

A Service for Queue Prediction and Job Statistics

Warren Smith

**Texas Advanced Computing Center
University of Texas at Austin**

Overview

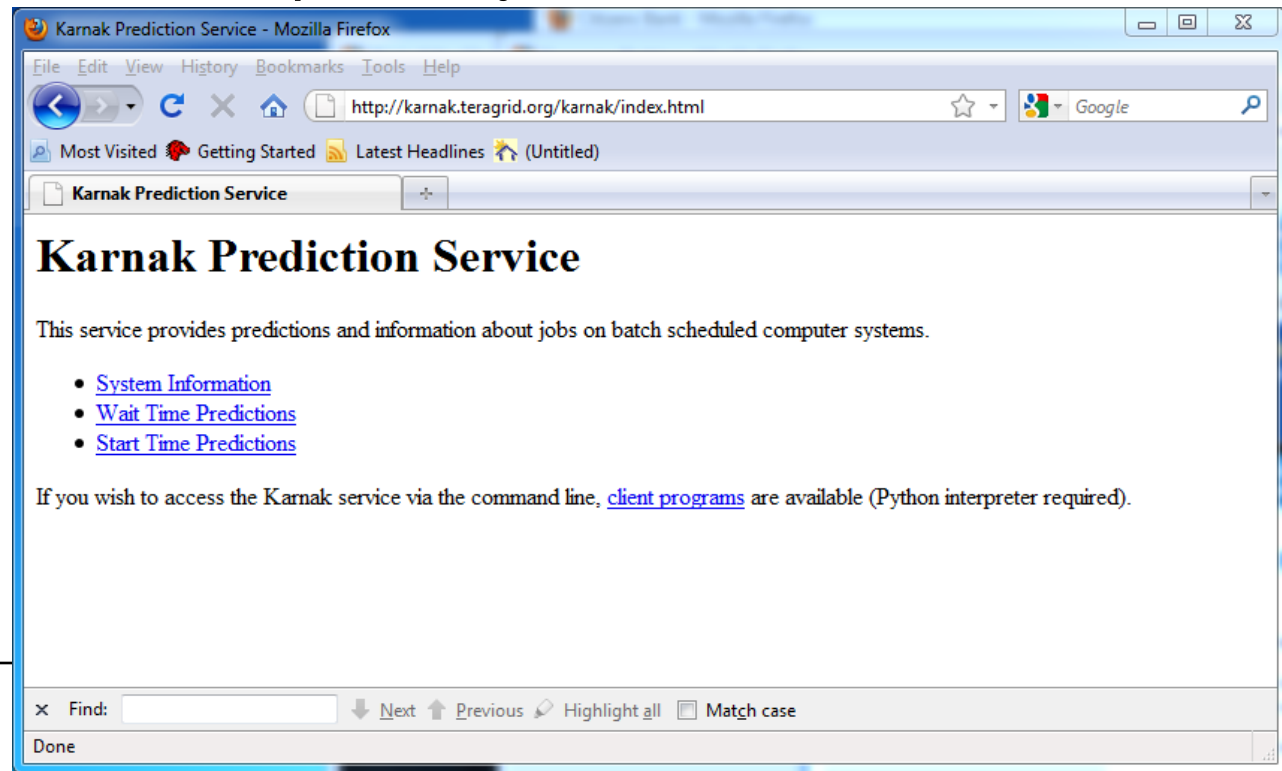
- Motivation
- Functionality
- Interfaces
- Performance
- Implementation

Motivation

- Many compute resources are managed by batch schedulers
 - Computations can wait in a queue for a non-trivial amount of time before they start
- Many gateways and users have access to multiple compute resources
- Help gateways and users:
 - Decide where to request allocations
 - Decide where to submit computations
 - Plan their work around submitted computations
- Deploy for production use in TeraGrid

Functionality

- Historical job information
- Current job information
- Queue predictions of potential jobs
- Queue predictions for queued jobs



Systems Overview

- Idea of how busy systems are
- Only a subset of TeraGrid systems shown
 - More on why later

Systems - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://karnak.teragrid.org/karnak/system/

Most Visited Getting Started Latest Headlines (Untitled)

Systems

System Status

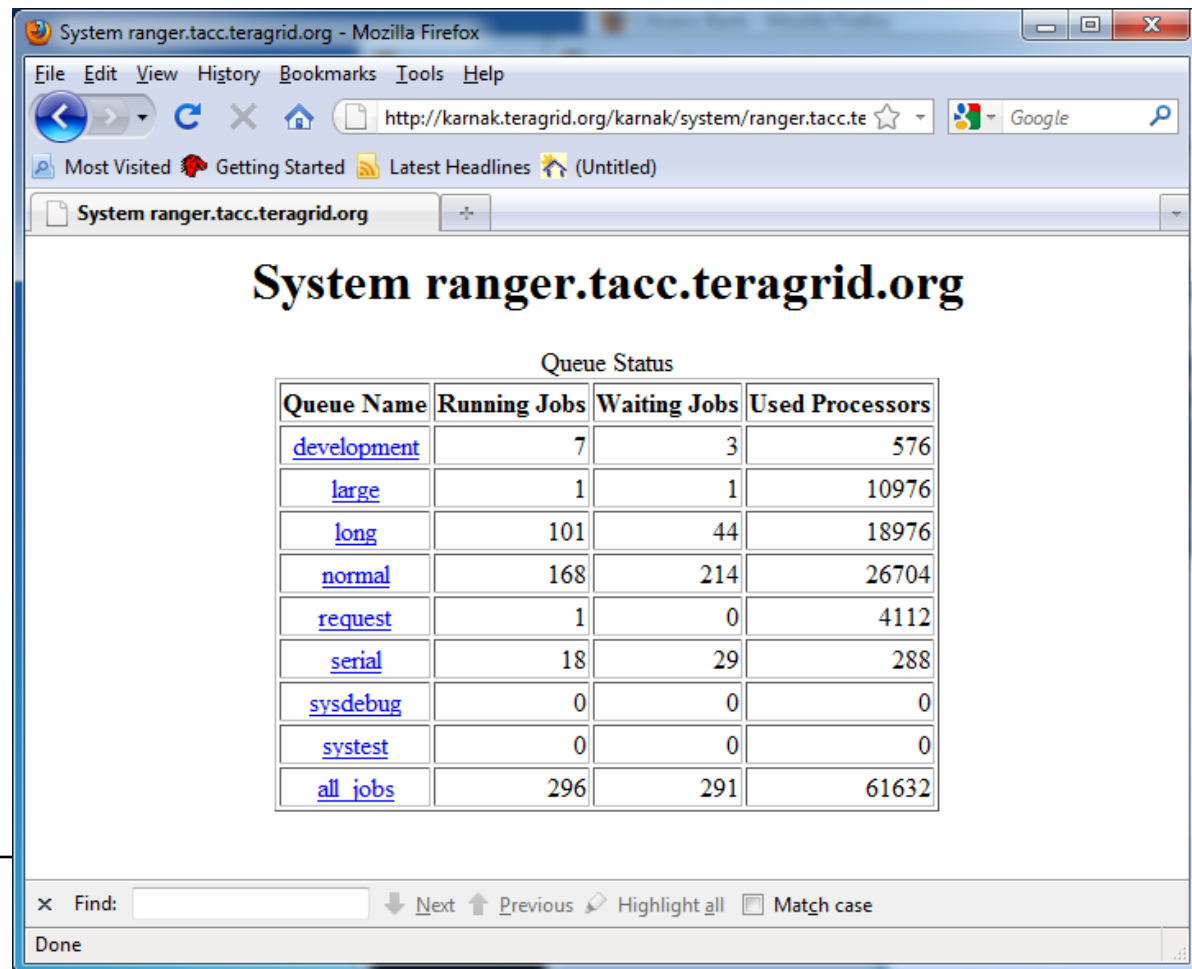
System Name	Running Jobs	Waiting Jobs	Used Processors
abe.ncsa.teragrid.org	172	70	5646
cobalt.ncsa.teragrid.org	50	223	963
lonestar.tacc.teragrid.org	50	84	4153
nstg.oml.teragrid.org	0	0	0
people.psc.teragrid.org	40	40	544
ranger.tacc.teragrid.org	296	291	61632

× Find: Next Previous Highlight all Match case

Done

Single System Overview

- Current jobs



System ranger.tacc.teragrid.org - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://karnak.teragrid.org/karnak/system/ranger.tacc.te

System ranger.tacc.teragrid.org

System ranger.tacc.teragrid.org

Queue Status

Queue Name	Running Jobs	Waiting Jobs	Used Processors
development	7	3	576
large	1	1	10976
long	101	44	18976
normal	168	214	26704
request	1	0	4112
serial	18	29	288
sysdebug	0	0	0
systest	0	0	0
all jobs	296	291	61632

Find: Next Previous Highlight all Match case

Done

Queue Overview

- Current status
- Historical information

The screenshot shows a web browser window with the URL `http://karnak.teragrid.org/karnak/system/ranger.tacc.teragrid.org/queue/all_jobs/s`. The page title is "System ranger.tacc.teragrid.org Queue all_jobs".

Status

Running Jobs	Waiting Jobs	Used Processors
296	291	61632

Started Jobs

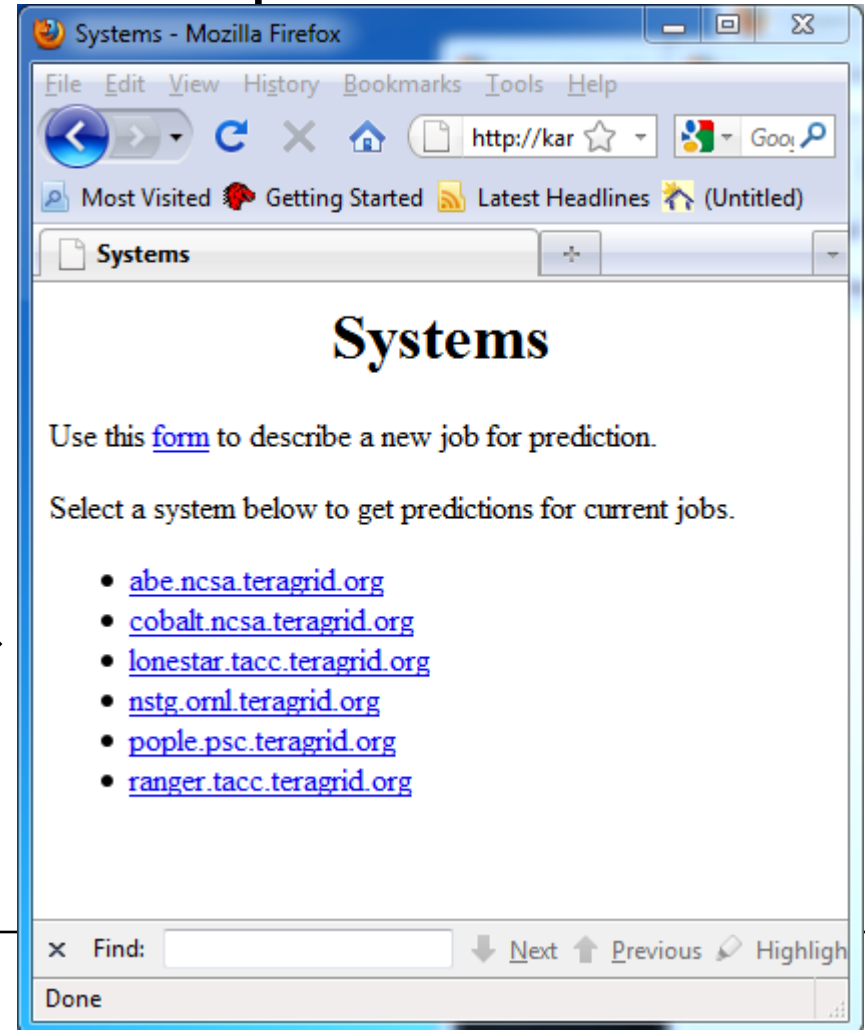
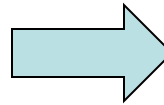
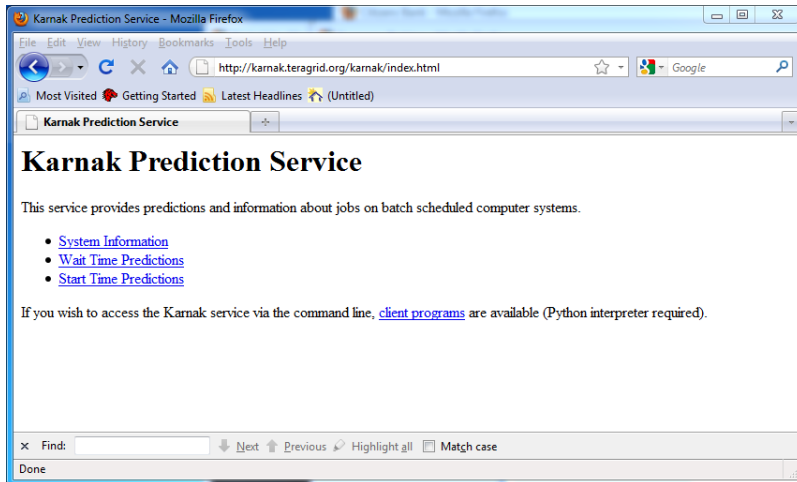
When	Number of Jobs	Mean Processors	Mean Requested Wall Time (hours:minutes:seconds)	Mean Wait Time (hours:minutes:seconds)
last hour	30	108	17:36:59	11:42:17
last four hours	189	423	11:05:51	08:04:47
last day	879	278	13:33:13	09:14:16
last week	6399	233	15:06:59	06:56:46

Completed Jobs

When	Number of Jobs	Mean Processors	Mean Requested Wall Time (hours:minutes:seconds)	Mean Wall Time (hours:minutes:seconds)
last hour	28	102	17:05:40	05:31:23
last four hours	169	457	10:11:04	05:02:09
last day	893	271	14:55:12	08:48:27
last week	6075	234	14:27:34	07:20:34

Wait Time Predictions

- How long a job will wait in the queue before it starts
- Describe a new job
- Identify an existing job



Predicting Waiting Jobs

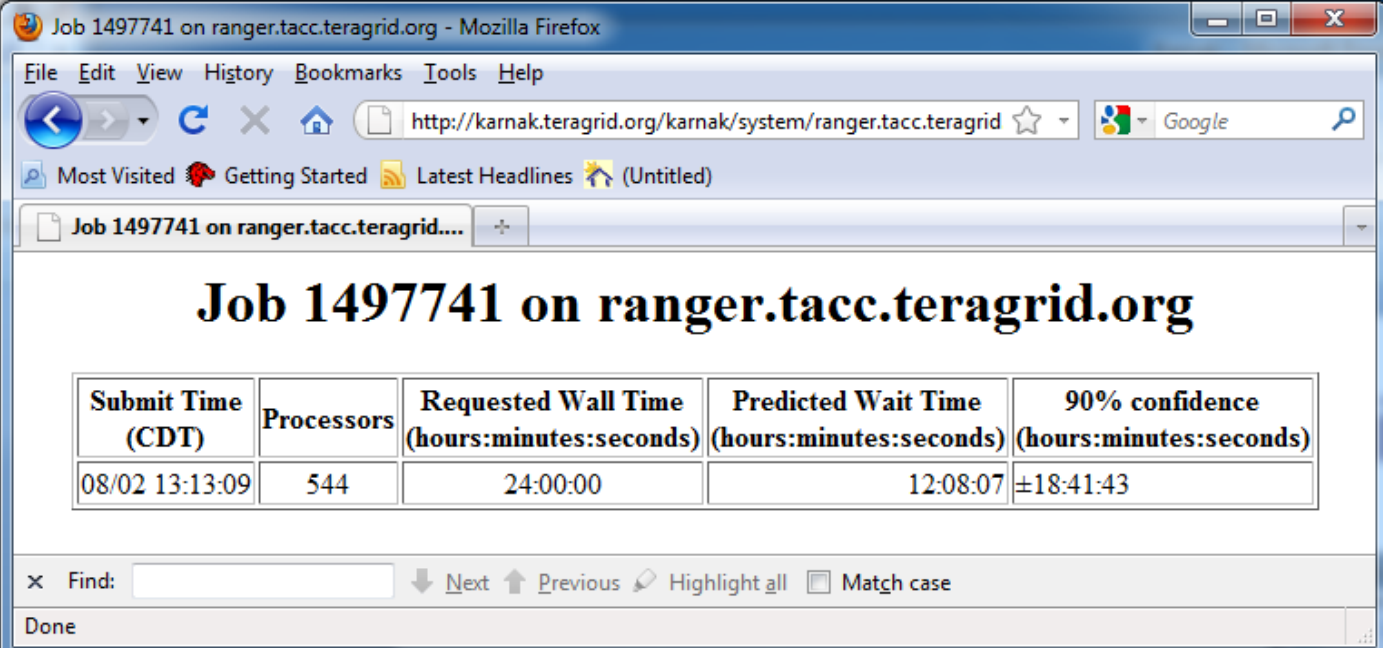
- A little information is presented about each job
- Jobs are presented in queue order
 - May not be schedule order
- This is accessible to anyone
 - No user names
 - No project names

Waiting Jobs on ranger.tacc.teragrid.org

Job Identifier	Submit Time (CDT)	Processors	Requested Wall Time (hours:minutes:seconds)
1497911	08/02 14:56:12	10976	06:00:00
1497629	08/02 12:05:07	8192	01:00:00
1498218	08/02 18:38:53	8192	00:10:00
1495584	08/01 06:28:28	16	00:10:00
1497674	08/02 12:23:26	400	24:00:00
1498143	08/02 17:28:17	768	00:50:00
1498050	08/02 16:08:58	32	24:00:00
1488245	07/27 13:03:28	128	02:00:00
1498269	08/02 20:19:35	192	12:00:00
1498144	08/02 17:29:19	768	00:50:00
1497142	08/02 08:51:15	16	24:00:00
1497765	08/02 13:32:07	48	03:00:00
1498264	08/02 20:10:29	96	01:29:00
1498261	08/02 20:02:50	64	08:00:00
1493645	07/30 14:04:34	128	02:00:00
1498057	08/02 16:11:18	32	24:00:00
1498301	08/02 21:39:24	64	24:00:00
1497938	08/02 15:11:19	1024	24:00:00
1497741	08/02 13:13:09	544	24:00:00
1497643	08/02 12:13:38	128	24:00:00
1498118	08/02 17:07:59	256	02:00:00
1498130	08/02 17:14:50	224	24:00:00
1498145	08/02 17:30:22	768	00:50:00
1497933	08/02 15:09:38	512	24:00:00

Wait Time Prediction

- Two components in the prediction
 - Predicted wait time
 - A confidence interval
 - Provides information about the expected accuracy of the prediction



Job 1497741 on ranger.tacc.teragrid.org - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://karnak.teragrid.org/karnak/system/ranger.tacc.teragrid

Most Visited Getting Started Latest Headlines (Untitled)

Job 1497741 on ranger.tacc.teragrid.org

Submit Time (CDT)	Processors	Requested Wall Time (hours:minutes:seconds)	Predicted Wait Time (hours:minutes:seconds)	90% confidence (hours:minutes:seconds)
08/02 13:13:09	544	24:00:00	12:08:07	±18:41:43

Find: Next Previous Highlight all Match case

Done

Start Time Predictions

- Same web pages until the prediction

Systems - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://kar

Systems

Systems

Use this [form](#) to describe a new job for prediction.

Select a system below to get predictions for current jobs.

- abc.ncsa.teragrid.org
- cobalt.ncsa.teragrid.org
- lonestar.tacc.teragrid.org
- nstg.ornl.teragrid.org
- people.psc.teragrid.org
- ranger.tacc.teragrid.org

Find: [] Next Previous Highlight

Done



Waiting Jobs on ranger.tacc.teragrid.org - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://karnak.teragrid.org/karnak/waitti

Waiting Jobs on ranger.tacc.teragrid...

Waiting Jobs on ranger.tacc.teragrid.org

Job Identifier	Submit Time (CDT)	Processors	Requested Wall Time (hours:minutes:seconds)
1497911	08/02 14:56:12	10976	06:00:00
1497629	08/02 12:05:07	8192	01:00:00
1498218	08/02 18:38:53	8192	00:10:00
1495584	08/01 06:28:28	16	00:10:00
1497674	08/02 12:23:26	400	24:00:00
1498143	08/02 17:28:17	768	00:50:00
1498050	08/02 16:08:58	32	24:00:00
1488245	07/27 13:03:28	128	02:00:00
1498269	08/02 20:19:35	192	12:00:00
1498144	08/02 17:29:19	768	00:50:00
1497142	08/02 08:51:15	16	24:00:00
1497765	08/02 13:32:07	48	03:00:00
1498264	08/02 20:10:29	96	01:29:00
1498261	08/02 20:02:50	64	08:00:00
1493645	07/30 14:04:34	128	02:00:00
1498057	08/02 16:11:18	32	24:00:00
1498301	08/02 21:39:24	64	24:00:00
1497938	08/02 15:11:19	1024	24:00:00
1497741	08/02 13:13:09	544	24:00:00
1497643	08/02 12:13:38	128	24:00:00
1498118	08/02 17:07:59	256	02:00:00
1498130	08/02 17:14:50	224	24:00:00
1498145	08/02 17:30:22	768	00:50:00
1497933	08/02 15:09:38	512	24:00:00

Find: [] Next Previous Highlight all Match case

Done

Start Time Prediction

- Provides a start time instead of a wait time

Job 1497741 on ranger.tacc.teragrid.org

Submit Time (CDT)	Processors	Requested Wall Time (hours:minutes:seconds)	Predicted Start Time (CDT)	90% confidence (hours:minutes:seconds)
08/02 13:13:09	544	24:00:00	08/03 10:23:53	±18:36:31

Find: Next Previous Highlight all Match case

Done

Potential Job

- Describe a potential job
 - Systems & queues
 - Processing cores
 - Wall time
 - Size of confidence interval

The screenshot shows a web browser window titled "Wait Time Predictions - Mozilla Firefox" with the address bar at "http://karnak.teragrid.org". The page content is titled "Wait Time Prediction Form" and includes a section for selecting systems and queues, followed by fields for job description, processing cores, wall time, and confidence interval size, and a "Submit" button.

Select one or more systems and queues:

System	Queues
<input checked="" type="checkbox"/> abe.ncsa.teragrid.org	<input type="checkbox"/> debug <input type="checkbox"/> long <input checked="" type="checkbox"/> normal <input type="checkbox"/> wide
<input checked="" type="checkbox"/> cobalt.ncsa.teragrid.org	<input type="checkbox"/> debug <input type="checkbox"/> long <input checked="" type="checkbox"/> standard
<input checked="" type="checkbox"/> lonestar.tacc.teragrid.org	<input type="checkbox"/> development <input type="checkbox"/> high <input checked="" type="checkbox"/> normal <input type="checkbox"/> serial
<input checked="" type="checkbox"/> nstg.ornl.teragrid.org	<input checked="" type="checkbox"/> batch
<input checked="" type="checkbox"/> pople.psc.teragrid.org	<input type="checkbox"/> batch_1 <input checked="" type="checkbox"/> batch_r <input type="checkbox"/> debug
<input checked="" type="checkbox"/> ranger.tacc.teragrid.org	<input type="checkbox"/> development <input type="checkbox"/> large <input type="checkbox"/> long <input checked="" type="checkbox"/> normal <input type="checkbox"/> serial

Describe your job:

Processing cores:

Requested wall time: : (hours:minutes)

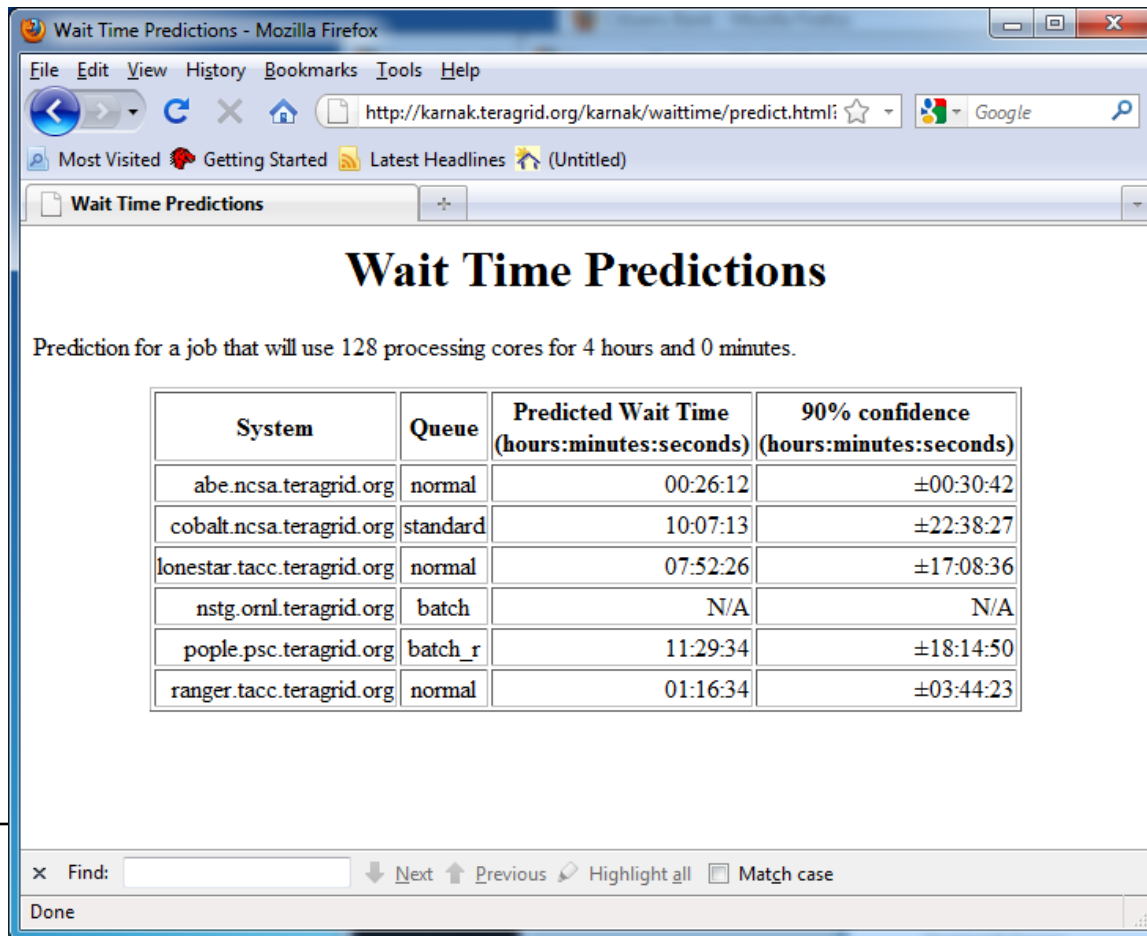
Confidence interval size: %

Find: Next Previous Highlight all

Done

Prediction for a Potential Job

- Prediction provided for each system/queue
 - N/A if the job can't be submitted to a queue



Wait Time Predictions - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://karnak.teragrid.org/karnak/waittime/predict.html

Wait Time Predictions

Wait Time Predictions

Prediction for a job that will use 128 processing cores for 4 hours and 0 minutes.

System	Queue	Predicted Wait Time (hours:minutes:seconds)	90% confidence (hours:minutes:seconds)
abe.ncsa.teragrid.org	normal	00:26:12	±00:30:42
cobalt.ncsa.teragrid.org	standard	10:07:13	±22:38:27
lonestar.tacc.teragrid.org	normal	07:52:26	±17:08:36
nstg.oml.teragrid.org	batch	N/A	N/A
pople.psc.teragrid.org	batch_r	11:29:34	±18:14:50
ranger.tacc.teragrid.org	normal	01:16:34	±03:44:23

Find: Next Previous Highlight all Match case

Done

Service Interfaces

- REST-style interactions
- HTTP is the base protocol
- A few options for encoding data
 - HTML
 - Simple web interface described previously
 - Plain text
 - Supports thin command line clients
 - XML
 - Supports integration with tools
 - JSON (partial)
 - Supports integration with web-based tools

Command Line Programs

- Roughly follows the web pages
 - `ksystems` – system summary
 - `ksystem` – jobs in queues on a system
 - `kqueue` – current and historical job information
 - `kjobs` – info about waiting jobs
 - `kwait`, `kstart` – predict waiting job
 - `kwouldwait`, `kwouldstart` – predict potential job
- Implemented in Python
- Provides output as HTML, XML, or text
 - JSON in the near future
- Downloadable via the TeraGrid Karnak web page

Potential Job Performance

- Post-analysis of performance
 - Event driven simulation
 - Job data recorded from TeraGrid systems
- Workload has 2 event types
 - Request a prediction for each job using system state just before it is submitted
 - Insert information about a job wait time as the job starts
- Process:
 - Perform insertions only for X days
 - Perform insertions and predictions for the next Y days
 - Output performance
- Optimized prediction parameters

Potential Job Workloads

	Train			Evaluate			
System	Start	End	Inserts	Start	End	Inserts	Predicts
Ranger	8/15	8/23	6,683	8/23	9/1	15,093	15,210
Abe	8/15	8/23	39,700	8/23	9/1	42,527	43,021
Lonestar	8/15	8/23	16,131	8/23	9/1	36,087	36,047
Cobalt	5/1	8/1	15,690	8/1	9/1	5,976	5,871
Pople	5/1	8/1	8,285	8/1	9/1	2,387	2,402

Potential Jobs

System	Number of Jobs	Mean Wait Time (hours)	Accuracy	Mean Confidence Interval (hours)	Percent in 90% Confidence Interval
Ranger	15,210	2.63	0.43	10.74	80.16
Abe	43,021	0.83	0.57	0.53	80.38
Lonestar	36,047	1.12	0.53	1.51	79.05
Cobalt	5,871	7.14	0.39	12.00	79.49
Pople	2,402	6.92	0.34	15.86	85.55

$$accuracy = \begin{cases} 1 & \text{if } E_{wt} = A_{wt} \\ A_{wt}/E_{wt} & \text{if } E_{wt} > A_{wt} \\ E_{wt}/A_{wt} & \text{if } E_{wt} < A_{wt} \end{cases}$$

Potential Jobs Discussion

- Accuracies are lower than we would like
 - Seem to be in line with other methods
 - e.g. QBETS
 - Difficult to compare
 - Different workloads
 - Different metrics
 - Think we can improve accuracies a bit more
- Confidence intervals need to be improved

Potential Jobs Discussion II

- Can take some time to optimize prediction parameters
 - 24 hours on 16 cores for larger data sets
 - Different processes for each data set
- Started to develop configurations that are generally ok
 - Less time doing off-line optimizations
- Starting to investigate on-line optimization techniques
 - Adjust predictor as jobs start

Submitted Job Performance

- Post-analysis of performance
 - Event driven simulation
 - Job data recorded from TeraGrid systems
- Workload events
 - Request a prediction for each job using system state just after it is submitted
 - Insert information about a job wait time as the job starts
- Process:
 - Perform insertions only for X days
 - Perform insertions and predictions for the next Y days
 - Output performance
- Optimized prediction parameters

Submitted Job Workloads

	Train			Evaluate			
System	Start	End	Inserts	Start	End	Inserts	Predicts
Ranger	8/15	8/23	5,987	8/23	9/1	9,970	9,861
Abe	8/15	8/23	9,249	8/23	9/1	11,102	11,520
Lonestar	8/15	8/23	11,920	8/23	9/1	14,395	14,238
Cobalt	5/1	8/1	8,373	8/1	9/1	3,774	3,513
Pople	5/1	8/1	5,498	8/1	9/1	1,140	1,114

Submitted Jobs

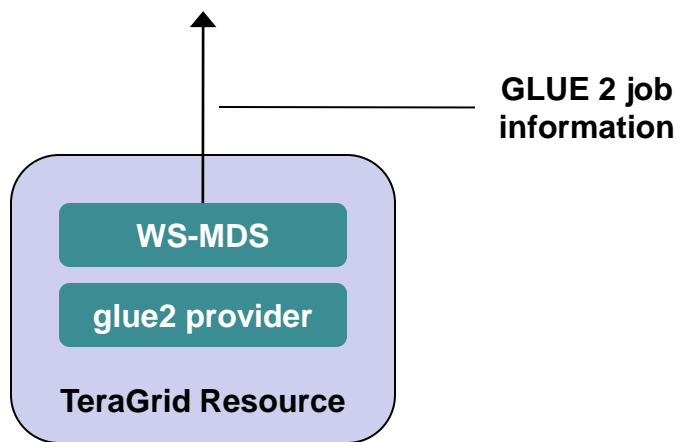
System	Number of Jobs	Mean Wait Time (hours)	Accuracy	Mean Confidence Interval (hours)	Percent in 90% Confidence Interval
Ranger	9,861	3.11	0.33	8.49	84.36
Abe	11,520	2.56	0.40	2.76	69.61
Lonestar	14,238	2.06	0.45	3.32	82.27
Cobalt	3,513	11.80	0.36	14.12	73.33
Pople	1,114	11.42	0.44	25.89	90.04

Submitted Jobs Discussion

- Fewer jobs in these workloads
 - Service does not see jobs as waiting if they start quickly
 - Service gets periodic snapshots of queue states from TeraGrid
- Slightly lower accuracy than for potential jobs
 - Probably due to jobs that start quickly being easier to predict
- Accuracies again lower than we would like
 - Think we can improve a bit
- Confidence intervals need to be improved

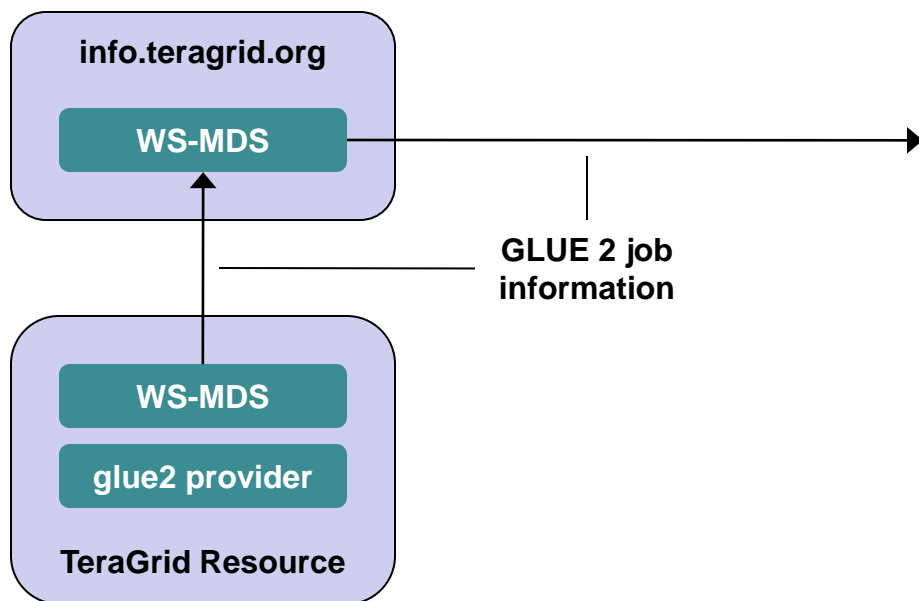
Implementation

- glue2 provider generates job information
 - Periodic snapshots of queued and running jobs
 - XML in a TeraGrid realization of the GLUE 2 schema
- WS-MDS picks up this information
- No GLUE 2 job information, no predictions



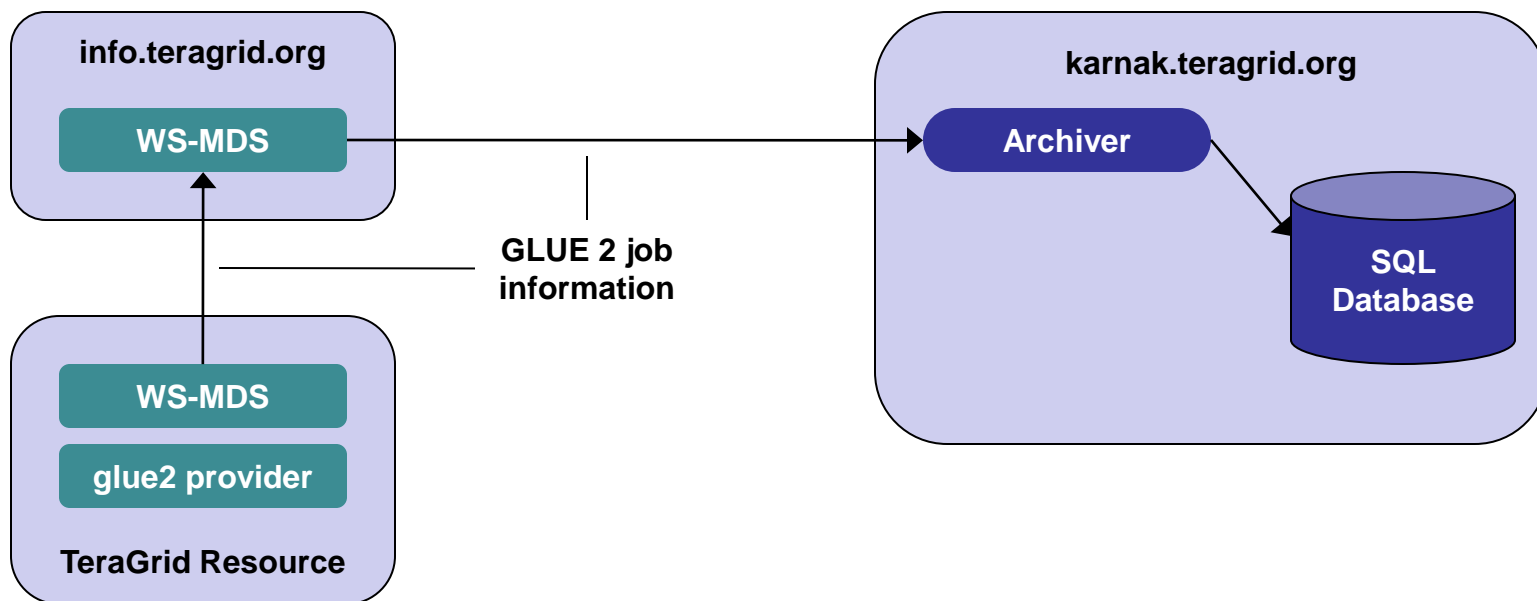
Implementation

- GLUE 2 job information published to centralized WS-MDS
 - Only authenticated access
 - Short list allowed to access



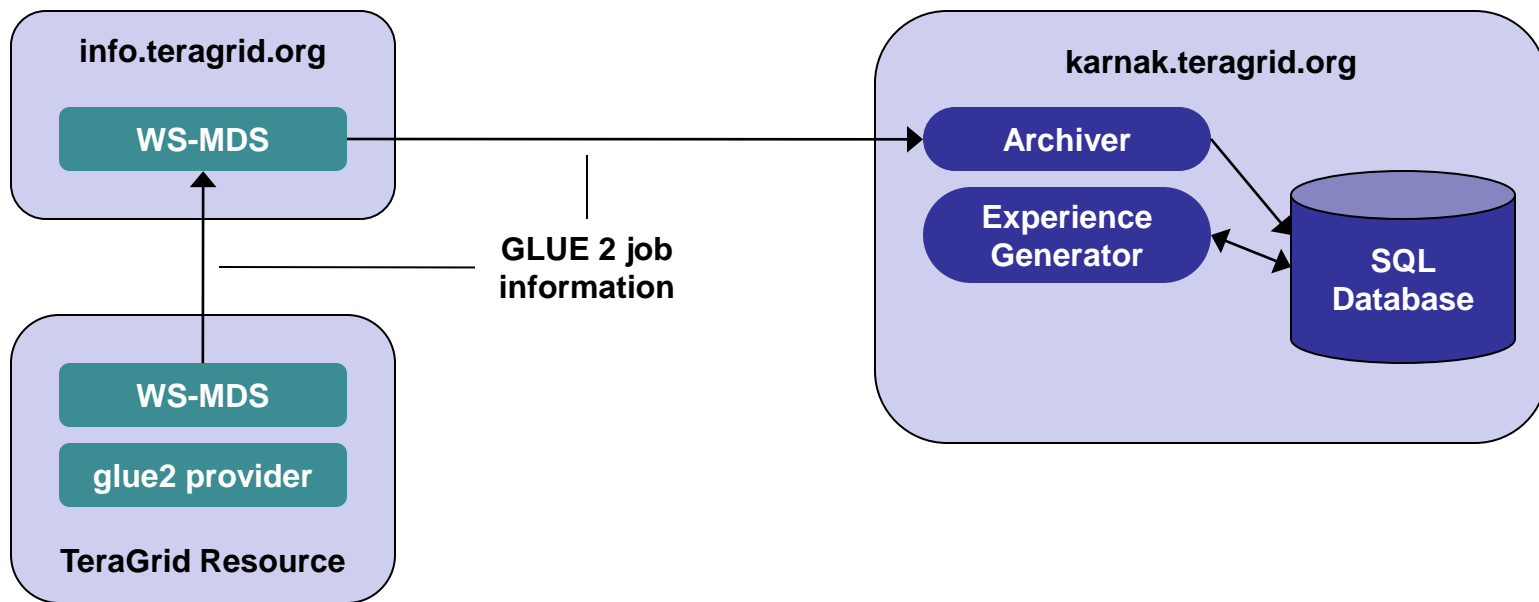
Implementation

- Archiver process
 - Polls WS-MDS for job information
 - Stores job ordering
 - Stores job information (cores, req. wall time, user, ...)
 - Generates events for each job (submit, start, complete)



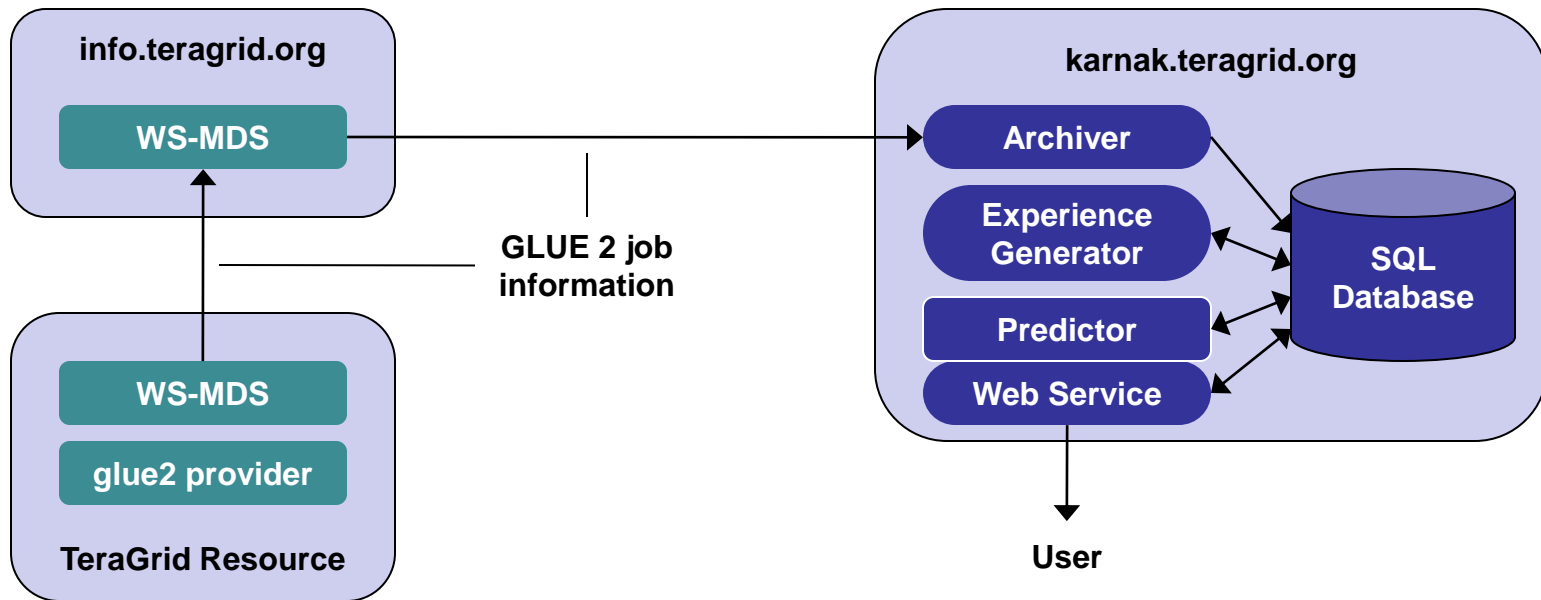
Implementation

- Experience Generator
 - Performs scheduling simulations and stores results
 - Constructs experiences
 - job description, queue position, simulated start time



Implementation

- Web service interacts with users
 - Job information from database
 - Embedded predictor to form predictions



Current Status

- Beta version available at <http://karnak.teragrid.org>
- Harder than I expected!
 - Fancy prediction algorithms existed so were easy
 - Turning queue snapshots into experiences
 - Cleaning data
 - Handling gaps in data
 - Within a constrained resource environment

Future Work

- Improve quality of predictions
- Improve optimization process
- Get more TeraGrid resource providers to publish job information
- Move service into production on TeraGrid
- Enhance service to meet user needs
 - User-specified searches of historical job data
 - Provide additional predictions such as job run times and file transfer times