

A Reconfigurable High-Performance Optical Data Center Architecture

Chong Liu, Maotong Xu, and Suresh Subramaniam

Department of Electrical and Computer Engineering, The George Washington University

Email: cliu15@gwu.edu, htly8927@gwu.edu, suresh@gwu.edu

Abstract—Optical data center network architectures are becoming attractive because of their low energy consumption, large bandwidth, and low cabling complexity. In [1], an AWGR-based passive optical data center architecture (PODCA) is presented. Compared with other optical data center architectures, e.g., DOS [2], Proteus [3], and Petabit [4], PODCA can save up to 90% on power consumption and 88% in cost. Also, average latency can be low as $9 \mu\text{s}$ at close to 100% throughput. However, PODCA is not reconfigurable and cannot optimize the network topology to dynamic traffic.

In this paper, we present a novel, scalable and flexible reconfigurable architecture called RODCA. RODCA is built on and augments PODCA with a flexible localized intra-cluster optical network. With the reconfigurable intra-cluster network, racks with mutually large traffic can be located within the same cluster, and share the large bandwidth of the intra-cluster network. We present an algorithm for DCN topology reconfiguration, and present simulation results to demonstrate the effectiveness of reconfiguration.

I. INTRODUCTION

The number of data-intensive applications is rapidly increasing in data center networks. These applications, such as MapReduce, Hadoop, and Dropbox, require low latencies and high throughput and bring new challenges for future data center networks (DCNs). Data-intensive computing platforms typically use high speed communications switches and networks, which allow the data to be partitioned among the available computing resources and processed independently to achieve performance and scalability based on the amount of data.

In order to address these problems, recent research has focused on novel interconnect topologies for data center networks. The typical design is to place 20-50 servers in a rack, with an aggregation (“Top of Rack”, or ToR) switch in each rack. Further, only a few ToRs are hot in DCNs and most of the traffic from these ToRs goes to a few other ToRs [5], requiring DCNs to be rapidly reconfigurable. Conventional DCNs, such as fat tree [6], flattened butterfly [7], and VL2 [8], use commodity electrical switches to optimize the limited bandwidth available. However, these DCNs require a large number of links and switches, thus leading to rapidly increasing wiring complexity as the network scales. Moreover, electrical switches are also power-hungry devices. By contrast, optical DCNs provide the advantage of reduced power consumption

and network cost as the network can be constructed using predominantly passive components [1], [9], [10].

A. Related Work

Existing optical data center networks are commonly based on optical switching, e.g., Semiconductor Optical Amplifier (SOA)-based switch, Micro-Electro-Mechanical Systems (MEMS) switches and Arrayed Waveguide Grating Routers (AWGR). MEMS switch, used in c-through [11] and Helios [12], is a power-driven reconfigurable optical switch and its reconfiguration time could be on the order of a few milliseconds [3], and is therefore not well suited for fast packet switching in DCN applications. Nevertheless, optical MEMS switches could be reconfigured at coarse time scales for switching large volumes of data. Dynamically provisioning lightpaths for relatively stable traffic between server racks makes optical networks a cost-effective solution for allocating large bandwidth on-demand across the data center. Also, high-speed MEMS switches with switching times on the order of microseconds [9], [13] are on the horizon, and are expected to be commercialized in the near future. Despite the relatively slow switching times, MEMS switches are scalable. FARBON [14], RODA [15] and Wavecube [16] are three recent reconfigurable optical DCN architectures that use specialized fast switches or expensive wavelength selective switches that do not scale well.

AWGR is a passive optical device that does not require reconfiguration, and can achieve packet contention resolution in the wavelength domain. The cyclic routing characteristic of the AWGR allows different inputs to reach the same output simultaneously by using different wavelengths. Recently, a few AWGR-based DCN architectures have appeared in the literature such as DOS and Petabit. They employ tunable wavelength converters (TWCs), which are power-hungry devices [9]. Moreover, TWCs significantly increase the total cost of the architectures. We presented our own AWGR-based passive optical data center architecture (PODCA) in [1]. There are three versions of the PODCA architecture suited to small (S), medium (M), and large (L) DCNs. Compared with DOS and Petabit, PODCA-L employs a large-scale AWGR as a central switch, and can easily accommodate over 2 million servers. We also presented algorithms for wavelength assignment and

for scheduling packets in PODCA in [1]. We showed in [1] that average packet latencies (excluding protocol overhead) can be as low as $9 \mu s$ at close to 100% throughput. However, PODCA is not reconfigurable, and cannot adapt to fluctuations in traffic that are common in DCNs today.

B. Our Approach and Contributions

In this paper, we concurrently employ AWGR and MEMS-based optical switches to develop a scalable and reconfigurable architecture called RODCA. RODCA augments PODCA-L with a flexible intra-cluster optical network, which is a Clos multi-stage MEMS-based network, to adapt the network to traffic dynamics. The backbone of RODCA is a hierarchical optical DCN topology – several ToRs are interconnected through an AWGR to form a *cluster*, and several clusters can be interconnected through a higher-level AWGR. The design for the intra-cluster network includes a reconfigurable switching network that can be reconfigured at relatively coarse time scales (i.e., reconfiguration times are comparable to or larger than packet transmission times), so that racks with mutually large traffic can be located within the same cluster, and enjoy the large bandwidth of the intra-cluster network.

The following contributions are made in this paper:

- We employ passive AWGRs and MEMS switches to develop a flexible and hierarchical DCN architecture.
- We present an algorithm to trigger cluster reconfiguration based on traffic fluctuations.
- We present extensive performance results from simulations exploring the effects of various algorithm and component parameters.

The rest of this paper is organized as follows: Section II presents a brief description of the PODCA architecture. Section III presents the proposed reconfigurable optical data center architecture, and an algorithm for topology reconfiguration in response to dynamic traffic changes. In Section IV, we present performance evaluation results. Finally, we conclude the paper in Section V.

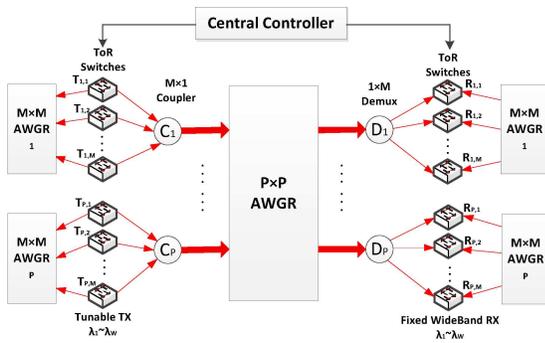


Figure 1. The PODCA-L architecture [1].

II. BACKGROUND

In this section, we briefly review PODCA-L [1], which is the architecture RODCA builds on. As shown in Fig. 1, PODCA-L is a hierarchical, AWGR-based architecture. Suppose the total number of racks is S and a $P \times P$ AWGR is available. We split the S racks into P clusters and each cluster has M racks, where $M = \lceil \frac{S}{P} \rceil$. M racks of the same cluster connect to an input port of the $P \times P$ AWGR through a $M \times 1$ coupler and connect to an output port of the AWGR through a $1 \times M$ demultiplexer. The signal from an output port of a demultiplexer can be either a fixed wavelength or a fixed range of wavelengths. The $P \times P$ AWGR is for inter-cluster communication. Within each cluster, there is an $M \times M$ AWGR for intra-cluster communication. We denote $W = F \cdot P$ as the number of wavelengths, where $F \geq 1$ is an integer. The AWGR routes wavelengths from an input port to a specific output port in a cyclic way; the c^{th} wavelength λ_c is routed from input port i to output port [9]:

$$[(i + c - 2) \bmod P] + 1, 1 \leq i \leq P, 1 \leq c \leq W. \quad (1)$$

Each ToR has one or more fast tunable transmitters and fixed wide-band receivers. PODCA is a time-slotted system, where the time to transmit a packet is one time slot. Packets arriving to a ToR and needing to be transmitted to another ToR are placed in a virtual buffer in the ToR (one for each destination ToR). In each time slot, a central controller schedules packet transmissions for the next time slot and follows three scheduling constraints. First, a tunable transmitter or receiver can only transmit (respectively, receive) one packet at a time. Second, because of the cyclic wavelength routing property of the AWGR, at most F packets can be transmitted from an input port of the AWGR to an output port of the AWGR in a time slot. Third, tunable transmitters connecting to the same AWGR port need to transmit on distinct wavelengths. The central controller first selects packets for transmission based on these three scheduling constraints, and then uses a packet scheduling algorithm to tune wavelengths and schedule packets for transmission.

In PODCA-L, some packets may need two hops to arrive at their destinations – packets may need to be first routed to a ToR different from the destination ToR but in the same cluster as the destination ToR, by inter-cluster transmission, and then utilize intra-cluster transmission to reach the destination ToR. For each selected packet, the packet scheduling algorithm first checks if there is any available wavelength for transmitting directly to the packet’s destination. If there is more than one available wavelength, a round robin method is used to choose one of the available wavelengths. If direct transmission is not possible, then the algorithm checks if a two-hop transmission is possible on any available wavelength. If this is not possible either, the packet waits in the source buffer until the next slot when the above steps are repeated. The interested reader can refer to [1] for full details of the architecture and the scheduling algorithm.

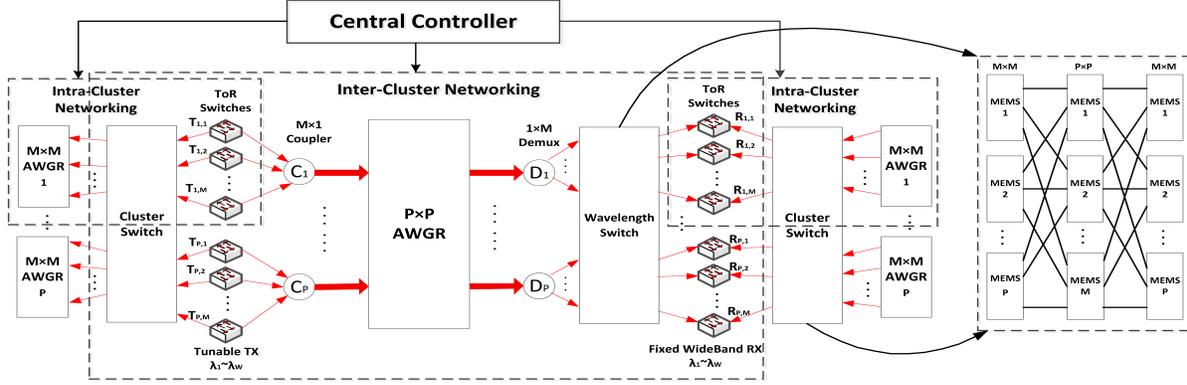


Figure 2. The reconfigurable optical data center architecture (RODCA).

III. RODCA ARCHITECTURE AND RECONFIGURATION

In this section, we present the proposed reconfigurable optical data center architecture (RODCA), and an algorithm for reconfiguration. Fig. 2 shows the proposed architecture.

A. RODCA Design

The architecture, shown in Fig. 2, augments PODCA with two switches (switching networks, in fact) – a *wavelength switch* (this is not to be confused with a wavelength-selective switch) and a pair of *cluster switches*. The cluster switches are used to reconfigure cluster memberships dynamically, i.e., to partition the set of rack transmitters and receivers into clusters based on traffic demands so that rack pairs with large traffic demands are placed in the same cluster. The use of the wavelength switch is explained below.

Currently, the size of the fast (microsecond-level) MEMS optical switch mentioned in [9] is limited to a few tens of ports. In order to scale the network to large sizes (hundreds or thousands of racks), we use a Clos multi-stage network of MEMS optical switches to build large-sized cluster and wavelength switches. The path diversity and non-blocking nature of the Clos network enables the routing of arbitrary traffic patterns with no loss of throughput [7], and plays an important role in the scalability of RODCA.

We now explain why the wavelength switch is needed. During the topology reconfiguration process, RODCA reconfigures the wavelength switch to make each cluster to be able to receive all W wavelengths from an output port of the inter-cluster AWGR. Without the wavelength switch, some racks might not be able to receive certain wavelengths. To illustrate this point, in Fig. 3, consider a network of four racks, $R_{1,1}$, $R_{1,2}$, $R_{2,1}$, and $R_{2,2}$, and suppose two wavelengths are available. Without the wavelength switch, $R_{1,1}$ and $R_{2,1}$ can only receive λ_1 , and $R_{1,2}$ and $R_{2,2}$ can only receive λ_2 (as per the design of PODCA [1], which uses fixed-band receivers). After reconfiguration, suppose $R_{1,1}$ and $R_{2,2}$ are placed into one cluster. Then, if the wavelength switch is not used, this

cluster cannot receive any packet from racks that are connected to the second/bottom input port of the inter-cluster AWGR, since the second input port of the inter-cluster AWGR can reach the first/top output port of the inter-cluster AWGR only by using λ_2 , but this wavelength cannot be received by $R_{1,1}$. Similarly, the second input port of the inter-cluster AWGR can reach the second output port of the inter-cluster AWGR by using λ_1 , but the wavelength cannot be received by $R_{2,2}$. However, by using the wavelength switch, we can connect $R_{1,1}$ and $R_{2,2}$ to the second output port of the inter-cluster AWGR. Now, $R_{1,1}$ and $R_{2,2}$ can receive λ_1 and λ_2 , respectively. The second input port of the inter-cluster AWGR can reach this cluster by sending packets to $R_{1,1}$ by using λ_1 ; it can also reach it by sending packets to $R_{2,2}$ on λ_2 .

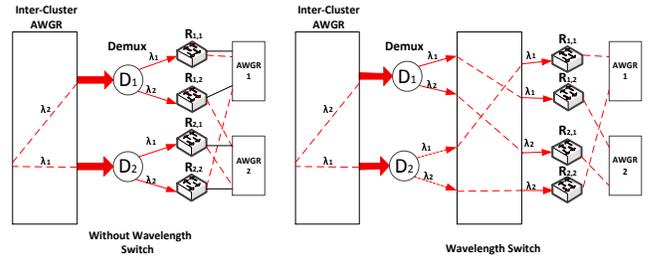


Figure 3. The need for the wavelength switch.

B. Reconfiguration Algorithm

We use the number of packets in the virtual buffers to decide *when* the RODCA topology should be reconfigured. In particular, we consider the total number of packets waiting for inter-cluster transmission (L_{inter}), the total number of packets waiting for intra-cluster transmission (L_{intra}), and use a threshold parameter (β) to make the reconfiguration decision. For this purpose, every rack is required to send the number of packets waiting in each of its virtual buffers to the central controller. We then propose to perform reconfiguration if:

$$\beta \cdot L_{intra} \leq L_{inter}, \quad (2)$$

Thus, reconfiguration is triggered if (a multiple of) the number of intra-cluster packets falls below the number of inter-cluster packets. Since sampling the buffers and reporting the buffer occupancy to the controller causes control message overhead, we introduce a sampling interval (SI) to balance the overhead with performance. $SI = \delta$ means that the buffers are sampled every δ time slots, and the reconfiguration condition above is checked.

If the condition is satisfied, the controller suspends all inter- and intra-cluster packet transmissions in order to reconfigure the network based on the number of packets waiting for transmission in the virtual buffers. We denote a set containing all $P \cdot M$ racks as Θ and call the number of packets between two racks as the *mutual number of packets*. A simple greedy heuristic is used to form clusters. We first find two racks with the largest mutual number of packets in Θ , and remove these two racks from Θ . Then, we find another rack in Θ with the largest mutual number of packets with the previous two chosen racks and remove this rack from Θ . We repetitively find a rack from Θ owning the largest mutual number of packets with all chosen racks in the previous two steps and remove the rack from Θ , and so on, until we have collected M racks to generate a cluster. We repeat the above steps until we have P clusters. The pseudocode of this reconfiguration algorithm, which is executed every time slot, is shown in Algorithm 1. In Algorithm 1, we use a counter to indicate if it is time to reconfigure.

Algorithm 1: RECONFIG

```

1: if counter == SI and  $\beta \cdot L_{intra} \leq L_{inter}$  then
2:   counter = 0
3:   for  $p = 1 : P$  do
4:     find a pair of racks with the largest mutual number of
       packets
5:     remove this pair of racks from  $\Theta$ 
6:     for  $m = 1 : M$  do
7:       find a rack with the largest mutual number of packets
         for all racks removed in iteration  $p$ 
8:       remove the rack from  $\Theta$ 
9:     end for
10:  end for
11: else
12:   counter++
13: end if

```

Once reconfiguration is complete, we resume packet transmission by following the *packet scheduling for PODCA-L* algorithm [1]. The central controller schedules both intra-cluster and inter-cluster transmission based on *scheduling constraints*. Recall that some packets can be transmitted in one hop, and some need two hops to arrive at their destinations. For example, suppose P equals 2 and W equals 8. There are 8 racks within each cluster, i.e., $M = 8$. On each rack, there is one tunable transmitter and one wide-band receiver for intra-cluster

communication. Also, there is one tunable transmitter and one wide-band receiver for inter-cluster communication. Suppose a packet is from $T_{1,1}^1$ to $R_{2,1}^1$. The only wavelength $R_{2,1}^1$ can receive, within inter-cluster transmission, is λ_1 . However, based on the routing characteristics of the AWGR, the receivable wavelengths from the first input port of the AWGR to the second output port of the AWGR can only be $\lambda_2, \lambda_4, \lambda_6$ and λ_8 . Therefore, λ_1 transmitted from $T_{1,1}^1$ cannot arrive at the second output port of the AWGR in a single hop. Thus, a two-hop transmission is needed. We first choose one wavelength from $\{\lambda_2, \lambda_4, \lambda_6, \lambda_8\}$; suppose we choose λ_2 . The packet is transmitted to $R_{2,2}$ by using λ_2 , and then in the next time slot, $R_{2,2}$ can transmit the packet to $R_{2,1}$ by using an intra-cluster transmission.

Algorithm 2: RODCA_PACKET_SCHEDULING

```

1: for packets at the head of virtual buffers do
2:   RECONFIG ()
3:   Load-balance between inter- and intra-cluster network by
     using PODCA-L_PACKET_SCHEDULING () [1]
4: end for

```

If the source and destination racks are in the same cluster, we can use either intra-cluster transmission or inter-cluster transmission. Here, we define a threshold to determine if the packet uses intra-cluster transmission or inter-cluster transmission. If the number of packets waiting for intra-cluster transmission is less than the threshold, then we place the packet at the tail of the waiting queue of the intra-cluster transmission. Otherwise, we use inter-cluster transmission to transmit that packet. To transmit more than one packet in a time slot, each ToR can have more than one tunable transmitter and wide-band receiver for intra-cluster transmission, or inter-cluster transmission, or both.

The operation of RODCA follows the algorithm whose pseudocode is shown in Algorithm 2.

IV. PERFORMANCE EVALUATION

In this section, we conduct simulations to evaluate the latency and throughput performance of RODCA. For space reasons, we fix the network configuration and conduct experiments by varying the algorithm and component parameters only. Both the inter-cluster and intra-cluster AWGR sizes are set to be 30×30 ; thus, the DCN is made up of 900 racks, and we assume 48 servers per rack for a total of 43,200 servers in the DCN. The number of wavelengths is set to 120. On each rack, there are 3 tunable transmitters for intra-cluster transmission and 1 tunable transmitter for inter-cluster transmission. Each rack has $P \cdot M = 900$ virtual queues and each virtual queue buffers packets for each destination rack. The buffer size of a ToR is 5 Mb. The transmission rate of each tunable transmitter is assumed to be 10 Gbps, and its tuning time is 8 ns [1]. Since the network only connects racks with each other, we assume

traffic arrives to racks, and do not model server traffic in our simulations. Following [14], we assume that tasks arrive to the data center according to a Poisson process with a mean rate of 3 tasks/sec. Each task arrives to a random source rack, and triggers κ flows, where $\kappa \in [1, 900]$ is an integer chosen based on a uniform distribution. We model two types of traffic flows, *mice flows* and *elephant flows*. Following [17], we use two Gaussian distributions to model mice and elephant flow rates. (If the Gaussian distribution gives a negative value, we set the value to be 0.) Mice flow arrival rates for each server follow a Gaussian distribution with mean 0.01 Mbps, and elephant flow arrival rates for each server follow a Gaussian distribution with mean 40 Mbps. Thus, mice arrival rates for each ToR follow a Gaussian distribution with mean 0.48 Mbps (48×0.01 Mbps), and elephant flow arrival rates for each ToR follow a Gaussian distribution with mean 1.92 Gbps (48×40 Mbps). We dynamically change the number of elephant flows and mice flows to make the ratio between total elephant arrival rate and total mice arrival rate to be 9:1.

Recall that β is a threshold parameter that is used by our algorithm to trigger network reconfiguration. Each packet is assumed to be 1500 bytes long, which implies that each slot is $1.2 \mu\text{s}$ long (assuming 10 Gbps wavelength capacity and transmitter rates). The time to reconfigure the topology, RT , is set to 1 slot (fast switches) or 10 slots (relatively slow switches). In all of our simulations, we observed no packet drops and thus the throughput is 100%, except for the case when $RT = 10$ and $\beta = 0$. In this case, we observed a 5% packet drop rate, suggesting that too frequent reconfiguration can be detrimental to performance when RT is not small.¹ We therefore focus on packet latency, and make references to throughput only when it is less than 100%.

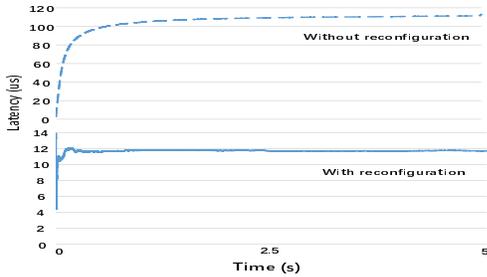


Figure 4. The effect of reconfiguration on packet latency.

We first compare the throughput and latency performance with and without reconfiguration. We set $SI = 1$, $\beta = 15$, and $RT = 10$, and simulate the DCN for 5 seconds (corresponding to approximately 600 million packet arrivals). Our results show that no packets are dropped with reconfiguration, whereas approximately 4% of the packets are dropped when the network does not reconfigure in response to traffic changes. Fig. 4

¹We will see later that the latency is also negatively impacted in this scenario.

shows the evolution of packet latency with time as the network starts from an empty state. The latency quickly reaches steady state and settles around $12 \mu\text{s}$ with reconfiguration, whereas it increases rapidly to around $110 \mu\text{s}$ when no reconfiguration is done. This 90% improvement in latency clearly shows the benefits of adapting the network topology to dynamic traffic variations.

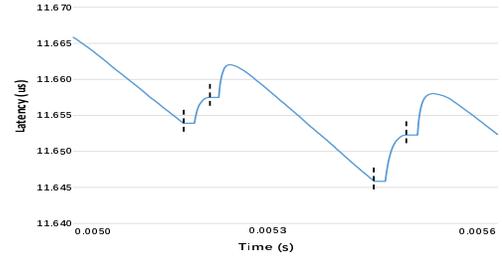


Figure 5. Latency dynamics upon reconfiguration.

In Fig. 5, we zoom into a 0.6 ms window of Fig. 4 to better illustrate the behavior of latency when reconfiguration happens. Reconfiguration initiation times are marked with a dashed vertical line in the figure. When the reconfiguration condition is satisfied, RODCA initiates reconfiguration, and suspends all packet transmissions. During this time, no packets are received by any destination rack and thus the packet latency remains constant. RODCA resumes packet transmission after reconfiguration is complete, and packet latency starts to increase initially and then goes down. The initial increase in latency is due to the fact that packets that are backlogged during reconfiguration contribute to an increase in latency, but soon after, the latency starts decreasing because of the optimized topology. Fig. 5 also shows that two reconfigurations could occur back to back, and the effect of the first reconfiguration on packet latency is never noticed. This is a consequence of choosing a small value of SI . In our experiments, we have not seen this occur when SI and β are large.

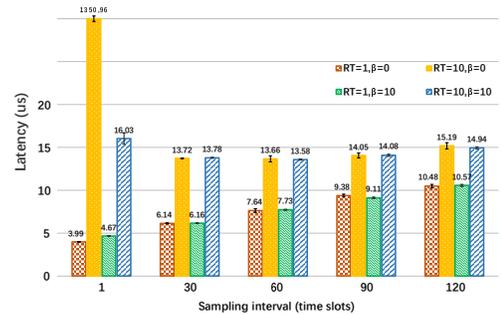


Figure 6. Latency vs. sampling interval for various values of β and RT .

We next show the latency for different SI , β , and RT values in Fig. 6. For this and subsequent figures, we simulate 100 million packet arrivals, and show 95% confidence intervals from 30 trials for each data point. Fig. 6 shows that when $RT = 1$,

small SI and β give the best latency performance. Intuitively, when RT is small, there is minimal penalty for reconfiguring the network (which is facilitated by small values of SI and β) and optimizing the topology helps boost performance significantly. Further, when RT = 10, as the SI increases, the latency first decreases and then increases. As the SI increases, the frequency of reconfiguration decreases, and as Fig. 5 illustrates, some reconfigurations are not useful at all when SI is small. Thus, when SI increases from 1 to 60, the penalty from suspension of packet transmissions decreases and the number of useless reconfigurations decreases. However, when SI increases from 60 to 120, the benefit from reconfigurations also decreases and the latency increases. Eventually, as SI approaches infinity, reconfiguration is rarely triggered, and the latency would approach that for the no reconfiguration case. Fig. 6 also illustrates how β affects the packet latency. A large β implies a reluctance to reconfigure the network, and this is particularly useful when SI is small, e.g., 1. It can be seen that a large β tremendously decreases the packet latency in this case.

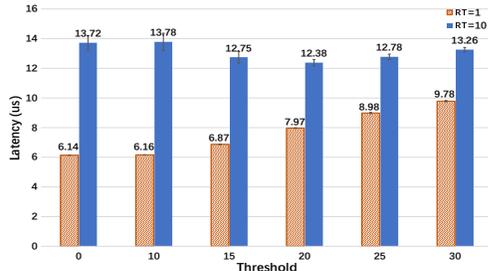


Figure 7. Latency vs. threshold for various values of RT

Fig. 7 presents packet latency for different values of the threshold β , for SI = 30. When RT = 1, as β increases, the latency always increases. When RT = 10, as β increases, the latency first decreases and then increases, as explained above. Fig. 7 shows that once an SI is chosen, we can tune the β to get the smallest latency. Even though both SI and β affect the latency, they are introduced for different purposes. Recall that SI is a measure of how frequently the virtual buffers are sampled.² This frequency places a load on the controller, with smaller SI values requiring more frequent control messages for gathering the occupancy information on the virtual buffers. The actual overhead due to such messages is not modeled in this paper, and is left for future work. In contrast to SI, β is an algorithm parameter for controlling the latency for a given SI.

V. CONCLUSIONS

In this paper, we present a scalable and flexible reconfigurable architecture called RODCA. RODCA is built on and augments PODCA-L with a flexible intra-cluster optical network. With the reconfigurable intra-cluster network, racks with

²We have also looked into averaging the sampled values over a time window, and found it to be not very useful.

mutually large traffic can be located within the same cluster, and share the large bandwidth of the intra-cluster network. We present an algorithm for DCN topology reconfiguration, and present simulation results to demonstrate the effectiveness of reconfiguration. Our results show that packet latencies around 10–12 μ s are achievable with judicious topology reconfiguration in response to dynamic traffic changes, while latencies could be an order of magnitude larger if the topology is not reconfigured.

VI. ACKNOWLEDGMENTS

This work was supported in part by NSF award # 1618487.

REFERENCES

- [1] M. Xu, C. Liu, and S. Subramaniam, "PODCA: A passive optical data center architecture," in *2016 IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.
- [2] X. Ye, Y. Yin, S. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: A scalable optical switch for datacenters," in *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, 2010, p. 24.
- [3] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, 2010.
- [4] K. Xia, Y.-H. Kaob, M. Yangb, and H. Chao, "Petabit optical switch for data center networks," *Polytechnic Institute of New York University, New York, Tech. Rep.*, 2010.
- [5] S. Kandula, J. Padhye, and P. Bahl, "Flyways to de-congest data center networks," in *Proceedings of HotNets*, 2009.
- [6] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. ACM SIGCOMM*, 2008, pp. 63–74.
- [7] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: a cost-efficient topology for high-radix networks," *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2, pp. 126–137, 2007.
- [8] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: a scalable and flexible data center network," in *Proc. ACM SIGCOMM*, 2009, pp. 51–62.
- [9] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," *IEEE Comm. Surveys & Tutorials*, vol. 14, no. 4, pp. 1021–1036, 2012.
- [10] J. Chen, Y. Gong, M. Fiorani, and S. Aleksic, "Optical interconnects at the top of the rack for energy-efficient data centers," *IEEE Communications Magazine*, vol. 53, no. 8, pp. 140–148, 2015.
- [11] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. Ng, M. Kozuch, and M. Ryan, "c-through: Part-time optics in data centers," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 327–338, 2011.
- [12] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," in *Proc. ACM SIGCOMM*, 2011, pp. 339–350.
- [13] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat, "Integrating microsecond circuit switching into the data center," in *Proc. ACM SIGCOMM*, 2013, pp. 447–458.
- [14] Z. Guo and Y. Yang, "Augmenting data center networks with a fast reconfigurable optical multistage interconnect," in *Proceedings of IEEE Globecom*, 2014, pp. 2198–2204.
- [15] A. Pal and K. Kant, "RODA: A reconfigurable optical data center network architecture," in *Proc. IEEE LCN*, 2015, pp. 561–569.
- [16] K. Chen, X. Wen, X. Ma, Y. Chen, Y. Xia, C. Hu, and Q. Dong, "Wavecube: A scalable, fault-tolerant, high-performance optical data center architecture," in *Proc. IEEE INFOCOM*, 2015, pp. 1903–1911.
- [17] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: Measurements & analysis," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*. ACM, pp. 202–208.