Large Data Center Interconnects Employing Hybrid Optical Switching

Matteo Fiorani, Maurizio Casoni and Slavisa Aleksic

Large Data Center Interconnects Employing Hybrid Optical Switching

Matteo Fiorani and Maurizio Casoni
Department of Engineering "Enzo Ferrari"
University of Modena and Reggio Emilia
Via Vignolese 905, Modena, Italy
{matteo.fiorani, maurizio.casoni}@unimore.it

Slavisa Aleksic
Institute of Telecommunications
Vienna University of Technology
Favoritenstrasse 9-11/389, Vienna, Austria
slavisa.aleksic@tuwien.ac.at

***Abstract*—Current data centers (DCs) networks rely on electronic switching and point-to-point interconnects. When considering future DC requirements, point-to-point interconnects will lead to poor network scalability and large power consumption. For this reason several optical switched interconnects for DCs have been recently proposed. However, the proposed optical switching solutions suffer from low flexibility and are not able to provide service differentiation. Furthermore, very few studies evaluate possible improvements in energy efficiency offered by optical switching solutions. In this paper we introduce a novel architecture of interconnects for DCs based on hybrid optical switching (HOS). HOS combines three different optical switching paradigms, namely circuit, burst and packet switching within the same network. Furthermore, HOS envisages the use a two parallel optical switches, a slow and low power consuming switch for the transmission of data using circuits and long bursts, and a fast switch for the transmission of packets and short bursts. The possibility of choosing between circuits, bursts and packets ensures the flexibility required by future DCs. At the same time, the option to select the most suitable switch technology for each data flow guarantees high transmission efficiency and low power consumption.**

I. Introduction

A data center (DC) refers to any large, dedicated cluster of computers that is owned and operated by a single organization. The exponential growth of the Internet traffic, mainly driven by emerging applications such as social networking and cloud computing, has created the need for more powerful DCs. Over the years DCs have evolved to include tens of thousands of servers in a single facility to support data-intensive and computing-intensive applications. These applications pose a significant challenge to the networking of the DCs creating the need for highly efficient interconnection systems. The interconnection network of future DCs must provide high flexibility, high throughput, low latency and low power consumption. A current DC consists of multiple racks hosting the servers connected through the DC interconnection network [1]. Each rack contains 20 to 80 servers, usually in the form of blades. The blade servers can be either web servers, application servers or databases. Current DCs networks are usually organized in a fat-tree 3-Tier architecture. The 3 tiers of a DC network are the edge tier, the aggregation tier and the core tier. The edge tier consists of the Top-of-Rack (ToR) switches that connect the servers to the DCs network fabric. The aggregation tier consists of aggregation switches that interconnect the ToR switches in the edge layer. Finally, the core tier consists of core routers that connect the DC to the wide area network (WAN). Current DCs networks rely on electronic switching solutions and point-to-point interconnects. The electronic switching is realized by commodity switches that are interconnected using either electronic or optical point-to-point interconnects. Electrical point-to-point interconnects suffer from high crosstalk as well as distance dependent attenuation due to dielectric losses coupled with high frequencies. As a consequence, very high data-rates over electrical interconnects can be hardly achieved. Because of the limited link data-rate, a large number of copper cables are required to interconnect a high-capacity DC, thereby limiting the scalability and increasing the power consumption of the interconnection network. It has been demonstrated [2] that current data centers networks based on copper cables consume around 23% of the total IT power, which corresponds to several MWs of power in large DCs operated by service providers. Optical transmission technologies are generally able to provide higher data rates over longer transmission distances than electrical transmission systems, leading to increased scalability and reduced power consumption. Hence, recent high-capacity DCs are increasingly relying on optical point-to-point interconnection links. According to an IBM study [3] only the replacement of copper-based links with VCSEL-based point-to-point optical interconnects can reduce the power consumption of the DC network by almost a factor of 6. However, the energy efficiency of point-to-point optical interconnects is limited by the power hungry electrical-to-optical (E/O) and optical-to-electrical (O/E) conversion required at each node along the network since the switching is performed using electronic packet switching.

When considering future DC requirements, architectures based on point-to-point interconnects and electronic switching will lead to poor network scalability, large latency and low power efficiency. In this context, optical switched interconnects that make use of optical switches and wavelength division multiplexing (WDM) technology, can be employed to provide high communication bandwidth while reducing significantly the power consumption. Especially in large DCs for cloud computing managed by online service providers such as Google, Microsoft, and Amazon, which host up to

100K servers, the use of power efficient interconnects is of paramount importance. It has been demonstrated in several research papers that solutions based on optical switching can improve both scalability and energy efficiency with respect to point-to-point interconnects [4,5] In the research literature several optical switched interconnect architectures for DCs have been recently presented [6-11] Some of the proposed architectures [6,7] are based on hybrid switching with packet switching in the electronic domain and circuit switching in the optical domain. The others are based on all-optical switching elements, and rely either on optical circuit switching [9,11] or on optical packet/burst switching [8,10]. Only a few of these studies evaluate the energy efficiency of the optical interconnection network and make comparison with existing solutions based on electronic switching [7,11], while the others are mainly focused on latency reduction. Furthermore, none of these studies analyze the flexibility of the proposed network solution, i.e. the capability of the proposed network to adapt to current traffic conditions in order to serve each DC application with the required service quality. With the continuous rise of new DC applications with different traffic characteristics, flexibility is becoming a primary requirement for future DC networks. The future DC network should be able to serve each application with the required service quality while achieving efficient resource utilization and low energy consumption. To achieve high flexibility, in telecommunication networks hybrid optical switching (HOS) approaches have been recently proposed [12,13]. HOS combines optical circuit, burst and packet switching on the same network and maps each application to the optical transport mechanism that best suits to its traffic requirements, thus enabling service differentiation directly at the optical layer. Furthermore, HOS envisage the use of two parallel optical switches, a slow optical switch for the transmission of circuits and long bursts and a fast optical switch for the transmission of packets and short bursts. Consequently, employing energy aware scheduling algorithms, it is possible to dynamically choose the best suited optical switching element while switching off or putting in low power mode the unused ones. In this paper we propose a novel DC network based on the HOS switching paradigm. We present the architecture of such a network and evaluate through simulations its performance in terms of data losses, latency and energy efficiency. We will demonstrate that HOS has potential for satisfying the requirements of future DCs.

The rest of the paper is organized as follows. In Section II we introduce the proposed HOS optical interconnect for DCs. In Section III we present the modeling approach for analyzing the proposed HOS network. In Section IV we present results in terms of average loss rates, average delays and energy consumption. Finally, in V conclusions are drawn.

II. HOS Interconnects for Data Centers

A. Traffic Characteristics

The proposed HOS interconnection network supports three different optical data types: circuits, bursts and packets.

A circuit is a long-lived optical path between source and destination servers established employing slow and low power consuming optical switching elements. We assume that circuits are time division multiplexed and employ a two-way reservation mechanism. Circuits are scheduled with highest priority within the HOS core switch and ensure no data losses, no jitter and very low latency. As a consequence, circuits are well suited for DC's applications with high service requirements and generating long-term point-to-point bulk data transfer, such as virtual machine migration and reliable storage. However, due to relatively high reconfiguration times, optical circuits introduce low flexibility and are not suited for applications generating bursty traffic.

Optical burst switching has been widely investigated in telecommunication networks for its potential in providing high flexibility while keeping costs and power consumption bounded. In optical burst switching, before a burst is sent a control packet is generated and sent toward the destination to make an one-way resource reservation. In HOS networks, the burst itself is sent after a fixed delay called offset-time. The offset-time gives to bursts a kind of prioritized handling in comparison to packets and enables for the implementation of different service classes. In this paper we distinguish between two types of bursts, namely short and long bursts, which can support two different service classes. Long bursts are transmitted using higher offset-times and slow optical switching elements, while short bursts are forwarded using shorter offset-times and fast optical switching elements. The main drawback of optical burst switching within data centers is the relatively high latency. In fact, due to the offset-time and to the time required for the burst assembly, bursts introduce relatively high end-to-end delays. Optical bursts can be then associated to data-intensive and latency-insensitive DC's applications, such as MapReduce, Hadoop, and Dryad.

Optical packets are transmitted toward the HOS core switch without any resource reservation in advance. As a consequence packets show a higher contention probability with respect to bursts, but on the other hand they also experience lower delays. The HOS core switch employs fast optical switching elements whose switching time is in the order of nanoseconds to forward optical packets. Optical packets are mapped to DC's applications requiring low latency and generating small and rapidly changing data flows. Examples of such applications are those that are based on parallel fast Fourier transform (MPI FFT) computation, such as weather perdition and earth simulation. MPI FFT requires data-intensive all-to-all communication and periodic global synchronization among the servers, and consequently requires frequent exchange of small data entities. For a more detailed description of the HOS data types the reader is referred to [12,13].

B. System Architecture

The architecture of the proposed HOS optical switched network for DCs is shown in Fig. 1. The topology is a traditional fat-tree 3-Tier in which the aggregation switches are replaced by the HOS edge nodes and the core electronic routers are

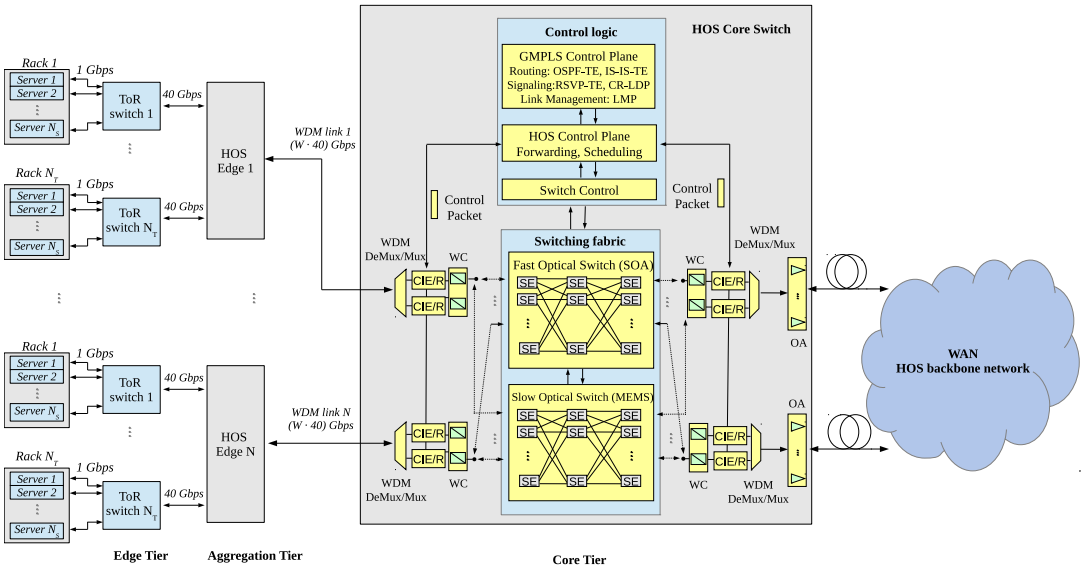


Fig. 1. Architecture of a DC employing a HOS interconnection network. ToR: top of the rack, CIE/R: control information extraction/reinsertion, WC: wavelength converters, OA: optical amplifiers.

replaced by a HOS core node. Each rack is equipped with N_S blade servers each of which generates traffic independently and is connected to a ToR switch through a 1 Gbps link. In future DCs, the servers will be probably connected using higher capacity links. However, the majority of current DCs employ 1 Gbps links. Also, the proposed network should scale easily to support higher capacity per server port.

Each ToR switch is in turn connected through a 40 Gbps link to a HOS edge node that is used to perform traffic aggregation and classification. As many as N_T ToR switches are connected to a single HOS edge node. Each HOS edge node is connected to the HOS core node through a WDM link composed of W wavelengths channels operated at 40 Gbps. Assuming that the core node interconnects as much as N HOS edge nodes, its total switching capacity is $N \cdot W \cdot 40$ Gbps. The HOS core node is directly connected to a WAN. For increased overall performance and energy efficiency we assume that the HOS core node is connected to a HOS WAN [13,14] but in general the core node could be connected to the Internet using any kind of network technology.

The HOS core node can be logically divided into three building blocks: the control plane, the switching fabric and the other active components. The control logic processes the control information and schedules data transmissions. It comprises three modules: the GMPLS control plane, the HOS control plane and the switch control unit. The GMPLS control plane is used to ensure the interoperability with the other core nodes connected to the WAN. Hence, the GMPLS control module is needed only if the HOS core node is connected to a GMPLS-based WAN, such as the WAN proposed in [13], [14]. In case the core node is connected to a WAN that does not employ GMPLS, the GMPLS control module is not needed.

The HOS control plane manages the transmission of circuits, bursts and packets. Three different scheduling algorithms are employed, one for each different data type. In order to increase the resource utilization and reduce the packet loss probability, the HOS control plane fills unused TDM-slots of circuits with optical packets with the same destination [12]. Finally, the switch control unit creates the optical paths through the switching fabric. The switching fabric is composed of two optical switches, a slow switch for the transmission of circuits and long bursts and a fast switch for the transmission of packets and short bursts. The fast optical switch is based on semiconductor optical amplifiers (SOA) and its switching elements are organized in a non-blocking three-stage Clos network. The slow optical switch is realized using 3D micro electromechanical systems (MEMS). Finally, the other active components include optical amplifiers (OAs), tunable wavelength converters (TWCs), and control information extraction/reinsertion (CIE/R). TWCs can convert the signal over the entire range of wavelengths, and are used to solve data contentions. A more detailed description of the structures and functions of the above mentioned modules can be found in [12].

The HOS edge nodes are electronic switches used for traffic classification and aggregation. The architecture of a HOS edge node is shown in Fig. 2. In the direction toward the core switch the edge node comprises three modules, namely classifier, traffic assembler and resource allocator. In the classifier, packets coming from the ToR switches are classified basing on their application layer requirements and are associated with the most suited optical transport mechanism. The traffic assembler is equipped with virtual queues for the formation of optical packets, short bursts, long bursts and circuits. Finally, the resource allocator schedules the optical data on the output wavelengths according to specific scheduling algorithms that

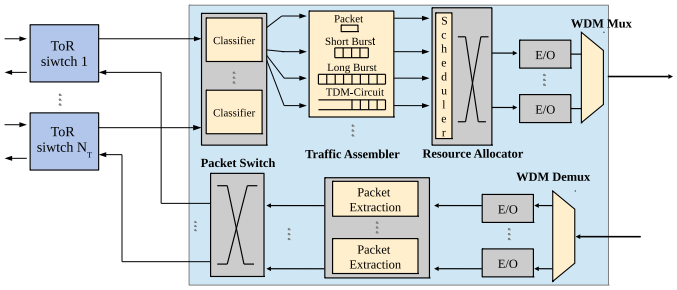


Fig. 2. HOS edge node architecture.

aim at maximizing the bandwidth usage. In the direction toward the ToR switches a HOS edge node comprises packet extractors, for extracting packets from the optical data units, and an electronic switch for transmitting packets to the destination ToR switches.

In DCs with many thousands of servers, failures in the interconnection network may lead to losses of a high amount of important data. Therefore, resilience is becoming an increasingly critical requirement for future large-scale DC networks. However, the resilience is out of scope of this study and we do not address it in this paper, leaving it as an open issue for a future work.

III. Modeling Approach

To evaluate the proposed HOS DC network we developed an event-driven C++ simulator. The simulator takes as inputs the size of the DC, i.e. the number of servers per rack (N_S), number of ToR switches per HOS edge node (N_T), number of WDM fiber links (N), number of wavelengths per fiber (W), and the DC traffic characteristics, i.e. the inter-arrival rate distribution of the packets generated by the severs. As regards the size, DCs can be categorized in three classes: university campus DCs, private enterprise data centers, and cloud-computing DCs. While university campus and private enterprise DCs have usually up to a few thousands of servers, cloud-computing DCs, operated by large service providers, are equipped with up to 100K servers. In this paper we analyze a large cloud computing DC. We set $N_S = 48$, $N_T = W = 64$ and $N = 32$. We then obtain that the DC is equipped with 98,304 servers and that the HOS core switch has a total switching capacity of 81.9 Tbps. The proposed HOS network can be easily scaled to support a higher number of servers by increasing N and/or W. An interesting feature is that when using more wavelength channels per fiber, i.e. increasing W, we can expect an improvement in performance. This is due to the scheduling algorithms employed in the resource allocators and the HOS core node, which achieve higher performance for a higher number of wavelengths per fiber, i.e. a higher W.

In the literature, we have not found any comprehensive theoretical model of DC network traffic. However, there are several research papers that report some measured data on traffic characteristics in real DCs [15-16]. Basing on the information collected in these papers, the inter-arrival rate distribution of the packets arriving at the DC network can be modeled with a positive skewed and heavy-tailed distribution. This highlights the difference between the data center environment and the wide area network, where a long-tailed *Poisson* distribution typically offers a best fit with real traffic data. The *lognormal* and *weibull* distributions represent both a good model for the DC network traffic. The *lognormal* distribution fits better to cloud-computing and private enterprise DCs, while the *weibull* distribution fits better to university campus DCs. As a consequence, in our simulations the inter-arrival rate of the packets generated by the servers belongs to a *lognormal* distribution. In order to analyze the performance at different network loads, we run simulations with different mean and standard deviation values of the *lognormal* distribution.

In the proposed HOS network, the flows between servers in the same rack are handled by the ToR switches and thus are not transmitted throughout the aggregation and core tiers. Conversely, all the flows between servers located in different racks are transmitted throughout the HOS edge and core switches. We define here the *intra-rack (IR) traffic ratio* as the ratio between the traffic directed to the same rack and the total generated traffic. According to [15-16], the IR ratio fluctuates between 20% and 80% depending on the DC category and the applications running in the DC. Since the IR ratio has a strong impact on the HOS network performance, we run simulations with different values for IR. The total traffic generated by the servers is given by the sum of the IR traffic and the traffic between different racks, i.e. the traffic that passes through the HOS edge and core switches. We refer to the network load as to the ratio between the total traffic generated by the servers and the simulation time.

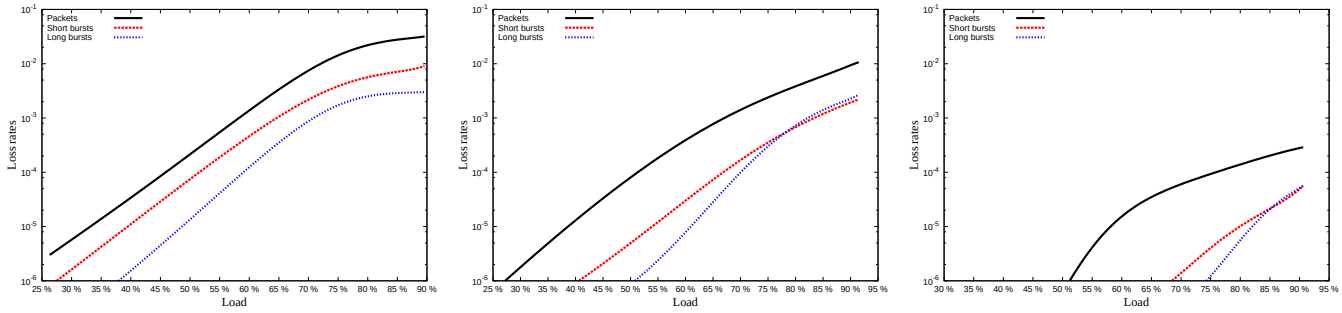
Another important observation is that although a single rack can generate as much as 48 Gbps, the ToR switches are connected to the HOS edge nodes by 40 Gbps links, leading to an over-subscription ratio of 1.2. Over-subscription relies on the fact that very rarely servers transmit at their maximum capacity because very few applications require continuous communication. It is often used in current DC networks to reduce the overall cost of the equipment and simplify DC network design. As a consequence, the aggregation and core tiers of a data center are designed to have a lower capacity with respect to the edge tier.

IV. Numerical Results

In this Section, we investigate the performance of the proposed HOS DC network by analyzing the following performance metrics: average data loss rates, average end-to-end delays, and total energy consumption.

A. Loss Rates

In this Section we show and discuss the average data loss rates. We assume that the electronic switches are equipped with electronic buffers with unlimited capacity and thus they do not introduce data losses. As a consequence, losses may happen only in the HOS core switch. The HOS core switch does not employ optical buffers to solve data contentions in



(a) Average loss rates with IR=0.2. (b) Average loss rates with IR=0.4. (c) Average loss rates with IR=0.6.

Fig. 3. Average data loss rates as a function of the input load for different intra-rack ratios.

the time domain, but it is equipped with TWCs for solving data contentions in the wavelength domain. We define the packet (burst) loss rate as the ratio between the number of dropped packets (bursts) and the total number of packets (bursts) that arrive at the HOS core switch. Similarly, the circuit establishment failure probability is defined as the ratio between the number of negative-acknowledged and the total number of circuit establishment requests that arrive at the HOS core switch. We model the traffic generated by the servers so that about 25% of the flows arriving at the edge nodes require the establishment of a circuit, 25% are served using long bursts, 25% are served with short bursts, and the remaining 25% are transmitted using packet switching.

Due to the space reason we decided not to consider in this paper the impact of different traffic patterns, i.e. different percentage of traffic served by circuits, long bursts, short bursts and packets. This effect was already evaluated for core networks in [12], and will be evaluated for DC networks in a future work.

Fig. 3 shows the average loss rates as a function of the input load for three different IR ratios, namely 0.2, 0.4 and 0.6. By comparing Figs. 3(a), 3(b), and 3(c), it can be observed that the higher the IR ratio the lower are the data loss rates. This is due to the fact that a higher IR ratio leads to a lower amount of traffic passing through the core switch, thus leading to a lower probability of data contentions. The difference between the case with IR=0.2 and IR=0.6 is around two orders of magnitude for all three transmission types (packets, short bursts and long bursts). For a fixed value of the IR ratio, the packet loss rates are always higher than the burst loss rates. This is due to the fact that for packets there is no resource reservation in advance. Due to shorter offset-times, the short bursts show higher loss rates with respect to long bursts, especially for low and moderate loads. At high input loads the advantage of having longer offset time is reduced and long bursts show almost the same loss rates as short bursts. In all cases shown in Fig. 3 the circuit establishment failure probability is null. The data loss rates shown in Fig. 3 are acceptable for today's DC applications. Applications having stringent requirements in terms of data losses can be mapped on TDM-circuits or long bursts, while loss-insensitive applications can be mapped on optical packets or short bursts.

B. Delays

In this Section we address the network latency. We assume that the IR traffic is forwarded by the ToR switches with negligible delay, and thus we analyze only the delay of the traffic between different racks, i.e. the traffic that is handled the HOS edge and core switches. The end-to-end delay is defined as the time between a data packet is generated by the source server and the time in which the data packet is received by the destination server. The *end-to-end delay* is given by the sum of the queuing delay and the propagation delay, i.e. $D = D_q + D_p$. The *queuing delay* includes the queuing time at the ToR switch and the delays introduced by the traffic assembler and resource allocator in the HOS edge switch ($D_q = D_{ToR} + D_{as} + D_{ra}$). The HOS optical core switch does not employ optical buffers and thus does not introduce any queuing delay. The *propagation delay* D_p depends only on the physical distance between the servers. Since we aim at comparing the delays introduced by the different HOS transport mechanisms under different traffic situations, but for the same size and architecture of the DC, we exclude the propagation delay from our analysis. The propagation delay is just an additive constant that should be added to the obtained values. We refer to the packet delay as to the average delay of data packets that are transmitted through the HOS core node using packet switching. Similarly, we define the short (long) burst delay as the average delay of data packets that are transmitted through the HOS core node using short (long) burst switching. Finally, the circuit delay is the the average delay of data packets that are transmitted through the HOS core node using circuit switching.

In Fig. 4 the packet, short burst, long burst, and circuit delays are shown as a function of the network load and for different IR ratios. Since there are differences of several orders of magnitude between the delays of the various traffic types, we plotted the curves using a logarithmic scale.

The figure shows that the higher the IR ratio the lower are the delays of packets, short bursts and circuits. This is due to the fact that a higher IR ratio leads to a lower amount of traffic passing through the ToR and HOS edge switches, thus leading to lower D_q. This effect is more evident for packets.

The figure also shows that circuits introduce the lowest

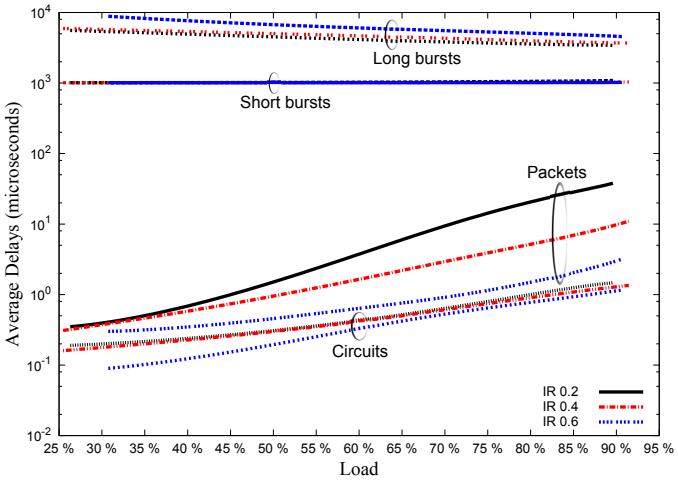


Fig. 4. Average end-to-end delays as a function of the input load for different intra-rack ratios.

delay. For circuits the assembly delay D_{as} is related to the circuit setup delay. Since in our network the circuit setup delay is several orders of magnitude lower then the circuit duration, its effect on the end-to-end delay is negligible. Furthermore, circuits are scheduled with the highest priority by the the resource allocator resulting in negligible D_{ra}. As a consequence, the circuit delay is given only by the delay at the ToR switches D_{ToR}. As can be seen from Fig. 4, circuits ensure an average delay below 1.5 μs even for network loads as high as 90% and thus are suitable for DC applications with strict delay requirements. Packets also do not suffer from any assembly delay, i.e. $D_{as} = 0$, but they are scheduled with low priority in the resource allocator resulting in non negligible values of D_{ra}. However, it can be observed that at 50% load the average packet delays are lower than 1 μs, even when IR=0.2, and that increasing the load up to more than 90% the average packet delays remain bounded to a few tens of μs. As a consequence, packets are suitable for the majority of today's delay-sensitive DC applications. Short and long bursts are characterized by very high traffic assembler delays D_{as}. In case of bursts, D_{as} is given by the sum of the time required for the burst assembly and the offset-time. The traffic assembler delay is several orders of magnitude higher than D_{ToR} and D_{ra} and thus the end-to-end delay can be approximated with D_{as}. Short bursts employ a timer-based assembly algorithm, with average assembly time of 500 μs, and have an average offset-time of 500 μs. This results in an average D_{as} of 1 ms, independently of the network load and the IR ratio. As a consequence, as shown in Fig. 4, the short burst delays with IR ratios of 0.2, 0.4 and 0.6, are almost identical and constant with respect to the network load. Long bursts employ a mixed timer/length assembly algorithm, in which a minimum length for the burst is required before starting the timer. We refer to this minimum required length as to the long burst threshold. The higher is the rate of the traffic arriving at the HOS edge switch the lower is the time required for reaching the long burst threshold and starting the timer. As a consequence, the higher is the rate of the traffic arriving at the HOS edge the lower is the long burst assembly delay D_{as}. This effect is shown in Fig. 4, where it is clear that the long burst delay decreases while increasing the network load and while decreasing the IR ratio. Due to high assembly times and offset-times, around 1.5 ms in average, long bursts introduce high delays in the range of several milliseconds. The majority of current DC applications do not accept such a high latency. This raises the question if it is advisable or not to use long bursts in future DC interconnects. On the one hand long bursts have the advantage of introducing low loss rates, especially at low and moderate loads, and reducing the total power consumption, since they are forwarded using slow and low power consuming switching elements. On the other hand, it may happen that a DC provider does not have any suitable application to map on long bursts due to their high latency. If this is the case, the provider could simply switch off the long burst mode and run the data center using only packets, short bursts and circuits. In this way, the DC can be adapted to changing network requirements ensuring maximum flexibility.

C. Energy Consumption

In this Section we evaluate and discuss the energy efficiency and the greenhouse gas emissions of the proposed HOS interconnect. The energy consumption of the DC network is given by the sum of the energy consumed by all of its active elements. In our analysis we exclude the power consumption of the servers and we consider only the power consumption of the network elements. We make the assumption that the power consumption of all the electronics components in the DC is always constant and does not depend on the network load. This is due to the fact that current electronic switches do not yet support dynamic switching off or putting in low power mode temporarily inactive ports. On the contrary, we assume that the optical ports of the HOS core node can be switched off when they are inactive. This is possible because when two parallel switches are in use, only one must be active to serve traffic from a particular port. In addition, because circuits and bursts are scheduled a priori, the incoming traffic is more predictable, so the switch-control unit can schedule inactive ports to be switched off for some period of time.

The ToR switches are conventional electronic Ethernet switches. Several large companies, such as HP, Cisco, IBM and Juniper, offer specialized Ethernet switches for use as ToR switch in large cloud computing DCs. We estimated the ToR power consumption by averaging the values found in the data sheets released by these companies. The power consumption of the HOS edge node is calculated by summing the estimated power consumption of all its building blocks, namely classifier, assembler, resource allocator, packet extractor, and electronic switch. It is assumed here that the classifier and assembler are realized using two large field programmable gate array (FPGA) devices. The power consumption of the resource allocator is determined by estimating the power consumptions of the scheduler and the electronic switch, which is used to forward data to the selected output wavelength. The power consump-

TABLE I
SUMMARY OF THE POWER CONSUMPTION OF THE COMPONENTS WITHIN A DC INTERCONNECT.

Components	Power [W]
Top of the Rack Switch (*ToR*)	650
HOS edge switch	
Classifier (1 × port)	62
Assembler (1 × port)	62
Resource Allocator (1 × node)	296
Packet Extractor (1 × port)	25
Switch (1 × port)	8
HOS core switch	
Control logic (1 × node)	49,638
SOA switch (1 × port)	20
MEMS switch (1 × port)	0.1
Tunable Wavelength Converter (1 × port)	1.69
Control Info Extraction/Re-insertion (1 × port)	17
Optical Amplifiers (1 × port)	14
Electronic core switch	
Control logic (1 × node)	27,096
Electronic CMOS switch (1 × port)	8
Line card (1 × port)	300

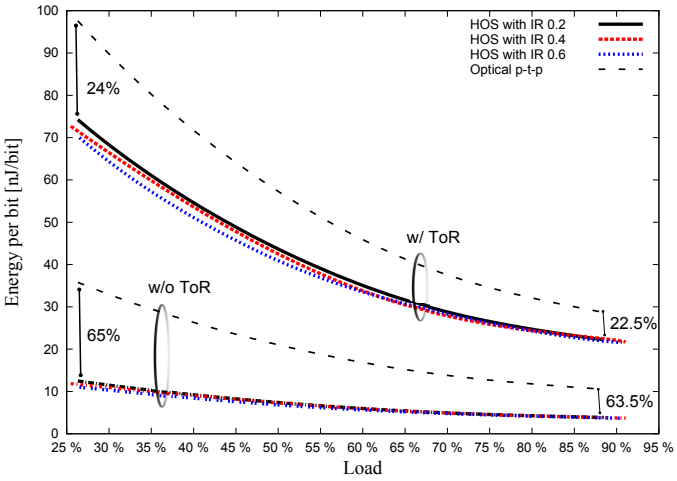


Fig. 5. Energy consumption per bit as a function of the network load for a HOS and an optical point-to-point DC interconnect.

tion of the HOS core node is obtained by summing the power consumption values of the control logic, the switching fabric and the other active components. The power consumption of the control logic is the sum of the power consumed by the GMPLS control module, the HOS control layer and the switch control unit. The power consumption of the switching fabric is given by the sum of the power consumed by the fast and the slow optical switch. Finally, the power consumed by the other active components (TWCs, CIE/R and OA) is added to the HOS core node's power consumption. We assume that the optical switch ports and the TWCs can be switched off when inactive, thus the power consumption of the HOS core node depends on the network load. For a detailed description of the power consumption model that we developed for evaluating the power consumption of the HOS core node the reader is referred to [12-14]. To highlight the improvement with respect to current technologies, we compare the energy efficiency and greenhouse gas emissions of the proposed HOS interconnect with the energy efficiency and greenhouse gas emissions of a conventional DC based on optical point-to-point interconnects and electronic switching. The considered optical point-to-point DC network is organized a 3-Tier topology, but with currently available electronic aggregation switches and core switches. The aggregation switch architecture is similar to the HOS edge node architecture, except that it does not include the traffic assembler, the resource allocator and the packet extractor. In fact, since electronic packet switching is used in the core switch these functions are not needed. The core electronic switch is composed of two building blocks: the control logic and the electronic switching fabric. The control logic includes the route processors, the management cards and the switch control unit. The electronic switching fabric is based on fast electronic switching elements and interconnects a large number of electronic line cards (LCs). A detailed description of the architecture of the all-electronic core switch can also be found in [12-14]. In Table I we report a summary of the power consumption values for the components mentioned above. The values were obtained by collecting data from several commercially available components and modules for conventional large-scale switching and routing systems as well as from research papers.

In Fig. 5 we show the energy consumption per bit as a function of the network load for a HOS and an optical point-to-point DC interconnect. For the HOS interconnect we consider three different values of the IR traffic ratio, while the energy consumption of the optical point-to-point DC does not depend on the IR ratio. Firstly, we considered the overall energy consumption of the DC network and thus we included in our analysis the power consumption of the ToR switches. It can be observed from Fig. 5 that when including the ToR switches, the HOS network provides energy savings in the range between 22.5% and 24%, depending on the network load. In this case, it could be possible to save almost one fourth of the power consumed in a DC network. The improvement in energy efficiency provided by HOS is limited by the power consumption of the electronic ToR switches. In fact, the ToR switches consume more then 80% of the total energy in a HOS DC network. In order to evaluate the relative improvement in energy efficiency provided by the use of HOS edge and core switches instead of using conventional solutions, we present also the energy consumption per bit obtained without considering the ToR switches. It can be seen that the relative gain offered by HOS is between 63.5% and 65% in this case. The electronic ToR switches limit then by almost three times the potential of HOS in reducing the DC power consumption, raising the issue for a more energy efficient ToR switch design. Fig. 5 shows that in all cases the energy consumption per bit decreases while increasing the network load, i.e. the energy efficiency increases while increasing the network load. This is explained by the fact that the electronic switches always consume the same amount of power independently of the network load. A higher network load leads to a higher amount of traffic that crosses the DC network, thus to an increase

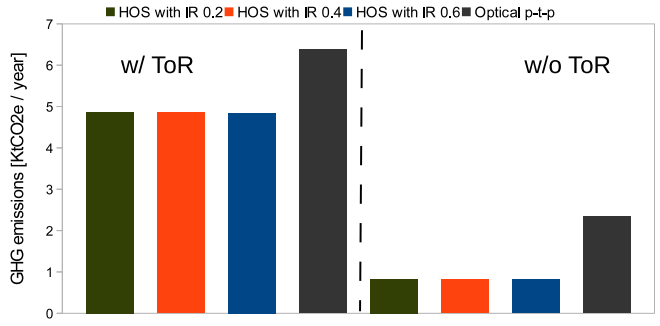


Fig. 6. Total greenhouse gas emissions of a HOS and an optical point-to-point DC network at 50% of network load.

in energy efficiency of both the electronic switches and the overall DC network. Only the HOS optical core node is able to adapt its power consumption to the current network usage. As a consequence, the advantage of using a HOS interconnect instead of an optical point-to-point interconnect is more evident at low and moderate loads, and it reduces slightly while increasing the network load. The values reported in Fig. 5 show that by increasing the load from 25% to 90% the difference between the energy consumption of the HOS and the point-to-point DC networks reduces by 1.5%. Finally, Fig. 5 additionally shows that the energy efficiency of the HOS network depends only marginally on the IR traffic ratio. While increasing the IR ratio the energy consumption decreases because a higher IR ratio leads to a lower amount of traffic crossing the HOS core node.

In Fig. 6 we show the greenhouse gas (GHG) emissions per year, expressed in metric kilotons (kt) of carbon dioxide equivalent (CO_{2e}), of a HOS and an optical point-to-point DC network. The GHG emissions of the HOS DC network are computed at 50% of network load and for three different values of the IR traffic ratio. The GHG emissions of the optical point-to-point DC network are independent from the network load and the IR ratio. We consider both the cases with and without ToR switches. To evaluate the GHG emissions, we first compute the total network energy consumption and then we apply the conversion factor of 0.356 $kgCo_{2e}$ emitted per kWh, which was found in [17]. Fig. 6 highlights that HOS reduces the overall GHG emissions by more then 1.5 $ktCO_{2e}$ per year. Again, if we do not consider the contribution of the ToR switches we achieve a much higher relative gain.

V. Conclusions

In this paper we introduced a novel optical switched interconnect network for DCs based on hybrid optical switching (HOS). HOS integrates optical circuit, burst and packet switching within the same network. Different DC applications are mapped to the optical transport mechanism that best suits to their traffic characteristics, ensuring high flexibility and high bandwidth utilization. Furthermore, HOS envisages the use of two parallel core optical switches, of which one is a slow and low power consuming switch for the transmission of circuits and long bursts while the other is a fast switch for the transmission of packets and short bursts. The use of two parallel optical switches enables switching off or putting in low power mode a number of switch ports, thus providing high transmission efficiency and low power consumption. The proposed HOS DC network is organized in a traditional fat-tree 3-Tier topology. In the aggregation tier, electronic HOS edge nodes are used to perform traffic classification and assembly, while in the core tier, a large HOS optical core node forwards data among different edge nodes and connects the DC to the Internet. We analyzed the proposed HOS DC network by evaluating average loss rates, average delays and energy consumption. Our results indicate that the loss rates are relatively low and acceptable for today's DC applications. Circuits and packets ensure low latency and are suitable for delay-sensitive DC applications. Bursts show higher delays and can be used for applications without stringent requirements in terms of latency. Finally, the results show that HOS is able to considerably increase the energy efficiency and reduce GHG emissions with respect to a conventional DC network based on electronic switching and optical point-to-point interconnects.

References

[1] C. Kachris, I. Tomkos, "A Survey on Optical Interconnects for Data Centers", IEEE Communications Surveys & Tutorials, Vol. 14, Issue 4, pp. 1021-1036, Fourth Quarter 2012.
[2] Where does power go? GreenDataProject, available online at: http://www.greendataproject.org, 2008.
[3] A. Benner, "Optical Interconnect Opportunities in Supercomputers and High End Computing," OFC 2012, OTu2B.4.
[4] N. Fehratovic, S. Aleksic, "Power Consumption and Scalability of Optically Switched Interconnects for High-Capacity Network Elements," Proc. OFC 2010, pp. JWA84-1-JWA84-3, Los Angeles, USA, March, 2010.
[5] S. Aleksic, N. Fehratovic, "Requirements and Limitations of Optical Interconnects for High-Capacity Network Elements," Proc. ICTON 2010, pp. We.A1.2-1-We.A1.2-4, Munich, Germany, June, 2010.
[6] G. Wang, et al., "c-Through: Part-time Optics in Data Centers," Proc. ACM SIGCOMM 2010, pp. 327-338.
[7] N. Farrington, et al., "Helios: a hybrid electrical/optical switch architecture for modular data centers," Proc. ACM SIGCOMM 2010, pp. 339-350.
[8] X. Ye, et al., "DOS: A scalable optical switch for datacenters," Proc. ACM/IEEE ANCS'10, 2010, pp. 24:1-24:12.
[9] A. Singla, et al., "Proteus: a topology malleable data center network," Proc. ACM SIGCOMM, 2010, pp. 8:1-8:6.
[10] K. Xia, et al., "Petabit Optical Switch for Data Center Networks," Technical report, Polytechnic Institute of NYU, 2010.
[11] O. Liboiron-Ladouceur, et al., "Energy-efficient design of a scalable optical multiplane interconnection architecture," IEEE J. Sel. Topics Quantum Electron., no. 99, pp. 1-7, 2010.
[12] M. Fiorani, M. Casoni, S. Aleksic, "Performance and Power Consumption Analysis of a Hybrid Optical Core Node," OSA/IEEE J. Opt. Comm. Netw., Vol. 3 , Issue 6, pp. 502-513, June 2011.
[13] M. Fiorani, M. Casoni, S. Aleksic, "Energy-Efficient Internet Core Employing Hybrid Optical Switching", IEEE Internet Computing Magazine, Vol. 17, Issue 1, pp. 14-22, Jan./Feb. 2013.
[14] M. Fiorani, M. Casoni, S. Aleksic, "Analysis of a GMPLS enabled hybrid optical switching network," Proc. of ONDM 2012, pp. 17-20, April 2012, Colchester, England.
[15] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," Proc. ACM workshop on Research on enterprise networking, 2009, pp. 65-72.
[16] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," Proc. IMC 2010, 2010, pp. 267-280.
[17] Guidelines to Defra/DECCs GHG Conversion Factors for Company Reporting, Am. Economics Assoc. (AEA), 2009.