Modelling Standard and Randomized Slimmed Folded Clos Networks

postprint version

Camarero, C., Corral, J., Martínez, C., & Beivide, R. (2020, August). Modelling Standard and Randomized Slimmed Folded Clos Networks. In European Conference on Parallel Processing (pp. 185-199). Springer, Cham.

https://link.springer.com/chapter/10.1007/978-3-030-57675-2_12

https://doi.org/10.1007/978-3-030-57675-2_12

Cristóbal Camarero[1], Javier Corral[1], Carmen Martínez[1], and Ramón Beivide[12]
[1]Computer Science and Electronics Department, University of Cantabria
[2]Barcelona Supercomputing Center

Abstract

Fat-trees (FTs) are widely known topologies that, among other advantages, provide full bisection bandwidth. However, many implementations of FTs are made slimmed to cheapen the infrastructure, since most applications do not make use of this full bisection bandwidth. In this paper Extended Generalized Random Folded Clos (XGRFC) interconnection networks are introduced as cost-efficient alternatives to Extended Generalized Fat Trees (XGFT), which is a widely used topological description for slimmed FTs. This is proved both by obtaining a theoretical model of the performance and evaluating it using simulation. Among the results, it is shown that a XGRFC is able to connect 20k servers with 27% less routers than the corresponding XGFT and still providing the same performance under uniform traffic.

Folded Clos, Extended Generalized Fat-tree, Random topologies.

1 Introduction

Nowadays, high-end supercomputers and datacenters are becoming extremely big, connecting hundreds of thousands servers. In consequence, the interconnection networks employed in these systems are becoming more costly and important. With such large sizes, the network cost can be a significant fraction of the total system cost. Deployment network cost includes NICs, routers and wires. The cost of large networks tends to be dominated by the cost of the required wires, but for raw comparisons, the number of network routers can be employed, as the number of wires linearly depends on it.

Fat-trees [1] (FTs), a popular instance of the folded Clos network [5], have been utilized in many high-end systems. The use of FTs entails important benefits. In theory, they can manage any admissible traffic at full rate; they are equipped with a very simple deadlock free routing; they are robust; and, their partitioning is easier than in other networks. Nevertheless, the high cost of FTs becomes prohibitive in very large deployments. In addition, depending of the application, the nature of its communications can be quite different. There is an important class of classic number-crunching applications showing a high degree of communication locality for which FTs reveal overprovisioned [8, 9]. Having in mind these applications, and in order to reduce the high cost of FTs, many deployed big systems have used different slimmed versions of them. Most of slimmed fat-trees can be studied under the model of Extended Generalized Fat Trees (XGFT) introduced in [11]. But nowadays, there is another important set of applications as those coming from big data and analytics that employ global communications and really require the capacity and redundancy of a FT; many of them require all-to-all (uniform) traffic. Moreover, other HPC applications, such as spectral codes perform a 3D Fast Fourier Transform, utilizing large all-to-all communications [12].

In [3] and [2] randomized versions of folded Clos networks were introduced. These topologies are more scalable, allow for easy upgrades of the system and entail less cost. Therefore, it has sense to consider the slimmed variants of such topologies, which are introduced in this paper and denoted as Extended Generalized Random Folded Clos

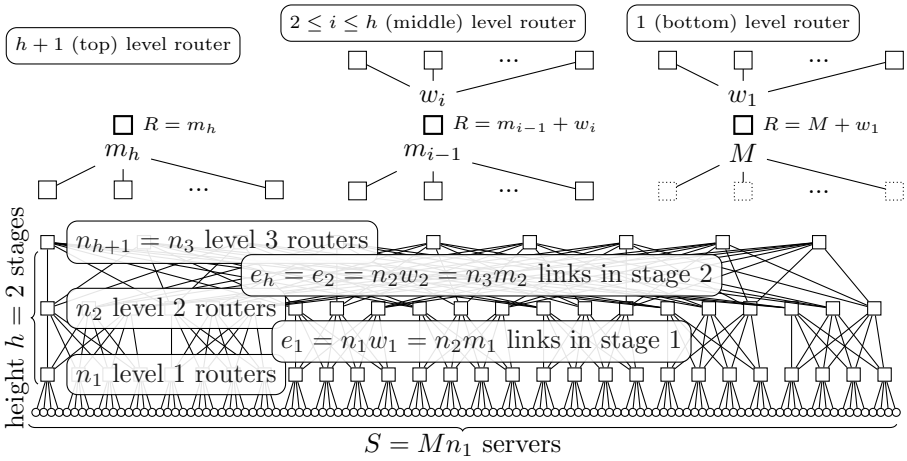

Figure 1: Graphical representation of the notation using XGFT$(2; 4, 7; 3, 3)$.

(XGRFC). We compare XGRFCs and XGFTs both in their topological merits and performance. We will show that XGRFCs inherit the scalability among other properties of random folded Clos, thus providing cost gains respect to slimmed fat-trees. Corresponding performance, we will firstly make a theoretical model that relates the communication pattern of the application, the fitness ratio of the slimmed topology and the performance. For illustrating it, we select a synthetic traffic pattern as to resemble applications needing global communications. This approach is validated using experimental simulation. As it will be shown, just randomizing the stages of slimmed FTs provides similar performance (throughput and fairness) but at smaller cost. In fact, there are outstanding cases such as uniform traffic, in which XGRFCs provide 38% more throughput than its XGFT counterpart.

This paper is organized as follows. In Section 2 folded Clos interconnection networks are summarized and Extended Generalized Fat-trees are revisited. In Section 3 Extended Generalized Random Folded Clos are introduced. In Section 4 a wide experimentation is presented to prove our results. Finally, in Section 5 the main achievements of the paper are summarized.

2 Folded Clos Networks

Folded Clos interconnection networks are widely considered for datacenters [5]. These interconnection networks are indirect, since there are two different kind of routers: those which are connected to servers, and the ones that are only connected to other routers. The routers are arranged into *levels* such that the links that join two different levels constitute a *stage*. Typically, the first level or Level 1 is the one that contains the routers directly connected to servers, known as *leaf routers*. We will consider that in the last level the *spine routers* have all their links in the last stage, that is, the network is *folded*. If an indirect network has l levels of routers it is said to have *height* $h = l - 1$. A common example of folded Clos networks are FTs [1]. In Table 1 the notation used along the paper is summarized and a example illustrating it is graphically represented in Figure 1.

Table 1: Notation

Symbol	Meaning
R	Router radix.
M	Servers per leaf router.
S	Total number of servers.
γ	Average injection rate per server.
n_i	Number of routers at level i.
m_i	Number of links from level $i+1$ to i in each router.
w_i	Number of links from level i to $i+1$ in each router.
e_i	Total number of links connecting routers of levels i and $i+1$.

Folded Clos networks are typically considered being *up/down connected* that is, for every pair of leaves, there is a path beginning with some up-links followed by the same number of down-links. Then, a simple deadlock-free routing can be made following these paths, which is one of the main advantages of Clos networks over other kind of networks. All networks considered in the paper are up/down connected (with very high probability when probabilistic).

Let us denote by w_i the number of links from each router in level i to routers in level $i+1$ and m_i the number of links from each router in level $i+1$ to routers in level i. Let us denote by n_i the number of routers in level i. Then, the number of links e_i that constitute stage i can be calculated as

$$e_i = n_i w_i = n_{i+1} m_i, 1 \leq i \leq h. \tag{1}$$

It follows that any n_i can be calculated from n_1:

$$\frac{n_{i+1}}{n_1} = \prod_{k=1}^{i} \frac{w_k}{m_k}. \tag{2}$$

Our study will be restricted to those networks that are built with identical routers, that is, the following *Regularity Equations* 3 are fulfilled:

$$\begin{aligned} &R = m_i + w_{i+1},\ 1 \leq i \leq h-1, \\ &R = m_h, \\ &R = w_1 + M, \end{aligned} \tag{3}$$

where M denotes the number of servers per leaf router.

A further assumption that can be made is to have, in all non-top levels, the same ratio of up-links. Note that the top level does not have up-links, so it cannot be included. Although from a theoretical point of view this *Constant Radix Ratio Property* seems very natural, it is not necessarily the best choice. In fact, later we consider some examples that fulfill it and others that do not. This assumption is formally stated as

$$\frac{m_i}{w_{i+1}} = \frac{M}{w_1},\ 1 \leq i \leq h-1. \tag{4}$$

This ratio is called *fitness ratio* in [8] and *contention factor* in [7], making both references the constant ratio assumption. Another notation for the same concept is *blocking ratio* in [13]. This assumption can be rewritten using the Regularity Equations 3 as

$$\begin{aligned} &m_i = M, && 1 \leq i \leq h-1, \\ &w_i = R - M, && 1 \leq i \leq h. \end{aligned} \tag{5}$$

Additionally, since $\frac{S}{e_h} = \frac{M}{w_1} \prod_{i=1}^{h-1} \frac{m_i}{w_{i+1}}$, for topologies with constant radix ratio the amount $\text{gmr} = \sqrt[h]{S/e_h}$ is the fitness ratio. And for all topologies, regardless of fulfilling the constant ratio property, gmr is the geometric mean of those ratios.

Remark 1 *Note that for our convenience, n_1 denotes the number of leaf routers and $S = Mn_1$ is the total number of servers, although in [11] the authors directly use n_1 as the number of servers.*

Most of the folded Clos in the industry roughly fit into the Extended Generalized Fat Tree (XGFT) network topology [11]. This definition allows to consider alternative topologies based on fat-tres but with less cost, what in some publications are named as *slimmed fat-trees* [7], *fit-trees* [8], *tapered fat-trees* [9] and other variations. The original definition in [11] was recursive. Next, in Definition 1 a new definition avoiding recursion is given, which simplifies the analysis of the topology.

Definition 1 *The Extended Generalized Fat Tree* $\text{XGFT}(h; m_1, \ldots, m_h; w_1, \ldots, w_h)$ *topology of height h consists on n_k routers in level k, with k ranging from 1 to $h+1$, where n_k is computed as*

$$n_k = \prod_{i=1}^{k-1} w_i \prod_{i=k}^{h} m_i,\ 1 \leq k \leq h+1. \tag{6}$$

Routers are connected to contiguous levels such that, the router at position x of level k, $1 \leq k \leq h$ is connected with the routers at position y of level $k+1$ if and only if there are integer numbers q, r, t, and u satisfying

$$\begin{aligned} &x = (qm_k + r)g + u, \quad 0 \leq r < m_k, \quad 0 \leq u < g, \\ &y = (qg + u)w_k + t, \quad 0 \leq t < w_k, \quad 0 \leq q < \prod_{i=k+1}^{h} m_i, \end{aligned} \tag{7}$$

where $g = \prod_{i=1}^{k-1} w_i$.

Note that both Equation 1 and Equation 6 are not independent. If the system of equations is reduced, then it is obtained that it is equivalent to Equation 1 with the restrictions corresponding to Level 1:

$$n_1 = \prod_{i=1}^{h} m_i.$$

Table 2: Topologies evaluated and their topological parameters.

Scenario	Topology	Servers	Routers	e_1	e_2	gmr
A	XGFT$(2;18,36;18,18)$	11664	1620	11664	11664	1
B	XGFT$(2;22,36;14,14)$	17424	1492	11088	7056	1.57
B	XGRFC$(2;22,36;14,14;792,504,196)$	17424	1492	11088	7056	1.57
B	XGRFC$(2;18,36;11,18;684,418,209)$	17100	1311	7524	7524	1.51
C	XGFT$(2;26,36;10,10)$	24336	1396	9360	3600	2.60
C	XGRFC$(2;26,36;10,10;936,360,100)$	24336	1396	9360	3600	2.60
C	XGRFC$(2;18,36;5,18;720,200,100)$	22320	1020	3600	3600	2.51

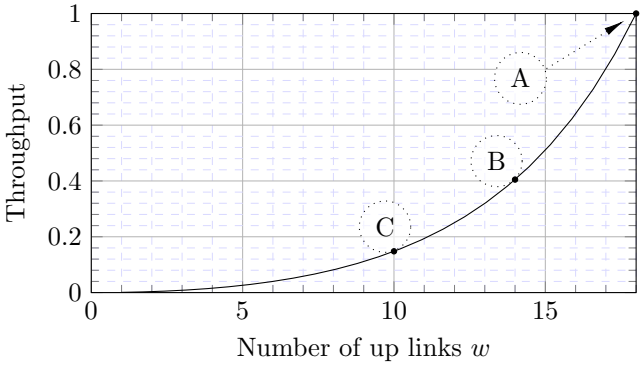

Figure 2: Theoretical maximum throughput with constant ratio of w up-links from the 36 total links when all packets reach level 3. Labels A, B, and C denote topologies in Table 2.

Example 1 *Let us consider router radix $R = 36$, very commonly used by industry, which will be used henceforward in the paper. Then,* XGFT$(2;18,36;18,18)$ *is the FT for this router radix and 3 levels. In the remainder of the paper* XGFT$(2;22,36;14,14)$ *and* XGFT$(2;26,36;10,10)$ *will be also considered. Note that these topologies are slimmed variants of the FT, with respective fitness ratios $\frac{22}{14} = 1.57$ and $\frac{26}{10} = 2.6$, which will imply different performance. Their topological properties, together other topologies that will later introduced in the paper, are summarized in Table 2.*

The high cost of a non-blocking FT interconnection network is better exploited when the application packets reach the routers in the $h + 1$ level. However, scientific applications constitute particular communication patterns. In fact, by means a thorough study, in [8] the authors demonstrated that a significant percentage of scientific applications send most traffic to near neighbours. Thus, it may be worth to dimension the topology for these applications, which would allow to reduce costs. In the following it is established a relation between links in the different stages and the injection rate per server.

Let γ be the average injection rate per server. Thus, there is a total of γS phits (packet units) that are being created on each cycle. Let p_i be the fraction of packets which reach some router at level i, potentially going further up. Clearly, it is hold that, $1 = p_1 \geq p_2 \geq \cdots \geq p_{h+1} \geq 0$. Then $\gamma S p_i$ is the total rate of packets reaching the routers in level i, giving the following immediate bound on the injection rate

$$\gamma S p_{i+1} \leq e_i. \tag{8}$$

In the case of using the constant ratio in Equations 5, we have the following nice expression:

$$\gamma p_{i+1} \leq \left(\frac{R-M}{M}\right)^i, \ 1 \leq i \leq h. \tag{9}$$

Example 2 *As an example of the previous bound we consider an extreme scenario in which all packets reach level 3, that is, $p_3 = 1$. Then, the maximum throughput is $(w/M)^2$, which is represented in Figure 2 for router radix $R = 36$. In this figure, it can be seen that the throughput is maximum when half of the ports go upward ($R = 2w$) and it decreases acutely (in fact hyperbolically) with reductions on w. Thus, it is clear that reducing the cost of a folded Clos by reducing the w_i terms has great impact on performance; at least for applications that have relatively many global communications. Note that the three points A, B and C in the figure correspond to the interconnection networks summarized in Table 2.*

In an XGFT, for any given leaf router, there are exactly $m_1 \cdots m_t$ leaf routers at distance at most $2t$, including itself. From this, it follows that, in a uniform traffic pattern the probability that a packet reaches level i is, in the XGFT, $p_i = 1 - \prod_{k=i-1}^{h} m_k^{-1}$. For the particular case of radix $R = 36$ we get that $p_{h+1} = 35/36 = 0.972$, which is very

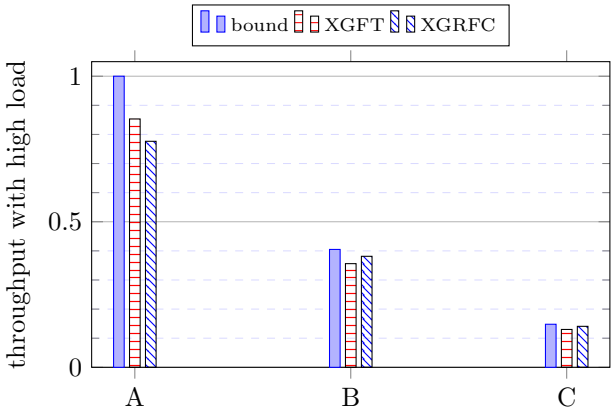

Figure 3: Achieved throughput under heavy load and global communication pattern. Labels A, B, and C denote topologies in Table 2.

close to 1. This means that the traffic pattern considered in Example 2 in fact closely resembles the uniform traffic pattern in the XGFT, but this does not hold for an arbitrary folded Clos. In the general case, assuming that there are not multiple links between any pair of routers we get $1 - \frac{m_1 w_1}{n_1} \leq p_3 \leq 1 - \frac{m_1}{n_1}$ for uniform traffic. If $m_1 = 22$, $w_1 = 14$ and $n_1 = 792$, then we obtain $0.611 \leq p_3 \leq 0.972$.

The formulas in Equation 9 can be validated by experimental simulation. We use parameters as in Example 1 and Example 2, and the predicted throughput is shown in Figure 2. For each of the points A, B, and C we compare the bound with the simulation values of the corresponding XGFT and the simulation values of a random analogue topology. These random topologies, called XGRFCs, are folded Clos with the links in each stage wired randomly; they are detailed in the next section. We have performed the simulations with a synthetic traffic pattern designed to reproduce the $p_3 = 1$ assumption. Therefore, all the packets reach the last level routers, so links in the last stage are widely used. Specifically, each time a packet is generated, a server at maximum distance is selected in a random uniform way as the destination of the packet. As it can be seen in Figure 3, where both the simulation results and the values predicted are represented, the theoretical model accurately estimates the achieved throughput. In a more deep analysis, it has to be noticed that XGFT is always at the same relative distance to the upper bound provided by the theoretical model. However, the greater the fitness ratio, the tighter the difference between the simulated throughput and its theoretical bound in XGRFCs.

3 Extended Generalized Random Folded Clos

Random Folded Clos (RFC) networks were introduced in [3] as an alternative to FTs that increases scalability, facilitates graceful expansion and reduces cost. These interconnection networks can be roughly described as folded Clos networks in which each level is randomly interconnected. Next, a generalization of these networks, in the same flavour that XGFTs, is presented.

Definition 2 *Let us define a Extended Generalized Random Folded Clos, and denote it by* XGRFC$(h; m_1, \ldots, m_h; w_1, \ldots, w_h; n_1, \ldots, n_{h+1})$*, a random multi-stage interconnection network selected among all the possible with the given parameters chosen near-uniformly. The parameters need to satisfy Equations 1 and 3 like any other multistage network.*

An implementation of almost uniform random bipartite graphs was presented in [3]. This algorithm was used to construct RFCs, and equivalently it can be used to build XGRFCs.

Although XGFTs are always up/down connected, in the case of XGRFCs this fact has to be verified. In [3], the conditions under a RFC is up/down connected were proved. Using the same techniques it might be proved that XGRFCs tend to be up/down connected with probability $e^{-e^{-x}}$ for

$$x = n_{h+1}^{-1} \prod_{i=1}^{h} w_i^2 - \ln \binom{n_1}{2}.$$

Proving this result is out of the scope of this paper, both because its mathematical complexity and because it is possible to compute the up/down condition directly. The up/down distances can be quickly computed with a slight modification of the Breadth First Search, which shows if the network is actually up/down connected. Although in the networks used in our examples such probability is so close to 1 that the check is not necessary, the computation is going to be performed anyway to populate the routing tables. In cases closer to the threshold, i.e., with the x in the

Table 3: Topological Parameters of Topologies in Example 4.

Levels	**Topology**	**Radix**	**Servers**	**Routers**	**links**	gmr
4	XGFT$(3; 54, 54, 92; 38, 38, 38)$	92	1.45M	644K	22.4M	1.42
3	XGRFC$(2; 54, 92; 38, 38;$ $268272, 188784, 77976)$	92	1.45M	535K	17.4M	1.42
5	XGFT$(4; 10, 10, 10, 16; 6, 6, 6, 6)$	16	160K	36.1K	208K	1.67
4	XGRFC$(3; 10, 10, 16; 6, 6, 6;$ $16000, 9600, 5760, 2160)$	16	160K	33.5K	188K	1.67

probabilistic formulae close to 0, checking the up/down connectivity would be necessary. If the check fails, then we have just to generate again a different network with another seed for the random number regenerator.

Example 3 *Let us consider* $R = 36$ *and the topologies* XGRFC$(2; 18, 36; 11, 18; 684, 418, 209)$ *and* XGRFC$(2; 18, 36;$ $5, 18; 720, 200, 100)$. *Their topological properties are summarized in Table 2. As it can be seen,* XGRFC$(2; 18, 36;$ $11, 18; 684, 418, 209)$ *connects 2% less servers but with 12% less routers of the corresponding XGFT. In the case of* XGRFC$(2; 18, 36; 5, 18; 720, 200, 100)$*, cost gain is more important, since 27% less routers are needed to connect 8% less servers. When applying the probabilistic formulae we get values* $x \geq 80$*, which means that the probability of being up/down-connected is greater than* $1 - 10^{-x/\ln(10)} \geq 1 - 10^{-34}$*, which is almost 1. Therefore, it is practically impossible to generate a XGRFC up/down disconnected with these parameters. However, once the XGRFC is generated, its up/down connectivity is verified, and in the case it is not fulfilled, it is just a matter of generating another one.*

Example 4 *For large networks, sometimes it is possible to find an up/down XGRFC connecting the same amount of servers than a XGFT with the same radix and fitness ratio, but having one less level. The size of the network for which this is possible grows with the fitness ratio. The topologies and their properties in this example are summarized in Table 3. To illustrate a 4 to 3 level reduction, we can consider the* XGFT$(3; 54, 54, 92; 38, 38, 38)$ *of radix 92 that connects 1.4M servers. Then, the* XGRFC$(2; 54, 92; 38, 38; 268272, 188784, 77976)$ *is a random analogue with one less level that is up/down connected with a probability around 0.92. As an example of a 5 to 4 level reduction, we can consider the* XGFT$(4; 10, 10, 10, 16; 6, 6, 6, 6)$ *of radix 16 that connects 160K servers. Then, the* XGRFC$(3; 10, 10, 16;$ $6, 6, 6; 16000, 9600, 5760, 2160)$ *is a random analogue with one less level that is up/down connected with a probability around 0.95. These cases suppose an improvement in latency from the lesser height in addition to the cost reduction by having less routers and cables.*

Another property that XGRFCs inherit from the RFCs is the expandability. In a fully populated XGFT, that is with all the Mn_1 servers, expanding the system implies making drastic changes, such as increasing the height, changing the fitness ratio or replacing the routers with others with greater radix. On the contrary, in a XGRFC this is possible by just adding some routers in each level and randomly rewiring some of the links in each stage. Note that the number of routers added in the first level must be a multiple of $\prod_{k=1}^{h} \frac{m_k}{\gcd(w_k, m_k)}$ in accordance with Equation 1. Otherwise, some routers would have unwired ports. This provides a simple way to gradually increment the capacities of a system based on a XGRFC.

4 Evaluation

To conclude the study, this section is devoted to the experimental evaluation. In Subsection 4.1 the experimental set up is described, including topologies evaluated, simulator, traffic patterns, etc. In Subsection 4.2 experimental results for the simulated topologies are shown.

4.1 Experimental Set Up

Next, different topologies are evaluated by simulation. The experiments have been done using the functional simulator in [10]. The simulations have been performed considering a router with 4 virtual channels, input buffers of length 4 packets and virtual cut-through as flow control. Every packet has 16 phits. Both link latency and router arbitration take 1 cycle.

For the experiments, we use the topologies summarized in Table 2. As asserted before, most of the folded Clos in the industry are XGFTs, thus we compare XGRFCs and XGFTs. Firstly, we evaluate the family B with smallest fitness ratio (other than 1). Later, we compare the results with the ones denoted by C. In both cases, one XGFT and two different XGRFCs are compared. The first XGRFC is always done using the same resources as the XGFT, that is, the same number of servers, routers and cables. On the contrary, the second one has been selected to provide lower cost and the same performance under global traffic patterns. This has been done by enforcing the same number of links in the last level and reducing the ones in the first level to the minimum possible.

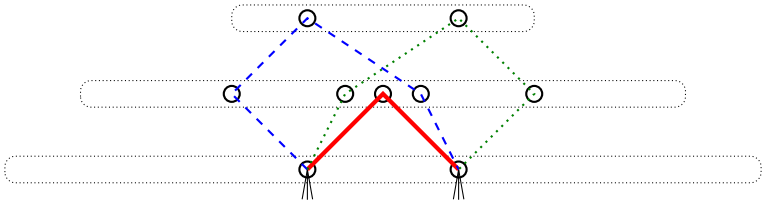

Figure 4: Up/down paths of different lengths: one of length 2 and two of length 4.

All these topologies are compared in terms throughput, average latency and Jain's fairness index [6]. The throughput and average latency are common measures, with throughput being the injection rate from the servers and the average latency being the average number of cycles required to consume the packet. Jain's fairness index is a function of the coefficient of variation on the injection across the servers. A value of this index of $\frac{k}{S}$ is compatible with having k servers generating the same amount of traffic and the $S-k$ remaining servers generating no traffic at all. Some compatible scenarios with a $\frac{8}{10}$ index would be to have $S = 130$ servers from which either only 104 are working or 81 are working with rate 16 and the remaining 49 with the lower rate 9. Thus, a bad Jain index may mean that a few servers have important issues or that many servers have a poor performance, both being inadmissible.

The experiments have been done using three synthetic traffic patterns, that have been slightly adapted from [4]. These traffic patterns have been selected to represent typical application behaviour, which are:

- *Uniform*: each generated packet has as destination a random compute node selected uniformly.
- *Random-pairing*: the set of switches is initially divided into pairs in a random uniform way. Each compute node generates packets with destination any of the compute nodes in the switch paired to its switch. This traffics pattern is a case of a random permutation of the switches, which is more adversarial than a permutation of the compute nodes.
- *Fixed-random*: at the beginning, each switch selects a different switch in a random uniform way. During the simulation each node generate packets towards the selected compute node. It is not a permutation since some compute nodes in different switches can have selected the same destination.

An up/down route is constructed as follows: first taking up links till a common ancestor is reached, and then going down. Only up/down routes are considered, and when various routes exist, one is selected randomly. Unlike what happens in XGFTs, in XGRFCs for some pairs of leaf routers, there are up/down routes of different lengths, as illustrated in Figure 4. In this schematic example, we have tried to show a situation that it is common in XGRFCs. Two routers (in the leaf level) can communicate using different routes. The one in solid red is minimum, that is, it provides distance 2. However, at least two more routes are possible in this example, those depicted in dashed blue and dotted green, but in this case with longer length, providing distance 4. Thus, in these topologies it is possible to use two different routing algorithms: *minimal_routing*, in which only minimum up/down routes are considered, and *all_paths_routing*, using all possible up/down routes. In the next sections, when illustrative, both routings are used for XGRFCs.

4.2 Experimental Results

First, let us consider the results of the simulations for the topologies with a smaller fitness ratio, that is, scenario B. In Figure 5, the results for both throughput and latency of the different interconnection networks under uniform traffic pattern are shown. In these graphs, random topologies have been evaluated using minimal_routing and all_paths_routing. As it can be seen, random networks benefit from minimal routing under uniform traffic pattern. In this case, the XGRFC with the same resources as the XGFT provides 38% more throughput. However, as it will be seen later, in XGRFCs all the up/down routes should be used under other non-uniform traffic patterns. With all_paths_routing, all the topologies provide almost the same throughput. Note that the latency graph perfectly corresponds with the one being expected from the observed throughput. Since this happens across all experiments, latency graphs are no longer shown in favour of showing Jain's fairness results.

Concerning the random pairing traffic pattern, the results of the experiments can be seen in Figure 6. As mentioned before, in random topologies, restricting to minimal routes not only constitutes a disadvantage for throughput but also has harmful consequences on fairness. Considering the throughput results for all_paths_routing, the XGFT provides the best performance. For the corresponding XGRFC with the same resources, this performance falls a 5%. The cheaper XGRFC with 12% less routers provides 27% less throughput.

Finally, the results for the fixed random traffic pattern are shown in Figure 7. In this case only the evaluation with all the routes is shown, since restricting only to minimal routes has the same problems that has already been observed in random pairing traffic. It can be seen that XGRFC provides 20% more throughput than XGFT when it

uses the same resources. However, the cheaper version provides 7% less performance than XGFT. Nevertheless, when analyzing fairness, it can be observed that both random topologies have an excellent behaviour, and XGFT exhibits an important problem.

Now, let us analyze what happens for a greater fitness ratio, that is, scenario C. In this case, only the results for uniform traffic are shown in Figure 8, since the other traffic patterns provide similar outcomes. As it can be observed, the behaviour is almost the same as the one shown in Figure 5. The only difference that can be highlighted is that the discrepancy between throughput measured for both minimal_routing and all_paths_routing, has been decreased with respect to the topologies denoted by B. Note that this happens because more fitness ratio implies less routers in the top level, thus providing less path diversity.

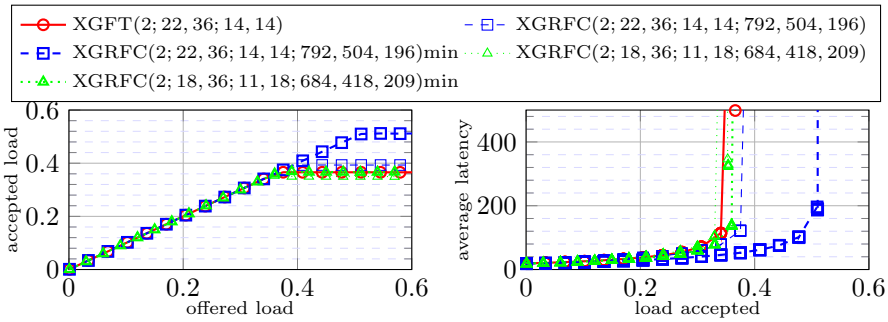

Figure 5: Uniform traffic: average accepted load and average latency.

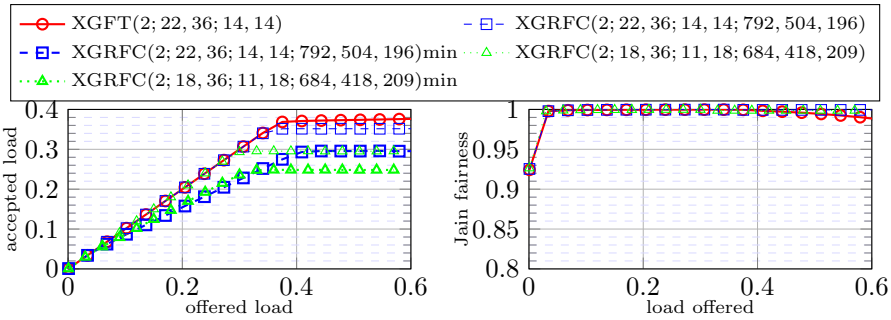

Figure 6: Randompairing traffic: average accepted load and Jain fairness index.

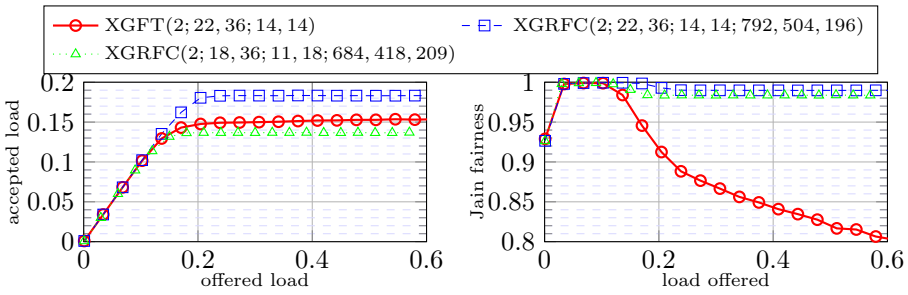

Figure 7: Fixedrandom traffic: average accepted load and Jain fairness index.

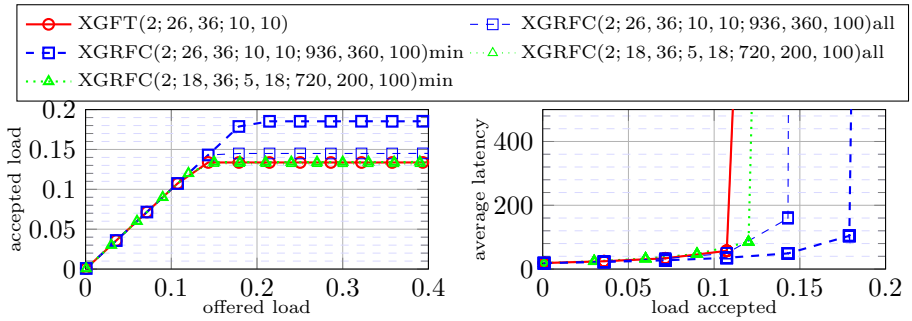

Figure 8: Uniform traffic: average accepted load and average latency.

5 Conclusions

It has been proved that the performance of a slimmed folded Clos, both standard or with random interconnection, can be estimated in terms of the nature of communications of the application. Although other cases can be considered, we have selected an application with a totally global traffic pattern, in which all its communications use the links at the last stage. This global traffic is almost uniform traffic in the considered topologies. We have measured the impact of different fitness ratios on the performance, both with a theoretical model and corroborated by simulation. We have shown that the information provided by the model would be of great interest for systems designers to make a better usage of their procurement budget. Moreover, random topologies provide greater cost savings, since it is possible to build them with fewer resources, in exchange for an assumable degradation of the performance and an improvement in fairness. Furthermore, extended random folded Clos topologies provide higher scalability, great expansion and better fault tolerances than the extended fat trees counterparts do.

Acknowledgements

This work has been supported by the Spanish Ministry of Science, Innovation and Universities under contract TIN2016-76635-C2-2-R (AEI/FEDER, UE). The first author is supported by the Spanish Minister of Science and Innovation programme Juan del Cierva Formación reference FJCI-2017-31643.

References

[1] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A scalable, commodity data center network architecture. In *Proceedings of the ACM SIGCOMM 2008 conference on Data communication*, SIGCOMM '08, pages 63–74, New York, NY, USA, 2008. ACM, ACM.

[2] C. Camarero, C. Martínez, and R. Beivide. On random wiring in practicable folded clos networks for modern datacenters. *IEEE Transactions on Parallel and Distributed Systems*, 29(8):1780–1793, Aug 2018.

[3] Cristóbal Camarero, Carmen Martínez, and Ramón Beivide. Random folded Clos topologies for datacenter networks. In *Proceedings of the 23rd IEEE Symposium on High Performance Computer Architecture*, HPCA '17, pages 193–204, 2017.

[4] Dong Chen, Noel Eisley, Philip Heidelberger, Sameer Kumar, Amith Mamidala, Fabrizio Petrini, Robert Senger, Yutaka Sugawara, Robert Walkup, Burkhard Steinmacher-Burow, Anamitra Choudhury, Yogish Sabharwal, Swati Singhal, and Jeffrey J. Parker. Looking under the hood of the IBM Blue Gene/Q network. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 69:1–69:12, Los Alamitos, CA, USA, 2012. IEEE Computer Society Press.

[5] Charles Clos. A study of non-blocking switching networks. *Bell System Technical Journal, The*, 32(2):406–424, March 1953.

[6] R. Jain, D.M. Chiu, and W. Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. In *DEC Research Report TR-301.*, pages 598–607, June 1984.

[7] A. Jokanovic, J. C. Sancho, J. Labarta, G. Rodriguez, and C. Minkenberg. Effective quality-of-service policy for capacity high-performance computing systems. In *2012 IEEE 14th International Conference on High Performance Computing and Communication 2012 IEEE 9th International Conference on Embedded Software and Systems*, pages 598–607, June 2012.

[8] S. Kamil, L. Oliker, A. Pinar, and J. Shalf. Communication requirements and interconnect optimization for high-end scientific applications. *IEEE Transactions on Parallel and Distributed Systems*, 21(2):188–202, February 2010.

[9] E. A. León, I. Karlin, A. Bhatele, S. H. Langer, C. Chambreau, L. H. Howell, T. D'Hooge, and M. L. Leininger. Characterizing parallel scientific applications on commodity clusters: An empirical study of a tapered fat-tree. In *SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 909–920, November 2016.

[10] Javier Navaridas, José Miguel-Alonso, Jose Antonio Pascual, and Francisco Javier Ridruejo. Simulating and evaluating interconnection networks with INSEE. *Simulation Modelling Practice and Theory*, 19(1):494–515, 2011.

[11] S. R. Ohring, M. Ibel, S. K. Das, and M. J. Kumar. On generalized fat trees. In *Proceedings of 9th International Parallel Processing Symposium*, pages 37–44, April 1995.

[12] Germán Rodríguez. *Understanding and Reducing Contention in Generalized Fat Tree Networks for High Performance Computing*. PhD thesis, Facultad de Informática, U. Politécnica de Cataluña, 2011.

[13] A. Shpiner, Z. Haramaty, S. Eliad, V. Zdornov, B. Gafni, and E. Zahavi. Dragonfly+: Low cost topology for scaling datacenters. In *2017 IEEE 3rd International Workshop on High-Performance Interconnection Networks in the Exascale and Big-Data Era (HiPINEB)*, pages 1–8, 2017.