Understanding Torus Network Performance through Simulations

Sandeep Palur
Dept. of Computer Science
Illinois Institute of Technology
psandeep@hawk.iit.edu

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

Abstract

A number of supercomputers on the TOP500 list use 3D Torus networks[1] (e.g. IBM's BlueGene/L and BlueGene/P, and the Cray XT3). We benchmark the Torus network through appropriate performance metrics under different workloads using the ROSS(Rensselaer's Optimistic Simulation System)simulator. We have studied the communication imbalance generated by the common static single path routing in Torus interconnects

Motivation

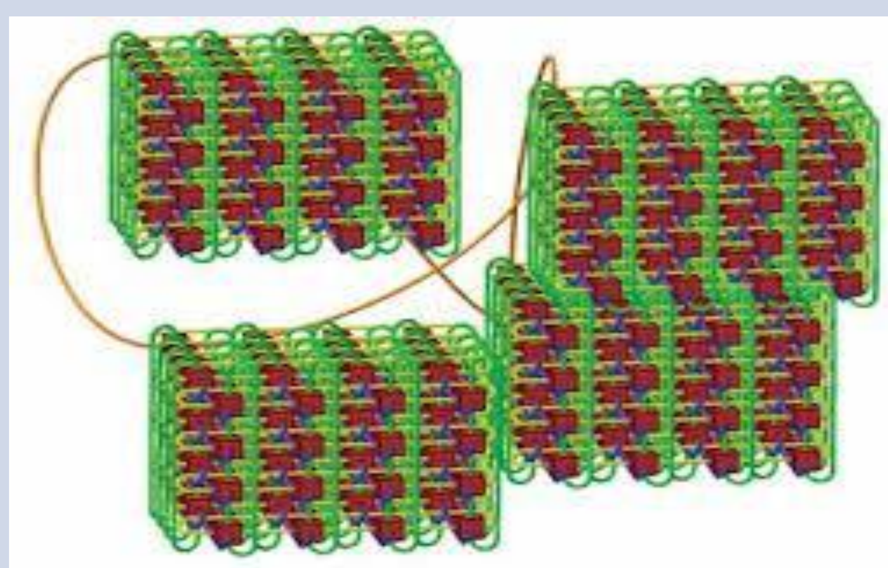
- Technology developments in the storage and processing of data spurred the development of distributed computing with distributed compute-clusters and supercomputers processing massive data
- A Torus interconnect is a network topology for connecting processing nodes in a parallel computer system

Torus Interconnect

- Network topology for connecting processing nodes in a parallel computer system
- In 3D Torus interconnect, the nodes are imagined in a three dimensional lattice in the shape of a rectangular prism
- Each node 3D Torus interconnect is connected with its 6 neighbors, with corresponding nodes on opposing faces of the array connected and higher dimension add another pair of nearest neighbor connections to each node

Advantages

- High speed and low latency
- Linear scalability
- Switch-less configuration
- Avoids bottleneck
- Hardware cost reduction
- Less energy consumption
- Regular and hidden wiring
- Lower energy usage for communication



Torus Interconnect in Blue Gene - IBM

ROSS(Rensselaer's Optimistic Simulation System)

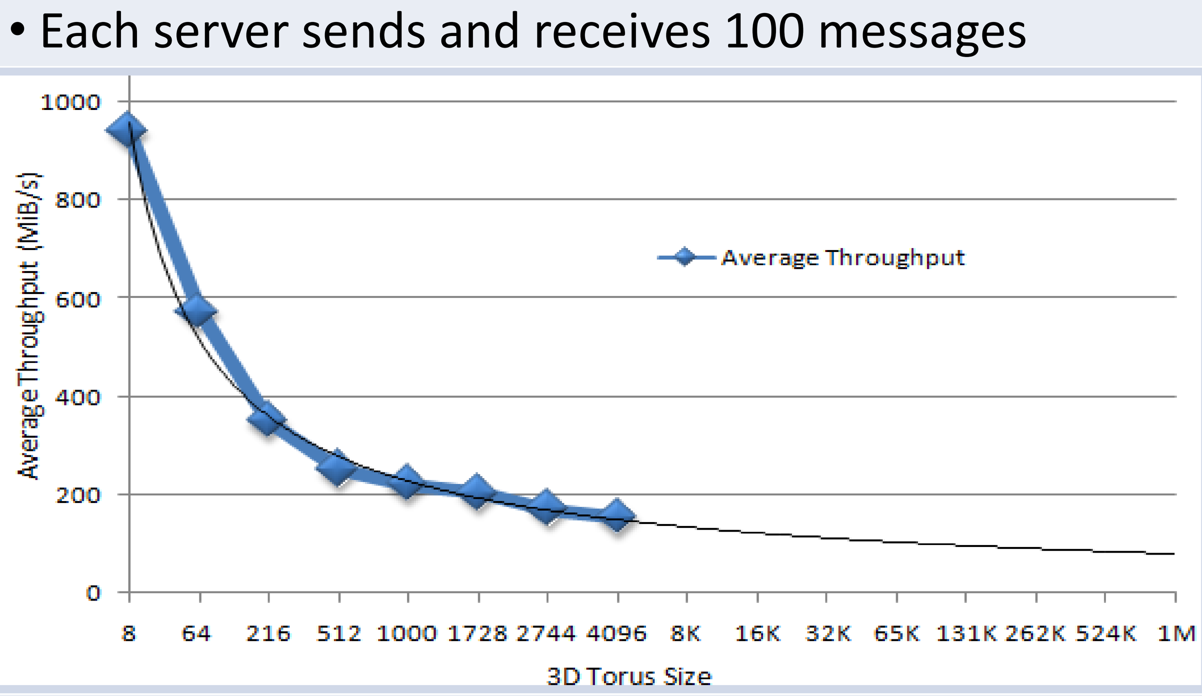
- Parallel discrete-event simulator that executes on shared-memory multiprocessor systems
- The synchronization mechanism is based on Time Warp[2,3,4]
- Collection of *logical processes,* each modeling a distinct component of the system being modeled
- Works in 3 modes: Sequential , Parallel Conservative and Parallel Optimistic

CODES

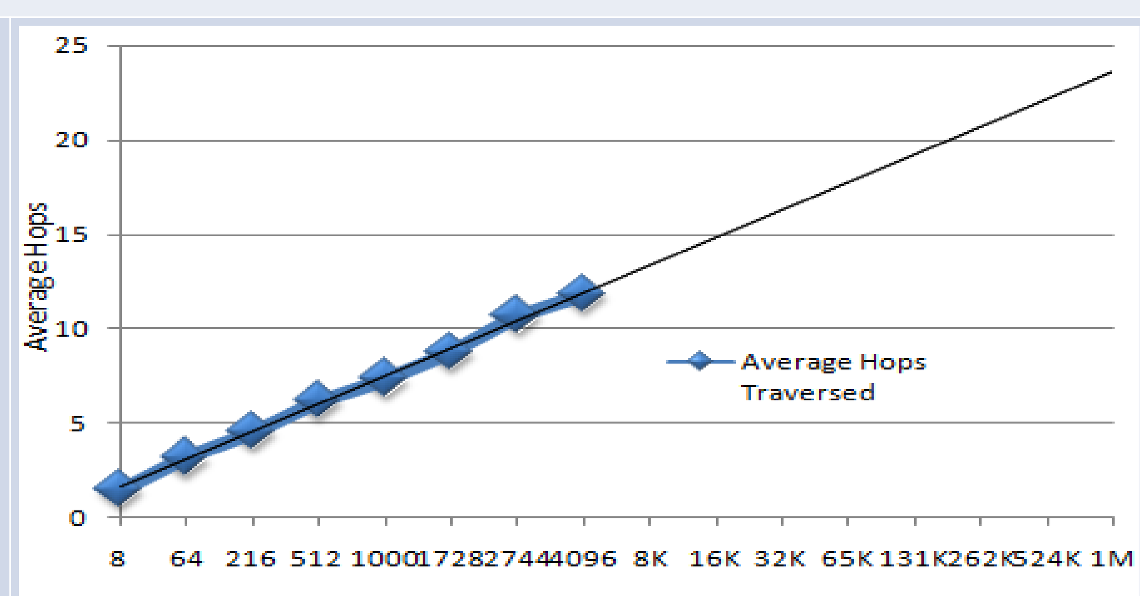
- Accurate and highly parallel simulation toolkit for exascale storage and is built on ROSS
- Divided into codes-base and codes-net.
- Codes-base is the utility library for construction of storage models
- Codes-net is collection of network interconnect models and shared abstraction layer.

Experiments & Results

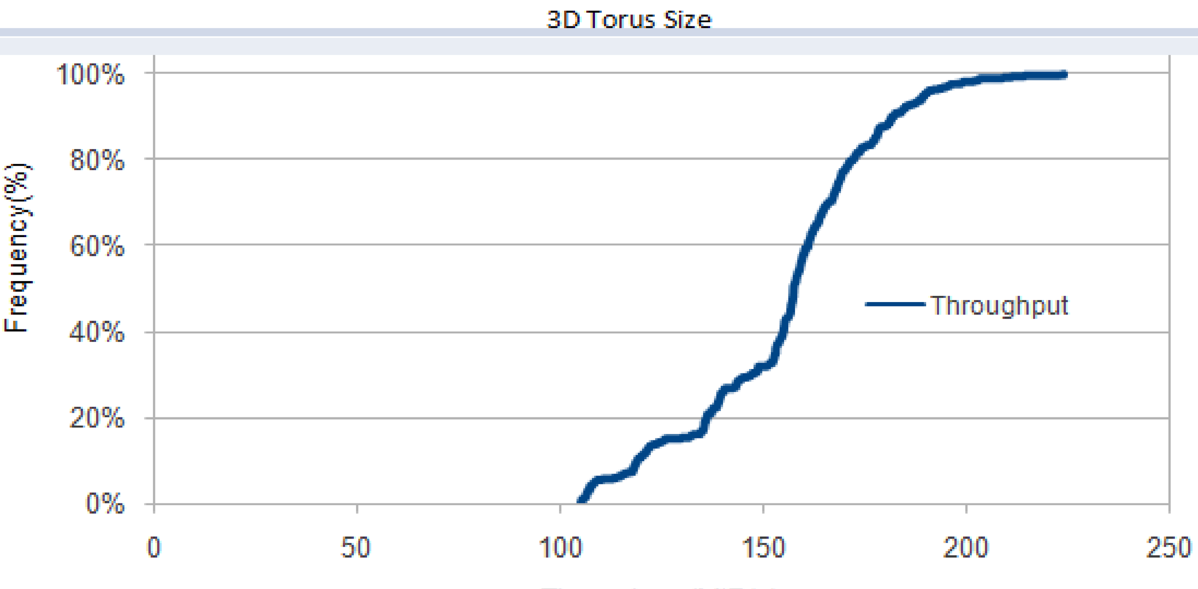
- We ran experiments on 48 cores 250 GB ram machine with x86_64 architecture
- Used ROSS simulator in parallel optimistic mode
- Each server in the torus network communicates with its own pair. Server pairs are generated by Fisher–Yates shuffle algorithm
- Each server sends and receives 100 messages



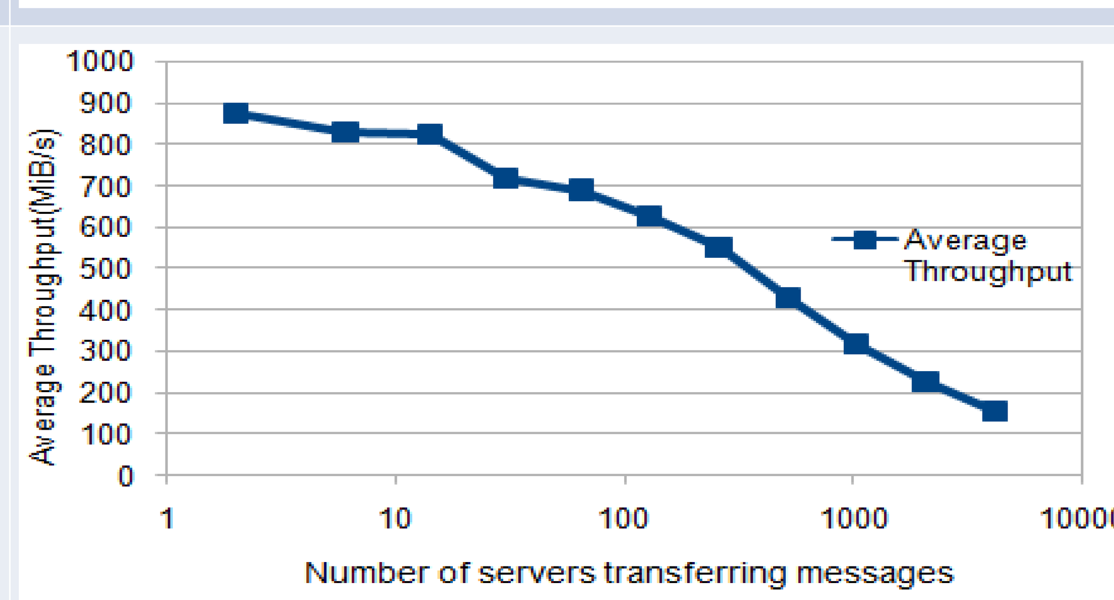
The average throughput decreases with increase in the size of the network



The average hops increases with increase in the size of the network



CDF plot shows there is a lot of difference between the throughput of each of 4096 nodes in the network



The average throughput decreases with increase in the Number of servers transferring messages

Conclusion

- Through synthetic benchmarks, we have studied the communication imbalance generated by the common static single path routing in Torus interconnects
- In torus network latency increases and throughput decreases as the size of the torus network and number of servers participating in message transfer increase
- Since torus uses static single path routing, transferring messages between random server pairs leads to a lot of congestion at some intermediate nodes via which most of the messages pass through.

Future Work

- Design and develop a monitoring framework to monitor the network state and indicate the hot spots.
- Demonstrate that multi-path dynamic routing could have positive impact on both the end-to-end application performance as well as the aggregate system wide performance.

References

[1] Narasimha R Adiga, Matthias A Blumrich, Dong Chen, Paul Coteus, Alan Gara, Mark E Giampapa, Philip Heidelberger, Sarabjeet Singh, Burkhard D Steinmacher-Burow, Todd Takken, et alBlue Gene/L torus interconnection network. IBM Journal of Research and Development.

[2] D. R. Jefferson and H. Sowizral. Fast concurrent simulation using the Time Warp mechanism,part I: Local control. Technical Report N-1906-AF, RAND Corporation, December 1982.

[3] D. R. Jefferson. Virtual time. ACM Transactions on Programming Languages andSystems, 7(3):404–425, July 1985.

[4] R. M. Fujimoto. Parallel discrete-event simulation. Communications of the ACM, 33(10):30–53,October 1990.