TCP traffic carrying capabilities of OBS-based hypercubes for datacenters

Pablo Jesus Argibay-Losada[a], Chunming Qiao[b], Limei Peng[c] and Wan Tang[d]

[a]Telematic Engineering, University of Vigo, Campus Universitario s/n 36310 Vigo Spain
[b]Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260-2000 USA
[c]Electronic and Information Engineering, SooChow University, Suzhou City, Jiangsu, China
[d]Computer Science, South-Central University for Nationalities, Wuhan, Hubei 430074, China

ABSTRACT

High power consumption due to O/E/O processing in many conventional electronic datacenter networks can be greatly decreased by a suitable use of all-optical switching technologies. OBS is specially suited to perform this task given its bursty-data traffic friendly mode of operation. In this paper, we evaluate through analysis and simulation the performance of both OBS and electronic networks when used to carry TCP flows inside a 6D-hypercube, a highly symmetrical topology representative of datacenter networks.

Keywords: Optical burst-switching, datacenter, TCP, hypercube

1. INTRODUCTION

The high power requirements of current electronic networks present concerns that are specially significant in datacenter environments due to the high switching densities there. All-optical technologies are being studied as a means to avoid O/E/O processing; however, they continue to be of an experimental nature due to the hurdles they have to overcome in terms of needed hardware, cost, algorithmics and, ultimately, final performance with common applications. Optical Burst Switching (OBS)[1] is an optical switching technology that is appealingly simple and flexible due to strong functional similarities to today's packet-switching. The fact that transmissions in a datacenter are "short haul", in terms of both distance and hops, has also made it a lot easier to implement OBS. In this paper we present an study of the performance of TCP flows in a two-class OBS network inside a highly-symmetrical topology representative of datacenter environments, with and without deflection routing (DR).

In OBS, packets —called bursts in OBS terminology— are sent in all-optical form from source to destination and are composed of a variable number of conventional packets aggregated at the OBS edge. Edge nodes create the bursts and schedule them for transmission; once transmitted, bursts do not undergo O/E conversion until destination, performing instead cut-through across the core switches. Contention is conceptually similar to the one found in packet-switching environments. OBS can be used with traffic profiling, in such a way that in-profile and out-of-profile bursts coexist, and, therefore, arbitrary policies can be defined in the core switches to guide the contention resolution processes; this adds a significant degree of flexibility to the base OBS network in case the network operator needs it. OBS presents decentralized, essentially datagram packet-switched operation, with the flexibility, simplicity and efficiency gains of those environments. Interestingly, a commonly cited downside of OBS architectures is that, like the majority of its datagram analogues, it leads to packet losses that can be exacerbated by the lack of optical buffers inside the core switches. In this paper we show, through simulation and analytical tools, that this conventional wisdom is not well-founded, since the bufferless characteristics of OBS networks allow them to manage traffic patterns considerably smoother than their electronic counterparts, and the effect of TCP congestion control allows to have operating points suitable for end-user applications. These facts, in turn, ultimately lead to requirements for network dimensioning that are reasonable in comparison with the ones for their electronic counterparts, while presenting both the power consumption benefits of all-optical technologies and the simplicity of decentralized operation compared with other all-optical proposals. OBS is, therefore, a viable candidate for further research in high-bandwidth network environments where power savings are at a premium.

Photonics and Optoelectronics Meetings (POEM) 2011: Optical Communication Systems and Networking, edited by Zisen Zhao, Richard Penty, Chester Shu, Tao Jiang, Proc. of SPIE Vol. 8331, 83310A · © 2012 SPIE · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.917098

This paper is organized as follows: in Section 2, we propose an analytical model for studying TCP performance in a symmetrical two-class OBS 6D-hypercube network without wavelength conversion with and without deflection routing; we also complement this study by means of simulation. Section 3 presents a comparative analysis with an equivalent electronic counterpart; this also serves as a tool for dimensioning OBS networks like the ones described before. Section 4, finally, concludes the paper.

2. MODEL DESCRIPTION

The main performance measures in an OBS network can be estimated by taking into account that the absence of optical buffers in the core leads to traffic patterns less bursty than those in their electronic counterparts. This allows to make a suitable use of Markovian models to estimate burst blocking probabilities. We also assume that wavelength conversion is not available and that TCP —or a TCP-friendly transport protocol— is used to manage congestion. Let's denote by L and R the sets of links and routes in the network, m the number of fibers in each link, $R(l)$ the set of routes traversing link l, $x(r,l)$ the ordinal number associated to link l in route r, N the number of TCP flows in each route, T the average value of the TCP retransmission timer,[2] $t_{\rm pkt}$ the average burst size in bytes, and C the wavelength bandwidth. In addition, denote by $a_l^i, b_l^i \quad i \in$ {HI,LOW} the (unknown) offered traffic intensity in Erlangs and burst loss probability (BLP) of class-i bursts at link l (by default, 80% of bursts offered to any route are HI), a_l and b_l the (unknown) offered traffic intensity and average BLP at link l (based on a_l^i, b_l^i), and t_r, h_r, λ_r, p_r and a_r, respectively, the average RTT, hop count, sending rate in packets per unit of time, packet loss probability and traffic intensity for packet flows transported in bursts following route r in the OBS network. Then, in a symmetric topology with conservative schedulers we can write

$$b_l = E_B(m, a_l) \tag{1}$$

$$p_r = 1 - (1 - b_l)^{h_r} \tag{2}$$

$$\lambda_r = N \frac{1}{t_r \sqrt{\frac{2p_r}{3}} + T \min(1, 3\sqrt{\frac{3p_r}{8}}) p_r (1 + 32p_r^2)} \tag{3}$$

$$a_r = \lambda_r \frac{8 \cdot t_{\rm pkt}}{C}, \qquad a_l = \sum_{r \in R(l)} a_r \cdot (1 - b_l)^{x(r,l)-1} \tag{4}$$

for each $l \in L$, $r \in R$. Eq. (1) gives the dependence of the losses with the traffic dynamics through an Erlang system (conservative assumption), while Eq. (2) and Eq. (3) give the end-to-end performance measures experienced by applications in terms of loss and TCP throughput, respectively, and Eq. (4) relates the traffic intensities offered to routes and links as a function of the routing processes.

In the above equations, the main unknowns are b_l and a_r, interrelated by the Erlang-B (denoted by E_B) and TCP congestion control formulas. For example, in a 6D-hypercube with 64 nodes (each with 6 links), this results in a system with 14400 equations (a_l and b_l for the HI, LOW traffic and their averages/total traffic in each of the 384 links plus the λ_r, p_r and a_r for each of the 4032 routes). Solving them, performance measures such as the BLP (or throughput) of the TCP flows, as a function of the number of TCP flows per route, N, can be obtained.

The main assumption of this model are the presence of equivalent traffic patterns at the input of each link —usually a conservative assumption— and the Markovian assumption for those processes, which in networks without buffers is an adequate assumption.

To determine the effect of DR, due to symmetrical considerations in a 6D-hypercube, we note that any given link l will get the offered load z corresponding to the non-DR case, plus the blocked $zb_l/4$ from each of the 4 links that can deflect traffic to that link, resulting in an additional load of zb_l; from there, a conservative, upper bound for the end-to-end loss probability in a given route can be estimated by allowing only a single deflected route for each original one: for the 1-hop paths, we establish a deflected path with two additional hops, and for the 2,3,4,5 and 6-hop paths, we can establish a deflected path with the same length as the original *.

*The potential impact of reorder-related problems on TCP throughput is minimal in this DR-enabled OBS network.

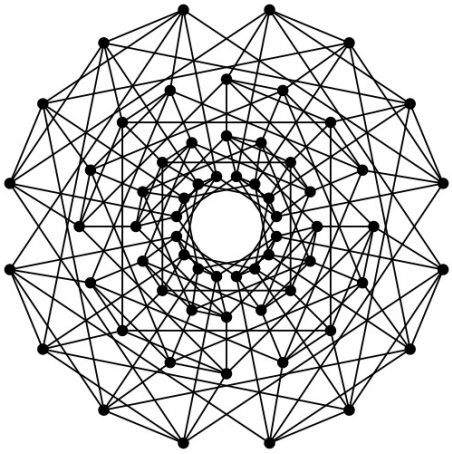


Figure 1. A 6D-hypercube

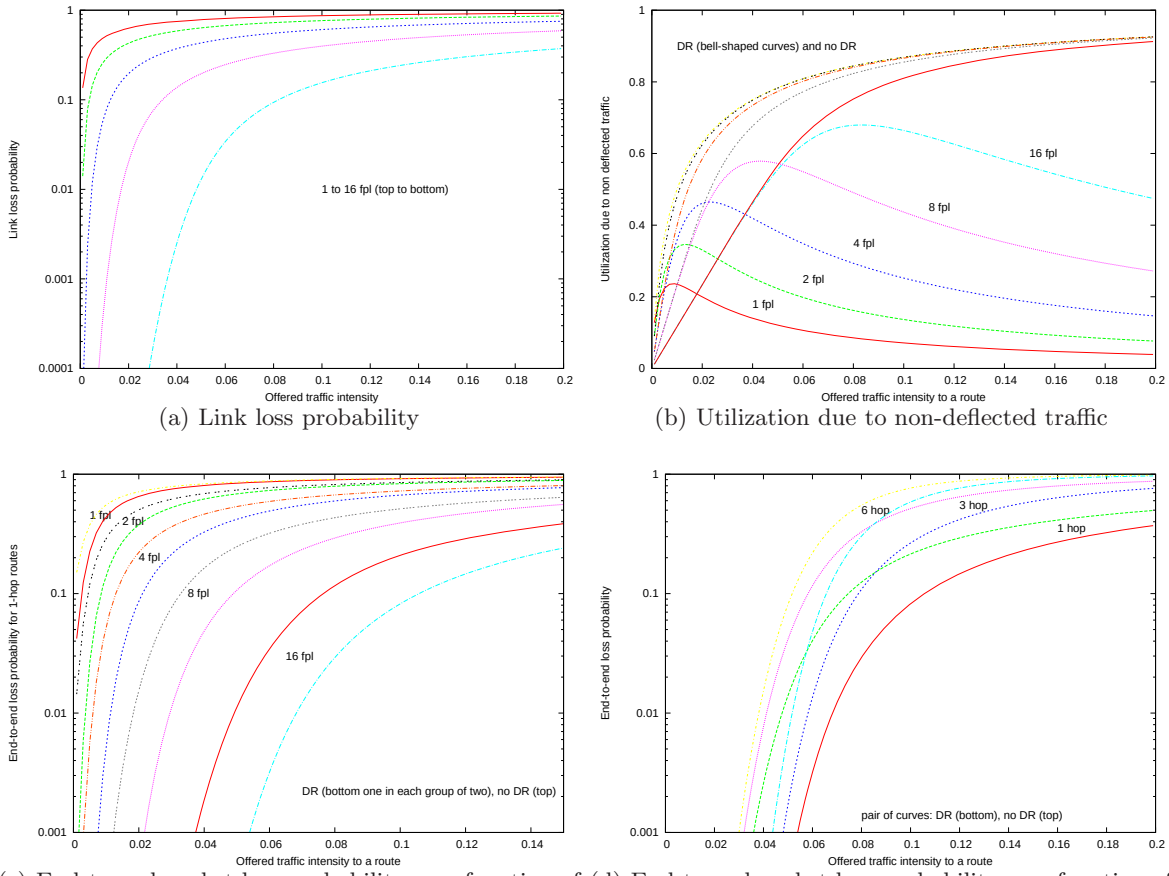
Let's consider an extension of the service offered by a plain OBS network by defining traffic profiles for traffic engineering, and resources provisioning purposes. Bursts are classified as HI and LOW according to whether they follow the profile or not. In this way we can assign a high priority to HI bursts and optimize their performance. If schedulers are configured to drop LOW traffic deterministically whenever it collides with HI traffic, then the performance of the HI and LOW streams can be estimated extending the previous model with:

$$b_l^{\mathrm{HI}} = E_B(m, a_l^{\mathrm{HI}}) = \frac{\frac{\left(a_l^{\mathrm{HI}}\right)^m}{m!}}{\sum_{i=0}^{m} \frac{\left(a_l^{\mathrm{HI}}\right)^i}{i!}}, \qquad b_l = E_B(m, a_l = a_l^{\mathrm{HI}} + a_l^{\mathrm{LOW}}) = \frac{\frac{a_l^m}{m!}}{\sum_{i=0}^{m} \frac{a_l^i}{i!}} \qquad b_l^{\mathrm{LOW}} = \frac{a_l b_l - a_l^{\mathrm{HI}} b_l^{\mathrm{HI}}}{a_l^{\mathrm{LOW}}}$$

where we take into account that the HI traffic has strict priority over the LOW one, allowing HI bursts to preempt LOW bursts, and that schedulers are conservative, allowing us to compute the average blocking of the combined HI and LOW streams in a straightforward way; from it, we can compute the LOW blocking probability using b^{HI}. The main assumptions of this model, apart from the ones present in the previous one, are the conservative system, which can lead to slight discrepancies depending on the burst size distribution, and the effect of preempted LOW bursts whose packets have already been transmitted from a given node; this last effect, however, is small and can therefore be considered negligible to estimate network performance.

Fig. 1 shows the datacenter network under consideration: a 6D-hypercube topology, with 64 nodes connected in a totally symmetrical way, giving rise to 4032 different routes. Each node connects one or more servers to the datacenter. Fig. 2(a) depicts the link burst loss probability (BLP) as a function of the offered traffic intensity to the link, for several values of the number of fibers per link; this figure is independent of the transport layer protocol used —i.e., TCP, UDP or other; i.e., we are using Eq. (1) to analyze the performance.

Qualitatively, we can see that the BLP performance is easily modified by the number of fibers in each link. For example, taking into account that many high-capacity links in the Internet are dimensioned for utilizations around 10-20%, if we use those values we achieve around 0.1%-1% losses when there are 4 fibers per link, and the BLP rapidly decreases with more fibers: with 16 per link, the offered traffic intensity can be 0.5 if losses are to remain under 0.1%, i.e., an average utilization of around 50%. There are also techniques that can be used in OBS to try to improve performance, like deflection routing (DR), where a given burst is routed through an available output link when the desired one is busy. Fig. 2(b) to 2(d) show the performance with and without DR. Given that DR forces bursts to take longer paths than the minimum in exchange for improving contention, it is expected that, as load increases, the majority of traffic in a given link will be due to this new, deflected traffic, as we can see in Fig. 2(b). Nevertheless, the net effect of DR is positive when taking a look at the average end-to-end losses in Fig. 2(c), although somewhat small; this is direct result of the fact that deflected traffic uses more network capacity than the non-deflected one. Fig. 2(d) show the gains for the case of 16 fibers per



(a) Link loss probability

(b) Utilization due to non-deflected traffic

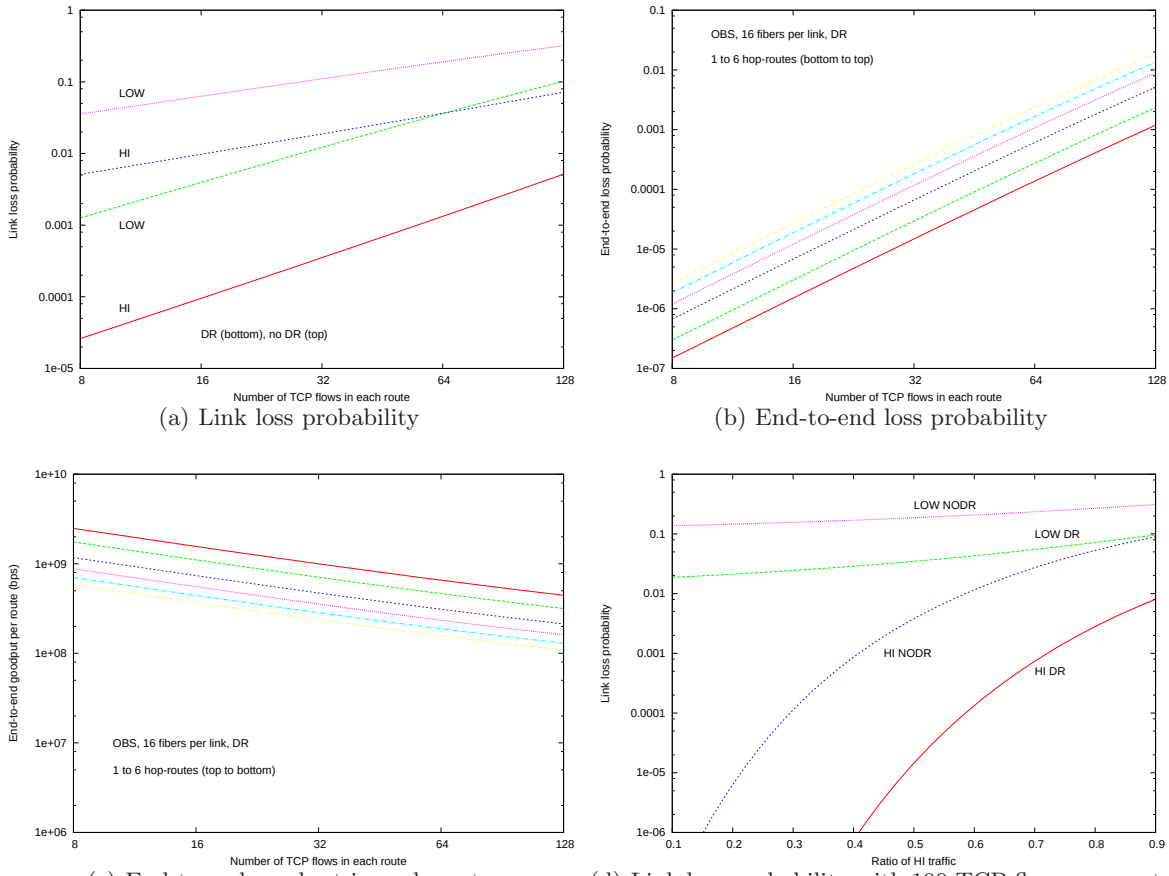
(c) End-to-end packet loss probability as a function of m for 1-hop routes

(d) End-to-end packet loss probability as a function of h_r

Figure 2. Performance in a transport-protocol agnostic way with one-class OBS

link as a function of the path length, showing the same qualitative behavior as before, but with the gains more noticeable for the shorter paths.

As in the vast majority of network environments, TCP, in some of its multiple variants, is used in the servers to provide flow and congestion control, and reliable transmission if needed. Therefore, once we have studied the basic performance measures in OBS in a protocol-agnostic way, now we can take into account the specific behavior of TCP to check which operating points, from the ones shown in the previous figures, it stabilizes around: as TCP peforms its congestion control, it adapts its transmission rate with the result that the BLP is within a reasonable range, as we will see next. These operating points will be sensitive to a myriad of variables both in the end-point TCP stacks and in the network configuration, and here we study performance when the network is the single element determining it. For the time being, we assume that there is a fixed number of TCP flows in each of the 4032 routes. Also, as said before, it is easy to add flexibility to a plain OBS network if we introduce the idea of traffic profiles for the traffic entering the network. Traffic that follows the profile is marked as HI, and the remainder as LOW. Arbitrary policies defined by users or the network operator decide which bursts are marked which way. When under contention in core switches, HI bursts have priority over LOW ones, preempting them. Therefore, as a function of the policy to mark bursts, the BLP of the original OBS network will branch into two components: a lower one for the HI traffic, b^{HI}, and a higher one for the LOW traffic, b^{LOW}. Adjusting the proportions of traffic marked as HI and LOW, the network has a flexible knob to achieve the desired performance for HI bursts, while providing service to the LOW traffic when network load allows to do so. In addition, and appealingly for TCP flows, if we consider the option of dispatching packets from the same flow into both classes of bursts, we can achieve an indirect and powerful method for controlling the throughput



(a) Link loss probability

(b) End-to-end loss probability

(c) End-to-end goodput in each route

(d) Link loss probability with 100 TCP flows per route (403200 total)

Figure 3. TCP performance with two-class OBS

of the individual flows —i.e., if a given flow sends more packets in HI bursts than another flow, it receives more throughput, and this is a function of the specific percentage of packets marked as HI; this is not only due to less losses, but also to the reaction of TCP congestion control to those losses. This leads to a two-pronged approach to control throughput inside the network: by means of the proportion of HI to LOW bursts, and by means of the proportion of packets from a given flow that are sent through each class of burst.

Figs. 3(a),3(b) and 3(c) show the link loss, end-to-end losses and end-to-end goodput as a function of N for all route lengths, when 80% of the traffic is marked as HI. The figures assume an RTT of 10ms in each route, including processing times in the end-points —we assume therefore that all routes have the same delay irrespective of the number of hops due to the absence of buffers and the small geographical distances involved in datacenters, which cause delays to be usually dependent on factors independent of route length, like load in the end servers. We can see how easily the HI bursts suffer more than one order of magnitude less losses than the LOW ones. Moreover, we can estimate the performance of each kind of traffic in a more general way by means of the percentage of HI traffic in the total mix. The plot in Fig. 3(d) shows that dependence: the evolution of b^{HI} and b^{LOW} when that percentage changes and we have 100 flows per route (403200 total). We can see that it is easy to achieve very low loss probabilities for the HI traffic by means of tightening the marking policy at the network edge; for example, if 50% of all traffic is HI, there is a difference of more than 3 orders of magnitude between them, essentially implying loss-free transmission for HI bursts. All the while, LOW traffic continues to be routed in a more best effort basis, being able to use available bandwidth when it is free. This is achieved in conjunction with the power savings that can be expected of an all-optical network and of the easiness of operation of a datagram-like network architecture.

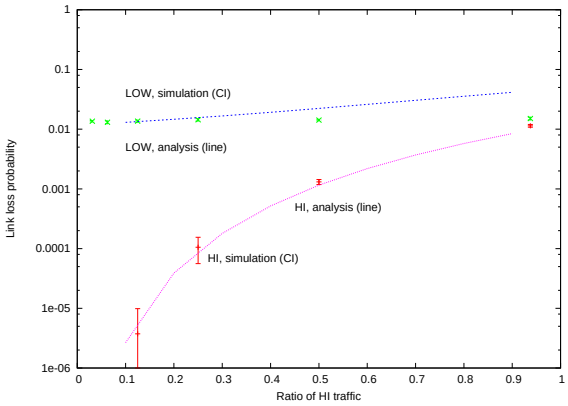


Figure 4. Simulation vs. Analysis

In addition, taking into account that TCP is a bandwidth-probing protocol, that usually injects packets into the path until losses are encountered, and it performs well with losses ranging from 0.1% to 10%, we can see from the figure that the OBS hypercube allows a large numbers of TCP flows with good performance. For example, at 1% losses, we can have 100 full TCP flows in each route (a total of 403200 concurrent, long-lived ones in the full topology). Comparing Figures 3(a) with 2(a) we can see that TCP stabilizes its rate in an operating point where it is able to take advantage of the presence of DR while the number of flows is not too large.

In summary, TCP in a DR-enabled, two-class OBS network is able to work in operating points whose performance is adequate for end-user applications.

2.1 Simulation

To complement our previous analytical study, we also show in Fig. 2.1 a simulation study of the two-class OBS network previously described. We are using offset times of 10ms, control packet processing times of 1μs, $m = 4$, and we have that preempted LOW bursts continue to use their allotted time downstream and participate in contention processes. Bursts are generated in all 4032 routes each millisecond and, in order not to improperly delay packets in the datacenter, bursts consist of a single packet. According to measurements of large datasets in the Internet[6] and in order to have a relatively conservative assumption, we model the aggregation of the TCP flows in each route as an M/G/∞ process with a strong degree of temporal dependence given by a Hurst parameter $H = 0.9$. Fig. 2.1 depicts the 95% confidence intervals for b^{HI} and b^{LOW} as computed by simulation of this OBS network, and the analytic results for the case of using 8 flows. We can see good correspondence between analysis and simulation, validating our previous results: OBS achieves good performance for end-user applications, and simple traffic profiling at the edge and simple contention resolution schemes in the core allow for flexible performance differentiation by means of two-class OBS, if needed by the network operator.

3. OBS VS. ELECTRONIC

To complement the previous analyses, it is also useful to compare the performance of the OBS hypercube with the one in an equivalent electronic hypercube carrying the same number of TCP flows. In order to estimate the loss probabilities and delays inside the buffers, we have to note that, as always in the cases involving buffers, they are highly dependent on the traffic processes injecting traffic into the network and on the configuration of all the elements inside it. As there is no single practical analytical tool that can be used to study any arbitrary case, we are going to resort to widely available analytical models that are able to abstract the input traffic characteristics in order to consider the effect of the correlations in the input traffic and the buffer size. Specifically, we are going to make use of gaussian processes, because they are specially suitable when traffic results from the aggregation of many sources, as expected in large datacenter networks, and because they are amenable to analysis under arbitrary autocorrelation structures. The ones we use here are based on a refinement[4] of the well known maximum variance asymptotic (MVA) for estimation of the overflow probability in infinite buffer queues.[5] This refinement, also called the MVA for loss, modifies the estimates given by the MVA to take

into account the effect of a finite buffer size in order to compute the corresponding packet loss probability. In this situation, for a link with bandwidth c, buffer size k, gaussian input traffic with average intensity A, standard deviation σ and autocorrelation function $r(i)$, the packet loss probability can be estimated by means of:

$$p(k) = F(A) \exp\left(-\frac{m_k(A)}{2}\right)$$

with

$$F(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(\frac{(c-x)^2}{2\sigma^2}\right) \int_c^\infty (r-c) \exp\left(-\frac{(r(1-x/c))^2}{2\sigma^2}\right) \mathrm{d}r$$

$$m_k(x) = \frac{1}{\max_{n\geq 1} \gamma\left(k, c, x, n\right)}$$

$$\gamma(k, c, x, n) = \frac{n\sigma^2 + 2\sigma^2 \sum_{i=1}^{n-1} (n-i) r(i)}{(k + (c-x)n)^2}$$

The availability of buffers inside the network will cause the formation of packet bursts, naturally inducing several kinds of correlations as a function of the exogenous processes feeding the entire network. In order to sweep a broad range of possible traffic structures, we are going to consider the following:

- short-range dependent (SRD) processes with strong autocorrelation structures. Specifically, we are going to use an autoregressive process (AR) whose first lag has a value of 0.9, as representative of a strong SRD case —i.e, slow decay rate for the autocorrelation function—

$$r(i) = w^{|i|}, w = 0.9 \tag{5}$$

- long range dependent (LRD) processes with strong autocorrelation structures. Specifically, we are going to use a fractional Brownian motion (fBm) process with Hurst parameter $H = 0.9$ as representative of strong LRD

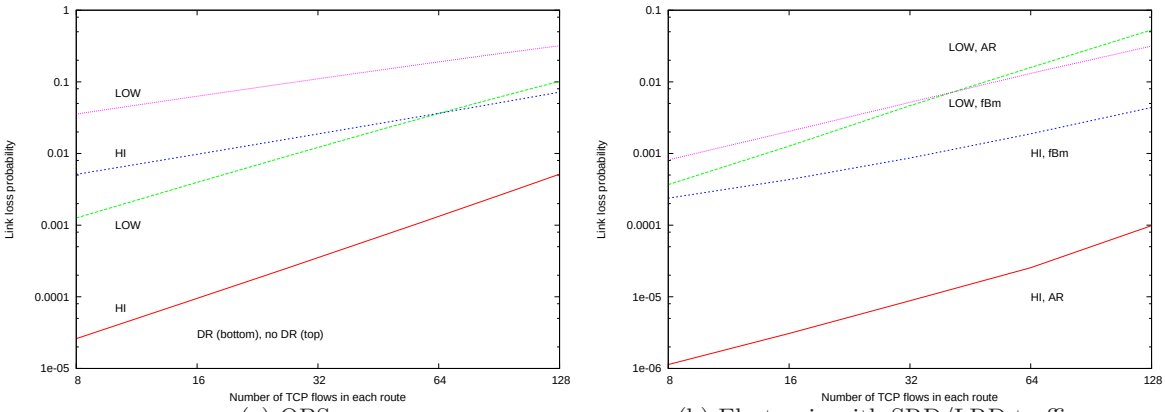
$$r(i) = \frac{1}{2}\left(|i-1|^{2H} + |i+1|^{2H} - |i|^{2H}\right)$$

We can also make an estimation for the buffer delay taking into account that, due to packets arriving in bursts, many of them will find near full queues, so the worst case delay will be given by the time used to service a full system:

$$d_2 = (k+1)s$$

Accordingly, Fig. 5(b) shows the link packet loss probabilities in a conventional electronic network —which uses buffers instead of DR and additional fibers for contention resolution— for the SRD and LRD cases described above. The main conclusion that we can extract here is that the maximum number of TCP flows to achieve a given packet loss probability is roughly equivalent to the OBS case, plotted in Fig. 5(a). For example, OBS losses are in the range 0.1%-10% for LOW bursts, as N goes from 8 to 128, and these values are similar to those we can see in the electronic SRD and LRD cases. Moreover, losses for HI bursts are around 1.5 orders of magnitude lower than those for LOW bursts in OBS, a level right between the one for the electronic LRD (around 0.5 orders of magnitude lower) and SRD cases (2.5 orders of magnitude lower). Here we can observe one of the features of OBS that plays at its advantage: the absence of buffers leads to a weaker autocorrelation structures for the arriving bursts,[3] which leads to smoother, less bursty traffic processes more easily manageable.

In summary, an OBS network can be reasonably configured —by making use of DR and a reasonable value for the number of fibers per link— to offer a performance level equivalent to the one provided by a conventional electronic network; this constitutes a specially noteworthy result, given lingering concerns in the literature regarding potential loss-induced performance problems in bufferless optical networks like OBS.



(a) OBS (b) Electronic with SRD/LRD traffic

Figure 5. OBS vs. Electronic

4. CONCLUSION

OBS-based transmission is a viable technology to respond to power constraints in the datacenter. With the current transport protocol of choice, TCP, it is able to lock into operating points with effective performance for end-user applications in a resource-efficient way. Existing concerns about the suitability of bufferless all-optical architectures like OBS are not well founded, since they are based on results obtained without taking into account the interrelation between TCP congestion control and the smooth traffic dynamics found in bufferless networks. Our study shows that a reasonably dimensioned OBS network can achieve a performance level equivalent to the one of an electronic counterpart. Therefore, the choice between the two will be given by the maturity of optical technology vs. power savings, and not by concerns about the levels of service that applications can experience in OBS networks.

Acknowledgment

This research is supported in part by the National Natural Science Foundation of China (61103248).

REFERENCES

[1] Y. Chen, C. Qiao and X. Yu, "Optical burst switching: a new area in optical networking research," IEEE Network, vol. 18, no. 3, pp. 16–24, may 2004.

[2] J. Padhye, V. Firoiu, D. Towsley and J. Kurose, "Modeling TCP Reno performance: a simple model and its empirical validation," IEEE/ACM Transactions on Networking, vol. 8, no. 2, pp. 133–145, apr. 2000.

[3] M. Grossglauser, J.C. Bolot, "On the relevance of long-range dependence in network traffic," IEEE/ACM Transactions on Networking, vol. 7, no. 5, oct. 1999.

[4] H.-S. Kim and N. Shroff, "Loss probability calculations and asymptotic analysis for finite buffer multiplexers," IEEE/ACM Transactions on Networking, vol. 9, no. 6, pp. 755–768, dec. 2001.

[5] J. Choe, N.B. Shroff, "A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks," IEEE/ACM Transactions on Networking, vol. 6, no. 5, pp. 659–671, 1998.

[6] T. Karagiannis, M. Molle, M. Faloutsos and A. Broido, "A nonstationary Poisson view of Internet traffic," IEEE INFOCOM 2004.