Design and Evaluation of a Flexible-Bandwidth OFDM-Based Intra Data Center Interconnect

Philip N. Ji, Dayou Qian, Konstantinos Kanonakis, Christoforos Kachris and Ioannis Tomkos

Abstract—Data center networks are facing growing challenges to deliver higher bandwidth efficiency, lower latency, better flexibility and lower cost. Various optical interconnect schemes have been proposed to take advantage of the high bandwidth capacity and low power consumption offered by optical switching. However, these schemes cannot offer flexible bandwidth sharing due to the large granularity in optical circuit switching, and they require costly optical components. In this paper, we introduce a novel data center network architecture based on cyclic arrayed waveguide grating device and multiple-input-multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) technology with parallel signal detection (PSD). This architecture offers flexible bandwidth resource sharing at fine granularity. Other features include high-speed switching, low and uniform latency, and the ability to change the data rates dynamically. By eliminating costly optical components and keeping the core optical router passive and static, the power consumption, hardware cost and operation cost are reduced. The fine granularity bandwidth sharing and MIMO switching through PSD are verified experimentally. We also propose and evaluate efficient subcarrier allocation schemes to achieve high bandwidth utilization. Finally, we present the implementation of an efficient scheduler for the bandwidth allocation of the proposed scheme.

Index Terms—Optical interconnects, MIMO OFDM, data center networks.

I. INTRODUCTION

As the global Internet traffic growing exponentially, the data centers, which host many Internet application servers, are also facing rapid increase in bandwidth demands. Due to emerging applications like cloud computing, next generation data centers need to achieve low latency, high throughput, high flexibility, high re-source efficiency, low power consumption, and low cost. Furthermore, as more and more processing cores are integrated into a single chip, the communication requirements between racks in the data centers will keep increasing significantly. By integrating hundreds of cores into the same chip (e.g. Single-chip Cloud Computer-SCC [1]) we can achieve higher processing power in the data center racks. However these cores require a fast and low-latency interconnection scheme to communicate with the storage system and the other servers inside or outside of the rack.

Optical technology has been adopted in data center networks (DCN) due to its high bandwidth capacity. However, it is mainly used for point-to-point links, while the intra Data Center Network (DCN) interconnection is still based on electrical switching fabrics, which have high power consumption and limited bandwidth capacity [2]. Currently, the power consumption of data center networks accounts for 23% of the total IT power consumption [3]. However, due to the high communication requirements of the future data center networks, it is estimated that the data center networks will account for much higher percentages of the overall power consumption [4]. Therefore it is expected that data center networks may evolve to all-optical networks, much like telecommunication networks themselves which have gradually evolved from electrical to opaque and finally transparent, using all-optical switches.

In recent years, several hybrid optical/electrical or all-optical interconnect schemes for DCN have been proposed [5], [6], [7], [8], [9], [10]. Many of them rely on large scale fiber cross-connects (FXC) ([5], [6], [9]) or multiple wavelength-selective switches (WSS) [7], which are though costly and imply slow switching rates (at the millisecond timescale). Having a large scale FXC also present an undesirable single source-of-failure. A recent work in [8] uses silicon electro-optic mirroring WSS and semiconductor optical amplifier-based switch to achieve nanosecond scale switching, making all-optical packet level routing possible. However the key components are not commercially available and have low scalability. Other architectures use tunable wavelength converters ([9], [10]). They are also costly and do not allow bandwidth resource sharing among the connections. Some of them also require electrical or optical buffer.

In this paper we propose and experimentally demonstrate a novel all-optical DCN interconnect which, for the first time to our knowledge, combines a passive cyclic arrayed waveguide grating (CAWG) core router with the technologies of orthogonal frequency division multiplexing (OFDM) modulation and parallel signal detection (PSD). The proposed architecture achieves fast switching (nanosecond speed), low latency, low power consumption, multiple-input-multiple-output (MIMO) switching, while allowing fine granularity bandwidth sharing without the need for FXC, WSS, or tunable wavelength converters. Moreover, we investigate methods of efficiently performing resource allocation in this OFDM-based interconnect, evaluate their delay and throughput performance and study their feasibility for hardware implementation.

II. MIMO OFDM FLEXIBLE BANDWIDTH INTRA-DCN ARCHITECTURE

A. MIMO OFDM ARCHITECTURE

A key technology for this DCN architecture is MIMO OFDM. OFDM is a modulation technology to achieve high
spectral efficiency transmission by parallel transmission of spectrally overlapped, lower rate frequency-domain tributaries where the signals are mathematically orthogonal over one symbol period (Fig. 1). Originally applied for copper and wireless communications, OFDM technology has been adopted in optical communication network applications in the past few years as the high speed digital signal processing and broadband DAC/ADC became feasible [11][12]. Because of the advantages such as better tolerance to fiber dispersion and the ability to perform one-tap equalization at the frequency domain, OFDM has been demonstrated to be a good candidate for long distance transmission [13][14][15]. OFDM technology has also been proposed for optical access network to take advantage of its flexibility to share the spectrum among multiple users, such as in the application of OFDMA-PON (orthogonal frequency division multiple access passive optical network) [16]. The feasibility of using OFDM technology for data center application has been discussed in [17], but no actual network architecture has been proposed for the intra-data center network.

There are mainly two types of OFDM implementation in optical transmission. The first type is to generate the OFDM signal electrically and modulate the signal to an optical carrier [13][14][15]. This is referred to as the optical OFDM (O-OFDM). The receiver can use direct detection or coherent detection techniques. The second type is to generate the orthogonal subcarriers (or orthogonal frequency division multiple access passive optical network) [16]. This is referred to as the all-optical OFDM (AO-OFDM).

The DCN architecture proposed in this chapter is based on an O-OFDM implementation. It demonstrates network-level MIMO operation because each rack can send the same OFDM signal to multiple destination racks simultaneously, and multiple racks can send the signal to one destination rack at the same time by modulating data on different OFDM subcarriers in the RF domain. At each receiver, a common photo-detector (PD) can simultaneously detect multiple O-OFDM signals from many sources in different optical wavelengths, provided that there is no contention in the OFDM subcarriers and the WDM wavelengths. This is referred to as the parallel signal detection (PSD) technology [19], and has been demonstrated in OFDM WDM-based optical networks [20].

B. Cyclic Arrayed Waveguide Grating

The key optical component for the proposed DCN architecture is a CAWG. An \(N \times N\) CAWG (also called an AWG router or a cyclic interleaver) is a passive optical multiplexer/demultiplexer that routes different wavelengths from \(N\) different input ports to \(N\) different output ports in a cyclic manner. Fig. 2 illustrates the cyclic wavelength arrangement of an \(8 \times 8\) CAWG. CAWG is usually constructed using planar lightwave circuit technology. The cyclic wavelength arrangement characteristic avoids the wavelength contention and eliminates the need for large scale core FXC or multiple WSS units. Several other DCN architectures also use CAWG as the core optical router (9)[10].

C. The Proposed DCN Interconnect Architecture

The schematic of the proposed MIMO OFDM DCN interconnect architecture is illustrated in Fig. 3. It contains \(N\) racks, each accommodating multiple servers connected through a top-of-the-rack switch (ToR). Inter-rack communications are performed by interconnecting these ToRs through the DCN.

The inter-rack signals at each rack are aggregated and sent to a transmitter, which contains an OFDM modulator that modulates the aggregated signals into \(K\) OFDM data streams with appropriate subcarrier assignments, where \(K\) is the number of destination racks that the signals from this source rack need to travel to, so \(0 \leq K \leq N\). Each source rack can be connected to a different number of destination racks at each time (e.g. \(K\) can be different for each rack). These OFDM data streams are converted to \(K\) WDM optical signals through an array of \(K\) directly modulated lasers (DMLs) or \(K\) sets of laser/modulator with different wavelengths. If these lasers are of the fixed wavelength type, \(N\) units will be needed since the signal from each ToR might be switched to any destination rack potentially.

If the number of the racks increases, it is not cost-efficient to install \(N\) lasers at each transmitter. Moreover, it is not likely that each rack needs to communicate with all other racks simultaneously. Therefore the \(N\) fixed wavelength lasers in the transmitter can be replaced with fewer tunable lasers.
These O-OFDM signals are then combined through a WDM combiner to form an OFDM-modulated WDM signal and sent to the $N \times N$ CAWG. Due to the cyclic non-blocking wavelength arrangement of the CAWG, the WDM channels are routed to the respective output ports for the destination racks. Each optical Rx receives one WDM channel from each input port and all WDM channels can be simultaneously received by a single PD using the PSD technology. Through a centralized OFDM subcarrier allocation scheme that will be described below, it is ensured that there is no subcarrier contention in any WDM channel at the receiver side. The received OFDM signal is then demodulated back to the original data format and send to the appropriate servers through the destination ToR.

When a new switching state is required, the OFDM modulators execute the new subcarrier assignments determined by the centralized scheduler, and the respective lasers turn on and off to generate the new OFDM WDM signals. Some servers in the DCN may constantly have large communication volumes with certain servers. It will be less efficient for them to go though the ToR before exiting the rack. In some cases, the large traffic volume from these servers might even congest the ToR. To serve these “super servers” more effectively, the MIMO OFDM DCN architecture can be extended to reserve dedicated OFDM WDM transmitters and dedicated CAWG ports for them. These servers can bypass the ToR and connect to the transmitters directly, as illustrated in Fig. 4.

**D. DCN Interconnect Architecture Features**

Comparing it to other optical or hybrid intra-DCN architectures proposed so far, the proposed architecture offers the following advantages:

**MIMO switching:** Conventional optical DCN architectures use optical circuit switching. Each rack can only talk to one rack at a time. It needs to wait for the current connection to complete before another connection can be established for the same rack. Due to the MIMO OFDM operation in this architecture, each rack can communicate with multiple racks simultaneously. Thus the waiting time is eliminated and high interconnect efficiency can be achieved.

**Flexible bandwidth allocation and sharing:** By flexibly selecting the number of subcarriers at each O-OFDM transmitter and sharing the total available subcarriers at each receiver among multiple sources, this architecture allows providing different bandwidth allocations for different source-destination pairs at the same time, which can be dynamically modified over time. Such a feature is extremely suitable for DCN applications where there are frequent setting up and tearing down of connections and large fluctuation in bandwidth demands.

**Fine granularity switching:** Since the O-OFDM signal is generated electrically, the switching granularity is much finer than the current optical DCN technologies. For example, in direct optical point-to-point link, the granularity is one fiber; in a regular WDM system, the granularity is one WDM channel, which typically carries 10 Gb/s, 40 Gb/s or 100 Gb/s of data; in the AO-OFDM system, the granularity is one optically generated OFDM subcarrier, which is typically 10 Gb/s or higher. The switching granularity in the O-OFDM system is a single electrically-generated OFDM subcarrier, which is typically in the order of tens of Mb/s or less. Having a finer granularity allows more flexible bandwidth allocation and thus more efficient spectrum utilization.

**Flexible modulation format and data rate:** OFDM modulation also provides the capability to change the modulation order to adjust the amount of data to be carried within the same subcarrier (or group of subcarriers). For example, the OFDM signal in each subcarrier can be modulated using BPSK, QPSK, 16-QAM, 64-QAM, etc. This allows a variable amount of data to be packed within the same subcarriers as these modulation formats encode a different number of data bits in each symbol. This feature could be potentially used to
solve the congestion issue at the destination racks. Note also that within the same OFDM signal different subcarriers can use different modulation formats.

**No guard band:** In PSD-based OFDM system, guard bands are usually required between subcarrier groups from different sources due to the synchronization difficulty and impairments during transmission, such as dispersion and OSNR degradation. However, such guard bands are not required in the intra-DCN case because the transmission distance is short (typically from tens of meters up to 2 km). This allows maximum bandwidth utilization at each receiver.

**Fast switching:** This architecture performs optical circuit routing by turning respective lasers on and off. This can be achieved at sub-nS level, making this system feasible for packet level operation. If tunable lasers are used, the switching speed will then be determined by both the laser on/off speed and tuning speed, which can also be realized at ns level [21]. All switched signals take exactly one hop (i.e. passing through the core optical router only once), providing the opportunity for low and uniform latency as opposed to the multiple buffering levels required in several electrical switch-based DCN architectures. Moreover, due to the use of MIMO operation and bandwidth sharing via OFDM subcarrier allocation, no optical buffers are required (like for instance in [9]).

**Scalability:** The key factor that determines the available scale of this DCN is the port count of the CAWG. A 400-channel AWG based on 6-inch silica waver with 25GHz spacing has been reported in [22] and a 512-channel AWG based on 4-in silica wafer have been demonstrated more than a decade ago [23]. With the recent advancement in silicon photonics technology, even higher port count CAWG can be expected because silicon waveguide can achieve higher core/cladding index contrast and thus allow waveguide bending radius to be several order of magnitude lower than silica waveguide [24]. Therefore, although it is not trivial to increase the number of channels with decreasing the channel spacing, the number of supported channels of current CAWGs (400 or 512) is still much larger than current Ethernet switches with 64 or 128 ports. Furthermore, as opposed to the system proposed in [9], there is no need for multiple receivers at each port.

**Simplified control:** Although the subcarrier contention restriction at each receiver of course exists, the proposed architecture allows the same OFDM subcarriers to be used by different OFDM signals generated by the same transmitter. Therefore the subcarrier allocation problem can be considered independently for each receiver, thus greatly reducing the associated algorithmic complexity.

**Low power consumption:** Because the core optical router in this architecture is completely passive and static, the optical components have lower power consumption compared with other optical DCN architectures that require switching through WSS or FXC. The heat dissipation is also lower.

**Low cost:** Since this architecture does not require FXC, WSS or tunable wavelength converter, the optical component cost is low. Having low power consumption and low heat dissipation also reduces the cooling requirement and the operation cost.

### TABLE I

<table>
<thead>
<tr>
<th>Test</th>
<th>From ToR 1</th>
<th>From ToR 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SC 100-450</td>
<td>SC 300-450</td>
</tr>
<tr>
<td></td>
<td>16QAM</td>
<td>64QAM</td>
</tr>
<tr>
<td></td>
<td>QPSK A</td>
<td>QPSK D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SC 200-600</td>
<td>SC 50-150</td>
</tr>
<tr>
<td></td>
<td>16QAM</td>
<td>700-900</td>
</tr>
<tr>
<td></td>
<td>QPSK B</td>
<td>QPSK C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**III. EXPERIMENTAL DEMONSTRATION OF FLEXIBLE MIMO OFDM OPTICAL INTERCONNECT**

The flexible MIMO interconnect capability of the proposed architecture is demonstrated on a lab testbed. The optical core router is an 8 × 8 CAWG with 100 GHz spacing by Enablence [25]. Each transmitter contains two tunable external cavity lasers with 10 GHz intensity modulators. The OFDM signals are generated by an arbitrary waveform generator. The main advantage of the demonstrated testbed is that the PSD does not require guard band between subcarrier groups from different sources, because of the short distance in DCN compared to other optical networks. Furthermore, different modulation formats for different subcarrier groups can be used in this testbed based on the traffic requirements.

Each signal consists of 1200 subcarriers, each occupying a bandwidth of 5 MHz. Two scenarios using different wavelength and subcarrier assignments are tested (Table I). Due to the cyclic wavelength routing arrangement in the CAWG, λ3 and λ4 from ToR1 are routed to ToR3 and ToR6 respectively, while λ4 and λ7 from ToR2 are routed to ToR3 and ToR6 respectively.

Different modulation formats, including QPSK, 16QAM and 64QAM, are used for different OFDM subcarrier groups highlighted in black, producing per-subcarrier data rates of 10 Mb/s, 20 Mb/s and 30 Mb/s respectively. A single-ended, 10 GHz bandwidth direct-detection PD is used at each CAWG output to receive the WDM signal and convert it to an electrical OFDM signal. This OFDM signal is then captured and digitized by a real time oscilloscope and processed using an offline computer to recover the data from the OFDM signal. No optical amplification was used in this experiment.

The RF spectra of the OFDM signals are measured at each OFDM transmitter output each the receiver output for different tests (Fig. 5). They show that different OFDM signals from the same transmitter can overlap (such as subcarrier group A and subcarrier group B in Test 1) or non-contiguous subcarriers (such as subcarrier group E), and through the PSD technology each receiver can successfully detect and receive OFDM signals from multiple sources, provided that these OFDM subcarriers do not overlap. No guard band is required when assigning the OFDM subcarriers, as shown at the receiver of ToR 3 in Test 1. Besides allocating subcarriers,
the centralized controller also balances the subcarrier powers among signals from different inputs.

The optical spectra of the WDM signals at the CAWG input ports from ToR 1 and ToR 2 transmitters and CAWG output ports to ToR 3 and ToR 6 receivers are shown on Fig. 6. They confirm the non-blocking cyclic routing operation of the CAWG.

The PSD receiver performance (solid symbols), represented by the bit error rates (BER) under different received optical power levels, are measured and compared with single channel receiver (hollow symbols) for different OFDM signals at each test (Fig. 7). While the absolute BER value varies with the modulation format and per-subcarrier power level, no significant degradation between single channel detection and PSD detection is observed in any tests, and OFDM signals with different modulation formats can all be detected successfully. This shows that PSD technology is feasible for receiving multiple OFDM WDM signals simultaneously using a single PD, and thus realizing MIMO switching.

IV. PERFORMANCE EVALUATION

A. Simulation Model and Traffic Assumptions

In order to evaluate the performance of the proposed architecture, a custom simulation model was built using OPNET Modeler. The model simulated an OFDMA-based switch operating as described above, consisting of 8 ToRs. Each ToR receiver used 1000 subcarriers, offering a bitrate of 10 Mbps each, hence the total ToR-ToR capacity is 10 Gbps. The sizes of the produced packets followed a tri-modal distribution, with sizes of 64, 192 and 1408 Bytes appearing with probabilities of 0.05, 0.5 and 0.45 respectively. The latter was shown in [26] to constitute a good approximation, based on real data center network traffic measurements. Note that all arriving packets were segmented in 64-Byte cells, while traffic produced by each ToR was destined with equal probability to all other ToRs. Two different traffic models were considered: In the first one (most commonly used in switching literature), each ToR produces packets according to a Bernoulli distribution at a constant rate (the latter defined by the average offered load and packet size) - hence the inter arrivals of packets towards a given ToR follow a geometric distribution. A buffer of 128 kB was dedicated to each source-destination ToR pair. In the second traffic model, self-similar traffic was created at each ToR using a multiplicity of ON-OFF sources, with Pareto-distributed ON and OFF periods in order to simulate the bursty nature of real data center network traffic (also reported in [26]). The Hurst parameter was 0.8 while and the average OFF/ON ratio was 10. In this case, the buffer size per ToR pair was chosen to be 10 MB due to the increased traffic burstiness.

B. Subcarrier Allocation Schemes

In this section we propose and evaluate two different subcarrier allocation schemes for exploiting the considered architecture:

1) Fixed Subcarrier Allocation (FSA): In FSA, virtual transmission pipes of fixed bandwidth are created for each ToR pair by statically assigning to each of them a number of subcarriers. Since in this work we considered uniform average traffic, each ToR pair was assigned 125 subcarriers (i.e. 1000/8). The main advantage of FSA is its implementation simplicity. However, to the negative side, it cannot adapt to bursty traffic and in such cases it is expected to lead to increased packet delay. Note that a similar scheme could be realized in the time domain by serving the input ToRs in a round-robin fashion and dedicating to each of them one timeslot (cell) per output ToR in each round. The packet delay in that case would obviously be roughly the same as in FSA; However, FSA can scale much better to higher bitrates per output port, since the electronics controlling the communication per ToR pair can operate at 1/N of the output port rate.

2) Dynamic Subcarrier Allocation (DSA): We propose this scheme as a feasible approach to achieve assignment of subcarriers for each ToR pair according to their actual traffic needs. In that respect, DSA is executed periodically (the scheduling period is denoted as T) and adjusts the current assignment based on traffic information collected from all ToRs. More specifically, a ratio (denoted as f) of the available subcarriers is distributed equally to all ToR pairs in a fixed manner. The rest are assigned each time in a weighted fashion, with the weights decided according to the traffic rate in each of them during the previous scheduling period. Note that T should be selected carefully, since on the one
hand it should be relatively short to avoid affecting QoS (i.e. inaccurate decisions should be corrected quickly) and on the other hand, a very short $T$ would impose challenges in the relevant electronics and could produce inaccurate measurements (depending on the exact traffic characteristics). As a final note, it should be mentioned that FSA could also be considered as a subset of DSA, with $f$ equal to 1 and $T = \infty$ (since it is executed only once in the beginning).

Moreover, we also obtained results from two electrical switching schemes and compared them with FSA and DSA. The first one, which uses virtual output queues at the transmitter side (like FSA and DSA do) is called Dynamic Timeslot Allocation (DTA). DTA operates similarly to DSA, however it dynamically assigns a number of timeslots (equal to the cell duration at 10Gbps) to each input port during each scheduling period $T$. Each ToR input port can transmit to an output port at 10Gbps, but only during the timeslots assigned to it. The second scheme follows an Output Queueing (OQ) paradigm. In OQ, arriving packets from any ToR towards a specific destination ToR are transmitted in FIFO order using the full 10Gbps output port rate. In that way, no bandwidth resources are wasted and thus OQ is expected to offer the minimum possible average packet delay. This makes it very useful for performing the benchmarking of the aforementioned algorithms. Note though that (as it is well-known) OQ is not realistic, since it implies an internal speed-up of $N$ (obviously making it not scalable even for moderate-sized switches).

C. Simulation Results

Performance of the proposed algorithms was evaluated for different network loading conditions, using the models described above. Note that the load values shown here, indicate the average aggregate load as a ratio of the total available switch capacity.

Fig. 8 shows the average packet delay for all the considered subcarrier allocation schemes under the self-similar traffic model. First of all it is clear that, as expected, OQ and FSA offer the best and worst performance respectively. In particular, FSA proves to be completely inadequate since it exceeds OQ by far, even at moderate loads (i.e. around three orders of magnitude for 0.5 load). Moreover, FSA resulted in heavy packet loss at higher load (around 0.3% and 2% for 0.5 and 0.7 loads respectively - while loss in DSA was 0 in all scenarios), rendering it practically unusable by most applications.

At the same time, the use of DSA (a $T$ value of 1ms was used) offers delay values that are much closer to OQ over a wider range of loading conditions. The use of a hybrid DSA/FSA approach by means of a non-zero $f$ value (0.25 here) seems to improve performance, though only at low to moderate load. For higher loads, the static assignment of part of the subcarriers results in increased average delay.

Note, however, that the exact traffic profile (besides the average offered load) is also expected to heavily influence DSA performance. For example, Fig. 9 depicts the actual packet delay (averaged over 1ms measurement intervals) for DSA at a load of 0.5, with different $f$ values and $T = 1ms$. From Fig. 9(a) it is obvious that $f = 0$ results in much lower delay during extreme traffic peaks (i.e. at around 0.2s and 0.5s), due to its ability to reallocate the available subcarriers to the ToR pairs that temporarily require them. However, a closer examination during the rest of the time (as shown in Fig. 9(b) for the period between 0.6s and 0.7s) indicates that an increased $f$ actually results in lower delay, since there is a transmission pipe available for serving a packet upon its arrival. As a result, it can be generally deduced that the burstier the expected traffic patterns, the lower the recommended $f$ value.

Next, in Fig. 10 and Fig. 11 we depict average packet delay
results under the Bernoulli traffic model (including the DTA scheme this time), with $T = 10 \mu s$ and $100 \mu s$ respectively. For DSA and DTA a relatively high $f$ value was chosen 0.375), due to the smoother traffic used. It is evident from both figures that (as expected from the discussion above) the difference between FSA and DSA is not as prominent as in the case of self-similar traffic. A very important finding though, is that DSA always performs better than DTA. The reason is that DTA is much more affected by the choice of scheduling period $T$ (DTA delay increases significantly for $T = 100 \mu s$). The results indicate the DSA can operate well even by performing the scheduling process less frequently. As a result, it can scale well for much higher bitrates without prohibitive increase in the complexity of the electronics.

Finally, Fig. 12 and Fig. 13 depict packet throughput (i.e. aggregate bitrate corresponding to correctly received packets) for $T = 10 \mu s$ and $T = 100 \mu s$ respectively. First of all, it is important to point out that, even with the limited buffering assumed, DSA throughput remains at reasonably high levels in all cases. The throughput performance of OQ almost approaches the offered load, while DTA again performs much worse than DSA - especially for $T = 100 \mu s$. It should also be noted that, although at high load FSA performs similarly to DSA in terms of average delay, its throughput is evidently lower. This can be attributed to the fact that DSA handles traffic spikes better, thus reducing the possibility of buffer overflows.

D. Scheduler Implementation

In this section we describe the implementation of the subcarrier assignment algorithm for the proposed architecture. As it was described above, the most efficient scheme is DSA, in which the number of subcarriers is assigned dynamically based on the traffic rate of each port. The high-level architecture of the subcarrier assignment process is depicted in Fig. 14. Congestion management using dynamic subcarrier allocation is performed using specialized control packets, as it is described in [27].

Each ToR switch first evaluates the required number of subcarriers for a specific port based on the traffic rate of the packets targeting this port. Then the ToR switches send the requested subcarriers through the CAWG to the destination port through control packets (shown in Fig. 14 as 1). The receiver extracts the requested subcarriers and feeds the scheduler 2 which calculates the subcarriers allocated for each input port. Of course, the total number of allocated subcarriers for all the transmitters should always be lower than the maximum number of supported subcarriers (e.g. 1024). The scheduler
The scheduler could also be configured to work in the same way using the utilization of buffers instead of the traffic rate. In this case, more subcarriers would be requested and assigned to ports that have almost full buffers.

As the MIMO OFDM receivers can accept simultaneously the subcarriers from all wavelengths, the implementation of the scheduler is straightforward. The scheduler must divide the requested subcarriers of each port by the total requested subcarriers for this port and then to assign accordingly these subcarriers. Assuming a 32x32 switch with 1024 subcarriers for this port and then to assign accordingly these subcarriers to the corresponding ToR switch [3], again using control packets. Finally, each ToR sends the data based on the number of allocated subcarriers though the transmitters [4].

The scheduler has been also implemented in hardware and has been ported to an FPGA. In this case, 32 dividers are used in parallel to accelerate the scheduling time as it is depicted in Fig. 15. The result of each divider is added to the fixed number of subcarriers \( f \) in order to calculate the total number of assigned subcarriers. The maximum clock frequency after place-and-route, using a Xilinx Virtex 5 FPGA, is 500MHz. Thus, the execution time of the hardware scheduler is \( 32 \text{div/port} \times 34 \text{cc} \times 2 \text{ns} \approx 2 \mu s \). In the case of a commercial product the arbiter could be implemented in an ASIC reducing the latency by 3x to hundreds of nanoseconds [29].

The main advantage of a hardware implementation of the scheduler is that it can scale efficiently to a high number of ports and subcarriers. Table II shows the scheduling time \( T \) for the software-based and the hardware-based scheduler for different number of ports. As it is shown in the table, the software-based scheduler can support up to 128 ports (since the execution time is \( O(N^2) \)), where \( N \) is the number of ports) while the hardware-based scheduler can efficiently support a much higher number of ports since it scales linearly with the number of ports \( O(N) \).

For determining the minimum scheduling period \( T \), we also have to take into account the communication latency between the requested subcarriers and the allocated subcarriers. As it was described above, the requested subcarriers are sent to the destination port and the destination port sends back the allocated subcarriers. Thus the total scheduling period is \( T \geq rtt + t_{sch} \), where \( rtt \) is the round-trip time and \( t_{sch} \) the processing delay of the scheduler. However, the control packets that carry information about the subcarriers (requested and allocated) can have higher priority than data packets, thus eliminating the need for buffering into the memory of the transmitters. Therefore, \( rtt \) is practically fixed and depends only on the delay of the transmitter, the receiver and the CAWG. Hence, the minimum scheduling period is basically determined by the processing time of the scheduler, which (as mentioned earlier) scales linearly with the number of ports.

### E. Power consumption

The main advantage in the proposed scheme is that there is no need for an aggregate switch since the switching is
performed directly at the optical domain through the use of the CAWG. The CAWG is a passive optical component and therefore it does not consume any power. The switching is performed by tuning accordingly the DML-based transmitters. Therefore we can eliminate the power consumption of the crossbar switches, the electrical-to-optical and optical-to-electrical transceivers (e.g. SFP+) at the aggregate switches. The only overhead of the proposed scheme is that the optical transceivers at the ToR switches have to be replaced by new transceivers that also perform the FFT/IFFT processing.

A simple 10Gbps optical multi-mode transceiver (e.g. SFP+) dissipates around 1.5W [30]. The power consumption of the IFFT/FFT processing and the subcarrier modulation that is used in OFDM is estimated to 50mW/Gbps [31]. The OFDM system requires also a high speed ADC and DAC modules that dissipates around 1W and 1.2W respectively [31]. Finally, the optical transponder consumes 2.4W at 40Gbps [32]. Therefore, the total power consumption of each OFDM transceiver used in the ToR switch is around 6.6W for a 40Gbps transceiver.

Table III shows the power consumption of a reference design and the MIMO OFDM scheme for a typical data center with 40 racks. The MIMO OFDM scheme uses 8 40Gbps OFDM transceivers per rack that can provide low latency and eliminate the congestion [33]. The reference design is based on commodity switches using a fat-tree topology. To evaluate the power consumption of the reference design we used the Cisco 3048 Ethernet Switch as the ToR and the Cisco 5020 as the aggregate switch. The Cisco 3048 has 48 Ethernet links at 1Gbps and 4 SFP+ ports at 10Gbps that are used to connect with the aggregate switch. The typical power consumption of this ToR switch is 124W [34]. The Cisco 5020 aggregate switch has 48 Ethernet ports at 10Gbps and the typical power consumption is 480W [35]. As it is shown the power consumption of the MIMO OFDM-based network can achieve up to 25% less power consumption that is due to the elimination of the aggregate switches. The switching at the aggregate level is performed in the optical domain through the configuration of the OFDM transceivers thus reducing significantly the overall power consumption.

V. CONCLUSIONS

We propose a novel DCN architecture utilizing OFDM and PSD technologies. This architecture offers fine grain bandwidth allocation, high switching speed, and low and uniform latency. We experimentally demonstrate the MIMO OFDM switching and fine granularity flexible bandwidth sharing features. We also develop efficient subcarrier allocation algorithms for the fine grain bandwidth allocation. The simulation results shows that the MIMO OFDM-based architecture provides low latency and high-throughput switching. Furthermore, the implementation of the scheduler shows that it can support a high number of nodes and subcarriers. Therefore the proposed architecture could be a promising solution for all-optical interrack and inter-server communication in next generation data center networks.

REFERENCES
