BIOTEAM
Enabling Science

ENABLING DYNAMIC SCIENCE WITH FLEXIBLE INFRASTRUCTURE

Anushka Brownley, Senior Scientific Consultant
Aaron Gardner, Senior Scientific Consultant

GALAXY COMMUNITY CONFERENCE 2014

BIOTEAM
Enabling Science

AGENDA

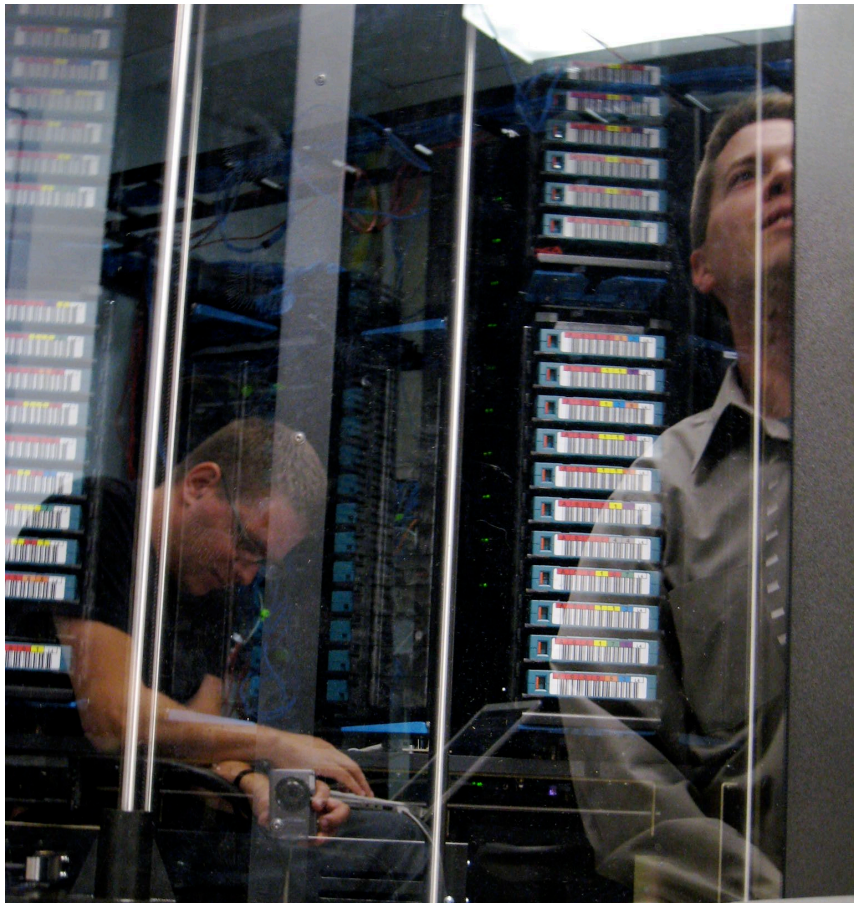
Galaxy

- Who We Are
- SlipStream Galaxy Appliance
- Science vs Infrastructure
- Hybrid Computing: Flexibility and Scale
- Looking Ahead

Over a Decade of Life Sciences IT Consulting

- Staffed by **scientists** forced to learn IT to get research done

- Served over **400** organizations
  - Academic, Non-profit
  - Government, Military
  - Pharm, AgBio, Biotech
  - Cloud & Datacenter Providers



Bridging the IT Gap

- Encapsulate IT best-practices expertise to eliminate redundant effort spent building IT systems and installing software

- Reduce the barrier to entry into data analysis by improving accessibility of the Galaxy platform

OFFICIAL APPLIANCE PROVIDER FOR THE GALAXY PROJECT

BIOTEAM
Enabling Science

SLIPSTREAM GALAXY APPLIANCE

Galaxy

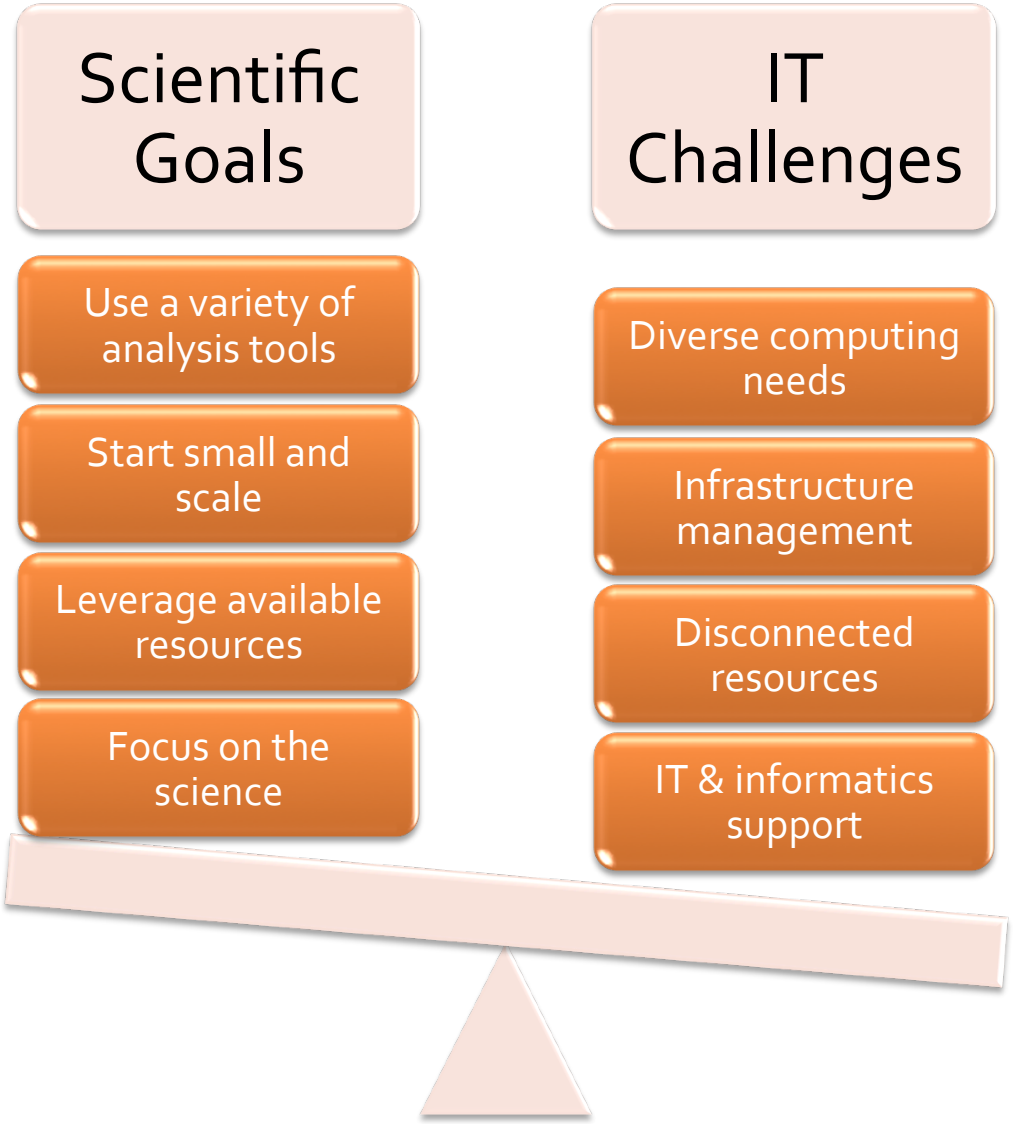
SLIPSTREAM GALAXY EDITION
APPLIANCE
Galaxy made easy.

Powerful dedicated desktop server pre-configured with a fully operational production instance of Galaxy

BIOTEAM
Enabling Science

SCIENCE VS INFRASTRUCTURE

Galaxy



HYBRID COMPUTING: FLEXIBILITY AND SCALE

The Problem

- Enable users to utilize additional resources available to them beyond those in SlipStream Galaxy Appliance
  - Local resources
  - Cloud resources

HYBRID COMPUTING: FLEXIBILITY AND SCALE

The Goal

- Make SlipStream Galaxy a central gateway to additional resources
- Keep things simple

What We Did (Example 1)

- Customer wants to leverage existing SGE environment and resources
- Jobs should spill over once the appliance is "busy"
- **Solution:** Cross-mount storage and implement transfer queue

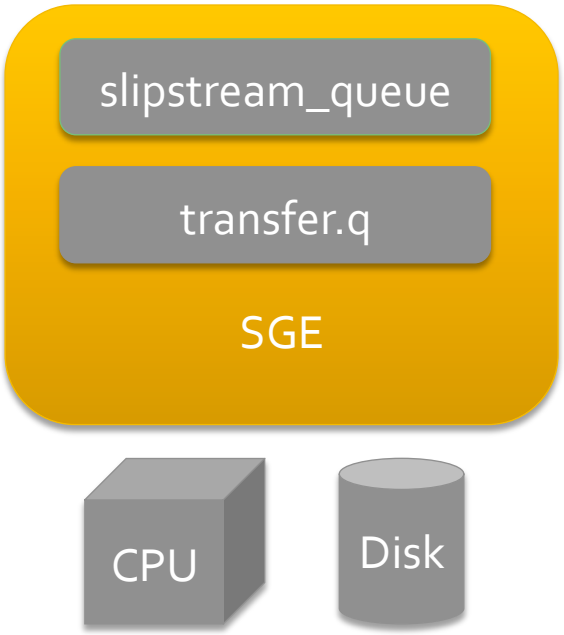
BIOTEAM
Enabling Science

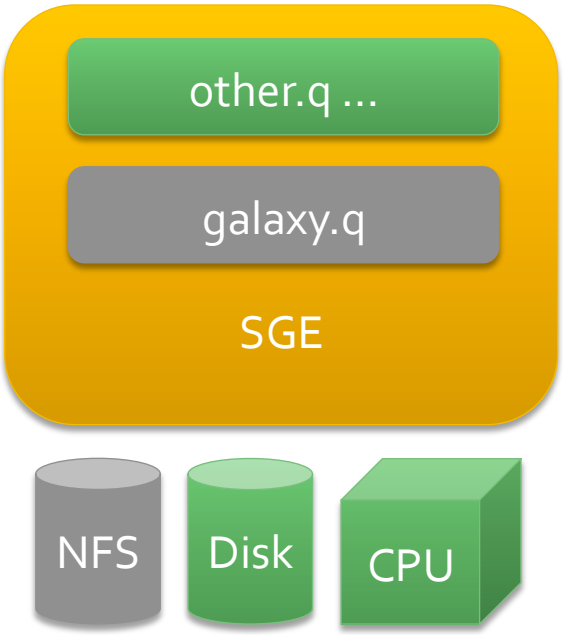
HYBRID COMPUTING: FLEXIBILITY AND SCALE

Galaxy

Customer SlipStream Appliance

Customer "Rocks" Cluster





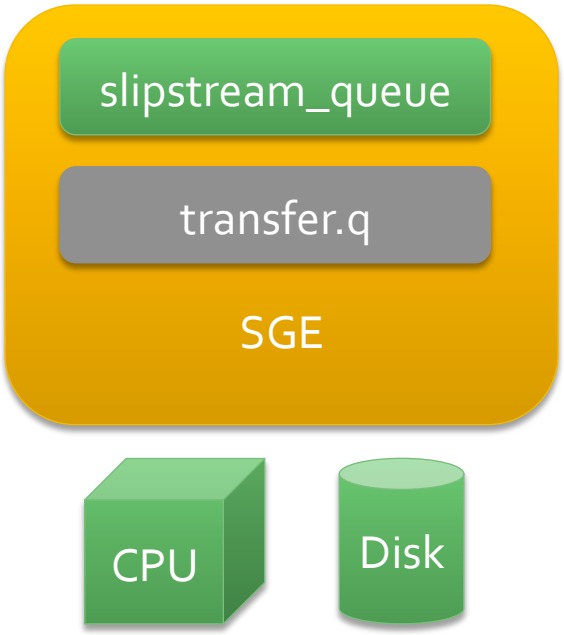
BIOTEAM
Enabling Science

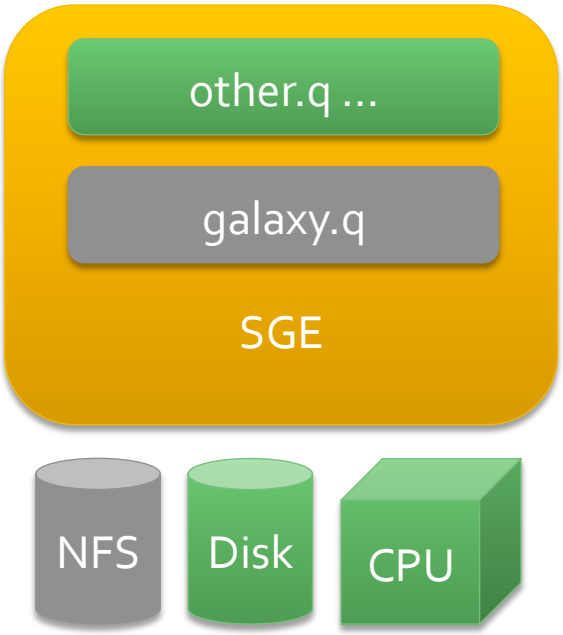
HYBRID COMPUTING: FLEXIBILITY AND SCALE

Galaxy

Customer SlipStream Appliance

Customer "Rocks" Cluster





Life is Good…

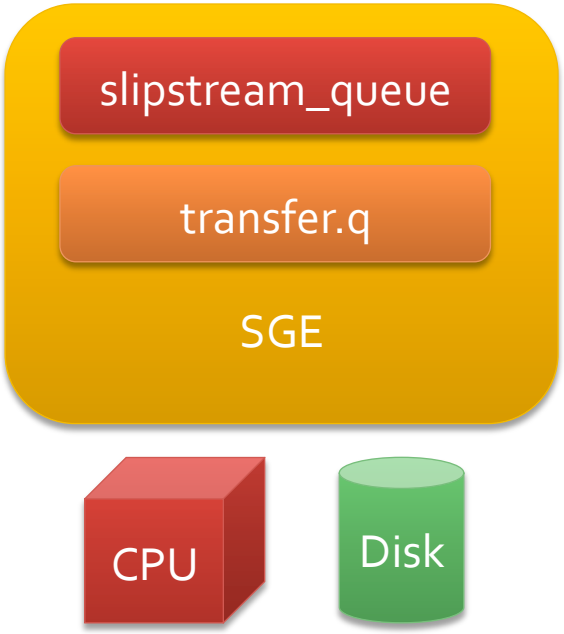
BIOTEAM
Enabling Science

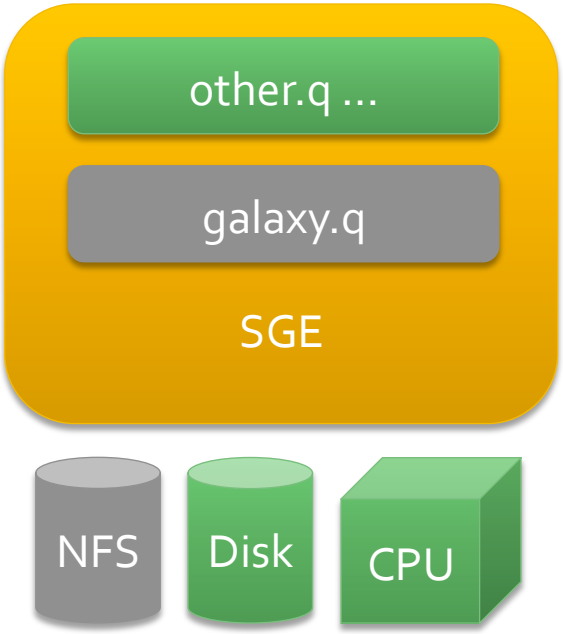
HYBRID COMPUTING: FLEXIBILITY AND SCALE

Galaxy

Customer SlipStream Appliance

Customer "Rocks" Cluster





Need more resources!

BIOTEAM
Enabling Science

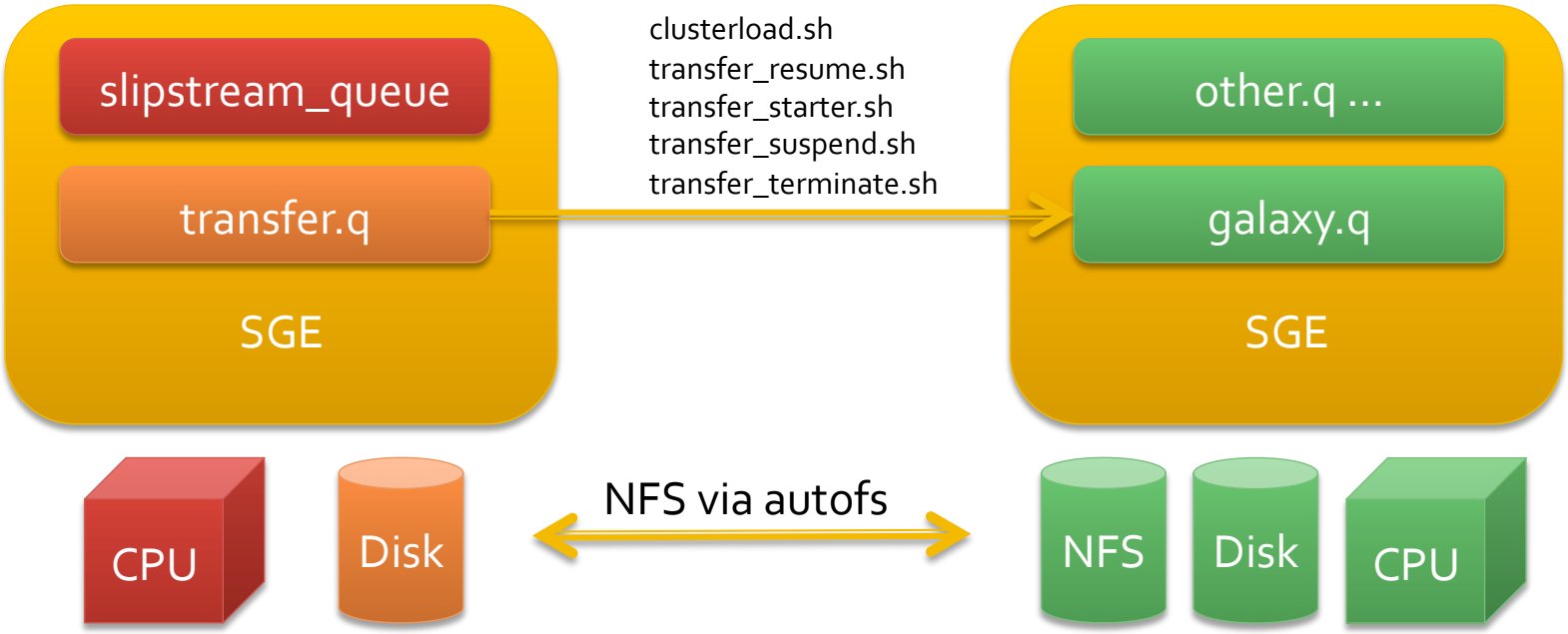
HYBRID COMPUTING: FLEXIBILITY AND SCALE

Galaxy

Customer SlipStream Appliance

Customer "Rocks" Cluster

Hooks...
clusterload.sh
transfer_resume.sh
transfer_starter.sh
transfer_suspend.sh
transfer_terminate.sh



Load sensor trips, jobs start to transfer to the cluster

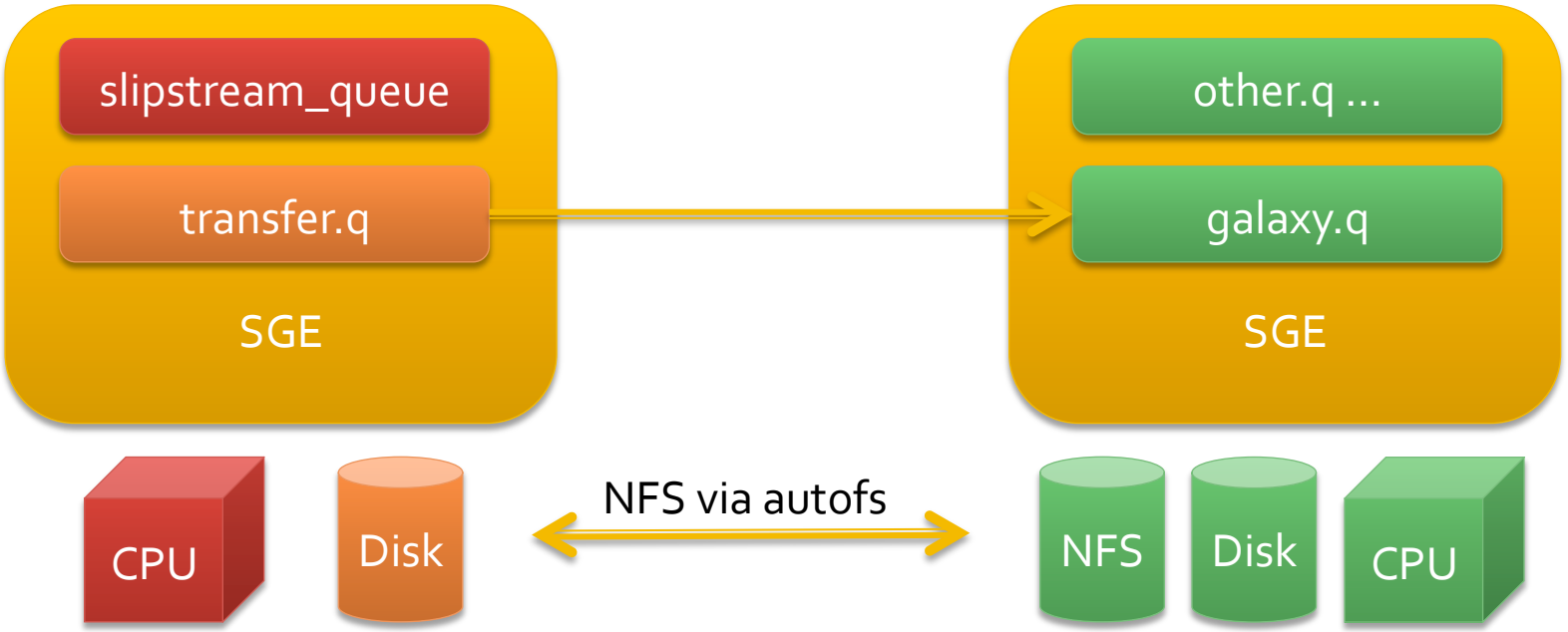
BIOTEAM
Enabling Science

HYBRID COMPUTING: FLEXIBILITY AND SCALE

Galaxy

Customer SlipStream Appliance

Customer "Rocks" Cluster



Life is better!

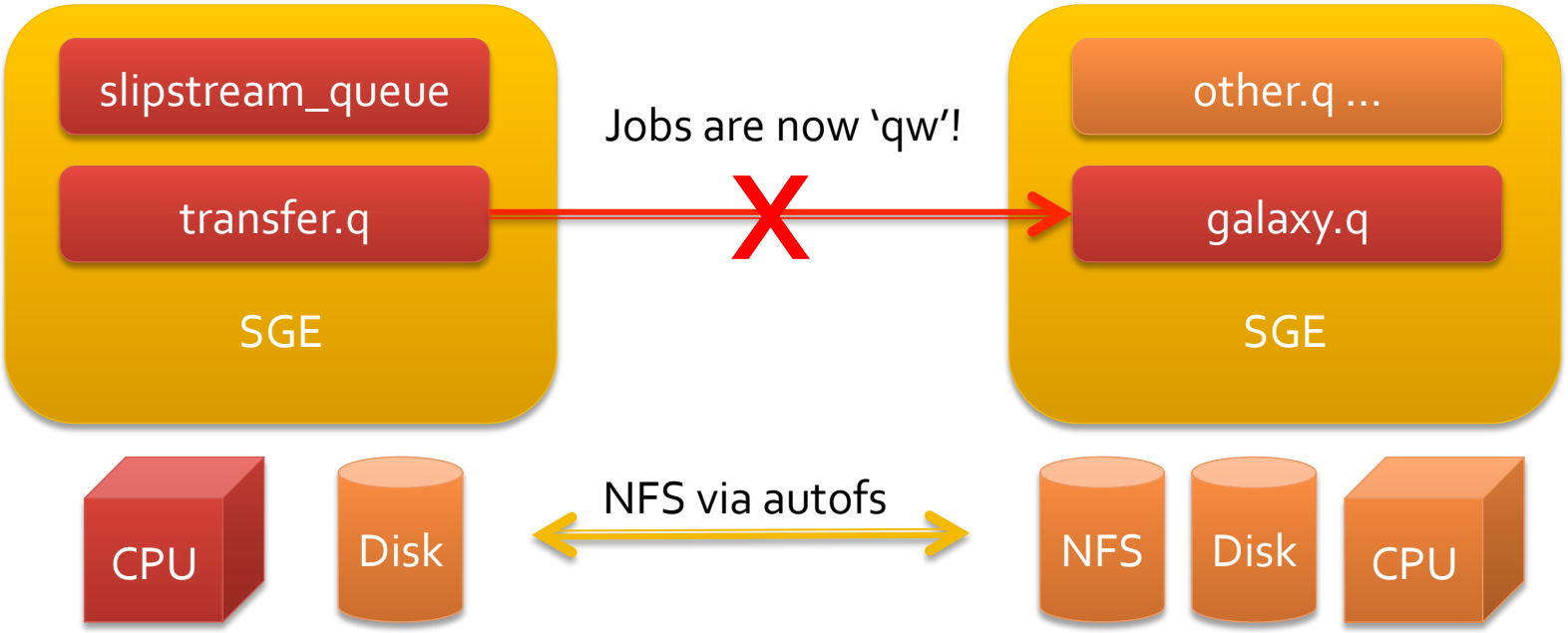
BIOTEAM
Enabling Science

HYBRID COMPUTING: FLEXIBILITY AND SCALE

Galaxy

Customer SlipStream Appliance

Customer "Rocks" Cluster



Maximum number of jobs transferred reached… need more resources? **Now what?**

BIOTEAM
Enabling Science

HYBRID COMPUTING: FLEXIBILITY AND SCALE

Galaxy

HYBRID COMPUTING: FLEXIBILITY AND SCALE

BIOTEAM Enabling Science

Galaxy

Customer SlipStream Appliance

amazon web services™ &

StarCluster

Modified StarCluster AMI

Hooks...
clusterload.sh
transfer_resume.sh
transfer_starter.sh
transfer_suspend.sh
transfer_terminate.sh

slipstream_queue
transfer.q
SGE

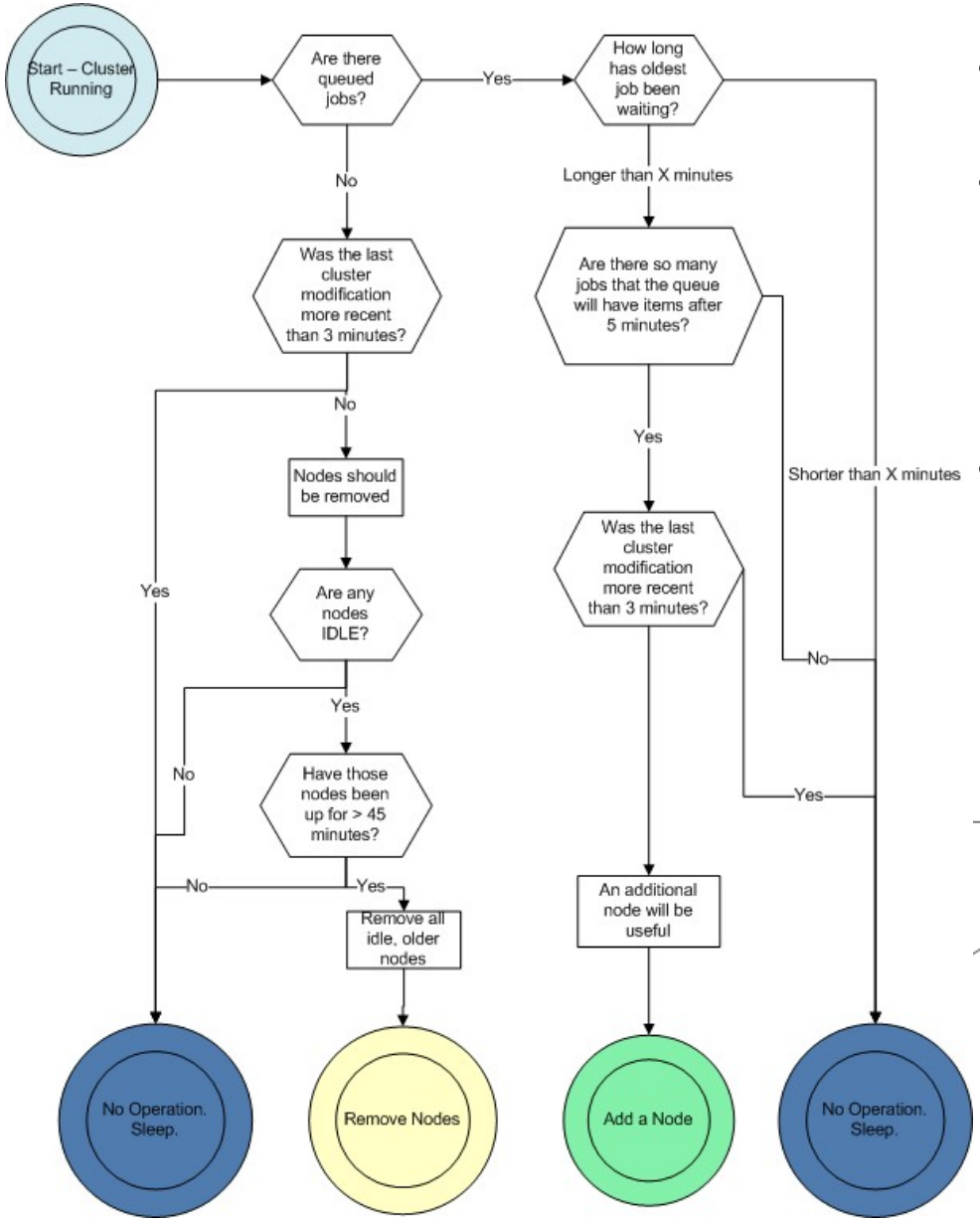
other.q ...
galaxy.q
SGE

CPU
Disk

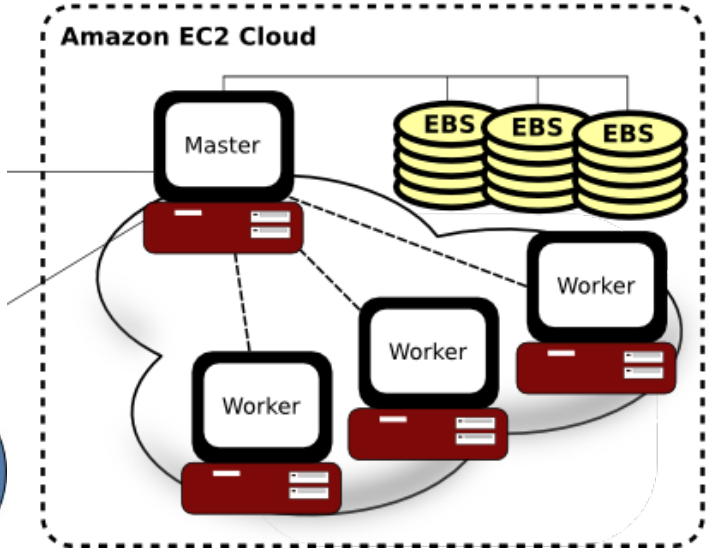
NFS sync

NFS
EBS
CPU

HYBRID COMPUTING: FLEXIBILITY AND SCALE



- Star Cluster has a powerful decision engine
- Leave the head node running and StarCluster will automatically spin up and spin down workers depending on level of "bursting" happening.
- Configuration management with transfer queue scripts to provision head node automatically…



The Challenges with Bursting

- **Users want to run jobs from the CLI as well as Galaxy that use external resources**
- Tool compatibility in heterogeneous environments (Rocks cluster is RHEL, Appliance is Ubuntu)
- Share storage between resources
- Again, keep things simple

How We Solved Them

- **Users want to run jobs from the CLI as well as Galaxy that can use external resources**

- By using a transfer queue, users interact with SGE in a familiar way from the CLI as well as through Galaxy

How We Solved Them

- **Tool compatibility in heterogeneous environments (Rocks cluster is RHEL, Appliance is Ubuntu)**

- Some toolshed tools with precompiled binaries are designed to be compatible between RHEL/ CentOS & Ubuntu
- Others must be built carefully
- StarCluster is Ubuntu-based so it is easier to maintain tool compatibility

How We Solved Them

- **Share storage between resources**
- NFS solution was the customer's preferred starting point
- Delivering NFS in cloud could benefit using asynchronous caching (ex. Avere)

How We Solved Them

- **Again, keep things simple.**
- Using existing SGE environment, NFS, etc. minimizes integration effort
- Customer's interaction with appliance from the CLI and Galaxy doesn't change

HYBRID COMPUTING: FLEXIBILITY AND SCALE

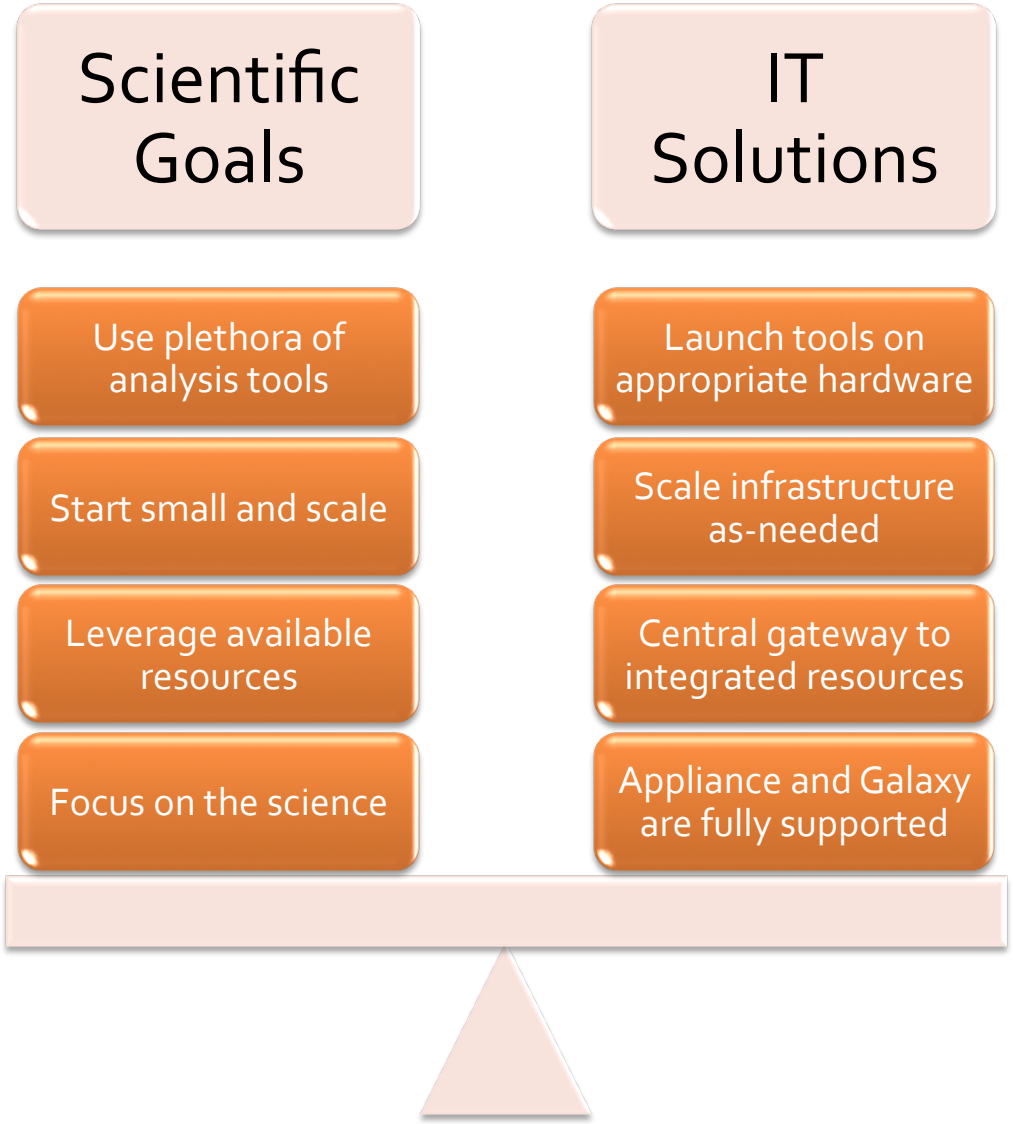
Good start… but how do we expand this concept in the future?

- Light Weight Runner to increase abstraction for resource aware scheduling
- Docker/LXC to provide isolation and portability of tools
- Apache Mesos for resource aware meta-scheduling

BIOTEAM
Enabling Science

SCIENCE VS INFRASTRUCTURE

Galaxy



LOOKING AHEAD

New Reference Design

Collaboration with SGI and Intel to provide an even more powerful, affordable appliance

Enhanced Galaxy Support

Partnership with BioStar Genomics to develop additional Galaxy support and service offerings

Scalable Infrastructure

Continue to build infrastructure integrations that dynamically fit scientific computing needs

CONCLUSION

SlipStream Appliance: Galaxy Edition

A high performance solution for data analysis

Why SlipStream Galaxy

- 10+ years of Life Science IT expertise
- Dedicated, flexible, scalable resource
- Infrastructure and Galaxy administration support

BIOTEAM
Enabling Science

THANK YOU!

Galaxy

SLIPSTREAM GALAXY EDITION
APPLIANCE
Galaxy made easy.

ONGOING EARLY ACCESS PROGRAM
(Limited Availability)

Visit the BioTeam booth for more information!!

www.bioteam.net/slipstream/galaxy-edition