Design Space Exploration of the Dragonfly Topology

Min Yee Teh[1(✉)], Jeremiah J. Wilke[2], Keren Bergman[1], and Sébastien Rumley[1]

[1] Lightwave Research Laboratory, Columbia University, New York, NY 10027, USA
mt3126@columbia.edu

[2] Scalable Modeling and Analysis, Sandia National Labs, Livermore, CA 94551, USA

Abstract. We investigate possible options of creating a Dragonfly topology capable of accommodating a specified number of end-points. We first observe that any Dragonfly topology can be described with two main parameters, *imbalance* and *density*, dictating the distribution of routers in groups, and the inter-group connectivity, respectively. We then introduce an algorithm that generates a dragonfly topology by taking the desired number of end-points and these two parameters as input. We calculate a variety of metrics on the generated topologies resulting from a large set of parameter combinations. Based on these metrics, we isolate the subset of topologies that present the best economical and performance trade-off. We conclude by summarizing guidelines for Dragonfly topology design and dimensioning.

Keywords: Topologies · Dragonfly · Optical interconnects

1 Introduction

The Dragonfly topology, introduced by Kim et al. [1], is a direct topology, in which every router accommodates a set of *terminal* connections leading to end-points, and a set of *topological* connections leading to other routers. The Dragonfly concept fundamentally relies on the notion of *groups*. A collection of routers belonging to the same group are connected with *intra-group* connections, while router pairs belonging to different groups are connected with *inter-group* connections. In practical deployments, routers and associated end-points belonging to a group are assumed to be compactly colocated in a very limited number of chassis or cabinets. This permits connections between routers and terminals within a group to be implemented using short-distance, low-cost electrical transmission links. Meanwhile, *inter-group* connections are based on optical equipment capable of spanning inter-cabinet distances in the range of tens of meters.

Modularity is one of the main advantages provided by the dragonfly topology. Owing to the clear distinction between intra- and inter-group links, the wiring within a group is independent of the total number of groups in the topology.

J.M. Kunkel et al. (Eds.): ISC High Performance Workshops 2017, LNCS 10524, pp. 57–74, 2017.
https://doi.org/10.1007/978-3-319-67630-2_5

Vendors can therefore propose all-included, all-equipped cabinets corresponding to a group, while supercomputer operators are free to decide how many such groups/cabinets they want to acquire. For instance, the XC40 architecture proposed by Cray consists of 1 to 241 groups [3]. The fixed intra-group wiring also makes upgrading a dragonfly based supercomputer relatively straightforward from a hardware point-of-view, as only existing inter-group links may have to be reorganized. In some cases, incumbent inter-group links can even be kept in place, and simply complemented with additional inter-group links connecting the incumbent groups with several interconnected new groups.

A dragonfly topology also guarantees a large path diversity between endpoints, enabling various flavors of adaptive, non-minimal routing schemes [1]. In the presence of congestion between two groups, traffic can first be deflected to third party groups, then forwarded to the correct destination. This feature allows the bandwidth available between two groups to be virtually multiplied by a factor of up to $g - 2$, where g is the number of groups.

Besides its modularity and capability to leverage non-minimal routing schemes, the Dragonfly topology also clearly distinguishes optical from electrical cables connecting the routers. Although the price gap is shrinking, optical links are still generally more expensive than their electrical counterpart, and thus represent a considerable fraction of an interconnect's total cost. There is therefore a motivation to allow fine-tuning of the expensive "optical bandwidth". A dragonfly cleanly separates the most expensive fraction of the bandwidth (optical) outside the cabinets while leaving the least expensive part (electrical) "hard-wired" inside the cabinets. As not all parallel applications require the same balance between bandwidth and computation, being able to adapt the bandwidth available at procurement time is an interesting feature. For instance, supercomputer operators interested in compute power and less concerned with bandwidth-intensive workloads can save on the "optical-bandwidth" and invest in additional cabinets.

All these interesting features make the Dragonfly topology the default choice for the whole XC series of Cray [4], and is thus widely adopted in the largest supercomputing platforms. The dragonfly concept also triggered sustained interest from the scientific community, with research papers addressing congestion in dragonflies [5] or optimizing throughput [6], and possible inclusion of optical switching [7].

One can note across literature, however, the varying ideas of what constitutes a Dragonfly. Here we aim to clarify the definition of the Dragonfly and then show what a Dragonfly can and cannot be. We first make the relatively trivial but important statement that a Dragonfly with fully-meshed intra-group connectivity can be assimilated into a *2-dimensional Flattened Butterfly* (2D-FB) [2], but with partial connectivity in one dimension (the one wired with optical cables). We then show that a Dragonfly topology can be described by a) the varying sizes of the two dimensions of the underlying 2D-FB, and b) the number of links in the optical dimension. Having reduced the shape of a Dragonfly topology to these two parameters, we perform a thorough exploration of

the Dragonfly design space. We finally analyze the value of the identified designs by means of a cost model. Our analyses are related to the those reported by Camarero et al. [8], but with a focus on practical insights rather than graph theory.

2 Dragonfly Variants Description and Construction

2.1 Definitions

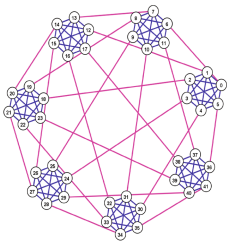
We begin by introducing a notation much inspired by the one originally given by Kim et al. [1]. We consider a Dragonfly as being made of g groups with a routers in each group, therefore with a total of $S = ag$ routers. Each router accommodates p *terminal* connections to end-points. Because we uniquely consider Dragonflies with fully-meshed intra-group connectivity in this paper, each router also accommodates $a - 1$ intra-group connections to the other $a - 1$ routers of the group. Finally, each router has h inter-group connections to routers located in other groups. We immediately remark that under these assumptions, each router must offer at least $radix = p + h + a - 1$ ports and that the topology can scale to $N = Sp = agp$ terminals. The topology is also made of $ga(a-1)/2$ bi-directional electrical links, and $gah/2$ optical ones.

We additionally introduce Δ as the global average distance in the Dragonfly graph, i.e. the average of the minimal number of hops separating every possible node pair (a node in the graph represents a router). We note that Δ is a function of the a, g and h parameters, nevertheless we privilege the Δ notation to $\Delta(a, g, h)$ for brevity. Next to the global average distance Δ, we also introduce δ_i as the minimal distance separating node i from another node on average, which relates to Δ as $\Delta = \frac{1}{S} \sum_{i=1}^{S} \delta_i$.

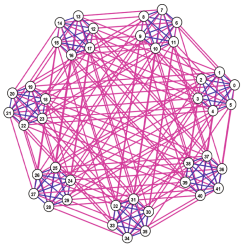
We set the *imbalance* coefficient $b \in [-1, 1]$ to represent the relative size mismatch between the optical and electrical dimensions, and the *density* coefficient $d \in [0, 1]$ to represent the degree of connectivity in the optical dimension. These two parameters will be further described in Sect. 2.4. Finally, because we are interested in comparing Dragonflies of similar scales, we introduce $S_{desired}$ as a parameter imposing a minimal number of routers (hence $S \geq S_{desired}$), and $N_{desired}$ to impose a minimal number of end-points ($N \geq N_{desired}$).

2.2 Dragonfly Construction

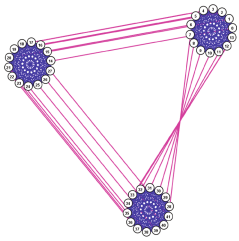
Six examples of Dragonflies all made of $S = S_{desired} = 42$ nodes are illustrated in Fig. 1. We call the case drawn in Fig. 1a the *canonical* design. We take this case as the starting point for our explorations. A Dragonfly is said to be *canonical* when $g = a + 1$ and $h = 1$. In that case, the number of *inter-group* connections associated to a group is $ha = g - 1$, i.e. a group is exactly connected once to every other group. This is in contrast with the case shown in Fig. 1b, which has the same g and a values as the *canonical* case but has $h = 6$ inter-group



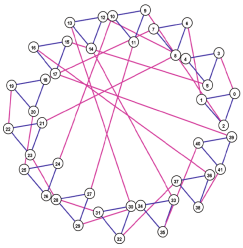
(a) "Canonical" Dragonfly with $a = 6$, $g = 7$, $h = 1$.



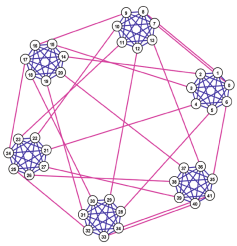
(b) Dragonfly variant with $a = 6$, $g = 7$, and $h = 6$



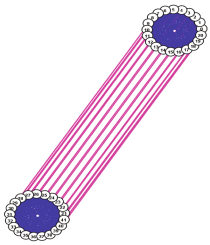
(c) Dragonfly variant with $a = 14$, $g = 3$, and $h = 1$



(d) Dragonfly variant with $a = 3$, $g = 14$, and $h = 1$



(e) Dragonfly variant with $a = 7$, $g = 6$, and $h = 1$



(f) Dragonfly variant with $a = 21$, $g = 2$, and $h = 1$

Fig. 1. Examples of $S_{\text{desired}} = 42$ Dragonfly variants parameterized using different combinations of a, g, and h. Purple links represent inter-group optical links, while blue links represent intra-group electrical links (Color figure online).

links per router. In this case, not only is every group connected to every other group, but every router is directly connected to every other group (as $h = g-1$). As a result, the Dragonfly becomes effectively a 2D-FB with a maximal optical dimension. Through this example, we see that every router can be characterized by a point described by coordinate (x, y) in a 2D-lattice, with x giving the router's position in the electrical dimension (i.e. within a group) and y giving the group the router belongs to. We further remark that the size of the *electrical* dimension is a (as $x \in [0, a-1]$), and the size of the *optical* dimension is g ($y \in [0, g-1]$). The optical dimension is minimally populated when $h = 1$ and maximally populated with $h = g - 1$. We also note that the cases in Fig. 1a and b have similar sizes in both the optical and electrical dimensions, with Fig. 1b having maximal optical connections (note $h = g - 1 = 6$) while Fig. 1a has minimal optical connections (note $h = 1$). We can therefore describe the *canonical* dragonfly as a case with minimal optical wiring (since $h = 1$), in which routers are identically distributed across both electrical and optical dimensions. Note that this *canonical* construction still allows every group pair to be directly connected.

Figure 1c shows a case of great discrepancy between electrical and optical dimensions, with the electrical dimension ($a = 14$) much larger than the optical one ($g = 3$). We note that each group has $ah = 14$ inter-group links, the total number of inter-group links is $gah/2 = 21$, and that each pair of groups is

connected through 7 connections. This means that exactly half of the routers in, say, group 0 are connected to group 1, and the other half to group 2.

Figure 1d shows an opposite case with a small electrical dimension ($a = 3$, $g = 14$). Since only one inter-group link is allocated to each router, the number of inter-group links leaving each group is only $ah = 3$, which does not permit full inter-group connectivity. Also note that it is not straightforward to pick which 3 among 13 other groups to form an inter-group connection with, since there are many such possible combinations. A similar problem of links/group-mismatching is faced in the example shown in Fig. 1e: each group has $ah = 7$ inter-group links at its disposal, whereas only $g - 1 = 5$ neighboring groups must be reached. To allocate inter-groups links in these "inharmonious" cases, a wiring algorithm is introduced in the next subsection. Finally, Fig. 1f shows a case of when $h = g-1$. Due to this equality, the resulting topology is a 2D-FB, and although $h = 1$, it is incidentally also equals to $g-1$, and thus cannot be scaled larger. Through these examples, we see that the design space for a Dragonfly with $S = 42$ is already quite wide, demonstrating the richness of designs when S scales to $1,000$ or higher.

2.3 Dragonfly Graph Wiring Algorithm

As discussed in the previous subsection, in order to explore the entire design space, we need to be able to generate a Dragonfly topology described by any arbitrary combination of a, g, and h parameters. Given this set of parameters, we would like to distribute the inter-group links between groups such that the diameter and global average distance Δ are minimized, while maintaining fairness by avoiding unevenly-connected nodes (indicated by high variance of δ_i).

The problem of distributing inter-group links is that to achieve optimal fairness, diameter or Δ (or a combination thereof) is NP-hard. Instead of targeting global optimality, the wiring algorithm we introduce is a greedy heuristic. The algorithm starts by considering every group as a vertex in a secondary graph $G = (V, E)$, and by allocating $a \times h$ links to each vertex $V_k \in V$, effectively creating an inter-group topology. The destination group V_i of each newly added link is chosen by considering the sum of two factors: (a) the total number of connections V_i has with every other vertex in G, and (b) the number of connections V_i has with the target group, V_k, specifically. To maintain wiring fairness and minimize diameter, the V_i that corresponds to the lowest sum of the aforementioned two factors is picked. As a result of this policy, the algorithm may select V_i even though one or more links have already been awarded to the (V_k, V_i) pair. Once the link has been allocated to said group pair, the algorithm then identifies the routers within groups k and i with the least number of connections so far, and connects these two routers.

When the graph G is sparsely occupied by edges, every group is equally likely to be picked to form a link with V_k, and inter-group link allocation resembles the *relative global link* arrangement as discussed in E. Hastings et al. [11]. As G becomes more saturated with edges, the algorithm tends to distribute links

Algorithm 1. Dragonfly Wiring Algorithm

define $G := (V, E)$, s.t V is set of all the Dragonfly groups and E is the set of inter-group links
initialize $\eta_{ij} := 0, \forall\ i, j \in V$
for $k \in V$ **do**
  for $d := 0, ..., a \times h$ **do**
    for $i \in V$ where $i \neq k$ **do**
      define $\mu_i := \eta_{ik} + \sum_{j \in V} \eta_{ij}, \forall\ i, j$ s.t $j \neq k$
      pick i s.t $\mu_i = \min_{i' \in V} \mu_{i'}$ and $\sum_{j \in V} \eta_{ij} < (a \times h)$
      $\eta_{ik} := \eta_{ik} + 1$
    end for
  end for
end for

in a fair way by selecting groups currently with the lowest number of formed connections, thus making inter-group link arrangement seem more random.

In the preceding pseudocode, η_{ij} is used to represent the total number of inter-group links connecting group i to group j. Since G is an undirected graph, symmetry dictates that $\eta_{ij} = \eta_{ji}$. μ_i denotes the "score" of the group i, which is used to account for the sum of both how many inter-group links the current target group k shares with destination group i (accounted for by η_{ik} term), and how many inter-group links destination group i currently shares with other groups (accounted for by $\sum \eta_{ij}$ term).

We evaluated the topologies obtained with our wiring algorithm in terms of global average distance Δ, diameter, and fairness. To measure wiring fairness, we consider two metrics: the first identifies δ_{min} and δ_{max} among all δ values, i.e. the average distances seen from the best and worst connected node, respectively, and calculate the greatest percentage difference, d, using $d = 100(\frac{\delta_{max} - \delta_{min}}{\delta_{min}})$. The second metric calculates the squared coefficient of variation across the δ_i set. Results for a set of topologies with at least $S_{desired} = 1000$ are displayed in Fig. 2. We observe that global average distances Δ generally decreases as more links are added to the optical dimension. In general, the larger the groups, g (thus smaller group sizes, a), the more reliant the Dragonfly is on optical

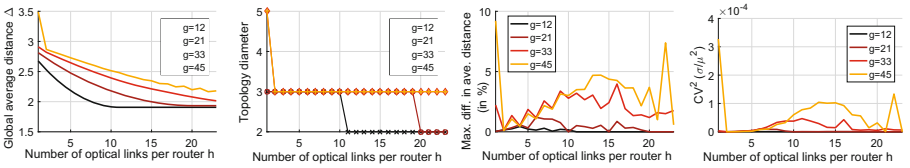


Fig. 2. (a) Global average distance Δ, (b) topology diameter, (c) maximum difference between smaller and larger node average distance δ_i, and (d) squared coefficient of variation of δ_i.

links to "reach" routers in other groups, as opposed to reaching them directly via the intra-group electrical links. This translates into larger Δ values for the same h. Note that ripples appear for $g = 45$, revealing some limitations in the wiring algorithm. More importantly, when $a \times h$ reaches or exceeds $g - 1$, both dimensions are fully populated, and we obtain a 2D-FB topology with diameter of 2. At this point, additional inter-group links are parallel to existing links, which does not affect Δ. In contrast, when $g = 45$, and $a = \lceil \frac{S_{\text{desired}}}{g} \rceil = 23$, the diameter is 5 for $h = 1$ as shown in Fig. 2b. Hence, with $ah = 23$ inter-group links per group, all-to-all group connectivity cannot be guaranteed anymore.

Figure 3 shows the sorted δ_i values for 16 datapoints of Fig. 2. The maximum difference d between δ_{min} and δ_{max} is also displayed. For $(g = 12, a = 84, h = 15)$, $(g = 21, a = 48, h = 15)$ and $(g = 21, a = 48, h = 10)$, the average distance δ_i is the same for all nodes and d is therefore null (ideal fairness). In the first case, h is larger than $g - 1$ leading to a saturation of the connectivity in the optical dimension thus to a 2D-FB topology. In the second case, each group has $a \times h = 48 \times 15 = 720$ inter-group links, which is a round multiple of $g - 1 = 20$. Every group pair is thus awarded $720/20 = 36$ links. The fact that these 36 links must be further allocated to the $a = 48$ routers composing each group is not causing unfairness, a fact that validates the viability of the wiring algorithm. The same situation occurs in the third case $(g = 21, a = 48, h = 10)$: there are 480 inter-group links per group, which is also a round multiple of 20.

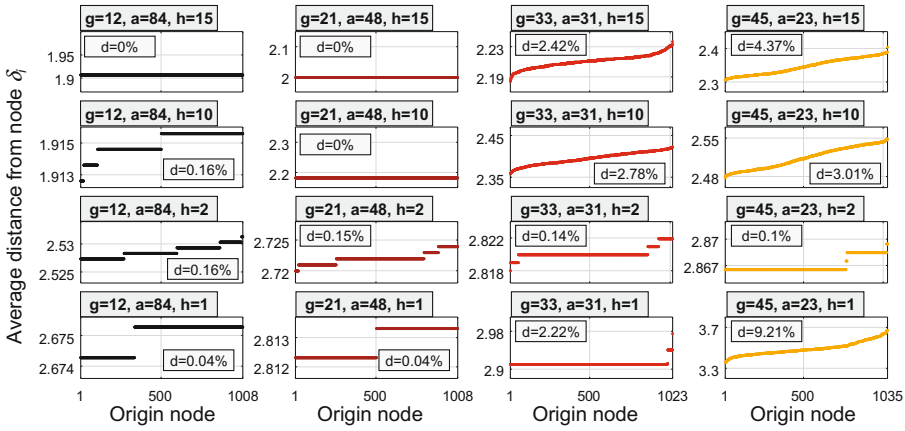


Fig. 3. Distributions of average distances of graph as viewed from each node. d in each plot denotes the percentage of greatest difference in average distance, δ_i

When $a \times h$ (the number of inter-group links per group) is not a multiple of $g - 1$, some group pairs receive extra links (the "remainder" links). These specific routers that are given the "remainder" links are consequently favored. Looking at the general behavior on Fig. 2(c-d), we observe that unfairness tend to grow with large h values, and with the number of groups g. In general, the

more remainder links and group pairs, the harder it is to maintain fairness. Also note the bottom right cases on Fig. 3 ($h = 1$, $g = 33$ or 45): with less than one inter-group link per group pair on average, all-to-all inter-group connectivity is not maintained, causing the diameter to be 5. Such cases are also subject to increased unfairness.

2.4 Exploring the Dragonfly Using Imbalance and Density Parameters

As mentioned above, we introduce two parameters to control the shape of a Dragonfly topology. The *imbalance* coefficient $b \in [-1, 1]$ represents the relative size mismatch between the optical electrical dimensions, and the *density* coefficient $d \in [0, 1]$ represents to what extent the optical dimension is inter-connected. The density d parameter implicitly controls h through:

$$h = max(0, \lfloor 1 + d(g-2) \rfloor), \text{ where } 0 \leq d \leq 1 \text{ and } g > 1 \tag{1}$$

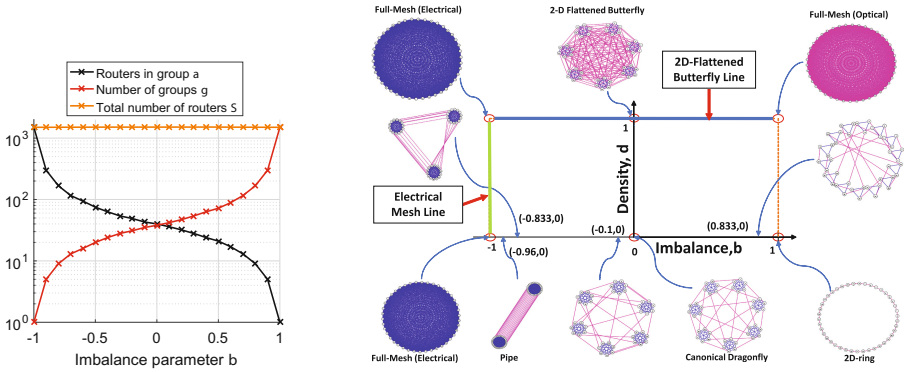
For $d = 0$, h is always equal to one (minimal inter-group connectivity). In contrast, for $d = 1$, $h = g - 1$, each router is connected to its counterpart in every other group, and the topology is thus a 2D-FB (maximal inter-group connectivity). For the imbalance parameter, $b = 0$ should reflect a situation as close to the *canonical* dragonfly as possible with $g = a - 1$. We define $b = -1$ as the case where the optical dimension is down-sized to $g = 1$, i.e. the topology is made of a single, large group with $a = S$ routers. On the other extreme, we define $b = 1$ to describe a topology with $g = S$ groups, each composed of a single router ($a = 1$). In order to control a and g using b, we first need to identify the sizes of the electrical and optical dimensions of a *canonical* Dragonfly corresponding to $S_{desired}$. Noting that $ag \geq S_{desired}$ and that $g = a + 1$, we can write $S_{desired} \geq a(a+1)$. Equality is achieved when $a_{canonical} = \frac{-1+\sqrt{1+4S_{desired}}}{2}$. From there we can define:

$$a = \begin{cases} \lceil a_{canonical} - b(S_{desired} - a_{canonical}) \rceil & \text{when } -1 \leq b < 0 \\ \lceil 1 + (1-b)(a_{canonical} - 1) \rceil & \text{when } 0 \leq b \leq 1 \end{cases} \tag{2}$$

$$g = \lceil S_{desired}/a \rceil \tag{3}$$

The above equations do permit us to obtain (i) $a = S_{desired}$ and $g = 1$ when $b = -1$; (ii) $a = 1$ and $g = S_{desired}$ when $b = 1$; and (iii) a construction close to one of the *canonical* dragonflies for $b = 0$. In the last case, taking for instance $S_{desired} = 2000$, we have $a_{canonical} \simeq 45.22$ thus $a = \lceil a_{canonical} \rceil = 46$ and $g = \lceil S_{desired}/a \rceil = 44$.

However, for negative b values, a linear control of a with b is ineffective. Hence, for $-1 < b < -0.5$, Eq. 2 returns $S_{desired} - 1 > a > S_{desired}/2$. When introduced into Eq. 3, these values all return $g = 2$. To avoid this pitfall, we use b to control g instead of a for negative b values. First, we similarly obtain $g_{canonical} = \frac{1+\sqrt{1+4S_{desired}}}{2}$. We then modify Eq. 3 into:



(a) (b)

Fig. 4. (a) Effect of imbalance parameter b on Dragonfly parameters. (b) Illustration of the Dragonfly design space. Each point within this space represents a unique Dragonfly variant.

$$g = \lceil 1 + (b+1)(g_{canonical} - 1) \rceil, a = \lceil S_{desired}/g \rceil \text{ when } -1 \leq b < 0 \quad (4)$$

$$a = \lceil 1 + (1-b)(a_{canonical} - 1) \rceil, g = \lceil S_{desired}/a \rceil \text{ when } 0 \leq b \leq 1 \quad (5)$$

Fig. 4a shows the obtained a, g and S values for $S_{desired} = 1500$ as a function of b. Defined this way, Eqs. 4 and 5 allow b to control a and g values while minimizing $ag - S_{desired}$.

Having introduced the mapping of (b, d) to (a, g, h), we can represent the Dragonfly design space as a rectangular space with $x \in [-1, 1]$ and $y \in [0, 1]$. The corner cases in the design space are drawn in Fig. 4b. Along the $b = -1$ line, the obtained topology is an electrical full-mesh. Since the optical dimension is non-existent, topologies along this line are not affected by density d. At coordinate $(1, 0)$ we find an optical ring. An optical full-mesh appears at coordinate $(1, 1)$. Finally, along the $d = 1$ line, we find all the 2D-FB constructs of size $S_{desired}$, except for $b = -1$ or $b = 1$ where either g or a, respectively, equals 1. We can also reverse-evaluate the *imbalance* and *density* coefficients of the designs shown in Fig. 1. In Fig. 1a, the canonical Dragonfly logically maps to $(0, 0)$ while the 2D-FB in Fig. 1b maps to $(0, 1)$. The other topologies of Fig. 1 are also reproduced in Fig. 4b along with their corrresponding coordinates in the design space.

Figure 5a and b depict how the ratio of optical links is affected by the two parameters b and d. As expected, when imbalance is $b = -1$ or $b = 1$ the topology has only one dimension, which is either fully electrical or optical. Figure 5c shows how the topology diameter is influenced by the density and imbalance. For $b = -1$, the topology is an electrical full-mesh of diameter 1. For $b = 1$ with densities $d = 0.5$ and $d = 0.8$, the resulting topologies are not 2D-FB, but the wiring density is large enough to always conserve one of the two 2-hop paths between each node pairs that a regular 2D-FB offers, resulting in diameter 2 topologies. When density $d = 0$ and $b = 1$, the topology becomes a ring with a diameter of 750. Figure 5d and e depict the impact of parameters on the global

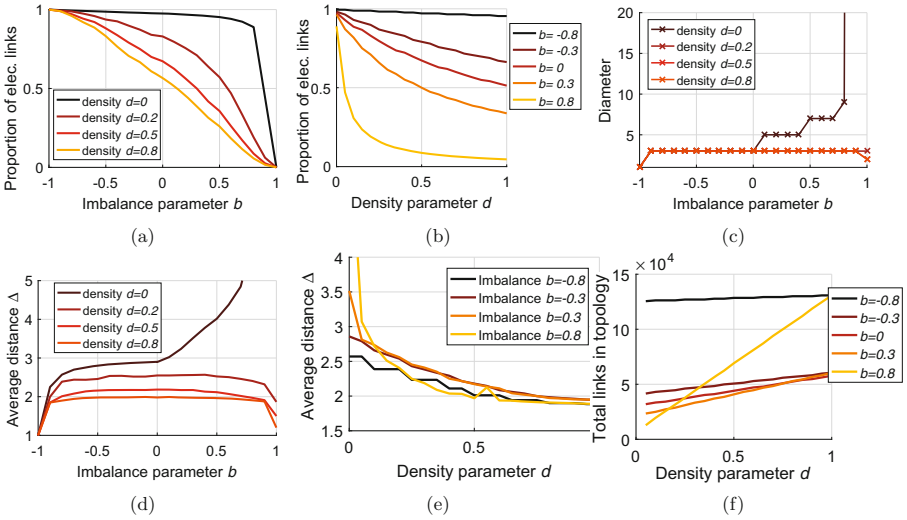


Fig. 5. Characteristics of Dragonfly topologies accommodating at least $S_{desired} = 1500$ routers.

average distance. As the imbalance leans toward negative values, Δ decreases, which is expected since more routers can be reached in 1 hop owing to the large intra-group electrical-mesh. Interestingly, positively imbalanced topologies also show lower Δ's than strictly balanced ones, provided enough *density* is given. This is mostly due to the high value that h can take when the number of groups g increases (as $h = \max(0, \lfloor 1 + d(g-2) \rfloor)$). Looking closer at the $b = 0.8$ case, we observe that the topology made from 167 groups translates into $h = 83$ when $d = 0.5$. The many inter-group links cause the vast majority of node pairs to be separated by two hops (electrical-optical, optical-electrical, and optical-optical). When $d = 1$ (2D-FB cases), graph diameter is at most 2, hence Δ converges to 1 as imbalance grows and the topology approaches a full-mesh.

These analyses highlight the diversity of Dragonfly designs, notably in terms of the proportion of optical links, average distance and diameter. However, this diversity also translates into a highly-varied total topological bandwidth (i.e. total number of links shown in Fig. 5f). Each topology thus possesses the ability to support different number of terminals (Fig. 5f) and corresponds to different implementation costs. In order to compare the diversely dense and balanced Dragonflies, we first show in the next section how to adapt our exploration space to exclusively identify topologies capable of accomodating a given number of terminals, $N_{desired}$. Then, in Sect. 4, we introduce a cost model to evaluate the cost of each design and elaborate on topologies supporting $N_{desired}$ terminals.

3 Constructing Dragonflies for a Minimal Number of End-Points

In our explorations so far, we have let the parameter p which denotes the number of terminals per router untouched. p is, however, a key factor in the Dragonfly construction, as it determines not only the final scalability of the topology, but also the required router radix. Moreover, we observe in Fig. 5f that the total number of links employed in the Dragonflies explored greatly varies with b and d, and consequently the available bandwidth of each topology also varies significantly. If a substantial amount of bandwidth is available within the topology, e.g. when the Dragonfly is heavily electrically-balanced ($b = -0.8$ as in Fig. 5f), we can populate the S routers with more terminals to ideally exploit the available bandwidth.

We can make the number of terminal attached to a router, p, proportional to the number of links attached to this same router $p \approx (a - 1 + h)$. This is the approach used in Kim et al. original Dragonfly proposal [1]. Since a Dragonfly is a diameter 3 topology, each transmitted bit is, in the worse case, forwarded twice onto a local link, once onto a global link, and once onto the destination's terminal link. This relationship gives us $p = \frac{a}{2} = h$. This approach, however, is too limited in our case, as our wiring algorithm may return topologies of variable diameter. Furthermore, for topologies strongly negatively-balanced (highly-negative b and large electrical groups), much of traffic remains within the groups, which contradicts the worst case assumption that every bit transits across groups.

To obtain a number of terminals p most suited to each of our designs, we start by remarking that the total traffic carried over a topology is proportional to the average path lengths (assuming no locality – every node pair have equal probability to exchange traffic). Thus, the total bandwidth made available by the topology should be proportional to Δ, and the number of traffic injectors should be inversely proportional to Δ. Since we cannot easily add bandwidth over the topology, we compensate Δ by changing p. This relationship can be expressed as follows:

$$p \approx \frac{S(a - 1 + h)}{\Delta} \tag{6}$$

In applying the methodology proposed by Rumley et al. [9], we can pick p such that the total traffic injected under uniform traffic must not exceed the total bandwidth installed, i.e. $N\Delta \leq S(a - 1 + h)$ which can be rewritten as:

$$p = \frac{N}{S} \leq \frac{(a - 1 + h)}{\Delta} \tag{7}$$

If we target an almost saturated topology under uniform traffic, $p_{selected} = \lfloor (a - 1 + h)/\Delta \rfloor$ terminals should be connected to every router. Note that the resulting network utilization (still under uniform traffic assumption) can be written as:

$$H = \frac{p_{selected}}{(\frac{(a-1+h)}{\Delta})} \tag{8}$$

If equality is reached in Eq. 7, utilization is maximal (100%). In constrast, when equality in Eq. 7 is not met, $p_{selected}$ is smaller than $\frac{a-1+h}{\Delta}$ due to rounding, and utilization is consequently driven down.

Equation 7 is not entirely satisfying as it implies that the number of routers, S, best suited to support N terminals is already known – either dictated by a, g and h, or, when using our exploration mechanisms, given as a parameter alongside b and d. The resulting total number of terminals supported $N = pS$ might thus clearly differ from the original $N_{desired}$ goal. We can circumvent this limitation by iteratively testing a sequence of p values. As soon as p is fixed, $S_{desired}$ can be obtained as $S_{desired} = \lceil N_{desired}/p \rceil$, a Dragonfly topology of parameters b, d and S can be produced and its global average distance Δ subsequently obtained, which ultimately permits us to evaluate the bandwidth utilization (Eq. 8). The value $p_{selected}$ for which the utilization is the closest to 1 should be retained. To find $p_{selected}$, we note that the utilization necessarily grows with p. Hence, for very small p values, the number of routers S is large, which results greater number of links. As p is increases, the Dragonfly topology shrinks and so does its bandwidth. There is necessarily a p_{excess} for which utilization exceeds 1. Finding the p that maximizes the utilization can thus simply be achieved by considering incremental integer p values until reaching p_{excess}. This is computationally acceptable as p is typically smaller than 50 for most Dragonfly designs. One may also cap p by the limiting the router radix which equals $p+h+a-1$. Most modern routers available in the market today (year 2017) are limited to radices of ≈ 100. Meanwhile, Δ can be easily obtained as a side product of the wiring algorithm.

It is important to recognize the limitations of Eq. 7, as it only considers p such that the *total* bandwidth can support a uniform traffic, but does not guarantee that this bandwidth is available where the highest congestion occurs. For instance, Eq. 7 would not hold when the topology is one with two large groups connected by a single optical link, since the single optical link would need to support roughly half the traffic. Even with uniform traffic injection, the optical link would be subjected to extreme congestion, bottlenecking the network bandwidth at a lower bound than what the right-hand side of Eq. 7 provides. To prevent such situations, the utilization of each link could be individually evaluated and p selected in a way that would ensure that every link's utilization is below 1.

Figure 6 reports the properties of many Dragonflies generated with the technique described above, all of which capable of supporting at least $N_{desired} = 10,000$ terminals. We first observe how the value p corresponding to highest utilization, H, varies across designs (Fig. 6a). Through the $S = \lceil N_{desired}/p \rceil$ relationship, the number of routers S (Fig. 6b) is also affected and not stable as previously seen in Fig. 4a. Notice that the changing of S and density parameter also significantly affects the shape of the a and g curves of Fig. 6c.

We observe that the global average distances Δ in Fig. 6d is very much comparable to the constant $S_{desired}$ case depicted in Fig. 5d. This is because the average distance is mostly related to the structure of the topology, hence to b

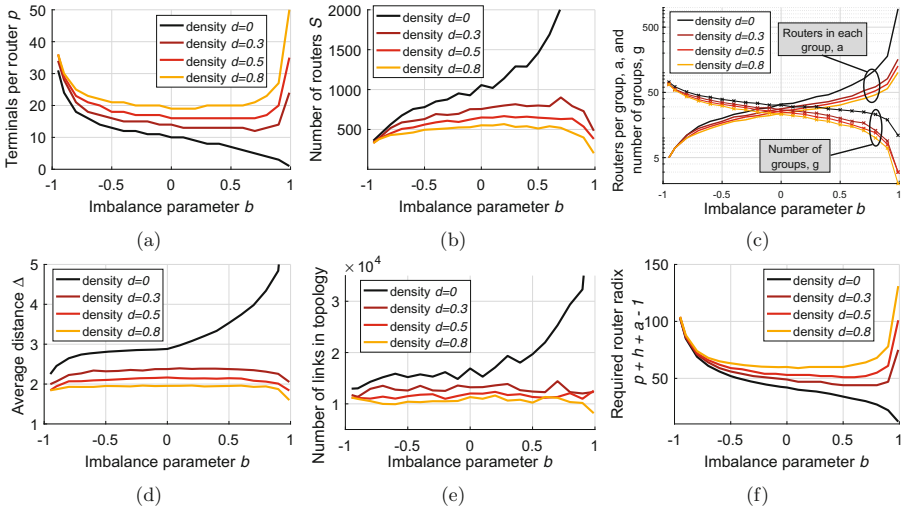


Fig. 6. Characteristics of Dragonfly topologies accommodating at least $N_{desired} = 10,000$.

and d, and only marginally related to its size. The shapes of the Δ curves propagates into the ones of p (Fig. 6a), as p is inversely proportional to Δ, and finally into the shapes of S. The number of links present in each topology (Fig. 6e) is also roughly proportional to Δ, and is overall less affected by the Dragonfly-"shaping" parameters b and d than previously when we explored topologies with a constant $S_{desired}$.

Figure 6f finally shows the impact of *imbalance* and *density* on the required radix. We note that when density is maximal, the radix requirements is minimized when topologies are balanced, which is a known property of Flattened-Butterflies. When density decreases, positively imbalanced topologies tend to favor low-radix routers. For minimal density $d = 0$, the required radix constantly decreases until the topology becomes a ring. It is interesting to note that designs with high b and low d becomes more favorable due to their low radix requirements. Figure 6b supports this as it shows that low router radices are required when there are more numerous routers in the Dragonfly. To clarify the value of these different option, we introduce in the next section a cost model for routers and links.

4 Design Selection via Cost Comparison

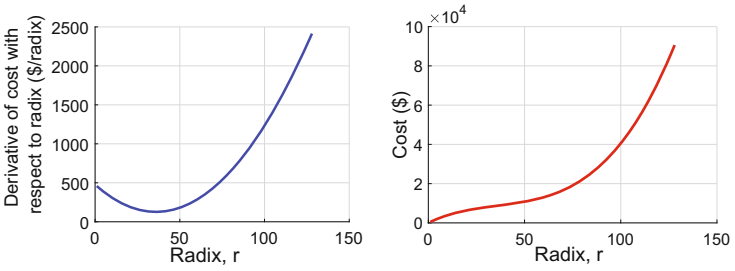
In this section we aim at estimating the cost a high-end HPC packet router switch of any radix. Based on pricing information available on ColfraxDirect [10], we considered a low-tier 24-port router currently priced at $7095, and a high-tier 48-port router at $10455, taken from the same supplier and both working at 100 Gb/s. These two data points are used to derive the following cost model.

We assume the marginal cost of adding a port to an existing router to be a U-shaped quadratic function with a minimum point at $radix = 36$. The rationales are the following: adding a port would benefit from economics of scale, but is also subject to technical complexity; the global minimum of the U-shaped curve correspond to the port count where the two effects negate each other. We place the minimum marginal cost in the middle of the low-tier and high-tier designs, assuming that with more resources, the supplier may incorporate a "mid-tier" 36-port router into its product line. Since this is not the case, two designs equally distant from the optimal cost will fulfill the market demands better. This causes the derivative of our cost model to be written as $\frac{d}{dr}cost(r) = c_1(r-36)^2 + c_2$, where c_1 and c_2 are constants. Solving for the polynomial constants using the discussed price points, we arrive at the following cost model:

$$cost(r) = 0.0901r^3 - 9.73r^2 + 477r \tag{9}$$

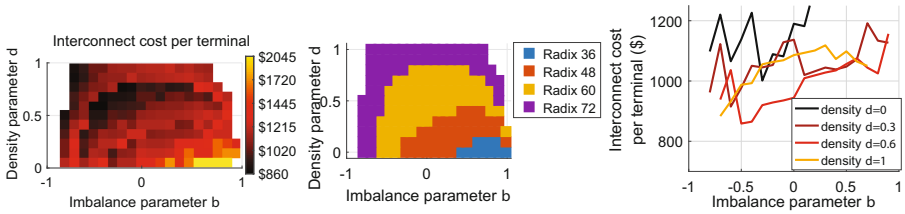
where r is the radix/port count of the router, and cost(r) is in the units of $'s. The resulting cost and it's derivative with respect to port-count for port counts between 0 and 128 are shown in Fig. 7a and b. We emphasize here that obtaining a model with a growing marginal cost per port is necessary to ensure that the router radix is not infinitely scalable. If the cost of a router is simply assumed a linear function of the number of ports, the cheapest topology becomes the one consisting of a single router with $N_{desired}$ ports. Provided that routers always have a radix multiple of 8 or 12, we then use this cost model to pinpoint the cost of routers with a range of radices. Logically, our model returns $7095 and $10,460 for 24-port and 48-port routers, respectively ($296 and $218 per port). A putative 64-port router is $14,320 ($228 per port). For 96 ports, this price grows to $35,884 ($374 per port).

For links, we consider a 100 Gb/s electrical link to be $80 [10]. As we are interested in analyzing the impact the optical/electrical cost ratio has on the Dragonfly topology selection, we consider optical links to have cost comprised between $80 (same as electrical) and $800 (ten times more expensive). As of today (2017), optical links are about five times more expensive than their electrical counterparts.



(a) (b)

Fig. 7. Cost model for predicting router price as a function of radix/port count



(a) (b) (c)

Fig. 8. Cost analysis of Dragonflies accomodating at least $N_{desired} = 10,000$ terminals

Results of the cost analysis are depicted in Fig. 8 for $N_{desired} = 10,000$, and considering radices of $[36, 48, 60, 72]$. Figure 8a shows how the cost evolves with the design space when considering \$400 for optical links. We note a correlation between Fig. 8a and b. The cheapest solutions are the ones that make the best use of the ports available. Figure 8c shows that the cheapest design found in our exploration is obtained for $b = -0.5$ and $d = 0.6$, which correspond to $g = 17$ groups of $a = 32$ routers, $p = 19$ terminals per router and $h = 10$ inter-group links per router. The proportion of electrical links to all links is 76%. We note that this cheapest design requires 60-port routers and dominates all designs requiring 72 ports. As expected, it is found in the negatively-balanced region that favors electrical links.

Figures 9a, b and c illustrate the cost per terminal by considering an optical link price of \$80, \$400 and \$800, respectively. We note that as the price of optics increases, negatively-balanced designs tend to become cheaper. Interestingly, in the presence of optical links that are equally expensive as electrical ones, six designs that achieve the cheapest cost are found at a cost of \$733.86 per terminal, with *densities* of 0.7 or 0.8, and *imbalance* spanning from -0.2 to 0.7. In the \$800 case, the cheapest design is a strongly imbalanced case ($b = -0.8$, $d = 0.5$) with only 10 groups made of 45 routers per group, and 23 terminals per router.

We complete our analysis by exploring designs supporting $N_{desired} = 25,000$ terminals (Fig. 9d). Here we assume radices of $[48, 64, 80]$ are available. We note first that the cost per terminal is slightly higher than that of the $N_{desired} = 10,000$ case, as the larger network scale incurs a cost premium. Even though we consider here \$400 for each optical link, it is still surprising to see the cheapest design being positively-balanced ($b = 0.2$). Our analyses show that for very large scale topologies, the positively-balanced designs emerge as among the cheapest options due to their lower radix requirements (as visible in Fig. 6f). In the $N_{desired} = 25,000$ case, the cheapest design found ($b = 0.2$ with a moderate density of $d = 0.3$) has 43 groups, 34 routers per group, $h = 13$ inter-group links per router, and $p = 18$ terminals per router. It still guarantees a high proportion of electrical links (72%), and requires routers with radix of 64.

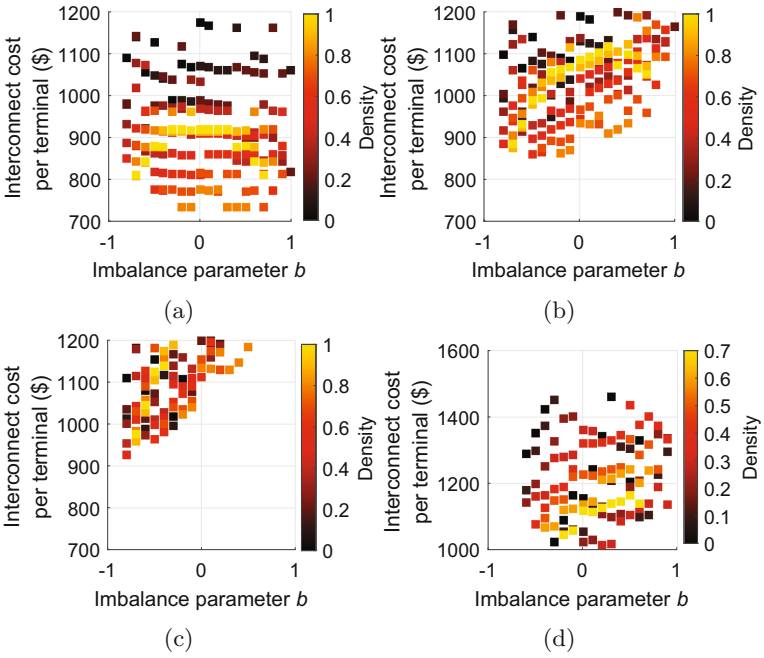


Fig. 9. Cost analysis for when optical links are set to (a) \$80, (b) \$400, (c) \$800 with $N_{desired} = 10,000$ and when optical links set to (d) \$400 when $N_{desired} = 25,000$

5 Conclusion

The Dragonfly topology, while having recently garnered much attention from the HPC community, have been subjected to different interpretations across literature. In this paper, we aim at formalizing the definition of a Dragonfly topology. To do so, we first state that any Dragonfly variant can be represented as 2D-Flattened Butterfly. In other words, a router can be represented in terms of its (x, y) coordinate in a 2-D lattice, where x (electrical dimension) represents the router's position in a group, and y (optical dimension) represents the group said router belongs to. Next, we introduce two Dragonfly-shaping parameters, namely: (a) the *imbalance* parameter, $b \in [-1, 1]$, which controls the relative sizes of the optical dimension to the electrical dimension, and (b) the *density* parameter, $d \in [0, 1]$, which controls a router's inter-group connectivity in the optical dimension. The space spanned by b and d creates the Dragonfly design space.

Using the wiring algorithm presented in Sect. 2.2, we generated various dragonflies in the design space, and subsequently identified several interesting designs. By studying dragonflies with 1500 routers, we found that as long as $d \neq 0$, the average global distance of the network remains fairly constant over the range of the *imbalances*, and only tending towards 1 when either the optical or electrical dimensions get downsized to 1. This is due to the topology approaching a

full-mesh (when optical dimension is downsized to 1) or a flattened-butterfly (when electrical dimension is downsized to 1). In general, the number of links in the topology also increases as b becomes more negative due to the larger electrical dimension, in which more router pairs are directly-linked as a result of the larger full-mesh intra-group topology.

We found that topologies with a *density* of 0 exhibit poor network characteristics, since each router only has one inter-group link at its disposal. This minimal connectivity in the optical dimension incurs a higher global average distance on these topologies, an effect that is even more pronounced as the optical dimension expands (*imbalance* tending to more positive values). Our results in Sect. 4 indicate that if given access to routers with higher radices, it is generally worth maximizing the utilization of the allocated port counts to obtain interconnect designs of more optimal costs. This can be done either by (a) expanding bandwidth in the electrical dimension by opting for more nagatively-balanced Dragonflies or by (b) expanding bandwidth in the optical dimension by opting for Dragonflies with higher *densities*.

Finally, the effects of varying the cost of optical links relative to electronic links on cost-efficiency are explored on dragonflies supporting $10,000$ terminals. Our results show that as the cost of optical links increases, negatively-balanced Dragonfly variants tend to be more cost-efficient due to their larger electrical dimension. A similar exploration done on topologies with $25,000$ terminals, however, showed that positively-balanced Dragonfly offer more cost-efficient designs, despite considering optical links at $5\times$ the cost of electrical links. These results unanimously indicate that *density* should generally be greater than 0 to yield cost-efficient designs with reasonable global average network distances. On the other hand, it is difficult to draw a conclusion on the range of *imbalance* that yields the most cost-optimal Dragonfly designs. We recognize that the regions in the design space corresponding to the most cost-optimal dragonflies vary significantly based on the targeted system scale (defined by number of terminals), the available router radix, and the cost of the network components (e.g. links and routers). However, the methodology employed to study cost-efficiency is valid, and we plan to investigate the ideal *imbalance* to scale relationship in the future by means of workload simulations.

References

1. Kim, J., Dally, W.J., Scott, S., Abts, D.: Technology-driven, highly-scalable dragonfly topology. In: 2008 International Symposium on Computer Architecture, pp. 77–88, June 2008
2. Kim, J., Dally, W., Abts, D.: Flattened butterfly: a cost-efficient topology for high-radix networks. In: Proceedings of the 34th Annual International Symposium on Computer Architecture, ISCA 2007, New York, NY, USA, pp. 126–137 (2007)
3. Alverson, B., Froese, E., Kaplan, L., Roweth, D.: Cray XC series network (2012), http://www.cray.com/sites/default/files/resources/CrayXcnetwork.pdf

4. Faanes, G., Bataineh, A., Roweth, D., Court, T., Froese, E., Alverson, B., Johnson, T., Kopnick, J., Higgins, M., Reinhard, J.: Cray cascade: a scalable HPC system based on a dragonfly network. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, SC 2012, Los Alamitos, CA, USA, pp. 103:1–103:9. IEEE Computer Society Press (2012)
5. Bhatele, A., Jain, N., Livnat, Y., Pascucci, V., Bremer, P.T.: Analyzing network health and congestion in dragonfly-based supercomputers. In: 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 93–102, May 2016
6. Jain, N., Bhatele, A., Ni, X., Wright, N.J., Kale, L.V.: Maximizing throughput on a dragonfly network. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2014, Piscataway, NJ, USA, pp. 336–347. IEEE Press (2014)
7. Wen, K., Samadi, P., Rumley, S., Chen, C.P., Shen, Y., Bahadroi, M., Bergman, K., Wilke, J.: Flexfly: enabling a reconfigurable dragonfly through silicon photonics. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2016, Piscataway, NJ, USA, pp. 15:1–15:12. IEEE Press (2016)
8. Camarero, C., Vallejo, E., Beivide, R.: Topological characterization of hamming and dragonfly networks and its implications on routing. ACM Trans. Archit. Code Optim. **11**, 39:1–39:25 (2014)
9. Rumley, S., Glick, M., Hammond, S.D., Rodrigues, A., Bergman, K.: Design methodology for optimizing optical interconnection networks in high performance systems. In: Kunkel, J.M., Ludwig, T. (eds.) ISC High Performance 2015. LNCS, vol. 9137, pp. 454–471. Springer, Cham (2015). doi:10.1007/978-3-319-20119-1_32
10. http://www.colfaxdirect.com/. Accessed 16 Apr 2017
11. Hastings, E., Rincon-Cruz, D., Spehlmann, M., Meyers, S., Bunde, D.P., Leung, V.J.: Comparing global link arrangements for dragonfly networks. In: 2015 IEEE International Conference on Cluster Computing, Chicago, IL, USA, pp. 361–370 (2015)