# Open MPI on the Cray XT

LEADERSHIP
COMPUTING FACILITY
NATIONAL CENTER FOR COMPUTATIONAL SCIENCES


*presented by*

**Richard L. Graham**

**Galen Shipman**

Oak Ridge National Laboratory
U.S. Department of Energy

# Open MPI Is…

- Open source project / community
- Consolidation and evolution of several prior MPI implementations
- All of MPI-1 and MPI-2
- Production quality
- Vendor-friendly
- Research- and academic-friendly

# Current Membership

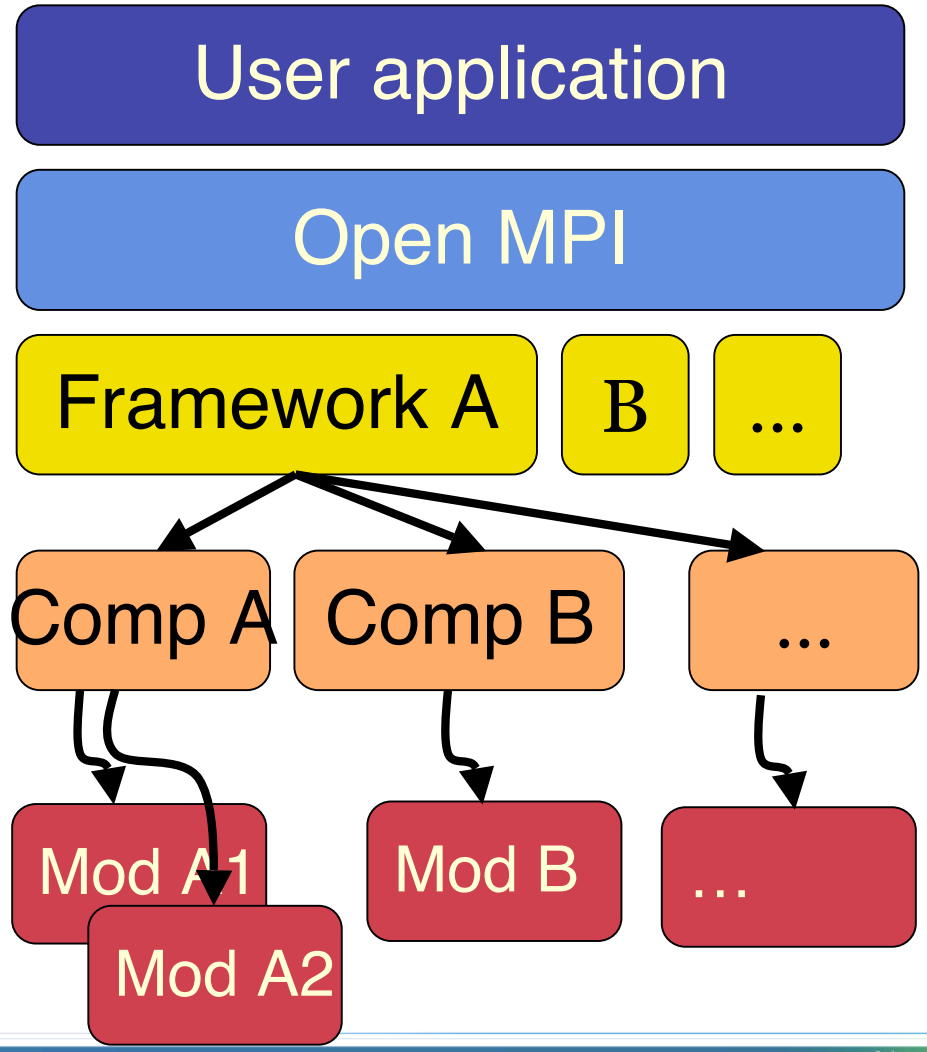
- 14 members, 9 contributors, 1 partner
  - **4 US DOE labs**
  - **8 universities**
  - **10 vendors**
  - **1 individual**

# Some Current Highlights

- Production MPI on SNL's Thunderbird
- Production MPI on LANL's Road Runner
- Working on getting it up on TACC (Ranger)
- The MPI used for the EU QosCosGrid: Quasi-Opportunistic Complex System Simulations on Grid
- Tightly integrated with VampirTrace (vs 1.3)

# Modular Component Architecture

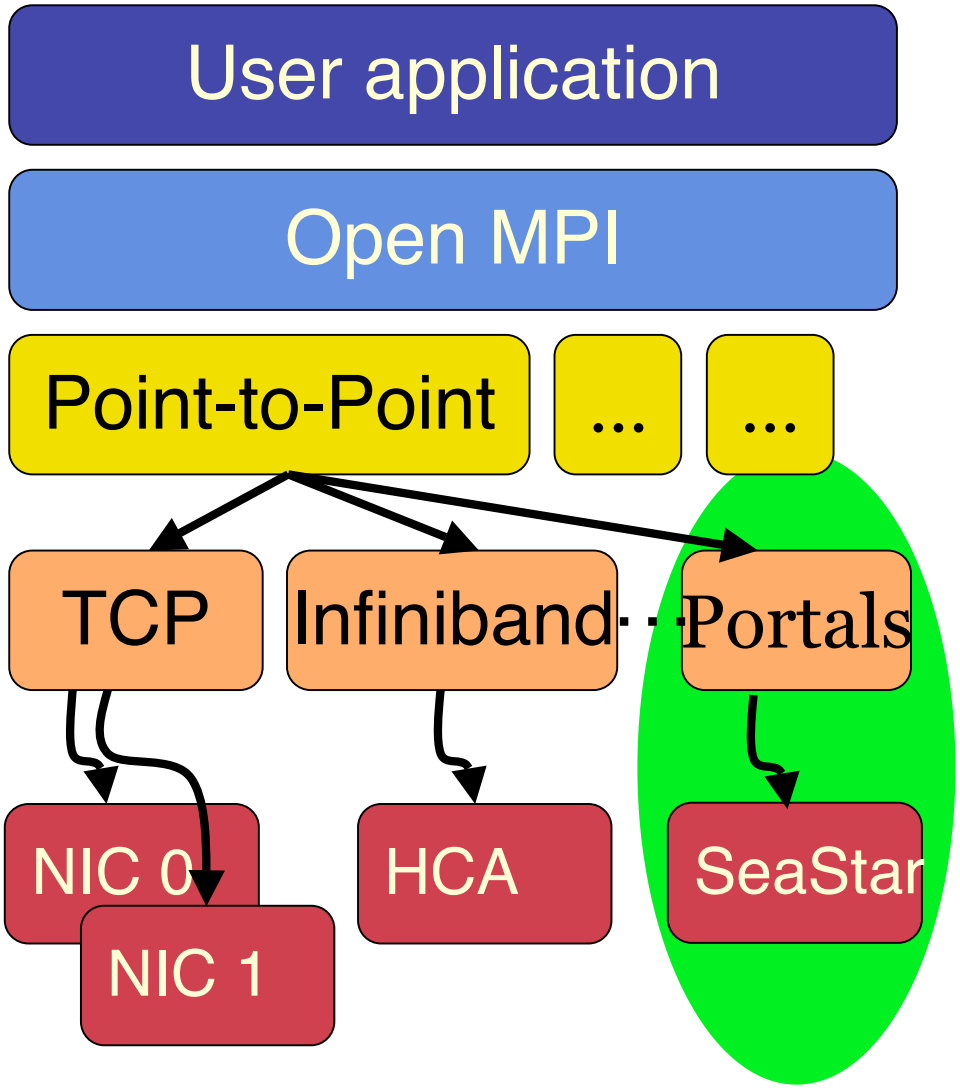
- **Framework**:
  - **API targeted at a specific task**
    - **PTP message management**
    - **PTP transfer layer**
    - **Collectives**
    - **Process startup ….**
- **Component**:
  - **An implementation of a framework's API**
- **Module**:
  - **An instance of a component**



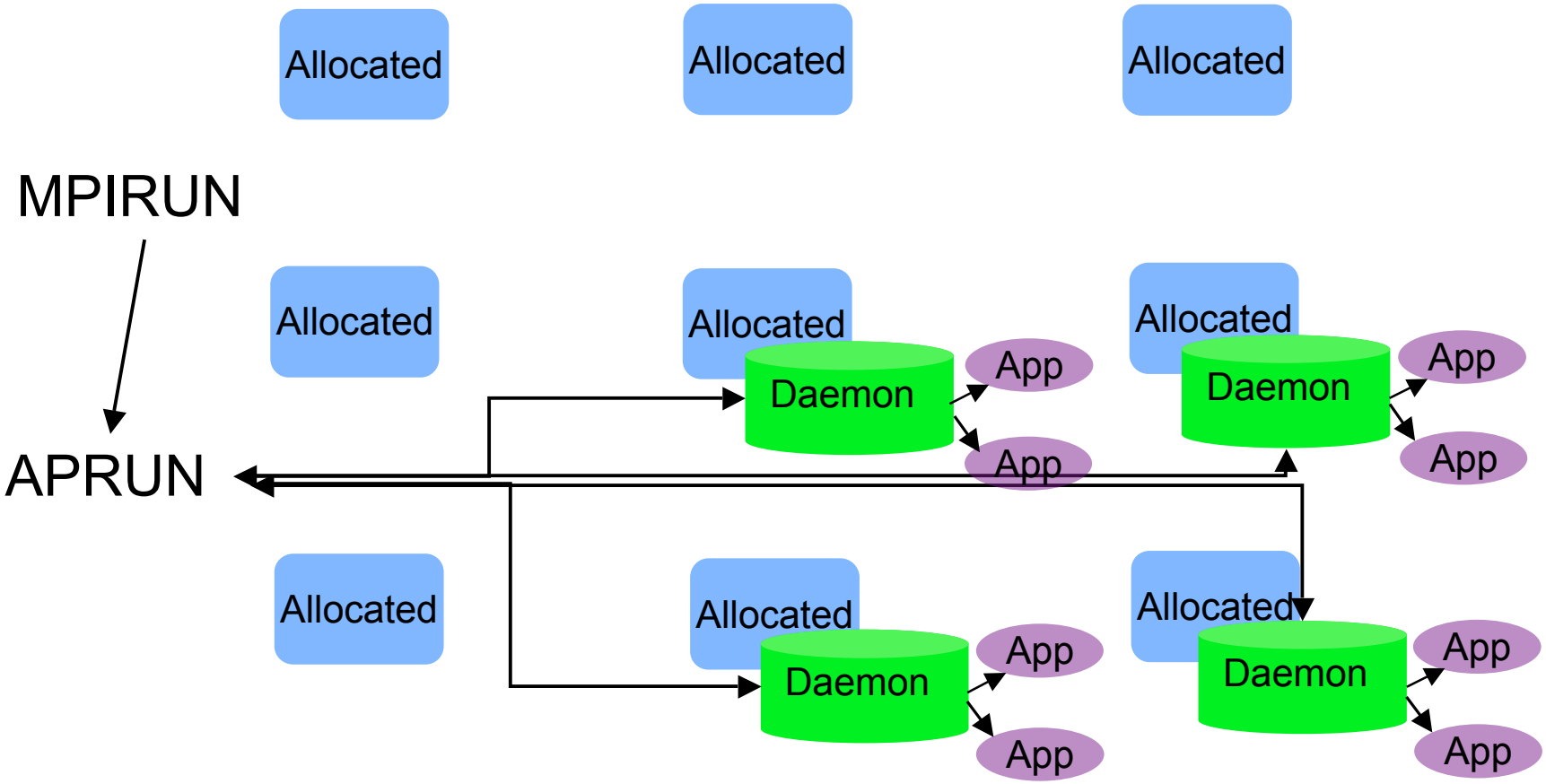
# Open MPI's CNL Port

- Portals port from Catamount to CNL
  - Enhance Point-to-Point BTL component
- ALPS support added
  - Add process control components for ALPS
    - mpirun wraps multiple calls to APRUN to
  - Support MPI-2 dynamic process control
  - Support for recovery from process failure
  - Support arbitrary number of procs per node (even over subscribe)
- Pick up full MPI 2.0 support

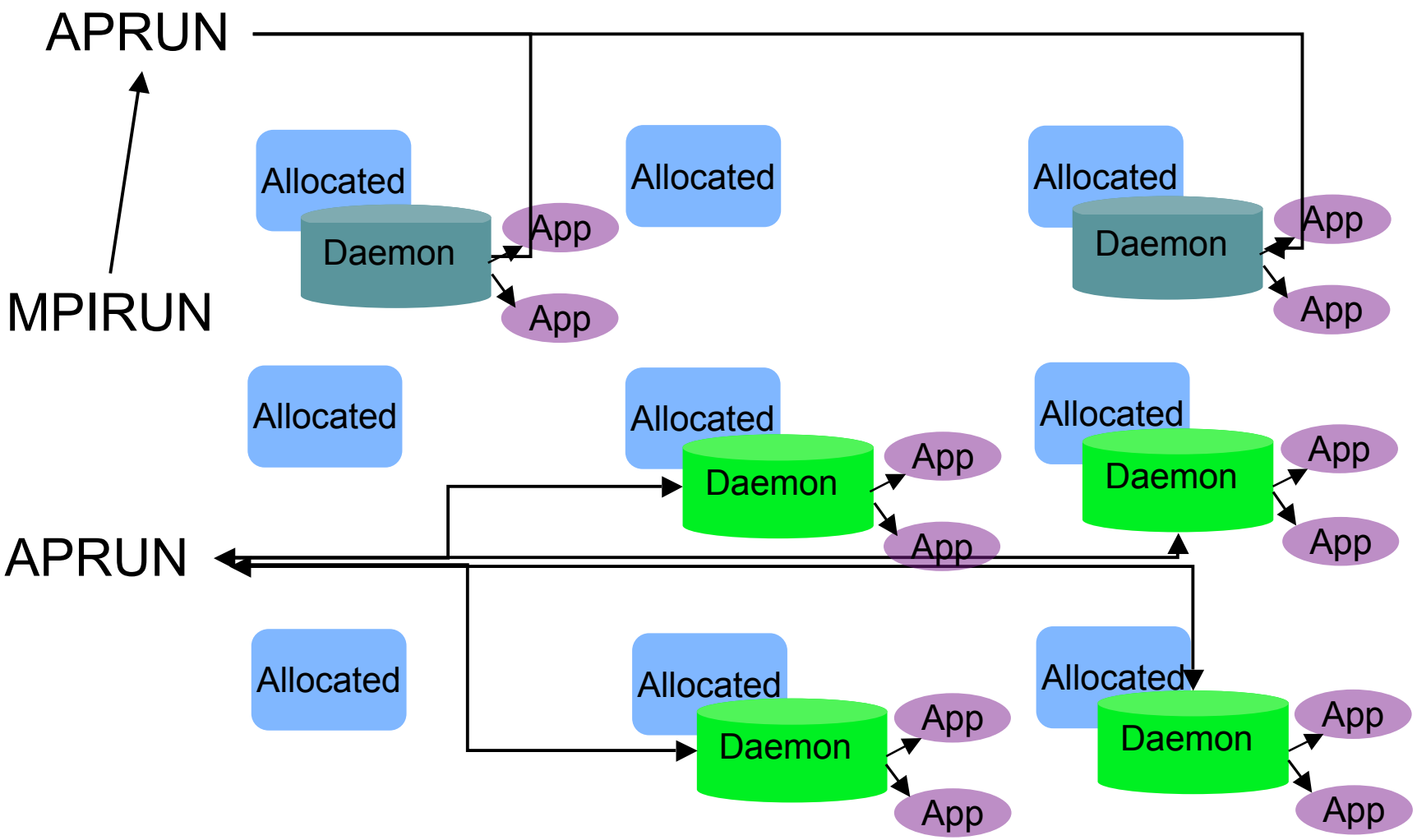
# Modular Component Architecture - Data Transfer



# Process Startup on CNL - Start



# Process Startup on CNL - Spawn



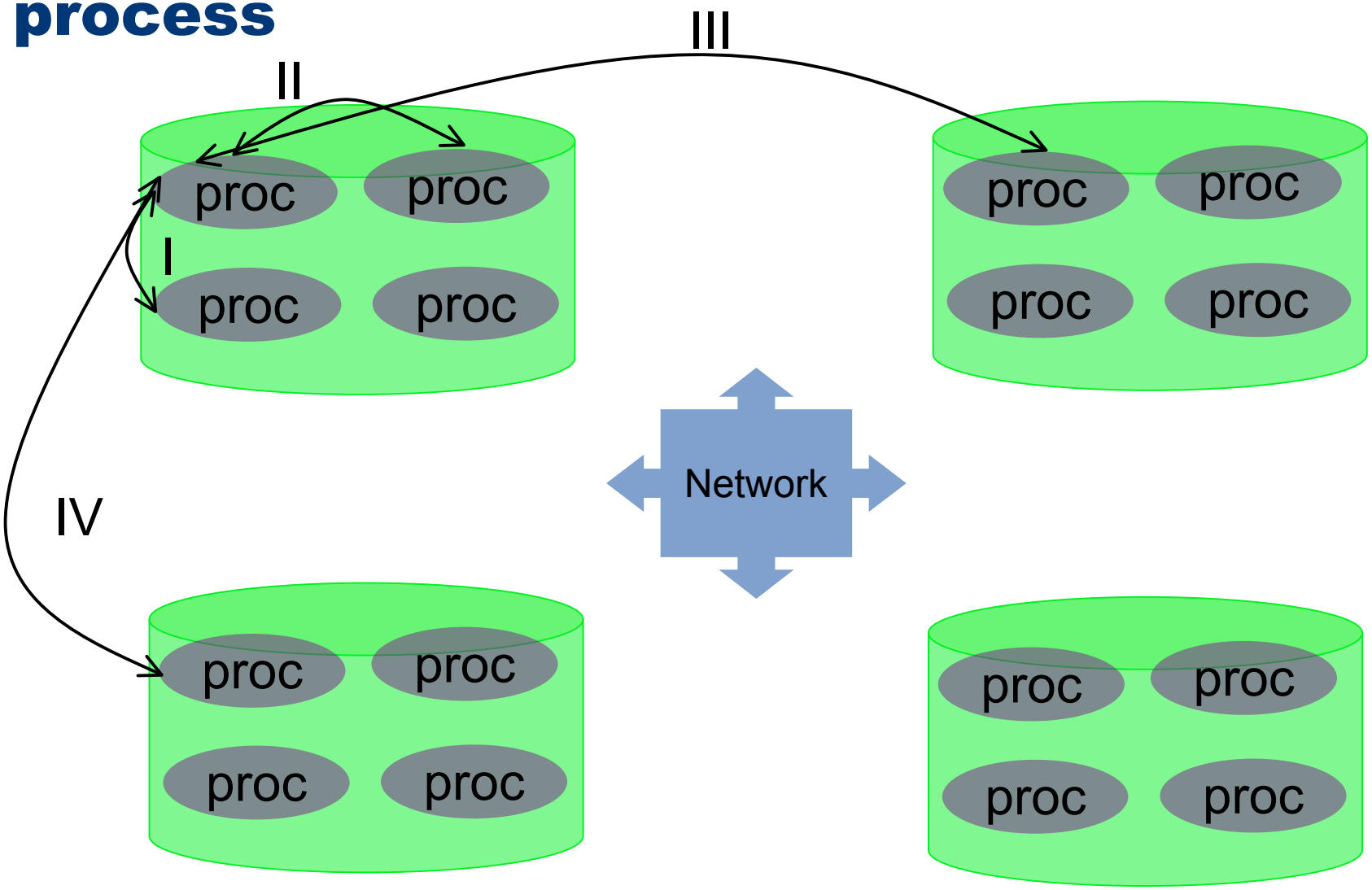
# Features in Open MPI for Multi-Core Support

- Shared Memory point-to-point communications
  - On par with other network devices
  - Does not use any network resources
- Shared Memory Collective optimizations
  - On-host-communicator optimization
  - Hierarchical collectives on the way

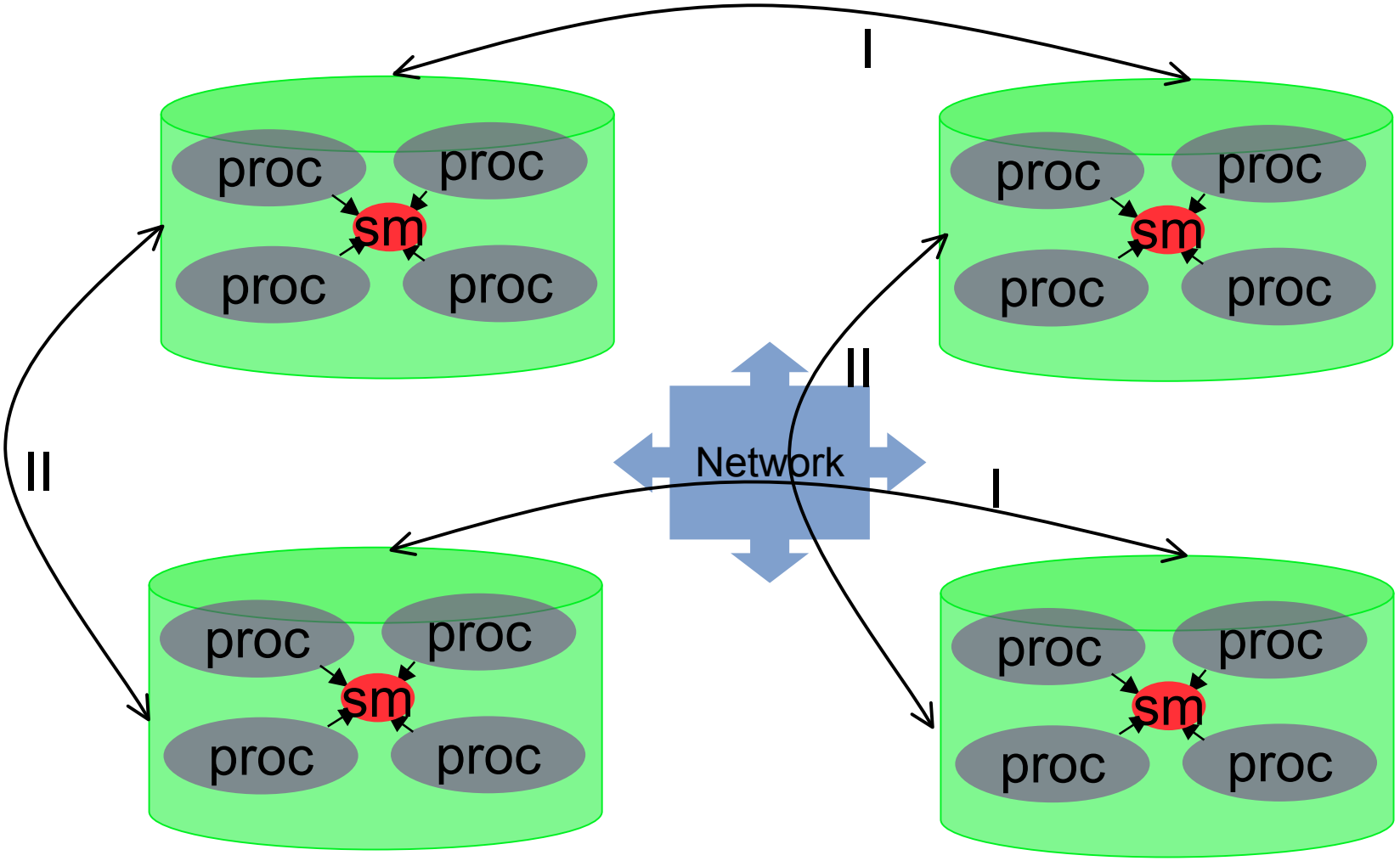
# Hierarchical Collectives

- Exist in the code base (HLRS/University of Houston)
- Need to be tested with the new shared-memory module
- Need to be optimized

# Collective Communication Pattern - per process



# Collective Communication Pattern - Total Interhost traffic



# Performance Data

# Ping-Pong 0 byte MPI latency : Inter-node

| MPI / Protocol | Latency (uSec) |
|---|---|
| Open MPI / CM | 6.18 |
| Open MPI / OB1 | 8.65 |
| Open MPI / OB1 - no ack | 7.24 |
| Cray MPT (3.0.7) | 7.44 |

# Ping-Pong 0 byte MPI latency

**CM**

0 Bytes - 6.18 uSec
16 Bytes - 6.88 uSec
17 Bytes - 9.69 uSec (measured on different system)

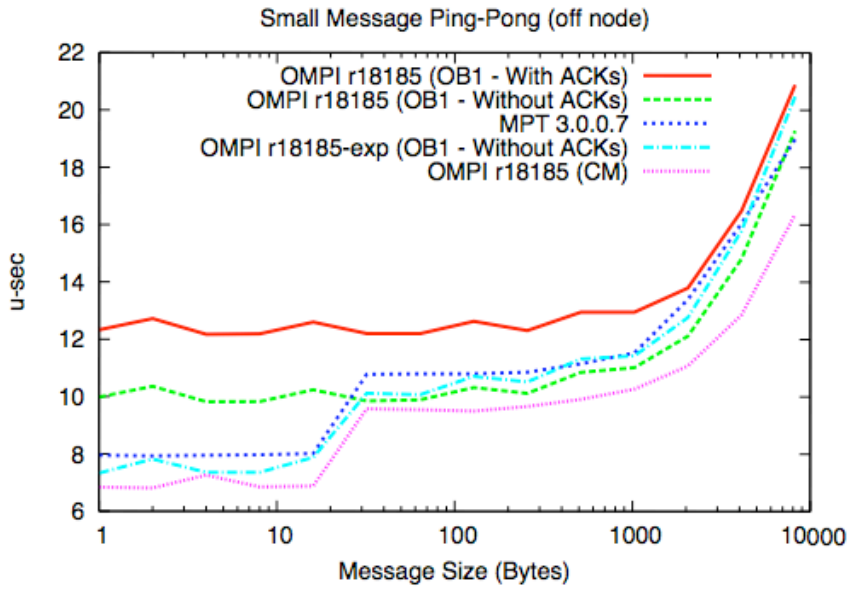
**OB1**

0 Bytes - With ACK: 8.65 uSec
0 Bytes - Without ACK: 7.24 uSec
1 Byte - Without ACK: 10.14 uSec (measured on different system)

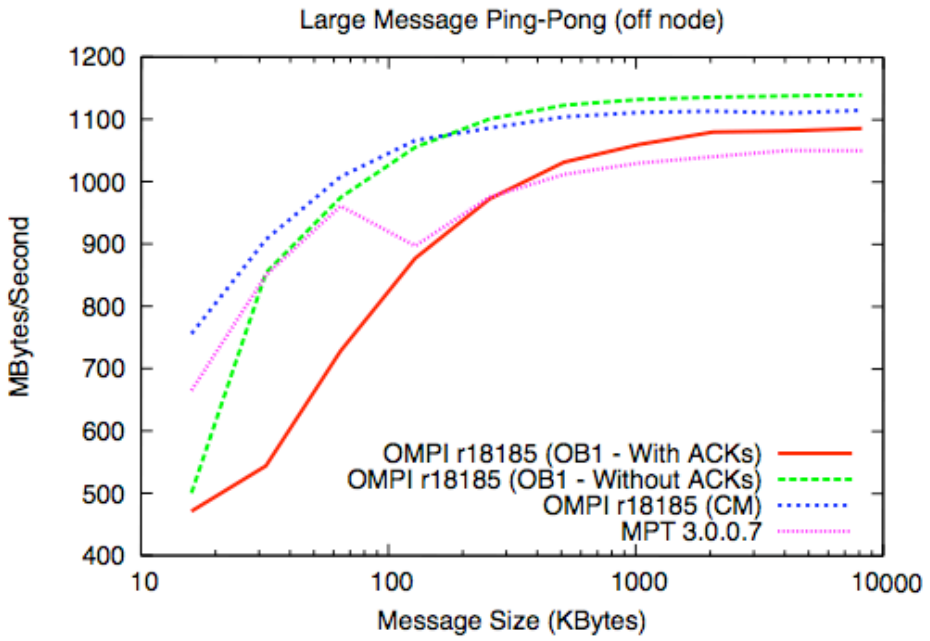
# Ping-Pong 0 byte MPI latency : Intra-node

| MPI / Protocol | Latency (uSec) |
|---|---|
| Open MPI / CM | |
| Open MPI / OB1 | 0.64 |
| Open MPI / OB1 - no ack | |
| Cray MPT (3.0.7) | 0.51 |

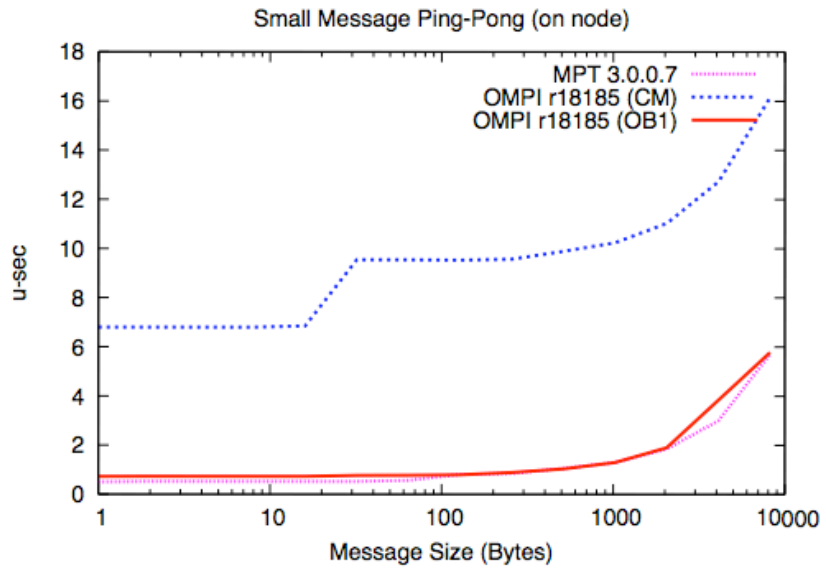
# Ping-Pong Latency Data - Off Host



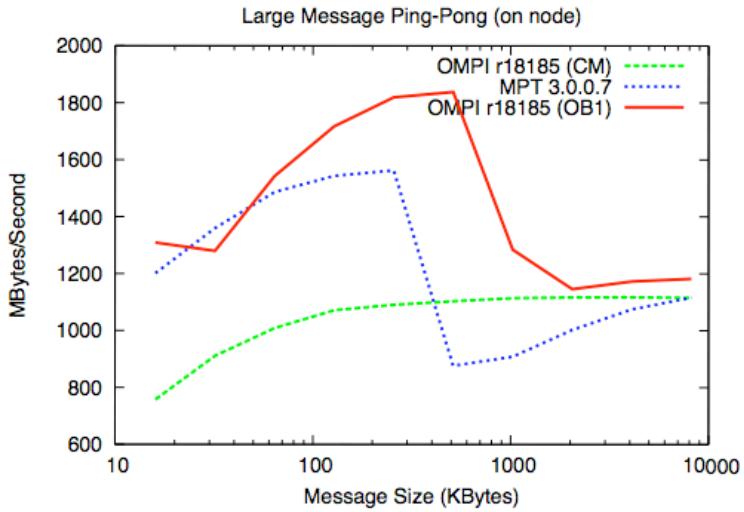
# Ping-Pong Data - Off Host



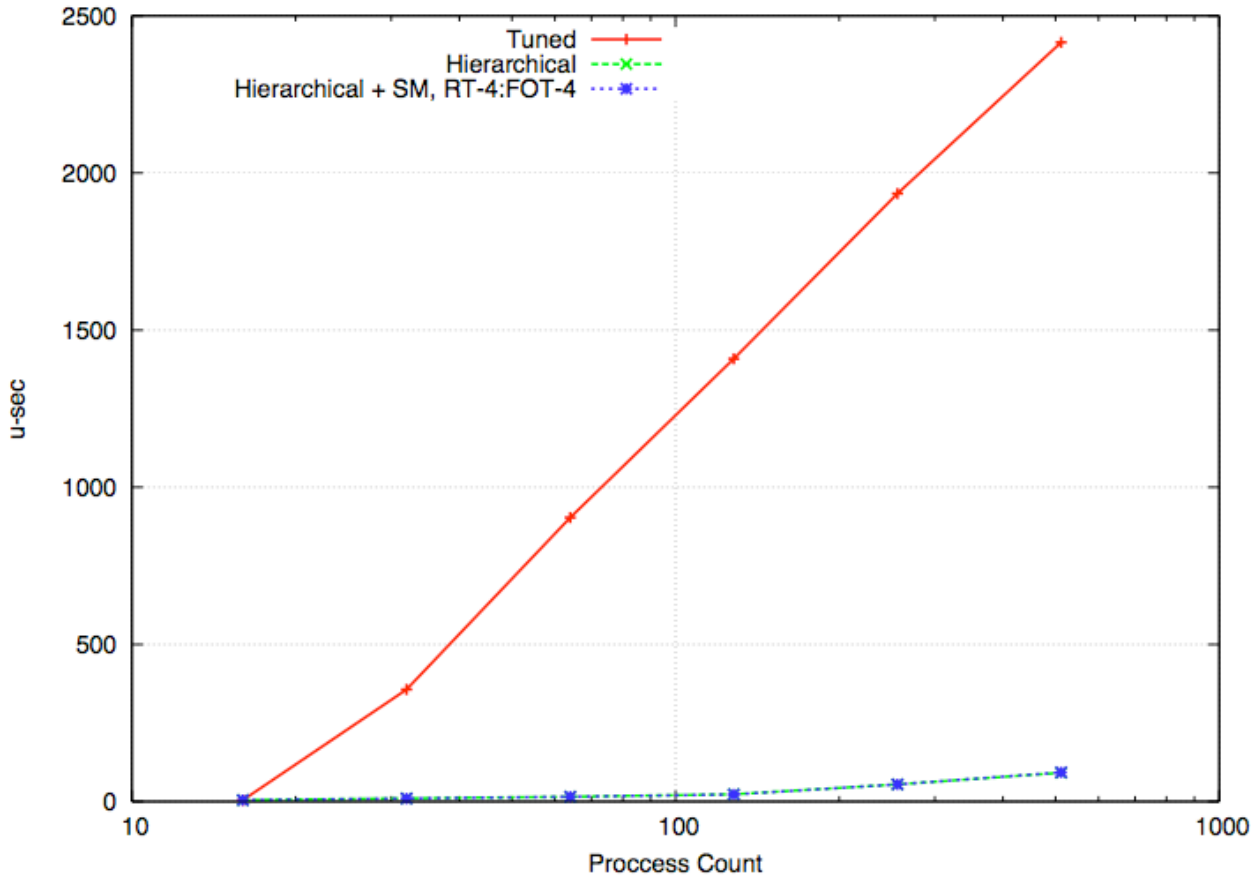
# Ping-Pong Data - On Host



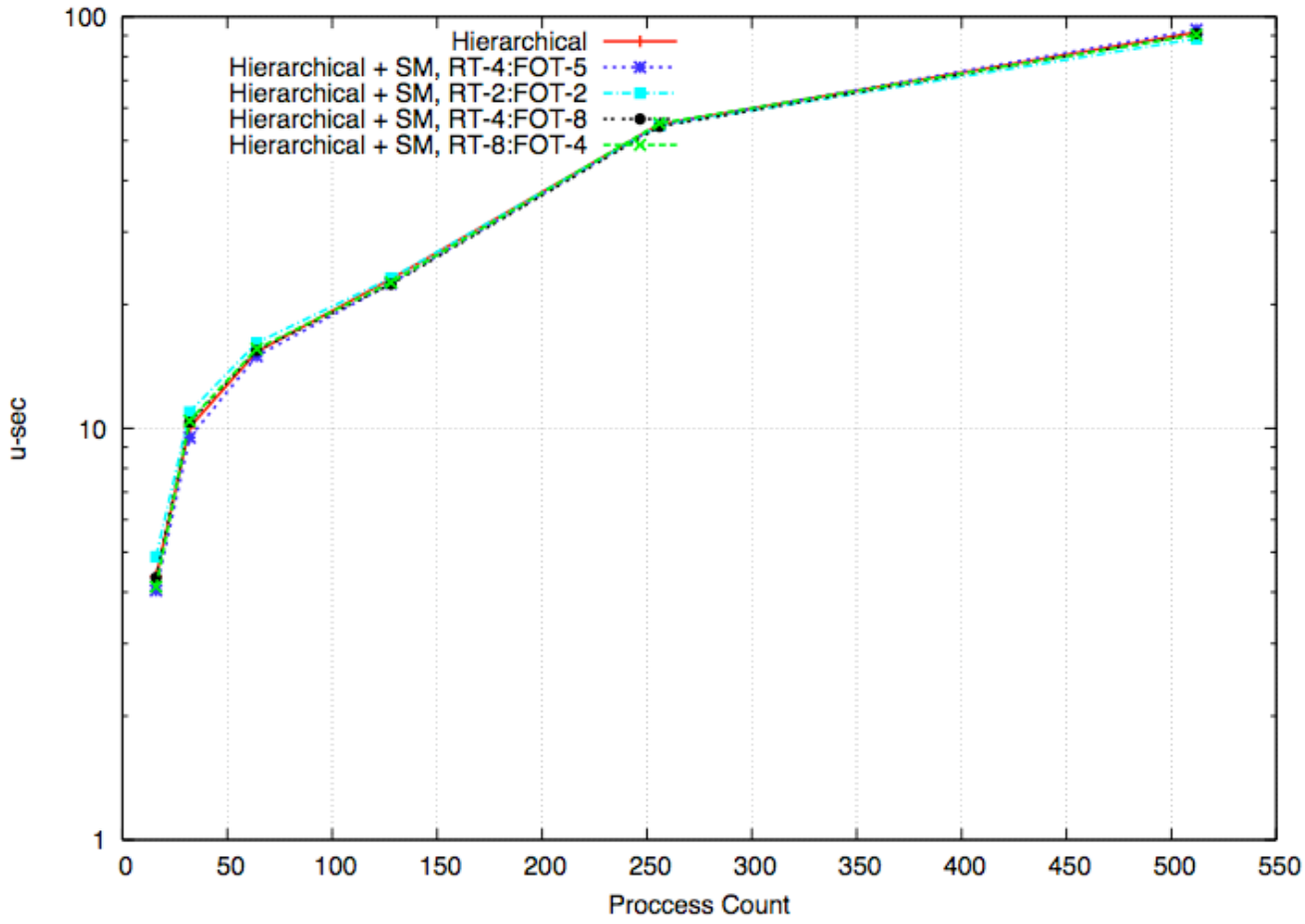
# Ping-Pong Bandwidth Data - On Host



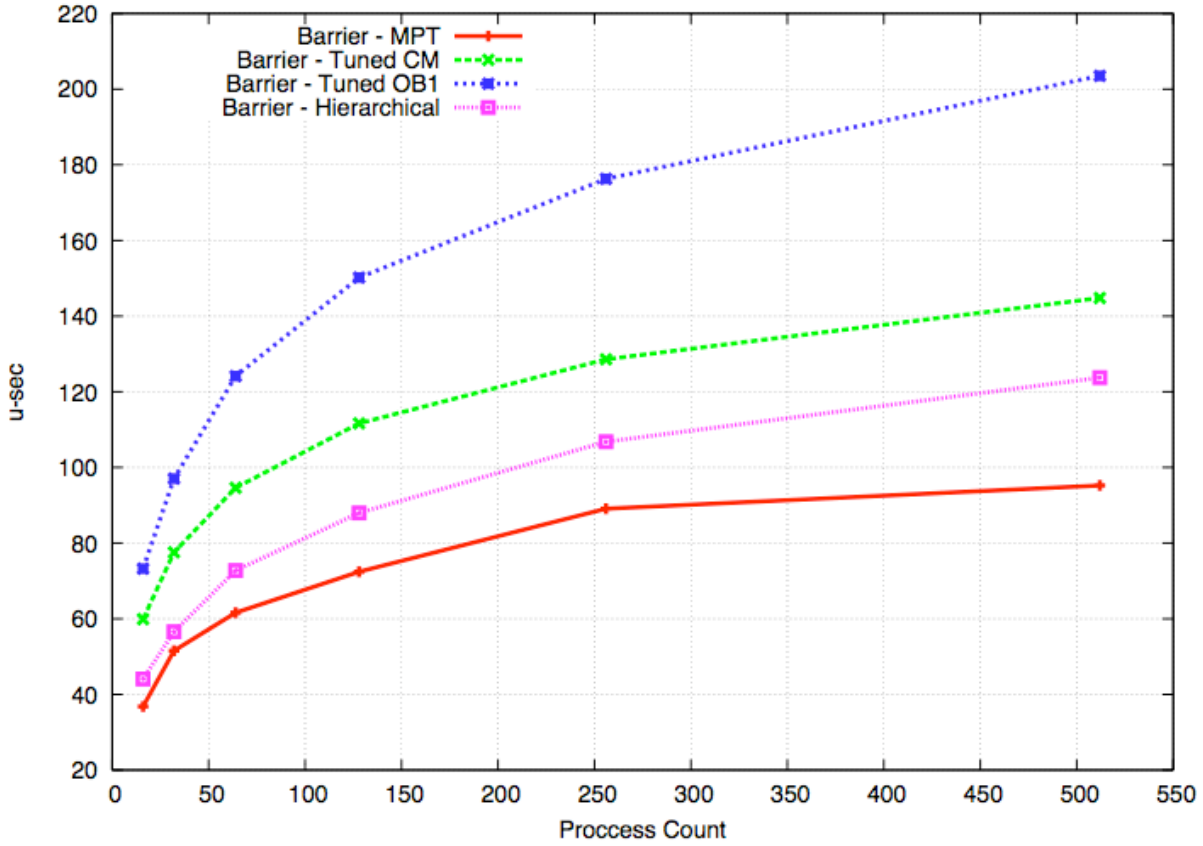
# Barrier - 16 cores per host



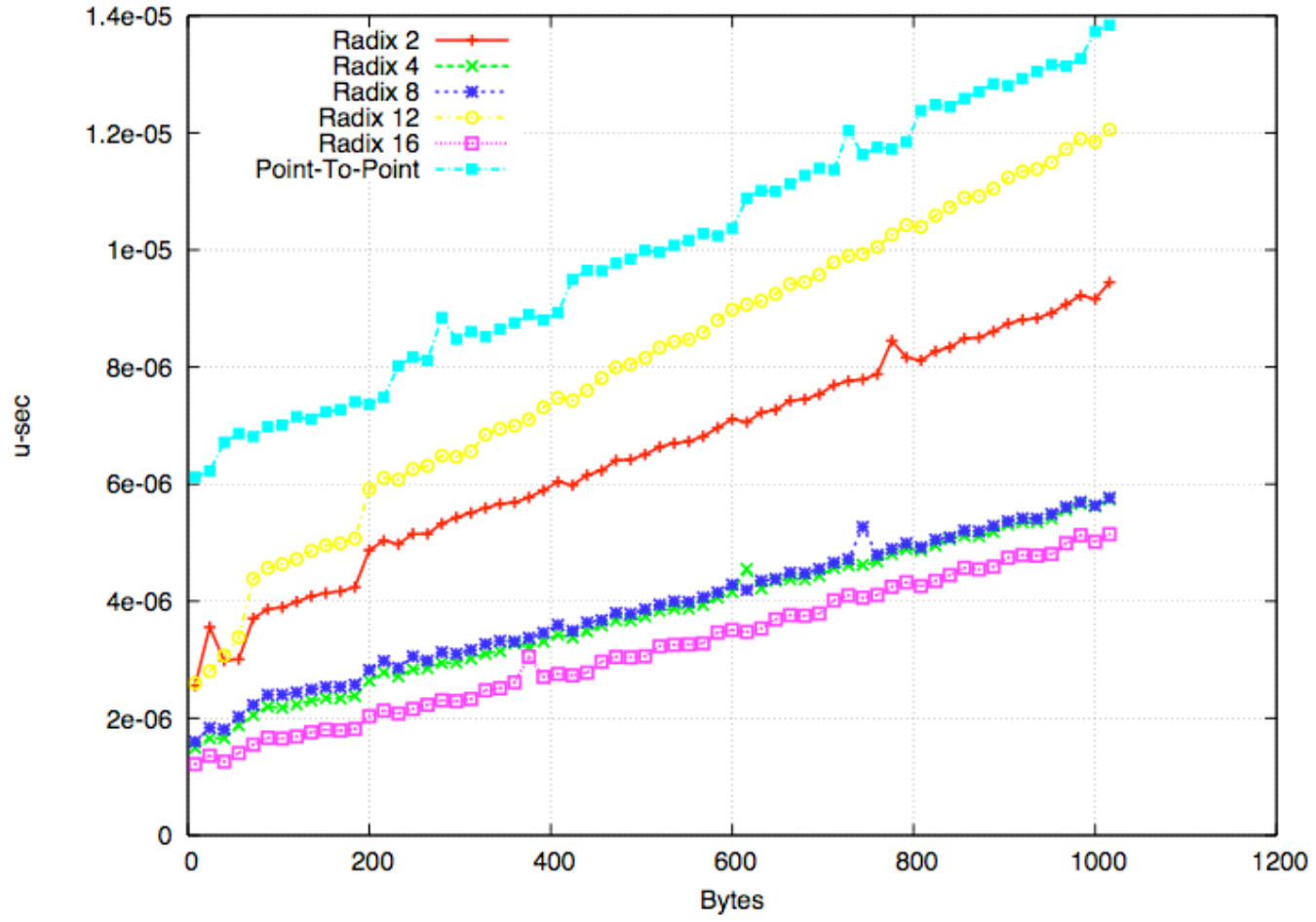
# Barrier - 16 cores per host - Hierarchical



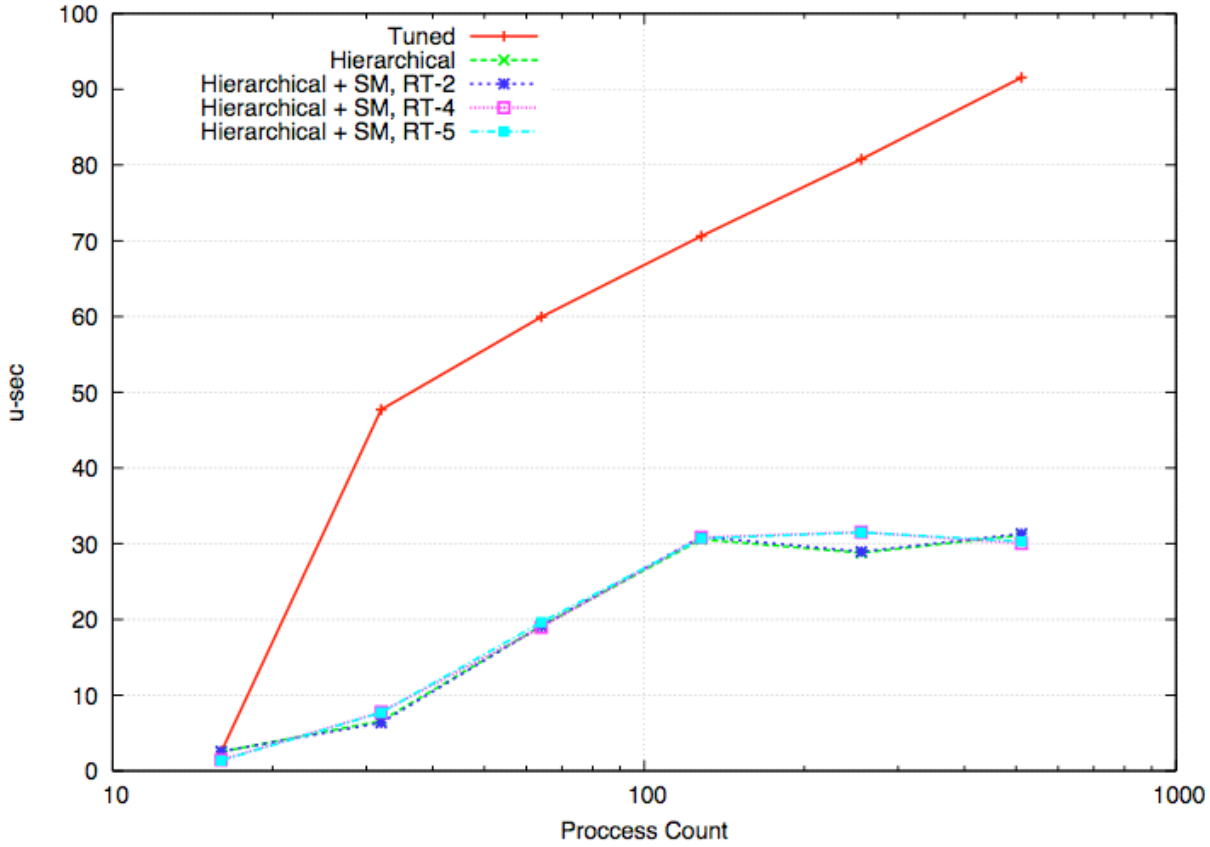
# Barrier - XT


Barrier - MPT
Barrier - Tuned CM
Barrier - Tuned OB1
Barrier - Hierarchical
u-sec
Proccess Count
220
200
180
160
140
120
100
80
60
40
20
0
50
100
150
200
250
300
350
400
450
500
550

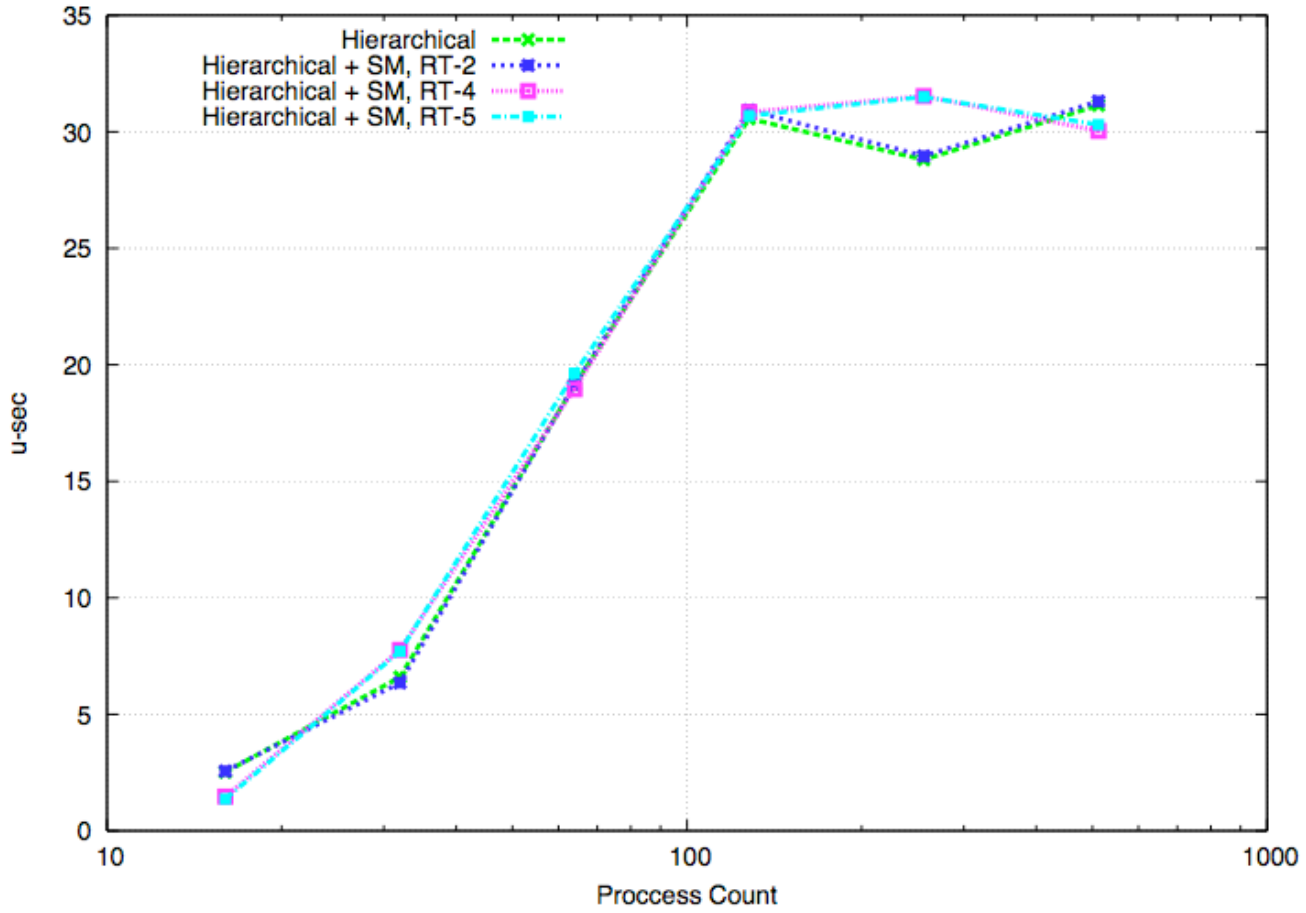
# Shared-Memory Reduction - 16 processes



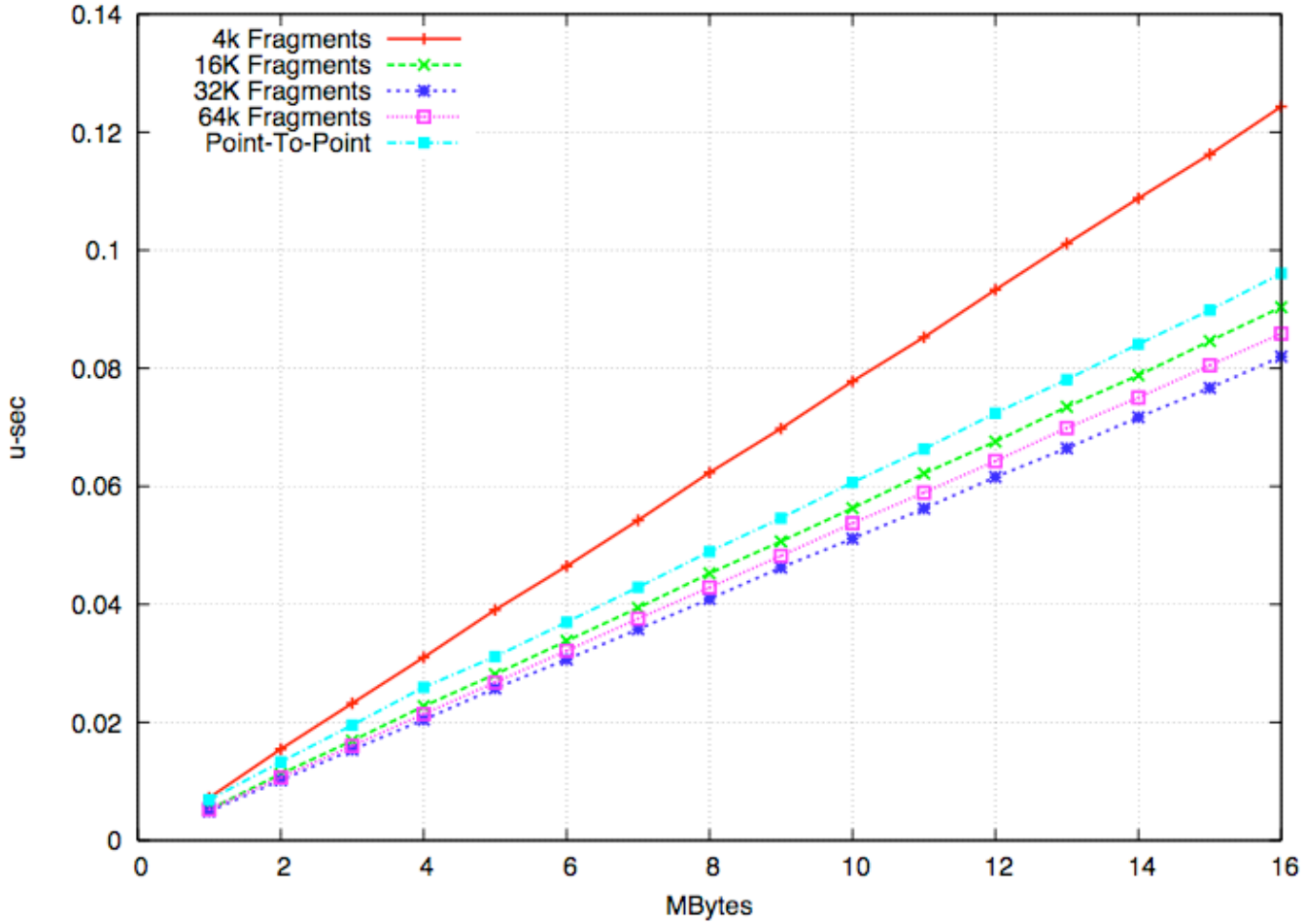
# Reduction - 16 core nodes - 8 Bytes



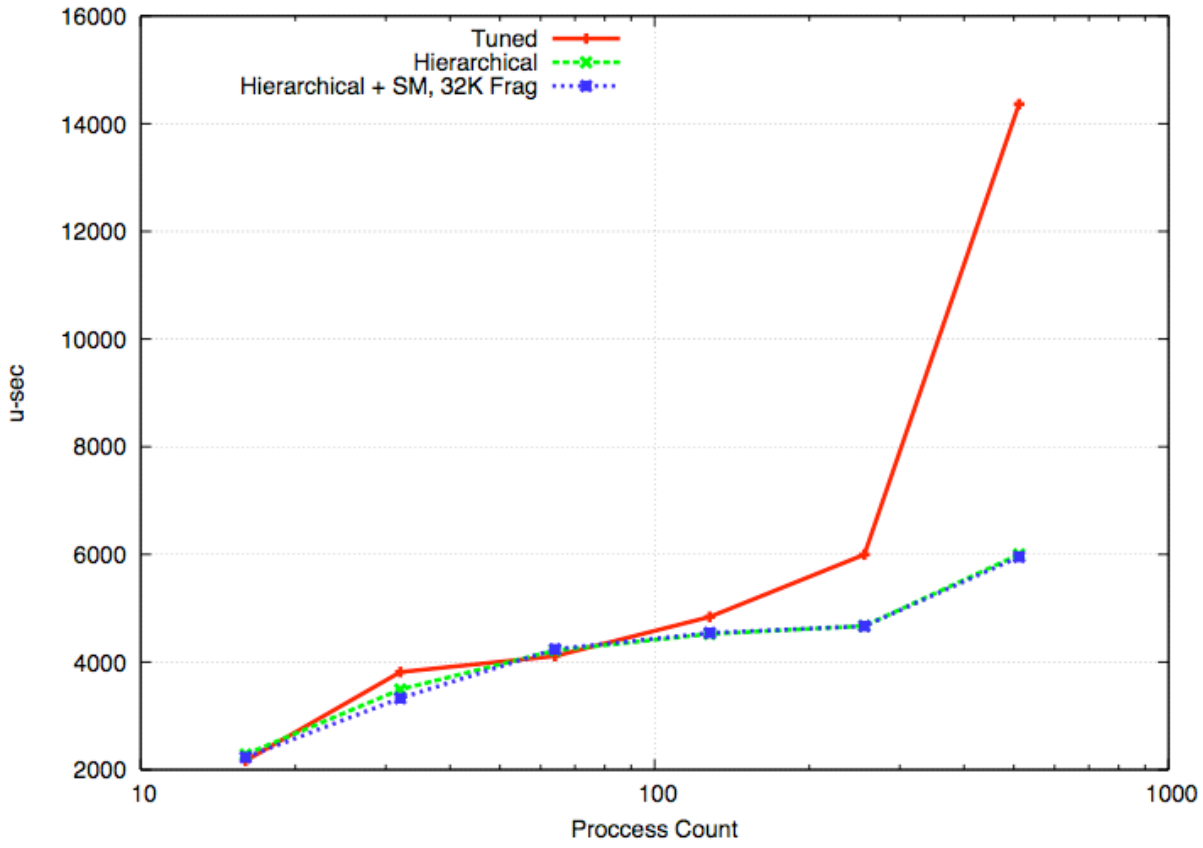
# Reduction - 16 core nodes - 8 Bytes - Hierarchical



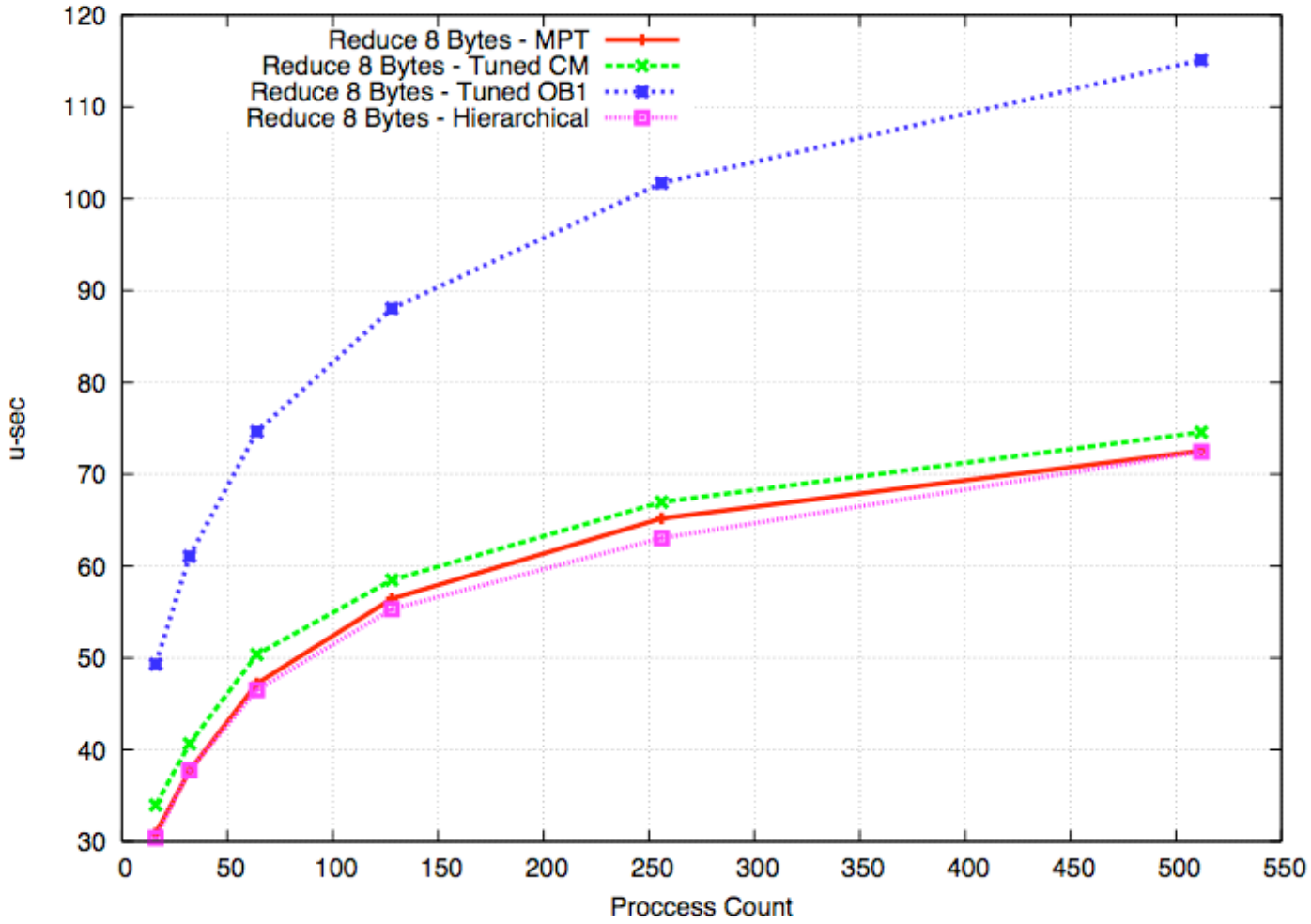
# Shared-Memory Reduction - 16 Processes



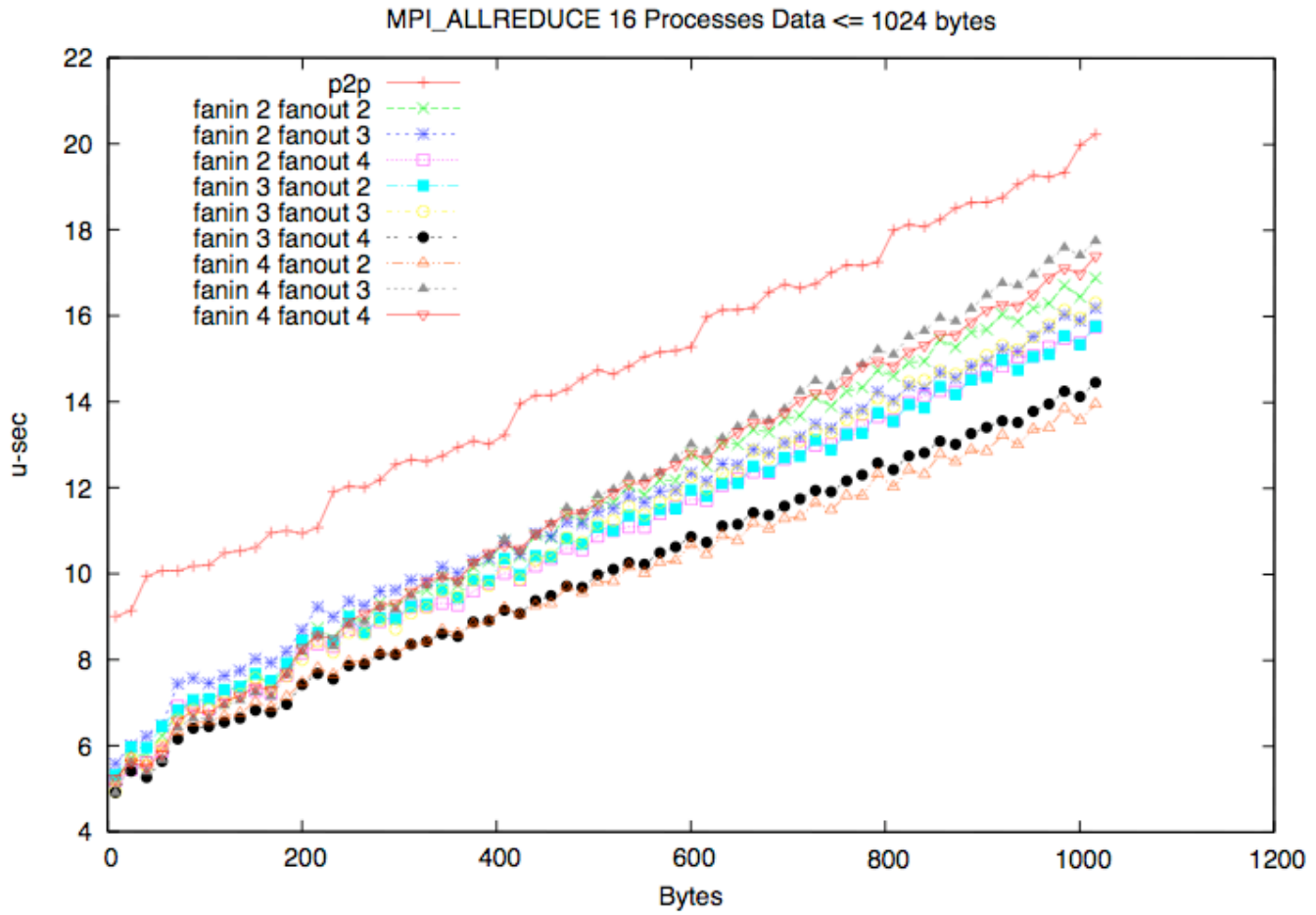
# Reduction - 16 core nodes - 512 KBytes



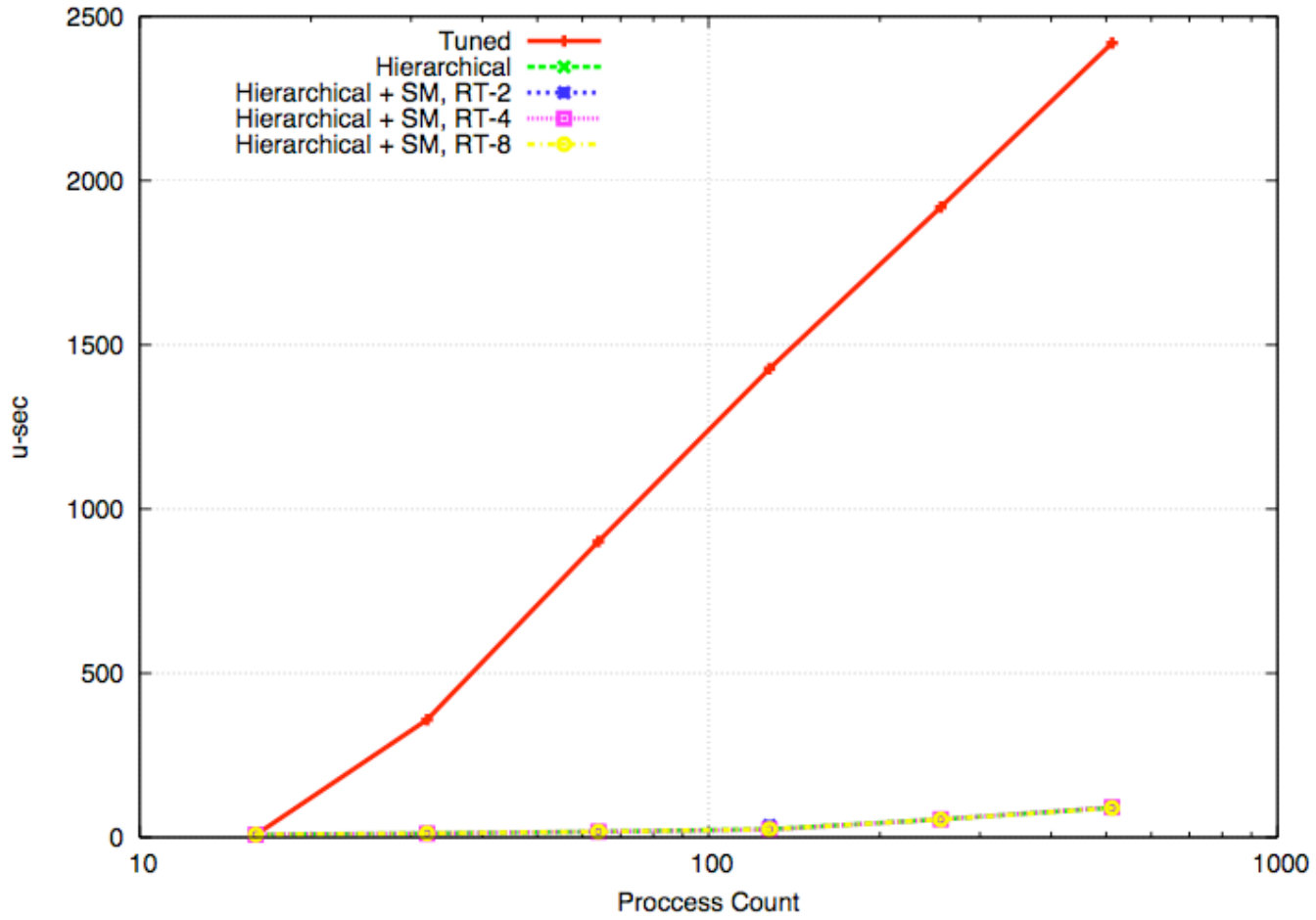
# Reduction - XT



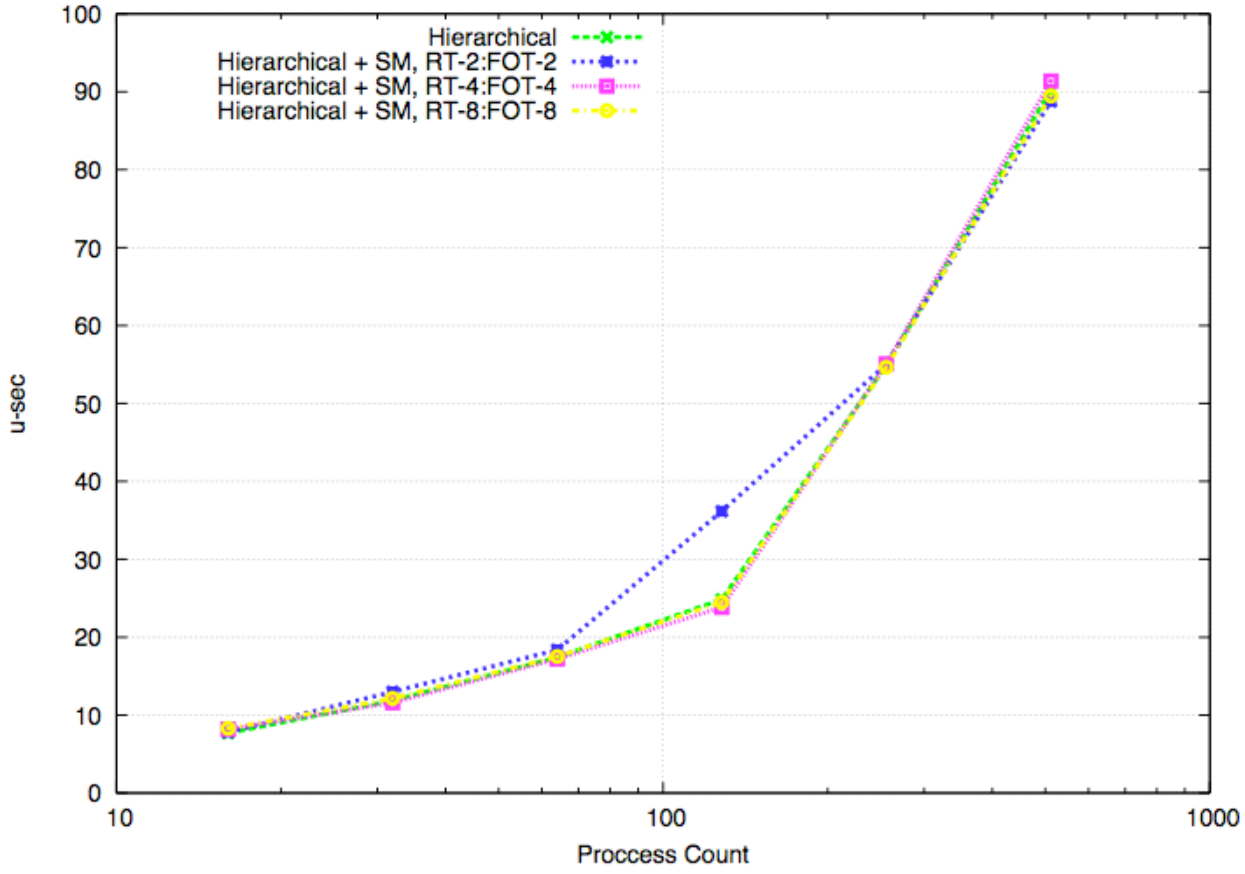
# Shared Memory Allreduce - 16 processes



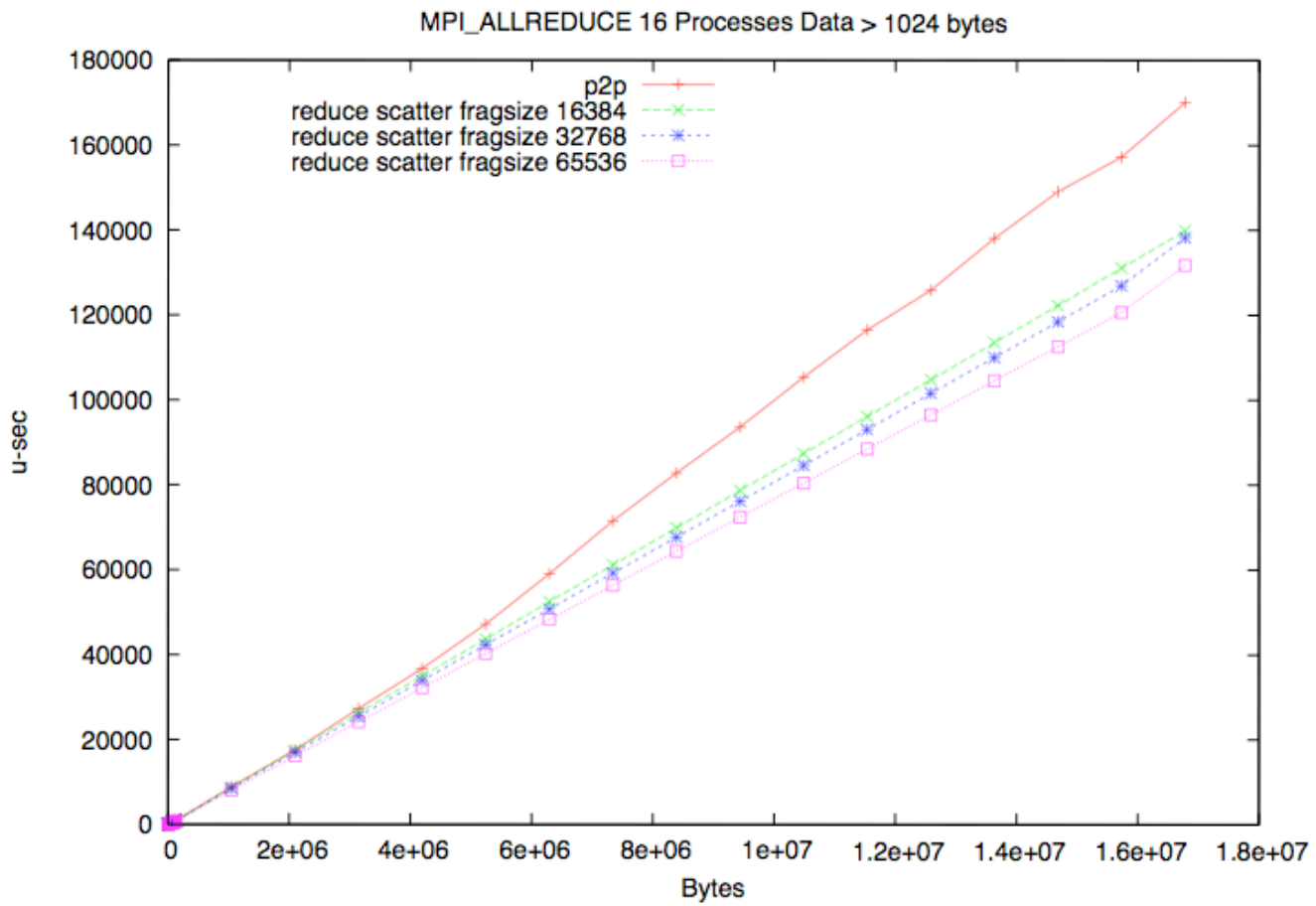
# Allreduce - 16 cores per node - 8 Bytes



# Allreduce - 16 cores per node - 8 Bytes - Hierarchical



# Shared Memory Allreduce - 16 processes



# Allreduce - XT