

IS TOPOLOGY IMPORTANT AGAIN?

Effects of contention on message latencies in
large supercomputers

Abhinav S Bhatele and Laxmikant V Kale
Parallel Programming Laboratory, UIUC



E-mail: bhatele, kale @ illinois.edu
Webpage: <http://charm.cs.uiuc.edu>



Outline

Why should we consider topology aware mapping for optimizing performance?

Demonstrate the effects of contention on message latencies through simple MPI benchmarks

Case Study: OpenAtom



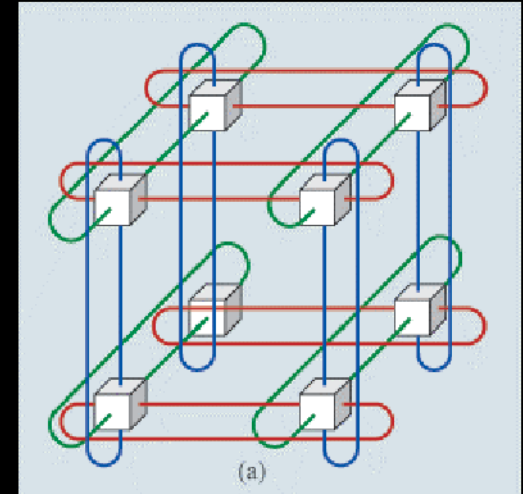
The Mapping Problem

- Given a set of communicating parallel “entities”, map them onto physical processors
- Entities
 - COMM_WORLD ranks in case of an MPI program
 - Objects in case of a Charm++ program
- Aim
 - Balance load
 - Minimize communication traffic



Target Machines

- 3D torus/mesh interconnects
- Blue Gene/P at ANL:
 - 40,960 nodes, torus - 32 x 32 x 40
- XT4 (Jaguar) at ORNL:
 - 8,064 nodes, torus - 21 x 16 x 24
- Other interconnects
 - Fat-tree



Motivation

- Consider a 3D mesh/torus interconnect
- Message latencies can be modeled by

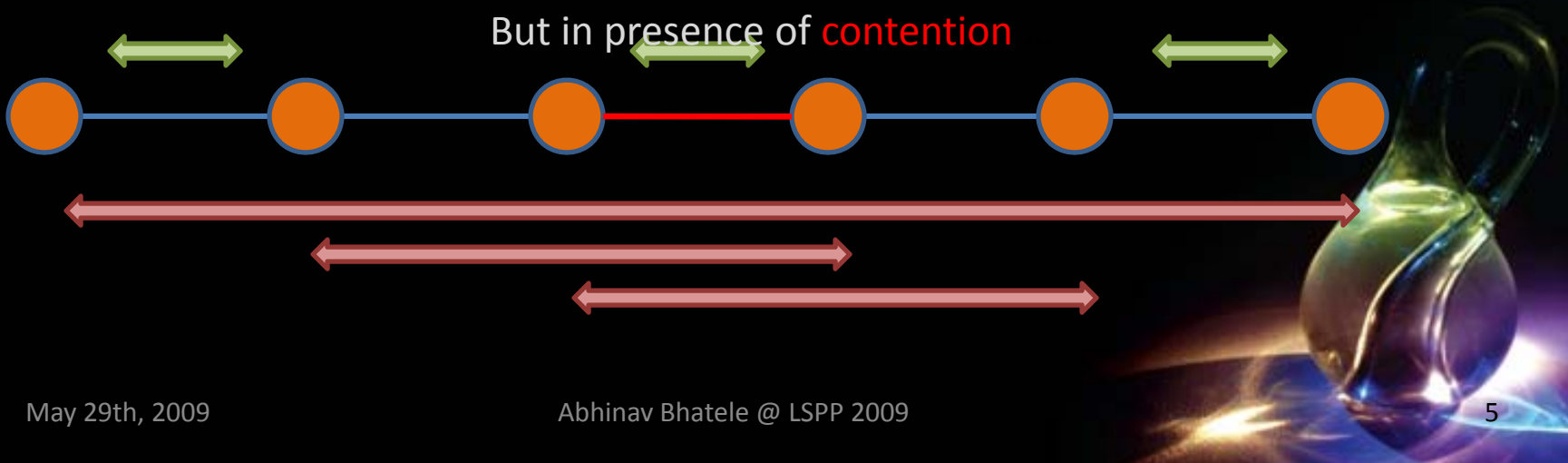
$$(L_f/B) \times D + L/B$$

L_f = length of flit, B = bandwidth,

D = hops, L = message size

When $(L_f * D) \ll L$, first term is negligible

But in presence of contention



MPI Benchmarks†

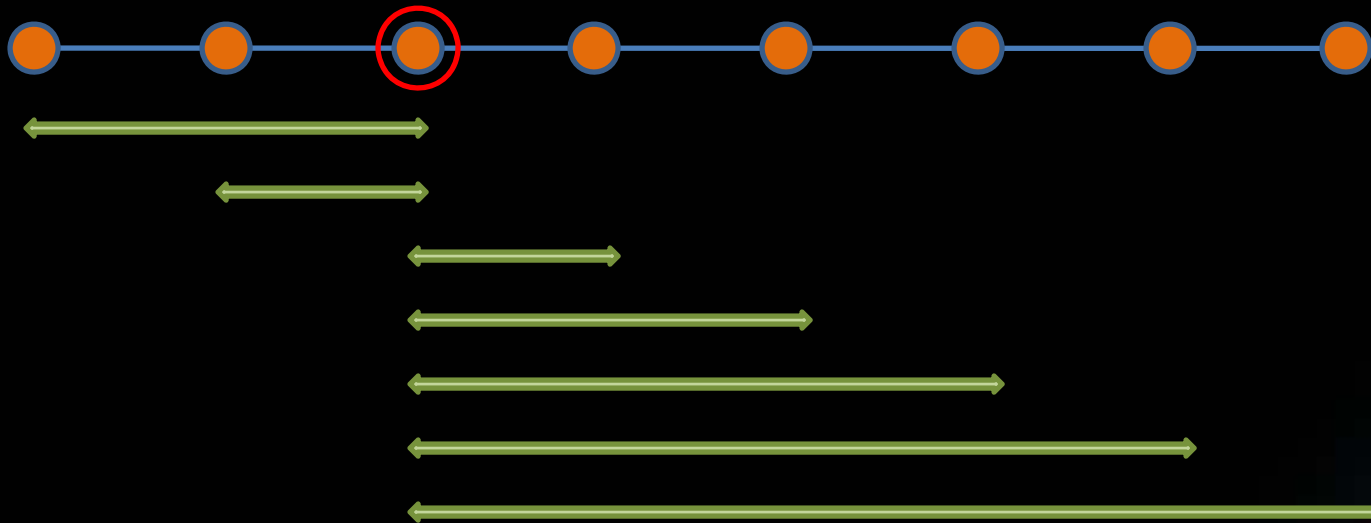
- Quantification of message latencies and dependence on hops
 - No sharing of links (no contention)
 - Sharing of links (with contention)

† <http://charm.cs.uiuc.edu/~bhatele/phd/contention.htm>



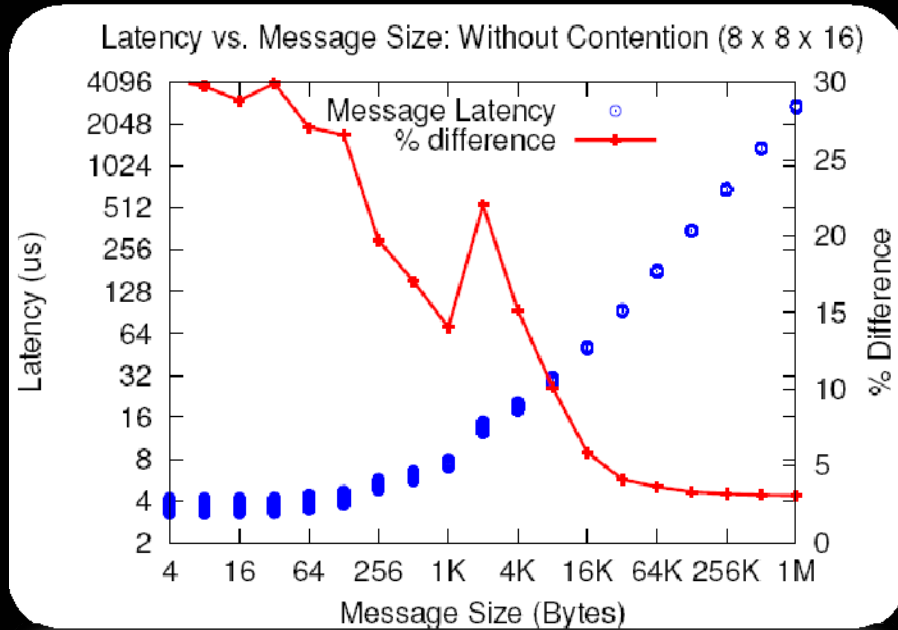
WOCON: No contention

- A master rank sends messages to all other ranks, one at a time (with replies)

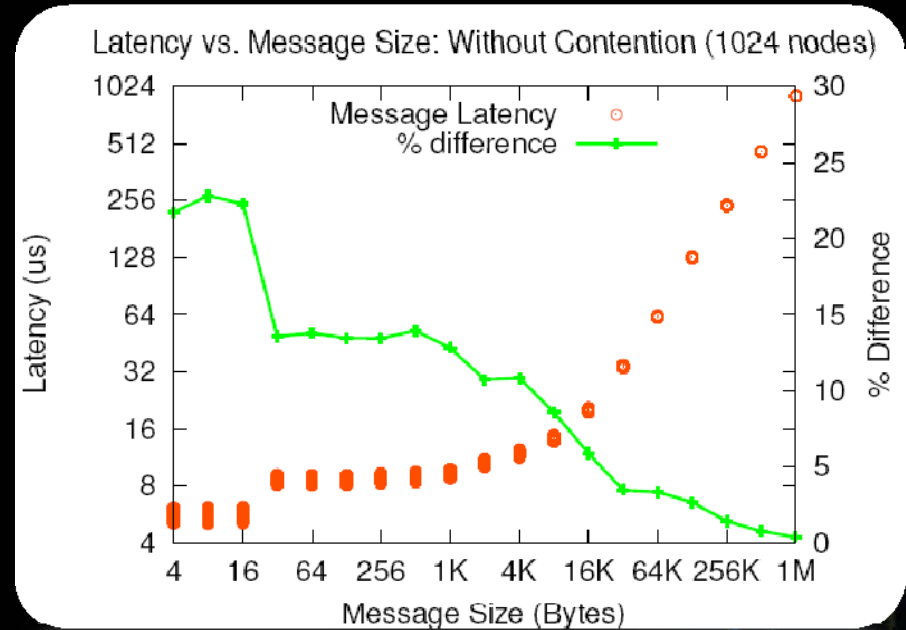


WOCON: Results

$$(L_f/B) \times D + L/B$$



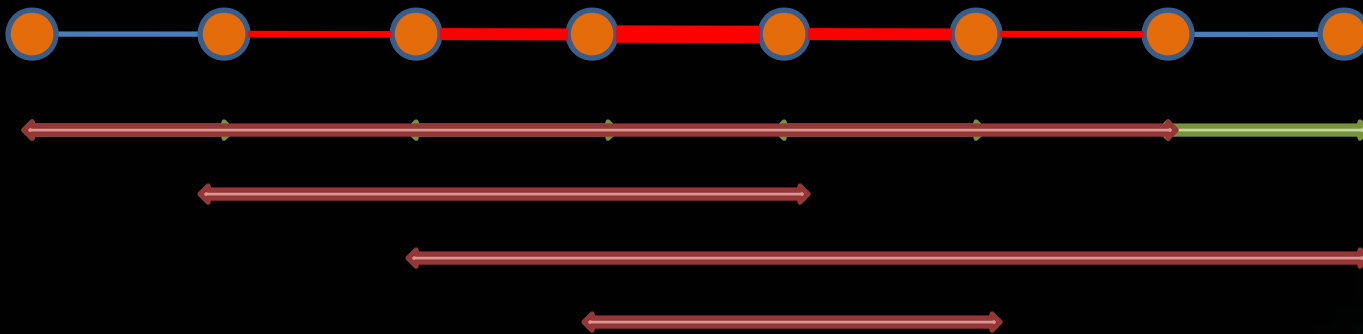
ANL Blue Gene/P



PSC XT3

WICON: With Contention

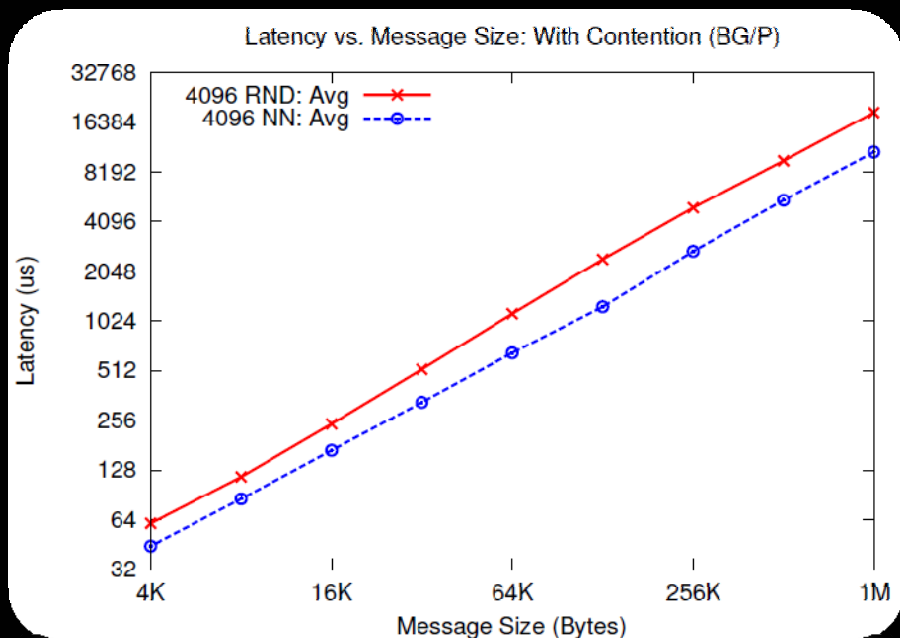
- Divide all ranks into pairs and everyone sends to their respective partner simultaneously



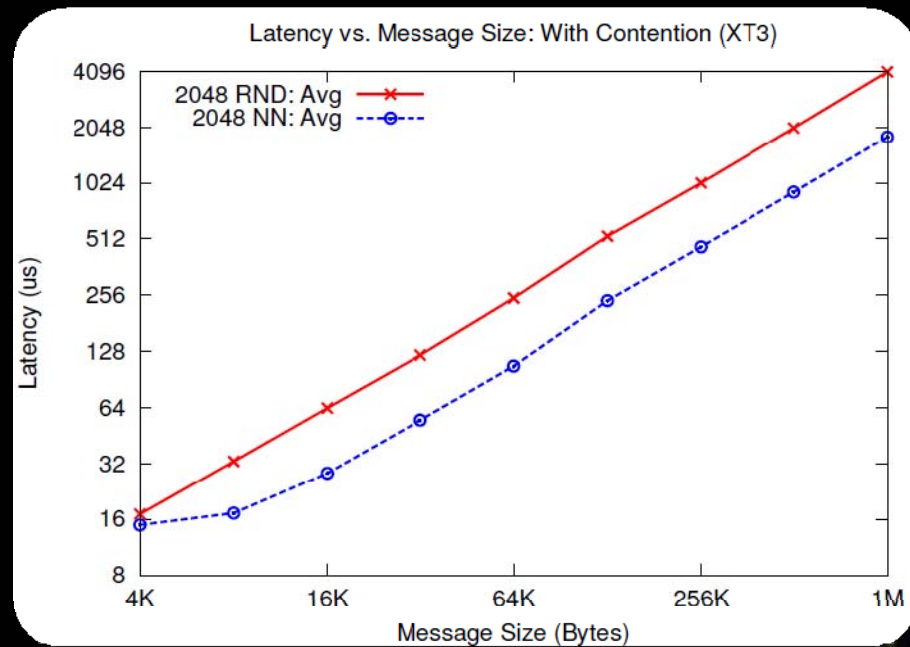
Read Neighbor: NN



WICON: Results



ANL Blue Gene/P

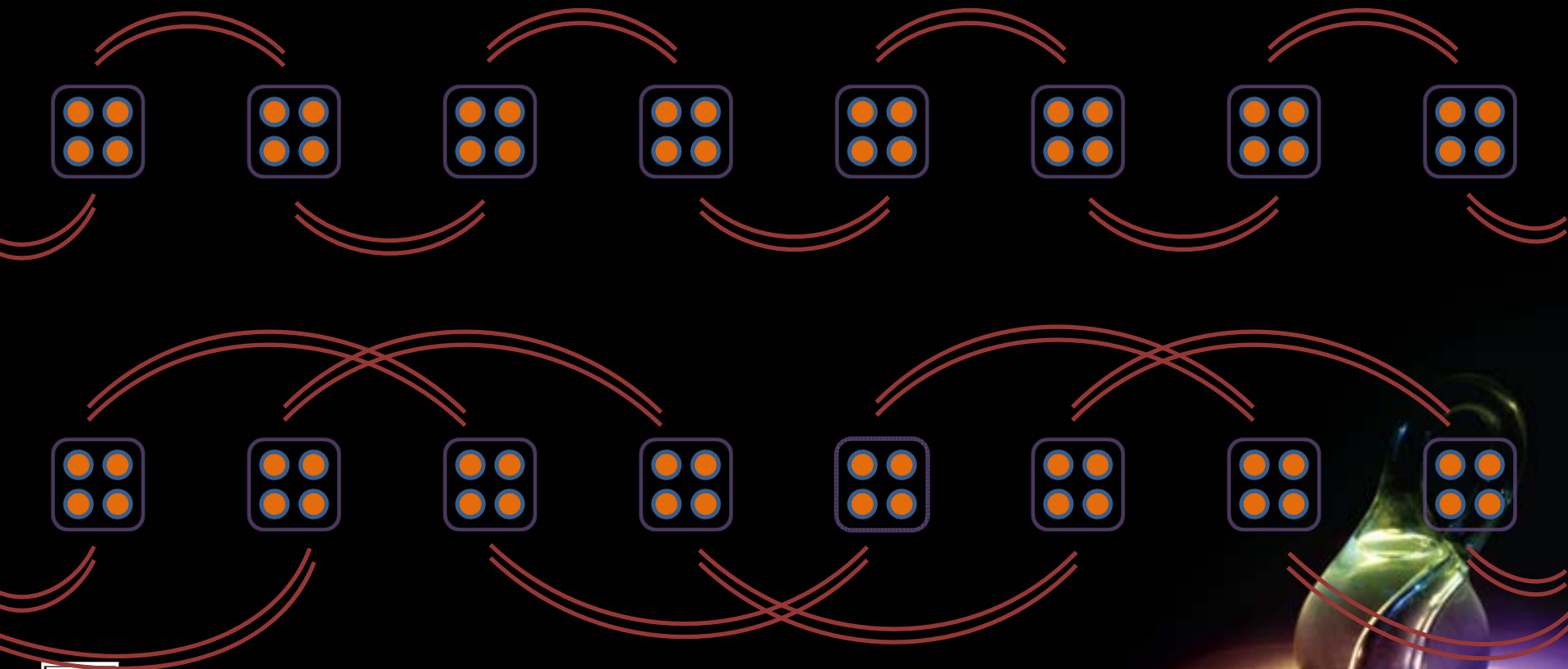


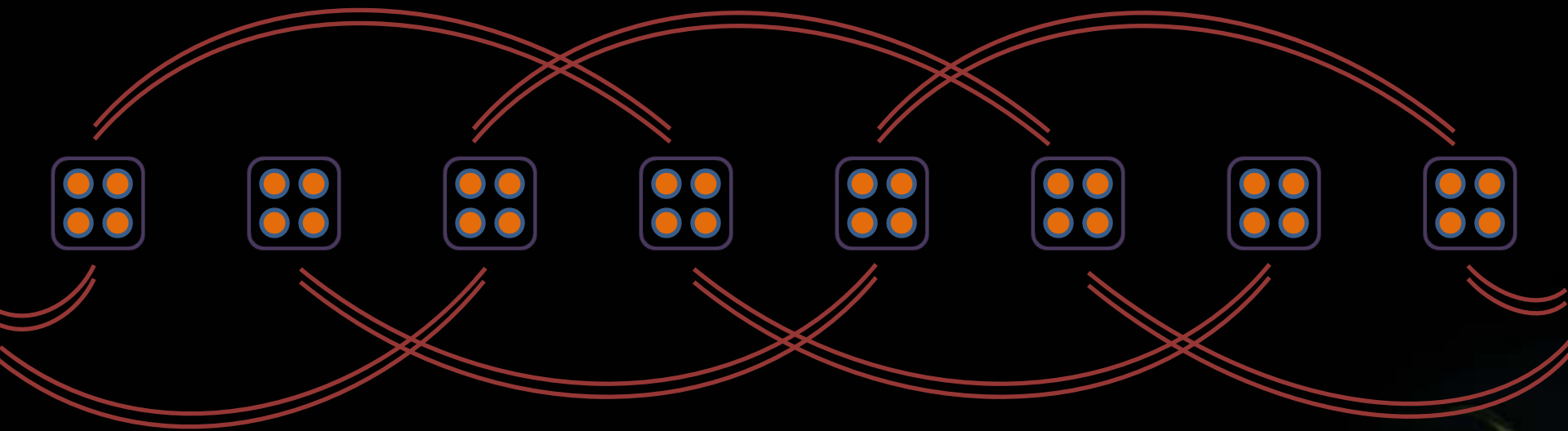
PSC XT3



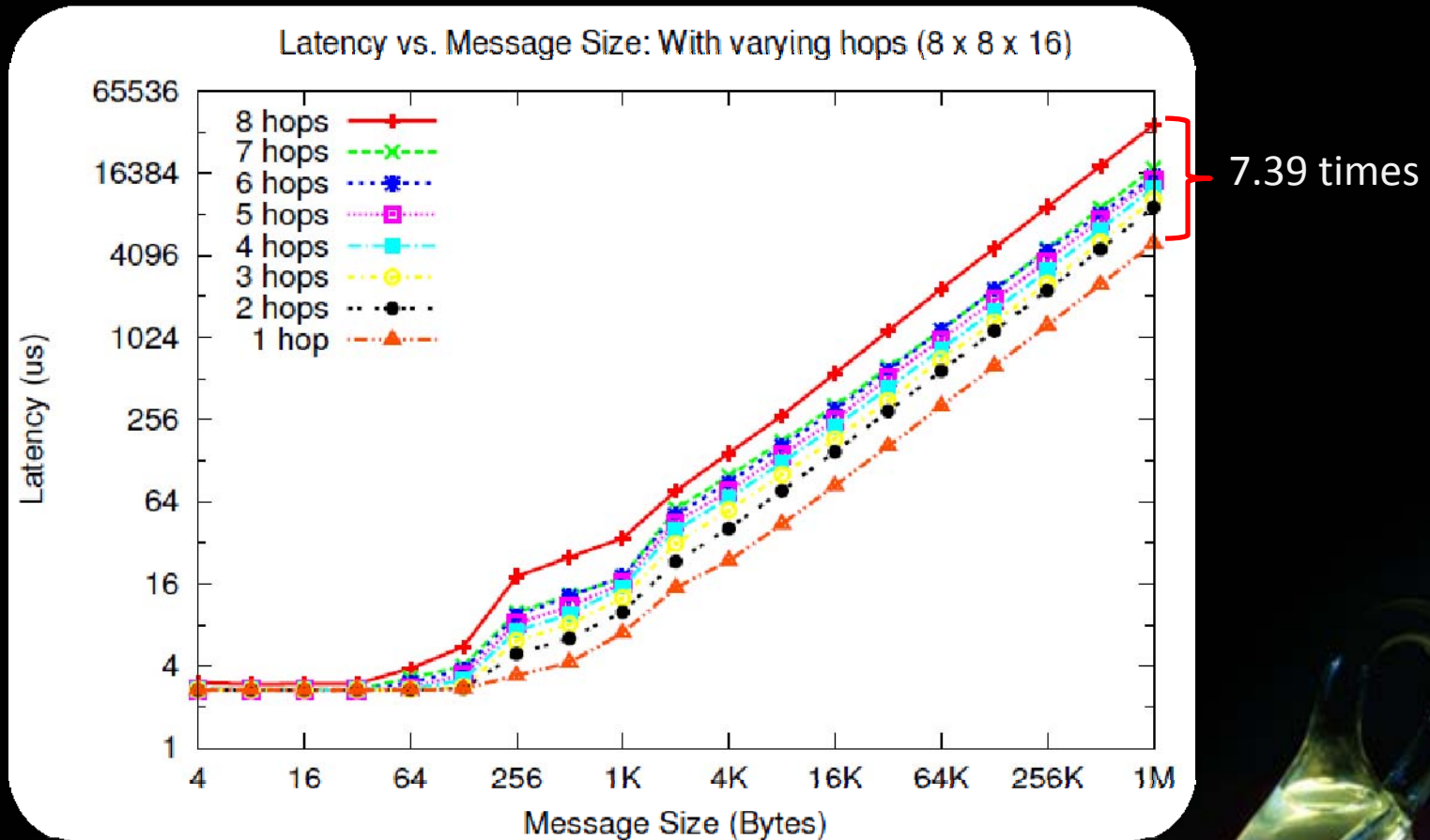
WICON2: Controlled Contention

- Pair each rank with a partner which is 'n' hops away

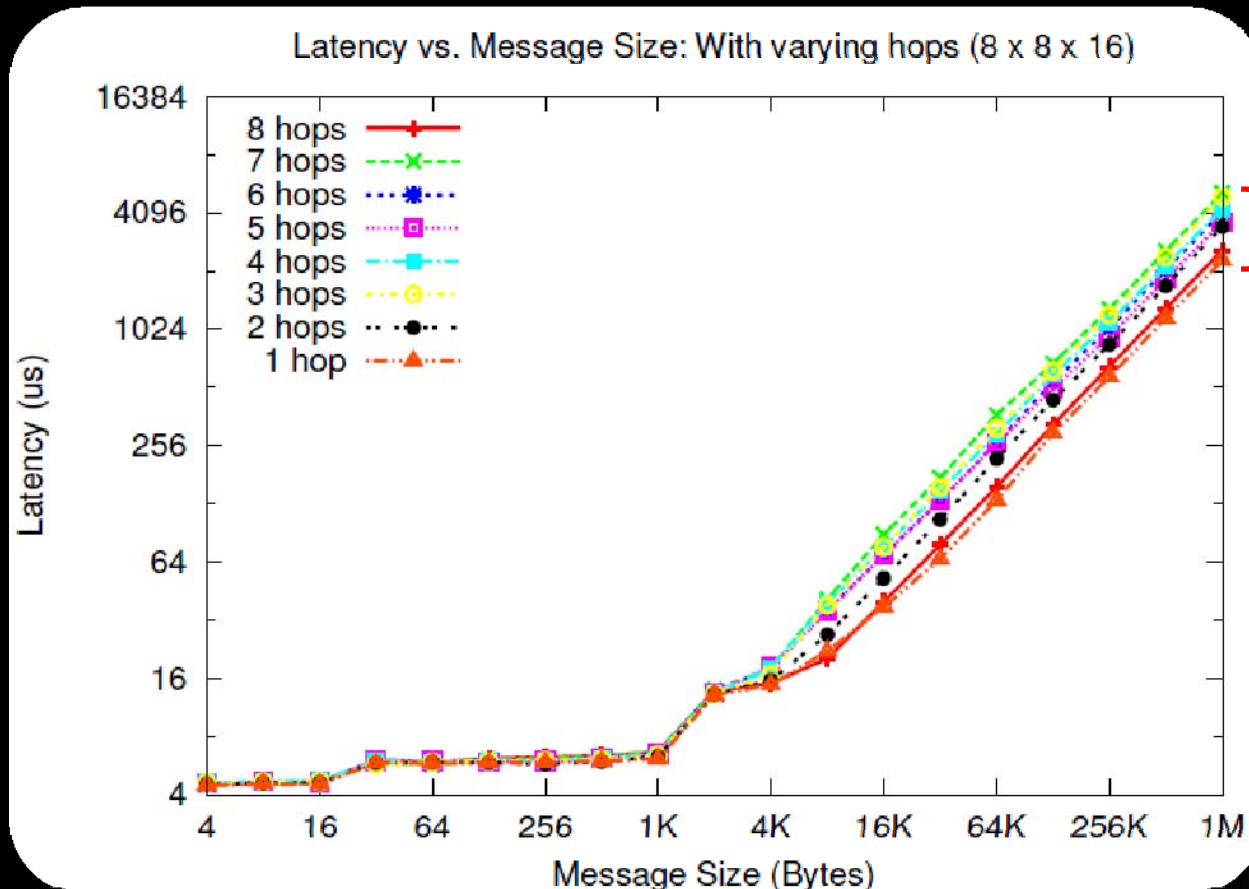




WICON2: Blue Gene/P

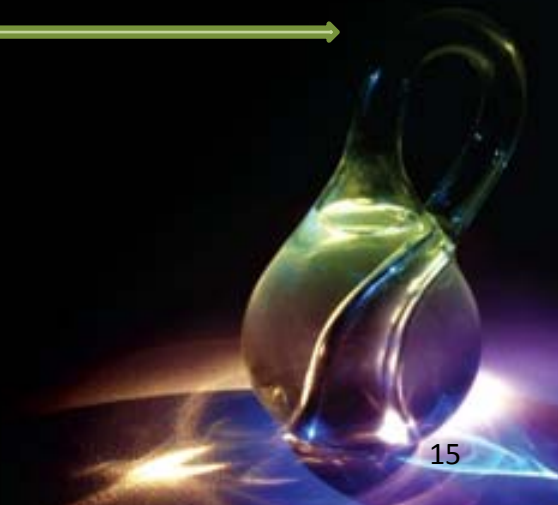
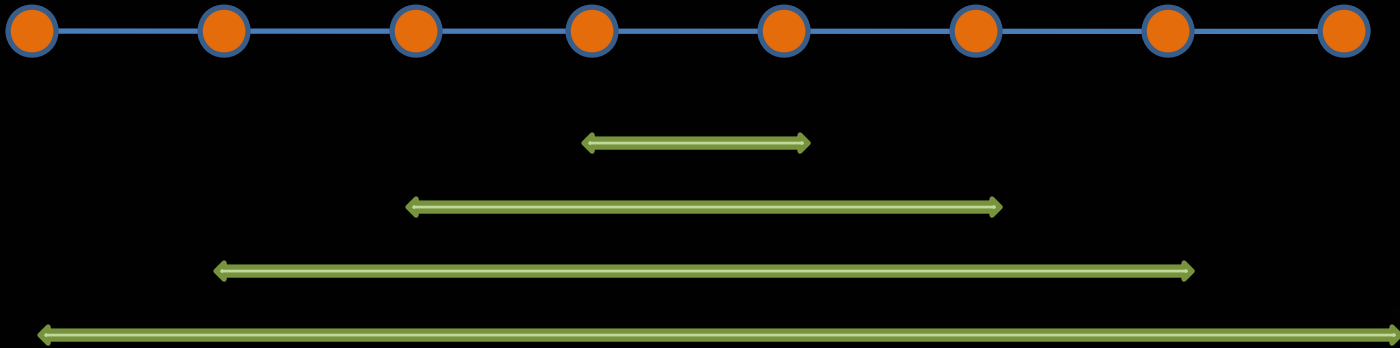


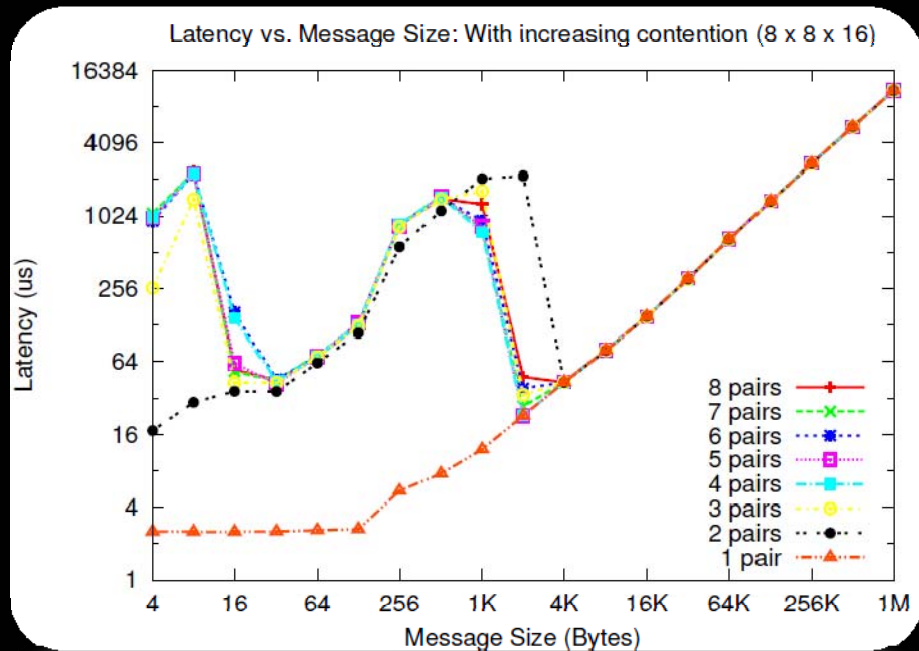
WICON2: Cray XT3



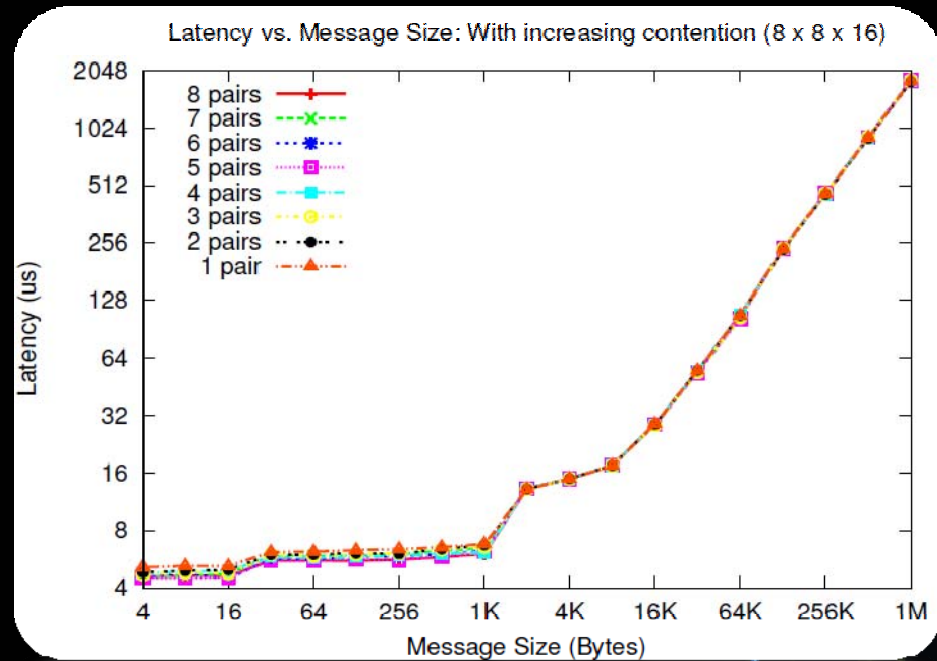
2.23 times

More tests ...



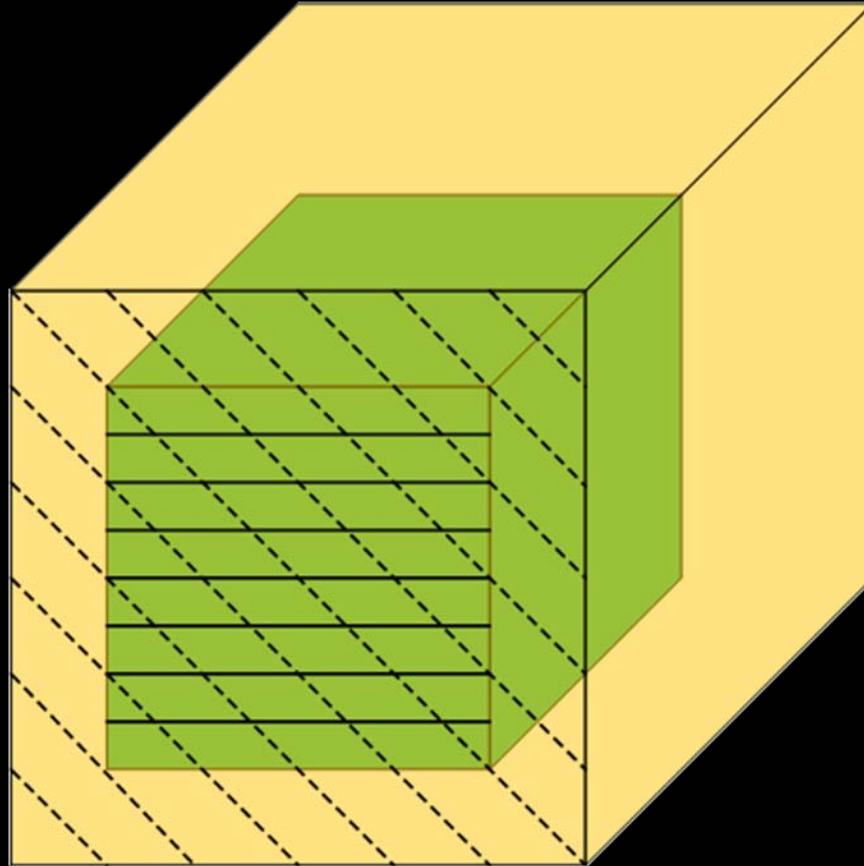


ANL Blue Gene/P

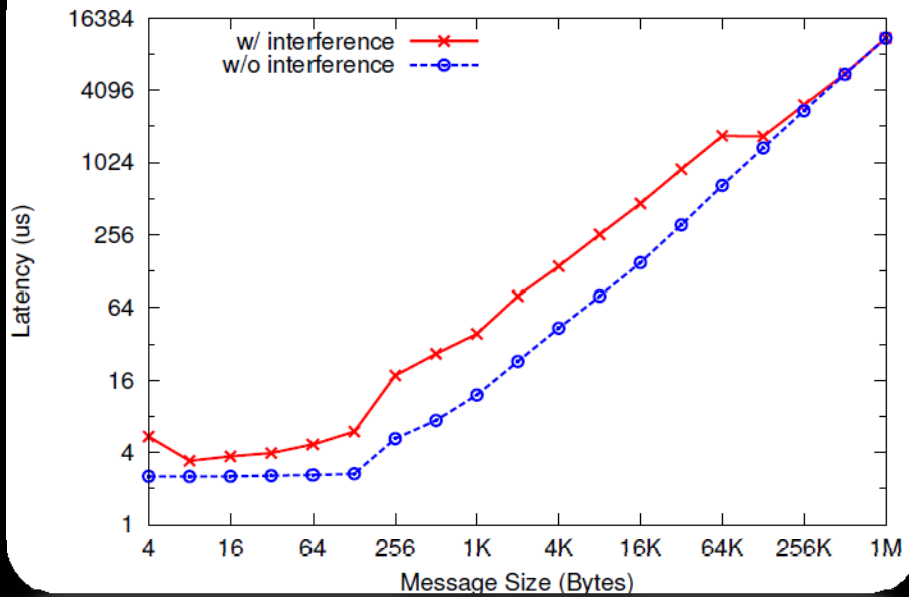


PSC XT3

More tests ...

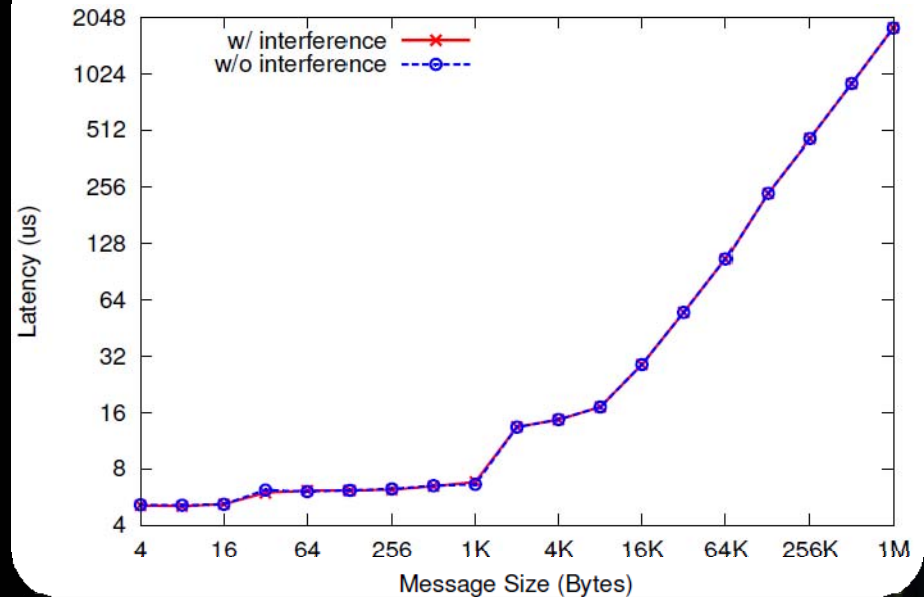


Latency vs. Message Size: Job Interference (8 x 8 x 8)



ANL Blue Gene/P

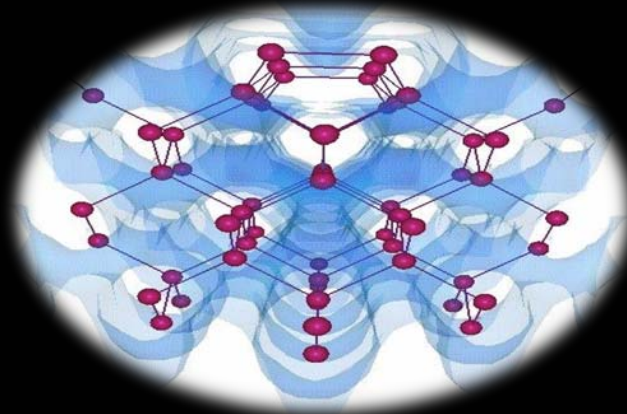
Latency vs. Message Size: Job Interference (8 x 8 x 8)



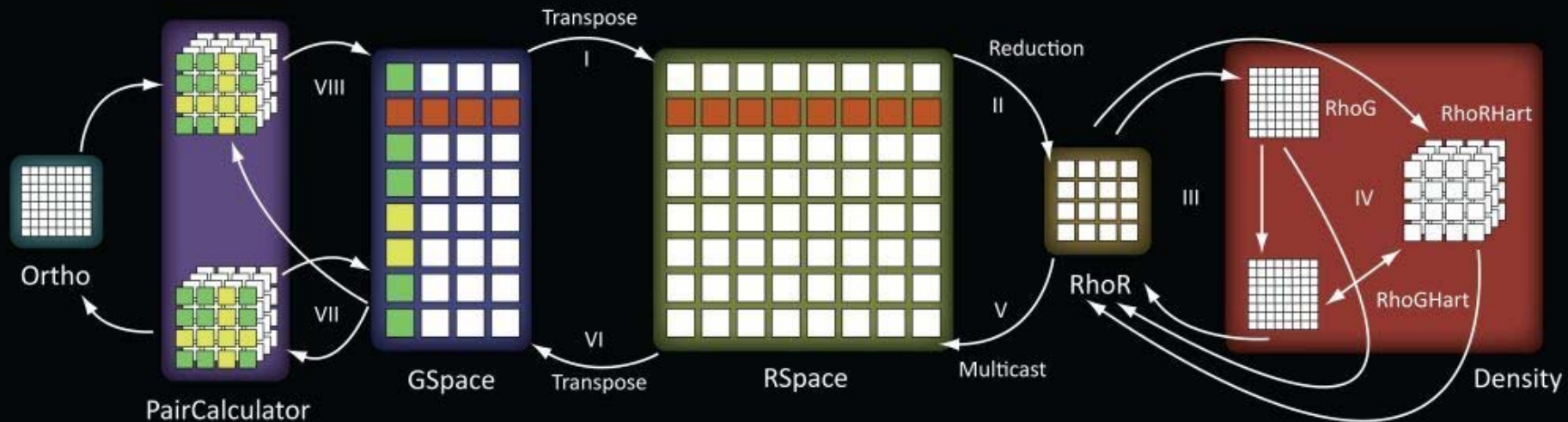
PSC XT3

OpenAtom

- Ab-Initio Molecular Dynamics code
- Communication is static and structured
- Challenge: Multiple groups of objects with conflicting communication patterns



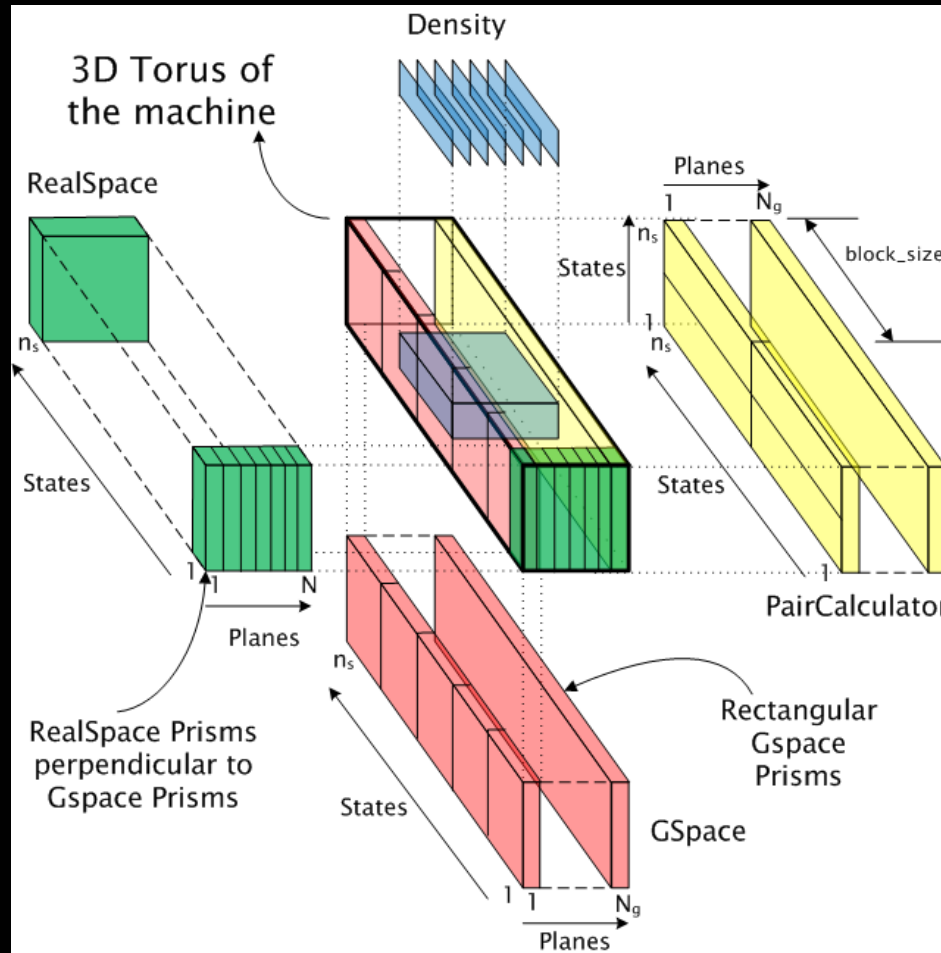
Parallelization using Charm++



[10] Eric Bohm, Glenn J. Martyna, Abhinav Bhatele, Sameer Kumar, Laxmikant V. Kale, John A. Gunnels, and Mark E. Tuckerman. **Fine Grained Parallelization of the Car-Parrinello ab initio MD Method on Blue Gene/L.** *IBM J. of R. and D.: Applications of Massively Parallel Systems*, 52(1/2):159-174, 2008.



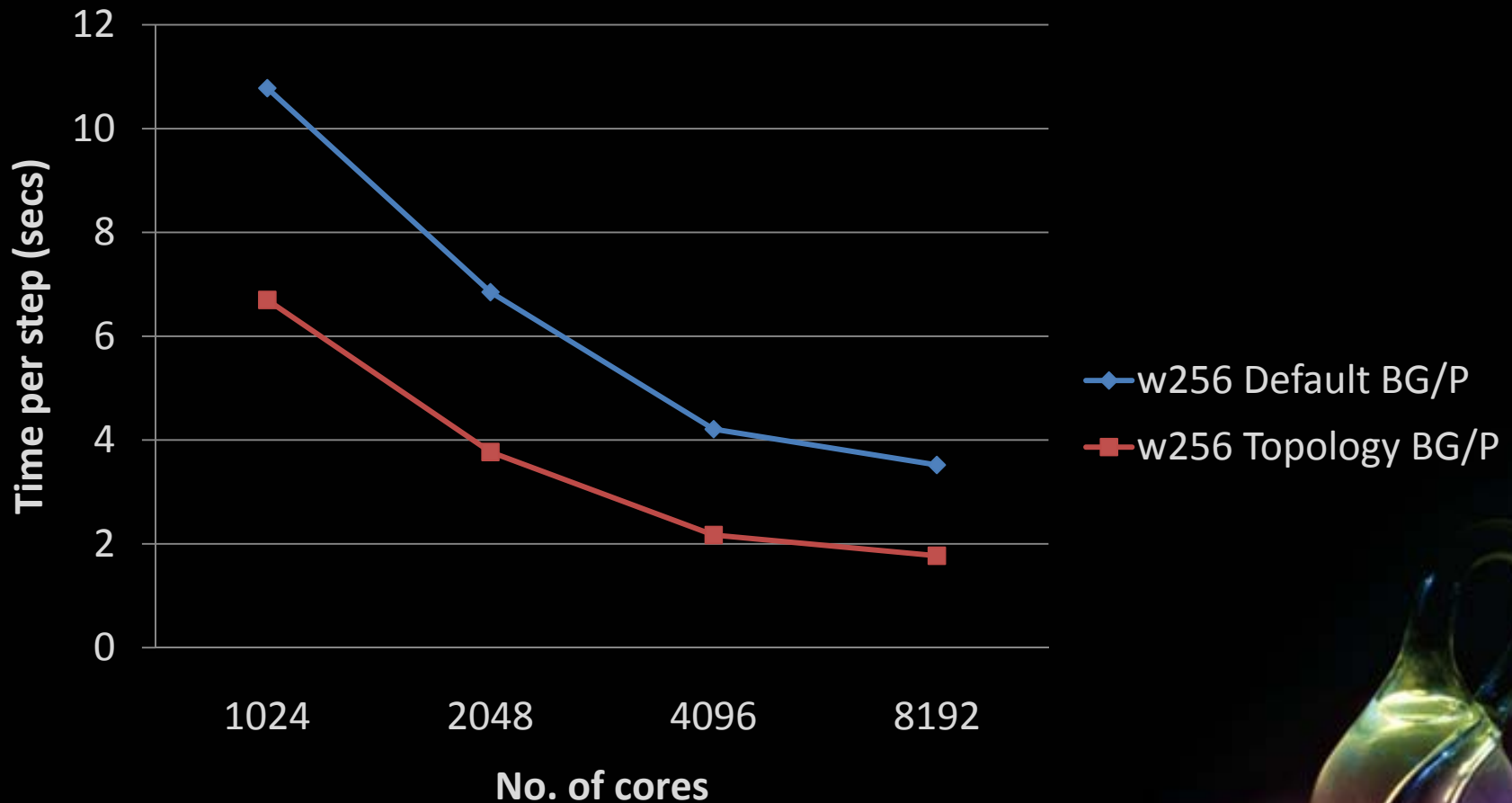
Topology Mapping of Chare Arrays



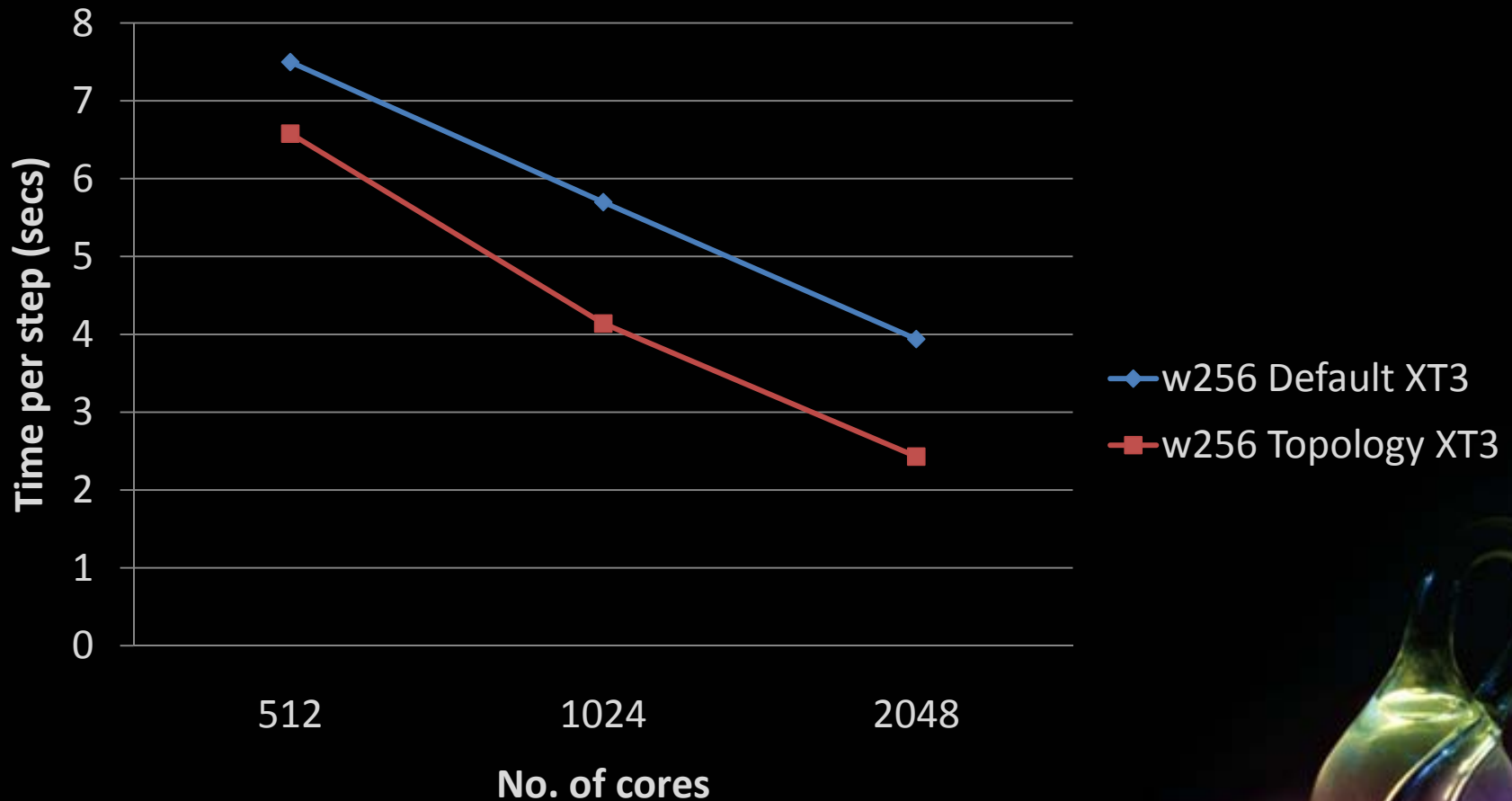
State-wise communication

Plane-wise communication

Results on Blue Gene/P (ANL)



Results on XT3 (BigBen@PSC)



Summary

1. Topology is important again
2. Even on fast interconnects such as Cray
3. In presence of contention, bandwidth occupancy effects message latencies significantly
4. Increases with the number of hops each message travels
5. Topology Manager API: A uniform API for IBM and Cray machines
6. Case Studies: OpenAtom, NAMD, Stencil
7. Eventually, an automatic mapping framework

Acknowledgements:

1. Argonne National Laboratory: Pete Beckman, Tisha Stacey
2. Pittsburgh Supercomputing Center: Chad Vizino, Shawn Brown
3. Oak Ridge National Laboratory: Patrick Worley, Donald Frederick
4. IBM: Robert Walkup, Sameer Kumar
5. Cray: Larry Kaplan
6. Matt Reilly

References:

1. Abhinav Bhatele, Laxmikant V. Kale, **Dynamic Topology Aware Load Balancing Algorithms for MD Applications**, *To appear in Proceedings of International Conference on Supercomputing (ICS)*, 2009
2. Abhinav Bhatele, Eric Bohm, Laxmikant V. Kale, A Case Study of Communication Optimizations on 3D Mesh Interconnects, *To appear in Proceedings of Euro-Par*, 2009
3. Abhinav Bhatele, Laxmikant V. Kale, **Benefits of Topology-aware Mapping for Mesh Topologies**, *Parallel Processing Letters (Special issue on Large-Scale Parallel Processing)*, Vol: 18 Issue:4, Pages: 549-566, 2008



E-mail: bhatele, kale @ illinois.edu
Webpage: <http://charm.cs.uiuc.edu>

