Designing HPC Solutions

Onur Celebioglu

Dell Inc
Agenda

• HPC Focus Areas

• Performance analysis of HPC Components
  – Compute
  – Interconnect
  – Accelerators
  – And many more

• Best Practices

• Designing better HPC solutions
  – Domain specific Appliances
HPC at Dell

• Evaluate new HPC technologies and selectively adopt for Integration
• Share our findings with the broader HPC community.
• Analyze decision points to obtain the optimal solution to the problem at hand.

• Decision Points include but not limited to
  – Compute Performance
  – Memory Performance
  – Interconnect
  – Accelerators
  – Storage
  – Power / Energy Efficiency
  – Software Stack
  – Middleware

• Focus Areas
  – Define best practices by analyzing each and every component of an HPC cluster
  – Use these best practices to develop plug and play solutions targeted at specific HPC verticals such as Life sciences, Fluid Dynamics, High frequency trading etc.
Compute, Memory & Energy Efficiency
12G – Optimal BIOS Settings

<table>
<thead>
<tr>
<th>System Profile</th>
<th>Balanced configuration</th>
<th>Performance focused</th>
<th>Energy Efficient configuration</th>
<th>Latency sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Power Mgmt</td>
<td>Performance Per Watt</td>
<td>Performance focused</td>
<td>Energy Efficient configuration</td>
<td>Latency sensitive</td>
</tr>
<tr>
<td>Turbo Boost</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Disabled</td>
<td>Disabled</td>
</tr>
<tr>
<td>C States &amp; C1E</td>
<td>Enabled</td>
<td>Disabled</td>
<td>Enabled</td>
<td>Disabled</td>
</tr>
<tr>
<td>Monitor/ Mwait</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Disabled</td>
</tr>
<tr>
<td>Logical Processor</td>
<td>Disabled</td>
<td>Disabled</td>
<td>Disabled</td>
<td>Disabled</td>
</tr>
<tr>
<td>Node Interleave</td>
<td>Disabled</td>
<td>Disabled</td>
<td>Disabled</td>
<td>Disabled</td>
</tr>
</tbody>
</table>
Ivy Bridge vs. Sandy Bridge Single Node

Performance Gain with Ivy Bridge (12 core) over Sandy Bridge (8 core)

- HPL: 46%
- ANSYS Fluent: 25%
- LS-DYNA: 16%
- Simulia Abaqus 6.1.2- S4B: 3%
- Simulia Abaqus 6.12- E6: 12%
- LAMMPS: 26%
- MUMPS: 37%

- E5-2670 8C 2.6 Ghz (SB) vs E5-2697 V2 12C 2.7 GHz (IVB)
Decision: Processor selection. Criteria: Performance

Performance across four nodes using multiple IVB processors

- 2 x E5-2697-v2 @ 2.7 GHz 12c 130W does the best in most cases.
- All tests done on fully subscribed 4 servers with FDR interconnect.
Decision: Processor selection. Criteria: Power

Energy efficiency across four nodes using multiple IVB processors

- 2 x E5-2697-v2 @ 2.7 GHz 12c 130W does the best in most cases.
- All tests done on fully subscribed 4 servers with FDR interconnect.
Decision: Memory selection. Criteria: Performance

- Dual Rank memory modules give best performance
- All tests done on fully subscribed 4 servers with FDR interconnect.
Interconnect Performance
OSU Latency and Bandwidth (FDR vs 40 GigE RoCE)

- How do benchmarks, synthetic kernels and micro benchmarks behave at scale?
- Can micro benchmark performance explain application’s performance at a larger scale?
RoCE vs IB vs TCP

MVAPICH2-2.0b
RoCE vs IB vs TCP

WRF Conus 12Km

Nodes - Cores

Performance relative to FDR (Rating)

1.00 1.00 1.02 0.94 0.94 0.84 0.83

FDR 40GigE-ROCE 40GigE-TCP

MILC Intel Data set

Nodes - Cores

Performance relative to FDR (Rating)

1.00 1.00 0.98 0.98 0.67 0.59 0.49 0.29 0.30 0.29 0.32 0.40 0.57 0.40

FDR 40GigE-ROCE 40GigE-TCP

MVAPICH2-2.0b WRF 3.5 MILC 7.6.2
Interconnect Summary

- InfiniBand is still performs higher than other network fabrics in this study for HPC workloads
- For some workloads, RoCE performs similar to InfiniBand and may be a viable alternative.
  - Haven’t seen wide adoption of RoCE in production yet.
  - Mileage will vary based on application’s communication characteristics
  - Needs switches with DCB support for optimal lossless performance.
- Ethernet with TCP/IP stops scaling after 4-8 nodes.
Accelerator Performance
Power and Performance: K20 vs K40

HPL performance on single-node

Power & energy efficiency of an eight node cluster.
MV2 performance with GPU Direct : OMB

Device to Device Latency

Small message latency

- 62 %

Small message Bandwidth

- 2.6x

Intel SandyBridge (E5-2670), NVIDIA Tesla K20m GPU, Mellanox ConnectX-3 FDR, CUDA 6.0, OFED 2.2-1.0.0 with GPU Direct RDMA Beta
Domain specific solutions
Dell Genomic Analysis Platform

- Mellanox SX6036 IB Switch
- Dell PowerEdge R320
- HSS 4.5 Management Nodes
- 2 x Dell PowerEdge R420 BCM Login Nodes
- Dell PowerEdge R820
- Fat Node With 1.5 TB Memory
- Dell PowerVault MD3260
- NSS 4.5 HA Storage
- Dell PowerVault MD3220
- HSS 4.5 Lustre Metadata Storage Target
- Dell PowerVault MD3250
- HSS 4.5 Object Storage Target
- Dell PowerEdge M1000e Chassis With 32 Dell PowerEdge M420s
- Dell Force10 S55 Switch
- Dell PowerEdge R320 CIFS Gateway
- 2 x Dell PowerEdge R420
- BCM Head Nodes
- 2 x Dell PowerEdge R620
- NSS 4.5 HA Head Nodes
- 2 x Dell PowerEdge R620
- HSS 4.5 Lustre Metadata Servers
- 2 x Dell PowerEdge R620
- HSS 4.5 Lustre Object Storage Servers
- Dell PowerVault MD3260
- HSS 4.5 Object Storage Target
Dell Genomic Analysis Platform (Continued)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Results and Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken for analyzing 30 samples</td>
<td>19.5 Hours</td>
</tr>
<tr>
<td>Energy Consumption for analyzing 30 samples</td>
<td>222.77 kWh</td>
</tr>
<tr>
<td>kWh/Genome</td>
<td>7.42 kWh / Genome</td>
</tr>
<tr>
<td>Genomes/day</td>
<td>37</td>
</tr>
</tbody>
</table>

Advantages

- Metrics relevant to the domain instead of GFLOPs.
- Energy Efficient
- Plug and Play
- Scalability
- What used to take 2 weeks now takes less than 4 hours.
- More to follow..
Collateral
Future Work and Potential Areas of Research

- Deployment tools
- Use of virtualization and cloud (Openstack) in HPC
  - Linux Containers and Docker
- Hadoop
- Lustre FS
- Accelerators
Storage Blogs

- HTSS + DX Object Storage
  › [http://dell.to/zJqiTK](http://dell.to/zJqiTK)
- Dell HPC NFS Storage Solution with High Availability -- Large Capacity Configuration
  › [http://dell.to/GYWU5x](http://dell.to/GYWU5x)
- Dell support for XFS greater than 100 TB
  › [http://dell.to/GUjXRq](http://dell.to/GUjXRq)
- NSS-HA 12G Performance Intro
  › [http://dell.to/NFUafG](http://dell.to/NFUafG)
- NSS4.5-HA Solution Configurations
  › [http://dell.to/10xLxJV](http://dell.to/10xLxJV)
- Dell Fluid Cache for DAS performance with NFS
  › [http://dell.to/15KnsDc](http://dell.to/15KnsDc)
- Achieving over 100000 IOPs with NFS Async
  › [http://dell.to/16yE3bP](http://dell.to/16yE3bP)
- Dell | Terasca HPC Storage Solution Part I
- Dell | Terasca HPC Storage Solution Part 2
- DT-HSS3 Performance and Scalability
Storage Blogs Continued

- Dell | Terascale HPC Storage Solution - HSS5
  - [http://dell.to/1gpVVyN](http://dell.to/1gpVVyN)

- NSS overview

- NSS-HA overview

- NSS-HA XL configuration

- Dell HPC NFS Storage Solution - High Availability Solution NSS5-HA configurations
  - [http://dell.to/1eZU0xL](http://dell.to/1eZU0xL)
Copro Processor Acceleration Blogs

- GPUDirect Improves Communication Bandwidth Between GPUs on the C410X
  › http://dell.to/AnLz5
- Comparing GPU-Direct Enabled Communication Patterns for Oil and Gas Simulations
  › http://dell.to/JsWqWT
- Accelerating ANSYS Mechanical Simulations with M2090 GPU on the R720
  › http://dell.to/JT79KF
- Accelerating High Performance Linpack (HPL) with GPUs
  › http://dell.to/MrYw8q
- Faster Molecular Dynamics with GPUs
  › http://dell.to/PEaFaF
- Deploying and Configuring Intel Xeon Phi Coprocessor with HPC Solution
  › http://dell.to/14GtFRv
Best Practices Blogs

- 12G HPC Solution with ROCKS+ from StackIQ
  - http://dell.to/xGmSHO

- HPC mode on Dell PowerEdge R815 with AMD 6200 Processors
  - http://dell.to/MMGG4s

- Optimal BIOS settings for HPC workloads
  - http://dell.to/PkkMG1

- CFD Primer
  - http://dell.to/UwJQum

- OpenFOAM
  - http://dell.to/Rga3hS

- PowerEdge M420 with single Force10 MXL Switch
  - http://dell.to/Zjnhjz

- Active Infrastructure for HPC Life Sciences
  - http://dell.to/18eaDSJ

- Dell HPC Solution Refresh: Intel Xeon Ivy Bridge-EP, 1866 DDR3 memory and RHEL 6.4
  - http://dell.to/18U3Aki
Performance Blogs

- HPC I/O performance using PCI-E Gen3 slots on the 12th Generation (12G) PowerEdge Servers
  › [http://dell.to/wzdV0x](http://dell.to/wzdV0x)

- HPC performance on the 12th Generation (12G) PowerEdge Servers
  › [http://dell.to/zozohn](http://dell.to/zozohn)

- Unbalanced Memory Configuration Performance
  › [http://dell.to/UQ1kQu](http://dell.to/UQ1kQu)

- Performance analysis of HPC workloads
  › [http://dell.to/STbE8q](http://dell.to/STbE8q)
Questions?