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

Ehsan Totoni

Laxmikant V. Kale

{totoni2, kale}@illinois.edu

Communication algorithms play a crucial role in the performance of large-scale parallel systems. They are implemented in runtime systems and used in most parallel applications as a critical component. As vendors are willing to design new custom networks with significantly different performance properties for their new supercomputers, designing new efficient communication algorithms is an inevitable challenge. This task is desirable to be done before the machine comes online since inefficient use of the system before the new algorithm's availability is a huge waste of a possibly hundreds of millions of dollars resource. In this poster, we demonstrate the usability of our simulation framework, BigSim, in meeting this challenge. Using BigSim, we observe that the commonly used Pairwise-Exchange algorithm for all-to-all communication pattern is suboptimal for the PERCS network. We designed a new all-to-all algorithm for it and predict a five-fold performance improvement for large message sizes using this algorithm.

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

PARALLEL PROGRAMMING LAB

DEPT. OF COMPUTER SCIENCE, UNIVERSITY OF ILLINOIS

PPL UIUC

http://charm.cs.illinois.edu

PERCS Network

New innovative networks (such as PERCS) are becoming popular for larger machines!

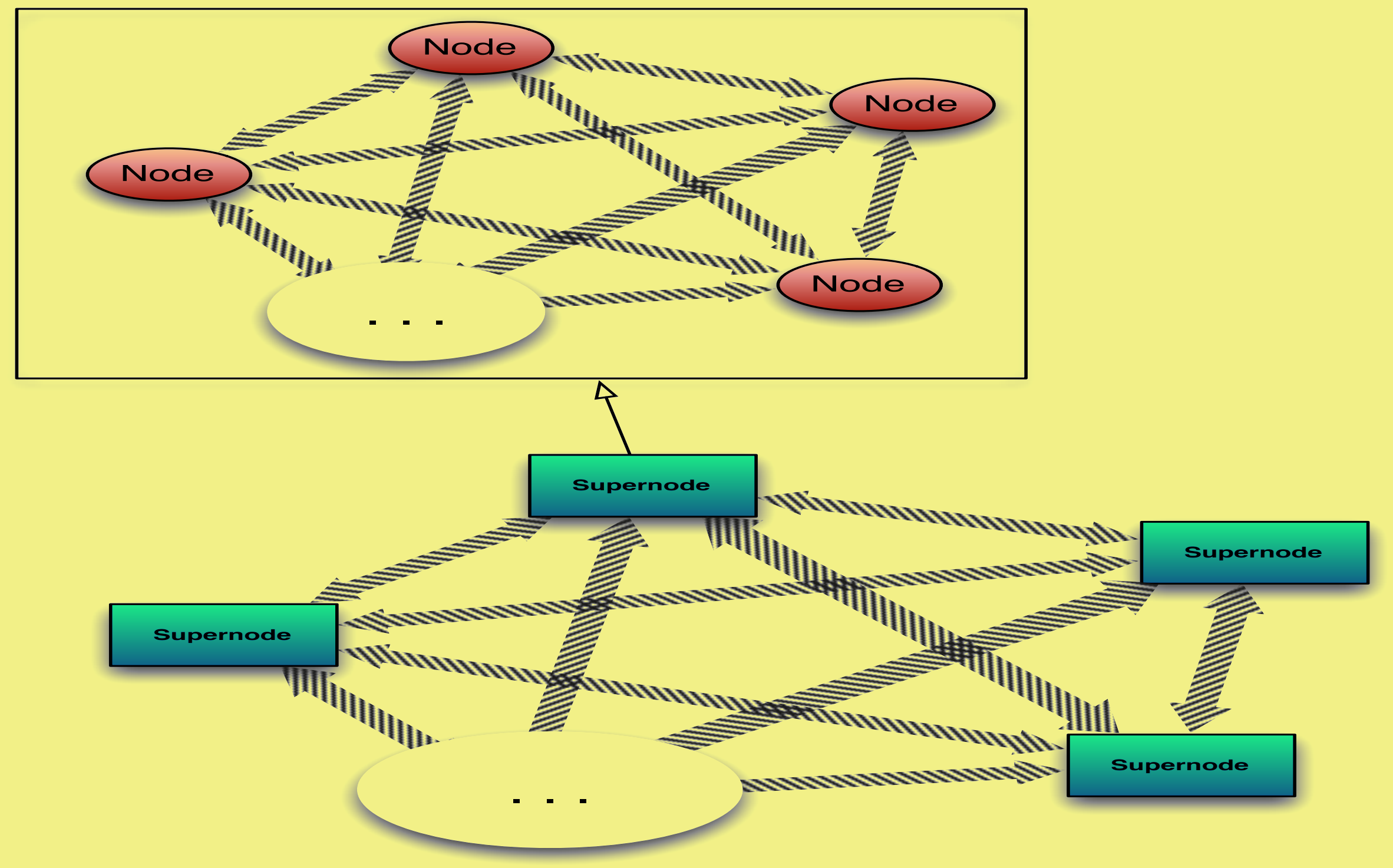
- Rather than well known ones such as 3D Torus

No legacy theory or experience of them!

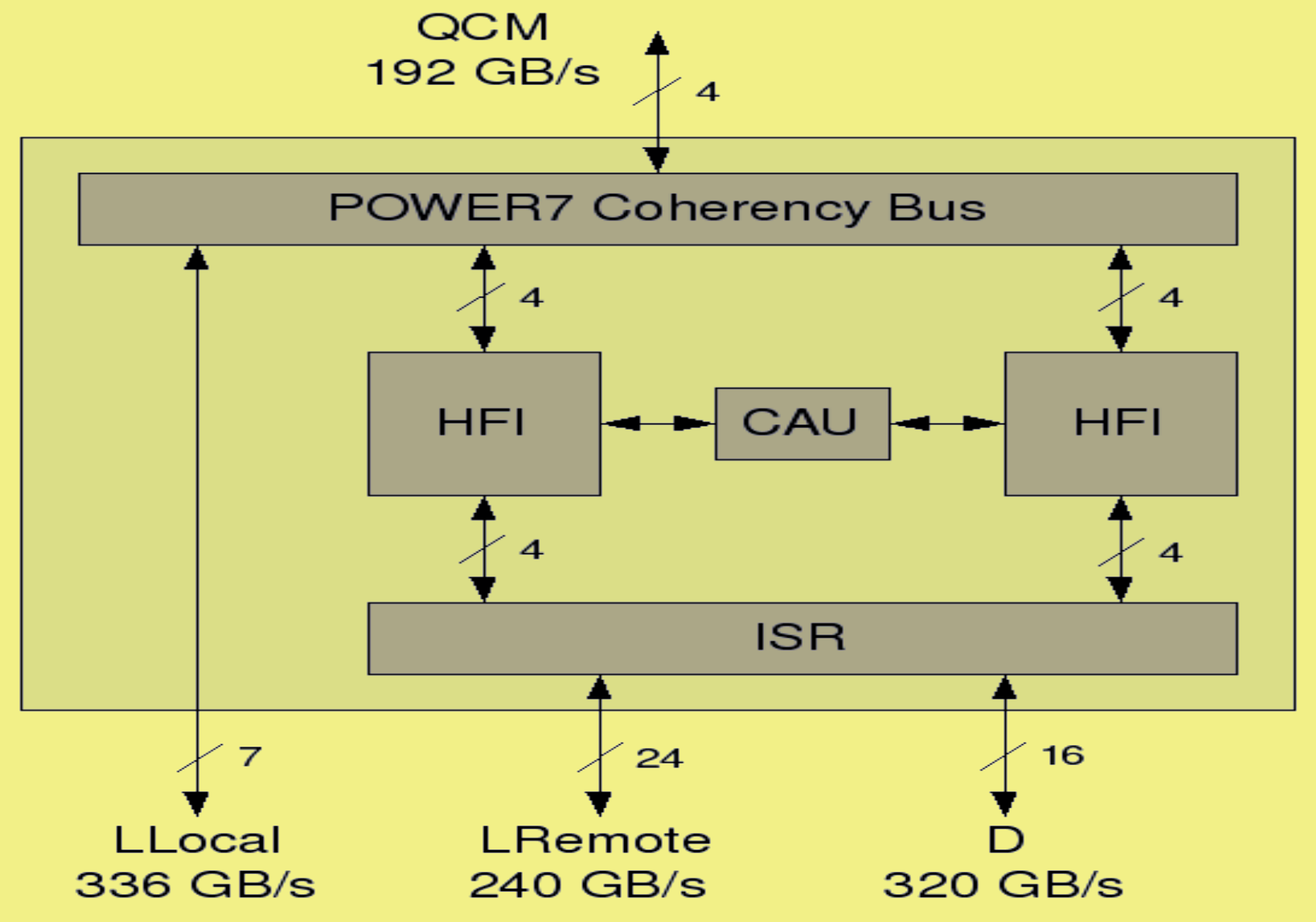
We need to prepare for them before the hundred million dollor machines come online!

PERCS Network will be used in many IBM machines

- It is a two level (**complicated**) network!
- Fully connected "Supernodes": 32 fully-connected nodes
- 24-GB/s LLocal (LL) links for nodes within a drawer (8 nodes)
- 5-GB/s LRemote (LR) links across drawers within a Supernode, 10-GB/s D links across Supernodes



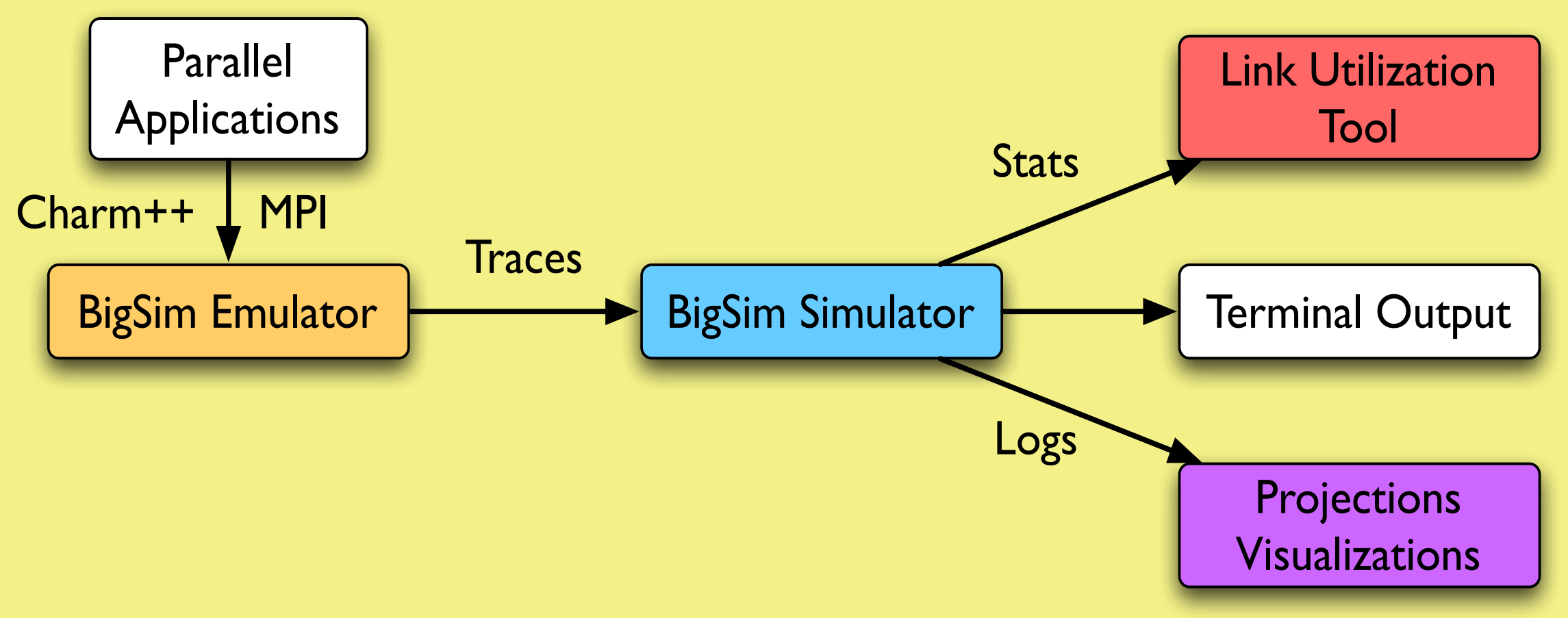
- 4 POWER7 chips, each with 8 cores form a compute node 192 GB/s to a Hub Chip in a Quad Chip Module (QCM)
- Different components inside the Hub Chip: HFI, ISR ...



BigSim

BigSim is a simulation-based framework used for simulating the behavior of real applications on large parallel machines

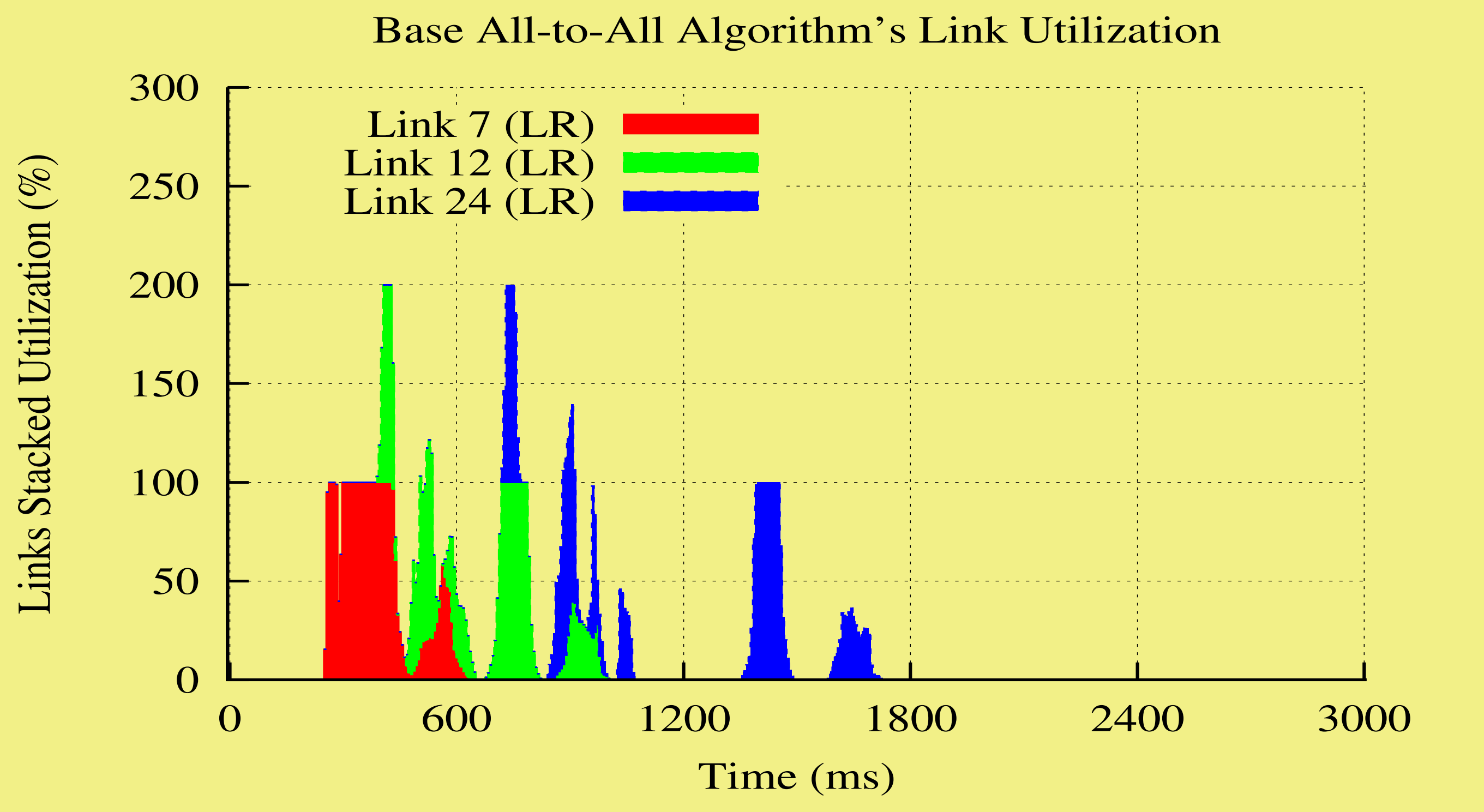
- Charm++ or MPI application is run on the BigSim emulator
- BigSim is built on Charm++, which allows it to use Charm++'s processor virtualization ability to emulate multiple target processors on each physical processor



BigSim simulator uses a network model to get actual times

- These network models include objects that represent processors, nodes, network interface cards, switches, and network links and can simulate contention in the network
- We have built a BigSim model of the PERCS network and validated it against IBM's MERCURY simulator and hardware prototypes

BigSim shows that links are not utilized efficiently using the common Pairwise Exchange algorithm

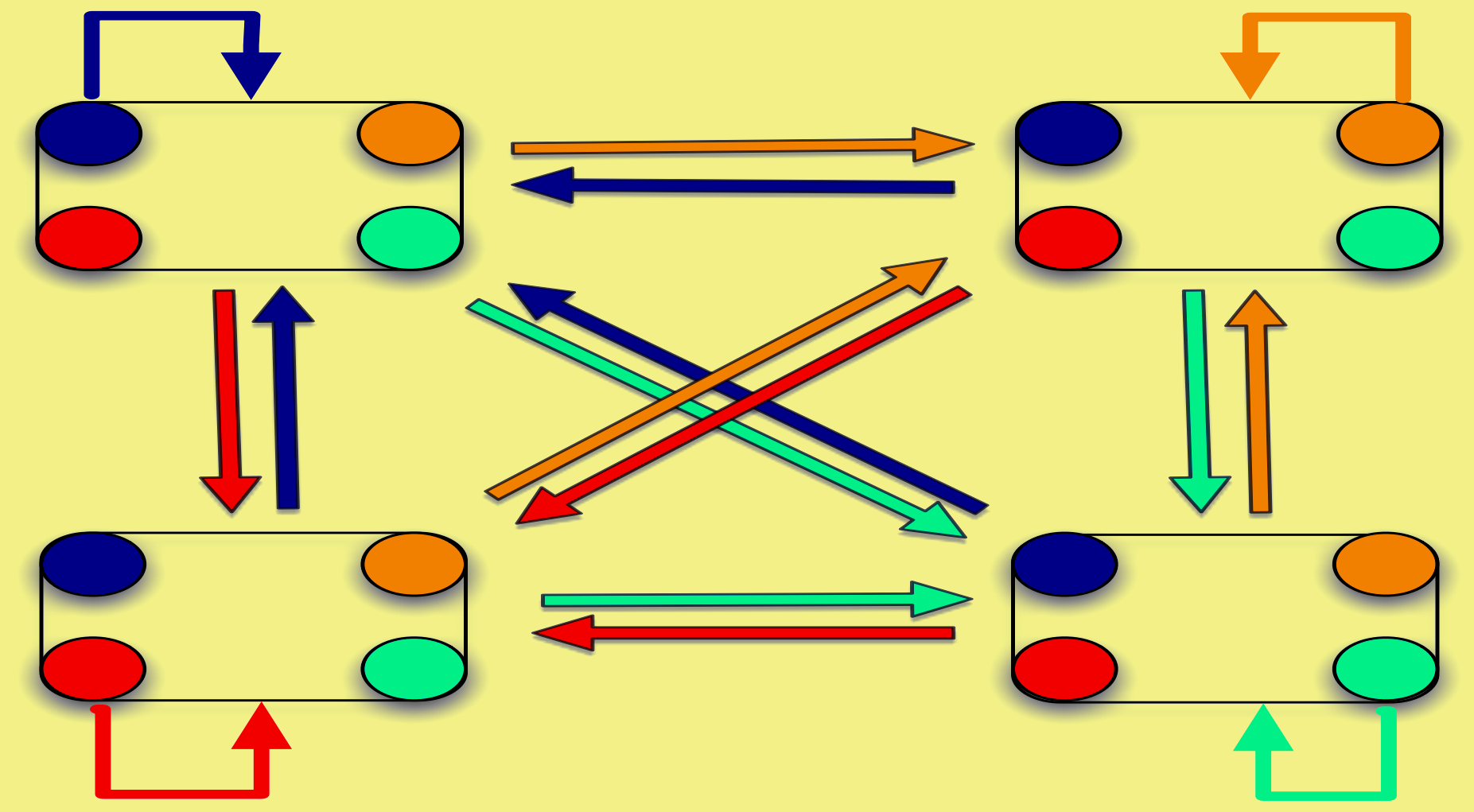


Alltoall Optimization

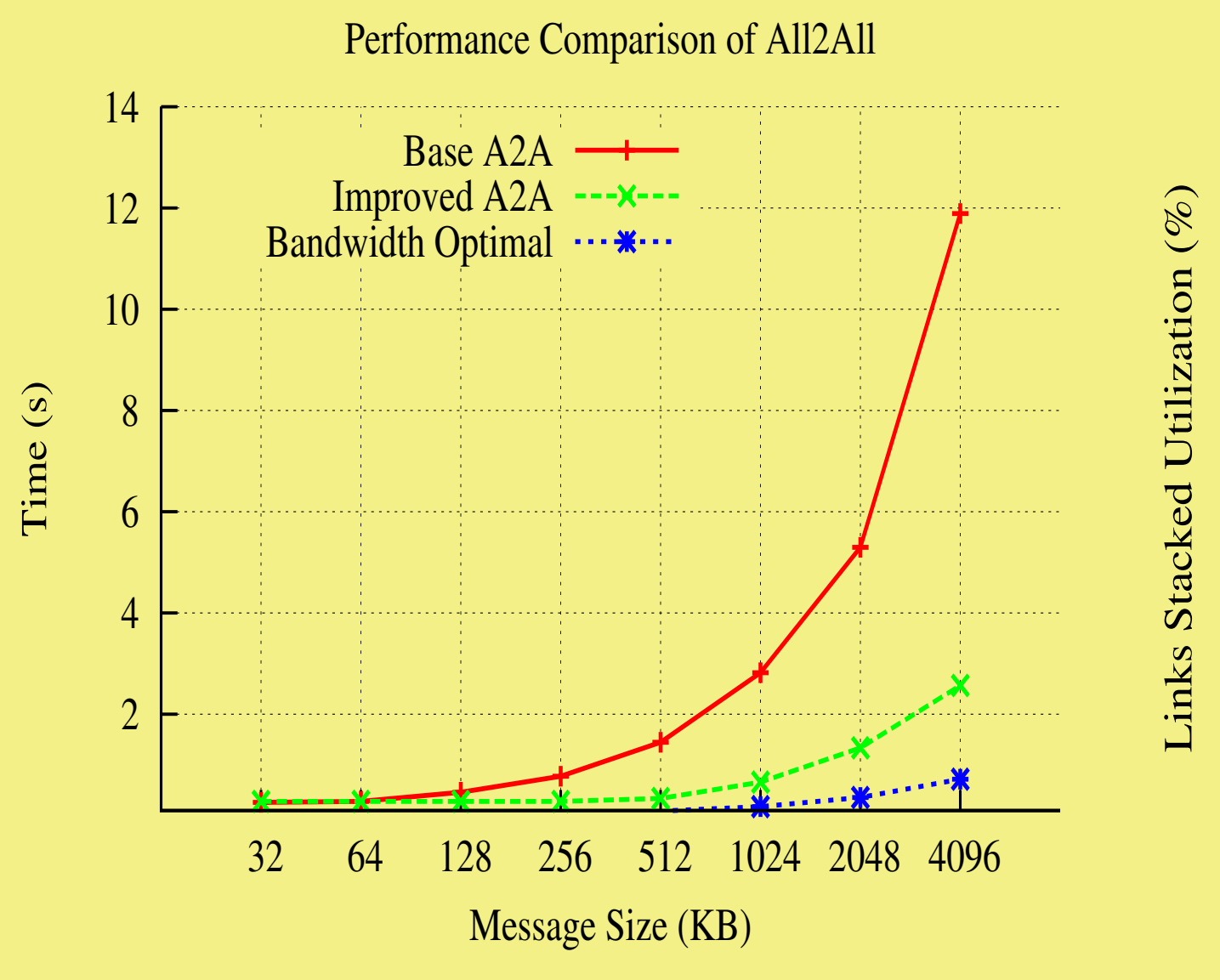
- We propose the following carefully-determined ordering to:
  (a) simultaneously utilize links stemming from a QCM
  (b) avoid the undesired congestion and link contention.

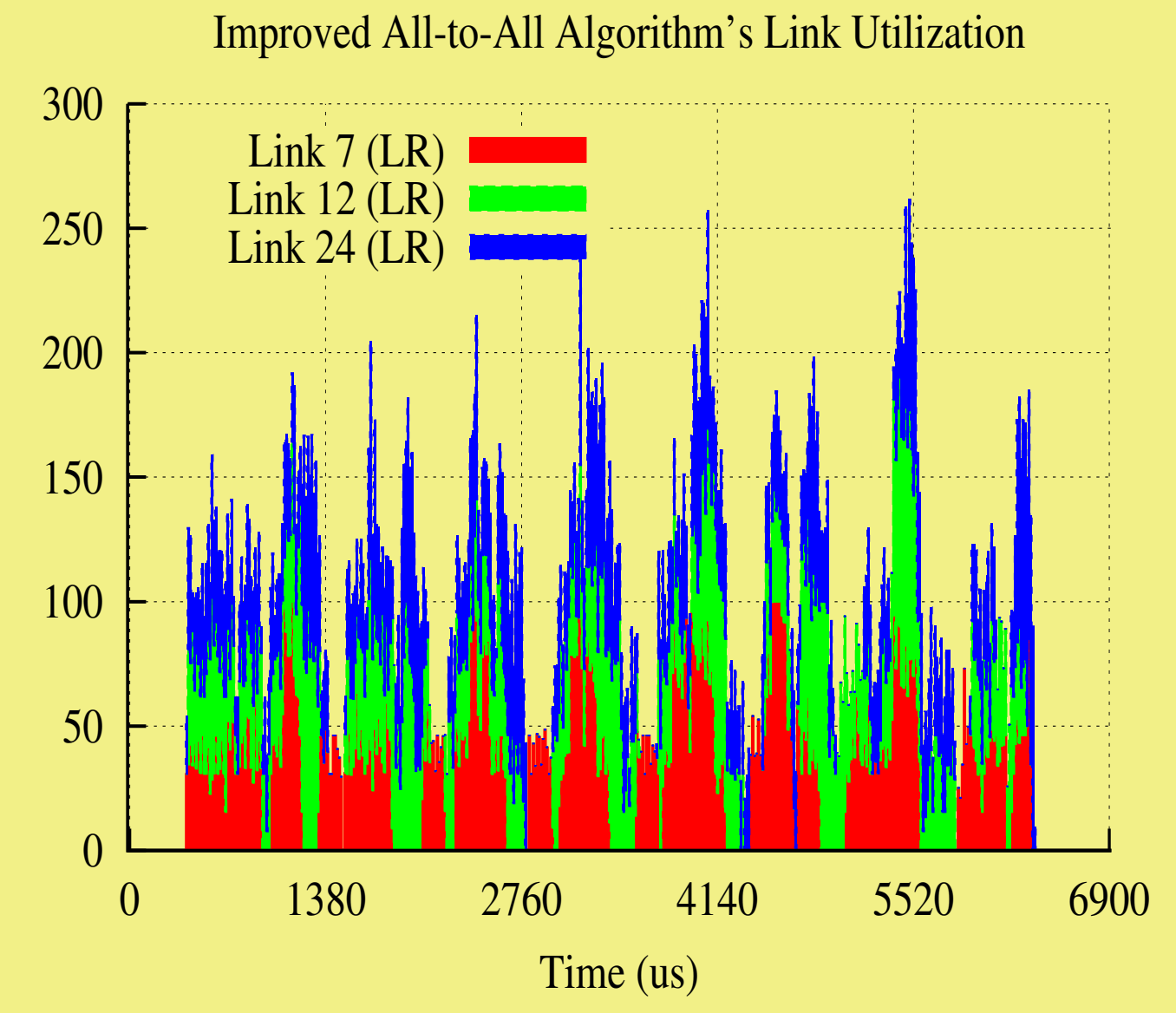
Each core sends to a different node in each phase!

- Formally, assume a node-level all-to-all network with *n* nodes, each containing *c* cores:
  1) Consider a list of *t* = *n***c* tasks running on *n* nodes with *c* cores each.
  2) Each task has to send t − 1 messages, of which sets of c destination cores lie on a given node. Any core can reach a particular set of c cores by using the direct link between the destination node and its home node.
  3) In phase i (0≤i≤n−1), core j (0≤j≤c−1) on every node sends data to the set of cores residing in the ((j + i) mod n)th node.



Up to 5 times improvement for large messages!
Links are now utilized simultaneously


Performance Comparison of All2All
Base A2A
Improved A2A
Bandwidth Optimal
Time (s)
Message Size (KB)


Improved All-to-All Algorithm's Link Utilization
Link 7 (LR)
Link 12 (LR)
Link 24 (LR)
Links Stacked Utilization (%)
Time (us)