An Accelerated Implementation of Portals on the Cray SeaStar

Ron Brightwell **Kevin Pedretti** **Keith Underwood**

Sandia National Laboratories

Center for Computation, Computers, Information, and Mathematics

Trammell Hudson

OS Research

Cray Users' Group Annual Technical Conference

May 11, 2006

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

Sandia National Laboratories

Outline

- SeaStar
- Portals
- CORPSE
- Performance
- Future Work

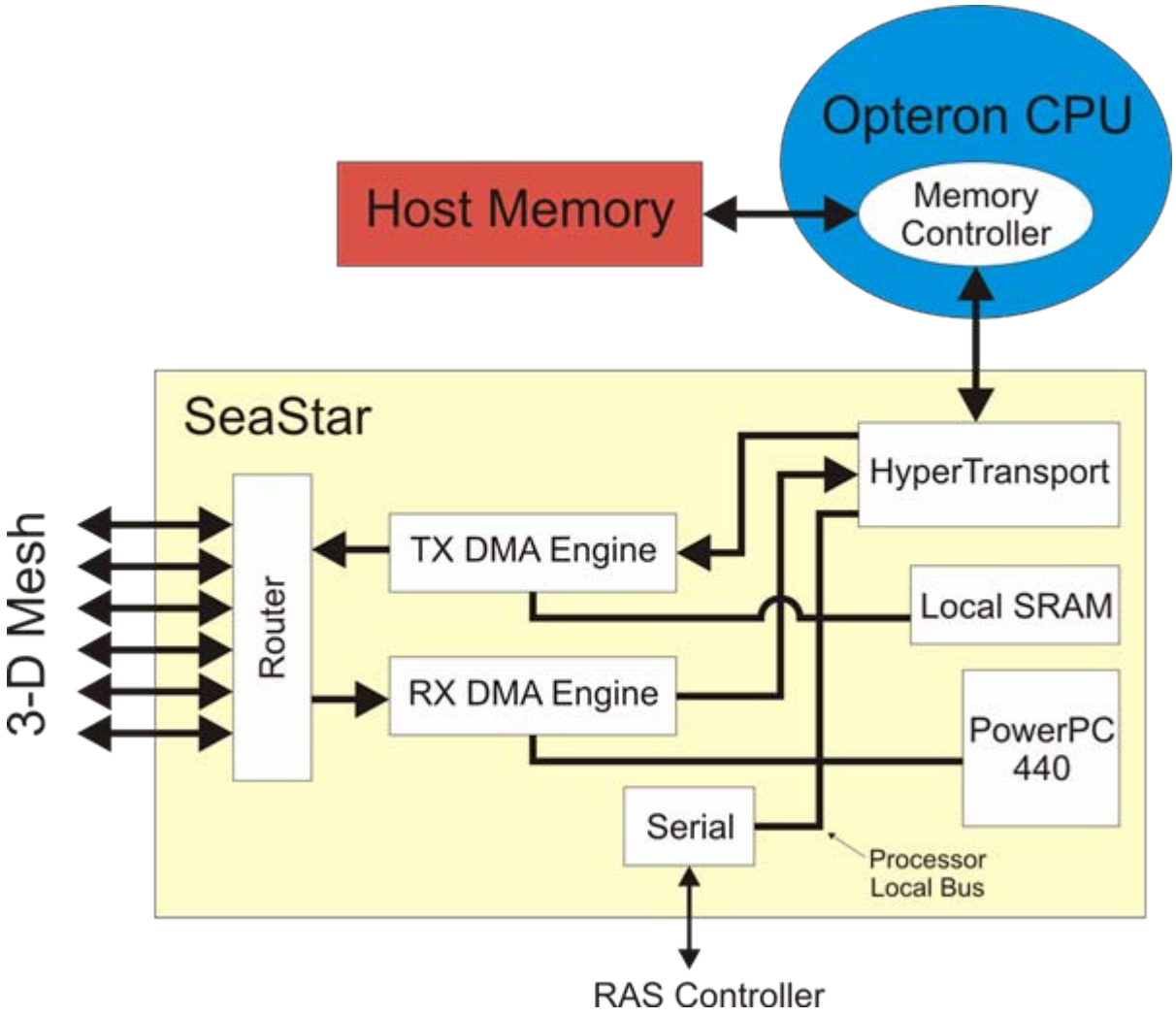
Sandia National Laboratories

Cray SeaStar NIC/Router

- 16 1.6 Gb/s HyperTransport to Opteron
- 500 MHz embedded PowerPC 440
- 384 KB on-board scratch RAM
- Seven-port router
- Six 12-channel 3.2 Gb/s high-speed serial links

Sandia National Laboratories

SeaStar Block Diagram



Portals 3.3 for SeaStar

- Cray started with Sandia reference implementation
- Needed single version of NIC firmware that supports all combinations of
  - User-level and kernel-level API
  - NIC-space and kernel-space library
- Cray added bridge layer to reference implementation to allow NAL to interface multiple API NALs and multiple library NALs
  - qkbridge for Catamount applications
  - ukbridge for Linux user-level applications
  - kbridge for Linux kernel-level applications

Sandia National Laboratories

SeaStar NAL

- Portals processing in kernel-space
  - Interrupt-driven
  - “generic” mode
- Portals processing in NIC-space
  - No interrupts
  - “accelerated” mode

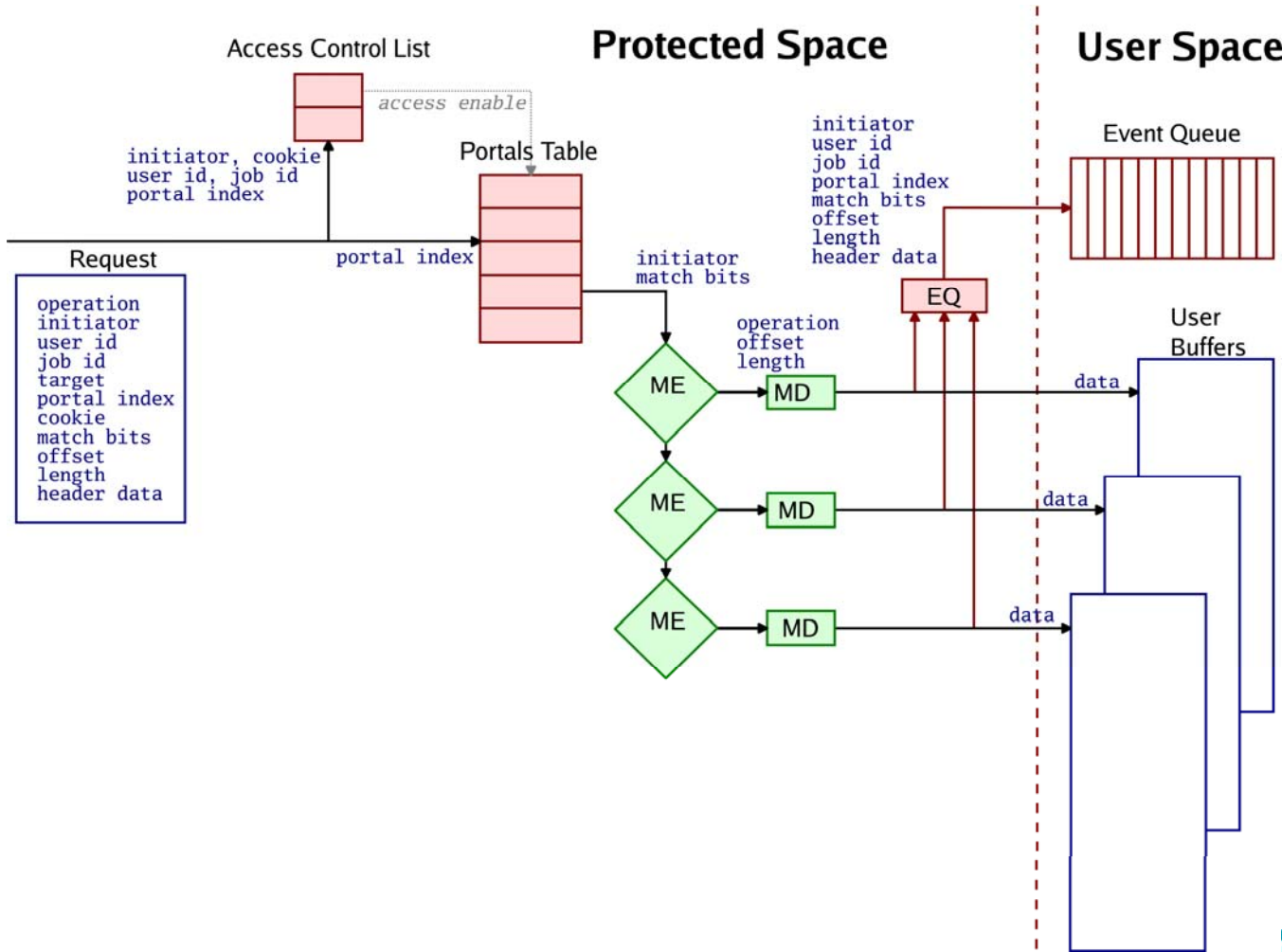
Sandia National Laboratories

Prototype NIC-based Network Stack

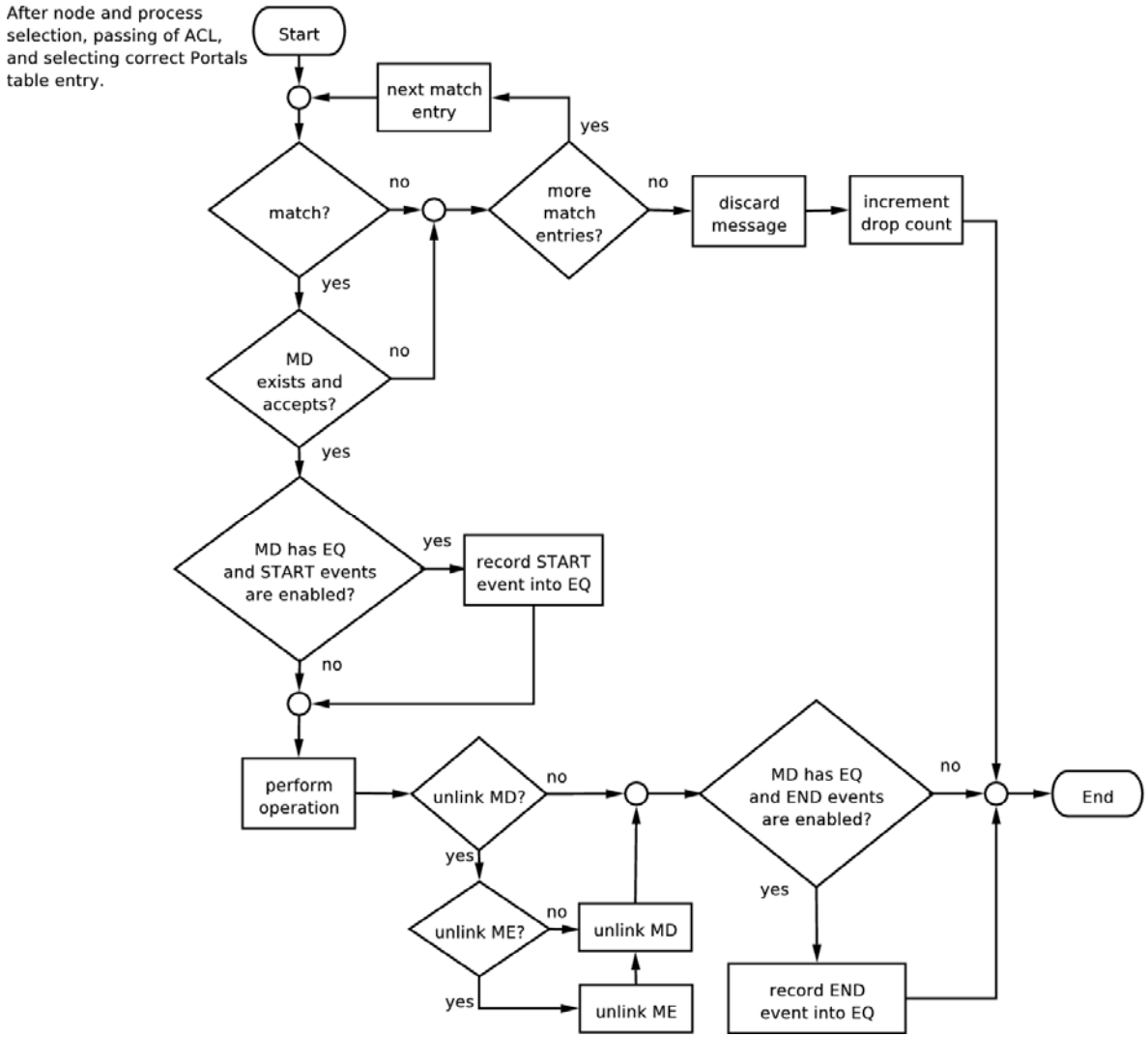
- Allowed characterization of
  - Impact of interrupts on latency
  - Impact on throughput (Messages/second)
  - NIC vs. host CPU matching speed
  - Penalty of having multiple NIC mailboxes
  - NIC resource requirements of each CPU core

Sandia National Laboratories

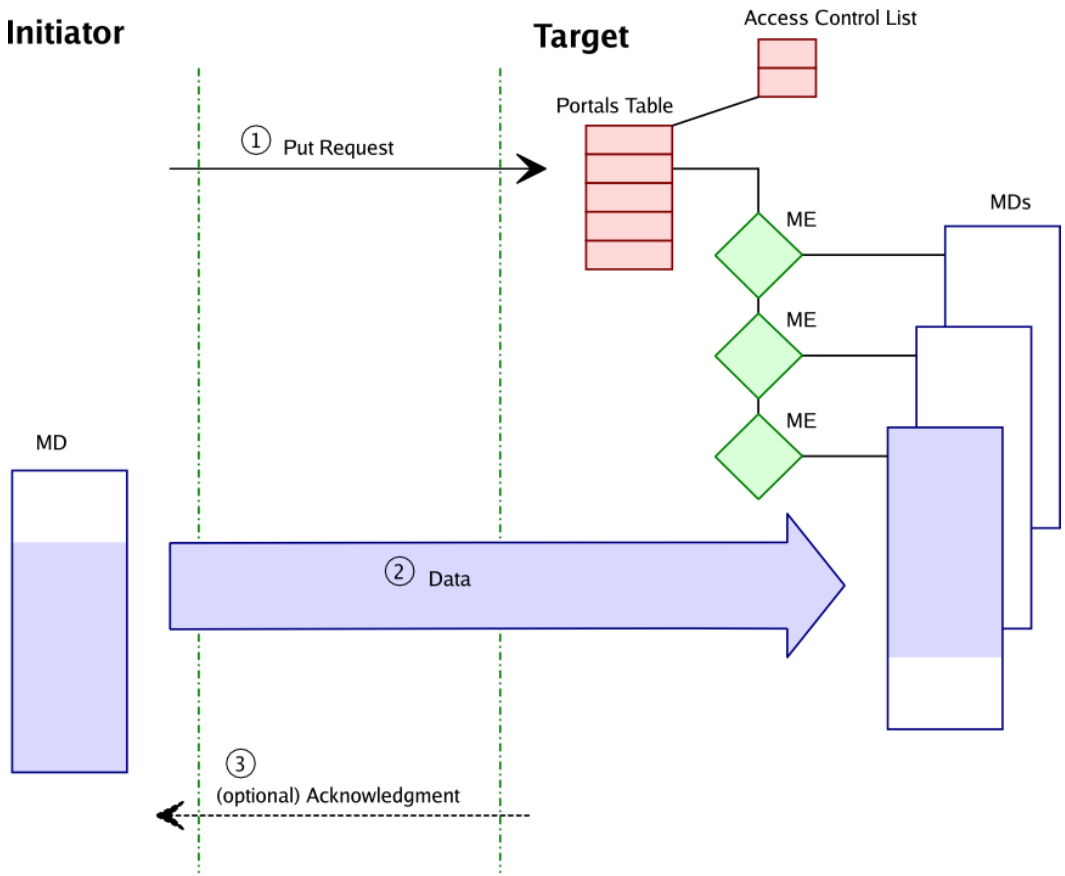
Addressing



Steps for Address Translation

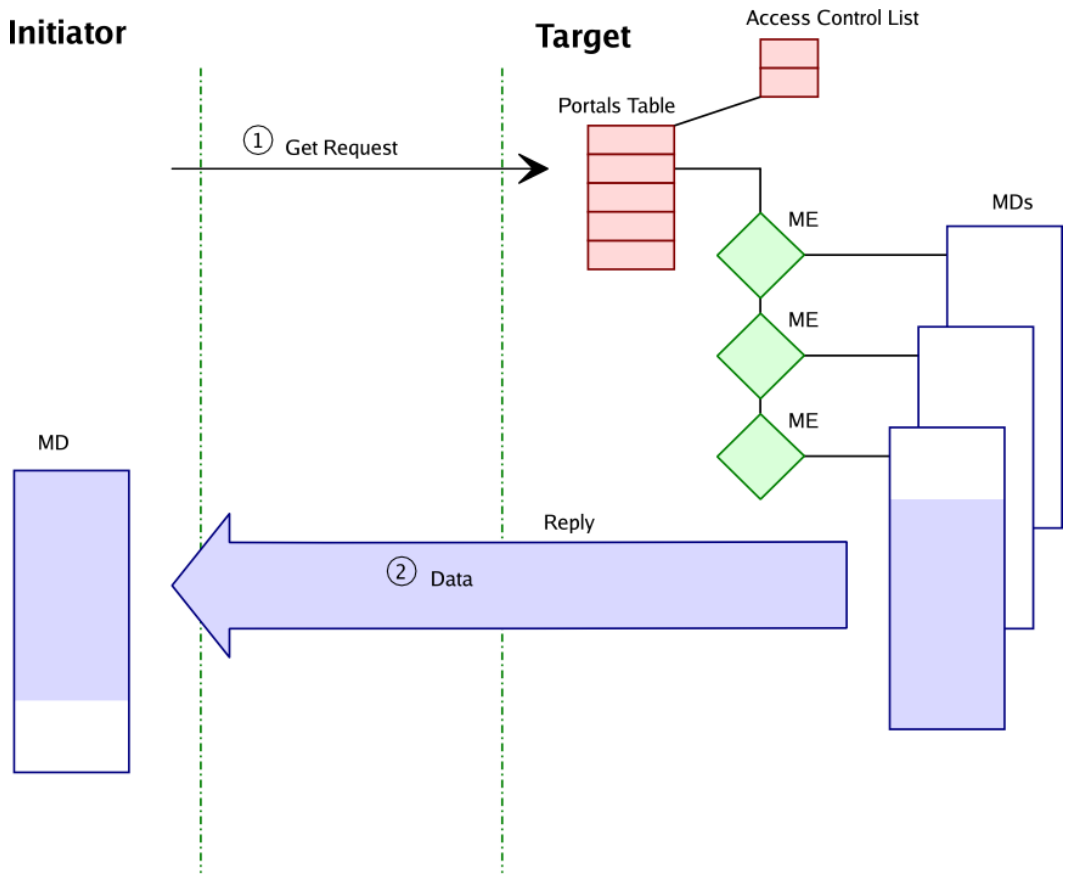


Put Operation

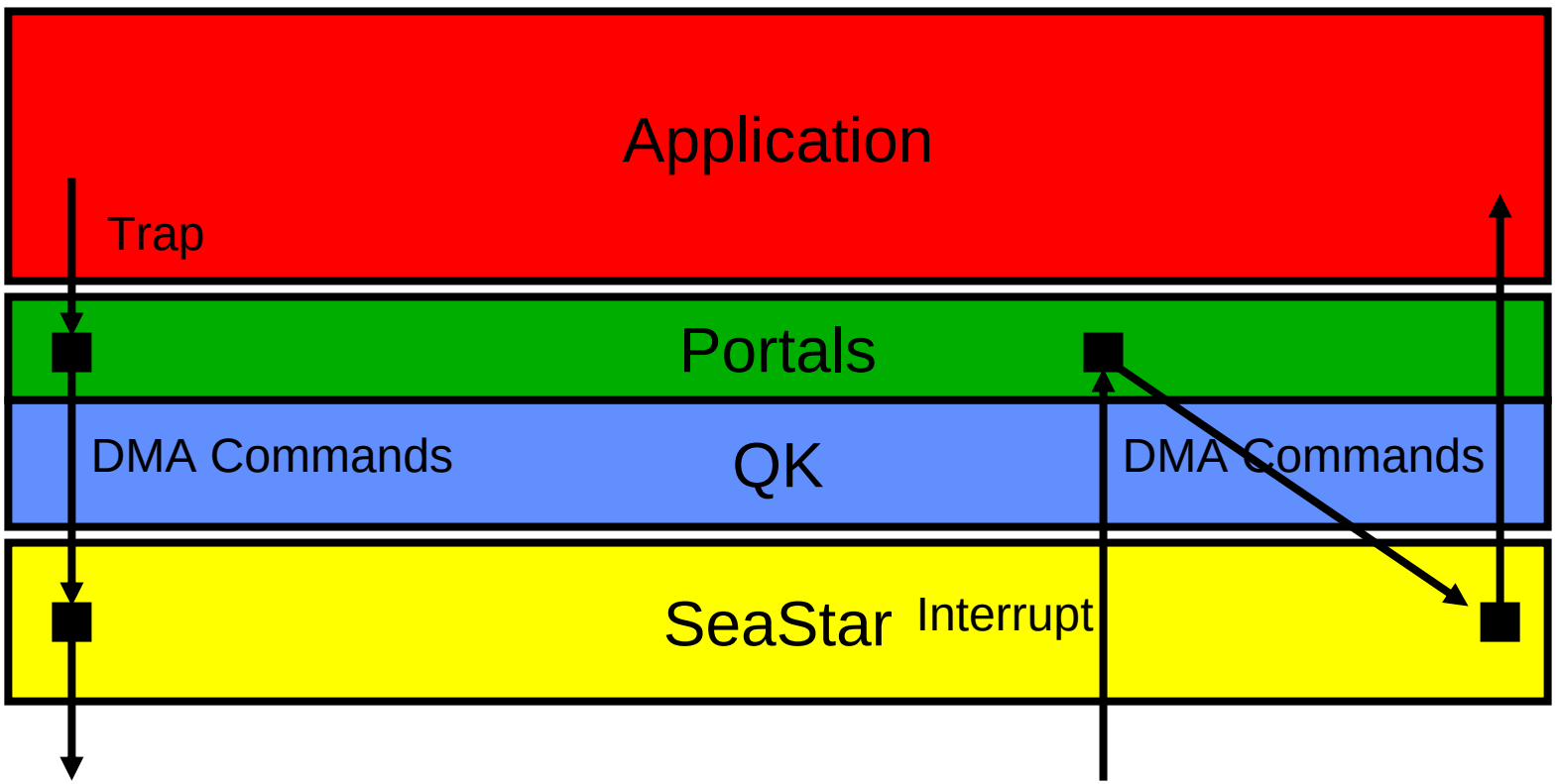


Sandia National Laboratories

Get Operation

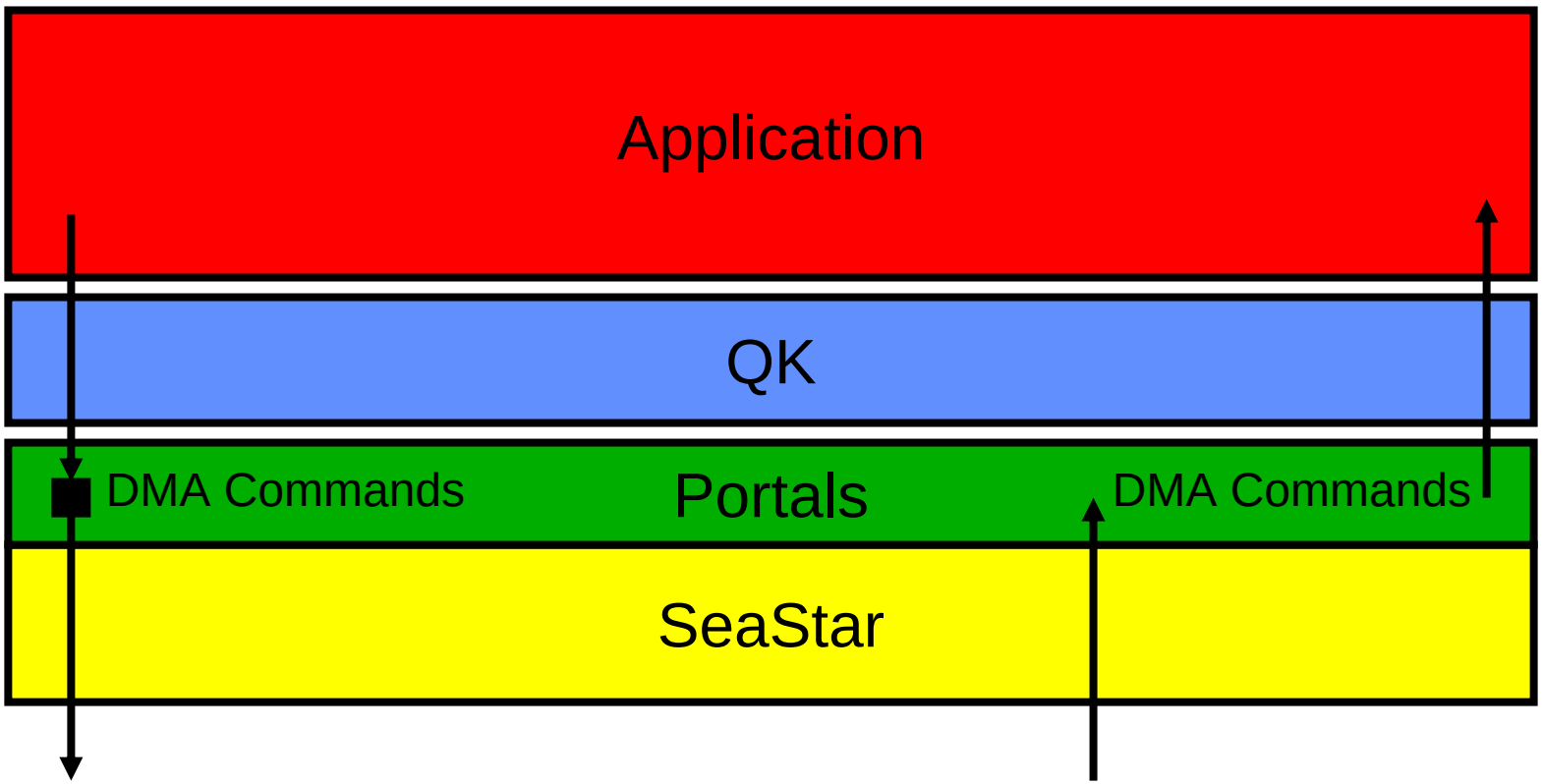


Host-Based Portals Implementation



Sandia National Laboratories

NIC-Based Portals Implementation



Sandia National Laboratories

Portals Application Mailbox

- Untrusted mailbox between application and firmware
- Initialization is via kernel mailbox
  - Maps the processes address space
  - Physically contiguous so whole mapping is done
  - Only works with Catamount
- Application mailbox
  - All other Portals command are delivered directly to the SeaStar
  - Trusted header
  - Sending from SeaStar memory is prohibitive

Sandia National Laboratories

Flow Control

- CAM Overflow Remediation Protocol SystEm (CORPSE)
  - Sandia's protocol that runs entirely on the SeaStar
- CAM Overflow Protocol
  - Cray's protocol that runs entirely on the Opteron

Sandia National Laboratories

Changes to Portals/MPI for Accelerated

- MPI receive posting was slow
  - Posting a receive involves a round trip to the SeaStar (1 us) relative a kernel trap (65 ns)
  - Combine PtlMEAttach(), PtlMDAttach(), PtlMDUpdate int PtlMEMDPost()
- Reduce HT transfers on the send side
  - Move send-side MD creation out of fast path
  - Create three MDs that cover all of data, stack, and heap
  - Improves message rate and fixes the amount of resources on the send-side

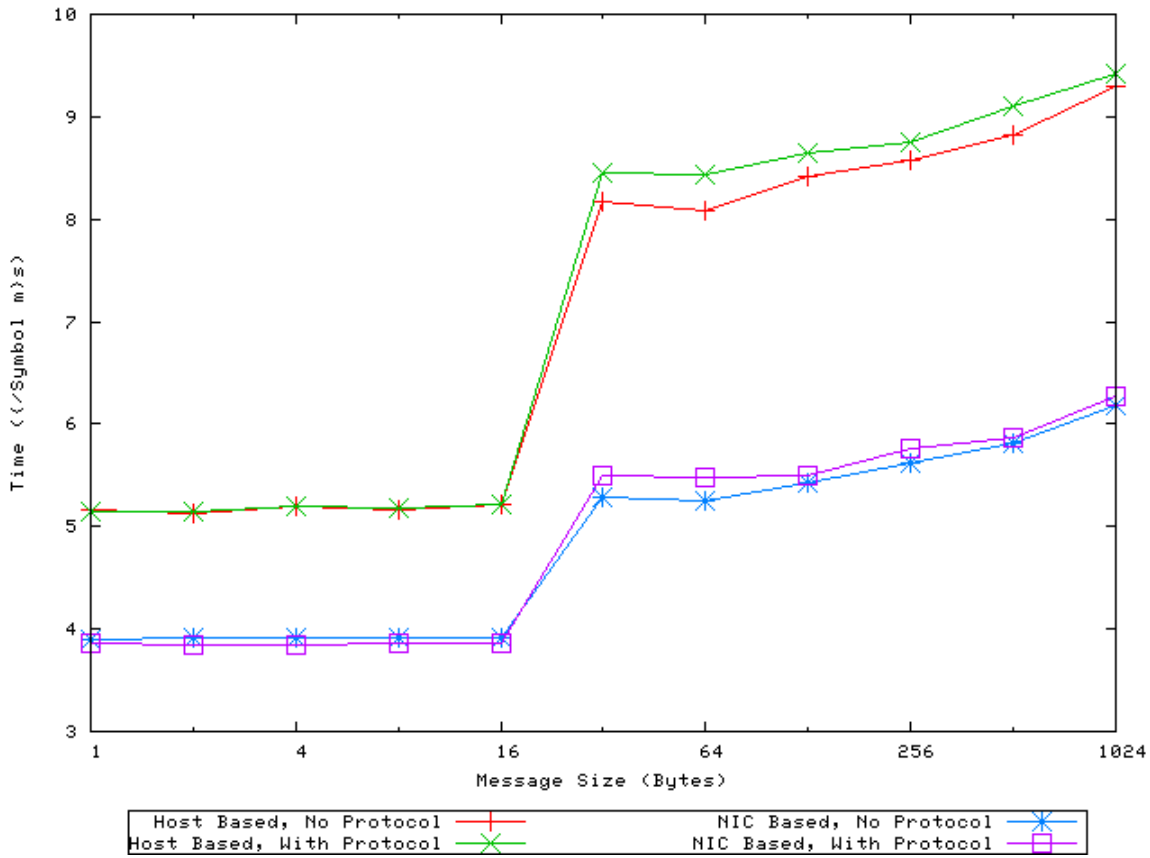
Sandia National Laboratories

Benefits of Accelerated

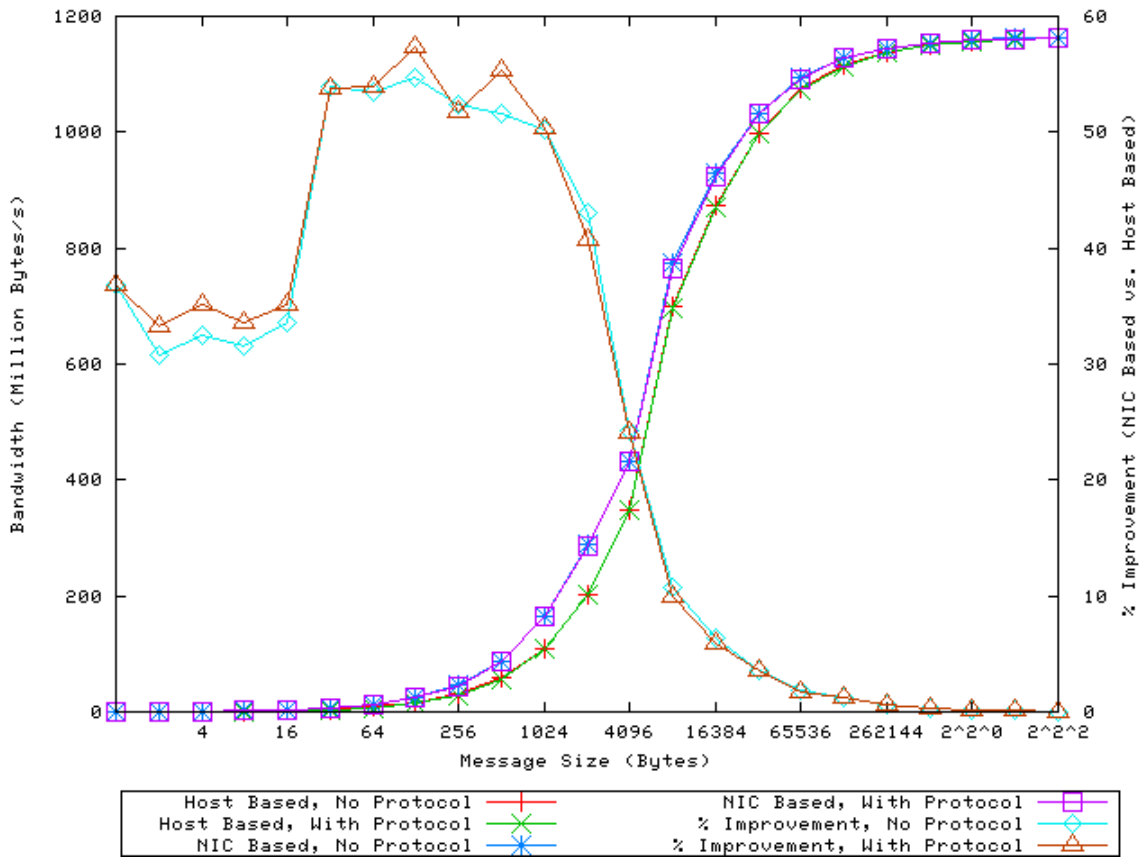
- OS does not run on message arrival
- No context switch overhead
- Portals address translation done by SeaStar

Sandia National Laboratories

Latency

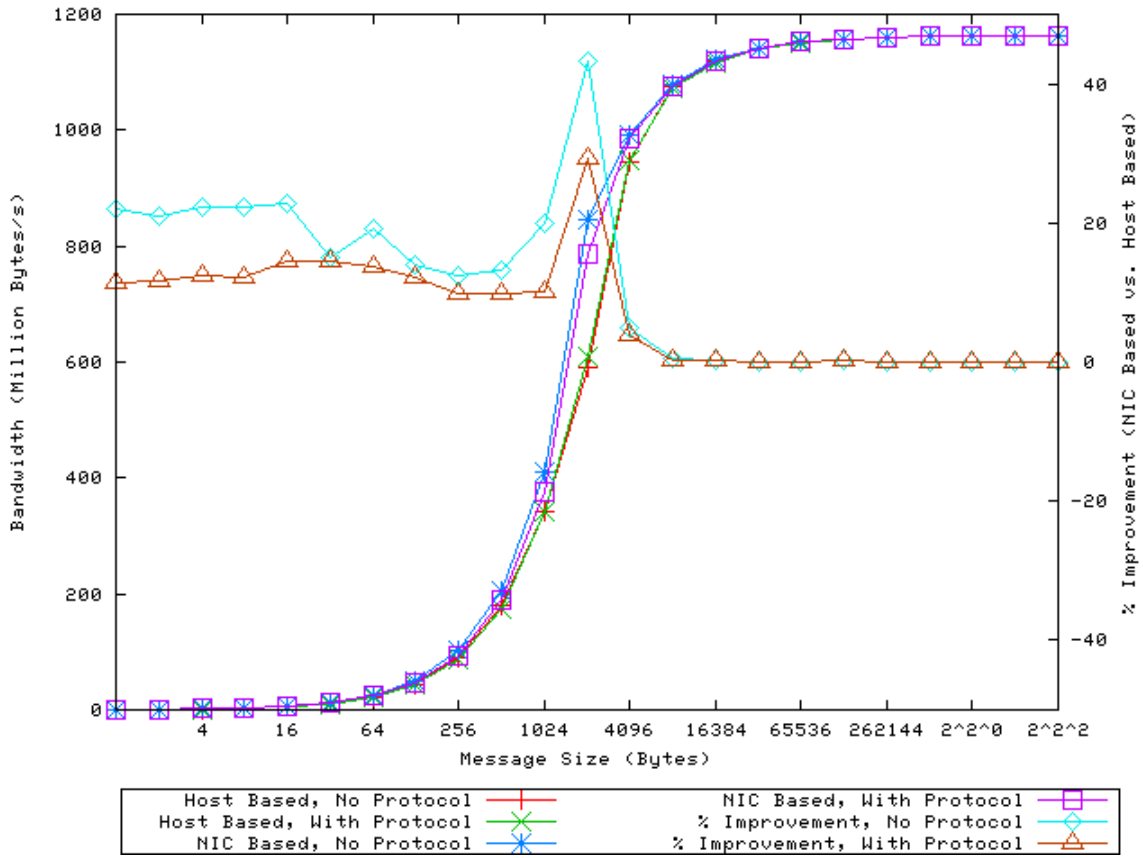


Bandwidth



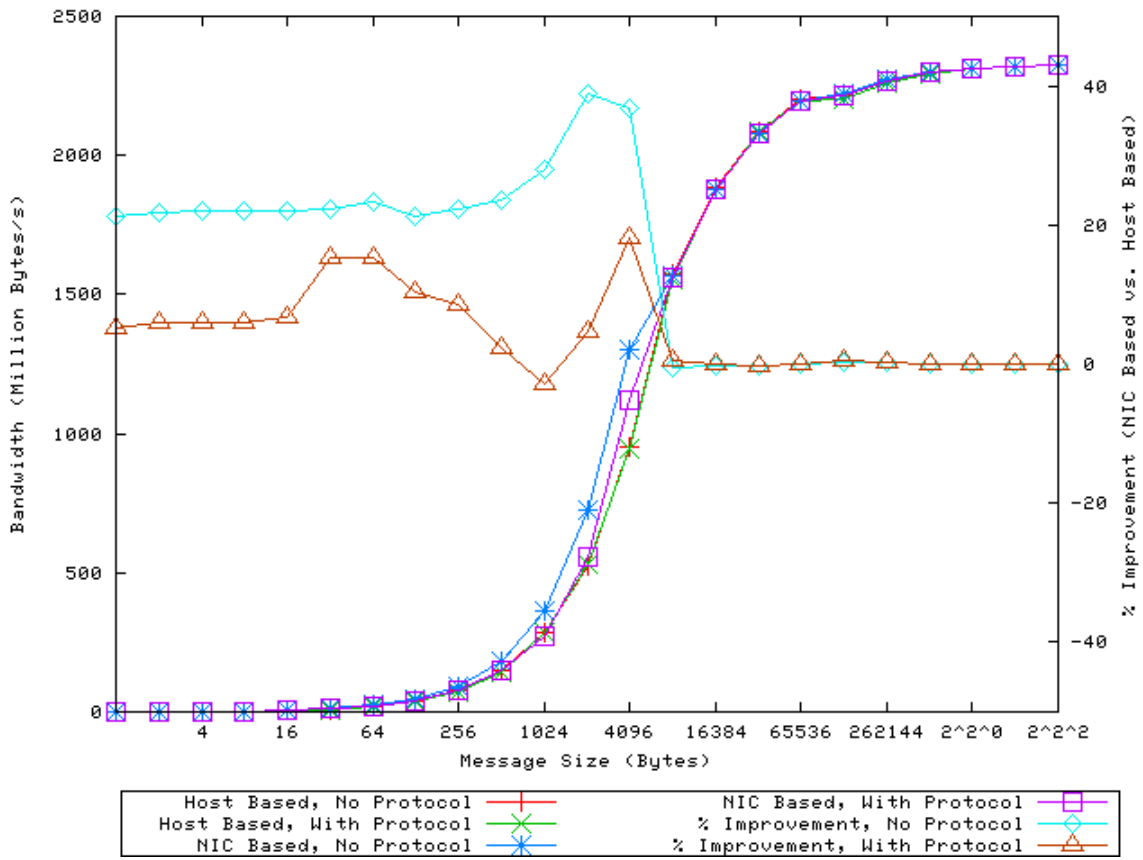
Sandia National Laboratories

Streaming Bandwidth



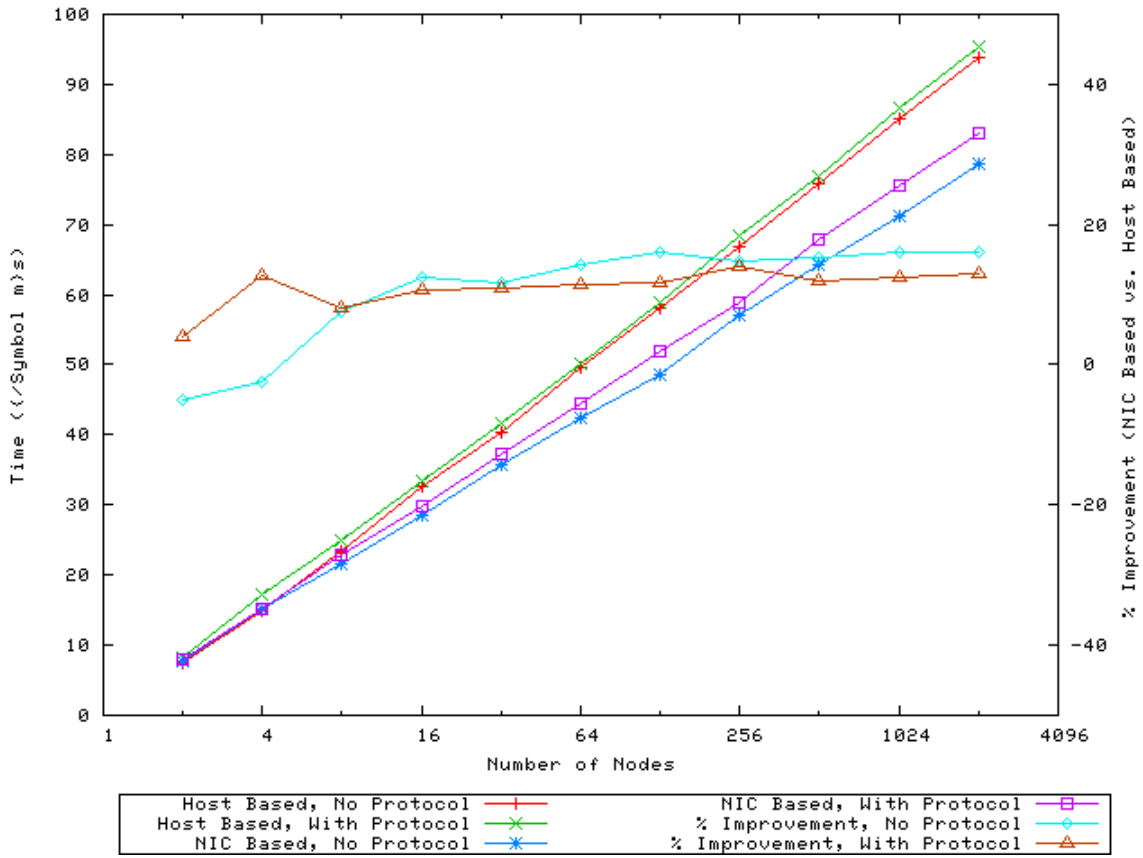
Sandia National Laboratories

Bi-Directional Streaming Bandwidth



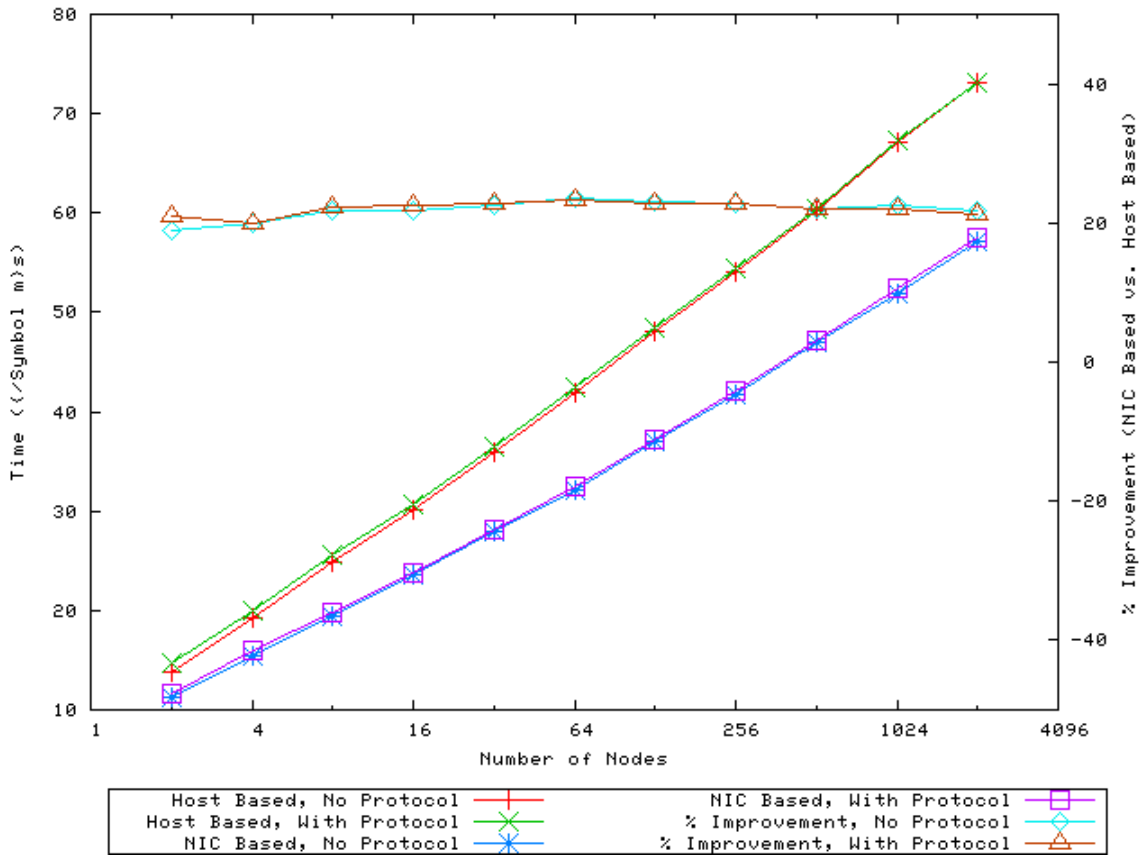
Sandia National Laboratories

Barrier



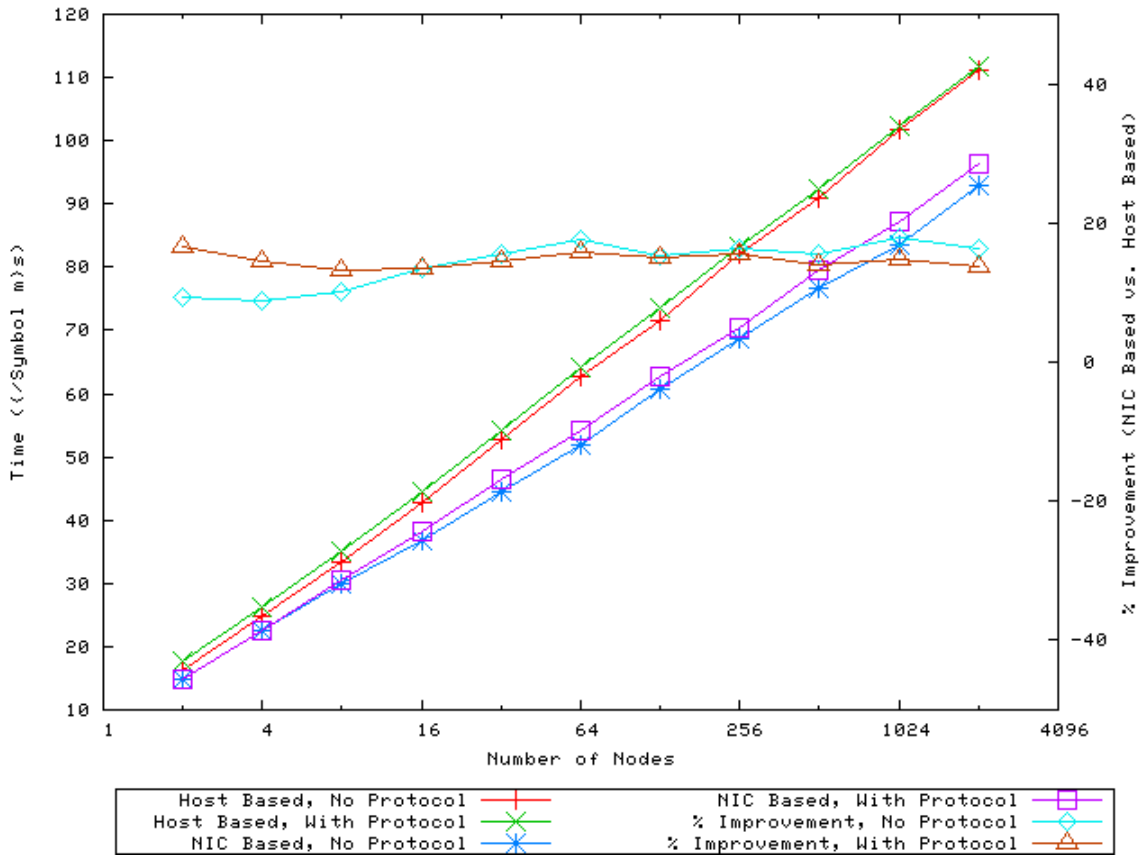
Sandia National Laboratories

Allreduce



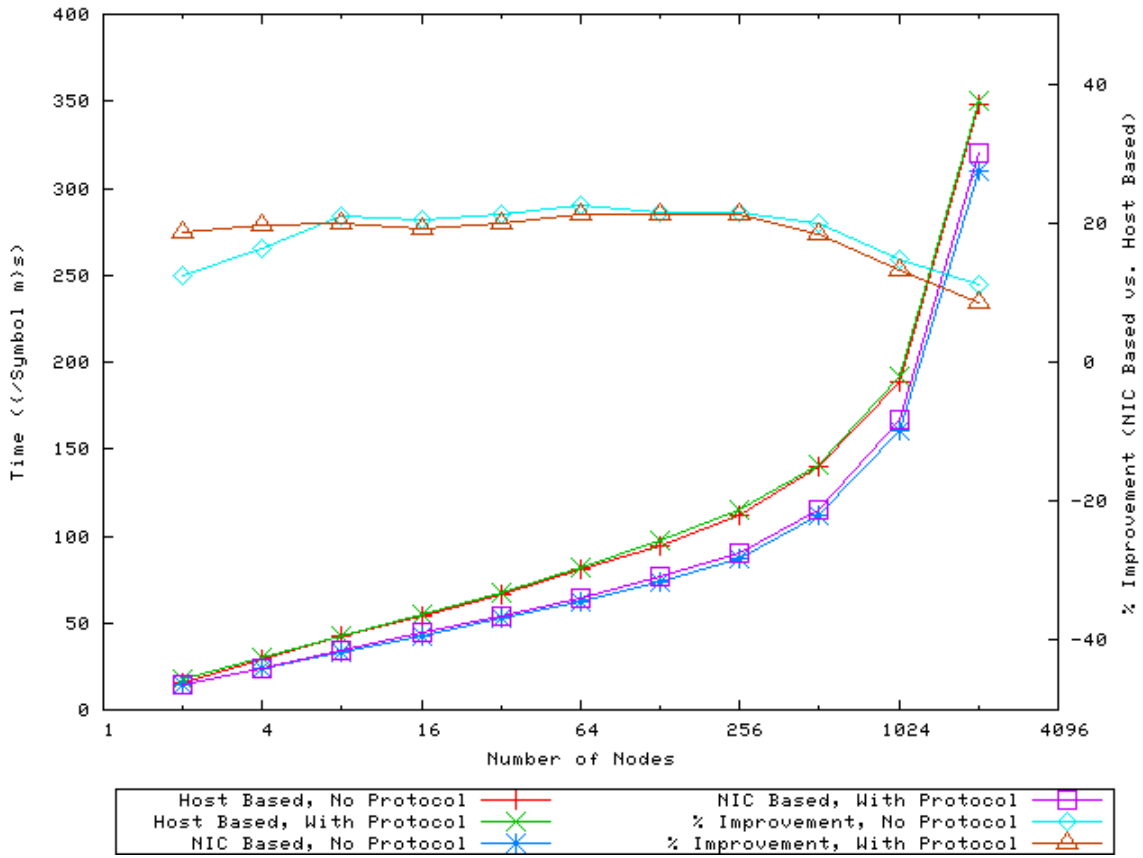
Sandia National Laboratories

Reduce



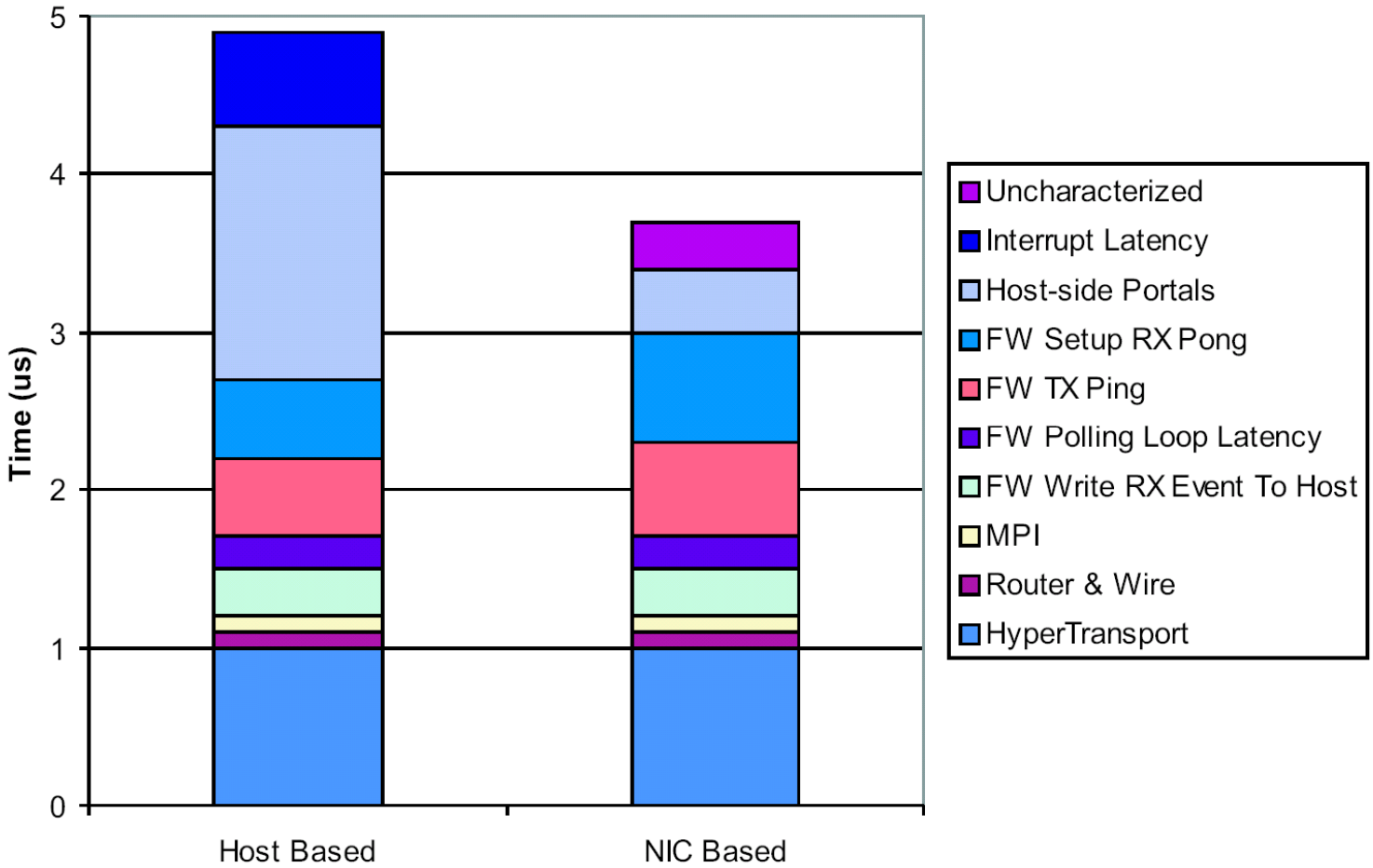
Sandia National Laboratories

Allgather



Sandia National Laboratories

Where the Time Goes



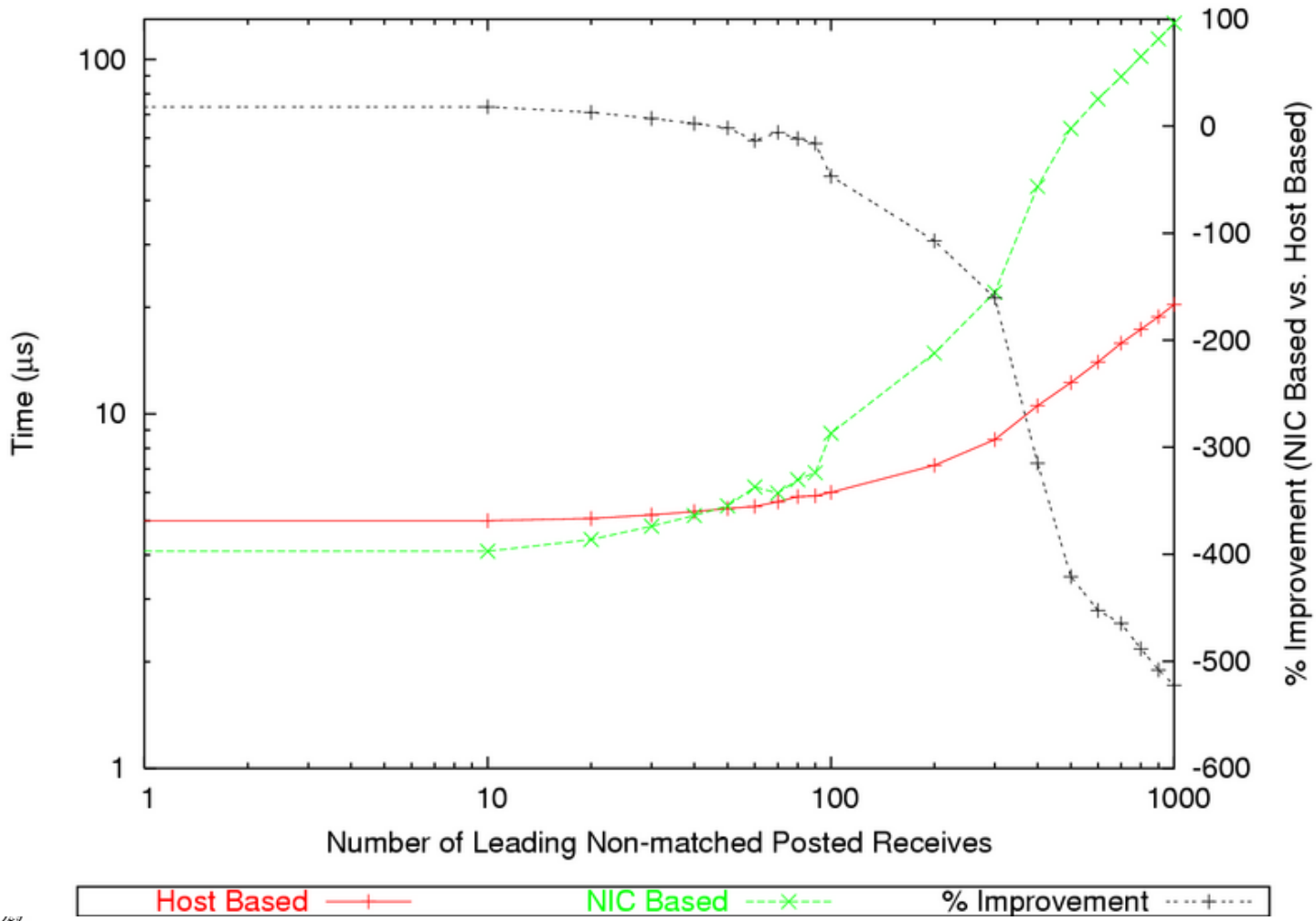
Sandia National Laboratories

Impact of the Length of the Posted Receive Queue

- Most CPU intensive part of Portals is traversing the list of match entries
- Opteron can do this at 2 GHz
- PowerPC can do this at 500 MHz

Sandia National Laboratories

MPI Tag Matching Speed: Host vs. SeaStar



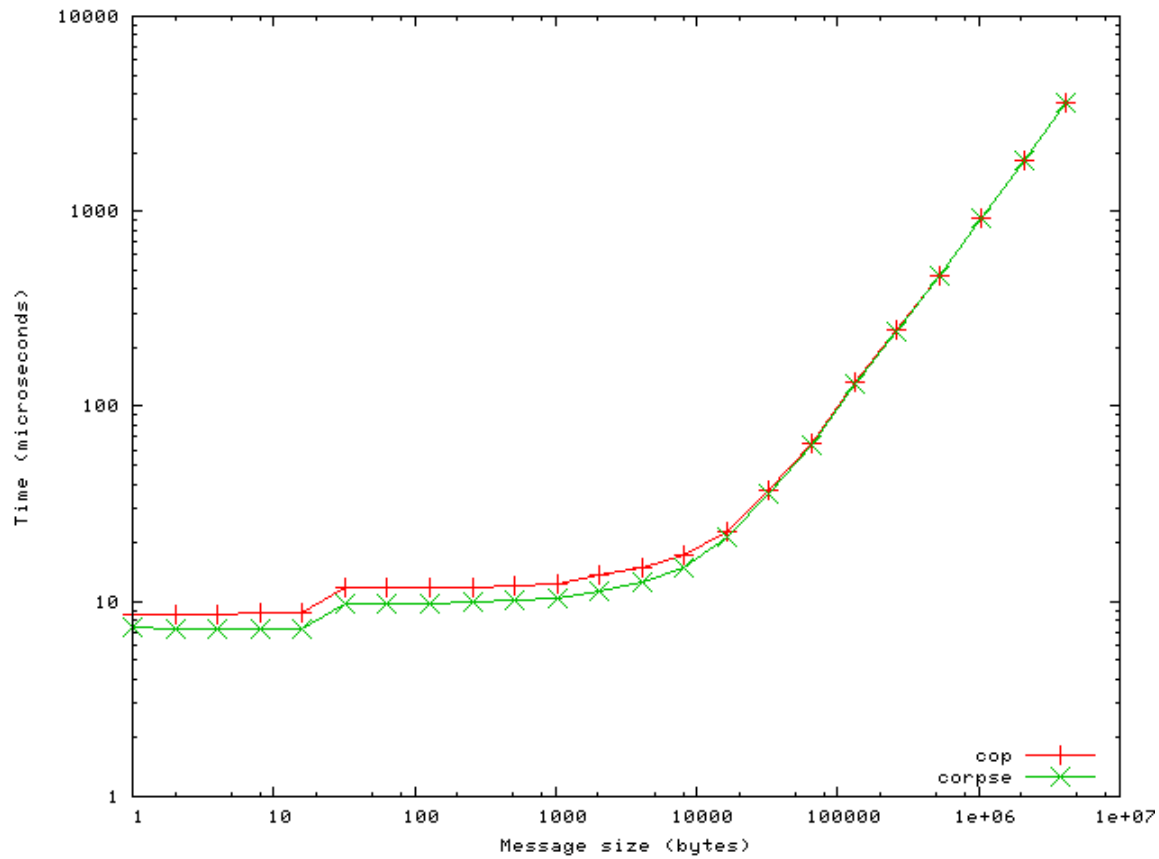
Sandia National Laboratories

Overhead Comparison of Flow Control

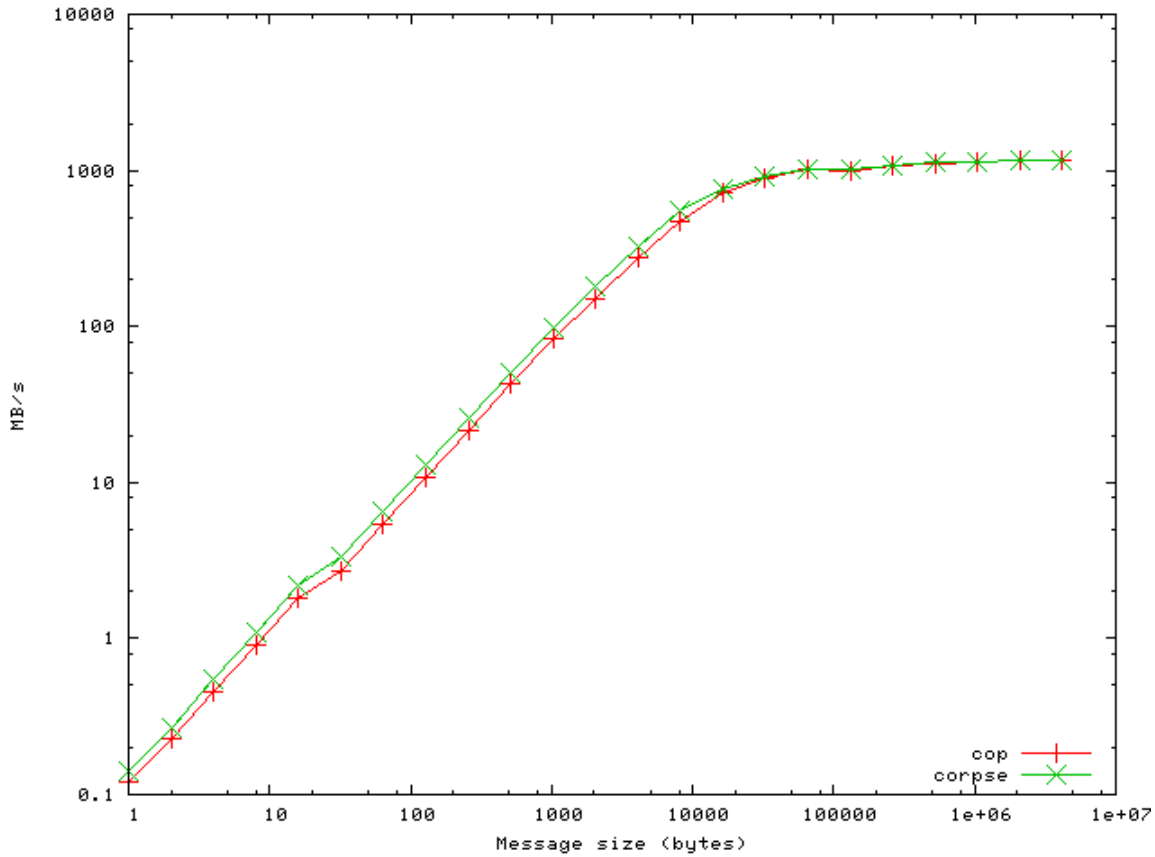
- **Is there an advantage for doing NIC-based flow control versus host-based flow-control?**

Sandia National Laboratories

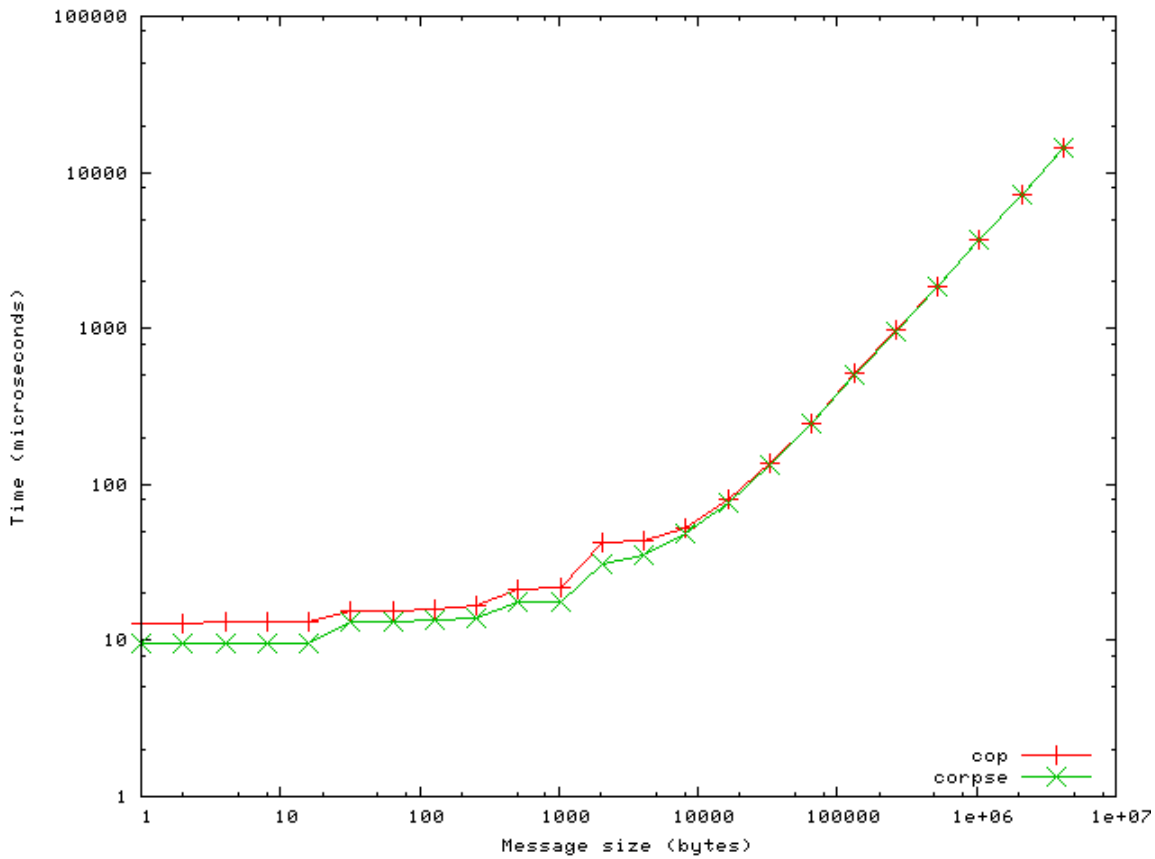
Latency



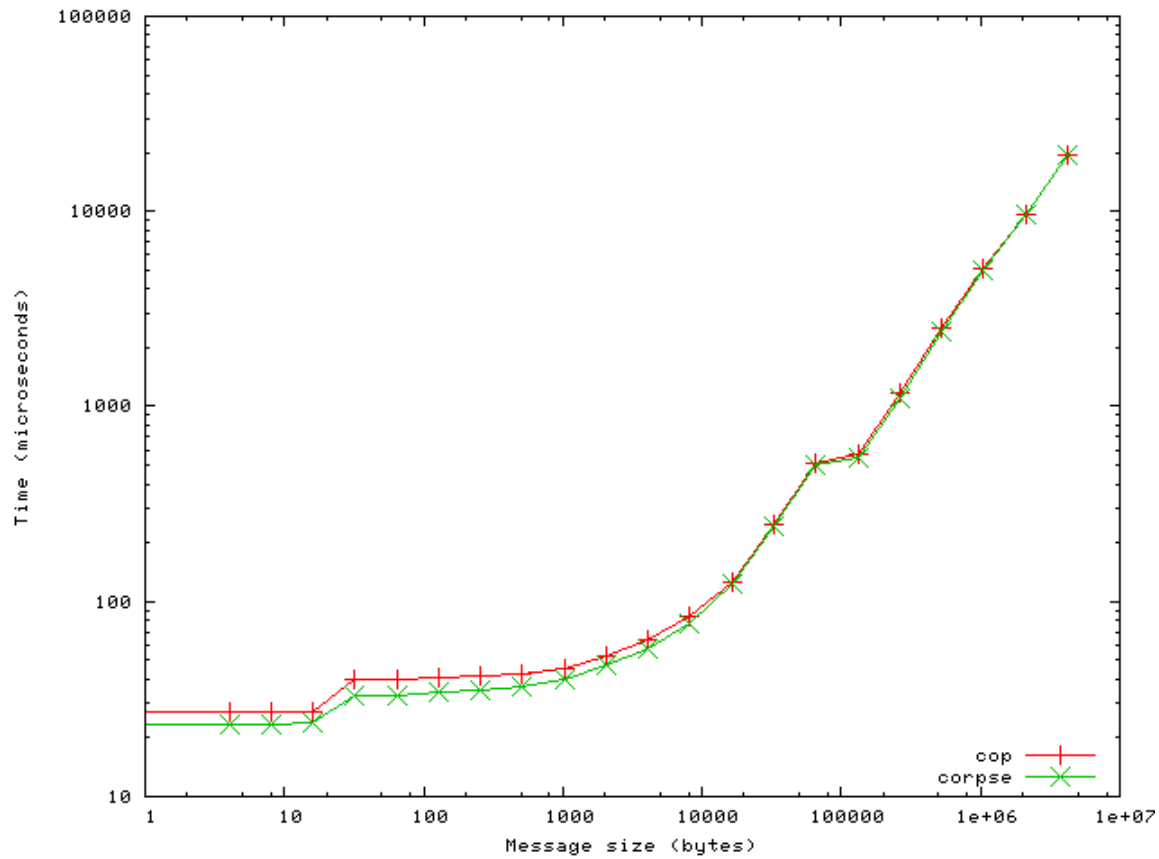
Bandwidth



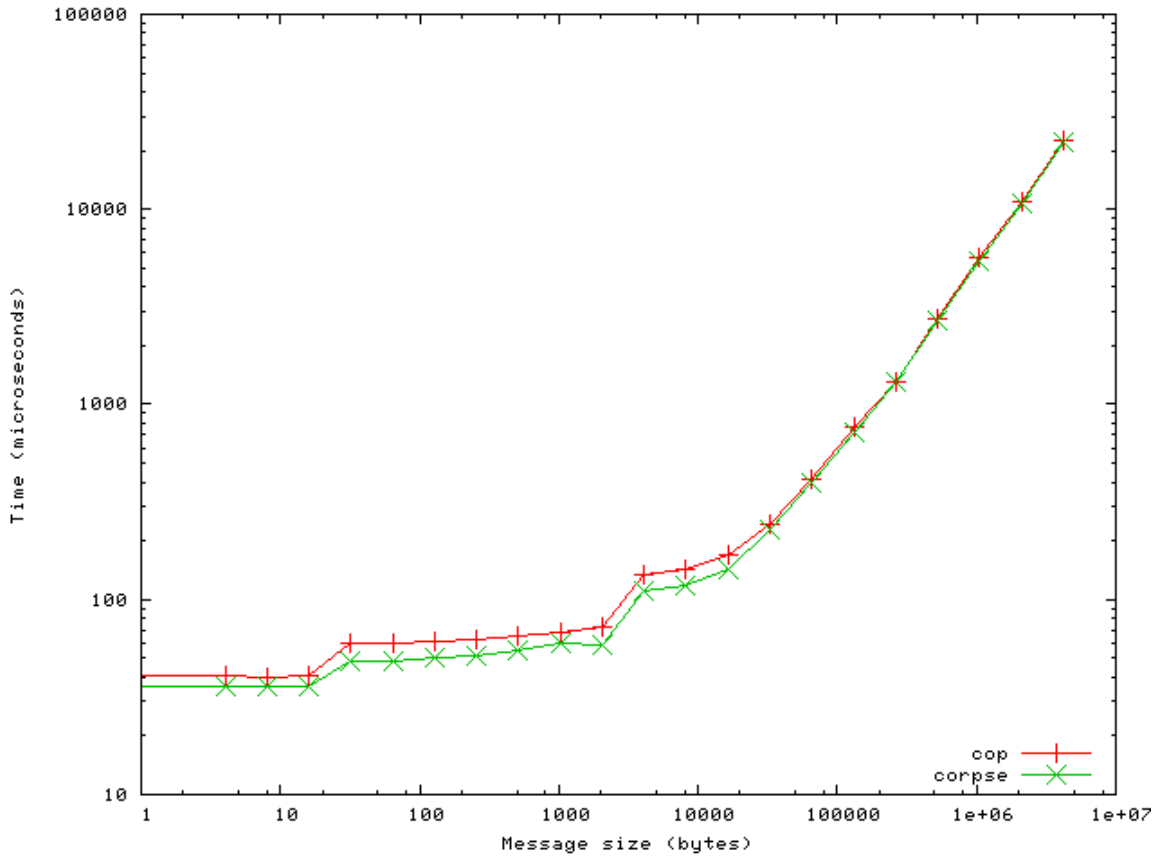
Broadcast



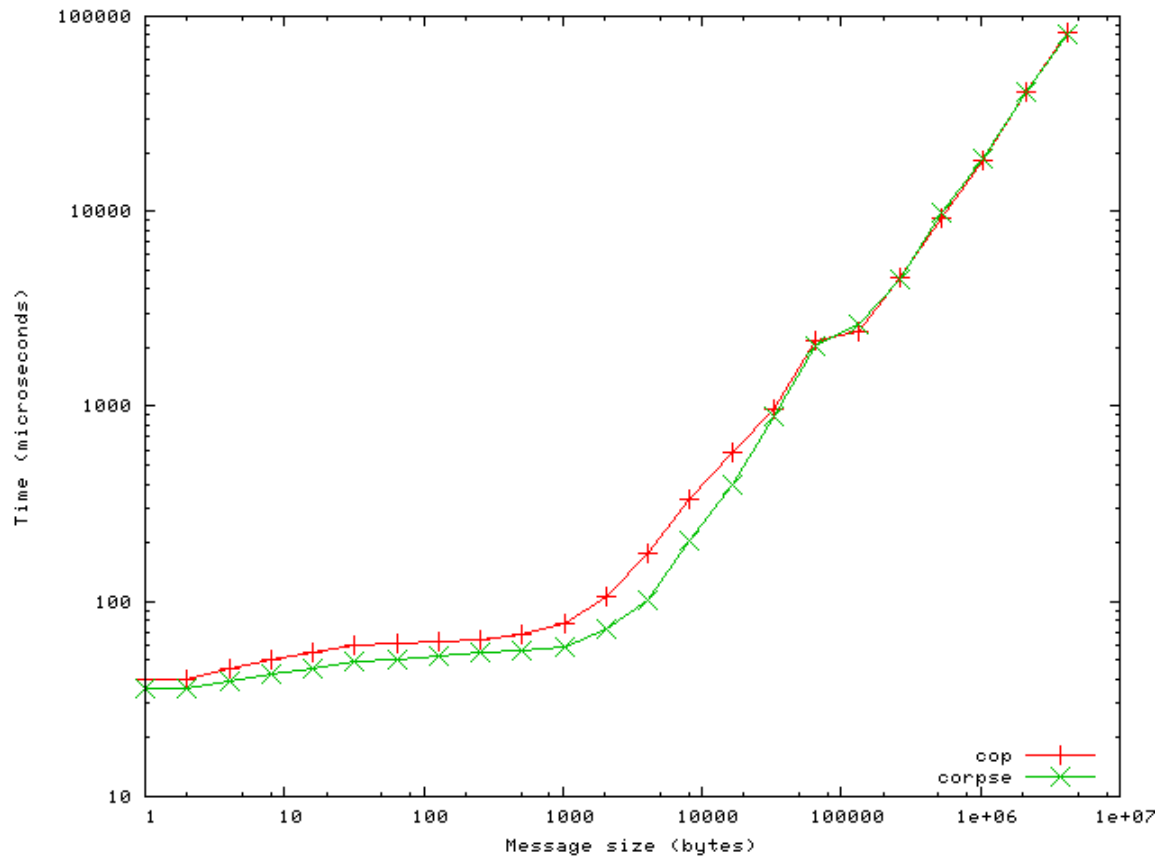
Reduce



Allreduce



Allgather



More Details

- **“Cray’s SeaStar Interconnect: Balanced Bandwidth for Scalable Performance”, Ron Brightwell, Trammell Hudson, Kevin Pedretti, Keith Underwood, IEEE Micro, May/June 2006.**

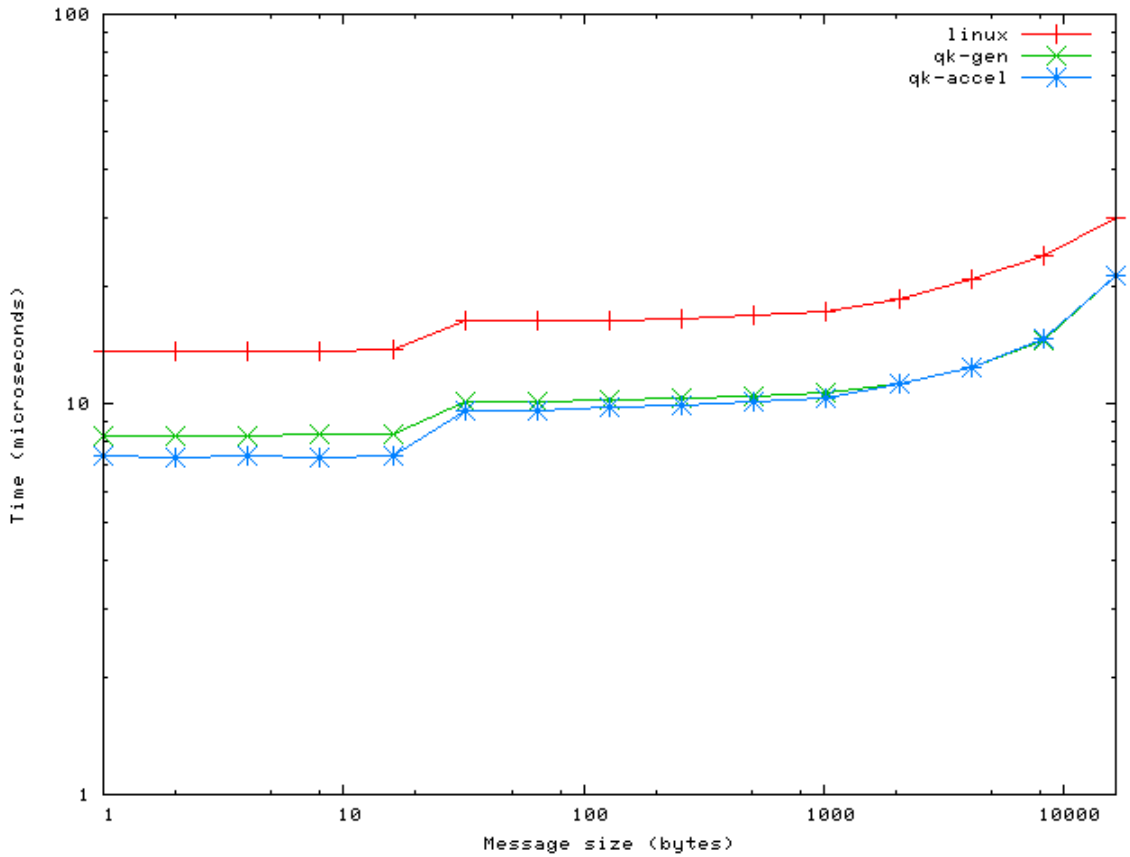
Sandia National Laboratories

Ongoing Work

- Accelerated Portals (AP) being integrated into Cray's development tree for a 1.5 release
- More extensive measurements of the impact of AP
- Portals collective library
  - Collective operations built on top of Portals
- Non-blocking collective functions
  - Collective operations integrated into Portals
  - SeaStar can support offloading collective operations
  - Barrier proof-of-concept is done and working

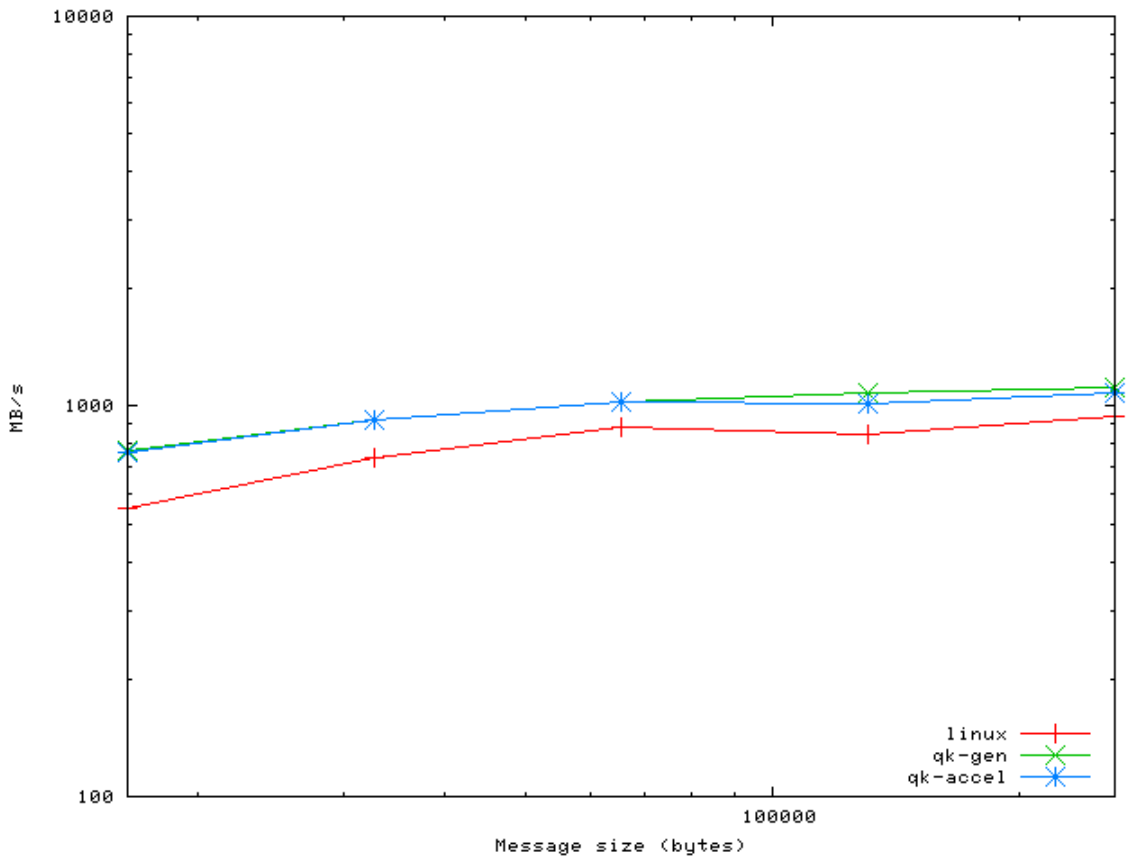
Sandia National Laboratories

Impact of Linux on Latency



Sandia National Laboratories

Impact of Linux on Latency



Sandia National Laboratories