



A New MPI Implementation for Cray SHMEM

Ron Brightwell

Scalable Computing Systems

Sandia National Laboratories

Albuquerque, New Mexico, USA

EuroPVM/MPI

September 21, 2004



Outline

- **Motivation**
- **Implementation**
- **Performance**
- **Limitations**
- **Memory polling strategies**



Motivation

- **Tool for analyzing impact of MPI protocol processing on host processor versus network interface (NIC) processor**
- **Extensive comparisons using identical hardware and similar software stack (QsNet)**
- **SHMEM semantics similar to RDMA capability provided by IB verbs, uDAPL, IWARP, etc.**



Related Work

- **SHMEM device for Cray T3D**
 - No alignment restriction
 - No explicit cache management
- **Similar to strategy used for RDMA-based implementations**
 - MVAPICH from Ohio State



Cray SHMEM Semantics

- **One sided transfers – put/get**
- **Relies on symmetric memory**
 - **Global variables**
 - **Static variables**
 - **Shared heap variables**
- **Strict SPMD model**



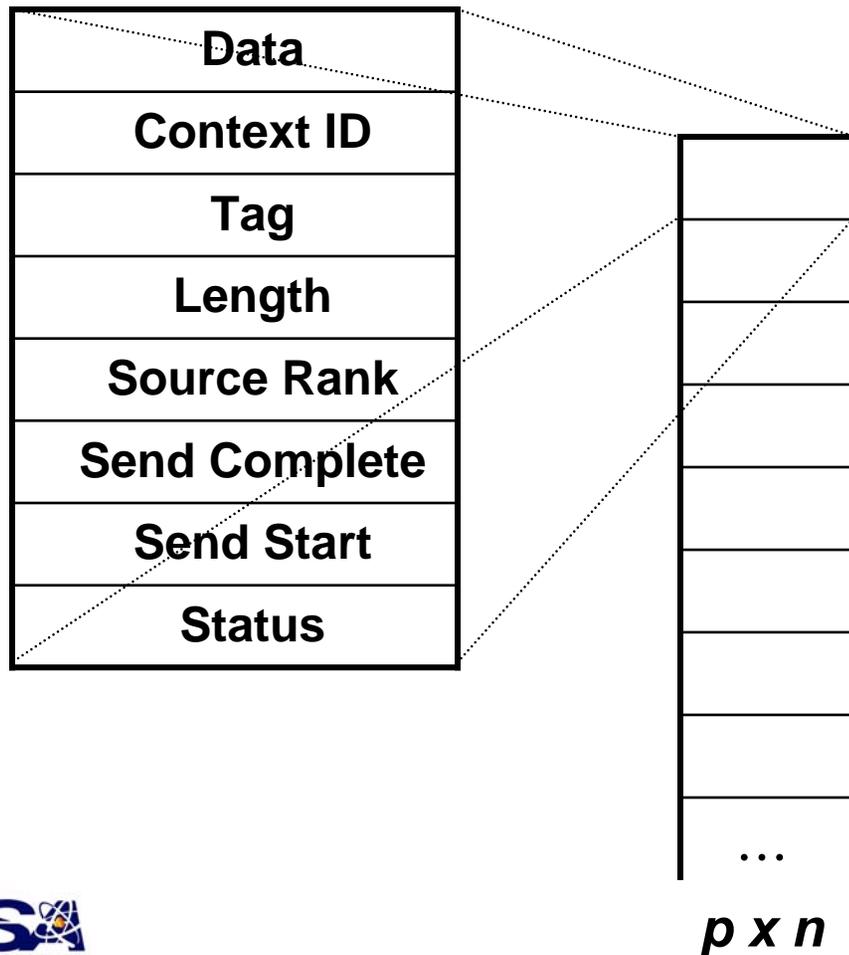
Implementation

- **Native device for MPICH 1.2.5**

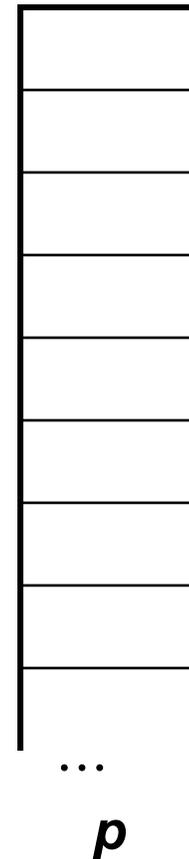


Send Side

Send Packets



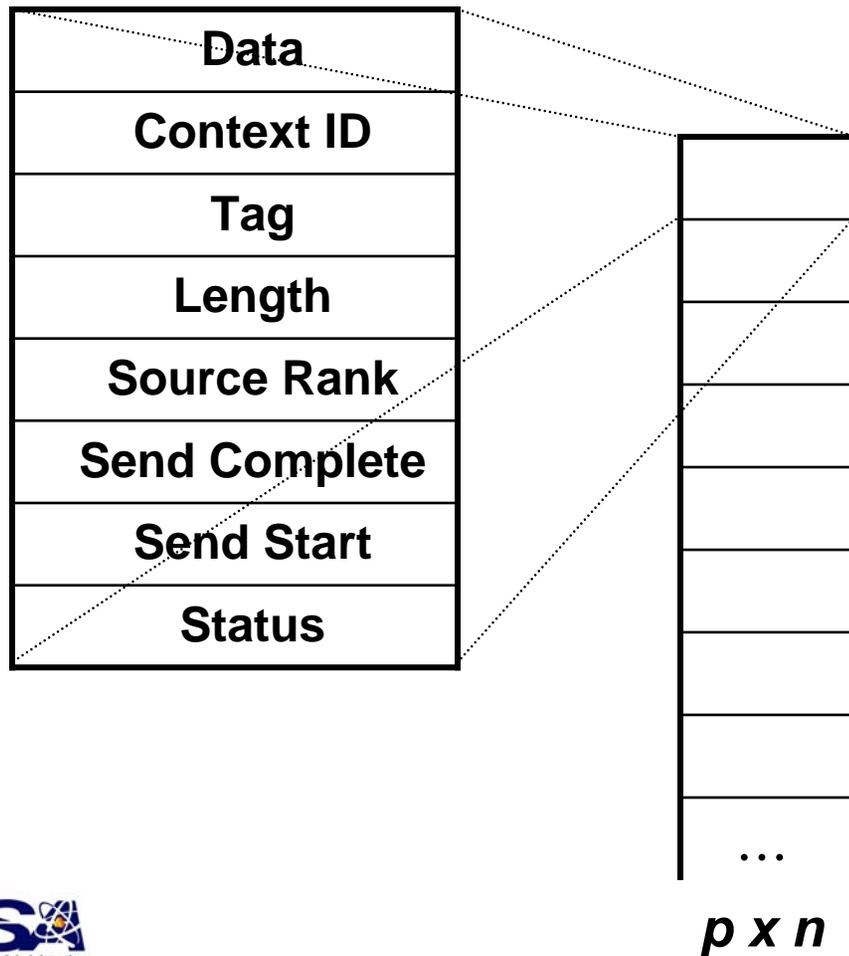
Send Packet Counter



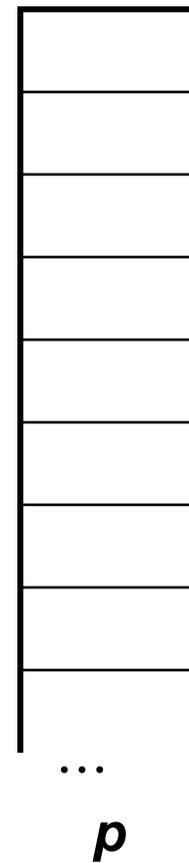


Receive Side

Recv Packets



Recv Packet Counter



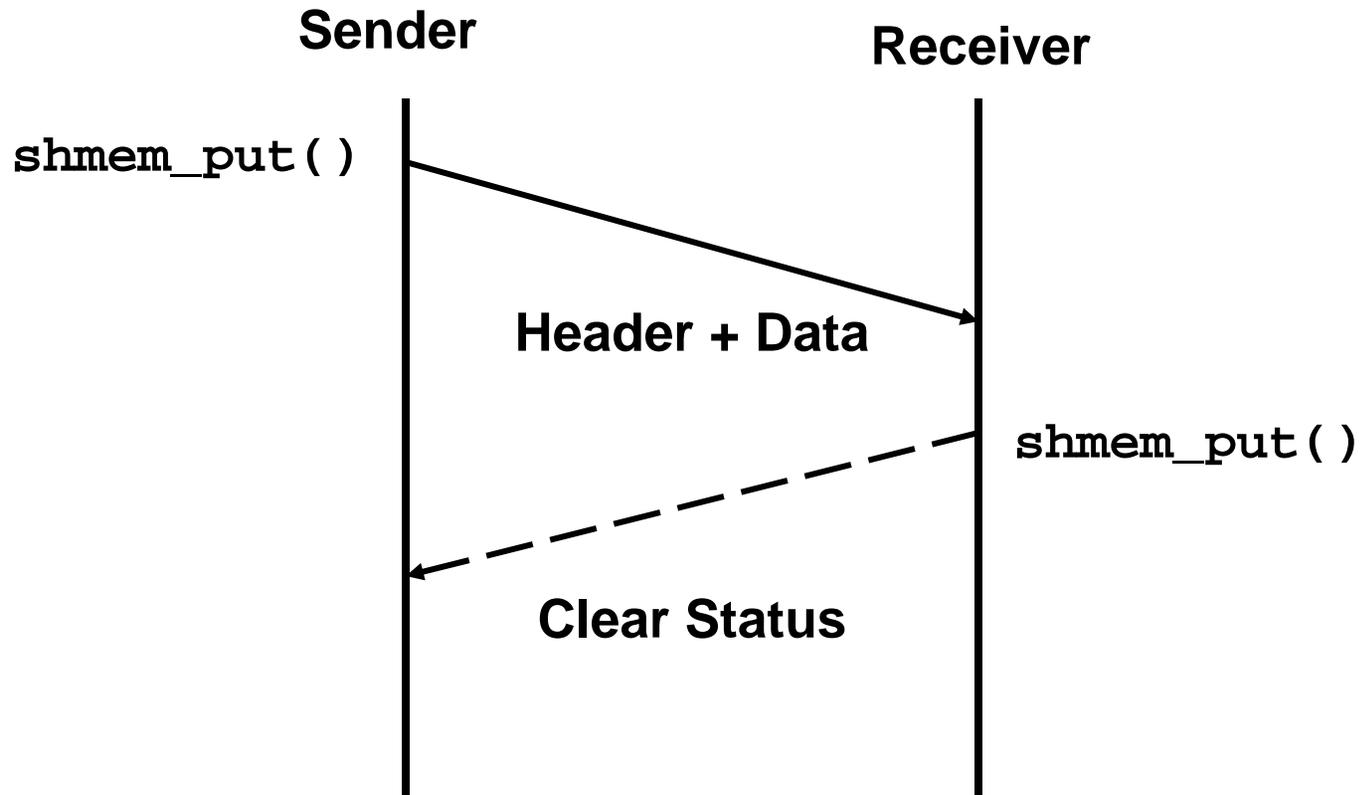


Protocols

- **Short messages**
 - Single packet
 - Send is complete when the packet is sent
- **Long messages**
 - “Send Start” is the address of buffer to be sent
 - “Send Complete” is the address of the completion flag in the request handle
 - No data is sent
 - Receiver pulls data
 - Sets remote completion flag

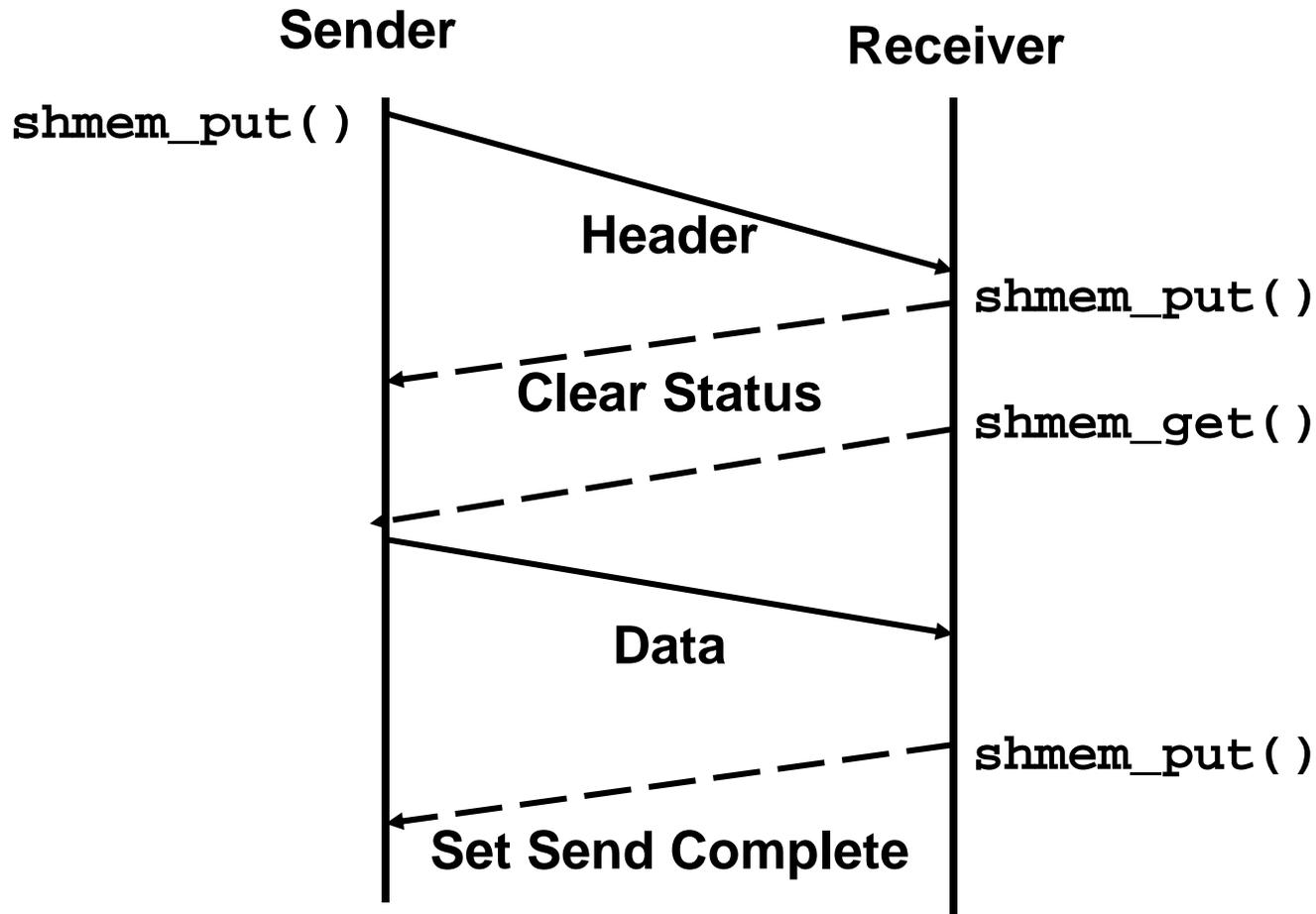


Short Protocol



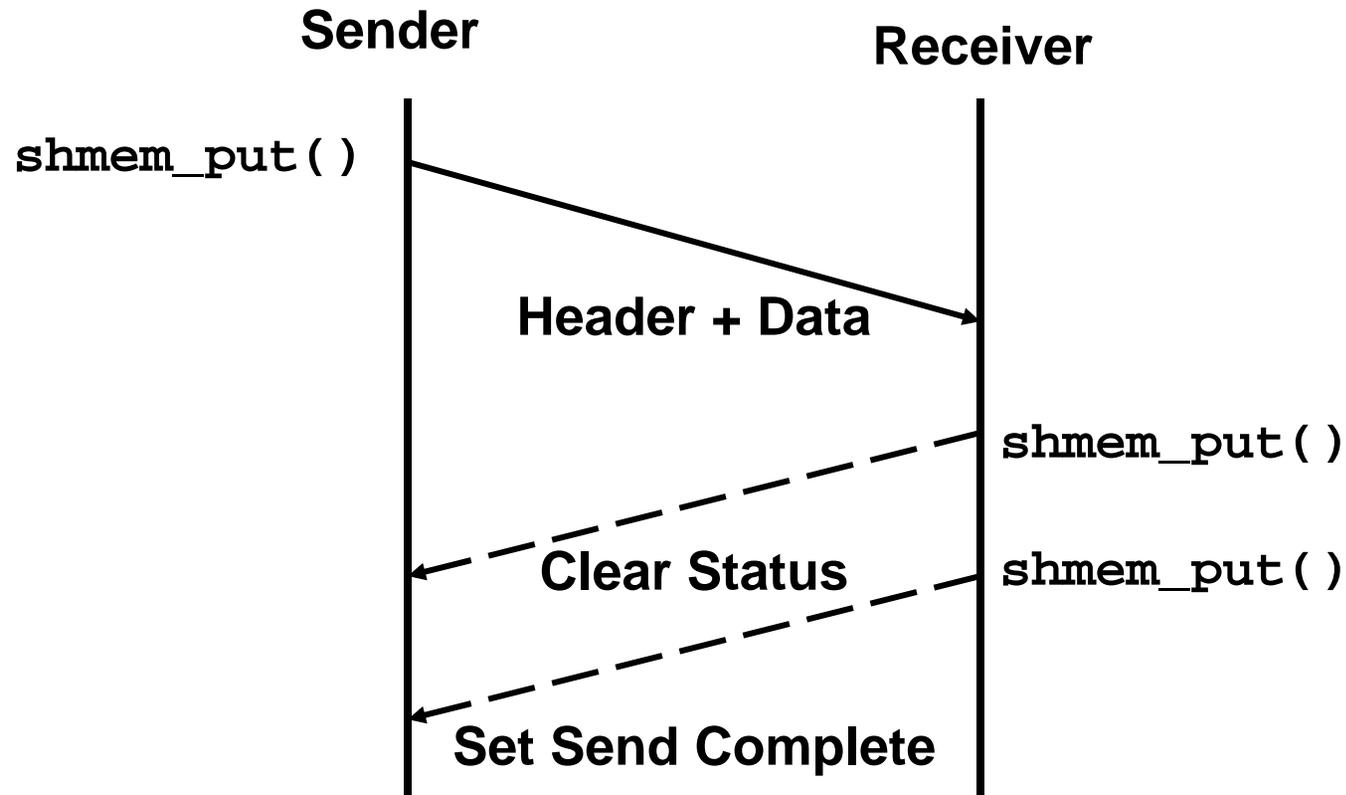


Long Protocol





Short Synchronous Protocol



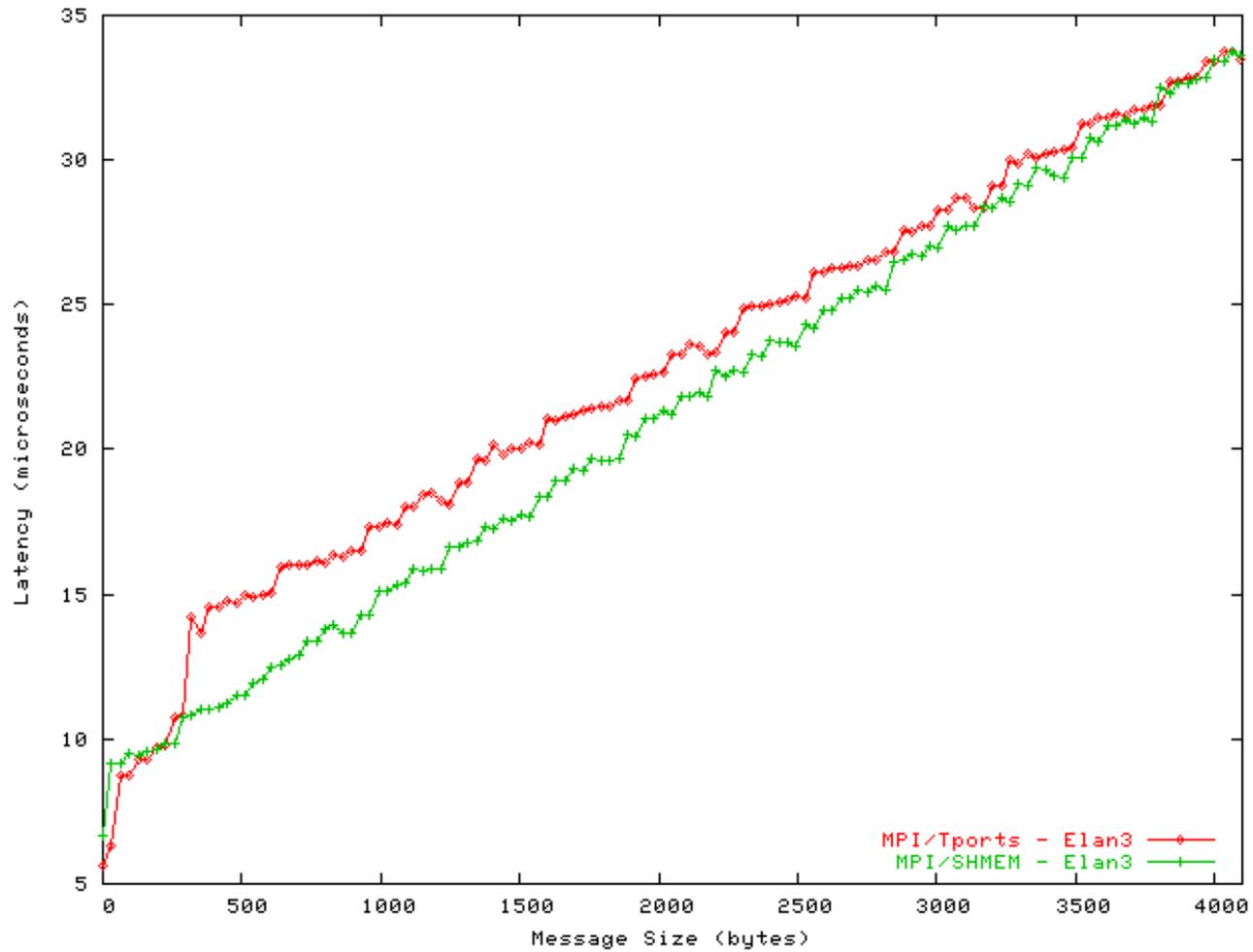


Platforms

- **32-node cluster at LANL**
 - Dual 1 GHz Intel Itanium-2 CPUs
 - 2 GB main memory
 - 2 Quadrics QsNet (Elan-3) NICs
 - Linux 2.4.21
- **128-node cluster at Sandia**
 - Dual 2.0 GHz AMD Opteron
 - 2 GB main memory
 - 1 Quadrics QsNet-II (Elan-4) NIC
 - Linux 2.4.21

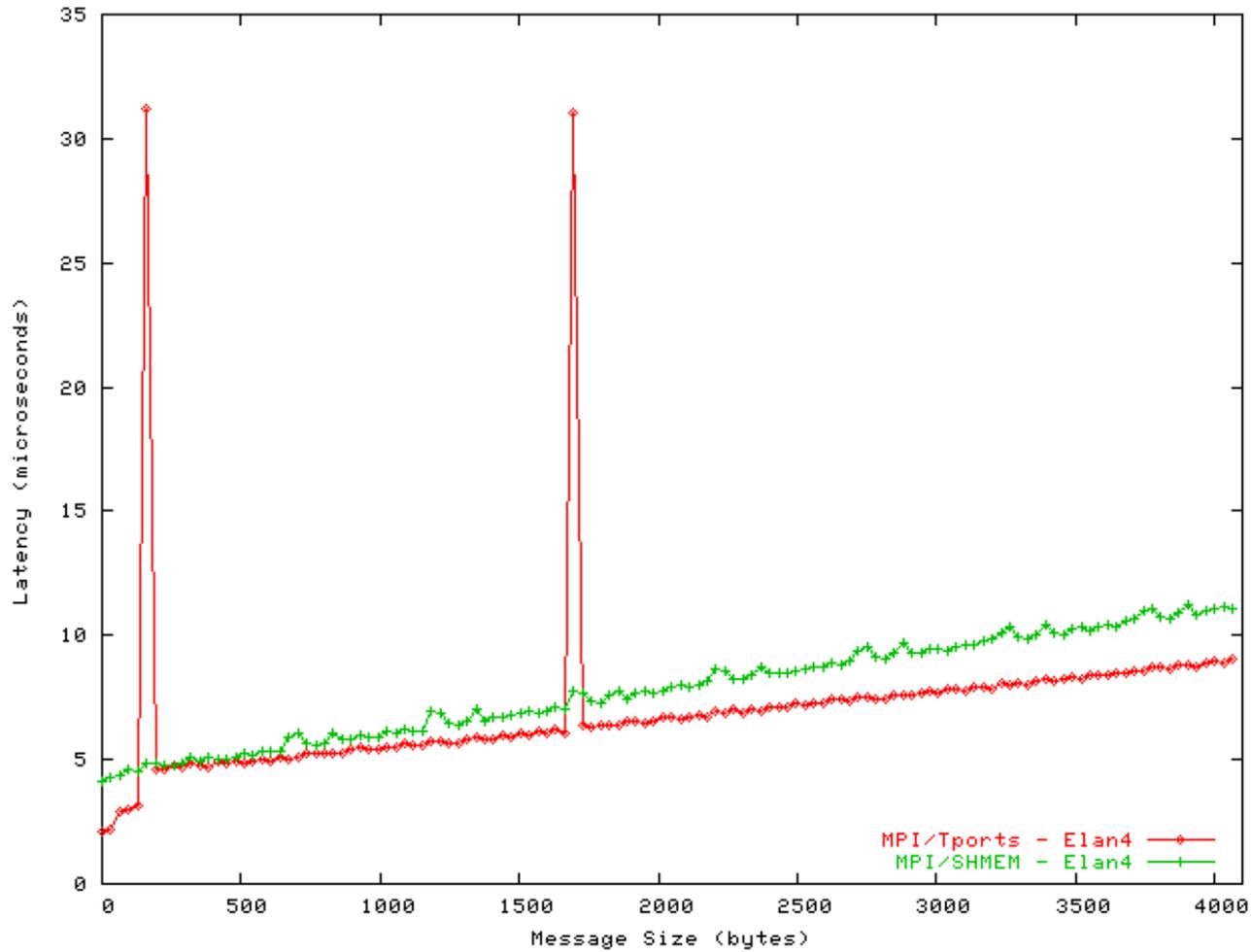


Elan-3 Latency



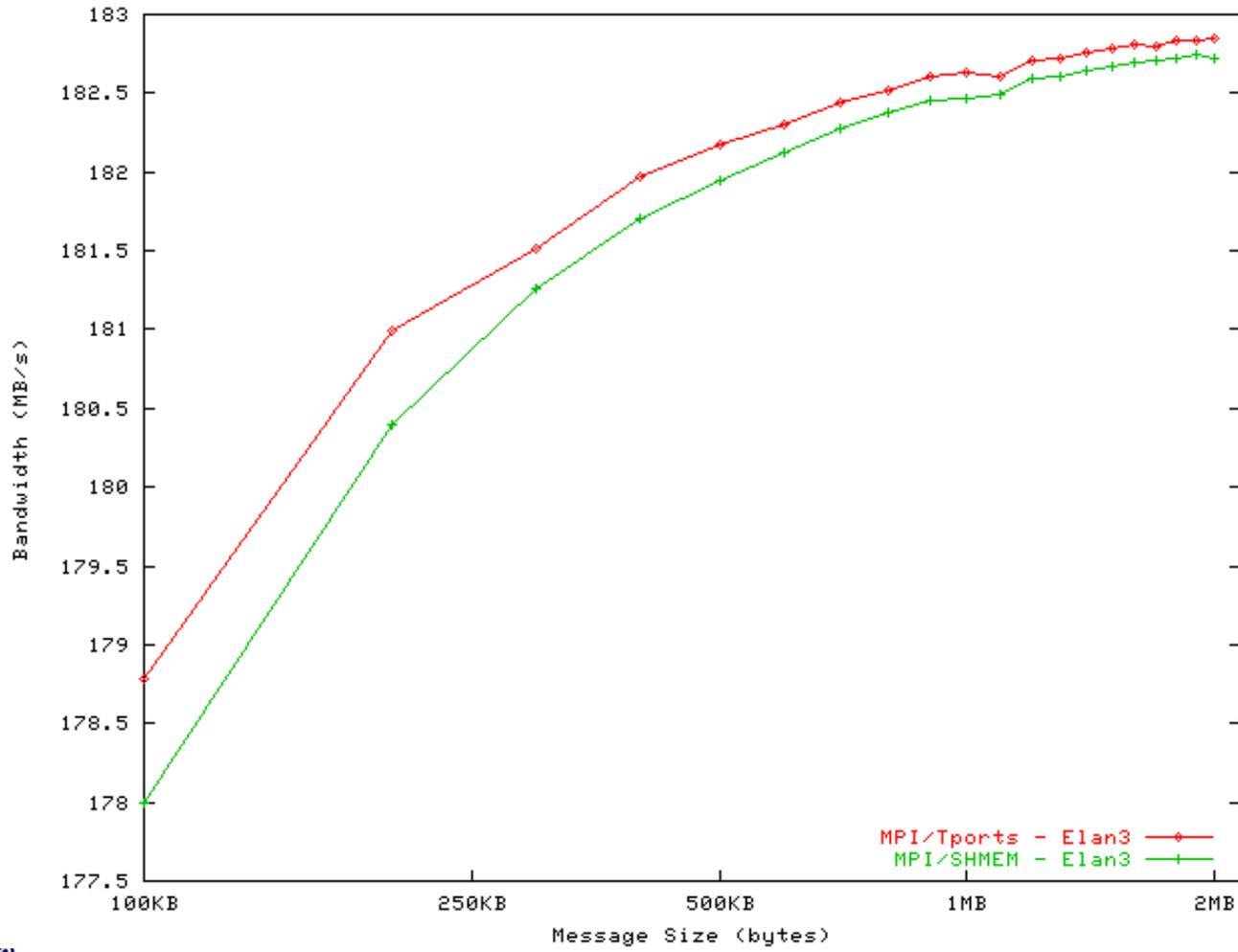


Elan-4 Latency



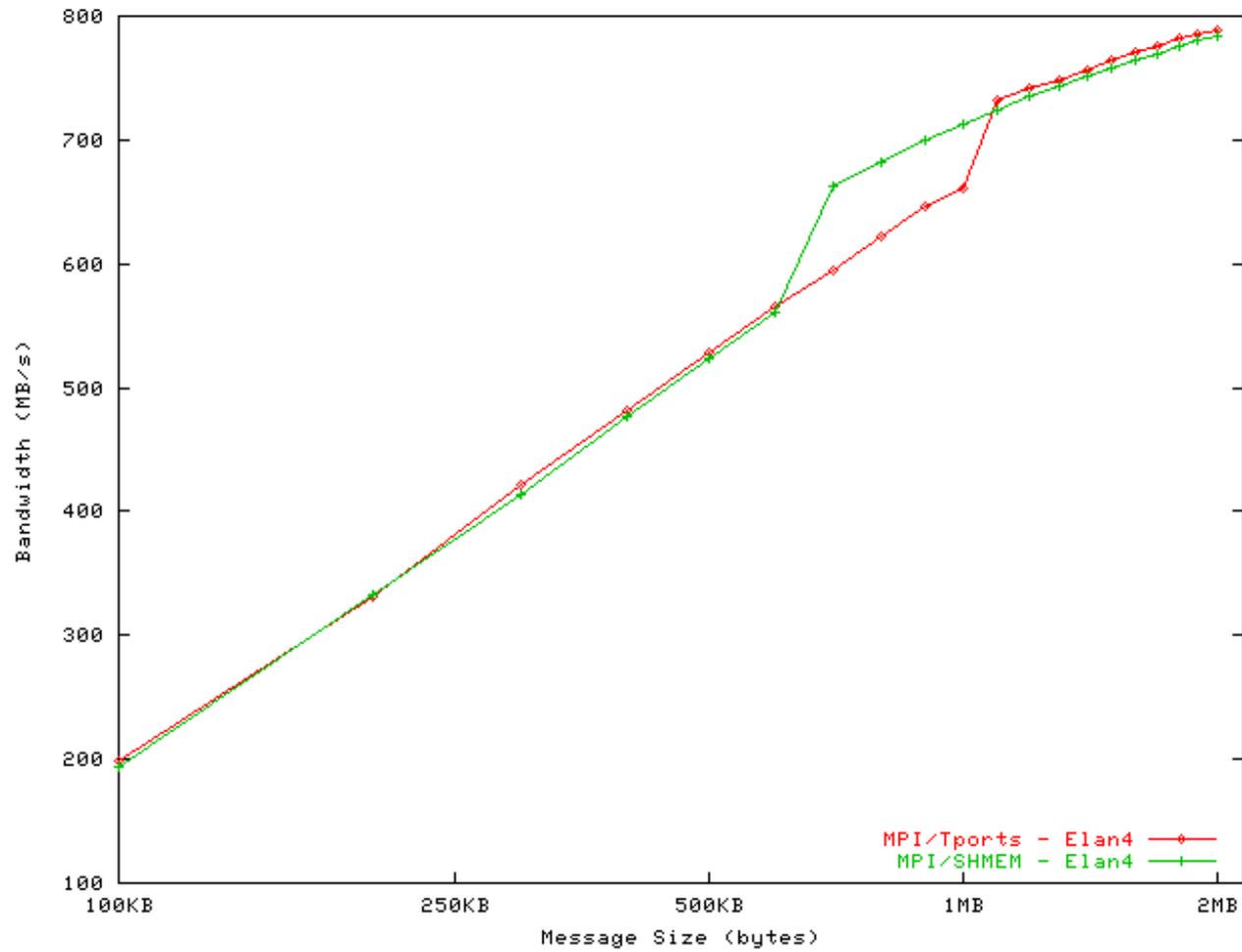


Elan-3 Bandwidth



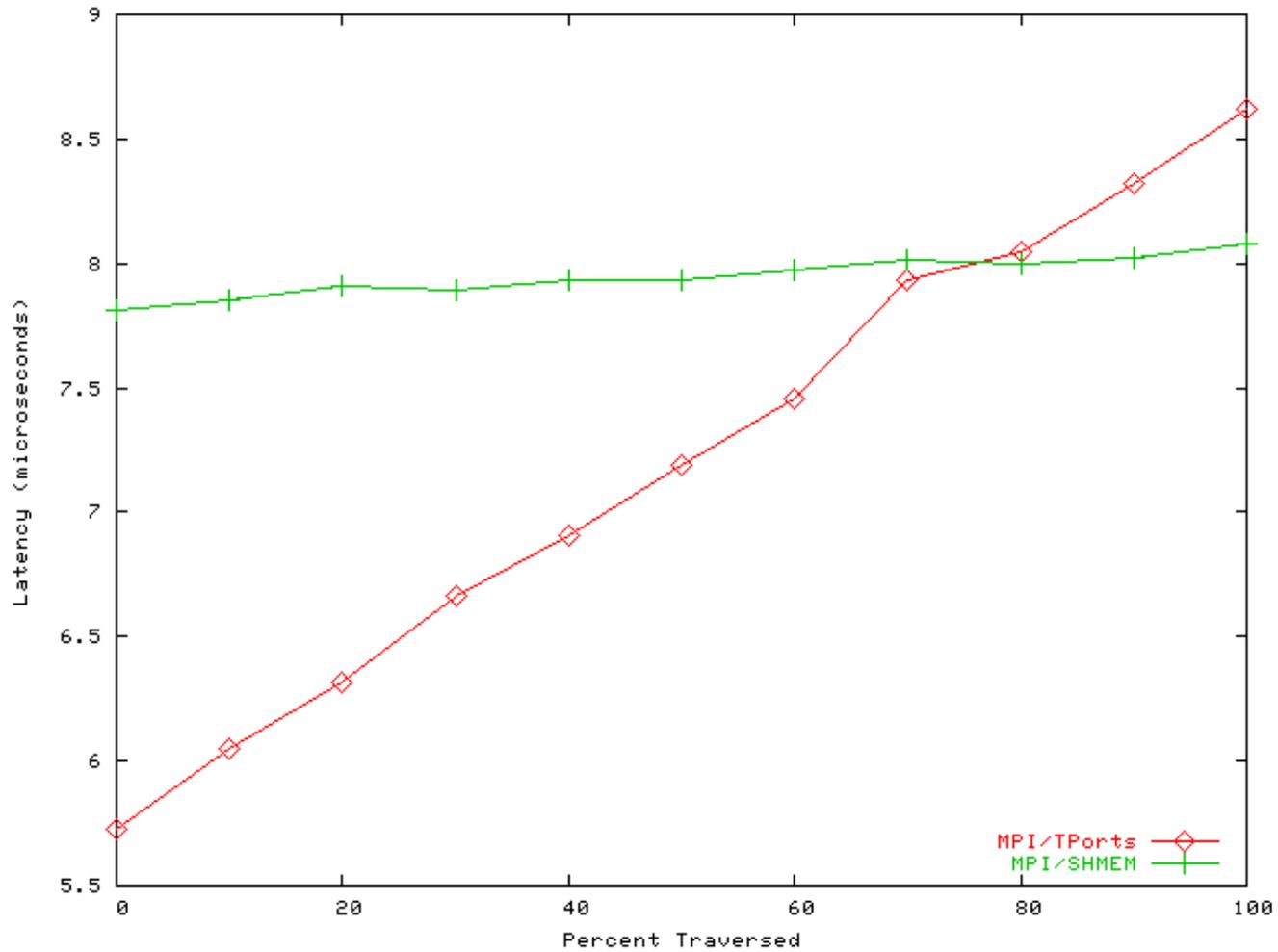


Elan-4 Bandwidth



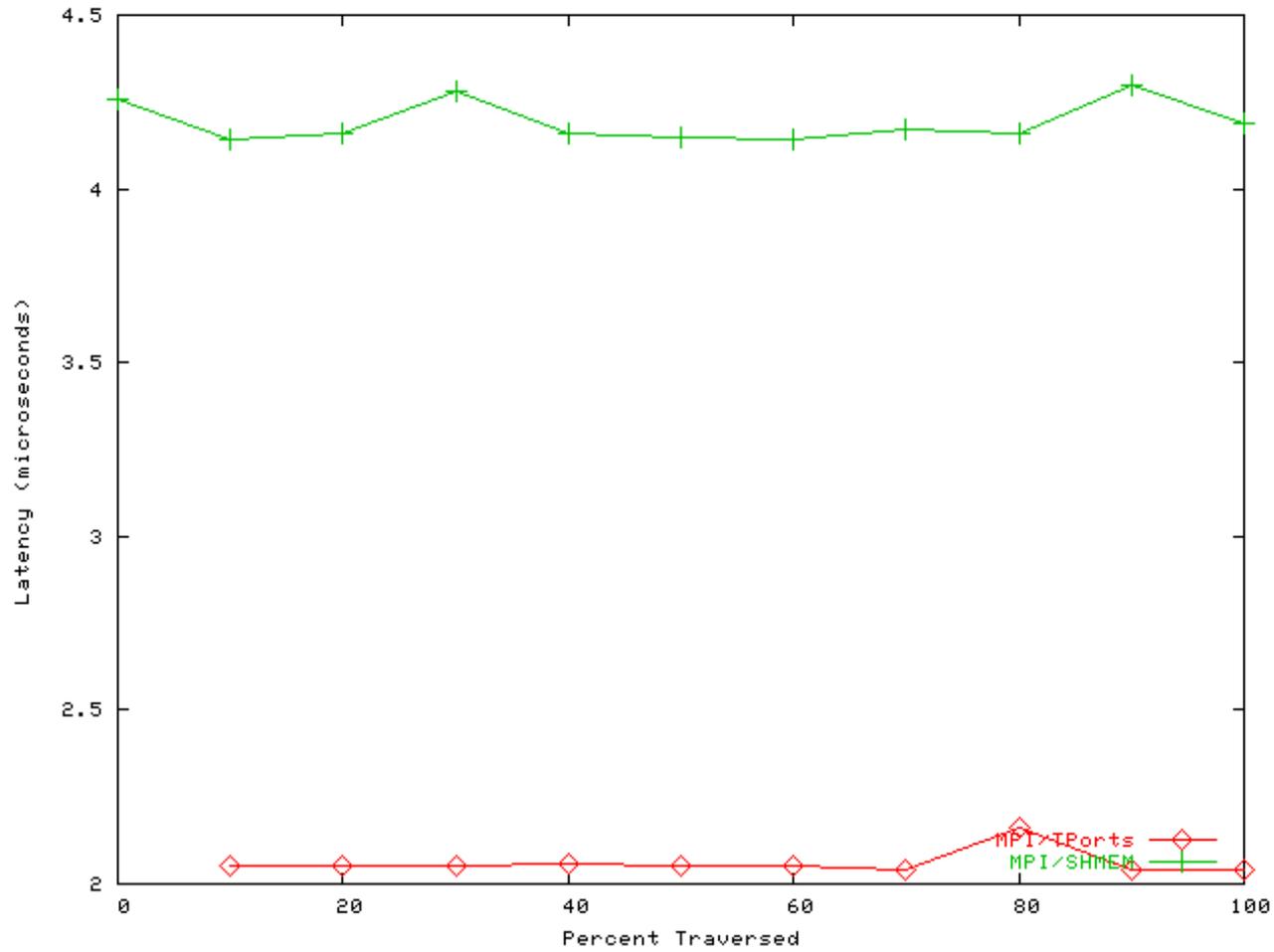


Elan-3 Pre-Posted Latency (10 Entries)





Elan-4 Pre-Posted Latency (10 Entries)





Limitations

- Amount of host memory scales linearly with number of processes in job
- Does not support independent progress
 - MPI library calls must be made in order for long messages to move
 - Need to use a user-level thread for progress
- Non-blocking puts/gets are not used
- Limited to the SPMD model
- Looking for incoming messages is not very efficient



Memory Polling Strategies

- **Two polling**
 - Optimization for ping-pong benchmark 😊
- **Naïve polling**
 - Start with rank 0 slot and loop through all ranks
- **Fair polling**
 - Start with the rank beyond where last msg found
- **Cached polling**
 - Cache the N most popular ranks
- **Posted queue polling**
 - Use posted receive queue as a hint



Related Papers

- **“An Initial Analysis of the Impact of Overlap and Independent Progress for MPI,”** Ron Brightwell, Keith Underwood, and Rolf Riesen, in *Proceedings of the 11th EuroPVM/MPI Users’ Group Meeting*, September 2004.
- **“The Impact of MPI Queue Usage on Message Latency,”** Keith Underwood and Ron Brightwell, in *Proceedings of the 2004 International Conference on Parallel Processing*, August 2004.
- **“An Analysis of the Impact of MPI Overlap and Independent Progress,”** Ron Brightwell and Keith Underwood, in *Proceedings of the 18th Annual ACM International Conference on Supercomputing*, June 2004.