Grid Job Submission and Condor-G Scalability Issues

Xin Zhao
Brookhaven National Lab

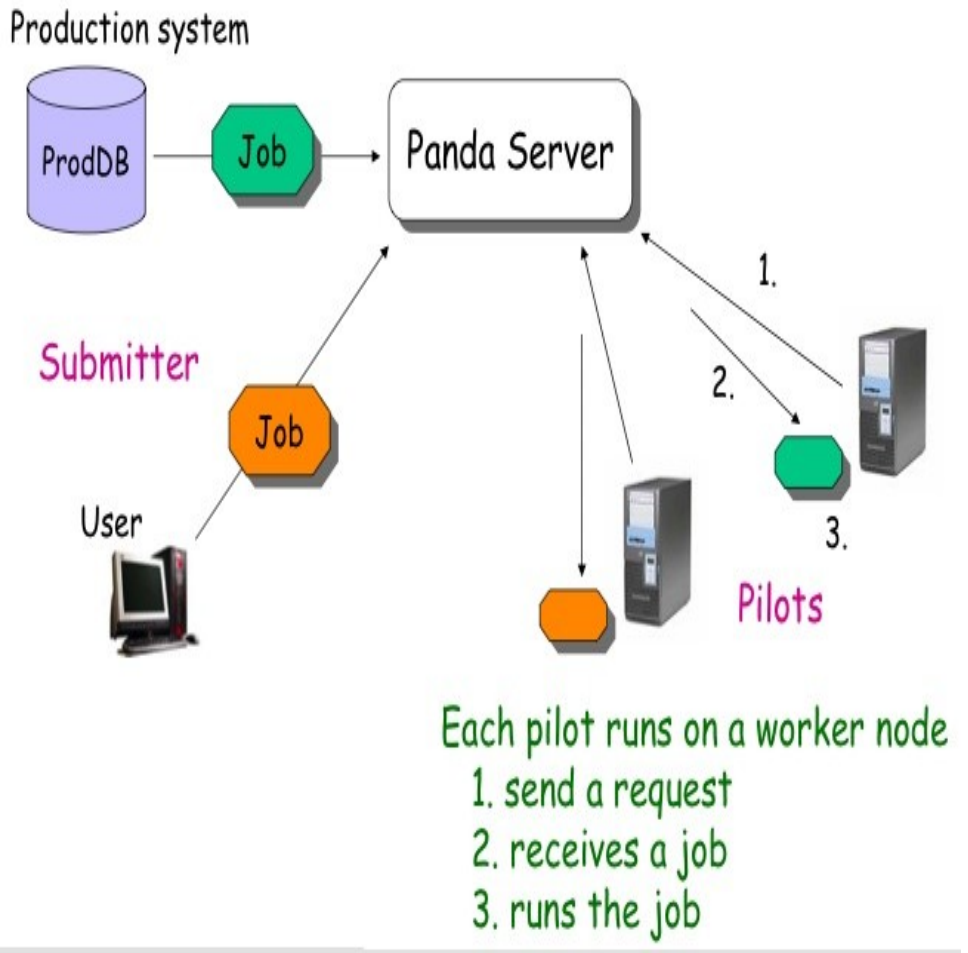
USATLAS T2/T3 Workshop
UTA, November 2009

Outline

- Condor-G in PanDA
- Condor-G in action
- Condor-G improvements
- Condor-G stress test
- Future plan

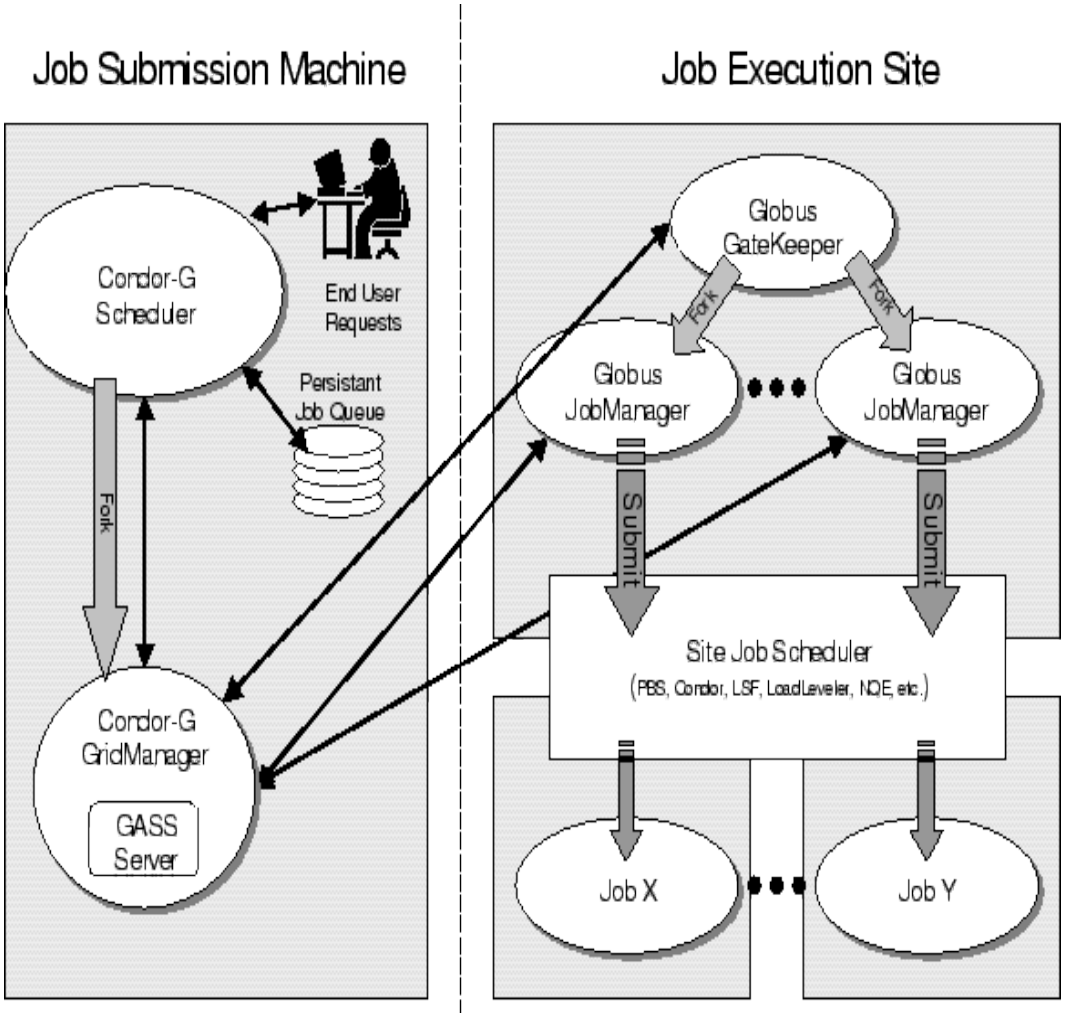
Condor-G in PanDA

- PanDA autopilot sends pilots to all production sites, using Condor-G.
- It's critical to maintain steady, high volume pilot streams to all sites.



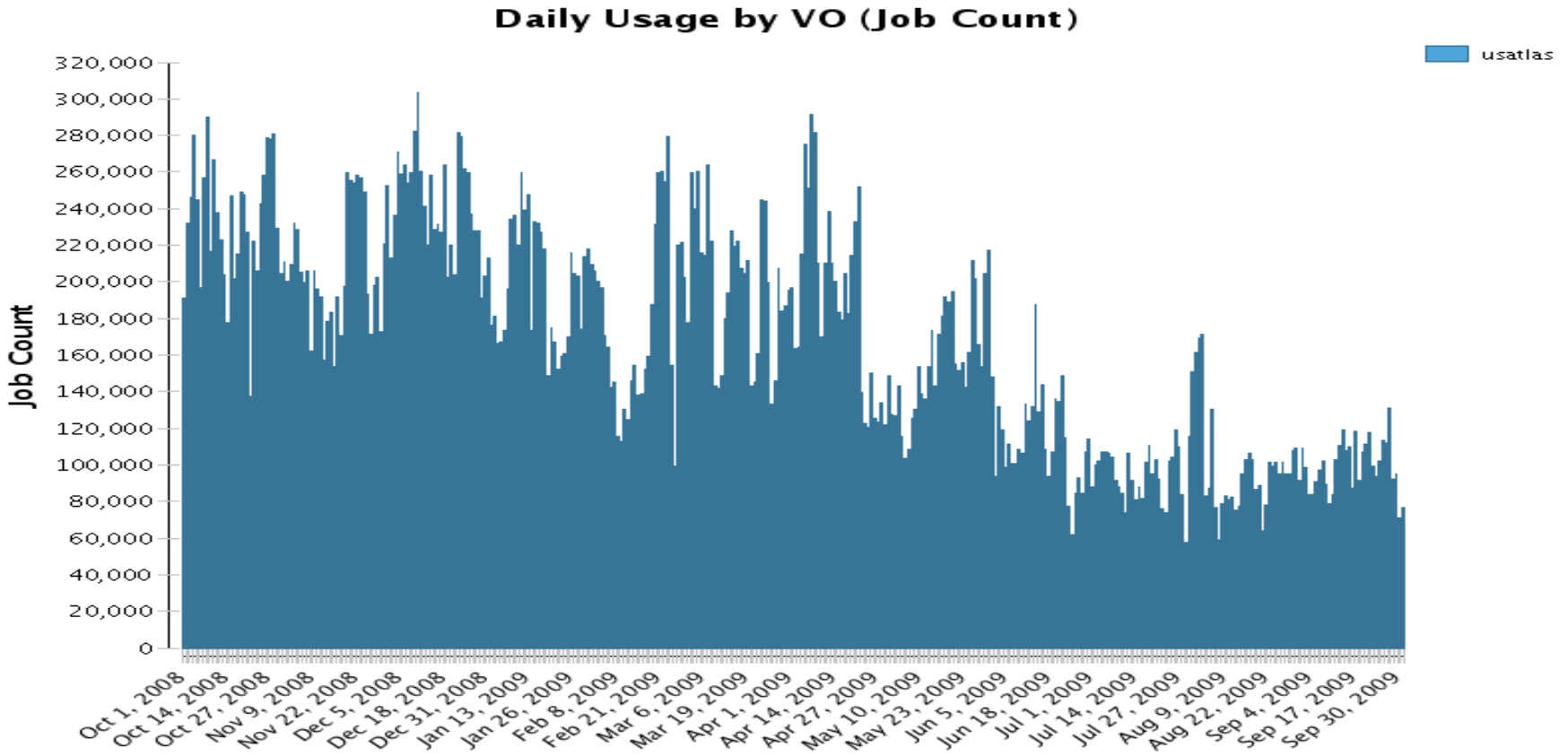
Condor-G in PanDA (cont.)

- Condor-G
  - “Grid job scheduler”
  - Job management from condor
  - Inter-domain resource management by Globus

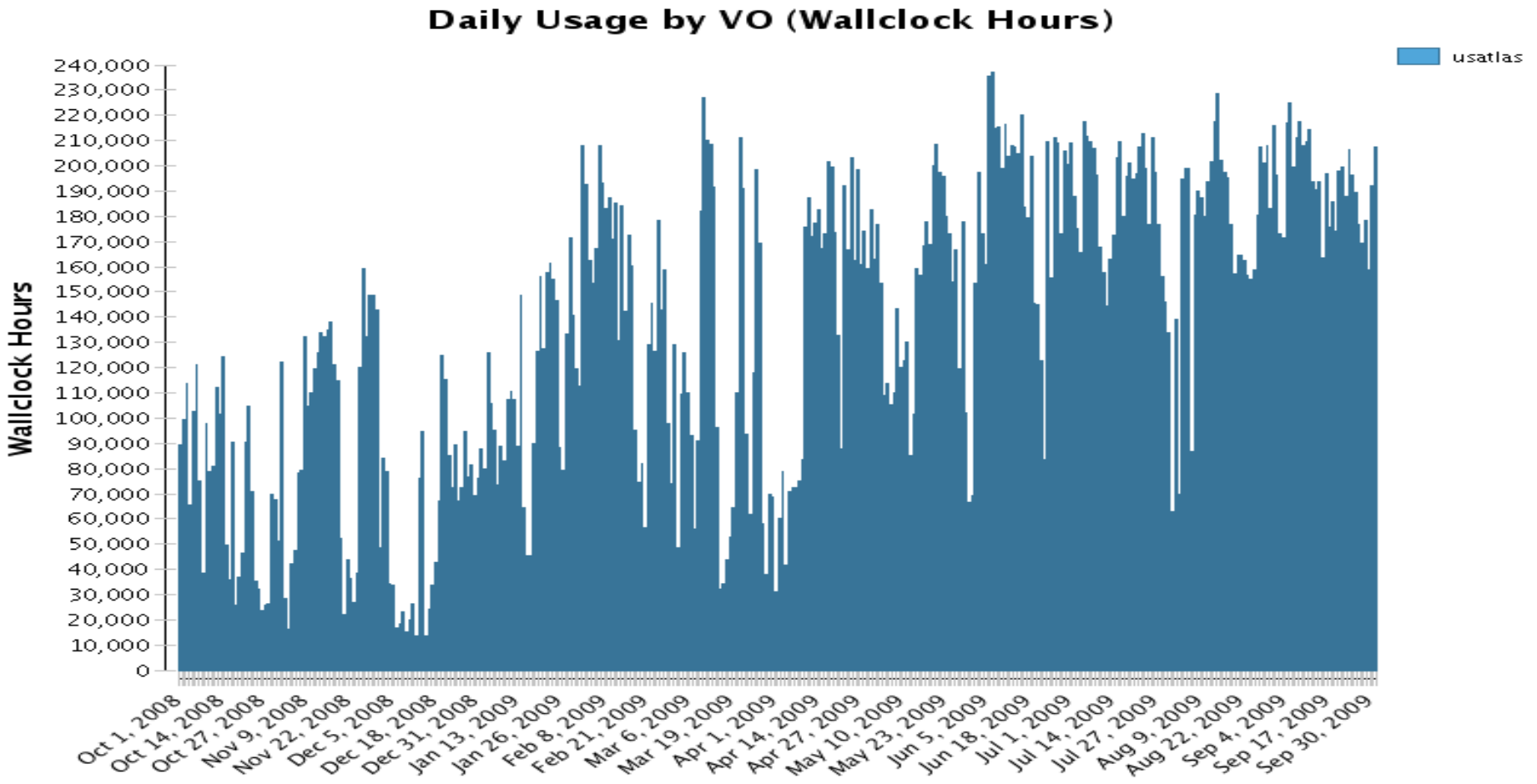


Condor-G in action

- USATLAS jobs(pilots) run on OSG (from OSG gratia report)
  - Run ~280,000 jobs a day at peak on OSG (most of them are pilots)



Condor-G in action (cont.)



Condor-G in action (cont.)

- Condor-G submit hosts at BNL
  - Feed pilots to all queues on USATLAS sites, and some EGEE sites as well
  - 5 submit hosts
    - Quad-core, 8~16GB memory nodes
    - >500GB local disk
    - Condor version 7.3.2 with some newer patches to Condor-G codes

Condor-G in action (cont.)

- Issues in condor-g submission
  - “No pilots while there are many activated jobs?”
  - “Why many jobs are in unsubmitted state?”
  - “Why many jobs stay in hold for a long time?”
  - “Why is job status not updated promptly ?”
  - Can't keep all job slots occupied for all sites?
  - Slow ramp-up after a downtime?
  - No output files found for some completed jobs?
  - .....

Condor-G in action (cont.)

- More submit hosts are needed to keep up with the production rate
  - Manual operation and cron job often needed to clean up stuck jobs and restart condor-g processes
- Hotline to Condor team almost every time production volume increases

Condor-G improvements

- Condor team is very responsive to our concerns and requests. In the past several months, we have weekly meeting with them to tackle some of the long standing issues in condor-g job submission
- Many improvements and enhancements are implemented in condor-g.
- Below are some highlights of the improvements

Condor-G improvements (cont.)

- Nonessential job attribute
  - Condor usually treats every job as important, will hold and retry them in case of failure
    - Too many hold jobs clogs condor-g, prevent it from submitting new jobs
    - Pilots are job sandboxes, not as essential as real job payloads
  - The Nonessential attribute is enabled in pilot job submit file
    - Condor-g will abort the problematic job in stead of putting it on hold, potentially leaving debris on the remote CE.

Condor-G improvements (cont.)

- GRID_MONITOR_DISABLE_TIME
  - Previous version of condor-g will wait for one hour, after a grid monitor job failure, to submit a new grid monitor job
    - Intention is not to flood remote CE when there is a problem
    - Wait time too long
      - Job submission rate can't sustain at high rate level
      - Slow down job status update
  - New condor-g configuration setting is introduced
    - Now we set this to 5 minutes

Condor-G improvements (cont.)

- Separate gridmanager process for each remote site
  - Previous condor-g starts a gridmanager instance for each user account, which covers all remote CE sites for this user
    - One problematic site could affect the overall throughput to all sites
  - Now it's possible to configure condor-g to start a new gridmanager instance per remote site per user
    - Better isolation of site issues from the overall production

Condor-G improvements (cont.)

- Better understanding and handling of failures grid monitor encounters
  - Gridmanager receives callbacks on grid monitor job status from jobmanager
  - Better globus-url-copy error handling in grid monitor
  - Cache of completed jobs in grid monitor
  - Don't give up on grid monitor job after 5 minutes
  - Cancel grid monitor job when gridmanager gives it up
  - Grid monitor outer script sends heartbeat more reliably

Condor-G improvements (cont.)

- Separate throttle on limiting jobmanager processes based on their role
  - Previous 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 break in half, one for job submission, the other for job completion/cleanup

Condor-G improvements (cont.)

- Globus bug fixes
  - GRAM client used by condor-g occasionally stops receiving connections from remote jobmanager for job status update
  - Memory leak in the globus client
  - We used to run a cronjob to periodically restart gahp server on the submit host
  - Fixed buy Globus team and new condor-g binary is compiled with the new library.

Condor-G improvements (cont.)

- Gridmanager publishes resource classads to collector, for users to query grid job submission status to all sites

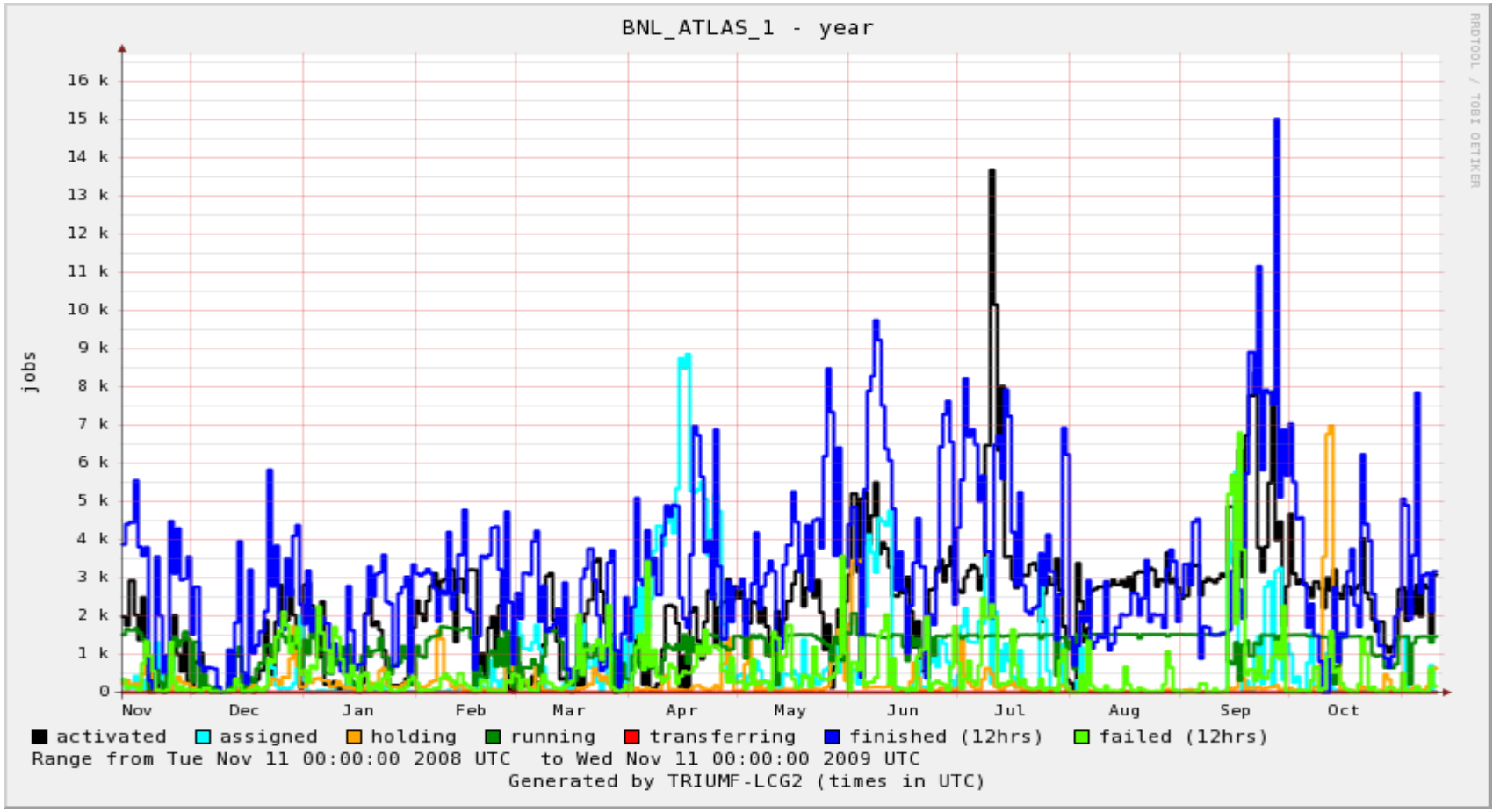
```
$> condor_status -grid
        Name                Job Limit   Running    Submit Limit  In Progress
        gt2 atlas.bu.edu:211   2500        376         200             0
        gt2 gridgk04.racf.bn  2500        1           200             0
        gt2 heroatlas.fas.ha  2500        100         200             0
        gt2 osgserv01.slac.s  2500        611         200             0
        gt2 osgx0.hep.uiuc.e  2500        5           200             0
        gt2 tier2-01.ochep.o  2500        191         200             0
        gt2 uct2-grid6.mwt2.  2500        1153        200             0
        gt2 uct3-edge7.uchic  2500        0           200             0
```

Condor-G improvements (cont.)

- Some best practices in condor-g submission
  - Reduce frequency of voms-proxy renewal on the submit hosts
    - Condor-g pushes renewed proxy to remote jobs aggressively
  - Avoid hard-kill (-forcex) jobs from client side
    - Leave debris on the remote gatekeeper
  - Separate condor schedd for managed-fork job manager on the gatekeeper
    - Not compete for scheduling cycle between grid monitor job and real production jobs

Condor-G improvements (cont.)

- pilots submission is much more stable today
  - Two submit hosts in production, each manages ~7500 jobs



Condor-G improvements (cont.)

- Current condor-g config settings on the BNL submit hosts
  - GRIDMANAGER_SELECTION_EXPR = GridResource
  - GRIDMANAGER_MAX_PENDING_REQUESTS = 500
  - GRIDMANAGER_MAX_SUBMITTED_JOBS_PER_RESOURCE = 20000
  - GRIDMANAGER_MAX_PENDING_SUBMITS_PER_RESOURCE = 1000000
  - GRIDMANAGER_JOB_PROBE_INTERVAL = 30
  - GRID_MONITOR_DISABLE_TIME = 300
  - GRIDMANAGER_MAX_JOBMANAGERS_PER_RESOURCE = 50
  - GRIDMANAGER_CHECKPROXY_INTERVAL = 1800
  - GRIDMANAGER_MINIMUM_PROXY_TIME = 180
  - ENABLE_GRID_MONITOR = TRUE

Condor-G stress test

- Stress test setup
  - One submit host at BNL
  - CE side
    - Four gatekeepers at Wisconsin, in front of a condor pool of ~7000 nodes
  - Test job
    - Sleep 1200
    - 500KB input and output for staging
  - Condor development release

Condor-G stress test (cont.)

- Comfortable zone 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
  - All stress test improvements are included in the just-released condor 7.4.0 release
    - Not used on our production submit hosts yet

Future plan

- Continue the stress test
  - Push higher on the total number of jobs one submit host can handle ?
  - Compare the performance between a lightweight gatekeeper and a fully loaded OSG gatekeeper
  - Expectations from ATLAS production ?
    - 100,000 jobs per submit host ??
- Improve monitoring and debugging tools for condor-g
  - Not the easiest thing to automate checking for, but will try