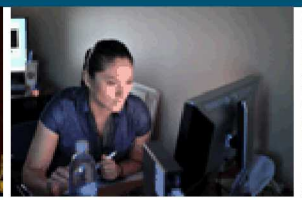
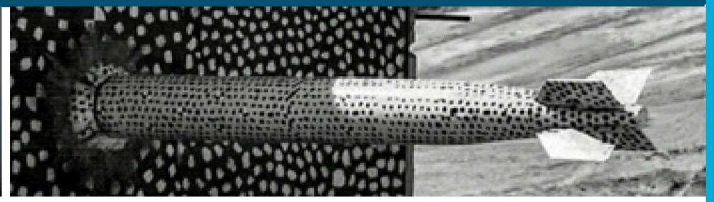
SAND2018-6282PE

Portals 4: Status of Specification and Implementation





PRESENTED BY

Andrew J. Younge

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

New specification updates

- Portals 4.1 released in 2017 – a large "minor" revision
- Multi-party agreement on features and behavior
  - Bull BXI is Portals 4.1 Compliant

New MPI performance optimizations

- Wide support for offloaded collectives in Open MPI
  - Use triggered operations to offload collectives
  - Complicated but practical engineering
    - Demonstrates how to expand such schemes for custom collectives
- Many Performance and Feature/Stability improvements (OMPI)

Portals 4 Specification

- Many changes in version 4.1
- Some changes for setup of future large changes
  - userID -> UsageID - moving to having multiple uids in Portals such that multiple isolation groups can be used in a single app
- Others impact features of current generation hardware
  - Difference atomics (PTL_DIFF) added
  - Many improvements around Nis that can support atomics that are coherent with host
- Important note:
  - PTL_DIFF operations may return the completed op type of PTL_SUM on BXI hardware if the atomic was performed
    - If not performed (overflow list) then it will be listed correctly
    - Not an issue as op type is normally not checked for priority list recvs

Portals 4 Specification – Future

- Many changes in version 4.2 as well, upcoming this year
  - Some new features – ptl_triggered_me_append etc
  - Many clarifications and notes to users
  - Important note: changes coming to atomics and unexpected list matches, atomics will not be performed on unexpected list matches
- Version 4.3 in development as well
  - More sweeping changes that can impact next generation hardware or hardware refresh
  - Exploring new features
    - Operation caching – reduce data movement by allowing common operations to be setup on NIC with almost all command fields filled out
    - Example: Put operation with similar options is used many times, only change destination, buffer and size, cache the rest of the options on the NIC

Portals 5 – The near horizon

- Current exploration for Portals 5 concentrates on moving compute closer to network data
  - Enhance existing offloads (triggered ops) to enable significantly more complex operation on data at the NIC
  - New methods being explored to allow for In Network Program Execution
  - Need non-stream data processing (no deadlines for working on data like in stream models like SPiN)
    - Streaming methods have limited future utility as eventual line rates will cause deadlines to become too short to be useful
- Other work concentrates on more efficient data movement and notification
  - Exploring alternative completion notification that will allow fast wake from sleep notification of the proper waiting core

Portals – Community Effort

Many stakeholders in Portals Networks

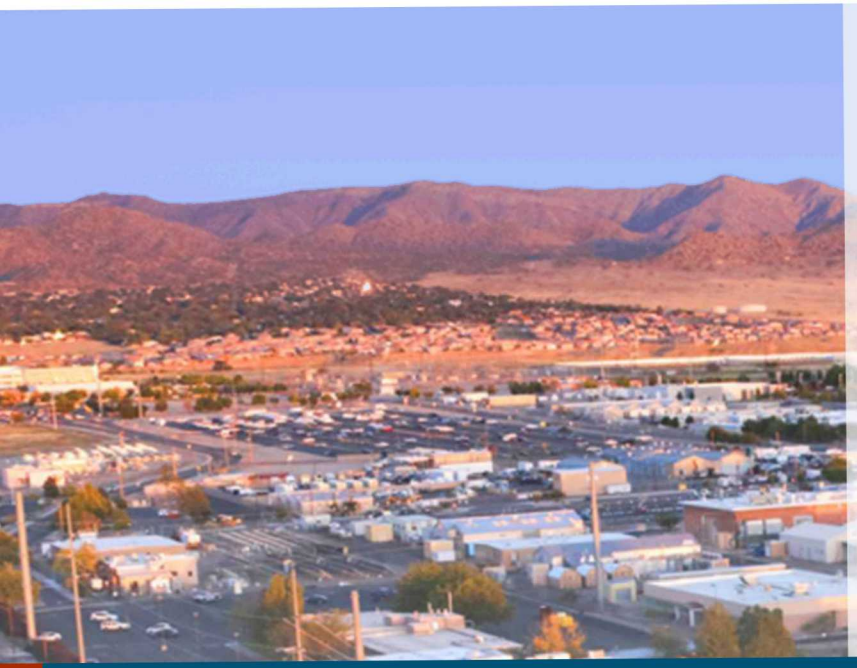
- We are building a community
- Stakeholders nearing stage where public discussion of Portals specification is possible
- Version 4.3 timeframe for community participation
- Users and sites with Portals hardware are welcomed to join
- Community will be invite only

Portals – Implementation Status

Support for Portals 4 is widespread

- Sandia OpenSHMEM is the leading SHMEM implementation
  - Includes both Portals 4 and OFI support
  - Under active development and production hardening with Intel Corp.
- Open MPI continues to be the recommended Portals MPI implementation
  - Active development and bug fixes for Portals components
- MPICH supports Portals netmod

Collaborative Questions



Questions?

Email: regrant@sandia.gov