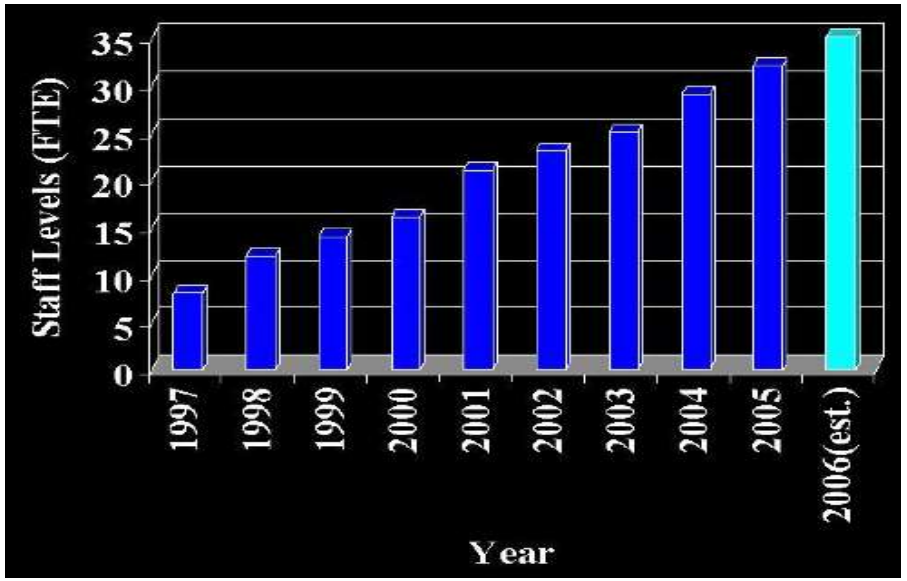
Condor Experience at Brookhaven National Laboratory

Alexander Withers
RHIC/US ATLAS Computing Facility
CondorWeek 2006

(Brief) Facility Overview



- One of a handful of Laboratories supported and managed by the U.S. gov't through DOE.
- RHIC/ATLAS Computing Facility is operated by BNL Physics Dept. to support the scientific computing needs of two large user communities.
  - RCF is the "Tier-0" facility for the four RHIC expts.
  - ACF is the Tier-1 facility for ATLAS in the U.S.
  - Both are full-service facilities.
- >2400 Users, 31 FTE.
- RHIC Run6 (Polarized Protons) started March 5th.

Computing Facility Resources

- Full service facility: central/distributed storage capacity, large Linux Farm, robotic system for data storage, data backup, etc.
- 9+ PB permanent tape storage capacity.
- 920+ TB central/distributed disk storage capacity.
- ~4000 CPUs available to Condor.
- 1.8 million SpecInt2000 aggregate computing power in Linux Farm.

Batch Computing Overview

- All reconstruction and analysis batch systems have been migrated to Condor, except STAR analysis ---which still awaits features like global job-level resource reservation --- and some ATLAS distributed analysis (these use LSF 6.0).
- Configuration:
  - Five Condor (6.6.x w/6.7.x startd) pools on two central managers.
  - 113 available submit nodes.
  - One monitoring/Condorview server and one backup central manager.

Condor Monitoring

- Nagios and custom scripts provide live monitoring of critical daemons.
- Place job history from ~100 submit nodes into central database.
  - This model will be replaced by Quill.
  - Custom statistics extracted from database.
- CondorView being replaced by custom websites using rrdtool and quill.

Condor Monitoring, cont.

- Custom startd, schedd, and “startd cron”. ClassAds allow for quick viewing of the state of the pool using Condor commands.
  - Some information accessible via web interface.
- Custom startd ClassAds allow for remote and peaceful turn off of any node.
  - Note that the “condor_off -peaceful” command (v6.8) cannot be canceled (?).

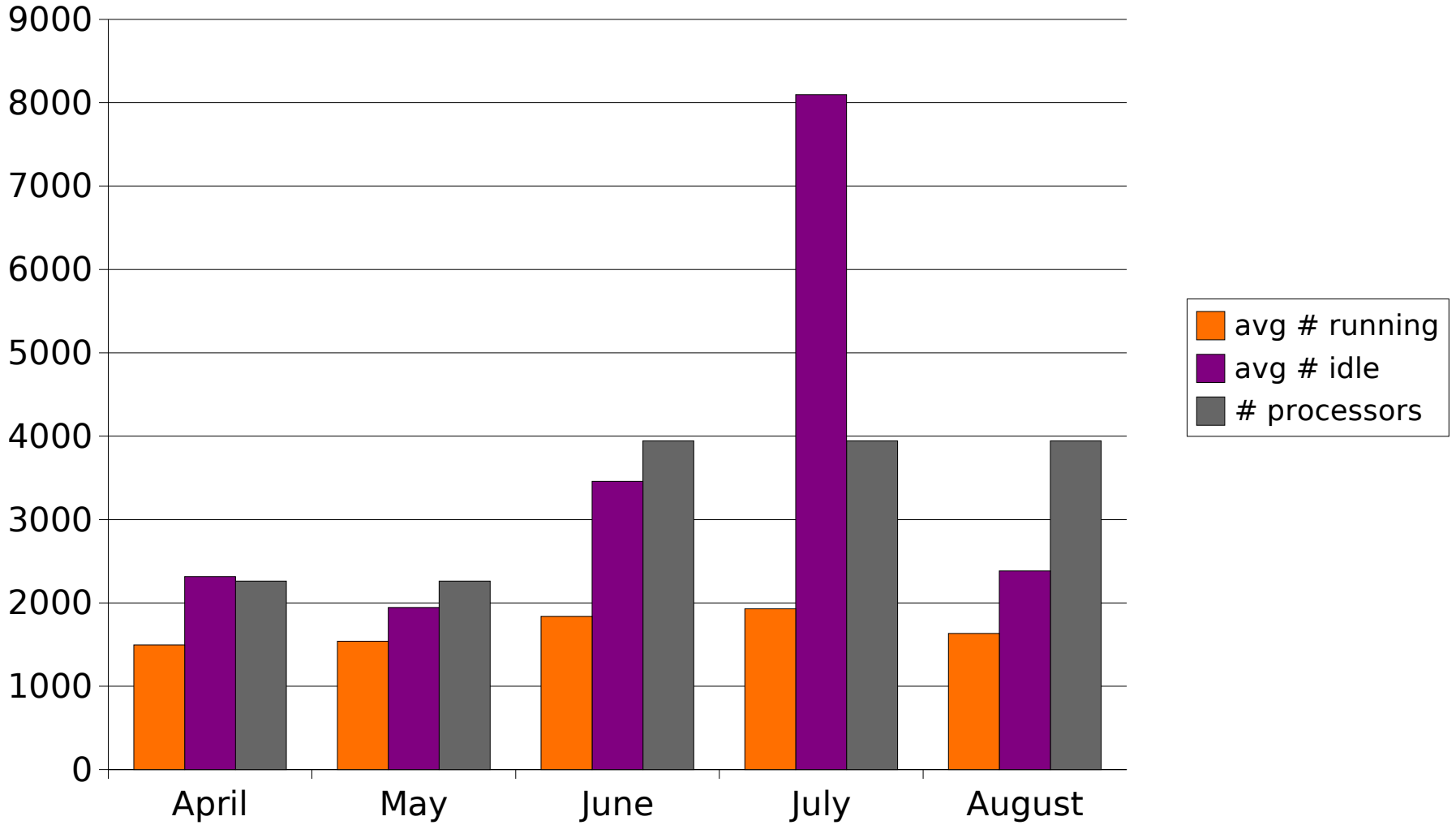
Outline of Configuration

- Two configuration models.
  - Priority scheme is vertical.
    - 8 VMs where VMn+1 > VMn in terms of priority.
    - Suspension used to enforce priority.
    - Only two jobs running at a time.
  - Priority scheme is horizontal.
    - Machines divided into groups.
    - Each group gives a specific job type higher priority.
    - RANK and MaxJobRetirementTime enforces priority.
- Distinction between these two models is sometimes blurred.

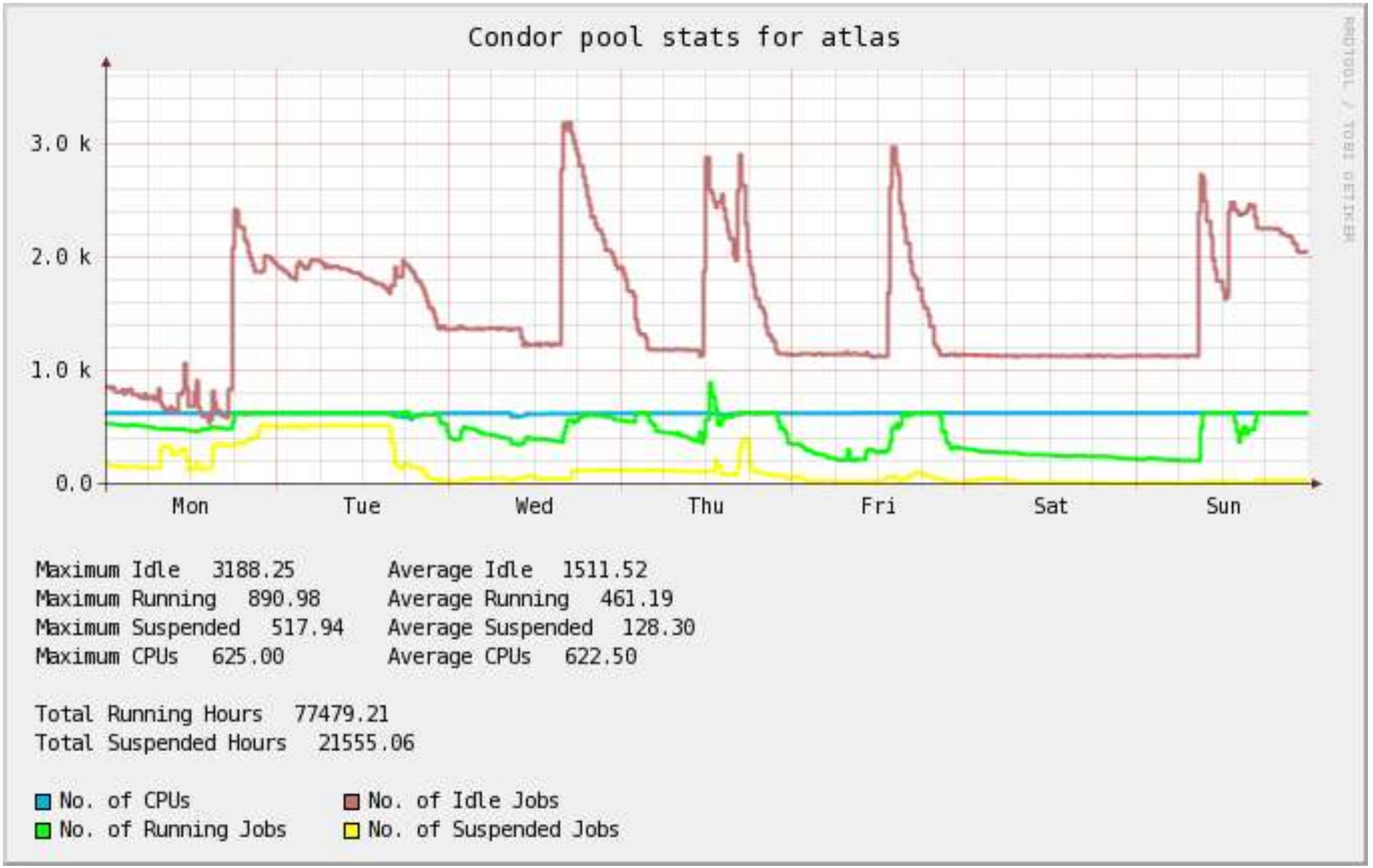
	CAS 1 CPU_Speed == 1	CAS 2 CPU_Speed == 2	CAS 3 CPU_Speed == 3
Highest Priority	Short Jobs (Local Users Only) +RACF_Group = "short"	Short Jobs/Dial Jobs/Software Testing +RACF_Group = "short"	Production Jobs (available to OSG)
	BNL Local Users Jobs +RACF_Group = "bnl-local"	USATLAS Grid and Local Jobs (available to OSG) +RACF_Group = "usatlas"	USATLAS Grid and Local Jobs (available to OSG) +RACF_Group = "usatlas"
	USATLAS Grid and Local Jobs (available to OSG) +RACF_Group = "usatlas"	ATLAS Grid Jobs (Available to OSG and LCG)	ATLAS Grid Jobs (available to OSG and LCG)
Lowest Priority	ATLAS Grid Jobs (available to OSG and LCG)	Non-ATLAS Grid Jobs/General Queue (available to OSG)	Non-ATLAS Grid Jobs/General Queue (available to OSG)
Node Allocation	56 nodes, 112 CPUs (Dell 3.4 GHz)	126 nodes, 252 CPUs (Dell 3.4 GHz) 46 nodes, 92 CPUs (PC 3.06 GHz)	100 nodes, 200 CPUs (Dell 3.4 GHz)

Things run well but...

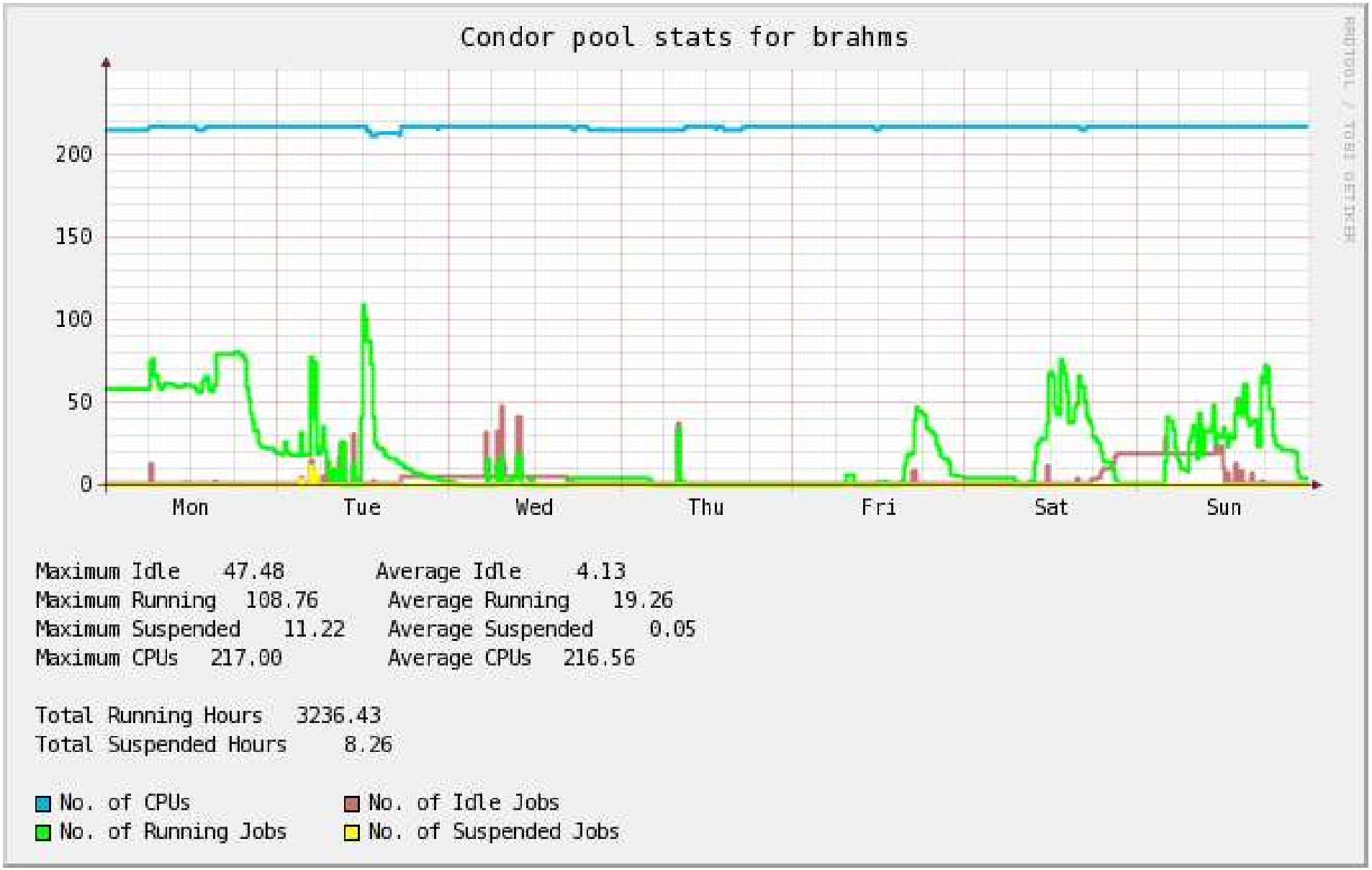
Usage History of all Condor Pools 2005



One pool is well utilized...



...and another is not.



Sharing resources with a “general queue”

- Two ways of sharing considered:
  - Collapse all five pools into one large pool.
    - Using 6.6 but would like to use HAD for this option.
    - Would require altering administrative and monitoring tools.
    - Worried about the increased load.
    - Plan on doing this in the future.
  - Use flocking to effectively combine pools.
    - Easy to implement without drastic changes.
    - Easy to turn off in case things go bad.

Policy for the “general queue”

- Those who own the resources don't want their jobs interfered with.
- RANK jobs from the pool owners higher.
- The conclusion: evict immediately.
- However, have all jobs flow towards unused and least loaded nodes.

Statistics from 03-31-2005 to 03-31-2006

no. jobs completed

destination

	phenix	phobos	star	brahms	atlas	total
phenix		3857	200	1284	161	5502
phobos						
star						
brahms	36	1154	2	106	1	1299
atlas	1185	2951	602	1264		6002

no. jobs evicted before completion

destination

	phenix	phobos	star	brahms	atlas	total
phenix		281	38	176	101	596
phobos						
star						
brahms	3	18				21
atlas	448	1540	45	821		2854

total effective runtime hours consumed by completed jobs

destination

	phenix	phobos	star	brahms	atlas	total
phenix		26866.85	947.45	4735.41	136.16	32685.87
phobos						
star						
brahms	2.84	61.24	0.08	680.22	0.08	744.46
atlas	8468.01	49967.12	137.22	18169.28		76741.63

total ineffective runtime hours consumed (including jobs removed)

destination

	phenix	phobos	star	brahms	atlas	total
phenix		1617.04	160.66	1741.47	2984.32	6503.49
phobos						
star						
brahms	211.81	188.18	34.2	1411.83	0.01	1846.03
atlas	3148.81	29372.01	249.81	12501.66		45272.29

Issues with Flocking

- Schedd hangs on contacting the negotiator when networking issues were present.
  - Not really a flocking issue but exasperated by large FLOCK_TO list.
- Scheduling jobs
  - Very fast for local pool.
  - Can take up to ~20 minutes for a foreign pool when resources are available,
  - Users report problems prematurely.

Some Grid Activities

- PANDA
  - Analysis jobs and production jobs.
  - Submitted through the grid via Condor-G.
  - Restricted to subset of CPUs.
  - Can also run on the “general queue” (in testing).
- OSG and other grid jobs
  - Running on unused atlas nodes.
  - Also runs on the “general queue”.

General Issues Resolved

- Scheduling latency dramatically decreased.
  - SIGNIFICANT_ATTRIBUTES
  - Tweaked various timeouts (neg. cycle, etc.)
- MaxJobRetirementTime makes for a happier user (if less efficient).
- Useful features much needed:
  - CLAIM_LIFETIME and SYSTEM_PERIODIC_*.

Features Needed

- Need job ClassAd which gives user's primary group.
- Transfer output files for debugging when job is evicted.
- Interested in Condor on Demand (COD), but lack of features prevents more usage.
  - Users would like Condor to do scheduling.
  - One reason why LSF is still being used.
- Need more cluster management tools friendly to the vanilla universe.