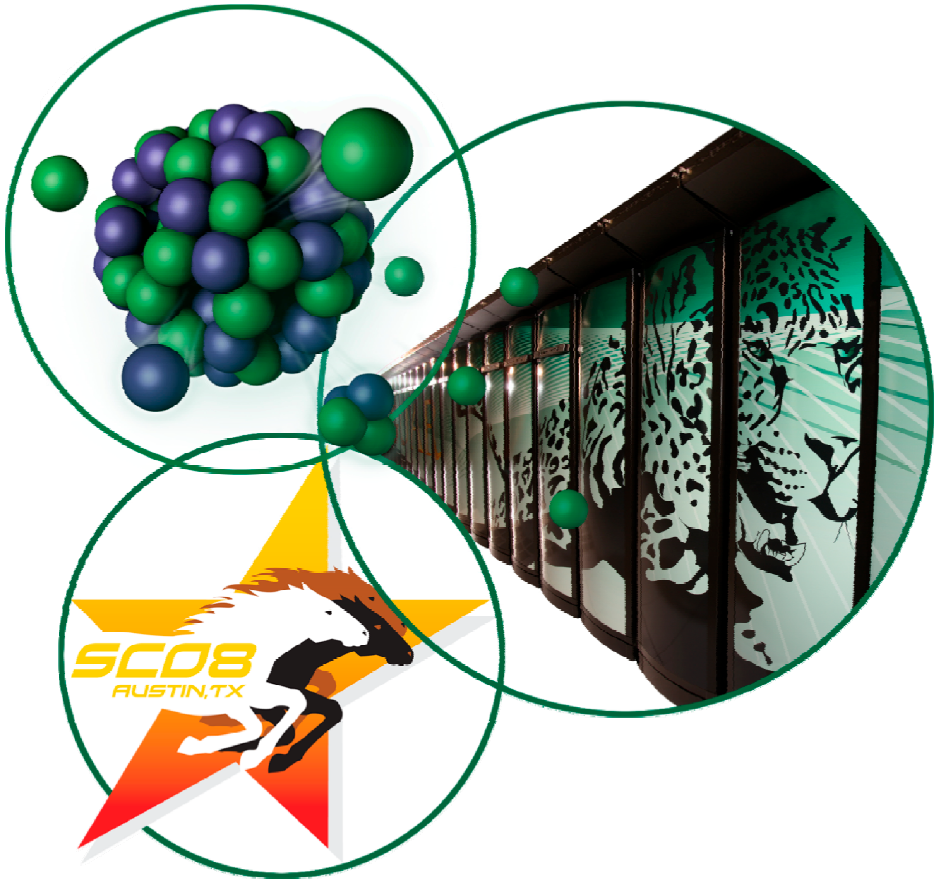
Open MPI on the Cray XT



Presented by

Richard Graham
Galen Shipman

Oak Ridge National Laboratory

Why does Open MPI exist?

- Maximize all MPI expertise:
  - Research/academia
  - Industry
  - Elsewhere
- Capitalize on (literally) years of MPI research and implementation experience
- The sum is greater than the parts

Current membership

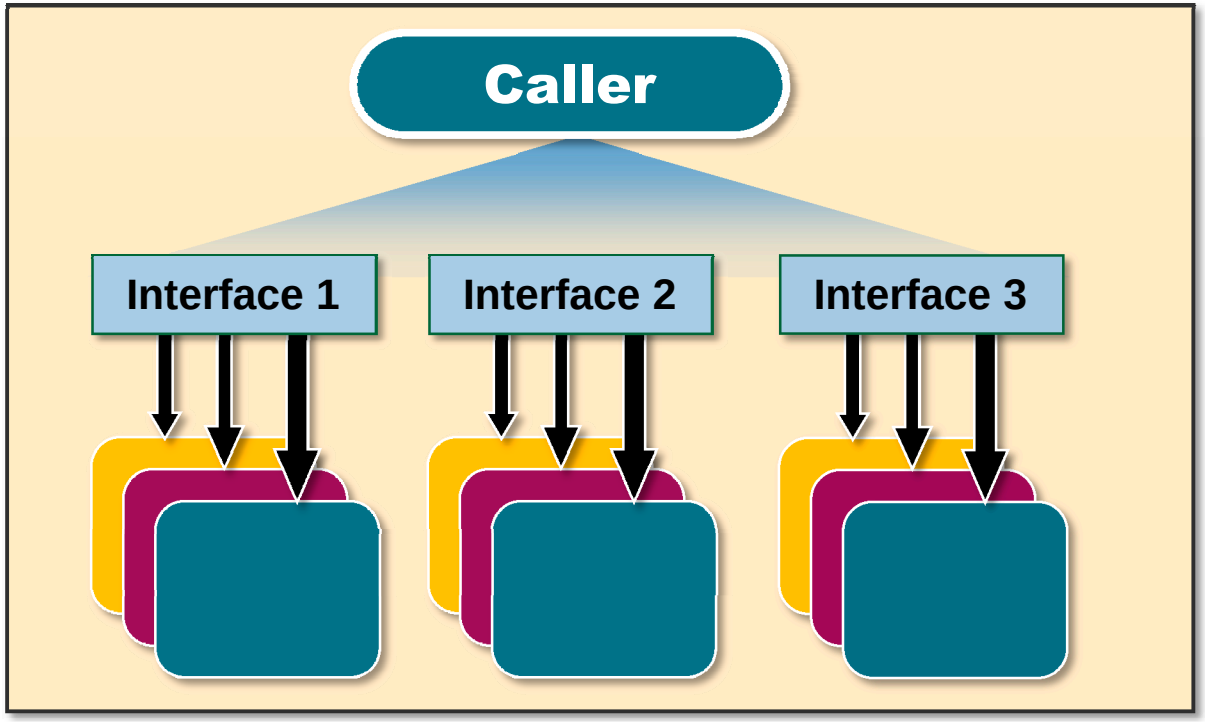
14 members, 9 contributors, 1 partner

- 4 US DOE labs
- 8 universities
- 10 vendors
- 1 individual

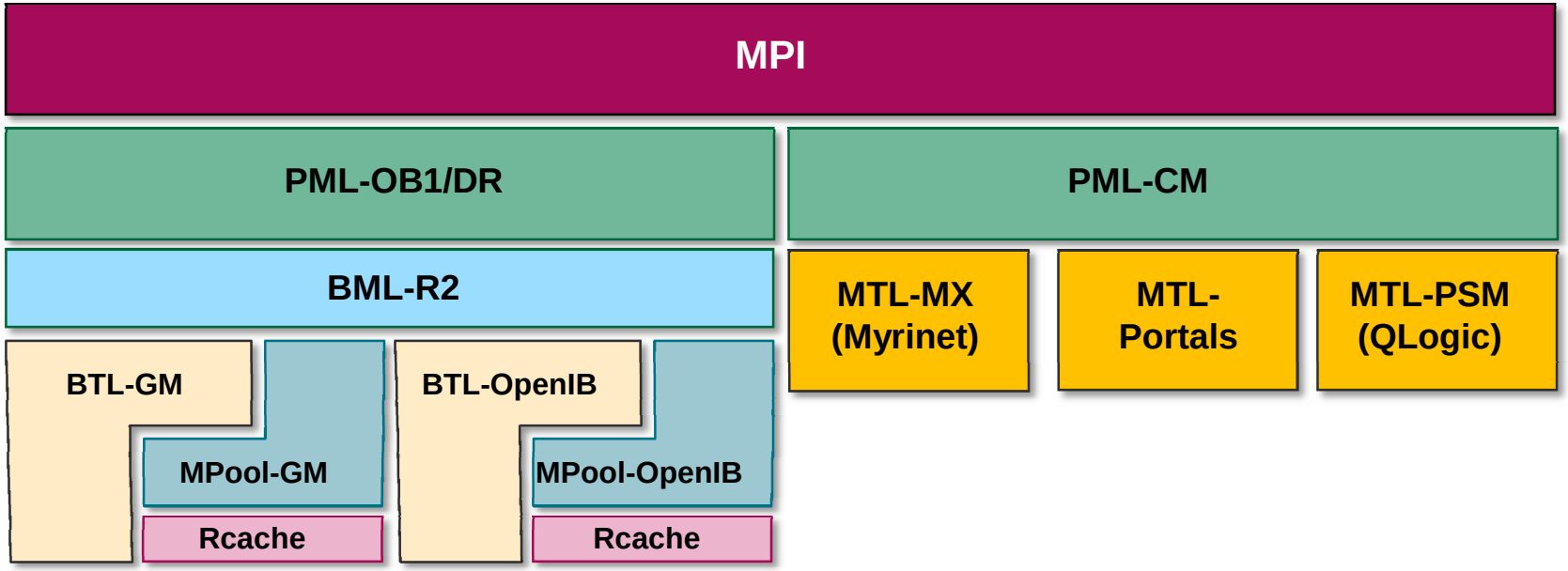
Key design feature: Components

Formalized interfaces

- Specifies “black box” implementation
- Different implementations available at runtime
- Can compose different systems on the fly



Point-to-point architecture



Portals port: OB1 vs. CM

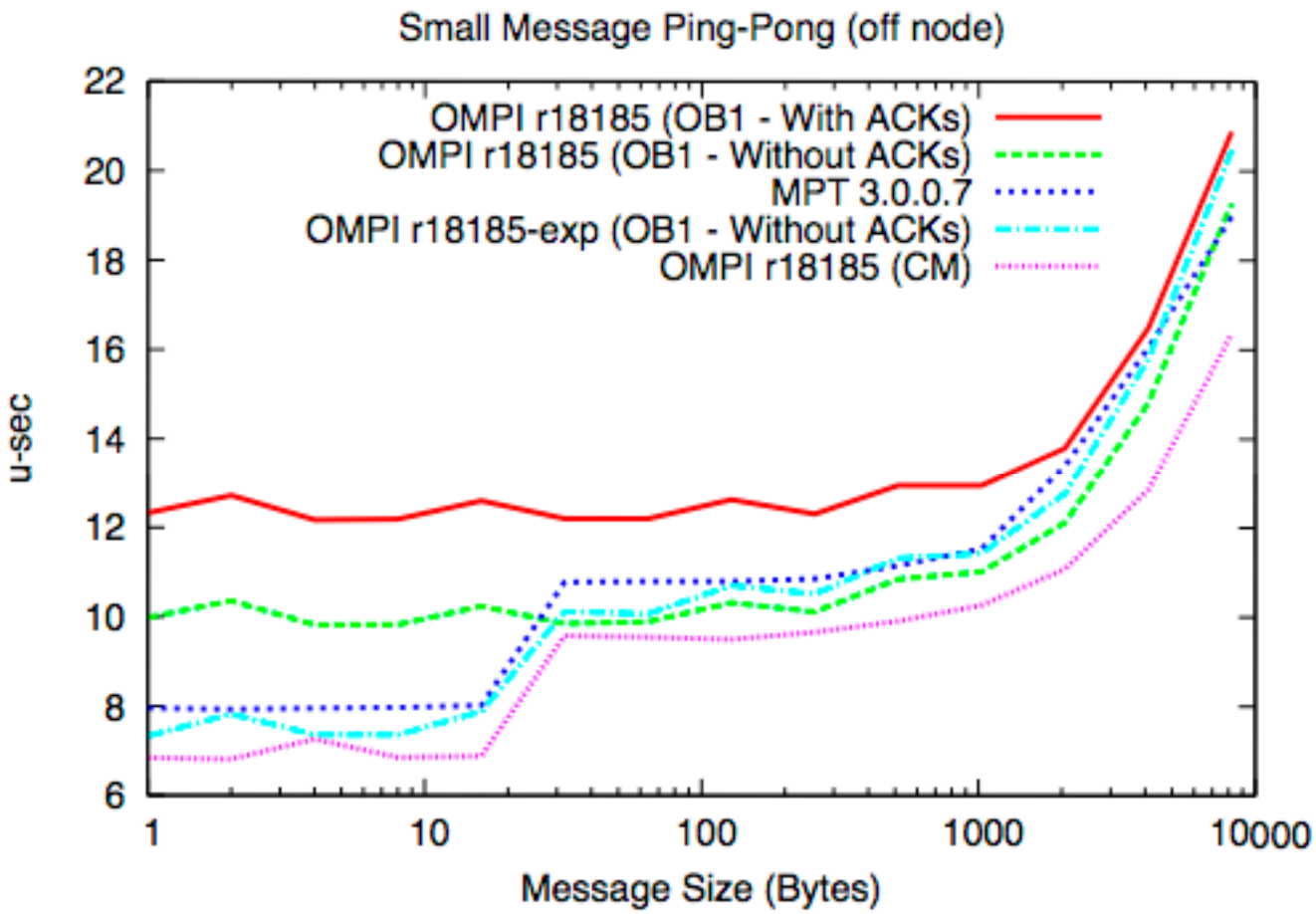
OB1

- Matching in main-memory
- Short message: eager, buffer on receive
- Long message: rendezvous
  - Rendezvous packet: 0 byte payload
  - Get message after match

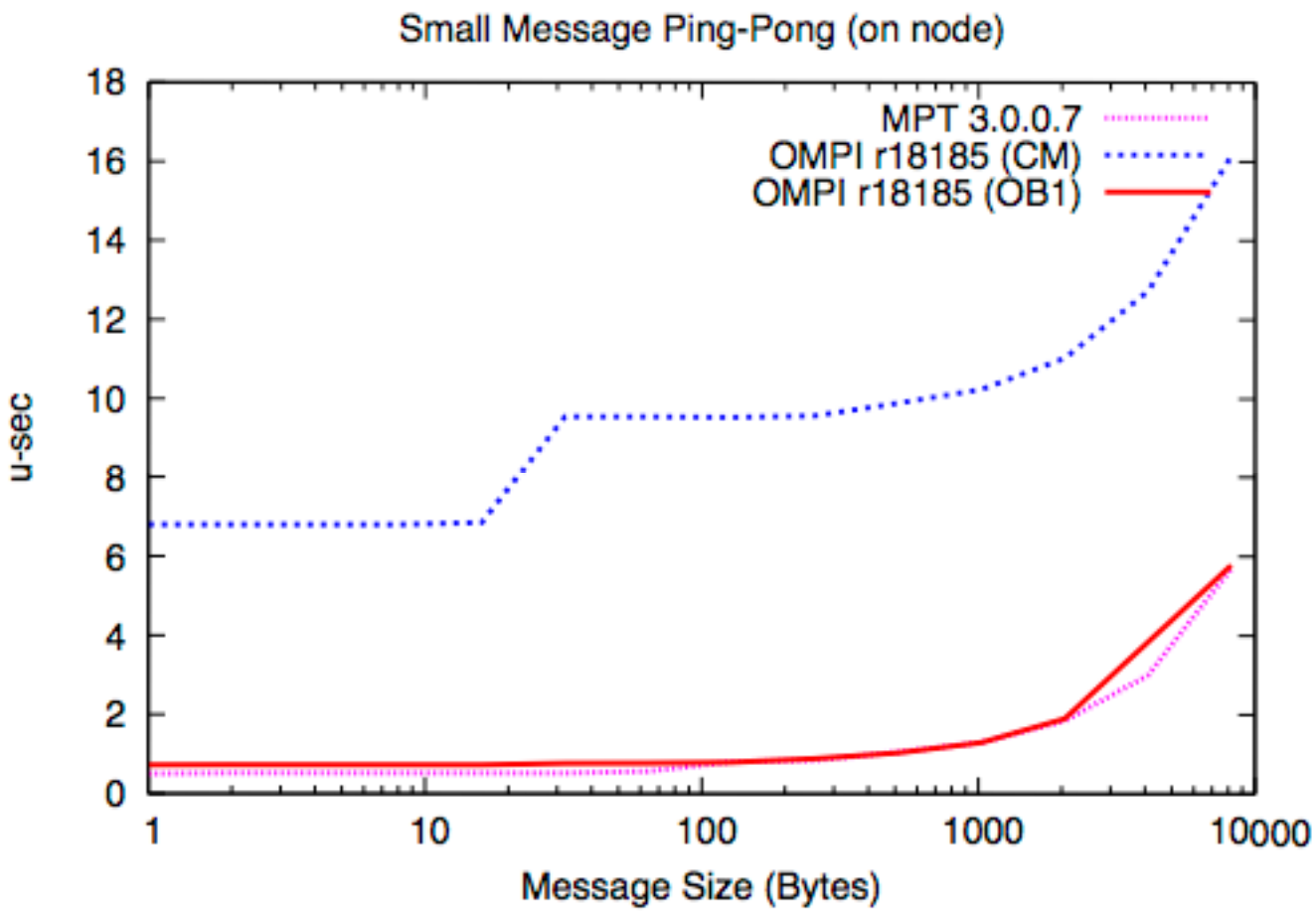
CM

- Matching maybe on NIC
- Short message: eager, buffer on receive
- Long message: eager
  - Send all data
    - If Match: deliver directly to user buffer
    - No Match: discard payload, and get() user data after match

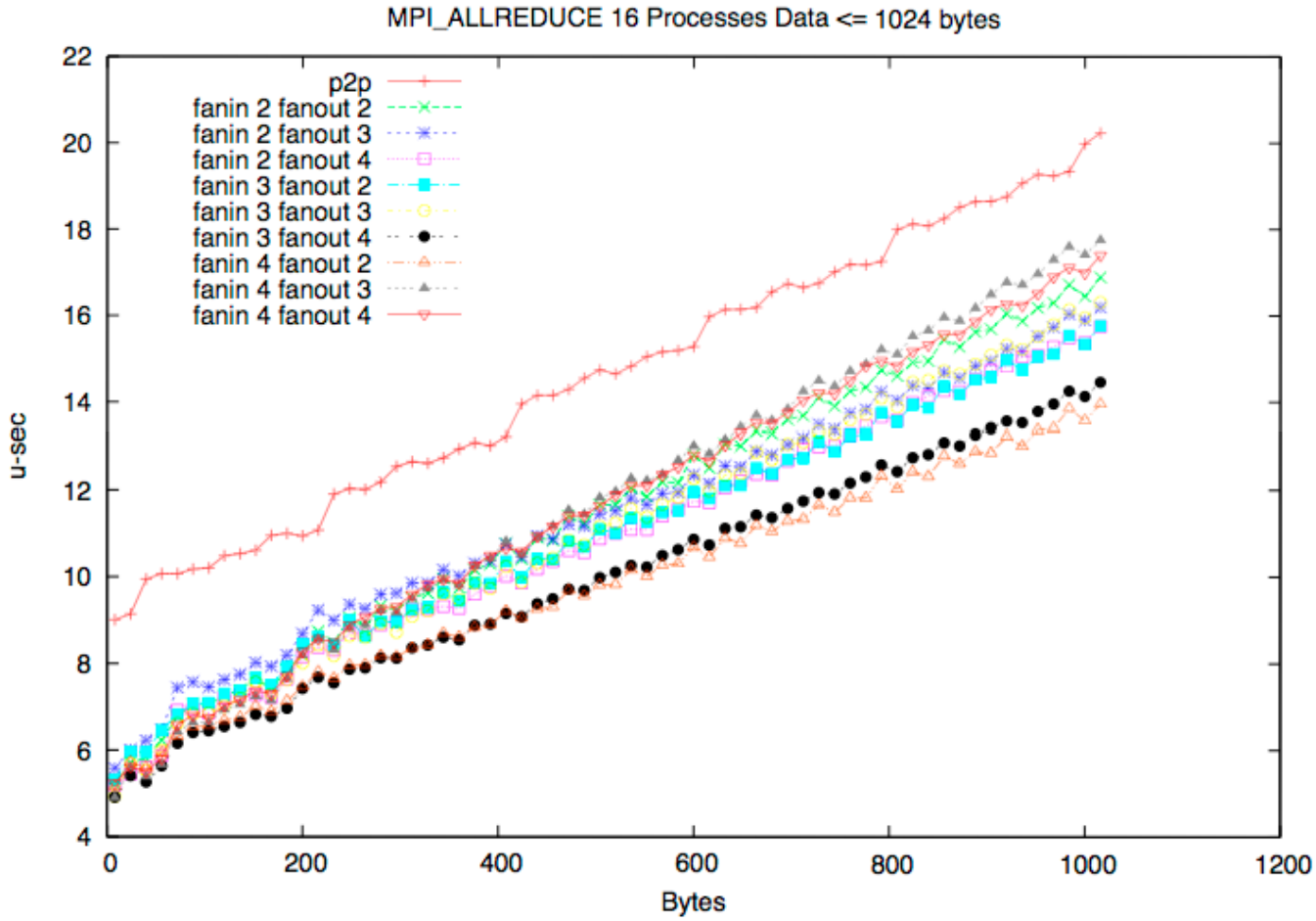
Ping-pong latency data—off Host



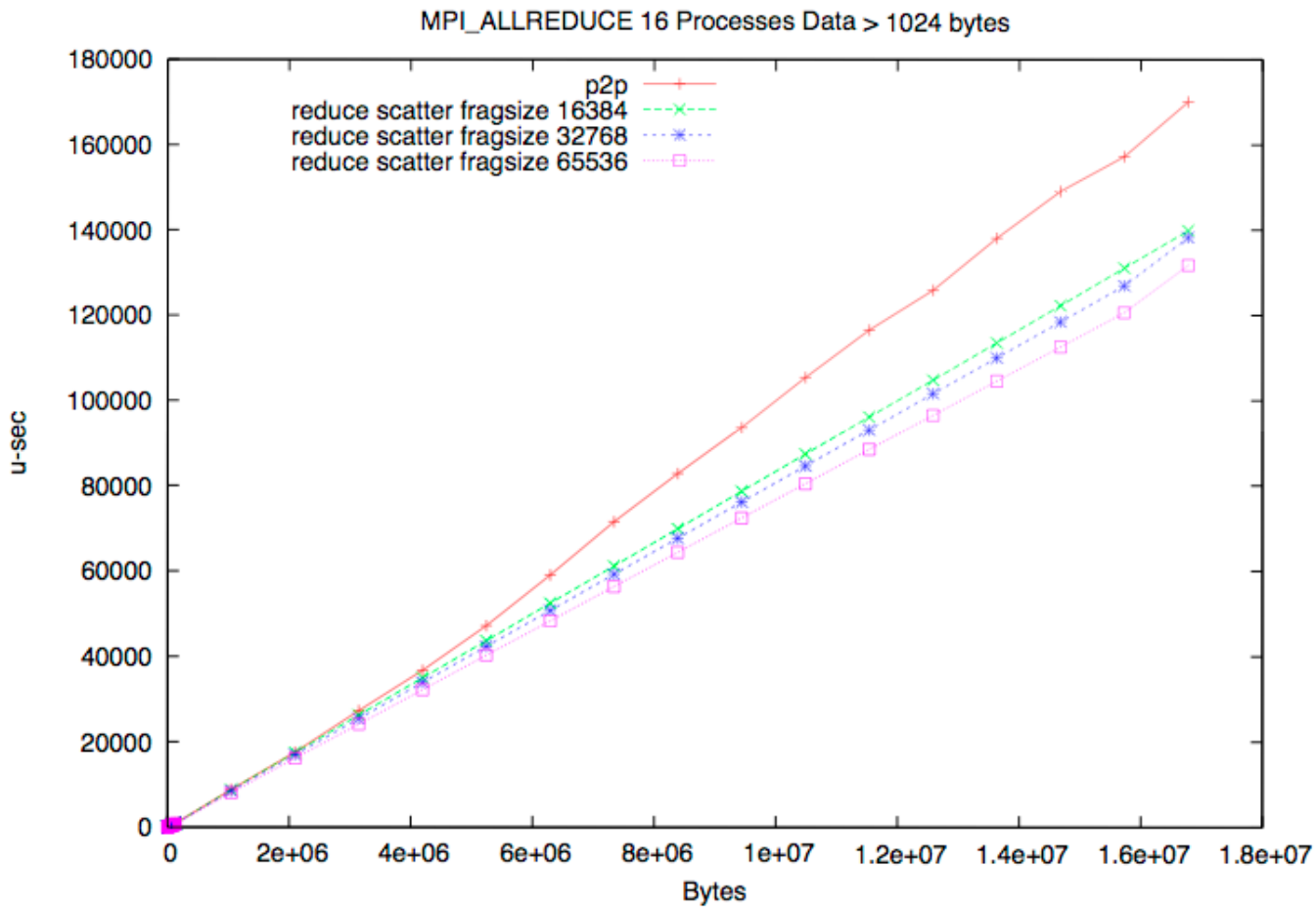
Ping-pong data—on host



Collective optimizations: Shared memory allreduce–16 processes



Collective optimizations: Shared memory allreduce—16 processes



Contact

Richard L. Graham

Application Performance Tools
Computer Science and Mathematics Division
(865) 356-3469
rlgraham@ornl.gov