Analyzing Network Congestion on a Production Dragonfly-based System

Joy Kitson[†,*], Sudheer Chunduri[*], Abhinav Bhatele[†]

[†]Department of Computer Science, University of Maryland, College Park, MD 20742 USA
[*]ALCF, Argonne National Laboratory, Lemont, IL 60439 USA
Email: jkitson@umd.edu, sudheer@anl.gov, bhatele@cs.umd.edu

I. Introduction

As the HPC community continues along the road to exascale, and HPC systems grow ever larger and busier, the question of how network traffic on these systems affects application performance looms large. In order to fully address this question, the HPC community needs a broadened understanding of the behavior of traffic on production systems. We present an analysis of communications traffic on the Theta cluster at Argonne Leadership Computing Facility (ALCF), with a focus on how congestion is distributed - in both space and time - across the system.

II. Methodology

A. Data Collection

The Theta cluster is a Cray XC40 machine consisting of 4392 compute nodes, and is used for a variety of applications. For communication, it uses an Aries network with 1098 routers (one for every four compute nodes) with a Dargonfly topology. This layout consists of fully connected electrical groups spread across two cabinets each. These groups are in turn all directly connected to each other by optical links spread across the routers in each group. As a result, any given router is at most three hops from any other router, provided no routers are down. The links connecting routers come in three colors - blue, green, and grey - representing inter-group, intra-group column, and intra-group row connections, respectively. In addition, there are eight links connecting each router to four connected processors. To collect data on the operation of this system, we used the Lightweight Distributed Metric Service, (LDMS)[1], running continuously on the system for a period of two weeks. During that time, LDMS recorded the values of a variety of hardware counters every minute.

B. Data Processing

Over the two week period, LDMS captured around ten thousand snapshots of the system state. To pare this down to a more manageable number and to smooth over missing data, we separated the snapshots into bins. For the full data we used hour-long bins to give a coarse view of the data, and later used smaller bins to give a finer view of particular regions of interest. We selected the first complete snapshot from each bin to represent the start of that one hour (or shorter) period. In order to compute metrics, we compared the counters at the start of subsequent hours.

TABLE I
Counters used to collect each metric on each type of tile

Counter	Metric	Tile Type
AR_RTR_<r>_<c>_INQ_PRF_INCOMING_FLIT_VC{0-8}	Flits	Network
AR_RTR_<r>_<c>_PT_INQ_PRF_INCOMING_FLIT_VC{0,4}	Flits	Processor
AR_RTR_<r>_<c>_INQ_PRF_ROWBUS_STALL_CNT	Stalls	Network
AR_RTR_<r>_<c>_PT_PRF_ROWBUS_STALL_CNT	Stalls	Processor

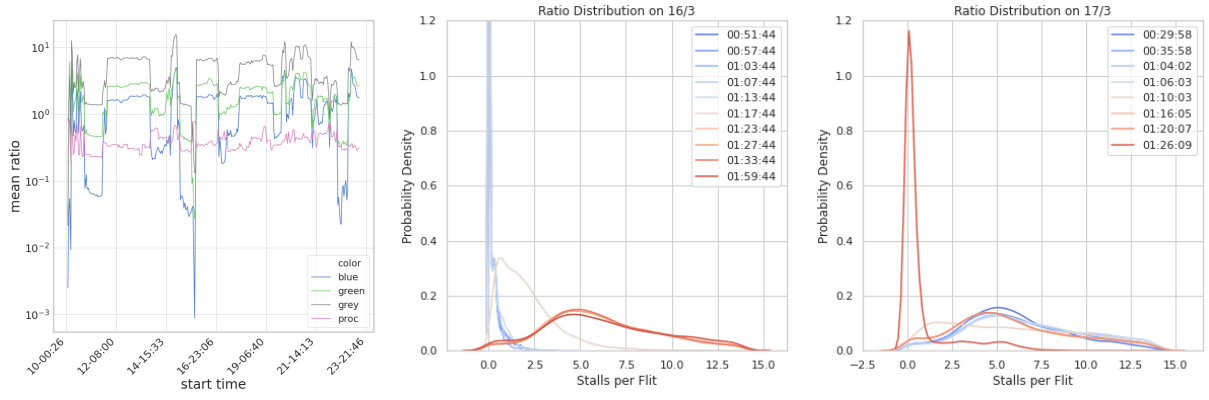
In particular, we examined several counters that together represent the number of flits - atomic units of communication [2] - sent by each tile on each router, and the number of cycles those tiles spent stalled. We found the raw counts of flits and stalls for each tile from the counters shown in Table I. We then calculated the rate of change by taking the difference between the corresponding counters from subsequent bins, and dividing by the time elapsed between them. While usually about an hour, this time period varied, mostly due to missing data. We used the ratio of stalls to flits as a proxy for the level of congestion because a high ratio indicates that a number of messages were left waiting to be sent [2]. We generally separate out the values by tile color when analyzing the data.

III. Results

A. Spatial Distribution

The LDMS data have two variables that describe how they are distributed in space: routers and tiles. Given the relatively flat topology of the system, these variables don't give a good quantitative sense of the distances between routers, but rather a qualitative one; the closeness of routers is better captured by questions like "Are two routers in the same group?" than ones like "What is the difference between two router ids?". As a result, we consider the spatial dimensions by partitioning routers into groups and tiles into colors.

Of these two variables, the effects of tile colors were by far the strongest. As shown in Figure 1(a), the mean stall to flit ratio of different colors of tiles were sometimes an



(a) (b) (c)

Fig. 1. (a) The average stall to flit ratio by tile color for two weeks of LDMS data and the PDF of that ratio on grey tiles for 5-minute increments for about half an hour on either side of (b) the start of job 416088 at 01:21 on the 16th and (c) the end of the job at 00:59 on the 17th. Gaps are due to missing data.

order of magnitude apart. This is not to say that the colors behaved entirely differently, as the three colors of network tile generally trended up and down together. The mean stall and flit values of each color of network tile on a router were highly correlated, particularly in log space where grey and green flits and stalls had an R^2 value near one. Conversely, there was little variation between different groups on the system. At any given time, the mean ratio of tiles of a certain color was generally about the same across the system. Only the grey average showed any noticeable, albeit still quite small, degree of variation. Interestingly, this appears to imply that the level of congestion is roughly uniform across Theta. This contradicts the assumptions of analysis techniques based on congestion regions, which have proven useful on toroidal systems [3].

B. Temporal Distribution

Unsurprisingly, we saw a large degree of variation over time in the data. The length of any given period of high congestion varied wildly, as can be seen in Figure 1(a). Over the two weeks of data, we observed three separate periods of consistent high congestion which lasted for at least a day.

Since it is difficult to generalize about the behavior of the system congestion over time, we present an investigation into one sustained period of congestion as a case study.

1) Case Study: We started by looking through the ratio distribution plots for each color on each day for any hour whose distribution visibly diverged from the others of the same color on the same day. One anomalous distribution we found consisted of the grey data from around midnight on the 17th. In order to investigate this period of time, we zoomed in on it by calculating the stall, flit, and ratio statistics for all data from about 23:00 on the 16th through 1:00 on the 17th[1]. We then recreated the distribution plots for each time in the fine-grained data to verify that the data was similarly distributed at this resolution. Once we had confirmed this, we examined the job logs to identify all jobs running in this period and plotted the number of nodes used by each job. We found one such job, with an id of 416088, which used over 2000 nodes, about 500 more than the next largest job. We then used the job logs to identify when this job started and finished, and then found the LDMS data from immediately before and after both times. This data, aggregated in 5-minute bins, shows how the distribution of the ratios shifts towards higher values after job 416088 begins and back towards smaller values after the job is over, as seen in Figure 1(b-c). The shift after the job ends takes roughly 20 minutes to occur, whereas the one after the job starts is almost immediate. After comparing the start and end times of the job to the distribution of the ratio over time, seen in Figure 1(a), we noticed that the duration of the job matched a day-long congestion period we had noticed earlier.

[1]In order to sidestep the binning process mentioned earlier, we placed each time in its own bin and calculated the stall and flit rates accordingly.

IV. Conclusion

Our work revealed several patterns in how congestion is distributed spatially on Theta. Most congestion is focused on the network tiles, with the greatest intensity occurring on grey (intra-group column) links. We did not observe a similar concentration of congestion on any subset of the groups on the system, which implies that the entire network of routers experiences similar levels of congestion at any given time.

As for the temporal distribution of congestion, we found that there was a high degree of variability in the system state. Some periods of high congestion lasted on the order of days, while others were very short-lived. In at least one day-long period of congestion, the system stayed congested for some time after the appearent cause (a large job) was gone.

Acknowledgment

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. This work was supported by funding provided by the University of Maryland College Park Foundation.

REFERENCES

[1] A. Agelastos, B. Allan, J. Brandt *et al.*, "The lightweight distributed metric service: a scalable infrastructure for continuous monitoring of large scale computing systems and applications," in *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2014, pp. 154–165.

[2] J. M. Brandt, E. Froese, A. C. Gentile *et al.*, "Network performance counter monitoring and analysis on the cray xc platform." Sandia National Lab.(SNL-CA), Livermore, CA (United States); Sandia National ..., Tech. Rep., 2016.

[3] S. Jha, A. Patke, J. Brandt *et al.*, "Measuring congestion in high-performance datacenter interconnects," in *17th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 20)*, 2020, pp. 37–57.