A Comparison of Application Performance Using Open MPI and Cray MPI

LEADERSHIP COMPUTING FACILITY
NATIONAL CENTER FOR COMPUTATIONAL SCIENCES

presented by

Richard L. Graham

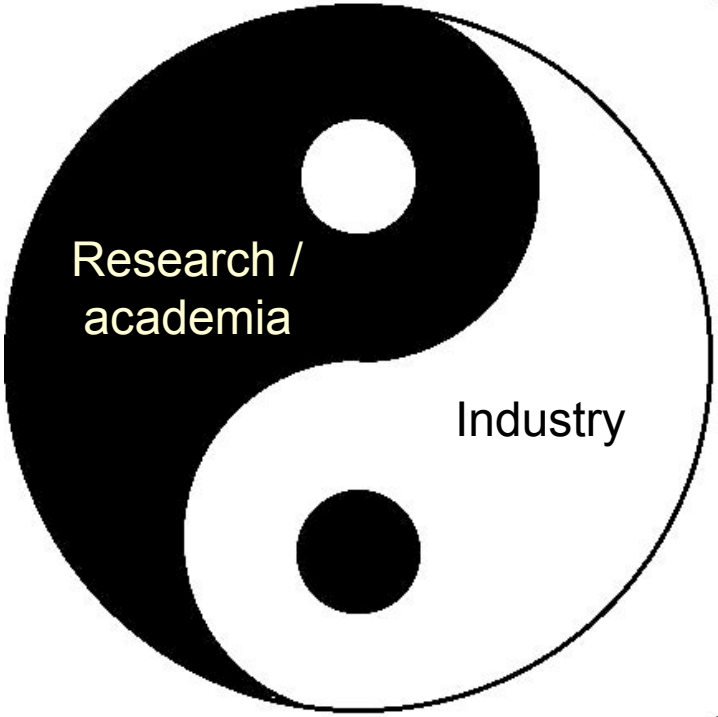
Oak Ridge National Laboratory
U.S. Department of Energy

Outline

- Open MPI overview
- Technical Design
  - Point-to-Point design
  - Collective Design
- Benchmark results
  - Bandwidth and Latency
  - Applications
- Future work

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, 6 contributors
  - **4 US DOE labs**
  - **8 universities**
  - **7 vendors**
  - **1 individual**

IBM
OAK RIDGE National Laboratory
ZIH Center for Information Services & High Performance Computing
CISCO
VOLTAIRE The Grid Backbone
Los Alamos NATIONAL LABORATORY EST. 1943
Myricom
Sandia National Laboratories
AIST
CHEMNITZ UNIVERSITY OF TECHNOLOGY
ICL UT
pervasivetechnologylabs AT INDIANA UNIVERSITY
Mellanox TECHNOLOGIES
QLOGIC
Sun microsystems
CS@UH
HLRS

Current projects

- "Open MPI Project" is an umbrella organization for multiple projects
  - **OMPI: Open MPI**
  - **ORTE:Open Run-Time Environment**
  - **PLPA: Portable Linux Processor Affinity**
  - **MTT: MPI (Middleware) Testing Tool**

Success stories

- OFED + Open MPI
  - **Thunderbird Sandia cluster**
    - **#6 in Top 500**
  - **Road Runner Los Alamos cluster**
    - **16k Opteron cores + 16k cell broadband engines**
  - **Coyote Los Alamos cluster**
    - **2580 Opteron cores**
- Sun ClusterTools v7

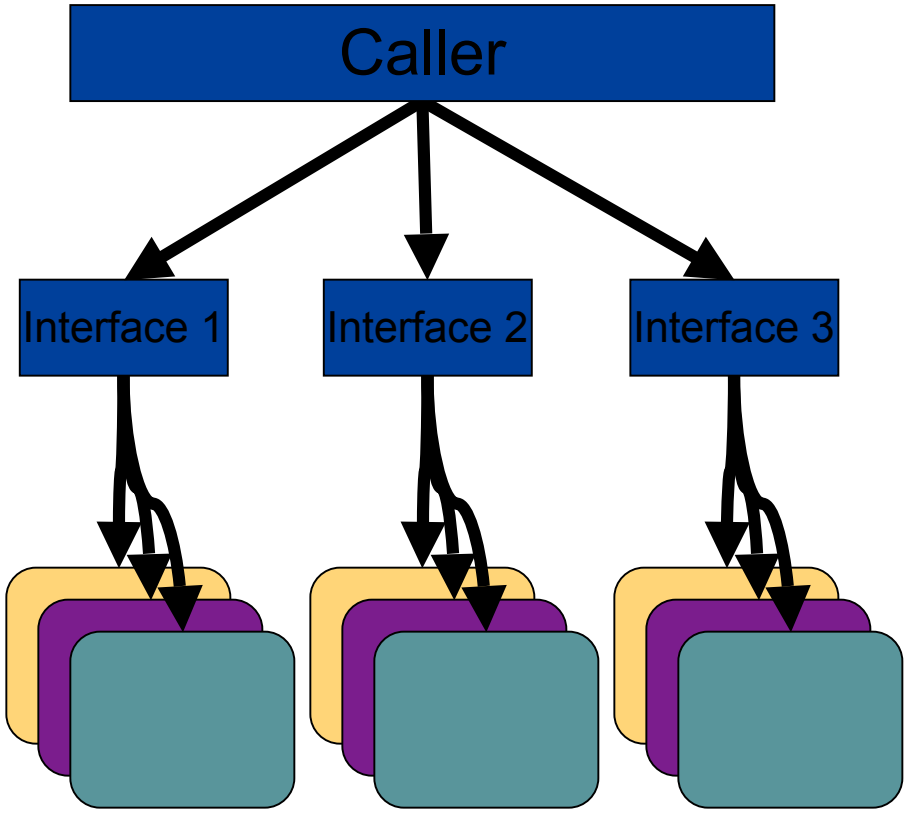
Roadmap

- 1.2 series is current stable
  - **v1.2.1 latest release**
- 1.3 series tentatively targeted at end of year
  - **Checkpoint / restart (and other FT)**
  - **Integration with debuggers**
  - **Windows support (*)**
  - **MPI collectives performance improvements**
  - **LSF integration**

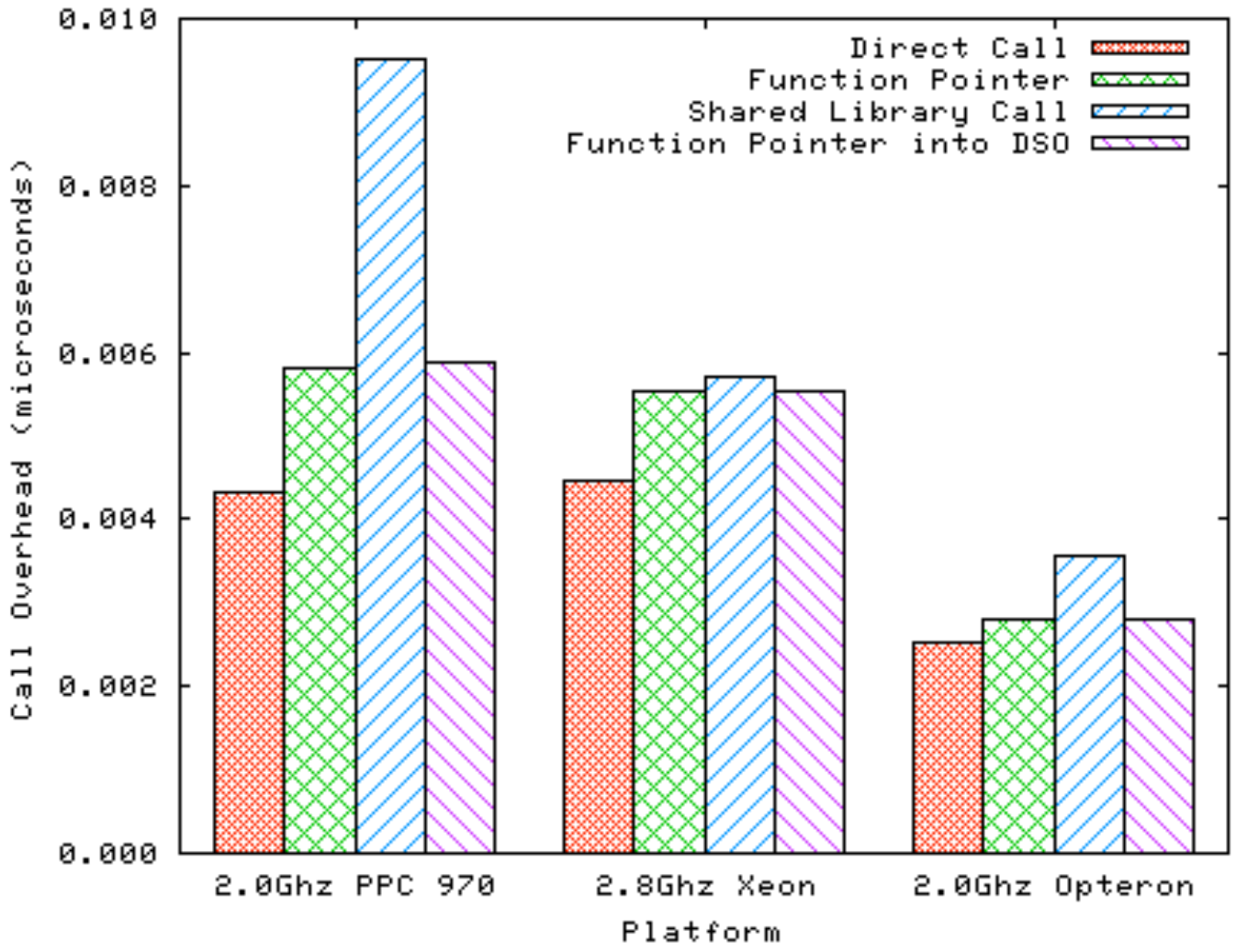
Technical Background

Key Design Feature: Components

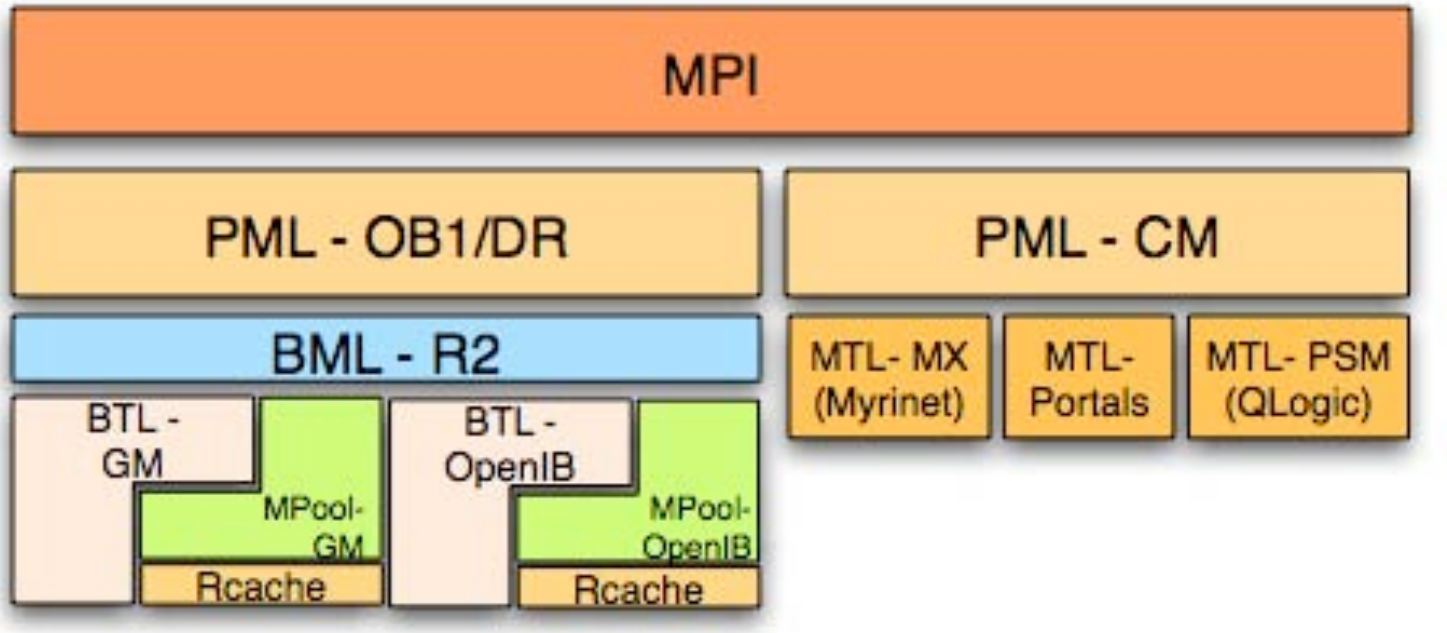
- Formalized interfaces
  - **Specifies "black box" implementation**
  - **Different implementations available at run-time**
  - **Can compose different systems on the fly**



Performance Impact



Point-To-Point Architecture



Portals Port: OB1 vs. CM

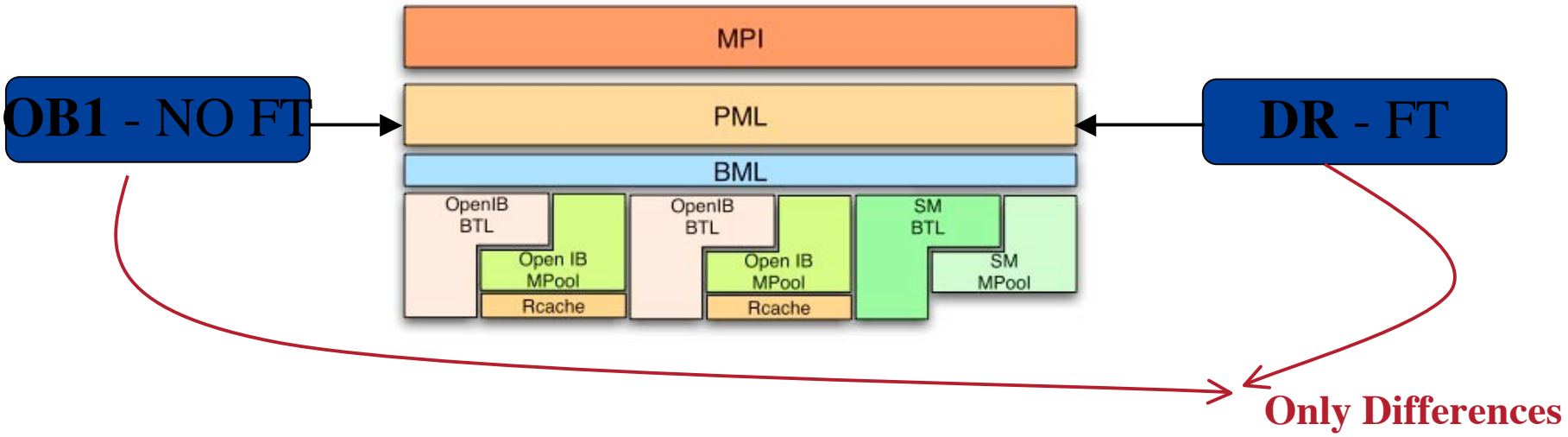
OB1

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

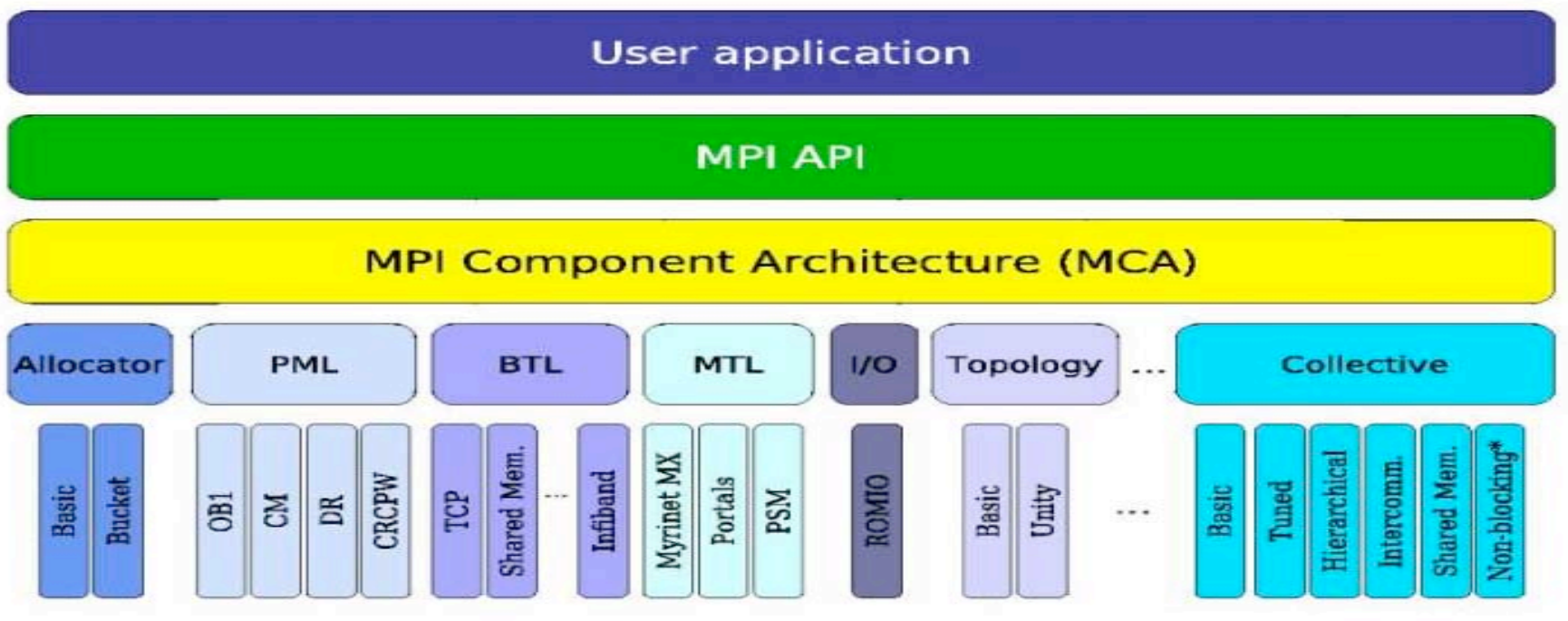
CM

- 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

Network Fault-Tolerance



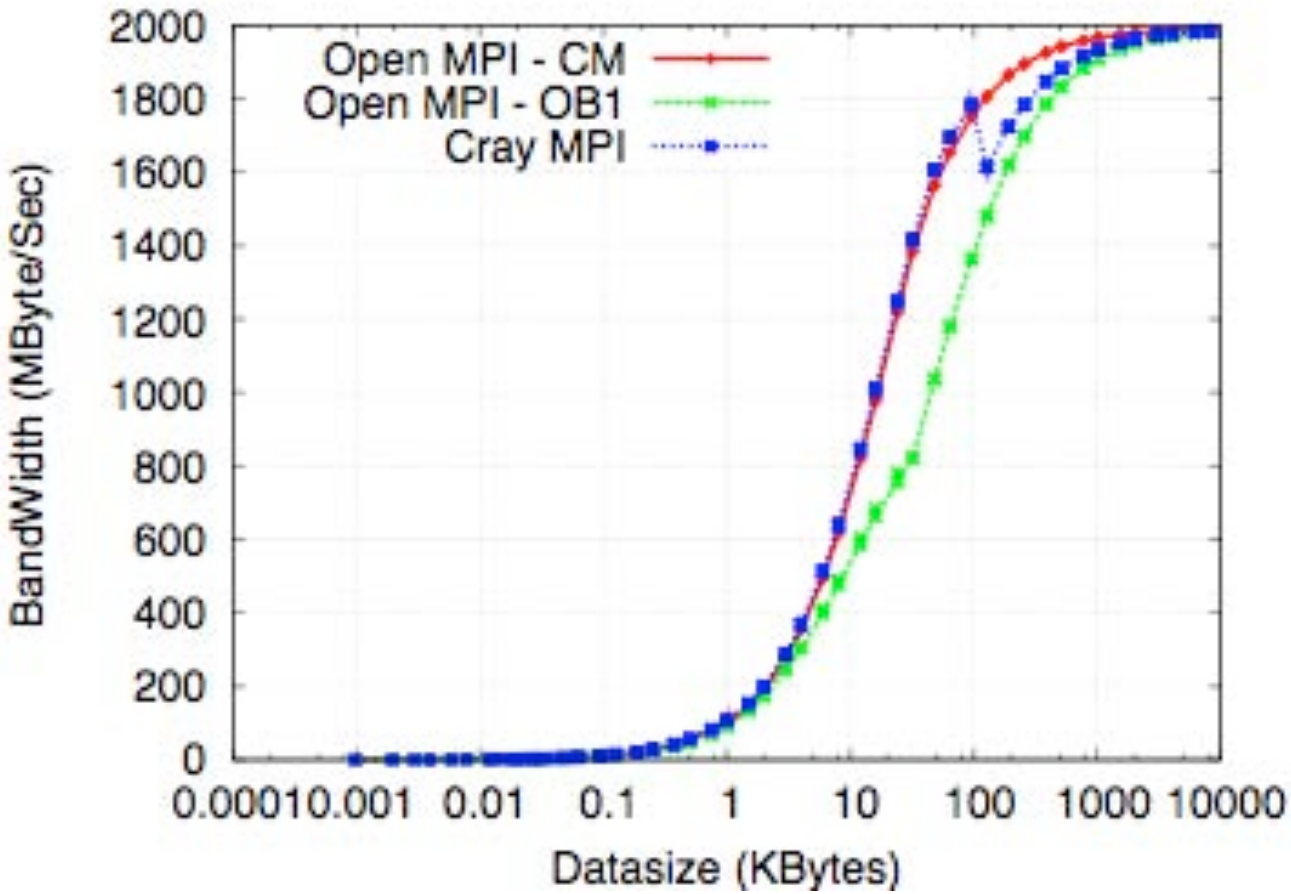
Collective Communications Component Structure



Benchmark Results

Latency and Bandwidth Data

NetPipe Bandwidth Data (MB/sec)

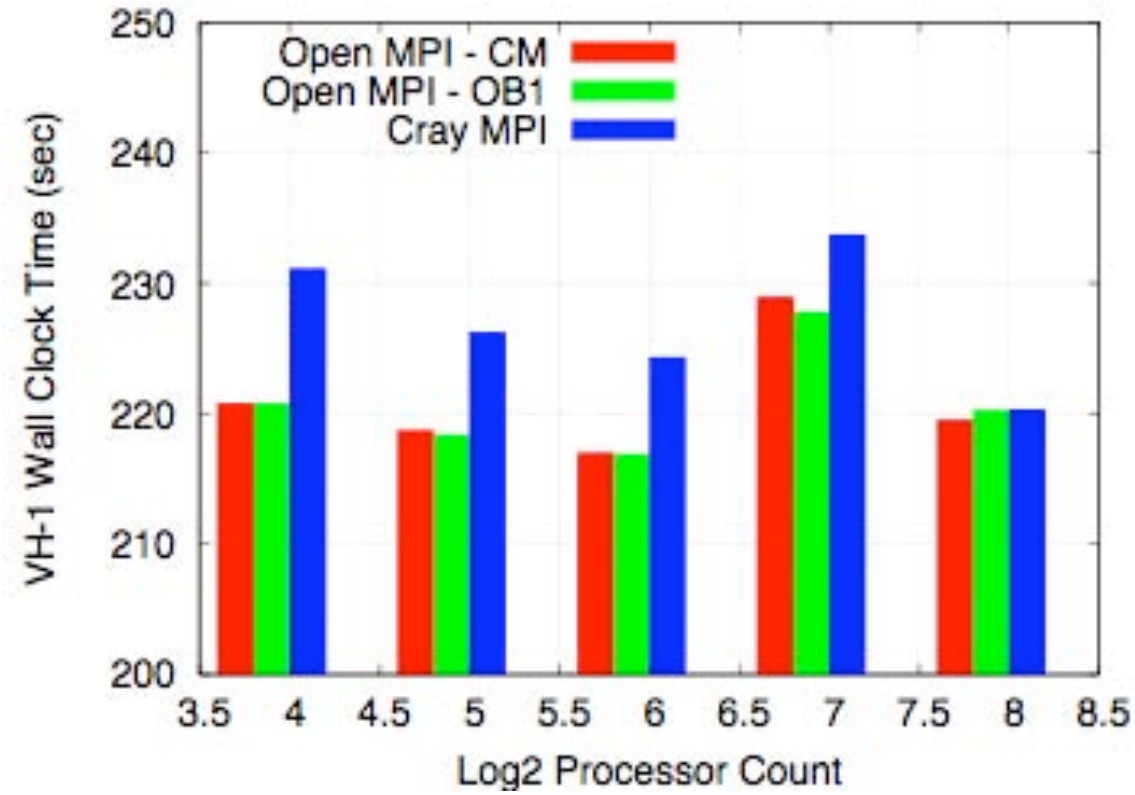


Zero Byte Ping-Pong Latency

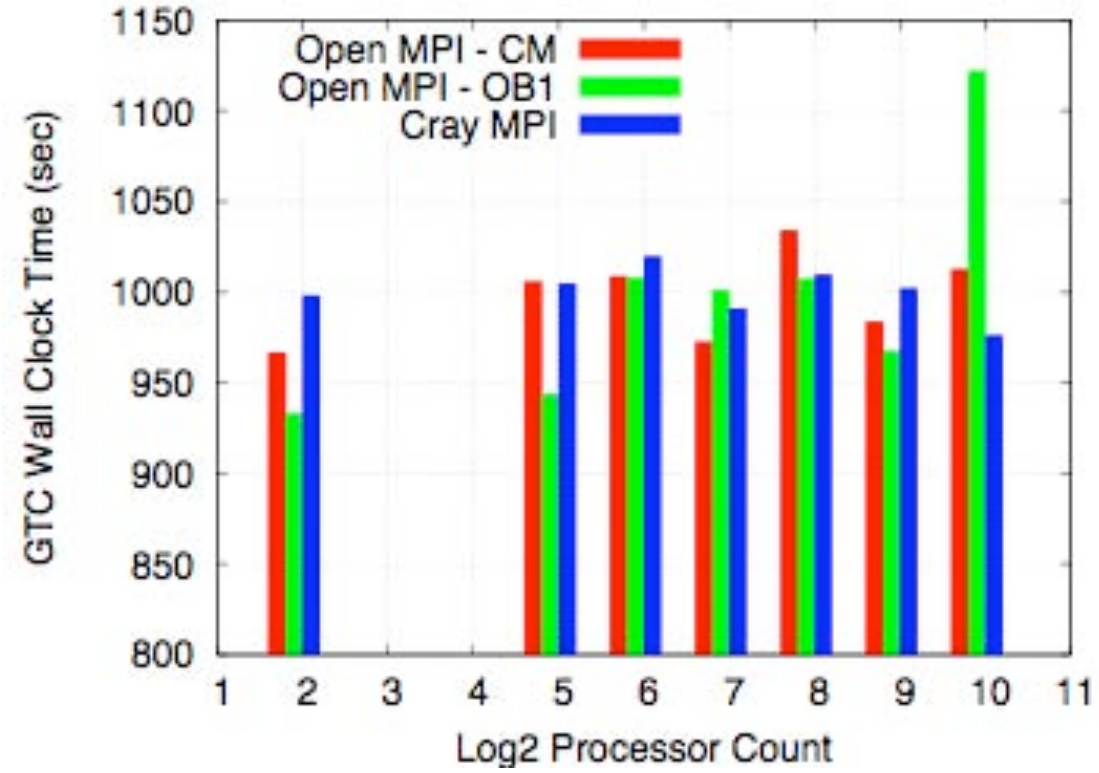
Open MPI - CM	4.91 usec
Open MPI - OB1	6.16 usec
Cray MPI	4.78 usec

Application Benchmarks

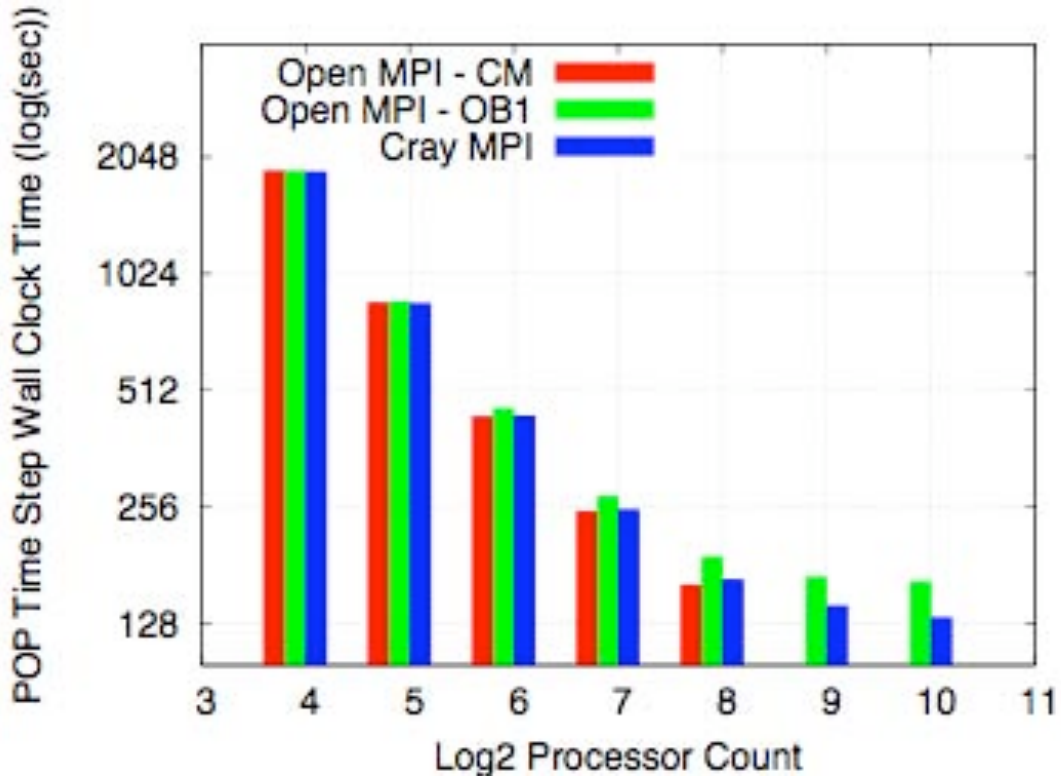
VH1 - Total Runtime



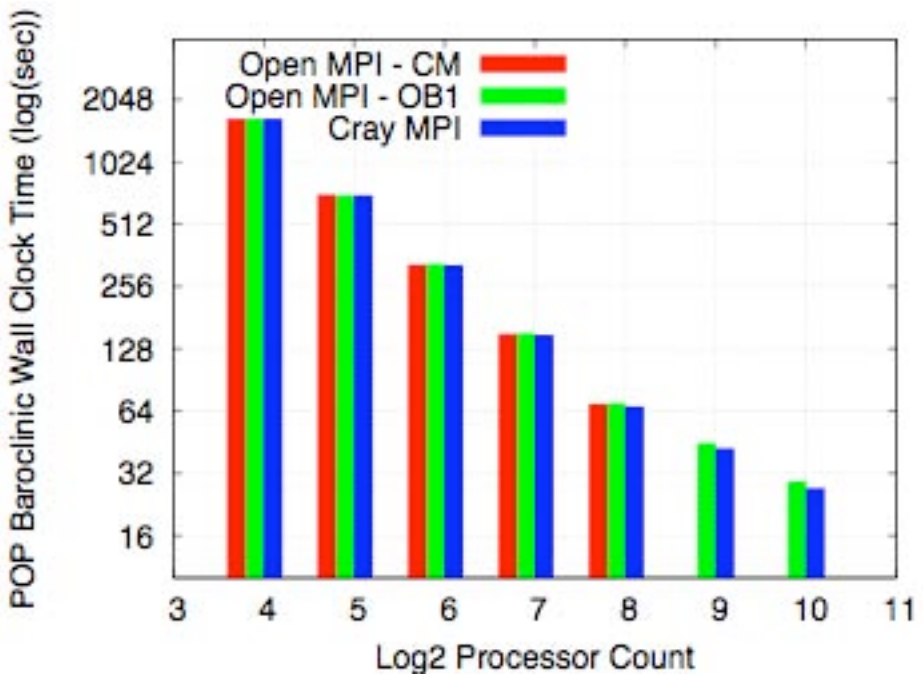
GTC - Total Runtime



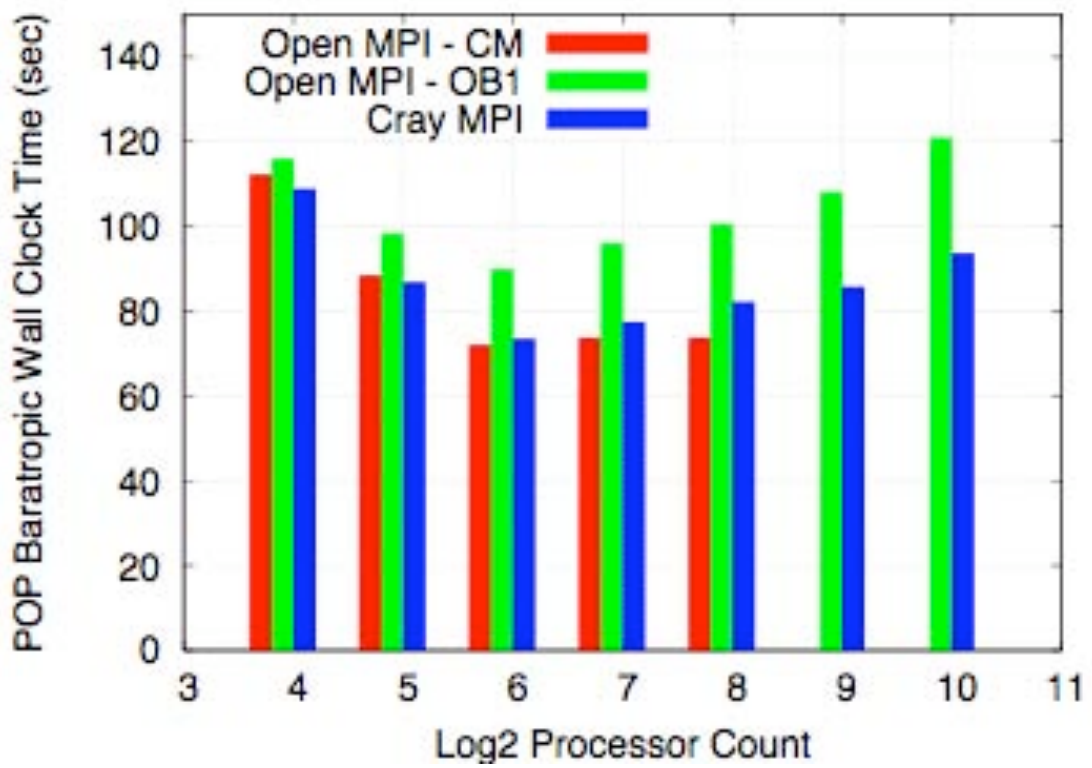
POP - Step Runtime



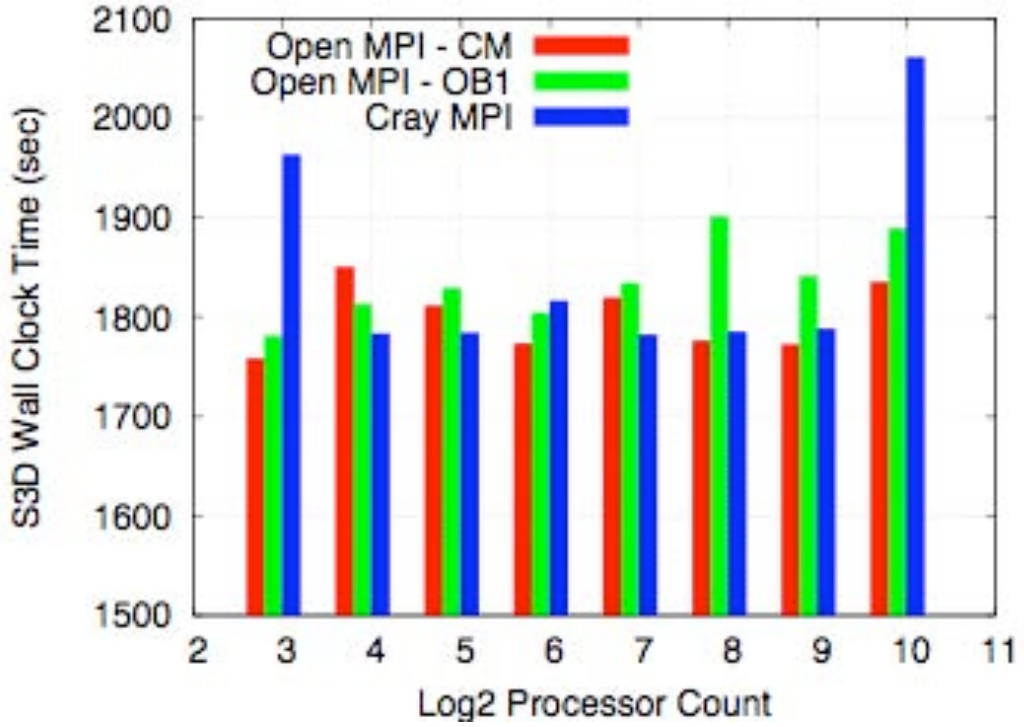
POP - Baroclinic Phase Totaltime



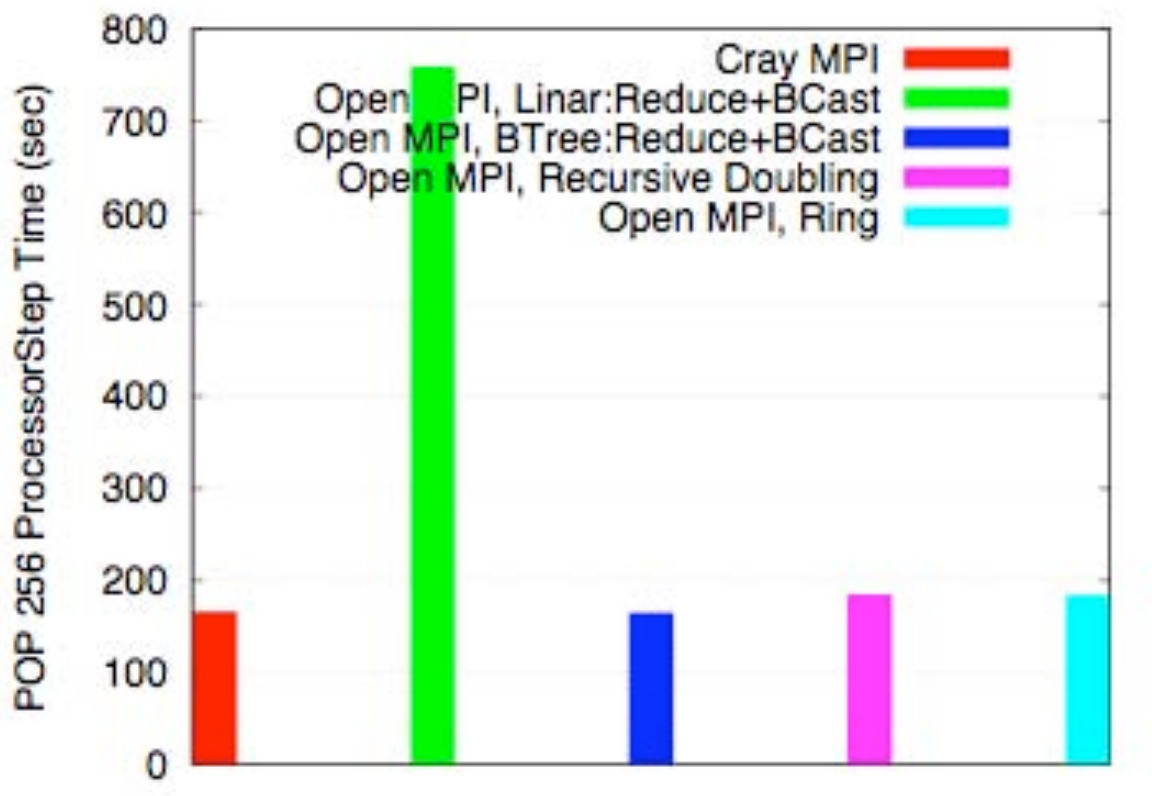
POP - Barotropic Phase Totaltime



S3D - Total Time



POP - Total Time at 256 Procs vs. Collective Algorithm



Future Directions

- ALPS port
- OB1 optimization
- Topology aware collectives