Building Optical Packet Networks without Buffering, Signalling or Header Processing

Thomas Bonald, Davide Cuda
Telecom ParisTech, France
{thomas.bonald, davide.cuda }@telecom-paristech.fr

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

Ludovic Noirie
Alcatel-Lucent Bell Labs, France
ludovic.noirie@alcatel-lucent.com

***Abstract*—We propose a technique for building optical packet networks that does not require any buffering, any signalling or header processing. Contentions are solved by means of an optical device that allows the first packet to go through while blocking others. Blocked packets are redirected back to their source nodes, thus notifying the latter about the packet status. We describe the design principles of the corresponding all-optical networks, assess their performance and power consumption and give examples of application in the context of access networks and data centers.**

***Index Terms*—Dynamic optical combiner, packet loss, blocking, random access.**

I. Introduction

Today's optical technologies such as wavelength division multiplexing (WDM) are mainly used to provide high-capacity point-to-point links between electronic network nodes. While optical transmission is able to cope with increasing traffic demands, electronic switching is currently reaching fundamental limits in terms of processing speed, energy requirements and port count [1]. The gap between high-speed optical transmission and limited electronic processing can be bridged by moving some switching functionalities to the optical domain.

Optical switching techniques differ with respect to the granularity of switched units, i.e., wavelengths [2], bursts [3] or packets [4], [5]. Ideally, to match IP traffic and achieve high utilization, optical switching should be performed at packet level. Proposed Optical Packet Switching (OPS) solutions mimic electronic IP networks. At each node, the information contained in the packet header is extracted and processed electronically to make the forwarding decision; meanwhile, the payload is stored optically using fiber delay lines (FDL). Unfortunately, contentions cannot be solved in the time domain, as in electronic networks, due to the lack of optical random access memory (RAM). Moreover, the OPS technology is limited by the power consumption required to perform O/E/O conversion of packet headers. Recent studies are questioning the ability of OPS to reduce power consumption when compared to present-day electronic routers [1].

In this paper, we propose an approach for building optical packet networks without resorting to any optical buffering, signalling or header processing. The proposed solution is particularly suitable for corporate networks, access networks and data centers. Buffering and routing operations are performed electronically at the network edge, while packets are transmitted end-to-end in the optical domain. Contentions are resolved through a simple first-come first-served policy: the first optical packet to arrive goes through while the others are blocked and sent back to their source node. This contention resolution method can be implemented by means of a new optical device, that we refer to as a Dynamic Optical Combiner (DOC). Importantly, the cost, complexity and power consumption of a DOC is practically independent of the data rates.

The feedback mechanism used by the DOC is similar in principle with the concept of protection against collision described in [7] for star-coupler based WDM systems. The feasibility of the feedback mechanisms has recently been demonstrated experimentally in the context of array waveguide grating router (AWGR)-based optical interconnects [8]. In this paper, we apply this idea to build optical packet switches that can be deployed in any network topology through wavelength routing, as long as a virtual tree topology per wavelength can be established over the physical network. We also propose a distributed Medium Access Control (MAC) which, together with the DOC, provides efficient and fair sharing of resources like transmitters and wavelengths.

We assess network performance through some variants of the Engset model [9] at the packet level and processor-sharing queues at the flow level. Our results show that performance is comparable to that of electronic networks of same capacity, both in terms of throughput and packet delay. The key advantages are the scalability in the input data rates and a huge gain in power consumption, by one order of magnitude or more.

The rest of the paper is organized as follows. The following two sections describe the optical devices and the network architecture. In Section IV, we present some variants of the Engset model that characterize the way optical resources are shared. These models are applied in the following three sections to derive key metrics on throughput, packet delay and power consumption, respectively. The results are illustrated on the practically interesting examples of access networks and data centers in Section VIII. Section IX concludes the paper.

II. Optical Devices

We first describe the Dynamic Optical Combiner (DOC), which is able to handle packet contention in the optical domain and to notify the sources about packet collisions, if any.

The authors are members of LINCS (Laboratory of Information, Network and Communication Sciences), Paris, France. See www.lincs.fr.

A. Dynamic Optical Combiner

A DOC has N input ports and a single output port. At any given time, at most one of the input optical signals is transmitted to the output. Specifically, the first input to become active is allowed to pass; during the transmission, any other input signal is blocked and redirected to the source.

Figure 1 depicts a possible implementation of a DOC, which requires only low-cost, off-the-shelf components: N 90/10 splitters, N photodiodes, N FDLs, N optical switches, one gate controller and one optical coupler at the output.

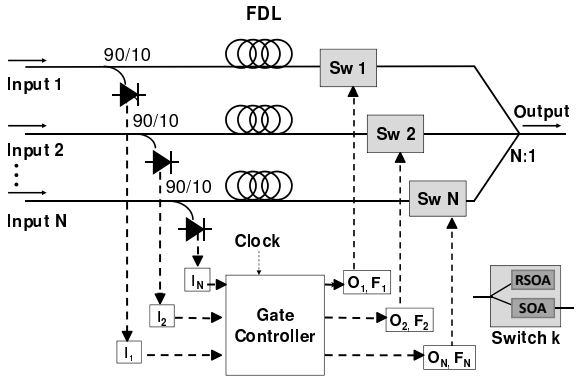

Fig. 1. A $N \times 1$ Dynamic Optical Combiner (DOC). The continuous lines represent optical signals while the dotted lines correspond to electrical signals.

Each optical switch is used to forward any incoming signal to the output or to reflect the signal back into the input port if the output is occupied. It can be implemented using one 90/10 splitter, and two ON/OFF optical gates, the first one being classical Semiconductor Optical Amplifier (SOA) and the second one a Reflective Semiconductor Optical Amplifier (RSOA). By default, the two gates are in the OFF position. When the input becomes active, only one of the two gates is switched ON, depending on the DOC output state, the other gate blocking the signal. If the first gate (SOA) is ON, the incoming signal is forwarded to the output of the DOC. If the second gate (RSOA) is ON, the incoming signal is redirected back to the input of the DOC. Each gate is thus SOA-based, which allows switching times in the order of nanoseconds [10], and compensation of signal losses due to splitting and coupling, as explained in §VIII-A below.

FDLs are needed to delay packets while the optical switches are configured. Since configuration times are of the order of nanoseconds, the required length of FDLs is around 1m. This is very different from classical OPS where FDLs of hundreds of meters are needed to store the payload while the header is processed electronically.

The gate controller only performs elementary operations and thus can be implemented using a simple programmable electronic device such as FPGA. For instance, a synchronous implementation of the gate controller (with timeslots of the order of nanoseconds) would use the following variables:

- **Input variables**: I_k describes the state of input k, equal to 1 if input k is active and to 0 if input k is idle;

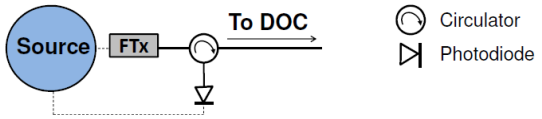

Fig. 2. Source node able to detect the feedback received from a DOC. Continuous and dotted lines represent optical and electrical signals, respectively.

- **Output variables**: O_k and F_k give the respective states of the SOA gates (O) and of the RSOA gate (F) in switch k. O and F are equal to 1 if the gate is open and to 0 if the gate is closed; since both gates cannot be open at the same time, we have $O_k F_k = 0$;
- **Internal variable**: S gives the input port that is selected for transmission, if any; the default value is 0 (no input is selected).

The value of the internal variable S determines the values of the output variables O_k and F_k. For all k, we have $O_k = 1$ if and only if $S = k$, and $F_k = I_k \wedge \bar{O}_k = I_k \cdot (1 - O_k)$. The input port $S(n)$ that is allowed to pass at time slot n is then selected as follows. Assume that all inputs are initially idle. If input k becomes active at time n, i.e., $I_k(n) = 1$, $S(n)$ is set to k so that $O_k(n) = 1$: the signal coming from input k is transmitted to the DOC output [1]. If another input, say input j, becomes active while input k is still active, i.e., $I_j(n) = I_k(n) = 1$ and $S(n-1) = k$, $S(n)$ is set to k yielding $F_j(n) = 1$: the signal coming from input j is blocked and redirected back to the source. Now assume that the transmission on input k ends while input j is still active, i.e., $I_j(n) = 1$, $I_k(n) = 0$ and $S(n-1) = k$. The controller sets $S(n)$ to 0, yielding $F_j(n) = 1$: the signal from input j is still redirected back to the source avoiding the transmission of an incomplete packet.

Since any blocked signal is sent back to the source, each source must be able to detect the feedback received from the DOC. To this end, each source is equipped with a transmitter, a circulator and a photodiode, as depicted in Figure 2. The circulator transmits signals received from the transmitter to the DOC and signals received from the DOC to the photodiode. The photodiode is used to detect whether or not an optical signal is received from the DOC. The presence of a signal corresponds to a collision, indicating that the packet must be retransmitted.

To enable source nodes to detect a collision before the end of the packet transmission, the round-trip propagation delay (RTD) between the source and the DOC must be less than the packet duration. To ensure this, it is necessary to impose a *minimum packet size*. For instance, at 1 Gbit/s, a short-range network with a maximum path length of 100m would require a minimum packet size of 125 B. On detecting a collision, a source node stops the transmission and waits for a random backoff time before retransmitting the packet. To allow for longer path lengths, larger minimum packet sizes should be imposed. Alternatively, explicit acknowledgements

[1] In the rare case where several inputs become active at the same time slot, the signal to be sent is arbitrarily selected, e.g., at random.

on a separate channel could be used; the signal would not have to be reflected back to the source in case of collision.

B. Optical Switch-Combiner

DOCs can be used to build a wavelength-routed optical packet switch, that we call *optical switch-combiner*. We consider an $N \times N$ optical switch in which every input fiber carries W wavelength channels. Each wavelength on each input fiber is assigned to one or several output fibers. The signal transmitted on each output wavelength is obtained by dynamically combining all input signals transmitted on the same wavelength by means of an $N \times 1$ DOC. Although we described the DOC with a dedicated gate controller, there may be a single gate controller common to all DOCs in practice.

Figure 3 depicts a simple 2×2 switch-combiner with $W = 4$ wavelengths on each input fiber. Input wavelengths are statically[2] assigned to specific output ports: wavelength 1 is routed to output port 1, wavelengths $2, 3$ are routed to output port 2 and wavelength 4 is routed to both output ports 1 and 2. The WDM signal received on each input fiber is first demultiplexed into W different optical signals, one for each wavelength channel. The optical signals are then separated into W wavelength planes, i.e., signals transmitted on the same wavelength are grouped together and sent to a DOC. The latter will allow only one of the 2 input signals to pass at any given moment. The output signal of each DOC is an aggregation of the data transmitted on all input fibers, on a given wavelength channel. This signal is then transmitted to specific output port(s) and thus to specific destination(s).

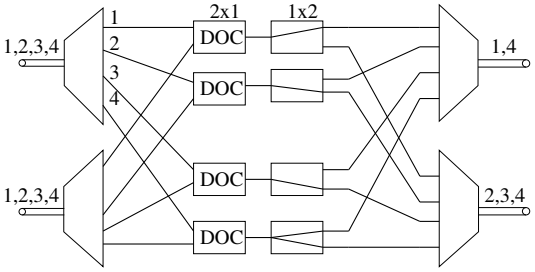

Fig. 3. A 2×2 optical switch-combiner with $W = 4$ wavelengths per fiber.

When cascading several switch-combiners, each switch-combiner must be able to detect the feedback received from an upstream switch-combiner. Similarly to source nodes, each output port of a switch-combiner must be equipped with one circulator. The role of the circulator is to allow traffic exiting the switch-combiner to be sent upstream and to allow feedback received from an upstream switch-combiner to be sent to the gate controller. The received feedback signal must first be demultiplexed and then fed to the gate controller. When the controller detects a collision on a wavelength w, it enforces the state $S(t) = 0$ to the DOC associated to wavelength w (i.e., all signals transmitted on w are blocked and redirected to the sources). If optical amplification is needed, an Erbium Doped Fiber Amplifier (EDFA) is added between the output of the switch-combiner and the circulator.

[2]In practice, to provide flexibility and adapt to changes in the traffic matrix, the switching matrix should be slowly reconfigurable (in the order of minutes).

III. Network Architecture

We now present the key design principles of the proposed all-optical networks.

A. Wavelength Routing

The network consists of a set of edge nodes that transmit and receive data in the form of optical packets and a set of optical switch-combiners that multiplex and forward these packets, possibly dropping some, as described in Section II.

Routing inside the network is wavelength-based. Specifically, a source node selects the destination(s) of a packet by transmitting it on the appropriate wavelength, say w. The packet is then sent into the network and forwarded by successive switch-combiners which are configured to deliver wavelength w to the destination(s) associated to it.

Each edge node is equipped either with a *tunable* transmitter so as to select the appropriate wavelength, depending on the destination of the optical packet, or with a simpler, less expensive, *fixed* transmitter, configured on the appropriate wavelength, if the edge node transmits data to a single set of destinations. Each transmitter is characterized by some bit rate R, typically equal to 1 or 10 Gbit/s.

Edge nodes are also equipped with one or several fixed receivers, depending on the wavelengths on which they receive data. Note that by allowing each wavelength channel to be shared in reception by several nodes, it is possible to increase the number of destinations beyond the number of available wavelengths (see the example in Section VIII-C). In this case, an optical packet sent by some source on wavelength w to some specific destination is usually received by a set of nodes, all receiving data on wavelength w. These edge nodes convert the received signal to the electronic domain and process the packet header in order to identify the packets addressed to them.

Note that the packet headers are only processed at the network edge to determine the destination; the switch-combiners multiplex and forward packets based on the activity of the input ports on the different wavelengths, without any header processing.

B. Access Scheme

The edge nodes generate traffic in the form of *flows* (e.g., data transfers, voice calls, video streams) that compete for access to the network. A Medium Access Control (MAC) is thus required to enforce some form of fairness in the sharing of network resources. We consider a simple access scheme in which every flow delays the transmission of each of its packets by some random backoff time, as in IEEE 802.11 networks for instance. Unless otherwise specified, we consider that the backoff time distribution is the same for all flows, enforcing fair sharing. Service differentiation can easily be introduced by decreasing the backoff times of high-priority flows.

The proposed MAC regulates access to each transmitter and to each wavelength. Note that it is fully distributed and, as such, does not prevent from packet collisions inside the network; it is the role of switch-combiners to resolve the contentions by letting one packet pass and by redirecting the colliding packets back to the corresponding sources. These packets are then retransmitted after some random backoff time, according to the above access scheme.

IV. Contention Models

We here describe contention models that will prove useful for evaluating the performance and power consumption of the all-optical networks described above. Specifically, we consider a fixed number of n *sources* transmitting packets through some common *channel*. Each source will later be interpreted either as an individual flow or as a set of flows, while the channel will be either a transmitter or a wavelength. We are interested in the utilization of the channel and in the packet blocking probability, assuming a static traffic matrix of persistent flows.

A. The Engset Model

We first consider the case where each source can *sense* the channel, meaning that it can immediatly determine whether the channel is occupied. Each source waits for a random backoff time and transmits a packet if the channel is sensed idle. Otherwise, the packet is blocked and the source restarts immediately a new backoff. This model corresponds to the behaviour of n flows sharing a common transmitter, according to the access scheme described in §III-B.

Homogeneous sources: We first assume that the backoff times have an exponential distribution with parameter ν for all sources. We also assume that transmission times have an exponential distribution with parameter μ. We shall see later that the results are in fact valid for any distribution of the transmission time with mean μ. We say that results are *insensitive* to the distribution beyond its mean.

Let $X(t)$ be the channel state at time t, equal to 1 if occupied and to 0 if idle. The stochastic process $X(t)$ is a Markov process with transition rates $n\nu$ from state 0 to state 1 and μ from state 1 to state 0. We deduce that, in steady state, the channel is utilized a fraction of time:

$$U(n) = \frac{n\nu}{n\nu + \mu} = \frac{n}{n+b}. \tag{1}$$

where $b = \mu/\nu$ denotes the mean normalized backoff time. In addition, the probability that a packet is blocked is equal to the probability that the channel is occupied in the presence of $n-1$ sources:

$$B = \frac{n-1}{n-1+b}.$$

The system is nothing more than Engset's model with n sources and a single circuit [9]. This model is known to have the insensitivity property so that the above results remain valid for any non-lattice distributions of transmission times and backoff times with respective means $1/\mu$ and $1/\nu$ [11].

Heterogeneous sources: We now assume that sources have exponential backoff times with different means. Specifically, the backoff times of source i are exponential with parameter ν_i. The channel utilization becomes:

$$U(n) = \frac{\sum_{i=1}^{n} \nu_i}{\sum_{i=1}^{n} \nu_i + \mu} = \frac{n}{n+m},$$

where m the harmonic mean of the normalized backoff times:

$$m = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{b_i} \right)^{-1},$$

with $b_i = \mu/\nu_i$. The utilization of source i is proportional to ν_i, that is inversely proportional to b_i. Moreover, the blocking probability of source i is equal to the probability that the channel is occupied by one of the $n-1$ other sources, that is:

$$B_i = \frac{\sum_{j \neq i} \nu_j}{\sum_{j \neq i} \nu_j + \mu} = \frac{\sum_{j \neq i} \frac{1}{b_j}}{\sum_{j \neq i} \frac{1}{b_j} + 1}.$$

B. The Generalized Engset Model

We now consider the generalized Engset model where sources *cannot* sense the channel. This is, for instance, the case of n flows originating from different edge nodes and sharing the capacity of a common wavelength at the output of a DOC: in the worst case where the RTD is larger than the packet duration, sources cannot detect the collision before the end of the packet transmission.

Homogeneous sources: We first consider the case where the mean backoff times are the same for all sources. Each source is either *idle*, *transmitting* or *blocked* the sense that it is transmitting a packet which is blocked at the switch - combiner. We denote by $X(t)$ the channel state at time t (equal to 1 if the channel is occupied and to 0 otherwise) and by $Y(t)$ the number of blocked sources at time t. Under the assumption of exponential packet transmission and backoff times, the process $(X(t), Y(t))$ forms a Markov process. The model is no longer tractable and we need some approximations to get explicit expressions.

A standard approximation consists in computing the mean *actual* backoff time of each flow, say $1/\tilde{\nu}$, accounting for the potential blocked transmission, and then applying the Engset model [12]. Denoting by $\tilde{b} = \mu/\tilde{\nu}$ the mean normalized actual backoff time and by $\tilde{B}$ the corresponding blocking probability, we get:

$$\tilde{b} = b + \tilde{B},$$

with:

$$\tilde{B} = \frac{n-1}{n-1+\tilde{b}}.$$

Solving these equations yields:

$$\tilde{b} = \frac{1}{2} \left(\sqrt{(n+b+1)^2 - 4(b+1)} - n + b + 1 \right). \tag{2}$$

The channel is utilized a fraction of time given by:

$$U(n) = \frac{n}{n+\tilde{b}}. \tag{3}$$

This utilization grows from $1/(1+b)$ to 1 when n goes from 1 to ∞, as illustrated by Figure 4. The expression (1) derived from the Engset model provides an upper bound, which is tight for large b only, i.e., when the transmission time is negligible with respect to the backoff time.

The simulation results obtained for exponential backoff times and exponential or constant packet transmission times show both the accuracy of the approximation and the insensitivity property. Indeed, the utilization is almost the same for both distributions of the transmission times, and very close to the approximation (3).

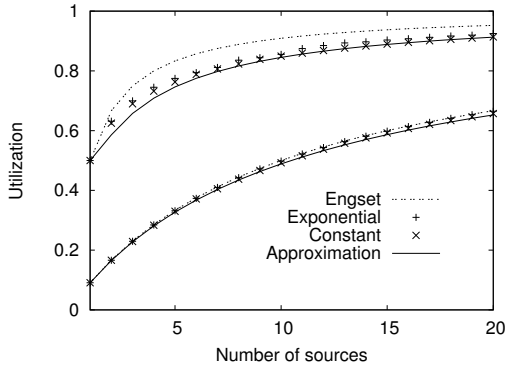

Fig. 4. Utilization of the channel by homogeneous sources for mean normalized backoff times $b = 1$ (top curves) and $b = 10$ (bottom curves).

Heterogeneous sources: The same approximation applies to the case of heterogeneous sources. We denote by $\tilde{b}_i$ the mean *actual* backoff time of source i. The fixed-point approximation becomes:

$$\tilde{b}_i = b_i + \tilde{B}_i, \tag{4}$$

where $\tilde{B}_i$ is the corresponding blocking probability:

$$\tilde{B}_i = \frac{\sum_{j\neq i} \frac{1}{\tilde{b}_j}}{\sum_{j\neq i} \frac{1}{\tilde{b}_j} + 1}. \tag{5}$$

These equations have a unique solution that can be easily computed by iteration. We obtain the channel utilization:

$$U(n) = \frac{n}{n + \tilde{m}} \tag{6}$$

where $\tilde{m}$ is the harmonic mean of the *actual* backoff times:

$$\tilde{m} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\tilde{b}_i} \right)^{-1}.$$

The utilization of source i is inversely proportional to $\tilde{b}_i$.

Figure 5 shows the results obtained for two types of sources, characterized by the respective mean backoff times $b_1 = 1$ and $b_2 = 10$, and the same number of sources. The total utilization per type of source is shown against the number of sources of each type. The approximation is quite accurate, particularly when n is large, and the results are almost insensitive to the packet size distribution.

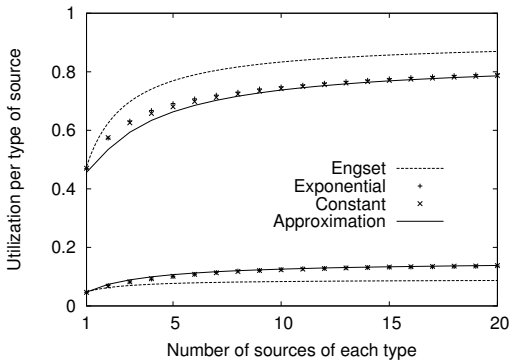

Fig. 5. Utilization of the channel by two types of sources with mean normalized backoff times $b_1 = 1$ (top curves) and $b_2 = 10$ (bottom curves).

V. Throughput Performance

In this section, we apply the above contention models to estimate the throughput achieved by each data flow in the considered all-optical networks. We first present the traffic model and the throughput metric, then analyse the wavelength sharing and consider the respective impacts of the number of ports and of the transmitters.

A. Traffic Model

The throughput achieved by each flow critically depends on the random number of flows that share the same network resources. We focus on some wavelength at the output of a switch-combiner and assume that flows arrive according to a Poisson process of intensity λ. We shall see that this assumption is in fact not essential due to the insensitivity property. We denote by σ the mean flow size in bits. The traffic intensity is then defined as the product $\lambda \times \sigma$ (in bit/s). Since packets are transmitted at rate R, the *load* of the wavelength is given by:

$$\rho = \frac{\lambda\sigma}{R}. \tag{7}$$

B. Flow Throughput

Let $\pi(n)$ be the probability that n flows share the considered wavelength in steady state. From Little's law, the mean flow duration is given by $\mathrm{E}(n)/\lambda$ with $\mathrm{E}(n) = \sum_{n\geq 1} n\pi(n)$. We define the *flow throughput* as the ratio γ of the mean flow size to the mean flow duration, in bit/s. In view of (7), we obtain:

$$\gamma = \frac{\rho R}{\mathrm{E}(n)}. \tag{8}$$

In the ideal case where the wavelength is shared in a fair and fully efficient way, the system behaves as a processor-sharing queue and the stationary distribution of the number of flows is geometric, given by $\pi(n) = (1-\rho)\rho^n$ [13]. We obtain $\mathrm{E}(n) = \rho/(1-\rho)$ and, in view of (8):

$$\gamma = R(1-\rho). \tag{9}$$

This is representative of electronic networks where flows have full access to the link capacity. In the limiting case $\rho \to 0$, each

flow is typically alone when active and gets the link capacity so that $\gamma \to R$.

In our case, the MAC prevents a single flow from using the whole wavelength capacity. We shall see using the contention models of Section IV that the corresponding throughput loss is limited, however. Specifically, the flow throughput is approximately given by:

$$\gamma \approx \frac{R}{b+1}(1-\rho).$$

The maximum flow throughput now depends on the mean backoff time b and there is still a linear decrease of the flow throughput when the wavelength load ρ grows from 0 to 1.

It is worth noting that, by the insensitivity property of the processor-sharing queue, the results are valid for any flow size distribution with mean σ. It is not even necessary to assume Poisson flow arrivals; the results hold under the more general assumption of Poisson *session* arrivals, each session corresponding to a series of flows and silence periods [11].

C. Wavelength Sharing

Consider a wavelength shared by n data flows at the output of the switch-combiner. Since the flows are not aware of the occupancy of the wavelength, the generalized Engset model applies. Specifically, the wavelength utilization $U(n)$ is given by (3), with a fair sharing between ongoing flows. The system then behaves as a processor-sharing queue with state-dependent service rate [14]. Using (7), we deduce the stationary distribution of the number of flows:

$$\pi(n) = \pi(0)\frac{\rho^n}{U(1)\dots U(n)}. \tag{10}$$

Note that, like in electronic networks, the number of flows is unbounded and may well grow to infinity. Since $U(n)$ tends to 1 when $n \to \infty$, the stability condition is $\rho < 1$. This means that the wavelength can be fully utilized: the number of flows remains finite as long as the traffic intensity is less than R. The flow throughput then follows from (8).

Again, the results are insensitive to the traffic statistics beyond their means. The only necessary assumption is that the mean flow size is sufficiently large so that the packet-level dynamics (as described by the generalized Engset model) are much faster than the flow-level dynamics (the variations of n), so that the wavelength utilization $U(n)$ achieved by n flows is indeed given by (3).

Figure 6 shows the results for different values of the mean backoff time b. In this and subsequent figures, the flow throughput is expressed as a fraction of the wavelength capacity R. The analytical results derived from (10) are compared with simulation results for a switch-combiner of $N = 100$ input ports with even traffic distribution. Unless otherwise specified, the simulation results are obtained for a simulated time of 5mn, a typical timescale over which traffic can be considered as stationary; flows arrive according to a Poisson process, have exponential sizes with mean 100 kB and constant packet sizes of 1 kB.

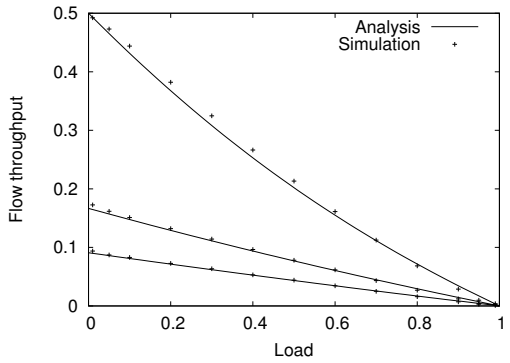

Fig. 6. Flow throughput at the output of a switch-combiner with $N = 100$ input ports ($b = 1, 5, 10$, from top to bottom).

Note that the maximum value of the flow throughput is obtained for $\rho \to 0$ and given by:

$$\gamma \to \frac{R}{b+1}. \tag{11}$$

This is due to the random access scheme that prevent any single flow from fully occupying the wavelength. When ρ grows to 1, the number of flows increases and the flow throughput gradually decreases to 0.

For large mean backoff times, the flow throughput can be shown to be approximately linear in the load. Indeed, we know that the Engset model provides a good approximation in this case (refer to Figure 4). Assuming b is an integer, we get from (1) and (10):

$$\pi(n) = \pi(0)\binom{n+b}{n}\rho^n. \tag{12}$$

It then follows from (8) that:

$$\gamma = \frac{R}{b+1}(1-\rho). \tag{13}$$

The flow throughput decreases linearly from its maximum value (11) to 0 as the wavelength load ρ goes from 0 to 1. We verify on Figure 6 that the flow throughput is indeed approximately linear for a mean backoff time $b = 10$.

D. Cascading multiple switch-combiners

We have so far assumed that each flow occupies a different input port. When cascading several switch-combiners, the input of an upper-layer switch-combiner may now correspond to the aggregate traffic going out of a lower-layer switch-combiner (see the examples of Section VIII and Figure 13). Each input port of the upper-layer switch-combiner is thus shared by several flows. Let n_i be the number of ongoing flows on input port i, for $i = 1, \dots, N$. In view of (3), the input port i can be viewed as a source with a state-dependent backoff b_i which is a function of n_i. In the limiting regime $n_i \to \infty$ for all i, the output channel after the lower-layer switch-combiner is always occupied and so $b_i \to 0$ and the utilization follows from (3) and (2) with $n = N$ and $b = 0$. We

deduce that the maximum utilization depends on the number of ports N and is given by:

$$U^\star(N) = \frac{N}{N + \frac{1}{2}(\sqrt{(N+1)^2 - 4} - N + 1)}. \quad (14)$$

This is an increasing function, as illustrated by Figure 7, with $U^\star(2) \approx 0.76$ and $U^\star(N) \geq 0.9$ for $N \geq 8$. For a large number of ports, say $N \geq 64$, the maximum utilization is very close to 1 so that the loss of efficiency is negligible, as expected in view of the simulation results of Figure 6.

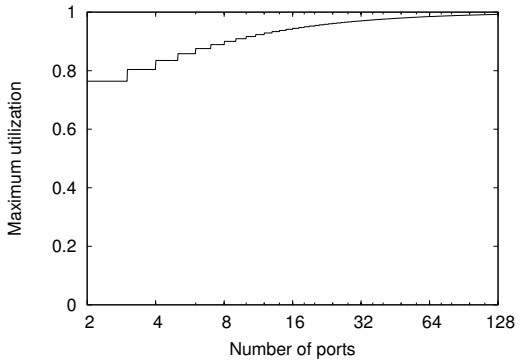

Fig. 7. Maximum utilization with respect to the number of ports.

Figure 8 shows the results for an 8×1 switch-combiner, each input port connecting a switch-combiner that aggregates the traffic of 100 edge nodes. Note that $U^\star(8) \approx 0.9$, so that the network is unstable at loads higher than 0.9 (the number of flows grows continuously on each input port). The accuracy of the approximation is thus confirmed by the simulation results that give a null flow throughput for loads greater than 0.9.

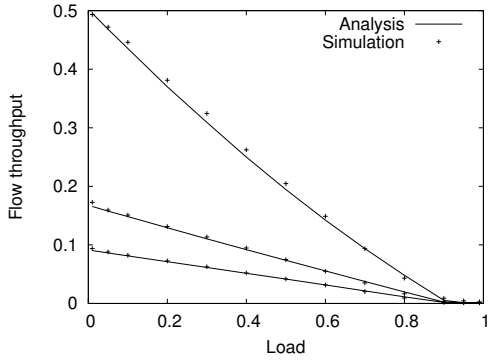

Fig. 8. Flow throughput at the output of a switch-combiner with $N = 8$ input ports ($b = 1, 5, 10$, from top to bottom), each input port connecting a switch-combiner that aggregates the traffic of 100 edge nodes.

E. Impact of Transmitters

As described in Section III, each edge node is typically equipped with a single transmitter that must be shared by the ongoing flows. The key difference with the problem of wavelength sharing considered so far is that flows can check the transmitter availability: the standard Engset model (with sensing) applies. In particular, the utilization of the transmitter is given by (1) in the presence of n flows and the flow throughput imposed by the transmitter is given by (13), where ρ now corresponds to the transmitter load instead of the wavelength load. The results are shown in Figure 9. In practice, the transmitter load is typically much lower than the wavelength load (see the examples of Section VIII) so that its impact can be neglected.

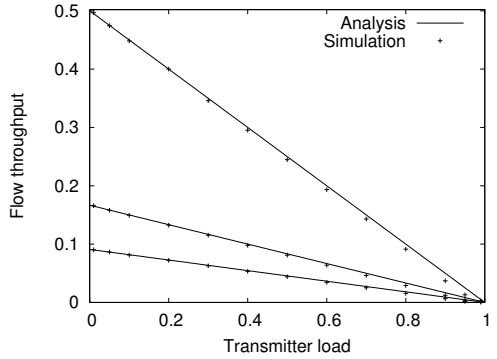

Fig. 9. Flow throughput due to the transmitter constraint ($b = 1, 5, 10$, from top to bottom).

VI. Delay Performance

We have so far considered data traffic only. This section is devoted to the performance of real-time traffic like media streaming or voice applications.

Consider a real-time flow with mean packet inter-arrival time δ. Note that typical values of δ range from 1 ms to 100 ms and thus are much higher than the packet transmission time, which is in the order of the microsecond (1kB at 10Gbit/s, say). Each packet can thus be assumed to see the network in steady state at its arrival.

A. Impact of Blocking

Consider a wavelength at the output of a switch-combiner. Denoting by ρ the corresponding load, a packet of the real-time flow is blocked if and only if the wavelength is occupied. By the conservation law, this occurs with probability ρ (since the system is stable, all incoming traffic is eventually transmitted, and the wavelength is occupied a fraction of time ρ). Thus the number of trials needed to successfully transmit a packet has a geometric distribution with mean $1/(1-\rho)$.

Figure 10 shows the probability that a real-time packet waits longer than 1 ms as a function of the wavelength load, for a single switch-combiner of $N = 100$ input ports and a data rate $R = 1$ Gbit/s. The simulation results are obtained for a mix of data traffic and real-time traffic. The proportion of real-time traffic is 10% and real-time flows are generated according to a Poisson process, each consisting of 1000 packets of 1kB arriving every $\delta = 10$ ms. For a short backoff time $b = 1$, the probability that the packet delay exceeds 1 ms is negligible (less than 10^{-5}) whenever $\rho < 0.8$; note that the analysis

is slightly optimistic in this case since successive blocking events of the same packet are not mutually independent. The analysis is more accurate for a long backoff time $b = 10$; since the packet delay is a linear function of the backoff time, the probability to exceed 1 ms is now negligible whenever $\rho < 0.5$ and then increases gradually with the network load.

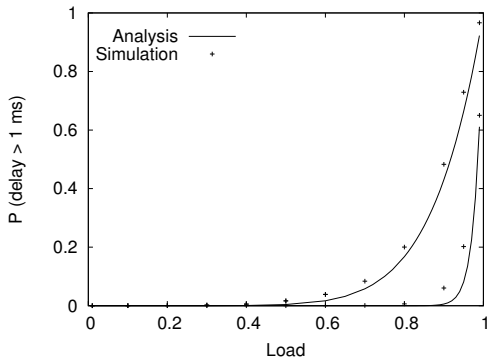

Fig. 10. Probability that packet delay exceeds 1 ms for a single switch-combiner with $N = 100$ input ports ($b = 10, 1$, from top to bottom).

B. Cascading switch-combiners

We have seen in §V-D that a slight loss of efficiency occurs when cascading several switch-combiners. The maximum utilization is given by (14) for N ports. The equivalent load is given by $\rho/U^\star(N)$, yielding a geometric distribution with mean $1/(1 - \rho/U^\star(N))$ for the number of trials per packet, whenever $\rho < U^\star(N)$.

Figure 11 shows the probability that a real-time packet waits longer than 1 ms for a switch-combiner of $N = 8$ input ports, each input port connecting a switch-combiner that aggregates the traffic of 100 edge nodes. The results are comparable to those of Figure 10 except when the wavelength load exceeds the maximum utilization $U^\star(8) \approx 0.9$. In this case, the analytical model is no longer applicable and the packet delays are indeed very high.

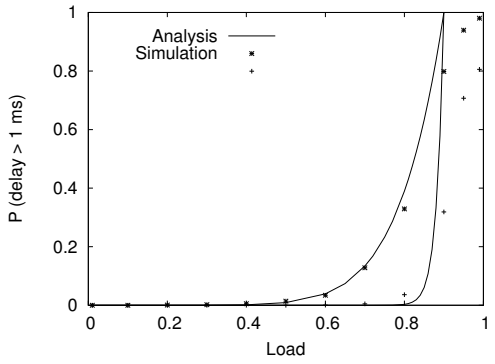

Fig. 11. Probability that packet delay exceeds 1 ms for a switch-combiner with $N = 8$ input ports ($b = 10, 1$, from top to bottom), each input port connecting a switch-combiner that aggregates the traffic of 100 edge nodes.

VII. Power Consumption

A. Maximum Power Consumption

The only active elements inside the DOC are the SOA, the RSOA and the gate controller. The default position of the SOA and of the RSOA is the OFF state in which no power is consumed [15]. When an input is active, either the SOA or the RSOA is turned ON so as to allow the signal to pass through or to redirect it back to the source, respectively. As the network load increases, more and more inputs become active causing the consumption of the switch-combiner to increase, as explained below. The power consumption of a switch-combiner is maximum only when *all* inputs and *all* wavelengths are active simultaneously.

Given a power consumption per SOA/RSOA of $P_{(\mathrm{R})\mathrm{SOA}} = 0.1$ W, the *maximum* power consumption of a switch-combiner with $N = 100$ input fibers and $W = 80$ wavelengths per fiber at rate $R = 10$ Gbit/s is 0.8 kW for a switching capacity of 800 Gbit/s. Regarding the gate controller, we just need few logical gates to implement our control algorithm; FPGA can be used, with a power consumption of 3W or even lower [6].

As a comparison, the Cisco Calalyst 6513 switch consumes 3.2 kW for a switching capacity of 720 Gbit/s [16].

B. Impact of Network Load

The key advantage of the switch-combiner over today's electronic switches is that its power consumption strongly depends on the network load and is typically much lower than the maximum power consumption. Consider a wavelength at the output of a switch-combiner with N input ports. Denote by ρ the wavelength load. There is one SOA and one RSOA associated with each input port; The SOA is active when a packet is transmitted through the DOC, while the RSOA is active when a packet is blocked by the DOC and redirected back to the source. Note both cannot be simultaneously active. To compute the average power consumption of these two gates, we need to estimate the fraction of time the input port is active. Assuming equal load distribution over the N input ports, each packet on the considered input port is blocked with probability $\rho(1 - 1/N)$. We deduce the fraction of time each input port is active: $\rho/(N(1 - \rho) + \rho)$. The total power consumption of the DOC due to the SOAs is then given by:

$$P = P_{(\mathrm{R})\mathrm{SOA}} \frac{N\rho}{N(1-\rho)+\rho}.$$

Figure 12 illustrates the power consumption of a DOC due to the SOAs for $N = 10$ or 100 ports. Results are normalized to the maximum power consumption. We observe that the average power consumption is typically much smaller than the maximum power consumption. For $\rho = 0.5$ for instance, less than 1 SOA is active in the whole DOC on average, independently of N. In the example of §VII-A, we deduce that the contribution of the SOAs and RSOAs to the power consumption is equal to 8 W only (0.1 W per wavelength); adding the power consumption of the gate controllerwe obtain a total power consumption of 11 W only.

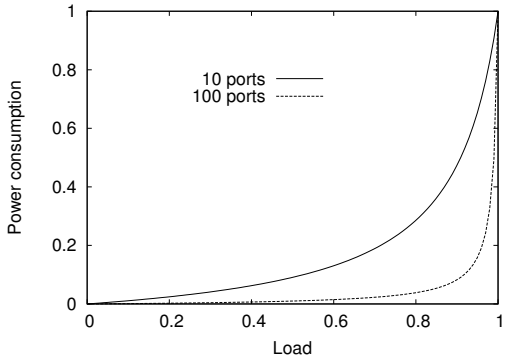

Fig. 12. Average power consumption of a DOC due to the SOAs and RSOAs.

VIII. Case Studies

We now give two examples of all-optical networks based on the proposed architecture. We start with a scalability analysis.

A. Scalability

Physical impairments like signal attenuation, noise and crosstalk limit the scalability of both the DOC and the switch-combiner in terms of number of ports. We here use the methodology presented in [17] and [18] to analyse the scalability of the DOC; the same approach applies to the switch-combiner, see [19] for details. A receiver can correctly decode an optical signal if the quality of the latter, both in terms of received power and Optical Signal to Noise Ratio (OSNR), is enough to guarantee a Bit Error Rate (BER) less than some target.

1) Power Sensitivity: We take a transmission power of 3 dBm. For a target BER of 10^{-12}, today's best receivers have a power sensitivity of -26 dBm at rate $R = 10$ Gbit/s. The power sensitivity to bit rate ratio is typically equal to 13.5 dBm per decade [20], yielding a power sensitivity of -39.5 and -17.8 dBm at rates 1 Gbit/s and 40 Gbit/s, respectively.

Any optical signal forwarded by the DOC goes through two 90/10 splitters, a FDL, a SOA and a coupler, see Figure 1. We use the model of [17] to assess the signal attenuation caused by these devices. We add a margin of 3 dB to account for component aging, laser misalignment and other effects that could degrade the optical signals and assume that the SOA can compensate for power losses up to 25 dB. The signal can then be correctly received after the DOC provided the number of ports does not exceed 310, 220 and 160 at $1, 10$ and 40 Gbit/s, respectively.

2) Optical Signal to Noise Ratio: Noise is mainly introduced by the SOAs. When in the ON state, the SOA amplifies both the useful signal and the noise; it also adds noise through spontaneous emission, whose power is typically equal to 9 dB [17]. When in the OFF state, a small part of the signal still goes through the SOA. The ratio between the output power of an SOA in the ON and OFF states is called the extinction ratio (ER). This leads to in-band crosstalk contributions coming from the blocked input ports. Our analysis shows that the ER is the most critical factor limiting the DOC scalability. Specifically, adopting the methodology presented in [21], [17], we modeled the effect of crosstalk as a power penalty defined for an optimized decision-threshold receiver for a given target BER [19]. Table I presents the maximum number of input ports of the DOC for different values of the SOA ER, assuming a transmitter OSNR of 40 dB. For instance, a DOC can have up to 100 ports at 10 Gbit/s for an SOA ER of 40 dB.

Bit rate	$ER = 35$ dB	$ER = 40$ dB	$ER = 45$ dB
1 Gbit/s	64	150	185
10 Gbit/s	57	100	115
40 Gbit/s	46	67	75

TABLE I
Maximum number of ports of the DOC.

B. Access Networks

Consider a point-to-point Fiber To The Home (FTTH) access network providing access to the Internet. Each customer is connected to the central office through a direct fiber and is equipped with a fixed transmitter at rate $R = 1$ Gbit/s and a fixed receiver. Downlink transmission is typically based on broadcasting data to all stations and is thus contention-free. We consider here only uplink transmission in which contentions may occur.

In today's optical access networks, a central controller dynamically allocates slots to customers, which requires signalling, synchronisation and scheduling. Instead, we propose to use DOC to resolve contentions. For instance, a switch-combiner can collect the traffic of $N = 100$ customers using 2 wavelengths, allowing a maximum traffic per user of 20 Mbit/s. The switch-combiner has only two DOCs, one per wavelength. Assuming that an optical fiber carries $W = 80$ wavelengths, then the traffic of 40 switch-combiners can be collected onto a single fiber by means of a passive combiner. At the output of this combiner, the optical signal has a power of -23 dBm and an OSNR of 22 dB, which are sufficient to correctly recover the signal. The access network serves a total of 4,000 customers having an aggregate capacity of 80 Gbit/s.

If each customer has a traffic intensity of 10 Mbit/s, the load per wavelength is equal to 50%. For a mean normalized backoff time $b = 1$ (that is, around 10 μs), user performance is excellent for both throughput and delay: the results of Sections V and VI show that the flow throughput is approximately equal to 200 Mbit/s (corresponding to the mean throughput of a user when active) while packet delays are lower than 1 ms with very high probability. The power consumption of each switch-combiner is essentially due to the gate controller and less than 4 W. As a comparison, the SUN-GE 8100 Optical Line Terminal has a typical power consumption of 20 W [22].

C. Data Centers

We now consider the case of a three-tier data center, see Figure 13. The data center consists of 120 clusters of $N = 100$ servers each; clusters are grouped into 15 islands of 8 clusters each. Each server is equipped with a tunable transmitter at $R = 1$ Gbit/s and several fixed receivers and has an average traffic intensity of 50 Mbit/s both in upstream and downstream.

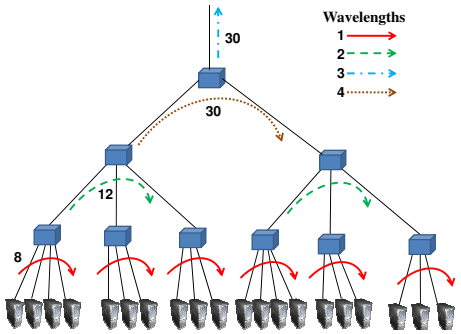

Fig. 13. An all-optical data center.

We consider a typical traffic matrix for a cloud data center. Specifically, we assume that 80% of the traffic is confined to the cluster within which it is generated [23]. The other 15% is destined to servers that are in the same island and the remaining 5% is equally distributed between the Internet and the other islands. To ensure that each wavelength is loaded at 50%, the 80 wavelengths are allocated as follows: 8 wavelengths are allocated to each cluster for internal traffic, 12 wavelengths are used inside each island for internal traffic and 30 wavelengths are allocated to traffic destined to the Internet. The remaining 30 wavelengths can be used to address the 15 islands of the data center, i.e., 2 wavelengths are associated to each island in reception.

Note that all the clusters and all the islands of the data center utilize the same wavelengths for their internal communication. This wavelength reutilization allows to increase the total amount of traffic sustained by the network. Our data center sustains a total rate of 6 Tbit/s; without reutilization, the network capacity would be $W \times C = 80$ Gbit/s.

For a mean backoff time of 10 μs, the wavelength load of 50% guarantees excellent performance for both throughput and delay, as in the case of access networks. Regarding the physical impairments, we consider the worst case of a signal crossing the entire data center. In this case, EDFAs are needed to recover from losses, see [19] for details. With the current technology, the optical signal arrives at the receiver with a power of -3 dBm and an OSNR of 18 dB, i.e., it can be correctly decoded at the receiver.

Finally, we evaluate the power consumption of the data center network and compare it to that of an Ethernet-based network. For instance, 48-port Cisco Nexus 5548 data center switches can be used to interconnect the 12,000 servers. In total, 263 such Ethernet switches disposed over 3 hierarchical layers are required to build the data center, yielding a total consumption of 158 kW [16], regardless of the network load. Using switch-combiners, the maximum power consumption is equal to 121 kW. This value is attained only at 100% load. In the considered scenario with a load of 50 %, the power consumption of each switch-combiner is equal to 11 W. Adding a power consumption of 11 W per EDFA, we obtain a power consumption of the whole data center less than 4 kW.

IX. Conclusion

In this paper, we have proposed a novel optical device based only on off-the-shelf components, able to resolve packet contention optically without requiring any electronic signalling or header processing. We have shown that this simple component can be used as a building block for all-optical packet networks. The practical interest of these networks has been illustrated in the context of access networks and data centers. Future work will be focused on the comparison with existing techniques to assess the potential further gains of limited packet processing and wavelength conversion at intermediate nodes.

References

[1] R. Tucker, "Scalability and energy consumption of optical and electronic packet switching," *Lightwave Technology, Journal of*, vol. 29, Aug. 2011.

[2] K. Zhu and B. Mukherjee, "Traffic grooming in an optical wdm mesh network," *Selected Areas in Communications, IEEE Journal on*, vol. 20, no. 1, pp. 122 –133, Jan. 2002.

[3] Y. Chen, C. Qiao, and X. Yu, "Optical burst switching: a new area in optical networking research," *Network, IEEE*, vol. 18, no. 3, pp. 16 – 23, May-June 2004.

[4] M. J. O'Mahony, D. Simeonidou, D.K. Hunter and A. Tzanakaki, "The application of optical packet switching in future communication networks," *Communications Magazine, IEEE*, vol. 39, no. 3, pp. 128 –135, March 2001.

[5] S. Di Lucente, N. Calabretta, J. A. C. Resing and H. J. S. Dorren, "Scaling Low-Latency Optical Packet Switches to a Thousand Ports," in *JOCN*, vol. 4, Sept 2012.

[6] Xilinx 7 Series FPGA Power Benchmark Design Summary, Xilinx, July 2012, pp. 13-14.

[7] B. Glance, "Protection-against-collision optical packet network," *Lightwave Technology, Journal of*, vol. 10, no. 9, pp. 1323 –1328, Sept. 1992.

[8] R. Proietti, Y. Yawei, Y. Runxiang, Y. Xiaohui, C. Nitta, V. Akella and S.J.B.Yoo, "All-optical physical layer nack in awgr-based optical interconnects," *Photonics Technology Letters, IEEE*, vol. 24, no. 5, pp. 410 –412, March 2012.

[9] T. Engset, "On the calculation of switches in an automatic telephone system," in *Tore Olaus Engset*, A. Myskja and O. Espvik, Eds., 1998.

[10] C. Gallep and E. Conforti, "Reduction of semiconductor optical amplifier switching times by preimpulse step-injected current technique," *Photonics Techn. Letters, IEEE*, vol. 14, no. 7, pp. 902 –904, July 2002.

[11] T. Bonald, "Insensitive traffic models for communication networks," *Discrete Event Dynamic Systems*, vol. 17, 2007.

[12] H. L. Vu, A. Zalesky, E. Wong, Z. Rosberg, S. Bilgrami, M. Zukerman, and R. Tucker, "Scalable performance evaluation of a hybrid optical switch," *Journal of Lightwave Technology*, vol. 23, no. 10, pp. 2961 – 2973, Oct. 2005.

[13] S. Ben Fredj, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," in *Proc. of ACM SIGCOMM*, 2001.

[14] R. Serfozo, *Introduction to Stochastic Networks*. Springer, 1999.

[15] V. Eramo and M. Listanti, "Power consumption in bufferless optical packet switches in soa technology," *IEEE/OSA Journal of Optical Comm. and Net.*, vol. 1, no. 3, pp. B15 –B29, August 2009.

[16] Cisco Systems, "Cisco Nexus 5548 Data Sheets," 2012.

[17] B. Ramamurthy, G. Rouskas, and K. Sivalingam, *Next-Generation Internet: Architectures and Protocols*. Cambridge Univ. Press, 2011.

[18] D. Cuda, R. Gaudino, G.A. Gavilanes, F. Neri, G. Maier, C. Raffaelli and M. Savi, "Capacity/cost tradeoffs in optical switching fabrics for terabit packet switches," in *Proc. of ONDM*, feb. 2009, pp. 1 –6.

[19] T. Bonald, D. Cuda, R.-M. Indre, and L. Noirie, "Feasibility of optical switch-combiners," LINCS, Tech. Rep., 2012. [Online]. Available: http://perso.telecom-paristech.fr/∼bonald/Pub/doc.pdf

[20] E. Sackinger, *Broadband Circuits for Optical Fiber Communication,*. New York,: NY: Wiley, 2005, vol. 1.

[21] G. P. Agrawal, *Fiber-Optic Communication Systems*. Wiley, 2002.

[22] Sun Telecom., "SUN-GE8100 OLT Data Sheet." 2012.

[23] T. Benson, A. Akella, and D. Maltz, "Network traffic characteristics of data centers in the wild," in *Proc. of IMC*, 2010.