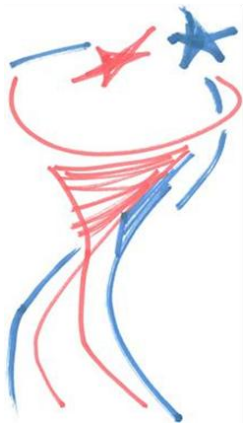
ICT - Information and Communication Technologies

SEVENTH FRAMEWORK PROGRAMME



PhoxTroT

Photonics for High-Performance, Low-Cost & Low-Energy Data Centers, High Performance Computing Systems: Terabit/s Optical Interconnect Technologies for On-Board, Board-to-Board, Rack-to-Rack Data Links

Collaborative Project
Grant Agreement Number: 318240

Optical Packet router architectures for HPC and Data Centers
D11.1

Deliverable number:	D11.1	**Work package number:**	11
Due date of deliverable:	01.06.2014 (M20)	**Actual submission date:**	07.07.2014
Start date of the project:	01.10.2012 (M01)	**Duration:**	48 months
Nature:	Report	**Dissemination level:**	PU

Lead beneficiary: DAS
Contact person: Andreas Hakansson
Address: DAS Photonics, Camino de Vera, s/n Ed. 8F, 46022 Valencia, Spain
Phone: +34 61 753 74 27
Email: ahakansson@dasphotonics.com
Author(s): Nikos Pleros, Dimitris Apostolopoulos, Kostas Christodoulopoulos, Manos Varvarigos, Pavlos Maniotis, Stella Markou, Siokis Apostolos
Contributing beneficiaries: CTI, ICCS/NTUA, XR, CERTH/ITI, TE

Abstract:
This document gives an overview of the state-of-the-art and the challenges related to data center architectures. It is then shown how optical packet switch and transport (OPST) in combination with advanced modulation formats can improve the communication. Specifically, the technology developed and used in PhoxTroT is envisaged.

Keywords: Data Center, Advanced Modualtion, HPC, POADM

Project Information

PROJECT	
Project name:	Photonics for High-Performance, Low-Cost & Low-Energy Data Centers, High Performance Computing Systems: Terabit/s Optical Interconnect Technologies for On-Board, Board-to-Board, Rack-to-Rack data links
Project acronym:	PhoxTroT
Project start date:	01.10.2012
Project duration:	48 months
Contract number:	318240
Project coordinator:	Dr. Tolga Tekin - Fraunhofer
Instrument:	Large-scale integrating project - CP-IP
Activity:	ICT-8-3.5 - Core and disruptive photonic technologies
DOCUMENT	
Document title:	PhoxTroT – D11 1 Optical Packet router architectures for HPC and Data Centers.doc
Document nature:	Report
Deliverable number:	D11.1
Due date of delivery:	01.06.14 (M20)
Calendar date of delivery:	07.07.14
Editor:	Andreas Hakansson
Author(s):	Nikos Pleros, Dimitris Apostolopoulos, Kostas Christodoulopoulos, Manos Varvarigos, ppmaniot@csd.auth.gr, Stella Markou, Siokis Apostolos
Lead beneficiary:	DAS
Contributing beneficiaries:	CTI, ICCS/NTUA, XR, CERTH/ITI, TE
Dissemination level:	PU
Work package number:	11
Work package title:	Data Center Interconnect Platforms
Date created:	
Updated:	
Version:	v1
Total number of pages:	34
Document status:	

PU = Public ; PP = Restricted to other programme participants (including the Commission Services) ; RE = Restricted to a group specified by the consortium (including the Commission Services) ; CO = Confidential, only for members of the consortium (including the Commission Services)

Table of Contents

1 Executive Summary

This document gives an overview of the state-of-the-art and the challenges related to data center architectures. It is then shown how optical packet switch and transport (OPST) in combination with advanced modulation formats can improve the communication. Specifically, the technology developed and used in PhoxTroT is envisaged.

2 Introduction

The objective of this deliverable is to investigate Advanced Modulation formatted traffic at inter-rack communication level.

2.1 Document structure

The present deliverable is split into five major chapters:

- State of the art in Data center architectures
- PhxTroT OPST Architectures
- Main OPST node Building Blocks
- Ring-Star Architecture
- Ring-Ring Archtiecture
- Overview of the critical parameters
- PhoxTroT technology in HPC environment

2.2 Audience

This document is public.

3 State of the art in Data center architectures

The proliferation of the cloud application-, platform- and infrastructure-as-a-service models is motivating the construction of new and more powerful datacenters [1]. This is raising the bar in communication requirements not only among the cloud datacenters, but also within them.

Today's datacenters are typically designed with a fat-tree or oversubscribed fat-tree interconnection topology. Two approaches are followed in fat-tree networks: (i) the traditional fat-tree approach using higher rate ports towards the root of the tree, or (ii) using low-port-rate commodity switches in a folded Clos topology so as to provide low-rate multiple paths between the endhosts (servers) [2]. The former is considered impractical for big datacenters since supporting many endhosts requires huge bandwidth ports at the root of the tree, while the cost increases vastly as we move from commodity to high-end equipment. Moreover, a network built in this way cannot be expanded to support more endhosts and suffers from single point of failure problems. Thus, most popular in real-life datacenters is either the latter (Clos) approach, which restrains costs by using many but cheap commodity equipment at the higher interconnection levels and employs multipaths which solve the link, port and switch failure problems, or a combination of the two (a Clos topology of higher rate switches than the server rate). Figure 1 shows a fat-tree network built out of commodity low-rate switches in a Clos topology.

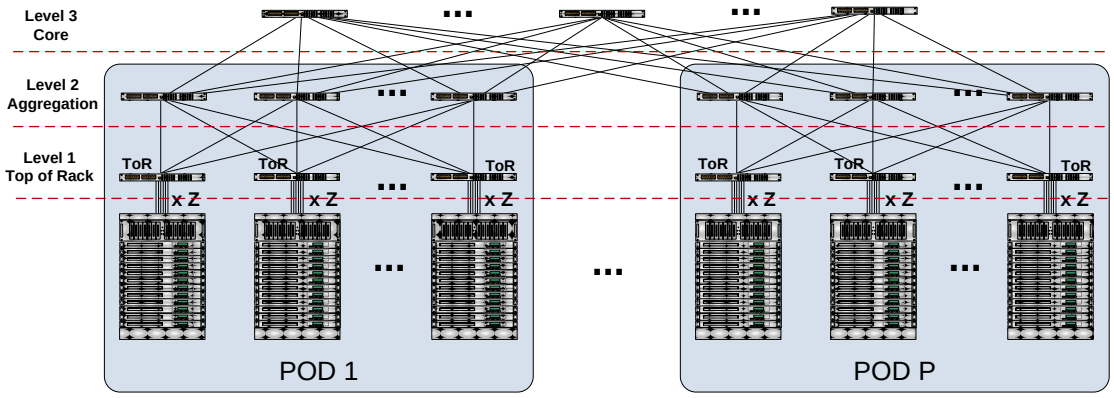


Figure 1: Fat-tree network built out of commodity low-rate switches in a Clos topology

Today's datacenters contain tens to hundreds of thousands servers, placed in racks of typically between 20 and 40 servers. A number of server racks can be logically or physically grouped into pods; logically meaning that they belong to the same subnetwork and are interconnected with the same pod switches (second level of the tree) while physically means that they are placed in the same room-container. Companies like HP, IBM, Sun, etc. pack servers and networking equipment into shipping containers that can be used as a standalone small datacenter or can be connected as a module of a larger datacenter. The number of racks per physical pod is typically between 40 and 60, but even denser values can be found.

Each rack is typically equipped with a Top-of-Rack switch (ToR) that provides southbound interfaces to the rack's servers and northbound interfaces to the higher levels of the tree. Assuming *N* hosts in the datacenter and *k*-radix packet switches, the depth of the fat-tree is ——— . It follows that scaling the number of endhosts in this architecture comes with the requirement to install an additional fat-tree level once a number of servers is reached. Today datacenters have tens to hundreds of thousands of servers and require 3 to 4 fat-tree levels to achieve full bisection bandwidth. Hence, although the cost of Clos is lower than the traditional fat-tree approach, it still scales super-linearly. Moreover, the cabling of the huge number of switches becomes quite cumbersome and is error prone during installation and maintenance. The use of a large number of active packet switches contributes hugely to the energy consumption of the whole system (64-port 10GbE switches consume 150 to 350 Watts). Note that roughly 90% of this energy consumption is independent of the load and thus savings are impossible to come from any load balancing/scheduling method. Finally, upgradability is a big issue with fat-tree architectures: (i) adding more racks-servers requires connecting free ports of several and scattered switches, assuming that free ports are available, i.e. we have not reached the server limit for the given tree depth, but once the limit is reached then a huge number of new switches is required to populate the new level, (ii) upgrading the communication rate of the servers requires a complete new fat-tree network, and in most cases the previous switches cannot be reused.

Fat-tree networks are under-utilized most of the time [3]. As a result full bisection bandwidth is not needed, since it seldom happens that all servers talk simultaneously at full speed. This is partially resolved by building an over-subscribed fat tree network i.e. a tree with less bandwidth at higher levels, which does not provide full bisection bandwidth but is cheaper. Even oversubscribed trees are utilized lightly on average: [3] reports less than 20% average utilization for real oversubscribed tree systems, but there are certain instants where congestion in the form of hotspots is created and packets are lost. The problem is

the rigid allocation of the available (reduced) bandwidth to the servers that cannot facilitate re-allocation of the hardware resources among endhosts according to their network needs.

Latency depends on the load of the system, since congestion adds queuing delays and might lead to packet drops that are handled by the higher layers (TCP) or the application. Even when lossless operation is guaranteed (e.g. using Infiniband or Ethernet's extension Datacentre Bridging), congestion still increases latency. We can however measure the longest path latency of an empty (zero load) network, where queuing delay is zero and also propagation delay is neglected. Typical figures of Ethernet switch latencies are between 500 nsec and 1.5 μsec. At light network load the order of latency is hundreds of μsec to msec.

In short, the main limitations of current datacenter architectures are:

- Super linear scaling
- High energy consumption independent of load
- Cable spaghetti
- Huge waste of capex and opex since network is underutilized on average
    - full fat-tree is very expensive
    - oversubscribed fat-tree (still underutilized on average) can exhibit congestion (hotspot) problems
- Rigid bandwidth allocation
- Large latencies for east-west traffic which is dominant (need to travel north-south)

3.1 *The challenge of disaggregation*

These limitations are further exacerbated by the emerging concept of Resource Disaggregation which is rapidly gaining momentum. So far, in the conventional datacenter model, the computing, memory, storage and communication resources are fixed for the servers which compose the datacenters. In fact, each server consists of a fixed combination of computing, memory, storage and communication resources all incorporated, or "aggregated", in the same enclosure. The basic idea of disaggregation is to share these resources among the datacenter's racks and use them on-demand.

There are multiple benefits from the transition to disaggregated datacenters. Modularity of the infrastructure leads to more efficient operation and improved performance. This physical decoupling of resources allows for more fine-grained resource provisioning as well as higher utilization with statistical multiplexing of the available resources [16]. Furthermore, this modularity also enables independent evolution, as the individual types of resources follow different trends and constraints. From this point of view the operators are able to adopt easily the state-of-the-art in any particular module independently of the other resources, e.g. upgrading to a new generation of storage system leaves all compute, memory and network elements unaffected and thus save unnecessary costs. In addition, the vendors are more flexible to develop innovative components.

The concept of disaggregation is championed by Intel and Facebook, the main motivators behind the Open Compute Project (OCP) that aims to build a broad industrial consensus. The proposed architecture, called Disaggregated Rack-Scale Server (DRS), is optimized for large address space, in-memory databases and analytics. The OCP DRS goal is the separation of computing, storage and communication hardware components within the rack and the interconnection between them with distributed switching functions [2]. OCP anticipates a 24% reduction of costs and an efficiency increase of about 38% with this new disaggregated rack paradigm [4]. Although the disaggregation concept is still at its infancy and only recently were the first products showcased [5][6], expectations are that it will rapidly become mainstream: According to Mark Roenigk, COO of multi-billion IT hosting company Rackspace, within three years the level of adoption will be "between 35 and 50 per cent of new installations of servers" [7].

Despite the unique benefits of disaggregation in terms of virtualization, optimum resource provisioning and infrastructure upgradeability, it comes with a major hurdle: Disaggregating the system resources causes inordinate requirements to the network interconnecting them. As a result, network interconnects are facing the challenge to meet the skyrocketing demands for high bandwidth and low latency across the physical distance of the distributed components into the datacenter. Current implementations based on conventional fat-trees and present-generation switches simply cannot scale to support these traffic demands. For the high-bandwidth transport of data, silicon photonics is advocated as an enabling technology and development actions are underway globally [7]. As far as efficient switching and routing of this vast amount of information is concerned, photonics appears again as the enabling technology. Using the optical layer not only as a forwarding plane but also for switching can offload the processing burden from power-hungry electronic switches and also enable flatter network architectures. Much like the recent paradigm of optical telecom networks, the concept is not to replace electronics with photonics rather than to use them synergistically, so as to take the best of both worlds. This notion has fuelled a number of research actions on hybrid electro-optical networks that are considered as a viable migration path in order to meet the rising networking demands in the datacenter.

3.2 *Optical switching datacenter network architectures*

Optical switching has been investigated for transferring aggregated traffic between racks or collections of racks, partly or entirely replacing the higher levels of the electronic tree networks [8-10]. These proposals leverage interesting features exhibited by optical switches, such as protocol- and rate-transparency, relatively low cost per port, and reconfigurability. Reconfiguration is crucial since it can improve the efficiency and reduce the cost of the network by providing a hardware bandwidth-on-demand interconnect, avoiding the over-provisioning and rigid bandwidth allocation of (oversubscribed or not) fat-tree networks. To ensure a realistic deployment scenario, commercial off-the-shelf (COTS) optical switches have been mainly investigated such as MEMS, tunable lasers and AWGs.

A notable class of optical-switched datacenter networks relies on commodity Micro Electro-Mechanical Systems (MEMS) switches owing to the matureness of the technology. MEMS switches have long reconfiguration times that typically range in the order of tens to hundreds of milliseconds, making them suitable for circuit-switched flows. To

accommodate packet-switched traffic as well, the optical network in these implementations is used in tandem with an electrical packet-switched network. Thus, long-lived flows are handled by the optical switched network whereas short, bursty traffic is accommodated by the electrical packet switched network. This MEMS-based hybrid electronic-optical network approach is followed by Helios [8], Calient [11], and REACToR [12]. The optical circuit switched fabric provides essentially unlimited bandwidth that scales without the need for equipment upgrades as network speeds increase. On the other hand, a considerable portion of the traffic in the datacenter is short-lived and cannot reap the benefits of the optical-switched network. In addition, the requirement for classification is quite demanding since it involves traffic monitoring-prediction over the extremely large network. The control plane that serves to handle the network reconfiguration adds considerable delays [13], which can go up to the seconds' timescale. One of the control bottlenecks is that reconfiguration, apart from the MEMS switch, involves informing end-host or switches of how to split the traffic into circuits. Finally, since the radix of MEMS switches is quite limited (up to 320 port switches are commercially available) and building higher-port switches out of smaller ones is very complex (due to losses and synchronization issues) the hybrid solution based on MEMS exhibits scalability problems.

In the attempt to reduce the optical network reconfiguration times, wavelength-switching concepts have been considered as a viable optical packet switched option. Such concepts rely on the fast tuning capabilities of tunable lasers which, followed by wavelength-selective elements like Arrayed Waveguide Grating Routers (AWGRs), can route the incoming optical data to the respective output of the AWGR according to their wavelength. Wavelength-switched implementations with AWGRs and other passive optical devices are tolerant to crosstalk and introduce no further impairments due to the switching operation [14]. A number of initiatives has investigated this concept in different realizations, such as DOS LIONS [15][16], Petabit [17] and IRIS [18]. Despite the speed, modularity and low contention benefits that the aforementioned architectures offer to a datacenter network, they face significant scalability issues. The total number of wavelengths for the optical links is limited (approximately 80 in the C-band), which hinders possible extension of the network to large-scale datacenter networks.

Another hybrid datacenter architectural approach that aims to mitigate the long reconfiguration times of optical MEMS switches is developed in Mordia [19]. Mordia introduces the use of WDM transmission with fast Wavelength Selective Switches (WSSs) for network reconfiguration. The Mordia WSSs achieve switching times in the order of 10 μs which is substantially faster than the speed of MEMS but still not compatible with Ethernet-packet granularity. In addition to the architecture and physical layer implementation, Mordia has also researched algorithms to ensure fast reconfiguration of the underlying network infrastructure. Although the proposed algorithms are 2-3 orders of magnitude faster than traditional approaches, they still cannot scale to huge datacenters of thousands of racks. The main limitations of Mordia remain its scalability and cost. Both the architecture and the control plane functionalities have not been validated in a realistic sized datacenter, whereas the cost of the expensive WSS components is shared among a very small number of endhosts (4 in the project's demonstrations and generally limited by the available ports of a WSS, which typically is not above 9) [20].

Finally, as an alternative to the hybrid electro-optical approach and in an attempt to shed the electrical switches completely from the datacenter, the Lightness project [21] is developing a datacenter network that can handle both packet and circuit flows in the optical domain. An optical technology-based data plane is adopted, which relies on a

flattened architecture integrating both optical packet switching (OPS) and optical circuit switching (OCS) technologies. OPS switches are used for handling short-lived flows with low latency requirements, while OCS switches are for long-lived traffic flows. Servers are interconnected to the hybrid OPS/OCS datacenter networks via electrical TOR switches, which aggregate traffic and also classify short-lived and long-lived traffic. The TOR switch is electrical (FPGA-implemented) and designed to support the hybrid OCS/OPS with 100 Gb/s capacity and low latency. The OPS switch provides WDM operation and is based on WSSs implemented with AWGs followed by large SOA arrays in order to perform the OPS between different ToR switches. This SOA-based switching concept has been well-investigated in the past in the context of OPS-telecom networks [22], however it still involves significant hurdles, such as the relatively high power consumption and power dissipation of the switches, as well as the lack of an established supply chain (switching components are not commercialized and validated in any operating environment). The OCS is realized through Architecture on Demand (AoD) nodes where an optical backplane of large port count MEMS is connected to several signal processing modules as well as to the input/outputs of the node offering flexibility as the components are not hardwired but can be interconnected together in an arbitrary manner. The latter provides additional network services where required, but also results in redundant optical equipment that increases overall equipment cost and management overheads. Finally, since MEMS switches do not scale, scalability is also a huge problem in this AoD-OCS concept.

3.3 *Metro/telecom optical network architectures in the data center*

The Packet Optical Add-Drop Multiplexing (POADM) architecture, also called Ecoframe, [23-26] for metropolitan networks, is based on a unidirectional ring topology, Wavelength Division Multiplexing (WDM) and fixed length packets (slotted operation). The POADM switch can insert a packet at different wavelengths, using fast tunable lasers, and each switch is equipped with one or several such tunable lasers. As seen in Figure 2, each POADM switch listens to specific wavelengths (fixed wavelength receivers), and fast 1x2 switches are used to drop the packets at the destination. All optical channels on the WDM ring are divided into time slots of equal duration on a synchronous basis. The time slot duration is in the order of μsec. An adaptation layer is responsible for adapting the size of the upper layer packets into the fixed-size packets, by means of concatenation and possibly fragmentation. The equipment at the nodes constraint the number of packets that can be added or dropped in each time slot. Wavelength contention and packet collisions are avoided by a control protocol running on a separate wavelength channel. The control channel is synchronized with the data channels and carries information relative to each data packet carried in each time slot. The control protocol collects and disseminates the information regarding the occupancy of each time slot on each wavelength. Based on such information, each node can identify the packets to receive and drop them. Once freed, the time slot can be re-used for transmitting a locally generated packet to any ring node receiving on such wavelength. The control plane for virtual circuit allocation in Ecoframe ring is outlined in [24]. A hub node is also introduced there, which interconnects several rings and plays the role of the centralized scheduler regarding resources. So in that approach the hub implements a central reservation mechanism based on the requests it collected from the nodes located on the same ring. If the resources are available, it allocates the requested time slots and informs each node about the reserved time slots.

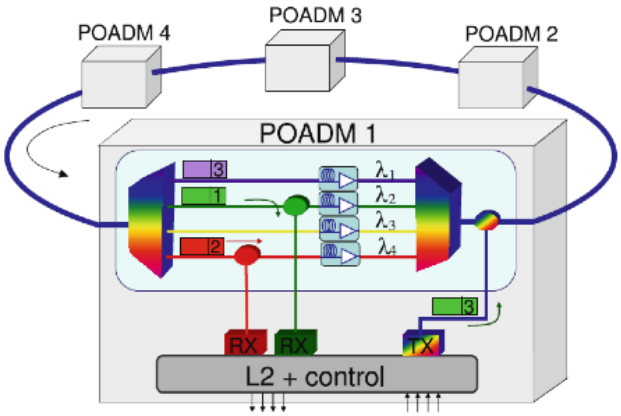


Figure 2. The Packet Optical Add-Drop Multiplexing (POADM) –Ecoframe- switch architecture.

Optical Packet Switch and Transport (OPST) [27][28], also used for metro networks, is a networking paradigm that collapses network layers 1 to 2 under the same control plane using tunable lasers. The tunable transmitters perform two traditional functions (transmission of light and layer 2-switching of logical packet flows) in the same device (Figure 3). The control plane runs internally inside a ring network of OPST devices (nodes). Each node contains a fixed wavelength filter which is the wavelength address of the ports of the system. That is, each node listens to a specific wavelength. A scheduler forms bursts from the VoQ (Virtual Output Queues) of packets and sends it using tunable laser transmitter whose wavelength is tuned to the wavelength of the destination. Then the burst of packet is sent out on the ring. The OPST system (contrary to the Ecoframe approach) is based on a distributed scheduling system that ensures fair access onto the ring. More specifically, an Optical Media Access Control system (MAC) employing Carrier Sense Media Access with Collision Avoidance (CSMA-CA) is used. This asynchronous access system avoids the need for ring-wide synchronisation. Note that detailed information on the MAC, the tunable lasers and the switches architectures is not widely available, since they comprise proprietary information.

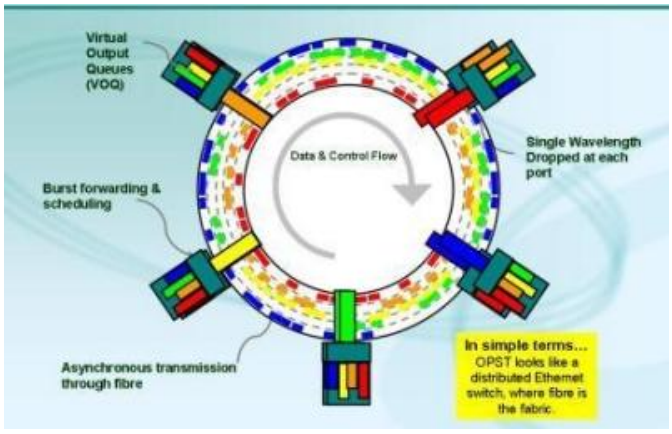


Figure 3. The Optical Packet Switch and Transport (OPST) architecture.

Although the POADM and OPST approaches seem very promising for metro telecom networks there are certain issues that need to be considered when such architecture is used in a datacenter network environment.

Capacity bottleneck: Both the aforementioned architectures are based on a single WDM ring. Assuming 80 wavelengths, and e.g. 40 Gbps per wavelength, the total capacity is 3.2

Tbps. Considering a rack in a datacenter with 30 servers and 10 Gbps interfaces per server, a rack produces about 300 Gbps. Even if we consider oversubscription, so that each rack gets a fraction the total capacity that it produces to communicate with other racks, it turns out that a WDM ring provides enough capacity for a few tens of racks. If the ring is not directly connected to the racks, but is higher in the hierarchy of the interconnect (e.g. between pods-clusters of racks), the limited capacity problem is still present. A solution to such problem is to employ multi-fiber rings, the number of which matches the capacity requirements of the racks or pods. Dividing the traffic in these multi-fiber rings is not straightforward, and extensions to the control plane approaches of the POADM and OPST are needed.

Ring dimension: Both aforementioned architectures target metro networks where the number of end-points is low, in the order of tens of nodes or even lower. Adding a huge number of end-nodes, apart from capacity issues (see the previous comment), would create problems with respect to the quality of the optical signal (transparently bypassing nodes yields loss, but also contributes to various physical layer impairments, such as dispersion, interference, etc) and in the bandwidth allocation process (the proposed control plane solutions do not scale to high number of nodes). Since the number of racks in a datacenter is much higher, we need to examine bridges or hub nodes that connect two or more rings together. Such bridges are mentioned in both POADM and OPST, but there is no detailed-specific architecture provided, nor has any evaluation been performed. Based on the number of endpoints supported by the metro architectures, using such architecture at the higher levels of the datacenter hierarchy, to e.g. interconnect pods-clusters of racks together, seems more viable.

Subsytems requirements: The subsystems used to build the POADM and OPST nodes, include tunable lasers, fast 1x2 space switches and mux and demux modules, or fast 1x2 space and wavelength switches (such as wavelength selective switches – WSS). Based on the subsystems developed in Phox-trot, alternative solutions for implementing tunable lasers have to be found, such as using a bank of WDM VCSELS and choosing the transmitting one. The approach of fast 1x2 space switches and mux and demux modules, and not the approach of using WSSs, fits well with the subsystems developed in Phox-trot.
As a conclusion, from the literature review of the optical packet switching architectures for metro ring telecom networks, it seems that such architectures are applicable in a datacenter interconnect environment. Based on basic dimensioning calculations, it seems that in a datacenter environment multiple rings have to be used, while rings are better suited at the higher levels of the hierarchy of the network, e.g. to interconnect pods-clusters of racks. Phox-trot components such as a bank of WDM VCSELS, fast 1x2 space switches, mux and demux modules are required for such network.

4 PhoxTroT OPST Architectures

We investigate new approaches which will be able to support optical packet switch and transport (OPST) architectures in combination with wavelength division multiplexing (WDM) techniques and advanced modulation formats (e.g. PAM-4, PAM-8, 16QAM). The WDM techniques, which have already been employed in OPST architectures, have the capability to increase the offered bandwidth by parallel transmissions of WDM data packets. However exploiting WDM data packets only, implies that packets are always processed at every node of the network, regardless if they want to bypass a certain ring-

node and continue propagating within the backbone network in order to reach the desired destination node. The aggregate traffic processing increases the power consumption, leading us to design efficient networks with reduced power consumption and low cost. Following this roadmap, we propose two architectures, namely the Ring-Star Architecture (RSA) and the Ring-Ring Architecture (RRA), based on a WDM optical ring network that also supports packet switching. In both architectures, the WDM packet traffic can bypass transparently the nodes until reaching its final destination. This suggests that the packets will be processed only by the destined node. In order to achieve that, a Packet-Optical Add/Drop Multiplexer (POADM) has been introduced to drop the packets to each destined node without implementing any electronic processing at the WDM traffic. The POADM incorporation in both of the aforementioned OPST architectures exploits very useful benefits that are presented below:

1. The POADM technology deflects the aggregate traffic recirculating within the backbone network in the optical domain by offering optical transparency. The employment of optical transparency overcomes the speed limitation of electronics, allowing the transportation of huge amounts of data at high transmission rates.
2. Moreover, exploiting optical transparency without any electronic processing of the data traffic facilitates reduced size of the ring nodes by simplifying the hardware. Minimizing the use of electronic components inside the ring nodes can thus be translated to a subsequent cost and energy reduction.
3. One additional advantage is that the OPST technology handles the optical traffic at a fine granularity on a packet-by-packet basis. The packet granularity provides flexibility and efficient use of the bandwidth through the combination of time division and wavelength division multiplexing techniques.

Figure 4 illustrates the backbone communication network for both the RSA and RRA, which consists of a number of nodes connected through an optical fiber in a ring topology. Each ring node connects a pod network to the backbone network and processes only the data packets that will be dropped in to the pod or will be added from the pod to the backbone traffic. The non-dropped packets transparently pass the ring nodes and continue propagating within the backbone network until reaching their destined nodes. Both the RSA and RRA employ WDM data traffic that incorporates the data wavelengths and the control channel and can also support multi level modulation formats. This in turn indicates that each node can transmit data packets on any wavelength and receive data packets on multiple wavelengths, shared by other nodes. Each packet has a fixed duration and each time slot can be allocated to one fixed-size optical data packet per wavelength. The control channel is also a packet-synchronous channel that transports the headers of the data packets and is divided in time slots too. Each time slot of the control channel is dedicated to one of the data packets that are transported on the wavelengths of the WDM optical ring.

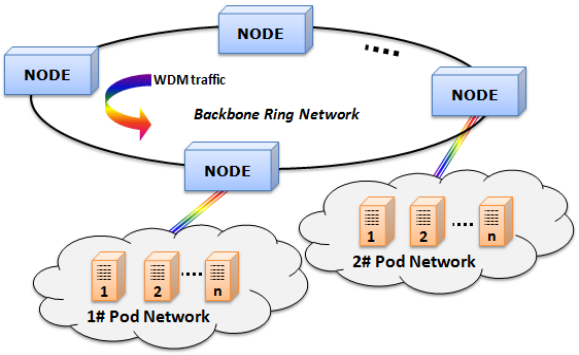


Figure 4: The backbone WDM optical ring network.

Moreover the ring nodes can add new data packets to the empty time slots of the backbone traffic. Each node can recognize if any packets have been dropped thus identifying the empty time slots and the corresponding wavelengths, in order to transmit new data packets within the WDM optical ring. Figure 5 presents step by step the logical procedure and the way packets are added to the WDM optical ring. For simplicity step 1 depicts an example with only three wavelengths, namely blue, green and red, constituting the WDM packet data traffic recirculating within the backbone ring network. The packets depicted in transparent blue and green color represent the empty slots, whereas the red outlined packet is the packet that will be dropped to the pod.

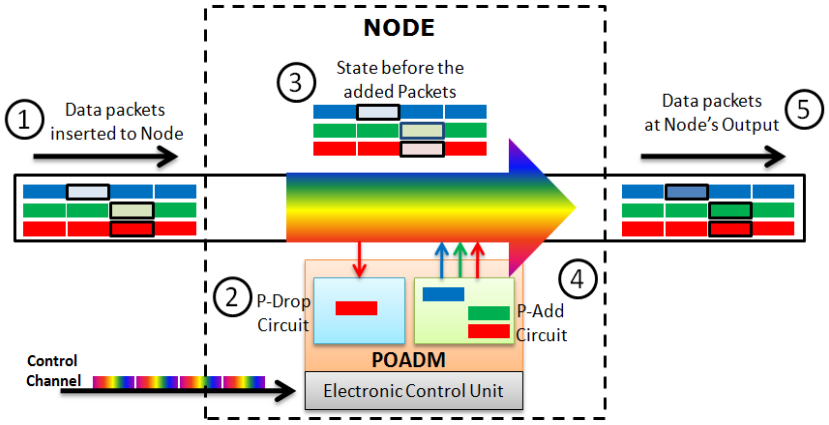


Figure 5: A schematic representation of the step-by-step Drop/Add mechanism at the Node

In step 2, the control channel is inserted to the Electronic Control Unit (ECU) which in turn drives the P-Drop Circuit in order to drop the red packet to the pod and release the time slot. Once the time slot is released, it can be re-used by the node in order to transmit a locally generated packet to the backbone network. Step 3 illustrates the new traffic state inside the node that stems from step 2, including the three available empty time slots, one per wavelength. The P-Add circuit now can add data packets at the slots of each wavelength and embed them to the remaining non-dropped data traffic, as depicted in step 4. Finally step 5 presents the output data traffic emerging from the node, as a combination of the recirculating non-dropped packets and the new generated packets added by the node. The latter are marked again with a black outline.

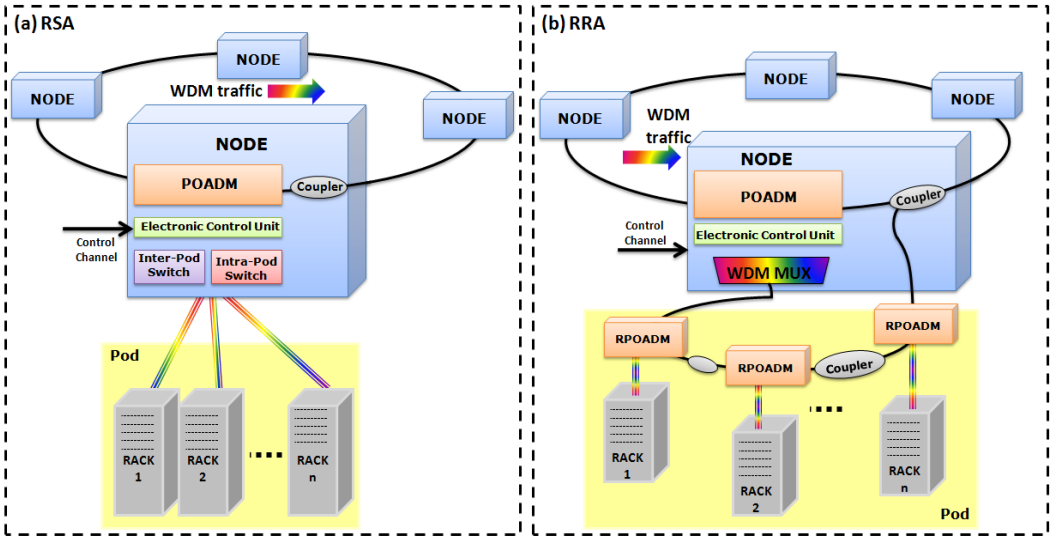


Figure 6: The proposed WDM optical ring - packet switching architectures: (a) The Ring Star Architecture (RSA) and (b) The Ring Ring Architecture (RRA)

Figure 6 (a) and (b) illustrate RSA and the RRA node-architecture respectively, proposed within PhoxTrot, with both architectures relying on the same backbone WDM optical ring network. The RSA ring nodes comprise of a POADM optical interface, an ECU, the Inter-Pod switch and the Intra-Pod switch. Each ring node is connected via a star topology with a pod and handles the drop/add traffic to and from the pod through the POADM. The decision about which packets will be dropped to the pod or will be added to the network is carried out through the ECU, which drives the POADM according to the control channel information. Then the ECU drives the Inter-Pod-switch in order to send the dropped packets to the corresponding destination rack within the pod and notifies when the rack can add a new packet and on which wavelength in to the WDM optical ring traffic. The local communication within the pod is carried out through the Intra-Pod switch which is based on the same technology with the Inter-Pod-switch and will be discussed in details later in this report.

In case of the RRA, shown in figure 6 (b), each ring node is connected through a ring topology with a pod, where each rack has one dedicated POADM optical interface. The RRA ring nodes consist of a POADM, an ECU, which controls which packets will be dropped to the pod, and a WDM multiplexer. The dropped packets are multiplexed though a WDM multiplexer and are fed to the optical ring of the pod though a fiber. The dropped WDM data packets are then propagating into the pod until reaching a POADM that is connected with a rack (RPOADM). Then the RPOADM controls which of the packets will be dropped to the rack and which will continue propagating to the other racks. At this point it should be mentioned that the RPOADMs not only control the data traffic from other pods but also manage the intra-pod communication relaxing the need for any additional mechanism or equipment. This is ensured by the fact that each rack can add new data packets to WDM traffic utilizing the same transmitter irrespective of the final destination, since both data packets intended to the rack inside a local pod or to the racks of the other pods will be transmitted to the local RPOADM. The aggregate data traffic that is generated by all the racks inside a pod and is destined to reach racks out of the pod is inserted to the WDM backbone network through an optical coupler. As the RRA does not necessitate any switch, it benefits from reduced hardware complexity and cost of the node, however it is expected to introduce higher network latency due to the fact that each packet should first travel through the optical ring nodes until reaching its final destination.

5 Main OPST node Building Blocks

The proposed OPST architectures comprise a backbone network that consists of an optical ring fiber and ring nodes. Each ring node is composed of different processing building blocks properly connected in order to achieve the drop and add operations. The role of each node building block is described individually below:

5.1 *Packet Optical Add/Drop Multiplexer (POADM)*

The POADM architecture is presented in figure 7. The WDM data packets are first preamplified through an optical amplifier before being inserted synchronously to the POADM. The incoming data packets are optically demultiplexed by means of an optical WDM demultiplexer and are launched to the optical gates. At the same time the control channel that carries the information about which packets should be dropped to the pod or not is inserted to the ECU. The ECU in turn drives the optical gates either in "OFF" state to drop the packets or in "ON" state when the packets have to transparently bypass the ring node. In this way the drop and pass operations are imprinted to the "OFF" and "ON" state of the optical gates respectively. The non-dropped packets are multiplexed again via a WDM multiplexer and are coupled with the new packets of the pod. The new packets are generated during the available free time slots and are embedded to the non-dropped WDM traffic. Then, the aggregate traffic is amplified before propagating to the backbone network. At this point it should be mentioned that if new data packets are not added to the backbone traffic, the ring nodes generate dummy packets in order to maintain constant peak power levels for each wavelength at the output of the optical amplifier. Two different POADM switching architectures are proposed here, both based on technologies deployed within PhoxTroT. In the first case the optical gate comprises a Mach–Zehnder Interferometer (MZI) whereas in the second case, the optical gate consists of a coupler followed by a tunable All-Pass ring.

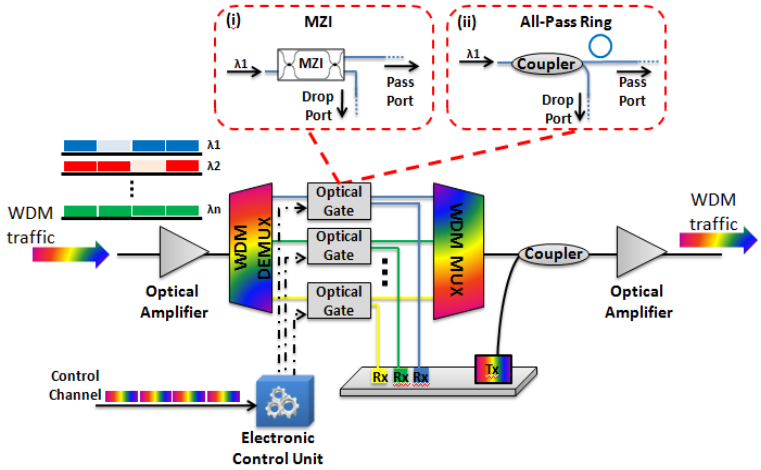


Figure 7: POADM architecture: (i) based on MZIs and (ii) based on All-Pass Rings

5.1.1 POADM based on Mach–Zehnder Interferometers (MZIs)

The inset (i) of figure 7 presents the POADM architecture based on MZIs. The switching operation in MZI is controlled by the ECU according to the information of the control wavelength-channel. In case a drop operation is intended, the ECU transmits an electrical control signal to the MZI optical gate to switch its operation, setting it in the "OFF" state. Thus the packets to be dropped will emerge at the MZI's switched port, hereby termed as Drop port, while at the same time the corresponding time slots are released. In case the ECU drives the MZI in "ON" state, the packets will exit through the unswitched port of the MZI, namely the Pass Port and will transit transparently through the node. The MZIs will rely on Silicon-on-Insulator (SOI) technology and will employ either the Carrier Depletion (CD) or Silicon Organic Hybrid (SOH) switching mechanism. Both MZI designs consist of straight and bends waveguides and two multimode interference couplers which provide a perfect -3dB splitting ratio (50% of the input power in each arm), but their basic difference is the mechanism that the phase shifters employ for the switching operation. In the case of the CD-based MZIs, the switching operation relies on a waveguide technology with a vertical p-n junction within the core that depletes the available carriers and thus changes the refractive index values, when an electrical control signal is applied. On the other hand, the SOH based MZIs exploit slotted waveguides filled with a polymer of high electro-optic effect. The characteristics for both MZI switching technologies have been extracted in WP6 and are summarized in table 1 below (see also deliverable 6.1).

Table 1: MZIs Characteristics

	SOI based MZI	SOH based MZI
Drive Voltage	3 V	3 V
Interaction Length	1250 μm	1000 μm
Insertion Loss	1.55 dB	3.0 – 5.0 dB
Footprint	0.080 mm^2	0.060 mm^2
Power Consumption	16 mW	2.6 mW
Switching Time	<50ps	< 50 ps

5.1.2 POADM base on Rings

Ring resonators have also been considered as an alternative to the MZIs due to their ability to achieve high extinction ratios under low drive voltages. The inset (ii) of figure 7 presents the POADM architecture based on tunable all-pass rings. In this case the optical gate comprises a coupler followed by a tunable all-pass ring. The coupler continuously drops a tap of the incoming light for all recirculating WDM data packets through the Drop port. However only those packets of the tap that must be dropped are processed by the node, while the rest of the packets are discarded from the node. On the other hand, for the remaining optical power travelling within the transit line of the coupler, the optical data packets cross a tunable all-pass ring which is driven by the ECU. Once an optical packet has reached its destination node, the ECU drives the all-pass ring in "OFF" state by detuning the ring to a corresponding off-resonant-wavelength, so as to block the packet passing through the node. In case a data packet is destined to another node, the ECU drives the all-pass ring in "ON" state by tuning the resonant wavelength of the ring, so as to let the packet to pass though the node via the Pass Port. The tunable all-pass rings will again rely

on SOI technology and can be designed either CD- or SOH-based electro-optic rings, following the design specifications of [30] with compact rings featuring high ER (>20 dB) and large FSR (>15nm) values The physical layer specifications for each waveguide technology have been extracted in WP6.

5.2 *Intra-Pod and Pod Switch*

The two switches employed within the RSA, namely the Inter-Pod switch and the Intra-Pod switch, are photonic nxn switching matrices that connect multiple inputs to multiple outputs. These photonic nxn switches consist of multiple cascaded 2x2 switching elements in a Benes architecture. The 2x2 switching elements are MZI switches exploiting either the CD- or the SOH-switching mechanism. Figure 8 presents such a 4x4 non-blocking switching matrix consisting of 6 symmetric single-arm MZI-based switching elements arranged in a Benes topology, while many 4x4 switching matrices can be combined in larger switching topologies with a higher port count, towards implementing the Inetr- and Intra-Pod photonic NxN switching matrices.

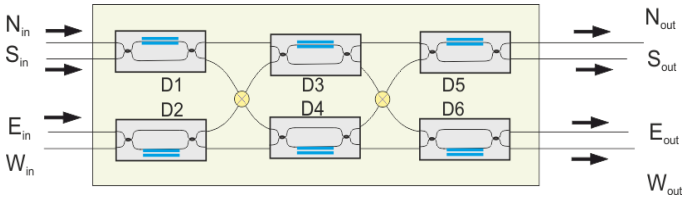


Figure 8: 4x4 non-blocking switching matrix

The overall 3D router architecture, that will incorporate a 4x4 switching matrix and is being currently developed within PhoxTrot, will route a stream of 12 multiplexed signals per port of the 4x4 photonic non-blocking switch, leading to 4x12 input and 4x12 output multiplexed signals. The electrical interface of the 3D Router consists of a 12-PDs array, the cache memory, the electronic processor and a 12-VCSELs array. Each VCSEL of the 12-VCSELs array emits at different wavelength from 1520nm to 1580nm and supports multi-level modulation formats with bit rate up to 40 Gb/s. The VCSELs are fabricated by Vertilas and their specifications have been summarized in the deliverable D4.2. The 3D router can perform two different levels of routing: the coarse and the fine operation.

The coarse operation of the 3D router is depicted in figure 9, where all the 12 incoming traffic channels entering the 12 input pins of one input port will exit through the corresponding 12 output pins of one output port. In detail, the 12 input channels are multiplexed through an Array Waveguide Grating (AWG) and fed into the first input port of the 4x4 switch. Then the multiplexed channels are routed to the 4th output port of the 4x4 switch and demultiplexed by means of an AWG to the corresponding output pins. For example the red, blue and green wavelengths will be inserted to the 1st, 11th and 12th pin of the 1st input port and will exit respectively through the 13th, 23rd and 24th pin, which are the 1st, 11th and 12th pin of the 4th output port.

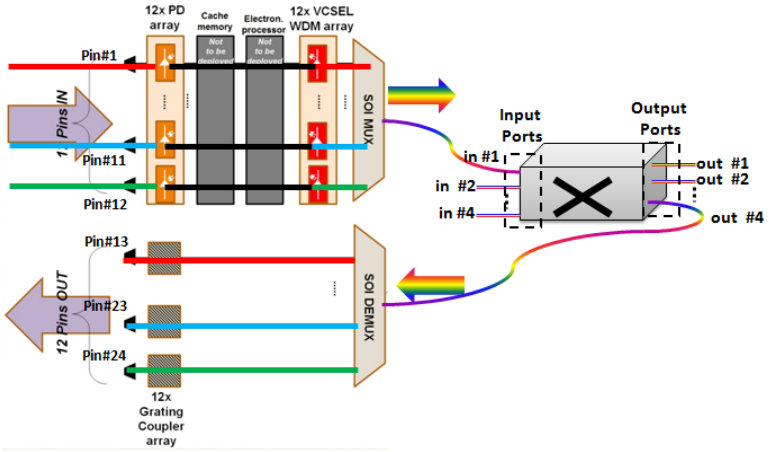


Figure 9: Coarse Operation of the nxn non-blocking switching matrix

The case of fine operation supports wavelength conversion and is presented on figure 10. Every traffic stream entering in one of the twelve pins of an input port of the 4x4 switch can leave from any of the twelve output pins of an output port. For example the red optical signal is now inserted to the photodiode element of the 1st input pin and after its electrical conversion is buffered to the cache memory until the electronic processor decides the wavelength to which the optical stream will be converted, namely the destination address. Then the electronic processor turns on the VCSEL with emission frequency corresponding to the chosen destination wavelength, marked with the blue wavelength in our case. The blue wavelength now is carrying the information and is fed to the photonic 4x4 switch the AWG multiplexer. Finally, the blue wavelength will be routed to the corresponding output port and will exit the switch trough the output pin which supports the blue wavelength, namely 24th pin out.

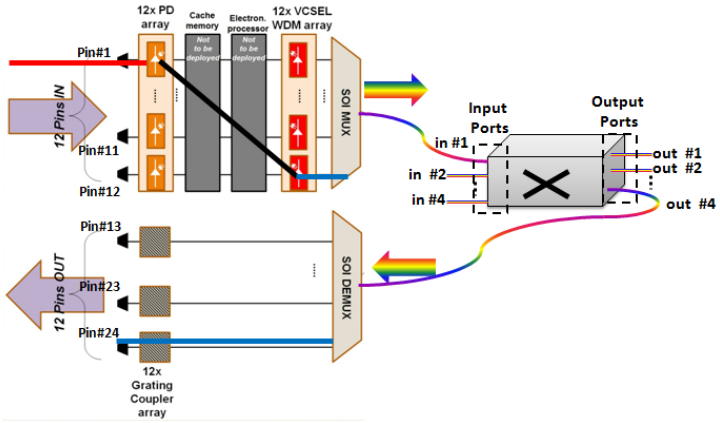


Figure 10: Fine Operation of the nxn non-blocking switching matrix

In both operation cases, each pin can support up to 40 Gb/s line rate and consequently each nxn switch port can reach an aggregate traffic up to 480Gb/s (12 pins x 40 Gb/s per pin). Thus the total traffic that can be supported by the nxn switch can be calculated by multiplying the number of input ports n by the 480 Gb/s aggregate throughput per port.

5.3 *Add-Packet Circuit*

The RSA can recognize an empty time slot in a fixed wavelength and add a new packet to the WDM optical ring through the Add Packet circuit that is presented in figure 11. The packets that will be added to the WDM optical ring are sent to the Add Packet circuit by means of active optical cables (AOCs). The optical packets are converted to electrical packets at the receivers of the AOC and are held in electrical buffers until the next empty time slot. The electrical buffer reads the headers of the packets and when a time slot in a fixed wavelength, for example λ_i, is available, it transmits the packet to the driver that directly modulates the VCSEL with emission frequency the λ_i. All optical packets are then multiplexed though a WDM multiplexer and are coupled with the remaining non-dropped traffic, thus allowing the aggregate traffic emerging at the output of the node to propagate within the WDM optical ring without requiring any additional electrical/optical buffering until reaching the egress node.

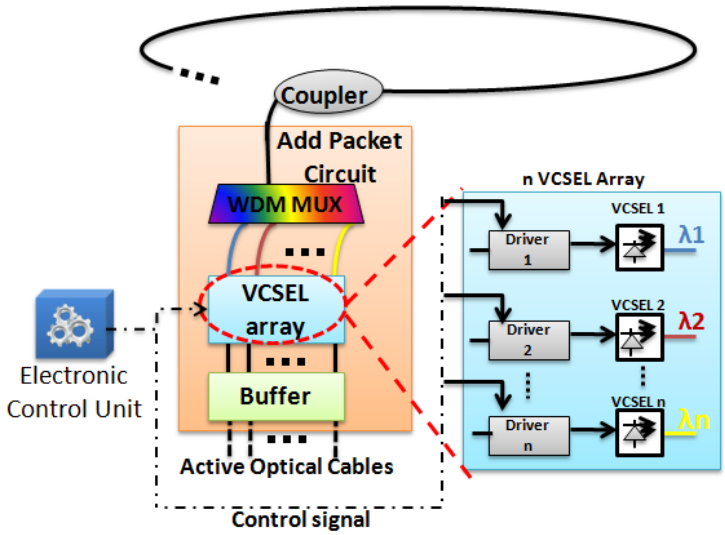


Figure 11: Add Packet Circuit architecture

5.4 *Top-of-Rack Unit*

The Top-of-Rack (ToR) units, namely receivers (Rx), AOC and transmitters (Tx), sit at the very top of the pod racks and are utilized for both the local communication within the pod (intra-pod traffic) and external communication between the pods (backbone traffic), as depicted in figure 12.

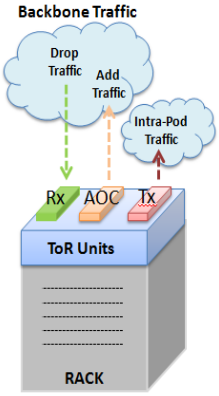


Figure 12: Top of Rack Units

The Rx unit consists of a n-PD array and receives optical packets that either derive from the Inter-Pod Switch or the Intra-Pod switch. The optical packets are converted then to electrical data that are buffered in an electronic cache memory before being transmitted into the rack. The AOC unit connects each rack with the Add-Packet Circuit of the ring node and is used only for sending the new packets to the Add-Circuit in order to be embedded to the non-dropped packet traffic before propagating to the WDM optical ring. The AOC architecture will follow the design of the AOC that has been proposed in PhoxTroT and will utilize a number of wavelengths that will be sufficient for serving the rack traffic needs. The Tx unit is used by a rack for transmitting a packet to another rack in to the same pod, implying that the Tx unit is used only in local communication. In detail, an electrical packet is forwarded to the Tx unit of a rack and is inserted into a drive of the n-Driver array which directly modulates a VCSEL of the n-VCSEL array. The output optical signal at each VCSEL is then propagated to the corresponding input pin of one input port of the nxn Intra-Pod switch and is routed to its destination rack.

6 Ring-Star Architecture

Figure 13 illustrates in detail the design of the RSA. The WDM data traffic is inserted to the ring node through the POADM and is demultiplexed via the WDM demultiplexer of the drop-packet circuit. At the same time the control channel is launched to the ECU which in turn drives the optical gates and the Inter-Pod-switch. In case the ECU drives the optical gates to "OFF" state the packets at each transit line are dropped to the nxn Inter-Pod switch, whereas in case of "ON" state the packets bypass transparently through the node. The nxn Inter-Pod switch receives the dropped packets and routes them to the output port that is connected with the corresponding destined rack. The non-dropped packets are multiplexed again by means of a WDM multiplexer and are propagated to the network. On the other hand, when the racks of a pod transmit data packets to the backbone traffic, they should first send these packets to the Add-Packet circuit through the AOCs. The Add-Packet Circuit then recognizes the available time slots per wavelength and transmits the new data packets after multiplexing them via a WDM multiplexer. The new multiplexed packets are coupled together with the remaining non-dropped traffic and are propagated to the backbone network.

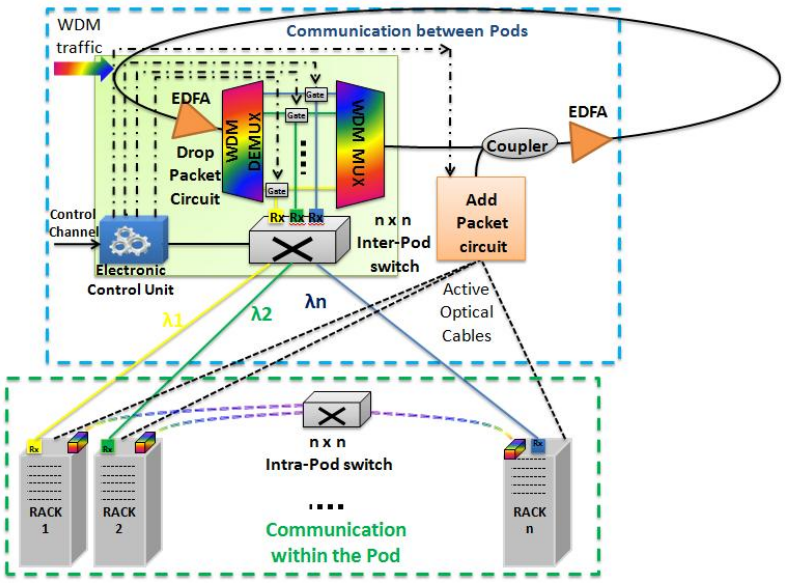


Figure 13: RSA with Pod and Intra-Pod switches

At this point it should be mentioned that the Add-Packet circuit recognizes the released time slots derived either from the dropped packets or the available time slots that are subsisted in to the backbone traffic. Apart from the communication between the pods, the racks inside a pod communicate with each other by exchanging data packets. Each rack can transmit a packet destined to another intra-pod rack through the Tx of the ToR units. The Tx unit sends the packet to the Intra-Pod switch which in turn routes the packet to the destined rack. Thus, the Intra-Pod communication is carried out via the nxn Intra-Pod switch without the need of any additional equipment and is totally independent from the Inter-Pod communication.

7 Ring-Ring Architecture

The RRA is a packet architecture that comprises a WDM optical ring and ring nodes, where each ring node connects a pod of racks to the WDM optical ring and consists of a POADM, an Electronic Control unit and a WDM multiplexer. Each pod of racks is also a WDM ring and each rack is connected to the optical fiber through a RPOADM as depicted on figure 14. The data packets that are targeting a pod are dropped by the corresponding node and are propagated to pod ring through a WDM multiplexer. The dropped packets are circulated to the ring until reaching the 1st RPOADM. The 1st RPOADM drops the packets that are destined to its rack whereas the non-dropped packets are multiplexed again via a WDM multiplexer and continue propagating into to the optical ring within the pod until reaching their final destination. Each rack can recognize the time slot and the fixed wavelength of a received packet and thus can add a new packet to the released time slot of the fixed wavelength. The new generated packets from each rack are multiplexed though a WDM multiplexer and are coupled with the non-dropped packet traffic generated by the pod in order to be transmitted to the optical ring within the pod. Each rack can transmit new packets destined either to the racks in the same pod, namely local communication or to racks of another pod, namely external communication. In case of local communication, the new data packets circulate in to the optical ring within the pod until reaching their final destination. In case of external communication, the new packets should first traverse the ring within the pod before being coupled with the non-dropped

backbone traffic. The aggregate traffic is amplified by means of an EDFA and is propagated to the backbone ring network.

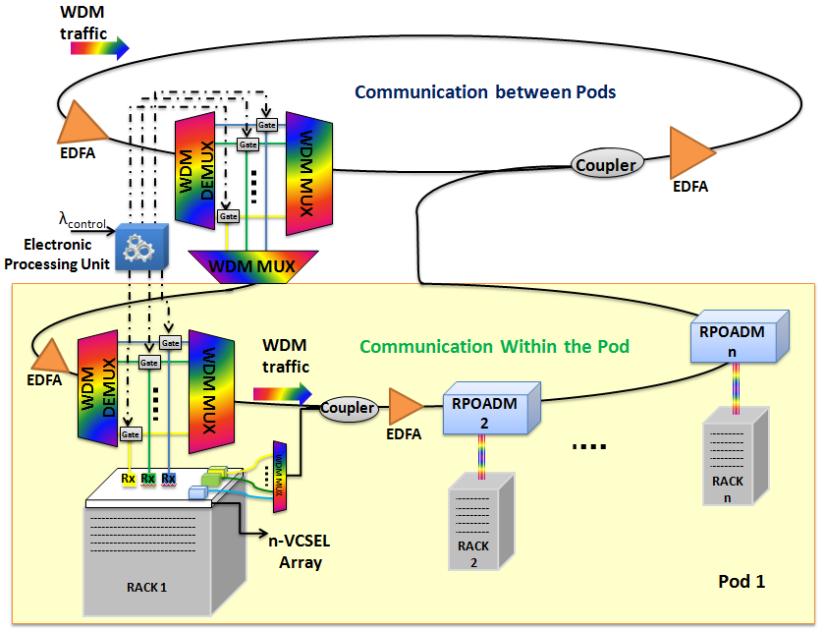


Figure 14: RRA with RPOADM design

8 Overview of the critical parameters

In this section we discuss the critical parameters of the proposed architectures and investigate how these parameters affect the cascadability, the capacity and the cost of the network. Cascadability issues stem from the fact that the POADMs induce losses as the optical signal passes through the nodes, leading in this way in a RSA or RRA with only few nodes. The employment of EDFAs at the input and the output of each POADM is an acceptable low cost solution for amplifying the optical signals and managing the power budget. Compensating the insertion losses of the optical data streams with the total gain of the EDFAs towards the utilization of a high number of nodes is however one side of the story. The additive noise that is introduced by the cascade of multiple EDFAs within the backbone ring packet network might degrade the overall signal quality and should be treated carefully towards a fully operational network design. The final number of the cascade nodes will thus depend on the quality of the transmitted signals, the degradation of the signals, when passing through the EDFAs, which is directly related to the optical signal to noise ratio (OSNR), and finally the sensitivity of the employed PDs. Another critical parameter directly related with the number of the nodes is the Extinction Ratio (ER) of the optical signals and the On/Off switching operation of the Drop/Add circuits. If the ER of the packets to be dropped is low, the released slots of the WDM data traffic (recirculating within the backbone ring) will contain some remaining low-power packets as crosstalk for the newly added packets.. This will degrade the quality and consequently the ER of the transmitted packets, limiting in this way the number of usable nodes. On the other hand a high ER for the switching operation of the Drop/Add circuits will ensure a high ER and signal quality for the newly added packets, which will be able to traverse more nodes.

Another critical parameter is the number of the wavelengths and the channel spacing in the ring are expected to greatly impact the performance of the network. The total number of wavelengths should be sufficient for serving the traffic inside the network and is a trade off between the number of racks in each pod and the wavelength capacity. The wavelength's capacity should be totally exploited by different data flows that are destined to different nodes succeeding in this way the minimization of the wavelengths. The data capacity also depends on the guard bands time between the data packets, which is directly related by the switching time of the nxn switches. For example, if the nxn switch has low switching time, then the guard bands time should be high so that the switch will be able to route the data packets before the arrival of new ones. A switch with slow switching time limits the data wavelength capacity. On the other hand, a high switching time enables the utilization of low guard bands time and consequently improved exploitation of the wavelength capacity. Towards enhanced wavelength capacity, multilevel modulation formats are also exploited within the proposed architectures, which in turn implies that the capacity is multiplied in comparison with one level modulation formats.

Finally the power consumption and the number of the nodes are two parameters that affect the cost of the RSA and RRA and for this reason the appropriate number of nodes should be decided in the light of cost and power consumption minimization.

9 PhoxTrot technology in HPC environment

In this section we examine the impact of the application of Phoxtrot AOCs and router chips on the system performance of a number of existing HPC systems. In particular, we examine the interconnection networks of existing HPC systems that are high in the top 500 list [31] (IBM Blue Gene Q is third, and K supercomputer is fourth, as of June 2014). For these HPC systems we examine how their interconnection network would change if we used Phoxtrot AOCs and router chips, focusing mainly in rack-to-rack communication.

- ☑ **Cray Jaguar**

Cray XT [32][32] systems are constructed from dual socket Opteron nodes. Cray compute nodes are organized in a 3D Torus (every node is directly connected to 6 neighbours). In Cray XT5 series a compute node consists of two AMD Opteron 2000-series processors (dual or quad core), each coupled with its own memory and dedicated Cray SeaStar2+ communication ASIC (electronic router chip). HyperTransport technology enables a 6.4 GB/sec direct connection (3.2 GB/s uni-directional = 25.6 Gbps) between the computing elements and the Cray SeaStar2+ router chip. Seastar2+ provides 6 switch ports at 9.6 GB/sec each (4.8 GB/s unidirectional = 38.4 Gbps). Each Cray XT5 blade includes four compute nodes (1 x 2 x 2). Four nodes are packaged on a blade card, 16 in a rack (cabinet), for a total of 96 nodes (192 processor sockets) per rack and a 1x4x16 network. Jaguar, a Cray XT5 system, has 200 racks arranged in 8 rows of 25 racks resulting in a 25x32x16 system network. Figure 15 shows the 3D torus interconnect of Jaguar system.

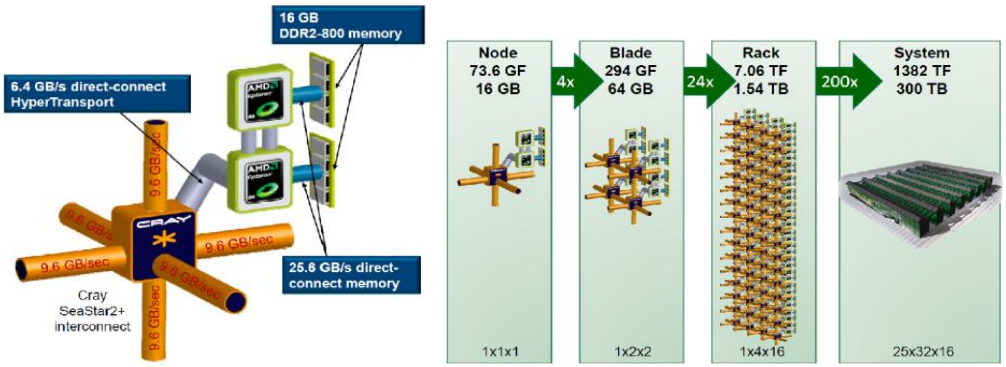


Figure 15: Cray XT5 (a) node (b) packaging hierarcy. Courtesy of Cray

If we consider a rack, this corresponds to a 1x4x16 2D torus subnetwork of the 25x32x16 3D torus total system. A rack has 160 torus links/channels leaving the rack: 64 to adjacent racks per direction in dimension *x* (2 neighbouring racks in dimension *x*), 16 to adjacent racks per direction in dimension *y* (2 neighbouring racks in dimension *y*). As stated above each torus link is of 4.8 GB/s capacity.

Applying Phoxtrot AOCs (8 lanes of 160 Gbps, 4 lanes per direction=640 Gbps/direction) for rack-to-rack communication we have:

- In dimension *x*, we have (64*4.8 GB/s)=2457,6 Gbps capacity to adjacent rack (per direction). Thus, we need =4 AOCs for rack-to-rack communication per direction in *x* dimension.
- In dimension *y* we have (16*4.8 GB/s)=614,4 Gbps capacity to adjacent rack (per direction). Thus, we need =1 AOCs for rack-to-rack communication per direction in *y* dimension.

The speedup[1] of Cray is calculated as follows:

The system has 25x32x16=12800 nodes and bisection width (*Bw*): 800 (bi-directional) channels (calculated based on the specific 3D torus topology). The bisection bandwidth (*Bb*) is *Bb=Bw bi-directional channel capacity*, and the traffic crossing the *Bw* is found for uniform traffic, assuming that the processors inject 25.6 Gbps traffic per node.

☑ **Replacing Cray routers with Compass EoS routers**

We now examine how the architecture would change if we replaced the Cray SeaStar2+ router chips with PhoxTrot's CompassEoS router chips (168 bi-directional channels at 8 Gbps each). We would require waveguides of channels of 8 Gbps required for processors-to-router connection (+ 4 waveguides for receiver). The Compass EoS chip's

[1] Speedup: ratio of the available bandwidth of the bottleneck channel to the bottleneck channel load. Uniform random traffic it is a commonly used traffic pattern for evaluation. For Uniform random traffic, speedup equals the ratio of the bisection bandwidth to the half of the total generated traffic. The latter crosses the bisection channels. Speedup equal to 1 means that under ideal conditions (perfect routing, load balancing, infinite flow granularity) the network could accommodate the injected traffic with no congestion. Designing a network with speedup greater than 1, allows non-idealities in the implementation.

164 channels can be used for routing channels per dimension of the 3D torus. In this case, the resulting speedup would be:

And we would have 22 AOCs for rack-to-rack communication per direction in *x* dimension, and 6 AOCs for rack-to-rack communication per direction in *y* dimension.

☑ Adding more processors

The application of PhoxTrot's Compass EoS router chips in combination with AOCs would allow ideally the accommodation of more processors per router (assuming that they can still be packaged on a single board).

Assuming that we would want to build the same topology with speedup ≥ 1, we would have

Where *n* is the number of processor chips per router (or equivalently the number of times we can increase the injection bandwidth). With *n* = 2 we can use 2 times more processor chips per node (4 processor chips per node) using (8 Gbps) links per dimension of the 3D torus. We would require 21 and 6 AOCs for rack-to-rack communication per direction in dimensions *x* and *y*, respectively.

For keeping the speedup ≈ (close to the original system speedup),

Thus we can use 5 times more processors per node (10 processors), using (8 Gbps) links per dimension of the 3D torus, with system speedup close to that of the original system (0.375). We would require 20 and 5 AOCs for rack-to-rack communication per direction in dimensions *x* and *y*, respectively.

☑ IBM Blue Gene Q

Blue Gene/Q (BG/Q) [33] is the third generation of highly scalable, power efficient supercomputers in the IBM Blue Gene line, following Blue Gene/L and Blue Gene/P. BG/Q compute nodes are organized in a 5D Torus (every node is directly connected to 10 neighbours). Applications run on the compute nodes while file IO is shipped from a compute to an IO node, where it is then sent over a PCIe interface to a file system. A BG/Q compute node consists of the SoC singlechip module with associated memory. Each chip has 11 network ports. Each can transmit and receive at 2 GB/s (= 16 Gbps): 4GB/s totally for a bi-directional link. 10 links are used to form the 5D torus and 1 link is used to connect

to I/O node. 32 compute nodes are electrically interconnected to form a 2x2x2x2x2 grid on a node card. 16 node cards comprise a 512-node midplane and two midplanes stack vertically to form a 1024-node rack, with electrical links within midplanes and optical links between midplanes. A single rack contains a 4x4x4x8x2 subnetwork of the system 5D torus. Figure 16 shows the hierarchical construction of the 5D torus of the IBM BG/Q.

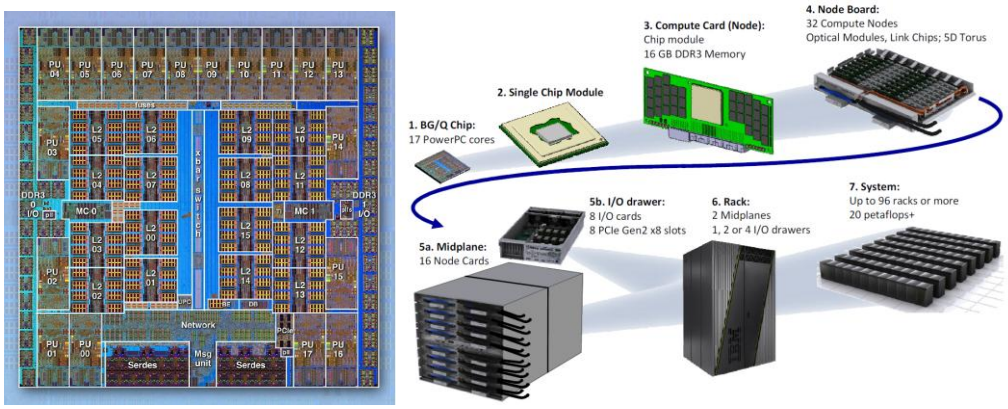


Figure 16: Blue Gene/Q (a) chip (b) packaging hierarcy. Courtesy of IBM

We will examine the Sequoia - BlueGene/Q system of 64 racks: a 16x16x16x8x2 torus (65536 nodes).

A rack corresponds to a 4x4x4x8x2 torus subnetwork (rack) of the 16x16x16x8x2 whole system torus and has 1536 torus links/channels leaving the rack: 512 to each of dimensions 1, 2, 3, that is, 256 channels to adjacent racks per direction of dimensions 1, 2, 3 (since we have 2 adjacent racks per dimension). Applying Phoxtrot AOCs (640 Gbps/direction) for rack-to-rack communication we would require:

We have (256*16 Gbps)= 4096 Gbps capacity per direction, in each dimensions 1, 2 and 3. Thus, we need =7 AOCs for rack-to-rack connection in each direction of each dimensions 1, 2 and 3, of the 5D torus network.

The speedup of Blue Gene/Q is calculated as follows:

The system has 16x16x16x8x2=65536 nodes and bisection width (*Bw*): 8192 (bi-directional) channels (calculated based on the specific 5D torus topology). The bisection bandwidth (*Bb*) is equal to *Bb=Bw bi-directional channel capacity*, and the traffic crossing the *Bw* is found for uniform traffic, assuming that the processors inject 16Gbps traffic per node.

☑ **Replacing routers with Compass EoS routers**

We now examine how the architecture would change if we replaced the routing chips of BG/Q with the PhoxTrot's CompassEoS router chips (168 bi-directionalchannels of 8 Gbps each). We would need 2 waveguides (channels of 8 Gbps) for processor-to-router connection (+ 2 waveguides for receiver). The CompassEoS chip's 166 channels can be used for routing channels per dimension of the 5D torus.

So, we would require (256*8*17 Gbps)= 34816 Gbps per direction of each of the 1,2,3 dimensions. Thus, we would need =55 PhoxTrot AOCs for rack-ro-rack communication per direction of the 1,2,3 dimensions of the 5D torus interconnect.

☑ Adding more processors

The application of PhoxTrot's Compass EoS router chips in combination with AOCs would allow ideally the accommodation of more processors per router (assuming that they can still be packaged on a single board).

Assuming that we would want to build the same topology with speedup ≥ 1, we would have:

Where *n* is the number of processor chips per router (or equivalently the number of times we can increase the injection bandwidth). The results *n* = 4, means that we can use 4 times more processors per node (4x1=4 processors) using (8 Gbps) links per dimension of the 5D torus interconnect. In this case we would require 13 AOCs for rack-to-rack communication per direction of the 1,2,3 dimensions of the 5D torus interconnect. For keeping speedup = (close to the original system speedup)

Thus we can use 8 times more processors per node (8x1=8 processors), using (8 Gbps) links per dimension of the 5D torus interconnect, with system speedup close to the speedup of the original system (0.5). We would require 12 AOCs for rack-to-rack communication per direction of the 1,2,3 dimensions of the 5D torus interconnect.

☑ K computer

K-Computer [34][35] is a massively parallel computer system developed by Fujitsu and RIKEN. A compute node consists of a single CPU and an interconnect controller. 10 links are used for inter-node connection at 10 GB/s per link (5GB/s unidirectional). The topology is the Tofu interconnect: a 6D Mesh/Torus topology. A position in the 6D Torus is given by six-dimensional coordinates: X, Y, Z, A, B, C. Tofu interconnect is actually a structure with ABC groups of 3D Mesh/Tori of 2 x 3 x 2 connected by a XYZ- 3D Torus. Within ABC structures (or Tofu units), each node has a constant degree of 4. Of the total 10 available links, 4 of them are used for ABC interconnection and the remaining 6 for XYZ. 4 CPU chips on a board are interconnected with the A- and C-axes links. 3 boards in a Tofu unit are interconnected with the B-axis links. Tofu units are interconnected with the X-, Y- and Z-axes links which form a 3D torus. Each pair of adjacent ABC Mesh/Torus is connected

with 12 links. 4 nodes are accommodated on a system board and 24 system boards (96 nodes) comprise a rack. There are 96/12=8 tofu units in a rack, thus Z- axis = 8. Each Tofu link can transmit and receive at 5 GB/s (=40 Gbps): 10GB/s totally for a bi-directional link. Figure 17 shows the internal structure of the Tofu interconnect.

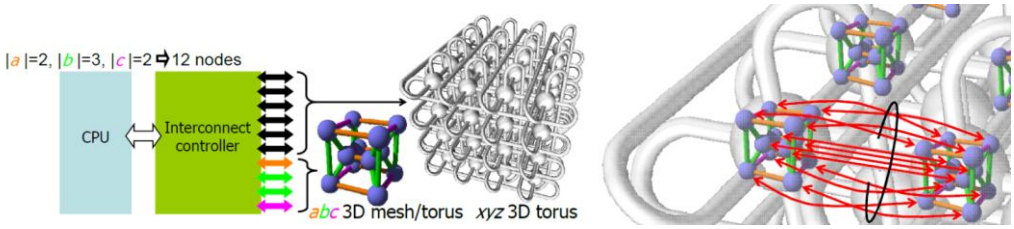


Figure 17: (a) Compute node and (b) system architecture of K computer. Courtesy of Fujitsu.

We will examine the K supercomputer system of 256 racks: a 16x16x8x2x3x2 mesh/torus (24576 nodes).

If we focus on a rack, this corresponds to a 1x1x8x2x3x2 torus subnetwork of the 16x16x8x2x3x2 whole system torus. A rack has 384 links/channels leaving the rack. More specifically, a 1x1x8 subtorus of the 16x16x8 3D XYZ torus has 32 torus links/channels leaving the rack. A single node of the XYZ torus is an ABC (Tofu) unit and each pair of adjacent ABC Mesh/Torus is connected with 12 links. Thus 32x12 = 384 links/channels leave a single rack: 192 to each of dimensions X, Y, that is, 96 channels to adjacent racks per direction of dimensions X, Y (since we have 2 adjacent racks per dimension).

Applying Phoxtrot AOCs (640 Gbps/direction) for rack-to-rack communication we would require: (96*40 Gbps)= 3840 Gbps per direction. Thus, we need =6 AOCs for rack-to-rack communication per direction of the XYZ torus.

Speedup of K computer is calculated as follows:

The system has 16x16x8x2x3x2=24576 nodes and bisection width (Bw): 256x12= 3072 (bi-directional) channels (256 is the bisection width of a 16x16x8 XYZ torus x 12 links for every XYZ link). The bisection bandwidth (*Bb*) is equal to *Bb=Bw bi-directional channel capacity*, and the traffic crossing the *Bw* is found for uniform traffic, assuming that the processors inject 40Gbps traffic per node.

☑ **Replacing routers with Compass EoS routers**

We now examine how the architecture would change if we replaced K router chips with PhoxTrot's CompassEoS router chips (168 bi-directional channels of 8 Gbps each). We would need 5 waveguides (channels of 8 Gbps) for processor-to-router connection (+ 4 waveguides for receiver). CompassEoS chip's 163 channels can be used for routing per dimension of the 6D mesh/torus.

And we would require 20 PhoxTrot AOCs for rack-to-rack communication per direction of the XYZ torus.

☑ Adding more processors

The application of PhoxTrot's Compass EoS router chips in combination with AOCs would allow ideally the accommodation of more processors per router (assuming that they can still be packaged on a single board).

Assuming that we would want to build the same topology with speedup **≥** 1:

Where *n* is the number of times we can increase the injection bandwidth (or equivalently the number of processors per router). This means that we could not increase the number of processors of the system if we wanted a system with speedup **≥** 1. For keeping speedup **≈** (close to the original system speedup)

Thus we can use 3 times more processors per node (3x1=3 processors), using (8 Gbps) links per dimension of the 5D torus, with system speedup close to the speedup the original system (0.5). We would require 18 AOCs for rack-to-rack communication per direction of the XYZ torus.

Concluding, we examined several existing HPC system interconnects of systems that are high in the top 500 list and examined how they would change based on the use of PhoxTrot's AOCs and router chips. We calculated that the numbers of AOCs required in the examined HPC systems for rack-to-rack communication are very low, and thus the use of PhoxTrot's AOCs would simplify their cabling. What was even more interesting was that we examined the performance of the systems assuming the replacement of the router chips of the examined HPC systems with PhoxTrot's CompassEoS routers. We observed that when keeping the same interconnect topology we would have much more bandwidth available for processors communication. So we would obtain much higher network speedup. Assuming networks with the same speedup as the original HPC system interconnects, the advantages of using the PhoxTrot's CompassEoS routers could be translated into using more processors per rack and the whole system. We calculated that we could increase the number of processors by 5, 8 and 3 times, for the Cray XT5, Sequoia Blue Gene/Q, and K supercomputer respectively. Note that the new systems with the higher number of processors would have the same speedup, meaning that they would have similar average network performance as the original systems. In all these cases the number of PhoxTrot's AOC required per rack was found to be also small. The above finding manifest the suitability of the PhoxTrot components (AOCs and router chips) for building HPC interconnects based on traditional well established topologies.

10 References

[1] Cisco, Cisco Global Cloud Index: Forecast and Methodology, 2012-2017
[2] Moor Insight & Strategy, Intel's Disaggregated Server Rack, 2013
[3] T. Benson, A. Akella, D. A. Maltz, "Network traffic characteristics of datacenters in the wild", Conference on Internet measurement (IMC), pp. 267-280, 2010
[4] "Open Compute Project," 2014. [Online]. Available: http://www.opencompute.org/
[5] "Facebook Shatters the Computer Server Into Tiny Pieces," 2013. [Online]. Available: http://www.wired.com/2013/01/facebook-server-pieces/
[6] http://www.wired.com/2011/11/calxedas-arm-for-the-cloud/
[7] http://www.zdnet.com/open-compute-does-the-data-center-have-an-open-future-7000013012/
[8] N. Farrington, G. Porter, S. Radhakishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen and A. Vahdat, "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," in SIGCOMM'10, New Delhi, India, 2010
[9] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. Eugene Ng, M. Kozuch and M. Ryan, "c-Through: Part-time Optics in Data Centers," in SIGCOMM'10, New Delhi, India, 2010
[10] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen and Y. Chen, "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibilty," IEEE/ACM Transactions on Networking, 2013
[11] C. 3. MEMS, "The Software Defined Hybrid Packet," 2013
[12] H. Liu, F. Lu, R. Kapoor, A. Forencich, G. M. Voelker, G. Papen, A. C. Snoeren and G. Porter, "REACToR: A REconfigurable pAcket and Circuit ToR Switch," in IEEE Photonics Society Summer Topical Meetings, Waikoloa, HI USA, 2013
[13] K. Christodoulopolos, K. Katrinis, M. Ruffini, D. O'Mahony, "Accelerating HPC Workloads with Dynamic Adaptation of a Software-Defined Hybrid Electronic/Optical Interconnect", paper Th2A.11, OFC 2014
[14] D. Siracusa, G. Maier, V. Linzalata and A. Pattavina, "Scalability of Optical Interconnections based on the Array Waveguide Grating in High Capacity Routers," in 15th International Conference on Optical Network Design and Modeling (ONDM), Bologna, Italy, 2011
[15] T. Benson, A. Akella, D. A. Maltz, "Network traffic characteristics of datacenters in the wild", Conference on Internet measurement (IMC), pp. 267-280, 2010
[16] R. Proietti, Yawei Yin, R. Yu, C. J. Nitta, V. Akella, C. Mineo and S. Yoo, "Scalable Optical Interconnect Architecture Using AWGR-Based TONAK LION Switch With Limited Number of Wavelengths," Journal of Lightwave Technology, vol. 31, no. 24, pp. 4087-4096, 2013
[17] K. Xi, Y.-H. Kao, M. Yang and H. J. Chao, "Petabit Optical Switch for Data Center Networks," 2010
[18] T. Niwa, H. Hasegawa and K.-i. Sato, "A 270 x 270 Optical Cross-connect Switch Utilizing Wavelength Routing with Cascaded AWGs," in OFC/NFOEC, Anaheim CA USA, 2013
[19] N. Farrington, A. Forencich, G. Porter, P.-C. Sun, J. E. Ford, Y. Fainman, G. C. Papen and A. Vahdat, "A Multiport Microsecond Optical Circuit Switch for Data Center Networking," IEEE Photonics Technology Letters, vol. 25, no. 16, pp. 1589-1592, 2013
[20] N. Farrington, P.-C. Sun, A. Forencich, J. Ford, Y. Fainman, G. Papen, A. Vahdat and G. Porter, "A Demonstration of Ultra-Low-Latency Data Center Optical Circuit Switching," in SIGCOMM'12, Helsinki, Finland, 2012
[21] A. Predieri, M. Biancani, S. Spadaro, G. Bernini, P. Cruschelli, N. Ciulli, R. Monno, S. Peng, Y. Yan, N. Amaya, G. Zervas, N. Calabretta, H. Dorren, S. Iordache, J. C. Sancho, Y. Becerra, M. Farreras, C. Liou and I. Hussain, "Lightness Deliverable D2.2: Design Document for the proposed network architecture," 2013
[22] R. R Grzybowsky et. al., "The OSMOSIS Optical Packet Switch for Supercomputers: Enabling Technologies and Measured Performance", Photonics in Switching, 2007, pp. 21-22.
[23] B. Uscumlic, et al. "Impact of peer-to-peer traffic on the efficiency of optical packet rings." International Conference on Broadband Communications, Networks and Systems (BROADNETS), 2008.
[24] C. Cadéré, et al. "Virtual circuit allocation with QoS guarantees in the ECOFRAME optical ring." Conference on Optical Network Design and Modeling (ONDM), 2010.
[25] B. Uscumlic, et al. "WDM optical packet ring performance insights: Scheduling and capacity." IEEE Symposium on Computers and Communications (ISCC), 2012.
[26] B. Uscumlic, I. Cerutti, A. Gravey, P. Gravey, D. Barth, M. Morvan, P. Castoldi, "Optimal dimensioning of the WDM unidirectional ECOFRAME optical packet ring", Springer Photonic Network Communications, 2011.
[27] http://www.intunenetworks.com/home/shape-up/core_innovation/opst_technical_introduction/
[28] J. Dunne, Tom Farrell, and Jim Shields. "Optical packet switch and transport: A new metro platform to reduce costs and power by 50% to 75% while simultaneously increasing deterministic performance levels." Transparent Optical Networks, 2009. ICTON'09. 11th International Conference on. IEEE, 2009.
[29] H. Liu, et al. "Circuit Switching Under the Radar with REACToR." Proceedings of the 11th ACM/USENIX Symposium on Networked Systems Design and Implementation (NSDI), Seattle, WA. 2014
[30] Shijun Xiao, Maroof H. Khan, Hao Shen and Minghao Qi, "Compact silicon microring resonators with ultra-low propagation loss in the C band", OSA, Optics Express 14475, 29 October 2007, Vol. 15
[31] http://www.top500.org/list/2014/06/?page=1
[32] Bland, Arthur S., et al. "Jaguar: The world's most powerful computer." Memory (TB) 300.62 (2009): 362.
[33] Chen, Dong, et al. "Looking under the hood of the ibm blue gene/q network." Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. IEEE Computer Society Press, 2012.
[34] Ajima, Yuuichirou, et al. "Tofu: Interconnect for the K computer." Fujitsu Sci. Tech. J 48.3 (2012): 280-285.
[35] Maeda, Hideki, et al. "System packaging technologies for the K computer." Fujitsu Scientific and Technical Journal 48.3 (2012): 286-294