

# Fall 2005 - Spring 2006 HEPiX Report

---

Technology Meeting  
May 22, 2006

Robert Petkus

RHIC/USATLAS Computing Facilities  
Brookhaven National Laboratory

# Overview

---

This talk will summarize site and technology reports from HEPiX Fall 2005 and Spring 2006 that pertain to current areas of interest at the RCF/ACF including:

- **Power and cooling**
- **Centralized and distributed storage**
- **Recent procurements**
- **Farm management**
- **Security**

# BNL Site Report

## BNL Fall 2005

- Purchase of 867 dual-Xeon systems
- Local farm storage exceeds 670 TB
- Transition from NIS -> LDAP
- Centralized logging
- Grid development: GUMS, SUMS, OSG
- dCache testing and deployment for PHENIX and USATLAS

## BNL Spring 2006

- HPSS: (2x) SL8500 with LTO3 drives gearing for production, 30 dual-core Opteron IBM movers, DDN S2A FC disk arrays
- Central storage: Retirement of 100TB NFS storage, Panasas problems
- Condor Flocking
- Nagios Monitoring
- Preparation for SC4
- Power – cooling concerns a major factor in purchasing
  - Dual-core CPU tests
  - Installation of Liebert rack cooling modules

# CASPUR Site Report

---

CASPUR Spring 2006

- **Centralized / Distributed Storage:**
  - Migration from IBM SANFS to GPFS
  - RAID 6 Infortrend disk systems for AFS
  - AFS - OSD
  - 24 TB of Polyserve
  - Investigating Terragrid, Lustre, GFS, PVFS2
- **Cooling concerns**

# CERN Site Report

---

## CERN Fall 2005

- Refurbishment of data center – limited power
- System monitoring with Lemon
- 6 LTO-3 drives in STK silos
- Use of SMART to ferret out iffy disks
- Software license management with LicMon
- PXE-based install made default

## CERN Spring 2006

- SC3 – Castor2 achieved >750MB/s sustained over a week
- Migrate all experiments from Castor1 -> Castor2
- Added 1200 dual cpu farm nodes
- Data center refurbishment near completion
- Mar 26/27: temperatures exceed 100° in data center! Systems shut down.
- Plan for massive SLC4 upgrade by 10/2006
- LAN upgrade – Force10 Routers

# SLAC Site Report

---

## SLAC Fall 2005

- Added 360 dual-core Opteron servers
- Retiring Netra T1 and VA Linux clusters
- HPSS migration from 4.5 -> 5.1

## SLAC Spring 2006

- Completed replacement of 30 year old AC unit
  - Needed to run data center with all fans running and doors open
- Added 350 dual-core Opteron servers
- Added 230 TB storage on Sun 3511 disk arrays – mostly for Xrootd
- Testing RT for project management – linked child tickets for each subtask

# RT at SLAC Spring 2006

---

- SLAC needed a powerful, searchable ticketing system with e-mail, web, and cli interfaces
- Current ticketing system, Remedy, was inadequate
- Can be extended using Asset Manager plugin
- Pros: userbase generally pleased with RT. 50k tickets currently in system
- Cons: Poor interaction with Remedy ticket system. Multiple user ids. Can't unmerge merged tickets.
- Verdict: good enough for continued deployment. Possible addition of enhanced RT CLI created at DESY.

# RT at SLAC Spring 2006

RT at a glance - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://www-rt.slac.stanford.edu/rt3/

BBHome SysCpDocs Google eBooks RT me Hot Seat Monitoring

SLAC SCCS Preferences | Logout  
Logged in as bartlett

RT for SLAC New ticket in admin-mtg Search

Home RT at a glance

Tickets

RTFM

Tools

Configuration

Preferences

Approval

Statistics

Using RT

X 10 highest priority tickets I own...					
#	Subject	Priority	Queue	Status	
34387	www7	80	unix-admin	open	
40108	Perl update plans?	80	unix-admin	open	
42343	FW: [RHSA-2006-0199-01] Critical: mozilla security update	67	unix-admin	open	
43544	lynx problem	40	unix-admin	open	
44026	FW: [RHSA-2006-0266-01] Important: gnupg security update	26	unix-admin	open	
33021	RT bug: error in Ticket_Overlay.pm when adding URL reference with a leading space	20	RT	open	
23183	apache/cgi stuff	0	Projects	open	

X My 10 top tickets...					
#	Subject	Queue	Status	Created	
42550	ApacheServerTokenNotSet	Scans-Security	open	6 weeks ago	
41272	python 2.4.2	unix-admin	open	2 months ago	
35990	need a machine or two for The Written Word	unix-admin	open	4 months ago	
34387	www7	unix-admin	open	5 months ago	
32630	Czar list again unavailable	www-admin	new	6 months ago	

X Quick search				
Queue	New	Open	Stall	
admin-mtg	9	5	2	
ChangeLog	0	0	0	
Incident	0	2	1	
Installation	2	10	0	
Internal	0	0	0	
mail-admin	1	7	0	
Monitoring	26	9	0	
PrivilegeRequests	1	5	1	
Projects	2	0	2	
raid	0	0	0	
RT	9	10	3	
Scans-Security	2	45	0	
Security	9	21	4	
SpaceRequests	1	3	0	
Tech-Coord	2	2	0	
test2	3	1	0	
unix-admin	26	173	52	
www-admin	63	38	1	

Don't refresh this page. Col

https://www-rt.slac.stanford.edu/rt3/ www-rt.slac.stanford.edu

# RT at SLAC Spring 2006

## Asset Manager

The screenshot shows a web browser window with the URL <https://webcs02.slac.stanford.edu/n3/AssetTracker/Index.html>. The page title is "RT test server" and it is logged in as "bartelt". The main content area is titled "Asset Tracker" and includes a "Quick search" table and a "10 Most recently updated assets..." table.

**Quick search**

Asset Type	production	spare	all
Desktop	1240	2	1242
Servers	2418	0	2418
Storage	337	0	337
all	3995	2	3997

**10 Most recently updated assets...**

Name	Description	Type	Status
mbr-xfer01	(No description)	Servers	production
lan002	(No description)	Servers	retired
come098	(No description)	Servers	production
noma0357	(No description)	Servers	production
noma0090	(No description)	Servers	production
lari0123	(No description)	Servers	production
opi08rfa00	(No description)	Desktop	production
noma0301	(No description)	Servers	production
sysdev13	(No description)	Desktop	production
mccelog-mgt	(No description)	Desktop	production

John Bartelt - SLAC - RI

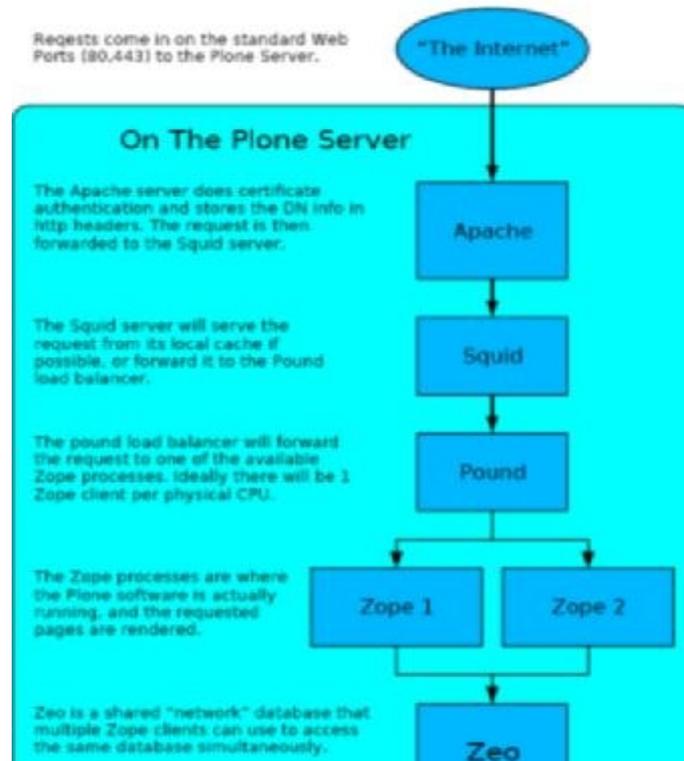
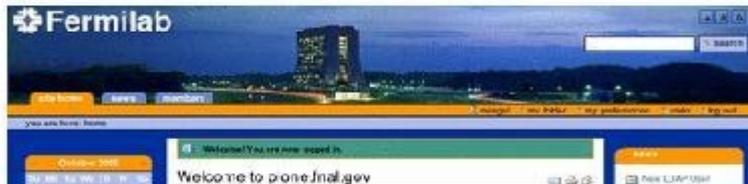
# PLONE at Fermi Fall 2005

- Customizable, full-featured CMS out-of-the-box
- Platform for all interactive content on site
- Tools/features include: wiki, form management, trouble ticket systems, editor support (WebDAV), and searching (RSS)
- Custom additions @Fermi: SSL auth, logbook

• Before:



• After:



# JLAB Site Report

---

## JLAB Fall 2005

- Added 25TB of Panasas
- Cooling concerns – problems with generator/UPS systems
  - Using SiteView to monitor AC & UPS status in real time

## JLAB Spring 2006

- Additional 8TB Panasas shelf
- LAN upgrade to 10GbE using Foundry BigIron Rx-8
- WAN upgrade to OC-192
- New data center in operation
  - UPS backup only – no generator
  - Need IPMI for shutdown / startup
- Investigating VOMS, GUMS, Wiki, and Subversion

# SiteView at JLAB

Control Panel:

**Air Conditioner: cchvac07**

**System**  
 Manufacturer:  Location:   
 Description:  Contact:

**Temperature (degrees Farenheit)**  
 Set Point:  Reading:  High Threshold:  Low Threshold:

**Temperature (degrees Centigrade)**  
 Set Point:  Reading:  High Threshold:  Low Threshold:

**Humidity (percent Relative)**  
 Set Point:  Reading:  High Threshold:  Low Threshold:

**State**  
 Powered:  YES Cooling:  YES Heating:  NO Humidifying:  NO Dehumidifying:  NO

Active Alarms:

• No Active Alarms

Current Status:

**Celsius Reading**  
 Value: 20 Degrees Celsius  
 12.4 deg C | 18 deg C | 37.5 deg C | 32 deg C

**Humidity Reading**  
 Value: 47 Percent Relative Humidity  
 30 pct | 35 pct | 70 pct | 65 pct

**Temperature Reading**  
 Value: 69 Degrees Fahrenheit  
 35 deg F | 65 deg F | 100 deg F | 90 deg F

Last updated: 2005-10-03 17:02:37

Control Panel:

**Three Phase UPS: qcdups01**

**System**  
 Manufacturer:   
 Model:   
 Output Source:   
 Percent Load:  Percent  
 Heat Dissipation:  Tons

**Input**

Line 1	Line 2	Line 3
Frequency: <input type="text" value="60"/> Hertz	<input type="text" value="60"/> Hertz	<input type="text" value="60"/> Hertz
Voltage: <input type="text" value="470"/> Volts	<input type="text" value="470"/> Volts	<input type="text" value="470"/> Volts
Current: <input type="text" value="38.5"/> Amps	<input type="text" value="38.3"/> Amps	<input type="text" value="38.3"/> Amps

**ByPass**

Line 1	Line 2	Line 3
Voltage: <input type="text" value="470"/> Volts	<input type="text" value="472"/> Volts	<input type="text" value="467"/> Volts

**Output**

Line 1	Line 2	Line 3
Frequency: <input type="text" value="60"/> Hertz	<input type="text" value="60"/> Hertz	<input type="text" value="60"/> Hertz
Voltage: <input type="text" value="208"/> Volts	<input type="text" value="208"/> Volts	<input type="text" value="208"/> Volts
Current: <input type="text" value="32.4"/> Amps	<input type="text" value="33.8"/> Amps	<input type="text" value="29.7"/> Amps
Power: <input type="text" value="3000"/> Watts	<input type="text" value="3000"/> Watts	<input type="text" value="3000"/> Watts

**Battery**  
 Status:   
 Voltage:  Volts  
 Current:  Amps  
 Temperature:  Deg. C  
 Time Used:  Seconds  
 Time Remaining:  Minutes  
 Charge Remaining:  Percent

Active Alarms:

• No Active Alarms

Current Status:

**Battery Temperature**  
 Value: 30 Degrees Centigrade  
 0 deg C | 40 deg C | 30 deg C

**Bypass 1 Voltage**  
 Value: 470 RMS Volts  
 432 volts | 456 volts | 528 volts | 504 volts

**Bypass 2 Voltage**  
 Value: 472 RMS Volts  
 432 volts | 456 volts | 528 volts | 504 volts

**Bypass 3 Voltage**  
 Value: 467 RMS Volts  
 432 volts | 456 volts | 528 volts | 504 volts

**Charge Remaining**  
 Value: 100 Percent  
 0 pct | 99 pct | 100 pct

**Input 1 Frequency**  
 Value: 60 Hertz  
 59.5 Hz | 61 Hz | 60.5 Hz

**Input 1 Voltage**  
 Value: 470 RMS Volts  
 432 volts | 456 volts | 528 volts | 504 volts

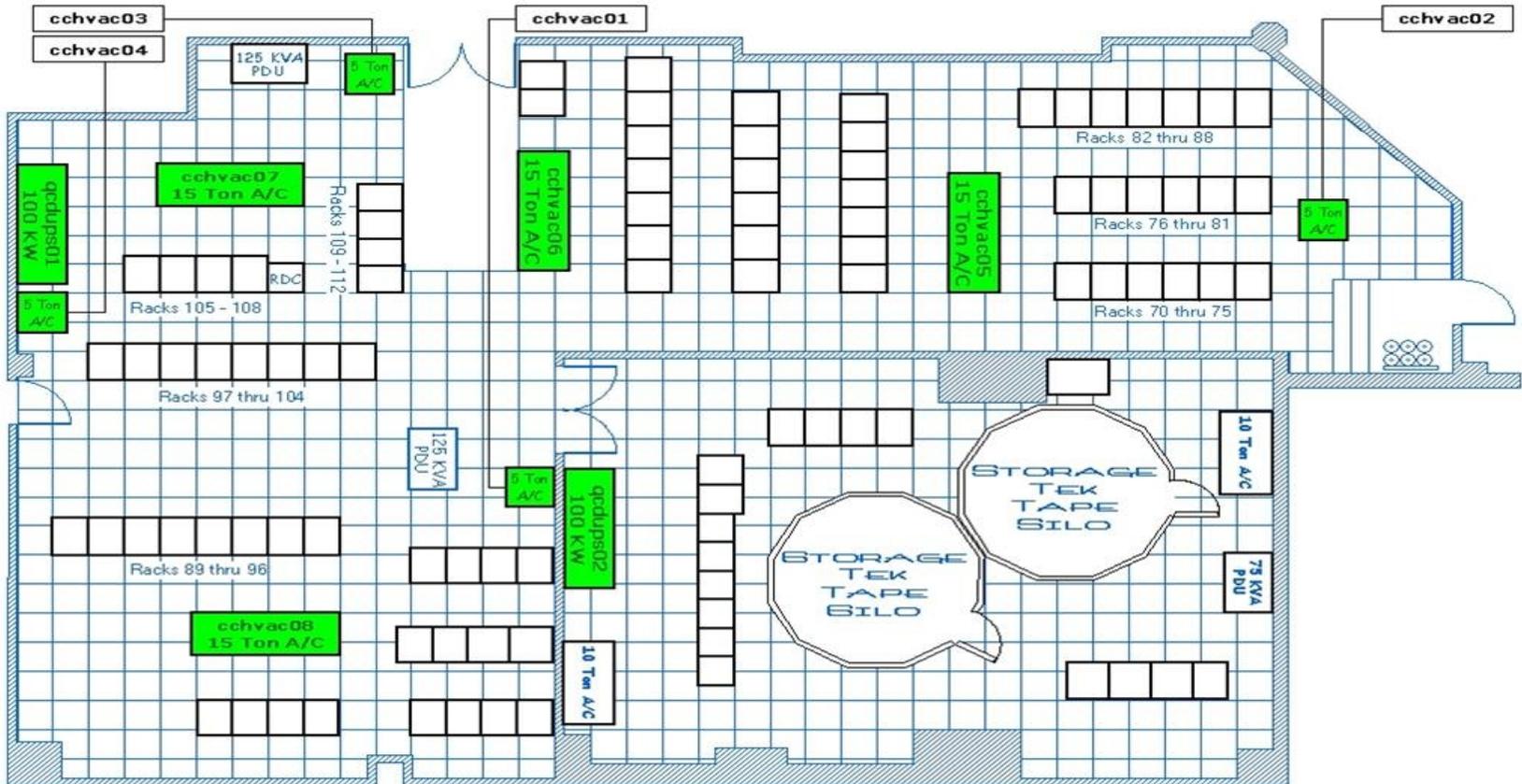
**Input 2 Frequency**  
 Value: 60 Hertz  
 59.5 Hz | 61 Hz | 60.5 Hz

**Input 2 Voltage**  
 Value: 470 RMS Volts  
 432 volts | 456 volts | 528 volts | 504 volts

# Site View at JLAB

## CENTRAL COMPUTING FACILITY

### CEBAF Center - Second Floor



# TRIUMPF Site Report

---

## TRIUMPF Fall 2005

- Request Tracker (RT) in use as a ticketing system
- Work toward 10GbE connectivity with CERN

## TRIUMPF Spring 2006

- 10GbE – upgrade connectivity to CERN using Foundry BigIron FX-4
- Cooling Issues:
  - Combined (2x) 11-ton ACs
  - Added heating coil
  - Blades for ATLAS (30% less heat, 20% less TCO)
  - Modular water cooling systems from HP
  - Open rack doors

# SC3 Results as Reported by TRIUMPF

---

*Tier0 → Tier1 Tests Apr 3 → 30*

<b>Site</b>	<b>Disk-Disk</b>	<b>Disk-Tape</b>
ASGC	100	75
<b>TRIUMF</b>	<b>50</b>	<b>50</b>
BNL	200	75
FNAL	200	75
NDGF	50	50
PIC	60*	60
RAL	150	75
SARA	150	75
IN2P3	200	75
FZK	200	75
CNAF	200	75

# GSI Site Report

---

## GSI Fall 2005

- Upgrade of OS on batch farm to Debian Sarge
- Farm uses fanless CPU servers without thermal difficulty

## GSI Spring 2006

- New air-cooled racks for the batch farm – doors can be closed with no thermal issues
- Content with WD 320 SATA II disks for mass storage
  - Problems with Maxtor III Maxline Nearline disks

# Air Cooled Racks at GSI



## Batch farm upgrade

- new AIR Cooled Racks!
  - Powerfull Fans
  - 47 HE capacity
- 100 new boxes 1 HE
  - Tyan B2891
  - Dual Core Operon 2.4 Ghz
  - 4/8GB, 4 hot swap slots
  - 2 x 160 GB disks
  - No fans on board
- Air cool system
- Air cooled/closed doors
- No thermal problems
- in: 16 C° constant, Floor
- 38 C° Top, back side



Walter Schön, GSI

# NERSC Site Report

---

## NERSC      Fall 2005

- Migrate from LSF to SGE 6.0 release 4
- Investigating SL4
- Migrate toward jumbo frame network
- Lustre and GPFS installed and in production
- OTP (Steve Chan)

# Air-Cooled Cabinet Cooling (Bill Watts, Intel)

---

## Findings

- Air flow efficiency is measured by the amount of heat that can be removed
- Rack cable support arms for wire management actually block airflow and act as heat sinks
- Intel worked with cabinet vendors to increase the air-cooling capacity of their racks:
  - Increased cabinet depth and plexiglass cabinet door produce a “chimney flue” for heated server exhaust.
  - Chimney flue connected to ceiling return space
  - Room cooling returns also connected to ceiling return
  - Such racks cost ~16K each

# CPU Technologies (Bernd Panzer-Steindel)

---

- Talk focused on the CPU roadmaps of Intel and AMD – smaller fabrication (64 nm), power efficiency, and multicore systems (8 core by 2009)
- Game processors: XBox 360 (1k GFLOPS), Playstation 3 (cell proc, 8 cores, 1.8k GFLOPS). Can HEP harness their power and low cost?
- Multi-core systems require more memory (4 core == 10GB RAM)

# CPU Power Comparisons (Yannick Perret)

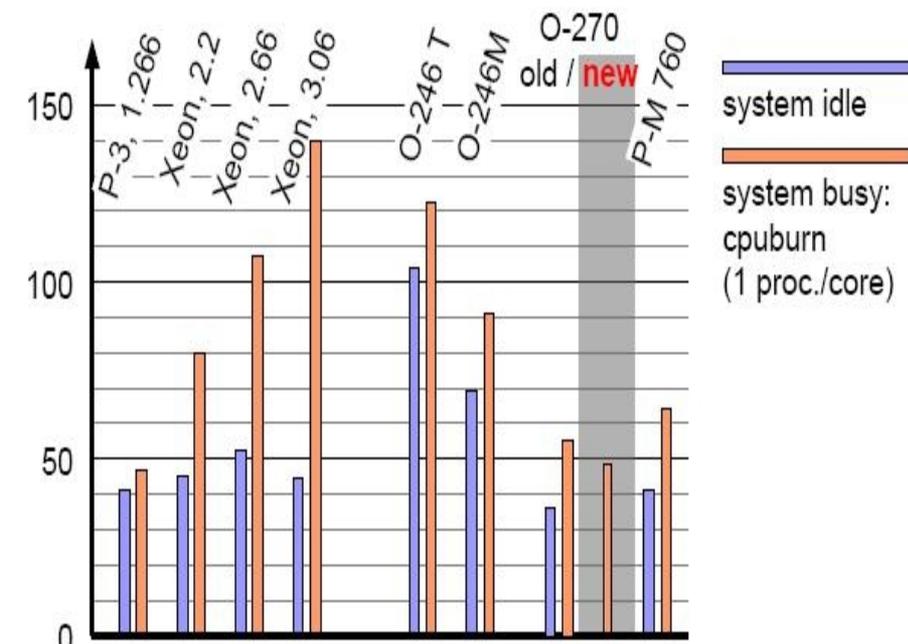
---

- Serious cooling and power issues during the summer forced CC-IN2P3 to reevaluate requirements for future procurements
- Need systems that consume less power and generate less heat
- Conclusions:
  - Opteron superior to Xeon (CPU and power)
  - Dualcore superior to hyperthreading
  - One power supply better than two or more
  - Big companies (SUN, IBM) better than smaller
  - Blade systems better than others

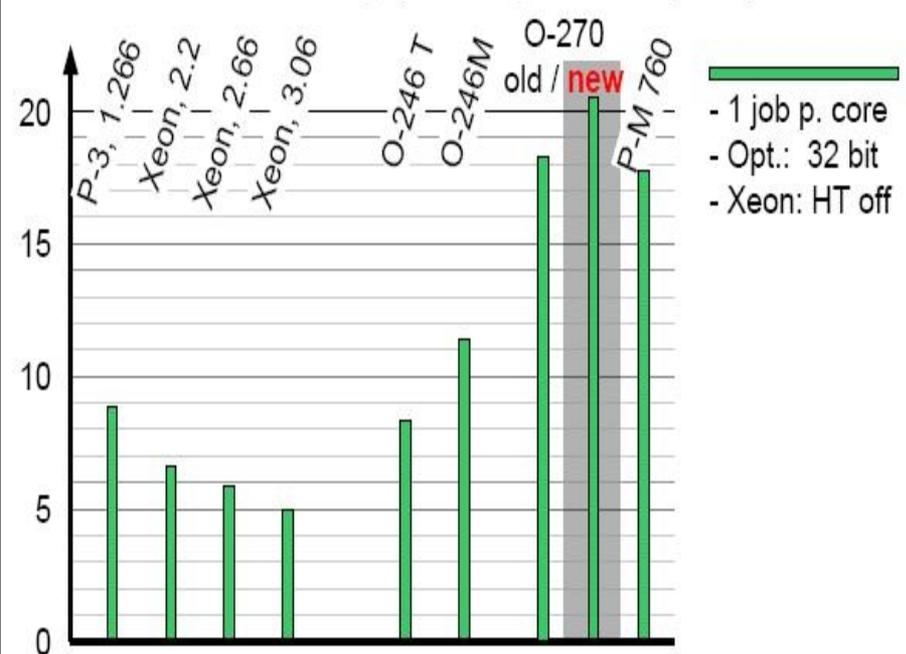
# Dual-core Batch Nodes at GridKa (Manfred Alef)

- Concluded that, again, Opteron processors provide greater performance and less power consumption than their Intel counterparts.

Electric Power Consumption (Whole System, W per CPU Core)



Thermic Efficiency (SPECint\_base2000 per W)



# AMD64 vs. EMT64 (Ian Fisk)

---

## Conclusions

- Dual-core offers double the computing power of single-core at the same electricity consumption
- 64bit architecture offers a big performance boost for ported applications
- 32bit applications run fine without modification on 64bit processors in either 32bit mode or 64bit compatibility mode
- Applications tested were Pythia, Root, Oscar, and Orca (Digi and DST)

# 10GbE TCP Performance Optimization (Tiziana Ferrari)

---

- Provided a thorough and comprehensive testing of 10GbE optimization in preparation of future LHC service challenges
  - Comparison of 10GbE NICs
  - HW and SW tunable parameters for max single flow throughput
  - TCP stack comparisons
    - Reno (Linux 2.4): 5 streams @ 6.2 Gb/s
    - BIC – Binary Increase Congestion Control (Linux 2.6): 5 streams @ 6.8 Gb/s
      - better performance over long-distance paths
      - more efficient process scheduling algorithms results in superior multiple-stream performance
- 10GbE transmission is CPU-bound

# Tape Technology (Don Petravick)

---

- At Fermi:
  - Tape growth ~1 PB/year
  - 16 PB moved/year; 20 TB/day
  - 3-4 bytes read for each byte written
- Tape capacities double every 18-24 mos (current LTO-3 @ 400GB)
- Tape density roadmaps face few fundamental engineering challenges vs. disk
- Tape offers easy and reliable expansion
- However, will tape remain viable in the market?

# Disk Storage, Interconnects, Protocols (Martin Gasthuber)

---

- Discusses various storage implementations such as FC SAN, SCSI/FC external RAID, Internal PCI RAID, NAS
- Flat disk technology outlook for next 2-4 years:
  - Little improvements (IOps, bandwidth, seek)
  - No FC growth
  - SaS will shrink to the benefit of the DB market (8 spindles per CPU core)
  - 24/7 datacenter SATA drives at higher capacity, lower RPM
- RAID 6 offers protection from disk/controller issues.
- High-end SATA controllers are needed soon for:
  - real disk error handling (parity inside block)
  - RAID scrubbing
  - Better SMART analysis

# GPFS and StoRM at INFN Tier-1

## (Luca Dell-Agnelo)

---

- GPFS in use at INFN for >2 years.
  - Stable, reliable, fault tolerant, fully POSIX compliant
  - Free for academic use but difficult to obtain IBM support
  - Non-invasive – doesn't require any kernel mods
  - All farm nodes need passwordless root access via rsh or ssh – if one node is compromised, they all are.
- Lustre outperformed GPFS but is too intrusive requiring kernel mods.
- StoRM is a storage-based resource manager optimized for use with GPFS that is under investigation. It may be modified to interface with Lustre should that product be tested again.

# Local File Systems (Peter Keleman)

---

- Investigates block-structured, extent-based, and journaling file-systems such as ext3, XFS, JFS, ReiserFS
- ext3 (block structured, journaled (metadata, metadata+data), 10k SLOC)
  - Active development at RedHat (RHEL4 inclusion)
  - Stable & widely used
- XFS (extent-based, multiple B+-trees, 100k SLOC)
  - Active development at SGI
  - Disabled in RHEL4
  - Fully 64-bit, geared for large files
- JFS (extent-based, multiple B+-trees, 30k SLOC)
  - Active development at IBM
  - Disabled in RHEL4

# Local File Systems (Peter Keleman)

---

- CERN opted for XFS which they included in SLC3
- ~650 TB XFS in production, however:
  - 4TB AGs unstable under load
  - v2 log replay mem allocation problems
- XFS in SL4 (kernel 2.6.9)
  - i386: with 4k stack, heavy load triggers overflow
  - However, OK with x86\_64
  - ext3 is catching up in performance – offers better local streaming speed, but <<slow>> deletes

# AFS/OSD Project (Ludovico Giammarino)

---

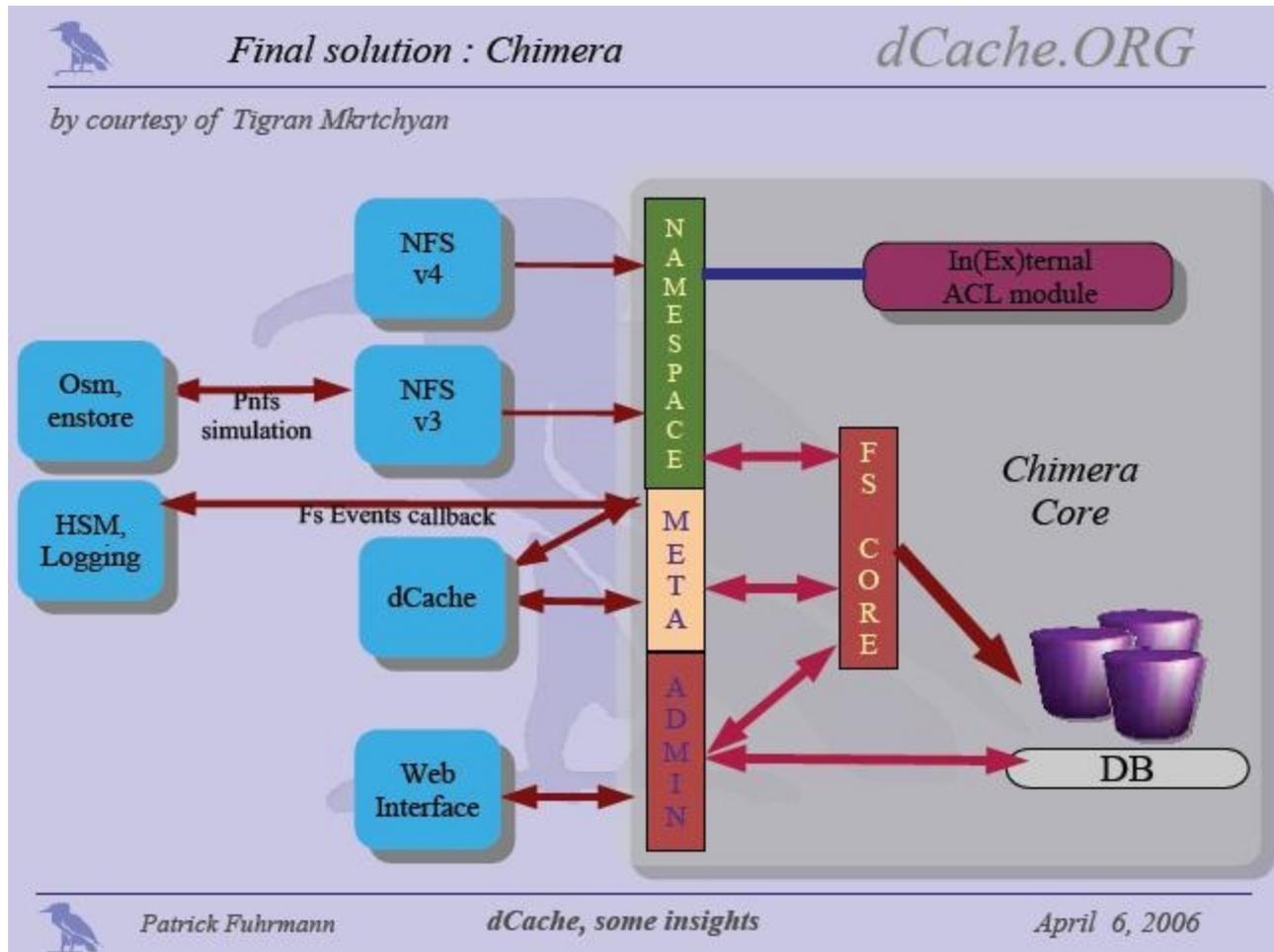
- Focus on improving AFS performance and scalability:
  - Single fileserver throughput is too low
  - AFS volumes limited to single disk/lun/filesystem
  - Inability to stripe large files across file servers
  - No RW replication
- AFS extended to support SCSI T10 OSD standard:
  - Panasas-like
  - Object based storage device (OBSD)
  - Command descriptor block (CBD): format for executing commands on OBSD
  - OBSD access via RX-protocol (RPC over UDP)
- Goal is for a high-performance prototype by late summer 2006
- Currently, however, volume replication is not supported

# dCache (Patrick Fuhrmann)

---

- A detailed discussion of the data and control flow of dCache
- The list of dCache developers is growing beyond the DESY/Fermi base
- A plethora of data transfer protocols
  - Local Area: (gsi)dCap, xRootd
  - Wide Area: (gsi)FTP, http (never used)
- VOMS integration for pluggable authentication/authorization
- Issues: Name Space Provides (Pnfs) seen as potential bottleneck
  - band-aid solution: Pnfs partitioning since dCache can run against multiple pnfs instances
  - long-term solution: Chimera
    - Eliminates OS dependency for name space provider (Current test running a pool on XP)
    - dCache can run with any JDBM enabled DB (not just gdbm or Postgres)

# dCache – A look at Chimera

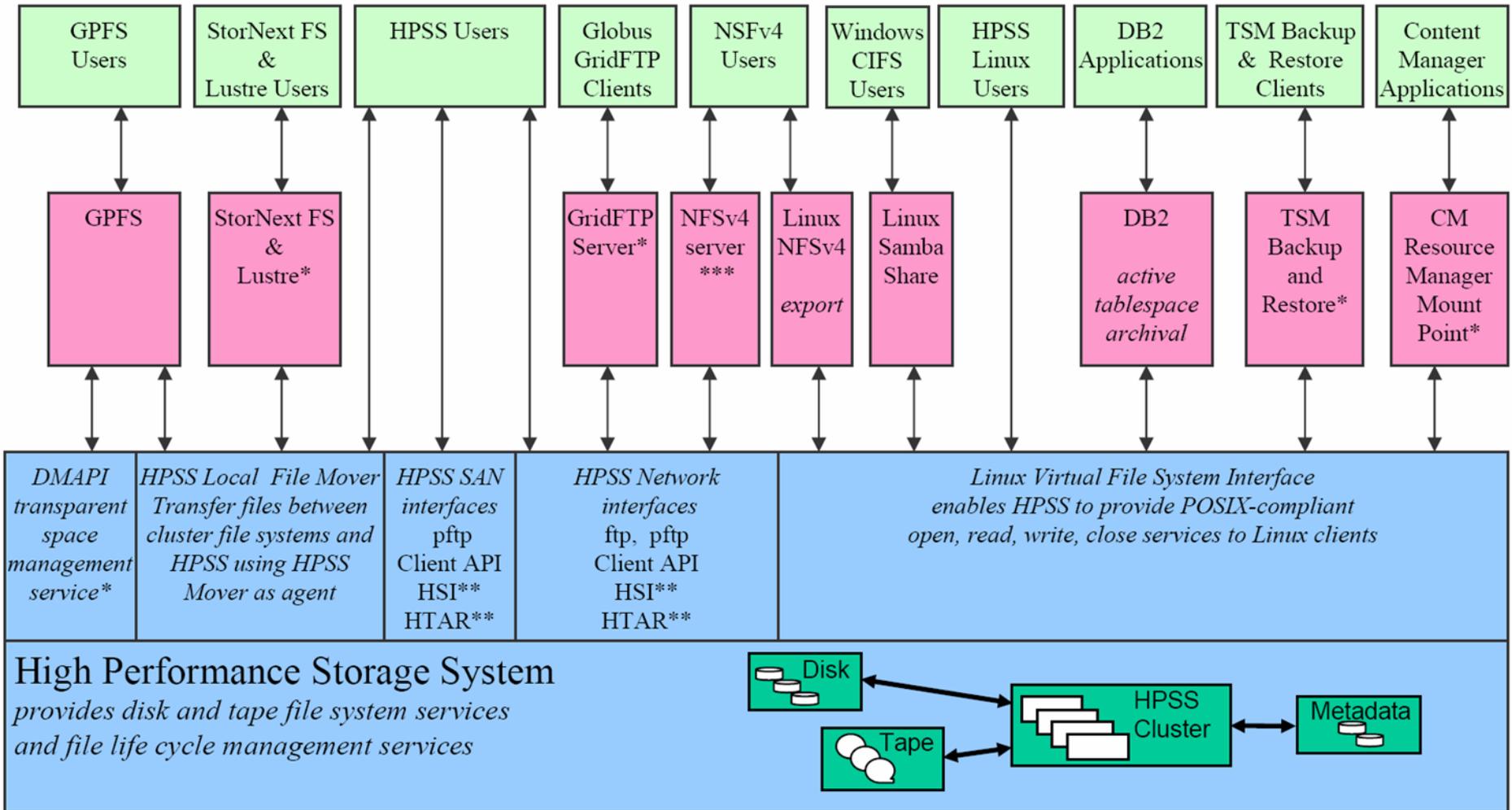


# HPSS (Andrei Moskalenko)

---

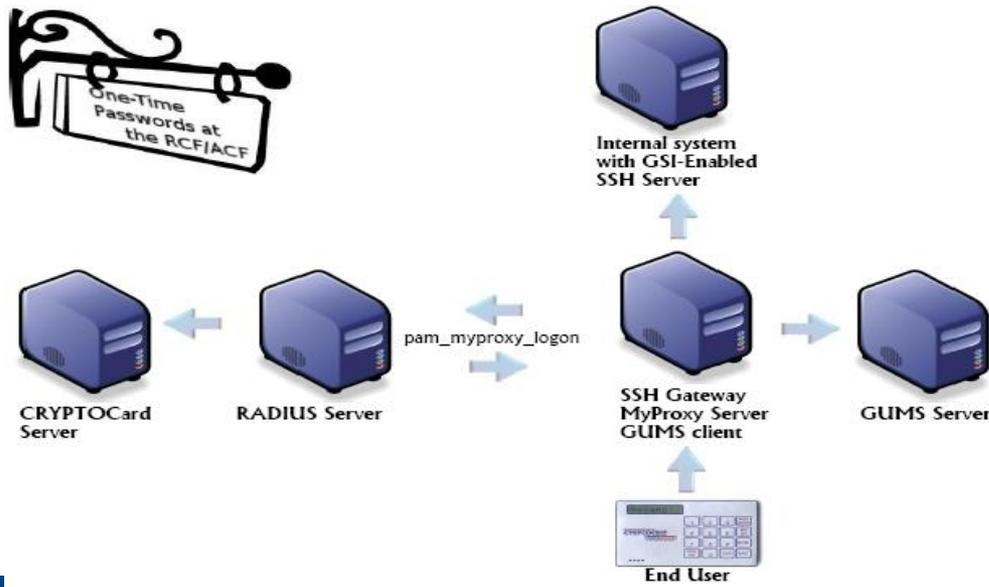
- HPSS is in use at IN2P3 and is seen as a robust, modular, scalable HSM system
- Can host 10s of PBs of data – 100s?
- Bottlenecks, irritants, qualms are discussed:
  - New data ends up on same media as old data (repack)
  - Easy to write, difficult to read (but write/reads = 15/85)
  - Volatile DB2 tables after repacks-n-deletes
  - Administrative idiosyncracies: weak tape error support, lame error messaging, non-dynamic configuration
- A host of HPSS 6.1 access methods provided

# HPSS 6.1 Client Access Points



# OTP – SSO at BNL (R. Petkus)

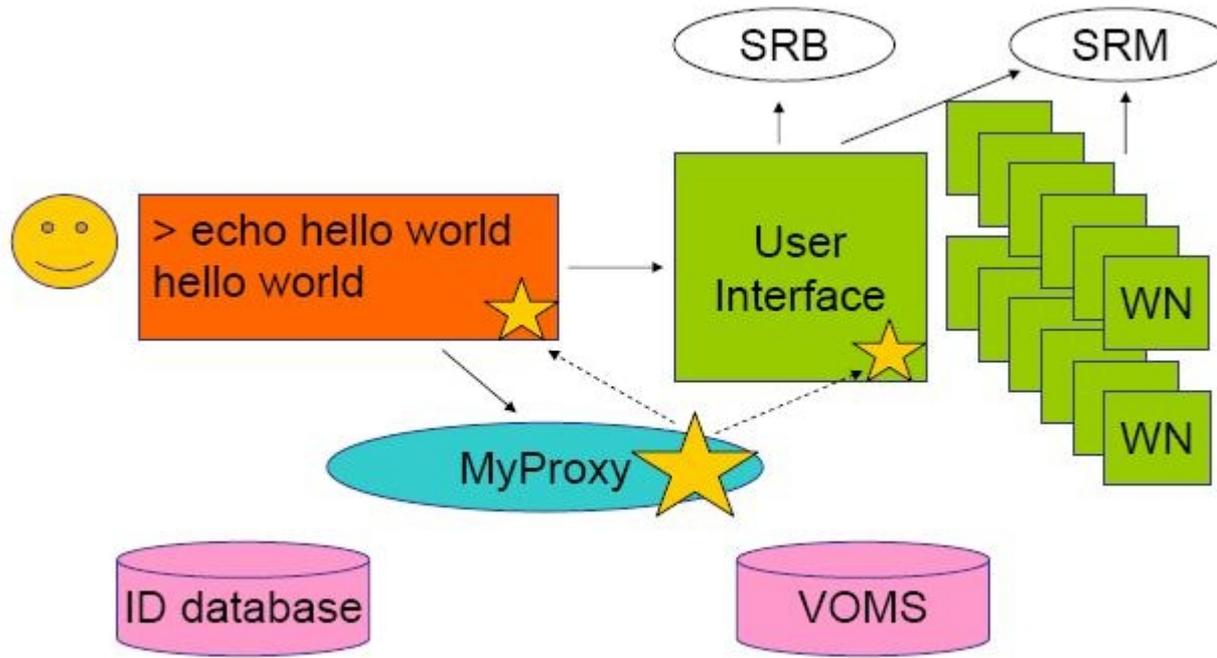
- Current offerings are SSO via Kerberos (AFS token via aklog) or ssh keys.
- Deficiencies in these models along with recent DOE mandates require integration with a OTP system.
- One scheme is introduced for use on an interactive ssh gateway that is undergoing tests:



# SSO to the Grid at RAL (Jens Jenson)

- SSO encompasses identity management, credential conversions (certificates, Active Directory, Kerberos), password validation.
- Different modes of SSO:
  - On-site w/federal ID (AD/Kerberos)
  - Offsite w/certificate loaded into browser
  - Otherwise, access w/username & password
- Use of Java SSHTerm enhanced with GSISsh plugin and hacked for MyProxy compatibility integrates nicely into SSO scheme at RAL – predictable integration w/AD

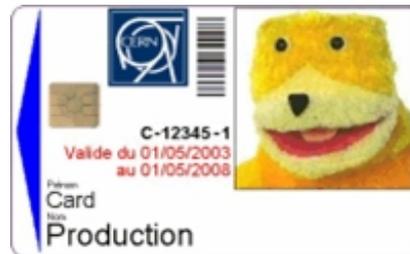
# JAVA SSHTerm



Jens G Jensen  
CCLRC e-Science

# Integrating PKI & Kerberos (Emmanuel Ormancy)

- CERN currently uses both PKI (Grid-related work) and Kerberos (Windows domains & AFS) and efforts are underway to integrate them.
- Plans to create a CA integrated w/Kerberos in 2006/2007
- Foreseen accreditation from European Grid Policy Management Authority
- Examples of web authentication and e-mail signing
- Future authentication via SmartCards (Certificate and private key in a HW token)
  - Allow users to map existing certificates (issued by trusted CA) to their Kerberos account



# Scientific Linux Roadmap (Troy Dawson)

---

- Current use of SL  $\geq$  16k installations
- Top 3 countries using SL are the US, UK, and Taiwan
- Preparations in place for SL5 – dependent on RedHat's schedule – but preliminary investigations underway using Fedora Core 5
- Removal of Bug Tracker from website
- Call for volunteers

# Future HEPiX Meeting Locales

---

- Fall 2006 (October 9) at Jefferson Lab, Virginia
- Spring 2007 at DESY, Hamburg