Stampede Hardware Overview

John Cazes
May 4, 2015

Texas Advanced Computing Center
The University of Texas at Austin

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

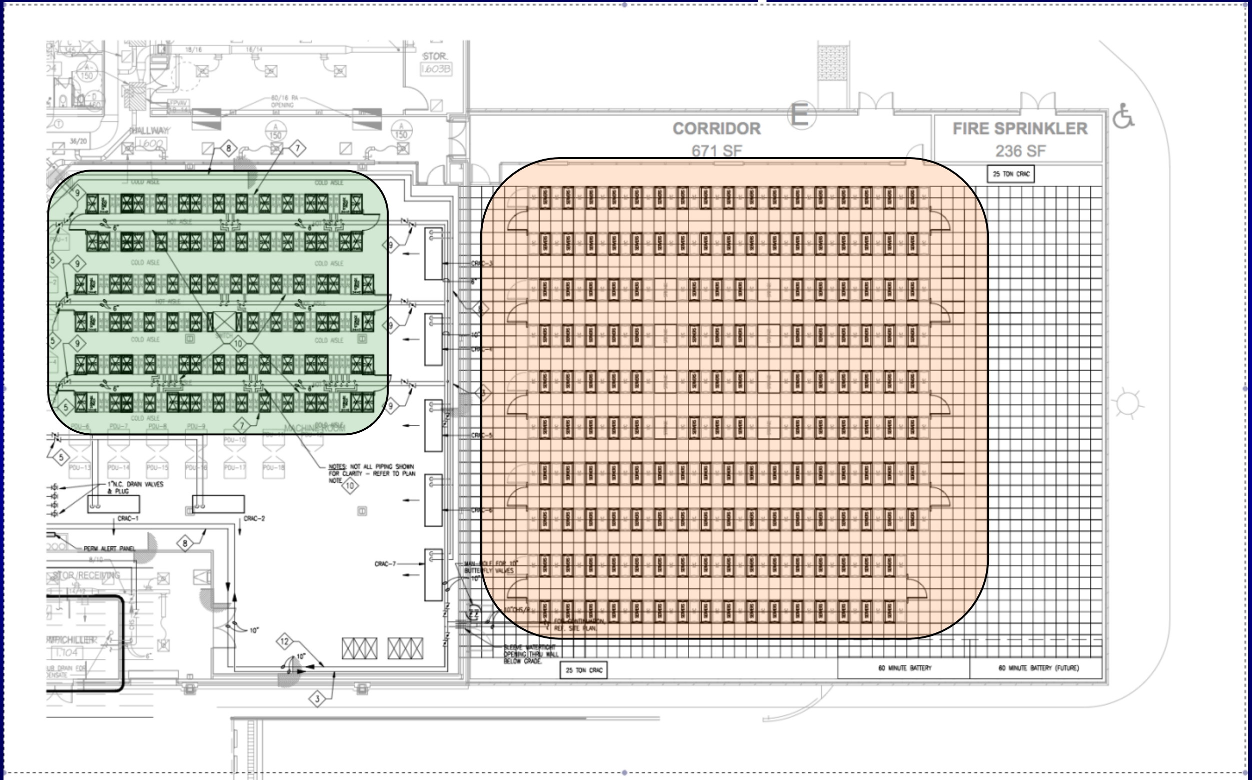
Stampede

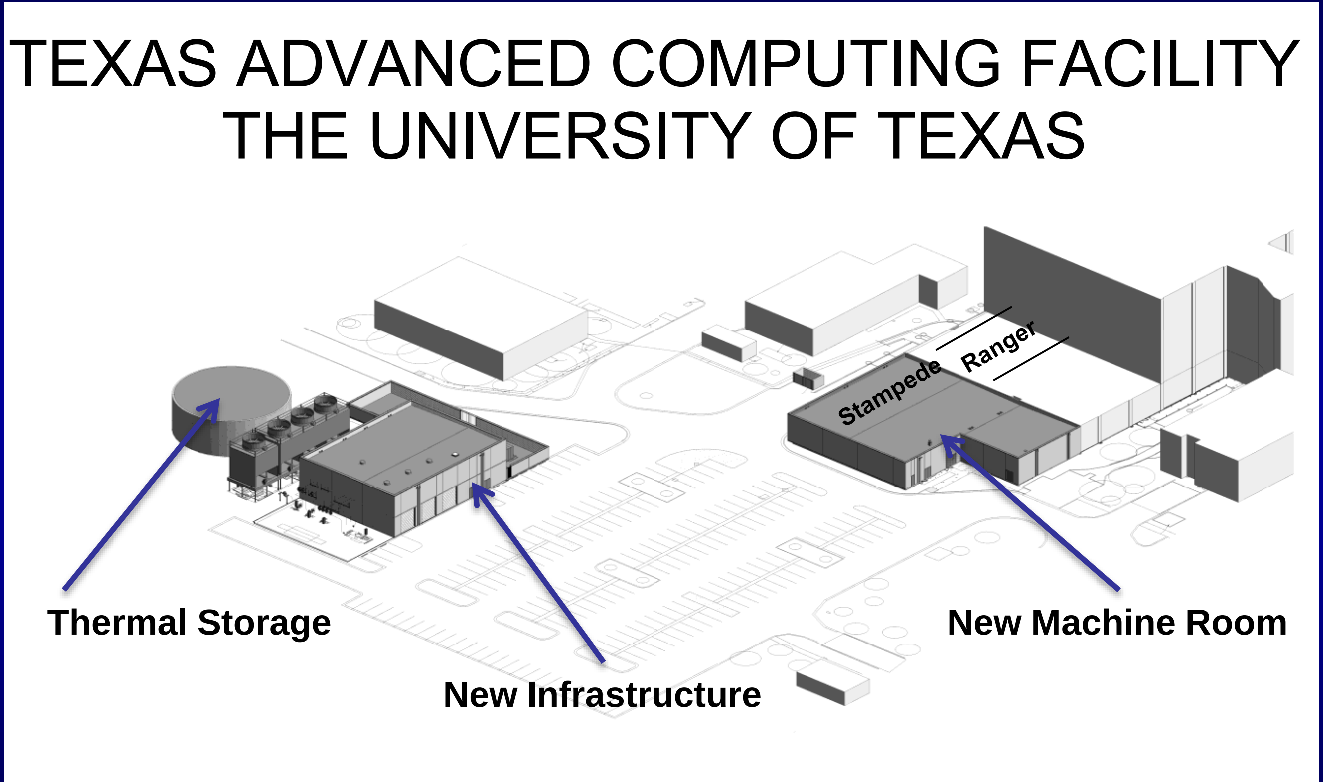
- NSF 11-511: “High Performance Computing System Acquisition: Enhancing the Petascale Computing Environment for Science and Engineering”
- Enable sustained petascale computational and data-driven science and engineering and provide an “innovative component”

Dell/Intel Partnership

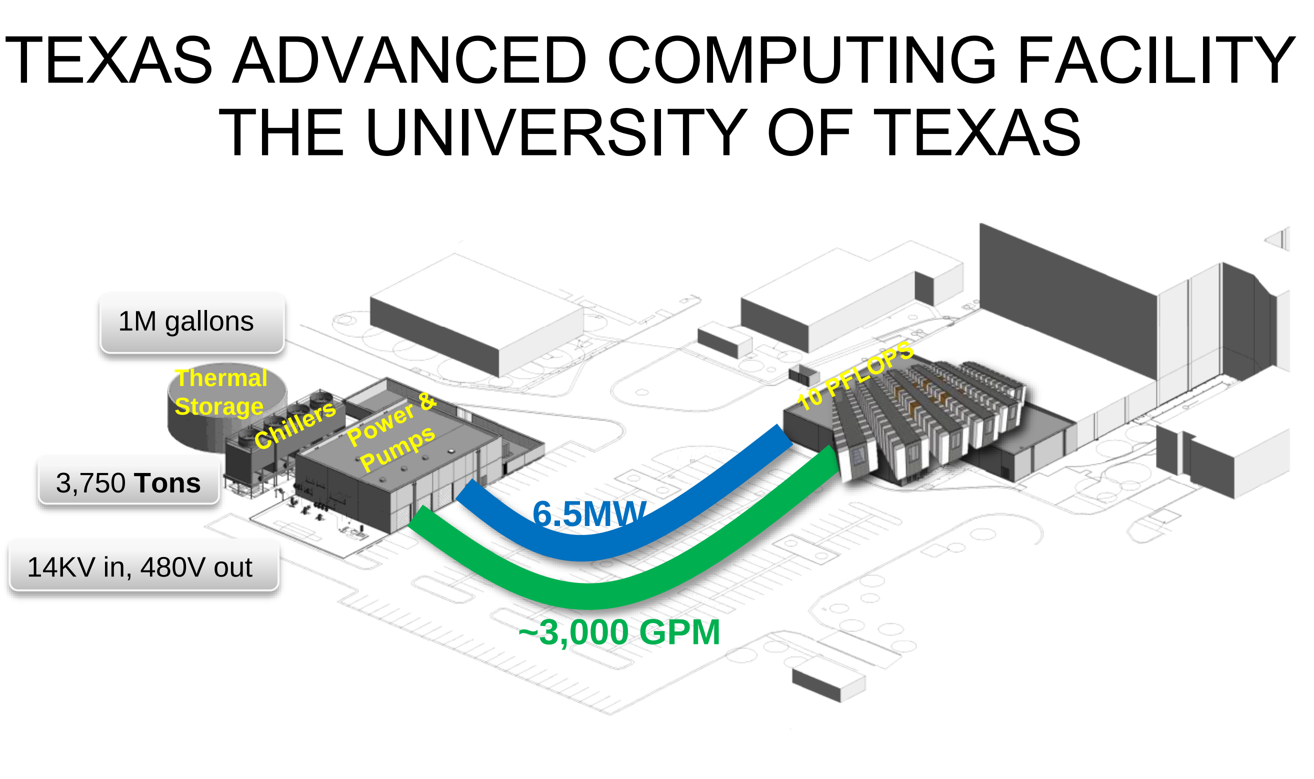
- TACC Partnered with Dell and Intel to design Stampede
- Intel MIC (Intel Xeon Phi) is the innovative component
  - High performance and low power per operation
  - Highly programmable

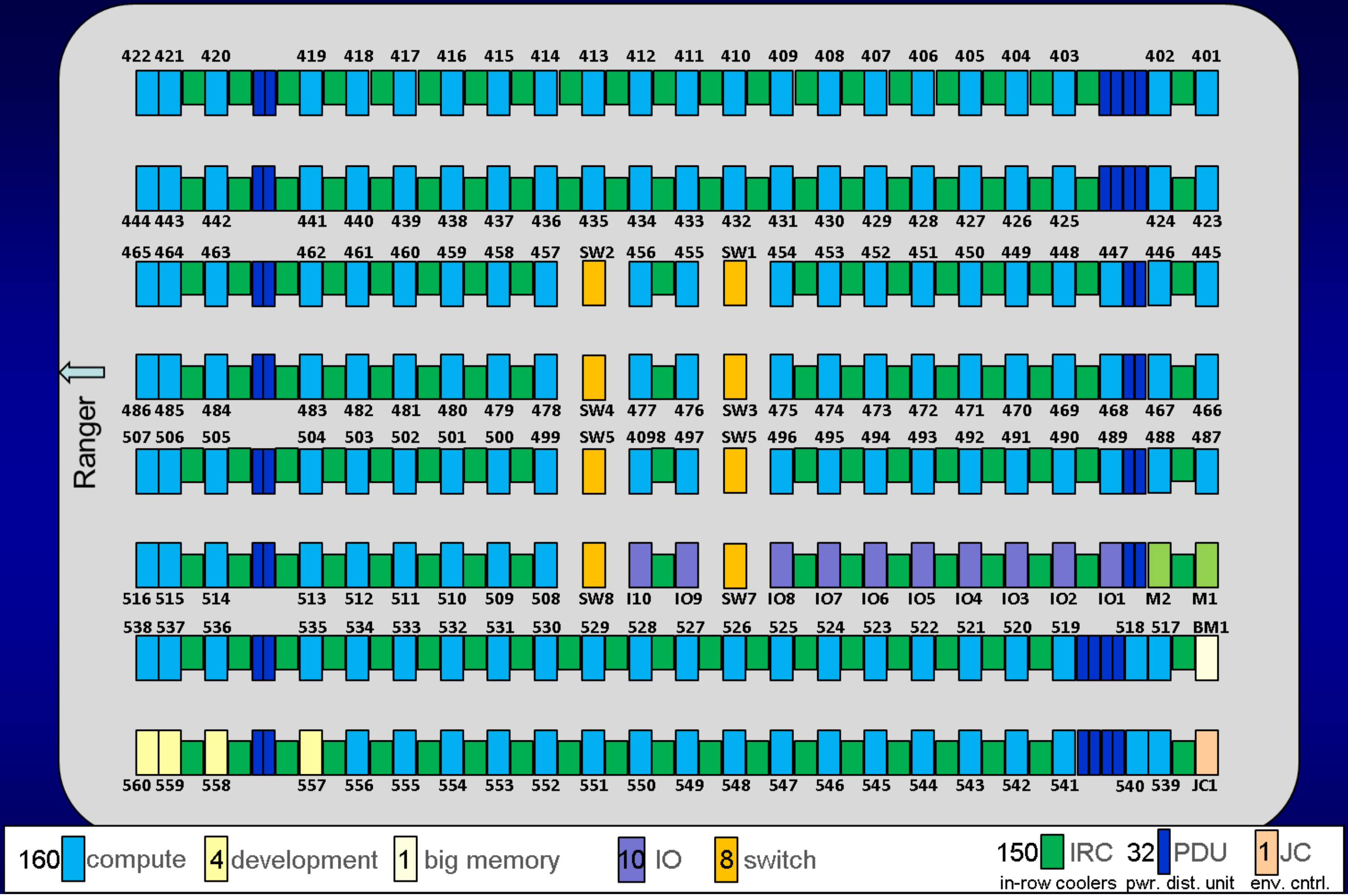
Datacenter Expansion




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

2x footprint, 2x power, 20x Capability (Stampede vs. Ranger)





Stampede consists of 182 48U cabinets.

Cooling and Electrical Infrastructure

Stampede Performance

Stampede debuted at #7 on the Top 500

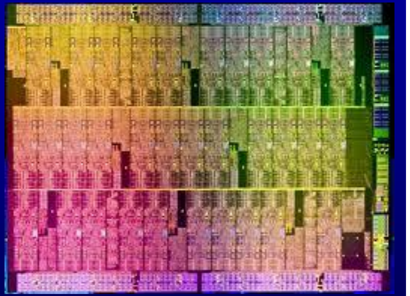
Stampede Overview

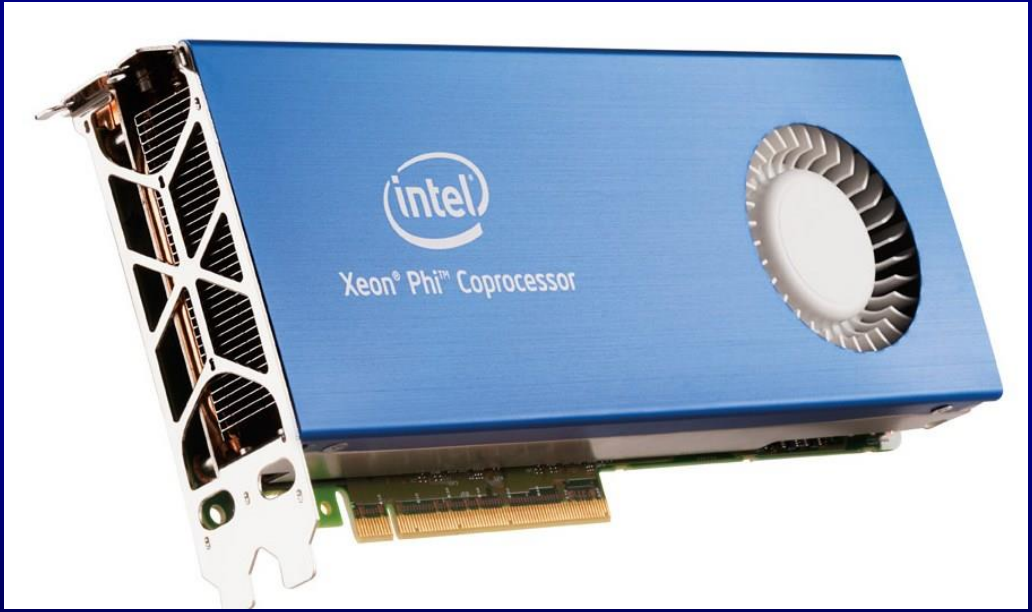
- $27.5M acquisition
- 10 petaflops (PF) peak performance
- 2+ PF Linux cluster
  - 6400 Dell DCS C8220X nodes
  - 2.7GHz 8 core Intel Xeon E5 (Sandy Bridge)
    - 102,400 total cores
  - 56Gb/s FDR Mellanox InfiniBand
  - 7+ PF Intel Xeon Phi Coprocessor
    - TACC has a special release: Intel Xeon Phi SE10P
  - 14+ PB disk, 150GB/s
  - 16 1TB shared memory nodes
  - 128 NVIDIA Tesla K20 GPUs

Processor Specs

Arch. Features	Xeon E5	Xeon Phi SE10P
Frequency	2.7GHz +turbo	1.0GHz +turbo
Cores	8	61
HW threads/core	2	4
Vector size	256 bits 4 doubles 8 singles	512 bits 8 doubles 16 singles
Instr. Pipeline	Out of Order	In Order
Registers	16	32
Caches	L1:32KB L2:256KB L3:20MB	L132KB L2:512KB
Memory	2 GB/core	128 MB/core
Sustained Memory BW	75 GB/s	170 GB/s
Sustain Peak FLOPS	1 thread/core	2 threads/core
Instruction Set	x86 + AVX	x86 + new vector instructions

MIC Details

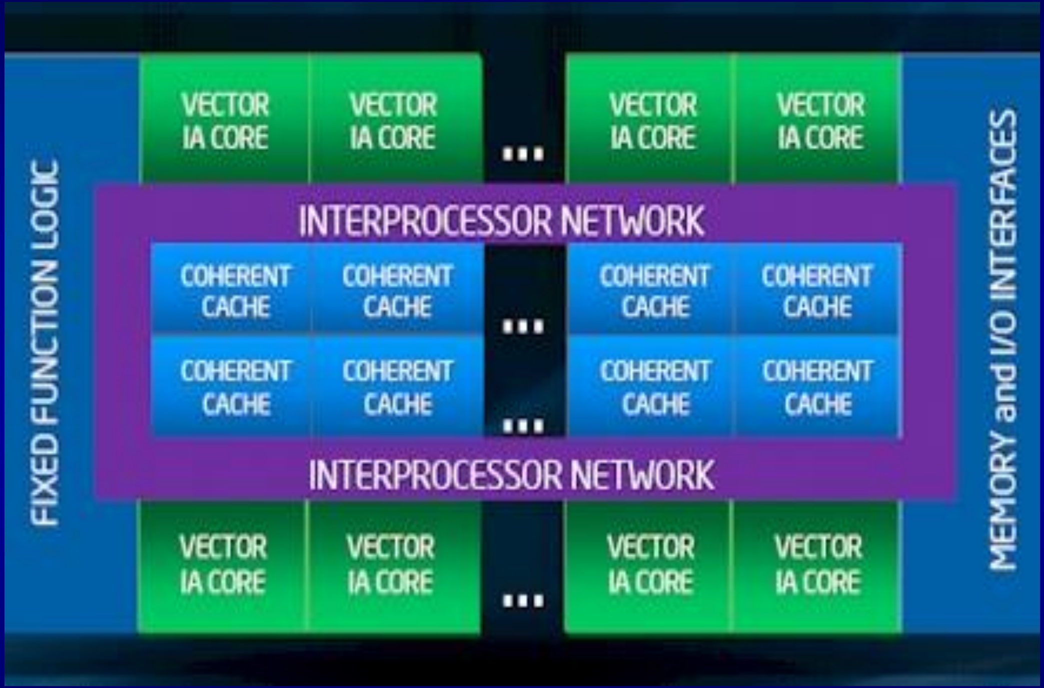




What is a MIC

- 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



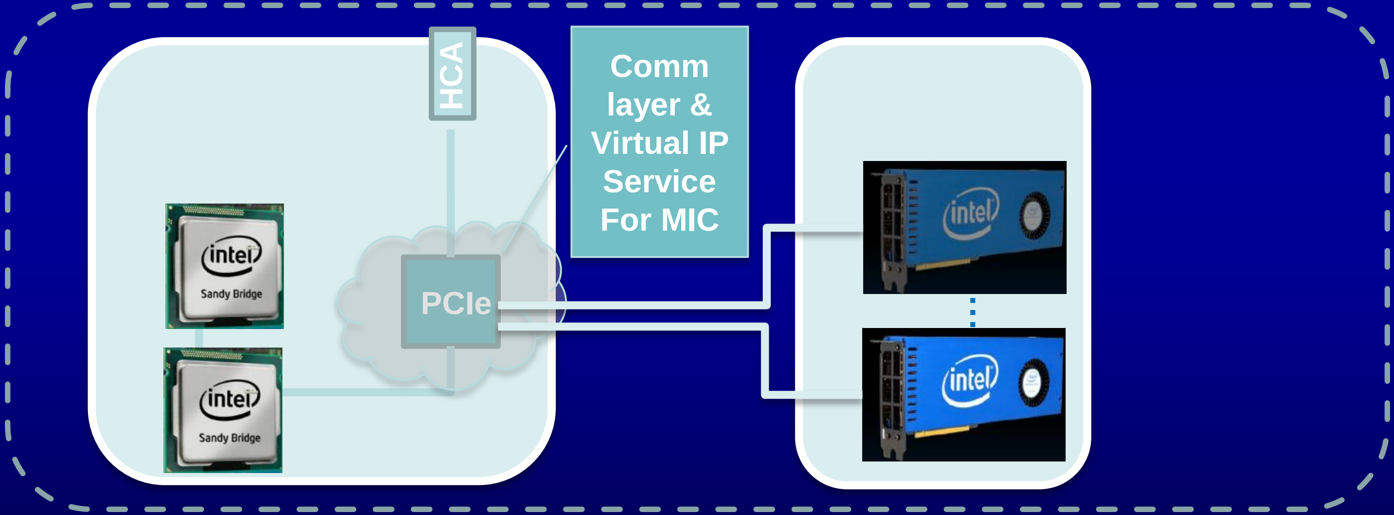
- Many cores on the die
- L1 and L2 cache
- Bidirectional ring network
- Memory and PCIe connection

Dell DCS C8220z Compute Node

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

Compute Node Configuration

CPUs and MIC appear as separate HOSTS ("symmetric" computing)



Stampede Filesystems

Storage Class	Size	Architecture	Features
Local (each node)	Login: 1TB Compute: 250GB Big Mem: 600 GB	SATA SATA SATA	432GB on /tmp 80GB on /tmp 398GB on /tmp
Parallel	Total: 8 PB $HOME: .5 PB $SCRATCH: 7.4 PB	Lustre	372 OST 72 OST 300 OST
Parallel(Center wide)	$WORK: 19 PB	Lustre	672 OST 112 OSS 2 MDS

Stampede Filesystems

- **$HOME**
  - Quota : 5GB, 150K files
  - Filesystem is backed up
- **$WORK**
  - Quota: 1 TB, 3M files
  - NOT backed up
  - Use `cdw` to change to $WORK
- **$SCRATCH**
  - No Quota
  - Total size 7.4 PB
  - 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 & 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

Queue Structure

Queue Name	Max Runtime	Max Nodes/Procs	Max Jobs in Queue	SU Charge Rate	Purpose
normal	48 hrs	256 / 4K	50	1	normal production
development	2 hrs	16 / 256	1	1	development nodes
largemem	48 hrs	4 / 128	4	2	large memory 32 cores/node
serial	12 hrs	1 / 16	8	1	serial/shared_memory
large	24 hrs	1024 / 16K	50	1	large core counts (access by request)
request	24 hrs	--	50	1	special requests
normal-mic	48 hrs	256 / 4k	50	1	production MIC nodes
normal-2mic	24 hrs	128 / 2k	50	1	production MIC nodes with two co-processors
gpu	24 hrs	32 / 512	50	1	GPU nodes
gpudev	4 hrs	4 / 64	5	1	GPU development nodes
vis	8 hrs	32 / 512	50	1	GPU nodes + VNC service
visdev	4 hrs	4 / 64	5	1	Vis development nodes (GPUs + VNC)

John Cazes

cazes@tacc.utexas.edu

For more information:

www.tacc.utexas.edu

TACC