TACC

TEXAS ADVANCED COMPUTING CENTER

STAMPEDE 2 UPDATE

STAMPEDE 2

- Funded by NSF as a renewal of the original Stampede project.
- Follow the legacy of success of the first machine as a supercomputer for a *broad* range of workloads, large and small.
- Install without ever having a break in service – in the same footprint.

STAMPEDE 1 RECAP

- Awarded by NSF as an XSEDE resource in September 2011.
- Stampede was constructed in 2012, and went into production on January 7th, 2013.
  - Through 5 years, more than 8M simulations and more than 3B hours delivered to 13,000+ users on more than 3,500 projects.
  - #6 on Nov. 2012 top 500 list – stayed in top 10 for 7 lists (fell to 17 in Nov. 2016 at end of original life)
  - **Request rate to XSEDE 5-6x available capacity *every* quarter.**
  - A true national resource – UT accounted for <7% of usage; 400+ institutions represented.
  - **TACC staff answered more than 15,000 tickets.**

A DATA/WORKLOAD DRIVEN DESIGN

- We keep a massive amount of data about what runs on our system *and* how well it runs.
  - TACC Stats
    - Low level performance counter data, sampled at a very course grain, every 10 minutes, for last ~9 million jobs
    - (Now integrated with XDMOD reporting)
  - XALT
    - Binary/shared library tracking for life of Stampede.
  - Lustre instrumentation
    - Metadata traffic, other filesystem instrumentation, for life of Stampede.

STAMPEDE-2 DESIGN

- Support the high-end, MPI user:
  - Majority of cycles on Stampede consumed by MPI-based binaries
    - 70% of computational capacity on Stampede 2 in KNL/Xeon Phi
- But that user isn't everyone, and not all codes run well on KNL
  - Some users running serial/small scale codes or scripting languages that want high clock rate
    - 30% in Xeon processor – but wait for the new generation, since it will be used for four years
- Broader trend towards Exascale is more cores – so we didn't build a "head in the sand" system towards the future.

STAMPEDE 2 DESIGN

- Stampede 2 supports a wide range of use cases – but not all of them
- However, innovation in the software/ operations approach lets us support many more things than a "traditional" leadership machine
  - Stampede 2, like all TACC systems, has a REST API that will support gateways/web applications/automated workflows
  - Through Singularity, we have increased our support for Life Sciences codes to more than 2,000 applications
  - Software Defined Vis for in-situ vis work

STAMPEDE 2 -- COMPONENTS

- Phase 1 – June 2017
  - 4,204 Intel Xeon Phi 7250 ”Knights Landing” (KNL) nodes
  - ~20PB (usable) Lustre Filesystem (Seagate), 310GB/s to /scratch.
  - Intel OmniPath Architecture (OPA) Fabric – Fat Tree topology
  - Ethernet fabric and (some) management infrastructure.
- Phase 2 – December 2017
  - 1,736 Intel Xeon Platinum 8160 “Skylake” two-socket nodes
  - (Associated rack level networking, but core in phase 1).
  - Balance of management hardware, new Skylake servers
- Phase 3 – 2nd half 2018
  - 3D Xpoint NVDIMMS as an experimental component in a small subset of the system.

HARDWARE OVERVIEW

- Stampede 2 Phase 1 compute nodes, 285,882 cores
  - 924 Dell C6320P chassis, 4 nodes per chassis
    - 3,696 total compute nodes
      - Intel Xeon Phi 7250 CPU, 68 cores, 1.4GHz
      - 96 GB (6x16GB) 2400MHz DDR4
      - 200 GB SSD
    - Redundant 1600W power supplies
  - 126 Intel PCSD chassis, 4 nodes per chassis (originally Stampede 1.5)
    - 508 total compute nodes
      - Intel Xeon Phi 7250 CPU, 68 cores, 1.4GHz
      - 96 GB (6x16GB) 2400MHz DDR4
      - 120 GB SSD

STORAGE SUBSYSTEM (PHASE 1)

- Seagate (now Cray) ClusterStor 300
  - 35 Scalable Storage Units (SSU)
    - Pair of servers configured for high availability with active/active failover
    - 82 10TB drives, 41 drives per LUN in declustered parity (GridRAID), two drives act as filesystem external journal
    - Each SSU designed to provide ~10GB/s of performance
  - 3 Metadata Management Units (MMU)
    - Pair of Lustre meta-data servers with active/active failover
    - Disk to support up to 4 billion inodes per MMU
  - 2 System Management Units (SMU)
    - Pair of management servers, primary and secondary
    - Used to configure and manage the filesystems
  - 6 racks with two GigE and two OPA switches per rack

STORAGE FILESYSTEMS

- Seagate storage provides two Lustre filesystems
  - Home: 2 SSUs, 1 MMU, 1 SMU; quota and backed up to archive
  - Scratch: 33 SSUs, 2 MMUs, 1SMU; no quota but purged, designed for >300GB/s bandwidth
- Stockyard provides /work site-wide filesystem
  - DataDirect Networks 25PB Lustre filesystem

OPA FABRIC TOPOLOGY

- Fat-tree topology design with 7:5 oversubscription
- Each top of rack switch connects to twenty different line cards to flatten topology
  - Up to sixteen ToR switches per director class core switch line card
  - Switches in the same rack always connect to same line card
- Adjacent racks connected to same line card as much as possible
  - E.g. first 8 racks connect to same line card
- I/O switches spread across line cards to avoid I/O bottlenecks
- Custom cable management panels to allow for easy cabling of core switches

RESULTS SO FAR – REALLY BROAD GENERALIZATIONS

- Everything runs on KNL, but. . .
  - Carefully tuned codes are doing pretty well, but with work.
  - “Traditional” MPI codes, especially with OpenMP in it do relatively well, but not great.
  - Some codes, particularly, not very parallel ones, are pretty slow, and probably best run on regular Xeon processors.
- The Intel Xeon Scalable Processors are far exceeding original performance expectations

OUR EXPERIENCE WITH XEON PHI

- ***Xeon Phi looks to be the most cost and power efficient way to deliver performance to highly parallel codes.***

- In many cases, it will not be the fastest. For things that only scale to a few threads, it is *definitely* not the fastest.

- But what is under-discussed:
  - A dual-socket Xeon node costs 1.6x what a KNL node costs, even after discounts.
    - A dual-socket, dual GPU node is probably >3x a Xeon Phi node.
  - A KNL node uses 100 less watts per node than a dual-socket Xeon node.

POWER FROM TOP 500

- List from June at ISC17 in Frankfurt
- Stampede-2 uses half the power of a roughly equivalent performance system (see 11 vs 12)

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
10	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	301,056	8,100.9	11,078.9	4,233
11	United Kingdom Meteorological Office United Kingdom	Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect Cray Inc.	241,920	7,038.9	8,128.5	3,629
12	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede2** - PowerEdge C6320P, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Dell	285,600	6,807.1	12,794.9	1,890

DEPLOYMENT SUMMARY

- All phase 1&2 racks/chassis/nodes installed and operational as part of Stampede2
- Updated BIOS and firmware have resolved all major stability and performance issues encountered during the phase 2 deployment
- Performance of phase 2 nodes exceeding all expectations

INTEL XEON “SKYLAKE” PERFORMANCE

- Platinum 8160 processor exceeding expectations
- STREAM: expected 175 GB/s, measuring >200 GB/s
- HPL: expected 1.9 TFlops, measuring 2.3-2.4 TFlops
- Latency: expected 0.8 μs, measuring < 0.5 μs
- One limitation, single core memory bandwidth: 13GB/s

- Processor frequency range, 1.6 GHz – 3.7 GHz!
- Frequency depends on cores active *AND* instruction set compiled/executed in application.

THANKS!

QUESTIONS?

Tommy Minyard
minyard@tacc.utexas.edu