Simulating and Visualizing Traffic on the Dragonfly Network

Abhinav Bhatele[†], Nikhil Jain[*], Yarden Livnat[‡], Valerio Pascucci[‡], Peer-Timo Bremer[†,‡]

[†]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551 USA
[*]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA
[‡]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT 84112 USA
E-mail: [†]{bhatele, ptbremer}@llnl.gov, [*]nikhil@illinois.edu, [‡]{yarden, pascucci}@sci.utah.edu

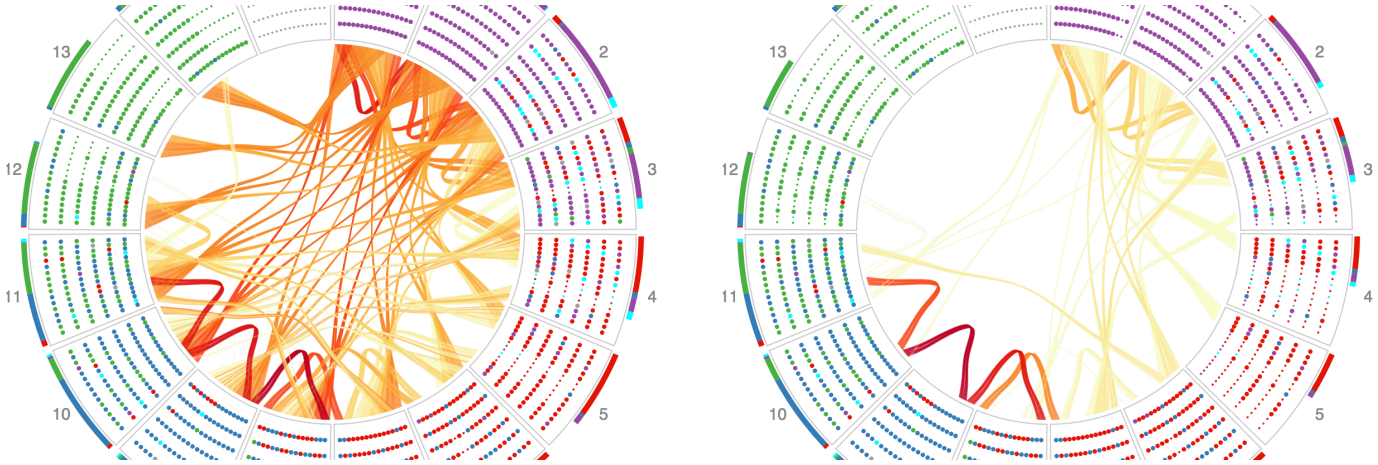

Fig. 1. Visualization of the inter-group links on a dragonfly network for a parallel workload comprised of four jobs (colored circles). The left configuration has fewer inter-group cables connecting the groups which leads to more network hot spots in comparison to the network configuration on the right.

I. Introduction

It has been expected that highly-connected network topologies such as the dragonfly coupled with adaptive routing strategies would greatly reduce or eliminate the effects of inter-job interference [1], [2]. However, preliminary experiments on Edison at NERSC have shown that for communication-heavy applications, inter-job interference and thus presumably job placement remains an important factor.

In this poster, we explore the effects of parallel workloads and network configurations on network throughput to better understand inter-job interference. In particular, we focus on the Edison supercomputer installed at NERSC/Lawrence Berkeley National Laboratory for studying pre- and post-deployment system configuration issues. Edison is a 30-cabinet (15 groups) Cray Cascade system [3] based on the dragonfly topology [4].

II. Experiments and Results

We use a simulation tool called Damselfly to estimate the steady state behavior of the dragonfly network for various parallel workloads and network configurations. The simulation tool provides: 1) Access to hardware counters for all routers on the system, something impossible to gather in production; and 2) The ability to easily study the impact of removing or adding network cables on network throughput.

We use the network connectivity provided to us by NERSC system administrators for the Edison installation. We use a placement policy similar to that used on Edison where we try to assign most nodes of a job as close to one another as possible within the same group or nearby groups. We use five different communication patterns that are representative of some of the application codes run at NERSC: 2D Stencil, 4D Stencil, Many-to-many, Spread and Unstructured Mesh. DragonView is an interactive visual analytics tool that we have developed to evaluate the impact of job placement policies, network interference and system configuration on the network utilization. We perform three kinds of experiments:

Individual Jobs: We simulated the five communication patterns described above for different job sizes ranging from 16,384 to 131,072 cores. Each simulation consisted of a single job occupying a part of the machine.

Parallel Workloads: We simulated workloads consisting of four parallel jobs, each of size 32k cores as shown in the table in the poster. For each workload, we pick four out of the

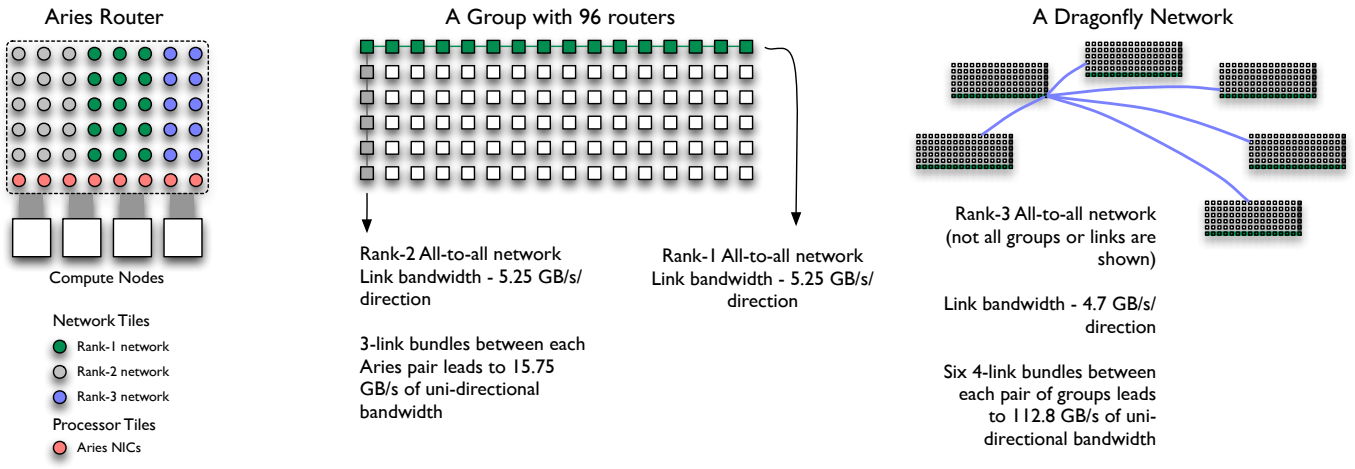

Fig. 2. Network configuration of the Cray Cascade (dragonfly/Aries) installation at NERSC – Edison.

five patterns one of which is always Spread to ensure some background traffic or congestion.

Modifying the Network: The number of cables to deploy on a smaller instance of the full-scale dragonfly design is a tradeoff between extra bandwidth and performance versus monetary cost. In our analysis, we observed that black links (see Figure 2) are seldom the most congested links for any of the workloads that we simulated. So we designed some experiments to remove some black cables from the system and analyze the impact on network throughput.

We observe that black links are usually not contended for. Depending upon the application pattern, the bottleneck is either on inter-group (blue) or intra-group (green) links. We show that when multiple jobs run in a parallel workload, the communication of each job gets restricted to fewer links to provide a fair share of bandwidth to other jobs. However, this leads to higher maximum traffic on the links. Again, this is observed on blue links for some patterns and green links for other patterns.

Finally, we performed some experiments that change the number of network cables (black and blue) on the dragonfly system. We find that removing one out of the three black links per router pair only has a small impact on the overall congestion in the network. However, adding a blue link and removing a black link per router pair can lower the hot-spots on inter-group connections.

III. Summary

The procurement, installation and operation of supercomputers at leadership computing facilities is expensive in terms of time and money. It is important that we understand and evaluate various system parameters that can impact overall system utilization and performance. In this poster, we touch upon two aspects of system utilization – inter-job interference and the impact of the network configuration on congestion. Using the network configuration of a production supercomputer (Edison) as a baseline and five different communication patterns, we evaluated the impact of one job's traffic on other jobs. We presented a simulation tool called Damselfly and a visual analytics system called DragonView, both of which can be useful tools for machine architects, system administrators and end users to understand network performance.

Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This work was funded by the Laboratory Directed Research and Development Program at LLNL under project tracking code 13-ERD-055 (LLNL-ABS-676088).

References

[1] A. Bhatele, K. Mohror, S. H. Langer, and K. E. Isaacs, "There goes the neighborhood: performance degradation due to nearby jobs," in *ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '13. IEEE Computer Society, Nov. 2013, LLNL-CONF-635776.

[2] J. Brandt, K. Devine, A. Gentile, and K. Pedretti, "Demonstrating improved application performance using dynamic monitoring and task mapping," in *Proceedings of the 1st Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications*, ser. HPCMASPA '14, 2014.

[3] G. Faanes, A. Bataineh, D. Roweth, T. Court, E. Froese, B. Alverson, T. Johnson, J. Kopnick, M. Higgins, and J. Reinhard, "Cray cascade: A scalable hpc system based on a dragonfly network," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '12. Los Alamitos, CA, USA: IEEE Computer Society Press, 2012.

[4] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," *SIGARCH Comput. Archit. News*, vol. 36, pp. 77–88, June 2008.