A CASE STUDY OF COMMUNICATION OPTIMIZATIONS ON 3D MESH INTERCONNECTS

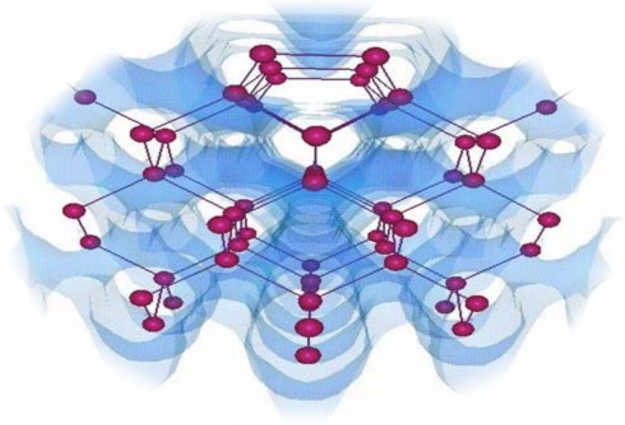
Abhinav Bhatele, Eric Bohm, Laxmikant V. Kale
Parallel Programming Laboratory

Outline

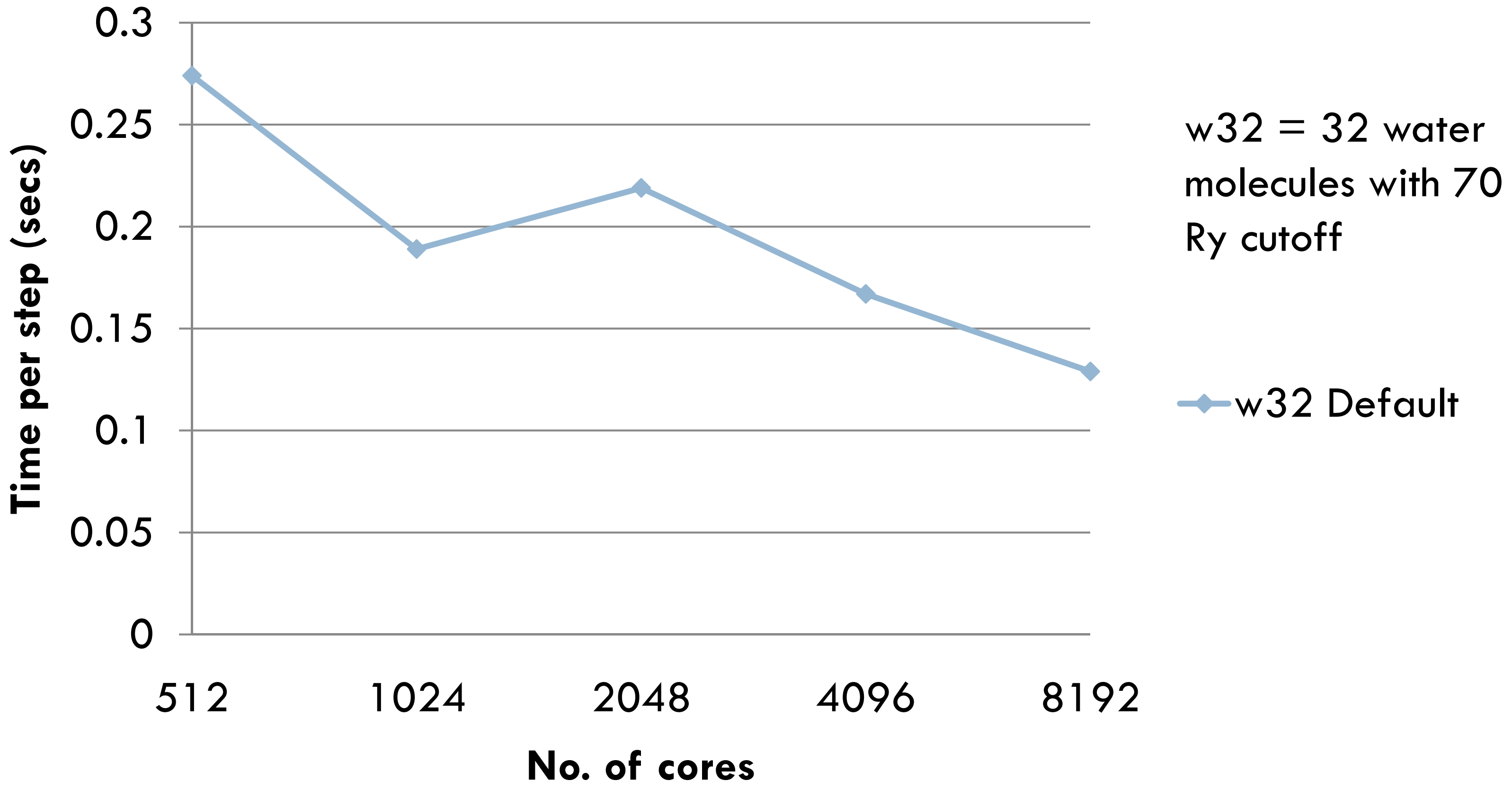
- □ Motivation
- □ Solution: Mapping of OpenAtom
- □ Performance Benefits
- □ Bigger Picture:
  - Resources Needed
  - Heuristic Solutions
- □ Automatic Mapping

OpenAtom

- □ Ab-Initio Molecular Dynamics code
- □ Consider electrostatic interactions between the nuclei and electrons
- □ Calculate different energy terms
- □ Divided into different phases with lot of communication



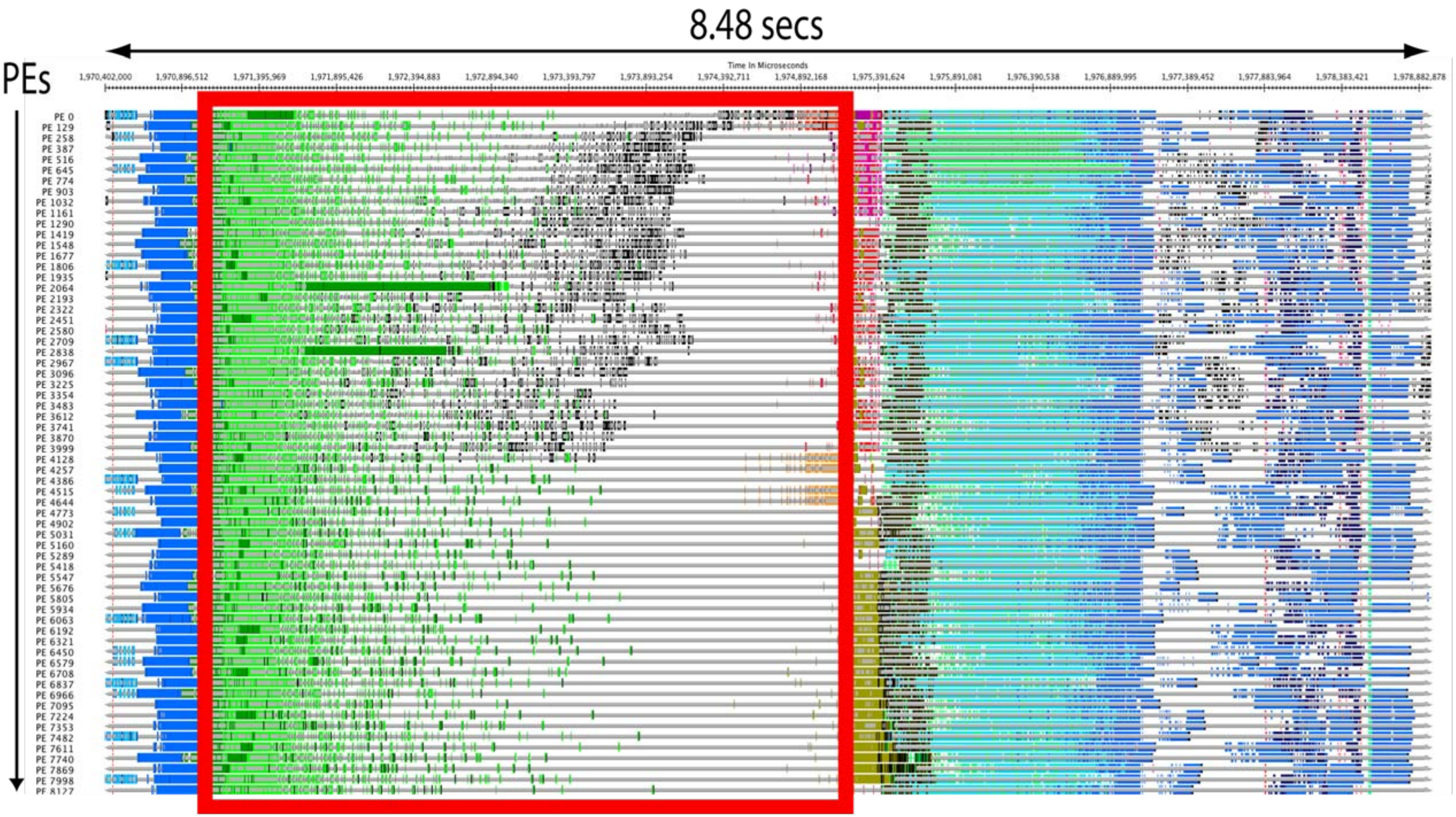
OpenAtom on Blue Gene/L



w32 = 32 water molecules with 70 Ry cutoff

Runs on Blue Gene/L at IBM T J Watson Research Center, CO mode

The problem lies in …



Performance Analysis and Visualization Tool: Projections (part of Charm++) – Timeline View

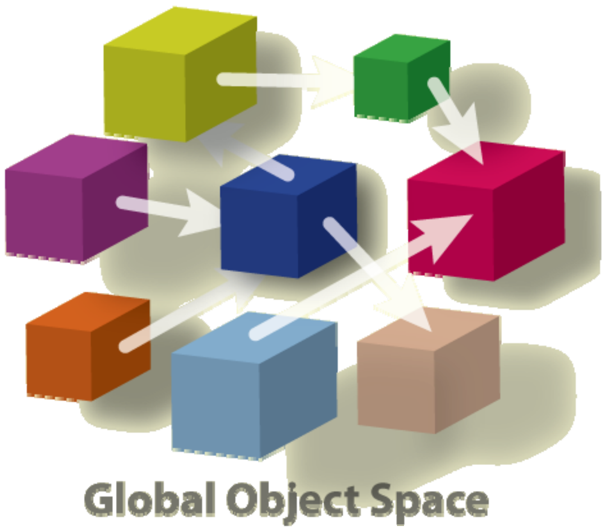
Solution –

Topology Aware Mapping

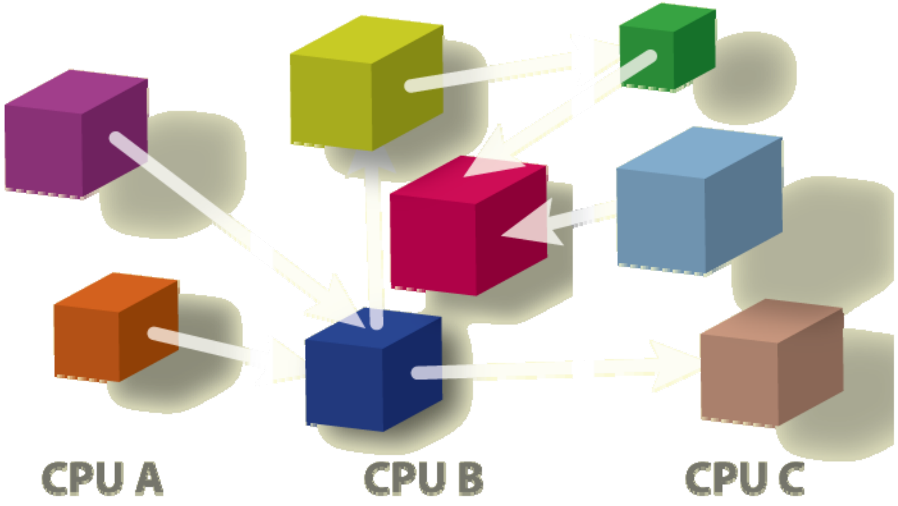
Processor Virtualization

Programmer: Decomposes the computation into objects

Runtime: Maps the computation on to the processors



User View



System View

Benefits of Charm++

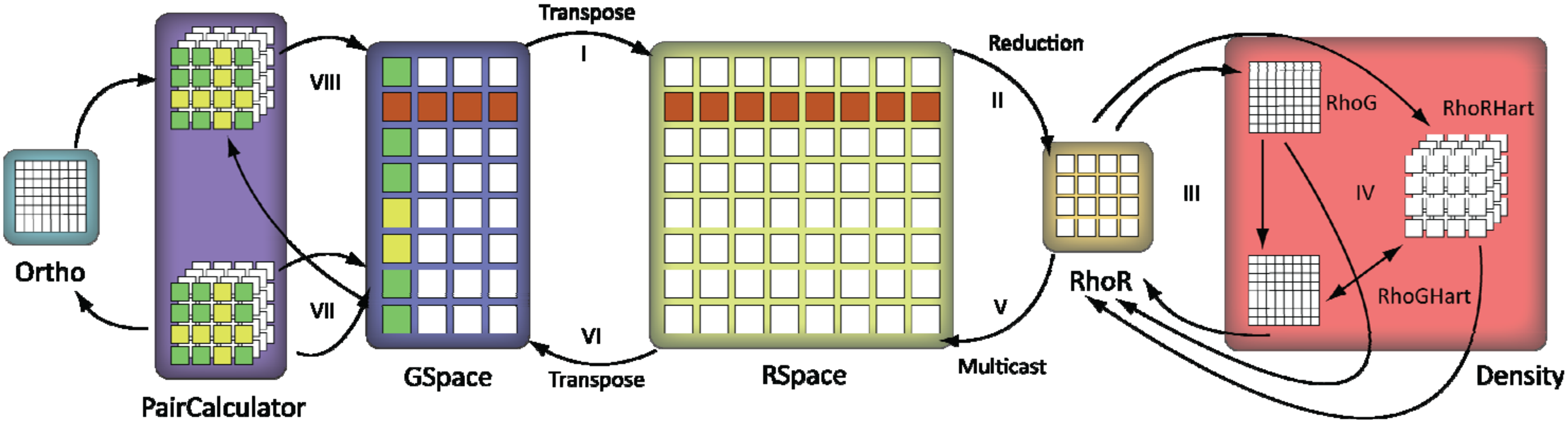
- Computation is divided into objects/chares/virtual processors (VPs)
- Separates decomposition from mapping
- VPs can be flexibly mapped to actual physical processors (PEs)

Topology Manager API†

- The application needs information such as
  - Dimensions of the partition
  - Rank to physical co-ordinates and vice-versa
- TopoManager: a uniform API
  - On BG/L and BG/P: provides a wrapper for system calls
  - On XT3/4/5, there are no such system calls
  - Provides a clean and uniform interface to the application

† http://charm.cs.uiuc.edu/~bhatele/phd/topomgr.htm

Parallelization using Charm++

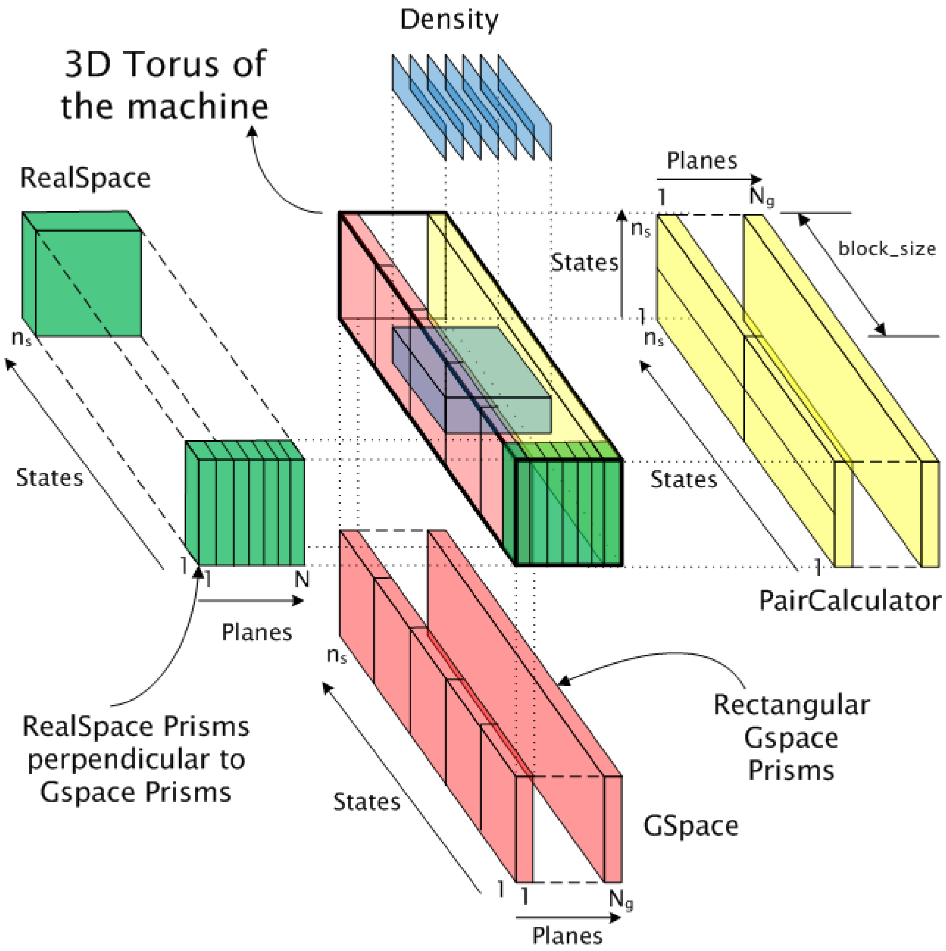


Eric Bohm, Glenn J. Martyna, Abhinav Bhatele, Sameer Kumar, Laxmikant V. Kale, John A. Gunnels, and Mark E. Tuckerman. **Fine Grained Parallelization of the Car-Parrinello ab initio MD Method on Blue Gene/L**. *IBM J. of R. and D.: Applications of Massively Parallel Systems, 52(1/2):159-174*, 2008.

Mapping Challenge

- Load Balancing: Multiple VPs per PE
- Multiple groups of communicating objects
  - Intra-group communication
  - Inter-group communication
- Conflicting communication requirements

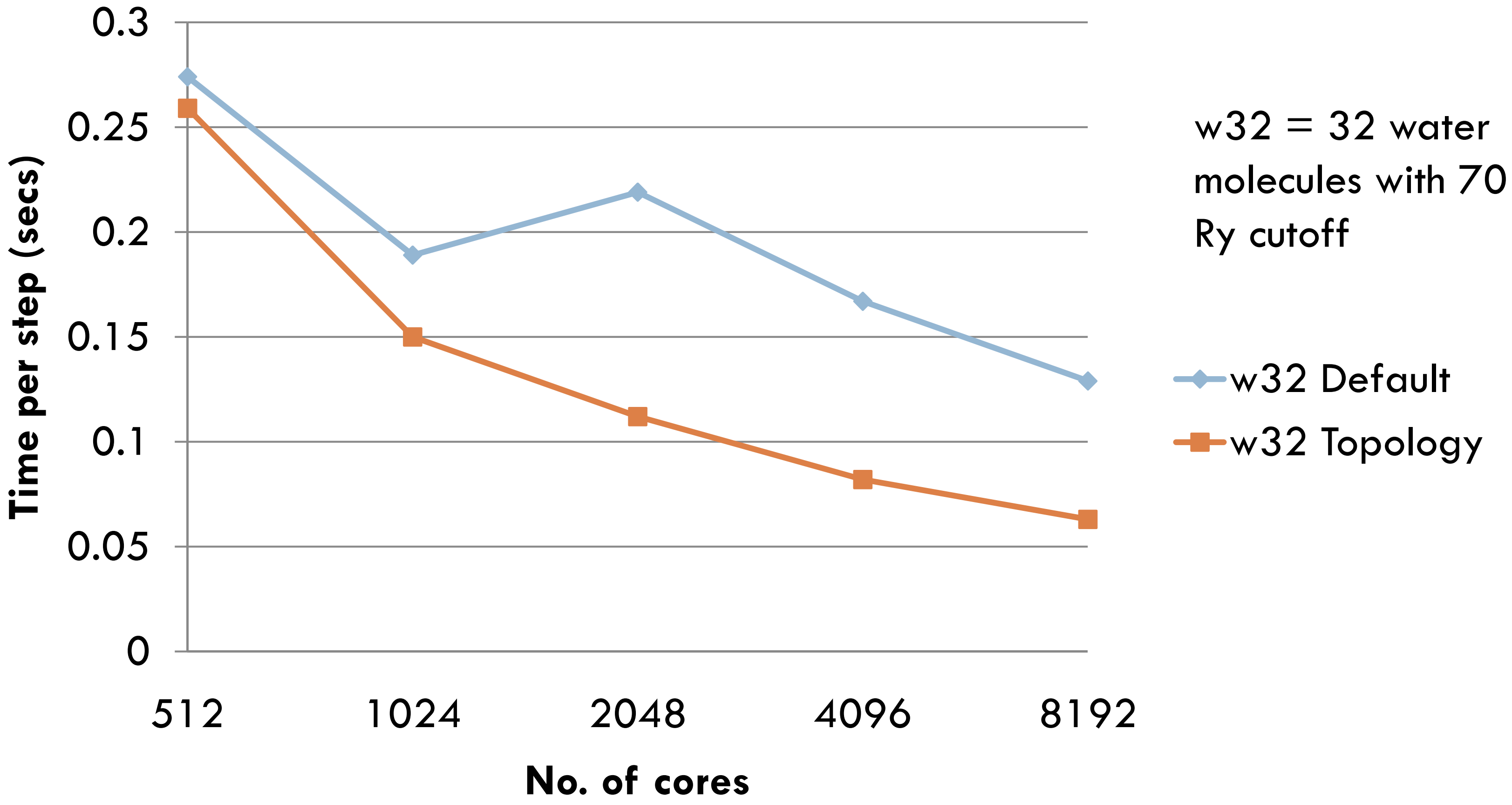
Topology Mapping of Chare Arrays



RealSpace and GSpace have state-wise communication

Paircalculator and GSpace have plane-wise communication

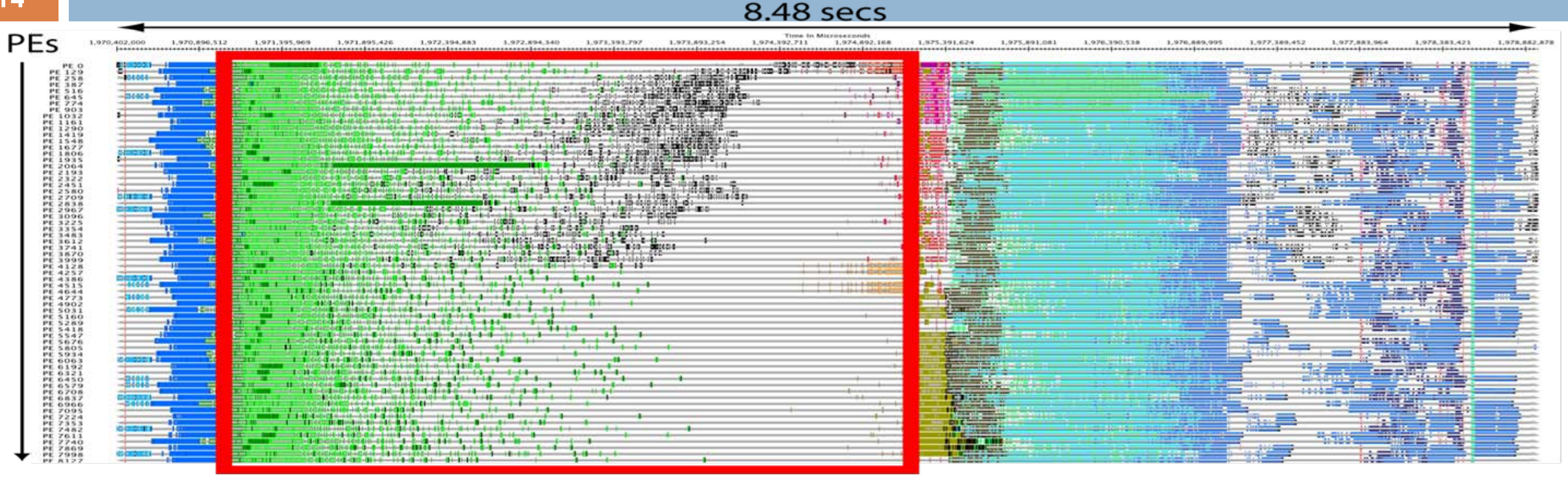
Performance Improvements on BG/L

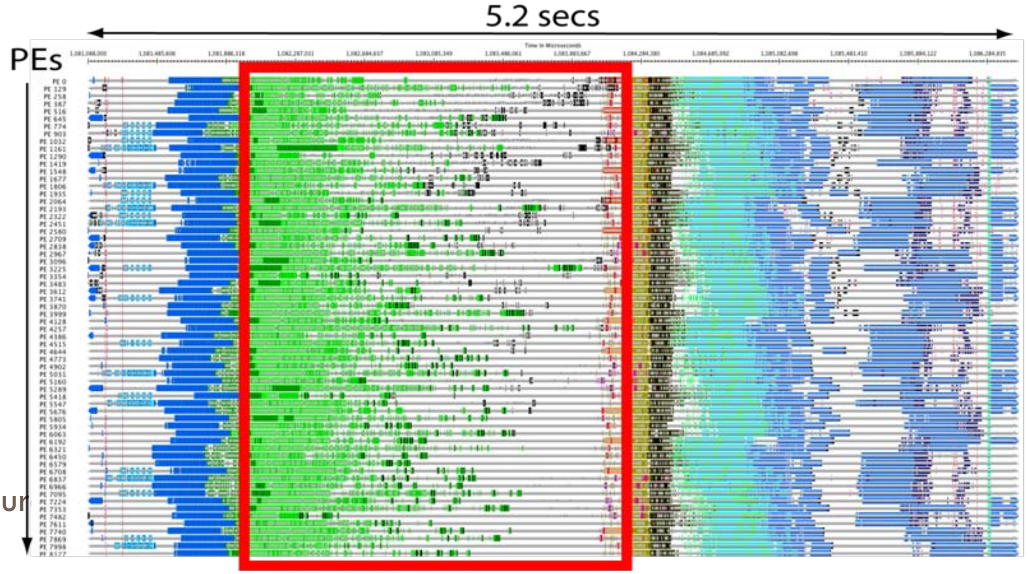


w32 = 32 water molecules with 70 Ry cutoff

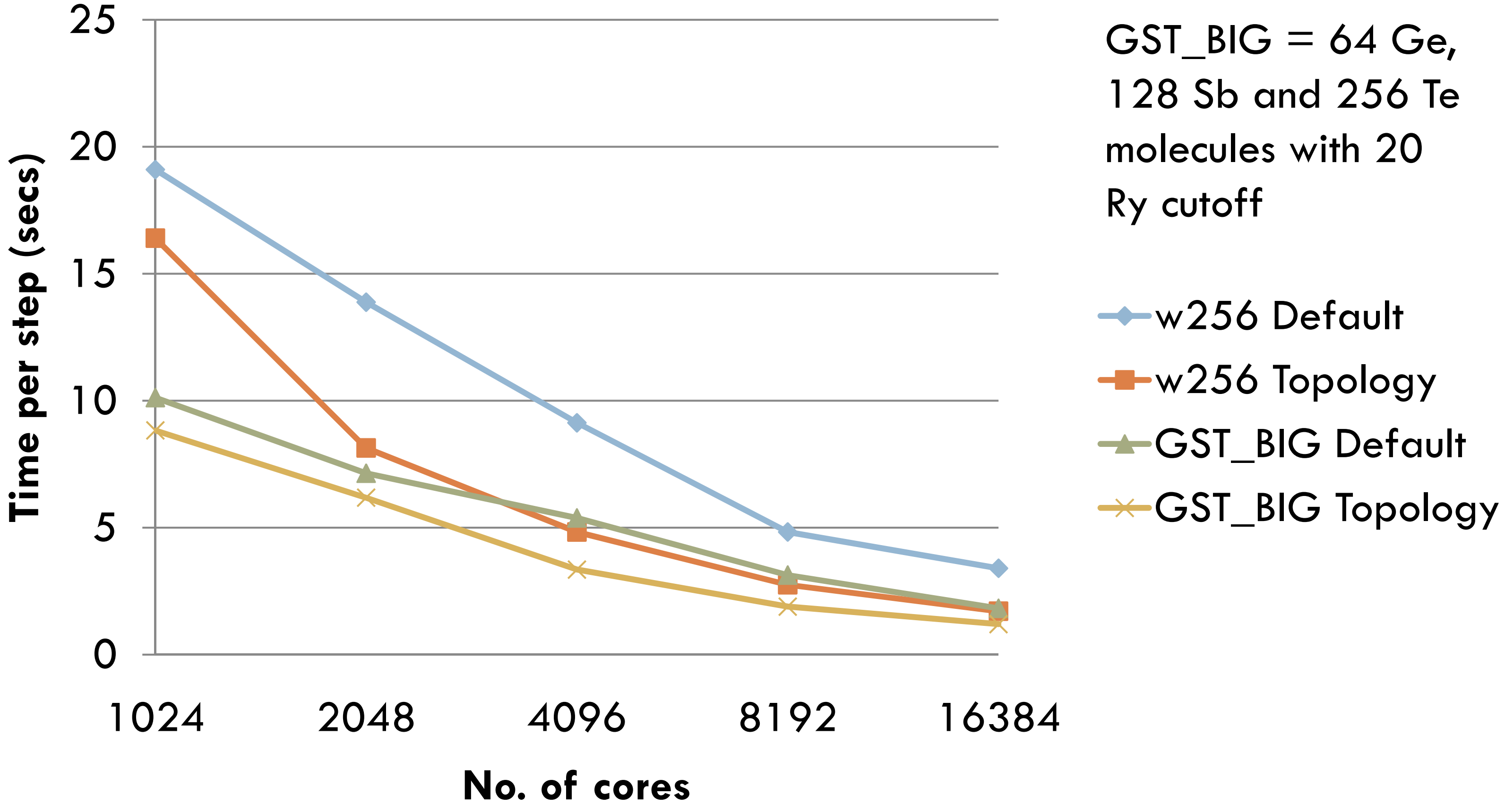
Runs on Blue Gene/L at IBM T J Watson Research Center, CO mode, Year: 2006

Improved Timeline Views





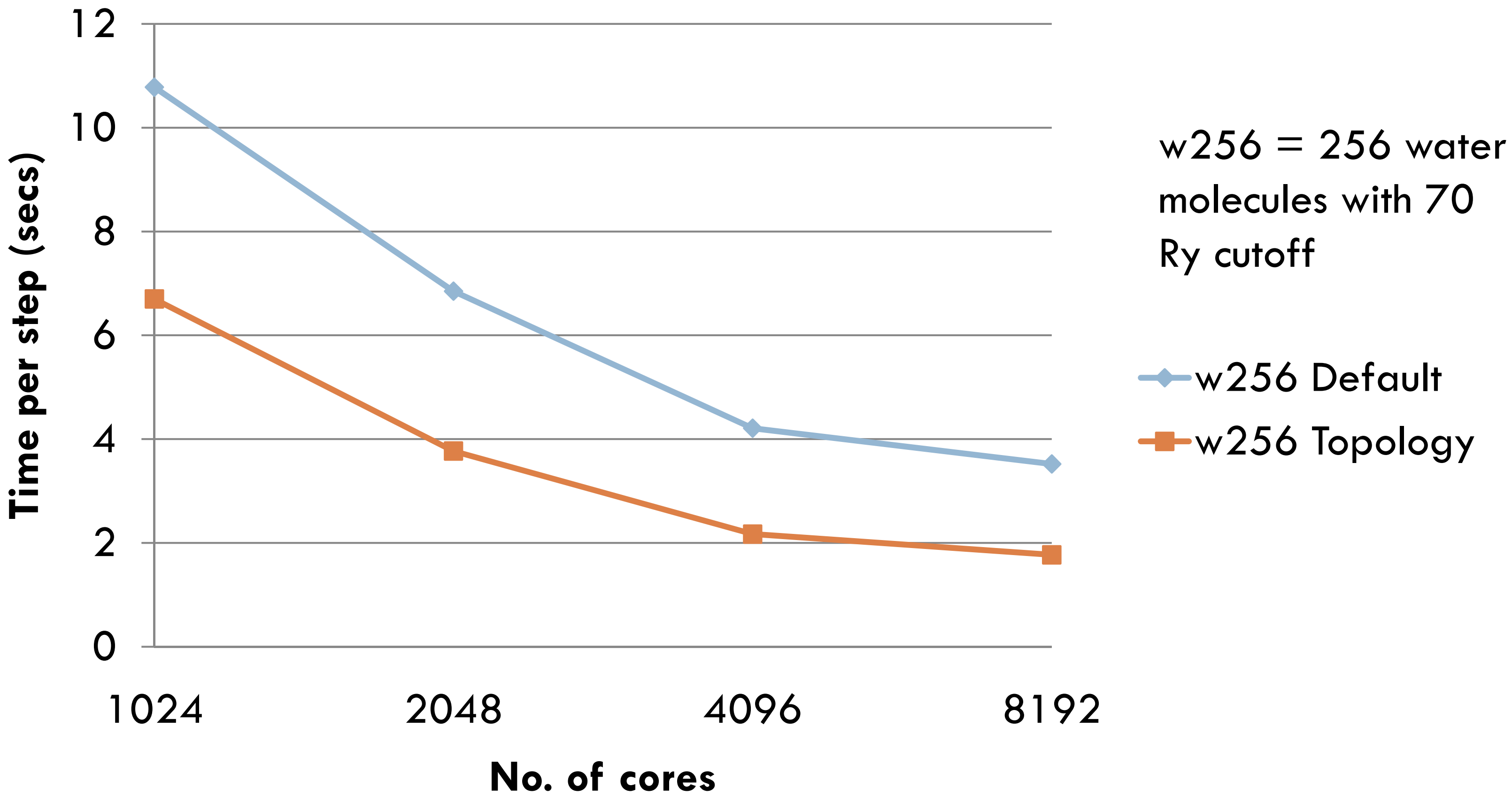
Results on Blue Gene/L



GST_BIG = 64 Ge, 128 Sb and 256 Te molecules with 20 Ry cutoff

Runs on Blue Gene/L at IBM T J Watson Research Center, CO mode

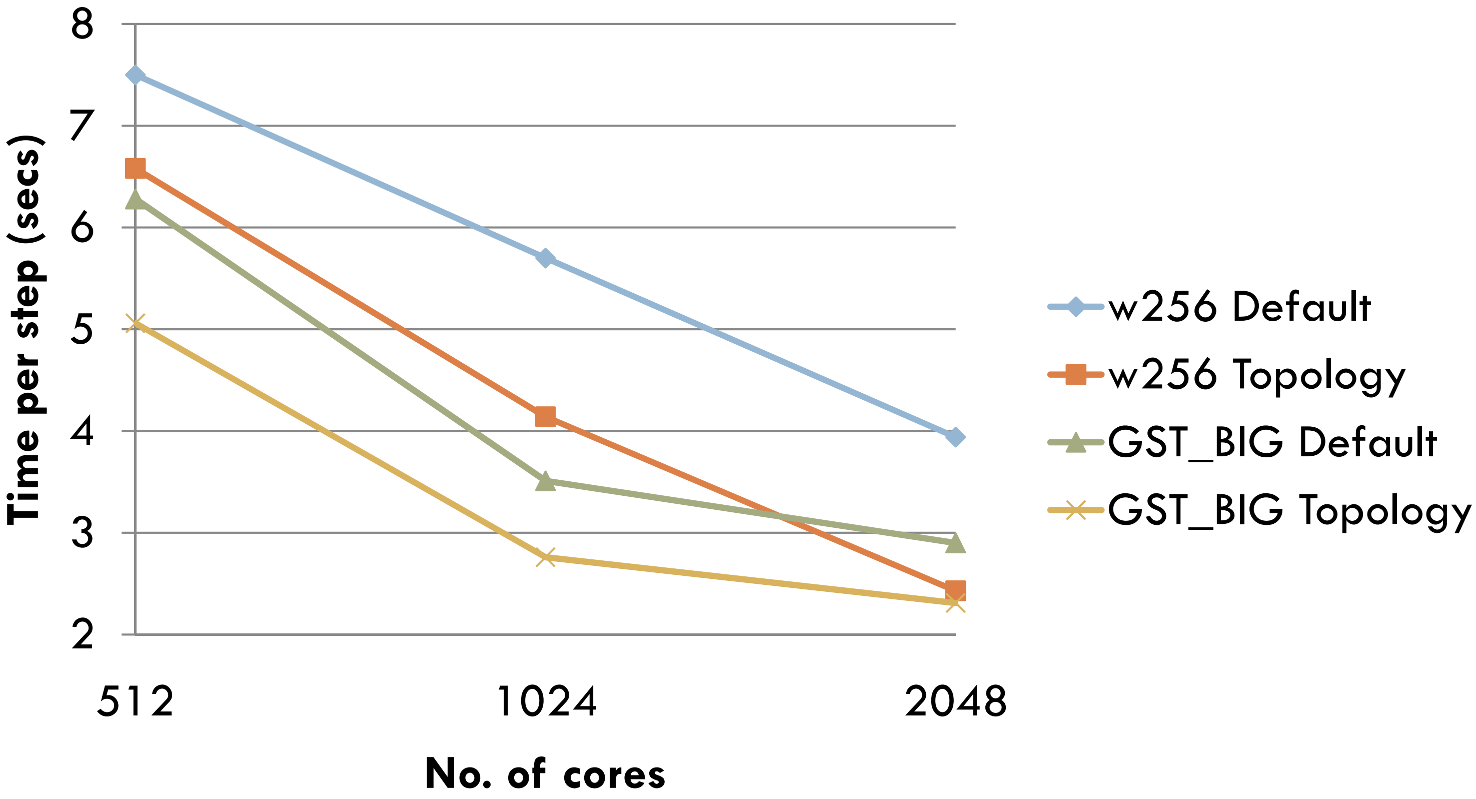
Results on Blue Gene/P



w256 = 256 water molecules with 70 Ry cutoff

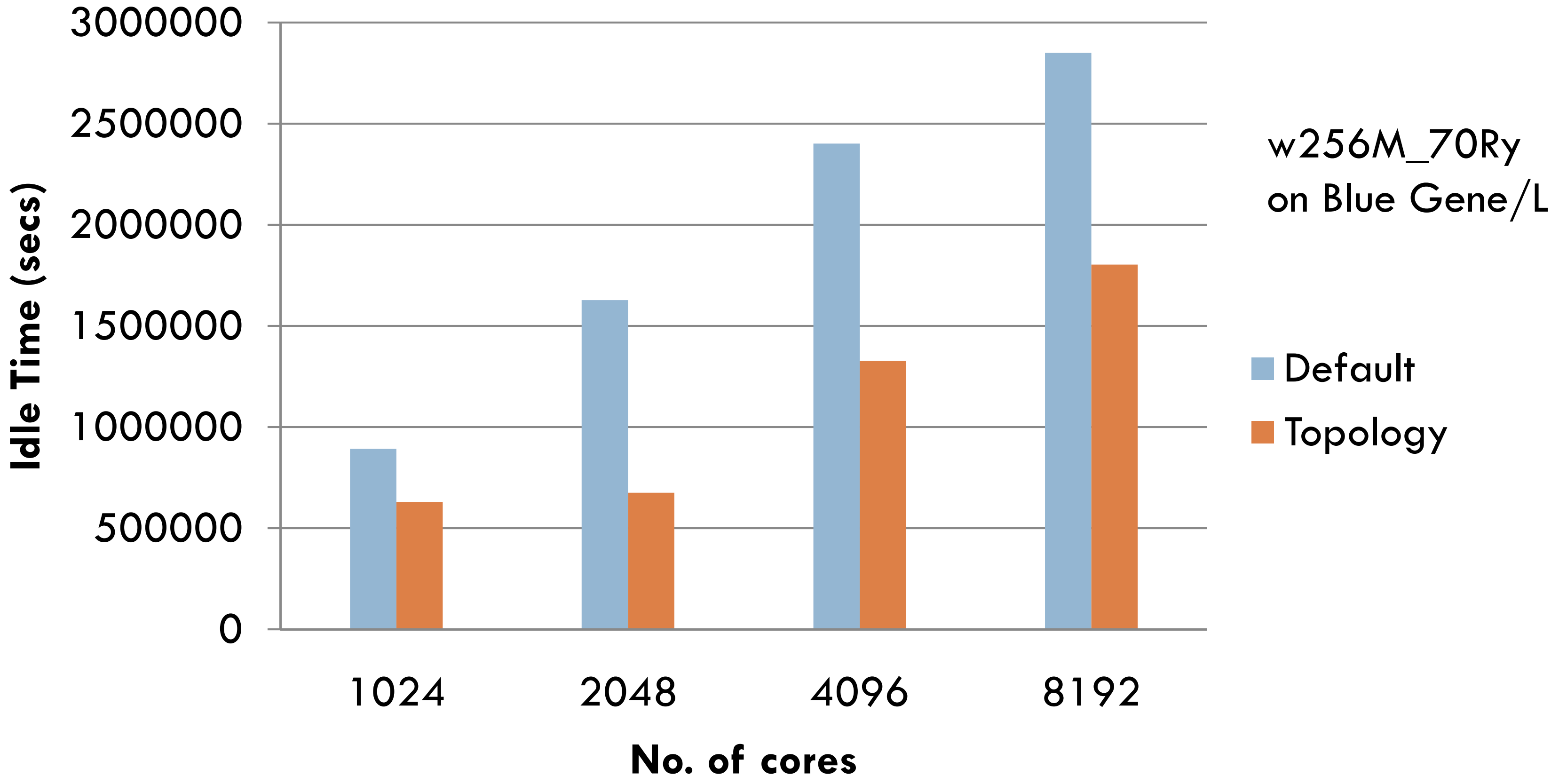
Runs on Blue Gene/P at Argonne National Laboratory, VN mode

Results on Cray XT3



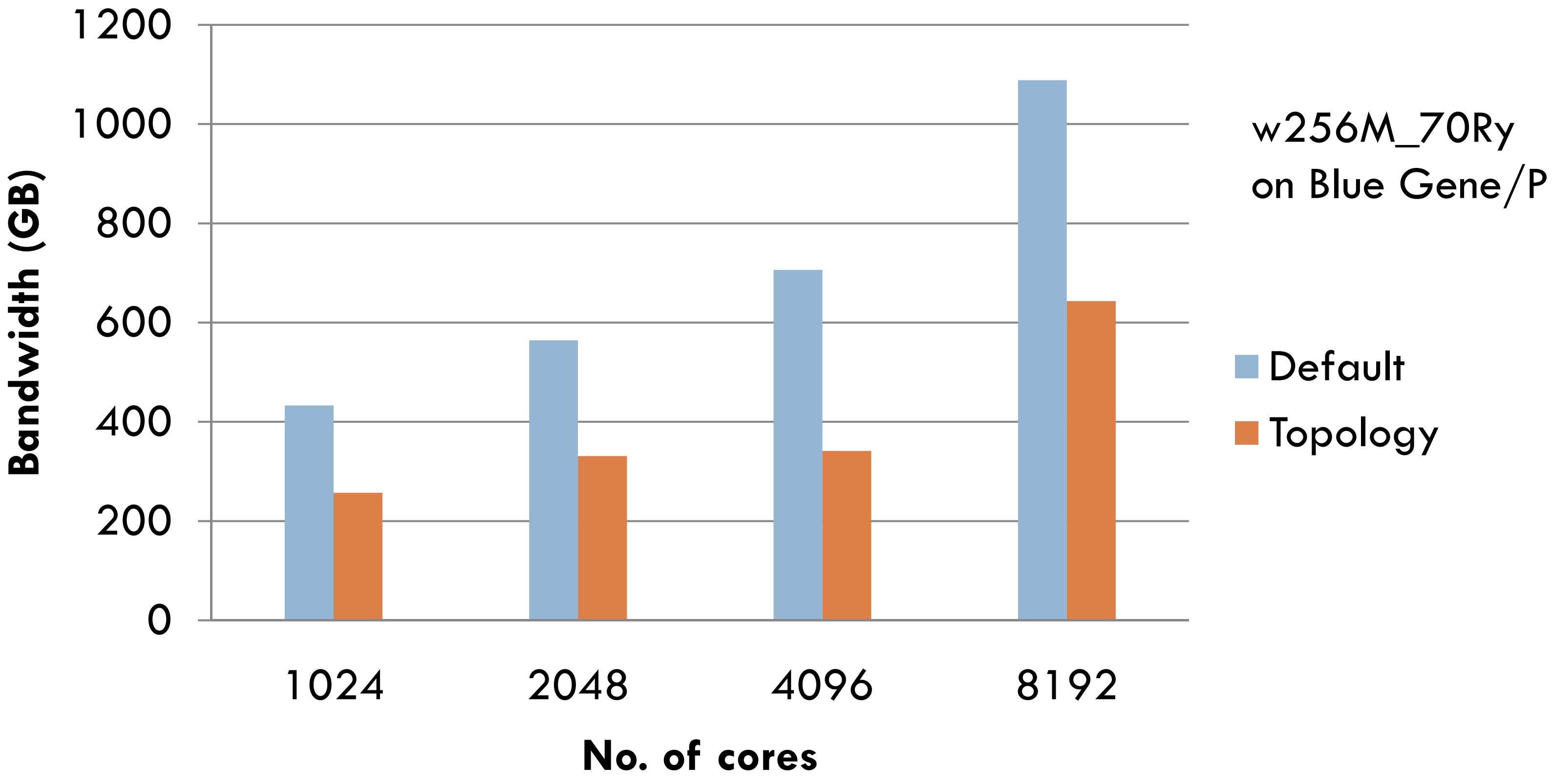
Runs on Cray XT3 (Bigben) at Pittsburgh Supercomputing Center, VN mode (with system reservation to obtain complete 3d mesh shapes)

Performance Analysis



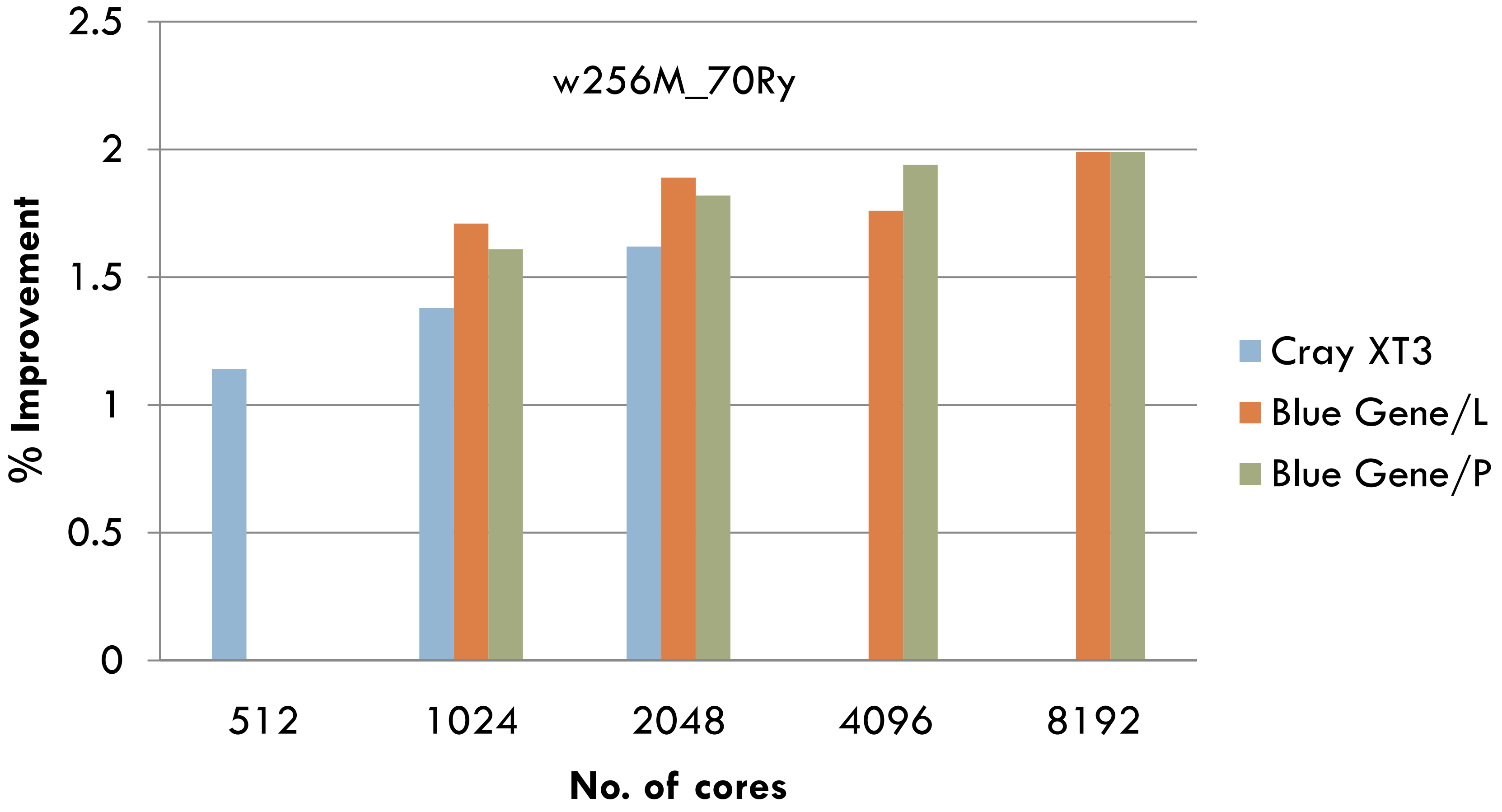
Performance Analysis and Visualization Tool: Projections – Idle time added across all processors

Reduction in Communication Volume



Data obtained from Blue Gene/P's Uniform Performance Counters

Relative Performance Improvement



Bigger picture

- Different kinds of applications:
  - Computation bound
  - Communication bound
    - Latency tolerant
    - Latency sensitive
- Technique:
  - Obtain processor topology and application communication graph
  - Heuristic Techniques for mapping

Why does distance affect message latencies?

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

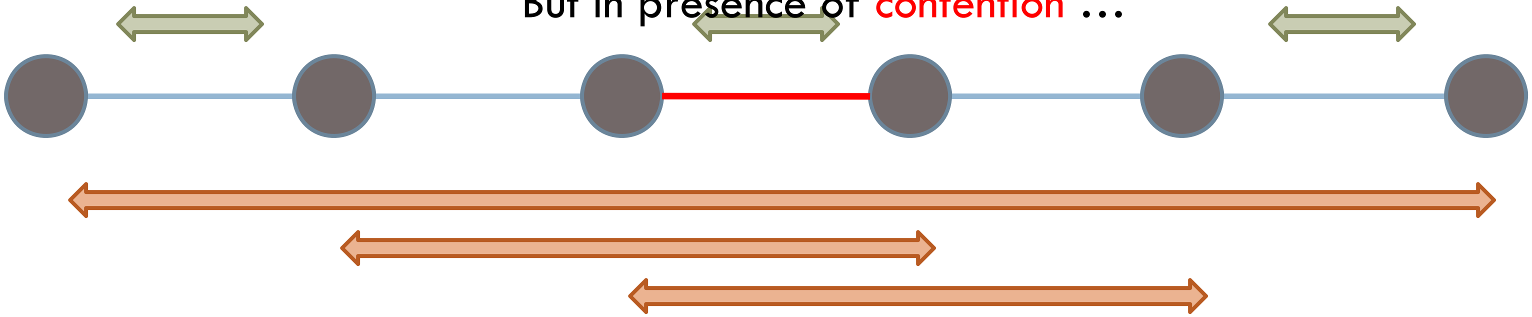
$$(L_f/B) \times D + L/B$$

L_f = length of flit, B = bandwidth,

D = hops, L = message size

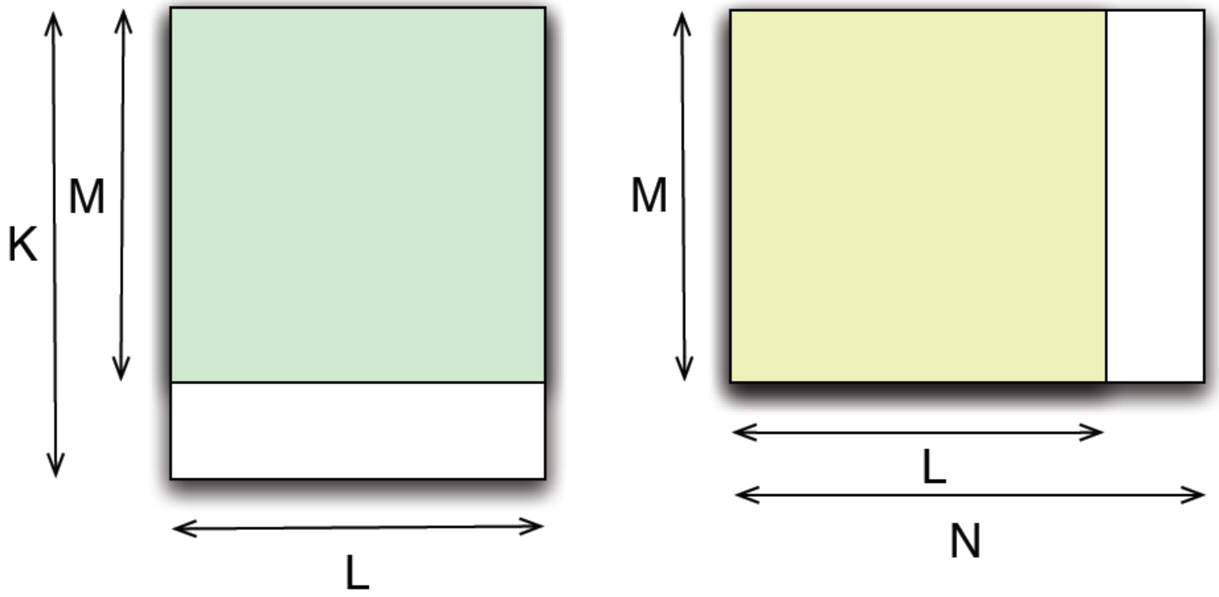
When (L_f * D) << L, first term is negligible

But in presence of contention …



Automatic Topology Aware Mapping

- Many MPI applications exhibit a simple two-dimensional near-neighbor communication pattern
- Examples: MILC, WRF, POP, Stencil, …



Acknowledgements:

- ✓ Shawn Brown and Chad Vizino (PSC)
- ✓ Glenn Martyna, Sameer Kumar, Fred Mintzer (IBM)
- ✓ Teragrid for running time on Bigben (XT3)
- ✓ ANL for running time on Blue Gene/P

E-mail: bhatele@illinois.edu
Webpage: http://charm.cs.illinois.edu

Funding

DOE Grant B341494 (CSAR), DOE Grant DE-FG05-08OR23332 (ORNL LCF) and NSF Grant ITR 0121357