Enhancements to Condor-G for the ATLAS Tier 1 at BNL

John Hover
Group Leader
Experiment Services (Grid Group)
RACF, BNL
Outline

- Background
- Problems
- Solutions
- Results
- Acknowledgements
Our (Odd?) Situation

• **ATLAS Pilot-based Grid Workload System: PanDA**
  (Production and Distributed Analysis)
  – Individual (pilot) jobs are identical.
  – Individual (pilot) jobs are not valuable.
  – Jobs can, unpredictably, be very short (~3-5 minutes).

• **Brookhaven National Laboratory's Role:**
  – BNL Tier 1 responsible for sending pilots to all ATLAS sites in OSG (U.S. Cloud).
  – Central PanDA services located at CERN.
PanDA Autopilot

• Runs on top of Condor-G.
  – One automatic scheduler process for each PanDA 'queue'.
  – Each running `condor_q` and parsing output, and `condor_submit`.
  – (Nearly) all run as single UNIX user.
  – Each minute:
    • Queries Condor-G for job status (per queue per gatekeeper).
    • Queries Panda Server for current `nqueue` value.
    • Decides how many pilots to submit.

• At BNL ATLAS Tier 1:
  – 5 Submit hosts. (3 primary)
  – Serving 92 PanDA queues at 43 gatekeepers (some overlap).
Condor-G Interaction Diagram

1. Submission
2. Job I/O, Control
4. Job runs and finishes.
5. GridMonitor restarts JM.
6. JM does stage-out, quits.

Condor Team
Globus Team

Communication
Fork/Exec
Problems (Opportunities) 1

• ~5000 job ceiling
  - General scaling issues in Condor-G in 6.8.x.
  - Manual operation and cron job often needed to clean up stuck jobs and restart Condor-G processes.

• HELD jobs
  - Held jobs “clog” queue and interfere with further submission.

• GRAM <-> Condor communication glitches
  - Condor-G loses track of jobs at site. Requires gaehp_server restart. Slow/no job status update.
  - Memory leak in Globus client?

• Inter-site effects
  - Error condition on one site/gatekeeper can affect another.
Problems 2

• **Grid Manager <-> Grid Monitor Issues**
  - When problem occurs, a new Grid Monitor is not started for an hour.

• **Difficulty troubleshooting**
  - `condor_status` info oriented toward local batch.
Solutions 1

• Establish the goal and set up coordination between BNL and Condor Team members.
  – Formal meeting at BNL to discuss plans.
  – Full login access for Condor devs on BNL production hosts.
  – Frequent email and phone communication to track progress.
  – Clear problem list and action items for teams.
  – This was a pre-requisite for all further progress.

• Ultimately, establish stress testbed at U.Wisc. to which we submit.
Solutions 2

• **Internal efficiency fixes:**
  - Jaime found loops (that cycle through internal data structures) that were inefficient at 5000+ job scales. Fixed.

• **HELD jobs never needed. Pilots are expendable.**
  - `+Nonessential = True`
  - When pilot jobs fail, we don't care. Just remove and discard them rather than saving them for later execution.
  - Unconditional removal and cleanup enabled.

• **Grid Monitor restart behavior fix**
  - Made this configurable: `GRID_MONITOR_DISABLE_TIME`
  - But required refining the error handling on the Grid Manager side to avoid accidentally flooding site with Grid Monitors.
Solutions 3

• **Grid Manager Tweaks**
  - Previously, one GridManager per user on submit host. Since all sites served by a single user, only one started.
  - `GRIDMANAGER_SELECTION_EXPR = GridResource`
  - Determines how many GridManagers get started, by providing an expression used to hash resources. Now we have a separate Gridmanager per gatekeeper, per user on submit host.

• **GAHP Server fixes**
  - Frequent source of communication errors.
  - Jaime worked with Globus dev (Joe Bester) to integrate upstream fixes into GAHP.
Solutions 4

• **Separate throttle on limiting jobmanager processes based on their role:**
  
  – Previously Condor-G had one throttle for the total number of jobmanagers invoked on the remote CE
    • A surge in job completions/removals will stall new job submission, and vice-versa.
  
  – Now the throttle limit is broken in half, one for job submission, the other for job completion/cleanup
  
  – Sum controlled by:
    GRIDMANAGER_MAX_JOBMANAGERS_PER_RESOURCE
  
  – ( Might be nice to have distinct settings.)
Solutions 5

- **Improved `condor_status -grid` output:**

```
[root@gridui11 condor-g-probe]# condor_status -grid

<table>
<thead>
<tr>
<th>Name</th>
<th>NumJobs</th>
<th>Allowed</th>
<th>Wanted</th>
<th>Running</th>
<th>Idle</th>
</tr>
</thead>
<tbody>
<tr>
<td>gt2 abitibi.sbgrid.org:2119</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>gt2 cmsosgce3.fnal.gov:2119</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>gt2 cobalt.uit.tufts.edu:2119</td>
<td>90</td>
<td>90</td>
<td>0</td>
<td>38</td>
<td>52</td>
</tr>
<tr>
<td>gt2 fester.utdallas.edu:2119</td>
<td>119</td>
<td>119</td>
<td>0</td>
<td>80</td>
<td>39</td>
</tr>
<tr>
<td>gt2 ff-grid3.unl.edu:2119</td>
<td>162</td>
<td>162</td>
<td>0</td>
<td>0</td>
<td>162</td>
</tr>
<tr>
<td>gt2 gate01.aglt2.org:2119</td>
<td>2398</td>
<td>2398</td>
<td>0</td>
<td>2017</td>
<td>381</td>
</tr>
<tr>
<td>gt2 gk01.atlas-sw2.org:2119</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>gt2 gk04.swt2.uta.edu:2119</td>
<td>535</td>
<td>535</td>
<td>0</td>
<td>510</td>
<td>25</td>
</tr>
<tr>
<td>gt2 gridgk05.racf.bnl.gov:2119</td>
<td>1410</td>
<td>1398</td>
<td>0</td>
<td>648</td>
<td>737</td>
</tr>
</tbody>
</table>
```
Solutions 6

- **Establish a stress testbed to explore limits.**
  - One submit host at BNL.
  - Four gatekeepers at Wisconsin, in front of a Condor pool of ~7000 nodes.
  - Test job:
    - Sleep 1200
    - 500KB input and output for staging
  - Runs Condor development release.
Solutions (Summary)

• Generally, over a ~6 month period (mid 2009 to early 2010)

Jaime and the Condor team:
  – Responded promptly to problem reports.
  – Actively helped us troubleshoot mysterious behavior.
  – Rapidly developed fixes and tweaks to address issues.
  – Provided us with pre-release binaries to test.
  – Made sure we understood how to leverage newly-added features.
Results 1

- **Scalability**
  - ~5000 job ceiling now up to ~50000(?) per submit host.
  - We are now limited by contention issues and concern about hardware failures more than raw performance.

- **Functionality**
  - Nonessential jobs enabled.
  - HELD job behavior. Unconditional removal.

- **Configurability**
  - Tunable via new configuration variables.

- **“Monitor-ability”**
  - Enhancements to 'condor_status -grid' help us notice and solve problems.
Results 2

• **Stress test results:**
  
  – Comfortable limit reached.
  
  – Manage 50,000 jobs from one submit host.
  
  – Submit 30,000 jobs to one remote gatekeeper.
    • Gatekeeper runs only GRAM/GridFTP, no other OSG services running on it.
    • 30,000 is a hard limit, restricted by the number of subdirs allowed by the file system. Now exceeded at BNL with BlueArc NFS appliance.
  
  – All stress test improvements are included in the just-released condor 7.4.0 release
    • Now used on our production submit hosts.
Results 3: The numbers.

- ATLAS jobs (pilots) run on OSG (from OSG Gratia report) ~280,000 jobs a day.
Results 4: Nov '08 - Oct '09
Results 5

• **Generally all-around improved reliability.**
  - Fewer crashed processes
  - Fewer communication failures.
  - Less mysterious anomalies.

• **We all sleep better at night.**
The Future

• **Continue to refine 'condor_status -grid' output.**
  - More info, laid out in intuitive fashion.

• **Add time-integrated metric information to augment instantaneous info.**
  - How many jobs were submitted in the last 10 minutes to Site X?
  - How many finished in the last 10 minutes?
  - *Rates* rather than absolute numbers.

• **Finer-grained internal queue categorization to avoid contention.**
  - When multiple queues are served by one GridResource: PanDA considers them separate, while Condor-G thinks they are the same.
Acknowledgments: Thanks!!

- **Jaime Frey**
  - Condor-G lead developer.

- **Todd Tannenbaum**
  - Condor lead developer.

- **Xin Zhao**
  - BNL OSG Gatekeepers, Condor-G and PanDA Autopilot wrangler.

- **Miron Livny**
  - Condor Team Leader