NEPHELE: An End-to-End Scalable and Dynamically Reconfigurable Optical Architecture for Application-Aware SDN Cloud Data Centers

Paraskevas Bakopoulos, Konstantinos Christodoulopoulos, Giada Landi, Muzzamil Aziz, Eitan Zahavi, Domenico Gallico, Richard Pitwon, Konstantinos Tokas, Ioannis Patronas, Marco Capitani, Christos Spatharakis, Konstantinos Yiannopoulos, Kai Wang, Konstantinos Kontodimas, Ioannis Lazarou, Philipp Wieder, Dionysios I. Reisis, Emmanouel (Manos) Varvarigos, Matteo Biancani, and Hercules Avramopoulos

Interaction of the optically switched data plane with a software-defined control and orchestration framework, meeting current common practices in data centers, necessitates the design of custom network control algorithms and software modules as well as the integration of novel functionalities. The authors present the approach of the European project NEPHELE, offering an end-to-end solution that addresses the optical data plane, the control plane, and its interaction with the application layer.

Abstract

The efficient integration of optical switching in data center networks is being studied as a means to cope with surging traffic demands. Optically switched, flatter network architectures more efficiently handle the east-west traffic profiles of modern data centers. Limitations in the port count and reconfiguration speed of optical switches require novel network designs offering network scalability and dynamicity. Interaction of the optically switched data plane with a software-defined control and orchestration framework, meeting current common practices in data centers, necessitates the design of custom network control algorithms and software modules as well as the integration of novel functionalities. The approach of the European project NEPHELE is presented, offering an end-to-end solution that addresses the optical data plane, the control plane, and its interaction with the application layer.

Introduction

Data centers are the hubs of our content-centric Internet. The proliferation of fifth generation (5G) mobile and cloud applications, video distribution, and the emerging Internet of Things is bringing data center traffic on a steep growth reaching 25 percent annually [1]. This soaring traffic demand is outpacing progress in network infrastructure, which generally follows Moore's law [2], threatening a capacity crunch inside the data center. The largest portion of this traffic concerns communications between servers and storage inside the data center, exacerbating the challenge. Following an east-west traffic profile, intra-data-center communication is dwarfing north-south traffic from/to the Internet, thus putting more stress on legacy data center network architectures that are inherently more appropriate for north-south flows (i.e., fat-tree topologies).

Optical switching is gaining momentum as a potential path for gracefully scaling data center networks, due to its inherent speed, energy efficiency, and transparency to bit rate and protocol. Having established itself in long-haul communication networks, the technology is now being advocated for deployment not only between data centers, but also inside them. A multitude of solutions have been proposed leveraging the most prominent optical switching technologies, such as space switching (e.g., using micro-electro-mechanical systems — MEMS or semiconductor optical amplifiers — SOAs [3–5]), wavelength switching (through combination of tunable lasers with arrayed waveguide grating routers — AWGRs [6, 7]), or a combination thereof (e.g., using wavelength-selective switches — WSSs [8, 9]). Introduction of optical switching in the data center has proven to be a nontrivial task due to the idiosyncracy of optical switches that differ from their electronic counterparts (it is not possible to retrofit optical switches into the existing infrastructure), and practical deployment stumbles on the following challenges.

Speed vs. size trade-off of mature optical switch technologies: High-port-count optical switches, like MEMS, typically offer millisecond reconfiguration times, whereas nanosecond-speed optical switches like polarized lead zirconate titanate (PLZTs) strive to exceed the dimensions of an 8×8 matrix, thus inhibiting network scalability. Hybrid architectures are proposed to tackle this challenge, relying on the combination of slow optical switches for long-lived "elephant" flows with optical [4] or electronic [9] packet switches for short-lived "mice" flows. Network dynamicity and scalability are thus interwoven with the underlying optical switching technology.

The buffer-less nature of optical switches: To avoid contention, a means of traffic scheduling is essential for the entire optical network. Scheduling flows in the buffer-less optical network effectively shifts all buffers toward the end hosts [9],

Paraskevas Bakopoulos and Eitan Zahavi are with Mellanox Technologies; Konstantinos Christodoulopoulos and Konstantinos Kontodimas are with the University of Patras; Muzzamil Aziz and Philipp Wieder are with GWDG; Domenico Gallico and Matteo Biancani are with Interoute; Richard Pitwon and Kai Wang are with Seagate UK; Konstantinos Tokas, Christos Spatharakis, Ioannis Lazarou, and Hercules Avramopoulos are with the National Technical University of Athens; Giada Landi and Marco Capitani are with Nextworks; Konstantinos Yiannopoulos is with the University of Peloponnese; Emmanouel (Manos) Varvarigos is with Monash University and the University of Patras.

Digital Object Identifier:
10.1109/MCOM.2018.1600804

0163-6804/18/$25.00 © 2018 IEEE

raising concerns regarding scheduler latency and buffer size. An interesting side-effect is that it can enable lossless networks that are of particular interest in certain data center types, like high-performance computing (HPC).

Insufficient integration of optical switching into the ubiquitous software-defined networking (SDN) paradigm: Recently, efforts have begun toward abstracting optical switch functionalities and integrating them into the control plane and orchestration platforms [10]. Since this task is strongly technology-dependent, further work is required to integrate novel architectures into the SDN ecosystem, with the ultimate target being to extend SDN's programmability to the optical layer.

The European project NEPHELE (www.nephele project.eu) is developing a dynamic optical network infrastructure for scale-out, disaggregated data centers that leverages optical switching with SDN control and orchestration to overcome current data center challenges. The project follows a vertical development approach extending from the data center architecture to the overlaying control plane and its interface with the application in order to deliver a fully functional networking solution, extending network virtualization to the optical layer. This multidisciplinary research brings the following innovations:

- A scalable data plane architecture, leveraging mature/commercial off-the-shelf (COTS) photonic component technologies. To enable dynamic and efficient sharing of resources, the NEPHELE network operates in a slotted time-division multiple-access (TDMA) manner.
- An SDN control and orchestration framework capable of managing the underlying data plane elements. NEPHELE's framework is the first to extend prominent SDN platforms with TDMA functionality, adding the capability to dynamically assign network resources directly at the optical layer. Fast resource allocation (scheduling) algorithms are being developed and integrated as add-ons to the SDN platform.

In the rest of this article, the main design routes of NEPHELE are outlined.

NEPHELE Dynamically Reconfigurable Data Plane Architecture

NEPHELE Network Overview

The NEPHELE network architecture is illustrated in Fig. 1a. The main building block is the pod, hosting a number of racks, accommodating a few thousand disaggregated resources (e.g., storage, compute) called "innovation zones"; hence, the pod is effectively a small-scale data center. Each rack is administered by a top-of-rack (ToR) switch, and all ToR switches are interconnected to the pod in a star topology, using one port per ToR. Each ToR port is equipped with a tunable laser and a burst mode receiver. For traffic destined within the same pod (intra-pod), switching is performed passively by means of optical filtering elements. To further scale the NEPHELE network, multiple pods are interconnected into a ring topology, which allows the use of small-port-count optical switches. Each NEPHELE ring carries wavelength-division multiplexed (WDM) traffic and consists of multiple fibers to supply the necessary capacity between pods. Communication between servers of different pods (inter-pod) is a combination of wavelength and space switching, allowing reuse of wavelengths among pods, and thus enabling network scalability beyond the typical wavelength count of dense WDM (DWDM) systems. Add/drop multiplexing to and from the NEPHELE ring is performed on a per-wavelength basis; hence, despite its ring physical topology, the network's logical topology is a mesh. The ensemble of a NEPHELE ring along with its corresponding pod switches and ToR ports is called a NEPHELE optical plane. To scale network capacity, additional and independent optical planes are deployed. This involves installing additional NEPHELE pod switches and connecting them through new rings, as well as populating additional ports in the ToR switches to connect to the newly added pod switches, as shown in Fig. 1b. Addition of new optical planes does not affect the existing ones, ensuring scalability and enabling pay-as-you-go deployment. The reference values of the NEPHELE architecture parameters (used for scalability and techno-economic studies throughout this article) are summarized in Table 1.

The NEPHELE data plane operates in a slotted TDMA manner, where "slots" are time segments that can be accessed by a single rack-to-rack communication. Slots (and therefore network resources) can be assigned dynamically to communicating racks, and the NEPHELE network can attain close to full utilization of the network capacity, leading to both energy and cost savings. The slotted operation of NEPHELE and its scalability using optical planes significantly expand on current demonstrations of optical data centers, while relying on mature photonic components [3–11]. In contrast to approaches based on elastic spectrum allocation [4], NEPHELE's TDMA approach provides dynamic assignment of network capacity without the need for complicated flex-grid hardware that would dramatically increase deployment cost. For very dynamic traffic scenarios dominated by mice flows, a hybrid electronic-optical implementation is considered, with the two networks interfacing at the ToR level.

The NEPHELE topology is a two level (tier) network: the first level comprises the ToR switches, and above them, there is a single level of pod switches. To support more servers, the network expands in the east-west direction, suiting much better the east-west type of traffic that flows in current data centers. Thus, in a sense, the NEPHELE network is flat, compared to legacy fat-tree networks that route traffic via several tree levels, the number of which depends on the number of servers. Note that the required network equipment scales linearly in NEPHELE, while the fat-tree network requires the addition of switches at all levels, and after a point the addition of a new level, yielding super-linear scaling of the number of servers.

To support more servers, the network expands in the east-west direction, suiting much better the east-west type of traffic that flows in current datacenters. Thus, in a sense, the NEPHELE network is flat, compared to legacy fat-tree networks that route traffic via several tree levels, the number of which depends on the number of servers.

NEPHELE Network Modules

The basic building blocks of the NEPHELE data plane are the ToR and the pod switch.

NEPHELE ToR Switch: Each NEPHELE ToR switch interconnects the devices in the data cen-

The pod switch is responsible for handling both intra-pod and inter-pod traffic in the optical domain, using different switching approaches for each scenario in order to achieve a combination of switching speed and scalability suitable for practical deployment in datacenter installations with realistic size and traffic dynamicity.

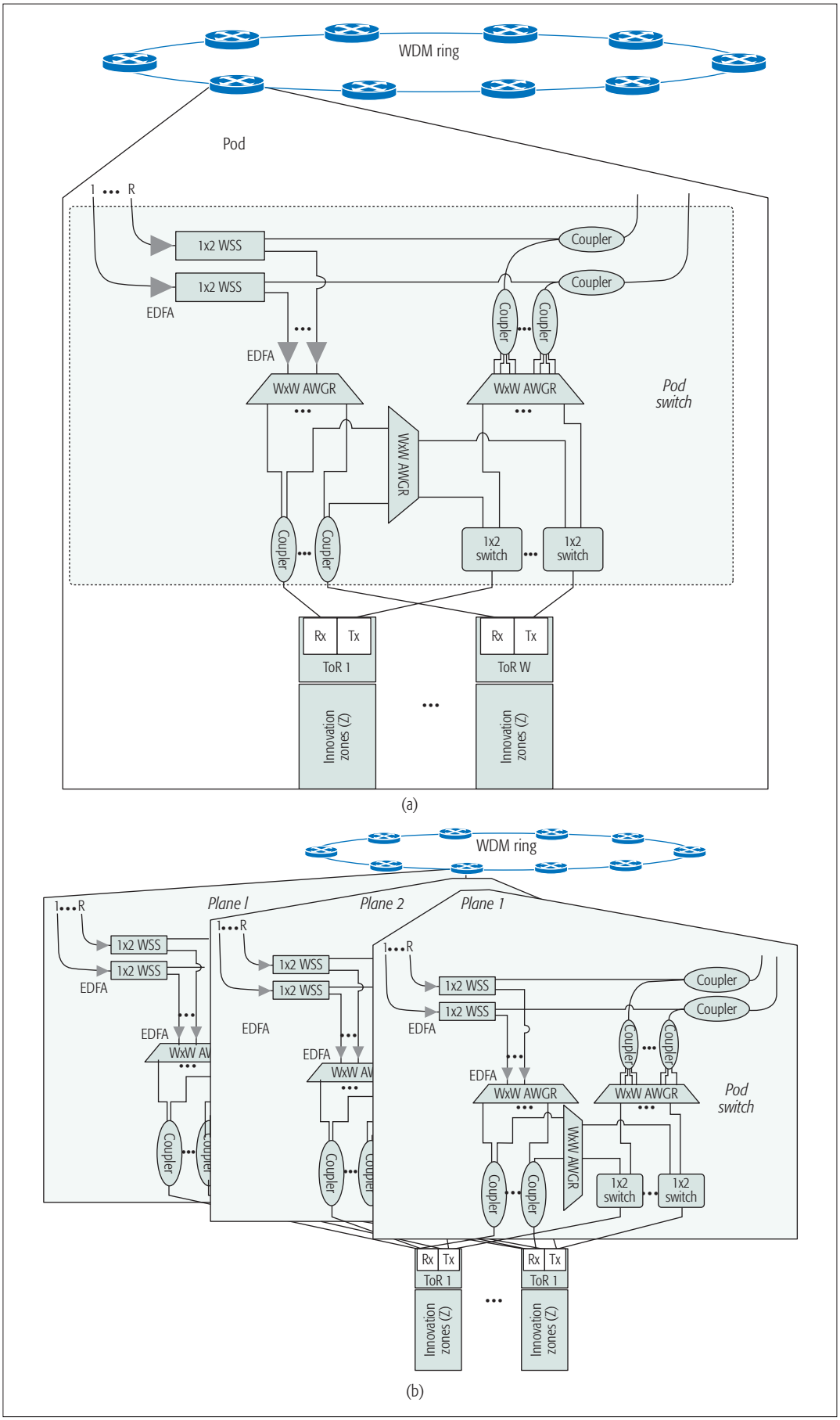

Figure 1. a) The NEPHELE optical data center network architecture; (b) scaling the NEPHELE network with the addition of optical planes.

ter racks among themselves as well as to the higher network tier, handled by the pod switches. The NEPHELE ToR is an extension of a standard Ethernet switch, with extra functionalities aiming to format the traffic originating from the end hosts and align it with the TDMA operation of the NEPHELE network, and to properly adjust the emission wavelength of the transmitted data, so as to enable wavelength switching. At its south interface, the NEPHELE ToR is equipped with standard Ethernet ports, for compatibility with Ethernet hosts. At its north ports, custom burst-mode optical transmitters and receivers are used to handle TDMA traffic. To maximize dynamicity, the ToR transmitters use fast tunable lasers: broadly available laser technologies such as MG-Y and DS-DBR have demonstrated wavelength tuning times below 50 ns across the entire C-band, for example [12]. The ToR gathers Ethernet frames from the end hosts connected to its south ports and assembles them into NEPHELE frames; each frame has a single destination, occupies a NEPHELE TDMA slot, and is assigned a wavelength and an optical plane for routing purposes.

NEPHELE Pod Switch: The pod switch resides at the top level of the NEPHELE network and is interconnected with its underlying ToRs in a star topology. Separate pod switches are used for each optical plane, collectively forming the individual pods. The pod switch is responsible for handling both intra-pod and inter-pod traffic in the optical domain, using different switching approaches for each scenario in order to achieve a combination of switching speed and scalability suitable for practical deployment in data center installations with realistic size (Table 1) and traffic dynamicity ([13]). Intra-pod traffic is switched solely according to its wavelength information, by means of a $W \times W$ AWGR. The selected wavelength for a communication between two ToRs depends on their location inside the corresponding pods. The number of ToRs in a pod equals the number of wavelengths, and wavelengths are reused for intra- and inter-pod communication. For inter-pod traffic, WSSs are used to drop traffic from the ring to the destination pod. One 1×2 WSS per fiber is used, operating in a TDMA manner so as to drop only the slots destined to the pod's racks. To constrain the guard periods between consecutive TDMA slots, fast WSSs are considered, for example, based on DLP technology offering switching times on the order of 10 μs [9]. Further routing of the dropped traffic to the destination ToRs is performed with an AWGR, according to the signal's wavelength, as in the case of intra-pod traffic. A third AWGR is used to add traffic to the ring, distributing groups of wavelengths coming from the ToRs to the available fibers in the ring. To distinguish between inter- and intra-pod traffic, a fast 1×2 space switch is used for each ToR port, routing up-bound traffic to the corresponding AWGR. In the NEPHELE network prototype under development, all optical switches in the pod (WSSs, 1×2 switches) are controlled by field programmable gate arrays (FPGA) boards executing the control plane's commands.

Routing in the NEPHELE Data Center Network

Optical routing in the NEPHELE data plane is depicted in Fig. 2. TDMA traffic flows originating from a ToR are first switched through the 1×2 switch according to their locality; if the traffic flow is destined to a ToR inside the same pod, it remains within the pod switch; otherwise, it is routed toward its east port. After the 1×2 switch, intra-pod traffic enters a $W \times W$ AWGR where it is passively routed. The AWGR's routing characteristics are static and depend on the wavelength of the incoming traffic and the input port from which it enters the AWGR (Fig. 3). Inter-pod traffic is routed via the fast 1×2 switch toward a second $W \times W$ AWGR followed by couplers for combining multiple AWGR outputs (typical: 4) into each fiber of the NEPHELE ring. After propagation in the ring, traffic is dropped at the destination pod's WSS on a per-fiber, per-wavelength, and per-slot basis according to the control plane's instructions. All the outputs of the WSSs — corresponding to all the pod optical planes — are introduced into a $W \times W$ AWGR and are passively routed to the ToRs of the destination pod. The combined routing characteristics of the two AWGRs involved in inter-pod communication yield at least one wavelength for reaching any destination ToR from any source ToR, ensuring non-blocking operation, while wavelength conflicts are avoided by proper allocation. Using optical switches with low port count in NEPHELE vouches for the scalability of its architecture and allows the use of COTS optical switches with a reasonable number of I/O ports and fast reconfiguration speed. Figure 2 insets show experimental results from a proof-of-concept experiment involving one source and two destination ToRs, connected in the same (Fig. 2a) or adjacent (Fig. 2b) pods. Data packets at 10 Gb/s with 200 μs duration and 10 μs guard time were successfully routed and received.

Parameter	Meaning	Typical value
Z	Number of innovation zones per ToR switch	4
S	Number of innovation zones' ports per ToR switch	20
W	Number of racks and ToRs per pod; also number of wavelengths in the system	80
R	Number of fiber rings per optical plane	20
P	Number of pods	20
I	Number of NEPHELE optical planes	20

Table 1. Dimensions of the NEPHELE reference network data plane architecture.

NEPHELE Techno-Economics

To support the adoption of the NEPHELE data center network (DCN) architecture, we have performed detailed scalability and techno-economic studies.

The scalability of the NEPHELE data plane was investigated through system simulations, using VPItransmissionMaker™. All data plane elements were modeled according to the commercial components' specifications, and two erbium-doped fiber amplifiers (EDFA) were used in each pod to compensate for the insertion losses. The interconnection of $P = 20$ pods was evaluated, and it was confirmed that the optical performance was acceptable under worst case transmission scenarios.

The selected wavelength for a communication between two ToRs depends on their location inside the corresponding pods. The number of ToRs in a pod equals the number of wavelengths, and wavelengths are re-used for intra- and inter-pod communication. For inter-pod traffic WSSs are used to drop traffic from the ring to the destination pod.

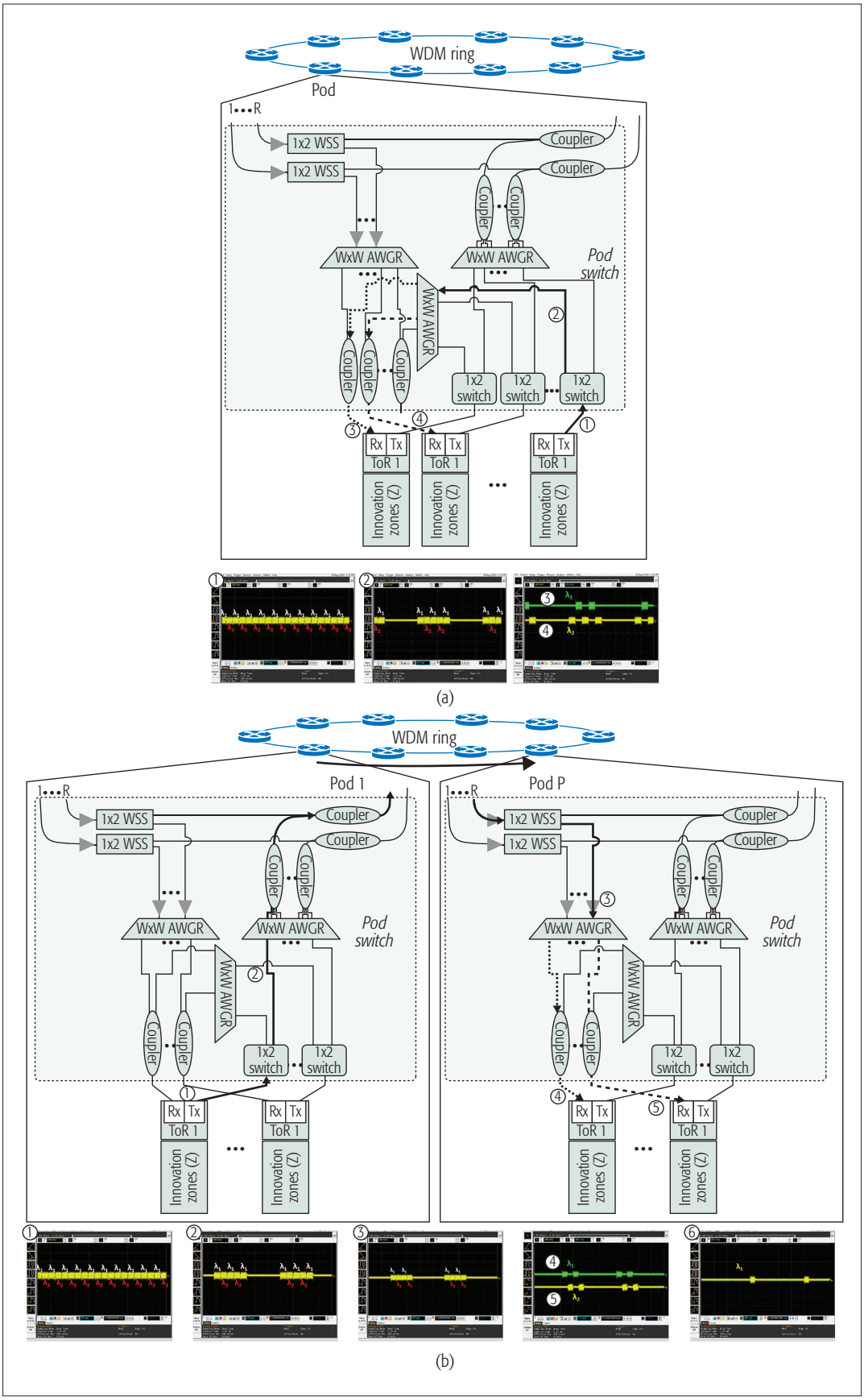

Figure 2. Routing in NEPHELE architecture: (a) intra-pod communication: (1) tunable-λ transmitter, (2) fast 1 × 2 switch routing traffic inside the pod, (3), (4) traffic delivered to ToR receivers according to its wavelength, after the $W \times W$ AWGR; b) inter-pod communication: 1) tunable-λ transmitter, (2) fast 1 × 2 switch routing traffic outside the source pod, (3) WSS dropping traffic in destination pod, (4),(5) traffic delivered to ToR receivers according to its wavelength after the $W \times W$ AWGR.

For the techno-economic study we calculated the cost of the NEPHELE network reference architecture as shown Fig. 1. The ToR NEPHELE switch consists of a 40-port electronic switch, and W tunable Txs, while the POD switch consists of W 1×2 WSS, W 1×2 space switches, 3 W×W AWGRs, and 2.R EDFAs (2 per ring). Note that the required number of these components is linear to the number of supported ports ($W.P.S$). For these components we obtained reference market prices and also projected their price evolution, taking into account the learning curve due to mass production and the simplification of specifications that will arise when applying these — currently telecom oriented — components into data center applications. We also calculated the cost of an equal sized fat-tree network that also provides full bisection bandwidth, assuming a folded Clos topology [14] and the use of 64-port Ethernet switches. The cost of the fat-tree network is linear for a range of supported ports, but after the upper limit a new tree level is deployed, and the cost increases linearly but with a higher slope. For the reference NEPHELE dimension (32K supported ports) and the projected component prices, the NEPHELE network was calculated to be about two times more expensive than the equivalent (three-level) fat-tree. However, as the number of supported ports increases, the difference decreases. The key reason for this is the linear increase of the cost of the NEPHELE network as opposed to the super-linear cost increase of the fat tree. For 256K ports the projected cost of the NEPHELE network is the same as the cost of an equivalent (four-level) fat tree. It is worth noting that the energy of the reference NEPHELE network (32K supported ports) is less than half of the equivalent fat-tree, and the benefits improve further as the size of the network increases.

NEPHELE Network Operation

Bandwidth Allocation Algorithms for Slotted Data Center Network Operation

The NEPHELE network aims to provide its resources in a dynamic fashion. To this end, time is divided in (time) slots, and dynamic slot allocation is performed in a periodic manner, with each period including T slots. ToR switches periodically report their bandwidth requests to the network controller, or applications report their communication requirements to the controller. The controller constructs a traffic matrix (TM) of size $W \cdot P \times W \cdot P$ for each period. A TM entry with coordinates (s, d) corresponds to the number of slots requested for the communication between ToR source s and ToR destination d.

The resource allocation algorithm (also referred to as scheduling algorithm) takes the TM as input and allocates slots and optical planes to these communicating pairs. This allocation must be performed in a coordinated manner taking into account the transmitter/receiver capabilities and avoiding wavelength collisions on the shared optical rings. The resource allocation is achieved by expressing (decomposing) the TM as a sum of I·T binary matrices, called permutation matrices (PMs), that conform to certain architecture-related constraints. Each PM represents the network configuration for a single slot and a single optical plane, and a binary entry of a PM with coordinates (s, d) indicates whether the (s, d) source-destination ToR pair communicates in the slot and plane that corresponds to the PM at hand. Figure 4 presents an example of the resource allocation process.

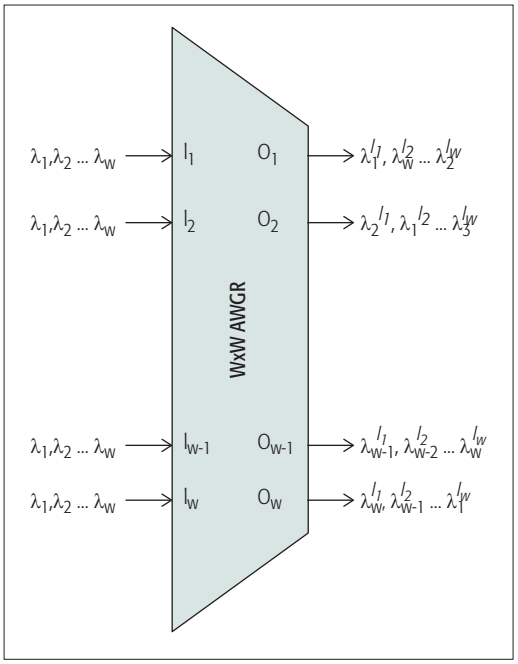

Figure 3. AWGR's routing characteristics.

Following the above, dynamic resource allocation becomes a TM decomposition problem that can be solved in an optimal manner using the Birkhoff-von Neumann theorem and bipartite graph matching [15]. Even though this approach ensures full utilization of the available slots and optical planes, the execution time of the optimal algorithm becomes prohibitive due to the high number of interconnected ToRs (tens of seconds for the reference network dimensions listed in Table 1 and the algorithm implemented in Matlab and executed on an Intel i5 laptop).

Given the limited applicability of optimal decomposition, we have explored a number of heuristic decomposition algorithms that achieve a trade-off between resource utilization and execution time. The proposed heuristics utilize previous decomposition solutions and appropriately modify them, given the updated traffic. This is performed in an *incremental* fashion and only traffic that was modified is taken into account. Consequently, the execution time is vastly improved and depends on the traffic dynamicity (i.e., how fast the traffic pattern changes in consequent periods) rather than the network size. For example, an incremental greedy algorithm was shown in simulations to be stable and achieve maximum throughput for load ≤ 0.8 (opposed to load = 1 for the optimal decomposition algorithm), and exhibit execution time lower than 0.2 s for the same setting (network dimensions, Matlab, and Intel i5 laptop) [13]. A parallel implementation of a greedy heuristic in an FPGA was shown to further reduce by one order of magnitude the execution time, while algorithmic solutions based on hierarchical control approaches (see the discussion on SDN controllers below) are also under examination.

The proposed heuristics utilize previous decomposition solutions and appropriately modify them, given the updated traffic. This is performed in an incremental fashion and only traffic that was modified is taken into account. Consequently, the execution time is vastly improved and depends on the traffic dynamicity rather than the network size.

The NEPHELE network control framework is based on the SDN controller and is composed of SDN applications implementing the algorithms and logic of the NEPHELE datacenter network. The SDN controller is a centralized entity that is in charge of configuring the data plane for deploying the virtual networks requested by the cloud orchestration framework.

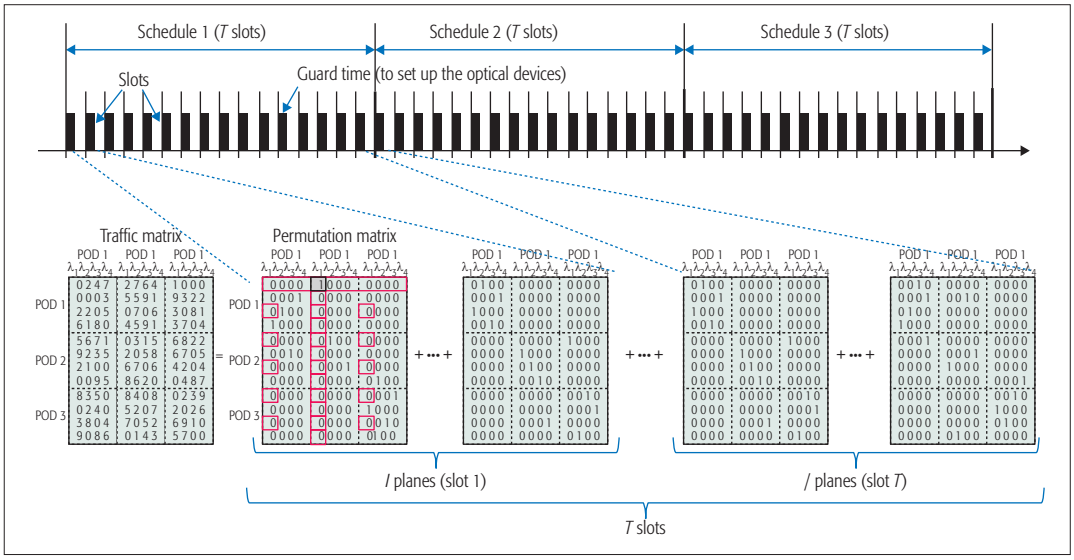

Figure 4. The NEPHELE dynamic resource allocation problem: The time is divided into (time) slots, and resource allocation is performed periodically. The traffic matrix of a period is decomposed into I · T permutation matrices (PMs), each corresponding to the configuration of the NEPHELE network for a specific slot and optical plane. In particular, the (s, d) entry of a PM indicates whether the (s, d) source-destination ToR pair communicate in the slot and optical plane to which the PM corresponds. The architectural constraints are translated into related PMs' constraints (for the ToR pair communication in blue, the ToR pairs indicated in red cannot be scheduled in the same permutation matrix — the same optical plane and slot).

Synchronization

Slotted operation of the NEPHELE data plane necessitates precise time synchronization to smoothly implement the calculated schedule in the bandwidth allocation process. Effective network synchronization includes the following functionalities:

- Clock distribution, allowing all elements to share a common accurate clock of the same frequency (syntonization)
- Time synchronization, meaning that all elements also share the same time reference
- Propagation delay estimation, so that each data plane component has an estimate of its relative delay to all the remaining components on every possible light path

The following approaches are considered for synchronization in NEPHELE.

Absolute Timing Synchronization: In this approach, all the network elements are synchronized to a high-precision time reference, so they all share the exact same local time. Clock distribution is performed through a dedicated optical link that transports a signal from a reference pod switch to the entire network. For efficient demultiplexing of the reference signal, the clock wavelength is considered to be in a different waveband (e.g., in the O-band). To obtain time synchronization, the reference pod transmits structured frames carrying synchronization patterns and information on its counters. Identification of each data plane element's local time is achieved by estimating propagation delay from the reference pod, using a timing protocol running over the control network (e.g., reverse Precision Time Protocol — PTP) or over the synchronization wavelength (e.g., a PTP-based timing protocol). The delay matrix for all possible source-destination combinations is estimated during network initialization, when data plane elements exchange timestamped messages through the link. Although somewhat complex, this concept is compatible with established practices and timing protocols, which makes it interesting for practical deployments where generic solutions are sought.

Relative Timing Synchronization: This approach relies on a floating time reference, travelling at the speed of light along the NEPHELE rings. Clock distribution follows the same concept of a dedicated wavelength; however, time synchronization requires the data plane elements to be aware only of their local time (i.e., a local counter, related to the TDMA slot arriving at the particular element). The master broadcasts frames instructing the data plane elements when to start their counters. For effective network operation, the network controller feeds the data plane elements with a schedule referring to their local counter's value. Hence, this concept bypasses the need for a detailed link delay matrix of the entire network and is therefore the choice followed in NEPHELE. To mitigate fault effects regarding loss of synchronization, the design implements a "slot counter" on a stable clock domain. Periodic control messages from the master notify the data plane elements to recalibrate their slot counter, which in turn inform the controller in case of lost synchronization.

NEPHELE Control and Orchestration Framework

Functional Architecture for a NEPHELE SDN-Based Data Center

The resource programmability offered by SDN is gaining attention for data center operation management, since the fine-grained control required to manage the data center resources resides at the control and orchestration level.

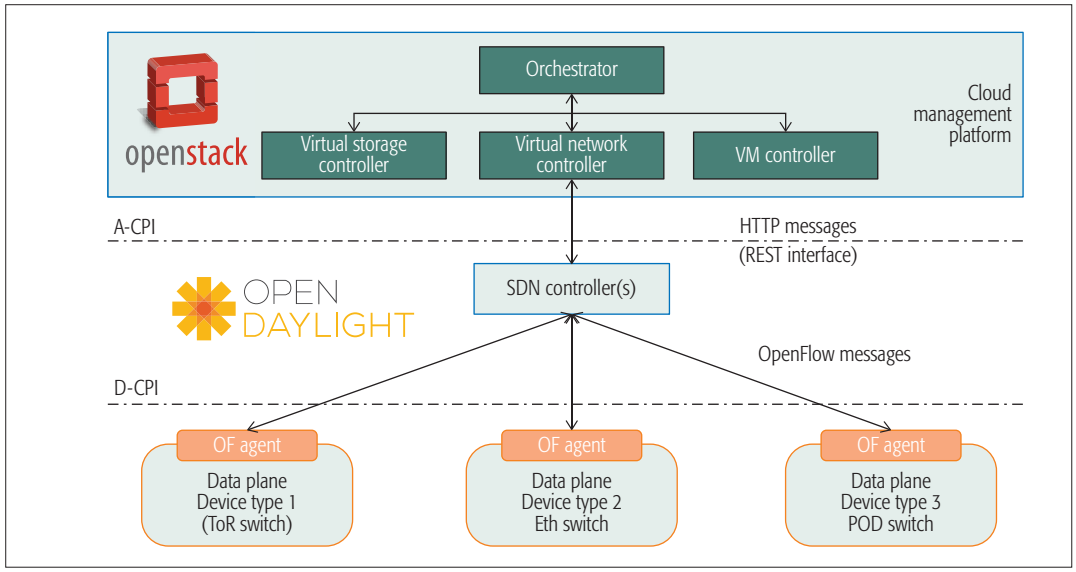

Figure 5. Overall architecture of the NEPHELE data center control infrastructure.

The prototype SDN agent is able to act as a proxy for both legacy Ethernet and novel optical switching devices. This is why it implements the parsing mechanism for both standardized and extended (NEPHELE-specific optical extensions) OpenFlow 1.3 commands.

The NEPHELE architecture consists of the following components (Fig. 5):

- A *cloud orchestration framework*, in charge of operating the whole NEPHELE data center infrastructure
- A *network control framework*, in charge of operating the NEPHELE network with the required level of efficiency and flexibility

The cloud orchestration framework manages all data center resources (computing, memory, network, and storage) within the Innovation Zones. It is implemented through a cloud management platform like OpenStack. Its main function is the on-demand delivery of the virtual environments requested by the customers, while guaranteeing the efficiency of the data center utilization from the operator's perspective. This task is carried out by OpenStack's Heat component, which ensures the right order in the deployment, provisioning, and configuration of the different virtual resources. Moreover, it takes care of the whole life cycle of a virtual resource, like managing on-demand modifications or automated scaling actions. Other specific components are responsible for configuring each type of virtual resource: computing (Nova), storage (Swift, Cinder, and Glance), and network (Neutron).

The NEPHELE network control framework is based on the SDN controller and is composed of SDN applications implementing the algorithms and logic of the NEPHELE DCN. The SDN controller is a centralized entity that is in charge of configuring the data plane for deploying the virtual networks requested by the cloud orchestration framework. Tight integration with the upper layer orchestrator allows the NEPHELE architecture to advance the approach used by state-of-the-art solutions for TDMA-based resource allocation in optical data centers (like the ones proposed in [4, 12]), introducing application awareness at the network level. In fact, the NEPHELE scheduling algorithms driving resource allocation described above take as input the connection requirements of the cloud applications, as declared in the requests issued by the cloud orchestrator in order to continuously update the traffic matrix and compute the network allocation solution. The latter is then automatically translated into a set of OpenFlow-based commands sent to the agents of the data plane devices and configured on their FPGAs.

NEPHELE adopts the OpenFlow protocol for the interaction between the SDN controller and the data plane, with extensions for the configuration and advertisement of optical devices. Three types of interaction are defined:

- Advertisement of the data plane devices capabilities (e.g., active ports, switching capabilities, available wavelengths and time slots)
- Operational configuration of the devices (e.g., adding a flow entry, creating a cross-connection with time slots and wavelength specification)
- Data plane monitoring, including asynchronous notifications from the data plane to the controller and retrieval of traffic counters from the controller to the data plane

The above extensions are realized in the form of an SDN agent, whose key responsibility is to receive the OpenFlow commands from the controller, translate them, and forward them to the data plane devices. The prototype SDN agent is able to act as a proxy for both legacy Ethernet and novel optical switching devices. This is why it implements the parsing mechanism for both standardized and extended (NEPHELE-specific optical extensions) OpenFlow 1.3 commands. Furthermore, it has the ability to detect out-of-order arrival of control plane commands, and re-order them in a particular schedule before pushing the flows to the device FPGA. The design of the SDN agent is made extensible to accommodate any vendor-specific device instructions/extensions in the future; that is, only the translation mechanism needs to be updated, without disturbing the other modules.

Deployment Models for SDN Controllers in NEPHELE Data Centers

Scalability is a key feature in NEPHELE, achieved through a modular approach based on optical planes and pods easily scalable to large data cen-

The network topology is maintained at the "child" SDN controllers, while only aggregated details are transferred to the centralized "parent" controller through abstraction procedures, thus reducing dimension and complexity at the parent level and improving control plane scalability. In this case the resource allocation problem is typically solved through the cooperation of parent and child controllers.

ters. At the control plane, in small-size contexts a single centralized controller, typically deployed in a redundant manner for high-availability purposes, covers the entire data center network. This controller has the full knowledge of the network topology, including pod and ToR switches and innovation zones, each of them supporting a certain number of virtual machines (VMs). This detailed view, with network nodes' capabilities, ports status, traffic load, and resource allocation in terms of wavelengths and time slots, allows implementing effective algorithms to estimate the optimal solution for the global resource allocation problem. Based on this, the network is periodically re-optimized following the evolution of the global TM representing the network's traffic load.

However, target values for the NEPHELE network scale up to 400 pod switches, 1600 ToR switches, and 6400 network interface cards (NICs) at the innovation zones, for a total of 8400 network devices and around 150 million flows; these values may overload the centralized controller. A possible solution is based on a distributed model with SDN controllers responsible for specific network partitions and hierarchically coordinated through a parent controller. The network topology is maintained at the "child" SDN controllers, while only aggregated details are transferred to the centralized "parent" controller through abstraction procedures, thus reducing dimension and complexity at the parent level and improving control plane scalability. In this case the resource allocation problem is typically solved through the cooperation of parent and child controllers.

Different hierarchical approaches can be applied in NEPHELE environments.

Per-layer SDN controllers with child controllers dedicated to ToR and pod switches, respectively. This model fits classical data center network approaches with core-leaf separation and allows adopting technology-specific controllers. Intra-ToR traffic is managed exclusively by the ToRs' controller, while inter-ToR and inter-pod traffic is handled through decisions at the pods' controller and the parent controller.

Per-optical-plane SDN controllers with child controllers responsible to operate all the network devices belonging to single optical planes. The main limitation of this model stems from its misalignment with the hierarchy of traffic flows and the complexity at the parent controller. Moreover, physical ToR switches need to be partitioned in logical devices assigned to different controllers, since their ports are associated with different optical planes.

Per-pod SDN controllers with child controllers responsible for all the pod and ToR switches in a given pod. Each child controller manages intra-pod traffic and, since most of the traffic will stay intra-pod, the hierarchical coordination at the parent is simplified. This model is preferred in NEPHELE deployments, since it guarantees a fair load balancing between child controllers, with devices and traffic flows equally distributed in different network partitions, and it reflects the logical distribution of the traffic among servers.

NEPHELE Control Plane Prototype Implementation

NEPHELE supports the following list of functionalities and services in its control and orchestration framework.

- Application Affinity: Collect the application/service level agreement (SLA) requirements from the orchestrator and translate them into network configuration decisions.
- Data Center Virtualization: Abstract the network resources to help the promotion of concepts of network resource partitioning and network as a service (NaaS), and the creation of virtual data centers.
- Monitoring: Collect monitoring information from the underlying network and make decisions to improve network performance and efficiency.
- Dynamic Bandwidth Allocation: Control the data plane devices in a coordinated manner to avoid conflicts and allocate the resources efficiently.

The application affinity service in a dynamically reconfigurable DCN has been functionally validated through the implementation of a proof-of-concept prototype of the NEPHELE controller, based on the OpenDaylight Lithium version. The application affinity service constitutes an SDN application (written in Java) that exposes a REST-based northbound Interface to the data center orchestrator (e.g., OpenStack) to enable requests for network connections with specific application requirements. To communicate with the NEPHELE data plane devices, the prototype extends the OpenFlow plugin to support wavelengths and time slots in OpenFlow messages, and includes a set of new SDN applications implementing the logic of the application affinity service, the creation of the traffic matrix, and the computation of the resource allocation solution. Development considerations for the other aforementioned services were also carried out: collection of monitoring information from the data plane devices can rely on the OpenFlow messages for retrieval of counters related to the established flows, while data center virtualization can be performed in a REST northbound interface similar to application affinity.

The controller prototype has been tested over a simple network emulated with the Mininet tool and including two planes, with three pod switches organized in a double ring, each of them connected to four ToR switches. More extensive tests integrating the SDN agents that would translate the OpenFlow commands and forward them to prototype data plane devices (ToRs and pod switches — Fig. 2), along with a more dense network topology are planned as future work.

Conclusion

Optical switching is gaining traction as a promising enabler for scaling data center networks beyond the trajectory of Moore's law. The European project NEPHELE is developing an end-to-end optical infrastructure for scale-out, disaggregated data centers. Efforts are focusing toward the development of scalable optical network implementations that are compatible with the characteristics and limitations of current photonic technologies, enabling rapid deployment. In order to make a real impact on data center networks, the NEPHELE architecture employs slotted network operation offering dynamic allocation of resources. Efficient algorithms for rapid scheduling are under development with close to optimal performance.

Control of NEPHELE's optical data center network is performed through an SDN cloud orchestration and network control framework that extends popular open source implementations with essential functionalities for efficient interaction with the optical data plane. Deployment models are investigated enabling graceful scaling of the NEPHELE network.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 645212 (NEPHELE).

References

[1] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2014–2019," 2015; http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.pdf, accessed 30 Nov. 2016.
[2] Ethernet Alliance, "The 2016 Ethernet Roadmap," 2016; http://www.ethernetalliance.org/roadmap/, accessed 30 Nov. 2016.
[3] H. Liu *et al.*, "REACToR: A Reconfigurable Packet and Circuit ToR Switch," *Proc. IEEE Photonics Soc. Summer Topical Meeting Series*, 2013, pp. 235–36.
[4] G. M. Saridis *et al.*, "Lightness: A Function-Virtualizable Software Defined Data Center Network with All-Optical Circuit/Packet Switching," *J. Lightwave Technol.*, vol. 34, 2016, pp. 1618–27.
[5] L. Xu *et al.*, "A Hybrid Optical Packet and Wavelength Selective Switching Platform for High-Performance Data Center Networks," *Optics Express*, vol. 19, no. 24, 2011, pp. 24,258–67.
[6] K. Sato *et al.*, "A Large-Scale Wavelength Routing Optical Switch for Data Center Networks," *IEEE Commun. Mag.*, vol. 51 no. 9, Sept. 2013, pp. 46–52.
[7] R. Proietti *et al.*, "Scalable Optical Interconnect Architecture Using AWGR-Based TONAK LION Switch with Limited Number of Wavelengths," *J. Lightwave Technol.*, vol. 31, no. 24, 2013, pp. 4087–97.
[8] G. M. Saridis *et al.*, "Lightness: A Function-Virtualizable Software Defined Data Center Network with All-Optical Circuit/Packet Switching," *J. Lightwave Technol.*, vol. 34, no. 7, 2016, pp. 1618–27.
[9] G. Porter *et al.*, "Integrating Microsecond Circuit Switching into the Data Center," *Proc. SIGCOMM*, 2013, pp. 447–58.
[10] Shuping Peng *et al.*, "Multi-Tenant Software-Defined Hybrid Optical Switched Data Centre," *IEEE/OSA J. Lightwave Technol.*, vol. 33, no. 15, pp. 3224–33.
[11] J. Wang *et al.*, "Energy-Efficient Optical HPC and Datacenter Networks Using Optimized Wavelength Channel Allocation," *SPECTS*, 2015.
[12] J. M. Fabrega *et al.*, "Modulated Grating Y-Structure Tunable Laser for Routed Networks and Optical Access," *IEEE J. Selected Topics in Quantum Electronics*, vol. 17, no. 6, 2011, pp. 1542–51.
[13] K. Christodoulopoulos *et al.*, "Bandwidth Allocation in the NEPHELE Hybrid Optical Interconnect," *ICTON* 2016.
[14] M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," *SIGCOMM*, 2008.
[15] J. E. Hopcroft and R. M. Karp, "An n5/2 Algorithm for Maximum Matchings in Bipartite Graphs," *SIAM J. Computing*, 1973.

Biographies

Paraskevas Bakopoulos [M] (paraskevasb@mellanox.com) received his Diploma and Ph.D degrees from the National Technical University of Athens (NTUA), Greece. He worked as a senior researcher at NTUA and is currently a senior staff engineer at Mellanox Technologies. His research interests focus on optical interconnects and optical switching in data center networks. He has authored or co-authored more than 110 peer-reviewed articles and conference papers and holds 2 patents. He is actively involved in several European projects and has participated in Technical Program Committees of photonics-related events.

Konstantinos Christodoulopoulos (kchristodou@ceid.upatras.gr) received a Diploma from the School of Electrical and Computer Engineering (ECE), NTUA, an M.Sc. in advanced computing from Imperial College London, and a Ph.D from the Department of Computer Engineering and Informatics (CEID), University of Patras. He worked as an adjunct assistant professor at CEID, as a senior researcher at the Computer Technology Institute and Trinity College Dublin, and as a contractor for IBM Research Ireland. He recently started to work as a researcher at ECE. His research interests are in the areas of algorithms and protocols for communication and computer networks.

Giada Landi (g.landi@nextworks.it) received her degree in telecommunication engineering from the University of Pisa, Italy, in 2005. She is an R&D project manager at Nextworks. Her research areas include SDN, NFV, cloud, and service orchestration, with participation in several research and industrial projects. Some of her past activities focused on ASON/GMPLS, PCE, control plane for wireless access networks, and inter-technology mobility. She is currently active in the H2020 NEPHELE, BlueSpace, 5G Crosshaul, and 5G TRANSFORMER projects.

Muzzamil Aziz (muzzamil.aziz@gwdg.de) is a scientific researcher at the eScience Group, GWDG. His areas of expertise are software defined networking, cloud computing, and distributed systems. He is currently involved in the NEPHELE European project to lead the design of an SDN control plane for advanced data centers. Throughout his career, he has worked in various R&D projects for the telecommunication industry related to machine-to-machine communication and next generation networks. Recently, he received his Ph.D. from RWTH University, Aachen, Germany.

Eitan Zahavi (eitan@mellanox.com) is a distinguished architect and co-founder of Mellanox. He leads the Mellanox network architecture group focusing on cluster-level performance. He also acts as a Co-Chair of the IBTA Management working group. He teaches logic design automation for VLSI systems in the Technion Electrical Engineering Department. He received his Ph.D. on forwarding in compute clusters in 2015 and graduated in 1987, from/at the Electrical Engineering Department of the Technion, Israel institute of technology.

Domenico Gallico (domenico.gallico@interoute.com) has a degree in telecommunication engineering from Universitá degli Studi di Pisa. Currently he works as a project manager at Interoute and he is involved in EC Projects for the Interoute R&D Department. The main focus is on virtualization and the cloud environment. Currently he is involved in the FP7 COSIGN Project, and H2020 COGNET, CYCLONE, PRISMACLOUD, and NEPHELE. Previously he worked at Hermes Trade s.r.l as an IT consultant for GE.

Richard Pitwon [M] (richard.pitwon@seagate.com) received his B.Sc. and M.Sc. from the University of St. Andrews and is a Chartered Engineer. He currently leads the photonics research and development group at Seagate UK, and his research interests include optical circuit boards, photonic integrated circuits, and passive and active optical connectors and transceivers. He holds 46 patents and has authored 34 publications in this field, one book chapter, and two international standards.

Konstantinos Tokas (ktok@mail.ntua.gr) obtained his Diploma in electrical and computer engineering from NTUA. His thesis was carried out at the Photonics Communications Research Laboratory (PCRL) and concerned high-speed optical interconnects using advanced modulation schemes. He is currently a member and Ph.D. candidate of PCRL, and his research activities include, among others, optical data center network sub-systems and architectures.

Ioannis Patronas (johnpat@phys.uoa.gr) received a B.S. from the Physics Department and M.S. (2015) in computing and control from the Physics Department of the National & Kapodistrian University of Athens, Greece, where he is currently a Ph.D. candidate writing a thesis, *Parallel Processing and Architectures for Nodes and Elements of Optical Networks and Data Centers*.His research interests include parallel algorithms and architectures, mapping techniques, real-time systems, optimization of networks protocols, cloud computing, and data centers.

Marco Capitani (m.capitani@nextworks.it) received his M. S. in mathematics from the University of Pisa in 2016, with a curriculum heavily invested in mathematical logic. Currently, he is a software engineer at Nextworks. His research is mainly focused on SDN. He is active in the European projects H2020 NEPHELE, 5G Crosshaul, and ARCFIRE.

Christos Spatharakis (cspatha@mail.ntua.gr) received his Diploma in electrical and computer engineering from NTUA in 2010 and his M.Sc. in analog and digital IC design from Imperial Col-

The European project NEPHELE is developing an end-to-end optical infrastructure for scale-out, disaggregated datacenters. Efforts are focusing toward the development of scalable optical network implementations that are compatible with the characteristics and limitations of current photonic technologies, enabling rapid deployment.

lege London in 2011. He is a member of the Photonics Communications Research Laboratory (PCRL) pursuing his Ph.D., and his main research activities include FPGA design and development, digital signal processing for coherent optical transceivers, and multiformat optical QAM signal generation for telecom systems. His work in these fields is carried out through active participation in the FP7 ICT-SPIRIT, and H2020 ORCHESTRA and NEPHELE European projects. He has authored or co-authored more than 20 publications in IEEE and OSA peer-reviewed journals and conferences.

Konstantinos Yiannopoulos [S'03, M'05] (kyianno@uop.gr) received a Diploma and a Ph.D. in electrical and computer engineering from NTUA. He has extensively studied the physical and network layers of optical networks, and is currently serving as an assistant professor at the University of Peloponnese, Greece, with a research focus on optical wireless communications. He has published over 60 papers in peer-review journals/conferences and has received over 500 independent citations.

Kai Wang received his M.Sc. in microelectronic systems and telecommunications from Liverpool University and his Ph.D. degree in optical engineering from University College London (UCL), United Kingdom. He became a research fellow in 2004 in the Department of Electronic and Electrical Engineering, UCL. His research interests include computer modeling in LCDs, backlights, multimode waveguides, and design OPCBs. He joined Xyratex Technology Ltd. (acquired by Seagate Technology PLC in March 2014) as a senior engineer in 2012 and worked on the design and characterization of optical backplane and optical interconnects for next generation storage and server systems.

Konstantinos Kontodimas (kontodimas@ceid.upatras.gr) received a Diploma and M.Sc. from the Department of Computer Engineering and Informatics, University of Patras. He is currently a Ph.D. candidate in the School of Electrical and Computer Engineering, NTUA. His research interests are in the areas of performance evaluation of interconnection networks and optimization.

Ioannis Lazarou (ilazarou@mail.ntua.gr) received a B.Sc. and an M.Sc. degree from the Computer Science Department of Aristotle University of Thessaloniki in 2008 and 2010, respectively. In 2015, he received his Ph.D. degree from NTUA focusing on advanced modulation formats for next-generation optical networking. Currently, he is a senior researcher at the Photonics Communications Research Laboratory of NTUA, specializing in high-speed optical telecommunication concepts for telecom and datacom networks

Philipp Wieder (philipp.wieder@gwdg.de) is leading the eScience group at GWDG, Germany. His research interests span from scheduling over data management to distributed systems. He was and is involved in multiple national and European projects. He received his M.S. from RWTH Aachen and his Ph.D. from TU Dortmund University.

Dionysios I. Reisis (dreisis@phys.uoa.gr) received a B.S. from the Electrical and Computer Engineering Department of the University of Patras, and his M.S. and Ph.D. in computer engineering from the Electrical Engineering Department of the University of Southern California in 1989. His research interests include parallel processing, embedded systems, and applications in image, video, and signal processing. He is an associate professor in the National & Kapodistrian University of Athens.

Emmanouel (Manos) Varvarigos (manos@ceid.upatras.gr) received a Diploma in electrical and computer engineering from NTUA in 1988, and M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology in 1990 and 1992, respectively. He has held tenured faculty positions at the University of California, Santa Barbara, Delft University of Technology, the University of Patras, NTUA, and Monash University. His research activities are in optical networking, data centers, and smart grids.

Matteo Biancani (matteo.biancani@interoute.com) has a degree in telecommunications engineering from Universitá degli Studi di Pisa). Currently he is a sales director at Interoute responsible for enterprise business in Italy. He looks after the organization of sales and sales engineering in order to develop and grow markets and to improve the quality of the support provided to sales forces. In addition to customer/market related activities, he is deeply involved in Interoute's R&D initiatives on SDN, cloud, data centers, and RINA. He is coordinator for the H2020 CYCLONE projects. Previously, he was coordinator of the FP7 projects GEYSERS, LIGHTNESS, and DOLFIN in which he worked on cloud and network integration. Previous job experiences: Telecom Italia/IT Department; Netesi SpA (application service provider). Publication: "All-Optical Packet/Circuit Switching-Based Data Center Network for Enhanced Scalability, Latency and Throughput," *IEEE Network* Special Issue on Optical Networks in Cloud Computing, published December 2013.

Hercules Avramopoulos [M] (hav@mail.ntua.gr) received his B.S. in physics, M.S. in applied optics, and Ph.D. for studies of nonlinear effects in lasers and optical fibers from Imperial College London. He is a full professor at the School of Electrical and Computer Engineering, NTUA. He also heads the PCRL (www.photonics.ntua.gr), which he founded 20 years ago. From 1989 to 1993 he joined the Optical Computing/Digital Optics Research Department at AT&T Bell Labs, Holmdel, New Jersey, where he worked on optical signal processing. His research interests include photonic integration, optical interconnects, and optical coherent communication systems. He has authored or co-authored more than 350 peer-reviewed articles and conference papers and holds two patents. He has served on several panels, including committees for research program definition and as an evaluator for the European Commission. Currently, he is a member of the Technical Program Committee of the European Conference in Photonic Communications.