SAND2006-7142P

Accelerated Portals

Integration into XT3 Code Base

November 12, 2006

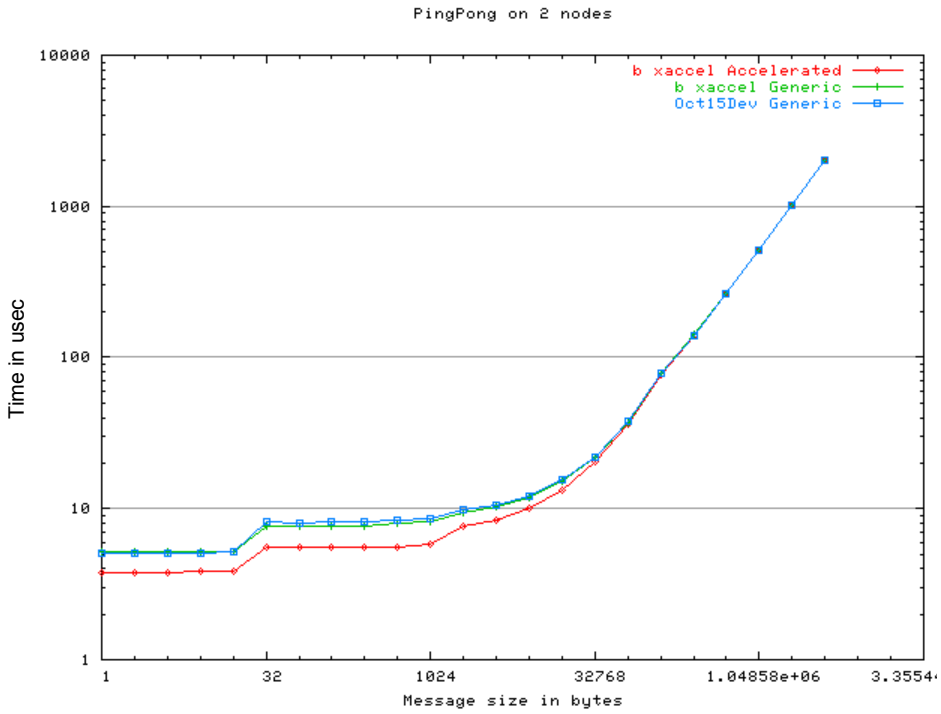
Sue Kelly
smkelly@sandia.gov
505-845-9770

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

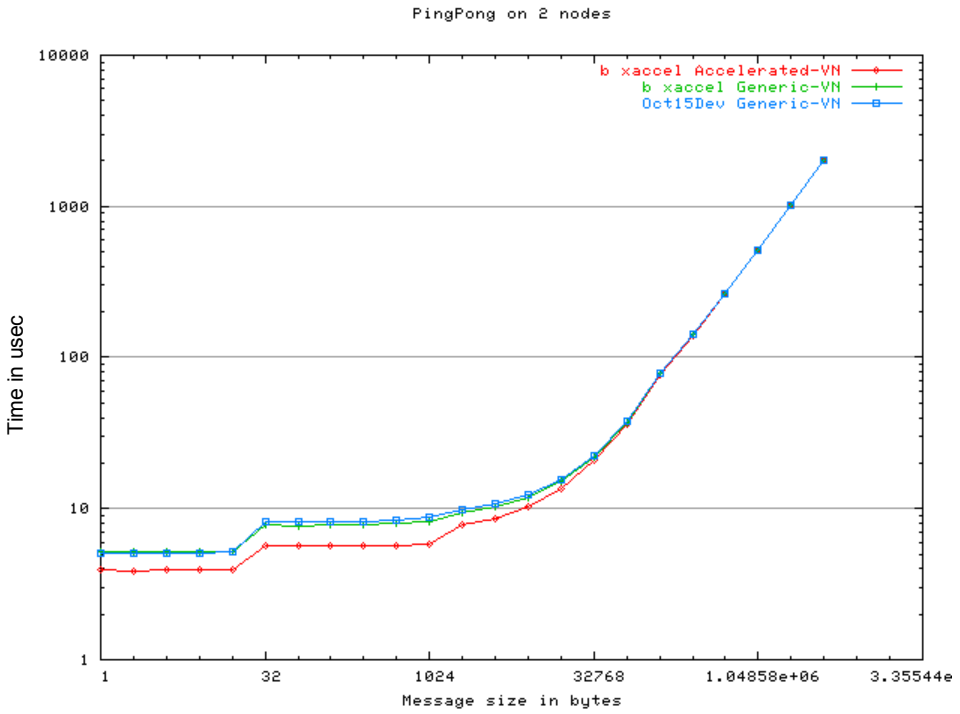
Background

- Portals – low level networking software designed for scalability and message passing
  - http://gaston.sandia.gov/cfupload/ccim_pubs_prod/portals33.pdf
- Generic portals (Host based mode) – an implementation of the V3.3 portals specification for the XT3 that places most of the protocol processing in the kernel.
  - http://gaston.sandia.gov/cfupload/ccim_pubs_prod/Brightwell_paper.pdf
- Accelerated portals (NIC-based mode) – an implementation of the V3.3 portals specification for the XT3 that places most of the protocol processing in the Seastar NIC (Network Interface Chip)
  - http://doi.ieeecomputersociety.org/10.1109/MM.2006.65

Accelerated Portals has Lower Latency



SN mode

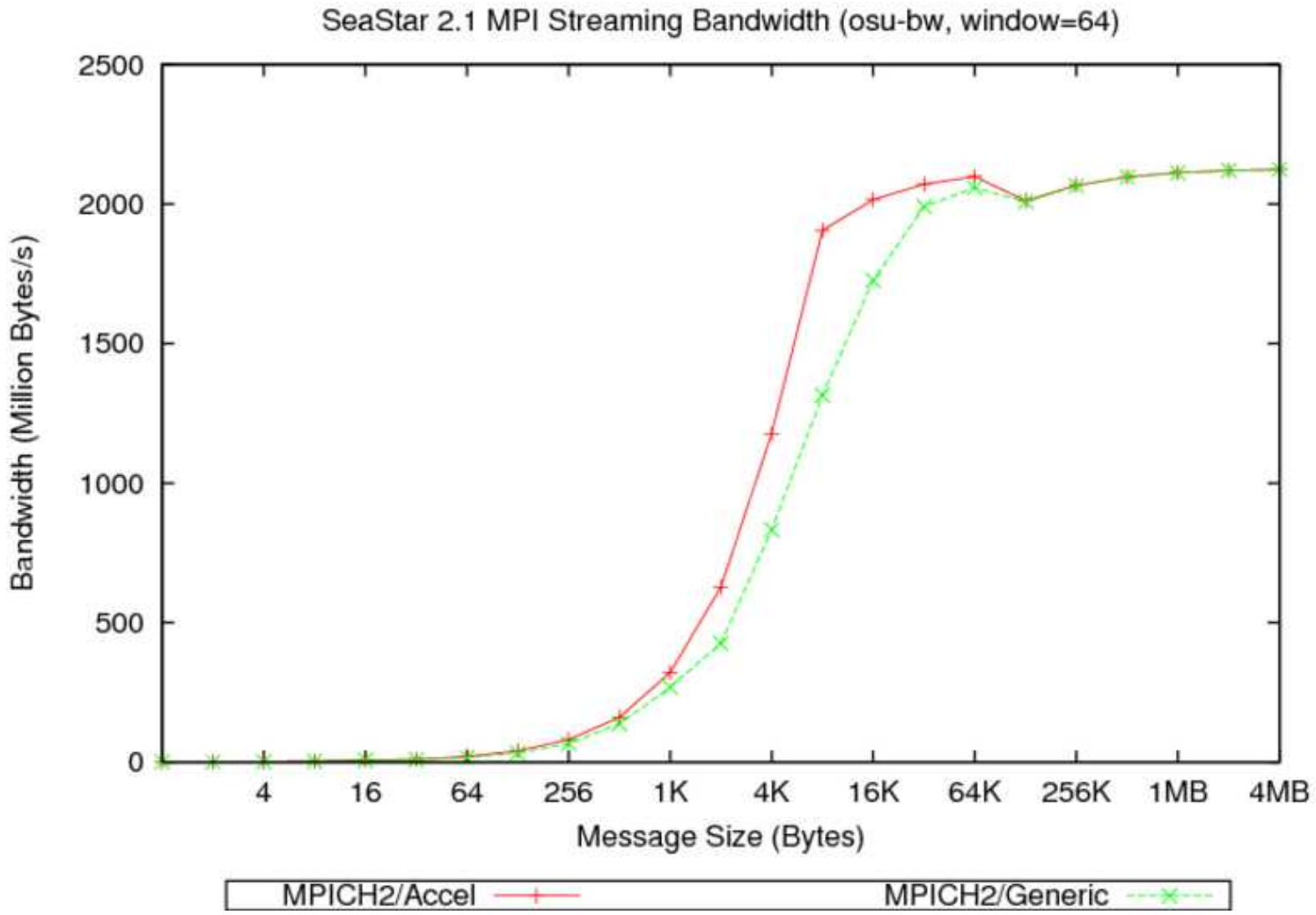


VN mode

HPCC Random Access showed 27% improvement in SN mode and **90%** in VN mode

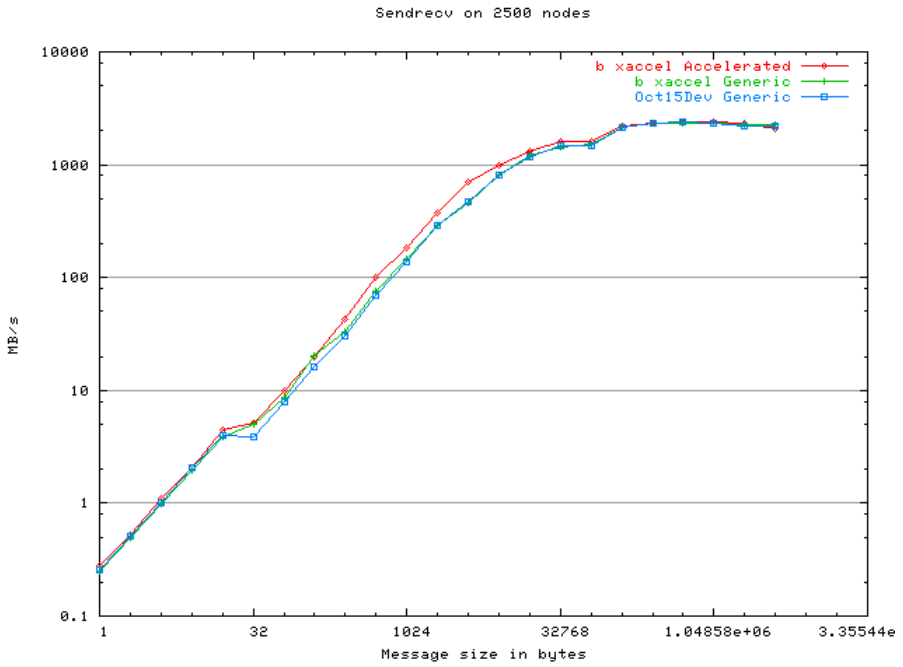
Sandia National Laboratories

Accelerated Provides Improved Throughput on Small (<32KB) Messages

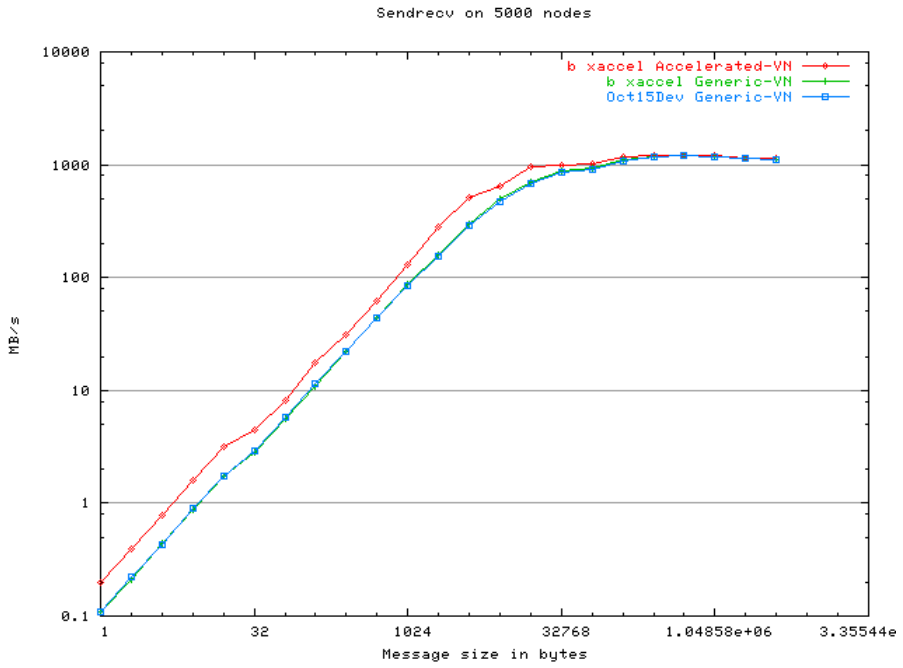


Sandia National Laboratories

Accelerated Offers Improved Performance in VN Mode due to Protocol Offload



SN mode



VN mode

Sandia National Laboratories

Accelerated Constraints

- Only works on Catamount
  - Useful for MPI between Catamount nodes
  - File I/O will continue to use generic
  - Yod/pct load and fan-out protocols will use generic
- The number of outstanding match entries is limited to 1024 (currently 2048 in generic).
- Applications that will benefit from accelerated:
  - Those with latency-sensitive messaging
  - Those able to overlap computation and communication

Testing Status

- Ran Cray Silver Suite in October on Cray test system
  - No regressions
- Tested on Red Storm on 10/30-11/1
  - Tested with Pallas, HPCC, CTH and Sage
  - Ran SN and VN modes three times each (generic, generic with accelerated in code base, and accelerated)
  - Max application size was 2500 (5000 virtual) nodes
  - Ran on as many as 7500 nodes at one time.
  - No regressions to generic portals
  - Latency-insensitive apps (i.e. CTH and Sage) did not show improvement with accelerated
  - One application (CTH), using accelerated portals, hung on 2500 nodes. App was killable and nodes were re-usable. No time to debug.

Code Status

- To be checked into Cray's 2.0 code base on 11/9/06
- Accelerated code is not enabled
  - Requires ENV support in MPICH
- No CAM protection provided (yet)
  - Current COP/BEER protocol is host-level and can't detect dropped accelerated messages
  - Accelerated currently coded to abort node on CAM overflow
  - Hope to integrate Sandia's NIC-level CAM protection; date TBD
- Support to non-Sandia sites is targeted for a later 2.0.x release