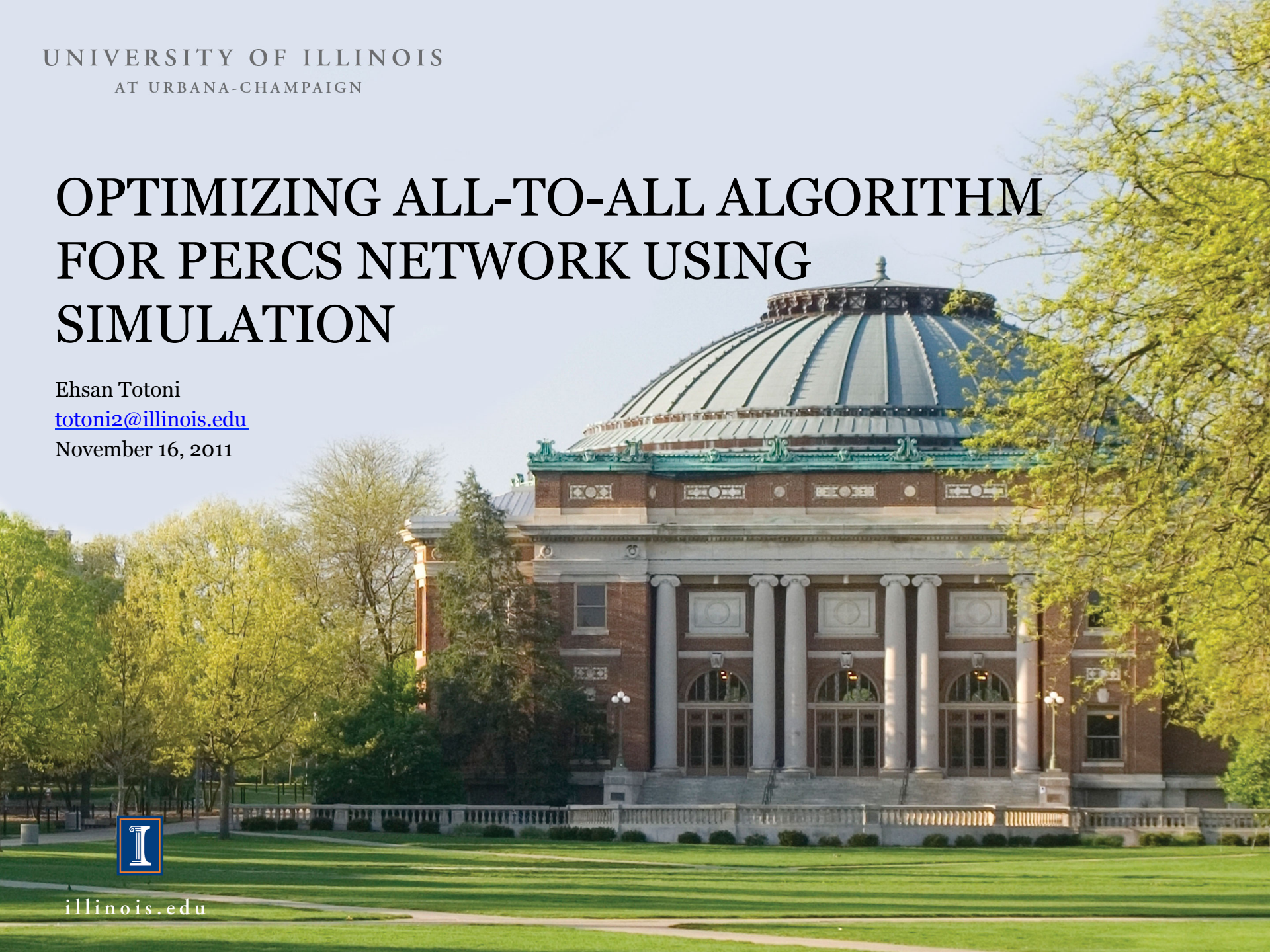
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

OPTIMIZING ALL-TO-ALL ALGORITHM FOR PERCS NETWORK USING SIMULATION

Ehsan Totoni
totoni2@illinois.edu
November 16, 2011



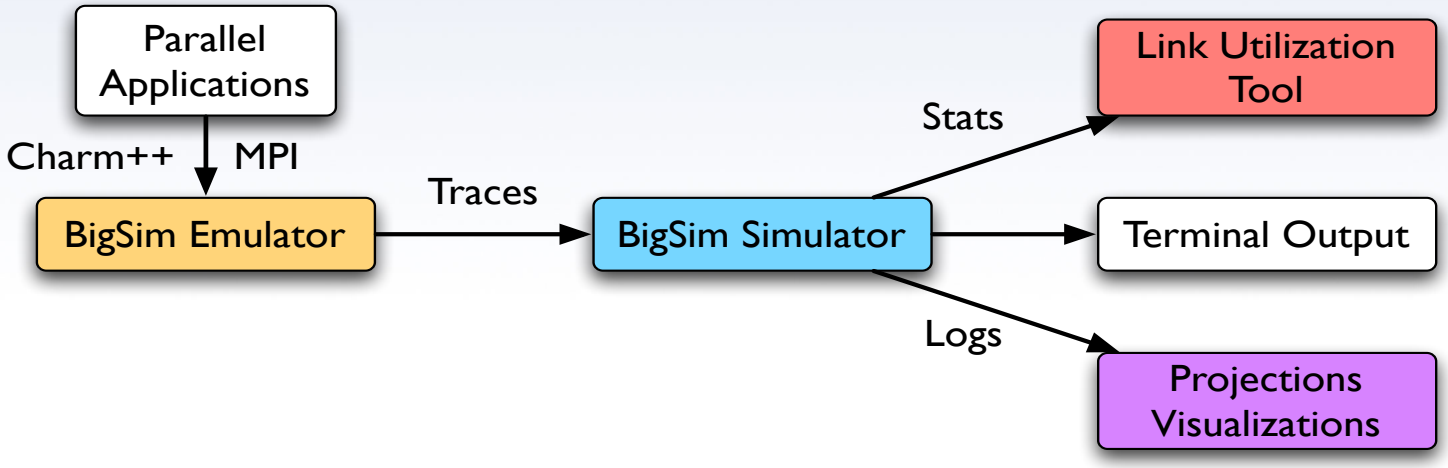
illinois.edu

Introduction

- Hundreds of millions of dollars are spent on large supercomputers
  - Tuning runtimes and applications can take months to years
    - Specially with new networks e.g. 6D Torus, PERCS
  - Huge waste of resources!
- Our approach: use simulation to tune before the machine comes online (or even after)
  - Case study: MPI_Alltoall on PERCS

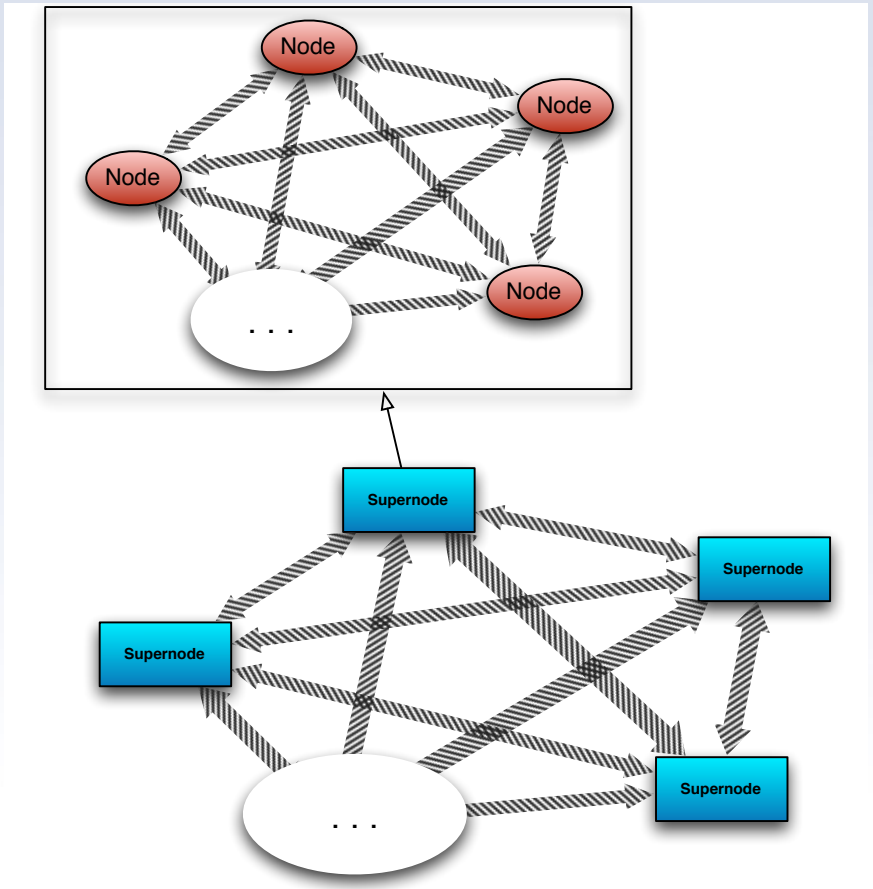
BigSim

- Emulate Charm++ or MPI application at scale with processor virtualization
- Accurate packet-level or simple network simulation
  - Different outputs, e.g. time prints, link utilization
  - PERCS model with IBM



PERCS Architecture

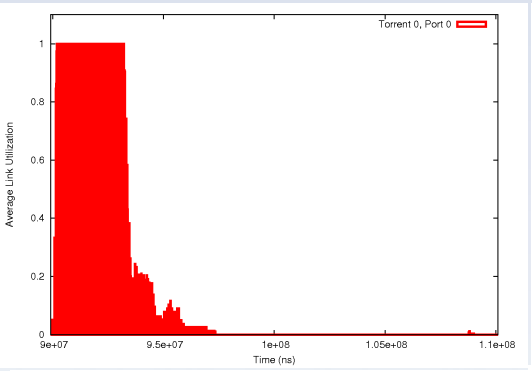
- DARPA machines, IBM Power 775
- Two level network
  - 32 fully-connected nodes-> Supernode
  - Fully-connected supernodes
- A lot of details:
  - 24-GB/s LL-links, 5-GB/s LR links, 10-GB/s D-links
  - 4 Power7 chips have 192-GB/s to Hub Chip

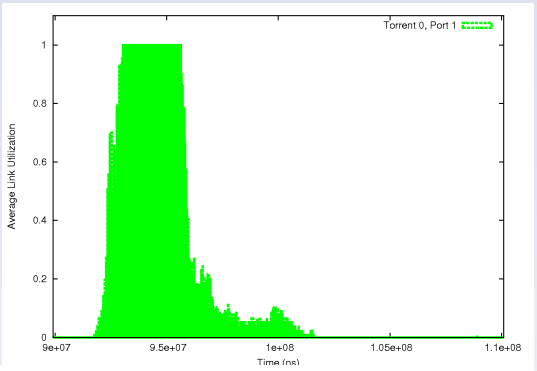


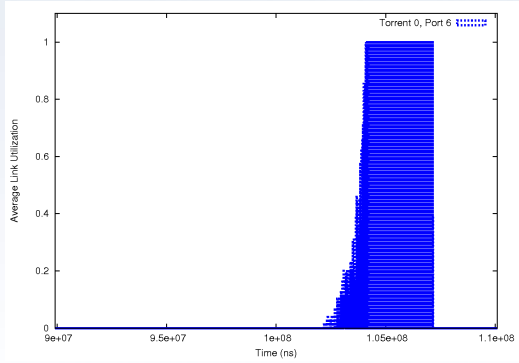
MPI_Alltoall on PERCS

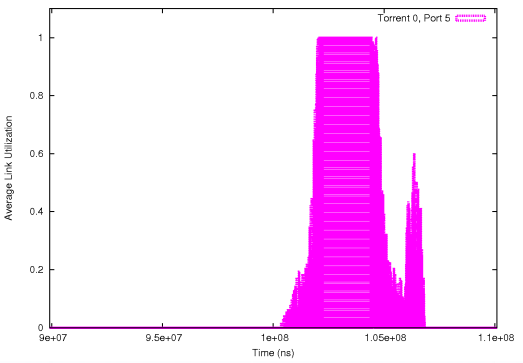
- Collectives are essential for HPC applications
- Many applications can benefit from optimizing communication library
- MPI_Alltoall inside a Supernode is important (with fully-connected links)
  - Data exchange in groups in FFT
- BigSim shows links are not utilized efficiently with Pairwise Exchange algorithm

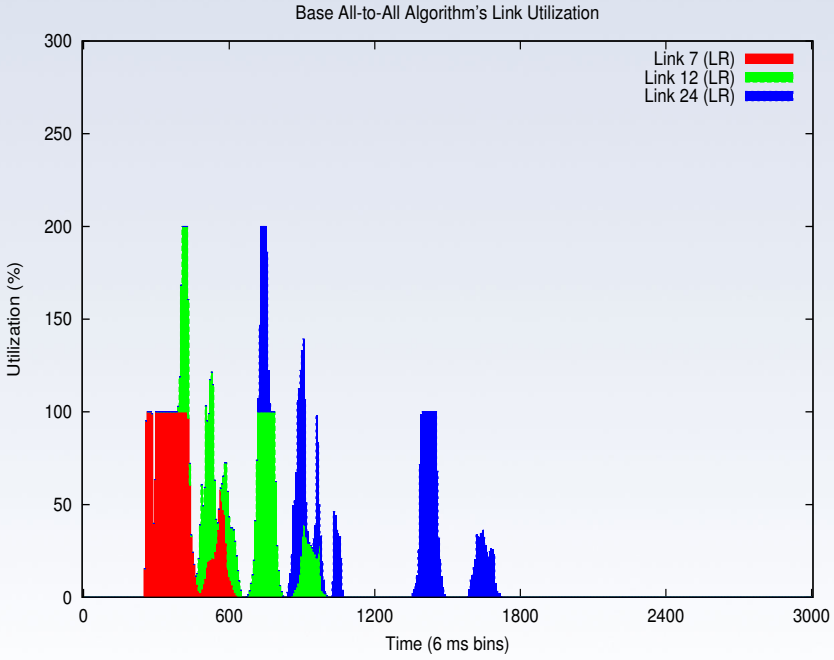
BigSim Output











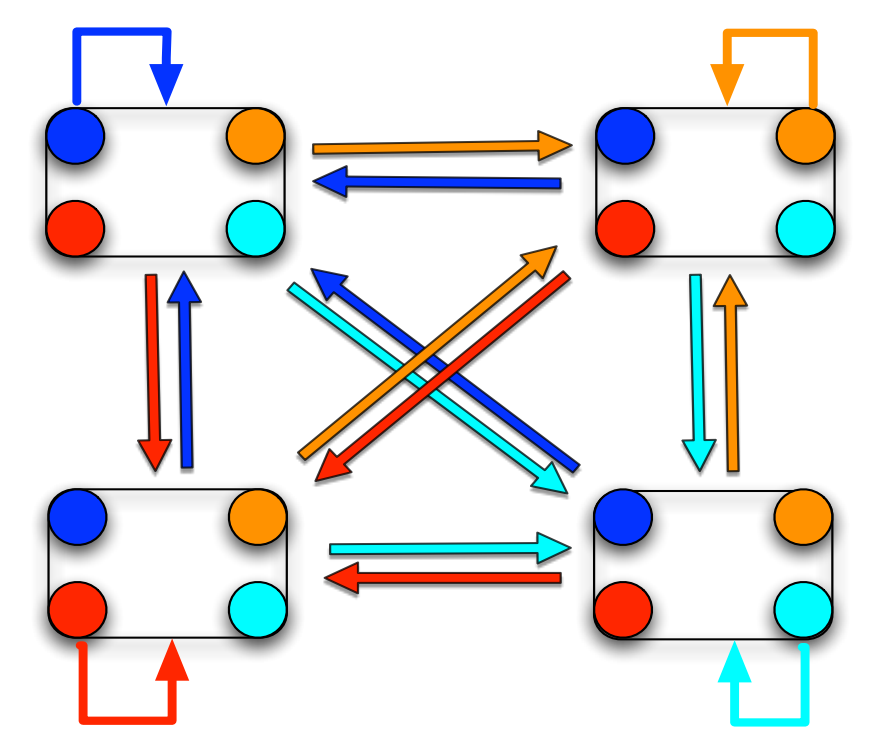
illinois.edu

Methodology

- Simulate algorithm and analyze different output
  - Link utilizations over time for communication algorithms
- Evaluate according to performance criteria
  - Each link is highly utilized during the whole Alltoall period
- Change algorithm if not satisfactory and go back to step one
  - Try to exploit network and node properties

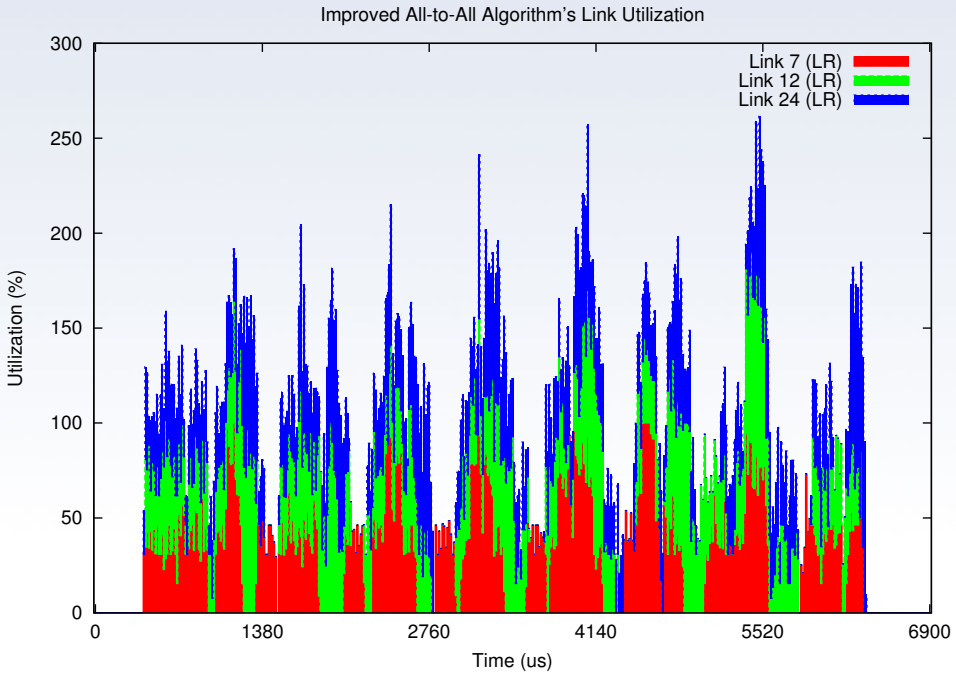
New Alltoall Algorithm

- Use all links, avoid link congestion
- Solution: each core send to a different node



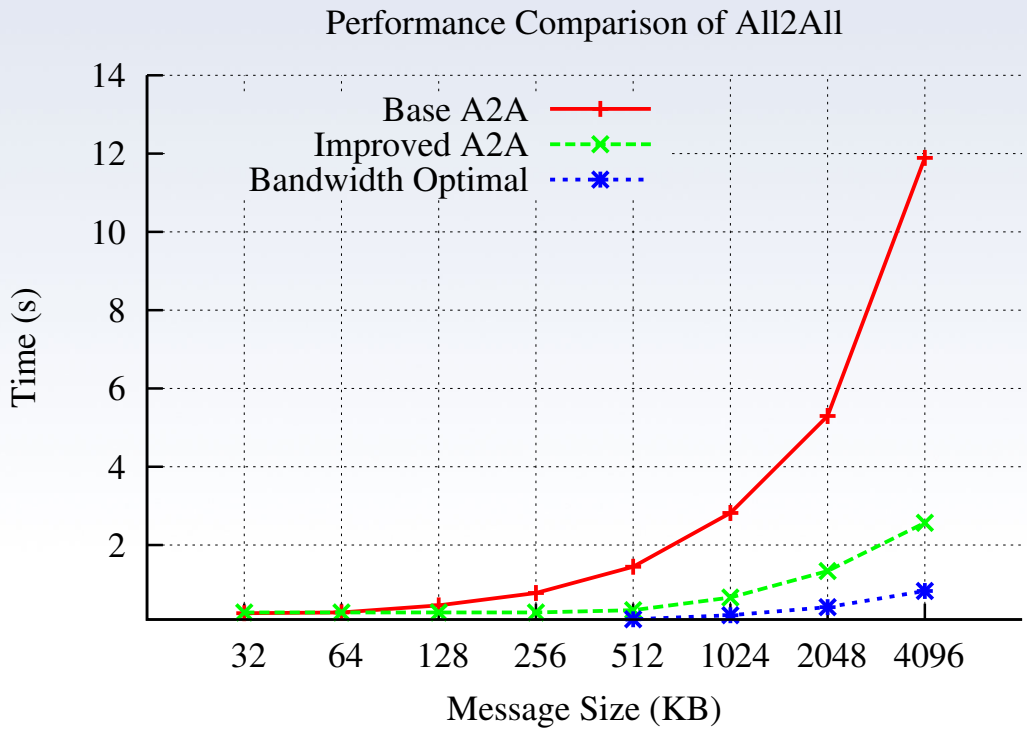
Utilization Improvement

- Links are now utilized simultaneously for the whole Alltoall duration



Performance Result

- Up to 5 times improvement for large messages!



Reference

- **”Simulation-based Performance Analysis and Tuning for a Two-level Directly Connected System”,** E. Totoni, A. Bhatele, E. J. Bohm, N. Jain, C. Mendes, R. Mokos, G. Zheng, L. V. Kale, to appear in ICPADS 2011
- http://charm.cs.illinois.edu