SAND2017-12241C

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,
[2] Scalable Modeling and Analysis, Sandia National Labs, Livermore, CA 94551, USA
mt3126@columbia.edu

Abstract. We investigate possible options to create a Dragonfly topology able to accommodate 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 introduce an algorithm that generates a dragonfly topology taking the desired number of end-points and these two parameters as input. We calculate a variety of metrics on generated topologies resulting from a 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 *group*. 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 to implement intra-group and terminal connections with short-distance, lower-cost electrical transmission links. In return, *inter-group* connections are based on optical equipment that is capable of spanning the tens of meters inter-cabinet distances.

Modularity is one of the main advantage provided by the dragonfly topology. Thanks to the clear distinction between intra- and inter-group links, the final number of groups present within one supercomputer does not affect the wiring within a group. Vendors can therefore propose all-included, all-equipped cabinets corresponding to a group, letting supercomputer operators 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/cabinets [2]. The fixed intra-group wiring also makes upgrading a dragonfly based supercomputer relatively straightforward from an hardware point-of-view, as only existing inter-group links might 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 the new ones, and the new ones among themselves.

A dragonfly topology also guarantees a large path diversity between end-points, enabling various flavors of adaptive, non-minimal routing schemes [1]. In presence of a congestion between two groups, traffic can 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 up to $g-2$, where g is the number of groups.

However, besides its modularity and capability to leverage non-minimal routing schemes, the Dragonfly topology takes into account the difference between electrical and optical cables. Although the price gap is shrinking, optical links are still largely more expensive than their electrical counterpart, and thus represent a considerable fraction of an interconnect 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 of the cabinets whereas the least expensive part (electrical) is “hard-wired” inside the cabinet. As not all parallel applications require the same balance between bandwidth and compute, adapting the bandwidth available 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 caused the Dragonfly topology to be the default choice for the whole XC series of Cray [3], and thus to be 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 [4] or optimizing throughput [5], and possible inclusion of optical switching [6]. One can note across the literature, however, varying ideas of what constitutes a Dragonfly. Here we aim at clarifying the Dragonfly definition, and at showing 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 is essentially a 2D-Flattened Butterfly (2D-FB), one dimension of which has been thinned (the one wired with optical cables). We then show that a Dragonfly topology can be described by a) varying sizes of the two dimensions of the “underlying” 2D-FB, and b) by the thinning of the optical dimension. Having reduced Dragonfly to two parameters, we scan many designs and perform thorough exploration of the Dragonfly design space. We finally analyze the value of the identified designs by mean of a cost model. Our analyses are related to the those reported by Camarero et al.[7], but with focus on practical insights rather than graph theory.

2 Dragonfly variants description and construction

2.1 Definitions

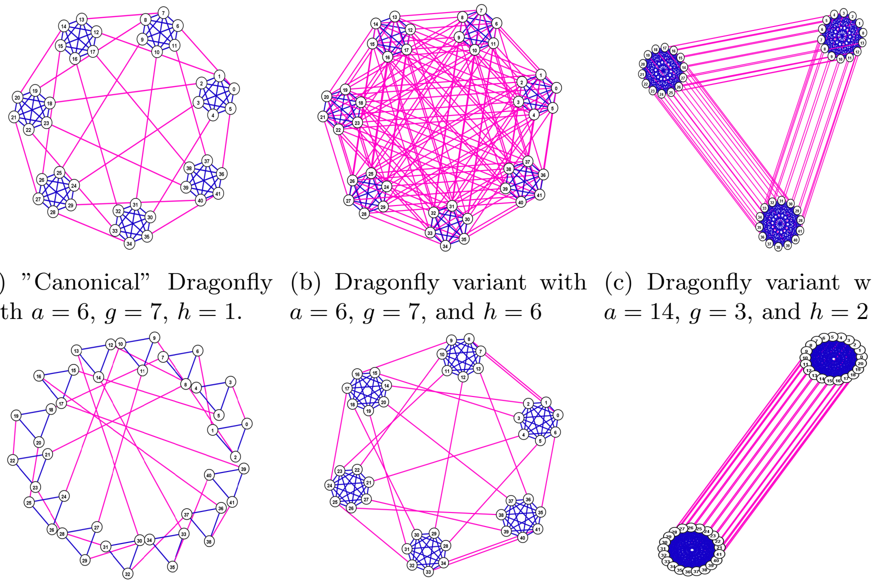
We begin by introducing a notation, much inspired by the one originally given by Kim et al. [1]. We consider a Dragonfly as being made of g groups with a routers in each group, therefore with a total of $S = ag$ routers. Each router is accommodating p *terminal* connections to end-points. Because we uniquely consider Dragonflies with fully-meshed intra-group connectivity, 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 average distance in the Dragonfly graph, i.e. the average of the minimal number of hops separating every possible node pair. 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 average global distance Δ, we also introduce δ_i as the average distance separating node i from other nodes.

We set the *imbalance* coefficient $b \in [-1, 1]$ to represent how the sizes of the optical and electrical dimensions match or mismatch, and $d \in [0, 1]$ that captures the extent to which the optical dimension is interconnected. These two parameters will be further described in subsection 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 groups. This contrasts with the case shown in Fig.1b which has the same g and a values but has $h = 6$ inter-group links per router. In this way, not only is every group connected to every other group, but additionally every router is directly connected to every other group (as $h = g - 1$). In that case, the Dragonfly becomes effectively a 2D-FB. Through this example, we see that every router can be characterized by a (x, y) coordinate, x giving the router position in the electrical dimension (i.e. within a group) and y 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 notes that the cases in Fig 1a and Fig1b have similar optical and electrical dimension sizes. We can therefore describe the *canonical* case as minimal optical wiring with routers identically distributed across electrical and optical dimensions.



(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 $h = 2$

(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 2-level Dragonfly variants parameterized using different combinations of a, g, and h but with $S_{\text{desired}} = 42$. Purple links represent inter-group optical links, while blue links represent intra-group electrical links.

Fig. 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 thus that each pair of group 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.

Fig. 1d shows an opposite case of very small electrical dimension ($a = 3$, $g = 14$). We note that even though more than 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. It is not straightforward to pick which 3 among 13 other groups to form an inter-group connection with, since there are many possible combinations. A similar problem of links/group mismatching is faced in the example of 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 sub-section. Finally, in the case shown in Fig. 1f, although h equals 1, is incidentally equals to $g-1$. The resulting topology is therefore a 2D-FB, and h can also not be made larger. Through these examples, we see that the design space for getting a Dragonfly with $S = 42$ is quite wide already, demonstrating the richness of designs when S scales to $1,000$ or higher.

2.3 Dragonfly Graph Wiring Algorithm

As discussed in the previous section, in order to explore the entire design space, we need to be able to generate a Dragonfly topology using any arbitrary combination of a, g, and h. Given this set of parameters, we would like to distribute the global links between groups as evenly as possible, in particular such that the diameter and the average global distance Δ are minimized, and maintain fairness by avoiding "unevenly' connected' nodes with varying δ_i average distances.

The problem of allocating inter-group links that achieving 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 V_i of each newly added link is taken in the set of vertices with the least number of connections so far. Note that the algorithm may select V_i even though one or more links have been awarded to the (V_k, V_i) pair. Once the link has been allocated to a group pair (V_k, V_i), the algorithm identifies the router with the least number of connections so far within groups k and i and allocates the link to said router. When the graph is sparsely occupied by edges, every group is equally likely to be picked to from a link with V_k, and global link allocation resembles the *relative global link* arrangement[10]. As the graph becomes more saturated with edges, the algorithm tends to distribute links in a fair way by selecting target groups with lowest global "reachability", thus making global link arrangement more random in all cases other than when $g = ah + 1$.

In the pseudocode listed below, η_{ij} is used to represent the total number of inter-group links connecting group i to group j (and is symmetric for i and j). μ_i denotes the score of the group i, which is used to account for both how many inter-group links the current target group k has already formed with group i (accounted for by $\sum \eta_{\mathrm{ij}}$ term), and how densely group i is globally-linked to other groups (η_{ki}).

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 $\eta_{\mathrm{ij}} := 0, \forall i, j \in V$
for $k \in V$ do
    for $d = 0, ..., a \times h$ do
        for $i \in V$ where $i \neq k$ do
            define $\mu_i := \sum_{j \in V} \eta_{\mathrm{ij}}, \forall\ i, j$ s.t $j \neq k$
            pick $i$ s.t $\mu_i := \min_{i' \in V} \mu_{i'}$ and $\sum_{j \in V} \eta_{\mathrm{ij}} < (a \times h)$
            $\eta_{\mathrm{ik}} := \eta_{\mathrm{ik}} + 1$
        end for
    end for
end for
```

We evaluated the topologies obtained with our wiring algorithm, in terms of average global distance Δ, diameter, and fairness. To measure the fairness level, we consider two metrics: the first identifies δ_{min} and δ_{max} among all δ values, i.e. the average distances seen from the most and least centered node, respectively, and calculate the greatest difference in percent ($d = 100\frac{\delta_{max} - \delta_{min}}{\delta_{min}}$). The second metric calculates the squared coefficient of variation across

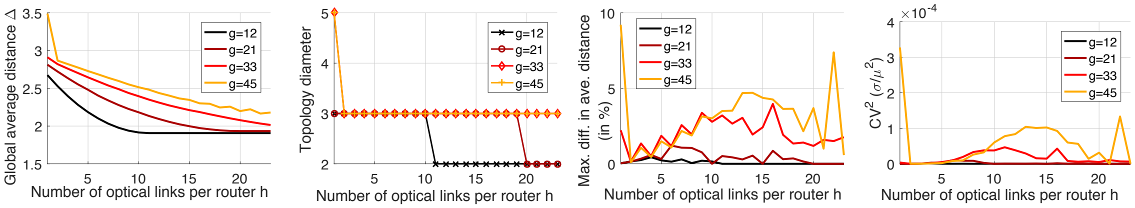


Fig. 2: (a) Global average distance $\varDelta$, (b) topology diameter, (c) maximum difference between smaller and larger node average distance δ_i, and (d) squared coefficient of variation of δ_i.

the δ_i set. Results for a set of topologies with at least $S_{desired} = 1000$ are displayed in Fig. 2. We observe that average distances $\varDelta$ 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 links to "reach" routers in other groups, as opposed to reaching them directly via the intra-group electrical links. This translates into a larger $\varDelta$ 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 $\varDelta$. 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.

Fig. 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 $(g = 12, h = 15)$, $(g = 21, h = 15)$ and $(g = 21, h = 10)$ the average distance 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 $(g = 21, h = 10)$: there are 480 inter-group links, which is also a round multiple of 20.

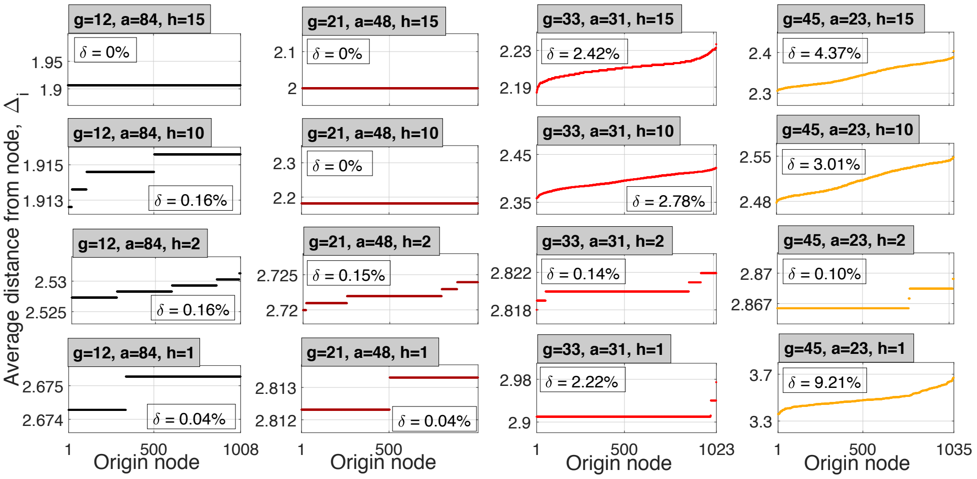


Fig. 3: Distributions of average distances of graph as viewed from each node. δ in each plot denotes the coefficient of variation for the average distance.

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). The routers present in these pairs are consequently favored. Looking at the general behavior on Figures 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 how the sizes of the optical and electrical dimensions mismatch, and the density coefficient $d \in [0, 1]$ captures to which extent the optical dimension is interconnected. The density d parameter implicitly controls h through:

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

We observe that for $g = 1$, h is always null (no inter-group links). 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 by extension g) by b, we first need to identify the sizes of the ideal 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 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 later case, taking for instance $S_{desired} = 2000$, we have $a_{canonical} \simeq 45.22$ thus $a = \lceil a_{canonical} \rceil = 46$ and $g = \lceil S_{desired}/a \rceil = 44$.

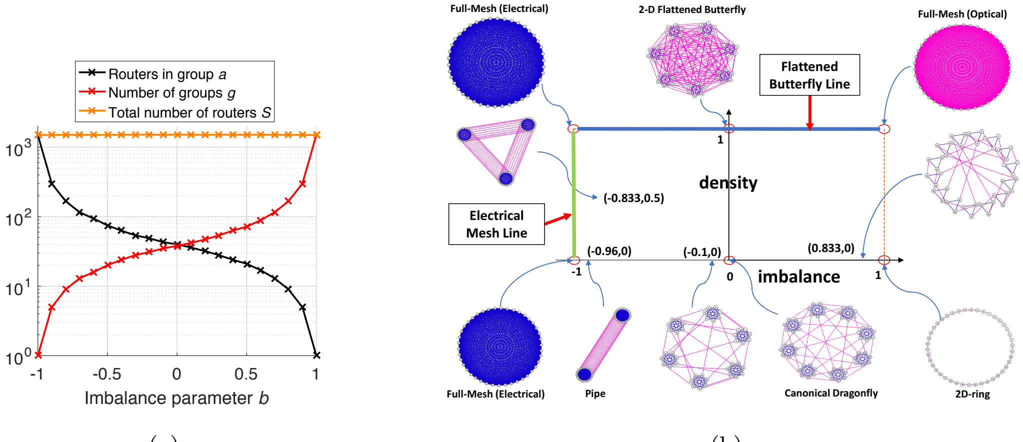
For negative b values, a linear control of a with b was ineffective. Hence, for $-1 < b < -0.5$, eq. 2 returns $S_{desired} - 1 > a > S_{desired}/2$. When introduced into eq. 3, these values all return $g = 2$. To avoid this pitfall, we use b to control g instead of a for negative b values. First, we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

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

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

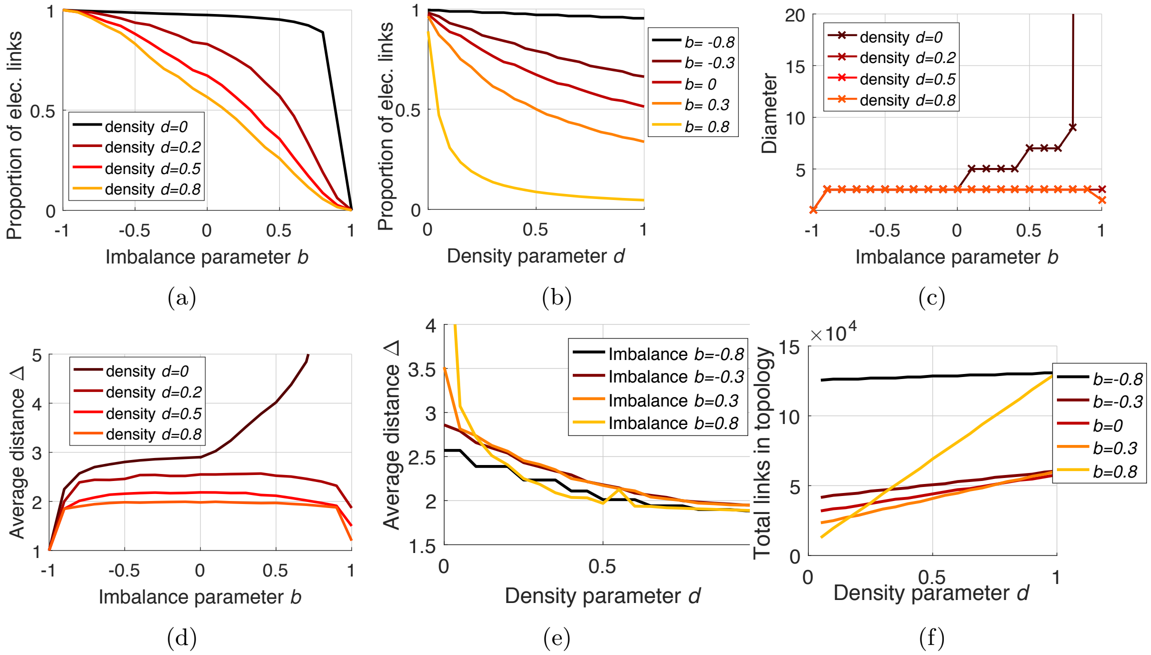
Fig. 4a shows obtained a, g and S values for $S_{desired} = 1500$, as a function of b. Defined this way, Eqn 4 and 5 allows b to control a and g values while minimizing $ag - S_{desired}$.

Having introduced the mapping of (b, d) to (a, g, h), we can represent the Dragonfly design space as a rectangle from $(-1, 0)$ to $(1, 1)$. The corner-cases of these design spaces 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. At coordinate $(1, 0)$ we find an optical ring. Finally, an optical full-mesh appears at coordinate $(1, 1)$. We can also reverse evaluate the imbalance and density coefficients of the designs shown in 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 with their corrresponding coordinates.



(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".



(a) (b) (c)

(d) (e) (f)

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

Fig. 5a and 5b 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. Fig. 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-hops paths between each node pairs that a regular 2D-FB offers, resulting into a diameter of 2. When density $d = 0$ and $b = 1$, the topology becomes a ring with a diameter of 750. Fig. 5d and 5e depict the impact of parameters on the average distance. As the imbalance leans toward negative values, Δ decreases, which is expected: more routers can be reached in 1 hop through the large intra-group electrical-mesh. Interestingly, positively imbalanced topologies also show a lower Δ 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 case $b = 0.8$, 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, and $\varDelta$ 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, each of which possessing 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 includes different topologies that are all capable of accommodating a similar number of terminals $N_{desired}$. Then, in Section 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 d and b, and consequently so does the bandwidth made available. If a substantial amount of bandwidth is available within the topology, e.g. when the Dragonfly is clearly electrically balanced ($b = -0.8$ as in Fig. 5f), it is interesting to 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]. A Dragonfly being of diameter 3, 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. From this stems that $p = \frac{a}{2} = h$. This approach, however, is too limited in our case, as our wiring algorithm may return topologies of variable diameter. Furthermore, for topologies with strongly negative (large electrical groups) imbalance b, much of traffic remains within the groups which contradicts the worse 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 pairs have equal probability to exchange traffic). Thus, either the total bandwidth made available by the topology should be proportional to $\varDelta$, or the number of traffic injectors should be inversely proportional to $\varDelta$. Since we cannot easily add bandwidth over the topology, we compensate $\varDelta$ by changing p:

$$p \approx \frac{S(a - 1 + h_{eff})}{\varDelta}, \text{ where } h_{eff} = k \times h \tag{6}$$

Here we introduce the optical link redundancy factor, k, which represents how many optical cables connect two routers. Note that the introduction of the redundancy factor does not affect the characteristics of a given topology aside from introducing more bandwidth between two routers. In applying the methodology proposed by Rumley et. al[8], we can pick p such that the total traffic injected under uniform traffic must not exceed the total bandwidth installed $N\varDelta \leq S(a - 1 + h)$ which can be rewritten as:

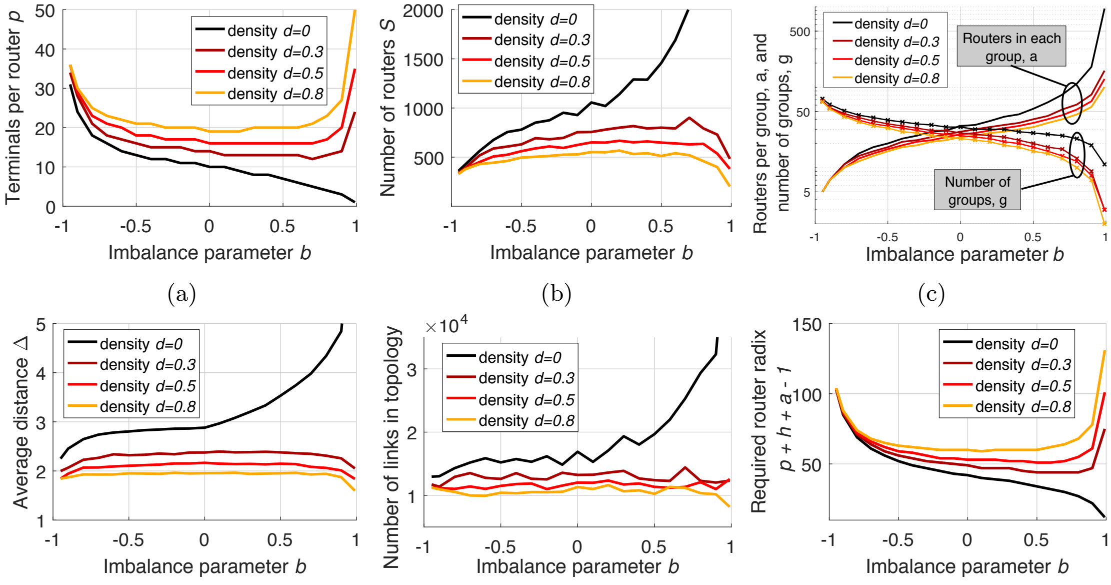
$$p = \frac{N}{S} \leq \frac{(a - 1 + h_{eff})}{\varDelta} \tag{7}$$

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

$$H = \frac{p_{selected}}{\left(\frac{(a-1+h_{eff})}{\Delta} \right)} \tag{8}$$

If equality is reached in Eq. 7, utilization is maximal (100%). Otherwise, $p_{selected}$ is the largest integer smaller than $\frac{a-1+h_{eff}}{\Delta}$, in which case utilization is less than 1.

Eq. 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 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, its average distance Δ can be 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 increased, 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 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 router radix which equals $p + h_{eff} + a - 1$. Most modern routers available in the market today (year 2017) is limited to radices of ≈ 100. Meanwhile, Δ can be easily obtained as a side product of the wiring algorithm.



(a)

(b)

(c)

(d)

(e)

(f)

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

It is important to recognize the limitations of Eq. 7, as it only considers p such that the *global* 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 is subjected to extreme congestion, bottlenecking the network bandwidth at a lower bound than what the right side of Eq. 7 provides. To prevent such situations, the bottleneck link should be identified and p chosen in such a way that ensures the utilization of the said link is below 1.

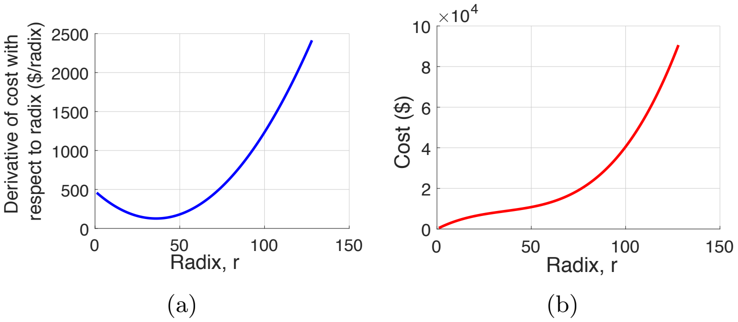
Fig. 6 reports the properties of many Dragonflies generated with the technique described above, all of which capable of supporting at least $N_{desired} = 10,000$ terminals. We first observe how the maximal number of terminals per router p varies across designs (Fig. 6a). Through the $S = \lceil N_{desired}/p \rceil$ 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 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 and d, and marginally related to its size. The shapes of the Δ curves propagates into the ones of p (Fig. 6a), p being 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 overall less affected by the Dragonfly "shaping" parameters b and d than previously when exploring topologies with constant $S_{desired}$.

Fig. 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 lower radix routers. For minimal density $d = 0$, the required radix constantly decreases until the topology becomes a ring. It is interesting then to note that designs with high b and low d becomes more favorable due to their limited radix requirements. Fig. 6b supports this as it shows that low router radix 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[9], 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 working at 100Gb/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 in 36. The rationales are the following: adding a port would benefit from economics of scale, but is also subject to technical complexity; the 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 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



(a)

(b)

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

using the discussed price points, we arrive at the following cost model:

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

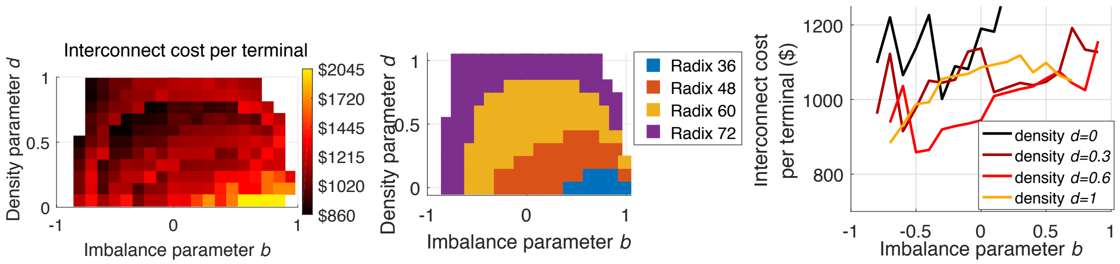
where r is the radix/port count of the router, and cost(r) is in the units of \$'s. The resulting cost and it's derivative with respect to port-count for port counts between 0 and 128 are shown in Fig. 7a and 7b. We emphasize here that obtaining a model with a growing marginal cost per port is necessary 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 pin-point the cost of a range of routers. 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 100Gb/s electrical link to be \$80 [9]. 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.

Results of the cost analysis are depicted in Fig. 8 for $N_{desired} = 10,000$, and considering radixes of $[36, 48, 60, 72]$. Fig 8a shows how the cost evolves with the design space when considering \$400 for optical links. We note a correlation between Fig 8a and Fig. 8b. The cheapest solutions are the ones that make the best use of the ports available. Fig. 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, $p = 19$ terminal per router and $h = 10$ inter-group link per router. The proportion of electrical 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 negative imbalance region that favors electrical links.

Figures 9a, 9b and 9c illustrate the cost per terminal considering an optical link price of \$80, \$400 and \$800. We note that as the price of optics increases, negatively imbalanced designs tend to become cheaper. Interestingly, in the presence of equally expensive electrical and optical links, six designs achieve the cheapest cost found (\$733.86), with densities of 0.7 or 0.8, and imbalance spanning from -0.2 and 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, and 23 terminals per router.

We complete our analysis by exploring designs supporting $N_{desired} = 25,000$ terminals (Fig. 9d). Here we assume radixes of $[48, 64, 80]$ are available. We note first that the cost per terminal is slightly higher than for 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 can be surprising to see the cheapest design to be positively balanced ($b = 0.2$). Our analyses show that for very large scale topologies, the positively balanced designs emerges as among the cheapest options due to their lowe radix requirements (as visible in Fig. 6f). In the $N_{desired} = 25,000$ case, the cheapest



(a)

(b)

(c)

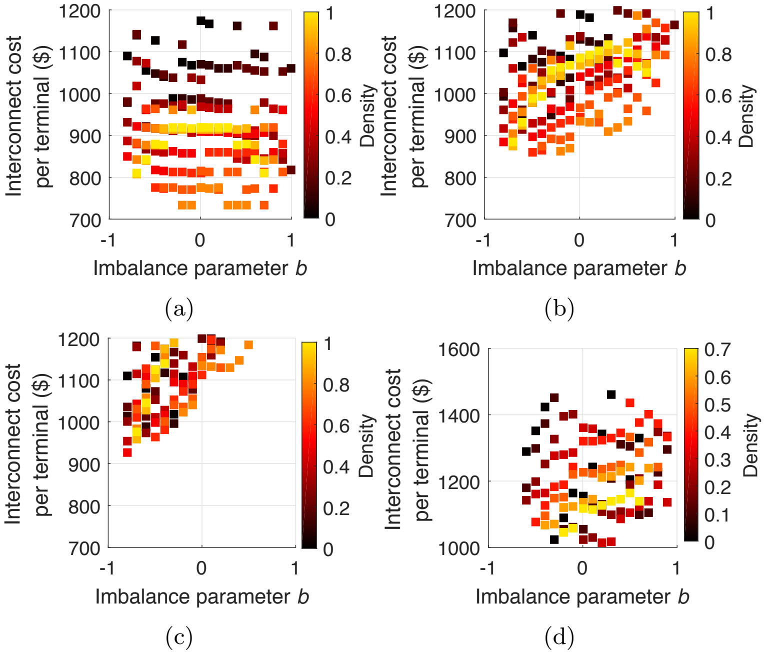
Fig. 8: Cost analysis of Dragonflies accomodating at least $N_{desired} = 10,000$ terminals

design found ($b = 0.2$ combined with a moderate density of $d = 0.3$) has 43 groups, 34 routers per group, $h = 13$ inter-group links, and $p = 18$ terminals per router. It still guarantees a high ratio of electrical links (72%), and requires a radix of 64.

5 Conclusion

The Dragonfly topology, though recently being widely studied due to its low diameter and versatile characteristics, have not been well-formalized.In this paper, we first introduce an algorithm that connects the all the routers in the Dragonfly topology of any arbitrary a, g, and h combinations as introduced by Kim et. al[1] in the fairest possible way. Then, we introduced two network-size-independent parameters that describes all the variations of the Dragonfly topology: the imbalance parameter, b, and the density parameter, d. The imbalance parameter controls how large a Dragonfly group is, and by extension, the proportion of electrical links that is used in the entire topology. The density parameter controls the proportion of optical links utilized by the topology. These two parameters will then be mapped into a, g, and h parameters which the wiring algorithm will utilize to form the described Dragonfly topology. Using this methodology, we presented the results of analyzing the topological characteristics of a Dragonfly network with $S_{desired} = 1500$ for various combinations of b and d within the design space.

Next, we introduce a cost model that predicts the price for routers of any arbitrary size based on commercially available price points with known radices. Using this model, we explored the various Dragonfly network sizes that supports a minimal number of terminals, $N_{desired}$, subject to the constraint that the network must have sufficient global bandwidth to support a uniform traffic injection by $N_{desired}$ terminals. Using $N_{desired} = 10,000$, our results show that the most cost-effective Dragonflies tends to be slightly negatively-imbalanced with larger groups and larger electrical dimension, as shown in Fig 8c. However, the same cannot be said when the model is used on $N_{desired} = 25,000$, in which case slightly positively-balanced topologies with smaller router groups and larger optical dimension are favored.



(a)

(b)

(c)

(d)

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 $N_{desired} = 25,000$

6 Acknowledgments

This worked was supported by Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy?s National Nuclear Security Administration under contract DE-NA-0003525.

References

1. J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *2008 International Symposium on Computer Architecture*, pp. 77–88, June 2008.
2. B. Alverson, E. Froese, L. Kaplan, and D. Roweth, *Cray xc series network.* 2012. http://www.cray.com/sites/default/files/resources/CrayXcnetwork.pdf.
3. G. Faanes, A. Bataineh, D. Roweth, T. Court, E. Froese, B. Alverson, T. Johnson, J. Kopnick, M. Higgins, and J. Reinhard, "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 '12, (Los Alamitos, CA, USA), pp. 103:1–103:9, IEEE Computer Society Press, 2012.
4. A. Bhatele, N. Jain, Y. Livnat, V. Pascucci, and P. T. Bremer, "Analyzing network health and congestion in dragonfly-based supercomputers," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 93–102, May 2016.
5. N. Jain, A. Bhatele, X. Ni, N. J. Wright, and L. V. Kale, "Maximizing throughput on a dragonfly network," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, (Piscataway, NJ, USA), pp. 336–347, IEEE Press, 2014.
6. K. Wen, P. Samadi, S. Rumley, C. P. Chen, Y. Shen, M. Bahadroi, K. Bergman, and J. Wilke, "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '16, (Piscataway, NJ, USA), pp. 15:1–15:12, IEEE Press, 2016.
7. C. Camarero, E. Vallejo, and R. Beivide, "Topological characterization of hamming and dragonfly networks and its implications on routing," *ACM Trans. Archit. Code Optim.*, vol. 11, pp. 39:1–39:25, Dec. 2014.
8. S. Rumley, M. Glick, S. D. Hammond, A. Rodrigues, and K. Bergman, *Design Methodology for Optimizing Optical Interconnection Networks in High Performance Systems*, pp. 454–471. Cham: Springer International Publishing, 2015.
9. http://www.colfaxdirect.com/, Accessed: 2017-04-16.
10. E. Hastings, D. Rincon-Cruz, M. Spehlmann, S. Meyers, D. P. Bunde, and V. J. Leung, "Comparng Global Link Arrangements for Dragonfly Networks," in *2015 IEEE International Conference on Cluster Computing*, (Chicago, IL, USA), pp. 361-370, 2015