RODA: A Reconfigurable Optical Data Center Network Architecture

Amitangshu Pal and Krishna Kant
Computer and Information Sciences, Temple University, Philadelphia, PA 19122
E-mail:{amitangshu.pal,kkant}@temple.edu

Abstract- In this paper, we introduce a novel all-optical Data center networking (DCN) fabric, by leveraging the reconfigurability of the optical transceivers and switches, while dynamically changing the end-to-end optical routes to match the varying traffic demands. The dynamic flow scheduling along with their wavelength assignment turns out to be a NP-hard problem. We propose centralized heuristics for the inter-rack flow scheduling, by exploiting the optical wavelength division multiplexing, while minimizing the number of intermediate optical hops. Through extensive simulations, we show that the proposed architecture and flow scheduling reduces the network congestion by a factor of 15-18, compared to state-of-the-art part-time optical DCNs. For most of the traffic patterns, the proposed scheme can deliver >90% of the inter-rack traffic through *direct* optical communication.

I. Introduction

Data centers (DCs) house a high volume of computation and storage resources, along with an infrastructure to network them, so that a huge amount of such resources can be quickly and effectively communicated. Data center networking (DCN) is thus an emerging field that is getting a significant research attention in both industry and academia, due to its growing importance in supporting various Internet-based applications, social networking, video content hosting and distribution as well as various high-volume, intensive computation applications.

Limitation of existing schemes: Traditional DCN interconnects the server racks by electrical switching, in a multi-tier interconnection architecture to provide full connectivity. Such interconnection networks either (a) provides low-cost, but poor performance, due to oversubscribed links at the higher levels, or (b) are expensive, over-provisioned solutions [1], [2], [3] that provides all-time high capacity among all the racks. Due to the dynamic nature of the inter-rack traffic, over-provisioning is the only way to provide a good worst-case performance for a static, electrical network. However, some recent studies [4], [5] reveal that, in a DCN only few racks and inter-racks links are hot, and so building such over-provisioned solutions is an overkill. Another downside of electrical DCNs is that, the traditional copper wires experience high electrical loss at higher data rates, that makes them unacceptable for distances over 10 meters for 10 GigE links [6].

To cope with the above mentioned problems of electrical DCNs, optical networking technology has been introduced in c-Through [7], Helios [8], Mordia [9], OSA [6], WaveCube [10]. Optical switches along with wavelength division multiplexing (WDM) provides reconfigurable and dynamic capacity allocation across the DCN racks. Optical cables also support higher data rates over long distances compared to copper wires. Although using optics in DCN have shown a significant improvement in terms of available flexibility and network bandwidth, these schemes are either MEMS (Micro-Electro-Mechanical Systems) based [7], [8], [9], [6], or increases the hop-counts of the inter-rack optical communications [10]. A central MEMS limits the scalability of a DCN design, due to its low port density. The solution of a single MEMS can be avoided by connecting multiple MEMS switches in the form of a multi-stage fattree. However such structure has the problems [10] of (a) limited connectivity, as a MEMS allows pairwise, bipartite connectivity among its ports, (b) significant increase in cost, due to the use of a large number of MEMSes, (c) additional latency while synchronizing multiple MEMSes in a dense DCN. On the other hand, optical communication with multiple intermediate hops incurs (a) additional latency due to optical-electrical-optical (O-E-O) conversion each hop, and (b) additional transceivers to receive and forward the optical signals at intermediate nodes. To overcome these issues, we have designed a MEMS-free ***R**econfigurable **O**ptical **D**ata Center Network **A**rchitecture (RODA)*, by extending the ideas and optical equipments used in [6], [10], while eying to reduce the inter-rack communication hops, with the efficient usage of limited rack-switch ports and optical wavelengths.

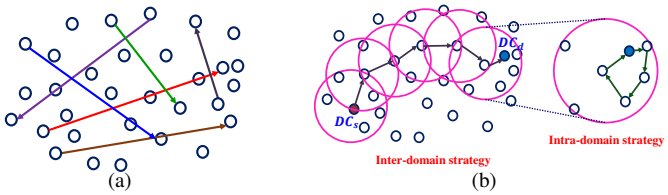


Fig. 1. (a) Point-to-point logistics, vs (b) Shared logistics.

Motivation from logistic networks: The RODA architecture mainly stems from our recent efforts for building a worker-friendly, efficient logistic network. In a typical logistics network, the trucks deliver several products in between different distribution points as shown in Fig. 1(a). Such a point-to-point delivery (a) lengthens the driver's away home time, and at the same time (b) reduces the transportation efficiency due to the fact that, very often the trucks go half-empty [11]. To cope with this, we proposed a worker-friendly and more efficient logistics system [12], shown in Fig. 1(b), where the trucks have their own operating zones. If the source and destination points lie in different operating zones, the products are hopped through multiple trucks of different zones, at the exchange points of the zones. This significantly reduces the driver's away home time and can significantly enhance

This research was supported by the NSF grant CNS-1414509.

the transportation efficiency, since within its operating zone a truck carries products of multiple distribution companies, which it loads/unloads at the corresponding distribution points. The cost, of course, is the extra delay and handling at the distribution points. This paper extends this idea of *shared* logistics to develop an all-optical DCN architecture, where the lightpaths can be selectively added or dropped at any intermediate racks in an on-demand basis, which is identical to truck loading/unloading at the intermediate nodes in the shared logistics.

Our contributions: Interestingly, there is a clear relation between the number of trucks in logistics networks, and wavelengths in DCNs. To illustrate this point, we consider a scenario, where two Top-of-Rack (ToR) switches a and b need to transfer some data to c, as shown in Fig. 2. In presence of just one truck/wavelength, the truck/lightpath is loaded at a, and needs to be re-loaded at b and finally unloaded at c, as shown in Fig. 2(a). Whereas in presence of multiple trucks/wavelengths, two trucks/lightpaths can be independently routed to c, as shown in Fig. 2(b). While it sounds simple, this procedure is difficult to imitate in a real DCN scenario. A reconfigurable unit (distribution point) needs to be attached at b, which takes the decision to either drop a lightpath at b, to pass it, or to deflect it to other directions in case b is attached to multiple nodes. Building such a flexible and reconfigurable interface architecture is one of our key contributions.

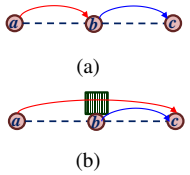


Fig. 2. Multi-hop vs single-hop

Inter-rack flow scheduling, along with their wavelength assignment, on top of the RODA architecture is challenging, which turns out to be NP-hard. We propose a polynomial time approximation scheme that identifies a subset of flows and assign them direct optical hops, while others are assigned multi-hop routes by minimizing the maximum network load. The proposed approximation scheme is shown to be fairly accurate and faster compared to the optimal solution obtained from typical commercial solvers.

The paper quantifies the effectiveness of the proposed inter-rack communication model using extensive simulations, which show that RODA reduces the overall network congestion by a factor of **15-18**, compared to part-time optical DCNs, like c-Through. We also show that, in most cases, >90% of inter-rack flows can be routed by direct optical hops, without any intermediate multi-hopping, in presence of 40 optical wavelengths.

Organization: The rest of the paper is organized as follows. Section II explains the detailed RODA architecture, which is developed by integrating different optical devices that are available in the market. We then propose an inter-rack communication and wavelength assignment scheme on top of RODA in section III. Extensive simulations are presented in section IV. Related proposals and relevant discussions are summarized in section V.

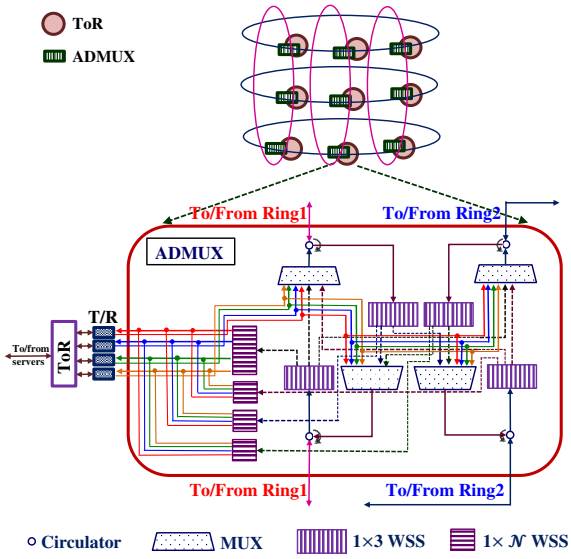


Fig. 3. (a) Proposed architecture of RODA and an ADMUX unit.

II. Network Architecture

A. Optical devices used

A brief overview of the optical devices used for our RODA architecture are listed as follows.

Transceiver: An optical transceiver consists of a transmitter and a receiver. The transmitter end takes an electrical signal and convert it to an optical signal, which is transmitted through an optical fiber. The receiver end receives the optical signal and then converts it to electrical. The transceivers can be tuned to different wavelengths. Common optical transceivers are SFP, SFP+, XFP, X2, Xenpak, GBIC etc [13].

Circulator: Circulators are non-reciprocal optical devices that direct an optical signal from one port to the next, in only one direction, i.e. optical signals that enter at port 1 exits from port 2, similarly port 2 signals are directed to port 3 and so on. Circulators are used to achieve bi-directional optical transmission over a single fiber, because they can separate the optical signals traversing in opposite directions [14].

Multiplexer: Optical WDM multiplexers are $N\times1$ optical devices, which couple multiple optical signals of different wavelengths, from N input ports to a single output port. Multiplexers allow different optical carriers to be transmitted on a single optical fiber, without interfering among each others.

Wavelength selective switch (WSS): WSS is the heart of the RODA design. WSSs are wavelength selective, i.e. they can switch signals depending on their wavelengths. A $1\times N$ WSS has one incoming port and N outgoing ports, which is capable of directing any wavelength from the incoming port to any of the N outgoing ports. This is a very attractive feature, as it allows adding and dropping single wavelengths from a multi-wavelength beam, without the need of electronically process the whole signal.

B. RODA Architecture details

RODA architecture consists of ToR switches that are connected in any regular architecture, as shown in Fig. 3(a). Although we explain our model using a torus topology in

Fig. 3(a), the concepts are applicable in any general regular architecture, such as cube, hypercube or tree. The ToR switches are connected to the servers on one side, and the other side is connected to the optical add-drop multiplexer (ADMUX) units. A ToR switch along with its ADMUX is defined as a *node*. The ADMUX units are connected in ring structures, where signals are transmitted in optical domains.

An ADMUX unit consists of multiplexers (MUXs), wavelength selective switches (WSSs) and circulators, which is shown in Fig. 3(b). Circulators are used to support bidirectional communications. In RODA architecture, we assume that a TOR switch has I ports. Some of them are connected to the servers through direct or some hierarchical tree based multi-tier connection. The rest are connected to tunable transceivers, which are used to provide inter-rack communications. We assume that each ToR switch has $\mathcal{N}$ transceivers, each of them is can be tuned dynamically to separate and distinct wavelengths. The transceivers are connected to the MUXs and the WSSs of the ADMUX unit. The optical signals that enter in the ADMUX units are (a) either dropped by the WSS to the transceivers (b) or passed (c) or deflected to the neighboring ring in case of inter-ring forwarding. All these three operations are possible by dynamically reconfiguring the WSSs and the transceivers to drop and receive different wavelengths. While transmitting at a specific wavelength, (a) the transceiver is tuned to that wavelength, and (b) then the lightpath is added to the ring by selecting any one of the four MUXs, depending on which ring, and which direction (clockwise or anti-clockwise) the node wants to transmit. The transceivers are tunable or reconfigurable in nanoseconds [15]. The WSS units can be built using 2D-MEMS technology which offers microsecond switching time at small port count (1×4 or 1×8), or can be 3D-MEMS where the port counts are higher at the cost of relatively slow reconfiguration time ($\sim$10s of milliseconds), as reported in [9]. In RODA, WSSs that are directly connected to the circulators require small port count (1×3), whereas WSSs that are connected to the transceivers need higher port counts ($1\times\mathcal{N}$), which makes this architecture a bit expensive. However, we expect that the constantly advancing optical technology will make this design cheaper and faster.

Similar to OSA [6], the link capacity is flexible in RODA, i.e. if a source-destination (SD) ToR pair wants to communicate among themselves at a rate which is w times the line speed of a single port, they can establish w lightpaths in between themselves, each of which is associated with a distinct transceiver. In RODA, the fiber cannot carry two flows with same wavelength in the same direction, whereas bidirectional transmission of same wavelength across a fiber is possible. A transceiver can simultaneously send and receive on the same wavelength. In each ring, two transceivers of the same ToR can transmit/receive on the same wavelength, one in clockwise and another in anti-clockwise direction. In torus topology, as the transceivers are associated with two neighboring rings, a total of four such simultaneous transmissions/receptions corresponding to a ToR switch on a single wavelength are possible, as shown in Fig. 4. It is important to note that even though a transceiver can transmit/receive at the same time on the same wavelength, the transmission and reception speed is independent of each other, which is limited by the line speed of a single port.

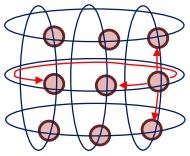


Fig. 4. 3×3 torus topology.

III. Flow Scheduling with Wavelength Assignment

In this section, we describe the flow scheduling along with their wavelength assignment, on top of RODA. Before going into the details, let us define few terminologies that are relevant in the context of data forwarding in the optical domain. In the optical domain, a route between a source-destination pair of ToR switches is known as a *lightpath*. A direct lightpath is also called an *optical* edge/link. In the absence of intermediate wavelength conversion, the *valid available wavelength (VAW)* of a lightpath is defined by the wavelengths that are free at all the physical links of that lightpath.

We describe our proposed flow scheduling scheme in three steps. First, a static, connected optical topology is constructed, which is presented in section III-A. This static topology can be used by any ToR switch to communicate with any other. Small and bursty traffic flows can be routed using this static backbone, without consulting with the central controller. Also when a flow initiates, it is not known whether it will become sufficiently intense. Thus this static topology is used initially to route the flows, whenever the flow grows intense and becomes *significant*, the central controller is communicated to route this flow. Other than the high-volume flows, some delay-sensitive, higher priority, moderate flows may also be considered as significant flows. The switches send their significant flows to the controller, which tries to assign direct lightpaths to the SD pairs periodically, as presented in section III-B. All the remaining flows are sent through multi-hop optical communication, as discussed in section III-C. We assume that all the switches can be configured dynamically by the central controller. In this paper, all the flows are considered as significant.

A. Compute a static, connected topology:

We assume that each ToR has $\mathcal{N}$ transceivers, among them at most n transceivers are reserved for generating a static, connected topology, whereas others are configured dynamically based on the traffic demands. The static topology is constructed in such a way that the optical hop-counts of all the source-destination pairs are minimized under the constraint that the incident edges at any switch is at most n. Any regular, connected graph of bounded node degree (such as cube, butterfly etc) can be constructed for this static optical backbone. However, in our simulations, we generate multiple random graphs, while keeping the number of transceivers below n and pick the one which is connected and has lowest overall hop count. We denote this static graph of optical links as G_l, whereas the actual physical topology is denoted as G_p. While generating G_l, we also ensure that the physical path

length, corresponding to an optical link, is less than some threshold τ. The main intuition behind this is that the racks mostly transfer data in their close vicinity. We assume τ = 10 for our simulation in this paper. As this static configuration is done offline, this brute force technique is reasonable.

The next step is to assign the wavelengths to the edges of G_l, such that no two edges passing through a fiber have the same wavelength. To do this, we first construct a *conflict* graph $\bar{G}_l$, where (a) the edges in G_l are represented as the vertices of $\bar{G}_l$, and (b) there exists an edge in between two vertices in $\bar{G}_l$ if they have a common physical edge in G_p. Hence, if there is a link between two vertices in the conflict graph $\bar{G}_l$, then those two vertices share at least a common physical edge, thus we have to assign different wavelengths to these two vertices. This is similar to *vertex coloring problem* [16], which is a NP-complete problem, but fast heuristics are known. We use Brelazs DSATUR [17] coloring scheme to assign the colors to the vertices of $\bar{G}_l$. The colors are then mapped to appropriate wavelengths, which are then assigned to the static transceivers of the switches. Thus any ToR can transfer its data to another one, via multi-hop optical communication by using this static topology.

TABLE I
TABLE OF NOTATIONS

		Indices
i, j	$\triangleq$	Index for intermediate nodes (1, ..., V)
f	$\triangleq$	Index for flows (1, ..., F)
k	$\triangleq$	Index for the candidate routes (1, ..., $\mathcal{K}$)
t	$\triangleq$	Index for wavelengths (1, ..., W)
		Binary Variables
$\mathcal{S}_f^i \in 0,1$	$\triangleq$	Whether or not node i is the source of the f-th flow
$\mathcal{D}_f^i \in 0,1$	$\triangleq$	Whether or not node i is the destination of the f-th flow
$S_f \in 0,1$	$\triangleq$	Whether or not the f-th flow is admitted
$P_f^k \in 0,1$	$\triangleq$	Whether or not the k-th path is chosen corresponding to the f-th flow
$\mathring{P}_f^k \in 0,1$	$\triangleq$	Whether or not the k-th *optical* path is chosen corresponding to the f-th flow
$y_{ft}^k \in 0,1$	$\triangleq$	Whether or not the k-th path for the f-th flow is assigned wavelength λ_t
$\mathring{y}_{ft}^{ijk} \in 0,1$	$\triangleq$	Whether or not the optical link i-j of the k-th path for the f-th flow is assigned wavelength λ_t
$\mathring{w}_t^{ij} \in 0,1$	$\triangleq$	Whether or not the optical link i-j is assigned wavelength λ_t
		Other Variables
R_f	$\triangleq$	Flow rate requested by the f-th flow
$\mathcal{K}$	$\triangleq$	Number of candidate routes in between the SD racks
$\mathcal{N}$	$\triangleq$	Number of transceivers per ToR
W	$\triangleq$	Number of wavelengths

B. Routing and wavelength assignment for direct optical flows

The ToR switches send their significant flows as well as their predicted flow rates to the centralized controller, which periodically assigns the routes as well as their wavelengths based on the overall traffic demands. In the centralized version of this *direct flow with wavelength assignment (DFWA)*, the central manager stores $\mathcal{K}$ shortest paths in between all the SD pairs, and then chooses the routes among those $\mathcal{K}$ paths, to maximize the number of requests served. We use Yen's algorithm [18] to find out the $\mathcal{K}$ shortest paths in between any SD pair. The notations are summarized in Table I. Assume that x_f^{ijk} = 1 when the k-th path from requested flow f, goes through link $i \rightarrow j$, and 0 otherwise. The problem is to maximize the total demands satisfied across all the racks through direct lightpaths, from a list of given F flow requests as well as their $\mathcal{K}$ candidate paths. We assume that all the flows have demands which is less than the line speed of a single ToR port. If an ToR wants to initiate a connection which is w times the line speed of a single port, it requests w flows, each one has a demand equal to the line rate. This constrained optimization can be captured by the following integer linear program (ILP).

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{f=1}^{F} S_f \times R_f \\
\text{subject to} \quad & \sum_{f=1}^{F} S_f . \left(\mathcal{S}_f^i + \mathcal{D}_f^i\right) \leq \mathcal{N} \ \ \forall i \\
& \sum_{f=1}^{F} \sum_{k=1}^{\mathcal{K}} y_{ft}^k . x_f^{ijk} \leq 1 \ \ \forall i, \forall j, \forall t \\
& P_f^k = \sum_{t=1}^{W} y_{ft}^k \ \ \forall f, \forall k \qquad \sum_{k=1}^{\mathcal{K}} P_f^k = S_f \ \ \forall f
\end{aligned}
\tag{1}
$$

The objective function of the above ILP is to maximize the total inter-rack demands that can be admitted using single-hop optical communication. The first constraint states that the number of active transceivers for each ToR switch is limited by its available set of transceivers. The second constraint states that a link/fiber cannot carry more than one wavelength in the same direction. The third constraint says that a chosen light-path is assigned at most one wavelength. The fourth constraint ensures that each chosen flow is assigned a lightpath among its candidate routes. The problem formulation in equation(1) assumes bidirectional inter-rack communications, i.e. if there is a lightpath from $a \rightarrow b$, then the reverse path $b \rightarrow a$ is also *reserved* for communicating in the reverse direction. Thus for bidirectional communication x_f^{ijk} is symmetric. However, the model can also be extended for unidirectional inter-rack communications, in that case x_f^{ijk} is going to be asymmetric. For our simulation, we assume a bidirectional traffic pattern where the flow rates in both directions are same, which ensures a complete *bijective* traffic pattern.

TABLE II
COMPARISON ON OPTIMAL DFWA AND IT'S APPROXIMATION VERSION.

Topology	DFWA (LP relaxed)		DFWA (heuristic)	
	Flows admitted	Time (secs)	Flows admitted	Time (secs)
6×6	576	19.6	573	0.146
7×7	784	98.3	780	0.308
8×8	1024	482	1022	0.527
10×10	1600	7352.3	1594	1.5

The above integer linear program (ILP) is a special case of routing and wavelength assignment problem in optical networks, which is proven to be NP-complete [19]. Because

of its complexity, typical solvers takes a significant amount of time of generate the solution of such ILPs. We consider a DCN where the ToR switches are connected in a torus architecture. We assume that each ToR has $\mathcal{N}$ = 32 transceivers, and the total number of wavelengths W = 40. We assume that each ToR generates one unit of flow to all the other ToR switches. $\mathcal{K}$ is assumed to be 1, and all the flows have the same demands. We solve this problem using GLPK [20], which is a package for solving linear programming (LP) and mixed integer programming (MIP) related problems. The solver runs on a Intel Core i7 @ 3.4 GHz processor, with 16 GB RAM. Table II shows the solution and the computation time of the LP-relaxed version of the above ILP, which also yields the upper-bound of the optimal solution. Considering its complexity and computation time, we propose a heuristic to solve the above ILP.

Algorithm 1 Direct flow and wavelength assignment (DFWA) scheme

```
INPUT : ℝ = f_i ∀i is the set of all requests, 𝒦 lightpaths corresponding to each request, the available wavelengths for each request.
OUTPUT : Chosen lightpaths and their wavelength assignment.
visited[i] = 0, ∀ i ∈ ℝ;
Sort f_i = (s_i, d_i), ∀ i ∈ ℝ in decreasing order of their flow priorities/traffic demands;
while ℝ ≠ NULL OR there is at least an available wavelength for a lightpath in ℝ do
    ℙ = NULL;
    for i = 1; i < |ℝ| ; i++ do
        if visited[i] == 0 AND (s_i, d_i) have available transceivers then
            Choose the shortest lightpath s_i → d_i that has some common VAW with the lightpaths of ℙ;
            if Such a lightpath is found then
                ℙ = ℙ ∪ f_i;   ℝ = ℝ \ f_i;   visited[i] = 1;
            end if
        end if
    end for
    Assign ℙ with the first common VAW, say λ_j;
    Remove λ_j from all the physical edges of the lightpaths in ℙ;
end while
```

An approximation algorithm: The idea behind this heuristic is to find the maximum edge disjoint paths (MEDPs) [21] corresponding to the requests and assign an available wavelength that is common to all those paths. Then another MEDPs from the remaining requests are constructed and an available wavelength is assigned. The same process is continued until all the requests are met or the wavelengths are exhausted. As the MEDP problem is an NP-hard problem [21], we propose an approximation scheme which is shown in Algorithm 1. Initially all the flows are marked as unvisited. Then the requested flows are visited in decreasing order of their traffic demands. In case of a tie, the tie is broken by the hop-counts. In presence of multiple priorities, higher priority flows are visited ahead of others. In Algorithm 1, $\mathbb{R}$ is the set of all unvisited flows and $\mathbb{P}$ records the set of flows that have common available wavelengths. In each visit of flow f_i, the shortest path in between $s_i \rightarrow d_i$ is chosen, such that the path has some common VAWs with the existing lightpaths in $\mathbb{P}$. This set of lightpaths in $\mathbb{P}$ are assigned a wavelength that is common *first* VAW to all of them. This process goes on until all the requests are visited or no wavelength is available for the remaining flows.

To compare its approximation accuracy and computation time, we implemented it in MATLAB R2015a [22] and have run the simulation in similar setting. The results are shown in Table II, which shows that the proposed heuristic is several order faster than the ILP, but still the number of admitted flows is very close to that of the optimal solution. We believe that the computation can be made even faster, if the flows that are independent to each other can be computed concurrently.

Observation 1: *By using DFWA, all the flows can be assigned a direct optical hop, as far as (a) at least one transceiver at the SD switches is available, (b) the number of available channels are infinite.*

Infact in [23], the authors have claimed that with orthogonal frequency-division multiplexing (OFDM) transponders and coherent optics, it is feasible to have 1000 channels, each at 40Gbs, over a short distance fiber. In such a situation with large number of wavelengths, DFWA can accommodate direct optical paths for almost all the flows, as far as the transceivers at the SD ends are available.

C. Compute the routes of the remaining flows:

The remaining flows in $\mathbb{R}$ are assigned multi-hop optical routes on top of the connected topology made by the existing static lightpaths, as well as the admitted dynamic lightpaths. For the remaining flows, the routes are chosen among $\mathcal{K}$ candidate *optical* routes, that are least congested or utilized. We assume that $\mathring{x}_f^{ijk}$ = 1 when the k-th lightpath from requested flow f, goes through optical link $i \rightarrow j$, and 0 otherwise. The following ILP can be developed to model this optimization problem.

$$
\begin{aligned}
\text{Minimize} \quad & \mathbb{T} \\
\text{subject to} \quad & \sum_{k=1}^{\mathcal{K}} \mathring{P}_f^k = 1 \quad \forall f \\
& \mathring{P}_f^k . \mathring{x}_f^{ijk} = \sum_{t=1}^{W} \mathring{y}_{ft}^{ijk} \quad \forall f, \forall i, \forall j, \forall k \\
& \mathring{y}_{ft}^{ijk} \leq \mathring{w}_t^{ij} \quad \forall f, \forall i, \forall j, \forall k, \forall t \\
& \sum_{f=1}^{\hat{F}} \sum_{k=1}^{\mathcal{K}} \mathring{y}_{ft}^{ijk} . \mathring{x}_f^{ijk} . \mathring{w}_t^{ij} . R_f \leq \mathbb{T} \quad \forall i, \forall j
\end{aligned}
\tag{2}
$$

The objective is to minimize the maximum load at any lightpath. The first set of constraints state that one of the candidate routes are chosen for all the flows. The second constraints ensure that all the optical links, corresponding to a flow f, are assigned one wavelength. The third set of constraints state that an optical link i-j of flow f, uses wavelength λ_t only if that wavelength is available at i-j. The final set of constraints say that the total load at any optical link is less than the objective. The above ILP is NP-hard, because of its similarity with the network-flow problem. Considering its complexity, we propose a simple heuristic, which is mentioned in Algorithm 2, and is self-explained. We also compare the accuracy of this simple heuristic, against its ILP version, on the static optical backbone developed in section III-A. We assume that each ToR sends

one unit of traffic to all others. The results are shown in Table III, which confirms that the proposed heuristic is fairly accurate, but still several times faster compared to the optimal ILP.

Algorithm 2 Routing to minimize the maximum load

```
INPUT : ℝ = f_i ∀i is the set of all requests not assigned in DFWA, 𝒦 lightpaths corresponding to each request.
OUTPUT : Chosen lightpaths corresponding to the remaining flows.
while ℝ ≠ NULL do
    for i = 1; i < |ℝ| ; i++ do
        Choose the lightpath for flow s_i → d_i that is least loaded;
        Update the load of the corresponding lightpaths;
        ℝ = ℝ \ f_i;
    end for
end while
```

TABLE III
COMPARISON ON OPTIMAL MIN-MAX LOAD AWARE ROUTING AND IT'S APPROXIMATION VERSION.

Topology	min-max load (LP relaxed)		min-max load (heuristic)	
	Max. load	Time (secs)	Max. load	Time (secs)
4×4	23	1	23	.27
5×5	43	5.5	43	1.43
6×6	67	5345.1	67	5.89

IV. SIMULATION RESULTS

We evaluate the performance of our proposed flow scheduling and wavelength assignment schemes on RODA using Matlab simulations. Unless otherwise mentioned, we assume 64 ToR switches that are placed in an uniform 8×8 torus architecture. This architecture can accommodate 2688 servers on a 64 racks, assuming 42 servers fit in a rack. Each ToR has I = 64 ports, 32 of them are connected to the servers through direct or hierarchical connection, whereas 32 others are connected to the optical transceivers. This needs four 1×3 WSSs and another four 1×32 WSSs per ToR, that are connected to the transceivers. $\mathcal{K}$ is assumed to be 5. All the flow requests are assumed to be of one unit.

Traffic patterns: We use the following synthetic traffic patterns for our evaluation purpose:

1) **Random SD pair:** We randomly choose 600 random inter-rack flows, i.e. on an average a rack transfers data to ∼15% of the number of racks.

2) **Hotspot based:** Many real study [4] of DCN traffic pattern reveals that in a DCN, only few ToR switches are hot, and most of their traffic goes to a few other ToR switches. We model this type of traffic pattern by randomly choosing 15 hotspots (∼25% of the number of racks), each one sends data to 10 other randomly chosen ToR switches.

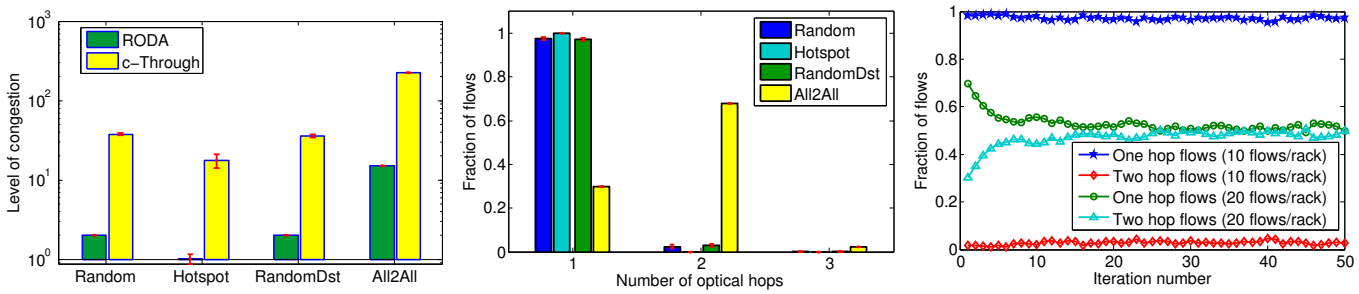
3) **Random destination (RandomDst):** Unless otherwise mentioned, in this traffic pattern, we assume that each one of the ToR switches sends data to 10 randomly chosen ToR switches among others, which is ∼15% of the total number of racks. For some simulations, we also consider 20 flows/rack with RandomDst traffic pattern.

4) **All to all (All-2-All):** In this traffic pattern, all the racks transmit data to all other racks, which results in a highly congested scenario.

We built a simulator in MATLAB, that runs iteratively. In each iteration, flows from the previous iteration stays with 50% probability. The remaining slots are filled up with the newly generated flows, based on the chosen traffic pattern. We use *non-preemptive* scheduling policy for our simulations, i.e. the routes and assigned wavelengths of the existing flows are not changed in any new iteration, however infrequent preemption can be allowed in real scenarios. All the simulations are averaged over fifty such iterations.

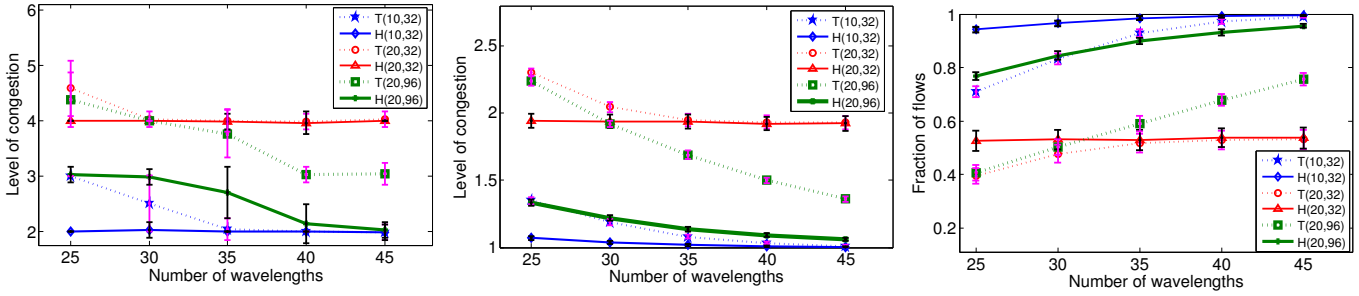
Comparison with different traffic patterns: We first define *Level of congestion (LoC)* of a lightpath, as the number of flows that are carried on that lightpath, i.e. the load of the lightpath. As an example in Fig. 2(a), the LoC for lightpath $b \rightarrow c$ is two, as it is carrying both a and b's data to c. The worst LoC is the maximum of LoCs of all the lightpaths. Fig. 5(a) compare the worst LoCs for all different traffic patterns, where the ToR switches are placed in a 8×8 torus topology. The number of wavelengths are assumed to be 40, which is the approximate number of channels allowed in the C-band with 0.4 nm channel spacing [8]. From Fig. 5(a), we observe that Random and RandomDst performs very similar to each other, as their overall number of flows are analogous. In both patterns, more than 97% of the flows are scheduled by direct optical inter-rack connection, as shown in Fig. 5(b). In Hotspot scenario, the worst LoC is reduced by almost 100%, compared to Random and RandomDst, and all the flows are assigned direct optical flows. In All-2-All traffic pattern, the worst LoC is increased by almost seven times, compared to Random and RandomDst patterns, because of accommodating a significantly large number of flows. Because of such a high volume of data flow, ∼68% of the traffic are transferred in two hops, after all the single hops are used up. Fig.5(c) shows the fraction of flows, that are routed through one or multiple optical hops, in different iterations, in case of RandomDst traffic pattern. From Fig.5(c), we can observe that even if non-preemptive scheduling is adopted, most of the traffic can still be scheduled using direct optical communications, when the number of flows/rack is limited to 10. We then increase the number of flows/rack to 20, and have noticed that the number of direct flows starts reducing with successive iterations. Thus we can infer that with heavy inter-rack load, infrequent preemption may be needed. This can be done either periodically, or when a certain fraction of flows starts through multiple hops.

We also compare our proposed RODA, with a part-time optical DCN architecture named c-Through [7]. We assume that one core switch is connected to 8 aggregate switches, each one of them is then connected to 8 ToR switches. All the electrical links in between the core and the aggregate switches are assigned bandwidths of four units, whereas the links connected in between the aggregate switches and the ToR switches have unit bandwidth. The LoC is defined as the number of flows assigned per unit bandwidth. In the optical part, a ToR can be connected to atmost one other ToR through an optical MEMS. Given the traffic matrix, we first run maximum matching [24] to schedule the optical flows,



(a) (b) (c)

Fig. 5. Comparison of the proposed flow scheduling scheme with different traffic patterns, (a) variation of LoCs, and (b) the fraction of the traffic transferred by different number of optical hops. (c) Fraction of traffic routed by different number of hops, in successive iterations.



(a) (b) (c)

Fig. 6. Comparison of (a) maximum LoCs, (b) average LoCs and (c) the fraction of the traffic transferred by direct optical hops, in torus and hypercube RODA structure. T(i, j) and H(i, j) are torus and hypercube topology with i flows/rack and j transceivers/rack respectively.

rest of the flows are satisfied by the electrical subnetwork. We observe that RODA reduces the worst LoC by factors of 15-18, compared to c-Through. The main reason is the use of optical WDM to multiplex several lightpaths in a single fiber, which significantly reduces the LoC.

Comparison with different number of wavelengths: We vary the number of wavelengths from 25 to 45, the effect of this varying wavelengths is reported in Fig. 6. We use RandomDst for these set of graphs. From Fig. 6 we can observe that, in case of 10 flows/racks, with the increase in number of wavelengths from 25 to 45, ∼30% more flows can be routed through direct optical lightpaths, due to the increasing ability to multiplex more flows through an optical link. This also reduces the average load by ∼25%, and the maximum load by ∼33%, for a 64 nodes torus topology.

Effect of different topologies: We also compare the RODA architecture with different interconnecting topologies. Fig. 6 shows the comparison of RODA where the ToR switches are connected in torus and hypercube, for a RandomDst traffic generation pattern. A hypercube graph Q_n is a regular graph with 2^n vertices, $2^{n-1}.n$ edges, and degree n. In our case, n = 6, which results in 64 ToR switches. From Fig. 3 we can observe that with lesser number of wavelengths, the hypercube structure significantly reduces the LoCs, which admits more direct optical flows. In presence of 25 optical wavelengths and 10 flows/rack, the number of direct flows is ∼25% more in hypercube topology, while the LoC goes down by ∼50%. The primary reason is that hypercube structure has more vertex degree, compared to torus, which results in more edges. As an example, with a 64 node hypercube has a vertex degree of 6, with 192 edges, while a 8×8 torus has a vertex degree of 4, that results in a total of 128 edges. The network diameter is also 6 in case of a hypercube, compared to 8 for torus. Thus a hypercube RODA can distribute the traffic across more edges, which results in reduced LoC, and the increase in direct optical flows. With the increase in number of wavelengths, the difference between these two topologies starts shrinking, as more number of flows are routed and multiplexed through direct optical hops, for both torus and hypercube.

To show the improvement of a hypercube topology further, we increase the per-rack flows from 10 to 20, which results in the increase in ∼12% of direct flows, and a reduction of ∼10% of network load, compared to torus, in presence of 25 wavelengths. Still we observe that the difference starts shrinking, with the increase in number of wavelengths. The primary reason for this is that, the number of transceivers/rack is limited to 32, which puts a limit on the number of light-paths. Thus even if in case of a more connected topology in hypercube, the number of lightpaths becomes similar to that of a torus topology. We then increase the number of transceivers/rack to 96. As a result, a clear improvement is observed for a hypercube architecture, as shown in Fig. 3. With more number of transceivers/rack, the hypercube architecture increases the direct optical hops by ∼20-35%, whereas the network congestion is reduced by ∼23-40%, which shows a clear advantage of a highly connected DCN topology. Also we can observe from Fig. 3(c), that with higher number of wavelengths and better connectivity, almost all the flows can be routed through direct optical hops, which validates our intuitive claim (observation 1) in section III.

Comparison of different flow priorities: We bring the flow priorities into account by considering 50% of the flows as

high priority (HP) traffic, whereas the other half is considered as low priority (LP). The HP flows are first considered for assigning direct flows. We simulate such scenario in a 64 node torus architecture, and the results are shown in shown in Fig. 7. We assume RandomDst traffic pattern for Fig. 7. From Fig. 7 we can observe that with 10 flows/rack, almost all the flows are routed through direct hops. For 20 flows/rack and 32 transceivers/ToR, ∼62% of the HP flows follow direct hops, whereas other are routed through two-hops, due to the limited number of transceivers per rack. With increased number of per-rack transceivers, almost 85% of the HP flows are direct, whereas ∼65% of the LP flows are direct too.

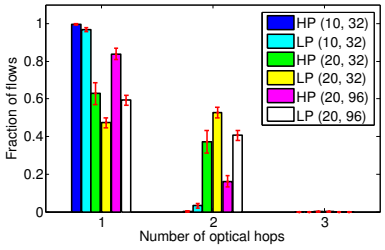


Fig. 7. Comparison of HP vs LP traffic. (i, j) is defined by i flows/rack and j transceivers/rack.

Comparison of different network sizes: We next vary the network size from 8×8 to 12×12 torus architecture. The 12×12 architecture accommodates 144 racks and 6048 servers. We assume that each ToR is connected to 32 transceivers. We also assume 10 flows/rack for RandomDst traffic pattern. The worst LoCs corresponding to these topologies are shown in Fig. 8. From Fig. 8 we can observe that for Random and Hotspot traffic scenarios, the worst LoCs remains almost similar for all network dimensions, mainly because the total number of flows remains the same for these two traffic patterns. For RandomDst and All2All traffic patterns the worst LoC starts increasing due to the increase in total number of flows. For RandomDst pattern the LoC increases by ∼50%, whereas in case of All2All it increases by a factor of 4, when the topology varies from 8×8 to 12×12 torus. In case of All2All traffic pattern the number of inter-rack flows increase significantly as the network grows, which results in a drastic increase in LoC as seen from Fig. 8.

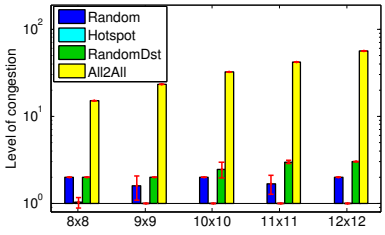


Fig. 8. Comparison of worst LoCs with different network size.

V. Related Works and Discussions

A. Related Works

The design of a flexible, cost-effective and efficient DCN architecture is well researched, and so a large number of proposals exist in the literature. In this section, we classify them in the following categories, and then present each of them separately.

Electrical switching based DCN: In electrical switching based DCN [1], [2], [3], the ToR switches are connected by multi-tier, tree-based, interconnection architectures. Due to the hierarchical architecture, the intrinsic limitation of a typical tree based DCN architecture is its over-subscription problem. In order of ensure that the aggregation/core layer of the DCN is not oversubscribed, a significantly large number of switches and physical wires are needed [8], which drastically increases the hardware cost and wiring complexity.

Wireless DCN: In [4], [25] the authors proposed a DCN architecture that utilizes the 60 GHz wireless spectrum for providing supplemental data transfer through wireless medium, in addition to the traditional wired links. The major drawbacks of this 60 GHz wireless communications is that (a) the wireless links are limited by line-of-sight and so can be blocked by small obstacles, and (b) potential wireless interference severely limits concurrent transmissions in a dense DCN. To alleviate these problem, in [26] the authors proposed a design that puts mirrors on the DC ceilings, from where the wireless signals are reflected to establish an indirect line-of-sight communication in between any two racks in a data center. In the similar line, a free-space-optics (FSO) based wireless DCN is designed in [27], where the optical signals are bounced back to the receiver, after reflecting from the ceiling mirrors. Compared to the wireless/RF technologies, the FSO has the advantages of (a) lower interference, (b) longer range, and (c) higher bandwidth.

Hybrid electrical-optical DCN: Hybrid electrical-optical DCN architecture is proposed in c-Through [7], Helios [8] by exploiting the advantages of low-bandwidth, fast electrical switching and high-bandwidth, slow optical switching. In these proposals, the delay-sensitive, bursty data flows are transmitted through the electrical switches, whereas the relatively steady and longer flows are routed through reconfigurable optical switches. The key limitation of these schemes is that, at a time, a ToR can *only* connect to one other ToR switch, and the communication among them is limited by the line speed of a single port.

All optical DCN: In OSA [6], the authors proposed an all-optical switching architecture, where the ToR switches are connected to a MEMS switch, through optical MUX/DEMUX and switching components. This scheme dynamically adjusts the capacities of the optical links, to satisfy the changing traffic demands through the use of WSS. They have also used optical circulators to accommodate simultaneous bidirectional transmission over the circuits. Communication in between any two ToR switches is established by stitching multiple optical hops. In [10], the authors proposed a MEMS free DCN architecture, to overcome the limitations of low port density of a MEMS switch. They have proposed a DCN architecture, named WaveCube, where the ToR switches are placed in a multi-dimensional cube structure, and have developed a flow scheduling scheme that adjusts the link bandwidths dynamically based on the traffic demands.

The RODA architecture proposed in this paper is all-optical and MEMS-free, which is mostly related to and influenced by WaveCube. However, our architecture uses *tunable* transceivers, that switch in between different wavelengths

dynamically, which brings more flexible and dynamic lightpath creation. In addition to that, we developed a reconfigurable ADMUX unit that has the capability to pass/drop/deflect lightpaths in between the source and destination switches, which (a) provides more flexibility in assigning the lightpaths, and at the same time (b) can utilize the ToR transceivers efficiently.

B. RODA flexibility vs cost

One practical concern of RODA is its cost, as it uses a large number of expensive optical components for flexible and dynamic lightpath assignment among the source-destination racks. The primary cost component of RODA is the large number of reconfigurable and expensive WSS ports ($1000 per port as reported in [6], [10]) to drop different wavelengths dynamically, which makes RODA more expensive than other less-flexible architectures. However, cost is a function of the technological development, and because of rapid decrease in Optics costs (e.g., 90% cost reduction in optical transceivers in the last decade [28]), Optical networks is expected to become more affordable in the future. Interestingly, RODA has a built-in tradeoff that can be exploited to deal with the cost issue. RODA includes a statically routed optical network for transfer of smaller flows which can be built using cheap (non-reconfigurable) DEMUX units. Using more ports for the static network in the short-run provides the required tradeoff between flexibility and cost.

VI. Conclusions

In this paper, we proposed a reconfigurable DCN architecture, named RODA, based on optical WDM, and devised a flow admission scheme for rack-to-rack communication on top of RODA. We showed that through this flow scheduling, we can dynamically admit most of the inter-rack data flows using direct optical hops, while reducing the amount of network congestion. In future, we plan to develop a partly *distributed* inter-rack communication scheme on RODA, which will significantly reduce the dependency on a centralized controller. Although major/large flows in a DCN need to be scheduled in a *centralized* fashion, to avoid localized congestions, black-holes or deadlocks, some of the small flows can be routed by a distributed approach. This reduces the dependency and control message exchanges in between the switches and a centralized controller. Whereas such concepts are explored in DevoFlow [29], [30] they are limited to electrical DCNs, where insignificant/small flows are routed using schemes like equal-cost multi-path (ECMP) [31]. In an optical DCN, exploring such schemes are even more challenging, as the optical bursts are expected to be forwarded with minimum intermediate O-E-O conversion. In future we also plan to extend the RODA architecture using free space optics (FSO), which will greatly reduce the wiring complexity and cost of our proposed architecture.

References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *ACM SIGCOMM*, 2008, pp. 63–74.

[2] A. G. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: a scalable and flexible data center network," *Commun. ACM*, vol. 54, no. 3, pp. 95–104, 2011.

[3] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "Bcube: a high performance, server-centric network architecture for modular data centers," in *ACM SIGCOMM*, 2009, pp. 63–74.

[4] S. Kandula, J. Padhye, and P. Bahl, "Flyways to de-congest data center networks," in *ACM HOTNETS*, 2009.

[5] K. J. Barker, A. Benner, R. Hoare, A. Hoisie, A. K. Jones, D. K. Kerbyson, D. Li, R. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. Stunkel, and P. Walker, "On the feasibility of optical circuit switching for high performance computing systems," in *ACM/IEEE Conference on Supercomputing*, 2005.

[6] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, "OSA: an optical switching architecture for data center networks with unprecedented flexibility," *IEEE/ACM Transactions on Networking*, vol. 22, no. 2, pp. 498–511, 2014.

[7] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. P. Ryan, "C-through: part-time optics in data centers," in *ACM SIGCOMM*, 2010, pp. 327–338.

[8] 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 *ACM SIGCOMM*, 2010, pp. 339–350.

[9] G. Porter, R. D. Strong, N. Farrington, A. Forencich, P. Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat, "Integrating microsecond circuit switching into the data center," in *ACM SIGCOMM*, 2013, pp. 447–458.

[10] K. Chen, X. Wen, X. Ma, Y. Chen, Y. Xia, C. Hu, and Q. Dong, "Wavecube: A scalable, fault-tolerant, high-performance optical data center architecture," in *IEEE INFOCOM*, 2015.

[11] B. Montreuil, "Toward a physical internet: meeting the global logistics sustainability grand challenge," *Logistics Research*, vol. 3, no. 2-3, pp. 71–87, 2011.

[12] A. Pal and K. Kant, "Efficient distribution of food packages in fresh food physical Internet," in *Technical Report, Temple University*, 2015.

[13] http://www.router-switch.com/Price-cisco-optics-modules_c8.

[14] http://www.fiberstore.com/Optical-Circulator-Tutorial-aid-455.html.

[15] R. A. Berry and E. Modiano, "On the benefit of tunability in reducing electronic port counts in WDM/TDM networks," in *IEEE INFOCOM*, 2004.

[16] http://en.wikipedia.org/wiki/Graph_coloring.

[17] http://armanboyaci.com/?p=487.

[18] http://www.mathworks.com/matlabcentral/fileexchange/32513-k-shortest-path-yen-s-algorithm.

[19] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: an approach to high bandwidth optical wan's," *IEEE Transactions on Communications*, vol. 40, no. 7, pp. 1171–1182, 1992.

[20] https://www.gnu.org/software/glpk/.

[21] P. Manohar and V. Sridhar, "Routing, wavelength assignment in optical networks using an efficient and fair EDP algorithm," in *ICCS*, 2004, pp. 1178–1184.

[22] http://www.mathworks.com/products/matlab/.

[23] A. Gumaste, B. M. K. Bheri, and A. Kshirasagar, "FISSION: flexible interconnection of scalable systems integrated using optical networks for data centers," in *IEEE ICC*, 2013, pp. 3963–3968.

[24] http://www.mathworks.com/matlabcentral/fileexchange/.

[25] D. Halperin, S. Kandula, J. Padhye, P. Bahl, and D. Wetherall, "Augmenting data center networks with multi-gigabit wireless links," in *ACM SIGCOMM*, 2011, pp. 38–49.

[26] 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 *ACM SIGCOMM*, 2012, pp. 443–454.

[27] N. H. Azimi, Z. A. Qazi, H. Gupta, V. Sekar, S. R. Das, J. P. Longtin, H. Shah, and A. Tanwer, "Firefly: a reconfigurable wireless data center fabric using free-space optics," in *ACM SIGCOMM*, 2014, pp. 319–330.

[28] J. Berthold, "Optical networking for data center interconnects across wide area networks," 2009.

[29] A. R. Curtis, J. C. Mogul, J. Tourrilhes, P. Yalagandula, P. Sharma, and S. Banerjee, "Devoflow: scaling flow management for high-performance networks," in *ACM SIGCOMM*, 2011, pp. 254–265.

[30] J. C. Mogul, J. Tourrilhes, P. Yalagandula, P. Sharma, A. R. Curtis, and S. Banerjee, "Devoflow: cost-effective flow management for high performance enterprise networks," in *ACM HotNets*, 2010.

[31] http://www.ieee802.org/1/pages/802.1bp.html.