Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

OSG Submission Tutorial

Ken Herner

2016 FIFE Workshop

21 June 2016

FABRIC FOR FRONTIER EXPERIMENTS

June 21, 1893: First Ferris Wheel

Introduction

- For this talk, I will assume you're familiar with jobsub_client, dCache, IFDH, job submission in general, CVMFS, and the existence of the Open Science Grid.
  - If you are not, please refer to the respective talks in the workshop and the FIFE Redmine material
- So what will we talk about today?
  - How to actually modify your submission to run outside of Fermilab
  - How to maximize your chances of a successful job with well-written scripts

How to submit to the OSG

- If you're already submitting to GPGrid it's quite easy
  - make sure OFFSITE is in your --resource-provides=usage_model option
- For maximum available resources you can mix onsite and offsite options
  - e.g. DEDICATED,OPPORTUNISTIC,OFFSITE
- Possible to steer to specific site(s) with --site option
- Since this is so easy, what's left to talk about?
  - Majority of problems we see relate to scripts not being sufficiently robust to operate outside of GPGrid

What are the OSG gotchas?

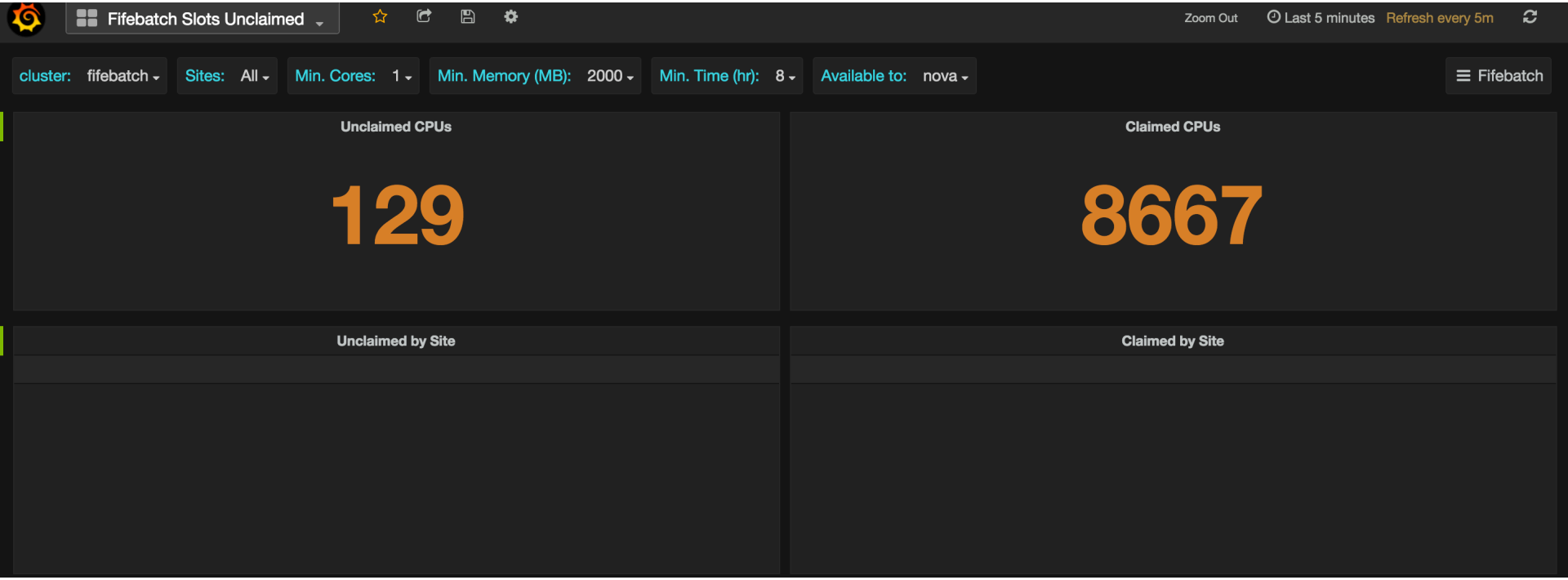
- Main difference is pre-emption
  - If other jobs with higher priority come into the cluster, your job may be evicted in favor of those.
    - Risk increases with longer-running jobs.
  - Job will automatically go back in the queue and restart (from the beginning) eventually. The stdout/stderr is lost though.
  - Condor log may say “job was evicted”, but might also show up as a disconnection. We have no control over that.
- **BlueArc areas are not mounted** (but you shouldn’t really run code directly from there anyway.)
- Some system libraries (e.g. X11) may not be installed
- We are seeing linux kernel versions >= 3 in the wild; can be a problem for UPS products and/or SRT builds if workarounds are not in place.

Other things to note on the OSG

- Some sites only support certain experiments
- This is all opportunistic usage. **We have no control whatsoever over your job's priority at a remote site**
- Offsite priorities are not adjusted between experiments. First come, first served is the order of the day
  - That is, between jobs that make equal resource requests
- Your job's resource request (CPU, disk, memory, run time) is the single biggest driver of your wait time
  - People who submit after you may run first if their resource requests are smaller; depends on what's available at a site
- We may not have any glideins running at a given site all the time. Glideins can take several hours (or longer) to start. Best to have jobs in the queue all the time...

What's available at a site?

- FIFEMON has a good tool for this
- https://fifemon.fnal.gov/monitor/dashboard/db/fifebatch-slots-unclaimed



Picking sites before submitting

- So does that mean I should look at FIFEMON and tailor my submission to point to sites that have free resources?

Picking sites before submitting

- So does that mean I should look at FIFEMON and tailor my submission to point to sites that have free resources?
  - NO.

Picking sites before submitting

- So does that mean I should look at FIFEMON and tailor my submission to point to sites that have free resources?
  - NO.
  - Absolutely not.

Picking sites before submitting

- So does that mean I should look at FIFEMON and tailor my submission to point to sites that have free resources?
  - NO.
  - Absolutely not.
  - That will not buy you anything. Do not do this.

Picking sites before submitting

- So does that mean I should look at FIFEMON and tailor my submission to point to sites that have free resources?
  - NO.
  - Absolutely not.
  - That will not buy you anything. Do not do this.
  - Seriously, do not do it.

Picking sites before submitting

- So does that mean I should look at FIFEMON and tailor my submission to point to sites that have free resources?
  - NO.
  - Absolutely not.
  - That will not buy you anything. Do not do this.
  - Seriously, do not do it.
- Unless you need to run a certain job at a certain site, you are **much** better off submitting to all sites and letting the system match you. Things change too rapidly to keep up.

Picking sites before submitting

- So does that mean I should look at FIFEMON and tailor my submission to point to sites that have free resources?
  - NO.
  - Absolutely not.
  - That will not buy you anything. Do not do this.
  - Seriously, do not do it.
- Unless you need to run a certain job at a certain site, you are **much** better off submitting to all sites and letting the system match you. Things change too rapidly to keep up.
- What WILL help you? Profiling your jobs and requesting only the resources that you need. Don't just stick with the default because it will work for you. If you only need 2 hours or 1 GB of RAM, modify the requests. It WILL make a huge difference

Other ways to maximize success

- Preemption probability increases with run time. Aim for shorter (few hour) jobs. Two three-hour jobs instead of one 6 hour job, etc. That also makes more sites available to you You can see the difference here:

https://cdcvs.fnal.gov/redmine/projects/fife/wiki/Information_about_job_submission_to_OSG_sites

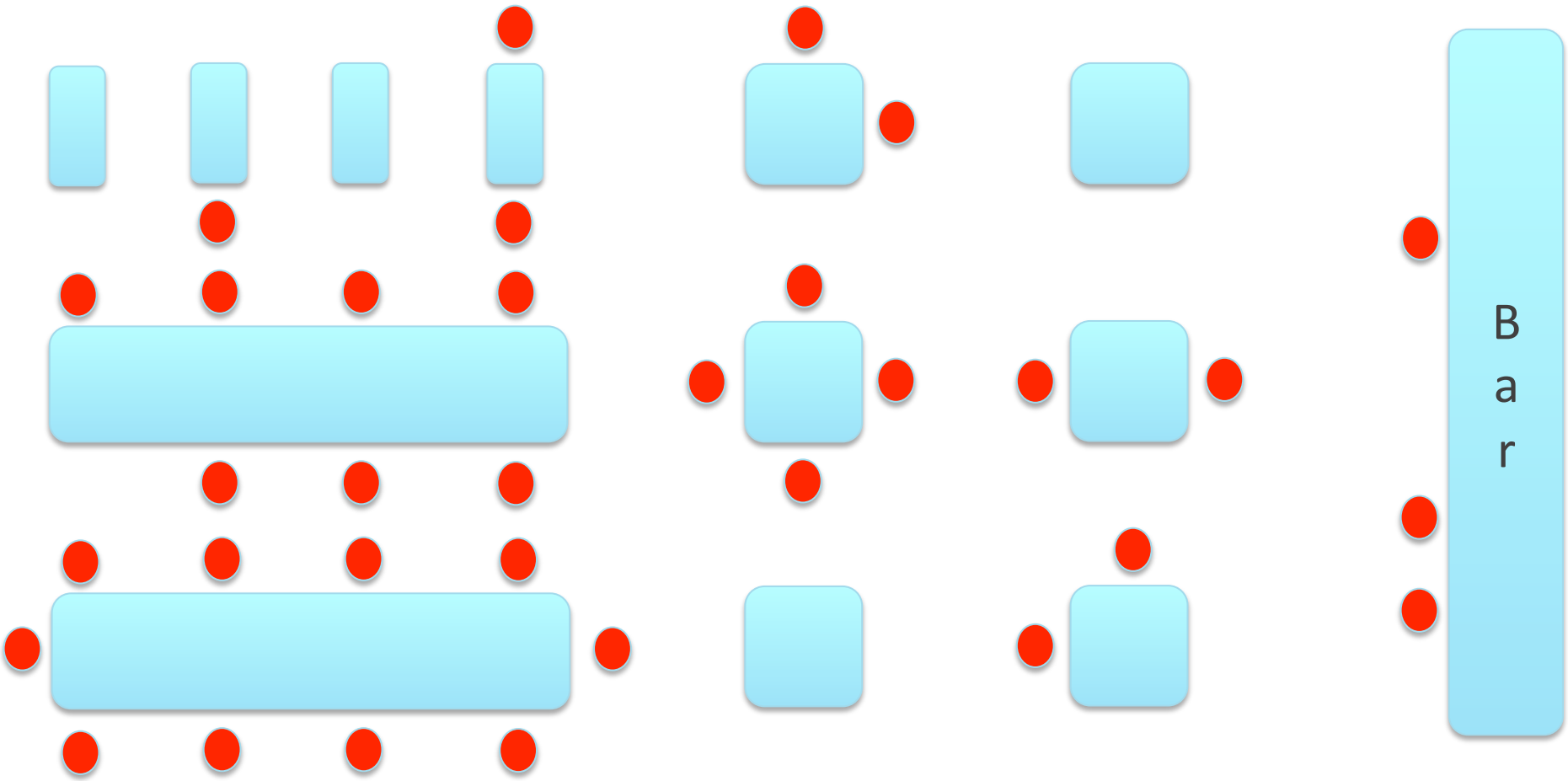
Site Name	--site option (sorted)	Supported Experiments	Known limitations	Maximum memory	Estimated maximum job lifetime
Brookhaven National Laboratory	BNL	All	jobsub defaults are OK	3000MB	24 h
Caltech T2	Caltech	All	jobsub defaults are OK	3000MB	25 h
Clemson	Clemson	All	jobsub defaults are OK	14000MB	24 h
Cornell	Cornell	All	jobsub defaults are OK	2500MB	unknown
GPGrid	FNAL	All+CDF+LSST+D0	new GPGrid accessed via FNAL Feb 2016, but job could end up on CMS T1	16000MB	96 h
FNAL CMS Tier 1	FNAL	All	jobsub defaults are OK	16000MB	24 h
Czech Academy of Sciences	FZU	NOvA only	request --disk=20000MB or less	2500MB	unknown
University of Washington	Hyak_CE	All	available resources vary widely	Tested to 7000MB	3.5 h
University of Manchester	Manchester	uboone only	disk default is OK Ask for one CPU for every 2 GB of memory requested (e.g. 2 CPUs for memory request between 2 and 4 GB.)	Tested to 8000MB	between 12-24 h
ATLAS Great Lakes Tier 2 (AGLT2)	Michigan	All	jobsub defaults are OK	2500MB	approx. 8 h
MIT:http://www.cmsaf.mit.edu/condor4web/	MIT	All + CDF	jobsub defaults are OK	2500MB	unknown
Midwest Tier2	MWT2	All	jobsub defaults are OK single core jobs will take a very long time to run if requesting more than 1920 MB of memory. Running custom 3.x linux kernels, probably need to set UPS_OVERRIDE to setup products.	Tested to 7680MB	5 h
Red/Sandhills	Nebraska	All	jobsub defaults are OK; Some slots are SL6 Docker containers on SL7 hosts	Tested to 8000MB	48 h

More sites below…

Fermilab

Shorter jobs on the OSG (from Monday)

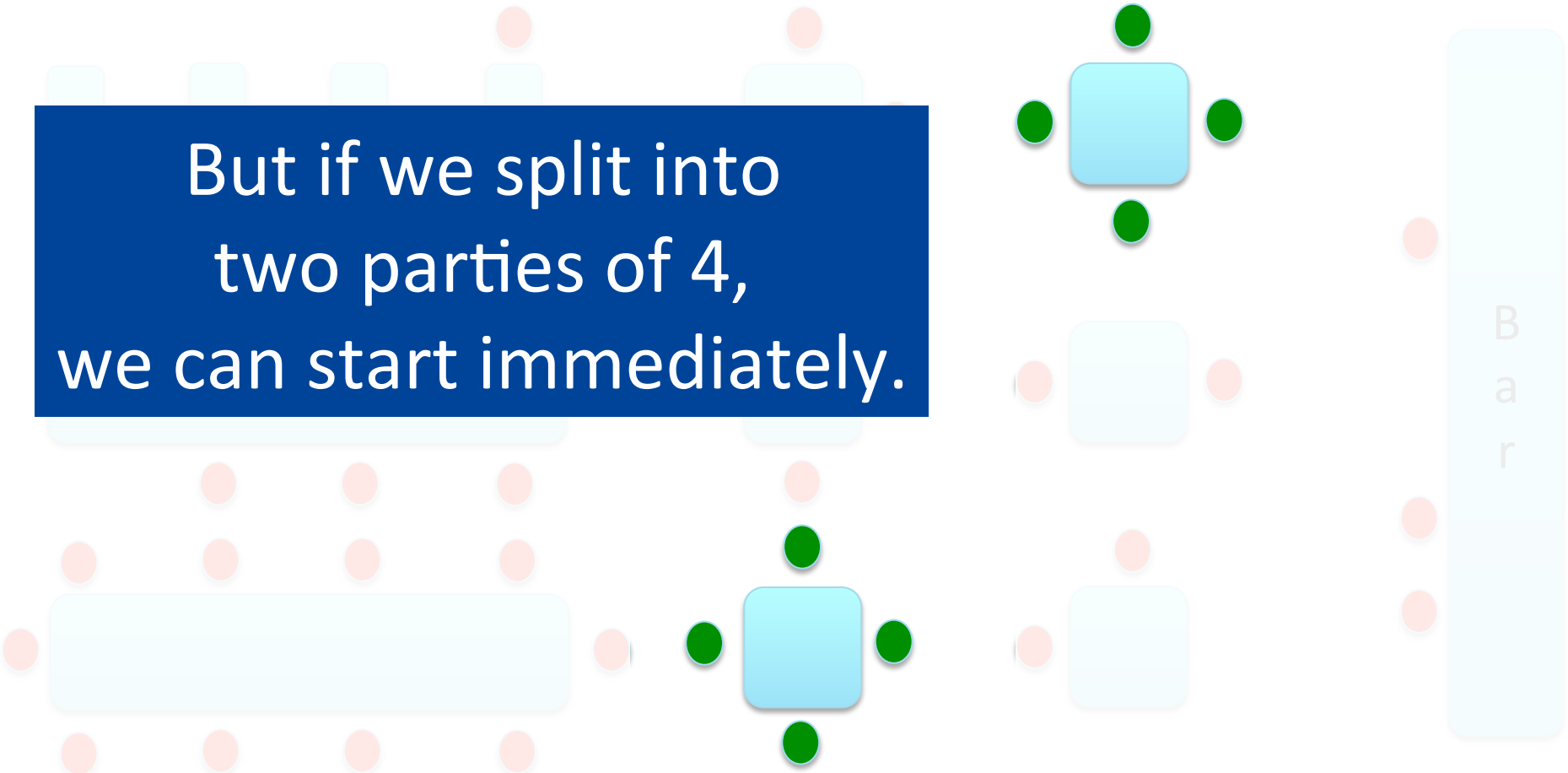
- Let's say we go as a party of 8 (they don't take reservations)
- Here, we will have to wait for the big tables to open up



Another scenario

- Let's say we go as a party of 8 (they don't take reservations)
- Here, we will have to wait for the big tables to open up

But if we split into two parties of 4, we can start immediately.



Writing robust scripts

- The best way to ensure success on the OSG is to write a good script. A good script should
  - Check for the existence of all needed CVMFS repositories before starting to run
  - Check OS environment and set env variables as needed (such as UPS_OVERRIDE)
  - Check the exit status of data transfer commands and executables
  - Avoid hard-coding any paths for input or output files (CVMFS paths are OK though) and assumptions about what username and/or home directory the job sees
- Let's look at and critique some example scripts

Job script #1

```
#!/bin/bash

. /grid/fermiapp/products/common/etc/setup
setup ifdhc
. /grid/fermiapp/products/larsoft/etc/setup
setup mycode v1_2_3_4 -q whatever

/myexperiment/app/user/${USER}/mycode/myexe myargs

ifdh cp -D myoutfile1 myoutfile2 /myexperiment/data/users/$USER/
myoutputs/
echo “Done!”
```

Job script #2

```
#!/bin/bash

. /cvmfs/fermilab.opensciencegrid.org/products/common/etc/setup
setup ifdhc
. /cvmfs/fermilab.opensciencegrid.org/products/larsoft/etc/setup
setup product1
setup product2
setup mycode v1_2_3_4 -q whatever
myexe myargs
MYCONFIG=~/.myconfig
ifdh cp -D myoutfile1 myoutfile2 /pnfs/myexperiment/scratch/users/$USER/myoutputs/
echo “Done!”
```

Job script #3

```
#!/bin/bash

export HOME=$PWD
. /cvmfs/fermilab.opensciencegrid.org/products/common/etc/setup
setup ifdhc
. /cvmfs/fermilab.opensciencegrid.org/products/larsoft/etc/setup
setup mycode v1_2_3_4 -q whatever
setup product1
setup product2
ifdh cp -D /pnfs/myexperiment/scratch/users/${GRID_USER}/myinput.tar.gz ./
tar zxf myinput.tar.gz ./
MYCONFIG=${HOME}/myconfig  # assuming myconfig is part of myinput.tar.gz
./myexe myargs
ifdh cp -D myoutfile1 myoutfile2 /pnfs/myexperiment/scratch/users/$GRID_USER/myoutputs/
echo “Done!”
```

Job script #4 (Getting closer to the mark now)

```
#!/bin/bash
/cvmfs/grid.cern.ch/util/cvmfs-uptodate /cvmfs/fermilab.opensciencegrid.org
if [ ! -d /cvmfs/myexperiment.opensciencegrid.org ]; then
echo “experiment CVMFS repo seems to not be present. Sleeping and then exiting.”
sleep 1000
exit 1
fi
. /cvmfs/fermilab.opensciencegrid.org/products/common/etc/setup
setup ifdhc
. /cvmfs/fermilab.opensciencegrid.org/products/larsoft/etc/setup
setup product1
setup product2
setup mycode v1_2_3_4 -q whatever
ifdh cp -D /pnfs/myexperiment/scratch/users/${GRID_USER}/myinput.tar.gz ./
IFDH_RESULT=$?
ifdh log  “ifdh cp of myinput.tar.gz for job $JOBSUBJOBID finished at `date` with status $IFDH_RESULT”
if [ $IFDH_RESULT -ne 0 ]; then echo “Error copying input tar” ; exit 2 ; fi
tar zxf myinput.tar.gz ./
MYCONFIG=${HOME}/myconfig  # assuming myconfig is part of myinput.tar.gz
./myexe myargs
RESULT=$?
if [ $RESULT -ne 0 ]; then exit $RESULT ; fi
ifdh cp -D myoutfile1 myoutfile2 /pnfs/myexperiment/scratch/users/$GRID_USER/myoutputs/
RESULT=$?
if [ $RESULT -ne 0 ]; then echo “ifdh failed with status $RESULT ; fi
echo “Done”
```

Job script #5 (Closer still)

```
#!/bin/bash
echo “Job starting on `uname -a`”
echo “Site is $GLIDEIN_Site”
/cvmfs/grid.cern.ch/util/cvmfs-uptodate /cvmfs/fermilab.opensciencegrid.org
case `uname -r` in
   3.*) export UPS_OVERRIDE="-H Linux64bit+2.6-2.12";;
   4.*) export UPS_OVERRIDE="-H Linux64bit+2.6-2.12";;
esac
if [ ! -d /cvmfs/myexperiment.opensciencegrid.org ]; then
echo “experiment CVMFS repo seems to not be present. Sleeping and then exiting.”
sleep 1000
exit 1
fi
. /cvmfs/fermilab.opensciencegrid.org/products/common/etc/setup
setup ifdhc
. /cvmfs/fermilab.opensciencegrid.org/products/larsoft/etc/setup
setup product1
setup product2
setup mycode v1_2_3_4 -q whatever
setup my_extra_libs vX_Y -z /cvmfs/myexperiment.opensciencegrid.org/products # perhaps a UPS package in your experiment CVMFS area?
ifdh cp -D /pnfs/myexperiment/scratch/users/${GRID_USER}/myinput.tar.gz ./
ifdh log  “ifdh cp of myinput.tar.gz for job $JOBSUBJOBID finished at `date` with status $IFDH_RESULT”
if [ $IFDH_RESULT -ne 0 ]; then echo “Error copying input tar” ; exit 2 ; fi
tar zxf myinput.tar.gz ./
MYCONFIG=${HOME}/myconfig  # assuming myconfig is part of myinput.tar.gz
./myexe myargs
RESULT=$?
if [ $RESULT -ne 0 ]; then exit $RESULT ; fi
ifdh cp -D myoutfile1 myoutfile2 /pnfs/myexperiment/scratch/users/$GRID_USER/myoutputs/
RESULT=$?
if [ $RESULT -ne 0 ]; then echo “ifdh failed with status $RESULT” ; exit $RESULT ; fi
echo “Done”
```


Fermilab

Job script #5 (Closer still)

```
#!/bin/bash
echo “Job starting on `uname -a`”
echo “Site is $GLIDEIN_Site”
/cvmfs/grid.cern.ch/util/cvmfs-uptodate /cvmfs/fermilab.opensciencegrid.org
case `uname -r` in
   3.*) export UPS_OVERRIDE="-H Linux64bit+2.6-2.12";;
   4.*) export UPS_OVERRIDE="-H Linux64bit+2.6-2.12";;
esac
if [ ! -d /cvmfs/myexperiment.opensciencegrid.org ]; then
echo “experiment CVMFS repo seems to not be present. Sleeping and then exiting.”
sleep 1000
exit 1
fi
. /cvmfs/fermilab.opensciencegrid.org/products/common/etc/setup
setup ifdhc
. /cvmfs/fermilab.opensciencegrid.org/products/larsoft/etc/setup
setup product1
setup product2
setup mycode v1_2_3_4 -q whatever
setup my_extra_libs vX_Y -z /cvmfs/myexperiment.opensciencegrid.org/products # perhaps a UPS package in your experiment CVMFS area?
ifdh cp -D /pnfs/myexperiment/scratch/users/${GRID_USER}/myinput.tar.gz ./
ifdh log  “ifdh cp of myinput.tar.gz for job $JOBSUBJOBID finished at `date` with status $IFDH_RESULT”
if [ $IFDH_RESULT -ne 0 ]; then echo “Error copying input tar” ; exit 2 ; fi
tar zxf myinput.tar.gz ./
MYCONFIG=${HOME}/myconfig  # assuming myconfig is part of myinput.tar.gz
./myexe myargs
RESULT=$?
if [ $RESULT -ne 0 ]; then exit $RESULT ; fi
ifdh cp -D myoutfile1 myoutfile2 /pnfs/myexperiment/scratch/users/$GRID_USER/myoutputs/
RESULT=$?
if [ $RESULT -ne 0 ]; then echo “ifdh failed with status $RESULT” ; exit $RESULT ; fi
echo “Done”
```

Everyone: be sure that cetpkgsupport v1_10_02 or greater is set up! Experiments using SRT packages (NOvA) may require additional variables to be checked and/or reset

Running this onsite

- Will any of these changes be a problem on-site?
  - NO! Anything that runs on the OSG will run on GPGrid.
  - Best to write your job scripts with OSG in mind from the beginning. Then you only need to write them once.
- Things to keep in mind
  - Use CVMFS for your code instead of /grid/fermiapp, even if you’re only running on-site. Less load on BlueArc!

Always test it first!!!

- Do NOT send many jobs without sending tests first.
- A fermicloud node is available to run tests. Copy your script there and run it first! If it works, send a single test job to a remote site, then some more, etc. There is also a node called gpgtest.fnal.gov where you an test workflows (just run a few events for example)

Summary

- Increased use of OSG resources key to meeting experiment computing requests
- We've recapped the basics of job submission to the OSG
- The keys to success include carefully requesting resources and writing robust scripts
- Good scripts include the following elements:
  - Portability (no dependence on absolute paths other than CVMFS)
  - Good logging and exit code checking
  - No dependence on user ID
- We are always happy to look things over for you and provide suggestions!

FABRIC FOR FRONTIER EXPERIMENTS