Exploring Dynamic Multipath Routing in 3D Torus Networks through Simulations

Daniel Parker
Illinois Institute of Technology/
University of Chicago
Departments of Computer Science
dkparker@uchicago.edu

Sanjiv Kapoor
Illinois Institute of Technology
Department of Computer Science
kapoor@iit.edu

Ioan Raicu
Illinois Institute of Technology
Department of Computer Science
iraicu@cs.iit.edu

1. ABSTRACT

Torus Networks are an important and widely used architecture in modern supercomputers. However, these supercomputers frequently employ static routing protocols, which have poor worst case performance under heavy network traffic. In this paper, we first characterize bottlenecking behavior under the current standard protocol, then demonstrate that even simplistic dynamic routing protocols offer lower, more scalable latencies over a 3D torus network. As supercomputing moves towards exascale, such scalability will be necessary to cope with unprecedented workloads. These results were generated using the CODES highly parallel simulator, built on top of the ROSS parallel discrete event simulator.

2. INTRODUCTION

Modern supercomputers have high inter-node networking requirements. This means that supercomputer architectures must allow for fast inter-node communication. One answer to this need has been the torus network architecture, used in multiple modern supercomputers. Torus networks have the advantages of fast nearest-neighbor communications and diversity of paths between any two given nodes within the network.

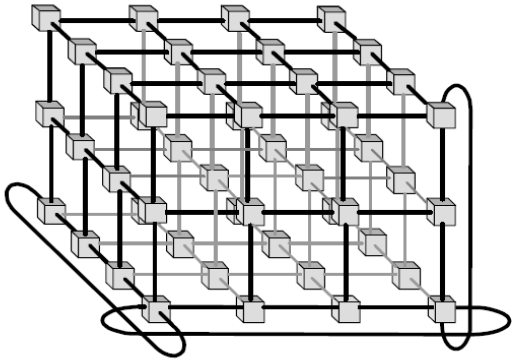


Figure 1. A 3D Torus network interconnect. Notice how the edges wrap around in each dimension.

However, modern routing protocols over torus networks do not take full advantage of this path diversity. Many modern systems employ Dimension Order Routing (DOR), which routes packets across the network using a static, predetermined path. Specifically, packets will traverse the network by going as far as necessary in one direction before beginning to move in the next. We theorize that this may lead to packets passing through areas of excessive traffic when more optimal paths exist. This would lead to some packets experiencing excessive, unnecessary latencies over the network. We propose that dynamic and multi-path routing schemes can mitigate this effect.

Previous work in routing spans network scales. Investigations by Cho et al demonstrate improvement over DOR using similar routing schemes to those in this paper, but across Networks on a Chip [4]. This work is the introduction of such methods to the context of supercomputing. Another similar routing scheme was proposed by Valiant, the difference being that his algorithm routes packets to distinct random locations before sending them to the destination, whereas ours routes packets randomly at each step [5]. Wide Area Networks have also been making use of complex dynamic multipath routing, as discussed in Devetak et al and Shin et al [2][3], but the context of supercomputing imposes unique challenges. Any implementation of dynamic multipath routing for use in supercomputing must introduce significant latency improvement, while being simple enough to avoid excessive overhead, and must be sufficiently robust to scale well to large network sizes and heavy network traffic.

In this work, we first demonstrate the need for such a protocol, by examining the poor latency scaling using the current DOR method. We then give an example of the improvements that dynamic, multipath routing offers through a simple implementation, called Randomized Dynamic Routing. We discuss the improvements in scalability and average latency that this scheme offers.

These experiments are carried out over 3D torus networks simulated using CODES, a highly parallel simulator designed for use in supercomputer co-design at extreme scales [1]. First, the routing protocols used in real supercomputers are often difficult to impossible to modify by the end user. This complicates the study of alternate routing methods. Supercomputing time is also expensive and valuable. Preliminary evidence of the effectiveness of dynamic multipath routing may justify overcoming the difficulties of further investigation using either existing or novel systems.

3. PROCEDURE

Experiments were carried out on a 3D 10x10x10 torus network simulation. A certain number of nodes within the network were randomly designated as "off" and received no work; the rest were considered "on," and were evenly divided into senders and receivers, then randomly paired in a 1 to 1 mapping. The sender nodes then generated 8 kilobyte ping messages and sent them through the network to the corresponding receiver node, which sent back an acknowledgement pong message back to the sender. The time from the transmission of the ping to the receipt of the pong was timed as latency. The sender nodes would repeat this a total of ten times and record the average latency experienced over the course of the experiment. Such experiments were carried out with the number of "on" nodes ranging from 100 to 1000.

Experiments measuring bandwidth were calculated similarly, the difference being that the sender nodes would send 10 8 kilobyte messages before receiving an acknowledgement message. This time was again taken as latency, then the total amount of data transmitted (80 kilobytes) was divided by this latency to measure bandwidth.

Following this experimentation on a DOR network, comparable experiments were run using Randomized Dynamic Routing. The procedure of this routing method is as follows.

RDR:

- If at destination: DONE
- Else, for each dimension, determine whether it would be advantageous to travel in the positive direction, negative direction, or not move, using the half length of the network. If travel over that dimension is viable, mark that dimension VIABLE.
- Select a dimension marked VIABLE uniformly at random, travel one node in the appropriate direction, then call RDR again.

Essentially, a packet randomly selects a dimension of travel at each step, rather than proceeding as far as possible in each dimension sequentially. This protocol was chosen for its simplicity as well as its potential to avoid network crowding, by means of random distribution of traffic.

4. RESULTS

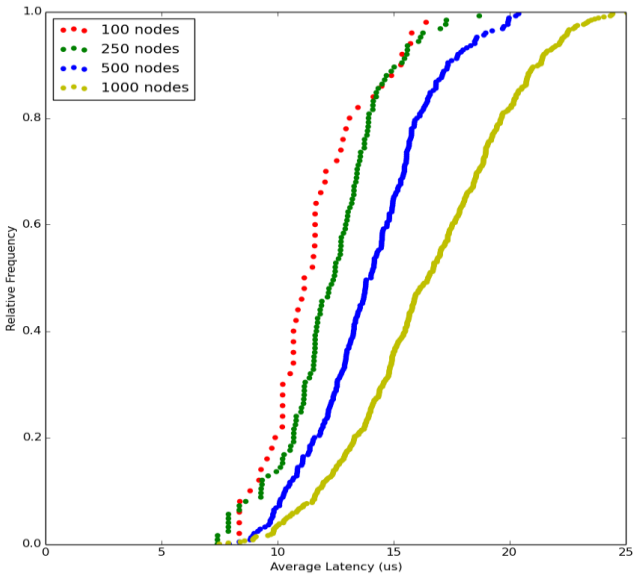


Figure 2. CDF plots of average latencies over 3D Torus Networks using the DOR protocol with 100, 250, 500, and 1000 active nodes.

The key observation from these first trials is that as the network becomes busier, with more nodes in communication, latencies increase significantly. This indicates that the latencies on the busier network are affected by bottlenecking. We theorized that this effect would be ameliorated with alternate routing methods.

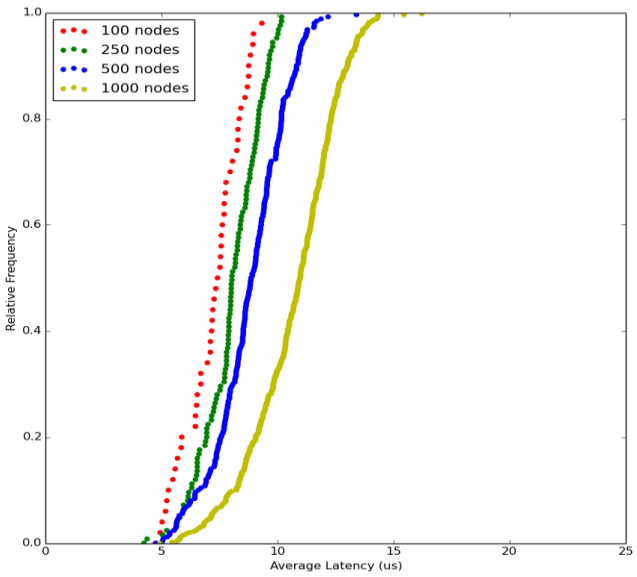


Figure 3. CDF plots of average latencies over 3D Torus Networks using the RDR protocol with 100, 250, 500, and 1000 active nodes

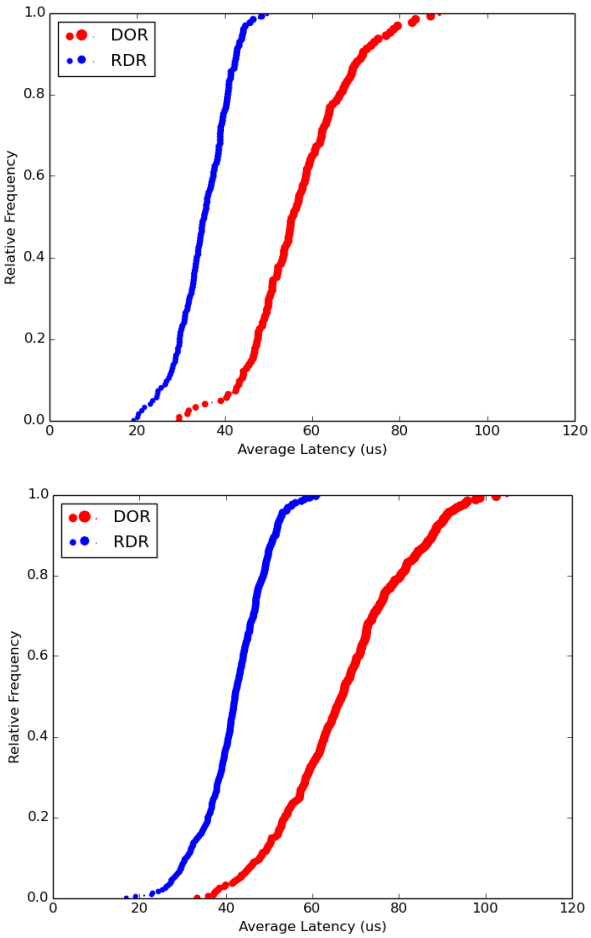
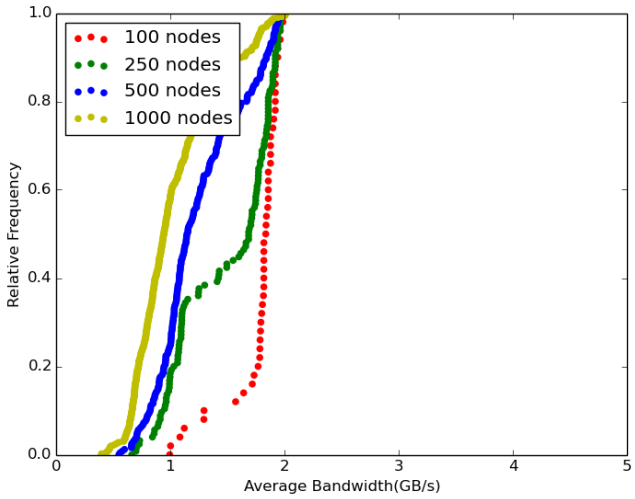


Figure 4. Comparison of latency CDF plots; first at 500, then 1000 active nodes.

From the comparisons, we observe that as network traffic increases, the RDR latencies have a smaller range and steeper slope, an effect that becomes more apparent under higher traffic. Thus we see that less nodes experience excessive latencies under an RDR scheme, which indicates that dynamic routing lessens the effects of high network traffic.



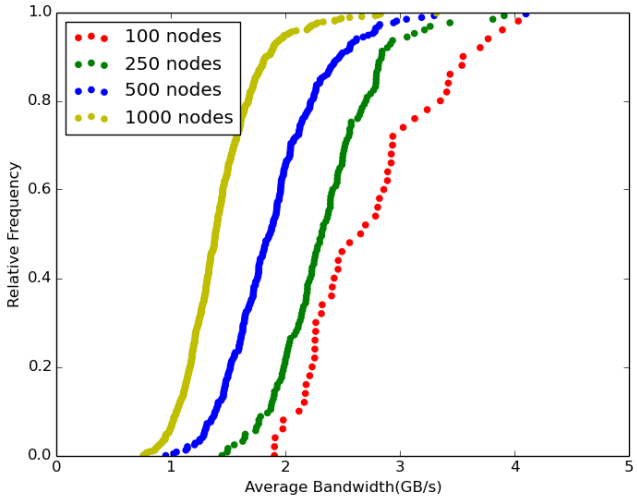


Figure 5. CDF plots of first DOR, then RDR routing. Notice how the RDR bandwidths exceed the per link bandwidth limit of 2 GB/s.

We observe during bandwidth trials that the network using DOR is unable to exceed the maximum link bandwidth of 2 GB/s. However, under RDR, the network can vastly exceed this amount, at up to 4 GB/s. This demonstrates the efficacy of multipath routing at increasing network bandwidth.

5. CONCLUSION

Our experiments with dynamic multipath routing provide solid evidence for the potential of such routing schemes to significantly improve latency in modern supercomputer topologies. As supercomputing continues to grow into extreme scales, every possible optimization will be needed to achieve optimal performance.

There are many opportunities for further work in this area. Experimentation has indicated that CODES does not currently take into account the overhead required for sending small messages. In order to more accurately simulate real systems, we would need to implement this functionality into CODES.

In this vein, we also wish to run validation experiments on a real system like the Blue Gene-Q at Argonne National Laboratory. That system uses a 5D Torus interconnect, which may require rerunning these experiments on such a network. Validation would make clear the correspondence between our simulation results and real systems, providing further evidence of the viability of dynamic multipath routing on multi-node systems. Further, experimentation over other network topologies like Dragonfly networks could give evidence of dynamic multipath routing's viability on a greater set of large-scale systems.

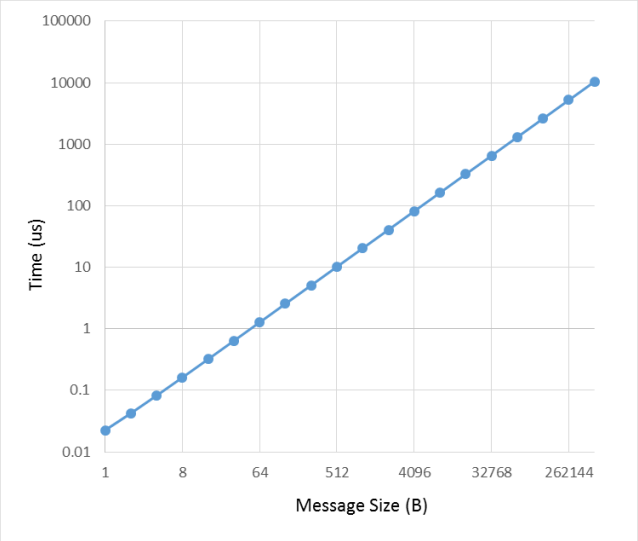


Figure 6. Latency plotted against message size for nearest neighbor communication. Latency scales linearly with message size even at very small message sizes, indicating that baseline message transmission time is not accurately considered.

More work needs to be done with regard to testing maximum possible throughput over our simulation network. Theoretically, inter-node communications should attain maximum bandwidths of 6 GB/s, so identifying a routing protocol that can get closer to this upper bound than our current maximum of 4 GB/s would give further evidence of dynamic multipath routing's viability. As [4] describes multiple other oblivious routing protocols of suitable complexity, experimenting with more of these could lead to even better results.

6. REFERENCES

[1] Cope, Jason, Ning Liu, Sam Lang, Phil Carns, Chris Carothers, and Robert Ross. "Codes: Enabling co-design of multilayer exascale storage architectures." In *Proceedings of the Workshop on Emerging Supercomputing Technologies*, pp. 303-312. 2011.

[2] Devetak, F. et al. 2011. Minimizing Path Delay in Multipath Networks. *2011 IEEE International Conference on Communications (ICC)*. (2011).

[3] [2]Junghwan Shin, et al. 2013. Concurrent multipath routing over bounded paths: Minimizing delay variance. *2013 IEEE Global Communications Conference (GLOBECOM)*. (2013).

[4] Myong Hyon Cho, Mieszko Lis, Keun Sup Shim, Michel Kinsy, and Srinivas Devadas. 2009. Path-based, randomized, oblivious, minimal routing. In *Proceedings of the 2nd International Workshop on Network on Chip Architectures* (NoCArc '09). ACM, New York, NY, USA, 23-28.

[5] Valiant, Leslie G. "A scheme for fast parallel communication." *SIAM journal on computing* 11.2 (1982): 350-361.