How the Results of Summit and Sierra are Influencing Exascale

Al Geist
Oak Ridge National Laboratory

HPC Forum
September 9-11, 2019
ORNL Summit: IBM System Overview

System Performance
- Peak of 200 Petaflops (FP\textsubscript{64}) for modeling & simulation
- Peak of 3.3 ExaOps (FP\textsubscript{16}) for data analytics and artificial intelligence

The system includes
- 4,608 nodes
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM file system transferring data at 2.5 TB/s

Each node has
- 2 IBM POWER9 processors
- 6 NVIDIA Tesla V100 GPUs
- 608 GB of fast memory (96 GB HBM2 + 512 GB DDR4)
- 1.6 TB of non-volatile memory
LLNL Sierra: IBM System Overview

System Performance

- Peak of 125 Petaflops (FP\textsubscript{64}) for modeling & simulation
- Peak of 2.0 ExaOps (FP\textsubscript{16}) for data analytics and artificial intelligence

The system includes

- 4,320 nodes
- Single plane Mellanox EDR InfiniBand network w/ 2 to 1 tapered Fat Tree
- 154 PB IBM file system transferring data at 1.5 TB/s
- 2 IBM POWER9 processors
- 4 NVIDIA Tesla V100 GPUs
- 576 GB of fast memory (64 GB HBM2 + 512 GB DDR4)
- 1.6 TB of non-volatile memory

(Same parts different architecture)
Summit Node
- 3 GPU per CPU
- Coherent memory across entire node
- 1.6 TB of on-node NVM

Sierra Node
- 2 GPU per CPU
- 75 GB/s GPU-CPU links

Titan Node
- 1 GPU per CPU
  AMD CPU / Nvidia GPU

6 GPU, 2 CPU Summit Node

TF 42 TF (6x7 TF)
HBM 96 GB (6x16 GB)
DRAM 512 GB (2x16x16 GB)
NET 25 GB/s (2x12.5 GB/s)
MMmsg/s 83

HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.
Summit has 27,648 NVIDIA Volta GPUs each with optimized AI Performance

Each Volta GPU can perform:

- 7.5 FP\textsubscript{64} TFLOPS | 15 FP\textsubscript{32} TFLOPS | 120 FP\textsubscript{16} TFLOPS
- Tensor cores do mixed precision multiply-add of 4x4 matrices

\[
D = \begin{pmatrix}
A_{0,0} & A_{0,1} & A_{0,2} \\
A_{1,0} & A_{1,1} & A_{1,2} \\
A_2,0 & A_{2,1} & A_{2,2}
\end{pmatrix}
+ \begin{pmatrix}
B_{0,0} & B_{0,1} & B_{0,2} \\
B_{1,0} & B_{1,1} & B_{1,2} \\
B_{2,0} & B_{2,1} & B_{2,2}
\end{pmatrix}
+ \begin{pmatrix}
C_{0,0} & C_{0,1} & C_{0,2} \\
C_{1,0} & C_{1,1} & C_{1,2} \\
C_{2,0} & C_{2,1} & C_{2,2}
\end{pmatrix}
\]

- The Modeling & Simulation community can benefit by utilizing mixed / reduced precision algorithms
- AI community can do ML training at 120 FP\textsubscript{16} TFLOPs
As supercomputers got larger and larger, we expected them to be more specialized and limited to just a small number of applications that can exploit their growing scale.

Summit’s architecture with powerful, multiple-GPU nodes with huge memory per node seems to have stumbled into a design that has broad capability across:

- Traditional HPC modeling and simulation
- High performance data analytics
- Artificial Intelligence
Summit Excels Across Simulation, Analytics, AI

- Data analytics – CoMet bioinformatics application for comparative genomics. Used to find sets of genes that are related to a trait or disease in a population. Exploits cuBLAS and Volta tensor cores to solve this problem 5 orders of magnitude faster than previous state-of-art code.
  - **Has achieved 2.36 ExaOps** mixed precision (FP_{16}-FP_{32}) on Summit
- Deep Learning – global climate simulations use a half-precision version of the DeepLabv3+ neural network to learn to detect extreme weather patterns in the output
  - **Has achieved a sustained throughput of 1.0 ExaOps (FP_{16})** on Summit
- Nonlinear dynamic low-order unstructured finite-element solver accelerated using mixed precision (FP_{16} thru FP_{64}) and AI generated preconditioner. Answer in FP_{64}
  - **Has achieved 25.3 fold speedup** on Japan earthquake – city structures simulation
- Half-dozen apps >25x speedup on Summit vs. Titan (a couple around 100x speedup)
Summit Displays Its Balanced Design
Achieves #1 on TOP500, #1 on HPCG, and #1 Green500 (level 3)

122 PF HPL
Shows DP performance

2.9 PF HPCG
Shows fast data movement

13.889 GF/W
Shows energy efficiency
What Makes Summit Architecture Better Than Titan? (These same concepts are being carried over into Frontier)

- Many fewer nodes
- Much more powerful nodes
- Much more memory per node and higher memory bandwidth
- Faster interconnect
- Much higher bandwidth between CPUs and GPUs
- Much larger and faster file system
- 7x more performance for only slightly more power (HPL 122 PF run was 8.8 MW)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Titan</th>
<th>Summit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak FLOPS&lt;sub&gt;64&lt;/sub&gt;</td>
<td>27 PF</td>
<td>200 PF</td>
</tr>
<tr>
<td>Max possible Power</td>
<td>9 MW</td>
<td>13 MW</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>18,688</td>
<td>4,608</td>
</tr>
<tr>
<td>Node performance</td>
<td>1.4 TF</td>
<td>42 TF</td>
</tr>
<tr>
<td>Memory per Node</td>
<td>32 GB DDR3</td>
<td>512 GB DDR4</td>
</tr>
<tr>
<td></td>
<td>6 GB GDDR5</td>
<td>96 GB HBM2</td>
</tr>
<tr>
<td>NV memory per Node</td>
<td>0</td>
<td>1.6 TB</td>
</tr>
<tr>
<td>Total System Memory</td>
<td>0.7 PB</td>
<td>2.8 PB + 7.4 PB NVM</td>
</tr>
<tr>
<td>System Interconnect</td>
<td>Gemini (6.4 GB/s)</td>
<td>Dual Rail EDR (25 GB/s)</td>
</tr>
<tr>
<td>Interconnect Topology</td>
<td>3D Torus</td>
<td>Non-blocking Fat Tree</td>
</tr>
<tr>
<td>Bi-Section Bandwidth</td>
<td>15.6 TB/s</td>
<td>115.2 TB/s</td>
</tr>
<tr>
<td>Processors on node</td>
<td>1 AMD Opteron™ 1 NVIDIA Kepler™</td>
<td>2 IBM POWER9™ 6 NVIDIA Volta™</td>
</tr>
<tr>
<td>File System</td>
<td>32 PB, 1 TB/s, Lustre®</td>
<td>250 PB, 2.5 TB/s, GPFS™</td>
</tr>
</tbody>
</table>
Frontier Continues the Accelerated Node Design begun at ORNL with Titan and continued with Summit

Frontier Node -- 4 GPUs per CPU:
- One purpose-built AMD EPYC™ processor
- Four HPC and AI optimized AMD Radeon Instinct™ GPUs
- Fully connected with high speed AMD Infinity Fabric links
- Coherent memory across the node
- 100 GB/s node injection bandwidth
- On-node NVM storage

Partnership between ORNL, Cray, and AMD
The Frontier system will be delivered in 2021
Peak FP\textsubscript{64} Performance greater than 1.5 EF
Max Power Consumption 29 MW
Cray Shasta cabinets Connected by Slingshot™ interconnect
• with adaptive routing, congestion control, and quality of service
Questions?

ORNL / Cray / AMD Partnership