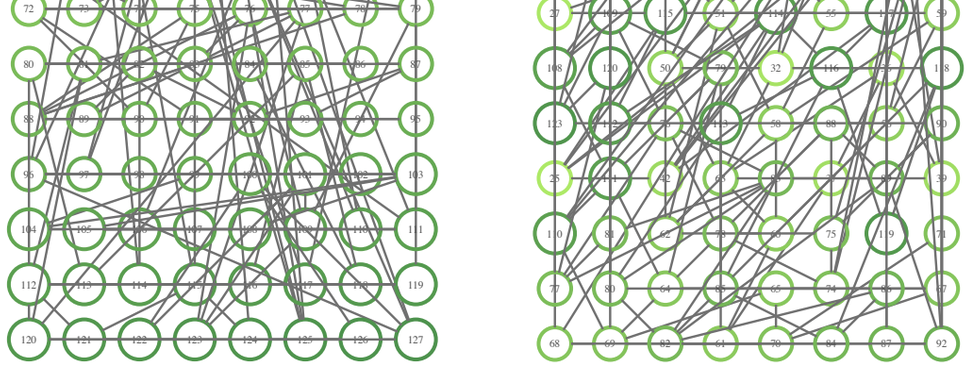
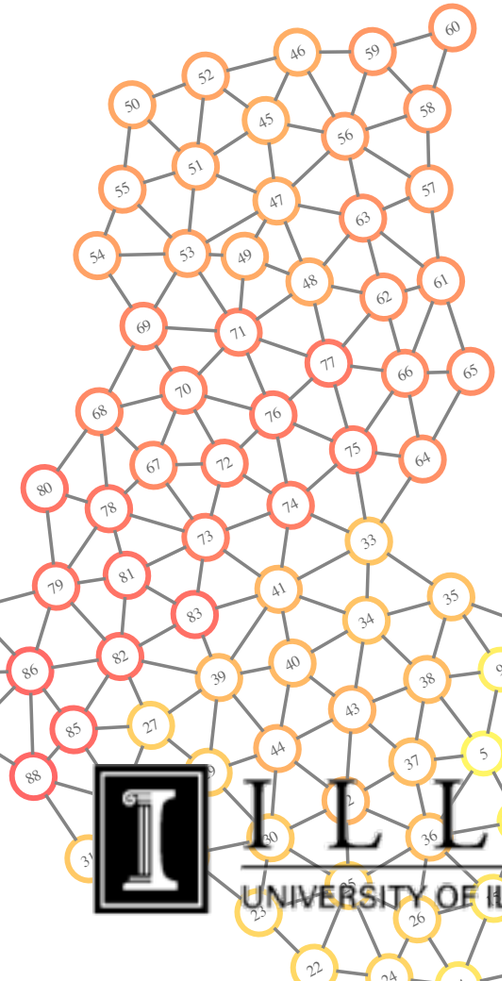
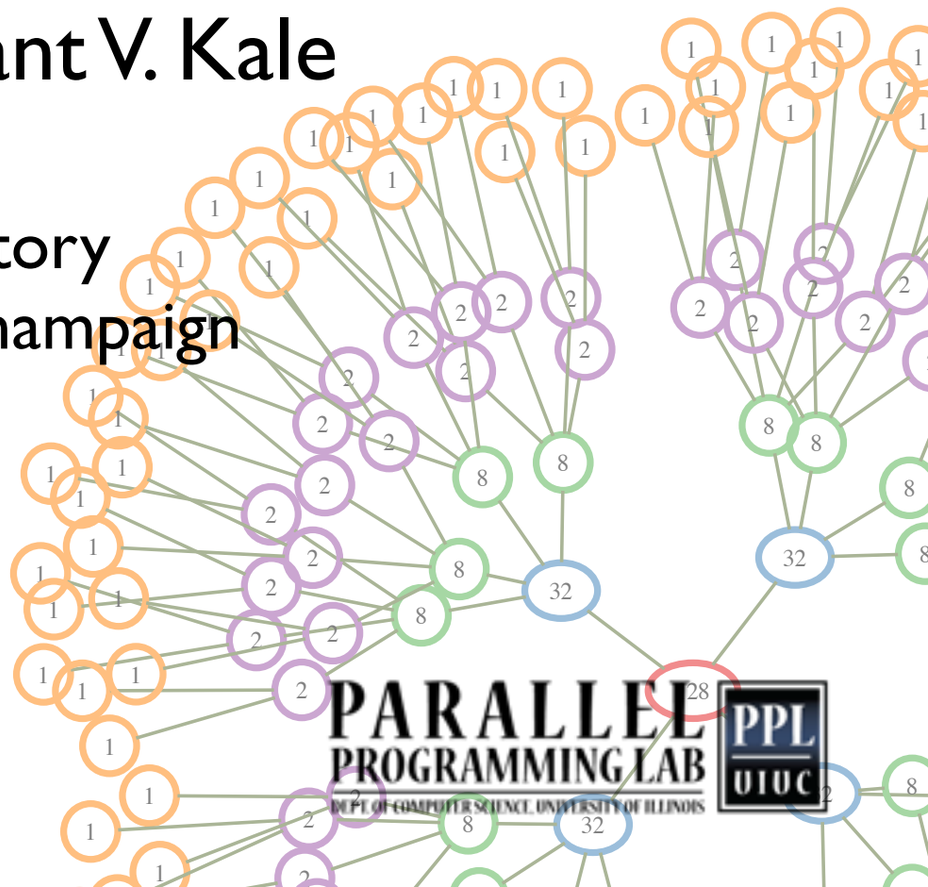


Mapping parallel applications on the machine topology: Lessons learned

Abhinav Bhatele and Laxmikant V. Kale

Parallel Programming Laboratory
University of Illinois at Urbana-Champaign



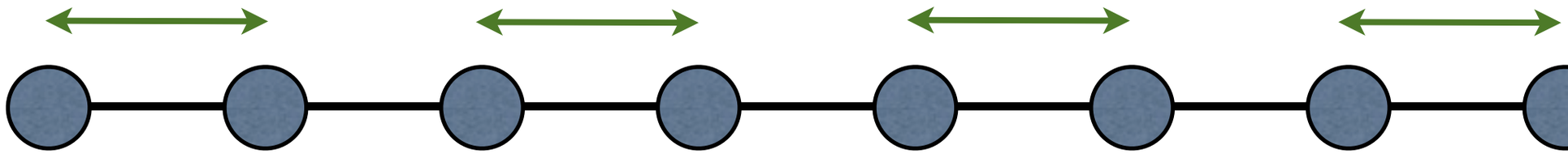


ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

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

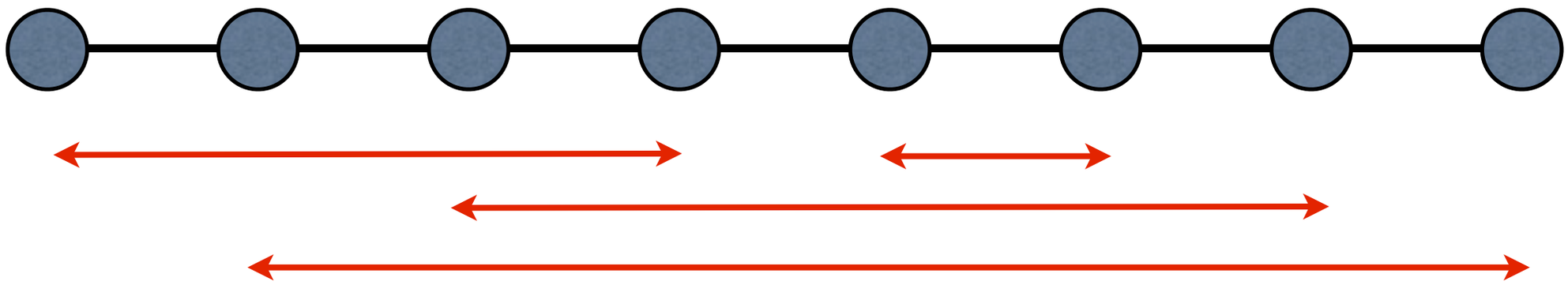
Motivation

- Running a parallel application on a linear array of processors:



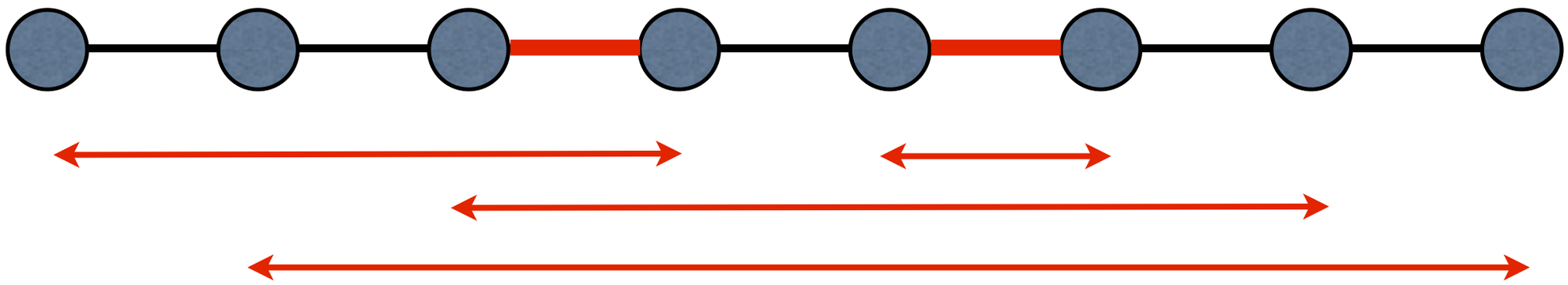
Motivation

- Running a parallel application on a linear array of processors:



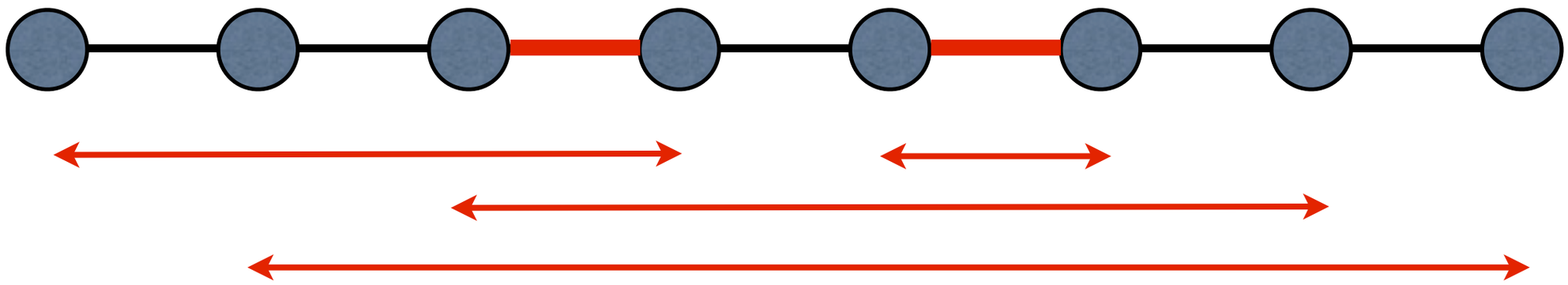
Motivation

- Running a parallel application on a linear array of processors:



Motivation

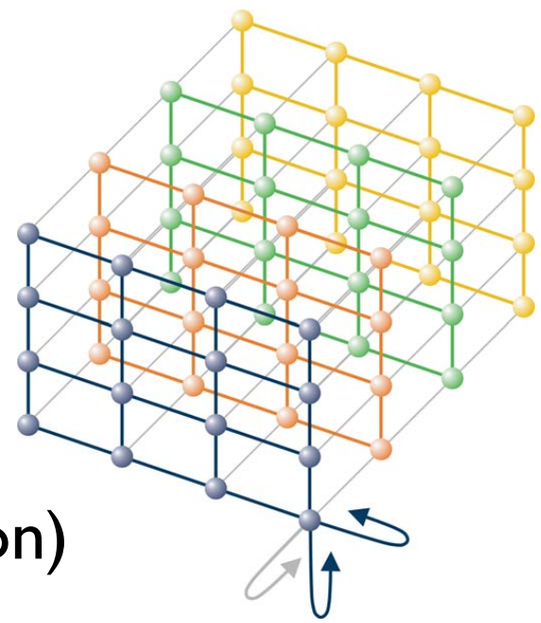
- Running a parallel application on a linear array of processors:

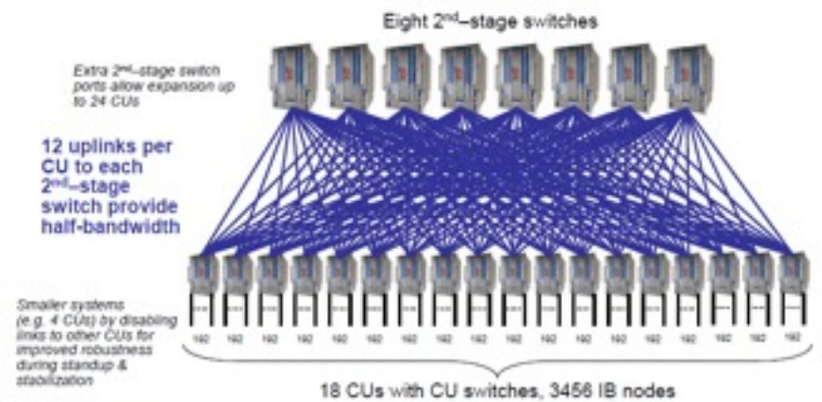


- Typical communication is between random pairs of processors simultaneously

Interconnect Topologies

- Three dimensional meshes
  - 3D Torus: Blue Gene/L, Blue Gene/P, Cray XT4/5
- Trees
  - Fat-trees (Infiniband) and CLOS networks (Federation)
- Dense Graphs
  - Kautz Graph (SiCortex), Hypercubes
- Future Topologies?
  - Blue Waters, Blue Gene/Q

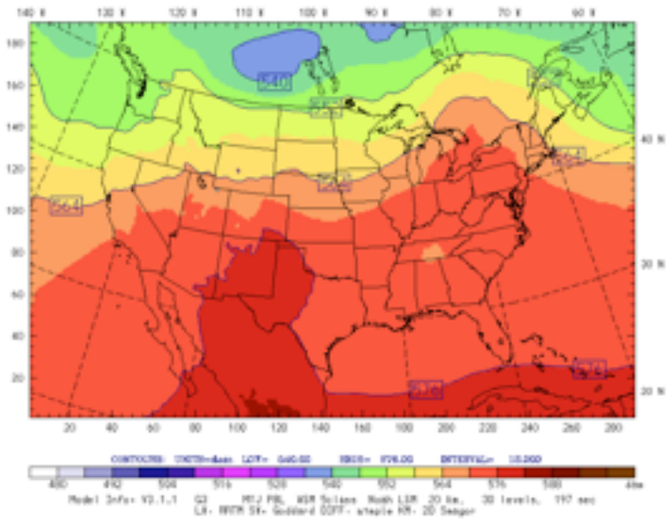


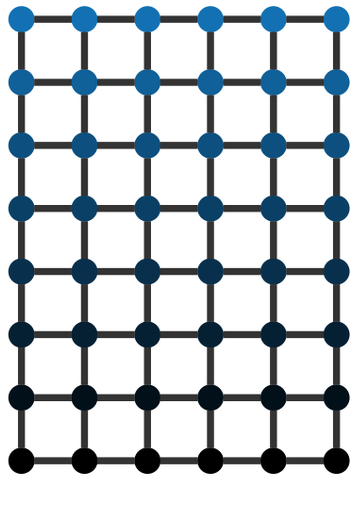


Roadrunner Technical Seminar Series, March 13th 2008, Ken Koch, LANL

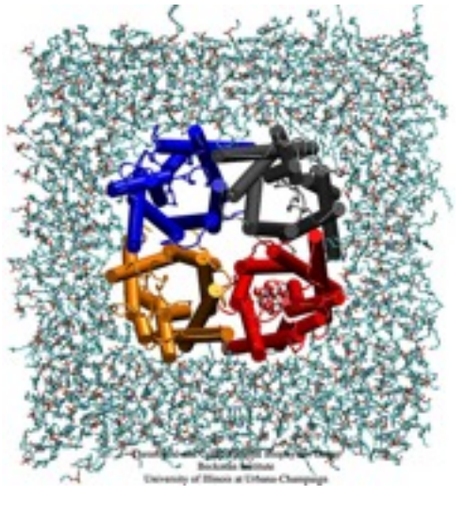
Application Topologies

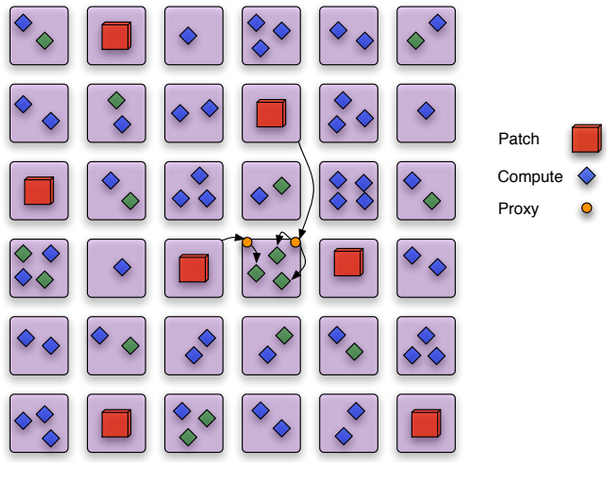
http://wrf-model.org/plots/realtime_main.php

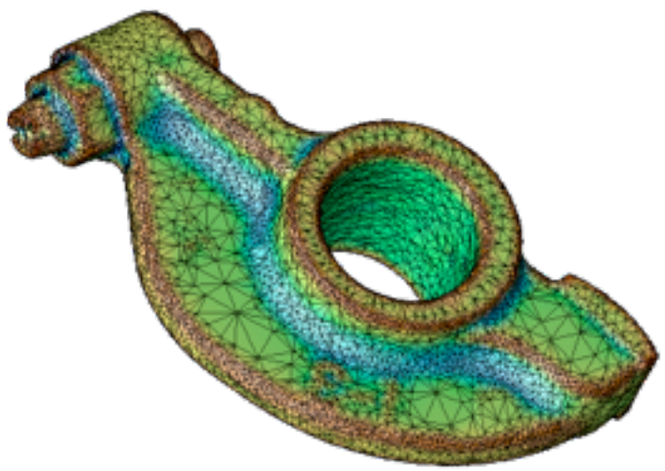




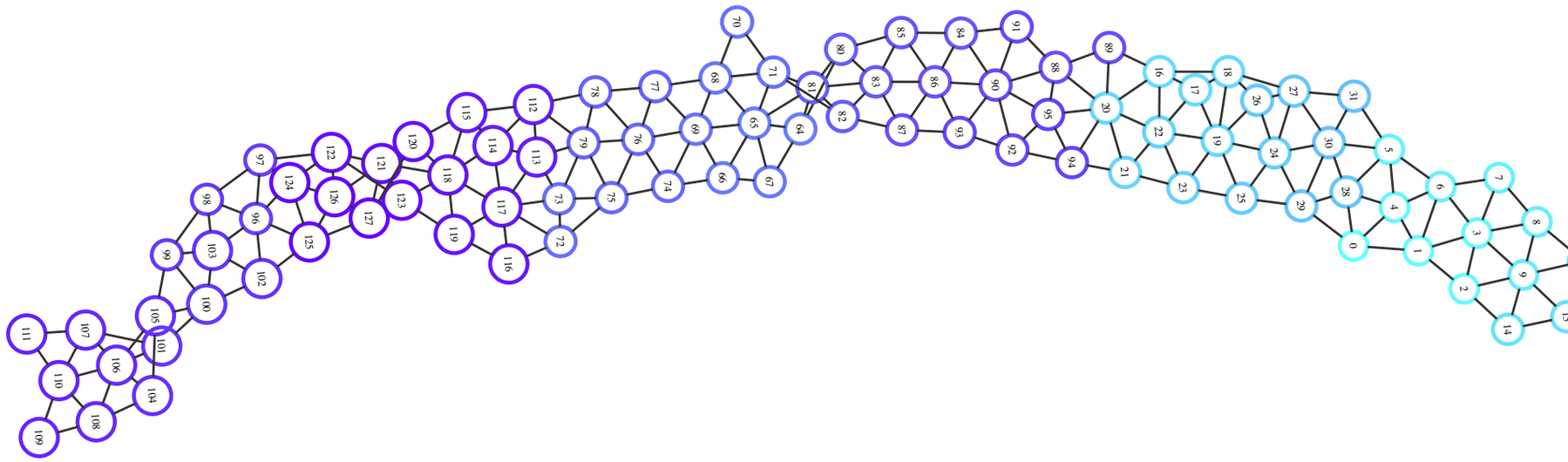
http://www.ks.uiuc.edu/Gallery/Science/





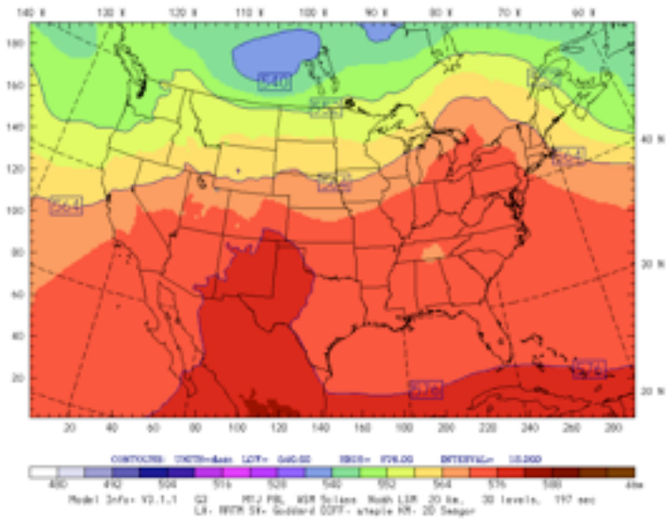


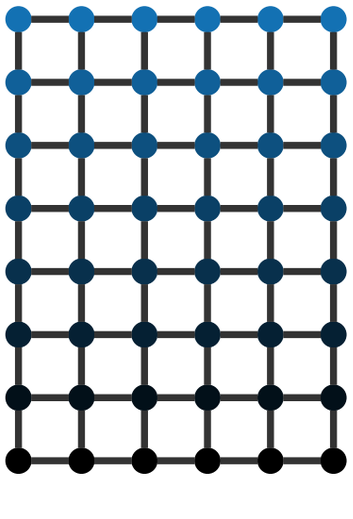
http://math.lanl.gov/Research/Projects/meshing.shtml



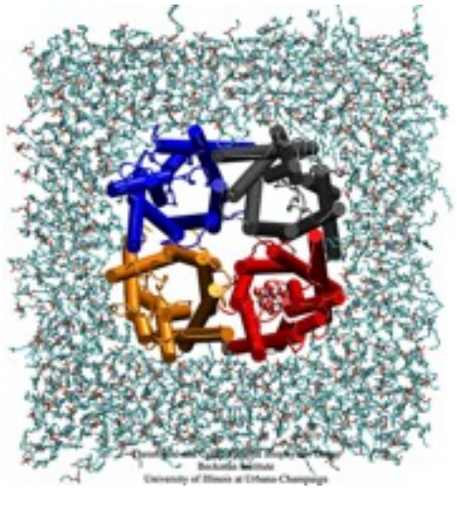
Application Topologies

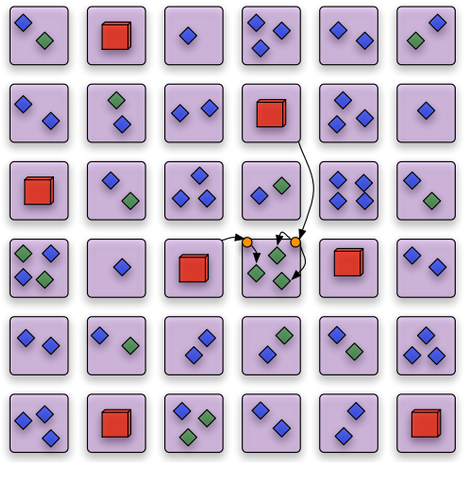
http://wrf-model.org/plots/realtime_main.php

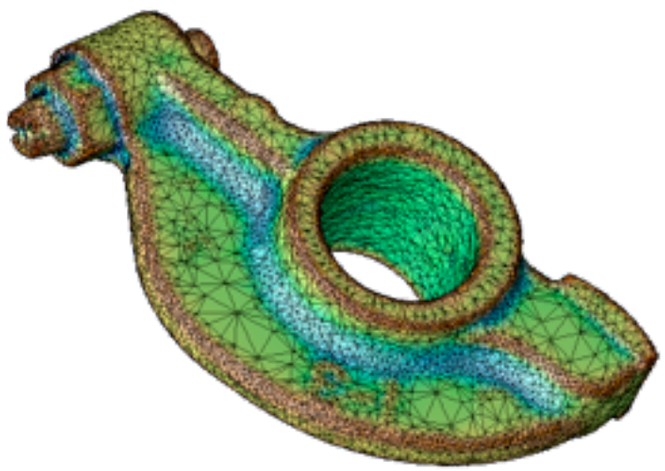




http://www.ks.uiuc.edu/Gallery/Science/







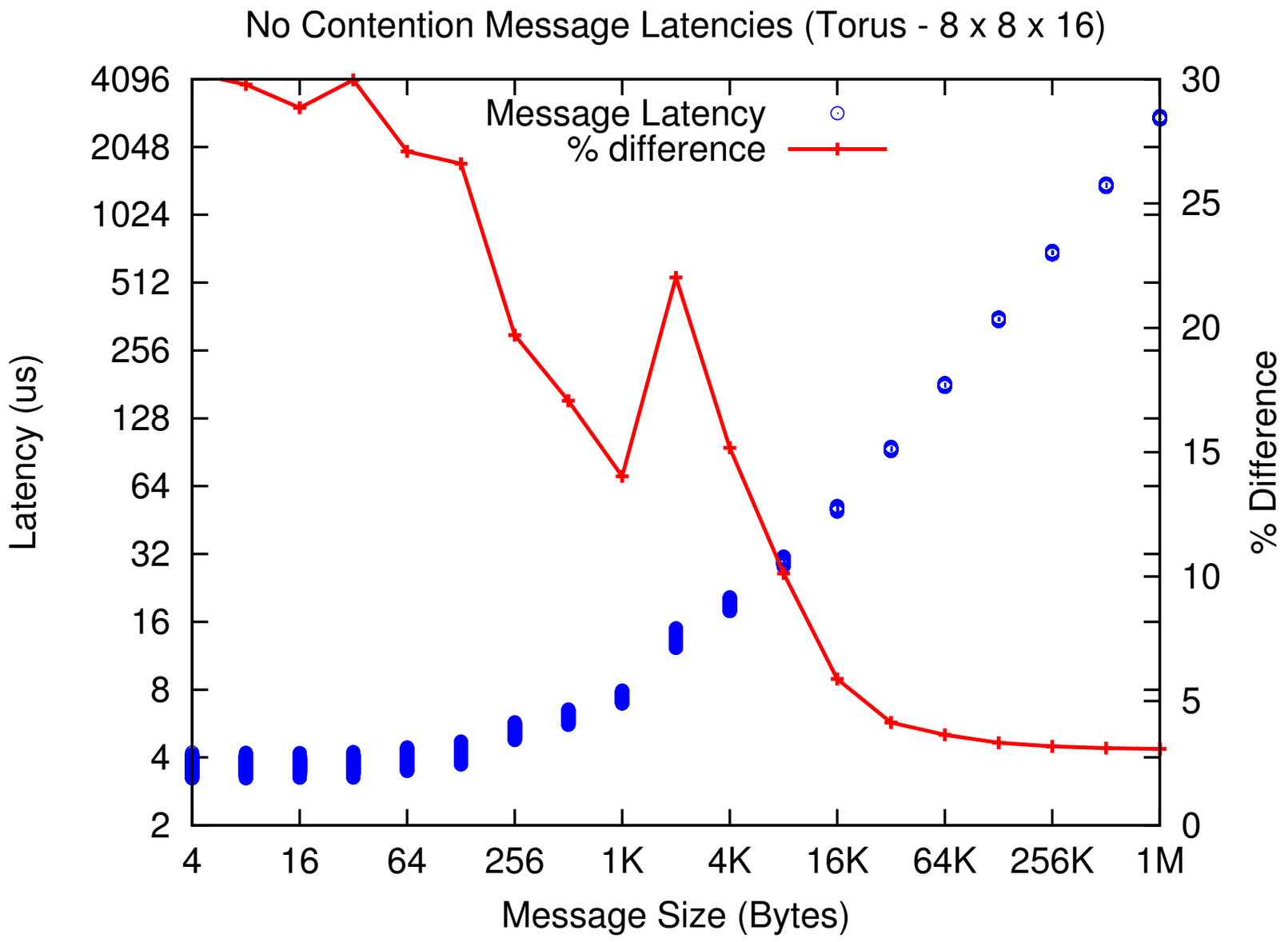
http://math.lanl.gov/Research/Projects/meshing.shtml

We want to map communicating objects closer to one another

The Mapping Problem

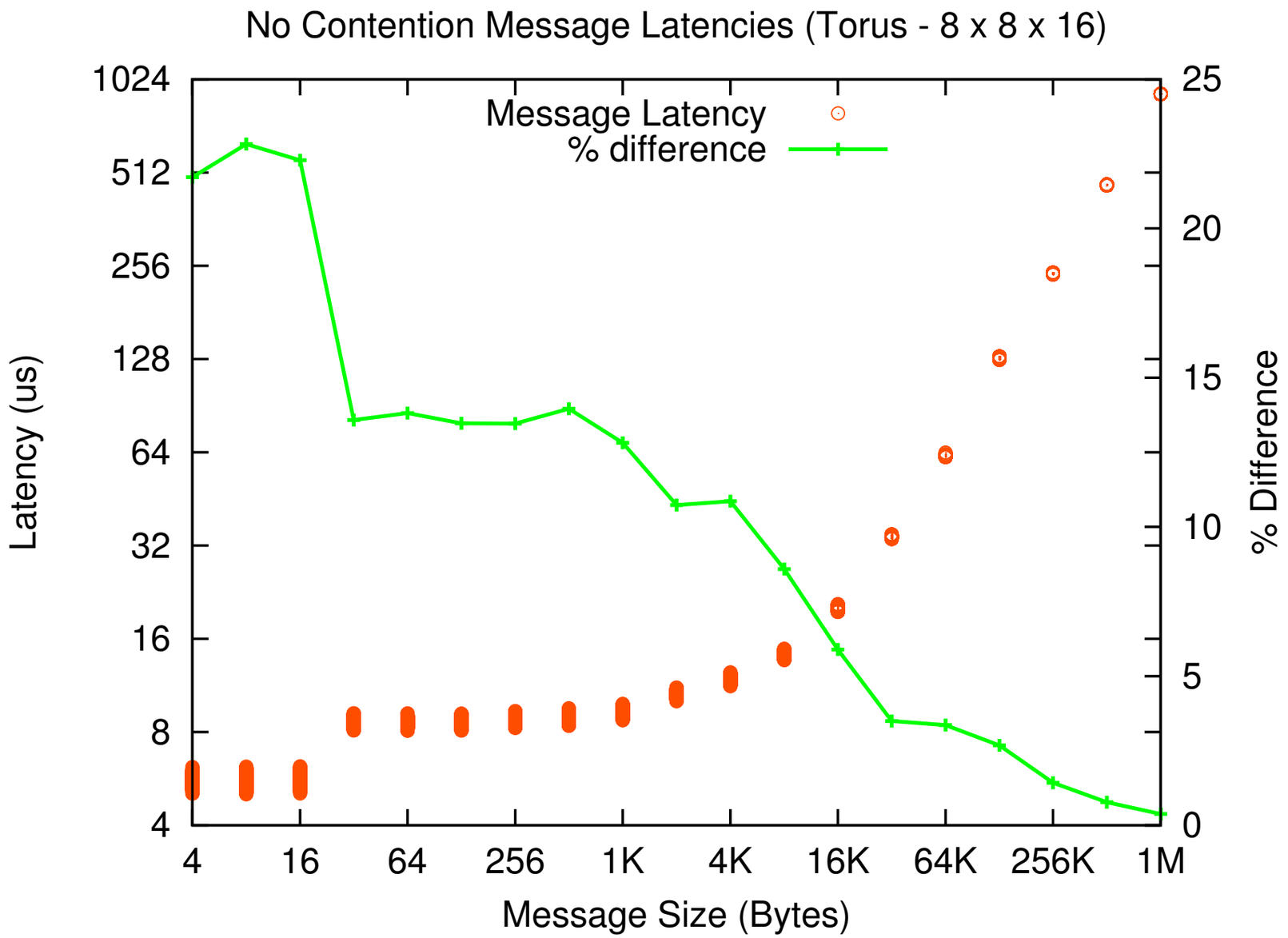
- Applications have a communication topology and processors have an interconnect topology
- Definition: Given a set of communicating parallel "entities", map them on to physical processors to optimize communication
- Goals:
  - Minimize communication traffic and hence contention
  - Balance computational load (when n > p)

No Contention Runs



Blue Gene/P

No Contention Runs



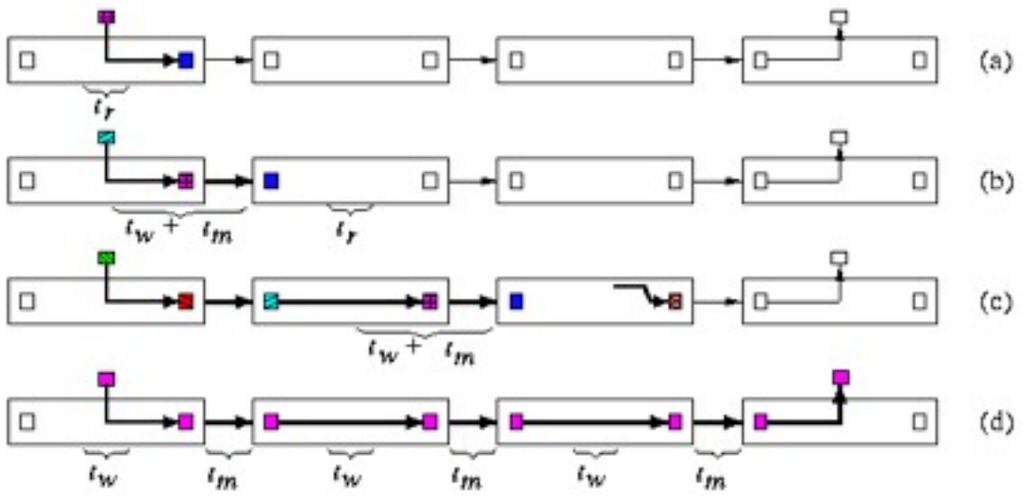
XT3 (BigBen)

Wormhole Routing

- Ni et al. 1993; Oh et al. 1997 - Equation for modeling message latencies:

$$\frac{L_f}{B} * D + \frac{L}{B}$$

L_f = length of flit, B = bandwidth,
D = hops, L = message size

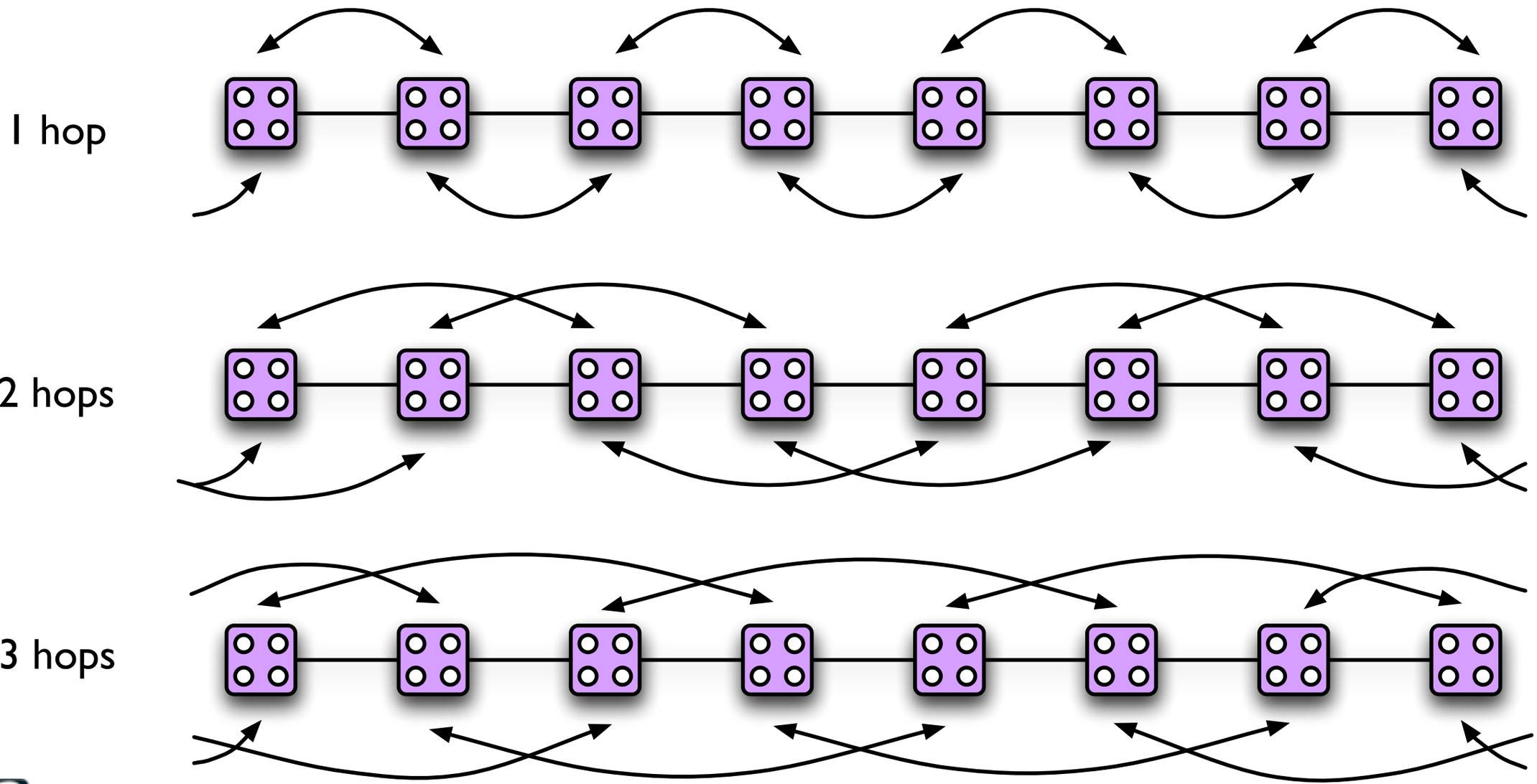


http://pages.cs.wisc.edu/~tvrdik/7/html/Section7.html

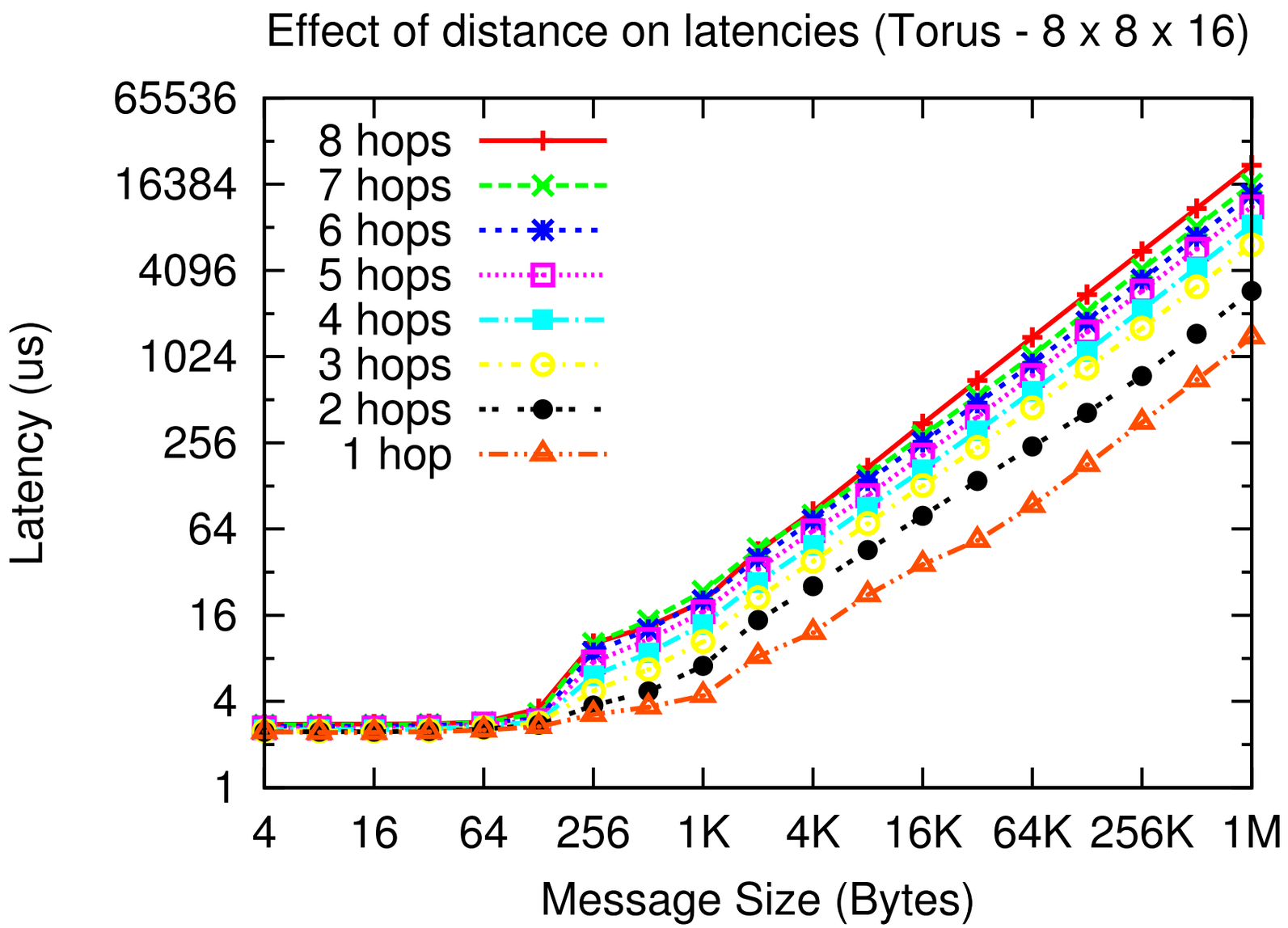
- Relatively small sized supercomputers
- It was safe to assume message latencies were independent of distance

Benchmark Creating Artificial Contention

- Pair each processor with a partner that is *n* hops away



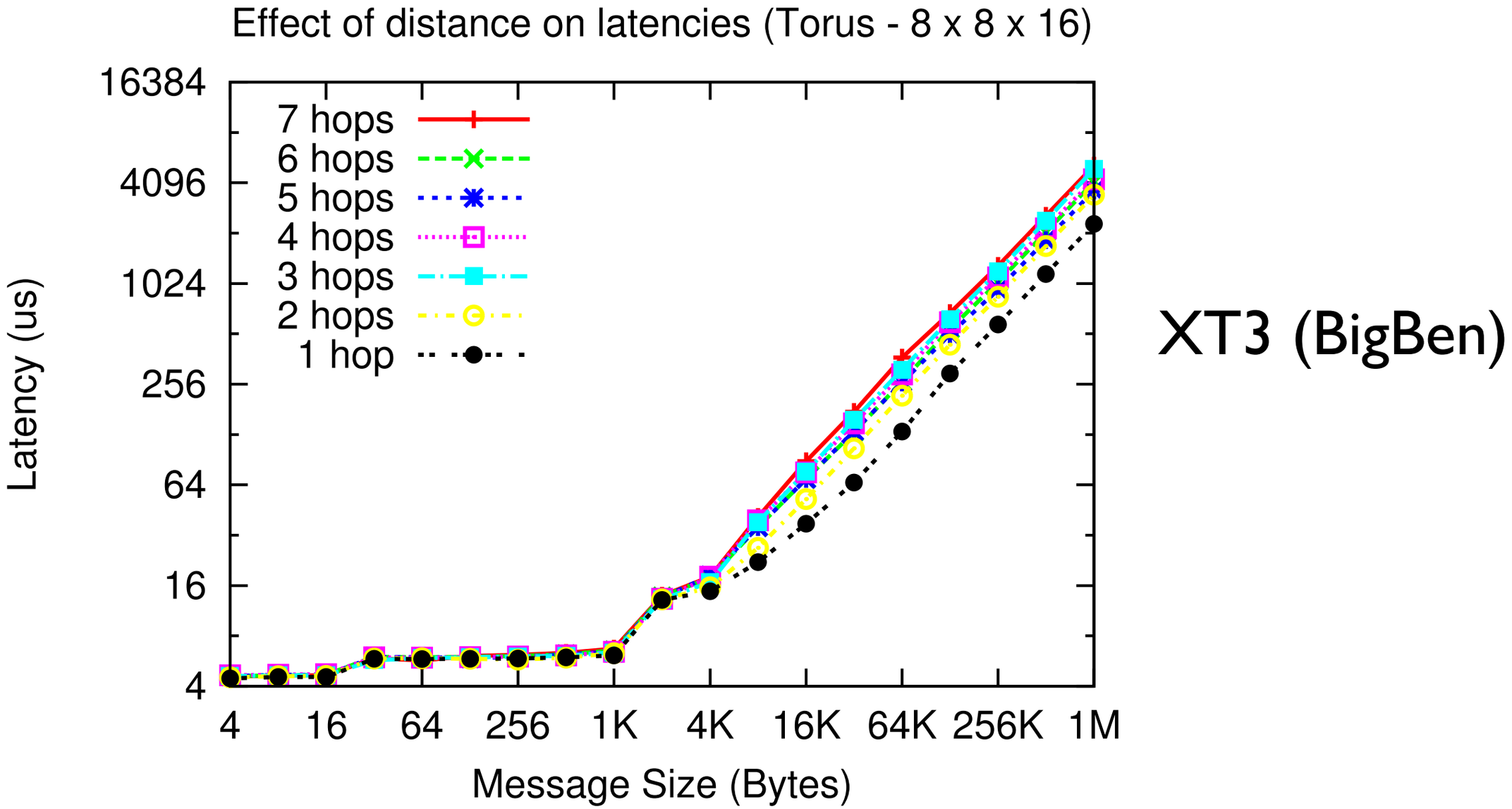
Results: Contention



Blue Gene/P

Bhatele A., Kale L.V., Quantifying Network Contention on Large Parallel Machines, *Parallel Processing Letters (Special Issue on Large-Scale Parallel Processing)*, 2009. Best Poster Award, ACM Student Research Competition, Supercomputing 2008, Austin, TX.

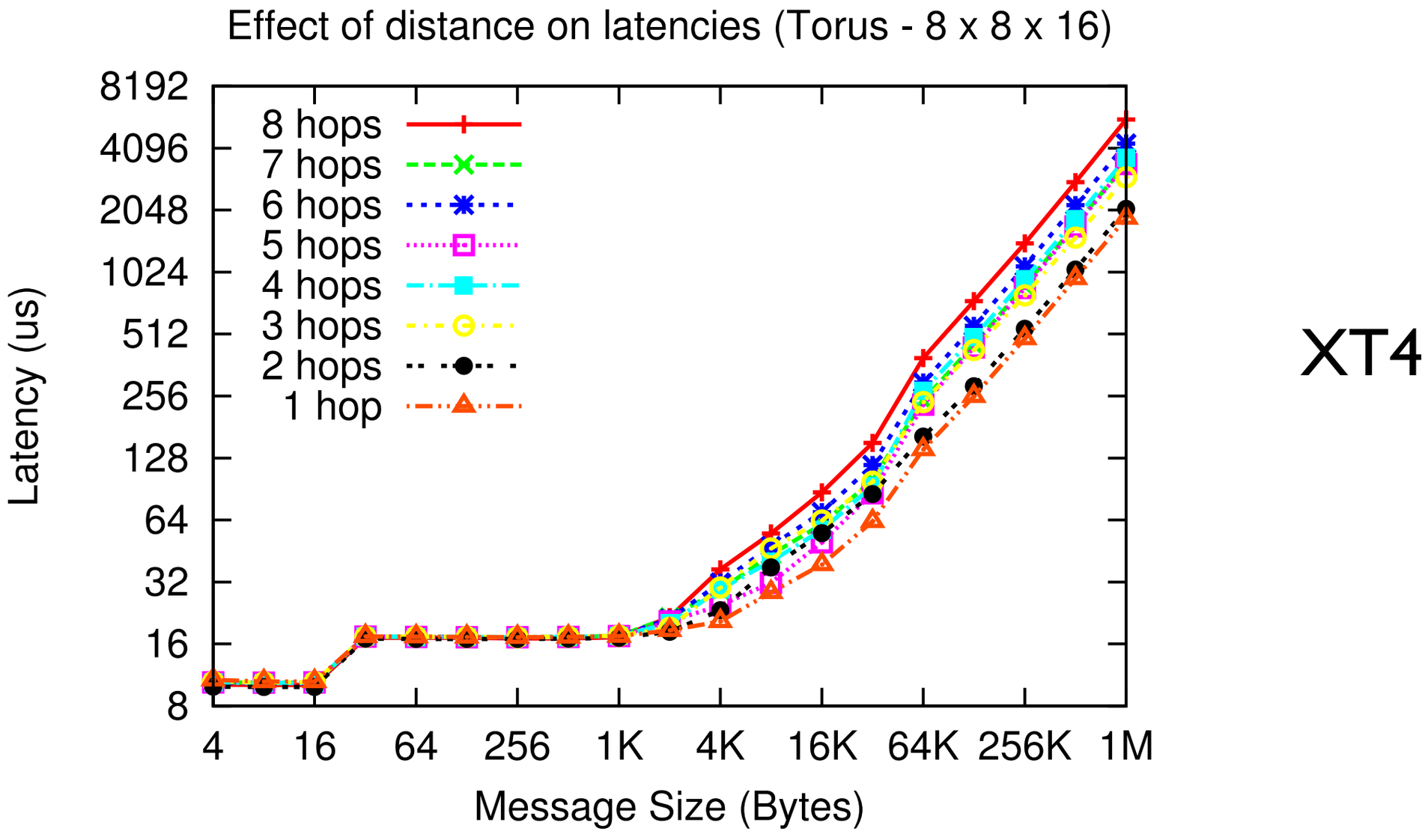
Results: Contention



XT3 (BigBen)

Bhatele A., Kale L.V., Quantifying Network Contention on Large Parallel Machines, *Parallel Processing Letters (Special Issue on Large-Scale Parallel Processing)*, 2009. Best Poster Award, ACM Student Research Competition, Supercomputing 2008, Austin, TX.

Results: Contention



XT4

Bhatele A., Kale L.V., Quantifying Network Contention on Large Parallel Machines, *Parallel Processing Letters (Special Issue on Large-Scale Parallel Processing)*, 2009. Best Poster Award, ACM Student Research Competition, Supercomputing 2008, Austin, TX.

Obtaining Topology Information

Topology Discovery

- Topology Manager API: for 3D interconnects (Blue Gene, XT)
- Information required for mapping:
  - Physical dimensions of the allocated job partition
  - Mapping of ranks to physical coordinates and vice versa
- On Blue Gene machines such information is available and the API is a wrapper
- On Cray XT machines, there is no easy way to obtain topology information

Cray XT machines

- Get nid (node ID) corresponding to an MPI rank:
  - XT3: cnos_get_nidpid_map
  - XT4/5: PMI_Portals_get_nid
- Get physical coordinates corresponding to nid:
  - rca_get_meshcoord
- Translate the origin and provide this information through the Topology Manager API

Bigben @ PSC

- Bigben: The first Cray XT3 system in the world
  - Officially unveiled on July 20, 2005 (ranked 44 in the top500 list) and decommissioned on March 31, 2010
  - Initially had 2.4 GHz single core Opterons (upgraded to 2.6 GHz dual-core nodes in late 2006) - 4,180 cores 21.5 TF
  - SeaStar interconnect (3D torus of size 11 X 12 X 16)

PITTSBURGH SUPERCOMPUTING CENTER

Bigben @ PSC

- Bigben: The first Cray XT3 system in the world
  - Officially unveiled on July 20, 2005 (ranked 44 in the top500 list) and decommissioned on March 31, 2010
  - Initially had 2.4 GHz single core Opterons (upgraded to 2.6 GHz dual-core nodes in late 2006) - 4,180 cores 21.5 TF
  - SeaStar interconnect (3D torus of size 11 X 12 X 16)

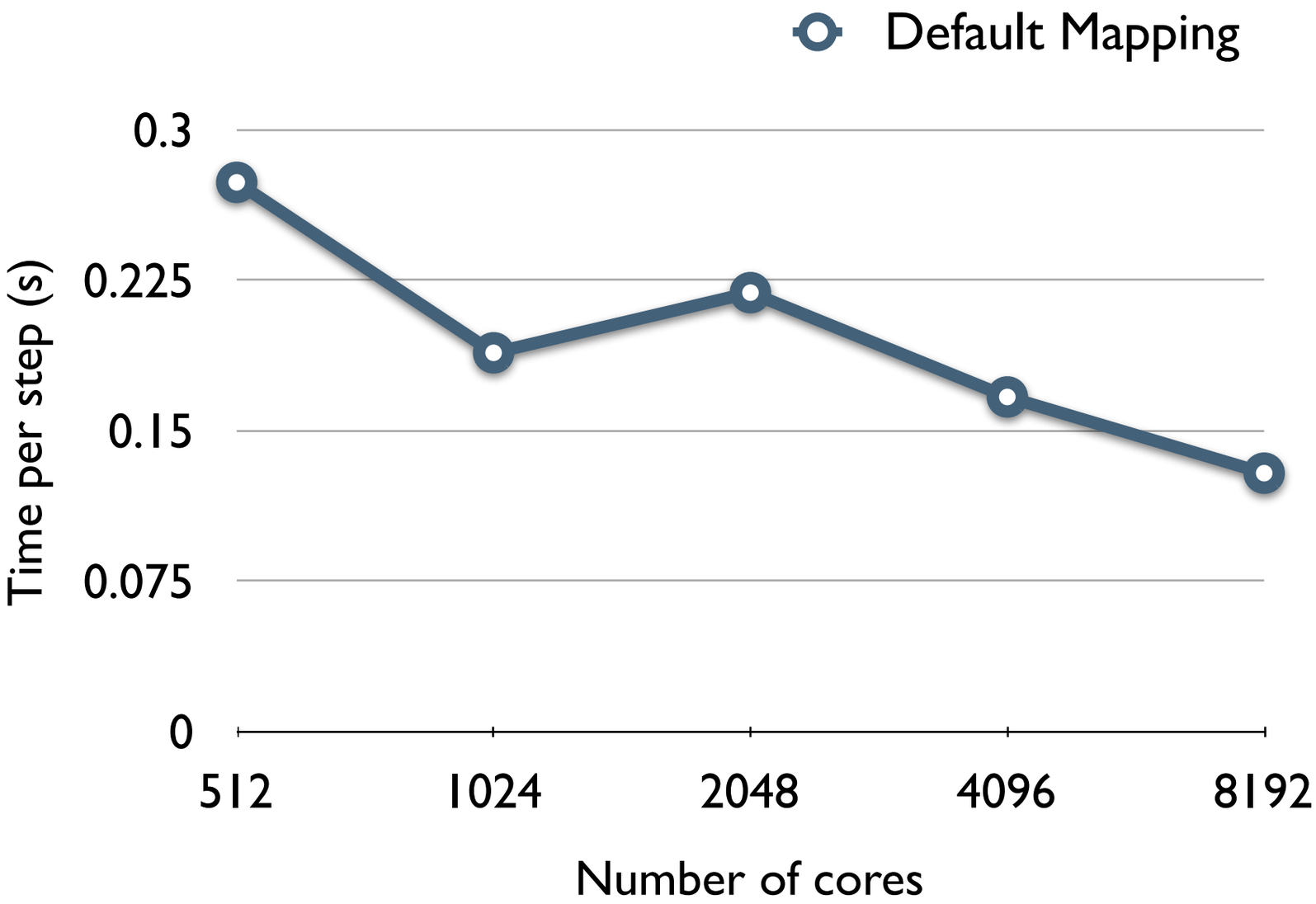
PITTSBURGH SUPERCOMPUTING CENTER

Thanks to Chad Vizino and Shawn Brown

Application Case Studies

Case Study 1: OpenAtom

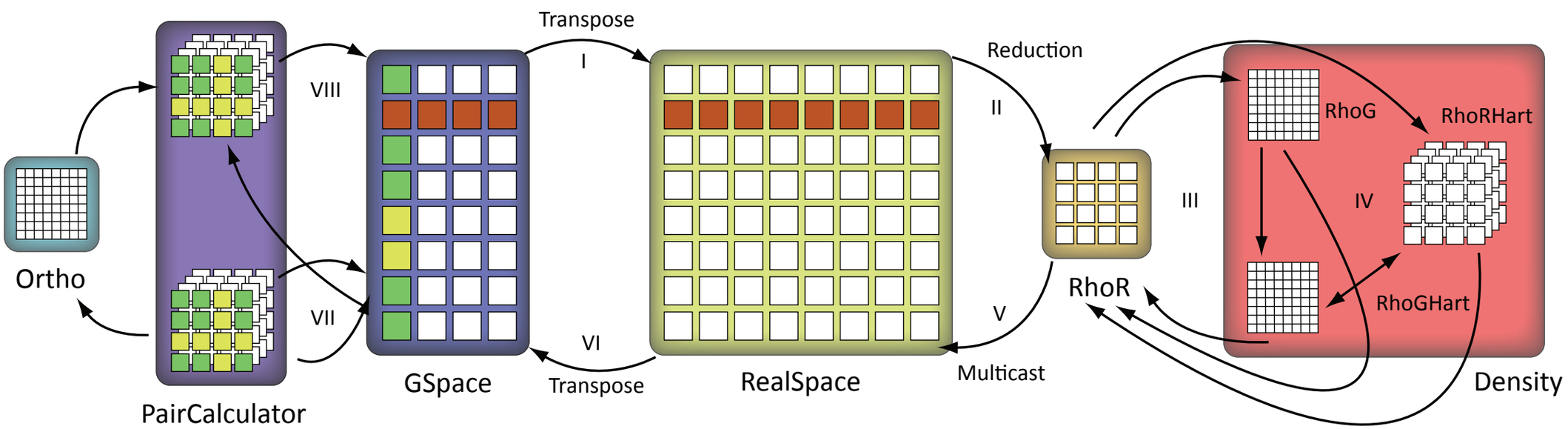
Performance on Blue Gene/L



Diagnosis

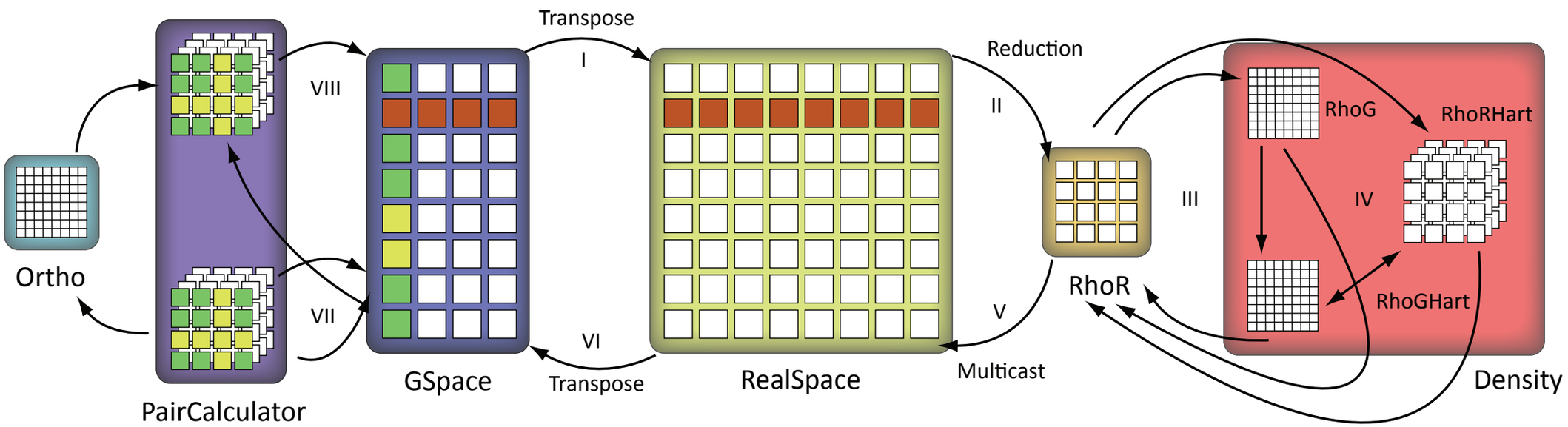
Timeline view (OpenAtom on 8,192 cores of BG/L) using the performance visualization tool, Projections

Mapping of OpenAtom Arrays



A. Bhatele, E. Bohm, and L.V. Kale. A Case Study of Communication Optimizations on 3D Mesh Interconnects. In Euro-Par, LNCS 5704, pages 1015–1028, 2009. Distinguished Paper Award, Feng Chen Memorial Best Paper Award

Mapping of OpenAtom Arrays

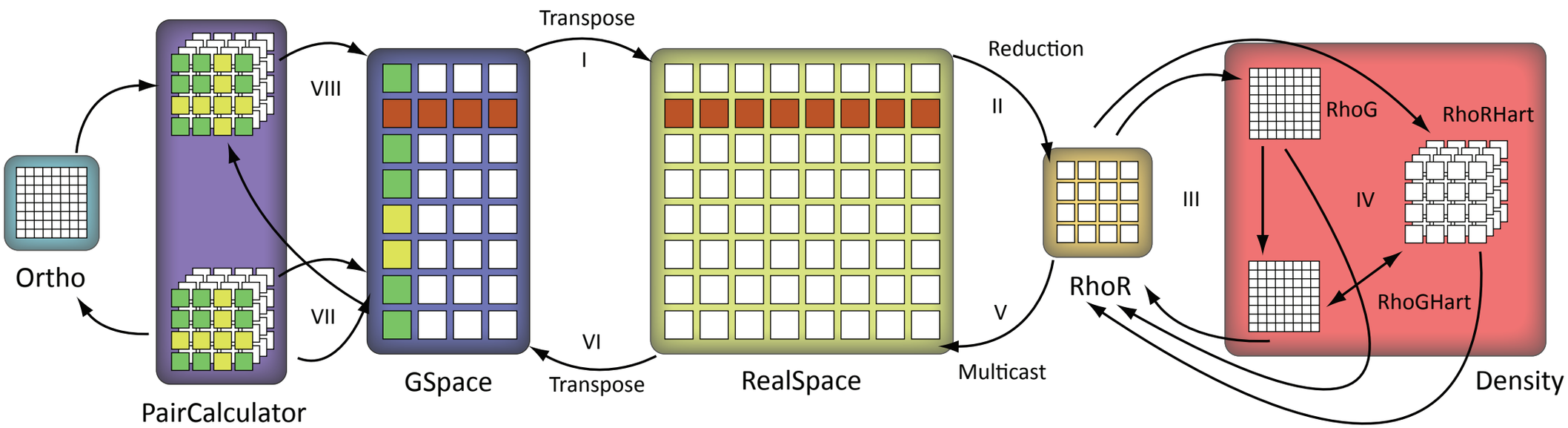


Paircalculator and GSpace have plane-wise communication

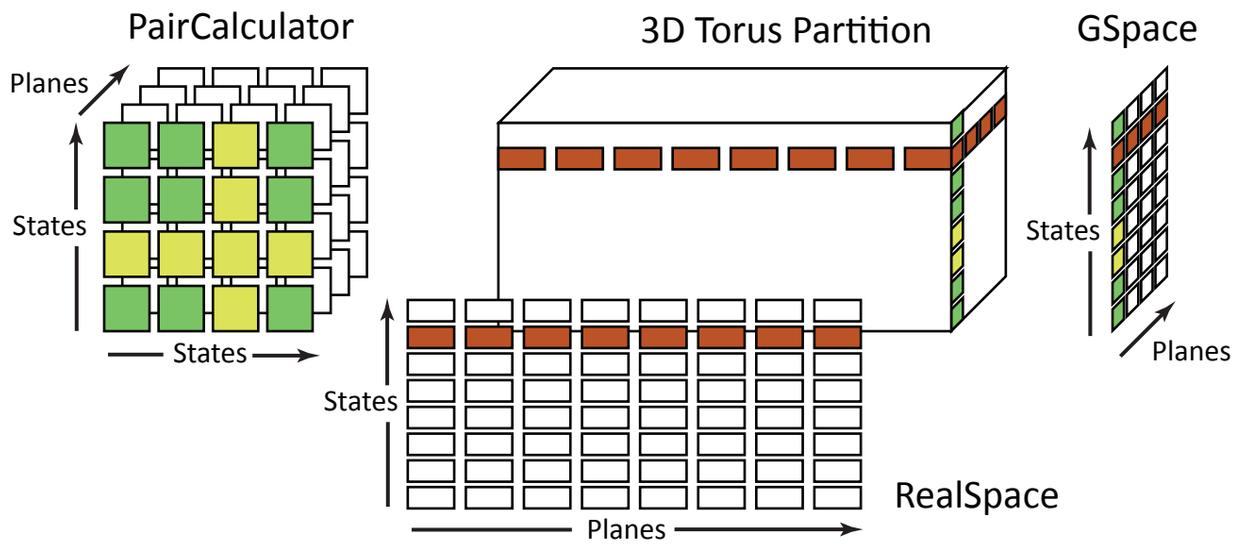
RealSpace and GSpace have state-wise communication

A. Bhatele, E. Bohm, and L.V. Kale. A Case Study of Communication Optimizations on 3D Mesh Interconnects. In Euro-Par, LNCS 5704, pages 1015–1028, 2009. Distinguished Paper Award, Feng Chen Memorial Best Paper Award

Mapping of OpenAtom Arrays



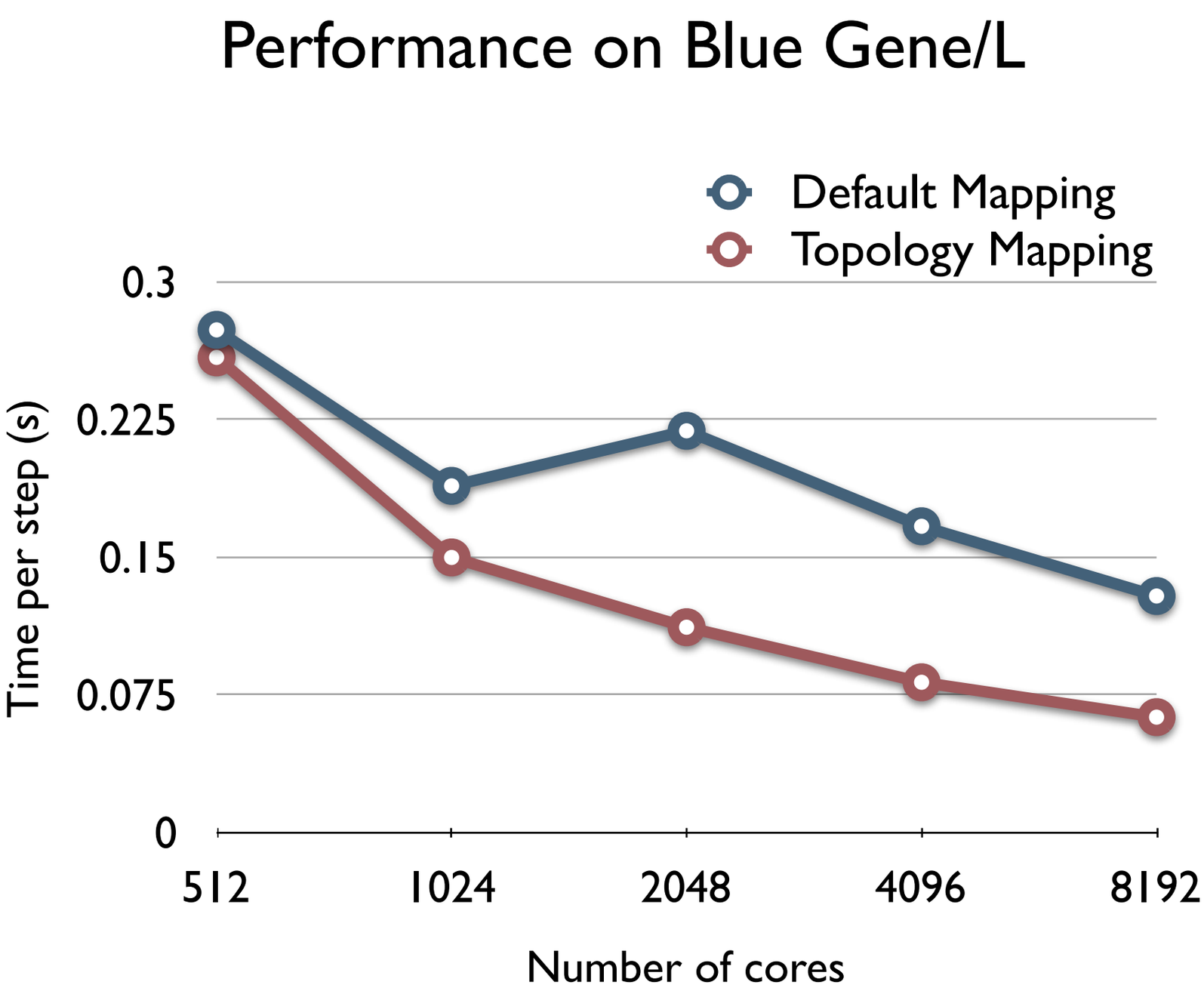
Paircalculator and GSpace have plane-wise communication



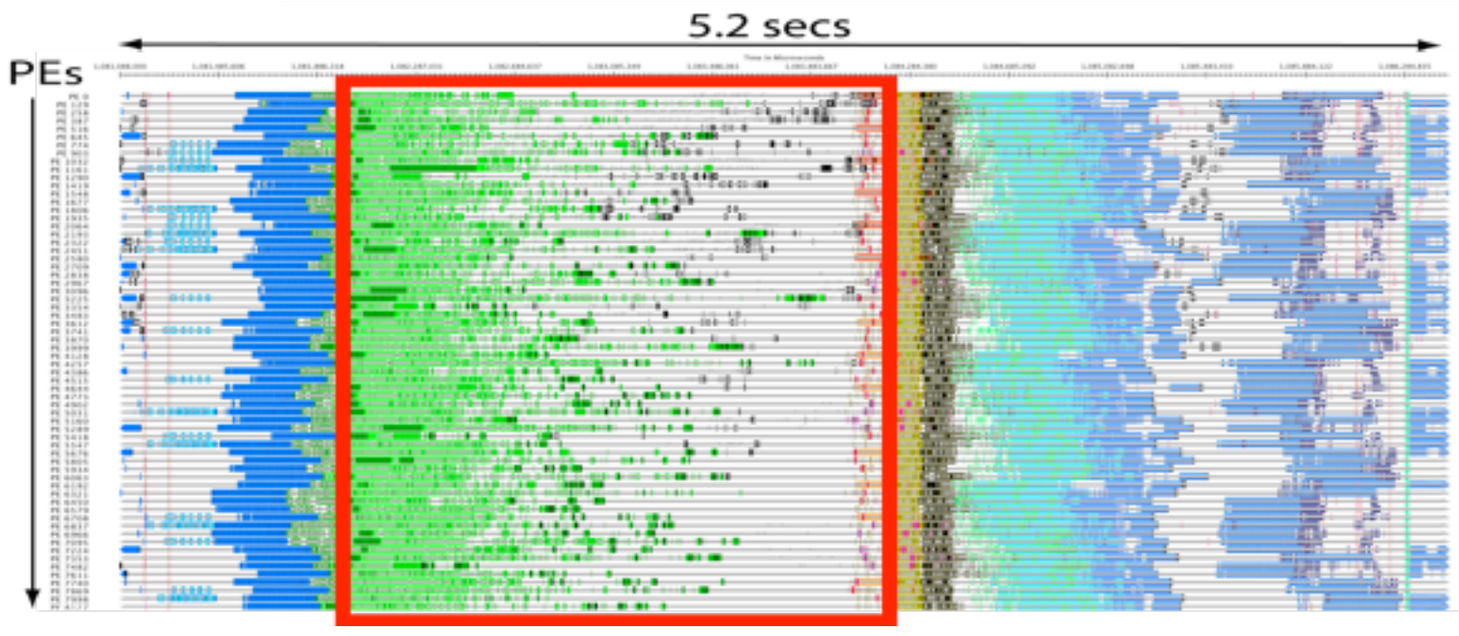
RealSpace and GSpace have state-wise communication

A. Bhatele, E. Bohm, and L.V. Kale. A Case Study of Communication Optimizations on 3D Mesh Interconnects. In Euro-Par, LNCS 5704, pages 1015–1028, 2009. Distinguished Paper Award, Feng Chen Memorial Best Paper Award

Performance Benefits from Mapping



Diagnosis of Improvement

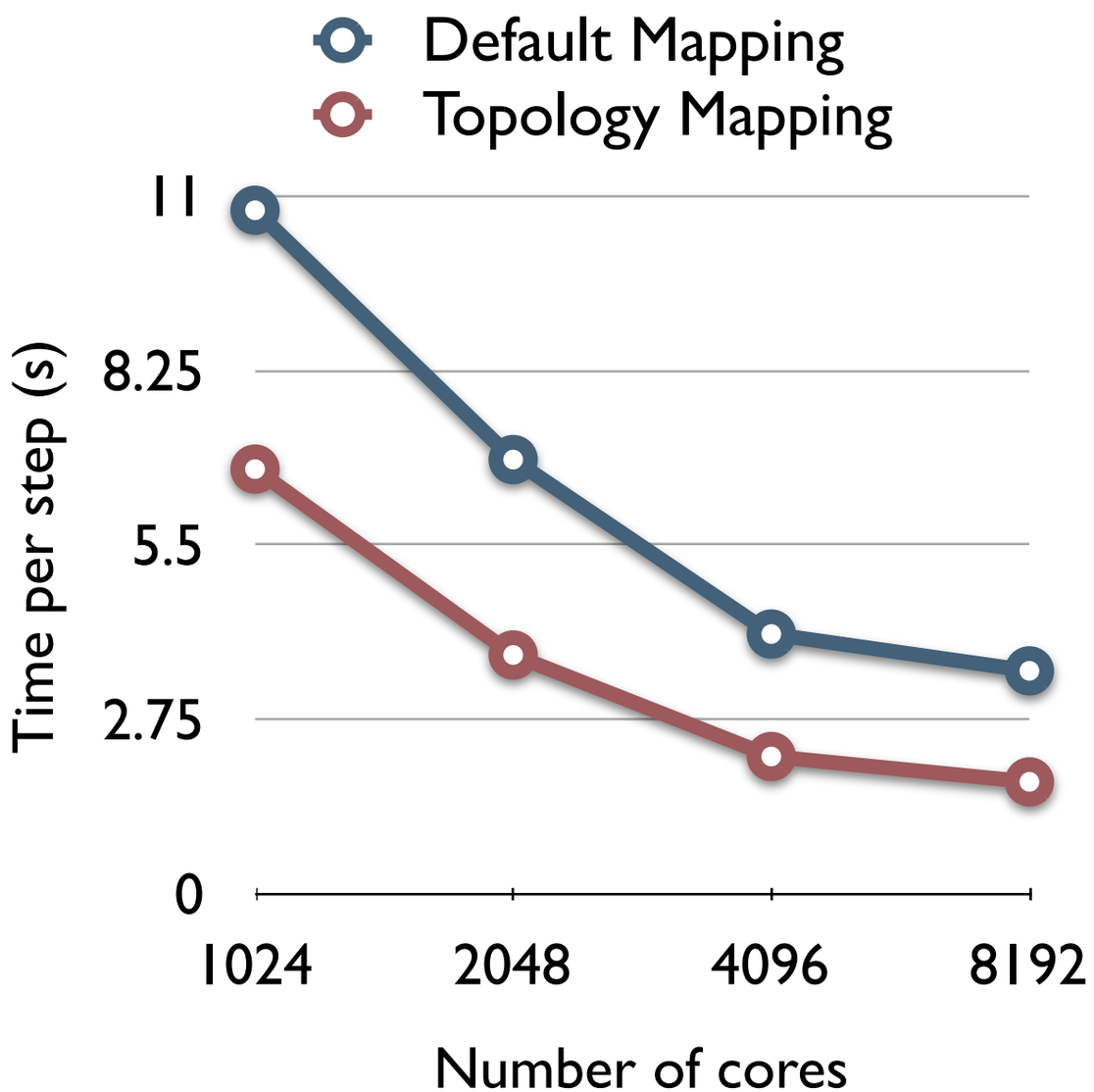


Timeline of 1 iteration of OpenAtom running WATER_256M_70Ry on 8192 cores of BG/L

Timeline view using the performance visualization tool, Projections

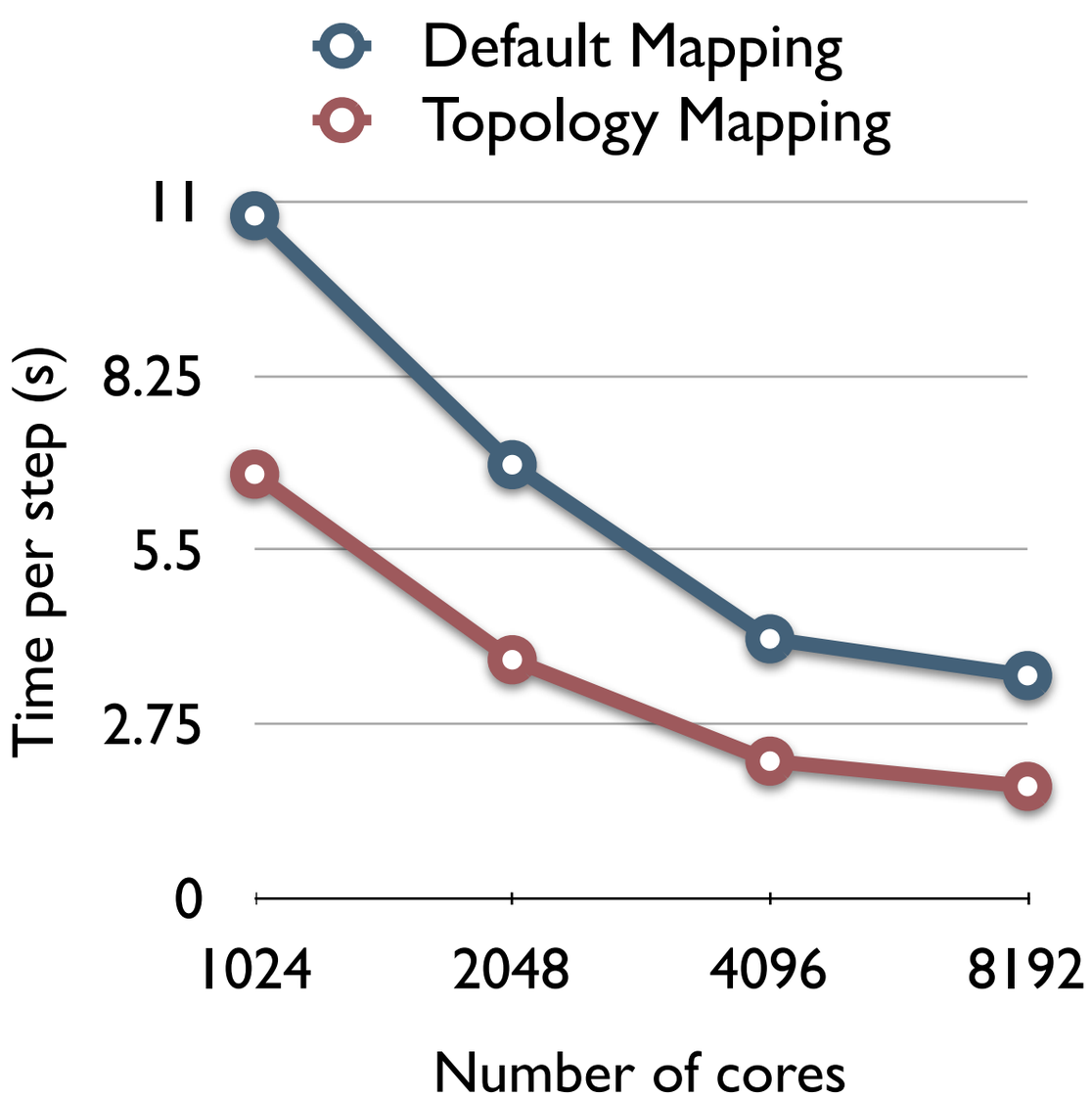
OpenAtom Performance on Blue Gene/P

Application Performance

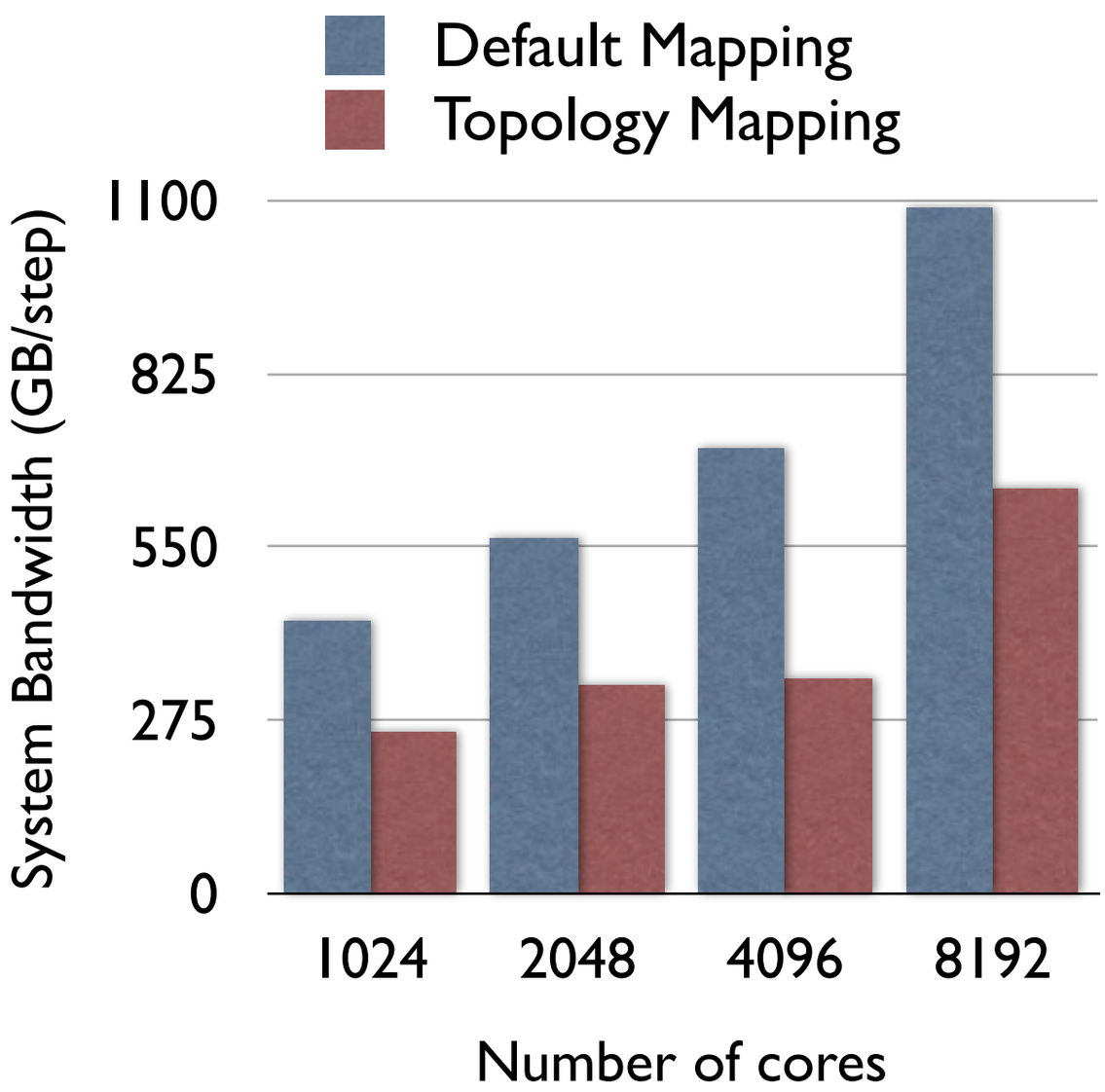


OpenAtom Performance on Blue Gene/P

Application Performance

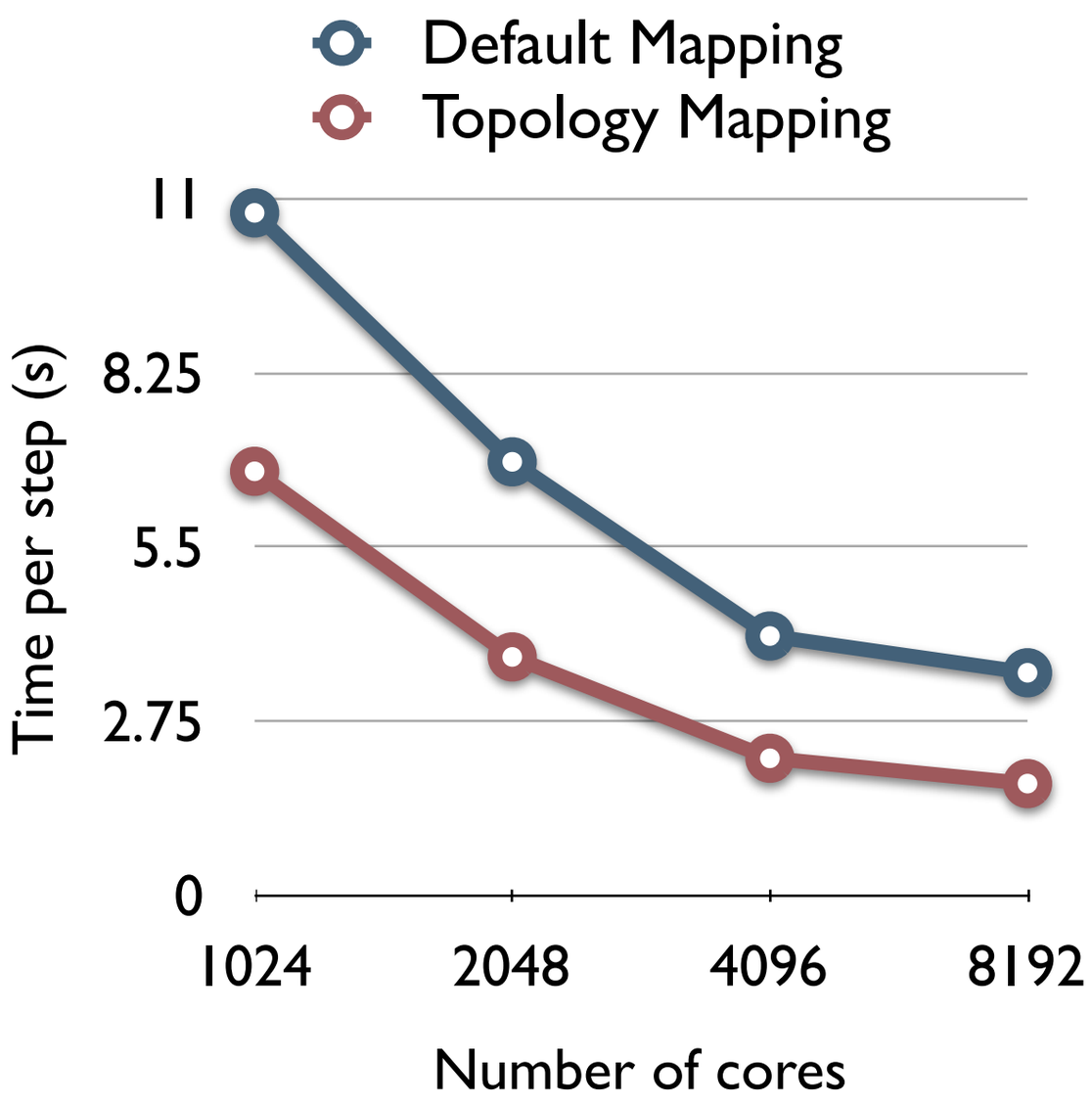


Performance Counters

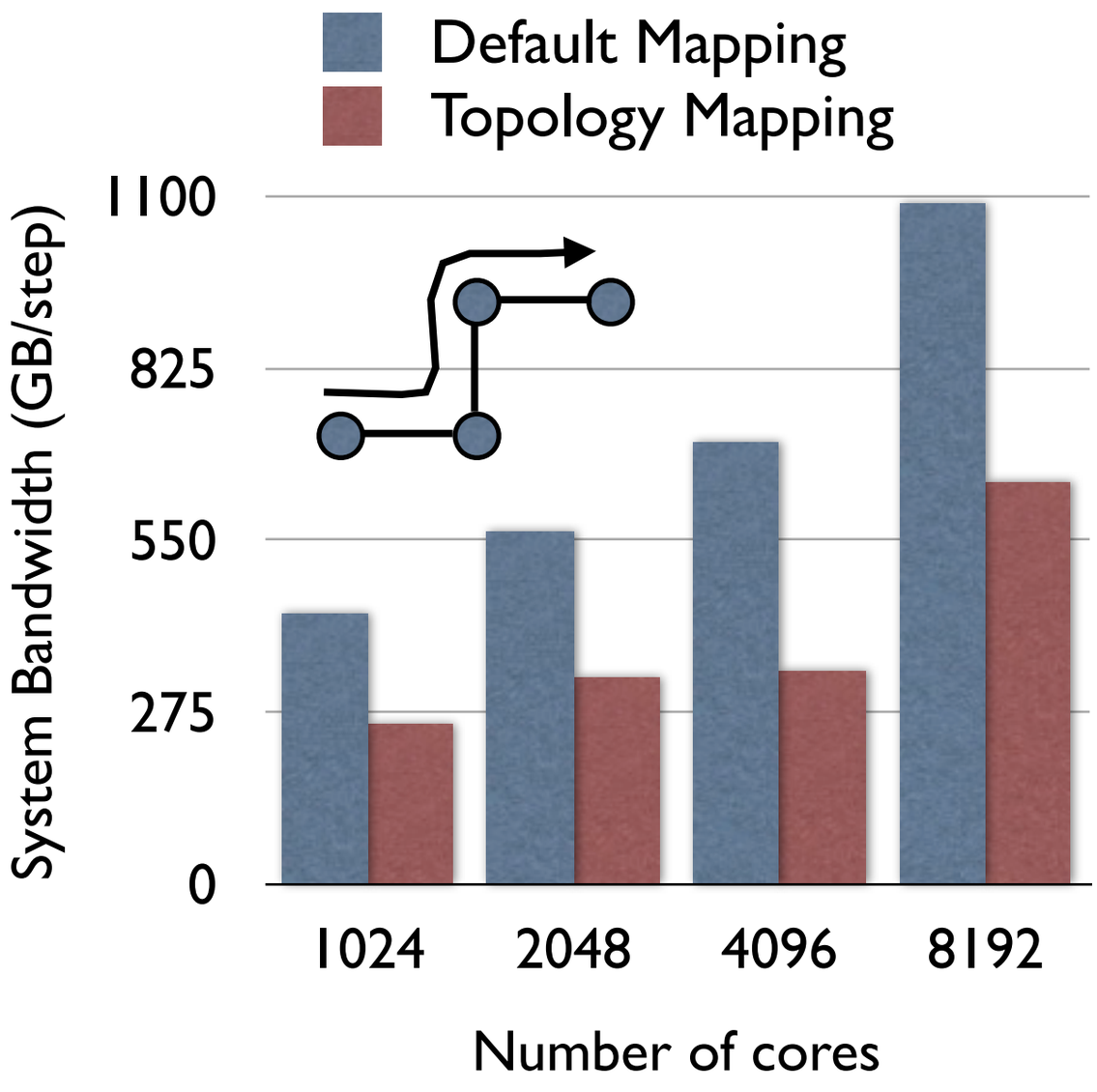


OpenAtom Performance on Blue Gene/P

Application Performance



Performance Counters



OpenAtom Performance on Cray XT3

OpenAtom Performance on Cray XT3

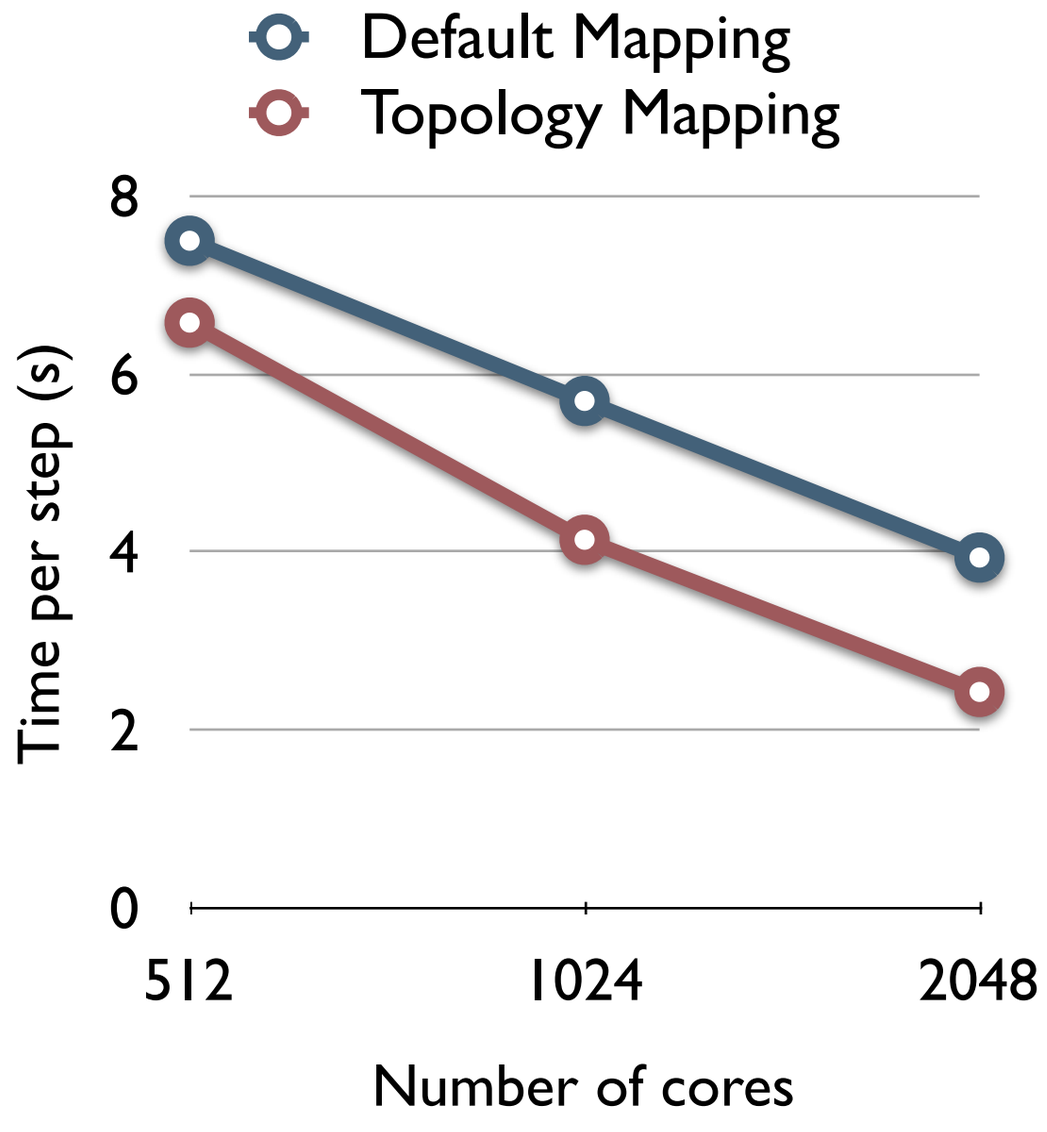
- Cray XT3:
  - Link bandwidth - 3.8 GB/s (XT3), 0.425 (BG/P), 0.175 (BG/L)
  - Bytes per flop - 8.77 (XT3), 0.375 (BG/P and BG/L)

OpenAtom Performance on Cray XT3

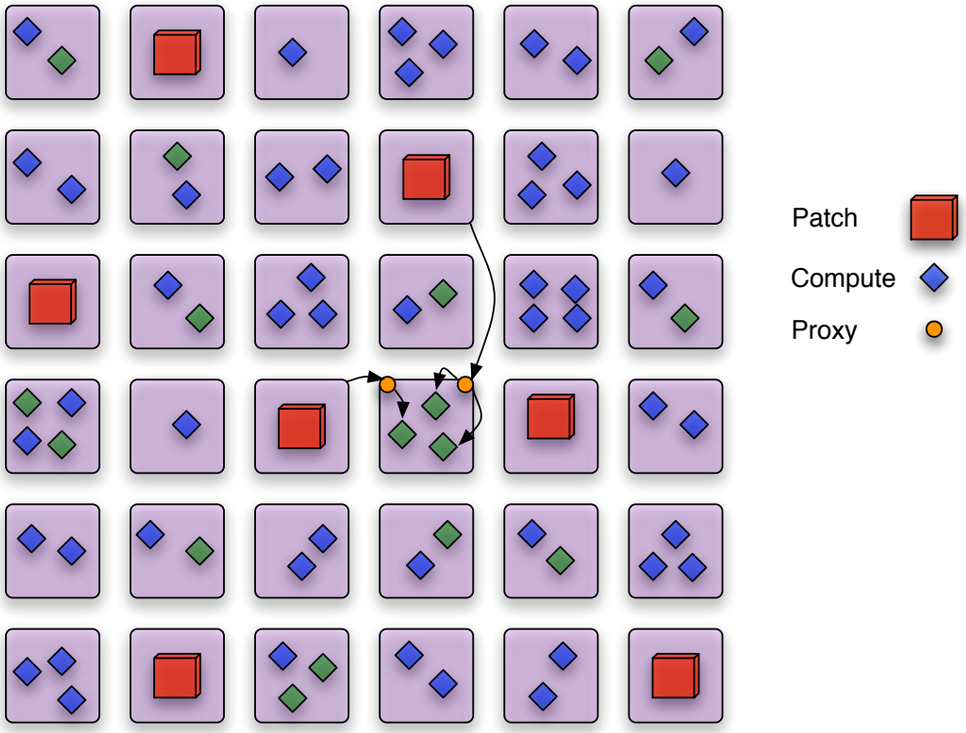
- Cray XT3:
  - Link bandwidth - 3.8 GB/s (XT3), 0.425 (BG/P), 0.175 (BG/L)
  - Bytes per flop - 8.77 (XT3), 0.375 (BG/P and BG/L)
- Job schedulers on Cray are not topology aware

OpenAtom Performance on Cray XT3

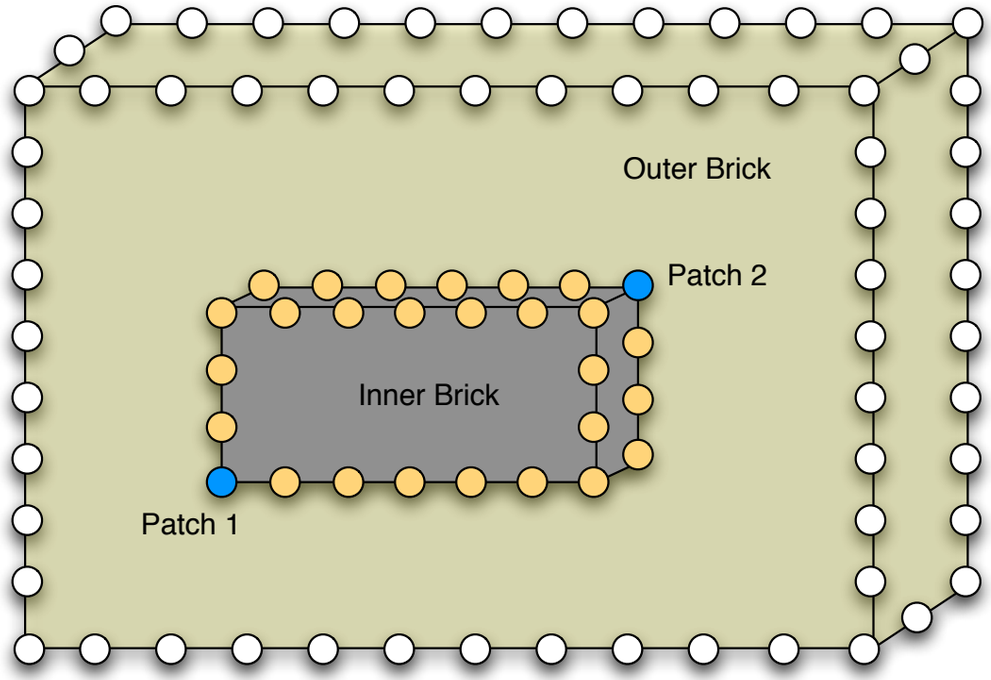
- Cray XT3:
  - Link bandwidth - 3.8 GB/s (XT3), 0.425 (BG/P), 0.175 (BG/L)
  - Bytes per flop - 8.77 (XT3), 0.375 (BG/P and BG/L)
- Job schedulers on Cray are not topology aware
- Performance Benefit at 2048 cores: 40% (XT3), 45% (BG/P), 41% (BG/L)



Case Study II: NAMD



Communication between patches and computes

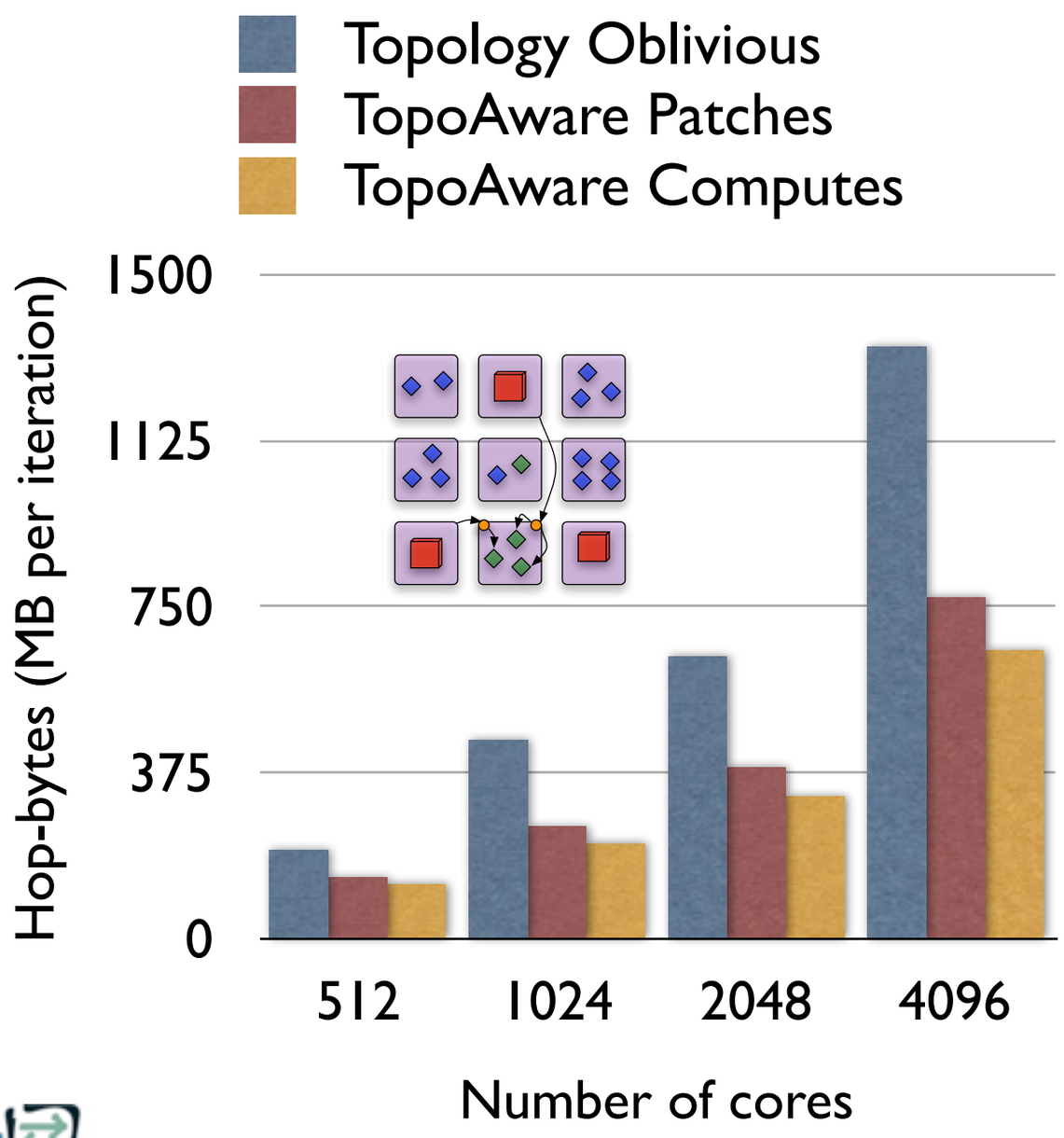


Topology aware placement of computes

A. Bhatele, L.V. Kale and S. Kumar, Dynamic Topology Aware Load Balancing Algorithms for Molecular Dynamics Applications, In 23rd ACM International Conference on Supercomputing (ICS), 2009.

NAMD Performance on Blue Gene/P

Measured Hop-bytes



- Evaluation Metric: Hop-bytes

$$HB = \sum_{i=1}^{n} d_i \times b_i$$

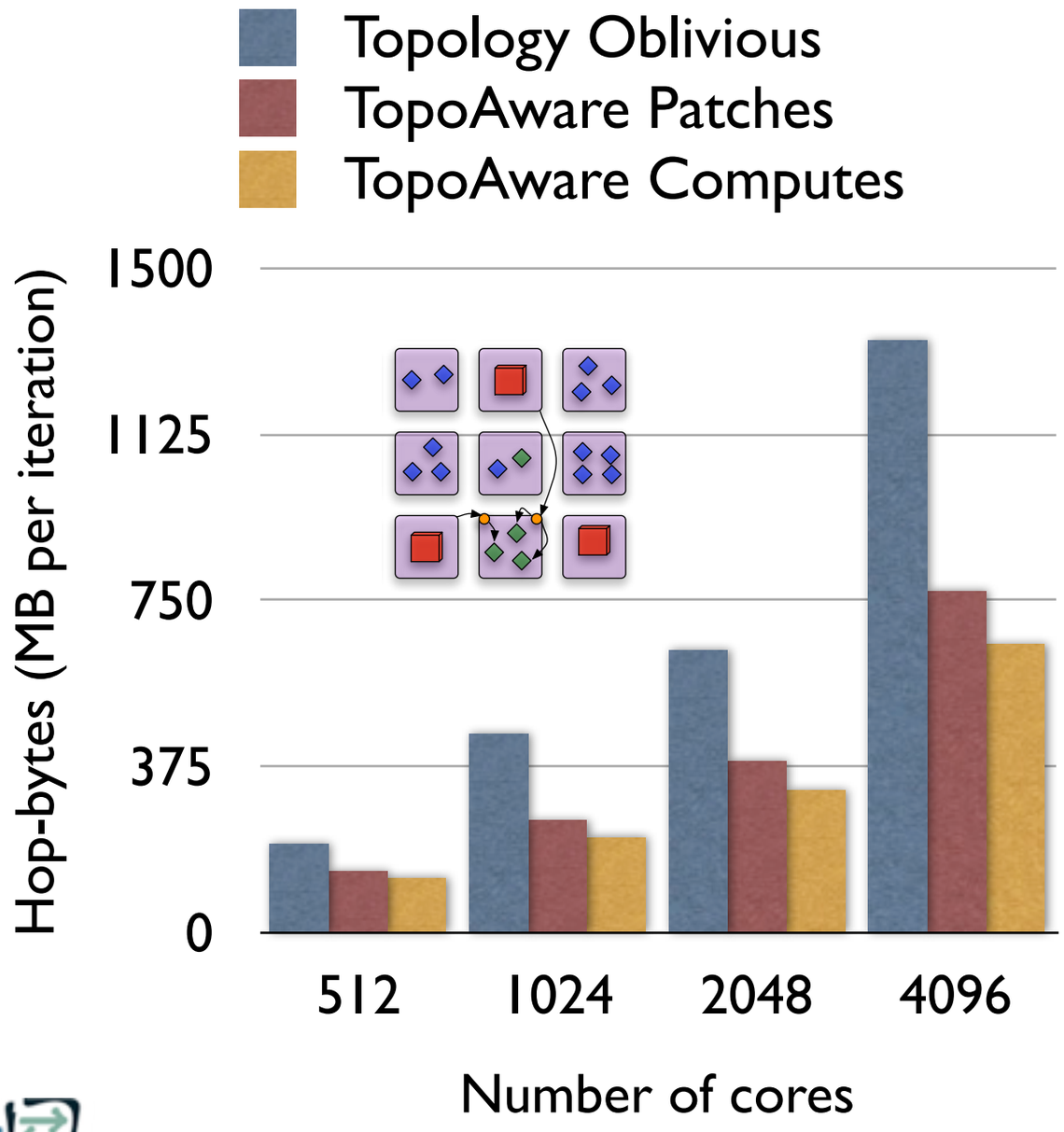
d_i = distance
b_i = bytes
n = no. of messages

- Indicates amount of traffic and hence contention on the network
- Previously used metric: maximum dilation

$$d(e) = max\{d_i | e_i \in E\}$$

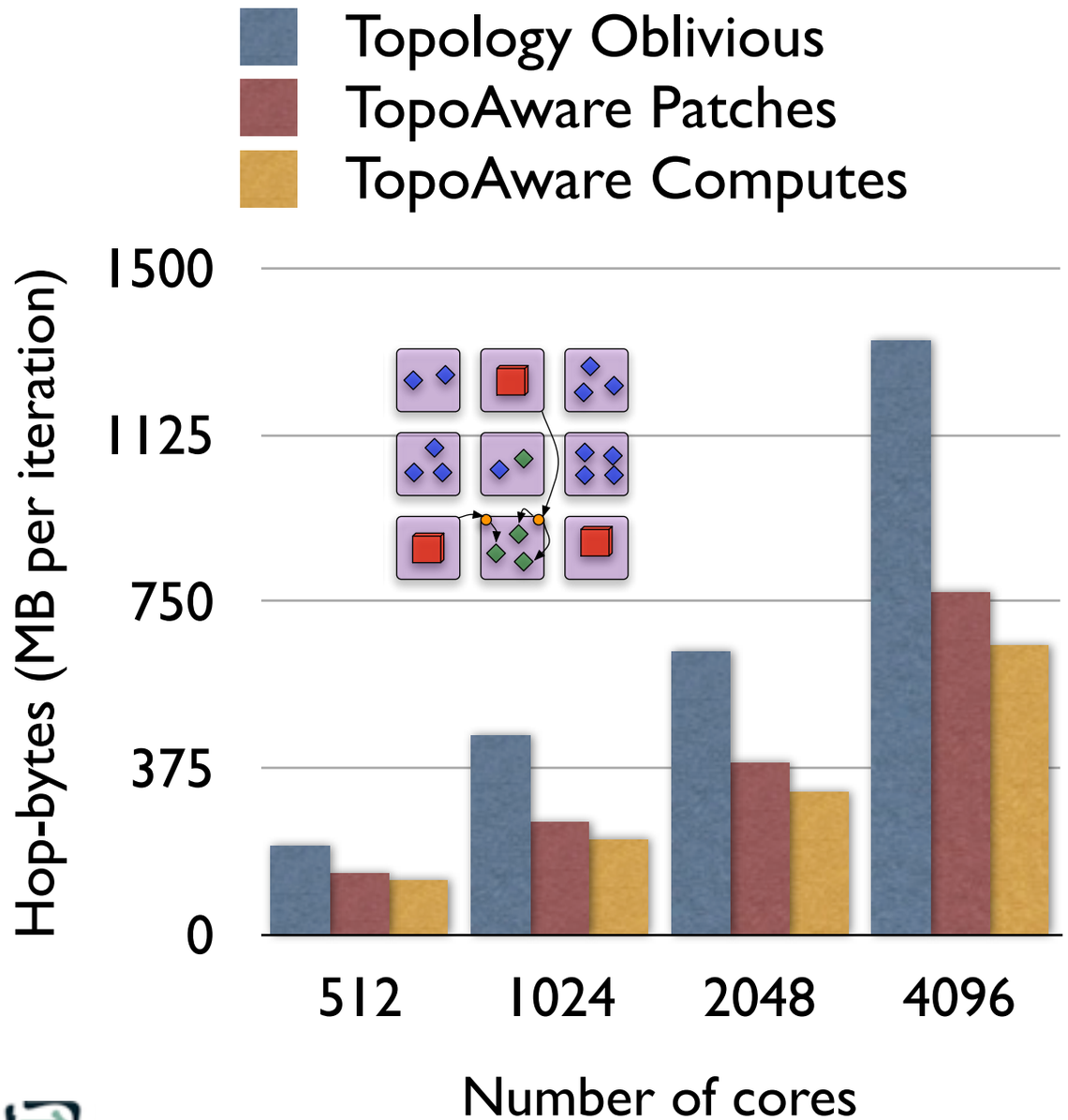
NAMD Performance on Blue Gene/P

Measured Hop-bytes

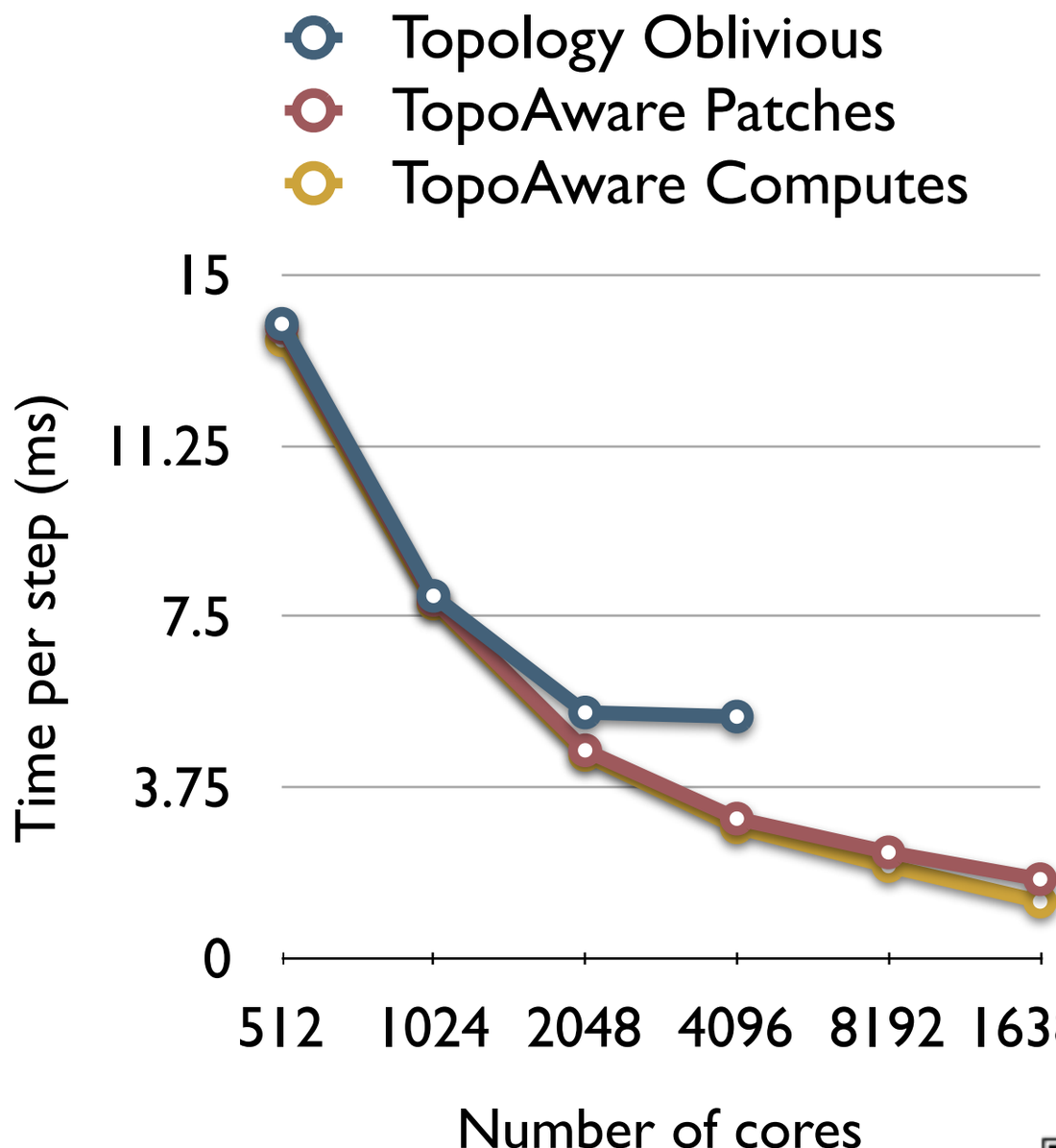


NAMD Performance on Blue Gene/P

Measured Hop-bytes

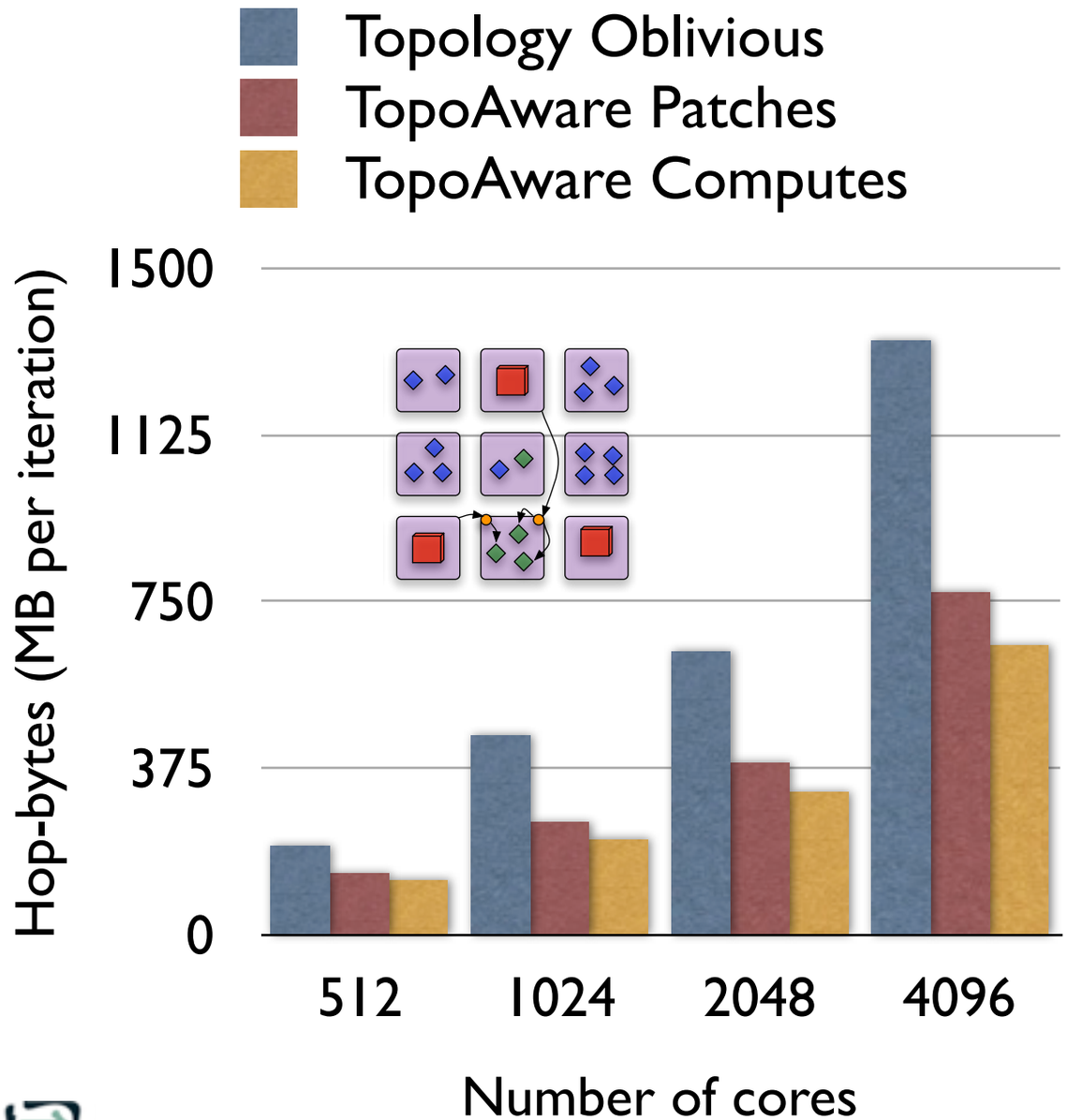


Application Performance

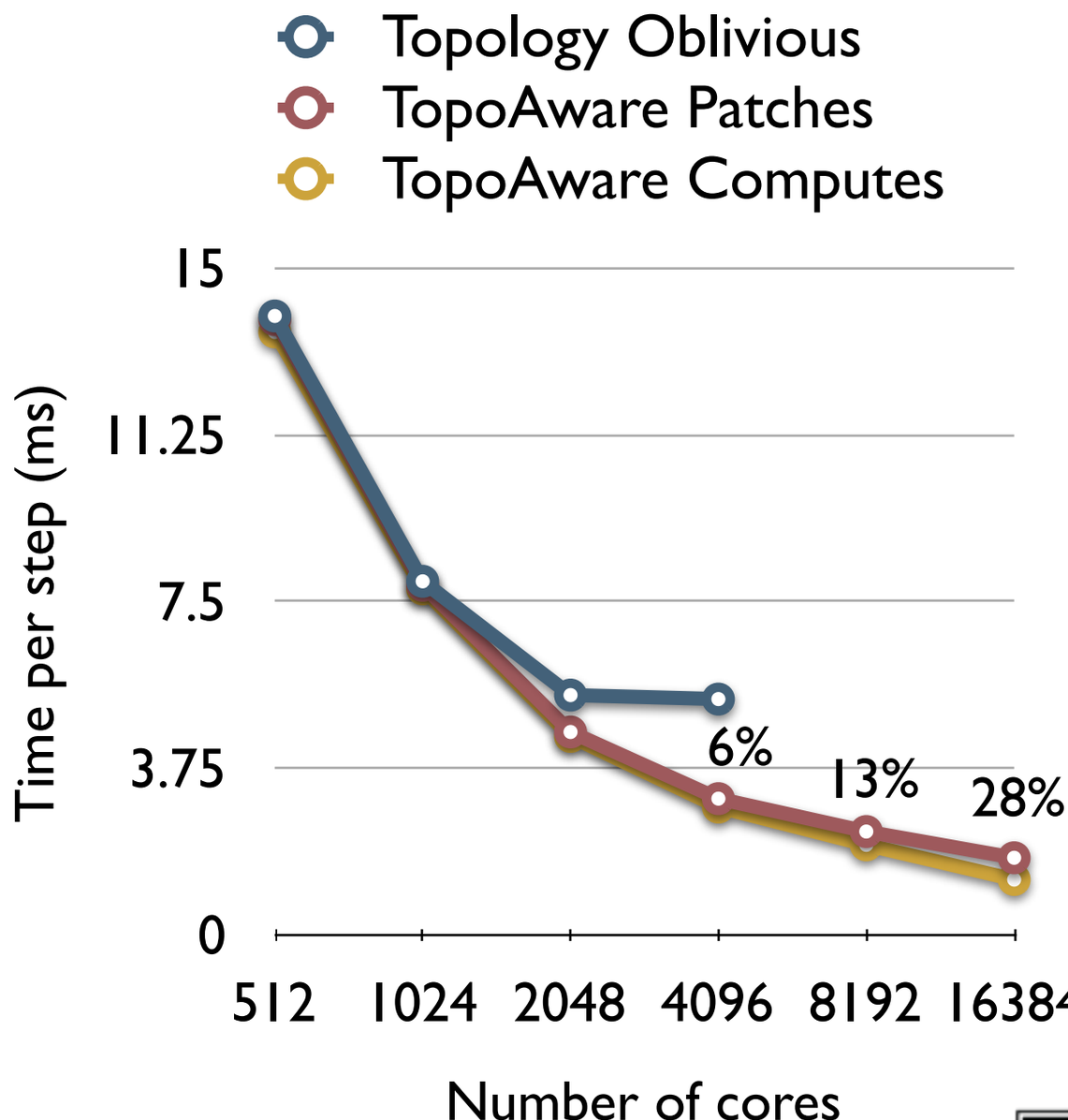


NAMD Performance on Blue Gene/P

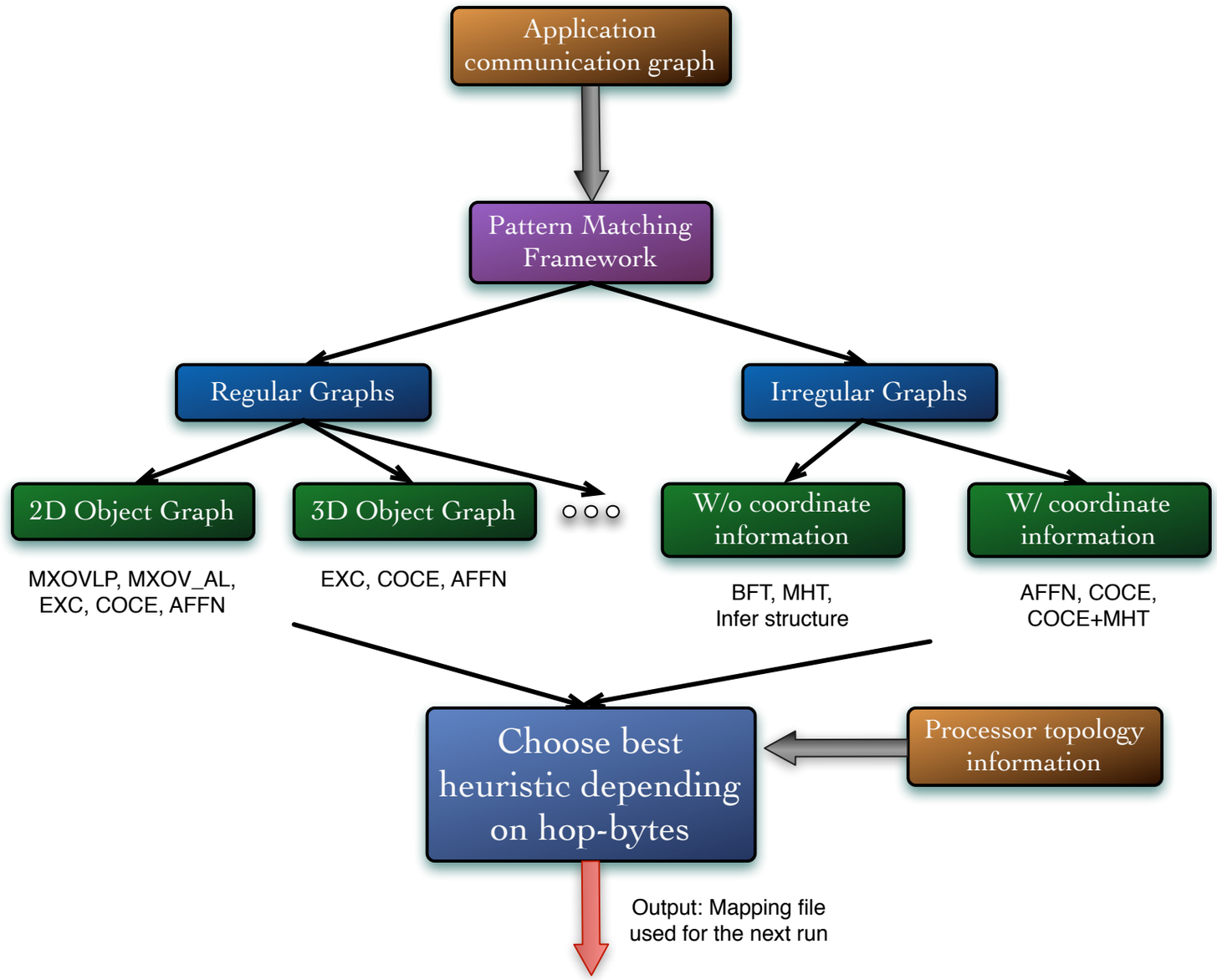
Measured Hop-bytes



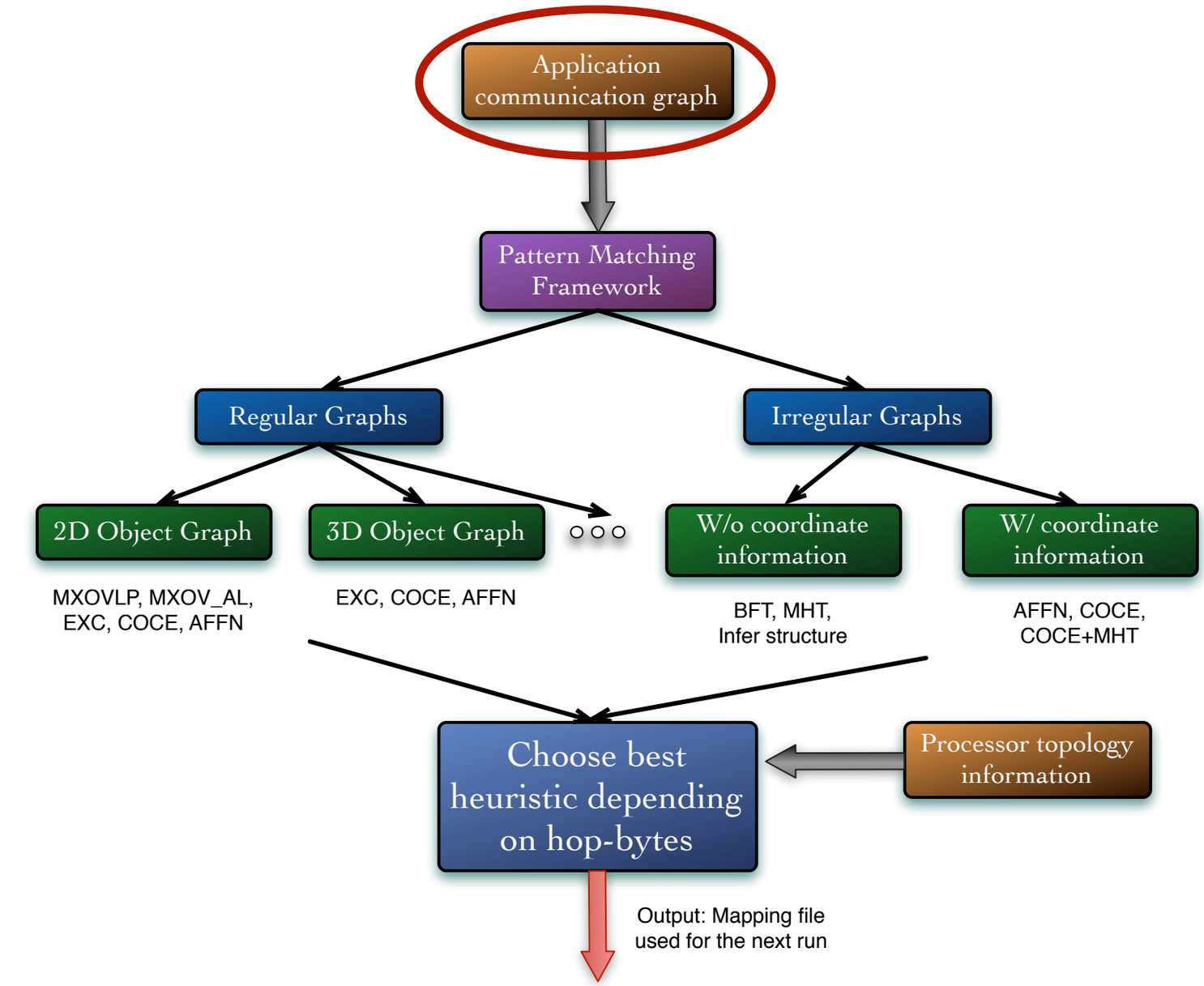
Application Performance



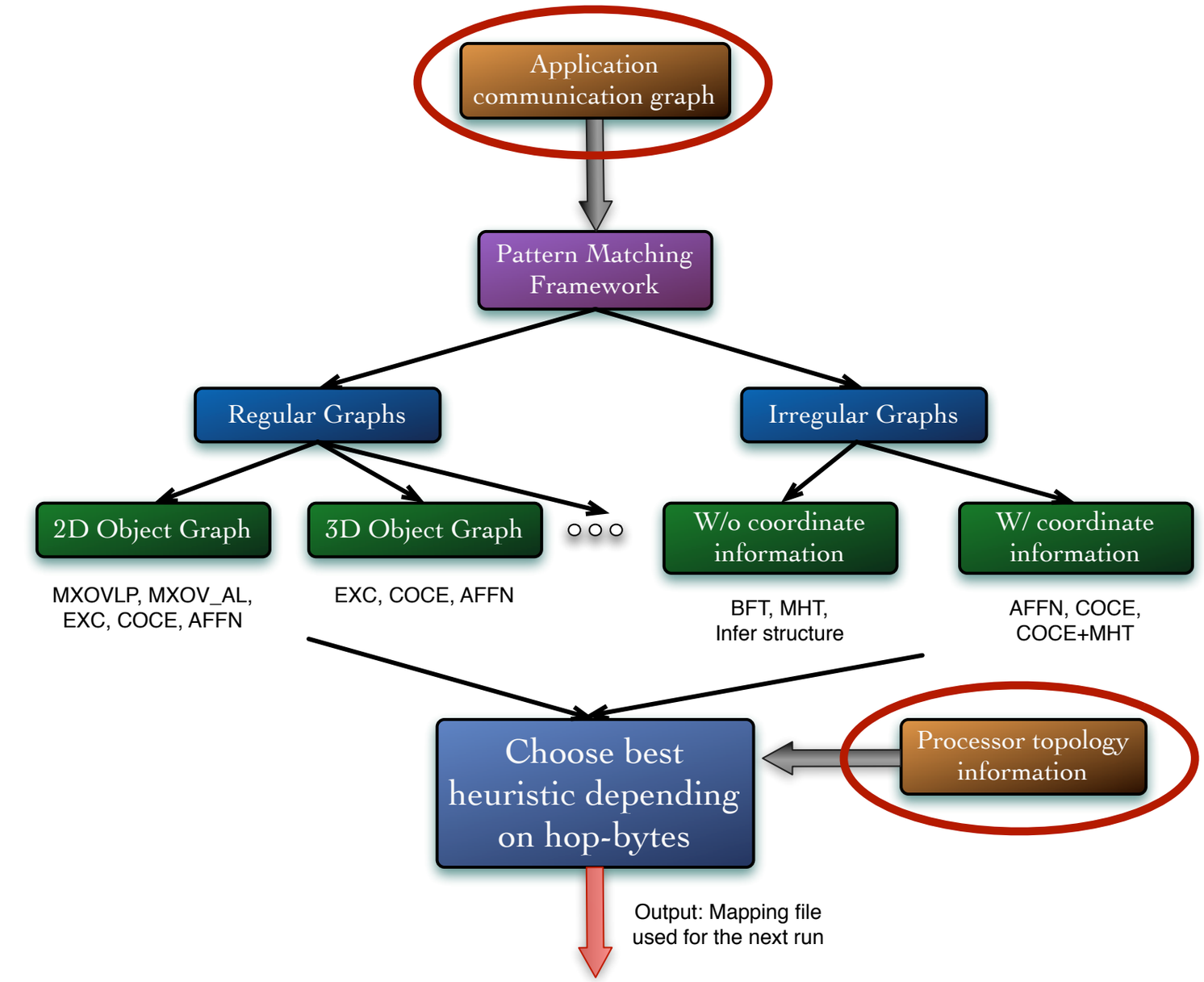
Automatic Mapping Framework



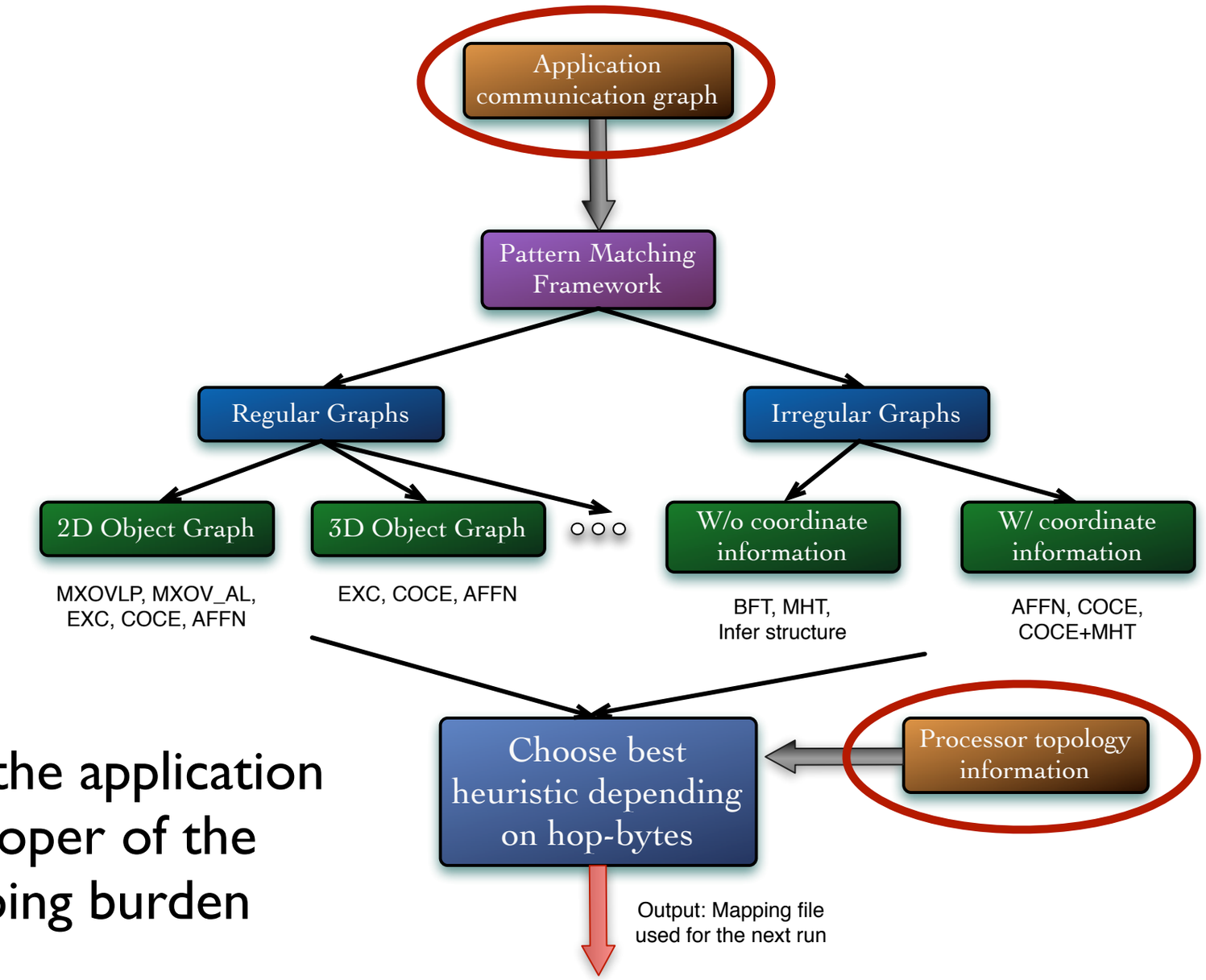
Automatic Mapping Framework



Automatic Mapping Framework

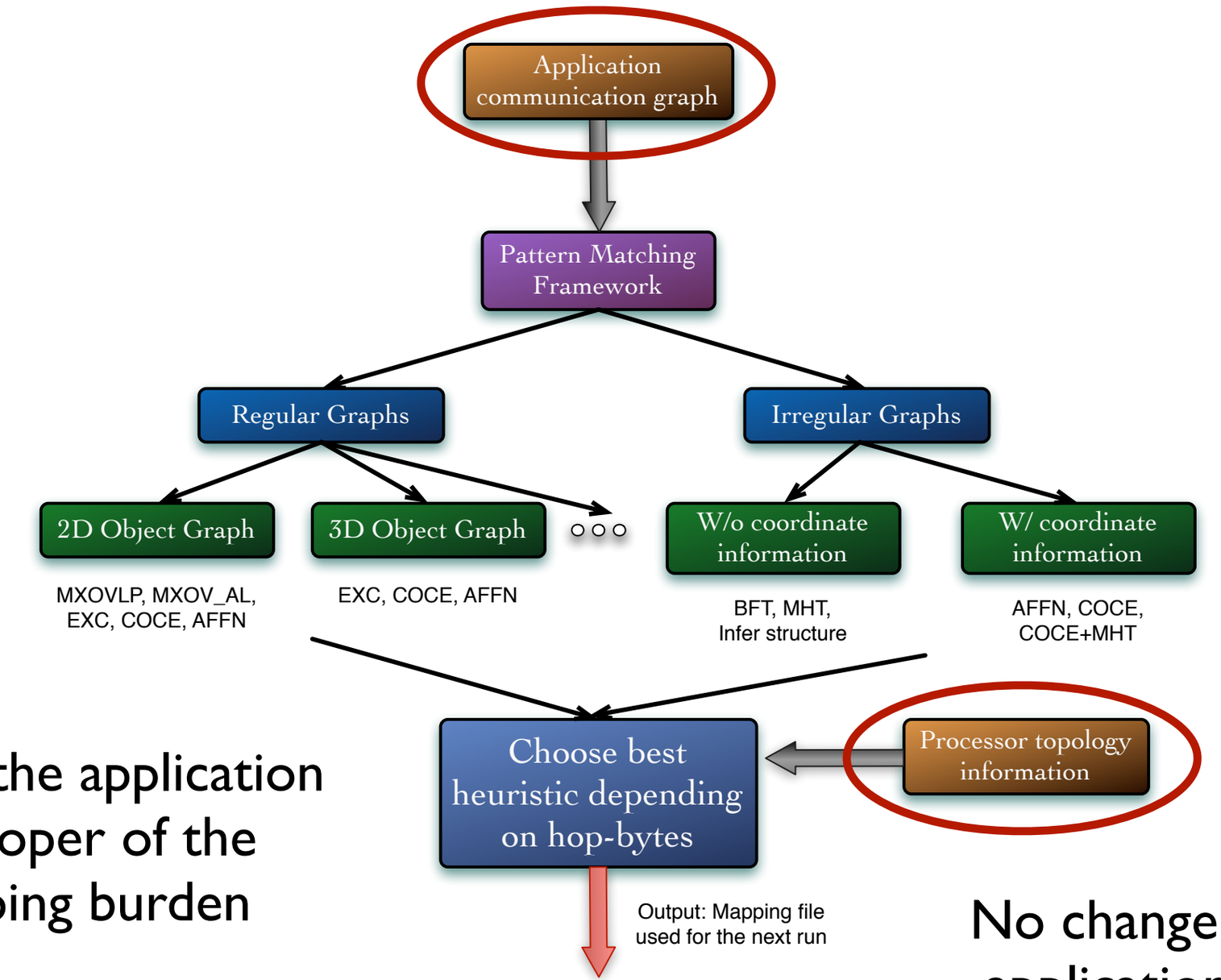


Automatic Mapping Framework



Relieve the application developer of the mapping burden

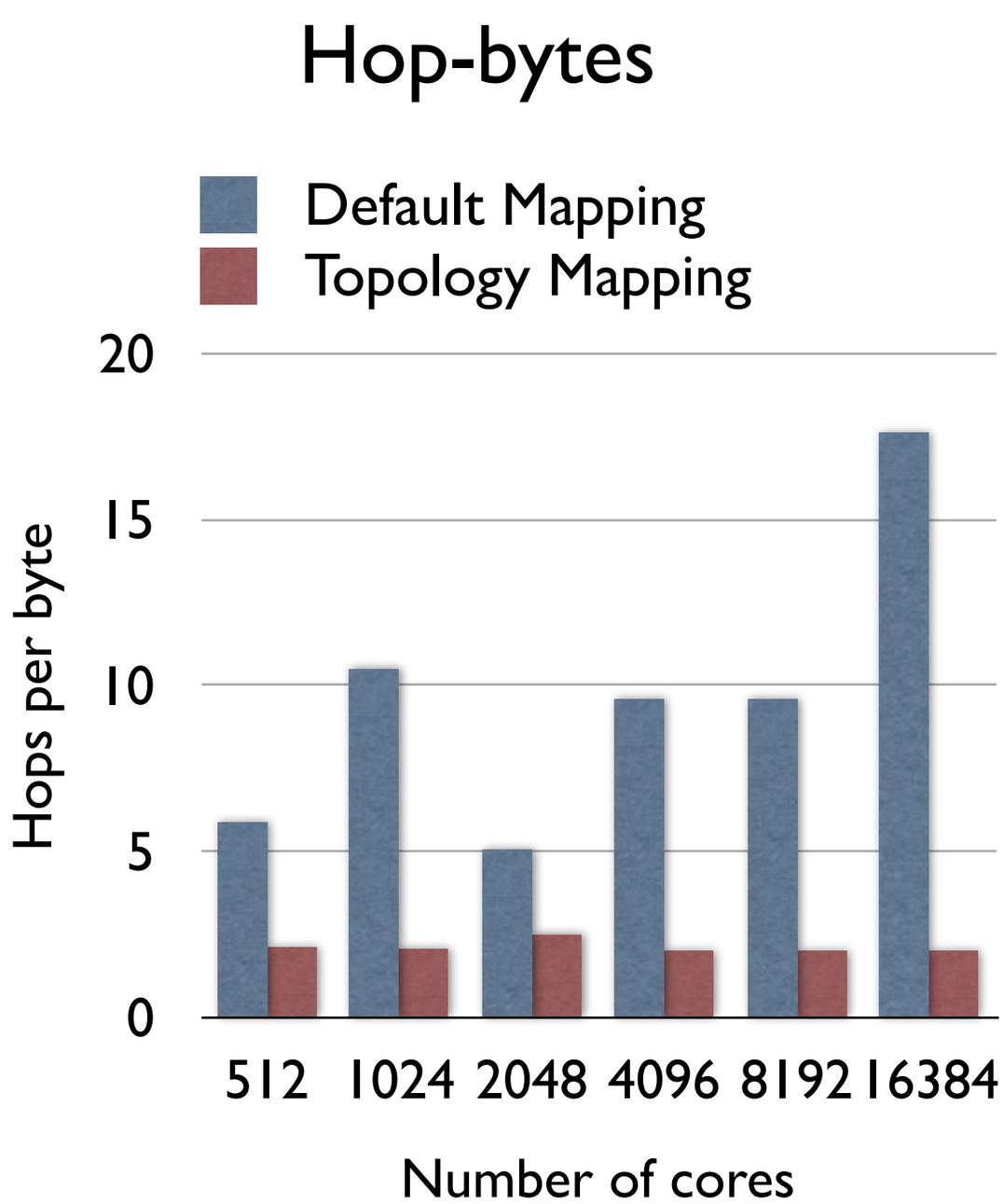
Automatic Mapping Framework



Relieve the application developer of the mapping burden

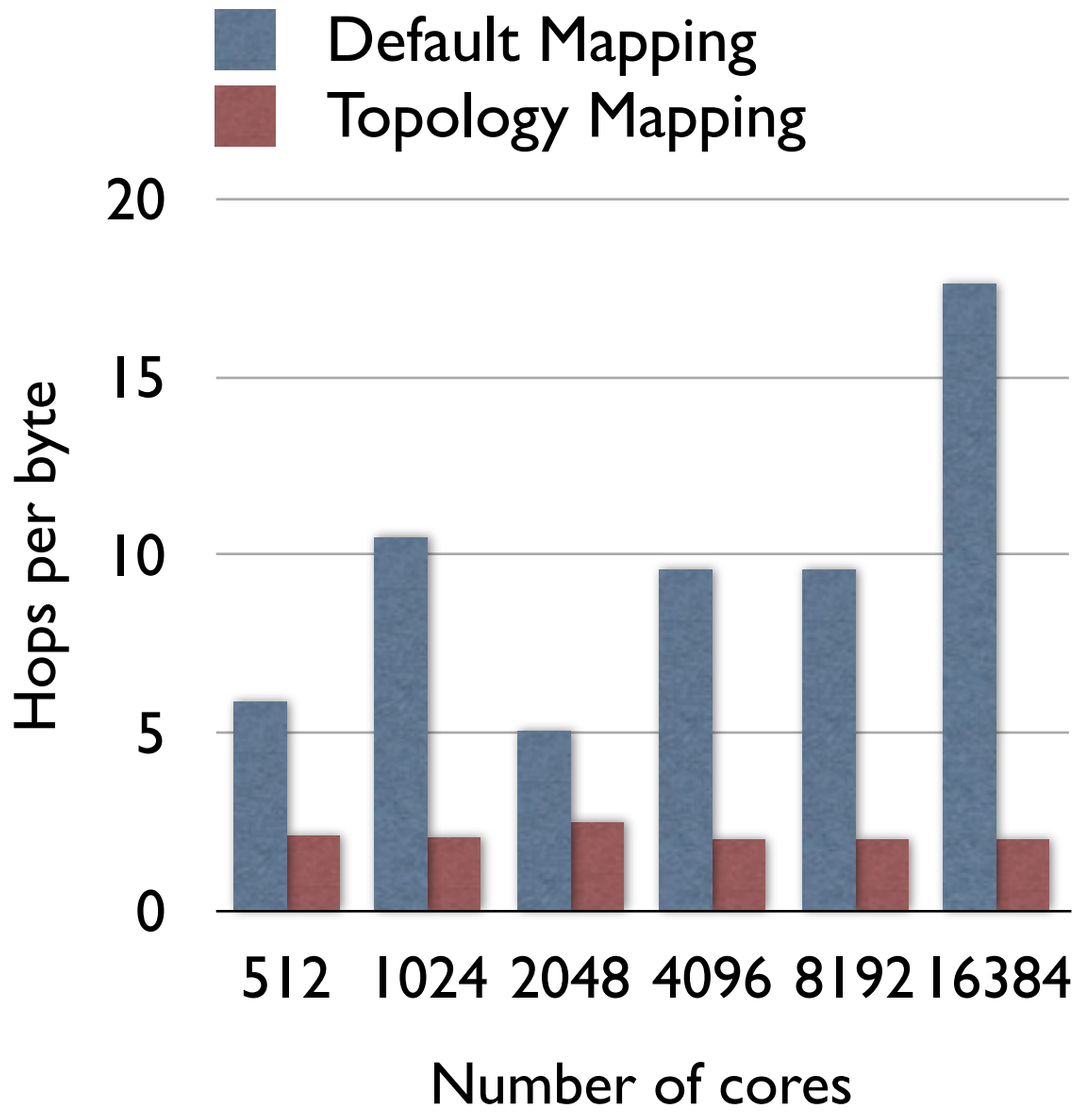
No change to the application code

Results: 2D Stencil on Blue Gene/P

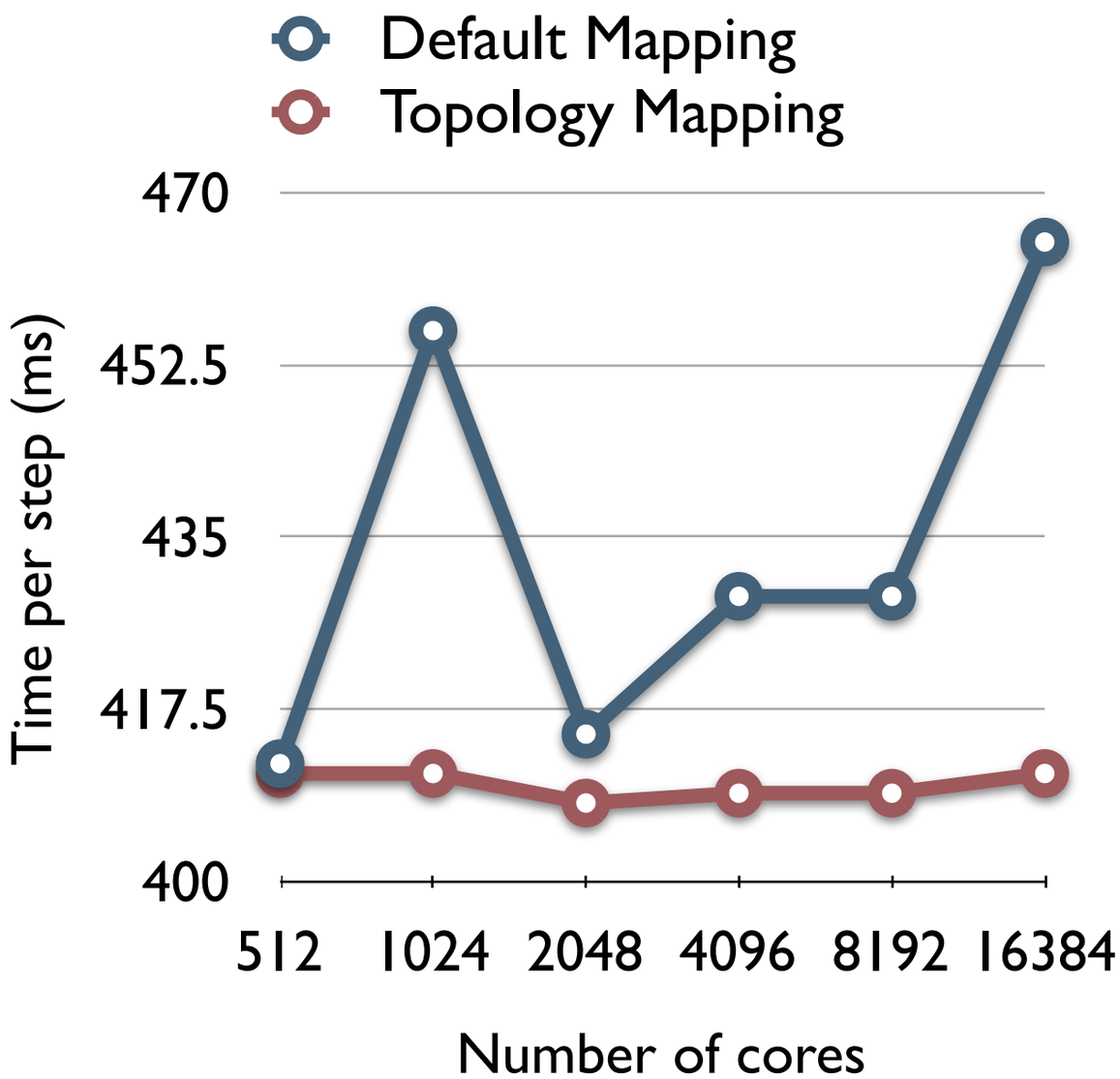


Results: 2D Stencil on Blue Gene/P

Hop-bytes

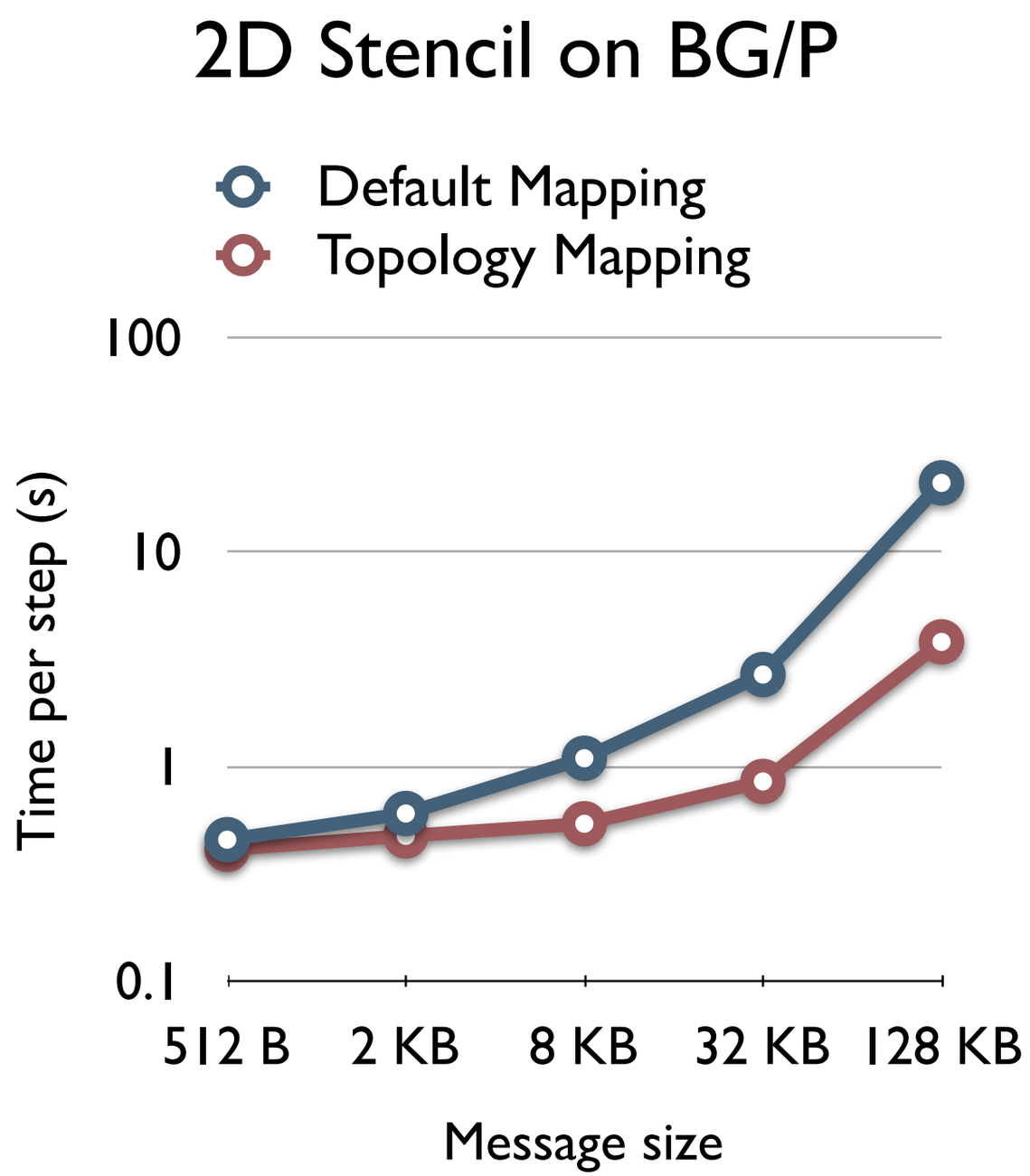


Performance

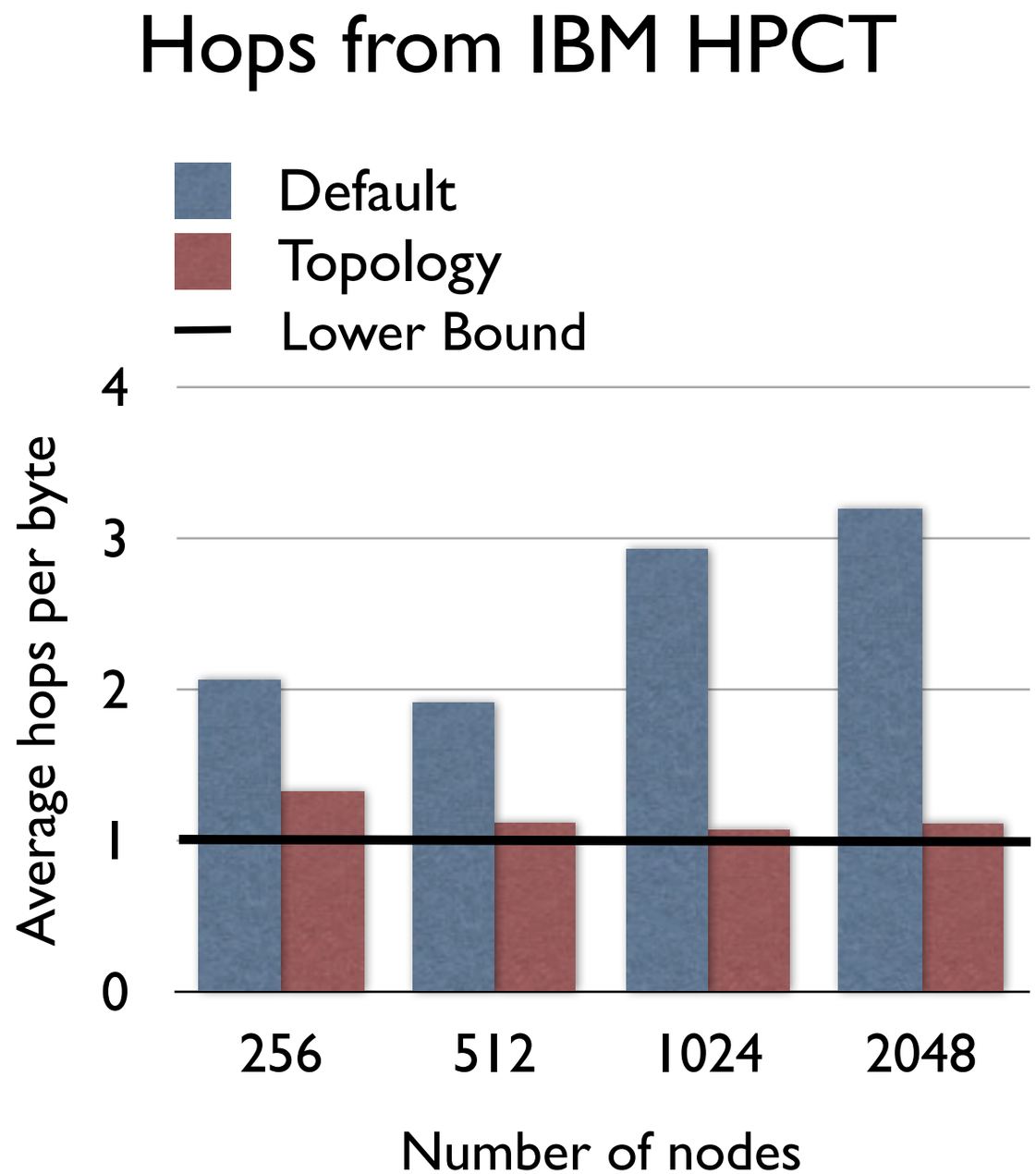


Increasing communication

- With faster processors and constant link bandwidths
  - computation is becoming cheap
  - communication is a bottleneck
- Trend for bytes per flop
  - XT3: 8.77
  - XT4: 1.357
  - XT5: 0.23

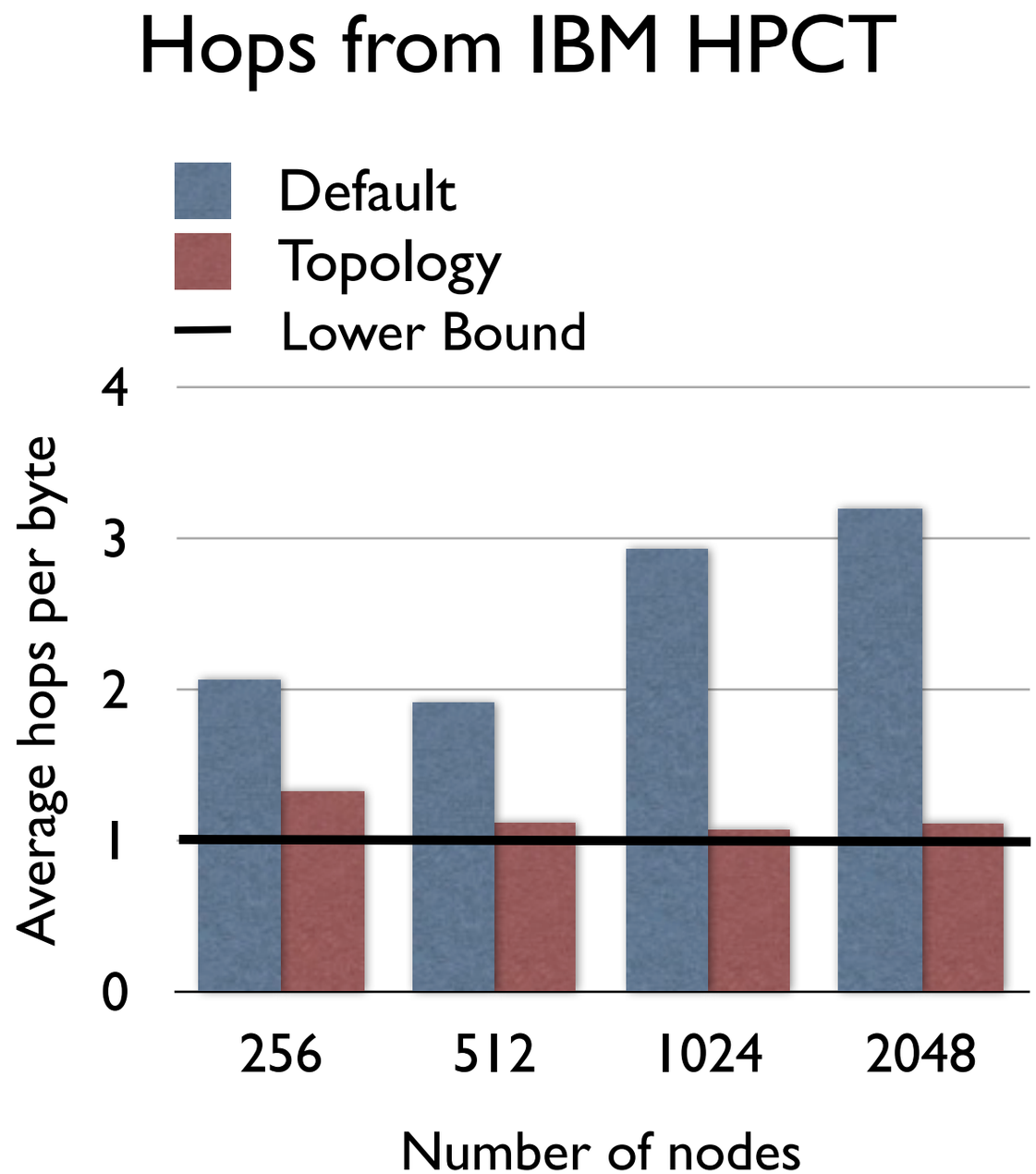


Results: WRF on Blue Gene/P



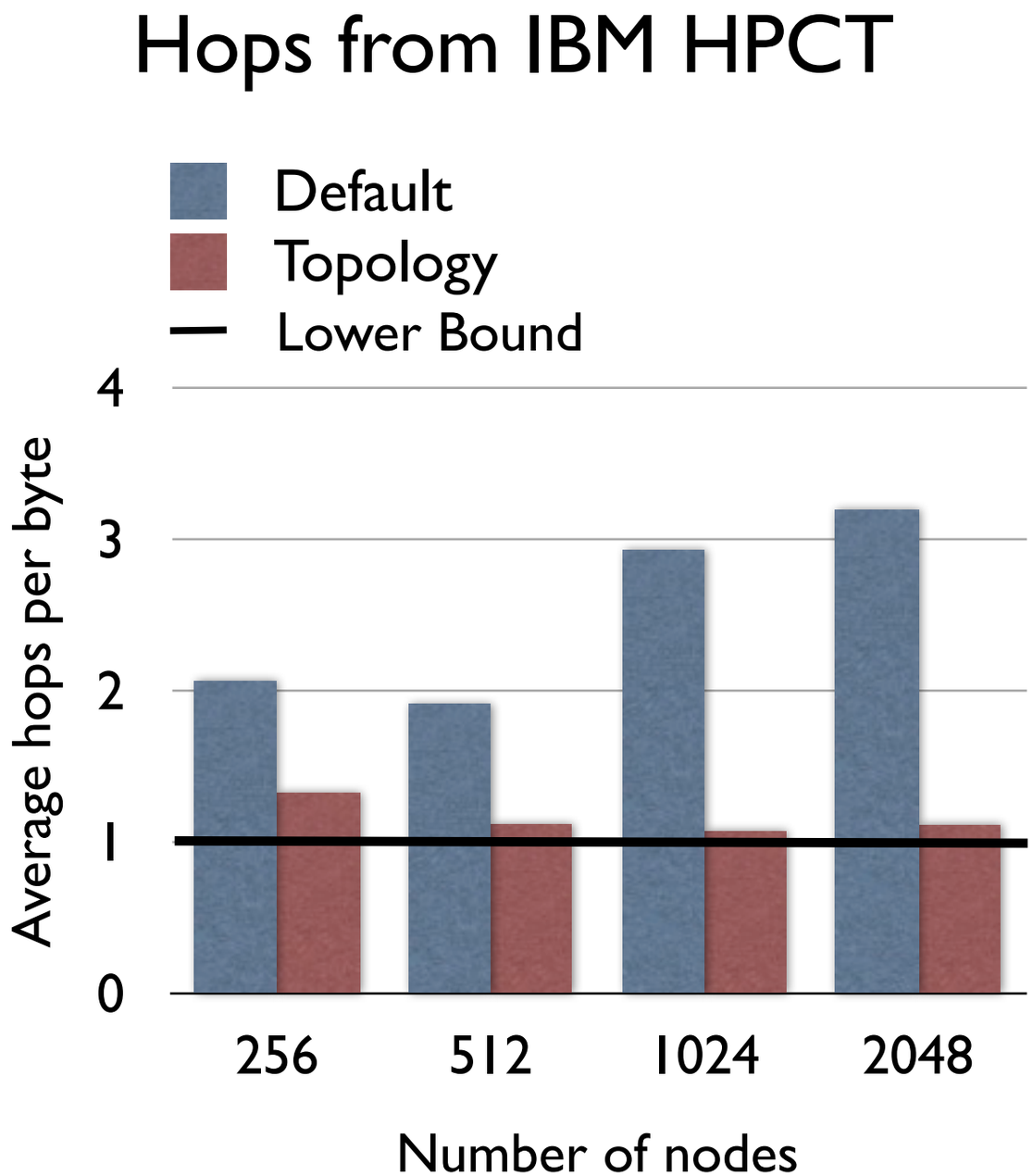
Results: WRF on Blue Gene/P

- Performance improvement negligible on 256 and 512 cores



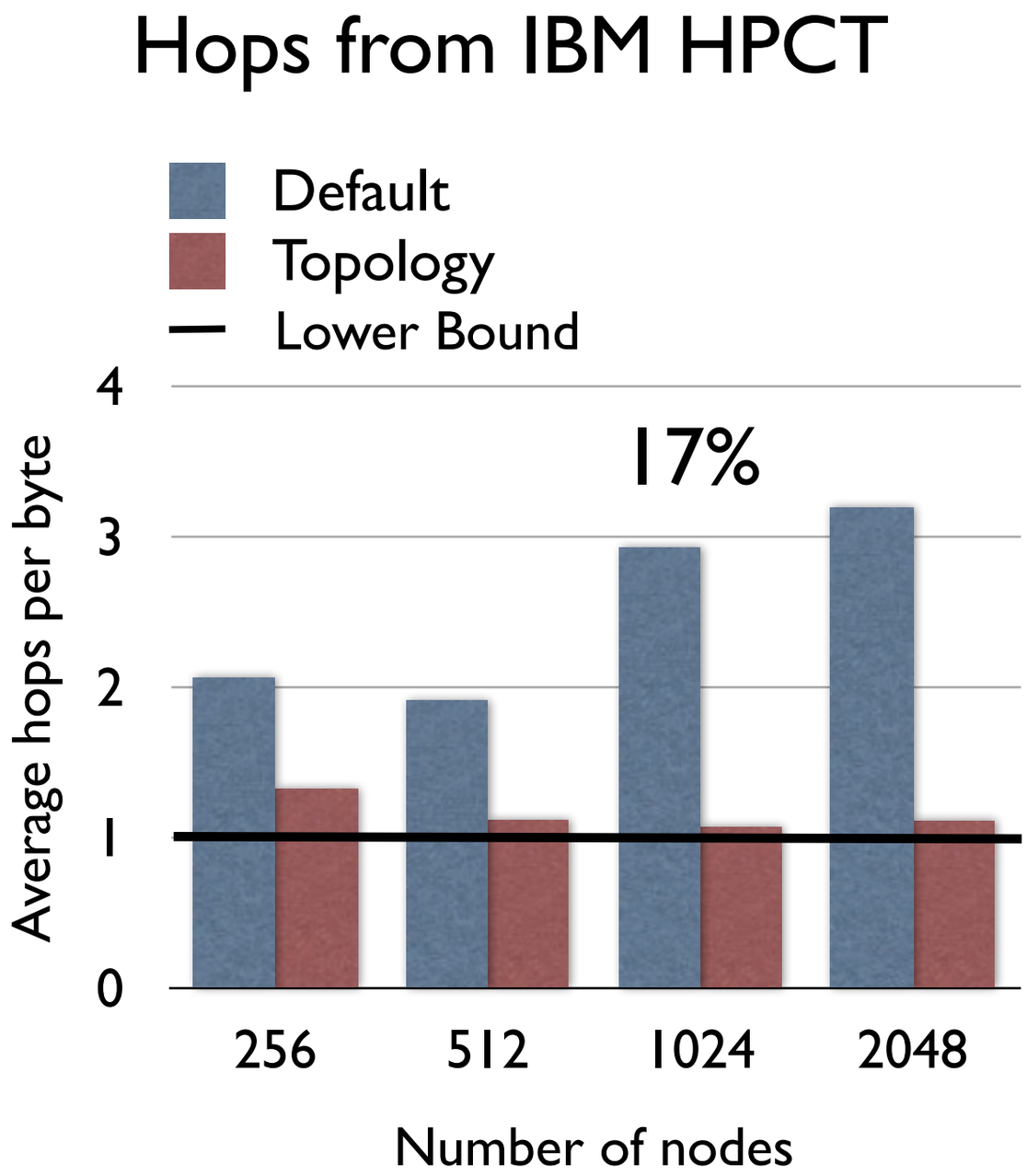
Results: WRF on Blue Gene/P

- Performance improvement negligible on 256 and 512 cores
- On 1024 nodes:
  - Hops reduce by: 64%
  - Time for communication reduces by 11%
  - Performance improves by 17%



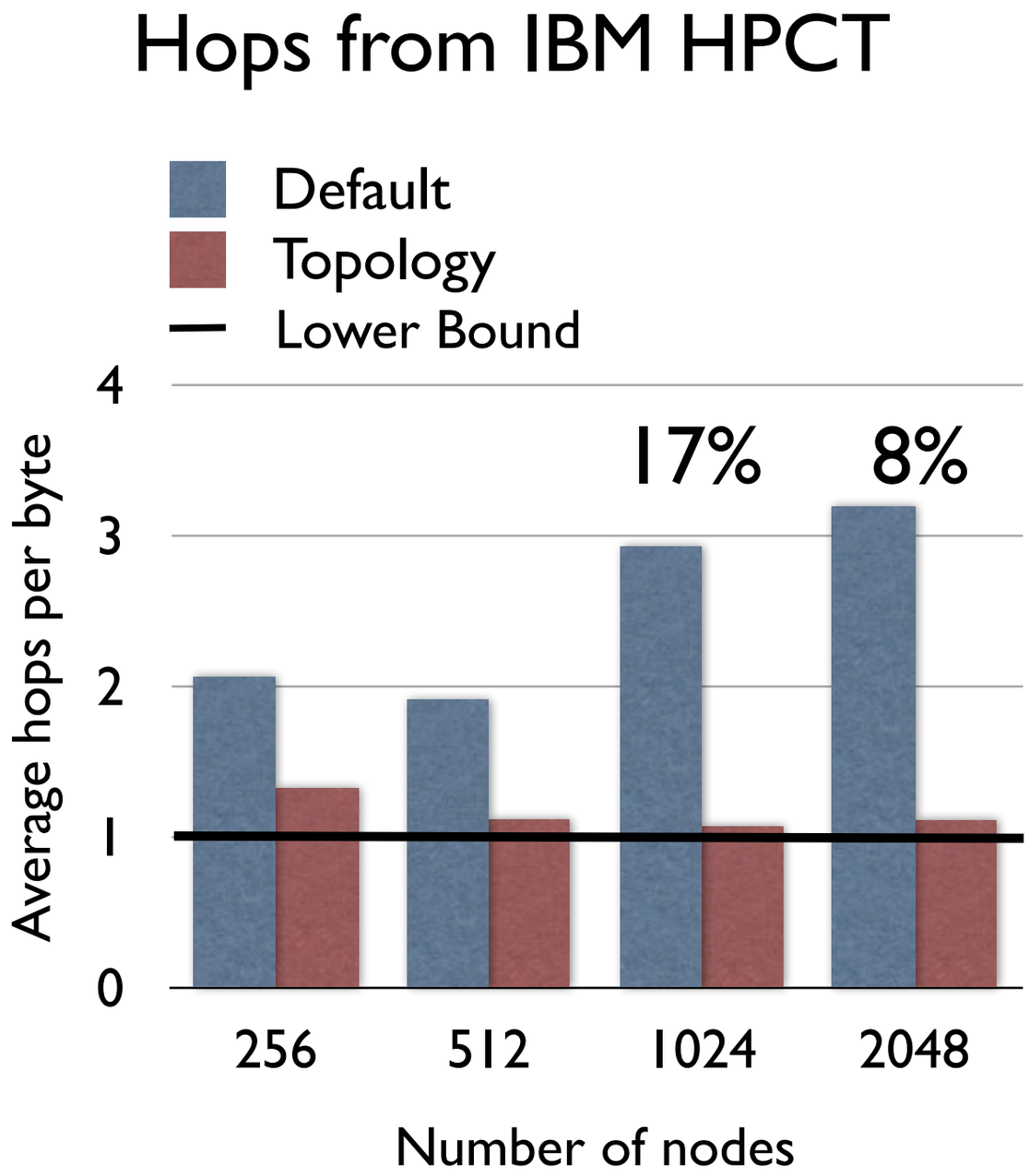
Results: WRF on Blue Gene/P

- Performance improvement negligible on 256 and 512 cores
- On 1024 nodes:
  - Hops reduce by: 64%
  - Time for communication reduces by 11%
  - Performance improves by 17%

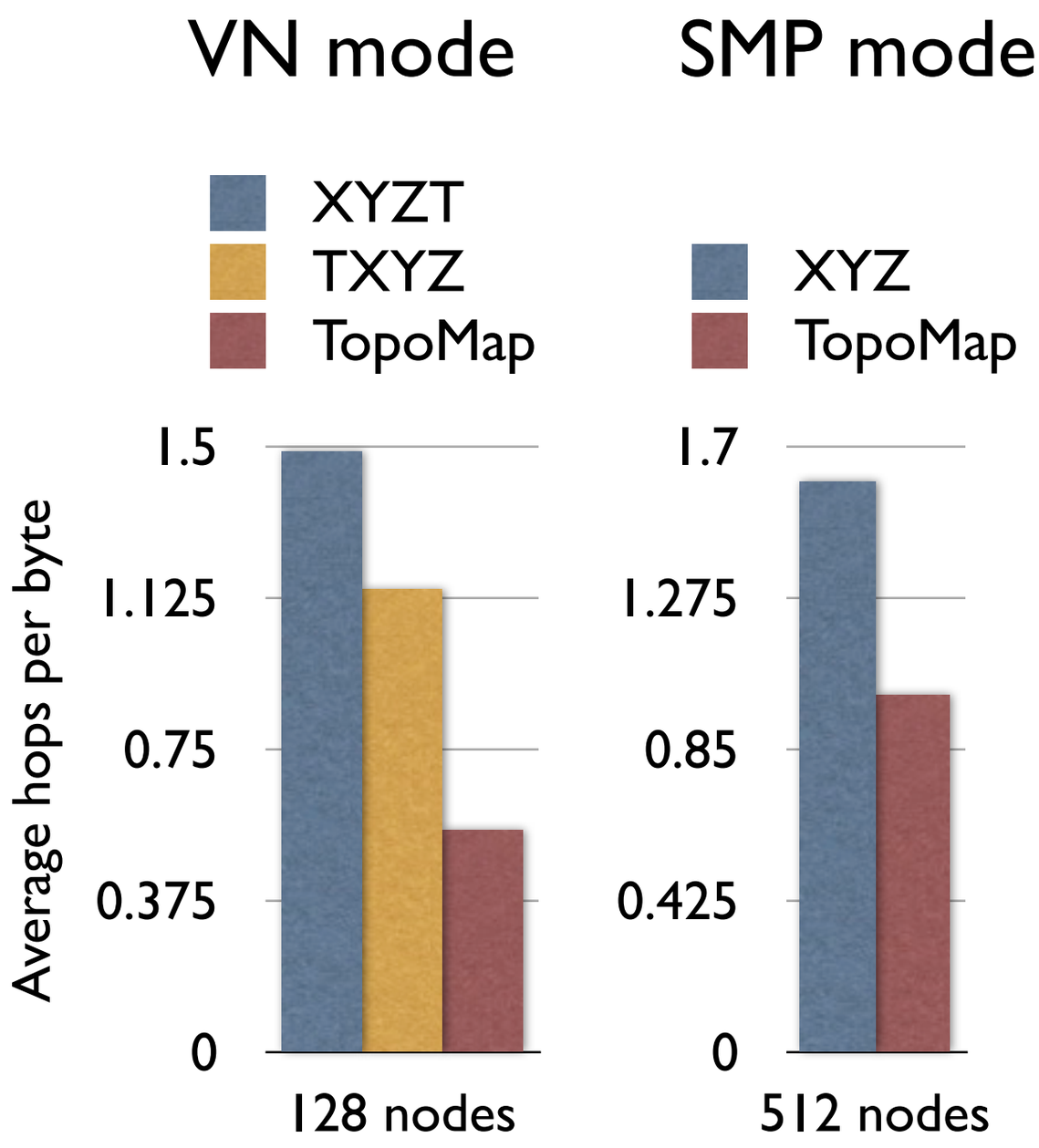


Results: WRF on Blue Gene/P

- Performance improvement negligible on 256 and 512 cores
- On 1024 nodes:
  - Hops reduce by: 64%
  - Time for communication reduces by 11%
  - Performance improves by 17%

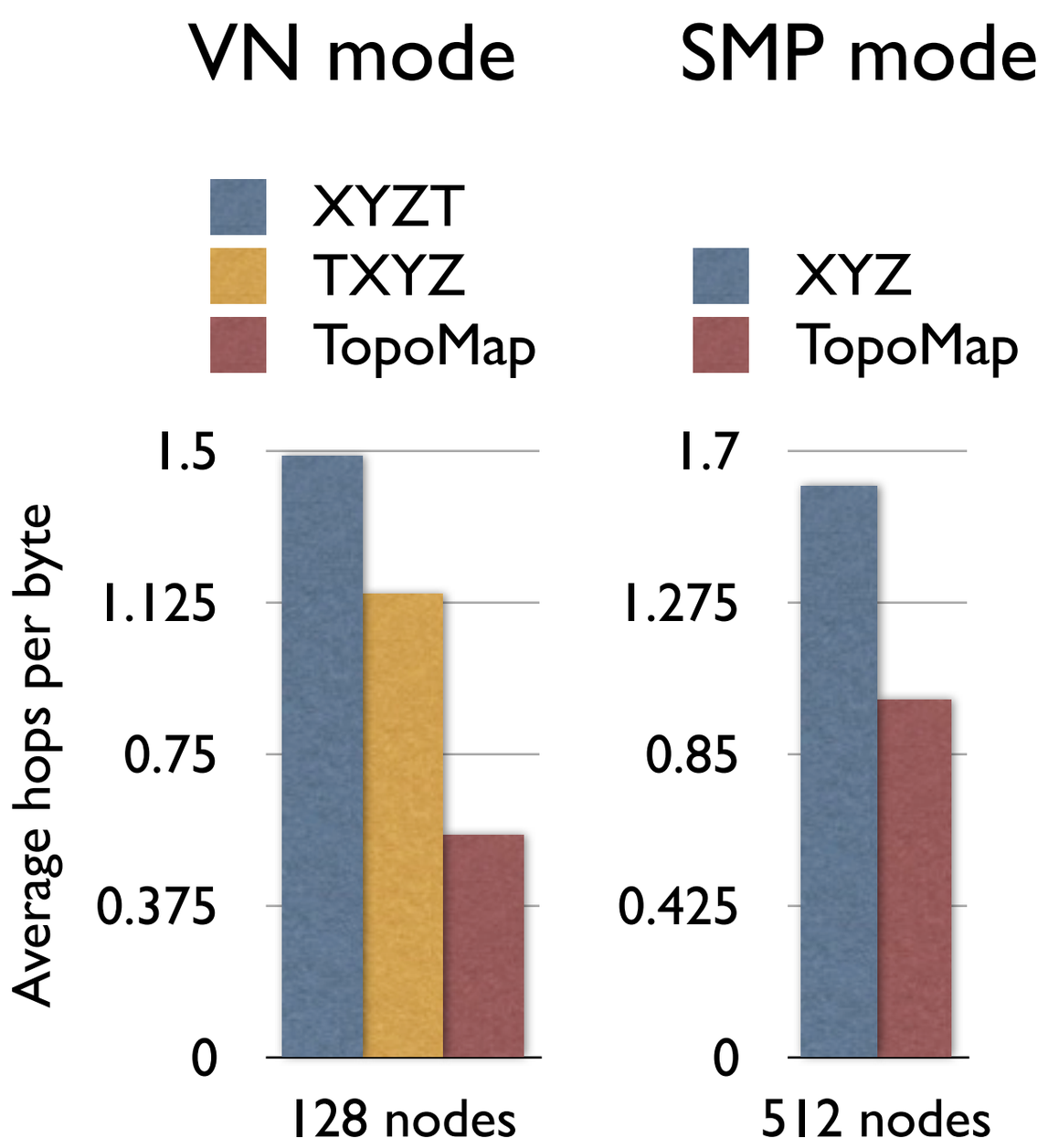


Results: POP on Blue Gene/P



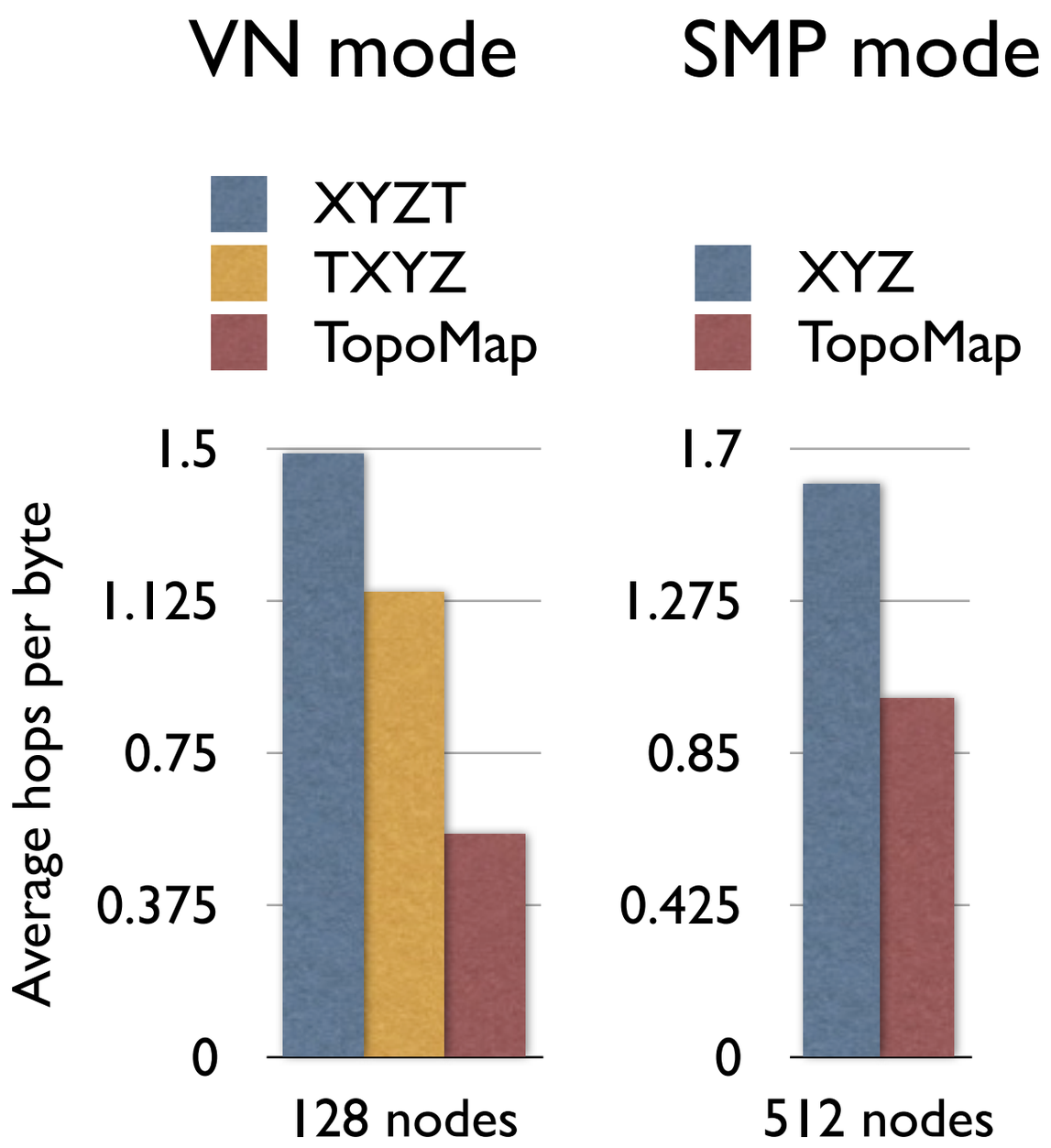
Results: POP on Blue Gene/P

- In VN mode (using all 4 cores per node):
  - Reduction in hops: 60%
  - No improvement in overall performance



Results: POP on Blue Gene/P

- In VN mode (using all 4 cores per node):
  - Reduction in hops: 60%
  - No improvement in overall performance
- In spite of POP spending 55% time in communication
  - MPI_Waitall and MPI_Allreduce



Summary

- Contention in modern day supercomputers can impact performance: makes mapping important
  - Even for high bandwidth interconnects such as Cray
- Certain classes of applications (latency sensitive, communication bound) benefit most
  - OpenAtom shows performance improvements of up to 50%
  - NAMD - improvements for > 4k cores
- Developing an automatic mapping framework
  - Relieve the application developer of the mapping burden

PPL UIUC

Questions?

Acknowledgements:

IBM Watson Research Center (Blue Gene/L): Fred Mintzer, Glenn Martyna
Pittsburgh Supercomputing Center (Cray XT3): Chad Vizino, Shawn Brown
Argonne National Laboratory (Blue Gene/P): Pete Beckman, Charles Bacon
Oak Ridge National Laboratory (Cray XT4/5): Donald Frederick, Patrick Worley

Funded in part by the Center for Simulation of Advanced Rockets (Univ. of Illinois) through DOE Grant B341494

E-mail: bhatele@illinois.edu