COMPUTING FOR THE ENDLESS FRONTIER

Dan Stanzione
Executive Director
Associate Vice President for Research

SC Asia - Singapore

March 2019

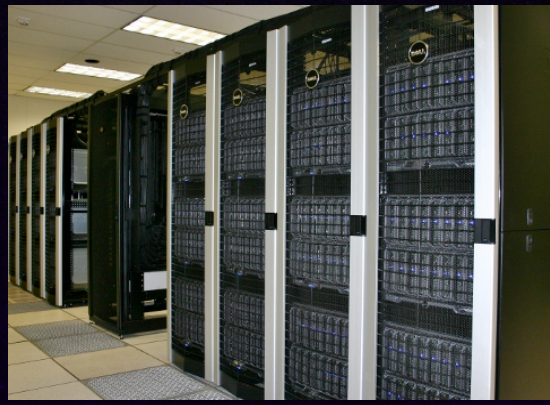
FRONTERA SYSTEM --- PROJECT

- *A new, NSF supported project to do 3 things:*
- Deploy a system in 2019 for the largest problems scientists and engineers currently face.
- Support and operate this system for 5 years.
- Plan a potential phase 2 system, with 10x the capabilities, for the future challenges scientists will face.

A BIT OF HISTORY ABOUT TACC

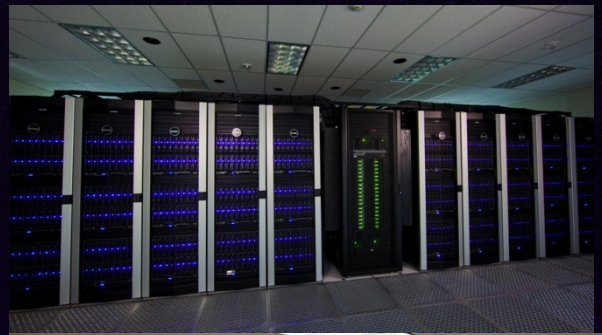
TACC LAUNCHED IN JUNE, 2001 AFTER EXTERNAL REVIEW

- Original HPC effort launched by Hans Mark as System initiative in 1986 at $30M
  - Passed between Austin and System several times over next 15 years.
- In 2001, budget of $600k staff of 12 (some shared).
- 50GF computing resource (1/200,000th of the current system).



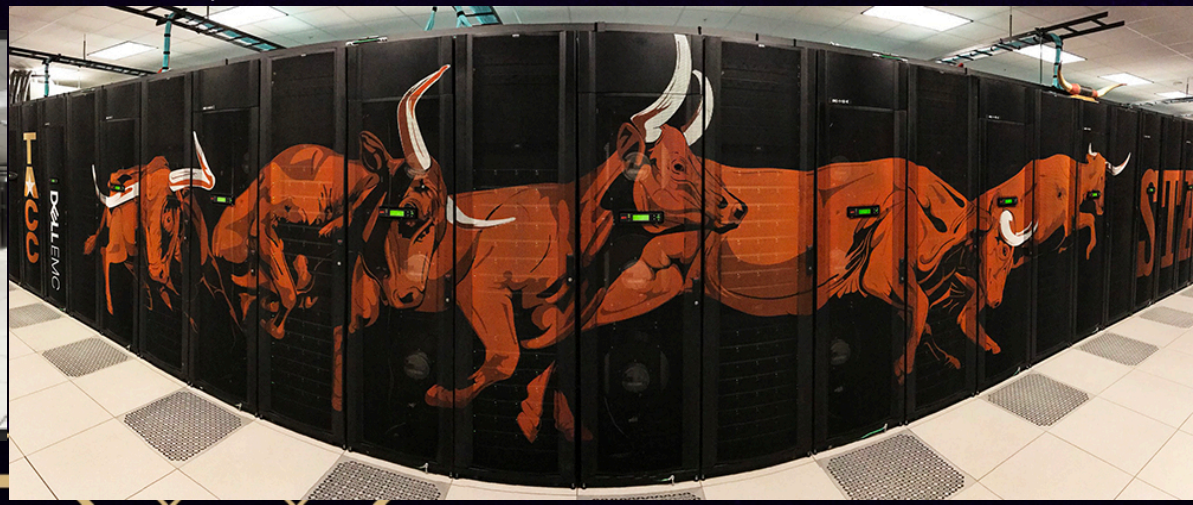
RAPID GROWTH FROM THEN TO NOW…

- 2003 – First Terascale Linux cluster for open science (#26)
- 2004 – NSF funding to join the Teragrid
- 2006 – UT System Partnership to provide Lonestar-3 (#12)
- **2007 - $59M NSF award – largest in UT history – to deploy Ranger, the world's largest open system (#4)**
- 2008 – funding for new Vis software and launch of revamped visualization lab.
- 2009 - $50M iPlant Collaborative award (largest NSF bioinformatics award) moves a major component to TACC, life sciences group launched.
  - In 2009, we reached, 65 employees.

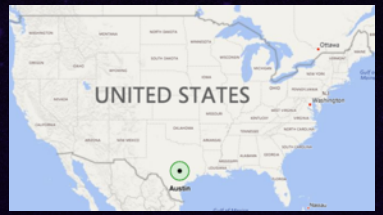


NOW, A WORLD LEADER IN CYBERINFRASTRUCTURE

- 2010, TACC becomes a core partner (1 of 4) in XSEDE, the TeraGrid Replacement
- 2012, Stampede replaces Ranger with new $51.5M NSF Award
- 2013, iPlant is renewed, expanded to $100M
- 2015, Wrangler, first data intensive supercomputer is deployed.
- 2015, Chameleon cloud is launched
- 2015, DesignSafe, the cyberinfrastructure for natural hazard engineering, is launched.
- 2016 Stampede-2 awarded the largest academic system in the United States, 2017-2021.



TACC AT A GLANCE



Personnel

175 Staff (~70 PhD)

Facilities

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

Systems and Services

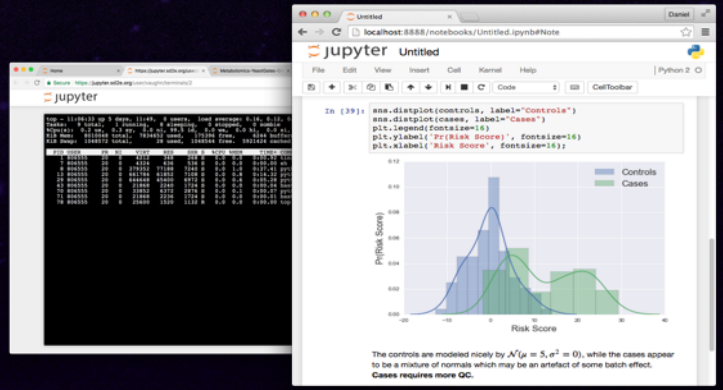
Two Billion compute hours per year
5 Billion files, 75 Petabytes of Data,
Hundreds of Public Datasets

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

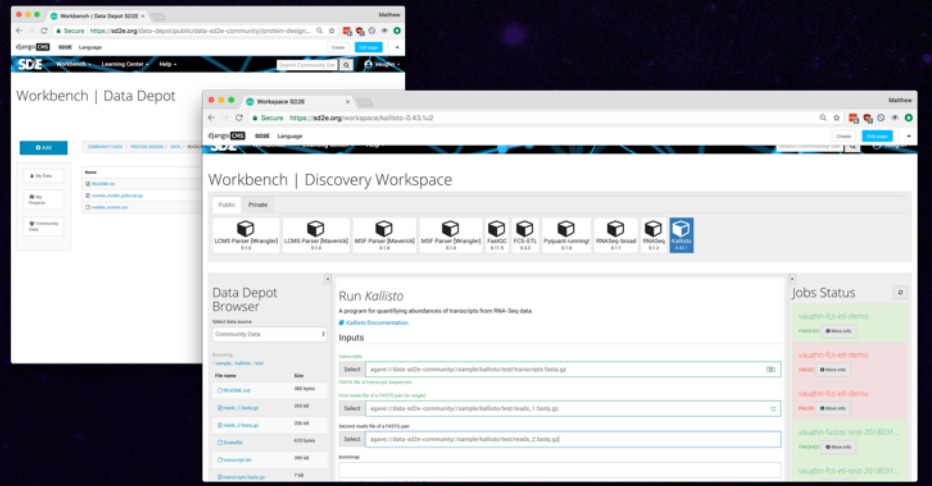
HPC DOESN'T LOOK LIKE IT USED TO. . .

HPC-Enabled Jupyter Notebooks
Narrative analytics and exploration environment



Event-driven Data Processing
Extensible end-to-end framework to integrate planning, experimentation, validation and analytics

Web Portal
Data management and accessible batch computing



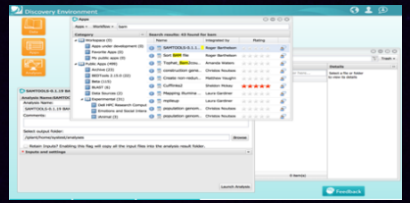
From Batch Processing and single simulations of many MPI Tasks – to that, plus new modes of computing, automated workflows, users who avoid the command line, reproducibility and data reuse, collaboration, end-to-end data management,

- **Simulation** where we have models
- **Machine Learning** where we have data or incomplete models

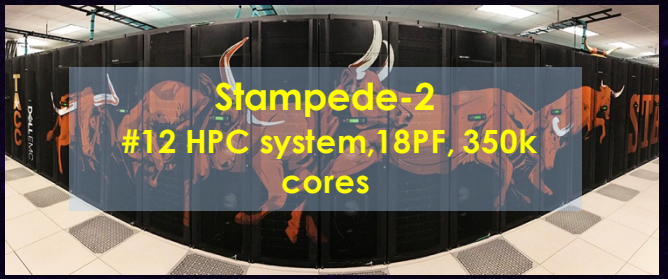
And most things are a blend of most of these. . .

SUPPORTING AN EVOLVING CYBERINFRASTRUCTURE

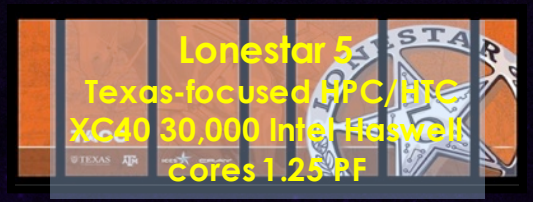
- Success in Computational/Data Intensive Science and Engineering takes more than systems.
- Modern Cyberinfrastructure requires many modes of computing, many skillsets, and many parts of the scientific workflow.
  - Data lifecycle, reproducibility, sharing and collaboration, event driven processing, APIs, etc.
- Our team and software investments are larger than our system investments
  - Advanced Intefaces – Web front ends, Rest API, Vis/VR/AR
  - Algorithms – Partnerships with ICES @ UT to shape future systems, applications and libraries.



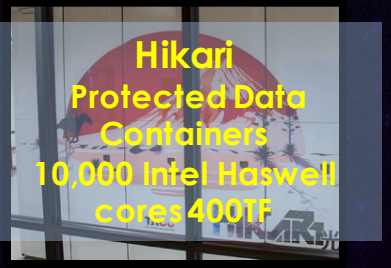
AN ECOSYSTEM FOR EXTREME SCALE SUPERCOMPUTING



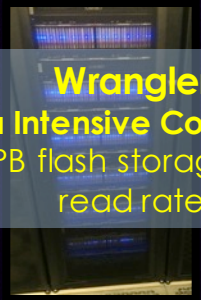
Stampede-2
#12 HPC system,18PF, 350k cores



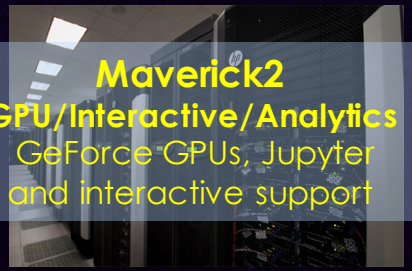
Lonestar 5
Texas-focused HPC/HTC
XC40 30,000 Intel Haswell cores 1.25 PF



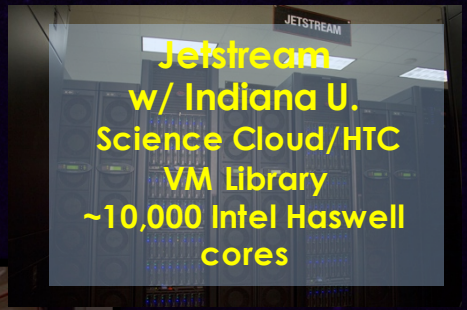
Hikari
Protected Data Containers
10,000 Intel Haswell cores 400TF



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



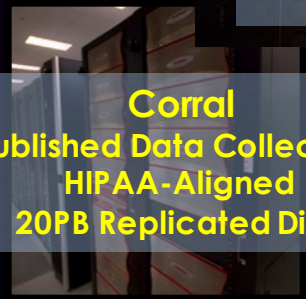
Maverick2
GPU/Interactive/Analytics
GeForce GPUs, Jupyter and interactive support



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

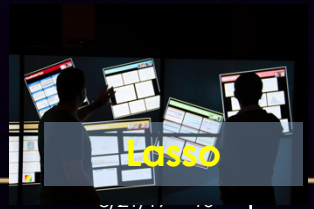
Stockyard
Shared Storage Across TACC
30PB, Lustre

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



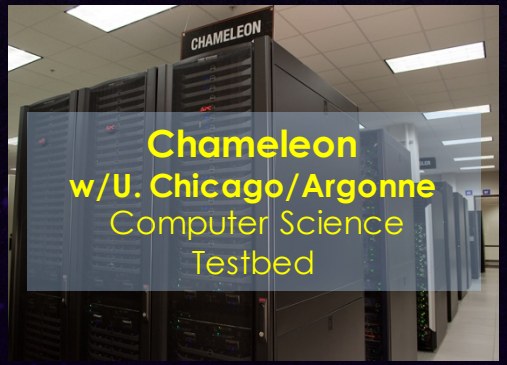
Corral
Published Data Collections
HIPAA-Aligned
20PB Replicated Disk,

Rodeo

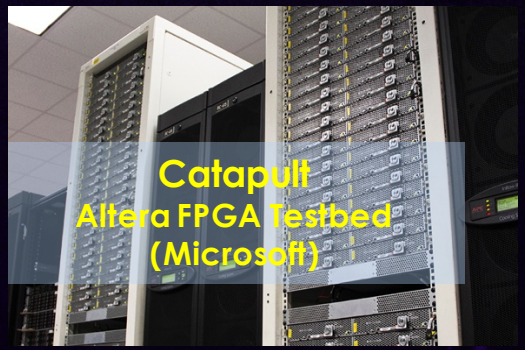


Lasso

EXPERIMENTAL SYSTEMS



Chameleon
w/U. Chicago/Argonne
Computer Science
Testbed



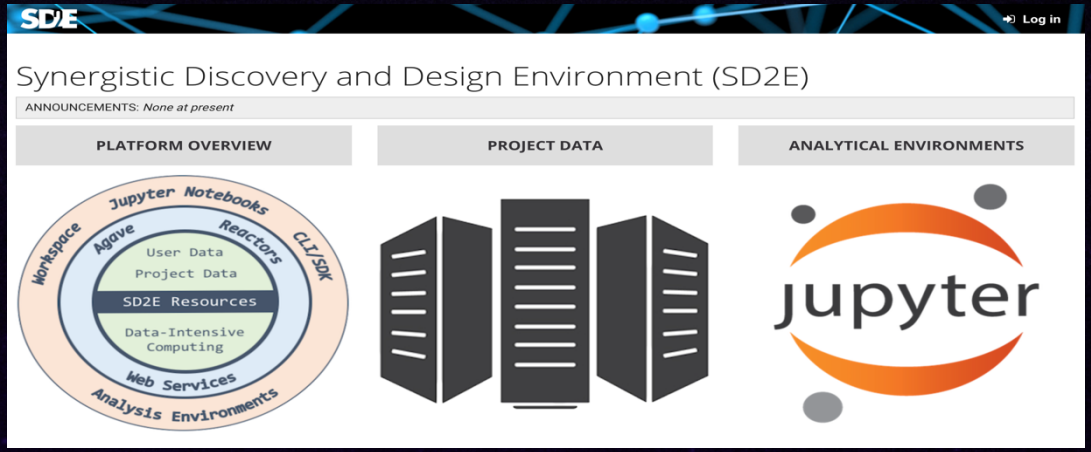
Catapult
Altera FPGA Testbed
(Microsoft)

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

Discovery
New Processor/Storage
Benchmarking

Rustler
Object Storage Testbed

AN EXEMPLAR PROJECT – SD2E

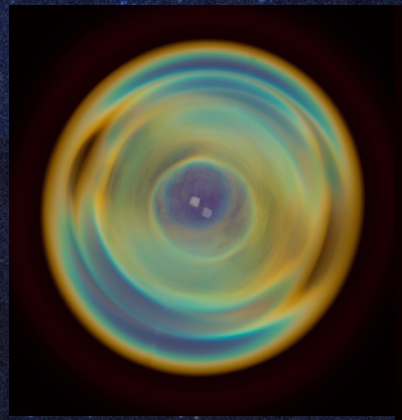


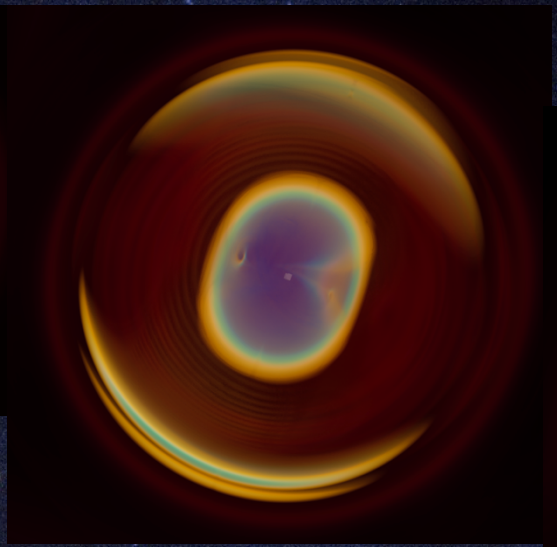
- DARPA – "*Synergistic Discovery and Design* (SD2)"
- Vision: to "develop data-driven methods to accelerate scientific discovery and robust design in domains that lack complete models."
- Initial focus in synthetic biology; ~six data provider teams, ~15 modeling teams, **TACC for platform**
- Cloud-based tools to collect, integrate, and analyze diverse data types; Promote collaboration and interaction across computational skill levels; Enable a reproducible and explainable research computing lifecycle; **Enhance, amplify, and link the capabilities of every SD2 performer**

TACC SUPPORTS AN INCREDIBLE AMOUNT & DIVERSITY OF RESEARCH

- Our request rate (NSF Sytems) continues to be about 5-10x what we can deliver
  - More than 2,000 unique users run jobs in any given month (Stampede2)
    - (More than 12k people have run in production on Stampede2, spanning 3,500+ projects).
  - 2+ **million** successful jobs last year.
  - We estimate well 0ver 35,000 use TACC systems via Web or API.

COSMOS GRAVITATIONAL WAVES STUDY





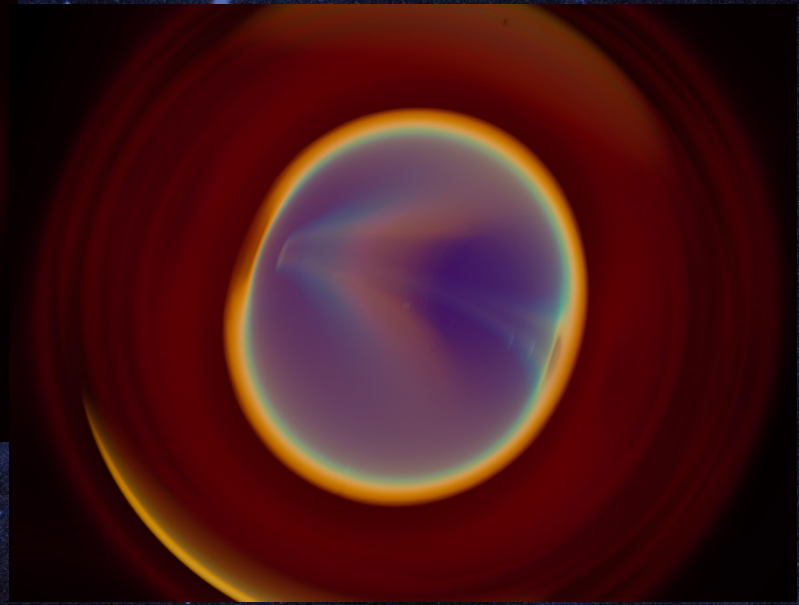


Image Credits:

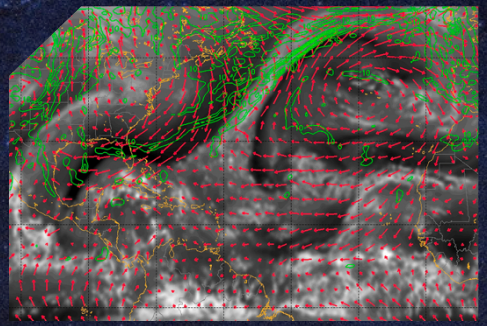
Greg Abram – TACC

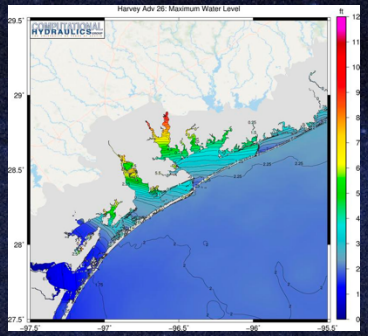
Francesca Samsel – CAT

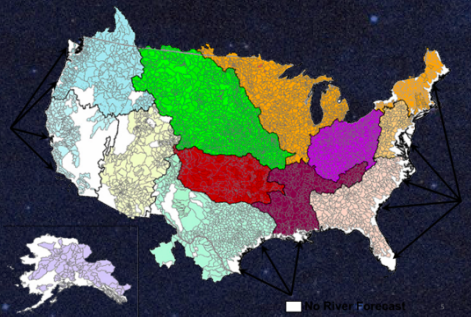
Carson Brownlee - Intel

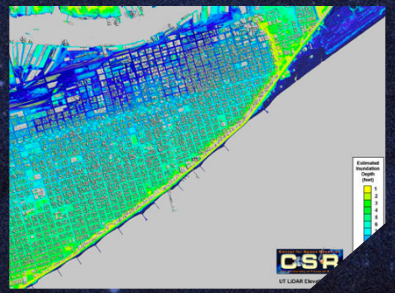
Markus Kunesch, Juha Jäykkä, Pau Figueras, Paul Shellard

Center for Theoretical Cosmology, University of Cambridge









HARVEY

- Next Generation Storm Forecasting (with Penn State)
- Storm Surge Modeling (with Clint Dawson UT Austin)
- Preliminary river flooding and inundation maps (David Maidment UT Austin)
- Remote Image Integration and Assimilation (Center for Space Research, UT Austin)

MASSIVE DATA SET WORTHY OF ROSS ICE SHELF ITSELF

TACC partners with Lamont-Doherty Earth Observatory (LDEO) to host for one of the country's largest earth sciences data collections

- Managing hundreds of TB using Stampede2, Corral, and Ranch: storage, provenance, visualization, and public access
- Achieved 10x workflow speedup by moving to TACC (from 50 hrs down to 5 hrs for transfer and analysis tasks)

"...partnership...with TACC shows [it's] possible to manage…this level of data in a cost-effective, user-friendly and easily accessible manner…"

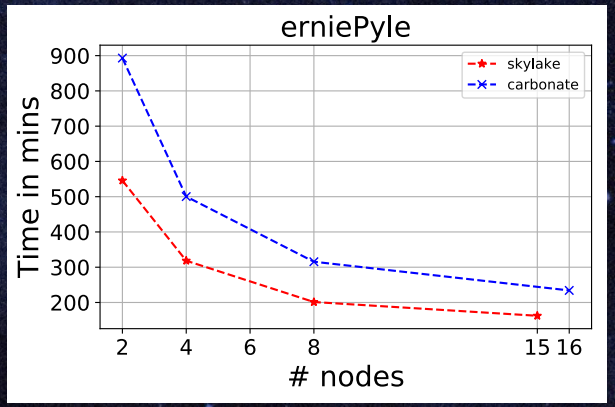
Image courtesy Oceanwide Expeditions.

TACC News Release

PI Lingling Dong, Columbia University
XSEDE support to multidisciplinary, multi-institutional Rosetta project

PHOTOGRAMETRY ON KNL

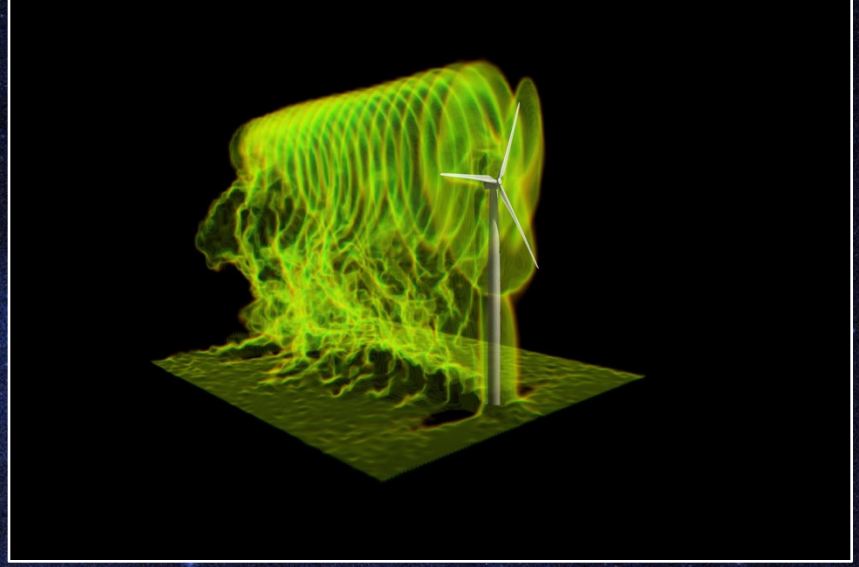
- Effort lead by IU (Wernert, McCombs, Ruan, Tuna)
- Create 3d point cloud & Mesh Model of texture/color map using tiled 2d images
  - Camera panoramas, Drone Survey
  - Future underwater shipwrecks/reefs
- Using Agisoft Photoscan software
  - More speedup from larger datasets
- Exploring OpenSource alternatives
  - Adding MPI layers needed



REAPING POWER FROM WIND FARMS

Multi-Scale Model of Wind Turbines

- Optimized control algorithm improves design choices
- New high-res models add nacelle and tower effects
- Blind comparisons to wind tunnel data demonstrate dramatic improvements in accuracy
- Potential to increase power by 6-7% ($600m/yr nationwide)



"TACC…give[s] us a competitive advantage…"

Graphic from Wind Energy, 2017.

Christian Santoni, Kenneth Carrasquillo, Isnardo Arenas-Navarro, and Stefano Leonardi

UT Dallas, US/European collaboration (UTRC, NSF-PIRE 1243482)

TACC News Release

"Using commodity HPC servers...the time to data-driven discovery is reduced and overall efficiency can be significantly increased." (Niall Gaffney, TACC)

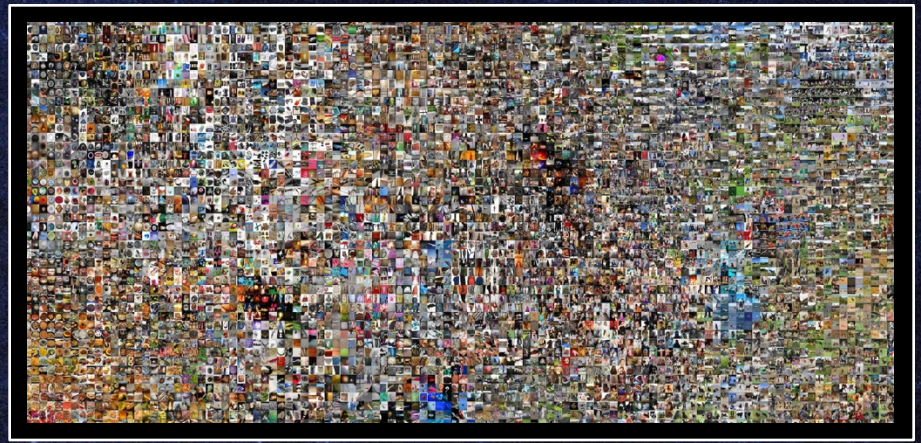
RECORD ACHIEVED ON AI BENCHMARK

TACC, Berkeley, Cal Davis collaborate on large-scale AI runs

- Research demonstrating the potential of commodity hardware for AI
- Skylake ImageNet benchmark: (100 epochs, 11 min, 1024 nodes) -- fastest result at time of publication
- Knights Landing ImageNet benchmark (90 epochs, 20 min, 2048 nodes) – 3x faster than Facebook, with higher large-batch accuracy

[TACC News Release]

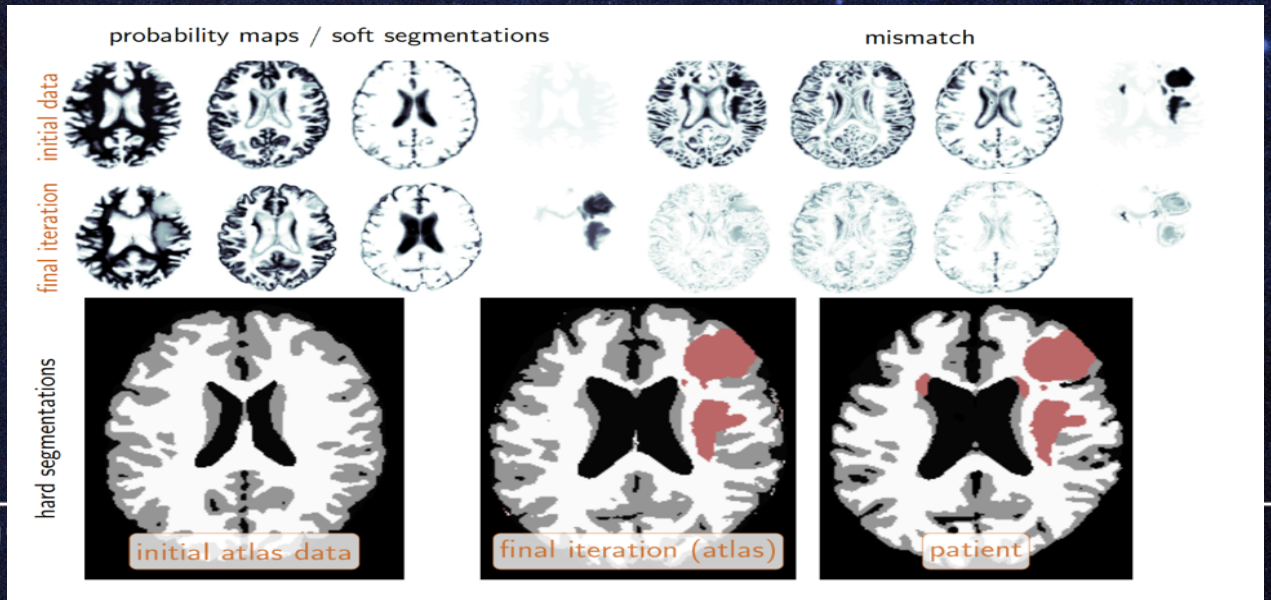
Graphic credit Andrej Karpathy



Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, Kurt Keutzer

BRAIN TUMOR SEGMENTATION

- A team of researchers led by George Biros from The University of Texas at Austin scored in the top 25% of participants in the Multimodal Brain Tumor Segmentation Challenge 2017 (BRaTS'17) enabled by Stampede2 and other TACC resources.
- In the challenge, research groups presented methods and results of computer-aided identification and classification of brain tumors, as well as different types of cancerous regions.
- The team's method combined biophysical models of tumor growth with machine learning algorithms for the analysis of Magnetic Resonance imaging data of glioma patients.



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

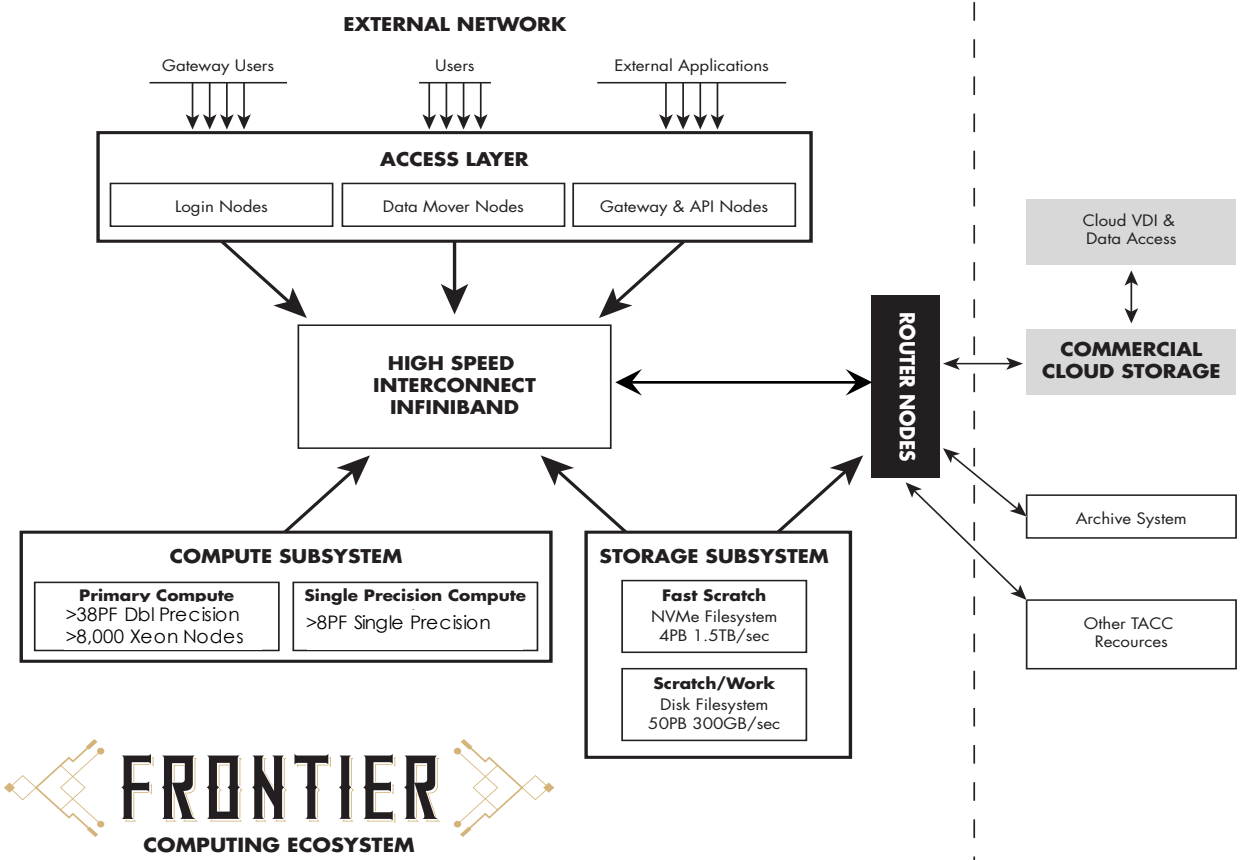
DESIGN DECISIONS - PROCESSOR

- The architecture is in many ways "boring" if you are an HPC journalist, architect, or general junkie.
  - We have found that the way users refer to this kind of configuration is "useful".
- No one has to recode for higher clock rate. We have abandoned the normal "HPC SKUS" of Xeon, in favor of the Platinum top bin parts – the ones that are 205W per socket.
  - Which, coincidentally, means the clock rate is higher on every core, whether you can scale in parallel or not.
  - Users tend to consider power efficiency "our problem".
  - This also means there is *no* air cooled way to run these chips.
- Versus Stampede2, we are pushing up clock rate, core count, and main memory speed.
  - This is as close to "free" performance as we can give you.

DESIGN DECISIONS - FILESYSTEM

- Scalable Filesystems are always the weakest part of the system.
  - Almost the only part of the system where bad behavior by one user can affect the performance of a *different* user.
- Filesystems are built for the aggregate user demand – rarely does one user stress *all* the dimensions of filesystems (Bandwidth, Capacity, IOPS, etc.)
- We will divide the "scratch" filesystem into 4 pieces
  - One with very high bandwidth
  - 3 at about the same scale as Stampede, and divide the users.
- Much more aggregate capability – but no need to push scaling past ranges at which we have already been successful.
  - Expect higher reliability from perspective of individual users
  - Everything POSIX, no "exotic" things from user perspective.

ORIGINAL SYSTEM OVERVIEW



FRONTERA SYSTEM --- INFRASTRUCTURE

- Frontera will consume almost 6 Megawatts of Power at Peak
- Direct water cooling of primary compute racks (CoolIT/DellEMC)
- Oil immersion Cooling (GRC)
- Solar, Wind inputs.

TACC Machine Room Chilled Water Plant

THE TEAM - INSTITUTIONS

- Operations: TACC, Ohio State University (MPI/Network support), Cornell (Online Training), Texas A&M (Campus Bridging)
- Science and Technology Drivers and Phase 2 Planning: Cal Tech, University of Chicago, Cornell, UC-Davis, Georgia Tech, Princeton, Stanford, Utah
- Vendors: DellEMC, Intel, Mellanox, DataDirect Networks, GRC, CoolIT, Amazon, Microsoft, Google

SYSTEM SUPPORT ACTIVITIES
THE "TRADITIONAL"

- Stuff you always expect from us:
  - Extended Collaborative Support (under of course yet another name) from experts in HPC, Vis, Data, AI, Life Sciences, etc.
  - Online and in person training, online documentation.
  - Ticket support, 24x7 staffing
  - Comprehensive SW stack – the usual ~2,000 RPMs.
  - Archive access – scalable to an Exabyte.
  - Shared Work Filesystem – same space across the ecosystem.
  - Queues for very large and very long – plus small and short, and backfill tuned so that works OK.
  - Reservations and priority tuning to give Quality of Service guarantees when needed.

SYSTEM SUPPORT ACTIVITIES
THE "TRADITIONAL"

- Stuff that is slightly newer (but you should still start to expect from us) :
  - Auto-tuned MPI stacks
  - Automated Performance Monitoring, with data mining to drive consulting
  - Slack channels for user support (it's a much smaller user community).

NEW SYSTEM SUPPORT ACTIVITIES

- Full Containerization support (this platform, Stampede, and *every other* platform now and future.
- Support for Controlled Unclassified Information (i.e. Protected Data)
- Application servers for persistent VMs to support services for automation.
  - Data Transfer (ie. Globus)
  - Our native REST APIs
  - Other service APIs as needed – OSG (for Atlas, CMS, LIGO)
  - Possibly other services (Pegasus, perhaps things like metagenomics workflows)

NEW SYSTEM SUPPORT ACTIVITIES

- Built on these services, Portal/Gateway support
  - Close collaboration at TACC with SGCI (led by SDSC).
  - "Default" Frontera portals for: (not all in year 1).
    - Job submission, workflow building, status, etc.
    - Data Management – not just in/out and on the system itself, but full lifecycle – archive/collections system/cloud migration, metadata management, publishing and DOIs.
    - Geospatial
    - ML/AI Application services.
    - Vis/Analytics
    - Interactive/Jupyter
  - And, of course, support to roll your own, or get existing community ones integrated properly.

PHASE 2 PROTOTYPES

- Allocations will include access to testbed systems with future/alternative architectures
  - Some at TACC, e.g. FPGA systems, Optane NVDIMM, {as yet unnamed 2021, 2023}.
  - Some with partners – a Quantum Simulator at Stanford.
  - Some with the commercial cloud – Tensor Processors, etc.
- **Fifty nodes with Intel Optane technology will be deployed next year in conjunction with the production system**
  - Checkpoint file system? Local checkpoints to tolerate soft failures? Replace large memory nodes? Revive "out of core" computing? In-memory databases?
- Any resulting phase 2 system is going to be the result, at least in part, of actual users measured on actual systems, including at looking at, what they might actually *want* to run on.
- Eval around the world – keep close tabs on what is happening elsewhere (sometimes by formal partnership or exchange – ANL, ORNL, China, Europe).

LEVERAGE THE ECOSYSTEM

- At TACC:
  - /work shared file system between platforms.
  - Ranch archive system
  - Corral data collections system
  - Rodeo VM Farm
  - Agave tenants
- And at other places around the country:
  - OSN (both "Open" and "Oklahoma").
  - Public data repositories
  - Data Transfer Software (i.e. Globus)
  - Google Dataset search, community portals.
  - Public cloud providers (Microsoft, Amazon, Google)
    - Options to publish data in the cloud, use innovative cloud services in scientific workflows, and access to new technologies each year as we plan phase 2.

STRATEGIC PARTNERSHIP WITH COMMERCIAL CLOUDS

- Cloud/HPC is *not* an either/or. (And in many ways, we are just a specialized cloud).
- Utilize cloud strengths:
  - Options for publishing/sustaining data and data services
  - Access to unique services in automated workflow; VDI (i.e. image tagging, NLP, who knows what. . .)
  - Limited access to *every* new node technology for evaluation
    - FPGA, Tensor, Quantum, Neuromorphic, GPU, etc.
  - We will explore some bursting tech for more "throughput" style jobs – but I think the first 3 bullets are much more important. . .

THE BROADER TACC ECOSYSTEM
DISCOVERY SCIENCE AT ALL SCALES

FRONTERA

Leadership/Discovery Science

Longhorn

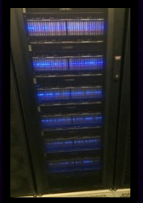
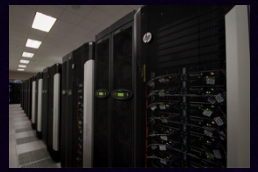
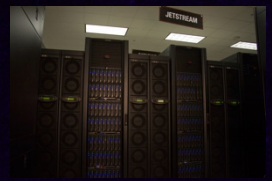
AI/ML/DL @ Scale

Testbeds

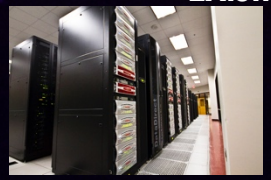
Catapult (Upgrade)
Non-Volatile Memory
Quantum
Future . . .

Existing TACC Computing Systems



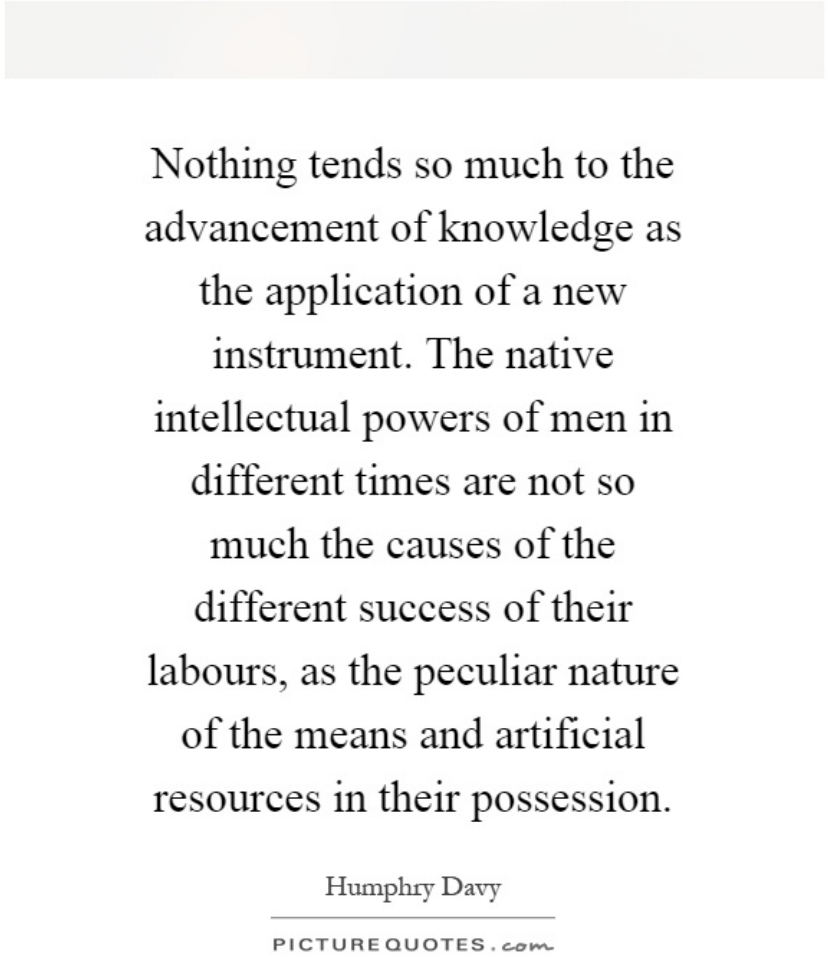
Existing TACC Storage Systems



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

- **Humphry Davy, Inventor of Electrochemistry, 1812**
- (Pretty sure he was talking about our machine).




FRONTERA
TACC
NSF
TEXAS