A Study of Network Congestion in Two Supercomputing High-Speed Interconnects

Saurabh Jha*, Archit Patke*, Jim Brandt†, Ann Gentile†, Mike Showerman§, Eric Roman‡,
Zbigniew T. Kalbarczyk*, Bill Kramer*, and Ravishankar K. Iyer*

*University of Illinois at Urbana-Champaign
†Sandia National Labs
§National Center for Supercomputing Applications
‡National Energy Research Scientific Computing Center

Abstract—**Network congestion in high-speed interconnects is a major source of application runtime performance variation. Recent years have witnessed a surge of interest from both academia and industry in the development of novel approaches for congestion control at the network level and in application placement, mapping, and scheduling at the system-level. However, these studies are based on proxy applications and benchmarks that are not representative of field-congestion characteristics of high-speed interconnects. To address this gap, we present (a) an end-to-end framework for monitoring and analysis to support long-term field-congestion characterization studies, and (b) an empirical study of network congestion in petascale systems across two different interconnect technologies: (i) Cray Gemini, which uses a 3-D torus topology, and (ii) Cray Aries, which uses the DragonFly topology.**

I. INTRODUCTION

Despite years of innovation in network routing, congestion avoidance, and mitigation algorithms across generation high-performance interconnects, extreme-scale applications running on high-performance computing systems continue to suffer from performance variation and scaling challenges [1]–[3] due to (i) frequent exposure to congestion; and (ii) the inability to automatically optimize resource parameters (such as placement, rank mapping, and application scheduling) to improve network utilization. The current interconnects suffer from congestion that can occur because of (i) bad application placement (e.g., tightly packed ranks versus ranks spread across the network) [4], [5] and presence of bully applications [6]; (ii) the presence of large numbers of failed links [7] and new link failures, which force adversarial traffic shaping/flow on the network; and (iii) inherent congestion susceptibility due to network design choices (e.g., directional order routing in torus networks) or design bugs/gaps (e.g., incorrect/bad dynamic routing policy).

Growing body of work on network congestion [1], [3]–[8] reflect the challenges in achieving high application performance, and the difficulty of providing timely monitoring and diagnosis of network congestion. However, the prior studies and tools are limited to proxy applications, simulations and benchmarks and hence are not reflective of production characteristics and issues. This is mainly due to inability to monitor, collect, analyze data on network congestion. To fill this gap, we provide (a) an end-to-end monitoring and analysis tool for understanding congestion causes and support long-term field-congestion characteristics, and (b) an empirical study of network congestion in high-performance computing systems. In particular, we present empirical measurements and insights obtained from two generations of Cray interconnects by using the Lightweight Distributed Metric Service (LDMS) [9] as a monitoring tool, and Monet [10] as a diagnosis tool. The studied high-speed interconnects are: (a) a Cray Gemini network deployed on Blue Waters [11] which uses a 3D torus-based topology and direction order routing; and (b) a Cray Aries network deployed on Edison [12] and 2-cabinet experimental system that uses a DragonFly-based topology and dynamic routing. The study provides empirical data and insights to influence research directions in application runtime, networking and system development.

Our contributions include the following:

- Measurements obtained from a production system running production workloads,
- Demonstration of end-to-end monitoring and analysis framework at scale. The proposed end-to-end framework uses LDMS [9] for monitoring and Monet [10] for diagnosis of network congestion to generate actionable insights for application developers, system managers, and network designers. LDMS collects performance-related information on links via Cray's gpcdr [13] kernel module, whereas Monet uses a combination of data science and machine learning techniques on LDMS-collected data to enable online diagnosis, data summarization, and visualization.
- An empirical study of network congestion in petascale systems across two different interconnect technologies: (i) Cray Gemini, which uses a 3-D Torus topology; and (ii) Cray Aries which uses the DragonFly topology.

The key results obtained from empirical measurements are:

- Despite the use of low-level flow control and routing algorithms, hotspots in the network are common and exist for long duration of time.
- Heterogeneity in link bandwidth across different link-types (electrical and optical links) increases the susceptibility to congestion.
- Use of adaptive routing and a low-diameter network

	Blue Waters	**Edison**
Flow control	Credit-based	Credit-based
Technology	Cray Gemini	Cray Aries
Topology	3D Torus	Dragonfly
Routing	Directional-order routing	Adaptive
Number of Nodes	27,648	5,586

Table I: Differences between production systems

topology (such as the Cray Aries DragonFly topology) significantly improves congestion avoidance and mitigation compared to non-adaptive routing protocols, such as directional order routing in a high-diameter network topology (such as Cray Gemini Torus networks).

- Identification of design gaps in HPC interconnects. Using our monitoring and analysis, we found that routing algorithms in the production network may not choose the least congested path among symmetrical paths.

More detailed results on Blue Waters Cray Gemini network can be found in [10].

II. NETWORK AND DATA DESCRIPTION

In this paper, we demonstrate our monitoring tool on two production systems, Blue Waters and Edison, which use the Cray Gemini and Cray Aries interconnects respectively. Subsection II-A provides system description, Subsection II-B provides dataset description and Subsection II-C defines the congestion metric used in the study.

A. System and Interconnect Description

Table I shows differences and similarities between the two production system with respect to network interconnect technology features.

NCSA's (National Center for Supercomputing Applications) Blue Waters system is composed of 27,648 nodes and has a large-scale (13,824 x 48 port switches) Gemini 3D torus (dimension 24x24x24) interconnect. The available bandwidth on a particular network link is dependent on the link type (i.e., electrical vs optical) and number of tiles in the link. Where multiple links from a Gemini switch connect in the same direction, it is convenient to consider them as a *directionally aggreggated link* which we will henceforth call *link*, one in each of the 6 directions, X+/-, Y+/-,Z+/-, in the torus. For large XE/XK systems [14] all such aggregated X links have an aggregate bandwidth of 9.4 GB/s, Y links alternate between 9.4 GB/s and and 4.7 GB/s, and Z links are predominantly 15 GB/s with 1/8 of them at 9.4 GB/s. Gemini interconnect uses directional-order routing which is predominantly static.

NERSC's (National Energy Research Scientific Computing Center) Edison system is composed of 5,586 nodes and has an (1,440 x 48 port switches) Aries DragonFly interconnect with 15 electrical groups. Electrical links (*Green* and *Black*) connect Aries switches within the group, where as, optical links (*Blue*) form group to group connections. Optical and electrical links have a bandwidth of 1.56 GB/s and 1.75 GB/s

respectively. Adaptive routing is used to route packets on non-minimal paths to alleviate congestion on the minimal path.

B. Field-Congestion Datasets

Network performance on links is exposed via Cray's *gpcdr* kernel module. These are not collected nor made available for analysis via vendor-provided collection mechanisms. This data is collected and transported off the system for storage and analysis via the Lightweight Distributed Metric Service (LDMS) monitoring framework [9]. LDMS daemons synchronize their sampling (node to node time skew not accounted for) in order to provide coherent snapshots of network state across the whole system. The resolution of sampling is one second on Edison and sixty seconds on Blue Waters. In this work, we demonstrate the capability using one week of production data; that amounts to 7.7 TB for Edison and 370 GB for Blue Waters.

C. Congestion Metric

In this paper, we use *Percent Time Stalled* (*PTS*) as a congestion metric to quantify network congestion. It is a suitable metric for interconnect networks that use credit-based flow control algorithms. In credit-based flow control networks, a source is allowed to send packets to the destination only if the source has sufficient credits. If sufficient credits are not available, then the link stalls, and the percentage of time spent in stalled state per unit time quantifies the extent of congestion. In this paper, we refer to this metric (i.e., percentage of time spent in stalled state per unit time) as the *Percent Time Stalled*.
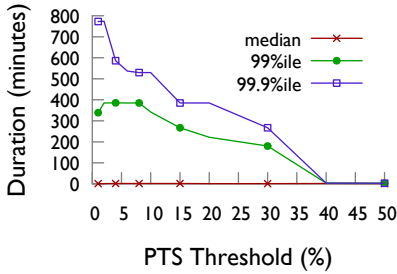
III. TOOL DEMONSTRATION

In this section, we describe the following two results obtained from Monet [10] tool on field-congestion data.

- impact of routing algorithms on congestion (see Subsection III-A)
- impact of heterogeneity in link-bandwidth on congestion (see Subsection III-B)
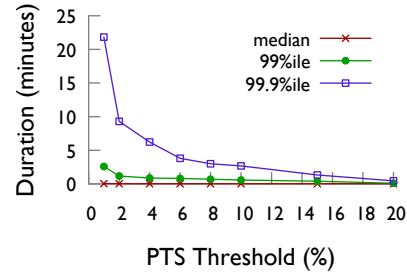
A. Impact of Routing Algorithms

Figure 1 shows the quantile values for different congested link durations, i.e., durations for which the *PTS* value on the link is above a fixed threshold (PTS_{th}). The figure leads to the following insights:
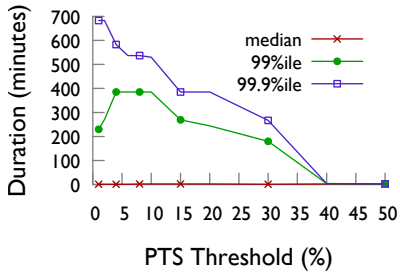
- *Use of the dragonfly topology and adaptive routing has led to improvement in congestion control between two generations of Cray interconnects.* The Dragonfly topology used in Aries has a low global diameter of one hop, which helps to contain the back pressure of congested links. Furthermore, adaptive routing allows packets to take a longer but less congested path, which helps to alleviate congestion on the minimal path. Figure 1 provides empirical evidence for that observation. For every PTS_{th} threshold, the congested link duration in Aries is an order of magnitude less than in Gemini. For example, if the threshold for congestion is fixed at 15% *PTS*, while the median duration is close to zero in both systems,
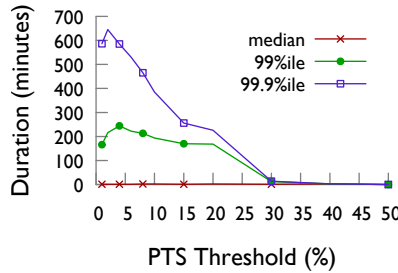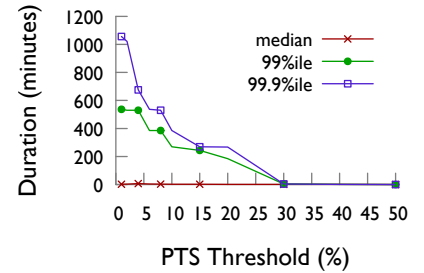
(a) Gemini

(b) Aries

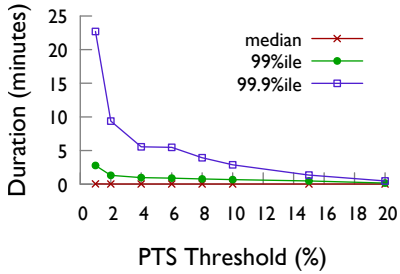**Figure 1:** Congested link durations vs. PTS threshold for Blue Waters (Gemini) and Edison (Aries)
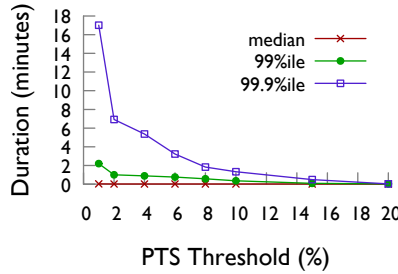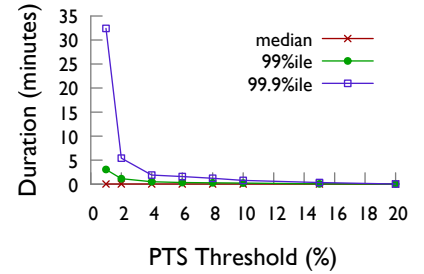

(a) X+ and X-

(b) Y+ and Y-

(c) Z+ and Z-

Figure 2: Congested link durations for different link types in Gemini

(a) Green

(b) Black

(c) Blue

Figure 3: Congested link durations for different link types in Aries

the 99.9th percentile duration is approximately 1 minute for Edison and 400 minutes for Blue Waters. However, while Aries manages long bouts of congestion better than Gemini does, application runtime variability due to network performance remains a concern [15].

- *Detection of long-duration congestion using traffic measurements can facilitate intervention such as rank remapping or rescheduling of bully jobs [6].* The 99.9th percentile congested link duration observed in both systems for $PTS_{th} \leq 20\%$ is greater than a minute. Such long duration congestion allows us to tolerate greater latency for detection and diagnosis in real time. Moreover, a diagnosis can be converted to actionable feedback to be used by tools such as TopoMesh [16], which can remap MPI ranks or the scheduler to reschedule bully jobs.

B. Impact of Heterogeneity in Link-bandwidth

Heterogeneity in link bandwidth across different link types (electrical and optical links) increases the susceptibility to congestion. Figure 2 (a), Figure 2 (b) and Figure 2 (c) respectively show congested link durations at different quantile values for *X*, *Y* and *Z* directional links of Cray Gemini interconnect in Blue Waters, and Figure 3 (a), Figure 3 (b) and Figure 3 (c) respecitvely show the congested link durations at different quantile values for *Green*, *Black* and *Blue* links of Cray Aries interconnect in Edison. In Gemini, for higher PTS_{th} thresholds ($\geq 20\%$), links along the *X* direction have longer lasting congestion than those on the *Y* and *Z* direction links. Similarly, in Aries, optical links (*Blue*) have shorter and less severe bursts of congestion than the electrical links (*Green* and *Black*). *Thus, mismatch and heterogeneity in link-*

bandwidth leads to varying levels of congestion along network path.

IV. Conclusion

In this work, we demonstrated the use of Monet [10] to conduct long-term characterization of field-congestion data obtained from petascale systems across two different interconnect technologies: (i) Cray Gemini, which uses a 3-D torus topology, and (ii) Cray Aries, which uses the DragonFly topology. Future work will include an in-depth analysis of field-congestion data and methods to alleviate congestion issues in HPC interconnects.

V. Acknowledgement*

References

[1] A. Bhatele, K. Mohror, S. H. Langer, and K. E. Isaacs, "There goes the neighborhood: performance degradation due to nearby jobs," in *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, 2013, pp. 41:1–41:12.

[2] N. Jain, A. Bhatele, S. White, T. Gamblin, and L. V. Kale, "Evaluating HPC networks via simulation of parallel workloads," in *High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for*. IEEE, 2016, pp. 154–165.

[3] T. Hoefler, T. Schneider, and A. Lumsdaine, "Characterizing the influence of system noise on large-scale applications by simulation," in *Proc. ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2010, pp. 1–11.

[4] T. Agarwal, A. Sharma, A. Laxmikant, and L. V. Kalé, "Topology-aware task mapping for reducing communication contention on large parallel machines," in *Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*. IEEE, 2006, pp. 10–pp.

[5] M. Mubarak, P. Carns, J. Jenkins, J. K. Li, N. Jain, S. Snyder, R. Ross, C. D. Carothers, A. Bhatele, and K.-L. Ma, "Quantifying I/O and communication traffic interference on dragonfly networks equipped with burst buffers," in *Cluster Computing, 2017 IEEE Int'l Conf. on*. IEEE, 2017, pp. 204–215.

[6] X. Yang, J. Jenkins, M. Mubarak, R. B. Ross, and Z. Lan, "Watch out for the bully!: job interference study on dragonfly network," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Press, 2016, p. 64.

[7] S. Jha, V. Formicola, C. Di Martino, M. Dalton, W. T. Kramer, Z. Kalbarczyk, and R. K. Iyer, "Resiliency of HPC Interconnects: A Case Study of Interconnect Failures and Recovery in Blue Waters," *IEEE Transactions on Dependable and Secure Computing*, 2017.

[8] S. Jha, J. Brandt, A. Gentile, Z. Kalbarczyk, and R. Iyer, "Characterizing supercomputer traffic networks through link-level analysis," in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2018, pp. 562–570.

[9] A. Agelastos *et al.*, "Lightweight Distributed Metric Service: A Scalable Infrastructure for Continuous Monitoring of Large Scale Computing Systems and Applications," in *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2014, pp. 154–165.

[10] S. Jha, A. Patke, B. Lim, J. Brandt, A. Gentile, G. Bauer, M. Showerman, L. Kaplan, Z. Kalbarczyk, W. T. Kramer, and R. Iyer, "Measuring congestion in high-performance datacenter interconnects," in *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, Feb 2020.

[11] "Blue Waters," https://bluewaters.ncsa.illinois.edu.

[12] "http://www.nersc.gov/users/computational-systems/edison/."

[13] Cray Inc., "Managing System Software for the Cray Linux Environment," Cray Doc S-2393-5202axx, 2014.

[14] K. Pedretti, C. Vaughan, R. Barrett, K. Devine, and S. Hemmert, "Using the Cray Gemini Performance Counters," in *Proc. Cray User's Group*, 2013.

[15] D. Wang, A. Bhatele, and D. Ghosal, "Performance variability due to job placement on edison," *Poster presented at SC14, Nov*, pp. 16–21, 2014.

[16] T. Agarwal, A. Sharma, and L. V. Kalé, "Topology-aware task mapping for reducing communication contention on large parallel machines," in *Proc. Int'l IEEE Parallel and Distributed Processing Symposium*, 2006.