This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

The Global Packing Number of a Fat-tree Network

Yuan-Hsun Lo, Yijin Zhang, *Member, IEEE*, Yi Chen, Hung-Lin Fu, *Member, IEEE* and Wing Shing Wong, *Fellow, IEEE*

***Abstract*—Data centers play an important role in today's Internet development. Research to find scalable architecture and efficient routing algorithms for data center networks has gained popularity. The fat-tree architecture, which is essentially a folded version of a Clos network, has proven to be readily implementable and is scalable. In this paper, we investigate routing on a fat-tree network by deriving its *global packing number* and by presenting explicit algorithms for the construction of optimal, load-balanced routing solutions.**

Consider an optical network that employs Wavelength Division Multiplexing in which every user node sets up a connection with every other user node. The global packing number is basically the number of wavelengths required by the network to support such a traffic load, under the restriction that each source-to-destination connection is assigned a wavelength that remains constant in the network. In mathematical terms, consider a bidirectional, simple graph, G and let $N \subseteq V(G)$ be a set of nodes. A path system $\mathcal{P}$ of G with respect to N consists of $|N|(|N|-1)$ directed paths, one path to connect each of the source-destination node pairs in N. The global packing number of a path system $\mathcal{P}$, denoted by $\Phi(G, N, \mathcal{P})$, is the minimum integer k to guarantee the existence of a mapping $\phi : \mathcal{P} \to \{1, 2, \ldots, k\}$, such that $\phi(P) \neq \phi(\widehat{P})$ if P and $\widehat{P}$ have common arc(s). The global packing number of (G, N), denoted by $\Phi(G, N)$, is defined to be the minimum $\Phi(G, N, \mathcal{P})$ among all possible path systems $\mathcal{P}$. In additional to Wavelength Division optical networks, this number also carries significance for networks employing Time Division Multiple Access (TDMA).

In this paper, we compute by explicit route construction the global packing number of $(\mathbf{T}_n, N)$, where $\mathbf{T}_n$ denotes the topology of the n-ary fat-tree network, and N is considered to be the set of all edge switches or the set of all supported hosts. We show that the constructed routes are load-balanced and require minimal link capacity at all network links.

***Index Terms*—Global packing number, Fat-tree networks, Clos networks, Latin squares, Load-balancing**

This work was supported in part by a project from the Huawei Technologies Co. Ltd. under Grants YB2015100060 (Y.-H. Lo, Y. Zhang, Y. Chen and W. S. Wong), the National Natural Science Foundation of China under Grants 11601454 (Y.-H. Lo) and 61301107 (Y. Zhang), the Natural Science Foundation of Fujian Province, China under Grant 2016J05021 (Y.-H. Lo), the Fundamental Research Funds for the Central Universities in China under Grant 20720150210 (Y.-H. Lo), the Ministry of Science and Technology, Taiwan under Grant MOST 104-2115-M-009-009 (H.-L. Fu), and the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University, China under Grant 2017D09 (Y. Zhang).

Y.-H. Lo is with the School of Mathematical Sciences, Xiamen University, Xiamen 361005, China (e-mail: yhlo0830@gmail.com).

Y. Zhang is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, and also with National Mobile Communications Research Laboratory, Southeast University, Nanjing, China (e-mail: yijin.zhang@gmail.com).

Y. Chen is with the School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), and also with Shenzhen Research Institute of Big Data, Shenzhen, China (e-mail: chelseachenyi@gmail.com).

H.-L. Fu is with the Department of Applied Mathematics, National Chiao Tung University, Hsin Chu, Taiwan 30050 (e-mail: hlfu@math.nctu.edu.tw).

W. S. Wong is with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: wswong@ie.cuhk.edu.hk).

I. Introduction

As data center networks (DCNs) assume an increasingly important role in the Internet, interests on their architecture have built up steadily in the research community. Three models stand out among a host of proposed options: three-tier, fat-free, and DCell [1]. The fat-tree architecture is a folded version of a Clos network, which was originally designed for circuit switching [2]. The fat-tree architecture has subsequently been applied to parallel computing [3]. Its application to DCN has been proposed by S. B. Alexander et al. in [4]. Compared to the legacy three-tier model, the Clos network design enjoys a significant cost benefit and is scalable in regard to the number of connected servers. Google for example has adopted the Clos network framework in its DCN architecture design [5].

In this paper, we investigate the routing problem for an arbitrary fat-tree network under the assumption that every leave node in the tree has a communication connection with identical data rate to any other leave node. Although this fully loaded traffic scenario may not arise frequently, it can be regarded as a worst-case scenario as well as a reference point for more realistic traffic load analysis.

There are two main results reported in this paper. First, we derive the *global packing number* of a fat-tree network. Second, we explicitly construct routing solutions which achieve the global packing number while possessing good load-balanced property.

The global packing number was originally motivated by an interest to evaluate the minimum number of wavelengths required for an all-optical network [4], [6]. The network employs Wavelength Division Multiplexing (WDM) so that multiple communication connections can share an optical link at the same time by using distinct wavelengths. Although wavelength conversion is feasible, in the majority of all-optical networks there is a wavelength continuity constraint so that the same wavelength will be used for transmission over all the traversed links from source to destination. A route with a fixed wavelength will be referred to as a lightpath.

Given a network traffic pattern, one can construct different sets of lightpaths to satisfy the requirement. The global packing number is the minimum number of wavelengths required to support all the lightpaths. For a ring network, the global packing number has been solved in [7] and later in [8], assuming a uniform load for all source-destination pairs. In this paper, we compute the global packing number of a fat-tree network through explicit construction of routing paths and wavelength assignment. Moreover, we will show that the global packing number for a fat-tree network is identical to the maximum of all its link loads. For more global packing number results, please refer to static Routing and Wavelength

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

Assignment (RWA) problems [9], [10], [11], [12].

The global packing number has relevance beyond all-optical networks and is connected to network flows for example. A flow in a network refers to a sequence of packets from a given source-destination pair travelling over a fixed route. Packet forwarding of the flow can be implemented through a flow identifier embedded in the header of each individual packet in the flow. An intermediate node reads the flow identifier and looks up a flow forwarding table to determine how the packet should be forwarded to the next node. The flow identifiers need not be uniquely assigned to all flows. As long as two flows do not share any common link, it is in principle feasible for them to share the same identifier. The global packing number can be viewed as the maximum number of flow identifiers required for a given traffic pattern.

A third interpretation of the global packing number comes from zero-queueing delay routing as investigated under the DCN model in [13]. Consider a DCN in which the links between routers are divided into periodic frames composed of a fixed number of equal duration time slots. Due to the layered structure of a fat-tree network, one can group time slots from different links and assign them to the same flow so that packets from the flow can travel from source to destination without being buffered at any intermediate nodes, in other words, with zero-queueing. The global packing number can be used to define the minimal frame size so that, all flows enjoy the zero-queuing property.

We conclude this section by noting that the blocking problem on a 3-stage Clos network has been extensively studied in the literature, under various definitions of blocking, (such as strictly nonblocking, wide-sense nonblocking and rearrangeably nonblocking,) link condition, (such as single rate or multirate), and traffic models, (such as permutation routing.) We refer the readers to [14], [15], [16], [17] and [18], and the references contained therein for details. Although a small part of the results reported here concerning the global packing number could be derived based on these classical results, we would like to emphasise on two major differences. First, the fat-tree networks we consider in this paper are folded 5-stage or 7-stage Clos networks, so traditional 3-stage network results may or may not carry over. Second, our paper focuses on constructive, load-balanced solutions, which are not explicitly guaranteed by applying classical results. To make the load-balanced concept precise, we introduce the definition of a *well-balanced* routing solution in the following section.

The structure of the paper is as follows. After the introduction in Section I, we specify the notation used in the paper and provide a preliminary discussion of relevant basic concepts in Section II, along with some quick reference to useful combinatorial tools. Routing solution construction and global packing number computation will be presented in Section III and Section IV, where Section III focuses on the Edge-to-Edge traffic model and Section IV on the Host-to-Host traffic model. Section V is devoted to discussion of the relationship between global packing number and the maximum link load. A conclusion is presented in Section VI.

II. Preliminaries

A. Global packing

Let G be a connected simple graph with node set $V(G)$ and a link set $E(G)$. By putting two opposite arcs on each link of G we derive a bidirected graph. Let $A(G)$ be the set of arcs in the bidirected graph. Obviously, $|A(G)| = 2|E(G)|$. If there is no danger of confusion, we use $G = (V, A)$ to denote the bidirected graph and $G = (V, E)$ to denote the underlying simple graph. The global packing number is defined on a bidirected graph.

Let $G = (V, A)$ be a bidirected graph and $N \subseteq V(G)$ be a set of nodes. For any two nodes, $a, b \in N$, one can assign a particular directed path (or *dipath*) starting from node a to node b and denote the set of arcs in the dipath by $P_{(a,b)}$. Node a and node b are called the starting node and ending node of $P_{(a,b)}$, respectively. Note that $P_{(a,b)}$ is sometimes set to be one of the shortest paths from a to b. Let

$$\mathcal{P} := \{P_{(a,b)} : a, b \in N, a \neq b\}$$

denote a set of assigned dipaths, one for each distinct ordered node pair. We say $\mathcal{P}$ is a *path system* of G with respect to N, and let $\mathfrak{P}_{G,N}$ denote the collection of all such path systems.

Definition 1. Given a path system $\mathcal{P} \in \mathfrak{P}_{G,N}$, a *global packing* of $(G, N, \mathcal{P})$ is a mapping ϕ from $\mathcal{P}$ to a set of k distinct labels, such that for any two dipaths $P, \widehat{P} \in \mathcal{P}$, $\phi(P) \neq \phi(\widehat{P})$ provided that P and $\widehat{P}$ have one or more than one arc in common. The *global packing number* of $(G, N, \mathcal{P})$, denoted by $\Phi(G, N, \mathcal{P})$, is the minimum integer k to guarantee the existence of a global packing. The global packing number of (G, N), denoted by $\Phi(G, N)$, is defined to be the smallest global packing number of $(G, N, \mathcal{P})$ among all path systems $\mathcal{P} \in \mathfrak{P}_{G,N}$; i.e.,

$$\Phi(G, N) := \min_{\mathcal{P} \in \mathfrak{P}_{G,N}} \Phi(G, N, \mathcal{P}).$$

A path system $\mathcal{P}$ with $\Phi(G, N, \mathcal{P}) = \Phi(G, N)$ is said to be *ideal*.

A path system is said to be *symmetric* if $P_{(a,b)}$ and $P_{(b,a)}$ contain the same links but opposite directions. A global packing is said to be symmetric if both $P_{(a,b)}$ and $P_{(b,a)}$ receive the same label. The symmetric global packing number over all symmetric path systems can be simply defined on the underlying simple graph $G = (V, E)$, by considering both $P_{(a,b)}$ and $P_{(b,a)}$ as a non-directed path $P_{\{a,b\}}$. Let $\Phi^s(G, N)$ denote the symmetric global packing number of (G, N). Since a symmetric global packing is also a global packing by viewing a link as two opposite arcs, we have

$$\Phi(G, N) \leq \Phi^s(G, N). \tag{1}$$

The global packing number and symmetric global packing number for a ring network have been solved in [7] and [8] respectively. The results show that the equality in (1) may not hold in some cases. In fact, for a ring of n nodes, R_n, $\Phi(R_n, V(R_n)) = n^2/8$ and $\Phi^s(R_n, V(R_n)) = 1+n^2/8$ when n is doubly even.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

All terminologies and notations on graph theory used throughout this paper can be referred to the textbook written by D. B. West [19].

B. Maximum link load and natural bounds

Let $G = (V, A)$ be a bidirected graph and N be a subset of $V(G)$.

Definition 2. For $\mathcal{P} \in \mathfrak{P}_{G,N}$ define the *link load* of $\mathcal{P}$ by

$$L(\mathcal{P}) := \max_{a \in A(G)} \sum_{P \in \mathcal{P}} \mathbf{1}_{P \cap a \neq \emptyset},$$

where $\mathbf{1}_s = 1$ if s is true and $\mathbf{1}_s = 0$ otherwise. Then, the *maximum link load* (or, *required capacity*) of G with respect to N is defined by

$$C(G, N) := \min_{\mathcal{P} \in \mathfrak{P}_{G,N}} L(\mathcal{P}).$$

The maximum link load is a natural lower bound of global packing number, that is, for a graph G and any node set $N \in V(G)$ we have

$$\Phi(G, N) \geq C(G, N). \tag{2}$$

Let $\mathcal{P}$ be a path system of (G, N) and ϕ be a global packing of $(G, N, \mathcal{P})$. Then, for each label i, the preimage of i under ϕ, $\phi^{-1}(i)$, forms a set of mutually arc-disjoint dipaths. Another natural lower bound of $\Phi(G, N, \mathcal{P})$ is given as follows.

Proposition 3. *Let $\mathcal{P} \in \mathfrak{P}_{G,N}$ be a path system of (G, N). Then,*

$$\Phi(G, N, \mathcal{P}) \geq \left\lceil \frac{\sum_{P \in \mathcal{P}} |A(P)|}{|A(G)|} \right\rceil. \tag{3}$$

Proof. Assume $\Phi(G, N, \mathcal{P}) = k$. Let ϕ be a global packing of $(G, N, \mathcal{P})$ with k labels, say $0, 1, \ldots, k-1$. Since any set of mutually arc-disjoint dipaths consists of at most $|A(G)|$ arcs in total, we have

$$\sum_{P \in \phi^{-1}(i)} |A(P)| \leq |A(G)|, \tag{4}$$

for $i = 0, 1, \ldots, k-1$. On the other hand, each dipath is assigned a unique label under ϕ, so it is clear that

$$\sum_{P \in \mathcal{P}} |A(P)| = \sum_{i=0}^{k} \left(\sum_{P \in \phi^{-1}(i)} |A(P)| \right). \tag{5}$$

Combining (4) and (5) yields $\sum_{P \in \mathcal{P}} |A(P)| \leq k|A(G)|$. □

C. Fat-tree networks

A *fat-tree* topology is an efficient and economical architecture for data center networks. In an n-ary fat-tree there are n^2 $2n$-port *core switches* and $2n$ pods, each pod contains two layers of n $2n$-port switches: *aggregation switches* in the upper layer and *edge switches* in the lower layer. The core switches can be separated evenly into n groups. For $0 \leq i < n$, each switch of the i^{th} core group has one port connected to the i^{th} aggregation switch of each of $2n$ pods. That is, each aggregation switch is connected to n core switches. Each of the remaining n ports of aggregation switch is connected to the n edge switches in the same pod. Finally, each edge switch is directly connected to n *hosts*, which form a *subnet*. Therefore, an n-ary fat-tree network can support up to $2n^3$ hosts. There are three layers of arcs in a fat-tree. From top to bottom they are: between core and aggregation switches, between aggregation and edge switches, and between edge switches and hosts. We group them into three sets:

(i) $\mathcal{C}_1$: The set of arcs between core and aggregation switches.
(ii) $\mathcal{C}_2$: The set of arcs between aggregation and edge switches.
(iii) $\mathcal{C}_3$: The set of arcs between edge switches and hosts.

Obviously, $|\mathcal{C}_1| = |\mathcal{C}_2| = |\mathcal{C}_3| = 4n^3$ in an n-ary fat-tree. Fig. 1 illustrates the topology of 3-ary fat-tree. Note that in [20] an n-ary fat-tree refers to an n-port topology, instead of $2n$-port in this paper.

Let $\mathbf{T}_n$ represent an n-ary fat-tree network so that the link between any two nodes represents a communication channel. There are four classes of nodes in $\mathbf{T}_n$: core-nodes for core switches, aggregation-nodes for aggregation switches, edge-nodes for edge switches, and host-nodes for hosts. We use $c_{i,j}$ to denote the j^{th} core-node in the i^{th} core group, $a_{i,j}$ to denote the j^{th} aggregation-node in pod i, $e_{i,j}$ to denote the j^{th} edge-node in pod i, and $h_{t,i,j}$ denote the j^{th} host-node in the subnet under the edge-node $e_{t,i}$. Denoted by C_n, A_n, E_n and H_n the set of core-nodes, aggregation-nodes, edge-nodes and host-nodes in the fat-tree $\mathbf{T}_n$, respectively. Obviously, $|C_n| = n^2$, $|A_n| = |E_n| = 2n^2$ and $|H_n| = 2n^3$.

For a path system $\mathcal{P}$ of G with respect to N and a corresponding global packing ϕ, the *packing set* of arc $a \in A(G)$, denoted by ϕ_a, is defined to be the set of labels it receives under ϕ. That is,

$$\phi_a = \{\phi(P) : P \cap a \neq \emptyset, P \in \mathcal{P}\}.$$

Definition 4. Let ϕ be a global packing of a path system of $\mathbf{T}_n$. ϕ is called *well-balanced* if for any two arcs a and b in the same group $\mathcal{C}_i$, $i = 1, 2, 3$,

$$\phi_a = \phi_b.$$

D. Latin squares

For any positive integer n, let $\mathbb{Z}_n := \{0, 1, 2, \ldots, n-1\}$ denote the ring of residues modulo n. Let $\mathfrak{S}_n$ be the set of permutations on $\mathbb{Z}_n$. We adopt the notation $\pi = [\pi_0\, \pi_1\, \pi_2\, \cdots \pi_{n-1}]$, where $\pi_i = \pi(i)$ for $i \in \mathbb{Z}_n$, to denote a permutation $\pi \in \mathfrak{S}_n$.

An n by n array with entries in $\mathbb{Z}_n$ is called a *Latin square* of order n if none of integers in $\mathbb{Z}_n$ occurs twice within any row or column. In other words, all rows and columns are permutations in $\mathfrak{S}_n$. A Latin square of order n, denoted by $\mathsf{L} = (\ell_{i,j})$ for $i = 0, 1, \ldots, n-1$ and $j = 0, 1, \ldots, n-1$, is called *unipotent* (or *fixed diagonal*) if $\ell_{i,i} = 0$, $\forall i \in \mathbb{Z}_n$.

An n by n array defined on $\mathbb{Z}_n$ is a *partial Latin square* of order n if each integer in $\mathbb{Z}_n$ occurs at most once within any row or column. A partial Latin square $\mathsf{L} = (\ell_{i,j})$ is called *diagonal-free* provided cell $\ell_{i,j}$ is filled if and only if $i \neq j$. We use diagonal-free square to denote a partial diagonal-free Latin square for short. It is worth mentioning that a diagonal-free square is different from a *holey Latin square* of type 1^n,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

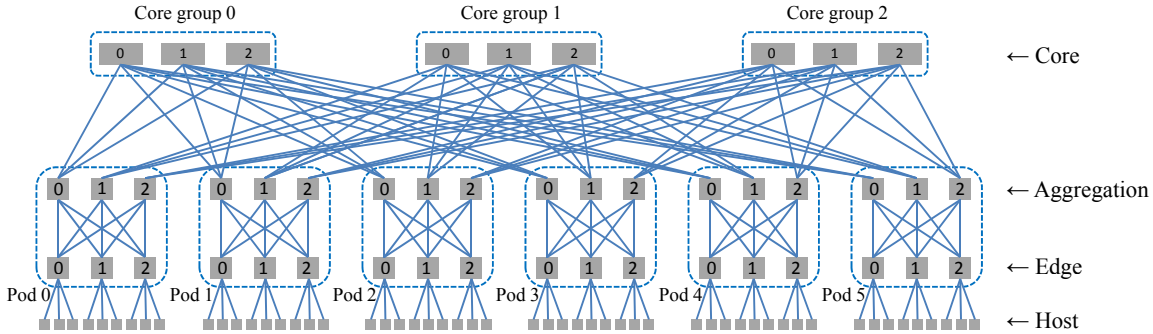

Fig. 1. The 3-ary fat-tree network.

because it is always possible to extend a holey Latin square to a Latin square, but not for a diagonal-free square.

See Fig. 2 for examples of Latin square, unipotent Latin square and diagonal-free square of order 4.

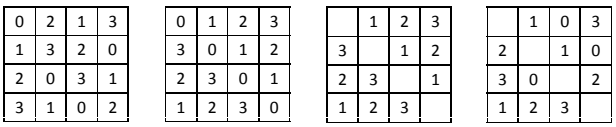

Fig. 2. (a) Latin square, (b) unipotent Latin square, and (c)–(d) diagonal-free squares of order 4.

Denote by $\mathrm{LS}(n)$ the set of all distinct Latin squares of order n, $\mathrm{ULS}(n)$ the set of all distinct unipotent Latin squares of order n, $\mathrm{DFS}(n)$ the set of all distinct diagonal-free squares of order n, and by $\mathrm{ls}(n)$, $\mathrm{uls}(n)$, $\mathrm{dfs}(n)$ their respective cardinalities. Obviously, we can construct a diagonal-free square by deleting all diagonal entries of a unipotent Latin square. See Fig. 2(b) and Fig. 2(c) for instance. Therefore, $\mathrm{uls}(n) \le \mathrm{dfs}(n)$. Please refer to A002860 and A000479 in [21] for the values of $\mathrm{ls}(n)$ and $\mathrm{uls}(n)$, respectively. See [22] for more information on (unipotent) Latin squares.

Consider an n-ary fat-tree network $\mathbf{T}_n$. This paper investigates the global packing number of $(\mathbf{T}_n, N)$ for two cases: $N = E_n$ and $N = H_n$. The approach used here is to find lower bounds for the two cases by (2) and (3), and then to construct corresponding path systems by means of Latin squares, unipotent Latin squares and diagonal-free squares such that these lower bounds are achieved. It is worth mentioning that the use of such permutation structures on engineering problems is an old fashion and can be found in literatures, such as [23], [24] on recursive circulant networks (RCNs), [25] on reliable data broadcasting by independent spanning trees (ISTs), and [26] on cognitive radio networks (CR networks).

III. Edge-to-edge traffic

Recall that E_n is the set of edge-nodes in T_n. This section is devoted to the computation of global packing number of $(\mathbf{T}_n, E_n)$.

In E_n there are two classes of edge-node pairs: (i) both edge-nodes are in the same pod; and (ii) edge-nodes are in different pods. Let $P_{(t,i),(t',i')}$ denote a dipath starting from edge-node $e_{t,i}$ to $e_{t',i'}$. Any path system of $\mathbf{T}_n$ with respect to E_n can be partitioned into two sets by

$$\mathcal{E}_1 := \{P_{(t,i),(t,i')} : t \in \mathbb{Z}_{2n}, i, i' \in \mathbb{Z}_n, i \neq i'\}, \tag{6}$$

$$\mathcal{E}_2 := \{P_{(t,i),(t',i')} : t, t' \in \mathbb{Z}_{2n}, i, i' \in \mathbb{Z}_n, t \neq t'\}. \tag{7}$$

Note that $|\mathcal{E}_1| = 2n^2(n-1)$ and $|\mathcal{E}_2| = 2n^3(2n-1)$.

Lemma 5. *For any integer $n > 1$, we have $\Phi(\mathbf{T}_n, E_n) \ge 2n$.*

Proof. Since an edge-to-edge directed path does not contain any arc in $\mathcal{C}_3$, we have $\Phi(\mathbf{T}_n, E_n) = \Phi(\mathbf{T}'_n, E_n)$, where $\mathbf{T}'_n$ is the graph obtained from $\mathbf{T}_n$ by cutting off all host-nodes. Let $\mathcal{P} = \mathcal{E}_1 \uplus \mathcal{E}_2$ be an ideal path system of $(\mathbf{T}'_n, E_n)$, where $\mathcal{E}_1, \mathcal{E}_2$ are shown in (6) – (7). If $P \in \mathcal{E}_1$, then $|A(P)| \ge 2$; otherwise, $|A(P)| \ge 4$. So we have

$$\sum_{P \in \mathcal{P}} |A(P)| \ge 2|\mathcal{E}_1| + 4|\mathcal{E}_2| = 4n^2\big(4n^2 - n - 1\big). \tag{8}$$

By the fact that $|A(\mathbf{T}'_n)| = 8n^3$, we derive from Proposition 3 that

$$\begin{aligned}\Phi(\mathbf{T}'_n, E_n) &= \Phi(\mathbf{T}'_n, \mathcal{P}, E_n) \\ &\ge \left\lceil \frac{\sum_{P \in \mathcal{P}} |A(P)|}{|A(\mathbf{T}'_n)|} \right\rceil = \left\lceil \frac{4n^2 - n - 1}{2n} \right\rceil \\ &= 2n,\end{aligned}$$

whenever $n \ge 2$. Hence $\Phi(\mathbf{T}_n, E_n) = \Phi(\mathbf{T}'_n, E_n) \ge 2n$. □

Construction 1 (First canonical path system)**.** *Let $\mathsf{D} = (d_{i,j})$ be a diagonal-free square of order n, $\mathsf{L} = (\ell_{i,j})$ be a Latin square of order n, and π be a permutation on $\mathbb{Z}_n$. The first $(\mathsf{D}, \mathsf{L}, \pi)$-based canonical path system $\mathcal{P}^* = \mathcal{E}_1 \uplus \mathcal{E}_2$ of $\mathbf{T}_n$ with respect to E_n, where $\mathcal{E}_1, \mathcal{E}_2$ are as defined in* (6)–(7)*, is defined as*

(i) for $P_{(t,i),(t,i')} \in \mathcal{E}_1$ let

$$P_{(t,i),(t,i')} = e_{t,i} \to a_{t,d_{i,i'}} \to e_{t,i'}; \tag{9}$$

(ii) for $P_{(t,i),(t',i')} \in \mathcal{E}_2$ let

$$P_{(t,i),(t',i')} := e_{t,i} \to a_{t,\ell_{i,i'}} \to c_{\ell_{i,i'},\pi_i} \to a_{t',\ell_{i,i'}} \to e_{t',i'}. \tag{10}$$

See Fig. 3–4 for an illustration.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

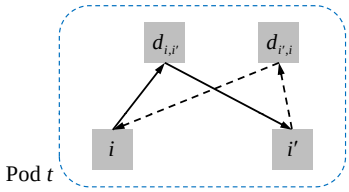

Fig. 3. The dipath setting in the same pod of Construction 1: solid dipath for $P_{(t,i),(t,i')}$ and dash dipath for $P_{(t,i'),(t,i)}$.

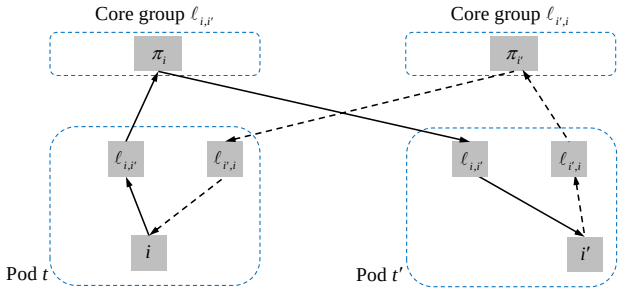

Fig. 4. The dipath setting between different pods of Construction 1: solid dipath for $P_{(t,i),(t',i')}$ and dash dipath for $P_{(t',i'),(t,i)}$.

Example 1. Consider a 3-ary fat-tree. Let D and L be the diagonal-free square and Latin square shown in Fig. 5, and let $\pi = [1\,2\,0]$ be a permutation in $\mathfrak{S}_3$.

Following Construction 1, in Pod 1 the edge-node $e_{1,1}$ connects to $e_{1,2}$ via $a_{1,1}$, while $e_{1,2}$ connects to $e_{1,1}$ via $a_{1,0}$ in reverse. And, the dipaths connecting $e_{0,1}$ and $e_{2,2}$ are set to be

$$e_{0,1} \to a_{0,1} \to c_{1,2} \to a_{2,1} \to e_{2,2}$$

and

$$e_{2,2} \to a_{2,2} \to c_{2,0} \to a_{0,2} \to e_{0,1},$$

since $\pi_1 = 2$ and $\pi_2 = 0$.

Lemma 6. *For any integer $n > 1$, we have $\Phi(\mathbf{T}_n, E_n) \le 2n$.*

Proof. Let $\mathcal{P}^*$ be a $(\mathsf{D}, \mathsf{L}, \pi)$-based canonical path system of $\mathbf{T}_n$ with respect to E_n for some $\mathsf{D} \in \mathrm{DFS}(n)$, $\mathsf{L} \in \mathrm{LS}(n)$ and $\pi \in \mathfrak{S}_n$. It suffices to find a global packing from $\mathcal{P}^*$ to a set of $2n$ distinct labels, say $\{0, 1, \ldots, 2n-1\}$.

Recall that $\mathcal{P}^* = \mathcal{E}_1 \uplus \mathcal{E}_2$, where $\mathcal{E}_1$ is the set of dipaths whose endpoints are in the same pod and $\mathcal{E}_2$ otherwise. Let $\mathsf{F} = (f_{i,j})$ be a unipotent Latin square of order $2n$. The mapping ϕ^* is defined by

$$\phi^*(P) = \begin{cases} 0, & \text{if } P \in \mathcal{E}_1; \quad (11a) \\ f_{t,t'}, & \text{if } P = P_{(t,i),(t',i')} \in \mathcal{E}_2. \quad (11b) \end{cases}$$

Now, we claim ϕ^* is a global packing of $(\mathbf{T}_n, E_n)$. Suppose $P \neq \widehat{P}$ have one common arc, κ, and $\phi^*(P) = \phi^*(\widehat{P}) = \lambda$.

Case 1. $\lambda = 0$. In this case, κ is an arc between edge- and aggregation-nodes in some pod t. By symmetry, let $\kappa = (e_{t,i}, a_{t,k})$ for some i and k. Assume $e_{t,j}$ and $e_{t,\hat{j}}$ are the ending nodes of P and $\widehat{P}$, respectively, where $i, j, \hat{j}$ are all distinct. By (9), $d_{t,j} = k = d_{t,\hat{j}}$, which is a contradiction since D is a diagonal-free square.

Case 2. $\lambda > 0$. In this case, κ may be an arc in $\mathcal{C}_1$ or $\mathcal{C}_2$. First, consider $\kappa \in \mathcal{C}_2$. Assume by symmetry the case that $\kappa = (e_{t,i}, a_{t,k})$ for some t, i and k; that is, P and $\widehat{P}$ have common starting node. Since F is a unipotent Latin square, there exists a unique column index, say t', such that $f_{t,t'} = \lambda$. Therefore, the ending nodes of P and $\widehat{P}$ are both in pod t'. Assume $e_{t',j}$ and $e_{t',\hat{j}}$ are these two ending nodes. We now consider two subcases: $j = \hat{j}$ and $j \neq \hat{j}$. If $j = \hat{j}$, we have $P = \widehat{P}$ because dipaths are uniquely determined by their starting and ending nodes by (10). If $j \neq \hat{j}$, we have $\ell_{i,j} = k = \ell_{i,\hat{j}}$, a contradiction to that L is assumed to be a Latin square. It remains to consider that $\kappa \in \mathcal{C}_1$. Again, consider by symmetry the case that $\kappa = (a_{t,k}, c_{k,x})$ for some t, k and x. Notice that P and $\widehat{P}$ have the same aggregation-node $a_{t,k}$, so they are from the same pod t. Assume the two starting nodes are $e_{t,i}$ and $e_{t,\hat{i}}$ for some i and $\hat{i}$. By (10), we have $\pi_i = x = \pi_{\hat{i}}$, and thus $i = \hat{i}$. Therefore, P and $\widehat{P}$ have a common arc in $\mathcal{C}_2$, which yields a contradiction by the argument above. Hence we complete the proof. □

	1	0
2		1
1	0	

D

0	1	2
2	0	1
1	2	0

L

Fig. 5. A diagonal-free square and a Latin square of order 3.

Theorem 7. *For any integer $n > 1$, we have $\Phi(\mathbf{T}_n, E_n) = 2n$. Moreover, the packing sets of all arcs in $\mathbf{T}_n$ under the global packing ϕ^* defined in* (11a), (11b) *are of minimal size, and*

(i) for any arc $a \in \mathcal{C}_1$, $|\phi^*_a| = 2n - 1$;
(ii) for any arc $a \in \mathcal{C}_2$, $2n - 1 \le |\phi^*_a| \le 2n$.

Proof. The conclusion that $\Phi(\mathbf{T}_n, E_n) = 2n$ directly follows from Lemma 5 and Lemma 6. Given a global packing ϕ of a path system $\mathcal{P}$, one can study the load-balancing performance of it by counting the cardinality of the set

$$\Pi_{\phi,\mathcal{P}} := \{(a, \ell) : a \in A(\mathbf{T}_n), \text{ arc } a \text{ receives label } \ell \text{ under } \phi\}.$$

For a fixed dipath $P \in \mathcal{P}$, every arc receives a unique label, so P can support a number of $|A(P)|$ ordered pairs to the set $\Pi_{\phi,\mathcal{P}}$. By going through all dipaths, by (8), we have

$$|\Pi_{\phi,\mathcal{P}}| = \sum_{P \in \mathcal{P}} |A(P)| \ge 4n^2\big(4n^2 - n - 1\big). \quad (12)$$

Now, consider the structure of the first canonical path system, $\mathcal{P}^*$, and the global packing, ϕ^*, defined in Lemma 6. By (10) and (11b), every arc in $\mathcal{C}_1 \cup \mathcal{C}_2$ receives a label from $\{1, 2, \ldots, 2n-1\}$. Since label 0 is assigned to dipaths in $\mathcal{E}_1$, where all arcs are from $\mathcal{C}_2$, we have

$$\phi^*_a = \{1, 2, \ldots, 2n-1\}, \text{ for } a \in \mathcal{C}_1.$$

It suffices to consider the arcs in $\mathcal{C}_1$ which do not receive label 0, in other words, do not appear in $\mathcal{E}_1$. By (6) and (9), they are those of the form $(e_{t,i}, a_{t,d_{i,i}})$ or $(a_{t,d_{i,i}}, e_{t,i})$, where $\mathsf{D} =$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

$(d_{i,j})$ is the given diagonal-free square of order n. Therefore, for any arc $a \in \mathcal{C}_2$,

$$\phi_a^* = \begin{cases} \{1,2,\ldots,2n-1\}, & \text{if } a = (e_{t,i}, a_{t,d_{i,i}}) \text{ or} \\ & (a_{t,d_{i,i}}, e_{t,i}) \text{ for some } t,i, \\ \{0,1,2,\ldots,2n-1\}, & \text{otherwise.} \end{cases}$$

Note that there are exactly $4n^2$ arcs in $\mathcal{C}_2$ having packing set $\{1,2,\ldots,2n-1\}$. It is also clear that for $a \in \mathcal{C}_3$, $\phi_a^* = \emptyset$. By $|\mathcal{C}_1| = |\mathcal{C}_2| = 4n^3$, we have

$$\begin{aligned} |\Pi_{\phi^*,\mathcal{P}^*}| &= 4n^3(2n-1) + 4n^2(2n-1) + (4n^3 - 4n^2)2n \\ &= 4n^2(4n^2 - n - 1), \end{aligned}$$

which attains the minimum value in (12). □

Corollary 8. *For any integer $n > 1$, there exist at least*

$$\big(\mathrm{dfs}(n)\big)^{2n}\big(\mathrm{ls}(n)\cdot n!\big)^{2n(2n-1)}$$

ideal path systems of $\mathbf{T}_n$ with respect to E_n.

Proof. Lemma 6 proved that any $(\mathsf{D},\mathsf{L},\pi)$-based canonical path system $\mathcal{P}^* = \mathcal{E}_1 \uplus \mathcal{E}_2$ is ideal. In fact, we essentially have proved (in Case 1 of Lemma 6) that all dipaths in the same pod are mutually arc-disjoint if the connection is based on a diagonal-free square. Then, we can use a different diagonal-free square for each pod t. Similarly, we have proved (in Case 2 of Lemma 6) that all dipaths between any two pods are mutually arc-disjoint if the connection is based on a Latin square and a permutation. So, we can use a different Latin square and a different permutation for each pair of distinct pods t, t', respectively. Hence the result follows from the fact that there are $2n$ pods in $\mathbf{T}_n$.

□

IV. Host-to-host traffic

Recall that H_n is the set of host-nodes in $\mathbf{T}_n$. This section is devoted to the computation of global packing number of $(\mathbf{T}_n, H_n)$. We first derive its lower bound.

Lemma 9. *For any positive integer n, we have $\Phi(\mathbf{T}_n, H_n) \geq 2n^3 - 1$.*

Proof. Any host-to-host dipath with a given starting node, say $h_{t,i,j}$, must pass through the edge-node $e_{t,i}$ and the arc $(h_{t,i,j}, e_{t,i})$. So, the maximum link load of $\mathbf{T}_n$ respect to H_n is at least $2n^3 - 1$, which implies that $\Phi(\mathbf{T}_n, H_n) \geq C(\mathbf{T}_n, H_n) \geq 2n^3 - 1$ by (2). □

In H_k there are three classes of host-node pairs: (i) both host-nodes are in the same subnet; (ii) both host-nodes are in the same pod but not in the same subnet; and (iii) host-nodes are in different pods. Let $P_{(t,i,j),(t',i',j')}$ denote a dipath from host-node $h_{t,i,j}$ to $h_{t',i',j'}$. Any path system of $\mathbf{T}_n$ with respect to H_k can be partitioned into three sets by

$$\mathcal{H}_1 := \{P_{(t,i,j),(t,i,j')} : t \in \mathbb{Z}_{2n}, i,j,j' \in \mathbb{Z}_n, j \neq j'\}, \quad (13)$$

$$\mathcal{H}_2 := \{P_{(t,i,j),(t,i',j')} : t \in \mathbb{Z}_{2n}, i,i',j,j' \in \mathbb{Z}_n, i \neq i'\}, \quad (14)$$

and

$$\mathcal{H}_3 := \{P_{(t,i,j),(t',i',j')} : t,t' \in \mathbb{Z}_{2n}, i,i',j,j' \in \mathbb{Z}_n, t \neq t'\}. \quad (15)$$

Note that $|\mathcal{H}_1| = 2n^3(n-1)$, $|\mathcal{H}_2| = 2n^4(n-1)$, and $|\mathcal{H}_3| = 2n^5(2n-1)$.

Now, we define the host-to-host version of canonical path system of $\mathbf{T}_n$ with respect to H_n.

Construction 2 (Second canonical path system)**.** *Let $\mathsf{L} = (\ell_{i,j})$ be a Latin square of order n and $\pi \in \mathfrak{S}_n$. The second (L,π)-based canonical path system $\mathcal{P}^\sharp = \mathcal{H}_1 \uplus \mathcal{H}_2 \uplus \mathcal{H}_3$ of $\mathbf{T}_n$ with respect to H_n, where $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ are defined in* (13) – (15)*, is defined as*

(i) for $P_{(t,i,j),(t,i,j')} \in \mathcal{H}_1$ let

$$P_{(t,i,j),(t,i,j')} := h_{t,i,j} \to e_{t,i} \to h_{t,i,j'}; \quad (16)$$

(ii) for $P_{(t,i,j),(t,i',j')} \in \mathcal{H}_2$ let

$$\begin{aligned} P_{(t,i,j),(t,i',j')} := \; & h_{t,i,j} \to e_{t,i} \to \\ & a_{t,j} \to e_{t,i'} \to h_{t,i',j'}; \text{ and} \end{aligned} \quad (17)$$

(iii) for $P_{(t,i,j),(t',i',j')} \in \mathcal{H}_3$ let

$$\begin{aligned} P_{(t,i,j),(t',i',j')} := \; & h_{t,i,j} \to e_{t,i} \to \\ & a_{t,\ell_{i,i'}} \to c_{\ell_{i,i'},\pi_i} \to \\ & a_{t',\ell_{i,i'}} \to e_{t',i'} \to h_{t',i',j'}. \end{aligned} \quad (18)$$

See Fig. 6 for an illustration of (17)*.*

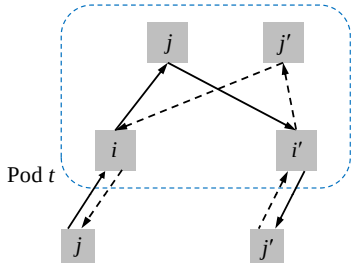

Fig. 6. The dipath setting in the same pod of Construction 2: solid dipath for $P_{(t,i,j),(t,i',j')}$ and dash dipath for $P_{(t,i',j'),(t,i,j)}$.

Comparing (18) *with* (10)*, the dipaths in $\mathcal{H}_3$ can be realized by appending two host-nodes to the associated edge-to-edge dipath $P_{(t,i),(t',i')} \in \mathcal{E}_2$ defined in Construction 1. Let $\mathcal{P}^*$ be a $(\mathsf{D},\mathsf{L},\pi)$-based canonical path system of $\mathbf{T}_n$ with respect to E_n for some $\mathsf{D} \in DFS(n)$. One can rewrite* (18) *in*

$$P_{(t,i,j),(t',i',j')} := h_{t,i,j} \to P_{(t,i),(t',i')} \to h_{t',i',j'}. \quad (19)$$

Lemma 10. *For any positive integer, we have $\Phi(\mathbf{T}_n, H_n) \leq 2n^3 - 1$.*

Proof. Let $\mathcal{P}^\sharp$ be an (L,π)-based canonical path system of $\mathbf{T}_n$ with respect to H_n, for some $\mathsf{L} \in \mathrm{LS}(n)$ and $\pi \in \mathfrak{S}_n$. It suffices to find a global packing from $\mathcal{P}^\sharp$ to a set of $2n^3 - 1$ distinct labels, say $\{1,2,\ldots,2n^3-1\}$.

Observe that $\mathcal{P}^\sharp$ can be partitioned into three parts: $\mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{H}_3$ according to Construction 2. Let $\mathsf{F} = (f_{i,j})$ be a unipotent Latin square of order n, $\mathsf{G} = (g_{i,j})$ be a unipotent Latin square of order $2n$, and $\mathsf{M}^{(1)} = (m^{(1)}_{i,j})$, $\mathsf{M}^{(2)} = (m^{(2)}_{i,j})$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

$\mathsf{M}^{(3)} = (m^{(3)}_{i,j})$ be three Latin squares (may be the same) of order n. The mapping $\phi^\sharp$ is defined as

(i) for $P = P_{(t,i,j),(t,i,j')} \in \mathcal{H}_1$ let

$$\phi^\sharp(P) = f_{j,j'}; \tag{20}$$

(ii) for $P = P_{(t,i,j),(t,i',j')} \in \mathcal{H}_2$ let

$$\phi^\sharp(P) = n \cdot f_{i,i'} + m^{(1)}_{j,j'}; \text{ and} \tag{21}$$

(iii) for $P = P_{(t,i,j),(t',i',j')} \in \mathcal{H}_3$ let

$$\phi^\sharp(P) = n^2 \cdot g_{t,t'} + n \cdot m^{(1)}_{i,i'} + m^{(1)}_{j,j'}. \tag{22}$$

Now, we claim that $\phi^\sharp$ is a global packing of $(\mathbf{T}_n, H_n)$. Suppose $P \neq \widehat{P}$ have one common arc, κ, and $\phi^\sharp(P) = \phi^\sharp(\widehat{P}) = \lambda$. Since, by (20) – (22), the images of $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ under $\phi^\sharp$ are $\phi^\sharp(\mathcal{H}_1) = \{1, 2, \ldots, n-1\}$, $\phi^\sharp(\mathcal{H}_2) = \{n, n+1, \ldots, n^2-1\}$ and $\phi^\sharp(\mathcal{H}_3) = \{n^2, n^2+1, \ldots, 2n^3-1\}$, we have the following three cases according to the value of λ.

Case 1. $1 \leq \lambda \leq n-1$. In this case, we have $P, \widehat{P} \in \mathcal{H}_1$, and thus κ is a host-to-edge arc in $\mathcal{C}_3$. By symmetry, assume $h_{t,i,j}$ is the common starting node of P and $\widehat{P}$, whose ending nodes are $h_{t,i,j'}$ are $h_{t,i,\hat{j}'}$ respectively, for some t, i, j, j' and $\hat{j}'$. Note that $j \neq j' \neq \hat{j}'$. By (20), $f_{j,j'} = \kappa = f_{j,\hat{j}'}$, which is a contradiction to that F is a unipotent Latin square.

Case 2. $n \leq \lambda \leq n^2 - 1$. In this case, we have $P, \widehat{P} \in \mathcal{H}_2$, and thus κ may be an arc in $\mathcal{C}_2$ or $\mathcal{C}_3$. Assume $P = P_{(t,i,j),(t,i',j')}$ and $P = P_{(t,\hat{i},\hat{j}),(t,\hat{i}',\hat{j}')}$ for some $t, i, j, i', j', \hat{i}, \hat{j}, \hat{i}'$ and $\hat{j}'$. Since $\phi^\sharp(P) = \phi^\sharp(\widehat{P})$, by (21) we have

$$\begin{cases} f_{i,i'} = f_{\hat{i},\hat{i}'}, \text{ and} & \text{(23a)} \\ m^{(1)}_{j,j'} = m^{(1)}_{\hat{j},\hat{j}'}. & \text{(23b)} \end{cases}$$

If $\kappa \in \mathcal{C}_3$, then either $i = i', j = j'$ or $\hat{i} = \hat{i}', \hat{j} = \hat{j}'$. Either case will imply $(i, i', j, j') = (\hat{i}, \hat{i}', \hat{j}, \hat{j}')$ by (23a)–(23b), namely, $P = \widehat{P}$, a contradiction. The subcase that $\kappa \in \mathcal{C}_2$ can be dealt with in the same way, and the proof is omitted.

Case 3. $n^2 \leq \lambda \leq 2n^3 - 1$. In this case, we have $P, \widehat{P} \in \mathcal{H}_3$. Assume $P = P_{(t,i,j),(t',i',j')}$ and $P = P_{(\hat{t},\hat{i},\hat{j}),(\hat{t}',\hat{i}',\hat{j}')}$ for some $t, i, j, t', i', j', \hat{t}, \hat{i}, \hat{j}, \hat{t}', \hat{i}'$ and $\hat{j}'$. Since $\phi^\sharp(P) = \phi^\sharp(\widehat{P})$, by (22) we have

$$\begin{cases} g_{t,t'} = g_{\hat{t},\hat{t}'}, & \text{(24a)} \\ m^{(2)}_{i,i'} = m^{(2)}_{\hat{i},\hat{i}'}, \text{ and} & \text{(24b)} \\ m^{(3)}_{j,j'} = m^{(3)}_{\hat{j},\hat{j}'}. & \text{(24c)} \end{cases}$$

Since P and $\widehat{P}$ are assumed to have common arcs, we only need to consider the case that $\{t, t'\} = \{\hat{t}, \hat{t}'\}$ by (24a). Without loss of generality, suppose $t = \hat{t}$ and $t' = \hat{t}'$. Notice that P and $\widehat{P}$ are obtained by appending host-nodes to associated edge-to-edge dipaths which are defined in (10). Since we have proved in the case 2 of the proof of Lemma 6 that, all edge-to-edge dipaths between any two fixed pods are mutually arc-disjoint, it remains to consider that κ is a host-edge arc. Then, we have either $i = i'$, $j = j'$ or $\hat{i} = \hat{i}'$, $\hat{j} = \hat{j}'$. By (24b) – (24c), we have $(i, i', j, j') = (\hat{i}, \hat{i}', \hat{j}, \hat{j}')$, namely, $P = \widehat{P}$, a contradiction. Hence we complete the proof. □

Lemma 10 not only shows that each (L, π)-based canonical path system defined in Construction 2 is ideal, it also highlights that the path system is load-balancing-wise "optimal" as being made precise in the following Theorem:

Theorem 11. *For any positive integer n, we have $\Phi(\mathbf{T}_n, H_n) = 2n^3 - 1$. Moreover, the global packing defined in* (20)–(22) *is well-balanced and the packing sets of all arcs in $\mathbf{T}_n$ are of minimal size with*

(i) for any arc $a \in \mathcal{C}_1$, $|\phi^\sharp_a| = n(2n^2 - 1)$;
(ii) for any arc $a \in \mathcal{C}_2$, $|\phi^\sharp_a| = n^2(2n - 1)$; and
(iii) for any arc $a \in \mathcal{C}_3$, $|\phi^\sharp_a| = \Phi(\mathbf{T}_n, H_n) = 2n^3 - 1$.

Proof. $\Phi(\mathbf{T}_n, H_n) = 2n^3 - 1$ is directed from Lemma 9 and Lemma 10. From the definition of $\phi^\sharp$ and the structure of the second canonical path system $\mathcal{P}^\sharp$, one has

$$\phi^\sharp_a = \begin{cases} \{n^2, n^2+1, \ldots, 2n^3-1\}, & \text{if } a \in \mathcal{C}_1, \\ \{n, n+1, \ldots, 2n^3-1\}, & \text{if } a \in \mathcal{C}_2, \\ \{1, 2, \ldots, 2n^3-1\}, & \text{if } a \in \mathcal{C}_3. \end{cases}$$

Then, the result follows by using the same counting method as in the proof of Theorem 7. □

Corollary 12. *For any positive integer n, there exist at least*

$$\big(\mathrm{ls}(n) \cdot n!\big)^{2n(2n-1)}$$

ideal path systems of $\mathbf{T}_n$ with respect to H_n.

Proof. The proof is similar to that of Corollary 8, and is omitted. □

V. Global packing numbers v.s. Maximum link load

Our results in previous sections reveal that the both sides of equation (2) are identical for the case that $G = \mathbf{T}_n$ and $N = E_n$ or H_n. That is,

$$\Phi(\mathbf{T}_n, E_n) = C(\mathbf{T}_n, E_n)$$

and

$$\Phi(\mathbf{T}_n, H_n) = C(\mathbf{T}_n, H_n).$$

In fact, it has been conjectured in [10] that $\Phi(G, V(G)) = C(G, V(G))$ for any bidirected graph G, and the conjecture is proved to be true for some particular graphs: cycles [7], trees [11], and some Cartesian product grpahs [9], [12]. Based on our results, it would be interesting to consider if $\Phi(G, N) = C(G, N)$ for any bidirected graph G and for any set of nodes $N \subseteq V(G)$.

When it comes to symmetric global packing numbers, following (1) and (2) we also have

$$\Phi^s(G, N) \geq C(G, N). \tag{25}$$

It would be more difficult to make the equality hold in (25) than in (2), since there may be a gap between symmetric global packing number and global packing number. Here, we take the complete binary tree as an example.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2710184, IEEE Transactions on Information Theory

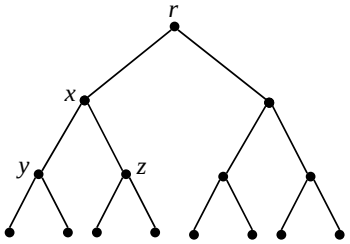

Fig. 7. Complete binary tree of height 3.

Let T be a complete binary tree of height 3, see Fig. 7. It is not hard to see that $C(T, V(T)) = 56$ and (r, x) or (x, r) is the arc where the maximum link load occurs. There is a unique path system of $(T, V(T))$, since there is only one path between any two nodes. Let us consider $\Phi^s(T, V(T))$. Suppose $\Phi^s(T, V(T)) = 56$ and ϕ is an ideal symmetric global packing. Let $\mathcal{P}_1$ be the set of these 56 paths that contain the common link $\{x, r\}$. Let $\mathcal{P}_2$ is the collection of paths connecting the nodes of the subtree under y and that under z. Obvious, $|\mathcal{P}_2| = 9$. Since any path in $\mathcal{P}_2$ contains links $\{x, y\}$ and $\{x, z\}$, the nine paths must receive distinct symbols under ϕ, i.e., $|\phi(\mathcal{P}_2)| = 9$. On the other hand, among $\mathcal{P}_1$ there are exactly 8 paths that do not contain links $\{x, y\}$ and $\{x, z\}$, namely, those paths connecting x with r and the nodes on the right subtree under r. So the remaining paths in $\mathcal{P}_1$ must receive different labels from $\phi(\mathcal{P}_2)$. Therefore, we have $|\phi(\mathcal{P}_1 \cup \mathcal{P}_2)| \geq 56 - 8 + 9 = 57$, a contradiction to $\Phi^s(T, V(T)) = 56$. Note that in [11] it has been shown that $\Phi(T, V(T)) = C(T, V(T))$. Hence in this example we conclude that $\Phi^s(T, V(T)) > \Phi(T, V(T)) = C(T, V(T))$.

VI. Concluding remarks

The global packing number is a newly defined index that characterizes the capacity required to support a uniformly loaded communication network. This paper focuses on the fat-tree networks, which are popular choices for data center communication architecture. Based on the notion of Latin square, we compute the global packing number for a fat-tree network with respect to the set of all edge switches or hosts, and provide a construction algorithm for ideal path systems. Our result also shows that the global packing number for a fat-tree network is identical to the maximum link load.

Acknowledgement

The authors would like to express their gratitude to the referees for their valuable comments and suggestions in improving the presentation of this paper.

References

[1] K. Bilal, S. U. Khan, L. Zhang, H. Li, K. Hayat, S. A. Madani, N. Min-Allah, L. Wang, D. Chen, M. Iqbal, C.-Z. Xu, and A. Y. Zomaya, "Quantitative Comparisons of the State of the Art Data Center Architectures," Concurrency and Computation: Practice and Experience, vol. 25, no. 12, pp. 1771–1783, August 2013.

[2] C. Clos, "A Study of Non-Blocking Switching Networks," Bell System Technical Journal, vol. 32, no. 2, pp. 406–424, March 1953.

[3] C. E. Leiserson, "Fat-trees: universal networks for hardware-efficient supercomputing," IEEE Transactions on Computers, vol. 34, no. 10, pp. 892–901, October 1985.

[4] S. B. Alexander, R. S. Bondurant, D. Byrone, V. W. S. Chan, S. G.Finn, R. Gallager, B. S. Glance, H. A. Haus, P. Humblet, R. Jain, I.P. Kaminow, M. Karol, R. S. Kennedy, A. Kirby, H. Q. Le, A. A.M. Saleh, B. A. Schofield, J. H. Shapiro, N. K. Shankaranarayanan, R.E. Thomas, R. C. Williamson, and R. W. Wilson, "A pre competitive consortium on wide-band all-optical networks," J. Lightwave Technol., vol. 11, no. 5/6, pp. 714–735, May/June 1993.

[5] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart and A. Vahdat, "Jupiter Rising: A Decade of Clos topologies and centralized control in Googles Data center," in Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, pp. 183–197, August 2015, London, UK.

[6] A. Saleh and J. M. Simmons, "All-optical networkingevolution, benefits, challenges, and future vision," Proceedings of the IEEE, vol. 100, no. 5, pp. 1105-1117, March 2012.

[7] G. Wilfong, "Minimizing wavelengths in an all-optical ring network," in Proceedings of International Symposium on Algorithms and Computation, pp. 346–355, December 1996.

[8] Y.-H. Lo, Y. Zhang, W. S. Wong and H.-L. Fu, "The global packing number for an optical network," arXiv: 1509.07029.

[9] B. Beauquier, "All-to-all communication for some wavelength-routed all-optical networks," Networks, vol. 33, pp. 179–187, 1999.

[10] B. Beauquier, J.-C. Bermond, L. Gargano, P. Hell, S. Pérennes and U. Vaccaro, "Graph problems arising from wavelength-routing in all-optical networks," 2nd Workshop on Optics and Computer Science (WOCS), Geneve, Switzerland, April 1997.

[11] L. Gargano, P. Hell and S. Perennes, "Colouring all directed paths in a symmetric tree with applications to WDM routing," in Proceedings of the 24th International Colloquium on Algorithm, Languages and Programming (ICALP 97), vol. LNCS 1256, pp. 505–515.

[12] H. Schröder, O. Sýkora and I. Vrt'o, "Optical all-to-all communication for some product graphs," in Proceedings of the 24th Seminar Current Trends in Theory and Practice of Information, 1997, vol. LNCS 1338, pp. 555–562.

[13] J. Perry, A. Ousterhout, H. Balakrishnan, D. Shah and H. Fugal, "Fastpass: A centralized zero-queue datacenter network," in Proceedings of the 2014 ACM Conference on Special Interest Group on Data Communication, vol. 44, no. 4, pp. 307-318, October 2014, Chicago, Illinois.

[14] V. E. Benes, Mathematical Theory of Connecting Networks and Telephone, Academic Press, New York, 1965.

[15] J. R. Correa and M. X. Goemans, "Improved bounds on nonblocking 3-stage Clos networks," SIAM Journal on Computing, vol. 37 no. 3, pp. 870–894, 2007.

[16] R. Melen and J. S. Turner, "Nonblocking multirate networks," SIAM Journal on Computing, vol. 18, pp. 301-313, 1989.

[17] H. Q. Ngo and V. H. Vu, "Multirate rearrangeable Clos networks and a generalized edge coloring problem on bipartite graphs," SIAM Journal on Computing, vol. 32 no. 4, pp. 1040–1049, 2003.

[18] J. S. Turner and R. Melen, "Multirate Clos networks," IEEE Communication Magazine, vol. 41, no. 10, pp. 38–44, October 2003.

[19] D. B. West, Introduction to graph theory, Prentice Hall, Upper Saddle River, NJ 07458, 2001.

[20] M. Al-Fares, A. Loukissas and A. Vahdat, "A scalable, commodity data center network architecture," in Proceedings of the 2008 ACM Conference on Special Interest Group on Data Communication, pp. 63–74, August 2008, Seattle, WA, USA.

[21] OEIS Foundation Inc. The on-line encyclopedia of integer sequences. http://oeis.org

[22] J. Denes and A. D. Keedwell, Latin squares and their applications, Academic Press, New York, 1974.

[23] I. Chung, "Application of the special latin squares to the parallel routing algorithm on hypercube," Journal of Korean Information Science Society, vol. 19, no. 5, 1992.

[24] S. Kim and I. Chung, "Application of the special Latin square to a parallel routing algorithm on a recursive circulant network," Information Processing Letters, vol. 66, no. 3, pp. 141–147, May 1998.

[25] J.-S. Yang, S.-M. Tang, J.-M. Chang and Y.-L. Wang, "Parallel construction of optimal independent spanning trees on hypercubes," Parallel Computing, vol. 33, no. 1, pp. 73–79, February 2007.

[26] K. Bian, J.-M. Park and R. Chen, "Control channel establishment in cognitive radio networks using channel hopping," IEEE Journal on

Selected Areas in Communications, vol. 29, no. 4, pp. 689–703, April 2011.

Yuan-Hsun Lo received the B.S., M.S., and Ph.D. degrees in applied mathematics from National Chiao Tung University, Hsinchu, Taiwan, in 2004, 2006, and 2010, respectively. He is currently an Assistant Professor with the School of Mathematical Sciences, Xiamen University, Xiamen, China. His research interests include combinatorics, graph theory, and combinatorial designs and their applications.

Yijin Zhang (M'14) received the B.Eng. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004, the M.S. degree from the Southeast University, Nanjing, China, in 2007, and the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong, in 2010, all in Information Engineering. He is now an Associate Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology. His research interests include sequence design and resource allocation in communication networks.

Yi Chen received the B.S. degree in Communication Engineering from Beijing University of Posts and Telecommunications in 2007, and Ph.D. degree in Information Engineering from The Chinese University of Hong Kong in 2012. She is currently a Research Assistant Professor with the School of Science and Engineering, the Chinese University of Hong Kong (Shenzhen). Her research interests include wireless communication, resource allocation and machine learning.

Wing Shing Wong (M'81-SM'90-F'02) received the combined masters and bachelors degrees (summa cum laude) from Yale University, New Haven, CT, USA, in 1976, and the M.S. and Ph.D. degrees from Harvard University, Cambridge, MA, USA, in 1978 and 1980, respectively. After working at AT&T Bell Laboratories for ten years, he joined the Chinese University of Hong Kong, Hong Kong, in 1992, and is now a Professor of information engineering. His research interest includes mobile communication, nonlinear filtering, and networked control. He was the Chairman of the Department of Information Engineering from 1995 to 2002 and the Dean of the Graduate School from 2005 to 2014, and served as the Science Advisor at the Innovation and Technology Commission of the HKSAR Government from 2003 to 2005.

Hung-Lin Fu (M'15) received the B.S. degree in Mathematics from National Taiwan Normal University in 1973 and the Ph.D. degree in mathematics (major in combinatorics) from Auburn University, Auburn, AL, in 1980. Currently, he is a Fellow of Institute of Combinatorics and Its Applications, and a Professor of Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan (since 1988). His research interests are graph theory, combinatorial designs, and their applications.