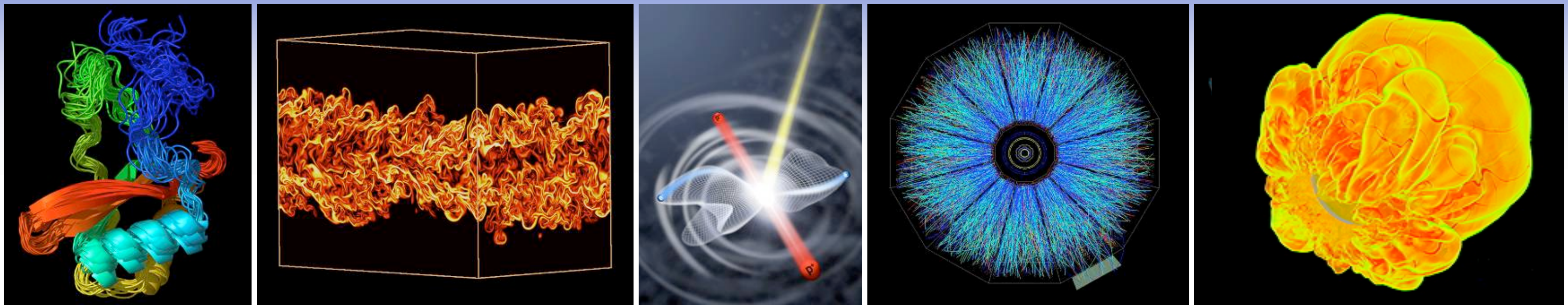


Integrating Grid Services into a Cray XT4 Environment

Hwa-Chun Wendy Lin and Shreyas Cholia
National Energy Research Scientific Computing Center
(NERSC/LBL)

CUG 2009, Atlanta, GA

What Is a Grid?

"A grid is a system that coordinates resources that are not subject to centralized control, using standard, open, general-purpose protocols and interfaces, to deliver nontrivial qualities of service."

-- Ian Foster

What Is Globus Toolkit?

- Globus Toolkit/GT: an implementation of grid services standards/protocols
  - Core: Security Services
    - Grid Security Infrastructure (GSI)
      - Authentication (Who you are)
      - Authorization (What you can do on my system)
  - Three pillars (primary components)
    - Information Services (MDS)
    - Resource Management (GRAM)
    - Data Management (GridFTP)

What Is Open Science Grid (OSG)?

- Originally a High Energy Physics Grid
- Data source: the LHC (Large Hadron Collider) @CERN
- Data relaying: Tier-1 sites
- Data processing: Tier-2 sites
- Virtual organization (VO): CMS, Atlas, etc
- Non-LHC VOs added: STAR, ITER, RENCI, LIGO, etc
- Parallel resources desirable

OSG Stack for CE

- VDT (Virtual Data Toolkit)
- Globus Toolkit
  - GSI (Authentication & Authorization)
  - GRAM (Job submission)
  - GridFTP (Data management)
- OSG specific for Compute Element (CE)
  - CEMon (Resource descriptions)
  - RSV (Resource availability)
  - Gratia (Accounting)

Franklin Specifics

- Designated grid node: alias franklingrid
- Production system shared with local users
  - Privilege separation important
  - OSG software installed on /usr/common/osg as the globus user
  - OSG cron jobs run as the globus user
- Shared-root environment
  - Specialized for the franklingrid node
    - /etc/xinetd.d/gsiftp -> /.shared/base/node/256/etc/xinetd.d/gsiftp
    - /etc/xinetd.d/gsigatekeeper -> /.shared/base/node/256/etc/xinetd.d/gsigatekeeper
    - /etc/init.d/rc3.d/K03xinetd, /etc/init.d/rc3.d/S20xinetd
    - /etc/grid-security -> /usr/common/osg/grid-security

Franklin Specifics (cont.)

- **Jobmanager-pbs**
  - **Aprun with mppwidth, mppnppn conversions**
- **CEMon resource discovery**
  - **Finds system characteristics about franklingrid, a service node**
    - **Need to override to provide compute nodes info**
- **Gratia probes**
  - **PBS server runs on the SDB node**
    - **Accounting data are copied over from server's private /var to /usr/common daily**
  - **Filter out entries about local jobs**

NERSC Specifics

- Requirement of individual accounts
  - DOE requirement
  - No VO support
- Short-lived proxy certificate issued by NERSC CA
  - NERSC-wide setup
  - X.509 Public Key Infrastructure (PKI) certificate management painful
  - Handled by the online MyProxy credential management service
    - myproxy-logon

Why Use the Grid?

- **Job can be managed remotely without users' knowing about batch system specifics**
  - **mpiexec vs. aprun vs. poe**
  - **qsub vs. llsubmit**
  - **qstat vs. llq**
  - **pbsnodes vs. llstatus**

Batch Job Submission

qsub qsub.cmd

```
#PBS -l mppwidth=4
#PBS -o test.out
#PBS -e test.err
cd test_dir
aprun -n 4 ./test_application
```

llsubmit llsub.cmd

```
#@ job_type=parallel
#@ cpus=4
#@ output=test.out
#@ error=test.err
#@ queue
poe test_dir/test_application
```

What Is a Grid job?

- Job specifics, such as resource requirements, are specified in RSL (Resource Specification Language), directly or indirectly

- Job submits to a Globus gatekeeper, directly or indirectly

Grid Job Submission: Globus

globusrun

```
globusrun -r franklingrid.nersc.gov/jobmanager-pbs -f cmd.rsl
```

```
& (count=4)
   (jobtype=mpi)
   (directory=test_dir)
   (executable=test_application)
   (stdout=x-gass-cache://$(GLOBUS_GRAM_JOB_CONTACT)stdout anExtraTag)
   (stderr=x-gass-cache://$(GLOBUS_GRAM_JOB_CONTACT)stderr anExtraTag)
```

globus-job-submit

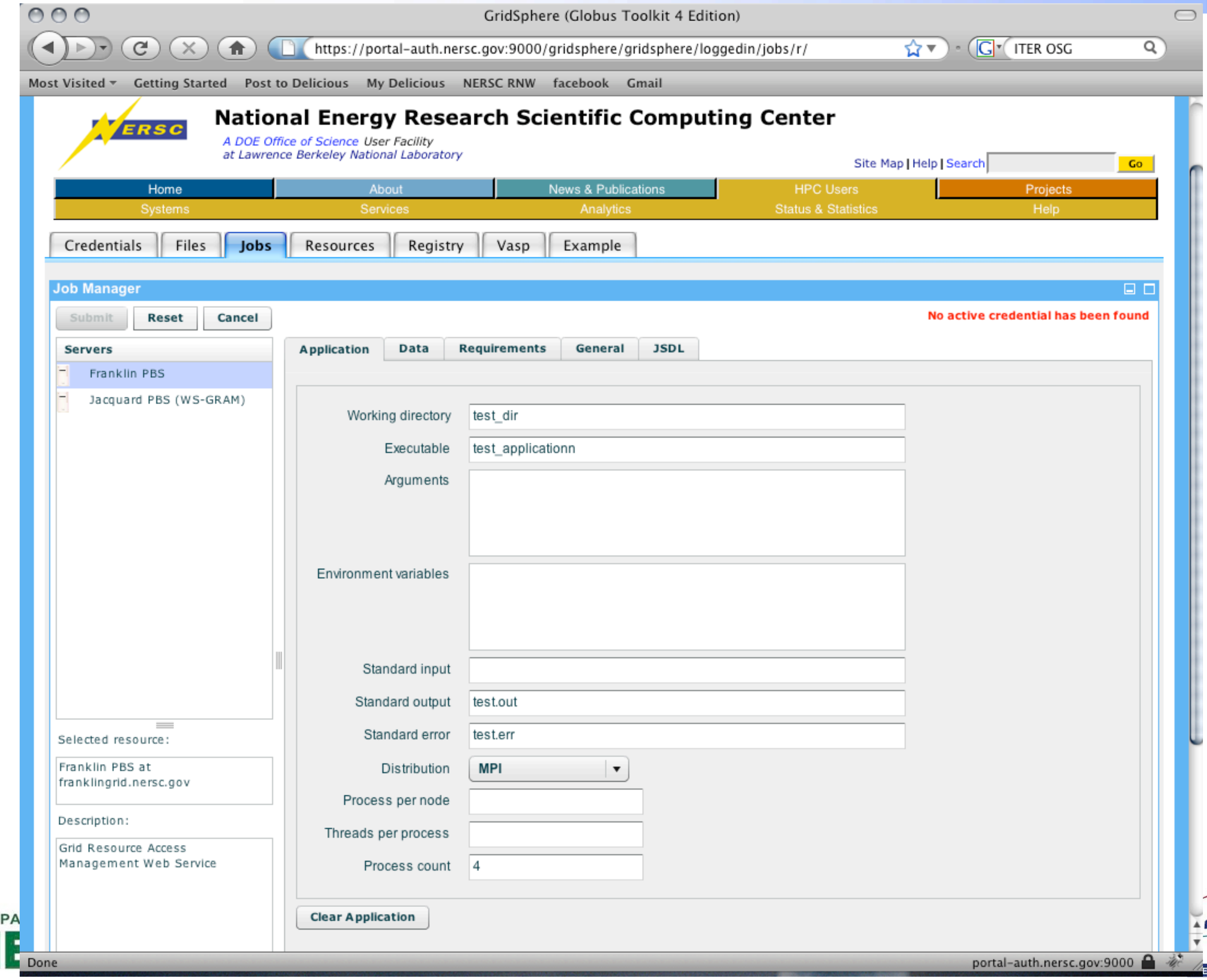
```
globus-job-submit franklingrid.nersc.gov/jobmanager-pbs -np 4
        -x‘&(jobtype=mpi)’ test_dir/test_application
```

Grid Job Submission: Condor-G

condor-submit test.cmd

```
Universe = grid
Executable = test_dir/test_application
transfer_executable = false
grid_resource = gt2 franklingrid.nersc.gov/jobmanager-pbs
globus_rsl = (jobType=mpi) (count=4)
output = test.out
error = test.err
Queue
```

Grid Job Submission: Portals/Science Gateways



GridSphere (Globus Toolkit 4 Edition)

https://portal-auth.nersc.gov:9000/gridsphere/gridsphere/loggedin/jobs/r/

National Energy Research Scientific Computing Center

A DOE Office of Science User Facility at Lawrence Berkeley National Laboratory

Job Manager

No active credential has been found

Selected resource:

Franklin PBS at franklingrid.nersc.gov

Description:

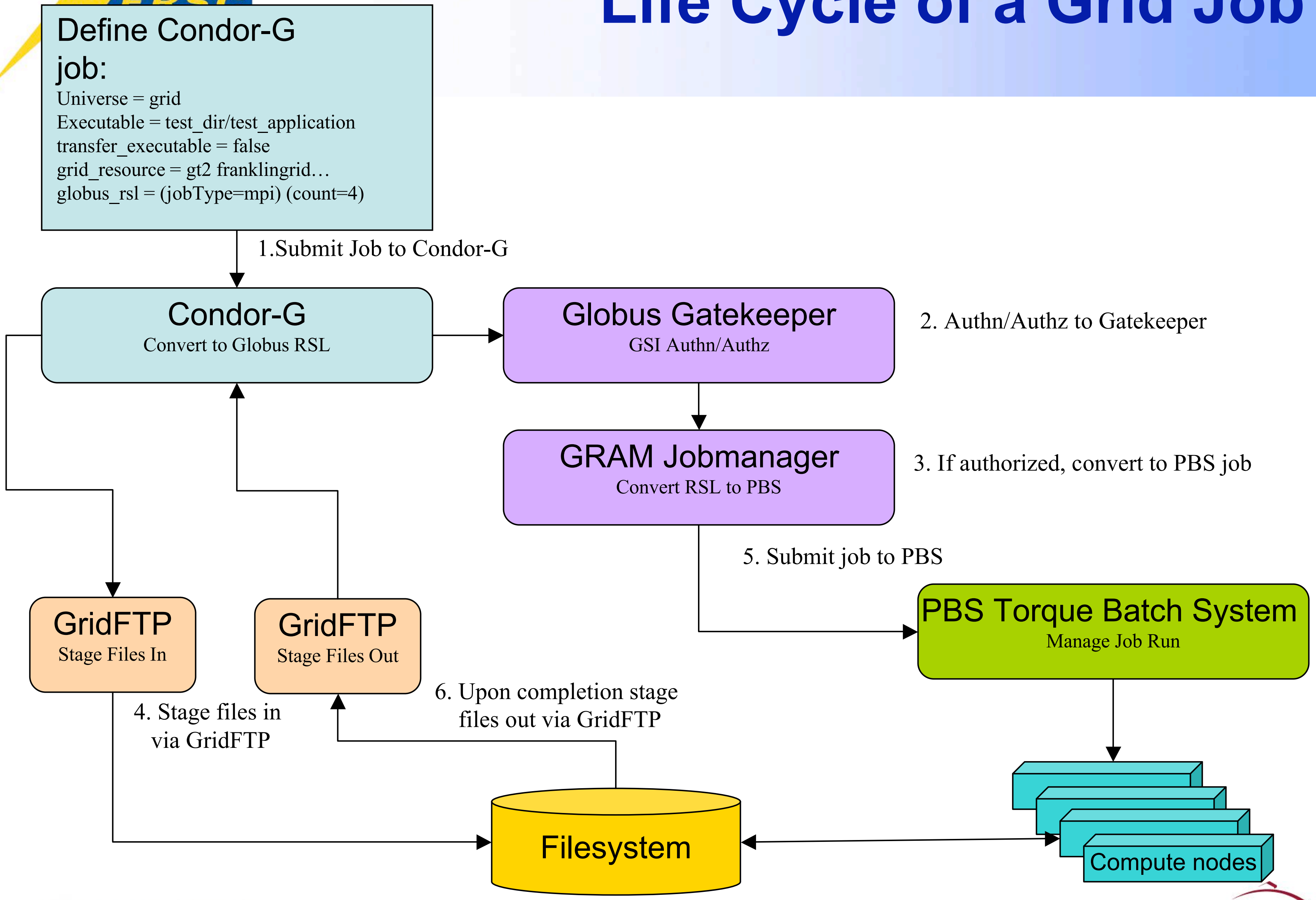
Grid Resource Access Management Web Service

Working directory	test_dir
Executable	test_applicationn
Arguments	
Environment variables	
Standard input	
Standard output	test.out
Standard error	test.err
Distribution	MPI
Process per node	
Threads per process	
Process count	4

Done

portal-auth.nersc.gov:9000

Life Cycle of a Grid Job



Define Condor-G job:

```
Universe = grid
Executable = test_dir/test_application
transfer_executable = false
grid_resource = gt2 franklingrid…
globus_rsl = (jobType=mpi) (count=4)
```

1.Submit Job to Condor-G

Condor-G
Convert to Globus RSL

Globus Gatekeeper
GSI Authn/Authz

2. Authn/Authz to Gatekeeper

GRAM Jobmanager
Convert RSL to PBS

3. If authorized, convert to PBS job

4. Stage files in via GridFTP

GridFTP
Stage Files In

GridFTP
Stage Files Out

5. Submit job to PBS

PBS Torque Batch System
Manage Job Run

6. Upon completion stage files out via GridFTP

Filesystem

Compute nodes

Work in Progress

- The Project Account Project
  - Satisfy users' desire to share data and work
  - Satisfy DOE's requirement for tracking individuals' use of resources
  - Add the VO support afterwards
- The esLogin Project
  - Provide external login capability for franklin
  - Move grid stuff to an external login node
    - Simplify the shared-root environment
    - Increase the grid node stability

Conclusion

- Useful in running production codes
  - Developers build codes for specific platforms
  - Users use the codes provided
- Not useful in Top 500 LinPack runs
- Overall performance vs. individual runs performance

Acknowledgements

- DOE for supporting NERSC
- Follow-up e-mail:
  - scholia@lbl.gov
  - hclin@lbl.gov