Performance of Applications on Nurion and Neuron Utilizing MVAPICH2

Minsik Kim, Ph.D.
Supercomputing Infrastructure Center, KISTI

9th Annual MVAPICH User Group Meeting (MUG’21)
Introduction to KISTI-5 Supercomputer and Future Roadmap
KISTI Supercomputing Center

- The National Supercomputing Center in Korea
- Provide computational resources and its support to R&D communities in Korea
- Nurion (KISTI-5): CPU system, Neuron: GPU system
- KISTI-6 production will start in 2023
KISTI-5 Procurement & Deployment

- **'15.06**: Data Center Building constructed
- **'15.07.07**: Approved from Preliminary feasibility study
- **'16.03**: RFI and BMT Code release
- **'16.10~12**: 1st Bidding
- **'17.02~05**: 2nd Bidding
- **'17.06~07**: Cray Inc. won the bid and Tech/Price Negotiation
- **'17.08**: Contract finalized (49M USD)
- **'17.11**: Pilot system(16nodes) delivered
- **'17.12~'18.4**: Main system delivery and deployment
- **'18.05~09**: BMT, Functional and Stability Test
- **'18.07~10**: Early Access on Pilot system
- **'18.10~11**: Main system Beta service
- **'18.12~**: Production

Cray CS500 25.7PFlops

KISTI Facility PUE 1.35
KISTI-5 Compute Nodes

The Largest KNL/OPA based commodity cluster System
Rpeak 25.7PFlops, Rmax 13.9PFlops

**Compute nodes**

- 8,305 KNL Computing modules, 116 Racks, 25.3PF
  - 1x Xeon Phi KNL 7250, 68Cores 1.4GHz, AVX512
  - 3TFlops Peak, ~0.2 Bytes/Flops,
  - 96GB (6x16GB) DDR4-2400 6 channel RAM,
  - 16GB HBM (460GB/s)
  - 1x 100Gbps OPA HFI, 1x On-board GigE Port

**CPU-only nodes**

- 132 Skylake Computing modules, 4 Racks, 0.4PF
  - 2x Xeon SKX 6148 CPUs, 2.4GHz, AVX512
  - 192GB (12x 16GB) DDR4-2666 RAM
  - 1x Single-port 100Gbps OPA HFI card
  - 1x On-board GigE (RJ45) port
KISTI-5 OPA Interconnect

KNL Compute Nodes CS500 (Qty 8,305=25.3PF)
- 1x Intel KNL 7250 (3.06TFLOPS/node peak)
- 6x 16GiB DDR4-2400 DIMM (96GiB)
- 1x Intel OPA 100Gbps HFI

CPU-only Nodes CS500 (Qty 132=0.4PF)
- 2x Intel SKL 6148 (3.072TFLOPS/node peak)
- 12x 16GiB DDR4-2666 DIMM (192GiB)
- 1x Intel OPA 100Gbps HFI

OPA Switch Group (Intel OPA, Fat-Tree, 50% Blocking)
- 8x Intel OPA 24-slot Director Switch
  (19x 32P Leaf module/switch)

Service Nodes (Qty 23)

Intel OPA 48P Edge Switch
x 277 (16Up/32Down)

DDN IME Burst Buffer
80x OPA ports

DDN Lustre ExaScaler
49x OPA ports
2:1 Blocking OPA Interconnect

- K#: KNL Compute Node (x 8305)
- S#: SKL CPU Only Node (x 132)
- V#: Service Node (x 223)
- IME#: IME Node (x40, 2x HCA/node)
- LV#: Edge Storage Service Node (x 9)
- E#: ExaScaler Storage (x 40)

Diagram showing the interconnect between different nodes and switches, with various links and connections.
## Benchmark Performance Result

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
<th># of nodes</th>
<th>Score</th>
<th>World Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPL</td>
<td>Large-scale Dense Matrix Computation Used for Top500</td>
<td>8,174(KNL)</td>
<td>13.93PF</td>
<td>31&lt;sup&gt;st&lt;/sup&gt; (Jun 2021)</td>
</tr>
<tr>
<td></td>
<td>+ 122(SKX)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPCG</td>
<td>Large-scale Sparse Matrix Computation Similar to normal user applications</td>
<td>8,250(KNL)</td>
<td>0.39PF</td>
<td>22&lt;sup&gt;nd&lt;/sup&gt; (Jun 2021)</td>
</tr>
<tr>
<td>Graph500</td>
<td>Breadth-First Search, Single-Source Shortest Paths</td>
<td>1,024(KNL)</td>
<td>1,456GTEPS</td>
<td>11&lt;sup&gt;th&lt;/sup&gt; (Jun 2021)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>337GTEPS</td>
<td>4&lt;sup&gt;th&lt;/sup&gt; (Jun 2021)</td>
</tr>
<tr>
<td>IO500</td>
<td>Various IO Workloads</td>
<td>2,048(KNL)</td>
<td>282.45</td>
<td>12&lt;sup&gt;th&lt;/sup&gt; (Jun 2021)</td>
</tr>
</tbody>
</table>
Neuron

- **GPU system in KISTI (1.24 PFlops in 2020)**
  - Intel Xeon Ivy Bridge (IVY), Skylake (SKL), NVIDIA K40, V100
  - 12 GB (K40), 16GB, 32GB (V100) GPU memory
  - Mellanox FDR, EDR InfiniBand

- **Extends system size in every year**
  - GPU servers with NVIDIA A100 in 2021 (0.93 PFlops)
  - New storage in 2021

- **Testbed for KISTI-6**
  - AMD Instinct MI100, NVIDIA A100
Performance Test on Nurion and Neuron
Performance of Applications on Nurion and Neuron

• Preliminary feasibility study for KISTI-6
• Prepare benchmark candidate and expected performance for KISTI-6
  – Update performance results with recent MVAPICH2 library
  – Do similar performance test on Neuron if possible

• Benchmarks
  – OSU Micro-Benchmarks (OMB)
  – NAS Parallel Benchmarks (NPB)

• Applications
  – Direct Numerical Simulation – Turbulent Boundary Layer (DNS-TBL)
  – Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)
  – CosmoFlow

• Experimental environment
  – Intel Fortran/C++ compiler 19.1.2
  – MVAPICH2 2.3.6
  – NVIDIA CUDA 10.0
  – MVAPICH2-GDR 2.3.4
OMB: Collective communication on Nurion

- For the verification of the MVAPICH2 2.3.6 installation
- Collective communications on 2 – 2048 nodes (Message size: 64 bytes)
- Similar performance compared to MVAPICH2 2.3.1
OMB: P2P & Collective communication on Neuron

- P2P inter-node communications on 2 GPU nodes (V100)
- Collective communications on 4 GPU nodes (V100)
- Better performance on small message compared to MVAPICH2
NPB: NAS Parallel Benchmarks on Nurion

• Benchmarks are derived from computational fluid dynamics applications

• Experimental environment
  – Evaluation on 256-1296 nodes on normal queue (cache mode)
  – Problem Class = F (Grid size: 2560 × 2560 × 2560), strong scaling
  – 1 OpenMP thread, PPN = 64, 16384-82944 MPI processes

• Performance evaluation
  – MVAPICH2 2.3.6 is 3% - 10% faster on NPB benchmark compared to MVAPICH2 2.3.1
  – MVAPICH2 2.3.6 shows better performance on the small number of nodes
  – MVAPICH2-X shows better performance on the large number of nodes (XPMEM)
DNS-TBL

- Direct Numerical Simulation – Turbulent Boundary Layer*
- Velocity solver and pressure solver (Application of turbulent flow)
- Solve the continuity and incompressible **Navier-Stokes equation**
  - Second-order finite difference scheme on the 7-point stencil
  - Discretized into hepta-diagonal matrix in 3D, and broke into 3 tridiagonal matrices
  - Transformed pressure Poisson’s equation into a single tridiagonal matrix by 2D FFT
- Optimized code for Intel® Xeon Phi™ Processor, Fortran, FFTW 3.3.7 library

DNS-TBL: Experiment Results on Nurion

- **Experimental environment**
  - Evaluation on 64-256 nodes on normal queue (cache mode)
  - 8193 X 401 X 8193 grids, 800 Reynolds number, 10 time step, strong scaling
  - 8 OpenMP thread, PPN = 8, 512-2048 MPI processes

- **Performance evaluation**
  - MVAPICH2 2.3.6 shows better performance on velocity and pressure solver
  - Similar performance results on update velocity and pressure and others
  - MVAPICH2 2.3.6 shows better performance on DNS-TBL
LAMMPS

- Large-scale Atomic/Molecular Massively Parallel Simulator*
- Molecular dynamics simulator
- Rhodo benchmark: Rhodopsin protein in solvated lipid bilayer
- Pair, Bond, K-space, Neighbor, Output, Modify, and others
- Version: 3Mar20

* S. J. Plimpton *et al*., Particle-Mesh Ewald and rRESPA for Parallel Molecular Dynamics Simulations (SIAM Conference on Parallel Processing for Scientific Computing, 1997)
**LAMMPS: Experimental Results on Nurion**

- **Experimental environment**
  - Evaluation on 2-64 nodes on normal queue (cache mode)
  - 10X10X10 with 32M atoms, strong scaling
  - 1 OpenMP thread, PPN = 64, 128-4096 MPI processes

- **Performance evaluation**
  - MVAPICH2 2.3.6 is 1% - 5% faster on LAMMPS compared to MVAPICH2 2.3.1
LAMMPS: Experimental Results on Neuron

- Experimental environment
  - Evaluation on 1-3 V100 GPU nodes
  - 8x8x8 with 16M atoms, strong scaling
  - 1 OpenMP thread, PPN = 20 for CPU part

- Performance evaluation
  - It shows some performance gain and scalability even in small problem
  - Hard to fit large problem on V100 HBM memory
CosmoFlow Benchmark on Neuron and Neuron

- MLPerf HPC v0.5
- Nurion: KNL 64-1024 nodes
- Neuron: 4x V100 1-4 nodes
- Network: 3D CNN
- Dataset: cosmoUniverse_2019_02_4 (2.1 TB)
- Local batch size: 1, 4

*A. Mathuriiya et al., CosmoFlow: using deep learning to learn the universe at scale (International Conference for High Performance Computing, Networking, Storage, and Analysis, 2018)*
Conclusion & Future Plan

• Additional performance improvements with MVAPICH2 version upgrade
  – There was performance improvements due to MPI communication improvement.
  – It could be reflected in the performance prediction of KISTI-6

• Neuron on MVAPICH2-GDR
  – Due to the system scale, MPI communication performance has less effect on the application performance compared to Nurion.
  – Additional experiment will be proceed after the system extension is completed

• Prepare benchmark candidates for KISTI-6
  – HPC in-house codes are quite difficult to test due to program porting issues
  – Although HPC application has been ported to the GPU, there is a limit to performing large-scale tasks due to the GPU memory capacity
  – ML applications is already good fit to GPU system
THANK YOU