Site Debugging

by Jeff Dost (UCSD)

Types of Issues we debug

- Validation Errors
- Held Glideins
- RunDiff
- Authentication issues
- Glideins not starting
- Black Hole WNs
- Communication Errors
- Internal errors

Validation Errors

- Detected with analyze_entries daily email report
- Percent validation errors are listed per site over last 24 hours

```
frontend_UCSDCMS_cmspilot:

                                 strt fval 0job |  val idle  wst badp |  waste     time   total
CMS_T3_MX_Cinvestav_proton_work   40%   7%  78% |   8%  17%  25%  94% |     77      306  |  155
CMS_T2_US_Nebraska_Red_gw1        35%   1%  89% |   2%   5%   7%  90% |     70      923  |  214
CMS_T2_US_Nebraska_Red_gw2        50%   3%  93% |   4%   7%  10%  86% |     60      557  |  210
CMS_T2_US_Nebraska_Red            34%   0%  80% |   1%   4%   5%  49% |     45      902  |  146
```

Validation Errors

- Validation error message can typically be found in glidein stderr / stdout logs
- Worker node hostname is listed in stdout log

Handling Validation Errors

- Most of the time a validation error identifies a problem with the expected environment setup on the workernode
- If this is the case we open a GOC or Savannah ticket
  - CMS prefers we open Savannah tickets whenever possible if it was a CMS glidein
- Typically validation errors are due to simple misconfiguration. Admins quickly fix the issue once notified.

Held Glideins

- Debugging held glideins is generally more difficult
- Hard to determine if a glidein is held due to a local problem or a problem at the remote side at the site
- GRAM hold reasons are well defined but CREAM hold reasons are difficult to interpret

Factory Held Glidein Removal Policy

- Glideins held with particular GRAM hold reasons are known to never be recoverable
  - It is safe to remove these
- The factory constantly checks for these GRAM errors and just removes the stuck glideins if possible.
- The factory doesn't touch held glideins it doesn't know what to do with
  - We have to check these manually and act if needed

*__note__ there's no CREAM equivalent

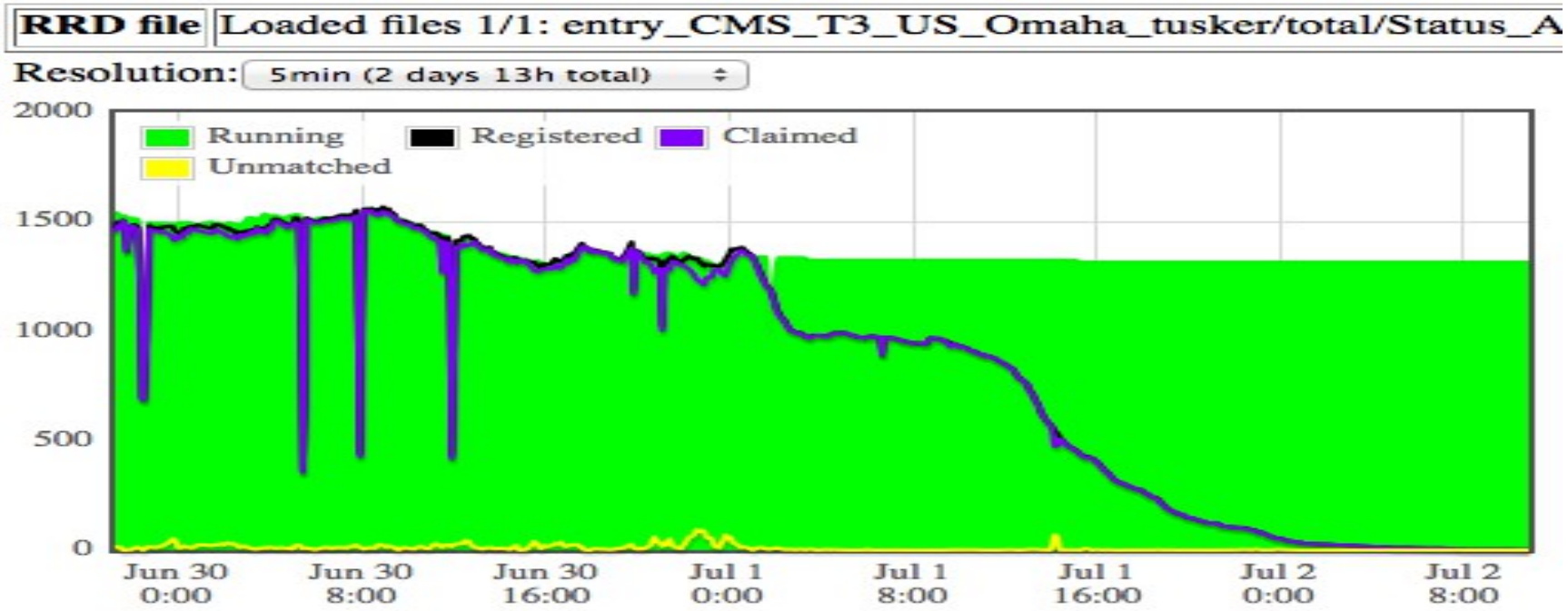
Globus Hold Reasons

Globus Error Code	Held Reason	Job is Recoverable
10	globus_xio_gsi: Token size exceeds limit. Usually happens when someone tries to establish a insecure connection with a secure endpoint, e.g. when someone sends plain HTTP to a HTTPS endpoint without	No
121	the job state file doesn't exist	No
126	it is unknown if the job was submitted	Yes
12	the connection to the server failed (check host and port)	Yes
131	the user proxy expired (job is still running)	Maybe
17	the job failed when the job manager attempted to run it	No
22	the job manager failed to create an internal script argument file	No
31	the job manager failed to cancel the job as requested	No
3	an I/O operation failed	Yes
47	the gatekeeper failed to run the job manager	No

RunDiff

- Phenomenon where factory loses track of running glidiens
- Can be seen in web monitoring
  - Factory reports total number of glideins submitted that are still running at a site and haven't terminated
  - Frontend reports how many of the running glideins are registered (known at the user collector)
  - RunDiff = # Running - # Registered

RunDiff



- RunDiff is one of the hardest issues to debug because it requires digging to determine the real picture

RunDiff Causes

- If site has Condor jobmanager and has a policy that calls condor_hold
  - Gridmanager loses track because there is no GRAM state to represent “held” jobs
  - Even on release the glideins really go idle and eventually start up again, but the gridmanager never gets the message
  - From the factory point of view these jobs “run” indefinitely and never recover
  - Temporary solution – set periodic hold removal in glidein submit files for condor sites

RunDiff Causes

- Site comes back up out of downtime
- Factory thinks there are still glideins running on the queue from before the downtime and is unaware that the admins cleared them out
- If this looks obvious we just remove them
- It is not always easy to remove them because it is not always clear from factory which glideins are “RunDiff” glideins and which are registered

RunDiff Causes

- Network issues between glidein and user collector
- Typically one can see packet loss errors in glidein StartdLogs
- A common cause for this was if sites block UDP packets at firewall level
- We now have factories running 100% TCP connections so this is less common in practice
  - Still see on occasion if sites have other advanced firewall or packet filtering policies in place

Authentication Issues

- Two common issues:
  - Glidein proxy denied at site; Goes held with:

    Globus error 7: an authorization operation failed
    - Site may have changed VO policy
    - Pilot proxy lifetime may have expired
  - Factory not correctly whitelisting new frontend
    - Typically just human error on Frontend or Factory config during registration process

Glideins not starting

- First it helps to define two different types of idle:
  - Waiting – idle on the factory machine; hasn't made it to remote side yet
  - Pending – made it to remote side but is idle in the site job manager
  - The sum is the total number idle as reported in our monitoring:

		Status:							
Entry Name		Running	Idle	Waiting	Pending	Staging in	Staging out	Unknown	Held
CMS_T2_EE_Estonia_europa	↑	4	78	65	13	0	0	0	54
CMS_T3_US_UMiss_umiss001	↑	0	55	55	0	0	0	0	0

Glideins not starting

- Stuck Pending for a long time
  - Typically means a site issue, we open a ticket in this case
- Stuck Waiting for a long time
  - Glidein never made it to the site and is likely a local problem on the factory node
  - Can simply happen if site is down for maintenance
  - May mean site is permanently decommissioned
  - Sometimes gridmanagers just get stuck. In this case we remove old glideins and restart them

Black Hole WNs

- glidein stderr / stdout logs are empty
- It is likely a black hole if the glidein terminates minutes or even seconds after it started
  - Glideins try to run for at least 20 minutes even on failure
- We typically open a ticket to notify site and give relevant lines in condor logs
- Unfortunately no information can tell us which WN it occurs on since logs are empty

Communication Errors

- Between Factory and Frontend
  - Most common is authentication errors described earlier
  - Likely see errors in Factory CollectorLog
- Between glidein and User Pool
  - As discussed in RunDiff topic, likely to see errors in glidein StartdLog

Internal Errors

- Neither Condor nor GlideinWMS are perfect; we discover bugs from time to time
- We have working relationships with both Condor and GlideinWMS developers
- As soon as we can verify we have found a bug we contact the relevant support team
- Bug fixes are usually pushed within a few version updates
  - If not and fix is urgent, developers supply us with temporary patches