBaud, has been experimentally demonstrated.

Besides, in the real data centers, temperature changes over time [180], which can lead to the length of the silica fibers alter by 4.1×10^{-7} mK^{-1} [181]. Moreover, this change can affect the refractive index of the fiber and the degree of the effect is related to the properties of the core, such as radius and original index [182]. Consequently, these changes may cause the latency or skew as well as the dispersion of the fiber change. Given the characteristics of MCFs, temperature fluctuations can have a similar effect on these fibers, for example, it has been experimentally shown in [183] that the higher the temperature, the longer the skew, provided that evolution of IC-XT inside the MCF is associated with the skew and structural variations along the fiber. Therefore, it is clear that the temperature fluctuations across the MCF can affect the behaviour of the IC-XT. Not only the temperature, pseudo random binary sequence (PRBS) pattern, which is one of the most important parameters for bit error rate (BER) evaluations in the transmission testbeds in the labs or for practical applications, is also a practical factor that need to be taken into account since it can fundamentally change the properties of the power spectral density (PSD) of the modulated signals [184, 185], which may potentially alter the behaviour of IC-XT. In addition, as different PRBS patterns can be easily generated and have been used by different research groups or applications, an in-depth study on the relationship between PRBS pattern and the IC-XT behaviour is necessary. However, there have been limited studies on the impact of temperature variation on IC-XT [174] and no study on the effect of PRBS pattern on IC-XT so far.

Apart from the aforementioned transmission parameters that can fundamentally affect the IC-XT itself, the value or the accuracy of the observed IC-XT can also be influenced by the configuration of the devices, such as the averaging time of the power meter, and the measurement time window (i.e. how long the IC-XT is measured) [186], in the real world experiment process or application. The differences on the observation results in terms of the average IC-XT, the IC-XT variations and the worst-case IC-XT resulted from these parameters can significantly affect the MCF applications. In the existing researches, different averaging times and time windows have been used [175, 175, 186–188] and in particular, most of the investigations on the IC-XT behaviour using power meter for measurements were limited in determining IC-XT fluctuations at sub-Hz levels, which might cause considerable differences. However, to the best of the author's knowledge, few studies have shown the impacts of these parameters. Although, the time window effect was analyzed in [186], only the narrow band CW source light had been experimentally demonstrated. There is still no reference showing that what kind of power meter should be used or for how long the measurement is sufficient to achieve a good IC-XT accuracy. Therefore, an exploration of the relationships between these parameters and the observed IC-XT is inevitable.

This chapter presents a comprehensive study on the variations and accuracy of IC-XT in a TA-MCF. For the first time, the effects of temperature and PRBS length on IC-XT is investigated. Besides, a variety of advanced and practical modulation formats, such as QAM with different cardinalities and a commonly used format in data center, PAM-4, operating at different baud rates are adopted to extend the studies in [176]. Moreover, I evaluate the impact of the number of the excited neighbouring cores on the IC-XT behaviour of the target core and analyze it with the consideration of decorrelation time. To provide a full picture of IC-XT over a wide band, the IC-XT was measured with a time window of up to 12 hours for a ultra-wide wavelength window spanning 250 nm (O-S-C-L bands). Additionally, in order to validate the existing theoretical IC-XT models, the obtained results are compared to the theoretical estimations. At the end of this chapter, the relationships between the accuracy of the observed IC-XT and the measurement time window as well as the averaging time for different signalling sources are studied. In the meanwhile, an novel analysis on the IC-XT step distribution is presented.

3.2 Static and Dynamic IC-XT

For simplicity, in this chapter the IC-XT is classified as static IC-XT and dynamic IC-XT, where the former one stands for the average IC-XT for a long term (> 1

hour), which can be regarded as the mean of the short term (e.g. 25 ms) average IC-XT over a time window of 1 to 12 hours. Furthermore, it is found that the value of static IC-XT is approximately equal to the statistical mean IC-XT, which can be calculated by multiplexing $h(\kappa, C_p)$ (Eq. (2.1)) with the fiber length L in a homogeneous MCF [144], as expressed in Eq. 3.1. Therefore, static IC-XT can also be seen as the statistical mean IC-XT.

$$IC\text{-}XT_{mean} = h(\kappa, C_p)L = 2\frac{\kappa^2 R}{\beta C_P}L \tag{3.1}$$

From the above equation, it is easy to find that if the bending radius (R), propagation constant (β), core pitch (C_p) and fiber length are fixed, the static IC-XT is linearly related to the square of the coupling coefficient (κ^2). According to Eqs. (2.2) and (2.3), κ is dependent on the operational wavelength (λ), therefore, IC-XT is wavelength dependent. This wavelength dependence can also be explained by that the mode field diameter of each core enhances with the increase of the wavelength, as a consequence, the mode field area overlapping between the neighbouring cores becomes wider, triggering higher power leakage between the cores. Moreover, since the coupling coefficient is also related to the refractive index contrasts between the core, cladding and trench, temperature changes that can alter the refractive index may also affect the static IC-XT accordingly.

In contrast, dynamic IC-XT is defined as the range of IC-XT fluctuation for a relative long time window ($\geq$ 1 hour), which is equal to the difference between the maximal and minimal observed IC-XT. To analyze the dynamic IC-XT behaviour, a commonly used analytical model for IC-XT in frequency domain is adopted, which can be expressed as:

$$A_n(z_l, \omega) = A_n(0, \omega) - jK_{nm}\sum_{l=1}^{N} e^{-j[\Phi_l(t)+sz_l\omega]}A_m(z_{l-1}, \omega) \tag{3.2}$$

This equation was derived and first published in [175] based on [173], and then cited and modified in [176]. Afterward, the time variant t was added in [189]. $A_n(z_l, \omega)$ in the equation represents the complex amplitude of the target core (n) at the l-th PMP. ω is the angular frequency and s denotes the difference between the group delays of the target core and the excited core (m) for a unit of fiber length and sL is referred to as inter-core skew [176]. N stands for the total number of PMPs, which can be roughly calculated by $N \approx L\gamma/\pi$, where γ is the twist rate. $\Phi(t)$ denotes the time-dependent phase offset between every two continuous PMPs, which randomly varies between 0 to 2π. z_l is the distance of the l-th PMP along the fiber from the

start of the fiber and K_{nm} represents the discrete coupling coefficient between the two cores, which has been modelled as:

$$|K_{nm}| = \sqrt{\frac{2\pi\kappa^2 R}{\beta C_p \gamma}} \tag{3.3}$$

If the IC-XT is adequately low ($A_n(z_l,\omega) \ll 1$), $A_n(0,\omega) = 0$ and $A_m(z_{l-1},\omega) = 1$, Eq. (3.2) can be simplified as:

$$A_n(L,\omega) \approx -jK \sum_{l=1}^{N} e^{-j[\Phi_l(t)+sz_l\omega]} \tag{3.4}$$

According to the theoretical analysis in [176], if the skew-distance-baud rate product ($sz_l\omega$) is much bigger than $\Phi(t)$, either the skew between the two cores (s) is large or the signals are modulated with high baud rates, the IC-XT will feature high stability (i.e. small dynamic IC-XT). On the contrary, if the baud rates for the modulated signals are low or the the skew between the two adjacent cores is small, $\Phi(t)$ will be much larger than $sz_l\omega$ and then dominantly determines the IC-XT variance. Consequently, big dynamic IC-XT will be observed.

3.3 Experimental Setup for Measuring IC-XT

To explore the IC-XT behaviour under different conditions, many experiments were conducted with a 8-core TA-MCF, which had been previously showcased in [190]. The experimental setup utilized to measure the IC-XT and its variations is presented in Fig. 3.1, while the cross-sectional area of the employed TA-MCF is shown in the right of the figure. It can be seen that, the setup can be divided into two parts, signal generation and power measurement, among which the signal generation consists of three subsystems that can generate various types of source signals to guarantee a fully study on the behaviour of IC-XT .

As seen in the subsystem (a), in order to explore the behaviour of IC-XT induced by intensity modulated signals over an ultra-wide wavelength band, two tunable lasers, one covered O-band wavelengths (i.e. 1260-1360 nm) with 500 kHz linewidth and the other one covered S-, C-, and L-band wavelengths (i.e. 1480-1630 nm) with 200 kHz linewidth, were used to generate CW seed lights, seperately. The generated CW lights were then modulated by a Mach-Zehnder modulator (MZM) driven by an electrical modulation signal containing either OOK or PAM-4 signals. Both the OOK and PAM-4 signals were generated via a pulse pattern generator (PPG), which could operate with either 10 or 25 GBaud signalling rates. In addition, the PPG enabled a reconfigurable PRBS pattern with lengths of 2^i-1 (i = 7, 9, 10, 11, 15, 20, 23 and 31). In contrast, in the subsystem

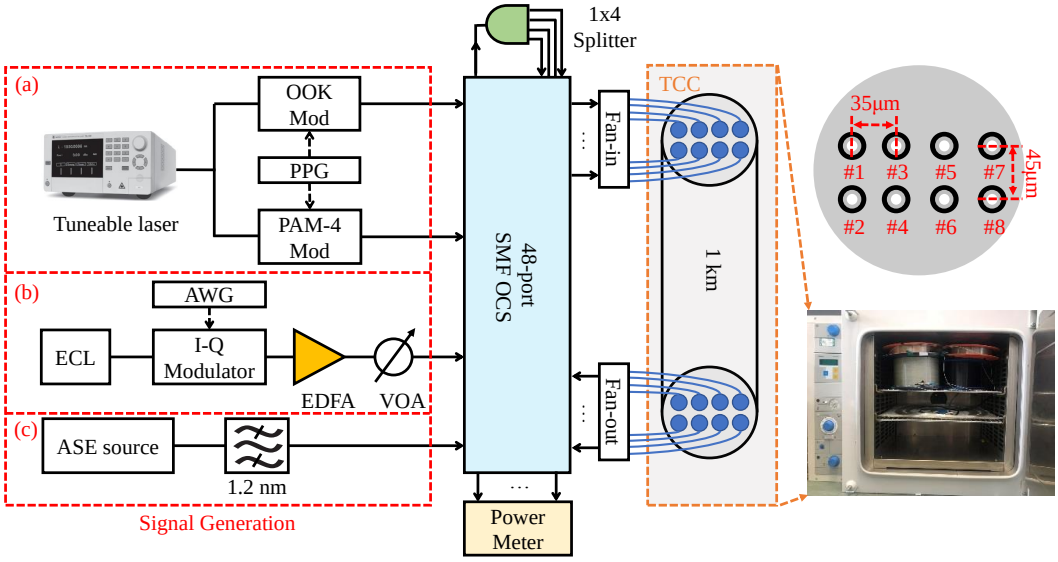


Figure 3.1: Experimental setup for IC-XT measurement and profile of the fabricated 8-core TA-MCF

(b), in order to investigate the behaviour of the IC-XT induced by different I-Q modulated signals, a dual-polarization I-Q modulator driven by a 92 GS/s arbitrary waveform generator (AWG) modulated the C-band CW signals generated by an external cavity laser (ECL) with 100 kHz linewidth. The AWG could generate different m-ary (m = 4, 16, 64 and 256) DP-QAM formats and each QAM format could operate at a variety of signalling rates ranging from 15 to 80 GBaud. The I-Q modulated QAM signals were then amplified by an erbium-doped fiber amplifier (EDFA), where its output power was then controlled by a variable optical attenuator (VOA) before being launched into the following device. The measured OSNR at the output of the VOA was ranging from 46.3 to 48.6 dB. In the subsystem (c), a C-band amplified spontaneous emission (ASE) source was generated by an EDFA and its output was passed through a 1.2 nm band-pass filter to generate a broadband signal with 150 GHz bandwidth prior to being launched into the following optical switch.

It can be seen that the output(s) of each subsystem was connected to the input(s) of a 48-port optical circuit switch (OCS) through multiple SMFs. Moreover, the input of the TA-MCF was also connected to the switch via a fan-in device and thus, the generated source signals could be launched into any group or individual core in the MCF via the fan-in device with or without passing through a 1x4 splitter, which enabled the replication of the same signal source. The 8-core TA-MCF employed here had a total length of 1 km with a cladding diameter of 180 μm and a bending radius of 0.17 m. The core pitch between any two adjacent

cores in horizontal was 35 μm and in vertical was 45 μm. In order to investigate the effect of temperature on the IC-XT behaviour, the TA-MCF was resided within a temperature-controlled chamber (TCC), which could shift the temperature from 20 to 80 °C or keep a given temperature in this range for any time period. The observed IC-XT was the ratio of output power of the crosstalk core to that of the excited core, and to measure the output power of the excited core and crosstalk core, the output of the MCF was connected to the fan-out device and fanned-out into eight SMF links. The SMF links corresponding to the exited core(s) and crosstalk core were then connected to the ports of the power meter via the OCS. In this work, the power meter adopted had an averaging time of 25 ms and could detect power levels from -80 to $+10$ dBm [191]. The switch, splitter and fan-in/out device had 1 dB, 6 dB and 3 dB of insertion loss, respectively. The crosstalk between the utilized switch ports was $<$ -80 dB. To avoid any nonlinearities being induced in the experimental process, the output power for each subsystem was set to 4 dBm before being launched into optical switch, by controlling the output power of either the VOA or the lasers.

3.4 Transmission Parameters that Influence IC-XT

As mentioned in the previous sections, the characteristics of the static and dynamic IC-XT can be influenced by many transmission parameters, including signalling source, baud rate, modulation format, temperature, wavelength, PRBS length and number of excited cores. This section presents the obtained experimental results considering all these factors and for each of them, a corresponding analysis or discussion is provided.

3.4.1 Static IC-XT between cores

To check the accuracy of the experimental process and the observed results, the observed core to core (pairwise) static IC-XT is compared to the theoretical value estimated using Eq. (2.1), and the results are shown in Fig. 3.2. In this figure, the IC-XT was induced by a 1550 nm ASE light source and each bar is the average result over a time window of 1 hour. During the measurement process, only one core was chosen as the excited core (e.g. core 1) at one time, and then the IC-XT between this core and the other seven cores (e.g. cores 2-8) were measured sequentially and individually. As seen, the observed static IC-XT between the cores varies in the range of -71 dB to -44.9 dB, roughly following the expectation according to Eq. (2.1), the bigger the core pitch between the two investigated cores, the lower the IC-XT. Moreover, the estimated IC-XT between a core pair with different core pitches is also shown in the figure with dash lines and the comparisons indicate that

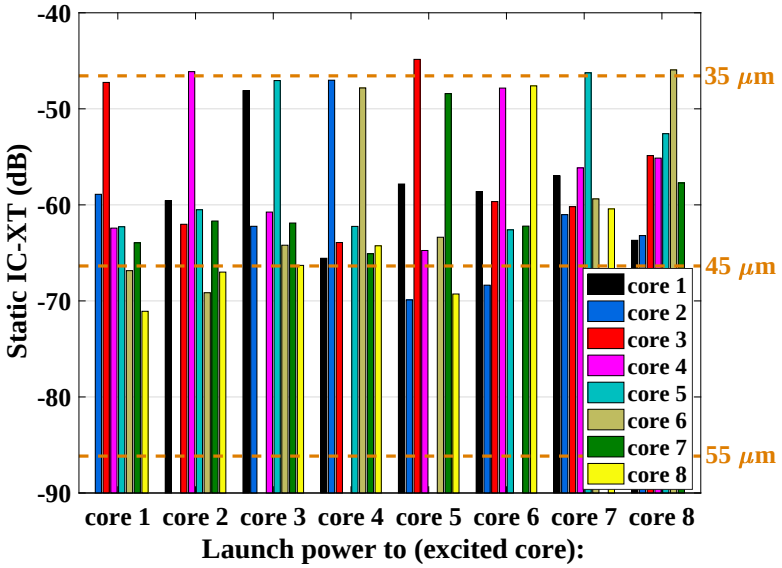


Figure 3.2: Measured and estimated pairwise IC-XT between cores

when the core pitch between the cores is 35 μm, e.g. core 1 and core 3, core 6 and core 8, the observed static IC-XT is almost equal to the theoretical value, confirming the accuracy of the experiment. However, when the core pitch goes up to 45 μm, e.g. core 1 and core 2, core 5 and core 6, slight higher IC-XT than the estimated value can be observed. What is worse, the measured IC-XT is much higher then the predicted value when the core pitch reaches 55 μm, e.g. core 1 and core 4, core 6 and core 7. To explore the reason behind them, the crosstalk from the fan-in/out devices were measured. The obtained results, ranging from -72 to -55 dB, indicate that this kind of crosstalk contributed to the observed IC-XT slightly when the core pitch was 45 μm, however, it dominated the observed value when the core pitch exceeded 55 μm. Therefore, to ensure that the observed IC-XT was dominated by the IC-XT from MCF, in the following subsection 3.4.2 and subsection 3.4.3, core 1 and core 3, with 35 μm core pitch, were utilized for investigations. Among which, core 1 was the excited core while core 3 was the target core. As for the subsection 3.4.4, since there were more than one excited cores, the core number will be presented in detail inside the subsection.

3.4.2 Light source, modulation format and baud rate

Up to now, a variety of source signals have been adopted to stimulate and study the behaviour of IC-XT, among which the CW, ASE and OOK signals are the most popular types [176,188,192]. Therefore, the normalized IC-XT over a time window of 12 hours for these three types of signals is firstly presented in Fig. 3.3, in which every IC-XT sample is the short term average IC-XT over 25 ms. As expected, due to the bend, twist and structural fluctuations along the MCF, the observed IC-

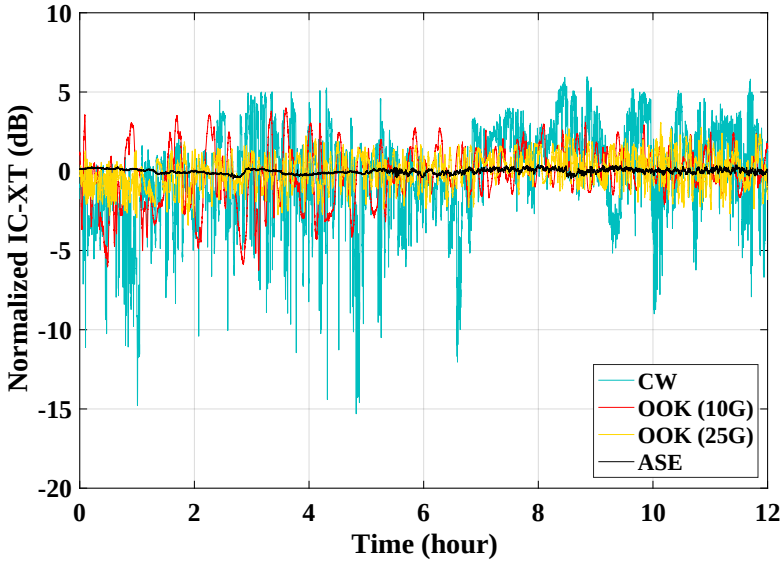


Figure 3.3: Normalized IC-XT over time in MCF for various signalling sources

XT fluctuates rapidly over time. In addition, the degree of the fluctuation changes with the the type of source signals and/or the baud rate of the modulated signals. It can be easily seen that with the narrow band CW light source, the induced IC-XT provides the highest level of dynamicity, i.e. 22.58 dB. On the contrary, with the wide band ASE light source, the induced IC-XT is quite stable around its statistical mean and has a fluctuation range of merely 0.92 dB. This can be explained by the fact that with wider bandwidth, the IC-XT can be better averaged. Compared these results with that in [176] (i.e. 0 dB, 16 dB and 6 dB dynamic IC-XT for ASE, CW and 10G OOK sources, respectively), it is found that higher degree of dynamicity was observed in our setup. This is because the MCF employed in this work is designed and fabricated for data center usage, therefore, it is much shorter than the one utilized in [176]. Moreover, the fiber type, i.e. trench-assisted rather than step-index, layout and core pitch, which are the parameters that can significantly affect the behaviour of IC-XT, are also different. In contrast, the stability of the IC-XT induced by OOK signals is between the cases with CW and ASE signals and when the baud rate of the signals increases from 10 GBaud to 25 GBaud, a 5 dB reduction on the dynamic IC-XT can be observed. Besides, by analyzing the static IC-XT over different time windows, it is found that, 1-hour static IC-XT for ASE and OOK sources is within $\pm$0.5 dB to that of the 12-hour results. Moreover, in parts of the network (e.g data centers, metro) optical circuits/connections may only last just seconds, minutes or hours. Therefore, most of the results in the following sections were obtained over an 1-hour time window to study the IC-XT behavior in this short period of time. Additionally, the impact of the observation time on

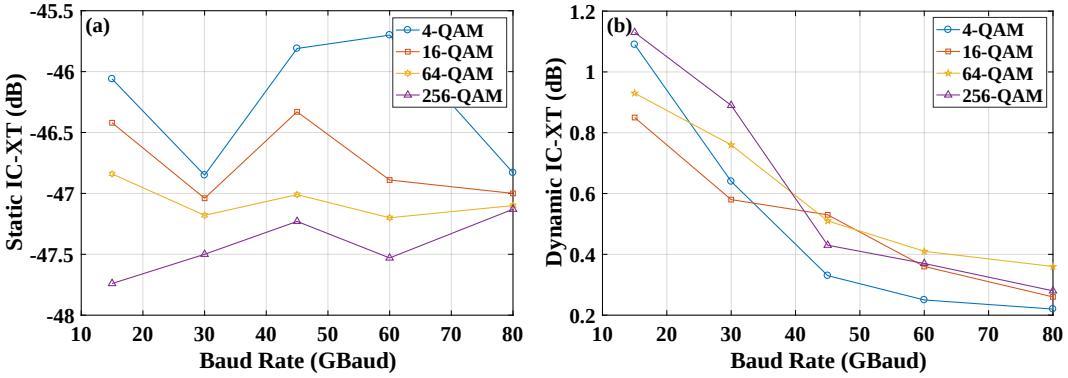


Figure 3.4: (a) Static and (b) dynamic IC-XT for QAM formats with various baud rates

the value of the observed static and dynamic IC-XT will be described in detail in Section 3.5.

Not only the aforementioned three types of source signals, I-Q modulated signals with various baud rates, ranging from 15 to 80 GBaud, were also used to stimulate IC-XT. Each point on both the figures was obtained over a time window of 1 hour. As shown in Fig. 3.4(a), when the order of QAM format increases from 4 to 256, nearly 1.5 dB reduction on the static IC-XT can be observed. To explore the reason behind it, I checked the OSNR of the QAM signals and found that due the configuration of the setup, the OSNR decreased with the QAM order increase. When the format changed from 4-QAM to 256-QAM, over 2 dB OSNR decrease were observed, indicating that observation results were due to the OSNR variation rather than that the IC-XT was intrinsically affected. Therefore, as OSNR difference can be observed on any transmission and network experiments, similar observations should not be wrongly perceived as the IC-XT change. Moreover, it can also be seen that for every QAM format, when the baud rate changes, the static IC-XT fluctuates slightly around its mean highlighting that the static IC-XT is baud rate independent. On the contrary, distinctive decreasing trends for the dynamic IC-XT can be observed in Fig. 3.4(b), when the baud rate rises from 10 to 80 GBaud, which follows the expectation of theoretical analysis in Section 3.2. Since an similar experimental demonstration on BPSK signals has been done in [176], the work in this figure with higher order I-Q modulated signals can be seen as an extending experimental validation of the theory in [176].

The dynamic IC-XT over a time window of 12 hours for all the investigated source signals, including CW, intensity modulated formats (i.e. OOK and PAM-4), phase modulated formats (i.e. m-QAM, m = 4, 16, 64, 256) and ASE, are summarized in Fig. 3.5(a), where all the signals were operated at 1550 nm wavelength. It can be seen that, since the wider the bandwidth of the signals, the

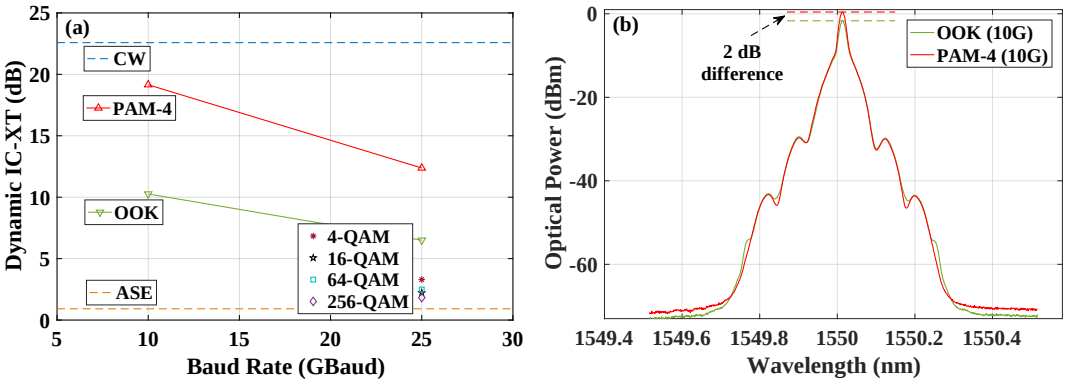


Figure 3.5: (a) Dynamic IC-XT for different signals over 12 hours and (b) Optical spectrum of OOK and PAM-4 signals (resolution 0.02 nm)

better the IC-XT being averaged and in turn the higher the stability [174], the observed dynamic IC-XT for the narrow band CW source and the broadband ASE source set the upper and lower bounds, separately. The intensity modulated formats, PAM-4 and OOK, perform the second and third highest variations, respectively. The reason behind is that these formats have strong optical carriers or high carrier-to-signal power ratios, which can result in strong interference at the PMPs. Therefore, since there are more non-zero intensity levels exist in PAM-4 format leading to a 30% increase (estimated) in the optical carrier-to-signal ratio against that of the OOK format, higher dynamic IC-XT was observed for PAM-4 signals over the baud rates of both 10 and 25 GBaud. Experimental validation on this analytical analysis has also been done by looking into the optical spectrum of the two types of signals. As shown in Fig. 3.5(b), when the baud rate was 10 GBaud, the PAM-4 signals performed 2 dB higher carrier-to-signal ratio. In terms of the phase modulated signals, as the optical carrier is completed removed, the observed 12-hours dynamic IC-XT for the QAM signals is quite low (i.e. less than 3.5 dB), among which 4-QAM and 256-QAM schemes achieve the highest and lowest dynamicity, respectively.

3.4.3 Temperature

According to the analysis in Section 3.2, temperature changes can affect the behaviour of IC-XT. In addition, calculations using Eqs. (2.3) and (3.1) show that even a slight change in the refractive index contrasts between the core and the cladding or cladding and trench can lead to considerable IC-XT increase. For example, a 5×10^{-5} variation in refractive index contrast between the core and the cladding can contribute to a $>$1 dB static IC-XT increase between the cores with 35 μm core pitch in the employed TA-MCF. To experimentally explore the relationship between them, Fig. 3.6(a) is presented, where an 1550 nm ASE source

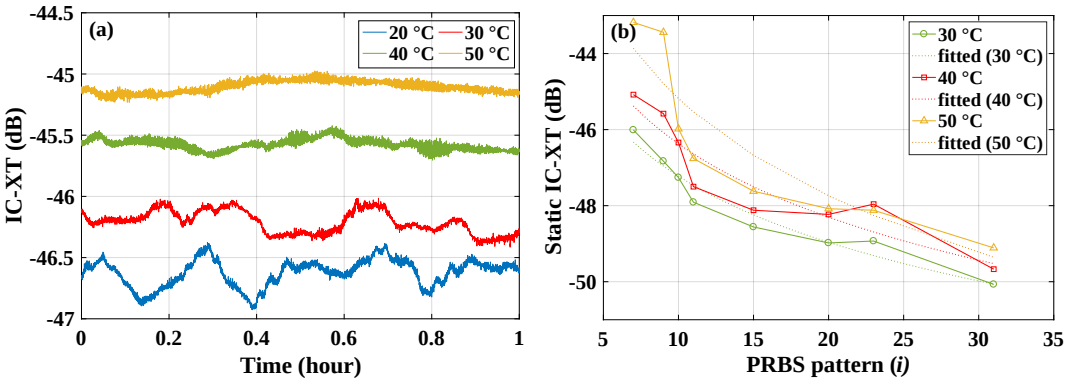


Figure 3.6: Effect of (a) temperature and (b) PRBS length on static IC-XT

was utilized as the source light since it can provide the most stable IC-XT. As seen, with the increase of the temperature, the IC-XT increases slowly. Particularly, a 30 °C temperature increase, from 20 to 50 °C, leads to a 1.5 dB static IC-XT rise, which can be translated into a static crosstalk-temperature coefficient of 0.05 dB/K. The reason for this dependence is that the refractive indexes of the core, cladding and trench change with the temperature increase [86, 182, 193], in the meanwhile, as the radius, original indexes and other properties of them are different, the degree of the effects of temperature on them are different [182], leading to the index contrasts change and in turn the IC-XT increase.

On the contrary, it is found that the 30 °C temperature increase leads to a 0.3 dB reduction on dynamic IC-XT (i.e. −0.01 dB/K), which can be explained by the fact that the inter-core skew increases with the temperature increase [183], contributing to better IC-XT stability as explained in Section 3.2. Apart from the static and dynamic IC-XT, the decorrelation time of the IC-XT was also evaluated based on the method in [136, 188]. For the four temperatures, the decorrelation times of the IC-XT varied in the range of 3.7 to 10.2 mins, however, no distinctive relation to the temperature was found.

3.4.3.1 PRBS length

The following Fig. 3.6(b) depicts the effects of both temperature and PRBS length on the static IC-XT, where the practical signalling source, 25 GBaud OOK operating at 1550 nm wavelength, was adopted. Each point on the figure is the 1-hour static IC-XT and the decorrelation times for them were ranging from 1.3 to 3.8 mins. As seen, the static IC-XT is inversely proportional to the PRBS length: when the temperature is 30 °C, the static IC-XT reduces from −46.0 to −50.1 dB, following the PRBS length increase from 2^7-1 to 2^{31}-1; under the temperatures of 40 °C and 50 °C, the static IC-XT decreases from −45.9 to −48.7 dB and from −43.2 to −49.1 dB, respectively. It is found that the observed correlation between the IC-XT

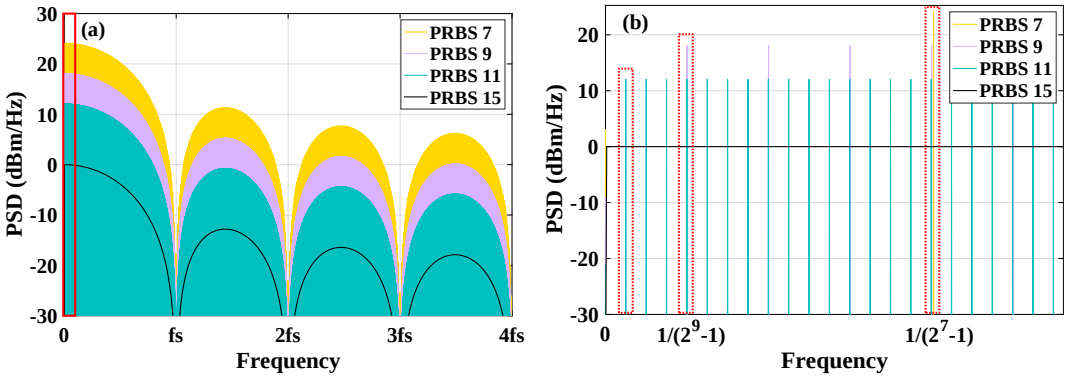


Figure 3.7: (a) PSD for signals with various PRBS pattern and (b) Zoomed-in figure of (a)

and PRBS pattern (i) can greatly fit the function, $IC\text{-}XT = a\log_2 i + b$, where the values of the coefficients, a and b, are -1.7 and -41.4, -1.9 and -40.0, -2.6 and -36.7, for the temperature of 30 °C, 40 °C and 50 °C, respectively. Attributing to the chronological order of the measurements that for each PRBS pattern the fiber was subjected to one heat cycle, the differences between the observed IC-XT and the fitting lines for each temperature have the same trend. For example, when PRBS pattern is 23, the observed IC-XT for each temperature is higher than the fitting one. On the contrary, the observed values for all the temperatures are lower than the fitting ones when the PRBS pattern is 11.

To explore the reasons for the relationship between the PRBS pattern and the static IC-XT, Fig 3.7(a) is presented, which shows the PSD of signals with different PRBS patterns. In this figure, the PRBS 15 is regarded as the benchmark, all the other patterns (PRBS 7, 9 and 11) were repeated to have the same sequence length with it. As seen, with the increase of the order of the PRBS pattern, the amplitude of the PSD envelope reduces. It can be explained by the zoomed-in figure (particularly for the part inside the red rectangle), Fig 3.7(b), that the number of the frequency components increases with the order increase, while the spacing between every two continuous components reduces [185]. To make sure the same total power for signals in all the patterns, the power of components for higher order patterns reduces accordingly. Therefore, the possible reasons for the correlation are: a) the randomness of the signals increases with the PRBS order in the transmission pattern, reducing the coherent interference and, b) the increase in the number of components and the reduction in the power of each component, altering the behaviour of the IC-XT.

3.4.3.2 Operational wavelength

The same signalling source, 25G-OOK, was also utilized to evaluate the impacts of temperature and operational wavelength on the static and dynamic IC-XT. It can be

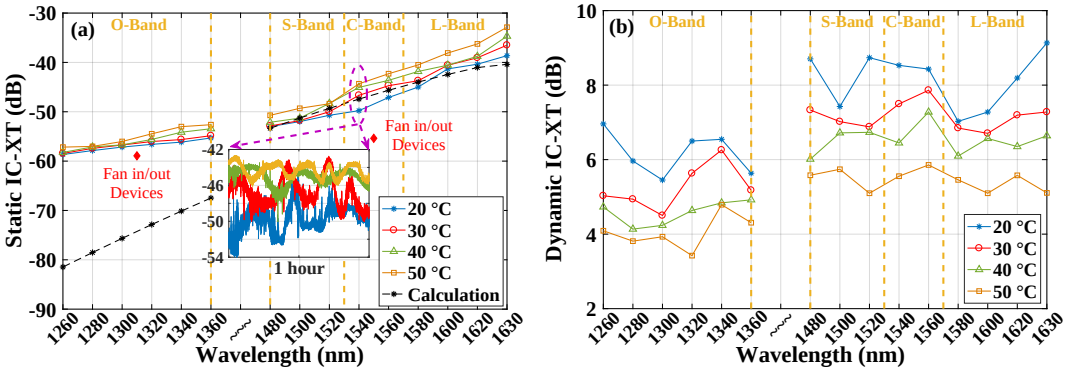


Figure 3.8: Impacts of temperature and wavelength on (a) static IC-XT and (b) dynamic IC-XT (25G-OOK)

seen in Fig. 3.8(a) that apart from the temperature, the static is also proportional to the wavelength. Particularly, when the wavelength increases in the O-band, from 1260 nm to 1360 nm, the static IC-XT performs an average increase of 6 dB for different temperatures. In contrast, when the wavelength goes up from the start of the S-band to the end of the L-band, the static IC-XT rises by 17 dB. Mathematically, the averaged crosstalk-wavelength dependence coefficient for O-band is 0.06 dB/K and for S-, C- and L-bands is 0.113 dB/nm. The reason for the wavelength dependence has been explained in Section 3.2 and the difference between the dependence coefficients in different bands can be explained by that the observed IC-XT in O-band was dominated by the crosstalk from the fan-in/out devices, while the observed IC-XT was mainly from the MCF for signals operating at S-, C- and L-bands wavelengths. It has also been proven by estimating the static IC-XT using Eq. (3.1), which is shown in the figure with the black dashed line. As seen, the estimated values can perfectly fit the observed IC-XT in S-, C- and L-bands, which is much bigger than the crosstalk from the fan/in out devices. However, the estimated values, from -81.5 to -67.5 dB, are much smaller than the crosstalk induced in the fan/out devices for O-band wavelengths, e.g. -58 dB at 1310 nm. It should be noted that, the slope of the indicator line for O-band is bigger than that of the S-, C-, L-bands, attributing to that the sensitivity of the coupling coefficient of the employed TA-MCF to the wavelength change varies in different bands.

From the inset in Fig. 3.8(a), it can be observed that the dynamic IC-XT also changes with the temperature increase. Moreover, compared to that observed in Fig. 3.6(a), the dynamic IC-XT for OOK signals are higher than that of the ASE scenarios. Therefore, to have a better understanding of the correlation between the dynamic IC-XT and wavelength as well as temperature, Fig. 3.8(b) is depicted. As seen, the dynamic IC-XT is relatively stable when the wavelength changes in S-, C-

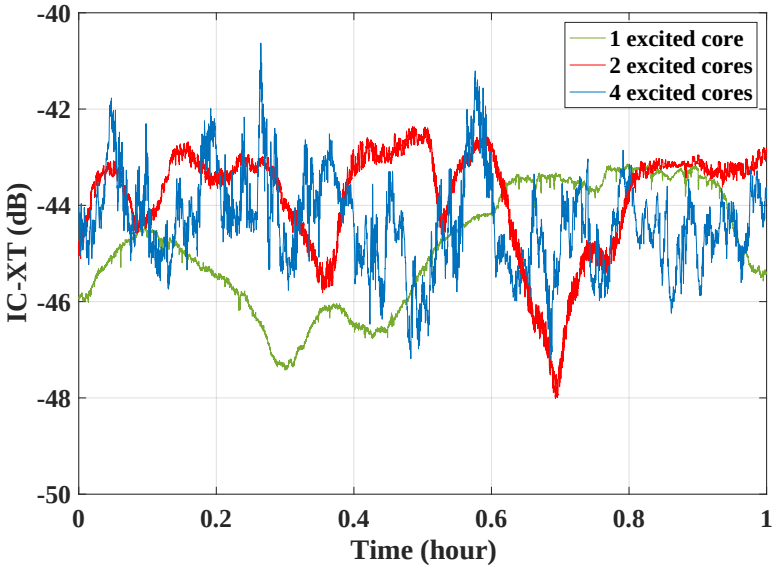


Figure 3.9: IC-XT over time for different numbers of excited cores

and L-bands, indicating the the dynamic IC-XT is wavelength independent. Note that, O-band cases are not discussed here since the IC-XT is dominantly from the fan/in out devices. In contrast, the dynamic IC-XT reduces with the temperature increase in the range of 20 to 50 °C, following the slope of -0.09 dB/K.

3.4.4 Number of excited cores

The previous studies focus on the behaviour of IC-XT between one excited core and one target core. In this subsection, behaviour of the total IC-XT from up to four exited cores on the target core is also investigated by using using a 1x4 splitter shown in Fig. 3.1. During the experimental process, core 5 was selected as the target core, where core 3 was the excited core for the 1 excited core scheme; cores 3 and 7 were the excited cores for the 2 excited cores scheme; cores 3, 6, 7 and 8 were the excited cores for the 4 excited cores scheme. Transparently, the total IC-XT increases with the number of excited cores. Moreover, it can be found that the speed of the IC-XT fluctuation, which can affect the design of crosstalk-tolerant adaptive techniques and indicates how often the service will be interrupted [136], also increases with the number of excited core increase, attributing to the increase in the total number of the PMPs. To evaluate the correlation between the fluctuation speed and the excited core number, the speed is characterized by counting the number of peaks and bottoms of the obtained IC-XT sequences. Results show that compared to the IC-XT fluctuation speed of the 1 excited core scheme, the 2 and 4 excited cores schemes offer 6 and 29 times higher speeds, respectively, indicating that the IC-XT fluctuation speed increases by a factor of 7.4 per core. Furthermore, by comparing the results in this work with

those in [183], it is found that although the employed MCF and experimental setup were different, similar correlation between the excited core number and the IC-XT fluctuation speed were observed. This is to say, the excited core number dominantly affects the IC-XT fluctuation speed. In addition, the decorrelation times of the three schemes are evaluated, which are 9.2 mins, 3.1 mins and 1 mins, for 1, 2 and 4 excited core schemes, respectively, validating the analytical prediction in [136] that the shorter the decorrelation time, the faster the IC-XT fluctuates.

3.5 Accuracy of Observed IC-XT

As mentioned in the Introduction, the behaviour of the observed IC-XT can not only be internally affected by the previous investigated parameters, but also be externally impacted by the experimental operations and setup, i.e. the time window of the measurement and the averaging time of the power meter or the interval time between the samples. In this section, the differences of the static IC-XT, the dynamic IC-XT and the worst-case IC-XT for a given time window and/or averaging time against those for the benchmarks (i.e. 12-hours time window and/or 25 ms averaging time), attributing to these parameters, are defined as inaccuracies. 25 ms averaging time is the benchmark since it is the shortest averaging time that can be achieved with the employed power meter, while the reason why choosing 12-hours observation time as the benchmark can be explained by Fig. 3.10. As seen in Fig. 3.10(a), the 12 hours' data shows a convergence value of 97% compared to that of the 30 hours' experimental data, which shows a steady behavior with oscillation of 0.3% in the last two hours (28-30 hours). According to the principle of convergence of uncertainties [194], anything above 90% convergence can be considered as an accurate representation of the parent population, therefore, the 12 hours' data is adequate for statistical analysis. The same analysis has been performed for the IC-

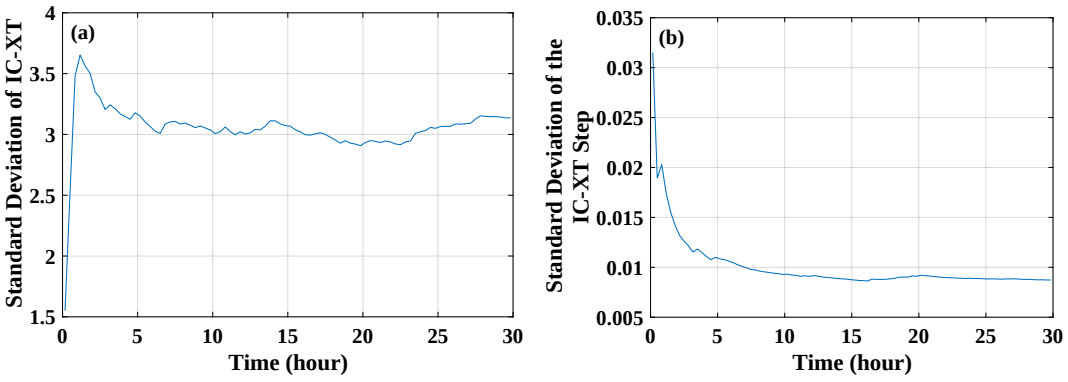


Figure 3.10: Standard deviation of (a) IC-XT and (b) IC-XT step for 30 hours' observation

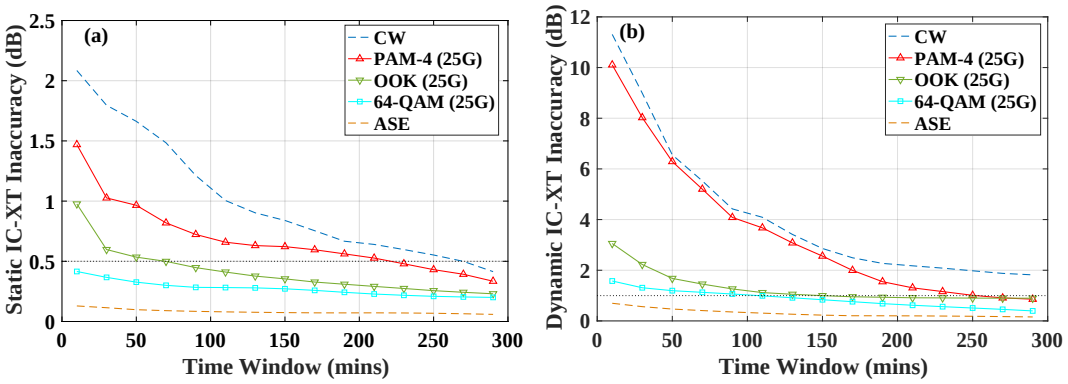


Figure 3.11: Effect of time window on (a) static IC-XT and (b) dynamic IC-XT for various signals

XT step (95% convergence compared to a 30 hours' experimental data), which is the difference between a pair of adjacent IC-XT samples at a discrete time interval and will be analyzed in subsection 3.6.

3.5.1 Time window

The influences of the observation time window on both the static and dynamic IC-XT for a variety of source signals are presented in Fig. 3.11. Obviously, inaccuracies of both the static and dynamic IC-XT reduce with the time window increase, among which the time window has higher degree of influence on the dynamic IC-XT than that on the static IC-XT. For instance, when the time window is 10 mins, the inaccuracy on the static IC-XT is 2.1 dB while it can be up to 11.3 dB on the dynamic IC-XT. Furthermore, it can also be observed from both the figures that the degree of the time window effect also depends on the type of the signalling source, following the trend of that the more stable the IC-XT induced by the source, the lower the degree of the time window effect on it. For instance, the IC-XT induced by the broadband ASE signals, which has the highest stability, achieves a static IC-XT inaccuracy of merely 0.13 dB and a dynamic IC-XT inaccuracy of 0.68 dB when the time window is 10 mins. On the contrary, even when the time window is extended to 300 mins, static and dynamic IC-XT inaccuracies for the narrow band CW source compared to the 12-hours benchmark are still 0.4 dB and 2 dB, respectively.

3.5.2 Averaging time

Figure. 3.12 shows the effects of the averaging time of the power meter on the accuracy of the observed dynamic IC-XT and the worst-case IC-XT, which can determine the power margin of the transmission system. Static IC-XT is not taken into consideration in this figure because no distinct changes had been observed

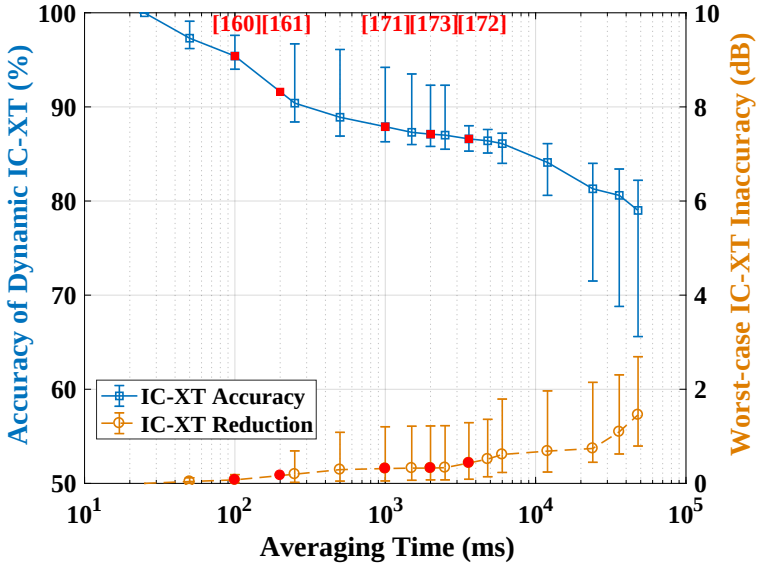


Figure 3.12: Effect of averaging time on dynamic IC-XT and the worst-case IC-XT

even when the averaging time changed from 25 ms to 48 s. In addition, since the dynamic IC-XT accuracy and the worst-case IC-XT inaccuracy vary with the type of the singalling source change, the results are shown as error bars. As seen, with the increase of the averaging time, the accuracy of the dynamic IC-XT reduces and the inaccuracy of the worst-case IC-XT increases accordingly. The figure indicates that to achieve 95%, 90% and 80% dynamic IC-XT accuracy, the averaging time of the power meter should be shorter than 100 ms, 250 ms and 36 s, respectively. Moreover, the averaging time should be shorter than 250 ms, 4.8 s and 36 s, when the worst-case IC-XT inaccuracies of 0.2 dB, 0.5 dB and 1 dB are required, respectively. Compared with the work demonstrated in the existing researches considering various averaging times, e.g. 100 ms [175], 200 ms [176], <1 s [186], 2 s [188] and 3.6 s [187], 5% - 13% higher accuracy on the dynamic IC-XT and 0.1 to 0.33 dB reductions on the worst-case IC-XT inaccuracy have been achieved in this work.

3.6 Distribution of IC-XT Step

The study on IC-XT accuracy can be a reference for IC-XT experiments in the lab. In real world applications, if the IC-XT can be predicted or the source signals can be classified based on the observed IC-XT using machine learning methods, the efficiency and security of the MCF-based systems will be significantly improved. Therefore, the possibility or feasibility of IC-XT prediction and source signal classification are investigated. First af all, the circular correlation, which is a computationally efficient method to measure auto-correlation of a sequence [195],

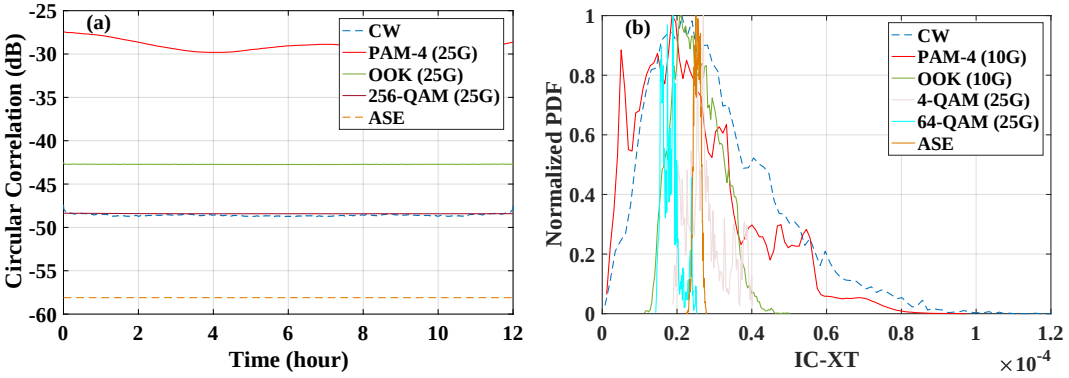


Figure 3.13: (a) Circular correlation and (b) PDF of IC-XT for various source signals

of the observed IC-XT sequences for different types of source lights have been investigated and the results are shown in Fig. 3.13(a). Unfortunately, it is found that the circular correlations of all the IC-XT sequences are negligible, i.e. less than -27 dB or 0.002, indicating that the the observed IC-XT sequence is completely random which is not suitable for IC-XT prediction or source signal classification.

However, the statistical analysis on the IC-XT distribution provides an opportunity for machine learning methods based source signal classification. According to the research in [174], the Chi-square distribution can be used to fit the IC-XT distribution and recent studies in [186] and [175] have experimentally validated the accuracy of this model by using CW and OOK sources, respectively. To check if this model can fit the distribution of IC-XT induced by other kinds of source signals, Fig. 3.13(a) is provided, which presents the normalized probability density function (PDF) of the 12-hours IC-XT for CW, PAM-4, OOK, 4-QAM, 6-QAM and ASE signals. It can be seen that, only the distribution of the IC-XT induced by CW signals can greatly fit a Chi-square distribution, however, for the IC-XT induced by other sources, the fitness to the distribution reduces accordingly. Especially, for the IC-XT induced by QAM signals, almost no existing models can be used for fitting. What is worse, it is found that even for the IC-XT induced by CW signals, a relative good distribution fitting to the Chi-square distribution can only be achieved when the time window is larger than 10 hours. With the decrease of the time window, the fitness reduces rapidly, which indicates that the IC-XT distribution is also not a good choice for source signal classification.

To find a better approach, the IC-XT is considered as a Pseudo-random walk process based on the following reasons: a) IC-XT sequence is a completely random process, b) the analysis in [188] of the auto-correlation extremely resembles the one of a random-walk (cumulative sum of stochastic random

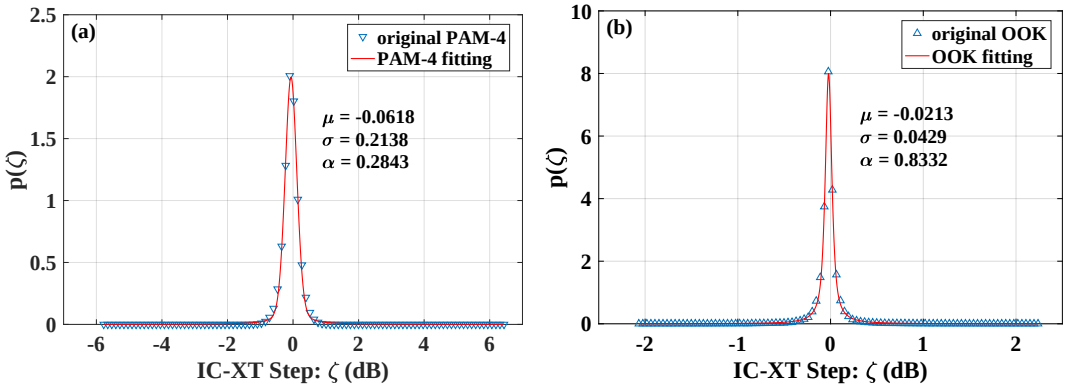


Figure 3.14: PDF of IC-XT step for a) PAM-4 (25G), b) OOK (25G)

variables), for which elements closer to each other in time will be more similar and therefore more correlated, and c) both the IC-XT and IC-XT step standard deviations are found to converge to a steady numerical value when the number of samples increases. The model can be expressed as:

$$S_{t+1} = S_t + \zeta \backsim p(\zeta|S_t) \tag{3.5}$$

In which, S_t stands for the IC-XT of a certain time and ζ represents the IC-XT step, which follows a particular distribution depending on the value of S_t. This process simplify the analysis to a single stochastic random variable ζ. For this reason, I model the PDF of ζ, and it is found that it can be expressed as:

$$p(\zeta) = \int_{-\infty}^{\infty} p(\zeta|S_t)p(S_t)\,dS_t \backsim PVP(\zeta;\mu,\sigma,\alpha) \tag{3.6}$$

In the equation, PVP represents the Pseudo-Voigt profile, which is a numerical approximation of a Voigt profile, therefore a convolution between a Cauchy-Lorentz distribution and a Gaussian distribution, which can be expressed as:

$$PVP(\zeta;\mu,\sigma,\alpha) = \frac{(1-\alpha)}{\sigma_g\sqrt{2\pi}} e^{\left[-(x-\mu)^2/2{\sigma_g}^2\right]} + \frac{\alpha}{\pi}\left[\frac{\sigma}{(x-\mu)^2+\sigma^2}\right] \tag{3.7}$$

Where μ and σ are the mean and standard deviation, respectively. $\sigma_g = \sigma/\sqrt{2\ln 2}$. The first part of the equation relates to the Gaussian distribution and the second to the Cauchy-Lorentz distribution, α is the scaling coefficient between the two distributions ($0 < \alpha < 1$).

To check the accuracy of this model, Fig. 3.14 is depicted, showing the

Table 3.1: Fitting coefficients and performance

Signalling Source	μ	σ	α	Fitting Accuracy
CW	-0.0880	0.3712	0.8396	99.56%
OOK	-0.0213	0.0429	0.8332	99.85%
PAM-4	-0.0618	0.2138	0.2843	99.74%
256-QAM	0.3675	0.0044	-0.0016	99.87%
ASE	-0.0053	0.0444	0.1998	99.33%

distribution of the observed IC-XT steps and the fitting results using this model for PAM-4 and OOK signals, respectively. As seen, both the two schemes achieve perfect fitting performance. Moreover, as shown in Table 3.1, the distribution of the observed IC-XT steps for the other signalling sources can also perfectly fit this model, mathematically, all the scenarios can achieve $> 99.3\%$ fitting accuracy. It can be seen that for each signalling source, the fitting model has unique coefficients, therefore, the source signal classification is potentially enabled.

As aforementioned, the fitness of the IC-XT distribution to the Chi-square distribution can reduce rapidly when the time window decreases. To explore the effect of the time window and averaging time on the fitting accuracy of the IC-XT step distribution to this model, Fig. 3.15 is shown, where the distribution accuracy is characterized as the similarity of the distribution of current IC-XT step samples to that of the benchmark samples (12-hours time window or 25 ms averaging time), which is evaluated through a R^2 score function [196]. It can be seen that, both the time window and averaging time have relative slight effects on the IC-XT step distribution accuracy. For instance in Fig. 3.15(a), even with 20 mins observation time, over 90% IC-XT step distribution accuracy is obtained compared to that of the 12-hours benchmark. Moreover, when the time window is extended to over 80 mins, the IC-XT step distribution accuracy can increase to 95%, which definitely satisfies the requirement for source signal classification purpose. According to the results shown in Fig. 3.15(b), the averaging times just need to be shorter than 8.5 s and 4 s, which can be easily realized with the commonly used power meters in the labs, in order to obtain 90% and 95% IC-XT step distribution accuracy, respectively. Compared with the approach of using IC-XT distribution (Chi-square) for signal source classification, adopting the IC-XT step distribution requires much shorter observation time window and/or less strict averaging time while providing good accuracy. The IC-XT induced by CW source can serve as an example, a time window of 50 minutes is sufficient to achieve over 90% IC-XT step distribution accuracy when the averaging time is 3.5 s. If the averaging time can be shorter than 0.625 s, IC-XT just need to be observed for 10 mins.

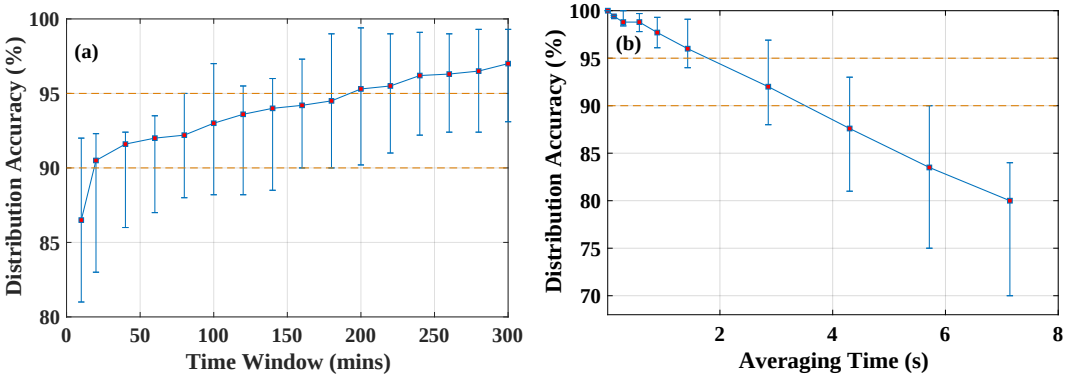


Figure 3.15: Effects of a) time window and b) averaging time on IC-XT step distribution

Furthermore, unlike the IC-XT distribution model, this model for IC-XT step distribution is suitable for all the investigated sources, including the IC-XT induced by ASE and QAM signals, showcasing that low complexity power monitors and short windows can be used to create a IC-XT step distribution that can be used to identify the signal propagated.

3.7 Summary and Conclusion

This chapter first provides a comprehensive study on the behaviour of the static and dynamic IC-XT, by conducting a mount of experiments with a TA-MCF fabricated for data center usage, taking the effects of lots of transmission parameters into account, including modulation format, baud rate, temperature, PRBS length, operating wavelength, and the number of excited cores. The correlations between the IC-XT and these parameters are characterized and summarized in Table 3.2. On the perspective of the static IC-XT, it is proportional to both the temperature and the wavelength while inversely proportional to the PRBS length. In the meanwhile, the dynamic IC-XT is inversely proportional to the baud rate, temperature whilst independent on the wavelength. These investigations signify the importance of temperature, PRBS length and modulation format on the IC-XT investigations. In terms of the modulation format, lower IC-XT dynamicity can be achieved when I-Q modulation signals are adopted compared to that of the intensity modulated signals. The IC-XT fluctuation speed increases with the number of excited cores increase. Furthermore, the experimental results relate to the dependence of static IC-XT on wavelength and the correlation between the IC-XT fluctuation speed and the decorrelation time can greatly fit the theoretical estimations and analysis in the existing research, significantly validating the accuracy of the models. Secondly, this chapter explores the impacts of the time window and the averaging time on the accuracy of the

Table 3.2: IC-XT dependence on the investigated parameters and static/dynamic IC-XT for various signalling sources (1550 nm, 25 GBaud)

Coefficient	**Temperature**		**Wavelength**	**Baud Rate** (x)	**PRBS Length** ($2^i - 1$)		
Wavelength (nm)	1550		1480-1630	1550	1550		
Signalling Source	ASE	OOK (25G)	OOK (25G)	OOK/PAM -4/m-QAM	OOK (25G)		
Temperature (°C)	20-50	20-50	20-50	23	30	40	50
Static IC-XT	0.05 dB/K	0.13 dB/K	0.113 dB/nm	independent	-1.7$log_2 i$ -41.4	-1.9$log_2 i$ -40	-2.6$log_2 i$ -36.7
Dynamic IC-XT	-0.01 dB/K	-0.09 dB/K	independent	QAM: $1.43 * 0.98^x$	-	-	-
Excited Cores	fluctuation speed increases by a factor of 7.4 per core						

Signalling Source	**CW**	**ASE**	**OOK**	**PAM-4**	**4-QAM**	**16-QAM**	**64-QAM**	**256-QAM**
Static IC-XT (dB)	-46.07	-45.95	-45.92	-46.9	-45.63	-46.84	-47.43	-48.61
Dynamic IC-XT (dB)	22.58	0.92	6.50	12.38	3.30	2.22	2.50	1.80

observed IC-XT, which can significantly benefit the studies on MCFs in the labs and in practical applications. At last the novel study on the IC-XT step distribution can serve future ML-based IC-XT classification. I expect that the comprehensive analysis, sophisticated understanding and accurate measurements of IC-XT levels demonstrated in this chapter can significantly benefit the design and scaling of future MCF-based data centers, metro networks or telecommunication systems.

Chapter 4

SDM-based Data Center Networking

4.1 Introduction

It has been mentioned in Chapter 1 that attributing to the increasing adoption of cloud services, video services and associated machine learning applications, the traffic demand inside data centers is increasing exponentially, which necessitates an innovated networking infrastructure with high scalability and cost-efficiency. However, the conventional solutions using SMF-based WDM over C+L-band may fall short in satisfying the requirements of high-performance DCNs, including network capacity, cabling complexity, spatial efficiency, cost and power consumption [42]. Therefore, as mentioned in Section 2.4.2, to address these challenges SDM solutions have been proposed and they are regarded as the first potential candidates for DCNs. Among which, the application of homogeneous MCFs is considered as the most feasible and efficient approach to realize SDM-based short reach DCNs, attributing to that the issue of IC-XT is not severe over short link spans ($<$ 1km) compared to that in long-haul transmission. In recent years, researches have shown that MCF-based SDM technique can lead to hardware, cost and energy savings through sharing transceiver digital signal processing (DSP) [197] and it is compatible to the cost and power efficient integrated technologies, e.g. vertical cavity surface emitting laser (VCSEL) array [198]. Moreover, it has been shown that silicon photonic on-board transceivers coupled on MCF can be used to support MCF transmission without requiring any fan-in/out or core pitch conversion devices, which can increase the front panel density while offering better panel space management [190]. In another research, the showcased beam steering MCF switch can route signals from multiple cores together to support purely MCF-based DCN links with SDM switching [199]. In addition, different topologies have been investigated in [42] for the purely SDM-based DCN using MCFs. Compared to the WDM-based DCN using SMFs, it provides much higher power and cost efficiency.

However, such networks are still vulnerable to the IC-XT in the MCF even over short distances. As a consequence, the IC-XT can limit the transmission reach of network and in turn the DCN' size. To alleviate this issue in conventional uni-directional (1di) long-haul networks, various routing and spectrum allocation (RSA) algorithms, e.g. the algorithms described in Section 2.5, have been proposed and developed to ensure the quality of the signal though allocating the resources in a proper manner [155]. Moreover, recent research has shown that IC-XT can be further suppressed by transmitting optical signals in opposite directions in the neighbouring cores over the MCF link [200], in particular, this significant IC-XT reduction between a core pair can be up to 20 dB. However, the benefits of applying the bi-directional (2di) transmission to DCN have not been explored so far. Before its application, new analytical models for the IC-XT in bi-directional MCF should be derived, moreover, corresponding RSA algorithms should be proposed. Furthermore, since in short-reach DCNs, IC-XT is not severe as that in standard backbone networks, the IC-XT reduction attributes to the bi-directional transmission may enable the MCF to have more cores with smaller core pitches, which can mitigate the cabling complexity and increase the spatial efficiency of the transmission systems.

This chapter demonstrates a high-capacity and highly spatial efficient optical DCN solution using the MCF-based SDM techniques and exploits the benefits contributed by the bi-directional MCF in DCN. At first, based on the Eq. 2.4, new IC-XT formulations for bi-directional MCFs are developed, considering different fiber profiles, i.e. step-index and trench-assisted, and layouts, i.e. hexagonal and rectangular. To reduce IC-XT and alleviate the associated computational complexity, new IC-XT aware RSA algorithms are also proposed based on the algorithms explained in Section 2.5 for the bi-directional transmission. The performance of these algorithms are then evaluated in DCNs with different topologies, i.e. spine-leaf topology, Facebook topology, and three-tier fat tree topology, in terms of blocking probability, network utilization, fragmentation and computational time. In the simulation process, two types of requests are generated and three types of multiplexing scenarios, i.e. pure WDM, pure SDM and SDM-WDM are considered for analyzing the network capacity and link spatial efficiency of the DCN. It was claimed in [201] that IC-XT could be ignored in intra data center interconnects, however, it is found that IC-XT can limit the transmission reach and the core number inside the fiber. Moreover, the limitation on the transmission distance varies with the fiber type change. Therefore, a variety of homogeneous MCFs are investigated, including hexagonal MCFs with 7, 19, 37, 61 cores and rectangular MCFs with 8, 12, 30, 52 cores. Furthermore, the network

performance of them in terms of network capacity and link spatial efficiency is contrasted and compared. To have an more comprehensive DCN investigation considering practical transmission requirements, IC-XT thresholds required by different real modulation formats, e.g. OOK and PAM-4, are taken into consideration.

4.2 New IC-XT Formulations for Bi-directional Homogeneous MCFs

In this section, a variety of new equations are formulated for different bi-directional homogeneous MCFs, including SI-MCF and TA-MCF with equal or unequal core pitches. In addition, the IC-XT dependence on wavelength is also considered.

4.2.1 SI-MCF with equal core pitch

Homogeneous MCF with a hexagonal layout or triangle lattice has been widely used in research experimentation and trials. As shown in Fig. 4.1, the core pitches (C_P) between any two adjacent cores in this kind of fibers are identical. In order to model the IC-XT reduction ($\Delta IC\text{-}XT_{-dB}$) contributed by the bi-directionality, a power reduction coefficient P_r is introduced to denote the the IC-XT suppression between the adjacent cores, which can be calculated by:

$$P_r = 10^{(-\frac{\Delta IC\text{-}XT_{-dB}}{10})} \tag{4.1}$$

Also, according to [200], P_r can be modelled as:

$$P_r = \frac{S\alpha_R}{2\alpha}[\frac{e^{\alpha L} - e^{-\alpha L}}{\alpha L} - 2e^{-\alpha L}] \tag{4.2}$$

In which, S is the recapture factor of the Rayleigh scattering component into the backward direction. α_R and α denote the attenuation coefficient results from Rayleigh scattering, and the fiber attenuation coefficient, respectively.

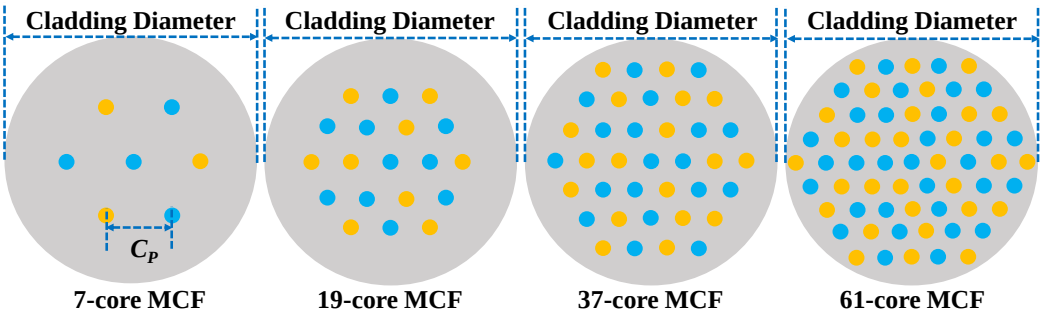


Figure 4.1: Homogeneous MCFs with hexagonal layout

As aforementioned in Section 3.2, the operational wavelengths of signals in two neighbouring cores can also influence the IC-XT between them and the correlation is that the longer the wavelength, the higher the IC-XT. Based on [147], the additional wavelength-dependent IC-XT ($\Delta IC\text{-}XT_{dB}$) attributing to the wavelength increases on the lower end of the C-band (λ_0 = 1530 nm) can be expressed as:

$$\Delta IC\text{-}XT_{dB}(C_P) = 10log_{10}(1-0.001256\Delta\lambda)^4 + 19.85\pi n_1\sqrt{2\Delta_1}\frac{\Delta\lambda C_P}{\lambda\lambda_0} \tag{4.3}$$

Where $\Delta\lambda$ is the difference between the transmission wavelength (λ) and λ_0. n_1 denotes the refractive coefficient of cladding and Δ_1 represents the refractive coefficient difference between the core and the cladding, which have been described in Section 2.4.2. This IC-XT increase can be transformed into a power increase coefficient P_i with the Eq. (4.4). Furthermore, by considering the wavelength dependence, the power reduction coefficient P_r is modified to $P_r{}'$, which can be expressed as:

$$P_i(C_P) = 10^{(\frac{\Delta IC\text{-}XT_{dB}}{10})} \tag{4.4}$$

$$P_r{}'(C_P) = 10^{(\frac{\Delta IC\text{-}XT_{dB}-\Delta IC\text{-}XT_{-dB}}{10})} \tag{4.5}$$

In a bi-directional MCF, as there are more than one cores assigned to carry signals in each direction, generalized equations derived based on Eqs. (2.4), (4.4), (4.5) are necessary to model the total IC-XT on the target core from the neighbouring cores carrying signals in the same [(Eq. 4.6)] and the opposite directions [Eq. (4.7)]:

$$IC\text{-}XT_{hex\text{-}SI\text{-}same} = \frac{P_i n_{c1}[1-e^{-(n_c+1)hL}]}{1+P_i n_c e^{-(n_c+1)hL}} \tag{4.6}$$

$$IC\text{-}XT_{hex\text{-}SI\text{-}oppo} = \frac{P_r{}' n_{c2}[1-e^{-(n_c+1)hL}]}{1+P_i n_c e^{-(n_c+1)hL}} \tag{4.7}$$

In which, n_{c1} and n_{c2} denote the number of neighbouring cores carrying data in the same and the opposite directions with the target core, respectively. The total number of adjacent cores for the target core (n_c) equals to the sum of n_{c1} and n_{c2}. Therefore, the total IC-XT of a target core in a bi-directional hexagonal SI-MCF can be calculated by:

$$IC\text{-}XT_{hex\text{-}SI} = IC\text{-}XT_{hex\text{-}SI\text{-}same} + IC\text{-}XT_{hex\text{-}SI\text{-}oppo} = \frac{(P_i n_{c1} + P_r{}' n_{c2})[1 - e^{-(n_c+1)hL}]}{1 + P_i n_c e^{-(n_c+1)hL}} \tag{4.8}$$

4.2.2 TA-MCF with equal core pitch

As described in Section 2.4.2, attributing to the introduction of the low-index trench, TA-MCF can facilitate lower IC-XT than SI-MCF. Moreover, the correlation between the mean IC-XT per meter of the TA-MCF (h') and that of the SI-MCF (h) can be expressed as:

$$h'(C_P) = h\frac{W_1}{[W_1 + (W_2 - W_1)\frac{w_t}{C_P}]}e^{-2(W_2-W_1)\frac{w_t}{a}} \tag{4.9}$$

Where the coefficients have been described in Eqs. (2.2), (2.3). By replacing h with h' in Eq. (4.8), the IC-XT model for bi-directional hexagonal TA-MCF can be derived and expressed as:

$$IC\text{-}XT_{hex\text{-}TA} = \frac{(P_i n_{c1} + P_r{}' n_{c2})[1 - e^{-(n_c+1)h'(C_P)L}]}{1 + P_i n_c e^{-(n_c+1)h'(C_P)L}} \tag{4.10}$$

4.2.3 SI/TA-MCF with unequal core pitch

Apart from the MCFs with hexagonal layout, IC-XT models for rectangular MCFs are also derived. As seen in Fig. 4.2, the neighbouring cores are arranged with square lattice and thus, more than one core pitches between the core pairs exist. Based on the previous equations for hexagonal MCFs, in the rectangular SI-MCF the overall IC-XT for a target core from the adjacent cores in the transmission and reception directions can be expressed as Eqs. (4.11) and (4.12), respectively.

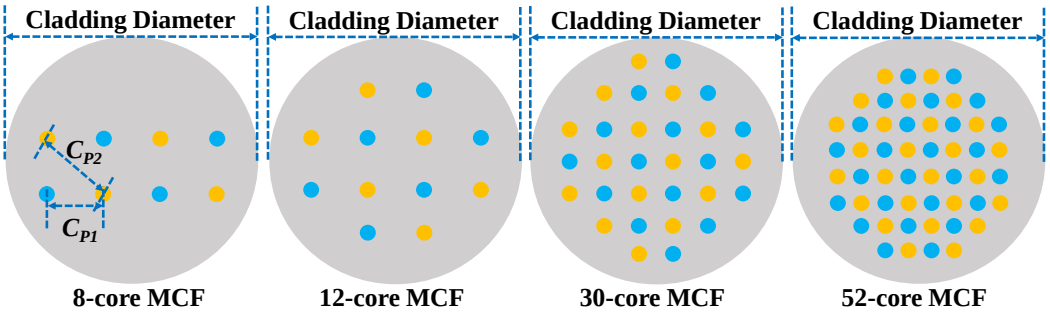


Figure 4.2: Homogeneous MCFs with rectangular layout

$$IC\text{-}XT_{rec\text{-}SI\text{-}same} = \frac{P_i(C_{P1})n_{c3}[1-e^{-(n_c+1)h(C_{P1})L}]+P_i(C_{P2})n_{c4}[1-e^{-(n_c+1)h(C_{P2})L}]}{1+P_i(C_{P1})(n_{c3}+n_{c5})e^{-(n_c+1)h(C_{P1})L}]+P_i(C_{P2})(n_{c4}+n_{c6})e^{-(n_c+1)h(C_{P2})L}]} \tag{4.11}$$

$$IC\text{-}XT_{rec\text{-}SI\text{-}oppo} = \frac{P'_i(C_{P1})n_{c5}[1-e^{-(n_c+1)h(C_{P1})L}]+P'_i(C_{P2})n_{c6}[1-e^{-(n_c+1)h(C_{P2})L}]}{1+P_i(C_{P1})(n_{c3}+n_{c5})e^{-(n_c+1)h(C_{P1})L}]+P_i(C_{P2})(n_{c4}+n_{c6})e^{-(n_c+1)h(C_{P2})L}]} \tag{4.12}$$

Where n_{c3} and n_{c4} stand for the quantities of the adjacent cores in the same direction with the target core spaced apart the target core by C_{P1} and C_{P2}, separately. In contrast, n_{c5}, n_{c6} represents the numbers of adjacent cores in the opposite direction with the target spaced apart the target core by C_{P1} and C_{P2}, respectively. Therefore, $n_c = n_{c3}+n_{c4}+n_{c5}+n_{c6}$. It can be seen in Fig. 4.2 that, for the fibers investigated in this work, the core pitch between the target core and the neighbouring cores in the opposite directions is always equal to C_{P1}, while for the adjacent cores in the same direction the core pitch is equal to C_{P2}. That is to say, n_{c3} and n_{c6} equal to zero. Finally, the total IC-XT for a given target core in a rectangular SI-MCF can be calculated by:

$$IC\text{-}XT_{rec\text{-}SI} = IC\text{-}XT_{rec\text{-}SI\text{-}same} + IC\text{-}XT_{rec\text{-}SI\text{-}oppo} \tag{4.13}$$

Note that, the IC-XT models for rectangular TA-MCFs can be easily obtained by changing h in the above two equations with h':

$$IC\text{-}XT_{rec\text{-}TA} = IC\text{-}XT_{rec\text{-}SI}(h = h') \tag{4.14}$$

In this chapter, Eqs. (4.8), (4.10) and (4.13) will be adopted to calculate the IC-XT for the links of the considered networks using the investigated MCFs. The total IC-XT over a path consists of several links connecting the source and destination nodes in DCN can be calculated by:

$$IC\text{-}XT_{path} = \sum_{i=1}^{LN} IC\text{-}XT_{link\ i}, \quad LN : link\ number \tag{4.15}$$

4.3 IC-XT Aware Allocation Algorithms

Several new IC-XT aware allocation algorithms for the bi-directional transmission, which can either reduce the IC-XT or speed up the allocation process, are

presented in this section, including a bi-directional core prioritization strategy and two spectrum-splitting approaches.

4.3.1 Bi-directional core priority mapping

The bi-directional core priority mapping is proposed to mitigate the IC-XT by pre-defining the sequence of the core usage, which avoids assigning contiguous blocks of adjacent cores for transmission in the same direction. During the process, to keep the fairness, the number of cores for each direction should be the same. In this work, two strategies have been proposed for the mapping in a bi-directional link model with two fibers per link. The first one is shown in Fig. 4.3, where the mapping starts from two cores in a single MCF, and it is denoted as *start 1*. In contrast, the other one, denoted as *start 2* and shown in Fig. 4.4, starts the mapping from two cores in two MCFs. An example of the whole process of core priority mapping inside a pair of MCFs is illustrated in Fig. 4.3, where the signals in green cores propagate in one direction and those in orange cores propagate in the opposite direction. The propagation direction of each core is pre-assigned to guarantee that signals in an adjacent core pair propagate in opposite directions. During the mapping process, the priorities of the cores in each direction are assigned independently.

As seen, at the beginning of the mapping (step 0), each core in the MCFs is assumed to have a core cost C_i, whose value is initialized to 0, and in which, i is the core index. In the following step, the algorithm will choose the cores with the lowest core costs to be the next cores in the priority sequences for each direction. In the meanwhile, the costs of the adjacent cores of the chosen cores will increase by 1. For instance, at the step 1 in Fig. 4.3, cores with indexes 4 and 7 are selected

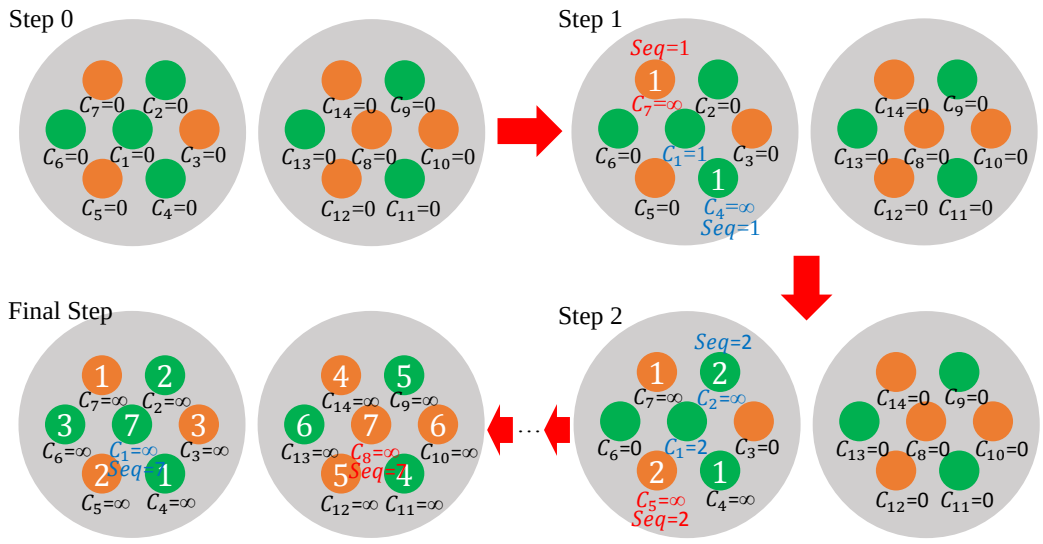


Figure 4.3: Core priority mapping starts from one MCF (*start 1*)

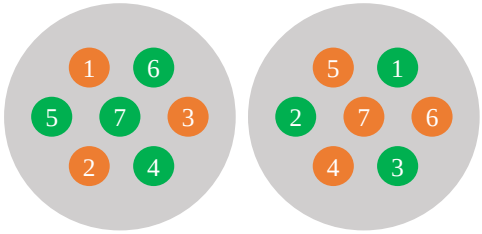


Figure 4.4: Core priority mapping starts from two MCFs (*start 2*)

as the first cores ($Seq = 1$) for each direction, respectively. In the same time, their costs change to infinity (C_4, $C_7 = \infty$) to avoid being assigned again and the cost of the core with index 1 increases to 1 ($C_1 = 1$). Since the core with index 1 is only in the same direction with the core with index 4, the cost only increases by 1 rather than 2. Subsequently, in the step 2, cores with indexes 2 and 5 are selected to be used in their directions ($Seq = 2$) since they have the lowest costs. Simultaneously, there costs are set to infinity and the cost of the core with index 1 increases again ($C_1 = 2$). If the two cores in different MCFs have the same cost, the core located in the same fiber as the core with higher priority will be chosen first. Only when the cost of a core in the second MCF becomes the lowest or if all cores in the first fiber are processed, the algorithm will start defining the cores in the second fiber. Following this rule and step by step, all the cores in both directions will be defined with sequence numbers, and the complete map for the *start1* scheme is depicted in the final step of Fig. 4.3.

Unlike the *start 1* scheme that the cores with the highest priority in each direction are in the same MCF, in the *start 2* scheme shown in Fig. 4.3, transmissions in the two directions start from the cores in different MCFs. Comparing the two schemes in practice, the *start 1* strategy can offer better modularity and scalability since the network can be extended by simply adding new fibers. In contrast, the *start 2* strategy has the potential to provide better network performance, as there is completely no IC-XT when traffic load is low.

4.3.2 Spectrum splitting scheme

In order to realize a low IC-XT transmission, the RSA algorithms should check the IC-XT for every new request based on the aforementioned equations. As a result, the computational complexity and time may increase rapidly. To alleviate this issue, the spectrum splitting scheme is proposed, which divides the whole spectrum (e.g, C-band and C+L-band, since extensive research and development related to WDM have been done in these bands) into two bands. With this scheme, only half of the spectral resources will be searched to find the available slots at one time, which can significantly reduce the computational time. Moreover, to further

Table 4.1: Pre-defined spectrum division for bi-directional transmission

Defined Sequence (One Direction)	First Division for Request Allocation	Defined Sequence (Opposite Direction)	First Division for Request Allocation
Core 1	D1	Core 1	D2
Core 2	D1	Core 2	D2
...	...	...	...
Core V	D1	Core V	D2
Core V+1	D2	Core V+1	D1
...	...	...	...
Core W	D2	Core W	D1

mitigate IC-XT, the spectral bands for the neighbouring cores are arranged in a non-overlapping fashion. This novel scheme is fully compatible with the bi-directional transmission and based on the splitting strategy, it can be categorized as soft spectrum splitting and hard spectrum splitting. The main procedures of the resource allocation process for the SDM-WDM scenario considering soft spectrum splitting, core priority mapping and core switching approaches are as follows:

1. Divide the whole spectrum with 100 frequency slots into two divisions in each core of the MCFs, among which D1 is used to represent the first division of 50 frequency slots and D2 is the second division of the remaining slots.

2. Based on the given core priority map and follow the defining rule shown in Table 4.1, define the initial allocation division for each core insides the MCFs in both directions. As presented in Table 4.1, W equals to the total number of cores in each direction, Core W means that the core is the W *th* core (Seq = W) for usage pre-defined using the core priority mapping scheme. According to the core priority map, Core 1 to Core V are not adjacent in each direction and thus, they will be allocated first. On the contrary, Core (V+1) to Core W are the neighbouring cores of previous allocated cores, therefore, they will be used later. An example of the final map for a 19-core MCF considering core priority and first division for request allocation is depicted in Fig. 4.5, where W = 19 and V = 13. The number inside each core is the core usage sequence, where spectrum division D1 is used first for the cores with black digits and the cores the white ones use D2 first.

3. As shown in Fig. 4.6, when a new request comes, the algorithm will first check the pre-defined first allocation division (i.e. D1 or D2) of the first core slot by slot. If sufficient available frequency slots are found, IC-XT will be

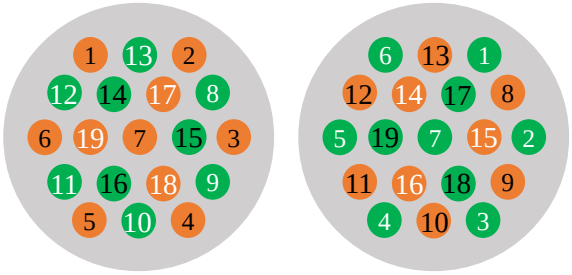


Figure 4.5: Core priority map with defined spectrum division for 19-core MCF

checked. If there are no enough frequency slots in the current division of the current core, the next core in the priority map will be checked. The same process will repeat until available and low IC-XT slots are found or all the slots in the pre-defined first allocation divisions of all the cores are checked. In the figure, the numbers inside the slots are the indexes of the frequency slots. The green frequency slots carry signals in one direction, the orange ones in the opposite direction, while the slots in white are unused slots.

4. If there are no sufficient frequency slots in the pre-defined first allocation divisions of all the cores in both directions, the two divisions in all the core will be swapped. This is to say, in the previous Fig. 4.6, the slots in D2 of the orange Core 1 to Core V can be scanned to check if they are suitable for

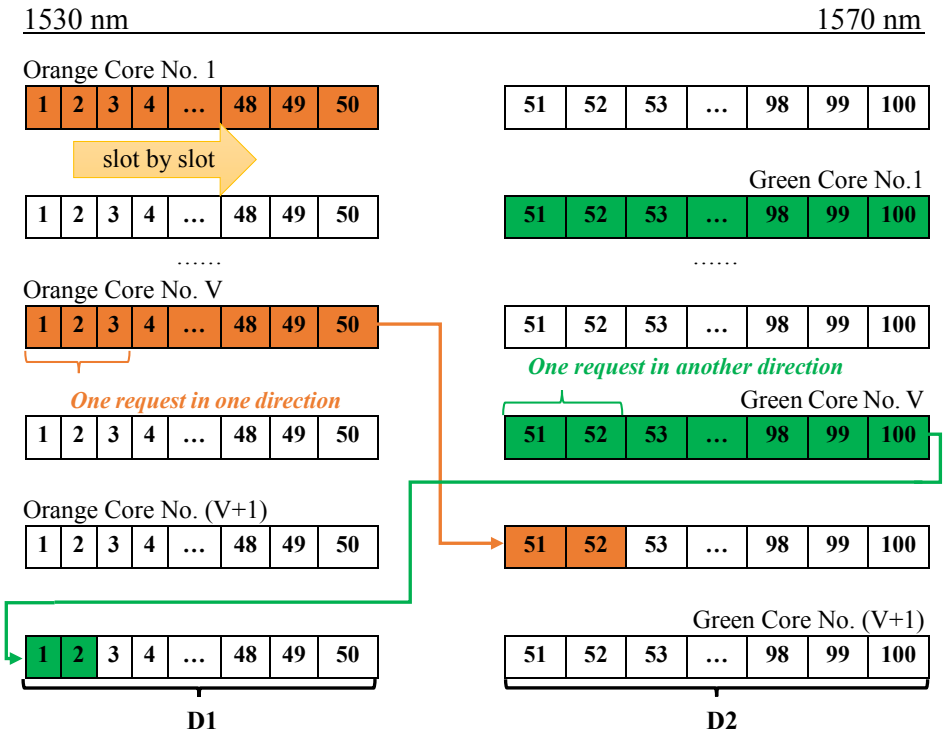


Figure 4.6: Procedures of resource checking in the pre-defined first allocation divisions

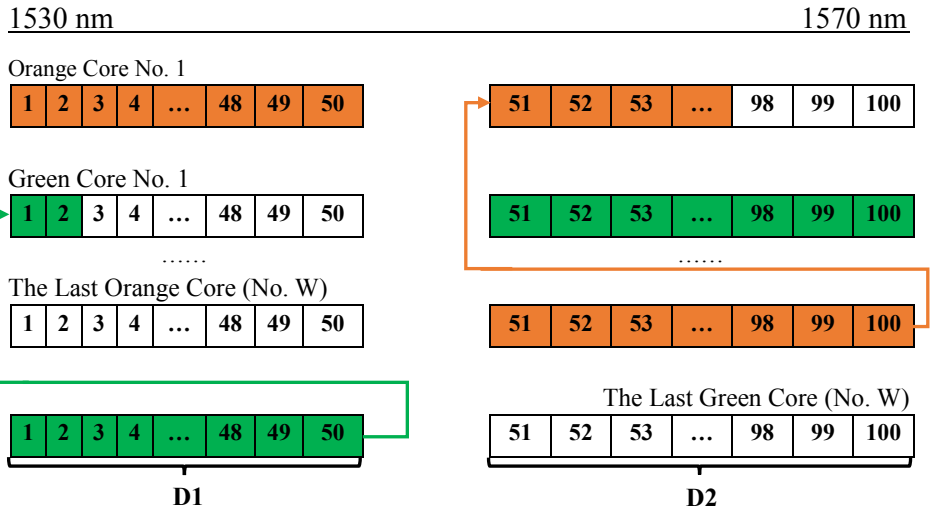


Figure 4.7: Division swap when the pre-defined first allocation divisions are saturated

allocation, while the green Core 1 to Core V can check the slots in D1. It can be seen in Fig. 4.7, when the pre-defined first allocation division of the last orange (i.e. D2) is fully occupied, the divisions swap allowing the second division of the first orange core (i.e. D1) to be scanned and checked.

5. When the the next request comes, the soft spectrum splitting approach will still scan and check the frequency slots in the pre-defined first allocation division of each core, since there may have some remaining unused cores. If no enough frequency slots are found, the divisions will swap again. Only when both the two divisions in all cores cannot offer sufficient frequency slots, the request will be blocked.

Figure 4.8 shows the allocation process of the hard spectrum splitting approach. Differently from the soft spectrum splitting approach, the hard one will directly block the request once there are no enough unused frequency slots can be found in the pre-defined first allocation divisions, in the meanwhile, the blocking probability of the network will be updated. If the blocking probability increases to the given threshold, e.g. 10%, 1%, or 0.1%, the two divisions in all the cores will be swapped permanently, which means the coming requests can only scan and check the slots in the new divisions. It can be found that the value of the threshold determines the slots utilization of the pre-defined first allocation divisions, particularly, the lower the threshold, the higher the possibility that the slots in that division are not be fully utilized. Obviously, this approach can reduce the execution time to provide better processing efficiency, since only one division will be checked for every request. However, it may block more requests when the network utilization is low. In contrast, the soft spectrum splitting approach is more

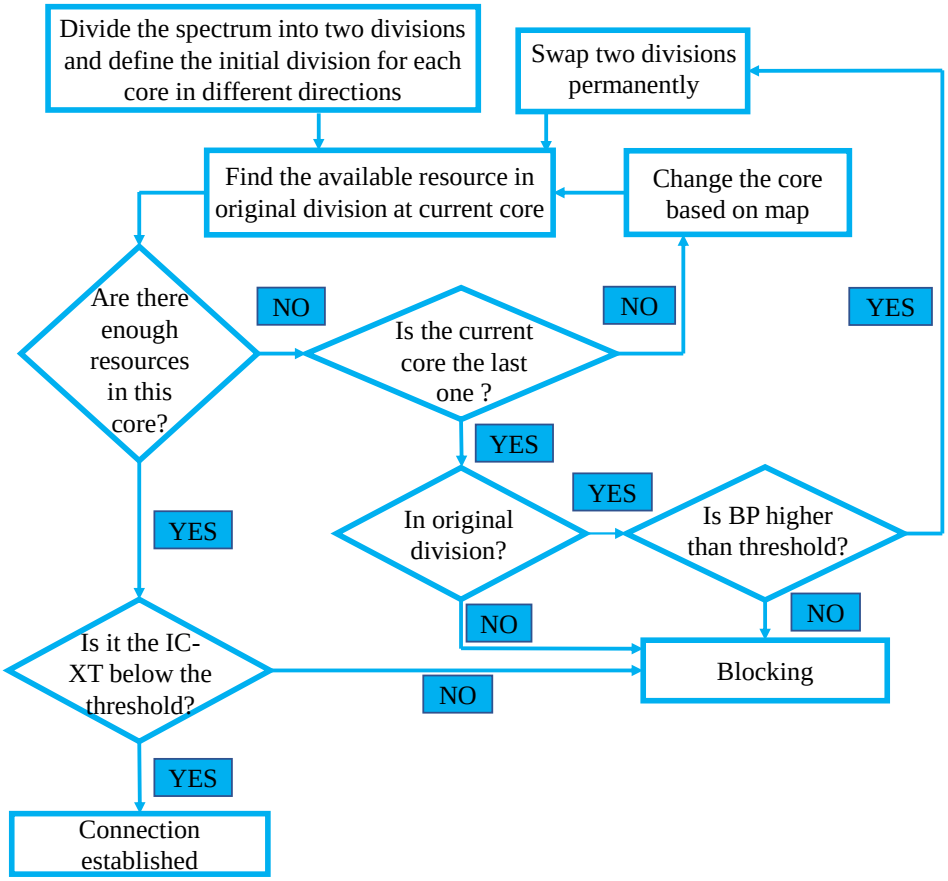


Figure 4.8: Flowchart of the hard spectrum splitting approach (BP: blocking probability)

flexible and the first blocked request will occur only when the whole spectrum has been scanned. However, it has higher computational complexity and requires longer processing time for checking the pre-defined first allocation divisions repeatedly, especially when the network utilization is high.

4.4 Simulation Environment

To evaluate the performance of the proposed algorithms in different DCNs, a simulator in Matlab is developed, which can support pure WDM, pure SDM and SDM-WDM networks with either SMFs or aforementioned eight types of MCFs. For the pure WDM and SDM-WDM schemes, each SMF or MCF core is assumed to have 100 of distinct 25 GHz frequency slots over the C-band, which can be realized by either a passive AWG [202] or an active but more flexible bandwidth-variable wavelength selective switch (BV-WSS) [203]. In addition, three popular data center topologies previously described in Section 2.3, i.e. spine-leaf topology, Facebook data center topology, 3-tier fat tree topology, are adopted. It is assumed that each of the topologies can support 20 racks and each

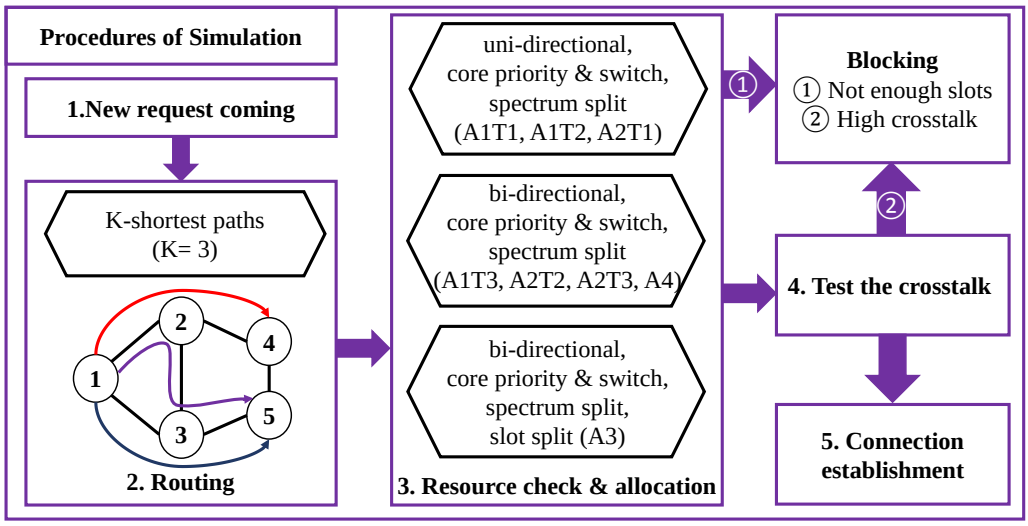


Figure 4.9: Procedures of simulation for each request

server is interconnected to the ToR with a single channel (in the pure SDM or pure WDM scheme) or a single core carrying several channels (in the SDM-WDM scheme). In addition, each link is assumed to consist of two MCFs and a variety of link lengths are considered for evaluation purposes.

4.4.1 Simulation procedures

The overview of the five main procedures of the simulation is shown in Fig. 4.9, including 1) request generation, 2) path routing, 3) resource allocation, 4) IC-XT checking and 5) connection establishment.

The traffic characteristics in DCNs vary according to the requirements of different applications. For example, the DCN traffic utilized in a research institution for carrying out complex computations will be different to that used for supporting video streaming. However, there are few studies on intra data center traffic from data center operators, which makes it difficult to model the data center traffic. Therefore, in this work, it is assumed that all the virtual machines (VMs) requests arrive following a Poisson distribution with an average inter-arrival time of 10 time units. Each request carries the information of start node, end node and bandwidth requirement and has a holding time of 200,000 time units to ensure an incremental traffic scenario. To have a more practical assumption, two types of VM requests are considered, which are shown in Table. 4.2. Uniform distribution of the number of requested frequency slots (bandwidth) is assumed for the first type request and different bandwidth corresponds to different data rates. The potential modulation formats to realize them and the corresponding IC-XT threshold at which the formats induces 1 dB of penalty at BER of 10^{-3} are presented [204, 205]. Note that, a 6 dB lower threshold than the PAM-4 format is

Table 4.2: Assumed modulation and multiplexing schemes of the requests

Bandwidth (GHz)	Capacity (Gb/s)	Possible Modulation	Threshold (dB)
25	10	OOK [207]	-14
50	100	DP-QPSK [208], PAM-4 [209]	-18
75 (25+50)	110 (100+10)	2 λ: OOK +PAM-4	-18/-14
100 (50 +50)	300	PAM-8 [210, 211]	-24
Type 1 request: Combination of 10, 100, 110, 300 Gb/s			
Type 2 request: Fixed data rate, bandwidth (300 Gb/s, 100 GHz)			
Type of multiplexing used on networks considered (fiber type, technique)	a) SDM (using MCFs, fiber switch)		
	b) WDM (using SMF, AWG/WSS)		
	c) SDM-WDM (using MCFs, AWG/WSS and fiber switch)		

assumed for the PAM-8 format based on [206]. This type of request refers to a DCN that goes through phased migration, resulting in some servers having 10 Gb/s transceivers and others with either 100 Gb/s or 300 Gb/s. Particularly, the study on the 110 Gb/s case reflects the sum of 10 Gb/s and 100 Gb/s client rates in two channels to represent DCN evolution where multiple rates could coexist. Type 2 request can apply to a maximum-capacity single-rate green-field DCN realization that with only 300 Gb/s transceivers. In addition, the three considered multiplexing schemes (i.e. pure WDM, pure SDM, and SDM-WDM) can be supported by SMFs with either AWGs or WSSs, MCFs with fiber switches, MCFs with AWGs/WSSs and fiber switches, respectively.

K-shortest paths algorithm, as described in Section 2.5, where K=3, is used for the path routing process. The IC-XT aware algorithm sets for resource allocation and the adopted mechanisms in each set are present in Table. 4.3. The core priority mechanism, which is not added in the table, is applied to all the algorithm sets. A1T1, denotes the algorithm 1-type 1 that considers only core priority mechanism using the *start 1* scheme, was developed for uni-directional transmission [212]. In this Chapter, it is regarded as the benchmark for other algorithms sets. Compared to A1T1, A1T2 utilizes the *start 2* scheme, while A1T3 takes the bi-directional transmission into account. In contrast, the soft spectrum split mechanism is introduced for A2T1, A2T2 and A2T3, among which A2T1 using the *start 1* scheme is for uni-directional transmission, while A2T2 and A2T3 using *start 1* and *start 2* schemes, respectively, are for bi-directional transmission. Compared to A2T3, the slot split mechanism, which has been explained in Section 2.5, is introduced to A3, while A4 replaces the soft spectrum split

Table 4.3: Algorithm sets in the simulations

Algorithm and Type	Direction (1di/2di)	Core Priority (start1/start2)	Spectrum Split (Soft/Hard/N)	Slot Split (Y/N)
A1T1	1di	*start 1*	N	N
A1T2	1di	*start 2*	N	N
A1T3	2di	*start 1*	N	N
A2T1	1di	*start 2*	Soft	N
A2T2	2di	*start 1*	Soft	N
A2T3	2di	*start 2*	Soft	N
A3	2di	*start 2*	Soft	Y
A4	2di	*start 2*	Hard	N

*A/T, Algorithm/Type; Y, with; N, without

mechanism with the hard one in A2T3. In the routing and resource allocation processes, only when sufficient slots with lower IC-XT than the threshold in the selected paths are found, will the connection be established, otherwise, the request will be blocked. The pseudo code for the allocation process with A2T3 is presented in Algorithm 1 (Appendix A).

4.4.2 Fiber characteristics

The detail of the characteristics for the assumed 8 types of MCFs (4 hexagonal and 4 rectangular), as well as the values of the parameters used for calculating IC-XT are shown in Table. 4.4 [113,147,204,213]. It can be found that, in order to arrange more cores in a single fiber, all these MCFs have bigger cladding diameters than that of the standard SMF, which is 125 μm. The coupling coefficients for various

Table 4.4: Detail of MCFs and parameters for IC-XT calculation

Type of MCF (hex)	Core Pitch (μm^2)	Cladding Diameter (μm)	Cladding Area (μm^2)	**Type of MCF (rec)**	C_{P1}-(C_{P2}) (μm^2)	Cladding Diameter (μm^2)	Cladding Area (μm^2)
7-core	30	140	15393.8	**8-core**	30 - ($30\sqrt{2}$)	180	25446.9
19-core	30	200	31415.9	**12-core**	30 - ($30\sqrt{2}$)	180	25446.9
37-core	30	260	53092.9	**30-core**	30 - ($30\sqrt{2}$)	260	53092.9
61-core	25	260	53092.9	**52-core**	25 - ($25\sqrt{2}$)	260	53092.9

Parameters	**Value (unit)**	**Parameters**	**Value (unit)**
β	4*10 (m^{-1})	w_t/a	1
R	0.05 (m)	λ	1530-1570 (nm)
n_1	1.45	Δ_1, Δ_2	0.35,-0.35 (%)

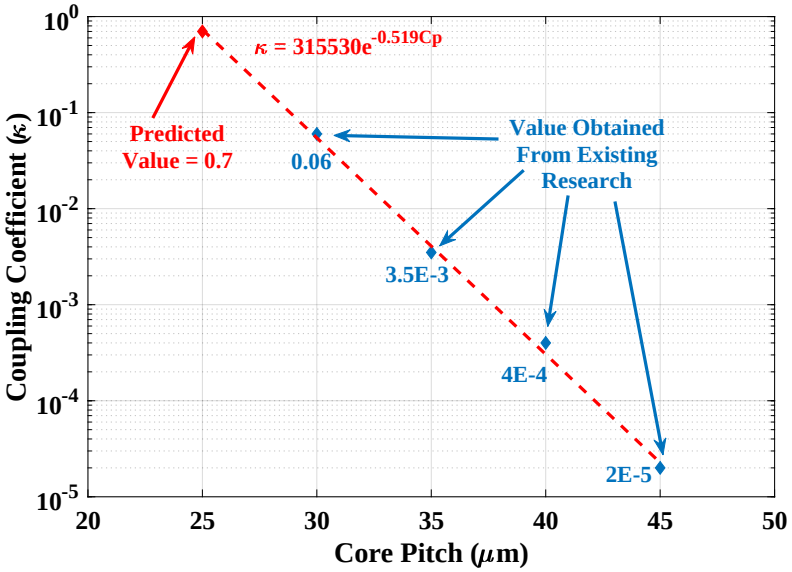


Figure 4.10: Coupling coefficient versus core pitch values

core pitches are depicted in Fig. 4.10, among which the value for 25 μm core pitch is predicted based on the existing researches.

4.5 Performance of the Proposed Schemes

This section firstly explore the IC-XT reduction on the different MCFs, contributing by the bi-directional core arrangement. Subsequently, all the proposed algorithm sets are compared in a DCN with the spine-leaf topology, in terms of blocking probability, network utilization, fragmentation and computational time. Then, the optimal algorithm set is used for comparing the network capacity and link spatial efficiency of the DCNs with three different topologies using different hexagonal MCFs. At last, the topology provides the best performance is adopted for the comparison between the hexagonal and rectangular MCFs.

4.5.1 IC-XT reduction due to bi-directional transmission

According to the experimental demonstration in [200], for a long haul transmission with 100 km link, up to 20 dB reduction in the IC-XT between a core pair could be achieved, attributing to the exploiting of the bi-directional scenario. For short-reach DCNs, since the impacts from some of the parameters may be negligible, the IC-XT suppression between the cores may be more than 20 dB, which means $P_r < 0.01$. By using the Eqs. (2.4), (4.8), and (4.10) to calculate IC-XT in different hexagonal fibers, Fig 4.11 is presented, which shows the effects of the bi-directional transmission and trench-assisted technique on IC-XT.

To explore the effects on the worst-case IC-XT in MCFs, the central core of the MCFs is considered. For example in Fig. 4.4, half of the surrounding core

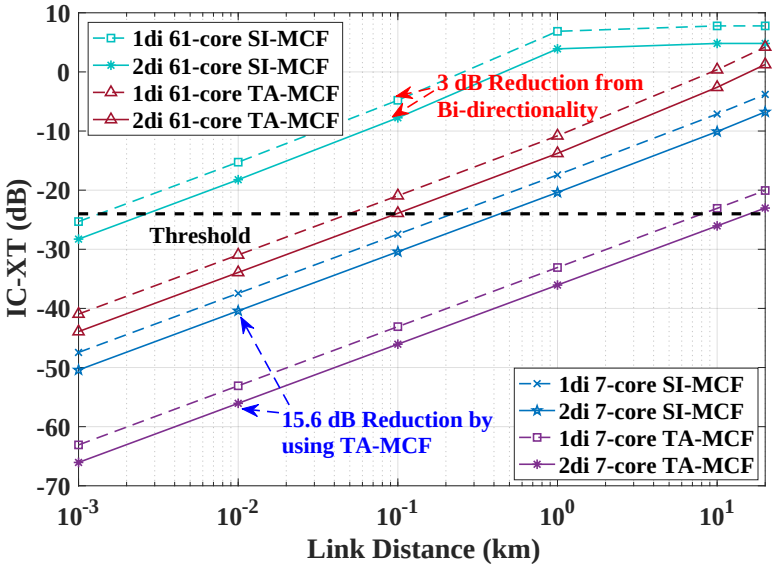


Figure 4.11: IC-XT reduction in the central core of various MCFs due to bi-directional transmission and trench-assisted technique

are carrying signals in the same direction with the central core, while the others carrying signals in the opposite direction. In addition, it is assumed that $P_r = 0.01$ when the bi-directional transmission is considered. As seen, by exploiting the bi-directional model to the two extreme cases, a) 7-core with largest core pitch and b) 61-core with smallest core-pitch, both of them achieve 3 dB IC-XT reductions on the central cores, which means half of the IC-XT is suppressed and indicates that the neighbouring cores on the same direction dominantly contribute the IC-XT. It is also clear by mathematically analyzing Eq. (4.8) that, when P_r is smaller than 0.01, $IC\text{-}XT_{hex\text{-}SI\text{-}oppo}$ contributes marginally to total IC-XT. Therefore, $P_r = 0.01$ is used for all the studies in the following sections. Besides, it can be easily found that a considerable reduction of 15.6 dB on IC-XT can be achieved by replacing the SI-MCFs with TA-MCFs. This 15.6 dB can significantly enlarge the transmission reach with IC-XT below the threshold, moreover, this benefit can be doubled by using the bi-directional model at the same time.

4.5.2 Comparison between the algorithm sets

4.5.2.1 Network behaviour

To compare the network performance of the algorithms, a DCN with spine-leaf topology is developed, where the 7-core MCF is adopted and the link length is 250 m. All the results in this subsection are obtained with the Type 1 requests and SDM-WDM scenario.

Firstly, the network performance of the uni-directional and bi-directional

transmissions is compared in Fig. 4.12(a). As seen, compared to the benchmark set (A1T1), the equivalent bi-directional set (A1T3) can considerably improve the network performance in terms of the overall blocking probability. For example, under blocking probability level of 0.01 and 0.1, the case using A1T3 can provide 8% and 7% higher network utilization, respectively. Furthermore, when the proposed soft spectrum splitting mechanism is incorporated (A2T2), the enhancements on the network utilization at the same blocking probability levels can be further improved to 17% and 16%, separately. Since the IC-XT in the MCF link with 250 m is quite low, the results indicate that the proposed resource allocation mechanism is beneficial even for DCNs with low-XT MCFs. It can also be seen that A2T2 using *start 1* scheme and A2T3 using *start 2* scheme provide the same performance in this DCN, therefore in the following investigations, only the *start 2* scheme will be considered.

The previous Fig. 4.12(a) proves that the exploited bi-directional model and the proposed spectrum splitting schemes can significantly reduce IC-XT in the MCF links and in turn improve the DCN performance. To find the optimal algorithm set for further SDM-based DCN investigations, the network performance of a 7-core MCF based DCN with the spine-leaf topology obtained by using five algorithm sets is presented and compared in Fig. 4.12(b), among which two algorithms sets are for uni-directional transmission (i.e. A1T2 and A2T1, dash lines) and the other three sets for bi-directional transmission (i.e. A2T3, A3, and A4, solid lines). As seen, the first blocking for DCN using A4 occurs at the second earliest time when the network utilization is around 45%, which attributes to that only half of the spectrum resources (one division) can be utilized at the beginning of the allocation process and when there are no sufficient slots for the coming request in that division, the hard spectrum splitting

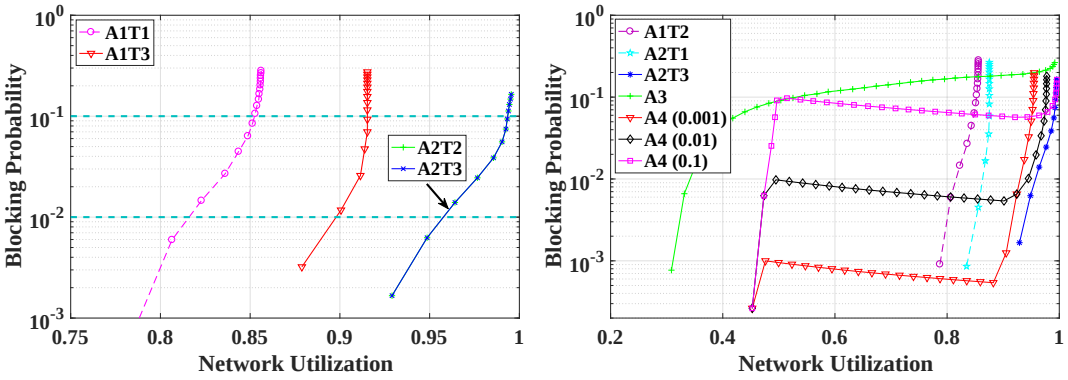


Figure 4.12: Comparison of (a) uni-directional and bi-directional transmissions and (b) all the algorithms in the spine-leaf topology

mechanism directly blocks the request. Moreover, it is found that due to the resource unavailability the blocking probability afterwards increases sharply until it reaches threshold, i.e. 0.001, 0.01 or 0.1, at which time the two divisions swap. Subsequently, the coming requests can use the new slots for allocation and therefore, the blocking probability decreases. However, when the network meets nearly 90%, the blocking probability rises again as the slots in the new division is not sufficient for the subsequent requests. In addition, it is found that the final network utilization can not reach the same level as DCN using the soft spectrum splitting mechanism (i.e. A2T3), which validates the prediction in Section 4.3.2.

In contrast, the first blocked requests of the DCN adopting the remaining three algorithm sets, A1T2, A2T1, A2T3, appear late until the network utilization reaches 79%, 82% and 92%, respectively. This is because these algorithm sets using soft spectrum splitting mechanism check the whole spectrum rather than only one division. Therefore, the blocking due to resource unavailability only occurs when the network saturates. In terms of the network performance of the two algorithm sets for unidirectional model, A2T1 provides considerable improvement against A1T2, which does not use any spectrum splitting mechanism. Moreover, this improvement is further enhanced by A2T3, which considers the bi-directional model. It can be seen that A2T3 shows the best performance in terms of blocking probability versus network utilization, while A3 provides the worst output. It can be explained by the fact that one split narrower band request after the slot split scheme can affect the path selection for all other split requests. For example, if one path of the three shortest paths found can provide sufficient slots with tolerable IC-XT for the first split narrower band request, it will be directly selected for the whole request, which may consist of several split requests, irrespective of whether or not this path can provide sufficient free slots for other split requests in the set. As a consequence, the whole request will be blocked if the requirements for any split request in the set are not satisfied, therefore, the first blocked request for the case using A3 occurs the earliest at around 31% network utilization.

4.5.2.2 Fragmentation

Apart from the blocking probability and network utilization, the performance in terms of spectrum fragmentation for four algorithm sets is also compared and presented in Fig. 4.13. This figure snapshots the states of the highest loaded link (i.e. 7-core hexagonal MCFs) for each set after processing 20,000 requests, where the vertical axis in each subplot stands for the core index and the horizontal axis represents the spectrum slot index. In addition, the slots in different colors are occupied by different kind of requests and white slots are unused. As seen, for the

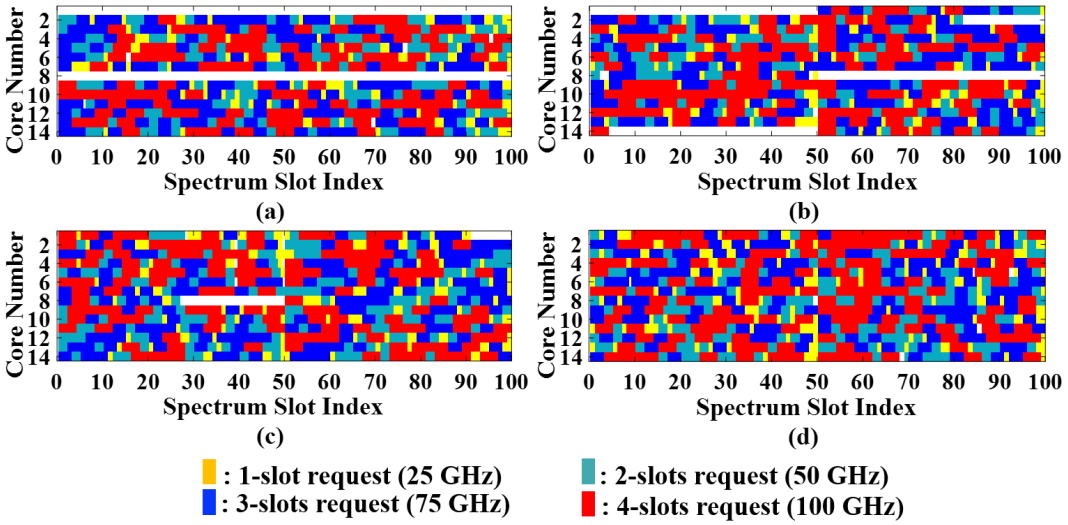


Figure 4.13: Spectrum fragmentation for four algorithm sets in 7-core hexagonal MCF: (a) A1T1, benchmark; (b) A4, hard split; (c) A2T3, soft split; and (d) A3, soft split and slot split

benchmark scheme shown in Fig. 4.13(a), since the central cores of the two MCFs suffer much higher IC-XT than the surrounding cores, these two cores are completely unused. In contrast, when the bi-directional transmission is applied in the remain three schemes, slots in the central cores are used accordingly. Moreover, a clear boundary between the first and last 50 slots can be observed in Figs. 4.13(b)-(d), since the spectrum splitting mechanism is adopted. Compared to A4 shown in Fig. 4.13(b), A2T3 provides 9.5% lower fragmentation, which explains why the network utilization for this scheme in Fig. 4.12(b) is higher than that of A4. Due to the slot splitting approach, A3 achieves the lowest fragmentation ($<$1% slots are unused) for a fully loaded link, however, it will not be used for further DCN investigation as its blocking performance is the poorest.

4.5.2.3 Execution time

Execution time is also an important KPI for evaluating an algorithm, therefore, Fig. 4.14 depicts the computational time for the same four algorithms sets in the process of dealing with 20,000 requests. As seen, since both the benchmark set and A2T3 need to scan and check the whole spectrum during the allocation process, they have similar performance. In contrast, A4 requires the shortest computational time while A3 needs to run for a longest time. The reasons are that for each request, at most only half of the frequency slots will be checked for allocation when A4 is used, however, each request will be separated into several requests and for each of the split requests, the whole spectrum will be checked when A3 is adopted. Taking the performance in terms of blocking probability, network utilization, fragmentation

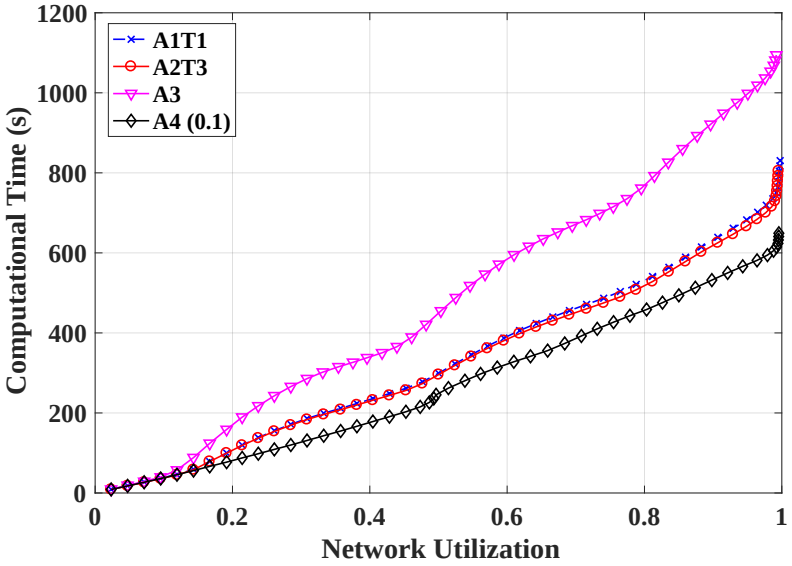


Figure 4.14: Computational time for different algroithm sets

and execution time into consideration, A2T3 is regarded as the optimal algorithm set for the investigations in the following sections.

4.5.3 Comparison between different topologies

All the previous results are obtained with spine-leaf topology, to investigate the network performance of the proposed models and algorithms in DCNs with different topologies, Fig. 4.15 is presented, where the link length is 250 m and A2T3 is utilized for resource allocation. As shown in Fig. 4.15(a), in terms of blocking probability versus network utilization, the spine-leaf topology provides the best performance, which shows 1.5% higher network utilization than both the Facebook topology and the 3-tier fat tree topology under the blocking probability level of 10%. This slight improvement can be attributed to the tier number suppression, however, it limits the scalability and connectivity of the topology. The two three-tier topologies provide similar performance that can achieve over 98% network utilization at 10% blocking probability level.

Since the IC-XT is dependent on the path length, the effect of the topology type on the IC-XT level and in turn the IC-XT induced blocking is clearer. Figure. 4.15(b) depicts the correlation between the percentage of IC-XT induced blocking over the total blocking and the link length for different topologies, in which four types of DCN classified by the link distance are considered: small scale DCN (<10 m; type A), intra-cluster network (10-250 m; type B), larger multi-cluster DCN (250-1000 m; type C) and building-to-building data center farm or metro-to-metro DCN (> 1000 m; type D). As seen, when the uni-directional SI-MCF is considered, the length of the link that the signals can be transmitted in

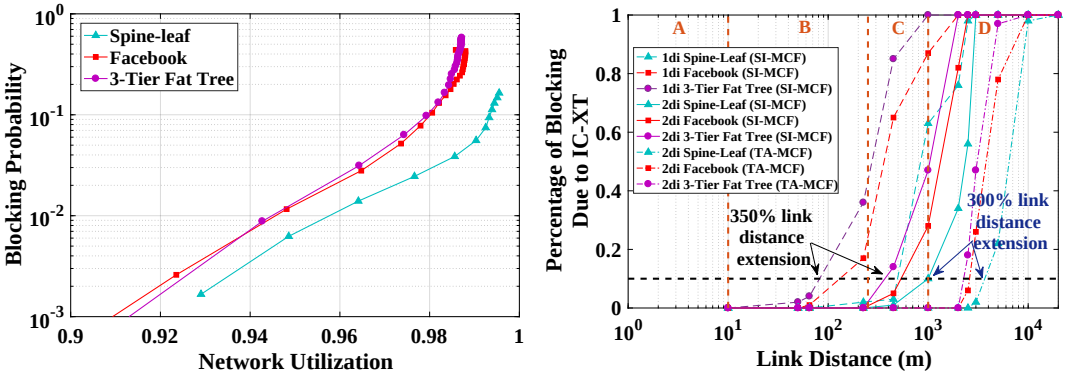


Figure 4.15: Comparison of (a) network behaviour and (b) percentage of blocking due to IC-XT for three topologies

without any IC-XT induced blocking for the spine-leaf topology is longer than that for the two 3-tier topologies. It is attributed to that a whole path between any two end nodes in the spine-leaf topology consists of only 2 links, however, the average link numbers for the whole paths are 3.6 and 4.1 for the Facebook and the 3-tier fat tree topologies, respectively. Taking the path distance into account, the two 3-tier topologies provide better performance. Moreover, it is transparent in Fig. 4.15(b) that when the bi-directional transmission is applied, the lengths of these links in all the topologies can be drastically extended, which has also been mentioned in Fig. 4.11. Specifically, taking the situation when the IC-XT induced blocking accounts for 10% of the total blocking as an example, the bi-directional transmission enables the link distance in the 3-tier fat tree topology extend from 82 m to 370 m, yielding a highly 350% improvement. The improvements for the spine-leaf and Facebook topologies are 100% and 320%, respectively, indicating that the bi-directional approach has greater influence on the multi-tier topologies. However, as shown in Fig. 4.15(b), even with the bi-directional approach the SI-MCFs can only support the realization of DCN types A and B considering either the Facebook topology or the 3-tier fat tree topology. Nevertheless, by using bi-directional TA-MCFs, the transmission distance can be further extended, e.g. 300% distance extension for a spine-leaf topology, enabling the realization of data center types C and D with any of the three topologies.

4.5.4 Comparison between different hexagonal MCFs and multiplexing schemes

4.5.4.1 Front panel density

Before the investigation on the network capacity for different fibers and multiplexing schemes, the benefit of the adoption of MCF on the front panel

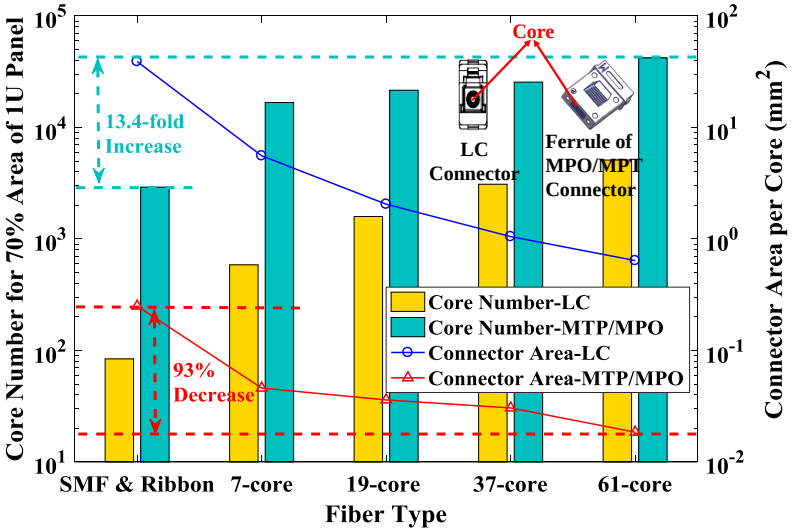


Figure 4.16: Front panel density for DCN with various fibers

density against the SMF or SMF ribbon is explored. To provide an more practical comparison between the SMF, SMF ribbon and MCF solutions, two kinds of standard commercial connector used for accommodating the fibers on a typical 1U rack mount shelf are considered. Moreover, it is assumed that only 70% front panel area of the 1U tray can be used for the connectors, while the remaining 30% area is for the ventilation and other devices. As presented in Fig. 4.16, the lucent connector (LC) connector [214] is adopted for the comparison between the SMF and MCF solutions, where the MCF connector is assumed to have the same surrounding area as that of the SMF one. It can be easily seen that the achievable core density in a 1U front panel, which is presented as yellow bar in Fig. 4.16, is linearly proportional to the core number inside a fiber. For example, the applications of 7-core and 37-core MCFs can achieve 7 times and 37 times core density than the case using SMF, respectively. The reason is that compared to effect of the surrounding area of the connector, which dominates its size, the impact of the fiber cladding area is limited. In order to compare the performance of the fiber ribbon and the MCF solutions, the highly dense 72-fiber multi-fiber termination push-on/multi-fiber push-on (MTP/MPO) connector [215] is considered. It is assumed that the size of the ferrule, in which the fiber alignment is dependent on the eccentricity and pitch of the fiber, and the alignment of pin holes for all fiber types are the same to ensure the same connector size. Therefore, the maximal fiber number that a connector can support can be calculated by:

$$Fiber\ Number\ per\ Connector = \frac{72 \times Cladding\ Area\ of\ SMF}{Cladding\ Area\ of\ MCF} \tag{4.16}$$

This equation indicates the effect of the cladding area of the fiber on the core density. Based on Eq. (4.16), the connector area a core occupies and the core density on a front panel can be calculated by Eqs. (15) and (16), respectively.

$$Connector\ Area\ per\ Core = \frac{Connector\ Size}{Fiber\ Number \times Core\ Number\ per\ Fiber} \tag{4.17}$$

$$Core\ Number\ per\ Front\ Panel = \frac{Front\ Panel\ Area\ Dedicated\ to\ Connectors}{Connector\ Area\ per\ Core} \tag{4.18}$$

These equations can be used to calculated the core density of any other type of panels using any MCFs. It can be seen that compared to the SMF solution, the fiber ribbon based SDM solutions can dramatically increase the core density on a front panel, moreover, the improvement can be further enhanced by adopting the MCFs and MTP/MPO connectors. In this work, the 61-core MCF solution can achieve the highest improvement, i.e 13.4-fold increase, on the core density against the fiber ribbon solution.

4.5.4.2 Network capacity and link spatial efficiency

Two more comprehensive comparisons of the overall achievable network capacity and link spatial efficiency are presented in Fig. 4.17, where different topologies, hexagonal MCFs (shown in Table. 4.4), types of requests and multiplexing schemes (shown in Table. 4.2) are considered. The network capacity refers to the sum of the capacity used on every link contributed by all accepted requests at the blocking probability level of 10%, while the link spatial efficiency or bandwidth density is defined as the capacity divided by the cross-sectional area of the MCF, measured in bits/s/μm^2, which is also a crucial KPI for data center deployment. Moreover, the link distances are short enough to avoid any IC-XT induced blocking in these cases. It should be noted that, type 1 requests can be used for both pure WDM and SDM-WDM schemes, since 100 frequency slots exist in each single core. However, for pure SDM scheme that considers only one channel per core, type 2 requests are used.

In terms of the topology, it can be found that the DCN with the Facebook topology can achieve better performance in both network capacity and link spatial efficiency than the other two topologies, which have similar performance. This can be explained by the fact that the Facebook topology consists of more links than the others (better connectivity), especially, there are four links connected to each end node. However, in both the spine-leaf and 3-tier fat tree topologies, the link

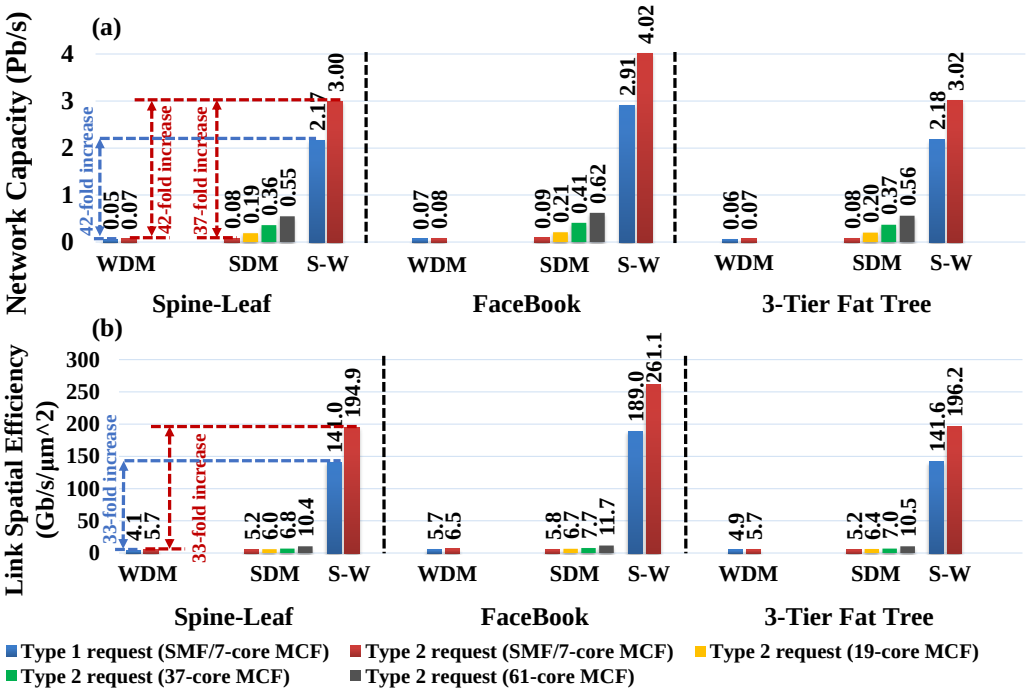


Figure 4.17: (a) Total network capacity and (b) link spatial efficiency obtained by A2T3 in different topologies and for different schemes (*S-W: SDM-WDM)

numbers are three. As shown in Fig. 4.17(a), by using different multiplexing schemes, a variety of network capacities are obtained. Taking the DCN with spine-leaf topology as an example, the adoption of the pure SDM scheme with the 7-core MCFs and type 2 requests offers 0.08 Pb/s network capacity, which is 14% greater than that of the pure WDM scheme with the SMFs obtained using a first-fit algorithm, proving that MCF-based SDM can provide benefits in DCNs against SMF-based WDM. Moreover, with the increase of the core number in the MCF, the benefit can be further enhanced. For example, when the 61-core MCF is adopted, 6.8-fold capacity increase over that of the pure WDM scheme is achieved. Furthermore, by combining the SDM and WDM techniques in the 7-core MCFs, up to 42-fold and 37-fold capacity increases on that of the SMF scenario are observed using type 1 and type 2 requests, respectively.

Unlike the network capacity, the link spatial efficiency for a pure WDM scheme with SMFs and the type 2 requests is 9.6% better than that of the pure SDM scheme with 7-core MCFs for all topologies. This can be attributed to that the cross-sectional area of the SMF is nearly 20% smaller than that of the 7-core MCF. However, this disadvantage of the link spatial efficiency for MCF-base SDM can be compensated by increasing the core number in MCF. As seen in Fig. 4.17(b), when the core number is 19, 37 and 61, the link spatial efficiency for the pure SDM scheme in spine-leaf topology is 17%, 19% and 82% better than that

of the SMF one, respectively. In the Facebook and 3-tier fat tree topologies, the enhancements for the same cases are 3%, 18%, 80% and 12%, 23% 84%, respectively. Similar to the network capacity, the combination of SDM and WDM techniques within a MCF also brings significant benefit on the link spatial efficiency over the SMF case. With either type 1 or type 2 requests, the SDM-WDM scheme with 7-core MCF offers 33-fold higher efficiency than the SMF-based WDM scheme for the DCN using spine-leaf topology.

The coming Fig. 4.18 displays the achievable network capacity for DCNs with different topologies and link distances as well as using different hexagonal MCFs. To have a more practical assumption showing that the DCN enables different transmission capabilities, type 2 requests are utilized. A2T3 is the algorithm set for resource allocation and the case using uni-directional 7-core MCF is the benchmark. It is clear in Fig. 4.18(a) that the maximal achievable capacity of the DCN is proportional to the number of cores in the SI-MCFs. Compared to the SMF-based pure WDM scheme shown in Fig. 4.17(a), all the schemes offer considerable improvement, especially, the 61-core SI-MCF provides the highest 379-fold increase when spine-leaf topology is considered. For the Facebook and the 3-tier fat tree topologies, the increases are 361- and 317-fold, respectively. These highly increases are contributed by both the usage of MCF and the proposed algorithms. Comparing the 7-core bi-directional SI-MCF case with the benchmark, the maximal capacity for them are the same. This can be explained

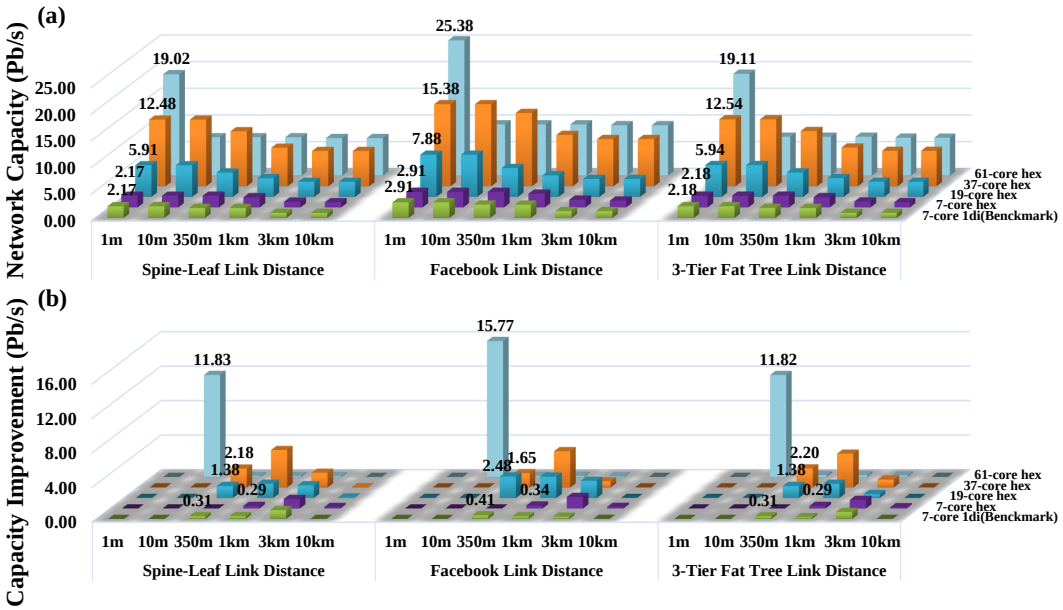


Figure 4.18: (a) Total network capacity as a function of link distance for different normal step-index fiber types and (b) improvement of network capacity by using TA-MCF over using SI-MCF

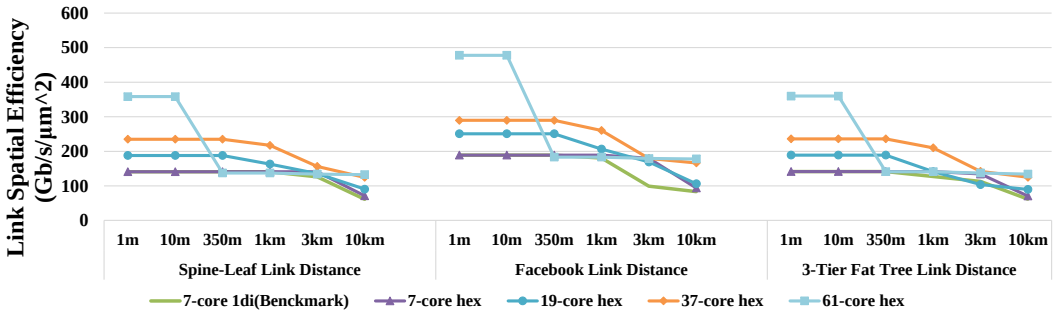


Figure 4.19: Total link spatial efficiency as a function of link distance for different TA-MCF types

by that the maximal capacity is achieved when there is no blocking due to IC-XT (i.e. IC-XT is lower than threshold), which means the IC-XT reduction contributed by bi-directional approach does not make sense on the value of the maximal capacity. However, the IC-XT suppression contributes to protecting the requests from being blocked due to IC-XT value over the threshold. Therefore, bi-directional MCF can keep the maximal capacity for longer distance (without IC-XT induced blocking), which will be discussed in the following content.

Figure. 4.18(b) shows that, by replacing the SI-MCFs with TA-MCFs, although the maximal capacity that a DCN can achieve keeps the same, the lengths of the links where the maximum capacities can be achieved are extended. For example, by using the 61-core TA-MCF, the DCN with spine-leaf topology can maintain the maximal capacity for a link distance of $>$10 m, yielding an over 730% link distance extension on that of the DCN using 61-core SI-MCF. It can be also found in the figure, when the link distance is between 10 m and 3 km, different levels of capacity improvements are obtained for all the investigated topologies. However, when the link distance reaches 10 km, the trench-assisted technique can not bring any benefits any more, as the IC-XT in all types of fibers is severe.

Figure. 4.19 presents the link spatial efficiency of the previous schemes using TA-MCFs. As seen, DCNs using any type of TA-MCF can achieve the optimal link spatial efficiency when the link distance is shorter than 10 m, among which the case using bi-directional 61-core MCF achieves more than 80-fold higher efficiency than that of the WDM solution using SMF shown in Fig 4.17(b) for all the three topologies. Moreover, compared to the new bi-directional 7-core model, the benchmark case is more sensitive to the link length. For example in 3-tier fat tree topology, when the link distance goes up from 1 m to 1 km, 11% efficiency (from 142 to 127 Gb/s/μm^2) decrease occurs. Eventually, it falls to 44% (62 Gb/s/μm^2) of the maximal efficiency. On the contrary, the proposed method can keep the optimal efficiency until around 3 km link distance and finally holds on

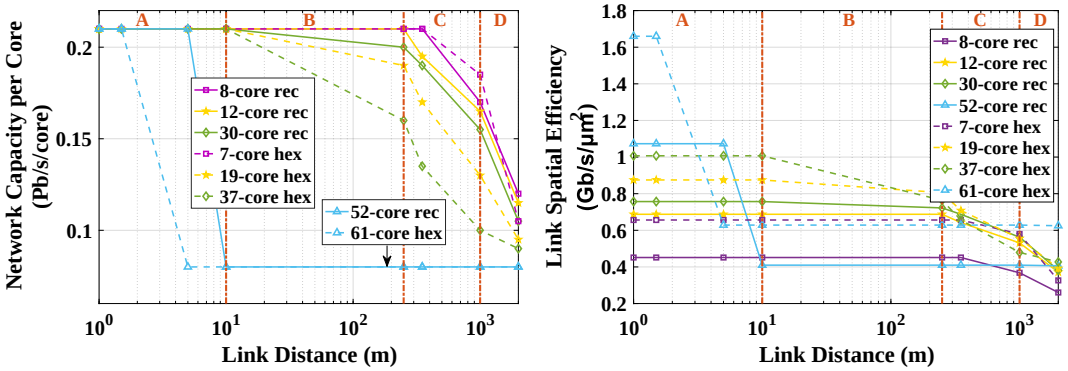


Figure 4.20: Comparison of (a) network capacity per core and (b) link spatial efficiency between hexagonal and rectangular SI-MCFs

nearly 50% (71 Gb/s/μm^2) of the optimal. Although, 61-core TA-MCF can achieve the highest link spatial efficiency, it can only maintain the maximum link spatial efficiency for up to 10 m. Therefore, 61-core TA-MCF may be more preferred when designing a small-scale DCN. However, all the fibers compared in the previous figures are hexagonal fibers with uniform core pitch, to provide a better reference for fiber selection in the data center deployment process, more fibers should be considered.

4.5.5 Comparison between hexagonal and rectangular MCFs

Figure. 4.20 compared the network capacity per core and link spatial efficiency of the DCNs using hexagonal and rectangular SI-MCFs, where the Facebook topology is utilized as it can provide better performance than the other two topologies. In terms of the network capacity per core in Figure. 4.20(a), the hexagonal 7-core MCF provides better performance than that of the rectangular 8-core MCF in DCN type C. For the remaining MCFs, rectangular 52-core MCF outperforms the hexagonal 61-core MCF for DCN type A, while the rectangular 12-core and 30-core MFCs outperform the 19-core and 37-core hexagonal MCFs for DCN types C and D. In addition, the rectangular 8-core MCF provides the best performance among all MCFs when the link distance is 2 km, indicating that the IC-XT inside rectangular MCFs is less sensitive to the distance changing. In contrast, in Fig. 4.20(b) the hexagonal 61-core MCF achieves the best performance among all the investigated MCFs for DCN type D, attributing to the large resource base. Moreover, the hexagonal 37-core and 19-core MCFs outperform other fibers for DCN types B and C, respectively. The results presented in Figs. 4.18, 4.19 and 4.20 could be a useful reference for MCF selection when designing, evaluating, and deploying MCFs for different applications, which depend on the requirements for both resource efficiency and the physical scale.

4.6 Summary and Conclusion

This chapter proposes a novel data center networking solution, considering bi-directional transmission and MCF-based SDM-WDM multiplexing technique. For the first time, a mount of equations for calculating the IC-XT in bi-directional MCFs are derived, with the consideration of different fiber types (i.e. TA-MCF or SI-MCF) and different layouts (hexagonal MCF or rectangular MCF). The derived model is based on several analytical results for MCF systems that have been experimentally proven, validating its application to real SDM network systems. Apart from the IC-XT models, several resource allocation mechanisms, i.e. bi-directional core priority mapping and spectrum splitting, are proposed to improve the performance of the proposed DCN solution. By examining different algorithm sets that combine various mechanisms in DCN, the bi-directional IC-XT aware algorithm set with core prioritization, soft spectrum splitting, and core switching mechanisms provides the best performance in terms of blocking probability and network utilization. Several homogeneous SI-MCFs and TA-MCFs are investigated in three different topologies with the objective to maximize capacity and spatial efficiency of the DCN, and find the best fit between fiber type and data center environment. Simulation results demonstrate that the bi-directional model can extend the transmission reach of MCF against that of the uni-directional model under the same conditions. In terms of topology, the Spine-Leaf topology shows a slight advantage in network behavior, while the Facebook topology provides the highest network capacity and link spatial efficiency. For the multiplexing schemes, the experimental results support the superiority of SDM over WDM networks. For SDM-WDM networks with MCF, more than 80 times higher link spatial efficiency and up to over 300 times increased capacity (61-core MCF) are attained compared with the SMF-based WDM solution for all three topologies. This indicates the potential of SDM solutions to support future high-capacity DCNs. Eventually, the studies carried out here on network capacity with respect to the link distance clearly highlight that crosstalk suppression enables highly dense MCFs, i.e., 61-core hexagonal SI/TA MCF and 52-core rectangular SI-MCF are the ideal candidates for small-scale data centers, e.g. intra-rack DCN, data center in a box, and points of delivery (PODs), with link spans ranging from few meters and metro-to-metro data centers with >10 km link distance. 37-core hexagonal SI/TA-MCF achieves the best results for intra-cluster network with 10s of meter link distances and 19-core hexagonal SI/TA-MCFs can be utilized for larger multi-cluster data centers. For building-to-building data center farms with link spans of up to 1000 s of meters, rectangular MCFs are more suitable since the IC-XT in this kind of fiber is less sensitive to the link distance change.

Critically, although the introduction of MCFs can considerably increase the DCN performance in terms of capacity and spatial efficiency, the complexity of the allocation process is considerably increased. In this work, DCN with only 20 end nodes (ToRs) was investigated, if it is scaled to have 1000 or more ToRs, the time required for resource allocation would be risen exponentially. Therefore, more algorithms that can simplify the resource allocation process are necessitated for the deployment of SDM techniques in real world. Moreover, optical switches and amplifiers play an important role in DCN. However, the existing MCF switches (even in research) can not perfectly realize core to core switching, which limits the flexibility of the DCN architecture, while the MCF amplifier is still under research. To convince data center operators, service providers and network vendors to deploy MCFs in DCN, all of the aforementioned challenges should be addressed.

Chapter 5

Disaggregated Data Center Networking

5.1 Introduction

The adoption of SDM techniques is a solution to improve the DCN performance only at the link level or bandwidth level. However, the challenges that the data center faces may also from other fields. It has been aforementioned in Section 2.2 that the conventional server-centric data center architectures, where each server tightly integrates a fixed amount of IT (i.e. CPU, memory, storage) and network (i.e. bandwidth) resources onto a single mainboard, present shortcomings in areas such as flexibility, adaptability, scalability and resource utilization. Particularly, IT resources cannot be fully utilized in each server due to the mismatch between fixed proportionality and diverse set of request requirements. To address these issues, disaggregated system has been proposed, which classifies IT resources into different resource pools, and an optical or opto-electronic network is utilized to interconnect all the resources. However, the main challenges for the disaggregated system are a) the traffic extension from a server range to the whole data center scale requires extremely high bandwidth and low latency [10] and b) the extra bandwidth requirements between the IT resources result in extra cost compared to that of the conventional data center. To overcome these challenges, a novel and fully developed resource-centric architecture for DDCs called the disaggregated recursive data center in a box (dReDBox) is proposed [46]. This architecture allows all IT elements in the topology to act as standalone entities with dynamic on chip packet/circuit switching capabilities, which can independently communicate with one another through a high capacity and low latency circuit switched optical network.

In this chapter, the proposed dReDBox architecture is first introduced at the rack-and-cluster scale with the consideration of three disaggregated rack

structures. Subsequently, a simulation platform has been developed to perform coordinated orchestration and allocation of IT resources together with reservation of their network bandwidth and interconnection for both TDC and DDC. It investigates the performance of these two data center architectures under six different types of input requests (including IT resource requirements and network requirements). All the requests arrive the system dynamically and can only be served when both of the IT resources and network resources are successfully allocated. During the simulation process, four different allocation algorithms for DDCs with different rack structures are compared and contrasted, the best algorithm-structure combination is then selected for the comparison between DDC and TDC. At last, a cost model is proposed to evaluate the cost efficiency of DDC in different network infrastructures.

5.2 Disaggregated Data Center Architectures

It can be seen in Fig. 5.1 that the dReDBox data center architecture is developed based on a 3-tier fat tree topology (shown in Fig. 2.5) and consists of disaggregated racks (dRacks), disaggregated boxes (dBoxes) and disaggregated bricks (dBricks), which refer to the racks, blades and bricks in TDC, respectively. The dRacks in the architecture are interconnected by the switch module called disaggregated data center optical switch module (dDOSM) and each dRack houses multiple dBoxes interconnected by the disaggregated rack optical switch modules (dROSMs). Inside each dBox, up to 16 dBricks can be resided and the dBricks can be arbitrary combinations of compute/memory/accelerator bricks according to the requirements. To support the networking inter dBoxes and/or intra a dBox, the

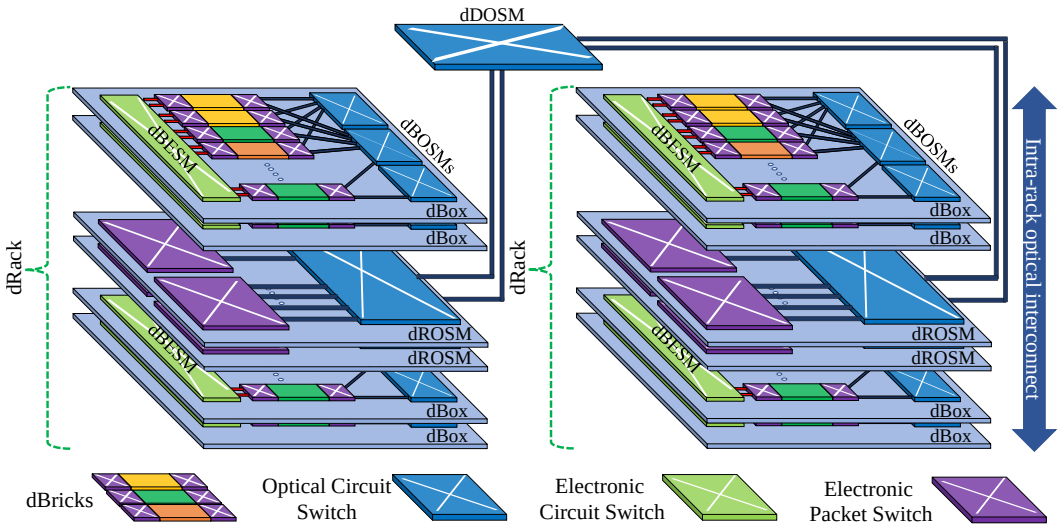


Figure 5.1: dReDBox data center architecture

dBricks are connected to the disaggregated box optical switch modules (dBOSMs). Apart from the optical switches, to realize fast communication between the dBricks placed in the same dBox, they are also interconnected to each other through a disaggregated box electronic switch module (dBESM). It should be noted that inter-dBox communication can only be realized through the dBOSM.

To deliver a high level of modularity and reflect the subscription ratio requirements, each dBOSM switch, which refers to the top of blade (ToB) switch in TDC, has a small port count (i.e., 48 or 96) with the highest density. In contrast, since lots of links are connected to the dROSMs (refers to ToR switch in TDC) and dDOSMs, large port count (i.e., 384 × 384) switches are required. Moreover, in order to offer transparent connectivity between different tiers, the dROSMs also provide access to pluggable electronic programmable packet switches. The packet switching services can be realized by using the programmable on-chip packet/circuit switching supported on dBricks or attached as pluggable modules on dROSMs. To minimize footprint and power consumption while maximizing bandwidth density of the architecture, on-board 200 Gb/s Silicon Photonic single-mode 1.3 μm transceivers will be used at each dBricks. The beam-steering switches at the switch modules and the Silicon Photonic transceivers contribute to a transparent brick-to-brick multi-hop network with minimum possible latency. The combination of transparent switches with protocol programmable system on chip allows for a function and topology programmable architecture.

To compare to the performance of the proposed DDC architecture with the

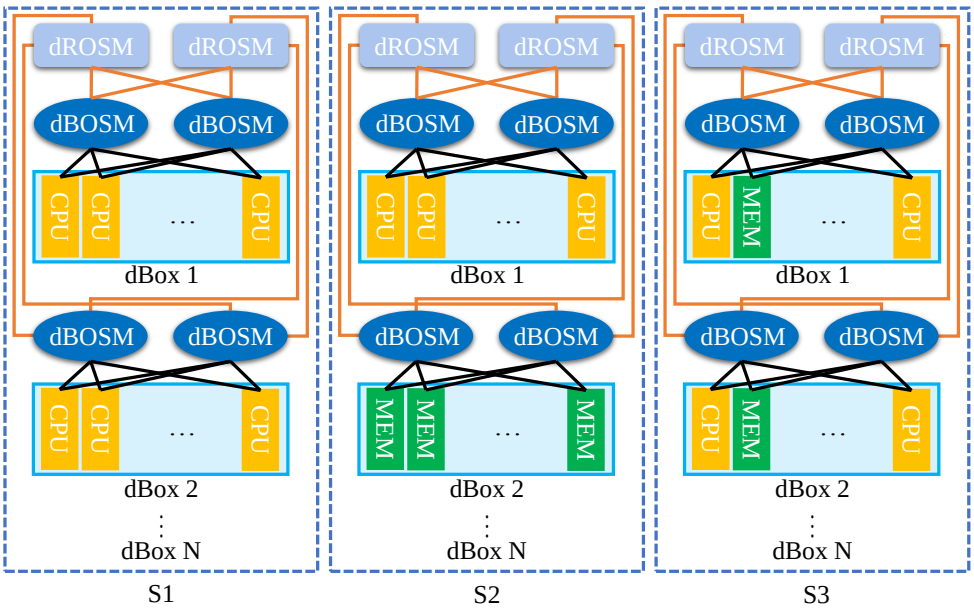


Figure 5.2: dRack structures for DDC architecture (*MEM: memory)

TDC architecture, three dRack structures are considered: structure 1 (S1), homogeneous dRack with homogeneous dBoxes, which means that a dRack (including all the dBoxs inside) can only accommodate one type of dBricks; structure 2 (S2), heterogeneous dRack with homogeneous dBoxes, where different IT resources are allowed in one dRack, but each dBox holds only one type of dBricks; structure 3 (S3), heterogeneous dRack with heterogeneous dBoxes, where each dBox contains various types of dBricks. In a DDC, when a request arrives, the system will elect the corresponding IT resources from several different dBricks and set links between these IT resources. For the DDC with S1 dRack, as shown on the left Fig. 5.2, the path selected between the CPU and memory dBricks for a request could be extremely long since these two types of dBricks are configured in different dRacks. In contrast, for the DDC with the middle dRack in Fig. 5.2, the path could be confined within rack scale owning to that the CPU and memory dBricks are distributed in the same dRack. Furthermore, as a dBox contains both CPU and memory dBricks for DDC with the right dRack in Fig. 5.2, the path could be even shorter, i.e. within the dBox. Transparently, the short-distance connections are more likely to be built for communications with S3 dRack, which is expected to realize the best performance among all the investigated structures.

On the contrary, for the racks in TDC, each server or brick houses both the CPU and memory resources. During the process of IT resource allocation, the CPUs and memory selected for a request should be allocated in the same server, which indicates that one request can only be served by one server or brick. However, in DDC a request can be served by different dBricks. Since investigation in this work focus on CPU and memory utilization in different data centers, disaggregated storage pools are utilized in all the data center structures during the simulation.

5.3 Simulation Environment and Proposed Resource Allocation Algorithms

A simulator in Matlab is developed to compare the performance of the TDC and DDC with various dRack structures. This simulator performs coordinated orchestration and allocation of IT resources together with the reservation of their network bandwidth and interconnection to serve VM requests, which can be divided into four stages: a) data center creation b) request generation, c) IT and network resources allocation and d) results collection.

5.3.1 Data center configurations

The assumed TDC and DDC configurations are depicted in Table 5.1. It is notable that some of the parameters are variable to fit different requirements. As presented

Table 5.1: Data center parameters and resource volumes

Parameter	Description			Value
nRacks	Total number of dRacks (racks)			12
nBoxes	Number of dBoxes (blades) per dRack (rack)			6
nBricks	Number of dBricks (slots) per dBox (blade)			8
nUnits	Number of units per dBrick (slot)			16
nROSM	Number of dROMS (ToR) switches per dRack (rack)			2
nBOSM	Number of dBOSM (ToB) switches per dBox (blade)			2
Brick Type	**DC Type**	**Volume of a brick**	**Unit Size**	**Total Volume**
CPU and Memory	TDC	8 CPU units + 8 MEM units	4 cores / 4 GB	12233 core + 12 TB
CPU	DDC	16 CPU units	4 cores	12288 cores
Memory	DDC	16 MEM units	4 GB	12 TB
Storage	Both	16 storage units	64 GB	192 TB

in the table, there are twelve racks for each type of data center structure, eight of them are used to assign CPUs and memory, and the other four are used for storage, which realizes 1/3 ratio of compute, memory and storage resources across the whole multi-rack system. Moreover, it depicts the unit size and volumes of IT resources. In TDC, a brick contains both CPU and memory resources while only one type of IT resource can be configured in a DDC dBrick. To keep the same IT resources volumes in these two types of data centers, CPU units or memory units in a TDC brick is half of that in a DDC dBrick. As aforementioned, the same volume of storage is used for TDC and DDC.

Table 5.2 depicts the connectivities inside the data centers, including the topologies utilized between different tiers, number of channel per link, channel bandwidth and link distances between the switches and bricks. Note that the number of channel will be decreased when investigating the impact of bandwidth on the performance. Additionally, the delay assumptions are used to measure the cost of latency in different situations. In TDC, only Tx & Rx delay is valid since

Table 5.2: Connectivity and delay assumptions

Level	Topology	Link Distance (m)	Channels
dROSM-dROSM (ToR-ToR)	Fully-connected	10	16
dROSM-dBOSM (ToR-ToB)	Spine-leaf	3	16
dBOSM-dBrick (ToB-brick)	Spine-leaf	0.25	8
Chanel Bandwidth	25 Gb/s	**Switch latency/ Tx & Rx delay**	5 ns/ 140 ns

Table 5.3: Requirements for the VM requests (*STO: storage)

Request	Type 0	Type 1	Type 2	Type 3	Type 4	Type 5
	random	high MEM	high CPU	half & half	more MEM	more CPU
CPU, MEM	1-32 core, 1-32 GB	1-8 core, 24-32 GB	24-32 core, 1-8 GB	Type 1 (50%), Type 2 (50%)	1-16 core, 17-32 GB	17-32 core, 1-16 GB

CPU-MEM Bandwidth (fixed for all requests)	5 Gb/s/unit	**MEM-STO Bandwidth (fixed for all requests)**	1 Gb/s/unit
CPU-MEM Round-trip Latency	180-480 ns	**MEM-STO Round-trip Latency**	480-780 ns

memory is directly attached to CPU. However, switch delay and fiber delay (i.e. 5 ns per meter) are involved for the communication between CPU and memory in DDC.

5.3.2 Request requirements and generation

The IT and network requirements for a single VM request are shown in Table 5.3. Six types of input requests are assumed to simulate different application environments. As described in the table, each request dynamically arrives following a Poisson distribution with an average inter-arrival time of 10 time units, containing the information of IT resources requirements (number of CPU cores, size of memory and size of storage, which is 128 GB for each request), network requirements (CPU-MEM or MEM-STO bandwidth and latency) and holding time. Note that the latency lower bound (180 ns) is the latency requirement from the transceivers, while the upper bound includes the network latency. The range is chosen to represent local and remote CPU-MEM communication latency requirements. The only differences between these types are the demands on CPU number and memory size. All other parameters for them are the same, either be constant or change randomly within a given range. As for the holding time, the first 100 requests will be held for 6300 time units, and the holding time increases by 360 time units for every 100 requests.

5.3.3 IT and network resources allocation

The flowchart of overall resource allocation process is depicted in Fig. 5.3, including both IT resource allocation and network resource allocation. When a new request comes, the simulator will firstly execute the IT resource allocation algorithm and then the network allocation algorithm. During the resource allocation process, two databases (IT resource database and network resource database) are utilized to record and update available resources. A request could be served only when sufficient IT and network resources are found, otherwise, the

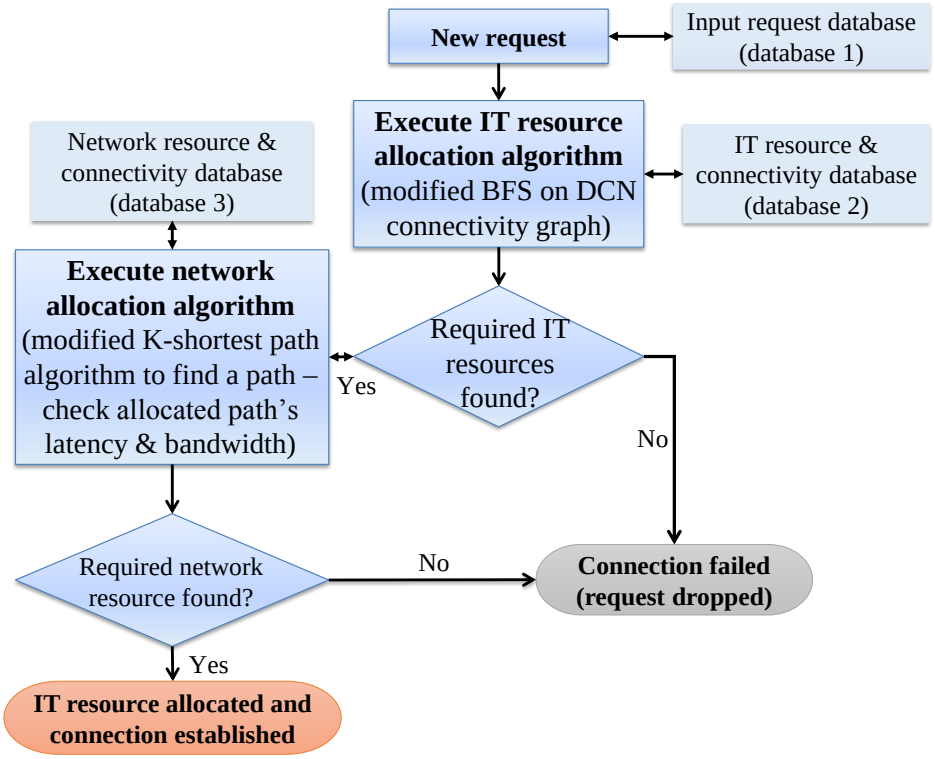


Figure 5.3: Flowchart of IT and network resource allocation

request will be dropped (blocking).

In TDC, memory is directly attached to the CPU and thus, there is no extra bandwidth and latency requirements on the network. In contrast, in DDC, IT resources could be placed far away from each other. To Maximize the IT and in particular CPU/memory resource utilization, the algorithm should carefully select IT resource nodes in case of failures result from high latency. Therefore, a well-designed allocation algorithm is necessary for boosting the efficiency of the system and decreasing the network load in DDC. Four different IT resource allocation algorithms are developed in this work for DDC (the pseudo code for the algorithms is presented in Appendix B):

1. *First-fit (FF) resource allocation algorithm* is simplest algorithm developed based on first-fit basis which has been aforementioned in Section 2.5. With this algorithm, when a request arrives, all the IT resource nodes will be scanned one by one and the request will be allocated to the first available nodes. The FF resource allocation does not consider the network resources around the allocated IT resource nodes, thus it is possible that the bandwidth between these IT resource nodes is insufficient and then blocking occurs. As a consequence, high blocking probability as well as low resource utilization could be seen when this algorithm is adopted in a DDC. Note that, the TDC

will only use this algorithm for IT resource allocation.

2. *Best-fit (BF) resource allocation algorithm* is an optimization of the previous one. Dissimilarly, it classifies IT resource nodes according to resource types (CPU, memory or storage), and searches each type of resource independently. In this case, the best possible combination of IT resource nodes could be found since one type of resource may be found in different racks. Therefore, the blocking probability can be considerably reduced.

3. *Network-unaware Locality Based (NULB) resource allocation* is proposed to realize globally resource allocation. In this algorithm, the concept of *Contention Ratio* for each type of IT resources is created, which relates to the ratio of the units required and the current available units in the whole data center. A high contention ratio indicates that this type of IT resource is highly demanded. When a request arrives, the algorithm will firstly search and allocate the resources with the highest contention ratio, and then search other types of IT resources around the allocated nodes via breath-first search (BFS). BFS is an algorithm for traversing and searching tree and graph data structures. It is utilized to find the nearest available node to a given node. Since one type of resource is allocated firstly, other types of resources nearby will be found. As all the required resources are allocated on the neighboring nodes, it largely releases latency pressure in the DDC.

4. *Network-aware Locality Based (NALB) resource allocation algorithm* is an optimization of the NULB resource allocation algorithm which uses a modified breadth-first search algorithm. It involves network factors (high bandwidth links have a higher priority) while searching nearby nodes after the nodes with highest contention ratio have been allocated. It pre-considers the network resources between all the allocated IT resources nodes before network allocation algorithm starts, which deceases the blocking probability to a large extent.

After the IT resource allocation, the network resource allocation will be executed. In this chapter, all the IT resource allocation algorithms follow the same network allocation algorithm: *modified K-shortest paths algorithm*. Compared to the traditional K-shortest paths algorithm described in Section 2.5, a new weight factor (W) considers both bandwidth and latency requirements is introduced to replace the original weight, distance.

$$W = f \times (1 - \frac{available\ bandwidth\ of\ the\ current\ link}{max\ bandwidth\ of\ one\ link}) + (1-f) \times \frac{distance\ of\ the\ current\ link}{max\ link\ distance} \quad (5.1)$$

The value of W can be calculated by Eq. (5.1), where f is a variable between 0 and 1. If the value of f is close to 1, the bandwidth is favored, whereas a value close to 0 means distance/latency is favored. In this work, f is assumed to be 0.5 to ensure that both bandwidth and latency are weighted equally.

5.4 Performance of the Data Centers

5.4.1 Comparison between DDCs with various dRack structures and algorithms

As aforementioned in the previous sections, three different dRack structures and four different IT resource allocation algorithms for DDC are proposed. Before comparing the performance of the DDC and TDC, the best structure and allocation algorithm combination for DDC should be found. Therefore, Fig. 5.4 is presented, which shows the number of blocked requests for different algorithm-structure combinations in DDC after processing 1000 Type 0 requests and the maximum IT as well as network resource utilization for each of them. It should be noted that among the three reasons which can lead to request blocking, IT resource (e.g. CPU) unavailability has the highest priority. This is because if there is no sufficient IT resource found, the request will be directly rejected and the network allocation process will not start. Blocking due to bandwidth or latency will only occur when sufficient IT resources are found. Therefore, each blocked request will only be

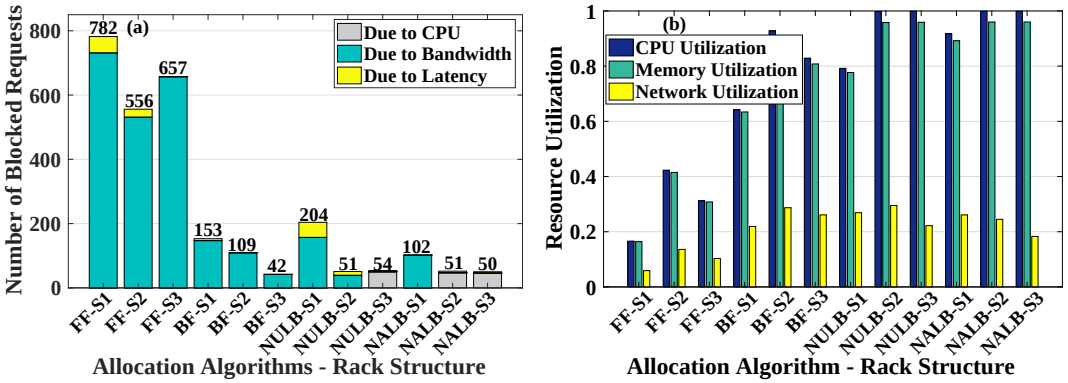


Figure 5.4: (a) Number of blocked requests and (b) resources utilization for different rack structure and algorithm type combinations

counted once. It can be seen in Fig 5.4(a) that a large number of requests are blocked when FF allocation algorithm is applied, while all other algorithms show better performance than it. Especially, the NALB allocation algorithm performs the best since it considers network resources while allocating IT resources. In terms of DDC structure, S2 and S3 present much less blocked requests than that of S1. This is because that a mass of network resources is needed for S1 to link IT resource nodes that are placed far apart (e.g. different dRacks).

BF-S3, NULB-S2, NULB-S3, NALB-S2 and NALB-S3 outperform other combinations in Fig 5.4(a), as such the best algorithm-structure combination should be found among them. To begin with, all of the blocked requests in BF-S3 and NULB-S2 shown in Fig 5.4(a) result from the bandwidth insufficiency and/or high latency, which indicates that the network utilization in these two combinations are higher (above 25% in Fig 5.4(b)) than the other three cases. Moreover, CPUs cannot be fully utilized with BF-S3 (approximately 82%) since blocking occurs due to network factors before IT resources saturate. However, in the remain cases, blocking occurs due to CPU unavailability (CPU resources are fully utilized). To clarify the best option among these three combinations we can only see that the NALB-S3 requires the least network resources (18%) to achieve the highest CPU utilization (100%), this keeps the added network cost and complexity to the minimum.

The round-trip (memory read/write transaction) network latency is also an important KPI for evaluating the network performance. Therefore, Fig. 5.5 is

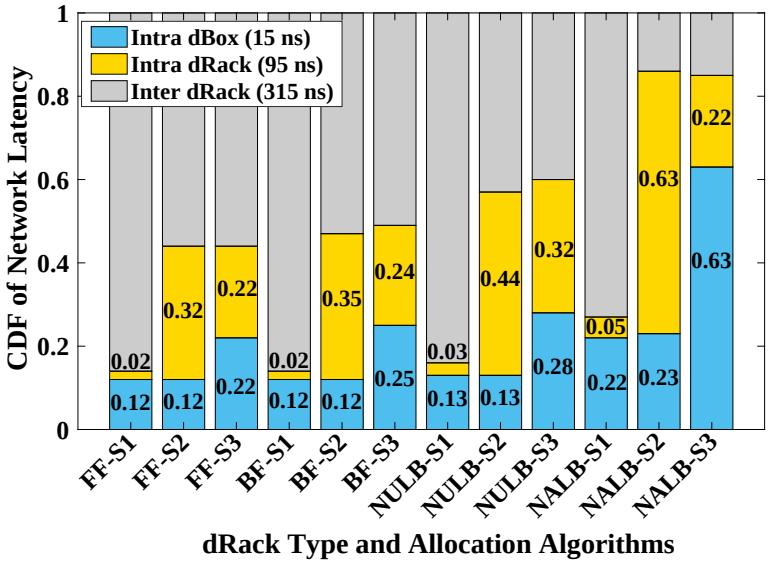


Figure 5.5: CDF of the round-trip network latency for different rack structure and algorithm type combinations

presented showing the cumulative distribution function (CDF) of network latency for all algorithm-structure combinations. The network latency is the sum of the propagation latency in the fibers and passing delay in the switches. As seen, in terms of the dRack structures, DDC with S2 and S3 dRacks outperform S1 for all the algorithms. This is because that for these two heterogeneous rack structures, up to 86% (NALB-S2) and 85% (NALB-S3) of the traffic happens with in the dRack. However, each request in DDC with S1 dRack requires at least three dRacks (CPU, memory and storage), resulting in 73% (NALB-S1) of the traffic is between the dRacks. Moreover, since heterogeneous dBoxes are considered in S3 dRacks, 174% more traffic is generated within the dBoxes than that of S2 with the NALB algorithm, contributing to a 28% average latency reduction (from 107.4 ns to 77.6 ns). This factor can also be used to explain the reduction in network utilization of NALB-S3 compared to NALB-S2 when all CPUs are utilized in Fig. 5.4(b). As for the algorithms with S3, the case using the NALB algorithm outperforms all the others, which achieves 63%, 58% and 52% average latency reduction on that of the cases using FF, BF and NULB algorithms, separately. To sum up, both the DCN structure and resource allocation process plays an important role on reducing the network latency. As NALB-S3 shows great benefits in terms of resource utilization and latency, it will be utilized as the benchmark for identifying the advantages of the DDC architecture against the TDC architecture.

5.4.2 Comparison of network performance between DDC and TDC

In this work, the total IT resources in these two data center architectures are assumed to be identical, and six types of input requests databases (Table 5.3) are utilized to imitate different application environments. Figure. 5.6(a) shows the blocking probability of the data centers with six types of input requests. Results show that with Type 0 requests, the first blocked request is observed at the 575th request in TDC, whereas no blocking occurs until the 925th request in DDC. Networks with Type 1 and Type 2 requests present the similar results, attributing to that the total number of CPU units and memory units are identical in each data center. Memory in the network with Type 1 request or CPU in that with Type 2 requests will be fully occupied at a similar time point. In these two cases, there could be extremely high demand for one specific IT resource (memory or CPU). When one type of IT resource is fully occupied, massive blocks will occur. As such, high blocking probability can be seen for networks with these two types of requests in both TDC and DDC. However, DDC performs better than TDC in these conditions. By combining these two conditions (Type 3), the blocking probability

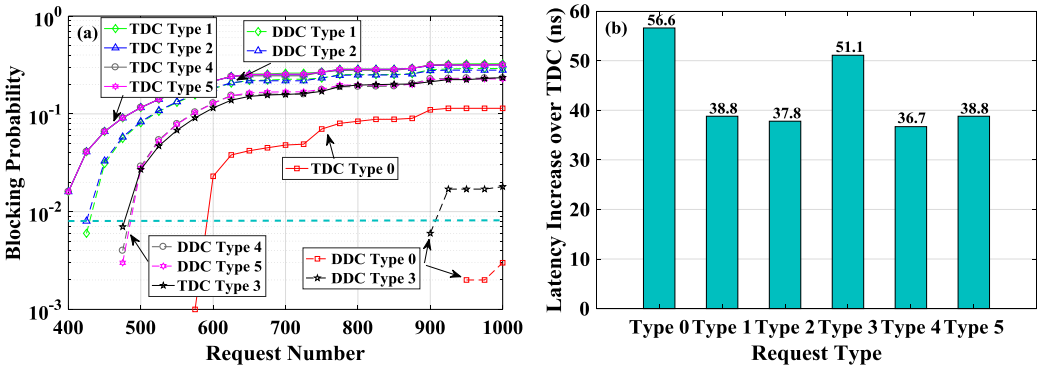


Figure 5.6: (a) Comparison of blocking probability and (b) average increased round-trip network latency of DDC over TDC using six different input requests

of DDC reduces considerably. In this case, high CPU demand request could be a complementary of high RAM demand request. As such, DDC performs much better with Type 3 requests than the previous cases. In contrast, TDC achieves little benefit since one request could only be allocated to one server. Once the memory or CPUs in a server is fully utilized by a single high RAM demand request or a single high CPU demand request, the server could not be utilized to serve other requests. Similar blocking probability are obtained with Type 4 and Type 5 requests.

Since adequate storage as well as network resources are provided in the model, all the blocks are caused by CPU or memory factors. Type 0 as well as Type 3 performs better than others because of IT resources are utilized in a more efficient way in these two types. It is notable that the blocking probability reduces when the input requests change from Type 1 to Type 4 in DDC, while the performance stays similar in TDC. When a request requires more than 16 GB of memory, a slot (containing 32 GB of memory) will only be able to serve one request in TDC as the remaining memory is less than 16 GB. Since all the requested memory above 16 GB, there is no difference between Type 1 and Type 4 for a TDC. The similar results are obtained between Type 2 and Type 5. To sum up, results show that DDC performs better than TDC under all these six types of input requests without considering IT and network utilization. To evaluate the exact performance of DDC, more factors should be involved.

Latency is an important indicator to evaluate the performance of DDC. As different types of IT resources are not directly accommodated in the same slot, extra latency will be involved in data transmission process. In this study, since disaggregated storage pools are utilized, only the latency between CPU and memory is compared between these two types of data centers. Fig. 5.6(b) presents the average increased round-trip latency between CPU and memory for 1000

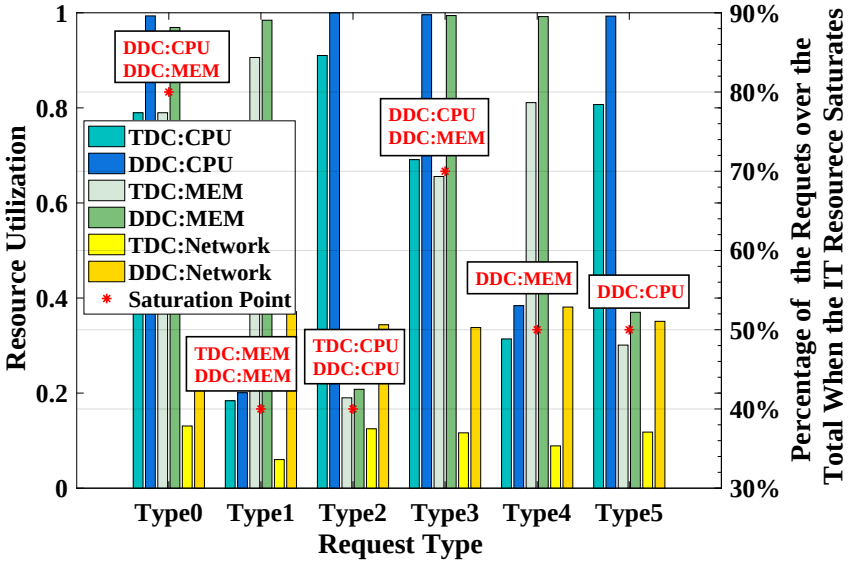


Figure 5.7: Comparison of the maximal resource utilization

requests in DDC comparing with the latency in the TDC (only Tx & Rx latency). It can be seen that the latency of DDC with Type 0 and Type 3 requests are relatively higher than the others. The reason is that only for these two cases, both CPU and memory resources are nearly fully occupied. It leads the algorithm more likely to find the IT resources in different racks and then increase the latency.

Fig. 5.7 depicts IT and network resource utilization under six different input requests. It can be observed that at least one type of IT resources could be fully occupied in DDC for all schemes. On the contrary, CPU and memory can hardly reach 90% utilization in TDC, which indicates that DDC could save IT resources. However, network (bandwidth) utilization in DDC is much higher than that in TDC. As seen, CPU utilization and memory utilization are 99% and 97%, respectively in DDC after processing 1000 Type 0 requests, but the value is 79% for both of them in TDC. However, 31% of network resources have been used in DDC, while the network utilization in TDC is 13%. For the network with Type 2 requests, since CPUs are largely demanded, CPU utilization saturates rapidly (40% of the total requests loaded). And the same outcome for memory utilization could be seen in the figure for the similar reason (Type 1). By combining previous two types of requests (Type 3), both of the CPU utilization and memory utilization could reach nearly 100% in DDC. However, in TDC, relatively lower CPU and memory utilization (approximately 70%) are observed while comparing to that with Type 0 requests. This is because when a server in TDC serves a request with high CPU demands, most of the RAM resources in this slot will be wasted. On the other hand, a waste of CPU resources occurs while serving requests with high RAM demands. Systems

with Type 4 and Type 5 requests provide similar performance to the systems with Type 1 and Type 2 requests, respectively. However, tiny differences are observed attributing to the slight differences in the requirements. Since adequate network resources (bandwidth) are provided, the network utilization is much lower than CPU utilization or memory utilization in all conditions.

5.5 Cost Model for Evaluating Network Gain

As mentioned earlier, DDC can save CPU and memory resources. However, it introduces additional cost in bandwidth and latency since CPU and memory are not directly connected. To clarify the benefits of disaggregation comprehensively, a cost model has been proposed by modifying the model reported in [216] to quantify the cost of latency, cost of bandwidth, CPU saving and memory saving due to disaggregation, which can be expressed as:

$$G = CS + MS - (CN + CL) \tag{5.2}$$

Where G is the total gain of the disaggregated system on the traditional system, CS and MS are CPU saving and memory saving achieved from pooling CPU and memory resources, respectively. CN is the cost of additional network resources, and CL is the cost of increased latency between CPU and memory.

5.5.1 CPU saving and memory saving

As stated in the former analysis, IT resource utilization of DDC is always higher than that of TDC. The unused IT resources in TDC could be regarded as a waste of money in hardware purchasing, thus the IT resource utilization differences (CPU and memory) stands for hardware saving due to disaggregation. In this work, CS and MS in DDC can be calculated with the equations below.

$$CS = CPU\ Utilization\ Difference \times Total\ Amount\ of\ CPU \times CPU\ Unit\ Price \tag{5.3}$$

$$MS = MEM\ Utilization\ Difference \times Total\ Amount\ of\ MEM \times MEM\ Unit\ Price \tag{5.4}$$

Where the utilization differences for all the resources are shown in Fig. 5.8(a), which are obtained from Fig. 5.7, and the total amount of resources are given in Table 5.1.

Figure. 5.8(b) illustrates the CPU and memory savings due to disaggregation,

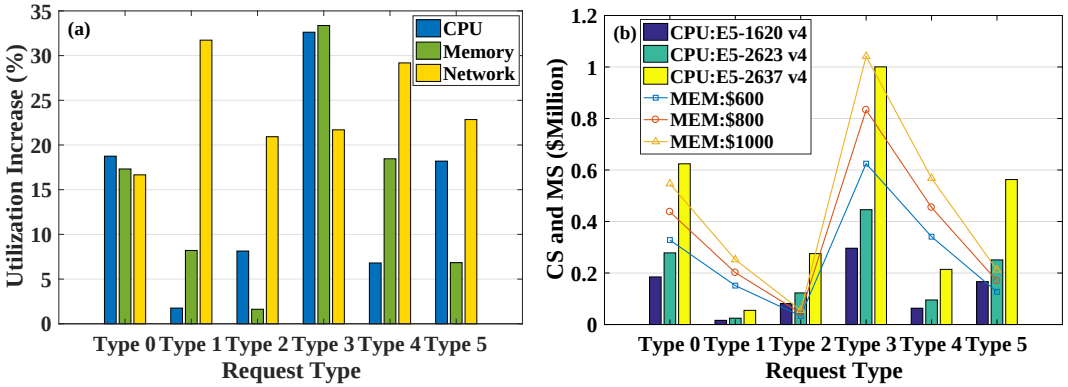


Figure 5.8: (a) Utilization increase and (b) CPU and memory savings of DDC over TDC

where three models of CPU (i.e. E5-1620 v4, E5-2623 v4 and E5-2637 v4) are adopted from Intel ARK [217] to investigated the influence of CPU prices on the total gain. The unit price of them are $295, $444 and $999, respectively. Moreover, three different unit prices, i.e. $600, $800 and $1000, for 4GB hybrid memory cube (HMC) [218] are assumed. It is transparent that a high CPU and/or memory utilization difference results in high CPU and/or memory saving, such as data center with Type 3 requests. Additionally, when high price CPU and/or memory model is adopted, the corresponding saving would be high. In summary, high IT resource utilization difference or high unit price will lead to high saving.

5.5.2 Cost of networks

In DDCs, traffic between CPU and memory moves to the whole DCN scale, which requires high bandwidth supporting from the networks. This section calculates the extra Cost of Networks (*CN*) of DDC comparing to TDC. Researchers in [219] has proposed a method to quantify the cost of bandwidth for HPC networks. Based on their research, network cost is calculated by summing up the price of the network equipment. In this study, three types of network equipment are considered, including Tx & Rx, optical switch and network interface card (NIC). The price of optical switches can be quantified by the cost of ports. The price of NIC is $250 per channel, however, NIC is needed only in TDC between each server and network, while other two equipment are used in both of the data centers.

In order to investigate the resource utilization in these two types of data centers in the former simulations, adequate and identical network resources (channels or bandwidth) were provided. In this section, the network resources are decreased in the two data centers according to their performance under each type of input requests. As stated in Table 5.2, there are three levels of channels in each data center, including dROSM-dROSM (ToR-ToR) channels, dROSM-dBOSM

Table 5.4: Optimized channel numbers under different inputs

Parameter	Channel Level	Original Number	Type 0	Type 1	Type 2	Type 3	Type 4	Type 5
DDC	dROSM-dROSM	16	7	2	2	8	2	2
	dROSM-dBOSM	16	11	3	5	11	4	5
	dBOSM-dBrick	8	8	4	5	7	4	5
TDC	ToR-ToR	16	4	1	4	3	2	3
	ToR-ToB	16	6	2	5	3	2	5
	ToB-brick	8	1	1	2	2	1	2

(ToR-ToB) channels and dBOSM-dBrick (ToB-brick) channels. To begin with, the numbers of dROSM-dBOSM (ToR-ToB) and dBOSM-dBrick (ToB-brick) channels are fixed, while the number of dROSM-dROSM (ToR-ToR) channels are being decreased one by one until blocking due to bandwidth unavailability occurs. The minimal number before the first blocking occurs is recorded. For example, the first blocking due to bandwidth occurs when the dROSM-dROSM (ToR-ToR) channel number is decreased to 6, then "7" is the new dROSM-dROSM (ToR-ToR) channel number. Subsequently, the same method are utilized to obtain the new dROSM-dBOSM (ToR-ToB) channel number and dBOSM-dBrick (ToB-brick) channel number. With this approach, we guarantee the performance of the data center with the minimum provision of network resources. Since the channel numbers go down, the costs for Tx & Rx, optical switches and NICs drops accordingly. Table 5.4 illustrates the optimized channel numbers under each types of input requests. Obviously, channel numbers in DDC are larger than that in TDC, and thus the increased *CN* in DDC could be calculated.

In the above analysis, all the switches are optical switches, and the cost of optical switches are measured by the number of the ports. To make the research more comprehensive, some optical switches are replaced with electrical switches and the networks cost are recalculated. Specifically, as there are two dBOSM (ToB) switches in each dBox (blade), one of them will be replaced with an electrical switch, which makes the network hybrid. It is assumed that the price of a 16x16 electrical switch is $1200. Since the channel numbers are different under each types of input requests (Table 5.4), the price of the electrical switch changes according to the channels needed.

The final *CN* is presented in Fig. 5.9, including all optical situation and hybrid (half-optical & half-electrical) situation. To make a comparison between these two situations, the Tx & Rx price per Gb/s as well as the optical switch port price are increased from $0.5\$50 to $3\$300 gradually, but the price for electrical switches

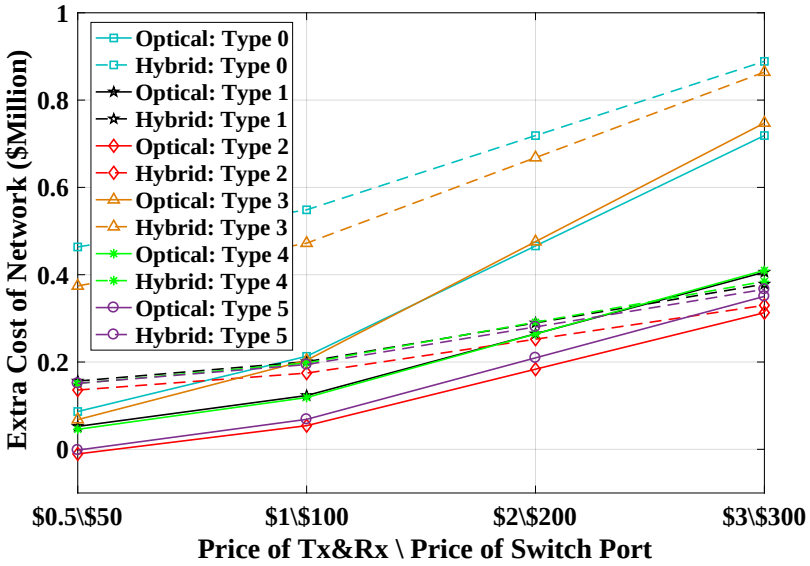


Figure 5.9: Extra cost of networks (*CN*) owning to disaggregation

and for NICs are constant. As shown in the figure, *CN* in all optical situations are lower than that in hybrid situations, however, with the increase of Tx & Rx and optical switch port price, the *CN* differences between all optical and hybrid shcemes decrease gradually. In terms of input request type, the costs of DCNs with Type 0 and Type 3 requests are relatively higher than others.

5.5.3 Cost of latency

Latency is a vital issue in DDC which influences the performance of the whole system. Since CPU and memory could be distributed in different bricks, there could be extra latency between CPU and memory. Resource allocation algorithm plays a critical role in latency decreasing. A well-designed allocation algorithm could usually select the most suitable IT resources and find the shortest path between them. Recently, a 2-approximation algorithm has been proposed to optimize the distance and minimize the latency between selected data centres [220]. Also, a performance-aware VM allocation approach has been designed to improve the VM bandwidth and network delay by tracking the performance of the applications [221]. Furthermore, a delay prediction mechanism [222] for the optimization of VM allocation represents reduced latency while compared to other approaches.

To measure the extra cost of latency (*CL*) in DDC, the model reported in [216] is utilized for calculation. In this model, the increased latency could be compensated by processors with higher performance (more expensive). In other words, *CL* could be presented as an increase of CPU price. Six benchmarks were used in [216] to present the relationship between the increased latency and CPU

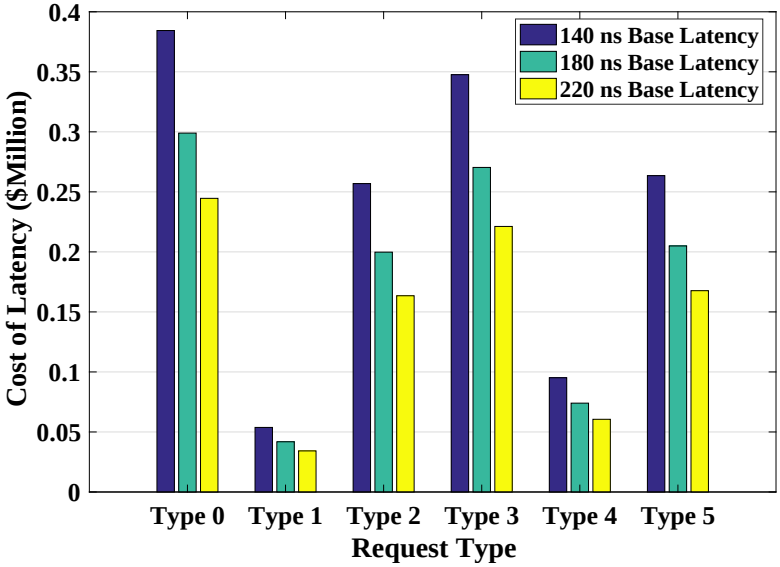


Figure 5.10: Extra cost of latency (*CL*) owning to disaggregation

price. In this work,one of the benchmarks (SPEC-FP on single threaded cores) is chosen for *CL* calculating. As disaggregated storage is utilized both in DDC as well as in TDC, only the latency between CPU and memory is considered. It is assumed that all the latency in TDC results from Tx & Rx (base latency). Thus, latency from optical switches and channels between CPU and memory in DDC is the extra part. According to the model, *CL* for each CPU unit can be expressed as:

$$Y = 310 \times X \tag{5.5}$$

Where Y is the increased price for each CPU unit (as a compensation for increased latency) and *310* is a coefficient in the chosen model (i.e. SPEC-FP on single threaded cores). X indicates the percentage of increased latency due to disaggregation comparing with TDC, which can be expressed as:

$$X = Increased\ Latency / Base\ Latency \tag{5.6}$$

In which, increased latency in DDC can be obtained from Fig. 5.6(b). Combining all the information above, *CL* can be calculated by:

$$CL = Y \times CPU\ Utilization \times Total\ Amount\ of\ CPU \tag{5.7}$$

Fig. 5.10 presents the *CL* under six types of input requests and three different values of HMC base latency, i.e. 140 ns, 180 ns and 220 ns. Since CPU utilization is a critical factor in calculating *CL*, the costs are relatively lower in low CPU utilization conditions such as DDC with Type 1 and Type 4 requests. In terms of

base latency, a decrease of *CL* could be observed with the increase of base latency. This is because that the value of X becomes small when base latency is high and then results in a decrease of *CL*.

5.5.4 Total gain from disaggregation

As aforementioned, the total gain contributed by disaggregation can be calculated by using Eq. (5.2). Since *CS*, *MS*, *CN* and *CL* have been calculated in the former analysis, the total gain could be obtained easily. Fig. 5.11 illustrates the total gain of the DDC adopting different CPU, memory, network prices and base latencies for each type of resources. In the figure, a benchmark price for each type of resources is set, which are that a CPU unit (E5-2637 v4) costs $996 and a HMC unit costs $600. As for the networks, all optical situation is used. The price of the Tx & Rx and optical port are $1 per Gb/s and $100, respectively. In addition, the HMC base latency is 180 ns. Based on the above common setup, one variable is set in each figure to investigate the relationship between a specific type of resource and the total gain.

It can be seen in Figs. 5.11(a) and 5.11(b) that for almost all cases the total gains are positive, which means that the proposed DDC architecture can save money. Especially, in high CPU or memory utilization difference situation, e.g. the

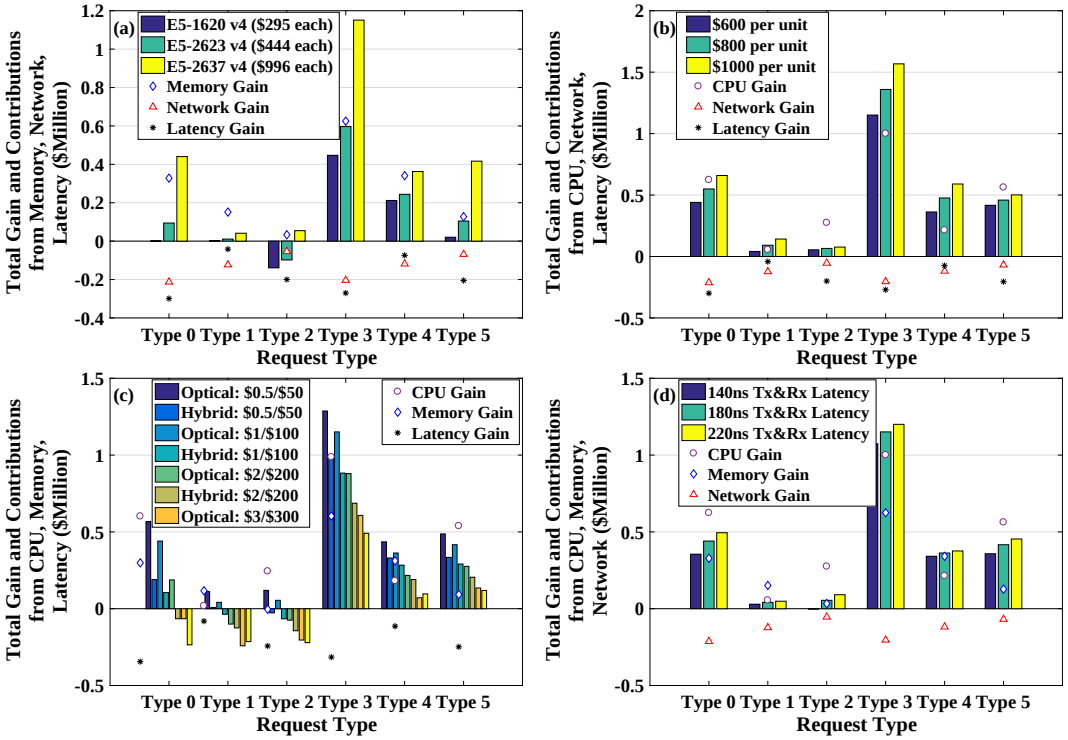


Figure 5.11: Effects of (a) CPU price, (b) memory price, (c) network price and (d) base latency on the total gain

network with Type 0 or Type 3 requests, large gains are achieved while using different processors or HMCs. In terms of networks [Fig. 5.11(c)], more positive total gains are achieved in DDC, especially for network with Type 3, 4 and 5 requests. However, the total gain decreases when the network equipment price raises. Moreover, total gain in all optical situations are higher than that in hybrid situations when the network equipment are cheap, but the reverse results occur with higher equipment price. Similar to Fig. 5.11(b), all the cases obtain positive total gains in Fig. 5.11(d). It is likely that the impact of latency is not as significant as the influence of networks. Furthermore, it poses little influence on the total gain in low CPU utilization types, in particular, when Type 1 and 4 requests are considered.

5.6 Scalability of the DDC Architecture

In the previous sections, the proposed DDC architecture has presented significant benefits over the TDC architecture, including blocking, resource utilization and cost-efficiency. Scalability is another important parameter need to be investigated for the new architecture. In this work, the scalability is evaluated by analyzing the relationship between the variables, including the number of ports per dBrick (PPB), the number of dBricks per dBox (BPB), the number of dBoxes per dRack (BPR), the number of dRacks, the port number of the optical switches (PPS) used and the (over)-subscription ratio, and the number of dBOSM, dROSM and dDOSM switches required in DDC architecture. Mathematically, the correlations between them can be expressed as:

$$dBOSM\ Number\ per\ dBox = PPB \times BPB \times (1 + \frac{1}{N})/PPS \tag{5.8}$$

$$dROSM\ Number\ for\ Single\ dRack = PPB \times BPB \times \frac{1}{N} \times BPR/PPS \tag{5.9}$$

$$dDOSM\ Number = PPB \times\ BPB \times \frac{1}{N} \times BPR \times \frac{1}{M} \times dRack\ Number/PPS \tag{5.10}$$

Where N (N $>$ 1) and M (M $>$ 1) are the (over)-subscription ratios at the first and second tiers, respectively.

The effect of the (over)-subscription ratio on the number of dROSMs required within a single dRack is shown in Fig. 5.12(a). It can be easily found that when the port numbers of the dROSM switches and dBricks are fixed, the number of

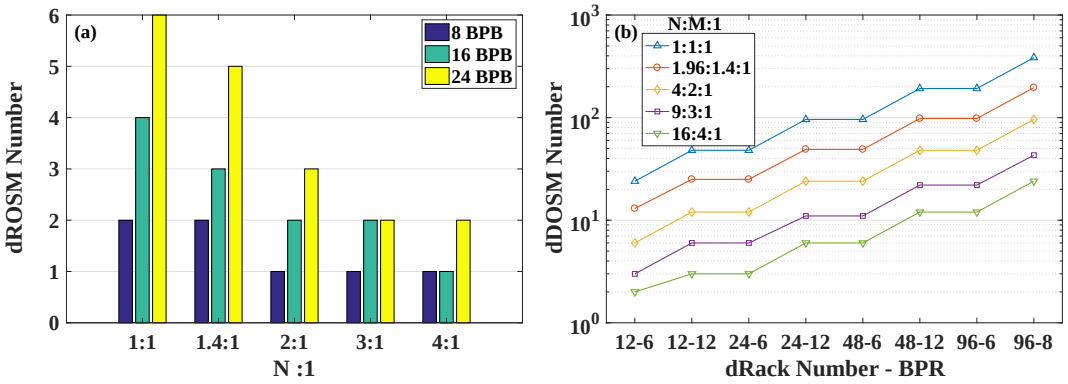


Figure 5.12: (a) dROSM number for single dRack and (b) overall dDOSM number for data center with multi-dRacks (384*384 switches, 16 ports per dBrick, 16 dBricks per dBox)

dROSMs needed is proportional to the number of dBricks per dBox and inversely proportional to the (over)-subscription ratio. That is to say, the higher the (over)-subscription ratio, the less the dROSMs are required for a single dRack. Figure. 5.12(b) illustrates the numbers of dDOSMs needed for a data center cluster with multiple dRacks, i.e. 12, 24, 48 and 96 dRacks, where the dROSM switches are connected to the dDOSM switches with a spine-leaf topology. Similar to the number of dROSMs in a single dRack, the number of dDOSMs required is inversely proportional to the (over)-subscription ratio. Moreover, results indicate that larger scale data centers with greater number of dBoxes per dRack or/and dRacks can be easily supported by adding more switches (i.e. dBOSM, dROSM and dDOSM switches).

5.7 Summary and Conclusion

A disaggregated data center architecture, i.e. dReDBox architecture, is proposed in this chapter. Moreover, three dRack structures and four resource allocation algorithms are developed and compared. In terms of the DDC architecture, the one with heterogeneous dRacks housing heterogeneous dBoxes presents the best performance. As for the resource allocation algorithm, the network-aware locality based resource allocation algorithm outperforms other three algorithms in DDC. Comparing with TDC, DDC can save IT (i.e. CPU, memory) resources at the cost of additional bandwidth and latency. Besides, a new cost model is developed and utilized to investigate and compare the cost efficiency of the data centers under six different application requirements. Results show that, DDC is more cost efficient than TDC (provides positive total gain) in most of the situations. In addition, CPU/memory price plays an vital role in the total gain, a high CPU/memory price can largely boost the CPU/memory saving in high CPU/memory utilization

conditions, however, high CPU/memory price also indicates high budget on the whole system. Cost of network varies with the prices of network equipment. High network equipment price may decrease the cost efficiency of the disaggregated data center, which even may lead the total gain to be negative. The impact of the base latency is not as significant as CPU, memory or networks. To sum up, CPU saving, memory saving or cost of network changes rapidly when applying different types of processors, different prices of memory or network equipment, but the effect of the cost of latency is more stable. It is believed that this promising cost effective and scalable disaggregated data center architecture will be widely applied in the future data center deployments.

Chapter 6

Scalable Topologies for Disaggregated Data Center

6.1 Introduction

As previously described in Chapter 5, DDC architecture has been proposed to overcome the shortcomings of TDC architecture, in terms of flexibility, resource utilization and adaptability. However, such architecture still presents a number of fundamental challenges that need to be overcome, e.g. it requires lower latency compared to the traditional direct-attached modular, the cost and power consumption need to be reduced whilst supporting a substantially higher bandwidth and bandwidth density, and the substrate and orchestration should enable the system to support various specific resource connections (e.g. CPU-memory, memory-storage) at low latency and cost. Therefore, the dReDBox architecture was proposed showing advantages in modularity, scalability and IT resource utilization maximization. As seen in Fig. 6.1 and aforementioned in Chapter 5, the dReDBox architecture consists of multiple dRacks housing multiple interconnected dBoxes. Since the structure of heterogeneous dRack with and heterogeneous dBoxes can provide the best performance in Chapter 5, it is adopted in this work. In addition, it uses a 3-tier fat tree topology (non-parallel) consisting of several spine-leaf topologies interconnecting the dBricks, dBOSMs, dROSMs and dDOSM. As a consequence, the communication between two end nodes may over up to 5 hops indicating a relative high network latency. Moreover, since the switches at higher tiers need to connect all lower-tier nodes, the bandwidth of the link or the port count of the optical switch should be high, i.e 384x384 switches, which may reduce the cost-efficiency of the system. To realize a higher scalable and modular DDC architecture with ultra-low latency, high capacity and cost/power-efficiency, new interconnects or topologies are required.

In this chapter, two parallel topologies for disaggregated data center network

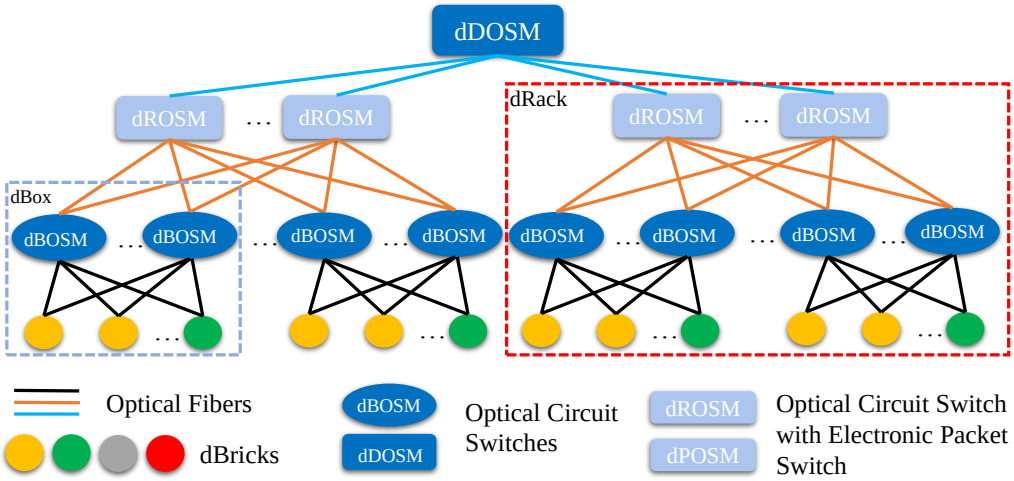


Figure 6.1: Disaggregated data center architecture with non-parallel topology

are proposed. Subsequently, the performance of the three topologies is evaluated and compared in terms of blocking probability, network and IT resources utilization, round-trip latency, switch cost as well as power consumption. Moreover, one of the parallel topologies is then adopted to explore the feasibility of a MCF-SMF hybrid optical network for disaggregated data centers. The benefit of introducing MCFs to the DDC architecture in power consumption is also evaluated in detail.

6.2 Proposed Parallel Topologies for DDC Network

The two proposed parallel topologies, i.e. Box-modular and Brick-modular topologies are depicted in Fig. 6.2 and Fig. 6.3, respectively. Compared to the non-parallel topology shown in Fig. 6.1, both the two topologies have only two

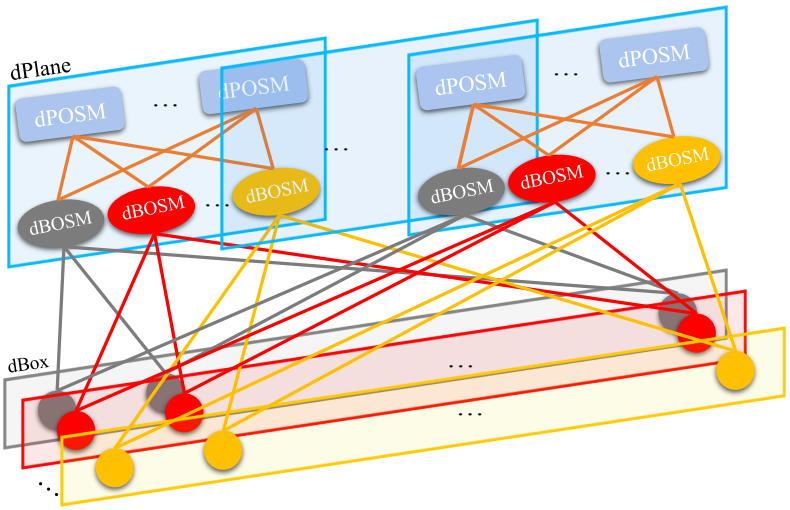


Figure 6.2: Disaggregated data center architecture with Box-modular topology

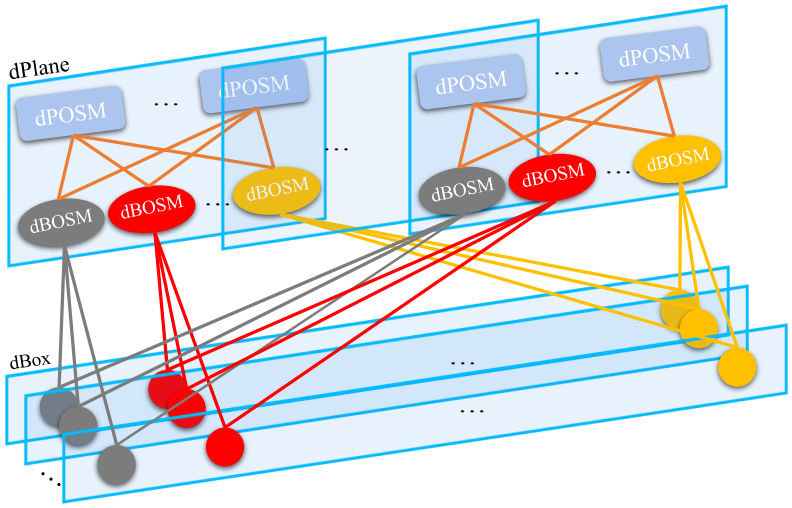


Figure 6.3: Disaggregated data center architecture with Brick-modular topology

tiers of switches, which can minimize the distance of the routed paths for communication between any two dBricks and in turn reducing network latency. As seen, in order to realize the structure, multiple disaggregated planes (dPlanes) are introduced and each of the dPlanes comprises two tiers of optical modules, i.e. disaggregated plane optical switch modules (dPOSMs) and dBOSMs. The dBOSMs are linked to the dPOSMs with a spine-leaf topology, while there is no link between the same type of modules. Since the dPlanes are not connected to each other and the dBOSMs in each dPlane connects all the dBricks to enable any to any dBrick communication via any individual dPlane, the parallel architecture is realized. In addition, every dPlane can support and switch one spatial channel with either single or multiple wavelengths per dBrick, which makes it a capacity modular parallel topology.

It can be found that the only difference between the two parallel topologies is the way how the dBricks are connected to the dBOSMs. In the Box-modular topology as shown in Fig. 6.2, each dBOSM in a dPlane is only responsible for connecting all the dBricks in one dBox, which ensures that the communication between any two dBricks in the same dBox can be achieved via one-hop (dBrick-dBOSM-dBrick) transmission. For example, all the dBricks in the yellow dBox are only connected to the yellow dBOSM in each dPlane. On the contrary, for the communication between any two dBricks in the different dBoxes, 3-hops (dBrick-dBOSM-dPOSM-dBOSM-dBrick) transmission is required. In contrast, in the Brick-modular topology as depicted in Fig. 6.3, each dBOSM in a dPlane connects only one dBrick in each dBox. The dBricks connected to the same dBOSM are placed at the same position of the dBoxes. As a consequence,

intra-dBox communications inside this topology requires both dBOSM and dPOSM switches, while the inter-dBox transmission can be easily executed via a dBOSM switch.

The number of dPlanes in both the two parallel topologies is determined by the physical channel numbers per dBrick and per link, i.e. channel number per dBrick / channel number per link. In the dBox-modular topology, the number of dBOSMs in a single dPlane is same as the number of dBoxes in the architecture, while it equals to the number of dBricks in the dBrick-modular topology. Therefore, the DDC architectures with these two topologies can be easily scaled by adding the dBOSMs. Moreover, small port count (e.g. 96 ports) switches rather than the ones with high port-count are demanded for the architecture realization, indicating the benefits that these parallel topologies bring, including scalability, flexibility and switch port utilization.

6.3 Configurations for the Architectures

The same Matlab simulator described in Section 5.3 is utilized in this work to evaluate and compare the performance of the proposed architectures. As aforementioned, the overall simulation procedure consists of four main stages: request generation, IT resource allocation, network resource allocation and connection establishment. The configurations of the architectures are shown in Table. 6.1 and it should be noted that the parameters that are not presented in the table, e.g. CPU/memory unit number per dBrick, are same as the ones described in Section 5.3. Heterogeneous dBoxes are adopted for all the architectures and each of the DCNs has 1/3 ratio of compute, memory and storage resources, respectively. Moreover, to simplify the fabric, each distinct resource element is configured at same position inside each dBox. In terms of the allocation process, NALB resource allocation algorithm, which shows the best performance in Chapter 5, is

Table 6.1: Simulation configurations for the architectures

Architecture	Non-parallel	Box-modular	Brick-modular
dRacks	4	-	-
dPlanes	-	8	8
dBoxes	6 per dRack	24	8
dBricks per dBox	8	8	24
dBrick-dBOSM	0.25 m	0.25 m	0.25 m
dBOSM-dR/POSM	3 m	3 m	3 m
dROSM-dDOSM	10 m	-	-
8 I/Os per dBrick @ 25Gb/s		5ns pass-through latency per switch	

utilized in the simulations.

6.4 Comparison between the Architectures

6.4.1 Network behavior

As shown in Fig. 6.4, the network behaviors of the architectures in terms of network resource utilization and blocking probability after processing 1000 VM requests are first evaluated and compared. It can be seen that all the three architectures provide the similar blocking probability performance, which is attributing to that all of them can support any to any connectivity between the end nodes, in addition, sufficient network resources are offered to the links. That is to say, almost all the blocking is resulted from the IT source unavailability, which has been proven by Fig. 6.5 that the first blocked request (around the 400th request) occurs when the IT sources almost saturate. In terms of the network utilization, the Box-modular topology outperforms the others, since it requires the least bandwidth to achieve the same performance in terms of blocking probability (Fig. 6.4) and IT resource utilization (Fig. 6.5) as the other two topologies. Particularly, after processing 1000 VM requests, it demands 9% and 27% less bandwidth resource than that of the non-parallel and Brick-modular topology, respectively. In contrast, the Brick-modular requires the most network resource, which will be explained by Fig. 6.6.

Figure. 6.6 presents the CDF of the round-trip latency and the average latency of the three topologies after proceeding 1000 VM requests. As shown, both the two parallel topologies guarantee less than 100 ns round-trip latency between the end nodes (i.e. CPU, memory, storage dBricks) on all established VMs, while the

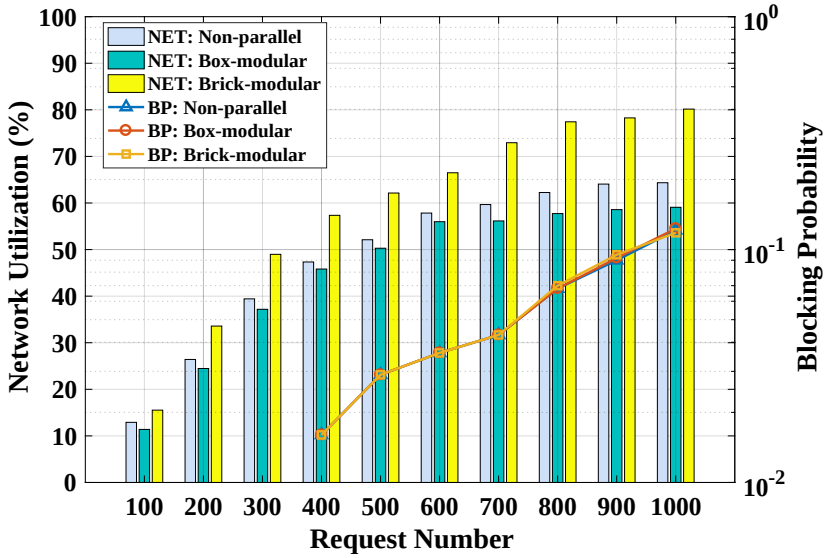


Figure 6.4: Overall comparison of network behavior for three architectures

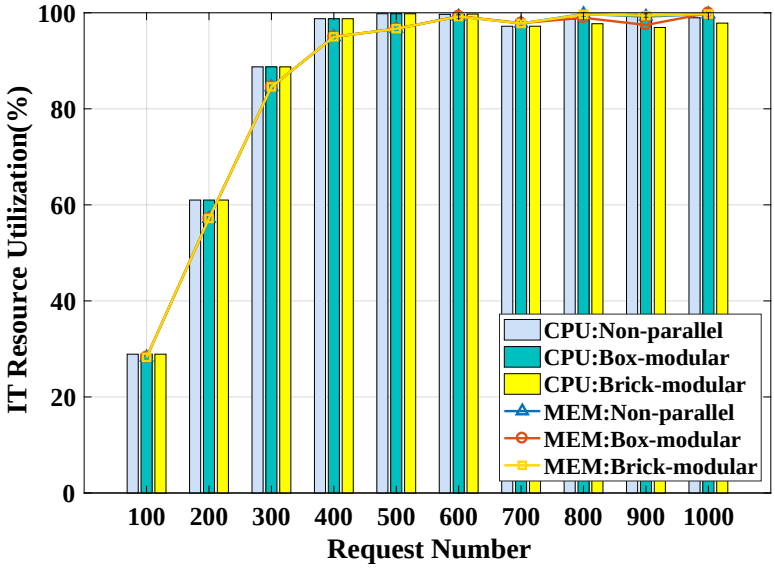


Figure 6.5: Overall comparison of IT resource utilization for three architectures

non-parallel presents up to 315 nm latency. The latency CDF indicates that for the non-parallel and Box-modular topologies, up to over 70% of the total traffic is inside the dBoxes (1-hop transmission) since different types of IT resources are configured. On the contrary, as the intra-dBox communication between different type of resources (e.g., CPU-memory, memory-storage) requires the participation of the dPOSM switch in the dBrick-modular (3-hops transmission), more network resources are required and higher average latency can be observed. Compared the non-parallel topology with the Box-modular topology, as 8% of the communication between IT resources is completed with the involvement of five switches (i.e. two

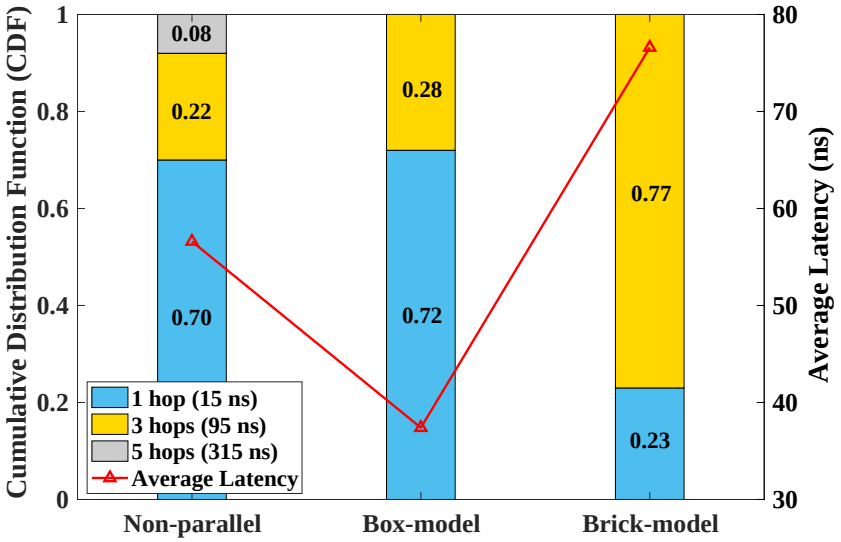


Figure 6.6: Overall comparison of IT round-trip latency for three architectures

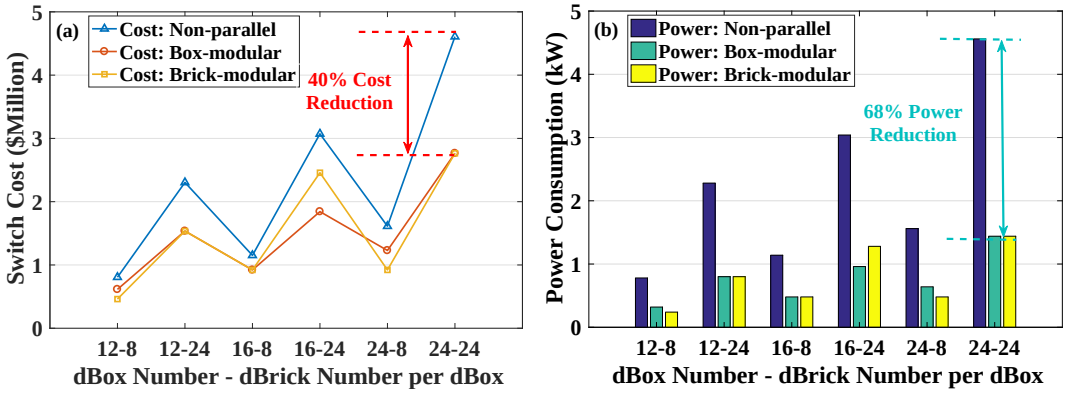


Figure 6.7: Overall comparison of (a) switch cost and (b) power consumption for three architectures

dBOSM, two dROSM and one dDOSM switches), 34% higher average latency than that of the Box-modular topology is observed.

6.4.2 Switch cost and power consumption

Subsequently, the switch cost and power consumption of the three topologies are compared and presented in Fig. 6.7. Since the topologies can be easily scaled by adding dBoxes and/or dBricks with the corresponding optical switches, different combinations of dBoxes and dBricks are considered. Moreover, it is assumed that each optical switch port costs 100 dollars, while the power consumption of a 96-ports switch and a 384x384 switch is 5 W and 100 W, respectively. As shown, both the two proposed parallel topologies are more cost-effective and power-effective than the non-parallel topology. Especially, for a data center with 24 dBoxes and each of them houses 24 dBricks, i.e. 576 dBricks in total, they can achieve 40% and 68% reductions on switch cost and power consumption, respectively. As for the two parallel topologies, the dBrick-modular topology outperforms the dBox-modular one for some scenarios (e.g. 16-24), and the opposite results for some other scenarios (e.g. 12-8 and 16-24), indicating that the topologies are suitable for data centers with different deployment requirements.

6.5 MCF-SMF Hybrid Architecture for DDC

6.5.1 Proposed hybrid architecture

As aforementioned in the previous chapters and sections, the transmission latency and link bandwidth density are two of the main challenges that the disaggregated data centers have faced. The proposed parallel topologies enable the transmission latency to be lower than 100 ns, which have significantly relieved the pressure on the latency. That is to say, the issue remain is the link bandwidth density.

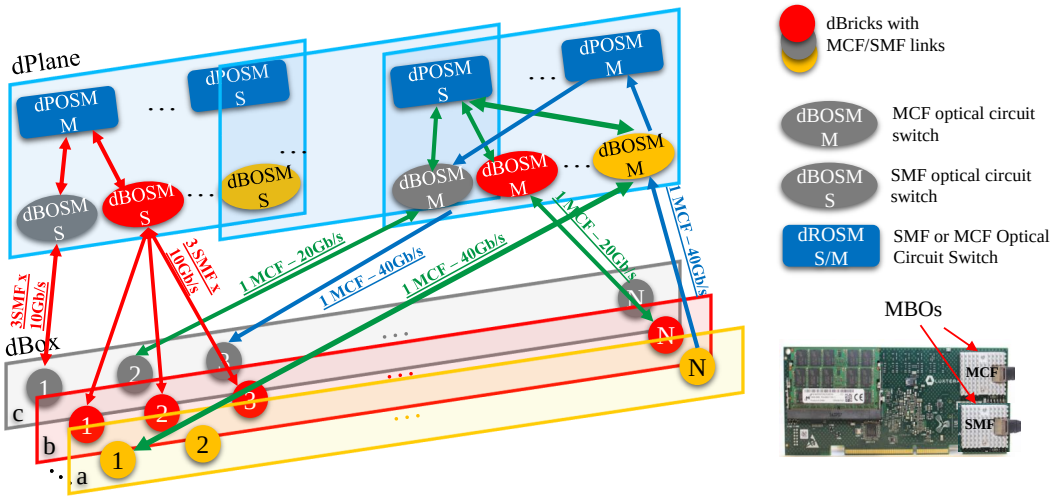


Figure 6.8: MCF-SMF hybrid architecture

According to the investigation in Chapter 4, MCF-based SDM technique is a perfect candidate to address the capacity and potential switch density requirements for the data center disaggregation. Therefore, a novel MCF-SMF hybrid DDC architecture is proposed and presented in Fig. 6.8. It can be found that this new architecture is developed based on the Box-modular topology since it offers the best performance in terms of both network behavior and transmission latency among the three topologies compared in Section 6.4. To realize the MCF transmission, MCF switches are configured at both the dBOSMs and dPOSMs. Moreover, high bandwidth multi transceiver silicon photonic mid board optics (MBOs) for SMFs and MCFs are integrated to the dBricks. This hybrid architecture enables dBricks to perform CPU-memory communication across many dBoxes, whilst preserving architectural flexibility/ VM diversity. Figure 6.8 also showcases how the architecture deals with different VM requests (CPU:memory), where the achievable capacity of each channel/core is assumed to 10 GB/s:

1. High performance VM (1:1): for a request demands one CPU (aN) and one memory (c3) with 40 Gb/s capacity, a pure 2-tier MCF transmission can be utilized, which has been demonstrated with the blue lines.

2. Memory dominant VM (1:many): for a request requires one CPU (c1) and multiple memories (b1, b2, b3) in same dBox with 10 Gb/s capacity, a hybrid SMF-MCF transmission is suitable, which has been demonstrated with the red lines; for a request requires one CPU (a1) and multiple memories (bN, c2) in different dBoxes with 20 Gb/s capacity, a pure 2-tier MCF transmission can be utilized, which has been demonstrated with the green lines.

3. CPU dominant VM (many:1): the connections for this case is same as the previous one as the positions of the CPU and memory dBricks can be swapped.

Obviously, this new architecture has all the advantages inherent in the Box-modular topology and the MCF-based SDM technique, such as low blocking, ultra-low latency, high flexibility, high bandwidth and front panel density. Moreover, the correct usage of ports (dPOSM and dBOSM selection) and the grooming and splitting in the hybrid 2-tier MCF-SMF network create a range of scenarios that the architecture can support, with reduced port counts and in turn, increases the power-efficiency or space-efficiency of the system.

6.5.2 Power consumption analysis

Figure. 6.9 showcases the power consumption reductions that can be achieved by the proposed hybrid architecture against a fully SMF-based architecture. For the calculations used, it is assumed that the architecture has 128 dBoxes. Each dBox consists of 16 dBricks, i.e. 8 CPU dBricks and 8 memory dBircks, and each dBricks has 16 individual optical channels. According to [223], the power consumption of each dBox, CPU dBrick and memory dBirck is 35 W, 19 W and 23 W, respectively. Moreover, to support a dBrick with 16 channels, 6 W of power is consumed by the MBOs [224]. In order to provide a comprehensive investigation on the hybrid architecture, different ratios of the MCF and SMF channels are considered at both the first and second tiers of the topology. Furthermore, the impact from the type (core number) of the employed MCF on the power consumption is also analyzed.

It can be seen in the figure that by introducing the MCF switches into the architecture, the power consumption can be considerably reduced in different magnitudes according to the ratios of SMF to MCF switches. Particularly, even if the MCF switches are only configured in the first tier (the second group), up to 50% power consumption reduction can be achieved. The reason behind is that all

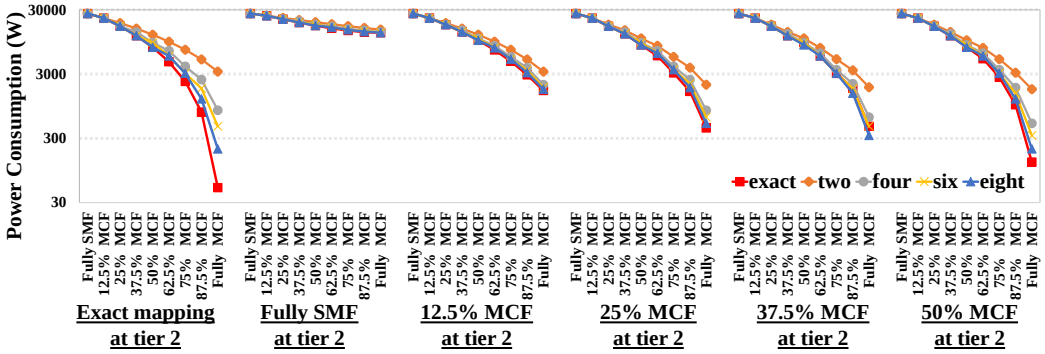


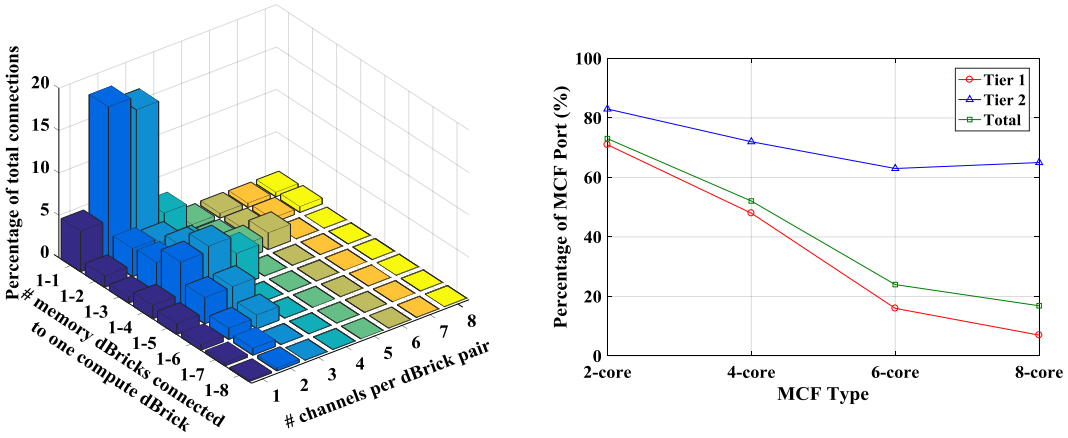
Figure 6.9: Total power consumption of the optical switching layer

the 16 channels from one dBrick can be supported by two 8-core MCFs (2 MCF ports) or four 4-core MCFs (4 MCF ports) rather than sixteen SMFs (16 SMF ports), while the fully connectivity can be realized by the SMFs at the second tier. Moreover, to further reduce the power consumption of the switching layer, the MCF switches should be also configured at the second tier, and the more the MCF switches are placed, the higher the power consumption reduction will be achieved. According to the results in Fig. 6.9, when half of the SMF switches at the second tier are replaced by the MCF switches (the sixth group), the reduction in power consumption will go up to 93-99%, however, the configurability of the architecture will become low. To find a great trade-off between the power consumption and the configurability is quite important for this architecture. Apart from the ratios between the switches, the type of the employed MCF can also affect the power consumption. As seen, for all the cases, with the increase of the core number in the MCF, the power consumption decreases since less switches ports are required. When the 16-core (exact) MCF is applied, over 99% power can be saved, however, this choice will result in the least amount of flexibility as it assumes all 16 channels on one dBrick will be directed to the other dBrick pair.

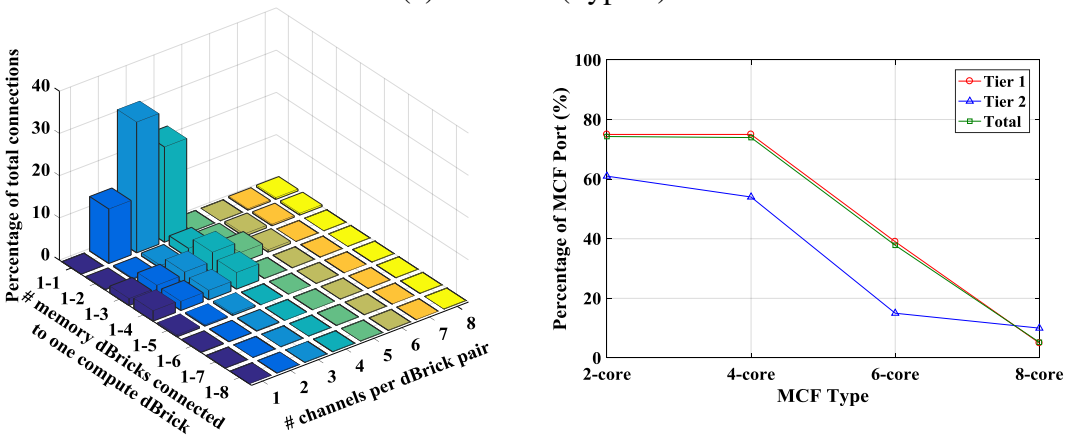
6.5.3 MCF-SMF port analysis

To evaluate the effectiveness of hybrid MCF-SMF architecture and explore how much MCF switches are demanded for different traffic requirements, the same network level simulator developed in Section 6.3 is utilized and the results are shown in Fig. 6.10. The configurations for these simulations are same as assumptions in Fig. 6.9, i.e. there are 128 dBoxes and each dBox houses 8 CPU dBricks and 8 memory dBricks. In addition, each CPU dBrick is assumed to have 64 CPU cores and each CPU dBrick contains 64 GB random access memory. To analyze the situation for different workload scenarios, three types of request described in Table 5.3 are taken into consideration, 1) random request (Type 0): 1–32 CPU cores and 1-32 GB memory; 2) high CPU request (Type 1): 24–32 CPU cores and 1-8 GB memory; 3) high memory request (Type 2): 1–8 CPU cores and 24-32 GB memory. Other configurations, assumptions, requirements and the simulation process are completely the same simulator illustrated in Section 6.3.

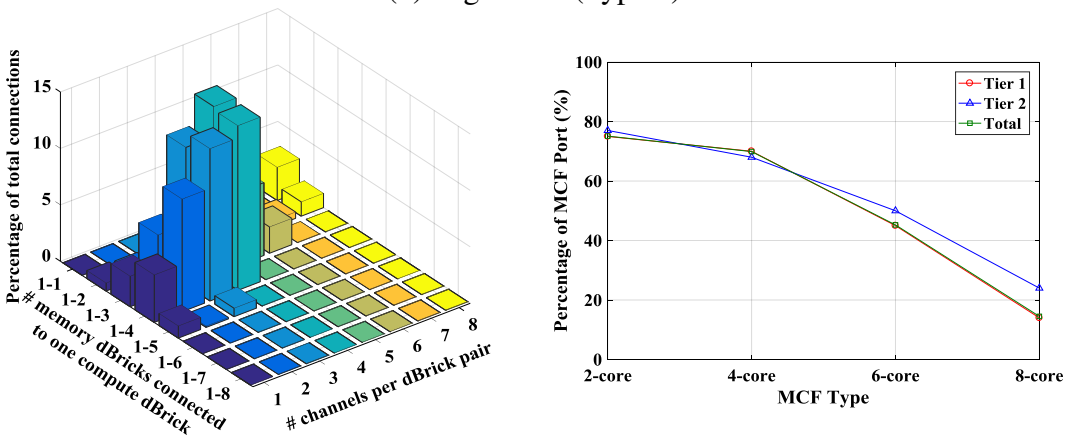
As seen, the left column of Fig. 6.10 presents the distribution of the number (from one to eight) of memory dBricks connected to one CPU dBricks and the number of optical channels required between any two dBricks. It can be found that with the increase of the demand on memory, the percentage of the cases that more than one memory dBricks are connected to one CPU dBrick increases accordingly. Especially, for the high memory workload scenario, the cases of only 1 memory



(a) Random (Type 0)



(b) High CPU (Type 1)



(c) High Memory (Type 2)

Figure 6.10: Distribution of connections between CPU and memory dBricks (left column) and required MCF/SMF links in the first and second tiers (right column) for a) Random, b) High CPU and c) High memory request scenarios

dBrick being connected to 1 CPU dBrick contribute marginally to total connections. As for the number of channels required between the dBrick pairs, the extreme case, i.e. only one channel is required between a CPU and memory dBrick, accounts for

only 10.9%, 4.9% and 8.7% of total connections required in the random, high CPU and high memory scenarios, respectively. On the contrary, the cases that require 2-4 optical channels between the dBrick pairs account for 71.77%, 86.28%, 74.82% of total connections after processing the random, high CPU and high memory requests, respectively. For these cases, MCF can be applied to significantly reduce the port number of the switch, and in turn, the spatial density and power consumption.

The right column indicates the the percentage of MCF connections/ports required for each workload scenario, with the consideration of a variety of MCF types. In order to reduce the port wastage, it is assumed that only when the required optical channel number is more than one and not smaller than half of the MCF core number, a MCF port is configured. For example, when the 6-core MCF is considered, the MCF port will be configured only when the channel number required is more than two, otherwise, one or two SMF ports are configured. Therefore, with the increase of the core number in the MCF, the percentage of of the MCF port required decreases. It can be seen that 2-core and 4-core MCFs are quite desirable for all the three cases (60%-80% of the ports can be MCF ports), while for higher core count MCFs, e.g. 6-core and 8-core MCFs, the high memory one can benefit the most, attributing to that multiple memory dBricks are required to attach to one CPU dBrick, thus requiring a larger mount of links between a dBrick pair. These figures also show the resultant percentage of MCFs in different tiers. As for the high CPU and memory scenarios, over 95% of the traffic happens inside the dBoxes (83% for the random request scenario), the total percentage of MCF ports needed in Figs. 6.10(b) and (c) is close to the value at Tier 1, respectively. By comparing these results to Fig. 6.9, the potential reductions in power consumption for all scenarios can be easily obtained. For instance, for the random scenario with the integration of 2-core MCFs, nearly 74% of power consumption reduction at the switching layer can be achieved. As for the high CPU scenario with the employment of 2-core or 4-core MCFs, up to 68 and 87% power consumption reduction can be achieved, respectively. In contrast, for the high memory scenario, the deployment of 2-core, 6-core and 8-core MCFs may lead to 81%, 68% and 28% power consumption reduction at the switching layer compared to case only adopting SMFs. To sum up, the integration of 2-core MCF in the proposed architecture for all the investigated workload scenarios can reduce the power consumption at the switching layer by 68%-81%. It should be noted that, all the the power consumption reduction is calculated based on port number reduction, therefore, the same magnitude of benefit can be achieved on the perspective of space (i.e. space-efficiency) since the switches space is linearly proportional to the port number.

6.6 Summary and Conclusion

Two parallel topologies, Box-modular and Brick-modular, are first proposed in this chapter for disaggregated optical data centers. By comparing the topologies with the original dReDBox topology (non-parallel), both of them can reduce the cost by 40% and increase the power-efficiency by up to 68% for a data center with 576 dBricks. Moreover, they can guarantee the round-trip transmission latency between the CPU and memory dBricks within 95 ns while delivering the maximum IT resource utilization. Among which, the Box-modular topology is the best one since it can also lead to resource saving and latency decreasing against the non-parallel topology. However, as the dPlane number of the two topologies are constrained by the port number of each dBrick, the flexibility of the switching layers is limited. Moreover, with the Brick-modular topology, the data center can only by scaled by adding dBricks in each dBoxes. As a consequence, the scalability (scale of DCN) may be limited by the physical parameter, such as power consumption per dBox. The investigation on latency distribution also indicates that even original dReDBox architecture can scale to cluster level distances, attributing to that the latency-aware algorithms can deliver 95 ns round-trip transmission latency for 92% of the VMs. Subsequently, a MCF-SMF hybrid DDC architecture is developed based on the Box-modular topology, which has the potential to show advantages in terms of network capacity, flexibility, scalability, bandwidth and front panel density. Comparing to the pure SMF-based architecture, the new architecture can achieve up to 68%-81% power/space efficiency.

Chapter 7

Conclusions and Future Work

7.1 Conclusion

The rapid growth in data center traffic demands necessitates the introduction or development of new technologies, topologies and network architectures, which have the potential to enable the data center network (DCN) to provide high performance in terms of network capacity, resource utilization, front panel/bandwidth density, cabling complexity, flexibility, scalability while reducing the latency, cost and power consumption. To achieve these goals, two advanced technologies were exploited and investigated in DCNs, which are a) space division multiplexing (SDM) technique using multi-core fibers (MCFs) and b) optically disaggregated data center (DDC) architecture.

Chapter 3 and Chapter 4 focus on scaling the DCN using MCF-based SDM technique. Before introducing the MCFs into the DCNs, the behavior of inter-core crosstalk (IC-XT), which is the main constraint of MCF, was comprehensively studied in Chapter 3. A variety of experiments were executed and the obtained results were analyzed in depth by either using the existing IC-XT models or comparing with previous works that had been done by other researchers. It was found that both the static and dynamic IC-XT can be impacted by the signalling source, baud rate, modulation format, operational wavelength, temperature, PRBS pattern and the number of excited cores. Moreover, some of the results validated the accuracy of the theoretical models. Besides, when the IC-XT was measured with different power meter averaging times and measurement time windows, the accuracy of the observed IC-XT was found to be different. Therefore, the correlation between them was carefully explored. Finally, a new model on the IC-XT step distribution was proposed and elaborated in Section 3.6, which achieved up to 99.87% fitting accuracy to the experiment results. The sophisticated studies and characterisation of the IC-XT presented can greatly benefit the MCF research in the labs and applications in the MCF-based data centers, especially, the

novel distribution model provides an opportunity for future IC-XT classification using machine learning methods, which may contribute to the security of MCF-based DCNs.

In Chapter 4, a SDM-based DCN was developed with the consideration of the bi-directional transmission model, which could significantly reduce the IC-XT in MCF. To explore the feasibility of the new architecture and optimize the signal transmission process, several core and spectrum allocation algorithms were proposed following with the derivation of multiple new IC-XT equations for different types of MCFs. To perform a more practical investigation, various data center topologies, modulation formats, types of MCF and types of request were considered in the simulation process. According to the results, by reducing the IC-XT in MCF, through the adoption of either the bi-direction model or the trench-assisted technique, the transmission reach of the MCF links in DCNs could be considerably extended. In terms of the multiplexing technique, the solution using MCF-based SDM slightly outperformed the wavelength division multiplexing (WDM) using standard single-mode fiber (SMF). Moreover, the scheme that combines SDM and WDM within the MCF enabled the DCN to achieve up to 300-fold greater network capacity and up to 80-fold higher link spatial efficiency compared to a WDM solution using SMF. Furthermore, the studies carried out in this chapter on network capacity with respect to the link distance could be a useful reference for MCF selection when designing, evaluating, and deploying MCFs for different DCN applications.

The studies on SDM were trying to improve the data center performance at link level while not changing the overall DCN architecture. In contrast, Chapter 5 and Chapter 6 concentrated on the study of completely changing the traditional data center (TDC) architecture to DDC architecture, with the aim of increasing the resource (i.e. CPU, memory and storage) utilization. A disaggregated recursive data center in a box (dReDBox) architecture was developed in Chapter 5 with three proposed rack structures and four DDC resource allocation algorithms. By comparing different structure-algorithm combinations, the one provided the best performance in terms of blocking probability and resource utilization was adopted for further comparison between the DDC and TDC. Moreover, a cost model for network gain was created to serve the comparison. Compared to the TDC architecture, the DDC architecture could considerably save the CPU and memory resources at the expense of network latency and extra bandwidth. However, the cost analysis considering practical device costs indicated that for most of the investigated situations, the DDC architecture has much higher cost-efficiency.

In order to reduce the network latency and bandwidth requirement while

further increasing the flexibility and scalability of DDC, two parallel topologies, the Box-modular and Brick-modular, were proposed in Chapter 6. Through the comparison between the parallel topologies and the original dReDBox topology, it was found that both the two topologies could reduce the cost by up to 40% and increase the power-efficiency by up to 68%. Especially, the Box-modular topology could reduce the latency by 34% against the dReDBox topology. Apart from it, a MCF-SMF hybrid DDC architecture built based on the Box-modular topology showed an up to 68%-81% higher power consumption/space-efficiency compared to the one only with SMFs.

Summarizing, from the research works that have been presented in this thesis two overarching conclusions can be drawn. First of all, the introduction of MCF-based SDM technique at the link level can significantly improve the DCN performance, including cabling complexity, front panel density, cost/power-efficiency and especially the network capacity as well as the link spatial efficiency. Moreover, a good approach to reduce the IC-XT in MCF, which is sensitive to many transmission parameters, can further enhance the performance of SDM-based DCN. Secondly, the DDC architecture enables the DCN to have better flexibility, deliver the maximal IT resource utilization and in turn, achieving higher cost-efficiency against the TDC architecture. By applying suitable topologies or introducing the SDM technique, the network resource utilization, round-trip latency and power consumption for DCN can be reduced, while the scalability can be upgraded.

7.2 Future Work

The research that has been presented in this thesis has given rise to a variety of interesting questions and directions, which have been briefly described in the following subsections.

7.2.1 Accurate IC-XT modelling

As described earlier in Chapter 3, IC-XT behavior in MCF can be internally influenced by core pitch, propagation constant, bending radius, and externally affected by signalling source, modulation format, baud rate, temperature, PRBS pattern, operating wavelength, and the number of excited cores. Therefore, to perform a good SDM-based network simulation considering more accurate or practical IC-XT in the resource allocation process, the mathematical model adopted should consider the above parameters as more as possible. However, the existing equations to date have only included a small part of the parameters [144, 173, 175, 189], necessitating the derivation of a more

comprehensive model based on the IC-XT characterization presented in Chapter 3. The new model can also significantly benefit the application of MCFs in data centers.

7.2.2 IC-XT classification using neural network

It has been proven in Section 3.5, the IC-XT in MCF cannot be predicted based on the observed data. However, the proposed model on IC-XT step distribution gives a chance for IC-XT classification by using neural network. That is to say, based on the received IC-XT, the source signals can be classified, which may contribute to the network security. For example, if someone launches a CW light into a core to attack the desired PAM-4 signals in another core, the CW source can be easily recognised according to the observed IC-XT and then the next thing is to filter it.

7.2.3 ML-based ultra-low IC-XT MCF design

According to the work in Chapter 4, the IC-XT reduction in MCF can extend considerably the transmission reach and in turn, enlarge the scale of the DCN. Moreover, by arranging more cores inside the MCF, the network capacity can be significantly increased. To achieve these goals, numerous of MCFs have been designed and fabricated, such as the TA-MCF [120], hole-assisted MCF [225] and rod-assisted MCF [226]. Generally, the MCF is designed based on comparison between several parameters combinations. However, even though wide exploration on the core structure and core distribution in MCF has already been made, there is no reliable method to search the minimal crosstalk, maximal core density and optimal core distributions for different core numbers. As in the fiber design process a large mount of parameters should be considered, e.g. core radius, refractive indexes of core, trench and cladding, widths of trench and cladding, machine learning method can be a good choice. For example, the particle swarm optimization algorithm may be suitable for this.

7.2.4 Demonstration of MCF switching and networking

Simulation results in Chapter 4 and Chapter 6 have shown that MCF solutions can benefit existing DCNs and disaggregated data center networks, respectively. However, none of them have been experimentally validated. Therefore, the most direct step on this work in the future is to demonstrate part of the work with real MCFs and devices. The devices, such as dBricks and MCF switches, or setup that have been manufactured or built for the dReDBox program may contribute to the demonstration, nevertheless, it a long way to go.

Appendix A

Pseudo Code for Algorithm in Chapter 4

Symbol	Description
LN	Number of links for end-to-end path
S_i	Source or ingress node of i-th link
T_i	Destination or egress node of i-th link
BW	Required bandwidth for the connection request
D	Spectrum division
$D1$	Division 1 of available spectrum resources (first 50 slots; 1–50)
$D2$	Division 2 of available spectrum resources (last 50 slots; 51–100)
CPM	Core priority map (specifying usage sequence)
CPM_1	Core priority map for direction 1
CPM_2	Core priority map for direction 2
SA	Available spectrum slots
$IC\text{-}XT$	Total crosstalk for the current request (initially 0)
$IC\text{-}XT_i$	Crosstalk value for the current request on the i-th link (IC-XT$_i$ = -1 indicates spectrum slot unavailability)
CS	Status of division switch (different between soft and hard); 1 means true and 0 means false
t	Indicator of the division being checked (0 is initialized and 1 indicates division switched)
W	Total number of cores
V	Index of the highest used core without adjacent cores carrying signals in the same direction

Algorithm 1 Procedure of resource allocation in MCF-based DCN with A2T3

1: **procedure** Routing result ($LN, S_i, T_i, i = 1, ..., LN$), Core priority map ($CPM_1$ & CPM_2)
2: **for each** i = 1 to LN **do**
3: $IC\text{-}XT_i \leftarrow 0, t \leftarrow 0$ ▷ initialization of parameter
4: **if** S_i less than T_i in the node index value **then**

5: $D \leftarrow D1, CPM \leftarrow CPM_1$
6: **else**
7: $D \leftarrow D2, CPM \leftarrow CPM_2$
8: **end if**
9: **for each** k = 1 to W **do**
10: **if** $k == V + 1$ && $D == D1$ **then**
11: $D \leftarrow D2$
12: **end if**
13: **if** $k == V + 1$ && $D == D2$ **then**
14: $D \leftarrow D1$
15: **end if** ▷ define the division for allocation
16: Check the whole D in k-th core to find SA
17: **if** $SA == BW$ **then**
18: Calculate $IC\text{-}XT_i$
19: **break**
20: **else**
21: $IC\text{-}XT_i \leftarrow -1$
22: **end if**
23: **end for** ▷ get the crosstalk for the current request
24: **if** $IC\text{-}XT_i == -1$ **then** ▷ spectrum resources are not available on current D
25: **if** $t == 0$ && $CS == 1$ **then**
26: $t \leftarrow 1$
27: **if** $D == D1$ **then**
28: $D \leftarrow D2$
29: **else**
30: $D \leftarrow D1$ ▷ division swapping
31: **end if**
32: Turn to line 9 ▷ search spectrum slots on the new D
33: **end if**
34: Block the request ▷ blocked due to resource unavailability
35: **break**
36: **else if** $IC\text{-}XT_i \neq -1$ && $IC\text{-}XT_i \geq$ threshold **then** ▷ Check the IC-XT for current link
37: Block the request ▷ blocked due to high IC-XT
38: **break**
39: **else** ▷ $IC\text{-}XT_i$ is lower than threshold
40: $IC\text{-}XT = IC\text{-}XT + IC\text{-}XT_1$ ▷ Calculate the total IC-XT of link 1 to link i
41: **if** $IC\text{-}XT_i \geq$ threshold **then**
42: Block the request ▷ blocked due to high IC-XT
43: **break**

```
44:             else                                   ▷ IC-XT is lower than threshold
45:                 if i == LN then    ▷ the IC-XT for the whole path (LN links) has been checked
46:                     Connection established
47:                 end if
48:             end if
49:         end if
50:     end for
51: end procedure
```

Appendix B

Pseudo Code for Algorithms in Chapter 5

B.1 First-fit resource allocation algorithm

Algorithm 2 First-fit resource allocation algorithm

Require: G: Data centre graphs, C: Data centre config, R: Request

1: **procedure** First-fit(G, C, R)
2: $available.CPU \leftarrow$ find slots $\geqslant 1$ in $C.CPULocations$
3: $available.Mem \leftarrow$ find slots $\geqslant 1$ in $C.MemLocations$
4: $available.Sto \leftarrow$ find slots $\geqslant 1$ in $C.StoLocations$
5: **if** $available.CPU < R.CPU$ **then** ▷ Inadequate CPU resources
6: $ITallocation \leftarrow Failure$
7: $IT\,failureCause \leftarrow CPU$
8: **else if** $available.Mem < R.Mem$ **then** ▷ Inadequate memory resources
9: $ITallocation \leftarrow Failure$
10: $IT\,failureCause \leftarrow Mem$
11: **else if** $available.Sto < R.Sto$ **then** ▷ Inadequate storage resources
12: $ITallocation \leftarrow Failure$
13: $IT\,failureCause \leftarrow Sto$
14: **else**
15: $loopIncrement \leftarrow C.nSlots \times C.nBlades$ ▷ Size of a rack
16: $slot \leftarrow startSlot$
17: **while** $slot \leqslant C.totalSlots$ **do** ▷ To try multiple combinations
18: $found.CPU \leftarrow 0$ ▷ Initialise resource counter variables
19: $found.Mem \leftarrow 0$
20: $found.Sto \leftarrow 0$
21: $index.CPU \leftarrow 1$ ▷ Initialise resource index variables
22: $index.Mem \leftarrow 1$

```
23:     index.Sto ← 1
24:     ITres ← [ ]                                  ▷ Initialise IT resource 'tracker'
25:     for each slot s in C.ITres do
26:         switch (C.ITresTypes(s)) do              ▷ Different resource types
27:             case CPU                             ▷ Find CPU slots
28:                 if found.CPU < R.CPU then
29:                     units ← C.ITres(s)
30:                     found.CPU ← found.CPU + units
31:                     ITres[1, index.CPU] ← {s, units}
32:                     index.CPU ← index.CPU + 1
33:                 end if
34:             case Mem                             ▷ Find memory slots
35:                 if found.Mem < R.Mem then
36:                     units ← C.ITres(s)
37:                     found.Mem ← found.Mem + units
38:                     ITres[2, index.Mem] ← {s, units}
39:                     index.Mem ← index.Mem + 1
40:                 end if
41:             case Sto                             ▷ Find storage slots
42:                 if found.Sto < R.Sto then
43:                     units ← C.ITres(s)
44:                     found.Sto ← found.Sto + units
45:                     ITres[3, index.Sto] ← {s, units}
46:                     index.Sto ← index.Sto + 1
47:                 end if
48:         end switch
49:     end for
50:     if R.CPU and R.Mem and R.Sto found then   ▷ All resources found
51:         ITallocation ← Success
52:         IT failureCause ← None
53:         break
54:     else
55:         ITallocation ← Failure
56:         Evaluate failure cause
57:         Update IT failureCause
58:     end if
59:     if ITallocation = Success then
60:         [NETallocation, NETres] ← NETALLOCATION(G, C, R, ITres)
```

```
61:                 if NETallocation = Success then
62:                     break
63:                 else
64:                     Remove failure nodes in C.ITres     ▷ Remove failure nodes
65:                 end if
66:             else
67:                 break          ▷ Failed due to unavailability of IT resources
68:             end if
69:             slot ← slot + loopIncrement                  ▷ Jump to next rack
70:         end while
71:     end if
72:     if ITallocation = Success and NETallocation = Success then
73:         Update C.ITres                    ▷ Remove allocated IT resources
74:         Update G.bMap                  ▷ Remove allocated network resources
75:     end if
        return ITallocation, ITres, NETallocation, NETres
76: end procedure
```

B.2 Best-fit resource allocation algorithm

Algorithm 3 Best-fit resource allocation algorithm

Require: G: Data centre graphs, C: Data centre config, R: Request

```
1:  procedure Best-fit(G, C, R)
2:      available.CPU ← find slots ⩾ 1 in C.CPULocations
3:      available.Mem ← find slots ⩾ 1 in C.MemLocations
4:      available.Sto ← find slots ⩾ 1 in C.StoLocations
5:      if available.CPU < R.CPU then              ▷ Inadequate CPU resources
6:          ITallocation ← Failure
7:          IT failureCause ← CPU
8:      else if available.Mem < R.Mem then         ▷ Inadequate memory resources
9:          ITallocation ← Failure
10:         IT failureCause ← Mem
11:     else if available.Sto < R.Sto then         ▷ Inadequate storage resources
12:         ITallocation ← Failure
13:         IT failureCause ← Sto
14:     else
15:         loopIncrement ← C.nSlots × C.nBlades                ▷ Size of a rack
16:         slotLimit ← max(C.CPULocations, C.MemLocations, C.StoLocations)
17:         slot ← 1
```

```
18:  while slot < slotLimit do                          ▷ To try multiple combinations
19:      found.CPU ← 0                                  ▷ Initialise resource counter variables
20:      found.Mem ← 0
21:      found.Sto ← 0
22:      index.CPU ← 1                                  ▷ Initialise resource index variables
23:      index.Mem ← 1
24:      index.Sto ← 1
25:      CPUStart ← 1                                   ▷ Initialise start slot numbers
26:      memStart ← 1
27:      stoStart ← 1
28:      ITres ← [ ]                                    ▷ Initialise IT resource 'tracker'
29:      for each slot s from CPUStart in C.CPUResources do
30:          if found.CPU < R.CPU then                  ▷ Find CPU slots
31:              units ← C.CPUResources(s)
32:              found.CPU ← found.CPU + units
33:              ITres[1, index.CPU] ← {s, units}
34:              index.CPU ← index.CPU + 1
35:          else
36:              break
37:          end if
38:      end for
39:      for each slot s from memStart in C.MemResources do
40:          if found.Mem < R.Mem then                  ▷ Find memory slots
41:              units ← C.MemResources(s)
42:              found.Mem ← found.Mem + units
43:              ITres[2, index.Mem] ← {s, units}
44:              index.Mem ← index.Mem + 1
45:          else
46:              break
47:          end if
48:      end for
49:      for each slot s from stoStart in C.StoResources do
50:          if found.Sto < R.Sto then                  ▷ Find storage slots
51:              units ← C.StoResources(s)
52:              found.Sto ← found.Sto + units
53:              ITres[3, index.Sto] ← {s, units}
54:              index.Sto ← index.Sto + 1
55:          end if
```

```
56:             end for
57:             if R.CPU and R.Mem and R.Sto found then   ▷ All resources found
58:                 ITallocation ← Success
59:                 IT failureCause ← None
60:                 break
61:             else
62:                 ITallocation ← Failure
63:                 Evaluate failure cause
64:                 Update IT failureCause and IT checkBreak
65:             end if
66:             if ITallocation = Success then
67:                 [NETallocation, NETres] ← NETALLOCATION(G, C, R, ITres)
68:                 if NETallocation = Success then
69:                     break
70:                 else
71:                     Remove failure nodes in C.CPUResources
72:                     Remove failure nodes in C.MemResources
73:                     Remove failure nodes in C.StoResources
74:                 end if
75:             else
76:                 if IT checkBreak = 1 then
77:                     break          ▷ Failed due to unavailability of IT resources
78:                 end if
79:             end if
80:             CPUStart ← CPUStart + loopIncrement          ▷ Increment all slot
    indices
81:             memStart ← memStart + loopIncrement
82:             stoStart ← stoStart + loopIncrement
83:             slot ← slot + loopIncrement                  ▷ Jump to next rack
84:         end while
85:     end if
86:     if ITallocation = Success and NETallocation = Success then
87:         Update C.ITres                          ▷ Remove allocated IT resources
88:         Update G.bMap                  ▷ Remove allocated network resources
89:     end if
        return ITallocation, ITres, NETallocation, NETres
90: end procedure
```

B.3 NULB resource allocation algorithm

Algorithm 4 NULB resource allocation algorithm

Require: G: Data center graphs, C: Data center config, R: Request

1: **procedure** NULB(G, C, R)
2: $available.CPU \leftarrow$ find slots $\geqslant 1$ in $C.CPULocations$
3: $available.Mem \leftarrow$ find slots $\geqslant 1$ in $C.MemLocations$
4: $available.Sto \leftarrow$ find slots $\geqslant 1$ in $C.StoLocations$
5: $cr.CPU \leftarrow \frac{R.CPU}{available.CPU}$ ▷ CPU contention ratio
6: $cr.Mem \leftarrow \frac{R.CPU}{available.Mem}$ ▷ Memory contention ratio
7: $cr.Sto \leftarrow \frac{R.CPU}{available.Sto}$ ▷ Storage contention ratio
8: $cr.Max \leftarrow max(cr.CPU, cr.Mem, cr.Sto)$ ▷ Find maximum contention ratio
9: $loopIncrement \leftarrow C.nSlots \times C.nBlades$ ▷ Size of a rack
10: **for each** c in cr **do** ▷ To try multiple contention ratios
11: $s \leftarrow cr.Max$
12: $cr.Max \leftarrow$ Update to new maximum contention ratio
13: **switch** s **do**
14: **case** CPU
15: **if** $available.CPU < R.CPU$ **then** ▷ Inadequate CPU resources
16: $ITallocation \leftarrow Failure$
17: $IT failureCause \leftarrow CPU$
18: **else**
19: **while** $slot \leqslant C.CPUResources$ **do**
20: $[ITallocation, ITres] \leftarrow$ BFS($G, slot, R$)
21: **if** $ITallocation = Success$ **then**
22: $[NETallocation, NETres] \leftarrow$ NETALLOCATION($G, C, R, ITres$)
23: **if** $NETallocation = Success$ **then**
24: **break**
25: **else**
26: Remove failure nodes in $G.ITres$
27: **end if**
28: **else**
29: Evaluate failure cause
30: Update $IT failureCause$
31: **break**
32: **end if**
33: $slot \leftarrow slot + loopIncrement$
34: **end while**

```
35:                 end if
36:                 if ITallocation = Success and NETallocation = Success then
37:                     break
38:                 end if
39:             case Mem
40:                 if available.Mem < R.Mem then          ▷ Inadequate memory resources
41:                     ITallocation ← Failure
42:                     IT failureCause ← Mem
43:                 else
44:                     while slot ⩽ C.MemResources do
45:                         [ITallocation, ITres] ← BFS(G, slot, R)
46:                         if ITallocation = Success then
47:                             [NETallocation, NETres] ← NETALLOCATION(G, C, R, ITres)
48:                             if NETallocation = Success then
49:                                 break
50:                             else
51:                                 Remove failure nodes in G.ITres
52:                             end if
53:                         else
54:                             Evaluate failure cause
55:                             Update IT failureCause
56:                             break
57:                         end if
58:                         slot ← slot + loopIncrement
59:                     end while
60:                 end if
61:                 if ITallocation = Success and NETallocation = Success then
62:                     break
63:                 end if
64:             case Sto
65:                 if available.Sto < R.Sto then     ▷ Inadequate storage resources
66:                     ITallocation ← Failure
67:                     IT failureCause ← Sto
68:                 else
69:                     while slot ⩽ C.StoResources do
70:                         [ITallocation, ITres] ← BFS(G, slot, R)
```

```
                    if ITallocation = Success then
                        [NETallocation, NETres] ← NETALLOCATION(G, C, R, ITres)
                        if NETallocation = Success then
                            break
                        else
                            Remove failure nodes in G.ITres
                        end if
                    else
                        Evaluate failure cause
                        Update IT failureCause
                        break
                    end if
                    slot ← slot + loopIncrement
                end while
            end if
            if ITallocation = Success and NETallocation = Success then
                break
            end if
        if ITallocation = Failure then                ▷ Inadequate IT resources
            break
        end if
    end switch
end for
if ITallocation = Success and NETallocation = Success then
    Update C.ITres                                    ▷ Remove allocated IT resources
    Update G.bMap                                     ▷ Remove allocated network resources
end if
 return ITallocation, ITres, NETallocation, NETres
end procedure
```

B.4 Modified BFS algorithm

Algorithm 5 Modified breadth-first search algorithm

Require: G: Data center graphs, s: Start vertex, R: Request

```
procedure modifiedBFS(G, C, R)
    bMap ← G.bMap                                     ▷ Copy original bandwidth map
    dMap ← G.dMap                                     ▷ Copy original distance map
    Remove all edges in bMap and dMap where e_b < min(R.b_cm, R.b_ms)
```

```
5:    for each vertex v in G do                          ▷ Initialise each vertex
6:        v.distance ← ∞
7:        v.parent ← null
8:    end for
9:    Q ← [ ]                                            ▷ Create an empty queue
10:   s.distance ← 0
11:   Q.enqueue(s)                                       ▷ Enqueue start vertex
12:   ITres ← Find IT resources on start vertex s
13:   while Q ≠ [ ] do                                   ▷ Run until queue is empty
14:       current ← Q.dequeue()
15:       neighbours ← All vertices adjacent to current  ▷ High bandwidth links
      have a higher priority (Network aware)
16:       neighbours' ← Sort neighbours in descending order
17:       for each vertex v in neighbours' do
18:           if v.distance = ∞ then
19:               v.distance ← current.distance + 1
20:               v.parent ← current
21:               Q.enqueue(v)                           ▷ Enqueue v^th neighbour
22:               ITres ← Find IT resources on vertex v
23:               if R.CPU and R.Mem and R.Sto found then ▷ Resources found
24:                   breakWhile ← True
25:                   break
26:               end if
27:           end if
28:       end for
29:       if breakWhile = True then
30:           ITallocation ← Success
31:           IT failureCause ← None
32:           break
33:       else
34:           ITallocation ← Failure
35:           IT failureCause ← Evaluate failure cause
36:           break
37:       end if
38:   end while
       return ITres, ITallocation
39: end procedure
```

Acronyms

ASE	Amplified spontaneous emission
AWG	Arrayed waveguide grating
AWG	Arbitrary waveform generator
BER	Bit error rate
BP	Blocking probability
BPSK	Binary phase shift keying
BV-WSS	Bandwidth-variable wavelength selective switch
CDF	Cumulative distribution function
CPU	Central processing unit
CWDM	Coarse wavelength division multiplexing
dBESM	Disaggregated box electronic switch module
dBOSM	Disaggregated box optical switch module
DCN	Data center network
DDC	Disaggregated data center
dDOSM	Disaggregated data center optical switch module
DP-16QAM	Dual-polarization 16-quadrature amplitude modulation
dPOSM	Disaggregated plane optical switch module
dReDBox	Disaggregated recursive data center in a box
dROSM	Disaggregated rack optical switch module

DSP	Digital signal processing
DWDM	Dense wavelength division multiplexing
ECL	External cavity laser
EDFA	Erbium-doped fiber amplifier
HMC	Hybrid memory cube
IC-XT	Inter-core crosstalk
IP	Internet protocol
IT	Information technology
KPI	Key performance indicator
LC	Lucent
MBO	Mid board optic
MCF	Multi-core fiber
MEMS	Micro-electromechanical system
MTP/MPO	Multi-fiber termination push-on/multi-fiber push-on
MZM	Mach-Zehnder modulator
NIC	Network interface card
OCS	Optical circuit switch
OOK	On-off keying
OSNR	Optical signal-to-noise ratio
PAM	Pulse amplitude modulation
PDF	Probability density function
PMP	Phase-matching point
PODs	Points of delivery
PPG	Pulse pattern generator

PRBS	Pseudo random binary sequence
PSD	Power spectral density
QAM	Quadrature amplitude modulation
QPSK	Quadrature phase shift keying
RSA	Routing and spectrum allocation
SDM	Space division multiplexing
SI-MCF	Step-index multi-core fiber
SMF	Single-mode fiber
SNR	Signal to noise ratio
TA-MCF	Trench assisted multi-core fiber
TCC	Temperature-controlled chamber
TDC	Traditional data center
ToB	Top of blade
ToR	Top of rack
VCSEL	Vertical cavity surface emitting laser
VM	Virtual machine
VOA	Variable optical attenuator
WDM	Wavelength division multiplexing
WSS	Wavelength selective switch

Bibliography

[1] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: a review," *Optica*, vol. 5, no. 11, pp. 1354–1370, Nov 2018.

[2] Cisco, "The zettabyte erab: Trends and analysis," *White Paper*, July 2016.

[3] Cisco, "Cisco global cloud index: Forecast and methodology 2016-2021," *White Paper*, Nov 2018.

[4] Cisco, "Cisco visual networking index: Forecast and trends, 2017–2022," *White Paper*, Feb 2019.

[5] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, "Energy efficiency in the future internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures," *IEEE Communications Surveys Tutorials*, vol. 13, no. 2, pp. 223–244, Second 2011.

[6] Y. Zhang, P. Chowdhury, M. Tornatore, and B. Mukherjee, "Energy efficiency in telecom optical networks," *IEEE Communications Surveys Tutorials*, vol. 12, no. 4, pp. 441–458, Fourth 2010.

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

[8] Greenpeace International, "Make it green: Cloud computing and its contribution to climate change," April 2010, https://storage.googleapis.com/planet4-international-stateless/2010/03/f2954209-make-it-green-cloud-computing.pdf.

[9] A. S. G. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.

[10] S. Han, N. Egi, A. Panda, S. Ratnasamy, G. Shi, and S. Shenker, "Network support for resource disaggregation in next-generation datacenters," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. New York, NY, USA: ACM, 2013, pp. 10:1–10:7.

[11] Q. Zhang, J. L. Hellerstein, and R. Boutaba, "Characterizing task usage shapes in google's compute clusters," in *In Large Scale Distributed Systems and Middleware Workshop*. LADIS, 2011.

[12] L. A. Barroso and U. Holzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, Dec 2007.

[13] NGMN Alliance, "NGMN 5G White Paper," Feb 2015, https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf.

[14] S. Aleksic, "The future of optical interconnects for data centers: A review of technology trends," in *2017 14th International Conference on Telecommunications (ConTEL)*, June 2017, pp. 41–46.

[15] X. Zhou, R. Urata, and H. Liu, "Beyond 1tb/s datacenter interconnect technology: Challenges and solutions," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2019, pp. 1–3.

[16] B. Abali, R. J. Eickemeyer, H. Franke, C. Li, and M. Taubenblatt, "Disaggregated and optically interconnected memory: when will it be cost effective?" *CoRR*, vol. abs/1503.01416, 2015, http://arxiv.org/abs/1503.01416.

[17] P. X. Gao, A. Narayan, S. Karandikar, J. Carreira, S. Han, R. Agarwal, S. Ratnasamy, and S. Shenker, "Network requirements for resource disaggregation," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'16. Berkeley, CA, USA: USENIX Association, 2016, pp. 249–264.

[18] Y. Cheng, R. Lin, M. D. Andrade, J. L. Wosinska, and J. Chen, "Disaggregated data centers: Challenges and tradeoffs," *IEEE Communications Magazine*, 2019.

[19] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[20] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, Jan 1949.

[21] M. Secondini and E. Forestieri, "Scope and limitations of the nonlinear shannon limit," *Journal of Lightwave Technology*, vol. 35, no. 4, pp. 893–902, Feb 2017.

[22] R. Dar, P. J. Winzer, A. R. Chraplyvy, S. Zsigmond, K. . Huang, H. Fevrier, and S. Grubb, "Cost-optimized submarine cables using massive spatial parallelism," *Journal of Lightwave Technology*, vol. 36, no. 18, pp. 3855–3865, Sep. 2018.

[23] R. Freund, L. Molle, F. Raub, C. Caspar, M. Karkri, and C. Weber, "Triple-(s/c/l)-band wdm transmission using erbium-doped fibre amplifiers," in *2005 31st European Conference on Optical Communication, ECOC 2005*, vol. 1, Sep. 2005, pp. 69–70 vol.1.

[24] F. Boubal, E. Brandon, L. Buet, S. Chernikov, V. Havard, C. Heerdt, A. Hugbart, W. Idler, L. Labrunie, P. Le Roux, S. A. E. Lewis, A. Pham, L. Piriou, R. Uhel, and J. . Blondel, "4.16 tbit/s (104 /spl times/ 40 gbit/s) unrepeatered transmission over 135 km in s+c+l bands with 104 nm total bandwidth," in *Proceedings 27th European Conference on Optical Communication (Cat. No.01TH8551)*, vol. 1, Sep. 2001, pp. 58–59 vol.1.

[25] M. Matsuura and N. Kishi, "Broadband wavelength conversion with s/c/l-band flexible operation using cross-gain-modulation in a single quantum dot soa," in *2011 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference*, March 2011, pp. 1–3.

[26] S. Pitris, M. Moralis-Pegios, T. Alexoudi, Y. Ban, P. De Heyn, J. Van Campenhout, and N. Pleros, "A 4×40 Gb/s O-Band WDM Silicon Photonic Transmitter Based on Micro-Ring Modulators," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2019, pp. 1–3.

[27] B. Schrenk, M. Hentschel, and H. Hübel, "O-Band Differential Phase-Shift Quantum Key Distribution in 52-Channel C/L-Band Loaded Passive Optical Network," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2019, pp. 1–3.

[28] Tianjian Zuo, A. Tatarczak, M. Iglesias Olmedo, J. Estaran, J. B. Jensen, Q. Zhong, X. Xu, and I. T. Monroy, "O-band 400 gbit/s client side optical transmission link," in *OFC 2014*, March 2014, pp. 1–3.

[29] R. Kubo, M. Matsunaga, T. Shobudani, T. Fujimoto, H. Tsuda, M. Sudo, T. Hajikano, Y. Tomomatsu, and K. Yoshizawa, "Demonstration of 10-gbit/s transmission over g.652 fiber for t-band optical access systems using quantum-dot semiconductor devices," *IEICE Electronics Express*, vol. 15, no. 18, pp. 20 180 810–20 180 810, 2018.

[30] N. A. Idris, N. Yamamoto, K. Akahane, K. Yoshizawa, Y. Tomomatsu, M. Sudo, T. Hajikano, R. Kubo, T. Uesugi, T. Kikuchi, and H. Tsuda, "A WDM/TDM Access Network Based on Broad T-Band Wavelength Resource Using Quantum Dot Semiconductor Devices," *IEEE Photonics Journal*, vol. 8, no. 1, pp. 1–10, Feb 2016.

[31] N. A. Idris, K. Yoshizawa, Y. Tomomatsu, M. Sudo, T. Hajikano, R. Kubo, G. Zervas, and H. Tsuda, "Full-mesh T- and O-band wavelength router based on arrayed waveguide gratings," *Opt. Express*, vol. 24, no. 1, pp. 672–686, Jan 2016.

[32] N. K. Fontaine, R. Ryf, Haoshuo Chen, A. Velazquez Benitez, J. E. Antonio Lopez, R. Amezcua Correa, Binbin Guan, B. Ercan, R. P. Scott, S. J. Ben Yoo, L. Grüner-Nielsen, Yi Sun, and R. J. Lingle, "30×30 mimo transmission over 15 spatial modes," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2015, pp. 1–3.

[33] T. Mori, T. Sakamoto, M. Wada, T. Yamamoto, and F. Yamamoto, "Six-lp-mode transmission fiber with dmd of less than 70 ps/km over c+l band," in *OFC 2014*, March 2014, pp. 1–3.

[34] R. Ryf, S. Randel, A. H. Gnauck, C. Bolle, A. Sierra, S. Mumtaz, M. Esmaeelpour, E. C. Burrows, R. Essiambre, P. J. Winzer, D. W. Peckham, A. H. McCurdy, and R. Lingle, "Mode-division multiplexing over 96 km of few-mode fiber using coherent 6×6 mimo processing," *Journal of Lightwave Technology*, vol. 30, no. 4, pp. 521–531, Feb 2012.

[35] K. Mukasa, K. Imamura, and R. Sugizaki, "Multi-core few-mode optical fibers with large aeff," in *European Conference and Exhibition on Optical Communication*. Optical Society of America, 2012, p. P1.08.

[36] Y. Sasaki, Y. Amma, K. Takenaga, S. Matsuo, K. Saitoh, and M. Koshiba, "Trench-assisted low-crosstalk few-mode multicore fiber," in *39th European Conference and Exhibition on Optical Communication (ECOC 2013)*, Sep. 2013, pp. 1–3.

[37] K. Igarashi, T. Tsuritani, I. Morita, and M. Suzuki, "114 space-division-multiplexed wdm transmission using 6-mode 19-core fibers," in *2015 IEEE Photonics Conference (IPC)*, Oct 2015, pp. 643–644.

[38] X. Chen, J. Cho, A. Adamiecki, and P. Winzer, "16384-qam transmission at 10 gbd over 25-km ssmf using polarization-multiplexed probabilistic constellation shaping," in *2019 European Conference on Optical Communication (ECOC)*, 2019.

[39] H. Yuan, A. Saljoghei, T. Hayashi, T. Nakanishi, E. Sillekens, L. Galdino, P. Bayvel, Z. Liu, and G. Zervas, "Experimental investigation of static and dynamic crosstalk in trench-assisted multi-core fiber," in *Optical Fiber Communication Conference (OFC) 2019*. Optical Society of America, 2019, p. W4C.2.

[40] H. Yuan, A. Saljoghei, T. Hayashi, T. Nakanishi, E. Sillekens, L. G. amd P. Bayvel, Z. Liu, and G. Zervas, "Experimental analysis on variations and accuracy of crosstalk in trench-assisted multi-core fibers (invited)," *Journal of Lightwave Technology*, 2019, under review.

[41] A. Ottino, H. Yuan, Y. Xu, E. Sillekens, and G. Zervas, "Using pesudo ransom walk to model step distribution of mcf crosstalk on cw/ase/ppk/pam4/m-qam signals," in *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, 2020, under review.

[42] Y. Liu, H. Yuan, A. Peters, and G. Zervas, "Comparison of sdm and wdm on direct and indirect optical data center networks," in *ECOC 2016; 42nd European Conference on Optical Communication*, Sept 2016, pp. 1–3.

[43] H. Yuan, M. Furdek, A. Muhammad, A. Saljoghei, L. Wosinska, and G. Zervas, "Space-division multiplexing in data center networks: on multi-core fiber solutions and crosstalk-suppressed resource allocation," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 4, pp. 272–288, April 2018.

[44] H. Yuan, A. Saljoghei, A. Peters, and G. Zervas, "Comparison of sdm-wdm based data center networks with equal/unequal core pitch multi-core fibers," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[45] G. Zervas, F. Jiang, Q. Chen, V. Mishra, H. Yuan, K. Katrinis, D. Syrivelis, A. Reale, D. Pnevmatikatos, M. Enrico, and N. Parsons, "Disaggregated

compute, memory and network systems: A new era for optical data centre architectures," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–3.

[46] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, and V. Mishra, "Optically disaggregated data centers with minimal remote memory latency: Technologies, architectures, and resource allocation [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A270–A285, Feb 2018.

[47] H. Yuan, A. Saljoghei, A.Peters, and G. Zervas, "Disaggregated optical data center in a box network using parallel ocs topologies," in *Optical Fiber Communication Conference*. Optical Society of America, 2018, p. W1C.2.

[48] A. Saljoghei, H. Yuan, V. Mishra, M. Enrico, N. Parsons, C. Kochis, P. De Dobbelaere, D. Theodoropoulos, D. Pnevmatikatos, D. Syrivelis, A. Reale, T. Hayashi, T. Nakanishi, and G. Zervas, "Mcf-smf hybrid low-latency circuit-switched optical network for disaggregated data centers," *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4017–4029, Aug 2019.

[49] R. Birke, L. Y. Chen, and E. Smirni, "Data centers in the cloud: A large scale performance study," in *2012 IEEE Fifth International Conference on Cloud Computing*, June 2012, pp. 336–343.

[50] A. Weissberger, "2013 IDC directions part ii- new data center dynamics and requirements," Mar 2013, http://viodi.com/2013/03/17/2013-idc-directions-part-ii-new-data-center-dynamics-and-requirements/.

[51] Tencent and Intel, "Tencent explores datacenter resource-pooling using Intel® rack scale architecture (Intel® RSA)," https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/rsa-tencent-paper.pdf.

[52] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," in *Proceedings of the Third ACM Symposium on Cloud Computing*, ser. SoCC '12. New York, NY, USA: ACM, 2012, pp. 7:1–7:13.

[53] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, Dec 2007.

[54] P. Teich, "Research note: Intel's disaggregated server rack," Aug 2013, http://www.moorinsightsstrategy.com/wp-content/uploads/2013/08/Intels-Disagggregated-Server-Rack-by-Moor-Insights-Strategy.pdf.

[55] A. D. Papaioannou, R. Nejabati, and D. Simeonidou, "The benefits of a disaggregated data centre: A resource allocation approach," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–7.

[56] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch, "Disaggregated memory for expansion and sharing in blade servers," in *Proceedings of the 36th Annual International Symposium on Computer Architecture*, ser. ISCA '09. New York, NY, USA: ACM, 2009, pp. 267–278.

[57] X. Meng, C. Isci, J. Kephart, L. Zhang, E. Bouillet, and D. Pendarakis, "Efficient resource provisioning in compute clouds via vm multiplexing," in *Proceedings of the 7th International Conference on Autonomic Computing*, ser. ICAC '10. New York, NY, USA: ACM, 2010, pp. 11–20.

[58] H. M. Mohammad Ali, T. E. H. El-Gorashi, A. Q. Lawey, and J. M. H. Elmirghani, "Future energy efficient data centers with disaggregated servers," *Journal of Lightwave Technology*, vol. 35, no. 24, pp. 5361–5380, Dec 2017.

[59] G. M. Saridis, Y. Yan, Y. Shu, S. Yan, M. Arslan, T. Bradley, N. V. Wheeler, N. H. L. Wong, F. Poletti, M. N. Petrovich, D. J. Richardson, S. Poole, G. Zervas, and D. Simeonidou, "Evros: All-optical programmable disaggregated data centre interconnect utilizing hollow-core bandgap fibre," in *2015 European Conference on Optical Communication (ECOC)*, Sep. 2015, pp. 1–3.

[60] Y. Yan, G. M. Saridis, Y. Shu, B. R. Rofoee, S. Yan, M. Arslan, T. Bradley, N. V. Wheeler, N. H. Wong, F. Poletti, M. N. Petrovich, D. J. Richardson, S. Poole, G. Zervas, and D. Simeonidou, "All-optical programmable disaggregated data centre network realized by fpga-based switch and interface card," *Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1925–1932, April 2016.

[61] D. Syrivelis, A. Reale, K. Katrinis, I. Syrigos, M. Bielski, D. Theodoropoulos, D. N. Pnevmatikatos, and G. Zervas, "A software-defined architecture and prototype for disaggregated memory rack scale systems," in *2017 International Conference on Embedded Computer*

Systems: Architectures, Modeling, and Simulation (SAMOS), July 2017, pp. 300–307.

[62] A. Saljoghei, V. Mishra, M. Bielski, I. Syrigos, K. Katrinis, D. Syrivelis, A. Reale, D. N. Pnevmatikatos, D. Theodoropoulos, M. Enrico, N. Parsons, and G. Zervas, "dreddbox: Demonstrating disaggregated memory in an optical data centre," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[63] M. Besta and T. Hoefler, "Slim fly: A cost effective low-diameter network topology," in *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov 2014, pp. 348–359.

[64] S. Tiyyagura, P. Adamidis, R. Rabenseifner, P. Lammers, S. Borowski, F. Lippold, F. Svensson, O. Marxen, S. Haberhauer, A. Seitsonen, J. Furthmüller, K. Benkert, M. Galle, T. Bönisch, U. Küster, and M. Resch, "Teraflops sustained performance with real world applications," *The International Journal of High Performance Computing Applications*, vol. 22, no. 2, pp. 131–148, 2008.

[65] S. Legtchenko, N. Chen, D. Cletheroe, A. Rowstron, H. Williams, and X. Zhao, "Xfabric: A reconfigurable in-rack network for rack-scale computers," in *Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation*, ser. NSDI'16. Berkeley, CA, USA: USENIX Association, 2016, pp. 15–29.

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

[67] W. M. Mellette, A. C. Snoeren, and G. Porter, "P-fattree: A multi-channel datacenter network topology," in *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, ser. HotNets '16. New York, NY, USA: ACM, 2016, pp. 78–84.

[68] 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.

[69] Z. Guo and Y. Yang, "On nonblocking multirate multicast fat-tree data center networks with server redundancy," in *2012 IEEE 26th International Parallel and Distributed Processing Symposium*, May 2012, pp. 1034–1044.

[70] Z. Han and L. Yu, "A survey of the bcube data center network topology," in *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*, May 2018, pp. 229–231.

[71] D. Li, J. Wu, Z. Liu, and F. Zhang, "Dual-centric data center network architectures," in *2015 44th International Conference on Parallel Processing*, Sep. 2015, pp. 679–688.

[72] E. Conrad, S. Misenar, and J. Feldman, "Chapter 5 - domain 4: Communication and network security (designing and protecting network security)," in *CISSP Study Guide (Third Edition)*, E. Conrad, S. Misenar, and J. Feldman, Eds. Boston: Syngress, 2016, pp. 219 – 291.

[73] G. Lee, "Chapter 4 - cloud data center networking topologies," in *Cloud Networking*, G. Lee, Ed. Boston: Morgan Kaufmann, 2014, pp. 65 – 85.

[74] A. Henmi, "Chapter 5 - defining a vpn," in *Firewall Policies and VPN Configurations*, A. Henmi, Ed. Burlington: Syngress, 2006, pp. 211 – 265.

[75] Cisco, "Cisco data center spine-and-leaf architecture: Design overview," *White Paper*, April 2016.

[76] W. Nelson, "Introduction to spine-leaf networking designs," *Lenovo Express*, Nov 2017, https://lenovopress.com/lp0573.pdf.

[77] E. Banks, "Data center network design moves from tree to leaf," Nov 2013, https://searchdatacenter.techtarget.com/feature/Data-center-network-design-moves-from-tree-to-leaf.

[78] M. Alizadeh and T. Edsall, "On the data path performance of leaf-spine datacenter fabrics," in *Proceedings of the 2013 IEEE 21st Annual Symposium on High-Performance Interconnects*, ser. HOTI '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 71–74.

[79] G. Kathareios, C. Minkenberg, B. Prisacari, G. Rodriguez, and T. Hoefler, "Cost-effective diameter-two topologies: analysis and evaluation," in *SC*

'15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Nov 2015, pp. 1–11.

[80] S. R. Ohring, M. Ibel, S. K. Das, and M. J. Kumar, "On generalized fat trees," in *Proceedings of 9th International Parallel Processing Symposium*, April 1995, pp. 37–44.

[81] S. Zafar, A. Bashir, and S. Chaudhry, "On implementation of dctcp on three-tier and fat-tree data center network topologies," *SpringerPlus*, vol. 5, 12 2016.

[82] A. Andreyev, "Introducing data center fabric, the next-generation facebook data center network," Nov 2014.

[83] A. Moorthy, "Connecting the world: A look inside facebook's networking infrastructure," https://www.cs.unc.edu/xcms/wpfiles/50th-symp/Moorthy.pdf.

[84] N. Farrington and A. Andreyev, "Facebook's data center network architecture," in *2013 Optical Interconnects Conference*, May 2013, pp. 49–50.

[85] N. Benzaoui, Y. Pointurier, and S. Bigo, "Latency in a 2d torus burst optical slot switching data center," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–3.

[86] T. Wang, Z. Su, Y. Xia, B. Qin, and M. Hamdi, "Novacube: A low latency torus-based network architecture for data centers," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 2252–2257.

[87] P. Xie, H. Gu, K. Wang, X. Yu, and S. Ma, "Mesh-of-torus: a new topology for server-centric data center networks," *The Journal of Supercomputing*, vol. 75, no. 1, pp. 255–271, Jan 2019.

[88] B. Lebiednik, A. Mangal, and N. Tiwari, "A survey and evaluation of data center network topologies," *CoRR*, vol. abs/1605.01701, 2016.

[89] T. Wang, Z. Su, Y. Xia, and M. Hamdi, "Rethinking the data center networking: Architecture, network protocols, and resource sharing," *IEEE Access*, vol. 2, pp. 1481–1496, 2014.

[90] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: A cost-efficient topology for high-radix networks," *SIGARCH Comput. Archit. News*, vol. 35, no. 2, pp. 126–137, Jun. 2007.

[91] T. Wang, Z. Su, Y. Xia, Y. Liu, J. Muppala, and M. Hamdi, "Sprintnet: A high performance server-centric network architecture for data centers," in *2014 IEEE International Conference on Communications (ICC)*, June 2014, pp. 4005–4010.

[92] S. Chaugule and A. More, "WDM and optical amplifier (Wavelength Division Multiplexing)," in *2010 2nd International Conference on Mechanical and Electronics Engineering*, vol. 2, Aug 2010, pp. V2–232–V2–236.

[93] O. E. DeLange, "Wide-band optical communication systems: Part II—Frequency-division multiplexing," *Proceedings of the IEEE*, vol. 58, no. 10, pp. 1683–1690, Oct 1970.

[94] H. Ishio, J. Minowa, and K. Nosu, "Review and status of wavelength-division-multiplexing technology and its application," *Journal of Lightwave Technology*, vol. 2, no. 4, pp. 448–463, August 1984.

[95] T. Miki and H. Ishio, "Viabilities of the Wavelength-Division-Multiplexing Transmission System Over an Optical Fiber Cable," *IEEE Transactions on Communications*, vol. 26, no. 7, pp. 1082–1087, July 1978.

[96] B. Mukherjee, "WDM optical communication networks: progress and challenges," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, pp. 1810–1824, Oct 2000.

[97] A. Marincic and V. Acimovic-Raspopovic, "Evolution of WDM optical networks," in *5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service. TELSIKS 2001. Proceedings of Papers (Cat. No.01EX517)*, vol. 2, Sep. 2001, pp. 473–480 vol.2.

[98] ITU-T G.694.2, "Spectral grids for WDM applications: CWDM wavelength grid," Dec 2003.

[99] ITU-T G.694.1, "Spectral grids for WDM applications: DWDM frequency grid," Feb 2012.

[100] A. Lord, Y. R. Zhou, R. Jensen, A. Morea, and M. Ruiz, "Evolution from wavelength-switched to flex-grid optical networks," in *Elastic Optical Networks: Architectures, Technologies, and Control*, V. López and

L. Velasco, Eds. Cham: Springer International Publishing, 2016, pp. 7–30.

[101] Y. Zhang, Y. Zhang, S. K. Bose, and G. Shen, "Migration from fixed to flexible grid optical networks with sub-band virtual concatenation," *Journal of Lightwave Technology*, vol. 35, no. 10, pp. 1752–1765, May 2017.

[102] 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.

[103] S. Ferdousi, T. Ahmed, S. Rahman, X. Yu, M. Tornatore, and B. Mukherjee, "Migrating from fixed grid to flexible grid optical networks," in *2018 IEEE Photonics Conference (IPC)*, Sep. 2018, pp. 1–2.

[104] M. Recalcati, F. Musumeci, M. Tornatore, S. Bregni, and A. Pattavina, "Benefits of elastic spectrum allocation in optical networks with dynamic traffic," in *2014 IEEE Latin-America Conference on Communications (LATINCOM)*, Nov 2014, pp. 1–6.

[105] 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.

[106] N. Yamamoto and H. Sotobayashi, "All-band photonic transport system and its device technologies," *Optical Metro Networks and Short-Haul Systems*, vol. 7235, Jan 2009.

[107] N. Yamamoto, Y. Omigawa, K. Akahane, T. Kawanishi, and H. Sotobayashi, "Simultaneous 3×10 Gbps optical data transmission in 1-μm, C-, and L-wavebands over a single holey fiber using an ultra-broadband photonic transport system," *Opt. Express*, vol. 18, no. 5, pp. 4695–4700, Mar 2010.

[108] H. Tsuda, "Large-scale arrayed-waveguide grating based photonic router using T- and O-band," in *2016 IEEE 6th International Conference on Photonics (ICP)*, March 2016, pp. 1–3.

[109] D. Richardson, J. Fini, and L. Nelson, "Space division multiplexing in optical fibres," *Nature Photonics*, vol. 7, pp. 354–362, 05 2013.

[110] P. J. Winzer, "Scaling optical networks through sdm technologies," in *2019 24th OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC)*, July 2019, pp. 1–1.

[111] S. Inao, T. Sato, S. Sentsui, T. Kuroha, and Y. Nishimura, "Multicore optical fiber," in *Optical Fiber Communication*. Optical Society of America, 1979, p. WB1.

[112] S. Berdagué and P. Facq, "Mode division multiplexing in optical fibers," *Appl. Opt.*, vol. 21, no. 11, pp. 1950–1955, Jun 1982.

[113] G. M. Saridis, D. Alexandropoulos, G. Zervas, and D. Simeonidou, "Survey and evaluation of space division multiplexing: From technologies to optical networks," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2136–2156, Fourthquarter 2015.

[114] Sumitomo Electric Lightwave, "Fiber bundles," https://www.sumitomoelectric.com/products/fiber-bundles/.

[115] OFS, "Rollable ribbon fiber optic cables," https://fiber-optic-catalog.ofsoptics.com/viewitems/fiber-optic-cables/ables-rollable-ribbon-fiber-optic-cables-1.

[116] C. Kachris, K. Bergman, and I. Tomkos, *Optical Interconnects for Future Data Center Networks*. Springer Publishing Company, Incorporated, 2012.

[117] L. Zhang, J. Chen, E. Agrell, R. Lin, and L. Wosinska, "Enabling technologies for optical data center networks: Spatial division multiplexing," *Journal of Lightwave Technology*, 2019.

[118] M.-J. Li, "New development trends in optical fibers for data centers," *2018 European Conference on Optical Communication (ECOC)*, pp. 1–3, 2018.

[119] T. Matsui, T. Sakamoto, Y. Goto, K. Saito, K. Nakajima, F. Yamamoto, and T. Kurashima, "Design of 125 μm cladding multi-core fiber with full-band compatibility to conventional single-mode fiber," in *2015 European Conference on Optical Communication (ECOC)*, Sep. 2015, pp. 1–3.

[120] T. Hayashi, T. Taru, O. Shimakawa, T. Sasaki, and E. Sasaoka, "Design and fabrication of ultra-low crosstalk and low-loss multi-core fiber," *Opt. Express*, vol. 19, no. 17, pp. 16 576–16 592, Aug 2011.

[121] B. Li, Z. Feng, M. Tang, Z. Xu, S. Fu, Q. Wu, L. Deng, W. Tong, S. Liu, and P. P. Shum, "Experimental demonstration of large capacity wsdm optical access network with multicore fibers and advanced modulation formats," *Opt. Express*, vol. 23, no. 9, pp. 10 997–11 006, May 2015.

[122] R. Ryf, R. Essiambre, S. Randel, M. A. Mestre, C. Schmidt, and P. J. Winzer, "Impulse response analysis of coupled-core 3-core fibers," in *2012 38th European Conference and Exhibition on Optical Communications*, Sep. 2012, pp. 1–3.

[123] S. O. Arik and J. M. Kahn, "Coupled-core multi-core fibers for spatial multiplexing," *IEEE Photonics Technology Letters*, vol. 25, no. 21, pp. 2054–2057, Nov 2013.

[124] R. Ryf, N. K. Fontaine, M. Montoliu, S. Randel, S. H. Chang, H. Chen, S. Chandrasekhar, A. H. Gnauck, R. . Essiambre, P. J. Winzer, T. Taru, T. Hayashi, and T. Sasaki, "Space-division multiplexed transmission over 3×3 coupled-core multicore fiber," in *OFC 2014*, March 2014, pp. 1–3.

[125] L. Gan, L. Shen, M. Tang, C. Xing, Y. Li, C. Ke, W. Tong, B. Li, S. Fu, and D. Liu, "Investigation of channel model for weakly coupled multicore fiber," *Opt. Express*, vol. 26, no. 5, pp. 5182–5199, Mar 2018.

[126] K. Takenaga, Y. Arakawa, Y. Sasaki, S. Tanigawa, S. Matsuo, K. Saitoh, and M. Koshiba, "A large effective area multi-core fibre with an optimised cladding thickness," in *2011 37th European Conference and Exhibition on Optical Communication*, Sep. 2011, pp. 1–3.

[127] M. Koshiba, K. Saitoh, K. Takenaga, and S. Matsuo, "Multi-core fiber design and analysis: coupled-mode theory and coupled-power theory," *Opt. Express*, vol. 19, no. 26, pp. B102–B111, Dec 2011.

[128] Y. Sasaki, Y. Amma, K. Takenaga, S. Matsuo, K. Saitoh, and M. Koshiba, "Investigation of crosstalk dependencies on bending radius of heterogeneous multicore fiber," in *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, March 2013, pp. 1–3.

[129] T. Ito, E. L. T. de Gabory, M. Arikawa, Y. Hashimoto, and K. Fukuchi, "Reduction of influence of inter-core cross-talk in mcf with bidirectional assignment between neighboring cores," in *2013 Optical Fiber*

Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC), March 2013, pp. 1–3.

[130] J. Sakaguchi, B. J. Puttnam, W. Klaus, Y. Awaji, N. Wada, A. Kanno, T. Kawanishi, K. Imamura, H. Inaba, K. Mukasa, R. Sugizaki, T. Kobayashi, and M. Watanabe, "305 tb/s space division multiplexed transmission using homogeneous 19-core fiber," *Journal of Lightwave Technology*, vol. 31, no. 4, pp. 554–562, Feb 2013.

[131] Y. Goto, K. Nakajima, T. Matsui, T. Kurashima, and F. Yamamoto, "Influence of cladding thickness on transmission loss and its relationship with multicore fiber structure," *Journal of Lightwave Technology*, vol. 33, no. 23, pp. 4942–4949, Dec 2015.

[132] M. Li, B. Hoover, V. N. Nazarov, and D. L. Butler, "Multicore fiber for optical interconnect applications," in *2012 17th Opto-Electronics and Communications Conference*, July 2012, pp. 564–565.

[133] T. Tsuritani, D. Soma, Y. Wakayama, Y. Miyagawa, M. Takahashi, I. Morita, K. Maeda, K. Kawasaki, T. Matsuura, M. Tsukamoto, and R. Sugizaki, "Field test of installed high-density optical fiber cable with multi-core fibers toward practical deployment," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2019, pp. 1–3.

[134] S. Matsuo, K. Takenaga, Y. Arakawa, Y. Sasaki, S. Taniagwa, K. Saitoh, and M. Koshiba, "Large-effective-area ten-core fiber with cladding diameter of about 200μm," *Opt. Lett.*, vol. 36, no. 23, pp. 4626–4628, Dec 2011.

[135] S. Matsuo, K. Takenaga, Y. Arakawa, Y. Sasaki, S. Tanigawa, K. Saitoh, and M. Koshiba, "Crosstalk behavior of cores in multi-core fiber under bent condition," *IEICE Electronics Express*, vol. 8, no. 6, pp. 385–390, 2011.

[136] T. M. F. Alves, R. S. Luís, B. J. Puttnam, A. V. T. Cartaxo, Y. Awaji, and N. Wada, "Performance of adaptive dd-ofdm multicore fiber links and its relation with intercore crosstalk," *Opt. Express*, vol. 25, no. 14, pp. 16 017–16 027, Jul 2017.

[137] T. Hayashi, T. Nakanishi, K. Hirashima, O. Shimakawa, F. Sato, K. Koyama, A. Furuya, Y. Murakami, and T. Sasaki, "125-μm-cladding 8-core multi-core fiber realizing ultra-high-density cable suitable for o-band short-reach optical interconnects," in *Optical Fiber Communication Conference Post Deadline Papers*. Optical Society of America, 2015, p. Th5C.6.

[138] A. Sano, H. Takara, T. Kobayashi, H. Kawakami, H. Kishikawa, T. Nakagawa, Y. Miyamoto, Y. Abe, H. Ono, K. Shikama, M. Nagatani, T. Mori, Y. Sasaki, I. Ishida, K. Takenaga, S. Matsuo, K. Saitoh, M. Koshiba, M. Yamada, H. Masuda, and T. Morioka, "409-Tb/s + 409-Tb/s crosstalk suppressed bidirectional MCF transmission over 450 km using propagation-direction interleaving," *Opt. Express*, vol. 21, no. 14, pp. 16 777–16 783, Jul 2013.

[139] B. J. Puttnam, R. S. Luis, W. Klaus, J. Sakaguchi, J. . Delgado Mendinueta, Y. Awaji, N. Wada, Y. Tamura, T. Hayashi, M. Hirano, and J. Marciante, "2.15 pb/s transmission using a 22 core homogeneous single-mode multi-core fiber and wideband optical comb," in *2015 European Conference on Optical Communication (ECOC)*, Sep. 2015, pp. 1–3.

[140] Y. Amma, Y. Sasaki, K. Takenaga, S. Matsuo, J. Tu, K. Saitoh, M. Koshiba, T. Morioka, and Y. Miyamoto, "High-density multicore fiber with heterogeneous core arrangement," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2015, pp. 1–3.

[141] Y. Sasaki, R. Fukumoto, K. Takenaga, K. Aikawa, K. Saitoh, T. Morioka, and Y. Miyamoto, "Crosstalk-managed heterogeneous single-mode 32-core fibre," in *ECOC 2016; 42nd European Conference on Optical Communication*, Sep. 2016, pp. 1–3.

[142] Y. Sasaki, K. Takenaga, K. Aikawa, Y. Miyamoto, and T. Morioka, "Single-mode 37-core fiber with a cladding diameter of 248 μm," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–3.

[143] K. Mukasa, "100-core fibers," in *24th Optoelectronics and Communications Conference/International Conference on Photonics in Switching and Computing 2019 (OECC/PSC)*, 2019.

[144] M. Koshiba, K. Saitoh, K. Takenaga, and S. Matsuo, "Analytical expression of average power-coupling coefficients for estimating intercore crosstalk in multicore fibers," *IEEE Photonics Journal*, vol. 4, no. 5, pp. 1987–1995, Oct 2012.

[145] F. Ye, J. Tu, K. Saitoh, H. Takara, and T. Morioka, "Wavelength-dependent crosstalk in trench-assisted multi-core fibers," in *2014 OptoElectronics and*

Communication Conference and Australian Conference on Optical Fibre Technology, July 2014, pp. 308–309.

[146] K. Okamoto, "Preface to the second edition," in *Fundamentals of Optical Waveguides (Second Edition)*, K. Okamoto, Ed. Burlington: Academic Press, 2006, pp. xv – xvi.

[147] F. Ye, J. Tu, K. Saitoh, K. Takenaga, S. Matsuo, H. Takara, and T. Morioka, "Wavelength-dependence of inter-core crosstalk in homogeneous multi-core fibers," *IEEE Photonics Technology Letters*, vol. 28, no. 1, pp. 27–30, Jan 2016.

[148] K. Takenaga, Y. Arakawa, S. Tanigawa, N. Guan, S. Matsuo, K. Saitoh, and M. Koshiba, "Reduction of crosstalk by trench-assisted multi-core fiber," in *2011 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference*, March 2011, pp. 1–3.

[149] K. Saitoh, "Multicore fiber technology," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2015, pp. 1–22.

[150] J. Tu, K. Saitoh, M. Koshiba, K. Takenaga, and S. Matsuo, "Design and analysis of large-effective-area heterogeneous trench-assisted multi-core fiber," *Opt. Express*, vol. 20, no. 14, pp. 15 157–15 170, Jul 2012.

[151] F. Ye, J. Tu, K. Saitoh, and T. Morioka, "Simple analytical expression for crosstalk estimation in homogeneous trench-assisted multi-core fibers," *Opt. Express*, vol. 22, no. 19, pp. 23 007–23 018, Sep 2014.

[152] D. Medhi and K. Ramasamy, "Chapter 1 - networking and network routing: An introduction," in *Network Routing (Second Edition)*, ser. The Morgan Kaufmann Series in Networking, D. Medhi and K. Ramasamy, Eds. Boston: Morgan Kaufmann, 2018, pp. 2 – 29.

[153] H. Zang and J. P. Jue, "A review of routing and wavelength assignment approaches for wavelength-routed optical wdm networks," *Optical Networks Magazine*, vol. 1, pp. 47–60, 2000.

[154] Y. Wang, X. Cao, and Y. Pan, "A study of the routing and spectrum allocation in spectrum-sliced elastic optical path networks," in *2011 Proceedings IEEE INFOCOM*, April 2011, pp. 1503–1511.

[155] A. Muhammad, G. Zervas, D. Simeonidou, and R. Forchheimer, "Routing, spectrum and core allocation in flexgrid sdm networks with multi-core fibers," in *2014 International Conference on Optical Network Design and Modeling*, May 2014, pp. 192–197.

[156] A. S. Tanenbaum and D. J. Wetherall, *Computer Networks*, 5th ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010.

[157] A. W. Brander and M. C. Sinclair, *A Comparative Study of k-Shortest Path Algorithms*. London: Springer London, 1996, pp. 370–379.

[158] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, Dec 1959.

[159] D. S. Johnson, "Fast allocation algorithms," in *13th Annual Symposium on Switching and Automata Theory (swat 1972)*, Oct 1972, pp. 144–154.

[160] J. M. Robson, "Worst case fragmentation of first fit and best fit storage allocation strategies," *The Computer Journal*, vol. 20, no. 3, pp. 242–244, 01 1977.

[161] A. Benguedach, S. Niar, and B. Beldjilali, "Online first fit algorithm for modeling the problem of configurable cache architecture," in *ICM 2011 Proceeding*, Dec 2011, pp. 1–6.

[162] Y. Hasan, W. Chen, J. M. Chang, and B. M. Gharaibeh, "Upper bounds for dynamic memory allocation," *IEEE Transactions on Computers*, vol. 59, no. 4, pp. 468–477, April 2010.

[163] X. Sun, Y. Li, I. Lambadaris, and Y. Q. Zhao, "Performance analysis of first-fit wavelength assignment algorithm in optical networks," in *Proceedings of the 7th International Conference on Telecommunications, 2003. ConTEL 2003.*, vol. 2, June 2003, pp. 403–409 vol.2.

[164] S. Fujii, Y. Hirota, H. Tode, and K. Murakami, "On-demand spectrum and core allocation for reducing crosstalk in multicore fibers in elastic optical networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 6, no. 12, pp. 1059–1071, Dec 2014.

[165] K. Igarashi, T. Tsuritani, I. Morita, and M. Suzuki, "Ultra-long-haul high-capacity super-nyquist-wdm transmission experiment using multi-core fibers," *Journal of Lightwave Technology*, vol. 33, no. 5, pp. 1027–1036, March 2015.

[166] L. Velasco, A. Castro, M. Ruiz, and G. Junyent, "Solving routing and spectrum allocation related optimization problems: From off-line to in-operation flexgrid network planning," *Journal of Lightwave Technology*, vol. 32, no. 16, pp. 2780–2795, Aug 2014.

[167] H. Tode and Y. Hirota, "Routing, spectrum, and core and/or mode assignment on space-division multiplexing optical networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 1, pp. A99–A113, Jan 2017.

[168] A. Pagès, J. Perelló, S. Spadaro, and J. Comellas, "Optimal route, spectrum, and modulation level assignment in split-spectrum-enabled dynamic elastic optical networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 6, no. 2, pp. 114–126, Feb 2014.

[169] B. J. Puttnam, R. S. Luis, T. A. Eriksson, W. Klaus, J. . D. Mendinueta, Y. Awaji, and N. Wada, "Impact of intercore crosstalk on the transmission distance of qam formats in multicore fibers," *IEEE Photonics Journal*, vol. 8, no. 2, pp. 1–9, April 2016.

[170] Q. Yao, H. Yang, H. Xiao, Y. Zhao, R. Zhu, and J. Zhang, "Crosstalk-aware routing, spectrum, and core assignment in space-division multiplexing optical networks with multicore fibers," *Optical Engineering*, vol. 56, no. 6, pp. 1 – 9 – 9, 2017.

[171] H. Yang, Q. Yao, A. Yu, Y. Lee, and J. Zhang, "Resource assignment based on dynamic fuzzy clustering in elastic optical networks with multi-core fibers," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3457–3469, May 2019.

[172] Q. Yao, H. Yang, R. Zhu, A. Yu, W. Bai, Y. Tan, J. Zhang, and H. Xiao, "Core, mode, and spectrum assignment based on machine learning in space division multiplexing elastic optical networks," *IEEE Access*, vol. 6, pp. 15 898–15 907, 2018.

[173] T. Hayashi, T. Taru, O. Shimakawa, T. Sasaki, and E. Sasaoka, "Design and fabrication of ultra-low crosstalk and low-loss multi-core fiber," *Opt. Express*, vol. 19, no. 17, pp. 16 576–16 592, Aug 2011.

[174] T. Hayashi, T. Sasaki, and E. Sasaoka, "Behavior of inter-core crosstalk as a noise and its effect on Q-factor in multi-core fiber," *IEICE Transactions on Communications*, vol. E97.B, no. 5, pp. 936–944, 2014.

[175] R. S. Luis, B. J. Puttnam, A. V. T. Cartaxo, W. Klaus, J. M. D. Mendinueta, Y. Awaji, N. Wada, T. Nakanishi, T. Hayashi, and T. Sasaki, "Time and modulation frequency dependence of crosstalk in homogeneous multi-core fibers," *Journal of Lightwave Technology*, vol. 34, no. 2, pp. 441–447, Jan 2016.

[176] G. Rademacher, R. S. Luís, B. J. Puttnam, Y. Awaji, and N. Wada, "Crosstalk dynamics in multi-core fibers," *Opt. Express*, vol. 25, no. 10, pp. 12 020–12 028, May 2017.

[177] A. V. T. Cartaxo, R. S. Luis, B. J. Puttnam, T. Hayashi, Y. Awaji, and N. Wada, "Dispersion impact on the crosstalk amplitude response of homogeneous multi-core fibers," *IEEE Photonics Technology Letters*, vol. 28, no. 17, pp. 1858–1861, Sep. 2016.

[178] T. M. F. Alves and A. V. T. Cartaxo, "Theoretical modelling of random time nature of inter-core crosstalk in multicore fibers," in *2016 IEEE Photonics Conference (IPC)*, Oct 2016, pp. 521–522.

[179] T. Hayashi, "Multi-core fibers for space division multiplexing," in *Handbook of Optical Fibers*, P. GD, Ed. Springer, August 2018.

[180] K. Clark, H. Ballani, P. Bayvel, D. Cletheroe, T. Gerard, I. Haller, K. Jozwik, K. Shi, B. Thomsen, P. Watts, H. Williams, G. Zervas, P. Costa, and Z. Liu, "Sub-nanosecond clock and data recovery in an optically-switched data centre network," in *2018 European Conference on Optical Communication (ECOC)*, Sep. 2018, pp. 1–3.

[181] R. Slavik, G. Marra, E. N. Fokoua, N. Baddela, N. V. Wheeler, M. Petrovich, F. Poletti, and D. J. Richardson, "Ultralow thermal sensitivity of phase and propagation delay in hollow core optical fibres," *Scientific Reports*, vol. 5, no. 15447, October 2015.

[182] M. Ismail, "Thermal effects in optical fibers," Sep. 2009, https://www.researchgate.net/publication/332911460_Metal-coated_fiber_sensor_for_laser_radiation_power_measurements.

[183] B. J. Puttnam, G. Rademacher, R. S. Luis, J. Sakaguchi, Y. Awaji, and N. Wada, "Inter-core skew measurements in temperature controlled multi-core fiber," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[184] M. Rice, "Single-carrier modulation," *Academic Press Library in Mobile and Wireless Communications*, 2016, https://www.sciencedirect.com/topics/engineering/power-spectral-density.

[185] Wiley Online Library, "Communication signals," https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119264422.app1.

[186] T. M. F. Alves, R. O. J. Soeiro, and A. V. T. Cartaxo, "Probability distribution of intercore crosstalk in weakly coupled mcfs with multiple interferers," *IEEE Photonics Technology Letters*, vol. 31, no. 8, pp. 651–654, April 2019.

[187] R. S. Luis, B. J. Puttnam, W. Klaus, J. M. D. Mendinueta, Y. Awaji, N. Wada, A. Kanno, T. Kawanishi, T. Nakanishi, T. Hayashi, and T. Sasaki, "Experimental evaluation of the time and frequency crosstalk dependency in a 7-core multi-core fiber," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2015, pp. 1–3.

[188] T. M. F. Alves and A. V. T. Cartaxo, "Characterization of the stochastic time evolution of short-term average intercore crosstalk in multicore fibers with multiple interfering cores," *Opt. Express*, vol. 26, no. 4, pp. 4605–4620, Feb 2018.

[189] G. Rademacher, B. J. Puttnam, R. S. Luis, Y. Awaji, and N. Wada, "Time-dependent crosstalk from multiple cores in a homogeneous multi-core fiber," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–3.

[190] T. Hayashi, A. Mekis, T. Nakanishi, M. Peterson, S. Sahni, P. Sun, S. Freyling, G. Armijo, C. Sohn, D. Foltz, T. Pinguet, M. Mack, Y. Kaneuchi, O. Shimakawa, T. Morishima, T. Sasaki, and P. D. Dobbelaere, "End-to-end multi-core fibre transmission link enabled by silicon photonics transceiver with grating coupler array," in *2017 European Conference on Optical Communication (ECOC)*, Sept 2017, pp. 1–3.

[191] Santec, "Multi-port optical power meter," https://www.santec.com/en/wp-content/uploads/MPM-210-C-E-v1.1web.pdf.

[192] T. Hayashi, T. Taru, O. Shimakawa, T. Sasaki, and E. Sasaoka, "Ultra-low-crosstalk multi-core fiber feasible to ultra-long-haul transmission," in *2011 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference*, March 2011, pp. 1–3.

[193] N. Yang, Q. Qiu, J. Su, and S. Shi, "Research on the temperature characteristics of optical fiber refractive index," *Optik*, vol. 125, no. 19, pp. 5813 – 5815, 2014.

[194] H. W. Coleman and W. G. Steele, *Experimentation, Validation, and Uncertainty Analysis for Engineers*. John Wiley & Sons, Ltd, 2009, pp. i–xvi.

[195] "Cross-correlation," https://ccrma.stanford.edu/$\sim$jos/mdft/Cross$_$Correlation.html.

[196] W. Rowe, "Mean squared error, r2, and variance in regression analysis," July 2018, https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/.

[197] W. Klaus, B. J. Puttnam, R. S. Luis, J. Sakaguchi, J. D. Mendinueta, Y. Awaji, and N. Wada, "Advanced space division multiplexing technologies for optical networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 4, pp. C1–C11, April 2017.

[198] B. Lee, D. Kuchta, F. Doany, C. Schow, C. Baks, R. John, P. Pepeljugoski, T. Taunay, B. Zhu, M. Yan, G. Oulundsen, D. Vaidya, W. Luo, and N. Li, "120-gb/s 100-m transmission in a single multicore multimode fiber containing six cores interfaced with a matching vcsel array," *2010 IEEE Photonics Society Summer Topical Meeting Series, PHOSST 2010*, pp. 223 – 224, 08 2010.

[199] H. C. H. Mulvad, A. Parker, B. King, D. Smith, M. Kovacs, S. Jain, J. Hayes, M. Petrovich, D. J. Richardson, and N. Parsons, "Beam-steering all-optical switch for multi-core fibers," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–3.

[200] F. Ye, K. Saitoh, H. Takara, R. Asif, and T. Morioka, "High-count multi-core fibers for space-division multiplexing with propagation-direction interleaving," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2015, pp. 1–3.

[201] M. Fiorani, M. Tornatore, J. Chen, L. Wosinska, and B. Mukherjee, "Spatial division multiplexing for high capacity optical interconnects in modular data centers," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 2, pp. A143–A153, Feb 2017.

[202] NTT Electronics, "Awg multi/demultiplexer," http://www.ntt-electronics.com/en/products/photonics/awg_mul_d.html/.

[203] Finisar, "9/1 × 20 flexgrid wavelength selective switch," https://www.finisar.com/roadms-wavelengthmanagement/10wsaaxxfll.

[204] A. Muhammad, G. Zervas, and R. Forchheimer, "Resource allocation for space-division multiplexing: Optical white box versus optical black box networking," *Journal of Lightwave Technology*, vol. 33, no. 23, pp. 4928–4941, Dec 2015.

[205] P. J. Winzer, A. H. Gnauck, A. Konczykowska, F. Jorge, and J. . Dupuy, "Penalties from in-band crosstalk for advanced optical modulation formats," in *2011 37th European Conference and Exhibition on Optical Communication*, Sep. 2011, pp. 1–3.

[206] IEEE P802.3bm 40 Gb/s and 100 Gb/s Fiber Optic Task Force, "Pam8 & fec options," Nov 2012, www.ieee802.org/3/bm/public/nov12.

[207] N. Amaya, M. Irfan, G. Zervas, K. Banias, M. Garrich, I. Henning, D. Simeonidou, Y. R. Zhou, A. Lord, K. Smith, V. J. F. Rancano, S. Liu, P. Petropoulos, and D. J. Richardson, "Gridless optical networking field trial: Flexible spectrum switching, defragmentation and transport of 10g/40g/100g/555g over 620-km field fiber," in *2011 37th European Conference and Exhibition on Optical Communication*, Sep. 2011, pp. 1–3.

[208] 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.

[209] Huawei Technologies Co., Ltd. and IEEE 802.3 400 GbE Study Group, "Opportunities for pam-4 modulation," Jan 2014, http://www.ieee802.org/3/400GSG/public/14_01/.

[210] M. A. Mestre, H. Mardoyan, A. Konczykowska, R. Rios-Müller, J. Renaudier, F. Jorge, B. Duval, J. Dupuy, A. Ghazisaeidi, P. Jennevé, and S. Bigo, "Direct detection transceiver at 150-gbit/s net data rate using pam 8 for optical interconnects," in *2015 European Conference on Optical Communication (ECOC)*, Sep. 2015, pp. 1–3.

[211] IEEE802.3 NG100GE PMD Study Group, "Update on technical feasibility for pam modulation," Mar 2012, http://www.ieee802.org/3/100GNGOPTX/public/mar12/plenary/.

[212] H. Tode and Y. Hirota, "Routing, spectrum and core assignment for space division multiplexing elastic optical networks," in *2014 16th International Telecommunications Network Strategy and Planning Symposium (Networks)*, Sep. 2014, pp. 1–7.

[213] K. Saitoh, M. Koshiba, K. Takenaga, and S. Matsuo, "Crosstalk and core density in uncoupled multicore fibers," *IEEE Photonics Technology Letters*, vol. 24, no. 21, pp. 1898–1901, Nov 2012.

[214] "Lc product specification outline," http://lcalliance.net/lcInterface/pdfs/LC-Product-Spec.pdf.

[215] USCONEC, "C9730 datasheet," http://www.usconec.com/images/drawings/C9730.pdf.

[216] B. Abali, R. J. Eickemeyer, H. Franke, C. Li, and M. Taubenblatt, "Disaggregated and optically interconnected memory: when will it be cost effective?" *CoRR*, vol. abs/1503.01416, 2015.

[217] Intel, "ARK | processor feature filter," http://ark.intel.com/search/advanced?s=t&FamilyText=IntelÂőXeonÂőProcessorE5v4Fami-ly&BornOnDate=Q1‖16&CoreCountMin=8&CoreCountMax=10.

[218] R. Urata, H. Liu, X. Zhou, and A. Vahdat, "Datacenter interconnect and networking: From evolution to holistic revolution," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2017, pp. 1–3.

[219] C. Minkenberg, "HPC networks: Challenges and the role of optics," in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2015, pp. 1–3.

[220] M. Alicherry and T. V. Lakshman, "Network aware resource allocation in distributed clouds," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 963–971.

[221] V. D. Justafort and S. Pierre, "Performance-aware virtual machine allocation approach in an intercloud environment," in *2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, April 2012, pp. 1–4.

[222] R. K. Sharma, P. Kamal, and S. P. Singh, "A latency reduction mechanism for virtual machine resource allocation in delay sensitive cloud service," in

2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Oct 2015, pp. 371–375.

[223] dReDBox, "D5.4–software and hardware system integration and evaluation," 2018.

[224] Luxtera, "Luxtera silicon photonics optical transceivers," http://www.luxtera.com/products/.

[225] T. Sakamoto, K. Saitoh, N. Hanzawa, K. Tsujikawa, L. Ma, M. Koshiba, and F. Yamamoto, "Crosstalk suppressed hole-assisted 6-core fiber with cladding diameter of 125 μm," in *39th European Conference and Exhibition on Optical Communication (ECOC 2013)*. IET, 2013, pp. 1–3.

[226] X. Xie, J. Tu, X. Zhou, K. Long, and K. Saitoh, "Design and optimization of 32-core rod/trench assisted square-lattice structured single-mode multi-core fiber," *Optics express*, vol. 25, no. 5, pp. 5119–5132, 2017.