BSMBench
Performance Benchmark and Profiling

January 2017

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
  - BSMBench performance overview
  - Understanding BSMBench communication patterns
  - Ways to increase BSMBench productivity
- **For more info please refer to**
  - http://www.bsmbench.org/
  - https://gitlab.com/edbennett/BSMBench

BSMBench

- **Open source supercomputer benchmarking tool**
- **Based on simulation code used for studying strong interactions in particle physics**
- **Includes the ability to tune the ratio of communication over computation**
- **Includes 3 examples that show the performance of the system for**
  - Problem that is computationally dominated (marked as Communications)
  - Problem that is communication dominated (marked as Compute)
  - Problem in which communication and computational requirements are balanced (marked as Balance)
- **Used to simulate workload such as Lattice Quantum ChromoDynamics (QCD), and by extension its parent field, Lattice Gauge Theory (LGT), which make up a significant fraction of supercomputing cycles worldwide**
- **For reference: technical paper published at the 2016 International Conference on High Performance Computing & Simulation (HPCS), Innsbruck, Austria, 2016, pp. 834-839**

Objectives

- **The presented research was done to provide best practices**
  - BSMBench performance benchmarking
    - MPI Library performance comparison
    - Interconnect performance comparison
    - Compilers comparison
    - Optimization tuning
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

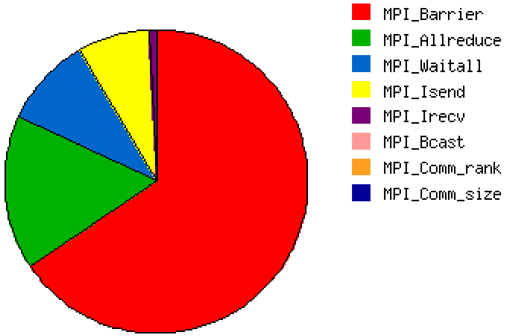
Test Cluster Configuration

- **Dell PowerEdge R730 32-node (1024-core) “Thor” cluster**
  - Dual-Socket 16-Core Intel E5-2697Av4 @ 2.60 GHz CPUs (BIOS: Maximum Performance, Home Snoop)
  - Memory: 256GB memory, DDR4 2400 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
  - OS: RHEL 7.2, M MLNX_OFED_LINUX-3.4-1.0.0.0 InfiniBand SW stack
- **Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters**
- **Mellanox Switch-IB SB7800 36-port EDR 100Gb/s InfiniBand Switch**
- **Intel® Omni-Path Host Fabric Interface (HFI) 100Gbps Adapter**
- **Intel® Omni-Path Edge Switch 100 Series**
- **Dell InfiniBand-Based Lustre Storage based on Dell PowerVault MD3460 and Dell PowerVault MD3420**
- **Compilers: Intel Compilers 2016.4.258**
- **MPI: Intel Parallel Studio XE 2016 Update 4, Mellanox HPC-X MPI Toolkit v1.8**
- **Application: BSMBench Version 1.0**
- **MPI Profiler: IPM (from Mellanox HPC-X)**

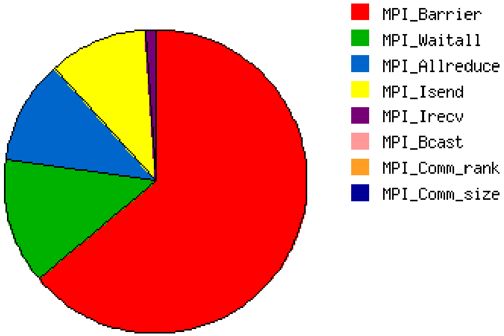
BSMBench Profiling – % of MPI Calls

- **Major MPI calls (as % of wall time):**
  - Balance: MPI_Barrier (26%), MPI_Allreduce (6%), MPI_Waitall (5%), MPI_Isend (4%)
  - Communications: MPI_Barrier (14%), MPI_Allreduce (5%), MPI_Waitall (5%), MPI_Isend (2%)
  - Compute: MPI_Barrier (14%), MPI_Allreduce (5%), MPI_Waitall (5%), MPI_Isend (1%)

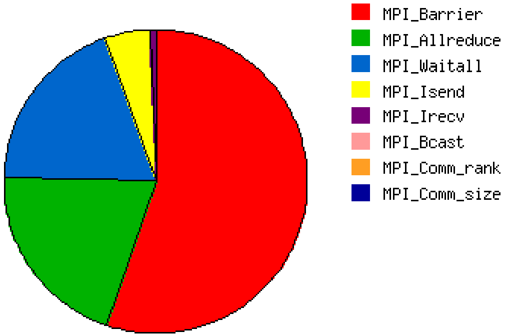
Balance



Communications



Compute

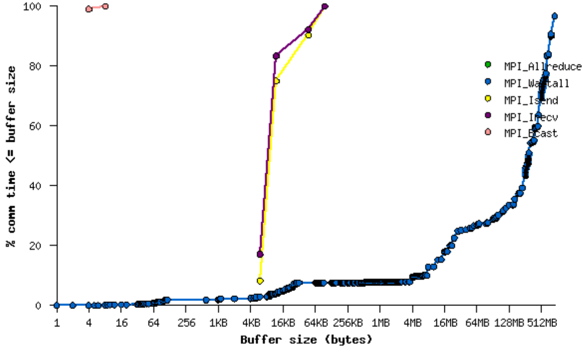


32 Nodes / 1024 Processes

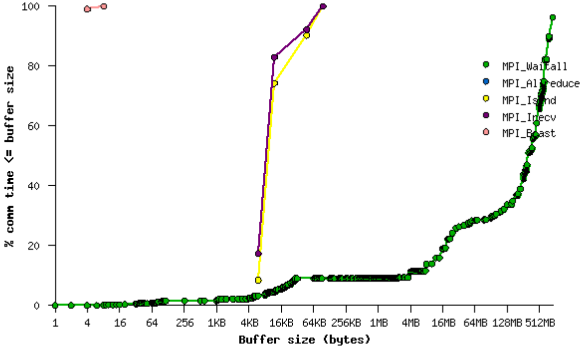
BSMBench Profiling – MPI Message Size Distribution

- **Similar communication pattern seen across all 3 examples:**
  - Balance: MPI_Barrier: 0-byte, 22% wall, MPI_Allreduce: 8-byte, 5% wall
  - Communications: MPI_Barrier: 0-byte, 26% wall, MPI_Allreduce: 8-byte, 5% wall
  - Compute: MPI_Barrier: 0-byte, 13% wall, MPI_Allreduce: 8-byte, 5% wall

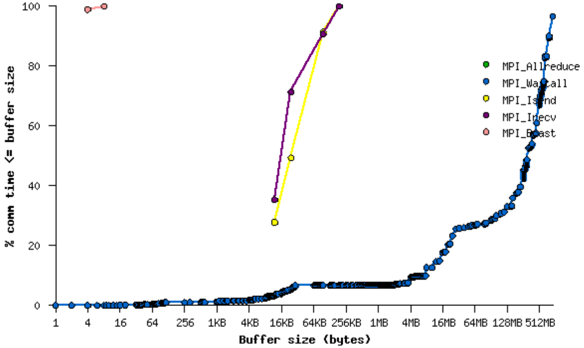
Balance



Communications



Compute

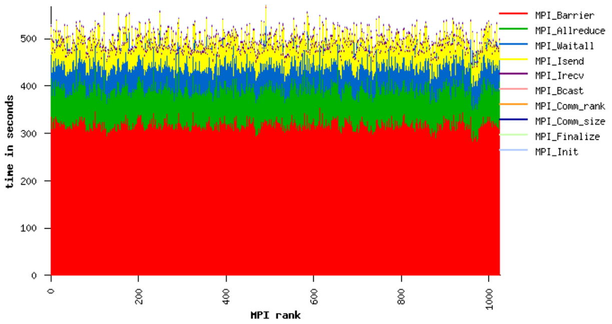


32 Nodes / 1024 Processes

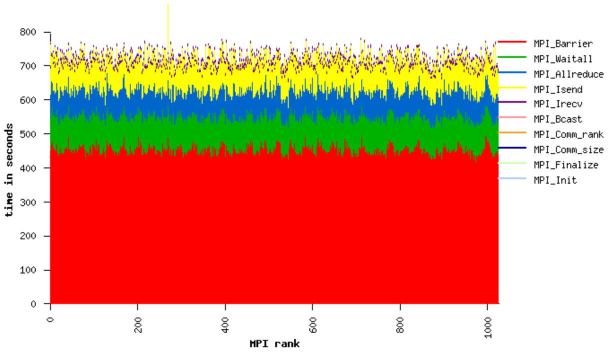
BSMBench Profiling – Time Spent in MPI

- **The different communications across the MPI processes is mostly balance**
  - Does not appear to be any significant load imbalances in the communication layer

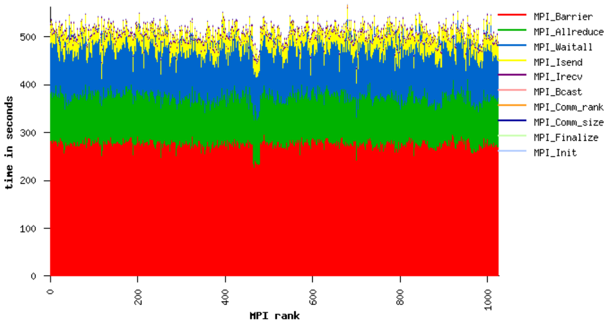
Balance



Communications



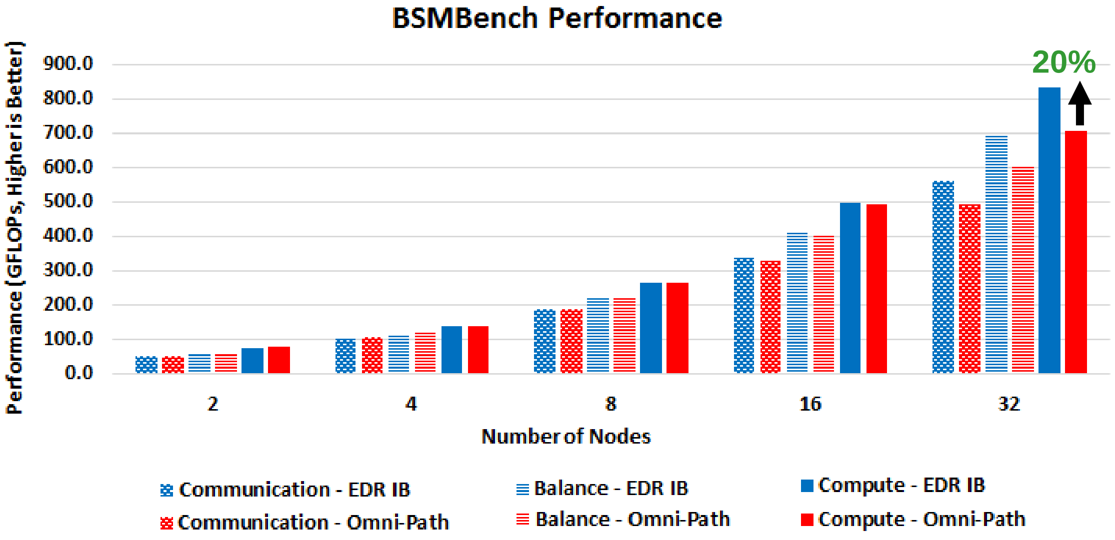
Compute



32 Nodes / 1024 Processes

BSMBench Performance – Interconnects

- **EDR InfiniBand delivers better scalability for BSMBench**
  - Similar performance between EDR InfiniBand and Omni-Path up to 8 nodes
  - Close to 20% performance advantage for InfiniBand at 32 nodes
  - Similar performance difference across the three different examples

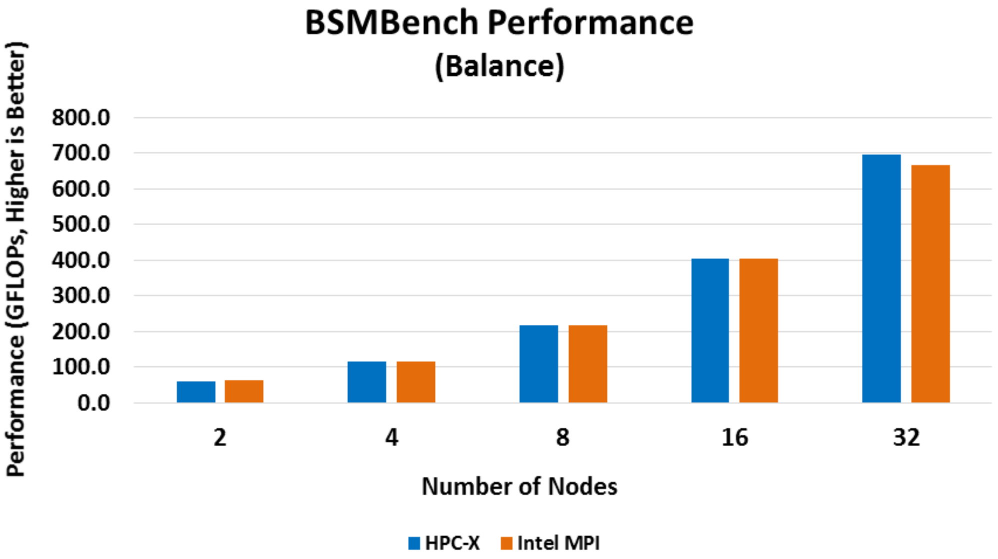


Higher is better

32 MPI Processes / Node

BSMBench Performance – MPI Libraries

- **Comparison between two commercial available MPI libraries**
- **Intel MPI and HPC-X delivers similar performance**
  - HPC-X demonstrates 5% advantage at 32 nodes



Higher is better

32 MPI Processes / Node

BSMBench Summary

- **Benchmark for BSM Lattice Physics**
  - Utilizes both compute and network communications
- **Fast network communication is important for scalability**
- **Interconnect comparison**
  - EDR InfiniBand demonstrates higher scalability beyond 16 nodes as compared to Omni-Path
  - EDR InfiniBand delivers nearly 20% higher performance 32 nodes / 1024 cores
  - Similar performance advantage across all three example cases
- **MPI Profiling**
  - Most MPI time is spent on MPI collective operations and non-blocking communications
    - Heavy use of MPI collective operations (MPI_Allreduce, MPI_Barrier)
  - Similar communication patterns seen across all three examples
    - Balance: MPI_Barrier: 0-byte, 22% wall, MPI_Allreduce: 8-byte, 5% wall
    - Comms: MPI_Barrier: 0-byte, 26% wall, MPI_Allreduce: 8-byte, 5% wall
    - Compute: MPI_Barrier: 0-byte, 13% wall, MPI_Allreduce: 8-byte, 5% wall

Thank You

HPC Advisory Council