

Projective Networks: Topologies for Large Parallel Computer Systems

Cristóbal Camarero, Carmen Martínez, Enrique Vallejo, and Ramón Beivide
Department of Computer Science and Electronics
Universidad de Cantabria, UNICAN, Spain

Abstract

The interconnection network comprises a significant portion of the cost of large parallel computers, both in economic terms and power consumption. Several previous proposals exploit large-radix routers to build scalable low-distance topologies with the aim of minimizing these costs. However, they fail to consider potential unbalance in the network utilization, which in some cases results in suboptimal designs. Based on an appropriate cost model, this paper advocates the use of networks based on incidence graphs of projective planes, broadly denoted as Projective Networks. Projective Networks rely on generalized Moore graphs with uniform link utilization and encompass several proposed direct (PN and demi-PN) and indirect (OFT) topologies under a common mathematical framework. Compared to other proposals with average distance between 2 and 3 hops, these networks provide very high scalability while preserving a balanced network utilization, resulting in low network costs.

1 Introduction

One current trend in research for the design of Exascale systems is to greatly increase the number of compute nodes. The cost and power of the network of these large systems is significant, which urges to optimize these parameters. Specifically, the problem is how to interconnect a collection of compute nodes using a given router model with as small cost and power consumption as possible. If the interconnection network is modelled by a graph, where nodes represent the routers and edges the links connecting them, the Moore bound [31] applies. This bound establishes the maximum network size for a given diameter k . The present paper deals with graphs attaining or approaching the generalized Moore bound [34]—which bounds the average distance for a given network size—while minimizing cost and power consumption.

Graph theory has dealt with very interesting topologies that have not yet been adopted as interconnection networks. One paradigmatic example are Moore graphs [31]. Hoffman and Singleton provided in [21] some few examples of regular graphs of degree Δ and diameter k having the maximum number of vertices; namely for $k = 2$ and $\Delta = 2, 3, 7$ and for $k = 3$ and $\Delta = 2$. They denoted such graphs as *Moore graphs* as they attain the upper bound for their number of

nodes, solving for these cases, the (Δ, k) -problem posed by E. F. Moore. Such graphs are optimal for interconnection networks as they simultaneously minimize maximum and average transmission delays among nodes.

In these interconnection networks, traffic is frequently uniform; when it is not, it can be randomized (using Valiant routing, [37]). Under uniform traffic, maximum throughput depends on the network average distance \bar{k} , rather than the diameter k . This promotes the search of generalized Moore graphs [34], which have minimum average distance for a given degree. This is attained when, from a given node, there are the maximum amount of reachable nodes at any distance lower than the diameter, with the remaining nodes at distance k .

As it will be shown in this paper, Moore and some generalized Moore graphs also minimize cost. With network cost dominated by the number of ports (SerDes in particular), minimizing graph average distance not only maximizes throughput but it can also minimize investment and exploitation expenses. Nevertheless, it is important to highlight that symmetric graphs [8] could be preferable as they guarantee the absence of bottlenecks that can compromise performance under uniform traffic. This paper shows examples of such topologies based on incidence graphs of projective planes and compares them with competitive alternatives. Incidence graphs of finite projective planes [7], [16] have been used to attain the Moore bound, but not only mathematicians have paid attention to this discrete structures. In fact, Valerio *et al.* already use them to define Orthogonal Fat Trees (OFT) [36], which are highly scalable cost optimal indirect networks. Brahme *et al.* [5] propose other topologies for direct networks for HPC clusters. Although the authors used perfect difference sets for their definition, it is shown in this paper that they can also be defined using projective planes. In this paper it is shown how incidence graphs of finite projective planes are suitable topologies for both direct and indirect networks for HPC systems.

Recently, three strongly related papers have been published. We summarize next their main achievements and bring to light how the results introduced in this paper improve them. In [33], a methodology based on minimizing average distance is proposed to identify optimal topologies for Exascale systems. Therefore, topologies close to the generalized Moore bound are searched, in particular graph products of known topologies. However, neither symmetry nor

link utilization are considered in this analysis and, therefore, their study omits some performance limitations of their proposal. In [2] the Slim Fly (SF) network is proposed. This topology provides very high scalability for diameter 2, approaching the Moore bound. However, SF is not symmetric and its links do not receive a balanced load. Therefore, the number of compute nodes per router must be adjusted in order to avoid oversubscription. Moreover, this lack of symmetry makes SFs more costly than projective networks with the same diameter, which also provide higher scalability. Finally, in [24] several diameter 2 topologies are studied: Stacked Single-Path Tree, Multi-layer Full-Mesh, Slim Fly and Orthogonal Fat Tree. Experimental results conclude that the Slim Fly and the OFT are the best direct and indirect topologies respectively. The present paper proves that topologies with diameter other than 2 such as projective networks are also interesting. Furthermore, a more accessible construction of the OFT and its relation with other topologies is given.

The rest of the paper is organized as follows. Section 2 introduces a cost model based on average distance and link utilization. The target will be to maximize the number of terminals with minimum average distance and balanced link utilization, which is related to the generalized Moore bound. In Section 3 *Projective Networks* are introduced, defined using incidence graphs of projective planes, with the smallest average distance for their size and balanced link utilization. In Section 4 a thorough analysis of how graph theoreticians have solved the generalized Moore bound for diameters 1–6 is done. This allows to present a complete comparison, in terms of our power/cost model, of all these topologies in Section 5, with special emphasis on the diameter 2 case. In Section 6 the case for indirect networks is considered. The cost model is adapted for indirect networks of diameter 2. As it will be shown, optimal topologies can also be obtained with our methodology to derive projective networks. Finally, in Section 7 the main achievements of the paper are summarized.

2 Power and cost optimization

The interconnection network constitutes a significant fraction of the cost of an High Performance Computing (HPC) or datacenter system, both in terms of installation and operation (dominated by energy costs). This section introduces a coarse-grain generic cost model based on the network average distance and average link utilization. This cost model will be employed to compare different topologies in next sections.

The network should provide the required bandwidth to the compute nodes with minimal latency, while scaling to the desired size. Measures of interest are throughput and average latency under uniform traffic. This uniform traffic not only determines the topological properties of the network, but also appears in multiple workloads (such as data-intensive applications or in many collective primitives) and

Parameter	Definition
T	Number of compute nodes or terminals.
R	Router radix (number of ports).
$G(V, E)$	Graph whose vertices V represent the routers and its edges E the connection between routers.
$N = V $	Number of routers.
Δ	Maximum degree of G .
Δ_0	Number of compute nodes attached to every router.
k	diameter of G .
\bar{k}	Average distance of G .
a	Load accepted by each router in saturation.
u	Average utilization of links.

Table 1: Notation used in the paper.

determines worst-case performance when using routing randomization [37].

An important figure in the deploying of a network is the number of ports in each router chip, also called router radix. This number is a technological constraint, and current 100 Gbps designs typically only support 32 to 48 ports [6, 30, 13, 23]. Different configurations of these switches provide more than a hundred ports but at lower speeds, typically 25 Gbps. Larger non-blocking routers are built employing multiple routing chips, at the cost of an increased complexity and at least triple switching latency [29, 22].

Thus, our goal will be to build a network for T computing nodes using routers of radix R , able to manage uniform traffic at the maximum injection rate of the nodes and minimizing its cost. Therefore, the use of the expression *optimal network* along this document refers to this optimization problem.

The notation used throughout the paper is presented in Table 1. For simplicity, all links are assumed to have the same transmission rate. Δ is employed to refer to the degree of a graph G ; when G is a Δ -regular graph, $2|E(G)| = N\Delta$. Similarly, Δ_0 is generally equal to all routers; in such case the router radix is $R = \Delta + \Delta_0$ and the number of compute nodes $T = N\Delta_0$.

2.1 Network Dimensioning and Cost Model

In this subsection a generic cost model for both power and hardware required by a direct network is introduced. An adaption of this model to indirect networks is explored in Section 6. This cost depends not only on the average distance of the topology, but also on the average utilization of the network links. Previous works such as [2, 33] do not consider the average link utilization in their calculations, what leads to suboptimal results, as it will be proved in Section 5.

First, the number of compute nodes Δ_0 which can be serviced per router is estimated. In this aim, *ideal routers* with minimal routing and a uniform traffic pattern will be assumed. As the load a increases, the saturation point is reached when some network link becomes in use all the cycles. When this happens, the network links will have an

average utilization $u \in (0, 1]$. If $u = 1$ then G is said *well-balanced*. Being G *edge-transitive* is a sufficient but not necessary condition to be well-balanced [8]. If the load injected per cycle per router at saturation is a ,¹ then the average utilization u is

$$u = \frac{\text{load}}{\#links} = \frac{aN\bar{k}}{2|E(G)|} = \frac{a\bar{k}}{\Delta}.$$

The load in terms of the utilization is $a = \Delta \frac{u}{\bar{k}}$. Therefore, the number of compute nodes per router Δ_0 which can be serviced without reaching the saturation point is:

$$\Delta_0 \leq \Delta \frac{u}{\bar{k}}. \quad (1)$$

Ideally, the equality should hold. If (1) does not hold, the network is said to be *oversubscribed*, what restricts maximum throughput under uniform traffic. In an oversubscribed network, the maximum link load is 1 and the average link load does not change respect to the case in which the equality holds. Conversely, for Δ_0 lower than the equality value, the network is oversized for the number of compute nodes connected, or *undersubscribed*. In this case, the network does not reach the saturation point.

Now, a generic estimation for the network cost per computing node C_{node} is considered, which is also particularized to economic or power terms ($C_{node-\$}$ and C_{node-W} in \$ and Watts, respectively). Generic average cost c_i per injection port, c_t per transit port, and c_r per router are assumed. The resultant cost per compute node is

$$C_{node} = \frac{N}{T} \cdot (c_i \Delta_0 + c_t \Delta + c_r) = \frac{c_i N \Delta_0 + c_t N \Delta + c_r N}{T}.$$

Considering the equality value in (1), $T = N \Delta_0$ and $R = \Delta + \Delta_0$, it results:

$$C_{node} = c_i + c_t \frac{\bar{k}}{u} + c_r \frac{1 + \bar{k}/u}{R}. \quad (2)$$

For the installation cost $C_{node-\$}$, router and transit links comprise the largest amounts. The router cost is roughly proportional to the number of ports, so it contributes a large amount to c_i, c_t and a small amount to c_r [2]. Regarding links, as network speed increases optics are expected to displace copper for even shorter distances, including both intra-rack and on-board communications [14]. When all network links are active optical cables their cost is largely independent of their length, since it is dominated by the optical transceivers in the ends. This leads to $c_i = c_t \gg c_r$, with $c_i = c_t$ approximately constant. Therefore, the largest component of the installation cost in (2) will be determined by the router ports, $C_{node-\$} \approx c_t(1 + \frac{\bar{k}}{u})$. A more detailed analysis considering different types of cables is presented in Section 5.

¹All routers are assumed to inject approximately the same load at saturation.

For the energy cost C_{node-W} , the most significant part are the router SerDes (which imply large c_i, c_t and small c_r); for example, the router design in [10] dedicates 87% of its power to SerDes. Again, this leads to the same result as for the installation cost, concluding that the best cost is obtained using topologies that minimize $\frac{\bar{k}}{u}$.

2.2 Moore Bounds

In this subsection limits of the network size and its cost will be studied. This will be done by considering the limits of the Moore bound for the relation between the diameter and network size, and the generalized Moore bound for the relation between the average distance and network size, both for a given degree.

Section 2.1 concludes that cost depends linearly on $(1 + \bar{k}/u)$. This expression is minimized in the complete graph K_N , which is symmetric—hence $u = 1$ —and has minimum average distance $\bar{k} = 1$. However, the complete graph has $\Delta_0 = N$, $R = 2N - 1$ and $T = N^2 = (\frac{R+1}{2})^2$. With a radix $R = 48$ the number of compute nodes would be only $T \approx 576$ nodes.

The Moore Bound [31] establishes that for a given diameter k the maximum network size for $\Delta \geq 3$ is bounded by

$$N \leq M(\Delta, k) = \frac{\Delta(\Delta - 1)^k - 2}{\Delta - 2}. \quad (3)$$

This bound requires the following distance distribution—the number $W(t)$ of vertices at distance t from any chosen vertex:

$$W(t) = \begin{cases} 1 & \text{if } t = 0 \\ \Delta(\Delta - 1)^{t-1} & \text{otherwise.} \end{cases}$$

Therefore, the average distance of a Moore graph is

$$\bar{k} = \frac{\sum_{t=1}^k tW(t)}{N - 1} = \frac{\sum_{t=1}^k \Delta(\Delta - 1)^{t-1}}{N - 1}.$$

Then, it is straightforward that $\lim_{\Delta \rightarrow \infty} \bar{k} = k$. There are good families of graphs approaching the Moore bound for low diameter, but they are restricted to very specific values in the number of nodes. Additionally, as derived from (2), the most important factor to minimize cost is the average distance \bar{k} , not the network diameter.

Generalized Moore graphs [34] reach the minimum average distance for a given router radix and number of vertices N . This is attained when there are the maximum amount of reachable nodes up to distance $k - 1$, with the remaining nodes being at distance k . That is, with the following distance distribution:

$$W(t) = \begin{cases} 1 & \text{if } t = 0 \\ \Delta(\Delta - 1)^{t-1} & \text{if } 1 \leq t \leq k - 1 \\ N - M(\Delta, k - 1) & \text{if } t = k. \end{cases}$$

With this generalization, the average distance can be approximated—for large Δ —as

$$\bar{k} \approx k - \frac{\Delta^{k-1}}{N}. \quad (4)$$

The generalized Moore bound determines the minimal average distance \bar{k} (hence cost, given a well-balanced topology) for a given number of nodes T and router radix R . Next, an expression relating these values and the diameter k is obtained. Following (1), there are $\Delta_0 = \Delta/\bar{k} = (R - \Delta_0)/\bar{k}$ compute nodes per router. Thus, $R = \Delta_0\bar{k} + \Delta_0 = \Delta_0(1 + \bar{k})$ and

$$\Delta_0 = \frac{R}{\bar{k} + 1}.$$

The degree is

$$\Delta = R - \Delta_0 = R \left(1 - \frac{1}{\bar{k} + 1}\right) = R \frac{\bar{k}}{\bar{k} + 1}.$$

The number of routers is

$$N = \frac{T}{\Delta_0} = \frac{T}{R}(\bar{k} + 1).$$

The difference $k - \bar{k}$ can be approximated using (4) by

$$k - \bar{k} \approx \frac{\Delta^{k-1}}{N} = \frac{\left(R \frac{\bar{k}}{\bar{k} + 1}\right)^{k-1}}{\frac{T}{R}(\bar{k} + 1)} = \frac{R^k}{T} \frac{\bar{k}^{k-1}}{(\bar{k} + 1)^k}.$$

Reordering terms, it is obtained the relation:

$$T \approx \frac{R^k \bar{k}^{k-1}}{(k - \bar{k})(\bar{k} + 1)^k} \quad (5)$$

This equation is used later as an upper bound for the number of compute nodes in direct topologies.

3 Projective Networks: A Topology Based on Incidence Graphs of Finite Projective Planes

In this section incidence graphs of finite projective planes are proposed to define network topologies attaining almost optimal values of average distance and average link utilization. These have been identified in the previous section as the target parameters to design optimal cost topologies. In Subsection 3.1 incidence graphs of finite projective planes are defined, which constitute a family of symmetric graphs with diameter 3 and average distance equal to 2.5 in the limit. In Subsection 3.2 such graphs are modified in such a way that their diameter and average distance both become 2. However, they are no longer symmetric although their link utilization equals 1 in the limit. These two families of graphs are used to define *Projective Networks* which, as it will be shown in Subsection 5.2, result in a competitive alternative to the recently proposed Slim Fly network [2]. Thus, in this section the methodology proposed in the paper is validated by a specific example.

3.1 Incidence Graph of Finite Projective Planes

A family of graphs with an average distance tending to 2.5 can be obtained as the incidence graph of finite projective planes. Next, an algorithmic description of these graphs is given, although a more geometrical approach is considered in Example 3.3. Since these graphs are defined in terms of finite projective planes, let us first introduce this concept.

A finite projection plane is a set of points and lines that contain these points, which obey certain relations. Let q be a power of a prime number. A canonic set of representatives of the finite projective plane over the field with q elements \mathbb{F}_q is

$$P_2(\mathbb{F}_q) = \{(1, x, y), (0, 1, x), (0, 0, 1) \mid x, y \in \mathbb{F}_q\}.$$

Remark 3.1. By a straightforward counting argument, it can be proved that $P_2(\mathbb{F}_q)$ has $q^2 + q + 1$ elements.

Two points $X, Y \in P_2(\mathbb{F}_q)$ are said *orthogonal* (written $X \perp Y$) if their scalar product is zero. The space $P_2(\mathbb{F}_q)$ contains also $q^2 + q + 1$ lines of exactly $q + 1$ points each. For every line, there is a unique point L which is orthogonal to all the points of the line. The coordinates of this point will be used to represent the line. Therefore, a line is incident to a point P (L contains P) if and only if P is orthogonal to L . This fact allows the definition of derived graphs based on the incidence relations in the finite projective plane, as follows.

Definition 3.2. Let q be a power of a prime number. Let $G_q = (V, E)$ be the graph with vertex set and edges set

$$V = \{(s, P) \mid s \in \{0, 1\}, P \in P_2(\mathbb{F}_q)\}$$

$$E = \{\{(0, P), (1, L)\} \mid P \perp L, P, L \in P_2(\mathbb{F}_q)\}.$$

G_q is said to be the incidence graph of the finite projective plane $P_2(\mathbb{F}_q)$.

It is clear that G_q has $2q^2 + 2q + 2$ vertices. Let us consider the following example to better understand this construction.

Example 3.3. In Figure 1 two different structures are represented. On the left side, a typical graphical representation of $P_2(\mathbb{F}_2)$, or the Fano plane, is shown. In this representation, both the 7 points and their incident lines of the Fano plane are labeled with their homogeneous coordinates. Note that the point 100 is incident to the line 001 since the scalar product of their coordinates is zero, that is, they are orthogonal. On the right side of the figure, a graphical representation of the incidence graph of the Fano plane, denoted by G_2 , is shown. There are two kinds of vertices, which are the points (left) and the lines of the Fano plane (right). Now, two vertices are adjacent if the corresponding point and line are incident. Therefore, since point 100 is incident to line 001 as we have seen before, in the graph there is an edge making them adjacent vertices. The other incident points to

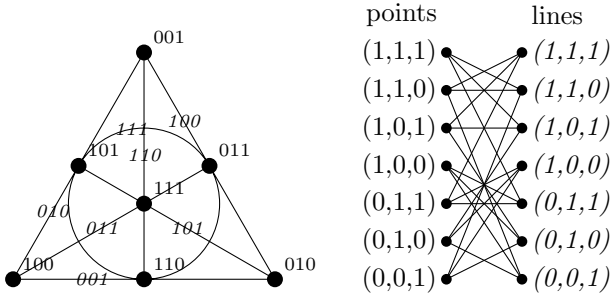


Figure 1: Left: the projective plane $P_2(\mathbb{F}_2)$, also known as the Fano plane. Right: the incidence graph G_2 , also known as Heawood graph.

this line, which are 110 and 010, are also neighbours to this line in the graph. As it can be seen, every vertex has degree 3 and there are minimal paths of lengths 1, 2 or 3.

Distance distribution, routing and symmetry in G_q are considered next. It is known that for any two different points $X, Y \in P_2(\mathbb{F}_q)$ there is a unique $Z \in P_2(\mathbb{F}_q)$ such that $X \perp Z$ and $Z \perp Y$. This implies that the half of the vertices $(0, X)$ of G_q are at distance 2 from $(0, (1, 1, 1))$ and the other half are at distance at most 3. $P_2(\mathbb{F}_q)$ also satisfies that there are $q + 1$ orthogonal points to any given one. Thus, in general G_q is a bipartite graph of degree $\Delta = q + 1$ with distance distribution

$$W(t) = \begin{cases} 1 & \text{if } t = 0 \\ q + 1 & \text{if } t = 1 \\ q^2 + q & \text{if } t = 2 \\ q^2 & \text{if } t = 3. \end{cases}$$

As a consequence, the average distance of G_q is

$$\bar{k} = \frac{5q^2 + 3q + 1}{2q^2 + 2q + 1} = 2.5 - \frac{2q + 1.5}{2q^2 + 2q + 1}.$$

Thus, the limit of \bar{k} is 2.5 and its diameter $k = 3$. The following lemma describes the minimal path calculation over these graphs.

Lemma 3.4 (Routing in G_q). *Let (s_1, P) and (s_2, Q) be a pair of vertices of G_q . Then,*

- If $s_1 = s_2$ then the unique path that joins them is $(s_1, P), (1 - s_1, P \times Q), (s_2, Q)$.
- If $s_1 \neq s_2$ then, or they are neighbours, or for any neighbour (s_2, U) of (s_1, P) , the path $(s_1, P), (s_2, U), (s_1, U \times Q), (s_2, Q)$.

Moreover, it can be proved that G_q is symmetric, which gives the optimal average link utilization.

Theorem 3.5. G_q is symmetric.

Proof. For any invertible matrix $M \in \mathcal{M}_3(\mathbb{F}_q)$, the application that maps the point P to the point MP is an automorphism of the projective plane $P_2(\mathbb{F}_q)$, since it maps subspaces to subspaces. As they preserve the incidence relation, they are also automorphisms of G_q .

Now, in order to prove both vertex-transitivity and edge-transitivity, let us prove that for any vertices $(0, P), (1, L), (0, P')$ and $(1, L')$ with $(0, P)$ adjacent to $(1, L)$ and $(0, P')$ adjacent to $(1, L')$ there is a graph automorphism that maps $(0, P)$ into $(0, P')$ and $(0, L)$ into $(0, L')$. This is equivalent to finding an automorphism φ of $P_2(\mathbb{F}_q)$ that maps the point P into P' and the line L into L' . By linear algebra there is an invertible matrix M such that $M[P, Q] = [P', Q']$. The induced automorphism is the one desired. To complete the vertex-transitivity note that mapping (s, P) into $(1 - s, P)$ is a graph automorphism. \square

Finally, a particular and interesting case of partitioning G_q graphs is presented: when $q = p^2$ is a square and p is a power of a prime. In this case, the projective plane $P_2(\mathbb{F}_{p^2})$ can be partitioned into $p^2 - p + 1$ subplanes $P_2(\mathbb{F}_p)$ [20]. This implies that G_{p^2} can be partitioned into $p^2 - p + 1$ graphs isomorphic to G_p , which leads to an straightforward layout of the network. Figure 2 shows the partitioning of G_4 as an example. Global links are represented with thin lines and local links with thick lines. The local links induce $3 = 2^2 - 2 + 1$ subgraphs isomorphic to G_2 . Note that for larger sizes the number of global links is almost the square of the local links.

3.2 Modified Incidence Graph of Finite Projective Planes

Vertex identification is a graph transformation in which the new graph has one vertex for each identified pair from the original graph and the new adjacencies combine the ones from both original vertices. In the previous graph G_q , each vertex $(0, P)$ can be identified with its pair $(1, P)$, for every $P \in P_2(\mathbb{F}_q)$, giving a graph of diameter 2 very close to the Moore bound. That is, it results a vertex P , which is connected to Q if $(0, P)$ is connected to $(1, Q)$, or equivalently, $(1, P)$ is connected to $(0, Q)$.

Definition 3.6. *Let q be a power of a prime number. Let $G_q = (V, E)$ be the graph with vertex set and edges set*

$$V = P_2(\mathbb{F}_q)$$

$$E = \{\{P, L\} \mid P \perp L, P \neq L, P, L \in P_2(\mathbb{F}_q)\}.$$

Independently and simultaneously, Brown in [7] and Erdős *et al.* in [15] gave similar definitions as the former. Interestingly, Brahme *et al.* have recently unknowingly reinvented these graphs with a different construction and proposed them for HPC clusters [5]. However, in this paper the Definition 3.6 will be considered as the network topology model, since it is based on projective planes.

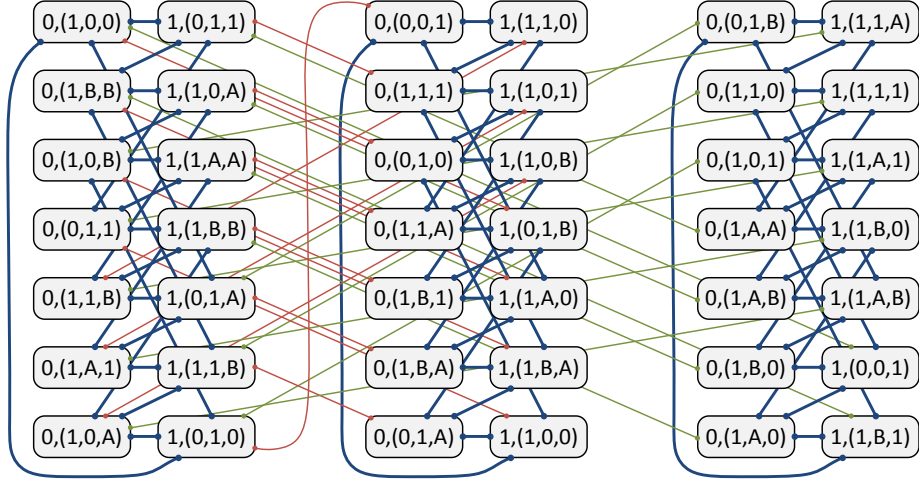


Figure 2: A layout of the Projective Network which relies on graph G_4 , based on the three subplanes of $P_2(\mathbb{F}_4)$. Representatives of \mathbb{F}_4 are denoted $\{0, 1, A, B\}$, with $B = A + 1 = A^2$. Links between subplanes 2 and 3 have been omitted for simplicity.

Clearly, \overline{G}_q has $q^2 + q + 1$ vertices. Now, since $P_2(\mathbb{F}_q)$ contains $q + 1$ points X such that $X \perp X$, this graph is a non-regular graph with degrees q and $q + 1$.² Hence, its number of vertices is $N = q^2 + q + 1 = \Delta^2 - \Delta + 1$, where $\Delta = q + 1$ is the maximum degree. Note that this expression is very close to the Moore bound $M(\Delta, 2) = \Delta^2 + 1$. In the next example it is shown how \overline{G}_2 is obtained from G_2 .

Example 3.7. In Figure 3 both G_2 and its modified graph \overline{G}_2 are represented. Now, vertex 001 is obtained identifying point 001 and line 001 in G_2 . In this case the adjacent vertices are the same than in the original graphs, that is, 100, 110 and 010, providing degree 3 again for this vertex. However, vertex 110 in \overline{G}_2 has degree 2. Note that point 110 and line 110 were adjacent in G_2 , and this edge is lost in the new graph because of the identification. In other case, it would imply the existence of a loop.

Existent paths, routing and link utilization are studied next.

Lemma 3.8. For each pair of vertices of \overline{G}_q there is a unique minimum path.

Proof. Let P, Q be two vertices in \overline{G}_q . If P and Q are adjacent, straightforwardly there is a unique edge joining them. On the contrary, if they are not adjacent, their vector product is adjacent to both, which gives a minimum path between them. If any other minimum path were exist, the two paths will form a square in the graph, which is not possible.

The nonexistence of a square can be proved as follows. Let the points P, Q be adjacent to the points X and Y . Let C be the cross point of the lines PQ and XY . Point C is adjacent to P and Q , since it is a linear combination

²It can look strange an element orthogonal to itself, but it happens in finite fields. Consider for example the field \mathbb{F}_3 , where $(1, 1, 1) \cdot (1, 1, 1) = 3 = 0$.

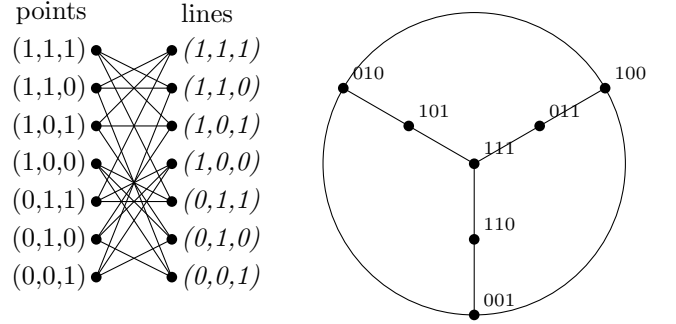


Figure 3: Left: The incidence graph G_2 (Heawood graph). Right: the associated modified incidence graph \overline{G}_2 .

of X and Y . In the same way it is adjacent to X and Y . Furthermore, C is adjacent to all the points in the lines PQ and XY , and hence to all the points in the plane, which contradicts the maximum degree being $q + 1$. \square

Corollary 3.9 (Routing in \overline{G}_q). For each pair of vertices P and Q of \overline{G}_q , one of the following holds:

- P and Q are neighbours or
- the unique path that joins them is $P, P \times Q, Q$.

Theorem 3.10. The average link utilization of \overline{G}_q is $u = \frac{2q^2 + q + 1}{2q(q + 1)}$.

Proof. The vector product of a vertex of degree q and a vertex of degree $q + 1$ is the vertex of degree q . It follows that there is no pair of adjacent vertices of degree q , since both should be their vector product. Thus, there are two types of edges: edges with endpoint degrees $q - (q + 1)$ and edges with endpoint degrees $(q + 1) - (q + 1)$. The remainder

of the proof consists on counting these links and the traffic over them.

First, let us consider edges of type $q-(q+1)$, denoted X and Y respectively. There are $q+1$ vertices of degree q and for each of these vertices there are q edges, all of this type. Therefore, there are $q(q+1)$ edges of this type. The traffic traversing the arc from X to Y is composed from the traffic from: 1 path from X to Y , $q-1$ paths from neighbours of X to Y , and q paths from X to neighbours of Y ; which gives a total of $2q$ paths.

Next, let us consider edges of type $(q+1)-(q+1)$. Let us denote the endpoints X and Y . The total number of edges in \overline{G}_q is $\frac{q(q+1)+(q+1)q^2}{2} = \frac{q(q+1)^2}{2}$. The number of edges of this type is then

$$\begin{aligned} \frac{q(q+1)^2}{2} - q(q+1) &= q \frac{(q^2 + 2q + 1) - (2q + 2)}{2} \\ &= \frac{q(q^2 - 1)}{2}. \end{aligned}$$

The vertices X and Y have a common neighbour $X \times Y$, whose traffic does not go through this edge. Thus, the traffic from X to Y is due to: 1 path from X to Y , $q-1$ paths from neighbours of X to Y , and $q-1$ paths from X to neighbours of Y ; which constitute a total of $2q-1$ paths.

The maximum load is therefore on the $q-(q+1)$ links. The average use of the links can be calculated as follows:

$$\frac{(2q)(q(q+1)) + (2q-1)\frac{q(q^2-1)}{2}}{\frac{q(q+1)^2}{2}} = \frac{2q^2 + q + 1}{q+1}.$$

Finally, the average link utilization at the saturation point is equal to the average use between the maximum use, this is,

$$u = \frac{\frac{2q^2+q+1}{q+1}}{2q} = \frac{2q^2 + q + 1}{2q(q+1)}.$$

□

Notation 3.11. *Previous families of graphs constitute the topological models of Projective Networks (PN). We will refer to PN when the graph G_q is considered, and to demi-PN when the graph \overline{G}_q is selected.*

3.3 Adverse Traffics

When the traffic is uniform, routing using the shortest paths gives the optimal performance. However, the situation is quite different for other traffic patterns. In this subsection adverse traffic patterns for projective networks are considered. Firstly, the worst case traffic pattern for demi-PN networks, without endpoint congestion, is characterized. This traffic, which is the same worst traffic pattern as for the Slim Fly [2], can be defined as follows. Let us consider any pair of adjacent vertices P and Q , both with degree $q+1$. Let us select Δ_0 compute nodes among the neighbours of P but different to Q and $P \times Q$. Note that, by Corollary 3.9,

each of these neighbour routers has a unique path towards Q , which uses the link (P, Q) . Now, to define the traffic, these Δ_0 compute nodes send packets towards Q . Similarly, the Δ_0 compute nodes in P send traffic towards Δ_0 compute nodes selected among the neighbours of Q other than P and $P \times Q$. Thus, there is a total of $2\Delta_0$ flows going through the link (P, Q) , so the maximum throughput is $\frac{1}{2\Delta_0}$ times the throughput under uniform traffic. Secondly, for the PN the situation is more or less the same. Since shortest paths of distance 2 are unique, the previous definition constitutes a traffic as adverse in the PN as for the demi-PN. However, in this network distance 3 shortest paths exists and are not unique, which might contribute with additional flows to the link.

There are different mechanisms to address adverse traffic patterns. A simple solution is the one offered by Valiant and Brebner [37], which consists on misrouting to a random intermediate router, making the traffic uniform but effectively doubling path lengths. However, applying the Valiant mechanism reduces the maximum throughput by half. The UGAL mechanism [35] decides to route minimally or using Valiant upon packet injection depending on the status of the queues. Specific strategies for the projective networks could also be considered. Anyway, graphs near the Moore bound have low path diversity. Note that attaining a huge number of nodes, but with a bounded degree and small diameter, does not allow to give alternate paths. In particular, generalized Moore graphs are such that shortest paths of length $k-1$ are unique. As a consequence, in this kind of networks adaptivity and fault-tolerance should be managed using misrouting techniques.

4 Topologies Near the Moore Bound

As stated in previous sections, our aim is to find topologies being optimal according to Equations (2) and (5). That is, for a given \bar{k} and R , the goal is to find well-balanced topologies with maximum number of terminals T . Thus, in Subsection 4.1, topologies with small average distance are considered, that is, $\bar{k} \leq 2$. The MMS graph has been proposed for interconnection networks with the name of Slim Fly and for this reason it is analyzed in depth in Subsection 4.2. Although the MMS graph is a generalized Moore graph with diameter 2 and $\bar{k} = 2$, its link utilization converges to 8/9, so it does not reach the bound in Equation (5). In Subsection 4.3 some other projective constructions of a greater average distance than the ones presented in Section 3 are summarized.

4.1 Topologies with Small Average Distance

In this subsection graph constructions approaching the generalized Moore bound and average distance between 1 and

2 are considered. Straightforwardly, the only graphs with $\bar{k} = 1$ are the complete graphs, which are indeed Moore graphs. As stated in Section 2, complete graphs are the optimal topologies as long as routers with enough radix are available. There are many other generalized Moore graphs with \bar{k} between 1 and 2, for example: the Turán graph, the Paley graph and the Hamming graph of dimension 2, which are described next.

The *Turán graph* [9] $\text{Turán}(n, r)$ is a complete multipartite graph on n vertices. Let s_1, \dots, s_r be r subsets of $\{1, \dots, n\}$ with cardinal number $\lfloor n/r \rfloor$ or $\lceil n/r \rceil$. Then, two vertices are connected if and only if they are in different subsets. Note that the Turán graph contains the complete bipartite graph as a special case: $\text{Turán}(2n, 2) \cong K_{n,n}$. In the limit the Turán graph has average distance $\lim_{N \rightarrow \infty} \bar{k} = 1 + \frac{1}{r} = 1.5, 1.\bar{3}, 1.25, 1.2, 1.1\bar{6}, \dots$

The *Paley graph* [4] is a graph with $\lim_{N \rightarrow \infty} \bar{k} = 1.5$ very similar to the complete bipartite graph. Let q be a prime power satisfying $q \equiv 1 \pmod{4}$. Then, the Paley graph $\text{Paley}(q)$ is the graph whose vertices are the elements of the finite field of q elements \mathbb{F}_q . Two vertices $a, b \in \mathbb{F}_q$ are connected in $\text{Paley}(q)$ if the difference $a - b$ has its square root in \mathbb{F}_q , i.e., if there is $x \in \mathbb{F}_q$ such that $a - b = x^2$. A notable property of this graph is that it is *self-complementary*: it is isomorphic to the graph that connects vertices if they are not connected in the Paley graph. The Paley graph will appear again later as subgraph of the MMS graph (yet to be introduced).

The *Hamming graph* [32] of side n and dimension 2 is defined as the Cartesian graph product of two complete graphs, $K_n \square K_n$. Two vertices are adjacent if their Hamming distance is 1. It is denoted in recent networking literature as flattened butterfly [25]; other names it has received are rook's graph, generalized hypercube [3] and K-cube [27]. It has diameter $k = 2$, average distance $\bar{k} = 2 - \frac{2}{n} - \frac{1}{n^2}$ and size $N = n^2 = \Delta^2/4 + \Delta + 1$ (a factor 1/4 from being asymptotically a Moore graph). Nevertheless, it is a generalized Moore graph.

4.2 Slim Fly

Slim Fly is the name given by Besta and Hoefler [2] to network topologies based on the McKay–Miller–Širán (MMS) graphs [28]. The MMS is a family of graphs of diameter 2 reaching asymptotically $\frac{8}{9}$ of the vertices given by the Moore bound. When degree $\Delta = 7$ is considered, the MMS graph coincides with the Hoffman–Singleton graph [21], which is a Moore graph. Thus, for small number of vertices it is a very good option although it gets slightly worse for larger ones. Figure 4 shows how the number of vertices of the MMS graph converges to $\frac{8}{9}$ the cardinal given by the Moore bound for $k = 2$. Note that the graph attaining value 1 in the ordinates is the Hoffman–Singleton graph, which is a Moore graph.

Let us now give a schematic definition of this graph based on the ideas in [19]. Let q be a prime power other than 2. Then, for some $\varepsilon \in \{-1, 0, 1\}$, $q \equiv \varepsilon \pmod{4}$. As q is a

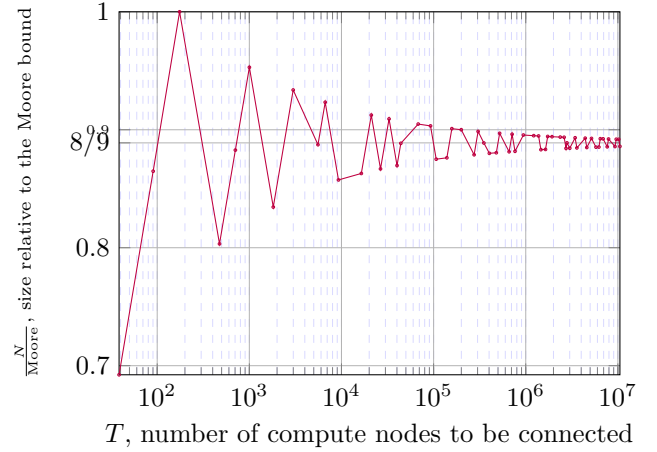


Figure 4: Convergence on the number of vertices in the MMS graph to $\frac{8}{9}$ of Moore bound for diameter 2.

prime power there is a (unique) finite field of q elements, which is denoted by \mathbb{F}_q . The set of vertices is defined as

$$V(\text{MMS}(q)) = \{(s, x, y) \mid s \in \{0, 1\}, x, y \in \mathbb{F}_q\}.$$

Thus, $\text{MMS}(q)$ is a graph with $2q^2$ vertices. In order to define the set of adjacencies a *primitive element* $\xi \in \mathbb{F}_q$ has to be found, that is, an element ξ satisfying $\{\xi^i \mid i \in \mathbb{Z}\} = \mathbb{F}_q \setminus \{0\}$. This implies that $\xi^{q-1} = 1$. Now, let us first define the sets

$$X_0 = \begin{cases} \{1, \xi^2, \dots, \xi^{q-3}\} & \text{if } \varepsilon = 1, \\ \{1, \xi^2, \dots, \xi^{\frac{q-1}{2}}, \xi^{\frac{q+1}{2}}, \dots, \xi^{q-2}\} & \text{if } \varepsilon = -1, \\ \{1, \xi^2, \dots, \xi^{q-2}\} & \text{if } \varepsilon = 0, \end{cases}$$

and $X_1 = \xi X_0$. Later it will be used that $|X_0| = \frac{q-\varepsilon}{2}$, $X_0 \cup X_1 = \mathbb{F}_q \setminus \{0\}$ and

$$X_0 \cap X_1 = \begin{cases} \emptyset & \text{if } \varepsilon = 1, \\ \{1, -1\} & \text{if } \varepsilon = -1, \\ \{1\} & \text{if } \varepsilon = 0. \end{cases}$$

The adjacencies are defined as follows:

1. (s, x, y_1) is adjacent to (s, x, y_2) for all $s \in \{0, 1\}$, $x, y_1, y_2 \in \mathbb{F}_q$ such that $y_1 - y_2 \in X_s$.
2. $(0, x_1, y_1)$ is adjacent to $(1, x_2, y_2)$ for all $x_1, x_2, y_1, y_2 \in \mathbb{F}_q$ such that $y_1 - y_2 = x_2 x_1$.

Thus, each vertex has $|X_0|$ incident edges by the first item and q incident edges by the second item. Therefore, the degree of $\text{MMS}(q)$ is $\Delta = \frac{3q-\varepsilon}{2}$. For convenience, let us call the edges by item 1), *local edges* and the edges by item 2), *global edges*.

The MMS has diameter 2. Let us study the minimum paths to prove this, and further, to count the use of local

and global edges. The possible routes between two vertices could be ll , lg , gl or gg ; where l means a local edge and g a global edge. Let (s_1, x_1, y_1) be the origin vertex and (s_2, x_2, y_2) the destination. If $s_1 = s_2$ and $x_1 = x_2$ then the minimum routes are ll , the same as in Paley graphs. Half of the vertices (s_1, x_1, y_m) can be used as the middle vertex. If $s_1 = s_2$ but $x_1 \neq x_2$ then the minimum route is gg with some middle vertex $(1 - s_1, x_m, y_m)$. The adjacency exists if $y_1 - y_m = (1 - 2s_1)x_mx_1$ and $y_2 - y_m = (1 - 2s_1)x_mx_2$. Hence, the vertex in the middle is unique and can be calculated by $x_m = (1 - 2s_1)(y_1 - y_2)/(x_1 - x_2)$ and $y_m = y_1 - (1 - 2s_1)x_mx_1$. If $s_1 = 1 - s_2 = s$ then half of the minimum routes are lg and the other half gl . The equations for a middle vertex (s, x_1, y_m) are $y_m = y_2 + (1 - 2s)x_1x_2$ and $z = y_1 - y_2 - (1 - 2s)x_1x_2 \in X_s$, while that for a middle vertex $(1 - s, x_2, y_m)$ they are $y_m = y_1 - (1 - 2s)x_1x_2$ and $z = y_1 - y_2 - (1 - 2s)x_1x_2 \in X_{1-s}$. Thus, routing is performed by computing $z = y_1 - y_2 - (1 - 2s)x_1x_2$. If $z = 0$ there is a global edge from the origin to the destination, otherwise, as $X_s \cup X_{1-s} = \mathbb{F}_q \setminus \{0\}$, either $z \in X_s$ or $z \in X_{1-s}$. If $z \in X_s$ use the middle vertex (s, x_1, y_m) and if $z \in X_{1-s}$ use the middle vertex $(1 - s, x_2, y_m)$. The uniqueness depends, therefore, in $X_s \cap X_{1-s}$; if $\varepsilon = 1$ then it is always the empty set and the route is unique, otherwise there are some pairs for which there are two minimal paths. As summary, the number of routes gg is asymptotically the sum of the number of routes lg plus routes gl . Thus, 3 global links are used per each local link used.

The analysis in [2] does not consider the utilization of the two different types of links and concludes that $\Delta_0 = \frac{\Delta}{2}$ terminals per router are required to maximize global bandwidth. However, under such configuration local links receive less load, and the network links are on average under-utilized as shown next. As proved above, the number of global links is about 2 times the number of local links, but the load over the total of global links is about 3 times the load of the local links. Thus, each global link receives about $3/2$ of the load received by a local link. Hence, saturation is reached when global links receive load 1 and local links receive $2/3$. Then, the link utilization is $u = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot \frac{2}{3} = \frac{8}{9}$.³ Therefore, by Equation (1), $\Delta_0 = \frac{4}{9}\Delta$ would provide the correct dimensioning of the network.

Figure 5 shows this convergence of the link utilization to $\frac{8}{9}$. Again, note that this is an asymptotic behaviour; for the case $q = 5$ —the Hoffman–Singleton graph—all links receive the same load and the utilization is $u = 1$ since it is a symmetric graph. The situation is a little worse if $\varepsilon \neq 1$, where there are non-unique minimal paths and, if the routing is deterministic, there are a few links that are used exclusively for messages between their endpoints.

³The value $8/9$ equals the quotient of its number of vertices to the Moore bound. This is a coincidence, it does not hold in the great majority of graphs.

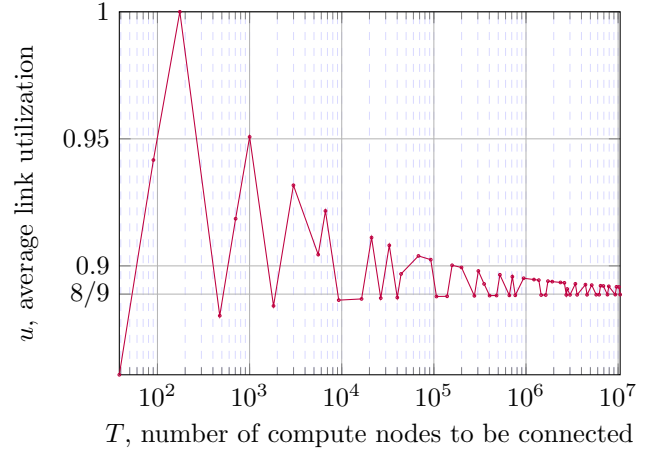


Figure 5: Convergence on the link utilization in the MMS graph to $\frac{8}{9}$.

4.3 Projective Networks of Higher Average Distance

In Section 3 two projective networks of average distances 2 and 2.5 were presented. There are also graphs based on projective spaces which attain the bounds for greater average distances. In this subsection they are enumerated. They are not described in a great detail since such an amount of terminal nodes is beyond the horizon of current network topologies.

The incidence graph over a generalized quadrangle or hexagon, instead of the projective plane, results in a generalized Moore graph with average distance tending to 3.5 and 5.5 respectively [17]. Alike happens to G_q , generalized quadrangles and hexagons exist whenever q is a prime power. Their number of vertices is the double of the number of points in their spaces, respectively $P_3(\mathbb{F}_q)$ and $P_5(\mathbb{F}_q)$.

Furthermore, these graphs allow for a modification similar to \bar{G}_q , as it was proved by Delorme [12]. In the case of quadrangles the resulting average distance tends to 3 and for hexagons it tends to 5. In both cases, the number of vertices is asymptotically close to the Moore bound. However q must be an odd power of 2. Hence, they exist only for a very reduced amount of sizes. Otherwise, Delorme's graph on quadrangles, that is the modified incidence graph on the quadrangles over $P_3(\mathbb{F}_q)$, would have been a very good alternative to current dragonfly topology. These graphs are denoted as *Delorme's graph* in the remainder of the paper. By default this notation will refer to the construction using generalized quadrangles, unless specified otherwise.

5 Comparison of the Topologies

In this section a comparison of the topologies presented in previous Sections 3 and 4 is done in terms of the cost model presented in Section 2. It first considers the complete pic-

Graph	k	$\lim_{N \rightarrow \infty} \bar{k}$	$\lim u$
Complete graph K_N	1	1	1
Turán(N, r)	2	$1 + \frac{1}{r}$	1
Complete bipartite graph $K_{n,n}$	2	1.5	1
Hamming graph 2D	2	2	1
Demi-projective network \bar{G}_q	2	2	1
Slim Fly MMS for $q = 4w + \varepsilon$	2	2	8/9
Projective network G_q	3	2.5	1
Dragonfly	3	3	1
Delorme's graph on quadrangles	3	3	1
Hamming graph 3D	3	3	1
Incidence graph of generalized quadrangles	4	3.5	1
Delorme's graph on hexagons	5	5	1
Incidence graph of generalized hexagons	6	5.5	1
Hypercube C_2^n	n	$n/2$	1

Table 2: Topological parameters of optimal topologies and some references.

ture of all the networks with diameters from 1 to 6, using other topologies such as the dragonfly [26] or 3D Hamming graph as references. Next, it presents a detailed comparison among projective networks and Slim Fly. Finally, it considers different implementations for two specific numbers of compute nodes.

5.1 General Comparison

Table 2 summarizes the fundamental parameters of the graphs presented in Section 4: the diameter and the limit values of average distance and utilization. Table 3 contains the parameters relevant to a network implementing the topology. Both tables present these values for the optimal graphs, other graphs which are close to be optimal and other graphs, such as the hypercube, to take as a reference.

Figure 6 illustrates the cost of networks implementing different topologies using routers with at most 64 ports. Other values of R give similar plots. The thick black curve is the average distance corresponding to an ideal generalized Moore graph with $u = 1$ (like Equation (5)), which is a lower bound for the values of the other curves. Each other curve corresponds to a topology, which is built for all possible radix up to 64. Δ_0 has been assigned trying to be a natural number, but sometimes this condition has been relaxed to avoid under/over-subscription, which would distort the figure. The ordinates axis shows the value \bar{k}/u which, according to Equation (2), is a measure of cost associated to the topology. Thus, curves that attain the bound are the optimal topologies, which are: the complete graph, the Turán graphs, the 2D Hamming graph, demi-PN, PN and Delorme's graph $P_3(\mathbb{F}_q)$. Note that the Delorme's graph over $P_3(\mathbb{F}_q)$ intersects the curve in the limit. However, it only exists when $\Delta - 1$ is a odd power of 2 which means that there are only two points in the range $R \leq 64$. The MMS graph does not attain the bound because of its asymmetry; as we have seen in Section 4, the MMS has $u = 8/9$ in the limit. Hence, the curve is about 9/8 the one of demi-PN. For greater average distances the dragonflies do scale very well, although not attaining the bound. As it can be observed the 3D Hamming graph is completely superseded

by the dragonfly.

Figure 7 indicates which topologies are realizable for a given number of terminals T and available router radix R . It holds that solid lines are sorted by average distance divided by link utilization. Hence, the optimal topology is the solid line immediately above the desired (R, T) point.

5.2 Projective Networks vs Slim Fly

This subsection explains in more detail the advantages of PN and demi-PN with respect to the SF MMS in the design of new high scale interconnection networks, showing the importance of link utilization in the network cost. For this explanation, Figure 8 will be used. In this figure \bar{k} and $\frac{\bar{k}}{u}$ are shown for the three topologies PN, demi-PN and SF MMS. Note that both curves coincide for PN since the graphs G_q are symmetric, as proved in Theorem 3.5.

The smallest average distance is given by SF MMS. However, its maximum size is $\frac{8}{9}$ smaller than the maximum possible one, which is attained by the demi-PN construction; notice the logarithmic abscissas axis. However, when link utilization is considered in the network cost model, for more than 1000 compute nodes demi-PN exhibits as the best alternative both in scalability and cost. Finally, PN is an alternative to scale to a larger amount of compute nodes reaching almost 10^5 compute nodes with minimum cost.

5.3 Cases of Use

To exemplify the use of the topologies, in this subsection different specific networks that connect a given amount of compute nodes are shown. Two approximate network sizes have been selected: 10,000 compute nodes and 25,000 compute nodes. Even for the small case of $T \approx 10,000$, the complete graph would require a router radix of about $R \approx 200$, which is currently unrealistic. Hence, the topologies to be considered will be the Hamming graph, the demi-PN, the SF MMS, the PN and the dragonfly. Tables 4 and 5 show the network parameters for each of the selected topologies in the small and large cases, respectively.

The calculations assume that nodes are arranged into fully electrical groups and cables outside them are optical. These groups are the closest possible to 500 compute nodes, while trying to maximize the connections inside a group. An electrical group size marked with asterisk in the tables indicates the size for most electrical groups, with a few smaller groups.

For a fair comparison, we have employed the cost models from [2] using speeds of 40 Gbps, avoiding the extra costs of 100G routers and cables which are still in their market introduction stage. An average intra-rack distance of 1m is assumed, from which it is obtained a price of 0.985\$/Gbps for the average electrical cable. The average length of the optical inter-rack cables is approximately the average distance of a mesh of same dimensions plus 2m of overhead. In the 10,000 nodes case, an average cost per optical cable of 7.7432\$/Gbps is computed, and in the 25,000 case of 7.9178\$/Gbps. The cost per router is modelled as

Graph	T	R	N	Δ	Δ_0
Complete graph K_N	N^2	$2N - 1$	N	$N - 1$	N
Turán(N, r)	$N^2 \frac{r-1}{r+1}$	$N \frac{(r-1)(2r+1)}{r(r+1)}$	N	$N \frac{r-1}{r}$	$N \frac{r-1}{r+1}$
Complete bipartite graph $K_{n,n}$	$4n^2/3$	$5n/3$	$2n$	n	$2n/3$
Hamming graph 2D of side n	n^3	$3n - 2$	n^2	$2(n - 1)$	n
Demi-projective network \bar{G}_q	$q^3/2 + q^2 + q + 1/2$	$3(q + 1)/2$	$q^2 + q + 1$	$q + 1$	$(q + 1)/2$
Slim Fly MMS for $q = 4w + \varepsilon$	$4/9q^2(3q - \varepsilon)$	$13/18(3q - \varepsilon)$	$2q^2$	$(3q - \varepsilon)/2$	$2/9(3q - \varepsilon)$
Projective network G_q	$4/5(q^3 + 2q^2 + 2q + 1)$	$7(q + 1)/5$	$2(q^2 + q + 1)$	$q + 1$	$2(q + 1)/5$
Dragonfly with h global links per router	$4h^4 + 2h^2$	$4h - 1$	$4h^3 + 2h$	$3h - 1$	h
Delorme's graph on generalized quadrangles	$(q + 1)^2(q^2 + 1)/3$	$4/3(q + 1)$	$q^3 + q^2 + q + 1$	$q + 1$	$(q + 1)/3$
Hamming graph 3D of side n	n^4	$4n - 3$	n^3	$3(n - 1)$	n
Incidence graph of generalized quadrangles	$4/7(q + 1)^2(q^2 + 1)$	$9/7(q + 1)$	$2(q^3 + q^2 + q + 1)$	$q + 1$	$2(q + 1)/7$
Delorme's graph on generalized hexagons	$1/5(q^4 + q^2 + 1)(q + 1)^2$	$6/5(q + 1)$	$q^5 + \dots + q + 1$	$q + 1$	$(q + 1)/5$
Incidence graph of generalized hexagons	$4/11(q^4 + q^2 + 1)(q + 1)^2$	$13/11(q + 1)$	$2(q^5 + \dots + q + 1)$	$q + 1$	$2(q + 1)/11$
Hypercube C_2^n	2^{n+1}	$n + 2$	2^n	n	2

Table 3: Structural parameter of optimal known topologies and some references.

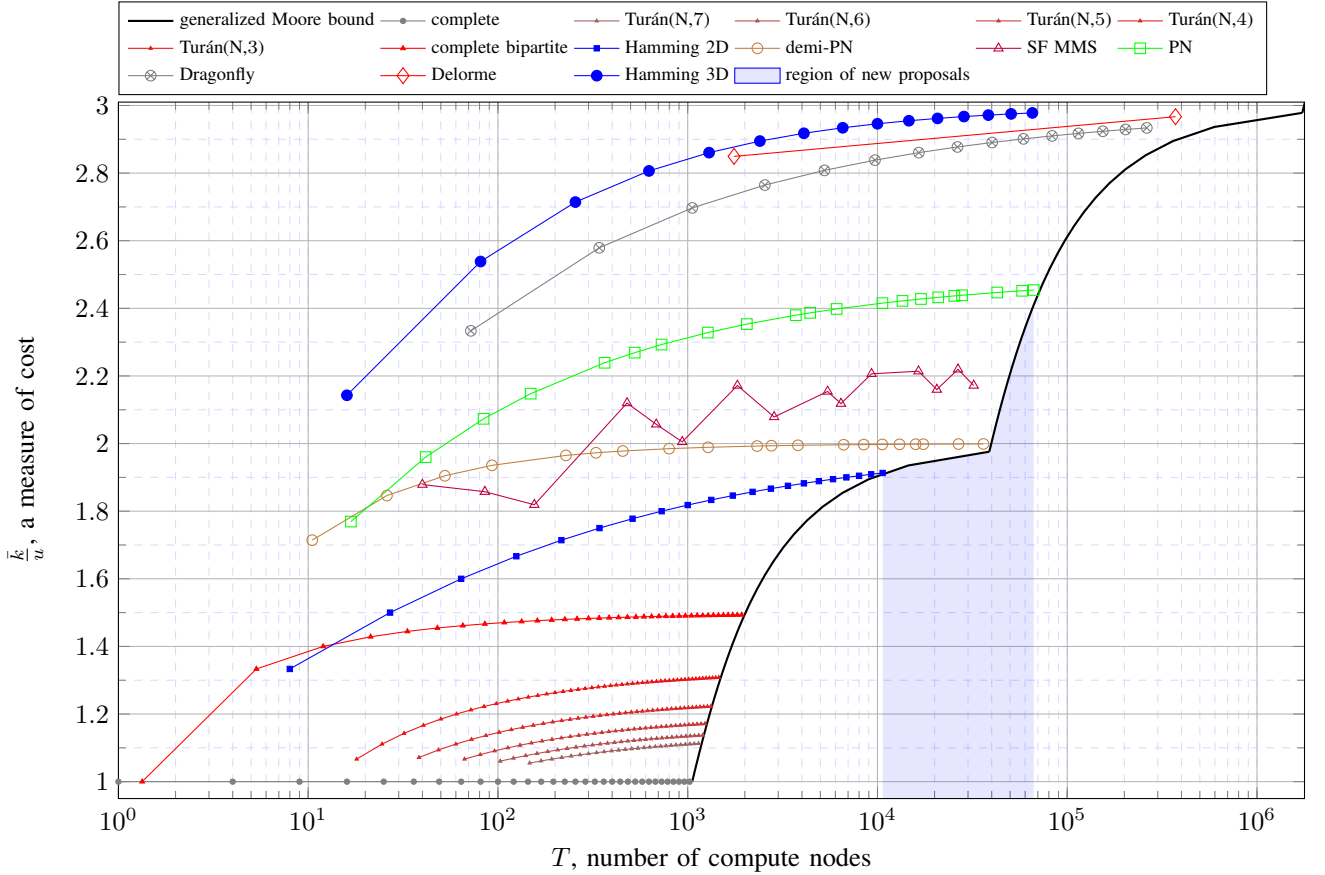


Figure 6: The measure of cost \bar{k}/u in realizations of topologies with a given number of compute nodes using routers with maximum radix 64.

350.4R – 892.3 \$/router. The only power considered is the consumed by the SerDes, which is approximated to 2.8 watts per port.

Tables 4 and 5 show cost and power per node for the topologies studied. The lowest results are obtained in both cases with a 2D Hamming graph. However, the required radix exceeds currently available designs, so it implies using slower links or using multi-chip switches with higher

latency, as discussed in Section 2. Next, we consider designs realizable with full speed and a single switch chip per router. With $T \approx 10,000$ nodes, the demi-PN provides the lowest cost and power, 1% and 7% respectively lower than SF MMS. For $T \approx 25,000$, a diameter 3 network is required using switches up to 48 ports. In this case, the PN provides the lowest power, 10.9% less than the dragonfly. A layout of a projective network requires more optical cables when com-

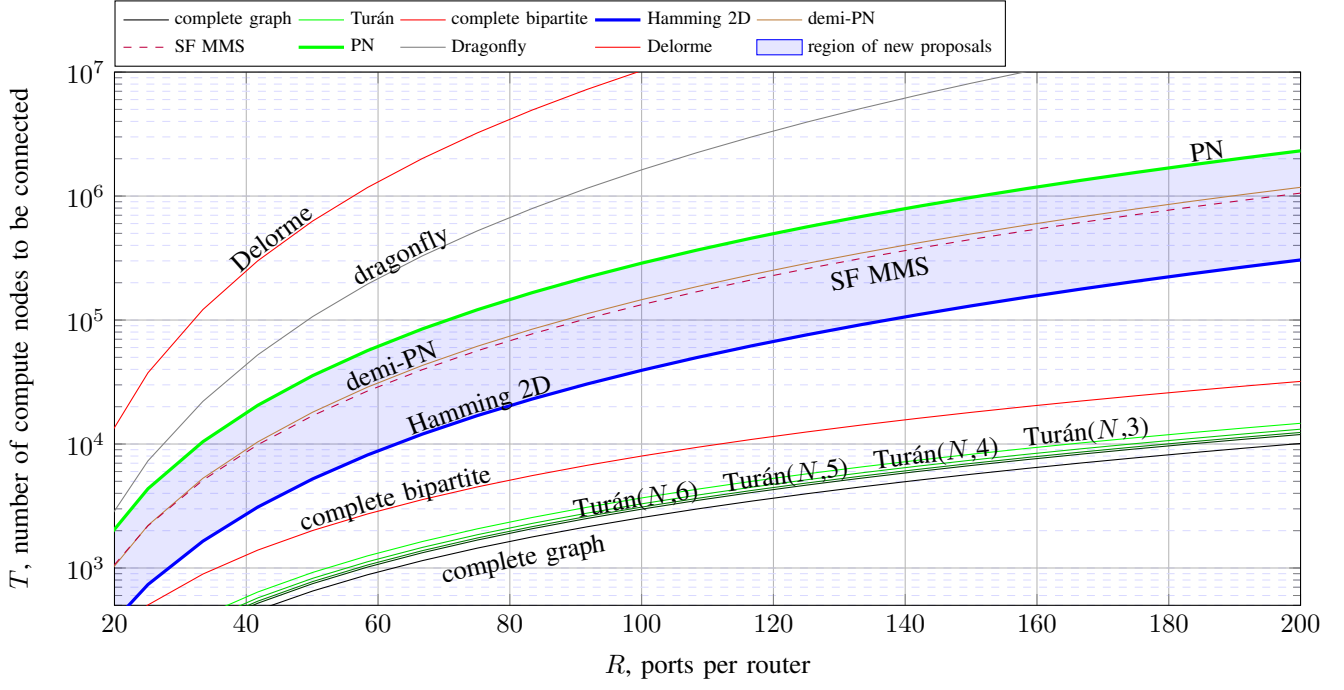


Figure 7: Scalability of the different topologies.

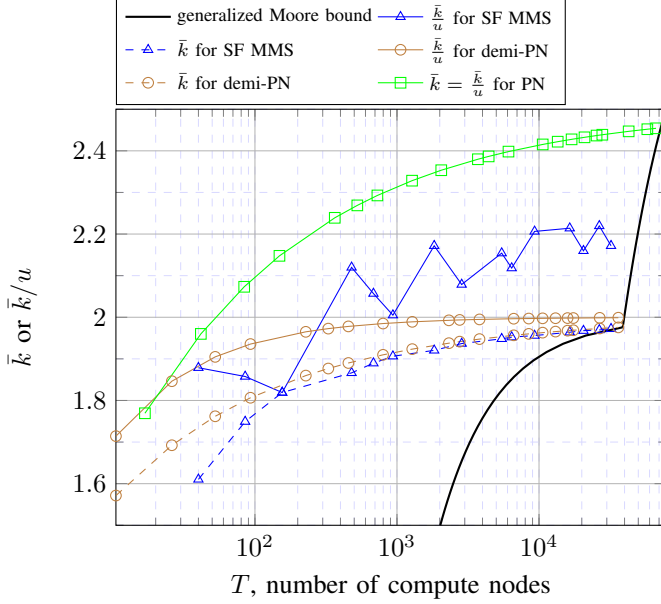


Figure 8: Average distance \bar{k} and cost \bar{k}/u given a number of terminals for SF MMS, PN and demi-PN. Using routers with maximum radix 64.

pared with SF MMS or dragonfly, so in this case the cost of the dragonfly is 2.6% lower because of its reduced number of optical cables. Note that, for an all-optical system such as PERCS [1], projective networks provide significantly better power and cost per node than the alternatives in the tables.

6 Indirect Networks

In this section it is explored how the cost model presented in this paper could be adapted to indirect networks. Moreover, the cost-optimal diameter 2 indirect network, which is the Two-Level Orthogonal Fat Tree [36], can also be obtained using the incidence graph of a projective plane. Hence, in this section it is also illustrated how the previous theoretical graph models for obtaining optimal direct networks can also be applied when dealing with indirect networks.

A *indirect network* has two types of routers. There are *spine routers*, which are connected only to other routers and *leaf routers*, which are also connected to compute nodes. Typically, all routers use the same hardware, so it can be assumed a fixed radix R . In addition, it will be assumed that all leaf routers have the same number Δ_0 of attached compute nodes. Therefore, the graph defined by the routers has two kind of vertices: leaf vertices of degree Δ and spine vertices with degree R , which clearly implies that it cannot be vertex-transitive. Note that the relation $R = \Delta + \Delta_0$ considered for direct networks still holds in the case of indirect networks. In the following, the number of leaf routers will be denoted by L and the number of spine routers by S . Thus, the total number of routers will be $N = L + S$.

Topology	Hamming K_{22}^2	demi-PN(27)	SF MMS(19)	PN(23)	dragonfly(7)
T	10648	10598	9386	9954	9702
R	64	42	42	33	27
N	484	757	722	1106	1386
Δ_0	22	14	13	9	7
subscription	1.002	0.999	0.991	0.921	0.994
Size of electrical group	484	504*	494	396*	490*
Number of groups	22	22	19	26	20
Electrical cables	5082	556	3971	1907	8926
Optical cables	5082	10028	6498	11365	4514
Cost per node (\$)	1145.41	1282.59	1294.51	1546.83	1404.42
Power per node (W)	8.15	8.40	9.05	10.27	10.80

Table 4: Example networks with about 10,000 compute nodes and electrical groups of about 500 nodes.

Topology	Hamming K_{29}^2	demi-PN(37)	SF MMS(27)	PN(31)	dragonfly(9)
T	24389	26733	26244	25818	26406
R	85	57	59	45	35
N	841	1407	1458	1986	2934
Δ_0	29	19	18	13	9
subscription	1.001	0.999	0.976	1.003	0.996
Size of electrical group	435*	532*	486	520*	486*
Number of groups	58	51	54	51	55
Electrical cables	5684	620	10935	3381	25101
Optical cables	17864	26094	18954	28395	13041
Cost per node (\$)	1237.43	1314.29	1344.11	1497.77	1457.39
Power per node (W)	8.21	8.40	9.18	9.70	10.89

Table 5: Example networks with about 25,000 compute nodes and electrical groups of about 500 nodes.

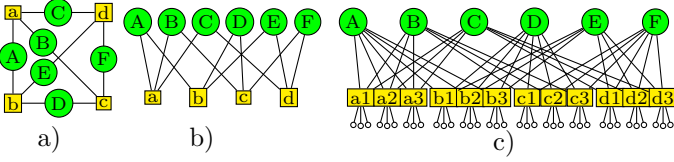


Figure 9: Incidence graph of K_4 and the Multi-layer Full-Mesh (MLFM) network.

When considering the graph model to study indirect networks, the main difference with the direct case lies on the diameter and average distance calculation. In this case, the distances of interest are the ones between leaves, so that a great distance between some leaf and some spine routers becomes irrelevant. Thus, instead of the diameter, the maximum distance among leaves is considered; and instead of average distance, the average distance between leaves, still denoted by \bar{k} . The remainder of the section shows how the graph theoretical techniques presented in previous sections can be used to infer indirect network topologies with good properties.

A first example considers the indirect Multi-layer Full-Mesh (MLFM) topology presented in [18]. This topology is obtained from the incidence graph of a complete graph K_n . Figure 9 reflects an example of the network construction using the incidence graph of K_4 . In Figure 9 a) a standard representation of the incidence graph of K_4 is shown. Square shaped vertices are the vertices of the complete graph and circles represent their incidence. In Figure 9 b) a different depiction of this graph is shown, where vertices on the bottom (leaves) are the vertices in K_4 and the vertices on the top (spines) are edges. Finally, in order to equalize the radix of the routers, leaves are replicated and compute nodes are added, as depicted in Figure 9 c). In general, such a configuration can be obtained from any K_n , thus obtaining an indirect network topology with $\binom{n}{2}$ spine routers and $n(n-1)$ leaf routers, each one connected to $n-1$ compute nodes. Therefore, $\Delta = n-1$, $\Delta_0 = n-1$ and $R = 2\Delta$. However, as it will be shown next, this topology is far from being the cost-optimal one among all the indirect topologies of diameter 2.

An analysis for cost and power optimization as the one done in Section 2 is unfeasible due to, among other reasons, the hardness of calculating Moore bounds on irregular graphs. Nevertheless, it is possible to infer a similar formula assuming maximum distance between leaf routers 2, as in the previous case of the MLFM. In such case, there might be links from leaf to spine routers or leaf to leaf routers; links between spines are possible only for diameter $k \geq 3$. Let δ denote the number of links from each leaf router to other leaves, which is again assumed to be constant. Note that $\delta = \Delta$ in direct topologies and $\delta = 0$ in fully indirect topologies, but there are some intermediate topologies. Now, since the maximum distance between leaf routers is 2, every of the R links in a spine router must go to leaf

routers. Thus, counting the links between leaf routers and spine routers it is obtained the following expression

$$L(\Delta - \delta) = SR.$$

Now, the maximum number of leaves in a graph with maximum distance between leaves being 2, can be expressed in terms of (δ, Δ, R) as follows:

$$L \leq 1 + \delta^2 + (\Delta - \delta)(R - 1), \quad (6)$$

Note that this is a Moore bound calculation but considering only leaf vertices. Also, if $\delta = \Delta$ then it becomes the original Moore bound $M(\Delta, 2)$ presented in (3). The optimal value for the number of compute nodes is obtained when

$$\Delta_0 = \frac{u}{k}(2\Delta - \delta),$$

which generalizes (1). Now, the cost per compute node is, analogously as it was done in (2),

$$\frac{\text{\#ports}}{\text{\#compute nodes}} = \frac{NR}{L\Delta_0} = \frac{R + \Delta - \delta}{\Delta_0} = 1 + \frac{\bar{k}}{u}.$$

This surprisingly implies that the cost per node does not depend on δ . Hence, the most interesting value for δ would be the one giving the best scalability, since it provides the maximum number of compute nodes for the same cost. The maximum for (6) is obtained when $\delta = 0$, which is the typical situation in indirect networks. That is, $L \leq 1 + \Delta(R - 1)$.

The *Orthogonal Fat Tree* (OFT) presented in [36] asymptotically attains this bound for $k = 2$. This was already experimentally proved in [24]. Next, a different construction than the one given in that work is presented, illustrating how also OFTs can be obtained from projective finite planes.

OFTs were constructed in [36] using orthogonal Latin squares. As the author already remarked in that paper, there is an intimate relation between orthogonal Latin squares and finite projective planes. That is, there are $n-1$ mutually orthogonal n -by- n Latin squares if and only if there is a finite projective plane of order n [11]. Therefore, in the following definition, OFTs are built directly using projective spaces instead of manipulating mutually orthogonal Latin squares.

Definition 6.1. Let q be a power of a prime number. Let $\hat{G}_q = (V, E)$ be the graph with vertex set and edge set:

$$V = \{(s, P) \mid s \in \{0, 1, 2\}, P \in P_2(\mathbb{F}_q)\}$$

$$E = \{\{(0, P), (1, L)\}, \{(1, P), (2, L)\} \mid P \perp L\}.$$

\hat{G}_q is said to be the *orthogonal fat tree of $P_2(\mathbb{F}_q)$* .

In an OFT network, vertices $(1, P)$ correspond to spine routers and the rest to leaf routers. As an example, let us consider Figure 10. In this figure black circles represent routers and white circles compute nodes. As it can be seen, the routers are displayed into three columns of $q^2 + q + 1 = 7$ routers, since the total number of routers is $N = 3(q^2 + q + 1)$.

Topology	MLFM 22	MLFM 30	OFT 16	OFT 23
T	9702	25230	9282	26544
R	42	58	34	48
N	693	1305	819	1659
Δ_0	21	29	17	24
cables	9702	25230	9282	26544
Cost per node	1297.18	1321.76	1282.19	1312.14
Watts per node	8.40	8.40	8.40	8.40

Table 6: Example Multi-Layer Full-Mesh and OFT networks with about 10,000 and 25,000 compute nodes.

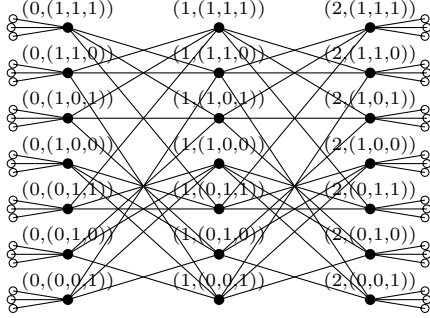


Figure 10: Orthogonal Fat Tree \hat{G}_2

1) = 21. The column in the middle would correspond to spine routers and the other two to leaf routers. It can also be seen that $\Delta = \Delta_0 = q + 1$ and $T = 2(q + 1)(q^2 + q + 1)$. Indirect networks are not vertex-transitive, but OFT is edge-transitive, so the utilization is exactly $u = 1$. The average distance between leaves is exactly $\bar{k} = 2$, since for any two leaves the minimal path connecting them is of length 2. Note that for each leaf there are several spine routers at distance 3. Finally, it is worthwhile to note that two G_q projective networks are embedded in any \hat{G}_q , thus connecting these two different topologies. Moreover, it can be seen that this network has the same cost than the demi-PN and almost the same scalability of the PN, since $T_{PN} = 0.29R^3$ and $T_{OFT} = 0.25R^3$.

Table 6 presents the cost and power per node for OFT and MLFM networks with sizes about 10000 and 25000 compute nodes. A typical layout of indirect networks is done without electrical groups, which implies that every cable has been considered to be optical for the calculations. The MLFM results are similar to the demi-PN with slightly higher power. With respect to the OFT, on the one hand its scalability is slightly lower than PN, since with a slightly greater radix router it connects almost the same number of terminals. On the other hand, OFT has the same cost and power per node than the demi-PN.

7 Conclusions

Projective networks have been proposed in this paper for large systems using direct networks. These networks are built using incidence graphs of projective planes. Our pro-

posal has been done by means of a coarse-grain cost model based on minimizing the average distance of the network while maintaining a uniform link utilization. The optimal networks under this cost model are those generalized Moore graphs which have uniform link utilization and, in particular, those being symmetric. By a complete a study of all the actually known families of generalized Moore graphs, for a given radix router and a number of compute nodes it is possible to choose the optimal network, using this cost model. In particular, projective networks have been proved to be a feasible alternative to the recently proposed Slim Fly. Finally, a first approach to the indirect networks' case has been considered. Our cost model has been adapted to this situation only for diameter two networks, since a general model for any diameter seems unfeasible. As it has been shown, optimal indirect networks for this case are the two-level Orthogonal Fat Trees, which can be also obtained by means of incidence graphs of projective planes.

8 Acknowledgments

This work has been supported by the Spanish Science and Technology Commission (CICYT) under contracts TIN2013-46957-C2-2-P and TIN2016-76635-C2-2-R, the European Union FP7 under Agreement ICT-2013-10-610402 (Mont Blanc 2), and the European HiPEAC Network of Excellence.

References

- [1] Baba Arimilli, Ravi Arimilli, Vicente Chung, Scott Clark, Wolfgang Denzel, Ben Drerup, Torsten Hoefler, Jody Joyner, Jerry Lewis, Jian Li, Nan Ni, and Ram Rajamony. The PERCS high-performance interconnect. In *IEEE Symposium on High Performance Interconnects*, pages 75–82. IEEE Computer Society, 2010.
- [2] Maciej Besta and Torsten Hoefler. Slim Fly: A cost effective low-diameter network topology. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '14*, pages 348–359. IEEE Press, 2014.
- [3] Laxmi N. Bhuyan and Dharma P. Agrawal. Generalized hypercube and hyperbus structures for a computer network. *Computers, IEEE Transactions on*, C-33(4):323–333, April 1984.
- [4] Béla Bollobás. *Random Graphs*. Cambridge studies in advanced mathematics, 2nd edition, 2001.
- [5] Dhananjay Brahme, Onkar Bhardwaj, and Vipin Chaudhary. SymSig: A low latency interconnection topology for HPC clusters. In *International Conference on High Performance Computing (HiPC)*, pages 462–471, December 2013.

- [6] Broadcom. High-density 25/100 Gigabit Ethernet StrataXGS Tomahawk Ethernet switch series. 2014.
- [7] William G Brown. On graphs that do not contain a Thomsen graph. *Canad. Math. Bull.*, 9(5):281–285, 1966.
- [8] Cristóbal Camarero. *Distance and Symmetry Properties of Graphs and their Application to Interconnection Networks and Codes*. PhD thesis, University of Cantabria, March 2015.
- [9] Gary Chartrand and Linda Lesniak. *Graphs and Digraphs*. California Wadsworth and Brooks, 2nd edition, 1986.
- [10] Nikolaos Chrysos, Cyriel Minkenberg, Mark Rudquist, Claude Basso, and Brian Vanderpool. SCOC: High-radix switches made of bufferless Clos networks. In *International Symposium on High Performance Computer Architecture (HPCA)*, pages 402–414, February 2015.
- [11] Charles J Colbourn and Jeffrey H. Dinitz. *Handbook of Combinatorial Designs*. CRC press, 2nd edition, 2007.
- [12] Charles Delorme. Grands graphes de degré et diamètre donnés. *European Journal of Combinatorics*, 6(4):291–302, 1985.
- [13] S. Derradji, T. Palfer-Sollier, J.-P. Panziera, A. Poudes, and F.W. Atos. The BXI interconnect architecture. In *Annual Symposium on High-Performance Interconnects (HOTI)*, pages 18–25, 2015.
- [14] Fuad E. Doany. High density optical interconnects for high performance computing. In *Optical Fiber Communication Conference*, page M3G.1. Optical Society of America, 2014.
- [15] Paul Erdős and Alfréd Rényi. On a problem in the theory of graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 7:623–641, 1962.
- [16] Paul Erdős, Alfréd Rényi, and Vera T. Sós. On a problem of graph theory. *Studia Sci. Math. Hungar.*, 1:215–235, 1966.
- [17] Geoffrey Exoo and Robert Jajcay. Dynamic cage survey. *Electron. J. Combin.*, 2013.
- [18] Fujitsu Laboratories. Fujitsu laboratories develops technology to reduce network switches in cluster supercomputers by 40%. <https://www.fujitsu.com/global/about/resources/news/press-releases/2014/0715-02.html>, 2014.
- [19] Paul R. Hafner. Geometric realisation of the graphs of McKay–Miller–Širáň. *Journal of Combinatorial Theory, Series B*, 90(2):223–232, 2004.
- [20] James William Peter Hirschfeld. *Projective Geometries over Finite Fields*. Clarendon Press Oxford, 1979.
- [21] Alan J. Hoffman and Robert R. Singleton. On Moore graphs with diameters 2 and 3. *IBM Journal of Research and Development*, 4(5):497–504, November 1960.
- [22] Intel. Omni-Path Director Class Switches 100 series product brief. Nov 2015.
- [23] Intel. Omni-Path Fabric Edge Switches 100 series product brief. Nov 2015.
- [24] Georgios Kathareios, Cyriel Minkenberg, Bogdan Prisacari, German Rodriguez, and Torsten Hoefler. Cost-effective diameter-two topologies: Analysis and evaluation. In *International Conference for High Performance Computing, Networking, Storage and Analysis, SC '15*, pages 36:1–36:11, November 2015.
- [25] John Kim, William J. Dally, and Dennis Abts. Flattened butterfly: a cost-efficient topology for high-radix networks. In *International Symposium on Computer Architecture*, ISCA '07, pages 126–137. ACM, 2007.
- [26] John Kim, William J. Dally, Steve Scott, and Dennis Abts. Technology-driven, highly-scalable dragonfly topology. In *International Symposium on Computer Architecture*, pages 77–88, 2008.
- [27] Laurence E. LaForge, Kirk F. Korver, and M. Sami Fadali. What designers of bus and network architectures should know about hypercubes. *Computers, IEEE Transactions on*, 52(4):525–544, April 2003.
- [28] Brendan D McKay, Mirka Miller, and Jozef Širáň. A note on large graphs of diameter two and given maximum degree. *Journal of Combinatorial Theory, Series B*, 74(1):110–118, 1998.
- [29] Mellanox. CS7500 switch system product brief. 2015.
- [30] Mellanox. SB7790 switch system product brief. 2015.
- [31] Mirka Miller and Jozef Širáň. Moore graphs and beyond: A survey of the degree/diameter problem (2nd ed). *The Electronic Journal of Combinatorics*, 5 2013.
- [32] Henry Martyn Mulder. Interval-regular graphs. *Discrete Mathematics*, 41(3):253–269, 1982.
- [33] Sébastien Rumley, Madeleine Glick, Simon D. Hammond, Arun Rodrigues, and Keren Bergman. Design methodology for optimizing optical interconnection networks in high performance systems. In *ISC High Performance conference*. 2015.
- [34] Michael Sampels. Vertex-symmetric generalized Moore graphs. *Discrete Applied Mathematics*, 138(1–2):195–202, 2004. Optimal Discrete Structures and Algorithms.

- [35] Arjun Singh. *Load-Balanced Routing in Interconnection Networks*. PhD thesis, Stanford University, 2005.
- [36] Marcos Valerio, Louise E. Moser, and P. Michael Melliar-Smith. Recursively scalable fat-trees as interconnection networks. In *IEEE International Phoenix Conference on Computers and Communications*, pages 40–46, April 1994.
- [37] Leslie G. Valiant and Gordon J. Brebner. Universal schemes for parallel communication. In *ACM symposium on Theory of computing*, STOC '81, pages 263–277. ACM, 1981.