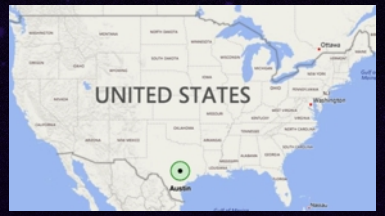
TACC Resources

Tim Cockerill
Director of User Services
cockerill@tacc.utexas.edu

Petascale Institute
August 23, 2019

TACC AT A GLANCE



Personnel

185 Staff (~70 PhD)

Facilities

12 MW Data center capacity
Two office buildings, Three Datacenters, two visualization facilities, and a chilling plant.

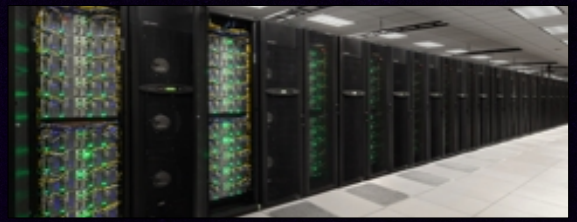
Systems and Services

Two Billion compute hours per year

5 Billion files, 75 Petabytes of Data, Hundreds of Public Datasets

Capacity & Services

HPC, HTC, Visualization, Large scale data storage, Cloud computing
Consulting, Curation and analysis, Code optimization, Portals and Gateways, Web service APIs, Training and Outreach

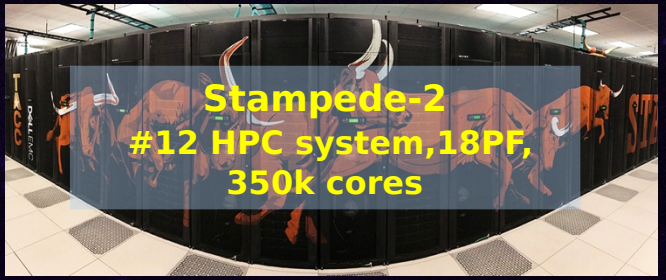


WE’RE HERE FOR YOU!

TACC’s most valuable resource – our staff

- Consulting
- Date curation and analysis
- Code optimization
- Portals and Gateway
- Web service APIs
- Training and Outreach

AN ECOSYSTEM FOR EXTREME SCALE SUPERCOMPUTING



Stampede-2
#12 HPC system,18PF, 350k cores

Lonestar 5
Texas-focused HPC/HTC XC40 30,000 Intel Haswell cores 1.25 PF

Hikari
Protected Data Containers 10,000 Intel Haswell cores 400TF

Wrangler
Data Intensive Computing
0.6 PB flash storage 1 TB/s read rate

Maverick2
GPU/Interactive/ Analytics
GeForce GPUs, Jupyter and interactive support

Jetstream
w/ Indiana U.
Science Cloud/HTC VM Library ~10,000 Intel Haswell cores

Stockyard
Shared Storage Across TACC 30PB, Lustre

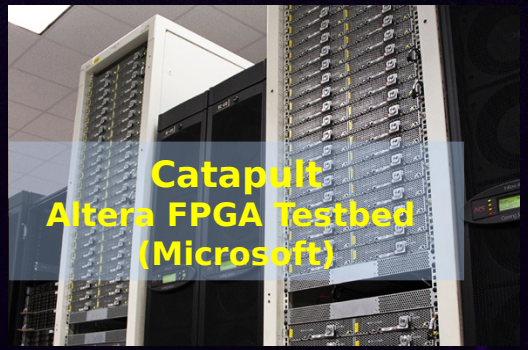
Ranch
Archive HIPAA-Aligned 30PB Disk Cache, 0.5EB Tape

Corral
Published Data Collections HIPAA-Aligned 20PB Replicated Disk,

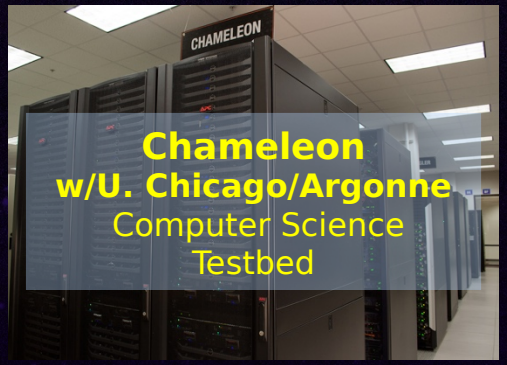
Rodeo

Lasso

EXPERIMENTAL SYSTEMS



Catapult
Altera FPGA Testbed
(Microsoft)



Chameleon
w/U. Chicago/Argonne
Computer Science
Testbed

Fabric
Alternate
Architectures
(IBM,CAPI,FPGA, GPU)

Discovery
New
Processor/Storage
Benchmarking

Rustler
Object Storage
Testbed

THE BROADER TACC ECOSYSTEM
DISCOVERY SCIENCE AT ALL SCALES

FRONTERA

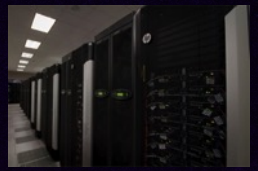
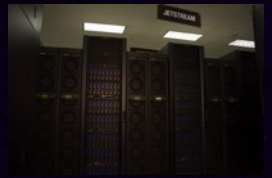
Leadership/ Discovery Science

Longhorn
IBM Power 9 +GPU
400+ Nvidia V100s
AI/ML/DL @ Scale

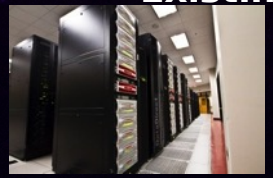
Testbeds
Catapult (Upgrade)
Non-Volatile Memory
Quantum
Future . . .

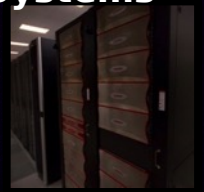
Existing TACC Computing Systems



Existing TACC Storage Systems

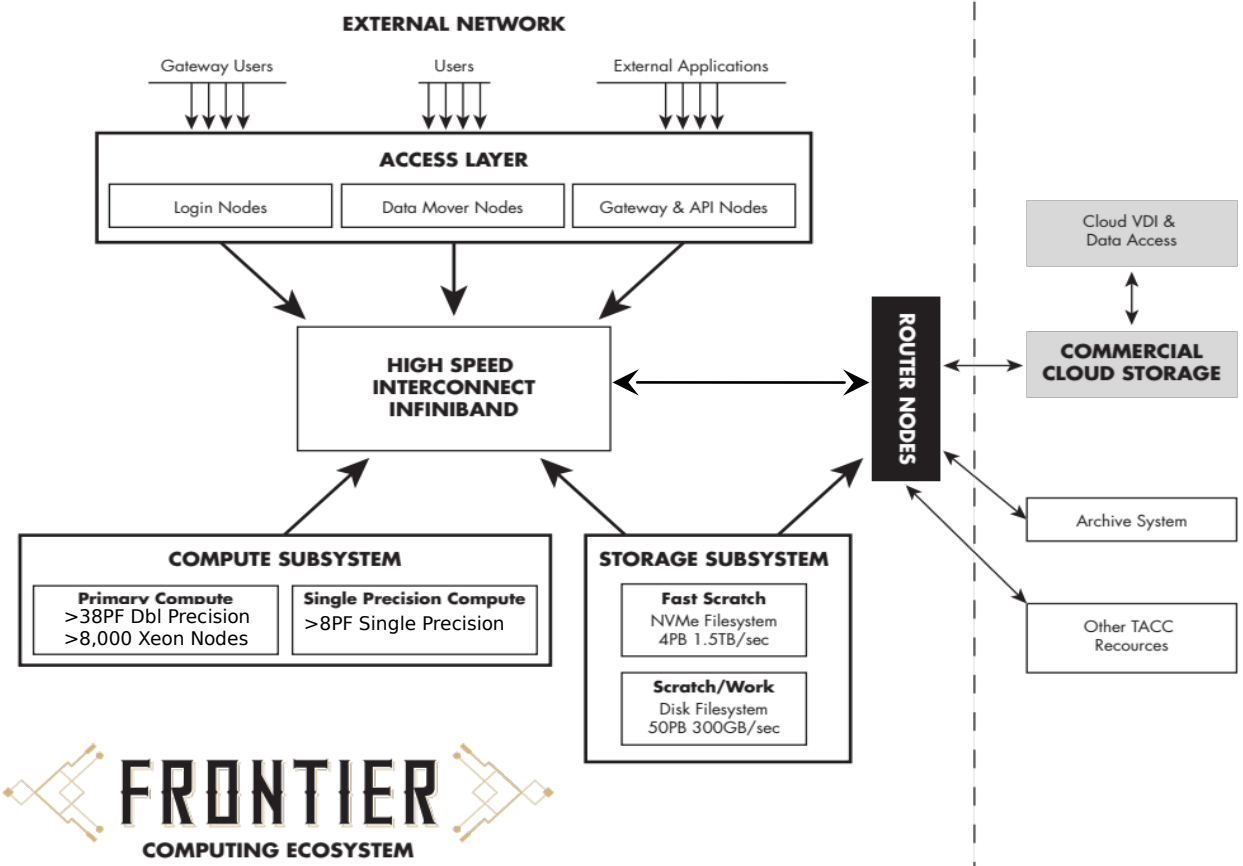




FRONTERA SYSTEM --- HARDWARE

- Primary compute system: DellEMC and Intel
  - 39PF PetaFlops Peak Performance -- 8,000+ nodes of Intel Cascade Lake
- Interconnect: Mellanox HDR and HDR-100 links.
  - Fat Tree topology, 200Gb/s links between switches.
- Storage: DataDirect Networks
  - 50+ PB disk, 3PB of Flash, 1.5TB/sec peak I/O rate.
- Single Precision Compute Subsystem: Nvidia
- Front end for data movers, workflow, API

ORIGINAL SYSTEM OVERVIEW



PROCESSORS

- “Main” Compute Partition: 8,008 nodes
- Node: Dual–socket, 192GB, HDR-100 IB interface, local drive.
- Processor: Intel 8280 “Cascade Lake” *Intel 2nd generation scalable Xeon blah blah*
  - 28 Cores
  - 2.7Ghz clock “rate” (sometimes)
  - 6 DIMM Channels, 2933Mhz DIMMS
- Core count+15%, clock rate +30%, memory bandwidth +15% vs. Skylake

- Why? They are universal, and not experimental

INTERCONNECT

- Mellanox HDR , Fat Tree topology
- 8008 nodes = 88*91 = 91 Compute Racks
- Mellanox ASICS == 40 HDR ports. Chassis switches have 800 ports.
- Each rack is divided in half, with it’s own TOR switch:
  - 44 compute nodes at HDR-100 == 22 HDR ports
  - 18 uplink 200Gb HDR ports, 3 lines (600Gb) to each of 6 core switches.
- No oversubscription in higher layers of tree (11-9 in rack).
- No oversubscription to storage, DTN, service nodes (all connected to all 6 switches).
- 8200+ cards, 182 TOR switches, 6 core switches, 50 miles of cable.
- Good news: 8,008 compute nodes use only 3,276 fibers to connect to core.

FILESYSTEMS

- We no longer need to scale filesystem size to scale Bandwidth.
- The size of the filesystem is mostly to support concurrent users – Bandwidth is the limit for individual user (or IOPS for pathological ones).
- So – we aren't going to build one big filesystem any more.
- /home1 , /home2, /home3
- /scratch1, /scratch2, /scratch3 (initial assignment round robin)
- Flash will be a separate filesystem with some clever name, like /flash.
  - This will require you to request access; or to be identified by our analytics as maxing a filesystem.
- Roughly 100GB/s to each scratch, 1.2TB/sec to /flash
  - The code on the previous slide can trash, at most, 1/7th of the available filesystems.
  - (Seriously, we have put in some tools to limit those; we may ask you to use a library we have that wraps Open(), and limits the number per second).

LARGE “MEMORY”, OR FASTER I/O

- One experimental piece we will add soon:
- ~Sixteen additional compute nodes, same Intel 8280
- Quad-socket, 384GB RAM
- Twenty-four 256GB NVDIMMS (6TB per node) – Intel “Optane”

GPU SYSTEMS

- We took a, umm, non-traditional approach with NVIDIA
- We have split the GPU system into two components:
  - *Longhorn*, which will run as a separate system, but all Frontera users will have access to it.
  - The Frontera GPU queue, which will use “workstation” GPUs, for single (or less) precision only.

LONGHORN

- If you have used DOE's *Sierra* system at Livermore, it's that.
- If you haven't:
  - IBM Power9 nodes, 4 V100 GPUs per node.
  - Mellanox EDR interconnect.
- 112 total nodes, 448 GPUs, separate scratch filesystem.
- Should be available by July.
- Running separately, because none of the binaries on Frontera (x86) will work here (your Frontera filesystems will be mounted and visible, but you will have a separate home and scratch).

FRONTERA-GPU SINGLE PRECISION SUBSYSTEM

- Shared filesystems with Frontera
- Mineral oil-cooled immersion
- Prototype (Maverick):
  - 24 (x86) nodes, 96 1080TI GPUs
- These cards give *amazing* Single Precision Flops/$
- But they don’t have tensor cores – NVIDIA really wanted us to have tensor cores
- New hardware will use Quadro 5000 RTX cards (Volta-generation silicon)
- Adding 90 additional nodes with 360 of these cards
- Available in September

CLOUD SERVICES YOU WILL SEE SOON

- Not quite ready yet, but hopefully September as we work out credentials management, etc
- From the Frontera portal, select data in home, work, or scratch, and send to your cloud account in AWS, GCP, or Azure
  - Optionally, add some curation data and get a DOI issued
  - We will pick up initial bills against your archive allocation
- Use credits issued against your Frontera allocation to use time on the cloud for TPU (Google), and other new hardware as it emerges

ARCHIVE

- For existing TACC Users, it’s still called “Ranch”, and still $ARCHIVE. You can move data to it via SCP or Globus.
- New hardware:
  - 30PB of disk cache (small files hopefully stay on disk for ~2 years).
  - Quantum tape infrastructure – 100PB initial capacity online, but can grown to 1EB.
  - Separate IBM tape infrastructure for encrypted/secure/replicated archive services.
- We have run versions of this archive continuously since 1986.

ALLOCATIONS (PROPOSED APPROACH)

- Four Tracks for Allocation:
  - **PRAC** – Very Large (>5% of total annual cycles) for projects immediately ready to run at scale (twice a year competitions).
  - **Pathways** – Projects that may become PRACs, but are not yet at that scale (quarterly, shorter proposals).
  - **Large Community Support Partnerships (LCSP)** - For key NSF investments (LHC, LIGO, etc) or community resources (Software Institutes, Large Gateways) where categorizing specific, individual experiments is hard, but aggregate demand is still large.
  - **Discretionary** – Still the catch all (Education, Industry, Emergency, etc.)
- We will also support “referral” proposals to/from XSEDE.
- We hope to open the next PRAC submission window in *October*.

HOW DO I GET ON?

- What scale are you today?
- A good starting point might be this other little, architecturally similar machine we have:
(apply  SEDE) .
- Ready to run at 10k+cores? Ask for a Discretionary today (or Pathways come October) to get some time to demonstrate readiness – use those results to apply for PRAC or LCSP (and get some work done in the process).

THANKS!!

- **The National Science Foundation**
- The University of Texas
- Peter and Edith O’Donnell
- Dell, Intel, and our many vendor partners
- Cal Tech, Chicago, Cornell, Georgia Tech, Ohio State, Princeton, Texas A&M, Stanford, UC-Davis, Utah
- **Our Users – the thousands of scientists who use TACC to make the world better**.
- All the people of TACC

Tim Cockerill
Director of User Services
cockerill@tacc.utexas.edu