Multi-Path Routing in the Jellyfish Network

Zaid ALzaid Saptarshi Bhowmik Xin Yuan
Department of Computer Science
Florida State University
Tallahassee, Florida, USA
{alzaid, bhowmik, xyuan}@cs.fsu.edu

***Abstract*—The Jellyfish network has recently been proposed as an alternative to the fat-tree network for data centers and high-performance computing clusters. Jellyfish uses a random regular graph as its switch-level topology and has shown to be more cost-effective than fat-trees. Effective routing on Jellyfish is challenging. It is known that shortest path routing and equal-cost multi-path routing (ECMP) do not work well on Jellyfish. Existing schemes use variations of k-shortest path routing (KSP). In this work, we study two routing components for Jellyfish: path selection that decides the paths to route traffic, and routing mechanisms that decide which path to be used for each packet. We show that the performance of the existing KSP can be significantly improved by incorporating two heuristics, randomization and edge-disjointness. We evaluate a range of routing mechanisms, including traffic oblivious and traffic adaptive schemes, and identify an adaptive routing scheme with noticeably higher performance than others.**

***Keywords*-Interconnection network, Jellyfish, KSP routing, rKSP routing, EDKSP routing, rEDKSP routing**

I. Introduction

The Jellyfish interconnect has recently been proposed for data centers and high performance computing (HPC) clusters [1]–[3]. It adopts the random regular graph (RRG) as its switch-level topology and has good topological properties such as high bisection bandwidth and low average path length. Jellyfish has shown to be more cost-efficient than the widely used fat-trees [1].

Due to the random connectivity in Jellyfish, the traditional shortest path routing and equal-cost multi-path routing (ECMP) do not perform well [1]. Singla et al. proposed k-shortest path routing (KSP) to select paths and explore path diversity in the network. A variation of KSP, called LLSKR, was later developed [2]. Both KSP and LLSKR suffer from two issues that can degrade their performance on Jellyfish. First, both routing schemes try to use "shortest" paths without considering the potential link usage. When a link is in a short path, it tends to be in other short paths as well. Hence, routing traffic with such paths can introduce load imbalance in the network and lower the performance. Second, vanilla KSP algorithms such as the ones that use node id to break the tie between two paths of the same length have biases in selecting paths. This can cause severe load imbalance problems in Jellyfish because Jellyfish tends to have many paths of the same length between a pair of nodes.

Beside *path selection* that decides the set of paths to be used to route traffic, another vital component in routing is determining a particular path for each packet. We will call this component the *routing mechanism*. While some conventional routing mechanisms can directly apply to multi-path routing on Jellyfish, this component has not been thoroughly investigated for Jellyfish; and it is unclear which routing mechanism is effective on Jellyfish.

In this work, we study both path selection and routing mechanisms for multi-path routing in Jellyfish. For path selection, we find that two heuristics, randomization, and edge-disjointness, can significantly improve the quality of the paths selected when they are incorporated in KSP. We investigate a range of routing mechanisms, including traffic oblivious and traffic adaptive schemes. We perform an extensive evaluation to compare the path selection methods and the routing mechanisms using performance modeling, flit-level simulation, and application simulation. The conclusions of this study include the following:

- For path selection, we find that the paths computed using the KSP with randomization and edge-disjointness heuristics achieve higher performance than the vanilla KSP across all routing mechanisms and all traffic patterns in the study. The improvement is significant, especially when the best performing adaptive routing scheme is used.
- For routing mechanism, we find a new adaptive routing scheme, which we call *KSP-adaptive*, is most effective for multi-path routing on Jellyfish. *KSP-adaptive* selects a path for each packet by randomly obtaining two candidate paths from the k paths for the source-destination pair and choosing the one with a smaller estimated latency to route the packet. Our evaluation indicates that *KSP-adaptive* achieves higher performance than various forms of universal globally adaptive load balance routing (UGAL) [4] on Jellyfish.

The rest of the paper is structured as follows. Section II presents the background. Section III describes our new proposed routing schemes. Section IV reports the performance study. The related work is discussed in Section V. Finally, Section VI concludes the paper.

II. JELLYFISH

Singla et al. [1] proposed the Jellyfish topology as a flexible and cost-efficient topology for large scale interconnection networks. The switch-level topology of Jellyfish is a random regular graph (RRG), where all switches have the same degree but are randomly connected. We consider the case when all switches are connected with the same number of processing nodes and have the same number of ports.

Jellyfish can be specified with three parameters [1]: the number of switches (N), the number of ports in each switch (x), and the number of ports in each switch that connect to other switches (y): each switch connects to $x-y$ processing nodes. We will use $RRG(N,x,y)$ to denote a Jellyfish topology with N switches, x switch ports, and y ports in each switch connecting to other switches. Figure 1 illustrates $RRG(N=15,x=4,y=3)$. Each switch connects to $4-3=1$ processing node. Note that each instance of an *RRG* is different from other instance *RRG*. However, when N and y are sufficiently large, different instances will have very similar network characteristics, and performance [2].

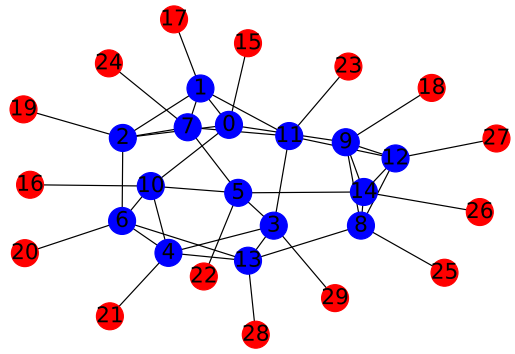


Figure 1: An example $RRG(N=15,x=4,y=3)$

Even though RRG provides high network capacity, it is known that single path routing and equal-cost multi-path routing performs poorly on Jellyfish [1]. As a remedy, Singla et al. [1] proposed the K-shortest path routing (KSP) scheme to explore the topology's path diversity. Due to the random nature of Jellyfish, the lengths of the K paths between different pairs of switches may vary. Yuan et al. [2] observe that using the plain KSP in Jellyfish has some limitations, including (1) not using all "short" paths when the number of short paths between two switches is large, and (2) using long paths when the number of short paths between two switches is small. Based on the observations, they proposed Limited Length Spread k-shortest Path Routing (LLSKR), which attempts to overcome the aforementioned limitations. LLSKR allows a variable number of paths between two switches, which allows more short paths to be used in comparison to KSP and can control the number of long paths to be used.

Both KSP and LLSKR rely on a k-shortest routing algorithm to compute the paths. Finding loopless k shortest paths in a given topology is a generalization of the shortest path problem. This problem has been studied extensively [5]–[7]. The baseline algorithm used in this work is the Yen's algorithm [6], which is shown in Figure 2. Since our heuristics are added over this algorithm, we describe the algorithm for completeness.

```
   Input: Graph, Source and Destination
   Output: K-shortest paths between Source and Destination
1  Function searching K-Shortest paths
2      A = [] a set to store the K- shortest path
3      B = [] a set to store the potential kth shortest path
4      A[0] = the shortest path from the Source to the
         Destination.
5      for i = 1 to i = k − 1 do
6          foreach node j in A[i − 1] without the Destination
             do
7              spurNode = j
8              rootPath = A[i − 1] from Source to spurNode
9              foreach path P in A do
10                 if rootPath = P[Source : spurNode] then
11                     remove edge
                         P[spurNode : spurNode + 1] from the
                         Graph.
12                 endif
13             end
14             foreach node ∈ rootPath except spurNode do
15                 remove node from the Graph.
16             end
17             spurPath = Dijkstra(Graph, spurNode,
                 Destination)
18             totalPath = rootPath + spurPath
19             if totalPath ∉ B then
20                 add totalPath to B
21             endif
22             restore the original Graph.
23         end
24         if B is not empty then
25             A[i] = Shortest path in B
26             B = []
27         else
28             end the loop.
29         endif
30     end
31 end
```

Figure 2: Yen's algorithm [6]

Yen's algorithm uses two containers A and B: Container A stores the k-shortest paths found, and container B stores potential shortest paths. The algorithm first finds the shortest path between the source and destination using Dijkstra's shortest path algorithm (or any other algorithm that finds the shortest path). The algorithm then loops for $k-1$ iterations, finding one shortest path in each iteration (Lines 5 to 30). The algorithm iterates over all nodes on the last shortest path in each iteration, except the destination node. For each node, the algorithm finds the shortest path to the destination

from that node (*spurPath* in Figure 2). To explorer a new shortest path *spurPath*, the Yen algorithm removes all nodes from the source up to the current node on the shortest path and the edge from the current node and the node after it on the shortest path. The algorithm then forms the *totalPath* by concatenating the *rootPath* (the shortest path from the source node up to the current node) and the *spurPath* as a candidate shortest path. After visiting all intermediate node along the shortest path, the algorithm selects the shortest path in B as the next shortest path and move it to A. The process is repeated until all k shortest paths are found.

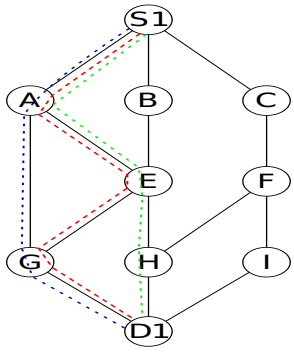
Besides path selection, the *routing mechanism*, which decides the path for each packet, is also an important component in multi-path routing on Jellyfish. This component has not been sufficiently investigated for the Jellyfish topology. Existing work has assumed some mechanisms that can take advantage of the multiple paths such as multi-path TCP (MPTCP) [8]. There are, however, other adaptive routing mechanisms, such as the Universal Globally Adaptive Load-balancing routing (UGAL) [4], that have not been thoroughly studied in this context.

III. Proposed Multi-path Routing Schemes for Jellyfish

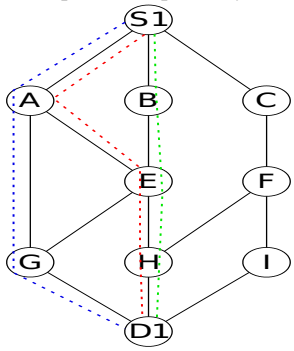
A. Path selection

Both KSP and LLSKR select "shortest" k-paths. Additionally, the vanilla version of KSP relies on Dijkstra's algorithm and is by default deterministic: when there is more than one path that can be selected, the tie is broken deterministically using some means such as preferring nodes whose ranks are higher or lower. This can easily result in low-quality paths for KSP in Jellyfish as illustrated in the example in Figure 3. In this example, we consider KSP that finds 3 shortest paths from $S1$ to $D1$. From $S1$ to $D1$, there is one 3-hop path $P0 = S1 \rightarrow A \rightarrow G \rightarrow D1$, and six 4-hop paths, $P1 = S1 \rightarrow A \rightarrow E \rightarrow G \rightarrow D1$, $P2 = S1 \rightarrow A \rightarrow E \rightarrow H \rightarrow D1$, $P3 = S1 \rightarrow B \rightarrow E \rightarrow G \rightarrow D1$, $P4 = S1 \rightarrow B \rightarrow E \rightarrow H \rightarrow D1$, $P5 = S1 \rightarrow C \rightarrow F \rightarrow H \rightarrow D1$, and $P6 = S1 \rightarrow C \rightarrow F \rightarrow I \rightarrow D1$.

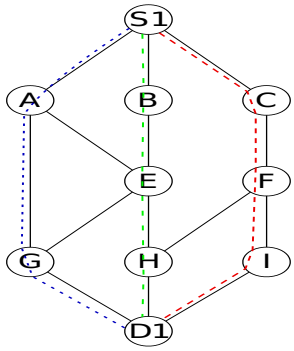
Let us assume that the textbook Dijkstra's algorithm is used in KSP, where nodes with smaller ranks are explored first to find the shortest path. We will call this version of KSP the vanilla KSP. The vanilla KSP will first find the three-hop path ($P0$). After that, it will find a 4-hop path with a bias toward nodes with smaller ranks (in Figure 3, such a path is the leftmost feasible path). As a result, the second path that will be found by KSP is $P1 = S1 \rightarrow A \rightarrow E \rightarrow G \rightarrow D1$; and the third path to be found is $P2 = S1 \rightarrow A \rightarrow E \rightarrow H \rightarrow D1$. These three paths are showed in Figure 3(a). As can be seen in the figure, all three paths share the link $S1 \rightarrow A$. Although there are three paths to carry traffic from $S1$ to $D1$, the effective bandwidth from $S1$ to $D1$ is equivalent to only having a single path. Note that there typically exist many paths of the same length from a source switch to a destination switch in an RRG [2]. The biases in the deterministic vanilla KSP algorithm can have serious problems for the Jellyfish topology.



(a) 3-shortest paths computed by vanilla KSP



(b) 3-shortest paths computed by KSP with randomization



(c) 3-shortest paths computed by KSP with edge-disjointness

Figure 3: Paths computed by KSP with different heuristics

This problem can be alleviated by incorporating randomness in KSP. The randomized KSP can be achieved by using a randomized Dijkstra's shortest path algorithm: when exploring the next node, the randomized Dijkstra algorithm does not use a deterministic mechanism, such as the node rank, to break the tie. Rather, it randomly selects the next node when there are multiple options. In the Yen's algorithm in Figure 2, this is done by replacing the Dijkstra() routing in Lines 17 and 4 by a randomized_Dijkstra() routine. Using

this heuristic, in the example in Figure 3, after the KSP finds the 3-hop path, it will randomly select two 4-hop paths out of the 6 candidate paths. For example, it may select: $P2 = S1 \rightarrow A \rightarrow E \rightarrow H \rightarrow D1$ and $P4 = S1 \rightarrow B \rightarrow E \rightarrow H \rightarrow D1$. This is shown in Figure 3(b). As can be seen from the figure, each link is at most shared by 2 paths, and the effective bandwidth from $S1$ to $D1$ is equivalent to having 2 independent paths, which is better than the paths computed by the vanilla KSP.

Sharing links among the selected paths reduces the total achievable bandwidth from the source to the destination. Randomization does not eliminate link-sharing among the selected paths. The edge-disjoint heuristic completely eliminates link sharing. In our implementation, the edge-disjoint heuristic follows the Remove-Find (RF) method [9]. The RF method consists of two steps: (1) finding the shortest path from the source to the destination, and (2) removing all edges associated with the shortest path. The algorithm repeats these two steps k times or until the source and destination node are disconnected. Using this heuristic, the 3 paths found for the example in Figure 3 are $P0$, $P4$, and $P6$, as shown in Figure 3(c). As can be seen from this figure, all of the three paths are now link disjointed. Using the three paths, the total bandwidth from $S1$ to $D1$ is equivalent to using three independent paths.

There are two potential problems with the edge-disjoint heuristic. First, there may not exist a sufficient number of edge-disjoint paths between two nodes. Second, the lengths of edge-disjoint paths can be much larger than the paths computed by the vanilla KSP, which increases the network resources to carry a packet and can decrease the performance when the network is under high load. However, as will be shown in our evaluation, in practical Jellyfish networks, y is much larger than k. In our experiments, with $k = 8$ and $k = 16$, edge-disjoint paths between all pairs of switches exist in all of the topologies. Moreover, the average path length computed by KSP with the edge-disjoint heuristic is similar to that computed by the vanilla KSP on practical Jellyfish networks.

In the rest of the paper, we will use KSP to denote the vanilla KSP, rKSP for randomized KSP, EDKSP for edge-disjoint KSP, and rEDKSP for KSP with both randomization and edge-disjoint heuristics. Note that rEDKSP will not improve the total throughput for each source-destination pair. However, randomizing path selection when there are more possibilities results in better load balancing for traffic patterns where multiple sources communicate with multiple destinations.

B. Routing mechanism

Given the multiple paths that can be used to carry a packet, the routing mechanism decides the path for the packet. We investigate a range of routing mechanisms for the multi-path routing in Jellyfish. In particular, we consider schemes to incorporate the Universal Globally Adaptive Load-balanced routing (UGAL) [4], which has demonstrated effectiveness on other topologies such as mesh and Dragonfly [4], [10]–[14].

UGAL distinguishes between two types of paths, minimal paths and non-minimal paths. The minimal path is the shortest path from the source to the destination in Jellyfish. A non-minimal path is formed by two minimal paths, one from the source to a (randomly selected) intermediate node and the other one from the intermediate node to the destination. To route a packet from a source to a destination, UGAL considers two paths, one minimal path and one random non-minimal path (with a random intermediate node), compares the estimated packet latency of the two paths sing queue length and chooses the path with the smaller latency for the packet. This scheme, which will be called *vanilla-UGAL*, can be directly applied to Jellyfish. Note that vanilla-UGAL does not need to use KSP to compute paths. The vanilla-UGAL uses more paths than the paths found by the KSP algorithm. One can restrict the non-minimal paths to only the ones computed by the KSP algorithm. In this case, the shortest path from the source to the destination will be the minimal path, and the rest of the k shortest paths are candidates for the random non-minimal paths. We will call this algorithm *KSP-UGAL*. Note that depending on the traffic and network condition, restricting non-minimal paths to the k shortest path can have advantages over the vanilla UGAL that may use longer non-minimal paths. In the situation when k paths can provide sufficient path diversity, using short paths reduce the link usage, which improves the performance. In the Jellyfish topology, most of k shortest paths are of similar length. Thus, when only the k paths are used, minimal and non-minimal paths are similar and may not need to be differentiated. Based on this observation, we propose a new UGAL-like adaptive routing scheme, which we call *KSP-adaptive*. KSP-adaptive randomly selects any two paths among the k shortest paths and chooses the one with a smaller estimated latency for the packet.

We compare the performance of these traffic adaptive routing schemes among one another and against other traffic oblivious schemes, including random routing (*random*) that randomly selects a path out of the k paths to route each packet, and *round-robin* that uses the k paths for each source-destination pair in a round-robin fashion to route packets.

IV. Evaluation

We perform comprehensive modeling and simulation studies to evaluate the proposed path selection and routing mechanisms on a number of Jellyfish topologies. In the following, we first describe our experimental methodology and then report the evaluation results.

A. Methodology

Topology: We report results on three Jellyfish topologies, a small topology $RRG(36, 24, 18)$ with 36 switches and 216

compute nodes, a medium-sized topology $RRG(720,24,19)$ with 720 switches and 3600 compute nodes, and a large-sized topology $RRG(2880,48,38)$ with 2880 switches and 28800 compute nodes. Table I summarizes important parameters of the topologies. As shown in [2], although random instances of RRG are different, when N and y are sufficiently large, any of the RRG topology with the same parameters will yield very similar performance and topological metrics. As such, we report the results for one random instance in our simulation studies.

Topology	Switch size	No. of switches	No. of compute nodes	Average shortest path len.
$RRG(36,24,16)$	24	36	288	1.54
$RRG(720,24,19)$	24	720	3600	2.57
$RRG(2880,48,38)$	48	2880	28800	2.59

Table I: Jellyfish topologies used in the experiments

Path selection and routing mechanisms: We compare four different path selection schemes for computing multiple paths for each pair of source and destination: KSP, rKSP (randomized KSP), EDKSP (edge-disjoint KSP), and rEDKSP (randomized edge-disjoint KSP). We will use the notations KSP(k), rKSP(k), EDKSP(k), and rEDKSP(k) to denote KSP, rKSP, EDKSP, and rEDKSP with k paths, respectively. The routing mechanisms considered are the following, which are described in Section III-B: single path routing (*SP*), *random*, *round-robin*, *vanilla-UGAL*, *KSP-UGAL*, and *KSP-adaptive*.

Performance metrics: We evaluate the proposed schemes in three different ways. First, we use a throughput model [2] that approximates the throughput for a given traffic pattern and a topology. Second, we use Booksim 2.0 [15], a cycle-accurate interconnection network simulator, to evaluate the aggregate throughput and packet latency for different traffic conditions. Finally, we use the CODES 1.0.0 [16] to measure the communication times for common communication patterns in HPC applications.

Throughput model: The throughput model [2] estimates the throughput for multipath routing schemes with multipath TCP (MPTCP [8]) like protocol where each flow is realized by multiple (k) sub-flows. Given a communication pattern, the model first computes the number of times each link is used by all sub-flows in the pattern. It then finds the link loads. For any link l with capacity C that is used X times; the link load defined as $load_l = \frac{X}{C}$. The model sets the rate for each sub-flow in the pattern to be equal to the reciprocal of the maximum load in the links along the path, $\frac{1}{max_{l \in path_n(s,d)} load_l}$. Finally, the model adds the rates for the k sub-flows to obtain the total rate for the flow.

$$T(s,d) = \sum_{n=1}^{k} \frac{1}{max_{l \in path_n(s,d)} load_l}. \quad (1)$$

Simulator modification and settings: Both Booksim and CODES do not have native support for the Jellyfish topology. We extended the software of both simulators to include Jellyfish. Four path calculation methods (KSP, rKSP, EDKSP, and rEDKSP) are added for both Booksim and CODES. For Booksim, five routing mechanisms are added: *random*, *round-robin*, *vanilla-UGAL*, *KSP-UGAL*, and *KSP-adaptive*. For CODES, two routing mechanisms are added: *random* and *KSP-adaptive*.

Booksim parameters are similar to those used in [14], [17] for simulating HPC systems. The UGAL routing techniques are set to have no bias towards MIN or VLB paths. We assume single-flit packets and a factor 2.0 router speedup, because our main focus is on evaluating routing performance, rather than on flow control and router delays. The latency of the channels are set to 10 cycles. To avoid deadlocks, we increase the virtual channel number every time a packet takes a new hop, so the total number of virtual channels is equal to the diameter of the network. The buffer size is 32 for each virtual channel. For each data point, Booksim warms up for 500 cycles and then collects the results over a window of 5000 cycles divided into 10 samples, each sample of a window of 500 cycles. Booksim considers the network to be saturated when the average packet latency of a sample exceeds 500 cycles. We record the last injection rate before the network reaches the saturation point as the network throughput.

CODES has more control parameters than Booksim. We set the router delay, soft delay, copy per byte and nic delay as 0, and other key parameters to be the same as those in Booksim to make sure that the evaluation with CODES has no extra latency than that in Booksim. The link bandwidth is set to 20GBps in the simulation. The packet size is 1500 Bytes and the buffer size is 64 packets.

Traffic patterns: For the throughput model, four different traffic patterns are used in the evaluation: random *permutation*, random *shift-N*, *Random*(X), and *All to All*. With the *permutation* pattern, each processing node communicates with at most one other processing node. In a *shift-N* pattern, a processing node i communicate with processing node $(i+N)$ *mod number_of_processing_node*. A random permutation is a randomly generated permutation traffic pattern while a random shift-N is a randomly generated shift pattern. With *Random(X)* pattern, each processing node sends traffic to X randomly picked destinations. In the *All to All* pattern, each processing node communicates with all other processing nodes in the system.

For Booksim, three traffic patterns are used in the evaluation: random *permutation*, random *shift-N*, and *uniform-random*. With *uniform-random*, the probability of sending a packet to each destination is equal.

For CODES, the communication times of four Stencil communication patterns are evaluated: : 2D Nearest Neighbor (2DNN), 2D Nearest Neighbor with diagonal

(2DNNdiag), 3D Nearest Neighbor (3DNN), 3D Nearest Neighbor with diagonal (3DNNdiag). These nearest neighbor communications are widely used in HPC applications. We collect DUMPI traces [18] for these applications, making sure that the trace size is same as the network size. For example, for $RRG(720,24,19)$, the network has 3600 compute nodes in total. So we collect the traces for 3600 ranks and keep the dimension size of 2DNN and 2DNNDiag as 60×60. Similarly, for 3DNN and 3DNNDiag, we keep the dimension size as $16 \times 15 \times 15$. For all of the applications, each process sends a total of 15MB data, which is divided among the flows from the process. The main motivation behind sending this amount of data is to make the network significantly utilized. For example, in 2DNN, each process sends to 4 neighbors. Thus, each neighbor will receive $\frac{15}{4} = 3.75MB$ of data. Physical communication patterns are also affected by the process-to-node mapping. In our study, two mappings for these applications are simulated, linear mapping where processes are mapped to the nodes in the network sequentially and random mapping where processes are randomly mapped to the nodes in the network.

B. Properties of the multiple paths selected

Before we present our evaluation results, let us first discuss relevant properties of the paths computed with different path selection schemes. These properties will help us understand the performance results since the paths selected to carry traffic will have a profound impact on the communication performance.

Table II shows the average path length for $RRG(36,24,16)$, $RRG(720,24,19)$ and $RRG(2880,48,38)$ with KSP(8), rKSP(8), EDKSP(8), and rEDKSP(8). Intuitively, randomization should not statistically increase the average path length while edge-disjointness will result in longer paths although the exact impact of the heuristics on the path length depends on the topology. The table shows that both randomization and edge-disjointness do not increase the average path length on $RRG(36,24,16)$ and $RRG(2880,48,38)$, while there is about 4.6% average path length increase with EDKSP(8) and rEDKSP(8) on $RRG(720,24,16)$, which is small. Overall, the impact on the path length by the two heuristics is small, indicating that these heuristics will result in better paths without significantly increasing the path length.

Topology	KSP(8)	rKSP(8)	EDKSP(8)	rEDKSP(8)
$RRG(36,24,16)$	2.06	2.06	2.06	2.06
$RRG(720,24,19)$	3.02	3.02	3.16	3.16
$RRG(2880,48,38)$	2.94	2.94	2.94	2.94

Table II: Average path length ($k = 8$)

Tables III and IV give the load-balance properties of the k-paths computed with different schemes. Tables III shows the percentage of switch pairs whose k paths do not share any link. As can be seen in the table, with KSP and rKSP, the percentage is very low for all topologies, while EDKSP and rEDKSP guarantee that the paths are link disjoint. With a fixed k, link-disjoint paths offer higher throughput between a pair of switches than non-link-disjoint paths for the pair. Hence, the quality of the paths computed by EDKSP and rEDKSP is higher than that by KSP and rKSP. Table IV shows the maximum number of paths for a pair of switches sharing one link. With KSP, the value is 6 for the small topology and 7 for the two larger topologies. For the large topologies, this means that there exists at least one pair of switches whose 7 out of 8 paths share one link. Hence, for this pair, the bandwidth capacity is equivalent to 2 paths instead of 8. This demonstrates the impact that the biases in KSP can have on the path selection. The table also shows that randomization does not solve this problem, but the edge-disjoint heuristic solves the issue. Overall, the results indicate that KSP with the randomization and edge-disjoint heuristics yields higher quality paths than the vanilla KSP.

Topology	KSP(8)	rKSP(8)	EDKSP(8)	rEDKSP(8)
$RRG(36,24,16)$	56%	59%	100%	100%
$RRG(720,24,19)$	2%	3%	100%	100%
$RRG(2880,48,38)$	9%	22%	100%	100%

Table III: Percentage of switch pairs whose k paths do not share any link ($k = 8$)

Topology	KSP(8)	rKSP(8)	EDKSP(8)	rEDKSP(8)
$RRG(36,24,16)$	6	3	1	1
$RRG(720,24,19)$	7	7	1	1
$RRG(2880,48,38)$	7	6	1	1

Table IV: Maximum number of times one link is shared by the k paths for one single switch pair ($k = 8$)

C. Throughput modeling Results

Figures 4, 5, and 6 show the average modeling throughput per flow for $RRG(36,24,16)$, $RRG(720,24,19)$ and, $RRG(2880,48,38)$ respectively, on four different traffic patterns.

In this experiment, we create 10 random samples for each RRG topology. For shift traffic, random-permutations traffic, and random(X=50) traffic, 50 different random instances have been used for each topology sample. The average of the random samples for four routing schemes KSP(k=8), rKSP(k=8), EDKSP(k=8), and rEDKSP(k=8) are reported. The throughput value in the figure is the per node normalized throughput. A value of 1 means the flows from a node can communicate at the full link speed; a value of 0.8 means that the flow can communicate at 80% of the link speed (because of the communication bottleneck in other links). There are several interesting observations in these figures.

First, randomization has significant impacts on both KSP and EDKSP, and noticeably improve the modeled throughput. For example, for *Random*(50)) traffic on $RRG(36,24,16)$ (Figure 4), the average throughput for KSP(8) is 0.80 while for rKSP(8) is .89, 11.1% higher. For $RRG(2880,24,16)$ in Figure 6, the average model throughput for random permutation traffic with EDKSP(8) is 0.76 while that with rEDKSP is 0.88, 15.8% higher. The results indicate that systematic biases in path selection can cause serious performance issues on Jellyfish and randomization is an effective method to alleviate the problem.

Second, *rEDKSP* consistently achieves the highest performance, out-performing all other path selection schemes. For example, for random permutation on $RRG(36,24,16)$ in Figure 4, the average throughput is 0.82 for rKSP(8) and 0.86 for rEDKSP(8), 4.9% higher. For shift traffic on $RRG(2880,24,16)$ in Figure 6, the average throughput is 0.51 for rKSP(8) and 0.55 for rEDKSP(8), 7.8% higher. In our more detailed simulation, rEDKSP out-performs rKSP even more, which indicates that using edge-disjoint paths is effective in achieving load balancing and rEDKSP is the best performing path calculation method for multi-path routing on Jellyfish.

Finally, multi-path routing schemes consistently outperform single-path routing to a large degree, which is inline with earlier results [1], [2].

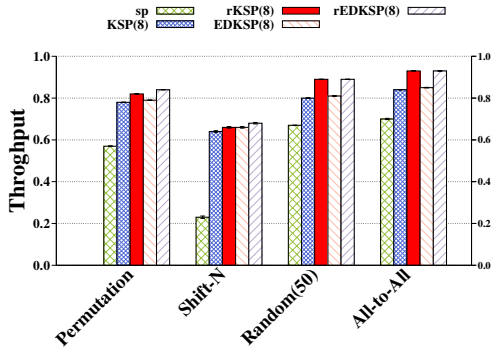


Figure 4: The average model throughput on $RRG(36,16,8)$

D. Results from Booksim

With the cycle-level simulation, Booksim results are more accurate than the model. Using Booksim, we investigate the performance of KSP, rKSP, EDKSP, and rEDKSP with different routing mechanisms.

Figures 7 and 8 show the average saturation throughput for the random permutation pattern with different path calculation and different routing mechanisms on $RRG(36,24,19)$ and $RRG(720,24,19)$, respectively. The results are the average of ten random patterns on each topology. We observe the following. First, across different routing mechanisms, rEDKSP consistently achieves the highest performance among the path selection schemes. The throughput is very significantly better than that for KSP. For example, using KSP-adaptive, on $RRG(720,24,19)$, rEDKSP achieves a throughput of 0.98, 11.3% higher than the throughput of 0.88 for KSP. This demonstrates the effectiveness of randomization and edge-disjointness. Second, the routing mechanism has a significant impact on the performance; adaptive routing (UGAL, KSP-UGAL, and KSP-adaptive) performs better than traffic oblivious routing (Round-Robin and Random). Between the adaptive routing scheme, KSP-adaptive is significantly better than KSP-UGAL for all path selection schemes. For example, with rEDKSP, KSP-adaptive achieves a throughput of 0.98, 10.0% higher than the throughput of 0.89 for KSP-UGAL. Moreover, KSP-adaptive and KSP-UGAL both have higher performance than the vanilla UGAL. Although vanilla UGAL is more flexible in selecting non-minimal paths, restricting the paths to be "short" paths results in higher performance on Jellyfish.

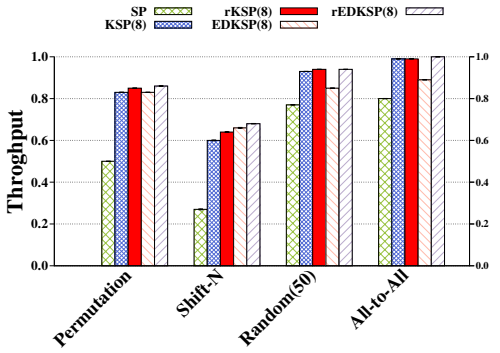


Figure 5: The average model throughput on $RRG(720,24,19)$

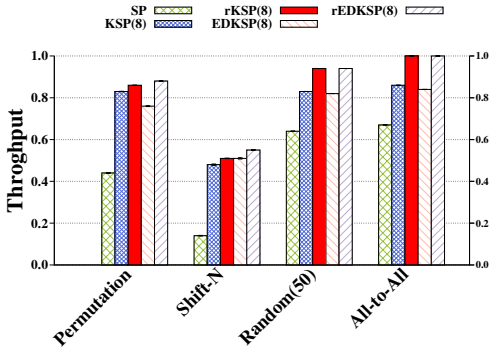


Figure 6: The average model throughput on $RRG(2880,48,38)$

Figures 9 and 10 show the average saturation throughput for running the previous experiment with random shift traffic patterns. The results are similar to that for permutation, except that the performance difference is larger. This is

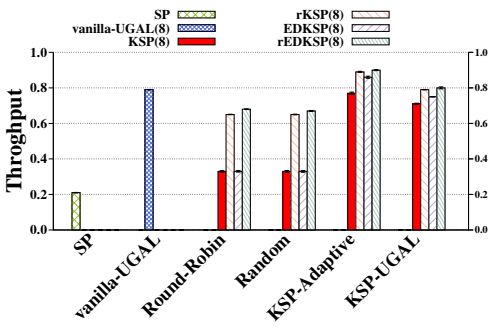


Figure 7: The average throughput of random permutations on $RRG(36,24,16)$

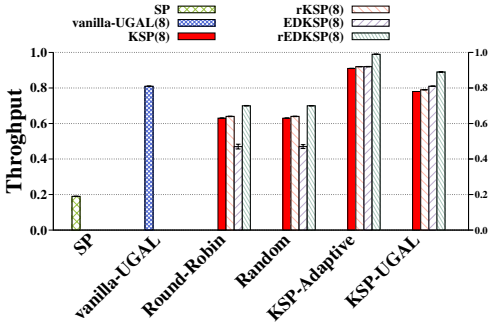


Figure 8: The average throughput of random permutations on $RRG(720,24,19)$

because an average shift traffic pattern is more demanding than an average permutation. The results indicate that KSP-adaptive is the best performing routing mechanism. With rEDKSP(8), KSP-adaptive achieves a throughput of 0.72, 20.0% higher than the 0.6 throughputs for KSP-UGAL. With KSP-adaptive, rEDKSP(8) achieves a throughput of 0.72, 20.0% higher than the 0.6 throughputs with KSP(8).

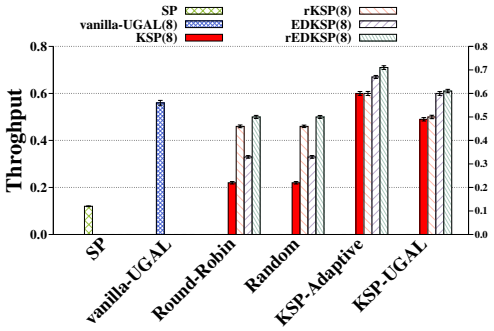


Figure 9: The average throughput of random shift on $RRG(36,24,16)$

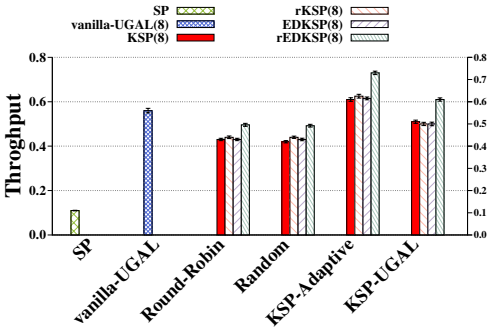


Figure 10: The average throughput of random shift on $RRG(720,24,19)$

Figures 11, 12 and 13 show the average packet latency as the offered load increases for $RRG(720,19,5)$ with three traffic conditions, the random-uniform traffic, a random permutation traffic, and a random shift traffic, respectively. The routing mechanism is KSP-adaptive. The results are consistent with all three traffic patterns. At low loads, all path selection schemes have similar latency. At high loads, rEDKSP achieves higher throughput and lower latency near saturation. This is because randomization and edge-disjointedness heuristics provide a better load-balance compared to KSP. The results demonstrate the effectiveness of rEDKSP, especially in comparison to the vanilla KSP.

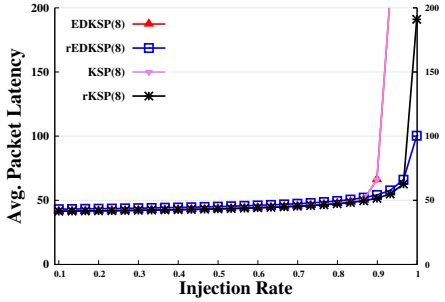


Figure 11: The average packet latency of random uniform on $RRG(720,19,5)$ on random routing scheme

E. Results from CODES

The communication times for Stencil communication patterns, which are very common in HPC application, are studied using CODES. Such results give a good indication about how the network performs in a more realistic setting. Table V shows the average communication times for each of the four applications with linear process-to-node mapping; The experiments are on $RRG(720,24,19)$, the link bandwidth is set to 20GBps. For each application, each process sends a total of 15MB data. The second column is the communication time with rEDKSP(8). The third column is

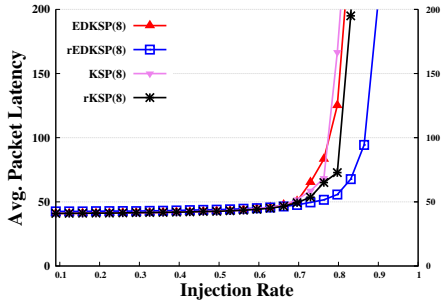


Figure 12: The average packet latency of a random permutation pattern on $RRG(720, 24, 19)$ on adaptive routing scheme

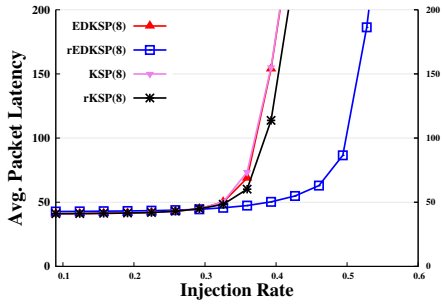


Figure 13: The average packet latency of a random shift-N pattern on $RRG(720, 24, 19)$ on adaptive routing scheme

the communication with KSP(8). The fourth column is the improvement percentage of rEDKSP(8) over KSP(8). The fifth column is the communication with rKSP(8). The sixth column is the improvement percentage of rEDKSP(8) over rKSP(8). With linear mapping, rEDKSP out-performs KSP and rKSP for all four Stencil patterns. On average, rEDKSP(8) improves over KSP(8) by 7.6% and over rKSP(8) by 4.5%. This is consistent with our results from the model and Booksim.

Applications	rEDKSP(8)	KSP(8)		rKSP(8)	
	time	time	imp.	time	imp.
2DNN	0.83	0.91	9.6%	0.88	6.0%
2DNNdiag	1.07	1.20	12.1%	1.15	7.5%
3DNN	0.90	0.95	5.6%	0.93	3.3%
3DNNdiag	1.01	1.04	3.0%	1.02	1.0%
Average			7.6%		4.5%

Table V: Communication time in milliseconds for different routing schemes with linear mapping of processes to nodes on $RRG(720, 24, 19)$

Table VI shows the results with random process-to-node mapping. The trend is similar to that with the linear mapping. On average rEDKSP(8) out-performs KSP(8) by 9.0%, and rKSP(8) by 0.8%. For 2DNNdiag, rEDKSP(8) performs slightly worse than rKSP(8). For the random mapping, rEDKSP performs very similar to rKSP. We attribute this to the random traffic pattern resulted from the random mapping. Overall, rEDKSP is still slightly better rKSP in this condition. In summary, for the Stencil patterns, rEDKSP consistently achieves higher performance than other path selection methods.

Applications	rEDKSP(8)	KSP(8)		rKSP(8)	
	time	time	imp.	time	imp.
2DNN	0.92	0.99	7.6%	0.94	2.2%
2DNNdiag	0.86	0.92	7.0%	0.84	-1.5%
3DNN	0.88	0.95	8.0%	0.88	0.0%
3DNNdiag	0.76	0.86	13.2%	0.78	2.6%
Average			9.0%		0.8%

Table VI: Communication time in milliseconds for different routing schemes in random mapping of processes to nodes one $RRG(720, 24, 19)$

V. Related work

Singla et al. [1], [19] propose Jellyfish for data centers to compete with the fat-tree topology [20]. Follow-up studies conclude that Jellyfish is more scalable than fat-tree for large HPC systems [2]. Variations of KSP have been suggested for Jellyfish [1], [2]. KSP allows randomization and link-disjointness heuristics to be incorporated. However, the effectiveness of the heurstics is topology dependent and it has not been thoroughly examined on Jellyfish. Our study indicates that, for the Jellyfish topology, the randomization and link-disjoint heuristics yield better paths, and significantly improve the routing performance. Routing mechanisms for multi-path routing on Jellyfish have not been systematically investigated. An ad hoc study was carried out where Jellyfish is compared with other topologies [21]. Like the Dragonfly topology where UGAL has many forms [10]–[14], multi-path routing on Jellyfish also allows for UGAL variations. We evaluate their performance and identify an effective routing mechanism for Jellyfish.

VI. Conclusion

We study two components of multi-path routing on Jellyfish, path selection and routing mechanism. We show that the current KSP routing for Jellyfish suffers from the load imbalance problem. We introduce two heuristics, randomization and edge-disjointness, to overcome this issue, and demonstrate that these two heuristics yield significantly better performance than the vanilla KSP scheme. We investigate various routing mechanisms for multi-path routing on Jellyfish. Our results indicate that using paths computed by KSP with the randomization and edge-disjointness heuristics and the KSP-adaptive scheme significantly improves Jellyfish's communication performance compared to existing schemes.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grants CICI-1738912, CRI-1822737, and SHF-2007827. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This work used the XSEDE Bridges resource at the Pittsburgh Supercomputing Center (PSC) through allocations ECS190004 and CCR200042.

REFERENCES

[1] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking data centers randomly," in *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, 2012, pp. 225–238.

[2] X. Yuan, S. Mahapatra, W. Nienaber, S. Pakin, and M. Lang, "A new routing scheme for jellyfish and its performance with hpc workloads," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2013, pp. 1–11.

[3] A. Singla, P. B. Godfrey, and A. Kolla, "High throughput data center topology design," in *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'14. USA: USENIX Association, 2014, p. 29Ű41.

[4] A. Singh, "Load-balanced routing in interconnection networks," Ph.D. dissertation, Stanford University, 2005.

[5] W. Hoffman and R. Pavley, "A method for the solution of the n th best path problem," *Journal of the ACM (JACM)*, vol. 6, no. 4, pp. 506–514, 1959.

[6] J. Y. Yen, "Finding the k shortest loopless paths in a network," *management Science*, vol. 17, no. 11, pp. 712–716, 1971.

[7] E. Q. Martins and M. M. Pascoal, "A new implementation of yens ranking loopless paths algorithm," *Quarterly Journal of the Belgian, French and Italian Operations Research Societies*, vol. 1, no. 2, pp. 121–133, 2003.

[8] H. Han, S. Shakkottai, C. V. Hollot, R. Srikant, and D. Towsley, "Multi-path tcp: A joint congestion control and routing scheme to exploit path diversity in the internet," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1260–1271, 2006.

[9] Y. Guo, F. Kuipers, and P. Van Mieghem, "Link-disjoint paths for reliable qos routing," *International Journal of Communication Systems*, vol. 16, no. 9, pp. 779–798, 2003.

[10] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Proceedings of the 35th Annual International Symposium on Computer Architecture*, ser. ISCA '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 77–88. [Online]. Available: http://dx.doi.org/10.1109/ISCA.2008.19

[11] N. Jiang, J. Kim, and W. J. Dally, "Indirect adaptive routing on large scale interconnection networks," in *Proceedings of the 36th Annual International Symposium on Computer Architecture*, ser. ISCA '09. New York, NY, USA: ACM, 2009, pp. 220–231. [Online]. Available: http://doi.acm.org/10.1145/1555754.1555783

[12] P. Faizian, J. F. Alfaro, M. S. Rahman, M. A. Mollah, X. Yuan, S. Pakin, and M. Lang, "TPR: traffic pattern-based adaptive routing for dragonfly networks," *IEEE Trans. Multi-Scale Computing Systems*, vol. 4, no. 4, pp. 931–943, 2018.

[13] M. A. Mollah, W. Wang, P. Faizian, M. S. Rahman, X. Yuan, S. Pakin, and M. Lang, "Modeling universal globally adaptive load-balanced routing," *ACM Trans. Parallel Comput.*, vol. 6, no. 2, Aug. 2019. [Online]. Available: https://doi.org/10.1145/3349620

[14] M. S. Rahman, S. Bhowmik, Y. Ryasnianskiy, X. Yuan, and M. Lang, "Topology-custom ugal routing on dragonfly," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '19. New York, NY, USA: ACM, 2019, pp. 17:1–17:15. [Online]. Available: http://doi.acm.org/10.1145/3295500.3356208

[15] N. Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D. E. Shaw, J. Kim, and W. J. Dally, "A detailed and flexible cycle-accurate network-on-chip simulator," in *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, April 2013, pp. 86–96.

[16] J. Cope, N. Liu, S. Lang, P. Carns, C. Carothers, and R. Ross, "Codes: Enabling co-design of multilayer exascale storage architectures," in *the Workshop on Emerging Supercomputing Technologies*, 2012.

[17] J. Won, G. Kim, J. Kim, T. Jiang, M. Parker, and S. Scott, "Overcoming far-end congestion in large-scale networks," in *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*, Feb 2015, pp. 415–427.

[18] C. L. Janssen, H. Adalsteinsson, S. Cranford, J. P. Kenny, A. Pinar, D. A. Evensky, and J. Mayo, "A simulator for large-scale parallel computer architectures," *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 1, no. 2, pp. 57–73, 2010.

[19] S. A. Jyothi, A. Singla, P. B. Godfrey, and A. Kolla, "Measuring and understanding throughput of network topologies," in *SC'16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2016, pp. 761–772.

[20] X. Yuan, W. Nienaber, Z. Duan, and R. Melhem, "Oblivious routing in fat-tree based system area networks with uncertain traffic demands," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1439–1452, 2009.

[21] M. A. Mollah, P. Faizian, M. S. Rahman, X. Yuan, S. Pakin, and M. Lang, "A comparative study of topology design approaches for hpc interconnects," in *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. IEEE, 2018, pp. 392–401.