IS TOPOLOGY IMPORTANT AGAIN?

Effects of Contention on Message Latencies in Large Supercomputers

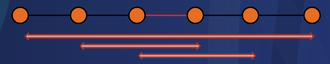
Abhinav Bhatelé
Laxmikant V. Kalé

Wormhole routing has historically minimized the impact of distance on message latencies. With machines having large diameters (such as the ANL Blue Gene/P and ORNL XT4) , these problems have remerged. For messages on these machines connected by a 3D torus or mesh interconnect, the first term in this equation is no longer insignificant:

$$(L_f/B)*D + L/B$$

More importantly, this equation models message latencies correctly only when links are not being shared between messages:

When multiple messages share links on the network, contention for resources reduces the effective bandwidth significantly:



This leads to a significant increase in message latencies depending on the number of links suffering from contention. To avoid this situation, one needs to map communicating objects topologically on physically nearby processors.

Much work was done on topology aware mapping in the 80s but:

Then	Now
Mainly directed towards theoretical object graphs on hypercubes, shuffle exchange and other theoretical networks	We are using object graphs from real applications on 3D torus/mesh topologies which are used on real machines
Most techniques were used offline and were slow	Our attempts are directed towards fast, runtime solutions
Demonstrated on graphs with tens or hundreds of nodes	We are developing scalable techniques for very large (petascale) machines
Number of nodes in the object and processor graph were the same	We handle the case of multiple objects per processor (load balancing issues)
Not tested with real applications on actual machines – theoretical work	Targeted at production codes and strategies being tested against real applications

This poster will demonstrate and quantify the effects of contention on message latencies for two large parallel machines: ANL's Blue Gene/P and PSC's XT3 (BigBen) through simple MPI benchmarks. It will also present improvements resulting from topology aware mapping for a production quantum chemistry application called OpenAtom.

References:

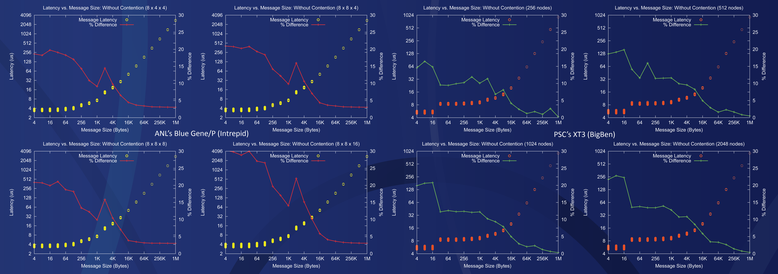
[1] Bhatele, A., Kale, L. V., Dynamic Topology Aware Load Balancing Algorithms for MD Applications, submitted to Phil. Trans. Roy. Soc. A, 2009
[2] Bhatele, A., Kale, L. V., Benefits of Topology-aware Mapping for Mesh Topologies, LSPP special issue of Parallel Processing Letters, 2008
[3] Bhatele, A., Kale, L. V., Application-specific Topology-aware Mapping for Three Dimensional Topologies, Proceedings of LSPP (held as part of IPDPS '08), 2008
[4] PhD Thesis Webpage: http://charm.cs.uiuc.edu/~bhatele/phd

E-mail: bhatele@illinois.edu

PARALLEL PROGRAMMING LAB
DEPT OF COMPUTER SCIENCE, UNIVERSITY OF ILLINOIS
PPL UIUC

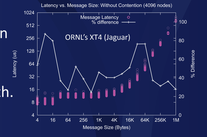
Benchmark I: No Contention

A master rank is chosen from all of the COMM_WORLD ranks. The master rank sends messages to all other ranks one at a time and receives a reply in return. The size of the message being sent is varied and for each message size, the time for a message send from the master rank to all other ranks is recorded.



Findings:

1. Message latencies depend on the distance (hops) traveled for small messages even in the absence of contention.
2. The effect is more pronounced for large torus sizes because of larger diameters.
3. It is observed both on IBM and Cray machines regardless of the available bandwidth.
4. Because of non-contiguous job allocation on XT4, the data is hard to interpret.

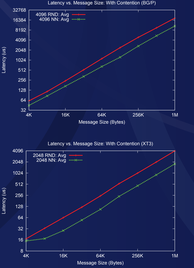


Benchmark II: Random Contention

All COMM_WORLD ranks are divided into pairs. The partner for each rank is either one hop away (near-neighbor or NN mode) or a random processor (random or RND mode). All pairs exchange messages simultaneously.

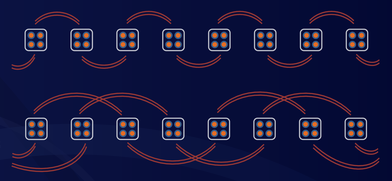
Findings:

1. In presence of contention, message latencies can increase by a factor of two.
2. The difference between the NN and RND modes is a factor of 1.75 on Blue Gene/P and 2.25 on XT3.
3. The phenomenon is observed for XT3 even though the interconnect is much faster and the absolute latencies are better than on Blue Gene/P.



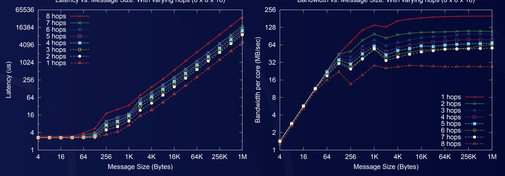
Benchmark III: Controlled Contention

All COMM_WORLD ranks are divided into pairs. The partner for each rank is 'n' hops away. Messages are sent only along one dimension (Z) of the torus and n is varied in different experiments.



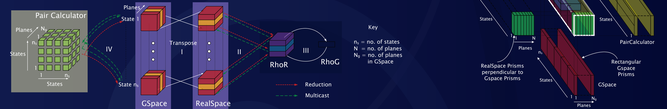
Findings:

1. Sharing of links between messages leads to contention and increased message latencies.
2. Difference in latencies when all messages travel one hop versus eight hops is as much as eight times!
3. Topology aware mapping is important for some applications.

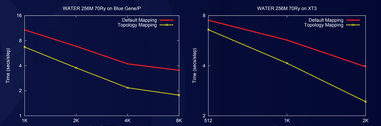


OpenAtom: A Case Study of Topology Aware Mapping

OpenAtom is a complex quantum chemistry application written in Charm++. It is heavily communication bound and has multiple object graphs which communicate with each other along orthogonal dimensions.



The challenge in mapping of OpenAtom objects comes from the conflicting requirements for optimizing communication between different groups of objects. Heuristic mapping schemes for torus and mesh interconnects give improvements up to 40% on Blue Gene/P and Cray XT3.



Summary

Link contention in 3D interconnects can affect message latencies significantly.
It is important to consider the topology of the machine to optimize performance.