IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2017, accepted June 8, 2017, date of publication June 22, 2017, date of current version July 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2716345

Dynamic Flow Scheduling With Uncertain Flow Duration in Optical Data Centers

TRAM TRUONG-HUU, (Senior Member, IEEE), MOHAN GURUSAMY, (Senior Member, IEEE), AND SHARMILA TRANQUEBAR GIRISANKAR
Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583

Corresponding author: Tram Truong-Huu (eletht@nus.edu.sg)

This work was supported by Singapore Ministry of Education, Academic Research Fund Tier 2, Grant No. MOE2013-T2-2-135, NUS WBS No. R-263-000-B11-112.

ABSTRACT Optical switching based on wavelength division multiplexing has become a promising network technology to scale the performance of data centers. It provides high bisection bandwidth with low power consumption and low complexity of network wiring. However, it raises new challenges for the flow scheduling problem due to the dynamic arrival of traffic flows with unknown service duration combined with the circuit-switched nature of optical networks and wavelength continuity constraint. While the knowledge of flow service time helps to use resources in a better way to increase the revenue, in practice, the service time cannot be accurately specified. In this paper, we address the problem of flow scheduling in optical data centers considering the above challenges. We first develop an optimization formulation using Markov decision process that can estimate the flow termination time and revenue for cloud providers in a long run under the uncertainty in flow service time. Since solving the optimization formulation is mathematically intractable, we then develop heuristic scheduling algorithms for both scenarios: with known and with unknown flow service time. We use a probabilistic model to address the uncertainty due to unknown flow service time. We design a flow scheduling framework that integrates the proposed algorithms to perform flow scheduling in optical data center networks. We evaluate the proposed algorithms through comprehensive simulations and compare their performance against that of a baseline algorithm. The results show that the proposed algorithms achieve significant performance improvement compared with the baseline algorithm.

INDEX TERMS Traffic flow scheduling, WDM-based optical networks, optical data centers, software defined networks, Markov decision process, uncertainty in flow service time.

I. INTRODUCTION

This decade has witnessed the evolution in the design of data center networks towards hybrid design approaches, in which both wavelength division multiplexing (WDM) based optical-switched network and electrical packet-switched network co-exist to serve different types of traffic demands [1]–[6]. This evolution has been driven by the growth of the traffic demands in data centers, in which thousands of servers are hosted to meet the growing number of online services and internal data center applications such as streaming video, healthcare and government systems, data backup and virtual machine migration [7]. In comparison with electrical packet-switched networks, WDM-based optical networks provide very high bisection bandwidth but do not require several layers of electrical packet switches as in FatTrees [8], consume low power and reduce the cabling complexity. The key concept of optical networks is a *lightpath* that is defined as an all-optical connection routed in the optical domain along one or more fiber links [9]. At a given time instant, a Top-of-the-Rack (ToR) switch can simultaneously connect to a number of ToR switches decided by the number of wavelengths carried by the optical fiber. Dynamic reconfiguration of lightpaths brings in flexibility for network management and traffic engineering but suffers from inherent reconfiguration overhead (in the order of ms or μs). This makes WDM-based optical networks suitable only for circuit switching but not for packet switching.

Given a traffic flow request between two ToR switches, the flow scheduling problem is defined as the decision of creation of a lightpath between the two ToRs to carry the traffic flow, i.e., switching the optical circuit at a given time instant. Given the absence of wavelength converters, the

lightpath between two ToR switches traversing through the optical switch needs to use the same wavelength on all fibers traversed. This is called wavelength continuity constraint. The service duration of a lightpath depends on the service duration of the hosted traffic flow. The optical circuit (lightpath) is thus released when it is no longer used by any flow. In some scenarios, service duration of traffic flows might be unknown. Due to the re-configuration of lightpaths, a traffic flow might also be migrated from one wavelength to another wavelength during the flow lifetime. The problem of flow scheduling in optical cloud data centers is therefore much more challenging due to the wavelength continuity constraint and dynamic arrival of traffic flows. We need to consider not only the flow admission but also the wavelength assignment for the lightpaths allocated to traffic flows. A sub-optimal wavelength assignment will lead to the disconnectivity of ToRs in future, low resource (wavelength) utilization and affects the overall revenue of commercial cloud providers in a long run.

The problem of traffic scheduling in optical data centers has received attention from researchers recently [10]–[15]. The work presented in [10] considered packet-level scheduling in optical data center networks, which requires more frequent reconfigurations of the network logical topology. The work presented in [11] considered a different objective that aims at maximizing the lifetime of the optical switches rather than flow admission control and wavelength assignment. Further, the existing works did not consider the uncertainty in flow service time that has a significant impact on wavelength utilization, connectivity of the ToRs and revenue of cloud providers. In our previous work [12], we carried out a preliminary study of the flow scheduling problem in optical data centers, assuming that flows can provide their service duration at the submission instant. We extend this work further to consider the case where flow service time is not specified. We develop probability-based model to address the uncertainty in flow service time, develop new algorithms and carry out new sets of simulation experiments.

In this paper, we address the problem of flow scheduling in optical data center networks considering the uncertainty in flow service time. We first develop an optimization programming formulation that aims at maximizing the total revenue of cloud providers in a long run while satisfying the wavelength continuity constraint and bandwidth capacity constraint. The optimization formulation is modeled as a Markov Decision Process (MDP) that can estimate the expected revenue of cloud providers based on the probability model of flow service time. Solving the optimization programming formulation is computationally prohibitive due to the large size of problem even for a small data center and small number of input flows. We thus develop heuristic algorithms for both scenarios: with known and with unknown flow service time. We introduce the concept of *congestion factor* of a pair of ToRs, i.e., the source and destination of a flow. It is computed based on the number of flows that are accommodated between the pair of ToRs and the number of common wavelengths available on the fibers connecting the two ToRs to the core optical switch. We note that the availability of common wavelengths is affected by the wavelength continuity constraint. Accommodating more flows or creating new lightpaths between two ToRs that have high congestion factor may cause disconnectivity of ToRs in future, thereby increasing the rejection ratio.

Based on the congestion factor concept, we propose three heuristic algorithms: (i) Least Congestion and Probability-based Service Time algorithm (LC-PBST), (ii) Congestion-Based Round-Robin algorithm (CB-RRA), and (iii) Least Congestion and Shortest Service Time First algorithm (LC-SSTF). Algorithms LC-PBST and CB-RRA consider the scenario where flow service time is unknown. Algorithm LC-PBST assumes that flow service time follows a probabilistic model that will be used to estimate the probability of a flow terminating in the next time slot. Algorithm CB-RRA uses the congestion factor of each pair of ToRs to determine the round-robin scheduling order, thus bringing better fairness among flows. In addition, we propose an algorithm LC-SSTF, which assumes that the flows can provide the estimated service duration at the time of request submission. LC-SSTF can be considered as the lower-bound performance of other algorithms since it requires perfect information of input flows.

To address the wavelength assignment problem, we adopt the wavelength selection method first presented in our previous work [16]. The proposed algorithms integrate the wavelength selection method to determine the best wavelength to establish a new lightpath when required. The proposed algorithms are adaptive in the sense that the lightpaths are dynamically reconfigured after each time slot. Not only the lightpaths that are no longer used by any flow are removed from the logical network topology (formed by the set of lightpaths), *active* (existing) flows can also be migrated to a new lightpath with a different wavelength so as to release the current wavelength for a better future-connectivity of ToRs. In other words, we re-assign a different wavelength for the existing lightpaths so as to improve wavelength utilization and network connectivity.

In summary, the contribution of the paper is as follows:

- We develop an optimization programming formulation for the flow scheduling problem in optical data centers. The formulation is modeled as an MDP that takes into account the uncertainty in flow service time;
- We propose three heuristic algorithms: LC-PBST, CB-RRA and LC-SSTF for both scenarios: with known and unknown flow service time. LC-SSTF is considered as the lower-bound performance for other algorithms;
- We design a flow scheduling framework for optical data center networks. The framework integrates the proposed algorithms and leverages on the features of Software Defined Networking (SDN) for traffic monitoring, flow scheduling decision and lightpath reconfigurations;
- We evaluate the effectiveness of the proposed algorithms through comprehensive simulations and compare the

proposed algorithms against a baseline algorithm that is based on first come first served basis.

The rest of the paper is organized as follows. We discuss the related work in Section II. We present the statement of the flow scheduling problem in optical data centers in Section III. We present the MDP formulation in Section IV. We present the proposed methods for wavelength selection and re-assignment in Section V. We present the proposed algorithms and scheduling framework in Section VI. We present the performance study and discuss the results in Section VII. We make concluding remarks in Section VIII.

II. RELATED WORK

Network traffic flow scheduling has been extensively studied in the literature as it is an important issue of traffic engineering. The performance of scheduling algorithms directly affects the performance and resource utilization of the networks. In the wide-area network context, TeXCP [17] and MATE [18] are two algorithms that have been developed to perform dynamic traffic engineering across multiple paths in a wide area network by using explicit congestion notification packets sent by the switches. While these works considered the context of a wide-area electrical packet-switched network, we address the problem in optical data centers using a tightly-coupled central scheduler based on the features of SDN. Considering the data center environment, in [19], the authors presented Hedera, a dynamic flow scheduling algorithm applying to multi-rooted hierarchical tree architectures. The algorithm dynamically estimates loads on the network links and moves flows from heavily loaded links to less utilized links, thus guaranteeing the load balancing among network links. Our work also considers the data center networks however using optical fibers to connect the ToR switches to the core optical switch. Thus, we are constrained by the limitations of optical networks such as the wavelength continuity constraint and high degree of the core optical switch.

There have been also the works that considered the flow scheduling problem in the context of optical networks [10], [11], [13]–[15]. In [10], the authors presented a scheduling algorithm that computes the solutions based on the traffic condition in data center networks. Precisely, the authors assumed that multiple queues exist at each ToR switch to store traffic packets for different destination ToRs. The algorithm tries to schedule the lightpaths among ToRs to forward data packets so as to ensure the load balancing among ToR queues, considering the limited number wavelengths in the fibers. This leads to the fact that the optical data center network will be reconfigured very frequently due to the packet-switched approach (switching occurs with temporal dynamics of few nanoseconds). We advocate for flow level switching with temporal dynamics much slower than packet level. With the flow switching approach, a lightpath assigned to a flow will serve entire lifetime of the flow. Otherwise, the interruption at the middle of service will cause a penalty in terms of performance as well as resource efficiency. We also consider flow admission control whereas few flows may be dropped due to the limited number of wavelengths. In a hybrid data center context, these dropped flows can be accommodated in the electrical packet-switched network.

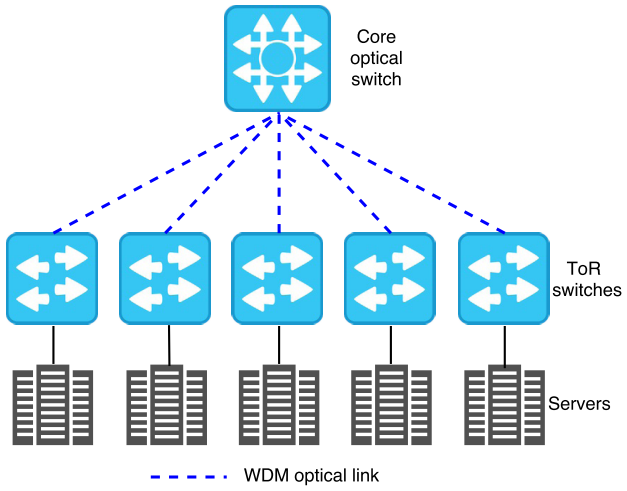

FIGURE 1. Optical data center architecture.

Also based on the flow switching approach, the work presented in [20], proposed heuristic algorithms for flow scheduling in a wide-area network where the routes between a pair of source and destination nodes have been pre-defined. Nevertheless, the heuristic proposed in this work does not consider flow service time, which is one of the important factors that affect the connectivity of optical networks. This work also does not consider the wavelength continuity constraint that requires an efficient method for wavelength assignment problem. We consider these issues in our work. In [11], the authors also considered the traffic scheduling problem in optical data centers. However, they considered a different objective that is to maximize the lifetime of switches. The authors proposed heuristic algorithms that minimize number of reconfigurations of the optical switch while ensuring balancing among flow queues at the ToRs. Considering hybrid optical and electrical data centers, the work presented in [21] proposed an algorithm for selecting flows to be routed by the optical network. While this work considers different objective from ours, it also does not consider the flow service time, which is a key contribution of our work.

III. PROBLEM STATEMENT

We consider a two-tier data center architecture as shown in Fig. 1. We assume that there is a total of M ToR switches in the data center. Each ToR switch is connected to the core optical switch by a fiber that carries up to W wavelengths. A ToR switch can simultaneously reach up to W other ToR switches through optical paths or lightpaths. Data from different ports of a ToR switch are multiplexed onto a WDM link, which is demultiplexed onto different wavelengths at the core optical switch and optically switched to appropriate output ports. Those wavelengths, which are destined to a ToR switch, are multiplexed at the optical switch and sent

through the WDM link connecting that ToR. In the absence of wavelength converters, the lightpath between two ToR switches traversing through the optical switch needs to be wavelength-continuous. The total number of ToR switches supported is limited by the number of ports of the optical switch. A commercial switch such as Cisco CRS-1 can provide up to 1000 ports [22], thus connecting up to $1000/W$ ToR switches where W is the number of wavelengths carried by a fiber. Although a ToR switch can reach only W other ToR switches simultaneously using the optical paths, due to the dynamic reconfiguration capability of the optical switch it can reach different sets of ToR switches at different times. For example, with a 1000-port optical switch and 8 wavelengths per fiber, 125 ToR switches can be connected. We can extend this architecture to connect increased number of ToR switches by using multiple optical switches forming a specified topology network such as ring.

Given a set of traffic flows that need to be accommodated in the data center network. The system needs to take a decision on admission control so as to maximize the total revenue of cloud providers in a long run. Since traffic flows once admitted will stay in the data center for their entire service duration, the admission decision of the current time slot will affect the admission decision of future time slots. We assume that an input flow is an aggregate flow between a pair of ToRs, i.e., it aggregates the traffic between all the servers under the pair of ToRs. A flow thus requires maximum bandwidth capacity of a wavelength, the number of flows accommodated between a pair of ToR switches must be less than the total number of wavelengths carried by the fibers that connect the two ToRs to the optical switch. Otherwise, a number flows will be dropped. Due to the dynamic arrival of flows, a pair of ToRs may receive many traffic flows at the current time slot and less flows in the next time slot. The number lightpaths allocated for a pair of ToR switches will be dynamically adjusted over time.

Let $\mathcal{F}_t$ be the set of all traffic flows that are active at the beginning of time slot t, i.e., it includes all the flows that have been accommodated in the network in time slot $t-1$ and all the flows that are submitted during time slot $t-1$. Each flow $f \in \mathcal{F}_t$ is represented by a tuple of (s_f, d_f, e_f) where s_f is the source ToR switch, d_f is the destination ToR switch, and e_f is the elapsed service time of flow f in the network, respectively. Obviously, when flow f is just submitted, e_f is set to 0. It is to be noted that once a flow has been accepted in the data center, it will not be dropped at the middle if it does not complete its service yet. Thus, the lightpath used by that flow will remain until the flow leaves the network.

While data centers have to maintain the lightpaths for the existing flows, at the beginning of every time slot, a set of new flows need to be scheduled. Thus, optimally solving the flow problem in data centers is computationally prohibitive or may be impossible due to the following reasons. First, the main reason is the uncertainty in flow duration that may be unknown to cloud providers at the admission time. Second, the size of the problem, i.e., the number of decision variables, is very large due the large number of traffic flows and the number of lightpaths in the data centers. Third, due to the dynamic arrival of input traffic flows, the admission decision of the current time slot will affect the admission decision in the future time slots, thereby affecting the overall revenue. Fourth, the wavelength selection for a lightpath between two ToR switches to accommodate a flow affects the future connectivity of ToRs. In the next section, we present a model that can solve the problem optimally based on Markov Decision Process (MDP).

IV. CONSTRAINED Markov DECISION PROCESS FOR FLOW SCHEDULING IN OPTICAL DATA CENTERS

In this section, we present a Constrained Markov Decision Process (CMDP) model that can handle the uncertainty in flow service time and estimate the revenue earned in the future time slots based on the decision of the current time slot. The model takes into account flow requests for a prediction window specified by a number of time slots.

We assume that the service duration of flow f denoted as u_f is an independent random variable and to be Pareto distributed with scale $\beta > 0$ and shape $\alpha > 0$, i.e.,

$$P(u_f > x) = \left(\frac{\beta}{x}\right)^{\alpha}, \quad x \geqslant \beta. \tag{1}$$

It is to be noted that Pareto distribution is a standard convention to model long range dependent heavy tailed service duration observed in the current Internet [23], [24].

A CMDP problem is defined as a four-tuple $(\mathbb{S}, \mathbb{A}, \mathcal{R}, P)$ where $\mathbb{S}$ is the system state space, $\mathbb{A}$ is the system action space, $\mathcal{R}$ is the rewards/revenue function, and P is the system transition probability. We present below the detailed description for each component.

A. STATE SPACE AND ACTION SPACE

A state $S_t \in \mathbb{S}$ is defined by the current network topology with the elapsed time for each lightpath that corresponds to the elapsed time of the flow carried by the lightpath. Thus, we define state S_t at time slot t as a three-dimensional matrix:

$$S_t = \left(E_{i,j,\lambda}\right) \in \mathbb{N}^{M \times M \times W}, \tag{2}$$

where $E_{i,j,\lambda}$ is the elapsed time of the lightpath between ToR i and ToR j using wavelength λ from the moment it has been created. Here, $E_{i,j,\lambda} = 0$ means that there is no lightpath between ToR i and ToR j using wavelength λ. By this definition, S_t solely depends on S_{t-1} since the topology of the network at time slot t can be determined by the current active flows and the newly admitted flows. In other words, S_t depends on the network topology at the end of time slot $t-1$ and the decision on new flow admissions taken at time slot t.

An action $A_t \in \mathbb{A}$ taken at time slot t is the admission decision and wavelength allocated to each lightpath for the newly admitted flows. We thus define action A_t as a three-dimensional binary matrix such that

$$A_t = \left(a_{i,j,\lambda}\right) \in \{0, 1\}^{M \times M \times W}, \tag{3}$$

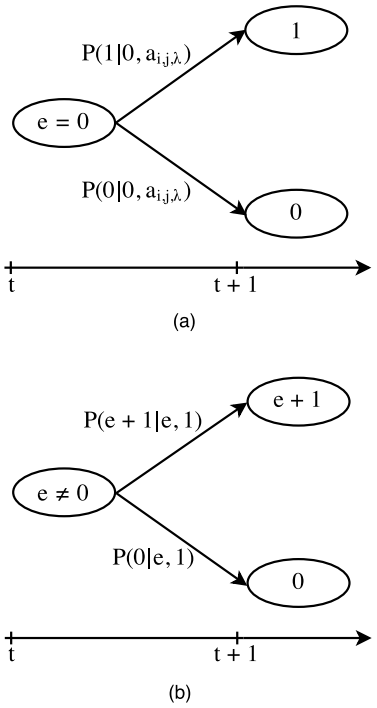

FIGURE 2. Lightpath state transition where e is defined as $E_{i,j,\lambda}$. (a) State transition of lightpaths for the newly-admitted flows. (b) State transition of lightpaths for the existing flows.

where $a_{i,j,\lambda} = 1$ means that we accommodate a flow between ToR i to ToR j in the lightpath using wavelength λ, and $a_{i,j,\lambda} = 0$ otherwise. It is to be noted that if a lightpath has been established between ToR i and ToR j then it is a bidirectional lightpath that can accomodate the flows from ToR j to ToR i, i.e., $a_{i,j,\lambda} = a_{j,i,\lambda}$.

B. TRANSITION PROBABILITY MATRIX

Given state S_t and action A_t taken at time slot t, we compute the transition probability of the system from state S_t at time slot t to state S_{t+1} at time slot $t+1$, denoted as $P(S_{t+1}|S_t, A_t)$. As defined in Eq. (2), each element in the three-dimensional matrix of state S_t, $E_{i,j,\lambda}$, is an integer number that indicates the service duration of the lightpath connecting ToR i and ToR j using wavelength λ. Since lightpaths are dynamically reconfigured, a lightpath will be removed from the logical network topology when it is no longer used by any flow. Thus, $E_{i,j,\lambda} = 0$ indicates that wavelength λ is not used to connect ToR i and ToR j currently. If $E_{i,j,\lambda} \neq 0$, this indicates that wavelength λ has been used for a lightpath between ToR i and ToR j for $E_{i,j,\lambda}$ time slots. Depending on the value of $E_{i,j,\lambda}$ and action taken at time slot t, the transition probability to the next state is computed.

In Fig. 2, we present different scenarios of lightpath state transition, each corresponding to a transition probability. If wavelength λ is available, i.e., $E_{i,j,\lambda} = 0$, a new lightpath using wavelength λ might be created at the beginning of the next time slot to serve a new flow (see Fig. 2a). This corresponds to action $a_{i,j,\lambda} = 1$, implying that $E_{i,j,\lambda} = 1$ in the next time slot. If no lightpath using wavelength λ will be created, it will remain available and $E_{i,j,\lambda}$ is still equal to 0 in the next time slot. The lightpath transition probability to different states in time slot $t+1$ is computed as follows:

$$P(1|0, a_{i,j,\lambda}) = \begin{cases} \left(\frac{\beta}{D}\right)^{\alpha}, & \text{if } a_{i,j,\lambda} = 1, \\ 0, & \text{if } a_{i,j,\lambda} = 0. \end{cases} \tag{4}$$

$$P(0|0, a_{i,j,\lambda}) = \begin{cases} 1, & \text{if } a_{i,j,\lambda} = 0, \\ 1 - \left(\frac{\beta}{D}\right)^{\alpha}, & \text{if } a_{i,j,\lambda} = 1, \end{cases} \tag{5}$$

where D is the duration of a time slot and the scale parameter β must satisfy the condition $\beta \leqslant D$.

For an existing flow, the lightpath must be kept until the flow completes its service (see Fig. 2b). Since a lightpath will be removed when its hosted flow completes the service, i.e., it is no longer used by any flow, the probability that the lightpath is released in the next time slot is equal to the probability of a flow that completes its service in the next time slot. As mentioned earlier, we assume that the service time of a flow follows Pareto distribution with scale $\beta > 0$ and shape $\alpha > 0$. Thus, the probability of the flow completing the service in the next time slot is computed as follows.

$$P(0|e, 1) = 1 - \left(\frac{e}{e+D}\right)^{\alpha}. \tag{6}$$

Thus, the probability that an existing flow continues its service in the next time slot is

$$P(e+1|e, 1) = \left(\frac{e}{e+D}\right)^{\alpha}. \tag{7}$$

Given the transition probability of each lightpath in the network topology of the data center computed by Eqs. (4)–(7), the system transition probability is defined as follows:

$$P(S_{t+1}|S_t, A_t) = \prod_{i=1}^{M} \prod_{j=1}^{M} \prod_{\lambda=1}^{W} P(E_{i,j,\lambda}^{t+1}|E_{i,j,\lambda}^{t}, a_{i,j,\lambda}), \tag{8}$$

where $E_{i,j,\lambda}^{t+1}$ is the state of the lightpath between ToR i and ToR j using wavelength λ extracted from the three-dimensional matrix of state S_{t+1}. Similarly, $E_{i,j,\lambda}^{t}$ is the state of the lightpath between ToR i and ToR j using wavelength λ extracted from the three-dimensional matrix of state S_t.

C. OBJECTIVE AND CONSTRAINTS

Given state S_t and action A_t at time slot t, the revenue function of cloud providers at time slot t, denoted as $\mathcal{R}(S_t, A_t)$, is defined as follows:

$$\mathcal{R}(S_t, A_t) = \frac{\sum_{f \in \mathcal{F}_t^{\text{at}}} c^{\text{unit}} E_{s_f, d_f, \lambda_f}}{\sum_{f \in \mathcal{F}_t^{\text{at}}} E_{s_f, d_f, \lambda_f}} + \sum_{f \in \mathcal{F}_t^{\text{in}}} c^{\text{unit}} a_{s_f, d_f, \lambda_f}, \tag{9}$$

where $\mathcal{F}_t^{\text{at}}$ is the set of all the flows that remain in the network from previous time slots, $\mathcal{F}_t^{\text{in}}$ is the set of input flows at

the current time slot. The first term in the above equation is the revenue obtained at the current time slot from the active flows that still remain in the network. The second term is the revenue obtained for the current time slot from the flows that are admitted. It is to be noted that E_{s_f,d_f,λ_f} is an element in the three-dimensional matrix of system state S_t. Similarly, a_{s_f,d_f,λ_f} is an element in the three-dimensional matrix of action A_t. c^{unit} is the unit cost of a lightpath per time slot. In a long run, the revenue of cloud providers has to be maximized. The average revenue of the providers is then defined as follows:

$$\mathcal{R} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left(\mathcal{R}(S_t, A_t)\right), \tag{10}$$

where $\mathbb{E}(\cdot)$ denotes the expectation of the revenue at each time slot for different states.

At time slot t, action A_t must satisfy the wavelength continuity and optical fiber capacity constraints. From the state at time slot t, $S_t = \left(E_{i,j,\lambda}\right) \in \mathbb{N}^{M\times M\times W}$, we construct the network topology of the data center as follows:

$$\mathcal{L}_{i,j,\lambda} = \begin{cases} 1, & \text{if } E_{i,j,\lambda} \neq 0, \\ 0, & \text{if } E_{i,j,\lambda} = 0. \end{cases} \tag{11}$$

The fiber capacity constraint ensures that the number of flows arriving or departing from ToR i must be less than the number of wavelengths carried by the fiber that connects ToR i to the optical switch. The constraint is represented as follows:

$$\mathcal{C}_i = \sum_{j=1}^{M} \sum_{\lambda=1}^{W} \left(a_{i,j,\lambda} + \mathcal{L}_{i,j,\lambda}\right) \leqslant W, \quad i = 1 \ldots M. \tag{12}$$

The wavelength continuity constraint ensures that two ToRs are connected through a lightpath using the same wavelength. It is presented as follows:

$$\mathcal{C}_{i,\lambda} = \sum_{j=1}^{M} \left(a_{i,j,\lambda} + \mathcal{L}_{i,j,\lambda}\right) \leqslant 1, \quad i = 1 \ldots M, \lambda = 1 \ldots W. \tag{13}$$

We assume that the port number associated to each wavelength in the Mux/DMux of each ToR is the same for all of them. Thus, if ToR i is connected to ToR j with wavelength λ, then port λ is occupied in both ToRs, i.e., $\mathcal{L}_{i,j,\lambda} = \mathcal{L}_{j,i,\lambda} = 1$. This simplifies the port number constraint of sending and receiving ToRs. This also simplifies the wavelength continuity constraint as presented in Eq. (13). For instance, if a lightpath using wavelength λ has been established between ToR 1 and ToR 2 ($L_{1,2,\lambda} = L_{2,1,\lambda} = 1$) then no lightpath that originates or destines ToR 1 and ToR 2 can use wavelength λ ($a_{1,j,\lambda} = a_{j,1,\lambda} = a_{2,j,\lambda} = a_{j,2,\lambda} = 0, \forall j \neq 1, 2$).

D. OPTIMIZATION FORMULATION

The revenue of cloud providers depends not only on the existing flows but also on the decision of cloud providers in admitting new flow requests and wavelength allocation. The optimal decision, i.e., optimal policy to admit new flow requests can be obtained by solving the CMDP formulation. A policy δ that is a mapping of a state $S \in \mathbb{S}$ to action $A \in \mathbb{A}$ is defined as $A = \delta(S)$. We consider a randomized policy in which action A to be taken at state S is randomly chosen according to the probability distribution denoted by $\mu(\delta(S))$ for which $\sum_{\delta(S)\in\mathbb{S}} \mu(\delta(S)) = 1$. In this case, $\mu(A = (a_{i,j,\lambda}) \in \{0, 1\}^{M\times M\times W})$ is the probability that a new lightpath will be established between ToR i and ToR j using wavelength λ if $a_{i,j,\lambda} = 1$. The solution of the CMDP formulation is referred to as the optimal policy denoted by δ^* that maximizes the revenue of cloud providers $\mathcal{R}(\delta)$ while maintaining the fiber capacity constraint and wavelength continuity constraint. The CMDP formulation of the flow scheduling problem is expressed as follows:

$$\text{Maximize:} \quad \mathcal{R}(\delta) \tag{14}$$

$$\text{subject to:} \quad \mathcal{C}_i(\delta) \leqslant W, \quad i = 1 \ldots M, \tag{15}$$

$$\mathcal{C}_{i,\lambda}(\delta) \leqslant 1, \quad i = 1 \ldots M, \lambda = 1 \ldots W. \tag{16}$$

To obtain the optimal policy δ^*, the CMDP formulation can be transformed into an equivalent linear programming (LP) problem [25]. Precisely, there is a one-to-one mapping between the optimal solution $\phi^*(\cdot)$ of the LP problem and optimal policy δ^* of the CMDP formulation. With the randomized policy, $\phi(S, A)$ denotes the steady state probability that action A is taken when the state is S. It is to be noted that the randomized policy is more general than the deterministic policy. Additionally, the randomized policy can be obtained directly by solving the LP problem that ensures the optimality of the solution. The LP problem corresponding to the CMDP formulation defined in Eqs. (14)–(16) is presented as follows:

$$\text{Maximize:} \quad \sum_{S\in\mathbb{S}} \sum_{A\in\mathbb{A}} \mathcal{R}(S, A)\phi(S, A) \tag{17}$$

$$\text{subject to:}$$

$$\sum_{S\in\mathbb{S}} \sum_{A\in\mathbb{A}} \mathcal{C}_i(S, A)\phi(S, A) \leqslant W, \quad \forall i, \tag{18}$$

$$\sum_{S\in\mathbb{S}} \sum_{A\in\mathbb{A}} \mathcal{C}_{i,\lambda}(S, A)\phi(S, A) \leqslant 1, \quad \forall i, \forall\lambda, \tag{19}$$

$$\sum_{A\in\mathbb{A}} \phi(S', A) = \sum_{S\in\mathbb{S}} \sum_{A\in\mathbb{A}} P(S'|S, A)\phi(S, A), \tag{20}$$

$$\sum_{S\in\mathbb{S}} \sum_{A\in\mathbb{A}} \phi(S, A) = 1, \phi(S, A) \geqslant 0 \tag{21}$$

for $S' \in \mathbb{S}$ where $P(S'|S, A)$ is the probability that the state changes from S to S' when action A is taken. This probability is computed using Eq. (8). The objective and constraints defined in Eqs. (17)–(19) correspond to those defined in Eqs. (14)–(16), respectively. The constraint in (20) satisfies the Chapman-Kolmogorov equation.

Let $\phi^*(S, A)$ denote the optimal solution of the LP problem defined in Eqs. (17)–(21). The optimal policy δ^* is a randomized policy that can be uniquely mapped from the optimal

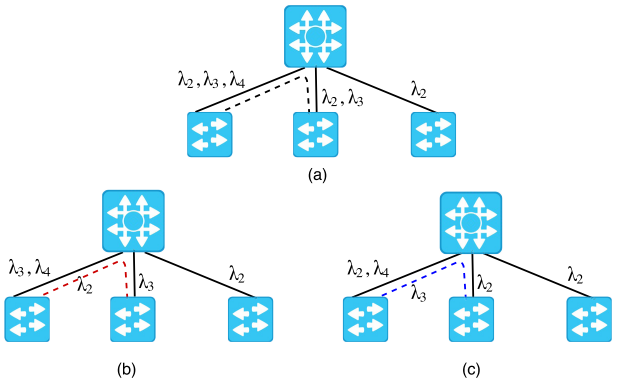

FIGURE 3. An example of wavelength selection: (a) Wavelength availability and lightpath to be created. (b) Use of λ_1 disconnects ToR 1 and ToR 3. (c) Use of λ_2 ensures good connectivity among ToRs.

solution of the LP problem as follows:

$$\mu(A = \delta^*(S)) = \frac{\phi^*(S, A)}{\sum_{A' \in \mathbb{A}} \phi^*(S, A')} \tag{22}$$

for $S \in \mathbb{S}$ and $\sum_{A' \in \mathbb{A}} \phi^*(S, A') > 0$. The optimal solution $\phi^*(S, A)$ is not mathematically tractable due to the large number of decision variables. For instance, given a data center with 4 ToRs, 4 wavelengths per fiber and each flow lasts 4 time slots on average, the number of decision variables of the problem is more than 4500. In the next sections, we develop several heuristic algorithms, considering the uncertainty in flow service time. The probabilistic models developed in this section will be used in the proposed heuristic algorithms to estimate the termination of each traffic flow. We also propose a heuristic algorithm that considers the ideal case where service time of traffic flows are known prior to the scheduling.

V. WAVELENGTH SELECTION AND DYNAMICS OF LIGHTPATH CREATION

We now present the approaches used in our heuristic algorithms to deal with the characteristics of optical data center networks. The methods allow lightpaths to be established dynamically, guaranteeing high resource efficiency and good connectivity of the ToRs in optical data centers.

A. WAVELENGTH SELECTION FOR LIGHTPATH CREATION

When a lightpath needs to be created between two ToRs, selection of a wavelength is an important problem to maximize wavelength utilization. Within a data center, the dynamic arrival of traffic flows and high degree of the optical switch make the impact of the continuity constraint more pronounced. Consider an example shown in Fig. 3 where a lightpath between ToR 1 and ToR 2 needs to be created. Since they have two common wavelengths, i.e., λ_2 and λ_3, any one of them can be used for the new lightpath. However, if λ_2 is used, ToR 1 cannot reach ToR 3 in future if communication is required between ToR 1 and ToR 3 (see Fig. 3b). Thus, it is better to use λ_3 to connect ToR 1 and ToR 2 and reserve λ_2 for the future as shown in Fig. 3c. To solve this problem, we use a function that computes the goodness value for each wavelength between a given pair of ToRs to choose the wavelength with the highest goodness value for a lightpath as in our earlier work [16]. The goodness function for a common wavelength λ between ToR i and ToR j is defined as follows:

$$F(\lambda) = \frac{1}{\mathcal{A}(\lambda)} \tag{23}$$

where $\mathcal{A}(\lambda)$ is the number of fibers on which wavelength λ is available. This is equivalent to choosing a wavelength that is used on most of the fibers, ensuring the highest connectivity of ToRs in the data center. It is to be noted that this goodness function is used not only for determining a wavelength for a new lightpath but also for evaluating the connectivity of a ToR. We present its usage in the next section.

B. RECONFIGURATION OF LIGHTPATHS

The flexibility of optical switch brings in the dynamic reconfiguration of lightpaths with two ways: (i) a lightpath can be removed dynamically when no longer required, and (ii) the wavelength assigned to the lightpath can be replaced by a different one, i.e., migrating traffic flows from a lightpath with a wavelength to another lightpath with a different wavelength. While the removal of free lightpaths releases wavelengths to accommodate newly-arriving flows, re-assignment of a wavelength could create a better connectivity for ToRs in the network. Take the example shown in Fig. 3 and let us assume that the current network state is shown in Fig. 3b, re-assigning wavelength λ_2 to the lightpath between ToR 1 and ToR 2 creates good connectivity among ToRs as shown in Fig. 3c.

It is to be noted that the re-assignment of wavelengths for the lightpaths is transparent to the users and they will not perceive any interruption in the data transmission of their applications. However, dynamic reconfiguration incurs additional overhead. We carry out reconfiguration at the beginning of each time slot. Precisely, when all the flows that have completed their service and leave the system and before we start scheduling newly arriving traffic flows, all the free lightpaths are removed. For the flows that have not yet completed their service, we algorithmically remove all the lightpaths and use the goodness function defined in Eq. (23) to determine the best wavelength to re-create the lightpath for each of them.

C. CONGESTION FACTOR OF LIGHTPATHS

Given the network state, i.e., the current usage of wavelengths for all the lightpaths in the network, the congestion factor of the lightpaths between a pair of ToRs, ToR i and ToR j, depends on two parameters: (i) the number of flows that have been accommodated between ToR i and ToR j, and (ii) the number of common wavelengths that are available in the fibers connecting ToR i and ToR j to the core optical switch. We denote the congestion factor of the lightpaths between ToR i and ToR j as $C(i, j)$, which is defined as a

production of two components as follows:

$$C(i,j) = \frac{F(i,j)}{|\mathcal{F}_t^{\text{at}}|}\left(1 - \frac{\mathcal{A}(i,j)}{W}\right) \quad (24)$$

where $F(i,j)$ is the number of flows that have been accommodated between ToR i and ToR j, and $\mathcal{F}_t^{\text{at}}$ is the set of traffic flows that have been accommodated in the network. $\mathcal{A}(i,j)$ is the number of common wavelengths that are available in the fibers connecting ToR i and ToR j to the core optical switch. The first component indicates that the more flows accommodated between two ToRs the more congested the fibers connecting the two ToRs to the core optical switch. However, in case where only few flows have been accommodated between two ToRs, it does not mean that the fibers of the ToRs are not congested. If they have used most of the wavelengths to create lightpaths to other ToRs, they may not have any common wavelengths, implying that it might not be possible to accommodate more flows between them.

VI. HEURISTIC ALGORITHMS

In this section, we develop heuristic algorithms to solve the problem of flow scheduling in optical data centers. We first present a heuristic algorithm that considers the case where the service time of flows is not known. We then describe an algorithm with the ideal case where the service time of flows is known. Finally, we present the design of the flow scheduling framework in optical data center networks with the support of the Software Defined Networking paradigm.

A. LEAST CONGESTION AND PROBABILITY-BASED SERVICE TIME

In this section, we present our first heuristic algorithm that considers the case where flow service time is unknown to the flow scheduler. Since service time of new input flows is unknown, we can schedule input flows based on only the current state of the data center network and existing active flows, i.e., the flows that are still remaining in the network and they will complete their service in future. As mentioned earlier, we assume that flow service time follows a probability distribution. The elapsed time of existing flows provides us the probability that they will terminate their service in the next time slot, thus releasing their lightpaths for future requests and reducing the congestion of the lightpaths between their source and destination ToRs. Based on this rational, we propose Least Congestion and Probability-based Service Time algorithm (LC-PBST) whose pseudo-code is presented in Algorithm 1.

Let us consider time slot t when we start the scheduling of all the new flows that arrive during time slot $t-1$, denoted as $\mathcal{F}_t^{\text{in}}$. Given the current state of the network and the set of the existing flows that are still remaining in the network, denoted as F_t^{at}, the algorithm performs lightpath reconfiguration as we described in the previous section (line 2). Given the set of the flows that arrive during time slot $t-1$, the algorithm repeatedly processes one by one until all the input flows have the admission decision. In each iteration, the algorithm determines the set of all (s,d) pairs that are the source and destination ToRs of the input flows, which have not been scheduled. For each (s,d) pair, the algorithm computes their congestion factor, using Eq. (24) (line 6). To estimate the connectivity of the involving ToRs in future, the algorithm computes the probability that the existing flows hosted by the lightpaths between s and d will complete their service in the next time slot. We denote $P^{\max}(s,d)$ as the probability of the flow that will mostly terminate in the next time slot (lines 8-14). Given the congestion factor and probability $P^{\max}$ of all the (s,d) pairs in set $\mathbb{O}$, the algorithm sorts them in the ascending order of their congestion factor. If two (s,d) pairs have the same congestion factor, they will be sorted in the descending order of the probability $P^{\max}(s,d)$, i.e., the (s,d) pair that has a flow finishing its service earlier will be selected first.

Algorithm 1 Least Congestion and Probability-based Service Time (LC-PBST)

Input: Network state.
Output: Admission and wavelength assignment decision.

1: **for** $t = 1 \ldots T$ **do**
2: Perform lightpath reconfiguration;
3: **while** $\mathcal{F}_t^{\text{in}} \neq \emptyset$ **do**
4: Get set $\mathbb{O} = \{(s,d)\}$ where (s,d) are the source and destination ToRs of input flow $f \in \mathcal{F}_t^{\text{in}}$;
5: **for** $(s,d) \in \mathbb{O}$ **do**
6: Compute $C(s,d)$ as defined in Eq. (24);
7: $P^{\max}(s,d) \leftarrow 0$;
8: **for** $f \in F_t^{\text{at}} \wedge s_f = s \wedge d_f = d$ **do**
9: Compute the probability that flow f completes its service in the next time slot as defined in Eq. (6), denoted as P_f;
10: **if** $P^{\max}(s,d) < P_f$ **then**
11: $P^{\max}(s,d) \leftarrow P_f$;
12: **end if**
13: **end for**
14: **end for**
15: Sort all the (s,d) pairs in set $\mathbb{O}$ based on their congestion factor and $P^{\max}(s,d)$;
16: Get flow $f \in \mathcal{F}_t^{\text{in}}$ from the first (s,d) pair;
17: **if** We can establish a lightpath between ToR s_f and ToR d_f **then**
18: Use Eq. (23) to determine the best wavelength;
19: Create the lightpath between ToRs s_f and d_f;
20: Accommodate flow f in the network;
21: Update the wavelength usage;
22: **else**
23: Inform the rejection message for flow f;
24: **end if**
25: $\mathcal{F}_t^{\text{in}} \leftarrow \mathcal{F}_t^{\text{in}} \setminus \{f\}$;
26: **end while**
27: **return** Admission control and wavelength assignment;
28: **end for**

Given flow f that belongs to the first (s, d) pair in the sorted list determined in the previous step, the algorithm verifies whether a lightpath can be established between ToR s and ToR d, i.e., there exist common wavelengths between ToR s and ToR d. If so, the algorithm uses the goodness function defined in Eq. (23) to determine the best wavelength for the new lightpath and updates the wavelength usage state. Otherwise, the flow is rejected due to the wavelength continuity or fiber capacity constraints. Flow f is then removed from the input flow set and the algorithm continues with the next input flow until all the flows are processed. If the first (s, d) pair has more than one input flow, flow f can be selected based on a random or first come first served basis.

It is to be noted that if an input flow is accommodated in the data center, i.e., a lightpath will be created/reserved for that flow, the congestion factor of the (s, d) pair that hosts the flow will be changed. Thus, at each iteration, the algorithm needs to re-compute the congestion factor for all the (s, d) pairs that still have input flows requiring to be scheduled.

B. LEAST CONGESTION AND SHORTEST SERVICE TIME FIRST

While algorithm LC-PBST considers the scenario in which flow service time is unknown to the scheduler, we present in this section a lower-bound algorithm, namely Least Congestion and Shortest Service Time First Algorithm (LC-SSTF). LC-SSTF assumes that flow service time is known to the scheduler and it is specified in the requests. While this assumption makes the scheduling simpler than LC-PBST, it still reflects realistic scenarios where cloud resources are requested as advance reservations with specific usage duration. Since the flow service time is known, instead of using the elapsed time of the existing flows to evaluate the future connectivity of the ToRs in data centers, we can use the service time of input flows to provide more accurate information on the completion time of flows, thus releasing the lightpaths for other flows.

As mentioned earlier, flow service time also affects the future connectivity of the ToRs. Longer the flow service time, longer the time the ToR is disconnected due to the wavelength continuity constraint. We thus give the priority to the flows that have shorter service time to be accommodated before other flows. Pseudo-code of LC-SSTF is presented in Algorithm 2. Similar to LC-PBST, LC-SSTF performs flow scheduling at the beginning of every time slot with a set of flows that arrive during the previous time slot, denoted as $\mathcal{F}_t^{\text{in}}$. The algorithm repeatedly processes one by one until all the input flows have the admission decision. In each iteration, the algorithm first computes the congestion factors for all the (s, d) pairs that are source and destination ToRs of the remaining input flows according to Eq. (24). The algorithm then selects flow f from the (s, d) pair with the least congestion factor and the shortest service time among the flows requesting for the same (s, d) pair. If there exist common wavelengths on the fibers that connect ToRs s_f and d_f to the core optical switch, the algorithm uses the goodness function defined in Eq. (23) to determine the best wavelength for the new lightpath. The algorithm updates the wavelength usage state to reflect the newly created lightpath in the next scheduling. Otherwise, the flow is rejected due to the wavelength continuity or fiber capacity constraints. The algorithm continues with the next input flow until all the flows are processed.

Algorithm 2 Least Congestion and Shortest Service Time First (LC-SSTF)

Input: Network state.
Output: Admission and wavelength assignment decision.

```
for t = 1 ... T do
    Perform lightpath reconfiguration;
    while F_t^in ≠ ∅ do
        Get set 𝕆 = {(s, d)} where (s, d) are the source and destination ToRs of input flow f ∈ F_t^in;
        for (s, d) ∈ 𝕆 do
            Compute C(s, d) as defined in Eq. (24);
        end for
        Get flow f from the (s, d) pair with the least congestion factor and f has the shortest service time;
        if We can establish a lightpath between ToR s_f and ToR d_f then
        Use Eq. (23) to determine the best wavelength;
            Create the lightpath between ToRs s_f and d_f;
            Accommodate flow f in the network;
            Update the wavelength usage;
        else
            Inform the rejection message for flow f;
        end if
        F_t^in ← F_t^in \ {f};
    end while
    return Admission control and wavelength assignment;
end for
```

C. CONGESTION-BASED ROUND-ROBIN ALGORITHM

Applying algorithms LC-PBST and LC-SSTF may create better connectivity for the network, thus increasing the number of traffic flows accommodated in the network. However, it may cause flow starvation such that some of flows will never be accommodated in the network due to dynamic arrival of new flows that have shorter service time. To provide better fairness among flows, we propose to use the round-robin approach instead of prioritizing the flows based on their service time. We develop an algorithm namely Congestion-Based Round-Robin Algorithm (CB-RRA). Precisely, after computing the congestion factor of all the (s, d) pairs that are source and destination of input flows, they are then sorted in the ascending order of their congestion factor. The obtained order will be used for the scheduling step. The round-robin scheduling is applied such that in each scheduling

TABLE 1. Example of scheduling order with different prioritizing methods.

Flow	Source	Destination	Service time	Congestion factor
f_1	ToR 1	ToR 2	3	0.3
f_2	ToR 1	ToR 2	4	0.3
f_3	ToR 2	ToR 3	5	0.4
f_4	ToR 2	ToR 3	6	0.4

round, a flow from each (s, d) pair starting with the (s, d) pair with the lowest congestion factor will be scheduled. CB-RRA brings better fairness for the input flows compared to algorithms LC-PBST and LC-SSTF since it keeps the flow selection order based the congestion factor of every pair of ToRs computed only once as discussed earlier. It is worth recalling that algorithms LC-PBST and LC-SSTF recompute the congestion factor of every pair of ToRs after each acceptance of flows.

In Table 1, we show an illustrative example with input flows that have different scheduling orders due to different prioritizing methods. When we apply the prioritizing method presented in the previous section using LC-SSTF, i.e., based on the congestion factor of (s, d) pairs and service time of flows, the scheduling order will be f_1, f_2, f_3, f_4. If we use the round-robin method, the scheduling order will be f_1, f_3, f_2, f_4. Given that the fibers connecting ToRs to the core optical switch have only 2 wavelengths available, then flows f_3 and f_4 will be rejected according to LC-SSTF. The round-robin method (CB-RRA) will reject flows f_2 and f_4, thus implying that both short and long flows in terms of service time have an equal chance to be accommodated in the network.

It is to be noted that except the prioritizing method, all the other steps of the algorithm are kept the same during the scheduling process as shown in Algorithm 2. We also use the goodness function defined in Eq. (23) to determine the best wavelength for a new lightpath when required. Thus, we skip the presentation of the pseudo-code of CB-RRA in this section.

D. COMPLEXITY OF ALGORITHMS

While the optimization formulation is mathematical intractable, the proposed heuristic algorithms have low polynomial complexity. To complete the scheduling for a time slot, algorithm LC-PBST has the worst complexity of $\mathcal{O}(N^2\ K)$ where N is the number of input flows and K is the number of active flows that still remain in the network. This is due to the fact that it has to consider not only the input flows but also active flows to predict their termination. Whereas, both algorithms LC-SSTF and CB-RRA have the worst complexity of $\mathcal{O}(N^2)$. We believe that such low complexity will not add any heavy overhead to the controller when performing the flow scheduling. Furthermore, with the advance of the Software Defined Networking paradigm, the proposed algorithms can be run in powerful servers and complete the execution in an acceptable time. This makes the proposed algorithms effective for data center networks, allowing short traffic flows that may last only few hundreds of μs still get serviced properly.

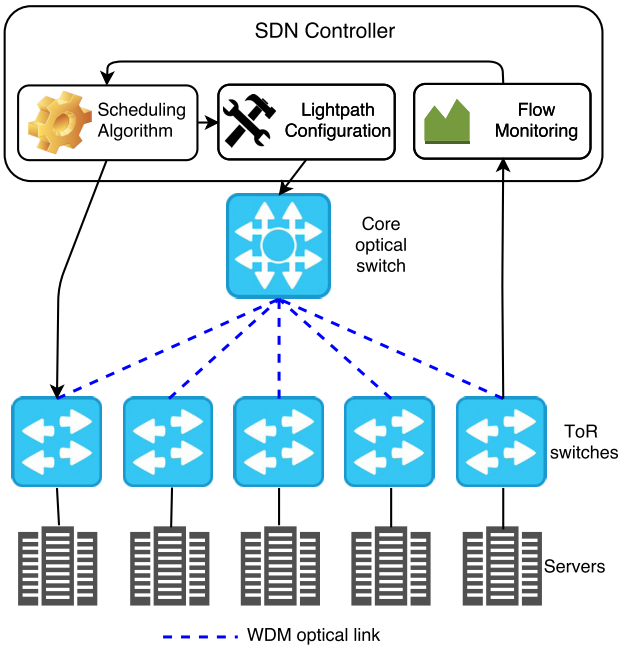

FIGURE 4. SDN-based framework for flow scheduling in optical data center networks. The arrows are the control and data flows among components of the SDN controller.

E. SCHEDULING FRAMEWORK IN OPTICAL DATA CENTER NETWORKS

We now present an SDN-based framework for flow scheduling in optical data center networks. The design of the framework is shown in Fig. 4. The `Flow Monitoring` module periodically monitors the information about traffic flows so as to estimate the service time of each flow. Upon arrival of a new flow request, the controller runs the `Scheduling Algorithm` to produce the admission decision and the wavelength assignment. Depending on the preference of cloud providers, LC-PBST, LC-SSTF or CB-RRA will be selected to generate the scheduling decision. The wavelength assignment is forwarded to the `Lightpath Configuration` module to invoke the circuit switching and establish the lightpaths among the ToR switches. The admission decision is sent to the ToR switches to start the data transmission for the admitted flows while other flows are dropped. It also depends on the preference of cloud providers to determine the frequency of running algorithms, i.e., duration of each time slot so as to maximize the performance. The total delay of the procedure is the running time of the scheduling algorithm and lightpath configuration. It might be possible that a newly admitted flow uses an existing lightpath such that wavelength reconfiguration is not required. Otherwise, a circuit switching time in the order of micro or milliseconds is required depending on the nature of the optical switch.

It is to be noted that with the features of SDN, the above framework can be realized with OpenFlow-enabled switches [26] that are used for the ToR switches. Recent literature shows the possibility of SDN control over optical networks [27], thus lightpath configuration can be performed flexibly. Indeed, enabling the SDN control (that operates at L2 and above) down to the photonic level operation of optical communications at L1 opens the possibilities for flexible adaption of the photonic elements, thus supporting optical networking functionalities. A Reconfigurable Optical Add-Drop Multiplexer (ROADM) is an important photonic switching device for optical networks [28]. Through wavelength selective optical switches, a ROADM can drop or add one or multiple wavelengths to an optical lightpath without requiring the conversion of the optical signal to electric signal. A management control plane has been designed in ROADM and provides OpenFlow protocols such that changes of wavelengths can be remotely controlled by the SDN controller.

It is also to be noted that the SDN controller is usually deployed in a powerful server so that it can run complex and sophisticate algorithms for network management and traffic engineering. Furthermore, to ensure the resiliency of the SDN controller, which represents the single point of failure in the system, backup (secondary) controller(s) can be additionally deployed. These backup controllers need not be in active mode during their lifetime as long as the primary controller is still in normal working condition. Upon a failure of the primary controller, a backup controller is made active and takes the responsibility of the primary controller on flow scheduling. In a large-scale data center, a distributed controller architecture can be used wherein multiple controllers co-exist, each managing a number of ToR switches for the purpose of control load balancing and fault tolerance. The controllers coordinate among themselves and in the event of a controller failure, the unaffected controllers can take over the responsibilities.

VII. PERFORMANCE STUDY

In this section, we evaluate the performance of the proposed algorithms. We first describe the settings of simulation parameters. We then analyze the obtained simulation results and compare the proposed algorithms against a baseline algorithm.

A. PARAMETER SETTINGS

We consider an optical data center network with the architecture as shown in Fig. 1. The core switch connects 48 ToR switches. The ToR switches are connected to the core switch with optical fibers each carrying 32 wavelengths. Input flows are generated with source and destination ToRs that are randomly chosen from the set of ToRs. The service time of each flow follows the Pareto distribution with the scale parameter set to 5 time slots. The average service time of flows is 10 time slots. As mentioned earlier, we assume that each flow requires the entire capacity of a wavelength.

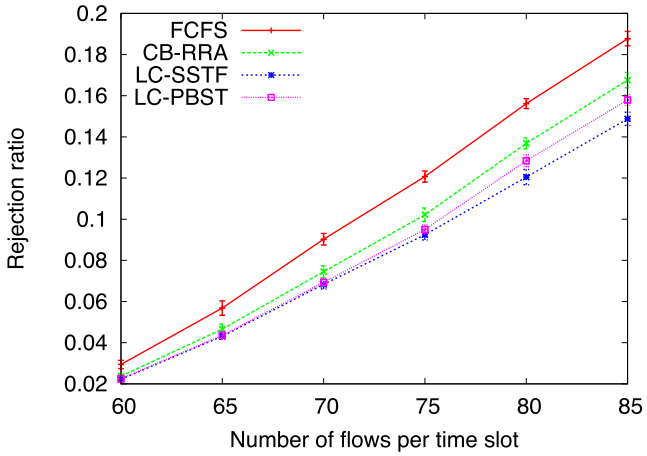

FIGURE 5. Rejection ratio for different arrival rates of flows.

We examine the performance of the four following algorithms.

- Least Congestion and Probability-based Service Time (LC-PBST) that uses a probabilistic model to address the uncertainty in flow service time;
- Least Congestion and Shortest Service Time First (denoted as LC-SSTF) that prioritizes the flows with short service time and low congestion factor of their (s, d) pair;
- Congestion-Based Round-Robin Algorithm (denoted as CB-RRA) that guarantees fairness among flows;
- Baseline algorithm: we use the first come first served basis (denoted as FCFS) such that input flows are scheduled based on their arrival order. Each flow will be admitted or dropped based on the network state, i.e., availability of lightpaths and wavelengths in the fibers that connect the source and destination ToRs of the flow.

All the algorithms are run with 2000 time slots. We use three following metrics to evaluate the performance of the algorithms. The results are plotted with 95% confidence interval.

- Rejection ratio: it is computed as the ratio of the total number of dropped flows over the number of input flows;
- Average revenue: it is computed according to Eq. (10);
- Wavelength utilization: The ratio of the number of wavelengths used over the total number of wavelengths in all fibers, taking the average over 2000 time slots.

B. RESULT ANALYSIS

1) PERFORMANCE COMPARISON AMONG PROPOSED ALGORITHMS

In Fig. 5, we present the rejection ratio generated by the algorithms with respect to the number of flows that arrive per time slot. We observe that even without the information about flow service time, algorithm LC-PBST still has better performance than the baseline algorithm FCFS and CB-RRA. In the best case, LC-PBST reduces the rejection ratio by up to 23% and 7% compared to that of FCFS and CB-RRA, respectively.

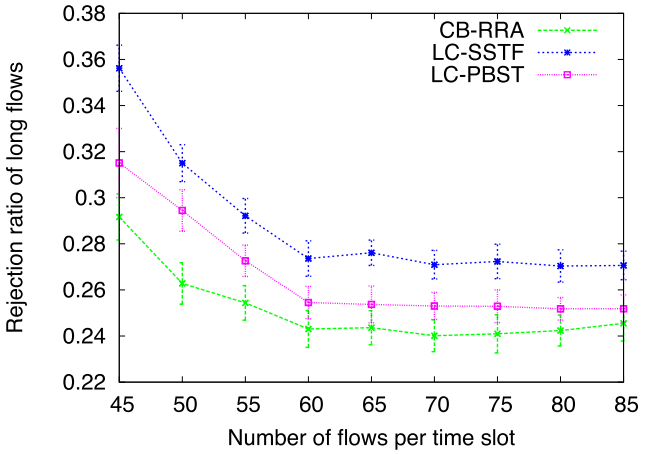

FIGURE 6. Rejection ratio of the flows with service time longer than or equal to 10 time slots.

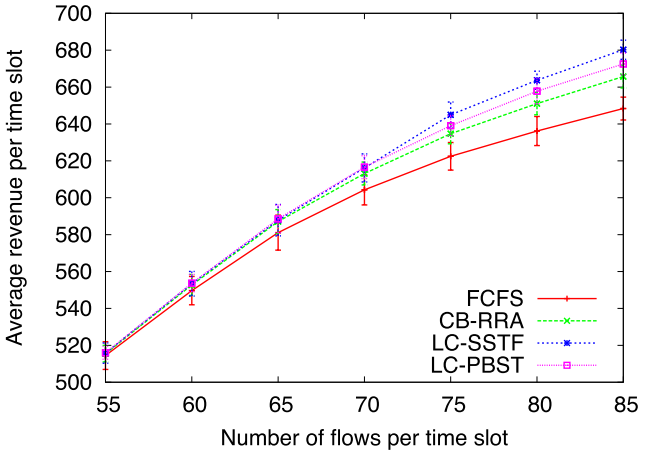

FIGURE 7. Average revenue.

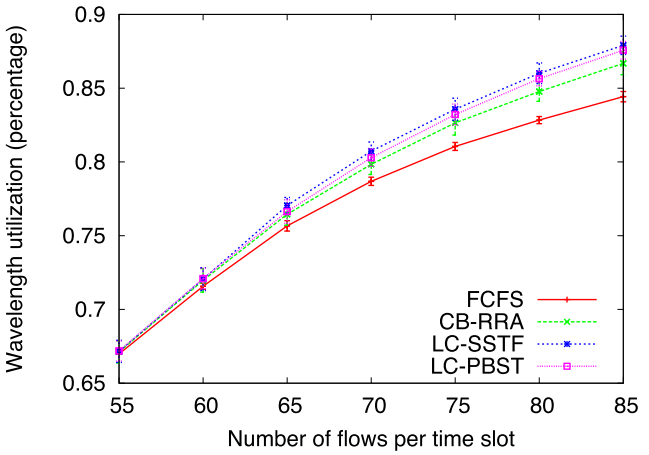

FIGURE 8. Wavelength utilization of the network.

Such performance improvement confirms that accommodating input flows without any prioritizing method will result in very low performance. Due to the wavelength continuity constraint, accommodating a certain flow will block the entire network and thus dropping all the flows that arrive later. The results also show the effectiveness of the probabilistic model used in the proposed algorithm. With perfect information of input flows, algorithm LC-SSTF expectedly has the best performance among the algorithms. In the best case, LC-SSTF reduces the rejection ratio by up to 25% and 12% compared to that of FCFS and CB-RRA, respectively. Compared to LC-PBST, LC-SSTF slightly reduces the rejection ratio since it has accurate information of input flows. In the best case, LC-SSTF reduces the rejection ratio by up 12% compared to algorithm LC-PBST. As we mentioned previously, algorithm LC-SSTF can be considered as the lower-bound of all the algorithms.

In Fig. 6, we present the rejection ratio of the flows with service time longer than or equal to 10 time slots. It is the ratio of the number of rejected flows with service time longer than or equal to 10 time slots over the total number of rejected flows. It is worth recalling that the service time of input flows follows the Pareto distribution with the mean value of 10 time slots. We assume that the flows with service time longer than or equal to 10 time slots are considered as long flows. The results show that the rejection ratio of long flows resulted by CB-RRA is significantly lower than that resulted by LC-SSTF. In the best case, CB-RRA reduces the rejections of long flows by up to 50% compared to LC-SSTF and LC-PBST. This implies that CB-RRA gives better fairness among traffic flows. We also observe that LC-PBST brings better fairness among flows than LC-SSTF. This is because without accurate information of input flows and using a probabilistic model, LC-PBST probably accommodates long flows rather than short flows.

In Fig. 7, we plot the average revenue obtained by cloud providers after 2000 time slots. The results show that all the proposed algorithms increase the average revenue compared to that of FCFS. In the best case, LC-SSTF increases the revenue by up to 5% compared to that of FCFS while CB-RRA and LC-PBST increase the revenue by up to 3% and 4% compared to FCFS, respectively. This demonstrates the effectiveness of the proposed algorithms that apply different intelligent techniques even though perfect information of input flows may not be available. It is also worth mentioning that LC-PBST has better performance than CB-RRA even though both LC-PBST and CB-RRA do not have accurate flow service time to be used in the algorithms. This shows that using a prediction technique based on a probabilistic model can provide approximate information on the flow service time. Even though the estimation may not be accurate, it is still better than CB-RRA that does not use this parameter at all.

In Fig. 8, we present the wavelength utilization of the proposed algorithms. The results show that the proposed algorithms achieve better performance. At the arrival rate of 70 flows per time slot, the proposed algorithms achieve the wavelength utilization of 84% to 88%, corresponding to an improvement of 4% compared to the baseline algorithm. This improvement comes from the prioritizing

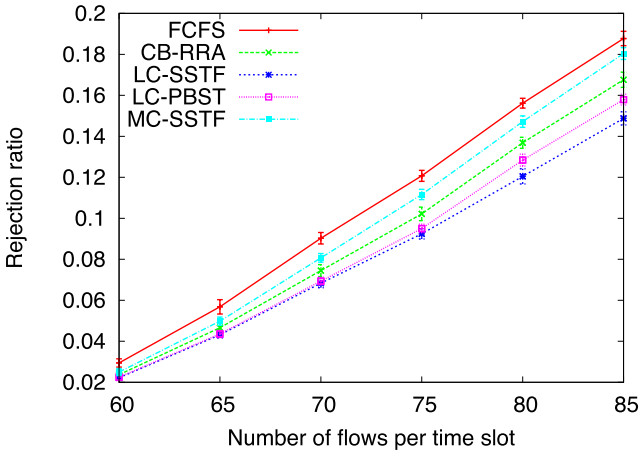

FIGURE 9. Performance of congestion model.

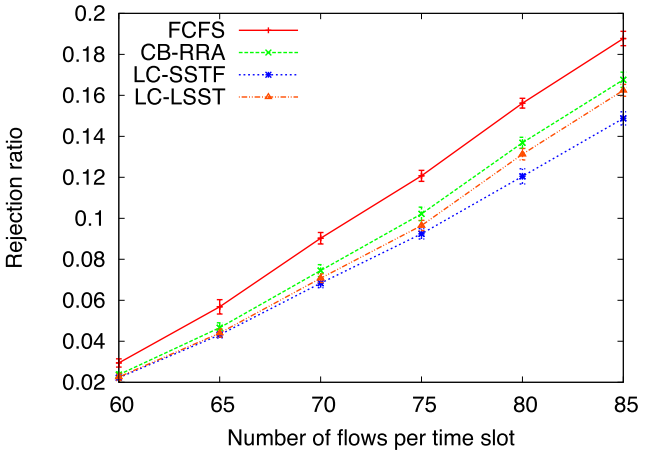

FIGURE 10. Short flows first vs. long flows first.

methods proposed in our work. Since each lightpath involves two fibers that connect the source and destination ToRs to the core optical switch, giving the priority for the flows with the least-congestion-first will better exploit the common wavelengths in the fibers of the involving ToRs. This will eliminate the situation that a wavelength is available in the fiber of the source ToR but not available in the fiber of the destination ToR.

2) PERFORMANCE OF CONGESTION MODEL

In this simulation, we evaluate the performance of the congestion model proposed in this work. We compare the performance of the algorithms that prioritize the flows with the least congestion of their source and destination ToRs against the algorithm that schedules the flows with the most congestion factor of their source and destination ToRs. The algorithm is denoted as MC-SSTF that is different from LC-SSTF only at the prioritizing method. In Fig. 9, we present the performance of the examined algorithms. The results show that exhausting the wavelengths in the fibers, i.e., prioritizing the flows involving congested ToRs will worsen the performance. In the worst case, MC-SSTF increases the rejection by up to 25% compared to that of LC-SSTF. It is worth mentioning that while MC-SSTF has worse performance than all the proposed algorithms, it still performs better than FCFS. This is because MC-SSTF gives the scheduling priority for short flows before long flows.

3) SHORT FLOWS FIRST VS. LONG FLOWS FIRST

In this simulation, we evaluate the prioritizing method based on flow service time. We compare the proposed algorithm LC-SSTF that gives priority to the short flows against LC-LSTF (Least Congestion and Longest Service Time First), which gives priority to the long flows. In Fig. 10, we present the performance of the examined algorithms. The results show that accepting long flows in the network leads to high rejection ratio. In the worst case, LC-LSTF increases the rejection ratio by up to 13%. However, by comparing LC-LSTF to the baseline algorithms FCFS and CB-RRA, LC-LSTF still has better performance. This demonstrates the effectiveness of the proposed priority approach based on flow service time. Additionally, one may think that accepting long flows may result in better revenue for cloud providers since a long flow generates more stable revenue than multiple short flows for a given time window. However, the results show that the average revenue per time slot obtained by LC-LSTF is less than that of LC-SSTF by up to 0.9%. This is due to the fact that accepting long flows blocks the network with a lightpath for long time. Even though wavelength allocated for lightpaths can be re-assigned, this still pronounces the impact of wavelength continuity constraint, leading to rejection of future flows.

4) PERFORMANCE OF WAVELENGTH RE-ASSIGNMENT

In this simulation, we evaluate the performance of the wavelength re-assignment approach. We run the proposed algorithms with two scenarios and measure the rejection ratio.

- With wavelength re-assignment: The algorithm performs wavelength re-assignment at the beginning of each time slot. The lightpaths that are no longer used by any flow are removed from the logical network topology. The active lightpaths used by the existing flows are also reviewed and possibly re-assigned with new wavelengths. The algorithms are denoted with suffix "WR";
- Without wavelength re-assignment: Only the lightpaths that are no longer used by any flow are removed. The algorithms are denoted with suffix "WoR".

In Fig. 11, we present the rejection ratio of the proposed algorithms running with the above scenarios. The results show that applying wavelength re-assignment significantly improves the performance of the algorithms. In the best case, the algorithms with wavelength re-assignment reduce the rejection ratio by up to 70% compared to the case without wavelength re-assignment. As mentioned earlier, wavelength re-assignment creates better connectivity among ToRs so as to accommodate more flows that arrive later. It also mitigates the effect of the wavelength continuity constraint in the network since it may increase the number of common

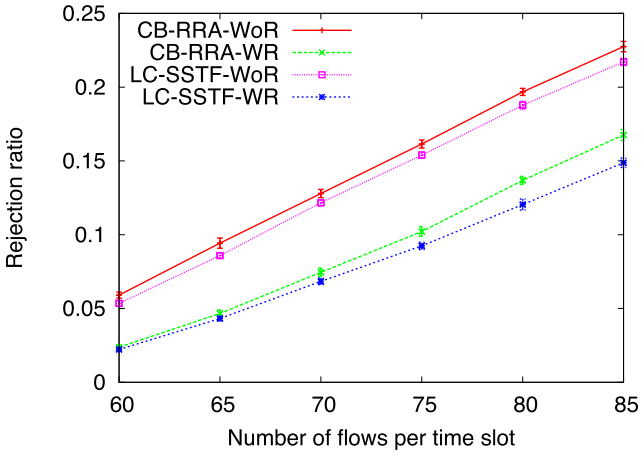

FIGURE 11. Rejection ratio of the algorithms with and without re-assignment of wavelengths.

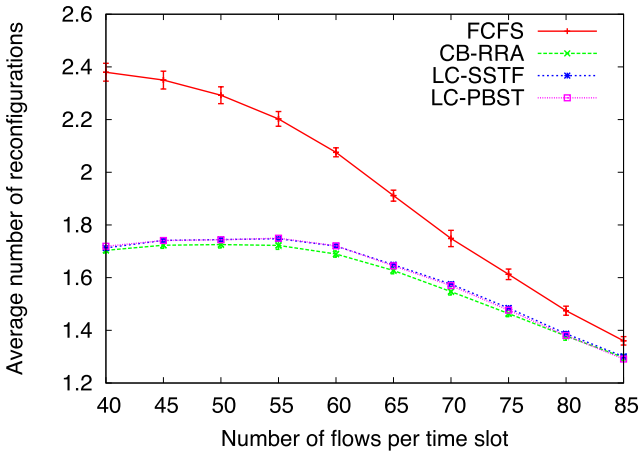

FIGURE 12. Average number of lightpath removals and wavelength re-assignment.

wavelengths that are available in the fibers for any pair of ToRs.

It is also to be noted that wavelength re-assignment incurs overhead for the data center network although we do not quantify such overhead in this simulation. Nevertheless, we present the average number of wavelength re-assignments per time slot in Fig. 12. The results show that the baseline algorithm FCFS performs more wavelength re-assignments than the proposed algorithms whereas only a few reconfigurations (< 2) are required by our algorithms. Such a small number of reconfigurations is still affordable given the large gain in the system performance. The results also show that when the arrival rate is low, more wavelength re-assignments happen than when the arrival rate is high. This is due to the fact that at low arrival rate, many wavelengths are available for a flow, the selected wavelength is the best at the flow submission instant but it may not be as good as other wavelengths, which are just released in the next time slot. On the other hand, there is not available wavelengths when arrival rate is high.

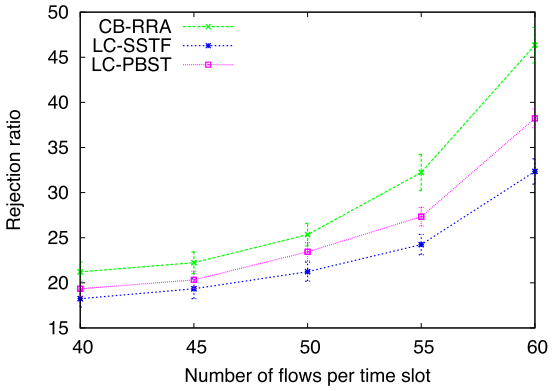

FIGURE 13. Performance of algorithms with incremental network topologies.

5) INCREMENTAL TOPOLOGY VS. RECONFIGURABLE TOPOLOGY

We also evaluate the performance of the proposed algorithms with incremental topology in which a lightpath will not be removed from the logical network topology even though it is no longer required. The wavelength re-assignment is also not done in the incremental topology scenario. We present the results in Fig. 13. As we expected, the rejection ratio generated by the proposed algorithms drastically increases in case of incremental topology. At the arrival rate of 40 flows per time slot, the rejection ratio reaches around 20% whereas no rejection occurs in case of reconfigurable network topology. This is due to the dynamic arrival of traffic flows as well as the randomness of the source and destination ToRs of traffic flows. The new flows may be submitted to the congested pair of ToRs that do not have any other available lightpaths. This demonstrates the effectiveness of lightpath reconfiguration that is feasible with reconfigurable optical switches.

VIII. CONCLUSION

In this paper, we studied the problem of flow scheduling in optical data center networks, considering the uncertainty in flow service time. We developed an optimization programming formulation for the problem that maximizes the revenue of cloud providers. We addressed the uncertainty of flow service time by applying the Markov Decision Process model that can estimate the termination of flows and expected revenue of cloud providers. As the problem is computationally prohibitive, we developed heuristic algorithms that efficiently schedule traffic flows in optical data center networks. We defined a new parameter called congestion factor for a pair of ToRs to determine the scheduling order of traffic flows in the algorithms. We adopted a probabilistic model that allows us to estimate the service time of traffic flows in the heuristic algorithms when the flow service time is not specified in flow requests. We also adopted a goodness function to determine the best wavelength for a lightpath to be created when required so as to ensure better connectivity of the ToRs. The proposed algorithms not only guarantee high revenue for cloud providers but also ensure improved

fairness among traffic flows. We evaluated the performance of the proposed algorithms through comprehensive numerical simulations. The results show that the proposed algorithms achieve significant performance improvement compared to a baseline algorithm by reducing the rejection ratio by up to 25%, thus achieving high revenue for cloud providers.

ACKNOWLEDGMENT

S. Tranquebar Girisankar was with NUS.

REFERENCES

[1] G. Wang *et al.*, "c-Through: Part-time optics in data centers," in *Proc. SIGCOMM*, New Delhi, India, May 2010, pp. 327–338.
[2] N. Farrington *et al.*, "Helios: A hybrid electrical/optical switch architecture for modular data centers," in *Proc. ACM SIGCOMM*, New Delhi, India, 2010, pp. 339–350.
[3] H. Wang, Y. Xia, K. Bergman, T. E. Ng, S. Sahu, and K. Sripanidkulchai, "Rethinking the physical layer of data center networks of the next decade: Using optics to enable efficient *-cast connectivity," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 3, pp. 52–58, Jul. 2013.
[4] M. Chen, H. Jin, Y. Wen, and V. C. M. Leung, "Enabling technologies for future data center networking: A primer," *IEEE Network*, vol. 27, no. 4, pp. 8–15, Jul. 2013.
[5] N. Hamedazimi *et al.*, "FireFly: A reconfigurable wireless data center fabric using free-space optics," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 319–330, Aug. 2014.
[6] A. Hammadi and L. Mhamdi, "Review: A survey on architectures and energy efficiency in data center networks," *Comput. Commun.*, vol. 40, pp. 1–21, Mar. 2014.
[7] "Cisco global cloud index: Forecast and methodology, 2013–2018," Cisco, San Jose, CA, USA, White Paper C11-738085-00, Oct. 2015.
[8] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, 2008.
[9] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: An approach to high bandwidth optical WAN's," *IEEE Trans. Commun.*, vol. 40, no. 7, pp. 1171–1182, Jul. 1992.
[10] C.-H. Wang, T. Javidi, and G. Porter, "End-to-end scheduling for all-optical data centers," in *Proc. IEEE INFOCOM*, Hong Kong, Apr. 2015, pp. 406–414.
[11] A. Bianco, P. Giaccone, and M. Ricca, "Scheduling traffic for maximum switch lifetime in optical data center fabrics," *Comput. Netw.*, vol. 105, pp. 75–88, Aug. 2016.
[12] S. T. Girisankar, T. Truong-Huu, and M. Gurusamy, "SDN-based dynamic flow scheduling in optical data centers," in *Proc. 9th Int. Conf. Commun. Syst. Netw.*, Bangalore, India, Jan. 2017, pp. 190–197.
[13] S. Bojja, M. Alizadeh, and P. Viswanath, "Costly circuits, submodular schedules and approximate Carathéodory theorems," in *Proc. ACM SIGMETRICS*, Antibes Juan-les-Pins, France, Jun. 2016, pp. 75–88.
[14] M. Ghobadi *et al.*, "ProjecToR: Agile reconfigurable data center interconnect," in *Proc. ACM SIGCOMM*, Florianopolis, Brazil, Aug. 2016, pp. 216–229.
[15] H. Liu *et al.*, "Scheduling techniques for hybrid circuit/packet networks," in *Proc. ACM CoNEXT*, Heidelberg, Germany, Dec. 2015, pp. 41:1–41:13.
[16] T. Truong-Huu, M. Gurusamy, and V. Girisagar, "Dynamic embedding of workflow requests for bandwidth efficiency in data centers," *Comput. Netw.*, vol. 108, pp. 184–198, Oct. 2016.
[17] S. Kandula, D. Katabi, B. Davie, and A. Charny, "Walking the tightrope: Responsive yet stable traffic engineering," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 253–264, Oct. 2005.
[18] A. Elwalid, C. Jin, S. Low, and I. Widjaja, "MATE: MPLS adaptive traffic engineering," in *Proc. IEEE INFOCOM*, Anchorage, AK, USA, Apr. 2001, pp. 1300–1309.
[19] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic flow scheduling for data center networks," in *Proc. USENIX NSDI*, San Jose, CA, USA, Apr. 2010, pp. 19–34.
[20] Z. Rosberg, J. Li, F. Li, and M. Zukerman, "Flow scheduling in optical flow switched (OFS) networks under transient conditions," *J. Lightw. Technol.*, vol. 29, no. 21, pp. 3250–3264, Nov. 1, 2011.
[21] I. K. Musa and S. Walker, "Hybrid optical and electrical network flows scheduling in cloud data centres," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 1–17, Apr. 2013.
[22] R. S. Tucker, "Optical packet-switched WDM networks—A cost and energy perspective," in *Proc. Opt. Fiber Commun. Conf.*, San Diego, CA, USA, Feb. 2008, paper OMG1.
[23] M. E. Crovella and A. Bestavros, "Self-similarity in world wide Web traffic: Evidence and possible causes," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 835–846, Dec. 1997.
[24] R. G. Addie, T. D. Neame, and M. Zukerman, "Performance evaluation of a queue fed by a Poisson Pareto burst process," *Comput. Netw.*, vol. 40, no. 3, pp. 377–397, Oct. 2002.
[25] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, Feb. 2005.
[26] A. Nygren, B. Pfaff, B. Lantz, and B. Heller, "OpenFlow switch specification," Open Netw. Found., Menlo Park, CA, USA, Tech. Rep. ONF TS-012, Oct. 2013.
[27] G. Parulkar, T. Tofigh, and M. D. Leenheer, "SDN control of packet-over-optical networks," in *Proc. OFC*, Los Angeles, CA, USA, Mar. 2015, pp. 1–27.
[28] A. S. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein, and W. Kellerer, "Software defined optical networks (SDONs): A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2738–2786, 4th Quart., 2016.

TRAM TRUONG-HUU (M'12–SM'15) received the M.Sc. degree from the Francophone Institute for Computer Science, Hanoi, Vietnam, in 2007, and the Ph.D. degree in computer science from the University of Nice Sophia Antipolis, France, in 2010. He held a Post-Doctoral Fellowship with the French National Center for Scientific Research, from 2011 to 2012. In 2012, he joined the National University of Singapore, where he is currently a Senior Research Fellow with the Department of Electrical and Computer Engineering. His current research interests include scientific workflows, grid, cloud and mobile computing, and networks. He received the Best Presentation Recognition at the IEEE/ACM UCC 2013.

MOHAN GURUSAMY (M'00–SM'07) received the Ph.D. degree in computer science and engineering from IIT Madras, Chennai, in 2000. In 2000, he joined the National University of Singapore, where he is currently an Associate Professor with the Department of Electrical and Computer Engineering. He has over 160 publications to his credit, including two books and three book chapters in optical networks. His research interests include cloud data center networks, software defined networks, and optical networks. He is currently serving as an Editor for IEEE TRANSACTIONS ON CLOUD COMPUTING, *Elsevier Computer Networks journal* and *Springer Photonic Network Communications Journal*. He has served as the Lead Guest Editor for two special issues of the IEEE *Communications Magazine*, 2005 and 2005, and as a Co-Guest Editor for a special issue of the *Elsevier Optical Switching and Networking journal*, 2008. He served as a TPC Co-Chair for several conferences, including the IEEE ICC 2008 (ONS).

SHARMILA TRANQUEBAR GIRISANKAR received the M.Sc. degree in electrical and computer engineering from the National University of Singapore in 2016. She was with a telecom company in India and has two years of experience in the mobile backhaul area. She was an Engineer in a Singapore-based Telecommunication Company. She has been part of many projects developing the layer three protocols in the routers and switches. Her areas of interests include traffic engineering, design of networking protocols, software define networking, and mobile communications.