

Generic Portals for Science Infrastructure (GPSI)

Thomas D. Uram, Michael E. Papka, Mark Hereld, Michael Wilde
Argonne National Laboratory, University of Chicago Computation Institute

GPSI

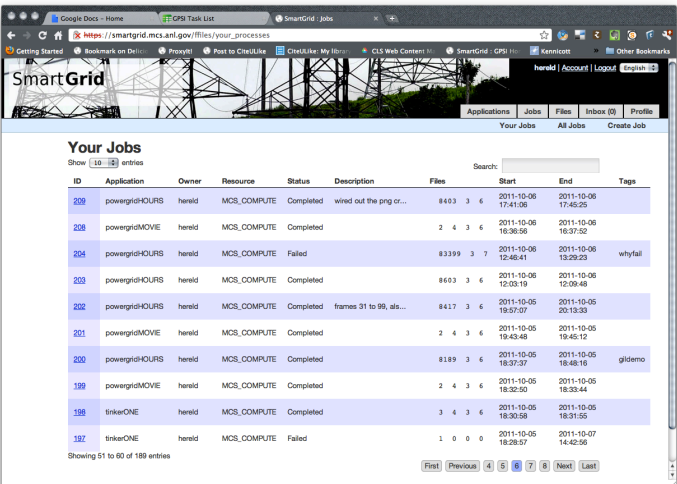
Generic Portals for Science Infrastructure

Approach

- Simplify execution and data management on compute resources
- Support reproducibility of results through execution and data provenance
- Enable collaboration among researchers working on a science campaign
- Capture details of scientists' idiosyncratic workflows

Research Questions

- Effective scheduling of jobs across resources
- Efficient data staging to and from compute resources
- Automated data analysis and feature extraction
- Data filtering for remote analysis



SmartGrid

Applications | Jobs | Files | Inbox | Profile

Your Jobs | All Jobs | Create Job

Your Jobs

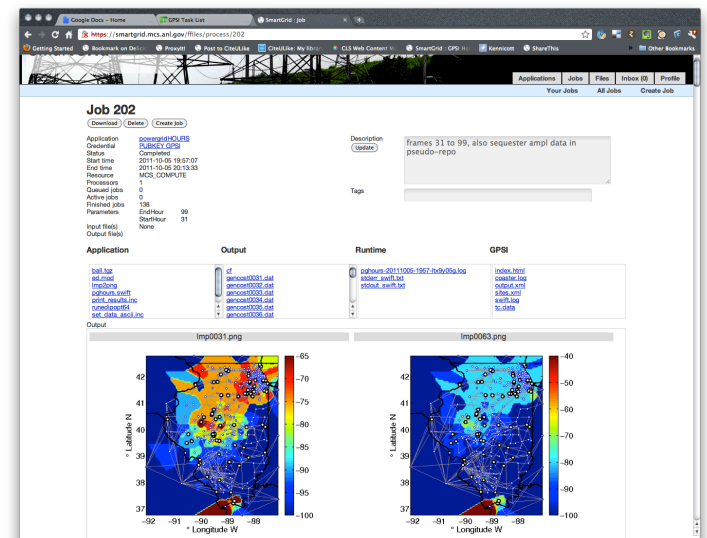
Show 10 entries

ID	Application	Owner	Resource	Status	Description	Files	Start	End	Tags
209	powergridHOURS	herald	MCS_COMPUTE	Completed	wired out the png or...	8453 3 6	2011-10-06 17:41:06	2011-10-06 17:45:25	
208	powergridMOVE	herald	MCS_COMPUTE	Completed		2 4 3 6	2011-10-06 18:36:56	2011-10-06 18:37:52	
204	powergridHOURS	herald	MCS_COMPUTE	Failed		83399 3 7	2011-10-06 12:46:41	2011-10-06 13:25:23	whyfail
203	powergridHOURS	herald	MCS_COMPUTE	Completed		8663 3 6	2011-10-06 12:08:19	2011-10-06 12:09:48	
202	powergridHOURS	herald	MCS_COMPUTE	Completed	frames 31 to 99, also...	8417 3 6	2011-10-05 19:57:07	2011-10-05 20:12:33	
201	powergridMOVE	herald	MCS_COMPUTE	Completed		2 4 3 6	2011-10-05 19:43:48	2011-10-05 19:45:12	
200	powergridHOURS	herald	MCS_COMPUTE	Completed		8189 3 6	2011-10-05 18:37:37	2011-10-05 18:48:16	glidem
199	powergridMOVE	herald	MCS_COMPUTE	Completed		2 4 3 6	2011-10-05 18:30:50	2011-10-05 18:33:44	
198	linkerONE	herald	MCS_COMPUTE	Completed		3 4 3 6	2011-10-05 18:30:58	2011-10-05 18:31:55	
197	linkerONE	herald	MCS_COMPUTE	Failed		1 0 0 0	2011-10-05 14:28:57	2011-10-07 14:42:56	

Showing 10 of 189 entries

First Previous 4 5 6 7 8 Next Last

View of job history



View of job details, including files and images produced

GPSI: Technologies

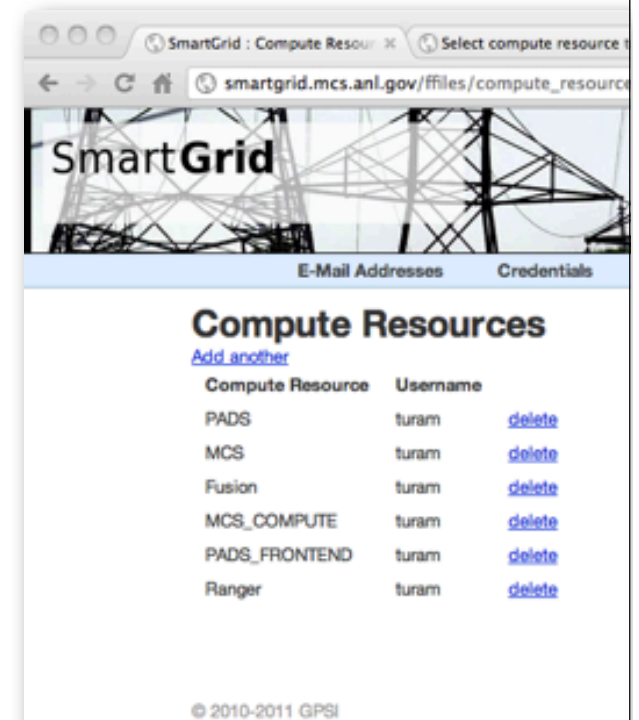
- Django core (hence Python)
 - Django is widely used in industry, and is the most popular Python web framework
 - Python is widely used in the scientific community
 - Uses a variety of database backends (MySQL, Sqlite)
- Collection of Django apps
 - Apps exist for (nearly) anything one needs
- GPSI is one more Django app
 - Can remain small, focus on only the gateway-related functionality
- Uses Swift for describing workflow and running jobs
- GPSI exposes opportunities for customization by science domains
 - User-driven gateways

GPSI: Overall Flow

- Introduce data
- Introduce applications
- Schedule jobs
- View/Analyze results
- Reproduce results
- Download
- Share

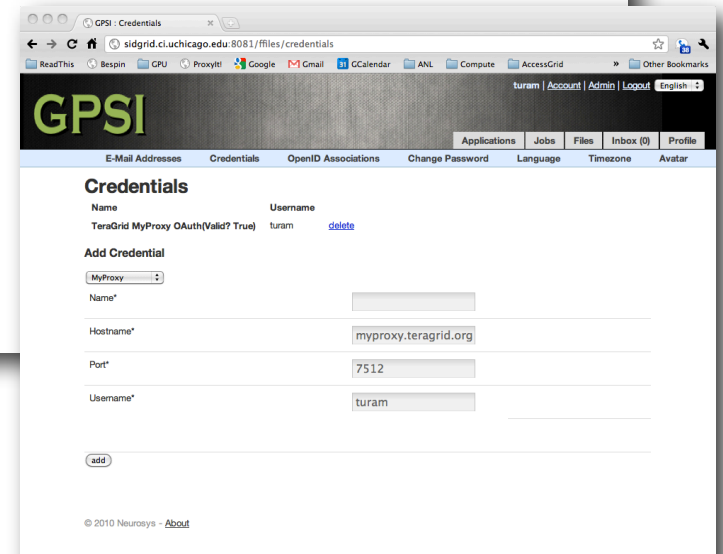
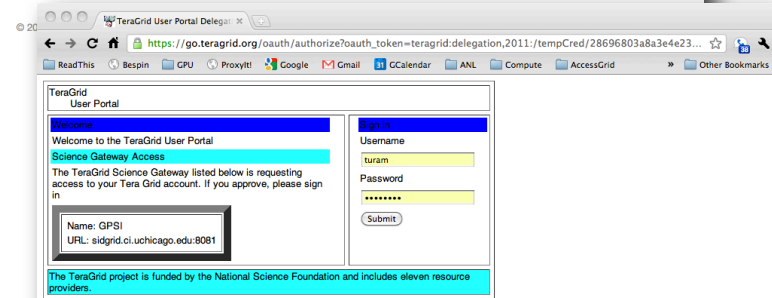
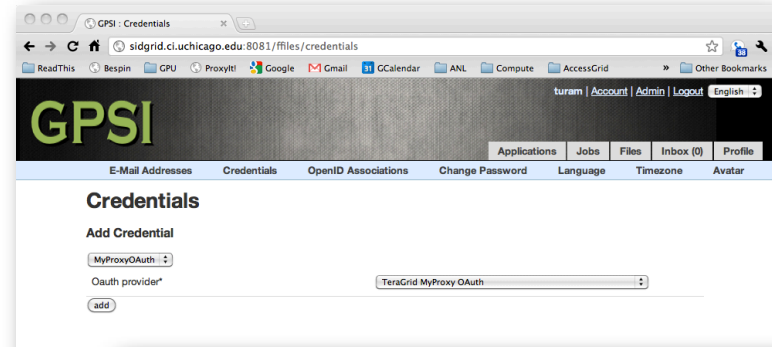
Compute Resource Setup

- Mechanisms supported by Swift
 - File transfer: gridftp, ssh, http
 - Execution: pbs, gram, sge, ssh, ec2, FutureGrid, ...
- Have run on a variety of resources
 - Ranger, Steele, Blacklight @ TeraGrid XSEDE
 - Compute machines at Argonne, including Fusion cluster
 - PADS at UChicago
 - Third-party resources
- Administratively controlled (for



Credential Management

- Credentials managed by user and associated with particular resources
- Credential types include
 - ssh
 - MyProxy
 - MyProxy server host, port, username and password
 - Credential renewed as needed
 - MyProxy OAuth
 - Avoids collection of MyProxy username and password by gateway
 - Have currently integrated pre-production TeraGrid MyProxy OAuth support



Introduce Data

- Upload relevant files, or ingest on server-side

Search across your files and those of your collaborators

Preview files inline

The screenshot shows the GPSI Files web interface. The search bar contains 'dpyr'. The results table lists files with columns for ID, Filename, Owner, Size, Date, and Tags. The files are sorted by date, showing a list of 'dpyr.out' and 'dpyr_im.out' files from June 15, 2009.

ID	Filename	Owner	Size	Date	Tags
13324	dpyr.out	admin	1237766116	2009-06-15 20:39:56	
13314	dpyr_im.out	admin	1289083325	2009-06-15 20:39:56	
13262	dpyr.out	admin	1237766116	2009-06-15 20:39:56	
13252	dpyr_im.out	admin	1289083325	2009-06-15 20:39:56	
13201	dpyr.out	admin	1237766116	2009-06-15 20:39:56	
13191	dpyr_im.out	admin	1289083325	2009-06-15 20:39:56	
13142	dpyr.out	admin	1237766116	2009-06-19 18:58:29	
13133	dpyr_im.out	admin	1289083325	2009-06-19 18:58:29	
13082	dpyr.out	admin	1238723983	2009-06-19	

The screenshot shows the inline preview of a file named 'memory.out'. The preview displays the output of a memory check, showing timestamps and node information for various jobs.

memory.out

Preview Download Update Delete

Owner: [admin](#)
Created: June 15, 2009
Size: 3946
Produced by: [Job 225](#)
Used by: None

Tags

Time stamp for mem check: Mon Jun 15 17:48:35 CDT 2009
NODE j88: 44M Mon Jun 15 17:48:38 CDT 2009
NODE j116: 44M Mon Jun 15 17:48:41 CDT 2009

Time stamp for mem check: Mon Jun 15 17:53:41 CDT 2009
NODE j87: 53M Mon Jun 15 17:53:45 CDT 2009
NODE j88: 46M Mon Jun 15 17:53:47 CDT 2009
NODE j116: 44M Mon Jun 15 17:54:10 CDT 2009

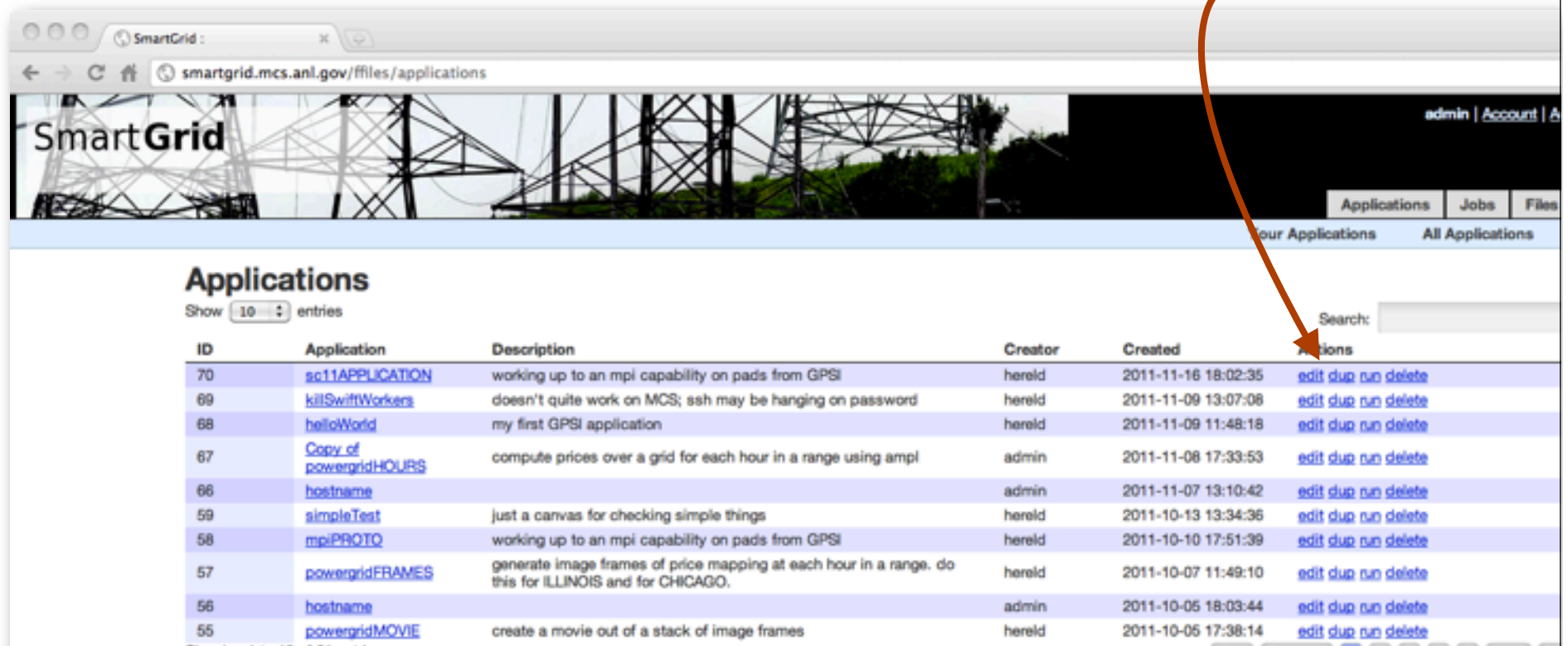
Time stamp for mem check: Mon Jun 15 17:59:10 CDT 2009
NODE j87: 53M Mon Jun 15 17:59:13 CDT 2009
NODE j88: 46M Mon Jun 15 17:59:15 CDT 2009
NODE j116: 44M Mon Jun 15 17:59:18 CDT 2009

Time stamp for mem check: Mon Jun 15 18:04:18 CDT 2009
NODE j87: 53M Mon Jun 15 18:04:22 CDT 2009
NODE j88: 46M Mon Jun 15 18:04:24 CDT 2009
NODE j116: 44M Mon Jun 15 18:04:29 CDT 2009

Introduce Applications

- Upload relevant files
- Application definition relies on required applications residing on compute resource
- Edit Swift script within GPSI

Edit, duplicate, run,
delete jobs



SmartGrid : smartgrid.mcs.anl.gov/ffiles/applications

admin | Account

Applications Jobs Files

My Applications All Applications

Applications

Show 10 entries

ID	Application	Description	Creator	Created	Actions
70	sc11APPLICATION	working up to an mpi capability on pads from GPSI	hereld	2011-11-16 18:02:35	edit dup run delete
69	killSwiftWorkers	doesn't quite work on MCS; ssh may be hanging on password	hereld	2011-11-09 13:07:08	edit dup run delete
68	helloWorld	my first GPSI application	hereld	2011-11-09 11:48:18	edit dup run delete
67	Copy of powergridHOURS	compute prices over a grid for each hour in a range using ampi	admin	2011-11-08 17:33:53	edit dup run delete
66	hostname		admin	2011-11-07 13:10:42	edit dup run delete
59	simpleTest	just a canvas for checking simple things	hereld	2011-10-13 13:34:36	edit dup run delete
58	mpiPROTO	working up to an mpi capability on pads from GPSI	hereld	2011-10-10 17:51:39	edit dup run delete
57	powergridFRAMES	generate image frames of price mapping at each hour in a range. do this for ILLINOIS and for CHICAGO.	hereld	2011-10-07 11:49:10	edit dup run delete
56	hostname		admin	2011-10-05 18:03:44	edit dup run delete
55	powergridMOVIE	create a movie out of a stack of image frames	hereld	2011-10-05 17:38:14	edit dup run delete

Showing 1 to 10 of 81 entries

Swift

- Language for parallel scripting
- Execution engine with interfaces to various schedulers

```
# simple script to execute hostname N times
type file;

# Swift app definition; each call will schedule a job
(file t)do_hostname() {
    app {
        hostname stdout=@filename(t);
    }
}
foreach i in [0:50] {
    file outfile <concurrent_mapper;prefix="stdout",suffix=".txt">;
    outfile = do_hostname();
}
```

For info on Swift, see <http://www.ci.uchicago.edu/swift>

Execute Jobs

- Web form generated by introspecting Swift script, including default values
- Uses server-side proxy certificate for single sign on

The image displays two overlapping browser windows. The left window, titled 'SmartGrid: Create Job', shows a web form for creating a new job. The right window, titled 'GPSI', shows a table of existing jobs.

SmartGrid: Create Job

smartgrid.mcs.anl.gov/ffiles/process/create?application=57

admin | Account

Create Job

Application: [dropdown]

Description: [text area]

User compute resource*: PADS [dropdown]

Credential: [dropdown]

Tags: [text input]

Parameters

StartHour	0
EndHour	0
DataDir	/home/hereld/Projec
DataSuffix	.dat
chiPrefix	chi
illPrefix	ill
imgSuffix	.png

GPSI

sidgrid.ci.uchicago.edu:8081/ffiles/processes

admin

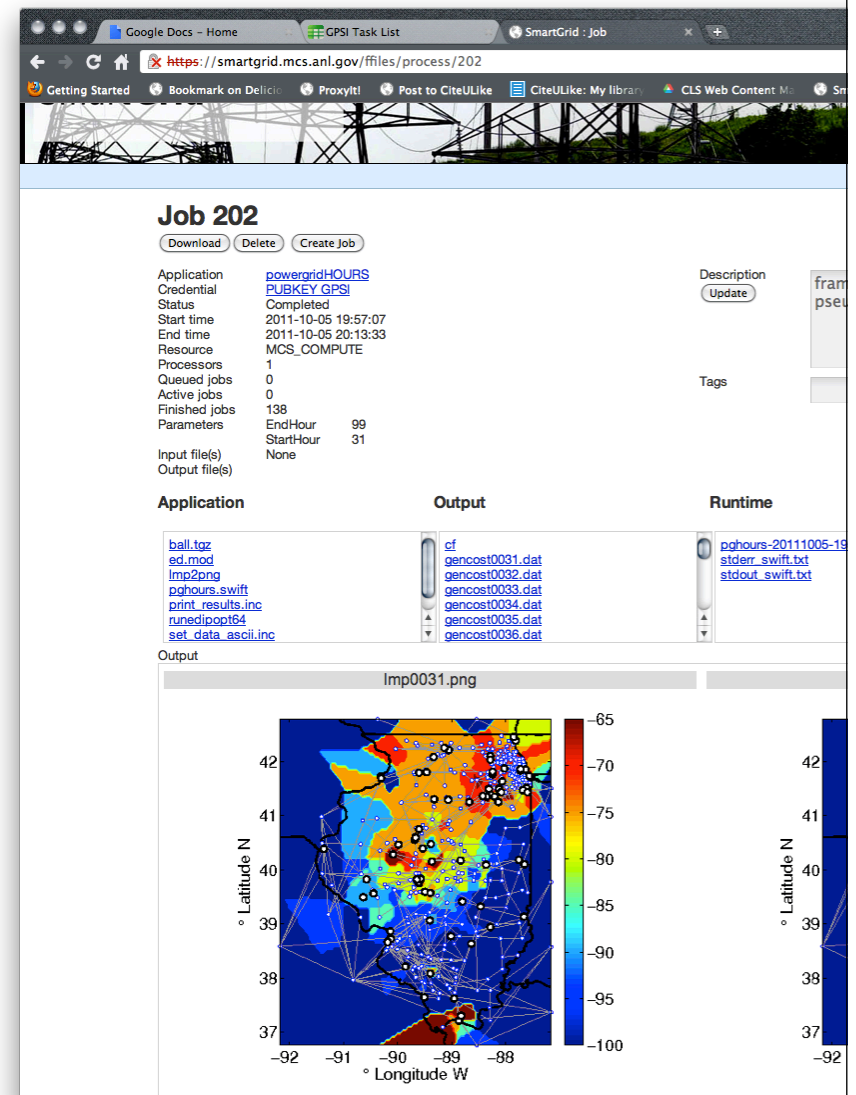
Jobs

Show 25 entries

ID	Application	Owner	Resource	Status
624	vi3.sh 1 30 624 dicom /gpfs/pads/scratch/turam/volume/27UP4XKH 1024x768	admin	PADS	Completed
623	vi3.sh 1 30 623 dicom /gpfs/pads/scratch/turam/volume/27UP4XKH 1024x768	admin	PADS	Completed
622	vi3.sh 1 30 622 dicom /gpfs/pads/scratch/turam/volume/27UP4XKH 1024x768	admin	PADS	Completed
591	vi3.sh 1 30 591 vigen sphere_32x32x32 1024x768	admin	PADS	Completed
589	vi3.sh 1 30 589 bin /gpfs/pads/scratch/turam/volume/zebrafish 512x384 1570x1450x100 16	admin	PADS	Completed
587	vi3.sh 1 30 587 dicom /gpfs/pads/scratch/turam/volume/27UP4XKH 512x384	admin	PADS	Completed
586	vi3.sh 1 30 586 dicom /gpfs/pads/scratch/turam/volume/27UP4XKH 240x160	admin	PADS	Completed
585	vi3.sh 1 30 585 vigen sphere_32x32x32 1024x768	admin	PADS	Completed
584	vi3.sh 5 30 584 bin /gpfs/pads/scratch/turam/volume/zebrafish/crop01-5.raw 1024x768 1000x1070x600 32	admin	PADS	Completed
583	vi3.sh 3 30 583 bin /gpfs/pads/scratch/turam/volume/zebrafish/crop01-5.raw 1024x768 1000x1070x600 32	admin	PADS	Completed
	vi3.sh 3 30 582 bin			

Generate, view, download results: Collaborate

- Well-known data types are presented graphically among job output
- Applications control which job outputs are included and how they're represented
 - XML designation of output files to include
 - XSLT transform of file input to representative HTML
- Users view each other's results, can run each other's apps



provenance

- Job details captured in one place
 - input, options, compute resource, credential
- Jobs can be rerun from within GPSI, using the original scenario

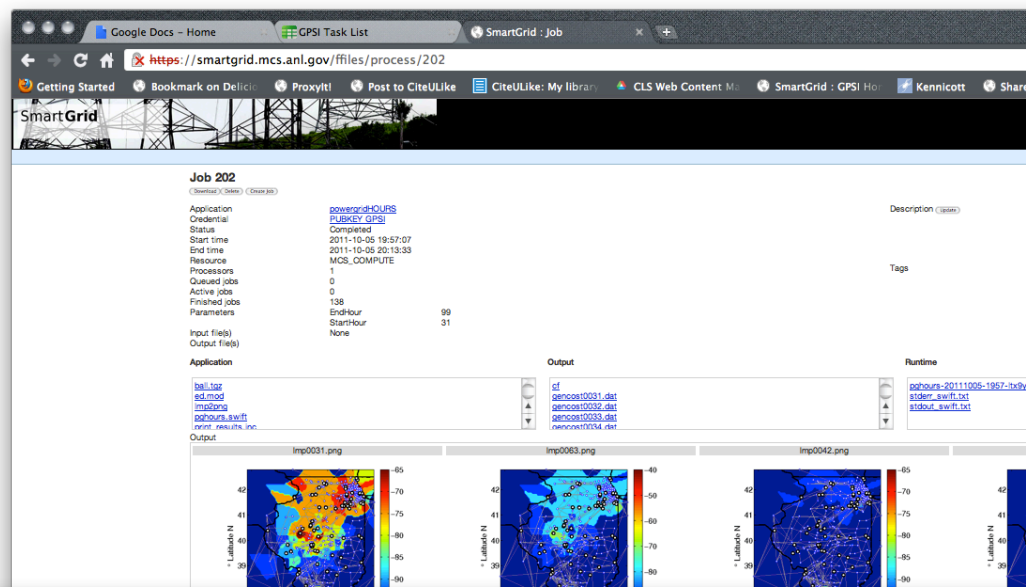
The screenshot displays the 'Job 202' page in the SmartGrid web application. The browser address bar shows the URL <https://smartgrid.mcs.anl.gov/files/process/202>. The page features a navigation bar with links for Applications, Jobs, Files, Inbox (0), and Profile. Below the navigation bar, the 'Job 202' section includes buttons for Download, Delete, and Create Job. The job details are as follows:

Application	powergridHOURS
Credential	PUBKEY GPSI
Status	Completed
Start time	2011-10-05 19:57:07
End time	2011-10-05 20:13:33
Resource	MCS_COMPUTE
Processors	1
Queued jobs	0
Active jobs	0
Finished jobs	138
Parameters	EndHour 99 StartHour 31
Input file(s)	None
Output file(s)	None

The 'Description' field contains the text: 'frames 31 to 99, also sequester ampl data in pseudo-repo'. The 'Tags' field is empty. Below the job details, there are four sections: Application, Output, Runtime, and GPSI. The 'Application' section lists files: ball_top, ed.mod, Imp2.png, pghours.swift, print_results.inc, runedipopt64, set_data_asciilinc. The 'Output' section lists files: cf, gencost0031.dat, gencost0032.dat, gencost0033.dat, gencost0034.dat, gencost0035.dat, gencost0036.dat. The 'Runtime' section lists files: pghours-20111005-1957-1tx9y05g.log, stderr_swift.txt, stdout_swift.txt. The 'GPSI' section lists files: index.html, coaster.log, output.xml, sites.xml, swift.log, tc_data. At the bottom, there are two heatmaps: 'Imp0031.png' and 'Imp0063.png'. The 'Imp0031.png' heatmap shows a color scale from -65 to -80, with a vertical axis labeled 'tude N' ranging from 40 to 42. The 'Imp0063.png' heatmap shows a color scale from -40 to -70, with a vertical axis labeled 'tude N' ranging from 40 to 42.

Where we're using it

- SmartGrid
- Materials
- Phylogenetics



The screenshot shows the GPSI web interface at the URL localhost:8082/files/file_search. The page features a search bar with the text "Code" and "contains" and a search button. Below the search bar, there is a grid of search results, each displaying a molecular structure and associated metadata. The results are organized into columns and rows, with each entry showing a date, a code, a description, and a search button. A red arrow points from the text "Triple-store for metadata integration and searching" to the search bar area.

Code	contains	Search
22-Dec-201 0	C6H8O3	Gaussian
08-Jul- 2010	C11H20O3	Gaussian
29-Sep-2009	C5H7N1	Gaussian
26-Aug-2010	C6H16O2S11	Gaussian
26-Jan-201 0	C6H17F1O2S11	Gaussian

Triple-store for
metadata integration
and searching

Future Work

- Integrated analysis
- Dynamic provisioning
- SCS integration
- Globus Online integration
- Data cataloging

Acknowledgements

- Michael E. Papka, Mark Hereld, Joe Insley, Eric Olson, and the Futures Laboratory at Argonne National Laboratory
- Mike Wilde and the Swift team
- Nancy Wilkins-Diehr, Suresh Marru, TeraGrid/XSEDE Gateways program
- Jim Basney, Jeff Gaynor for MyProxy OAuth
- NSF Award SCI-0503697
- XSEDE Science Gateways program

