Best Practices: Application Profiling at the HPCAC High Performance Center

Rich Graham
Community Outreach, Education & Enablement

- World-wide organization, not-for-profit collaborations
- >450 member companies / universities / organizations
- Bridges the gap between HPC usage and its potential
- Provides best practices and a support/development center
- Explores future technologies and future developments
- Leading edge solutions and technology demonstrations
164 Applications Best Practices Published

- Abaqus
- ABYSS
- AcuSolve
- Amber
- AMG
- AMR
- ANSYS CFX
- ANSYS Fluent
- ANSYS Mechanical
- BQCD
- BSMBench
- CAM-SE
- CCSM 4.0
- CESM
- COSMO
- CP2K
- CPMD
- Dacapo
- Desmond
- DL-POLY
- Eclipse
- FLOW-3D
- GADGET-2
- Graph500
- GROMACS
- Himeno
- HIT3D
- HOOMD-blue
- HPCC
- HPCG
- HYCOM
- ICON
- Lattice QCD
- LAMMPS
- LS-DYNA
- MiniFE
- MILC
- MSC Nastran
- MR Bayes
- MM5
- MPQC
- NAMD
- Nekbone
- NEMO
- NWChem
- Octopus
- OpenAtom
- OpenFOAM
- OpenMX
- OptiStruct
- PAM-CRASH / VPS
- PARATEC
- Pretty Fast Analysis
- PFLOTRAN
- Quantum ESPRESSO
- RADIOSS
- RFD tNavigator
- SNAP
- SPECFEM3D
- STAR-CCM+
- STAR-CD
- VASP
- WRF

### 38 Applications Installation Best Practices Published

- Adaptive Mesh Refinement (AMR)
- Amber (for GPU/CUDA)
- Amber (for CPU)
- ANSYS Fluent 15.0.7
- ANSYS Fluent 17.1
- BQCD
- Caffe
- CASTEP 16.1
- CESM
- CP2K
- CPMD
- DL-POLY 4
- ESI PAM-CRASH 2015.1
- ESI PAM-CRASH / VPS 2013.1
- GADGET-2
- GROMACS 5.1.2
- GROMACS 4.5.4
- GROMACS 5.0.4 (GPU/CUDA)
- Himeno
- HOOMD Blue
- LAMMPS
- LAMMPS-KOKKOS
- LS-DYNA
- MrBayes
- NAMD
- NEMO
- NWChem
- NWChem 6.5
- Octopus
- OpenFOAM
- OpenMX
- PyFR
- Quantum ESPRESSO 4.1.2
- Quantum ESPRESSO 5.1.1
- Quantum ESPRESSO 5.3.0
- TensorFlow 0.10.0
- WRF 3.2.1
- WRF 3.8

For more information, visit: [http://www.hpcadvisorycouncil.com/subgroups_hpc_works.php](http://www.hpcadvisorycouncil.com/subgroups_hpc_works.php)
Dell™ PowerEdge™ C6145 6-node cluster
Dell™ PowerEdge™ R815 11-node cluster
HPE Apollo 6000 10-node cluster
HPE ProLiant SL230s Gen8 4-node cluster
HPE Cluster Platform 3000SL 16-node cluster

Dell™ PowerEdge™ R730 GPU 36-node cluster
Dell PowerVault MD3420 Dell PowerVault MD3460
InfiniBand Storage (Lustre)
Dell™ PowerEdge™ M610 38-node cluster
IBM POWER8 8-node cluster

Dell™ PowerEdge™ C6100 4-node cluster
4-node GPU cluster
4-node GPU cluster
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
- HPC application performance overview
- Understanding HPC application communication patterns
- Ways to increase HPC application productivity
Test Clusters

- **HPE ProLiant DL360 Gen9 128-node (4096-core) “Hercules” cluster**
  - Dual-Socket 16-Core Intel E5-2697A v4 @ 2.60 GHz CPUs
  - Memory: 256GB memory, DDR4 2400 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
  - OS: RHEL 7.2, MLNX_OFED_LINUX-3.4-2.0.0.0 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) 100Gb/s Adapter, Intel® Omni-Path Edge Switch 100 Series
  - MPI: Intel MPI 2017, Open MPI 2.02

- **IBM OperPOWER 8-node “Telesto” cluster - IBM Power System S822LC (8335-GTA)**
  - IBM: Dual-Socket 10-Core @ 3.491 GHz CPUs, Memory: 256GB memory, DDR3 PC3-14900 MHz
  - Wistron OpenPOWER servers (Dual-Socket 8-Core @ 3.867 GHz CPUs. Memory: 224GB memory, DDR3 PC3-14900 MHz)
  - OS: RHEL 7.2, 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
  - Compilers: GNU compilers 4.8.5, IBM XL Compilers 13.1.3
  - MPI: Open MPI 2.0.2, IBM Spectrum MPI 10.1.0.2, MPI Profiler: IPM

- **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, MPI Profiler: IPM (from Mellanox HPC-X)
OpenFOAM Test Case
OpenFOAM Test case

- **Problem:** Lid Driven Cavity
- **Solver:** icoFoam
- **Two MPI Implementations used:**
  - Open MPI
    - Base implementation
    - With Mellanox accelerations
  - IntelMPI
Test System

- **HPE ProLiant DL360 Gen9 128-node (4096-core) “Orion” cluster**
  - Dual-Socket 14-Core Intel E5-2697 v3 @ 2.60 GHz CPUs
  - Memory: 256GB memory, DDR4 2133 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
  - OS: RHEL 7.2, MLNX_OFED_LINUX-3.2-1.0.1.1 InfiniBand SW stack
  - Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters, Mellanox Switch-IB2 SB7800 36-port EDR 100Gb/s InfiniBand Switch
MPI Description

- **Open MPI**
  - Base version: 1.10
  - Mellanox Acceleration: HPC-X version 1.6
    - MXM point-to-point acceleration
    - HCOLL collective acceleration

- **IntelMPI**
  - MPICH based
  - IntelMPI optimizations
    - Point-to-point (UDAPL)
    - Collectives
Open MPI Scalability Testing As a Function of Optimizations

OpenFOAM Performance
(Lid Driven Cavity, icoFoam)

Number of Nodes

Performance Rating

- Open MPI
- HPC-X (MXM only)
- HPC-X (MXM+HCOLL)
- HPC-X (SHArP)
OpenFOAM Performance (Lid Driven Cavity, icoFoam)

Performance Rating vs Number of Nodes

- Intel MPI
- Open MPI
- HPC-X (SHArP)
MPI Library Breakdown as a Function of Time
MPI Message Distribution as a Function of MPI Operation
SHARP Hardware Capabilities
SHARP Trees

- **SHAPR Operations are Executed by a SHARP Tree**
  - Multiple SHARP Trees are Supported
  - Each SHARP Tree can handle Multiple Outstanding SHARP Operations

- **SHARP Tree is a Logical Construct**
  - Nodes in the SHARP Tree are IB Endnodes
  - Logical tree defined on top of the physical underlying fabric
  - SHARP Tree Links are implemented on top of the IB transport (Reliable Connection)
  - Expected to follow the physical topology for performance but not required
SHARP Characteristics

- Reliable Scalable General Purpose Primitive, Applicable to Multiple Use-cases
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Many simultaneous outstanding operations in flight

Accelerating HPC applications

- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 32 / 64 bit
  - Repeatable results
  - Up to 256 bytes per reduction

- Significantly reduce MPI collective runtime
- Increase CPU availability and efficiency
- Enable communication and computation overlap
SHAPR MPI_Allreduce() Performance

Graphs showing Allreduce Latency vs Cluster Size (Nodes) for different sizes (8B, 128B, 1024B, 2048B) and two software implementations.
Thank You