"Don't Let the Stack Get Stuck": A Novel Approach for Designing Efficient Stackable Routers

Jose Yallouz*, Gideon Blocq†, Yoram Revah‡, Aviran Kadosh§, Ariel Orda¶
Marvell Israel and Technion Israel Institute of Technology
Email: {‡yoram@ §aviran@}.marvell.com and {*jose@tx †gideon@tx ¶ariel@ee}.technion.ac.il

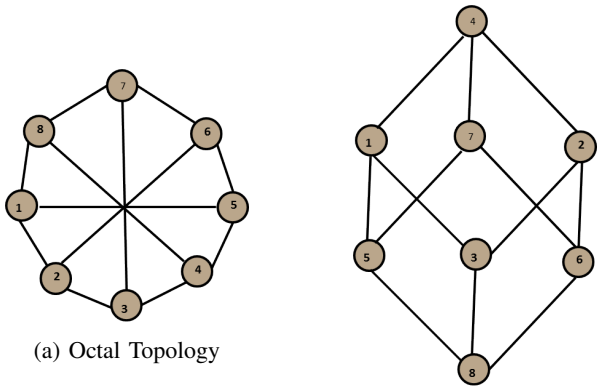
***Abstract*—Stackable Routers, i.e. a class of independent routing units operating together as a single router, constitute an affordable scalable approach for coping with the growing networking requirements of organizations. In this study, we investigate several design problems of stackable routers and develop novel schemes for improving their performance. First, we formalize a mathematical model for optimizing the network topology in terms of throughput and delay, while obeying constraints in the number of ports of each internal routing unit. We then consider the problem of minimizing the diameter of the interconnection topology, as a measure of maximum delay, and establish efficient near-to optimal (explicit) topologies. Furthermore, we also consider the problem of maximizing the throughput of a stackable router. We show its hardness and derive bounds for the optimal solution. While, traditionally, the different routing units of a stackable router are linked together in a ring topology, through simulations we show that a major improvement in the diameter of stackable routers can be accomplished even through the employment of randomly-generated topologies. Finally, we investigate the basic problem of constructing a feasible stackable router and establish some fundamental properties of the required structure of the routing units.**

I. Introduction

A. Background and Motivation

The networking infrastructure of an organization should be able to cope with its increase in networking requirements. While a simple approach is to replace and upgrade the whole networking equipment, a much more compelling alternative is to scale up the existing equipment in order to support higher demands. *Stackable Routers* are a class of standalone network routers configured to operate together so as to establish the characteristics of a single unit yet possessing the capacity of all devices. Hence, stackable routers constitute an affordable scalable approach for providing the growing networking requirements of organizations.

Stackable router architectures have been introduced by leading networking vendors, e.g. Cisco Catalyst 3750 StackWise Technology [1] and Juniper EX4200 Switches [2]. In such architectures, individual switches are connected together and share their routing information in order to create a single switching unit with high capacity. The stack sustains elasticity, i.e. switches can be added to and deleted from a working stack without affecting its performance. Historically, each switch in the stack contained two bidirectional high capacity ports designated for internal traffic. Accordingly, the switches of individual stackable routers have been connected together in a dual-ring topology. This specific stack structure enables a certain level of availability, since for each routing path in the system there exists an alternative path. However, as we will show in this paper, the dual-ring topology tends to significantly hurt the system performance when the amount of necessary switches increases. Recently, various stackable switches have been designed with several (i.e., more than 2) internal traffic ports [3]. Consequently, the topology of a stackable router should be reconsidered when interconnecting a large number of switches. Accordingly, in this work, we study the efficient design of such topologies.



(a) Octal Topology

(b) Hypercube Topology

Fig. 1: Comparison of topology delay for a given set of eight switches with three internal ports.

The performance of a single stackable router unit can be measured by looking at several major metrics, most notably the *diameter*. The diameter is defined as the longest shortest distance between every two switches in the stack, indicating the worst delay that a message incurs inside the router. Here, the distance is represented by the number of links. Clearly, this metric is directly affected by the interconnection of switches in the stack. Fig. 1 represents two possible feasible connections for a given set of eight switches each containing three internal ports. In the hypercube topology, depicted in Fig. 1b, we note that the distance between switch 2 and switch 5 is three links. Moreover, this is the longest shortest distance, hence the diameter in this case is 3 links. However, by connecting the routers in an octal topology as shown in Fig. 1a, we improve the distance between router 2 and router 5 to two links. Since

this is the longest shortest distance in this topology, the diameter is now two links. The above example shows that a judicious connection of the switching elements of a stackable router may result in a considerable improvement in its diameter. Moreover, the example gives rise to some fundamental questions about the efficiency of the connection of the stack elements. Specifically, given a set of switches with a constrained number of internal ports, we aim at finding an interconnection topology that minimizes the diameter of the stackable router. This problem, which we term as the *Port Constrained Optimal Diameter Problem (PCODP)*, is a main subject of this study.

The throughput in the network, which is defined as the total amount of traffic sustained by the system, constitutes another metric of major importance for designing an efficient stackable router. Accordingly, we also investigate a variant of PCODP, in which we aim at optimizing the throughput rather than the diameter. We term this problem as *Port Constrained Optimal Throughput Problem (PCOTP*).

We note that, while we focus on their importance in the context of stackable router design, the PCODP and PCOTP optimization problems are of interest also in the broader context of the design of (general) communication networks.

We mainly focus on the 'homogeneous" case, where the number of internal ports is unique. We then study the general, "heterogeneous" case, where the number of internal ports may vary and we investigate some fundamental connectivity characteristics of this problem.

B. Related Work

The PCODP problem is closely related to another problem, namely the *Degree-Diameter Problem (DDP)*, which has been extensively studied in the literature [4] [5] . The DDP problem aims at finding the largest order of a graph with given maximum degree p and diameter D. An upper bound, which can be achieved by the construction of a simple regular tree, is equal to

$$1 + p \cdot \sum_{i=0}^{D-1} (p-1)^i.$$

The graphs matching this bound were named *Moore graphs* and they have been the subject of much investigation, e.g. [4] . Those studies emphasized the rarity of Moore graphs; indeed, for diameter $D > 2$ and degree $p > 2$, Moore graphs do not exist. Recently, the practical aspects of the DDP and related problems attracted growing interest. Yet, the computational complexity of the DDP problem is still considered an open issue [6] and it is widely believed to be NP-Hard. Otherwise, the problem of finding graphs close to the Moore bound, which has been studied for more than fifty years [5], could be easily solved by a tractable solution to the DDP problem.

While the DDP problem has been widely studied, to the best of our knowledge no attention has been devoted to the related problem that we term as PCODP. Moreover, we provide an important insight to a much more complex, "heterogeneous" problem, where the maximum degree (number of ports) may vary among nodes.

The Port Constrained Optimal Throughput Problem (PCOTP) is related to the class of degree constrained max-flow problems, which were the subject of several recent studies [7] [8]. Specifically, [7] focuses on finding flows that minimize the network congestion where bounds are applied to the degree of the nodes in the network. Most related to PCOTP is [8], which considers a special version of the problem where bounds are applied to the number of paths at each node at the network rather than its degree.

C. Paper Outline

The rest of this paper is organized as follows. In Section II, we provide a formal definition of our model, within which we define the PCODP optimization problem which optimizes the diameter of a stackable router. In Section III, we provide an analytical lower bound on the solution to the PCODP problem. In view of the widely adopted assumption that DDP, hence also PCODP, are intractable problems, we analyze the diameter of several topology constructions, which aim at approximating the optimal solution. In Section IV, considering the throughput perspective, we formalize the PCOTP problem and establish its hardness. Consequently, we formulate a linear program to approximate its calculation and we derive lower and upper bounds on its integrality gap. Furthermore, in Section V, we show, through simulations, that these topologies considerably outperform the diameter of standard ring topology. Moreover, the simulations indicate that even a randomly-constructed topology provides a significant improvement over the ring topology. In Section VI, we consider the heterogeneous case, where each internal switch has a possible different number of ports, and we establish several fundamental properties for connecting a stackable router in several configurations. Due to space limits some proofs and details are omitted from this version and can be found (online) in [9].

II. Model and Problem Formulation

A *stackable router* is composed of a set of interconnected internal switches. Each *internal switch* is represented by a node v. A *connection* between a pair of internal switches u and v is represented by a link e or (u, v). The terms "internal switch" and "node" as well as "connection" and "link" are interchangeably used in this paper. We define a *Switch Set* V as a set of internal switches. Accordingly, a *stackable router* is represented by an undirected graph $G(V, E)$, where V is a switch set and E is the set of connections. A *path* is a finite sequence of nodes $\pi =< v_0, v_1, ..., v_h >$ such that, for $0 \leq n \leq h-1$, $(v_n, v_{n+1}) \in E$. A path is *simple* if all its nodes are distinct. Accordingly, we define the length of a simple path $l(\pi)$ as the number of edges in the path. Given a source node $s \in V$ and a destination node $t \in V$, the *set of all simple paths from s to t* is denoted by $P^{(s,t)}$.

We proceed with several definitions determining the limitation on the number of connections of an internal switch.

Definition 2.1: The *port quantity* of an internal switch v, p_v, is defined as the maximum number of connections that *can possibly be connected* to the internal switch v.

In this study we focus on the typical homogeneous case where the port quantity of each internal switch is equal for all nodes, i.e $\forall v \in V,\ p_v \equiv p$.

Definition 2.2: The *degree* of an internal switch v, deg_v, is defined as the number of connections actually connected to the internal switch v.

Actually, the port quantity of an internal switch constitutes an upper bound on its degree, i.e. $deg_v \leq p_v$, thus defining the maximum degree of the associated internal switch.

As a routing policy, we consider the widely employed shortest path routing. That is, the routing between two nodes in the network is performed through a shortest path between the nodes. In our context, the length of a path is taken to be the number of links along the path, which accounts for the routing delay in the stackable router system. In particular, we consider the worst case delay, which is accounted for by the diameter, defined as follows.

Definition 2.3: Given a stackable router $G(V, E)$, its *diameter* D_G is the number of hops in the longest shortest path between every two nodes $u, v \in V$, i.e. $D_G = \max_{u,v \in V} \min_{\pi \in P^{(u,v)}} l(\pi)$.

The above discussion formalizes the notion of the different routing components of a stackable router unit and its routing policy. For a given switch set, there might be several possibilities for connecting the different internal routing elements, and we are interested in those that provide the smallest diameter. This gives rise to the following optimization problem.

Definition 2.4: ***Port Constrained Optimal Diameter Problem (PCODP)***: Given is a switch set V where each internal switch $v \in V$ is associated with $p_v \in Z^+$. Find a set of connections E for the stackable router $G(V, E)$ such that

$$\min D_G$$

$$s.t.\ deg_v \leq p_v\ \forall v \in V$$

Another metric of major importance is the throughput of the stackable router $G(V, E)$, which is defined as the amount of traffic that is sent through the stackable router. In Section IV, we define an optimization problem, named Port Constrained Optimal Throughput Problem (PCOTP), in which we aim at maximizing the throughput of the stackable router.

In the following sections, we will discuss several aspects of the above optimization problems and provide efficient topological solutions.

III. Constructing the Stackable Router Topology: The Diameter Perpective

In this section, we consider the typical homogeneous case where the port quantity of each internal switch is equal to a value p, i.e $\forall v \in V,\ p_v \equiv p$. First, we consider the complexity of the PCODP problem (2.4) and derive a lower bound for its solution, i.e., optimal diameter. We then continue to present and analyze two near-to optimal topological solutions.

A. Problem Complexity

Recall that the Degree Diameter Problem (DDP) aims at finding the maximum number of nodes in a network with a given port quantity p and diameter k. Both the PCODP problem and the classical DDP problem belong to the same computational complexity class, since these problems can be reduced to the same decision problem as follows:

Definition 3.1: ***PCODP-DDP Decision Problem***:

Given are a port quantity p, a diameter bound k and a switch set size bound n. Does there exist a connected stackable router $G(V, E)$ such that its switch degree $deg_v \leq p\ \forall v \in V$, its diameter $D \leq k$ and its switch size $|V| \geq n$?

Note that the complexity of the DDP problem is still considered an open question [6], however, it is believed to be NP-hard (and consequently so is the PCODP problem).

B. Diameter Lower Bound

We proceed to establish a lower bound to the PCODP problem.

Theorem 3.1: Given is a stackable router $G = (V, E)$, with a switch set size $|V| = n$ and $\forall v \in V,\ p_v \equiv p > 2$. The optimal diameter is lower bounded by:

$$\log_{p-1}(n \cdot \frac{p-2}{p}) \leq D$$

.

Proof: Let v be a switch in $G(V, E)$ and let n_i denote the number of switches at distance $i \in [0, D]$ from v. A switch at distance $i \geq 1$ can be connected to at most $p - 1$ neighbors at distance $i + 1$. Therefore, we have that $n_{i+1} \leq (p - 1) \cdot n_i$ for $i \in [0, D - 1]$. Since $n_1 \leq p$ we get that $n_i \leq p(p - 1)^{i-1}$ for $i \in [0, D]$, which results in the famous Moore bound:

$$n \leq 1+p+p(p-1)+...+p(p-1)^{D-1} = 1+p\sum_{i=0}^{D-1}(p-1)^i =$$

$$1 + p\frac{(p-1)^D - 1}{p-2} \qquad ,p > 2$$

A simple manipulation on the above equation for the case $p > 2$ results in following inequality:

$$\frac{n(p-2)}{p} < \frac{n(p-2)+2}{p} = \frac{(n-1)(p-2)}{p} + 1 \leq (p-1)^D.$$

Finally, from the above, we derive the desired lower bound for the diameter, i.e. $\log_{p-1}(n \cdot \frac{p-2}{p}) \leq D$. ■

The above lower bound provides an instrument for analyzing the quality of different topologies. For this purpose, we define the approximation factor as follows.

Definition 3.2: A feasible solution S to the PCODP problem is a *k-approximation* if its diameter is not greater than a factor k from the optimal solution.

We proceed to propose two explicit near-to optimal constructions, namely the *Tree approach* and the *De Bruijn approach*.

C. Tree approach

A *balanced* tree is a widely implemented structure for providing short distances between its nodes. A $k-ary$ tree is a rooted tree where each node has at most k children. An *internally connected* k-ary tree is a tree whose internal nodes that are connected to a leaf have, each, exactly k children. A *balanced* k-ary tree is an internally connected tree for which all the leaves are at either level h or $h-1$. A *complete* k-ary tree is an internally connected tree for which all the leaves are at level h. k-ary trees have been widely investigated [10] and several interesting properties have been established.

In [9] we provide an explicit construction exploring the special properties of k-ary trees, namely the *Tree Construction Algorithm* (TCA). More specifically, the construction is based on the employment of p different complete balanced $(p-1)$-ary trees connected to a central root. The following theorem establishes the approximation factor of the TCA construction to the PCODP optimal solution.

Theorem 3.2: Given is a switch set V of size, with $|V| = n$ and $p > 2$. The Tree Construction Algorithm (TCA) is a 2-approximation to the PCODP problem.

Proof: See [9]. ∎

We note that the diameter can be improved by also connecting the leaves of the tree.

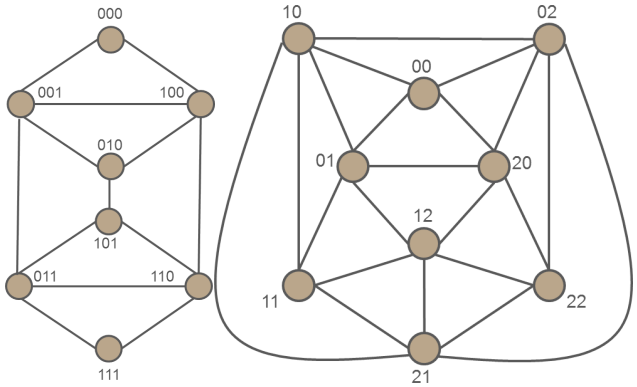
D. De Bruijn Approach

De Bruijn graphs are a well-known class of graphs with various applications in different areas, e.g. fault tolerant networks, grid networks and bioinformatics [11] [12]. A De Bruijn graph is composed of k^m nodes labeled by m-tuples over a k-character alphabet, i.e. all possible sequences of length m of the characters of the given alphabet. The edge between two nodes is set for every combination representing a left-shift operation on the node's label. Specifically, they are defined to be pairs of the form $((c_1...c_m),(c_2...c_{m+1}))$ where c_{m+1} is any character in the alphabet. Indeed, these graphs provide a promising network topology with a relatively small diameter. Moreover, each node in the graph will have a small degree. These properties turn De Bruijn graphs into a suitable topology for connecting a stackable router. Formally, a De Bruijn Graph is defined as follows:

Definition 3.3: Given are an alphabet $A = \{a_1, ..., a_k\}$ of k characters and a set of nodes $V = \{(c_1...c_m) : c_i \in A\}$ of size k^m, each labeled by a string of length m. A De Bruijn graph $DB(k,m)$ is a directed graph $G(V,E)$ whose $E = \{((c_1, ..., c_m),(c_2, ..., x)) : x \in A\}$.

Definition 3.4: Given a De Bruijn graph $DB(k,m)$, an undirected De Bruijn graph $UDB(k,m)$ is a modified version of $DB(k,m)$ that satisfies the following adjustments:

1) All self loop links are discarded.
2) If $u \rightarrow v$ is a directed link in $DB(k,m)$, then the undirected link (u,v) is a link in $UDB(k,m)$.
3) If both $u \rightarrow v$ and $v \rightarrow u$ are directed links in $DB(k,m)$, then there only exists one undirected link (u,v) in $UDB(k,m)$.



(a) $UDB(2,3)$ (b) $UDB(3,2)$

Fig. 2: Examples of undirected DeBruijn Graph

Fig. 2 shows two examples of undirected De Bruijn graphs, namely $UDB(2,3)$ and $UDB(3,2)$, where the labeled strings are attached to each node. Note that for $UDB(2,3)$ the maximum degree of any node is 3 and its diameter is 3. For $UDB(3,2)$, the maximum degree is 5 and its diameter is 2. We proceed to establish some important properties of an undirected De Bruijn graph $UDB(k,m)$.

Property 3.1: The diameter of an undirected De Bruijn graph $UDB(k,m)$ is upper bounded by m.

Proof: Assume two arbitrary nodes labeled by $v = (v_1, ..., v_m)$ and $u = (u_1, ..., u_m)$. There is a path between v and u following the sequence of nodes $\pi =< v, x_1, ...x_{m-2}, u >$ where $x_i = (v_i, ..., v_m, u_1, ..., u_{m-i})$ and its length is m. ∎

Property 3.2: The degree of each node in an undirected De Bruijn graph $UDB(k,m)$ is upper bounded by $2k$.

Proof: Assume a k-character alphabet $A = \{a_1, ..., a_k\}$ and a set of edges $E = \{((c_1, ..., c_m),(c_2, ..., x)) : x \in A\}$. There are k possibilities for selecting the x character, i.e. there are at most k outgoing links from node $v = (c_1, ..., c_m)$. Moreover, there are another k possibilities for nodes labeled $(c_1, ..., c_{m-1}, x)$, i.e. there are at most k incoming links to node $v = (c_1, ..., c_m)$. Since we consider the undirected version of De Bruijn graphs the degree of each node is upper-bounded by $2k$. ∎

The following theorem shows that an undirected De Bruijn graph $UDB(k,m)$ of size $n = k^m$, gives an approximation to the PCODP.

Theorem 3.3: Given is a switch set V of size $n = k^m$ composed of switches with an unique port quantity $p > 2$. A De Bruijn undirected graph $UDB(k,m)$ is a $\log_{\frac{p}{2}}(p-1) + \log_n(\frac{p}{p-2})$-approximation to the PCODP problem.

Proof: The number of nodes in the switch set is $n = k^m$. From Properties 3.1 and 3.2, we conclude that the number of ports is $p \leq 2k$ and the diameter D is bounded by m. Therefore, we have that $n \geq (\frac{p}{2})^D$ and by a simple manipulation we get that $D \leq \frac{\log(n)}{\log(\frac{p}{2})}$. According to Theorem 3.1, we have that $\log_{p-1}(n) \leq D + \log_{p-1}(\frac{p}{p-2})$. Accordingly,

the approximation factor can be calculated as the ratio between $\log_{\frac{p}{2}}(n) + \log_{p-1}(\frac{p}{p-2})$ and $\log_{p-1}(n)$ resulting in the desired approximation factor. ∎

Interestingly, from Theorem 3.3 we can conclude that the approximation ratio converges to 1 with the increase in the number of ports p. Moreover, the number of nodes n almost does not affect the approximation ratio, since the expression $\log_n(\frac{p}{p-2})$ rapidly converges to zero with the increase in the number of ports p and number of nodes n.

In [9] we show that the approximation ratio of the well known hypercube topology is $\log_2(p-1)$. However, in contrast to the De Bruijn topology, the hypercube approximation ratio grows with the increase of p.

Although we established several explicit constructions that provide near to optimal solutions to the PCODP, these topologies also restrict the selection of the switch set size. Specifically, the hypercube and the De Bruijn topologies allow only $n = 2^m$ and $n = k^m$ sizes, respectively. Therefore in Section V, through simulations, we further analyze the diameter of random topology of any size n.

IV. Throughput perspective

In this section we aim at finding the topology that maximizes the network's throughput. In doing so, we first somewhat amend our model. In particular, we are given a directed[1] graph $G = (V, E)$ (stackable router) with link capacities $c_{u,v} \equiv c,\ \forall (u,v) \in E$. We consider a set of commodities K, each commodity $i \in K$ being a tuple $i = (s_i, t_i, d_{s_i,t_i})$. Specifically, a commodity K_i represents a demand d_{s_i,t_i} that node s_i wishes to send to node t_i. Furthermore, each node v has a constraint on its port quantity, $p_v \equiv p,\ \forall v \in V$. The flow of commodity i on link (u,v) is denoted by $f^i_{u,v}$ and, consequently, the flow on each link is denoted by $f_{u,v} = \sum_i f^i_{u,v}$. A flow vector $\mathbf{f} = [f_{u,v}]_{\forall (u,v) \in E}$ is considered *feasible* if it abides by the following constraints:

Definition 4.1: ***Feasible Flow Constraints***

- **Demand constraint:**, $\sum_{v \in V} f^i_{s_i,v} \leq d_{s_i,t_i}\ \forall i \in K$
- **Capacity constraint:** $f_{u,v} \leq c,\ \forall (u,v) \in E$
- **Port quantity constraint:** $|\{v | f_{u,v} > 0\}| \leq p,\ \forall u \in V$
- **Kirchoff constraint:**, $\forall u \in V,\ \forall i \in K$,

$$\sum_{v \in V} f^i_{u,v} - \sum_{w \in V} f^i_{w,u} = \begin{cases} \sum_{v \in V} f^i_{s_i,v} & \text{if } u = s_i \\ -\sum_{v \in V} f^i_{s_i,v} & \text{if } u = t_i \\ 0 & \text{otherwise} \end{cases}$$

The *throughput* of the stackable router is given by $\sum_{i \in K} \sum_{v \in V} f^i_{s_i,v}$. Given a set of commodities, our goal is to find a set of links $\bar{E} \subseteq E$ in the stackable router, such that the total throughput in the system is maximized. To accomplish this goal we consider $G(V, E)$ to correspond to a *complete graph* and aim to maximize its throughput. The motivation behind a complete graph comes from the practical representation of the stackable router. In practice, each switch can possibly be connected to any of the internal switches of the switch set V. Nevertheless, only a finite amount of connections can be established per switch, namely p. We then solve the Port Constrained Optimal Throughput Problem (PCOTP) and consider the subset of links $\bar{E} \subseteq E$ with positive flow, i.e. $\bar{E} = \{(u,v) \in E | f_{u,v} > 0\}$. The set $\bar{E}$ represents the connections between the internal switches of the stackable router, which maximize the throughput while satisfying the feasibility constraints. Note that in the definition and lower bound analysis of PCOTP, we describe a more general problem which we call the Bounded Node-Degree Max Flow Problem (B-MFP). There we do not assume that $G(V, E)$ corresponds to a complete graph. Nevertheless, when proposing an upper bound on the maximum throughput, we do restrict $G(V, E)$ to a complete graph.

Definition 4.2: **Bounded Node-Degree Max Flow Problem (B-MFP)** Given is a graph $G = (V, E)$. Find the maximum feasible flow $\sum_{i \in K} \sum_{v \in V} f^i_{s_i,v}$.

Definition 4.3: **Port Constrained Optimal Throughput Problem (PCOTP)** Given is a complete graph $G = (V, E)$. Find the maximum feasible flow $\sum_{i \in K} \sum_{v \in V} f^i_{s_i,v}$ and return the set $\bar{E} = \{(u,v) \in E | f_{u,v} > 0\}$.

Obviously PCOTP is a specific instance of B-MFP. Thus any bounds that we find on B-MFP automatically apply to PCOTP.

Theorem 4.1: The Bounded Node-Degree Max Flow Problem is NP-hard.

Proof: We prove B-MFP's hardness through a reduction to the NP-hard *disjoint paths* problem [13]. The *disjoint paths* problem determines if a set of source-destination pairs in a general graph can be connected by a set of disjoint paths. Consider a graph, in which all edges have unit capacity, all K source-destination pairs have unit demand and each node has unit port quantity. Consider a node $v \in V$. Since $p_v = 1$, only a single edge e originating in v can carry a positive flow. Moreover, since $c_e = 1$, there can only be a single edge that terminates in v and carries positive flow. Thus, there does not exist more than a single path that carries flow from a source s_i to a destination t_i.

First, consider the case where all source-destination pairs can be connected by disjoint paths. Since B-MFP returns the maximum flow and all pairs have unit demand, in this case B-MFP will return a value of K.

On the other hand, if B-MFP returns a value of K, it is clear from our observation above that this must consist out of disjoint paths. Then solving B-MFP will also solve the disjoint paths problem. ∎

In [9], B-MFP is defined as an Integer Program. In order to find an approximation of the maximum feasible flow, we formulate the following linear program by relaxing the *port quantity constraint*:

Definition 4.4: ***LP-throughput***:

$$\text{maximize} \sum_{i \in K} \sum_{v \in V} f^i_{s_i,v}$$

$$\text{subject to} \sum_{v \in V} \frac{f_{u,v}}{c} \leq p \qquad , \forall u \in V \tag{1}$$

$$\text{Demand constraint} \, , \forall i \in K \tag{2}$$

[1]Here, each link represents an unidirectional connection.

$$\text{Capacity constraint, } \forall (u,v) \in E \tag{3}$$

$$\text{Kirchoff constraint, } \forall u \in V,\ \forall i. \tag{4}$$

In the LP-throughput description, the demand, capacity and Kirchoff constraints are as stated above.

We proceed to investigate the integrality gap between the LP-throughput linear program and the integer program representing B-MFP. First, it directly follows from [8] that we achieve a similar lower bound of $\Omega(\sqrt{|V|})$. (This is further elaborated in [9].) Accordingly, we proceed to find an upper bound for the integrality gap. In doing so, we focus on PCOTP and consider a complete graph (clique). Furthermore, in order to analyze PCOTP, we add two additional assumptions on the demands of the switches:

A1: $\sum_{v \in V} d_{s_i,v} \leq p \cdot c,$ for $i = 1 \dots K$.
A2: $d_{s_i,t_i} \leq c,$ for $i = 1 \dots K$.

Assumption A1 upper-bounds the maximum demand of each node by its maximal feasible capacity, while Assumption A2 provides an upper bound on the demand on each source-destination pair. Applying these two reasonable assumptions we proceed to provide an upper bound on the integrality gap.

Theorem 4.2: Given is a complete graph $G = (V, E)$ together with Assumptions A1, A2. The integrality gap of PCOTP is upper bounded by $\frac{|V|-1}{p}$.

Proof: We construct a solution for LP-throughput in which for each commodity $i = (s_i, t_i, d_{s_i,t_i})$ we send the entire demand d_{s_i,t_i} through the link $(s_i, t_i) \in E$, i.e. $f^i_{s_i,t_i} = d_{s_i,t_i}$. We call this approach the *one hop solution*. From Assumptions A1 and A2, it follows that this abides by all the constraints of LP-throughput. Moreover, when summing over all the nodes, it *de facto* brings about the maximum possible flow in the system. Therefore, it must be equal to the solution of our LP:

$$OPT_{LP} = \max \sum_{i \in K} \sum_{v \in V} f^i_{s_i,v} = \sum_{u \in V} \sum_{v \in V} d_{u,v}. \tag{5}$$

Although the one hop solution abides by the LP's constraints, its maximum flow might not be feasible, due to the relaxed port quantity constraint. Thus, to make our one hop solution feasible, per node, we pick the p outgoing links with the largest amount of flow. In other words: per node u, we connect to a set of nodes $P_u \subset V$, where $|P_u| = p$ and $\sum_{(u,v) \in P_u} d_{u,v} \geq \sum_{(u,v) \in \bar{P}_u} d_{u,v}$, for any set $\bar{P}_u \subset V$ such that $|\bar{P}_u| = p$. Thus, the total throughput in the graph decreases to $\sum_{u \in V} \sum_{v \in P_u} d_{u,v}$. It follows that we improve over the averaged case, i.e.,

$$\sum_{v \in P_u} d_{u,v} \geq \sum_{v \in V} d_{u,v} \cdot \frac{p}{|V|-1},\ \forall u \in V. \tag{6}$$

From (6) and by summing over all nodes, it follows that

$$\sum_{u \in V} \sum_{v \in P_u} d_{u,v} \geq \sum_{u \in V} \sum_{v \in V} d_{u,v} \cdot \frac{p}{|V|-1}. \tag{7}$$

Denote the optimal solution of PCOTP as OPT, from (5) and (7) we get that

$$\frac{OPT_{LP}}{OPT} \leq \frac{\sum_{u \in V} \sum_{v \in V} d_{u,v}}{\frac{p}{|V|-1} \cdot \sum_{u \in V} \sum_{v \in V} d_{u,v}} = \frac{|V|-1}{p}. \tag{8}$$

■

As a result of Theorem 4.2, our proposed LP-throughput solution is a promising approximation of PCOTP. We can solve the LP-throughput and return the set of links $(u, v) \in E$ such that $f_{u,v} > 0$, thus efficiently connecting the stackable router.

V. Simulation Study

Through comprehensive simulations, we will compare between the diameter of different random constructed topologies and the diameter of several explicit topologies. We will show that the simulated random graphs result in solutions that are close to optimal by a factor of approximately 1.5. Somewhat surprisingly, this approximation ratio is quite close to that of the constructions presented in Section III. Yet, we point out that the approximation ratios obtained in Section III are *proven guarantees*. Additionally, we will demonstrate that the widely employed ring topology exhibits significantly lower performance.

A. Setup

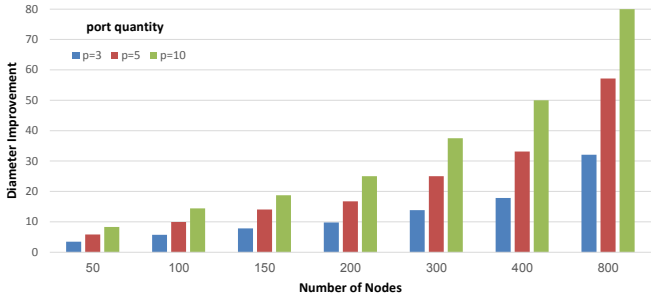
We generated a class of random regular network topologies in several configurations characterized by the number of nodes n and the port quantity p of each internal switch. Specifically, for each combination of the parameters n and p, we generated $1,000$ randomly connected regular network topologies.

For a given set V of n nodes with a homogeneous port quantity value p, we connect all the available ports randomly yet creating a connected component. For this purpose, we first created a connected component connecting all the original nodes of V together. Then, we connected the remaining unconnected ports randomly until all the ports are connected. (See [9] for details.)

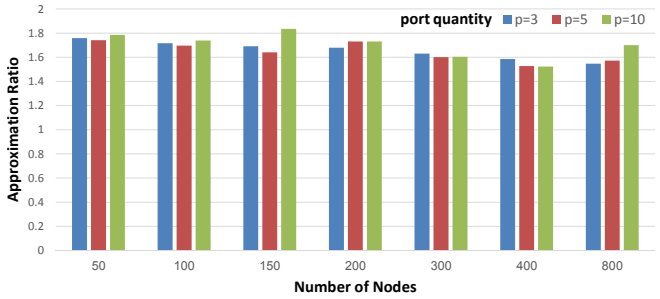
For each of the above randomly-generated topologies, we run the Breadth First Search (BFS) Algorithm in order to find the minimal distance between each pair of nodes in the graph. Accordingly, the diameter D of the graph is calculated as the maximum distance among these. Then, for each configuration where the number of nodes is $n \in [50, 100, 150, 200, 300, 400, 800]$ and the port quantity is $p \in [3, 5, 10]$, we calculate the average diameter $\bar{D}(n, p)$ for $1,000$ randomly generated (connected) topologies. Furthermore, we calculated the diameter of the traditional ring topology $R(n)$ and the value of the diameter's lower bound $LB(n, p)$ according to Theorem 3.1. Finally, we derived the diameter improvement ratio, which is defined as $\bar{\sigma}(n, p) \triangleq \frac{R(n)}{\bar{D}(n,p)}$ and the approximation ratio, which is defined as $\bar{\alpha}(n, p) \triangleq \frac{\bar{D}(n,p)}{LB(n,p)}$.

B. Results

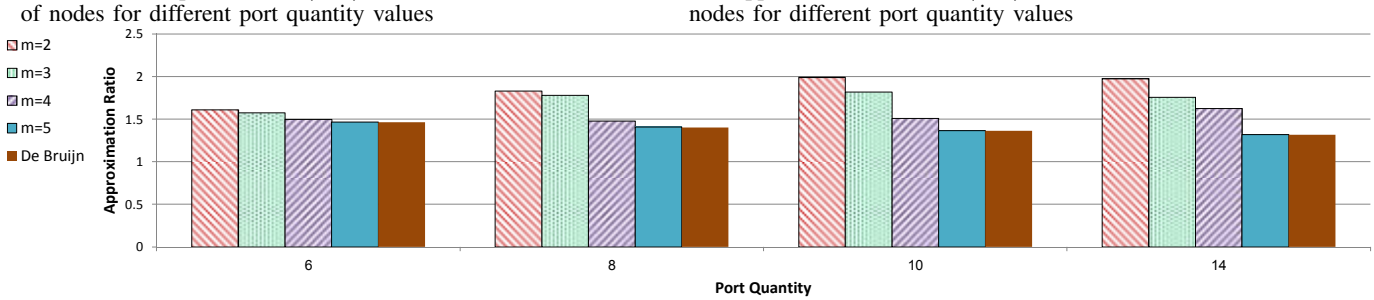
The simulation results are illustrated in Fig. 3. First, the graph depicted in Fig. 3a presents the diameter improvement ratio $\bar{\sigma}(n, p)$ as a function of the number of nodes $n \in [50, 100, 150, 200, 300, 400, 800]$ for different port quantity values $p \in [3, 5, 10]$. From the graph, we can observe a major improvement, as exhibited by the diameter improvement ratio for all simulated configurations. Moreover, this improvement tends to increase with the growth of the number of nodes in the graph.



(a) Diameter Improvement $\bar{\sigma}(n,p)$ as a function of the number of nodes for different port quantity values



(b) Approximation Ratio $\bar{\alpha}(n,p)$ as a function of the number of nodes for different port quantity values



(c) Approximation Ratio $\bar{\alpha}(n,p)$ as a function of the port quantity for different De Bruijn topologies

Fig. 3: Simulation Results

Most notably, these results indicate the inefficiency of the traditional, widely employed, ring topology, even when compared with randomly-generated topologies. Next, the graph depicted in Fig. 3b shows the approximation ratio $\bar{\alpha}(n,p)$ as a function of the number of nodes $n \in [50, 100, 150, 200, 300, 400, 800]$ for different port quantity values $p \in [3, 5, 10]$. The graph exhibits an approximation ratio in the range of $[1.5 - 1.83]$ for all simulated configurations.

Finally, we compare the approximation ratio of the simulated random topology with the ratio of the De Bruijn suggested explicit topology. As mentioned, the switch set size n of a De Bruijn topology take a value of the form $n = k^m$. Moreover, according to property 3.2, the degree of a De Bruijn graph of size $n = k^m$ is upper-bounded by $2k$. In order to incorporate the switch set size restriction of a De Bruijn topology, for different values of port quantity $p \in [6, 8, 10, 14]$ and a value $m \in [2, 3, 4, 5]$, we simulated random topologies of size $\hat{n} = (\frac{p}{2})^m$ and calculated its approximation ratio $\bar{\alpha}(\hat{n}, p)$. Accordingly, we calculated average value of the respective approximation ratio of a De Bruijn graph with a port quantity of $p \in [6, 8, 10, 14]$. Fig. 3c shows the approximation ratio of the different simulated configurations. From the graph, we conclude that the approximation ratio of random topologies converges to the approximation ratio of the explicit De Bruijn topology with the increase in the parameter m. Moreover, as expected, the approximation ratio of a De Bruijn Graph improves with the increase on the number of ports p.

We conclude by noting that the above findings support the claim in [14], namely that random topologies are efficient for data centers in terms of delay.

VI. On the Correct Structure of Stackable Routers

Several cost factors should be considered when selecting the components for designing a stackable router. In particular, the growth in the port quantity of an internal switch can significantly increase the cost of the component, consequently affecting the total price of the stackable router. Therefore, for the heterogeneous case where the port quantity p_v may vary among nodes, it is important to understand some fundamental properties in order to design internal switches with a proper port quantity. Namely, for each node we aim to find a minimum value that is just enough to sustain the construction of stackable routers with affordable connectivity. Accordingly, in this section we establish several conditions for avoiding unnecessary port redundancy.

We begin by providing conditions for the employment of a fully connected stackable router. Note that for a given switch set V, it is not always possible to establish a fully connected stackable router. For instance, assume V consist of 3 switches with port quantity of 1. Clearly, it is impossible to connect all three switches into a single stackable router. Accordingly, we provide necessary and sufficient conditions for establishing a connected graph out of the (entire) given set of internal switches. For this purpose, we split the switch set into two subsets V_1 and $V_{p\geq 2}$ such that $V = V_1 \bigcup V_{p\geq 2}$. Accordingly, we define $V_{p\geq 2} \subseteq V$ as the subset of switches whose port quantity is greater or equal to 2, i.e. $\{v|p_v \geq 2\}$ and $V_1 \subseteq V$ as the subset of switches whose port quantity is equal to 1, i.e. $\{v|p_v = 1\}$.

Theorem 6.1: Given are a switch set V composed of V_1 and $V_{p\geq 2}$. A necessary and sufficient condition for constructing a

connected stackable router $G(V,E)$ composed of all internal switches v in the switch set V is:

$$\sum_{v \in V_{p\geq 2}} p_v - 2 \cdot |V_{p\geq 2}| + 2 \geq |V_1| \tag{9}$$

Proof: The minimal structure in which the switch set V can be connected is a tree. The latter consists of $|V|-1$ edges and, consequently, the sum of its nodes' degrees is equal to $2\cdot(|V|-1)$. Thus, due to the minimality of the tree, a necessary and sufficient condition for connectivity is:

$$\sum_{v \in V} p_v \geq 2(|V|-1) \tag{10}$$

$$\sum_{v \in V_{p\geq 2}} p_v + \sum_{v \in V_1} p_v \geq 2(|V_{p\geq 2}| + |V_1| - 1) \tag{11}$$

$$\sum_{v \in V_{p\geq 2}} p_v - 2 \cdot |V_{p\geq 2}| + 2 \geq |V_1| \tag{12}$$

We note that expression (12) follows from the equality $\sum_{v\in V_1} p_v = |V_1|$. ∎

In the following discussion, we assume that a stackable router $G(V,E)$ does not contain any self loops. For a given switch set V, it is not always possible to connect all the ports of the internal routing components of the stackable router. For instance, assume that V consist of two switches with port quantities of 2 and 4. Although V is a Connectable Switch Set Instance, it is impossible to connect all ports, thus resulting in wasted resources. We proceed to provide necessary and sufficient conditions for connecting all the available ports of a stackable router.

Theorem 6.2: Given a switch set V the following conditions are necessary and sufficient for the connectivity of *all ports* of the stackable router $G(V,E)$:

1) $\sum_{v\in V} p_v$ is even.
2) $2 \cdot \max_{v\in V} p_v \leq \sum_{v\in V} p_v$.
3) $\sum_{v\in V_{p\geq 2}} p_v - 2 \cdot |V_{p\geq 2}| + 2 \geq |V_1|$.

Proof: We begin by proving that the conditions are necessary. We recall that p_v and deg_v represent the port quantity and the degree of a switch v. Since all ports of the stackable router $G(V,E)$ are connected, we have that $\sum_{v\in V} p_v = \sum_{v\in V} deg_v = 2 \cdot |E|$. The second equality follows from the fact that each edge in the graph $G(V,E)$ contributes exactly 2 to the summation. We proceed to show that $2 \cdot \max_{v\in V} p_v \leq \sum_{v\in V} p_v$. By subtracting $\max_{v\in V} p_v$ from both sides of the equation we shall prove that $\max_{v\in V} p_v \leq \sum_{v\in V\setminus(max_{v\in V})} p_v$. Now, assume by contradiction that $\max_{v\in V} p_v > \sum_{v\in V\setminus(max_{v\in V})} p_v$, i.e. the switch with maximum port quantity, $\arg\max_{v\in V} p_v$, cannot be connected to the rest of the switches, thus contradicting the assumption that all ports are connected. Moreover, the last condition straightforward from Theorem 6.1.

In [9] we prove that the conditions are sufficient by establishing an explicit construction of a connected stackable router $G(V,E)$ whose all of its ports are connected. ∎

Theorem 6.2 indicates that it is possible to connect all ports of the stackable router, if it contains more than 2 internal switches with maximum port quantity.

VII. Conclusions and Further Work

We investigated several design considerations in the construction of stackable routers. Accordingly, we formalized two problems, namely PCODP and PCOTP, which consider the optimization of the interconnection topology in terms of throughput and delay, while obeying constraints in terms of the number of ports of each internal routing unit. For PCODP, we provided a lower-bound on the diameter and established efficient near-to optimal (explicit) topologies. Furthermore, we formulated PCOTP as an Integer Program and provided an approximated solution. Through simulations we showed that a *major improvement* in the diameter of a stackable router system can be accomplished by departing from the traditional ring topology. Moreover, this can be achieved even through randomly-generated topologies. Finally, we devote some attention to the connectivity characteristics of the complex "heterogeneous" problem, where the number of ports may vary among the different routing units.

We believe that this work establishes an initial approach towards the construction of efficient stackable routers.

Acknowledgments: This research was supported by the Israeli Ministry of Science and Technology and Marvell Israel. Gideon Blocq is supported by the Google Europe Fellowship in Computer Networking. Jose Yallouz was in Marvell Israel during summer internship program.

References

[1] "Cisco stackwise and stackwise plus technology," White Paper, Cisco, 2010.
[2] "Juniper ex4200-24f datasheet," White Paper, Juniper, 2013.
[3] "Plexxi switch 2 datasheet," White Paper, Plexxi, 2013.
[4] M. Miller and J. Širán, "Moore graphs and beyond: A survey of the degree/diameter problem," *E. J. of Combinatorics*, vol. 61, pp. 1–63, 2005.
[5] E. Loz, H. Pérez-Rosés, and G. Pineda-Villavicencio, "Combinatorics wiki," http://combinatoricswiki.org/wiki.
[6] A. Dekker, H. Rosés, G. Villavicencio, and P. Watters, "The maximum degree & diameter-bounded subgraph and its applications," *J. Mathematical Modelling and Algorithms*, vol. 11, no. 3, pp. 249–268, 2012.
[7] P. Donovan, B. Shepherd, A. Vetta, and G. Wilfong, "Degree-constrained network flows," in *ACM STOC*, 2007.
[8] R. Cohen, L. Lewin-Eytan, J. S. Naor, and D. Raz, "On the effect of forwarding table size on sdn network utilization," in *IEEE Infocom*, 2014.
[9] ""Don't let the stack get stuck": A novel approach for designing efficient stackable routers," Tech. Rep., 2014. [Online]. Available: http://tx.technion.ac.il/%7ejose/TRStackableRouters.pdf
[10] R. Diestel, *Graph Theory, 4th Edition*, ser. Graduate texts in mathematics. Springer, 2012, vol. 173.
[11] R. Chikhi, A. Limasset, S. Jackman, J. T. Simpson, and P. Medvedev, "On the representation of de bruijn graphs," in *Research in Computational Molecular Biology*. Springer, 2014, pp. 35–55.
[12] M. Sridhar and C. Raghavendra, "Fault-tolerant networks based on the de bruijn graph," *IEEE Trans. Computers*, vol. 40, no. 10, pp. 1167–1174, 1991.
[13] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1990.
[14] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking data centers randomly." in *NSDI*, 2012.