HPC, CLOUD SERVICES AND THE EVOLUTION OF SCIENTIFIC COMPUTING

December 7th, 2017

Dan Stanzione, TACC

OUTLINE

- A bit about TACC
- HPC vs. the Cloud – is this a debate?
- HPC and the Cloud
- HPC is the Cloud?

TACC AT A GLANCE

Personnel

160 Staff (~70 PhD)

Facilities

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

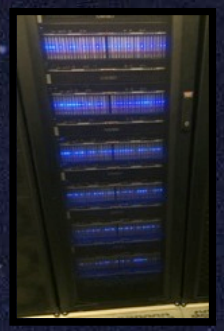
Systems and Services

A Billion compute hours per year
5 Billion files, 50 Petabytes of Data,
Hundreds of Public Datasets

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

EXTREME SCALE SUPERCOMPUTING



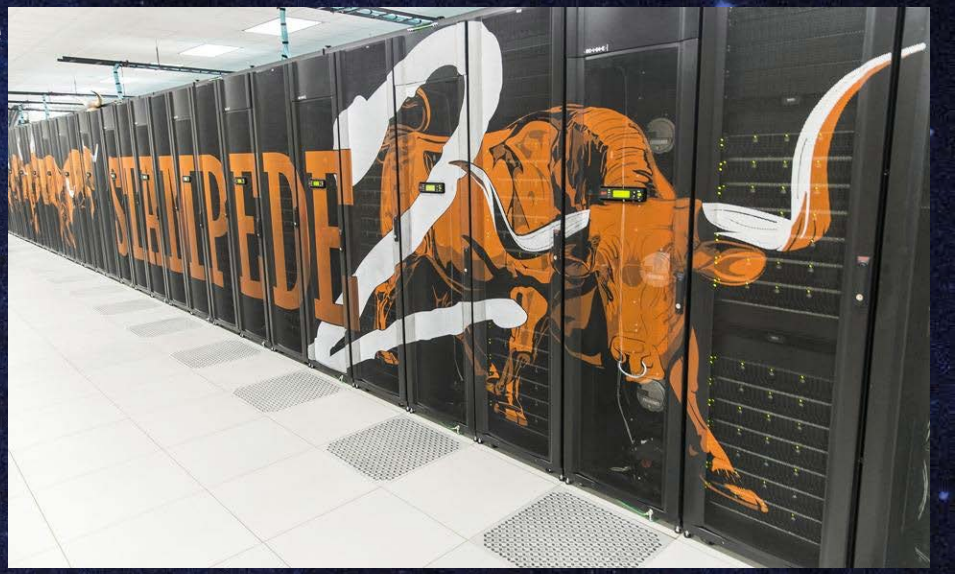
Stampede-2
•~ #12 HPC system in the world for computation 350k CPU core 18 PF
Lonestar 5
• Texas-focused Cray XC40 30,000 Intel Haswell cores 1.25 PF
Wrangler
• 0.6 PB usable DSSD flash storage w 1 TB/s read rate + 10 PB Lustre
Maverick
•132 Fat nodes w dual 10 core Ivy Bridge + NVIDIA Kepler K40 GPGPU
Chameleon & Jetstream Cloud
•1400 nodes OpenStack
Disk and Tape Storage
• 100+ PB storage in HIPAA-aligned data center

Hikari
- 380V DC Green computing system parternship with NEDO and NTT. 10k Haswell cores. HVDC and Solar (partial)
- Support for container ecosystem

STAMPEDE 2

STAMPEDE 2

- Funded by NSF as a renewal of the original Stampede project.
- The largest XSEDE resource (and largest university-based system).
- 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.



LIGO + COSMOLOGICAL SIMULATION

- One of the early users on Stampede 2 was the team from Stephen Hawking's lab in Cambridge.
- At an IXPUG In-situ visualization workshop at TACC last Ma several scientists from this group teamed up with the Intel/TACC Software Defined Visualization team to pair the GR-CHOMBO code on Stampede 2 with the new Ray tracing stack.
  - (Repeat workshop tomorrow, talk about this here earlier in the day).
- While the running simulation was still in RAM on S2, we simultaneously used some of the available processors to render the data for steering and new insight.

CAMBRIDGE COLLABORATION

And we've kept going from there, to explore more visualizations!

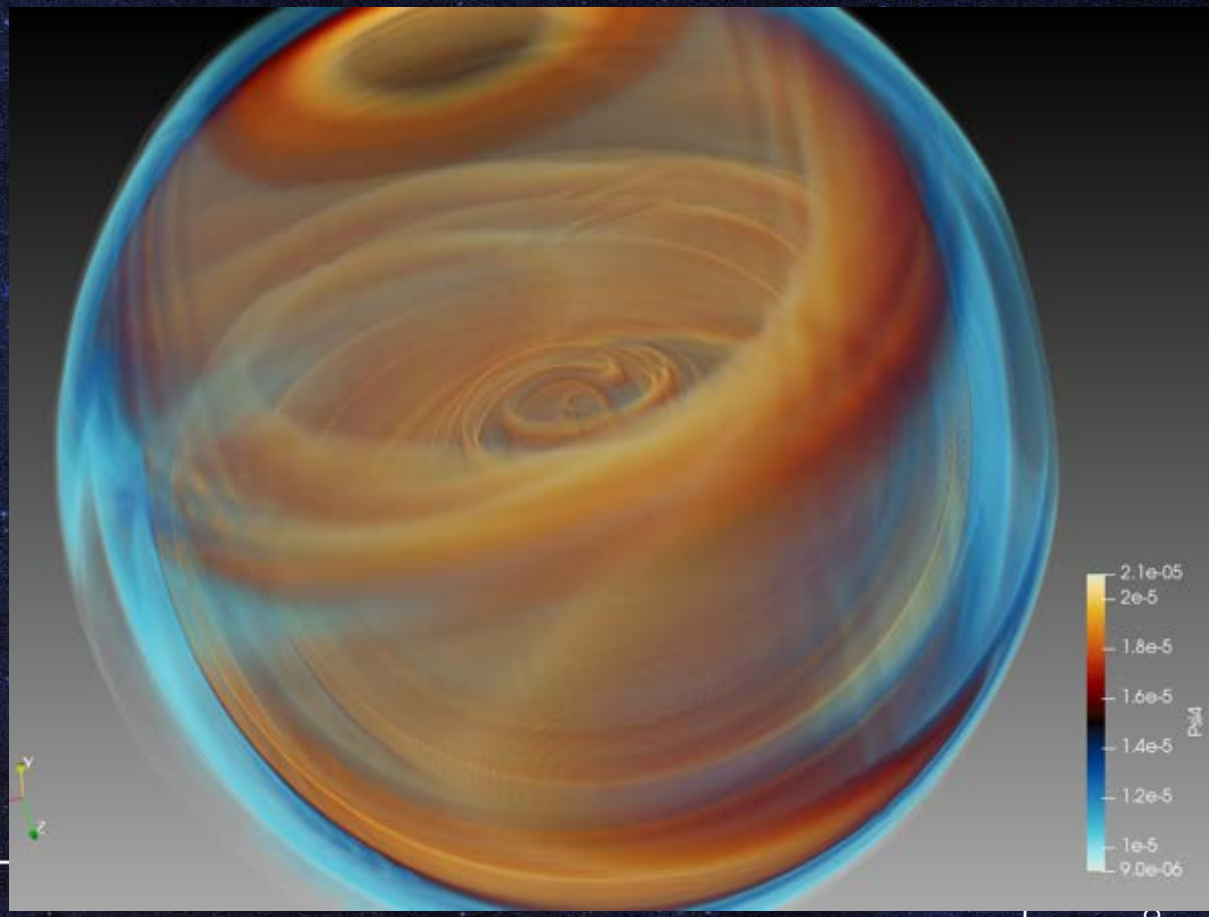


Image Credits:

Greg Abram– TACC

Francesca Samsel - CAT

Markus Kunesch, Juha Jäykkä, Pau Figueras, Paul Shellard

Center for Theoretical Cosmology, University of Cambridge

AN EXEMPLAR SCIENTIFIC COLLABORATION

- Lots of good things happened here – the way it should.
  - Two large teams got together to make something happen.
  - We successfully got good results.
  - Then we made them way, way more efficient, using modern techniques on modern hardware platforms.
  - Success led to learning from each other, and culture changes around software.
  - Success in this led directly to *more* success with *more* teams.
  - Everybody treated everyone else as colleagues – theoreticians, experimentalists, and computational experts (perhaps there is hope for peace in our time, after all).
  - Oh yeah, ***Along the way, groundbreaking science and a nobel prize.***

HPC VS. THE CLOUD

- People often equate HPC and Cloud Computing
  - HPC has a big datacenter full of servers.
  - The Cloud has a big datacenter(s) full of servers
    - In this sense, both Accounting and Particle Physics involve Math, so must be the same.
  - We both like the word "scalable" though we use it in very different ways.

HPC VS. THE CLOUD

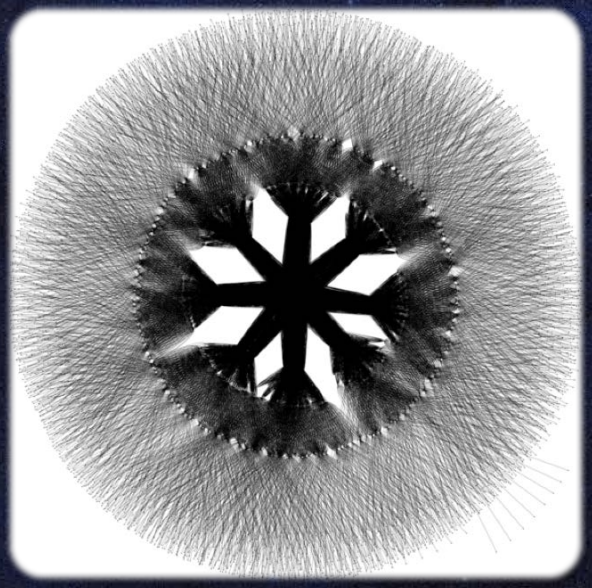
- This basic commonality actually does give us a LOT in common.
  - Cloud and HPC providers care about:
    - Power, cooling, and datacenter efficiencies.
    - Costs of hardware, and reliability of components
    - Standardization of hardware platforms, provisioning, etc.

BUT ALSO A LOT THAT'S DIFFERENT

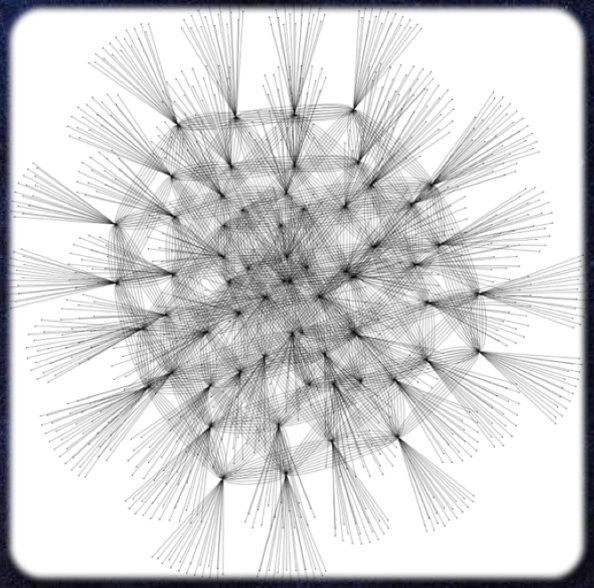
- HPC is about sustained Floating Point Performance.
- Usually, when an HPC person says "Scalable", they mean:
  - A Large number of synchronous, cooperating, tasks, working together to perform a single large computation.
  - Think global weather, colliding blackholes, jet airframe simulation
  - So what matters? Network *latency*; reproducible performance; vectorization, memory access times, I/O rate to a single file, etc.
- Usually, when a Cloud person says "Scalable" they mean:
  - More asynchronous transactions against what is likely a shared database.
  - Think scaling web site hits, search engines, airline reservation systems.
  - What matters? *Average* performance, cost per transaction, aggregate I/O, uptime, etc.

TOPOLOGY CONSIDERATIONS

- At scale, process mapping with respect to topology can have significant impact on applications



Full fat-tree (Stampede, TACC)

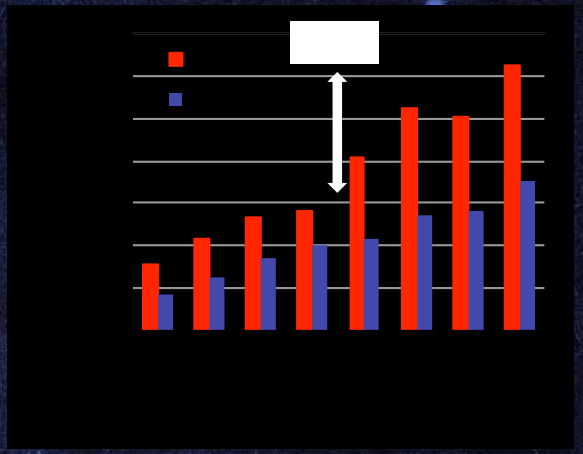


4x4x4 3D Torus (Gordon, SDSC)

TOPOLOGY CONSIDERATIONS

- Topology query service (now in production on Stampede) - NSF STCI with OSU, SDSC
  - caches the entire linear forwarding table (LFT) for each IB switch - via OpenSM plugin or *ibnetdiscover* tools
  - exposed via network (socket) interface such that an MPI stack (or user application) can query the service remotely
  - can return # of hops between each host or full directed route between any two hosts
  -

Nearest neighbor application benchmark from Stampede [courtesy H. Subramoni, SC 12]



- We will also be leveraging this service to perform topology-aware scheduling so that smaller user jobs will have their nodes placed closer together topologically
  - have created simple tool to create SLURM topology config file using above query service
  - works, but slows interactivity when users specifiy maximum # of switch hops desired during job submission

IS COST THE DIFFERENTIATOR?

- Because of these simlarities, the comparison of "HPC vs. the Cloud" or "HPC in the Cloud" is "Which costs less per core hour?"
- This is sort of a silly comparison. An inflatable raft is cheaper than a 200' yacht, but I know which one I would rather cost the Pacific in.
- But, since I brought it up. . .
  - HPC is probably cheaper.
  - Because we need to tightly couple applications, we might have a 10x, 100x, or 1,000x performance advantage *per hour* anyway.
  - On large genomics pipelines, we are an order of magnitude less than commercial clouds.
- This isn't a good measure because…

THINK ABOUT THE SOFTWARE

- Many cloud applications generate revenue – they are worth a lot, and are developed aggressively.
- Many scientific applications generate a *few* science results – they are developed *grudgingly*, by graduate students, over decades.
  - So, if you say "Let's replace POSIX with an Object store" then an airline reservation system will spend a few million dollars recoding.
  - But weather prediction will just stop. And most basic science.
- The Cloud solution to scaling POSIX-compliant parallel filesystems is "don't". That's a fair answer in that space. Not a good HPC answer.
- When colliding black holes, verification and validation is kind of hard. We can now get one experimental verification, for $2B.
  - For 15 years, LIGO codes had to produce *BIT FOR BIT* identical answers to change hardware platforms.

ULTIMATELY, THIS IS A SILLY COMPARISON

- OK, so HPC is good at delivering synchronous, tighly coupled simulations.
- The cloud really isn't.
  - So what? Why do we argue over the cost of using square pegs or round pegs?
- All of scientific computation is *MUCH BIGGER* than just the simulation.
- When I do a project, I also use Evernote, Trello, Slack, Box, etc. I don't try and run colllaboration tools on a supercomputer – and I've never done a cost analysis of why.


SCIENTIFIC COMPUTING THEN
C/C++/FORTRAN/PERL/SHELL
MPI
LAPACK/BLAS/PETSC
GRID ENGINE
UNIX*
X86/PPC/SPARC
COPROCESSORS

SCIENTIFIC COMPUTING NOW

LANGUAGES

- Python 2 & 3
- R
- Julia
- Perl
- Matlab
- Java
- Scala, Clojure, etc
- .NET
- C/C++
- Swift
- Haskell
- Go
- Javascript

FRAMEWORKS

- MapReduce Hadoop, Storm, Pachyderm, Cloudera
- Event & Streaming: Kinesis, Azure Stream Analytics, Camel, Streambase
- Deep/Machine Learning: Watson, Azure BI, Tensorflow, Caffe
- In-memory parsing: Kognito, Apache Spark
- Containers: Docker, Rocket, MESOS, Kubernetes
- Cloud: AWS, GCE, OpenStack, vCloud, Azure

HARDWARE

- Many-core computing - 50-100 threads/node*
  - Xeon / Xeon Phi
  - GPU
  - OpenPower
  - ARM
  - ShenWei
  - Google TPU
- Multi-level memory architecture
- Hierarchical storage
- FPGAs
- Quantum-like systems

THE NEW E-SCIENCE

- Many of the challenges we see today are largely driven by large scale computation – And computation centers like mine have of course focused on these kinds of problems for decades, with tremendous success – but many new kinds of problems are not just about computing.

- The new E-science is largely a problem of integrating, at scale, data collection, curation, and storage with advanced computing and analysis (mining, visualization, machine learning).

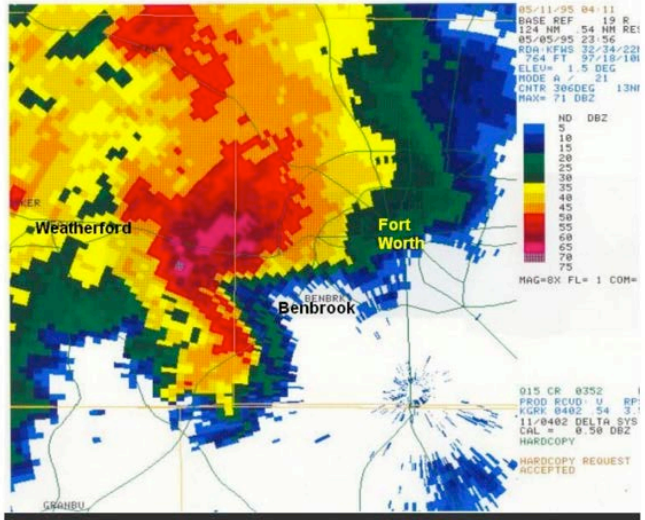
- A few examples of the "new" model:

E-SCIENCE ADVANCES RESEARCH PACE AND OUTCOMES

PREDICTING SEVERE HAIL STORMS

NSF-supported research at the University of Oklahoma uses supercomputers and simulations to improve storm forecasts

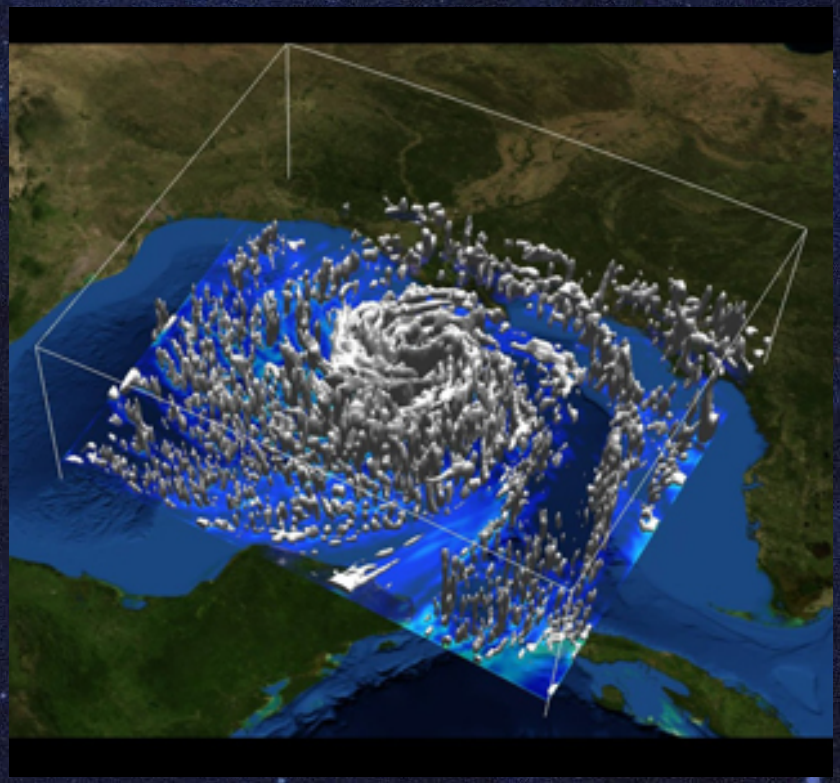
Published on March 22, 2016 by Aaron Dubrow



Radar imagery from 6:56pm shows a close-up of the Mayfest supercell centered west of Benbrook, Texas. The pink and darkest red colors represent radar indications of large hail with this storm. The storm impacted the Mayfest festival at 7:10pm. Credit: National Weather Service

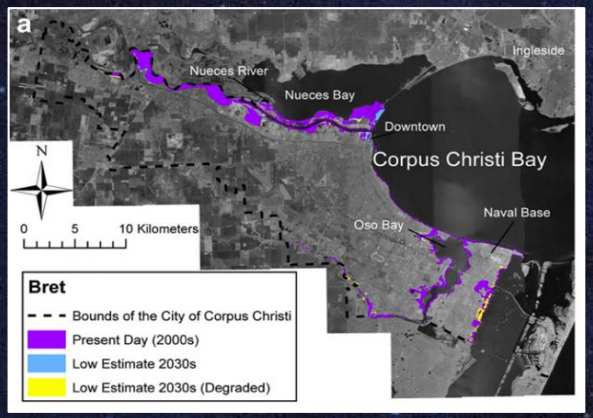
BLENDING ENSEMBLE SIMULATIONS, DATA ASSIMILATION+INTEGRATION, AND MACHINE LEARNING EXTENDS HAIL FORECASTS FROM 2 HOURS TO 24 HOURS

E-SCIENCE ADVANCES RESEARCH PACE AND OUTCOMES



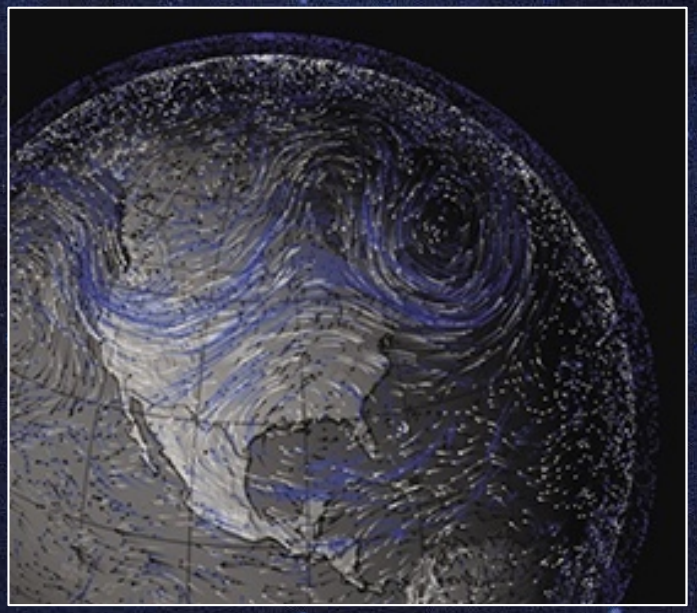
ENSEMBLE SIMULATIONS ON 40K CPU USING DATA STREAMED FROM NOAA IMPROVED PREDICTIONS OF LANDFALL LOCATION & INTENSITY FOR HURRICANE IKE

E-SCIENCE ADVANCES RESEARCH PACE AND OUTCOMES



SIMULATION COUPLED WITH CLIMATE PROJECTIONS HELPS CITY PLANNERS IN COASTAL CITY OF CORPUS CHRISTI, TX MITIGATE ECONOMIC, HEALTH, AND WELL-BEING IMPACT OF STORM SURGES

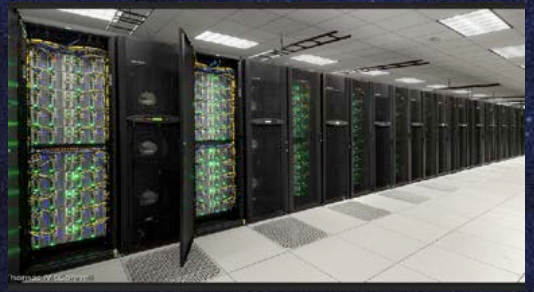
E-SCIENCE ADVANCES RESEARCH PACE AND OUTCOMES



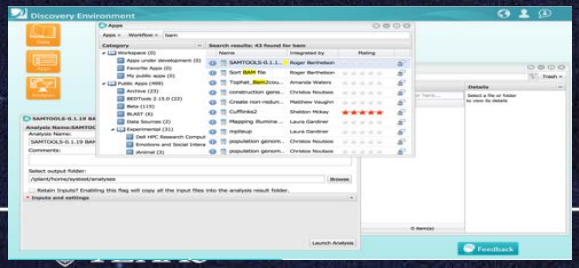
200+ TB DATA FROM NASA MERRA + WEB SERVICES + HADOOP + WEB WORKBENCH ENABLES RAPID ANALYSIS OF HISTORICAL, MULTISCALE SATELLITE DATA BY RESEARCHERS WORLDWIDE

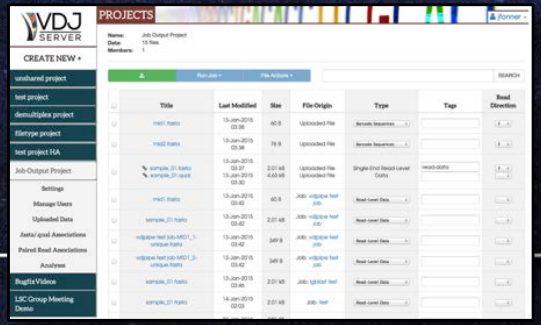
THE EVOLUTION OF A CYBERINFRASTRUCTURE

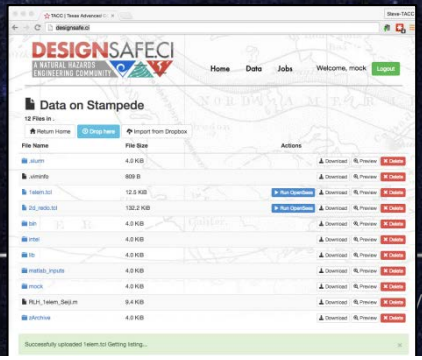
- Ten years ago, cyberinfrastructure was largely about building the hardware and networks to support large scale science.



- Today, it's about **new interfaces** to support **data analysis**, **collaboration** and **sharing**,**reproducibility** as well as easy access to simulation





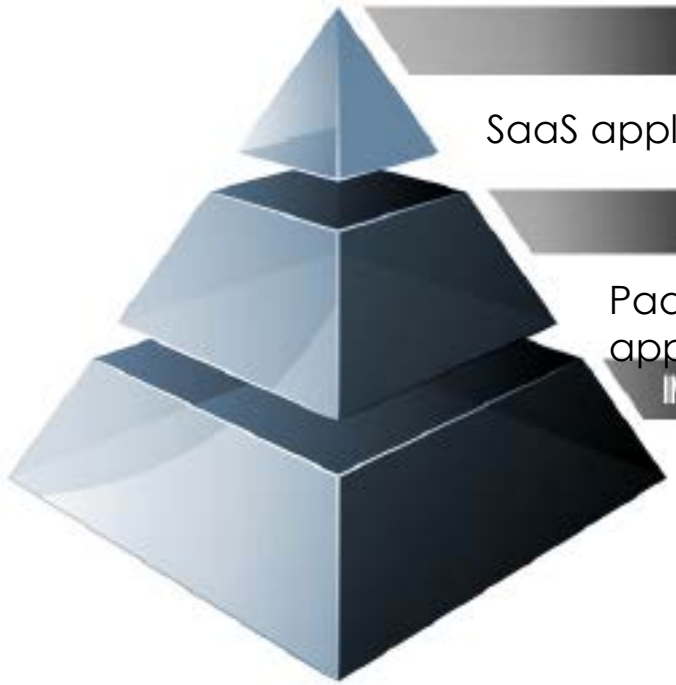


FOR THEIR RESEARCH, USERS NEED TO CONFIGURE…

- Interactive graphical research environments
- Orchestrated, heterogeneous science workflows
- Their own web applications, DBMS, AMQP services, etc.
- Specific operating system profiles
- Network architectures
- Newly emergent hardware

TACC aims to deliver the same quality of service for these use cases as for mainstream HPC

THE CLOUD COMPUTING STACK

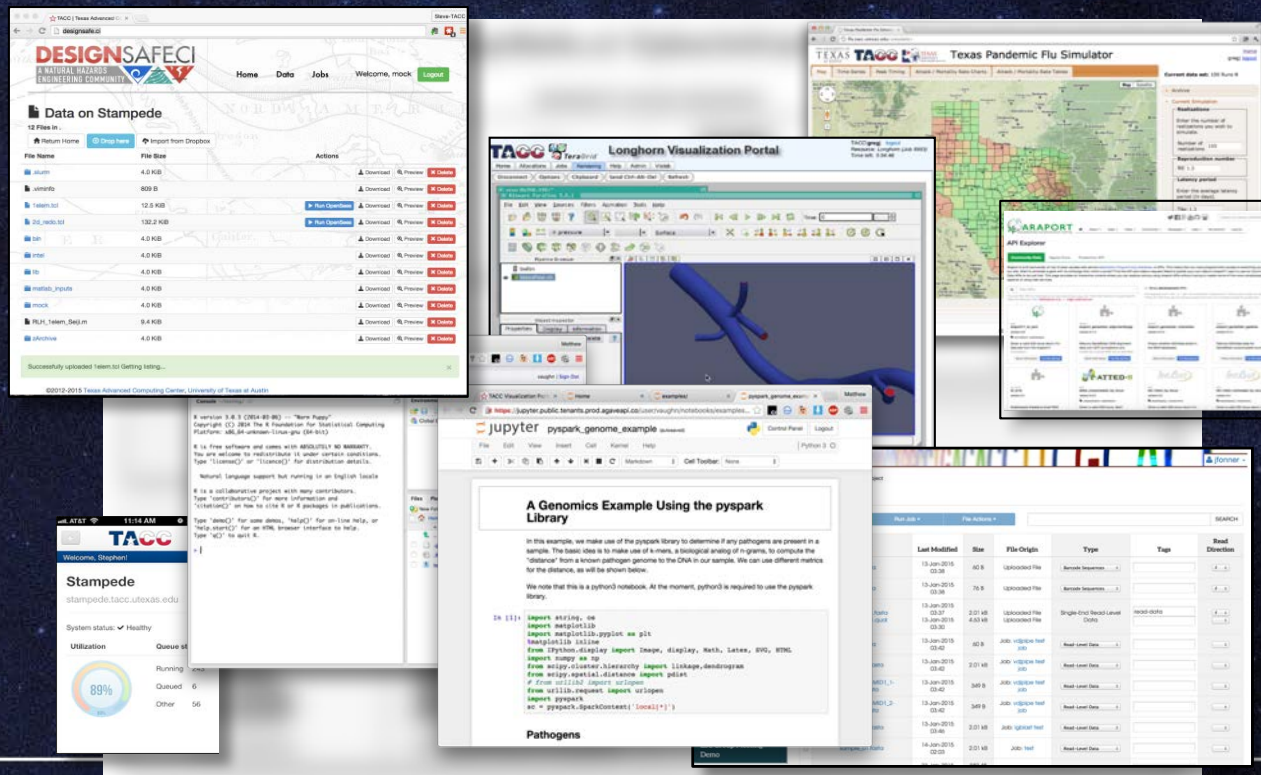


SaaS applications: Designed for end-users, delivered over the web

PaaS: Tools and services to help code and deploy web applications

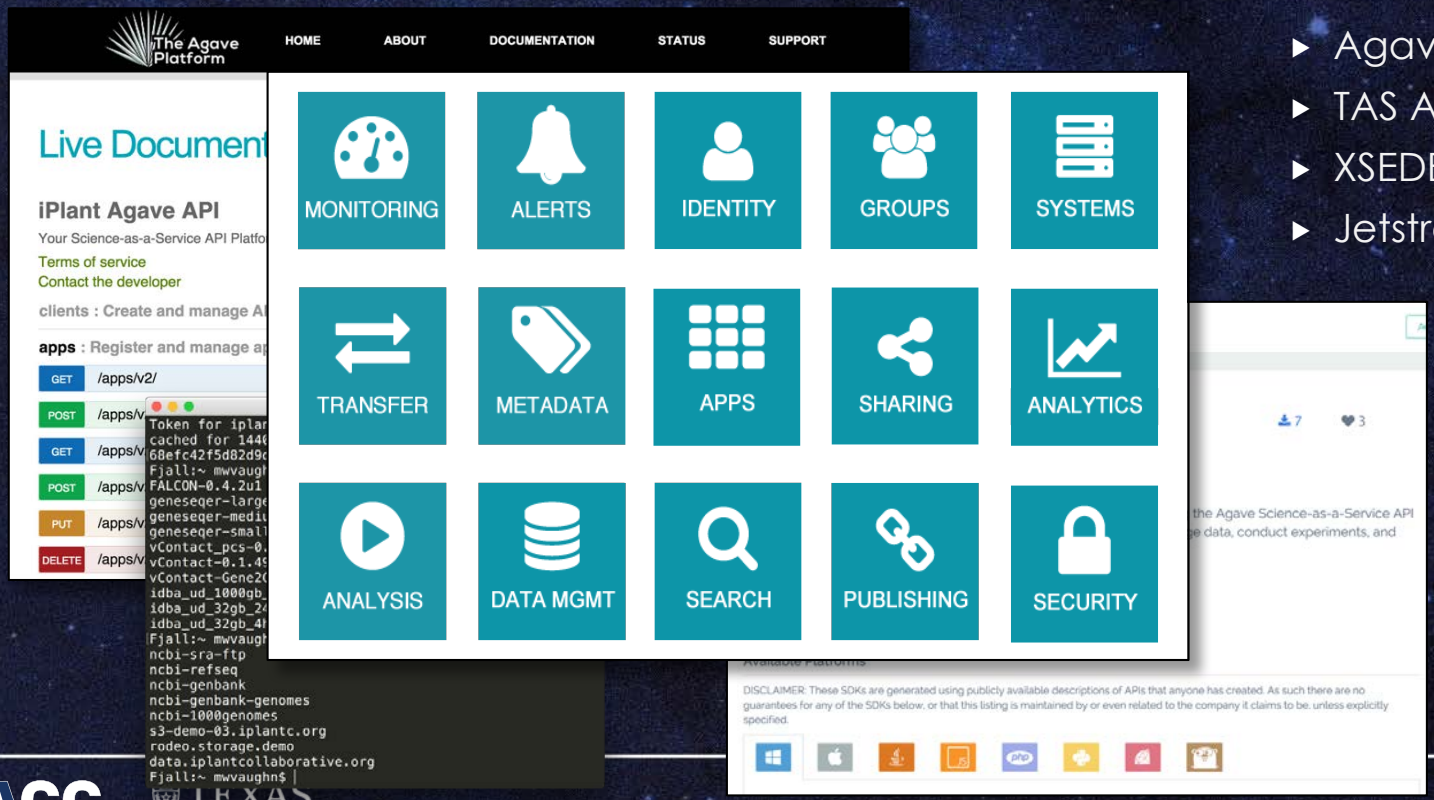
IaaS: Consumer-provisioned hardware and software to power it all

SAAS EXAMPLES



- DesignSafe Discovery Workspace
- Araport API Explorer
- VDJserver
- JupyterHub
- Rstudio
- TACC Vis Portal

PAAS: WEB SERVICE APIS EVERYWHERE



- Agave API
- TAS API
- XSEDE API
- Jetstream API

TACC

TEXAS
The University of Texas at Austin

http://jetstream-cloud.org/

IAAS: FLEXIBLE HARDWARE PROVISIONING

Roundup
VMware Internal Cloud (TACC production)

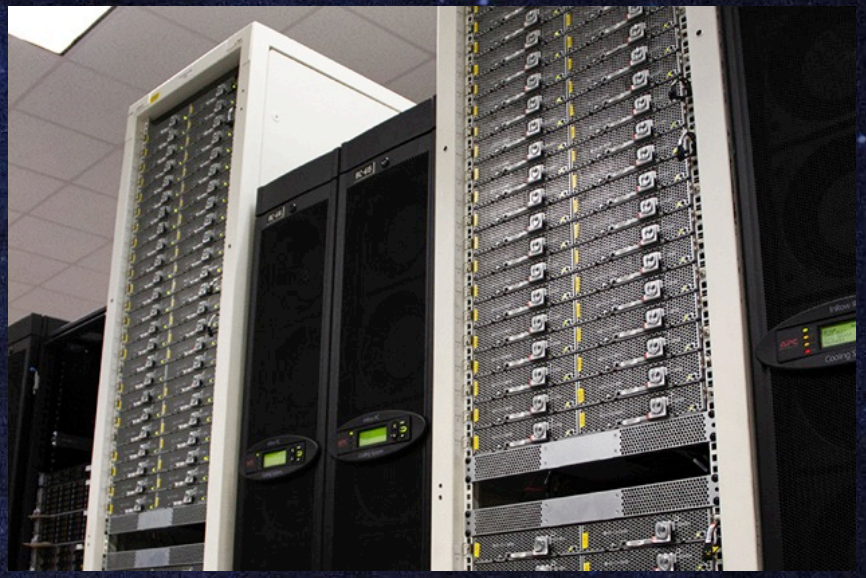
Chameleon
OpenStack Public Cloud
(CISE research)

Rodeo
OpenStack Internal Cloud
(TACC development)

Jetstream
OpenStack Public Cloud
(XSEDE research)

CATAPULT – AN EXPERIMENTAL PLATFORM FOR MACHINE LEARNING.

- A partnership with Microsoft and Altera, Catapult is an architecture to put Reconfigurable Hardware in the Datacenter.
- Implemented in the Bing search engine
- 432 node system at TACC is the first publicly accessible instance.
- Available for community use, focus on machine learning apps.



SO, IT'S NOT ABOUT HPC OR THE CLOUD

- Turns out big datacenters can be used in more than one way.
- It also turns out, to do science, we need Cloud *and* HPC.
  - We use HPC techniques when we need performance, and cloud techniques when we need APIs, failover, etc.
- In the end, to do science – we need both.

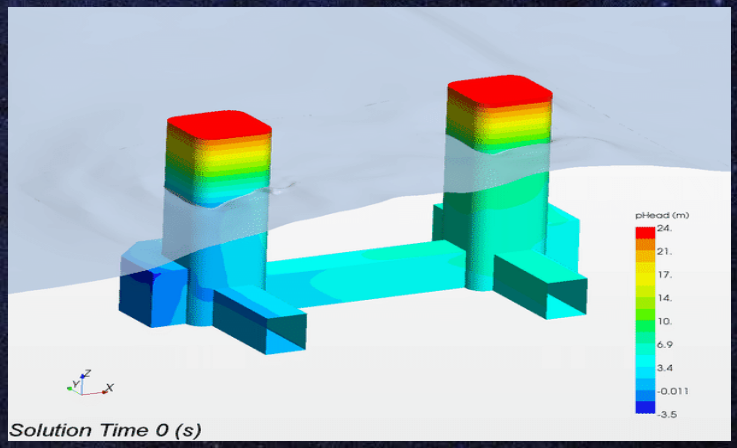
THERE ARE MORE SYNERGIES

- Recently, we have been able to scale "deep learning" (or as I like it to call it, HPC where the answers don't have to be right), to over 1,000 coupled nodes.
- Deep learning across nodes, it turns out, is doing a bunch of matrix operations with vector units, and synchronizing across them before advancing to the next layer.
  - This is remarkably like scaling up partial differential equation solvers.
- It's likely the performance features *HPC* needs will be the *future* performance features clouds need.

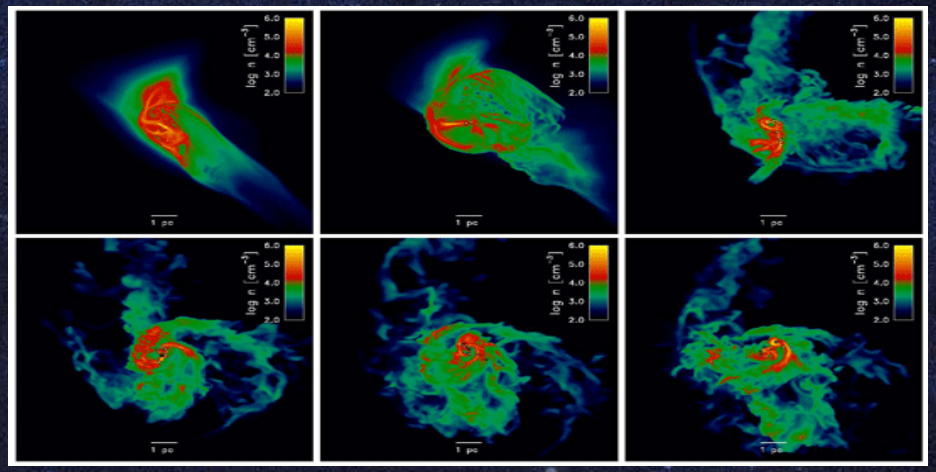
OUR SYSTEMS AND SOFTWARE ENABLE DESIGN AND DISCOVERY

From Rocket Engines designed by Firefly…

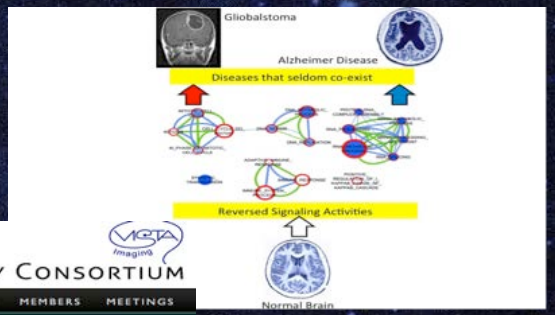
To offshore oil platforms with Technip…



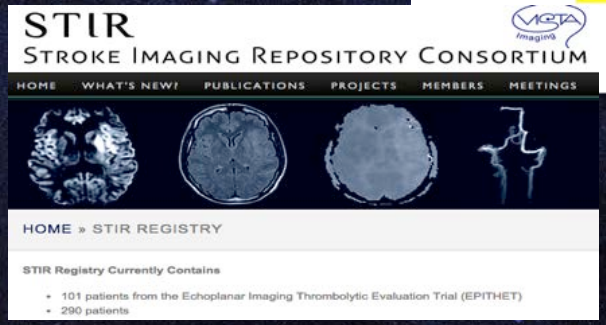
OUR SYSTEMS AND SOFTWARE ENABLE DISCOVERY



From the big mysteries of the universe…



To the tiny mysteries inside our cells…




TACC
TEXAS ADVANCED COMPUTING CENTER
TACC'S SYSTEMS AND EXPERTISE CHANGE HOW DISCOVERY IS DONE, AND CHANGE THE WORLD