Design Methodology for Optimizing Optical Interconnection Networks in High Performance Systems

Sébastien Rumley[1], Madeleine Glick[2], Simon D. Hammond[3], Arun Rodrigues[3] and Keren Bergman[1]

[1] Lightwave Research Laboratory, Columbia University, New York, USA
{rumley, bergman}@ee.columbia.edu
[2] College of optical science, University of Arizona, Tucson, USA
mglick@optics.arizona.edu
[3] Sandia National Laboratories, Albuquerque, USA
{sdhammo, afrodri}@sandia.gov

Abstract. Modern high performance computers connect hundreds of thousands of endpoints and employ thousands of switches. This allows for a great deal of freedom in the design of the network topology. At the same time, due to the sheer numbers and complexity involved, it becomes more challenging to easily distinguish between promising and improper designs. With ever increasing line rates and advances in optical interconnects, there is a need for renewed design methodologies that comprehensively capture the requirements and expose trade-offs expeditiously in this complex design space. We introduce a systematic approach, based on Generalized Moore Graphs, allowing one to quickly gauge the ideal level of connectivity required for a given number of end-points and traffic hypothesis, and to collect insight on the role of the switch radix in the topology cost. Based on this approach, we present a methodology for the identification of Pareto-optimal topologies. We apply our method to a practical case with 25,000 nodes and present the results.

Keywords. Topology · Network · HPC · Interconnect

1 Introduction

As aggregated computing power approaches the Exascale mark, and more importantly, as parallelism reaches unprecedented levels, modern interconnects need to provide ever growing bandwidths and connectivity. For rack-to-rack links, and in the near future, for all types of connections, this trend is likely to lead to the increased use of photonic networks. This transition provides an opportunity to re-examine interconnect design methodologies. Photonic systems differ in many aspects from conventional electrical ones. Depending on which optical technology is used, and how it is used, a particular design can be well suited or on the contrary fairly ill-adapted.

In this work, we examine the selection of the interconnect topology using two criteria: the switch radix, and the number of links. There is considerable research into increasing the port count of different flavors of optical switches [1, 2]. It would be helpful to system designers and component researchers to have a clearer view of the goals and trade offs. The situation is the same at the link level. There is considerable progress in components and also work on power efficient and/or power proportional systems. We aim to establish a methodology that can clearly expose the benefits and shortcomings of various topologies along these axes.

A great variety of topology descriptions are available in existing literature. Multi-dimensional tori and hierarchical structures (fat-trees) have been the dominant super-computer interconnects for many years (mid 1980s to mid 2000s) [3]. Tori fundamentally replicate the same structure as many of the simulations they support (2D or 3D stencils) [4]. They offer bandwidth in a very specific manner, which can be very efficient if the supporting code follows the same structure [5]. On the other hand, fat-trees can be constructed as virtual crossbars offering general purpose connectivity [6, 7]. Connectivity patterns, have, however, recently received renewed attention and these toroidal or hierarchical architectures have been progressively superseded by more complex schemes. This can be partly explained by the advent of two trends. On one hand, there is a growing imbalance between the energy costs of computation and communication. Tighter transistor integration allows computing circuits to be more efficient, but contributes little in decreasing bit displacement costs [8]. Networks are thus growing contributors to overall energy consumption - and subsequently dissipation, which is also a serious concern. On the other hand, increased expectations in terms of aggregated computing power, coupled with the progressive saturation of both CPU and chip computing power, leads to a general inflation of the interconnect size [9]. As underlined in several studies (e.g. [10]) as well as in what follows, the cost of traditional tori and fat-trees becomes discouraging at very large scale and under arbitrary traffic. More scalable alternatives, in particular, Flattened Butterfly [3], Dragonfly [11], Jelly-fish [12] and Slim-fly [13] have thus been proposed.

All these studies, however, address the question "what is the ideal topology?" by promoting a recipe to assemble a topology and by comparing it to other instances. In this paper, we dig deeper and propose an analytical approach to evaluate the fundamental connectivity requirements. Unless a radical breakthrough is achieved, larger interconnects will continue to be required, they will represent an increasing part of the operational costs, and they will integrate an increasing proportion of photonics. The motivation is thus to more clearly define metrics, requirements and options to gauge suitability of design options. Our proposed methodology, by allowing a rapid evaluation of the switch radix/capacity requirement, addresses these needs. It is also general enough to be applied in multiple contexts, even photonic-free ones.

Our approach is based on a capacity-flow analysis, and uses Generalized Moore Graphs. These graphs minimize the average routing distance. As such, they can be used as an optimal bound in terms of connectivity, as developed in Section 3. In Section 4, we further develop our approach in order to support a broader range of input cases. In particular, we integrate in our formulation a variable concentration factor, which represents the number of nodes or servers that share an entry switch. Varying

this factor allows us to explore the fundamental cost/performance/hardware requirements trade-offs that exists in terms of topologies, for various global interconnect sizes. We use these results to derive conclusions on fundamental future needs in terms of the switch radix. In Section 5, we evaluate several classical topologies against the bound, and show that Flattened Butterflies, despite their relative simplicity, achieve good performance. Since routing in these networks can be made very simple, they are still of interest for future large scale machines. Finally we sketch future research directions where our approach could be of further use, or further developed, and draw final conclusions in Sections 6 and 7.

2 Related work

Most classical means to interconnect multiple Processing Elements (PE) have been investigated in the 1970s-1980s. Preparata et al. proposed the Cube-Connected-Cycles network architecture [14], Bhuyan et al. summarized the properties of Hypercube networks [15] previously proposed by Pease [16] and Leiserson described the Fat-Tree concept [6]. Dally investigated the performance of n-dimensional tori [17]. These analyses targeting parallel computing were preceded by those realizing efficient automated switches for telephone networks [18]. The fundamental performance and cost metrics that apply to connection networks (average traffic delay, average traffic density, total connection cost, connections per node, reliability) are analyzed in the most "obvious" types of interconnection by Wittie [19].

Most of these references are over 20 years old. Since then, the computing power of supercomputers has generally doubled every year on average in two eras. Until the early 2000s, most of this increase was covered by the increase of the single CPU power (either through higher clocking or increased instruction level parallelism). The increase in terms of CPU (or core, as multi-core appeared) parallelism, although continuous, was less strong. Consequently, over this period, interconnect node count increased only modestly and results obtained in the literature cited above were sufficient to meet demand.

At the turn of the millennium, however, the scaling limitation of a single processor [9] became apparent. This triggered renewed interest in interconnection structures, in particular in those supported by high radix switches [20]. These switches, instead of providing only 8 to 24 bi-directional ports, pushed the port count to 64 [21] by 2006, and to more than 100 by today [22]. With such high port counts, more highly connected and compact topologies become feasible. The Flattened Butterfly [3], Dragonfly [11], and more recently, Slim-fly [13] architectures all enter in this new generation of topologies.

There are various reasons that might promote one topology over another. Some structures make routing decisions (Flattened Butterfly [3], Tori [17]), traffic-balancing (Dragonfly [11]) or incremental deployment (Exchanged Crossed Cube [23], Jellyfish [12]) easier. As the interconnect size grows, however, the economic argument becomes dominant, favoring designs involving as few resources as possible.

The aim of efficient resource utilization naturally leads one to a graph theoretic approach. The relation to the Moore Graph and Moore bound is mentioned by Von Conta [24], who also analyzes the structure of Tori and proposes other graphs (and methods to create them) to approach the Moore bound. The suitability of some Cayley graphs for designing interconnection networks is underlined by Akers and Krishnamurthy [25]. Hsieh and Hsiao showed k-valent Cayley graphs are maximally fault tolerant [26]. McKay, Miller and Siran propose a method (exploited to create the Slim-fly architecture) to construct non-Cayley graphs of unlimited sizes, all of diameter two (shortened as MMS) [27]. Most of these MMS graphs are the largest diameter 2 graphs known so far. A table of the largest graphs (of any diameter < 10) is provided in Reference 28.

The usage of optimal graphs as the Petersen or the Hoffman-Singleton graph has also been proposed [29]. Random search methods have been exploited to identify large graphs of small diameter [30, 31]. Random addition of links to rings [32] or simply random wiring of switches [12] has also been investigated. The optimal properties of the Generalized Moore Graph for comparison purposes have been previously exploited in the context of Metropolitan Network planning [33].

Topologies are generally compared and considered in the literature under maximal all-to-all traffic, as a proxy for worst conditions. In practice, observed utilization pattern almost always clearly differ from all-to-all, as illustrated by Vetter et al. [34] or Kandula et al. [35]. A significant effort is also invested to improve the matching between (actual or future) parallel codes requirements and topology capabilities, in particular through large-scale simulation [36, 37].

3 Identifying ideal connectivities

We start by defining terms and notations, and by clarifying the framework of this study. The decision problem can be summarized as follows: how does one connect N processing elements (PE) such that each PE pair can communicate (even minimally). We assume that each PE has a single bidirectional link with the outside world. We also assume that every link has a given, constant capacity, for example 10 Gb/s. In our calculations, all capacity and traffic measures are normalized to this reference capacity. Consequently, there is no need to retain its absolute value, and a PE can be assumed to have a communication capacity of 1 unit (u), in both directions. We also neglect how links are implemented. In practice links may be composed of multiple parallel lanes or even cables. We finally assume that PEs are attached to a single switch through a unique, bidirectional link of $1u$.

We denote r the number of ports available in each switch, i.e. the switch radix. All switches are assumed equivalent and capable of achieving 100% throughput as long as flows are (on the medium term) well balanced among the input and outputs. Unless r is larger than the number of PEs N, each switch is connected to at least one other switch to ensure global connectivity among all PEs. Two switches can be connected by more than one link in order to support a sustained traffic flow larger than $1u$.

We further assume that PEs are distributed as equally as possible among the switches, and hence that we have $S = \lceil N/C \rceil$ switches, where C is the concentration factor. With this assumption, we focus our study on direct interconnection networks as opposed to indirect ones. This assumption also allows us to consider all switches as equivalent [25].

We are interested in determining how many links must be placed between each switch pair. In raw optimization terms, this means finding $S(S-1)$ positive integer values, which become an unmanageable problem, especially if S is equal or larger than one thousand as it is the case in recent supercomputers.

We start by investigating how many switch pairs should be connected, or, stated slightly differently, how highly connected should the topology be. We define the connectivity of a topology, noted R, as the average number of direct neighbors a switch has, excluding the PEs. This is equivalent to the average vertex degree in graph theory. R can be as large as $S-1$, in which case the topology is a full-mesh. If the topology forms a ring, $R=2$. Topologies with a connectivity $R<2$ are possible but are not fault tolerant, a quality that we expect from HPC interconnects. Establishing how highly connected should the topology be means therefore finding the appropriate value of R between 2 and $S-1$.

R determines the global capacity of the topology. Hence, without further assumption about the way given PE pairs communicate, and since all switches are equivalently connected to PE, there is no first order reason to concentrate more connections around a particular switch or between a particular switch pair. If n links have to be installed between two switches to adequately support the traffic, n links will also be required between any of the connected pairs. This allows us to observe that, under this assumption, the total number of links is $nSR/2$ and the total installed capacity is nSR[1].

Considering the traffic demand, the amount of data that is injected into the interconnect, and the instants at which that data is injected is affected by multiple factors: node or server computing power and architecture, type of software executed, implementation of the software, or even input-data currently processed by the software. Quantifying the requirements is therefore a difficult task. To obtain an idea of the design requirements one generally defines a challenging scenario, as the maximal uniform traffic case: each of the PEs of the system uses its maximum networking capacities to distribute messages equally among all other PEs, i.e. each PEs sends a flow of $1/(N-1)\ u$ to each other PE. Under this traffic assumption, and that one PE is connected to each switch, i.e. $C=1$, we can formulate the total traffic flow as

$$N(N-1)\cdot\frac{1}{N-1}\cdot\Delta = N\Delta$$

which is the product of the number of flows, the flow magnitude and the average routing distance (in hops), Δ. Hence, the more hops a message has to travel, the more time it occupies links and switches. Note that the knowledge of the average distance is sufficient as long as all flows have the same magnitude.

[1] Each switch is connected on average to R others with n links, thus nSR ports are occupied in total. As each link connects two ports, the number of links is $nSR/2$. Since each link is assumed bidirectional, it represents two units of capacity so the total capacity is nSR.

We have shown that in a regular topology with S switches, each associated with 1 PE ($S=N$ as $C=1$) and under maximum uniform traffic, the total traffic is $N\Delta$. On other hand, the capacity obtained maintaining symmetry is $nSR = nNR$. In order to have the topology at least asymptotically supporting the traffic, the inequality $N\Delta \leq nNR$ must hold.

To evaluate the connectivity R, and to this aim, express Δ as a function of R, we assume that the topology will be organized such that distances between switches are kept short. We can expect the final topology, whatever its R being, to be closely related to an ideal one of same size that minimizes the average distance Δ. In the best case, this ideal topology is a Generalized Moore Graph (GMG - described in the Appendix) and the average distance $\Delta_{GMG}(R)$ between a switch and its $S-1$ neighbors (and thus between any node pair) can be written as

$$\Delta_{GMG}(R) = \frac{R + 2R(R-1) + 3R(R-1)^2 + \ldots + (D-1)R(R-1)^{D-2} + Dx}{S-1} \quad (1)$$

where $x = S - 1 - R - R(R-1) - R(R-1)^2 - \ldots - R(R-1)^{D-2}$ is the number of neighbors at a maximum distance D, and R is the maximum degree in the topology (D is also the diameter).

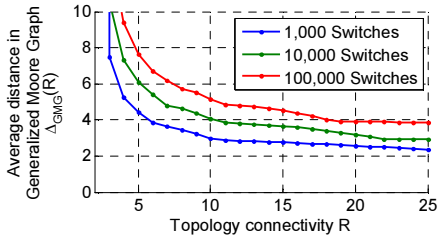

Fig. 1. Evolution of the average distance in a Generalized Moore Graph.

With Δ expressed as a function of R, we rewrite the inequality as $\Delta_{GMG}(R) \leq nR$. We simplify further by showing that n can be assumed to be equal to one. As we are interested in minimizing the topology costs, which we assume highly correlated with the number links, we also want to minimize the product nR. As shown in Figure 1, the average distance $\Delta_{GMG}(R)$ decreases with larger values R, and with it the total traffic. In contrast, changing n has no influence on the traffic. We can exploit this fact. Taking the case $n=2$ and $R=4$, thus $nR=8$. By rebalancing the factors using $n'=1$ and $R'=8$, the product is unchanged. On the contrary, the traffic side of the equation may be smaller but will never be larger. Therefore, under the assumptions listed so far, choosing $n=1$ never leads to less economical topologies.

Finally, this allows us to state that the smallest integer value R for which the inequality $\Delta_{GMG}(R) \leq R$ holds, R_{opt}, is the connectivity of the most economical topology that supports N PEs, each connected to a switch and exchanging maximum uniform traffic.

There is no evident closed-form expression for R_{opt}. However, as shown in Figure 2a, R_{opt} grows relatively slowly with $N=S$, it is therefore easy to find R_{opt} by calculating $\Delta_{GMG}(R)$ for increasing values of R. This approach allows one to also identify the

diameter of the most economical GMG, D_{opt} (Figure 2b). Note that D_{opt} is not monotonically increasing. When considering topologies of increasing sizes, at some points the larger size justifies a larger connectivity. As this increment of R allows more switches to be present at closer distances, it generally allows one to "remove" a level in the hierarchy, causing a drop in D_{opt}.

The evaluation of R_{opt} also allows us to analyze the capacity requirements of the most economical GMG topologies, which is given by $CAP_{opt}=NR_{opt}$, and plotted in Figure 2c. Notably, the total amount of resources increases supra linearly with the number of PEs, even in the optimal case: larger topologies thus induce a "complexity" penalty.

The values above are based on the conjecture that a GMG topology with N vertices and maximal degree R_{opt} exists. As in practice very few such graphs have been proven to exist, and even fewer have been constructed, these values must be considered as indicators and not as absolute goals. In Section 5 real topologies are compared to the GMG "bound" graph.

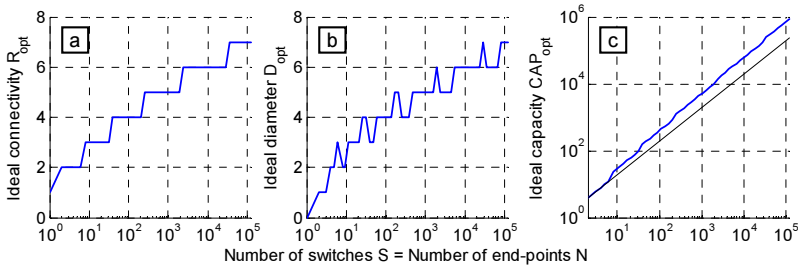

Fig. 2. Connectivity R_{opt} (a), diameter D_{opt} (b) and total capacity CAP_{opt} (c) of the ideal GMG topologies of increasing sizes with concentration factor $C=1$ (one end-point per switch, $S = N$). A strictly linear progression is drawn in (c) for reference.

4 Generalization to other concentrations and traffic patterns

As described so far, our approach indicates the ideal level of connectivity required to obtain a balanced topology under two assumptions: 1) there is only one PE per switch and, 2) the traffic is uniformly distributed among sources and destinations and injected at maximum rate. In this section we show how our approach can be extended to higher concentrations, and to other traffic assumptions.

Having a concentration factor $C>1$ modifies the distances between PEs. Supposing a GMG shaped interconnect, PE has now $(C-1)$ PEs at topological distance[2] 0, CR at distance 1, $CR(R-1)$ at distance 2 etc. The minimum average topological distance between PEs must be rewritten as follows:

$$\Delta_{GMG}(C,R) = \frac{CR + 2CR(R-1) + 3CR(R-1)^2 + \ldots + C(D-1)R(R-1)^{D-2} + CDx}{N-1} \quad (2)$$

[2] Topological distance refers to the number of hops achieved over the topology itself and excludes access and egress hops. A topological distance of 0 reflects the situation where messages are immediately forwarded to their final destination after hitting the first switch.

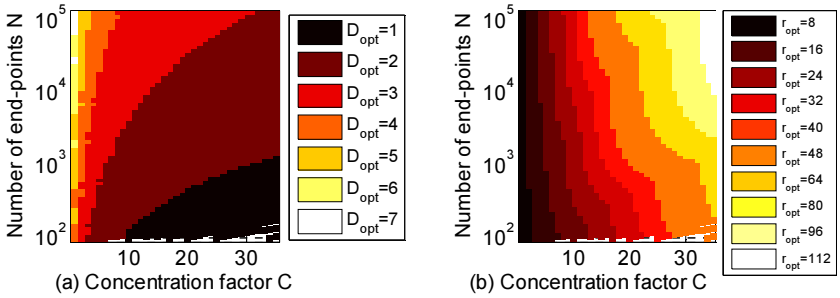

Fig. 3. Dependence of the ideal diameter D_{opt} (a) and of the ideal practical radix r_{opt} (b) on the number of end-points N and the concentration factor C, when considering Generalized Moore Graph topologies.

The total traffic is $N\Delta_{GMG}(C,R)$ and the resulting capacity is now nNR/C. For the same reasons described above, minimum capacity is guaranteed when $n=1$ which leaves us with the following inequality: $\Delta_{GMG}(C,R) \leq R/C$.

Larger concentrations lead to a reduced number of switches, thus to more compact topologies with smaller average topological distance. This in turn reduces the total traffic rate, which eventually translates into weaker requirements in terms of total capacities. Note also that $N(C-1)$ PE pairs have topological distance 0. If the concentration C is equal to N, there is a single switch in the topology so no inter switch links are required at all. One the other hand, large concentrations require the existence of switches with a large number of ports r, as r has to be greater or equal to $C + R$.

R_{opt}, D_{opt} and CAP_{opt} have been evaluated for several values of N up to 100,000, and different concentrations. Figure 3 shows the evolution of D_{opt} (a) and of the required switch radix $r_{opt} = C + R_{opt}$ (b) across this parameter space. To ease the rendering of the figure, and since real switches often have these number of ports, resulting r_{opt} values are rounded up to multiples of 8 up to 48, and to multiples of 16 above 48. Figure 3 shows that if higher concentrations require higher requirements in terms of radix, they also limit the diameter of the topology. As the diameter represents the longest shortest-path in the topology, it is a proxy for the largest zero-load latency. If one desires to maintain this diameter to 2 [13], while achieving ideal connectivity, concentrations larger than 2, 14 and 29 are required for $N=100$, $10{,}000$ and $100{,}000$ respectively.

By superimposing the data of Figures 3a and 3b, one obtains the Pareto front of the largest topologies with the smallest switch radix, for different diameters, as plotted in Figure 4a. To maintain a diameter 2 for $N \geq 100{,}000$ nodes, 96 ports are required. This requirement falls to 32 ports if diameter 3 is acceptable.

If both diameter and radix are crucial indicators of the performance and the technological requirement, they do not fully reflect the cost of the topology, in particular the operational cost (mainly energy) which can be expected to be proportional to the number of links. Figure 4b offers this perspective for three values of N and different concentrations. A fundamental trade-off exists between minimizing the switch radix and the number of links. The final choice of the concentration hence depends on cost

difference between low and high radix switches. There are three zones of interest. 1) Starting from very small r_{opt} values, increases in the first translates to substantial savings in terms of links. This suggests that unless links can be kept low cost at both CAPEX and OPEX levels, radixes of *16*, *24* and *32* at least should be considered for *N≥10,000*, *25,000* and *100,000* respectively. These radixes correspond to the least connected topology with diameter *3*. 2) Then follows a trade-off zone (tinted in the figure) in which capacity savings can be traded against larger radixes, until hitting the least connected topology of diameter *2*. 3) Past this point, an increased radix has little influence on CAP_{tot}. This suggests that building switches offering *48*, *64* or *96* can be taken as a good target for realizing interconnects supporting *10,000*, *25,000* and *100,000* PEs respectively. Under the assumptions considered, larger values will reduce the average distance and therefore the required capacities, but not in the same proportion.

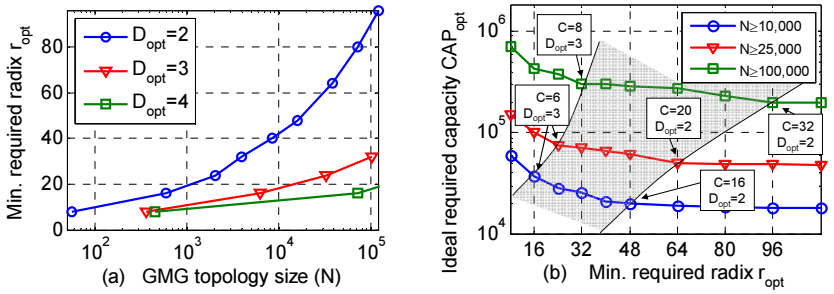

Fig. 4. (a) Evolution of the minimum practical radix r_{opt} required to interconnect N end-points for a given maximum diameter. Ensuring D_{opt}=2 induce high radixes requirements as N scales. Accepting larger diameters allows to decrease the absolute radix needs. (b) Concentration factors can be utilized to obtain different optimal capacity/radix trade-offs. Resulting topologies diversely populate the capacity/required radix Pareto front. Topologies with D_{opt}=*3* (tinted zone) appears as interesting trade-offs.

The approach presented here can also be used to analyze interconnect requirements under different traffic assumptions. Traffic predictions can be linearly reduced by assuming that each PE emits uniform traffic at a fraction of its maximum capacity. If this fraction is noted z, the inequality becomes $z\Delta_{GMG}(C,R) \leq R/C$. One can also examine more subtle traffic scenarios in which traffic flows between closely or remotely located PEs are imbalanced. Instead of estimating the traffic injected by a PE through the average distance, one can suppose that each PEs send a fraction p_i of its traffic to the other PEs located at distance i. The traffic sent by each PE thus become $p_1 + 2p_2 + ... + (D-1)p_{D-1} + Dp_D$ with the conditions that $0 \leq \sum_i p_i \leq 1$ and that each p_i is positive. If $p_1 = \frac{CR}{N-1}$, $p_2 = \frac{CR(R-1)}{N-1}$, ... $p_D = \frac{Cx}{N-1}$, i.e. that *(1/N-1)* of the traffic is sent to the *CR* PEs located at distance 1, to the *CR(R-1)* ones located at distance 2, etc, the traffic is uniform at maximal rate again. Another combination of interest is the one where $p_0 ... p_{D-1} = 0$ and $p_D = 1$. This situation assumes that every PE

sends at maximum rate messages to its most distant peers exclusively. It is therefore the worst-case scenario in terms of total traffic magnitude (adversarial traffic scenario).

Figure 5a and 5b show how these alternative traffic assumptions modify the optimal diameter value for various N and C. For uniform traffic at 50% load and $N=25,000$, diameter *2* topologies are equilibrated capacity wise for concentrations of *30* at least. In presence of adversarial traffic, this number falls to *19*.

Figure 5c compares the switch radix/installed capacity trade-offs across the three traffic cases (50% and 100% uniform, adversarial) for $N \geq 100,000$, which is in general the maximum scale desired for Exascale computers. As pointed out in several studies [3, 13], switches with port counts in this range are currently available, with line rates of several tens of Gb/s. However, Exascale HPC systems will require much larger line rates, in the range of the Tb/s [9]. In this context, the requirement of 100 port switches will become a challenge. As shown in Figure 5c, dimensioning the topology for 50% of the maximum injection rate drastically diminishes the radix requirements. If high radix switches cannot be developed for line rates of 1Tb/s or greater, then smaller radix ones supporting even higher rates may provide an alternative.

In this context, transparent optical switches may be an option. MEMS based switches with more than three hundred ports are already available, however, they suffer from low switching speed. Integrated optics-based ultra-fast switches [2], in contrast, are fundamentally capable of sub-nanosecond switching times. Recent results indicate that 32x32 optical switches based on silicon photonics ring resonators, supporting 20x10Gb/s parallel (WDM) signals, are feasible without recourse to intermediate amplification [1].

More generally, the analysis provided here indicates that realizing a 32-port switch capable of 1Tb/s line rate (32 Tb/s total bandwidth) is a reasonable minimum target for realizing an Exascale-class interconnect.

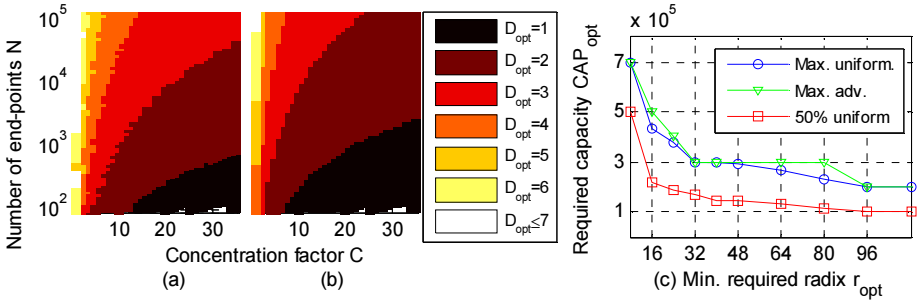

Fig. 5. Impact of the traffic assumption on the ideal diameter D_{opt} for various N and C a) with $z=0.5$ (50% uniform traffic) b) for worse-case traffic. c) Capacity/radix trade-off of GMG topologies supporting $N=100,000$ end-points using various concentration factors and different traffic assumptions

5 Identifying topologies close to the bound

As indicated above, all R_{opt}, CAP_{opt}, D_{opt} or r_{opt} values provided so far have been calculated on the premise that GMGs of any size and degree can be constructed. In reality, most of these GMG topologies either not yet been identified, or have been proven not to exist. In this section we therefore identify a host of topologies whose wiring is known, with adequate connectivity, and compare them to the ideal GMG ones. At the same time, we show how the knowledge of R_{opt} eases this identification process.

In order to find good candidate topologies to interconnect at least *25,000* PEs, able to support maximum uniform traffic, we start with an evaluation of R_{opt} for a range of realistic concentrations (typically 1 to 40), and plot these values against the number of switches corresponding to each concentration, i.e. $\lceil 25{,}000/C \rceil$. These values form the black line on Figure 6.

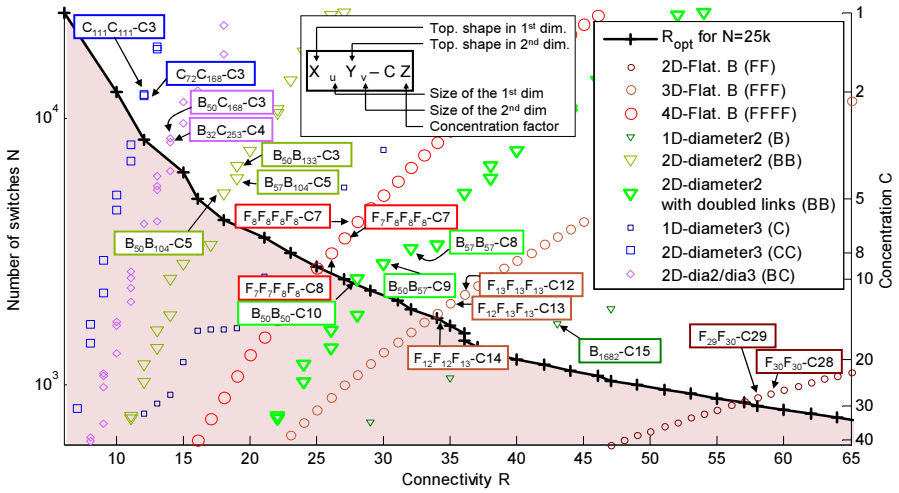

Fig. 6. Connectivity R of topologies of various types and sizes (but all supporting at least *N=25,000* end-points) reported against their number of these switches S (left y-axis) and concentration factor C (right y-axis). The solid curve represents the connectivity R_{opt} of ideal, GMG based topologies, and also delimits the feasibility region. Dots located close to the solid curve indicate potentially good candidate topologies.

The second step consists of evaluating the connectivity *R* of several known topologies such as Flattened Butterflies of different dimensions. We consider the 3D-Flattened Butterfly (3D-FB) as an example: switches are arranged in a 3-dimensional lattice. If concentration *C=1* is employed, this 3D lattice should be at least of size 29x29x30 = *25,230*. Therefore, each switch is connected to $R_{3D\text{-}FB,C=1}$ = *28 + 28 + 29 = 85* other switches. If *C=13*, the lattice is 12x13x13 = 2,038 (2,038 x 13 = 26,494) and switches have a connectivity $R_{3D\text{-}FB,C=13}$ = *11 + 12 + 12 = 35*. Hence, to each topology type (e.g. 3D-Flattened Butterfly) and concentration *C* there corresponds a

connectivity value (for a given N). This connectivity evaluation is shown in Figure 6 - for 2D, 3D and 4D-Flattened Butterflies [3] (denoted as FF, FFF and FFFF), for the largest known graphs of diameter 2 and limited degree (denoted as B) including ones exploited by Slim-fly [13], the largest ones of diameter 3 and limited degree (C), and for a 2-dimensional combination[3] of these largest known graphs (BB, BC and CC). We also included the connectivity of BB topologies whose links have been doubled to obtain a higher connectivity. Toroidal interconnects have not been included in Fig. 6 in an effort to not overload the image. If included, they would appear as vertical lines: for example, a 5D-torus has a constant connectivity R=10 for any number of switches. If the links of this 5D-torus are doubled, the connectivity becomes R=20.

All points located at the left of the R_{opt} curve can be excluded: they show a too weak connectivity which will oblige one to interconnect switches with more than one link ($n>1$). In contrast, practical topologies located close to the optimality curve, but to the right of it, i.e. the ones whose connectivity is slightly higher than the strict required minimum, are of great interest. Obviously, as "real" topologies, one can expect them to show a higher average distance than "virtual" GMG of similar size, but their higher connectivity might be sufficient to compensate this distance penalty.

Finally we further analyze the topologies of greatest interest (those indicated with boxes in Figure 6 plus several tori) by 1) constructing each topology 2) routing individual flows (using shortest-path routing only) over the topology, 3) looking for the most congested connection between two switches, 4) rounding up this congestion value and multiplying it by the number of edges in the topology (for multi-dimensional topologies, we do that for each dimension separately). By this method we obtain the capacity required in practice to secure a complete absorption of the traffic. This also allows us to determine the minimum required switch radix (rounded to multiples of 8 and 16 as previously). These values are represented on Figure 7.

Starting from the right side of the graph, the $F_{29}F_{30}$-C29 fails to approach the bound. The dimension of size 29 is not connected enough which obliges us to double its links to ensure sufficient capacity. This is a general problem with Flattened Butterflies. They are in theory capable of any connectivity R, but unless the lattice is equally sized in all dimensions, this total connectivity is not adequately balanced across the dimensions. In contrast, $F_{30}F_{30}$-C28 closely approximates the bound, but requires a large radix of *96*. The Slim-fly topology B_{1682}-C15 is the next Pareto dominant datapoint. Similarly to the 2D-FB, it is a diameter two topology, but due to its close to optimal wiring, it requires smaller radix than the 2D-FB. It lies ahead of the bound, however. As visible on Figure 6, there are relatively few known large graphs of diameter 2 (among them the MMS ones exploited by Slim-fly) which obliged us to consider a topology of higher connectivity than strictly required (43 instead of 35). This explains the distance to the bound. The $F_{13}F_{13}F_{13}$-C12 is also Pareto dominant, followed by the $B_{57}B_{57}$-C8. Although the base graph B_{57} is the largest one found so

[3] In the Flattened Butterfly topology, all vertices sharing a dimension in the lattice as interconnected (Full-Mesh). In a torus, all these vertices are connected along a ring. In our 2-dimensional construction, vertices sharing a dimension are interconnected by following the structure of the largest known graph for a given diameter and maximum degree.

far for degree 8, by using it in two dimensions *and* by doubling its edge we diluted its close to optimum properties. Still, it appears as a Pareto dominant topology. The 4D Flattened Butterfly can be realized with the same radix of 40, and for a slight increase in the capacity requirements. As routing and workload mapping is probably made easier with this topology than with the $B_{57}B_{57}$ one, it might also be considered, although not strictly part of the Pareto front. After the 4D-FB, the next Pareto dominant topology is the $B_{50}B_{133}$-C3 but as it lie far from the bound, it shows little value unless large radixes cannot be utilized. More generally, the Pareto front created by Pareto dominant topologies progressively diverges from the bound. This reveals a need for topologies close to the GMG bound, of diameter 3 and larger, and of diverse sizes. These topologies are also harder to emulate with symmetric constructions as the Flattened Butterfly or tori.

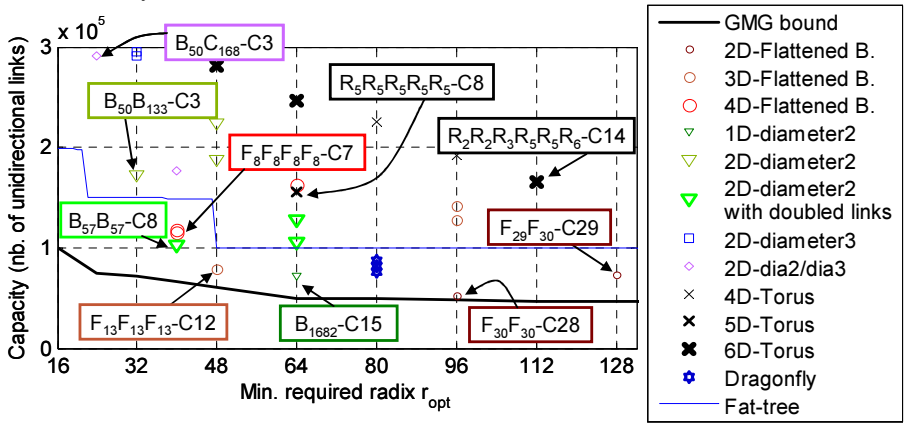

Fig. 7. Capacity/radix trade-off of practical topologies supporting at least 25,000 PEs.

Two additional datapoints have been highlighted on Fig. 7. They reflect topologies used in current dominant Supercomputers. $R_5R_5R_5R_5R_5$-C8 is a 5D-torus as the one available in Sequoia, while $R_2R_2R_3R_5R_5R_6$-C14 mimics the Tofu interconnect [38] (6D-torus) of the K computer. $R_5R_5R_5R_5R_5$-C8 is the best 5D-torus for N=25,000 maximal and uniform traffic. It dominates all other combinations of sizes and concentrations leading to realistic radixes. Still, its average distance Δ (just above 6) is more than three times larger than $\Delta_{GMG}(C=21,R=42)$ = 1.963 which would also require a radix of 64. This directly translate to greater than 3 times the capacity requirements (156,250, where the bound indicates 50,022). Other tori, including the best Tofu-like 6D-torus, lie even further from the bound, and no 3D-tori appears in the ranges of Fig. 7. This demonstrates that tori are fairly ill-suited to traffic conditions similar to uniform traffic at large scales. However, as mentioned in the introduction, tori can compensate this hindrance by providing bandwidth in a very structured way, a feature that is generally vastly exploited by workloads.

Although not directly related to our approach, fat-trees with no over-subscription (i.e. full-bisectional bandwidth) can also be characterized by a number of links and radix, and therefore be represented in Fig. 7. For all radixes, a fat-tree involves gener-

ally twice as many links than the bound. They are clearly outperformed by the Pareto optimal topologies, expect for small radixes. However, as pointed out above, other close to ideal topologies of larger diameter (not investigated here) might exist and outperform the fat-tree for smaller radixes, too.

6 Discussion and Future Directions

The methodology described here does not cover all facets of the topology selection process. The hypothesis that all links are equivalent, and that traffic is uniformly or arbitrarily distributed, excludes the Dragonfly topology which precisely assumes imbalanced traffic flows. All aspects related to embedding the topology onto backplanes and inter-rack/cabinet cables is also neglected although this may play an important role in the process of choosing one topology over another. The goal of this study is, however, not to provide a final choice of a topology but rather to define a framework allowing us to explore the main trade-offs and minimal requirements. More exhaustive comparisons will be realized in future publications. Still, Fig. 7, covering Tori, Fat-trees, Flattened Butterflies and Slim-flies of various sizes and shape, provide a rather inclusive set of "realistic" topologies that populate current Supercomputers.

No traffic simulations have been conducted on the practical topologies compared in the previous Section. However, if such simulations would be driven with uniform traffic, they would mainly confirm that no surge in latency occurs for loads < 1, as all topologies are precisely dimensioned for this aim. In contrast, simulating real workloads in large-scale architectures equipped with the different Pareto optimal topologies would be of high-interest. In particular, we would expect such experiments to confirm the statement that limited diameters are highly desirable. Our future plans include realizing such experiments.

Our approach also allows us to derive further insights on practical topologies. In results not shown here we have found that it allows one to identify how improper wiring, connectivity that is too large or too small, and an unfavorable number of endpoints each contribute to making a particular topology sub-optimal.

7 Conclusion

In this study, we concentrate on the relationship between traffic, capacities (i.e. number of links) and switch radixes, and do not deal with other aspects such as maintenance, organization in racks, incremental deployment, load balancing, etc. Surprisingly, even in this reduced scope, there is no clearly dominant recipe for building a topology. We conclude that the choice of the topology should not be made too early in the design process, and not on the sole assumption that one topology or another has been proved optimal in a given context. A final choice among Pareto dominant data points of Figure 7 (i.e. 5-10 options) can be made after comparison under real traffic with large scale simulators, or based on a detailed economical analysis once the costs of links and switches are known.

Results also show that if 2D-Flattened Butterflies and Slim-flies, both of diameter 2, land close to the bound, approaching the bound with larger diameters, corresponding to tighter radix availabilities, appears trickier. Provided that wide radixes might be hard to adapt to larger line-rates, there is a need for topologies of diameter 3 and 4, not necessarily of largest size, but showing close to optimal average routing distance. Similar topologies of diameter 2 would offer a good complement to the Slim-fly set which appears too scattered to adequately support the whole range of system sizes.

This study also shows that tori appear sensibly dominated by other topologies, although they account for a vast portion of modern Supercomputer interconnects. This suggests that for now workload structures can be mapped reasonably well over their hardware equivalent. However, with the advent of asynchronous parallel programming method, or adaptive workload balancing, traffic profiles may loose their structure. In this case, topologies as close to a GMG as possible seems the most logical choice.

Acknowledgement. This work has been realized in the context of Department of Energy (DoE) ASCR project "Data Movement Dominates". It has been partly supported by the U.S. Department of Energy (DoE) National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) program through contract PO1426332 with Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Appendix

A Generalized Moore Graph can be described as follows. Consider a vertex, *V*, in any graph of degree *R* (i.e whose vertices have never more than R incident links). *V* cannot have *more* than *R* direct neighbors. It also cannot have *more* than *R(R-1)* neighbors at distance *2* (each of its neighbors have *R* neighbors but *V* does not count as it is one of them), and generally cannot have *more* than $R(R-1)^{D-1}$ neighbors at distance *D*. A GMG is a graph which maximally uses this expansion possibilities offered by the degree *R*: in a GMG graph, each vertex has *exactly R* direct neighbors, *exactly* $R(R-1)^{i-1}$ neighbors at distance *i* ($i = 2..D-1$), and all the remaining vertices are at distance D. Figure 8 exemplifies the GMG concept. Because inner layers are maximally filled, there is no way to get a vertex closer without interchanging it with another vertex. This means that no distance between two vertices can be reduced, thus that the average distance in the graph is minimized.

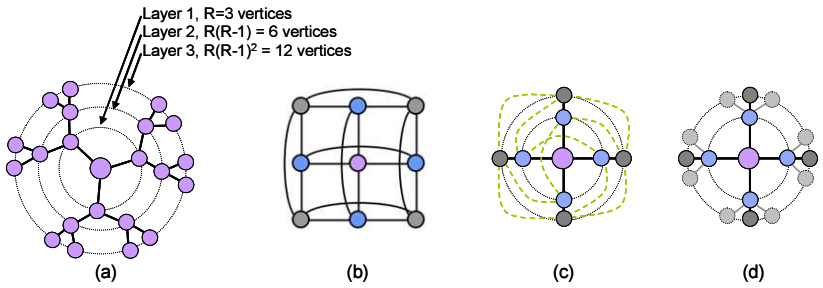

Fig. 8. a) Maximal expansion possibilities for connectivity/degree *R=3* and three layers. Generalized Moore Graphs follow this structure, except that the last layer does not have to be totally filled b) Example of Generalized Moore Graph (a 3x3 torus) c) The 3x3 torus reorganized to show the layers d) A representation of unfilled slots in the last layer

References

1. D. Nikolova, S. Rumley, D. Calhoun, Q. Li, R. Hendry, P. Samadi, K. Bergman, "Scaling silicon photonic switch fabrics for data center interconnection networks", Optics Express 23(2), 1159-1175 (2015).
2. B. G. Lee, N. Dupuis, P. Pepeljugoski, L. Schares, R. Budd, J.R. Bickford, C.L. Schow, “Silicon Photonic Switch Fabrics in Computer Communications Systems”, IEEE Journal of Lightwave Technology, in press.
3. J. Kim, W. J. Dally, D. Abts, "Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks", In Proc. of the International Symposium on Computer Architecture (ISCA), 126-137 (2007).
4. A. Bhatelé, L. V. Kalé, "Benefits of topology aware mapping for mesh interconnects", Parallel Programming Letters, 18(4), 549-566 (2008).
5. W. J. Dally, "Principles and Practices of Interconnection Networks", Morgan Kaufmann, 2004.
6. C.E. Leiserson, "Fat-trees: Universal networks for hardware-efficient supercomputing", IEEE Transactions on Computers, C-43(10), 892-901 (1985).
7. K. Sano, "Interconnection Network: Design Space Exploration of Networks for Supercomputers", Sustained Simulation Performance, Springer, 151-161, 2015.
8. S. Borkar, "Role of Interconnects in the Future of Computing", IEEE Journal of Lightwave Technology (JLT) 31(24), 3927-3933 (2013)
9. S. Rumley, et al. "Silicon Photonics for Exascale Systems", IEEE Journal of Lightwave Technology (JLT), 33(3), 547-562 (2015).
10. M. Bradonjic, I. Saniee, I. Widjaja, "Scaling of Capacity and Reliability in Data Center Networks", In Proc. SIGMETRICS, 2014.
11. G. Faanes, et al. "Cray cascade: a scalable HPC system based on a Dragonfly network", In Proc. of the International Conference on High Performance Computing, Networking Storage and Analysis (SC'12), (2012).

12. A. Singla, C.-Y. Hong, L. Popa, P.B. Godfrey. "Jellyfish: Networking data centers randomly". In Proc. of the USENIX Symposium on Networked Systems Design and Implementation (NSDI'12), 2012.
13. M. Besta, T. Hoefler, "Slim fly: a cost effective low-diameter network topology", In Proc. of the International Conference on High Performance Computing, Networking Storage and Analysis (SC'14), (2014).
14. F. P. Preparata, J. Vuillemin, "The cube-connected cycles: a versatile network for parallel computation", Communications of the ACM 24(5), 300-309, (1981).
15. L. N Bhuyan, D.P. Agrawal, "Generalized Hypercube and Hyperbus Structures for a Computer Network", IEEE Transactiosn on Computers C-33(4), 323-333 (1984).
16. M. C. Pease, "The Indirect Binary n-Cube Microprocessor Array", IEEE Transactions on Computers C-26(5) (1977).
17. W. J. Dally, "Performance Analysis of k-ary n-cube Interconnection Networks", IEEE Transactions on Computers 39(6), 775-785 (1990).
18. V.E. Benes, "Optical rearrangeable multistage connecting networks", Bell System Technical Journal 43(4), 1641-1656 (1964).
19. L. D. Wittie, "Communication Structures for Large Networks of Microcomputers", IEEE Transactions on Computers C-30(4), 264-273, (1981).
20. J. Kim, W. J. Dally, B. Towles, A.K. Gupta, "Microarchitecture of a high radix router", In Proc. of the International Symposium on Computer Architecture (ISCA), 420-431 (2005).
21. S. Scott, D. Abts, J. Kim, W. J. Dally. "The BlackWidow High-radix Clos Network", In Proc. of the International Symposium on Computer Architecture (ISCA), 16-28 (2006).
22. R. Barriuso and A. Knies. "108-Port InfiniBand FDR SwitchX Switch Platform Hardware User Manual" (2014).
23. K. Li, Y. Mu, K. Li, G. Min, "Exchanged Crossed Cube: A Novel Interconnection Network for Parallel Computation", IEEE Transactions on Parallel and Distributed Systems (TPDS) 24(11), 2211-2219 (2013).
24. C. Von Conta, "Torus and Other Networks as Communication Networks With Up to Some Hundred Points", IEEE Transactions on Computers 32(7), 657-666 (1983).
25. S.B. Akers, B. Krishnamurthy, "A group-theoretic model for symmetric interconnection networks", IEEE Transactions on Computers 38(4), 555-566 (1989).
26. S.-Y. Hsieh, T.-T. Hsiao, "The k-valent Graph: A New Family of Cayley Graphs for Interconnection Networks", In Proc. of the International Conference on Parallel Processing (ICPP) (2004).
27. B. D. McKay, M. Miller, J. Sirán, "A note on large graphs on diameter two and given maximum degree", Journal of Combinatorics 61, 1-63 (1998).
28. M. Miller, J. Sirán, "Moore graphs and beyond: A survey of the degree/diameter problem", Electronic Journal of Combinatorics, Dynamic Survey D 14 (2005).
29. W.-T. Boa, et al. "A High-Performance and Cost-Efficient Interconnection Network for High-Density Servers", Journal of computer science and Technology 23(2) (2014).

30. M. Sampels, "Large Networks with Small Diameter", Graph-Theoretic Concepts in Computer Science, Springer LNCS 1335, 288-302 (1997).
31. E. Loz, J. Sirán, "New record graphs in the degree-diameter problem", Autralasian Journal of Combinatorics 41, 63-80 (2008).
32. M. Koibuchi, H. Matsutani, H. Amano, D. F. Hsu, H. Casanova, "A Case for Random Shortcut Topologies for HPC Interconnects", In Proc. of the International Symposium on Computer Architecture (ISCA), 177-188 (2012).
33. K. C. Guan, V. W. S. Chan, "Cost-Efficient Fiber Connection Topology Design for Metropolitan Area WDM Networks", IEEE/OSA Journal of Optical Communication Networks 1(1) (2009).
34. J. Vetter, et al. "Quantifying Architectural Requirements of Contemporary Extreme-Scale Scientific Applications", in High Performance Computing Systems. Performance Modeling, Benchmarking and Simulation, Springer LNCS (2014).
35. S. Kandula, S. Sengupta, A. Greenberg, P. Patel, R. Chaiken, "The nature of data center traffic: measurements and analysis", in Proc. of the ACM conference on Internet measurement (IMC), 202-208 (2009).
36. G. Hendry, "The Role of Optical Links in HPC System Interconnects", In Proc. of the IEEE Optical Interconnects Conference (2013).
37. S. Hammond, et al. "Towards a standard architectural simulation framework", In Proc. of the Workshop on Modeling & Simulation of Exascale Systems & Applications, (2013).
38. Y. Ajima, T. Inoue, S. Hiramoto, T. Shimizu, "Tofu: Interconnect for the K computer", Fujistu Sci. Tech. Journal, 48(3), 280-285, 2012.