

RHIC Computing Facility

Michael Ernst

**DOE/Nuclear Physics Review of RHIC
Science and Technology**

18 - 20 July 2007

RHIC Computing Facility (RCF)

- Organizationally established in 1997
- Staffed as a Group in Physics Department
- Equipment physically located at Brookhaven Computing Facility (BCF)
 - BCF operated by ITD
- Currently co-located and co-operated with the ATLAS Computing Facility (ACF), the U.S. ATLAS Tier-1 Regional Center
 - ACF ramping up quickly, currently
 - ACF capacities are ~ 65% for processing, 121% for disk capacity
 - ACF staff level ~ 75% of RCF

RCF Mission and Scale

➤ Mission

- ❑ Online Recording of Raw Data
- ❑ Production reconstruction of Raw Data
- ❑ Primary Facility for Data Selection and Analysis
- ❑ Long time Archiving and Serving of all Data

➤ Scale

- ❑ Authorized staff of 20 FTE's
- ❑ Historically ~\$2M/year equipment replacement funding (25% annual replacement) – Last year limited to \$1.3M, current year limited to \$1.7M
 - Addressing obsolescence
- ❑ Growth beyond originally planned scale will require an increase in the funding

Experiment / RCF Interaction

- **Weekly Liaison Meeting**
 - ❑ Addressing operations issues
 - ❑ Review recent performance and problems
 - ❑ Plan for scheduled interventions
- **Experiments / RCF Annual Series of Meetings to develop Capital Spending Plan**
 - ❑ Estimate scale of need for current/coming run
 - ❑ Details of distribution of equipment to be procured
 - ❑ Most recent in early Spring for FY-07 funds
- **Periodic Topical Meetings, examples**
 - ❑ ~Annual Linux Farm OS upgrade planning
 - ❑ Replacement of Central Disk Storage
- **Other User Interactions**
 - ❑ Web site
 - ❑ Ticket System (Request Tracker (RT – Open Source))
 - Fully replaced in-house developed Trouble Ticket System (CTS)
 - ~3000 Tickets for RHIC & ATLAS Services (last 12 months)

Computing Requirements Estimate

- A Comprehensive Long Range Estimate done by PHENIX, RCF and STAR in Fall / Winter 2005
 - ❑ Conclusions published as part of “Mid-Term Strategic Plan: 2006-2011 For the Relativistic Heavy Ion Collider”, February 14, 2006
 - ❑ Needs to be revisited/updated
 - Lack of disk space has an obvious impact on analysis performance
- Input is Raw Data Volume for Each Species & Experiment by Year
- Model for Requirements Projection
 - ❑ Assume Facility resource needs scale with Raw Data volume
 - ❑ With adjustable parameters reflecting expected relative ...
 - Richness of data set (density of interesting events)
 - Maturity of processing software
 - Number of reconstruction passes

... for each experiment, species, and year

Computing Cost Estimate

- Requirements Model Output used as input to Cost Estimate
- Costing Model is based on
 - ❑ Recent procurements
 - ❑ Historic Trends (Moore's Law and similar technology based trends)
 - ❑ Use of inexpensive disk for bulk of storage
 - Linux processor farm distributed disk
 - Raid 6/ZFS based Storage Farms
 - ❑ Assume use of obvious technology evolution (multi-core processors), etc.)
 - ❑ For running scenarios considered, capacity growth associated with replacement of obsolete equipment meets increased capacity requirements in 2007 but increase of equipment funding is required in 2008 and beyond
 - Required capacities by year and a funding profile allowing them to be achieved are shown on following slide

Requirements Estimate for a Particular Running Scenario

	FY '06	FY '07	FY '08	FY '09	FY '10	FY '11
Annual Requirement						
<i>Real Data Volume (TB)</i>	1700	2700	3500	4500	7200	8600
<i>Reco CPU (KSI2K)</i>	600	1000	2900	4700	8500	9700
<i>Analys CPU (KSI2K)</i>	310	570	1800	2700	4600	5400
<i>Dist. Disk (TB)</i>	220	480	1500	1700	3000	3600
<i>Cent. Disk (TB)</i>	30	60	190	260	450	560
<i>Annual Tape Volume (TB)</i>	2000	3200	4200	5400	8700	10300
<i>Tape bandwidth (MB/sec)</i>	690	920	920	1700	2100	2300
<i>WAN bandwidth (Mb/sec)</i>	1400	2000	2100	4300	5700	6700
<i>Simulation CPU (KSI2K)</i>	110	200	610	1000	1800	2100
<i>Simulation Data Volume (TB)</i>	330	530	710	900	1400	1700
Installed Requirement						
<i>CPU (KSI2K)</i>	2100	2800	6800	11800	20800	27700
<i>Dist. Disk (TB)</i>	480	720	1900	2600	4300	5700
<i>Cent. Disk (TB)</i>	200	200	290	400	650	880
<i>Tape Volume (TB)</i>	4300	7500	11800	17200	25900	36200
<i>Tape bandwidth (MB/sec)</i>	920	1400	1600	2500	3300	4000
<i>WAN bandwidth (Mb/sec)</i>	1500	2700	3500	6100	8700	11000

Funding Profile (\$K)

	FY '06	FY '07	FY '08	FY '09	FY '10	FY '11
<i>CPU + Distributed Disk</i>	270	770	1360	790	1270	1960
<i>Central Disk</i>	150	250	400	330	450	310
<i>Tape Storage System</i>	640	590	250	1060	570	250
<i>LAN</i>	120	190	250	270	320	180
<i>Overhead</i>	130	200	250	270	290	300
Total Annual Cost	1310	2000	2510	2720	2900	3000

Principal RCF Services

- General Collaboration and User Support
- Processing Services (Linux Farm)
 - ❑ Programmatic Production processing
 - ❑ Individual and Group Analysis
- Online Storage (Disk)
 - ❑ Data storage for work area (Read / Write)
 - ❑ Data serving for Analysis (> 90% Read)
- Mass Storage (Robotic Tape System)
 - ❑ Raw Data recording and archiving
 - ❑ Derived Data Archiving
- Grid & Network Services

RCF Staff

- Current authorized staff level: 20 FTE's
- Excellent synergy in the context of a co-located ATLAS Tier-1 Center in terms of operations
 - ❑ Very high level of commonality
 - ❑ A dramatic divergence in technical directions could change this, but this seems very unlikely
- It does not allow for aggressive involvement in new technologies
 - ❑ Effort spent primarily on Integration and Operation

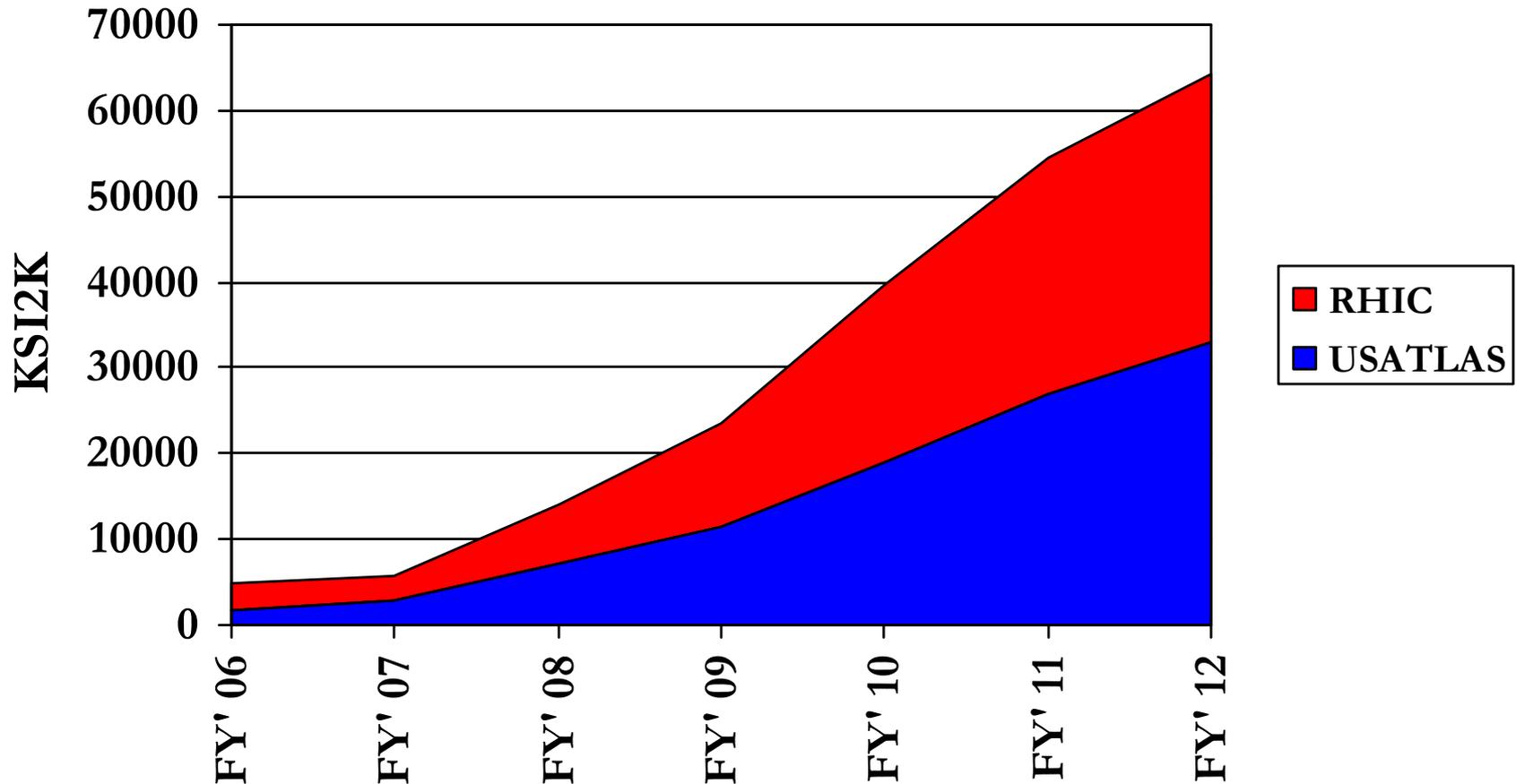
	Current FTE's	Target FTE's
Linux Farms	3.5	3.5
Mass Storage	4.2	4.2
Disk	2.6	2.6
User Support	2.9	2.9
Fabric Infrastructure	2.1	2.6
Wide Area Services	1.8	1.8
Admin	2.5	2.5
Total	19.5	20.0

Compute Servers

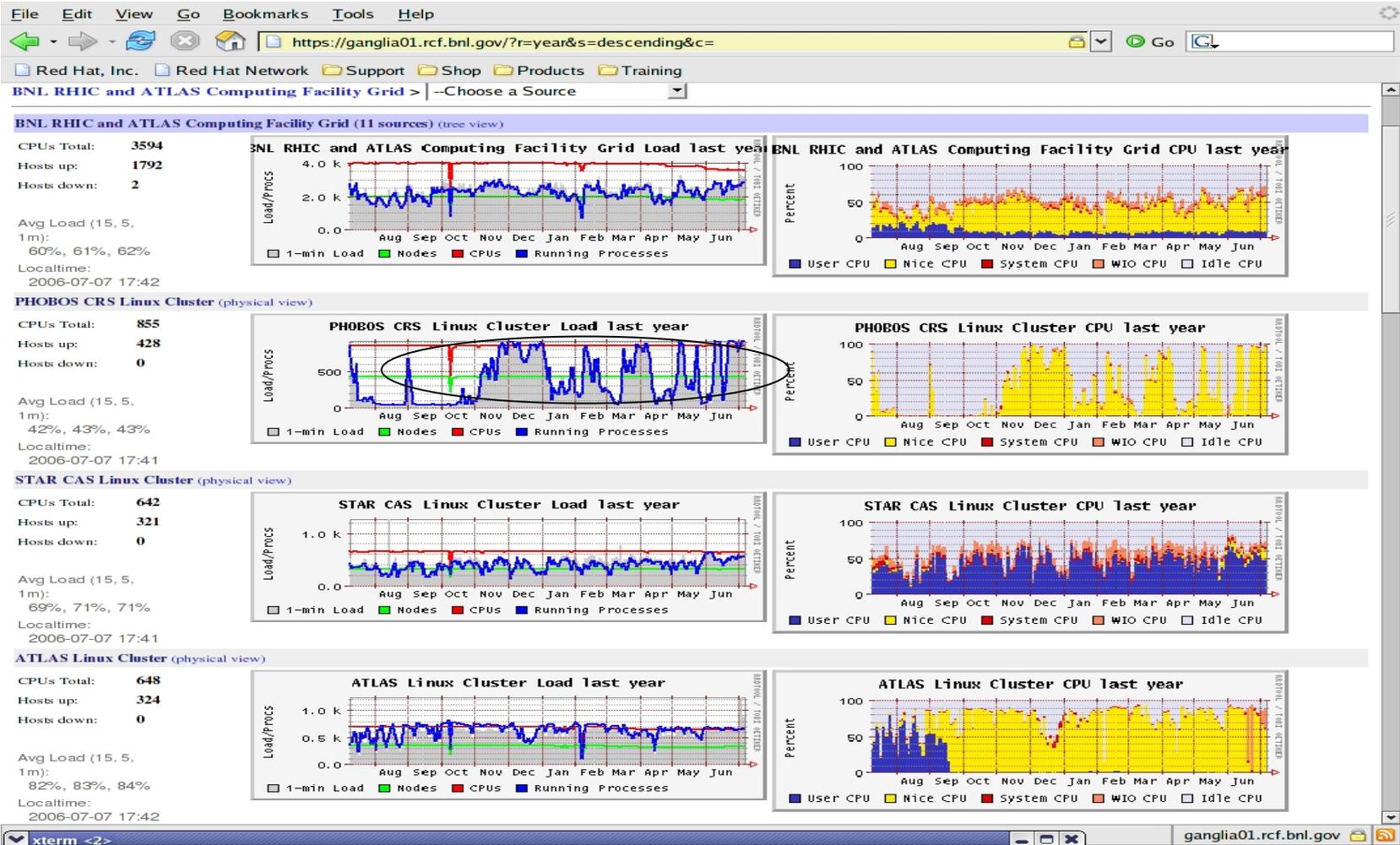
➤ Three Generations of Linux CPU rack mount systems

- ❑ Dual CPU (single core) systems (2600 SI2k per box, bought in 2002)
 - ❑ Dual CPU (dual core) systems (4600 SI2k – 10,000 SI2k per box)
 - ❑ Dual CPU (quad core) systems (20,000 SI2k per box)
- } x 8
- ❑ Currently 1,400 compute servers with 2,800 CPU's (4200 cores)
 - Lack of funding does not allow a timely “refresh” of equipment
 - Requires more space, power and cooling than anticipated
 - ❑ ~100 additional Dual CPU / Quad Core machines (8 cores / box) with 2 MSI2k
 - Delivery expected by end August
 - Multi-core CPU technology also addresses power/cooling barrier by finessing non-linearity of power consumption with clock speed
 - ❑ Expect to address future requirements by continuing to follow Moore's Law price/performance in commodity market (multi-core, 64 bit advances)

Expected Computing Capacity Evolution



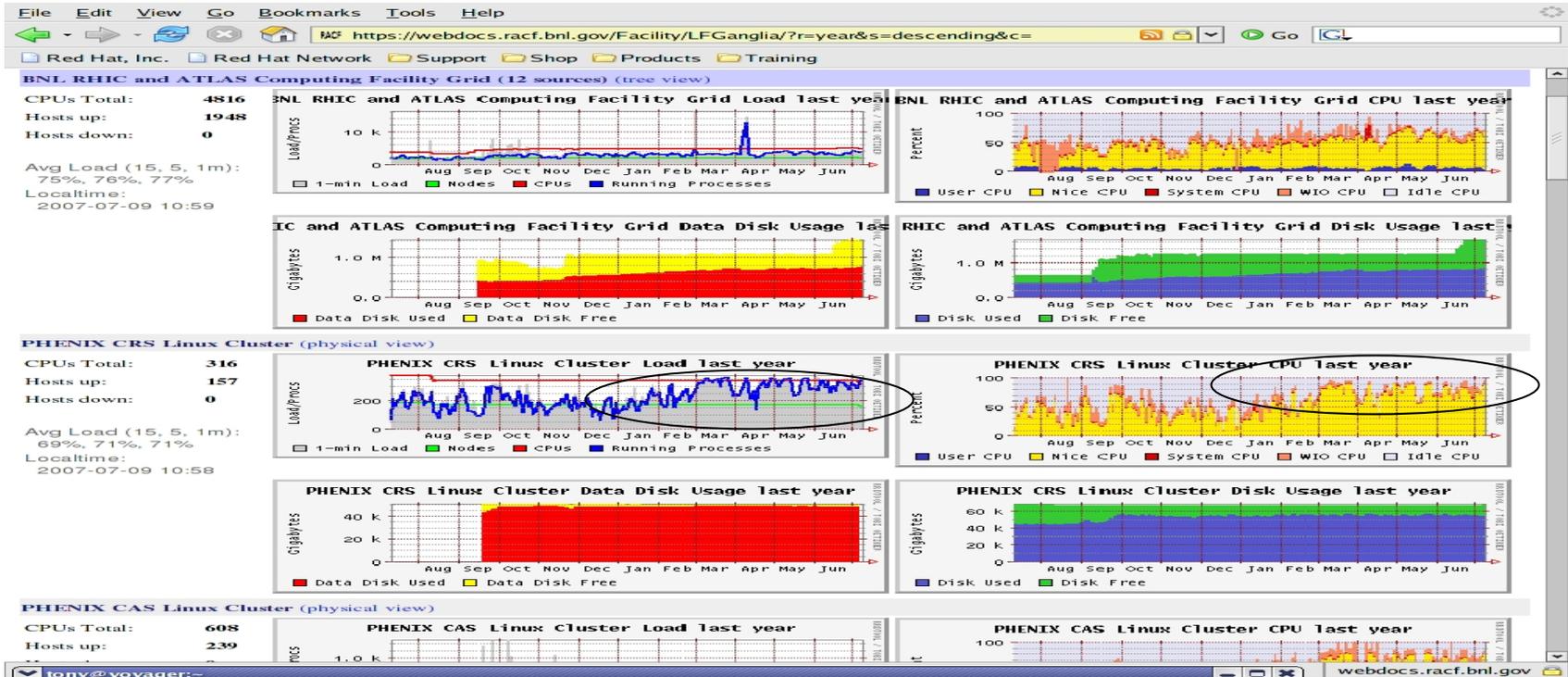
Resource Utilization Issues in 2006



Resource Sharing among Experiments

- Goal was to make idle cycles available in processor farms to other user communities without impact to “owner”
- Mechanism is to evict “guest” jobs when “owner” needs cycles
 - Consider extended rights for guests
 - Allow guest job to complete by grace period (implemented but currently not used)

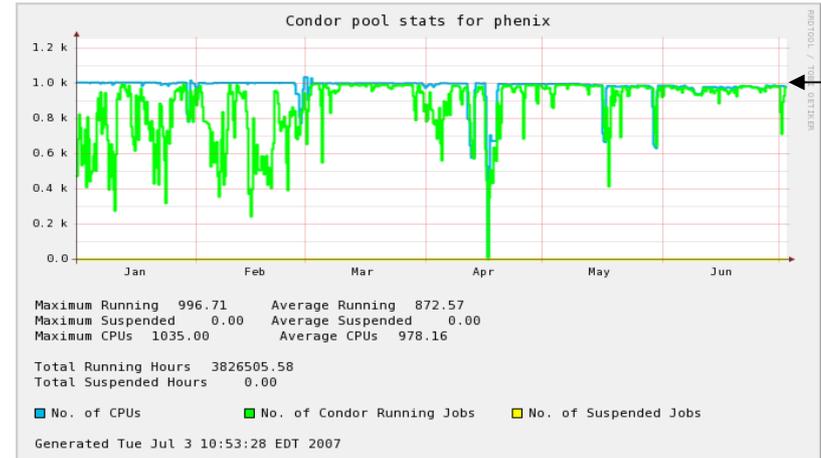
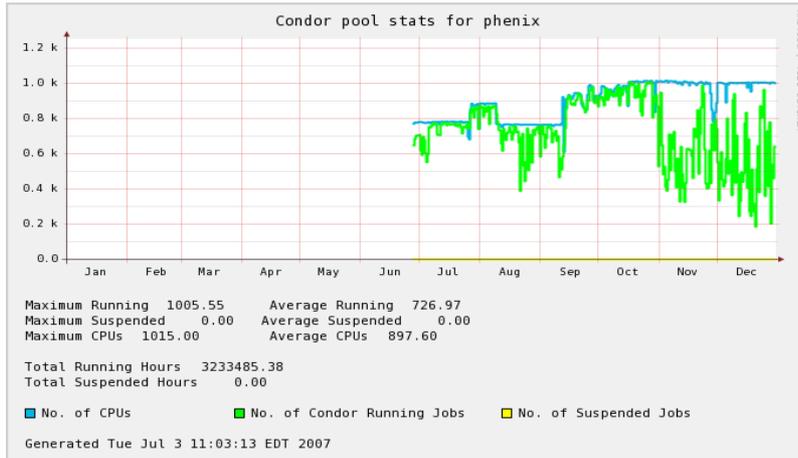
Resource Utilization in 2007



- Average load of 77% for the past 12 months.
- Average load of ~62% for 07/05 to 07/06 (2006 review).
- Excluding interactive nodes, maximum possible load is ~94%.

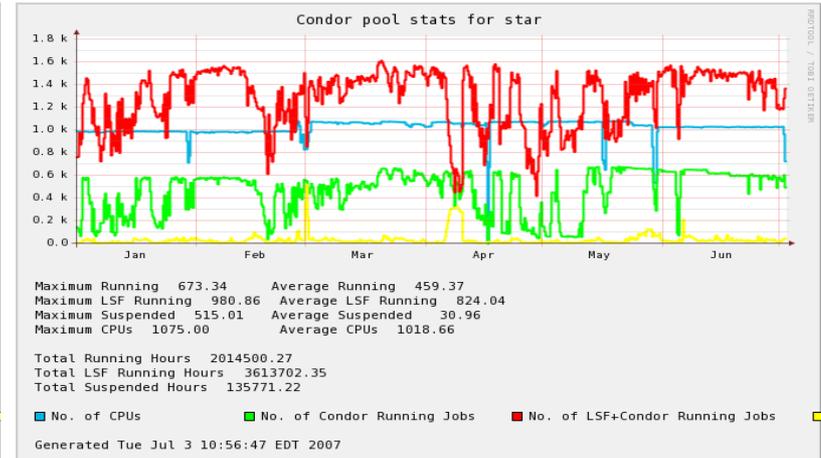
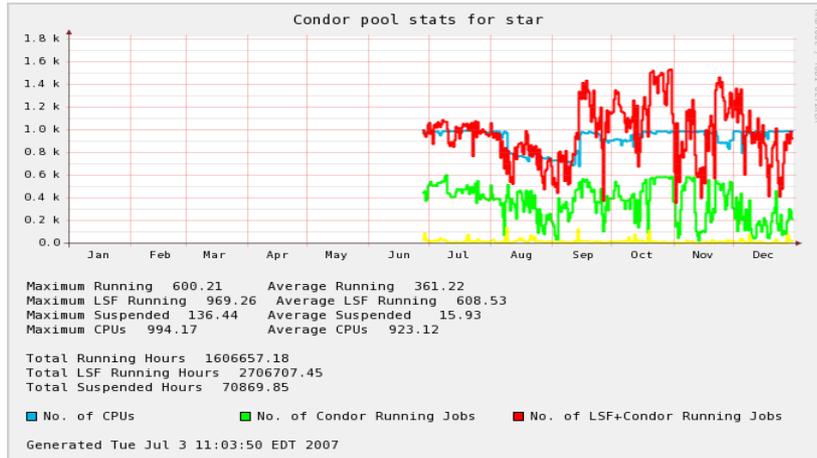
Condor Occupancy (PHENIX)

Upper Limit



- Left-hand plot is for late June'06 to 12/31/06.
- Right-hand plot is for 01/01/07 to 07/03/07.
- Occupancy rose from 81% to 89% between the two periods.

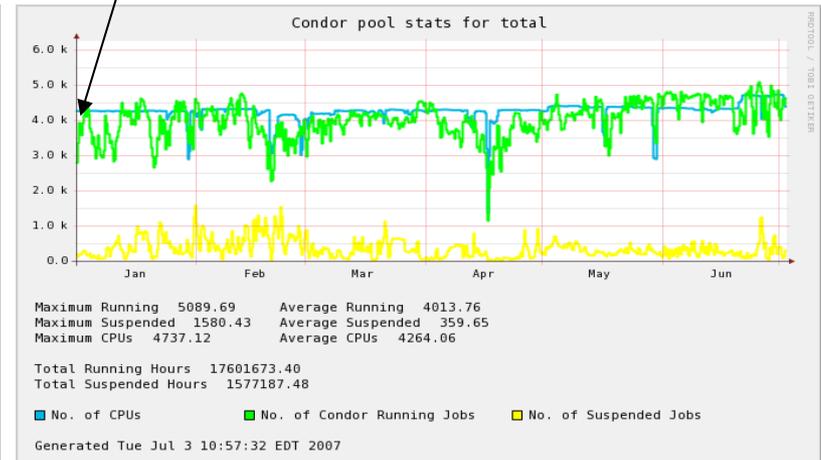
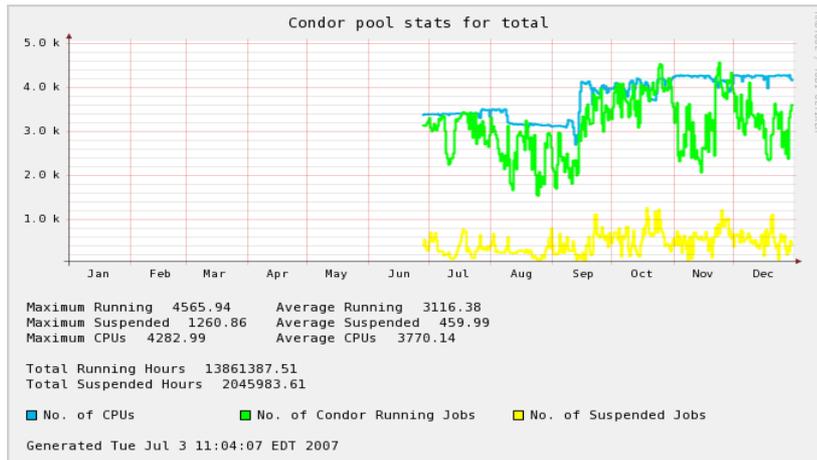
Condor/LSF Occupancy (STAR)



- Left-hand plot is for late June'06 to 12/31/06.
- Right-hand plot is for 01/01/07 to 07/03/07.
- Occupancy rose from 105% to 126% between the two periods.

Condor Occupancy (RACF)

4,200 Job Slots



- Left-hand plot is for late June'06 to 12/31/06.
- Right-hand plot is for 01/01/07 to 07/03/07.
- Occupancy rose from 83% to 94% between the two periods.
- Created general queue in 2006 to increase occupancy.

Online (Disk) Storage

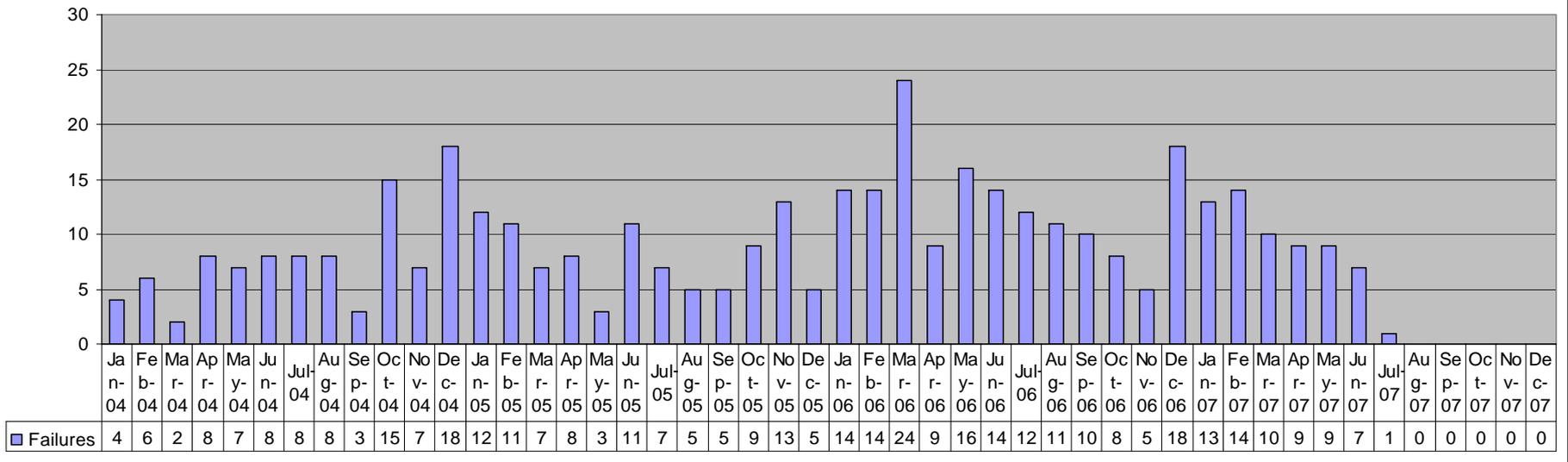
- Historic RCF model was Sun / NFS served RAID 5 SAN connected central disk for all storage areas
- Growth in demand drove disk costs to match and exceed CPU costs
- Current strategy: Differentiate disk technology by function
 - Central Disk
 - Limited amount of NFS “full function” (fully Posix compliant) disk for Read/Write
 - Working on a backup solution (selective)
 - “Read only” Disk
 - Majority on less expensive distributed disk (on Farm nodes) and integrated in storage farms for “mostly Read” of data on secure medium (tape)

“Full Function” Disk Service

- Read/Write (Posix compliant), reliable, high performance and high availability – NFS served RAID systems
 - ❑ Historically
 - ~150 TB of Sun served RAID 5 disk
 - ~70 TB of Panasas (appliance) served RAID 5 disk
 - ❑ Acquisition in 2006
 - ~100 TB of Nexsan & Aberdeen Linux served RAID 5/6 disk
 - ❑ Movement to lower Tier of RAID disk vendors last year
 - Product from expensive vendor failed to fulfill expectations
 - Inexpensive RAID systems unable to sustain the load
 - **Too many concurrent processes**
 - ❑ Very bad situation in early 2007
 - Many service disruptions due to old and unreliable equipment
 - Services distributed on too many different products
 - Negative impact on user efficiency (losing jobs, eventually losing data)
 - Two FTE’s constantly occupied to keep the service operational

Central Disk Failures over Time

GCE Failures 1/04 - 12/07



Consolidation of Central Disk

➤ Have initiated a Storage Evaluation Project

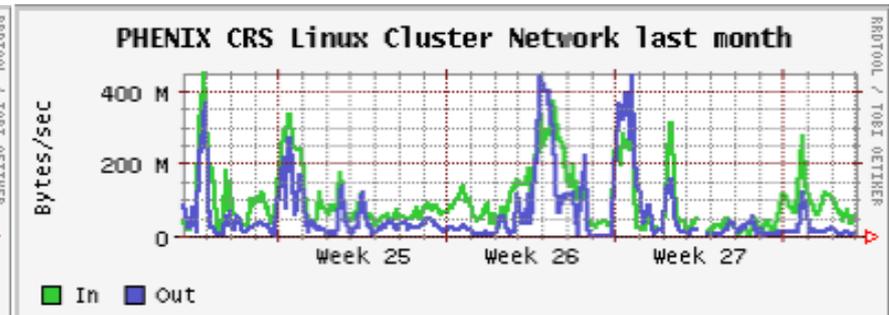
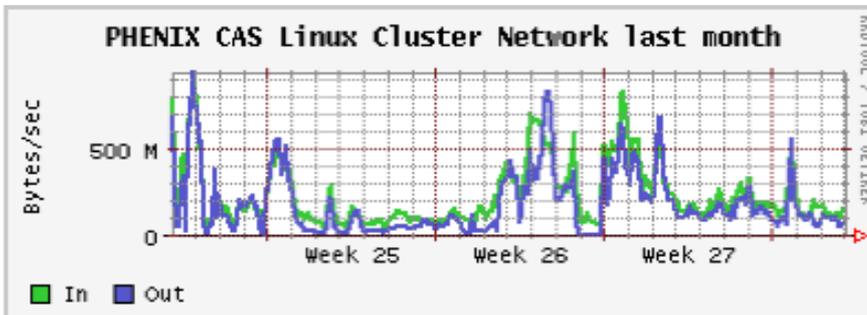
- ❑ Multiple vendors of appliances and disk storage backends were asked to bring in equipment for evaluation
- ❑ Ran relevant low level and (physics) application driven tests
 - Massive concurrency (100's of clients)
 - Read/Write performance oriented tests
 - Resiliency and fail-over tests
- ❑ Tests took longer than expected
 - Though recommended by the vendor of NAS Head the backend storage performance of SATA based disk backends was poor
 - **Unable to fix, despite massive amount of effort spent by vendors**
 - **FC disk based backend the only solution satisfying RACF's performance and resilience requirements**
- ❑ Purchasing a 200 TB system from BlueArc / Hitachi
 - To replace equipment older than 3 years (Panasas, MTI, Zyzsx)
 - Requires ~50% of FY'07 funds (not much left for processing)
 - High performance, high-availability storage at very competitive Cost (~\$3.5/GB)

“Mostly Read” Disk

- Disk deployed on Linux processor farm nodes
 - ~3 x less expensive than full function disk
 - No RAID, JBOD (Just a Bunch Of Disks)
- Requires additional storage management software
- Two such storage management systems currently in use at RCF
 - dCache – DESY/Fermilab developed Grid-aware S/W package
 - Scalable, robust, unified pool of independent storage components with integral Mass Storage backend, posix-like data access, ROOT support
 - ATLAS is major BNL user with 850 TB => 1,500 TB by end July 2007
 - Xrootd – SLAC, CERN, BNL + other community developers
 - STAR is major BNL user with ~300 TB managed capacity
 - **Heavily used for more than 2 years**

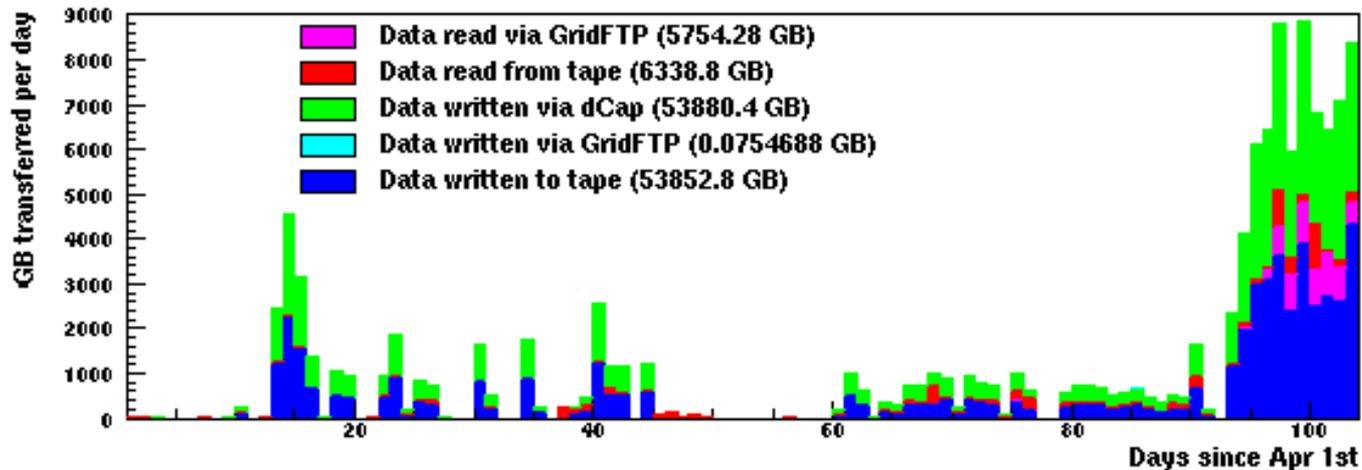
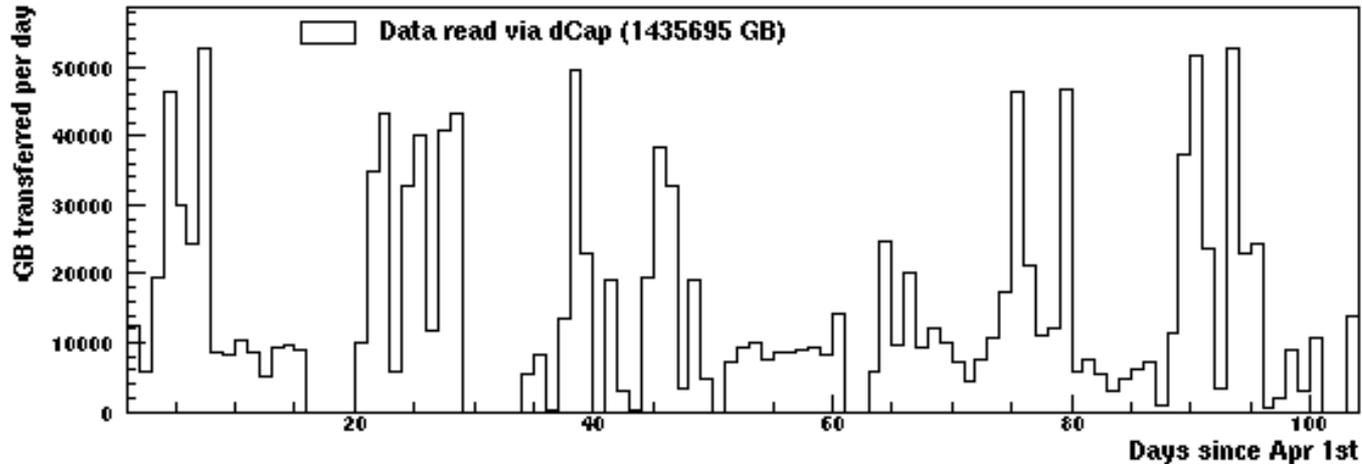
dCache Usage by PHENIX

- Usage is dominated by data transfer on LAN
 - ❑ Aggregate Throughput up to 1.5 GB/s
- Repository and Archiving mechanism for data production stream
- Integrated into “Analysis Train”
 - ❑ Aggregates user analysis jobs to run efficiently on common data subsets
 - ❑ Access restricted by policy to train “operators”
 - ❑ Increasing WAN transfer (to IN2P3)



PHENIX Transfer Statistics

Phenix dCache Statistics, Apr to Jul 2007



More Information on Storage Management

➤ STAR and Xrootd

□ Xrootd / Scalla rationales are

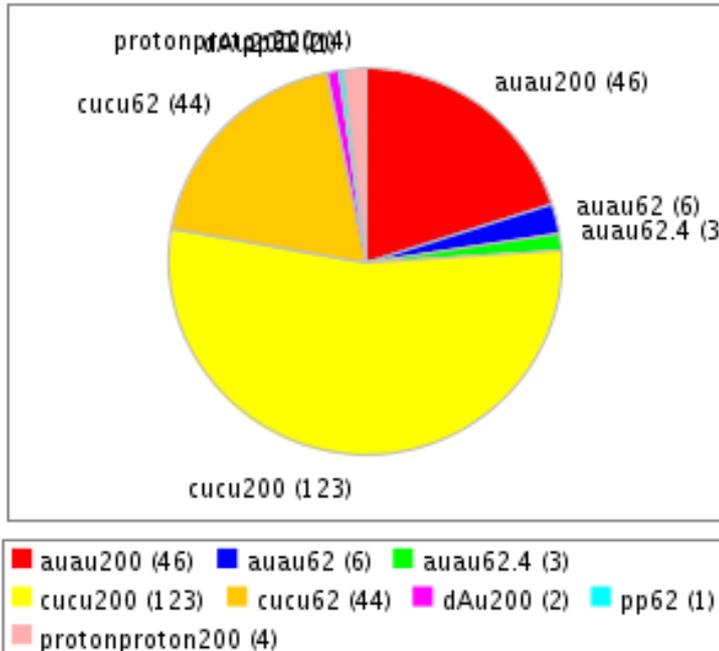
- Hope for better data access providing improved performance
- Growth of dataset size + budget constraints leading to difficult situation
 - **Use data compression (STAR tried, implemented)**
 - **Use even more inexpensive hardware**
 - **Access the data in smarter ways**
- STAR has the largest Xrootd deployment to date (still growing)

□ Issues

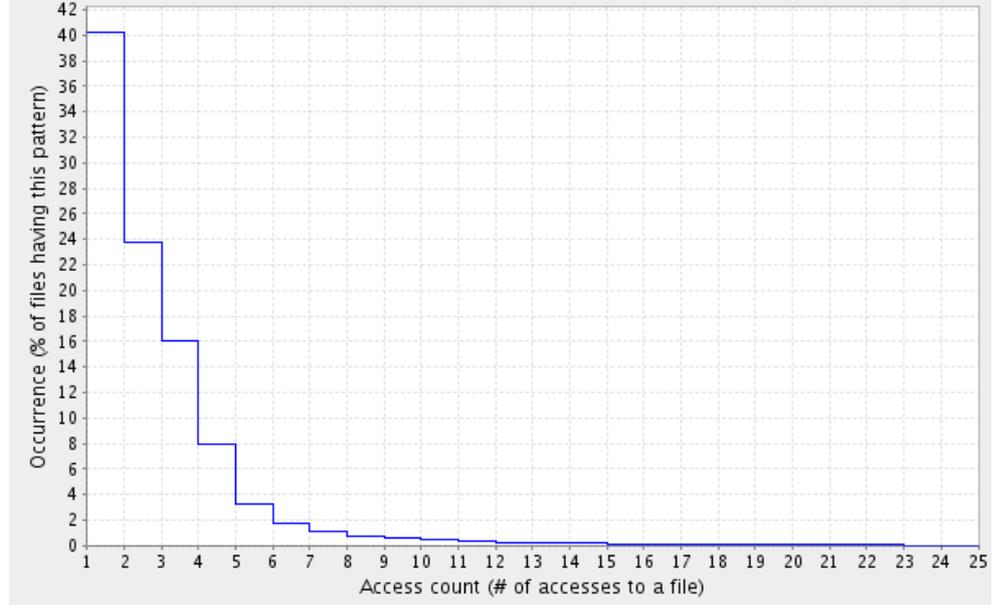
- Xrootd and dCache are still in R&D
 - **High backend MSS stability is required to utilize highly dynamic disk population model**
 - **Optimization non-trivial – STAR spent a fair amount of time to study data retrieval strategies assisted this year by RCF team**
- To make it work for RHIC effort is required from RHIC project
 - **STAR allocates out-sourced FTE to work on Scalla**
 - Not a long-term solution
 - Dedicated effort would be much more efficient
- Analysis relies on leading edge development and integration ⇔ Stability questionable
- Model in RHIC II era questionable w/o (more) integration effort now

Example of Dataset Usage and Access Pattern in STAR

Collision Key

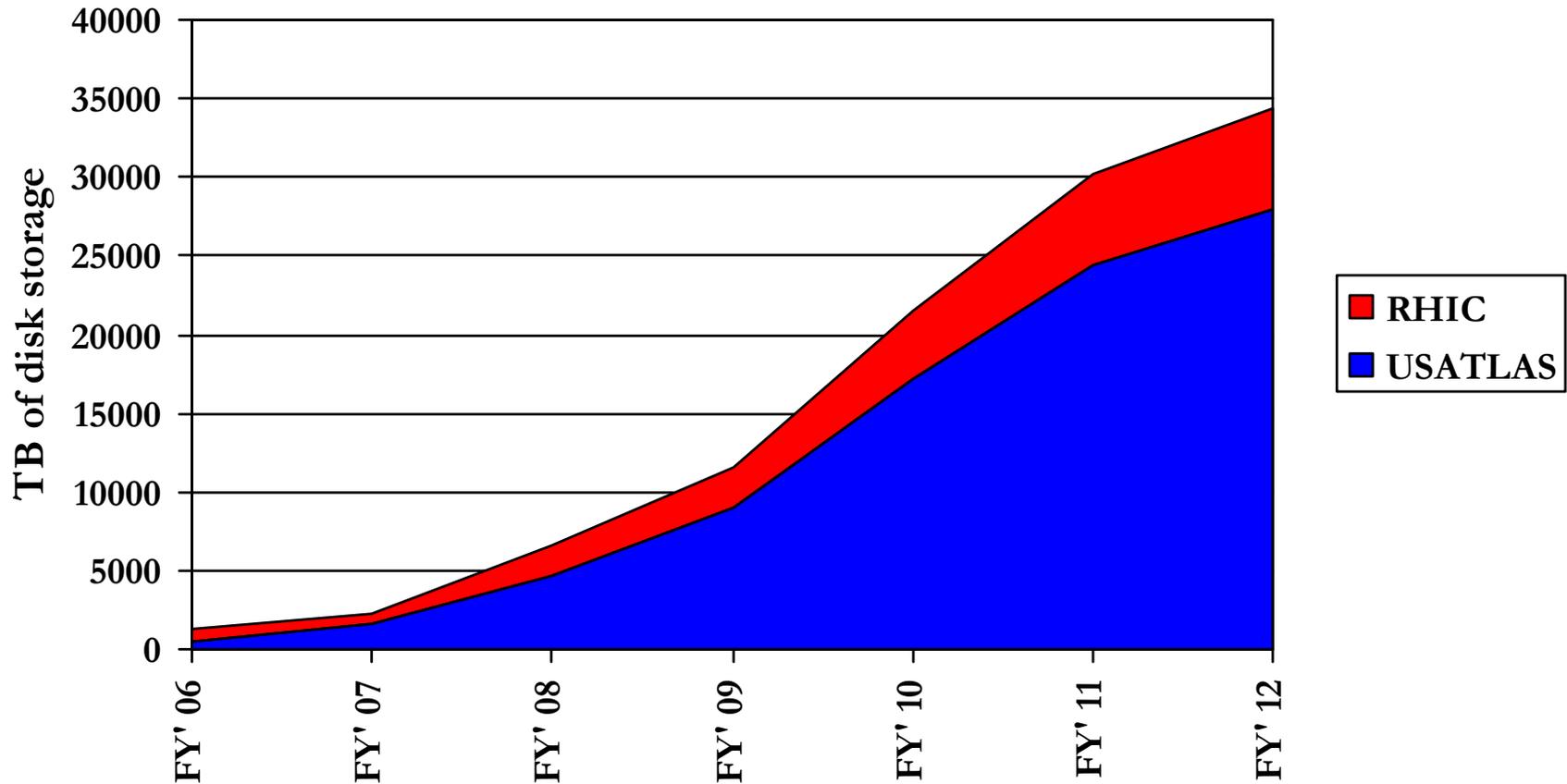


Access pattern of STAR files for month March



Rich variety of Physics Data to be concurrently analyzed leading to “threshing” of disk inventory

Expected Disk Storage Capacity Evolution

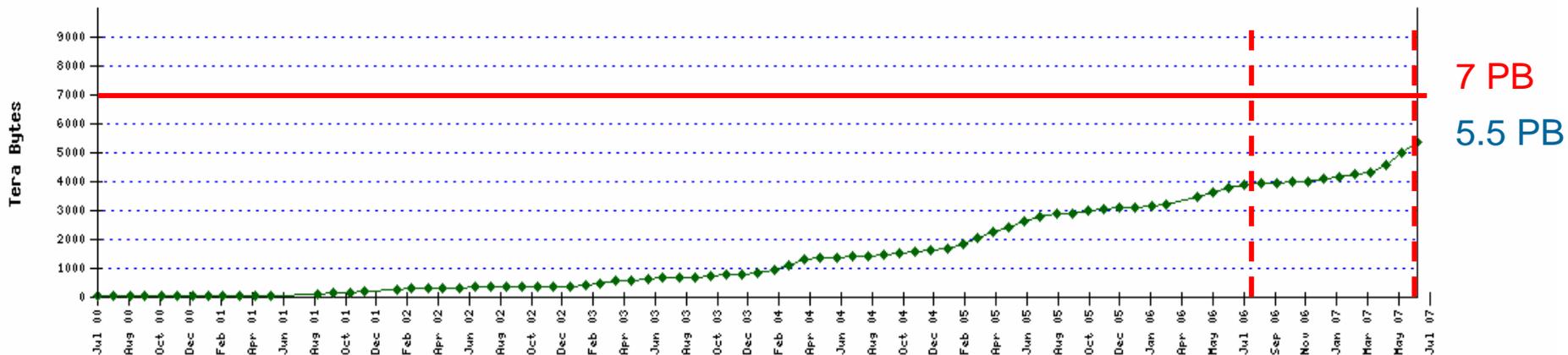


Disk capacity projection for RHIC as described in mid-term plan has foreseen far less space than ATLAS (despite the fact that U.S. ATLAS plans to keep all reconstructed events on Disk)

Mass Storage System

- HPSS (High Performance Storage System) Hierarchical Storage manager from IBM
 - ❑ Moving to version 6.2 in August
- Sun/StorageTek Robotic Tape Libraries
 - ❑ Four PowderHorn Silos
 - ❑ One SL8500 linear library (+1 SL8500 for ATLAS)
 - ❑ 7 PB total capacity

Volume of Data (TB) Stored in HPSS over Six Years
Last modified: Jul 01 2007 09:20:00

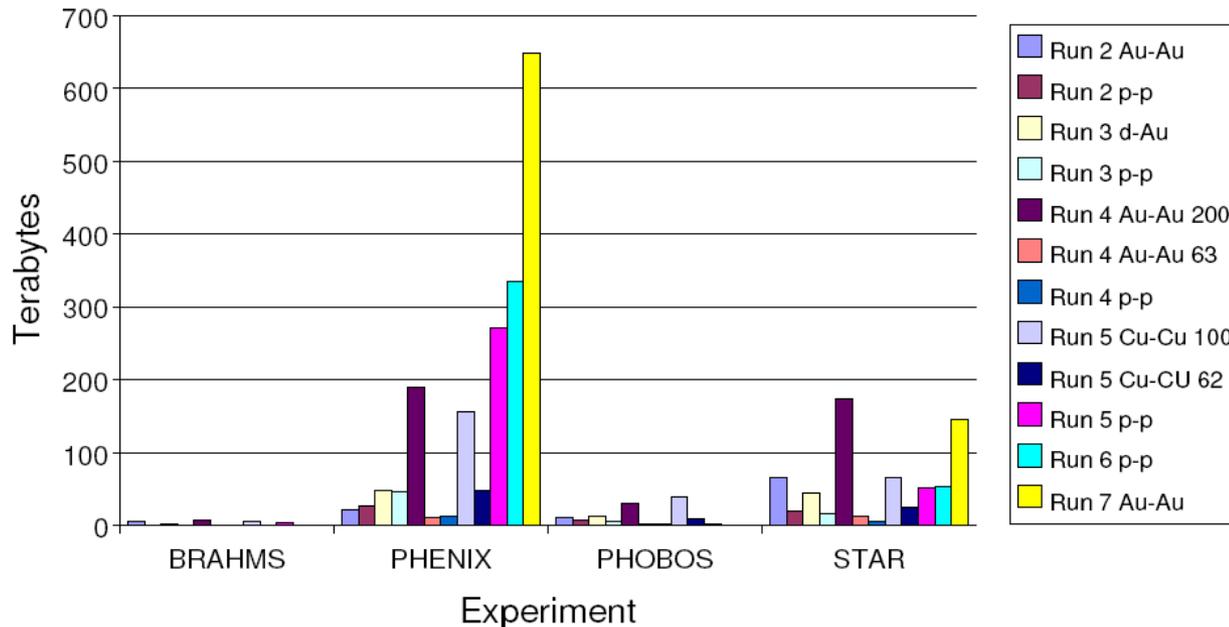


A lot more Raw Data this Year ...

Raw Data (TB)

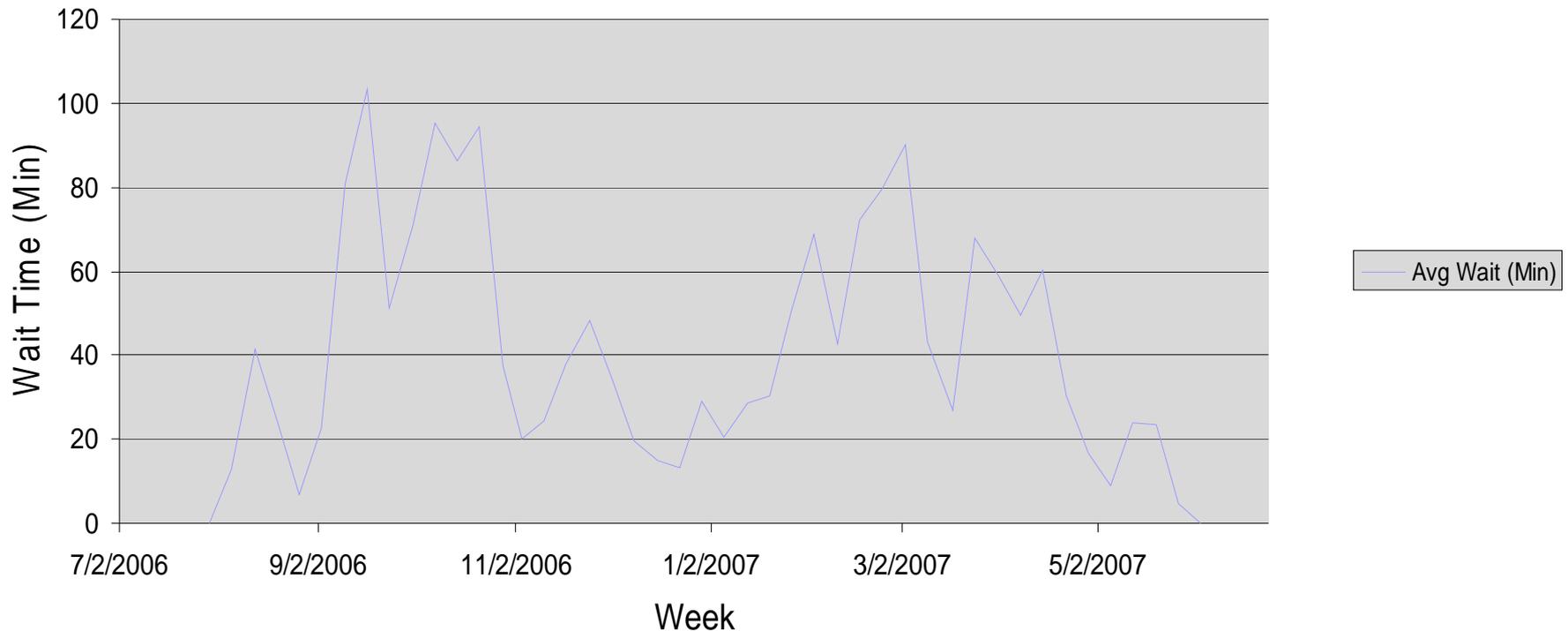
	Run 2 Au-Au	Run 2 p-p	Run 3 d-Au	Run 3 p-p	Run 4 Au-Au 200	Run 4 Au-Au 63	Run 4 p-p	Run 5 Cu-Cu 100	Run 5 Cu-CU 62	Run 5 p-p	Run 6 p-p	Run 7 Au-Au
BRAHMS	6.4	0.85	3.06	0.62	8.45	0.6	0.58	5.51	1.32	4.23	0.93	0.00
PHENIX	22.01	26.75	48.64	46.15	189.87	11.92	13.5	155.87	48.38	272.37	335.61	648.20
PHOBOS	11.26	7.25	13.53	5.7	30.35	2.23	2.61	39.75	10.5	2	0	0.00
STAR	65.91	19.69	45.6	16.92	174.55	13.79	5.94	66.91	26.43	51.60	54.10	145.29

Raw Data Collected in RHIC Runs



Latency (all Experiments)

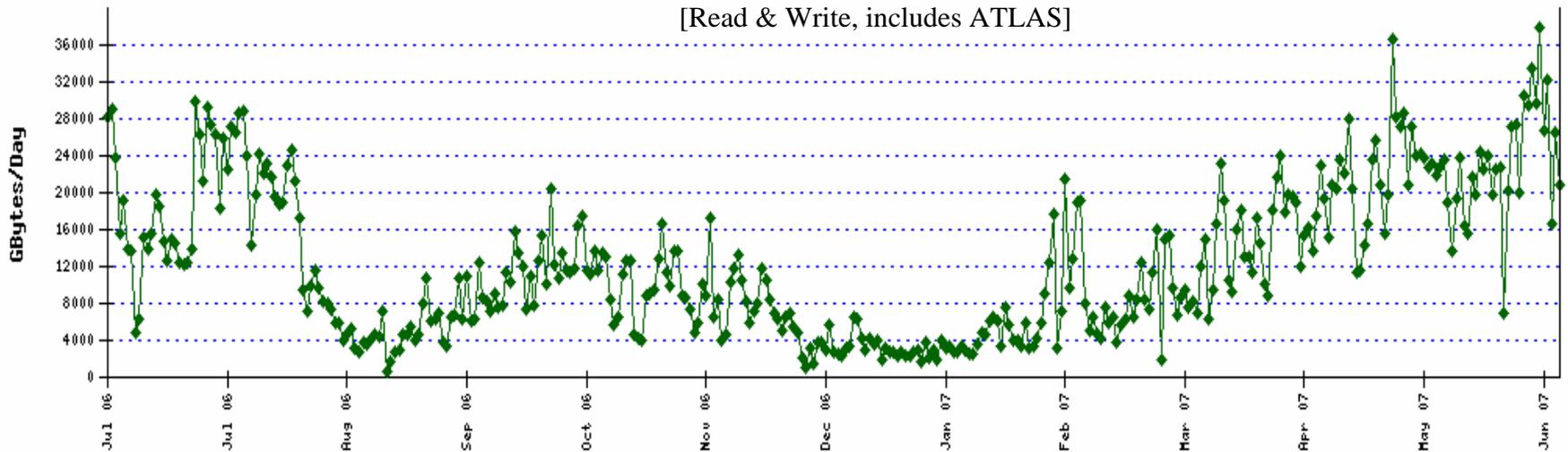
Average Wait Time (minutes)



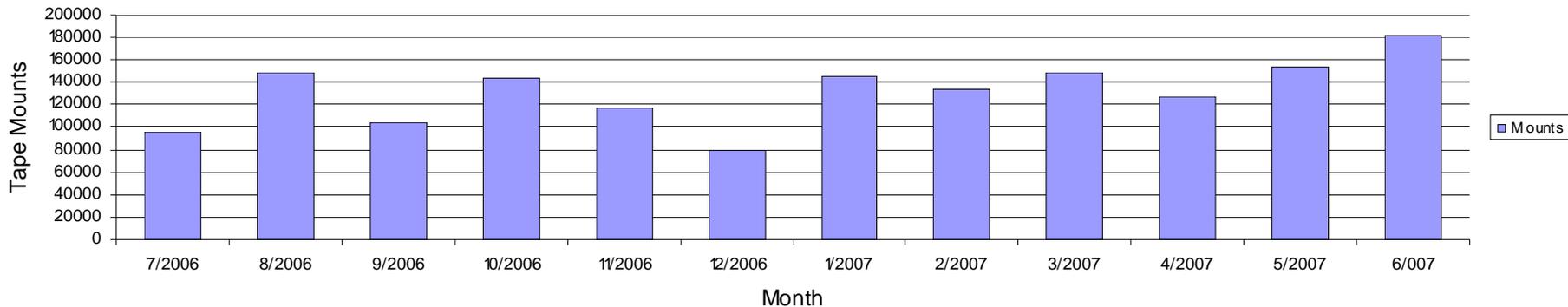
... an important parameter for planning purposes (number of Tape Drives in Robotic Library)

Tape Handling Performance

HPSS Data Handling Rate (GB/day) over Past Year by the Tape Drives
Last Day: June 30, 2007



Tape Mounts Per Month



Grid and Network Services

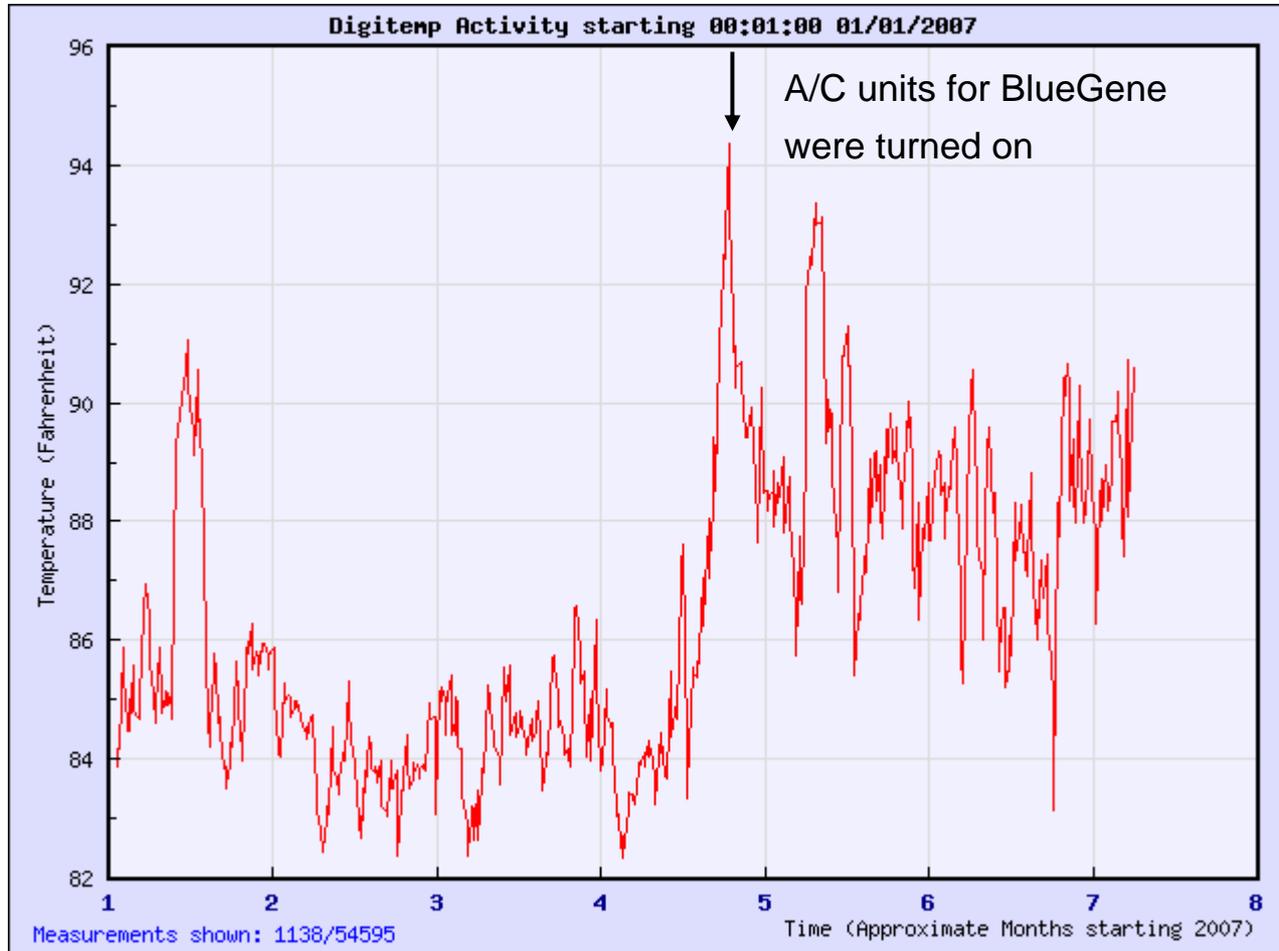
- **Computing models of RHIC Experiments predate the Grid**
 - Unlike ATLAS, they were not originally based on Grid Technology
 - Desire to utilize substantial distributed resources is driving evolution towards Grid Computing
 - Started with simulation, moving towards analysis
 - ◆ LBNL, Prague (working with ITD and ESnet on link), etc. for STAR
 - ◆ Riken, Vanderbilt, IN2P3, etc. for PHENIX
 - Same staff engaged in U.S. ATLAS Grid effort also supports RHIC wide area distributed computing with
 - ◆ Support for Grid tools and services as well as network expertise
 - GridFTP, SRM, ...
 - High volume network transfer optimization
 - ◆ Support for involvement (of STAR) in Open Science Grid (OSG)
 - OSG software deployment and integration of resources into OSG
 - OSG administration

Physical Infrastructure

- Major physical infrastructure improvements were made over the course of the past 12 months
 - ❑ 1.25 MW of local UPS / PDU systems added to support new procurements
 - ❑ New chilled water feed
 - ❑ Local rack top cooling for new procurements
 - ❑ Covered by GPP funds

- Have reached limit of available floor space
 - ❑ Without additional space RCF will not be able to accommodate the next robot (due in early spring 2008)
 - ❑ Reallocation of space to RCF/ACF allows 2007/8 expansion
 - Additional power & cooling is needed each year
 - ❑ Need expansion of space in 2009 and beyond
 - Working with ITD, BNL Plant Engineering and BNL Management on a plan
 - **Very tight schedule**
 - **Progress is not as good as we had hoped for**
 - Technical and organizational problems
 - **This is our top concern at the moment**

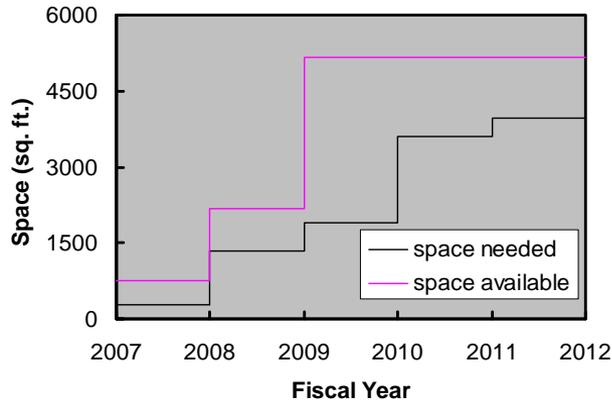
Severe Cooling Problems since April '07



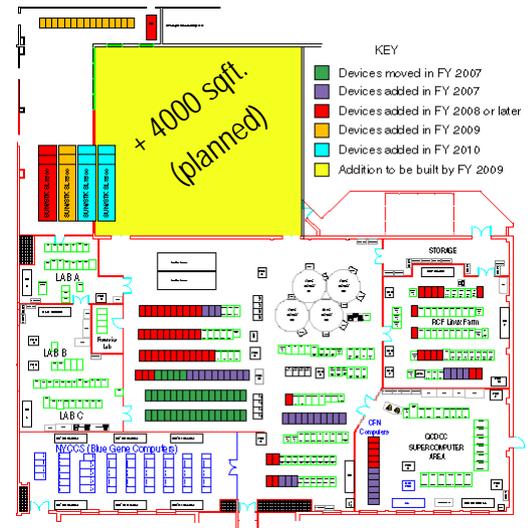
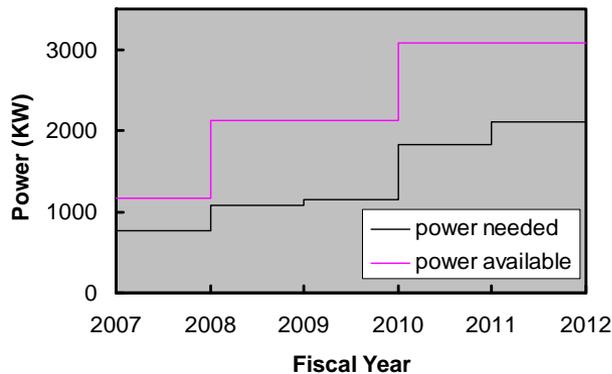
A lot of sediment was stirred up due to increased cooling flow, system never recovered so far

Infrastructure Planning at RACF

➤ Currently available space filled



➤ Soon running out of Power



Cyber Security

- Facility is a Firewall protected enclave within the BNL firewall protected site
- Most Services provided by the Facility have a single sign-on Kerberos based authentication infrastructure
- Major efforts
 - ❑ Contributing to BNL Cyber Security Program Plan
 - ❑ Deploying – Facility-wide – Ordo (BNL developed host based configuration tracking/auditing tool for Unix-like system)
- Concern remains of conflicts between User (Grid) requirements, regulatory requirements, and a cyber security policy/architecture which does not disrupt effective facility use

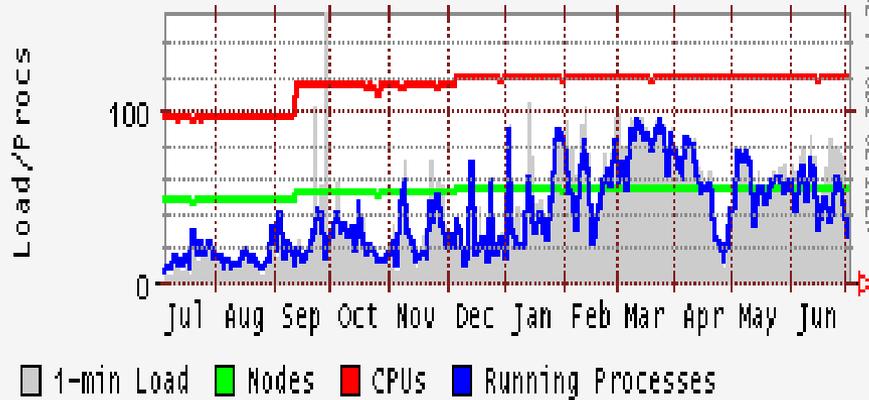
Conclusions

- Plans to evolve and expand facility services to meet expected needs
 - ❑ Are based on successful adjustments of technical directions
 - ❑ Requires agreed and planned for increases in 2008 and beyond
- Continued Slippage of Funding for Infrastructure / Facility creating difficult situation for the Experiments and the RHIC Computing Facility
 - ❑ 1/3 replacement per year impossible
 - ❑ Stretching equipment lifetime with bulk replacement potentially disruptive
 - ❑ Core Infrastructure (HPSS, High-end Disk) improvements need to be delayed
 - ❑ Further increasing burden on staff
- Physical infrastructure expansions and improvements are the top concern
 - ❑ Facility needs new space with appropriate characteristics and services for 2008 and beyond
- Grid technology is likely to change future RHIC computing
 - ❑ Are building on ATLAS experience
- Cyber Security is a major concern
 - ❑ Security versus Usability

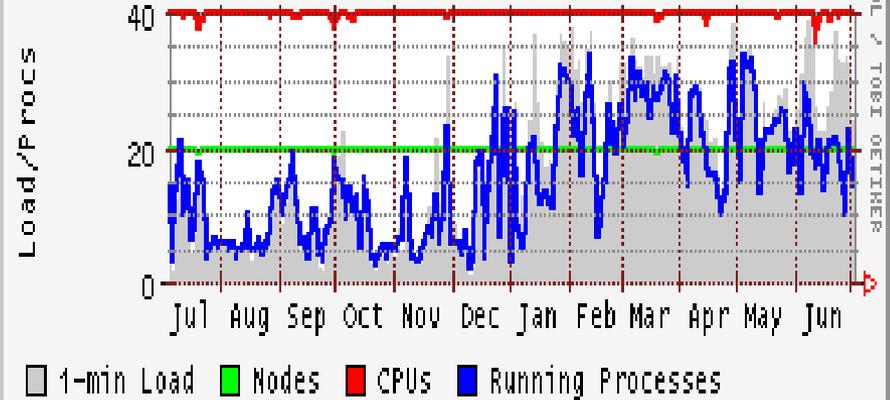
Backup Slides

Computing Resource Utilization BRAHMS & PHOBOS

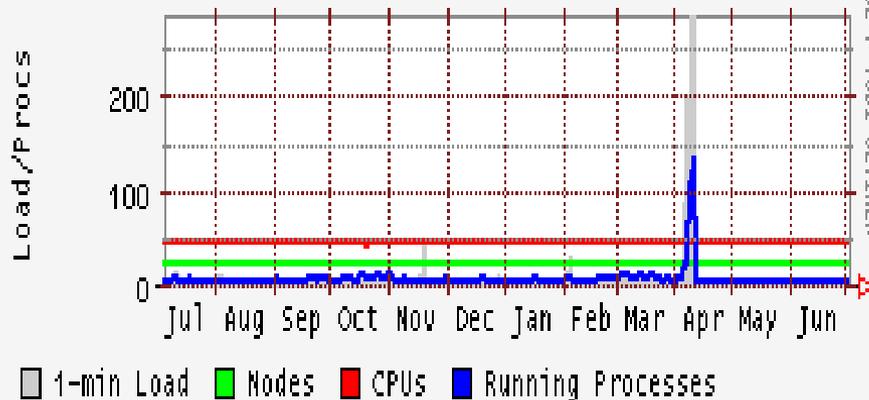
BRAHMS CAS Linux Cluster Load last year



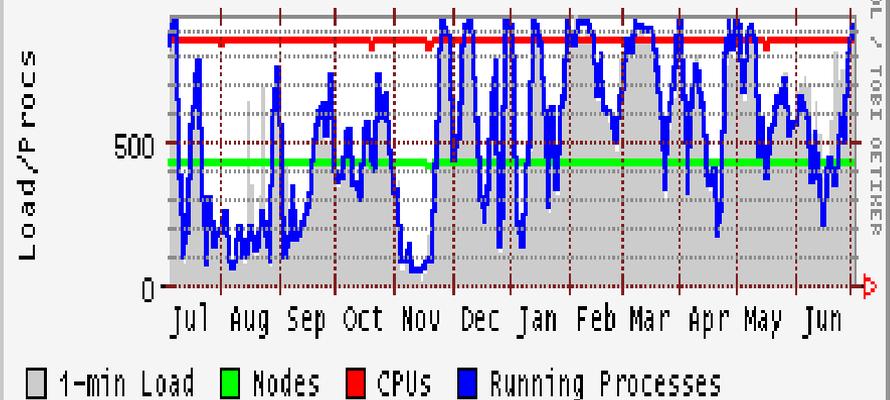
BRAHMS CRS Linux Cluster Load last year



PHOBOS CAS Linux Cluster Load last year

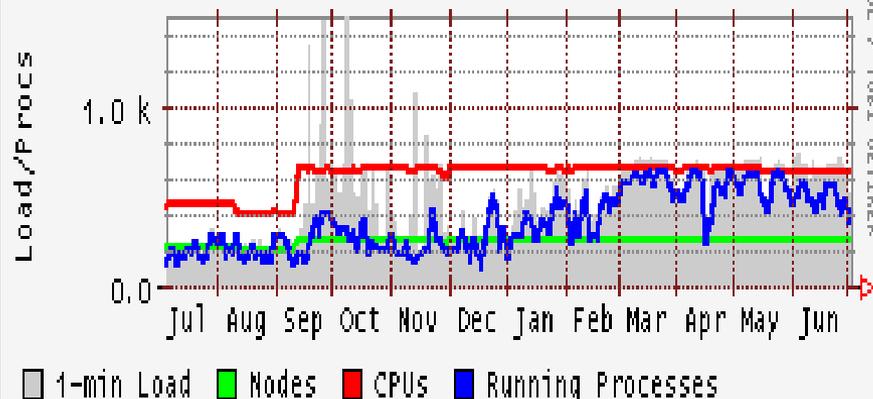


PHOBOS CRS Linux Cluster Load last year

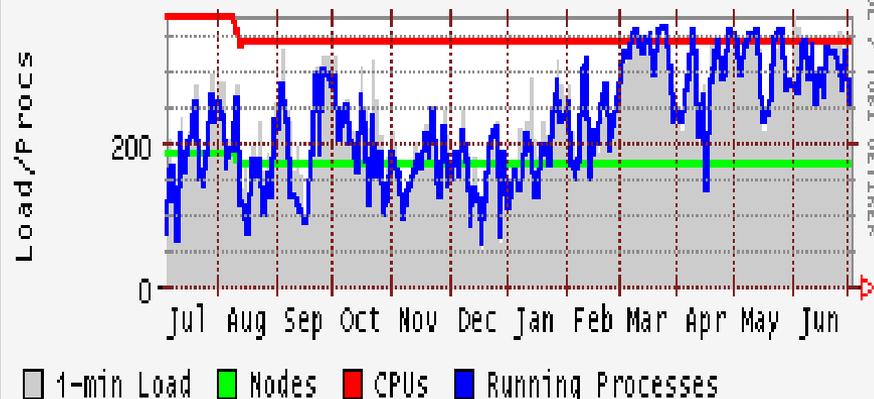


Resource Utilization PHENIX & STAR

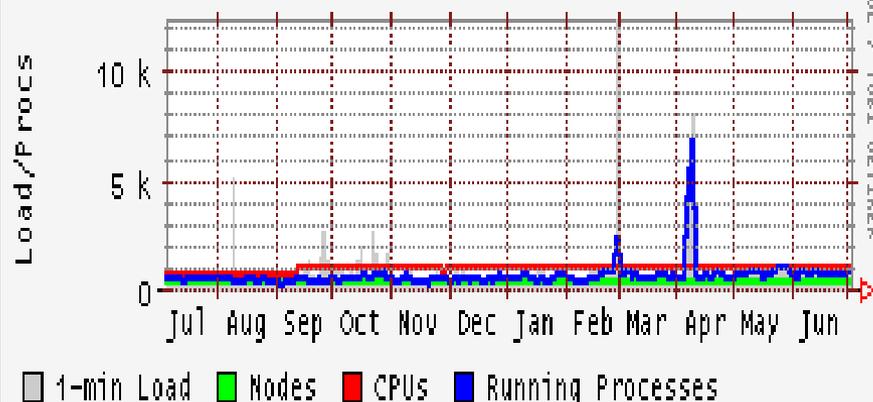
PHENIX CAS Linux Cluster Load last year



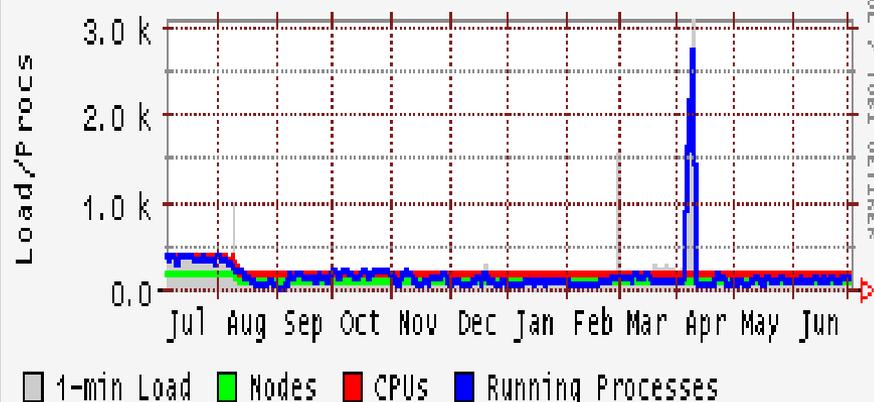
PHENIX CRS Linux Cluster Load last year



STAR CAS Linux Cluster Load last year



STAR CRS Linux Cluster Load last year



Condor Usage (2006)

Statistics from 07-06-2005 to 07-06-2006

no. jobs completed								no. jobs evicted before completion									
destination								destination									
	phenix	phobos	star	brahms	atlas	rcf	total		phenix	phobos	star	brahms	atlas	rcf	total		
	phenix	1683503	4252	955	3512	162	<u>1463</u>	1693847		phenix	<u>95880</u>	<u>301</u>	<u>184</u>	<u>468</u>	<u>101</u>	<u>57</u>	96991
source	phobos		443309					443309	source	phobos		<u>8345</u>					8345
	star	2137		24206	247	172	<u>164</u>	26926		star	<u>151</u>		<u>1058</u>	<u>7</u>	<u>2</u>	<u>18</u>	1236
	brahms	58	1183	2	138665			139908		brahms	<u>8</u>	<u>46</u>		<u>2833</u>			2887
	atlas	8500	16962	1427	11675	1659843	<u>3677</u>	1702084		atlas	<u>914</u>	<u>2163</u>	<u>240</u>	<u>921</u>	<u>31236</u>	<u>44</u>	35518
total effective runtime hours consumed by completed jobs								total ineffective runtime hours consumed (including jobs removed)									
destination								destination									
	phenix	phobos	star	brahms	atlas	rcf	total		phenix	phobos	star	brahms	atlas	rcf	total		
	phenix	1608830.32	27144.14	2984.96	9552.61	151.93	379.18	1649043.14		phenix	259578.7	482.99	787.13	3387.26	102.08	4.64	264342.8
source	phobos		2592983.57					2592983.57	source	phobos		126067.99					126067.99
	star	531.59		32417.35	4.83	1393.9	39.46	34387.13		star	9.51		419.96	0.2	0.04	1.78	431.49
	brahms	4.7	62.38	0.08	156659.66			156726.82		brahms	18.43	21.72		10296.49			10336.64
	atlas	14275.07	61648.79	1231.51	20724.46	4248720.05	1179.82	4347779.7		atlas	5468.55	21211.81	1118.55	10698.11	315962.75	560.2	355019.97

- Creation of general queue allows opportunistic usage of idle CPU's by user jobs not normally affiliated with CPU ownership
- General queue became default queue in late 2006. Users can override by specifying other queues
- General queue jobs were only 1.4% of all Condor jobs during this period

Condor Usage (2007)

Statistics from 07-05-2006 to 07-05-2007

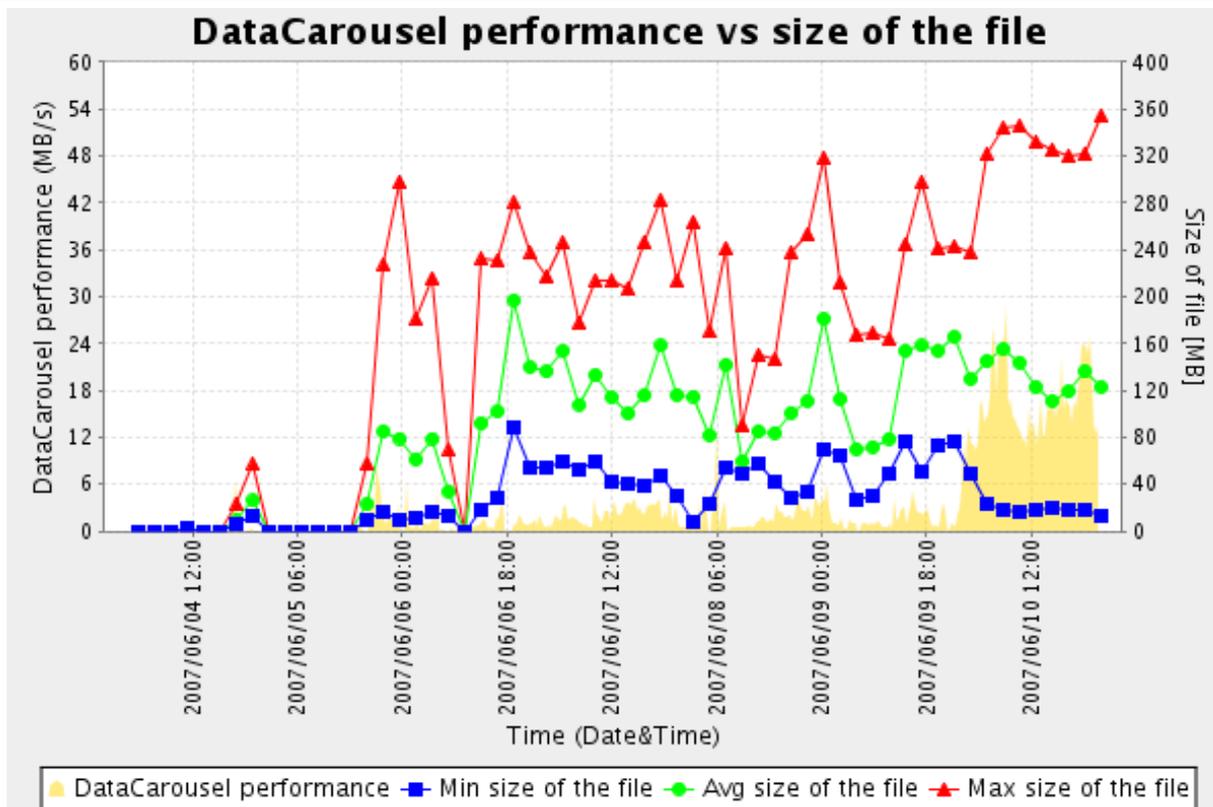
no. jobs completed								no. jobs evicted before completion								
destination								destination								
	phenix	phobos	star	brahms	atlas	rcf	total		phenix	phobos	star	brahms	atlas	rcf	total	
	phenix	2928260	803131	228589	589175	66466	<u>341200</u>	4956821	phenix	<u>228802</u>	<u>121837</u>	<u>70029</u>	<u>112348</u>	<u>5220</u>	<u>55616</u>	593852
source	phobos		834760						source	phobos		14143				14143
	star	7635	22994	419820	3143	39501	<u>2146</u>	495239	star	<u>2790</u>	<u>7546</u>	<u>39745</u>	<u>890</u>	<u>10375</u>	<u>979</u>	62325
	brahms	2837	4594	84	166832	737	<u>4687</u>	179771	brahms	<u>248</u>	<u>96</u>	<u>56</u>	<u>4142</u>	<u>9</u>	<u>396</u>	4947
	atlas	554	1021	4286	440	3506339	<u>7490</u>	3520130	atlas	<u>34</u>	<u>77</u>	<u>215</u>	<u>47</u>	<u>84495</u>	<u>152</u>	85020
total effective runtime hours consumed by completed jobs								total ineffective runtime hours consumed (including jobs removed)								
destination								destination								
	phenix	phobos	star	brahms	atlas	rcf	total		phenix	phobos	star	brahms	atlas	rcf	total	
	phenix	5072760.63	940249.06	120548.52	430851.94	12067.96	143432.31	6719910.42	phenix	365272.97	193778.52	53635.03	218839.65	11291.42	52392.1	895209.69
source	phobos		4251158.08					4251158.08	source	phobos		129556.63				129556.63
	star	12911.33	127644.94	1241432.97	10365.4	44923.35	12916.17	1450194.16	star	17672.03	43589.87	594317.94	11855.5	59759.56	19235.84	746430.74
	brahms	268.49	12820.29	104.8	52067.14	42.05	2799.49	68102.26	brahms	66.41	22.82	47.4	3784.24	16.52	155.2	4092.59
	atlas	58.14	744.29	266.34	1419.41	7155737.29	5171.66	7163397.13	atlas	55.09	229.76	767.46	72.75	752081.7	423.55	753630.31

- Condor usage grew by a factor of 3 (in terms of number of jobs) and by a factor of 4 (in terms of CPU time) over the past year.
- PHENIX executed over 40% of their jobs in the general queue.
- General queue efficiency is ~ 87% (i.e., only 13% ineffective use).
- General queue jobs amounted to 21% of all Condor jobs during this period.

PHENIX dCache Deployment (v1.7)

- 415 Read/Write Pools (shared), 36 external Write Pools or dedicated hosts
- 212 TB Storage, >750k files on disk
 - Adding 140 TB (usable) by end of July
- 3 GridFTP/SRM + 1 dCap door nodes, 1 admin node
- SL3+EXT3 on Read Pools, SL3/XFS+SL4/EXT3 on external Write Pools
- HPSS backend interface via HIS/Carousel/PFTP

STAR Mass Storage System File retrieval Performance



HPSS backend (DataCarousel) performance monitored versus file size in Xrootd / Scalla context

Mass Storage System – High Performance Storage System (HPSS)

➤ Tape Drives

- ❑ 37 StorageTek 9940B (30 MB/s)
- ❑ 30 LTO Gen3 (80 MB/s)

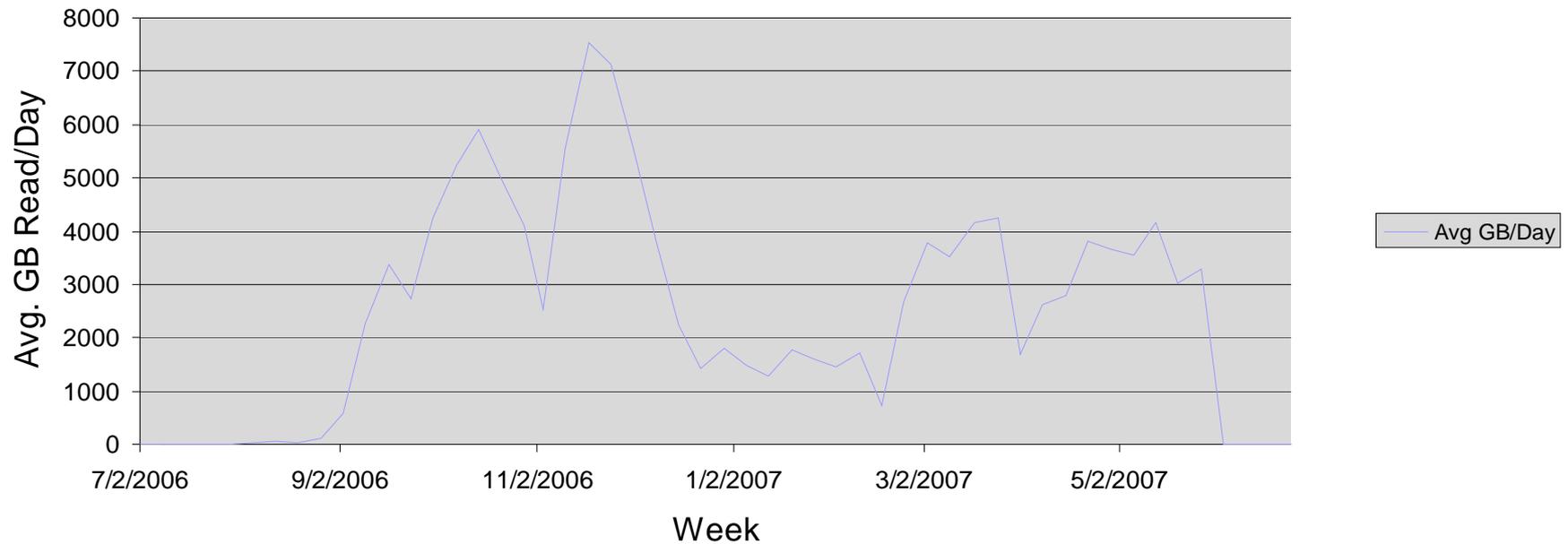
➤ 30 TB of HPSS Disk Cache

➤ In-house developed tape access optimization software

- ❑ Increases access efficiency by sorting requests according to data placement

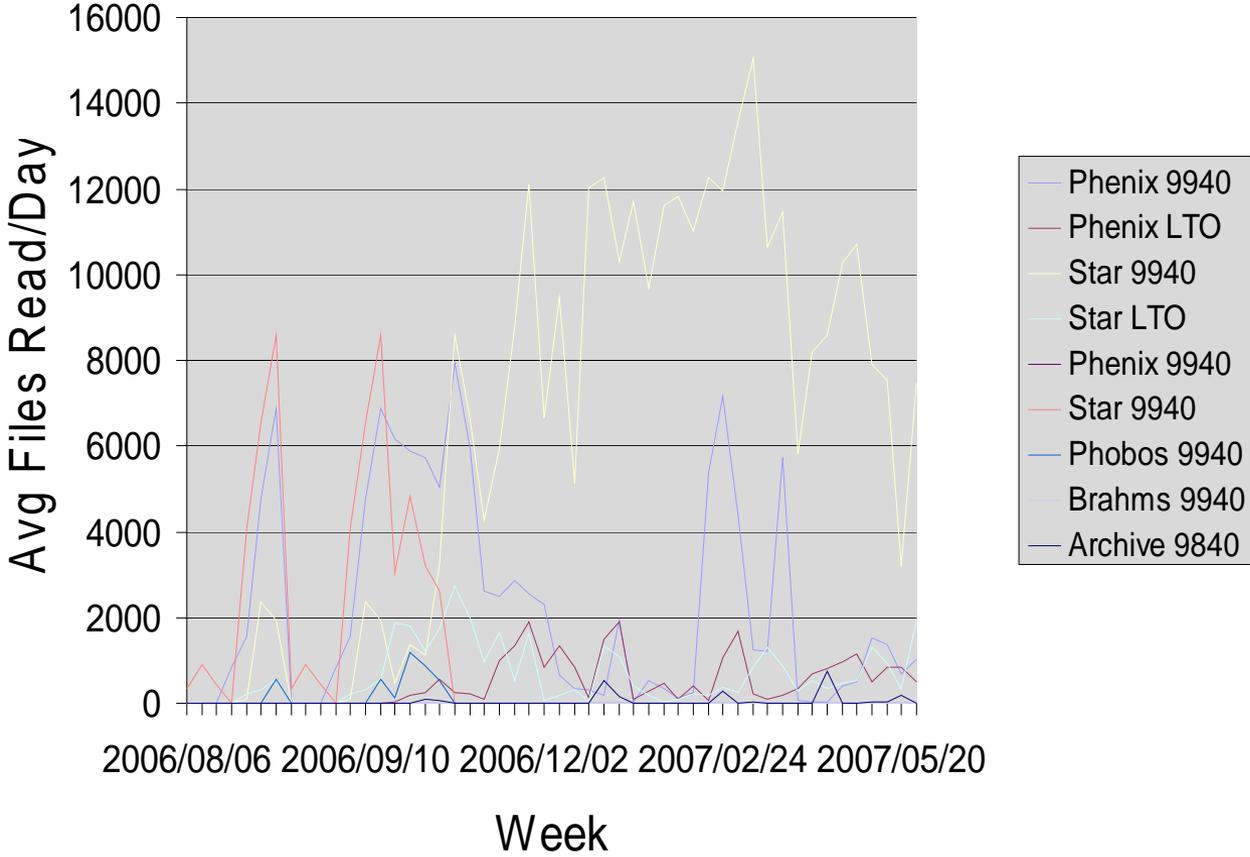
HPSS Read Performance (GB/day)

Average GB Read Per Day



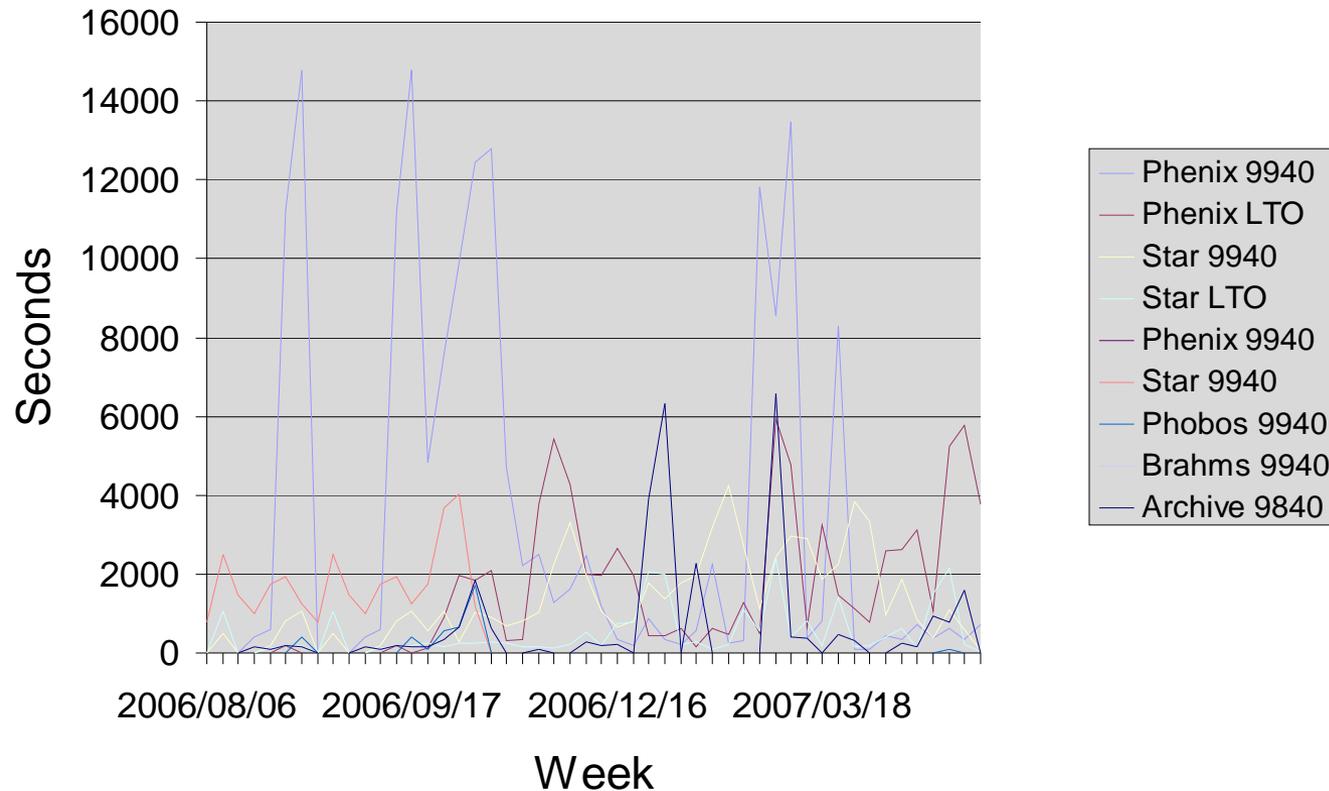
HPSS Read performance (# of Files)

Average Files Read Per Day



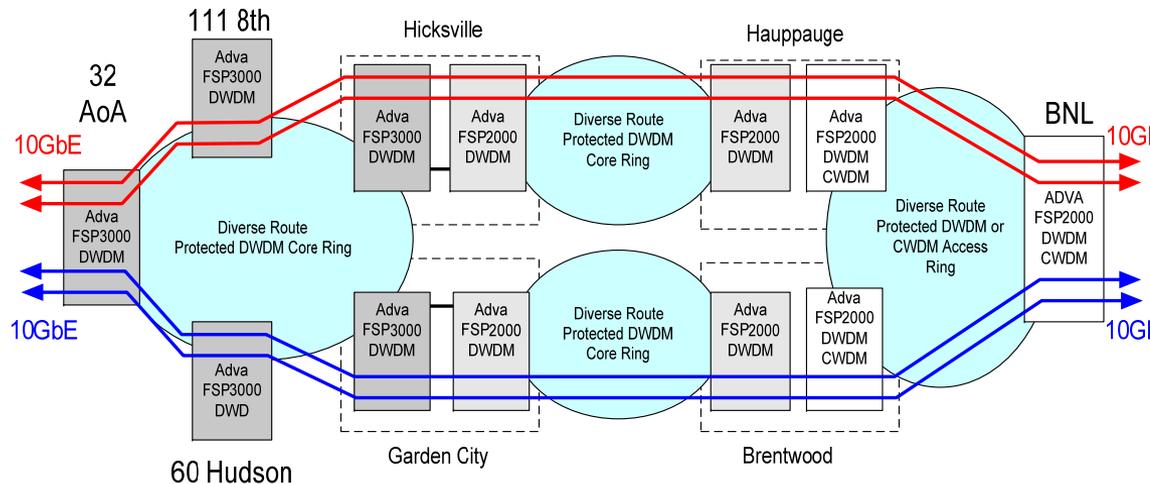
HPSS File Retrieval Latency (per Experiment)

Average File Retrieval Latency



Wide Area Network

- Jan '06 WAN last upgrade on BNL connectivity to 20 Gbps
- Funded in equal part by ESnet, DOE NP, DOE HEP and BNL
- Connection still lacks desired redundancy and diversity
 - ❑ Will require significant additional funding not yet identified



Wide Area Network Architecture

BNL 20 Gig-E Architecture

