Welcome to Stampede Training

Carlos Rosales
Texas Advanced Computing Center
carlos@tacc.utexas.edu
Thanks for Coming!

- Stampede is an exciting new system of incredible power.
- The next in a growing line of leadership systems at TACC
  - Ranger
  - Lonestar
  - Longhorn
TACC Mission & Strategic Approach

To enable discoveries that advance science and society through the application of advanced computing technologies.

– Resources & Services
  • Evaluate, acquire & operate world-class resources
  • Provide expert support via leading technology expertise

– Research & Development
  • Produce new computational technologies and techniques
  • Collaborate with researchers to apply advanced computing technologies in science projects
TACC Technology Focus Areas

• High Performance Computing (HPC)
  – Applications
  – Performance and Architectures
  – Software Tools

• Visualization
  – Scalable Visualization Technologies
  – Visualization Interfaces and Technologies

• Data Intensive Computing
  – Data Management and Collections
  – Data Mining and Statistics

• Advanced Computing Interfaces
  – Web and Cloud Services
  – Web and Mobile Applications
# Current TACC HPC/DATA Systems

<table>
<thead>
<tr>
<th>System</th>
<th>Stampede</th>
<th>Lonestar</th>
<th>Longhorn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>HPC</td>
<td>HPC</td>
<td>Data Analysis</td>
</tr>
<tr>
<td>Nodes</td>
<td>6,400</td>
<td>1,888</td>
<td>256</td>
</tr>
<tr>
<td>CPU/node x cores/CPU</td>
<td>2 x 8 + MIC</td>
<td>2 x 6</td>
<td>2 x 4 + 2GPUs</td>
</tr>
<tr>
<td>Total cores</td>
<td>102,400</td>
<td>22,656</td>
<td>2,048</td>
</tr>
<tr>
<td>CPUs</td>
<td>Intel Sandy Bridge 2.7GHz + Intel Xeon Phi SE10P</td>
<td>Intel Westmere 3.3GHz</td>
<td>Intel Nehalem +NVIDIA 2.5 GHz +Quadro Plex S4s</td>
</tr>
<tr>
<td>Memory</td>
<td>2GB/core</td>
<td>2GB/core</td>
<td>6GB/core (240 nodes) 18GB/core (16 nodes)</td>
</tr>
<tr>
<td>Interconnect</td>
<td>FDR IB</td>
<td>QDR IB</td>
<td>QDR IB</td>
</tr>
<tr>
<td>Disk</td>
<td>14PB Lustre (IB)</td>
<td>1PB Lustre (IB)</td>
<td>0.2PB Lustre (10GigE)</td>
</tr>
</tbody>
</table>
Storage Systems

High Speed Disk -- **Corral**
- 6 PB Data Direct Disk
- 5 PB, replicated GPFS
- 800TB Lustre File System
- 200TB Data Collections
- InfiniBand interconnect

Tape Storage -- **Ranch**
- 40PB capacity
- 640 TB cache
- 10Gb Ethernet interconnect
TACC Advanced Visualization Systems

- **Upgraded ACES Vislab**
  - 16x5 Tiled Display Wall, 328 MPixels, Nvidia GPUs
  - SONY 9MPixel Projector, 20ft x 11ft display
  - High-end Dell Workstations
  - Collaboration/conference room
  - Tiled touch display

- **Integrated visualization for remote sessions**
  - Lonestar: 16 GPU nodes
  - Stampede: 128 GPU nodes + 16 large shared memory nodes with GPUs

- **Longhorn**
  - 256-node, 512-GPU system for remote vis and HPC on GPUs
The Stampede Project

• A Complete Advanced Computing Ecosystem:
  – HPC Compute, storage, interconnect
  – Visualization subsystem (128 NVIDIA Kepler GPUs)
  – Large memory support (16 1TB nodes)
  – Integration with archive, and (coming soon) TACC global work filesystem and other data systems
  – People, training, and documentation to support computational science
• Hosted at an expanded building in TACC; massive power upgrades
  – 12MW new total datacenter and cooling power (5MW in use)
  – Thermal energy storage to reduce operating costs
• Will effectively replace Ranger (TACC) and Kraken (NICS) HPC systems
  – They expired in February 2013 (Ranger), and will expire in April 2014 (Kraken)
• It is now officially accepted, including the innovative component (MIC)
Stampede - Headlines

- **Initial System:**
  - 6,400 Dell C8220X nodes.
  - 102,400 cores of Intel Xeon E5 processors (dual socket nodes, 8 core processors)
  - 2GB per core, 200TB RAM in base system
- 6,880 Intel Xeon Phi SE10P Coprocessors
  - 61 cores, 244 threads per card (>1.5M threads).
  - 8GB per card, 50TB additional RAM
- **14PB of disk**
  - 76 Dell “Scorpion” chassis, 4,864 3TB drives
  - Lustre filesystems
- **FDR Infiniband** (56 Gbps, <1.2us latency) interconnected provided by Mellanox (Fat tree topology).
- Nearly 200 racks of compute hardware (10,000 sq ft)
- Upgrade with “Future Knight’s” in 2015
Stampede Will Enable New Scientific Discoveries Across Domains

1000+ projects, by thousands of researchers
Why did we include Xeon Phi?
The Power Problem in HPC

• We have lots of transistors… Moore’s law is holding; this isn’t necessarily the problem.

• We don’t really need lower power per transistor, we need lower power per *operation*.

• How to do this?
Intel’s MIC Approach: Coprocessor, Not Just Accelerator

• Since days of RISC vs. CISC, Intel has mastered the art of figuring out what is important about a new processing technology, and saying “why can’t we do this in x86?”

• Intel Many Integrated Core (MIC) architecture is (like GPUs) about large die, simpler circuit, much more parallelism… and in the x86 line
  – Easier to use (even as an accelerator) with standard, familiar tools (e.g. MPI, OpenMP)
  – More flexible programming modes, models
New Technologies/Milestones in the Stampede Project

• **Density:** will surpass 40KW/Cabinet
  – New Dell node designs to support multiple 300W expansion cards in single node ~1U

• **Total system power past 5 *megawatts***
  – Thermal storage technology incorporated.

• **Breakthrough price/performance and power/performance.**
  – Inclusion of Intel MIC; but we must program it.
  – Thermal storage to reduce power costs

• **Application concurrency past 1 *million* threads per application**
Stampede Footprint

Facilities at TACC

Stampede: 8000 ft\(^2\)
10 PF
6.5 MW

Ranger: 3000 ft\(^2\)
0.6 PF
3 MW

Capabilities: 20x
Footprint: 2x

Stampede InfiniBand (fat-tree)
~75 Miles of InfiniBand Cables
Some utilities are involved
Actually, way more utility space than machine space

Turns out the utilities for the datacenter costs more, takes more time and more space than the computing systems
Stampede: How Will Users Use It?

- 2+ PF Xeon-only system (MPI, OpenMP)
  - Many users will use it as an extremely powerful Sandy Bridge cluster—and that’s OK!
    - They may also use the shared memory nodes, remote vis
- 7+ PF MIC-only system (MPI, OpenMP)
  - Homogeneous codes can be run entirely on the MICs!
- ~10PF heterogeneous system (MPI, OpenMP)
  - Run separate MPI tasks on Xeon vs. MIC; use OpenMP extensions for offload for hybrid programs
The Stampede is Here!

• Stampede began production January 7, 2013.
  – Things are going really well
  – We still need and want your feedback on your experiences

• Thank to NSF, Intel, Dell
• Welcome to TACC!