PODCA: A Passive Optical Data Center Architecture

Maotong Xu, Chong Liu, and Suresh Subramaniam
Department of Electrical and Computer Engineering
The George Washington University
Email: htfy8927@gwu.edu, cliu15@gwu.edu and suresh@gwu.edu

***Abstract*—Optical interconnects for data centers can offer reduced power consumption, low latency, and high scalability, compared to electrical interconnects. However, active optical components such as tunable wavelength converters and Micro-Electro-Mechanical Systems (MEMS) switches suffer from high cost or slow reconfiguration times.**

In this paper, we propose three different Passive Optical Data Center Architectures (PODCAs), depending on the size of the network. Our key device is the Arrayed Waveguide Grating Router (AWGR), a passive device that can achieve contention resolution in the wavelength domain [1]. In our architectures, optical signals are transmitted from fast tunable transmitters and pass through couplers, AWGR, demultiplexers, and are received by wide-band receivers. Our architecture can easily accommodate over 2 million servers. Simulation results show that packet latency is below 9μs, and 100% throughput is achievable. We compare the power consumption and capital expenditure (CapEx) cost of PODCA with other recent optical data center network architectures such as DOS, Proteus, and Petabit. Results show that our architectures can save up to 90% on power consumption and 88% on CapEx.

I. INTRODUCTION

Power consumption, latency, and scalability are critical factors in designing data center network (DCN) architectures. Power consumption of data centers will reach 140 billion kilowatt-hours annually by 2020, and it will cost $13 billion annually [2]. On the other hand, interactive applications, e.g., web search, social network, and stock exchange, require low network latency; e.g., the acceptable latency range of stock exchange is 5-100 ms [3]. Furthermore, as data volume increases, the sizes of data centers scale up tremendously. Microsoft owns over 1 millions servers and its Chicago data center alone is estimated to contain over 250,000 servers [4].

Considering reduced power consumption, low latency, and high scalability, optical networking is a promising solution for current data centers [5]. Optical networks are commonly based on optical circuit switching, e.g., Micro-Electro-Mechanical Systems Switches (MEMS) and Arrayed Waveguide Grating Router (AWGR). MEMS is a power-driven reconfigurable optical switch and its reconfiguration time is in the order of a few milliseconds [6], and is not best suited for fast packet switching in DCN applications. Compared to MEMS, AWGR is a passive optical device that does not require reconfiguration, and can achieve 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. Previous AWGR-based architectures, e.g., DOS [1] and Petabit [7], employ Tunable Wavelength Converters (TWC), which are power-hungry devices [5]. Moreover, TWC highly increases the total capital expenditure (CapEx) of architectures.

Our solution is to employ passive optical devices, e.g., AWGR, coupler and demultiplexer, to reduce power consumption, while achieving low packet latency and high throughput. Moreover, our architecture is highly scalable and it can easily accommodate more than 2 million servers. The main contributions of our work are as follows:

- We propose three different-sized architectures – small-sized, medium-sized and large-sized passive optical architectures, and present the wavelength assignment and packet transmission algorithm for each.
- We obtain the packet latency and network throughput of the three architectures through simulations. Simulation results show that packet latencies of under 9μs and a throughput of 100% are achievable. We also present evaluation results for a number of different architectural parameters such as the the number of tunable transmitters per rack,[1] the type of architecture, as well as packet size.
- We compare the power consumption and CapEx of our architectures with DOS, Proteus [6], and Petabit. Results show that our architectures can save up to 90% on power consumption, and 88% on CapEx.

The rest of this paper is organized as follows: Section II provides three optical data center architectures and corresponding wavelength assignment and packet transmission algorithms. Section III describes the performance evaluation of the architectures and the simulation results. In Section IV, we compare the proposed architectures with three other all-optical architectures in terms of power consumption and CapEx. Finally, we conclude the paper in Section V.

II. PROPOSED ARCHITECTURES AND ALGORITHMS

Our architectures are hierarchical, and use an AWGR to interconnect racks. Racks that are connected to each port of the AWGR are said to form a *cluster*, and the entire DCN is a group of clusters. Suppose each cluster has M racks, i.e., $\lceil \frac{S}{P} \rceil = M$, where S is the total number of racks and P is the

[1]The term "rack" and "ToR" (for top-of-rack) are used interchangeably in the paper.

978-1-4799-6664-6/16/$31.00 ©2016 IEEE

number of ports of AWGR. We denote W as the number of available wavelengths and W equals $F{\cdot}P$, where $F{\geq}1$ is an integer. We denote the wavelength c as λ_c. The AWGR routes wavelengths from an input port to a specific output port in a cyclic way; λ_c is routed from input port i to output port [5]:

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

Each rack has a top of rack switch (ToR), and one or more fast tunable transmitters (tunable to any wavelength) and fixed wide-band receivers. A wide-band receiver is a simple photodetector receiver that can receive a signal on *any* wavelength as long as only one signal is directed to it. Let N be the number of tunable transmitters or wide-band receivers on each ToR. Let $\mathcal{T}^t_{i,j}$ and $\mathcal{R}^r_{i,j}$ denote the tth tunable transmitter and rth fixed-band receiver on the jth ToR connecting to the ith port of AWGR, respectively. Note that $\mathcal{T}_{i,j}$ and $\mathcal{R}_{i,j}$ are the corresponding transmission side rack and reception side rack, respectively.

We consider architectures for three cases: (a) Small DCN: $S{\leq}P$; P is typically around 50, but could be as large as 512; (b) Medium DCN: $P{\leq}S{\leq}W$, and (c) Large DCN: $W{<}S$. The architectures for the three cases are called PODCA-S, PODCA-M, and PODCA-L, respectively.

Packets arriving at a ToR and destined to another ToR are placed in a shared buffer. In this paper, we assume a central controller that schedules packet transmissions; while this might introduce scalability issues for large DCNs, it could be replaced by distributed scheduling with some loss in performance (reserved for future work). The system is time slotted, and a time slot includes includes both the packet transmission time (all packets are assumed to be the same size) and transmitter tuning time. In each time slot, the controller schedules transmissions for the next time slot. The packet scheduling algorithms for all architectures respect the following constraints, called *scheduling constraints*. First, a tunable transmitter can transmit at most one packet at a time. Second, a wide-band receiver can receive at most one packet at a time.[2] Third, due to 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 slot. Also, tunable transmitters connecting to the same AWGR port need to tune to distinct wavelengths. The first three constraints are satisfied by Algorithm 1 and the last one is satisfied by the wavelength assignment algorithm in each subsection. In Algorithm 1, c_{i_t,j_t}, c_{i_r,j_r}, and c_{i_t,i_r} represent the number of packets selected in $\mathcal{T}_{i_t,j_t}$, for $\mathcal{R}_{i_r,j_r}$, and from i_tth AWGR input port to i_rth AWGR output port, respectively.

In the following subsections, we present the architectures and the packet transmission algorithms.

A. PODCA-S

In this case, $S{\leq}P$, and each cluster is just a single rack. We connect each rack to a port of the AWGR. Let W be an integral

[2]Thus, any rack can transmit at most N packets and receive at most N packets in a time slot.

Algorithm 1: PACKET SELECTION

pkt is from $\mathcal{T}_{i_t,j_t}$ to $\mathcal{R}_{i_r,j_r}$
if $c_{i_t,j_t} < N$ and $c_{i_t,i_r} < F$ and $c_{i_r,j_r} < N$ **then**
  schedule tx from $\mathcal{T}^t_{i_t,j_t}$ to $\mathcal{R}^r_{i_r,j_r}$
  c_{i_t,j_t}++; c_{i_t,i_r}++; c_{i_r,j_r}++; t=c_{i_t,j_t}; r=c_{i_r,j_r};
else
  schedule next packet
end if

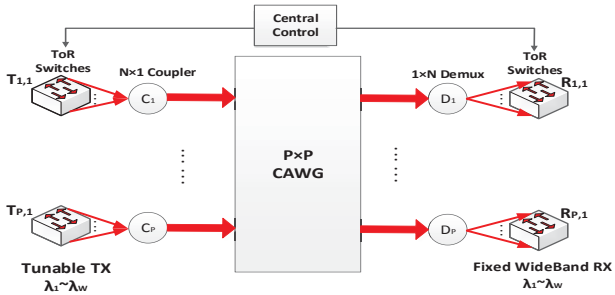


Figure 1. Architecture of PODCA-S.

multiple of N, i.e., $W{=}\alpha{\cdot}N$, where $\alpha{\geq}1$ is an integer. The architecture is shown in Fig 1. Note that the signal from an output port of a demultiplexer can be either a fixed wavelength or a fixed range of wavelengths. Since each output port of a demultiplexer connects to a wide-band receiver and the wide-band receiver can only receive one packet at a time, the central controller needs to guarantee that only one wavelength from the output port of a demultiplexer carries data at any given time.

Suppose a packet waiting for transmission is from $\mathcal{T}^t_{i_t,1}$ to $\mathcal{R}^r_{i_r,1}$. By deriving the modular inverse of (1), we know that wavelength assignment is determined by the input port number and output port number of the AWGR. If $i_t{\leq}i_r$, the tunable transmitter needs to tune to one of the wavelengths shown in (2) to successfully deliver the packet:

$$i_r - i_t + 1 + f{\cdot}P, \quad \forall f{\in}[0, F-1]. \quad (2)$$

If $i_t{>}i_r$, the tunable transmitter needs to tune to one of the wavelengths shown in (3) to successfully deliver the packet.

$$P + i_r - i_t + 1 + f{\cdot}P, \quad \forall f{\in}[0, F-1]. \quad (3)$$

Here, $f \leq F-1$ because the maximum number of available wavelengths is $W{=}F{\cdot}P$. The coupler combines different wavelengths from different tunable transmitters and outputs the combined signal to the AWGR. The output signal of the AWGR is evenly demultiplexed N ways, i.e., the first α wavelengths are demultiplexed to output port 1, the next α wavelengths to port 2, and so on. The pseudocode of the wavelength assignment algorithm is shown in Algorithm 2.

B. PODCA-M

The PODCA-M architecture is shown in Fig 2. In this case, $P{\leq}S{\leq}W$. We connect $M(>1)$ racks to each AWGR port. An $(N{\cdot}M){\times}1$ coupler connects all the N transmitters on each

Algorithm 2: PACKET SCHEDULING FOR PODCA-S

for packets at the head of output queues **do**
  PACKET SELECTION ()
  the packet is from $\mathcal{T}^t_{i_t,1}$ to $\mathcal{R}^r_{i_r,1}$
  if $i_t \leq i_r$ and $f^{i_r}_{i_t}$++ $<$ F-1 **then**
    tune to wavelength: i_r-i_t+1+$f^{i_r}_{i_t} \cdot P$
  else if $i_t > i_r$ and $f^{i_r}_{i_t}$++ $<$ F-1 **then**
    tune to wavelength: P+i_r-i_t+1+$f^{i_r}_{i_t} \cdot P$
  end if
end for

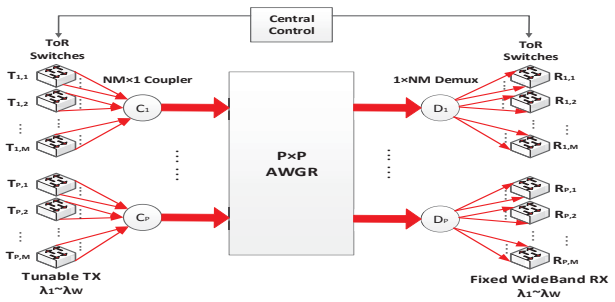


Figure 2. Architecture of PODCA-M.

of the M racks to an input port of the AWGR. $1\times(N\cdot M)$ demultiplexer connects an output port of the AWGR to M racks. Let W be an integer multiple of the number of tunable transmitters connecting to one demultiplexer, i.e., $W=\beta\cdot NM$, where $\beta\geq 1$ is an integer. The receivable wavelength range for the jth rack in a cluster is:

$$\left[(j-1)\cdot\frac{W}{M}+1, j\cdot\frac{W}{M}\right], \tag{4}$$

and we denote the above wavelength set as λ^s_j. Note that $S = M \cdot P \leq W = F \cdot P$, and so, $M \leq F$. From (4), we can see that the number of receivable wavelengths for any ToR is $\frac{W}{M}$. Also, $\frac{W}{M} \geq \frac{W}{F} = P$. Thus, the number of receivable wavelengths of a ToR is at least P. Also, based on (4), the P wavelengths are contiguous, which means that a ToR can receive packets from all the ports of the AWGR. Suppose the packet waiting for transmission is from $\mathcal{T}^t_{i_t,j_t}$ to $\mathcal{R}^r_{i_r,j_r}$. We tune the transmitters based on (2) and (3). Additionally, we need to guarantee that the tuned wavelength belongs to $\lambda^s_{j_r}$. The coupler combines distinct wavelengths from transmitters and outputs the combined signal to the AWGR. Note that if there are multiple available wavelengths, we use round robin method to choose one of them. The pseudocode of the wavelength assignment algorithm is shown in Algorithm 3.

C. PODCA-L

In this case, $W<S$. We use two different sized AWGRs, i.e., an $M\times M$ AWGR for intra-cluster transmission, and a $P\times P$ AWGR for inter-cluster transmission. In each intra-cluster network, we connect M tunable transmitters and M wide-band receivers to a $M\times M$ AWGR. For inter-cluster network, we connect M tunable transmitter, belonging to the same cluster,

Algorithm 3: PACKET SCHEDULING FOR PODCA-M

for packets at the head of output queues **do**
  PACKET SELECTION ()
  the packet is from $\mathcal{T}^t_{i_t,j_t}$ to $\mathcal{R}^r_{i_r,j_r}$
  if $i_t \leq i_r$ **then**
    tune to wavelength$\in$
    $\lambda^s_{j_r} \cap \{i_r$-$i_t$+1+$f\cdot P$, $\forall f \leq$F-1$\}$
  else if $i_t > i_r$ **then**
    tune to wavelength$\in$
    $\lambda^s_{j_r} \cap \{P$+i_r-i_t+1+$f\cdot P$, $\forall f \leq$F-1$\}$
  end if
end for

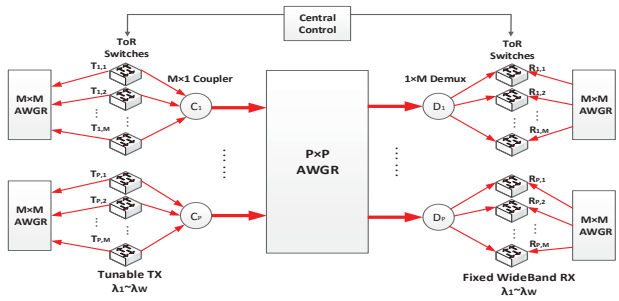


Figure 3. Architecture of PODCA-L

to a $M\times 1$ coupler and connect the output of the coupler to the $P\times P$ AWGR. Each output port of the $P\times P$ AWGR is connected to an input port of the $1\times M$ demultiplexer, and each output port of the demultiplexer connects to the M racks within a cluster. The architecture is shown in Fig 3.

Of course, coupler power losses may necessitate amplification if the cluster size becomes large. Nevertheless, PODCA-L is highly scalable and it can easily accommodate up to more than 2 million servers (assuming 48 servers per rack, 100-port AWGRs within the cluster, and a 512-port AWGR interconnecting clusters). Without utilizing reconfigurable devices in the architecture, we can achieve huge power savings. However, some packets 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 $\mathcal{T}^1_{1,1}$ to $\mathcal{R}^1_{2,1}$. The only wavelength $\mathcal{R}^1_{2,1}$ can receive, within inter-cluster transmission, is λ_1. However, based on the routing characteristics of AWGR, the receivable wavelengths from the first input port of the AWGR to the second output port of the AWGR can only be λ_2, λ_4, λ_6 and λ_8. So, λ_1 transmitted from $\mathcal{T}^1_{1,1}$ 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 first transmitted to $\mathcal{R}_{2,2}$ by using λ_2, and then in the next time slot, $\mathcal{R}_{2,2}$ can transmit the packet to $\mathcal{R}_{2,1}$ by using an intra-cluster transmission.

Next, we give the general description for the communication of packets. The central controller schedules both intra-cluster and inter-cluster transmission based on *scheduling constraints*. Suppose a packet waiting for transmission is from $\mathcal{T}^t_{i_t,j_t}$ to $\mathcal{R}^r_{i_r,j_r}$. If i_t=i_r, 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. Suppose the packet is transmitted by using intra-cluster transmission. If $j_r \geq j_t$, the tunable transmitter tunes to wavelength:

$$j_r - j_t + 1. \tag{5}$$

If $j_r < j_t$, the tunable transmitter tunes to wavelength:

$$M + j_r - j_t + 1. \tag{6}$$

Note, the number of wavelengths needed for intra-cluster transmission is M. The main reason is that there is only one tunable transmitter and wide-band receiver connecting to a port of the intra-cluster AWGR.

If $i_t \leq i_r$, the wavelength set ($\lambda^s_{i_t \leq i_r}$), which contains all available wavelengths from input port i_t of AWGR to output port i_r of AWGR, is:

$$i_r - i_t + 1 + f \cdot P, \quad \forall f \in [0, F-1]. \tag{7}$$

If $i_t > i_r$, the wavelength set ($\lambda^s_{i_t > i_r}$), which contains all available wavelengths from input port i_t of AWGR to output port i_r of AWGR, is:

$$P + i_r - i_t + 1 + f \cdot P, \quad \forall f \in [0, F-1]. \tag{8}$$

If $i_t \leq i_r$ and $\lambda^s_{j_r} \cap \lambda^s_{i_t \leq i_r} \neq \emptyset$, the packet only needs inter-cluster transmission. $\mathcal{T}^t_{i_t,j_t}$ can tune to a wavelength in $\lambda^s_{j_r} \cap \lambda^s_{i_t \leq i_r}$, which is the wavelength that can be received by $\mathcal{R}^r_{i_r,j_r}$ using inter-cluster transmission.

If $i_t \leq i_r$ and $\lambda^s_{j_r} \cap \lambda^s_{i_t \leq i_r} = \emptyset$, $\mathcal{T}^t_{i_t,j_t}$ tunes to one of the wavelengths in $\lambda^s_{i_t \leq i_r}$ and the packet needs both inter-cluster transmission and intra-cluster transmission to arrive at its destination. Suppose $\mathcal{T}^t_{i_t,j_t}$ tunes to λ_{c_1}, where $c_1 \in \lambda^s_{i_t \leq i_r}$. The packet arrives at $\mathcal{R}_{i_r, \lceil \frac{Mc_1}{W} \rceil}$ by using inter-cluster transmission. In another time slot, $\mathcal{R}_{i_r, \lceil \frac{Mc_1}{W} \rceil}$ will transmit the packet to $\mathcal{R}_{i_r,j_r}$ by using intra-cluster transmission.

If $i_t > i_r$ and $\lambda^s_{j_r} \cap \lambda^s_{i_t > i_r} \neq \emptyset$, the packet only needs inter-cluster transmission. $\mathcal{T}^t_{i_t,j_t}$ can tune to wavelength in $\lambda^s_{j_r} \cap \lambda^s_{i_t > i_r}$, and the packet can arrive at $\mathcal{R}^r_{i_r,j_r}$ by using only inter-cluster transmission.

If $i_t > i_r$ and $\lambda^s_{j_r} \cap \lambda^s_{i_t > i_r} = \emptyset$, $\mathcal{T}^t_{i_t,j_t}$ tunes to one of the wavelengths in $\lambda^s_{i_t > i_r}$ and the packet needs both inter-cluster transmission and inter-cluster transmission to arrive at its destination. Suppose $\mathcal{T}^t_{i_t,j_t}$ tunes to λ_{c_2}, where $c_2 \in \lambda^s_{i_t > i_r}$. The packet first arrives at $\mathcal{R}_{i_r, \lceil \frac{Mc_2}{W} \rceil}$. $\mathcal{R}_{i_r, \lceil \frac{Mc_2}{W} \rceil}$ will transmit the packet to $\mathcal{R}_{i_r,j_r}$ using intra-cluster transmission in another time slot. The wavelength assignment algorithm is shown in Algorithm 4. Note, if there are multiple available wavelengths, we use round robin method to choose one of the available wavelengths.

Algorithm 4: PACKET SCHEDULING FOR PODCA-L

```
for packets at the head of output queues do
    PACKET SELECTION ()
    the packet is from T^t_{i_t,j_t} to R^r_{i_r,j_r}
    if i_t=i_r and #pkt for intra-cluster < threshold then
        tune tunable transmitter within intra-cluster network
        if j_r≥j_t then
            tune to wavelength: j_r-j_t+1
        else
            tune to wavelength: M+j_r-j_t+1
        end if
        continue
    end if
    tune tunable transmitter within inter-cluster network
    if i_t≤i_r then
        if λ^s_{j_r} ∩ λ^s_{i_t≤i_r} ≠ ∅ then
            tune to wavelength∈ λ^s_{j_r} ∩ λ^s_{i_t≤i_r}
        else
            tune to wavelength∈λ^s_{i_t≤i_r}
        end if
    else
        if λ^s_{j_r} ∩ λ^s_{i_t>i_r} ≠ ∅ then
            tune to wavelength∈ λ^s_{j_r} ∩ λ^s_{i_t>i_r}
        else
            tune to wavelength∈λ^s_{i_t>i_r}
        end if
    end if
end for
```

To transmit more than one packet in a time slot, each ToR can have more than one tunable transmitters and wide-band receivers for intra-cluster transmission or inter-cluster transmission or both. For intra-cluster transmission, besides multiple tunable transmitters and wide-band receivers, each rack needs a multiplexer and a demultiplexer, as in PODCA-S. For inter-cluster transmission, we denote InterTx and InterRx as the number of inter-cluster tunable transmitters and wide-band receivers on each ToR, respectively. InterTx and InterRx equals the number of couplers connecting to the cluster, the number of demultiplexers connecting to the cluster, and the number of $P \times P$ AWGRs. The nth inter-cluster tunable transmitter and nth inter-cluster wide-band receiver on each rack connect to the nth $P \times P$ AWGR. Compared to [8], our architecture uses AWGR to connect racks within the same cluster and this design can save connection complexity, CapEx, and space. Also, [8] directly connects inter-connectivity AWGR to relay servers by using an optical fiber. This means that each AWGR port can only receive one packet in a time slot. In PODCA-L, each AWGR port can receive up to W packets in a time slot. Furthermore, in [8], there are m relay servers, within a cluster, directly connecting to m AWGR ports. The

scalability is limited when m is large, and throughput is limited when m is small. PODCA-L can achieve high scalability and near-optimal throughput at the same time, and reduce traffic volume from relay servers.

III. Simulation Results

In this section, we conduct simulations to evaluate the latency and throughput performance of PODCA. Packet arrivals follow a Poisson process. The transmission rate of a tunable transmitter (and wavelength capacity) is assumed to be 10 Gbps. The tuning time of tunable transmitters is 8 ns [9]. The latency consists of transmission time, queueing delay, and tuning time. The destination rack of packets is drawn from a uniform distribution. The shared buffer size of each ToR is 8MB. For PODCA-L, we set the threshold for determining if the packet uses intra-cluster transmission or inter-cluster transmission as 100 packets.

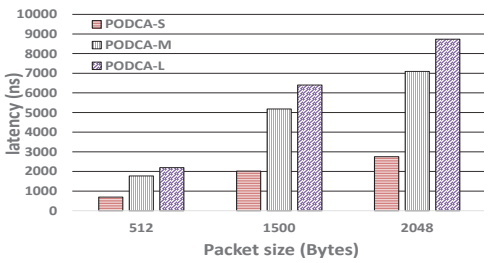


Figure 4. Comparing latency for different packet sizes.

We first compare the packet latency among PODCA-S, PODCA-M, and PODCA-L. The inter-cluster AWGR size for PODCA-S, PODCA-M and PODCA-L is 40×40 and the intra-cluster AWGR size of PODCA-L is 80×80. Thus, the size of PODCA-S, PODCA-M and PODCA-L are 40, 80, and 3200 racks, respectively. We set the packet arrival rate to be 600 pkt/ms per rack.[3] On each rack, there are 3 tunable transmitters for PODCA-S and PODCA-M. For PODCA-L, we set IntraTx=1 and InterTx=2. Figure 4 shows that as the architecture size increases, the latency of PODCA increases. The reason that PODCA-S has much smaller packet latency is that there is no inter-rack wavelength contention within a cluster. Figure 4 shows that the packet size has significant influence on packet latency. As we increase the size of the packet, the transmission time and queueing delay of packets increases, as expected.

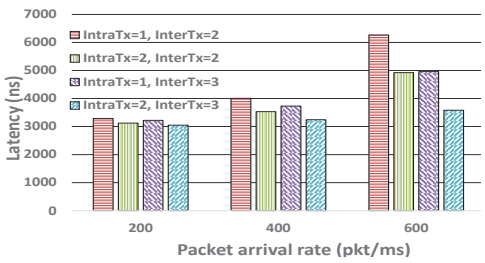


Figure 5. The end-to-end packet latency in PODCA-L.

[3]The transmission rate ranges from 2.4Gbps to 9.6Gbps.

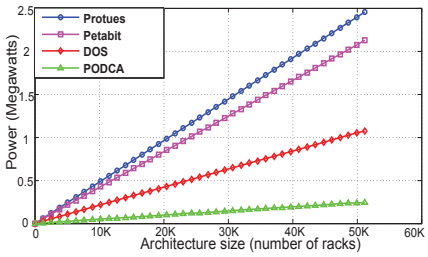


Figure 6. Power consumption comparison of different architectures.

We next evaluate the performance of PODCA-L. We set $W = 80$, $P = 40$, $M = 80$; thus, the size of PODCA-L is 3200 racks. We set the packet size as 1500 bytes. The packet arrival rate per rack is changed from 200 packet/ms to 600 packet/ms. Figure 5 shows that as the packet arrival rate increases, the packet latency increases. Also, as the total number of tunable transmitters per rack increases, packet latency decreases dramatically at high packet transmission rates. However, when the total number of tunable transmitters is fixed, there is almost no difference between increasing intraTx or interTx. The packet latency of our architecture is less than 7 μs. Also, our experiments show there is no packet loss and 100% throughput was achieved.

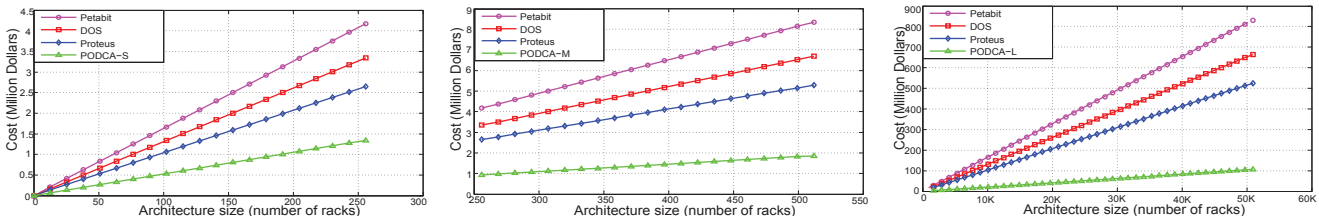
IV. Comparison of Optical DCN Architectures

In this section, we compare our proposed DCN architectures with three other optical DCN architectures (DOS, Proteus, and Petabit) in terms of power consumption and capital expenditure (CapEx). DOS consists of an AWGR, S TWCs and a loop-back shared buffer. Considering the small size of SDRAM, we ignore it in comparisons. Proteus is an all-optical architecture that consists of S WSS switches and an optical switching matrix based on MEMS. In Proteus, each ToR has N optical transceivers. Here, we assume that N is equal to 32, which is presented in the Proteus-2560 system [6]. Petabit is a scalable bufferless optical switch architecture. It adopts a three-stage Clos network and each stage consists of an array of AWGRs. Petabit includes TWC for wavelength routing in AWGRs.

The capacity of a wavelength is set to 10 Gb/s, which is also equal to the transmission rate of a tunable transmitter. The power consumption and unit cost values of the components are mainly taken from commercial product specifications from the literature [6], [10]. The tunable transmitter with 8 ns tuning time is not yet commercially available; in order to conduct the comparison, we assume a 5x price of commodity tunable transmitters. Besides, we assume that the number of transmitters and receivers on each ToR is 2 to trade off between better performance and lower cost. The simulation results show that when the number of transmitters and receivers on each ToR is 2, the performance is comparable to other architectures. We calculate the power consumption and CapEx of these network topologies by summing up the consumed power and dollar cost of each component. A summary of these individual optical components is shown in Table I.

Table I
POWER CONSUMPTION AND COST OF THE OPTICAL COMPONENTS IN DIFFERENT ARCHITECTURES

Component	Power(Watts)	Cost (Dollars)	DOS	Proteus	Petabit	PODCA		
						Small	Medium	Large
SFP transceiver	1.0	45	S	$32S$	0	0	0	0
Fast Wavelength tunable transmitter (WTT)	1.5	195×5	0	0	S	$2S$	$2S$	$2S$
Fixed wideband receiver (photodetector)	0.9	40	0	0	0	$2S$	$2S$	$2S$
Flexgrid WSS	15	1245	0	S	0	0	0	0
Tunable Wavelength Converter (TWC)	20	8000	S	0	$2S$	0	0	0
Arrayed waveguide grating (AWG) per port	0.0	15	0	0	$6S$	S	P	$S+P$
MEMS switch per port	0.25	255	0	$4S$	0	0	0	0
Circulator	0.0	105	0	$4S$	0	0	0	0
Coupler	0.0	195	0	S	0	S	P	P
Optical (DE)MUX	0.0	3000	S	$2S$	0	S	P	P
Fiber Delay Line	0.0	2000	S	0	0	0	0	0



(a) (b) (c)

Figure 7. CapEx comparison of different architectures: (a) PODCA-S. (b) PODCA-M. (c) PODCA-L.

Based on the details in Section II, we have three different architectures that could be employed for different sized networks. Note that we know P and W can both scale up to 512 [11]. Here, we set the size of PODCA-S to a maximum of 256 racks. The sizes of PODCA-M and PODCA-L are $256 < S \leq 512$ and $512 < S \leq 51200$, respectively. We assume that each rack consists of 48 servers. Therefore, the number of servers can vary from thousands to more than 2 million. Note, for PODCA-M, we set M=2 and S=2·P. Also, for PODCA-L we set M=100 and S=100·P.

Figure 6 presents the overall power consumption with increasing number of racks. Our main power-driven components include tunable wavelengths transmitters and fixed wide-band receivers. The fixed wide-band receiver consists of just a photodetecor and corresponding circuitry. We note that the power consumption of PODCA is dramatically lower compared to the Protues (saving 90%), Petabit (saving 87%) and DOS (saving 75%), when the number of racks increases up to 51K. The enormous power consumption of those architectures is caused by active components such as the TWC in DOS (and Petabit), and Flexgrid WSS in Proteus. Besides, multiple SFP transceivers required in each ToR cause huge power consumption in Proteus.

Figure 7(a), (b) and (c) show that our architectures can save CapEx as well. In our architectures, instead of choosing expensive devices such as TWC and WSS, we use relatively cheap devices, e.g., tunable wavelength transmitters and AWGRs. Figure 7(a) and 7(b) show that PODCA-S and PODCA-M outperform the other three architectures, and PODCA-S and PODCA-M can save at least 40% CapEx. The results shown in Figure 7(c) indicate that PODCA-L saves 88%, 84% and 80% CapEx compared with Petabit, DOS, and Proteus, respectively.

V. CONCLUSION

In this paper, we present architectures for a passive optical data center network. The key component of our architectures is the arrayed waveguide grating router (AWGR). Based on simulation results, the packet latency of our architectures is below 9μs and our architectures can achieve 100% throughput. Compared with other optical architectures in terms of power consumption and CapEx, results show that our architectures can save up to 90% power consumption and 88% CapEx.

REFERENCES

[1] 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*. ACM, 2010, p. 24.

[2] Pierre Delforge, Available online: "http://www.nrdc.org/", 2015.

[3] Available online: "https://www.citycloud.com/".

[4] Available online: "http://www.datacenterknowledge.com/".

[5] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," in *IEEE Comm. Surveys & Tutorials, 14(4):1021-1036*, 2012.

[6] 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*. ACM, 2010.

[7] 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.

[8] Z. Cao, R. Proietti, M. Clements, and S. Yoo, "Demonstration of scalable, flat, and high-throughput data center architecture based on arrayed waveguide grating routers," in *ECOC*, 2014.

[9] S. Matsuo and T. Segawa, "Microring-resonator-based widely tunable lasers," *IEEE J. Sel. Topics Quantum Electron*, vol. 15, no. 3, pp. 545–554, 2009.

[10] J. Chen, Y. Gong, M. Fiorani, and S. Aleksic, "Optical interconnects at the top of the rack for energy-efficient data centers," *Communications Magazine, IEEE*, vol. 53, no. 8, pp. 140–148, 2015.

[11] S. Cheung, T. Su, K. Okamoto, and S. Yoo, "Ultra-compact silicon photonic 512× 512 25 GHz arrayed waveguide grating router," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 20, no. 4, pp. 310–316, 2014.