All-Optical Packet/Circuit Switching-Based Data Center Network for Enhanced Scalability, Latency, and Throughput

Jordi Perelló, Salvatore Spadaro, Sergio Ricciardi, and Davide Careglio,
Universitat Politècnica de Catalunya
Shuping Peng, Reza Nejabati, George Zervas, and Dimitra Simeonidou, University of Bristol
Alessandro Predieri, and Matteo Biancani, Interoute
Harm J. S. Dorren, Stefano Di Lucente, Jun Luo, and Nicola Calabretta,
Technical University of Eindhoven
Giacomo Bernini, and Nicola Ciulli, Nextworks
Jose Carlos Sancho, Steluta Iordache, Montse Farreras, and Yolanda Becerra,
Barcelona Supercomputing Center
Chris Liou, and Iftekhar Hussain, Infinera
Yawei Yin, Lei Liu, and Roberto Proietti, University of California, Davis

Abstract

Applications running inside data centers are enabled through the cooperation of thousands of servers arranged in racks and interconnected together through the data center network. Current DCN architectures based on electronic devices are neither scalable to face the massive growth of DCs, nor flexible enough to efficiently and cost-effectively support highly dynamic application traffic profiles. The FP7 European Project LIGHTNESS foresees extending the capabilities of today's electrical DCNs through the introduction of optical packet switching and optical circuit switching paradigms, realizing together an advanced and highly scalable DCN architecture for ultra-high-bandwidth and low-latency server-to-server interconnection. This article reviews the current DC and high-performance computing (HPC) outlooks, followed by an analysis of the main requirements for future DCs and HPC platforms. As the key contribution of the article, the LIGHTNESS DCN solution is presented, deeply elaborating on the envisioned DCN data plane technologies, as well as on the unified SDN-enabled control plane architectural solution that will empower OPS and OCS transmission technologies with superior flexibility, manageability, and customizability.

Data centers (DCs) are currently the largest closed-loop systems in the information technology (IT) and networking worlds, continuously growing toward multi-million-node clouds [1]. DC operators manage and control converged IT and network infrastructures in order to offer a broad range of services and applications to their customers. Typical services and applications provided by current DCs range from traditional IT resource outsourcing (storage, remote desktop, disaster recovery, etc.) to a plethora of web applications (e.g., browsers, social networks, online gaming). Innovative applications and services are also gaining momentum to the point that they will become main representatives of future DC workloads. Among them, we can find high-performance computing (HPC) and big data applications [2]. HPC encompasses a broad set of computationally intensive scientific applications, aiming to solve highly complex problems in the areas of quantum mechanics, molecular modeling, oil and gas exploration, and so on. Big data applications target the analysis of massive amounts of data collected from people on the Internet to analyze and predict their behavior.

All these applications and services require huge data exchanges between servers inside the DC, supported over the DC network (DCN): the intra-DC communication network. The DCN must provide ultra-large capacity to ensure high throughput between servers. Moreover, very low latencies are mandatory, particularly in HPC where parallel computing tasks running concurrently on multiple servers are tightly interrelated. Unfortunately, current multi-tier hierarchical tree-based DCN architectures relying on Ethernet or Infiniband electronic switches suffer from bandwidth bottlenecks, high latencies, manual operation, and poor scalability to meet the expected DC growth forecasts [3].

These limitations have mandated a renewed investigation

0890-8044/13/$25.00 © 2013 IEEE

Architecture	Year	Elect./opt.	Circuit/packet	Scalability	Cap. limit	Prototype
c-Through	2010	Hybrid	Hybrid	Low	Tx/Rx	Yes
Helios	2010	Hybrid	Hybrid	Low	Tx/Rx	Yes
Proteus	2010	All-optical	Circuit	Medium	Tx/Rx	Yes
LIONS	2012	All-optical	Packet	Medium	TWC, AWGR	Yes
Petabit, IRIS	2010	All-optical	Packet	Medium	TWC, AWGR	No
Cascaded AWGRs	2013	All-optical	Circuit	High	TWC, AWGR	Yes
OSMOSIS	2004	All-optical	Packet	Low	SOA	Yes
Bidirectional	2009	All-optical	Packet	Low	SOA	Yes
Data Vortex	2008	All-optical	Packet	Low	SOA	Yes
Optical-OFDM	2012	All-optical	Packet	Medium	WSS	Yes

Table 1. *Summary of optical DCN solutions.*

into the introduction of ultra-high-bandwidth and low-latency optical technologies in the DCN, and the application of the control and management concepts from core/metro networks to also automate DCNs. Following these trends, LIGHTNESS [4], a European Framework Programme 7 (FP7) project started in November 2012, has the objective to design, implement, and experimentally demonstrate a high-performance DCN infrastructure for future DCs employing innovative optical switching and transmission solutions. Harnessing the power of optics will enable DCs to effectively cope with the unprecedented workload growth generated by emerging applications and services. With dense wavelength-division multiplexing (DWDM), more than 100 wavelength channels per fiber operating at 10, 40, 100 Gb/s, and beyond are feasible. This results in "unlimited" bandwidth capacities of multiple terabits per second per fiber link, which can be efficiently utilized through the combination of optical packet switching (OPS) and optical circuit switching (OCS) paradigms in the DCN data plane. A novel top-of-rack (ToR) switch will be designed and prototyped within LIGHTNESS to intelligently connect computing servers to the hybrid OPS/OCS DCN.

In this article, we review current solutions for DCs and HPC platforms, also surveying state-of-the art DCN solutions in the literature. Next, we highlight the main requirements that DCs and HPC platforms will have to address in the short- and mid-term future. These serve as motivations for the LIGHTNESS DCN solution. We describe the envisioned node architectures and technologies for the all-optical DCN data plane and the unified SDN-enabled control plane to provide the required flexibility, manageability, and customizability.

Current Solutions for Data Centers and HPC

HPC platforms address grand scientific challenges involving huge data sets and complex calculations, imposing high demands on CPU, memory, and communication resources. Providing a fast and low-latency network that interconnects the computational resources is required to scale HPC supercomputers to millions of computing cores, which can be needed to solve particularly complex scientific problems. Big data applications have also gained in popularity recently due to the vast amounts of information generated over the Internet. These applications fit parallel execution over distributed environments well, making use of the available resources in the DC. However, they require managing huge amounts of data moving across the network to allow fast computation before a completion deadline occurs. In this context, several solutions have emerged to perform task scheduling and load balancing to avoid network contention (e.g., MapReduce), a distributed file system to distribute data across nodes (e.g., Hadoop), and so on.

Inside DCs, operators are currently moving toward new architectures based on hierarchical clustering of thousands of virtualized servers arranged in racks around a converged DCN. Virtualization mechanisms and technologies (e.g., VMWare, Xen) are implemented to efficiently multiplex customers across physical servers. Such virtualization of DC resources and infrastructure is offered by DC and cloud operators as a service to their customers, to provide on-demand computing and cloud hosting with integrated applications.

Current Data Center Network Architectures and Technologies

In order to accommodate the vast amount of aggregated bandwidth between hundreds of thousands of servers inside DCs, the DCNs must be carefully designed to meet the following requirements:

- Scalability
- Agility
- Fault tolerance
- Self-optimization
- Cost effectiveness
- Power efficiency
- Backward compatibility
- The ability to provide high end-to-end throughput and low latency

The typical DCN architecture is based on a *two-tier* or *three-tier* hierarchical topology. In the two-tier DCN architecture, servers are arranged into racks describing the tier-one network, while the tier-two network is composed of switches providing server-to-server connectivity. Larger modern DCs commonly deploy three-tier DCN architectures, including core, aggregation, and access layers [3]. As the size and complexity of DCs continue to grow, however, scaling out the DCN infrastructure becomes challenging. The tree-like network topology has inherent disadvantages that cause bottlenecks in latency and bandwidth. To address this, a flattened

network infrastructure providing high bandwidth and low latency is desirable for next-generation DCNs. Optical DCN solutions have recently drawn attention due to their potential for providing high bandwidth and low latency. Table 1 summarizes the existing DCN architectures with a hybrid/all-optical data plane, with their main features and prototype availability.

c-Through (G. Wang *et al.*, 2010) and **Helios** [5] are the two major representatives of hybrid optical/electrical switching networks. c-Through adopts a hybrid packet and circuit switched (HyPaC) DCN architecture where ToR switches are connected to Ethernet and an optical circuit-based network. Similarly, Helios brings optical micro-electro-mechanical systems (MEMS) switches and WDM links into DCs, and integrates them with existing DC infrastructures. It uses existing commercially available optical modules and transceivers for optical communication. However, its main drawback concerns the inherent limitations of electronics. **Proteus** (A. Singla *et al.*, 2010) is an all-optical architecture based on wavelength selective switch (WSS) switching modules and MEMS, establishing direct optical connections between ToR switches for high-volume connections, or multihop connections in case of low-volume traffic. The slow reconfiguration time of MEMS limits its flexibility, though. The architectures of **LIONS** (previously named **DOS**) (Y. Yin *et al.*, 2012), **Petabit** (J. Chao *et al.*, 2010), and **IRIS** (J. Gripp *et al.*, 2010) are all based on arrayed waveguide grating routers (AWGRs) and tunable wavelength converters (TWCs). LIONS relies on a single-stage AWGR with multiple fast tunable transceivers per input port, and uses electrical loopback buffer or optical negative acknowledgment (NACK) technologies to handle packet contention. Conversely, the Petabit and IRIS projects employ multi-stage AWGR switching architectures. Specifically, Petabit adopts a three-stage Clos network, where each stage consists of an array of AWGRs used for the passive packet routing, while the IRIS three-stage switching network consists of two stages of partially blocking space switches and one stage of time switch that contains an array of optical time buffers. Both Petabit and IRIS are reconfigurable non-blocking switches. In addition, the 448 × 448 optical crossconnect (OXC) prototype and the 270 × 270 OXC with cascaded AWGRs (X. Ye *et al.*, 2012) that are bridged by delivery-and-coupling switches and WDM couplers shows the feasibility of modularizing AWGR switches. Another category is based on semiconductor optical amplifier (SOA) devices. The **OSMOSIS** (R. Hemenway *et al.*, 2004) switch is based on a broadcast-and-select (B&S) architecture using couplers, splitters, and SOA broadband optical gates. B&S architectures are very power-inefficient, though, since most of the signal power is broadcasted and blocked. Data Vortex (A. S. O. Liboiron-Ladouceur *et al.*, 2008) and Bidirectional (A. Shacham *et al.*, 2009) are both based on SOA 2 × 2 switching elements connected in a Banyan network. The Data Vortex topology is a fully connected directed graph with terminal symmetry. Its major advantage is that the single-packet routing nodes are distributed and do not require centralized arbitration. However, its significant complexity and latency do not favor scalability. The Bidirectional project overcomes the scaling limitation of Data Vortex. By using bidirectional SOAs, scaling to a large number of nodes is only constrained by the total required latency and congestion management. Scaling is not accompanied by an increased number of modules, thus achieving low power consumption. Finally, optical orthogonal frequency-division multiplexing (OFDM) technologies have also been brought into DCNs. Examples of this are the optical OFDM-based architectures proposed in C. Kachris *et al.*, 2012, and P. N. Ji *et al.*, 2012. Such architectures provide high spectral efficiency and fine-grained bandwidth allocation, but demand complex optical technologies.

DCN Control and Management

DCNs are mostly managed by DC operators as independent pools of resources to be bound to IT services and applications running in the servers. Current DCN control and management platforms focus on supporting efficient operations and management of DC fabrics by provisioning secured connectivity services, and proactively monitoring the DCN to detect performance degradation and failures. The aim is to provide visibility and control of the DCN infrastructure through a single management point, and ease the diagnosis and troubleshooting of DC outages.

These operations are commonly performed by deploying static or semi-automated control and management procedures where most actions, like DC network security or service provisioning, are performed by semi-automated procedures with human supervision and validation. However, DCs are becoming increasingly complex and massive, mainly due to the new emerging virtualization technologies, which add further levels of complexity while enabling higher workloads to be accommodated in the DCN. The management of such virtualized infrastructures (at both the IT and network levels) more than ever requires dynamicity, flexibility, and availability, as pointed out in the next section.

Requirements for Future Data Centers and HPC

Future DCs are required to provide more powerful IT capabilities, higher intra-DC bandwidth and energy efficiency, smaller time to market for new services to be deployed, and all these at lower cost. More specifically, DCs are expected to provide high flexibility and scalability not only in terms of computing and storage resource utilization, but also in terms of network infrastructure design and operation, including disaster recovery and security functions. Flexibility is already critical in present-day DC environments and will be imperative in the years to come. First, IT and network demands can vary depending on the specific services and applications, also fluctuating during the hours of the day, the day of the week, or specific business cycles (e.g., streaming of big sports events). Moreover, DCs also have to cope with more long-term variations such as customer growth and deployment of new IT services. Furthermore, provisioning and reservation of IT and network resources must be on-demand, letting customers access their resources when needed. This enhances resource utilization in the DC and allows pay-per-use accounting of customer resource usage. The joint optimization of converged IT and network resources infrastructures will also allow next generation DCs to provide business continuity to their customers.

Regarding the DCN, it will have to scale up without compromising performance or adding complexity. At the same time, maximizing the bandwidth capacity and throughput while minimizing the end-to-end latency will become tight requirements. To achieve the desired flexibility, currently static and manual control and management of the DCN will have to evolve toward automated solutions, able to dynamically and efficiently recover from network failures.

Focusing on HPC, it is expected that by 2018 supercomputers will have to achieve exascale performance (1000 times higher than todays' performance) by harnessing the power of millions of cores. This will exacerbate HPC technological demands, requiring network latencies below the microsecond and per-processor communication of tens of gigabytes per sec-

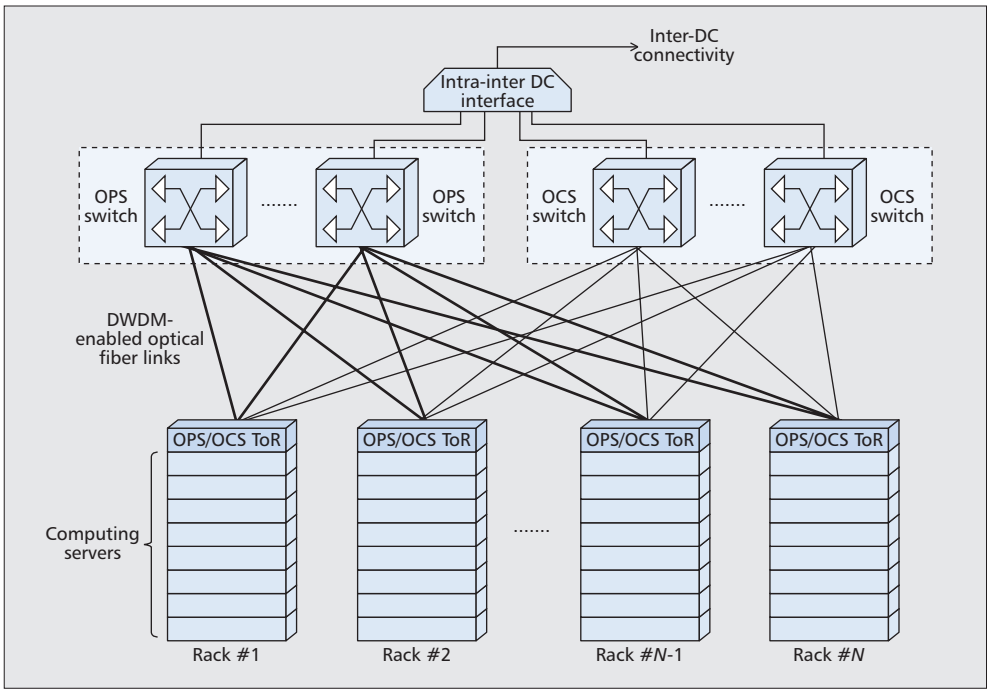

Figure 1. *LIGHTNESS approach: flattened DCN architecture.*

ond in the best cases (weak scaling mode). In the worst cases (strong scaling), critical network latencies below the nanosecond will be demanded, as well as per-processor communication of hundreds of gigabytes per second. To meet these figures, high-bandwidth network links, low-latency packet switching devices, and minimized DCN diameter are mandatory. Low energy consumption and housing costs will also be main requirements for HPC supercomputers. Indeed, there is a strong effort for exascale systems to decrease under a picojoule the energy needed to move one bit across the system, only achievable by photonic interconnects. To reduce housing costs, minimizing the number of components composing the HPC platform (e.g., switches in the DCN) seems to be a must.

LIGHTNESS DCN Data Plane

This section presents the data plane architecture and technology for the DCN envisioned within LIGHTNESS. This architecture brings innovation into current DCNs through the introduction of a hybrid OPS/OCS switching data plane that allows flattening current multi-tier hierarchical architectures for enhanced scalability, throughput, and latency. Transparent OPS and OCS technologies also avoid expensive optical-electrical-optical (OEO) conversions, optical transceivers, and cables, reducing the energy consumption and cost of current electrical solutions. The enabling OPS and OCS switching fabrics will be prototyped and experimentally demonstrated during the project.

In the DC environment, applications generating long-lived smooth data flows between servers coexist with applications exchanging short-lived data flows with tight latency requirements (typically < 1 μs). Employing a single optical switching technology to handle both long-lived and short-lived traffic compromises between throughput, packet loss, latency, and buffer size [5]. While short-lived packets require fast switching, long-lived data flows can be handled efficiently by low-speed optical switches. Moreover, long-lived and short-lived data flows going through the same switch buffers can make the latter ones experience unacceptably long latencies. Therefore, LIGHTNESS considers the DCN architecture shown in Fig. 1, that is, a flattened architecture integrating OPS and OCS switching technologies. The design of OPS switches targets high port count and low latency, which is employed for switching short-lived packet flows at 40 Gb/s. Conversely, OCS switches aim at handling long-lived data flows at 100 Gb/s. Computing servers are interconnected to the hybrid OPS/OCS DCN via the ToR switch, which performs traffic aggregation and application-aware classification to either short- or long-lived traffic. Moreover, OPS and OCS switches are connected to the intra-inter DC interface, providing inter-DC connectivity when required.

Scalability becomes critical when designing the DCN. This is addressed in the LIGHTNESS flattened DCN through the deployment of multiple OPS and OCS switches interconnecting the racks in the DC, which allows the number of input/output wavelengths to/from each ToR to be increased when the port count of OPS or OCS switches becomes a limiting factor. For even higher scalability, the configuration of clusters of racks in the DC (e.g., ToR switches within a cluster interconnected over point-to-point links or simple optical switch fabrics) offloading OPS-OCS switches from intra-cluster traffic could be an option.

ToR Switch Technology

ToR switches are interconnected to OPS and OCS nodes in the DCN. As shown in Fig. 2, the ToR switch will be designed and implemented based on high-speed field programmable gate array (FPGA) platforms, optoelectronic transceivers, and optical systems. Software and hardware programming techniques will be developed to handle 100 Gb/s DC flows. In particular, framing and processing techniques will be adopted for minimum processing and switching, as well as traffic aggregation mechanisms for maximum capacity processing and switching. The traffic generated from servers will be parsed over standardized protocols (e.g., fiber channel, Ethernet), and then mapped on optical packets with the minimum possible processing delay to maintain ultra-low latency.

The ToR switch will also include the electronic control interfaces with the OPS and OCS node and the management interfaces with the unified SDN-based LIGHTNESS control plane (detailed later). Effective algorithms in terms of clock cycle requirements and optimal load balancing performance

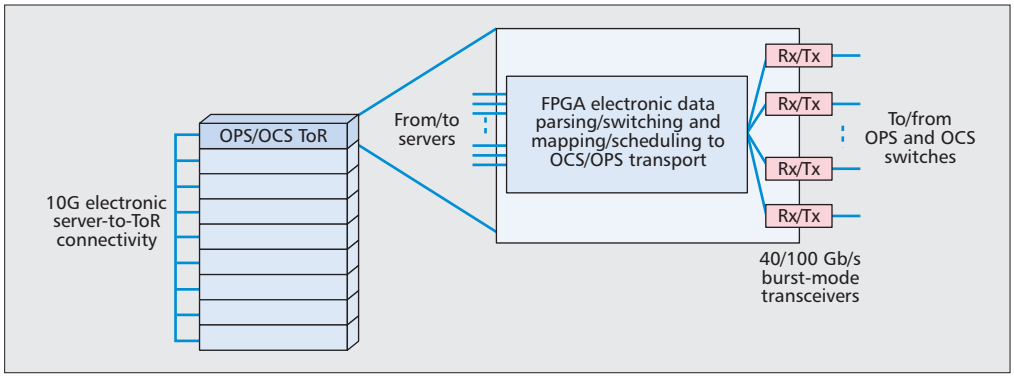

Figure 2. *ToR switch architecture for the hybrid OPS/OCS DCN.*

need to be investigated for implementing fast flow admission control operation for short-lived packets to eliminate optical packet contention. Burst mode transceivers at 40 and 100 Gb/s will be equipped to connect the ToR switch toward OPS and OCS switches, respectively. The flexible and programmable platform will be able to dynamically change the balance between intra-DC and inter-DC traffic processing and switching according to DC traffic demands.

OCS and OPS Switching Technologies

An OCS switch that scales to thousands of input/output ports allows interconnecting server racks in a flattened architecture. For example, MEMS switches or beam-steering switches can have reasonably high port counts and are capable of handling high-bit-rate data with low energy consumption, thus being the selected technology to implement the OCS data plane of the LIGHTNESS DCN, responsible for supporting long-lived data flows between servers.

Regarding the OPS technology, the expected features of the OPS fabric prototyped in LIGHTNESS are high port count (over 1000 ports) operating at 40 Gb/s with low end-to-end latency (< 1 μs) and high throughput. Scaling OPS switches over 1000 ports with low latency becomes challenging. Indeed, most OPS switches experimentally demonstrated so far are based on centralized control, with reconfiguration times increasing with the number of ports. Reference [6] reports that switch architectures with centralized control require at least $N \cdot \log 2N$ clock cycles to reconfigure the switch, N being the port count. Thus, if the number of ports exceeds a critical threshold, the switch is no longer capable of meeting the maximum allowed latency in the system.

LIGHTNESS introduces an OPS architecture with highly distributed control for port-count-independent reconfiguration time [6, 7]. This allows the proposed architecture, applied in short links with flow control, to provide a flattened interconnected network with sub-microsecond latency. Moreover, it enables high load operation while scaling to over 1000 ports. The investigated modular wavelength-division multiplexing (WDM) OPS Spanke-type switch architecture with distributed control is shown in Fig. 3 (top) [6, 7]. In the figure, the OPS switch has a total number of input ports $N = F \times M$, where F is the number of input fibers, each carrying M wavelength channels. Each of the M ToR switches has a dedicated electrical buffer queue and is connected to the OPS switch by optical fibers. Packets at each queue are electrical-optical (E-O) converted using burst mode opto-electronic interfaces. The OPS processes in parallel the N input ports by using parallel $1 \times F$ B&S switches, each with local control, and parallel $M \times 1$ wavelength selector contention resolution blocks (WSs), also with local control, enabling highly distributed control [6, 7]. Contentions occur only between the M input ports of each $M \times 1$ WS. Fixed wavelength converters (FWCs) located at the WSs output prevent contentions between packets destined to the same output fiber. Output fibers of the switch reach destination ToR switches through optical links, and positive flow control signals acknowledge the reception of packets. By implementing flow control between ToRs and the OPS switch, if multiple ToRs in the same input contend for the same OPS switch output, only one ToR is served, while the other ones are blocked and retransmitted. Upon retransmission, the OPS switch can successfully serve all buffered signals at the M ToRs thanks to the statistical multiplexing of the output resources. Preliminary experimental results on the realization of the presented OPS switch are reported in [8, 9]. As an example, an $F = 1$ $M = 4$ OPS switch laboratory prototype is illustrated in Fig. 3 (bottom).

Scaling the proposed OPS switch to over 1000 ports has to face some technological challenges. For example, a 1024 × 1024 OPS switch ($F = 32$ and $M = 32$) demands 32 optical modules (one per input fiber) including 32 1×32 photonic switches and 32 contention resolution blocks. For each of the 32 photonic switches, this translates into a 1×32 B&S switch (32 SOA gates) with 15 dB splitting losses. Regarding each contention resolution block, a 32-channel 100 GHz spaced AWG with crosstalk better than –30 dB and an FWC with broadband operation are needed. Although optical signal-to-noise ratio (OSNR) degradation might be an issue, the 15 dB splitting losses after the 1×32 B&S could be compensated by SOAs, while low-crosstalk AWGs and broadband FWC operation have already been demonstrated [8]. Therefore, the operation of a standalone 1×32 photonic switch and contention resolution block is challenging but technically feasible. However, the assembling of a single optical module requires more than 1024 SOAs and 32 FWCs. Clearly, building the overall OPS (32 optical modules) by using discrete optical components is impractical and demands photonic integration of the optical modules. Photonic integration of large chip size remains technologically demanding in fabrication and electrical wiring. Moreover, chip-coupling losses must also be considered. Improvements in photonic integrated technologies should allow the realization of large-scale photonic chips for scaling the optical modules with $M = 32$, and consequently assembling the $F = 32$ optical modules to realize a large-scale OPS switch.

Looking at the whole system, a 1024 × 1024 port OPS switch allows connectivity between 1024 ToR switches (one wavelength from each ToR toward the OPS switch). If wavelengths toward OPS operate at 40 Gb/s, we believe that 40 servers/rack are feasible, leading to medium-sized DCs of around 40,000 servers. Indeed, even in the worst and very infrequent case where all servers within a rack would demand inter-rack communication over OPS (note that OCS transport is also available), 1 Gb/s server–server in the ToR-to-OPS switch bottleneck is ensured. This speed can be increased by

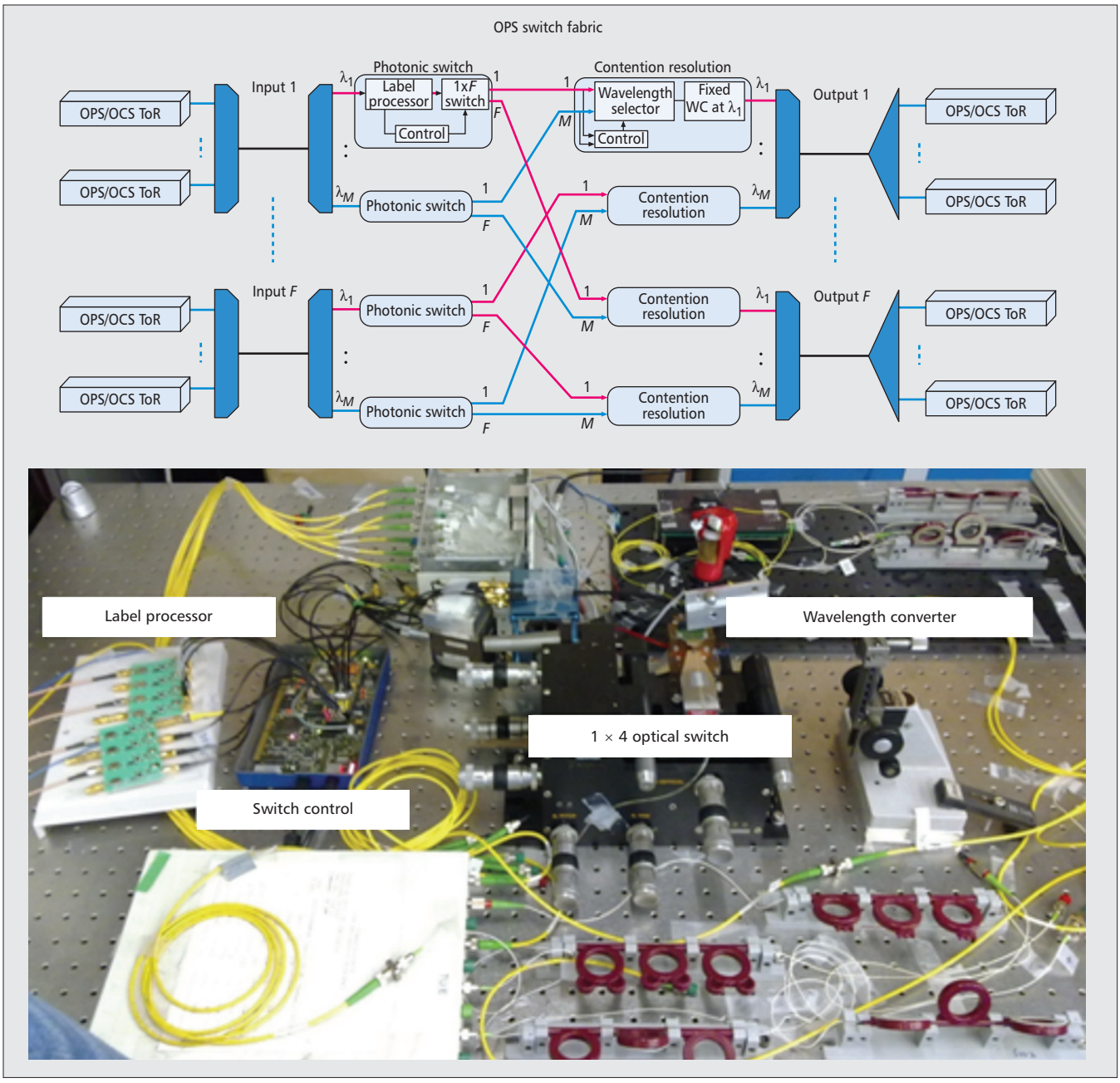

Figure 3. *OPS switch: architecture described as a block diagram (top); picture of an* F = *1* M = *4 OPS switch laboratory prototype built using discrete optical components (bottom).*

upgrading to 100 Gb/s, which would even allow increasing the number of servers per rack. For larger DC sizes, however, architectures with multiple OPS switches, as also depicted in Fig. 1, are necessary.

LIGHTNESS DCN Unified Control Plane

The unified network control plane allows current limitations of DCN management and control to be overcome. It arises as a means to implement automated procedures for setup, monitoring, recovery, and optimization of network connections spanning diverse optical technologies inside the DC (e.g., OPS and OCS), also matching the quality of service (QoS) requirements for IT services and applications. The unified network control plane aims to provide dynamic and flexible procedures to provision and reconfigure DCN resources, as well as to implement replanning and optimization functions according to performance and network usage statistics gathered from the DCN data plane. As a fundamental concept, it integrates functionalities offered by current DCN management frameworks to support on-demand provisioning of connectivity services, so as to substitute human actions and validations with automated procedures. Dynamicity and flexibility in the control framework also provide orchestration of long-lived and short-lived connectivity in the hybrid DCN optical data plane according to specific DC service requirements.

LIGHTNESS leverages software defined networking (SDN) concepts [10] to develop a unified control plane for advanced optical network connectivity in the DCN. Dynamicity, flexibility, and resiliency are combined to offer advanced bandwidth on-demand services to the DC providers. In addition, such a unified control plane is conceived to also support connectivity services among geographically distributed DCs, interconnected by means of optical core networks. A well-known choice to

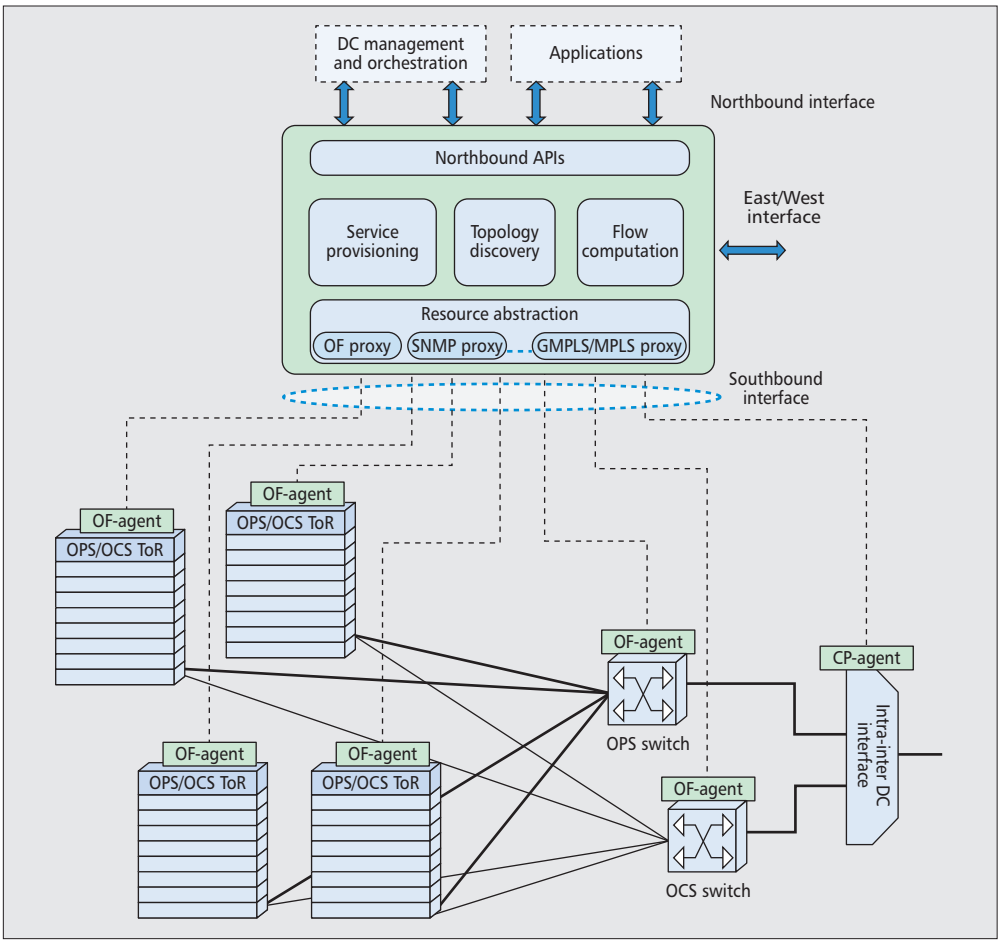

Figure 4. *Unified SDN-based control plane architecture.*

control such interconnection networks is the (generalized) multiprotocol label switching (GMPLS/MPLS) framework [11], which is a stable and mature protocol suite standardized by the Internet Engineering Task Force (IETF) for automatic provisioning of end-to-end connections in a fully distributed manner. Hence, a logical choice would be to equip the LIGHTNESS unified control plane with a flexible southbound interface to coordinate intra- and inter-DC dynamic connectivity (i.e., also data delivery among DCs) by interoperating with GMPLS/MPLS in the interconnection networks.

SDN-based Unified Control Plane Architecture

Current DCNs comprise a hierarchy of layer 2 and layer 3 technologies. However, as the size of traffic flows is exponentially increasing, DCNs are evolving to include optical network technologies used for handling long-lived and large data flows. DC operators need to deploy dynamic algorithms that can be continuously optimized over time for specific processes such as live virtual machine (VM) migration or load balancing. Vendor-specific control planes such as GMPLS/MPLS are neither open nor flexible enough (in terms of provisioning and routing procedures) for DC operators to deploy their proprietary and dynamic flow handling algorithms inside the DC.

SDN is defined as a control framework that supports the programmability of network functions and protocols by decoupling data and control planes, which are vertically integrated in most current network equipment. SDN can abstract the heterogeneous network technologies adopted inside DCs and represent them in a homogeneous way, which makes it a suitable candidate for the unified control plane of the DCN. OpenFlow (OF) [12] is an open standard vendor- and technology-agnostic protocol and interface that allows separating data and control planes. It is based on flow switching, and enables the execution of software/user defined flow-based routing, control, and management functions in an SDN controller decoupled from the data plane. In addition, to enhance the control plane flexibility, the SDN controller can support multiple protocols beyond OpenFlow for configuration and control of the DCN.

Therefore, an SDN-based control plane as depicted in Fig. 4 is an attractive solution for the control and management of the LIGHTNESS DCN, given its features:

- *Abstraction:* SDN can abstract the heterogeneous DCN resources (e.g., using OpenFlow) and be used to build a unified control plane over heterogeneous DCN technologies.
- *Programmability:* Flow computation is centralized inside the SDN controller, and flow tables can be programmed through an open northbound application programming interface (API).
- *Generic network controller:* SDN utilizes a generic network controller that can support any proprietary DC control, management, and orchestration application through the northbound interface.
- *Alliance:* SDN can benefit from well defined network functions and algorithms from the path computation element (PCE) framework standardized by the IETF [13], deployed as network applications on top of the SDN controller for enhanced routing performance. Moreover, leveraging the hierarchical PCE concept [14] can ease the control of connectivity services among remote DCs interconnected by optical core networks.

- *Interfacing:* The SDN controller can be equipped with flexible and extendable interfaces to interact with other control and management systems (e.g., at the southbound interface). If the inter-DC network is operated by a third-party service provider through GMPLS, programmable interfaces can be also be programmed to work as user-to-network interfaces (UNIs) to trigger inter-DC network services.

The unified SDN-enabled control plane is also conceived to easily communicate with the rest of the actors in the envisioned LIGHTNESS scenario, supporting the following interfaces and APIs:

- *Northbound:* Communicates with the upper-level centralized DC management and orchestration system, and processes on-demand connectivity service requests. It also allows the exchange of abstracted DCN-related information for monitoring and orchestration purposes. Moreover, it supports user-defined routing and path computation functions as applications on top of the controller for network intelligence and analytics.
- *Southbound:* Supports the communication with the DCN device controllers for the configuration and control of flows and connectivity services. It leverages the OF protocol, with proper optical extensions for OCS and OPS, to provide homogeneous management of the underlying DCN equipment, which hosts OF agents to process the incoming requests. Besides, the southbound interface is flexible enough to support vendor-specific interfaces, as well as further control/management protocols (e.g., Simple Network Management Protocol [15]) through the common resource abstraction layer. For inter-DC connectivity it interfaces with specific control plane agents (e.g., GMPLS/MPLS-based) at the DC boundaries.
- *East/West:* Enables cooperation with other SDN controllers potentially deployed in wide DCs for scalability purposes (e.g., each responsible for a specific technology segment or cluster of servers in the DC).

Conclusions

Emerging computationally and data-intensive applications to be served by current and future DCs are putting current DCN solutions in trouble, as they lack the flexibility, performance, and scalability needed to efficiently and cost-effectively deliver them. Leveraging the introduction of all-optical switching technologies inside the DC, LIGHTNESS aims at realizing a flexible and scalable DCN solution featuring ultra-high data throughput and low-latency server-to-server communication. Details of the LIGHTNESS data plane and unified control plane architectural solutions and technologies for future DCNs have been presented in this article, while highlighting their main features and benefits.

Acknowledgments

The authors would like to express their gratitude to Professor S. J. Ben Yoo from the University of California Davis for his valuable comments on the article. This work has been supported by the FP7 European Project LIGHTNESS (FP7-318606).

References

[1] ESG Research Report: Data Center Networking Trends, Jan. 2012.
[2] R. T. Kouzes *et al.*, "The Changing Paradigm of Data-Intensive Computing," *Computer*, vol. 42, no. 1, Jan. 2009, pp. 26–34.
[3] U. Hoelzle *et al.*, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, 1st ed., Morgan and Claypool, 2009.
[4] LIGHTNESS Project: http://www.ict-lightness.eu/
[5] N. Farrington *et al.*, "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," *ACM SIGCOMM 2010*.
[6] S. Di Lucente *et al.*, "Scaling Low-Latency Optical Packet Switches to a Thousand Ports," *J. Opt. Commun. and Net.*, vol. 4, no. 9, Sept. 2012, pp. A17–A28.
[7] H. J. S. Dorren *et al.*, "Scaling Photonic Packet Switches to a Large Number of Ports," *J. Opt. Commun. and Net.*, vol. 4, no. 9, Sept. 2012, pp. A82–A89.
[8] S. Di Lucente *et al.*, "Numerical and Experimental Study of A High Port-Density WDM Optical Packet Switch Architecture for Data Centers," *Optics Express*, vol. 21, no. 1, Jan. 2013, pp. 263–69.
[9] J. Luo *et al.*, "Low Latency and Large Port Count Optical Packet Switch with High Distributed Control," *Opt. Fiber Commun. Conf.*, Mar. 2012.
[10] "Software-Defined Networking: The New Norm for Networks," ONF white paper, Apr. 2012.
[11] E. Mannie *et al.*, "Generalized Multi-Protocol Label Switching (GMPLS) Architecture," IETF RFC 3945, Oct. 2004.
[12] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Comp. Comm. Rev.*, vol. 38, no. 2, Apr. 2008, pp. 69–74.
[13] A. Farrel, J. P. Vasseur, and J. Ash, "A Path Computation Element (PCE)-Based Architecture," IETF RFC 4655, Aug. 2006.
[14] D. King and A. Farrel, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS & GMPLS," IETF RFC 6805, Nov. 2012.
[15] D. Harrington, R. Presuhn, and B. Wijnen, "An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks," IETF RFC 3411, Dec. 2002.

Biographies

JORDI PERELLÓ (perello@ac.upc.edu) received his M.Sc. and Ph.D. degrees in telecommunications engineering in 2005 and 2009, respectively, both from the Universitat Politècnica de Catalunya (UPC), Spain. He is an assistant professor in the Computer Architecture Department at UPC. He has participated in FP-6 and FP-7 European research projects (LIGHTNESS, EULER, STRONGEST, DICONET, etc.). He has published more than 50 articles in international journals and conference proceedings. His research interests concern resource management and virtualization of future optical networks.

SALVATORE SPADARO (spadaro@tsc.upc.edu) received his M.Sc. (2000) and Ph.D. (2005) degrees in telecommunications engineering from UPC. He also received his Dr.Ing. degree in electrical engineering from Politecnico di Torino (2000). He is currently an associate professor in the Optical Communications group of UPC. He has participated in various European research projects (LIGHTNESS, EULER, STRONGEST, DICONET). His research interests focus on all-optical networks with emphasis on network control and management, resilience, and virtualization.

SERGIO RICCIARDI (sergio.ricciardi@ac.upc.edu) is a research associate in the Advanced Broadband Communications Center (CCABA) of UPC. He holds a Ph.D. in computer architecture from UPC, and has worked on several national and international projects (STRONGEST, LIGHTNESS, EULER, COST IC0804, CERN LHC ATLAS, etc.). His research interests are mainly focused on energy-aware RWA algorithms and protocols for telecommunication networks, and energy-oriented optimizations for grid and cloud computing.

DAVIDE CAREGLIO (careglio@ac.upc.edu) is an associate professor in the Department of Computer Architecture at UPC. He received his M.Sc. and Ph.D. degrees in telecommunications engineering from UPC in 2000 and2005, respectively, and his Dr. Ing. degree in electrical engineering from Politecnico di Torino in 2001. His research interests are in the field of network protocol and algorithm design for traffic engineering and quality of service provisioning.

SHUPING PENG (Shuping.Peng@bristol.ac.uk) received her B.S. in physics and Ph.D. in communications and information systems from Peking University. She is working as a research fellow with the High Performance Networks group at the University of Bristol. She is involved in multiple international, EU, and U.K. projects. Her research interests include network virtualization, network modeling, and algorithm design. She is an author and co-author of over 50 papers, and has served as TPC member and Session Chair of several IEEE/ACM conferences.

REZA NEJABATI (Reza.Nejabati@bristol.ac.uk) is a lecturer at the University of Bristol. His research interest is on the application of high-speed network technologies, design and control of software-defined, service-oriented, and programmable networks, cross-layer network design, and network architecture and technologies for e-science and cloud computing. He is an author and co-author of over 150 papers and three standardization documents. He is involved in several national/international projects. He has served as TPC member, Chair, and organizer of several IEEE conferences and workshops.

GEORGE ZERVAS (georgios.zervas@bristol.ac.uk) is a lecturer at the University of Bristol. He received his M. Eng. and Ph.D. degrees from, and was a research fellow and lecturer at the University of Essex. He has participated in more than

15 EC/U.K. projects. His research interests include flexible, multi-dimensional, programmable, and cognitive optical networks. He is a co-author of over 120 publications, and has served as TPC member of OFC, ACP, ONDM conferences. He has been involved in IETF and OGF and filed two patents.

DIMITRA SIMEONIDOU (dimitra.simeonidou@bristol.ac.uk) is a professor in the High Performance Networks group at the University of Bristol. She is a leading researcher in optical networks, future Internet research and experimentation (FIRE), and grid and cloud computing, and a founder of transport SDN. She has chaired a number of international conferences and committees across these technical fields. She is the author and co-author of over 350 publications, of which many have received best paper awards, and 11 patents and standards.

ALESSANDRO PREDIERI (Alessandro.Predieri@interoute.com) has a degree in telecomunications engineering (Politecnico di Milano). Currently he is a senior sales engineer at Interoute. He is supporting a sales force, involved in the entire process, including client contact, problem definition, technical and economic proposal writing, data research and analysis, definition of project costs, related P&L, delivery coordination, and troubleshooting of the projects. He is also currently involved in R&D activities, contributing to the FP7 LIGHTNESS and SmartenIT projects.

MATTEO BIANCANI (Matteo.Biancani@interoute.com) has a degree in telecommunications engineering (Università degli Studi di Pisa). He is the sales director at Interoute responsible for enterprise business in Italy: he looks after the organization of sales and sales engineering to develop and grow the market. He is involved in Interoute R&D initiatives, starting six years ago with the Digital Divide Satellite Offer funded by ESA. He was coordinator of the FP7 GEYSERS project, and is currently coordinating the FP7 LIGHTNESS project.

HARM S. J. DORREN (H.J.S.Dorren@tue.nl) is a professor and serves as the scientific director of COBRA. He joined Eindhoven University of Technology, the Netherlands, in 1998. In 2002 he was also a visiting researcher at the National Institute of Industrial Science and Technology (AIST) in Tsukuba, Japan. His research interests include optical packet switching, digital optical signal processing, and ultrafast photonics. He has coordinated several EU projects and (co)authored over 280 journal papers, and served as associate editor for the *IEEE Journal of Quantum Electronics*.

STEFANO DI LUCENTE (s.di.lucente@tue.nl) received his B.Sc. and M.Sc. degrees in electronic engineering from Roma Tre University, Italy, in 2006 and 2009, respectively. He is currently working toward his Ph.D. degree in the Electro-Optical Communication group, Eindhoven University of Technology. His research activities are focused on optical packet switching, optical labeling techniques, and optical packet switch control systems.

JUN LUO (j.luo@tue.nl) received his Ph.D. degree in optical communications from Tianjin University, China, in 2012. From 2010 to 2011 he was a visiting researcher at COBRA Research Institute, Eindhoven University of Technology, where he is currently working as a postdoctoral researcher. His research interests include optical packet switching and high-speed optical signal processing, and silicon photonics.

NICOLA CALABRETTA (N.Calabretta@tue.nl) received his M.S. degree in telecommunications engineering from Politecnico di Torino in 2000, and his Ph.D. degree from the COBRA Research Institute in 2004. He is currently with COBRA Research Institute. His fields of interest are all-optical signal processing for optical packet switching, semiconductor-based photonic integrated devices, advanced modulation formats, optical interconnects, and high-performance optical networks.

GIACOMO BERNINI (g.bernini@nextworks.it) received his Italian Laurea degree in telecommunication engineering from the University of Pisa in 2006. Currently he is R&D project manager at Nextworks. His research interests include SDN, cloud computing, cloud-to-network interface, NSI, ASON/GMPLS control plane, and PCE framework. He participated in design, development and demonstration activities in FP6 PHOSPHORUS, and FP7 GEYSERS, ETICS, and MAINS projects, as well as industrial projects. He is currently active in the FP7 LIGHTNESS, FP7 TRILOGY 2, and FP7 CONTENT projects.

NICOLA CIULLI (n.ciulli@nextworks.it), head of R&D activities at Nextworks, got a degree in telecommunication engineering in 1997 from the University of Pisa, and his Diploma degree from SSSUP S. Anna. His research and industrial activities focus on GMPLS control and management planes architectures for SDH networks. He has participated in several FP5, FP6, and FP7 projects (PHOSPHORUS, GEYSERS, ETICS, MAINS, CHANGE, etc.) and industrial projects (e.g., the Marconi ASTN/GMPLS project, where he coordinated the activities of Nextworks; and Alcatel projects on 4G mixed packet/TDM switches and T-MPLS).

JOSE CARLOS SANCHO (jose.sancho@bsc.es) received his Ph.D. degree in computer science from the Technical University of Valencia, Spain, in 2002. His Ph.D. thesis was focused on improving the performance of interconnection networks. In 2003, he joined Los Alamos National Laboratory, New Mexico, where he conducted research on fault tolerance and communication overlap. Currently, since 2010, he is a senior researcher at the Barcelona Supercomputing Centre, Spain, conducting research on interconnection networks for data centers.

STELUTA IORDACHE (steluta.iordache@bsc.es) is a postdoctoral researcher with the Department of Computer Sciences at Barcelona Supercomputing Center, where she is working on interconnect networks for supercomputer architectures. She holds a Ph.D. degree from UPC (2011) and an engineering degree from University Politehnica of Bucharest, Romania (2005). She has participated in several EU-funded projects: LIGHTNESS, NOVI, GEANT, GEYSERS, and N-CRAVE.

MONTSE FARRERAS (montse.farreras@bsc.es) received her Ph.D. degree in computer science at UPC (2008). She works as a collaborator professor at UPC, and she joined the Programming Models research line at BSC. In this research group she is conducting research about parallel programming models for high-performance computing, focusing on productivity, performance, and scalability. She has been collaborating with the Programming Models and Tools for Scalable Systems group at IBM T. J. Watson Research Institute since 2004.

YOLANDA BECERRA (yolanda.becerra@bsc.es) received a Ph.D. in computer science in 2006 from UPC. She is a full-time collaborator professor at the Computer Architecture Department of UPC and an associate researcher at BSC. In 2007 she joined the Autonomic Systems and eBussines Platforms research line at the BSC. In this research group she is conducting research about resource management strategies for big data applications in the cloud.

CHRIS LIOU (cliou@infinera.com) is a Fellow and the vice president of network strategy at Infinera, where he focuses on network architecture and solutions for network and service providers worldwide. He has previously held product management positions at Ciena and Cisco, along with systems architecture and engineering positions at Hewlett-Packard and Telcordia. He received his B.S.E. with high honors in electrical engineering with a minor in operations research and financial engineering from Princeton University, and his M.S. in electrical engineering and computer science from Stanford University.

IFTEKHAR HUSSAIN (IHussain@infinera.com) is a principal software engineer at Infinera Corporation, where he is involved in design of IP/MPLS control and data plane system architecture of ultra-high-capacity packet and optical transport systems for data center and core network applications. He has a Ph.D. degree in electrical and computer engineering from the University of California, Davis.

YAWEI YIN (yyin@ucdavis.edu) received his B.S. degree in applied physics from the National University of Defense Technology (NUDT), Changsha, China, in 2004 and his Ph.D degree in electrical engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2009. He is currently with NEC Laboratories America, Inc., Princeton, New Jersey. From 2009 to 2013, he was with the Next Generation Networking Systems Laboratory, University of California, Davis. His research interests include optical switching networks for data centers and high-performance computers, elastic optical networking frameworks, and SDN control and management technologies.

LEI LIU (leiliu@ucdavis.edu) received B.E. and Ph.D. degrees from BUPT in 2004 and 2009, respectively. From 2009 to 2012, he was a research engineer at KDDI R&D Laboratories Inc., Japan, where he was engaged in research and development of intelligent optical networks and their control and management technologies. He is now with the University of California, Davis.

ROBERTO PROIETTI (rproietti@ucdavis.edu) received his M.S. degree in telecommunications engineering from the University of Pisa in 2004 and his Ph.D. in electrical engineering from Scuola Superiore Sant'Anna in 2009.He is a postdoctoral researcher with the Next Generation Networking Systems Laboratory, University of California, Davis. His research interests include optical switching technologies, architectures for supercomputing and data center applications, high-spectrum efficiency coherent transmission systems, and elastic optical networking.