A Novel Approach to Optically Switching Inter-Pod Traffic in Datacenters

Limei Peng[1], Chunming Qiao[2], Wan Tang[3], Chan-Hyun Youn[1], Xinwan Li[4], Guiling Wu[4], Jianping Chen[4], Ting Wang[5]

[1]Department of Electrical Engineering, GRID Middleware Research Center, KAIST, Taejon, 305, Korea
[2] Department of CSE, SUNY Buffalo, Buffalo, New York 14260, USA
[3]College of Computer Science, South-Central University for Nationalities, Wuhan, 430074 P. R. China
[4]Department of EE, Shanghai Jiao Tong University Shanghai, 200240 China
[5]NEC Laboratories America, Princeton, NJ, USA
Emails: aurora_plm@kaist.ac.kr, qiao@computer.org

***Abstract*—In this paper, we propose and evaluate a novel approach to switching traffic between the pods (or containers) of servers all optically. We show that the proposed approach, called Labeled Optical Burst Switching with Home Circuit (or LOBS-HC) requires a fewer wavelengths (and transceivers) than existing approaches based on wavelength routing or optical circuit switching (OCS). In addition, it is also more suitable than optical burst switching (OBS) and electronic switching. Simulation results from OPNET confirm that LOBS-HC achieves good performance for dynamic inter-pod traffic that is common in a datacenter.**

Keywords- bursts; pods; home circuits; wavelengths; datacenter

I. INTRODUCTION

A large datacenter comprises of tens to hundreds of pods, each containing up to a thousand of servers. Each of the servers has a 1Gbps to 10Gbps Ethernet port, and may communicate with another server with a hard-to-predict and bursty traffic pattern. While it is feasible to interconnect all the servers in a pod using electronic (Ethernet) switches to provide full communication bandwidth, it is a daunting task to interconnect all the servers in this way. This is due to the fact that with electronic switching, a large number of wires and transceivers are needed to provide a full bisection bandwidth (i.e., with an oversubscription ratio of 1), which also implies a high cost and energy consumption. Even though exsiting datacenters with electronic switching can perform reasonably well with a larger oversubscription ratio, power consumption and potential communication bottleneck in the core are still major concerns.

Recently, hybrid optical and electronic switching architectures have been proposed wherein “mice” (i.e., short) flows between pods are switched electronically and “elephant” (i.e., long) flows are switched optically over wavelength circuits (also known as lightpaths) through a large MEMS switch whose switching time is on the order of a millisecond [1, 2]. Though such hybrid switchig approaches can alleviate some of the problems of all electronic switching, we believe they are only short-term compromise.

In this work, we propose a forward-looking approach that provides all-optical interconnects among the pods using multiple, small fast optical switching fabrics (whose switching time can be on the order of nanoseconds) to eliminate the needs for electronic switching outside the pods. More specifically, each of the *N* pods is connected to one or more such fast optical switching fabrics using WDM links, and there are *N* such optical switches in total, which form a cube-like topology among themselves.

Traffic from one pod to another is sent in units of *bursts* through these switches according to the *labeled optical burst swicthing with home circuits* (LOBS-HC) [3, 4]. LOBS-HC improves over OBS [5, 6] in that in OBS, burst contention will normally result in bust loss and hence a poor QoS guarantee. In LOBS-HC however, each source-destiantion pair of pods is allocated a HC, which provides a guaranteed long-term average bandwidth (say 10Gbps) to achieve *lossless* burst transmissions, and also supports bursty traffic (at say 100Gbps or 1Tbps). Compared to the approaches using all electronic packet-based switching, our proposed approach retains the benefits of agilitiy and statistical multiplexing gains as in OBS, but reduces the number of wires and tranceivers significantly comparing to OCS.

In addition, the proposed approach is much more wavelength resource efficient than approaches using MEMS-based optical circuit switching (OCS). This is mainly due to the *unique* feature/requirement of LOBS-HC that one can group multiple HCs (each of a sub-wavelength granularity) with the *same source and different destiantions* together*,* and assign them the same wavelength without requiring any O/E/O conversion (or tranceivers) at any intermediate nodes.

The corresponding challenge when applying the LOBS-HC concept is how to route and group multiple HCs so as to minimize the number of wavelengths needed to establish all the required HCs. The unique feature/requirement of LOBS-HC makes our problem different from the traffic grooming problem in OCS, and in fact, our problem has not been addressed for LOBS-HC networks of either irregular or regular topologies.

The paper makes the following three major contributions. To our best knowledge, this work is the first kind that proposes a future-proof datacenter networking architectures with fast optical switching for inter-pod traffic. Secondly, the paper proposes efficient algorithms to route HCs and assign them with wavelengths in rings, n-cubes and Generalized Hypercube (GHC), and shows that LOBS-HC uses a fewer wavelengths than OCS and a fewer switches, links and transceivers than electronic switching. Thirdly, the study evaluates the performance of the proposed LOBS-HC approach in the presence of dynamic traffic among pods via OPNET simulations and shows promising results.

The rest of the paper is organized as follows. In Section II, we first describe a *complementary HC assignment* (CHA) algorithm for a bidirectional ring (which is a generalized 2-dimensional cube) and then apply the concept of spanning balanced trees (SBT) to HC routing and grouping in a n-cube (where $n > 2$) and GHC. We also compare the proposed LOBS-HC approach with OCS and electronic switching, respectively, in terms of the number of wavelengths needed, and the number of switches, links and transceivers needed to provide sufficient bisection bandwidth. In Section III, we describe our simulation setup to evaluate the performance of LOBS-HC in the presence of dynamic traffic and discuss the results. We conclude the paper in Section IV.

II. Optimal HC Routing, Grouping and Wavelength Assignment in Cube-like Topologies

Compared to OCS and OBS, LOBS-HC has the unique feature of being able to provide bandwidth guarantee (as in OCS but not OBS) using HCs, while still supporting statistical multiplexing (as in OBS but not OCS) via HC multipelxing and opportunitistic burst transmissions over non-HCs. More specifically, in LOBS-HC, in-profile traffic for a source-destination (SD) pair, with respect to the guaranteed bandwidth for the SD pair, can be transmitted in a lossless fashion over a dedicated HC. In addition, multiple HCs from the same source to different destiantions can share the same wavelength. Moreover, when the HC from source S to destiantion D is busy, extra, i.e., out-of-profile, traffic from S to D can be sent in an opportunistic fashion over a HC from the same source S but to a different destination D that passes through D, *or a HC from a different source to the same destination that passes through this node*. Such out-of-profile traffic will be given a lower priority when competing with in-profile traffic of that HC, and hence will not affect the QoS guarantee for the in-profile traffic. As a result, LOBS-HC is expected to be more flexible and efficient that OCS and OBS.

Below, we discuss the HC routing, grouping and wavelength assignment for 2-dimensional cube/ring, n-cube (n>2) and generalized hypercube respectively.

A. Ring (Generalized 2-cube)

For ease of presentation, we start with a simple topology. Fig. 1 shows a datacenter with N pods interconnected in a bidirectional ring with N core optical switches, each of which can switch an optical burst of an arbitrary duration without O/E/O conversion.

In the following, we will assume that a pair of pods is provided with a guaranteed long-term average bandwidth of B_W, *say at* 10 Gbps (in a 100-pod datacenter, this implies that the total overall traffic through the core switches would reach almost 100Tbps). The burst rate at each pair of pods can be higher at e.g., 100Gbps. We denote by B the *normalized* bandwidth required by each HC, i.e., $B = B_W/C$, where C is the per wavelength capacity, which could be 40Gbps or 100Gbps. We also let $H = \lfloor 1/B \rfloor$ which denotes the number of HCs that can share one wavelength, and $G = \lceil N / H \rceil$ which denotes the number of groups of HCs that can share one wavelength.

The basic idea of the proposed HC grouping and wavelength assignment algorithm, called CHA, is to groom exactly H HCs from the same source pod i, whose hop lengths increase consecutively from jH+1 to $(j$+1$)H$, where $0 \le j \le G/2$-1 (assuming an even G), onto the same wavelength. The same wavelength is then reused spatially by source pods i+H, i+2H, …, i+$(G$-1$)H$ mod N in the same way. Assuming H=3, Fig. 2 is a partial example of using CHA in a ring consisting of N=12 pods (with G=4). It shows the two cases (or two stages of the algorithm) with j=0 and 1 respectively, where the dashed, dotted and solid wavelengths are used to establish the HCs needed by source pods 0, 3, 6 and 9 in the clockwise direction.

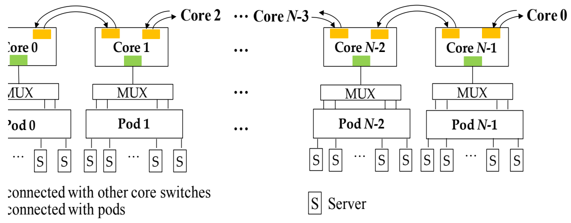

Figure 1. A LOBS-HC ring: the core switches (top) and the pods (below).

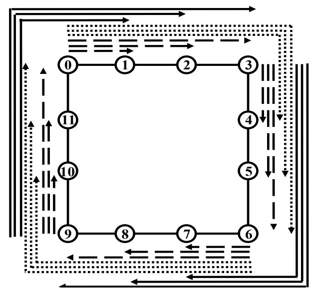

Figure 2. Wavelength Assignment for HCs using CHA

We can generalize the above idea into an algorithm that establishes all the HCs originating from source pods i, i+H, …, i+$(G$-1$)H$, whose hop lengths increase consecutively from jH+1 to $(j$+1$)H$ in stage/group j using j/2 wavelengths. More specifically, the total number of wavelengths required in all the G/2 (i.e., the G/2 groups in clockwise direction and another G/2 in the counter-clockwise direction can use the same wavelength sets) stages to establish all the HCs, denoted by W_G, can be calculated by (1).

$$W_G = \begin{cases} 1 + 2 + .. + \frac{G+1}{2} = \frac{(G+1)(G+3)}{8}, & \text{when } G \text{ is odd} \\ 1 + 2 + .. + \frac{G}{2} = \frac{G(G+2)}{8}, & \text{when } G \text{ is even} \end{cases} \quad (1)$$

We will omit the proof that W_G is also the minimum number of wavelengths needed, but suffice it to say that CHA is optimal since it wastes minimal wavelength resources.

Accordingly, we present the total number of wavelengths, denoted by W_{TTL}, needed to establish all the HCs originating from all N nodes is W_{TTL} by (2) without further proof. We also note that the total number of transceivers per pod is G, which makes the proposed approach quite feasible.

$$W_{TTL} = \begin{cases} H(G+1)(G+3)/8 + R(G+1)/2, & \text{when } G \text{ is odd} \\ HG(G+2)/8 + RG/2, & \text{when } G \text{ is even} \end{cases} \quad (2)$$

where $R = N \bmod H$.

B. N-cube and Generalized Hypercube (GHC)

We now extend our study by replacing the ring in Fig. 1 with a n-cube ($n > 2$) and n-dimensional k-ary generalized hypercube ($\mathrm{GHC}_{n,k}$) as shown in Figs. 3 and 4. Here, optimal HC routing, grouping and wavelength assignment is tricky since in a n-cube, there are n disjoint shortest paths between two nodes, and many more non-shortest paths to choose from.

In Fig. 3, the core switches are interconnected in a 3-cube topology, with each core having three direct neighbors, one in each dimension. Each node is associated with a binary number and assumed to have a pair of WDM fibers to carry outgoing/incoming traffic between pods. The optimal interconnection between each pod, e.g., node 000, and the core switches is to use one fiber to interconnect pod 000 to core switch 000, and another to interconnect pod 000 to core switch 111. Such a pod-to-core interconnection pattern can effectively increase the bisection bandwidth and reduce inter-pod communication latency as well as improve reliability, compared to a naive interconnection where all the two fibers are used to interconnect pod 000 to core 000 for example.

Our basic idea to minimize the wavelength resources needed is to route all the HCs along their shortest paths in such a way that *load balance* is achieved. In other words, from any given source pod, the number of its N-1 HCs should ideally be distributed evenly among the n outgoing (fiber) links. In this way, we can minimize the maximum number of wavelengths needed on any given link.

To achieve the load-balanced routing of the HCs, we adopt the concept of a spanning balanced tree (SBT) in n-cubes and GHCs [7, 8]. Assuming a 5-cube with N=32 nodes, each of which is represented by a binary number, Fig. 5 shows an example SBT rooted at node 00000 that contains 5 sub-trees (STs), each providing shortest paths to 6 or 7 destinations. This implies that there will be at most 7 HCs originating from 00000 going through any link in the 5-cube.

We omit the description on how to generate a SBT but suffice it to say that there may be more than one SBT. In addition, let the maximum number of destination nodes contained in a ST be denoted by max(n), which is lower bounded by (N-1)/n, we observe that the number of wavelength required by a SBT is $W_{SBT} = \lceil max(n)/H \rceil$. Given the symmetric property of a n-cube and a GHC, HCs originating from node h=$a_{n-1}a_{n-2}\dots a_0$, can reuse the same wavelengths as those originating from node e = $\overline{a_{n-1}a_{n-2}\dots a_1 a_0}$. Intuitively this states that 2 "diagonal"/exclusive nodes in a n-cube and k "diagonal"/ exclusive nodes in $\mathrm{GHC}_{n,k}$ can reuse the same wavelengths. Therefore, the total number of wavelengths needed by N SBTs denoted by W_{TTL} is calculated by (3), where d = 2 for n-cube, and d = k for $\mathrm{GHC}_{n,k}$. The number of transceivers per pod, denoted by P_{tx}, is given by (4) where l = n for n-cube and l = $n\times(k$-1) for $\mathrm{GHC}_{n,k}$. It can be seen that as C increases (to e.g., 1Tbps), H will increase linearly while these two numbers will both decrease linearly.

$$W_{TTL} = d^{n-1} \times W_{SBT} = d^{n-1} \times \lceil max(n)/H \rceil \quad (3)$$

$$P_{tx} = l \times \lceil max(n)/H \rceil \quad (4)$$

C. Numerical results

In this section, we compare the minimum number of wavelengths required using LOBS-HC and OCS, respectively, in a datacenter with a hypercube-like interconnection. We also compare the network cost in terms of the number of switches, links and transceivers required using LOBS-HC and electronic packet switching.

The number of wavelengths needed in a LOBS-HC ring using CHA with that needed in an OCS ring with wavelength routing is firstly compared. For OCS, we consider two cases, one with transient traffic grooming (TTG) and the other without TTG. Since TTG requires O/E/O conversion in OCS, the case with full TTG is similar to the case with electronic switching in terms of both the number of wavelengths and transceivers needed. In particular, with full TTG, each pod can multiplex outgoing traffic to different destination pods onto the smallest number of wavelengths, thus cutting down on the number of wavelengths and transceivers needed at each pod.

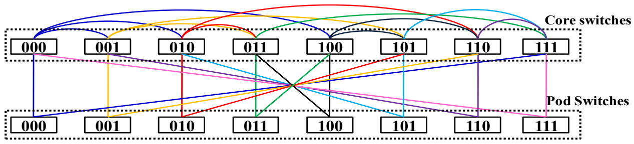

Figure 3. The logical interconnection for 3-cube.

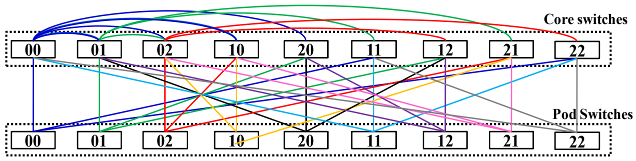

Figure 4. The logical interconnection for $\mathrm{GHC}_{2,3}$.

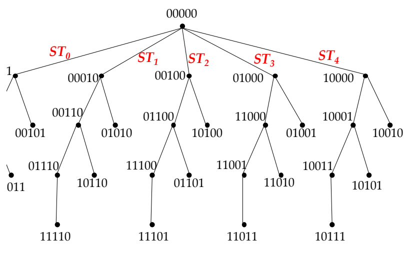

Figure 5. A SBT for node 00000 in the 5-cube

However, such an approach is more costly than the LOBS-HC ring due to its additional O/E/O transceivers at the core switches.

Table I compares the number of wavelengths needed at each pod in a datacenter consisting of 100 pods, where the sustained inter-pod bandwidth is 10Gbps and each wavelength capacity is 100Gbps. In such a datacenter, we have H=10 and G=10. The results in Table I shows that the OCS ring (without TTG) requires many more wavelengths (and accordingly transceivers at the pods) than the LOBS-HC ring. In addition, even with full TTG, the reduction of the number of wavelengths needed from that needed in LOBS-HC is not significant, implying that it may not justify the increase in the cost of O/E/O at the core switches due to the use of full TTG in OCS or with electronic switching.

Table II compares the number of wavelengths needed in a 64-pod datacenter using either OCS or LOBS-HC when the interconnection topology used is a bidirectional ring, 6-cube, or $GHC_{3,4}$. Assuming that each pod contains 1,024 10GE servers, and the same values of C=100Gbps, H=10 and G=10 as before, the oversubscription ratios in the 6-cube and $GHC_{3,4}$ are 2 and 8, resp., which is below most of the oversubscription ratios in existing datacenters.

Note that in the case of LOBS-HC n-cube, for example, each core switch needs 6 input/out fibers to connect to other core switches and 2 fibers to a pod. Table III compares the complexity of using the proposed LOBS-HC 6-cube and $GHC_{3,4}$ with that of using the existing 10Gbps Ethernet switching based on the Fat-tree and BCube topologies, in terms of the total number of required links (denoted by L_{TTL}), core switches (denoted by S_C) and transceivers (denoted by N_{tx}) needed for inter-pod connections (or at the highest level in the case of BCube) to provide a *comparable* bisection bandwidth. Note that even though we allow the Fat-tree and $BCube_3$ in [9, 10] to have an oversubscription ratio of 8, the results in Table III show the great potentials in reducing the number of wires and transceivers (as well as the cost and power consumption) when using the proposed LOBS-HC implementations.

TABLE I. WAVELENGTHS NEEDED IN A 100-POD RING

	W_{TTL}
CS without TTG	1275
CS with full TTG	128
OBS-HC	150

TABLE II. WAVELENGTHS NEEDED IN A 64-POD RING, CUBE OR GHC

	W_{TTL}
-ring with OCS	528
-ring with LOBS-HC	116
cube with OCS	416
cube with LOBS-HC	64
$HC_{3,4}$ with OCS	112
HC with LOBS-HC	16

TABLE III. NUMBER OF LINKS, SWITCHES AND TRANSCEIVERS NEEDED

	L_{TTL}	S_C	N_{tx}
cube	320	64	768
$HC_{3,4}$	384	64	578
t-Tree	8192	128	16384
$Cube_3$	8192	256	16384

III. PERFORMANCE EVALUATION

In this section, we evaluate the communication performance of a 32-pod datacenter network using LOBS-HC through OPNET simulation experiments. The topology used in simulation is a 5-cube (with 32 pods). Each pod is assumed to need a 10Gbps HC to each and every other pod as a guaranteed connectivity, and has 5 output fiber links, each consisting of 16 100Gbps wavelengths (according to (2)). The routing path of each HC between each SD pod pair is set up according to the spanning balanced tree (SBT) construction introduced earlier (see Section II B Fig.5).

In our simulations, we only focus on inter-pod traffic, which is the data (Ethernet frames) that need to be sent to another pod. The average size of Ethernet frames are generated according to the exponential distribution with a mean value of 4000bits. The arrival process of Ethernet frames also obeys the exponential distribution with an interval time being calculated according to the desired load. More specifically, we vary the interval time so the amount of traffic generated changes from anywhere between 0.1 to 1 with respect to the maximum network capacity, which is 16*16*100Gbps or 25.6Tbps. After each outgoing Ethernet frame is generated, multiple frames with the same destination pod are assembled into a burst with the mean size of each burst being 200Kbits.

In terms of the selection of the destination pod of each generated outgoing Ethernet frame, we consider two different traffic distributions among the 32 pods, i.e., the uniform versus non-uniform distribution of traffic demands. In datacenters, a pod is likely to store or access data from pods that are its neighbors. Hence, when the pod generates frames for the non-uniform traffic pattern in our simulation, pods that are closer to it (in terms of hop distance in the 5-cube) always have higher probabilities to be generated as destination pods. As a representative case, we use the decreasing geometric sequence to generate such probabilities. That is, the probabilities are about 51.6%, 25.8%, 12.9%, 6.45%, and 3.25%, respectively, for pods that are 1 hop, 2 hops, 3 hops, 4 hops, and 5 hops away, respectively, to be chosen as the destination pod.

We assume that at each pod, there is one queue for each destination pod. A burst will be sent out over a designated HC right away if the HC is available. If not, the burst will be queued for later transmission. If the queue length exceeds a certain threshold, the head-of-line burst will be sent as an out-of-profile burst using a non-HC (for that SD pod pair).

We have measured the goodput and queuing delay as a function of the offered load under both uniform and non-uniform traffic distributions, and the results are shown in Fig. 6 and Fig.7. From the results, we can see that both the goodput and queuing delay of the non-uniform case outperforms those of the uniform case. That is due to the fact that in the non-uniform distribution, more than half of the traffic demand is transmitted to pods that are 1-hop away, while in the uniform distribution, the average number of hops is more than 2.5. Accordingly, more data can be delivered in non-uniform traffic distribution. Similarly, the queuing delay is shorter in non-uniform distribution. Overall, when the traffic load is below 0.5, the performance in both uniform and non-uniform cases is sufficiently good for datacenter applications. With non-uniform traffic, even at 100% load, both the goodput and delay achieved are still respectable.

To obtain insights into LOBS-HC's unique feature, we have also measured and compared the performance of the in-profile (InP) and out-of-profile traffic (OoP) as shown in Fig. 8 and Fig. 9. These results are obtained by setting the queue-length threshold to be 0. That is, when a new burst is generated and the HC is not available (i.e., busying in transmitting a previous burst), the new burst will be sent out as out-of-profile traffic.

As can be seen, in-profile traffic under both uniform and non-uniform distribution performs well, and better than the out-of-profile traffic in terms of both goodput and queuing delay. These results are reasonable as the in-profile traffic has a lossless transmission over a designated HC. Therefore, their goodput in both scenarios are very close to 1. However, their queuing delay is not that much smaller than out-of-profile traffic since in-profile bursts need to be buffered in the same queue as out-of-profile bursts. Nevertheless, these results show that LOBS-HC is more effective than OBS, as in OBS, none of the traffic would be treated as well as in-profile traffic. In our simulation, several different queue-length threshold values have also been tried. Fig. 10 shows that when the threshold is set to 2Mbps, the ratio of traffic sent as in-profile traffic to out-of-profile traffic is higher than that with a zero threshold. In particular, at a low load (e.g., 0.2), all traffic is sent as in-profile-traffic with a threshold equal to 2Mbps whereas only half of the total traffic is sent as in-profile traffic when the threshold is zero. However, in both cases, the ratio decreases with network traffic load. This is because as the network load increases, more and more traffic has to be sent as out-of-profile traffic given the fixed and limited HC capacity.

Note that in addition to the results in Fig. 10, several different queue-length threshold values have also been tried in simulations. However, the value, as long as it is not as extreme as zero or infinite, seems to work well when the network load is low, but have little effect on the ratio of transmitted in-profile traffic (to the total generated traffic) when load increases and approaches 1. Specifically, when we set a queue-threshold which is large enough to guarantee all the traffic that has been sent out was transmitted as in-profile traffic, the amount of transmitted in-profile traffic is a little higher than when not using a threshold value, and all the rest traffic (where most of them has been sent out as out-of-profile traffic in the case without any queue-length threshold) is buffered in the queues. The fact is that there is a healthy amount of out-of-profile traffic implies that LOBS-HC is more effective than OCS in that OCS would have not been able to send the out-of-profile traffic at all.

We have also compared LOBS-HC with OBS as shown in Figs. 11 and 12. The simulation results have revealed that

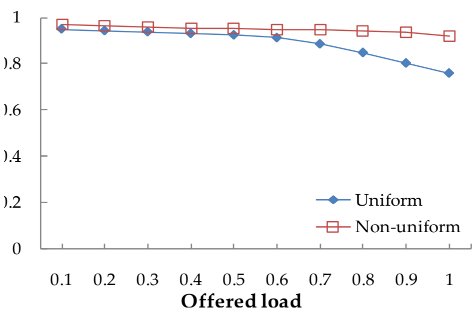

Figure 6. Goodput under different traffic pattern

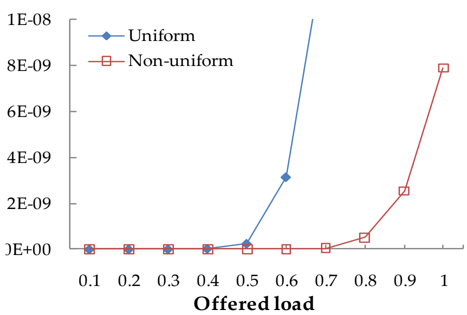

Figure 7. Queuing delay under different traffic pattern

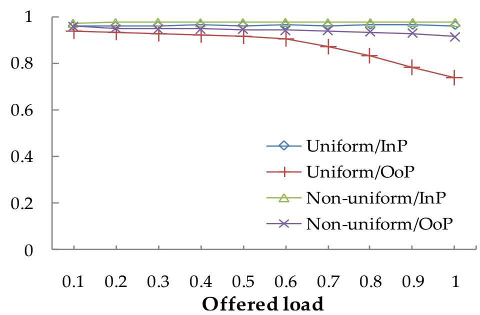

Figure 8. Goodput for in-profile and out-of-profile traffic

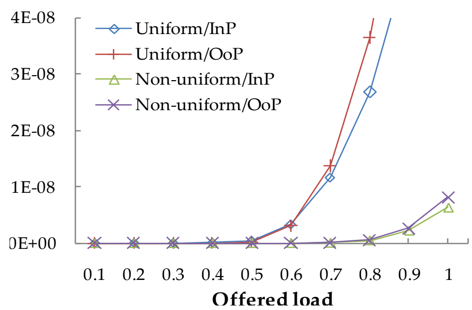

Figure 9. Queuing delay for in-profile and out-of-profile traffic

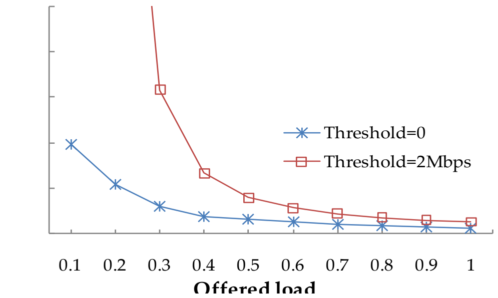

Figure 10. Ratio of in-profile traffic to out-of-profile traffic

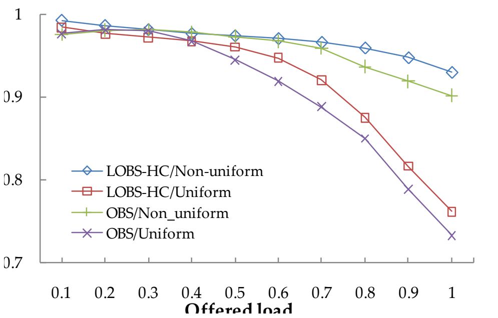

Figure 11. Goodput comparison between LOBS-HC and OBS

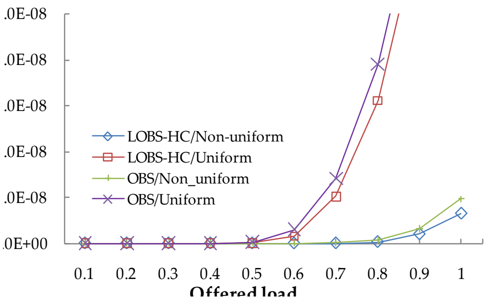

Figure 12. Queuing delay comparison between LOBS-HC and OBS

LOBS-HC performs slightly better than OBS under a high load under both the non-uniform and uniform distributions. The reason for LOBS-HC to be only slightly better than OBS is that the goodput and ETE delay do not distinguish one data frame from another as long as both are delivered, and in OBS, although it cannot deliver as many data frames between requesting pod and far-away storing pods as LOBS-HC, it can make up for some loss of goodput and ETE delay performance by delivering a few more data frames between requesting and storing pods that are close to each other.

IV. CONCLUSION

In this work, we have proposed a new paradigm called Labeled Optical Burst Switching with Home Circuit (LOBS-HC) for optical intra-datacenter networking and considered hypercube-like interconnection scheme among pods for such kind of datacenters. By taking advantages of the unique future of wavelength sharable in LOBS-HC and the minimum communication hop distance required in hypercube, efficient algorithms called CHA/SBT have been developed for HC and wavelengths assignment, which minimize the network resources (e.g., wavelengths) required. Dynamic network performance of such datacenters have also been studied and evaluated via simulation assuming either uniform or non-uniform communication among the pods. In addition, the network performance of in-profile and out-of-profile has been studied. The results have shown that proposed LOBS-HC approach can efficiently support intra-datacenter communications and in particular are more suitable than existing approaches based on either OCS or electronic switches or hybrid switching.

ACKNOWLEDGMENT

This work is partially supported by the R&D program of MKE/KEIT [KI001826 and 10035321], KAIST, Korea; the State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University (SJTU), China; and the Special fund for Basic Scientific Research of Central Colleges (ZZZ10003), South-Central University for Nationalities, China.

REFERENCES

[1] N. Farrington, G. Porter, S. Radhakrishnan, H. Hajabdolali Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," SIGCOMM ACM, New Delhi, India, Aug. 2010.

[2] G. Wang, D. G. Andersen, M. Kaminsky, M. Kozuch, T. S. Eugene Ng, K. Papagiannaki, and M. Ryan, "c-Through: part time optics in data centers," SIGCOMM ACM, New Delhi, India, Aug. 2010.

[3] C. Qiao, M. G. Ortega, A. S. González, X. Liu, and J C. López Ardao, "On the Benefit of Fast Switching in Optical Networks," OFC, USA, March 2010.

[4] M. G. Ortega, C. M. Qiao, A. S. González, X. Liu, and J C. López Ardao, "LOBS-H: an enhanced OBS with wavelength sharable home circuits," ICC, Capetown, South Africa, May 2010.

[5] C. Qiao and M. Yoo, "Optical burst switching (OBS)-a new paradigm for an optical internet," *Journal of High Speed Networks*, vol. 8, no. 1, pp.64-89, 1999.

[6] J. Li and C. Qiao, "Schedule burst proactively for optical burst switched networks," *Computer Networks*, vol. 44, no. 5, pp. 617–629, 2004.

[7] C. T. Ho and S. L. Johnsson, "Spanning balanced trees in boolean cubes, "Journal on Scientific and Statistical Computing archive, vol. 10, no. 4, 1989.

[8] S. G. Ziavras and S. Krishnamurthy, "Evaluating the communications capabilities of the generalized hypercube interconnection networks," Concurrency: Practice & Experience, vol. 11, no.6, pp.281-300,1999.

[9] M. Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," SIGCOMM, USA, Aug. 2008.

[10] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, et al., "Bcube: a high performance, server-centric network architecture for modular data centers," SIGCOMM, Spain, 2009.