Passive Optical Top-Of-Rack Interconnect for Data Center Networks

YUXIN CHENG

Licentiate Thesis in Information and Communication Technology
School of Information and Communication Technology
KTH Royal Institute of Technology
Stockholm, Sweden 2017

KTH School of Information and
Communication Technology
SE-164 40 Kista
SWEDEN

TRITA-ICT 2017:12
ISBN 978-91-7729-387-3

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framläggs till offentlig granskning för avläggande av licentiatexamen i Informations- och kommunikationsteknik måndagen den 12 juni 2017 klockan 10:00 i Ka-Sal C (Sal Sven-Olof Öhrvik), Electrum, Kungl Tekniska högskolan, Kistagången 16, Kista.

Tryck: Universitetsservice US AB

Abstract

Optical networks offering ultra-high capacity and low energy consumption per bit are considered as a good option to handle the rapidly growing traffic volume inside data center (DCs). Consequently, several optical interconnect architectures for DCs have already been proposed. However, most of the architectures proposed so far are mainly focused on the aggregation/core tiers of the data center networks (DCNs), while relying on the conventional top-of-rack (ToR) electronic packet switches (EPS) in the access tier. A large number of ToR switches in the current DCNs brings serious scalability limitations due to high cost and power consumption. Thus, it is important to investigate and evaluate new optical interconnects tailored for the access tier of the DCNs.

We propose and evaluate a passive optical ToR interconnect (POTORI) architecture for the access tier. The data plane of POTORI consists mainly of passive components to interconnect the servers within the rack as well as the interfaces toward the aggregation/core tiers. Using the passive components makes it possible to significantly reduce power consumption while achieving high reliability in a cost-efficient way.

Meanwhile, our proposed POTORI's control plane is based on a centralized rack controller, which is responsible for coordinating the communications among the servers in the rack. It can be reconfigured by software-defined networking (SDN) operation. A cycle-based medium access control (MAC) protocol and a dynamic bandwidth allocation (DBA) algorithm are designed for POTORI to efficiently manage the exchange of control messages and the data transmission inside the rack.

Simulation results show that under realistic DC traffic scenarios, POTORI with the proposed DBA algorithm is able to achieve an average packet delay below 10 μs with the use of fast tunable optical transceivers. Moreover, we further quantify the impact of different network configuration parameters (i.e., transceiver's tuning time, maximum transmission time of each cycle) on the average packet delay. The results suggest that in order to achieve packet-level switching granularity for POTORI, the transceiver's tuning time should be short enough (i.e., below 30% of the packet transmission time), while for the case of a long tuning time, an acceptable packet delay performance can be achieved if the maximum transmission time of each cycle is greater than three times of transceiver's tuning time.

Keywords: Optical communications, data center interconnects, MAC protocol, dynamic bandwidth allocation.

Sammanfattning

Optiska nätverk erbjuder extremt hög kapacitet och låg energikonsumtion per bit och anses därför vara ett bra alternativ för att klara den snabbt ökande trafikvolymen inuti datacenter (DCs). Till följd av detta så har flertalet arkitekturer för sammankoppling av optiska nätverk redan presenterats. De flesta hittills föreslagna arkitekturer fokuserar dock på aggregering av lager i datacenters nätverk (DCN) och förlitar sig konventionell top-of-rack (ToR) elektroniska paket switchar (EPS) i access lagret. Ett stort antal av ToR switchar i nuvarande DCNs leder till allvarliga begränsningar när det kommer till skalbarhet på grund av hög kostnad och energiförbrukning.

Vi föreslår och utvärderar en passiv optisk ToR sammankoppling (POTORI) arkitektur för accesslagret. Datalagret of POTORI består mestadels av passiva komponenter för att sammankoppla servrar i samma serverrack samt kommunikationsgränssnitt mot aggregations/grund lagren. Användning av passiva komponenter gör det möjligt att markant minska energiförbrukningen och samtidigt uppnå hög pålitlighet på ett kostnadseffektivt sätt.

Vår föreslagna POTORIs kontrollager är baserat på en centraliserad rackkontroller som är ansvarig för koordinering av kommunikation mellan servrarna i serverracket. Den går att konfigurera med software-defined networking (SDN) operationer. Ett cykelbaserat medium access control (MAC) protokoll och en dynamisk bandbreddsallokering (DBA) algorithm har designats för att POTORI ska kunna effektivt hantera utbyte av kontrollmeddelanden och dataöverföring inuti serverracket.

Resultat från simuleringar visar att under realistiska DC trafikförhållanden, POTORI med föreslagna DBA algoritmer kan uppnå en snitt paketfördröjning på under 10 μs vid användning av snabbt justerbara optiska transceivers. Dessutom kvantifierar vi påverkan av nätverkskonfigurations parametrar (t.ex. transceivers justeringstid, maximal sändningstid för varje cykel) på snitt paketfördröjningen. Resultaten visar på att för att kunna uppnå paketnivå switching granularitet för POTORI så måste justeringstiden för transceivern vara tillräckligt kort (under 30% av paketsändningstiden), medan för fallet med lång justeringstid, en acceptabel paketfördröjning prestanda kan uppnås om den maximala sändningstiden för varje cykel är större än tre gånger transceiverns justeringstid.

Keywords: Optiska kommunikation, data center sammankoppling, MAC protokoll, dynamisk bandbreddsallokering.

Acknowledgements

Study and work as a Ph.D. student at KTH is one of the best decisions I have ever made in my life.

Firstly, I would like to express my sincere gratitude to my supervisor Associate Professor Jiajia Chen for accepting me as her Ph.D. student and for all her guidance and support during these years. I also want to offer my special thanks to my co-supervisors Professor Lena Wosinska and Dr. Matteo Fiorani for the continuous support and countless invaluable discussions for my Ph.D. study. I feel really happy and lucky to work with my supervisors.

I would like to thank Associate Professor Markus Hidell for the advance review of my licentiate thesis with the insightful and helpful comments and feedbacks. I am also grateful to Dr. Qiong Zhang for accepting the role of opponent of my licentiate defense and my friend Kim Persson for helping with Swedish translation of the abstract.

I also like to express my appreciation to my colleges working in the VR Data Center project for their support and sharing their knowledge. Also I would like to thank all my friends and colleges in the Optical Network Lab (ONLab) for creating a friendly work environment.

Last but not the least, I would like to thank my family: my mother Huaixin Tao, my father Gang Cheng, and my girlfriend Xi Li for all their endless love, encouragement and support. Thank you.

Yuxin Cheng,
Stockholm, April 2017.

Contents

List of Figures

List of Tables

List of Acronyms

AWG	Arrayed waveguide gratings
BvN	Birkhoff-von-Neumann
BS	Base station
CAGR	Compound annual growth rate
CSMA/CD	Carrier sense multiple access with collision detection
DBA	Dynamic bandwidth allocation
DC	Data center
DCN	Data center network
E/O	Electrical-to-optical
EPON	Ethernet passive optical networks
EPS	Electronic packet switch
HEAD	High-efficient distributed access
LF	Largest first
MAC	Media access control
MPCP	Multipoint control protocol
MTBF	Mean time between failures
MTTR	Mean time to repair
OCS	Optical circuit switch
O/E	Optical-to-electrical
OLT	Optical line terminal
ONI	Optical network interface
ONU	Optical network unit
POI	Passive optical interconnect
POTORI	Passive optical top-of-rack interconnect
RBD	Reliability block diagram
RX	Receiver
SFP	Small form-factor pluggable transceiver
TD	Traffic demand
ToR	Top-of-Rack

WDM	Wavelength division multiplexing
WTF	Wavelength tunable filter
WTT	Wavelength tunable transmitter
WSS	Wavelength selective switch

List of Papers

Papers Included in the Thesis

Paper I. **Y. Cheng**, M. Fiorani, L. Wosinska, and J. Chen, "Reliable and Cost Efficient Passive Optical Interconnects for Data Centers," in *IEEE Communications Letters*, vol. 19, pp. 1913-1916, Nov. 2015.

Paper II. **Y. Cheng**, M. Fiorani, L. Wosinska, and J. Chen,"Centralized Control Plane for Passive Optical Top-of-Rack Interconnects in Data Centers," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016.

Paper III. **Y. Cheng**, M. Fiorani, R. Lin, L. Wosinska, and J. Chen,"POTORI: A Passive Optical Top-of-Rack Interconnect Architecture for Data Centers," in *IEEE/OSA Journal of Optical Communications and Networking (JOCN)*, to appear, 2017.

Chapter 1

Introduction

The overall data center (DC) traffic has been dramatically increasing since the last decade, due the continuously growing popularity of modern Internet applications, such as cloud computing, video streaming, social networking, etc. Fig. 1.1 shows Cisco statistics forecasting that DC traffic will keep increasing at a compound annual growth rate (CAGR) of 27% up to 2020, reaching 15.3 zettabyte per year [1]. It is also expected that by 2020 the majority, i.e. 77%, of the total DC traffic will stay within the DCs [**?**].

The rapidly increasing intra-DC traffic makes it important to upgrade the current data center network (DCN) infrastructures. For example, Facebook has upgraded their servers and switches to support 10 Gb/s transmission data rate [**?**]. Dell proposed DCN design for 40G and 100G Ethernet [**?**]. However, developing large (in terms of the number of ports) electronic packet switch (EPS) operating at high data rates is challenging, due to the power consumption and bottleneck of I/O bandwidth of the chip [**?**]. For large-scale DCs, there would be a high volume of the EPSs deployed in DCN to scale to a huge number of servers. This leads to a serious energy consumption problem [5]. It has been reported in [6] that the EPS in DCN accounts for 30% of the total energy consumption of the IT devices (including servers, storages, switches, etc.) in the DCs. One important reason of such high energy consumption of DCN is that there is a great number of power demanding electrical-to-optical (E/O) and optical-to-electrical (O/E) conversions deployed in DCN. Currently, optical fibers are used in DCNs only for the data transmission between the servers and switches. Small form-factor pluggable transceivers (SFP) are deployed on both server and switches for the E/O and O/E conversions, since EPSs are switching and processing data in the electronic domain.

1.1 Problem Statement

In this regard, optical interconnects are considered to be a promising solution to solve the power consumption problem of the DCNs. Comparing to the EPS, optical

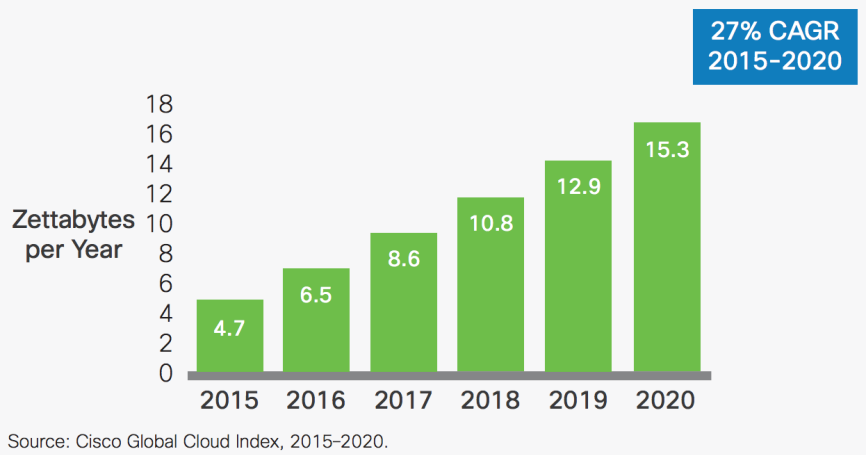


Figure 1.1: Global Data Center IP Traffic Growth [1]

interconnects are able to support high transmission rates and switching capacity in a cost- and energy-efficient way. By replacing the EPS with the optical interconnects, the overall cost and power consumption of the DCN will decrease dramatically, due the reduction of E/O and O/E conversions [7].

In recent years, different optical interconnect architectures for the DCNs have been proposed in the literature. Some of the proposed architectures (e.g., c-through [8] and HOS [9]) are hybrid solutions, where both the EPS and optical interconnect are used. Particularly, the EPS is used to transmit short-lived traffic flows (e.g., mice flows) and optical circuit switch (OCS) is used to transmit long-lived and bandwidth-consuming traffic flows (e.g., elephant flows). However, these architectures require the prediction or the classification of the data traffic to distinguish small and large flows so that the OCS can be properly configured, which might be challenging for DC operators.

The other proposed optical architectures (e.g., [10] - [12]) are all-optical, where optical switches are deployed in DCN to replace the EPS. However, most of the proposed all-optical interconnects mainly target the aggregation and core tier of the DCNs, where the access tier always relies on electronic top-of-rack (ToR) switches, i.e., one or multiple conventional EPSs are used to connect all the servers in one rack. Due to the strong locality of traffic pattern for some applications (e.g. MapReduce) in DCN [13], the access tier carries a large amount of overall data center traffic. The electronic ToR switches are responsible for the major part of cost and energy consumption [14].

So far, there are not too many works focusing on the optical architecture for the access tier of DCN. Therefore, it is essential to design efficient optical interconnect architectures for the access tier of DCN. Optical architectures have the advantages in terms of the cost and reliability comparing to the EPS, due to the less number of

power-hungry components used. The numerical results on the cost and reliability will further illustrate these advantages. Moreover, network performance (i.e., packet delay, packet drop ratio) of the optical interconnects should be evaluated. The network performance should be competitive with EPS, otherwise it will be hard to convince DC operators to deploy the optical architecture solutions at the expense of increasing packet delay or packet drop ratio.

1.2 Contribution of the Thesis

This thesis presents POTORI: a passive optical top-of-rack interconnect that is designed for the access tier of the DCNs. The data plane of POTORI is mainly based on passive optical components to interconnect servers in a rack. On the other hand, to avoid traffic conflict, POTORI requires a proper control protocol to efficiently coordinate the data transmission inside the rack. The contribution of the thesis can be divided into the design of the data plane and the control plane of POTORI.

1.2.1 Reliable and Cost-Efficient Data Plane Design of POTORI

Modern fault-tolerant data centers require very high availability of the overall infrastructure. As the result, the availability of the connection established for the communication among the servers should be even higher. In POTORI, the passive nature of the interconnect components brings the obvious advantages in terms of the cost, power consumption and reliability performance. **Paper I** of this thesis presents the data plane design of POTORI and provides a cost and reliability analysis. The results show that POTORI is able to achieve intra-rack connection availability higher than 99.995% in a cost-efficient way.

1.2.2 Centralized Control Plane Design of POTORI

Paper II and **III** of the thesis present a novel control plane tailored for POTORI. The control plane of POTORI is based on a rack controller, which manages the communications inside a rack. Each server exchanges control messages with the rack controller through a dedicated control link. A media access control (MAC) protocol for POTORI defines the procedure of the control message exchange and data transmission in the rack. Moreover, the rack controller is running the proposed dynamic bandwidth allocation (DBA) algorithm determining the resource (i.e., wavelength and time) allocation used by the servers.

1.3 Research Methodology

We apply a quantitative method in our research project. First, we propose the data plane of the passive optical interconnects (POIs) architecture which addresses the

aforementioned issues. Numerical results of cost and availability are calculated by applying the cost and reliability model to the different schemes of the POIs. Then, we design the control plane (including a media access control (MAC) protocol and a dynamic bandwidth allocation (DBA) algorithm) for the proposed POI. The performance (i.e., average packet delay, packet drop ratio) of the architecture is evaluated by a customized event-driven simulator. Finally, we are planning to experimentally evaluate the architecture, including both data plane and control plane, in the future work.

1.4 Sustainability Aspects

As academic researchers, we should contribute to a sustainable world. We consider three major types of the sustainability in our research: environmental, economic, and societal sustainability.

Environmental Sustainability

The increasing energy consumed by DCs is becoming a more and more challenging issue. A non-neglected proportion (about 4% to 12% [5]) of total power is consumed by DCN. By replacing the electronic packet ToR switches with the proposed POI in the thesis, the total power consumption of the DCN can be reduced significantly.

Economic Sustainability

Optical interconnects are considered more cost-efficient and reliable compared to the modern electronic packet switches [7]. As mentioned in the previous sections, DC operators are considering optical architectures for the aggregation and core tier of DCN. The work presented in this thesis proposed is the POI for the access tier of DCN, which can be integrated to the existing optical architectures seamlessly.

Societal Sustainability

Normally, the ordinary users will not own private DCs. However, by saving the bills on the cost and power consumption, DC operators can offer services with lower price, which makes all kinds of applications running in DCs more affordable by common users.

1.5 Organization of the Thesis

The thesis is organized as follows:

- Chapter 2 introduces different passive optical interconnect (POI) architectures that can be used as POTORI's data plane. Specifically, the cost and

reliability models as well as the corresponding numerical results of these POIs are presented.

- Chapter 3 presents the detailed control plane design of POTORI, including the proposed MAC protocol and DBA algorithm. In the simulation results, the performance in terms of the average packet delay and packet drop ratio is compared with the EPS. Moreover, the impact of different network configuration of POTORI on the average packet delay is presented.
- Chapter 4 concludes the thesis and highlights the possible future work.
- Finally, there is a brief summary of the papers included in the thesis along with the candidate's contributions to each paper.

Chapter 2

Reliable and Cost Efficient Data Plane of POTORI

Modern data center operators are upgrading their network devices (e.g. switches, routers) to higher data rates (e.g., 10 Gb/s) in order to serve the fast increasing traffic volume within data center networks [2], while in the future even higher data rates, i.e., 40 Gb/s and 100 Gb/s, are expected to be used [3]. As a result, the cost and energy consumption will increase dramatically in order to scale data center network to such high transmission capacity. On the other hand, the higher transmission rate, the greater volume of data center traffic will be affected in case of a network failure. A fault-tolerant data center infrastructure, including electrical power supply for servers and network devices as well as storage system and distribution facilities, should be able to achieve high availability (e.g., 99.995% [15]). Consequently, connection availability in data center networks (DCNs) should be higher than the required availability level of the total data center infrastructure, since the DCN is only a part of the overall service chain offered by the data center infrastructure. Different topologies (e.g., fat-tree[16], Quartz[17]) are proposed to improve the resiliency by providing redundancy in the aggregation and core tiers of DCN. However, the access tier is usually unprotected due to the high cost of introducing redundant ToR switches for every rack in data center.

Meanwhile, the use of an optical optical interconnect is a promising solution to solve the scalability problem brought by the conventional EPS DCN. Particularly, passive optical interconnects (POIs) are able to support ultra-high capacity in reliable, energy- and cost-efficient way due to the passive feature of the applied optical components (e.g., couplers, arrayed wavelength gratings (AWGs)). Many works have been done to show the advantage of POI in terms of cost and energy consumption [7] [14], but the reliability performance of POI is first addressed in the frame of this thesis.

This chapter presents and analyzes different reliable and cost-efficient POI based schemes that can be used as POTORI's data plane. Moreover, one of the schemes

can further enhance the reliability by introducing extra redundant components. The reliability and cost models of these schemes are described and the numerical results in terms of the cost and connection unavailability are shown and compared with the conventional EPS.

2.1 Passive Optical Interconnects

Paper I presents three POIs, see Fig. 2.1, that can be used as the data plane of POTORI. In these three POI schemes, each server in a rack is equipped with an optical network interface (ONI), which consists of a wavelength tunable transceiver. It allows one server to transmit and receive data on different wavelengths in a given spectrum range (e.g., C-band). The following paragraphs briefly introduce these three POI schemes.

Scheme I: AWG based POI

The POI shown in Fig 2.1 (a) uses an (N+K)×(N+K) arrayed waveguide grating (AWG) as the switching fabric where each wavelength tunable transmitter (WTT) and receiver (RX) in ONI is connected to a pair of input and output ports of the AWG, respectively. Here N is the maximum supported number of servers for a rack and K is the number of uplink ports that can be connected to the other racks or the switches in aggregation/core tier. This scheme is inspired by the POI proposed in [18]. In this scheme, N+K wavelengths are required to support intra-rack communications between any pair of servers within the rack and inter-rack communications between servers and uplink ports. Fig 2.2 gives a proper wavelength plan for Scheme I based on the cyclic property of the AWG. The grey fields in Fig 2.2 indicate that no wavelength is needed, since there is no traffic passing through the POI destined to the same source server (i.e. fields in the diagonal) or between K uplink ports (i.e. fields in the right bottom corner) or between different ports connecting to the outside of the rack.

Scheme II: (N+1)×(N+1) coupler based POI

Fig 2.1 (b) shows Scheme II. In this POI, an (N+1)×(N+1) coupler interconnects N servers in a rack. Similar to Scheme I, each ONI on server is connected to one pair of the input and output ports of the coupler. One input and output port of the coupler is reserved to connect to a wavelength selective switch (WSS). Unlike AWG-based Scheme (i.e., Scheme I) which requires a fixed predetermined wavelength plan, Scheme II has higher flexibility in wavelength allocation due the broadcast nature of the coupler. The WTT in ONI is able to transmit data traffic on any available wavelength. The data will be broadcast to all the output ports of the coupler. A wavelength tunable filter (WTF) in ONI is used to select the specific wavelength assigned to the communication and filter out the rest of signals. The

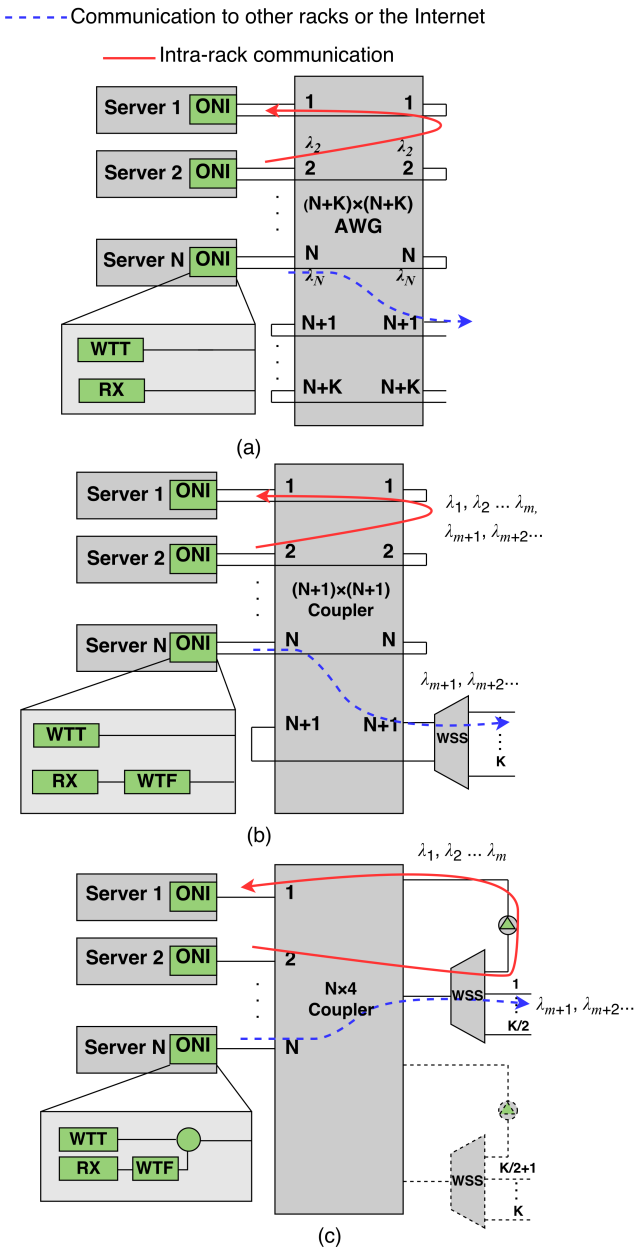


Figure 2.1: (a) Scheme I: (N+K)×(N+K) AWG based POI, (b) Scheme II: (N+1)×(N+1) coupler based POI and (c) Scheme III: N×4 coupler based POI (WTT: Wavelength Tunable Transmitter, AWG: Arrayed Waveguide Grating, ONI: Optical Network Interface, RX: Receiver, WTF: Wavelength Tunable Filter, WSS: Wavelength Selective Switch). © 2015 IEEE (**Paper I**)

From \ To		Within the rack				Outside of the rack		
		Server 1	Server 2	…	Server N	Interface 1	…	Interface K
Within the rack	Server 1		λ_2	…	λ_N	λ_{N+1}	…	λ_{N+K}
	Server 2	λ_2		…	λ_{N+1}	λ_{N+2}	…	λ_1
	…	…	…		…	…	…	…
	Server N	λ_N	λ_{N+1}	…		λ_{N-K}	…	λ_{N-1}
Outside of the rack	Interface 1	λ_{N+1}	λ_{N+2}	…	λ_{N-K}			
	…	…	…	…	…			
	Interface K	λ_{N+K}	λ_1	…	λ_{N-1}			

Figure 2.2: Wavelength plan for AWG based POI. © 2015 IEEE (**Paper I**)

WSS will also select the wavelengths assigned to the inter-rack communication and block the wavelengths for the intra-rack communication.

Scheme III: N×4 coupler based POI

Scheme III is shown in Fig 2.1 (c). It enhances the reliability of POI which is proposed in [7]. In this scheme, the ONI on server is connected to only one side of the coupler. The ports on another side of the coupler are connected to a WSS. All the traffic sent by servers is received first by the WSS, which loops back the wavelengths assigned to the intra-rack communication to the coupler, and forwards the wavelength assigned to the inter-rack communication through the rest of interfaces. Similar to the Scheme II, a WTF is needed in the ONI to select the signal destined to the corresponding server. In this scheme, WSS is the key component since all the traffic will pass it and it is responsible to separate intra- and inter-rack data traffic based on the wavelength assignment. WSS is an active component which has lower availability than the passive component (i.e., coupler). A backup WSS is introduced to further improve the reliability performance of this POI.

2.2 Reliability and Cost Model

In this section, we focus on the analysis of intra-rack communication. The same methodology can be applied to inter-rack communication or aggregation/core tier. Fig 2.3 shows the reliability block diagrams (RBDs) of the intra-rack communication for the EPS and the three POIs described in the previous chapter. RBD illustrates the availability model of a system or connection. Series configuration represent the system (or connection) which is available only and only if all the connected blocks are available. On the other hand, in parallel configuration at least one branch of connected blocks need to be available. Here, each block of RBD represents different active or passive component for the intra-rack communication. We compare the connection availability of Scheme I, Scheme II, and Scheme III (with and without protection) with connection availability of the regular EPS

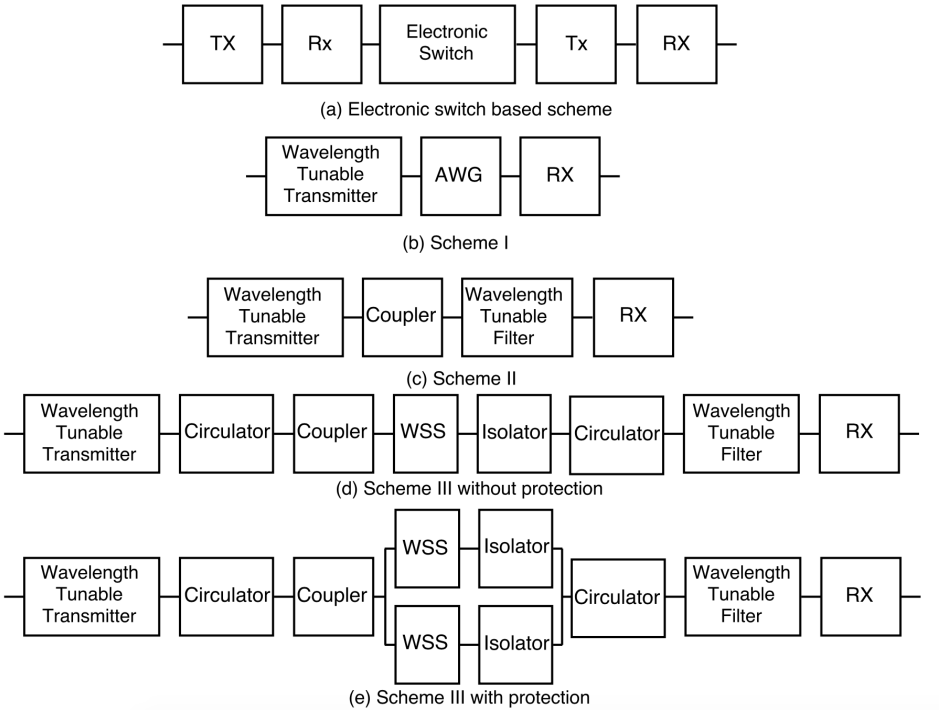


Figure 2.3: Reliability block diagrams. © 2015 IEEE (**Paper I**)

based scheme. Connection availability of a scheme is defined as the probability that the connection between two transceivers within a rack has not failed. In Fig. 2.3 (a) - (d), the overall availability of intra-rack communication can be derived by calculating the product of the availabilities of each individual component (block). In Fig. 2.3 (e) some redundant components are connected in parallel, so the overall connection availability is improved compared to the unprotected schemes. More details about the reliability models are given in **Paper I**.

The total cost of POI is calculated as the sum of the cost of all the network components inside a rack. First, the cost of a single ONI is the sum of the cost of the components that the ONI is built of (e.g., WTT, RX, WTF). Then, the total cost of ONIs in N servers inside a rack can be calculated as the cost of single ONI multiplied by N. Finally, the cost of a POI can be obtained by adding the cost of the remaining components (e.g., coupler, AWG, WSS, etc.) to the total cost of the ONIs. More details on the cost model can be found in **Paper I**.

Table 2.1: MTBF and cost of the network elements [**?**].

Components	**MTBF**[1]	**Cost**
10GE Electronic Switch	150 000 h	3 CU[2](port)
10Gbps Grey Transceiver	600 000 h	0.5CU
10Gbps Tunable Transceiver	500 000 h	1.3CU
WSS	300 000 h	8.3CU
AWG	4 000 000 h	0.1 CU(port)
Coupler	6 000 000 h	0.02 CU(port)
Isolator	12 000 000 h	0.3CU
Circulator	12 000 000 h	0.7 CU
Wavelength Tunable Filter	4 000 000 h	0.3 CU

1. Mean Time Between Failures
2. CU is the cost unit. 1 CU = 150 USD.

2.3 Performance Evaluation

With the reliability model and cost model presented in the previous section, we can evaluate performance of the proposed POIs in terms of connection unavailability and cost. We consider a rack with 48 servers, and the transmission data rate of 10 Gb/s. We compare POI based schemes to the EPS scheme. The results are shown in Fig. 2.4. Table 2.1 shows the mean time between failures (MTBF) and cost of each network component in the POIs and EPS. Note that for the y axis of Fig. 2.4, the unavailability of a system is defined as the probability that system fails at an arbitrary instant of time, and it can be defined as 1 - A, where A is the availability of the system. The calculation of the unavailability values is based on MTBF of components and mean time to repair (MTTR). The MTTR is dependent on the data center operator's maintenance policy. In Fig 2.4, we consider two type of MTTR (4h in (a) and 24h in (b)) representing fast and slow reparation time based on different policies.

Ideally, the data center operator would prefer a scheme with low cost and low unavailability of the connection. It can be seen that all of the proposed POIs show a great advantage compared to the EPS. Specifically, Scheme I and Scheme II perform better than other schemes, i.e. they have the lowest cost and also obtain much lower unavailability compared to the other schemes. On the other hand, the unprotected Scheme III shows the higher cost due to the extra circulator in each ONI, and it has the similar connection unavailability as the EPS. However, the protected Scheme III with a redundant WSS further improves the availability to the similar level as obtained by Scheme I and Scheme II, at the expense of a slightly increased cost. More detailed analysis and comparison of the performance can be found in **Paper I**.

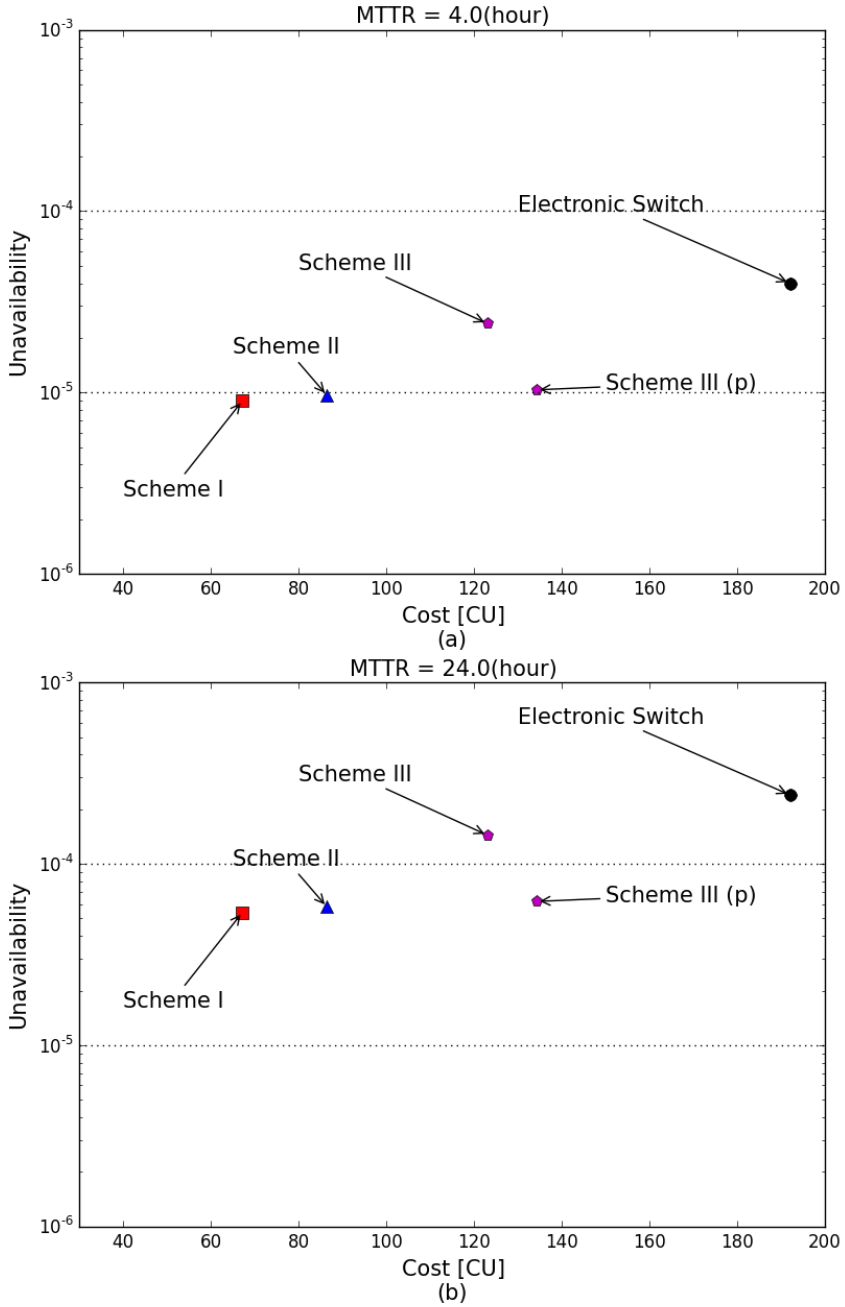


Figure 2.4: Unavailability v.s. total cost of three POIs for different MTTR values.

Chapter 3

Centralized Control Plane of POTORI

As mentioned in the previous chapter, the coupler-based passive POIs have more flexibility in the wavelength allocation for managing the communication inside a rack comparing to the AWG-based POI. Thus, in this chapter, we focus on the control plane for the couple-based POIs. Note that a similar control plane approach can be applied to AWG-based POI as well.

With wavelength tunable transceiver deployed in ONI, the data transmission in POTORI is done in the optical domain. Specifically, the transmitter of the source server and receiver of the destination server need to be tuned to the same wavelength so that a successful data transmission can be achieved. Moreover, concurrent communications inside the rack must be carried on different wavelengths in order to avoid collision in the coupler. As a consequence, a proper control plane design is needed for managing the intra-rack and inter-rack communication in both spectrum and time domains.

This chapter gives an overview of the control plane of POTORI. Our proposed POTORI's control plane relies on a centralized control entity namely rack controller. The rack controller is in charge of running resource allocation algorithm and communicating with servers by exchanging control messages. The tailored MAC protocol and dynamic bandwidth allocation (DBA) algorithms are shown in the following subsections. Finally, the performance of the proposed control schemes in terms of the average packet delay and packet drop ratio are illustrated and analyzed.

3.1 Overview of the control plane of POTORI

The centralized control plane of POTORI is referred to as a rack controller, which is shown in Fig 3.1. The rack controller and servers' ONIs are connected via dedicated control links. In order to coordinate the communications of servers, the rack

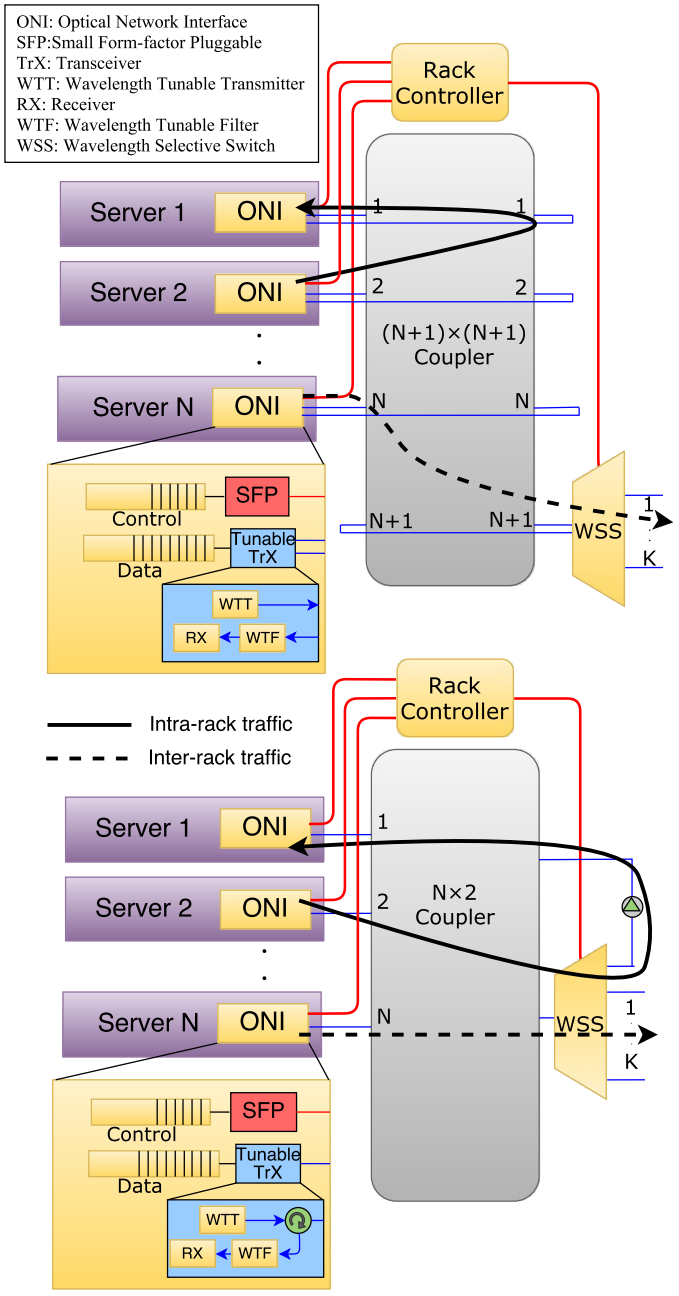


Figure 3.1: POTORI based on: (a) (N+1)×(N+1) Coupler, (b) N×2 Coupler © 2016 IEEE (**Paper II**)

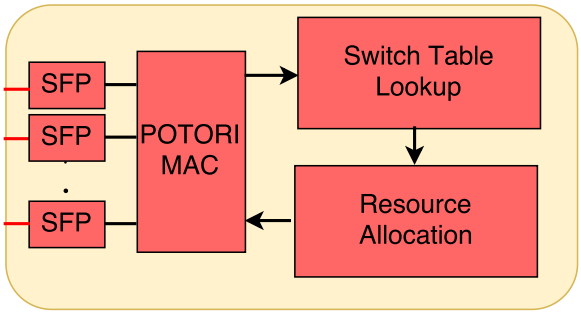


Figure 3.2: POTORI's Rack Controller © 2017 IEEE (**Paper III**)

controller needs to first collect relevant information (e.g. buffer size, destination server of data transmission, etc.) from each ONI. The POTORI's MAC protocol defines the format of control messages and message exchanging procedures for the rack controller and servers. After receiving the servers' information and doing the regular switch table lookup (i.e., map the source and destination server to the input port and output port of coupler), the rack controller runs a DBA algorithm, determining the wavelength and timeslot assigned for different data transmissions for all the servers in a rack. Finally, these decisions will be sent back to all the servers, and the servers transmit/receive data on the specified wavelengths and timeslots.

The control plane of POTORI can be easily integrated into the overall control architecture of the whole data center. The rack controller can be connected to a higher layer DC controller. Specifically, the rack controller can be equipped with a configurable switch table (e.g., an OpenFlow [19] switch table) and a configurable resource allocation module (see Fig. 3.2), so that the flow rules and employed DBA algorithm can be dynamically updated by the DC controller. In this thesis, we consider a simpler case of rack controller which only performs simple layer 2 function and runs a fixed DBA algorithm, and we leave the configurable modules for the future work.

3.2 Medium Access Control Protocol

Due the the broadcast nature of the coupler, a proper MAC protocol is required for POTORI in order to efficiently manage the communications without any collision inside the rack. Several MAC protocols have been proposed for different network scenarios, but as it is shown in the following subsection, the framework presented in this thesis is the first that can be directly applied to POTORI. In this regard, we propose a novel centralized cycle-based MAC protocol that is tailored for POTORI.

3.2.1 Related Work

Depending on whether a central controller is involved or not, the existing MAC protocols can be categorized as distributed and centralized. In the distributed MAC protocol, each node in the network makes its own decision on the resources about the data transmission based on the control information collected from other nodes. One example is the carrier sense multiple access with collision detection (CSMA/CD) [20], which is the standard MAC protocol of the old version of Ethernet (with 10 Mb/s, 100 Mb/s), but is not practical for Ethernet with a higher data rates (e.g., 10 Gb/s). Another interesting example is the high-efficient distributed access (HEAD) protocol proposed in POXN [21]. In the HEAD protocol, the control information is broadcast from one server to all the other servers in a rack. As the author stated, the collision of transmitting control information may occur during the broadcast. In the case of collision, the server needs to wait for a random back-off period, which introduces a significant control overhead and decreases the performance of the network (e.g., packet delay).

In the centralized MAC protocol, a centralized controller exchanges control information with the nodes inside a network and manage the data transmission among all nodes. One typical example is the IEEE 802.11 protocols [22] used for Wi-Fi, and another example would be the multipoint control protocol (MPCP) [23] used for the Ethernet passive optical networks (EPON). In EPON, each optical network unit (ONU) transmits both control information and data packets to the optical line terminal (OLT). However, these centralized MAC protocols are not applicable to POTORI, since they either do not support wavelength division multiplexing (WDM) scenarios, or do not support multipoint-to-multipoint communications, both of which are essential for POTORI.

3.2.2 The proposed MAC Protocol for POTORI

Fig. 3.3 shows the proposed MAC protocol. The centralized POTORI's MAC protocol is cycle-based, and it follows a Request-Grant approach. Thanks to the dedicated control channel between servers and rack controller, the control information exchanging and data transmission can be done in parallel. From the control plane's perspective, at the beginning of each cycle, each server needs to report to the rack controller in the Request message about the information of the stored data in its buffer. Based on the received Request message from all the servers in a rack as well as the interfaces towards the outside of the rack, the rack controller runs the DBA algorithm and computes the allocation of the wavelength and timeslot of the data transmission for the servers. These resource allocation decisions are then sent back to the servers with the Grant message. The Grant message contains all the necessary information for the data transmission of a server (e.g., the wavelength and timeslot used by the server to transmit/receive the data, the ending timestamp of the cycle, etc.).

On the other hand, servers transmit and receive data according to the Grant

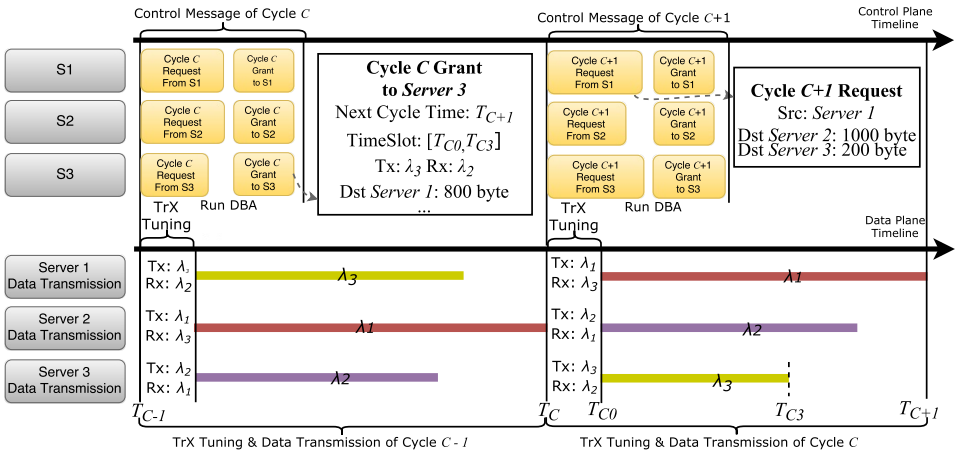


Figure 3.3: POTORI's MAC Protocol © 2017 IEEE (Paper III)

message received on the NEXT cycle (relative to the cycle mentioned for the control plane). At the beginning of the NEXT cycle, the servers first tune its transceiver to the specified wavelength. Here we consider that the tuning time of the transceiver is not negligible comparing to the transmission time of the data at high data rate. Then, each server transmits the data to the specified destination server. The cycle lasts until the ending timestamp, which is also the beginning of the following cycle. Then the whole procedure repeats. More detailed information on the MAC protocol as well as the structure of the Request and Grant message can be found in **Paper III**.

3.3 Dynamic Bandwidth Allocation Algorithm

As mentioned in the previous sections, in each cycle the rack controller needs to make decision on allocating resources in both wavelength and time domain for all servers in a rack. Thus, the DBA algorithm running at the rack controller has a great impact on the overall network performance (e.g., packet delay).

After receiving the Request messages from all servers and uplink interfaces, the rack controller can build a traffic demand (TD) matrix, where each row represents the amount of traffic (in bytes) addressed to different output ports (destination) reported by a certain input port (source). Based on the TD matrix of each cycle, a DBA algorithm computes a feasible solution of the wavelength assignment for the different traffic demands without any wavelength conflict. A wavelength conflict happens when different traffic demands are assigned with the same wavelength (i.e. data collision in coupler), or there is more than one wavelength assigned in a row

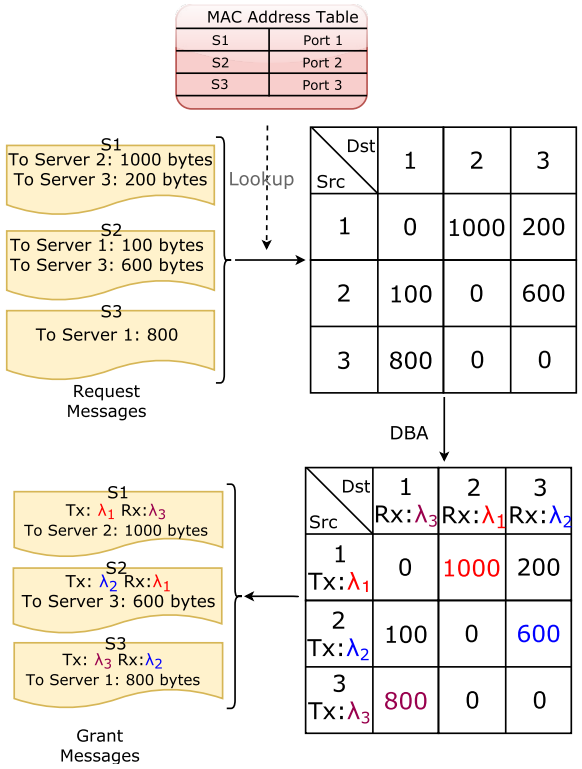


Figure 3.4: Traffic Demand Matrix

or column (i.e. wavelength clash at a transmitter or receiver). Fig. 3.4 gives an example of TD matrix and a feasible solution of DBA algorithm. The traffic demands that are assigned with wavelengths are shown in different colors. With this result gained from the DBA algorithm, the rack controller is able to form the Grant messages containing the relevant information and send them back to all the servers.

In this section we discuss different algorithms that can be applied to POTORI and propose a new heuristic algorithm, namely “Largest First”.

3.3.1 Related Work

The problem described above is similar to finding bipartite matching in a N-vertices graph [24], which has been widely studied and applied in the classical electronic packet switch scheduling problems. In the EPS, the incoming data are stored in the buffers at different input ports and then forwarded to the output ports. The traditional crossbar switch is able to forward data from multiple input ports to dif-

ferent output ports simultaneously without any collision. Different EPS scheduling algorithms have been proposed over decades to achieve high throughput and low packet delay.

An example is the algorithm based on the Birkhoff-von-Neumann (BvN) [25]. The BvN algorithm is able to decompose any TD matrix into a combination of different permutation matrix, where each permutation matrix can be used as a feasible solution for matching input and output ports. The authors in [26] have proposed an EPS based on the BvN algorithm. Similarly, the authors in [27] have applied the BvN algorithm to find scheduling solutions to configure the optical circuit switch in DCs. However, the high running-time complexity of matrix decomposition in the computation makes the BvN algorithm unsuitable for POTORI.

Another example for scheduling is the famous iSLIP algorithm [28], which has been widely used in the EPS. The iSLIP algorithm is an advanced round-robin algorithm and has a much lower running-time complexity compared to BvN. Moreover, it can be easily implemented in the hardware for the EPS. However, the iSLIP algorithm is not designed to support WDM. Therefore, we adapt the iSLIP algorithm with supporting wavelength allocation in POTORI as the benchmark DBA algorithm, and compare it to our proposed algorithm. More details on the iSLIP and its adaption to POTORI can be found in **Paper III**.

Algorithm 1 Largest First Algorithm

Input: M; **W**; **const R**
%Input: traffic demand matrix M, wavelength list W, transceiver data rate R
$tX \leftarrow [None, None...]$; $txTime \leftarrow [0, 0]$
$rX \leftarrow [None, None...]$; $rxTime \leftarrow [0, 0]$
List $\mathbf{T} \leftarrow \mathbf{M}.sort()$
repeat
    $D \leftarrow \mathbf{T}[0]$
    if $D.tX$ is **None** and $D.rX$ is **None then**
        $D.assigned \leftarrow$ **True**
        $tX[D.src] \leftarrow [\mathbf{W}[0] : [0, D.size/R]]$
        $rX[D.dst] \leftarrow [\mathbf{W}[0] : [0, D.size/R]]$
        $txTime[D.src] \leftarrow D.size/R$
        $rxTime[D.dst] \leftarrow D.size/R$
        delete $\mathbf{W}[0]$
    delete $\mathbf{T}[0]$
until T or **W** is Empty
return tX, rX

Figure 3.5: Largest First Algorithm

3.3.2 Largest First

We propose a greedy heuristic DBA algorithm, namely "Largest First" (LF). Fig. 3.5 gives the pseudo code of LF. The input of the algorithm is the TD matrix (M) of the current cycle, the number of available wavelength (W) for the assignment, and a constant value of the data rate (R) (Line 1 in Fig. 3.5). The LF algorithm first sorts the elements (traffic demands) of the TD matrix into a list T in the descending ordering (Line 5 in Fig. 3.5). Then starting from the first (i.e., largest) element in T, a wavelength is assigned to a traffic demand if and only if there is no wavelength clash at the transceiver, i.e., both the transmitter and receiver associated to this demand are not assigned with any wavelength in the current cycle (Line 8 in Fig. 3.5). The corresponding information (e.g., the assigned wavelength, transmitting/receiving timestamps) is updated in the transceiver list tX and receiver list rX (Line 10-14 in Fig. 3.5). In the case of wavelength clash, the traffic demand is skipped and left for the next cycle. The iteration stops when there are no more available wavelengths that can be assigned or the last traffic demand is served (Line 16 in Fig. 3.5). The output of the LF algorithm (i.e., trX and rX) is used by the rack controller to generate the Grant message.

3.4 Performance Evaluation

In this section, we mainly focus on the evaluation of two performance indicators of POTORI, i.e., the average packet delay and packet drop ratio, and we compare the simulation results with the conventional EPS. We also illustrate the impact of the tuning time of the transceiver (T_{Tu}) and the maximum transmission time of each cycle (T_M) on the performance of average packet delay under different load. In the simulation, we consider a rack with 64 servers and 16 uplink interfaces, and there are 80 available wavelengths. The data rate of the tunable transceiver is 10 Gb/s. The traffic model of the simulation is derived from [13] and [29], which is close to the real traffic pattern in DCs. More detailed aspects of the performance evaluation, including the parameters of the input traffic model, salability and impact of the number of wavelengths, can be found in **Paper III**.

A. POTORI vs. EPS

Fig 3.6 shows the average packet delay and packet drop ratio of POTORI with LF and iSLIP DBA algorithm as well as the EPS with iSLIP. The tuning time of the transceiver (T_{Tu}) is 50 ns [30] and the maximum transmission time of each cycle (T_M) is 1.2 μs. It can be seen in Fig 3.6 (a) that POTORI with LF has the best performance. Under the load lower than 0.5, LF can achieve an average packet delay lower than 10 μs, which is similar to the one obtained for EPS. Fig 3.6 (b) shows the packet drop ratio. It can be seen that LF performs slightly better than EPS (around 2% difference). The average packet delay of iSLIP under load lower

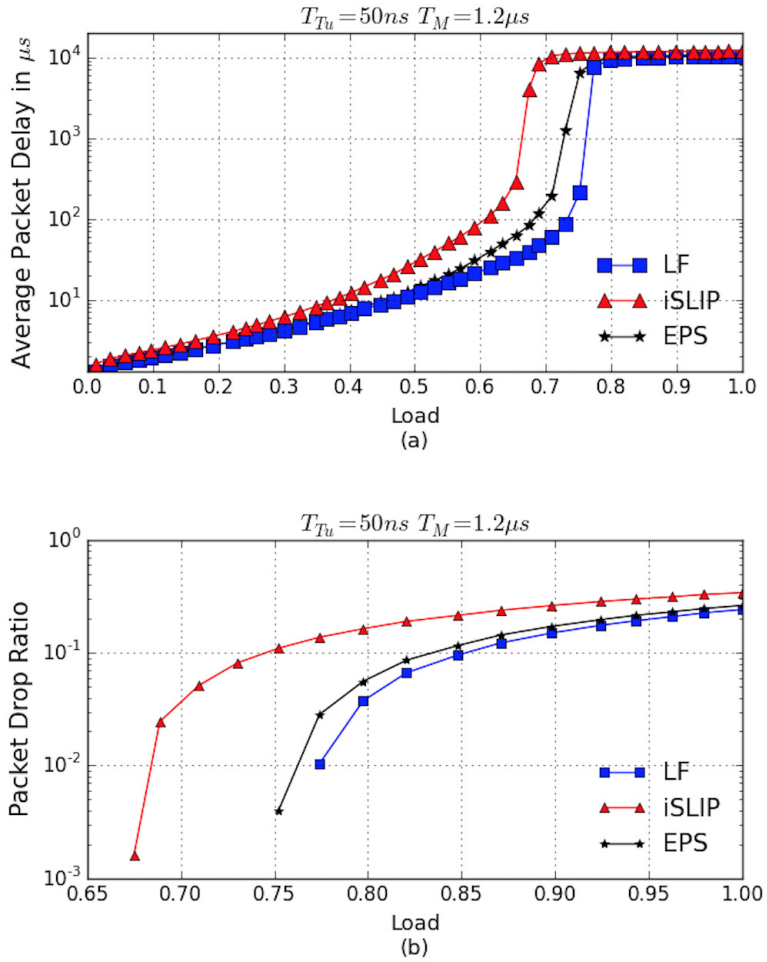


Figure 3.6: (a) Average Packet Delay (b) Packet Drop Ratio © 2017 IEEE (**Paper III**)

than 0.5 is double as high as that of LF. Moreover, POTORI with iSLIP also shows the highest packet drop ratio among all the tested schemes.

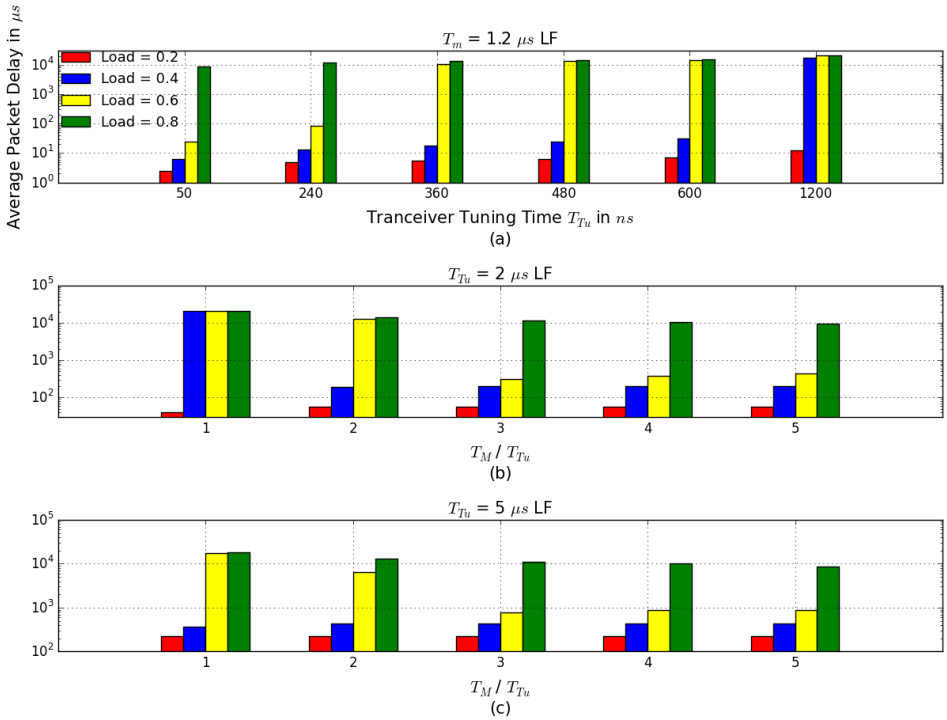


Figure 3.7: Average Packet Delay of (a) T_M = 1.2 μs with different T_{Tu}; (b) T_{Tu} = 2 μs (c) T_{Tu} = 5 μs with different ratio between T_M and T_{Tu}

B. Transceiver's Tuning Time v.s. Cycle's Maximum Transmission Time

As it is shown in the previous subsection, with an ultra-fast tuning time T_{Tu} =50 ns and a maximum transmission time of T_M = 1.2 μs, POTORI with LF is able to achieve a good performance similar to the EPS. Setting T_M = 1.2 μs for POTORI is equivalent to a packet-level switching granularity, since there is at most one packet transmitted in each cycle under the data rate of 10 Gb/s. However, if T_{Tu} increases, achieving packet-level switching granularity in POTORI while still maintaining average packet delay $<$ 100 μs is challenging, due to a larger tuning overhead. Fig 3.7 (a) shows the average packet delay with different T_{Tu} given T_M = 1.2 μs. With T_{Tu} of 50 ns and 240 ns, the average packet delay is lower than 100 μs under a load of 0.6. With a larger T_{Tu}, the performance is much worse under medium load (e.g., 0.6).

In order to achieve a better performance with larger T_{Tu}, the T_M should be increased so that the tuning overhead can be reduced. Figures 3.7 (b) and (c) show

the average packet delay of POTORI as a function of T_M/T_{Tu},where T_{Tu} is equal to 2 μs and 5 μs, respectively. If the ratio is as low as 1 or 2, the packet delay can be as high as 104 μs under load = 0.6. With a higher ratio (i.e., >=3), the performance is obviously better under load = 0.6. More detailed analyses on the T_{Tu} and T_M can be found in **Paper III**.

Chapter 4

Conclusions and Future Work

This chapter concludes the thesis and describes the extension work that is planned to be done in the future.

4.1 Conclusions

This thesis presents a passive optical interconnect designed for the access tier of DCN, referred to as POTORI. Compared to the conventional EPS, POTORI achieves lower power consumption and cost, and higher reliability while maintaining good network performance (e.g., average packet delay lower than 10 μs).

The overall POTORI architecture can be divided into data plane and control plane. The POTORI's data plane is based on passive components to interconnect the servers in a rack. The passive components in POTORI bring obvious advantages in terms of cost, energy consumption and reliability compared to the active EPS. The cost and connection availability of POTORI and EPS are evaluated. The results show that POTORI has lower cost and higher connection availability (i.e., >99.995% and beyond) at the high data rate (i.e., 10 Gb/s) and verify that POTORI is able to achieve high reliability in a cost-efficient way.

The control plane of POTORI is based on a centralized controller, which coordinates the intra-rack and inter-rack communications. A centralized cycle-based MAC protocol is proposed to manage the control message exchange and data transmission in the rack. Moreover, the rack controller runs a tailored DBA algorithm, namely Largest First (LF), which determines the resource (i.e., wavelength and time) used by servers. The simulation results show that with ultra-fast tunable transceiver, POTORI with LF is able to achieve an average packet delay lower than 10 μs, which outperforms the EPS. Moreover, the impact of transceiver's tuning time and maximum transmission time of each cycle on the average packet delay is evaluated. The results reveal that POTORI is able to achieve a packet-level switching granularity with a tuning time of the transceivers that is 30% shorter than the packet transmission time. If the tuning time is longer, increasing the maximum

transmission time of each cycle (i.e. to greater than 2 times of transceiver's tuning time) is still able to achieve an average packet delay under 100 μs.

4.2 Future Work

In the current work, the performance results are obtained from the simulation. In the future, we plan to experimentally validate the POTORI's control plane. A demo of POTORI's control plane will be developed, where the reconfigurable modules (e.g., switch tables, DBA algorithm) of the rack controller can be updated by a DC controller. This demo will be able to prove the concept of POTORI as well as investigate the possibility to integrate the POTORI's controller with the overall DCN's control plane.

On the other hand, the current POTORI control plane is designed for the access tier of DCN only. After integrating POTORI with the optical aggregation/core tier, a new control scheme coordinating the intra-DC traffic should be designed. It would also be interesting to consider inter-DC scenario, where all-optical connections between POTORIs at different DC are established. In this case, the data plane and the control plane of the optical aggregation/core tier would need to be carefully designed and evaluated.

Bibliography

[1] Cisco Global Cloud Index: Forecast and Methodology, 2015-2020, Cisco White Paper

[2] A. Andreyev, "Introducing data center fabric, the next-generation Facebook data center network."
https://code.facebook.com/posts/360346274145943, 2014.

[3] Data Center Design Considerations with 40 GbE and 100 GbE, Aug. 2013, Dell white paper.

[4] N. Binkert *et al.*, "The role of optics in future high radix switch design," in *Proc. IEEE ISCA*, 2011, pp.437-447.

[5] R.Priesetal. *et al.*,"Power consumption analysis of data center architectures," in *Green Communications and Networking*, 2012.

[6] C. Kachris *et al*, Optical Interconnects for Future Data Center Networks. 2013.

[7] M. Fiorani *et al.*, "Energy-efficient elastic optical interconnect architecture for data centers," in *IEEE Communications Letters*, vol.18, pp. 1531-1534, Sept. 2014.

[8] G. Wan *et al.*, "c-through: part-time optics in data centers," in *Proc. ACM SIGCOMM Conf.*, 2010, pp. 327-338.

[9] M.Fiorani *et al.*, "Hybrid Optical Switching for Data Center Networks," in *Hindawi Journal of Electrical and Computer Engineering*, Vol. 2014, Article ID 139213, 13 pages, 2014.

[10] F. Yan *et al.*, "Novel Flat Data Center Network Architecture Based on Optical Switches With Fast Flow Control," in *IEEE Photonics Journal*, vol. 8, number 2, April 2016.

[11] M. Yuang *et al.*, "OPMDC: Architecture Design and Implementation of a New Optical Pyramid Data Center Network," in *IEEE/OSA Journal of Lightwave Technology*, vol. 33, issue 10, pages 2019-2031, May 2015.

[12] M. Fiorani *et al.*, "Optical spatial division multiplexing for ultra-high- capacity modular data centers," in *Proc. IEEE/OSA Opt. Fiber Commun. Conf.* 2016, Paper Tu2h.2

[13] A. Roy *et al.*, "Inside the Social Network's (Datacenter) Network," in *Proc. ACM SIGCOMM Conf.*, 2015 pp. 123-237.

[14] J. Chen *et al.*, "Optical Interconnects at Top of the Rack for Energy- Efficient Datacenters," in *IEEE Communications Magazine*, vol. 53, pp. 140-148, Aug. 2015.

[15] Data center site infrastructure tier standard: topology", uptime institute, 2010.

[16] R.N. Mysore, *et al.*, "Portland: a scalable fault-tolerant layer 2 data center network fabric", in *Proc. of ACM SIGCOMM Computer Communication Review*, vol. 39, pp. 39-50, Oct. 2009.

[17] Y. Liu, *et al.*, "Quartz: a new design element for low latency DCNs", in *Proc. ACM SIGCOMM Conf.*, 2014

[18] Y. Yawei *et al.*, "LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers", in *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 2, p. 360- 409, Mar./Apr. 2013.

[19] N. McKeown *et al.*, "OpenFlow: Enabling innovation in campus net- works," in *ACM SIGCOMM Computer Communication*, Review 38, April 2008.

[20] "802.3-2012 - IEEE Standard for Ethernet".

[21] W. Ni *et al.*, "POXN: a new passive optical cross-connection network for low cost power efficient datacenters," in *IEEE/OSA Journal of Lightwave Technology*, vol. 32, pp. 1482-1500, Apr. 2014.

[22] "IEEE Standard for Information technology-Telecommunications and information exchange between systems Local and metropolitan area networks-Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications".

[23] L. Khermosh, "Managed Objects of Ethernet Passive Optical Networks (EPON)," RFC 4837, July 2007.

[24] T. Anderson, S. Owicki, J. Saxe, and C. Thacker, "High speed switch scheduling for local area networks," in *ACM Trans. Comput. Syst.*, vol. 11, no. 4, pp. 319-352, Nov. 1993.

[25] 25 G. Birkhoff. Tres Observaciones Sobre el Algebra Lineal. Univ. Nac. Tucuman Rev. Ser. A, 5:147-151, 1946.

[26] C. Chang, *et al.*, "Load balanced Birkhoff-von Neumann switches," *IEEE Workshop on High Performance Switching and Routing*, Dallas, TX, 2001, pp. 276-280.

[27] G. Poter *et al.*, "Integrating microsecond circuit switching into the data center," in *Proc. ACM SIGCOMM Conf.*, 2013 pp. 447-458.

[28] N. McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches," in *IEEE/ACM Trans. on Networking*, vol. 7, no 2, pp.188-201, 1999.

[29] S. Kandula *et al.*, "The Nature of Datacenter Traffic: Measurement and Analysis," in *Proc. ACM SIGCOMM Internet Eas. Conf.*, 2009, pp. 202-208

[30] S. Matsuo *et al.*, "Microring-resonator-based widely tunable lasers," in *IEEE J. Sel. Topics Quantum Electron*, vol. 15, no. 3, pp. 545-554, 2009.

Summary of the Original Works

Paper I. **Y. Cheng**, M. Fiorani, L. Wosinska, and J. Chen, "Reliable and Cost Efficient Passive Optical Interconnects for Data Centers," in *IEEE Communications Letters*, vol. 19, pp. 1913-1916, Nov. 2015.

In this paper, three schemes of passive optical interconnect (POI) for the access tier of the data center networks are presented. Moreover, these three schemes as well as electronic packet switch (EPS) based scheme are analyzed in terms of the cost and reliability. The results show that compared to the EPS scheme, POI schemes are able to achieve higher availability in a cost-efficient way.

Contribution of author: Cost and reliability calculations for different POI architectures, analysis and interpretation of results, preparation of the first draft and updated versions of the manuscript.

Paper II. **Y. Cheng**, M. Fiorani, L. Wosinska, and J. Chen,"Centralized Control Plane for passive Optical Top-of-Rack Interconnects in Data Centers," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016. In this work, we state our novel rain detection algorithm in the context of patent.

In this paper, the centralized control plane of POTORI is presented, including the MAC protocol and dynamic bandwidth allocation (DBA) algorithms. A rack controller is proposed to coordinate the communication in the rack by exchanging the control messages with servers and running the proposed DBA algorithm. The simulation results show that the proposed DBA algorithms outperform the benchmark algorithm in terms of the average packet delay and packet drop ratio.

Contribution of author: Proposing and implementation of the MAC protocol and DBA algorithms, development of the simulator, collection of simulation results, analysis and interpretation results, preparation of the first draft and updated versions of the manuscript, preparation of the presentation slides for the conference.

Paper III. **Y. Cheng**, M. Fiorani, R. Lin, L. Wosinska, and J. Chen,"POTORI: A Passive Optical Top-of-Rack Interconnect Architecture for Data

Centers," in *IEEE/OSA Journal of Optical Communications and Networking (JOCN)*, to appear, 2017.

This paper extends **Paper II.** II by introducing the following new contributions: (1) illustration of how POTORI as the access tier can be integrated into the overall data center network; (2) Extensive performance comparison among POTORI with different DBA algorithms as well as electronic packet switch; (3) Evaluating the impact of different network configurations on the performance of POTORI.

Contribution of author: Run simulation, collection of simulation results, analysis and interpretation results, preparation of the first draft and updated versions of the manuscript,

Paper I

Reliable and Cost Efficient Passive Optical Interconnects for Data Centers

Yuxin Cheng, Matteo Fiorani, Lena Wosinska, Jiajia Chen

IEEE Communications Letters, vol. 19, pp.1913-1916, Nov. 2015

Reliable and Cost Efficient Passive Optical Interconnects for Data Centers

Yuxin Cheng, Matteo Fiorani, Lena Wosinska, and Jiajia Chen

***Abstract*—To address the sustainability, scalability and reliability problems that data centers are currently facing, we propose three passive optical interconnect (POI) architectures on top of the rack. The evaluation results show that all three architectures offer high reliability performance (connection availability for intra-rack interconnections higher than 99.999%) in a cost-efficient way.**

***Index Terms*—Optical communications, data center interconnects, reliability analysis, cost models.**

I. Introduction

The growing popularity of cloud applications is drastically increasing the traffic volumes that data centers have to handle [1]. Consequently, the transmission capacity inside the data centers is rapidly growing. The majority of the servers today are equipped with 1Gbps or 10Gbps interfaces for communications, while in the future higher transmission rates are expected to be used (e.g., 40Gbps or 100Gbps per server) [2]. These trends lead to scalability problem for the network providing connectivity among severs inside the data center and toward the Internet.

Current data center interconnection networks include several tiers, such as edge, aggregation and core, and are based on electronic commodity switches. Scaling these networks to support very high transmission capacity may lead to dramatic increase in the total equipment cost and power consumption [3]. In this regard, optical communication has been considered as a promising technology for data center interconnects due to the ultra-high capacity that can be offered in cost- and energy- efficient way.

Several optical switching architectures have been recently proposed for data center networks [4][5][6][7]. They are based on either optical switches [4][5] or passive optical components [6][7] at the aggregation and core tiers. However, in these architectures the edge tier is still based on electronic top-of-rack (ToR) switches, which limits the overall cost and energy savings [3]. Paper [8] has explored different possibilities for optical interconnection solutions at ToR and identified that using POI for inter-server communication (such as the architectures proposed in [6][7]) can potentially offer significant energy saving while at the relatively low cost.

Meanwhile, the higher transmission rate, the larger the volume of traffic and number of cloud services can be affected in case of a failure in the network. The required availability of fault-tolerant data center infrastructure (including electrical power supply, storage and distribution facilities) should be higher than 99.995% [9]. Thus, the availability for any connection established within the data center needs to be even higher, since the communication system is only a part of the site infrastructure. Several topologies, e.g., fat-tree [10] and Quartz [11], have been proposed in order to improve the resiliency of large-scale data center networks. These topologies introduce redundancy in the aggregation and core tiers to increase reliability in the central part of the data center network. However, the edge tier is usually unprotected due to the high cost of introducing redundant ToR switches as well as due to the belief that edge tier can be self-healing (i.e., in case the connection to a certain server would be down, the task can be re-assigned and carried out by another server). Unfortunately, it may not be true in the scenario where the servers within the racks are highly loaded making it difficult to find resources to be allocated for a possible backup.

Therefore, the expected growth of traffic volume inside the data centers brings the need for highly reliable, yet cost and energy efficient, interconnection at the edge tier. In [3][8] several passive optical interconnects for the edge tier of data center networks have been presented showing that by replacing the electronic ToR switches with passive optical components it is possible to significantly reduce the overall energy consumption while offering high capacity interconnection.

On the other hand, such optical interconnects may lead to higher capacities needed in the aggregation and core tiers of the data center network because of the lack of statistical multiplexing in the optical domain. However, this problem can be mitigated by employing burst mode transceivers and a control protocol that is able to perform an efficient dynamic bandwidth allocation strategy [6][7].

Despite of the growing importance of fault tolerance of interconnects at ToR, this aspect has not been studied yet. In this letter, we focus on highly reliable and cost efficient passive optical interconnects and propose three architectures for interconnections at ToR. They can be integrated with any topology supporting large-scale data center networks, e.g., fat-tree and Quartz. We also evaluate the proposed architectures in terms of connection availability and cost. Our results verify that ultra-high connection availability, i.e., close to 5 nines (99.999%), can be achieved with both 1Gbps and 10Gbps server interfaces. In addition, the cost of the proposed passive optical interconnects scale more efficiently with the server capacity compared to electronic commodity switches.

II. Reliable Passive Optical Interconnects

In this section, three passive optical interconnects (POIs) for the edge tier are presented. The first one is based on an arrayed waveguide grating (AWG), while the other two are based on a coupler which broadcasts the traffic sent by one input port to all the output ports. The proposed POIs are shown in Fig. 1. In all three schemes servers are equipped with optical network interfaces (ONIs) sending and receiving optical signals. The communication can be either intra-rack (shown as red solid lines) or to outside the rack (inter-rack or to/from the Internet, shown as blue dashed lines).

It should be noted that the dynamicity and programmability for POIs are provided by the ONIs of the servers. Therefore, all the presented POIs have the ONIs equipped

Manuscript received June 18, 2015; revised August 10, 2015; accepted September 8, 2015. This work was supported by the Swedish Foundation for Strategic Research (SSF), Vetenskapsrådet, and Göran Gustafssons Stiftelse. The associate editor coordinating the review of this paper and approving it for publication was W. Fawaz. *(Corresponding author: Jiajia Chen.)*

All the authors are with KTH Royal Institute of Technology, Communication Systems Department (CoS), Electrum 229, SE-164 40 Stockholm, Sweden (e-mail: yuxinc@kth.se; fiorani@kth.se; wosinska@kth.se; jiajiac@kth.se).

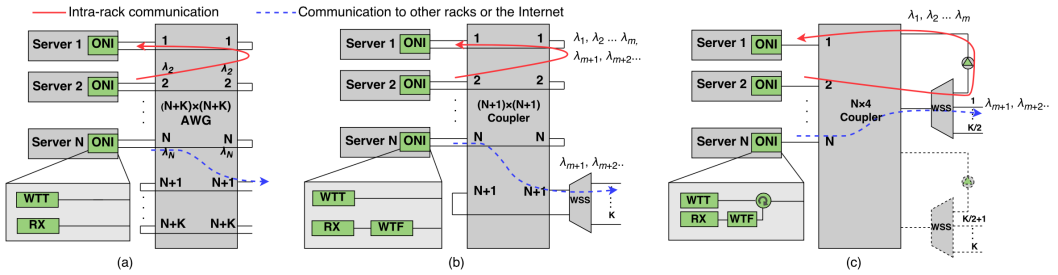


Fig. 1. (a) Scheme I: *(N+K)×(N+K)* AWG based POI, (b) Scheme II: *(N+1)×(N+1)* coupler based POI and (c) Scheme III: $N\times4$ coupler based POI (WTT: Wavelength Tunable Transmitter, AWG: Arrayed Waveguide Grating, ONI: Optical Network Interface, RX: Receiver, WTF: Wavelength Tunable Filter,WSS: Wavelength Selective Switch).

with wavelength tunability. It has been demonstrated in [6] and [7] that good network performance can be achieved with wavelength tuning speed in the magnitude of microseconds.

A. Scheme I: AWG based POI

Fig. 1(a) shows the structure of the first type of POI, which is based on *(N+K)×(N+K)* AWG interconnecting the servers within a rack as well as providing connection to the switches at aggregation/core tier. Here, N is the number of servers and K is the number of the links between ToR and the aggregation/core tier. This scheme is inspired by the POI proposed in [4]. Each ONI has two fibers connected to the AWG input and output ports. For this type of POI, in total N+K wavelengths are required. Thanks to the cyclic property of the AWG, a proper wavelength plan can be made (see Fig. 2) to set up a connection for any intra-rack or inter-rack communication without any conflicts in spectrum. Note that the grey fields in Fig. 2 represent no connection (i.e., there is no traffic passing through the POI destined to the same server or between two interfaces toward the aggregation/core tiers).

From \ To		Within the rack				Outside of the rack		
		Server 1	Server 2	…	Server N	Interface 1	…	Interface K
Within the rack	Server 1		λ_2	…	λ_N	λ_{N+1}	…	λ_{N+K}
	Server 2	λ_2		…	λ_{N+1}	λ_{N+2}	…	λ_1
	…	…	…		…	…	…	…
	Server N	λ_N	λ_{N+1}	…		λ_{N-K}	…	λ_{N-1}
Outside of the rack	Interface 1	λ_{N+1}	λ_{N+2}	…	λ_{N-K}			
	…	…	…	…	…			
	Interface K	λ_{N+K}	λ_1	…	λ_{N-1}			

Fig. 2. Wavelength plan for AWG based POI.

B. Scheme II: (N+1)×(N+1) coupler based POI

Fig. 1(b) shows the structure of the second proposed POI architecture. In Scheme II an *(N+1)×(N+1)* coupler is employed to interconnect N servers within the rack. Two ports of the coupler are connected to a wavelength selective switch (WSS) for inter-rack communications. The broadcast nature of the coupler provides higher flexibility in wavelength allocation than the AWG based scheme. In this scheme, the wavelengths can be dynamically assigned for the intra- and inter-rack communications, leading to high resource utilization. The wavelength tunable transmitters (WTTs) on the ONIs are able to use any available wavelength ($\lambda_1, \ldots \lambda_m$) for intra-rack communications. The data is broadcast to all the output ports of the coupler. The ONI consists of wavelength tunable filter (WTF) and receiver (Rx) for receiving the signal. Due to the broadcast-and-select manner, the WTF is needed to select the wavelength assigned to a specific communication, while the signals on other wavelengths are dropped. For the traffic to outside of the rack, the WSS switches the corresponding wavelengths ($\lambda_{m+1}, \lambda_{m+2}\ldots$) and forwards the traffic to the aggregation and core tires. In this architecture, we assume the use of a $2\times K$ WSS, as the one demonstrated in [12]. Multiple interfaces ($K\geq2$) can be reserved to connect to the aggregation/core tier and support any topology (e.g., fat-tree and Quartz) for high scalability and resiliency.

C. Scheme III: N×4 coupler based POI

Fig. 1(c) shows the third proposed POI architecture, which enhances the reliability performance of the coupler based POI proposed in [3]. In Scheme III, ONIs at the servers are connected to N input ports of an $N\times4$ coupler. By passing through a WSS, the wavelengths assigned for the intra-rack communications (i.e., $\lambda_1, \ldots \lambda_m$) are sent back to the coupler and then broadcasted to all the connected ONIs in the same rack. Like in Scheme II, the WTF is needed at the ONI to select the signal carried by the assigned wavelength. The wavelengths (λ_{m+1}, λ_{m+2} …) are allocated for traffic sent to/received from the outside of the rack. Similar as the other schemes, this approach can also reserve several interfaces to connect to the aggregation/core tier for resiliency and scalability enhancement. From resiliency perspective, WSS is critical in this POI structure, as the traffic for both intra-rack communications and to the outside of the rack needs to pass this component. Considering the fact that WSS is an active component having obvious lower availability than the passive devices, backup is proposed by introducing an additional WSS.

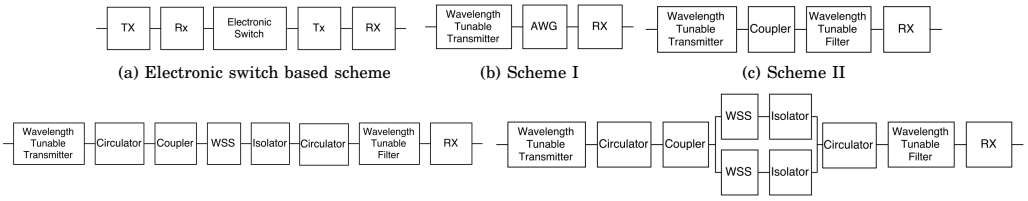
III. Reliability and Cost Models

In this section we present an analytical model for reliability and cost evaluations of the three proposed POI architectures and the electronic ToR switch scheme.

A. Reliability Analysis

In this letter, we focus on ToR interconnect design and hence we perform the reliability performance analysis for the intra-rack communication. However, the same methodology can be applied to the core/aggregation tier.

Figure 3 presents the reliability block diagrams (RBDs) for the intra-rack connections in the electronic switch based scheme and the three proposed POIs. RBD represents availability model of a system/connection, where series configuration corresponds to the case where all the connected blocks need to be available, while the parallel configuration means that at least one of the branches needs to be available. We can observe that in the three proposed POIs, the connections include several active and passive optical components. As a consequence, the connection availability (A) (i.e., the probability that the connection is operating) can be calculated by multiplying the availability of the cascaded components. We can then obtain the following formulas for availability of intra-rack connection in electronic switch based scheme (A_E), Scheme I ($A_{POI(I)}$), Scheme II ($A_{POI(II)}$), and Scheme III without protection ($A_{U-POI(III)}$):



(a) Electronic switch based scheme (b) Scheme I (c) Scheme II

(d) Scheme III without protection (e) Scheme III with protection

Fig. 3. Reliability block diagrams.

TABLE I
MTBF AND COST OF THE NETWORK ELEMENTS [8][13][14].

Components	MTBF[1]	Cost
1GE Electronic Switch	150 000 h	0.67 CU[2](port)
10GE Electronic Switch	150 000 h	3 CU
1Gbps Grey Transceiver	3 000 000 h	0.1 CU
10Gbps Grey Transceiver	600 000 h	0.5CU
1Gbps Tunable Transceiver	1 000 000 h	0.67CU
10Gbps Tunable Transceiver	500 000 h	1.3CU
WSS	300 000 h	8.3CU
AWG	4 000 000 h	0.1 CU(port)
Coupler	6 000 000 h	0.02 CU(port)
Isolator	12 000 000 h	0.3CU
Circulator	12 000 000 h	0.7 CU
Wavelength Tunable Filter	4 000 000 h	0.3 CU

1. Mean Time Between Failures
2. CU is the cost unit. 1 CU = 150 USD.

$$A_E = A_{GTRX}^2 \times A_{ES}, \tag{1}$$
$$A_{POI(I)} = A_{TRX} \times A_{AWG}, \tag{2}$$
$$A_{POI(II)} = A_{TRX} \times A_{CP} \times A_{WTF}, \tag{3}$$
$$A_{U-POI(III)} = A_{TRX} \times A_{CL} \times A_{CP} \times A_{WSS} \times A_{IS} \times A_{WTF}. \tag{4}$$

Here, A_{TRX} represents the availability of the tunable transceiver (note that transmitter and receiver are embedded on the same board). Meanwhile, we denote A_{GTRX}, A_{ES}, A_{AWG}, A_{CL}, A_{CP}, A_{WSS}, A_{IS}, and A_{WTF} as the availability of the grey transceiver, electronic switch, AWG, circulator, coupler, WSS, isolator and WTF, respectively. The availability of each component can be obtained as the ratio between the mean lifetime and the mean time between failures (MTBF) [15]. In the protected Scheme III, the reliability is improved by the redundancy of WSS and isolator comparing to the unprotected one. The availability can be obtained according to the following formula:

$$A_{P-POI(III)} = (1-(1-A_{WSS} \times A_{IS})^2) \times A_{TRX} \times A_{CL} \times A_{CP} \times A_{WTF}. \tag{5}$$

B. Cost Analysis

To calculate the equipment cost for the three proposed POIs, we employ a similar approach as in [3]. We define the total cost of a POI (C_{POI}) as the sum of all the network components inside the rack. As a consequence, the total cost for electronic switch based scheme (C_E), Scheme I ($C_{POI(I)}$), Scheme II ($C_{POI(II)}$), and the unprotected Scheme III ($C_{U-POI(III)}$), can be calculated according to the following formulas:

$$C_E = 2 \times N \times C_{GTRX} + C_{ES}, \tag{6}$$
$$C_{POI(I)} = N \times C_{TRX} + C_{AWG}, \tag{7}$$
$$C_{POI(II)} = N \times (C_{TRX} + C_{WTF}) + C_{CP} + C_{WSS}, \tag{8}$$
$$C_{U-POI(III)} = N \times (C_{TRX} + C_{CL} + C_{WTF}) + C_{CP} + C_{WSS} + C_{IS}. \tag{9}$$

Here, N is the number of servers in the rack and C_{TRX} is the cost of a tunable optical transceiver. Moreover, C_{GTRX}, C_{ES}, C_{WTF}, C_{AWG}, C_{CP}, C_{WSS}, C_{CL}, and C_{IS} are the cost of grey transceiver, electronic switch, WTF, AWG, coupler, WSS, circulator and isolator, respectively. In the protected Scheme III, additional WSS and isolator are used inside the rack to improve resiliency. Accordingly, the total cost of the protected Scheme III ($C_{P-POI(III)}$) can be obtained through the following formula:

$$C_{P-POI(III)} = N \times (C_{TRX} + C_{CL} + C_{WTF}) + C_{CP} + 2 \times (C_{WSS} + C_{IS}). \tag{10}$$

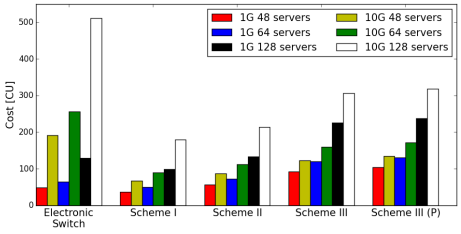


Fig. 4. Total cost for electronic ToR switch and the three proposed POI architectures

IV. NUMERICAL RESULTS

In this section, we evaluate and compare the cost and reliability of the proposed POIs. We consider a conventional electronic ToR switch as benchmark. The overall results cover the cases of 1Gbps and 10Gbps transmission capacity per server.

Table 1 shows the MTBF and the cost values for the involved components [8][13][14], which are used to evaluate the reliability and the total cost of the proposed POI architectures. Figure 4 shows the total cost of the proposed POIs and the conventional electronic ToR switch given the different number of total servers in a rack (48, 64 and 128) and 2 interfaces towards the aggregation/core tier. As reflected in the cost formulas, all the considered schemes show a linear increase in the total cost as a function of the number of servers. Considering the case with 1Gbps per server, Scheme I and Scheme II show a similar total cost as the electronic ToR switch. On the other hand, the cost of Scheme III without protection is almost doubled due to the use of an additional circulator in the ONI. However, the circulator makes cabling easier since one server only has one fiber port for interconnection. The protected Scheme III shows a small increase in the total cost, since additional WSS and the isolator are needed for backup. For the case with 10Gbps interface per server, the three POIs show great advantage in terms of cost comparing to electronic switches. The price of commercial electronic ToR switches operating at 10Gbps is much more expensive than the price of electronic ToR switches operating at 1Gbps. On the other hand, the increase of cost of three POIs from 1Gbps to 10Gbps is much lower than electronic switches. This is mainly due to the reason that for POIs the major cost increase is at transceivers side for higher data rate while the passive optical switch components remain the same. However, for solution based on commodity switches, cost increase occurs at both transceivers at the servers and the switches for interconnection.

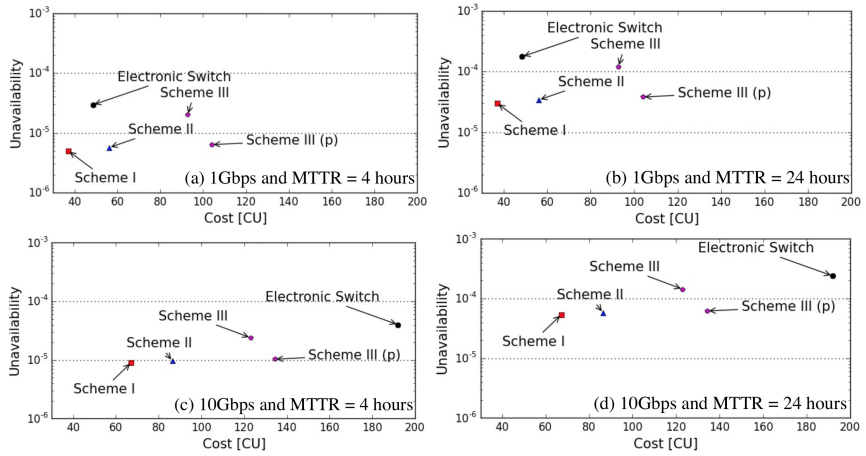


Fig. 5. Unavailability vs. total cost of three POIs for different MTTR values and server transmission capacities.

It is worth noting that, due to the lack of statistical multiplexing, the POIs may require more capacity, and thus higher costs, in the aggregation and core tiers. The cost results in [3] have shown that impact of the absence of statistical multiplexing in the edge tier on the overall data center network cost highly depends on the switching techniques adopted in core and aggregation tier. However, despite of the lack of additional statistical multiplexing in the edge tier, the cost reduction for the overall data center can still be achieved in many cases compared to the solution based on the conventional electronic switches [3].

Unavailability of a component or system is defined as the probability that it is failed and can be expressed as 1-A, where A denotes availability. Our calculation of the unavailability values of the three POIs is based on the MTBF of components presented in Table 1. Regarding mean time to repair (MTTR), it is dependent on the maintenance policy adopted by the data center operator. We assume two values of MTTR (4h and 24h) reflecting different fault management policies in respect to the length of reparation time. Figure 5 shows the unavailability of intra-rack connection versus total cost of 48 servers in a rack for the three considered optical intra-rack interconnects at 1Gbps and 10Gbps. It can be seen that Scheme I and Scheme II performs best, i.e., showing the lowest connection unavailability. In the case a fast reparation time (e.g., MTTR=4 hours), the intra-rack connection availability for Scheme I and Scheme II can reach 5 nines (99.999%), meeting the reliability requirement for fault-tolerant site infrastructure, i.e., availability of 99.995% [9]. On the other hand, unprotected Scheme III supports similar level of connection availability as its electronic counterpart, while with the proposed redundancy of the WSS, its intra-rack connection availability can reach 5 nines. Even for a relatively long reparation time (e.g., MTTR=24 hours), the availabilities up to 4 nines (99.99%) can be obtained for Scheme I, Scheme II and the protected Scheme III, which is much better than availability of ToR based on electronic switches. Advantages of POIs at 10Gbps on reliability performance compared to the electronic interconnect is more obvious than at 1Gbps. It is mainly because the reliability performance of the passive devices is not dependent on the data rate.

V. Conclusions

In this letter we proposed reliable optical interconnect architectures for the edge tier of data center interconnection networks. We have evaluated their cost and reliability performance and compared with the traditional commodity ToR switch. The results show that the proposed POIs outperform the electronic ToR switches in terms of both connection availability and cost. Compared to electronic commodity switches, the benefits of optical POIs are more obvious at the higher data rate. Therefore, our proposed POIs make it possible to reach the required connection availability of 99,995% and beyond at high data rates in the intra-data center networks.

References

[1] Cisco Global Cloud Index: Forecast and Methodology, 2013-2018, Sept. 2014, Cisco white paper.
[2] Data Center Design Considerations with 40 GbE and 100 GbE, Aug. 2013, Dell white paper.
[3] M. Fiorani, et al, "Energy-efficient elastic optical interconnect architecture for data centers", IEEE Communications Letters, vol.18, pp. 1531-1534, Sept. 2014.
[4] Y. Yawei et al., "LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers", IEEE J. Sel. Topics Quantum Electron., vol. 19, no. 2, p. 360-409, Mar./Apr. 2013.
[5] K. Chen et al., "OSA: An optical switching architecture for data center networks with unprecedented flexibility," IEEE/ACM Trans. Netw., vol. 22, no. 2, pp. 498511, Apr. 2014.
[6] W. Ni, et al., "POXN: a new passive optical cross-connection network for low cost power efficient datacenters", IEEE/OSA JLT, vol. 32, pp. 1482-1500, Apr. 2014.
[7] R.M. Indre, et al., "POPI: A passive optical Pod interconnect for high performance data centers", Proc. of ONDM, pp.84-89, May 2014.
[8] J. Chen, et al., "Optical interconnects at top of the rack for energy-efficient datacenters", IEEE Communications Magazine, to appear, 2015.
[9] "Data center site infrastructure tier standard: topology", uptime institute, 2010.
[10] R.N. Mysore, et al., "Portland: a scalable fault-tolerant layer 2 data center network fabric", Proc. of ACM SIGCOMM Computer Communication Review, vol. 39, pp. 39-50, Oct. 2009.
[11] Y. Liu, et al., "Quartz: a new design element for low latency DCNs", ACM SIGCOMM, 2014
[12] K. Fontaine, et al., "N×M Wavelength Selective Crossconnect with Flexible Passbands", Proc. of OFC, 2012
[13] EU project: Discus (The DIStributed Core for unlimited bandwidth supply for all Users and Services).
[14] K. Grobe, et al., "Cost and Energy Consumption Analysis of Advanced WDM-PONs", IEEE Communications Magazine, vol. 49, pp. s25-s32, February 2011.
[15] J. Chen, et al., "Analysis of Protection Schemes in PON Compatible with Smooth Migration from TDM-PON to Hybrid WDM/TDM PON", OSA Journal of Optical Networking, vol. 6, pp. 514-526, May, 2007

Paper II

POTORI: A Passive Optical Top-of-Rack Interconnect Architecture for Data Centers

Yuxin Cheng, Matteo Fiorani, Rui Lin, Lena Wosinska, Jiajia Chen

IEEE/OSA Journal of Optical Communications and Networking (JOCN), vol. 9, issue 5, pp. 401-411, May 2017

Centralized Control Plane for Passive Optical Top-of-Rack Interconnects in Data Centers

Yuxin Cheng, Matteo Fiorani, Lena Wosinska, and Jiajia Chen
Communication Systems Department, KTH Royal Institute of Technology, Stockholm, Sweden
{yuxinc, matteof, wosinska, jiajiac}@kth.se

***Abstract*—To efficiently handle the fast growing traffic inside data centers, several optical interconnect architectures have been recently proposed. However, most of them are targeting the aggregation and core tiers of the data center network, while relying on conventional electronic top-of-rack (ToR) switches to connect the servers inside the rack. The electronic ToR switches pose serious limitations on the data center network in terms of high cost and power consumption. To address this problem, we recently proposed a passive optical top-of-rack interconnect architecture, where we focused on the data plane design utilizing simple passive optical components to interconnect the servers within the rack. However, an appropriate control plane tailored for this architecture is needed to be able to analyze the network performance, e.g., packet delay, drop rate, etc., and also obtain a holistic network design for our passive optical top-of-rack interconnect, which we refer to as POTORI. To fill in this gap, this paper proposes the POTORI control plane design which relies on a centralized rack controller to manage the communications inside the rack. To achieve high network performance in POTORI, we also propose a centralized medium access control (MAC) protocol and two dynamic bandwidth allocation (DBA) algorithms, namely *Largest First* (*LF*)and *Largest First with Void Filling* (*LFVF*). Simulation results show that POTORI achieves packet delays in the order of microseconds and negligible packet loss probability under realistic data center traffic scenarios.**

***Index Terms*—Optical interconnect architectures, data center networks, medium access control (MAC), dynamic bandwidth allocation (DBA).**

I. Introduction

The network traffic inside the data centers has been increasing rapidly during the last decade, due to the growing popularity of applications such as cloud computing, social networking and video streaming [1]. In order to deal with this tremendous traffic increase, data center operators are upgrading the transmission rate and switching capacity of their network equipment [2]-[3]. However, current data center networks are based on electronic packet switches, whose cost and power consumption scale almost linearly with the transmission rate and switching capacity, bringing serious profitability and sustainability problems [4]-[5].

Optical interconnect architectures are considered as a promising solution to address the limitations of electronic packet switches in data centers thanks to their ability to support high transmission rates and switching capacities in a cost- and energy-efficient way. Several optical interconnect architectures for core and aggregation tiers have already been proposed in the literature, e.g., in [6]-[7]. They rely on the use of conventional electronic top-of-rack (ToR) switches in the access tier, i.e., the network segment in charge of interconnecting the servers in the same rack and the interfaces toward the other network tiers. The electronic ToR switches are usually responsible for the majority of the cost and power consumption in the data center network. This is due to the fact that the access tier carries a large amount of the overall data center traffic because most of the traffic stays locally in the same rack [9]. As a consequence, efficient optical interconnect architectures for the access tier in data centers are needed.

In [11] and [12] we proposed new passive optical top-of-rack interconnect data plane (POTORI data plane) architecture for the access tier in data centers. The POTORI data plane makes use of a passive optical coupler to interconnect the servers in the rack and employs the broadcast-and-select manner. The passive optical coupler offers relatively low cost and power consumption as well as high reliability. However, there is no control mechanism available for efficient management of the communications among the servers in the same rack as well as the interfaces to outside of the rack. Therefore, it is not possible to analyze the network performance, such as packet delay, packet drop rate, of the proposed POTORI data plane design.

In this paper we propose the POTORI control plane design enabling efficient management of the communications inside the rack. Some related works (e.g., [8]) introduced the control plane approaches based on distributed control mechanisms for passive optical interconnects. However, such solutions can bring a large control overhead. They may still be appropriate for the aggregation/core tiers of data center networks, handling the aggregated traffic from the ToR. However, they can result in a poor network performance of intra-rack interconnects, which is usually characterized by short-lived traffic flows. Therefore, in this paper we propose a centralized control plane for POTORI that has a separated rack controller responsible for the efficient management of the communications inside the rack and toward the other network tiers. We also introduce a new centralized medium access control (MAC) protocol and two dynamic bandwidth allocation (DBA) algorithms, namely *Largest First (LF)* and *Largest First with Void Filling (LFVF)*, to obtain high bandwidth utilization in POTORI. The simulation results verify that POTORI can achieve end-to-end packet delays in the order of microseconds and negligible packet loss probability under realistic data center traffic scenarios. The numerical results also show that the proposed DBA algorithms for POTORI outperform another popular switch scheduling algorithm from the literature [13].

The rest of the paper is organized as follows. In Section II, the POTORI data and control plane architectures are presented. In Section III the proposed centralized MAC is elaborated and in Section IV two novel DBA algorithms for POTORI are introduced. The simulation results are presented in Section V showing evaluation of the proposed DBA algorithms in terms of packet delay and packet drop ratio. Finally, the conclusions

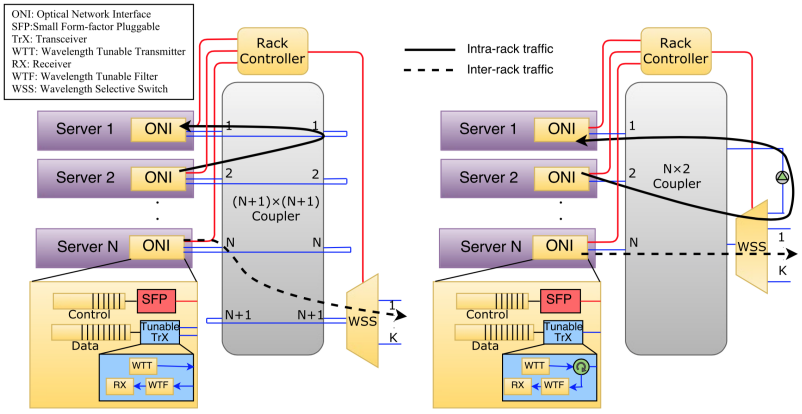


Fig. 1: POTORI based on: (a) *(N+1)×(N+1)* Coupler, (b) *N*×2 Coupler

are given in Section VI.

II. Passive Optical Top-of-Rack Interconnects (POTORI)

In POTORI both control plane and data plane architectures are considered. The control plane contains a rack controller while the data plane can have two options (see Fig. 1). In the POTORI architecture, each server is equipped with an optical network interface (ONI) composed of two optical transceivers. One is a tunable transceiver connected to the POTORI data plane. The other is a grey small form factor pluggable (SFP) transceiver connected to the POTORI control plane. In the following we describe the POTORI data and control plane architectures in detail. We define *N* as the number of servers in the rack.

Data Plane

The POTORI data plane is based on a passive optical coupler that interconnects the servers within the rack in a broadcast-and-select manner. The POTORI data plane was previously introduced in [11]-[12].

The first option, shown in Fig. 1(a), is based on an *(N+1)×(N+1)* coupler. Each ONI is connected to one input and one output port of the coupler using two separate fibers. The tunable transceiver in the ONI is composed of a wavelength tunable transmitter (WTT) connected to the input port of the coupler for sending data and a wavelength tunable filter (WTF) as well as a receiver connected to the output port of the coupler for receiving data. The coupler is equipped with an additional input and output port connected to a wavelength selective switch (WSS), which forwards the traffic to (from) the aggregation and core tiers.

The second option, shown in Fig. 1(b), is based on an *N*×*2* coupler. In this case each ONI is connected only to one input port of the coupler using a single bidirectional fiber. The tunable transceiver in the ONI is equipped with a circulator that forwards the traffic generated by the WTT toward the coupler and directs the traffic from the coupler toward the WTF and the receiver. All the traffic passes through the WSS that is responsible for separating the local flows that are looped back to the coupler and broadcast to the servers in the rack, from the traffic toward the outside of the rack. An isolator is used to guarantee the correct direction of the communications between the WSS and the coupler.

Although both architectures utilize the broadcast nature of coupler, the difference between these two is that in the first architecture, both intra- and inter-rack data transmitted by one server are broadcasted and received by all the servers within the same rack as well as the WSS connected to the coupler's output ports, while in the second architecture, only intra-rack traffic is filtered out by the WSS, and then broadcast to all the servers within the same rack.

Control Plane

Due to the broadcast nature of the coupler, the destination ONI needs to be tuned on the correct wavelength in order to receive the traffic from the desired source ONI. In addition, in order to avoid conflicts occurring in the coupler, concurrent communications inside the rack need to be assigned to different wavelengths. Consequently, a control plane that is able to schedule resources in both spectrum and time domains is required in POTORI for managing the communications inside the rack and toward the aggregation and core tiers.

We propose a centralized control mechanism for POTORI, referred to as a rack controller (see Fig. 1). It is equipped with *N* grey SFP transceivers to transmit (receive) the control information to (from) the servers. The rack controller is based on a centralized MAC protocol and a DBA algorithm to dynamically tune the WTT and WTF in each ONI to the proper wavelengths for transmitting and receiving the traffic while avoiding wavelength clash in the coupler. Specifically, the MAC protocol defines the controlling procedure and formats of the control message. The DBA algorithm is employed to compute the resources, including both the wavelengths and time duration, that are assigned to the servers for intra-rack and inter-rack communications. In the following sections, we concentrate on control plane and describe the proposed MAC protocol along with two tailored DBA algorithms for POTORI.

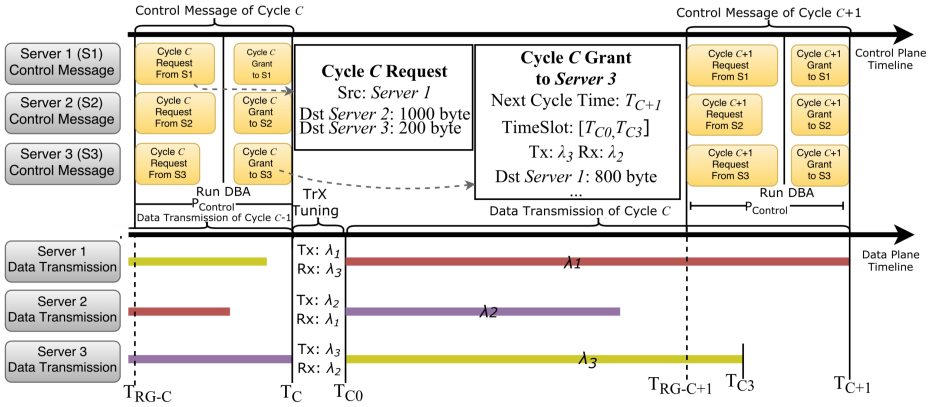


Fig. 2: The MAC Protocol and the Request-Grant Messages of POTORI

III. MAC Protocol

A proper MAC protocol is essential to efficiently manage the communications among the servers in POTORI. The existing MAC protocols that might be applied in POTORI can be classified as distributed and centralized. Distributed MAC protocols rely on the direct exchange of control information among the servers, i.e., without the mediation with a centralized controller. Typical examples are the carrier sense multiple access with collision detection (CSMA/CD) [15], which has been widely employed in the Ethernet, and the HEAD protocol which was proposed for the passive optical cross-connection network (POXN) architecture in [8]. However, the distributed MAC protocols bring a control overhead because the control messages need to be flooded to all the servers in the rack. This control overhead might be significant for the intra-rack communications that are characterized by a large amount of short lived traffic flows. As a consequence, we believe that they are not good candidates for POTORI.

The centralized MAC protocols rely on a centralized controller to manage the exchange of control information. Typical examples are the multipoint control protocol (MPCP) [17] which is used in Ethernet passive optical networks (EPONs) and the IEEE 802.11 standards [18] that are used in Wi-Fi networks. The MPCP protocol cannot be applied in POTORI because it supports only multipoint-to-point communications, while the POTORI requires multipoint-to-multipoint communications. Also the IEEE 802.11 standards cannot be employed in POTORI because they do not support wavelength division multiplexing (WDM) scenarios. With this in mind, we propose a new centralized MAC protocol for the POTORI control plane.

The proposed MAC protocol is depicted in Fig. 2. It follows a Request-Grant approach. The data transmission of servers is divided into different time cycles. Servers send a Request message to the rack controller containing the current buffered packets for different destination servers. After receiving the Request messages from all the servers, the rack controller generates the traffic matrix and runs the DBA algorithm to calculate the allocated wavelengths and transmission time for all the servers in the next time cycle. Afterwards, the controller informs the servers about the decisions using Grant messages. After receiving the Grant message, each server tunes the WTT and WTF to the assigned wavelengths for transmitting and receiving data during the specified time cycle. For example, in Fig. 2, Server 3 is informed by the Grant message to tune the transmitter to λ_3 and receiver to λ_2 in cycle C. The traffic from Server 3 is transmitted by using λ_3 from the moment T_{C0} to T_{C3} in this cycle. Thanks to the dedicated connection between each server and the rack controller, there is no collision during the control message exchange.

Note that each server only transmits the granted data packets according to the Grant message, which can be only a portion of the total number of buffered packets that was reported in the Request message. To make sure that the Grant messages arrive to each server in time, the Request and Grant messages for cycle C can be exchanged at the moment T_{RG-C} before the end of previous cycle C-1 (i.e., T_C), where T_{RG-C} is the starting time to transmit Request messages at each sever in cycle C (see Fig. 2). We consider the Request messages are sent at least the time period $P_{Control}$ before the end of the cycle, where $P_{Control}$ equals to the transmission time of two maximum Ethernet frames plus their propagation delay as well as computation time of DBA for the next cycle. Therefore, each server is guaranteed to receive the Grant messages before the data transmission starts in the next cycle.

A. Request Message

An example of the Request message is shown in Fig. 2. The first field of the Request message contains the identifier of the current time cycle (e.g., C shown in Fig. 2), which is used by the rack controller to check whether the received control messages are outdated or not. If yes, it is discarded. The second field contains the MAC address of the source server, i.e., the server that generates the Request message (e.g., Server 1). This is followed by *N* fields containing all the possible destination MAC addresses (*N-1* for the other servers in the rack plus one for the interface towards the aggregation/core tier) with the corresponding number of buffered bytes, i.e., the bytes that are ready to be transmitted at the source server. The

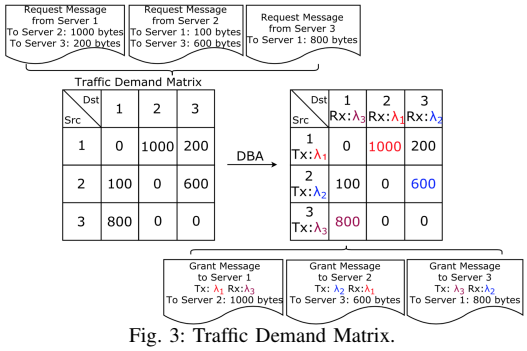


Fig. 3: Traffic Demand Matrix.

length in bytes of the Request message (L_R) can be calculated as:

$$L_R = L_{CH} + L_{SRCMAC} + N \times (L_{DSTMACi} + L_{Si}) \quad (1)$$

Where L_{CH} is the length of the first field for the identifier of the time cycle, L_{SRCMAC} and $L_{DSTMACi}$ are the length of source and the i_{th} destination MAC address (6 bytes for each MAC address), and L_{Si} is the length of reported size of the traffic sent to the i_{th} destination. Assuming L_{CH} = 8 bytes and L_{Si} = 4 bytes, one Request message as an Ethernet frame with a maximum size of 1518 bytes can support up to more than 140 servers in a rack, which is sufficient for a typical size of ToR interconnect in the data center.

B. Grant Message

An example of the Grant message is shown in Fig. 2. The first field contains the identifier of the current cycle (e.g., C shown in Fig. 2), while the second field contains the timestamp of the next cycle (e.g., T_{C+1} in Fig 2.) for which the Grant is valid. The third field contains the MAC address for the server for which the Grant message is intended (e.g., Server 3 in Fig 2.). The following three fields specify: (i) a time slot with starting timestamp (e.g., T_{C0} in Fig. 2) and ending timestamp (e.g., T_{C3} in Fig. 2); (ii) the assigned wavelengths for transmitter (e.g., λ_3 in Fig.2) and receiver (e.g., λ_2 in Fig.2) during this time slot; (iii) the destination MAC address (Server 1) as well as the granted size (e.g., 800 bytes in Fig.2) for the transmission during this time slot. The time slot can be either the entire time cycle or a part of the time cycle. This allows the DBA to assign multiple time slots to the same source server during a single time cycle. In this way the source server can send traffic to different destinations during the same time cycle improving the bandwidth utilization. As an example, in Fig. 2 Server 2 is assigned two time slots in the same time cycle: one for transmitting to Server 3 and one for transmitting to Server 1. Note that this might require reconfiguring the ONIs either at the source (i.e., tuning the WTT of Server 2 to λ_3) or at the destination server (i.e., tuning the WTF of Server 1 to λ_2).

The length of a Grant message L_G can be calculated as:

$$\begin{aligned} L_G = & L_{CH} + L_{SRCMAC} + L_{NCT} + \\ & K \times (2{\times}L_{TS} + 2{\times}L_{WI} + L_{DSTMAC} + L_S) \end{aligned} \quad (2)$$

where L_{NCT} and L_{TS} are the timestamp length of next cycle and starting/ending moment for data transmission, respectively (e.g. the timestamp of T_{C+1}, T_{C0} and T_{C3} in Fig. 2), K is the number of assigned time slots according to the DBA algorithm (which will be elaborated in the next section), and L_{WI} is the length of wavelength identifier. The other symbols have the same meaning as in the Request message. If we consider an 8-byte timestamp and 1-byte wavelength identifier, the maximum Ethernet frame (1518 bytes) can include up to 50 tunings in one cycle in the Grant message. Note that tuning the transceiver takes time, and too many tunings can decrease the bandwidth utilization in one cycle. Thus, we set the maximum allowable tuning times in one DBA cycle to less than 50.

IV. Dynamic Bandwidth Allocation Algorithms

In each cycle the rack controller collects the Request messages from the servers and accordingly builds the traffic matrix. An example of the traffic matrix generated by the rack controller is shown on the left side of Fig. 3. The rows and columns indicate the source and destination servers, respectively. Each element represents the amount of bytes that are ready to be transmitted between the corresponding source and destination servers. The DBA algorithm is responsible for assigning the available wavelengths to different elements in the matrix, while making sure that there is no collision in the network (i.e., the DBA needs to make sure that there is no wavelength clash and that the same transmitter/receiver is not assigned to more than one wavelength at the same time). A collision happens every time multiple elements within one row or one column are assigned the same wavelengths. On the right side of Fig. 3 there is an example of the results generated by the DBA algorithm. Note that assigning a wavelength to an element in the traffic matrix is equivalent to the tuning the WTT and WTF of the corresponding source and destination servers to this wavelength.

The problem described above is similar to the classical switch scheduling problem for which several optimal algorithms have already been proposed. These optimal algorithms are usually based on matrix decomposition. An example is the Birkhoff von Neumann (BvN) algorithm which was used in [7] for finding the optimal scheduling solutions to configure the circuit switches in data centers. However, due to the high complexity of the matrix decomposition process such algorithms are not suitable for POTORI.

Other suboptimal algorithms have been proposed to reduce the complexity of solving the switch scheduling problem. For example, one algorithm that has gained a lot of attention for application in electronic switches is iSLIP [13]. However, iSLIP is not designed to support WDM and thus it cannot be directly applied in POTORI. In this section, we first modify the iSLIP algorithm to adapt for POTORI as the benchmark, and then introduce two new greedy DBA algorithms designed specifically for POTORI, namely *Largest First (LF)* and *Largest First with Void Filling (LFVF)*.

A. Benchmark: Adapted iSLIP

In electronic packet switching scheduling, the iSLIP algorithm is an improved Round-Robin Matching (RRM) algorithm. It associates two Round-Robin (RR) schedulers with each input and output port of a crossbar switch fabric. There are three key steps in the iSLIP algorithm: Request, Grant, and Accept. (1) Each input port first sends the Request to the

Algorithm 1 Largest First Algorithm

1: **Input: M**; **W**; **const R**
2: *%Input: TDM M, wavelength list W, transceiver data rate R*
3: $tX \leftarrow [None, None...]$; $txTime \leftarrow [0, 0]$
4: $rX \leftarrow [None, None...]$; $rxTime \leftarrow [0, 0]$
5: List **T** $\leftarrow$ **M**.$sort()$
6: **repeat**
7: $D \leftarrow$ **T**[0]
8: **if** $D.tX$ is **None** and $D.rX$ is **None then**
9: $D.assigned \leftarrow$ **True**
10: $tX[D.src] \leftarrow$ [**W**[0] : [0, $D.size/R$]]
11: $rX[D.dst] \leftarrow$ [**W**[0] : [0, $D.size/R$]]
12: $txTime[D.src] \leftarrow D.size/R$
13: $rxTime[D.dst] \leftarrow D.size/R$
14: delete **W**[0]
15: delete **T**[0]
16: **until T** or **W** is Empty
17: **return** tX, rX

Fig. 4: Largest First Algorithm

output port according to its buffered traffic elements. (2) Each output RR scheduler chooses an input from all the requested elements to this output as a Grant according to the output arbiter. (3) Each input RR scheduler then Accepts one output from the replies of output RR scheduler according to the input arbiter to complete the match. The arbiters of the input and output RR are updated only if the match is completed. The iSLIP algorithm is easy to implement in the hardware, and it achieves 100% throughput for uniformly distributed Bernoulli arrivals, but may not be efficient for the other arrival traffic patterns such as lognormal distribution [13][14], which is often used to model the real traffic pattern for data centers. The iSLIP algorithm finds up to N matches of the input and output from a traffic demand matrix, where N is the size of the matrix. A wavelength can be assigned to each match. In the case where the total number of the available wavelengths W is lower than the number of matches N, we randomly pick W pairs from the result of iSLIP to assign the wavelengths.

B. Largest First

The Largest First (LF) is a greedy heuristic algorithm (Fig. 4). It prioritizes the largest elements in the traffic matrix. First, the matrix elements are sorted into a traffic list in a descending order according to the size (Line 5 in Fig. 4). Then, the first element is extracted from the list and an available wavelength is assigned to this element if there is no collision (Line 7-8 in Fig. 4). A collision occurs when either the transmitter or the receiver associated to the element has already been assigned to another wavelength. In this case, the element is skipped and left for the next cycle. If an element is assigned to a wavelength, the related information such as source, destination and transmission time will be updated in the Grant message (Line 10-14 in Fig. 4). The iteration stops when all the available wavelengths are assigned or the last element in the demand list is reached (Line 16 in Fig. 4). Note that the actual cycle length is decided by the data transmission time of the largest element scheduled for transmission.

C. Largest First with Void Filling

Since the size of each element in the traffic matrix is different, in LF algorithm the assigned wavelengths for the

Algorithm 2 Largest First with Void Filling Algorithm

1: **Input: M**; **W**; **const R**,$\mathbf{T}_t$,$\mathbf{T}_{tnMax}$**=50**; *%Same to the LF with the tuning time T_{tn} and max number of tuning 50*
2: *Same to the LF Line 3 - Line 16*
3: $T_{tn} \leftarrow [1, 1...]$
4: List **T'** $\leftarrow$ **M**.$sort()$
5: $T_{mac} \leftarrow$ **T'**[0]/**R**
6: **for** D' in **T' do**
7: **if** $D'.assigned$ is **False** and $\mathbf{T}_{tn}[D'.src] < T_{tnMax}$ **then**
8: **if** $txTime[D'.src]$,$rxTime[D'.dst] < \mathrm{T}_{max}$ **then**
9: $ts \leftarrow$max($txTime[D'.src]$,$rxTime[D'.dst]$)+T_t
10: $tX[D'.src].append(rx[D'.dst]:[ts,ts+D'.size/R])$
11: $rX[D'.dst].append(rx[D'.dst]:[ts,ts+D'.size/R])$
12: $txTime[D.src] \leftarrow ts+D'.size/R$
13: $rxTime[D.dst] \leftarrow ts+D'.size/R$
14: $T_{tn}[D'src]$++
15: **return** tX, rX

Fig. 5: Largest First with Void Filling Algorithm

elements lasts different duration of time within one cycle. For example, one wavelength can be assigned to an element with size of 1 Kbytes while another wavelength is assigned to an element with size of 600 bytes. In the channel of the latter wavelength, there is an idle time at least for 400 bytes until the next cycle. We refer to it as the void in the LF algorithm. The Largest First with Void Filling (LFVF) is an extension of the LF algorithm, which tries to fill in the void with the unassigned elements in the traffic matrix. A skipped element in the LF algorithm can be assigned a wavelength when both transmitter and receiver of this element are done with the previous transmission of the other elements (Line 6-8 in Fig. 5). Similar to the LF algorithm, the wavelength assignment is done in a descending ordering (i.e., starting from the largest skipped element). The Grant Message can be extended by including the new assigned demand with the information of the wavelength, transmission time as well as the destination (Line 10-13 in Fig. 5).

V. Performance Evaluation

In this section we evaluate the latency and packet drop performance of POTORI. We examine two proposed DBA algorithms, LF, LFVF and the benchmark adapted iSLIP using a custom-built discrete-event-driven simulator implemented in Python. All the results shown in the following subsections are with 95% confidence level.

Traffic Model

The traffic model utilized for the simulations is derived from [9] and [10]. The packet inter-arrival time follows a lognormal distribution and the size of the packets follows a bimodal distribution (i.e., with most of the packets having size of either 64-100 bytes or 1500 bytes) [9]. The data rate of the tunable transceivers on the ONIs is set to 10 Gb/s. The oversubscription rate (R_o) is set to 1:4. For example, if we consider 96 servers in one rack, there are 24 transceivers each with data rate of 10Gb/s as the uplink ports to handle the traffic from/to the aggregate/core layer. The servers and the uplink ports generate packets to different destinations. We assume that 80% of the traffic generated by the servers is the local intra-rack traffic, whose destination is uniformly distributed to all the other servers in the rack. The remaining 20% of the traffic

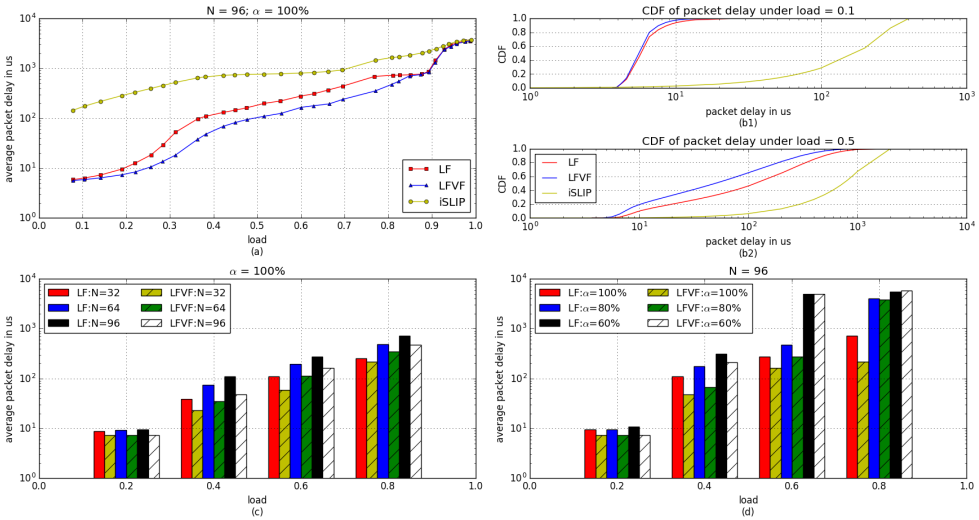


Fig. 6: (a) average packet delay of three DBA algorithms; (b) CDF of packet delay under (b1) load=0.1 and (b2) 0.5; (c) average packet delay with different N; (d) average packet delay with different α

is inter-rack traffic that is destined to the interfaces toward the aggregation/core tier, and is also uniformly distributed among all the uplink ports. We set the buffer size on each ONI to 5MB, the propagation delay to 50 ns (corresponding to 10m fibers for interconnection within the rack) and the fast tuning time of the transceivers to 200 ns [16]. We define two parameters N and α, where N is the number of servers in a rack and α is the ratio between the number of wavelengths and $N(1+R_o)$ (i.e., the sum of the number of servers and uplink port transceivers). For example, given the Ro=1:4, N=96, and α=100%, it means that the number of the total available wavelengths is 120.

A. Packet Delay

In Fig. 6 we present the packet delays under different load. Fig. 6(a) shows the average packet delays in a rack with 96 servers (i.e., N=96) and α=100%. In the iSLIP algorithm, a traffic demand in the Request messages is assigned a wavelength if and only if the arbiters of input RR scheduler and output RR scheduler are rotated to the source and destination of the demand, respectively. This may take several cycles even under the very low load, which results in the dramatic increase of the packet delay. LF and LFVF, on the other hand, outperform iSLIP, especially under low load condition. It can be seen that LFVF introduces significantly lower delays (up to 60%) compared to LF under medium loads, thanks to its ability of filling the voids in the data transmission. However, the difference in performance between LF and LFVF is insignificant if the load is very low (i.e., lower than 0.1) or very high (i.e., higher than 0.8). This is due to the fact that in the low load condition the traffic matrix is sparse and all the elements are assigned a wavelength by the LF algorithm, so that there are no or very few elements left that the LFVF can use to fill potential voids. On the other hand, in the high load condition most of the wavelengths are fully utilized even in the LF algorithm so that there are very few voids to fill in by the LFVF. In the real traffic scenario in the data centers, the load is normally between 0.1 and 0.5 and can reach peaks of 0.8 [10], so in the average working conditions LFVF outperforms the other considered algorithms.

Fig. 6(b1) and Fig. 6(b2) show the cumulative distribution functions (CDFs) of the packet delay under load = 0.1 and 0.5, which represent low and high load scenarios, respectively. Fig. 6(b1) shows that at load = 0.1 using both LF and LFVF all the packets experience delays lower than 20 μs. However, Fig. 6(b2) shows that at load = 0.5 the packet delays can vary from 10 μs to up to 1 ms. The long tail at load = 0.5 is due to the fact that both LF and LFVF prioritize the transmission of the largest traffic demands. This can potentially bring a problem for the network fairness (i.e., some packets that belong to small traffic demands may experience relatively long delays). We will address this issue in our future work.

Fig.6(c) shows the average packet delays for LF and LFVF as a function of the number of servers in the rack (i.e., N=32, 64, 96). As we may expect, the delay performance degrades with the increase of the number of servers in the rack. This is due to the fact that the number of elements in the traffic matrix increases significantly with the number of servers, while given a fixed value of α, the number of available wavelengths increases linearly with the number of servers. In low load condition (i.e., load = 0.2) the difference in the average packet delays for different values of N is very small. However, with higher load values (i.e., load = 0.4, 0.6 and 0.8) the average packet delays with N=96 become almost double as high as with N=32 and 50% higher than with N=64.

Fig. 6(d) shows the packet delay with different values of α. Given a fixed number of servers in the rack, i.e., N=96, we examine the impact of the number of available wavelengths on the packet delays, i.e., we vary α in the

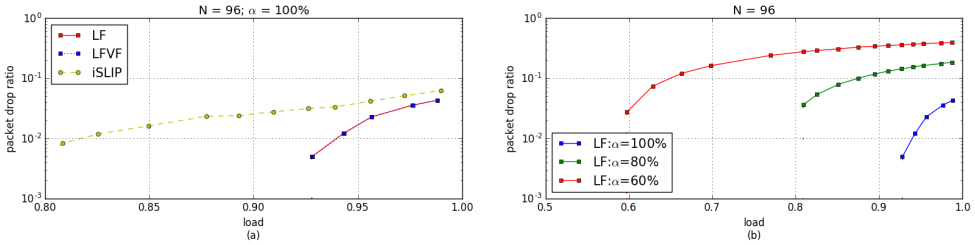


Fig. 7: Packet Drop Ratio of (a) three DBA algorithms with α= 100%; (b) α= 100%, 80%, 60% in the LF algorithm

range 100% (120 wavelengths), 80% (96 wavelengths) and 60% (72 wavelengths). Under low load (i.e., load = 0.2), the packet delays are almost the same for the considered values of α. This indicates that although in low load scenarios fewer wavelengths can be assigned to POTORI good performance in terms of packet delays (i.e., lower than 10 μs) can be maintained. However, with higher load values, the packet delays with α=60% are much higher than the ones with α=80% and α=100%.On the other hand, the difference in terms of average packet delays with α=80% and α=100% are not very significant.

B. Packet Drop Ratio

In POTORI packet drop can happen when the network is congested. In this situation new packets generated by the servers will be dropped at the ONIs. Note that in our simulation we set the size of the buffers at the ONIs to 5 MB.

In Fig. 7(a) we compare the packet drop ratio achieved with LF, LFVF and iSLIP assuming *N*=96 and α=100%. It can be observed that for load values lower than 0.9 the packet drop ratio of LF and LFVF ratio is negligible. LF and LFVF have the same performance due to the fact that at very high load they behave in the same way (the reason is that at very high load there are not that many voids to be filled, as explained previously). On the other hand, iSLIP shows higher packet drop ratio. It is worth noting that in the access tier of data centers the load usually does not reach values higher than 0.9 [9] so that LF and LFVF can achieve negligible packet drop ratio and 100% throughput in most cases.

Fig. 7(b) shows the impact of changing the parameter α on the packet drop ratio. Since the packet drop ratio of LFVF is the same as of LF, we are only showing the performance of LF. In the LF algorithm, when α equals to 60% (80%) the network gets congested and starts dropping packet at load=0.6 (0.8), which can also be observed by the sudden increase in the packet delays in Fig. 6.

VI. Conclusion

In this paper we proposed and evaluated a control plane for the passive optical ToR interconnect (POTORI) architecture. It is based on a centralized rack controller, which is responsible for the management of the communications inside the rack and toward the aggregation/core tires. A MAC protocol and two DBA algorithms, namely *Largest First* and *Largest First with Void Filling*, are also proposed aiming to achieve high network performance in POTORI.

The simulation results prove that POTORI obtain send-to-end packet delays in the order of microseconds and negligible packet loss probability under realistic data center traffic scenarios. In addition, the simulation results show that the proposed DBA algorithms for POTORI outperform the performance of a representative existing switch scheduling algorithm.

Acknowledgment

This work was supported by the Swedish Foundation for Strategic Research (SSF), Swedish Research Council (VR), Göran Gustafssons Stiftelse and National Natural Science Foundation of China (Grant No. 61550110240).

References

[1] Cisco Global Cloud Index: Forecast and Methodology, 2014-2019, October 2015, Cisco white paper.
[2] A. Andreyev, "Introducing data center fabric, the next-generation Facebook data center network." https://code.facebook.com/posts/360346274145943, 2014.
[3] Data Center Design Considerations with 40 GbE and 100 GbE, Aug. 2013, Dell white paper.
[4] R. Pries *et al*.,"Power consumption analysis of data center architectures," *Green Communications and Networking*, 2012.
[5] M. Fiorani *et al*., "Energy-efficient elastic optical interconnect architecture for data centers," *IEEE Communications Letters*, vol.18, pp. 1531-1534, Sept. 2014.
[6] K. Chen *et al*., "OSA: An optical switching architecture for data center networks with unprecedented flexibility," *IEEE/ACM Trans. Networking*, vol. 22, no. 2, pp. 498-511, Apr. 2014.
[7] G. Poter *et al*., "Integrating microsecond circuit switching into the data center," in *Proc. ACM SIGCOMM Conf.*, 2013 pp. 447-458.
[8] W. Ni *et al*., "POXN: a new passive optical cross-connection network for low cost power efficient datacenters," *IEEE/OSA Journal of Lightwave Technology*, vol. 32, pp. 1482-1500, Apr. 2014.
[9] A. Roy *et al*., "Inside the Social Network's (Datacenter) Network," in *Proc. ACM SIGCOMM Conf.*, 2015 pp. 123-237.
[10] S. Kandula *et al*., "The Nature of Datacenter Traffic: Measurement &Analysis," in *Proc. ACM SIGCOMM Internet Eas. Conf.*, 2009, pp. 202-208.
[11] J. Chen *et al*., "Optical Interconnects at Top of the Rack for Energy-Efficient Datacenters," *IEEE Communications Magazine*, vol. 53, pp. 140-148, Aug. 2015.
[12] Y. Cheng *et al*., "Reliable and Cost Efficient Passive Optical Interconnects for Data Centers," *IEEE Communications Letters*, vol. 19, pp. 1913-1916, Nov. 2015.
[13] N. McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches," *IEEE/ACM Trans. on Networking*, vol. 7, no 2, pp.188-201, 1999.
[14] T. Javadi *et al*., "A high-Throughput Algorithm for Buffered Crossbar Switch Fabric," *Proceedings IEEE ICC*, pp. 1581- 1591, June 2001.
[15] IEEE 802.3 ETHERNET WORKING GROUP, http://www.ieee802.org/3/
[16] R. Maher *et al*., "Fast Wavelength Switching 112 Gb/s Coherent Burst Mode Transceiver for Dynamic Optical Networks," In *ECOC*, 2012.
[17] L. Khermosh, "Managed Objects of Ethernet Passive Optical Networks (EPON)", RFC 4837, July 2007.
[18] "IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications".

Paper III

POTORI: A Passive Optical Top-of-Rack Interconnect Architecture for Data Centers

Yuxin Cheng, Matteo Fiorani, Rui Lin, Lena Wosinska, Jiajia Chen

IEEE/OSA Journal of Optical Communications and Networking (JOCN), to appear.

POTORI: A Passive Optical Top-of-Rack Interconnect Architecture for Data Centers

Yuxin Cheng, Matteo Fiorani, Rui Lin, Lena Wosinska, and Jiajia Chen

***Abstract*—Several optical interconnect architectures inside data centers (DCs) have been proposed to efficiently handle the rapidly growing traffic demand. However, not many works have tackled the interconnects at top-of-rack (ToR), which have a large impact on the performance of the data center networks (DCNs) and can introduce serious scalability limitations due to the high cost and power consumption. In this paper, we propose a passive optical ToR interconnect architecture (POTORI) to replace the conventional electronic packet switch (EPS) in the access tier of DCNs. In the data plane, POTORI relies on a passive optical coupler to interconnect the servers within the rack and the interfaces toward the aggregation/core tiers. The POTORI control plane is based on a centralized rack controller responsible for managing the communications among the servers in the rack. We propose a cycle-based medium access control (MAC) protocol to efficiently manage the exchange of control messages and the data transmission inside the rack. We also introduce and evaluate a dynamic bandwidth allocation (DBA) algorithm for POTORI, namely *Largest First* (*LF*). Extensive simulation results show that, with the use of fast tunable optical transceivers, POTORI and the proposed LF strategy are able to achieve an average packet delay below 10 μs under realistic DC traffic scenarios, which outperforms conventional EPSs. On the other hand, with slower tunable optical transceivers, a careful configuration of the network parameters (e.g., maximum cycle-time of the MAC protocol) is necessary to obtain a good network performance in terms of the average packet delay.**

***Index Terms*—Optical interconnect architectures, data center networks, medium access control (MAC), dynamic bandwidth allocation (DBA).**

I. Introduction

The growing popularity of modern Internet applications such as cloud computing, social networking and video streaming is leading to an enormous increase of data center (DC) traffic, including not only the north-south (client-server) traffic, but also the east-west (server-to-server) traffic exchanged within the DCs [1]. According to Cisco, the overall data center traffic will keep increasing at a compound annual growth rate (CAGR) of 25% up to the year 2019, reaching 10 Zettabytes per year [2]. Therefore, it is important to evolve the current data center network (DCN) infrastructure to support the continuously growing traffic demand.

Data center operators are addressing this problem by upgrading the transmission data rate and switching capacity of their network equipment. For example, Facebook has already deployed 10G Ethernet network interface cards (NICs) for all servers and Top-of-Rack (ToR) switches [3]. Optical fiber can be deployed in DCN to interconnect servers and switches in order to simplify cabling and avoid electromagnetic interference (EMI) [4]. Higher data rate and switching capacity (e.g., 40G, 100G) are also taken into consideration by the network operators in the future DC design [5]. However, it is hard to develop large electronic packet switches operating at high data rates, due to the bottleneck of I/O bandwidth and power budget of the chip [6]. As a consequence, a large amount of electronic switches need to be deployed to scale out the number of servers in the DC, which brings a serious scalability problem to the DCN in terms of cost and power consumption [7].

Optical interconnect architectures that are able to provide ultra-high transmission speed and switching capacity in a cost- and energy-efficient way, are considered to be a promising solution to address the limitations of electronic packet switches (EPSs) in DCs. By replacing EPSs with optical switches, the reduced power-demanding electrical-to-optical (E/O) and optical-to-electrical (O/E) conversion is expected to dramatically decrease the power consumption of data center networks [8]. Several optical interconnect architectures for DCs are proposed in literature in recent years, e.g., [13]—[16]. These architectures employ all-optical switches based on different topologies and technologies at aggregation/core layer, but rely on conventional EPSs at ToR to interconnect servers in the racks. However, the EPSs at ToR are responsible for a large amount of the overall DC traffic. For example, it is reported in [17] that in the DCs running extensive data exchange applications (e.g., MapReduce) around 80% of the total traffic is confined in the access tier. Moreover, the EPSs at ToR contribute to the majority of the overall DCN power consumption [21]. Therefore, efficient optical interconnect architectures for the access tier in DCs are required.

In our previous work (i.e., [8], [9], [21]), we proposed a concept of passive optical interconnect (POI) which uses mainly passive optical components for interconnection. The physical layer experiments [9] have shown that more than 500 ports can be supported in the passive optical interconnect at the capacity up to 5Tb/s. We also shown that the passive optical interconnect provides lower cost, lower energy consumption and higher scalability with respect to the conventional EPSs. Specifically, it has been demonstrated in [8] that the energy consumption per bit in the DCNs can be reduced by at least a factor of 7 by using passive optical interconnects at ToR compared to the ones using EPS. We also proposed a MAC protocol and a dynamic bandwidth allocation (DBA) algorithm, namely *Largest First* (*LF*), for achieving efficient bandwidth utilization when applying the passive optical interconnect at ToR [18]. This paper extends the work in [18] with a focus on passive optical ToR interconnect (POTORI) and introduces the following new contributions: (*i*) We illustrate how POTORI can be interconnected with other network architectures in the

The authors are with KTH Royal Institute of Technology, Department of Communication Systems, Electrum 229, SE- 164 40 Kista, Sweden (e-mail: {yuxinc, matteof, ruilin, wosinska, jiajiac}@kth.se).Corresponding author: Jiajia Chen.

aggregation/core tier to build large DCNs, where both the data plane and control plane are considered; (*ii*) We perform an extensive performance comparison among different DBA algorithms for POTORI; (*iii*) We study the impact of different network configuration parameters (e.g., tuning time of the optical transceivers, duration of the cycle time in the MAC protocol, etc.) on the performance of POTORI; and (*iv*) We compare the performance of POTORI with a conventional EPS in terms of the average packet delays and packet loss probability. The results show that using our proposed LF DBA algorithms along with ultra-fast tunable transceivers, POTORI can outperform the conventional EPS.

The rest of the paper is organized as follows. We present the related works on optical interconnect architectures for DC in Section II. In Section III, we illustrate the POTORI architecture, including both data plane and control plane. The proposed centralized MAC for POTORI is elaborated in Section IV, and in Section V we introduce and analyze the proposed DBA algorithm. The simulation results of POTORI and a conventional EPS are presented and discussed in Section VI. We conclude the paper in Section VII.

II. Related Works

Several optical interconnect architectures for DCN have been proposed in the literature. The c-through [10] and HOS [11] are two examples of hybrid electronic/optical interconnection solutions. In these hybrid interconnect architectures, optical circuit switches (OCS) are employed to provide high capacity for transmitting long-lived and bandwidth-consuming traffic flows (e.g., elephant flows) due to the long reconfiguration time of OCS, and EPS is used to transmit short-lived traffic flows that do not need large bandwidth (e.g., mice flows). These solutions require pre-knowledge or classification of traffic pattern in order to distinguish large and small flows and properly configure the OCS, which is challenging for DC operators.

On the other hand, there are some all-optical architecture solutions proposed recently. In [12], the authors demonstrated a new optical circuit switching (OCS) based architecture for DCN, which is based on a single comb-driven MEMS mirror and is able to achieve a switching time of 20 μs. However, such fast switching might still create a substantial delay in case of a small amount of data to be transmitted, making it not suitable to be employed at ToR where small bursts of intra-rack traffic need to be handled. The authors in [13] proposed a flat data center network architecture with fast flow control. Each ToR switch is connected to one intra-cluster optical switch as well as one inter-cluster optical switch. All the traffic is switched by optical switches according to the flow control mechanism, which is based on the packet header processing on each electronic ToR switch. OPMDC [14] is a three-tier architecture, where each tier is a set of reconfigurable optical add-drop multiplexers (ROADM) connected in ring topology, and the ROADM rings in the lower tier are connected to a ROADM in the upper tier. At the access tier each ToR switch is connected to a single ROADM. Space division multiplexing (SDM) is also considered in all-optical DCN solutions, e.g., [15][16] for improving capacity. In [15], four architectures based on SDM are proposed, and it is shown that these architectures are suitable to apply in different DCs depending on their size and work load. The authors in [16] reported an optical data center architecture based on multidimensional switching nodes connected in ring topology. These switching nodes are able to switch in space, wavelength and time domains, supporting the connections of different granularities. The ring topology reduces the number of physical links, simplifying the cabling management. Nevertheless, all these aforementioned architectures rely on optical switching only in the aggregation/core tires, while they are based on conventional electronic ToR switches in the access tier to interconnect the serves within the same rack. In [20], the authors proposed and demonstrated software-defined ubiquitous data center optical interconnection (SUDOI), which also considers optical switch at ToR. However, the main focus of SUDOI is on the control plane, where a service-aware schedule scheme is introduced to enable cross-stratum optimization of application and optical network stratum resources while enhancing multiple-layer resource integration. The concrete design of optical interconnects within DCs is not provided.

In [21] and [22] we proposed several passive optical ToR interconnect architectures for the access tier in DCs. These architectures use passive optical components (i.e., arrayed waveguide grating (AWG) and/or optical couplers) to interconnect the servers in the rack. It has been demonstrated that the passive optical components offer cost and power saving as well as high reliability. While in [21] and [22] we focused on the data plane architecture, in [18] we proposed a MAC protocol and novel DBA algorithms to achieve efficient bandwidth utilization in POTORI. The current paper extends our previous work by covering both data and control plane, and provides a complete design of POTORI architecture along with a detailed analysis of the network performance and an extensive comparison with the conventional EPS solutions.

III. POTORI: Passive Optical Top-of-Rack Interconnects

Fig. 1 illustrates the POTORI architecture, including both data plane and control plane. Each server is equipped with an optical network interface (ONI), which consists of two optical transceivers. The first one is a tunable transceiver connected to the POTORI data plane. The second one is a grey small form factor pluggable (SFP) transceiver connected to the POTORI control plane. In the following subsections, we elaborate the POTORI data and control planes, mapping POTORI to the use case of DCNs.

Data Plane

The POTORI data plane was proposed and introduced in our previous work [22]. The key component of POTORI data plane is an $(N+1)\times(N+1)$ passive coupler that acts as the switching fabric to interconnect all the servers in the rack. We define N as the number of servers in the rack. In each ONI, the tunable transceiver is composed of a wavelength tunable transmitter (WTT) for transmitting data, and a wavelength tunable filter (WTF) as well as a receiver for receiving data. The WTT and WTF are connected to one input port and one output port of the coupler, respectively. One additional pair of input and output ports of the coupler are connected to a wavelength selective switch (WSS), which forwards and receives the traffic to/from the aggregation and core tiers.

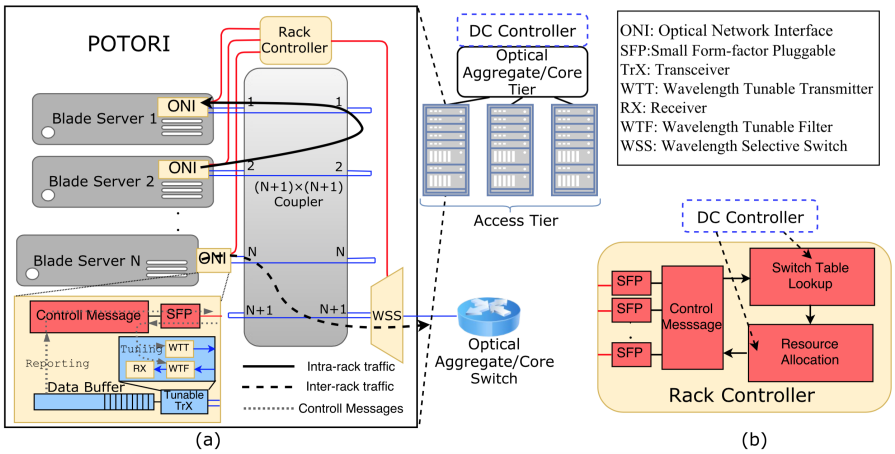


Fig. 1: (a) POTORI Architecture in Data Center, (b) Rack Controller

It can be seen that the POTORI's data plane is passive, except the WSS that is needed to dynamically filter out spectrum for the inter-rack communications. Actually, WSS can be replaced by the passive wavelength filter, in which a fixed configuration of the spectrum for intra- and inter-rack communications may result in low flexibility of resource usage. Due to the key component — coupler, the data transmission in POTORI follows the broadcast-and-select scheme. The traffic transmitted by one server is broadcast and received by all the other servers and the WSS. The destination server (or the WSS) then selects the traffic destined to it, while discarding the other traffic. In this way, the servers are able to send/receive traffic to/from each other in the rack (e.g., Server 2 sends traffic to Server 1 in Fig.1). The WSS receives and drops the intra-rack traffic while forwarding the inter-rack traffic to/from the upper tier (e.g., Server N sends traffic to the aggregation/core tier in Fig.1).

Control Plane

In order to successfully transmit data, both ONIs at source server and destination server need to be tuned to the same wavelength. To avoid data collision in the coupler, concurrent communications inside the rack can be carried on the different wavelengths. It calls for a proper control plane design to efficiently schedule resources in both spectrum and time domains for managing the intra-rack and inter-rack communications.

The proposed centralized control entity for POTORI, namely rack controller, is shown in Fig. 1. The rack controller exchanges control information with the ONIs using dedicated control links. The servers report the relevant information (e.g., buffer size) to the rack controller and tune the transceivers according to the feedback from the rack controller. The POTORI MAC protocol defines the exchanging procedure and the format of the control messages between the servers and the rack controller, which will be elaborated in Section IV. On the other hand, the rack controller collects the necessary traffic information from each server and creates the traffic matrix. Then it runs a DBA algorithm, determining the wavelength and time slots assigned for all the servers in the rack. Finally, it generates the control messages that include the outcome of the DBA and sends them to each server.

Application of POTORI in Data Center Networks

The POTORI architecture can be interconnected with any solution for the aggregation/core tiers to build large DCNs. In the data plane proper interfaces are needed to interconnect the POTORI with the aggregation/core switches. These interfaces can employ O/E conversion for connection to the conventional EPS in aggregation/core tier or they can be optical (e.g., to directly connect the POTORI to an optical core switch and realize an all-optical DCN [8]). In the latter case, a strategy for the joint allocation of the optical resources in the access and aggregation/core tires needs to be developed.

In the control plane the rack controller can be connected to a higher layer DC controller in a hierarchical control architecture (see Fig. 1(b)). In this way the DC operator can employ a single control infrastructure to manage all the resources in the DC. Depending on how the DC controller interacts with the rack controller, two different modes of operation can be defined, namely fixed-mode and programmable-mode. In the fixed-mode the DC controller is not able to influence the resource allocation inside the rack. The rack controller performs layer 2 functions, such as switch table lookup, and computes the resource allocation according to a deployed DBA algorithm. On the other hand, in the programmable-mode (see Fig. 1(b)) the DC controller can influence the resource allocation inside the rack, e.g., by changing the employed DBA algorithm dynamically. A possible way to realize a control plane operating in programmable-mode is to equip the rack controller with a configurable switch table (e.g., an OpenFlow [24] switch table) and a configurable resource allocation module (see Fig. 1(b)). Using a software defined networking (SDN) [23] DC controller is then able to dynamically change the flow rules and the DBA algorithm employed by the rack controller. In this paper, we consider only the control plane in fixed-mode, and leave the programmable-mode for the future work.

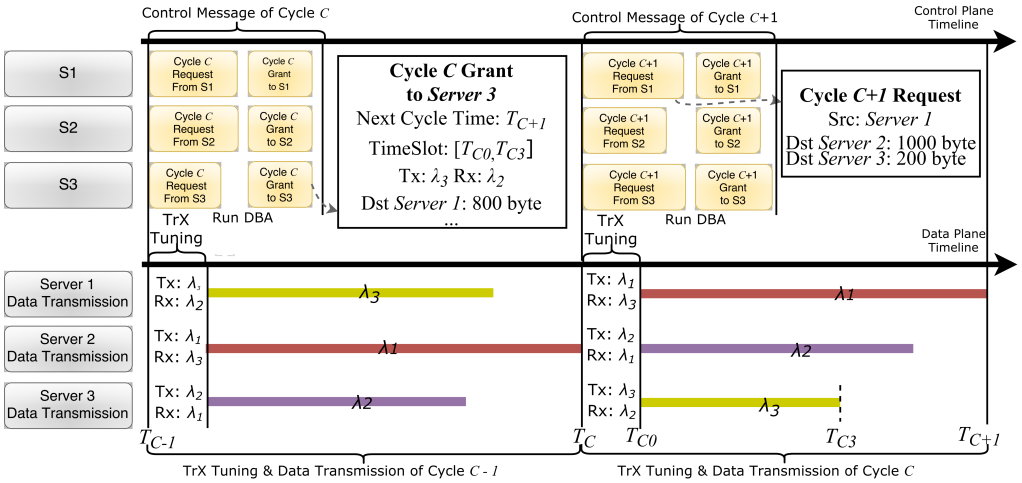


Fig. 2: The MAC Protocol and the Request-Grant Messages of POTORI

IV. MAC Protocol

Due to the broadcast property of the coupler, POTORI requires a proper MAC protocol to efficiently coordinate the data transmissions among the servers in a rack. The existing MAC protocols can be classified as distributed and centralized. In distributed MAC protocols, every server determines its own resources for data transmission, based on the direct exchange of control information with the other servers in the network and without the involvement of the centralized controller. There are some typical examples of distributed MAC protocols, such as the carrier sense multiple access with collision detection (CSMA/CD) [25] which has been widely employed in the old version of Ethernet (10 Mb/s and 100 Mb/s), and the high-efficient distributed access (HEAD) protocol which was proposed for the optical interconnect architecture POXN in [26]. However, the control overhead in the distributed MAC protocol should be taken into account. Given the large amount of short lived traffic flows within a rack in data centers, the control overhead brought by the distributed MAC protocol might be significant and the performance of the network (packet delay, packet drop ratio, etc.) will decrease. As a consequence, the distributed MAC protocols might not be good candidates for POTORI.

In contrary, a centralized controller is able to manage the exchange of control information and coordinate the data transmission among all the servers in the network. Typical examples are the IEEE 802.11 protocols [27], which are used in Wi-Fi networks, the multipoint control protocol (MPCP) [28] which is used in Ethernet passive optical networks (EPONs), and the time division multiple access (TDMA) Ethernet proposed in [29]. The TDMA Ethernet seems to be a good candidate for POTORI, as it allows one server at a time to use the entire available bandwidth for transmission in order to achieve low latency and low packet drop in the network. However, applying TDMA Ethernet to POTORI would result in a single server transmitting at a time over one wavelength and suspending the communications among the remaining servers. POTORI is able to support concurrent communications using wavelength division multiplexing (WDM), where multiple wavelengths are utilized to set up connections for different server pairs. The MPCP is another choice for POTORI, due to the similarity of the data plane in PON and POTORI. However, POTORI requires multipoint-to-multipoint communications where all the servers should be able to communicate with each other, while MPCP supports only multipoint-to-point communications, and thus it cannot be directly applied to POTORI. In addition, IEEE 802.11 standards do not support WDM, so that the control schemes utilized in Wi-Fi networks are not applicable to POTORI. In this regard, we propose a new centralized MAC protocol tailored for the POTORI architecture.

The proposed MAC protocol is shown in Fig. 2 and follows a Request-Grant approach. The time is divided in cycles. At the beginning of each cycle, each server sends a Request message to the rack controller. The Request message contains the information about the packets currently buffered at the server. After receiving the Request messages from all the servers, the rack controller is able to generate a traffic matrix and run the DBA algorithm to calculate the allocation of wavelengths and transmission time for all the servers. Afterwards, the controller informs the servers about the outcome of the resource allocation using Grant messages, which contain all the necessary information (wavelength, cycle time, etc.) for configuring the ONIs of the servers. At the beginning of the next cycle, each server will tune the WTT and WTF to the assigned wavelengths for transmitting and receiving data according to the information in the Grant message. For example, in Fig. 2, at the time T_{C-1}, all three servers (S1, S2, S3) report their buffer information to the rack controller through Request Messages for Cycle *C*, and the rack controller then responds with the Grant Messages in the same cycle. Meanwhile, all three servers tune their WTTs and WTFs, according to the Grant Message received during the previous cycle (i.e., Cycle *C-1*) and start transmitting and receiving traffic. At the time T_C (specified in the Grant Messages

received during Cycle *C*), all the servers tune the WTTs and WTFs accordingly. In the example shown in Fig. 2, Server 3 tunes the transmitter to λ_3 and the receiver to λ_2, and Server 1 tunes the transmitter to λ_1 and the receiver to λ_3 for Cycle *C*. Consequently, the traffic from Server 3 to Server 1 can be successfully transmitted in this cycle, which lasts from T_{C0}, to T_{C3}, according to the Grant message. Similarly, the Request and Grant Messages for Cycle *C+1* are exchanged in parallel with the data transmission of Cycle *C*. Thanks to the dedicated connection between each server and the rack controller, collisions among the Request and Grant Messages can be eliminated.

Note that each server only transmits the granted data according to the Grant message, which might be only a portion of the ones reported in the Request message. It is worth to mention that the granted traffic, which will be transmitted in the next cycle, is not reported in the next Report Message. For example, in Fig. 2 Server 3 receives the Grant Message of Cycle *C*, allowing to send 800 bytes to Server 1. At T_C, Server 3 subtracts the 800-byte traffic and reports the remaining data with the destination of Server 1 in the Request Message of Cycle *C+1*.

We further illustrate the structure of the Request and Grant Messages in the following sections.

A. Request Message

Fig. 2 shows an example of the Request message. The first field of the Request message contains the current time cycle identifier (e.g., Cycle *C+1* in Fig. 2), which is used by the rack controller to identify whether the received control messages are outdated or not. If a Request Message is not synchronized with the cycle identifier at the rack controller, it is discarded by the rack controller. The second field of the Request Message contains MAC address of the source server, i.e., the server that generates the Request message (e.g., Server 1 in Fig. 2). Besides, the request message should also contain in total *N* fields for all the possible destination MAC addresses (*N-1* for the other servers in the rack and one for the interface toward the aggregation/core tier), along with the corresponding number of buffered bytes, i.e., the bytes to be transmitted at the source server.

The Request Message can be encapsulated in an Ethernet frame. The length in bytes of the Request message (L_R) can be calculated as:

$$L_R = L_{CH} + L_{SRCMAC} + N \times (L_{DSTMACi} + L_{Si}) \quad (1)$$

where L_{CH} is the length of the first field in the time cycle, L_{SRCMAC} and $L_{DSTMACi}$ are the length of MAC address of the source and the i_{th} destination (6 bytes) server, and L_{Si} is the length of buffed packets size for i_{th} destination. If we assume L_{CH} = 8 bytes, L_{Si} = 4 bytes, a Request message as an Ethernet frame with the maximum size (1518 byte) can support up to 150 servers in a rack, which is sufficient for the access tier in DCs.

B. Grant Message

An example of the Grant message is shown in Fig. 2. Similar to the Request Message, the first field of Grant Message contains the current cycle identifier (e.g., Cycle *C* in Fig. 2). According to this field, the servers that newly join the network or lose synchronization to the rack controller can update their cycle identifier. The second field contains the timestamp that indicates the end of the cycle (e.g., T_{C+1} in Fig. 2). The third field contains the destination MAC address of the Grant Message, i.e. the server which the Grant message is destined to (e.g., Server 3 in Fig. 2.). The following three fields contain: (*i*) a time slot with the starting timestamp (e.g., T_{C0} in Fig. 2) and ending timestamp (e.g., T_{C3} in Fig. 2) for transmission; (*ii*) the assigned wavelengths for transmitter (e.g., λ_3 in Fig. 2) and receiver (e.g., λ_2 in Fig. 2) during the timeslot given in (*i*); (*iii*) the destination MAC address (e.g., Server 1 in Fig. 2) as well as the granted size (e.g., 800 bytes in Fig. 2) for the transmission during the time slot given in (*i*). The time slot can last either the entire time cycle or a part of the time cycle. Note that we define an extra parameter T_M as the maximum transmission time for each cycle. Each server cannot transmit data longer than T_M in each cycle, i.e., in Fig. 2 T_{C+1} - T_C should be less or equal to T_M. The value of T_M definitely affects the network performance, which will be discussed in Section VI.

The length of a Grant message L_G can be calculated as:

$$L_G = L_{CH} + L_{SRCMAC} + L_{NCT} + 2 \times L_{WI} + L_{DSTMAC} + L_s \quad (2)$$

where L_{NCT} is the length of timestamp, L_{TS} is the length of the starting/ending timestamp for data transmission, and L_{WI} is the length of the wavelength identifier. The remaining symbols are the same as the ones for the Request Message. If we would consider a timestamp of 8 bytes and a wavelength identifier of 1 byte, the length of a Grant message would be 52 bytes which is small enough to be encapsulated into one Ethernet frame.

V. Dynamic Bandwidth Allocation Algorithms

A traffic demand matrix can be built by the rack controller after receiving the Request Messages from all servers and uplink interfaces. An example is shown in Fig. 3. The rows and columns of the matrix indicate the input port (source) and output port (destination), respectively, and the matrix element represents the amount of traffic (in bytes) that needs to be transmitted from the source to the destination. The DBA algorithm should find a solution for assigning available wavelengths to the different traffic demands without any collision in the data transmission. A collision occurs when different traffic demands are assigned to the same wavelength at the same time. To avoid collisions, at most one traffic demand can be assigned to an available wavelength, i.e., each row and each column in the matrix should be associated to exactly one wavelength. The right side of Fig. 3 gives a feasible solution of the wavelength assignment without any collision. The wavelengths assigned for serving the different traffic demands are distinguished by colors. The rack controller forms the Grant Messages according to the DBA solution, indicating the wavelengths for transmitting and receiving at every server.

The problem described above is similar to the classical switch scheduling problem. A conventional electronic switch buffers incoming data traffic at input queues, and then forwards it to the output ports. The traditional crossbar switch fabric allows multiple input queues to transmit data traffic to different output ports simultaneously. Many scheduling solutions have been proposed over decades, aiming to find

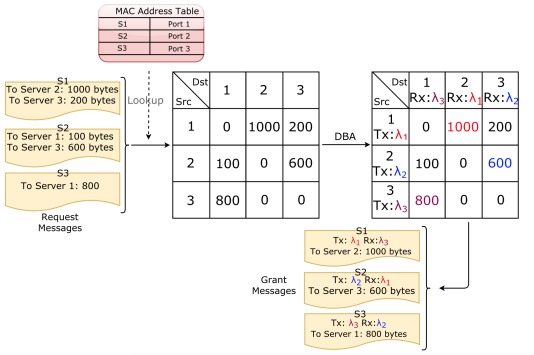


Fig. 3: Traffic Demand Matrix.

matches between input port and output port to achieve high throughput and low packet delay. Some solutions are based on matrix decomposition. For example, paper [30] uses Birkhoff von Neumann (BvN) algorithm to find optimal scheduling solutions to configure the circuit switches in DC. However, these solutions are not suitable for POTORI, since the high time complexity of matrix decomposition algorithms makes it only feasible for the scenario in [30], where the traffic demand for a long time period is known before running the algorithm. On the other hand, there are some other solutions having lower time complexity, such as iSLIP [19], which is one of the most widely used in electronic switches. However, iSLIP is not designed to support WDM. Thus we adapt the iSLIP algorithm to POTORI as a benchmark DBA algorithm to be compared with ourproposed novel DBA algorithm, referred to as *Largest First* (*LF*). We define N and W as the size of the matrix and the number of available wavelengths, respectively.

A. Benchmark: iSLIP Adapted to POTORI

In this paper we adapt iSLIP to be used in POTORI. The iSLIP algorithm is an improved Round-Robin Matching (RRM) algorithm. Each input and output port of a crossbar switch fabric is associated a Round-Robin (RR) scheduler. The detailed procedure of iSLIP algorithms can be found in [19]. Within N iterations, the iSLIP algorithm is able to find up to N matches between input and output from a traffic demand matrix. Given a traffic demand matrix, suppose that iSLIP algorithm finds N^* matches, where $N^* \leq N$. In the adapted iSLIP for POTORI, these found matches needs to be assigned different wavelengths. If $W \geq N^*$, then it is possible to assign every match with a unique wavelength. When $W < N^*$, we randomly pick W matches from iSLIP's result and assign them with different wavelengths. In original iSLIP algorithm, whenever an input-output match is found, the corresponding scheduler of this match is updated. The adapted iSLIP for POTORI updates the schedulers if and only if a match is assigned a wavelength.

The iSLIP algorithm is easy to implement in the hardware. It achieves 100% throughput for uniformly distributed Bernoulli arrivals, but may not be efficient for bursty arrival traffic patterns [19] [31], which is often more suitable to model the real traffic pattern in DCs [32].

Algorithm 1 Largest First Algorithm

1: **Input: M**; **W**; **const R**
2: *%Input: TDM M, wavelength list W, transceiver data rate R*
3: $tX \leftarrow [None, None...]$; $txTime \leftarrow [0, 0]$
4: $rX \leftarrow [None, None...]$; $rxTime \leftarrow [0, 0]$
5: List **T** $\leftarrow$ **M**.$sort()$
6: **repeat**
7: $D \leftarrow$ **T**[0]
8: **if** $D.tX$ is **None** and $D.rX$ is **None then**
9: $D.assigned \leftarrow$ **True**
10: $tX[D.src] \leftarrow$ [**W**[0] : $[0, D.size/R]$]
11: $rX[D.dst] \leftarrow$ [**W**[0] : $[0, D.size/R]$]
12: $txTime[D.src] \leftarrow D.size/R$
13: $rxTime[D.dst] \leftarrow D.size/R$
14: delete **W**[0]
15: delete **T**[0]
16: **until T** or **W** is Empty
17: **return** tX, rX

Fig. 4: Largest First Algorithm

B. Largest First

The Largest First (LF) is a greedy heuristic algorithm. It prioritizes the largest element in the traffic demand matrix, i.e., larger amount of traffic demand has higher probability to be assigned with a wavelength. First, the matrix elements are sorted in a descending order into a one-dimensional array (Line 5 in Fig. 4). Then, starting from the first element in the array, a traffic demand is assigned with a wavelength if and only if neither the transmitter nor the receiver associated to this demand have already been assigned another wavelength in the current cycle (Line 7-9 in Fig. 4). The corresponding information such as wavelength, source, destination and transmission time is used to generate the Grant Message (Line 10-13 in Fig. 4). If one of the transmitter or receiver of this demand is assigned another wavelength, the demand is not served and left for the next cycle (Line 15 in Fig. 4). The LF algorithm stops when all the available wavelengths are assigned, or the last traffic demand in the array is served (Line 16 in Fig. 4).

VI. Performance Evaluation

In this section, we evaluate the performance of POTORI and compare it with conventional electronic ToR packet switch (EPS) in terms of average packet delay and packet drop ratio. To be more specific, the packet delay consists of queuing time at the source node (servers and uplink interfaces), transmission time and propagation time. In POTORI, the ONI at servers drops packet when the buffer is full. Moreover, we examine the impact of different system configurations (e.g., selected DBA algorithms, tuning time of the transceivers, etc.) on the performance for POTORI. We build a customized discrete-event-driven simulator implemented in Java for the performance evaluation.

A. Traffic Model

The traffic model used in simulations is derived from [17][32]. Each server generates 10^6 packets where the packet inter-arrival time follows a lognormal distribution. The size of packets follows a bimodal distribution (i.e. most of packets are with size of either 64-100 bytes or around 1500 bytes), which is shown in Fig. 5(c). The data rate (R) per server is set to 10 Gb/s and we assume that the buffer size on the

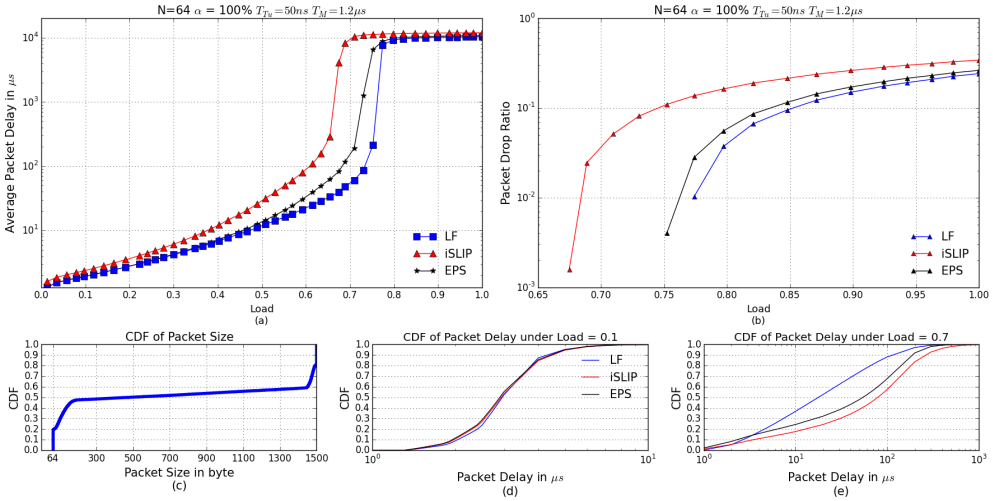


Fig. 5: (a) average packet delay; (b) packet drop ratio; (c) CDF of packet size for simulations (d) CDF of packet delay under load = 0.1; (e) CDF of packet delay under load = 0.7.

TABLE I: Table of Notations

N	Number of Servers in a rack
R	Data Rate of the Tunable Transceiver
O_R	Oversubscription Rate
α	The ratio between the number of available wavelengths and the sum of the number of servers and uplink transceivers
T_{Tu}	Tuning Time of the Transceiver
T_M	Maximum Transmission Time of Cycle
T_P	Transmission Time of an Ethernet Frame with size of 1518 bytes

ONI is 10 Mbytes. The propagation delay is set to 50 ns, which corresponds to 10 m fiber for interconnect within the rack. The network oversubscription rate O_R is set to 1 : 4 (i.e., if we consider 64 servers in a rack, there are 16 tunable transceivers for communication with the aggregation/core tiers, and the coupler with 64+16 = 80 pairs of input and output ports interconnecting all servers and uplink transceivers). Servers and uplink ports generate packets with random destinations. We assume that 80% of the traffic generated by servers is intra-rack traffic. The destination of the intra-rack traffic is uniformly distributed to all the other servers in the rack. The remaining 20% of traffic is inter-rack, whose destination is uniformly distributed among the uplink interface transceivers. Meanwhile, each uplink interface transceiver generates packets with destination uniformly distributed to all servers, representing the traffic from aggregation/core layer to the rack. We define N as the number of servers in a rack, and α as the ratio between the number of available wavelengths for a rack and $N(1+O_R)$ (i.e., the sum of the number of servers and uplink transceivers), reflecting the sufficiency of the wavelengths. In addition, the tuning time of the tunable transceivers is defined as T_{Tu}, and the maximum transmission time of a cycle is defined as T_M. Finally, Table I summarizes all the notations.

B. POTORI v.s. EPS

In this subsection, we compare the performance of POTORI using different DBA algorithms with the conventional EPS. We set the line rate to10Gb/s for both POTORI and EPS. The tuning time of the tunable transceiver for POTORI is 50 ns

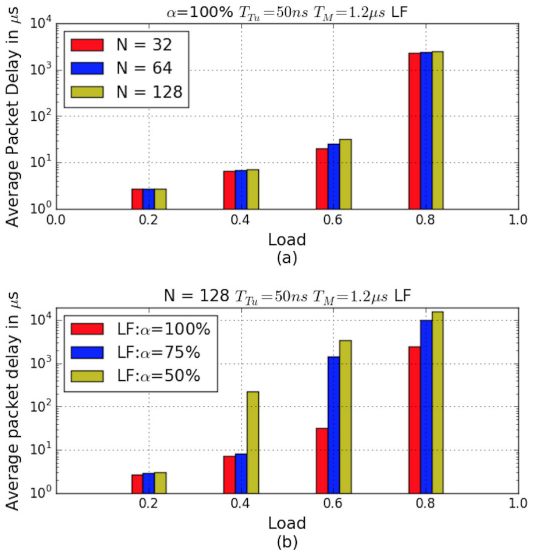


Fig. 6: Average packet delay with (a) different N; (b) different α.

(i.e., T_{Tu} = 50 ns) [29], and the maximum transmission time T_M for each cycle is set to 1.2 μs, allowing the server to transmit at most one packet with the maximum size of 1518 bytes. Fig. 5 (a) and (b) show the average packet delay and packet drop ratio for a rack with 64 servers (i.e., N = 64) and α = 100%, which means the total number of available wavelengths is $N\times(1+O_R)$ = 80).

It can be seen in Fig. 5(a) that when load is lower than 0.5, POTORI with the proposed LF DBA algorithm can achieve a

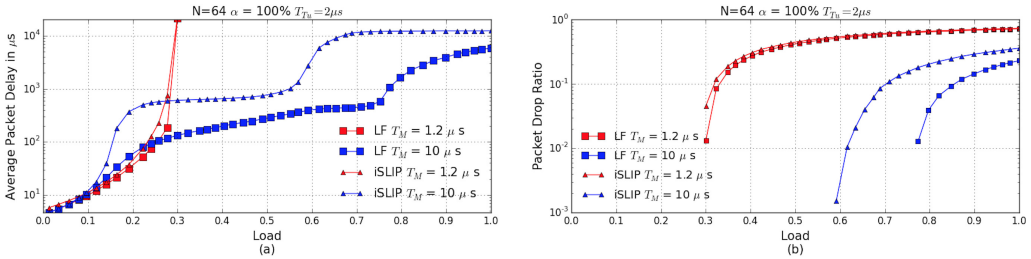


Fig. 7: Performance with T_{Tu} = 2 μs and different T_M: (a) average packet delay; (b) packet drop ratio.

packet delay lower than 10 μs, which performs similar as EPS. When the load is increased to 0.7, POTORI with LF is able to introduce up to 50% lower delays compared to EPS, thanks to the LF's feature of prioritizing the large traffic demand. Moreover, it performs slightly better (around 2% difference) than EPS in terms of the packet drop ratio. On the other hand, POTORI with the iSLIP DBA algorithms, has the worst performance. When the load is lower than 0.5, the average packet delay when employing iSLIP is double as high as that with LF. In addition, iSLIP shows highest packet drop ratio, which is greater than 10% when load is larger than 0.75.

In Figs. 5(d) and 5(e), we show the cumulative distribution function (CDF) of the packet delay at load = 0.1 and load = 0.7, which represent the cases with low load and high load, respectively. Under the load of 0.1, the difference between POTORI and EPS is negligible, and POTORI is able to achieve packet delay less than 10 μs for almost all packets (>99.99%) . Under the load of 0.7, there are 40% of the packets that are transmitted to the destination within 10 μs by POTORI with LF, and 80% of the packets have a delay lower than 100 μs, which outperforms POTORI with iSLIP and EPS.

C. Impact of Network Configuration

In this subsection, we investigate the impact of network configuration on performance and present the average packet delay as a function of *N* and α. The T_{Tu} and T_M are set the same in the previous subsection (i.e., T_{Tu} = 50ns and T_M = 1.2 μs). Because the proposed LF obviously outperforms iSLIP, we choose LF as the DBA algorithm for POTORI.

Fig. 6(a) shows the average packet delay of POTORI with different number of servers (i.e., *N* = 32, 64, and 128) in the rack with α = 100% (corresponding to 40, 80, and 160 available wavelengths, respectively). Under the low load (i.e., load = 0.2) and the heavy load (i.e., load = 0.8), the difference in the average packet delay is very small. It is because under the low load, the wavelength resources are sufficient and most of the traffic demands can be transmitted in one cycle regardless of *N*, while under high load the system gets statured, always resulting in high delay. For the medium load (i.e., load = 0.4 and 0.6), the packet delay performance degrades with the increase of the number of servers. This is due to the fact that the number of available wavelengths (which is equivalent to the maximum number of assigned traffic demands in each cycle) increases linearly with the number of servers, but the total number of traffic demands increases quadratically.

Fig. 6(b) shows the average packet delay with different values of α. We consider 128 servers per rack (i.e., *N* = 128), and

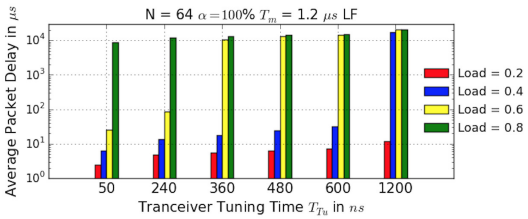


Fig. 8: Average packet delay with different T_{Tu}.

test α =100% (160 wavelengths), 75% (120 wavelengths), and 50% (80 wavelengths). The average packet delays are almost the same for all the considered values of α under the low load (i.e., load = 0.2), indicating that in this condition using less wavelengths in POTORI under the low load still maintains good packet delay performance (less than 10 μs). However, the packet delay with α = 50% increases dramatically under the higher load, while with α = 75%, the performance decreases significantly at load = 0.6.

D. Impact of Transceivers' Tuning Time T_{Tu}

In the previous subsections, we considered a tunable transceiver with an ultra-fast tuning time T_{Tu} = 50 *n*s and a maximum transmission time of T_M = 1.2 μs for each cycle. Note that with T_M = 1.2 μs, packet-level switching granularity is achieved by POTORI, since in each cycle there is at most one packet transmitted by a server. However, with the larger tuning time it may become challenging to efficiently realize the packet-level switching granularity. Thus, we relax the constraint of transceivers tuning time to 2 μs, which has been reported by the commercially available products [34]. With T_{Tu} = 2 μs, the performance of POTORI in terms of average packet delay and packet drop ratio with different DBA algorithms are shown in Fig 7. If keeping T_M = 1.2 μs, the packet delay of both DBA algorithms increase tremendously and the system has a high packet drop ratio, due to the increased tuning overhead. However, the performance can be much improved by increasing T_M. With T_M = 10 μs, the packet delay for POTORI with LF DBA algorithm employed can still be maintained below 100 μs when the load is lower than 0.3, and below 400 μs under the load up to 0.7. When employing iSLIP, on the other hand, POTORI performs worse, as the packet delay is around 600 μs under the medium load (load from 0.3 to 0.6). In addition, LF achieves lower packet drop ratio compared to iSLIP. Nevertheless, for both the LF

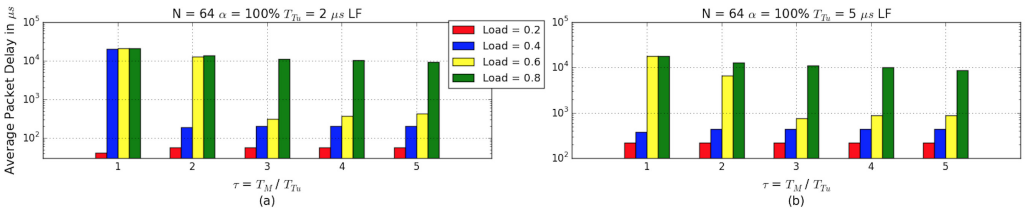


Fig. 9: Average packet delay with different ratio between T_M and T_{Tu}: (a) T_{Tu} = 2 μs; (b) T_{Tu} = 5 μs.

and iSLIP setting T_M = 10 μs outperforms the case with T_M = 1.2 μs. The reason for this is that with larger T_M, more packets can be sent by servers within one cycle, which reduces the tuning overhead.

One important question is "how to set the proper value of T_M for POTORI". Obviously, the proper value of T_M highly depends on the tuning time T_{Tu} of the transceiver. POTORI is able to achieve the packet-level switching granularity with the fast tuning transceivers. Fig. 8 shows that a moderate packet delay performance (i.e., $<$ 100 μs) when the load is lower than 0.6 can be achieved at the tuning times equal to 50 ns and 240 ns (corresponding to 4% and 20% of the maximum transmission time of one packet). The packet delay increases as we use transceivers with longer tuning time (i.e., 360ns, 480ns, 600ns, 1200ns, corresponding to 30%, 40% 50%, 100% of the maximum transmission time of one packet). We conclude that in order to obtain a good performance of packet delay (e.g., less than 100 μs under the load of 0.6) with packet-level switching granularity in POTORI, T_{Tu} should be less than 30% of maximum transmission time of one packet.

With a longer transceiver's tuning time, T_M should be increased to reduce the tuning overhead. In Fig. 9 we present the packet delay as a function of τ, which is defined as the ratio of T_M over T_{Tu}, given T_{Tu} = 2 μs and 5 μs, respectively. With a small τ (i.e., τ = 1 and 2), the packet delay is as high as $10^4\mu$s even under the medium load (i.e., load = 0.4 and 0.6) while obviously better packet delay performance can be achieved with a larger τ (i.e., τ = 3, 4, 5) under the same load. In addition, with τ = 5, the performance will decrease a little at load = 0.6. This is due to the fact that even if with larger τ, more traffic can be sent in one cycle, yet not all servers cannot fully utilize the cycle and may transmit much less traffic than is allowed in one cycle. It is caused by the bursty feature of the traffic generated by the servers, which may lead to quite different traffic demand per server in each cycle. The larger the T_M is, the larger can be the difference in traffic requested by the servers in one cycle. In order to use the resources more efficiently and achieve the acceptable delay performance, we conclude that given a tunable transceiver with long tuning time (i.e., in the scale of micro-seconds), the maximum transmission time of one cycle should be at least three times longer than the transceiver's tuning time.

VII. Conclusion

In this paper, we focus on POTORI, an efficient optical ToR interconnect architecture for DCs. The POTORI's data plane is based on a passive optical coupler, which interconnects the tunable transceivers of the servers within a rack. In the control plane, POTORI relies on a centralized rack controller, which is responsible for managing the intra-rack and inter-rack communications. A cycle based centralized MAC protocol and a DBA algorithm (*Largest First*) are proposed and tailored for POTORI, aiming to achieve the collision-free transmission with good network performance. POTORI can be applied in optical DCN with any aggregation/core tier architecture, given the proper design of interfaces.

The simulation results have shown that under the realistic data center traffic scenarios, POTORI with the proposed DBA algorithm obtains the average packet delay in the order of microseconds, which is superior to the performance of the conventional EPS. Moreover, we quantify the impact of network configurations (including the interconnect size and the number of available wavelengths) and transceiver tuning time on the packet delay. For POTORI, the packet-level switching granularity is feasible if the tuning time can be kept small enough (less than 30% of the packet transmission time). In the case of the short tuning time (i.e., in the magnitude of microseconds), setting the maximum transmission time of each cycle greater than three times of transceiver's tuning time is still able to achieve the acceptable packet delay performance.

Acknowledgment

This work was supported by the Swedish Foundation for Strategic Research (SSF), Swedish Research Council (VR), Göran Gustafssons Stiftelse, Celtic-Plus subproject SENDATE-EXTEND funded by Vinnova, Natural Science Foundation of Guangdong Province (Grant No. 508206351021) and National Natural Science Foundation of China (Grant No. 61550110240, 61671212).

References

[1] http://www.datacenterknowledge.com/
[2] Cisco Global Cloud Index: Forecast and Methodology, 2014-2019, October 2015, Cisco white paper.
[3] A. Andreyev, "Introducing data center fabric, the next-generation Facebook data center network." https://code.facebook.com/posts/360346274145943, 2014.
[4] FINISAR, "Cabling in the Data Center", https://zh.finisar.com/markets/data-center/cabling-data-center.
[5] Data Center Design Considerations with 40 GbE and 100 GbE, Aug. 2013, Dell white paper.
[6] N. Binkert *et al.*, "The role of optics in future high radix switch design," in *Proc. IEEE ISCA*, 2011, pp.437-447.
[7] R. Pries *et al.*,"Power consumption analysis of data center architectures," in *Green Communications and Networking*, 2012.
[8] M. Fiorani *et al.*, "Energy-efficient elastic optical interconnect architecture for data centers," in *IEEE Communications Letters*, vol.18, pp. 1531-1534, Sept. 2014.
[9] R. Lin *et al.*, "Experimental Validation of Scalability Improvement for Passive Optical Interconnect by Implementing Digital Equalization", European conference and exhibition on optical communication(ECOC), September 2016.
[10] G. Wan *et al.*, "c-through: part-time optics in data centers," in *Proc. ACM SIGCOMM Conf.*, 2010, pp. 327-338.
[11] M. Fiorani *et al.*, "Hybrid Optical Switching for Data Center Networks," in *Hindawi Journal of Electrical and Computer Engineering*, Vol. 2014, Article ID 139213, 13 pages, 2014.

[12] W. M. Mellette *et al.*, "A Scalable, Partially Configurable Optical Switch for Data Center Networks," in *IEEE/OSA Journal of Lightwave Technology*, vol. 35, no. 2, pp. 136-144, Jan. 2017.

[13] F. Yan *et al.*, "Novel Flat Data Center Network Architecture Based on Optical Switches With Fast Flow Control," in *IEEE Photonics Journal*, vol. 8, number 2, April 2016.

[14] M. Yuang *et al.*, "OPMDC: Architecture Design and Implementation of a New Optical Pyramid Data Center Network," in *IEEE/OSA Journal of Lightwave Technology*, vol. 33, issue 10, pages 2019-2031, May 2015.

[15] M. Fiorani *et al.*, "Optical spatial division multiplexing for ultra-high-capacity modular data centers," in *Proc. IEEE/OSA Opt. Fiber Commun. Conf.* 2016, Paper Tu2h.2

[16] V. Kamchevska *et al.*, "Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks," in *IEEE/OSA Journal of Lightwave Technology*, vol. 34, issue 8, April 2016.

[17] A. Roy *et al.*, "Inside the Social Network's (Datacenter) Network," in *Proc. ACM SIGCOMM Conf.*, 2015 pp. 123-237.

[18] Y. Cheng *et al.*, "Centralized Control Plane for Passive Optical Top-of-Rack Interconnects in Data Centers," in *Proc. IEEE GLOBECOM* 2016.

[19] N. McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches," in *IEEE/ACM Trans. on Networking*, vol. 7, no 2, pp.188-201, 1999.

[20] H. Yang *et al.*, "SUDOI: software defined networking for ubiquitous data center optical interconnection," IEEE Communications Magazine, vol. 54, no. 2, pp. 86-95, Feb. 2016.

[21] J. Chen *et al.*, "Optical Interconnects at Top of the Rack for Energy-Efficient Datacenters," in *IEEE Communications Magazine*, vol. 53, pp. 140-148, Aug. 2015.

[22] Y. Cheng *et al.*, "Reliable and Cost Efficient Passive Optical Interconnects for Data Centers," in *IEEE Communications Letters*, vol. 19, pp. 1913-1916, Nov. 2015.

[23] N. McKeown *et al.*, "OpenFlow: Enabling innovation in campus networks," in *ACM SIGCOMM Computer Communication*, Review 38, April 2008.

[24] Software-Defined Networking: The New Norm for Networks, Open Networking Foundation (ONF) White Paper, April 2012.

[25] "802.3-2012 - IEEE Standard for Ethernet".

[26] W. Ni *et al.*, "POXN: a new passive optical cross-connection network for low cost power efficient datacenters," in *IEEE/OSA Journal of Lightwave Technology*, vol. 32, pp. 1482-1500, Apr. 2014.

[27] "IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications".

[28] L. Khermosh, "Managed Objects of Ethernet Passive Optical Networks (EPON)," RFC 4837, July 2007.

[29] B. Vattikonda *et al.*, "Practical TDMA for Datacenter Ethernet," in *Proc. ACM EuroSys. Conf.*, pp. 225-238, 2012.

[30] G. Poter *et al.*, "Integrating microsecond circuit switching into the data center," in *Proc. ACM SIGCOMM Conf.*, 2013 pp. 447-458.

[31] T. Javadi *et al.*, "A high-Throughput Algorithm for Buffered Crossbar Switch Fabric," in *Proceedings IEEE ICC*, pp. 1581- 1591, June 2001.

[32] S. Kandula *et al.*, "The Nature of Datacenter Traffic: Measurement &Analysis," in *Proc. ACM SIGCOMM Internet Eas. Conf.*, 2009, pp. 202-208.

[33] S. Matsuo *et al.*, "Microring-resonator-based widely tunable lasers," in *IEEE J. Sel. Topics Quantum Electron*, vol. 15, no. 3, pp. 545-554, 2009.

[34] Keysight Technology: 81960A Fast Swept Compact Tunable Laser Source, 1505nm to 1630nm
http://www.keysight.com/en/pd-2038953-pn-81960A/fast-swept-compact-tunable-laser-source-1505nm-to-1630nm-new?cc=SE&lc=eng

Yuxin Cheng received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012 and M. Eng. degree in network services and systems from KTH Royal Institute of Technology, Sweden, in 2015. He is currently a Ph.D. candidate with Optical Networks Lab (ONLab), KTH Royal Institute of Technology, Sweden. His research interests include optical data center network and software defined networking.

Matteo Fiorani received the Ph.D. degree (February 2014) in ICT from University of Modena (Italy). From February 2014 to October 2016, he was Postdoc Researcher in optical networks at KTH (Sweden). He was visiting researcher at TU Vienna (Austria), UC Davis (USA) and Columbia University (USA). Since October 2016, he works as Senior Researcher in 5G networks at Ericsson (Sweden). He co-authored more than 50 papers published in leading technical journals and conference proceedings. He also submitted several patents and standardization contributions on 5G networks. He was co-founder and chair of an IEEE workshop on 5G transport networks.

Rui Lin received the B.Sc. degree in Electrical Information from Huazhong University of Science and Technology (HUST), and the Ph.D. degree in Communication System from KTH Royal Institute of Technology, Sweden in 2016. Her research interests include short-reach optical interconnect and datacenter network.

Lena Wosinska received her Ph.D. degree in Photonics and Docent degree in Optical Networking from KTH Royal Institute of Technology, Sweden, where she is currently a Full Professor of Telecommunication in the School of Information and Communication Technology (ICT). She is founder and leader of the Optical Networks Lab (ONLab). She has been working in several EU projects and coordinating a number of national and international research projects.

Her research interests include fiber access and 5G transport networks, energy efficient optical networks, photonics in switching, optical network control, reliability and survivability, and optical datacenter networks. She has been involved in many professional activities including guest editorship of IEEE, OSA, Elsevier and Springer journals, serving as General Chair and Co-Chair of several IEEE, OSA and SPIE conferences and workshops, serving in TPC of many conferences, as well as being reviewer for scientific journals and project proposals. She has been an Associate Editor of OSA Journal of Optical Networking and IEEE/OSA Journal of Optical Communications and Networking. Currently she is serving on the Editorial Board of Springer Photonic Networks Communication Journal and of Wiley Transactions on Emerging Telecommunications Technologies.

Jiajia Chen received the Ph.D. degree (May 2009) in Microelectronics and Applied Physics, and the docent degree (February 2015) in Optical Networking from KTH Royal Institute of Technology, Sweden. Her main research interests are optical interconnect and transport networks supporting future mobile system and cloud environment. She has more than 100 papers published in the leading international journals and conferences. She is Principle Investigator of several national research projects funded by Swedish Foundation of Strategic Research (SSF), Gran Gustafssons Foundation and Swedish Research Council (VR). Meanwhile, she has been involved and taken the leadership in various European research projects, including the European FP7 projects IP-OASE (Integrated Project-Optical Access Seamless Evolution) and IP-DISCUS (Integrated Project-the DIStributed Core for unlimited bandwidth supply for all Users and Services), and several EIT-ICT projects (e.g., Mobile backhaul, M2M and Xhaul).