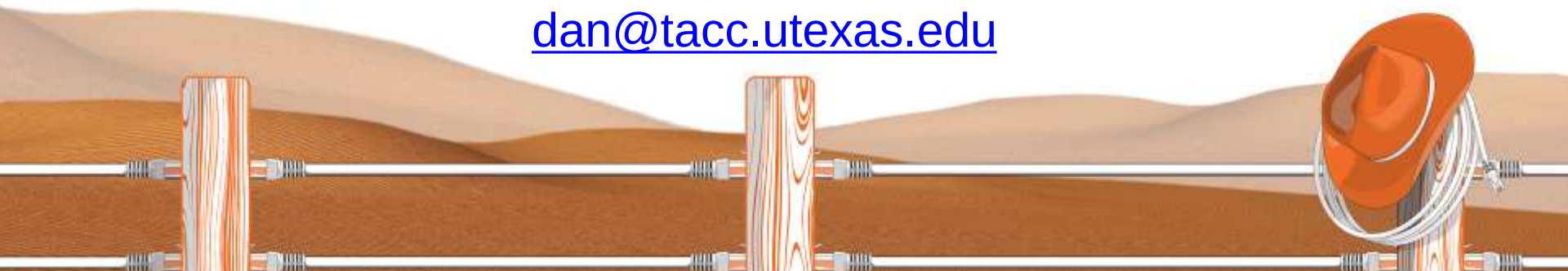
Wrangler : A New Generation of Data Intensive Cluster Computing

IDC HPC User Forum Meeting

Dan Stanzione, PI
Texas Advanced Computing Center

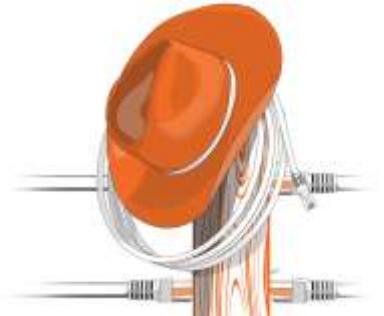
April 15th, 2015
Norfolk, Virginia

dan@tacc.utexas.edu



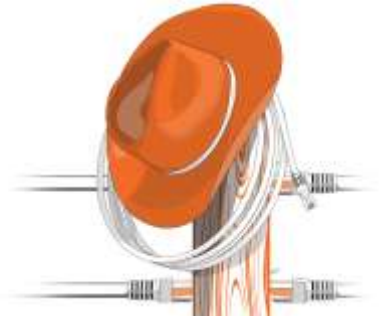
Acknowledgments

- The Wrangler project is supported by the Division of Advanced Cyberinfrastructure at the National Science Foundation.
  - Award #ACI-1447307 *"Wrangler: A Transformational Data Intensive Resource for the Open Science Community "*



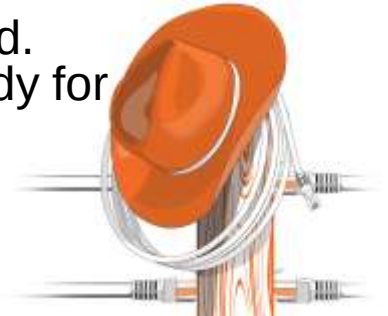
What is Wrangler

- Wrangler is a new data-intensive supercomputing system
- Built from the ground up for “data intensive” applications.
- HPC and “Big Data” have a lot in common
  - But the overlap isn’t 100% in all applications.
  - While Exascale computers will generate phenomenal amounts of data, not *every* data problem will map perfectly.
- New technologies can deal with the shortcomings in HPC Cluster architectures



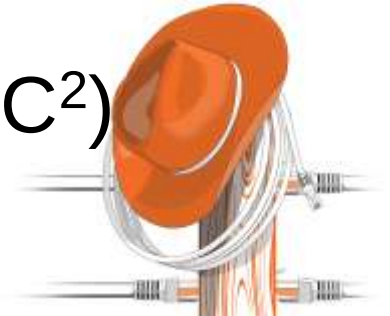
What Wrangler Enables

- Stampede is fantastic for tens of thousands of people
  - But has some limitations (metadata performance, I/O per node).
  - And makes assumptions about software (Distributed memory, MPI I/O for HPC, etc.).
- While *theoretically* we could fix all the software and workflows in the world to run well in this environment. . .
  - Practically, we will never have the time or resources.
  - Wrangler will simply lift up some of this “bad code”. Done your computation inside a SQL DB? No problem.
  - And the code that does get optimized will be able to do things we couldn’t imagine on Stampede. – Wrangler is *not* just the “bad code” system.
- On Wrangler, data isn’t just scratch; it’s a first class object to be protected, curated, kept on the system, and shared with the world. We think an enormous fraction of the scientific community is ready for this paradigm.

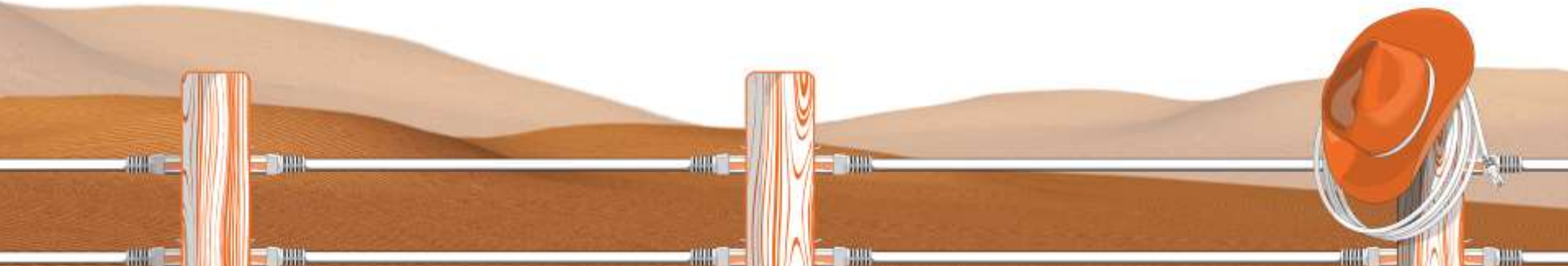


Project Partners

- Academic partners:
  - TACC – Primary system design, deployment, and operations
  - Indiana U. ; Hosting/Operating replicated system and end-to-end network tuning.
  - U. of Chicago: Globus Online integration, high speed data transfer from user and XSEDE sites.
- Vendors: Dell, DSSD (subsidiary of EMC^2)



[How we got here]



Once upon a time, most of us built garage-style clusters

- We'd cobble together some machines and call it a cluster.
- The computer scientist or engineer would build the cluster (often including the furniture), add user accounts, and in theory the users should take it from there.
- We've gotten a little more sophisticated since then…

Grendel, the second cluster in the Beowulf project, deployed in 1993, at the PARL Lab at Clemson

TACC

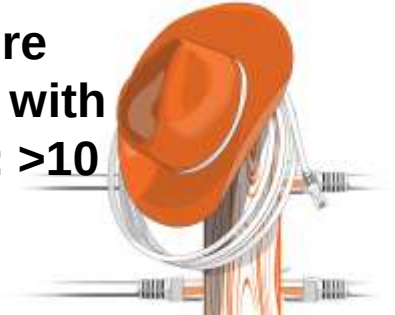
The Texas Advanced Computing Center: A World Leader in High Performance Computing

***1,000,000x** performance increase in UT computing capability in 10 years.*

Ranger: 62,976 Processor Cores, 123TB RAM, 579 TeraFlops, Fastest Open Science Machine in the World, 2008

Lonestar: 23,000 processors, 44TB RAM, Shared Mem and GPU subsystems, #25 in the world 2011

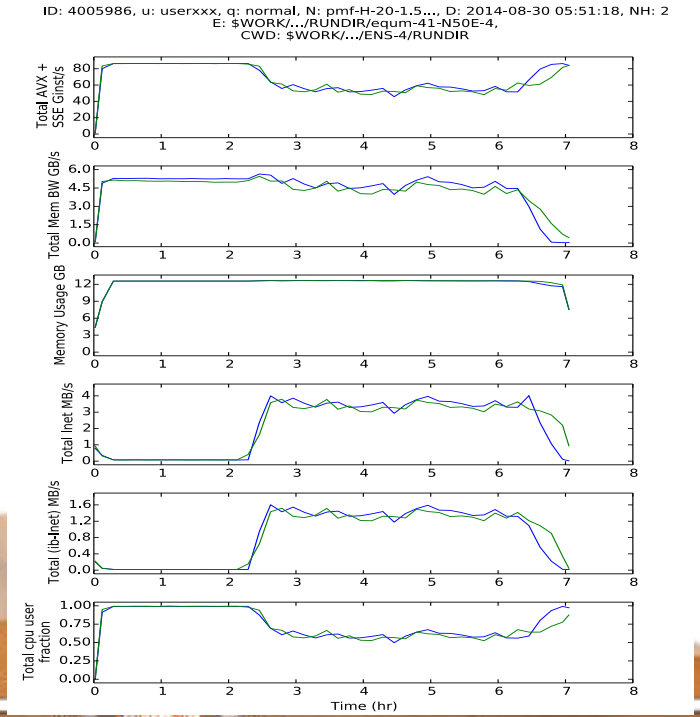
Stampede: #7 in the world, Somewhere around half a million processor cores with Intel Sandy Bridge and Intel MIC, Dell: >10 *Petaflops*.



We've gotten pretty good at making big, useful clusters.

- Through 28 months of Production Operation, Stampede by all measures is remarkably successful.
  - Over *1.5 Billion* Service Units delivered.
  - Over 5 **million** successful jobs
  - 2,283 distinct projects received allocations
  - 6,926 Individuals have actually run a job (~10,000 accounts).
  - 98% cumulative uptime (target: 96).
  - 5,006 User Tickets Resolved
  - 2,000+ users attended training last year.
- Formal requests from the community from XSEDE run ~500% available hours

Even in the workload we *already* have. we'd see jobs like this:

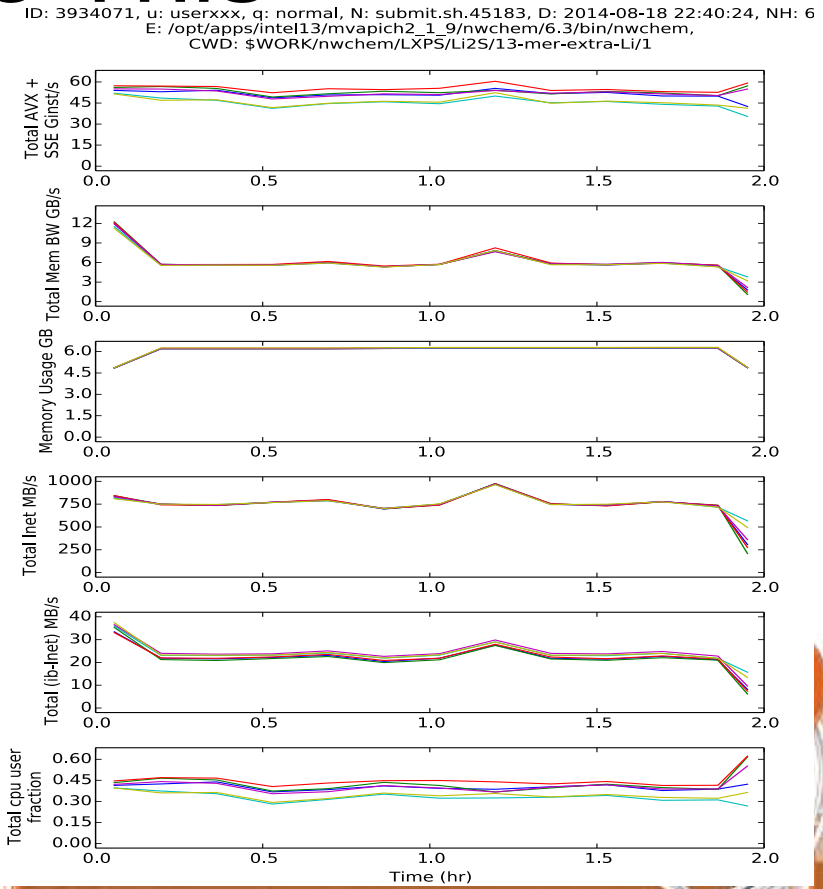


In this biophysics jobs, high metadata traffic starts to limit performance about 2 hours in, and in fact makes the job run 3 hours longer.

Or Like This

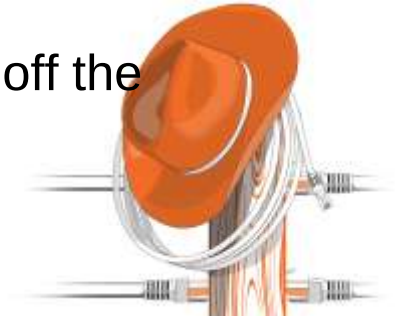
In this chemistry job, high metadata traffic throughout (a proxy for many small I/O operations) keeps performance at 1/3rd of peak throughout the 2 hour run.

How fast is this computer" is a multi-dimensional quantity, and in a world of small, random access, it means different things.

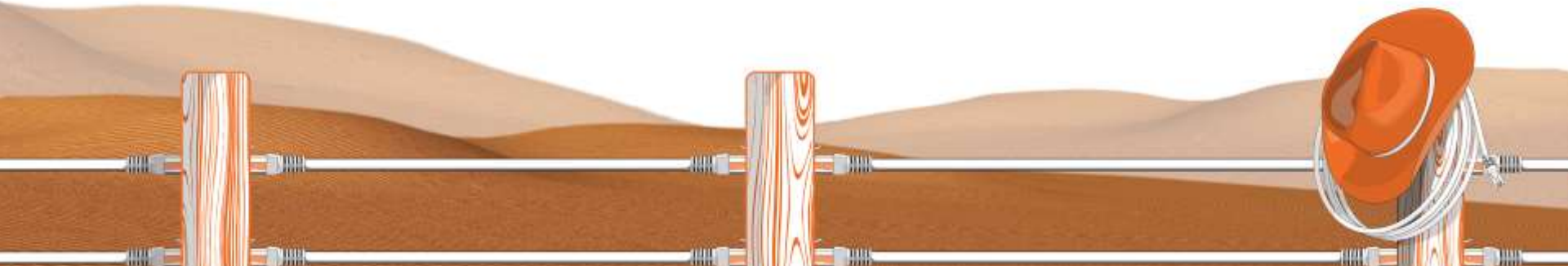


Hadoop

- In ~2011, we discovered an exciting new failure mode in large scale systems.
- We called this failure mode “Hadoop”.
- Recipe for success: take a big system
  - A huge central filesystem
  - Optimized for large, sequential, access
  - With a highly tuned, low level C interface
- And on that run software that:
  - Assumes a small, massively distributed filesystem.
  - Optimized for very small files.
  - With an untuned, well… Java.
- We deployed Rustler at our own expense to keep these users off the supercomputers.

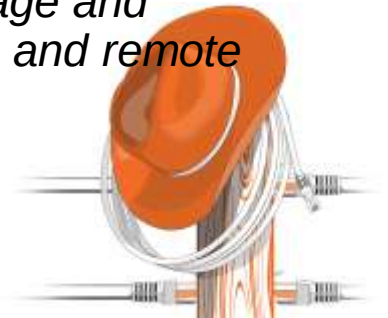


To address these problems, we proposed Wrangler



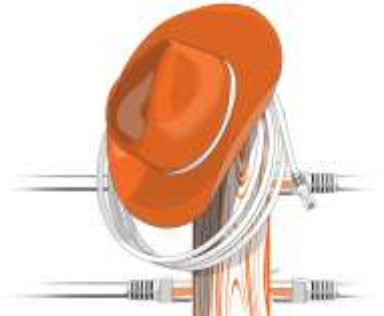
Goals of the Wrangler Project

- Respond directly to recent NSF and PCAST reports, which recommend that:
  - *High end ... data ... and sensor-based systems and the associated user support are needed for transformative science; ... networking, interoperable data systems and mining, including a focus on sustainability and extensibility.*
  - *[The NSF should] Serve scientific communities' data service requirements ... across a range of data types ... Such a service should NOT exclusively focus on large-scale or what could be referred to as "petabyte data" but rather include mid/small-sized research [and] working with the research community to positively and actively promote open access ...*
  - *The NSF should fund national facilities for at least short-term storage and management of data to support collaboration, scientific workflows, and remote visualization*

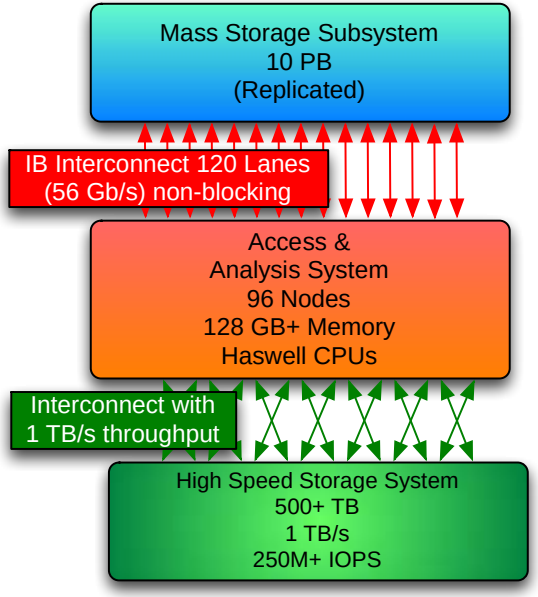


Goals of the Wrangler Project

- Our analysis of community needs indicated we needed:
  - To address the data problem in multiple dimensions
    - Big (and small), reliable, secure
    - Lots of data types: Structured and unstructured
    - Fast, but not just for large files and sequential access. Need high transaction rates and random access too.
  - To support a wide range of applications and interfaces
    - Hadoop, but not *just* Hadoop.
    - Traditional languages, but also R, GIS, DB, and other, perhaps less scalable things.
  - To support the full data lifecycle
    - More than scratch
    - Metadata and collection management support
- Wrangler is designed with these goals in mind.

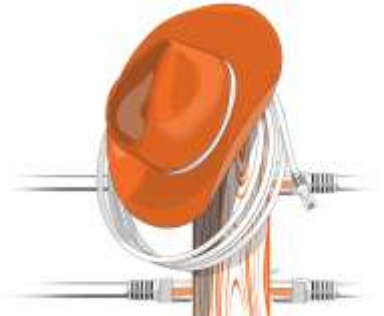


Wrangler Hardware

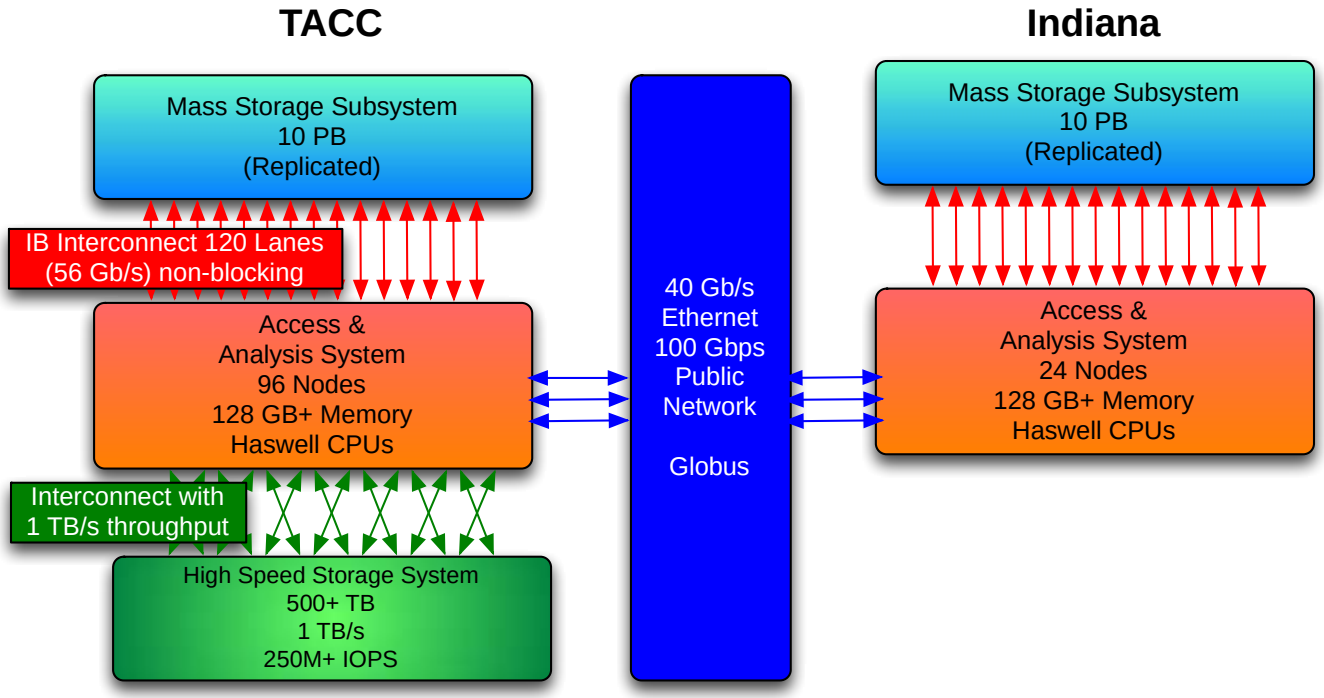


Three primary subsystems:

- A 10PB, replicated disk storage system.
- An embedded analytics capability of several thousand cores.
- A high speed global object store
  - 1TB/s
  - 250M+ IOPS

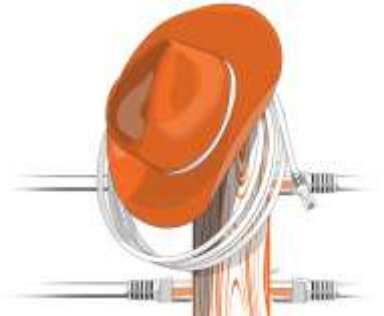


Wrangler At Large



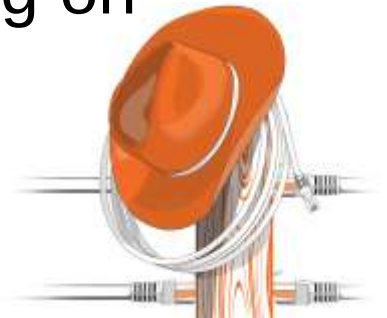
Storage

- The disk storage system will consist of more than 20PB of raw disk for “project-term” storage.
  - Geographically replicated between TACC and Indiana (more reliable than traditional scratch).
  - Ingest at either site.
  - Exposed to users on the system as a traditional filesystem



Analysis Hardware

- The high speed storage will be directly connected to 96 nodes for embedded processing.
  - Each node will have 24 Intel Haswell cores, and at least 128GB of RAM.
  - An additional 24 nodes will be at the replica site for data ingest, mirroring, processing on the bulk storage.

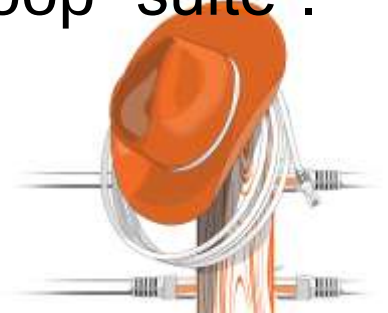


DSSD Storage

- The flash storage provides the truly “innovative capability” of Wrangler
- Not SSD ; a custom interface allows access to the NAND flash technology performance without the overhead of the traditional “disk” interface.
- Opportunity to explore APIs that integrate natively with apps (i.e. HDFS direct integration)
- Half a petabyte of usable space
  - Nearly 100k NAND flash dies
  - Initially 480, then 960 Gen3 x4 PCI links to the storage system.

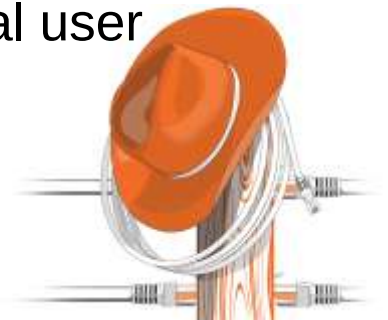
Wrangler Software and Use Cases

- The high speed storage will be visible in several ways:
  - As a traditional filesystem, for traditional applications
  - As an HDFS filesystem, for Hadoop and other Map Reduce applications.
  - As a SQL database
  - As an object store with a native API, for novel data applications
- In addition to our “traditional” HPC stack, we support R, databases, NoSQL databases, and the full Hadoop “suite”.

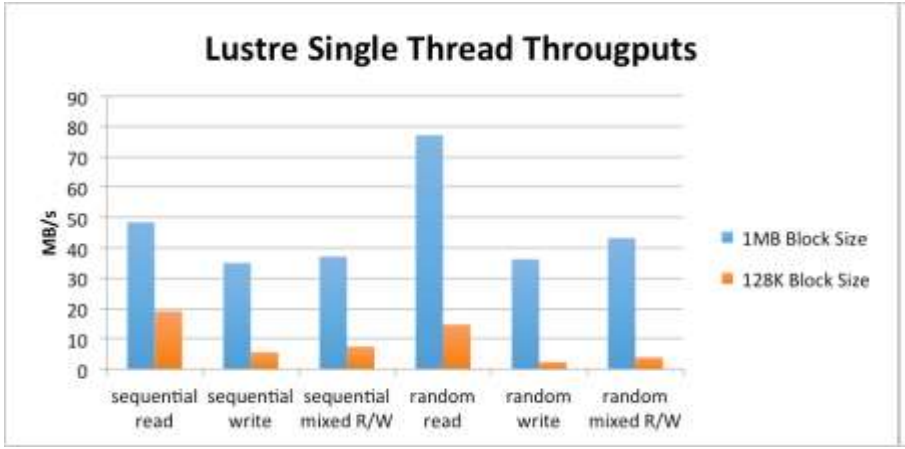


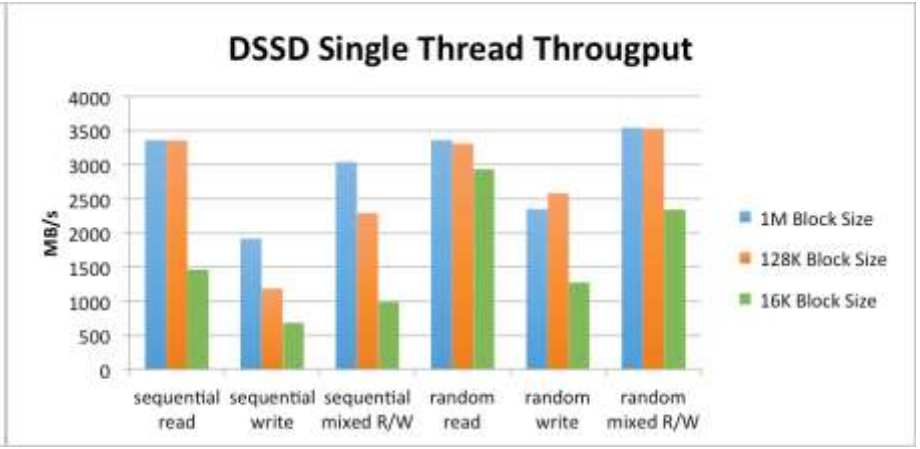
Status Today

- We are in early operations mode, with a number of projects on the system, and more being added each week (more on this in a moment).
- We have chosen to partition the Flash storage (to increase the number of compute nodes, and create some semi-persistent spaces).
  - Some nodes will still see all 10 DSSD devices.
  - Two pools of 48 nodes with >200TB available each
  - 16 additional nodes for persistent services.
  - Still experimenting with optimal configuration for actual user workloads

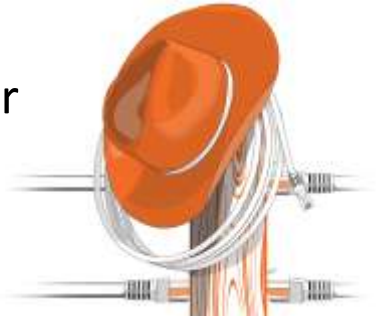


Where DSSD Really Shines



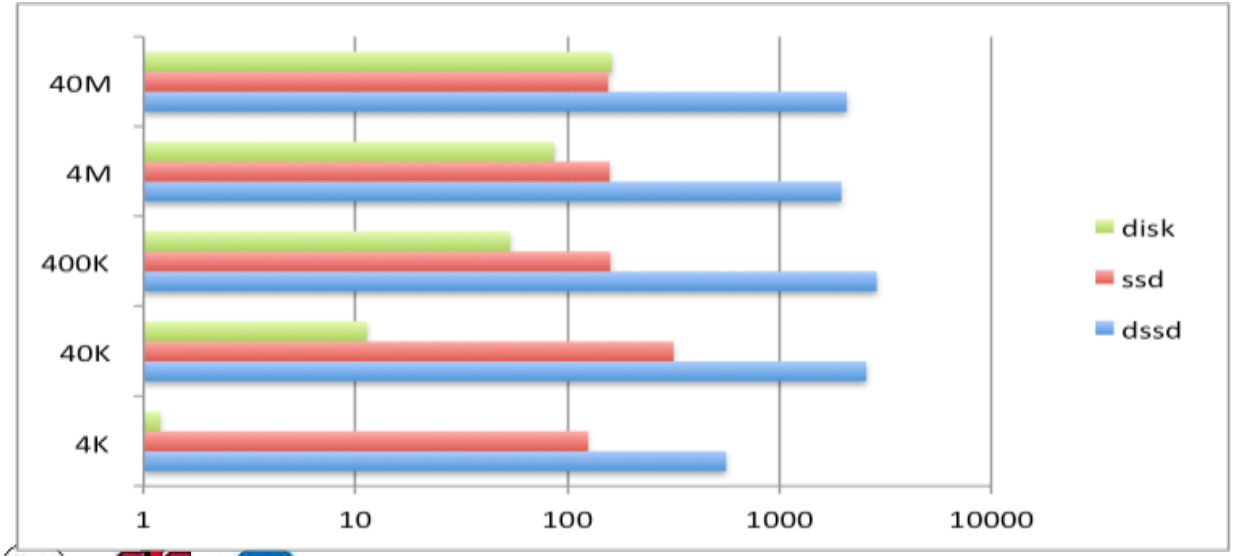


- Single thread IO for different block sizes
  - Flash is faster than single spinning disk (no surprise)
  - DSSD sustains most throughput for small block sizes and for sequential and random I/O paterns

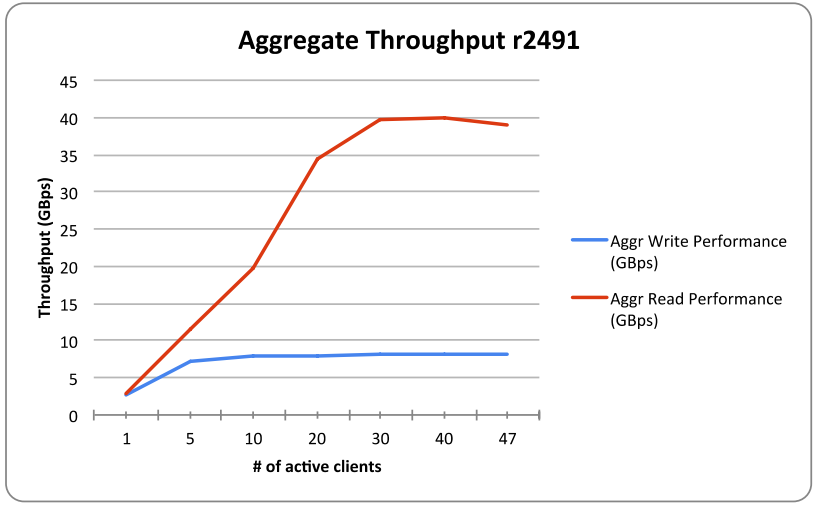


Really early data

- At big write sizes, we're 10x better than "conventional" disk.
- At small write sizes, we're ~400x better than disk .



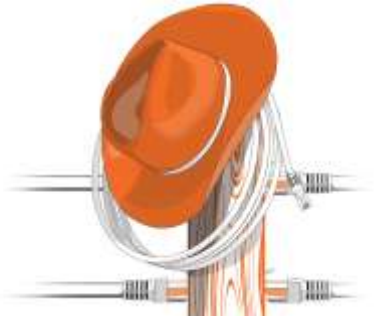
Object Store Rates



- Uses DSSD flood interface for direct access to Flash (no file systems)
- Current Object store can sustain 46+ GB/s read throughput

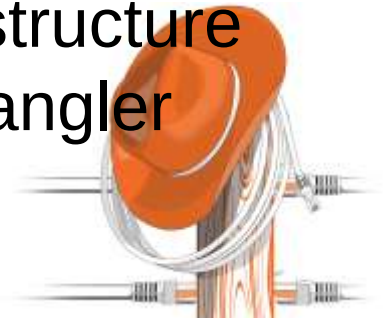
More Than the Parts

- We pride ourselves on our support of users…
  - but our HPC systems would fill if we weren't great at it.
  - A huge base of sophisticated, demanding users, some of whom have used HPC in their research for 30 years.
- With Wrangler, we are bringing in a myriad of new communities which will need much more support; and even the old ones will need to think differently.
- We've conceived the Wrangler *project* as more than just the system, but also a set of services.



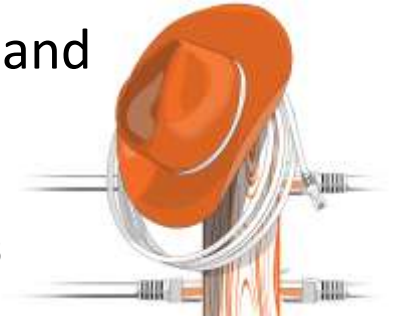
Collaborative Services

- Support for complex projects via XSEDE Extended Collaborative Data Support Services
  - Working directly with XSEDE staff
  - Focus on specific data management challenges
  - Work with teams to improve data workflows
  - Work to integrate tools into existing infrastructure
  - Transition of a projects processing to Wrangler



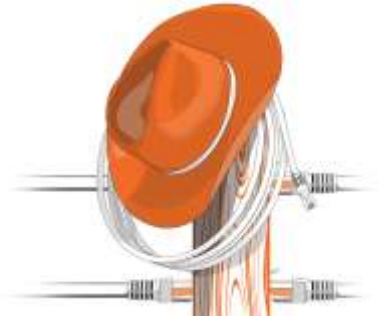
Managing Data

- Data Management & Autocuration Service
  - Leveraging Globus Online Dataset Services
  - Data Organization tying to research project to facilitate tracking of a projects data assets
  - Data Provenance leveraging emerging standards to ensure data accessibility and reusability
  - Data Fixity by automating extraction of technical metadata for files in the system
  - Data Usage by tracking overall size of a projects data and (internal and external) utilization information



A Flexible Comprehensive Software Environment

- My Hadoop style system
  - Working with Cloudera
  - More than just MapReduce (Spark? Mahout?)
- Apache Mesos
- RDB and noSQL databases with GIS integration
- System optimized R, Python, etc.

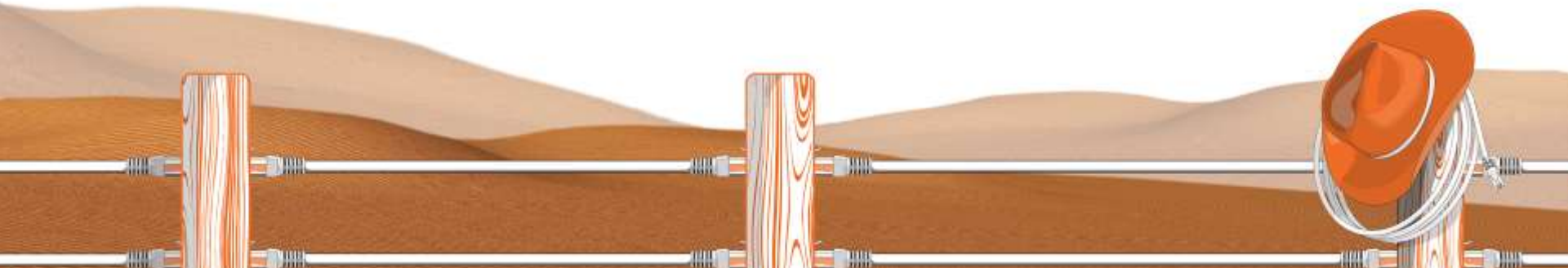


Moving Data Effectively

- End to End Network Performance Tuning
  - Leveraging I^2 and Globus based Perfsonar
    - XSP support to allocate bandwidth on demand
    - Software Defined Networking support via I2's AL2S
    - Network Performance Tools to monitor for bottlenecks and react to congestion
- Data Dock capability for the "last mile problem" where limited by network capabilities outside of I^2.

3,000 Rice genomes from the Philippines; yes, this is the actual box

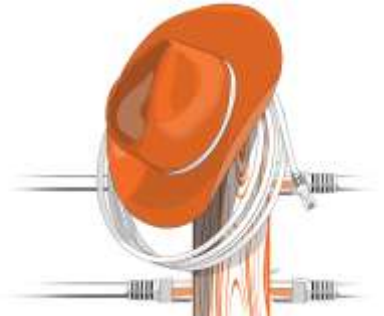
Early User Application Success Stories



OrthoMCL Science Case

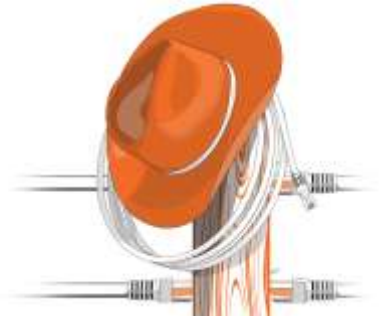
UT Austin Center for Computational Biology and Bioinformatics

- Dr Hans Hofmann, Dr. Rebecca Young
  - 6 and 8-species Gene expression comparison
  - Brain development/independent evolution of monogamous behavior
- Dr. Hans Hofmann, Dhivya Arasappan
  - *Rhazya stricta* gene family grouping
  - Medically important alkaloids



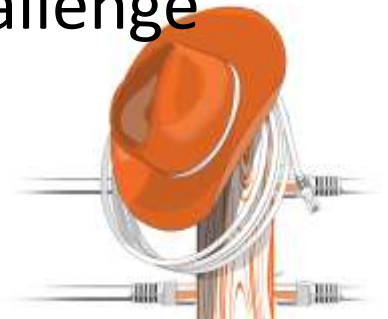
OrthoMCL Application

- "Orthologous Protein Grouping"
  - Multi-stage workflow
  - BLAST, protein grouping, results presentation
  - Protein grouping phase performed in-database
  - Both computational and I/O-Intensive
  - Order 10s of GBs databases typical



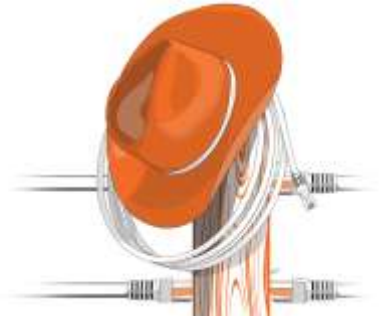
OrthoMCL Previous results

- Developers benchmark – 16-24 hours
- CCBB datasets on TACC Corral systems
  - Quad-socket servers, 64GB RAM, SAS-RAID6
  - Several steps took hours, some did not complete
  - Novel research data presented unique challenge



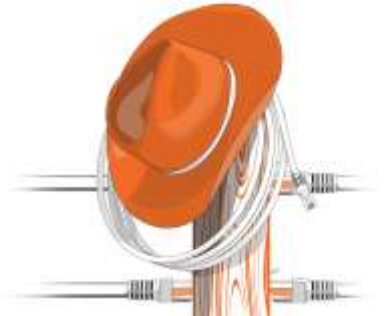
OrthoMCL Optimization Challenges

- Computational work performed in-database
- Ideal for ease-of-use but makes performance optimization difficult
- Data throughput/Random access are biggest factors in performance



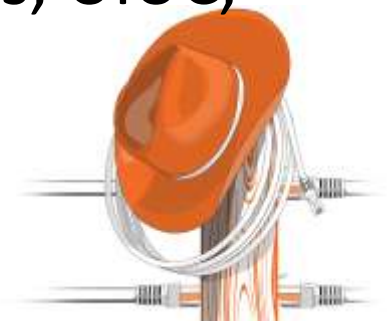
OrthoMCL/Wrangler Results

- Before Wrangler, TACC staff worked with researchers and OrthoMCL developers for > 1 month attempting to complete runs
- With Wrangler, multiple projects completed their research runs in less than a week
- All runs completed in 4-6 hours
- At least two publications in process



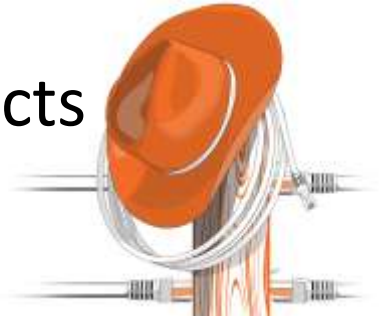
OrthoMCL Community

- Significant community of potential users
- At least one new project already allocated
- Sociogenomics RCN also planning to use OrthoMCL on Wrangler
  - Georgia Tech, Georgia State, Johns Hopkins, UIUC, Stanford, Harvard and UT Austin



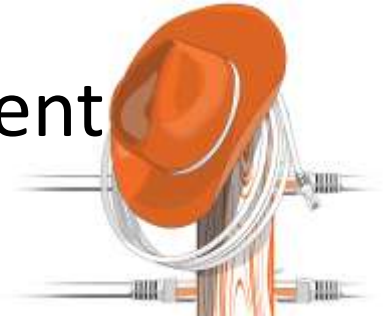
Genomics Projects and Applications

- iPlant and NCGAS projects
  - Key applications: BLAST and Trinity
  - DNA and RNA sequence processing
  - I/O Intensive due to large input datasets, need for global comparisons, and minimal reduction
  - Often the first step in longer life sciences workflows, used by many additional projects



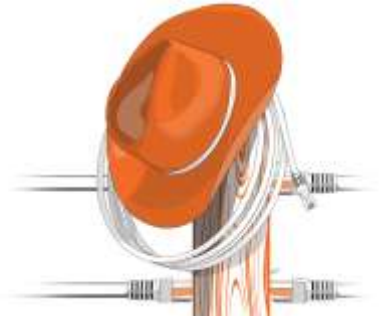
Improved Genomics Workflows

- BLAST and Trinity often run off of RAMdisk on current cluster systems
  - Wrangler eliminates this requirement
- Preliminary Trinity runs exhibit performance better than any existing TACC system
- Significant scope for further improvement



Early Use Cases in Progress

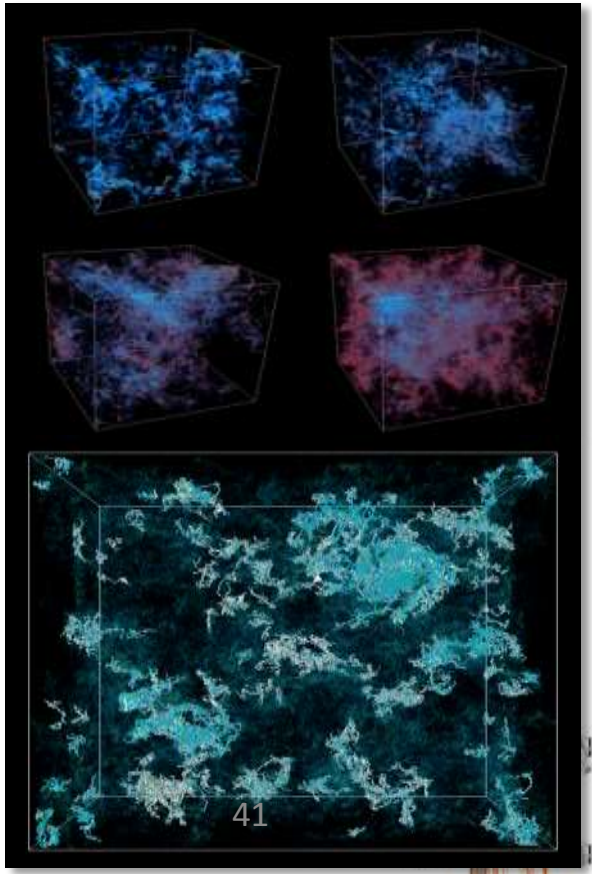
- Russ Poldrack, Stanford
  - fMRI Processing Pipelines
  - Freesurfer, R, SciPy, Hadoop/Spark
- Guatemalan National Police Historical Archive
  - 10 Million scanned document images
  - Large-scale image processing challenge



Early Use case: Analyzing Large Scale Turbulent Flow Data

P.K. Yeung, Georgia Tech; *Diego Donzis, Texas A&M; Kelly Gaither et al, TACC*

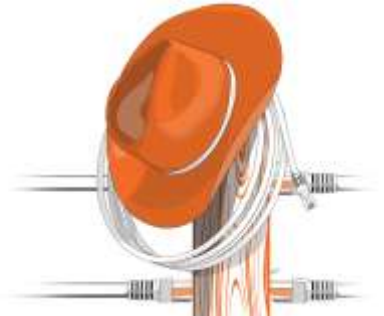
- Remote interactive visualization and data analysis of 17 time-steps (34 TB) of a turbulent flow simulation (4096^3)
- Equal parts data mining and remote interactive visualization – goal is to characterize flow behavior over time
- Data must be pre-processed and written to a Silo format to be useable for interactive manipulation
- With Longhorn, took ~3 hours to read the data set in for all 17 timesteps
- Theoretically possible to read data set in under 60 seconds on Wrangler



Gaither, K., Childs, H., Schulz, K., Harrison, C., Barth, W., Donzis, D., and Yeung, P.K., "Using Visualization and Data Analysis to Understand Critical Structures in Massive Time Varying Turbulent Flow Simulations," IEEE Computer Graphics and Applications, 32(4), Jul/Aug 2012.

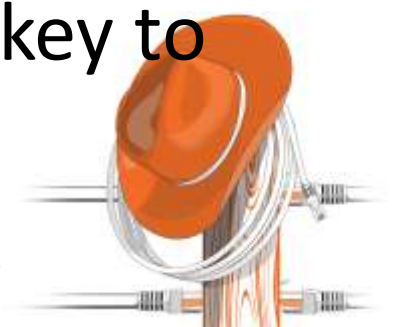
Early User Lessons Learned

- Process of allocating and configuring Flash storage is most complex task
- Reservation/Scheduling system crucial
- Use of global DSSD-based file system will make things even easier for users
- Many use cases will be generalizable



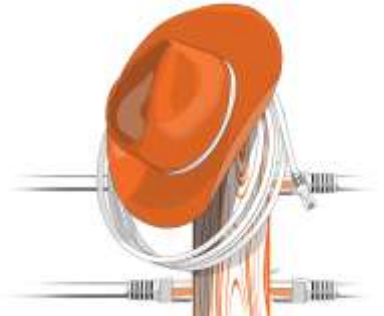
Early User Lessons Learned 2

- Wrangler is a unique resource
- Users will not always know if it is appropriate
- If it is appropriate, users will not necessarily know how best to make use of it
- Consulting and technical expertise are key to making effective use of Wrangler



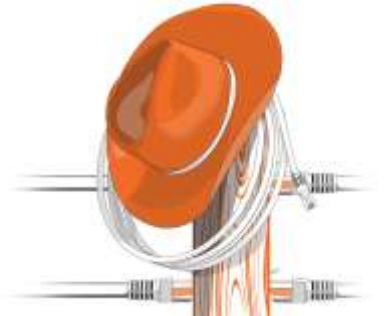
New User Communities

- A flood of new disciplines into computational sciences in last decade (led by life sciences).
  - Computational Economics requests as much time today as Computational Physics did in 2003!!!
- Most of these new areas have a few things in common
  - *Driven by data*, not equation based models
  - Mostly non-programmers
  - Less traditional languages, less performance tuning.



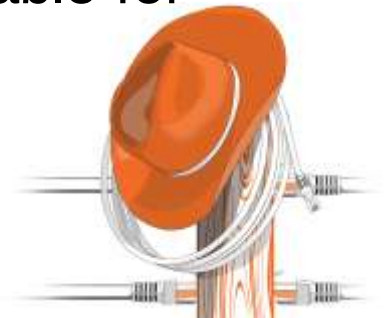
Wrangler in the TACC Ecosystem

- TACC is traditionally a provider of HPC, Visualization, and storage systems.
  - And we still are.
- But these new communities provide kinds of data-intensive problems our HPC systems just aren't built for
  - Run Hadoop on your favorite supercomputer to see what we need.
  - Or do a bunch of random access to a bunch of really small files.
- Wrangler is not to replace our supercomputer, vis, or cloud offerings; it supplements this environment.

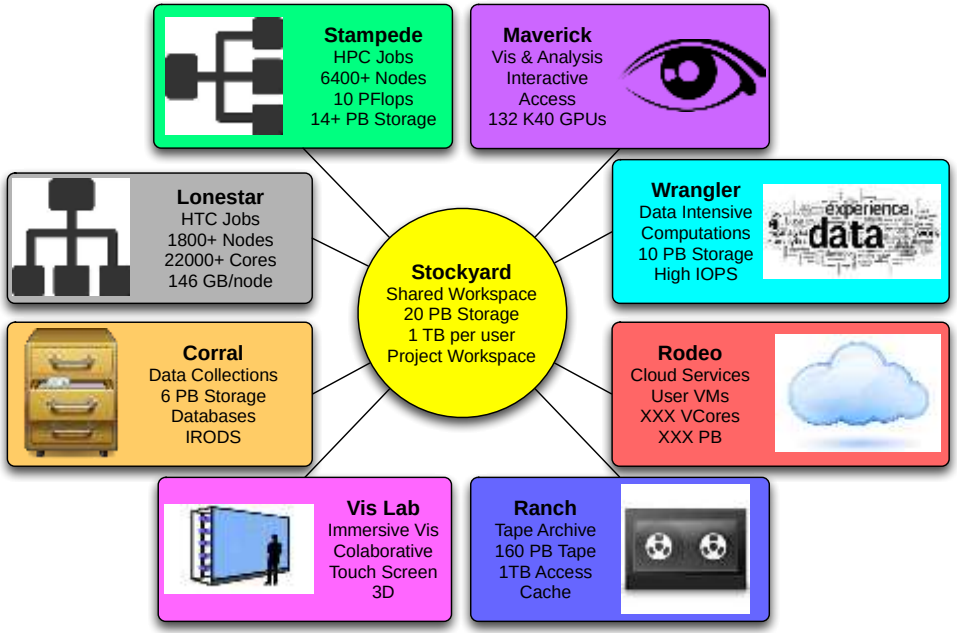


Wrangler in the TACC Ecosystem

- Wrangler will closely couple with related systems, most available to XSEDE users:
  - Stockyard, our global work filesystem, will allow easy data sharing between Stampede, Maverick, Rodeo, Lonestar, and soon Jetstream.
  - Rodeo will host VMs for persistent gateways, Maverick will enable large scale visualization, and Stampede might be the simulation data source for Wrangler users.
  - As always, our Ranch archive system will be available for long term data storage.



TACC Ecosystem



Thank You!

Dan Stanzione

dan@tacc.utexas.edu