Galaxy Farm

Kyle Ellrott, Dannon Baker
Current Clustering

Shared Filesystem and Queue
For High Performance

For large scale genomics using network shared drives doesn't scale

Galaxy needs a concept of data locality

Be able to do workflow end-to-end on remote node, only transfer back the required results
What kind of high performance

Applying standardized pipelines to massive data sets

Example:
Run standard re-alignment pipeline on every TCGA WGS
Farming

Workflow

Galaxy

Galaxy

Galaxy

Galaxy

Galaxy

Galaxy

Galaxy

Galaxy

Galaxy

Galaxy
Things that need to move around

- Work
- Workflows
- Tools
- Data
- Provenance
What do you need

Essentials

- Workflow Request Batches
- UUIDs (on everything)
- Data Referrals (UUID resolution)
- Standardized Job/Workflow/Tool/Provenance Transfer

Fancy Stuff

- Docker
- Gossip Protocols
- Peer Authentication
What's Done? What's needed?

Done:
● Dataset UUID are now on by default
● Workflow Batches Implemented
● Mesos based launching

Needed:
● Workflow Request Transfer
● Auto History transfer
Going forward?

Get a working prototype running

Do a large scale run

The tools that you need for farming are the same ones you need for federation...
Extra Slides
In the Galaxy DB, we need to separate the idea of a Workflow Request and a Workflow Instance.

Add the concept of Batches, that can be scheduled at the same time (with common scheduling configurations).
UUID resolution

14b228c0-72c5-4c25-aaa5-da8e5b13eef7
Data Referral

Cluster File System

Gene Torrent

Swift

14b228c0-72c5-4c25-aaa5-da8e5b13eef7
Job Offloading

Work and Data Transfers