Experiences from Running the PanDA Pilot Factory

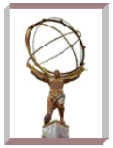
Xin Zhao

Brookhaven National Laboratory

OSG All Hands Meeting

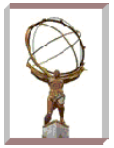
Harvard Medical School, Boston, MA

March, 2011



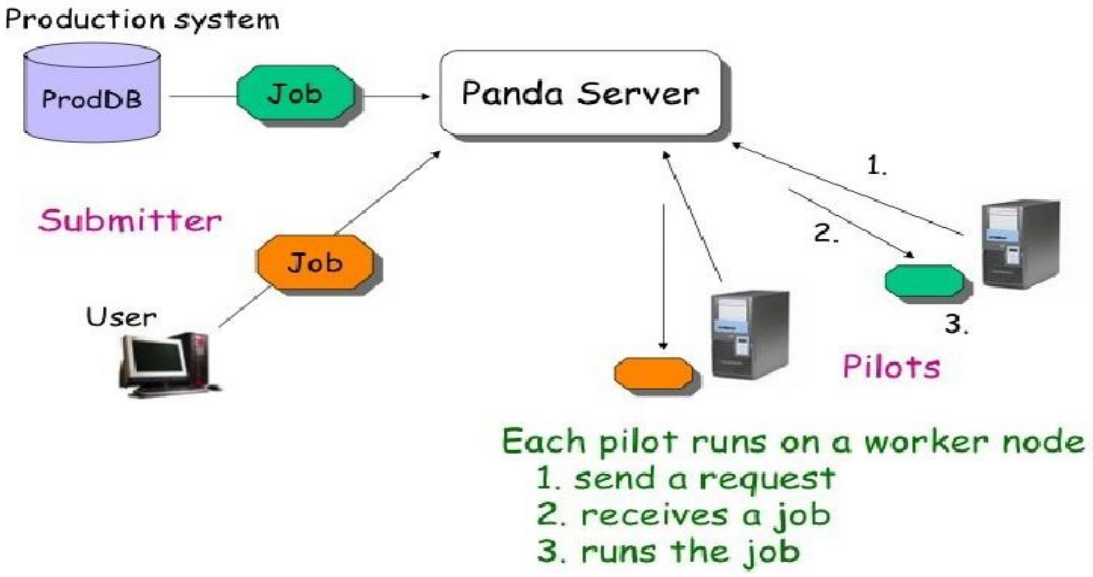
Outline

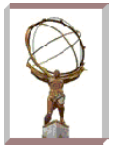
- PanDA Pilot Submission and Condor-G
- PanDA Pilot Submission at BNL ATLAS Tier1
- Scalability and Stability
- Monitoring



PanDA Pilot Submission and Condor-G

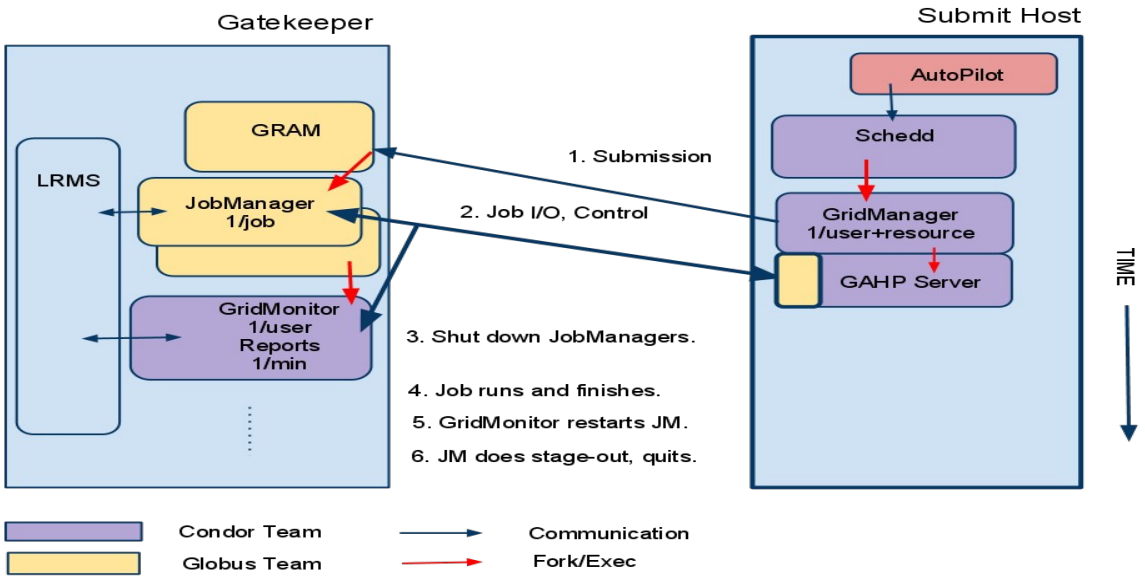
- PanDA (Production and Distributed Analysis)
  - ATLAS Workflow management system
  - Late-binding, pilot based system

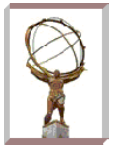




PanDA Pilot Submission and Condor-G

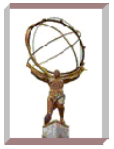
- Condor-G
  - PanDA Pilot Submission uses Condor-G as the grid job scheduler



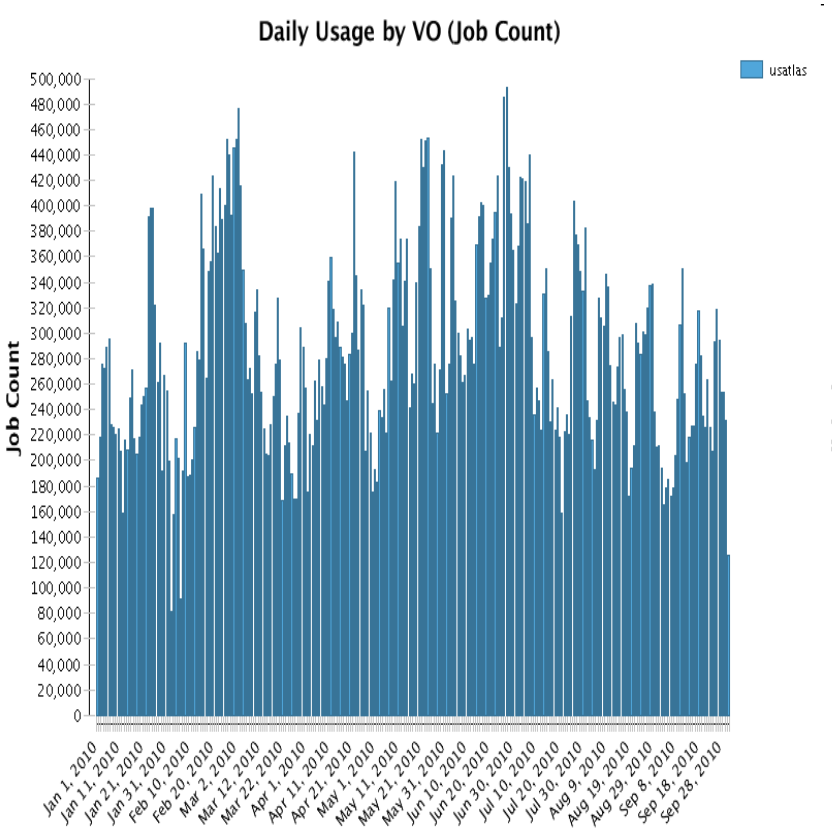


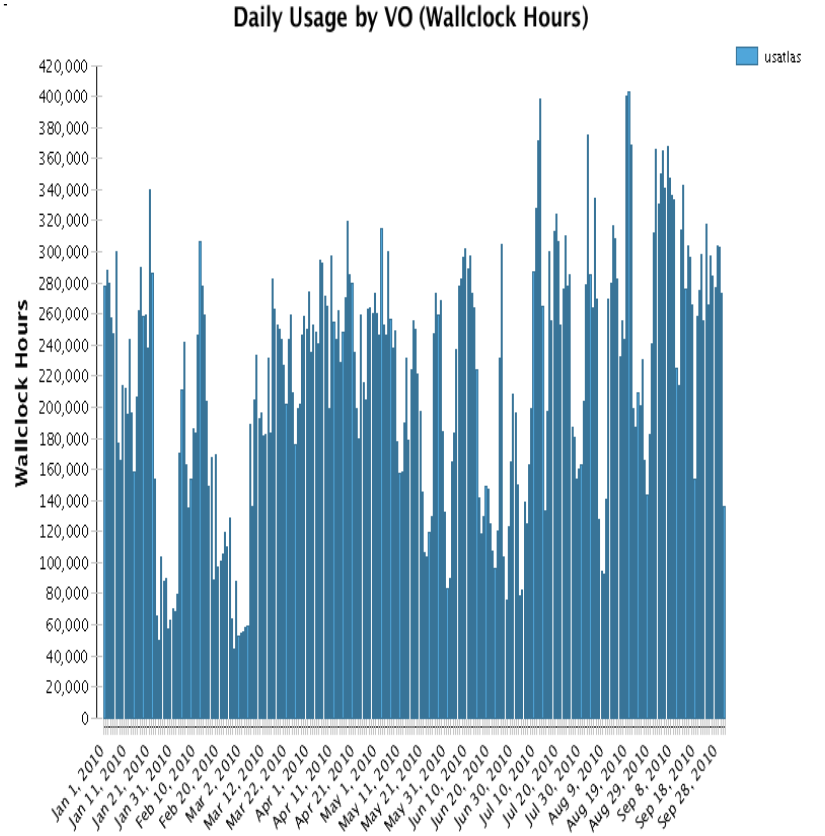
PanDA Pilot Submission at BNL T1

- Submit Pilots to ATLAS US Cloud
  - 92 PanDA queues at 43 gatekeepers, with HEPSPEC-06 ~ 111,000
  - Full scale production has >10K real jobs running, daily peaked at 480K pilots
  - 5 Condor-G submit hosts (3 primary), running Condor 7.4.4 right now.
  - Stress test results
    - 50K jobs managed on one submit host;
    - 30K jobs submitted to one remote GT2 gatekeeper



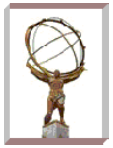
PanDA Pilot Submission at BNL T1





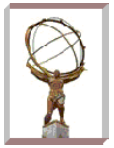
Scalability and Stability

- Many improvements and enhancements implemented to make the system more stable and scalable
  - Thanks to Condor team for their constant support and prompt response
  - Below I will go through some of the major improvements, and best practices, in a Q&A mode.



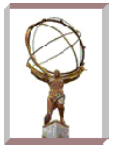
Scalability and Stability (1)

- Q: how to keep a stable pilots pool
  - Late-binding model
    - Whenever a real job is activated, there'd better be a pilot to pick up and run it. Latency as short as possible
  - Always keep certain number of pilots pending on local batch systems
  - Fluctuation of pilots pool
    - Changes of site status or job status
    - Dip needs to be refilled quickly



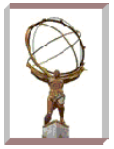
Scalability and Stability (1)

- A: aggressive Condor-G throttle settings
  - GRIDMANAGER_MAX_PENDING_REQUESTS = 500
  - GRIDMANAGER_MAX_JOBMANAGERS_PER_RESOURCE = 200
  - GRIDMANAGER_MAX_SUBMITTED_JOBS_PER_RESOURCE = 20000
  - GRIDMANAGER_MAX_PENDING_SUBMITS_PER_RESOURCE = 1000000
  - GRIDMANAGER_JOB_PROBE_INTERVAL = 60
  - GRID_MONITOR_DISABLE_TIME = 300
  - GRIDMANAGER_CHECKPROXY_INTERVAL = 1800
  - GRIDMANAGER_MINIMUM_PROXY_TIME = 180
  - CRED_MIN_TIME_LEFT = 120
  - ENABLE_GRID_MONITOR = TRUE



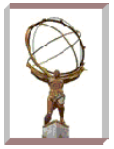
Scalability and Stability (2)

- Q: how not to overload/crash remote gatekeepers
  - Aggressive Condor-G throttle setting
    - potentially high load on remote CEs
  - GRIDMonitor helps reduce load
    - A lot of long running jobs is not an issue
  - Load still there when job starts up (stage-in) and finishes (stage-out)
    - CE will be hit hard if a lot of pilots come and go very fast, eg "empty" pilots or short real jobs.



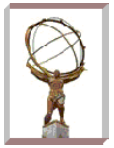
Scalability and Stability (2)

- A: adjust pilot submission rate
  - Autopilot adjuster script
    - Automatically adjust pilot submission limit based on availability of real jobs
    - Avoid overloading CE with a lot of "empty" pilots (no real jobs to run)
  - Multi-job pilot
    - Pilot is configured to stay alive for a minimum period of time (60minutes), given there are still real jobs to run
    - Avoid overloading CE with short jobs



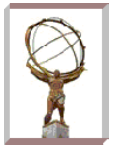
Scalability and Stability (3)

- Q: how to stabilize the overall pilot submission
  - Condor-G by default starts one gridmanager instance per user account, for all sites
    - Pilots are submitted using the same production account/certificate
  - "local" site issues can affect the overall throughput to all sites, served by the same submit host
    - A problematic CE
    - Regional network issue



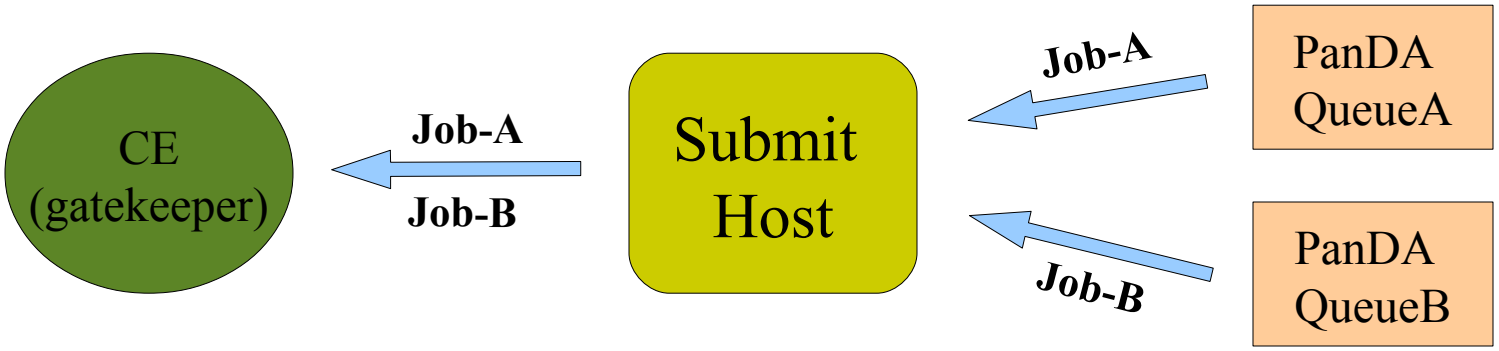
Scalability and Stability (3)

- A: Multiple GridManager instances
  - GRIDMANAGER_SELECTION_EXPR = <ClassAd expression>
  - GRIDMANAGER_SELECTION_EXPR = GridResource
    - A separate gridmanager instance per <remote site,user> pair
  - Throttle settings apply to each instance



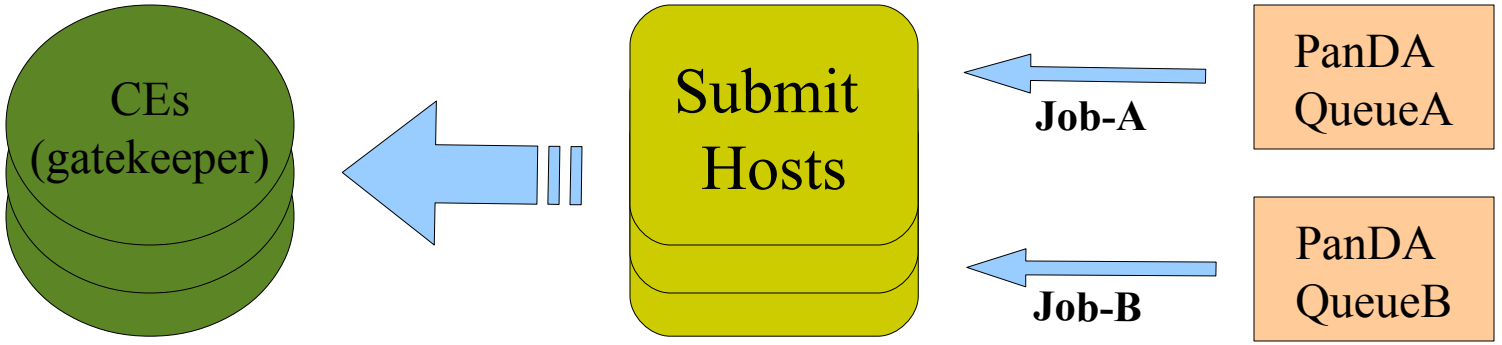
Scalability and Stability (4)

- Q: how to avoid contention among PanDA queues
  - PanDA queues may compete in the same Condor-G "channel"
    - Channel : one gridmanager instance, defined by **<CE,user,submit host>** value



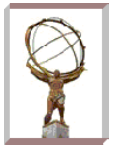
Scalability and Stability (4)

- A: create more "channels"
  - make more **<CE,user,submit host>** combinations
    - Multiple CEs per site, or multiple submit hosts per remote CE
    - assign PanDA queues among them



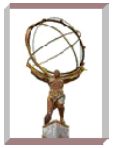
Scalability and Stability (4)

- A: another way to create more "channels"
  - Insert PanDA queue info into pilot job ClassAd and use GRIDMANAGER_SELECTION_EXPR expression
  - For example:
    - GRIDMANAGER_SELECTION_EXPR = PandaQueue
    - Will create a new gridmanager instance per PandaQueue value
  - In our to-do list, requires changes to autopilot factory, and needs more testing



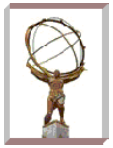
Scalability and Stability (5)

- Q: how to stabilize each submission channel
  - A submission channel can get stuck
    - Can't (or too slow to) get jobs status from remote sites
    - Can't (or too slow to) submit new pilots
    - ……
  - Many improvements and enhancements to tackle these problems



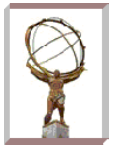
Scalability and Stability (5)

- A: deal with held jobs
  - Unconditional removal and cleanup of held pilots
    - Pilots are expendable, not real jobs
    - Not to waste cycle to retry them, which are often not recoverable
  - Introduce a new job attribute
    - +Nonessential = True



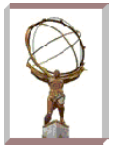
Scalability and Stability (5)

- A: Grid Monitor related improvements
  - Grid Monitor restart behavior adjustable
    - GRID_MONITOR_DISABLE_TIME
  - Cache jobs status in Grid Monitor
  - Patch globus jobmanagers to improve efficiency of jobs scanning and status update
  - Refined GridManager error handling to avoid flooding sites with the Grid Monitor jobs



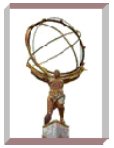
Scalability and Stability (6)

- Some best practices
  - Caching condor_q command results in autopilot factory
    - reduce load to condor schedd on submit hosts
  - Reduce frequency of proxy renewal
    - Condor-G pushes renewed proxy to all remote jobs aggressively
  - Avoid hard-killing jobs from submit hosts
    - leave debris on remote sites



Monitoring

- Condor command to display grid resources
  - condor_status -grid
- Home made monitoring dashboard
  - Add time-integrated metric information
    - Historical information of submission
    - Rate of submission
  - Connected to nagios
    - proactively monitor and report problems in the production pilot factory



Monitoring

- Monitoring dashboard snapshot (1)

RACF
Grid Group

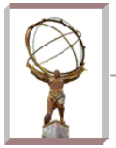
The RACF Production Dashboard

- Main Page
- Production System Overview
- Panda Production Pilots
  - Panda Production Queues
  - Current Panda Jobs
  - Queues on Submit Hosts
- BNL gatekeepers
  - BNL Gatekeepers
  - BNL Grid Ftp Servers
  - BNL Submit Hosts
  - NFS access time
  - BDII Status
- Other monitoring links:
  - RACF Nagios
  - RACF Nagios Services Table
  - Current RSV status

Current Panda Jobs

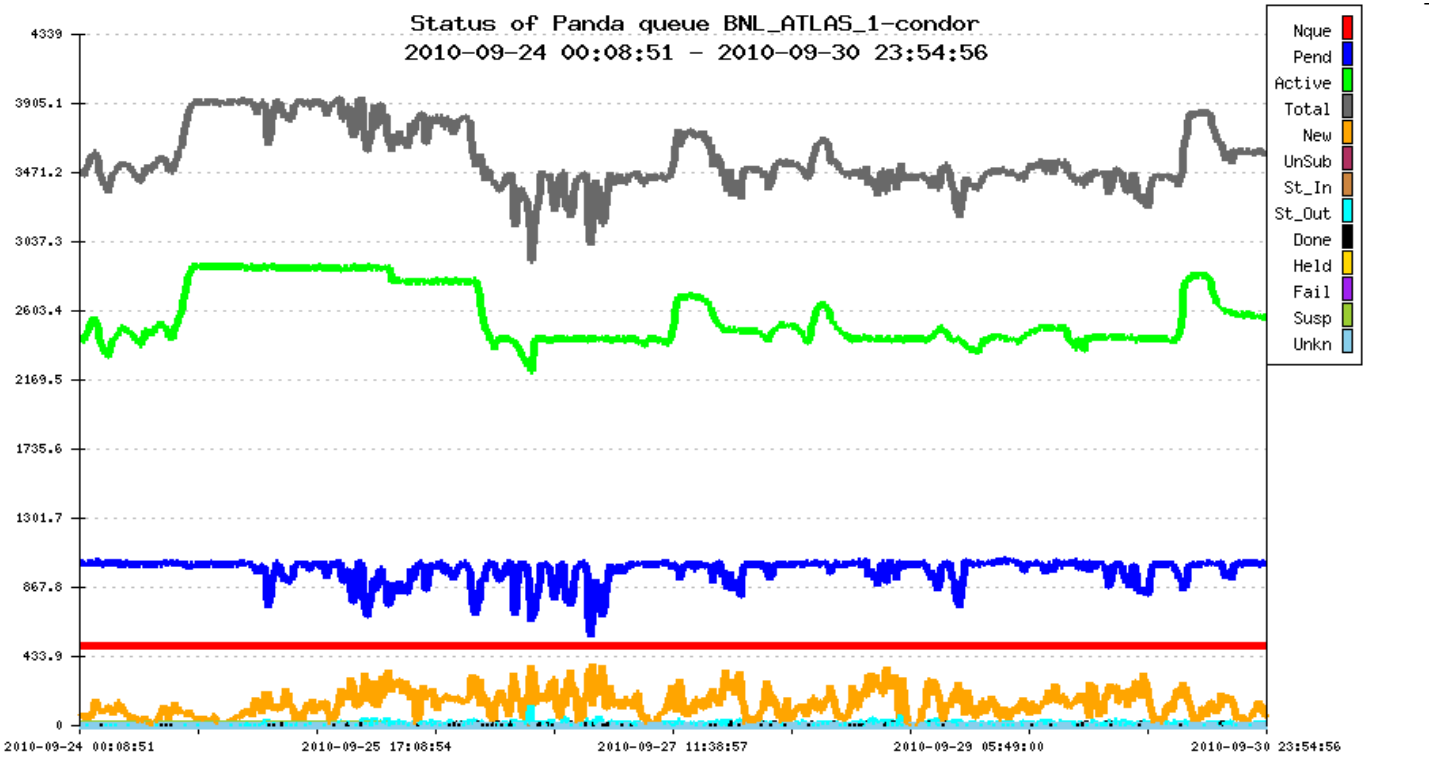
Queues on machine gridui07.usatlas.bnl.gov; status as of 2010-10-01 18:34:52

Panda_Queue_name	Total	New	UnSub	St_In	Pend	Active	St_Out	Done	Held	Fail	Susp	Unkn	Nque
AGLT2-OSG-condor	150	0	0	10	107	32	1	0	91	0	0	0	50
AGLT2-condor	563	8	0	0	110	453	0	0	0	0	0	0	110
ANALY_BNL_ATLAS_1-condor	451	0	0	0	309	142	0	0	0	0	0	0	250
ANALY_LONG_BNL_ATLAS	798	6	0	0	155	643	0	0	0	0	0	0	150
ANALY_Tufts_ATLAS_Tier3-lsf	80	10	0	0	56	24	0	0	0	0	0	0	10
BNL_ATLAS_1-condor	1726	91	0	0	464	1259	0	0	0	0	0	0	500
BNL_ATLAS_2-condor	354	7	0	0	190	161	0	0	0	0	0	0	200
BNL_ITB_ATLAS_TEST-condor	2	2	0	0	0	0	0	2	0	0	0	0	1
BNL_ITB_Test1-condor	8	6	0	0	2	4	2	0	0	0	0	0	4
BNL_SITE_GK02-condor	350	0	0	0	350	0	0	0	0	0	0	0	50
FIU-PG-condor	70	0	10	0	0	0	0	0	60	0	0	0	10
Firefly_SBGRID-pbs	10	9	1	0	0	0	0	0	0	0	0	0	10
Harvard-East_SBGRID-condor	12	12	0	0	2	10	0	0	0	0	0	0	10
IU_OSG-pbs	1	1	1	0	0	0	0	0	0	0	0	0	1
LBNL_DSD_ITB-condor	15	12	0	0	0	14	0	1	0	0	0	0	10
MWT2_IU-pbs	1153	17	0	0	512	640	1	0	0	0	0	0	512
OSG_LIGO_PSU-pbs	130	0	130	0	0	0	0	0	0	0	0	0	10
OUHEP_ITB-condor	19	16	0	0	6	10	1	0	0	0	0	0	10
SBGrid-Harvard-East-condor	10	0	10	0	0	0	0	0	0	0	0	0	10
SBGrid-Harvard-Exp-condor	90	0	90	0	0	0	0	0	0	0	0	0	10
SWT2_CPB-pbs	120	0	0	0	20	100	0	0	0	0	0	0	20
TTU_TESTWULF_ITB	1956	0	0	0	54	1902	0	0	0	0	0	0	10
Tufts_ATLAS_Tier3-lsf	88	20	0	2	46	38	0	0	0	0	0	0	20
UCITB_EDGE7-pbs	16	10	0	0	11	4	1	0	0	0	0	0	10
UConn-OSG-condor	2100	10	0	1971	0	72	23	0	2028	0	0	0	50
UFlorida-IHEPA-condor	130	0	130	0	0	0	0	0	0	0	0	0	10
UFlorida-IHEPA-hg-atlas-condor	130	0	130	0	0	0	0	0	0	0	0	0	10
UFlorida-PG-condor	130	0	130	0	0	0	0	0	0	0	0	0	10
UFlorida-PG-pg-atlas-condor	12	0	0	0	10	0	0	0	0	0	2	0	10



Monitoring

- Monitoring dashboard snapshot (2)



Future Plan

- New version of AutoPilot Factory coming
- CREAM vs Condor-G test
  - Some preliminary results
  - More scalability and performance tests to do