

Factory Ops Site Debugging

This talk shows detailed examples of how we debug site problems

By Jeff Dost (UCSD)

Overview

- Validation
- Rundiff
- Held
- Waiting
- Pending
- Unmatched

Overview

- **Validation**
- Rundiff
- Held
- Waiting
- Pending
- Unmatched

Validation

- `analyze_entries` shows 100% validation failure, mostly affecting UCSDCMS:

CMS_T2_US_Purdue_hadoop	str	fval	0job		val	idle	wst	badp		waste	time	total
	100%	100%	100%		100%	0%	100%	100%		72	72	207

- Use `entry_ls` to find logs:

```
$ entry_ls -fo CMS_T2_US_Purdue_hadoop fecmsucsd 20140605
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T2_US_Purdue_hadoop/job.2193146.0.out
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T2_US_Purdue_hadoop/job.2193193.0.out
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T2_US_Purdue_hadoop/job.2193720.0.out
```

- See if validation script generated an xml report:

```
$ cat_XMLResult.py
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T2_US_Purdue_hadoop/job.2193720.0.out
| less
```

Validation

- Looks like grid-proxy-init is not found:

```
<result>
  <status>ERROR</status>
  <metric name="TestID" ts="2014-06-03T13:20:56-04:00" uri="local">main/setup_x509.sh</metric>
  <metric name="failure" ts="2014-06-03T13:20:29-04:00" uri="local">WN_Resource</metric>
  <metric name="command" ts="2014-06-03T13:20:35-04:00" uri="local">grid-proxy-init</metric>
</result>
<detail>
  Validation failed in main/setup_x509.sh.

  grid-proxy-init command not found in path!
</detail>
```

- Find affected worker nodes:

```
$ entry_ls -fo CMS_T2_US_Purdue_hadoop fecmsucsd 20140605 | xargs xres | get_wns | sort | uniq
cms-e002.rcac.purdue.edu
```

- Problem seems isolated to a single node
- Open a ticket on site and report the error and affected WN hostname

Validation

- `analyze_entries` shows 100% validation failure, mostly affecting UCSDCMS:

CMS_T3_US_FIU_fiupg	str	fval	0job		val	idle	wst	badp		waste	time	total
	100%	100%	100%		100%	0%	100%	100%		87	87	260

- Use `entry_ls` to find logs:

```
$ entry_ls -fo CMS_T3_US_FIU_fiupg fecmsucsd 20140615
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T3_US_FIU_fiupg/job.2094373.0.err
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T3_US_FIU_fiupg/job.2094384.0.err
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T3_US_FIU_fiupg/job.2094520.0.err
```

- See if validation script generated an xml report:

```
$ cat_XMLResult.py
/home/gfactory/glideinsubmit/glidein_v3_0/client_log/user_fecmsucsd/entry_CMS_T3_US_FIU_fiupg/job.2098206.0.out |
less
```

Validation

- Unfortunately this one has no report:

```
<result>
  <status>ERROR</status>
  <metric name="TestID" ts="2014-06-15T22:53:41-04:00" uri="local">client/discover_CMSSW.sh</metric>
</result>
<detail>
  Validation failed in client/discover_CMSSW.sh.

  The test script did not produce an XML file. No further information available.
</detail>
```

- If no report the next best thing is to open the logs and look for any contextual info:

```
Signature OK for client:discover_CMSSW.e51fKp.sh.
cmsset_default.sh not found!\n
Looked in /cmsset_default.sh
and /osg/apps/cmssoft/cms/cmsset_default.sh
and /cms.cern.ch/cmsset_default.sh
and /cvmfs/cms.cern.ch/cmsset_default.sh
=== Validation error in /var/opt/condor/execute/dir_12458/glide_d12512/client/discover_CMSSW.sh ===
```

Validation

- In this case it is due to the glidein failing to find CMS Software. It was likely a misconfiguration at the site.
- If it is a FE validation script and you don't know how to interpret it, ask me or Igor
- If we don't know either, then we email the FE admin to ask for help

Validation

- Get list of worker nodes where failures occurred:

```
$ entry_ls -fo CMS_T3_US_FIU_fiupg fecmsucsd 20140615 | xargs xres | get_wns | sort | uniq  
compute-0-10.local  
compute-0-11.local  
compute-0-13.local  
compute-0-15.local  
compute-0-16.local
```

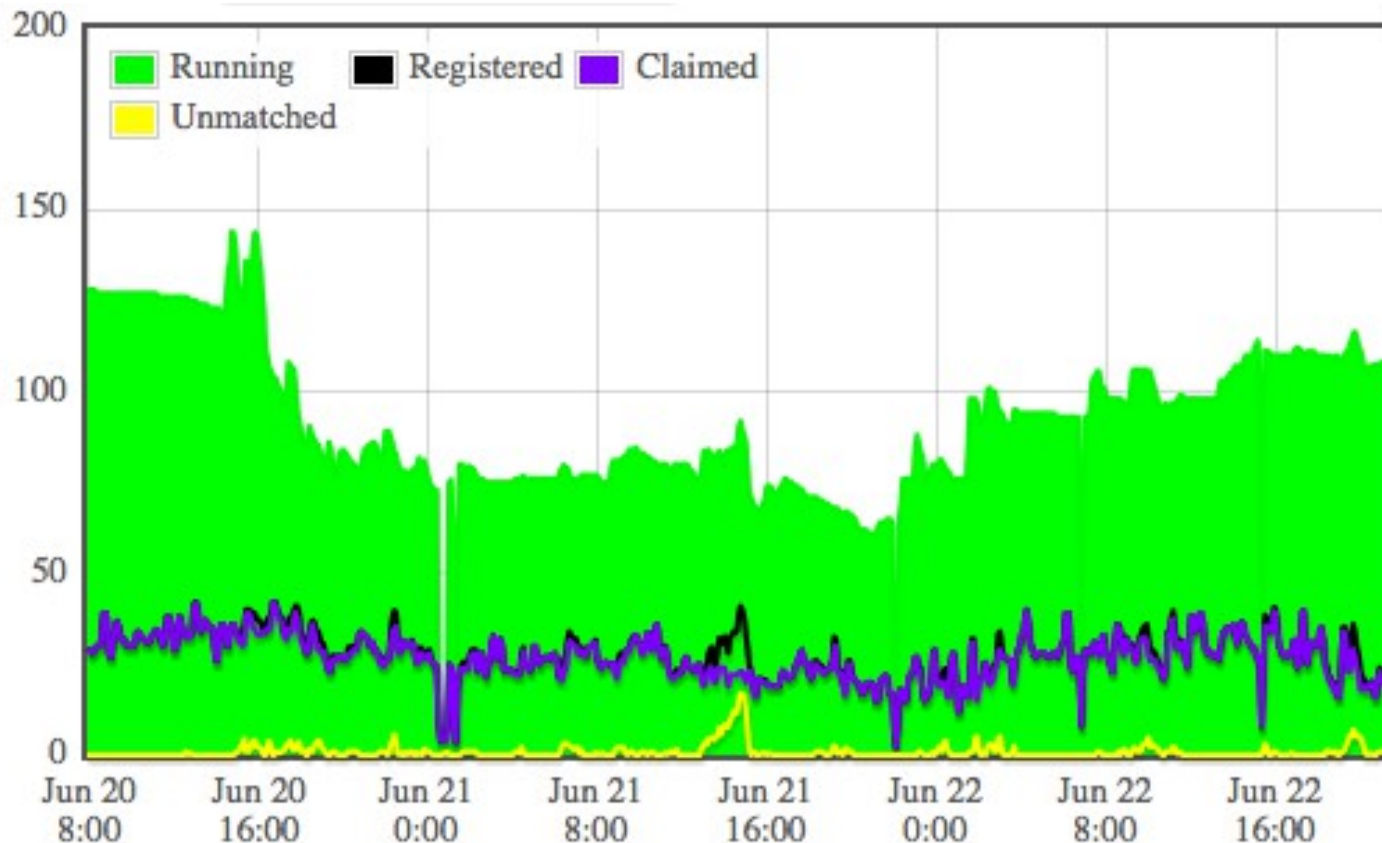
- Open a ticket with the site and provide the error and list of affected nodes

Overview

- Validation
- **Rundiff**
- Held
- Waiting
- Pending
- Unmatched

Rundiff

- CMS_T1_UK_RAL_arc_ce01- Registered significantly lower than running on factoryStatus



Rundiff

- Run **entry_q** to ensure they aren't just stale glideins (runtime >> maxwalltime)

```
$ entry_q CMS_T1_UK_RAL_arc_ce01 | head
```

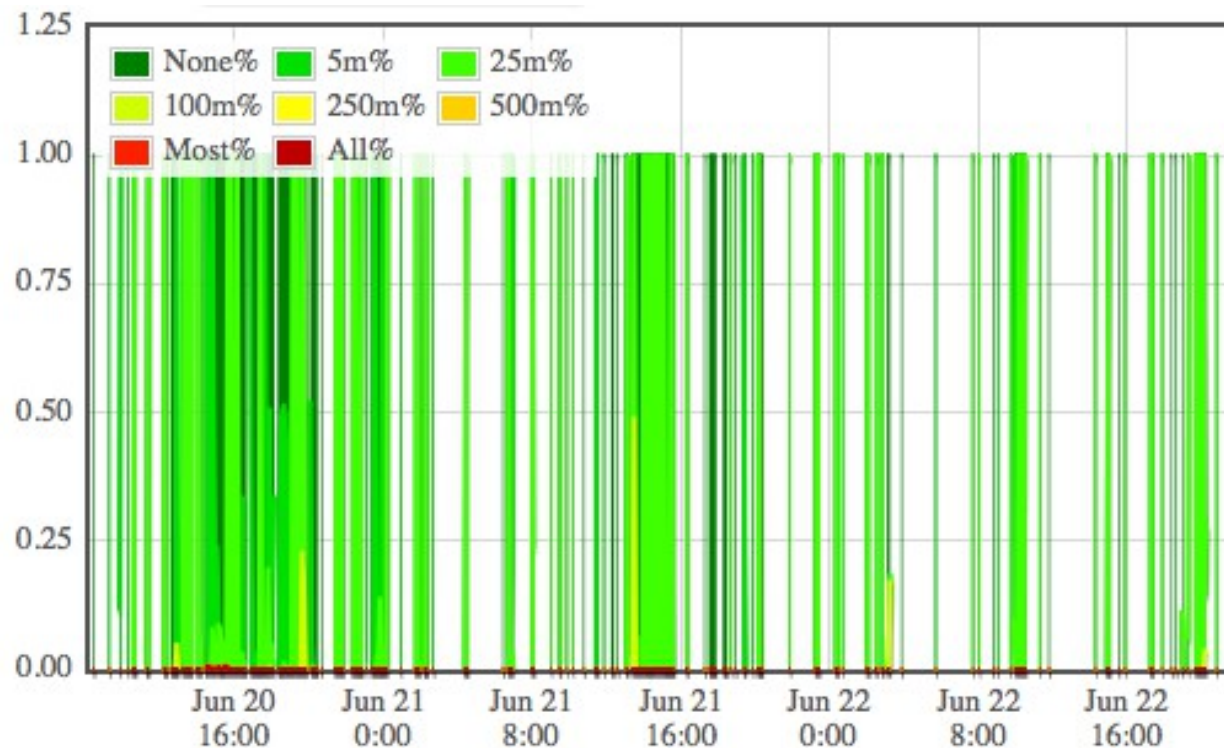
```
-- Schedd: schedd_glideins2@gfactory-1.t2.ucsd.edu : <169.228.38.36:50104>
```

ID	OWNER	SUBMITTED	RUN_TIME	ST	PRI	SIZE	CMD
2463093.6	fecmscern	6/18 11:33	2+07:21:49	R	0	0.0	glidein_startup.sh
2463106.0	fecmscern	6/18 11:41	2+07:21:49	R	0	0.0	glidein_startup.sh
2463106.1	fecmscern	6/18 11:41	2+07:18:38	R	0	0.0	glidein_startup.sh

Rundiff

- CompletedStats don't show obvious validation errors

Fraction wasted due to node validation



Rundiff

- Pick a few glideins and look at StartdLog:

```
$ cat_StartdLog.py job.2469098.0.err | grep '^06' | less
...
06/20/14 14:42:10 (pid:5080) condor_write(): Socket closed when trying to write 4096 bytes to collector
vocms97.cern.ch:9938, fd is 9
06/20/14 14:42:10 (pid:5080) Buf::write(): condor_write() failed
```

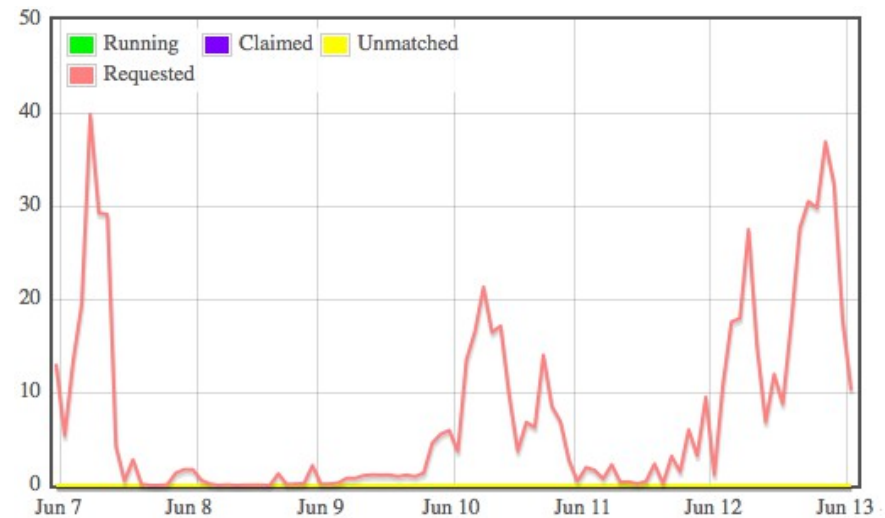
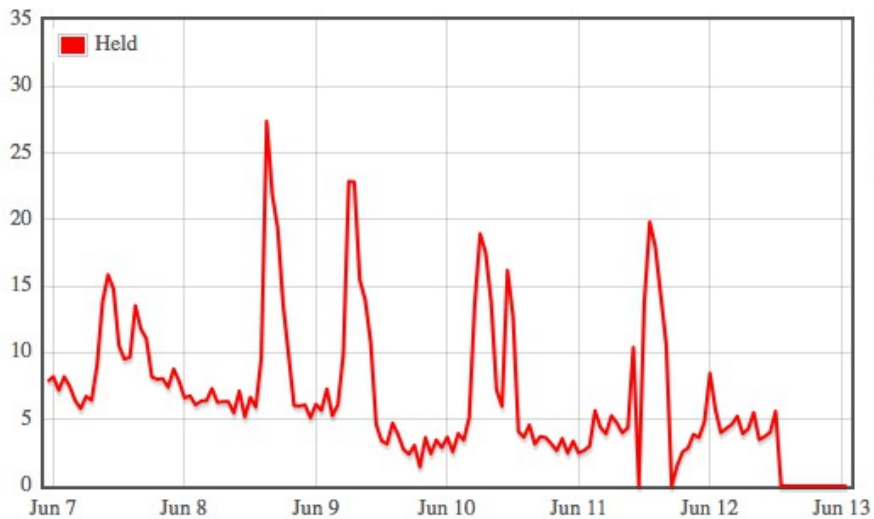
- TCP connections between the glidein (worker node) and User Collector are dropping
- Because this is not happening to glideins from the same VO at other sites it is likely occurring on the site side
- Common causes are NAT or firewall issues
- Open ticket with site and include found error

Overview

- Validation
- Rundiff
- **Held**
- Waiting
- Pending
- Unmatched

Held

- All UCSDCMS glideins going held at CMS_T3_US_FIU_fiupg



Held

- Check hold reason:

```
$ entry_q CMS_T3_US_FIU_fiupg -held
-- Schedd: schedd_glideins3@gfactory-1.t2.ucsd.edu : <169.228.38.36:46438>
ID      OWNER      HELD_SINCE  HOLD_REASON
2236925.0  fecmsucsd    5/27 21:14  Globus error 7: authentication with the remote server failed
2236935.0  fecmsucsd    5/27 21:09  Globus error 7: authentication with the remote server failed
2236938.1  fecmsucsd    5/27 21:09  Globus error 7: authentication with the remote server failed
```

- Site seems to be rejecting proxy. Check to make sure it hasn't expired
- Find the proxy on disk:

```
entry_q CMS_T3_US_FIU_fiupg 2236938.1 -format '%s\n' x509userproxy
/var/gfactory/clientproxies/user_fecmsucsd/glidein_v3_0/credential_UCSD-v1_0.main_160364
```

Held

- Proxy has not expired, must be a CE problem:

```
$ proxy_info fecmsucsd info -all /var/gfactory/clientproxies/user_fecmsucsd/glidein_v3_0/credential_UCSD-  
v1_0.main_160364  
subject   : /DC=ch/DC=cern/OU=computers/CN=cmspilot00/vocms0167.cern.ch/CN=proxy  
issuer    : /DC=ch/DC=cern/OU=computers/CN=cmspilot00/vocms0167.cern.ch  
identity  : /DC=ch/DC=cern/OU=computers/CN=cmspilot00/vocms0167.cern.ch  
type      : proxy  
strength  : 1024 bits  
path      : /var/gfactory/clientproxies/user_fecmsucsd/glidein_v3_0/credential_UCSD-v1_0.main_160364  
timeleft  : 71:55:06  
key usage : Digital Signature, Key Encipherment  
=== VO cms extension information ===  
VO        : cms  
subject   : /DC=ch/DC=cern/OU=computers/CN=cmspilot00/vocms0167.cern.ch  
issuer    : /DC=ch/DC=cern/OU=computers/CN=lcg-voms.cern.ch  
attribute : /cms/Role=pilot/Capability=NULL  
attribute : /cms/Role=NULL/Capability=NULL  
attribute : /cms/dcms/Role=NULL/Capability=NULL  
attribute : /cms/escms/Role=NULL/Capability=NULL  
attribute : /cms/itcms/Role=NULL/Capability=NULL  
timeleft  : 71:55:06  
uri       : lcg-voms.cern.ch:15002
```

- Open ticket with site with hold error and proxy info

Overview

- Validation
- Rundiff
- Held
- **Waiting**
- Pending
- Unmatched

Waiting

- CMS_T2_US_Purdue_cmstest1 – 100% Waiting, 0 Running

Entry Name		Running	Idle	Waiting	Pending	Staging in	Staging out	Unknown	Held
CMS_T2_US_Purdue_cmstest1	↑	0	124	124	0	0	0	0	0

- condor_activity log reveals CE is unreachable:

```
026 (1555490.000.000) 06/23 10:37:45 Detected Down Grid Resource
GridResource: gt5 cmstest1.rcac.purdue.edu:2119/jobmanager-condor
...
020 (1555490.001.000) 06/23 10:37:45 Detected Down Globus Resource
RM-Contact: cmstest1.rcac.purdue.edu:2119/jobmanager-condor
...
026 (1555490.001.000) 06/23 10:37:45 Detected Down Grid Resource
GridResource: gt5 cmstest1.rcac.purdue.edu:2119/jobmanager-condor
...
```

Waiting

- CMS_T2_US_Purdue_cmstest1 – 100%
Waiting, 0 Running

Entry Name		Running	Idle	Waiting	Pending	Staging in	Staging out	Unknown	Held
CMS_T2_US_Purdue_cmstest1	↑	0	124	124	0	0	0	0	0

- Confirm this with telnet:

```
$ telnet cmstest1.rcac.purdue.edu 2119
Trying 128.211.140.100...
telnet: connect to address 128.211.140.100: Connection refused
```

Waiting

- Look for downtime notice in RSV:

Purdue-Hadoop-TestCE CE

The screenshot shows a monitoring interface for 'Purdue-Hadoop-TestCE CE'. It features a notification box with a question mark icon stating: 'Status changed to UNKNOWN at Wed Jun 11 2014 15:37:11 GMT-0700 (PDT) 4 of 4 critical metrics have expired.' Below this is a 'Critical Metrics' section with a red 'X' icon and the text: 'CACert Expiry At Tue Jun 10 2014 09:15:11 GMT-0700 (PDT) (Expired)'. A detailed message below reads: 'Condor-G submission failed because the remote side is down. Make sure that the resource you are trying to monitor is online.'

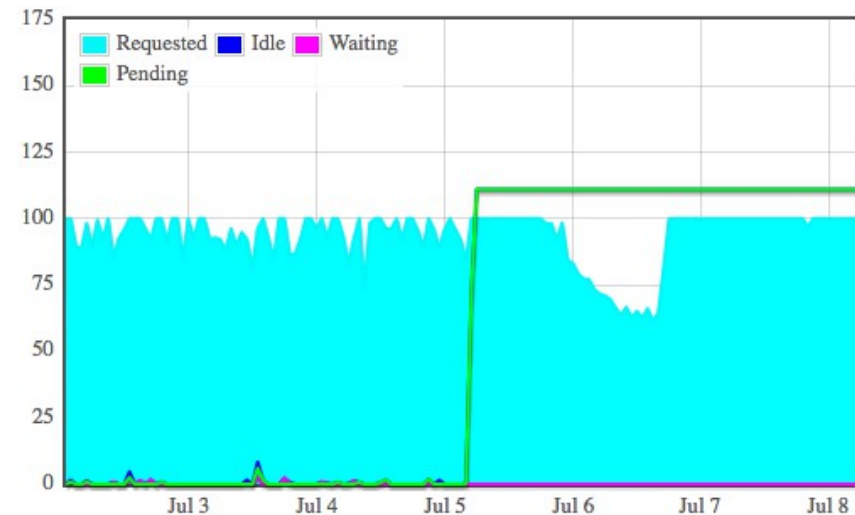
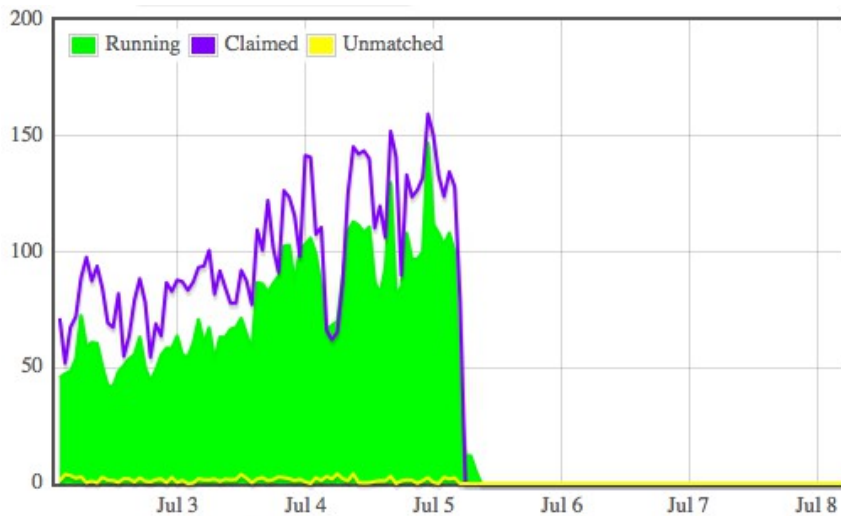
- Site in UNKNOWN status, lots of red X failures, and no sign of planned maintenance
- Open ticket on site, and provide downtime message, RSV link, and how long it has been down

Overview

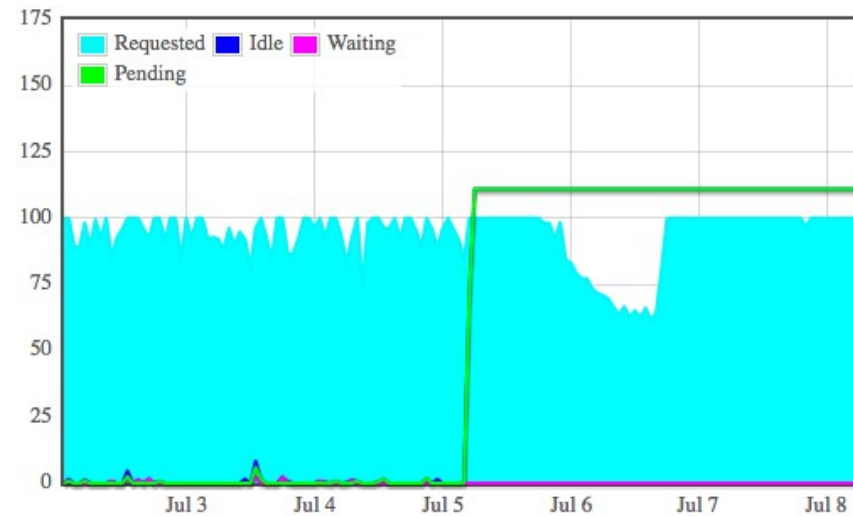
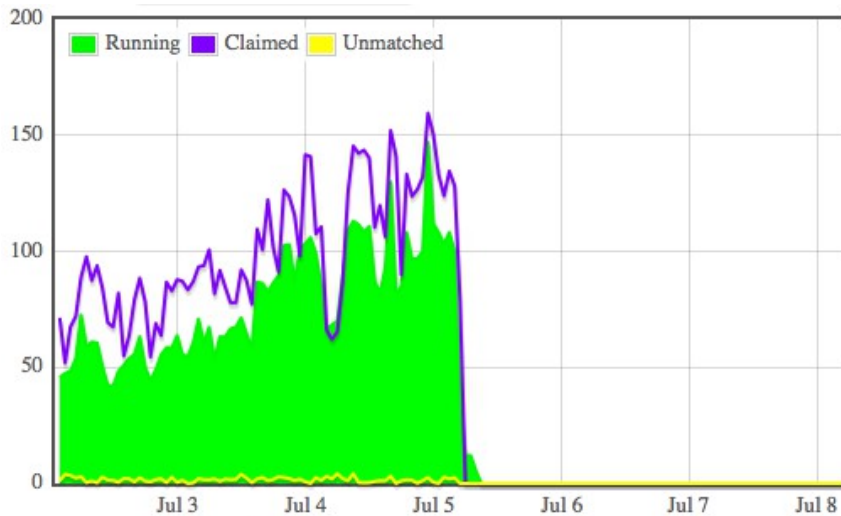
- Validation
- Rundiff
- Held
- Waiting
- **Pending**
- Unmatched

Pending

- OSG_US_Buffalo_u2-grid abruptly stopped working and went 100% pending



Pending



- The constant 0 slope of Pending is suspicious
- Typically this means we are no longer getting accurate updates back from the gatekeeper

Pending

- condor_activity shows no obvious issues other than jobs never getting past pending stage:

```
000 (1595555.009.000) 07/05 11:31:54 Job submitted from host: <129.79.53.27:32667>
...
017 (1595555.009.000) 07/05 11:32:07 Job submitted to Globus
  RM-Contact: u2-grid.ccr.buffalo.edu:2119/jobmanager-pbs
  JM-Contact: u2-grid.ccr.buffalo.edu:2119/jobmanager-pbs
  Can-Restart-JM: 1
...
027 (1595555.009.000) 07/05 11:32:07 Job submitted to grid resource
  GridResource: gt2 u2-grid.ccr.buffalo.edu:2119/jobmanager-pbs
  GridJobId: gt5 u2-grid.ccr.buffalo.edu:2119/jobmanager-pbs https://u2-
grid.ccr.buffalo.edu:52378/16433977007692051451/5166346122336875697/
...
```

- Try removing pending glideins to see if fresh ones start

Pending

- First find cluster id of last pending glidein:

```
entry_q OSG_US_Buffalo_u2-grid
```

```
-- Schedd: schedd_glideins3@glidein.grid.iu.edu : <129.79.53.27:32667>
ID      OWNER      SUBMITTED  RUN_TIME ST PRI SIZE CMD
1607244.0 feosgflock  7/5 16:04  0+00:00:00 I 0  0.0 glidein_startup.sh
1607244.1 feosgflock  7/5 16:04  0+00:00:00 I 0  0.0 glidein_startup.sh
1607244.2 feosgflock  7/5 16:04  0+00:00:00 I 0  0.0 glidein_startup.sh
...
1607268.9 feosgflock  7/5 16:11  0+00:00:00 I 0  0.0 glidein_startup.sh
```

- Remove all glideins up to and including this one:

```
entry_rm OSG_US_Buffalo_u2-grid -const 'clusterid<=1607268'
Jobs matching constraint (((GlideinFactory=?="OSGGOC")&&(GlideinName=?="v3_0")&&(GlideinEntryName=?="OSG_US_Buffalo_u2-grid"))&&(clusterid<=1607268)) have been marked for removal
```

Pending

- In this case, removing did the trick, and fresh glideins started running normally
- If that still doesn't work, open ticket with site, and provide pilot proxy info, as well as a few grid job ids along with submit times seen in activity log, e.g.:

```
000 (1595555.009.000) 07/05 11:31:54 Job submitted from host: <129.79.53.27:32667>
017 (1595555.009.000) 07/05 11:32:07 Job submitted to Globus
027 (1595555.009.000) 07/05 11:32:07 Job submitted to grid resource
    GridResource: gt2 u2-grid.ccr.buffalo.edu:2119/jobmanager-pbs
    GridJobId: gt5 u2-grid.ccr.buffalo.edu:2119/jobmanager-pbs https://u2-
grid.ccr.buffalo.edu:52378/16433977007692051451/5166346122336875697/
...
```

When Gridmanager Loses Track

- In the last example the gridmanager no longer was receiving updates from the CE
- One useful trick to determine this is by comparing the entry on this factory with the same entry on other factories
- If the problem is only observed on one factory, it is usually safe to assume it is just gridmanager ↔ CE issues
- In this case, the glideins can be considered “stale” and can be safely removed

Overview

- Validation
- Rundiff
- Held
- Waiting
- Pending
- **Unmatched**

Unmatched

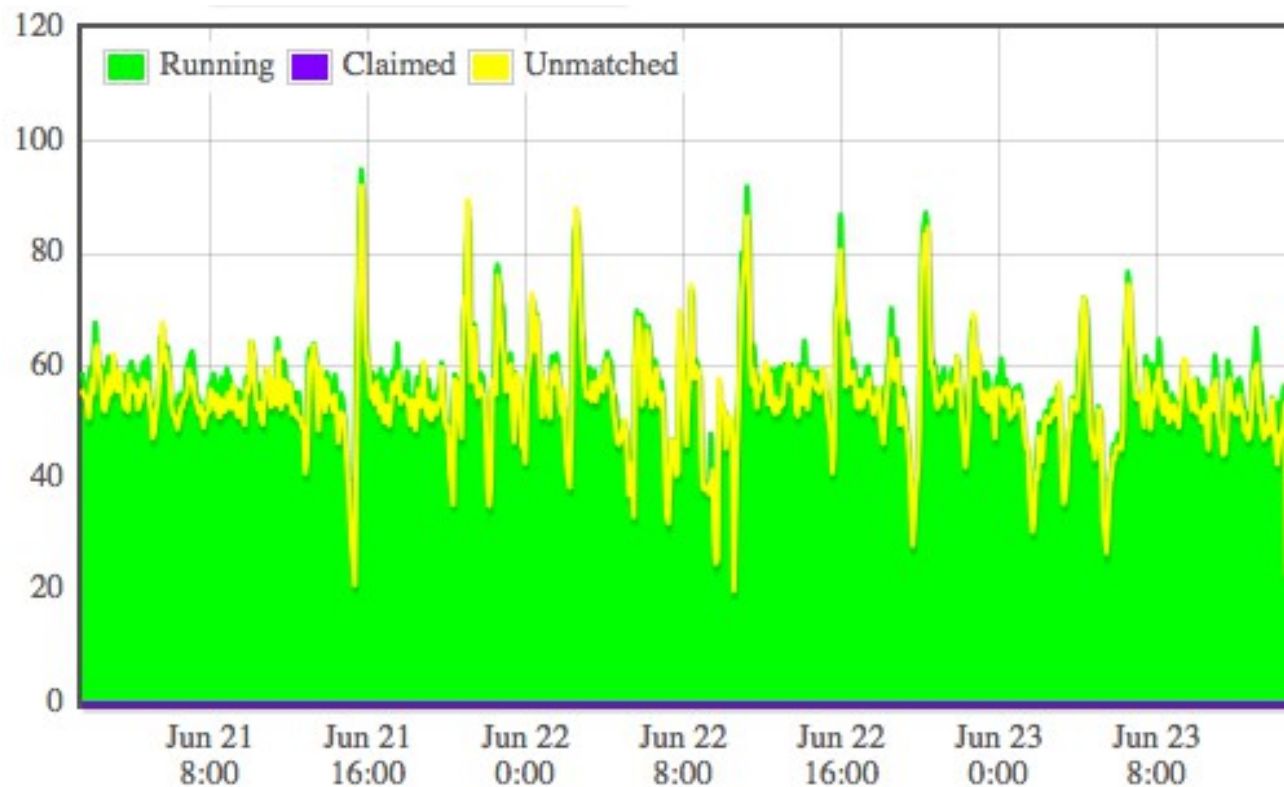
- Fermilab FE 100% waste in analyze_entries, no glideins running user jobs (100% idle):

	strt	fval	0job		val	idle	wst	badp		waste	time	total
CMS_T2_US_Nebraska_Red_gw1	0%	0%	100%		0%	100%	100%	100%		1259	1259	3296
CMS_T2_US_Nebraska_Red_gw2	0%	0%	100%		0%	100%	100%	100%		1181	1181	3095

- But condor isn't failing and there are no validation errors. Next, look at factoryStatus

Unmatched

- FactoryStatus shows Fermilab glideins are 100% Unmatched at Nebraska



Unmatched

- At this point it is likely a Frontend problem
- Glideins are being submitted on behalf of the FE users at Nebraska
- ...But the actual user jobs are not matching the glideins that started up
- Contact the FE admin and ask them to investigate their glidein start expressions
 - Give the list of sites we see this on, and for how long it has been happening