Shortest paths in Dragonfly systems

Ryland Curtsinger
Knox College
rbcurtsinger@knox.edu

David P. Bunde
Knox College
dbunde@knox.edu

***Abstract*— Dragonfly is a topology for high-performance computer systems designed to exploit technology trends and meet challenging system constraints, particularly on power. In a Dragonfly system, compute nodes are attached to switches, the switches are organized into groups, and the network is organized as a two-level clique, with an edge between every switch in a group and an edge between every pair of groups. This means that every pair of switches is separated by at most three hops, one within a source group, one from the source group to the destination group, and one within the destination group. Routing using paths of this form is typically called "minimal routing". In this paper, we show that the resulting paths are not always the shortest possible. We then propose a new class of paths that can be used without additional networking hardware and count its members that are shorter than or of equal length to these "minimal paths".**

I. Introduction

Dragonfly [1] is a network topology designed to meet the challenges of building an exascale computing system [2]. It is a hierarchical topology. Compute nodes are attached to networking switches. These are organized into groups, all members of which are connected with a *local link*. The groups are also all connected, with a single *global link* between each pair of groups. Figure 1 shows the switches and global links of a Dragonfly network; each switch is a solid box, each group is a dashed box, and we omit the compute nodes and local links to simplify the diagram. Group numbers are around the outside and the number of each switch within its group is the switch label.

The size of a Dragonfly system is determined by the following parameters:

- p number of compute nodes on each switch
- a number of switches in each group
- h number of global links per switch

A (p, a, h)-Dragonfly is one having those parameters. Since every pair of groups is connected by exactly one global link, it consists of $g = ah + 1$ groups, $ag = a(ah + 1)$ switches, and a total of $pag = pa(ah+1)$ nodes. We use $S_{i,j}$ to denote switch j of group i; the switch is numbered within the group.

The Dragonfly topology has attracted much attention because it has constant diameter; every pair of switches is connected by a path of length at most three. Such a path can be found using "minimal routing" [1]. To travel between switches, this algorithm takes the local link within the source group to the switch connected to the destination group, then the global link to the destination group, and finally the local link to the destination switch itself. We call this algorithm *HM routing* and its paths *HM paths* since they are "hierarchically minimal", i.e. minimal at the intra-group and inter-group levels. We also classify paths by the sequence of edge types used, so a longest HM path is a *local-global-local* (abbreviated *lgl*) path. Sometimes a local or global link is unnecessary, such as if the source switch is connected to the destination group. In this case, a *global-local* (*gl*) path suffices. Similarly, HM paths can be *local-global* (*lg*), *local* (*l*), or *global* (*g*).

To deal with adversarial traffic, a Dragonfly system can use Valiant routing [1], inspired by work of Valiant [3]. The idea is to route each packet to a randomly-chosen intermediate group and then to its destination. We call such a path a *Valiant path*. A Valiant path has length at most 5 hops; the packet may need a local and global hop to reach the intermediate group, a local hop within the intermediate group, and then a global and local hop to the destination switch. The total path would have type lglgl, with shorter paths possible if not all these hops are needed.

In practice, most systems would use an adaptive routing scheme. There are many of these (e.g. [4], [5], [6]), but a common framework is to assess the level of congestion on a per-packet basis, using the HM path when it is not highly congested but switching to a Valiant path if necessary.

Our work begins with the observation that HM paths are not guaranteed to be minimal because switches can have a two-hop *global-global* (*gg*) path and an lgl HM path. This was previously observed [7]. An example of it can be seen in Figure 1; switches $S_{0,1}$ and $S_{1,1}$ have a gg path through $S_{4,2}$ but the HM path uses switches $S_{0,0}$ and $S_{1,3}$.

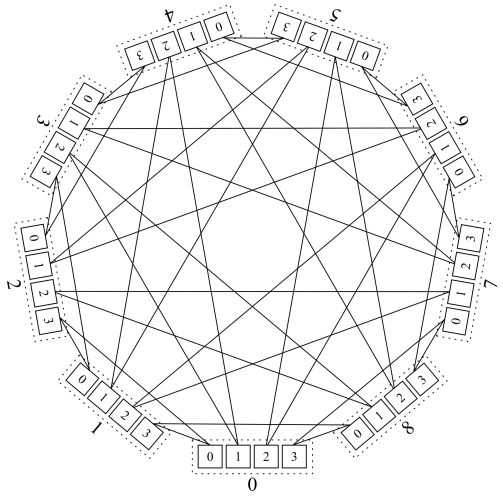

Fig. 1: $(p, 4, 2)$-Dragonfly with relative arrangement.

Arrangement	# destinations w/ shorter paths		# destinations w/ equal-length paths	
Relative	0	if $j = 0$ or $a - 1$	$2ah - a - h$	if $j = 0$ or $a - 1$
	$2h - 2$	otherwise	$4ah - 3a - 10h + 9$	otherwise
Absolute	0	if $j = \lfloor i/h \rfloor$ or $\lfloor (i-1)/h \rfloor$	$\geq h(a-1)$	
	$h - 1$	otherwise		

Fig. 2: Number of destination switches with an SV path from switch $S_{i,j}$ that is shorter or equal-length to the HM path

In this paper, we investigate the number of shortest paths not identified by HM routing. These paths might be strictly shorter, such as the gg path that replaces the lgl path above, or paths that are the same length. Even the latter are potentially useful to avoid congestion or support splitting traffic among shortest paths, following the ideas of [8].

We restrict our attention to paths that can be routed using 2 VCs on global links since that many is already required for Valiant routing (or any adaptive routing scheme incorporating it). These paths do not require extra hardware when using the standard deadlock avoidance strategy of increasing the VC index. They fall into four categories: gg paths, *global-global-local (ggl)* paths, *local-global-global (lgg)* paths, and *global-local-global (glg)* paths. We refer to paths in these categories that are no longer than HM paths as *SV paths*, for *short Valiant paths*, since they require only the hardware used for Valiant routing. In fact, each of these paths is a Valiant path with the correct choice of intermediate group.

The number of SV paths depends on the system's *global link arrangement*, the specific endpoints of each global link. The choice of global link arrangement was ignored when Dragonfly was first proposed, but Camarero et al. [7] identified three of them and Belka et al. [9] later proposed two more. Research on Dragonfly systems often does not specify the global link arrangement used, but typically uses either the relative or absolute arrangement (both defined later).

Figure 2 summarizes our results for these global link arrangements, presenting the number of destination switches from which switch $S_{i,j}$ has a SV path that is no longer than its HM path. For perspective, recall that there are $ag = a(ah + 1)$ switches in the system so the total number of possible destinations is one less than this. Thus, for all but two switches in each group, switches in the relative arrangement have strictly shorter SV paths to roughly $2/a^2$ of the other switches. In both of the relative and absolute arrangements, all switches have equal-length SV paths to between $\approx 1/a$ and $\approx 4/a$ of the others. For a system with a few thousand nodes, a is likely in the teens and these small fractions represent hundreds of nodes.

The other global link arrangements are more complicated to analyze than those shown in Figure 2. For the circulant arrangement, we give the number of shorter paths as the difference between two piece-wise linear functions (see Theorem 6 for details) and conjecture that the number of equal-length paths is $\approx 3ah$ (see Section V). For the other two arrangements, we outline a possible analysis and present the numbers of paths for specific sizes.

The rest of this paper is organized as follows. Section II defines terminology. Sections III–VII give results for each global link arrangement. Section VIII presents related work. Finally, Section IX discusses the potential use of SV paths in routing and future work.

II. Terminology

When defining the global link arrangements, we visualize the switches in a circle as shown in Figure 1, with the groups numbered in clockwise order. Conceptually, the switches in a group form a single *virtual switch* with ah ports connected with the virtual switches of other groups to form a clique. These ports are numbered $0 \ldots (ah - 1)$.

For simplicity, we use modular arithmetic on switch and group numbers so $S_{i,j}$, $S_{i,j+a}$, and $S_{i+g,j}$ all denote the same switch. We also define $\bmod$ so $a \bmod b$ is the smallest non-negative integer equal to $a - qb$ for integral q.

In addition, we define $R_x(S_{i,j})$ to be the set of switches reachable from $S_{i,j}$ using edges of type x. Similarly, we define $R'_x(S_{i,j})$ to be the set of groups with a switch in $R_x(S_{i,j})$. For example, $R_{lg}(S_{i,j})$ is the switches reachable from $S_{i,j}$ via an lg path and $R'_{lg}(S_{i,j})$ is their groups. Let $\delta_{<,x}(S_{i,j})$ be the number of switches other than $S_{i,j}$ whose x path from $S_{i,j}$ is shorter than the HM path from $S_{i,j}$. (Here "shorter" means fewer hops, regardless of the link types used.) If x is replaced by a list, then the function gives the number with any of those shorter paths. If x is omitted, the function counts all SV paths. Analogous functions using $=$ and $\leq$ instead of $<$ give the count of switches whose paths have the corresponding relationship with the length of the HM path respectively. For all these functions, we omit the source switch $S_{i,j}$ when it is clear from context.

III. Results for relative arrangement

Now we begin counting the SV paths for each kind of global link arrangement. In the *relative arrangement*[1] [7], port k of a group connects to the group $k + 1$ positions further along (i.e. port k of group i connects to group $i + k + 1$). Alternately, its port $ah - k$ connects to the group k positions behind it (i.e. to group $i - k$ from group i). By these two relationships, all neighbors of a group's j^{th} switch are the $(a - 1 - j)^{\text{th}}$ switches of their respective groups. For simplicity, we define $\bar{j} = a - 1 - j$ and observe that $\bar{\bar{j}} = j$. Figure 1 shows a sample of the relative arrangement.

Its symmetry makes the relative arrangement the easiest to analyze; rotating Figure 1 changes only the group numbers.

[1]Also called the *palmtree arrangement*; we use the name from [10]

We will show that switches at the edge of their group (i.e. those numbered 0 or $(a-1)$) have different numbers of SV paths than the others. We call the former *terminal* and the others *non-terminal*.

Now we identify switches reachable in one global hop:

$$R_g(S_{i,j}) = \{S_{x,\bar{j}} : i + jh + 1 \leq x \leq i + jh + h\} \quad (1)$$

The switch numbers come from the arrangement definition. For the group numbers, observe that a switch j connects to h groups starting $jh+1$ groups forward from its own.

A. 2-hop Valiant paths

To compute R_{gg}, apply Equation 1 twice. We observed above that $\bar{\bar{j}} = j$. The smallest group number comes from selecting the low end of the range both times. Thus, it is $i + jh + 1 + (a - 1 - j)h + 1 = i + ah - h + 2$, which is $i - h + 1$ since group numbers are taken mod g. Adding the analogous calculation for the largest group number gives

$$R_{gg}(S_{i,j}) = \{S_{x,j} : i - h + 1 \leq x \leq i + h - 1\} \quad (2)$$

From this, we count destinations with short Valiant paths:

Theorem 1: In the relative arrangement, if $S_{i,j}$ is terminal, then $\delta_{<,gg} = 0$ and $\delta_{=,gg} = 2h - 2$. If $S_{i,j}$ is non-terminal, $\delta_{<,gg} = 2h - 2$ and $\delta_{=,gg} = 0$.

Proof: $|R_{gg}| = 2h - 1$, but this includes $S_{i,j}$ so the number of other switches is only $2h - 2$.

To compare path lengths, observe that groups in R'_{gg} are all adjacent to the terminal switches of group i and global links connect switch 0 with switch $a - 1$ and vice versa. If $S_{i,j}$ is terminal, the HM path to members of R_{gg} is either a gl or an lg path. Either way, the HM path is the same length as the gg path. If $S_{i,j}$ is non-terminal, the HM path is 3 hops since the source and destination groups connect on terminal switches, but the source and destination switches are not terminal. ■

B. 3-hop Valiant paths

There are three kinds of three-hop Valiant paths: lgg, ggl, and glg. The first two kinds give the same set of destinations:

Lemma 1: In the relative arrangement from $S_{i,j}$,

$$R_{lgg} = R_{ggl} = \{S_{x,n} : x \in R'_{gg}, n \neq j\}$$

These are the $(2h-2)(a-1)$ switches in the groups of R'_{gg} not in R_{gg}.

Proof: For R_{lgg}, observe that a local hop reaches any switch $S_{i,n} \neq S_{i,j}$ and then two global hops reaches $R_{gg}(S_{i,n})$. R'_{gg} is the same for all switches in a group, so this reaches the switches $n \neq j$ in each of the groups in R'_{gg}.

For R_{ggl}, recall that two global hops reaches R_{gg}. Then, a local hop reaches any other switch in those groups. ■

Because HM paths have length at most three, none of these Valiant paths is shorter, but some have equal length.

Theorem 2: In the relative arrangement, $\delta_{=,ggl}$ is $(h-1)(a-1)$ if $S_{i,j}$ is terminal and $(2h-2)(a-2)$ if not.

Proof: We count the members of R_{ggl} with shorter HM paths and subtract these from the size given in Lemma 1.

All groups in R'_{ggl} are connected to group i at $S_{i,0}$ or $S_{i,a-1}$. First suppose $S_{i,j}$ is terminal and consider switches of R_{ggl} reachable using a two-hop HM path. None are reachable if the path begins with a local hop since it must go to the other terminal switch ($S_{i,\bar{j}}$) and its global edges all go to j^{th} switches, which are in R_{gg} and thus not in R_{ggl}. Starting with a global hop, the path can reach $h-1$ of the groups in R'_{ggl}. Any of the $a-1$ switches of that group in R_{ggl} is then reachable by ≤ 1 local hop. This gives $(a-1)(h-1)$ with a shorter HM path and proves the claim for terminal switches.

If $S_{i,j}$ is non-terminal, the HM paths to members of R_{ggl} begin with a local hop to a terminal switch. Each of these has a global hop to $h-1$ members of R_{ggl}, for $(2h-2)$ shorter paths in total. ■

The remaining three-hop Valiant path is glg:

Lemma 2: In the relative arrangement, $(2h-1)(a-1)$ switches are reachable in a glg path from $S_{i,j}$:

$$R_{glg} = \{S_{x,k} : i + (j + \bar{k})h + 2 \leq x \leq i + (j + \bar{k} + 2)h, \; k \neq j\}$$

Proof: The first global hop goes to a switch in R_g. The local hop then goes to a different switch in the same group, i.e. $R_{gl} = \{S_{x,n} : x \in R'_g, n \neq \bar{j}\}$. From there, we can go to any switch $k \neq j$ in a range of groups, those reachable from switch $\bar{k}$ in the groups of $R'_{lg} = R'_g$. The range given in the expression for R_{glg} above comes from composing the expression for the range in R'_g; the lower endpoint is the lower endpoint for R'_g, but replacing i with $i + jh + 1$ since that is the smallest group from which we take the last global hop.

To calculate $|R_{glg}|$, we note that the $(a-1)$ switches other than j are reached in a range of $(2h-1)$ groups. ■

To count 3-hop Valiant paths that are shortest, we first need some technical lemmas.

Lemma 3: In the relative arrangement, $|R_{glg}(S_{i,j}) \cap R_{ggl}(S_{i,j})|$ is $2h-3$ if $S_{i,j}$ is terminal and $2h-2$ if not.

Proof: Note that for R_{glg} and R_{gg} we compose Equation 1 on the same range of groups; from a given switch n in R'_g, we reach switch $\bar{n}$ in groups $i+(j+n)h+2$ through $i + (j + n + 2)h$. We call this X_n:

$$X_n(S_{i,j}) = \{S_{x,\bar{n}} : i+(j+n)h+2 \leq x \leq i+(j+n+2)h\}$$

Then $R_{gg} = X_{\bar{j}}$ and R_{glg} is the union of X_n for $n \neq \bar{j}$.

Observe that each switch number is reachable in $2h-1$ consecutive groups and that the ranges of groups reaching switches $n-1$ and n are offset by h.

To calculate the intersection, we examine the groups in R'_{gg}, following Lemma 1. By the discussion above, this is the range from which switch j is reachable and the ranges for neighboring switches are each offset by h. If j is non-terminal, it has both switches $j-1$ and $j+1$ as neighbors. Switch j does not contribute since it is not in R_{ggl}, but the other two ranges each overlap in $h-1$ groups, giving the claimed size of $2h-2$.

When j is terminal, the argument is similar except the overlap is slightly less on one side. For example, if $j = 0$

then switch $a-1$ is reachable from groups $i+2$ through $i+2h$, causing an overlap of only $h-2$ on the side, for $2h-3$ overall. ■

Lemma 4: In the relative arrangement, $|(R_g \cup R_{gl}) \cap R_{glg}|$ is $h-1$ if $S_{i,j}$ is terminal and $2h-2$ if not.

Proof: Together, these reach all switches in groups of R'_g (i.e. groups $i+jh+1$ through $i+jh+h$) and no others. Consider the intersection of these with X_n, the destinations of gg and glg paths. Switch $a-1$ is in X_n for groups $i+jh+2$ through $i+jh+2h$, an overlap of $h-1$ groups. Switch 0 is reachable from groups $i+(j+a-1)h+2 \equiv i+jh-h+1$ through $i+(j+a-1+2)h \equiv i+jh+h-1$, also an overlap of $h-1$ groups. No other switches in X_n are in the groups of R'_g since the range for switch 1 starts h groups after switch 0 starts, and the range for switch $a-2$ ends h groups before switch $a-1$ ends.

When j is non-terminal, glg paths reach switches 0 and $a-1$, so the intersection of R_{glg} with R_g and R_{gl} has $2h-2$ switches. When j is terminal, only the switch $\neq j$ of these is glg reachable so the intersection size is $h-1$.

Additionally, R'_g is a subset of R'_{gg} when j is terminal, except for one group (either $i+h$ or $i-h$) in which $\bar{j}$ is glg reachable, so all but one of the R_{glg} switches in R'_g is ggl reachable. ■

Lemma 5: In the relative arrangement, $|R_{glg} \cap R_{lg}|$ is $(a-2)(h-1)$ if $S_{i,j}$ is terminal and 0 if not.

Proof: When $S_{i,j}$ is terminal, the groups in R'_g are 1 through h away from i. In the relative arrangement, the range of groups that $S_{i,j}$ goes to is offset by one from the range of groups that switch j in adjacent group numbers goes to. As such, a global hop from a particular switch number in the R'_g groups reaches $h-1$ of the same destinations as a global hop from that switch number in group i. For switch $\bar{j}$, a global hop from that switch number in the R'_g constitutes a gg path and not a glg path. For switch j, a global hop from group i is a g and not an lg path. So, the overlap between R_{lg} and R_{glg} is $(a-2)(h-1)$.

When $S_{i,j}$ is non-terminal, the groups in R'_g are more than h away from $S_{i,j}$, so there isn't such an overlap. ■

Putting these together, we can compute $\delta_{=,glg}$:

Theorem 3: In the relative arrangement, $\delta_{=,glg}$ is $ah-h$ if $S_{i,j}$ is terminal and $2ah-a-4h+3$ if not.

Proof: For both parts, we start with Lemma 2 and subtract the switches reachable via shorter paths, i.e. g, gl, and lg paths, using Lemmas 4 and 5. Note that lg paths have no destinations in R'_g, which contains all members of R_g and R_{gl}, so no further terms are needed. ■

Finally, we can combine Theorem 3 with Lemmas 1 and 3 to get the total for all 3-hop Valiant paths:

Theorem 4: In the relative arrangement, $\delta_{=,glg,ggl}$ is $2ah-a-3h+2$ if $S_{i,j}$ is terminal and $4ah-3a-10h+9$ if not.

IV. Results for absolute arrangement

Groups have distinct roles in the *absolute arrangement*[2] [7]. In this arrangement, port k of group i goes to the k^{th} group other than i; the only dependence on i is that destination group numbers skip it. Formally, port k of group i links to group k if $k < i$ and group $k+1$ otherwise. Figure 3 shows a sample of this arrangement. The destination groups are the same as the relative arrangement for group 0, but not the others. The dependence on source group makes the arrangement less symmetric and complicates our analysis.

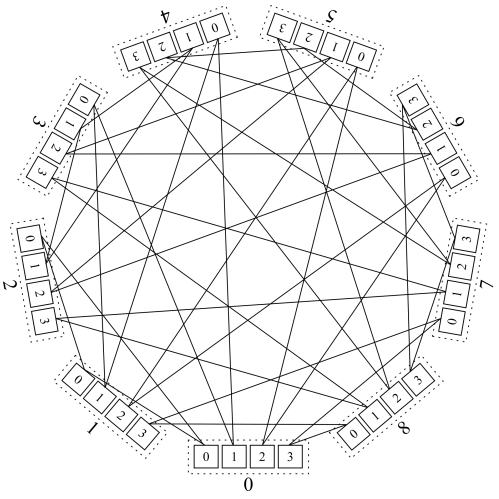

Fig. 3: $(p,4,2)$-Dragonfly with absolute arrangement

As with relative arrangements, we begin by looking at the switches reachable via a single global hop.

Lemma 6: In the absolute arrangement, the switches $R_g = R_g^+ \cup R_g^-$ are reachable from $S_{i,j}$, where

$$R_g^+ = \{S_{x,n} : x > i, jh < x \leq jh+h, n = \lfloor i/h \rfloor\}$$
$$R_g^- = \{S_{x,n} : x < i, jh \leq x < jh+h, n = \lfloor (i-1)/h \rfloor\}$$

Proof: In general, every group uses the same switch to communicate with a particular other group, and a switch is used for a contiguous set of group numbers. From a source switch $S_{i,j}$ the first global hop reaches h contiguous groups from jh to $jh+h$. The lower bound is inclusive when it is below the source group number, and the upper bound is inclusive when it is above the source group number, otherwise they are exclusive. This is because group destinations from a port are shifted by one for groups above the source.

The switch reached in one hop is generally $\lfloor \frac{i}{h} \rfloor$. However, when $i \bmod h \equiv 0$ and the destination group is $< i$, the destination switch is $\lfloor \frac{i}{h} \rfloor - 1 = \lfloor \frac{i-1}{h} \rfloor$; the shift in links on the destination group moves its source group link to a different switch. ■

A. 2-hop Valiant paths

The switch reached in one hop is the one that communicates with group i, which is also the switch used to communicate with $h-1$ of the adjacent-numbered groups to i, so the 2-hop destinations are in these groups. The 2-hop destination switch is switch j in the destination groups, since it connects directly to the same groups as $S_{i,j}$.

When $S_{i,j}$ connects with group numbers adjacent to i ($j = \lfloor i/h \rfloor$), there are h destinations. The R'_g groups jump port-numbering on the R_g switch, so the groups reachable from

[2]Also called the *consecutive arrangement*; we use the name from [10].

them will collectively span $h+1$ different groups including i, instead of h different groups including i.
When $j = \lfloor i/h \rfloor$,

$$R_{gg}(S_{i.j}) = \{S_{x,j} : \lfloor i/h \rfloor h \leq x \leq \lfloor i/h \rfloor h + h\}$$

When $j < \lfloor i/h \rfloor$,

$$R_{gg}(S_{i.j}) = \{S_{x,j} : \lfloor i/h \rfloor h < x \leq \lfloor i/h \rfloor h + h\}$$

When $j > \lfloor i/h \rfloor$,

$$R_{gg}(S_{i.j}) = \{S_{x,j} : \lfloor i/h \rfloor h \leq x < \lfloor i/h \rfloor h + h\}$$

Also, when groups numbered $< i$ use a different switch to connect to i than groups above it (i.e. when $i \equiv 0 \bmod h$), R'_{gg} varies among switches in i. When $j \geq \lfloor \frac{i}{h} \rfloor$, i is the lowest-numbered reachable group on switch $\lfloor i/h \rfloor$ from the R_g groups, and the two-hop destinations are as above. When j is smaller, group i is the highest-numbered group on the previous switch from R'_g , since R'_g jumped numbering at less than i. Switch $\lfloor \frac{i}{h} \rfloor - 1$ also links with adjacent groups, so it has h destinations as well as switch $\lfloor \frac{i}{h} \rfloor$.
When $j < \lfloor (i-1)/h \rfloor$ and $i \equiv 0 \bmod h$,

$$R_{gg} = \{S_{x,j} : \lfloor (i-1)/h \rfloor h < x \leq \lfloor (i-1)/h \rfloor h + h\}$$

When $j = \lfloor (i-1)/h \rfloor$ and $i \equiv 0 \bmod h$,

$$R_{gg} = \{S_{x,j} : \lfloor (i-1)/h \rfloor h \leq x \leq \lfloor (i-1)/h \rfloor h + h\}$$

When j is $\lfloor i/h \rfloor$ or $\lfloor (i-1)/h \rfloor$, the HM path to all destinations is one hop, since this is the switch number used to communicate with group i and the members of R_{gg}. Otherwise, the source/destination switch number does not connect to the source/destination groups, so the HM path has two local hops and total length three.

In group i, the switches numbered $\lfloor \frac{i}{h} \rfloor$ and $\lfloor \frac{i-1}{h} \rfloor$ have h unique 2-hop Valiant destinations, each with a strictly shorter HM path. There are 2 such switches for $i \equiv 0 \bmod h$ other than $i = 0$, and 1 such switch for other i. The other $a-1$ or $a-2$ switches have $h-1$ unique 2-hop Valiant destinations, all of which have a strictly longer HM path.

B. 3-hop Valiant paths

Because the absolute arrangement lacks symmetry, the number of destinations of 3-hop Valiant paths varies considerably by switch without a clear pattern. To illustrate this and to get a sense of the frequency of SV paths, we plot the values of $\delta_<$ and $\delta_\leq$ for a particular system in Figure 4. Recall that these are the number of destinations with strictly shorter and no longer SV paths, respectively. Also plotted using the lightest bar are the total number of SV path destinations; the portion of this bar not covered by the others is the number of destinations with a strictly shorter HM path. We looked at other system sizes as well; all had low values of $\delta_<$ and irregular but significant values for $\delta_\leq$.

Although we could not determine the number of SV paths, we give a bound:

Theorem 5: In the absolute arrangement, $\delta_{=,glg} \geq h(a-1)$ when $a > 2$.

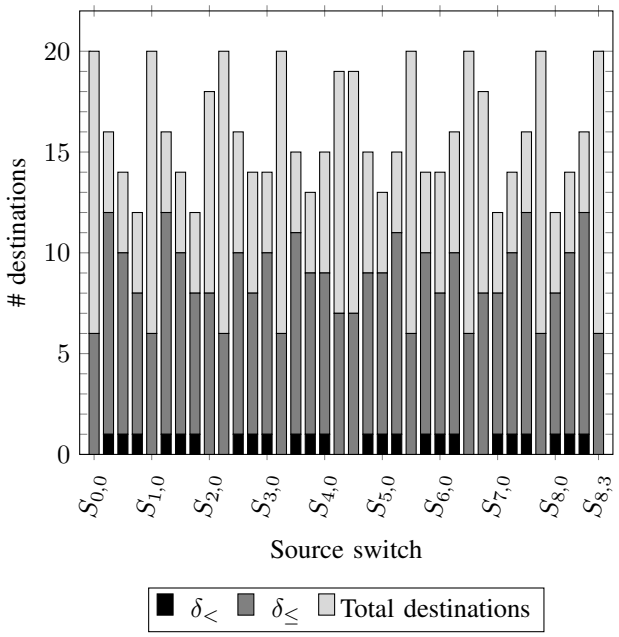

Fig. 4: Destinations of gg, lgg, glg, and ggl paths from each switch on a $(p, 4, 2)$-Dragonfly with absolute arrangement. The total number of possible destination switches is 35

Proof: From a given switch, each lg path goes to a distinct group since all their global links leave the source group. There are $h(a-1)$ lg paths. Thus, $|R_{glg}| \geq h(a-1)$; just start with an arbitrary global link.

When j is $\lfloor i/h \rfloor$ or $\lfloor (i-1)/h \rfloor$, the gg, g, and gl destinations are all in the same h groups ($h+1$ including group i). None of the glg destinations are in this range since the R'_g groups are linked through the g reachable switch, and the local hop leaves that switch. Therefore, at least $h(a-1)$ glg destinations are in the $h(a-1)$ other groups. Since each glg path through a particular intermediate group goes to a different group, each of these $h(a-1)$ groups has a glg destination through each of the h groups in R'_g. No switch in R_{glg} can link to both $S_{i,j}$ and all h groups in R'_g since it has only h global links. In each R'_{glg} group there must be a glg destination switch that is not adjacent to group i, and thus there cannot be an lg path to it.

For other j, the gg destinations are in different groups than the g and gl destinations. None of the $h(a-1)$ lg destinations from a given switch in $R_g(S_{i,j})$ are in the h groups of $R'_{gg}(S_{i,j})$ (including group i), or the starting group, so they must be across the other $h(a-1)$ groups. At most $h-1$ of the destinations can be in the other $h-1$ groups of R'_g, so there must be glg destinations in each of the remaining $h(a-2)+1$ groups, without g, gl, or gg paths. Those destinations also do not have lg paths; since $S_{i,j}$ does not communicate with neighboring group numbers, the groups of R'_g and group i are adjacent to different switches in the destination groups.

There are $(a-1)$ ggl-reachable switches in each of the $h-1$ gg-reachable groups. Since gg- and g-reachable groups differ, the only possible 2-hop HM path is lg, had by ≤ 1 switch per group. For $a > 2$, this means ≥ 1 switch per

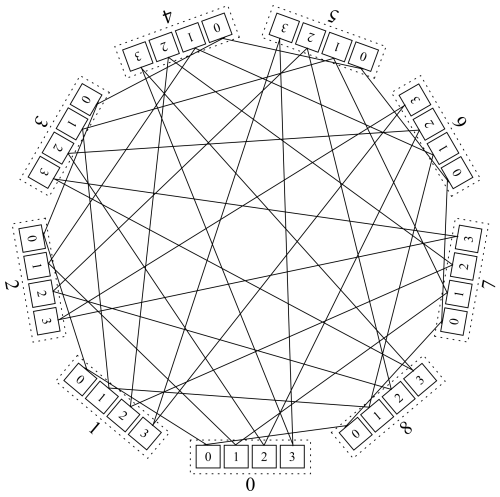

Fig. 5: $(p, 4, 2)$-Dragonfly with circulant-based arrangement

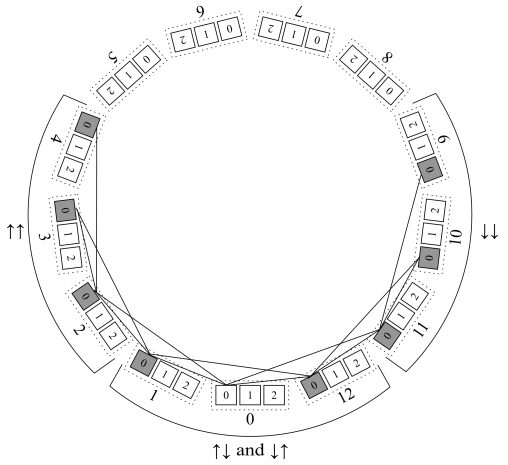

Fig. 6: gg destinations from $S_{0,0}$ on a $(p, 3, 4)$-Dragonfly

group has no shorter path.

Combining these, $\geq h(a-2)+1+h-1 = h(a-1)$ switches have no shorter path. ∎

V. Results for circulant-based arrangement

The *circulant-based arrangement* [7] assumes h is even and splits the edges leaving each switch into two categories. These categories, denoted $\uparrow$ and $\downarrow$, connect to the "next" higher-numbered and lower-numbered groups respectively. Thus, each switch has two reachable sets from $S_{i,j}$ depending on the category of edge used:

$$R_\uparrow = \{S_{x,j} : i + j(h/2) < x \leq i + j(h/2) + h/2\} \quad (3)$$
$$R_\downarrow = \{S_{x,j} : i - j(h/2) - h/2 \leq x < i - j(h/2)\} \quad (4)$$

Figure 5 shows a sample of this arrangement.

In the circulant-based arrangement each group's switches have a symmetric role like the relative arrangement, but the multiple edge categories complicate the analysis of multi-hop paths. Even analysis of the gg pattern splits into cases based on the edge types used: $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, or $\downarrow\downarrow$. Composing Equations 3 and 4 gives the destination sets for each pattern:

$$R_{\uparrow\downarrow} = R_{\downarrow\uparrow} = \{S_{x,j} : i - h/2 + 1 \leq x \leq i + h/2 - 1\}$$
$$R_{\uparrow\uparrow} = \{S_{x,j} : i + jh + 2 \leq x \leq i + jh + h\}$$
$$R_{\downarrow\downarrow} = \{S_{x,j} : i - jh - h \leq x \leq i - jh - 2\}$$

$R_{\uparrow\downarrow}$ and $R_{\downarrow\uparrow}$ are always the same. The three distinct sets are illustrated in Figure 6. If $R_{\uparrow\downarrow}$, $R_{\uparrow\uparrow}$, and $R_{\downarrow\downarrow}$ are disjoint, their combined size is $(h-2)+2(h-1)$ (excluding the source switch $S_{i,j}$, which falls within $R_{\uparrow\downarrow}$). The sets can intersect, however, depending on the value of j. Changes to j have no effect on $R_{\uparrow\downarrow}$, but increasing it shifts the range in $R'_{\uparrow\uparrow}$ upward (clockwise in Figure 6) and the range in $R'_{\downarrow\downarrow}$ downward (counterclockwise in Figure 6).

Working out all the details gives the following: (proof omitted for space)

Theorem 6: In the circulant-based arrangement, $\delta_{<,gg} = |R_{gg}| - s_{a,h,j}$, where

$$|R_{gg}| = \begin{cases} 2hj + 2h & \text{if } j \leq \frac{h-6}{2h} \\ 2hj - ah + 3h - 2 & \text{if } \frac{ah-h-1}{2h} < j \leq \frac{ah-3}{2h} \\ 2ah - 2hj - 2 & \text{if } \frac{2ah-3h+4}{2h} \leq j \\ 3h - 4 & \text{otherwise} \end{cases}$$

and

$$s_{a,h,j} = \begin{cases} h & \text{if } j = 0 \\ \frac{3jh}{2} - ah + \frac{3h}{2} & \text{if } \frac{2ah-3h+2}{3h} \leq j \leq \frac{2(a-1)}{3} \\ ah - \frac{3jh}{2} - \frac{h}{2} & \text{if } \frac{2(a-1)}{3} < j \leq \frac{2ah-h-2}{3h} \\ 0 & \text{otherwise} \end{cases}$$

The first case in the value of $|R_{gg}|$ occurs when $R_{\uparrow\uparrow}$ intersects the left end of $R_{\uparrow\downarrow}$ and $R_{\downarrow\downarrow}$ intersects the right end $R_{\uparrow\downarrow}$. The second case occurs when $R_{\uparrow\uparrow}$ and $R_{\downarrow\downarrow}$ intersect each other. The third occurs after they have crossed and each intersects the opposite end of $R_{\uparrow\downarrow}$ as in the first case. The "otherwise" case occurs when no sets intersect, either before or after $R_{\uparrow\uparrow}$ and $R_{\downarrow\downarrow}$ cross.

The function $s_{a,h,j}$ gives the number of switches in R_{gg} also in $R_\uparrow$ or $R_\downarrow$. (A local hop cannot help reach an element of R_{gg} in ≤ 2 hops since all its members are switch j.)

Because the "otherwise" case occurs when the sets are disjoint, $|R_{gg}|$ is largest in this case. Notice that this case is common on larger systems, i.e. those with large h. The first case can only happen when $j = 0$. The second case occurs when j is between $\approx a/2 - 1/2$ and $\approx a/2$. The third case occurs when j is at least $\approx a - 3/2$.

To consider 3-hop Valiant paths, we use Theorem 6 in a way similar to how we built on knowledge of R_{gg} for the relative arrangement. As then, R_{gg} consists entirely of switch j in the reachable groups. Thus, ggl paths lead to the other $(a-1)$ switches in each of these groups. Unlike in the relative arrangement, $R_{ggl} \neq R_{lgg}$ since R'_{gg} depends on the value of j in the circulant-based arrangement. We can, however, calculate $|R_{lgg}|$ by summing up the values of $|R_{gg}|$ from Theorem 6 over the other switches in the source group; the R_{gg} sets are disjoint for different j. By the discussion

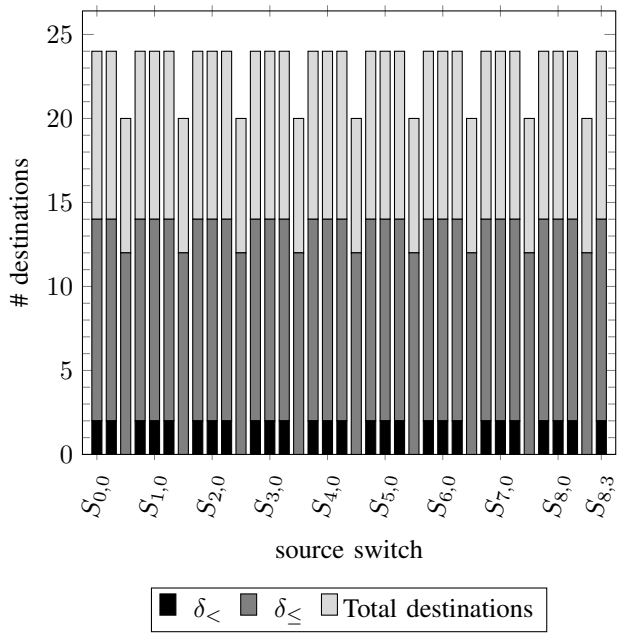

Fig. 7: Destinations of gg, lgg, glg, and ggl paths from each switch on a $(p, 4, 2)$-Dragonfly with circulant arrangement. The total number of possible destination switches is 35

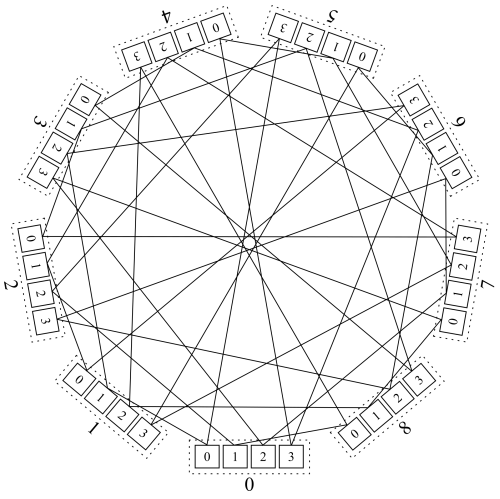

Fig. 8: $(p, 4, 2)$-Dragonfly with the helix arrangement

above on the frequency of each case, the sum is likely around $(a-1)(3h-4)$ for large h.

These estimates again lead to the prediction that $\delta_\leq$ is of order ah. Since we lack a closed-from solution, we again plot the values of $\delta_<$ and $\delta_\leq$ by switch; see Figure 7. There is much less variation in the values than in the absolute arrangement. $\delta_\leq$ is generally 35–40% of the $O(a^2h)$ possible destinations for small systems (a = 2–10, h = 2 or 4), dropping to around 10% at $a = 100$, $h = 50$.

VI. Results for helix arrangement

The *helix arrangement* [9] has the following reachable sets:

$$\begin{aligned} R_\uparrow &= \{S_{x,j+1} : i + j\lfloor h/2\rfloor < x \leq i + j\lfloor h/2\rfloor + \lfloor h/2\rfloor\} \\ R_\downarrow &= \{S_{x,j-1} : i - j\lfloor h/2\rfloor - \lfloor h/2\rfloor \leq x < i - j\lfloor h/2\rfloor\} \\ R_m &= \{S_{i+a\lfloor h/2\rfloor+j+1,a-1-j}\} \end{aligned}$$

Helix is very much like the circulant-based arrangement, but with two differences. First, rather than assume h is even, it adds a *mutual link* (*m*) connecting "middle switches" when h is odd. Second, the switch numbers at each end of a non-mutual link do not match; the switch number at the other end of an $\uparrow$ link is higher by one and that of a $\downarrow$ link is lower by one. Figure 8 shows a sample of this arrangement.

The analysis of gg paths in the helix arrangement has a similar flavor to that of the circulant-based arrangement except that both differences between the arrangements complicate the analysis. First, the mutual links add edge patterns $\uparrow m$, $\downarrow m$, $m \uparrow$, and $m \downarrow$. (mm paths form cycles and can be ignored.) Second, the switch number shifts break the symmetry between $\uparrow\downarrow$ and $\downarrow\uparrow$ patterns, which now yield distinct destination sets; traversing the first edge changes the switch number and thus the edges available for the second hop. Together, these complications prevented us from finding an analog of Theorem 6.

We plotted the values of $\delta_<$ and $\delta_\leq$ per switch, but omit the results for space. The overall look was similar to circulant-based (see Figure 7), with many ties and periodic dips in $\delta_\leq$, but the dips were slightly deeper and $\delta_<$ was more consistently positive (but still small).

VII. Results for nautilus arrangement

The *nautilus arrangement* [9] is defined with an incremental construction; switches are ordered by their group number and by their switch number within each group. Switches are visited in this order and, at each step, links are made to later switches so that the visited switch has h links. All links made from switch $S_{i,j}$ go to switch $i \bmod a$. The links are made to the next groups that are not already connected to group i in either the clockwise direction if j is even or the counterclockwise direction if j is odd. Figure 9 shows a sample of this arrangement.

The nautilus arrangement is the least regular of all the arrangements because the other end of a link from $S_{i,j}$ depends on whether the link was made when $S_{i,j}$ was visited or if it was made earlier. Even that the procedure generates a global link arrangement required proof [9]. Because of this irregularity, we do not determine the number of SV paths analytically, but computational testing was promising. Many gg paths were no longer than HM paths and most of their destinations from a given starting point were distinct; intuitively, this happens when $a = \Omega(h)$ since the first hop goes to distinct group numbers mod a, making the second hops land on distinct switches.

VIII. Related work

The work most directly related to ours is by Camarero et al. [7], who observe that gg paths can be shorter than HM paths, but do not count such paths or consider the use of non-HM paths for path diversity.

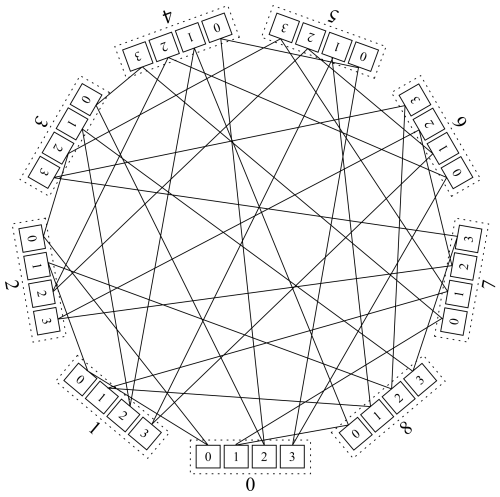

Fig. 9: $(p, 4, 2)$-Dragonfly with nautilus arrangement

As mentioned previously, the issue of Dragonfly global link arrangements has not received much attention. Camarero et al. [7] identified the relative, absolute, and circulant-based arrangements, but did not consider the difference important. Hastings et al. [10] compared them using bisection bandwidth and showed differences of up to 50%. Kaplan et al. [11] evaluated them with simulations, finding small performance differences in many cases, but some differences as high as 44%. Later, Belka et al. [9] introduced the helix and nautilus arrangements to improve bisection bandwidth in the case of high global bandwidth.

There has been quite a lot of research on routing algorithms for Dragonfly systems. It was found that the original Valiant algorithm could create hotspots on local edges [4], which led to algorithms using non-minimal paths even within a group [4], [5]. Valiant routing can also send messages in loops in other topologies [12], [13]. Despite these issues, Valiant continues to be used as a building block for adaptive routing (e.g. [6]).

Pascual and Navaridas [14] give algorithms for dynamic assignment of VCs, making it easier to change routing algorithms and adapt to hardware faults. They test on a routing algorithm that uses the shortest path (possibly an SV path) and one that spreads traffic across all shortest paths. Solomonik et al. [8] showed that the latter could significantly improve performance on a (3D torus) Blue Gene/P in concert with task mapping designed to exploit it.

IX. Discussion

We explored the frequency of HM paths being non-minimal. For most global link arrangements, the exceptions are fairly rare, but they do occur. Our class of SV paths also gives many paths with the same length; these paths made up at least $\approx 1/a$ of all paths in the relative and absolute arrangements and likely in the circulant-based arrangement as well. We also showed empirically that this ratio is around 35–40% for many switches on specific systems.

Now that SV paths have been shown to be relatively abundant, the obvious next question is how to use them. One idea is to recognize that SV paths are potentially-good Valiant paths so they could be preferentially included in the set of paths considered by adaptive schemes. This could be done without additional hardware. Other ideas are to replace HM routing with true shortest path routing, likely as part of an adaptive routing algorithm with a Valiant-like option for congestion, or to route traffic along all shortest paths, following the ideas of [8]. It would be interesting to evaluate these ideas experimentally.

Acknowledgments

We thank the reviewers for their thorough job and the many useful comments that have improved this paper. This work was partially supported by the National Science Foundation under grant CNS-1423413 and the Paul K. & Evalyn Elizabeth Cook Richter Memorial Fund.

References

[1] J. Kim, W. Dally, S. Scott, and D. Abts. Technology-driven, highly-scalable dragonfly topology. In *Proc. 35th Ann. Intern. Symp. Comput. Arch. (ISCA)*, pages 77–78, 2008.

[2] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R.S. Williams, and K. Yelick. Exascale computing study: Technology challenges in achieving exascale systems. Technical report, 2008.

[3] L.G. Valiant. A scheme for fast parallel communication. *SIAM J. Computing*, 11(2):350–361, 1982.

[4] M. García, E. Vallejo, R. Beivide, M. Odriozola, C. Camarero, M. Valero, G. Rodriguez, J. Labarta, and C. Minkenberg. On-the-fly adaptive routing in high-radix hierarchical networks. In *Proc. 41st Intern. Conf. Parallel Processing (ICPP)*, pages 279–288, 2012.

[5] M. García, E. Vallejo, R. Beivide, M. Odriozola, and M. Valero. Efficient routing mechanisms for dragonfly networks. In *Proc. 42nd Intern. Conf. Parallel Processing (ICPP)*, pages 582–592, 2013.

[6] P. Fuentes, E. Vallejo, M. García, R. Beivide, G. Rodríguez, C. Minkenberg, and M. Valero. Contention-based nonminimal adaptive routing in high-radix networks. In *Proc. 29th IEEE Intern. Parallel and Distributed Processing Symp. (IPDPS)*, 2015.

[7] C. Camarero, E. Vallejo, and R. Beivide. Topological characterization of Hamming and Dragonfly networks and its implications on routing. *ACM Trans. Architec. Code Optim.*, 11(4):39, 2014.

[8] E. Solomonik, A. Bhatele, and J. Demmel. Improving communication performance in dense linear algebra via topology aware collectives. In *Proc. Conf. High Performance Computing, Networking, Storagem, and Analysis (SC)*, page 77, 2011.

[9] M. Belka, M. Doubet, S. Meyers, R. Momoh, D. Rincon-Cruz, and D.P. Bunde. New link arrangements for dragonfly networks. In *Proc. 3rd Intern. Workshop High-Performance Interconnection Networks Towards the Exascale and Big-Data Era*, 2017.

[10] E. Hastings, D. Rincon-Cruz, M. Spehlmann, S. Meyers, A. Xu, D.P. Bunde, and V.J. Leung. Comparing global link arrangements for Dragonfly networks. In *Proc. IEEE Cluster*, pages 361–370, 2015.

[11] F. Kaplan, O. Tuncer, V.J. Leung, S.K. Hemmert, and A.K. Coskun. Unveiling the interplay between global link arrangements and network management algorithms on dragonfly networks. In *Proc. 17th IEEE/ACM Intern. Symp. Cluster Computing and the Grid*, 2017.

[12] D. Han, Z. Wang, and D.P. Bunde. Improving Valiant routing for Slim Fly networks. In *Proc. 10th Intern. Workshop Parallel Programming Models and Systems Software for High-End Computing (P2S2)*, 2017.

[13] P. Yébenes, J. Escudero-Sahuquillo, P.J. García, F.J. Quiles, and T. Hoefler. Improving non-minimal and adaptive routing algorithms in Slim Fly networks. In *Proc. 25th Ann. Symp. High-Performance Interconnects (HOTI)*, 2017.

[14] J.A. Pascual and J. Navaridas. High performance, low complexity deadlock avoidance for arbitrary topologies/routings. In *Proc. 32nd ACM Intern. Conf. Supercomputing (ICS)*, 2018.