Modeling and Comparison of Large-Scale Interconnect Designs

Md Atiqul Mollah

August 8, 2017

Abstract

Modern day high performance computing (HPC) clusters and data centers require a large number of computing and storage elements to be interconnected. Interconnect performance is one of the major contributing factors to the overall performance of such systems. Due to the massive scale of the network, interconnect designs are often evaluated and compared through models. My doctoral research is focused on modeling large-scale interconnections and applying such models to investigate the effect of different topology designs and routing scheme designs on the interconnect performance.

1 Introduction

In massively scaled HPC and data centers, the performance of the interconnect is a major potential bottleneck to the performance of the entire systems. Therefore, in order to design new HPC clusters/data centers or manage current systems efficiently, it is key to study and understand the concepts associated with interconnect performance. The performance of an interconnect primarily depends on the network topology, routing scheme, congestion control scheme and the traffic communication pattern. while some interconnect features may be measured and compared statically, a detailed analysis and comparison of interconnect designs typically involve simulation and modeling of the interconnect with practical traffic workloads. As the scale of the interconnection grows, the traditionally known ways of modeling the network behavior become computationally infeasible, making it impossible to profile and compare different interconnect designs at scale. Therefore, new modeling techniques are required to efficiently and accurately estimate the performances of current- and future-generation interconnects with thousands of switches and end-points.

My doctoral research is focused on devising scalable modeling methods to evaluate the performance characteristics of different large-scale interconnect designs. Developing such modeling methods have several benefits. Firstly, topology-specific models can be applied efficiently to determine how different routing schemes perform on a given topology design. Secondly, modeling allow us to effectively compare the performances of different interconnect designs under similar traffic conditions. Furthermore, interconnect models are often formulated as optimization problems, which can be used to study the performance bounds of different interconnect designs.

2 Research Highlights

In the following, I briefly describe the topics of my current and completed research. For further elaboration on the published works, the readers are encouraged to explore the citations appended to this article.

2.1 Fast calculation of max-min fair throughput in fat-tree networks[1]

Fat-tree based designs are widely popular among current HPC and data center interconnect fabrics. It consists of a large variety of hierarchical topologies and many supercomputing clusters, including the Tianhe-2, the No. 2 in the most recent TOP500 list of the world's fastest supercomputers[2], are interconnected using this topology.

Max-min fairness is often used in the performance modeling of interconnection networks. Existing methods to compute max-min fair rates for multi-commodity flows have high complexity and are computationally infeasible for large networks. In this research work, We show that by considering topological features, this problem can be solved efficiently for the fat-tree topology. We develop several efficient new algorithms for this problem. Using these algorithms, we demonstrate that it is possible to find the max-min fair rate allocation for multi-commodity flows in fat-tree networks that support tens of thousands of nodes. We evaluate the run-time performance of the proposed algorithms and show improvement in orders of magnitude over the previously best known method. Furthermore, demonstrate a new application of max-min fair rate allocation that is only computationally feasible using our new algorithms. We apply the rate allocation application to benchmark the performance of the popular destination-mod-k routing algorithm in large-scale fat trees.

2.2 Modeling UGAL routing for Dragonfly topology[3]

The Dragonfly topology has been proposed and deployed as the interconnection network topology for next generation supercomputers[4]. Practical routing algorithms developed for Dragonfly are based on a routing scheme called Universal Globally Adaptive Load-balanced routing with Global information (UGAL-G). While UGAL-G and UGAL-based practical routing schemes have been extensively studied, all existing results are based on simulation or measurement. There is no theoretical understanding of how the UGAL-based routing schemes achieve their performance on a particular network configuration as well as what the routing schemes optimize for.

In this work, We develop linear programming (LP) based throughput models for UGAL-G on the Dragonfly topology, and perform validation on those models. We identify a robust model that is both accurate and efficient across many Dragonfly variations. Given a traffic pattern, the proposed models estimate the aggregate throughput for the pattern accurately and effectively. Our contribution is two-fold. First, the throughput models that we have developed, can be used to accurately and efficiently predict the aggregate throughput for large scale Dragonfly networks. Second, the models reveal the implicit rate allocation in UGAL-G, which further our understanding of UGAL-based routing schemes.

2.3 Achieving adaptive routing performance on a non-adaptive SDN environment[5]

Recent years have seen emergence of Software Defined Networking (SDN) [6], a new networking paradigm based on logically centralized network control. SDN has shown promising performance in data centers networks, and there is a significant interest in the HPC community to adopt the SDN technology in HPC systems.

While current SDN technology have the potential to introduce a number of promising features to enhance HPC performance, it does not support adaptive routing, a pivotal scheme required in current HPC designs.This gives rise to the question whether SDN can achieve the performance that HPC systems expect with adaptive routing. In this work, we consider the current generation HPC interconnects with the Dragonfly topology and investigate possible methods to apply the SDN technology on such interconnects. We apply the rate control/allocation on the *elephant flows* in an SDN-enabled network with

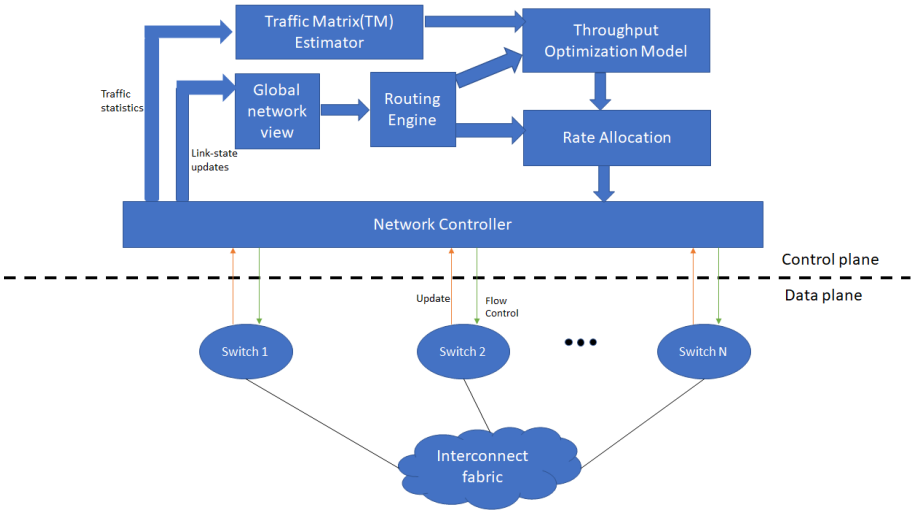

Figure 1: SDN based rate allocation scheme

an objective to maximize network throughput through load balancing. Then we compare the performance of SDN-style routing with that of a typical adaptive routing currently used in Dragonfly based systems.

Our research indicate that both SDN routing and adaptive routing have their own strengths: adaptive routing is an effective technique for HPC workloads and often results in higher performance than SDN.However, using the global network view, it is possible for SDN to compete with adaptive routing by allocating network resources more effectively.

2.4 Random Regular Graph and Generalized De Bruijn Graph with k-shortest Path Routing[7]

Random regular graphs (RRG) have considered for interconnecting future large scale data centers and HPC clusters. RRG is a special case of directed regular graph (DRG) where each link is unidirectional and all nodes have the same number of incoming and outgoing links. In this work, we establish bounds for DRG on diameter, average k-shortest path length, and a load balancing property with k-shortest path routing, and use these bounds to evaluate RRG. Our results indicate that RRG with k-shortest path routing is not ideal in terms of diameter and load balancing. We further consider the Generalized De Bruijn Graph (GDBG), a deterministic DRG, and prove that for most network configurations, GDBG is near optimal in terms of diameter, average k-shortest path length, and load balancing with a k-shortest path routing scheme. We further explore the strengths and weaknesses of RRG for different traffic conditions by comparing RRG with GDBG.

2.5 Study of Throughput performance Metrics[8]

In this study, we examine four commonly used interconnect throughput models and identify the cases when all models show similar trends, when different models yield different trends, and when different models produce contradictory results. Our study reveals important properties of the models and demonstrates the subtle differences among them, which are important for an interconnect designer to understand, in order to properly select a throughput model in the process of interconnect evaluation.

2.6 Study of Interconnect design approaches

A variety of approaches have been undertaken so far to design scalable high performance interconnects for next-generation HPC and data centers. Based on the topological features and the routing schemes, these current and prospective interconnect designs may be classified into two categories. The first category of interconnects is characterized by a low diameter that ensures low end-to-end latency on scale throughput the whole network. While a low diameter of these networks guarantees at least one short path among all end node pairs, the number of the diversity of short paths are not taken into consideration in this design approach. The prominent examples of such design approach are the Dragonfly and the Slimfly[9] topologies. On the other hand, The second category emphasizes on path diversity on a network rather than the diameter - to deliver high capacity among node pairs through several "short" paths. Examples of this category include the widely popular fat-tree network and its variants, random regular graphs and Generalized de Bruijn graphs. Our preliminary research show that each type of interconnect design have their advantages on certain traffic conditions. A more comprehensive study is currently underway.

References

[1] M. A. Mollah, X. Yuan, S. Pakin, and M. Lang, "Fast calculation of max-min fair rates for multi-commodity flows in fat-tree networks," in *2015 IEEE International Conference on Cluster Computing*, pp. 351–360, Sept 2015.

[2] E. Strohmaier, H. Simon, J. Dongarra, and M. Meuer, "Top500 supercomputing sites," June 2017.

[3] M. A. Mollah, P. Faizian, M. S. Rahman, X. Yuan, S. Pakin, and M. Lang, "Modeling ugal on the dragonfly topology," *Unpublished*, 2017.

[4] J. Kim, W. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Computer Architecture, 2008. ISCA '08. 35th International Symposium on*, pp. 77–88, June 2008.

[5] P. Faizian, M. A. Mollah, Z. Tong, X. Yuan, , and M. Lang, "A comparative study of sdn and adaptive routing on dragonfly networks," *accepted at SC17: International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov 2017.

[6] O. N. Foundation, "Sdn architecture." White Paper, ONF TR-521, February 2016. available at https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR-521_SDN_Architecture_issue_1.1.pdf.

[7] P. Faizian, M. A. Mollah, X. Yuan, S. Pakin, and M. Lang, "Random regular graph and generalized de bruijn graph with k-shortest path routing," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 103–112, May 2016.

[8] P. Faizian, M. A. Mollah, M. S. Rahman, X. Yuan, S. Pakin, and M. Lang, "Throughput models of interconnection networks: the good, the bad, and the ugly," *accepted at the 25th IEEE Annual Symposium on High Performance Interconnects (HotI)*, 2017.

[9] M. Besta and T. Hoefler, "Slim fly: A cost effective low-diameter network topology," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, (Piscataway, NJ, USA), pp. 348–359, IEEE Press, 2014.