Understanding the Performance Guarantee of Physical Topology Design for Optical Circuit Switched Data Centers

SHIZHEN ZHAO, PEIRUI CAO, and XINBING WANG, Shanghai Jiao Tong University, China

As a first step of designing **O**ptical-circuit-switched **D**ata **C**enters (ODC), physical topology design is critical as it determines the scalability and the performance limit of the entire ODC. However, prior works on ODC have not yet paid much attention to physical topology design, and the adopted physical topologies either scale poorly, or lack performance guarantee.

We offer a mathematical foundation for the design and performance analysis of ODC physical topologies in this paper. We introduce a new performance metric $\beta(\mathcal{G})$ to evaluate the gap between a physical topology $\mathcal{G}$ and the ideal physical topology. We develop a coupling technique that bypasses a significant amount of computational complexity of calculating $\beta(\mathcal{G})$. Using $\beta(\mathcal{G})$ and the coupling technique, we study four physical topologies that are representative of those in literature, analyze their scalabilities and prove their performance guarantees. Our analysis may provide new guidance for network operators to design better physical topologies for their ODCs.

ACM Reference Format:
Shizhen Zhao, Peirui Cao, and Xinbing Wang. 2022. Understanding the Performance Guarantee of Physical Topology Design for Optical Circuit Switched Data Centers. In *SIGMETRICS '22: ACM SIGMETRICS, June 06–10, 2022, Mumbai, India.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As data center traffic doubles every year [22], building Clos topologies [1, 11, 22] for data centers using electrical switches is becoming more and more expensive and power prohibitive [2]. In order to meet the growing demand at reduced energy cost, building ODCs is becoming a promising alternative for future data centers. An optical circuit switch, e.g. Calient [4], could offer at least hundreds of times higher switching capacity than an electrical switch, while its energy cost is hundreds of times lower (less than 45 Watts for an optical circuit switch with 320 Tx/Rx pairs).

While the eventual goal of ODC design is the full-optical data center, due to technological immaturity, existing designs of ODC have mostly adopted a hybrid design that involves both electrical switching and optical switching. The design of an ODC typically covers the following four aspects:

(1) Physical topology design: Determine how to interconnect hosts, electrical packet switches (EPS) and optical circuit switches (OCS).
(2) OCS control: Determine the configurations of all the OCS nodes in an ODC.
(3) EPS control: Set up the routing strategies for all the EPS nodes. Since OCS reconfiguration may affect the connected EPS nodes, the queuing and buffer management policies of the EPS nodes may also require careful redesign.
(4) Host control: Modify the host protocol stacks at different layers in accordance with the OCS/EPS control strategies to attain better end-to-end performance.

SIGMETRICS '22, June 06–10, 2022, Mumbai, India

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Existing works on ODC, e.g., Helios [9], c-Through [23], Mordia [21], REACToR [16], OSA [7], Solstice [17], MegaSwitch [8], FireFly [12], ProjectToR [10], RotorNet [19], Opera [18], etc., have primarily focused on the OCS/EPS/Host control aspects, but their physical topology designs are overly simple, which prevents their ODC architectures from supporting large-scale data centers. Only until recently, three scalable physical topology designs are adopted by Flexfly [24], Sirius [2] and 3D-Hyper-FleX-LION [15]. However, the performance guarantee of these physical topologies remain unclear. To the best of our knowledge, there is a lack of rigorous performance metrics that evaluate the "goodness" of a physical topology design, and there is no systematic study on how to design a good physical topology for ODC.

In this paper, we focus on the physical topology design for ODC, and offer a rigorous performance evaluation for different physical topologies. One challenge of this work is to design a performance metric for physical topologies. While much of the literature on ODC has adopted network throughput, flow completion time, max link utilization, etc., to evaluate the end-to-end performance of an ODC, these metrics may not accurately reflect the "goodness" of a physical topology. The reason is that, both the physical topology design and the network control policies may affect these metrics, and thus it is hard to evaluate the exact contribution of a physical topology. To eliminate the impact of network control policies, we introduce $\mu(\mathcal{G}, f)$, which is defined as the optimal throughput across all possible network control policies for a given physical topology $\mathcal{G}$ and a given data center demand pattern f (see §3.2). Based on $\mu(\mathcal{G}, f)$, we then define "β-optimality" for a given physical topology $\mathcal{G}$, i.e., $\mathcal{G}$ is said to be $\beta(\mathcal{G})$-optimal as long as $\mu(\mathcal{G}, f) \geq \beta(\mathcal{G})\mu(\mathcal{G}', f)$ for any physical topology $\mathcal{G}'$ and any demand pattern f (see §3.3). A physical topology $\mathcal{G}$ is said optimal if $\beta(\mathcal{G}) = 1$.

We first study how to design the optimal physical topologies with $\beta(\mathcal{G}) = 1$. A naive approach requires calculating $\mu(\mathcal{G}, f)$ for different f's and optimizing $\mu(\mathcal{G}, f)$ among different physical topologies. Unfortunately, this approach is computationally prohibitive. First, the formulation of $\mu(\mathcal{G}, f)$ is essentially an integer programming problem (see (5) in §3.2), which is NP-hard in general. Second, there are an uncountable number of different traffic patterns, and it is impossible to calculate $\mu(\mathcal{G}, f)$ one by one. To circumvent the above difficulty, we introduce a coupling technique (see Lemma 1 in §3.3) to compare two physical topologies. The intuition is that, if one physical topology can "imitate" all the control strategies of another physical topology, then the first physical topology will perform no worse than the second one. Using this coupling technique, we prove that the ideal physical topology and the uniform bipartite physical topology are both optimal. While these two physical topologies have been adopted by many existing ODC designs [5, 7, 9, 16–19, 21, 23] to build prototypes, they scale poorly to support large data centers.

We then study how to design scalable physical topologies for large ODCs. Along this direction, we first prove a negative result that no physical topology is optimal when the number of EPS nodes exceeds the number of ports of an OCS node. Hence, we have to seek for sub-optimal physical topologies, and calculate $\beta(\mathcal{G})$ as a performance guarantee of these physical topologies. We could still use the coupling technique to calculate $\beta(\mathcal{G})$ for these physical topologies. However, due to the lack of direct connectivity between certain node pairs, it may not always be possible for one physical topology to imitate the control decisions of another physical topology. To overcome this challenge, we enhance the above coupling technique using *overlay topologies*. If one physical topology allows constructing an overlay topology to "imitate" all the control strategies of another physical topology, then the first physical topology will perform no worse than the second one (see Lemma 2 in §3.3). Using the enhanced coupling technique, we successfully calculated $\beta(\mathcal{G})$ for three representative physical topologies.

The contributions of this work are summarized below:

(1) We introduce a new metric $\beta(\mathcal{G})$ to evaluate the goodness of a physical topology $\mathcal{G}$.

(2) We develop a coupling technique to calculate $\beta(\mathcal{G})$. This technique circumvents the computational complexity of calculating $\mu(\mathcal{G}, f)$ for any traffic pattern f.
(3) We design an optimal physical topology (with rigorous proof) for small-scale data centers.
(4) We prove that no physical topology is optimal for large-scale data centers.
(5) Motivated by Flexfly [24], 3D-Hyper-FleX-LION [15] and Sirius [2], we design three physical topologies that can scale to large ODCs, and prove their performance guarantees $\beta(\mathcal{G})$.

2 RELATED WORK

Existing works on optical circuit switched data centers have primarily focused on the overall network control. The adopted physical topology designs either scale poorly, or lack performance guarantee.

c-Through [23], Mordia [21], REACToR [16], OSA [7], Solstice [17] simply use one OCS to connect all the Top-of-Rack (ToR) switches. This physical topology design could only support tiny data centers with $\Theta(100)$ number of servers. To improve scalability, MegaSwitch [8] uses fiber rings for interconnection (a similar idea was also discussed in [21]). This physical topology architecture could support $\Theta(1000)$ number of servers, but still does not work for large-scale data centers with over 100k servers and thousands of ToRs.

Using free-space optics, FireFly [12] and ProjectToR [10] enable optical interconnection for thousands of ToRs. However, the free-space optical switching technology faces tremendous deployment complexity, as many environmental factors (e.g., vibration, dust, and humidity) may hinder the performance of the free-space optical links.

Helios [9], RotorNet [19], Opera [18], TROD [5] create a uniform bipartite graph between all the PoDs/ToRs and all the OCSs. This physical topology turns out to be optimal based on our analysis in §4. While this design scales well for the PoD-level interconnect, it scales poorly for the ToR-level interconnect. A large-scale data center may require hundreds of OCSs to connect over 100k servers. Unfortunately, a ToR switch only has tens of uplinks and thus it is impossible to establish a link between every pair of ToR and OCS. (In contrast, a PoD could have hundreds of uplinks.)

Recently, Flexfly [24], Sirius [2] and 3D-Hyper-FleX-LION [15] offered three physical topologies that can potentially scale to large-scale data centers. However, no performance guarantee is provided for any of the three physical topologies.

3 MATHEMATICAL MODEL

3.1 Basic Definitions

We study a network with N electrical-packet-switching (EPS) nodes $\mathcal{S} = \{S_1, S_2, ..., S_N\}$ and K optical-circuit-switching (OCS) nodes $\mathcal{O} = \{O_1, O_2, ..., O_K\}$. Each EPS node has L ports, each of which has a transceiver. Each OCS node has R ingress ports and R egress ports. Each EPS port can either connect to another EPS port, or connect to a pair of OCS ports. Note that when we connect an EPS port to OCS ports, we need to separate the transmitter and the receiver of the EPS port, have the transmitter connected to an OCS ingress port, and have the receiver connected to an OCS egress port. We do not allow connecting two OCS ports, because signals traversing two OCS nodes will experience too much optical loss. Since an OCS port capacity is typically much higher (100× and more) than an EPS port capacity, the capacity of each network link is determined by the EPS port capacity, and is denoted by B.

Traffic Matrix: Let f_{ij} be the relative traffic demand between the EPS node S_i and the EPS node S_j. Then, the network demand can be modeled using a traffic matrix $f = [f_{ij}, i = 1, 2, ..., N, j = 1, 2, ..., N]$.

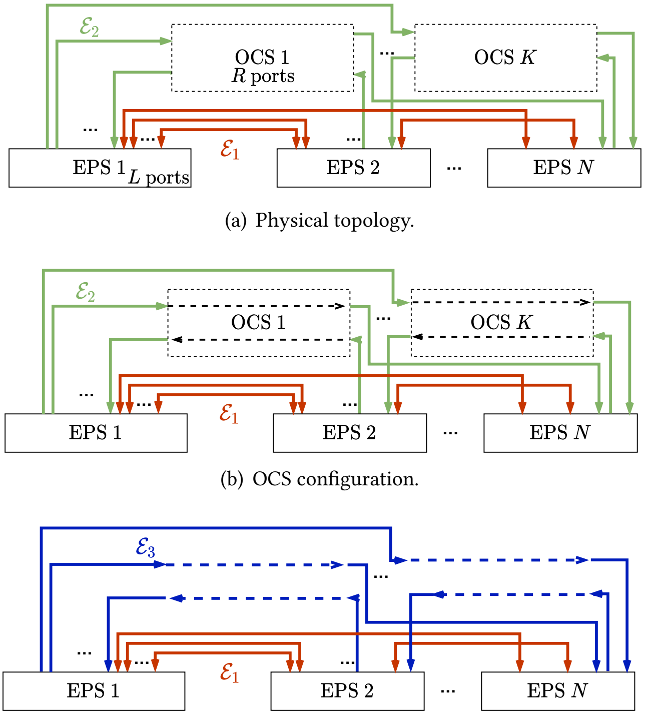


Fig. 1. Physical topology and logical topology. When an OCS configuration acts on a physical topology, the logical topology will be changed. In (a) physical topology, red arrowlines represent the set of bidirectional links that interconnect two EPS nodes and green arrowlines represent the set of unidirectional links that interconnect one EPS node and one OCS node. In (c) logical topology, blue arrowlines represent the set of logical links formed by the OCS configuration. Note that each EPS port can either connect to one red bidirectional arrowline, or connect to two green unidirectional arrowlines with opposite directions.

Physical Topology: We use $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{S} \cup \mathcal{O}$, to denote the physical topology among all the EPS nodes and all the OCS nodes. As shown in Fig. 1(a), the edges of $\mathcal{G}$ consist of two parts $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$, where $\mathcal{E}_1$ is the set of bidirectional links that interconnect two EPS nodes and $\mathcal{E}_2$ is the set of unidirectional links that interconnect one EPS node and one OCS node. The port count constraints of the EPS nodes and the OCS nodes are reflected by the degrees of the nodes in $\mathcal{G}$. Specifically, the in-degree and out-degree of each EPS node are no more than L and the in-degree and out-degree of each OCS node are no more than R. Note that there could be multiple physical links connecting two adjacent nodes.

OCS Configuration: Each OCS node has R ingress ports and R egress ports. As shown in Fig. 1(b), an OCS configuration is essentially a $1-1$ mapping from the R ingress ports to the R egress ports. In this paper, we denote by $\Pi_k, k = 1, 2, ..., K$ as the OCS configuration of O_k, and let $\Pi = \{\Pi_1, \Pi_2, ..., \Pi_K\}$. An OCS configuration helps create multiple *logical links* between EPS node pairs. For example, if an ingress port a is mapped to an egress port b in an OCS node, let S_i be the EPS node that connects to a and S_j be the EPS node that connects to b, then a unidirectional *logical link* is created between S_i and S_j. Since OCS nodes do not perform packet decoding, a packet traversing through a logical link will not be aware of the underlying OCS node.

Logical Topology: Given a physical topology $\mathcal{G}$ and an OCS configuration Π, a logical topology $\mathcal{G}_l = \mathcal{G}_l(\mathcal{G}, \Pi) = (\mathcal{V}_l, \mathcal{E}_l)$ is formed. In the logical topology $\mathcal{G}_l$ (also see Fig. 1(c)), the node set $\mathcal{V}_l = \mathcal{S}$ only contains the EPS nodes; the link set $\mathcal{E}_l = \mathcal{E}_1 \cup \mathcal{E}_3$, where $\mathcal{E}_1$ is the set of physical links among all the EPS nodes and $\mathcal{E}_3$ is the set of logical links formed by the OCS configuration Π. We use a non-negative integer $x_{ij}(\mathcal{G}_l)$ to denote the number of links from S_i to S_j in $\mathcal{G}_l$. Note that $x_{ij}(\mathcal{G}_l)$ could be larger than one.

3.2 Network Control

Given a traffic pattern f and a physical topology $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, network operators could perform different control strategies to optimize network throughput subject to various practical constraints. The purpose of this section is to introduce $\mu(\mathcal{G}, f)$ to measure the goodness of the physical topology $\mathcal{G}$ under an arbitrary traffic pattern f. From a high level, $\mu(\mathcal{G}, f)$ is the network throughput achieved under the optimal control policy. Next, we briefly discuss the network control strategies for ODC, and how these strategies affect network throughput.

We calculate network throughput based on the max-flow formulation. Network operators could perform topology engineering (ToE) and traffic engineering (TE) to optimize its network throughput.

Topology Engineering (ToE): ToE controls the logical topologies used to serve different traffic patterns. For each traffic pattern f, ToE could either generate one logical topology, or generate multiple logical topologies and perform time sharing among these logical topologies. Let $M \geq 1$ be the number of logical topologies generated by ToE. Each logical topology, denoted by $\mathcal{G}_l^{(m)}$, $m = 1, 2, ..., M$, corresponds to a set of K OCS configurations $\Pi^{(m)} = \{\Pi_1^{(m)}, \Pi_2^{(m)}, ..., \Pi_K^{(m)}\}$, and lasts for $\Delta^{(m)}$ amount of time. Note that reconfiguring OCSs could incur a reconfiguration latency δ, which ranges from hundreds of nanoseconds to tens of milliseconds, depending on the optical switching technology being used. Hence, the average bandwidth allocated to (S_u, S_v) can be computed as

$$B_{uv} = \left(\sum_{m=1}^{M} x_{uv}(\mathcal{G}_l^{(m)}) \Delta^{(m)} B \right) \Bigg/ \left(M\delta + \sum_{m=1}^{M} \Delta^{(m)} \right). \tag{1}$$

In addition, traffic patterns may change in an ODC. To model this fact, we impose the following constraint on $\Delta^{(m)}$:

$$M\delta + \sum_{m=1}^{M} \Delta^{(m)} \leq \Delta, \tag{2}$$

where Δ is the duration of the traffic pattern f.

Traffic Engineering (TE): Given a ToE solution, network operators can then perform traffic engineering to maximize network throughput. Here we adopt the max-flow formulation. For every flow from S_i to S_j, let $f_{ij}(u, v)$ be the amount of traffic allocated to (S_u, S_v). The traffic allocation $f_{ij}(u, v)$ must satisfy the following constraints:

(1) Flow conservation constraints: The total ingress traffic equals to the total egress traffic at every EPS node, i.e.,

$$\begin{cases} \sum_{w_1 \neq v} f_{ij}(w_1, v) = \sum_{w_2 \neq v} f_{ij}(v, w_2), \ \forall\ v \neq i, j, \\ \mu f_{ij} + \sum_{w_1 \neq i} f_{ij}(w_1, i) = \sum_{w_2 \neq i} f_{ij}(i, w_2), \\ \sum_{w_1 \neq j} f_{w_1 j}(u, j) = \mu f_{ij} + \sum_{w_2 \neq j} f_{ij}(j, w_2), \end{cases} \tag{3}$$

where μ is a scaling factor of the network traffic pattern f.

(2) Network capacity constraints: For any EPS node pair (S_u, S_v), the total amount of traffic allocated to (S_u, S_v) must be no larger than B_{uv}, i.e.,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}(u,v) \le B_{uv}. \tag{4}$$

Defining $\mu(\mathcal{G}, f)$: Note that we introduced a scaling factor μ in (3). We could optimize μ by solving the following optimization problem:

$$\boxed{\begin{aligned} \max\quad & \mu \\ \text{s.t.}\quad & \mathcal{G}_l^{(m)} = \mathcal{G}_l(\mathcal{G}, \Pi^{(m)}), \forall m = 1, 2, ..., M, \\ & \mu, B_{uv}, f_{ij}(u,v), \mathcal{G}_l^{(m)}, \Delta^{(m)} \text{ satisfy (1)(2)(3)(4).} \end{aligned}} \tag{5}$$

Then, $\mu(\mathcal{G}, f)$ is defined as the optimal objective value of the above optimization problem.
Remark: Even though we have defined $\mu(\mathcal{G}, f)$ rigorously, calculating $\mu(\mathcal{G}, f)$ can be still difficult. First, the number of OCS configurations $[\Pi^{(m)}, m = 1, 2, ..., M]$ grows exponentially with respect to the network size. Second, there are $\Theta(N^4)$ number of routing variables $[f_{ij}(u,v), i, j, u, v = 1, ..., N]$ to be optimized in (5). In fact, (5) is a large scale integer programming problem, which is extremely difficult to solve. We did try solving (5) in Section 7.2, and found that this approach fails to produce a solution in a few hours when there are more than 9 EPS nodes.

3.3 Objective of Physical Topology Design

Definition 1. *A physical topology $\mathcal{G}$ is said to be β-optimal if for every traffic pattern f,*

$$\mu(\mathcal{G}, f) \ge \beta\mu(\mathcal{G}', f)$$

*holds for any possible physical topology $\mathcal{G}'$. β is called the **performance ratio** of $\mathcal{G}$.*

Clearly, every physical topology $\mathcal{G}$ corresponds to a $\beta(\mathcal{G}) \in [0, 1]$. $\beta(\mathcal{G})$ can be viewed as a performance guarantee of the physical topology $\mathcal{G}$, and the larger the better. A physical topology is said to be optimal if its corresponding $\beta = 1$.

However, directly calculating $\beta(\mathcal{G})$ based on (5) and Definition 1 is computationally expensive. First, solving (5) is NP-hard. Second, there are uncountable number of traffic patterns f, and it is impossible to calculate $\mu(\mathcal{G}, f)$ one by one. To circumvent the above difficulty, we introduce a coupling technique to calculate $\beta(\mathcal{G})$. The following two lemmas are the key to this coupling technique.

Lemma 1. *Given two physical topologies $\mathcal{G}$ and $\mathcal{G}'$. If any logical topology $\mathcal{G}_l$ formed on top of $\mathcal{G}$ can be also formed on top of $\mathcal{G}'$, then for any traffic pattern f,*

$$\mu(\mathcal{G}', f) \ge \mu(\mathcal{G}, f).$$

The condition of Lemma 1 ensures that $\mathcal{G}'$ can imitate all the ToE+TE strategies designed for $\mathcal{G}$. Hence, any throughput value achieved by $\mathcal{G}$ is also achievable by $\mathcal{G}'$. Thus, $\mu(\mathcal{G}', f) \ge \mu(\mathcal{G}, f)$.

Lemma 1 can be also generalized to the concept of overlay topology, which is defined below:

Definition 2. *Given a logical topology $\mathcal{G}_l$. A topology $\mathcal{G}_{ol}$ can be formed as an overlay topology of $\mathcal{G}_l$ if we can reserve a number of paths for every source and destination pairs S_i, S_j, such that the following two conditions are met:*

(1) *From the overlay topology $\mathcal{G}_{ol}$ aspect, for every pair of EPS nodes S_i, S_j, the total reserved capacity among all paths is equal to $x_{ij}(\mathcal{G}_{ol})B$.*

(2) *From the underlay topology $\mathcal{G}_l$ aspect, for any EPS node pair (S_i, S_j), the total capacity offered for reservation does not exceed $x_{ij}(\mathcal{G}_l)B$.*

Lemma 2. *Given two physical topologies $\mathcal{G}$ and $\mathcal{G}'$. If for any logical topology $\mathcal{G}_l$ formed on top of $\mathcal{G}$, there is a logical topology $\mathcal{G}'_l$ formed on top of $\mathcal{G}'$, such that $\mathcal{G}_l$ can be formed as an overlay topology of $\mathcal{G}'_l$, then for any traffic pattern f,*

$$\mu(\mathcal{G}', f) \geq \mu(\mathcal{G}, f).$$

If the condition of Lemma 2 holds, $\mathcal{G}'$ can imitate $\mathcal{G}$ using overlay topologies. Note that the TE strategies of $\mathcal{G}$ should be also applied to the overlay topologies of $\mathcal{G}'$. Using this strategy, it is easy to verify that any throughput value achieved by $\mathcal{G}$ is also achievable by $\mathcal{G}'$. Thus, $\mu(\mathcal{G}', f) \geq \mu(\mathcal{G}, f)$.

3.3.1 An Ideal Physical Topology. Fix the number of links L and the link capacity B of an EPS node. Assume that each OCS node has the infinite number of ports. Then, we could connect all the EPS nodes to a single OCS node and obtain an ideal physical topology $\mathcal{G}^{\text{ideal}}_{L,B}$. It is easy to verify that any logical topology formed on top of any physical topology is realizable on top of $\mathcal{G}^{\text{ideal}}_{L,B}$. Then according to Lemma 1, $\mathcal{G}^{\text{ideal}}_{L,B}$ is optimal, as stated below:

Theorem 3. *For any fixed number of links L and fixed link capacity B, $\mathcal{G}^{ideal}_{L,B}$ is optimal.*

Based on the ideal physical topology $\mathcal{G}^{\text{ideal}}_{L,B}$, we can then define the ideal throughput $\mu^{\text{ideal}}_{L,B}(f)$ for an arbitrary traffic pattern f as

$$\mu^{\text{ideal}}_{L,B}(f) = \mu(\mathcal{G}^{\text{ideal}}_{L,B}, f).$$

Clearly, $\mu^{\text{ideal}}_{L,B}(f)$ scales linearly with respect to B. Further, as the number of links L of an EPS node increases, the ideal throughput $\mu^{\text{ideal}}_{L,B}(f)$ would never decrease.

4 OPTIMAL PHYSICAL TOPOLOGY FOR SMALL DATA CENTERS WITH $N \leq R$

Building an ideal physical topology requires $R \geq NL$. This only applies to tiny data centers with at most tens of EPS nodes[1]. In this section, we study a less restrictive case where $N \leq R$, which corresponds to hundreds of EPS nodes.

Before proposing our physical topology design, we first introduce the following lemma.

Lemma 4. *Given an $N \times N$ non-negative integer matrix $\mathbf{x} = \{x_{i,j}\}$, for any integer $H \geq 1$ and mutually disjoint sets $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_C$ satisfying $\cup_{c=1}^{C} \mathcal{A}_c = \{1, 2, ..., N\}$ and $\mathcal{A}_{c_1} \cap \mathcal{A}_{c_2} = \emptyset, \forall c_1 \neq c_2$, there must exist H non-negative integer matrices $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(H)}$ satisfying*

(1) $\mathbf{x} = \mathbf{x}^{(1)} + \cdots + \mathbf{x}^{(H)}$;

(2) $0 \leq \sum_{j=1}^{N} x_{i,j}^{(h)} \leq \left\lceil \frac{\sum_{j=1}^{N} x_{i,j}}{H} \right\rceil$, *for any $i = 1, ..., N$ and any $h = 1, 2, ..., H$;*

(3) $0 \leq \sum_{i \in \mathcal{A}_c} \sum_{j=1}^{N} x_{i,j}^{(h)} \leq \left\lceil \frac{\sum_{i \in \mathcal{A}_c} \sum_{j=1}^{N} x_{i,j}}{H} \right\rceil$, *for any $c = 1, ..., C$ and any $h = 1, 2, ..., H$;*

(4) $0 \leq \sum_{i=1}^{N} x_{i,j}^{(h)} \leq \left\lceil \frac{\sum_{i=1}^{N} x_{i,j}}{H} \right\rceil$, *for any $j = 1, ..., N$ and any $h = 1, 2, ..., H$;*

(5) $0 \leq \sum_{i=1}^{N} \sum_{j \in \mathcal{A}_c} x_{i,j}^{(h)} \leq \left\lceil \frac{\sum_{i=1}^{N} \sum_{j \in \mathcal{A}_c} x_{i,j}}{H} \right\rceil$, *for any $c = 1, ..., C$ and any $h = 1, 2, ..., H$.*

Lemma 4 can be proved by transforming it into a sequence of max-flow problems. It is actually a direct consequence of Theorem 3 in [25] by setting $\mathscr{A} = \mathscr{B} = \{\mathcal{A}_1, ..., \mathcal{A}_C, \{1\}, ..., \{N\}\}$. See Appendix A in [25] for the detailed proof.

[1] A practical OCS node has hundreds of ports and a practical ToR switches has tens of ports.

4.1 Physical Topology Design

Motivated by Lemma 4, we set the number of OCS nodes as $K = L$ and evenly distribute the L links of each EPS node across all the OCS nodes. In other words, the physical topology $\mathcal{G}$ becomes a uniform bipartite graph, with exactly one bidirectional physical link between every pair of EPS node and OCS node. (More specifically, the transmitter of the EPS port is connected to an OCS ingress port, and the receiver of the EPS port is connected to an OCS egress port.) Note that $N \le R$ ensures that the number of ports of each OCS node is not exhausted. We denote such a physical topology by $\mathcal{G}_{\text{uniform}}$, as shown in Fig. 2. We can prove that $\mathcal{G}_{\text{uniform}}$ is optimal.

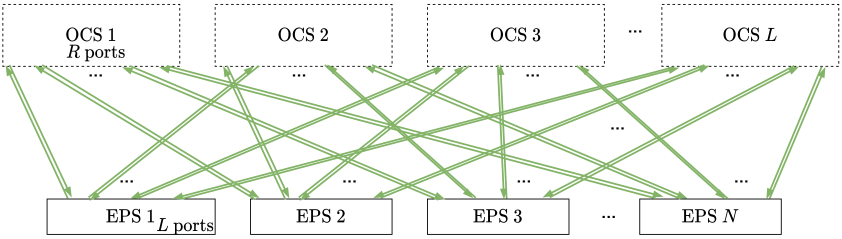


Fig. 2. $\mathcal{G}_{\text{uniform}}$ topology.

Theorem 5. *If $N \le R$, then $\mathcal{G}_{uniform}$ is optimal.*

Proof. We only need to prove that any logical topology $\mathcal{G}_l$ is realizable on top of the physical topology $\mathcal{G}_{\text{uniform}}$.

Since each EPS node has L ports, we must have

$$\begin{cases} \sum_{i=1}^{N} x_{ij}(\mathcal{G}_l) \le L, \ \forall\ j = 1, ..., N, \\ \sum_{j=1}^{N} x_{ij}(\mathcal{G}_l) \le L, \ \forall\ i = 1, ..., N. \end{cases} \tag{6}$$

Then, according to Lemma 4, we can decompose $\{x_{ij}(\mathcal{G}_l)\}$ into $H = K = L$ integer matrices $\{x_{ij}^{(h)}(\mathcal{G}_l)\}, h = 1, 2, ..., H$ satisfying the conditions in Lemma 4. Based on (6) and the conditions (2) and (4) of Lemma 4, it is easy to check that any $\{x_{ij}^{(h)}(\mathcal{G}_l)\}, h = 1, 2, ..., H$, must satisfy

$$\begin{cases} \sum_{i=1}^{N} x_{ij}^{(h)}(\mathcal{G}_l) \le 1, \ \forall\ j = 1, ..., N, \\ \sum_{j=1}^{N} x_{ij}^{(h)}(\mathcal{G}_l) \le 1, \ \forall\ i = 1, ..., N. \end{cases}$$

Since $x_{ij}^{(h)}(\mathcal{G}_l)$'s are non-negative integers, $x_{ij}^{(h)}(\mathcal{G}_l)$ is actually a permutation matrix. Note that each EPS node has one bidirectional physical link connected to the h-th OCS node, it is easy to verify that the permutation matrix $x_{ij}^{(h)}(\mathcal{G}_l)$ is realizable on the h-th OCS node.

□

5 NEGATIVE RESULT ON PHYSICAL TOPOLOGY DESIGN

Having designed an optimal physical topology for the case $N \le R$, we thus wonder if it is possible to design an optimal physical topology for large-scale data centers with $N > R$. Unfortunately, the following theorem demonstrates the impossibility of such a design.

Theorem 6. *If $N > R$, then no physical topology is β-optimal for $\beta > \frac{N+R-1}{2N-2}$. In other words, for any physical topology $\mathcal{G}$, there exists a traffic pattern f and a physical topology $\mathcal{G}'$, such that*

$$\mu(\mathcal{G}, f) \le \frac{N + R - 1}{2N - 2}\mu(\mathcal{G}', f).$$

Proof. We pick R EPS nodes from the N EPS nodes, and form a permutation $\mathcal{Z} = (k_1, k_2, ..., k_R)$ among the R EPS nodes. In total, we obtain $N(N-1)\cdots(N-R+1)$ different permutations.

For each permutation $\mathcal{Z} = (k_1, k_2, ..., k_R)$, we count the maximum number of unidirectional logical links that can be formed for $S_{k_1} \to S_{k_2}, S_{k_2} \to S_{k_3}, ..., S_{k_R} \to S_{k_1}$ under a physical topology $\mathcal{G}$, and denote this number by $N(\mathcal{G}, \mathcal{Z})$.

Step 1: We would like to prove that for any physical topology $\mathcal{G}$, there must exist a permutation $\mathcal{Z}^*$, such that

$$N(\mathcal{G}, \mathcal{Z}^*) \leq LR\frac{R}{N-1}.$$

Let $N_{i\to j}$ be the maximum number of links that can be created from the EPS node S_i to the EPS node S_j. We compute

$$\sum_{\mathcal{Z}} N(\mathcal{G}, \mathcal{Z}) = \sum_{\mathcal{Z}} \sum_{i,j:i\neq j} 1_{\{j \text{ is after } i \text{ in } \mathcal{Z}\}} N_{i\to j} = \sum_{i,j:i\neq j} N_{i\to j} \sum_{\mathcal{Z}} 1_{\{j \text{ is after } i \text{ in } \mathcal{Z}\}}.$$

Here, "j is after i in $\mathcal{Z} = (k_1, k_2, ..., k_R)$" means that either there exists $r = 1, 2, ..., R-1$, such that $k_r = i, k_{r+1} = j$, or $k_R = i, k_1 = j$. In order to compute the number of permutations in which j is after i, i.e., $\sum_{\mathcal{Z}} 1_{\{j \text{ is after } i \text{ in } \mathcal{Z}\}}$,

(1) we first find a permutation with $R-2$ elements excluding i and j, and the total number of such permutations is $(N-2)(N-3)\cdots(N-R+1)$;
(2) we then insert i and j to the right places of the above permutation, and the number of different inserting positions is R. (For example, when $R = 4$, $ij**$, $*ij*$, $**ij$ and $j**i$ are the 4 generated permutations.)

Hence,

$$\sum_{\mathcal{Z}} 1_{\{j \text{ is after } i \text{ in } \mathcal{Z}\}} = R(N-2)(N-3)\cdots(N-R+1).$$

We then focus on $\sum_{i,j:i\neq j} N_{i\to j}$. We first fix i and compute $\sum_{j:j\neq i} N_{i\to j}$. Assume that the EPS node S_i has at least one transmitter connected to the OCS node O_k. Let R_k be the number of links between the egress ports of the OCS node O_k and the receivers of all the EPS nodes excluding S_i. Then it is easy to check that the contribution of the OCS node O_k to $\sum_{j:j\neq i} N_{i\to j}$ is at most $R_k \leq R$. Since the EPS node S_i can connect to at most L different OCS nodes, we must have

$$\sum_{j:j\neq i} N_{i\to j} \leq LR, \text{ for any fixed } i.$$

Therefore,

$$\sum_{i,j:i\neq j} N_{i\to j} = \sum_{i=1}^{N} \sum_{j:j\neq i} N_{i\to j} \leq NLR. \tag{7}$$

Note that there are $N(N-1)\cdots(N-R+1)$ different permutations in total. According to the Pigeonhole Principle, there must exist a permutation $\mathcal{Z}^*$ such that

$$N(\mathcal{G}, \mathcal{Z}^*) \leq \frac{\sum_{\mathcal{Z}} N(\mathcal{G}, \mathcal{Z})}{N(N-1)\cdots(N-R+1)} \leq LR\frac{R}{N-1}.$$

Step 2: We construct a permutation traffic pattern f^* based on $Z^* = (k_1^*, k_2^*, ..., k_R^*)$, i.e.,

$$f_{ij}^* = \begin{cases} 1, \text{ if } i = k_r^*, j = k_{r+1}^*, \text{ where } r = 1, 2, ..., R-1; \\ 1, \text{ if } i = k_R^*, j = k_1^*; \\ 0, \text{ otherwise.} \end{cases}$$

Clearly, we can create a physical topology $\mathcal{G}(Z^*)$ by picking L OCS nodes and connecting these OCS nodes to the R EPS nodes $S_{k_1^*}, S_{k_2^*}, ..., S_{k_R^*}$ using a uniform bipartite graph. On top of this physical topology, we can then create L identical cycles $k_1^* \to k_2^* \to ... \to k_R^* \to k_1^*$ as the logical topology. It is easy to check that this logical topology offers the highest throughput for the traffic matrix f^* over the physical topology $\mathcal{G}(Z^*)$, and $\mu(\mathcal{G}(Z^*), f^*) = BL$.

We then compute the throughput $\mu(\mathcal{G}, f^*)$ for f^* over $\mathcal{G}$. Consider the optimal ToE+TE strategy that attains the throughput value $\mu(\mathcal{G}, f^*)$. Then, within a time period of Δ,

$$\mu(\mathcal{G}, f^*)\Delta \sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}^* = \mu(\mathcal{G}, f^*)\Delta R \tag{8}$$

amount of traffic is delivered from sources to destinations. The delivered traffic can be grouped into two classes:

(1) Class-1: traffic that is delivered via only one hop;
(2) Class-2: traffic that is delivered via at least two hops.

Class-1 traffic must occupy one of following types of logical links $S_{k_1^*} \to S_{k_2^*}, ..., S_{k_{R-1}^*} \to S_{k_R^*}, S_{k_R^*} \to S_{k_1^*}$. Since the total number of such links is at most $N(\mathcal{G}, \mathcal{Z}^*)$, we must have

$$\text{Class-1 traffic} \le N(\mathcal{G}, \mathcal{Z}^*)\Delta B. \tag{9}$$

Every byte of the Class-2 traffic occupies at least two bytes of the network capacity. Hence,

$$2 \times \text{Class-2 traffic} \le (LR - N(\mathcal{G}, \mathcal{Z}^*))\Delta B. \tag{10}$$

Based on (8)(9)(10), we then obtain

$$\begin{aligned}\mu(\mathcal{G}, f^*) &\le \frac{1}{R}\left(BN(\mathcal{G}, \mathcal{Z}^*) + \frac{B}{2}\left(LR - N(\mathcal{G}, \mathcal{Z}^*)\right)\right) = \frac{BL}{2} + \frac{B}{2R}N(\mathcal{G}, \mathcal{Z}^*) \\ &\le \frac{BL}{2} + \frac{BL}{2}\frac{R}{N-1} = \frac{N+R-1}{2N-2}\mu(\mathcal{G}(Z^*), f^*).\end{aligned}$$

This completes the proof. □

Discussion: Note that $\frac{N+R-1}{2N-2} < 1$ when $N > R + 1$. Then according to Theorem 6, no physical topology is optimal. Readers may wonder if it is possible to construct an optimal physical topology for the case where $N = R + 1$. The answer is no. When $N = R + 1$, $\frac{N+R-1}{2N-2} = 1$. This indicates that, if there exists an optimal physical topology, the "=" must hold for the inequality (7). This "=" requires that the $2R$ OCS ports (R ingress ports and R egress ports) must connect to $2R$ different EPS nodes. This is impossible because there are only $N = R + 1$ EPS nodes.

6 SCALABLE PHYSICAL TOPOLOGY DESIGNS WITH PERFORMANCE GUARANTEE

In this section, we propose three scalable physical topology designs and analyze their performance guarantees. The main results in this section are summarized in Table 1. Note that the maximum scale of different physical topologies depends on the OCS node port count R and the EPS node port count L. A commercially-available OCS node, e.g., the 3D-MEMS based OCS [3] and the Arrayed Wavelength Grating Routers [2], could have hundreds of ports. An off-the-shelf ToR EPS node, e.g., the CloudEngine 8800 [13], could have 32 ∼ 64 ports. Hence, the physical topologes studied in this section could support large-scale ODCs with thousands or even tens of thousands of EPS nodes.

6.1 EPS-Group based Physical Topology

Physical topology of $\mathcal{G}_{\textbf{eps-group}}$: Let $P = \lceil L/3 \rceil$. We create C EPS node groups, each of which contains P EPS nodes and is denoted by $\mathcal{A}_c$. Within each EPS node group, there are two links between every pair of EPS nodes. Each EPS node has $2(P-1)$ intra-group links, and the total number of links in each EPS node group is $P(P-1)$. Each EPS node also has P links connected

Table 1. The scalability and the performance ratios of different physical topologies.

Physical Topology	Maximum Number of EPS Nodes	Performance Ratio β
$\mathcal{G}_{\text{uniform}}$	R	1
$\mathcal{G}_{\text{eps-group}}$	$R\lceil L/3\rceil$	$\min_f \left\{\mu^{\text{ideal}}_{\lceil L/3\rceil,B}(f)/\mu^{\text{ideal}}_{L,B}(f)\right\}$
$\mathcal{G}_{\text{eps-mesh}}$	R^2	$\min_f \left\{\mu^{\text{ideal}}_{\lfloor L/3\rfloor,B}(f)/\mu^{\text{ideal}}_{L,B}(f)\right\}$
$\mathcal{G}_{\text{ocs-mesh}}$	$\lfloor R/2\rfloor\lfloor L/2\rfloor$	$\min_f \left\{\mu^{\text{ideal}}_{\lfloor L/2\rfloor,B}(f)/\mu^{\text{ideal}}_{L,B}(f)\right\}$

to the OCS nodes to establish inter-group connections. It is easy to verify that the degree of each EPS node is $3P - 2 \le L$. We use $K = P^2$ OCS nodes and divide these OCS nodes into P equal-sized groups. We number the EPS nodes in a group from 1 to P. The p-th EPS node in a group has one link connected to every OCS node in the p-th OCS node group, where $p = 1, 2, ..., P$. Since the number of ports of an OCS node is R, the number of EPS node groups of $\mathcal{G}_{\text{eps-group}}$ must satisfy $C \le R$. Thus, the total number of EPS nodes in $\mathcal{G}_{\text{eps-group}}$ satisfies $N = CP \le R\lceil L/3\rceil$.

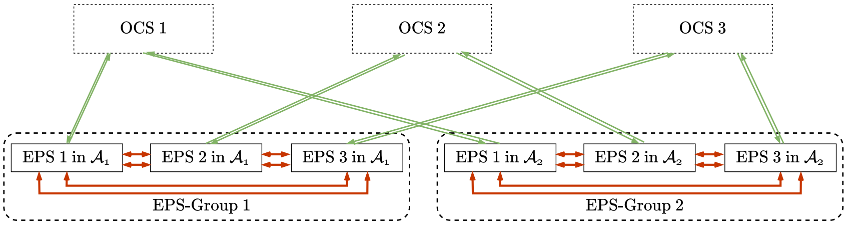


Fig. 3. Toy example of EPS-Group topology ($R = 2, N = 6$).

Theorem 7. *For any traffic pattern f, the optimal throughput under $\mathcal{G}_{eps\text{-}group}$ satisfies*

$$\mu(\mathcal{G}_{eps\text{-}group}, f) \ge \mu^{ideal}_{\lceil L/3\rceil,B}(f).$$

Proof. Let $\mathcal{G}^{\text{ideal}}_{\lceil L/3\rceil,B}$ be the ideal physical topology. According to Lemma 2, we need to show that any logical topology $\mathcal{G}_l$ formed over $\mathcal{G}^{\text{ideal}}_{\lceil L/3\rceil,B}$ can be realized as a overlay topology over $\mathcal{G}_{\text{eps-group}}$.

Since each EPS node in $\mathcal{G}^{\text{ideal}}_{\lceil L/3\rceil,B}$ has $P = \lceil L/3\rceil$ number of uplinks, the following constraints must be satisfied

$$\begin{cases} \sum_{i=1}^{N} x_{ij}(\mathcal{G}_l) \le \lceil L/3\rceil = P, \ \forall\, j = 1, ..., N, \\ \sum_{j=1}^{N} x_{ij}(\mathcal{G}_l) \le \lceil L/3\rceil = P, \ \forall\, i = 1, ..., N. \end{cases} \tag{11}$$

Note that it may not always be possible to establish a logical link between two EPS nodes in $\mathcal{G}_{\text{eps-group}}$, especially when the two EPS nodes are in different EPS node groups and these two nodes have different intra-group indices. Hence, we are interested in constructing $\mathcal{G}_l$ as an overlay topology on top of $\mathcal{G}_{\text{eps-group}}$. For every pair of EPS nodes S_i and S_j in $\mathcal{G}_{\text{eps-group}}$, we would like to create $x_{ij}(\mathcal{G}_l)$ number of virtual logical links in between. Here, a **virtual logical links** is essentially a path between two EPS nodes, with each path segment occupying exactly one logical link.

For every EPS node pair (S_i, S_j), we classify this pair's virtual logical links into $P = \lceil L/3\rceil$ types:

- If S_i and S_j are in the same EPS node group, the type-p virtual logical links are two-hop paths with intermediate node being the p-th EPS node in this EPS node group. It is possible that the p-th EPS node is exactly S_i or S_j. In this case, this "two-hop path" degenerates to a single-hop path.

- If S_i and S_j belong to different EPS node groups, the type-p virtual logical links are three-hop paths with intermediate nodes being the p-th EPS node in the EPS node group of S_i and the p-th EPS node in the EPS node group of S_j. It is possible that the p-th EPS node in the EPS node group of S_i is exactly S_i, or the p-th EPS node in the EPS node group of S_j is exactly S_j. In this case, this "three-hop path" degenerates to a two/one-hop path.

We use $x_{ij}^p, p = 1, 2, ..., P$ to denote the total number of type-p paths between S_i and S_j. According to Lemma 4, there exists an integer solution of $[x_{ij}^p]$ satisfying the following constraints:

(1) $x_{ij} = \sum_{p=1}^{P} x_{ij}^p, \forall i, j = 1, ..., N$;
(2) $0 \le \sum_{j=1}^{N} x_{ij}^p \le 1, \forall i = 1, ..., N, p = 1, ..., P$;
(3) $0 \le \sum_{i \in \mathcal{A}_c} \sum_{j=1}^{N} x_{ij}^p \le P, \forall c = 1, ..., C, p = 1, ..., P$;
(4) $0 \le \sum_{i=1}^{N} x_{ij}^p \le 1, \forall j = 1, ..., N, p = 1, ..., P$;
(5) $0 \le \sum_{i=1}^{N} \sum_{j \in \mathcal{A}_c} x_{ij}^p \le P, \forall c = 1, ..., C, p = 1, ..., P$.

Note that x_{ij}^p determines the underlying logical topology of the overlay topology $\mathcal{G}_l$. We need to show that this logical topology is compatible with the physical topology $\mathcal{G}_{\text{eps-group}}$.

Intra-group Logical Topology: If S_i and S_j are in the same EPS node group, then the number of links in between is

$$\sum_{u=1}^{N} x_{iu}^{(j \bmod P)} + \sum_{v=1}^{N} x_{vj}^{(i \bmod P)}.$$

Based on the constraints 2) and 4), it is easy to check that the above number is no greater than 2, which is compatible with intra-group physical topology design of $\mathcal{G}_{\text{eps-group}}$.

Inter-group Logical Topology: Fix $p = 1, 2, ..., P$. Let $S_{(c_1-1)P+p}$ and $S_{(c_2-1)P+p}$ be the p-th EPS node in the EPS node groups $\mathcal{A}_{c_1}$ and $\mathcal{A}_{c_2}$, respectively. Then the total number of links from $S_{(c_1-1)P+p}$ to $S_{(c_2-1)P+p}$ is

$$y_{c_1c_2}^p = \sum_{i \in \mathcal{A}_{c_1}} \sum_{j \in \mathcal{A}_{c_2}} x_{ij}^p.$$

Based on the constraints 3) and 5), it is easy to verify that

$$\sum_{c_1=1}^{C} y_{c_1c_2}^w \le P, \sum_{c_2=1}^{C} y_{c_1c_2}^w \le P.$$

Then, using the same arguments in the proof of Theorem 5, we can prove that the logical topology $y_{c_1c_2}^p, c_1, c_2 = 1, 2, ..., C$ is realizable on top of $\mathcal{G}_{\text{eps-group}}$. This completes the proof. □

Remark: Theorem 7 implies that $\beta(\mathcal{G}_{\text{eps-group}}) = \min_f \{\mu_{\lceil L/3 \rceil, B}^{\text{ideal}}(f) / \mu_{L,B}^{\text{ideal}}(f)\}$. $\mathcal{G}_{\text{eps-group}}$ is similar to the physical topology used by Flexfly [24]. In order to prove the performance guarantee in Theorem 7, we have restricted $\mathcal{G}_{\text{eps-group}}$ to use uniform mesh for its intra-group topology, which limits the scale of $\mathcal{G}_{\text{eps-group}}$. One could use flattened butterfly [14] for intra-group interconnect to increase the size of each EPS group of $\mathcal{G}_{\text{eps-group}}$, as suggested by Flexfly [24]. However, the performance guarantee would become much poorer.

6.2 EPS-Mesh based Physical Topology Design

Physical topology of $\mathcal{G}_{\text{eps-mesh}}$: Let $P = \lfloor L/3 \rfloor$. We arrange all the ESP nodes into a $C \times W$ mesh. The W EPS nodes in the c-th row forms an EPS node group, denoted by $\mathcal{A}_c$, $c = 1, 2, ..., C$. We index all the EPS nodes row by row, from 1 to CW. There are $2P$ OCS nodes dedicated for each EPS node group $\mathcal{A}_c$. For every EPS node in $\mathcal{A}_c$, there is one bidirectional link between this EPS node and each of the $2P$ OCS nodes. There are also P OCS nodes dedicated for the EPS nodes in each column.

For the w-th column, there is one bidirectional link between every EPS node in this column and every OCS node dedicated for this column. Since the number of ports of an OCS node is R, the number of EPS nodes in each row or each column must be no more than R. Thus, the total number of EPS nodes must satisfy $N = CW \leq R^2$.

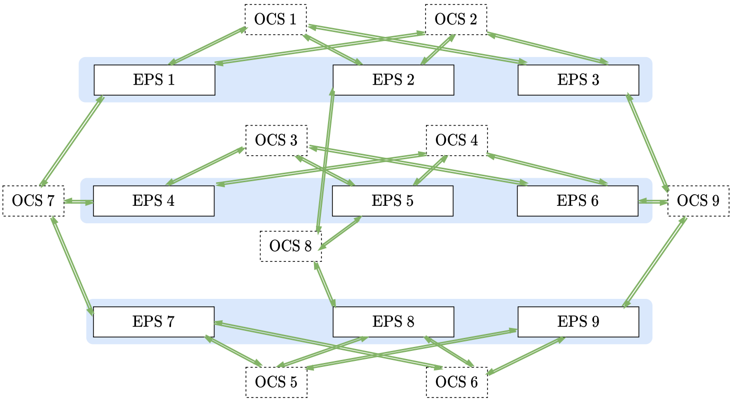


Fig. 4. Toy example of EPS-Mesh topology ($R = 3, N = 9$).

Theorem 8. *For any traffic pattern f, the optimal throughput under $\mathcal{G}_{eps\text{-}mesh}$ satisfies*

$$\mu(\mathcal{G}_{eps\text{-}mesh}, f) \geq \mu^{ideal}_{\lfloor L/3 \rfloor, B}(f).$$

Proof. Let $\mathcal{G}^{\text{ideal}}_{\lfloor L/3 \rfloor, B}$ be the ideal physical topology. According to Lemma 2, we need to show that any logical topology $\mathcal{G}_l$ formed over $\mathcal{G}^{\text{ideal}}_{\lfloor L/3 \rfloor, B}$ can be realized as a overlay topology over $\mathcal{G}_{\text{eps-mesh}}$.

Since each EPS node in $\mathcal{G}^{\text{ideal}}_{\lfloor L/3 \rfloor, B}$ has $P = \lfloor L/3 \rfloor$ number of uplinks, the following constraints must be satisfied

$$\begin{cases} \sum_{i=1}^{N} x_{ij}(\mathcal{G}_l) \leq \lfloor L/3 \rfloor = P, \ \forall \ j = 1, ..., N, \\ \sum_{j=1}^{N} x_{ij}(\mathcal{G}_l) \leq \lfloor L/3 \rfloor = P, \ \forall \ i = 1, ..., N. \end{cases} \tag{12}$$

Similar to $\mathcal{G}_{\text{eps-group}}$, it may not always be possible to establish a logical link between two EPS nodes in $\mathcal{G}_{\text{eps-mesh}}$ as well, especially when the two EPS nodes are in different rows and different columns. Hence, we will construct $\mathcal{G}_l$ as an overlay topology on top of $\mathcal{G}_{\text{large}}$. For every pair of EPS nodes S_i and S_j in $\mathcal{G}_{\text{eps-mesh}}$, we would like to create $x_{ij}(\mathcal{G}_l)$ number of virtual logical links in between, and classify these virtual logical links into W types:

- If S_i and S_j are in the same row, the type-w virtual logical links are two-hop paths with intermediate node being the w-th EPS node in this row.
- If S_i and S_j belong to different rows, the type-w virtual logical links are three-hop paths with intermediate nodes being the w-th EPS node in the same row as S_i and the w-th EPS node in the same row as S_j.

We use $x_{ij}^w, w = 1, 2, ..., W$ to denote the total number of type-w paths between S_i and S_j. According to Lemma 4, there exists an integer solution of $x_{ij}^w, w = 1, 2, ..., W$ satisfying the following constraints:

(1) $x_{ij}(\mathcal{G}_l) = \sum_{w=1}^{W} x_{ij}^w, \forall i, j = 1, ..., N;$

(2) $0 \leq \sum_{i \in \mathcal{A}_c} \sum_{j=1}^{N} x_{ij}^w \leq P, \forall c = 1, ..., C, w = 1, ..., W;$

(3) $0 \leq \sum_{i=1}^{N} \sum_{j \in \mathcal{A}_c} x_{ij}^w \leq P, \forall c = 1, ..., C, w = 1, ..., W.$

Note that x_{ij}^{w} determines the underlying logical topology of the overlay topology $\mathcal{G}_l$. We need to show that this logical topology is compatible with the physical topology $\mathcal{G}_{\text{eps-group}}$.

Inter-group Logical Topology: Fix $w = 1, 2, ..., W$. Let $S_{(c_1-1)W+w}$ and $S_{(c_2-1)W+w}$ be the w-th EPS node in the EPS node groups $\mathcal{A}_{c_1}$ and $\mathcal{A}_{c_2}$, respectively. Then the total number of links from $S_{(c_1-1)W+w}$ to $S_{(c_2-1)W+w}$ is

$$y_{c_1c_2}^{w} = \sum_{i\in\mathcal{A}_{c_1}}\sum_{j\in\mathcal{A}_{c_2}} x_{ij}^{w}.$$

Based on the constraints 2) and 3), it is easy to verify that

$$\sum_{c_1=1}^{C} y_{c_1c_2}^{w} \leq P, \sum_{c_2=1}^{C} y_{c_1c_2}^{w} \leq P.$$

Then, using the same arguments in the proof of Theorem 5, we can prove that the logical topology $y_{c_1c_2}^{w}, c_1, c_2 = 1, 2, ..., C$ is realizable on the P OCS nodes dedicated for the w-th row.

Intra-group Logical Topology: Fix $c = 1, 2, ..., C$. Let $S_{(c-1)W+w_1}$ and $S_{(c-1)W+w_2}$ be the w_1-th and the w_2-th EPS nodes in $\mathcal{A}_c$. Then the total number of links from $S_{(c-1)W+w_1}$ to $S_{(c-1)W+w_2}$ is

$$z_{w_1,w_2}^{c} = \sum_{u=1}^{N} x_{(c-1)W+w_1,u}^{w_2} + \sum_{v=1}^{N} x_{v,(c-1)W+w_2}^{w_1}.$$

According to Eqn. (13) and the constraints 2) and 3), we can verify that

$$\sum_{w_1=1}^{W} z_{w_1w_2}^{c} \leq 2P, \sum_{w_2=1}^{W} z_{w_1w_2}^{c} \leq 2P.$$

Take the first one for example:

$$\begin{aligned}\sum_{w_1=1}^{W} z_{w_1w_2}^{c} &= \sum_{w_1=1}^{W}\sum_{u=1}^{N} x_{(c-1)W+w_1,u}^{w_2} + \sum_{w_1=1}^{W}\sum_{v=1}^{N} x_{v,(c-1)W+w_2}^{w_1} \\ &= \sum_{i\in\mathcal{A}_c}\sum_{u=1}^{N} x_{iu}^{w_2} + \sum_{v=1}^{N} x_{v,(c-1)W+w_2}(\mathcal{G}_l) \leq 2P.\end{aligned}$$

Again, using the same arguments in the proof of Theorem 5, we can prove that the logical topology $z_{w_1w_2}^{c}, w_1, w_2 = 1, 2, ..., W$ is realizable on the $2P$ OCS nodes dedicated for the EPS node group $\mathcal{A}_c$. This completes the proof. □

Remark: Theorem 8 implies that $\beta(\mathcal{G}_{\text{eps-mesh}}) = \min_f\{\mu_{\lfloor L/3\rfloor,B}^{\text{ideal}}(f)/\mu_{L,B}^{\text{ideal}}(f)\}$. The design of $\mathcal{G}_{\text{eps-mesh}}$ shares a similar idea as the physical topology used by 3D-Hyper-FleX-LION [15]. Although 3D-Hyper-FleX-LION uses a 3D-mesh instead, it is easy to generalize $\mathcal{G}_{\text{eps-mesh}}$ from 2D-mesh to 3D-mesh to further improve its scalability. Notably, one critical design of $\mathcal{G}_{\text{eps-mesh}}$ is that the EPS ports used for intra-group interconnect is "two" times the EPS ports used for inter-group interconnect. This design allows us to attain the best performance guarantee.

6.3 OCS-Mesh based Physical Topology Design

Physical Topology of $\mathcal{G}_{\text{ocs-mesh}}$: We arrange C^2 OCS nodes into an $C\times C$ mesh, where $C = \lfloor L/2\rfloor$. The EPS nodes are arranged into C groups, each of which contains W EPS nodes and is denoted by $\mathcal{A}_c, c = 1, 2, ..., C$. For the EPS node group $\mathcal{A}_c$, we connect two directed links from the transmitters of each EPS node in $\mathcal{A}_c$ to the ingress ports of each OCS node in the c-th row; we also connect two directed links from the egress ports of each OCS node in the c-th column to the receivers of each EPS node in $\mathcal{A}_c$. Since each OCS node has R ingress/egress ports, the number of EPS nodes in each group must satisfy $W \le \lfloor R/2 \rfloor$. Hence, the total number of EPS nodes must satisfy $N = CW \le \lfloor L/2 \rfloor \lfloor R/2 \rfloor$.

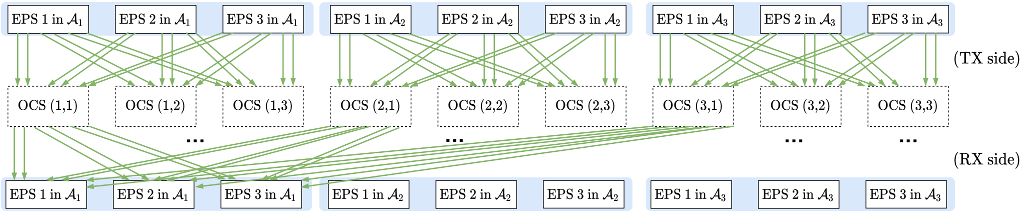


Fig. 5. Toy example of OCS-Mesh topology ($R = 6, N = 9$). Some RX side links are omitted for clear visualization. Each EPS node in $\mathcal{A}_2$ is connected to two egress ports of OCS(1,2), OCS(2,2) and OCS(3,2). Each EPS node in $\mathcal{A}_3$ is connected to two egress ports of OCS(1,3), OCS(2,3) and OCS(3,3).

Theorem 9. *For any traffic pattern f, the optimal throughput under $\mathcal{G}_{ocs\text{-}mesh}$ satisfies*

$$\mu(\mathcal{G}_{ocs\text{-}mesh}, f) \ge \mu^{ideal}_{\lfloor L/2 \rfloor, B}(f).$$

Proof. Let $\mathcal{G}^{\text{ideal}}_{\lfloor L/2 \rfloor, B}$ be the ideal physical topology. According to Lemma 2, we need to show that any logical topology $\mathcal{G}_l$ formed over $\mathcal{G}^{\text{ideal}}_{\lfloor L/2 \rfloor, B}$ can be realized as a overlay topology over $\mathcal{G}_{\text{ocs-mesh}}$.

Since each EPS node in $\mathcal{G}^{\text{ideal}}_{\lfloor L/2 \rfloor, B}$ has $C = \lfloor L/2 \rfloor$ number of uplinks, the following constraints must be satisfied

$$\begin{cases} \sum_{i=1}^{N} x_{ij}(\mathcal{G}_l) \le \lfloor L/2 \rfloor = C, \ \forall\, j = 1, ..., N, \\ \sum_{j=1}^{N} x_{ij}(\mathcal{G}_l) \le \lfloor L/2 \rfloor = C, \ \forall\, i = 1, ..., N. \end{cases} \tag{13}$$

Again, we construct $\mathcal{G}_l$ as an overlay topology on top of $\mathcal{G}_{\text{ocs-mesh}}$. For every pair of EPS nodes S_i and S_j in $\mathcal{G}_{\text{ocs-mesh}}$, we would like to create $x_{ij}(\mathcal{G}_l)$ number of virtual logical links. Each virtual logical link is essentially a two-hop path $S_i \longrightarrow S_k \longrightarrow S_j, k = 1, 2, ..., N$, and we use x_{ij}^k to denote the number of such virtual logical links.

Next, we would like to divide $x_{ij}(\mathcal{G}_l) = \sum_{k=1}^{N} x_{ij}^k$, such that the x_{ij}^k's are compatible with the physical topology $\mathcal{G}_{\text{ocs-mesh}}$. Unlike $\mathcal{G}_{\text{eps-group}}$ and $\mathcal{G}_{\text{eps-mesh}}$, we need to apply Lemma 4 multiple times to obtain a valid decomposition of $x_{ij}(\mathcal{G}_l)$.

Step 1: Decompose $x_{ij}(\mathcal{G}_l) = \sum_{c=1}^{C} x_{ij}^{(c)}$ such that the following constraints are met:

- $0 \le \sum_{j=1}^{N} x_{ij}^{(c)} \le 1, \forall i = 1, ..., N, c = 1, ..., C;$
- $0 \le \sum_{i=1}^{N} x_{ij}^{(c)} \le 1, \forall j = 1, ..., N, c = 1, ..., C.$

Step 2: For every $c_0 = 1, 2, ..., C$, decompose $x_{ij}^{(c_0)} = \sum_{w=1}^{W} x_{ij}^{(c_0-1)W+w}$ such that for any $w = 1, ..., W$, the following constraints are met:

- $0 \le \sum_{i \in \mathcal{A}_c} \sum_{j=1}^{N} x_{ij}^{(c_0-1)W+w} \le 1, \forall c = 1, ..., C;$
- $0 \le \sum_{i=1}^{N} \sum_{j \in \mathcal{A}_c} x_{ij}^{(c_0-1)W+w} \le 1, \forall c = 1, ..., C.$

Here $\mathcal{A}_c = \{(c-1)W + 1, (c-1)W + 2, ..., cW\}$.

Based on the above two steps, we obtain an integer decomposition $x_{ij}(\mathcal{G}_l) = \sum_{c=1}^{C} x_{ij}^{(c)} = \sum_{k=1}^{N} x_{ij}^k$ satisfying

(1) $0 \le \sum_{j=1}^{N} \sum_{k \in \mathcal{A}_c} x_{ij}^k \le 1, \forall i = 1, ..., N, c = 1, ..., C;$
(2) $0 \le \sum_{i=1}^{N} \sum_{k \in \mathcal{A}_c} x_{ij}^k \le 1, \forall j = 1, ..., N, c = 1, ..., C;$
(3) $0 \le \sum_{i \in \mathcal{A}_c} \sum_{j=1}^{N} x_{ij}^k \le 1, \forall c = 1, ..., C, k = 1, ..., N;$

(4) $0 \le \sum_{i=1}^{N}\sum_{j\in\mathcal{A}_c} x_{ij}^{k} \le 1, \forall c = 1,...,C, k = 1,...,N.$

With x_{ij}^k, the total number of logical links needed from S_i to S_j can be computed as $y_{ij} = \sum_{k=1}^{N} x_{kj}^i + \sum_{k=1}^{N} x_{ik}^j$. Next, we will show that y_{ij} is compatible with $\mathcal{G}_{\text{ocs-mesh}}$.

Recall that the OCS nodes in $\mathcal{G}_{\text{ocs-mesh}}$ are arranged as a 2D mesh. Consider the OCS node in the c_1-th row and c_2-th column. This OCS node connects the transmitters of the EPS node in group $\mathcal{A}_{c_1}$ to the receivers of the EPS node in group $\mathcal{A}_{c_2}$. Then, if the following constraints are satisfied

$$\sum_{j\in\mathcal{A}_{c_2}} y_{ij} \le 2, \sum_{i\in\mathcal{A}_{c_1}} y_{ij} \le 2, \tag{14}$$

the logical topology $y_{ij}, i \in \mathcal{A}_{c_1}, j \in \mathcal{A}_{c_2}$ will be realizable on this OCS. (14) can be verified below:

$$\sum_{j\in\mathcal{A}_{c_2}} y_{ij} = \sum_{j\in\mathcal{A}_{c_2}}\sum_{k=1}^{N} x_{kj}^i + \sum_{j\in\mathcal{A}_{c_2}}\sum_{k=1}^{N} x_{ik}^j \le 2;$$

$$\sum_{i\in\mathcal{A}_{c_1}} y_{ij} = \sum_{i\in\mathcal{A}_{c_1}}\sum_{k=1}^{N} x_{kj}^i + \sum_{i\in\mathcal{A}_{c_1}}\sum_{k=1}^{N} x_{ik}^j \le 2.$$

This completes the proof. □

Remark: Theorem 9 implies that $\beta(\mathcal{G}_{\text{ocs-mesh}}) = \min_f\{\mu_{\lfloor L/2\rfloor,B}^{\text{ideal}}(f)/\mu_{L,B}^{\text{ideal}}(f)\}$. The design of $\mathcal{G}_{\text{ocs-mesh}}$ shares a similar idea as the physical topology used by Sirius [2]. One critical difference between $\mathcal{G}_{\text{ocs-mesh}}$ and Sirius' physical topology is that we connect "two" links instead of one between each connected pair of EPS node and OCS node. This subtle design allows us to derive the performance guarantee in Theorem 9.

7 PERFORMANCE RATIO ANALYSIS

We have performed theoretical analysis for four physical topologies $\mathcal{G}_{\text{uniform}}$, $\mathcal{G}_{\text{eps-group}}$, $\mathcal{G}_{\text{eps-mesh}}$ and $\mathcal{G}_{\text{ocs-mesh}}$, and related their throughput metrics with that of the ideal physical topology. $\mathcal{G}_{\text{uniform}}$ is optimal but scales poorly. $\mathcal{G}_{\text{eps-group}}$, $\mathcal{G}_{\text{eps-mesh}}$ and $\mathcal{G}_{\text{ocs-mesh}}$ have better scalability but are sub-optimal. Accoring to Theorem 7, Theorem 8 and Theorem 9, It is easy to obtain

$$\begin{cases} \beta(\mathcal{G}_{\text{eps-group}}) = \min_f\{\beta(\mathcal{G}_{\text{eps-group}}, f)\}, \text{ where } \beta(\mathcal{G}_{\text{eps-group}}, f) = \mu_{\lceil L/3\rceil,B}^{\text{ideal}}(f)/\mu_{L,B}^{\text{ideal}}(f), \\ \beta(\mathcal{G}_{\text{eps-mesh}}) = \min_f\{\beta(\mathcal{G}_{\text{eps-mesh}}, f)\}, \text{ where } \beta(\mathcal{G}_{\text{eps-mesh}}, f) = \mu_{\lfloor L/3\rfloor,B}^{\text{ideal}}(f)/\mu_{L,B}^{\text{ideal}}(f), \\ \beta(\mathcal{G}_{\text{ocs-mesh}}) = \min_f\{\beta(\mathcal{G}_{\text{ocs-mesh}}, f)\}, \text{ where } \beta(\mathcal{G}_{\text{ocs-mesh}}, f) = \mu_{\lfloor L/2\rfloor,B}^{\text{ideal}}(f)/\mu_{L,B}^{\text{ideal}}(f). \end{cases} \tag{15}$$

In this section, we will calculate the approximated values of $\beta(\mathcal{G}_{\text{eps-group}})$, $\beta(\mathcal{G}_{\text{eps-mesh}})$ and $\beta(\mathcal{G}_{\text{ocs-mesh}})$. We focus on two cases below:

(1) ToE can generate multiple logical topologies to serve each traffic matrix. This setting was adopted by [2, 17–19].

(2) ToE generates only one logical topology (i.e., $M = 1$) for each traffic matrix. This setting was adopted by [5, 9, 12].

For the first case, we obtain some theoretical results in Section 7.1. The second case is hard to analyze, and we could only obtain some approximated numerical results in Section 7.2.

7.1 Calculating $\beta(\mathcal{G}_{\text{eps-group}})$, $\beta(\mathcal{G}_{\text{eps-mesh}})$ and $\beta(\mathcal{G}_{\text{ocs-mesh}})$ in Case 1

Definition 3. *A traffic pattern f is q-decomposable if there exists q permutation traffic matrices $Z_1, Z_2, ..., Z_q$ and $\phi_1 + \phi_2 + \cdots + \phi_q = 1, 0 \le \phi_i \le 1$, such that $f \le \sum_{i=1}^{q}\phi_i Z_i$.*

Definition 4. *A traffic pattern f is a normalized traffic pattern if the following condition is met*

$$\max\left\{\max_{i=1}^{N}\left\{\sum_{j=1}^{N} f_{ij}\right\}, \max_{j=1}^{N}\left\{\sum_{i=1}^{N} f_{ij}\right\}\right\} = 1.$$

Lemma 10. *For any normalized traffic pattern f, we must have $\mu_{L,B}^{ideal}(f) \leq BL$.*

Proof. Without loss of generality, we assume that the egress traffic from S_1 sums to 1, e.g., $\sum_{j=1}^{N} f_{1j} = 1$. Consider the network control policy that achieves the throughput value $\mu_{L,B}^{\text{ideal}}(f)$. Then, $\mu_{L,B}^{\text{ideal}}(f)\sum_{j=1}^{N} f_{1j} = \mu_{L,B}^{\text{ideal}}(f)$ amount of traffic can be delivered from the EPS node S_1 to other EPS nodes in a unit time period. On the other hand, the total egress bandwidth of S_1 is upper bounded by BL, because S_1 has L ports. Thus,

$$\mu_{L,B}^{\text{ideal}}(f) \leq BL.$$

□

Lemma 11. *Given a q-decomposable traffic pattern f that lasts Δ amount of time, if the ToE policy allows generating M OCS configurations, then*

$$\mu_{L,B}^{ideal}(f) \geq BL(1-\frac{q}{ML})(1-\frac{\delta M}{\Delta}).$$

Proof. We split Δ into M time slots, such that $\Delta^{(1)} = \Delta^{(2)} = \cdots = \Delta^{(M)} = \Delta/M - \delta$. In each time slot, L permutations can be formed because each EPS node has L bidirectional links connected the OCS. In total, we can form ML permutations.

Let $f \leq \sum_{i=1}^{q} \phi_i Z_i$ be the q-decomposition of f. For every permutation $Z_i, i = 1, 2, ..., q$, we configure OCSs such that Z_i is formed $\lceil (ML-q)\phi_i \rceil$ times. It is easy to verify that

$$\sum_{i=1}^{q} \lceil (ML-q)\phi_i \rceil \leq \sum_{i=1}^{q} ((ML-q)\phi_i + 1) = ML.$$

Hence, the above ToE strategy is feasible. Under this strategy, the bandwidth allocated to the ODC satisfies

$$\frac{1}{\Delta}\sum_{i=1}^{q} \lceil (ML-q)\phi_i \rceil B(\Delta/M-\delta) Z_i \geq \frac{(ML-q)B(\Delta/M-\delta)}{\Delta}\sum_{i=1}^{q}\phi_i Z_i \geq \frac{(ML-q)B(\Delta/M-\delta)}{\Delta} f.$$

Therefore, under the above ToE strategy and the one-hop routing, we can achieve a throughput value $\frac{(ML-q)B(\Delta/M-\delta)}{\Delta}$. Based on the definition of $\mu(\mathcal{G}, f)$, we must have

$$\mu_{L,B}^{\text{ideal}}(f) \geq \frac{(ML-q)B(\Delta/M-\delta)}{\Delta} = BL(1-\frac{q}{ML})(1-\frac{\delta M}{\Delta}).$$

□

Note that Lemma 11 holds for any M. According to the definitions of $\beta(\mathcal{G}_{\text{eps-group}})$, $\beta(\mathcal{G}_{\text{eps-mesh}})$ and $\beta(\mathcal{G}_{\text{ocs-mesh}})$ in (15), it is sufficient to consider normalized traffic patterns, because $\mu_{L,B}^{\text{ideal}}(f)$ increases linearly if we scales f linearly. Further, based on the Birkhoff and von Neumann theorem [6], any normalized traffic pattern f is q-decomposable with $q \leq N^2 - 2N + 2$. Combining all the above analysis with Lemma 10, we then obtain

Corollary 11.1. *Let $\gamma(M,l) = \left(1 - \frac{N^2-2N+2}{Ml}\right)\left(1-\frac{\delta M}{\Delta}\right)$, we have*

$$\frac{\lceil L/3 \rceil}{L}\max_{M}\{\gamma(M, \lceil L/3 \rceil)\} \leq \beta(\mathcal{G}_{eps\text{-}group}) \leq \frac{\lceil L/3 \rceil}{L}\left(\max_{M}\{\gamma(M,L)\}\right)^{-1},$$

$$\frac{\lfloor L/3\rfloor}{L}\max_{M}\{\gamma(M,\lfloor L/3\rfloor)\}\le\beta(\mathcal{G}_{\textit{eps-mesh}})\le\frac{\lfloor L/3\rfloor}{L}\left(\max_{M}\{\gamma(M,L)\}\right)^{-1},$$
$$\frac{\lfloor L/2\rfloor}{L}\max_{M}\{\gamma(M,\lfloor L/2\rfloor)\}\le\beta(\mathcal{G}_{\textit{ocs-mesh}})\le\frac{\lfloor L/2\rfloor}{L}\left(\max_{M}\{\gamma(M,L)\}\right)^{-1}.$$

When the OCS reconfiguration delay $\delta = 0$, $\gamma(M, l)$ can be arbitrarily close to 1 as we increase the number of OCS configurations M. In this case, we have

Corollary 11.2. *If $\delta = 0$, then $\beta(\mathcal{G}_{\textit{eps-group}}) = \frac{\lceil L/3\rceil}{L}$, $\beta(\mathcal{G}_{\textit{eps-mesh}}) = \frac{\lfloor L/3\rfloor}{L}$, $\beta(\mathcal{G}_{\textit{ocs-mesh}}) = \frac{\lfloor L/2\rfloor}{L}$.*

7.2 Calculating $\beta(\mathcal{G}_{\textbf{eps-group}})$, $\beta(\mathcal{G}_{\textbf{eps-mesh}})$ and $\beta(\mathcal{G}_{\textbf{ocs-mesh}})$ in Case 2 Where $M = 1$

When $M = 1$, the lower bound of $\mu^{\text{ideal}}_{L,B}(f)$ given by Lemma 11 can be extremely loose, and thus may not be useful for calculating $\beta(\mathcal{G}_{\text{eps-group}})$, $\beta(\mathcal{G}_{\text{eps-mesh}})$ and $\beta(\mathcal{G}_{\text{ocs-mesh}})$. In this case, we perform numerical analysis instead. Note that exhaustively enumerating all the different traffic matrices is not feasible. We simply generate random traffic matrices to perform the calculation. We use two approaches to generate random traffic matrices:

Approach 1: Randomly generate Q permutation traffic matrices $f^{(1)}, f^{(2)}, ..., f^{(Q)}$ and Q parameters $0 \le \gamma_1, \gamma_2, ..., \gamma_Q \le 1$ satisfying $\sum_{q=1}^{Q}\gamma_q = 1$. Let $f = \sum_{q=1}^{Q}\gamma_q f^{(q)}$. It is easy to verify that the traffic matrix f generated with this approach is a normalized traffic pattern.

Approach 2: Generate a random non-negative number for each entry $f_{ij}, i \neq j$, and then normalize f such that $\max\{\max_{j=1}^{N}\{\sum_{i=1}^{N} f_{ij}\}, \max_{i=1}^{N}\{\sum_{j=1}^{N} f_{ij}\}\} = 1$.

For every N and every L, we generate a number of normalized random traffic patterns using the above two approaches, and calculate throughput values $\mu^{\text{ideal}}_{\lceil L/3\rceil,B}(f)$, $\mu^{\text{ideal}}_{\lfloor L/3\rfloor,B}(f)$, $\mu^{\text{ideal}}_{\lfloor L/2\rfloor,B}(f)$, $\mu^{\text{ideal}}_{L,B}(f)$ for every traffic pattern f. We use three approaches to calculate the throughput values:

Joint Optimization: Directly solve the optimization problem (5) using Gurobi [20].

Decoupled Optimization: First, compute a logical topology $\bar{\mathcal{G}}_l^{(1)}$ that best approximates the traffic pattern f by solving the following optimization problem:

$$\boxed{\begin{aligned}&\max\sum_{i=1}^{N}\sum_{j=1}^{N}\left(Lf_{ij}-\lfloor Lf_{ij}\rfloor\right)x_{ij}(\bar{\mathcal{G}}_l^{(1)})\\&\text{s.t.}\lfloor Lf_{ij}\rfloor\le x_{ij}(\bar{\mathcal{G}}_l^{(1)})\le\lceil Lf_{ij}\rceil,\ \text{for all } i,j=1,2,...,N,\\&\quad\sum_{j=1}^{N}x_{ij}(\bar{\mathcal{G}}_l^{(1)})\le L,\ \text{for all } i=1,2,...,N,\\&\quad\sum_{i=1}^{N}x_{ij}(\bar{\mathcal{G}}_l^{(1)})\le L,\ \text{for all } j=1,2,...,N.\end{aligned}}\tag{16}$$

Then, solve the optimization problem (5) with the logical topology $\mathcal{G}_l^{(1)}$ being fixed as $\bar{\mathcal{G}}_l^{(1)}$.

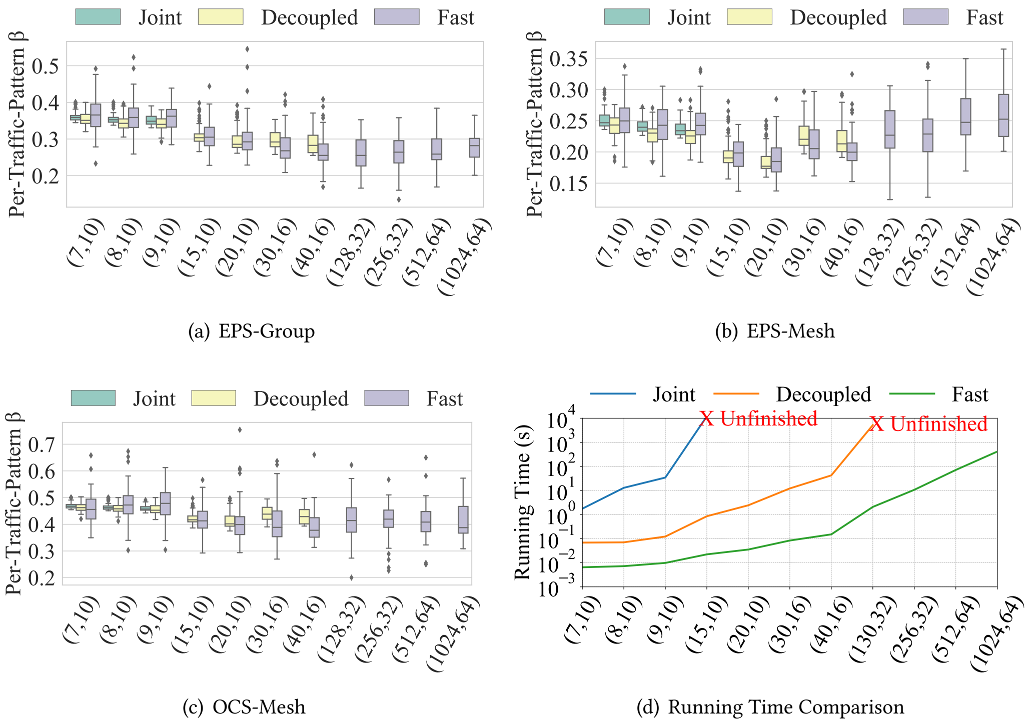
Fast Approximation: First, compute a logical topology $\bar{\mathcal{G}}_l^{(1)}$ by solving (16). Second, find all the shortest paths for every EPS node pair in $\bar{\mathcal{G}}_l^{(1)}$. Third, calculate the throughput value based on Equal-Cost Multi-Path (ECMP) routing.

The Joint Optimization approach solves topology and routing jointly, and gives the exact throughput values. However, this approach has the highest computational complexity. The Decoupled Optimization approach reduces algorithmic complexity by solving topology and routing separately. The Fast Approximation approach further reduces computational complexity by using a fixed

ECMP routing. Notably, the second and the third approaches can only compute approximated throughput and performance ratio values.

7.2.1 *Approximating $\beta(\mathcal{G}_{eps\text{-}group})$, $\beta(\mathcal{G}_{eps\text{-}mesh})$ and $\beta(\mathcal{G}_{ocs\text{-}mesh})$*.

We generate a list of (N, L) pairs: $(7, 10)$, $(8, 10)$, $(9, 10)$, $(15, 10)$, $(20, 10)$, $(30, 16)$, $(40, 16)$, $(128, 32)$, $(256, 32)$, $(512, 64)$ and $(1024, 64)$. For each (N, L) pair, we generate 60 normalized traffic patterns, and denote this set of traffic patterns by $\mathcal{F}_{N,L}$. For every traffic pattern $f \in \mathcal{F}_{N,L}$, we use the above three approaches to compute the per-traffic-pattern performance ratio values $\beta(\mathcal{G}_{\text{eps-group}}, f)$, $\beta(\mathcal{G}_{\text{eps-mesh}}, f)$ and $\beta(\mathcal{G}_{\text{ocs-mesh}}, f)$ as defined in (15).



(a) EPS-Group

(b) EPS-Mesh

(c) OCS-Mesh

(d) Running Time Comparison

Fig. 6. Calculating per-traffic-pattern performance ratios using Joint Optimization, Decoupled Optimization and Fast Approximation.

Since the Joint Optimization approach can compute the exact per-traffic-pattern performance ratio values, we consider $\beta_{\text{joint}}(\mathcal{G}) = \min_{f \in \mathcal{F}_{N,L}} \beta_{\text{joint}}(\mathcal{G}, f)$ as the best approximation to $\beta(\mathcal{G})$ for every $\mathcal{G} \in \{\mathcal{G}_{\text{eps-group}}, \mathcal{G}_{\text{eps-mesh}}, \mathcal{G}_{\text{ocs-mesh}}\}$. However, calculating $\beta_{\text{joint}}(\mathcal{G}, f)$ is computationally expensive. The Joint Optimization approach can only compute $\beta_{\text{joint}}(\mathcal{G})$ for the (N, L) pairs $(7, 10)$, $(8, 10)$ and $(9, 10)$. When $N > 9$, the Joint Optimization approach may not be able to compute a solution even after running a few hours. Hence, in the following, we study how to approximate $\beta_{\text{joint}}(\mathcal{G})$ using the other two approaches.

We first use the Decoupled Optimization approach to approximate $\beta_{\text{joint}}(\mathcal{G})$. For every $f \in \mathcal{F}_{N,L}$, we compute $\beta_{\text{decoupled}}(\mathcal{G}, f)$ using the Decoupled Optimization approach. Let $\alpha_{\text{decoupled}}(\mathcal{G}, p)$ be the p-th percentile value among all the values of $\beta_{\text{decoupled}}(\mathcal{G}, f)$. We aim to find p such that $\alpha_{\text{decoupled}}(\mathcal{G}, p) \approx \beta_{\text{joint}}(\mathcal{G})$. From Fig. 6, we can see that $p(\mathcal{G}_{\text{eps-group}}) \approx$ 25th, $p(\mathcal{G}_{\text{eps-mesh}}) \approx$ 35th

Table 2. The values of $p'(\mathcal{G})$ for different (N, L)'s. When $N \leq 9$, we pick $p'(\mathcal{G})$ such that $\alpha_{\text{fast}}(\mathcal{G}, p'(\mathcal{G})) \approx \beta_{\text{joint}}(\mathcal{G})$. When $9 < N \leq 40$, we pick $p'(\mathcal{G})$ such that $\alpha_{\text{fast}}(\mathcal{G}, p'(\mathcal{G})) \approx \alpha_{\text{decoupled}}(\mathcal{G}, p(\mathcal{G}))$.

	(7,10)	(8,10)	(9,10)	(15,10)	(20,10)	(30,16)	(40,16)
$p'(\mathcal{G}_{\text{eps-group}})$	36th	26th	23th	35th	33th	58th	56th
$p'(\mathcal{G}_{\text{eps-mesh}})$	28th	31th	20th	36th	36th	56th	61th
$p'(\mathcal{G}_{\text{ocs-mesh}})$	48th	31th	26th	46th	45th	65th	63th

Table 3. Approximating $\beta(\mathcal{G}_{\text{eps-group}})$, $\beta(\mathcal{G}_{\text{eps-mesh}})$ and $\beta(\mathcal{G}_{\text{ocs-mesh}})$ using the Fast Approximation approach.

	(7,10)	(8,10)	(9,10)	(15,10)	(20,10)	(30,16)	(40,16)	(128,32)	(256,32)	(512,64)	(1024,64)
$\beta(\mathcal{G}_{\text{eps-group}})$	0.34	0.34	0.33	[0.28,0.31]	[0.27,0.3]	[0.25,0.28]	[0.24,0.26]	[0.22,0.27]	[0.23,0.27]	[0.24,0.28]	[0.25,0.29]
$\beta(\mathcal{G}_{\text{eps-mesh}})$	0.24	0.23	0.22	[0.17,0.21]	[0.16,0.19]	[0.18,0.22]	[0.18,0.2]	[0.2,0.25]	[0.19,0.24]	[0.22,0.26]	[0.21,0.28]
$\beta(\mathcal{G}_{\text{ocs-mesh}})$	0.45	0.45	0.44	[0.39,0.43]	[0.36,0.42]	[0.36,0.42]	[0.35,0.4]	[0.37,0.45]	[0.39,0.44]	[0.38,0.43]	[0.37,0.45]

and $p(\mathcal{G}_{\text{ocs-mesh}}) \approx$ 25th when $(N, L) = (7, 10), (8, 10)$ or $(9, 10)$. Then, we could use $\alpha_{\text{decoupled}}(\mathcal{G}, p(\mathcal{G}))$ to approximate $\beta_{\text{joint}}(\mathcal{G})$ for larger data centers. However, even this Decoupled Optimization approach cannot scale beyond 40 EPS nodes. When $N > 40$, the Decoupled Optimization approach may not produce any solution after running a few hours.

We then use the Fast Approximation approach to approximate $\beta_{\text{joint}}(\mathcal{G})$. Similarly, let $\alpha_{\text{fast}}(\mathcal{G}, p')$ be the p'-th percentile value among all the values of $\beta_{\text{fast}}(\mathcal{G}, f)$. We aim to find p' such that $\alpha_{\text{fast}}(\mathcal{G}, p') \approx \beta_{\text{joint}}(\mathcal{G})$ when $N \leq 9$ and $\alpha_{\text{fast}}(\mathcal{G}, p') \approx \alpha_{\text{decoupled}}(\mathcal{G}, p(\mathcal{G}))$ when $9 < N \leq 40$. The values of $p'(\mathcal{G})$ vary for different physical topologies and different (N, L) pairs. The detailed values are summarized in Table 2. For every $\mathcal{G} \in \{\mathcal{G}_{\text{eps-group}}, \mathcal{G}_{\text{eps-mesh}}, \mathcal{G}_{\text{ocs-mesh}}\}$, we pick the minimum and the maximum values of $p'(\mathcal{G})$, and denote them by $p'_{\min}(\mathcal{G})$ and $p'_{\max}(\mathcal{G})$. Then, we can generate a range $[\alpha_{\text{fast}}(\mathcal{G}, p'_{\min}(\mathcal{G})), \alpha_{\text{fast}}(\mathcal{G}, p'_{\max}(\mathcal{G}))]$ to approximate $\beta_{\text{joint}}(\mathcal{G}) \approx \beta(\mathcal{G})$. We summarize the approximation results for all $N > 9$ in Table 3.

Note: The numerical analysis in this section is just our best-effort estimate for $\beta(\mathcal{G}_{\text{eps-group}})$, $\beta(\mathcal{G}_{\text{eps-mesh}})$ and $\beta(\mathcal{G}_{\text{ocs-mesh}})$. There is **no rigorous guarantee** that the true values of $\beta(\mathcal{G})$ are indeed within the ranges provided by Table 3. Nevertheless, we believe that our numerical analysis gives readers a sense on the magnitude of $\beta(\mathcal{G})$.

8 CONCLUSION

We study physical topology design for ODCs, and offer a novel methodology to analyze the performance guarantee for different physical topologies in this paper. Based on our methodology, we prove that the uniform bipartite graph based physical topology is optimal for small-scale data centers with $N \leq R$, while optimal physical topologies do not exist when $N > R$. We also design three physical topologies that support larger-scale ODCs, and prove their performance gaps with respect to the ideal physical topology. The methodology in this paper provides a theoretical foundation for ODC physical topology design, and may help network operators design physical topologies with a better guarantee.

REFERENCES

[1] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. 2008. A scalable, commodity data center network architecture. In *SIGCOMM*.

[2] Hitesh Ballani, Paolo Costa, Raphael Behrendt, Daniel Cletheroe, Istvan Haller, Krzysztof Jozwik, Fotini Karinou, Sophie Lange, Kai Shi, Benn Thomsen, and Hugh Williams. 2020. Sirius: A Flat Datacenter Network with Nanosecond Optical Switching. In *SIGCOMM*.

[3] Inc. CALIENT Technologies. [n.d.]. https://www.calient.net/.

[4] Inc. CALIENT Technologies. [n.d.]. S Series Optical Circuit Switch. https://www.calient.net/products/s-series-photonic-switch/.

[5] Peirui Cao, Shizhen Zhao, Min Yee Teh, Yunzhuo Liu, and Xinbing Wang. 2021. TROD: Evolving From Electrical Data Center to Optical Data Center. In *ICNP*.

[6] Cheng-shang Chang, Wen-Jyh Chen, and Hsiang-Yi Huang. 2000. Birkhoff-von Neumann Input-Buffered Crossbar Switches. In *INFOCOM*.

[7] Kai Chen, Ankit Singla, Atul Singh, Kishore Ramachandran, Lei Xu, Yueping Zhang, Xitao Wen, and Yan Chen. 2012. OSA: An optical switching architecture for data center networks with unprecedented flexibility. In *NSDI*.

[8] Li Chen, Kai Chen, Zhonghua Zhu, Minlan Yu, George Porter, Chunming Qiao, and Shan Zhong. 2017. Enabling wide-spread communications on optical fabric with megaswitch. In *NSDI*.

[9] Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Hajabdolali Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, and Amin Vahdat. 2010. Helios: a hybrid electrical/optical switch architecture for modular data centers. In *SIGCOMM*.

[10] Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Nikhil Devanur, Janardhan Kulkarni, Gireeja Ranade, Pierre Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel Kilper. 2016. ProjecToR: Agile Reconfigurable Data Center Interconnect. In *SIGCOMM*.

[11] Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, and Sudipta Sengupta. 2009. VL2: A Scalable and Flexible Data Center Network. In *SIGCOMM*.

[12] Navid Hamedazimi, Zafar Qazi, Himanshu Gupta, Vyas Sekar, Samir R. Das, Jon P. Longtin, Himanshu Shah, and Ashish Tanwer. 2014. FireFly: A Reconfigurable Wireless Data Center Fabric Using Free-Space Optics. In *SIGCOMM*.

[13] Co. Huawei Technologies. [n.d.]. https://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/ce8800.

[14] John Kim, William J. Dally, and Dennis Abts. 2007. Flattened Butterfly : A Cost-Efficient Topology for High-Radix Networks. In *ISCA*.

[15] Gengchen Liu, Roberto Proietti, Marjan Fariborz, Pouya Fotouhi, Xian Xiao, and S.J. Ben Yoo. 2020. Architecture and Performance Studies of 3D-Hyper-FleX-LION for Reconfigurable All-to-All HPC Networks. In *SC*.

[16] He Liu, Feng Lu, Alex Forencich, Rishi Kapoor, Malveeka Tewari, Geoffrey M Voelker, George Papen, Alex C Snoeren, and George Porter. 2014. Circuit Switching Under the Radar with REACToR. In *NSDI*.

[17] He Liu, Matthew K. Mukerjee, Conglong Li, Nicolas Feltman, George Papen, Stefan Savage, Srinivasan Seshan, Geoffrey M. Voelker, David G. Andersen, Michael Kaminsky, George Porter, and Alex C. Snoeren. 2015. Scheduling Techniques for Hybrid Circuit/Packet Networks. In *CoNEXT*.

[18] William M Mellette, Rajdeep Das, Yibo Guo, Rob McGuinness, Alex C Snoeren, and George Porter. 2020. Expanding across time to deliver bandwidth efficiency and low latency. In *NSDI*.

[19] William M. Mellette, Rob Mcguinness, Arjun Roy, Alex Forencich, and George Porter. 2017. RotorNet: A Scalable, Low-complexity, Optical Datacenter Network. In *SIGCOMM*.

[20] Gurobi Optimization. [n.d.]. https://www.gurobi.com/.

[21] George Porter, Richard Strong, Nathan Farrington, Alex Forencich, Pang Chen-Sun, Tajana Rosing, Yeshaiahu Fainman, George Papen, and Amin Vahdat. 2013. Integrating microsecond circuit switching into the data center. In *SIGCOMM*.

[22] Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, et al. 2015. Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network. In *SIGCOMM*.

[23] Guohui Wang, David G Andersen, Michael Kaminsky, Konstantina Papagiannaki, TS Ng, Michael Kozuch, and Michael Ryan. 2010. c-Through: Part-time optics in data centers. In *SIGCOMM*.

[24] Ke Wen, Payman Samadi, Sebastien Rumley, Christine P. Chen, Yiwen Shen, Meisam Bahadori, Keren Bergman, and Jeremiah Wilke. 2016. Flexfly: Enabling a Reconfigurable Dragonfly through Silicon Photonics. In *SC*.

[25] Shizhen Zhao, Rui Wang, Junlan Zhou, Joon Ong, Jeffery C. Mogul, and Amin Vahdat. 2019. Minimal Rewiring: Efficient Live Expansion for Clos Data Center Networks: Extended Version. In *NSDI https://ai.google/research/pubs/pub47492*.