Expanding Optical-Circuit-Switching Multi-Stage Networks to Ensure Admissible Blocking Probability in Data Centers

Eiji Oki[1], Ryotaro Taniguchi[1], Kazuya Anazawa[2,1], and Takeru Inoue[2]
[1]Kyoto University, Kyoto, Japan
[2]NTT Network Innovation Laboratories, NTT Corporation, Japan

***Abstract*—Future data centers are expected to integrate cutting-edge circuit switching technologies, particularly optical switching, providing superior transmission capacity and energy efficiency. A Clos network represents a multi-stage switching architecture featuring hierarchically connected switches. Its design is favored in data centers because it scales efficiently. Balancing the switching network size with the quality of an acceptable connection request presents a trade-off. It is crucial to tackle design challenges that aim to increase the switching network size while maintaining a specified admissible blocking probability. This paper studies the models for designing three-stage folded-type Clos networks, expanded from two-stage ones, with a blocking probability guarantee to maximize the switching network size. It considers both types of expanding: the input and output layer and the intermediate layer. We compare the performances of different models. Expanding the input-output layer effectively increases the switching network size, whereas expanding the intermediate layer does not.**

***Index Terms*—Clos network, optical circuit switching, data center, switching network size, blocking probability**

I. Introduction

A network in a data center, composed of switches and routers, is responsible for managing large-scale data processing. Future data centers are expected to integrate cutting-edge circuit switching technologies, particularly optical switching, recognized for their superior transmission capacity and energy efficiency [1]–[4]. Optical circuit switching technology provides consistent communication quality by dedicating a connection solely for data transfer and maintaining it throughout transmission. In data centers, a Clos network, a multi-stage switching architecture with hierarchically interconnected switches, is often employed [5]–[10].

Linking a transmitter and receiver pair to a common switch in a data center is advantageous [7]. A folded Clos network (F-Clos) is often utilized to achieve this. F-Clos typically has two layers: input-output and intermediate layers. In this setup, the transmitter and receiver pair are connected to the same switch within the input-output layer, the origin of the term *folded*.

Mano *et al.* [11], [12] introduced the twisted-folded Clos network (TF-Clos), which alleviates certain switch-port limitations by incorporating a *twisting* concept in the connection links between input-output layer switches and intermediate switches. The authors developed a TF-Clos design model using the limited usable number of identical $N \times N$ switches under the strict-sense non-blocking (SNB) condition to maximize the switching network size, i.e., the number of terminals connected to the network. They considered two-stage TF-Clos (2TF).

The SNB condition in the 2TF design prevents blocking but can still restrict the available switching network size. A network architecture that permits a certain degree of blocking could allow for more design flexibility. Taka *et al.* [13]–[15] explored how to design TF-Clos with a guaranteed admissible blocking probability to enhance the switching network size. The admissible blocking probability is the probability that a connection request from a terminal connected to an input port is blocked due to internal network congestion. Balancing the switching network size with the quality of an acceptable connection request presents a trade-off. Consequently, it is crucial to tackle design challenges that expand the switching network size while maintaining a specified admissible blocking probability [16].

Jajszcyk [17] addressed expanding two-stage F-Clos (2F) to three-stage F-Clos (3F) to increase the switching network size by mainly introducing two types of expanding: input and output layer and intermediate layer. The work assumed that the size of switches used for 3F could be flexibly customized and proved that both expansion types are isotopic. On the other hand, as assumed in [11], [12], considering a limited number of identical switches with the $N \times N$ size is more practical when we design a switching network in data centers.

Along this direction, Taniguchi *et al.* [18] developed a design model for a three-stage TF-Clos structure (3TF) with an admissible blocking probability guarantee to increase the switching network size beyond that of the 2TF design using a limited number of identical switches with the $N \times N$ size. They expanded the input and output layer for 3TF, an expansion type of the two [17], increasing the switching network size.

This paper presents the models for designing 3TF with a blocking probability guarantee to maximize the switching network size, considering both types of expansion: the input and output layer (type α) and the intermediate

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

layer (type β). We compare the performances of different models. Numerical results show that expanding the input-output layer (type α) effectively increases the switching network size, whereas expanding the intermediate layer (type β) does not. In this paper, we call 3TF with type α 3TF-α and 3TF with type β 3TF-β.

The rest of the paper is organized as follows. Section III describes the assumptions used in the design models. Section II describes the structure of 2TF. Sections IV and V present the design models for 3TF-α and 3TF-β, respectively. Section VI describes numerical results. Section VII concludes this paper.

II. Two-stage twisted-folded Clos network (2TF)

2TF comprises $k+m$ switches, each equipped with N input ports and N output ports, as shown in Fig. 1 [11], [12]. The input-output layer has k switches, labeled as S_i^1, where i ranges from 1 to k. Each switch in this layer connects to n transmitters, denoted by t_j^i with $i \in [1,k]$ and $j \in [1,n]$, at one side as well as n receivers, denoted by r_j^i with $i \in [1,k]$ and $j \in [1,n]$, at the other side. In the intermediate layer, there are m switches labeled as S_i^2, with $i \in [1,m]$. One/the other side of each switch in the input-output layer is connected to one/the other side of each switch in the intermediate layer via v links for incoming/outgoing connections.

Designing 2TF specifies the values of n, k, m, and v to maximize the switching network size of nk, given a total of a available $N \times N$ switches, where $k+m \leq a$ [11]–[15]. Mano *et al.* [11], [12] presented a 2TF design model that satisfies the SNB condition. Taka *et al.* [13]–[15] introduced alternative 2TF models that ensure an admissible blocking probability, aiming to expand the switching network size beyond what is achievable with the SNB condition.

III. Assumptions

We describe the assumptions used in this paper as follows. We design a Clos network for a data center using identical switches whose size is $N \times N$. The number of switches is at most a. Terminals, such as top-of-rack (ToR) or aggregation switches, in a data center network, can establish direct connections with optical circuit switching. A transmitter linked to a terminal generates a request, which is active with a probability of p and inactive with a probability of $1-p$. The generation of these requests follows an independent and identically distributed (i.i.d.) model. The destination of each request is not predetermined and can be arbitrary. The parameter ϵ represents the admissible blocking probability.

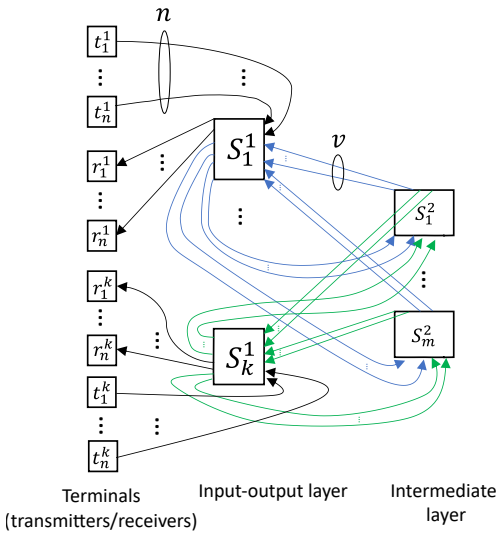

Fig. 1. Structure of 2TF.

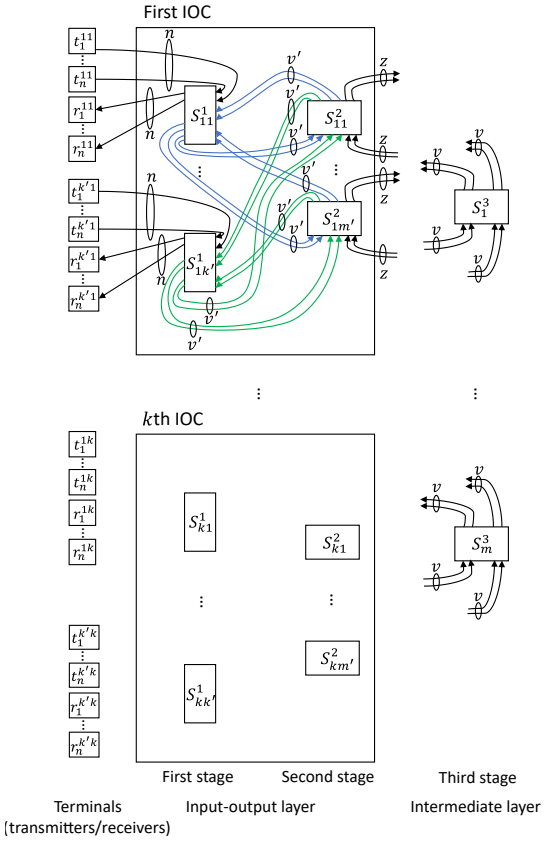

Fig. 2. Structure of 3TF with expanded input-output layer (3TF-α).

IV. Design for 3TF with expanding input and output layer (3TF-α)

A. Structure of 3TF-α

Fig. 2 shows the 3TF-α structure consisting of input-output and intermediate layers [18]. 3TF-α is composed

of $k(k'+m')+m$ switches, each equipped with N input and N output ports. The input-output layer encompasses the first and second stages, while the intermediate layer constitutes the third stage. The input-output layer includes k switching-network components, each with $k'+m'$ switches, and the intermediate layer contains m switches. We refer to a switching network component in the input-output layer as an IOC. The IOCs and third-stage switches are interconnected following the TF-Clos configuration, as are the switches in the first and second stages. $nk'k$ transmitters and $nk'k$ receivers are connected across all first-stage switches. The total switching network size for 3TF-α is given by $nk'k$. $S^1_{i_1 i_2}$ refers to the i_2th first-stage switch in the i_1th IOC, where $i_1 \in [1,k]$ and $i_2 \in [1,k']$. Similarly, $S^2_{i_1 i_2}$ refers to the i_2th second-stage switch in the i_1th IOC, where $i_1 \in [1,k]$ and $i_2 \in [1,m']$. The ith third-stage switch is denoted as S^3_i, where $i \in [1,m]$. Transmitters and receivers are denoted by $t^{i_1 i_2}_j$ and $r^{i_1 i_2}_j$, respectively, with $j \in [1,n]$, $i_1 \in [1,k']$, and $i_2 \in [1,k]$. The switch $S^1_{i_1 i_2}$ connects n transmitters and n receivers. Each $S^1_{i_1 i_2}$ is linked to $S^2_{i_1 i'_2}$ via v' outgoing and v' incoming links, where $i_1 \in [1,k]$, $i_2 \in [1,k']$, and $i'_2 \in [1,m']$. Additionally, $S^2_{i_1 i_2}$ is connected to S^3_i through v outgoing and v incoming links, where $i_1 \in [1,k]$, $i_2 \in [1,m']$, and $i \in [1,m]$. The switch $S^2_{i_1 i_2}$ supports up to z output ports to the third-stage switches and up to z input ports from the third-stage switches. The total number of output ports from each IOC to the third-stage switches is capped at zm', and the same applies to the input ports from the third-stage switches to each IOC.

B. Formulation of 3TF-α design model

The 3TF-α design problem with blocking probability guarantee is given by:

$$\max \quad nk'k \tag{1a}$$

$$\text{s.}t. \quad n + v'm' \le N \tag{1b}$$

$$v'k' + z \le N \tag{1c}$$

$$vk \le N \tag{1d}$$

$$vm \le zm' \tag{1e}$$

$$\sum_{w=n_1^{\text{snb}}+1}^{n} \binom{n}{w} p^w (1-p)^{n-w} \le \eta \tag{1f}$$

$$2\left\lfloor \frac{n_1^{\text{snb}}-1}{v'} \right\rfloor + 1 \le m' \tag{1g}$$

$$1-(1-\eta) \times \left(1 - \sum_{w=n_2^{\text{snb}}+1}^{nk'} \binom{nk'}{w} p^w (1-p)^{nk'-w}\right) \le \epsilon \tag{1h}$$

$$2\left\lfloor \frac{n_2^{\text{snb}}-1}{v} \right\rfloor + 1 \le m \tag{1i}$$

$$k(k'+m') + m \le a \tag{1j}$$

$$\eta \le \epsilon \le 1 \tag{1k}$$

$$n, n_1^{\text{snb}}, n_2^{\text{snb}}, k', k, v', v, m', m, z \in \mathbb{N} \tag{1l}$$

$$\eta \in \mathbb{R}^+, \tag{1m}$$

where $\mathbb{N}$ denotes a set of natural numbers, and $\mathbb{R}^+$ denotes a set of non-negative real numbers. Equation (1a) serves as the objective function to maximize the switching network size of 3TF-α, which corresponds to maximizing the number of terminals connected to the input and output ports of the switching network. Equations (1b)–(1d) define the constraints on the number of available ports for switches in the first, second, and third stages, respectively. Equation (1e) ensures that the total number of output (or input) ports from (or to) an IOC to (or from) third-stage switches is at least vm. Equation (1f) demonstrates that the blocking probability in the first stage is bounded by η. Equation (1g) specifies the SNB condition in this stage. Similarly, Equation (1h) shows that the blocking probability in the third stage is restricted by ϵ, with Equation (1i) detailing the corresponding the SNB condition. Equation (1j) ensures that the number of switches in the network does not exceed a. Equation (1k) defines the relationship between η and ϵ, ensuring that η is at most ϵ. Finally, Equations (1l) and (1m) define the types of decision variables.

V. Design for 3TF with expanding intermediate layer (3TF-β)

A. Structure of 3TF-β

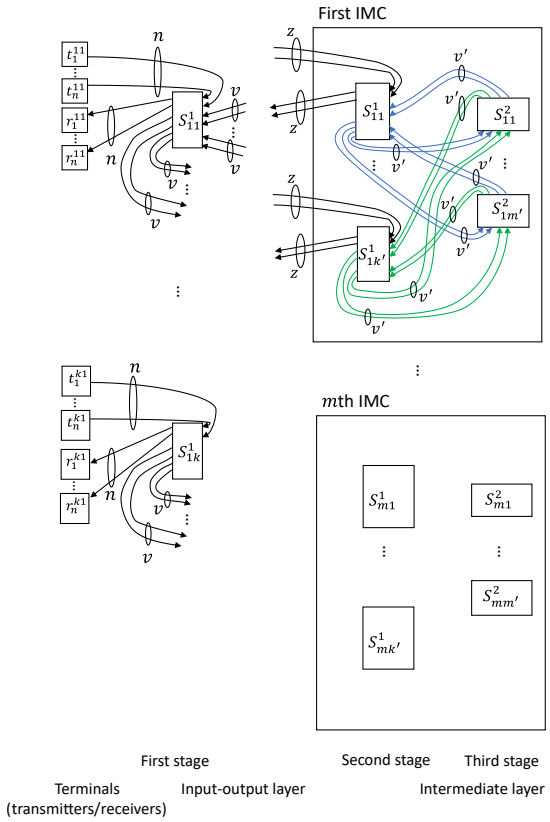

Fig. 3. Structure of 3TF with expanded intermediate layer (3TF-β.

Fig. 3 shows the 3TF-β structure consisting of input-output and intermediate layers. 3TF-β comprises $k + m(k' + m')$ switches, each of which has N input ports and N output ports. The input-output layer comprises k switches, and the intermediate layer comprises m switching-network components, each of which has $k' + m'$ switches. We call such a switching network in the intermediate layer an intermediate-layer switching network component (IMC). The first-stage switches and IMCs are connected in the way of TF-Clos, as are the second- and third-stage switches. nk transmitters and nk receivers are connected to all the first-stage switches. S_i^1 denotes the ith first-stage switch, where $i \in [1, k]$. $S_{i_1 i_2}^2$ denotes the i_2th second-stage switch of i_1th IMC, where $i_1 \in [1, m]$ and $i_2 \in [1, k']$. $S_{i_1 i_2}^3$ denotes the i_2th third-stage switch of i_1th IMC, where $i_1 \in [1, m]$ and $i_2 \in [1, m']$. t_j^i and r_j^i denote the jth transmitter and the jth receiver, respectively, where $j \in [1, n]$ and $i \in [1, k]$; S_i^1 connects n transmitters and n receivers. S_i^1 is connected to $S_{i_1 i_2}^2$ with v outgoing links and v incoming links, where $i \in [1, k]$, $i_1 \in [1, m]$, and $i_2 \in [1, k']$. $S_{i_1 i_2}^2$ is connected to $S_{i_1 i_2'}^3$ with v' outgoing links and v' incoming links, where $i_1 \in [1, m]$, $i_2 \in [1, k']$, and $i_2' \in [1, m']$. $S_{i_1 i_2}^2$ has at most z input ports incoming from first-stage switches, and at most z output ports outgoing to first-stage switches. The total number of input ports incoming from first-stage switches to each IMC is at most zk', and the total number of output ports from each IMC to first-stage switches is at most zk'.

B. Formulation of 3TF-β design model

The 3TF-β design problem with blocking probability guarantee is given by:

$$\max \quad nk \tag{2a}$$

$$\text{s.}t. \quad n + vm \le N \tag{2b}$$

$$v'm' + z \le N \tag{2c}$$

$$v'k' \le N \tag{2d}$$

$$vk \le zk' \tag{2e}$$

$$\sum_{w=n_1^{\text{snb}}+1}^{n} \binom{n}{w} p^w (1-p)^{n-w} \le \eta \tag{2f}$$

$$2\left\lfloor \frac{n_1^{\text{snb}}-1}{v} \right\rfloor + 1 \le m \tag{2g}$$

$$1 - (1-\eta) \times \left(1 - \sum_{w=n_2^{\text{snb}}+1}^{zk'} \binom{zk'}{w} p^w (1-p)^{zk'-w}\right) \le \epsilon \tag{2h}$$

$$2\left\lfloor \frac{n_2^{\text{snb}}-1}{v'} \right\rfloor + 1 \le m' \tag{2i}$$

$$k + m(k' + m') \le a \tag{2j}$$

$$\eta \le \epsilon \le 1 \tag{2k}$$

$$n, n_1^{\text{snb}}, n_2^{\text{snb}}, k', k, v', v, m', m, z \in \mathbb{N} \tag{2l}$$

$$\eta \in \mathbb{R}^+. \tag{2m}$$

Equation (2a) is the objective function to maximize the switching network size of 3TF-β, i.e., to maximize the number of terminals connected to the switching network's input and output ports. Equations (2b)–(2d) represent the constraints of the number of available ports of switches in the first, second, and third stages, respectively. Equation (2e) indicates that the total number of output (input) ports outgoing from (incoming to) an IMC to (from) third-stage switches is at least vm, respectively. Equation (2f) indicates that the blocking probability in the first stage is at most η. Equation (2g) expresses the SNB condition in the first stage. Equation (2h) represents that the blocking probability in the third stage is lower than or equal to ϵ. Equation (2i) expresses the SNB condition in the third stage. Equation (2j) expresses that the number of switches used in the network is at most a. Equation (2k) expresses the relationship between η and ϵ. Equations (2l) and (2m) describe the types of decision variables.

VI. Numerical results

Fig. 4 compares the switching network sizes among 3TF-α, 3TF-β, 2TF, and 2F depending on a with $N = 10$ under the SNB condition. The switching network size of 3TF-α increases with a up to $a = 150$, whereas those of 3TF-β and 2TF stop increasing at smaller a. The switching network size of 3TF-α is larger than those of 3TF-β and 2TF with a sufficiently large value of a. Moreover, the switching network size of 3TF-β is smaller than that of 2TF. The bottleneck of switch ports causes this limitation of 3TF-β due to its structure that expands the intermediate layer. Expanding the input-output layer effectively increases the switching network size, whereas expanding the intermediate layer does not.

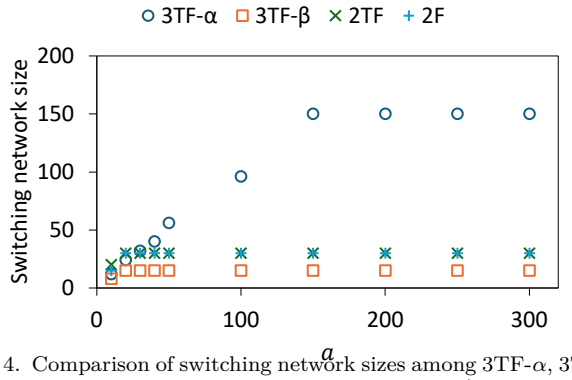

Fig. 4. Comparison of switching network sizes among 3TF-α, 3TF-β, 2TF, and 2F dependent on a with $N = 10$ and $\epsilon = 0$ (SNB condition).

We investigate the switching network sizes of 3TF-α, 2TF, and 2F dependent on a under the SNB condition, as shown in Fig. 5. We set $N = 40$. The larger a becomes, the greater the difference in the switching network size between two and three stages. Increasing the number of stages from two to three increases the value of a that peaks out. 3TF-α has a higher value of a that peaks out than 2TF and 2F. The switching network size of 2TF is higher than that of 2F.

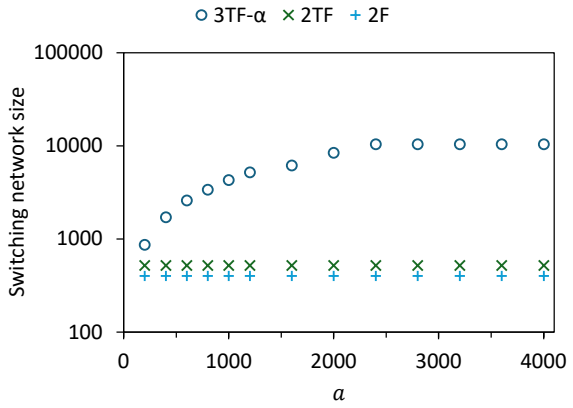

Fig. 5. Dependency of switching capacity on a with $N = 40$ and $\epsilon = 0$ (SNB condition).

Fig. 6 investigates the switching network size depending on ϵ with $N = 40$, $a = 2400$, and $p = 0.5$; that of the SNB condition is depicted as a reference. The larger ϵ is, the larger the switching network size of each design model is.

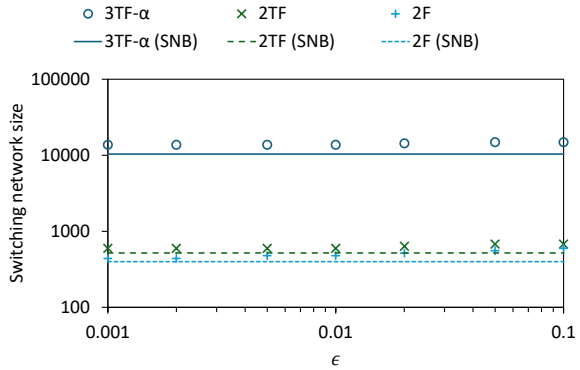

Fig. 6. Dependency of switching capacity on ϵ with $N = 40$, $a = 2400$, and $p = 0.5$.

VII. Conclusions

Upcoming data centers are expected to adopt advanced circuit switching methods, especially optical switching, to enhance transmission capacity and energy efficiency. A folded-type Clos network, characterized by its multi-level switch architecture with hierarchically interconnected switches, is a preferred design for data centers due to its scalability. There is a trade-off between the switching network size and the quality of a connection request. Addressing design challenges that seek to enlarge the switching network while keeping the blocking probability within acceptable limits is crucial.

This paper investigated the models for designing three-stage folded-type Clos networks, expanded from two-stage ones, with a blocking probability guarantee to maximize the switching network size. We designed a Clos network using identical switches whose size is $N \times N$ under the limited number of usable switches. We consider both expansion types to increase the number of stages from two to three: the input and output layer and the intermediate layer. We compared the performances of different models. Numerical results revealed that expanding the input-output layer (type α) effectively increases the switching network size, whereas expanding the intermediate layer (type β) does not.

References

[1] K.-I. Sato, "Optical switching will innovate intra data center networks [invited tutorial]," *J. Opt. Commun. Netw.*, vol. 16, no. 1, pp. 1–23, Jan. 2024.
[2] M. Taubenblatt, P. Maniotis, and A. Tantawi, "Optics enabled networks and architectures for data center cost and power efficiency," *J. Opt. Commun. Netw.*, vol. 14, no. 1, pp. A41–A49, Jan. 2022.
[3] Y. Shen, W. Wang, T. Liu, and J. Zhang, "Energy-efficient scaling of active electrical/optical switches in hybrid packet/circuit switched data center networks," in *Int. Conf. Opt. Commun. Netw.*, 2021, pp. 1–3.
[4] K. Anazawa, T. Inoue, T. Mano, H. Nishizawa, and E. Oki, "Efficient fiber-inspection and certification method for optical-circuit-switched datacenter networks," *J. Opt. Commun. Netw.*, vol. 16, no. 8, pp. 788–799, 2024.
[5] C. Clos, "A study of non-blocking switching networks," *Bell Syst. Tech. J.*, vol. 32, no. 2, pp. 406–424, 1953.
[6] A. Jajszczyk, "Nonblocking, repackable, and rearrangeable Clos networks: fifty years of the theory evolution," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 28–33, 2003.
[7] W. Kabaciński, *Nonblocking Electronic and Photonic Switching Fabrics*. Springer, 2005.
[8] E. Oki, Z. Jing, R. Rojas-Cessa, and H. Chao, "Concurrent round-robin-based dispatching schemes for Clos-network switches," *IEEE/ACM Trans. Netw.*, vol. 10, no. 6, pp. 830–844, 2002.
[9] E. Oki, N. Yamanaka, K. Nakai, and N. Matsuura, "Multi-stage switching system using optical WDM grouped links based on dynamic bandwidth sharing," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 56–63, 2003.
[10] E. Oki, H. Taka, and T. Inoue, "Enhancing capacity of optical circuit switching clos network in data center: Progress and challenges," in *2024 33rd Int. Conf. Compu. Commun. and Netw. (ICCCN)*, 2024, pp. 1–9.
[11] T. Mano, T. Inoue, K. Mizutani, and O. Akashi, "Increasing capacity of the Clos structure for optical switching networks," in *2019 IEEE Global Commun. Conf.*, 2019, pp. 1–6.
[12] ——, "Redesigning the nonblocking Clos network to increase its capacity," *IEEE Trans. Netw. Service Manag.*, vol. 20, no. 3, pp. 2558–2574, Sep. 2023.
[13] H. Taka, T. Inoue, and E. Oki, "Design of twisted and folded Clos network with guaranteeing admissible blocking probability," *IEEE Netw. Lett.*, vol. 5, no. 4, pp. 265–269, Dec. 2023.
[14] ——, "Twisted and folded Clos-network design model with two-step blocking probability guarantee," *IEEE Netw. Lett.*, vol. 6, no. 1, pp. 60–64, Mar. 2024.
[15] ——, "Design model of twisted and folded Clos network with multi-step grouped intermediate switches guaranteeing admissible blocking probability," *J. Opt. Commun. Netw.*, vol. 16, no. 3, pp. 328–341, Mar. 2024.
[16] E. Oki, H. Taka, and T. Inoue, "Toward increasing switching capacity of twisted and folded clos network guaranteeing admissible blocking probability," in *2024 24th Int. Conf. Transparent Opt. Netw. (ICTON)*, 2024, pp. 1–4.
[17] A. Jajszcyk, "On combinatorial properties of broadband time-division switching networks," *Comput. Netw. ISDN Sys.*, vol. 20, no. 1, pp. 377–382, 1990.
[18] R. Taniguchi, T. Inoue, K. Anazawa, and E. Oki, "Optical circuit switched three-stage twisted-folded Clos-network design model guaranteeing admissible blocking probability," *J. Opt. Commun. Netw.*, vol. 16, no. 11, pp. 1104–1115, Nov. 2024.