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 deployed with these goals in mind.
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.

• 11% of (5 million+) Stampede Jobs spend significant time on I/O (Processor blocked > 50%)
• These jobs have much higher Lustre BW and MDS ops on average than efficient jobs
• Key examples of the kinds of workflows that can take advantage of Wrangler’s advanced IO system
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 – original targets
  - 1TB/s
  - 250M+ IOPS
Wrangler At Large

TACC

- Mass Storage Subsystem
  - 10 PB
  - (Replicated)
- IB Interconnect 120 Lanes
  - (56 Gb/s) non-blocking
- Access & Analysis System
  - 96 Nodes
  - 128 GB+ Memory
  - Haswell CPUs
- Interconnect with
  - 1 TB/s throughput
- High Speed Storage System
  - 500+ TB
  - 1 TB/s
  - 250M+ IOPS
- 40 Gb/s Ethernet
- 100 Gbps
- Public Network
- Globus

Indiana

- Mass Storage Subsystem
  - 10 PB
  - (Replicated)
- Access & Analysis System
  - 24 Nodes
  - 128 GB+ Memory
  - Haswell CPUs
DSSD Storage

• The flash storage provides the truly “innovative capability” of Wrangler
• Not SSD; PCIe V3 connectivity to NAND flash giving performance without the overhead of the traditional “disk” interface.
• Used as traditional Block Devices
• APIs that integrate natively with apps (i.e. HDFS direct integration, native object store)
• 600 TB of **usable** storage in 10 servers
  - Nearly 100k NAND flash dies
  - 960 Gen3 x4 PCIe links to the storage system.
  - Fault tolerant internal RAID able to sustain writes without loss due to Memory Wear
Performance

<table>
<thead>
<tr>
<th>Acceptance Test</th>
<th>Value</th>
<th>Target or Comparison Value</th>
<th>Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS I/O</td>
<td>968 MB/s</td>
<td>33 MB/s</td>
<td>✓</td>
</tr>
<tr>
<td>TPC-H100</td>
<td>594</td>
<td>8951</td>
<td>✓</td>
</tr>
<tr>
<td>TPC-H300</td>
<td>2800</td>
<td>7284</td>
<td>✓</td>
</tr>
<tr>
<td>YCSB</td>
<td>11978</td>
<td>4000</td>
<td>✓</td>
</tr>
<tr>
<td>D5 Throughput</td>
<td>100.7 GB/s</td>
<td>100 GB/s</td>
<td>✓</td>
</tr>
<tr>
<td>D5 IOPS</td>
<td>25.6 M</td>
<td>25 M</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Ran the standard HPC tests, but they weren’t particularly interesting. . .
- YCSB – Yahoo Cloud Service Benchmark (vs. best reported number for MongoDB on Amazon SSD instances on web in 2014)
  - Based on the YCSB benchmark, Wrangler is worth $75M of Amazon time
- TPC – Database benchmark, mostly driven by HP, using PostgreSQL with 100GB and 300GB test cases, vs. Corral.
- DFS I/O is a Hadoop benchmark, vs. our older Hadoop cluster, beats it by 30 to 1.
Updated System Configuration

• Has two computing partitions
  – Normal partition – non-HDFS workflows using 240 TB GPFS flash
  – Hadoop partition – HDFS based jobs using 240 TB storage
• 26 service nodes for dedicated services & testing
  – Hosts persistent services (db, iRODS)
  – 120 TB for persistent data and testing new deliveries
  – Testing all to all & fat nodes
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
More Data Services

• Training
  – 7 Training sessions on Wrangler including specific topics such as DB, Hadoop, and R on Wrangler since December
  – Teaching Teachers in ECSS and Campus Champions

• Data Management & Autocuration Service (Ops Phase)
  – Leveraging Globus Online Dataset Services
  – Data Organization tying to research project to facilitate tracking of a projects data assets
  – Data: Provenance – Fixity & Usage
Science Successes on Wrangler
Just This Past Weekend

- **Cloudera Zika virus hackathon.**
  - ~30 attendees, a mix of Hadoop, Rstudio, Jupyter notebooks. Used Wrangler as the backend.

- **Eddie Garcia, Chief Security Architect, Cloudera, and Zika Hackathon co-organizer:**
  - "It's important ...to understand that there is this Wrangler platform, that we can work together and bring in more users because it is a massive and high performance system that is just hungry for different types of projects like Zika... This Zika hackathon is just one small example of how we can do it and for others to look at it, repeat it, and work closely. “
Hadoop-based Data-Intensive Apps
Chris Mattmann, Cal Tech/JPL

• Use of Wrangler to explore a variety of research areas and new Hadoop components
  – Square Kilometer Array data processing
  – DARPA Memex program: directed internet search/data mining for intelligence applications
  – More collaborations to come
Hadoop-based SKA processing

• Low Frequency Aperture Array = 1PB/day
• Testing of Apache Kafka stream processing
• Initial testing goal 10GB/sec by 2023
• First Wrangler test = 6.5GB/sec (!)
• Bottleneck in JVM, not hardware
• Now expect to exceed target >7 years early
Chris Mattmann (JPL) crawled the Apache SVN repository
- 133 Repositories, 3.5 Million files
- Used Distributed Release Audit Tool (DRAT) using Apache Tika and Object Oriented Data Technology (OODT)
  - Crawling 7x speed up
  - Indexing 2x speed up
  - Analysis 2-3x speed up
Mattmann Group Collaboration

• Extensive collaboration between PI, post-docs, graduate students, and TACC staff
• Understanding best use of Wrangler
• Optimizing configuration of Hadoop stack
• Very positive feedback:
  – “[feels like] a limitless environment”
  – “super slick and easy to work in”
Cosmological Structure Formation
Michael Norman, UCSD/SDSC

• Large-scale cosmological simulations
  – “Classic HPC” on Stampede and Blue Waters
  – Enzo & GADGET-2 cosmological hydrodynamics
  – Massive output data sets needing analysis/viz
  – 100TB+ of data to process on Wrangler
Cosmological Structure Visualization

• Initial dataset: 50TB Blue Waters -> Wrangler
• Using Globus Transfer, copied in <2 days
• Feature detection and projection
• 40% improvement in performance on computation-heavy, data-intensive processing
• Visualization time from days to < 1 day
Identified Population III Galaxies
Analysis of Market Data
Scott Murray, Georgia State University

– Performance of stocks with “lottery-like” features
– Understanding information used to value stocks
– Understanding investors “ambiguity attitudes”, e.g. relation of ambiguity to stock desirability
– Heavy focus on model fitting to empirical data

• Introduced to Wrangler through XSEDE Novel and Innovative Projects program
  – "the support that I have received ... has been phenomenal"
Analysis of Market Data 2

• Technical Details:
  – 2 Billion-element database of input data
  – ~2.5TB raw storage used for database
  – MariaDB “transient” database used for storage
  – SQL Queries generate input datasets for R tasks
Cooperatively-Developed Workflow

• Empirical verification of established theories & new hypotheses of market/investor behavior
• Each “run” consists of hundreds of R tasks with independent datasets (from SQL)
• TACC Launcher script used to fit jobs to cores
• End result is literally hundreds of times faster than previous analysis work
Electrical Engineering
Ali Yilmaz/Guneet Kanur, UT Austin

• Integral Equations applied to Electromagnetic Backscatter simulation

• Stampede runs generating excessive I/O due to datasets in the hundreds of TBs

• Port to Wrangler reduces I/O load on Stampede and accelerates application >500%
Electrical Engineering on Wrangler

• Even existing “traditional HPC” codes can take advantage of Wrangler if I/O fits the profile
• I/O intensive users on traditional Lustre file systems won’t be happy, nor will sysadmins
• Not expected to be a huge segment of users but very important to the ecosystem
• Demonstrates another route to use of Wrangler
Comparing DNA sequences to find common sources of biological characteristic

- MySQL DB stored on GPFS distributed over 4 D5s
- Wrangler did in **4 hours** what our other systems could not complete in multiple days
Genomics/Protein Grouping

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 Before Wrangler

- 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
    - Developer concerned data set was too complex to complete successfully on any system
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
FMRI Literature Mining
Russ Poldrack, Stanford University

• Exploring relationship between cognitive concepts and brain regions
• Mining a large corpus of existing research publications for possible inferences
• Initial database on Wrangler of 11,000 abstracts, 27,000 “brain regions”, 135,000 “cognitive concepts”
• Next stage will utilize machine learning to make inferences and learn relationships
Mathematica on Wrangler

- Wolfram staff participating in porting efforts
- Mathematica not parallel by nature
- Using flash file system as communication layer
- Enabled up to 360-way parallelism with a few weeks of effort from Wolfram
- Shows novel use of D5 hardware with GPFS
- Provides easy-to-use introduction to using “big iron”, smooth transition from desktop platforms
Key Lessons from Users So Far.

• Wrangler is “game-changing” for at least 3 categories of users:
  – XSEDE users accelerating data-intensive apps
  – XSEDE users running new applications
  – Users new to the XSEDE ecosystem

• Significant support efforts and partnership with users is essential to effective utilization
User Environment
Why Wrangler is Different

• Use cases not satisfied by a batch processing style environment
  – Many non-traditional users are not CLI capable (and don’t want to be)
  – Hadoop environment != mySQL != iterative data analysis != streaming analysis system

• Better suited for a more cloud-like provisioning system.
SLURM job environment

• Standard SLURM based job environment
  – Three queues, normal (GPFS storage), Hadoop (HDFS storage with reservation), and debug
  – Jobs limited to 48 hour runtime (4 hours in debug)

• Support reservations in normal and Hadoop queues
  – Reservations preserve storage between jobs
  – Reservation required to allocate backing HDFS file system for Hadoop jobs

• Projects charged for duration of Reservations and for all jobs not run under a reservation
Basic Information about the System

- System Overview
- Instructions on how to request an allocation
- Information about supported software and environments on Wrangler
- Users Authenticate using TACC or XSEDE credentials
  - Requires account be part of an allocation on Wrangler
Database Allocations

Users can request one instance of Postgres and MySQL hosted on the IU or TACC database service node.

Can specify the name of the schema and who will be the Administrator of the Database (has full permission on schema).

All users can read DB configurations.
iRODS configuration

Can request iRODS collection space with name and specified size

Validates project has enough storage allocation for collection

Users can get iRODS configuration parameters
Reservations

All users can create reservations for compute nodes and GPFS.
Checks that allocation does not exceed remaining allocation (considers all pending reservations as well).
Can start as soon as possible (within 5 minutes) or starting on a particular date and time.
Users can track reservation from Pending to Active to Completed.
Creator of reservation can cancel reservation.
Hadoop Allocation

Works like “normal” reservations except Hadoop system is configured and HDFS formatted for project on nodes in reservation

HDFS formatting takes 5 minutes to complete
Hadoop Details

• Cloudera 2.4 Hadoop installed on all compute nodes including Spark, Pig, Hive, and Mahout packages

• On demand Hadoop clusters created when reservations start
  – Email sent with final configuration to project about cluster configuration

• Users submit Hadoop or Spark jobs from any node in the reservation
  – Use iDev command to get interactive CLI session or VNC for web GUI

• At termination of job, HDFS file system is unmounted and is subject to purge
Wrangler GUIs

Visualization portal allows for graphical interactions with Wrangler

- VNC for desktop applications
- Proxied iPython/Jupyter and Rstudio sessions (pySpark, Spark R, and R Hadoop available)
Convergence of HPC and Big Data

- Comparing software stacks and functions
- Cloud as the merger?
Convergence of HPC and Big Data

Big Data Meets High Performance Computing

Intel® Enterprise Edition for Lustre® software and Hadoop combine to bring big data analytics to high performance computing configurations.

- Worked with Intel on POC around Lustre for Hadoop before Wrangler
  - Intel working on alternate schedulers for Hadoop as well
  - GPFS/IBM also has working solution
Moving forward on Wrangler

Wrangler provides a single system to explore merge of HPC and Big Data environments.

Users can transition between current less unified systems more transparently.
Wrangler in the TACC/XSEDE Ecosystem

• TACC, PSC, and XSEDE are all traditionally providers of HPC, Visualization, and storage systems.
  – And we still are.

• But many 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 complements them, and brings in new classes of users!
  – Come try it out!!
Thank You!

Dan Stanzione
dan@tacc.utexas.edu