Hierarchy Aware Blocking and Nonblocking Collective Communications-The Effects of Shared Memory Communications in the Cray XT Environment

Richard L. Graham, Joshua S. Ladd, Manjunath Venkata

Acknowledgements

- US Department of Energy FASTOS program

Outline

- Statement of the problem
- Design Overview
- Results
- Next steps

Problems being addressed

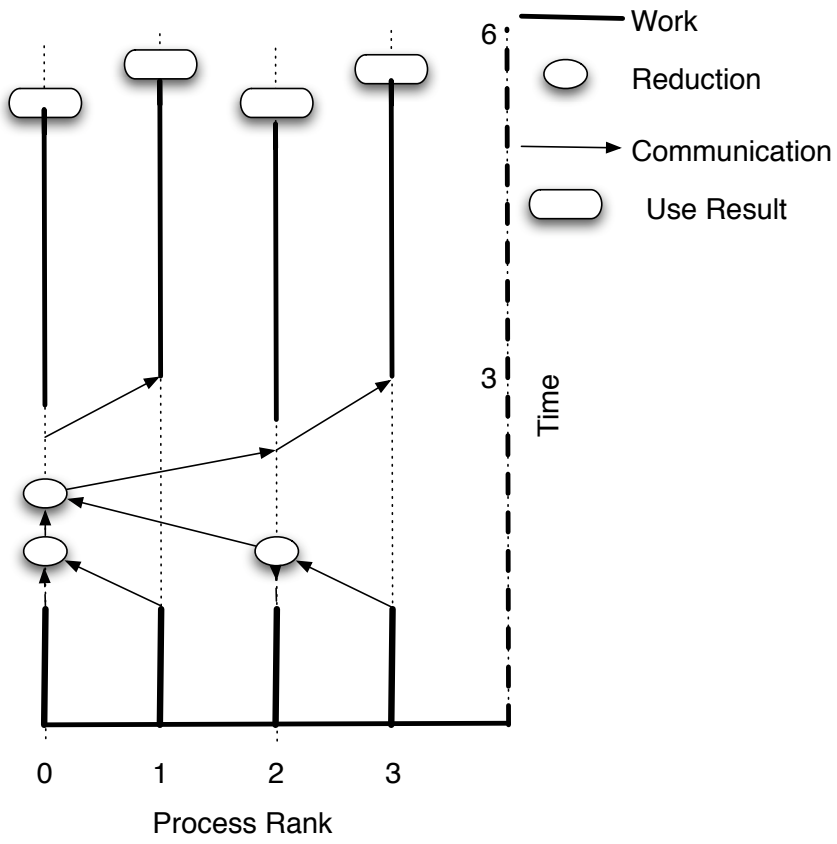
- Optimization of collective operations
- Implementation of extensible optimized collective operations
- Implementation of nonblocking collective operations

Why Optimize Collective Communications

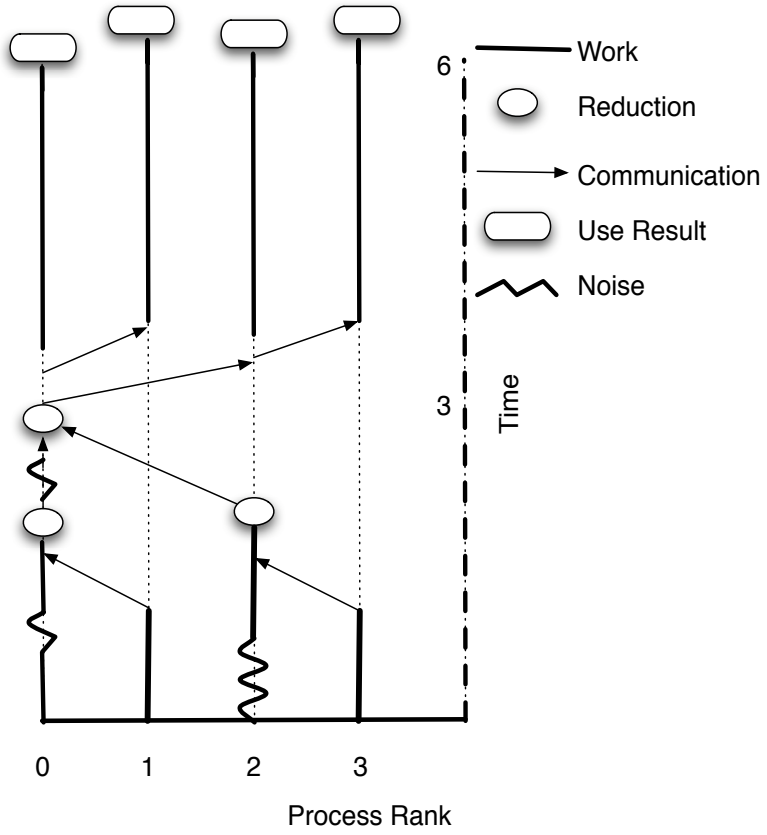
- Collective operations limit application scalability
- Communication pattern involving multiple processes (in MPI, all ranks in the communicator are involved)
- Optimized collectives involve a communicator-wide data-dependent communication pattern
- Data needs to be manipulated at intermediate stages of a collective operation
- Collective operations magnify the effects of system-noise

Scalability of Collective Operations

Ideal Algorithm

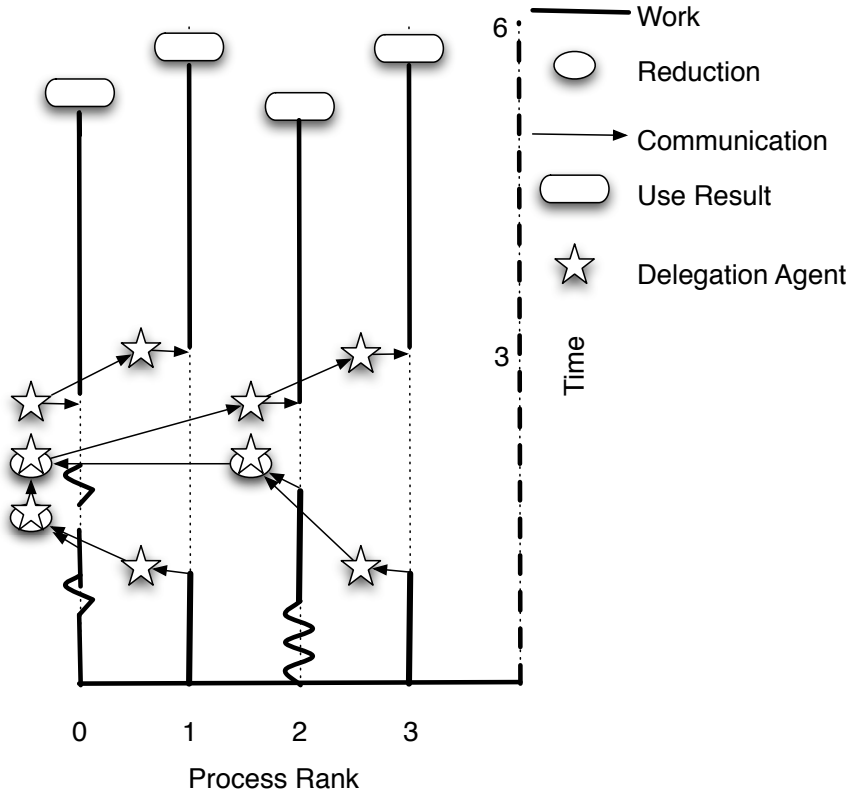


Impact of System Noise

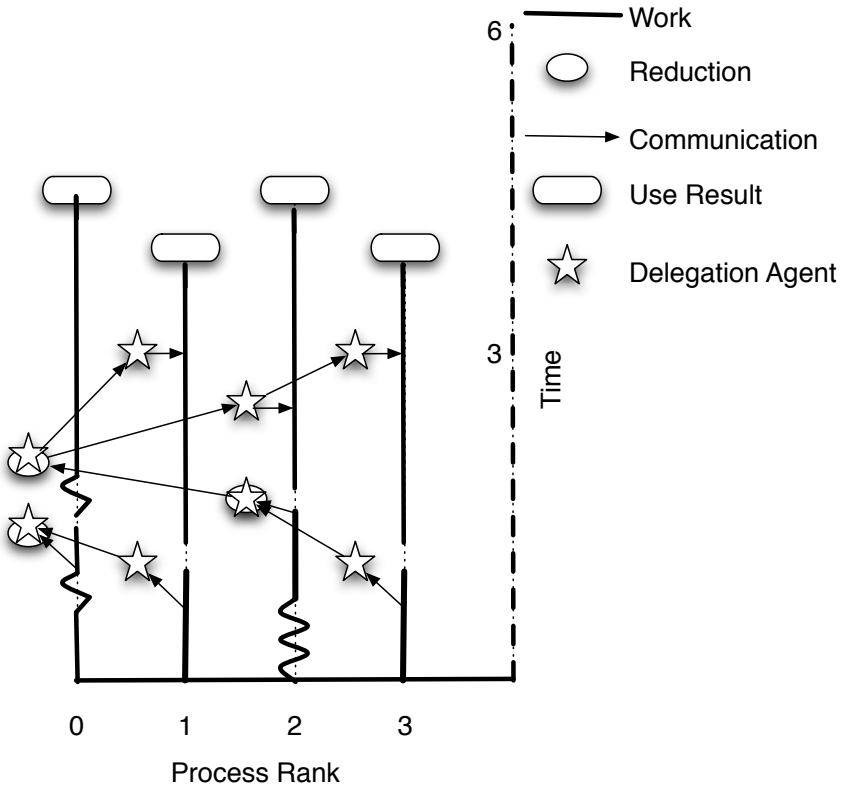


Scalability of Collective Operations - II

Offloaded Algorithm



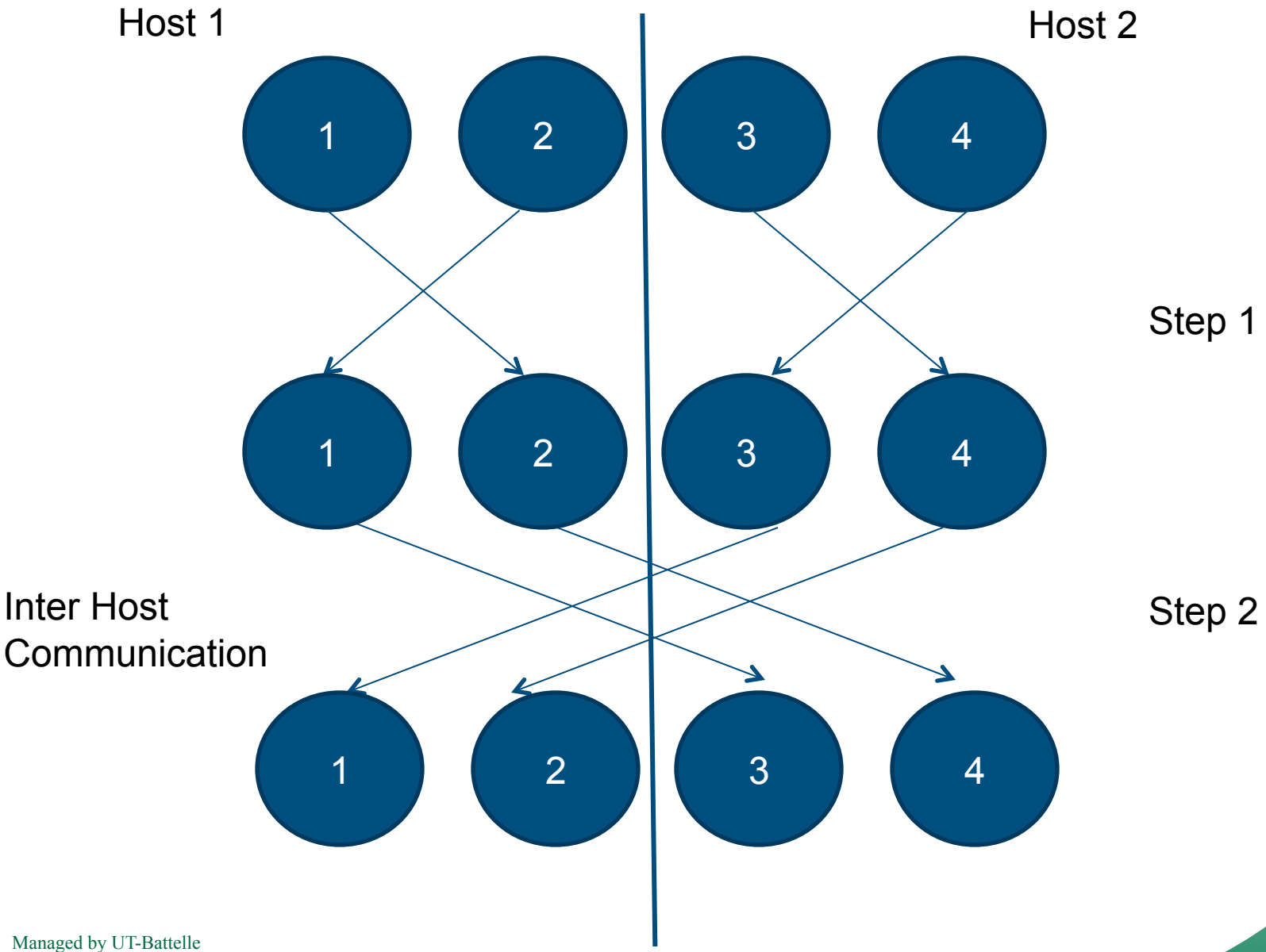
Nonblocking Algorithm



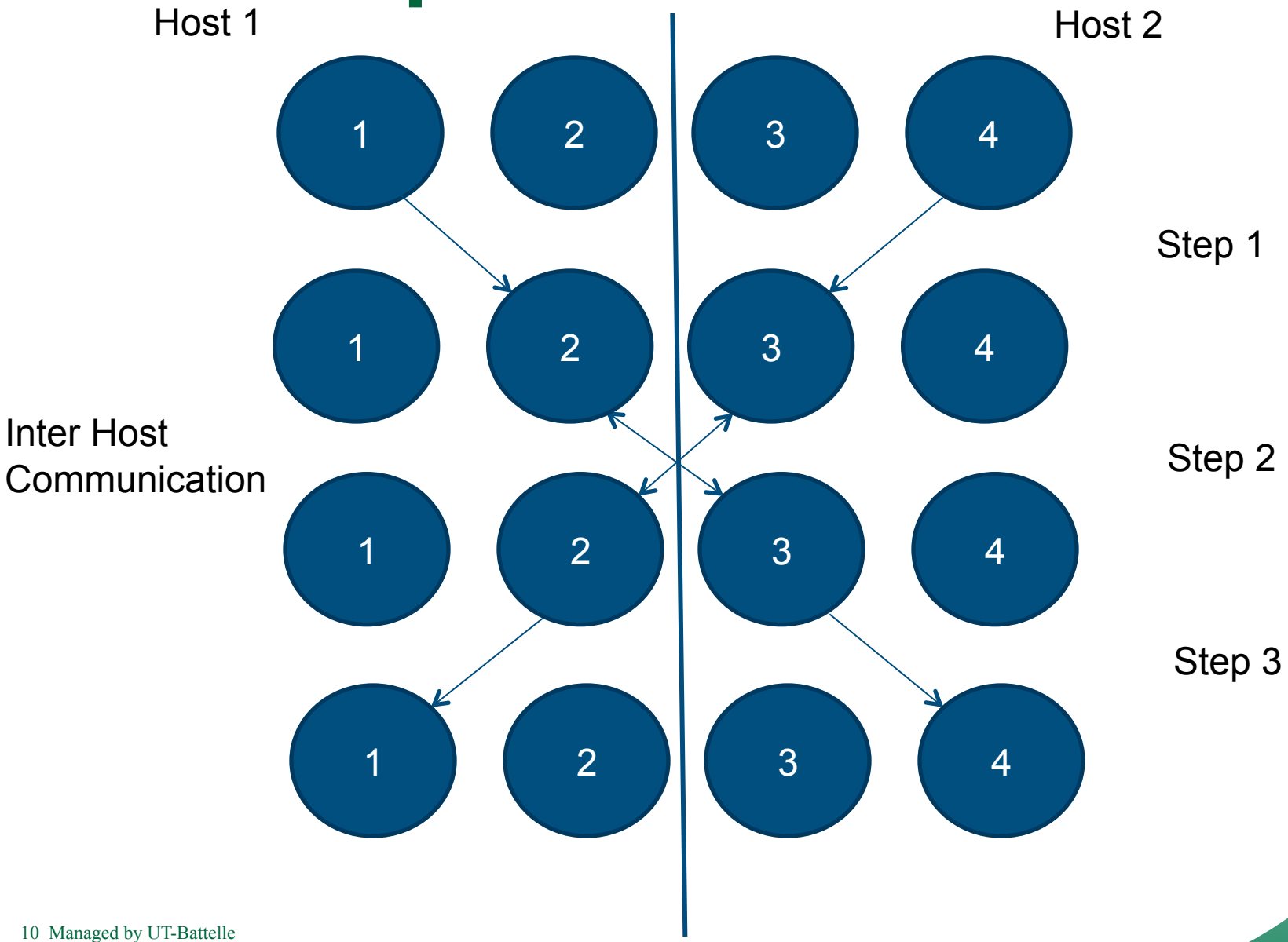
Mapping the collectives onto the system

- Consider communication hierarchies
- Schedule the network

Example – 4 Process Recursive Doubling



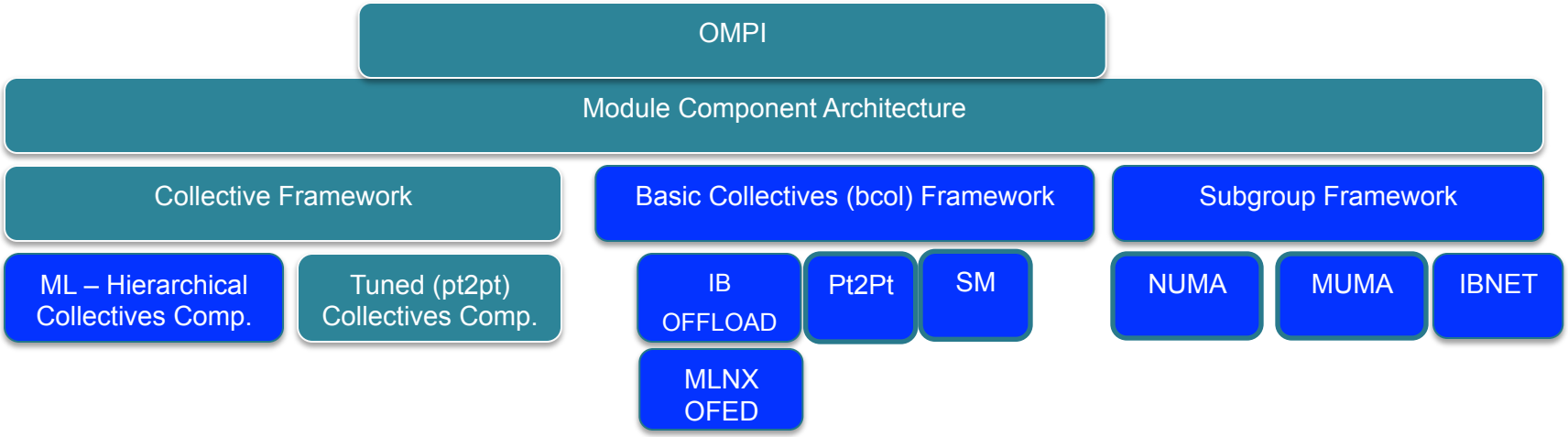
Example – 4 Process Recursive Doubling – On host optimization



Design strategy

- Decouple
  - Hierarchy detection
  - Network specific collective algorithm implementation (“single” level)
  - Full collective function implementation (hierarchical)
  - Basic building blocks from MPI level functions
- Share resources between levels w/o breaking the abstraction between layers

Collectives – Software Layers

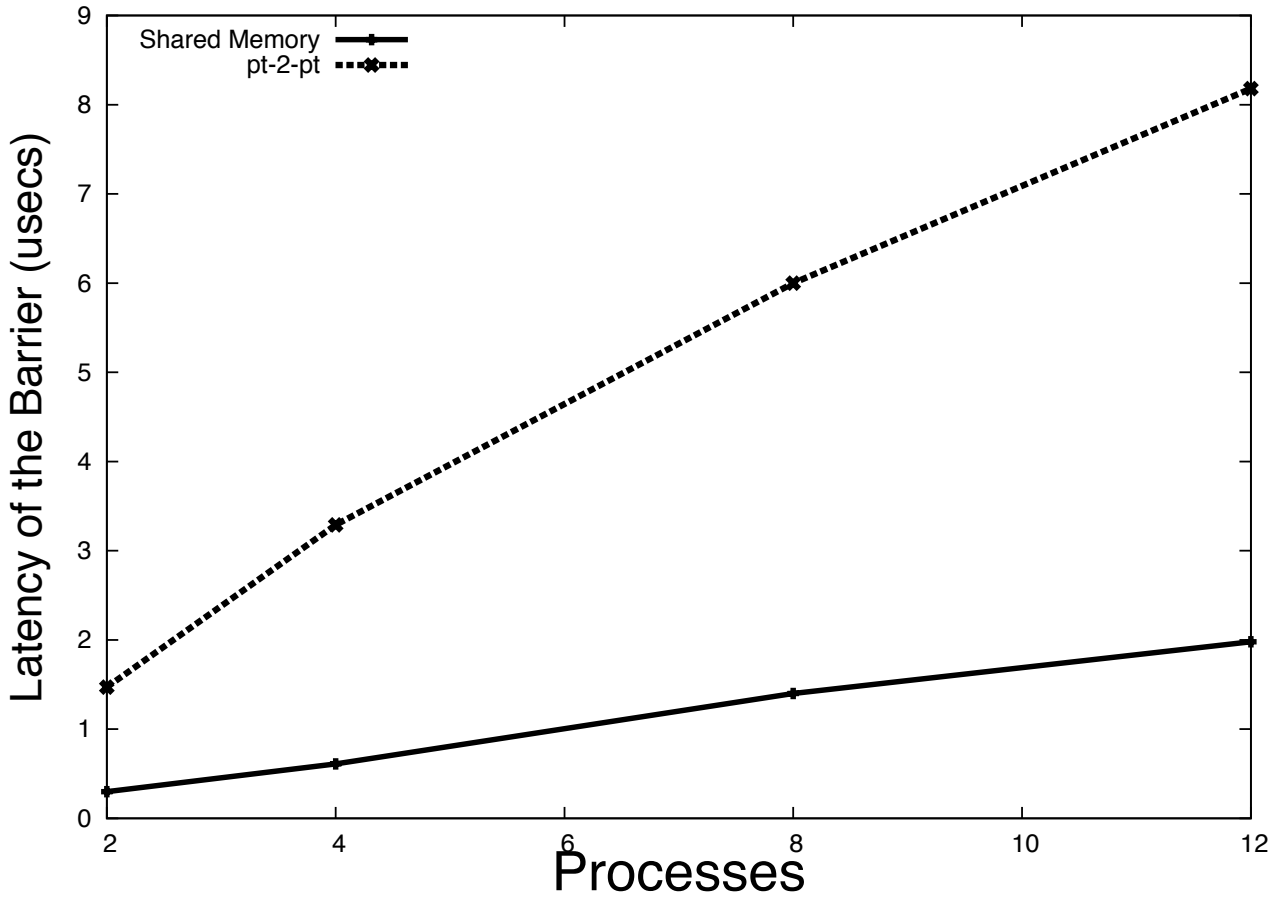


Benchmarks

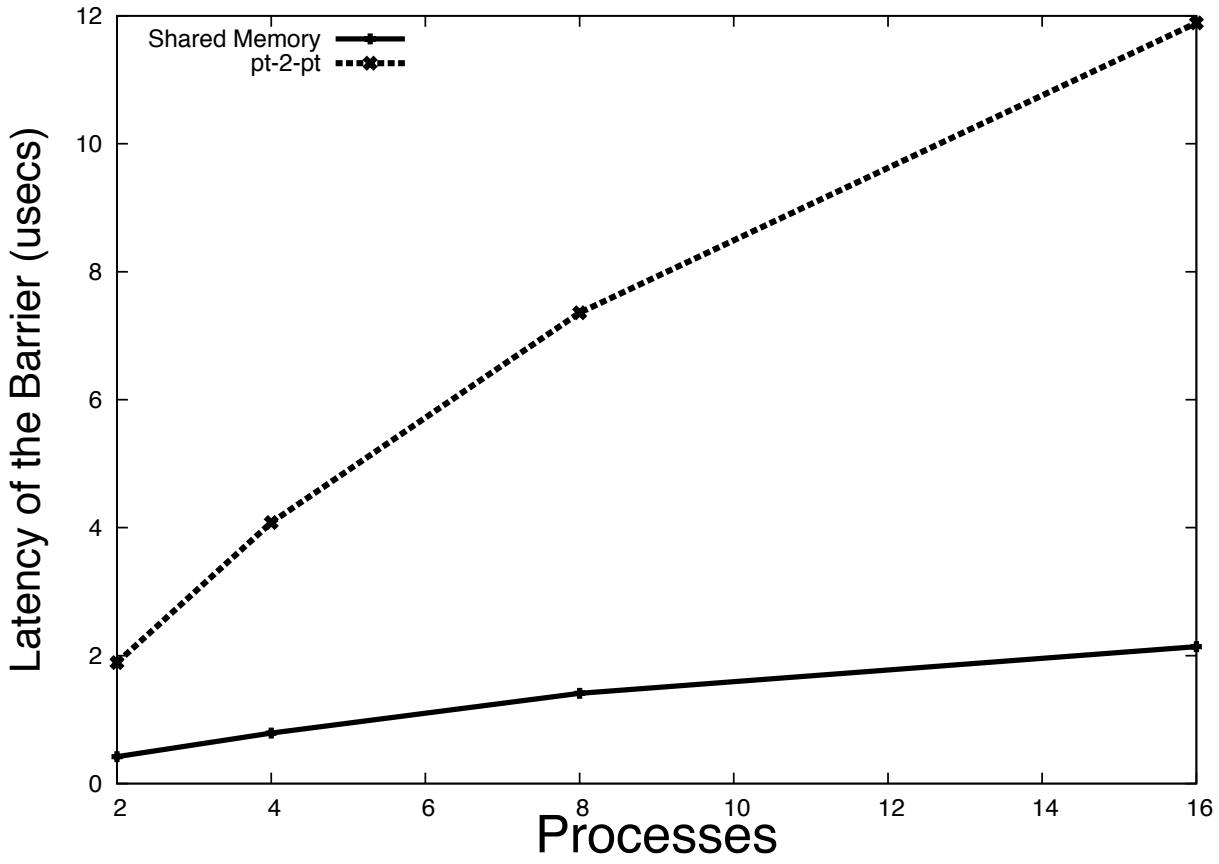
System setup

- Jaguar
  - 2.6 GHz Istanbul processor
  - Dual socket
  - Hex-core
- Smoky
  - 2.0 GHz Opteron
  - Quad socket
  - Quad core

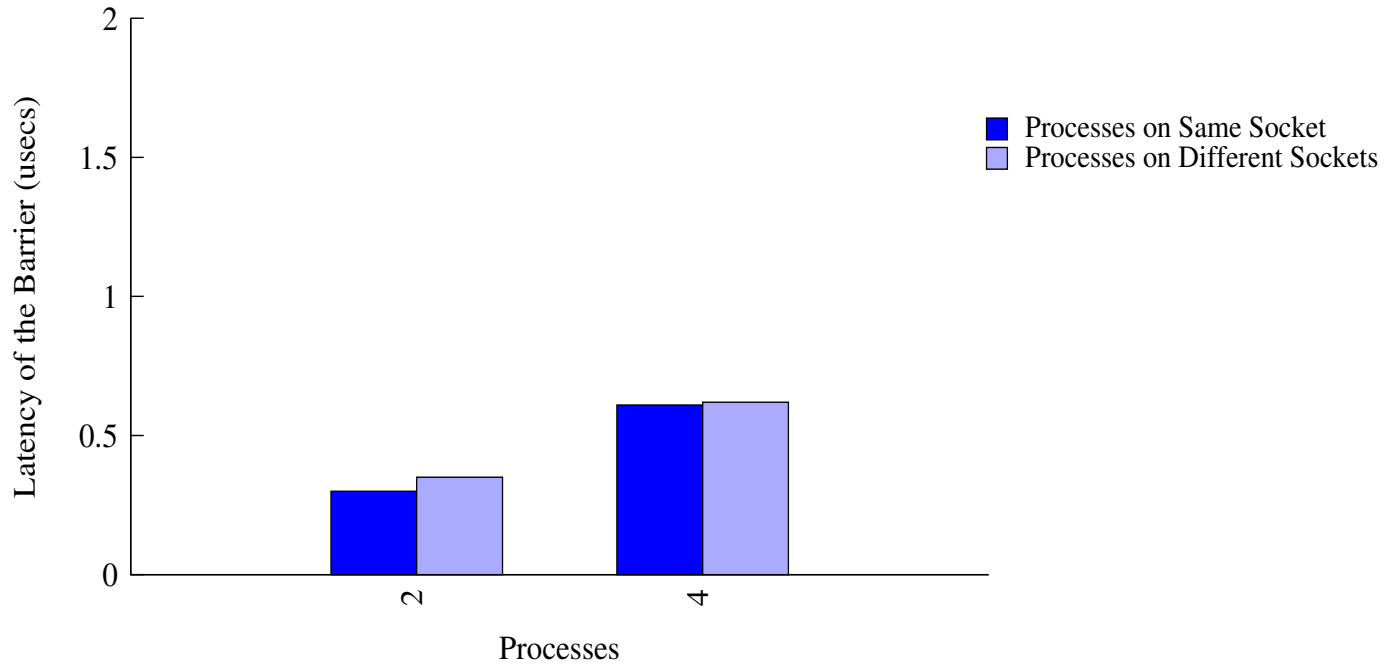
Barrier as a function of Process count – Jaguar – 2 Level hierarchy



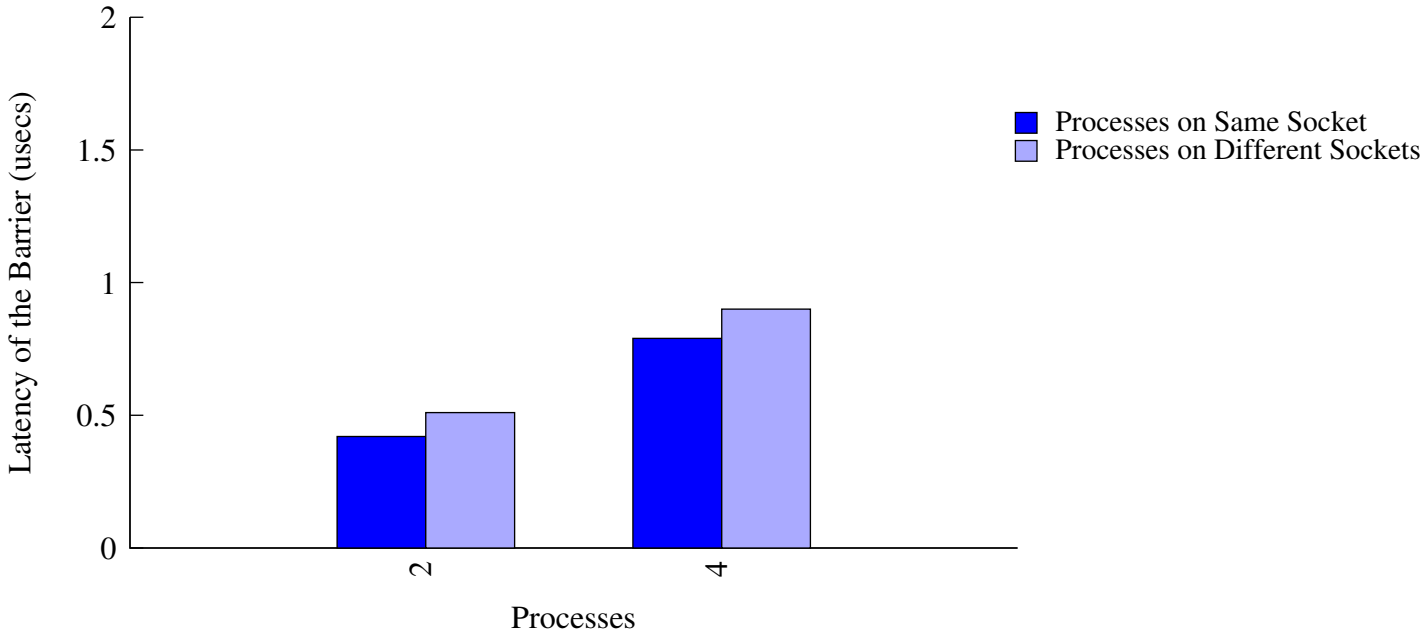
Barrier as a function of Process count – Smoky – 2 Level hierarchy



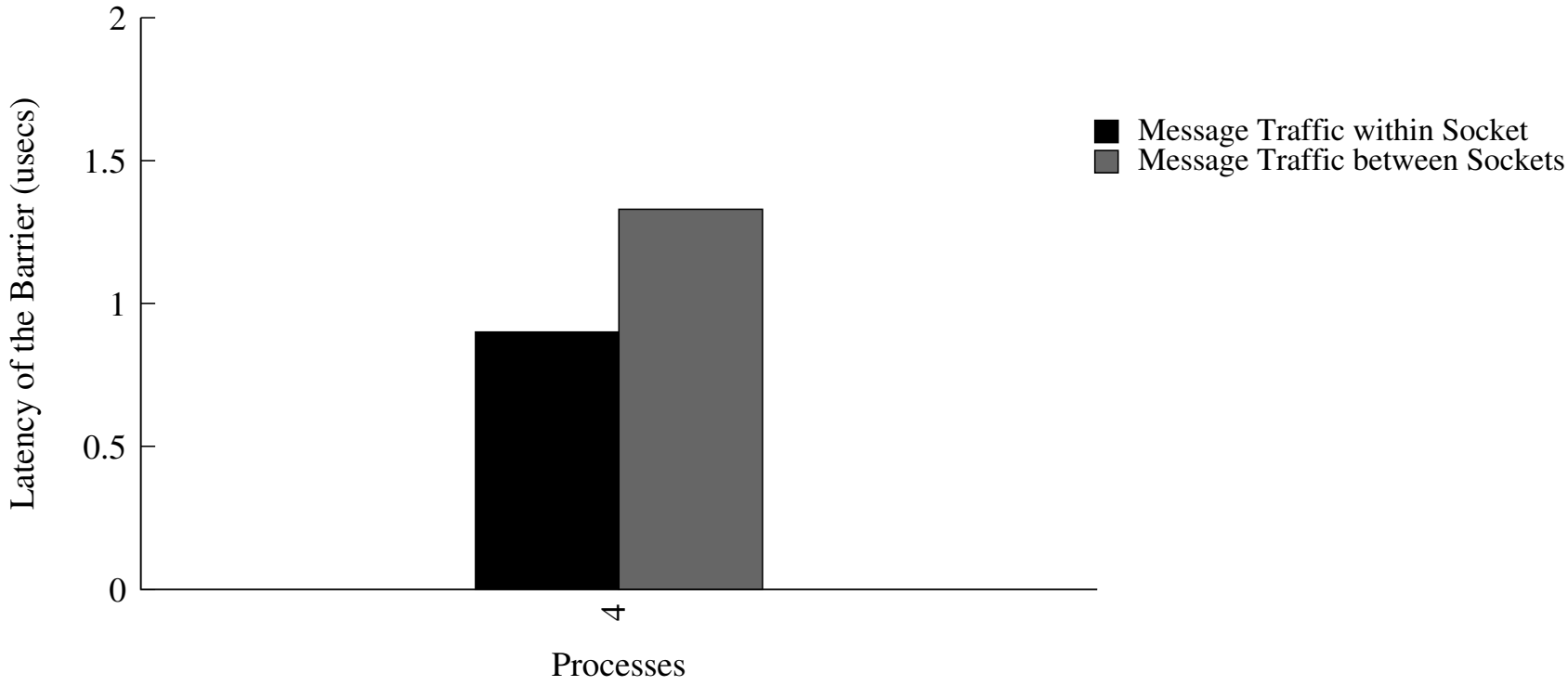
Barrier As a function of number of sockets - Jaguar



Barrier As a function of number of sockets (1,2) – Smoky



Barrier As a function of number of sockets (1,4) – Smoky



Summary

- Added hardware support for offloading collective operations
- Developed MPI-level support for asynchronous collectives
- Good barrier performance
- Good overlap capabilities
- Work is continuing