Towards 48-Fold Cabling Complexity Reduction in Large Flattened Butterfly Networks

Márton Csernai*, Florin Ciucu†, Ralf-Peter Braun‡ and András Gulyás§
*MTA-BME Future Internet RG †University of Warwick ‡Deutsche Telekom §MTA-BME Information System RG

***Abstract*—Amongst data center structures, flattened butterfly (FBFly) networks have been shown to outperform their common counterparts such as fat-trees in terms of energy proportionality and cost efficiency. This efficiency is achieved by using less networking equipment (switches, ports, cables) at the expense of increased control plane complexity. In this paper we show that cabling complexity can be further reduced by an order of magnitude, by reconfiguring the optical fully meshed components into optical "pseudo"-fully meshed components. Following established methods, optical star networks are obtained by exchanging the FBFly's regular (grey) optical transceivers for dense wavelength division multiplexing (DWDM or colored) optical transceivers and placing an arrayed waveguide grating router (AWGR) in the center. Depending on the data center configuration and equipment prices, our colored FBFly (C-FBFly) proposal yields lower capital expenditure than the original FBFly. The key advantage of our structural modification of FBFly, however, is that in large FBFly networks (e.g., > 50K nodes) it reduces the number of inter-rack cables by a factor as large as 48.**

I. Introduction

The management of cabling in data centers (DCs) is a complex task [1]. Large scale data centers can have hundreds of kilometers of cables deployed, among which long optical fibers can span more than hundred meters. Appropriate cabling management can minimize downtime, maximize space use, and reduce operational costs [2]. There are several ways to measure cabling complexity. One metric is the *data center entropy*, which is defined as the gradual increase of cabling complexity over time due to moves, additions, and changes in data center connectivity; along this process, cable trays can become overcrowded such that cables can take on tight bends, even resulting in pulled connectors [2]. A more explicit quantifiable measure of cabling complexity is the *number* of long inter-rack cables required to build a data center network [3]. A distinction is drawn between long inter-rack cables and short intra-rack cables or cables between adjacent racks, since short cables are easier to install and maintain, whereas long inter-rack cables require careful planning and routing through the physical structure in cable trays over the racks. A further disadvantage of long cables is that they are more difficult to test and replace after link failures, which increases the operational costs (OpEx) of the network.

When designing data center structures, reconciling their performance efficiency with cabling management has been challenging. The Folded Clos structure [4], also known as fat-tree or the leaf/spine architecture, was recently proposed as a high performance data center architecture [5]. It is currently the state of the art structure in new data center deployments, and is being advocated by data center networking vendors [6], [7]. Fat trees offer high performance by offering multiple redundant paths [3], [8]. Furthermore, they provide evenly distributed network capacity without enticing constraints on job placement in the network [9]. This efficiency is achieved, however, at high energy consumption per bandwidth, high total equipment cost, and significant increase of cabling complexity [3], [10].

A practical alternative to fat-trees is the flattened butterfly (FBFly) structure made possible by current high radix switches [8]. It maintains similar performance at lower cost by using fewer networking equipment and more complex (i.e., adaptive) routing [8], [10]. Since the dominant factor in the FBFly cost is determined by the cost of long inter-rack cables, there are proposals for architectural modifications of its structure to reduce the cabling complexity. This is achieved by trading off the high number of long inter-rack cables for further increased control plane complexity [11], [12].

In this paper we show that the cabling complexity in FBFly structures can be reduced by an order of magnitude *without* trading off an increase in the control plane complexity. The key idea is the transformation of the $k\,(k-1)/2$ long inter-rack cables, for all full meshes, within each dimension of the k-ary n-flat FBFly topology, into a "pseudo"-full mesh consisting of just k shorter cables. Concretely, this transformation is achieved by replacing the grey transceivers in the switches with dense wavelength division multiplexing (DWDM, or colored) transceivers. Furthermore, an optical arrayed waveguide grating router (AWGR [13]) is connected to all colored transceivers through a layer of multiplexers and demultiplexers on each end. The resulting structure is an optical star network with the AWGR in the center. The AWGR implements a logical full mesh on a star topology by resolving the contention of signals in the wavelength domain. From the switches point of view, this star topology is logically a fully meshed network, which we refer to as a "pseudo"-full mesh structure, whereas the resulting complete topology is referred to as a colored-FBFly or simply C-FBFly.

To quantify the cabling complexity reduction and cost efficiency of C-FBFly relative to the original FBFly, we use FBFly's floor space and cabling models [8]. We analytically show that C-FBFly achieves up to a 48-fold cabling complexity reduction compared to FBFly; the limit is (almost) achieved when $k \to 96$ (i.e., a very large DC of > 800K

servers). We point out that this cabling complexity reduction (in the number of long inter-rack cables) is achieved without increasing the total cable length, but on the contrary. These operational benefits are obtained at the expense of higher costs for colored interfaces (relative to the replaced grey interfaces) and additional purchases of optical equipments (AWGRs, multiplexers, and demultiplexers). By accounting for both the (cabling) cost reduction and additional expenses imposed by C-FBFly, we show that the overall capital optical cabling expenditures in large DCs can be smaller than in the original FBFly.

The rest of the paper is structured as follows. Sec. II overviews recent optical data center proposals and identifies our design choices. Sec. III describes the proposed pseudo-fully meshed DC structure (i.e., C-FBFly). Sec. IV discusses the floor space model from [8] and an estimation method for the average inter-rack cable lengths. Sec. V defines the optical cabling and the cost model for both FBFly and C-FBFly. The cabling complexity reduction of C-FBFly and cost efficiency numerical results are presented in Sec. VI. Sec. VII discusses our findings. Finally, Sec. VIII concludes the paper.

II. Related Work

Data center owners face continuous capacity growth and are forced to find new methods to reduce costs at the same time. These trends result in incentives to both reduce the number of switching equipment (switches, cables) and to reduce overall power consumption. To alleviate these problems, there have been numerous proposals using optical switching devices in DCs [29], [30], [31], [32]. All of these proposals under-provision the capacity of the optical network to achieve cost efficiency and dynamically route flows in the optical domain with complex optical control planes. c-through [29] and Helios [30] use MEMS (Microelectromechanical System) optical switches, and an optical control framework provisions optical circuits for flows in the system by positioning small mirrors in the MEMS device. DOS [31] and Petabit [32] both employ AWGRs and tunable wavelength converters (TWCs) to route packets to the destinations through the AWGR by dynamically changing the signals' wavelengths with TWCs. A drawback of these optical DC proposals is that they employ widely available optical switching devices, which switch in the order of milliseconds. However, micro-flow switching is prevalent in data center traffic patterns, i.e., microsecond flows are problematic to be handled with these available optical switching methods or current solutions do not scale well [26], [27], [28].

Contrarily to the already proposed optical switching concepts for DCs, C-FBFly allocates sufficient optical circuits in the wavelength domain for arbitrary traffic patterns, without introducing a complex optical control plane; switching in C-FBFly is solely done by the electrical switches. Since C-FBFly employs only passive optical parts, it does not place additional latency on the interconnect. Table I summarizes the properties of C-FBFly compared to current optical DC proposals in terms of optical switching techniques, control frameworks, and the support for micro-flow switching.

TABLE I
Comparison of Optical Data Centers

Architecture	Optical Switch	Controlled Optical Device	Micro-flow switching
c-through [29]	MEMS	Mirror	no
Helios [30]	MEMS	Mirror	no
DOS [31]	AWGR	TWC	no
Petabit [32]	AWGR	TWC	no
C-FBFly	AWGR	none	yes

The core idea of C-FBFly, that is to replace mesh cabling by switched star topologies, is widely known in networking. A multi-stage WDM/AWGR technology has been proposed for internet router architecture to scale up capacity to 100 Tbps and 640 ports [25]. A similar method has been proposed for barrier synchronization framework in parallel computers [33]. It achieves all-to-all connectivity by equipping the CPUs with DWDM capable optical transceivers and connecting them together through a multi-stage AWGR network while it eliminates the need for both electrical switches and tunable wavelength converters. We elaborate on these earlier and recent works of AWGR based star topologies. Our design choices for an optical data center can be summarized as the following:

- The system should eliminate the need to dynamically control optical devices.
- The system should employ commercially available electrical switches and optical equipment.
- The system should support data center specific performance requirements.
- The system should minimize the amount of switching devices and cabling.

In the next section we present our proposed modification of the FBFly [8] architecture that abides these above mentioned design goals.

III. Colored Flattened Butterfly

In this section we first summarize the main aspects of the original flattened butterfly (FBFly) topology. Then we compute the number of underlying full meshes in FBFly, which is needed to define the proposed C-FBFly thereafter. Finally, we argue on C-FBFly's feasibility by addressing inherent practical considerations.

A. Full Meshes in Flattened Butterfly (FBFly)

The flattened butterfly is a cost and energy efficient interconnection topology which utilizes fewer high degree switches [8], [10] and requires less physical links, relative to the folded-Clos (fat-tree) topology [4]. Moreover, by implementing less switching stages, FBFly can achieve better performance indicators such as latency. The drawback of FBFly is the requirement of adaptive routing to load balance arbitrary traffic patterns; in turn, fat-tree can employ simple static routing, even in the presence of complex traffic patterns, by making use of the available multiple physical paths [10].

We closely overview the FBFly structure following its description in [11], [10]. A regular k-ary n-fly butterfly topology contains k^n input and output terminals. It consists of n levels (rows) of butterfly interconnects: at level l, where $l = 0, 1, ..., (n-1)$, nodes are connected to others at distances that are multiples of k^l. A k-ary n-flat FBFly topology is constructed from a k-ary n-fly butterfly topology by combining the switches in each of the k^{n-1} columns of the butterfly topology into one single switch. The unidirectional input and output ports of the terminals are combined into bidirectional ports of $N = k^n$ servers. In general, when a butterfly is flattened, the transformation results in a $n-1$ dimensional array of $S = k^{n-1}$ switches. Each switch connects to k servers, and each switch is connected to each of the $k-1$ other switches that align with it in each of the dimensions. The total number of switch ports in a k-ary n-flat FBFly topology is $p = (n-1)(k-1)+k$.

Now consider a k-ary n-flat flattened butterfly as a graph $G(V, E)$, where V denotes the switches, and E denotes the uni-directional links between switches (we omit the terminal nodes and their connections to the switches). The number of full mesh subgraphs in G is denoted by $|G_{fm,k,n}|$ and can be computed as

$$|G_{fm,k,n}| = \frac{|E|}{|E_{fm}|} = k^{n-2}(n-1) \ , \qquad (1)$$

where $|E_{fm}|$ denotes the number of links in one full mesh.

B. Pseudo-full Mesh with AWGR (C-FBFly)

The main idea of the proposed C-FBFly is to substitute each of the previously computed full mesh by an optical pseudo-full mesh, similarly as in [25]. This transformation is achieved using a $M \times M$ AWG (aka. AWGR, short for Arrayed Waveguide Grating Router). This is a passive data-rate independent optical device that routes each wavelength of an input port to a different output port in a cyclic way: wavelength λ of input i is routed to output

$$[(i + \lambda - 2) \ mod \ M)] + 1, 1 \leq i \leq M, 1 \leq \lambda \leq \Lambda \ , \qquad (2)$$

where M is the number of input and output ports of the device and Λ is the total number of wavelengths [13]. The cyclic wavelength routing characteristics of the AWGR implements contention resolution in the wavelength domain, allowing for different input signals to reach the same output in parallel. In other words, an optical star network is being formed with the AWGR in the center which enables full connectivity [14]. We refer to this star structure as a "pseudo"-full mesh (C-FBFly).

At a high level, C-FBFly consists of k inter-rack fiber pairs instead of $k(k-1)/2$ inter-rack cables, as in the full mesh. Moreover, the same level of connectivity is guaranteed without an apparent bottleneck imposed by the AWGR. Practically, the switches use the same amount of ports as in FBFly's full meshes. The key difference is that the regular (*grey*) optical transceivers in FBFly (as proposed in [10]) are exchanged for DWDM capable (or also known as *colored*) transceivers. Such colored transceivers are commonly used in long reach telecommunication networks to take advantage of wavelength division multiplexing. In our case, we point out that while a grey optical device can only transmit and receive signals on a single standard wavelength, a DWDM transceiver can be set to any wavelength of the standardized ITU-T wavelength grid [15]; for this reason, the contention can be resolved in the wavelength domain and thus the AWGR is not a bottleneck.

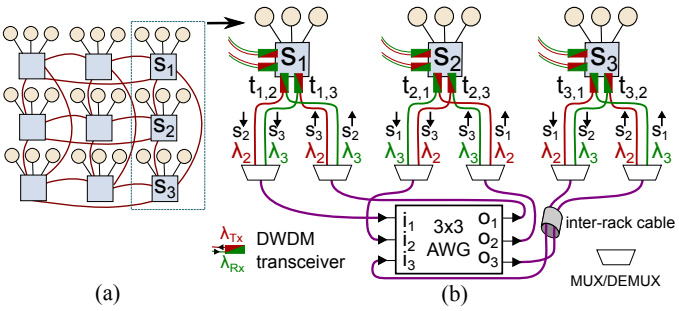

Fig. 1. a) Original 3-ary 3-flat flattened butterfly (FBFly). b) One of the original FBFly full mesh is transformed into an optical pseudo-full mesh; here, switch s_2 transmits to and receives from s_1 through transceiver $a_{2,1}$ on wavelengths λ_3 and λ_2, respectively. Wavelength λ_1 is looped back by the AWGR (Eq. 2) from i_m to o_m, $1 \leq m \leq M$, so it is not used for transmission.

Let us next describe C-FBFly from a cabling perspective (see Fig. 1b). Each transceiver a of a switch s is set to transmit on a particular wavelength such that the AWGR routes its signal to the appropriate destination switch given in Eq. 2, which is the switch t connected to s through port a in one of the original FBFly full meshes. The output fibers of the ports of switch s are optically multiplexed in the wavelength domain onto one fiber, which is then connected to the ith input port of the AWGR. Also, the ith output port of the AWGR is connected to a demultiplexer d with a single fiber. Note that the two unidirectional fibers between switch s and the AWGR can be merged into a fiber pair. Finally, according to the cyclic wavelength routing rule, the fiber exiting demultiplexer d is connected to the receiver side of transceiver a of switch s, which corresponds to the wavelength on which t is transmitting to s. The logical matching of the original fully meshed inter-rack channels to colored input and output signals is done for all colored transceivers of switch s.

This procedure is repeated for every full mesh of FBFly, and for every corresponding switch. In this way, every physical full mesh of the flattened butterfly topology is substituted for a pseudo-full mesh employing colored optical transceivers, multiplexer/demultiplexer (Mux/Demux) devices, and AWGRs. We point out that a pseudo-full mesh contains more fibers than a regular full mesh, since the number of cables between the switch ports and the Mux/Demux's is double the number of the fibers in a regular full mesh. However, as we will show in Sec. VI, the number of long inter-rack cables (i.e., the cabling complexity metric used in this paper) is reduced by a significant amount, which in turn reduces the total amount of fibers in the entire network.

C. Practical Considerations of AWGR

Commercially available AWGR devices have a limited number of input and output ports. This is due to the fact that the widely available DWDM optical parts are optimized for the optical C-band, which is roughly $B_{C-band} = 4800$ GHz wide. For example, by using the ITU-T [15] channel grid with $Ch_{spacing} = 50$ GHz channel spacing, the number of input/output ports M defined as

$$M = B_{C-band}/Ch_{spacing} \ , \tag{3}$$

is limited to 96. The custom manufacturing of AWGRs allows customers to decide on the desired number of input/output ports. According to Eq. 3 there is a trade-off between the number of input/output ports (M) and the optical bandwidth of each channel; note that higher M yields narrower channel widths. We finally mention that in lab environments, the number of input/output ports per AWGR can scale up to 512, along with 10 GHz channel spacing [16].

To achieve best performance in terms of latency and throughput, one needs to keep the number of dimensions of a flattened butterfly as small as possible [8], [10]. If we consider that today switches with up to 1600 10 Gigabit Ethernet (GE) ports are available [19], one could build a 800-ary 2-flat flattened butterfly topology with 640K servers by using 800 of these high degree switches. However, the constraints of the commercially available AWGR devices limit our focus to smaller radix flattened butterfly topologies. Since in our proposed structure the number of switches in one full mesh defines how many input and output ports are needed in a AWGR ($M = k$), we only consider k-ary 3-flat topologies with $k \leq 96$. This configuration scales up to $N = 884,736$ servers.

IV. Cable Length Calculation

Our optical cabling cost model is composed of different subunits. First of all, we investigate the current trends of *rack power density* in data centers. We take notice of the limitations of the maximum power that can be provided for a rack to calculate the server density per square meters. Using the justified density values we count the number of racks we require to accommodate the computing and switching equipment. Depending on the number of racks we estimate the *average inter-rack cable lengths*. Based on these considerations, the cabling cost in case of both the regular grey optics and the colored optics can be estimated in a realistic fashion.

A. Rack Power Density

The design of a power distribution and cooling system for a data center is a complex subject on its own. We limit our analysis only to determine what size of raised floor space our structure would need if it was deployed in a real-world scenario, considering the constraints of the power and cooling capacity of current data centers. We base our model on the power distribution capacity per each rack of a data center infrastructure and denote this metric as P_{rack}. Realistic values of P_{rack} can vary from 4kW/rack to 28kW/rack, the latter value being roughly the maximum power density that can be achieved with air based cooling [20].

For the arrangement of the servers and switches in the racks the following model is used. Let P_{ser} and P_{sw} be the power consumption of a server and of a switch, respectively. RU_{ser} and RU_{sw} denote the rack unit (RU) space required for one server and for one switch, respectively. The total space in RU of a rack is denoted by RU_{rack}. Along our analysis we set $RU_{rack} = 42$, which is the standard capacity of a data center rack. To keep our rack power density model simple and tractable, we make the following simplifying assumptions: a) the racks are filled with servers so that space and power surplus is left in *every* rack for a switch, and b) a switch is placed into a server rack such that the total amount of switch to server cables is minimized. Based on the previous assumptions we can calculate the number of servers per one rack as:

$$|Ser_{rack}| = \left\lfloor min\left(\frac{P_{rack} - P_{sw}}{P_{ser}}, \frac{RU_{rack} - RU_{sw}}{RU_{ser}}\right)\right\rfloor . \tag{4}$$

The height of a server is assumed to be $RU_{ser} = 1$ which can be considered as typical for current high-end data center servers [21]. The actual value of P_{sw} usually depends on the technical specification of the servers (number of processors, number of cores per processor, memory, I/O devices). Server power consumption values are in the range of 200-600 W based on a measurement study of twenty different production data center servers [22]. The switches are modeled after the Arista 7504 switch, which is 7 RU high and has 192 SFP+ ports. According to the data sheet [23] of the Arista 7504, a fully loaded chassis typically consumes about 2 kW (10 W per port).

B. Average Optical Cable Length

To calculate cable lengths we use a simple data center floor space model based on the packaging layout of the flattened butterfly topology [8]. We assume 23" $\times$47" ($\simeq$ 0.6 m$\times$1.2 m) racks which are placed into a *working cell* ([24], [8]), that is about two times deep as the depth of the rack to allow for accessing and cooling the racks. Similarly to the original flattened butterfly cabinet layout, we assume that the racks are aligned into rows of working cells in a raised floor area. The edge length of this area is estimated by $E_l = \sqrt{N/D}$, where D is the node density in $nodes/m^2$. D is calculated by dividing the number of servers per rack (Eq. 4) by the footprint of a working cell $D = |Ser_{rack}|/A_{wc}$. Then the average cable length between the switches of a FBFly network can be simply estimated as:

$$L_{avg} = E_l/3 \ . \tag{5}$$

Multi-gigabit links longer than 5 meters[1] cannot be efficiently built with electrical cables, so the common practice is that these longer links are created with optical cables [18]. In our model we consider optical cables of length $L_{avg} \geq$ 2 m, accounting thus for the fact that the inter-rack cables are

[1]According to the IEEE802.3ba standard, 40 Gbit/s and 100 Gbit/s transfer speeds with copper cable are supported up to 7 meters.

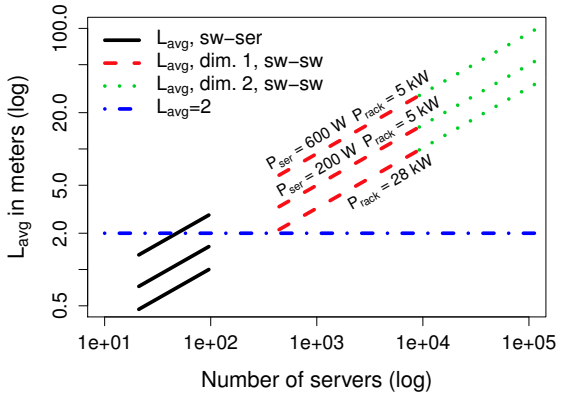

Fig. 2. L_{avg} values shown on a log-log plot when growing the size of three FBFly topologies. L_{avg} is smaller in case of 28 kW racks than in case of 5 kW racks. Also, in case of 5 kW racks, lower P_{ser} yields lower L_{avg}.

running in cable trays few feet over the racks [18]. Moreover, the endpoints of the links can be vertically anywhere in the racks [8], which adds about 3 meters to every inter-rack cable.

Fig. 2 shows L_{avg} for a) a k-ary 1-flat topology, b) a k-ary 2-flat topology, and c) a k-ary 3-flat topology adjusting the parameter k from 21 to 96. The (black) solid part of the lines corresponds to the average length of links between k servers and 1 switch in a k-ary 1-flat FBFly topology. The (red) dashed part corresponds to a 1-dimensional (k-ary 2-flat) FBFly's average cable length. The (green) dotted parts show the values of L_{avg} for links in both the first and second dimensions of a k-ary 3-flat FBFly. For each of the three topologies, we consider three power scenarios (corresponding to the parallel lines). In the first two we assume $P_{rack} = 5$ kW and also consider the server consumptions $P_{ser} = 600$ W and $P_{ser} = 200$ W. In the third we assume $P_{rack} = 28$ kW, in which case the actual server power consumption does not influence servers' placing into racks. We also consider the limiting case $L_{avg} = 2$ m. For topology a), the figure shows that $L_{avg} \leq 2$ m for links between switches and servers (denoted by 'sw-ser'), except for the scenario when $P_{ser} = 600$ W and $P_{rack} = 5$ kW corresponding to a legacy type data center building with low power density and current high-end servers. We argue that such scenarios should be avoided; instead, power distribution and cooling should be proportionally upgraded with server upgrades. For topologies b) and c), the figure illustrates inter-rack cables (denoted by 'sw-sw') with $L_{avg} \geq 2$ m.

We finally mention that in our floor space model, we always consider an optical cable to be a pair of single mode fibers[2].

[2]Links longer than 7 meters are usually connected with optical multi-mode fibers (MMF) because they can be combined with relatively cheap transceivers. Although MMFs have a range of about 200−300 meters, experts argue that distances ≥ 50 meters are better handled with single mode fibers (SMF). A SMF has a longer range but it requires more expensive transceivers. We also point out that DWDM optical systems have to use single mode fibers as well.

V. Optical Cabling Cost Model

In this section we calculate the total costs of *grey* and *colored* optical cabling in FBFly and C-FBFly topologies, respectively. In particular, we calculate the fiber cable costs and the transceiver costs, and also account for the extra optical devices (Mux/Demux and AWGR) in case of C-FBFly. We strive for an accurate cost comparison by accounting for the installation costs of both the inter-rack cables and extra optical devices. We omit the cost components, however, that are equivalent in case of both structures, so switch, server, etc. equipment and installation costs are not included in our model. The cost of cabling between servers and switches is omitted as well.

A. FBFly: Grey Cabling Costs

First we describe the cabling cost model of the original FBFly based on grey optics. Based on the parameters of a k-ary n-fly FBFly topology, the number of fiber links per switch equals $(n-1)(k-1)$, so the total number of inter-rack fiber links is

$$l_{grey} = \frac{S\,(n-1)(k-1)}{2}\ , \tag{6}$$

where S is the total number of switches. Recall that, by definition, l_{grey} is the cabling complexity of the original FBFly structure. The actual average cable length is $L_{real} = L_{avg} + L_{const}$, where L_{avg} is the average cable length in first and second dimension of FBFly and L_{const} is the required extra length of inter-rack cables (see Sec. IV-B). Then the total length of fiber cables for the grey structure is

$$L_{grey} = l_{grey}\, L_{real}\ . \tag{7}$$

For a cost c_{fiber} of a meter of fiber, cost of grey transceivers $c_{tr,grey}$, and also accounting for the inter-rack fiber cable installation cost $c_{long,inst}$, the total grey optical cost is

$$C_{grey} = L_{grey} c_{fiber} + 2\, l_{grey}\, c_{tr,grey} + l_{grey}\, c_{long,inst}\ . \tag{8}$$

This C_{grey} will represent the cost of cabling in FBFly in Sec. VI-A.

B. C-FBFly: Colored Cabling Costs

We now describe the cabling cost model for our proposed C-FBFly structure. The amount of required optical equipment is first calculated for one full mesh of the structure, then by considering optical equipment, cabling and installation costs, the calculation is done for the whole structure.

Number of AWGRs in a Full Mesh. Let us first recall that the key idea of C-FBFly is to substitute each full mesh of FBFly by a pseudo-full mesh with an $M \times M$ AWG (or AWGR) in the center. The number of input and output ports M of the AWGR is the number of switches in one dimension, which corresponds to the actual number of switches connected in a full mesh ($M = k$). In our next step we refer back to Sec. III-C, where we mentioned that it is possible to construct links with more than one sub-link. The AWGR based pseudo-full mesh can support this feature as well. We introduce the variable l_{sub} to set the number of sub-links encompassed in

one link. For the later calculations we always regard one link to contain only one sub-link (in particular, one 10 Gbit/s link corresponding to signal spectral efficiency of 0.2 Bit/s/Hz), however our model is general enough to support calculations for higher trunking factors. For this we first need to calculate the width of an AWGR channel:

$$B_{AWG,M} = B_{C-band}/M \ . \tag{9}$$

Since the whole C-band spectrum is divided into M distinct channels, each channel has a width equal to $B_{AWG,M}$. If each AWGR channel has $B_{AWG,M}$ width, then by using 50 GHz spaced ITU-T channels for the signals going through the AWG, one can fit

$$|Ch_{/AWG}| = \lfloor B_{AWG,M}/Ch_{spacing} \rfloor \tag{10}$$

number of signals on one AWG channel. Combining Eqs. 9 and 10, we obtain the number of channels per AWG channel on which we can send signals, i.e.,

$$|Sig_{AWG}| = \lfloor (B_{C-band}/M)/Ch_{spacing} \rfloor \ .$$

Based on the above equations we can finally calculate the number of $M \times M$ AWGs required per one full mesh

$$|AWG_{fm}| = \lceil l_{sub}/|Sig_{AWG}| \rceil \ ,$$

which also determines the number of multiplexers and demultiplexers as $|MuxDem_{fm}| = 2\,M\,|AWG_{fm}|$ per full mesh.

Number of Cables in a Full Mesh. In one full mesh, the number of long fiber pairs needed to connect the Mux/Demux devices with the AWGR is

$$l_{fm,MUX-AWG} = M\,|AWG_{fm}| \ .$$

In turn, the number short fiber pairs needed to connect the switch ports to the Mux/Demux devices is

$$l_{fm,sw-MUX} = M\,(M-1) \ .$$

Fig. 3 shows the arrangement of both short and long optical cables in C-FBFly.

Total Optical Equipment. Now we have the number of required colored equipment including the cables for one pseudo-full mesh in C-FBFly. We refer back to Eq. 1 where we calculated $|G_{fm,k,n}|$ as the total number of full mesh subgraphs in FBFly. Then, the total numbers of AWGRs, Mux/Demuxes, long and short cables in case of a given k-ary n-flat C-FBFly topology are:

$$\begin{aligned} |AWG| &= |AWG_{fm}|\,|G_{fm,k,n}| \\ |MuxDem| &= |MuxDem_{fm}|\,|G_{fm,k,n}| \\ l_{MUX-AWG} &= l_{fm,MUX-AWG}\,|G_{fm,k,n}| \\ l_{sw-MUX} &= l_{fm,sw-MUX}\,|G_{fm,k,n}| \ . \end{aligned} \tag{11}$$

Total Length of Optical Cables. Next we estimate the actual average cable lengths for the colored interconnection structure similarly as in the case of the grey cabling. However, we want to account for the fact that the AWGR device is about the size of a smart phone and it requires only couple of volts for its temperature control. Because of this, the AWGR device can easily be placed in the middle of a given full mesh, which means that a reasonable length estimation of the fiber pairs connecting the multiplexers/demultiplexers with the AWGR is $L_{real}/2$. For the length of the fibers connecting the switch ports with the multiplexers and demultiplexers, we use a conservative value $L_{sw-MUX} = 1$ m. Working with these average cable lengths, the total length of long cables running from the Mux to the AWGR and back to the Demux is given by $L_{long} = l_{MUX-AWG}\,L_{real}/2$. In turn, the total length of short cables between the switch and Mux/Demux is $L_{short} = l_{sw-MUX}\,L_{sw-MUX}$. We thus obtain the total length of cables in C-FBFly

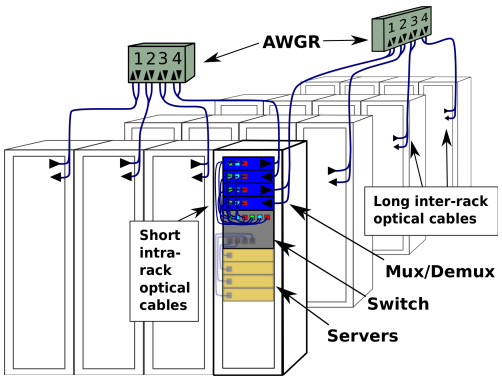

Fig. 3. Optical cabling in C-FBFly. Only one pseudo-full mesh is shown in each dimension.

$$L_{col} = L_{long} + L_{short} \ . \tag{12}$$

Total Optical Cost of C-FBFly. Having the number of AWGRs, Mux/Demuxes, and the total length of required fiber cables, we can finally calculate the total cabling cost $C_{colored}$ of C-FBFly. We denote the unit costs of the extra colored optical equipments, i.e., c_{AWG} for an $M \times M$ AWG and c_{MuxDem} for a Mux/Demux. Note that we calculate with the same value of c_{fiber} as in case of FBFly (Sec V-A). The unit cost of a colored transceiver is denoted as $c_{tr,col}$. Note that the total number of colored transceivers is the same as the total number of switch-Mux/Demux cables, and also the total number of grey transceivers, i.e., $T_{col} = l_{sw-MUX} = 2\,l_{grey}$. Finally, we consider the total installation costs

$$\begin{aligned} C_{inst} = l_{MUX-AWG}\,c_{long,inst} + l_{sw-MUX}\,c_{short,inst} + \\ + |AWG|\,c_{AWG,inst} + |MuxDem|\,c_{MuxDem,inst} \ . \end{aligned}$$

Here we account for the installation cost of the extra AWGRs and Mux/Demux devices as well as the cable installation costs. Based on the above details, adding up all the cost components we obtain

$$\begin{aligned} C_{colored} = |AWG|\,c_{AWG} + |MuxDem|\,c_{MuxDem} + \\ + L_{col}\,c_{fiber} + T_{col}\,c_{tr,col} + C_{inst} \ . \end{aligned} \tag{13}$$

This total colored optical cost will be compared against the total grey optical cost from Eq. 8 in the next section.

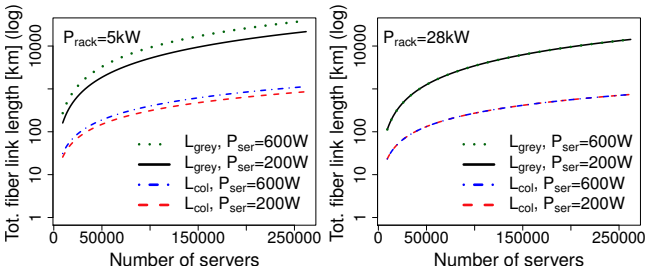

Fig. 4. Total length of grey (Eg. 7) and colored (Eg. 12) optical cables for different server power consumption and rack power availability. The difference between corresponding total lengths of grey and colored FBFly solely determines the cable cost reduction.

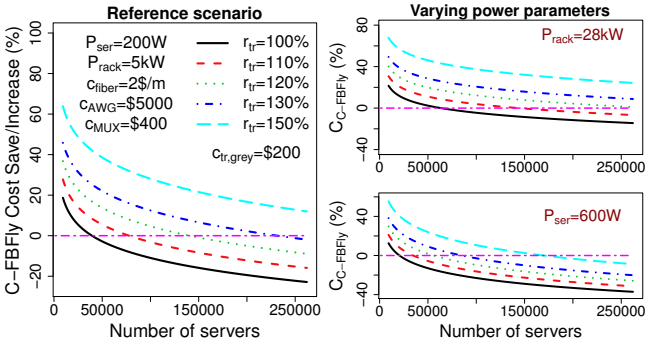

Fig. 5. Cost save (negative) or increase (positive) in case of implementing the C-FBFly structure in a reference scenario for different network size. On the right figures the server and rack power parameters are varied.

VI. Results

In this section we first numerically compare the capital expenditure (CapEx) cabling costs of the colored and the grey structure and we determine the sensitivity of cost saving achieved in C-FBFly relative to power parameters and equipment prices. Then the cabling complexity reduction achieved by C-FBFly is analytically quantified relative to FBFly, and some operational aspects of the reduction are discussed.

A. CapEx Optical Cost of C-FBFly vs. FBFly

Total Optical Cable Length Reduction. Based on the floor space model described in Sec. IV and the optical cabling model detailed in Sec V, the total length of optical cables is calculated for both FBFly and C-FBFly. The results are illustrated in Fig. 4 for different server power consumption and rack power availability. The figure clearly shows that the C-FBFly's total cable length is shorter than the grey FBFly's in every case. Moreover, one can observe a subtle dependence between the power density of the data center and the total cable length, i.e., given low rack power density, high power consuming servers must be spread across the raised floor beyond what the RU space in racks can accommodate. This relationship implies higher average cable length (for higher power consuming servers).

CapEx Optical Cabling Cost Balance. Next we compare the total optical cabling costs of C-FBFly and FBFly. We define the cabling cost of C-FBFly as the proportional increase (positive) or decrease (negative) of capital optical costs compared to the original FBFly optical cabling costs

$$C_{C-FBFly} = (C_{colored} - C_{grey})\,100\,/C_{grey}\ . \qquad (14)$$

$C_{C-FBFly}$ is calculated for FBFly data centers in the range of 9K to 260K servers. The cost of the grey and the colored FBFly's optical interconnection network is calculated based on web prices[3]. Actual cost prices may vary greatly, e.g., 40-60% of cost reduction of list prices is common for large orders. We assume cable installation cost of \$6.25 per inter-rack cable, and \$2.5 for short cables similarly to [18]. The installation cost of an AWGR or Mux/Demux device is assumed \$50, which is a conservative value. Here we considered the fact, however, that colored (DWDM) equipments are not common in data center deployments and they might incur some extra installation fees[4]. For the transceiver cost component, we define r_{tr} to denote the price of a colored transceiver compared to the price of a grey transceiver in terms of percentages:

$$r_{tr} = \frac{c_{tr,col}}{c_{tr,grey}}\,100\ . \qquad (15)$$

Gridlines of different r_{tr} values help to identify the ratio of colored vs. grey transceiver prices, at a given size of the structure, when the cabling complexity reduction results in cost reduction at the same time. Our purpose in this section is to give an overview of prices which result in cost saving for reasonable sized large scale data center networks. Fig. 5 left shows the cost savings for a C-FBFly reference scenario with realistic power parameters and optical equipment costs. Note that $C_{C-FBFly} < 0\%$ corresponds to capital cost savings, when using colored optics. In turn, when $C_{C-FBFly} > 0\%$, then the reduction in cabling complexity is achieved at a higher capital cost. For example, in case of a 200K server structure, implementing C-FBFly with 10% more expensive colored transceivers and reference optical prices, results in 12% cabling cost reduction compared to FBFly, which amounts to about \$14M in total cost saving.

CapEx Cost Sensitivity to Power Parameters Figs. 5 on the right show the cost balance for different server power consumption and rack power availability. Except for P_{rack} and P_{ser}, the power parameters and the optical equipment costs are the same as in the reference scenario. An interpretation of the results is that, in the case of a renewal of a legacy data center (e.g., when additional power density cannot be integrated in the building, or when increasing cooling efficiency is infeasible), then C-FBFly can be a cost effective choice for networks with > 50K servers. This efficiency is achieved for example by assuming the bottom right scenario in Fig. 5, using colored transceivers prices of 120% of grey prices, and the prices for additional colored equipment (shown on the left).

[3]The cable cost prices are estimated from http://www.fiberstore.com/. The transceiver and colored optical equipment prices may greatly vary depending on manufacturer and quantity of order (http://www.alibaba.com/).

[4]The total installation costs are negligible compared to all other cost components.

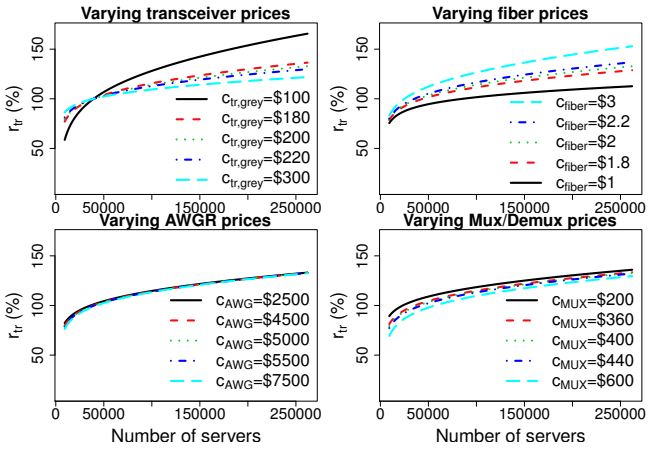

Fig. 6. Sensitivity analysis for transceiver cost balance r_{tr} when specifying $C_{colored} = C_{grey}$ and varying optical equipment prices. Note the difference of the transceiver cost balance sensitivity to fiber cable and grey transceiver prices vs. AWGR and Mux/Demux prices.

Cost Sensitivity to Equipment Prices We were interested in equipment prices when the colored and the grey cabling costs equal out each other for a given size of structure. This tells us the desired equipment prices when deciding to chose the colored structure over the grey one. Moreover, we wanted to indicate the sensitivity of the cost balance to the change in different equipments' prices, so that we assess which type of equipment dominates the balance. We emphasize that the price difference of colored and grey transceiver prices strongly determines the financial feasibility of C-FBFly, since we must use the same amount of transceivers in both cases. For these reasons we analyzed the cost ratio of transmitter prices r_{tr} for distinct equipment price values when the colored and the grey cabling costs are the same ($C_{colored} = C_{grey}$). This sensitivity analysis is done by increasing and decreasing each optical equipment price in turns by 10% and 50% relative to the reference scenario (Fig. 5 left). The first and second plot of the first row on Fig. 6 show that the cost saving achieved by C-FBFly is highly sensitive to grey transceiver and fiber cable prices. The second row indicates that the capital cost is less sensitive to AWGR and Mux/Demux price variance. We note that sensitivity of r_{tr} to all installation prices are similar to the sensitivity of the AWGR equipment price component and we omit these details for space constraints. The main message here is that the purchase and installation of extra optical devices do not contribute much to the CapEx cost of cabling in C-FBFly. On the other hand, high optical fiber prices and low transceiver prices greatly favor C-FBFly.

B. Cabling Complexity Reduction

The cost results presented so far are quite conservative, relative to C-FBFly, because they are exclusively based on CapEx costs. We point out that we did not quantify the additional operational cost component (i.e., OpEx), which is arguably significantly reduced given the very large cabling complexity reduction in C-FBFly. We elaborate on the OpEx cost components by quantifying the cabling complexity reduction as the reduction in the number of inter-rack cables.

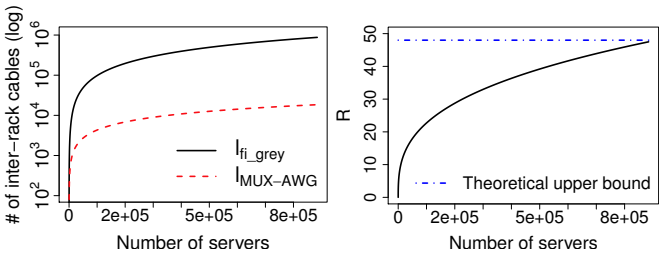

Fig. 7. Left: Total number of sw-sw links vs. number of servers in FBFly (from Eq. 6) and C-FBFly (from Eq. 11). Right: Ratio R of the number of sw-sw links in FBFly to sw-sw links in C-FBFly.

The C-FBFly cabling complexity reduction depends on four factors: the parameter k of the FBFly structure, the width of the optical C-band B_{C-band}, the channel spacing $Ch_{spacing}$, and the number of sub-links in a link l_{sub}. The cabling complexity reduction can be quantified by taking the ratio of l_{grey}, i.e., the number of inter-rack cables in the grey structure (Eq. 6), to $l_{MUX-AWG}$, i.e., the number of inter-rack cables in the colored structure (Eq. 11):

$$R = \frac{k-1}{2} \left\lceil \frac{\lfloor \frac{B_{C-band}}{Ch_{spacing}\, k} \rfloor}{l_{sub}} \right\rceil . \tag{16}$$

An immediate upper bound is

$$R_{upper} = \frac{48(k-1)}{k\, l_{sub}} \approx 48 . \tag{17}$$

The approximation is achieved when $k \to 96$, assuming one sub-link per link ($l_{sub} = 1$). Thus, 48 is a rough estimation of the theoretical upper limit of the cabling complexity reduction. Fig. 7 (left) shows the scaling of the inter-rack cables for FBFly and C-FBFly. Fig. 7 (right) illustrates R. We point out that the logarithmic behavior of R, in the number of servers N, is due to the exponential scaling of servers (Sec. III-A).

We argue that by localizing the fully meshed cabling into the racks, the overall cabling management is significantly simplified. The simplified cable management further reduces the risk of miswiring and unplanned downtime of the data center. The detailed quantification of these operational costs are regarded as future work.

VII. Discussion

Here we discuss C-FBFly's cabling complexity reduction results considering its control plane and network capacity while also mentioning some related technological considerations.

Our current approach reduces the number of long inter-rack cables without modifying routing in FBFly. In contrast, Dragonfly [12] achieves a two-fold cabling complexity reduction by increasing control plane complexity. While the Dragonfly is a fixed 3-level structure, our cabling complexity reduction results hold for arbitrary dimensional FBFly structures.[5] HyperX [11] generalizes the FBFly structure to achieve

[5]For the discussion on the trade-off between FBFly's number of dimensions and the number and length of inter-rack cables we point the reader to [8].

higher structural performance using less switching equipment. The authors in [18] develop an optimization framework for HyperX to find efficient designs given size and structural capacity constraints, focusing on reduction of overall costs, but without quantifying the resulting cabling complexity. We point out that no oversubscription is used in our calculations to achieve full bisection bandwidth in our topologies. However, the capacity of C-FBFly can be easily tailored by removing or adding inter-rack link capacity. Moreover, HyperX topology can also benefit from using colored pseudo-full meshes and the development of this concept is regarded as future work.

The key advantage of C-FBFly is that it opens up a seamless migration path towards higher data center capacities. Our approach can take advantage of higher capacity interfaces of servers and switches (e.g., 40 Gbit/s 40GBASE-FR interface standardized in IEEE 802.3bg). The proposed structure provides the opportunity to increase network capacity by only exchanging the interfaces for more advanced interfaces with higher spectral efficiencies. Hence the capacity of the structure can be increased without modifying the already installed DWDM fiber infrastructure. Moreover, C-FBFly does not limit cabling lengths due to the application of SMF cables, and this fact makes C-FBFly an attractive structure for distributed DCs.

The currently proposed architecture offers cost advantages in data centers using a large number of cost efficient 10 Gbit/s DWDM interfaces. For this reason, our proposal should be regarded as a strong argument behind the standardization of a cost-efficient DWDM capable (10GBASE-FR) one-lane interface using the optical C-Band on single mode fibers.

VIII. Conclusion

In this paper we have shown how to significantly reduce the cabling complexity, defined as the number of long inter-rack cables, in large flattened butterfly networks. The key idea was to use dense wavelength division multiplexing capable colored optical transceivers, in combination with arrayed waveguide grating routers, instead of grey transceivers. To evaluate the benefits of our new structure, we presented a cost model based on a data center floor space layout model with power density constraints. We applied this cost model to conduct a sensitivity analysis to optical equipment costs, and identified the required optical network equipment costs for our (colored) structure to be more cost efficient than the original (grey) structure. For example, if fiber cost prices are 2.8 $/m, and if colored and grey interfaces' prices are within 110% or 120%, then our structure lowers capital expenditure costs in networks with more than 50K or 75K servers, respectively. Moreover, our proposed structure additionally reduces operational costs by arguably a significant factor, given the very large cabling complexity reduction (e.g., by a factor of as much as 48).

References

[1] D. Abts and B. Felderman. A guided tour of data-center networking. *Communications of the ACM*, 55(6):44–51, June 2012.

[2] Corning Inc. Sustaining the cloud with a faster, greener and uptime-optimized data center. June 2012. Available online: http://goo.gl/fAloQ

[3] N. Farrington, E. Rubow, and A. Vahdat. Data center switch architecture in the age of merchant silicon. In *Proc. of HOTI*, 2009.

[4] C. E. Leiserson. Fat-trees: universal networks for hardware-efficient supercomputing. *IEEE Trans. on Computers*, 100(10):892–901, Oct. 1985.

[5] M. Al-Fares, A. Loukissas, and A. Vahdat. A scalable, commodity data center network architecture. In *Proc. of ACM SIGCOMM*, 2008.

[6] A. Bechtolsheim and J. Ullal. Evolutionary designs for cloud networking. Jan. 2009. Available online: http://goo.gl/rmCLJ

[7] N. Lippis. A simpler data center fabric emerges. Jun. 2010. Available online: http://goo.gl/CDPCU

[8] J. Kim, W. J. Dally, and D. Abts. Flattened butterfly: a cost-efficient topology for high-radix networks. In *Proc. of ACM/IEEE ISCA*, 2007.

[9] J. Hamilton. Data center networks are in my way. *Standford Clean Slate CTO Summit*, 2009. Available online: http://goo.gl/1SKH2

[10] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu. Energy proportional datacenter networks. In *Proc. of ACM/IEEE ISCA*, 2010.

[11] J. H. Ahn, N. Binkert, A. Davis, M. McLaren, and R. S. Schreiber. HyperX: topology, routing, and packaging of efficient large-scale networks. In *Proc. of SC*, 2009.

[12] J. Kim, W. J. Dally, S. Scott, and D. Abts. Technology-driven, highly-scalable dragonfly topology. In *Proc. of ACM/IEEE ISCA*, 2008.

[13] C. Kachris and I. Tomkos. A survey on optical interconnects for data centers. *IEEE Comm. Surveys & Tutorials*, 14(4):1021 –1036, Q4 2012.

[14] R.-P. Braun and D. Fritzsche. Cost effective scalable optical networks-transparent optically routed network (TOR-NET). *Photonische Netze*, ITG-Fachbericht Band 233, article no. 32. 2012.

[15] ITU-T G.694.1 Spectral grids for WDM applications: DWDM frequency grid. Online reference: http://goo.gl/h1YgK.

[16] K. Takada, M. Abe, M. Shibata, M. Ishii, and K. Okamoto. Low-crosstalk 10-GHz-spaced 512-channel arrayed-waveguide grating multi/demultiplexer fabricated on a 4-in wafer. *IEEE Photonics Technology Letters*, 13(11):1182–1184, Nov. 2001.

[17] IEEE Standard 802.3bg-2011. Part 3, Amendment 6: Physical Layer and Management Parameters for Serial 40 Gb/s Ethernet Operation Over Single Mode Fiber. Online reference: http://goo.gl/Bg0cv

[18] J. Mudigonda, P. Yalagandula, and J. C. Mogul. Taming the flying cable monster: A topology design and optimization framework for data-center networks. In *Proc. of USENIX ATC*, 2011.

[19] Alcatel-Lucent 7950 ERS. Online reference: http://goo.gl/9v1Sj

[20] R. L. Mitchell. Data center density hits the wall. Jan. 2010. Available online: http://goo.gl/0IcsP

[21] Dell PowerEdge R620 Server. Online reference: http://goo.gl/S48Vo

[22] A. Vasan, A. Sivasubramaniam, V. Shimpi, T. Sivabalan, and R. Subbiah. Worth their watts? - An empirical study of datacenter servers. In *Proc. of IEEE HPCA*, 2010.

[23] Arista 7500 Series DC Switch. Online reference: http://goo.gl/RWN6u

[24] M. Patterson, D. Costello, and P. F. G. M. Loeffler. Data center TCO; a comparison of high-density and low-density spaces. In *Proc. of THERMES*, 2007.

[25] I. Keslassy, S.-T. Chuang, K. Yu, D. Miller, M. Horowitz, O. Solgaard, and N. McKeown. Scaling Internet Routers Using Optics. In *Proc. of ACM SIGCOMM*, 2003.

[26] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, and A. Vahdat. Switching the optical divide: Fundamental challenges for hybrid electrical/optical datacenter networks. In *Proc. of ACM Symposium on Cloud Computing*, 2011.

[27] N. Farrington, G. Porter, P. Sun, A. Forencich, J. Ford, Y. Fainman, G. Papen, and A. Vahdat. A demonstration of ultra-low-latency data center optical circuit switching. In *Proc. of ACM SIGCOMM*, 2012.

[28] N. Farrington, G. Porter, Y. Fainman, G. Papen, and A. Vahdat. Hunting mice with microsecond circuit switches. In *Proc. of HotNets*, 2012.

[29] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. Ng, M. Kozuch, and M. Ryan. c-Through: Part-time optics in data centers. In *Proc. of ACM SIGCOMM*, 2010.

[30] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat. Helios: a hybrid electrical/optical switch architecture for modular data centers. In *Proc. of ACM SIGCOMM*, 2011.

[31] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella. DOS-A scalable optical switch for DCs. In *Proc. of ACM/IEEE ANCS*, 2010.

[32] K. Xia, Y.-H. Kaob, M. Yangb, and H. J. Chao. Petabit optical switch for data center networks. Technical Report, NYU-Poly, 2010. Available online: http://goo.gl/wZtyj

[33] X. Ye, S. J. B. Yoo, and V Akella. AWGR-Based Optical Topologies for Scalable and Efficient Global Communications in Large-Scale Multi-Processor Systems. *Journal of Optical Communications and Networking* 4(9):651–662. 2012.