Topologies and Adaptive Routing on large-Scale Interconnects

Shafayat Rahman
Department of Computer Science, Florida State University
Tallahassee, Florida 32306
rahman@cs.fsu.edu

ABSTRACT

The performance of the interconnect network is massively important in the modern day supercomputer and data centers. As a PhD student in the FSU CS EXPLORER (EXtreme-scale comPuting, modeLing, netwORking & systEms Research) lab under the supervision of Dr. Xin Yuan, my research activity revolves around the analysis, improvement and performance evaluation of a number of topology and routing schemes widely used in the field of high performance computing. Over the years, I worked in a number of projects that is briefly described over the next few paragraphs.

1 LOAD-BALANCED SLIM-FLY NETWORKS

The Slim Fly topology [2] has been proposed recently for future generation supercomputers. In this project, we investigated how the traffic is expected to disperse among the network, and discovered that in Slim Fly certain links are more likely to carry traffic than the rest of the links for both minimal and non-minimal routing. As a result, hot-spots are more likely to form in these links. To mitigate the issue, we came up with two different approaches. The first approach, which we call the *bandwidth-provisioning* scheme, is to modify the topology and increase the bandwidth of the overused links in such a rate so that the original load-imbalance goes away. For a given Slim Fly topology, we identify the links that needs to get the added capacity, and the amount of extra bandwidth needed. This approach eliminates the load-imbalance completely, but comes with implementation issues. The second approach, that we call the *Weighted non-minimal routing* scheme, is to modify the non-minimal routing to distribute the traffic in a more load-balanced fashion. Essentially, we assigned some weights to the non-minimal nodes to counter-balance the inherent routing bias that comes with the topology. We presented two strategies to tune the necessary weights to implement this approach. We validated our results with detailed analysis and simulation, and demonstrated that both the approaches result in a more effective Slim Fly network that its present form.

2 DRAGONFLY DESIGN SPACE: LINK ARRANGEMENT AND PATH DIVERSITY

Dragonfly [9] is a popular cost-effective interconnect design which has been used in a number of current and future generation supercomputers. In this topology, the nodes are grouped together into clusters, and the clusters are connected to each other to form a diameter-three network. The original Dragonfly paper left a few issues open, like how the nodes inside the groups will be connected, to which router and global channel should be connected to, or what happens when the total number of groups in the system is less than the maximum it can support. There has been efforts to answer these questions [5], and there are systems [4] that were built with deviations from the original Dragonfly design, but there are still a lack of formal answers of the open questions. In this project, we investigate the design space of Dragonfly. Our objective is two-fold. First, we want to figure out how the inter-group connections should be implemented to get optimum performance from the network, specially when the network contains less than maximum number of groups. Second, we analyze the minimal path distribution within the entire design space, and investigate the performance of the existing routing algorithms. This project is currently on-going and we are using a number of modeling, simulation and learning techniques to attain our objectives.

3 TRAFFIC-PATTERN BASED ADAPTIVE ROUTING IN DRAGONFLY NETWORKS

In Dragonfly topology, routing is done through minimal and non-minimal paths, and adaptive routing [13] is used to toggle between them. Traditionally, adaptive routing algorithms use locally available information, for example buffer queue length, to detect congestion scenario in the network. This poses a challenge for Dragonfly as the inter-group nodes may not have the most accurate information about congestion in the gateway routers. There have been a number of research [14] [8] on how to convey the congestion information efficiently to the local routers. For our project, we investigated the performance of adaptive routing in the Dragonfly network used in Cray Cascade [4] and presented a novel scheme to improve over it. The idea is to gather link usage statistics through performance counters within a certain history window, and use that information to infer the traffic pattern: benign or adversarial. Then the final routing decision is taken using both the traffic pattern and link load information. We performed simulation using a number of traffic patterns and demonstrated that our scheme achieves lower latency for benign traffic and higher throughput for adversarial traffic over the existing UGAL adaptive routing scheme. The first phase of the project only considered the intra-group communications. For the second phase, we are currently working to expand in to inter-group communications as well.

4 PERFORMANCE MODELING STUDIES

I worked in a number of projects over the years which analyzed and devised scalable modeling methods to evaluate various topologies, routings and performance metrics. In one project, we modeled

ICPP'18, August 2018, Eugene, OR, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/n

the UGAL[13] routing over Dragonfly topology to get a better theoretical understanding on how the routing works. UGAL is extensively used on various systems to implement adaptive routing, but there had been no theoretical understanding on its effectiveness; all the prior studies were simulation or experiment based. In our project, we modeled UGAL's performance with varying level of controls over the minimal and non-minimal paths, and compared the modeling results with simulation output. The verdict was that individual-level control on every path generally overestimates the system's output, full random control on all paths (basically treating each path as same) generally underestimates the output, and controlling the same-length paths as a group goes within 10% of simulation output. So this gives a good estimation of system output under UGAL routing while is also computationally feasible.

In another project, we evaluated a number of commonly-used throughput models and identified similar and contradictory trends in their performance. The models we studied were max-min fairness (MMF) [1], maximum concurrent flow (MCF) [12], Hoefler's method (HM) [6] and Jain's method (JM) [7]. We showed that even though the later three models approximate MMF, they have subtle differences and may even produce contradictory results depending on the topology and traffic pattern.

Finally, I worked in another project that studied the performance characteristics of a number of topologies that provide either low diameter (Dragonfly and Slim Fly), or high path diversity (fat-tree [11], Random Regular graph [10], and Generalized De Bruijn Graph [3]). We investigated the performance of some HPC applications paired with some routing schemes on each of the topologies. We observed that the adaptive routing techniques developed for low-diameter topologies are effective on high path-diversity topologies as well. Also, we determined that high path-diversity topologies generally perform better than equivalent-sized low diameter topologies.

I started the program in FSU in 2011. I spent a number of years taking various core and elective courses, going through the PhD comprehensive exam, and exploring a number of research areas. I started working with Dr. Yuan from 2014. I acquired the background knowledge necessary to excel in the field of HPC, completed my area exam and worked in a summer internship at Oak Ridge National Lab. I am on course to wrap up the current projects and complete my degree by the end of 2018.

ACM Reference Format:
Shafayat Rahman. 2018. Topologies and Adaptive Routing on large-Scale Interconnects. In *Proceedings of ACM Conference (ICPP'18).* Eugene, OR, USA , 2 pages. https://doi.org/10.1145/n

REFERENCES

[1] Dimitri P Bertsekas, Robert G Gallager, and Pierre Humblet. 1992. *Data networks.* Vol. 2. Prentice-Hall International New Jersey.
[2] Maciej Besta and Torsten Hoefler. 2014. Slim fly: a cost effective low-diameter network topology. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.* IEEE Press, 348–359.
[3] Ding-Zhu Du and Frank K Hwang. 1988. Generalized de Bruijn digraphs. *Networks* 18, 1 (1988), 27–38.
[4] Greg Faanes, Abdulla Bataineh, Duncan Roweth, Edwin Froese, Bob Alverson, Tim Johnson, Joe Kopnick, Mike Higgins, James Reinhard, et al. 2012. 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.* IEEE Computer Society Press, 103.
[5] Emily Hastings, David Rincon-Cruz, Marc Spehlmann, Sofia Meyers, Anda Xu, David P Bunde, and Vitus J Leung. 2015. Comparing global link arrangements for Dragonfly networks. In *Cluster Computing (CLUSTER), 2015 IEEE International Conference on.* IEEE, 361–370.
[6] Torsten Hoefler, Timo Schneider, and Andrew Lumsdaine. 2009. Optimized routing for large-scale InfiniBand networks. In *High Performance Interconnects, 2009. HOTI 2009. 17th IEEE Symposium on.* IEEE, 103–111.
[7] Nikhil Jain, Abhinav Bhatele, Xiang Ni, Nicholas J Wright, and Laxmikant V Kale. 2014. Maximizing throughput on a dragonfly network. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.* IEEE Press, 336–347.
[8] Nan Jiang, John Kim, and William J. Dally. 2009. Indirect Adaptive Routing on Large Scale Interconnection Networks. *SIGARCH Comput. Archit. News* 37, 3 (June 2009), 220–231. https://doi.org/10.1145/1555815.1555783
[9] John Kim, Wiliam J Dally, Steve Scott, and Dennis Abts. 2008. Technology-Driven, Highly-Scalable Dragonfly Topology. In *ACM SIGARCH Computer Architecture News*, Vol. 36. IEEE Computer Society, 77–88.
[10] Michihiro Koibuchi, Hiroki Matsutani, Hideharu Amano, D Frank Hsu, and Henri Casanova. 2012. A case for random shortcut topologies for HPC interconnects. In *Computer Architecture (ISCA), 2012 39th Annual International Symposium on.* IEEE, 177–188.
[11] Charles E Leiserson. 1985. Fat-trees: universal networks for hardware-efficient supercomputing. *IEEE transactions on Computers* 100, 10 (1985), 892–901.
[12] Farhad Shahrokhi and David W Matula. 1990. The maximum concurrent flow problem. *Journal of the ACM (JACM)* 37, 2 (1990), 318–334.
[13] Arjun Singh. 2005. *Load-balanced routing in interconnection networks.* Ph.D. Dissertation. Stanford University.
[14] J. Won, G. Kim, J. Kim, T. Jiang, M. Parker, and S. Scott. 2015. Overcoming far-end congestion in large-scale networks. In *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on.* 415–427. https://doi.org/10.1109/HPCA.2015.7056051