COMPUTING FOR THE ENDLESS FRONTIER

Joao Barbosa
Research Engineer
Scientific Visualization Group

IXPUG Annual Conference
September 2019

TACC AT A GLANCE

Personnel

135 Dedicated Staff (+25 students)

Facilities

12 MW Data center capacity

Two office buildings, Three Datacenters, two visualization facilities, and a chilling plant.

Systems and Services

A Billion compute hours per year

5 Billion files, 50 Petabytes of Data, Hundreds of Public Datasets

World Class Computing

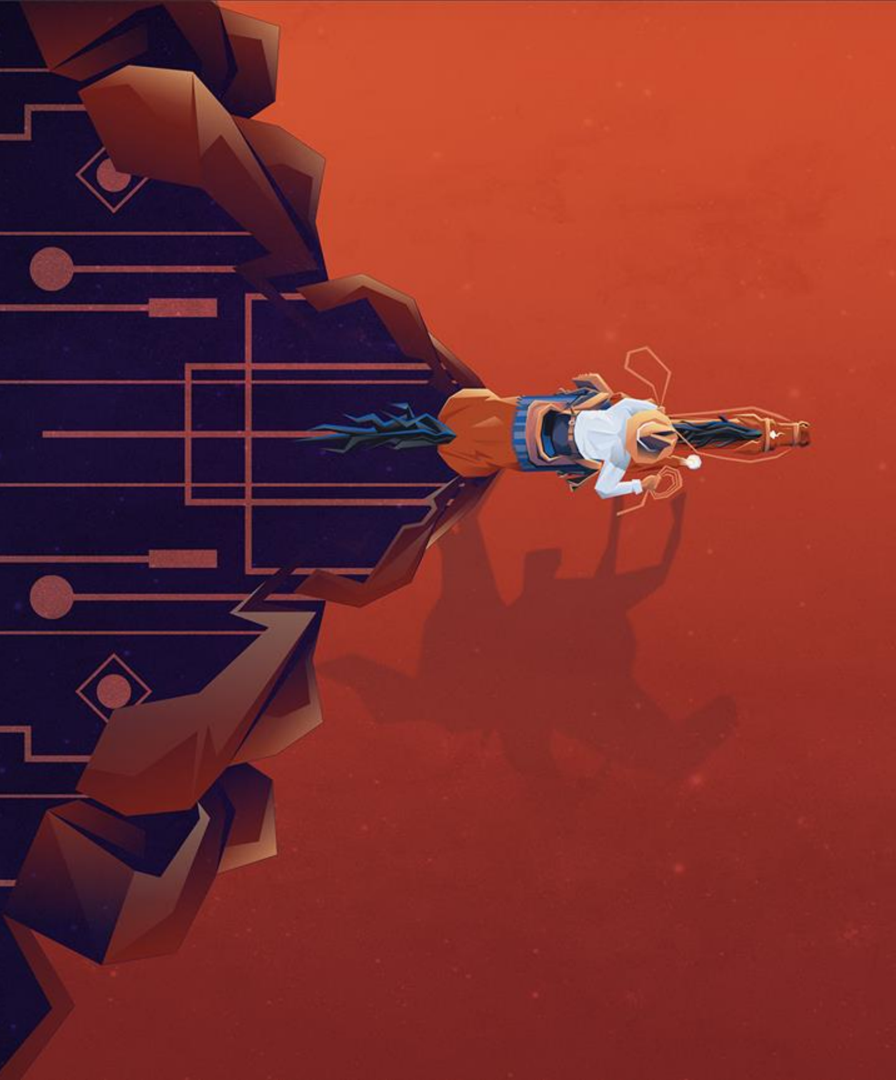
More than 15 supercomputers, data systems, cloud systems, visualization systems, machine learning systems, etc.

TACC SUPPORTS AN INCREDIBLE AMOUNT & DIVERSITY OF RESEARCH

- Since 2013…
  - Over *2 Billion* processor hours delivered to end users
  - 7+ **million** successful jobs
  - About 10,000 students, faculty, and staff use our Stampede directly
  - Over 30,000 more use it indirectly via portals and services
  - Peer-reviewed requests for time (via XSEDE) run ~500% available hours
- **Stampede alone** supports nearly 2,500 funded projects across the United States and abroad

FRONTERA

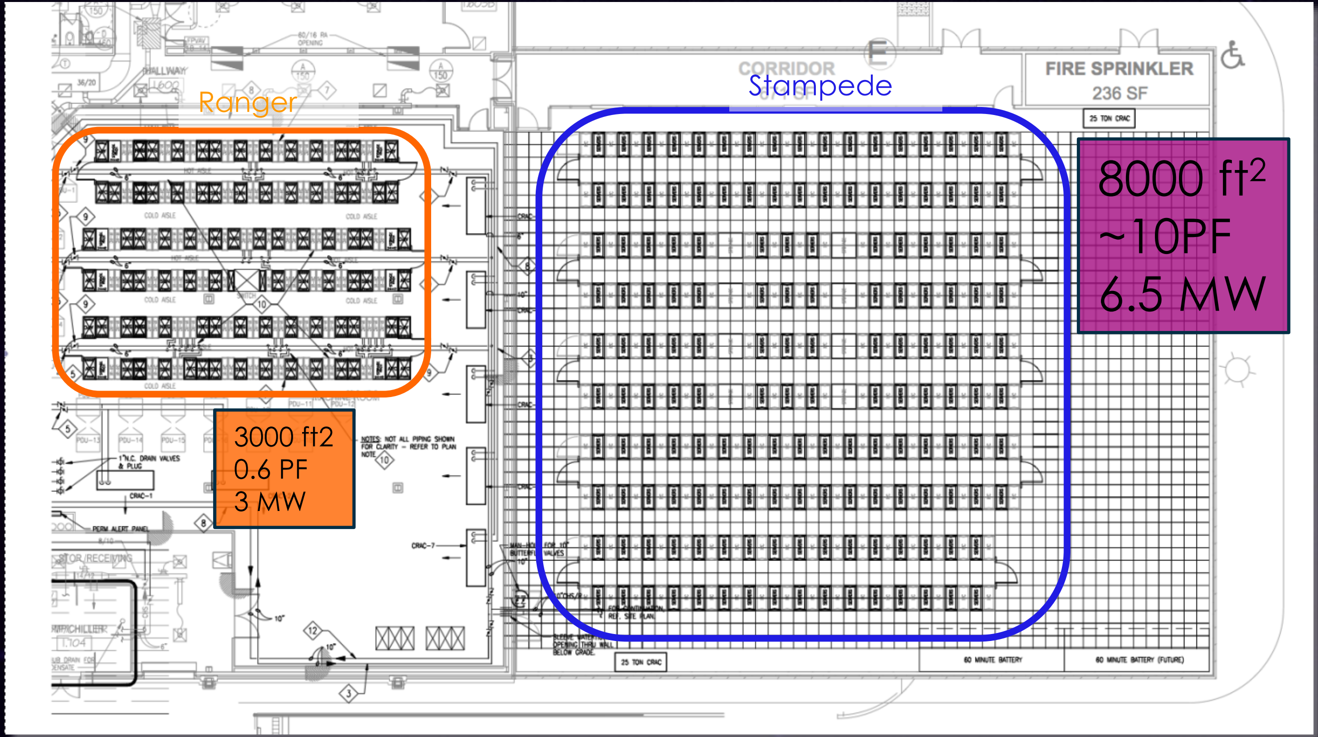
TACC | NSF | TEXAS



PI: Dan Stanzione TACC
Co-PIs: DK Panda, Ohio State
Omar Ghattas, UT-Austin
Tommy Minyard, John West, TACC

35
36

STAMPEDE FOOTPRINT



Ranger

3000 ft2
0.6 PF
3 MW

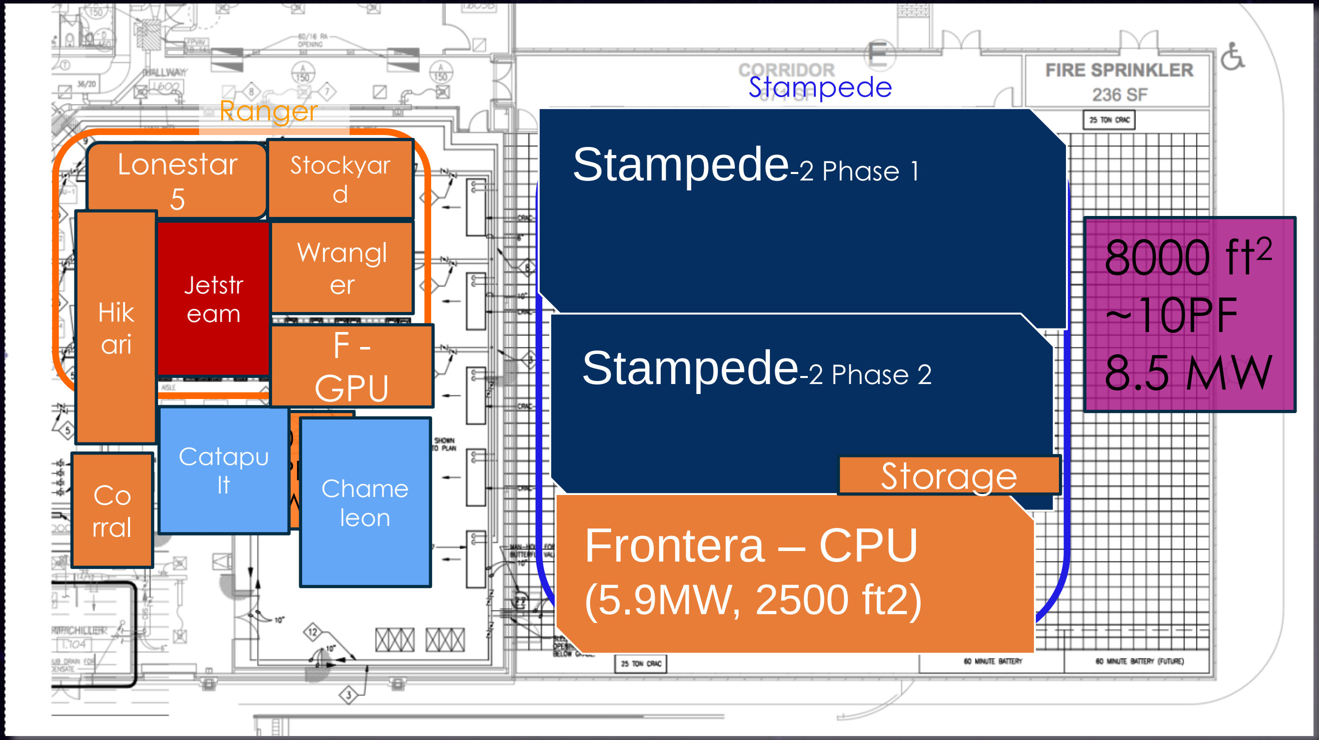
Stampede

8000 ft^2
~10PF
6.5 MW

Machine Room Expansion

Added 6.5MW of additional power

FRONTERA FOOTPRINT



Machine Room Expansion

Added 6.5MW of additional power

FRONTERA PROJECT(S) - SCOPE

- Frontera is made up of multiple NSF Cooperative Agreements:
  - **Acquisition – procure the system, everything up to acceptance and production. ($60M)**
  - Operations & Maintenance – (This proposal) from system production, expenses (mostly personnel) to operate and maintain ($12M/year)
  - Phase 2 planning – forward-looking and design/requirements activities towards a potential follow-on system with 10x the capabilities ($2M/year)

ACQUISITION TIMELINES

- Awarded August 27th, 2018
- Datacenter re-fit completed January
- SOW/Purchase Order (Dell) sent October.
- Storage/Network rack deliveries began February.
- Compute rack deliveries (orig. Feb) delayed until April, completed in May.
- First user jobs end of May, limited user access in June, all users granted access by early July; Acceptance and Production 3 months later.

- With the late start, we delayed the acceptance review ~1 month to debug all the problems in the system.

MILESTONES AS OF ISC 2019

- **Frontera is the #5 ranked system in the world.**
- **Fastest primarily Intel-based system**
- **Highest ranked Dell system ever.**
- **Highest ranked system at any university in the world**
- Frontera and Stampede2 are #1 and #2 among US Universities (and Lonestar5 is still in the Top 10).
- Early Science Period is now underway

FRONTERA SYSTEM --- HARDWARE

- Primary compute system: DellEMC and Intel
  - 35-40 PetaFLOPS Peak Performance
- 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

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*
    - 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

- Lustre, POSIX, and that’s it.
- Disk: 50PB
- Flash: 3PB
- We have come to believe that most user’s codes accessing the filesystem look like this:

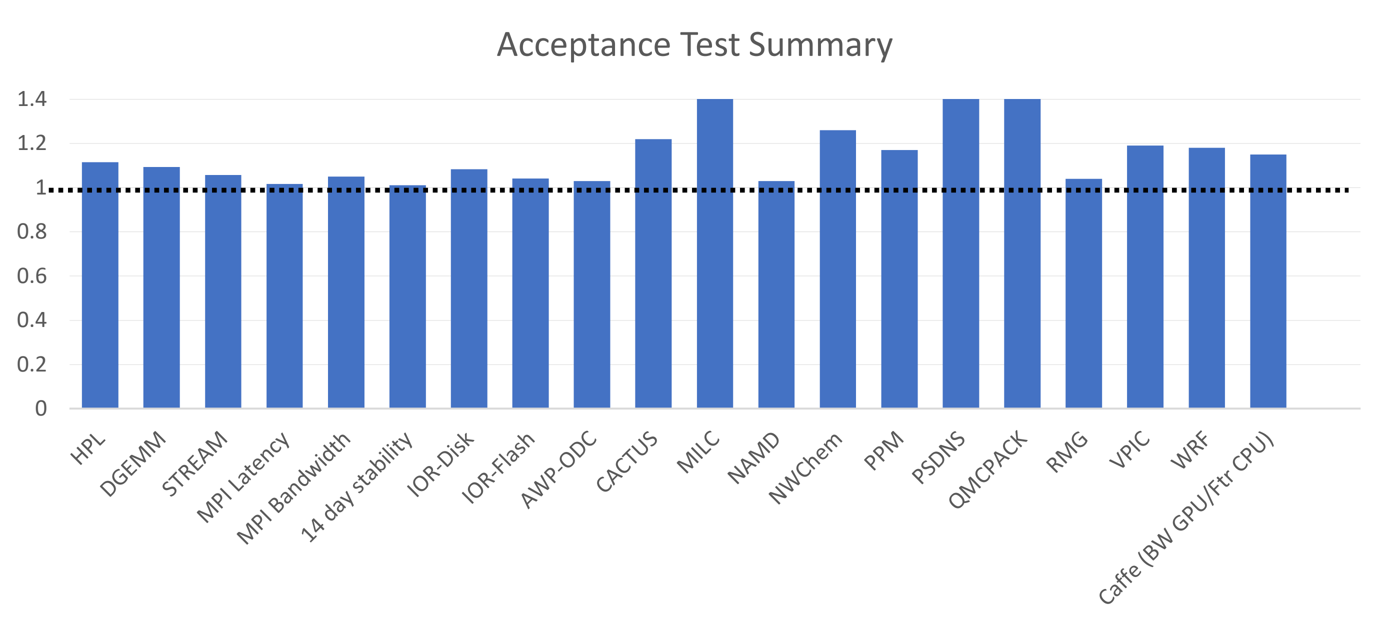
```
While (1) {
   fork();
   fopen();
   fclose(); //optional
}

Mpirun -np 80000 kill_the_filesystem
```

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).

STATUS



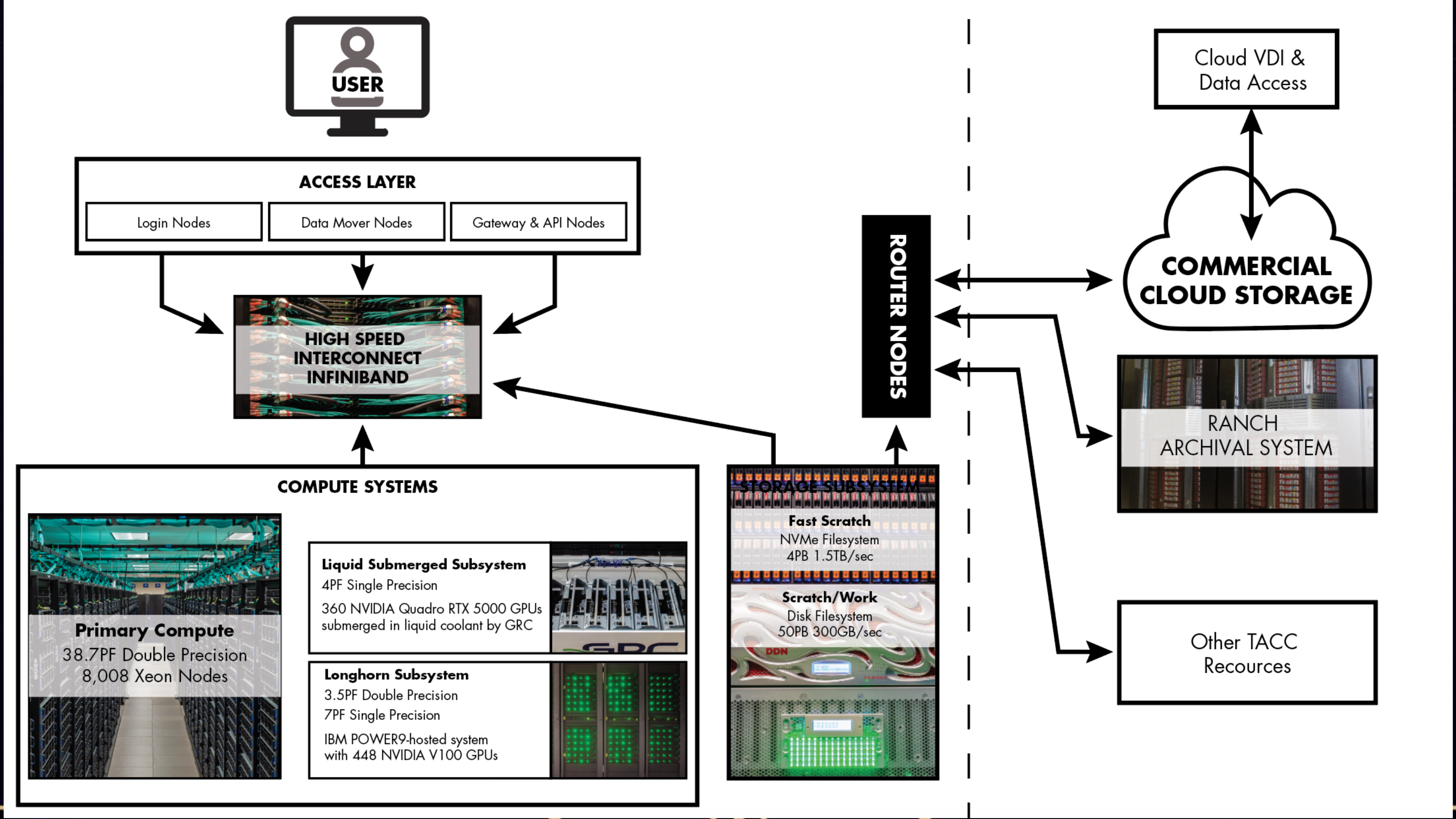
- **Of our 20 numerical measures of acceptance, as outlined in the proposal and project execution plan (PEP), we are “past the post” on all 20.**
- **This represents a mix of full applications, low level hardware performance, and system reliability.**

LARGE “MEMORY”, OR FASTER I/O

- *Panel note: This is technically Stampede2, but will be available as well.*

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

FRONTERA IS A GREAT MACHINE – AND MORE THAN A MACHINE



USER

ACCESS LAYER

Login Nodes

Data Mover Nodes

Gateway & API Nodes

HIGH SPEED INTERCONNECT INFINIBAND

ROUTER NODES

COMPUTE SYSTEMS

Primary Compute
38.7PF Double Precision
8,008 Xeon Nodes

Liquid Submerged Subsystem
4PF Single Precision
360 NVIDIA Quadro RTX 5000 GPUs submerged in liquid coolant by GRC

Longhorn Subsystem
3.5PF Double Precision
7PF Single Precision
IBM POWER9-hosted system with 448 NVIDIA V100 GPUs

STORAGE SUBSYSTEM

Fast Scratch
NVMe Filesystem
4PB 1.5TB/sec

Scratch/Work
Disk Filesystem
50PB 300GB/sec

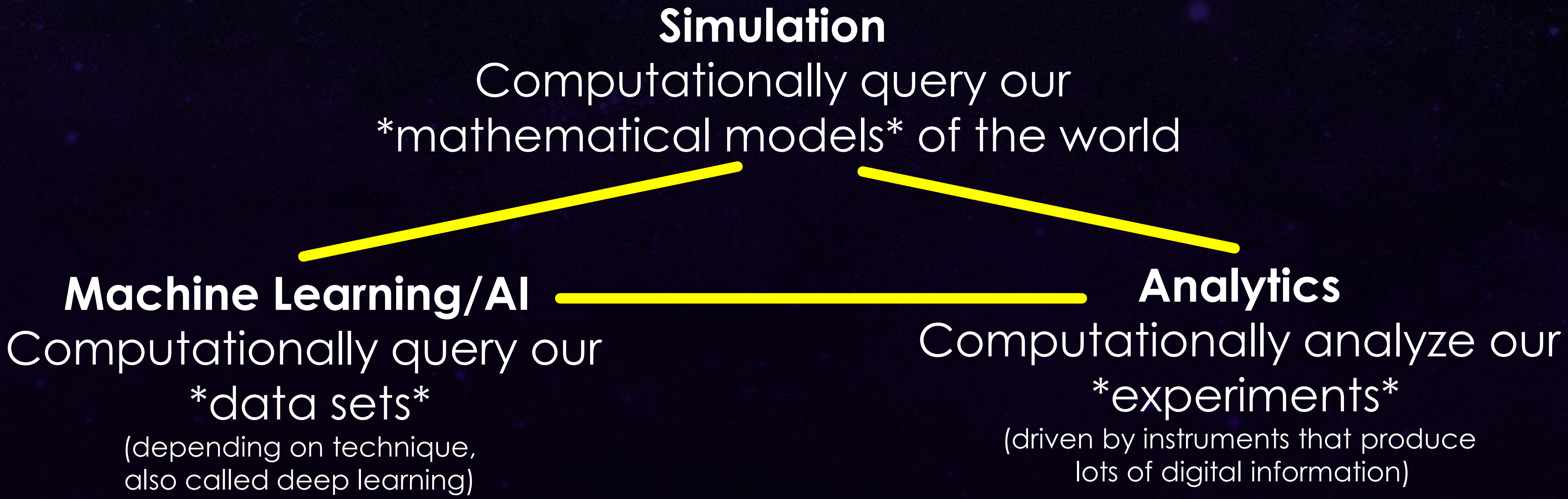
Cloud VDI & Data Access

COMMERCIAL CLOUD STORAGE

RANCH ARCHIVAL SYSTEM

Other TACC Recources

MODERN COMPUTATIONAL SCIENCE



Simulation
Computationally query our
mathematical models of the world

Machine Learning/AI
Computationally query our
data sets
(depending on technique,
also called deep learning)

Analytics
Computationally analyze our
experiments
(driven by instruments that produce
lots of digital information)

We would argue that modern science and engineering combine all three

SUMMARY

- We are >40k jobs in – this system is ready for production.
- It's amazing now – it will get better over time.
- GPU acceptance is up next.
- We are confident the firmware fix will improve reliability even further.

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

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
TACC
NSF
TEXAS