The **Stampede** is Coming
Welcome to Stampede Introductory Training

Dan Stanzione
Texas Advanced Computing Center
dan@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
Stampede: Solicitation

• US National Science Foundation (NSF) competition in 2011 for large HPC system
  – To support open science across all domains of science and engineering
  – To enable simulation-based & data-driven science

• Solicitation requirements & constraints
  – $25M production system, plus up to $5M for ‘innovative component’
  – Up to $6M/year for 4 years of O&M
  – In production by January 2013 (fingers crossed)
Stampede: NSF HPC Context

- System must be integrated into XSEDE
  - Allocations, integrated support, etc.
- Will effectively replace Ranger (TACC) and Kraken (NICS) HPC systems
  - They expire in February 2013 (Ranger), June 2014 (Kraken)
    - So probably had to bid at least 1½ - 2 PF
- Does not come with separate funding for networking, storage, visualization—but all are required!
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
  – Thermal energy storage to reduce operating costs
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
- A minimum of 6,400 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.

• Application concurrency past 1 *million* threads per application
Power/Physical

- Stampede spans 182 48U cabinets.
- Power density (after upgrade in 2015) will exceed 40kW per rack.
- Estimated 2015 peak power is 6.2MW.
Stampede Datacenter Features

- Thermal energy storage to reduce peak power charges
- Hot aisle containment to boost efficiency (and simply provide enough cooling).
- Total IT power to 9.5MW, total power ~12MW.
- Expand experiments in mineral oil cooling.
Facilities at TACC

Stampede: 8000 ft²
10 PF
6.5 MW

Capabilities: 20x
Footprint: 2x

Stampede InfiniBand (fat-tree)
~75 Miles of InfiniBand Cables

Ranger: 3000 ft²
0.6 PF
3 MW
Stampede Datacenter – February 20th
Stampede Datacenter – May 16th
Stampede Datacenter – June 20th
Stampede Datacenter, ~September 10th
Stampede Datacenter, ~September 10th
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
Will My Code Run on MIC?

• Yes

• That’s the wrong question, it’s:
  – Will your code run *best* on MIC?, or
  – Will you get great MIC performance without additional work?

• You’ll find out more about this over the next couple of days.
The Stampede is Here!

- Stampede just began production Monday night of this week.
  - You will be among the very first to use it!
  - In many ways, this is still the “shakedown cruise”; we need your feedback.

- Thank to NSF, Intel, Dell
- Welcome!