

Recent SAMGrid Problems Meeting

22 July 2008, 10:30-noon, FCC1, 88SAMDH

Present: Adam (typist), Robert, Steve Sherwood, Keith, Steve Timm, Parag, Andrew, Gabriele, Mike, Joel (virtually), Qizhong, Art, Heidi

Several problems with SAMGrid and the complete meltdown of production this past weekend triggered this meeting called by Adam.

Three problems were identified:

1) MC Jobs Held on the Queueing Node

A fraction Joel's jobs get held (never making it to the forwarding node) on the queueing node at various times of the day. Some of these held jobs are associated with the Schedd "blackout" while it is processing the periodic expressions, but other periods of held jobs cannot be easily explained. Many (most?) of these jobs are held with "Globus error 10" which is apparently caused by running out of available ports (but we don't think there is real evidence of running out of available ports). The rate at which jobs get held rose over the past week.

The current theory is that the queueing node becomes too busy to communicate effectively with all of its subordinate jobs running on the forwarding nodes. "Busy" here may mean too many jobs for Schedd to communicate with in a timely manner or network traffic on the machine making such communication wait (or keeping the CPU busy on other things). Parag changed these Schedd parameters on the queueing node (see DZSAM-156)...

```
GRIDMANAGER_JOB_PROBE_INTERVAL = 600
CONDOR_JOB_POLL_INTERVAL = 600
GRIDMANAGER_RESOURCE_PROBE_INTERVAL = 600
GRIDMANAGER_RESOURCE_PROBE_DELAY = 600
GRIDMANAGER_CONNECT_FAILURE_RETRY_COUNT = 10
GRIDMANAGER_MAX_PENDING_REQUESTS = 100
```

in order to make communications more resilient. It's a little too early to determine the effect of these changes, but jobs are still being held after the change.

ACTION ITEMS:

1.1) Andrew suggests a "no log file transfer" test, where we stop the transfer of log files from the forwarding nodes to the queueing node. The idea is to reduce the network and processing load on the queueing node to see if the held job rate decreases. The MC log files are large (2 GB!) and may be a source of problems. This item is DZSAM-160

assigned to Andrew.

1.2) Migrate to Condor 7. At least Condor 7 is not supposed to have the Periodic Expression blackout. This item is DZSAM-161. Steve sherwood thinks that d0srv047 is ready to be a test node. He'll check it out (subtask DZSAM-162).

2) Stale globus_jobmanager processes on the forwarding nodes

A globus_jobmanager process (gjp) on a forwarding node corresponds to a grid job on the queuing node. Each gjp takes one of the 100 slots available on the forwarding node (we limit the # of grid jobs to 100 per forwarding node). For some reason, some gjp's are no longer associated with a grid job and will live forever, taking up a slot. Over the course of a few weeks, the number of these stale gjp's can become large and effectively block a forwarding node. Furthermore, we seem to have periods where a large number of gjp's go stale and clog the forwarding node.

Our current mitigation is to run a script that removes gjp's older than 15 days (because the proxy runs out after 2 weeks, so no valid job can go beyond 15 days). But that does not work for the period where many gjp's go stale at once.

The reason why gjp's go stale is unknown. Theories are along the lines of problem #1 above - that for some reason the queuing node loses contact with the gjp and eventually gives up on it, putting the grid job in the held state. Maybe some callback is failing.

On Tuesday, a significant number of stale gjp's clogged samgfw01. Removing gjp's > 15 days old only cleaned up a few slots.

ACTION ITEMS

2.1) Come up with a smarter "stale gjp reaper" that looks at the queuing node to see if the parent grid job is marked as running. If it is not, then kill the gjp. Need to think how this affects Joel's jobs. This task is DZSAM-73 assigned to Adam.

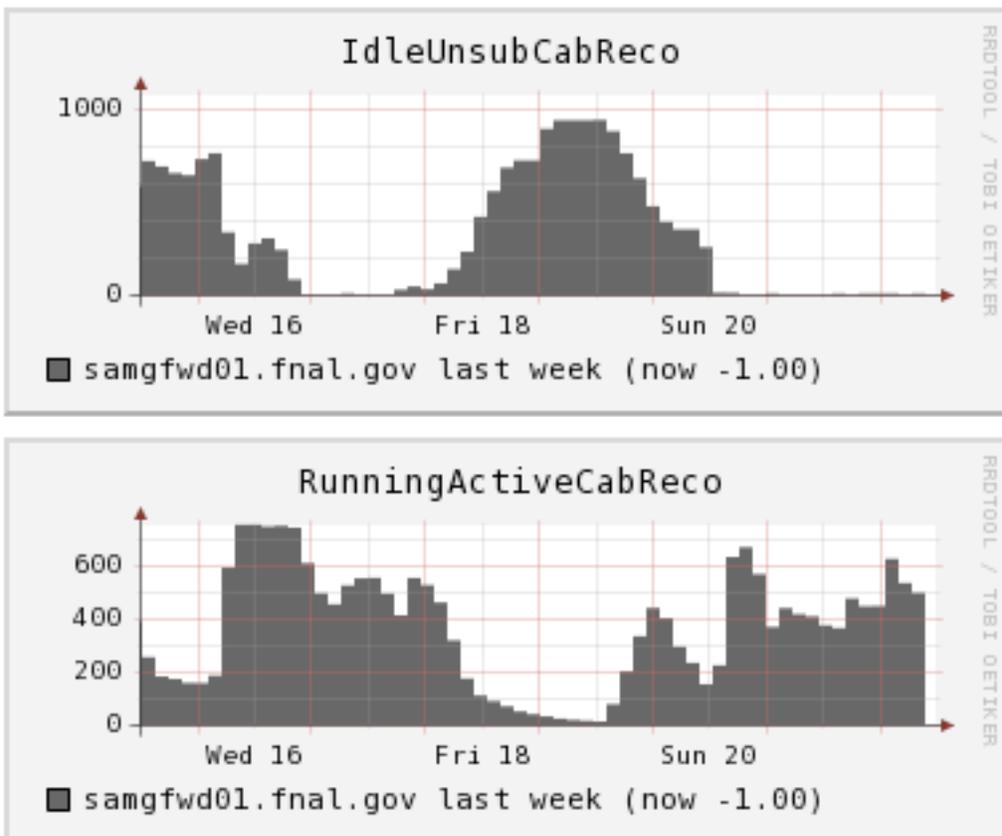
2.2) Mike does a bit of manual investigation for held reco jobs on the queuing node. Perhaps we can automate his check if a grid job actually has running batch jobs on CAB. DZSAM-171.

2.3) Turn up the logging level for Schedd on the queueing node to determine why it loses contact with gjp's. DZSAM-172

3) Forwarding nodes do not submit jobs to CAB even though they should

There were several periods last week where Idle jobs were sitting unsubmitted even though no limit had been reached. See plots below.

Problem was on samgfw01, but also appeared on samgfw02.



Usually idle jobs enter the system and then quickly switch to running when they are submitted to CAB. Big humps are bad, except for when there are already 750 jobs running on CAB (the MAX_GRID limit). But CAB was running few jobs in this case.

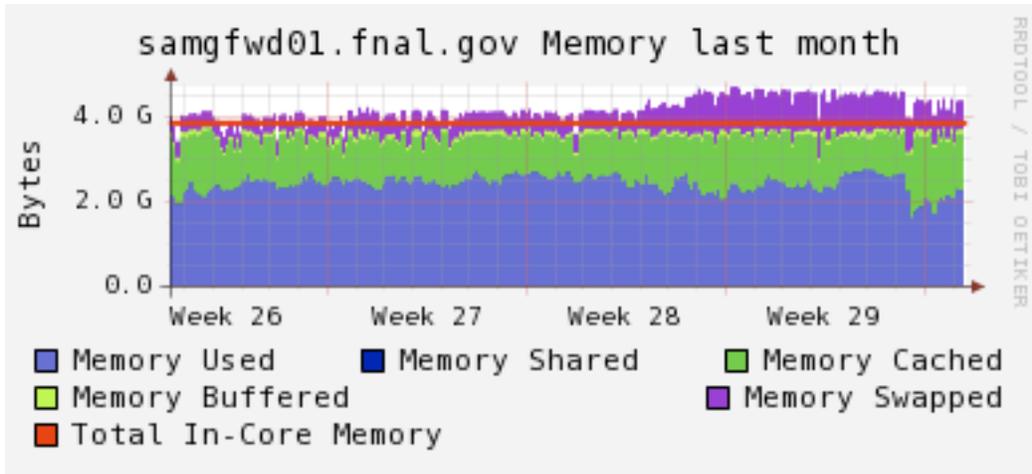
I "nuked" the system on Sunday (cleared Condor of all its information) on samgfw01,02. (See DZSAM-158).

Steve Timm could see the condor_gridmanager (cgm) considering idle jobs for submission, but then passed them over. There is no indication in the log as to why a job is passed over for submission. The theory is that the cgm was operating as if it had reached some limit when it in fact did not (though we have no hard evidence). Steve could also see d0cabosg2 sending signals to the forwarding node cgm saying that jobs were completed, but the cgm never noticed the signals.

Steve noticed that many jobs were in the "DONE" state - which should be rare. I restarted the cgm on Saturday - this helped a little but not completely.

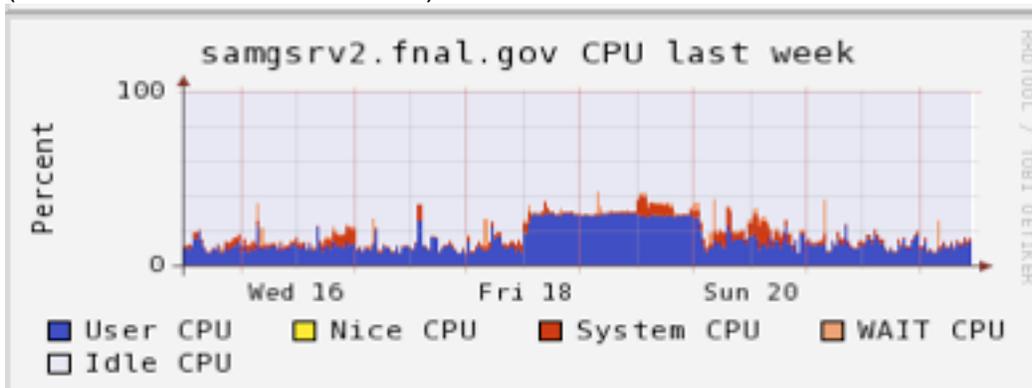
Parag also enlisted the help of Jamie (Condor developer) to look at the situation. He didn't have any theories as to what was going on. Jamie will supply us with a cgm with more debugging information.

Also, the memory usage of the forwarding nodes took a jump at the beginning of the week.



Not sure if that is indicative of a real problem.

A possible red herring - Schedd on the QUEUEING NODE was eating 100% CPU (we've never seen that before)...



Note that the CPU usage went down BEFORE I nuked the system (??). Steve also noted many timeouts in the Schedd log file.

We are confused about how some of the globus states contribute to the MAX_GRID limit (set to 750). E.g... Stage_out, Stage_in, Done, Active, Pending?

Also, how many jobs were in these states when things were really bad? What was the state of the XML DB during this time?

ACTION ITEMS

3.1) Add more plots to ganglia to count jobs in different globus states (e.g. StageIn, Done, ...) DZSAM-178 assigned to Adam.

3.2) Investigate the Condor spool directory that Adam moved out of the way during the

nuking. How many jobs were in stage-in, stage-out, done? DZSAM-179 assigned to Adam.

3.3) Investigate moving the station off of samgfw01. We don't think this causes any additional load, but we know that the project_masters take up memory. The fact that samgfw02 had the same problem vindicates the station and project_masters, but it may be good to move the station off anyway. DZSAM-180 assigned to Robert.

3.4) Do not submit jobs to sites that can never run them. Steve Timm notices that there are many jobs submitted to sites that are down or have problems. Why are we submitting there? Joel says that he does what ReSS says. Parag has a plan to mitigate this problem. DZSAM-182 assigned to Parag.

WHAT TO DO IF/WHEN THIS HAPPENS AGAIN:

at the very least...

- a) Try to install a debug version of the condor_gridmanager
- b) Look at the state of the XML DB process (is it eating CPU?)
- c) Look at the different globus states (Stage_in, ...)

APPENDIX:

Here is a time line of the production problems as noticed by Mike. Joel had even earlier noticed a larger number of held jobs than usual.

Tuesday ~12:00

Reported to d0sam-admin

Initial symptoms showed forwarding nodes blocked on CurMatches. CurMatches did not correspond to info on web page. Jobs in idle state waiting to be matched to forwarding node

[We had started looking]

Wednesday ~07:00

No response yet. Sent second note to d0sam-admin
Jobs now being matched to forwarding nodes, but several hundred slots on CAB still empty. Samgfw02 had load spike to ~30 on Tuesday when slowdown occurred that lasted a few hours and then went back to ~5. Jobs slowly going through samgfw02. Still looks limited by CurMatches.

Samgfw01 also had load spike to ~30-40, but never went back down. Lots of jobs now matched to samgfw01

but not going forward to CAB. Several hundred processes listed on web page as pending.

[Parag and Robert were looking at the problem]

Wednesday ~14:30

Still no response. Poked d0sam-admin again. Kin passes it on to experts.

Robert reports that he and Parag have noted inconsistency between number of processes samgfw01 thinks have been sent to cabsrv2 and number of processes PBS thinks are running. Parag consulting with condor experts.

Thursday

Large number of jobs suddenly go through around midnight. Not sure what action was taken. Managed to fill slots on CAB during most of day on Thursday, but still many processes in pending state on samgfw01.

Submission slots to crawl and slots start going idle again by late Thursday evening.

Friday ~12:30

Condor_schedd on samgrid.fnal.gov suddenly starts using 100% of a cpu. Makes retrieval of info from web page slow and spotty. This continues until after midnight on Saturday. Not clear if this is connected with forwarding problems.

Saturday ~08:30

Submissions slowing even more. Samgfw02 now also has many pending jobs. Not blocked on CurMatches. Jobs also now starting to fall into Held state reporting "Unspecified gridmanager error".

Restarted condor_gridmanager on both samgfw01 around 13:30. Some jobs started to go through, but at a very slow rate of < 100/hour. Lot of jobs remained pending on forwarding node. Also restarted condor_gridmanager on samgfw02 around 16:30

Number of running processes on CAB slowly increased to ~700 around midnight and then started to go back down. Both samgfw01 and 02 were blocked on CurMatches and both had large numbers of processes in

pending states.

Sunday ~11:30

Nothing improved overnight. Number of running processes on CAB back down below 400. Decide to nuke the system. Adam begins bombing. This clears everything quickly. We are back up to 1400 processes by ~15:00."

Monday - Parag changes samgrid Schedd parameters to improve the resilience of the communication of globus job managers