WiCOD: Wireless Control Plane Serving an all-Optical Data Center

Tugcan Aktaş, Chang-Heng Wang and Tara Javidi
Department of Electrical and Computer Engineering
University of California, San Diego
Email: {taktas, chw009, tjavidi}@ucsd.edu

***Abstract*—A novel architecture for the future data center networks with possibly up to a thousand of Top of the Rack (ToR) switches is proposed. The proposed architecture, WiCOD, relies on a wireless control plane serving an all-optical data plane. The first contribution of the work is the separation of the data and the control planes: while the data is switched between the ToR switches in an all-optical high rate network, the network state and control information is continuously conveyed to and from a central unit over an ultra low-latency wireless network. A proof of concept for this architecture is also presented by considering the initial design possibilities for each one of the planes. In order to obtain low packet delays, the data plane scheduling policies take non-zero reconfiguration and monitoring delays into consideration. The results prove that very low queueing delays are guaranteed for strictly frequent updates on the network state. Based on this observation, a technique for *monitoring* of ToR switch queue occupancy information is purposed. This monitoring technique uses mmWave wireless communications via a spatially adapted MIMO Orthogonal Frequency Division Multiple Access (MIMO-OFDMA) over a static frequency selective channel with large number of densely packed ToRs. The reduced monitoring delays, offered by this low-latency radio access technology, makes the fine-grained and adaptive circuit switching feasible and, in turn, enables a high utilization of optical switches.**

I. Introduction

It is well-known that internet style transportation of data puts the emphasis on distributed operation and scalability. On the other hand, as it has been recently observed in the networking literature, the data center networking can significantly depart from classical and Internet-inherited networking in order to allow fine-grain management and scheduling of the flows of data [1], [2]. This departure from classical networking has been motivated by the fact that any given data center is managed, more or less, by a single entity. Furthermore, the intra-data center networks consists of end nodes that are densely packed in a small area.

The overall objective of this work is to develop a framework, from first principles, which relies on the above unique attributes of data centers and advances in wireless and optical communications, to propose a transformative new networking architecture with increased level of efficiency and significantly smaller latency. In this paper, we pursue this goal by introducing a hybrid wireless/optical architecture in which mmWave wireless technology is used to provide a very fast, low-latency, moderate-aggregate-rate monitoring/control plane to enable dynamic end-to-end scheduling of optical switches (fine-grained circuit switching) as illustrated in Fig. 1. We name this architecture WiCOD: **Wi**reless **C**ontrol Plane Serving an all-**O**ptical **D**ata Plane. Additionally, we provide simple prototype solutions implementing the data and control planes. In our choice of these prototypes, we have restricted our attention to simple and fairly well-understood principles as a proof of concept. Our preliminary results are included which go beyond viability and underline the main challenges in terms of scalability, efficiency, and cost.

Our proposed architecture shares important attributes with recent networking solutions such as pFabric [1] and Fastpass [2]. Our proposed architecture also aims at (near-)zero in-network buffering. Both studies show the benefits of a zero-buffer networking solution extensively. In addition to these known gains identified in these studies, however, our emphasis on zero-in-network buffering is also motivated by our interest to integrate optical switching technologies in the data center. In particular, we are interested in utilizing fast optical connections without costly and unscalable optical-electronic-optical (O/E/O) conversion. When applied to optical switches, the proposed end-to-end scheduling has to also account for non-negligible switching overhead and reconfiguration penalty which is inherent to all-optical switch technology. As a simple prototype, we consider a simple variant of the well-known MaxWeight scheduler whose rate of switching is optimized to account for the switching penalty.

Our proposed architecture sharply deviates from prior work on (near-)zero in-network buffering in that we avoid using the data plane for monitoring and control. In particular, we push the monitoring and control functionalities (which are critical for fine-grained dynamic circuit switching) away from data-plane into an entirely separate wireless network. In other words, our monitoring/control plane functionalities are implemented in a manner quite similar to providing cellular wireless connection to a very large number of users. In particular, we envision each ToR switch to be equipped with a wireless transceiver which allows for a central controller/scheduler unit to monitor the traffic at each ToR across the data center. Since this single-hop wireless network is to realize a monitoring/control plane across the data center, achieving ultra-low latency is of paramount importance. The additional challange here is the very high spatial density of transmitter/receivers and an acute need for interference management. As a simple prototype for the monitoring/control plane, we provide a spatially adaptive Orthogonal Frequency Division Multiple Access (OFDMA) modulation combined with single-hop mmWave radio access technology (RAT) would satisfy the latency requirements.

When the combination of the proposed monitoring technique and scheduling policies is considered, we aim to present the first steps to transform today's data centers' over-designed under-utilized networks into the highly optimized large-scale distributed and parallel computation infrastruc-

978-3-9018-8273-9/15/ ©2015 IFIP

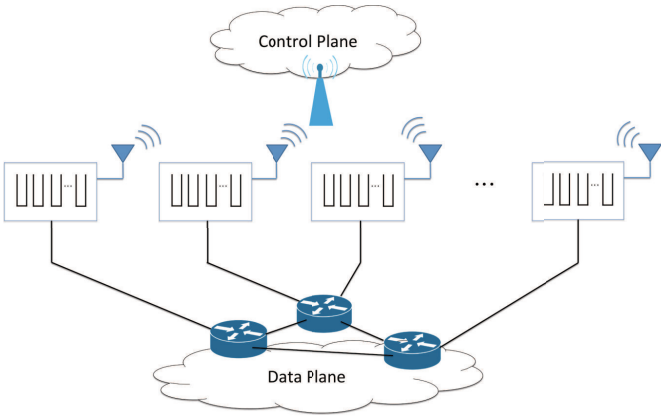

Fig. 1. Wireless control plane enabling all optical data plane

tures of tomorrow. Note that our proposed architecture is drastically different from the all-wireless data center design of [3], where the data is transmitted over wireless links which are inherently of lower capacity. Our proposed solution also differs from the wireless facility network of Angora [4]. Angora is a multi hop wireless network that provides an auxiliary network for facility bring up and installation and/or for forwarding table updates and/or reset hardware in response to electronic switch failures in the data plane.

The remaining of the paper is organized as follows. We start with basic definitions and details of the proposed architecture that covers both data and control planes in Section II. In Section III, we briefly identify the scheduling policy that is a strong candidate under non-zero reconfiguration time operation. We introduce a combination of techniques that yields a feasible monitoring plane in Section IV. This conceptual design is indeed responsible for informing the CU about the most recent state of the ToR switch queues so that the data plane is efficiently scheduled. In Section V, we present the improvements obtained by using the mentioned scheduling policy in comparison with the Traffic Matrix Scheduling (TMS) algorithm [5]. We also show that required small monitoring delays can be obtained by using the proposed monitoring prototype. Finally, Section VI concludes the paper with some possible future paths of study.

II. System Architecture: Wireless Monitoring Plane Serving an All-optical Data Plane

The proposed data center in this paper aims to allow for the flows to be optimally scheduled across the network via a clean-slate architecture and a rethinking of the protocol stack. The essential feature of this architecture is that it fully decouples the time scales associated with network monitoring and control and optical switching of data.

Modern data centers usually consist of hundreds to thousands of servers, and intensive data exchanges occur within a data center network, which is assumed to be operated by a single entity herein. Ever increasing data-rate requirements (40 Gbps, 100 Gbps, or beyond) and number of port counts have become bottlenecks for traditional electronic data switches. Optical switches have the advantages in scalability and lower power consumption. In addition, the ever decreasing switching time in optical switches (due to MEMS mirrors, etc) [5], [6] makes the boundary between circuit switching time scales and packet switching shift and blur quite a bit. We propose the data plane to be implemented using all optical switching where dynamic circuit switching maintains the operation basis. However, the optical switching comes with its own challenges to be faced. Buffering of information packets is not feasible in the optical domain. This means that utilizing optical switching in a data center requires fine-grain circuit switching, and hence, shifting the buffering to the ToR switches at the edge of the network. This in turn results in a network that is abstracted as a generalized switch with *non-zero reconfiguration and monitoring delays*. Therefore, we require an optimized circuit switching strategy that accounts for and schedules the outstanding traffic packets queued up at the edge of the network at each ToR switch with the ultimate goal of having *low delay* in the data plane. Furthermore, it will be clear in the subsequent sections that the low delay performance of the data plane switching strategy is highly sensitive not only to the reconfiguration delays but also to the monitoring delays.

The need for low-latency monitoring of the network state brings the other half of the proposed architecture into the picture: the *centralized* wireless monitoring/control plane that serves the data plane. More specifically in this paper, we argue that wireless technology, if carefully optimized across layers of the protocol stack, provides a cost effective solution for monitoring/control plane such that we establish a zero-buffer circuit switch at appropriate time scales.

Given the tight latency requirements, the wireless monitoring of a data center has unique challenges and opportunities. Considering the environment of densely packed racks in a data center with relatively short distances between communicating units, mmWave communication is a viable alternative from several perspectives. To start with, a large bandwidth, spanning several Ghz, has been allocated for unlicensed use around 60 GHz [7]. This has opened up new possibilities for realizing short-range, high-rate wireless communications [8], [9]. Moreover, the use of multiple antenna systems and beamforming is also practical for mmWaves, which may easily compensate for the very high propagation losses in this band. However, the stationarity of the racks in a data center results in little (if not zero) diversity in the time domain due to non-fading channels. Thus, in order to improve the reliability of communication, we are left with two opportunities: i. Multi-user diversity ii. Frequency diversity. The multi-user and frequency diversity resources are highly related to the physical orientation of the racks in the data center area. The ultimate challenge in such a densely packed multi-user environment is to manage these resources so that a large number of ToR switches can reliably convey their network state information to a centralized controller with low delays.

In the following sections, we make the features of the proposed data and monitoring/control planes clear and also explain the details of the decoupled operation with emphasis on performance implications of this novel architecture. An exemplary scheduling policy and a combination of wireless communication techniques are investigated for a prototype design.

III. Data Plane: Scheduling with non-zero Monitoring and Reconfiguration Delay

Our proposed architecture relies on an all-optical data plane to schedule flows between ToRs across the data center. We start with the description of the data plane system model, which is illustrated in Fig. 2. We envision a set of N Top

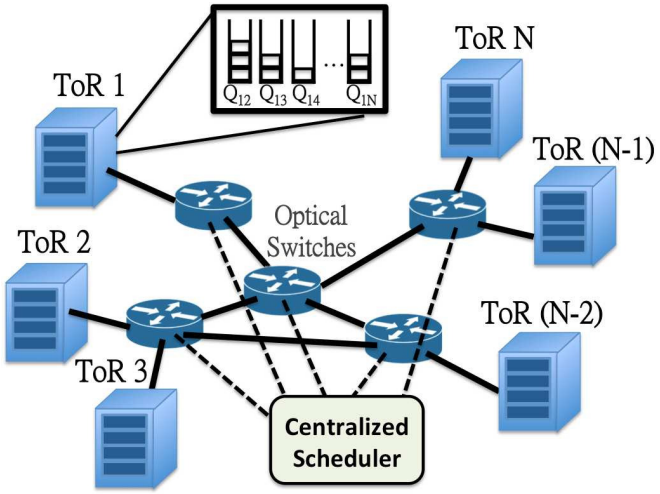

Fig. 2. An example dynamic circuit switched network of N ToRs

of Rack (ToR) switches, labelled by $\{1, 2, \ldots, N\}$, which are interconnected by an optical switched network. Each ToR switch can serve as a source and a destination simultaneously. We assume that there is no buffering in the optical network, but all the buffering is handled at the edge of the network, that is within the ToR switches. Therefore, each ToR switch maintains $N-1$ edge queues. Let us denote these queues by Q_{ij}, where $j \in \{1, 2, \ldots, N\}, i \neq j$: the packets destined from the ToR switch i to j are enqueued in the edge queue Q_{ij} before a scheduled transmission. Queues may be implemented either physically or virtually.

The system considered is assumed to be time-slotted, with the time indexed as $t \in \mathbb{N}_+ = \{0, 1, 2, \ldots\}$. Each slot duration corresponds to a transmission duration of a single data packet, which is taken to be a fixed value in this work. Let $L_{ij}(t)$ be the number of packets in the edge queue Q_{ij} at the beginning of the time slot t, and $\mathbf{L}(t) = [L_{ij}(t)]$, where $\mathbf{L}(t) \in \mathbb{N}_+^{N \times N}$.

Let $\mathbf{S}(t) \in \{0, 1\}^{N \times N}$ denote the end-to-end connectivity (also known as schedule) at time slot t, which indicates the optical circuits established between the ToR switches. Accordingly, $S_{ij}(t) = 1$ indicates that an optical circuit from ToR i to ToR j exists at time t, and $S_{ij}(t) = 0$ means that there is no connection from ToR i to ToR j. Note that $S_{ii}(t) = 0$ for all t and $i \in \{1, 2, \ldots, N\}$. We also assume at any t each ToR can only transmit to at most one destination, and can only receive from at most one source, *i.e.*, $\sum_i S_{ij}(t) \leq 1, \sum_j S_{ij}(t) \leq 1$.

Furthermore, $\mathbf{S}(t)$ should be such that for any pair of $S_{ij}(t), S_{i'j'}(t) > 0$, there exists non-blocking circuits available across the data center to transmit packets simultaneously from ToR switch i' to j'. Let $\mathfrak{S}$ be the set of all feasible schedules (respecting the bi-section bandwidth and parallel scheduling requirements).

Our proposed dynamic (fine-grained) circuit switching generalizes ideas from switch fabric scheduling to manage circuit scheduling at fairly fast and fine-grain time scales. In other words, let $\tilde{\mathbf{L}}(t) = [\tilde{L}_{ij}(t)]$ be the estimated state of ToR switches according to a centralized controller. The centralized scheduler, as a function of $\tilde{\mathbf{L}}(t)$, selects schedule $\mathbf{S}(t) \in \mathfrak{S}$ such as to minimize the latency at the edge queues. For a genie-aided centralized scheduler with no reconfiguration

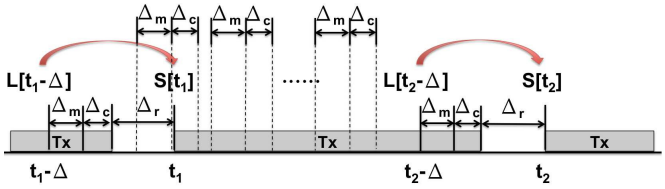

Fig. 3. Dynamic monitoring with frequently updated schedules

delay and switching penalty, $\tilde{\mathbf{L}}(t) = [\mathbf{L}(t)]$, and hence the problem becomes a simple generalization of cross-bar scheduling [10]. However, in our setting, the centralized scheduler is affected by three important challenges. Firstly, our centralized scheduler relies on the monitoring/control plane to estimate the network state information; this introduces delay and error in $\tilde{\mathbf{L}}(t)$ relative to $\mathbf{L}(t)$. Secondly, the centralized schedule has to compute and distribute the desired new schedule across the data center over the control plane. And, finally, every new schedule requires a collective reconfiguration of optical switches across the network, resulting in non-negligible down time in the operation of the network and flow of data packets. In this work, we map these challenges into the following delays:

- Δ_m is the monitoring delay for the CU to monitor edge queues.
- Δ_c is the computation delay for the CU to compute a schedule.
- Δ_r is the reconfiguration delay in activating a new schedule across the network and resumption of end-to-end transmission.

The above abstraction allows us to map the practical challenges in form of a delay $\Delta_m + \Delta_c + \Delta_r$; i.e., $\mathbf{S}(t)$ is only selected based on $\mathbf{L}(t - \Delta_m - \Delta_c)$ and will take effect at time $t + \Delta_r$. Fig. 3 illustrates this abstraction. Note that with a careful design of monitoring/control plane, as we will see in Section IV, we can ensure high reliability in the estimated values and low noise across any network, so long as we allow Δ_m to be sufficiently large. Furthermore, we assume CU has sufficient computational resource to ensure $\Delta_c \simeq 0$.

The problem of low complexity dynamic circuit switching with non-negligible monitoring and reconfiguration delays constitutes a topic of extensive research, building on prior work on cross-bar switch scheduling [10]. However, in this paper, we focus our attention on a proof of concept regarding the architecture. In particular, we limit our attention to a fairly straight-forward generalization of Maximum Weighted Scheduling, namely the Periodic MaxWeight (PMW) algorithm. Following the definitions from the Maximum Weighted Scheduling, the weight of a schedule $\mathbf{S}$ is defined as the sum of all the edge queues served by $\mathbf{S}$, that is $w = \sum_i \sum_j \mathbf{L}_{ij} \mathbf{S}_{ij}$. The PMW algorithm selects a time duration T and reconfigures the schedule to the maximum weighted schedule at time $t_k = kT$, $k = 0, 1, \ldots$, hence the schedules are given by

$$\mathbf{S}(t_k) = \arg\max_{\mathbf{S} \in \mathfrak{S}} \sum_i \sum_j \mathbf{L}_{ij}(t_k) \mathbf{S}_{ij}$$

The CU sets the parameter T so that the effective duty cycle $D = 1 - \frac{\Delta_r}{T}$ is larger than the traffic arrival rate to ensure finite expected queue lengths. Therefore, given the traffic arrival rate, we can determine a lower bound on T and to

optimize the delay performance by selecting an appropriate T value.

IV. Monitoring/Control Plane: Spatially Adaptive mmWave MIMO OFDMA

Our proposed architecture relies on a central unit (CU) to dynamically schedule flows between ToRs over an all-optical data plane across the data center (fine-grained circuit switching). More specifically, the CU requires monitoring the current instantaneous traffic demands across the data center (monitoring), calculating efficient schedules for packet transmissions, and making the resulting schedules available at the ToRs as well as the optical switches (control). Furthermore, in Section III, we saw that the performance of the CU critically depends on the latency of monitoring. In this section, we propose a single-hop wireless network design to implement the communication link to and from the CU (monitoring/control plane)[1].

While the proposed wireless network is the bridge between the ToRs and the CU for both monitoring and control functionalities, we focus on the monitoring objective. This is because distributing schedules (control) across the network can be achieved with relatively low rate (broadcasting a sparse set of end-end flow connectivities). In contrast, the monitoring plane is required to achieve low-latency and high reliability communication for hundreds of ToR switches densely packed in a small area. For this reason, we focus our attention on the design of monitoring plane (uplink).

A. Control Plane Messages: (Differential) Queue States

Since our focus is the network state monitoring functionality, we have to design the control and monitoring messages, to ensure that the CU has a low latency update regarding the backlog information across the network, *i.e.*$\{L_{ij}(t)\}_{i,j}$. In other words, each ToR i is responsible to update the CU on the amount of traffic it has for all other ToRs, *i.e.*the value of $L_{ij}(t)$, for all $j \neq i$. However, it is expected that the queue backlogs at time $t-1$ and t are highly correlated. We utilize this temporal correlation of size of a queue and design our monitoring messages to be that of differential queue occupancy information (instead of the exact queue sizes). At the same time since each ToR has the same number of edge queues, the message size is designed to be fixed across all ToRs and limited to $(N-1)b$ bits. In other words, the differential information $L_{ij}(t) - L_{ij}(t-1)$ is quantized into b bits that are sufficient to reconstruct the exact information at the CU if the monitoring plane is reliable.

Particularly, in case the monitoring frequency is high enough, the interval between two monitoring phases would be small so that differences in the queue sizes would also be represented by a very small number of bits. Once this message rate and the desired reliability of message transmission (usually in terms of bit-error-rate (BER)) are fixed over the network of ToRs, the ultimate goal in the design of the monitoring algorithms is to manage the resources spatially and minimize the monitoring delay. In other words, by taking the data center layout into consideration, we need to make use of the degrees of the freedom corresponding to each ToR's unique location in the data center in order to keep message transmission for all ToRs at a minimum.

[1]Such a wireless data-center-wide monitoring plane is expected to improve the throughput in both optical [5] and electrical [2] data plane implementations, although in this work we work on an all-optical data plane.

B. Wireless Radio Access: Enabling Technologies

We describe the monitoring operation as the CU collecting network state information from N ToR switches. Since each ToR's state information is composed of $(N-1)b$ bits, a total of $N(N-1)b$ bits are required to be received at the CU in a monitoring duration. Our ultimate goal is to maintain a frequent monitoring over a separate wireless medium. In this section, we identify the technologies that will enable us to reach this goal in an environment that is specific to data centers. In particular, the major challenge will be managing the aggregate rate for large data centers with potentially hundreds of ToRs.

In order to manage the rate requirement, we propose to use mmWave transmissions around 60 GHz for the radio access between a ToR and the CU. The mmWave-band communication has advantages in short distance communications: small channel delay spreads due to high path loss, large and unlicensed transmission bandwidth, and potential applications of massive Multiple Input Multiple Output (MIMO) antenna systems and beamforming. Although the propagation and the atmospheric losses are immense in the mmWave channel, use of narrow beams is a common method to solve the problem of low average received SNR values.

When combined with beamforming, the mmWave communication results in relatively small channel delay spreads; however, still a multipath propagation problem might arise in a dense scatterer environment like the one in a data center. In order to counteract the resulting intersymbol interference (ISI) problem, the digital modulation scheme is selected as a spatially adaptive version of Orthogonal Frequency Division Multiple Access (OFDMA) [11]. In addition to ISI mitigation, with an OFDM-type transmission, we have the opportunity to assign preferred subcarriers to users in a multi-user scenario. Moreover, considering the large number of ToRs communicating simultaneously, the simple receiver structure for OFDMA demodulation has a computational complexity advantage. On the other hand, one critical consideration with OFDM-type transmissions is the possible spectral efficiency loss due to usage of the cyclic prefix (CP). Therefore, we need to keep the number of subcarriers K large enough to make the duration of an OFDMA symbol much longer than that of CP.

In our proposed architecture, the latency of monitoring must be low enough to achieve efficient schedules. On the other hand, we also need to utilize channel codes for improving the end-to-end reliability. As a result, we are limited to channel codes of short blocklength with relatively higher rates. In this work, the channel code is selected as an irregular low-density parity-check (LDPC) code, since it has a BER performance close to Shannon capacity [12]. The parameters of the mentioned LDPC code are given in Section V-B.

C. Data Center Layout: Wireless Channel

The locations of the ToRs and other equipments in a data center are static. Owing to this static nature of the transmission medium, the impulse response function of the channel from each ToR to the CU can be estimated by the CU with high precision. In modelling the mmWave channels proposed to be used in this work, we rely on the empirical results presented in Table.I of [13]. In particular, we model the discrete-time equivalent wireless channel as an L-tap static frequency selective channel, denoted by $h_i[n]$ which

has exactly L taps (non-zero entries), h_{il}, $l \in \{1, 2, \ldots, L\}$. Since directed transmit beams are assumed, the first tap h_{i1} always represents the line-of-sight (LOS) path, while the remaining taps h_{il}, $l \in \{2, 3, \ldots, L\}$ corresponds to the non-LOS (NLOS) reflected paths. When we consider the fact that each ToR observes multipath taps at different delays and magnitudes, the resulting frequency selectivity pattern for that ToR becomes unique. This static and heterogeneous frequency selectivity across the ToRs is opportunistically utilized in the subcarrier allocation algorithm presented in Section IV-D.

Moreover, the Signal-to-Noise Power Ratio (SNR) corresponding to a ToR can be modelled as a constant value in time. This constant is a function of that ToR's distance to the CU and is found as the total power observed in multipath taps. Thus, the SNR value in dB scale for ToR i is safely approximated as

$$\text{SNR}_i = \text{SNR}_0 - 20\log\left(\frac{d_i}{d_0}\right) - 20\log\left(\frac{\beta_i}{\beta_0}\right), \quad (1)$$

where d_i is the distance of ToR i to the CU, β_i is the 3 dB beamwidth of the transmitter antenna pattern of the same ToR in degrees. In (1), d_0 and β_0 are the reference distance and 3 dB beamwidth values for which the received SNR value is SNR_0. The exact values of SNR_0, d_0, and β_0 utilized in the simulations are given in Section V.

D. Macro-level Multi-user Resource Allocation

A transmission mechanism that is adapted to the heterogeneity of the network of many ToRs is a key point in achieving reliable and low-latency communication for monitoring purposes. In that sense, we should allocate the OFDMA subcarriers to the ToRs carefully so that we can utilize the inherent frequency diversity that is described in Section IV-C. Other than frequency adaptivity, huge variation of the received SNR values (due to difference in the distances of the ToRs to the CU) should also be taken into consideration. Consequently, an adaptation of the modulation size and/or channel coding rate is required for reliable transmission of all ToRs. Assuming that we have in total M_{OFDM} symbols to transmit, the ideal management of the resources in the monitoring plane would require a joint optimization of Gaussian channel capacity for N ToRs by using a joint power, data rate and subcarrier allocation algorithm [14]. Then, one would maximize the following rate function under usual power constraints for transmitting ToRs.

$$R(\{c_{i,k,m}, p_{i,k,m}\}) = \sum_{i=1}^{N}\sum_{m=1}^{M_{\text{OFDM}}}\sum_{k=0}^{K-1} c_{i,k,m} log\left(1 + \frac{\gamma_i p_{i,k,m}|g_i[k]|^2}{N_0}\right), \quad (2)$$

where $g_i[k]$ is the known discrete time channel gain for ToR i at the subcarrier $k \in \{0, 1, \ldots, K-1\}$. In other words, $g_i[k]$ is the K-point Discrete Fourier Transform (DFT) of the channel impulse response $h_i[n]$ given in Section IV-C. In (2), $c_{i,k,m}$ denotes the indicator variable showing the allocation of subcarrier k to ToR i at the mth OFDMA symbol, i.e., $c_{i,k,m} \in \{0, 1\}$ with orthogonal transmission constraint $\sum_{i=1}^{N} c_{i,k,m} \leq 1$ for all (k, m) pairs. Furthermore, $p_{i,k,m}$ is the corresponding transmit power of ToR i at subcarrier k during the mth OFDMA symbol and N_0 denotes the variance of the zero mean circularly symmetric complex Gaussian noise signal at the CU. Also, $\gamma_i \triangleq (\beta_0/\beta_i)^2$ is the beamforming gain factor for ToR i.

In order to keep the investigation simple and a possible implementation efficient, we firstly assume that all ToRs transmit their message symbols at a constant beamformed power such that $\gamma_i p_{i,k,m} = P$. Moreover, instead of directly maximizing the total rate in (2) via fine-grained multi-user adaptation, we follow a macro-level resource allocation that includes two disjoint steps which are presented next.

1) Distance-based Rate Assignment: The first phase of spatial adaptation considers only the distance dependent SNR values of the ToRs in order to assign sensible modulation orders. Considering a fixed channel code rate r, a coded data rate of $rM_i < C_i$ is assigned to ToR i, where $C_i = \log_2(1 + \text{SNR}_i)$ is the AWGN channel capacity and 2^{M_i} is the constellation size. In order to simplify the demodulation at the CU, in this work, we assumed that only square Quadrature Amplitude Modulation (QAM) type constellations with cardinalities upto 1024 are assigned to ToRs by employing the following expression for calculating the constellation size.

$$2^{M_i} = \min\left\{1024, 2^{2\text{round}\left(\frac{\alpha C_i}{2r}\right)}\right\}, \quad (3)$$

where round(.) is the usual rounding function that maps its argument to the closest integer, and $\alpha < 1$ represents the achievable fraction of the capacity by using a channel code for any ToR. This fraction is exactly the *normalized rate* value in Fig. 15 of [15] and can be seen as a constant that defines the gap to the theoretical rate limit for channel codes. Since we are required to complete monitoring in extremely short durations, we are restricted to employ very short blocklength codes in this work. As a consequence, this necessitates the selection of relatively low α values in (3). The selected channel codes and the α value are given in Section V-B.

2) Greedy Frequency-time Resource Allocation for Low-magnitude Subcarrier Avoidance: After the modulation order and channel code rates are fixed for every ToR, the requested number of frequency-time resources are calculated in the second phase of adaptation. Considering ToR i, the number of requested resources is given by the number of symbols it should transfer to the CU and easily calculated as $\lceil (N-1)b/(rM_i) \rceil$, where $\lceil x \rceil$ is the smallest integer larger than or equal to x. Correspondingly, the total number of OFDMA symbols to be received by the CU is at least

$$M_{\text{OFDM}} = \left\lceil \frac{\sum_{i=1}^{N}(N-1)b/(rM_i)}{K} \right\rceil. \quad (4)$$

Hence we have KM_{OFDM} many frequency-time slots (resources) in a single monitoring round. The ToRs are then assigned these resources according to the greedy algorithm with the following steps:

- Sort all the ToRs according to the increasing received SNR value. Initiate the unassigned resources set to $\{1, 2, \ldots, KM_{\text{OFDM}}\}$.
- Starting with ToR i that has the weakest received SNR value, calculate its channel frequency response. By using this unique response, allocate $\lceil (N-1)b/(rM_i) \rceil$ many strongest frequency-time resources. More rigorously, the set of pairs assigned

to ToR i is defined as

$$T_i \triangleq \{(k,m) : \text{Resource } (k,m) \text{ is one of the } \lceil (N-1)b/(rM_i) \rceil \text{ strongest subbands in the unassigned set of resources}\}$$

Update unassigned set of resources by eliminating these newly assigned resources in it.

– Continue with the next weakest-SNR ToR following the rules in the previous step until all ToRs are assigned required number of resources.

As an example, in Fig. 4, we observe power levels in discrete frequency domain for 2 ToRs and 100 subcarriers. If we assume that 3 resources are required by two ToRs and ToR-1 has lower SNR than ToR-2, then the proposed algorithm would initially assign 3 most powerful subcarriers (numbered 1, 22, and 96) to ToR-1 as indicated by circles at these subcarriers. Then, for ToR-2, since the most strongest subcarrier is for $k = 1$ and it is already allocated, the algorithm will simply allocate the most powerful subcarriers among the remaining ones (numbered 44, 63, and 87) as shown by squares.

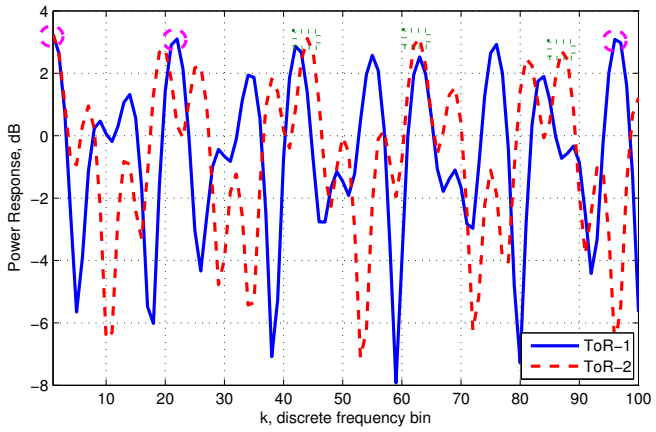

Fig. 4. Sample greedy allocation of subcarriers to ToRs 1 and 2

One final note is that many optimal and suboptimal techniques were developed for frequency subband assignment in uplink channels of OFDMA systems (see [14] and references therein). However, the main purpose of the proposed frequency-time resource assignment algorithm is to avoid the case that a ToR observes continuously the very same *bad* (low magnitude or even a frequency null) frequency subbands due to the static nature of the channel, rather than improving the rate with respect to the frequency non-selective case.

V. Simulation Results

A. Scheduling with Reconfiguration Delay and Effect of Monitoring Delay

In this section we provide the simulation results for the Periodic MaxWeight (PMW) [16] to be utilized in the data plane prototype and its comparison to a benchmark TMS [5] algorithm. The results presented here focuses on the mean queue lengths observed in the data center. Therefore, the average delays for a packet in the network can be deduced from these results simply by invoking Little's law.

To obtain Fig. 5, the simulations are conducted with the simulator built for the REACToR switch in [5]. The reconfiguration delay is taken to be $\Delta_r = 20\ \mu s$. In order to compare scheduling algorithms applied on the optical switches, we cease the electronic switches in the hybrid switch design in [5] and only utilize the optical switches. The number of ToR switches is $N = 100$ and the average traffic load of the network is 30% of the total throughput with uniform traffic pattern. The link data bandwidth at each host is $B = 100$ Gbps, and the packets are of the same size $p = 1500$ bytes (each takes $0.12\mu s$ for transmission). Each queue can store up to 1.67×10^5 packets, and incoming packets are discarded when the queue is full. For simplicity, the network topology is assumed to be rearrangeably nonblocking. We assume the arrival processes at each ToR queue Q_{ij} to be independent over $i, j \in \{1, 2, \ldots, N\}, i \neq j$. Each arrival process is i.i.d. and is Poisson distributed over time slots. The traffic is assumed to be admissible, i.e. the total arrival rate for packets to be sent from each ToR and to be received by each ToR are all less than 1.

In Fig. 5, the TMS monitors the queue lengths every $1500\ \mu s$ and determines 7 schedules to be used in the next $1500\ \mu s$. The PMW with $T = 200\ \mu s$ has comparable rate of schedule reconfigurations with the TMS, and outperforms the TMS under small monitoring delay. Under the constraint of duty cycle being larger than the traffic load, we set $T = 30\ \mu s$ to optimize the performance of the PMW algorithm. The performance improvement is substantial with difference nearly an order of magnitude when the monitoring delay is small. However, it is also shown that the monitoring delay is critical in this performance improvement.

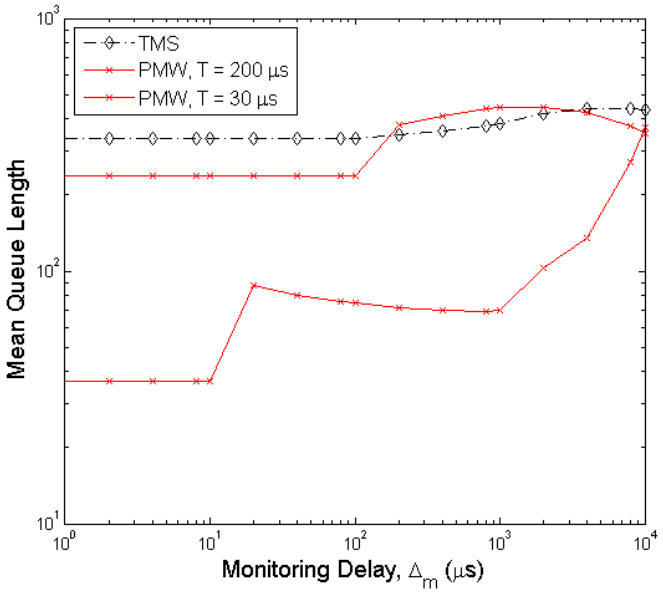

Fig. 5. Effect of monitoring delay on mean queue length

B. Wireless Monitoring Plane Performance Results

In the simulations, in order to determine the reference received signal level SNR_0, we make use of the Received Signal Strength (RSS) measurements for off-the shelf mmWave transcievers detailed in [3]. When we assume a reference transmit power of 10 dBm as in [17], $\text{SNR}_0 = 30$ dB is obtained at the reference distance of $d_0 = 1$ meter when a directed beam of reference 3-dB beamwidth equal to 30 degrees is used. Moreover, we make use of the static frequency selective channel model that varies across ToRs

given in Section IV-C. In order to simplify the simulations, we take the strongest two NLOS paths from Table.I of [13] in addition to the LOS path. For modelling the heterogeneity of the frequency selective channels across the ToRs, we generate two NLOS channel taps for each ToR according to the following methodology. For a given ToR i, there exist 3 discrete channel taps h_{ij}. The LOS tap is h_{i1} and its power level is taken as 0 dB reference level and its delay τ_{i1} is always 0. The NLOS tap h_{i2} is assumed to have relative (with respect to LOS tap) power equal to -20 dB and its delay τ_{i2} is a random variable uniformly distributed in $[5, 15]$ ns. Similarly the relative power of the other NLOS tap h_{i3} is fixed to -25 dB and its delay τ_{i3} with respect to the LOS tap is uniform in $[15, 25]$ ns.

We assume a $K = 1024$ subcarrier OFDMA transmission for improving the spectral efficiency over a channel bandwidth of 7 GHz. Then an OFDMA symbol duration is close to 146 ns. In comparison, the CP duration is selected as 20 ns, since the multipath Root Mean Squared (RMS) delay spread for the mmWave channels are typically on this order [18], [19].

The data center, which we investigate firstly, consists of $N = 100$ ToRs that are evenly distributed on a 10m by 10m area in the center of which the CU is located. The first thing to decide in such a network would be the modulation order assignment to the ToRs. Here, we assign the modulation orders to ToRs according to the expression given in (4). We select the normalized rate parameter as $\alpha = 0.62$. This selection is based on trial-and-error according to the achieved BER less than 10^{-5} for an irregular LDPC code of parameters (603, 301) [12] with the proposed greedy resource allocation algorithm. Following this procedure, it is found that 40 of the ToRs that are far away from the center are assigned 64-QAM, whereas the nearby 24 ToRs to the CU are given the chance to transmit using 1024-QAM. The remaining 36 ToRs are allowed to convey their data over 256-QAM constellation. We can approximate the monitoring delay for these assignments as follows. Assuming that each differential ToR queue occupancy information is composed of $b = 3$ bits and a code rate of $r = 302/603$ is utilized, a total of 101 64-QAM symbols has to be transmitted by a *low-SNR* ToR; whereas 76 256-QAMsymbols are to be conveyed by a *mid-SNR* ToR and only 61 frequency-time slots are consumed by a *high-SNR* ToR that uses 1024-QAM for modulation. Overall number of symbols to be transmitted is then equal to 8240 which corresponds to 9 OFDMA symbols even if some subbands are left empty out of $K = 1024$. Therefore, in such a large system, the proposed monitoring technique takes only $9(146ns+20ns) \sim 1.5\mu s$ to complete. This monitoring delay is favourable when we consider the mean queue length performance figures of the proposed scheduling policies given in Section V-A.

The second example data center we consider consists of $N = 900$ ToRs that are also distributed evenly on a 30m by 30m area. For this example, there are some ToRs with very low SNR values and so very low spectral efficiencies. The minimum spectral efficiency is calculated to be 1.76 bits/sec/Hz. Even if all the ToRs can efficiently transmit at their limiting spectral efficiencies, the monitoring delay can be shown to be close to $230\mu s$, which can affect the overall mean queue length (also packet delay) performance of the data center adversely. The spectral efficiency plot for this scenario is given in Fig. 7.

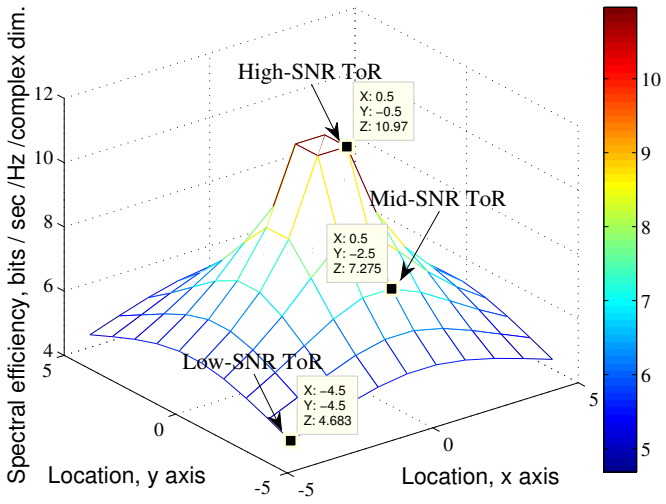

Fig. 6. Spectral efficiency plot for 10m by 10m data center (100 ToRs)

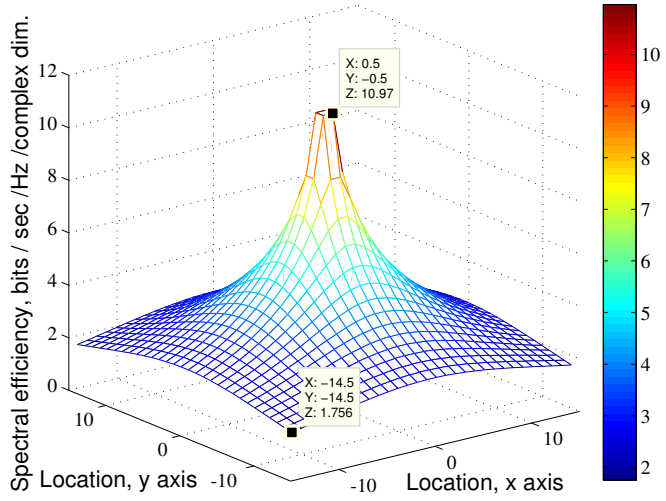

Fig. 7. Spectral efficiency plot for 30m by 30m data center (900 ToRs)

One way to improve the received average SNR values in the monitoring system would be utilizing directional transmissions by shaping very narrow beams at the ToR side. As an example, if we decrease the 3 dB beamwidth to 10 degrees in both elevation and azimuth directions, the total monitoring delay may be decreased to $113\mu s$, which is still high when compared to the values deduced from Fig. 5. The monitoring delays for two different beamwidth values are given with respect to the increasing number of ToRs in the data center in Fig. 8. We note that decreasing beamwidth three fold is equivalent to increasing transmit power by almost 9.5 dB. Clearly, in order to keep the monitoring delay below the $40\mu s$ limit, we need some other techniques if the data center size is larger than 550 and 450 for beamwidths of 10 and 30 degrees respectively.

To further reduce the monitoring delays in extremely large data centers, a possible approach would be supporting the CU with more than one directional antennas and related receiver chains. In this way, we can divide the overall area around the CU receivers into segments, in each of which, through parallel transmissions, many ToRs can simultaneously transmit using the same subcarriers with acceptable BER performance over the whole network.

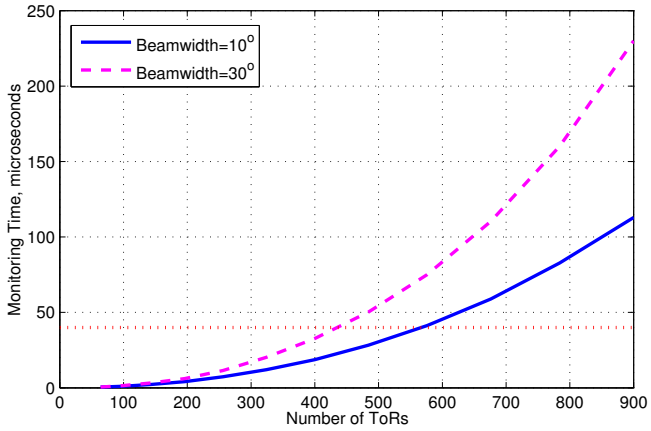

Fig. 8. Monitoring delay with respect to the number of ToRs

VI. Conclusion

A new hybrid wireless/optical architecture is introduced for realizing the control and the data plane functions in a very large data center. The decoupling of the control and the data planes is an essential part of the proposed architecture. Feasibility of the proposed all-optical data plane is demonstrated under non-zero reconfiguration and monitoring delays by utilizing a max-weight type policy that calculates efficient schedules only when the control plane supplies it with up-to-date network state information. Therefore, in order to improve the monitoring frequency for the network state, we presented a mmWave radio access technology with OFDMA signalling. The technique is adaptive based on the ToR-CU distances and the corresponding static but frequency selective channel responses so that each ToR is assigned a set of *good* subbands and an appropriate modulation scheme. The results presented for relatively large data centers prove the efficiency of the monitoring technique, which in turn satisfies the requirements of the scheduling policy for low mean-queue-length. The joint source-channel encoding of the queue occupancy information by making use of spatial correlations in addition to the temporal correlations in the network is a possible track for further reducing monitoring times. Another topic of future interest would be hierarchical and/or segmented monitoring of ToR switches for extremely large data centers.

Acknowledgment

This work has been partially supported by L-3 Communications and NSF Center for Integrated Access Networks (Grant EEC-0812072).

References

[1] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, and S. Shenker, "pfabric: Minimal near-optimal datacenter transport," in *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, SIGCOMM '13, (New York, NY, USA), pp. 435–446, ACM, 2013.

[2] J. Perry, A. Ousterhout, H. Balakrishnan, D. Shah, and H. Fugal, "Fastpass: A centralized "zero-queue" datacenter network," in *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, (New York, NY, USA), pp. 307–318, ACM, 2014.

[3] J.-Y. Shin, E. G. Sirer, H. Weatherspoon, and D. Kirovski, "On the feasibility of completely wireless datacenters," in *Proceedings of the Eighth ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, ANCS '12, pp. 3–14, ACM, 2012.

[4] Y. Zhu, X. Zhou, Z. Zhang, L. Zhou, A. Vahdat, B. Y. Zhao, and H. Zheng, "Cutting the cord: A robust wireless facilities network for data centers," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, MobiCom '14, (New York, NY, USA), pp. 581–592, ACM, 2014.

[5] H. Liu, F. Lu, A. Forencich, R. Kapoor, M. Tewari, G. M. Voelker, G. Papen, A. C. Snoeren, and G. Porter, "Circuit switching under the radar with reactor," in *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, NSDI'14, (Berkeley, CA, USA), pp. 1–15, USENIX Association, 2014.

[6] 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 *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, SIGCOMM '13, (New York, NY, USA), pp. 447–458, ACM, 2013.

[7] FCC Report, "Amendment of parts 2, 15 and 97 of the commissions rules to permit use of radio frequencies above 40 ghz for new radio applications," tech. rep., December 1995.

[8] P. Smulders, "Exploiting the 60 ghz band for local wireless multimedia access: prospects and future directions," *Communications Magazine, IEEE*, vol. 40, pp. 140–147, Jan 2002.

[9] R. Daniels and R. Heath, "60 ghz wireless communications: emerging requirements and design recommendations," *Vehicular Technology Magazine, IEEE*, vol. 2, pp. 41–50, Sept 2007.

[10] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE*, vol. 1, pp. 296–302 vol.1, Mar 1996.

[11] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser ofdm with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Sel Areas in Commun.,*, vol. 17, pp. 1747–1758, Oct 1999.

[12] A. Ramamoorthy and R. Wesel, "Construction of short block length irregular low-density parity-check codes," in *Communications, 2004 IEEE International Conference on*, vol. 1, pp. 410–414, June 2004.

[13] C. Gustafson, F. Tufvesson, S. Wyne, K. Haneda, and A. Molisch, "Directional analysis of measured 60 ghz indoor radio channels using sage," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pp. 1–5, May 2011.

[14] K. Kim, Y. Han, and S.-L. Kim, "Joint subcarrier and power allocation in uplink ofdma systems," *Communications Letters, IEEE*, vol. 9, pp. 526–528, Jun 2005.

[15] Y. Polyanskiy, H. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *Information Theory, IEEE Transactions on*, vol. 56, pp. 2307–2359, May 2010.

[16] C.-H. Wang, T. Javidi, and G. Porter, "End-to-end scheduling for all-optical data centers," in *INFOCOM, 2015 Proceedings IEEE*, April 2015.

[17] X. Zhou, Z. Zhang, Y. Zhu, Y. Li, S. Kumar, A. Vahdat, B. Y. Zhao, and H. Zheng, "Mirror mirror on the ceiling: Flexible wireless links for data centers," in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '12, (New York, NY, USA), pp. 443–454, ACM, 2012.

[18] T. Rappaport, E. Ben-Dor, J. Murdock, and Y. Qiao, "38 ghz and 60 ghz angle-dependent propagation for cellular amp; peer-to-peer wireless communications," in *Communications (ICC), 2012 IEEE International Conference on*, pp. 4568–4573, June 2012.

[19] H. Roufarshbaf, U. Madhow, and S. Rajagopal, "Ofdm-based analog multiband: a scalable design for indoor mm-wave wireless communication," in *Global Communications Conference, 2014 IEEE*.