OSG and some genomic researchers

Ian Fisk

Community

- For the last 8 months I have been trying to support data management and access for a community of genomic researchers
  - There are more sources of data than HEP, but the most advanced research sequencing machines are expensive and there are only a few
  - Communities collect data from individuals based on what they are trying to study
    - Regions, conditions, illnesses, etc.

What we set up?

- We established a VO (SCDA)
  - Currently ~50 members
- We put host certificates on resources at 3 sites for gridftp servers

File Access

- The machines produce a reasonable standard raw data format called BAM files
  - ~10GB file per person for whole exome (A study of about 1% of your DNA) Mutations here can have severe impact on the rest
  - >100GB file per person for whole genome sequencing. Modern machines can sequence the entire genome
- A study of a group might be a few hundred individuals
  - Raw data in the few TB range for exome and few hundred TB for full genome

Processing

- The processing steps are to apply a series of defined steps
  - The code for each step is often written by an independent party and shared
    - Techniques and code improve, groups often go back to the source data and reprocess

Kinds of Processing

- Align the data
  - Make sure you have the region you were looking for
- Variant Calls
  - Compare your sample to an agreed standard. Identifying the number of places that your sample varies from some “nominal”
- Then start looking for correlations
  - Frequently work is done with groups of related individuals

Differences with HEP

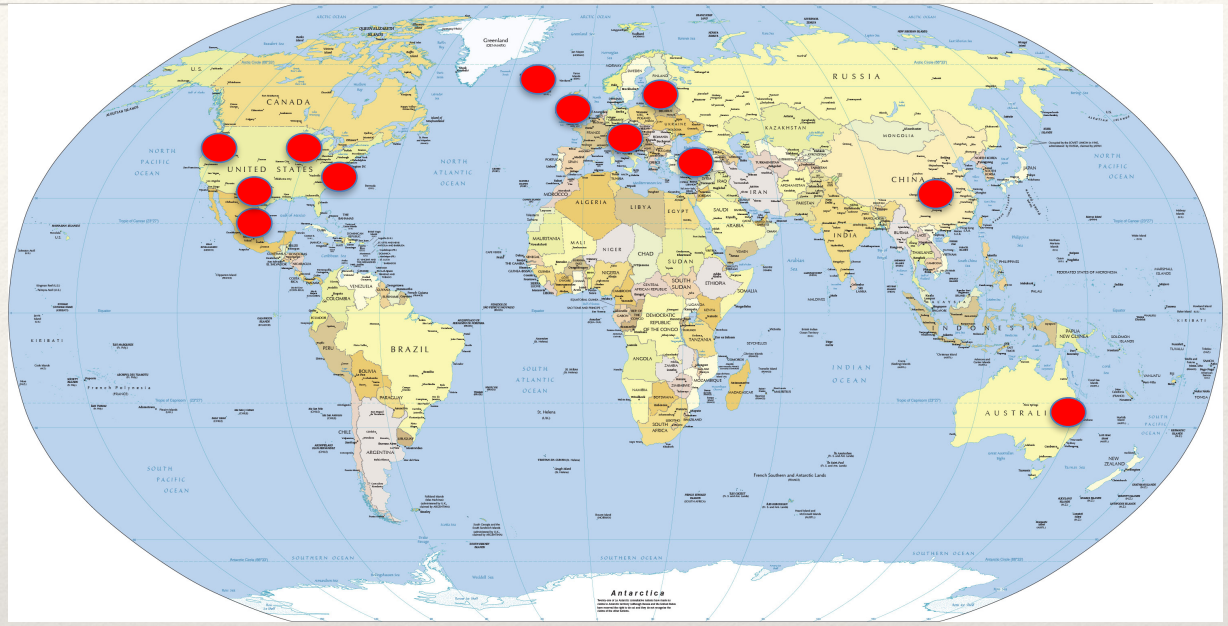
- Software is written by external people and there is a reluctance even to change where the default install it, let alone build in new IO libraries
  - Solutions that don't require recompilation like LD_PRELOAD and FUSE may get better adoption
- Steps are well defined and can often be seen as canned applications
  - The steps are updated so people do a full reprocessing from the raw

Size of the problem

- Groups were encouraged to host data in Amazon
  - 0.5PB even in the little local community
  - Advantages is that it's well understood how to serve to EC2 processors
  - Disadvantages is that it's expensive to export from Amazon and the government got nervous about storing sequencing data and decided to close the facility down and all data needed to come out

Current Distribution

- Currently this genomics community uses FNAL as an archival system
  - Recently imported ~400TB of data primarily from S3
  - 2 100TB samples were exported from FNAL using GridFTP to Iceland and Oregon for additional processing



- The community created and made publicly available 11TB of diversity project data
  - 300 people from all over the planet

The Challenge (1/2)

- There are about 40 entities that want samples between 10-100TB
- There is no real infrastructure for data management
  - File lists are sent with checksums in manifests
- These are labs with firewalls and data has grown much faster than expertise, so little community knowledge for how to move big samples around
  - Bare GridFTP is not completely user friendly nor is the entire grid certificate infrastructure

The Challenge (2/2)

- Path through the files in question is a semi-pathological
  - Unpacking internal buffers and retrieving objects across large swaths of the file
    - Access through the file during analysis is not linear and applications know nothing of training or pre-fetching

Hoarding

- Unlike a big central detector project where everyone is a member of the collaboration with rights to the data, these communities often share samples with detailed written agreements for specific periods of time
  - The convenience of being able to access anything will be weighted against the certainty of having the samples under your control

Things to explore

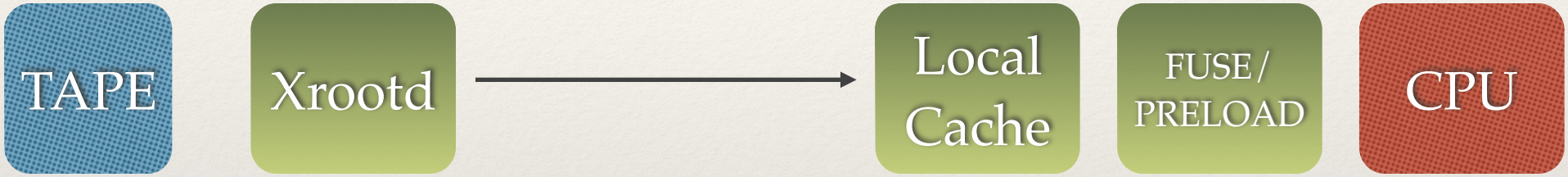
- We would like to process on OSG
  - A typical processing pass is a few hundred CPU months
- There are a few problems to solve
  - Environment
  - Data delivery
  - Data Management

Environment

- A lot of the code is written by others and canned
  - People are not used to changing how it's deployed (even to change the paths)
  - We will look at the common CVMFS
  - Maybe something like Docker would be a good working model

Data Delivery

- During the spring we will begin deploying something very simple



- Looking forward we have the potential of placing intermediate services for data serving
  - Currently the fully public samples are 10-15TB and could be replicated
  - Proposal for opportunistic site caches would be interesting

Data Management

- One aspect that is clearly lacking is data management
  - Datasets are defined by manifest lists (text files)
  - Where data physically is documented on web pages
- Works for a limited amount of data but will not scale for long

Outlook

- Looking forward to ramping up the scale
  - Data distribution is at a reasonable level
  - Wide area access through federation and opportunistic processing through OSG are goals