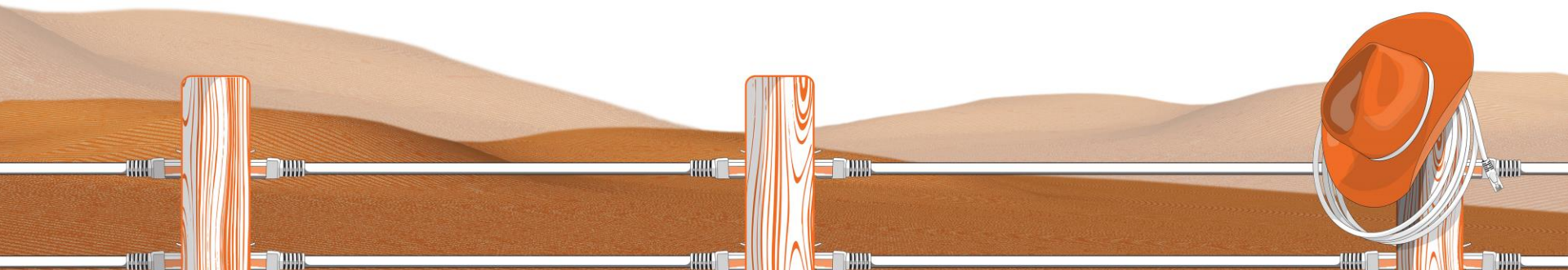
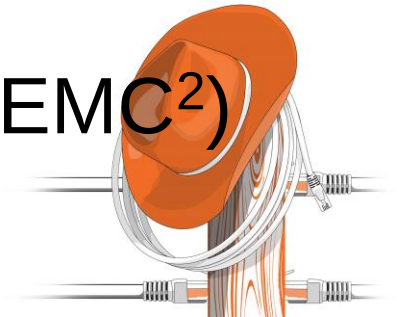
Wrangler: A New Generation of Data-intensive Supercomputing

Christopher Jordan, Siva Kulasekaran, Niall Gaffney

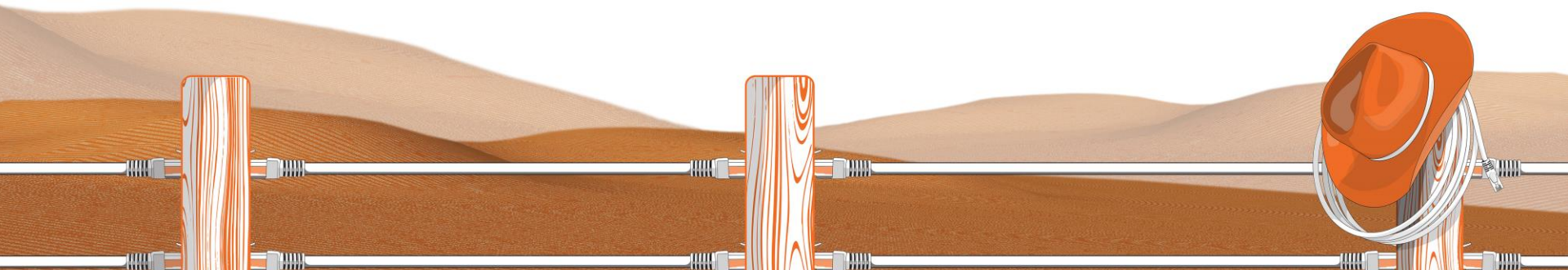


Project Partners

- 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)

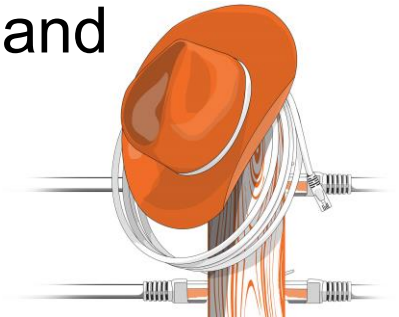


SYSTEM DESIGN AND OVERVIEW



Some Observations

- There are fundamental differences in data access patterns between Data Analytics and HPC
  - Small read access of many files vs. large sequential writes to several files
- Many Data Researchers want to work with Data not MPI, Lustre Striping, Vectorization, Code Optimization…
  - “What’s wrong with creating 4 Million 1K files and working with them at random?”

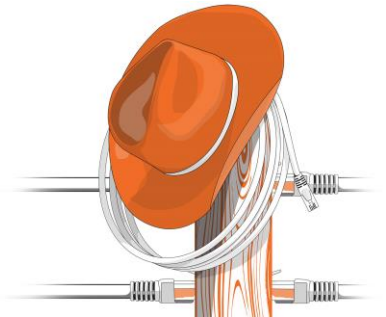


TACC

Goals for Wrangler

- To address the data problem in multiple dimensions
  - Data at large and small scales, 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, less HPC style performing workflows.
- To support the full data lifecycle
  - More than scratch
  - Metadata and collection management support

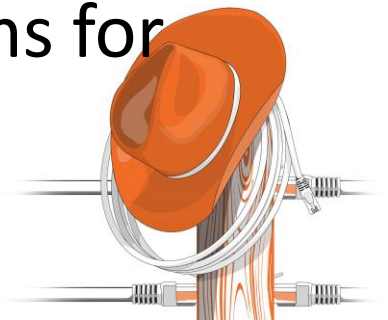
TACC



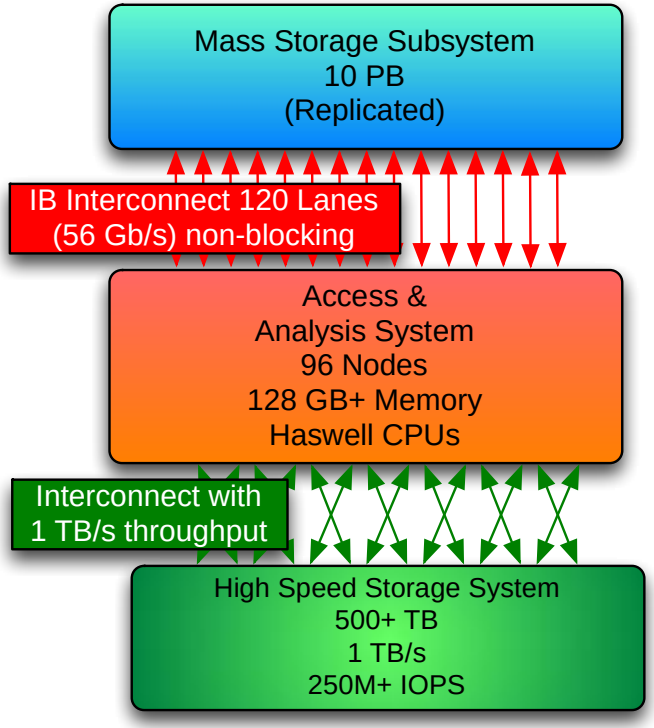
TACC Ecosystem

- Stampede – Top 10/Petaflop-class traditional cluster HPC system
- Stockyard and Corral – 25 Petabytes of combined disk storage for all data needs
- Ranch – 160 Petabytes of tape archive storage
- Maverick/Rustler/Rodeo – “Niche” systems for visualization, Hadoop, VMs, etc

TACC

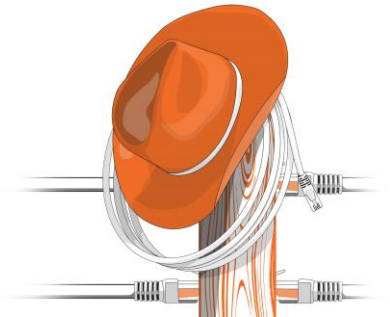


Wrangler Hardware

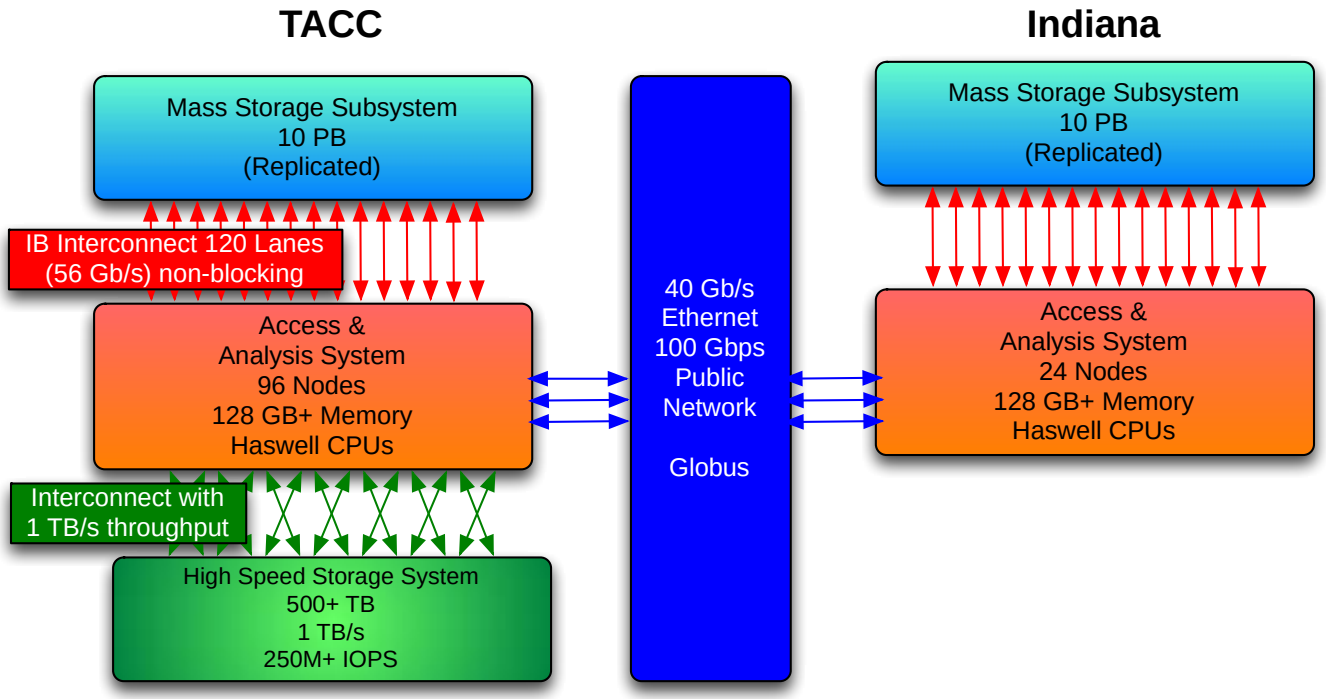


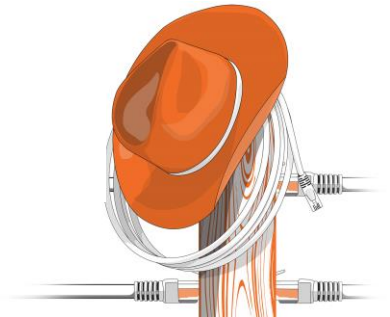
Three primary subsystems:

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



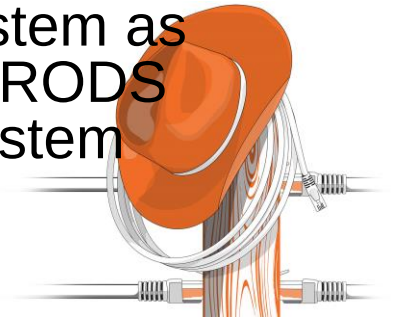
Wrangler At Large





Storage

- The disk storage system consists of more than 20 PB of raw disk for “project-term” storage.
  - ~75 GB/s sequential write performance
  - Lustre based file system with 34 OSS Nodes and 272 Storage Targets
  - Exposed to users on the system as a traditional filesystem and iRODS based data management system



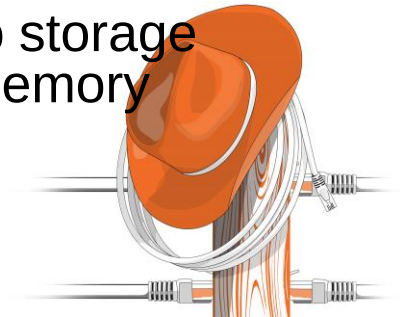
Analysis Hardware

- The high speed storage will be directly connected to 96 nodes for embedded processing.
  - Each analytics node will have 24 Intel Haswell cores, and 128GB of RAM, 40 GB Ethernet and Mellenox FDR networking.

TACC

DSSD Storage

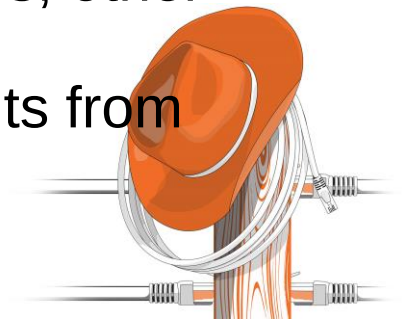
- The flash storage provides the truly “innovative capability” of Wrangler
- Not SSD; a direct attached PCI interface allows access to the NAND flash.
  - Not limited by 40 Gb/s Ethernet or 56 GB/s IB networking
- Flash storage not tied to individual nodes
  - Not PCI or SAS storage in a node
- More than half a petabyte of usable storage space once “RAIDed”
- Could handle continuous writes to storage for 5+ years without loss due to Memory Wear



TACC

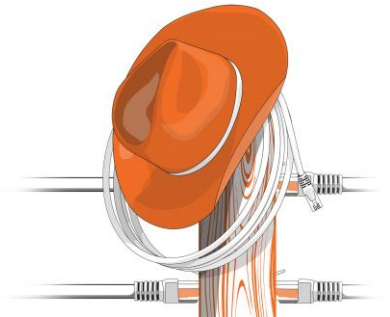
DSSD Storage (2)

- While the aggregate speed is great, the per node speed, and the speed for small, non-sequential transactions make this special for big data applications
  - We expect to get up to 12GB/s and 2 million+ IOPS to a single compute node
    - More than 100x a decent local hard drive
    - 5-6x a pretty good fileserver with a 48 drive RAID array.
    - I/O performance that used to require scaling to thousands of nodes will now require just a handful
- Great for traditional databases, some Hadoop apps, other transaction-intensive workloads
- Per-node performance is key – you can get benefits from Wrangler with e.g. Postgres on one node



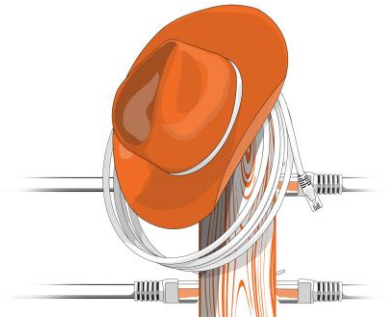
External Connectivity

- Wrangler will connect externally through both the existing public networks connections and the 100Gbps connections at both TACC and Indiana
- Fast network paths will be available to Stampede and other TACC systems for migration of large datasets
- Globus Online will be configured on Wrangler on day 1.



Wrangler File Systems

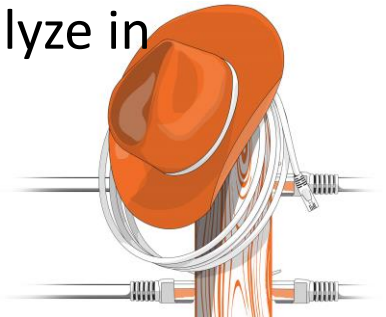
- /flash – GPFS-based parallel file system utilizing multiple DSSD units
- /data – Lustre-based 10PB disk file system
  - /data/published – Long term storage/publication area
- /work – TACC's 20PB global file system
- /corral-repl – TACC's 5PB Data management file system



Why Use Wrangler

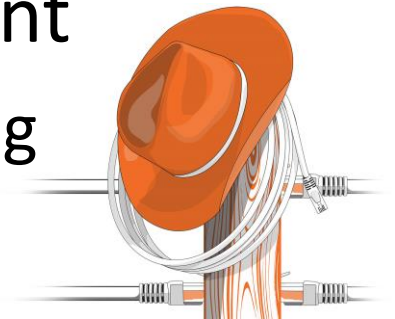
- Limited by storage system capabilities
  - Lots of small files to process/analyze
  - Lots of random I/O not sequential read/write
- Need to analyze large datasets produced on other systems quickly
- Need a more on-demand interactive analysis environment
- Need to work with databases at high transaction rates
- Have a Hadoop or Spark workflow with need for large high-performance backing HDFS datastore
- Have a dataset that many users will compute with or analyze in need of a system with data management capabilities
- Have a job that is currently IO bound

TACC

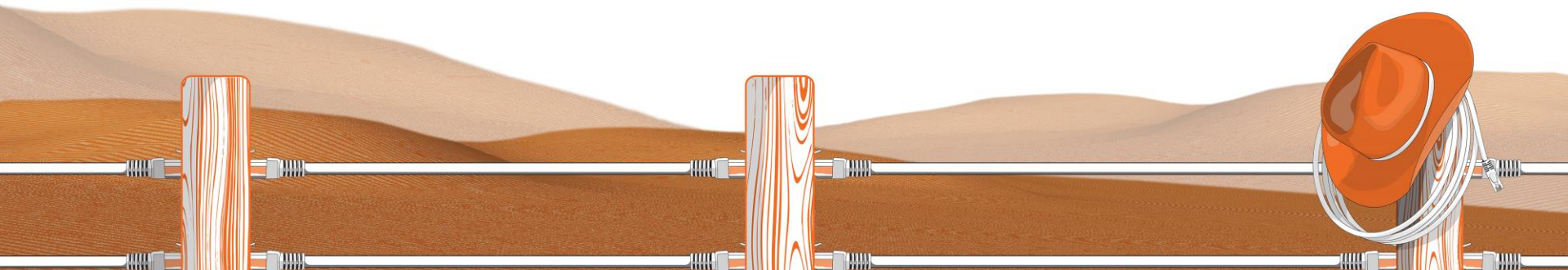


Why Not Wrangler

- Need a very specific compute environment
  - Cloud systems such as the upcoming Jetstream
- Need lots of compute capacity (>2K cores)
  - HPC systems like Stampede or Comet
- Need large shared memory environment
  - Stampede 1 GB nodes or Bridges is coming

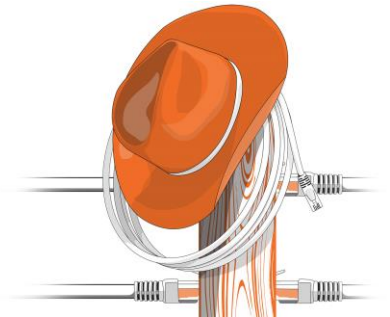


WRANGLER PORTAL AND SERVICES



Motivations

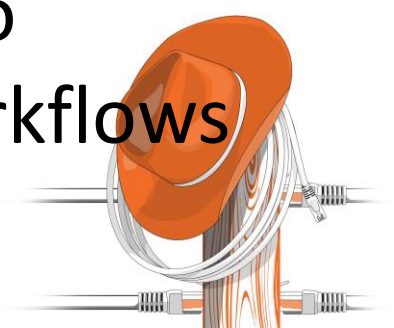
- Need to support:
  - Long-term reservations (months) for “traditional” and Hadoop job types
  - Persistent database provisioning
  - iRODS provisioning and data management service controls (fixity/audit/etc)
  - New workflows as they are developed



TACC

Portals for Everything

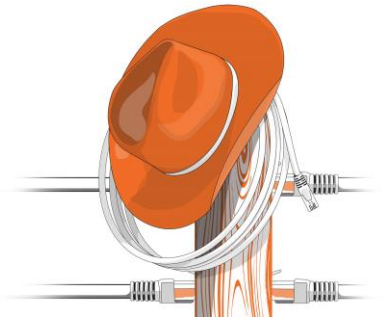
- TACC team has contributed significantly to our own portal and the XSEDE user portal
- Very successful with users
- Newly developed Wrangler portal provides interface and backend infrastructure to manage services, reservations and workflows



Wrangler Portal Architecture

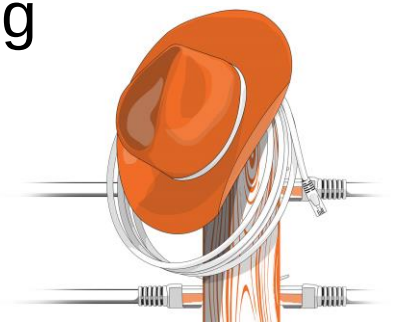
- Django-based website with backend MySQL
- RabbitMQ/AMQP-based messaging system
- Scheduler reservation and job scripts
- System-level deployment scripts
- Information services

TACC



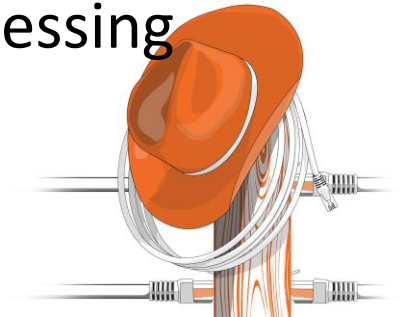
Wrangler Services

- “AutoCuration” - Data management services
- Data dock and other ingest services
- Globus Online Dataset Services
- Persistent and temporary service management
  - Postgres/MariaDB/MongoDB
  - Open web publishing/Web-based Data sharing
  - Other “science gateways” as appropriate



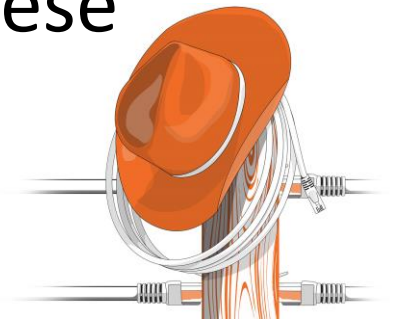
Wrangler and Databases

- Databases are a natural area of focus for Wrangler
- Persistent Databases
  - Data Collections/Resources, Stream Processing
- “Transient” Databases
  - Database used as temporary engine for processing
  - SQLite, other node-local options



Wrangler “Workflows”

- Hadoop framework needs special deployment steps at the time of reservation/execution
- Temporary database deployments require special setup process
- Once you have the framework to do these things …

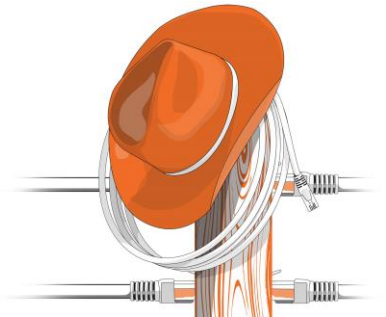


TACC

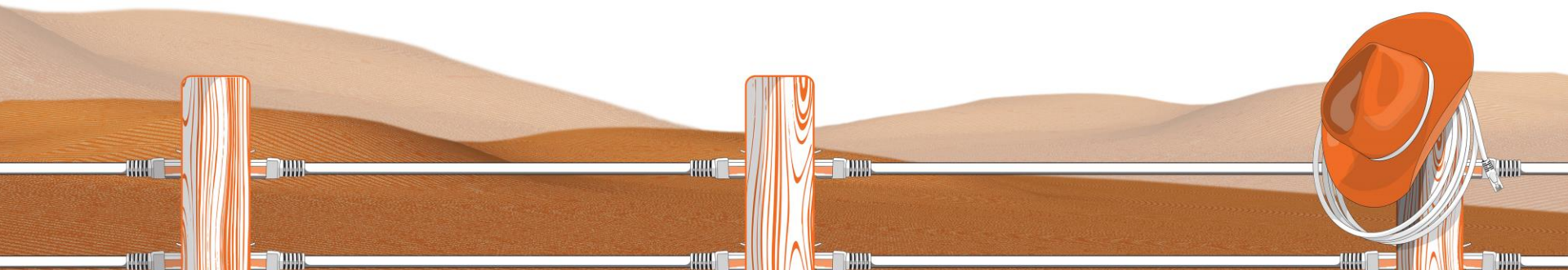
Wrangler Planned Workflows

- Bioinformatics
  - BLAST – Ubiquitous in genetic preprocessing
  - OrthoMCL – Protein grouping
  - iPlant integration to support community workflows
- Media curation:
  - Large-scale image analysis/conversion
  - Video and audio processing

TACC



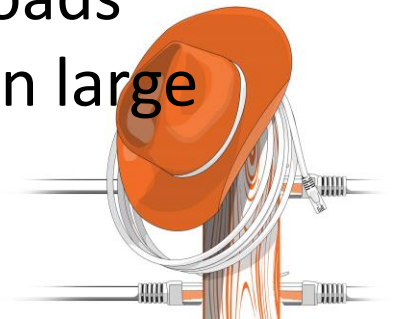
EARLY APPLICATION RESULTS



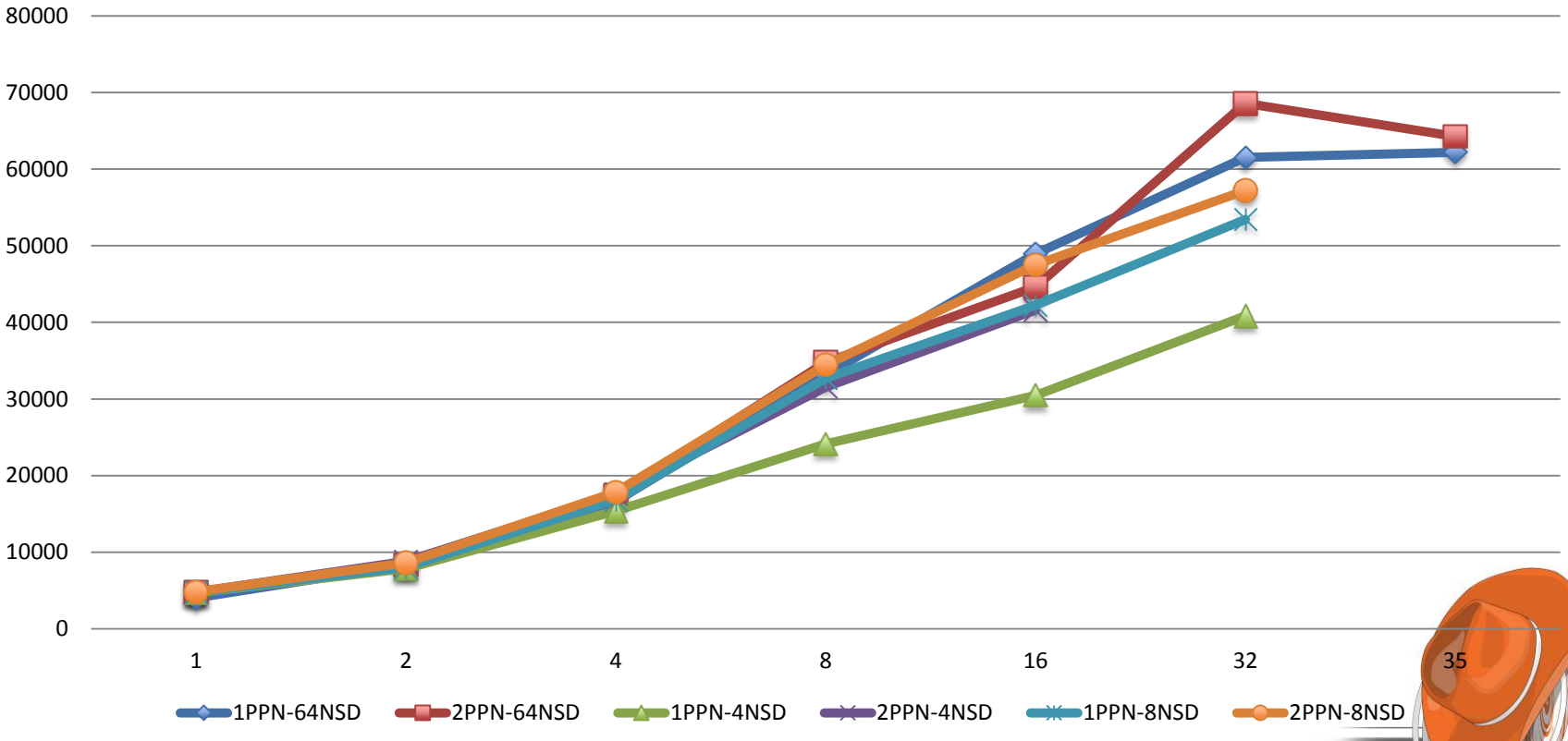
Some Application Modes

- Persistent Database – data hosting, shared curation, web application backends
- Temporary Database – SQL or NoSQL as application engine
- Traditional MPI – applications with IO-heavy requirements
- Hadoop – applications with IOPS-heavy workloads
- Bioinformatics – Perl/Python/Java operating on large datasets in a mostly serial fashion

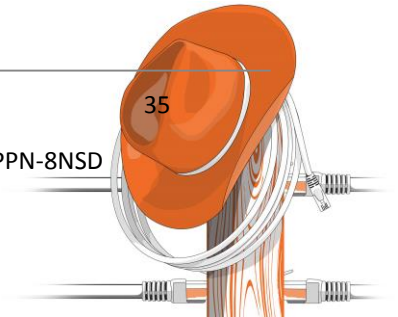
TACC



GPFS Parallel Read Performance

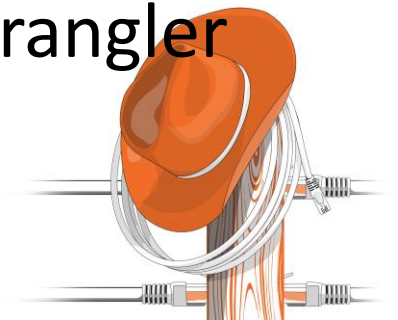


TACC

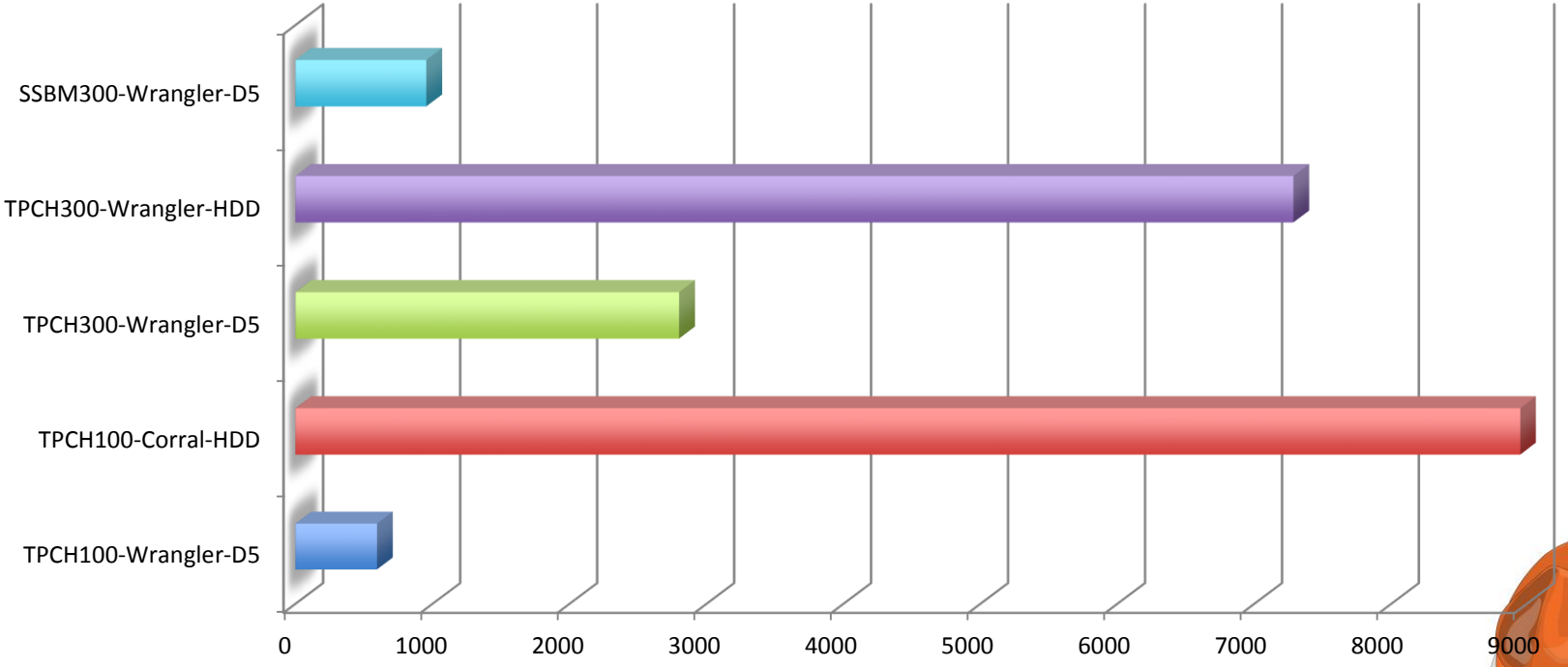


Persistent Database and Web

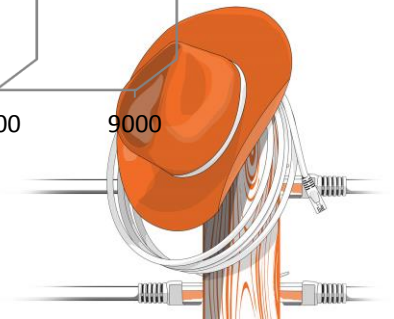
- Arctos collections website has >100GB database, ~10TB image collection
- Site performance directly attributable to database performance = storage performance
- Database to Flash, Images to Disk
- Also, intend to use multi-site nature of Wrangler to provide higher reliability



TPC-H & Star Schema BM
Postgres Seconds/Query

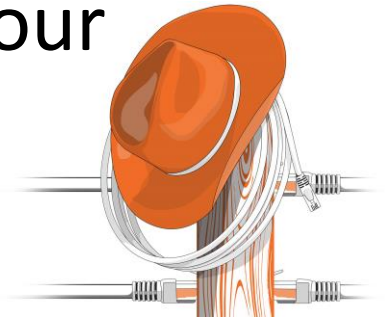


TACC



OrthoMCL

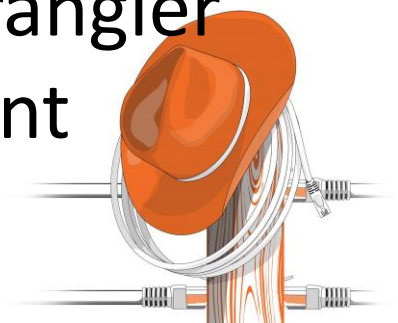
- Protein grouping application
- All significant computational effort expressed in SQL and performed in the database
- Not necessarily optimized SQL
- Execution time from >10 hours to ~1 hour moving from similar disk-based system



TACC

Image Collection Curation

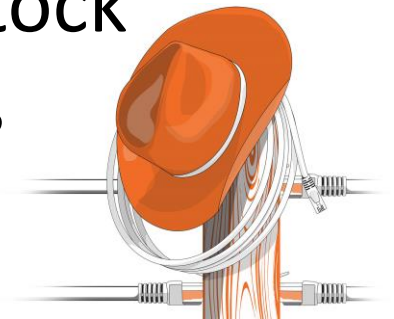
- Multiple projects working with 10s or 100s of thousands of high-quality images (~1-5GB)
- These images may require conversion, resizing, analysis, en mass
- Initial development on Stampede – Wrangler provides ~3X performance improvement



TACC

Social Science/Economics Analysis

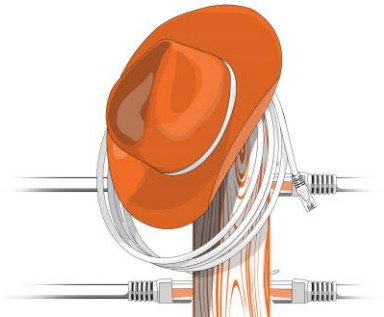
- Databases are commonly used as really big spreadsheets for survey results etc
- Data subsets are retrieved using R/Python/SAS/etc for further analysis
- Example: A researcher who has daily stock market activity data for last 100+ years



TACC

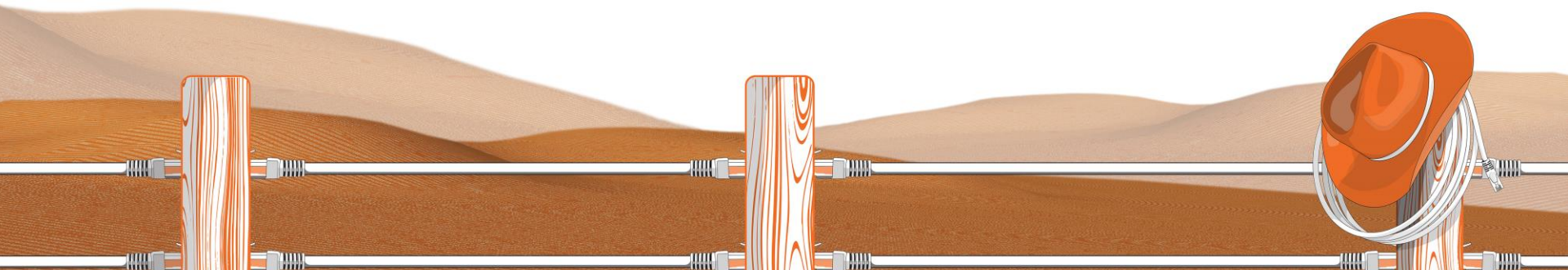
Stock Market Analysis Workflow

- Create and load “transient” database on flash file system
- Create derived databases with data subsets
- Save resulting database, restart on future job executions
- Save database state like checkpoints…



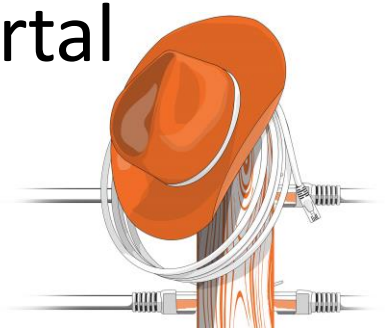
TACC

FUTURE DEVELOPMENT ON WRANGLER



Wrangler Status

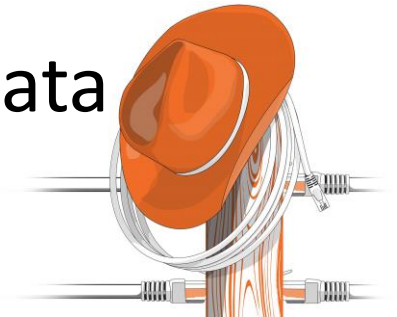
- DSSD products not in GA quite yet
- Dual-controllers should double performance
- Currently in “friendly user” mode
- Interested users can request startup/early user access through the XSEDE user portal



TACC

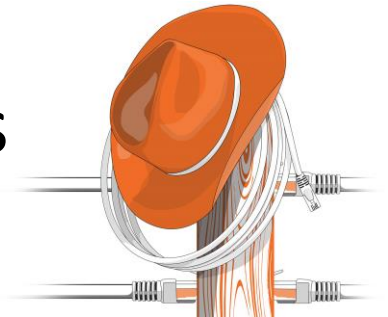
Ongoing Wrangler Tasks

- Development efforts always expected to continue beyond deployment
- Data Management/Curation task automation
- Data Publication – more flexible models
- Workflow capture and reproduction
- Gateway Hosting – bring your code + data



Importance of Data Publication

- Wrangler has an inherent notion of the data life cycle
- This includes long-term storage, publication
- Will provide mechanisms for acquiring DOIs, publishing one or more files
- Will support varying levels of openness



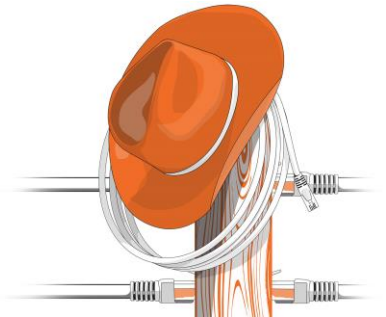
TACC

Acknowledgments

- The Wrangler project is supported by the Division of Advanced Cyber Infrastructure at the National Science Foundation.

  - Award #ACI-1447307 *"Wrangler: A Transformational Data Intensive Resource for the Open Science Community "*

TACC



Acknowledgments 2