An Analysis of Long-Tailed Network Latency Distribution and Background Traffic on Dragonfly+

Majid Salimi Beni(✉) and Biagio Cosenza

Department of Computer Science, University of Salerno, Salerno, Italy
{msalimibeni,bcosenza}@unisa.it

Abstract. Modern computing systems are highly affected by large performance variability, resulting in a long tail in the distribution of the network latency. For communication-intensive applications, the variability comes from several factors such as the communication pattern, job placement strategies, routing algorithms, and most importantly, the network background traffic. Although recent high-performance interconnects such as Dragonfly+ try to mitigate this variability by employing advanced techniques such as adaptive routing or topological improvements, the long tail is still there.

This paper analyzes the sources of performance variability on a large-scale computing system with a Dragonfly+ network. Our quantitative study investigates the impact of several sources, including the locality of job placement, the communication pattern, the message size, and the network background traffic. To tackle the difficulty in measuring the network background traffic, we propose a novel heuristic that accurately estimates the network traffic and helps to identify those highly-varying communications that contribute to the long tail. We have experimentally validated our proposed background traffic heuristic on a collection of pattern-based microbenchmarks as well as two real-world applications, HACC and miniAMR. Results show that the heuristic can successfully predict most of those runs in long-tail at job submission time on both microbenchmarks and real-world applications.

Keywords: MPI · Interconnect · Congestion · Dragonfly+ · Topology

1 Introduction

The growing gap between communication and computation in high-performance computing emphasizes the importance of optimized data communication. It is today well-understood that, to reach the Exascale, computing systems should provide high-performance network interconnects that deliver both high bandwidth and low latency.

The Dragonfly+ topology [47] is a modern hierarchical interconnect that has been recently introduced as an extended implementation of Dragonfly [30]. Such

A. Gainaru et al. (Eds.): Bench 2022, LNCS 13852, pp. 123–142, 2023.
https://doi.org/10.1007/978-3-031-31180-2_8

interconnect not only provides better network utilization and scalability in comparison to Dragonfly but also improves router buffer utilization [47]. However, despite Dragonfly+'s improvements compared to its predecessor, it still suffers from performance variability, especially with higher network congestion. Performance variability affects both system and applications' performance, and the batch scheduler must have a more precise estimation of applications' runtime to make accurate scheduling decisions [53,67].

Several users use large-scale compute clusters simultaneously, with different utilization patterns regarding program workflow, number of nodes, and data communication. While single-node computes units are typically not shared between users, the network is a shared resource. Network elements such as routers and links, shared among several jobs, are subject to contention. They negatively impact users' program performance by degrading I/O and slowing communication time. To address these issues, recent work has focused on monitoring, predicting, and balancing network traffic [12,32,33,58], as well as taking topological and network designing aspects into account [7,9,22,52]. In fact, the network has been identified as the main reason for performance variability [5,10,11,48].

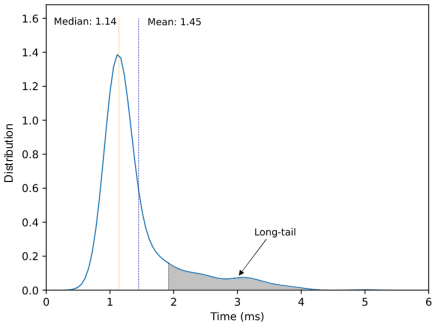


Fig. 1. Long-tail of the latency distribution on Dragonfly+.

1.1 Motivations

As performance variability is affected mainly by the network, it is essential to understand how network latency behaves on modern large-scale compute clusters. Figure 1 shows the frequency distribution of 1000 iterations of a latency test (MPI_Reduce in this case) on 16 nodes of the Marconi100 compute cluster with a Dragonfly+ topology. Interestingly, the results show a so-called *long-tailed* distribution. While a majority of the communication latencies are distributed around the median, more than 15% of the runs' latencies are larger than the 85th percentile (1.92 ms). The presence of such a long tail in the distribution also indicates that the distribution is not symmetric (e.g., not Gaussian), and there is a large gap between the mean and median. Also, the long tail negatively impacts the overall network performance by making the job execution highly unpredictable. While such performance variability is related to several

network-related factors, our work aims to analyze the main reasons behind such performance degradation, from the application's communication patterns to the external network traffic involving all users.

At the topology level, our work focuses on the Dragonfly+, which has better network utilization [47] than Dragonfly (known to suffer from performance variability [37,50]) and it is becoming a common topology in newly developed supercomputers [34,42].

1.2 Contributions

This paper conducts a performance variability study on a large-scale compute cluster with Dragonfly+ topology. The study comprises the analysis of several known sources of performance variability, in particular network-related aspects, including: different communications patterns, the impact of message size, the locality of job placement, and the effect of network background traffic generated by other users. The latter, in particular, is difficult to measure; to this end, we propose an easy-to-measure heuristic that estimates such traffic. As a part of the study, we further point out the effect of the adaptive routing strategy on the communication performance of Dragonfly+.

To the best of our knowledge, this is the first work that analyzes Dragonfly+ performance variability on a real supercomputer. While most related work relies on simulating background traffic [28,57], our approach is based on real-world data of background traffic extracted from a large-scale compute cluster. Insights from this analysis provide valuable feedback for job placement policy implementations on Dragonfly+ as well as network design for large-scale clusters.

The main contributions of this paper are:

- The first detailed **analysis of communication performance on a large-scale Dragonfly+ network based on real-world data:** We analyze different inter-node communication scenarios and show the performance variability of microbenchmarks with varying job placements.
- A **novel heuristics for background traffic estimation**, which is easy to measure and based on information known at job submission time.
- A comprehensive **correlation analysis between estimated background traffic and the communication performance**, with different communication patterns and message sizes.
- An **evaluation of the background traffic's impact on the long-tail of the latency distribution**.
- Further extension of the evaluation on **two communication-intensive real-world applications:** HACC[1] and miniAMR.

The rest of the paper is organized as follows: Sect. 2 and 3 introduce, respectively, related work and experimental setup. Section 4 presents our analysis of latency distribution, and Sect. 5 describes our background traffic measurement approach and its analysis. Section 6 is the discussion, and Sect. 7 concludes the paper.

[1] Hardware Accelerated Cosmology Code.

2 Related Work

A large part of the execution time of HPC applications is spent on transferring data between nodes; for this reason, considerable research efforts have been paid to investigating network topologies [4,20,26,39] and, on the application side, studying, analyzing, and optimizing communication on top of existing topologies [2,16,18,46,51,54,55].

Performance variability is often correlated with heavy-tailed distributions, which are probability distributions whose tails are not exponentially bounded [3]. In fact, when scaling up and increasing the complexity of a computing system, the tail of the latency distribution, which is not long in small systems, becomes more dominant at the large scale [14].

Bhatele et al. [5] analyzed the performance variability of Dragonfly with periodic system profiling of mini-applications; based on this analysis, they trained a machine learning model that predicts future executions. Groves et al. [19] studied the performance variability of the MPI_Allreduce collective in the Aries Dragonfly network and considered the relationship between different metrics such as process count, Aries counters, and message size with communication time, and showed the impact of background traffic on the performance.

Research on performance variability has investigated locality aspects and studied how topological locality and communication patterns affect different applications' performance [63]. Other research, however, considered other metrics such as network designs [13,44,60], routing strategies [8,15,27,38,40,50], congestion control [35,45] and background traffic [65]. Wilke et al. [61] discuss and compare existing challenges of Dragonfly and Fat-tree and show how different configurations and routing algorithms may affect QoS. They further illustrate the performance variability of Dragonfly while having various background traffic and different routing strategies. Alzaid et al. [1] have explored the Dragonfly network and measured the impact of different link arrangements between nodes and routing strategies on communication between nodes. They showed how data transfer through different links might be affected while the links tolerate different bandwidths.

Job allocation strategies have been recognized as a determinant factor in communication performance [29,36]. Level-Spread proposed by Zhang et al. [66] is a job allocation policy on Dragonfly that puts jobs in the minor network level that the current job can fit in to not only benefit from the node adjacency but also balance link congestion. Brown et al. [6] analyzed the relation between MPI communications and I/O traffic in Fat-tree networks; their analysis considers different parameters such as job allocation policies, message sizes, communication intervals, and job sizes. Wang et al. [59] have performed a comparative analysis of network interference on applications with nearest-neighbor communication patterns, considering various job placement strategies on Dragonfly. They show that having a trade-off between localized communication and a balanced network in job placement can reduce network interference and alleviate performance variability. In another work [58], they carried out an in-depth per-

formance analysis on Dragonfly and demonstrated how balanced network traffic and localized communication could impact different workloads.

Although related work has studied performance variability in Dragonfly, to the best of our knowledge, none of them have deeply investigated this variability in Dragonfly+. Moreover, we specifically show how background traffic affects different communication patterns, i.e., which collectives are more vulnerable to background traffic. Unlike most related work on background traffic, our analysis is based on real-world data (experiments have been conducted during a three-month time span at different times in order to have different background traffic) rather than simulations. Hence, the background traffic is generated by other users we have no control over, and we are not producing such traffic artificially.

3 Experimental Setup

Our analyses have been performed on a large-scale compute cluster, Marconi100 [34], available at the CINECA supercomputing center, which is currently ranked 18th in the TOP500 ranking [56].

3.1 Computing

The Marconi 100 cluster is an IBM Power System AC922 [43] consisting of 980 nodes, each of which is equipped with two IBM POWER9 AC922 multicore processors with 16 cores at 2.6 (3.1 turbo) GHz and four NVIDIA Volta V100 GPUs with 16GB, and 256 GB of per-node memory. All in all, the total number of CPU cores is 347,776, and it provides 347776 GB of memory.

3.2 Network

The internal interconnect of Marconi100 is a Mellanox InfiniBand EDR Dragonfly+. Figure 2 presents the Dragonfly+ topology implemented in this supercomputer. As shown, there are four large groups of nodes, each of which is called an

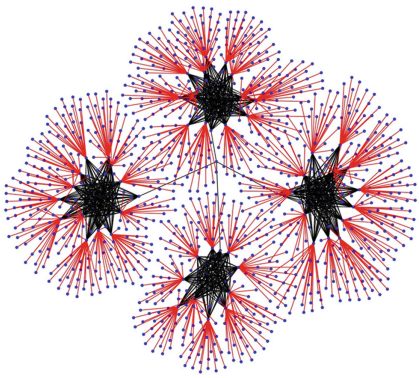


Fig. 2. The Dragonfly+ topology in Marconi100.

island. Within islands, there are smaller groups of nodes connected to one switch called *groups*. The main topological difference between Dragonfly and Dragonfly+ is that in Dragonfly+, intra-island routers are connected as a bipartite graph to improve the scalability.

It is worth mentioning that the Operating System is Red Hat Enterprise 7.6, IBM Spectrum-MPI 10.4 [25] is installed on the cluster, and SLURM [62] has the duty of resource management on this system. In addition, Adaptive Routing [17] is the default routing strategy used to prevent contention of the links and handle failures on the hardware.

3.3 Microbenchmarks and Applications

The main analysis and evaluation are done based on the OSU collection of microbenchmarks [41], which consists of three collectives, to which we added two real-world applications as summarized in Table 1. Moreover, to show the performance variability, each experiment is repeated in 1-millisecond intervals 1000 times in a loop (as suggested by [24] to perform at least 300 iterations), and, in all experiments, 1 MPI process is assigned to each physical node to leave other cores for the OS. Also, 16 physical nodes are allocated to the cluster in collective communications and application evaluations to partially involve all the islands in the communication.

Table 1. Benchmarks and applications used for the analysis.

Benchmark/App	Description	Evaluated sizes
Broadcast	Program calling Spectrum MPI_Bcast	$2^2, 2^{10}, 2^{15}, 2^{20}$(bytes)
Reduce	Program calling Spectrum MPI_Reduce	$2^2, 2^{10}, 2^{15}, 2^{20}$(bytes)
All-to-All	Representative of Spectrum MPI_Alltoall	$2^2, 2^{10}, 2^{15}, 2^{20}$(bytes)
HACC [21]	Includes various communication patterns	10M particles
miniAMR [23]	Includes various communication patterns	4K 3D blocks

4 Network Latency Distribution Analysis

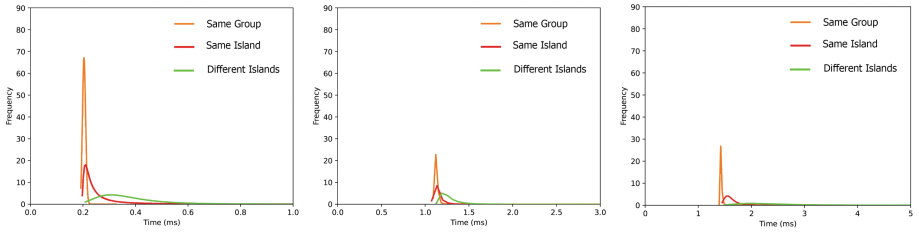
This section provides an analysis of the network latency on a Dragonfly+. First, we show the performance variability considering different locality levels for node allocation. Then, we show how the performance of microbenchmarks is affected when having different job allocation scenarios. Note that to make sure we are using the best-fitting distribution with minimum error in distribution plots, more than 100 different distributions have been fitted to the data.

4.1 Job Placement Locality and Performance Variability

Performance variability is the difference in an individual program's performance in consecutive executions. This section shows the impact of different job placement (node allocation) strategies on performance variability.

In our analysis, we consider three locality levels according to the Dragonfly+ topology and analyze the performance variability when having the following three node allocation scenarios:

a) **Same Group:** In this case, all required nodes are allocated in a single group. Therefore, only one network switch is involved in the communication between every two nodes.
b) **Same Island:** Nodes are allocated on one island, but they are distributed across different groups of that island. Hence, there is less locality than in the previous scenario.
c) **Different Islands:** Nodes are distributed on different islands. In this case, there is no limitation on allocating nodes; they are allocated everywhere on different islands and groups. In doing so, less locality is imposed.



(a) Broadcast (b) Reduce (c) AlltoAll

Fig. 3. Communication time frequency distribution of collective communications for 1000 iterations, with different allocation locality scenarios.

According to the defined locality levels, we focus on the role of both communication patterns and job placement on the performance variability and long-tail. In fact, we analyze different communication patterns to understand how they affect performance variability. The selected microbenchmarks include *one-to-all* (MPI Broadcast), *all-to-one* (MPI Reduce), and *all-to-all* (MPI AlltoAll).

We refined the analysis with a by-pattern study as shown in Fig. 3. This figure shows the frequency distribution of under-study collectives with different allocations on 16 nodes. For the *same group* job placement, all 16 nodes are allocated on the *same group* and connected through a single switch. For *different islands* mode, four nodes are allocated on each island in different groups. As illustrated, Broadcast (Fig. 3a) shows the best performance and shortest tail for all three allocation strategies; in fact, it benefits local communications more than other patterns, especially for the *same group*: it is not only faster than others (average time= 0.2), but also its peak is higher, which means that communication times of different iterations are very similar and there is a low performance variability. In Fig. 3a, the peaks of *different islands* and *same island* are 19 and 6, respectively, and they possess a peak much lower than the *same group* (68). However, they still show higher peaks than the correspondings in Reduce and

AlltoAll. For the Reduce (Fig. 3b), the average communication times of the *same group* and *same island* are almost the same (1.17 and 1.18 ms, respectively). However, with *different islands* we observe a slower average communication time (1.4 ms) and a much longer tail, reaching 10 ms. Finally, AlltoAll (Fig. 3c) is the slowest and most variable collective when all the nodes are on *different islands*. Its frequency distribution shows a very long tail (notice that the end of its tail is not shown in the figure), with a maximum observed communication time reaching 13 ms and a peak of 2.

Although allocating all nodes on the *same group* has been beneficial for collective communications, the number of nodes in each group of Dragonfly+ is limited (up to 20 nodes in Marconi 100), and the job scheduler cannot exclusively allocate to the *same group* more than the existing physical nodes. Even worse, large-scale compute clusters are typically used by several users that submit multiple jobs; in fact, very often, other nodes in the same group are already allocated by other users' jobs. In such cases, the job scheduler should necessarily allocate a job to nodes on different groups of that island or other islands unless we are willing to wait hours or even days until all the nodes in the same group are idle.

By default, SLURM [49] tries to place jobs on the currently idle nodes in the same group if the user does not specify particular nodes (in the host file). Because of the limited amount of idle nodes that can be found in the same group, SLURM's job scheduler looks for the switches (groups) with the fewest number of idle nodes and chooses the idle nodes connected to that switch, and repeats this process until it assigns all the requested nodes. So, based on the requested number of nodes by the user and the availability of cluster nodes, it may decide to assign jobs to nodes on different groups of the same island, or it spans over different islands, which the latter is the more probable scenario according to our observations.

5 Background Traffic Analysis

In real-world supercomputers, a single user does not operate on a dedicated system; instead, it submits jobs concurrently with other users. While resources such as computing nodes are typically allocated so that they are not shared between users at the same time, unfortunately, there is a resource for which some degree of contention is unavoidable: the network.

Intuitively, the larger the number of active jobs, the more probable the network congestion. More precisely, network congestion is more probable when users' jobs involve a larger number of nodes.

This section analyzes how the background traffic generated by other users' jobs affects the performance variability. In particular, we first define a simple heuristic that approximates the amount of network activity generated by other users' jobs. Successively, the analysis focuses on the correlation between background traffic with several communication patterns and message sizes.

5.1 Background Traffic Heuristic

The network congestion due to other users' activity is an essential cause of high-latency runs when using a large-scale compute cluster. We indicate with network background traffic: the external network traffic made by other users who are running their job simultaneously. To quantify how much such network activities impact the latency of our program communications, we have monitored the SLURM job queue before executing our jobs (i.e., we queried the *squeue* command before program execution).

In this way, we obtained information regarding the number of running and pending jobs, running jobs' runtime, as well as the number of nodes allocated by each job. Since we have no information about pending jobs and it is unclear when they will be running, they are not considered in our background traffic analysis. Besides, the running jobs that allocate only one node are excluded from our calculations because they have no communication with other nodes and, therefore, no effect on the background traffic (we experimentally observed many jobs that only allocate one node). Therefore, only jobs with the running status that allocate at least two nodes have been taken into account.

To better understand the background traffic with a simple and countable metric, we define a simple heuristic named *background network utilization* (b), which is defined as the number of unique nodes allocated by the running jobs and whose allocation includes at least two nodes over all the available nodes of the cluster. In other words, it shows the ratio of nodes contributing to communication to all the physical cluster nodes.

Formally, the background network utilization b ratio is defined as follows:

$$b = \frac{N_c}{N_t} \tag{1}$$

where:

N_c: number of unique nodes contributing to communication
N_t: total number of cluster physical nodes

In some cases, one node may be shared among different jobs by the scheduler in order to fully utilize its resources, e.g., each job takes a computation resource; which means that the node is being utilized by more than one communicating job, and we cannot count this node in our heuristic only once since the node produces higher background traffic. In order to take such cases into account, we count the shared node as many times it appears in the jobs' node lists that allocate more than two nodes. Hence, considering the appearance of some nodes more than once in the nodes list, the number of all running nodes can become larger than the cluster's physical nodes (N_t), which is a constant number. In an effort to resolve the problem and refine the heuristic, we consider the overhead of shared nodes by multiplying b by a new ratio which is: the number of nodes contributing to communication (consider some nodes might be counted more than once) to all the allocated running nodes (Similarly, we count each node as many times they appear in the jobs' nodes list). By doing so, we ensure that

we consider nodes contributing to different jobs and having communication with other nodes. Therefore, the refined version of the *background network utilization*, which will be considered in the rest of the paper, is defined as follows:

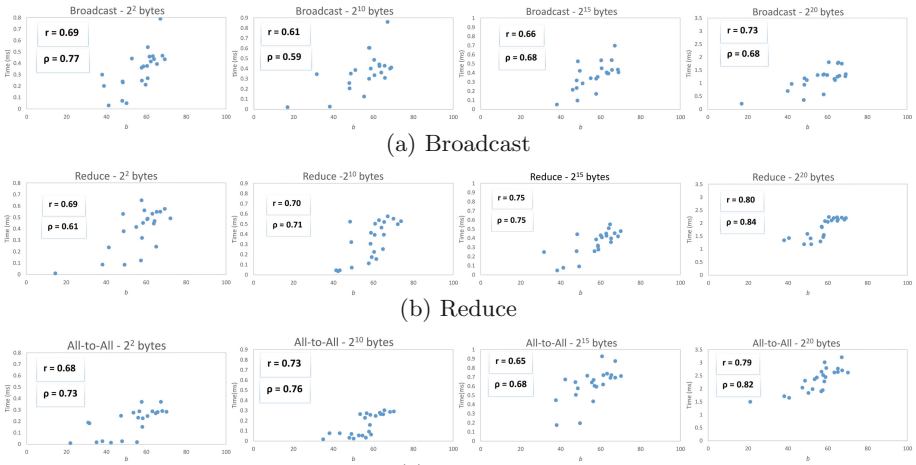
$$b = \frac{N_c}{N_t} * \frac{N'_c}{N_a} \tag{2}$$

where:

N'_c: the number of nodes contributing to communication (containing duplication)
N_a: all allocated running nodes (containing duplication)

Ideally, the value of b is 1 (or 100, if the percentage is taken into account) if running jobs allocate all the nodes and all of them are actively involved in communication, while b is 0 if non of the nodes are communicating or there is no active job at that moment. In order to make sure the measured b is showing a more accurate background network utilization and it has not changed during the microbenchmark's execution, we perform the *squeue* query also after the execution of each test and capture the b value only if the difference between two b values calculated is less than a threshold (5% in our experiments).

Note that some other network-related metrics, such as vendor-provided counters, can be also measured in some clusters to make precise network congestion measurement. However, not in all compute clusters are these counters available or accessible by non-admin users. Moreover, using such counters, the proposed method would not be portable to other clusters with different network infrastructure vendors. Therefore, we rely on data provided by SLURM, which is available on most clusters.



(a) Broadcast

(b) Reduce

(c) AlltoAll

Fig. 4. The relation between background traffic (b) and the average communication time of different collectives with different message sizes.

5.2 Correlation Analysis

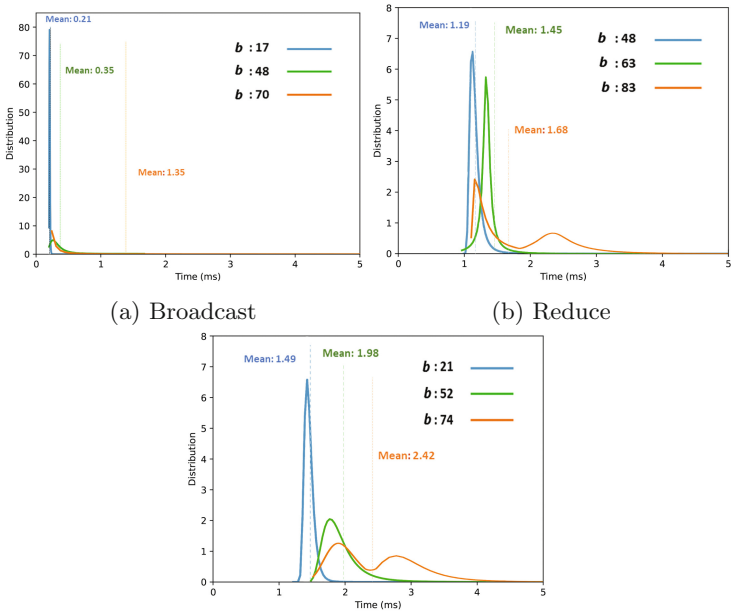
To evaluate how much the communication time is affected by the background traffic, we analyzed the correlation between the previously introduced b metric and the communication time over many runs with different workloads in terms of data sizes and communication patterns. In the evaluation, we used the Pearson Correlation Coefficient (r) [31] and Spearman Rank Correlation (ρ) [64] to analyze the relation between the two metrics. While Pearson's correlation shows if there is a linear relationship between data, Spearman's correlation evaluates the monotonic relationships in the data. In both, r, ρ: $r = +1$ or $\rho = +1$ means that there is a strong positive correlation between the variables, while $r = 0$ or $\rho = 0$ means independent variables. Figure 4 shows the correlation between background network utilization b and communication time for Broadcast (Fig. 4a), Reduce (Fig. 4b), and All-to-All (Fig. 4c) pattern, with different data sizes on 16 nodes allocated on *different islands*. We do not explore point-to-point communication here since it is not significantly affected by the background traffic. There are 22 points on each plot, and each point represents the average time of 1000 iterations. Experiments are performed in a three months time frame and represent experiments under different cluster utilization, i.e., different recorded background network utilization.

As shown in Fig. 4, the message transmission time is correlated with the *background network utilization* metric (b) and, overall, with increasing traffic, the communication time increases. In addition, as a general trend, with growing message size from 2^2, 2^{10}, and 2^{15} to 2^{20} bytes, the correlation between *background network utilization* and communication time becomes stronger, which means: the larger the data size is, the more the collective communication is affected by background traffic. Further, the correlations in Reduce collective for larger data (2^{15} and 2^{20} bytes) are higher than in others, meaning that in this collective, the communication time is highly dependent on the background traffic. Also, comparing the Pearson and Spearman correlations, Spearman shows a better fit for our use cases since it usually shows a more strong correlation.

It is worth mentioning that although background traffic is an essential factor that affects performance variability in communication-intensive jobs running on supercomputers, it is not the only player. Other reasons come from MPI itself, system activities, background daemons, garbage collection, queuing activities in intermediate servers and network switches, etc. [14,48]. Having said that, our *background network utilization* ratio is also an estimation relying on the obtainable information from other users. Hence, there might be possible errors in the measured runtimes, which is why some communications with smaller *background network utilization* have larger communication times, and the correlations are not +1.0 in Fig. 4.

5.3 The Impact of Background Traffic on Long-Tail

We have seen how performance variability is affected by the network background traffic for specific input sizes and communication patterns. In this section, we



(a) Broadcast (b) Reduce

(c) All-to-All

Fig. 5. Frequency distribution of communication times of 1000 iterations of Broadcast, Reduce, and All-to-All with different background network utilization.

go back to the motivation example and focus our analysis on the background traffic contribution to the long-tail effect. Figure 5 shows the frequency distribution of the execution time of 1000 iterations of 3 collectives on 16 nodes with message size 2^{20} bytes, with nodes allocated on *different islands*. For all three collectives, the higher the background network utilization, the lower the peak, and the longer the tail. For the Broadcast (Fig. 5a) and $b = 0.17$ (17%), the peak is very high, and there is a significant gap in the distribution of the higher and lower traffics; with higher background network utilization ($b = 0.70$), the tail of its corresponding distribution line is so long, which indicates that the communication performance is highly variable, ranging from 0.2 ms to 8 ms. Moreover, our experimental result reveals that the average execution time of 1000 iterations of Broadcast for $b = 0.70$ can be up to $6.4x$ larger than $b = 0.17$. Therefore, the Broadcast is highly affected by the background traffic, and, even if all the nodes are distributed on *different islands*, lower background traffic's performance can be as good as allocating all the nodes on the *same island*.

Similarly, in Figs. 5b and 5c, we observe that the distribution spreads at larger intervals with increasing background network utilization, and the tail becomes longer. For AlltoAll, especially when there is high background network utilization, the tail of the distribution is longer, the peak is lower, and the average communication time is larger than Broadcast and Reduce. Also, the mean of

distribution with $b = 0.74$ is around $1.6x$ larger than $b = 0.21$. In addition, unlike others, in AlltoAll, a significant shift in the peak of the charts (Median) of different background network utilizations is observed. In fact, this shift in the peak of different traffics is because of the All-to-All's inherent communication intensity: in this pattern, all nodes send their data to the others, and more data is sent through the network, making the network links more congested.

Besides, for higher background network utilization of Reduce and AlltoAll, the frequency distribution becomes dual (bimodal), which means that the higher amounts of iterations mainly happen at two different times instead of one. This behavior is related to the adaptive routing algorithm employed in this Dragonfly+ network. In adaptive routing, the router has multiple paths to choose from for each packet. In this way, some packets traverse on the shortest (minimal) path, and some go through an alternative, longer (non-minimal) one. Hence, some communications happen slower than the majority due to the penalty of selecting the non-minimal path. As demonstrated in Figs. 5b and 5c, when the network tolerates higher background network utilization, going through the non-minimal path becomes more probable that this either causes the distribution tail longer or makes it dual. Note that we cannot change the routing strategy since we are performing our experiments on a real compute cluster. Overall, it is clear how the background traffic pushes the tail. While the adaptive routing strategy helps mitigate the problem, there are cases where the problem still exists, particularly when there is very high background traffic.

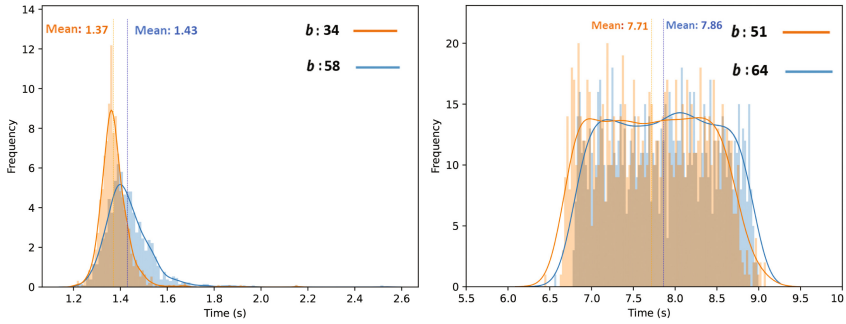
5.4 Application Analysis

So far, we have shown the impact of network background traffic and routing strategy on micro-benchmarks. In this section, we investigate the impact of background network utilization on two communication-intensive real-world applications that have shown to be affected by network congestion:

- HACC: a cosmology framework that performs n-body simulation to simulate the formation of structure in an expanding space.
- miniAMR: a mini-application that performs a stencil calculation on a unit cube computational domain.

Figure 6 shows the network latency distribution for HACC and miniAMR with both histogram and the frequency distribution. As shown in Fig. 6a for HACC, the average execution time and the peaks of $b = 34$ (the orange distribution) are 1.37 and 8.9, respectively. In contrast, for $b = 58$ (the blue distribution), the average time and peak reach 1.43 and 5.2, respectively. In other words, with a 24 percent increase in b, the average execution time increases by 4.4 percent. Moreover, both distributions in Fig. 6a are single and bell-shaped. However, the blue line is broadly distributed, and its tail reaches 2.5, while the orange line's tail is 2.1.

On the other hand, in Fig. 6b, when b changes from 51% to 64% and changes by 13, the average goes from 7.71 to 7.86 (2% increase). In contrast to all the



(a) HACC (b) miniAMR

Fig. 6. Frequency distribution of 1000 iterations of HACC and miniAMR applications with two different background network utilization.

observations, in this figure, both plots have multiple peaks, and a different behavior has been observed. Regarding the previous analysis on the two applications [65], in HACC, around 67% of the overall execution time of the application belongs to MPI operations. However, a tiny fraction (0.1%) is related to blocking collective communications. On the contrary, in miniAMR, 27% of total time belongs to MPI operations, in which 9.2% of the overall execution time belongs to only MPI_Allreduce, which means miniAMR performs more collective communications with the All-to-All pattern.

As we have demonstrated in Figs. 3 and 5, the All-to-All pattern is more prone to be affected by the network background traffic, and it has shown the flattest distribution when it is exposed to higher network background traffic in comparison to others. Moreover, the routing's effect can make its distribution bimodal. Looking over miniAMR's code, there are more than 10000 MPI_Allreduce operations which make the All-to-All pattern dominant. In Fig. 6b, the distribution becomes flat-topped that the main reason is because of its dominant All-to-All pattern, and its distribution is an aggregation of all of its dominant MPI_Allreduce communication latencies. Having said that, the routing algorithm will also play a role here because of the communication intensity of the All-to-All pattern, and we could expect a multi-modal distribution because of mixing many MPI_Allreduce distribution patterns.

6 Discussion

Our analysis of network latency distribution on a large-scale compute cluster with Dragonfly+ topology led to several insights. In terms of node allocation, there is a remarkable discrepancy between the *same group* and the two other allocation policies. When all the nodes are allocated to a single group, there is only one hop between every two nodes, which makes the communication minimally affected by the global background traffic. For the same reasons, in this

case, the minimal and non-minimal paths are the same for the adaptive routing (in contrast with the two other cases). So, it exhibits a latency distribution with the shortest tail and the higher peak. Hence, if there are enough available idle nodes on the *same group*, it is worth allocating all the required nodes there.

When analyzing the latency distribution according to the communication patterns, the Broadcast is the pattern that has significant benefit from the locality of the job allocation; in fact, results show that Broadcast has the shortest tail and higher peak and is faster than Reduce and All-to-All for both *same group* and *same island* allocations. However, when nodes are allocated on *different islands*, Broadcast is highly affected by the background traffic, showing a very long tail compared to the cases with lower background traffic. Moreover, when the background traffic is very low, Broadcast's allocation performance on *different islands* can be as variable as allocation on the *same group*. Nevertheless, since the introduced background network utilization has been between 0.40 and 0.70 most of the time, there is very little chance of being in this situation. On the other hand, All-to-All is the pattern with the most extended tail when the job placement expresses little locality on Dragonfly+. Although its distribution when allocating on the *same group* is similar to the Reduce on the *same group*, when performing All-to-All on *different islands*, the distribution tail becomes very long due to the higher amount of communication in All-to-All.

Among all possible sources of performance variability, it has been shown that the background traffic is the key factor in the performance variability of different collectives on Dragonfly+. Usually, with the increase in background traffic, the communication time of collectives takes longer. Additionally, collective communication increases with higher background traffic and larger message sizes.

On top of that, we have experimentally observed a two-peak distribution of the communication latency typically due to the adaptive routing algorithm, which offloads some packets to an alternative, longer path under congestion. Finally, when analyzing the latency distribution of a real-world communication-intensive application, the distribution is mostly affected by its dominant communication pattern, and the overall average execution time increases with an increment in the network background traffic.

7 Conclusion

In this paper, we showed the performance variability of Dragonfly+ and analyzed the impact of background traffic on the long-tailed distribution for different communication patterns. We proposed a novel network background traffic estimation method that relies on the data gathered from the job scheduler's execution queue. We further showed the relation between performance variability and message size and demonstrated how the adaptive routing algorithm impacts the distribution. Overall, this study considers different metrics, including communication patterns, message sizes, job placement locality, and background traffic, to show how they contribute to performance variability and long-tail. We have experimentally validated our proposed background traffic heuristic on a large-scale cluster, a collection of pattern-based microbenchmarks, and two real-world applications.

The insights coming of this paper can help either the user or the scheduler to make more optimal decisions by first, estimating the network congestion according to the user-level information, and second, submitting the job at an appropriate time to have the minimum network interference.

Acknowledgments. This research has been partially funded by the European High-Performance Computing Joint Undertaking (JU) under grant agreement No. 956137 (LIGATE project).

References

1. Alzaid, Z.S.A., Bhowmik, S., Yuan, X., Lang, M.: Global link arrangement for practical dragonfly. In: Proceedings of the 34th ACM International Conference on Supercomputing, pp. 1–11 (2020)
2. Aseeri, S.A., Chatterjee, A.G., Verma, M.K., Keyes, D.E.: A scheduling policy to save 10% of communication time in parallel fast Fourier transform. Concurr. Comput. Pract. Exp. e6508 (2021)
3. Beni, M.S., Cosenza, B.: An analysis of performance variability on dragonfly+topology. In: 2022 IEEE International Conference on Cluster Computing (CLUSTER), pp. 500–501 (2022). https://doi.org/10.1109/CLUSTER51413.2022.00061
4. Besta, M., et al.: Fatpaths: routing in supercomputers and data centers when shortest paths fall short. In: International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, pp. 1–18. IEEE (2020)
5. Bhatele, A., et al.: The case of performance variability on dragonfly-based systems. In: 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 896–905. IEEE (2020)
6. Brown, K.A., Jain, N., Matsuoka, S., Schulz, M., Bhatele, A.: Interference between I/O and MPI traffic on fat-tree networks. In: Proceedings of the 47th International Conference on Parallel Processing, pp. 1–10 (2018)
7. Brown, K.A., et al.: A tunable implementation of quality-of-service classes for HPC networks. In: Chamberlain, B.L., Varbanescu, A.-L., Ltaief, H., Luszczek, P. (eds.) ISC High Performance 2021. LNCS, vol. 12728, pp. 137–156. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78713-4_8
8. Chaulagain, R.S., Liza, F.T., Chunduri, S., Yuan, X., Lang, M.: Achieving the performance of global adaptive routing using local information on dragonfly through deep learning. In: ACM/IEEE SC Tech Poster (2020)
9. Cheng, Q., Huang, Y., Bahadori, M., Glick, M., Rumley, S., Bergman, K.: Advanced routing strategy with highly-efficient fabric-wide characterization for optical integrated switches. In: 2018 20th International Conference on Transparent Optical Networks (ICTON), pp. 1–4. IEEE (2018)
10. Chester, D., et al.: StressBench: a configurable full system network and I/O benchmark framework. In: IEEE High Performance Extreme Computing Conference, York (2021)
11. Chunduri, S., et al.: Run-to-run variability on Xeon Phi based cray XC systems. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–13 (2017)

12. De Sensi, D., Di Girolamo, S., Hoefler, T.: Mitigating network noise on dragonfly networks through application-aware routing. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–32 (2019)
13. De Sensi, D., Di Girolamo, S., McMahon, K.H., Roweth, D., Hoefler, T.: An in-depth analysis of the slingshot interconnect. In: International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, pp. 1–14. IEEE (2020)
14. Dean, J., Barroso, L.A.: The tail at scale. Commun. ACM **56**, 74–80 (2013). http://cacm.acm.org/magazines/2013/2/160173-the-tail-at-scale/fulltext
15. Faizian, P., et al.: TPR: traffic pattern-based adaptive routing for dragonfly networks. IEEE Trans. Multi-Scale Comput. Syst. **4**(4), 931–943 (2018)
16. Farmer, S., Skjellum, A., Grant, R.E., Brightwell, R.: MPI performance characterization on infiniband with fine-grain multithreaded communication. In: 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC), pp. 1102–1106. IEEE (2016)
17. Glass, C.J., Ni, L.M.: The turn model for adaptive routing. ACM SIGARCH Comput. Archit. News **20**(2), 278–287 (1992)
18. Grant, R.E., Dosanjh, M.G.F., Levenhagen, M.J., Brightwell, R., Skjellum, A.: Finepoints: partitioned multithreaded MPI communication. In: Weiland, M., Juckeland, G., Trinitis, C., Sadayappan, P. (eds.) ISC High Performance 2019. LNCS, vol. 11501, pp. 330–350. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20656-7_17
19. Groves, T., Gu, Y., Wright, N.J.: Understanding performance variability on the Aries dragonfly network. In: 2017 IEEE International Conference on Cluster Computing (CLUSTER), pp. 809–813. IEEE (2017)
20. Hashmi, J.M., Xu, S., Ramesh, B., Bayatpour, M., Subramoni, H., Panda, D.K.D.K.: Machine-agnostic and communication-aware designs for MPI on emerging architectures. In: 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 32–41. IEEE (2020)
21. Heitmann, K., et al.: The outer rim simulation: a path to many-core supercomputers. Astrophys. J. Suppl. Ser. **245**(1), 16 (2019)
22. Hemmert, K.S., et al.: Evaluating trade-offs in potential exascale interconnect technologies (2020)
23. Heroux, M.A., et al.: Improving performance via mini-applications. Sandia National Laboratories, Technical report. SAND2009-5574, vol. 3 (2009)
24. Hunold, S., Carpen-Amarie, A.: Reproducible MPI benchmarking is still not as easy as you think. IEEE Trans. Parallel Distrib. Syst. **27**(12), 3617–3630 (2016)
25. IBM Spectrum MPI, accelerating high-performance application parallelization. https://www.ibm.com/products/spectrum-mpi. Accessed 01 May 2022
26. Jeannot, E., Mansouri, F., Mercier, G.: A hierarchical model to manage hardware topology in MPI applications. In: Proceedings of the 24th European MPI Users' Group Meeting, pp. 1–11 (2017)
27. Kang, Y., Wang, X., Lan, Z.: Q-adaptive: a multi-agent reinforcement learning based routing on dragonfly network. In: Proceedings of the 30th International Symposium on High-Performance Parallel and Distributed Computing, pp. 189–200 (2020)
28. Kang, Y., Wang, X., McGlohon, N., Mubarak, M., Chunduri, S., Lan, Z.: Modeling and analysis of application interference on dragonfly+. In: Proceedings of the 2019

ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, pp. 161–172 (2019). ISBN 9781450367233
29. Kaplan, F., Tuncer, O., Leung, V.J., Hemmert, S.K., Coskun, A.K.: Unveiling the interplay between global link arrangements and network management algorithms on dragonfly networks. In: 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 325–334. IEEE (2017)
30. Kim, J., Dally, W.J., Scott, S., Abts, D.: Technology-driven, highly-scalable dragonfly topology. In: 2008 International Symposium on Computer Architecture, pp. 77–88. IEEE (2008)
31. Kirch, W.: Pearson's correlation coefficient. In: Encyclopedia of Public Health, pp. 1090–1091 (2008)
32. Kousha, P., et al.: INAM: cross-stack profiling and analysis of communication in MPI-based applications. In: Practice and Experience in Advanced Research Computing, pp. 1–11 (2021)
33. Liu, Y., Liu, Z., Kettimuthu, R., Rao, N., Chen, Z., Foster, I.: Data transfer between scientific facilities - bottleneck analysis, insights and optimizations. In: 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 122–131 (2019)
34. Marconi100, the new accelerated system. https://www.hpc.cineca.it/hardware/marconi100
35. McGlohon, N., et al.: Exploration of congestion control techniques on dragonfly-class HPC networks through simulation. In: 2021 International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS), pp. 40–50. IEEE (2021)
36. Michelogiannakis, G., Ibrahim, K.Z., Shalf, J., Wilke, J.J., Knight, S., Kenny, J.P.: Aphid: hierarchical task placement to enable a tapered fat tree topology for lower power and cost in HPC networks. In: 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 228–237. IEEE (2017)
37. Mollah, Md.A., Faizian, P., Rahman, Md.S., Yuan, X., Pakin, S., Lang, M.: A comparative study of topology design approaches for HPC interconnects. In: 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 392–401. IEEE (2018)
38. Mollah, Md.A., et al.: Modeling universal globally adaptive load-balanced routing. ACM Trans. Parallel Comput. **6**(2) (2019)
39. Navaridas, J., Lant, J., Pascual, J.A., Lujan, M., Goodacre, J.: Design exploration of multi-tier interconnection networks for exascale systems. In: Proceedings of the 48th International Conference on Parallel Processing, pp. 1–10 (2019)
40. Newaz, Md.N., Mollah, Md.A., Faizian, P., Tong, Z.: Improving adaptive routing performance on large scale Megafly topology. In: 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid), pp. 406–416. IEEE (2021)
41. OSU micro-benchmarks 5.8 (2021). https://mvapich.cse.ohio-state.edu/benchmarks/
42. Ponce, M., et al.: Deploying a top-100 supercomputer for large parallel workloads: the Niagara supercomputer. In: Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning), pp. 1–8 (2019)
43. POWER9 processor chip. https://www.ibm.com/it-infrastructure/power/power9
44. Rahman, Md.S., Bhowmik, S., Ryasnianskiy, Y., Yuan, X., Lang, M.: Topology-custom UGAL routing on dragonfly. In: Proceedings of the International Confer-

ence for High Performance Computing, Networking, Storage and Analysis, SC 2019. Association for Computing Machinery, New York (2019). ISBN 9781450362290
45. Rocher-Gonzalez, J., Escudero-Sahuquillo, J., Garcia, P.J., Quiles, F.J., Mora, G.: Efficient congestion management for high-speed interconnects using adaptive routing. In: 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pp. 221–230. IEEE (2019)
46. Ruhela, A., Xu, S., Manian, K.V., Subramoni, H., Panda, D.K.: Analyzing and understanding the impact of interconnect performance on HPC, big data, and deep learning applications: a case study with infiniband EDR and HDR. In: 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 869–878. IEEE (2020)
47. Shpiner, A., Haramaty, Z., Eliad, S., Zdornov, V., Gafni, B., Zahavi, E.: Dragonfly+: low cost topology for scaling datacenters. In: 2017 IEEE 3rd International Workshop on High-Performance Interconnection Networks in the Exascale and Big-Data Era (HiPINEB), pp. 1–8. IEEE (2017)
48. Skinner, D., Kramer, W.: Understanding the causes of performance variability in HPC workloads. In: IEEE International 2005 Proceedings of the IEEE Workload Characterization Symposium, pp. 137–149. IEEE (2005)
49. Slurm, Slurm's job allocation policy for dragonfly network (2021). https://github.com/SchedMD/slurm/blob/master/src/plugins/select/linear/select_linear.c
50. Smith, S.A., et al.: Mitigating inter-job interference using adaptive flow-aware routing. In: International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2018, pp. 346–360. IEEE (2018)
51. Subramoni, H., Lu, X., Panda, D.K.: A scalable network-based performance analysis tool for MPI on large-scale HPC systems. In: 2017 IEEE International Conference on Cluster Computing (CLUSTER), pp. 354–358. IEEE (2017)
52. Suresh, K.K., Ramesh, B., Ghazimirsaeed, S.M., Bayatpour, M., Hashmi, J., Panda, D.K.: Performance characterization of network mechanisms for non-contiguous data transfers in MPI. In: 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 896–905. IEEE (2020)
53. Tang, W., Desai, N., Buettner, D., Lan, Z.: Analyzing and adjusting user runtime estimates to improve job scheduling on the Blue Gene/P. In: 2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS), pp. 1–11. IEEE (2010)
54. Teh, M.Y., Wilke, J.J., Bergman, K., Rumley, S.: Design space exploration of the dragonfly topology. In: Kunkel, J.M., Yokota, R., Taufer, M., Shalf, J. (eds.) ISC High Performance 2017. LNCS, vol. 10524, pp. 57–74. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67630-2_5
55. Temuçin, Y.H., Sojoodi, A.H., Alizadeh, P., Kitor, B., Afsahi, A.: Accelerating deep learning using interconnect-aware UCX communication for MPI collectives. IEEE Micro **42**(2), 68–76 (2022)
56. Top500, MARCONI-100. https://www.top500.org/system/179845/. Accessed 01 May 2022
57. Wang, X., Mubarak, M., Kang, Y., Ross, R.B., Lan, Z.: Union: an automatic workload manager for accelerating network simulation. In: 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 821–830 (2020)
58. Wang, X., Mubarak, M., Yang, X., Ross, R.B., Lan, Z.: Trade-off study of localizing communication and balancing network traffic on a dragonfly system. In: 2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 1113–1122. IEEE (2018)

59. Wang, X., Yang, X., Mubarak, M., Ross, R.B., Lan, Z.: A preliminary study of intra-application interference on dragonfly network. In: 2017 IEEE International Conference on Cluster Computing (CLUSTER), pp. 643–644. IEEE (2017)
60. Wen, K., et al.: Flexfly: enabling a reconfigurable dragonfly through silicon photonics. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2016, pp. 166–177. IEEE (2016)
61. Wilke, J.J., Kenny, J.P.: Opportunities and limitations of quality-of-service in message passing applications on adaptively routed dragonfly and fat tree networks. In: 2020 IEEE International Conference on Cluster Computing (CLUSTER), pp. 109–118. IEEE (2020)
62. Yoo, A.B., Jette, M.A., Grondona, M.: SLURM: simple Linux utility for resource management. In: Feitelson, D., Rudolph, L., Schwiegelshohn, U. (eds.) JSSPP 2003. LNCS, vol. 2862, pp. 44–60. Springer, Heidelberg (2003). https://doi.org/10.1007/10968987_3
63. Zahn, F., Fröning, H.: On network locality in MPI-based HPC applications. In: 49th International Conference on Parallel Processing-ICPP, pp. 1–10 (2020)
64. Zar, J.H.: Spearman rank correlation. In: Encyclopedia of Biostatistics, vol. 7 (2005)
65. Zhang, Y., Groves, T., Cook, B., Wright, N.J., Coskun, A.K.: Quantifying the impact of network congestion on application performance and network metrics. In: 2020 IEEE International Conference on Cluster Computing (CLUSTER), pp. 162–168. IEEE (2020)
66. Zhang, Y., Tuncer, O., Kaplan, F., Olcoz, K., Leung, V.J., Coskun, A.K.: Levelspread: a new job allocation policy for dragonfly networks. In: 2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 1123–1132. IEEE (2018)
67. Zhou, Z., et al.: Improving batch scheduling on Blue Gene/Q by relaxing 5D torus network allocation constraints. In: 2015 IEEE International Parallel and Distributed Processing Symposium, pp. 439–448. IEEE (2015)