Automating Topology Aware Task Mapping on Large Parallel Machines

Abhinav S Bhatele
Advisor: Laxmikant V. Kale

University of Illinois at Urbana-Champaign

SC Portland, OR 2009
Computing For A Changing World.
November 14-20, 2009
Oregon Convention Center
Portland, Oregon

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

Current Machines and their Topologies

- 3D Mesh – Cray XT3/4/5
- 3D Torus – Blue Gene/L, Blue Gene/P
- Fat-tree, CLOS network – Infiniband, Federation
- Kautz Graph – SiCortex
- Future Topologies – Blue Waters, Blue Gene/Q?

Application Characteristics

- Computation-bound applications
- Communication-heavy applications
  - Latency tolerant
  - Latency sensitive

Motivation

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

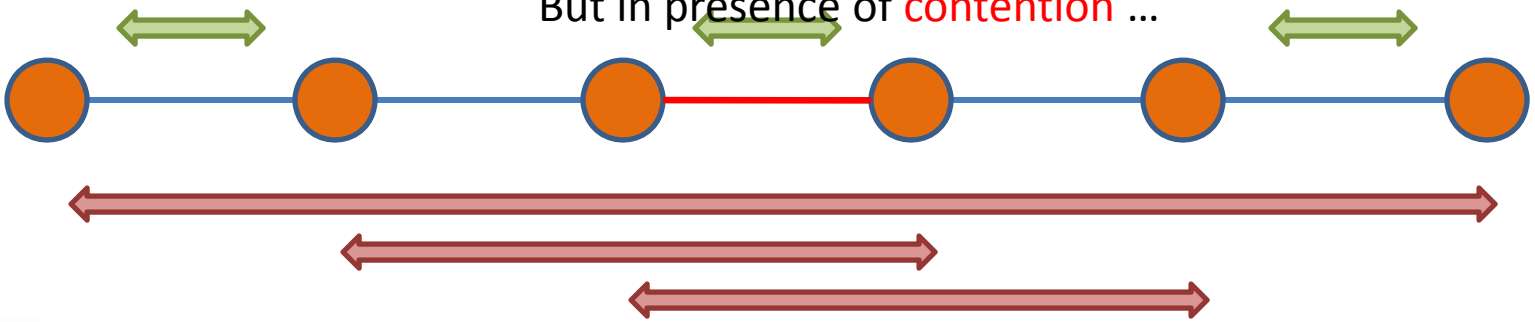
$$(L_f/B) \times D + L/B$$

L_f = length of flit, B = bandwidth,

D = hops, L = message size

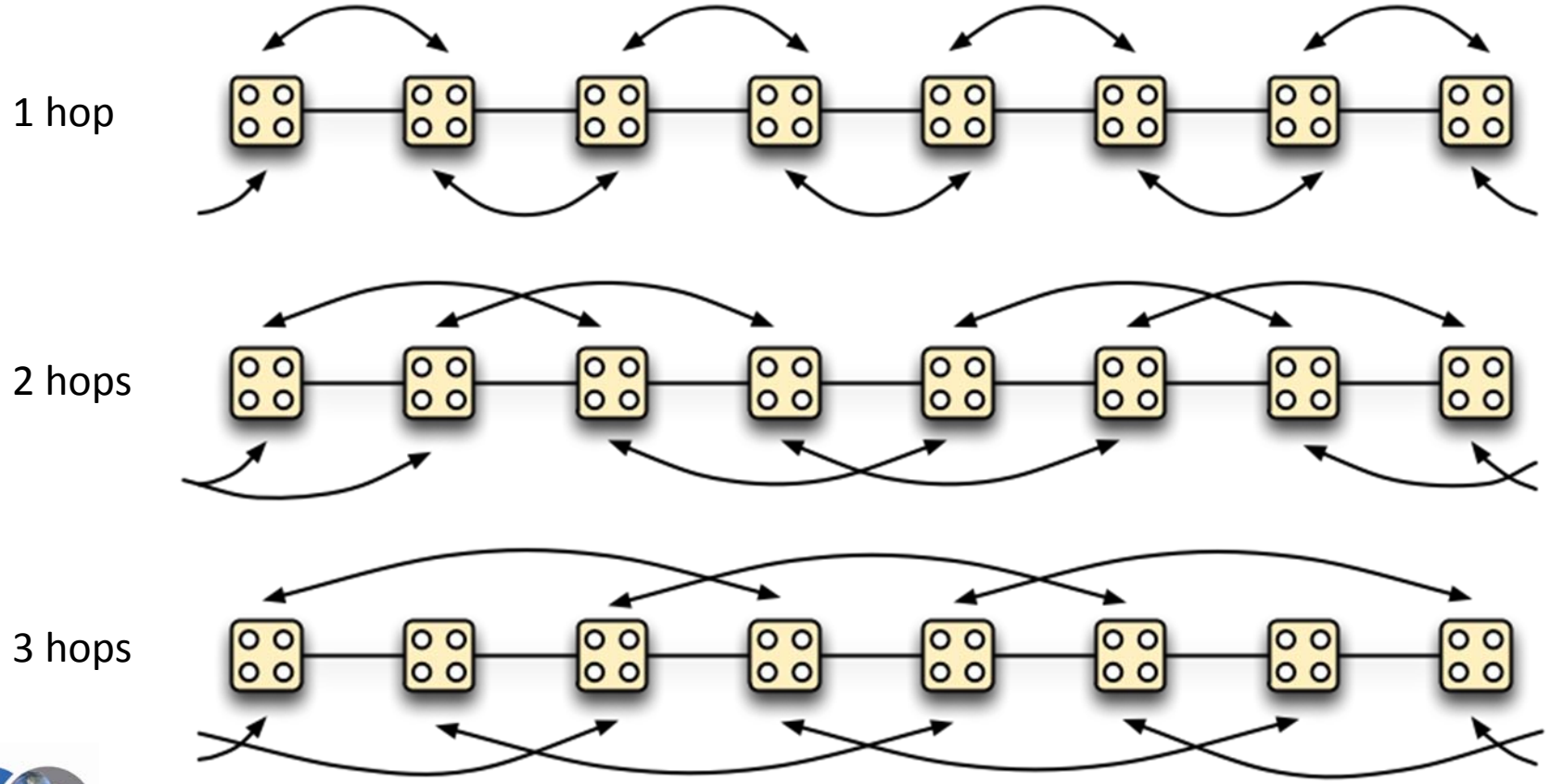
When $(L_f * D) << L$, first term is negligible

But in presence of contention …

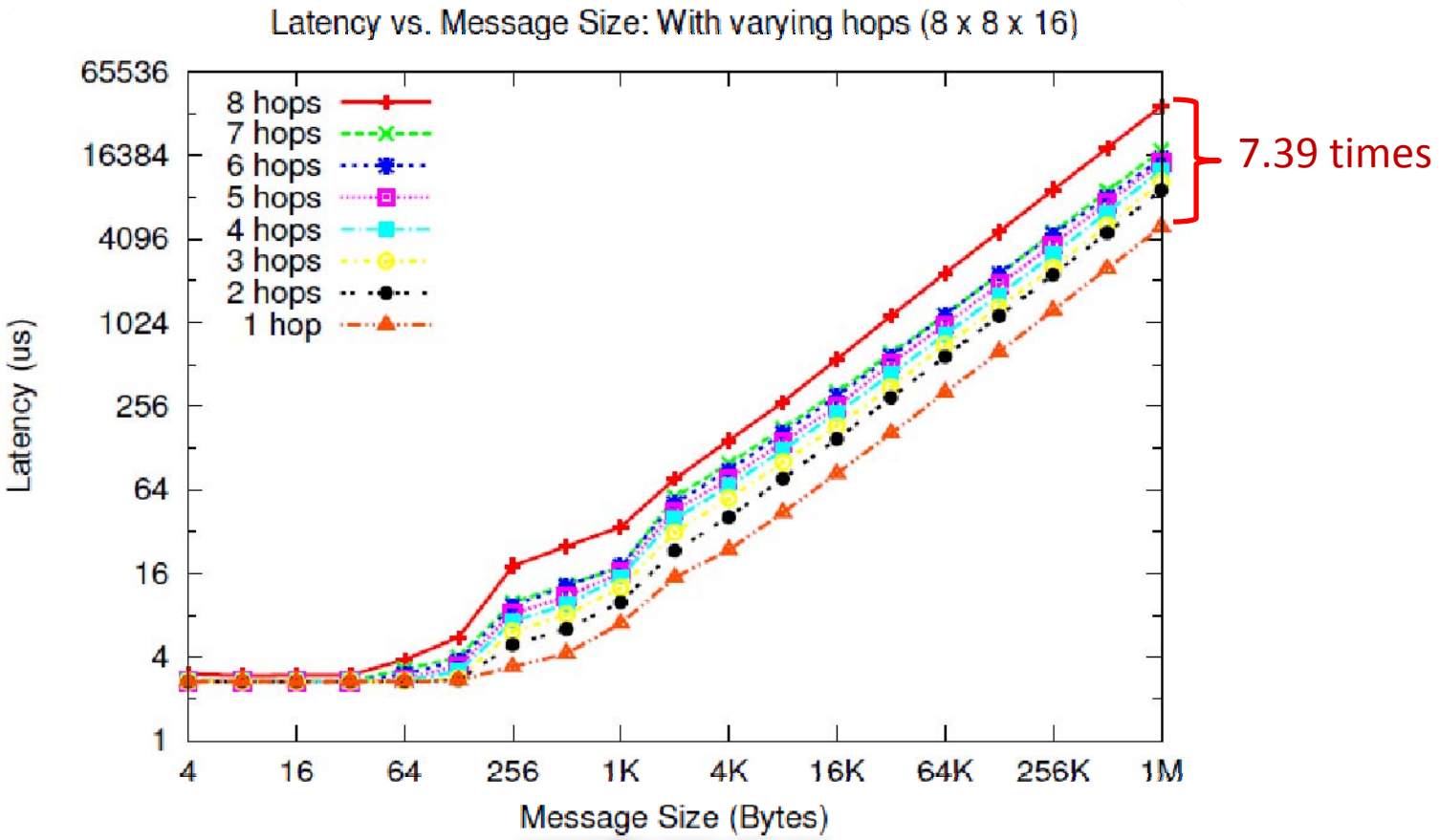


Equidistant-pairs Benchmark

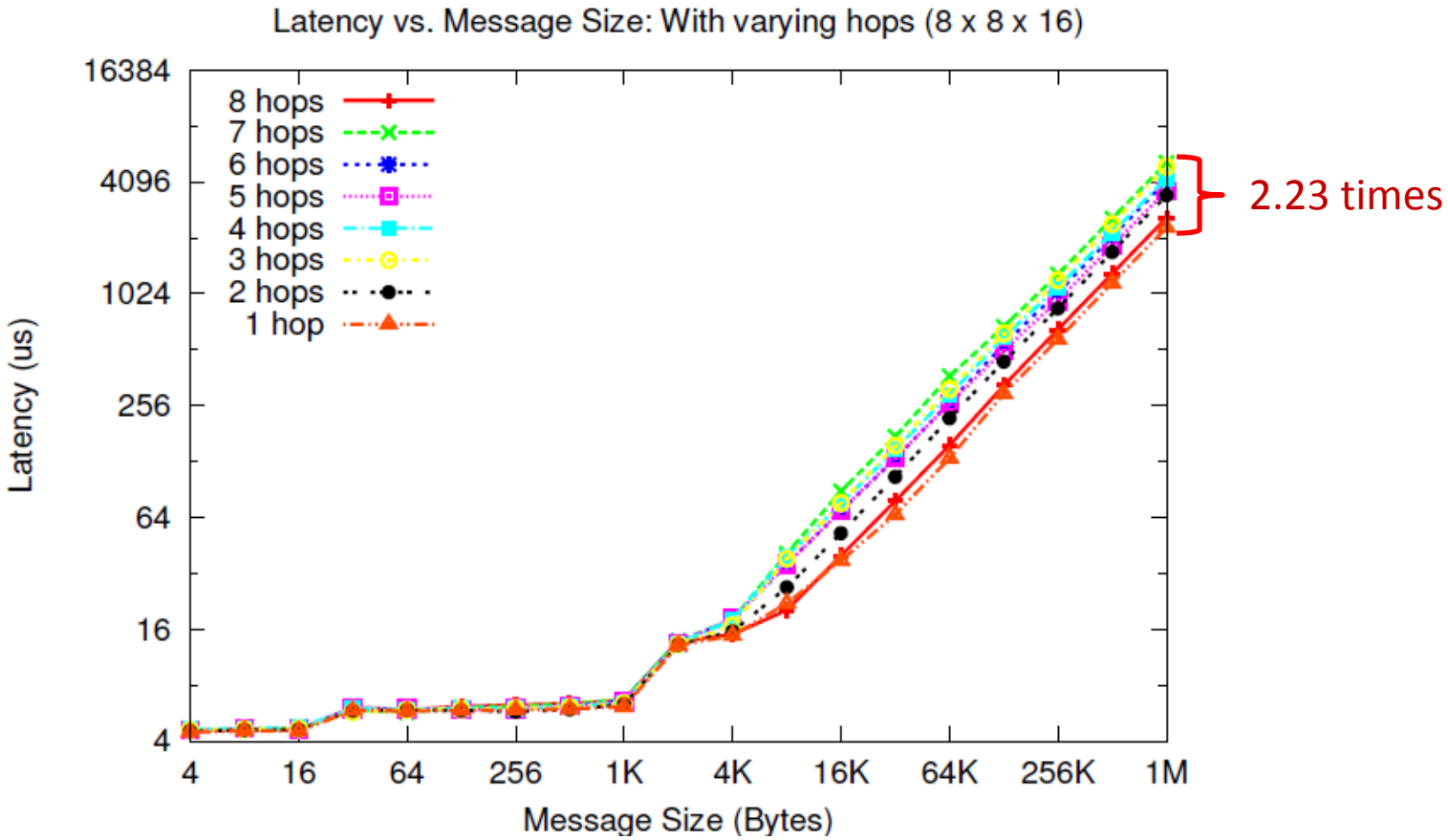
- Pair each rank with a partner which is ‘n’ hops away



Blue Gene/P


Latency vs. Message Size: With varying hops (8 x 8 x 16)
Latency (us)
65536
16384
4096
1024
256
64
16
4
1
8 hops
7 hops
6 hops
5 hops
4 hops
3 hops
2 hops
1 hop
4
16
64
256
1K
4K
16K
64K
256K
1M
Message Size (Bytes)
7.39 times

Cray XT3



Bhatele A., Kale L. V., Quantifying Network Contention on Large Parallel Machines, Parallel Processing Letters (Special Issue on Large-Scale Parallel Processing), 2009.

Automatic Mapping Framework

- Obtain the processor topology graph and communication graph for the application
- Pattern matching to identify 2D/3D/4D near-neighbor communication patterns
- Use different heuristics depending on the communication graph
  - Structured patterns
  - Irregular patterns

Topology Manager API†

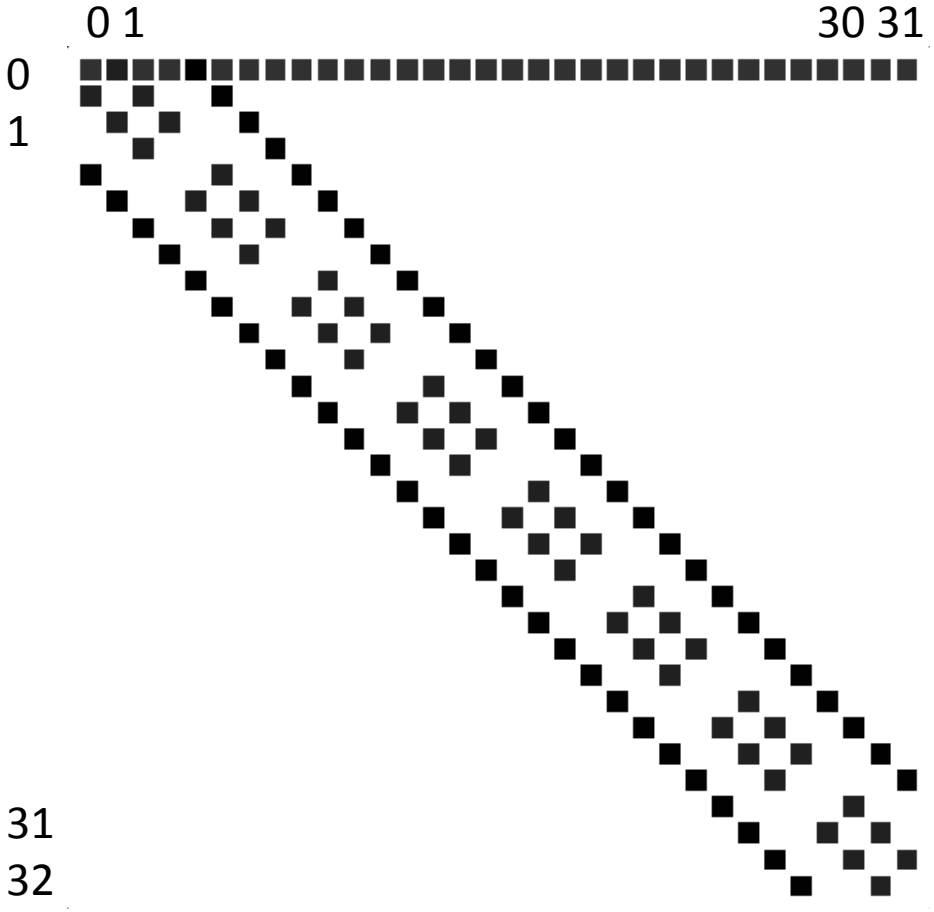
- The application needs information such as
  - Dimensions of the partition
  - Rank to physical co-ordinates and vice-versa
- TopoManager: a uniform API
  - On BG/L and BG/P: provides a wrapper for system calls
  - On XT3/4/5, there are no such system calls
  - Provides a clean and uniform interface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

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

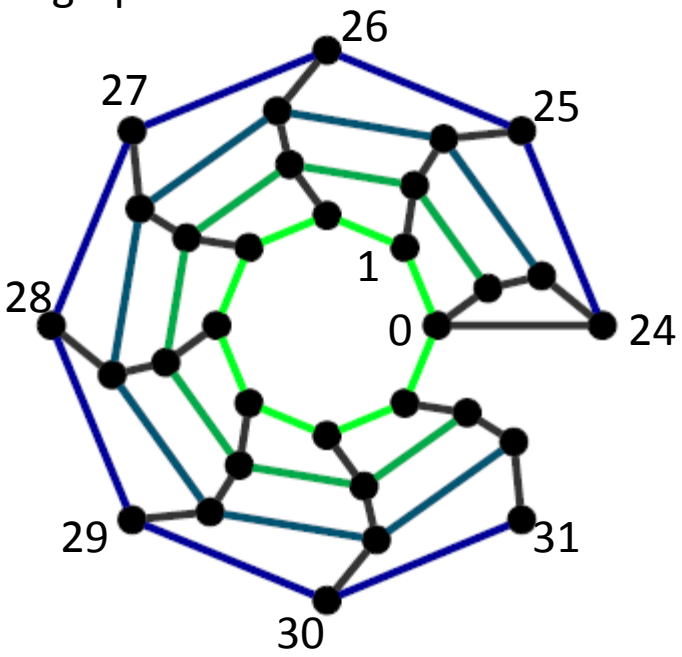
Object Communication Graph

- Obtaining this graph:
  - Manually
  - Profiling (e.g. IBM's HPCT tools)
  - Charm++'s instrumentation framework
- Visualizing the graph
- Pattern matching

WRF Communication Graph

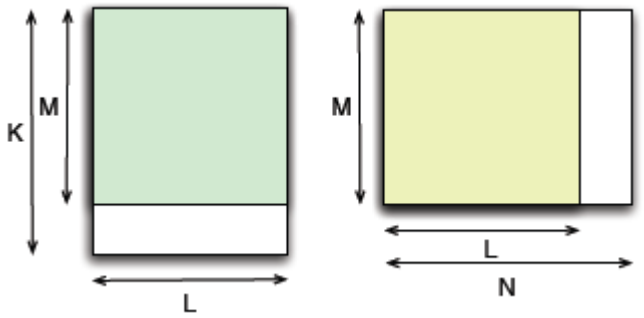


Pattern matching to find out if the communication graph is 2D and what are the dimensions of the graph?

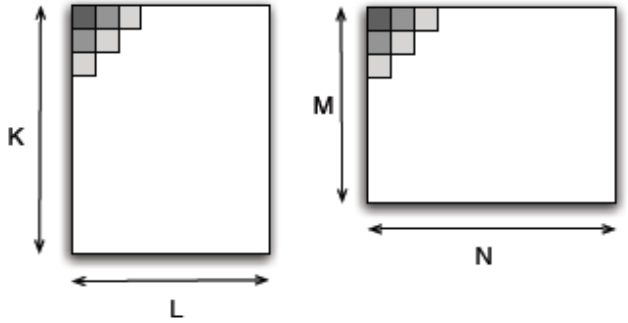


Mapping Heuristics

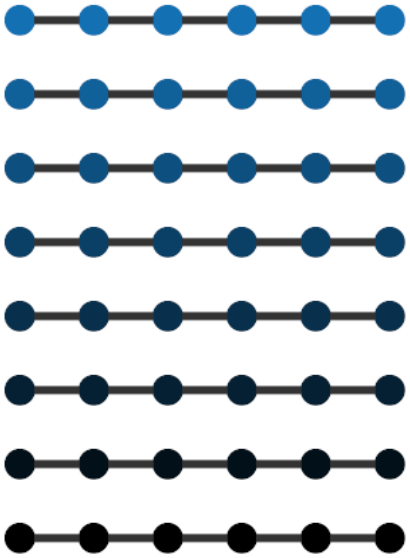
- Maximum Overlap

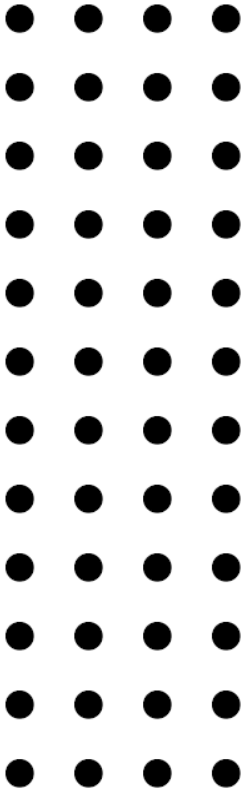


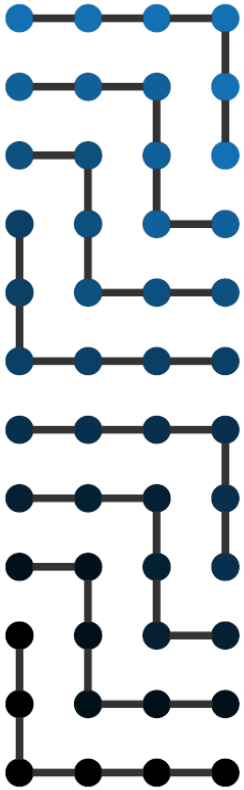
- Expand from Corners



Object Graph – 8 x 6
Processor Graph – 12 x 4

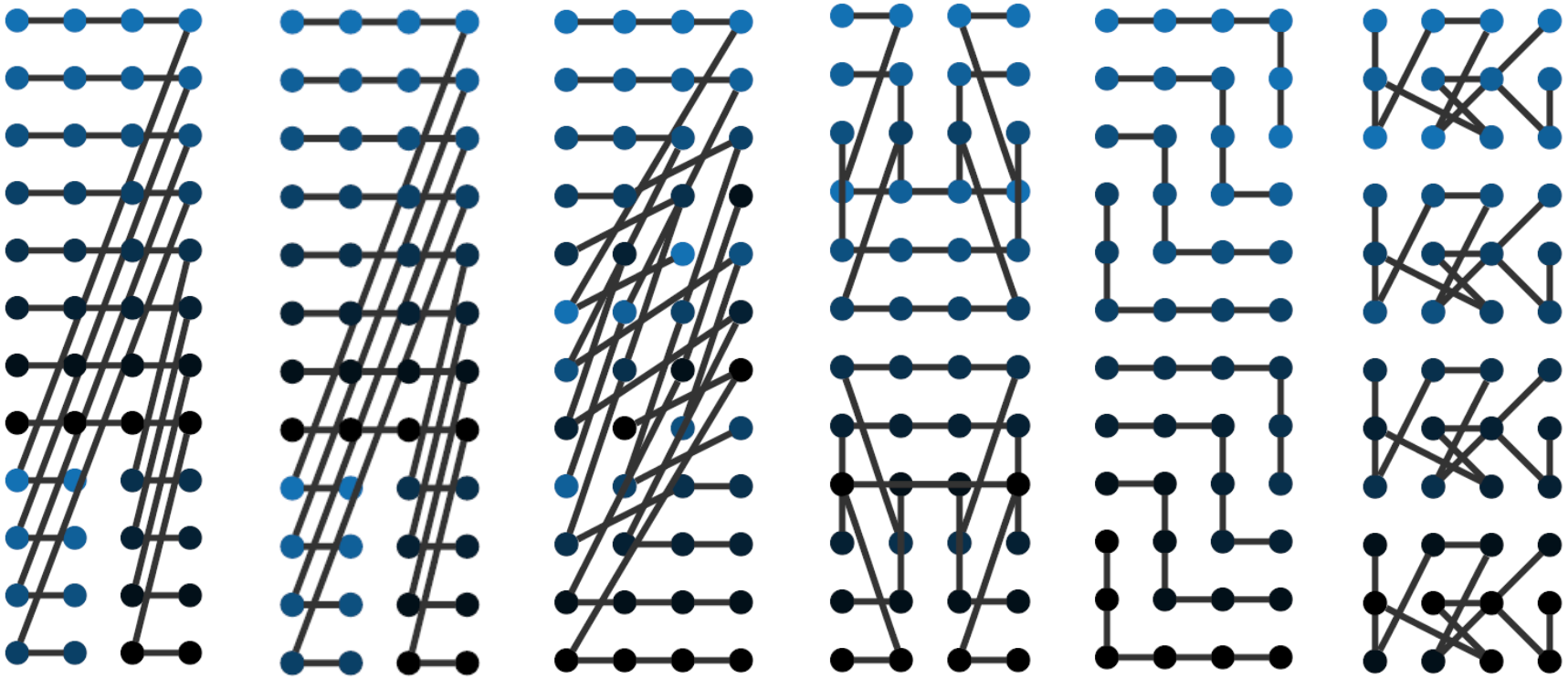






Aleliunas, R. and Rosenberg, A. L. On Embedding Rectangular Grids in Square Grids. IEEE Trans. Comput., 31(9):907–913, 1982

Different mapping heuristics



Bhatele A., Chung I., Kale L. V., Automated Mapping of Structured Communication Graphs onto Mesh Interconnects, in preparation, 2009.

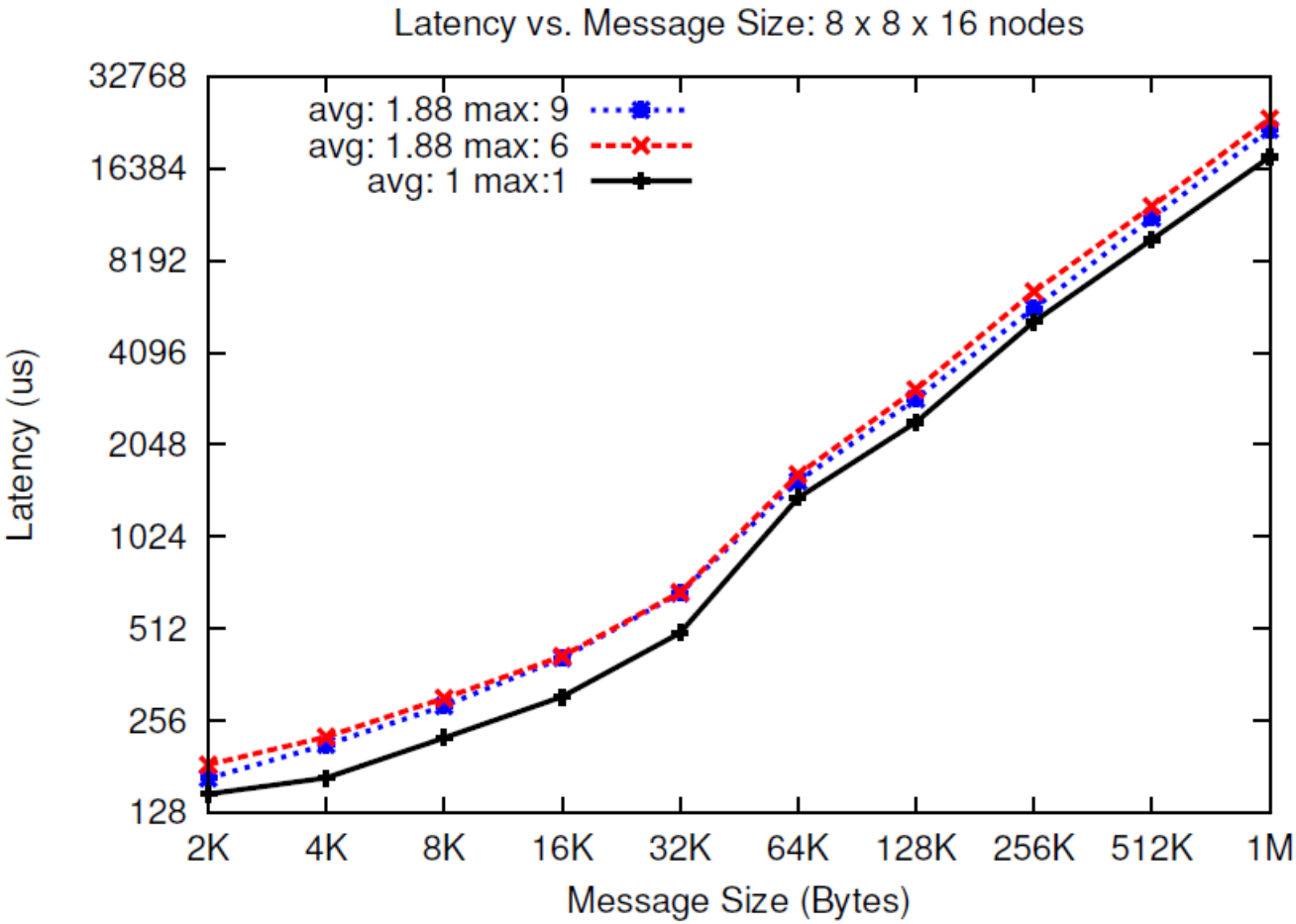
Evaluation Metric: Hop-bytes

- Weighted sum of message sizes where the weights are the number of links traversed by each message

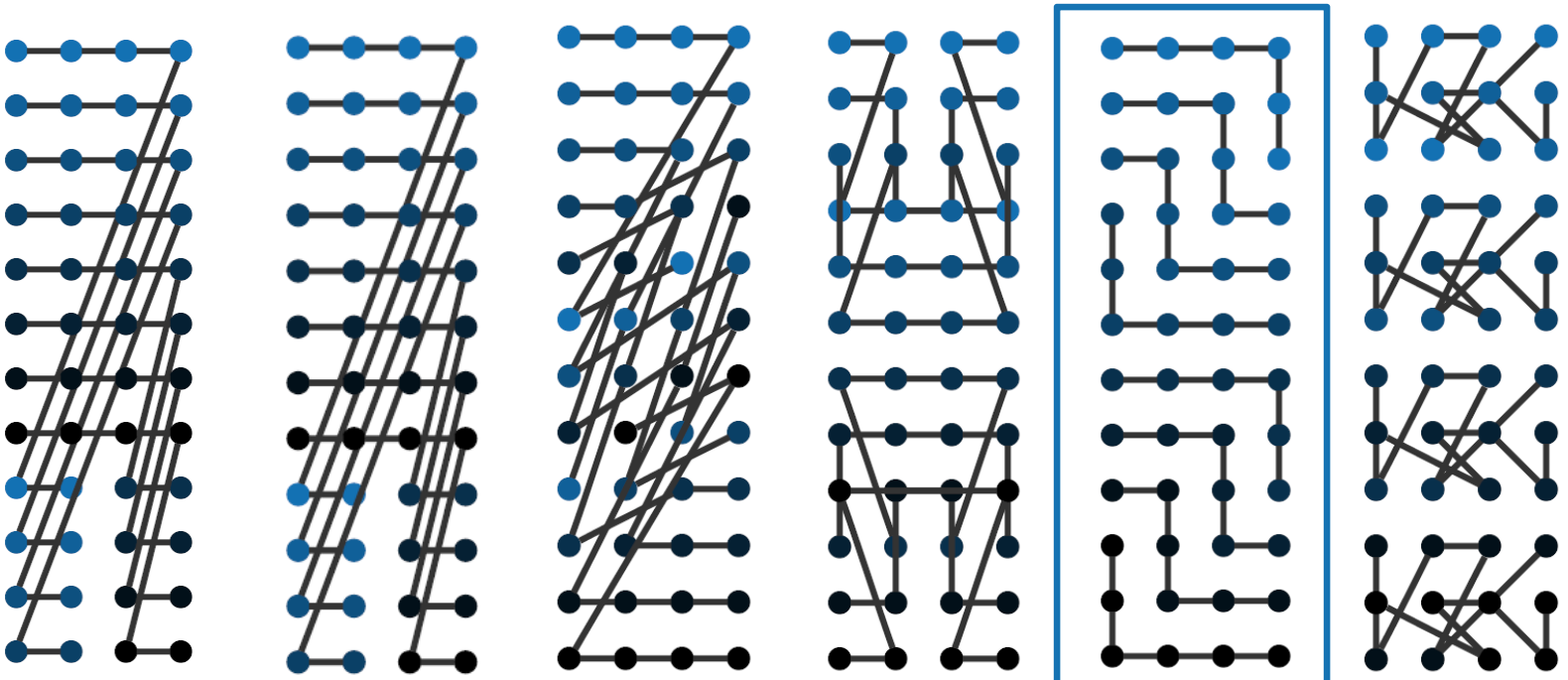
$$HB = \sum_{i=1}^{n} d_i \times b_i$$

- Indication of the communication traffic on the network
- Another metric: maximum dilation

Blue Gene/P (Intrepid)



Evaluation



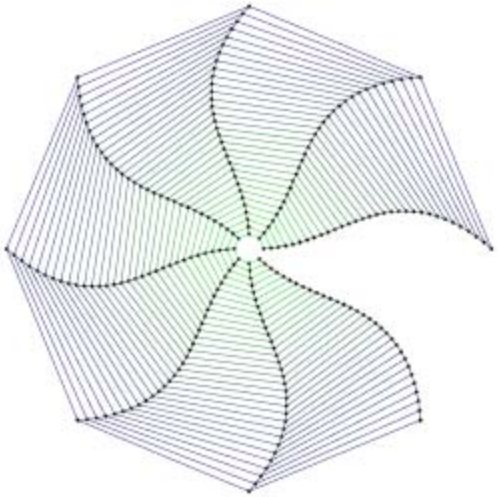
Hops	292	292	432	284	236	348
Dilation	11	11	8	7	3	5

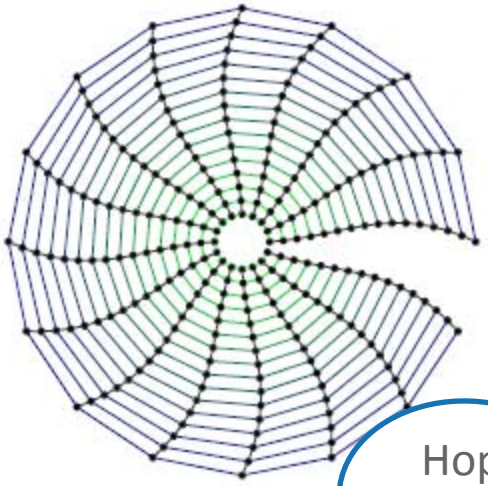
Mapping of MPI Applications

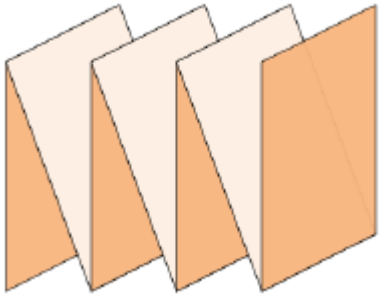
- Work with IBM (I-Hsin Chung)
  - Using HPCT to dump communication patterns
  - Derive a mapping offline and use in a subsequent run
- Applications: MILC, POP, WRF
  - Map 2D communication patterns to 3D tori of BG/P

Communication graphs for POP and WRF on 256 processors

Folding of 2D graph to 3D mesh





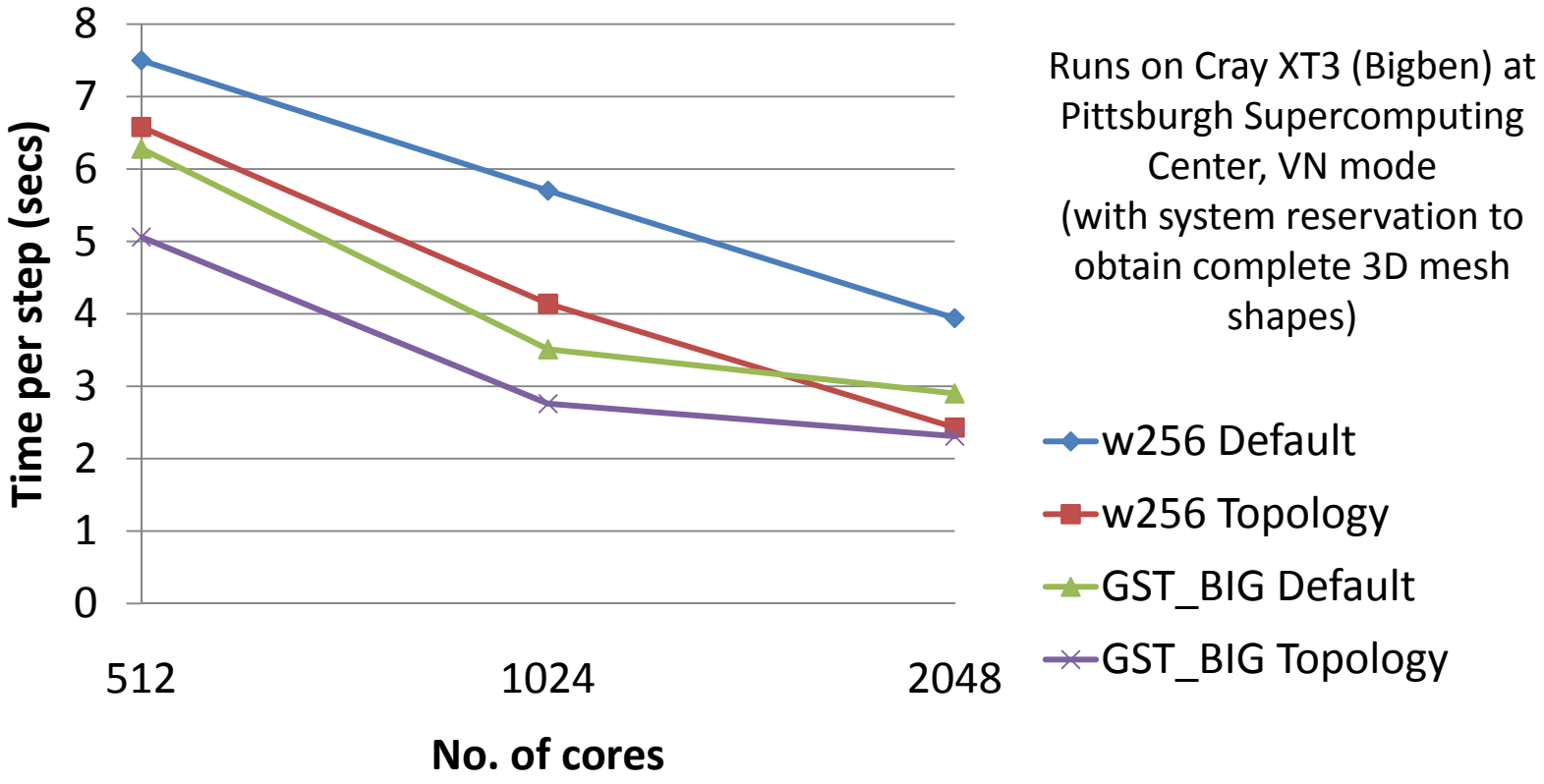


Application	Graph	VN mod XYZT	TXYZ				
MILC	$4 \times 4 \times 4 \times 4$	2304	2304				
POP	8×32	2208	1632	672			1056
POP	32×16	4032	2176	1088	3	2560	2200
WRF	16×16	1728	1152	832	1296	1344	1096
WRF	32×32	9216	4096	2688	11648	4608	4376

Hops Reduction – 64%
Communication Time Reduction – 45%
Performance Improvement - 17%

* FOLD - H. Yu, I.-H. Chung, and J. Moreira. Topology mapping for Blue Gene/L supercomputer. In SC '06: page 116, New York, NY, USA, 2006.

OpenAtom Performance on Cray XT3



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

Remaining and Future Work

- Consider weighted communication graphs
- Mapping of irregular communication graphs
  - Unstructured mesh applications, MD codes
- Future Work
  - Dynamic Load Balancing for MPI applications
  - Complex topologies of the future

I am on the job market …

Acknowledgements:

Prof. Laxmikant V. Kale
Prof. William D. Gropp
Prof. David A. Padua
Dr. Matthew H. Reilly

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, Tisha Stacey
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, kale @ illinois.edu

Webpage: http://charm.cs.illinois.edu/~bhatele