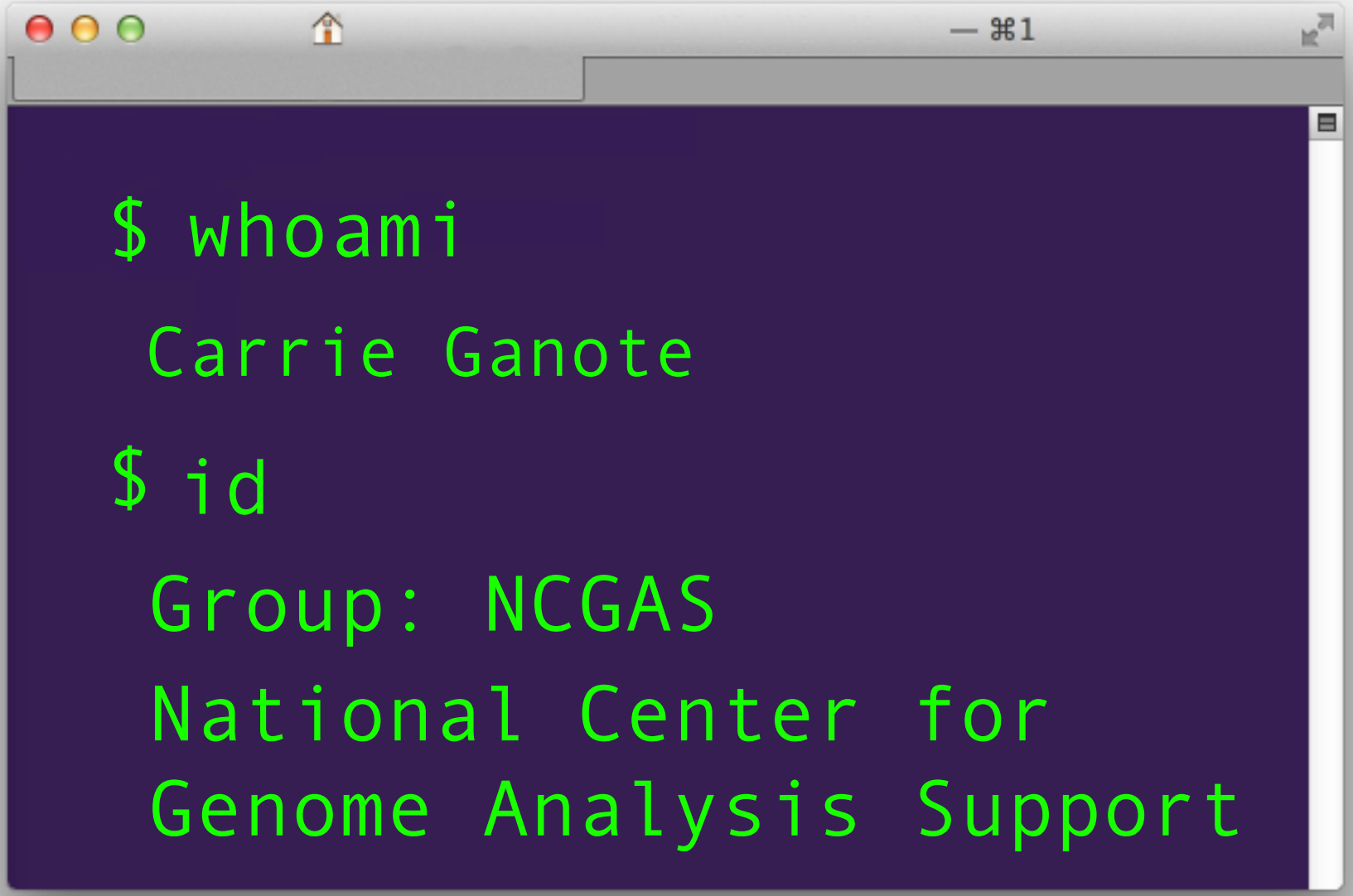
NATIONAL CENTER FOR
GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY



NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Galaxy Deployment on Heterogeneous Hardware

- Who we are and our computing environment
- How we set up Galaxy on multiple resources
- *Trials and tribulations we have encountered*
- *Directions moving forward*

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Who are these people

NCGAS is an NSF-funded organization with a focus on providing support and infrastructure for NSF-funded genomics projects (for free!).

We specialize in genome and RNA-Seq analysis for novel organisms, UNIX hand-holding, software installation, and advice from library prep to downstream software decisions

Based in Indiana University, we also support the IU community and install/ maintain biology software on the clusters (including Galaxy).

NCGAS is:

Yours Truly – Bioinformatics Analyst

Le-Shin Wu - Resident Computer Science Expert

Thomas Doak – Local Ciliate Biologist and Biology Liason

Rich LeDuc – Cat Wrangler, Statistician and Proteomics Guy

Bill Barnett – Director and man with the plan

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Mason
HP DL580 - Large memory machine for genomics, XSEDE Tier 2 resource

Meet the family

Big Red 2: Cray XE6/XK7 PetaFLOPS Supercomputer owned by IU

CRAY
THE SUPERCOMPUTER COMPANY

Quarry
IBM BladeCenter HS21 - General purpose Linux computing environment

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Meet the family

Extended

Distributed network of opportunistically available compute resources for US researchers

XSEDE

Extreme Science and Engineering Discovery Environment

Integrated collection of high-end compute resources for US researchers

Community of collaborative partners advancing biomedical research

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Our Galaxies

Penguin-on-Demand Galaxy

IU Galaxy

Hahn Lab Galaxy

Mockaitis Lab Galaxy

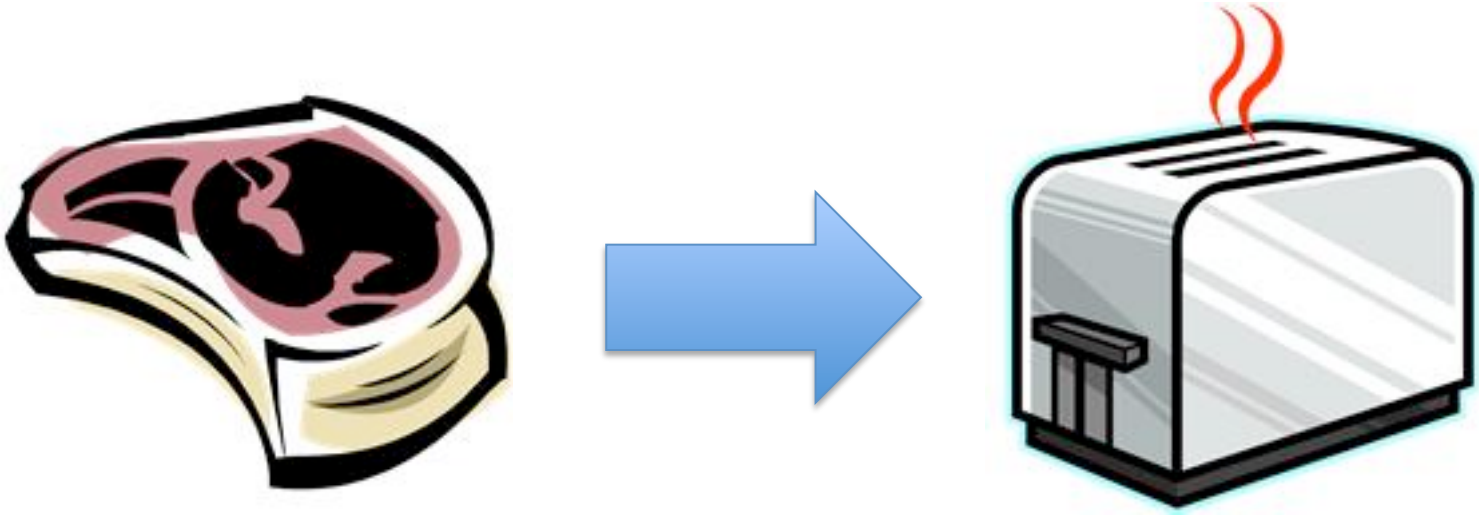
Test/Dev Galaxy

Trinity Galaxy

NCGAS Galaxy

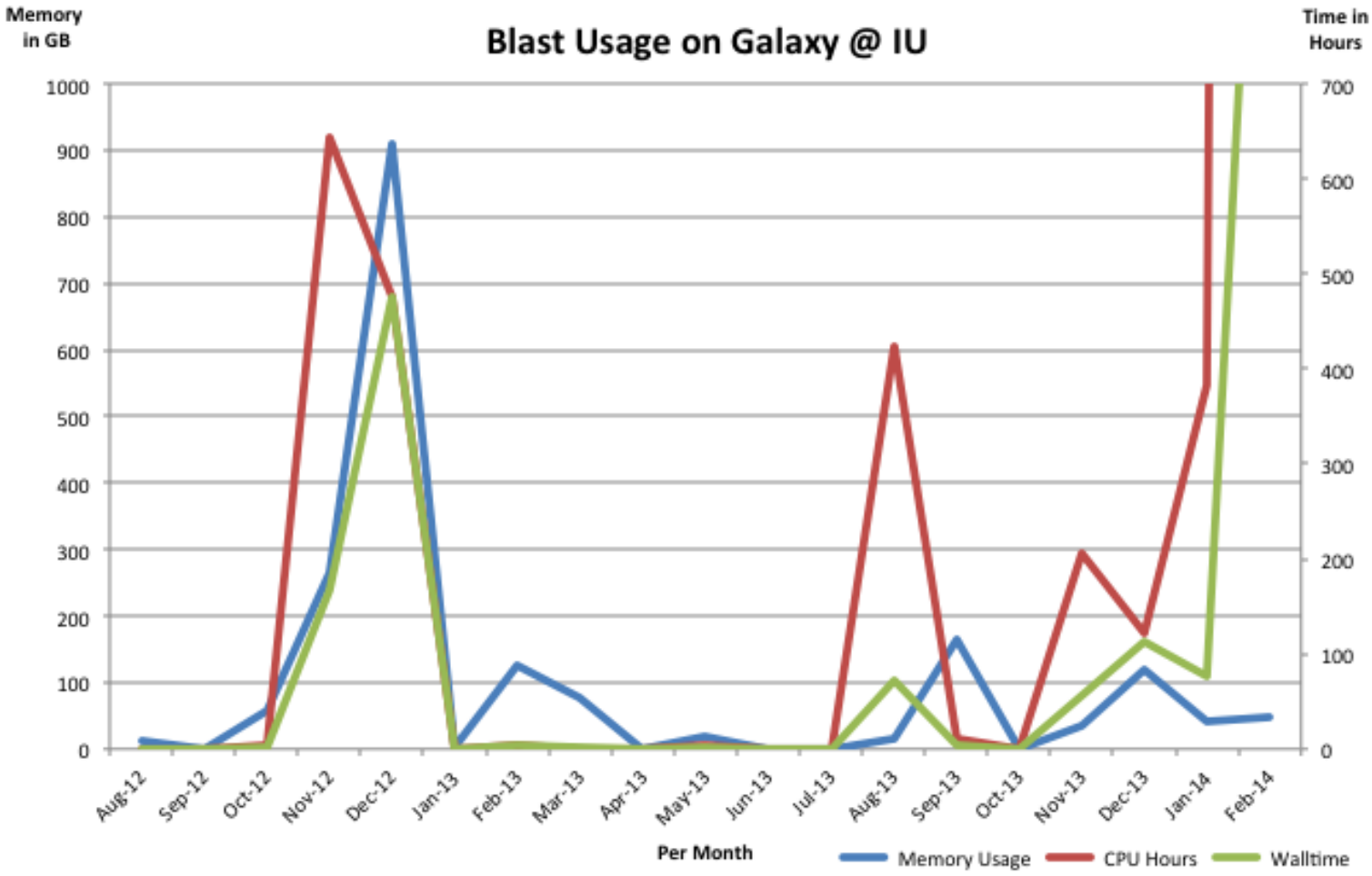
Photo of M109 spiral galaxy courtesy of NOAO/AURA/NSF

Why bother?



Just because an instrument WILL work doesn't mean it's the best suited for the job!

Too many BLAST+ jobs are landing on our big memory machine!



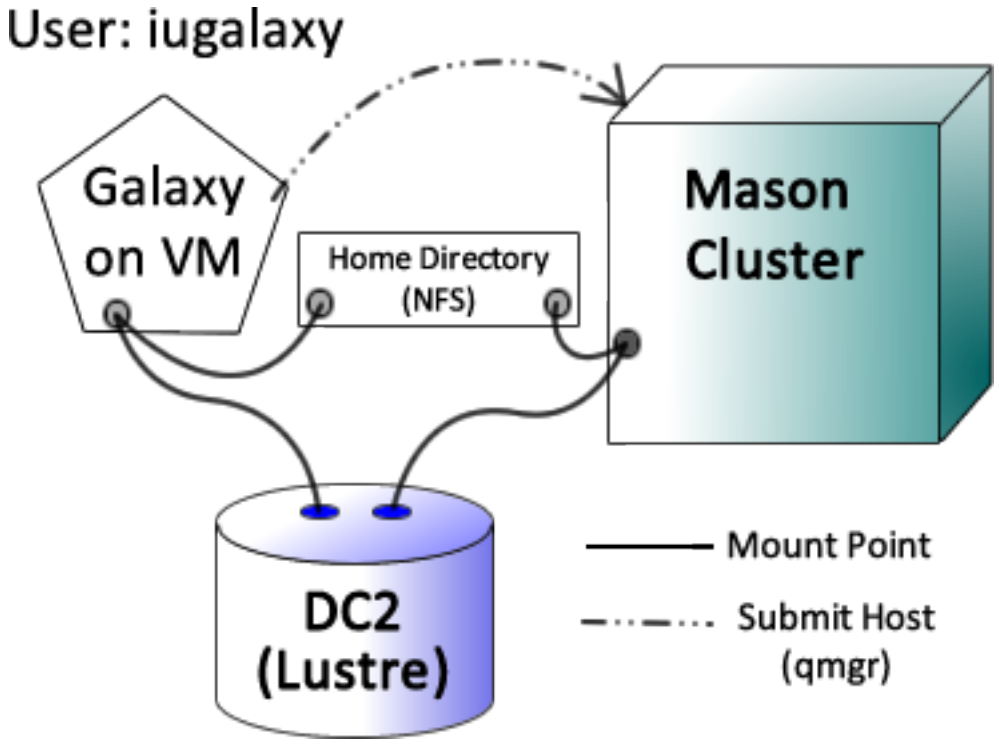
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

- Who we are and our computing environment
- How we set up Galaxy on multiple resources
- *Trials and tribulations we have encountered*
- *Directions moving forward*

WHERE TO START?

Starting Simple

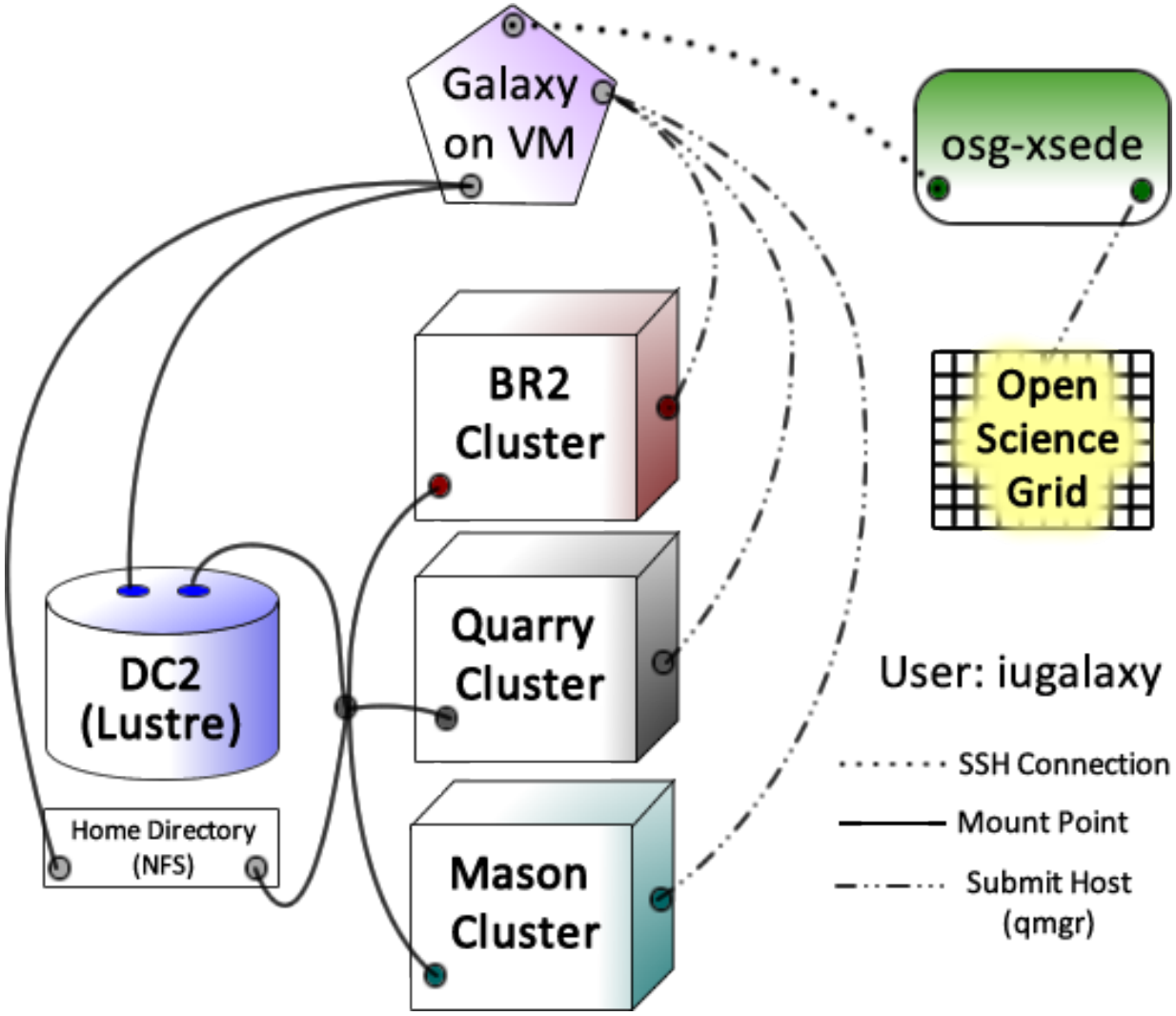


Here's how we got started:

- Galaxy is installed on a VM
- VM has access to shared filesystem with Mason
- VM is submit host to Mason
- Everything configured in Universe_wsgi.ini
- Everything runs as one Galaxy process

Torque

Less Simple

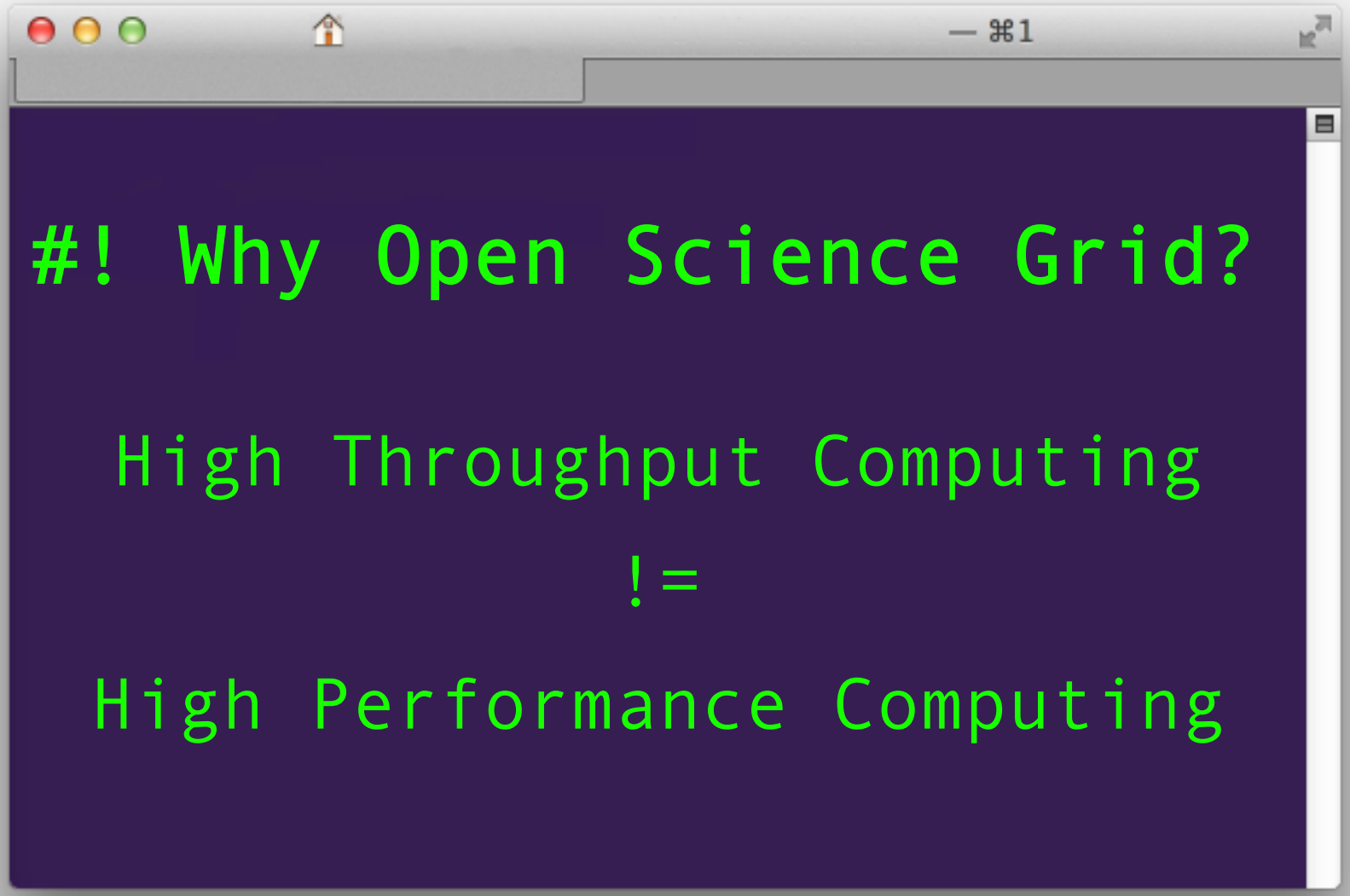


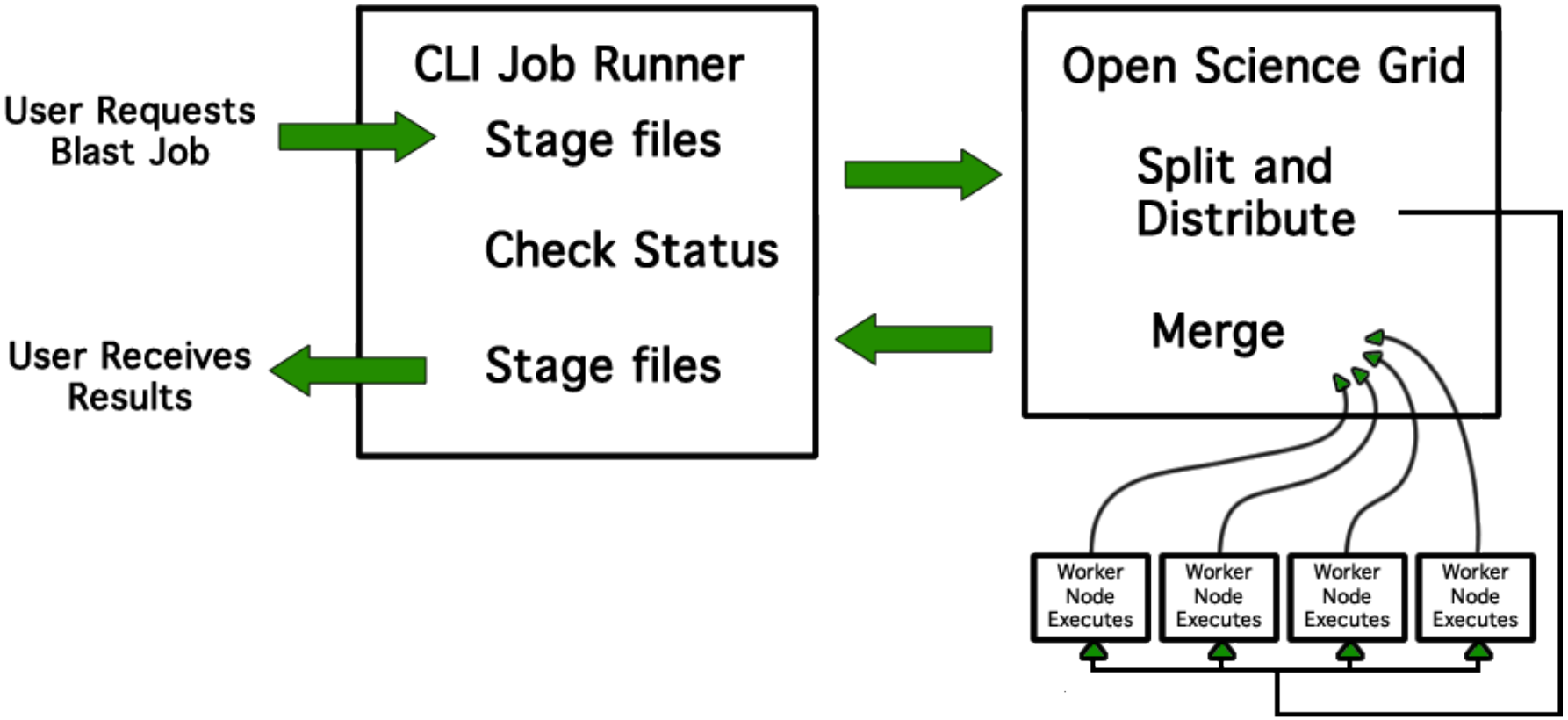
HTCondor High Throughput Computing

Torque

NATIONAL CENTER FOR
GENOME ANALYSIS SUPPORT

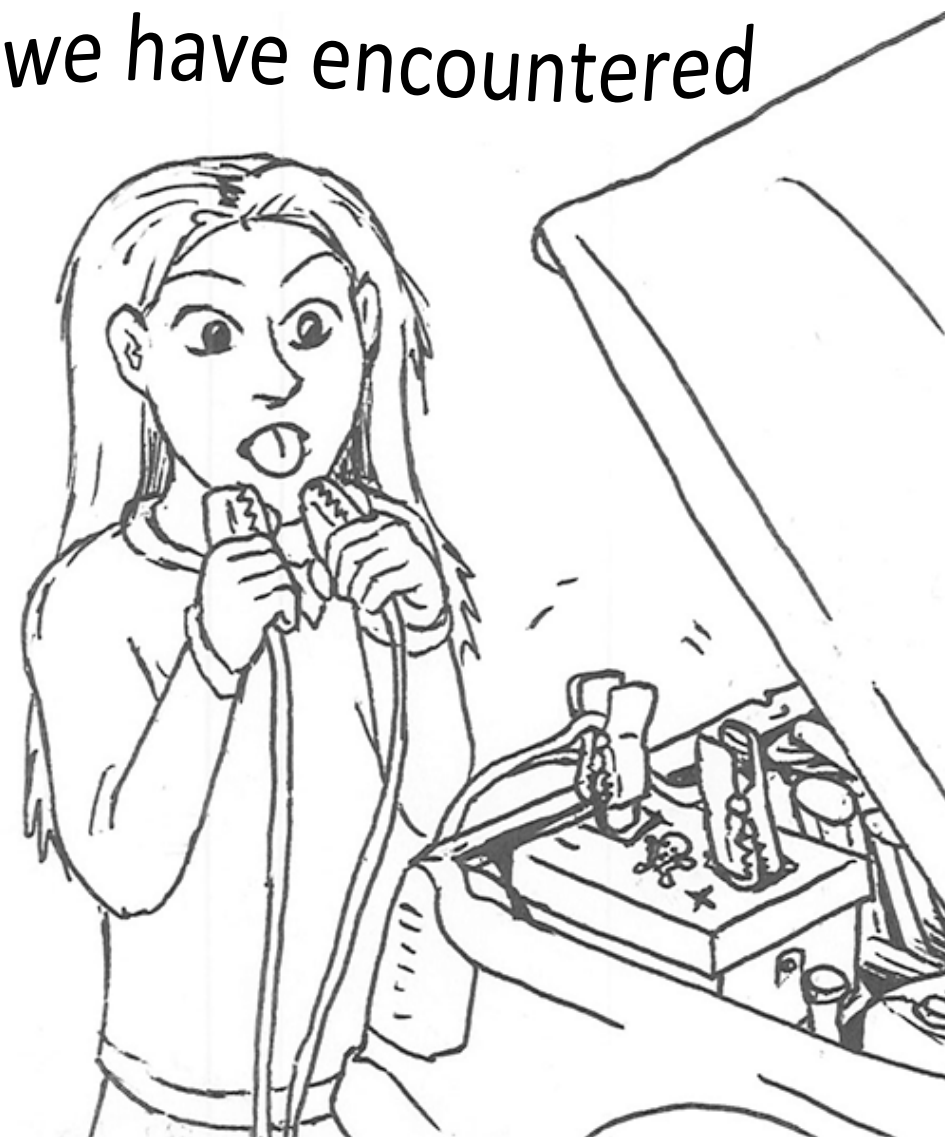
INDIANA UNIVERSITY



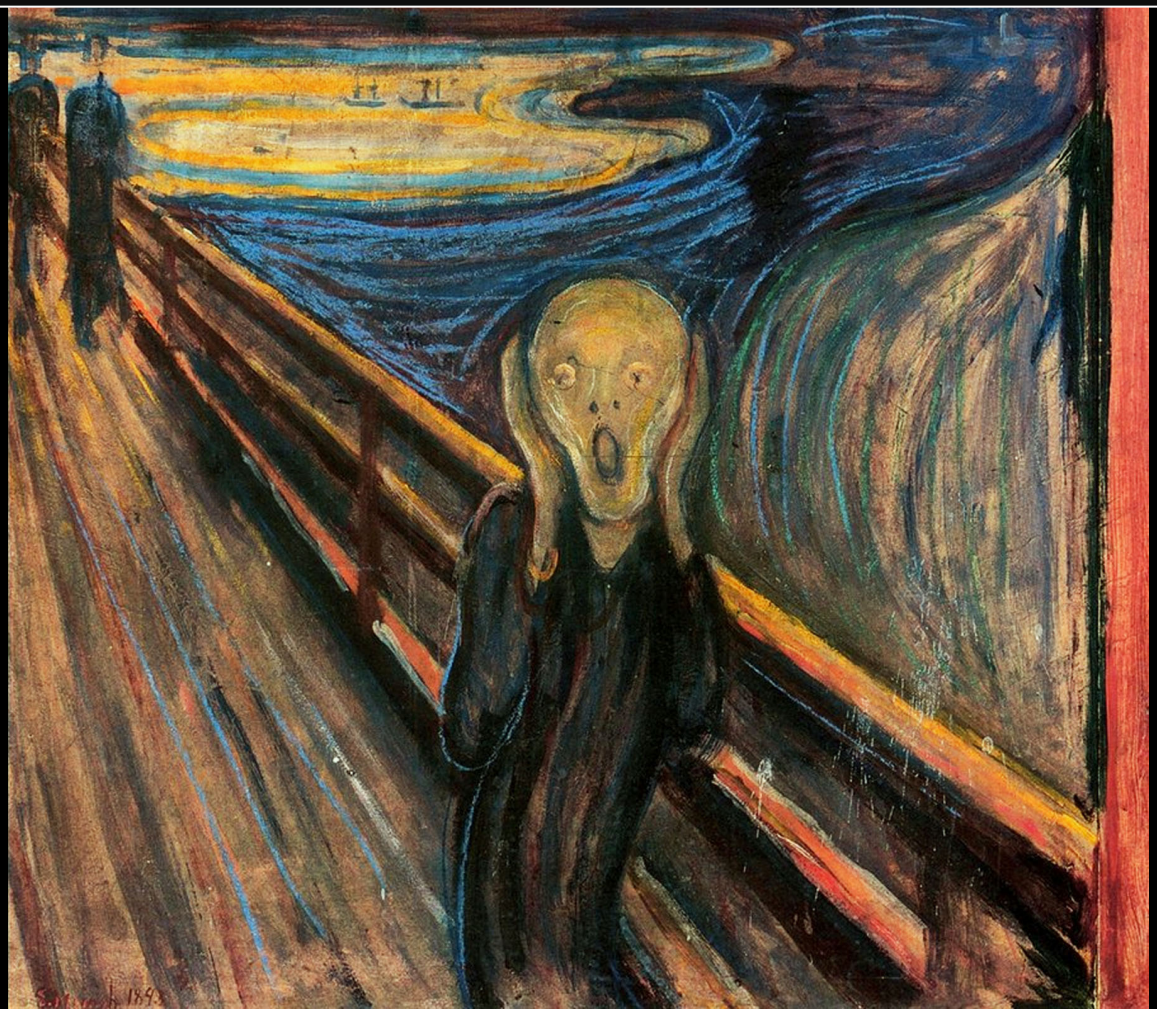

User Requests
Blast Job
CLI Job Runner
Stage files
Check Status
User Receives
Results
Stage files
Open Science Grid
Split and
Distribute
Merge
Worker
Node
Executes
Worker
Node
Executes
Worker
Node
Executes
Worker
Node
Executes

- *Trials and tribulations we have encountered*

What could possibly go wrong?



InternalException: code 1: (qsub) cannot access script file: Unauthorized Request MSG=can not authorize request (0-Success)



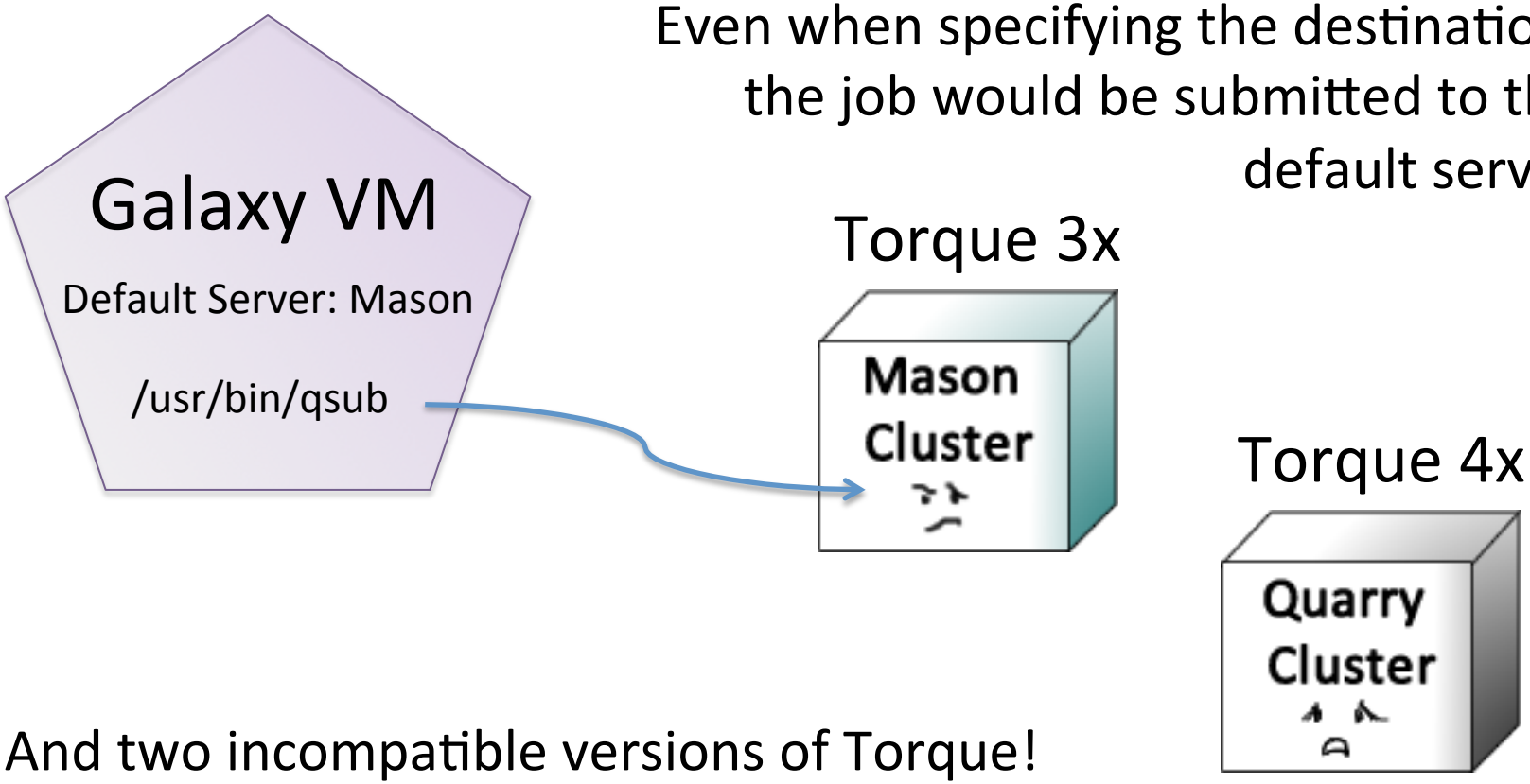
Edvard Munch, "The Scream", Public domain in the USA

This can happen because:

- There was a change to the PBS configuration on the server, but pbs_server was not restarted. Pause scheduling and restart pbs_server.
- You are trying to use DRMAA but Torque was not compiled with the - - enable_drmaa flag. Beg sys admins to recompile Torque.
- Library defines pbs_submit instead of submit_pbs_hash(). Install new pbs-drmaa library. (Thanks Oleksandr Moskalenko)
- For whatever reason, the hostname might be truncated, requiring two entries:
  set server submit_hosts += gw14.iu.xsede.org
  set server submit_hosts += gw14
- Faulty alignment of planets – wait a few days!

Issues Connecting to Quarry

Even when specifying the destination, the job would be submitted to the default server



And two incompatible versions of Torque!

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

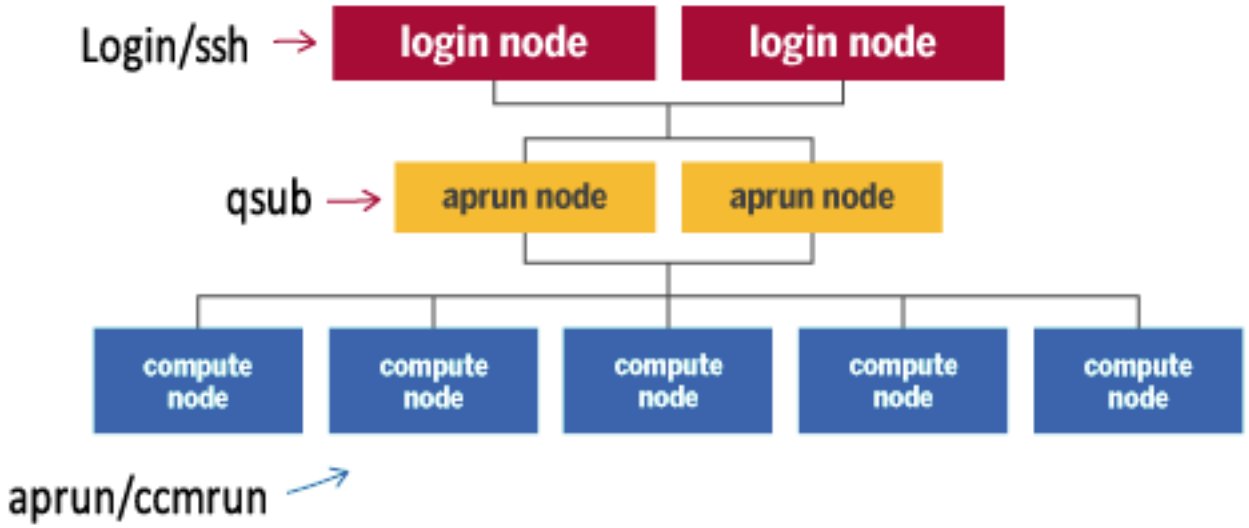
This is now fixed in Galaxy!

Now we can define destinations in the job configuration file.

Solution:

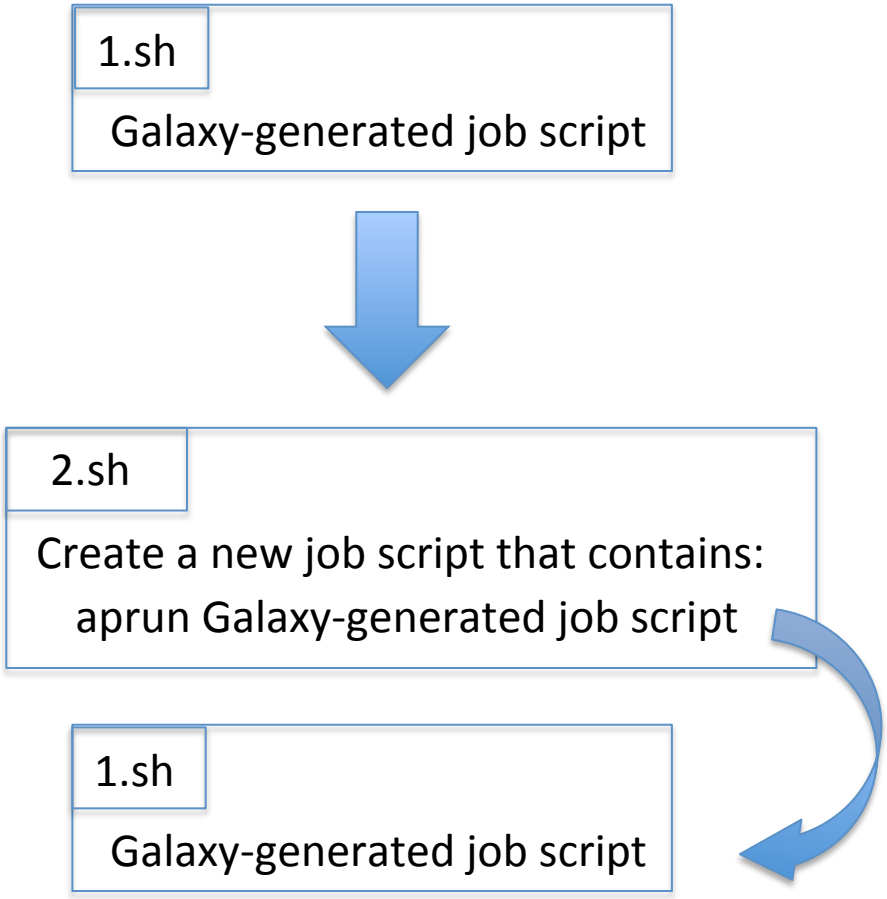
Dedicate a handler to Quarry,
hardcode path and destination for that handler only

Issues Dealing with Cray Architecture



Commands must start with “aprun” or “ccmrun” in order to be run on the compute nodes!

Solution:



Create a wrapper script for the Galaxy-generated shell script that will be submitted via qsub.

Working on a better solution for this!

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

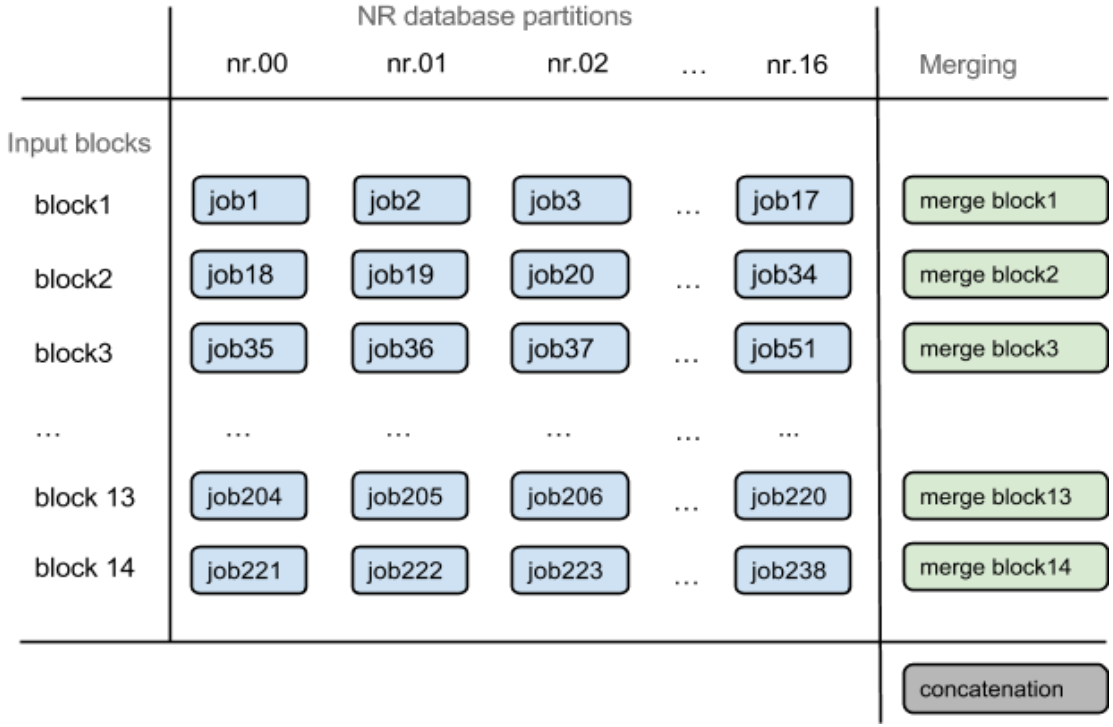
INDIANA UNIVERSITY

Eccentricities of Blast on the Open Science Grid

- Blast Databases exceed the size limits of many resources (NR is ~12GB compressed)
- Some resources may not be able to run the job (even if they say they can)
- Type of query/number of queries per worker can greatly affect performance

Solution:

- Splitting Database as well as Query is necessary, but this complicates e-value calculation which depends on the database size
- Disallow the same worker to pick up a job that it failed X times to complete
- Run a small number of queries first to attempt to calibrate number of queries per worker node

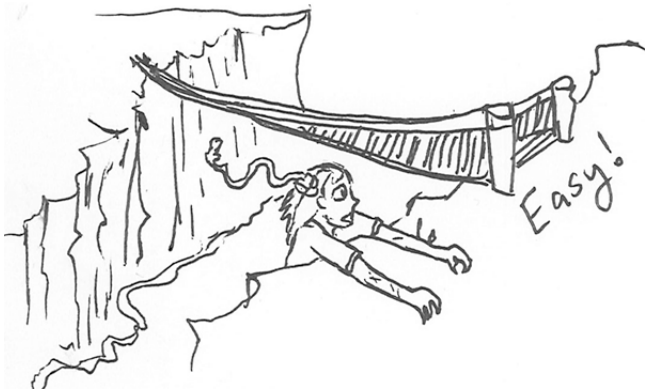


The Future

Intergalactic Communication

If Galaxies could talk, we could stand up one for each resource and specialize

OSG through LWR instead of CLI



NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Thanks for watching!

Special thanks to:

The NSF for our grant: #1062432

Soichi Hayashi with the Open Science Grid for our collaboration on Blast on the OSG

Brian Haas for collaboration on the Trinity RNA-Seq project

Of course, the Galaxy Team

Please visit us at ncgas.org
Contact NCGAS at help@ncgas.org
Contact me at cganote@iu.edu

BROAD logo used with permission