The Data Must Flow

Infrastructure, Architecture, and Organization
Data @ Scale @ Broad

Chris Dwan (cdwan@broadinstitute.org)
October 14, 2014
Big Data: Data that is so poorly organized and so fundamentally un-curated that the domain experts who created it cannot directly query that same data to answer questions within their own domain.

-Leslie Greengard, as overheard at the NY Genome Center
Chris Dwan

• Research Computing @ Broad
  – Linux and Storage Operations
  – Scientific Computing Services
  – Data Services
  – Scientific Pipeline Operations

•Previously
  – NY Genome Center
  – BioTeam
Conclusions

**Not overly concerned:** raw data volumes, CPU counts, and network bandwidths.

**Concerned:** Most genomic data will not be produced, owned, or hosted by any single project or at any single institution (not even the Broad!)

**Federation:** The most interesting scientific and clinical results will be achieved by researchers who are able to *seamlessly* leverage data *across multiple institutions*.

**This Requires Change:** Organizational and technical. At all levels.

**We are hiring:** Broad needs software and data engineers. Right now.
Three Topics

**Infrastructure**
- Shared Everything -> Shared Nothing
- Openstack
- Our road to the cloud
- Object Storage

**Data Architecture**
- BOSS: Broad’s Object Storage System
- Genomic Data “Vault”

**Organization**
- The genomic data “roadmap”
- Genome Bridge / Prometheus
- Data Sciences / Data Engineering
Scale @ Broad

**People**
- ~750 salaried research staff
- ~170 software engineers / computationalists
- ~2,500 “affiliate” members

**Technology**
- ~28PB usable in-house storage (1/2PB / month)
- ~18PB unique file data (mostly genomic sequence)
- ~8,000 cores High Throughput Computing environment

**Samples**
- ~700k samples in ~1,000 cohorts
Infrastructure:
Computing

(but first, dashboards)
Shared Everything

Shared services: IAM, various data registries, multiple job control systems

1,000+ linux Servers
(of varying ages, makes, models, and configurations)

10Gb/sec Ethernet

28P usable storage
(across more than a dozen filers of various capacities and performances)

Broad HPC Environment

Shared everything:

– Presents a “flat” environment to the user / developer
– Heavy burden on IT staff to make “everything” work “everywhere”
– Brittle, highly susceptible to “noisy neighbors.”
– Difficult to provide elastic resources or “move to the cloud.”
Localized shared storage (pods)

- **Localization Domain (Pod 1)**
  - Fast shared storage, dedicated to this pod
  - 40 – 80 servers (Coherent hardware configurations)

- **Pod 2**
  - 20 GB/sec Ethernet

- **Pod 3**
  - 10Gb/sec Ethernet

- **28P usable storage**
  (across more than a dozen filers of various capacities and performances)

Performance profiling and tuning of resource intensive jobs

“Shared everything” usage is still available

Performance improvements available to users willing to implement “stage-in” and “stage-out” (2-3 months effort)

Physical infrastructure improvements Cabling and hardware deployments become sane
Localized shared storage (pods)

Performance profiling and isolation of resource intensive jobs

“Shared everything” usage is still available

Performance improvements available to users willing to implement “stage-in” and “stage-out”

Physical infrastructure improvements

“Cloud pod” becomes reasonable, though still tricky
**Substantial changes** to end user codes (1 – 2 years effort)

“**Shared everything**” no longer sensible or even truly available

**Massive performance and cost improvements are** available, though still require detailed engineering

**Flexibility for job relocation:** “Containerized” codes can be scheduled in a “data aware” manner.
Auto Scaling Algorithms

• Applications are **aware** of underlying infrastructure
• Applications **deploy additional resources** as needed to maintain performance requirements

   **We are not here yet in any substantial way.**
   **Most likely, neither are you.**

• **Modern / web-scale** software development follows this model.
• **Beware** vendors who sell you auto-scaling tools without direct support for migrating your underlying algorithms.
Infrastructure:
Openstack
Openstack @ Broad

- Icehouse release is running on nearly 2,000 cores in our HPC environment
- Core genomics and cancer research workflows are migrating (mostly) seamlessly
- 2 month, “whole team” effort.
- Broad’s entire HPC environment will soon be virtualized in-place Openstack

Substantial thanks to Cisco, Intel, Red Hat, and Solina for making this possible
Openstack: “Seamless” Migration

Genomics Platform
Orchestration layer (Picard / Zamboni)

Cancer Program
Orchestration layer (Firehose)

Broad shared “farm”

We only asked Openstack to do exactly what we were already doing, and it was still a really hard project. Imagine a “simple” jump directly to a public cloud.
Openstack, open issues

Excluded from our project:
- Software defined networking (Neutron)
- “Cloud” storage (Cinder / Swift)
- Monitoring / Billing (Ceilometer, Heat)
- High Availability on Controllers

Custom:
- Most deployment infrastructure / scripting, including DNS
- Network encapsulation
- Active Directory integration
- All core systems administration functions

Core Message:
- **Do not change both what you do and how you do it at the same time.**
- Openstack could have been a catastrophe without rather extreme project scoping.
Virtual Private Cloud on Amazon’s Web Services

In House Job Scheduling: SGE / LSF

Increasing number of physical hosts managed by Openstack

Virtual Linux servers

External Job Scheduling: Cycle Computing’s CycleCloud

Virtual Linux servers

Virtual Private Cloud on Amazon’s Web Services

Job dispatch: Cycle Computing's SubmitOnce

Jobs must be tagged by the submitting user as “cloud appropriate,” “cost recoverable,” and in-house resources must be un-available.
Infrastructure: Data
File based data storage breaks down at scale

**Number of files:  10^9 Files**
- Directories must either get too wide or too deep
- Implicit access controls become unmanageable
- Data incoherencies can proliferate and hide for years

**Volume of data:  5P**
- We have enough integers to store more

**HOWEVER**

Shared-everything means all files have noisy neighbors
System maintenance / break-fix impact is simply too broad
Constantly *reacting to symptoms* of poor performance
Really quite difficult to get to outsourced storage in this model
Object storage is how information companies store data at scale

“Large” data volumes: 200P
“Impressive” data volumes: > 1E

Object storage must be
S3 compliant: Amazon remains the de-facto cloud standard
Geographically protected: Sustain loss of entire data centers
Cost effective at scale: You should see savings of 50% and more on storage hardware alone.

This changes the end-user relationship to the data
Substantial changes to end-user code
Stage-in / Stage-out rather than use in-situ
Data models are required to avoid merely transplanting a mess
REST access to data:
- Substantial rewrites to change from POSIX to HTTP semantics
- Concepts of “ownership” must fundamentally change

What was implicit must be explicit
- Data are owned by access / management applications rather than users
- Those access / management applications must handle IAM, logging, and many other functions

Adopting object storage imposes substantial work on the data owners to create usable data models.
Data Architecture
**BOSS: Broad’s Object Storage System**

- **“Simple” registry** of sequencing data produced by the Genomics Platform.
- Sequencing results accessed via a **“Broad ID”** rather than a file path.
- Allows storage team to **relocate data** from file to object storage, in-house or external, without disrupting user operations.
- Requires **substantial re-factoring** of end-user code to stage-in / stage-out rather than using data in-situ.
Genomic Data Vault

- **Indexing and Organizing** more than 12P of human genomic data
- **Designed from the beginning** to be an integration point for the world’s genomic data
- **Supporting queries** across consent, sample handling, and phenotype
- **This is the data model** required for the Broad to make effective use of object storage
Genomic Data Roadmap Committee

• 13 participants from across the Broad:
• Intensive effort for the summer of 2014, including interviews with dozen of thought leaders both in-house and external.
• High visibility to executive leadership

– **Capabilities** in the Broad of the future
– **Technologies** needed to implement those capabilities
– **Organization** to make it happen

*We recommended that Broad create an integrated Service Oriented Architecture for genomic data science, explicitly engineered for the needs of researchers worldwide.*
Project: Prometheus
Genome Bridge
- An *external brand and delivery vehicle* for Broad’s data and analysis services

Prometheus
- A *project* (60+ FTE, plus collaborators) to revolutionize how the Broad organizes, analyzes, and shares data

Data Science and Data Engineering (DSDE)
- A *new organizational unit* (25+ FTE) at the Broad that centralizes many infrastructure, computational biology, and software engineering functions
We propose creating a Broad where:

- Researchers have secure and compliant access to all available genotype/phenotype data

- Scientists spend the majority of their time answering research questions and developing new methods rather than wrangling data and IT

- Researchers build on each other’s work through frictionless discovery of colleagues’ data, methods and research

- Computing is elastic, scalable, and inexpensive

- Computational biologists (within and external to the Broad) share a unified conceptual model for analysis of variants and other derived data (e.g., expression) – yielding vastly improved and harmonized results and accelerating scientific discovery

- Results are delivered to researchers and clinicians in user friendly, interpretable ways
New organizational unit within the Broad

- ~25 FTEs, primarily software development and computational biology
- Integrated with a team of 60+ from across the Broad
- Dedicated to data engineering at scale

We are hiring:

- Senior software engineers, data scientists, and pipeline infrastructure dev/ops staff
Conclusions

**Not concerned:** raw data volumes, CPU counts, and network bandwidths.

**Concerned:** Most genomic data will not be produced or hosted by any single project or at any single institution (not even the Broad!)

**Federation:** The most interesting scientific and clinical results will be achieved by researchers who are able to *seamlessly* leverage data across multiple institutions.

**Change:** Organizational and technical. at all levels.

**We are hiring:** We need software and data engineers. Right now.
Thank You

Research Computing @ Broad
Cope Frazier*, Dave Gregoire, Eric Jones, Steve Kaplan, Kirill Lozinskiy, Paul McMartin, Matt Nicholson, Sam Novod, Ray Pete, Peter Ragone, Vince Ryan, Katie Shakun, Zach Shulte, Ken Streck*, Scott Sutherland, Kathleen Tibbetts, Charlotte Tolonen, Elsa Tsao, Jay Weatherell*, Ellen Winchester

Roadmap Committee
Eric Banks, Mary Carmichael, Kris Cibulskis, Chris Dwan, Tim Fennell, Jason Flannick, Eric Golin, Adam Kiezun, Daniel MacArthur, Karen Minn, Mike Noble, Anthony Philippakis, Michael Reich

Leadership @ Broad
David Altschuler, Eric Lander, Stacey Gabriel, Gad Getz, Samantha Singer, Matthew Trunnell

Technology Partners and Collaborators
Brocade, Cisco, Cycle Computing, EMC, Infinidat, Intel, Isilon, Red Hat, Scalable Informatics, Violin Memory,
Questions

• cdwan@broadinstitute.org