HPC Hardware Overview

John Cazes
September 20, 2013

Texas Advanced Computing Center
The University of Texas at Austin
Outline

• **Lonestar**
  – Dell blade-based system
  – InfiniBand (QDR)
  – Intel Processors

• **Longhorn**
  – Dell blade-based system
  – InfiniBand (QDR)
  – Intel Processors

• **Stampede**
  – Dell
  – Infiniband (FDR)
  – Intel Processors

• **Ranch & Corral** - Storage Systems
About this Talk

• As an applications programmer you may not care about hardware details, but...
  – We need to consider performance issues
    • Better performance means faster turnaround and/or larger problems
  – We will focus on the most relevant architecture characteristics

• Do not hesitate to ask questions as we go
High Performance Computing

• In our context, it refers to hardware and software tools dedicated to computationally intensive tasks

• Distinction between HPC center (throughput focused) and Data center (data focused) is becoming fuzzy

• High bandwidth, low latency
  – Memory
  – Network
Lonestar: Intel hexa-core system
Lonestar: Introduction

- Lonestar Cluster
  - Configuration & Diagram
  - Server Blades
- Dell PowerEdge M610 Blade (Intel Hexa-Core) Server Nodes
- Microprocessor Architecture Features
  - Instruction Pipeline
  - Speeds and Feeds
  - Block Diagram
- Node Interconnect
  - Hierarchy
  - InfiniBand Switch and Adapters
  - Performance
Lonestar Cluster Overview

lonestar.tacc.utexas.edu
# Lonestar Cluster Overview

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Components</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Performance</td>
<td>302 TFLOPS</td>
<td></td>
</tr>
<tr>
<td>Nodes</td>
<td>2 Hexa-Core Xeon 5680</td>
<td>1888 nodes / 22656 cores</td>
</tr>
<tr>
<td>Memory</td>
<td>1333 MHz DDR3 DIMMS</td>
<td>24 GB/node, 45 TB total</td>
</tr>
<tr>
<td>Shared Disk</td>
<td>Lustre parallel file system</td>
<td>1 PB</td>
</tr>
<tr>
<td>Local Disk</td>
<td>SATA</td>
<td>146GB/node, 276 TB total</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Infiniband</td>
<td>4 GB/sec P-2-P</td>
</tr>
</tbody>
</table>
Blade : Rack : System

- 1 node : 2 x 6 cores = 12 cores
- 1 chassis : 16 nodes = 192 cores
- 1 rack : 3 x 16 nodes = 576 cores
- 39⅓ racks : 22656 cores
Lonestar login nodes

• Dell PowerEdge M610
  – Intel dual socket Xeon hexa-core 3.33GHz
  – 24 GB DDR3 1333 MHz DIMMS
  – Intel QPI 5520 Chipset

• Dell 1200 PowerVault
  – 15 TB HOME disk
  – 1 GB user quota(5x)
Lonestar compute nodes

- 16 Blades / 10U chassis
- Dell PowerEdge M619
  - Dual socket Intel hexa-core Xeon
    - 3.33 GHz (1.25x)
    - 13.3 GFLOPS/core (1.25x)
    - 12 MB L3 cache (unified)
  - 24 GB DDR3 1333 MHz DIMMS
  - Intel QPI 5520 Chipset
  - 2x QPI 6.4 GT/s
  - 146 GB 10k RPM SAS-SATA local disk (/tmp)
Intel Xeon 5600 (Westmere)

- 32 KB L1 cache/core
- 256 KB L2 cache/core
- Shared 12 MB L3 cache
Cluster Interconnect

16 independent 4 GB/s connections/chassis
Lonestar Parallel File Systems: Lustre & NFS

- Lonestar Parallel File Systems
- Lustre & NFS
- Uses IP over IB
- InfiniBand
- Ethernet

$HOME
97 TB
1 GB/user

$WORK
226 TB
200 GB/user

$SCRATCH
806 TB

The University of Texas at Austin
Texas Advanced Computing Center
Longhorn: Intel Quad-core system
Longhorn: Introduction

• First NSF eXtreme Digital Visualization grant (XD Vis)
• Designed for scientific visualization and data analysis
  – Very large memory per computational core
  – Two NVIDIA Graphics cards per node
  – Rendering performance 154 billion triangles/second
## Longhorn Cluster Overview

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Components</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Performance (CPUs)</td>
<td>512 Intel Xeon E5400</td>
<td>20.7 TFLOPS</td>
</tr>
<tr>
<td>Peak Performance (GPUs, SP)</td>
<td>128 NVIDIA Quadroplex 2200 S4</td>
<td>500 TFLOPS</td>
</tr>
<tr>
<td>System Memory</td>
<td>DDR3-DIMMS</td>
<td>48 GB/node (240 nodes)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>144 GB/node (16 nodes)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13.5 TB total</td>
</tr>
<tr>
<td>Graphics Memory</td>
<td>512 FX5800 x 4GB</td>
<td>2 TB</td>
</tr>
<tr>
<td>Disk</td>
<td>Lustre parallel file system</td>
<td>210 TB</td>
</tr>
<tr>
<td>Interconnect</td>
<td>QDR Infiniband</td>
<td>4 GB/sec P-2-P</td>
</tr>
</tbody>
</table>
Longhorn “Large Mem” nodes

• Dell R710
  – Intel dual socket quad-core Xeon E5400 @ 2.53GHz
  – 144 GB DDR3 (18 GB/core)
  – Intel 5520 chipset

• NVIDIA Quadroplex 2200 S4
  – 4 NVIDIA Quadro FX5800
    • 240 CUDA cores
    • 4 GB Memory
    • 102 GB/s Memory bandwidth
Longhorn standard nodes

- Dell R610
  - Dual socket Intel quad-core Xeon E5400 @ 2.53 GHz
  - 48 GB DDR3 (6 GB/core)
  - Intel 5520 chipset

- NVIDIA Quadroplex 2200 S4
  - 4 NVIDIA Quadro FX5800
    - 240 CUDA cores
    - 4 GB Memory
    - 102 GB/s Memory bandwidth
Motherboard (R610/R710)

12 DIMM slots (R610)
18 DIMM slots (R710)

Quad Core Intel Xeon 5500

Quad Core Intel Xeon 5500

Intel 5520

42 lanes PCI Express or 36 lines PCI Express 2.0
Cache Sizes in Intel Nehalem

- L2: 256 KB
- L1: 32 KB

Memory Controller

- QPI 0
- QPI 1

L3: 8 MB

- Core 1
- Core 2
- Core 3
- Core 4

• Total QPI bandwidth up to 25.6 GB/s (@ 3.2 GHz)
• 2 20-lane QPI links
• 4 quadrants (10 lanes each)
Nehalem μArchitecture
Stampede: Intel 8 core system
Stampede

- Enable sustained petascale computational and data-driven science and engineering and provide an “innovative component”
Dell/Intel Partnership

- TACC Partnered with Dell and Intel to design Stampede
- Intel MIC (Intel Xeon Phi) is the innovative component
  - High performance and low power per operation
  - Highly programmable
Datacenter Expansion
TEXAS ADVANCED COMPUTING FACILITY
THE UNIVERSITY OF TEXAS

Thermal Storage
New Infrastructure
New Machine Room
TEXAS ADVANCED COMPUTING FACILITY
THE UNIVERSITY OF TEXAS

Thermal Storage
1M gallons
3,750 Tons
14KV in, 480V out

Chillers
Power & Pumps

6.5MW
~3,000 GPM
10 PFLOPS

1M gallons
3,750 Tons
14KV in, 480V out

2x footprint, 2x power, 20x Capability (Stampede vs. Ranger)
Stampede consists of 182 48U cabinets.
Key Datacenter Features

• Thermal energy storage to reduce peak power consumption
• Hot aisle containment to boost efficiency
• Datacenter power increased to ~12MW
• Expand experiments in mineral oil cooling
Cooling and Electrical Infrastructure
Stampede Performance

Stampede debuted at #7 on the Top 500
Stampede Overview

• $27.5M acquisition
• 9+ petaflops (PF) peak performance
• 2+ PF Linux cluster
  – 6400 Dell DCS C8220X nodes
  – 2.7GHz Intel Xeon E5 (Sandy Bridge)
    • 102,400 total cores
  – 56Gb/s FDR Mellanox InfiniBand
  – 7+ PF Intel Xeon Phi Coprocessor
    • TACC has a special release: Intel Xeon Phi SE10P
  – 14+ PB disk, 150GB/s
  – 16 1TB shared memory nodes
  – 128 NVIDIA Tesla K20 GPUs
## Processor Specs

<table>
<thead>
<tr>
<th>Arch. Features</th>
<th>Xeon E5</th>
<th>Xeon Phi SE10P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>2.7GHz +turbo</td>
<td>1.0GHz +turbo</td>
</tr>
<tr>
<td>Cores</td>
<td>8</td>
<td>61</td>
</tr>
<tr>
<td>HW threads/core</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Vector size</td>
<td>256 bits 4 doubles 8 singles</td>
<td>512 bits 8 doubles 16 singles</td>
</tr>
<tr>
<td>Instr. Pipeline</td>
<td>Out of Order</td>
<td>In Order</td>
</tr>
<tr>
<td>Registers</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Caches</td>
<td>L1:32KB</td>
<td>L1:32KB</td>
</tr>
<tr>
<td></td>
<td>L2:256KB</td>
<td>L2:512KB</td>
</tr>
<tr>
<td></td>
<td>L3:20MB</td>
<td></td>
</tr>
<tr>
<td>Memory</td>
<td>2 GB/core</td>
<td>128 MB/core</td>
</tr>
<tr>
<td>Sustained Memory BW</td>
<td>75 GB/s</td>
<td>170 GB/s</td>
</tr>
<tr>
<td>Sustain Peak FLOPS</td>
<td>1 thread/core</td>
<td>2 threads/core</td>
</tr>
<tr>
<td>Instruction Set</td>
<td>x86 + AVX</td>
<td>x86 + new vector instructions</td>
</tr>
</tbody>
</table>
MIC Details
What is a MIC

• Basic Design Ideas
  – Leverage x86 architecture (CPU with many cores)
    • X86 cores are simpler, but allow for more compute throughput
  – Leverage existing x86 programming models
  – Dedicate much of the silicon to floating point ops
  – Cache coherent
  – Increase floating-point throughput
  – Implement as a separate device
  – Strip expensive features (out-of-order execution, branch prediction, etc.)
  – Widen SIMD registers for more throughput
  – Fast (GDDR5) memory on card
  – Runs a full Linux operating system (BusyBox)
MIC Architecture

- Many cores on the die
- L1 and L2 cache
- Bidirectional ring network
- Memory and PCIe connection
# Stampede Login Nodes

<table>
<thead>
<tr>
<th>Component</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 login nodes</td>
<td>stampede.tacc.utexas.edu</td>
</tr>
<tr>
<td>Processors Sockets per Node/Cores per Socket</td>
<td>(2) E5-2680, 2.7 GHz – 8 cores</td>
</tr>
<tr>
<td>Motherboard</td>
<td>Dell R720, Intel QPI 600 Chipset</td>
</tr>
<tr>
<td>Memory Per Node</td>
<td>32GB DDR3-1333 MHz</td>
</tr>
<tr>
<td></td>
<td>Cache: 256KB/core L2: 20MB</td>
</tr>
<tr>
<td>$HOME /tmp</td>
<td>Global: Lustre, 5GB quota</td>
</tr>
<tr>
<td></td>
<td>Local: Shared, 432GB SATA 10K rpm</td>
</tr>
</tbody>
</table>
# Dell DCS C8220z Compute Node

<table>
<thead>
<tr>
<th>Component</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sockets per Node/Cores per Socket</td>
<td>2/8 Xeon E5-2680 2.7GHz (turbo, 3.5)</td>
</tr>
<tr>
<td>Coprocessors/Cores</td>
<td>1/61 Xeon Phi SE10P 1.1GHz</td>
</tr>
<tr>
<td>Motherboard</td>
<td>Dell C8220, Intel PQI, C610 Chipset</td>
</tr>
<tr>
<td>Memory Per Host</td>
<td>32GB 8x4GB 4 channels DDR3-1600MHz</td>
</tr>
<tr>
<td>Memory per Coprocessor</td>
<td>8GB DDR5</td>
</tr>
<tr>
<td>Interconnect Processor-Processor</td>
<td>QPI 8.0 GT/s</td>
</tr>
<tr>
<td>Processor-Coprocessor</td>
<td>PCI-e</td>
</tr>
<tr>
<td>PCI Express Processor</td>
<td>x40 lanes, Gen 3</td>
</tr>
<tr>
<td>PCI Express Coprocessor</td>
<td>x16 lanes, Gen 2 (extended)</td>
</tr>
<tr>
<td>250GB Disk</td>
<td>7.5 RPM SATA</td>
</tr>
</tbody>
</table>
Compute Node Configuration

CPUs and MIC appear as separate HOSTS (“symmetric” computing)
### Stampede Filesystems

<table>
<thead>
<tr>
<th>Storage Class</th>
<th>Size</th>
<th>Architecture</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local (each node)</td>
<td>Login: 1TB</td>
<td>SATA</td>
<td>432GB mounted on /tmp</td>
</tr>
<tr>
<td></td>
<td>Compute: 250GB</td>
<td>SATA</td>
<td>80GB mounted on /tmp</td>
</tr>
<tr>
<td></td>
<td>Big Mem: 600 GB</td>
<td>SATA</td>
<td>398GB mounted on /tmp</td>
</tr>
<tr>
<td>Parallel</td>
<td>14PB</td>
<td>Lustre</td>
<td>72 Dell R610 (OSS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 Dell R710 (MDS)</td>
</tr>
<tr>
<td>Ranch (Tape Storage)</td>
<td>60PB</td>
<td>SAM-FS</td>
<td>10GB/s connection through 4 GridFTP Servers</td>
</tr>
</tbody>
</table>
Stampede Filesystems

- **$HOME**
  - Quota: 5GB, 150K files
  - Filesystem is backed up

- **$WORK**
  - Quota: 400GB, 3M files
  - NOT backed up
  - Use `cdw` to change to $WORK

- **$SCRATCH**
  - No Quota
  - Total size ~8.5PB
  - NOT backed up
  - Use `cds` to change to $SCRATCH
  - Files older than 10 days are subject to purge policy

- **/tmp**
  - Local disk
  - ~80GB
Large Memory and Visualization Nodes

• 16 Large Memory Nodes
  – 32 cores
  – 1TB of memory
  – Used for data-intense applications requiring disk caching and large memory methods

• 128 Visualization Nodes
  – 16 cores
  – NVIDIA Tesla K20 with 8GB GDDR5 memory
Storage

• **Ranch** – Long term tape storage
• **Corral** – 1 PB of spinning disk
Ranch Archival System

- Sun StorageTek Silo
  - 10,000 - T10000B tapes
  - 6,000 - T10000C tapes
  - ~40 PB total capacity
  - 640 TB Disk cache
Corral

- 6PB DataDirect Networks online disk storage
- 8 Dell 1950 servers
- 8 Dell 2950 servers
- 12 Dell R710 servers
- High Performance Parallel File System
  - Multiple databases
  - iRODS data management
  - Replication to tape archive
- Multiple levels of access control
- Web and other data access available globally
References
Lonestar Related References

User Guide
www.tacc.utexas.edu/user-services/user-guides/lonestar-user-guide

Developers
Longhorn Related References

User Guide
www.tacc.utexas.edu/user-services/user-guides/lonestar-user-guide

General Information
www.intel.com/technology/architecture-silicon/next-gen/
www.nvidia.com/object/cuda_what_is.html

Developers
developer.intel.com/design/pentium4/manuals/index2.htm
developer.nvidia.com/forums/index.php
Stampede Related References

User Guide
http://www.tacc.utexas.edu/user-services/user-guides/stampede-user-guide

General Information
John Cazes
cazes@tacc.utexas.edu

For more information:
www.tacc.utexas.edu