BNL Site Report

HEPiX at DESY
Spring 2007

Robert Petkus

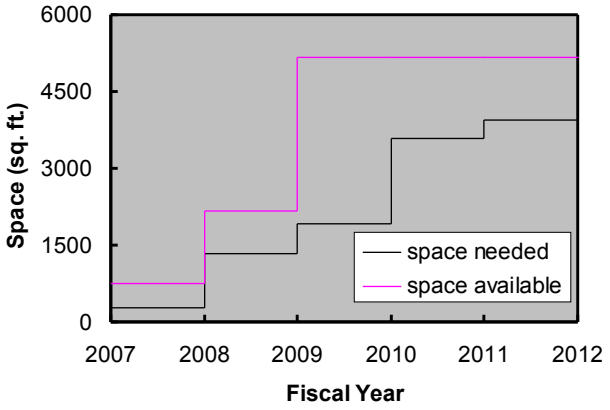
RHIC/USATLAS Computing Facility
Brookhaven National Laboratory

Facility Overview

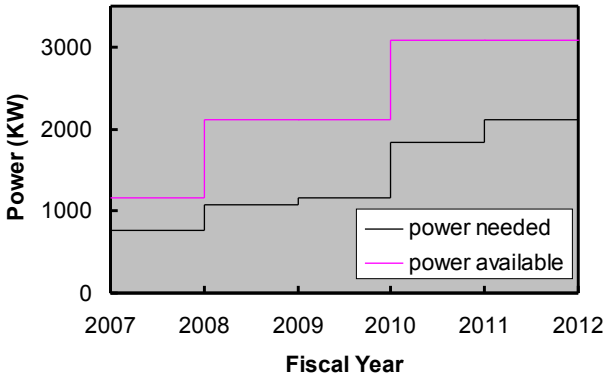
- RHIC/USATLAS Computing Facility is operated by the BNL Physics Department to support the computing needs of three user communities
  - RCF is the the “Tier-0” facility for the PHENIX, STAR, BRAHMS, and PHOBOS experiments of RHIC
  - ACF is the “Tier-1” facility for ATLAS in the U.S.
  - Growing computational component for the LSST (Large Synoptic Survey Telescope)
  - >2500 users, 37 FTEs
  - RHIC RUN-7 Au-Au Operation

Infrastructure Expansion

- Available space is saturated



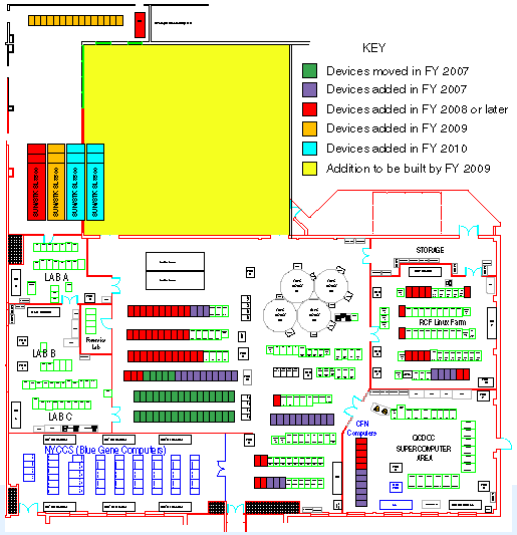
- Available power will be exhausted



A. Chan

Today

2009



Mass Storage

- Tape Robotics Library
  - Two Sun (StorageTek) SL8500s
  - 30 LTO-3 Drives (upgrading to 40)
  - DataDirect S2A 9500 + IBM DS4500 disk cache
  - 16 RHEL 4 Linux data movers
  - 3.9 PB stored data
- HPSS
  - v.5.1 to be upgraded to v.6.2 this summer.
  - Core server upgrade to IBM 8 processor PowerPC
- Oak Ridge Batch System
  - Improved monitoring system has resulted in optimized client behavior

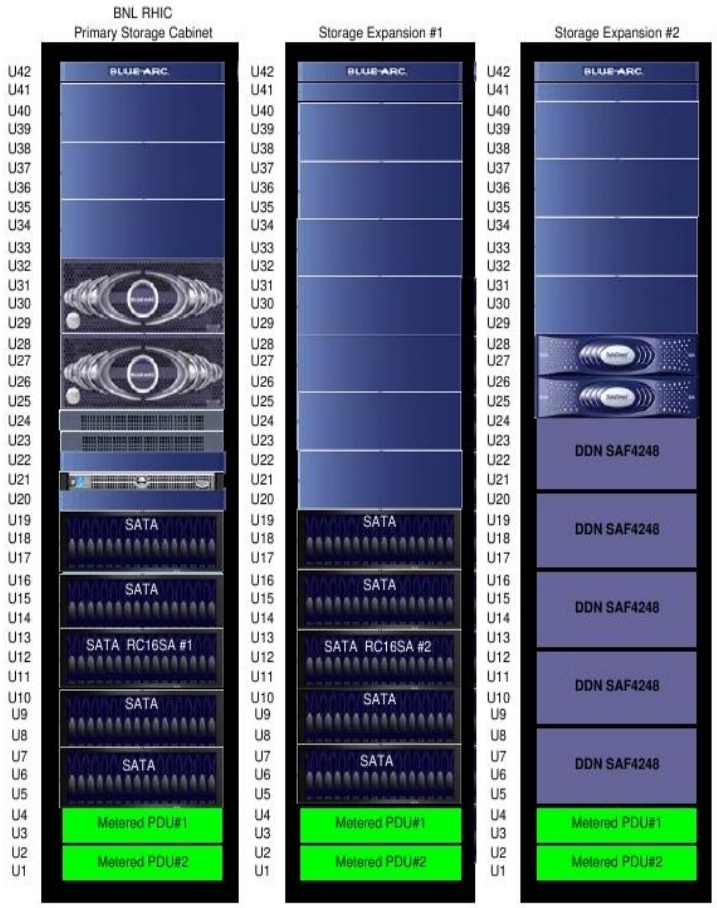
Central Storage

- **AFS**: RHIC and USATLAS cells
  - OpenAFS 1.4.2 / Solaris 10 on Sparc SunFire V240
  - Fileservers using ZFS file system – Issue exists for file systems > 1TB.
  - 8 TB of Fibre Channel DAS
  - TSM for backups are no longer supported
  - Strong desire to migrate to x86 architecture, either Linux or OpenSolaris

    (We need a new backup solution in order to move forward)
- **NFS / Panasas**
  - 100 TB Panasas Storage (20 shelves) accessed via DirectFlow and NFS
    - Imminent warranty expiration. No plans to renew.
  - >200 TB NFS storage. Mixed Solaris 9/SAN and Linux DAS
  - Planned retirement of all Panasas and Solaris NFS storage in the near term

BlueArc Evaluation

- Currently evaluating a BlueArc Titan Cluster to replace Panasas and Solaris NFS (>170 TB usable)
- Advantages:
  - High performance native NFS – no special client-side configuration needed
  - Robust user/group quota implementation
  - Clustering and failover capability
- Evaluation system:
  - (2) Titan 2200s, each capable of 10Gbs
  - (4) Disk subsystems:
    - (5) shelves each SATA LSI and DDN Storage
    - (2) Nexsan SataBeasts
    - (5) shelves FC LSI storage (for comparison)

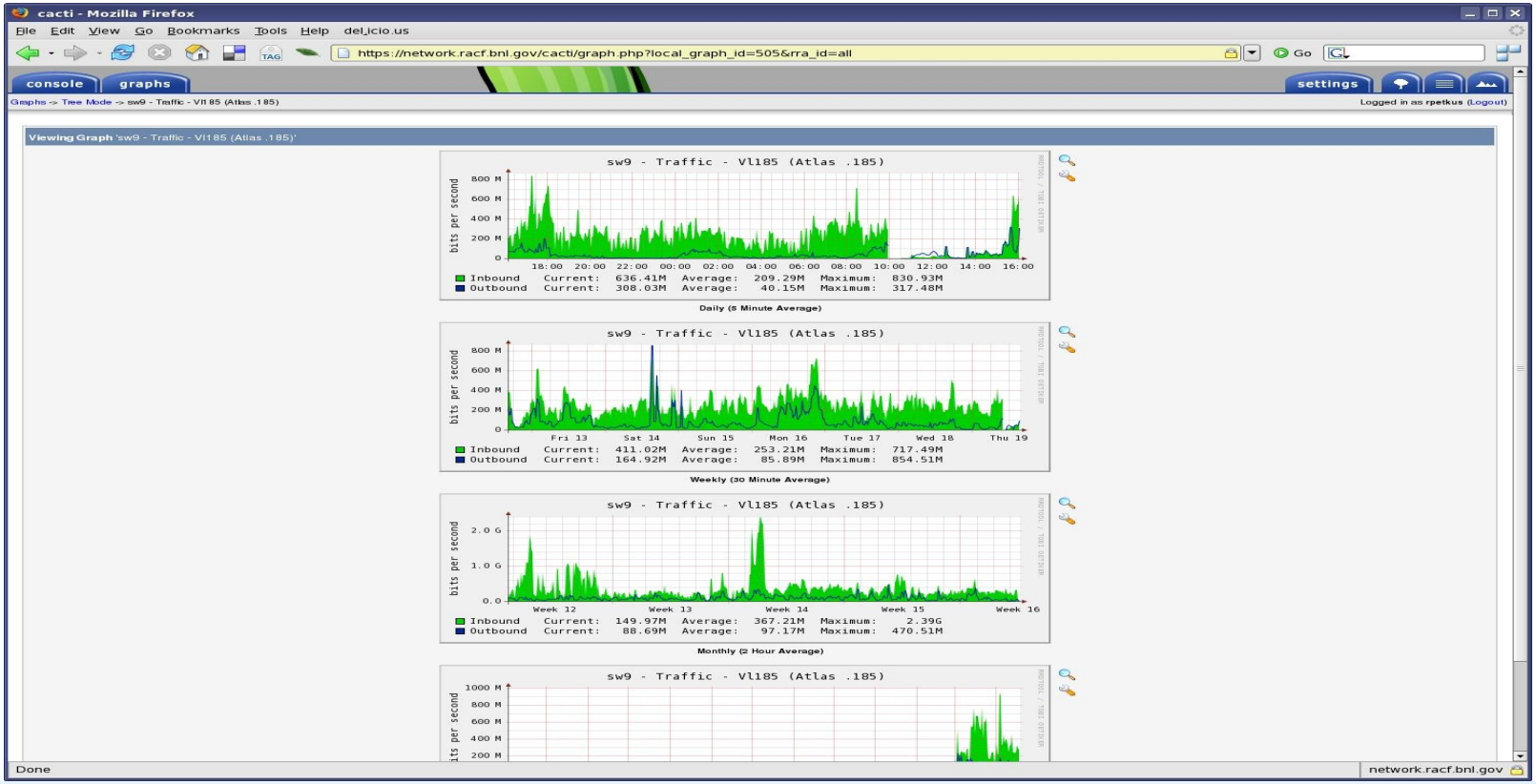


Linux Farm

- >4700 CPUs, >1.3PB local storage
- SL 3.0.5 on RHIC (upgrading to SL 4.x, SL 5?), SL 4.4 on USATLAS
- Procuring 120 nodes for USATLAS
  - Dell 2940 2.66GHz dual core Intel Woodcrest processors
  - 1.21M SI2k
  - 540 TB local storage
    - 6 750GB SATA disks per server, Hardware RAID 0
- Small Xen test deployment
  - Assess overhead and functionality
  - Containerize services
  - SL 4.4 with kernel.org kernel
    - Xen RPMs did not provide adequate source code to compile OAFS modules

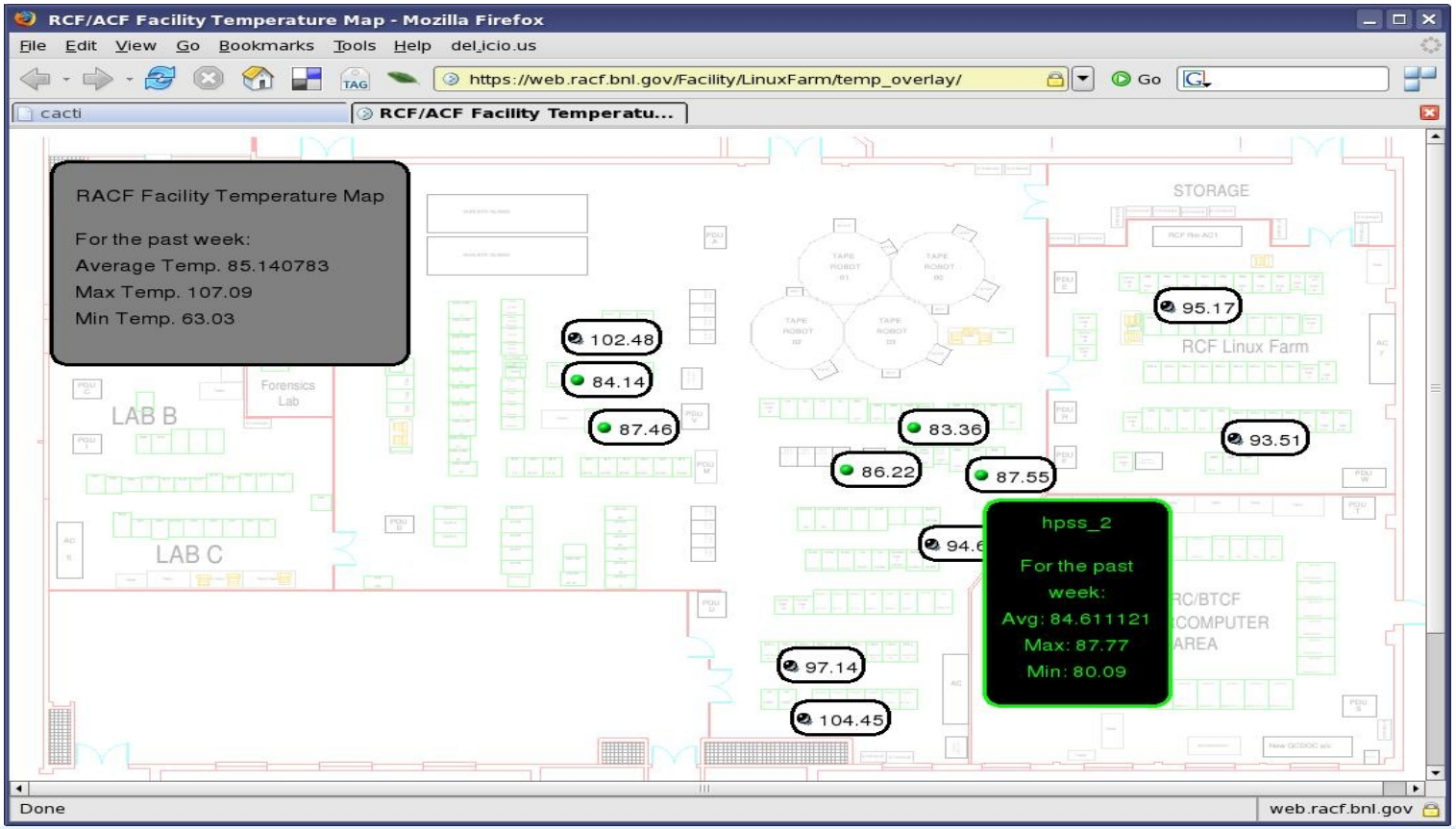
Monitoring

- Linux farm and infrastructure servers monitored primarily with Nagios
  - ~2000 hosts with an average of 8 services/host (dCache service checks added)
- Ganglia used to monitor Linux farm performance
- Cacti deployed for network bandwidth monitoring



Temperature Monitoring

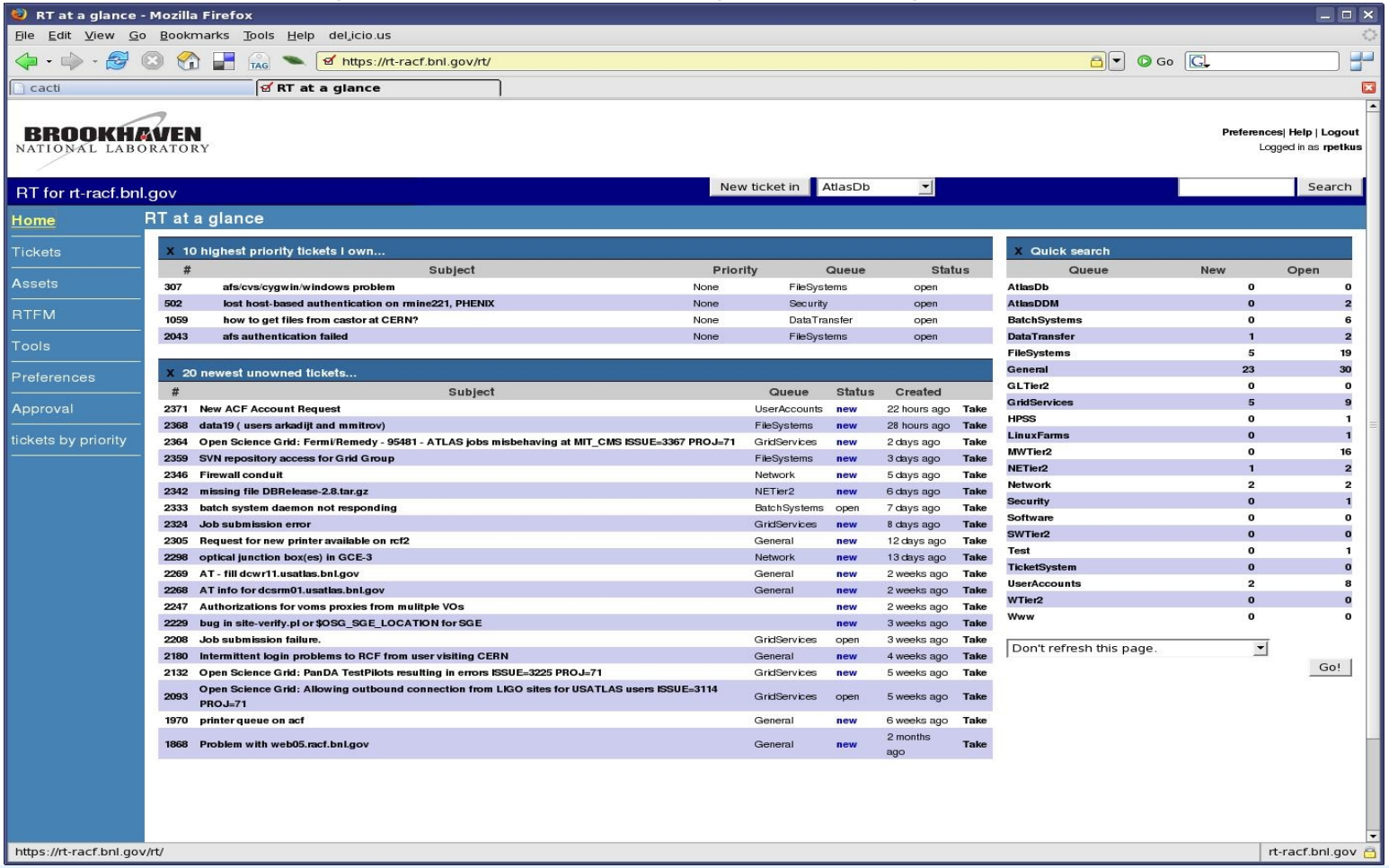
- Temperature is monitored across the entire data center via strategically distributed sensors
- Alerts automatically generate a ticket with the RT ticketing system
- Visual aiding: information is mapped to geography



A. Withers ->

Request Tracker

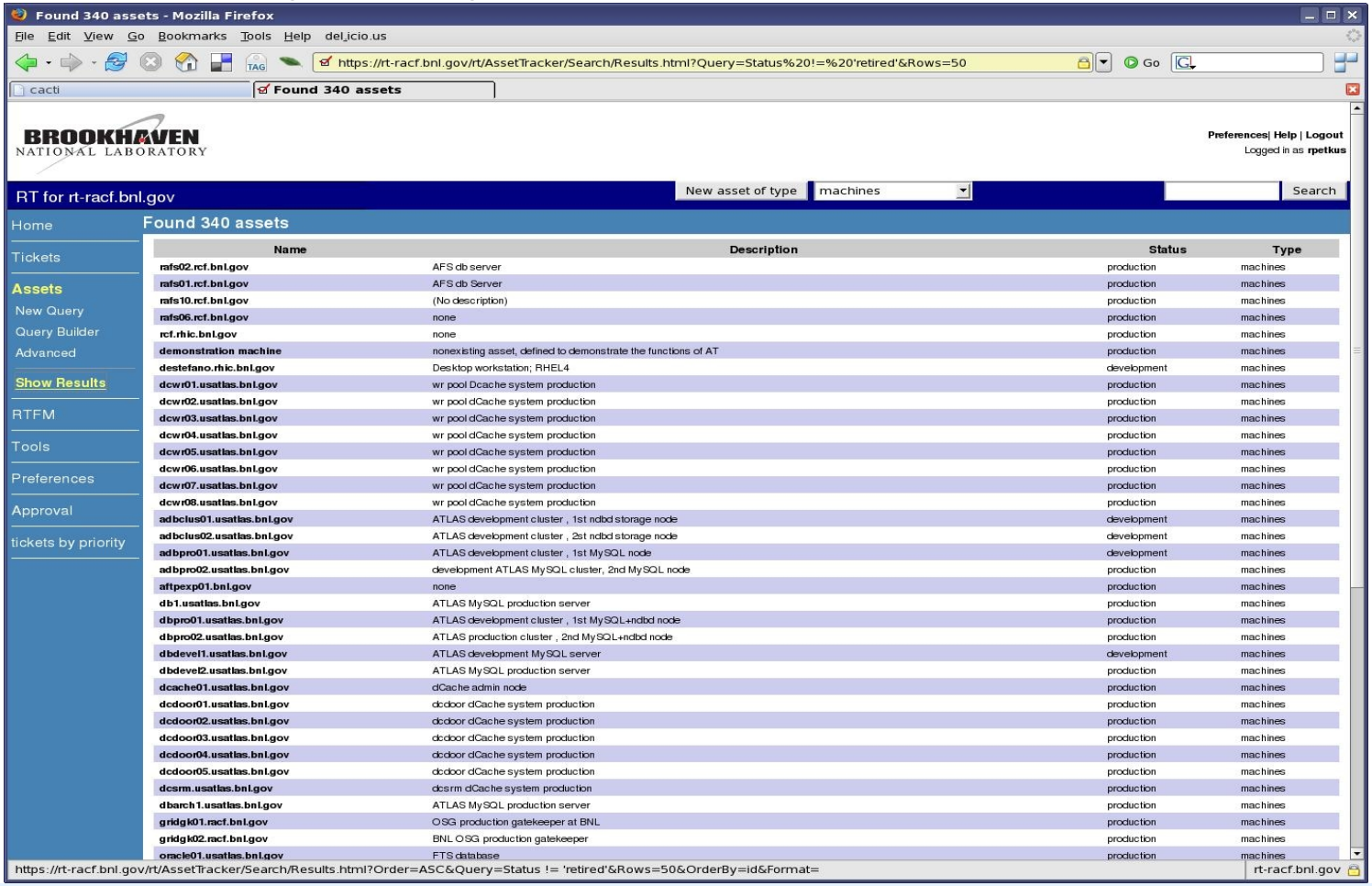
- RT (Request Tracking): an open source package for tracking user requests



T. Wlodek ->

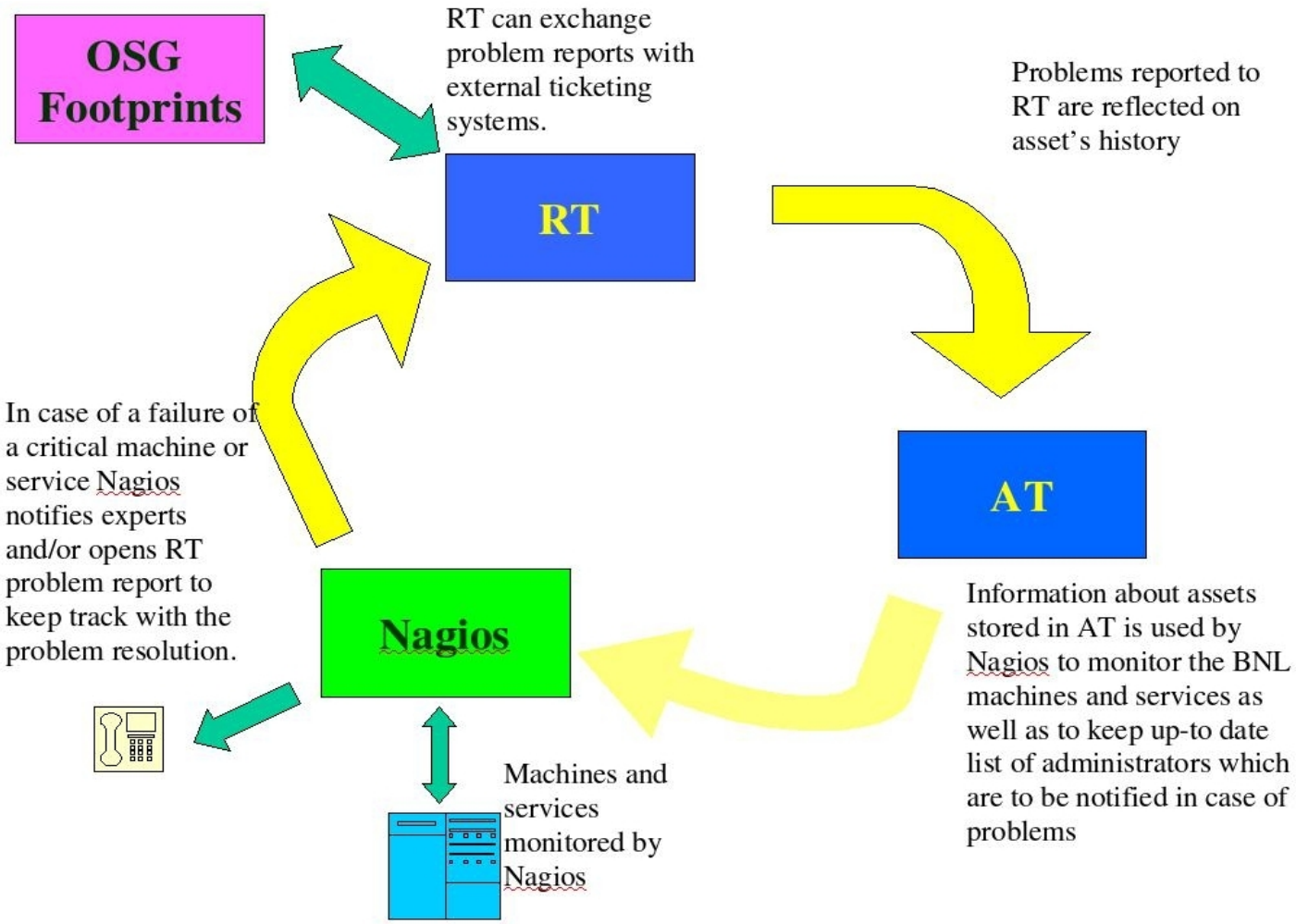
Asset Tracker

- AT (Asset Tracking): a RT plug-in to track assets (HW, SW, services, licenses, etc.)



T. Wlodek ->

RT – AT – Nagios Integration (T. Wlodek)



Batch Systems

- Upgrade of all nodes to Condor 6.8.4
- Quill database performance testing. A number of HW/SW configurations were evaluated to boost performance
  - Optimal performance/price using (2) dual-core Dell 2950, 8GB RAM, HW RAID 10 with 6 SAS drives

	SATA	SAS			
	altus1300_raid10 4 drives, SW raid	pe2950_raid5 3 drives, HW raid	pe2950_raid10 4 drives, HW raid	pe2970_raid5 3 drives, HW raid	pe2970_raid10 4 drives, HW raid
Avg TPS 10 clients 100 transactions	342.6	300.7	1157.6	674.3	370.9
Avg TPS 100 clients 100 transactions	314.6	544.1	1391.9	704.4	520.8

A. Withers

Distributed Storage

- Dcache Read Pools
  - Phenix: >200TB on 365 servers / 450 pools
  - Atlas: >430TB on 460 servers / pools
- Dcache Write Pools:
  - Atlas: 13 write pool nodes
    - Poor performance/throughput. Currently evaluating disk/server systems to satisfy demand. Need sustained 600MB/sec in, 600MB/sec out.
    - Current testing using IBM DS4700 SATA storage
    - Other promising contenders are the SunFire x4500 and Nexsan Satabeast
- Xrootd
  - STAR: >270TB on 420 servers (Largest deployment of Xrootd)
    - Carefully tuned Xrootd load-balancing policy allows co-existence of computation and data store on the same node
  - ATLAS: Small 10 node test-bed deployed to explore functionality and performance within the ATLAS analysis framework

ATLAS Tier-1 Activities

- BNL Tier-1 is the largest ATLAS Tier-1 and is delivering capacities consistent with this role

WLCG Accounting: ATLAS Tier-1's + CERN Apr - Oct 2006

	CPU use		disk occupancy		tape occupancy	
	KSI2K-days	% of total	TB at end of period	% of total	TB at end of period	% of total
CERN Tier-0 + CAF	95,858	28%	182	48%	469	35%
ASGC	13,413	4%	20	5%	13	1%
BNL	88,184	26%	48	13%	357	27%
CC-IN2P3	24,264	7%	15	4%	153	12%
CNAF	20,108	6%	18	5%	95	7%
FNAL	4,619	1%	-	0%	-	0%
FZK-GridKA	23,195	7%	26	7%	115	9%
NDGF	18,761	6%	28	7%	-	0%
NL LHC/Tier-1	14,574	4%	10	3%	18	1%
PIC	6,207	2%	8	2%	54	4%
RAL	27,672	8%	14	4%	54	4%
TRIUMF	1,876	1%	7	2%	-	0%
TOTAL	**338,731**	**100%**	**376**	**100%**	**1,328**	**100%**

M. Ernst

ATLAS Tier-1 Activities, cont.

- OSG 0.4x, gLite 3.02 (partial deployment)
- FTS MyProxy
  (File Transfer Services + Authorization)
- SiteBDII (Information service provider)
- ATLAS Panda for production and analysis
  - Development of “Pilot” factory system using Condor-G/C
- GUMS (Grid User Management System)
  - OSG sponsored
  - All admin tools integrated into web interface
  - New privilege classes (ACLs) added
- Terapaths
  - End-to-End virtual network paths with QoS guarantees
  - Budget allocated for integration at (5) Tier-2 sites
- GSTAT
  - Publish dynamic distributed storage space allocation