POPI: A Passive Optical Pod Interconnect for High Performance Data Centers

Raluca-Maria Indre
Orange Labs, France
ralucamaria.indre@orange.com

Jelena Pesic
Inria, France
jecapesic@gmail.com

James Roberts
IRT SystemX, France
james.roberts@irt-systemx.fr

***Abstract*—We propose an original approach for realizing an all-optical data center interconnect. This consists in interconnecting racks of servers by WDM channels via an optical coupler and applying an EPON-like dynamic bandwidth allocation (DBA) algorithm. We apply this approach to design a switch-free, pod-sized data center interconnect that we call POPI for Passive Optical Pod Interconnect. Bandwidth sharing is realized by a central controller that implements an original MAC protocol and DBA algorithm. We evaluate latency and throughput performance, accounting for the dynamic, stochastic nature of data center traffic.**

I. Introduction

The rapid expansion of data centers in number and size, coupled with increasing concerns about their energy footprint, has spurred recent research into the design of more efficient interconnects. In particular, there is much interest in extending the use of optical technology in the data center network. Most proposals would supplement or replace the interconnect between top-of-rack (ToR) switches using fast optical circuit switching. Optical packet switching is envisaged but proposed designs remain futuristic and not ready for commercial deployment. In this paper, we present a novel interconnect design based on the use of mature *passive optical networking* technology.

Passive optical networking technology is already widely deployed in the access network and a number of proposals exist for realizing a metropolitan area network. It has not, however, to the best of our knowledge, been proposed for the data center interconnect. Our proposal is to apply passive optical technology to construct the interconnect for a pod-sized network (i.e., a data center of around 1000 servers that would fit into a shipping container).

An alternative approach for a large data center would be to retain the ToR switches and interconnect them using a passive optical network. This would be more readily deployable since it does not require updates to servers. On the other hand, to replace ToR and aggregation switches by passive optical technology in the pod subnetworks leads to the greatest energy savings. We therefore limit present scope to the latter solution, reserving the evaluation of alternative passive optical interconnect designs for future work.

Our main objective is to demonstrate the feasibility and potential advantages of applying passive optical networking in the data center context. These advantages naturally include energy efficiency since the interconnect currently accounts for some 23% of data center power consumption [12]. Precise burst scheduling avoids the need for buffering and offers scope for efficient traffic control without relying on problematic end-system congestion control algorithms. Wavelength channels are fully shared by a large number of servers ensuring high levels of performance and facilitating traffic engineering. These advantages distinguish POPI from alternative interconnect designs, as discussed later in Section VI.

Our main contributions are as follows.

- We propose a novel passive optical pod interconnect (POPI, pronounced "poppy") whose topology interconnects servers directly without electrical switching.
- We have designed a flexible MAC protocol to control the dynamic sharing of the wavelength channels.
- We propose a flow-aware dynamic bandwidth allocation (DBA) algorithm and demonstrate its excellent performance.

The next section reviews the requirements of data center interconnects. The topology and components of POPI are presented in Section III where we also give a rough estimation of potential power savings. Section IV specifies the proposed MAC protocol and the flow-aware DBA algorithm and their performance evaluation is presented in Section V. Related work is discussed in Section VI before we draw the main conclusions in Section VII.

II. Data center interconnects

We briefly discuss the data center interconnect and the nature of its traffic.

A. Structure

In this paper we consider the data center to be a set of servers arranged in racks of 20 to 40 devices. Most data centers currently have the topology depicted in Figure 1. Servers in each rack are connected to a top of rack (ToR) switch. Racks are grouped in "pods" in which ToR switches are interconnected via aggregation switches. There are around 30 racks in a pod for a total of some

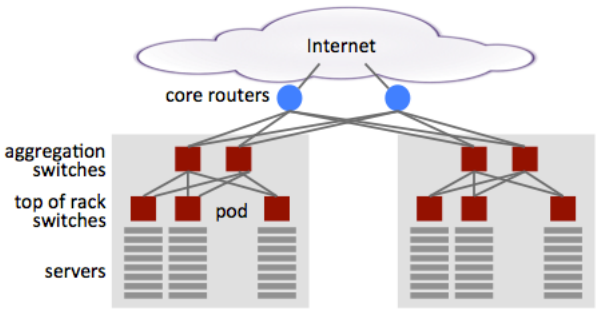


Fig. 1: Traditional data center network

1000 servers. Pods are typically interconnected via core routers which also constitute gateways to the Internet. The number of pods in a data center is extremely variable with the largest "data warehouses" now containing more than 100 000 servers.

B. Traffic characteristics

Data center traffic characteristics have been described in a number of papers without there yet emerging a commonly accepted generic traffic model. The types of application run on the servers varies widely from one center to another [6]. Most generate multi-task jobs and this impacts the nature of interconnect traffic with the generation of multiple flows, in parallel or sequentially, between a potentially large number of distinct servers. Flow sizes have been shown to be highly variable, most flows being very small but most traffic being contained in large flows [5]. The flow arrival process is bursty, as would be expected given the multi-task nature of jobs.

The server-to-server traffic matrix is sparse with pronounced locality patterns [13]. Average server-to-server traffic in a 10 s interval ranges from 0 to around 50 Mb/s (cf. Fig. 3 in [13]). Most data center links are lightly used with a median traffic per server of around 4 Mb/s (cf. Fig. 2 in [7]).

Data center traffic has been loosely categorized as a mix of background flows and query flows [5]. Background flows are relatively large and their performance requirement is basically high throughput. Query flows typically relate to interactive web applications and low latency is critical. Responses must meet a tight deadline between 10 and 100 ms to be useful.

III. A passive optical pod interconnect (POPI)

We present the network topology and identify its principle components before discussing energy requirements.

A. Network topology

The envisaged pod interconnect is illustrated in Figure 2. Servers, controllers and gateways are interconnected by one incoming and one outgoing fibre in a tree topology rooted on a passive star coupler. The figure shows a gateway interconnecting the pod to other pods and to the Internet. There could be more than one gateway, if necessary.

Routing is wavelength based. Each server or gateway has a designated incoming wavelength that it generally shares with several other devices. It receives all optical signals on that wavelength and must filter out its own packets using data in their header after conversion to electronic form.

Servers and gateways are equipped with fast tunable transmitters enabling them to send optical bursts to any destination by the appropriate choice of wavelength (Fig. 2b). They must maintain a forwarding information base (FIB) indicating the wavelength to which any destination address corresponds.

The capacity of wavelength channels and their number are design options and depend on network size and traffic. Given the traffic characteristics discussed in Section II, a relatively small number of wavelengths (around 16) appears adequate for a typical pod interconnect, even at a channel rate of only 1 Gb/s. Filters extract particular wavelength channels to be converted to electronic form and read by the destination device.

To avoid collisions, bursts must be scheduled precisely and this is the role of the interconnect controller. The controller receives reports from servers and gateways on a report control channel (orange in Fig. 2a) and issues grants on a grant channel (blue in the figure). To protect against failure of this vital element, the network is equipped with an active standby controller.

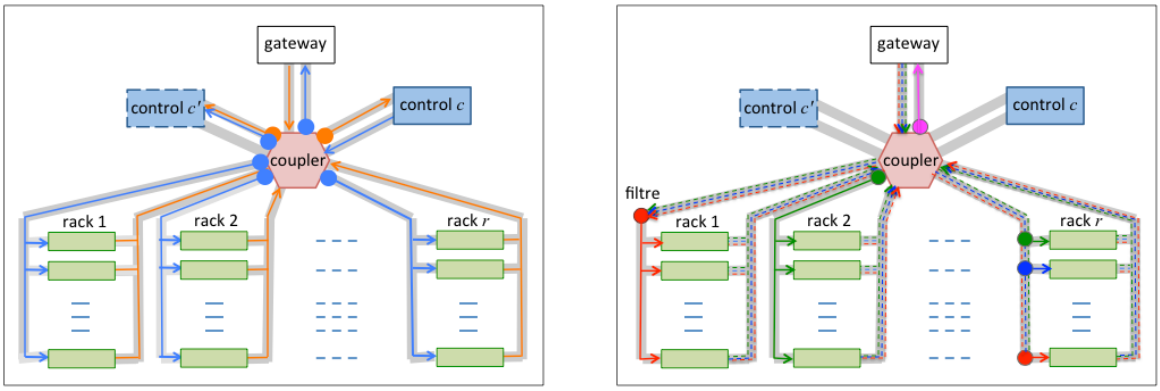
B. Components

1) Controller: The controller is a processor that computes the precise schedule of emissions on all wavelengths, as explained later. The controller is able to compute such a schedule by playing the role of the OLT in an EPON access network.

2) Server network interface: The servers must be equipped with an evolved network interface fulfilling functions like those of an ONU in EPON. The interface formulates and emits reports to the controller and executes the grants received in return. The latter operation consists in emitting bursts of precise duration starting at a precise instant. For incoming traffic, like the ONU, each server converts optical bursts to electronic form and discards packets destined to other servers.

3) Gateways: POPI allows all-optical communication between thousands of servers. To scale to higher data center capacities and to communicate with the outside, the network includes one or more gateways. A gateway must fulfill the same role as a server to communicate within the pod. It performs optical-to-electrical conversion and vice-versa and performs buffering to enable communication with the outside.

4) Optical technology: Passive couplers are sufficient to distribute light signals, even for the worst case of the grant channel where some 1000 servers each receive a



(a) Control wavelengths (b) Data wavelengths

Fig. 2: Passive optical pod interconnect: r racks of servers and a gateway, interconnected via a coupler using time-shared WDM channels, with active and standby controllers c and c'. The figure illustrates possible alternative positions for filters allowing wavelengths to be common to a rack (racks 1 and 2) or assigned individually to servers (rack r).

fraction of the signal emitted by the controller. Filters are required to block all but one or two specified wavelength channels. Transmitters must be rapidly re-tunable to allow successive burst transmissions on different channels.

C. Ranging, resource discovery and routing

The EPON ranging procedure (cf. IEEE 802.3av) can be applied in POPI with the interconnect controller playing the role of the OLT. Time stamp exchanges on the control channels enable the controller c to measure the round trip time (RTT) rtt_{ci} between itself and a server or gateway, i, and to set the remote device clock precisely to controller clock time minus the (unknown) one way propagation delay. This is sufficient to enable the controller to assign non-conflicting transmission slots to sources *as if* it were the destination. Since the slots would not collide at the controller, they would not collide either at the coupler and since they would not collide at the coupler they would not collide either at their final destination. These affirmations derive from the tree topology. The active standby controller c' in Figure 2 can participate in the above procedure and derive propagation delays, like the operational controller c, thus enabling it to immediately take control if c fails.

A new server can be added to the network using the EPON resource discovery protocol specified in IEEE 802.3av while wavelength assignments and filter settings can be accomplished using a protocol like GMPLS. The control plane would also incorporate a routing protocol to populate server and gateway FIBs.

D. Power consumption

A strong motivation for using an optical interconnect is to reduce power consumption. The following is a rough evaluation intended to illustrate the potential for economies through implementing POPI.

Assuming a pod of 32 racks and 1000 servers, an Ethernet interconnect requires 32 ToR switches and at least 2 aggregation switches. Assuming each switch consumes 600 W [2], this amounts to more than 20 kW.

In place of the switches, POPI adds a star coupler, some filters, two controllers and the control channels. We count one filter per rack at 1 W per filter [3]. The controllers are broadly equivalent to an EPON OLT and therefore consume around 20 W each [4]. The control channel requires one extra transceiver per server which, at 1 W each for short range transmission [1], contributes a total of 1 kW. We add up to 2 W per server for additional ONU-like network interface functions.

The comparison is thus largely in favour of the optical interconnect with some 3 kW instead of 20 kW.

IV. MAC protocol and algorithms

We propose a medium access control (MAC) protocol and dynamic bandwidth allocation (DBA) algorithm for POPI. We refer generically to servers and gateways as sources, while destinations are assimilated to their assigned wavelength.

A. Report and grant signalling

Reports are emitted in a static TDMA cycle. Each source in turn generates a burst including a fixed size report for each wavelength. The interval between successive reports is at least $S(Wb_r/C_r + \Delta_g)$ where S is the number of sources, W the number of wavelengths, b_r the number of bytes in each source-wavelength report, C_r the report channel rate in byte/s and Δ_g the inter-burst guard time. The guard time between successive grants is necessary to account for laser tuning delay and any residual timing imprecision. For example, assuming $b_r = 2$, $W = 16$, $S = 1000$, $C_r = 125$ MB/s and $\Delta_g = 200$ ns, the inter report interval is at least 456 μs.

Grants are broadcast by the controller using continuous mode transmission. Each grant identifies source and wavelength and specifies a start time and duration for a total size b_g of around 10 bytes. Assuming an average

burst size of G bytes, W wavelengths and a grant channel rate C_g equal to the data channel rate, the grant channel utilization is less than Wb_g/G. For example, for $b_g = 10$, $W = 16$ and $G > 1$ KB, the grant channel utilization is less than 16%.

B. Scheduling

We have adapted the scheduling algorithm proposed in [11] for a wide area network. Let $g_w(n)$ be the instant according to its local clock that the controller c decides to emit the n^{th} grant to use wavelength w. The n^{th} grant is sent to some source i and allocates a burst of duration $d_w(n)$ starting at time $s_w(n)$ *as given by the local clock at source* i. Let free_i denote the end of the last reservation by any wavelength for the transmitter of source i, as measured by the source i local clock. These start times and grant times are calculated recursively as follows:

1) at time $g_w(n)$ compute $s_w(n) = g_w(n) + \Delta_o - rtt_{ci}$,
2) **if** $(s_w(n) > \mathsf{free}_i)$, send the grant with start time $s_w(n)$ and duration $d_w(n)$; set the next grant time $g_w(n+1) = g_w(n) + d_w(n) + \Delta_g$,
3) **else**, send the grant to i with start time free_i; set the next grant time $g_w(n+1) = \mathsf{free}_i - \Delta_o + rtt_{ci}$,
4) update $\mathsf{free}_i = \mathsf{free}_i + d_w(n) + \Delta_g$.

In the above, Δ_g is the guard time and Δ_o an offset that compensates for differences in round trip times. The offset must be greater than $\max_j(rtt_{cj}) + \tau$ where τ is the maximum delay that can occur before the grant message is sent on the grant channel ($\tau = (W-1)b_g/C_g$, in the notation of Section IV-A). This offset ensures the grant always arrives at the source in time to start transmission at $s_w(n)$.

C. Dynamic bandwidth allocation

The following flow-aware DBA algorithm is adapted to the nature of data center traffic.

1) Report content, grant duration: We identify flows at a given source by the destination of packets waiting to be sent and any other criterion that distinguishes separate flows for the same destination (e.g., their application). We introduce quantum, Q, as the unit of allocation in bytes. Flows are distinguished as follows: "greedy flows" have more than Q bytes of data waiting while "singleton flows" at most Q. A quantum of 10 KB, for instance, would be a useful transmission unit for background transfers (a 10 μs burst at 1 Gb/s) while ensuring most queries are considered as singletons (cf. [5]). Reports from each source i communicate the current number of greedy flows, $r^g(i,w)$, for each wavelength w, together with the total number of bytes, $r^s(i,w)$, that have been generated by singleton flows since the last report was sent. The controller maintains a record of the latest reported number of greedy flows, $n^g(i,w)$, and its count of waiting singleton traffic in bytes, $t^s(i,w)$. When the controller emits a grant for source i to use wavelength w, it sets $d_w(n) = (n^g Q + t^s)/C$, where C is the data channel rate in byte/s, and resets t^s to zero. When a report arrives at the controller, n^g for that source wavelength pair is reset to r^g and t^s is incremented by r^s.

2) Server choice: We suppose sources with non-zero waiting packets ($n^g + t^s > 0$) are identified in a cyclic linked list. At grant times $g_w(n)$, the controller picks a starting point in this cycle at random and iteratively seeks the first source with a free transmitter. If the transmitters of all sources in the list are busy, the grant is attributed to the source whose transmitter is free first, i.e., to source i such that $\mathsf{free}_i - s_w(n)$ is minimal.

3) Burst assembly: When a source fulfills a grant at the designated starting time $s_w(n)$, it proceeds as follows. The burst is composed first by packets from all singleton flows in their order of arrival. If there is remaining time, this is used for quanta of greedy flows, considered in round-robin order until there is no room left. This service discipline gives local priority to singleton flows and therefore reduces their latency. However, latency is small thanks mainly to the fact that the number of simultaneous flows using a given wavelength is typically very small, as confirmed in the performance evaluation in Section V.

It is important to note that bandwidth sharing is controlled at MAC layer meaning performance is largely independent of the particular transport protocols implemented by the various server applications.

V. Performance evaluation

We evaluate the throughput and latency performance of the proposed DBA under a traffic model including background and query traffic. The results are indicative of the excellent performance expected from POPI under actual data center traffic.

A. Traffic model

Flows for each source-wavelength (SW) pair arrive according to a Poisson process. Background flows are "greedy" and have a geometrically distributed number of 1 KB packets and mean size 10 MB. Query flows consist of exactly one 1 KB packet. Eighty percent of traffic is due to background flows. We simulate a POPI configuration with 16 data channels, a report channel and a grant channel, all at 1 Gb/s. This channel rate would be sufficient for 1000 servers with an average traffic less than 16 Mb/s (cf. Sec. II). Traffic is perfectly symmetrical and generated by 16 or 64 independent active sources (results shown below allow us to predict the performance when more sources are active). We set all propagation times to 1 μs, corresponding roughly to fiber hops of 200 m (only the maximum time is significant). The quantum Q is equal to 1 KB (i.e., one packet) so that query flows are "singletons". We assume reports of 2 B are emitted every 500 μs and grants are 10 B long. POPI is simulated using Omnet++ (cf. www.omnetpp.org).

1) Background flow throughput: Figure 3 plots the throughput of background flows against channel load, ρ_w. By throughput we mean the mean flow size divided by

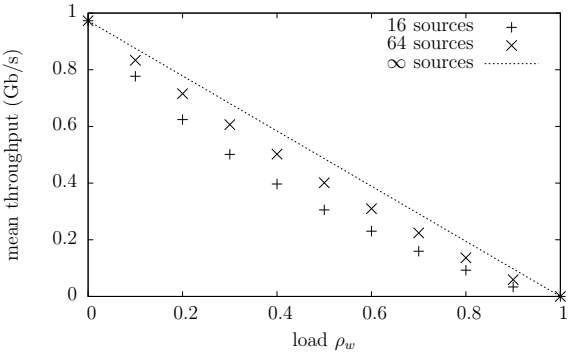


Fig. 3: Mean background flow throughput against wavelength channel load

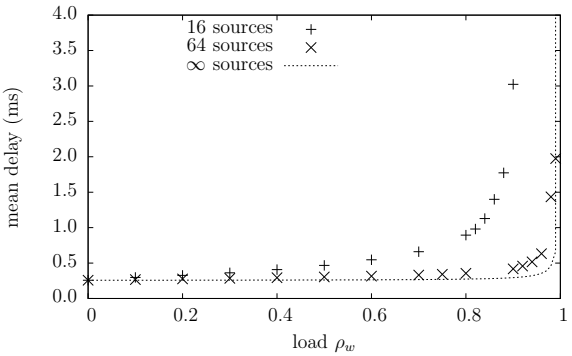


Fig. 4: Mean query packet delay against wavelength channel load

the mean flow completion time. It ranges from nearly 1 Gb/s at negligible load (channel rate reduced by a per-quantum guard time overhead: $C' = CQ/(Q + C\Delta_g)$ or 0.975 Gb/s for a 200 μs guard time.) to 0 as demand attains network capacity. Throughput is lower when traffic is concentrated on only 16 sources due to contention for transmitters. The straight line is the throughput for a very large number of sources. Each source then has only one active flow (with probability 1) and the wavelength is shared perfectly fairly between active flows. Throughput is therefore that of a processor sharing server of capacity C', given by $C'(1 - \rho_w)$. The figure suggests convergence to the processor sharing model is rapid as the number of active sources increases.

2) Packet latency: Figure 4 is a plot of the mean latency of singleton packets (the query traffic). The figure shows latency is mainly determined by signalling time until load ρ_w gets close to one. The signalling time is composed as follows: a new query packet must wait (250 μs on average) before being announced in a report, the report is sent to the controller ($\sim 2\mu$s propagation), the source waits for the grant to arrive ($\sim 2\mu$s propagation) and sends the packet ($\sim 1\mu$s transmission + $\sim 2\mu$s propagation), for a total of 257 μs.

Delay due to congestion is lower as the number of active sources increases due to the reduced impact of transmitter sharing. The curve for "infinite" sources is derived as follows: the mean number of (greedy and singleton) flows in progress is equal to the mean number of customers in a processor sharing queue, $\rho_w/(1-\rho_w)$, a new query packet waits for the signalling time and then, roughly, on average, for half the other flows to transmit one quantum yielding latency $\approx 257\ \mu s + (Q + C\Delta_g)/(2C)\rho_w/(1 - \rho_w)$. The congestion delay is negligible until demand is very close to capacity because the quantum transmission time Q/C is very small.

B. Impact of data center traffic

The above results for Poisson flows are in fact representative of much more general traffic models. This is due to the insensitivity properties of processor sharing. Mean throughput and query latency are the same for a traffic model defined as follows: *jobs* arrive as a Poisson process and generate tasks *successively*; each task initiates a flow on some wavelength with a generally distributed size; the interval between the end of one task and the start of the next is generally distributed; intervals and task sizes can be correlated; the number of tasks per job has a general distribution (cf. Sec. 3.1 in [9]). The impact of flows initiated *in parallel* is also known to be slight in realistic traffic [8], suggesting the following broad conclusions are generally valid:

- the system is stable until load attains maximum capacity 1,
- background flow throughput is excellent, even at high load,
- the latency of query flows is very low in normal operating conditions.

Performance depends essentially only on channel load and this can be controlled if necessary by reassigning servers to wavelengths.

VI. Related work

The recent survey on optical interconnects by Kachris and Tomkos constitutes a valuable state of the art [12]. Some proposed hybrid interconnects allow optical circuits to be created dynamically to bypass congested paths in the regular electrical interconnect. All-optical interconnects cited in [12] are frequently based on advanced technologies that are unlikely to be cost-efficient in the near future. POPI represents a novel approach with respect to this state of the art, applying mature passive optical technology to efficiently share wavelength channels at burst scale.

Some more recently proposed schemes would create a dynamically reconfigurable mesh of optical circuits between ToR switches using MEMS switching and WDM. The OSA proposal of Chen *et al.* [10] would change network topologies and link capacities dynamically to meet evolving demand, using multi-hop routing. Porter *et al.* [15] propose an enhanced hybrid interconnect using microsecond scale circuit switching. Both proposals

require complex optimization to design efficient circuit configurations in real time based on comprehensive traffic monitoring. In contrast POPI fully shares channel capacity realizing one-hop server-to-server paths. Interconnect reconfiguration to match capacity to demand is therefore both simple and rarely needed.

A major difference between our proposal and the above is, of course, that POPI directly interconnects servers and gateways without the need for ToR switches. In this respect, POPI is similar to the use of TDMA over Ethernet, as proposed by Vattikonda *et al.* [18]. We share the advantages identified in that paper while gaining efficiency thanks to the 802.3av ranging procedures which enable more precise timing than is possible with the 802.3x flow control protocol used in [18].

Since POPI can employ alternative MAC/DBA algorithms, there is much related work that could be cited in this area. Schemes proposed for the TWIN architecture are particularly relevant [16], [17]. The flow-aware, centrally controlled algorithm proposed for POPI is inspired by our prior application of passive optical technology to the wide-area network [11]. It has been necessary to adapt this algorithm to the specific context of the data center interconnect.

A very recent paper is quite closely related to our proposal. Ni *et al.* [14] also propose to use passive optical devices to realize a data center interconnect. They define a distributed MAC and DBA protocol where all nodes share a common view of the current request status and individually determine compatible grant timings.

VII. Conclusions

The present paper demonstrates that passive optical networking technology holds considerable promise for the realization of high performance data center interconnects. It would significantly reduce the energy footprint and simplify network management by eliminating at least two layers of electrical switching.

Our proposal is based on mature optical technology though some development is required to integrate ONU-like functions in the server network interface. The feasibility of the proposed MAC protocol and associated DBA algorithm hardly needs demonstrating, given the widespread penetration of EPON and GPON in the access.

The performance evaluation of POPI takes account of the stochastic nature of data center traffic. While this traffic is still imperfectly understood and varies significantly from one data center to another, our simulation and analytical results have quite general interpretations. They notably reveal that the proposed flow-aware DBA algorithm has excellent and predictable performance under simple network engineering rules.

We have several directions for future work. The proposed flow-aware DBA algorithm is just one possibility and it is important to more fully explore the requirements of particular data center instances in order to develop specifically adapted algorithms. Finally, passive optical technology has potential for much wider application in the data center than that explored here. It could, for instance, be used to interconnect POPI instances or, alternatively, to create an energy efficient interconnect between ToR switches.

Acknowledgment

This work was partially funded by the Celtic-Plus European project SASER-SAVENET.

References

[1] Product Data Sheet. Cisco 10-gigabit Transceiver Modules.
[2] Product Data Sheet. Cisco Nexus 5000 Series Switches.
[3] Product Data Sheet. Sercalo Tunable Optical Filter with Control Board.
[4] Product Data Sheet. SUN-GE8100 Optical Line Terminal.
[5] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan. Data center TCP (DCTCP). In *Proceedings of the ACM SIGCOMM 2010 conference*, SIGCOMM '10, pages 63–74, 2010.
[6] T. Benson, A. Akella, and D. A. Maltz. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, IMC '10, pages 267–280, 2010.
[7] T. Benson, A. Anand, A. Akella, and M. Zhang. Understanding data center traffic characteristics. *SIGCOMM Comput. Commun. Rev.*, 40(1):92–99, Jan. 2010.
[8] T. Bonald and L. Massoulié. Impact of fairness on internet performance. In *Proceedings ACM SIGMETRICS*, pages 82–91, New York, NY, USA, 2001.
[9] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Syst. Theory Appl.*, 53(1-2):65–84, June 2006.
[10] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen. Osa: An optical switching architecture for data center networks with unprecedented flexibility. *Networking, IEEE/ACM Transactions on*, PP(99):1–1, 2013.
[11] D. Cuda, R.-M. Indre, E. Le Rouzic, and J. Roberts. Building a low-energy transparent optical wide area network with "multipaths". *Optical Communications and Networking, IEEE/OSA Journal of*, 5(1):56–67, 2013.
[12] C. Kachris and I. Tomkos. A survey on optical interconnects for data centers. *Communications Surveys Tutorials, IEEE*, 14(4):1021–1036, 2012.
[13] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken. The nature of data center traffic: measurements & analysis. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 202–208, New York, NY, USA, 2009.
[14] W. Ni, C. Huang, Y. Liu, W. Li, K.-W. Leong, and J. Wu. Poxn: A new passive optical cross-connection network for low-cost power-efficient datacenters. *Lightwave Technology, Journal of*, 32(8):1482–1500, April 2014.
[15] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat. Integrating microsecond circuit switching into the data center. In *Proceedings ACM SIGCOMM 2013*, 2013.
[16] P. Robert and J. Roberts. A flow-aware mac protocol for a passive optical metropolitan area network. In *Proceedings of the 23rd International Teletraffic Congress*, ITC '11, pages 166–173. ITCP, 2011.
[17] K. Ross, N. Bambos, K. Kumaran, I. Saniee, and I. Widjaja. Scheduling bursts in time-domain wavelength interleaved networks. *IEEE JSAC*, 21(9):1441–1451, 2003.
[18] B. C. Vattikonda, G. Porter, A. Vahdat, and A. C. Snoeren. Practical tdma for datacenter ethernet. In *Proceedings of the 7th ACM european conference on Computer Systems*, EuroSys '12, pages 225–238, New York, NY, USA, 2012.