CERN IT Department

HEPiX Report

Helge Meinhard, Steve Murray,
Miguel Coelho dos Santos / CERN-IT

Computing Seminar / After-C5

21 November 2008

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Outline

- Meeting organisation, site reports, data centre track (Helge Meinhard)
- Storage (including Castor), benchmarking (Miguel Coelho dos Santos)
- Operating systems and applications, virtualisation, network and security, miscellaneous (Steve Murray)

HEPiX

- Global organisation of service managers and support staff providing computing facilities for HEP
- Covering all platforms of interest (Unix/Linux, Windows, Grid, …)
- Aim: Present recent work and future plans, share experience; provide technical advice to IHEPCCC (now suspended)
- Meetings ~ 2 / y (spring in Europe, autumn traditionally in North America)

HEPiX Autumn 2008 (1)

CERN IT Department

- Held 20 – 24 October at Academia Sinica in Taipei, Taiwan
  - Second meeting in Asia ever…
  - First one was in Tsukuba (JP) in 1991!
  - Intention: Offload US and CA, attract participants from Asia
  - Excellent organisation (Simon Lin, Vicky Huang, Stella Shen, Jill Lin et al.) set very high standards
    - Difficult to meet for other sites…

HEPiX Autumn 2008 (2)

- Format: Pre-defined tracks with conveners and invited speakers per track
  - Still room for spontaneous talks – either fit into one of the tracks, or classified as ‘miscellaneous’
  - Again proved to be the right approach
  - Judging by number of submitted abstracts, storage remains a very hot topic
- Full details, slides and tons of photographs: http://indico.twgrid.org/internalPage.py?pageId=1&confId=471
  - Alan Silverman’s trip report available, too

HEPiX autumn 2008 (3)

- 86 registered participants, of which 12 from CERN
  - Andreeva, Cass, Coelho dos Santos, Duellmann, Field, Lopienski, Lo Presti, Meinhard, Murray, Otto, Ponce, Silverman
  - Other sites: **ASGC,** CASPUR, CEA, DESY Hamburg, DESY Zeuthen, FNAL, GSI, **HAII Bangkok,** IN2P3, INFN, JSI Ljubljana, **KEK, KISTI (South Corea),** KIT, LAL, LBNL, **Manila, U Melbourne,** U Michigan, NIKHEF, NDGF, NSC Sweden, **Punjab (Pakistan), Putra (Malaysia),** Prague, RAL, SLAC, **U Taipei, Tata, Tokyo,** TRIUMF, U Victoria
  - Compare with St Louis (autumn 2007): 62 participants, of which 11 from CERN

HEPiX Autumn 2008 (4)

CERN IT Department

- 64 talks, of which 16 from CERN
  - Compare with St Louis: 59 talks, of which 15 from CERN
- Next meetings:
  - Spring 2009: Umeå (Sweden, 25 – 29 May 2009)
  - Autumn 2009: Not yet fixed (LBNL interested)
  - Further application: GSI

Site Report Highlights (1)

- Cooling, power, space
  - – Shortfalls mentioned by all large and most small centres
  - – Solutions include new data centres (RAL), temporary structures (like Blackbox – not without issues), refurbishing rooms/halls built for other purposes, using remote rooms or commercial hosting, upgrading structures in place
  - – Too little focus (IMHO) on power-efficient systems
  - – More problems due to harmonics reported
  - – Increasing number of sites retire old stuff aggressively

Site Report Highlights (2)

- CPU worker nodes
  - Mixture of blades, twins and traditional 1U pizza boxes (sometimes even with redundant PSUs)
  - Mostly Xeons (Harpertowns), some Opterons (driver issues for chipset)
- Disk storage
  - Mixture of storage-in-a-box (either “white boxes” or Thumpers), Infortrend (or similar) arrays with FC (2400 TB in INFN SAN), larger units (e.g. DataDirectNet in Karlsruhe – 16 PB over 3 years)
    - Separating storage from serving nodes appears to be architectural criterion at some sites
- Tape storage: SL8500, LTO drives

Site Report Highlights (3)

CERN IT Department

- File systems
  - Increasing interest in cluster file systems
  - A number of sites mentioned GPFS
    - Mostly qualitative statements, no performance data
  - Lustre
    - Detailed reports by DESY and GSI
- Windows: Still some (a lot of?) reluctance concerning Vista (in particular 64-bit) and Office 2007

Site Report Highlights (4)

CERN IT Department

- Communications
  - – More and more sites considering 10 GigE to storage nodes
  - – VoIP and unified messaging at FNAL
  - – IPv6: Dedicated talks, but little activity
- Software
  - – Quattor, Lemon mentioned a few times
  - – InDiCo used at most major sites (including ASGC)
  - – DPM
  - – Castor
  - – Subversion taking off (and over from cvs)
  - – MAUI: Scaling problems; LSF (and GPFS): Issues with core-based licencing

Site Report Highlights (5)

- Miscellaneous
  - Number of name changes, structure changes, personnel changes, …
  - ITIL mentioned by a number of sites, FNAL going for ITIL V2 and ISO 20000 certification within two years
  - Some sites reported to start (only) now deploying IPMI
  - Operational coverage: Some Tier1 sites working with 7 hours/working day operator coverage only

Tracks

- Storage issues
- Castor
- **Data centres**
- Operating systems and applications
- Virtualisation
- Network and security
- Benchmarking
- Miscellaneous

Data Centre Track (1)

- Presentations:
  - New CERN computer centre (Tony Cass)
  - Push-pull registry for discovery of OWL-S Web services (Hafiz Farooq Ahmad)
  - Data centre thermodynamic CFD (Enrico Mazzoni)
  - Roma Green computing centre (Giovanni Organtini)
  - High-density displays in CluMan project (Miguel Coelho dos Santos, for Sebastian Lopienski)

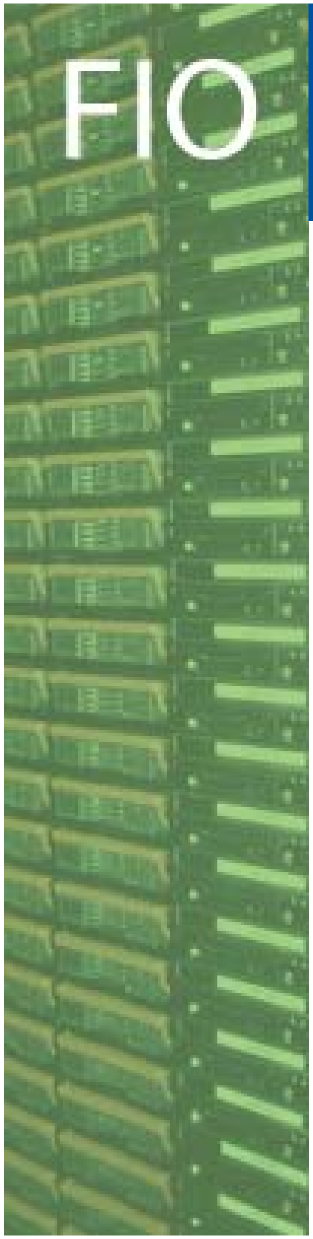
Data Centre Track (2)

- Data centre thermodynamic CFD (Enrico Mazzoni)
  - Report about work in progress
  - Motivation: higher energy densities require more detailed knowledge
  - Input: machine room design parameters (CAD model), server properties from data sheets
  - CFD results have correct tendency
    - But are a few degrees to high
    - Will iterate: change input parameters to make the results match better with room measurements
  - CFD is a valid tool to obtain useful information about cooling

Data Centre Track (3)

- Roma Green computing centre (Giovanni Organtini)
- – Rather small layout with 14 racks (Knürr 17 kW)
  - Battery UPS with 120 kVA feeding racks and water pumps
  - 3000 l water at 12 deg C
- – Currently at 45 kW for servers, 100 kW total
- – Estimate for full load: 90 kW for servers, 59 kW for services
  - Total / servers = 1.65
  - Estimate for air cooling: ~ 70 kW more

[redacted]



Fabric Infrastructure
and Operations

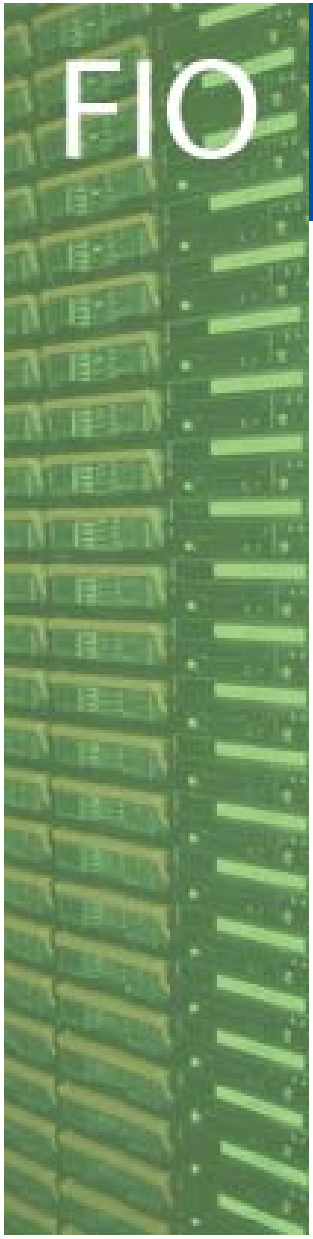
CERN IT
Department

HEPiX Fall 2008
ASGC (Taiwan)

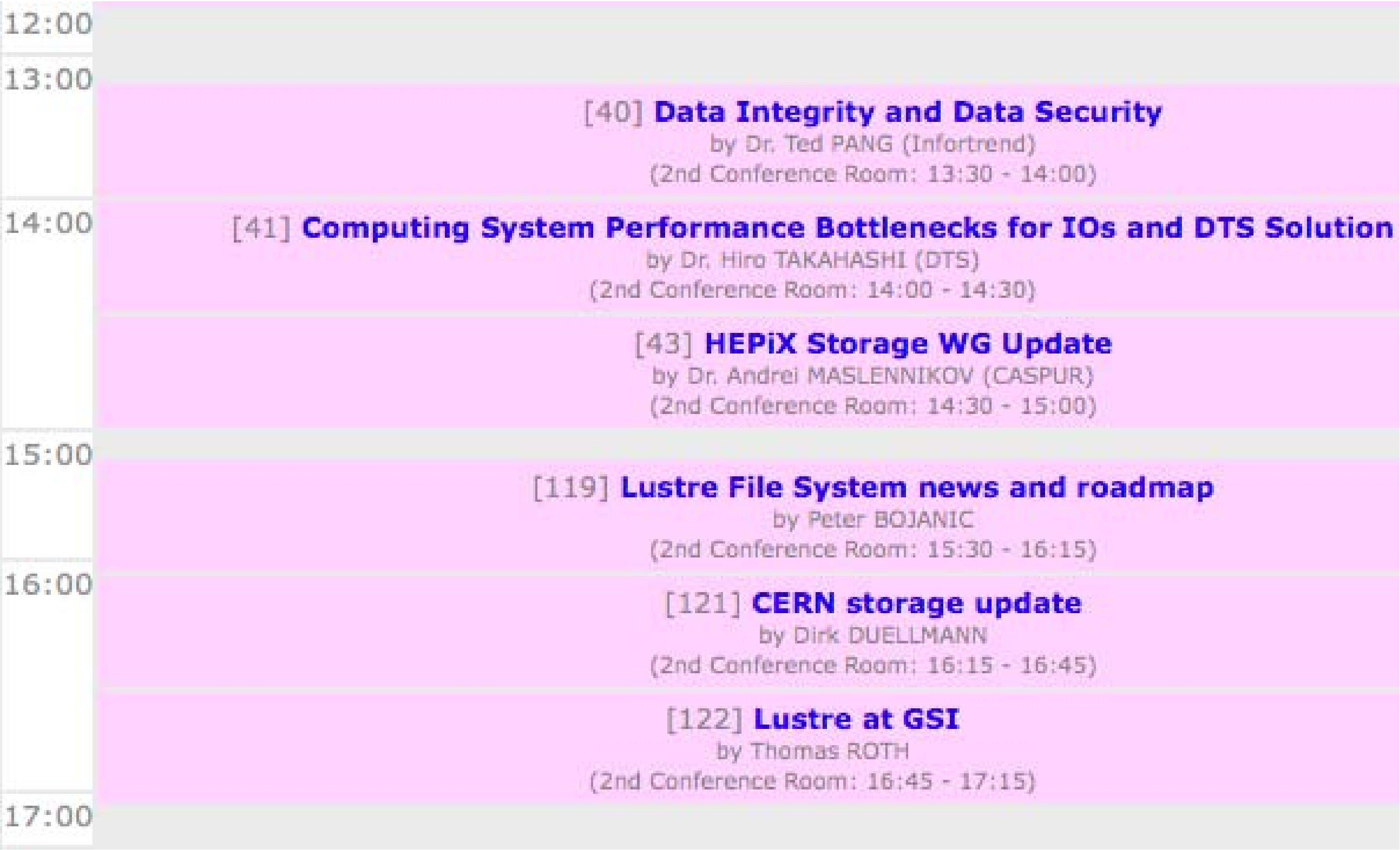
Storage issues, Castor and Benchmarking report

Miguel Coelho dos Santos
FIO/FS

November 2008



Storage track – day one



12:00

13:00

[40] **Data Integrity and Data Security**
by Dr. Ted PANG (Infortrend)
(2nd Conference Room: 13:30 - 14:00)

14:00

[41] **Computing System Performance Bottlenecks for IOs and DTS Solution**
by Dr. Hiro TAKAHASHI (DTS)
(2nd Conference Room: 14:00 - 14:30)

[43] **HEPiX Storage WG Update**
by Dr. Andrei MASLENNIKOV (CASPUR)
(2nd Conference Room: 14:30 - 15:00)

15:00

[119] **Lustre File System news and roadmap**
by Peter BOJANIC
(2nd Conference Room: 15:30 - 16:15)

16:00

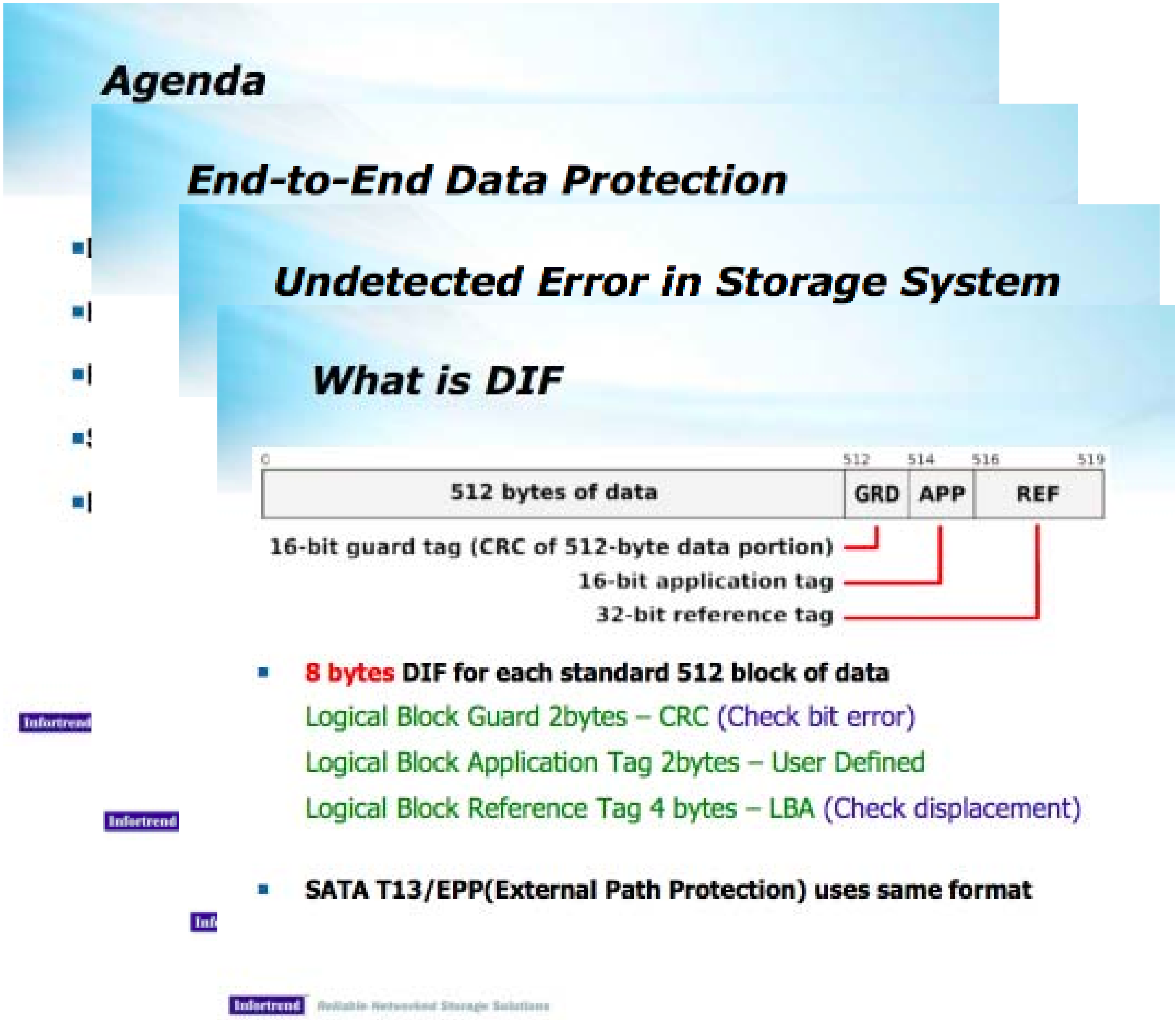
[121] **CERN storage update**
by Dirk DUELLMANN
(2nd Conference Room: 16:15 - 16:45)

[122] **Lustre at GSI**
by Thomas ROTH
(2nd Conference Room: 16:45 - 17:15)

17:00

Infortrend

CERN IT Department



Agenda

End-to-End Data Protection

Undetected Error in Storage System

What is DIF

0	512	514	516	519
512 bytes of data	GRD	APP	REF	

16-bit guard tag (CRC of 512-byte data portion)

16-bit application tag

32-bit reference tag

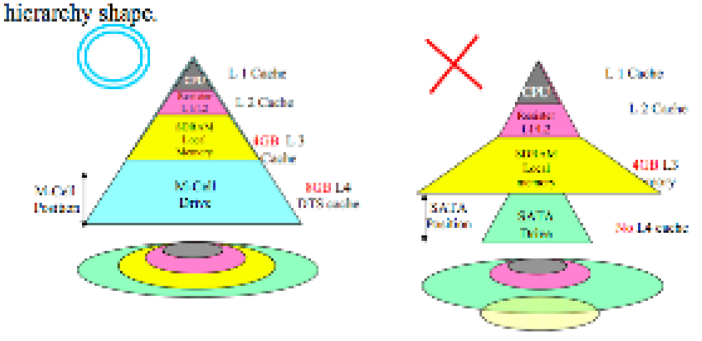
- 8 bytes DIF for each standard 512 block of data
  Logical Block Guard 2bytes – CRC (Check bit error)
  Logical Block Application Tag 2bytes – User Defined
  Logical Block Reference Tag 4 bytes – LBA (Check displacement)
- SATA T13/EPP(External Path Protection) uses same format

Infortrend Reliable Networked Storage Solutions

DTS

Proposed concept: Balanced Memory Architecture

- Unbalanced performance shape creates the bottleneck of computer utilization.
- DTS basis memory management maintains balanced performance hierarchy shape.

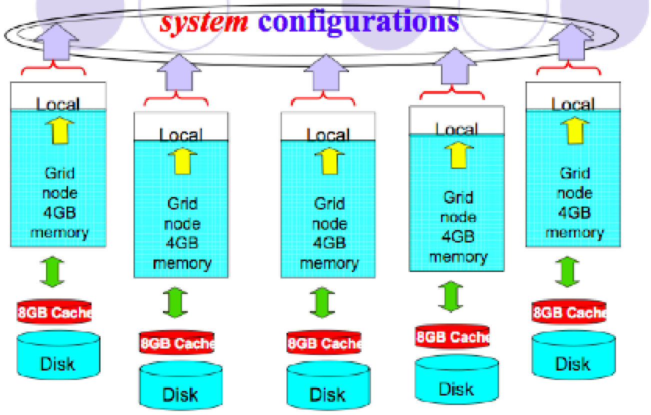


DTS memory basis memory hierarchy.

Traditional Computer memory hierarchy.

15

Proposing *DTS L4 cache memory system* configurations

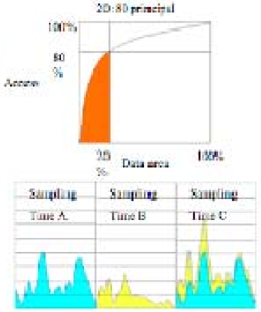


Reduce a nuclear destruction semiconductor memory error on the local memory.

17

DTS Intelligent caching algorithm

- Decide the most busy block address area.
- As 80 /20 principal, DTS intelligent cache keep these data on DTS cache table memory effectively.
- The DTS system enable most effective IO transfer performance by own cache technology and DTS working table's physical speed.



22

IO performance test result: DTS VS traditional computer

We researched write IOPS under DTS memory configuration and standard computer configuration as well. The result was very positive. For instance, 512 100%Write0%Read Random result is 211.51 IOPS but DTS based is 33354.32 IOPS under eight multiple accesses situation. Also 1K 80%Write20%Read Random result is 237.14 IOPS but DTS based is 32020.75 IOPS under eight multiple accesses situation. (Refer figure 6)

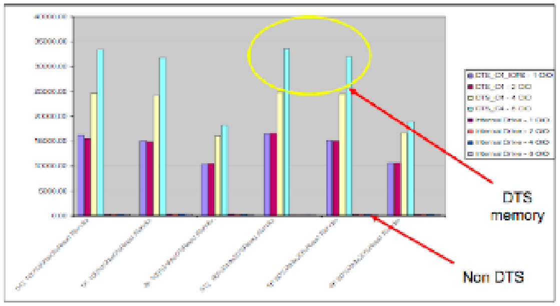


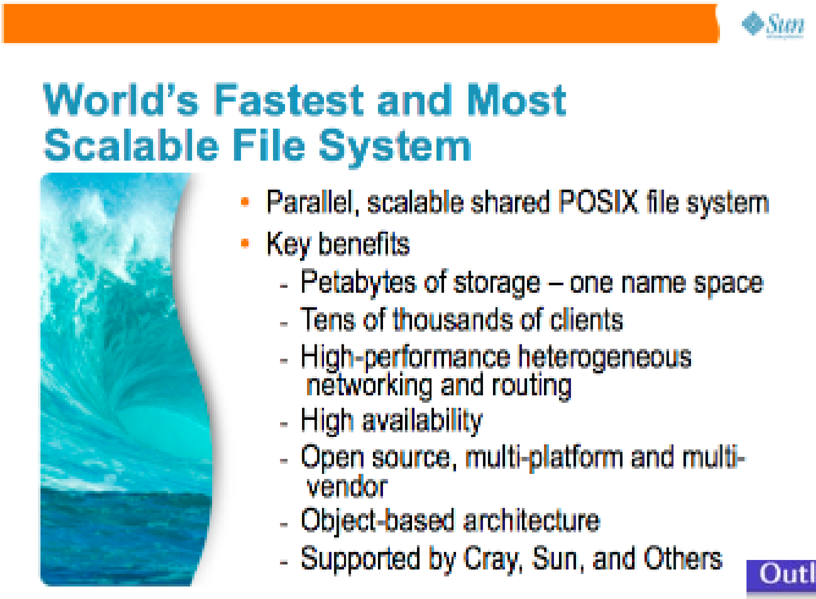
Figure 6: DTS type vs Traditional computer IOPS test result.

(Intel Xeon 2.4 GHz, 4GB System memory, 4GB Working Table, Target File Size = 2GB Write Back Policy, Windows2000 Advance SP4 Server Target Disk size 4GB)

25

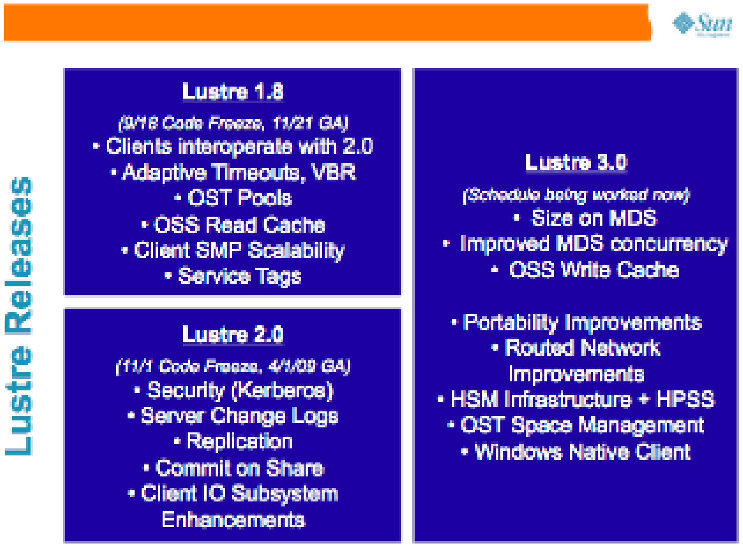
FIO

Lustre



World's Fastest and Most Scalable File System

- Parallel, scalable shared POSIX file system
- Key benefits
  - Petabytes of storage – one name space
  - Tens of thousands of clients
  - High-performance heterogeneous networking and routing
  - High availability
  - Open source, multi-platform and multi-vendor
  - Object-based architecture
  - Supported by Cray, Sun, and Others



Lustre Releases

Lustre 1.8

(9/18 Code Freeze, 11/21 GA)

- Clients interoperate with 2.0
- Adaptive Timeouts, VBR
- OST Pools
- OSS Read Cache
- Client SMP Scalability
- Service Tags

Lustre 2.0

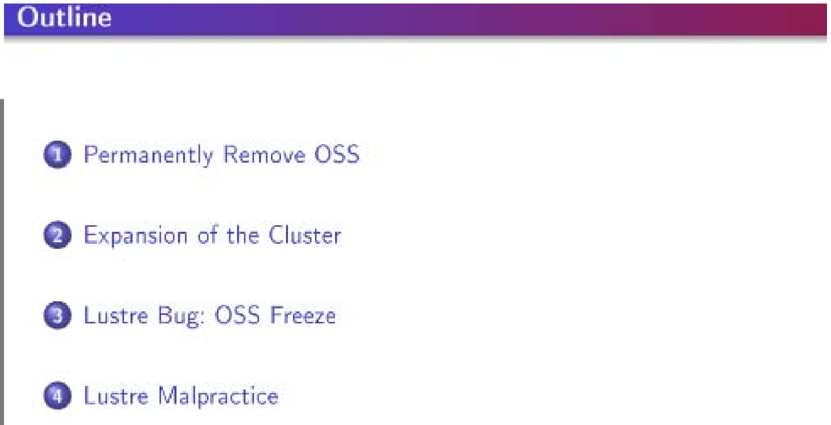
(11/1 Code Freeze, 4/1/09 GA)

- Security (Kerberos)
- Server Change Logs
- Replication
- Commit on Share
- Client IO Subsystem Enhancements

Lustre 3.0

(Schedule being worked now)

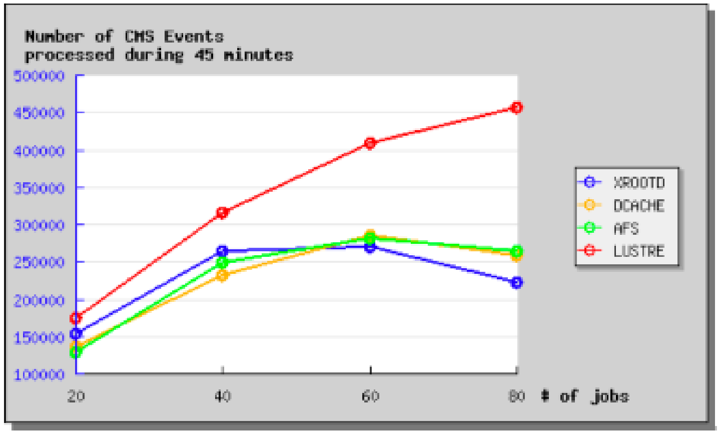
- Size on MDS
- Improved MDS concurrency
- OSS Write Cache
- Portability Improvements
- Routed Network Improvements
- HSM Infrastructure + HPSS
- OST Space Management
- Windows Native Client



Outline

1. Permanently Remove OSS
2. Expansion of the Cluster
3. Lustre Bug: OSS Freeze
4. Lustre Malpractice

First results (16/10/2008)

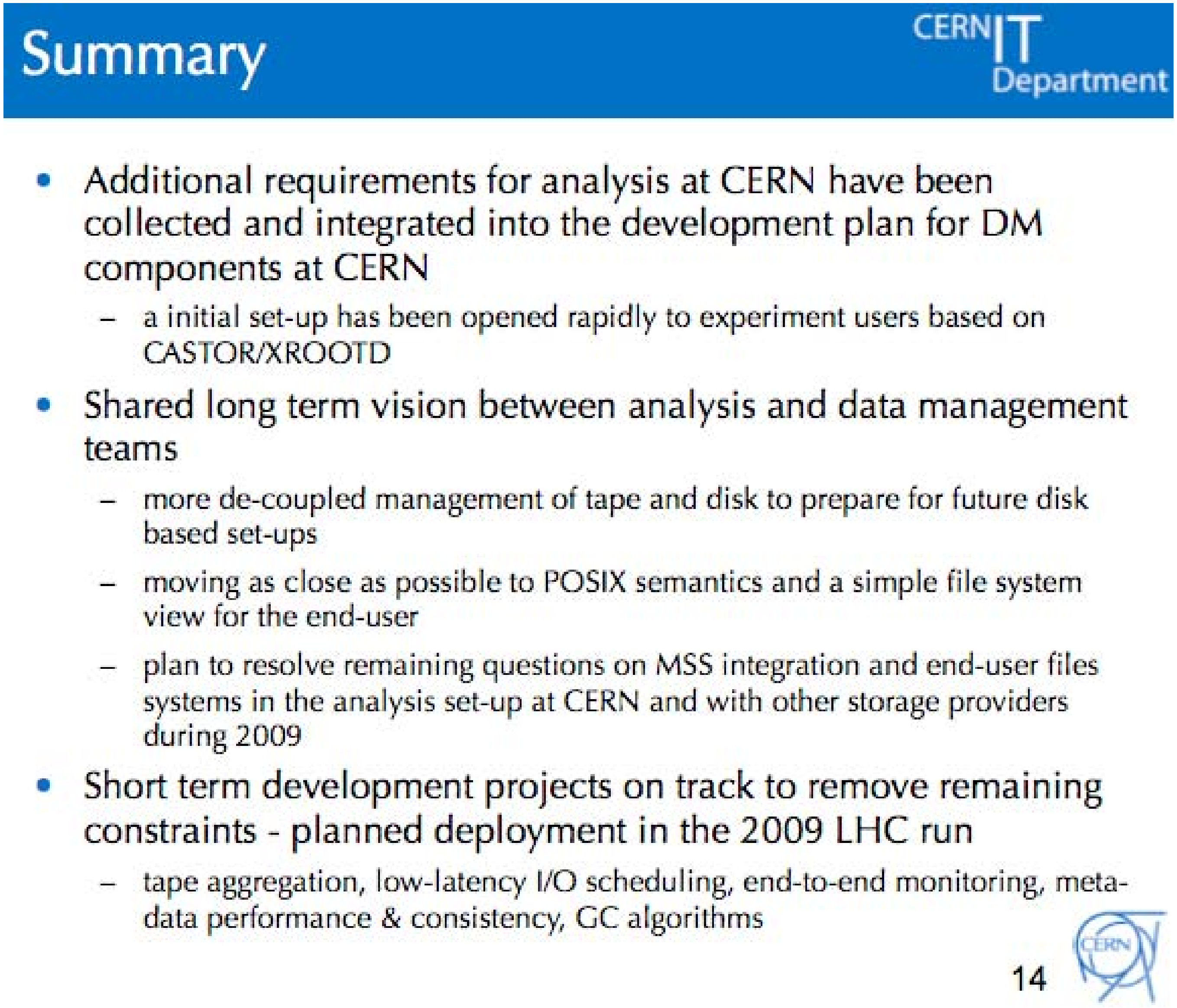


In this summary graph, Lustre seems to be almost twice as efficient compared to the other methods. As AFS, dCache and Xrootd seem to be very close to each other, they may have had a common blocking factor such as the local file system (and NOT the data access protocol).

FIO

CASTOR

CERN IT Department

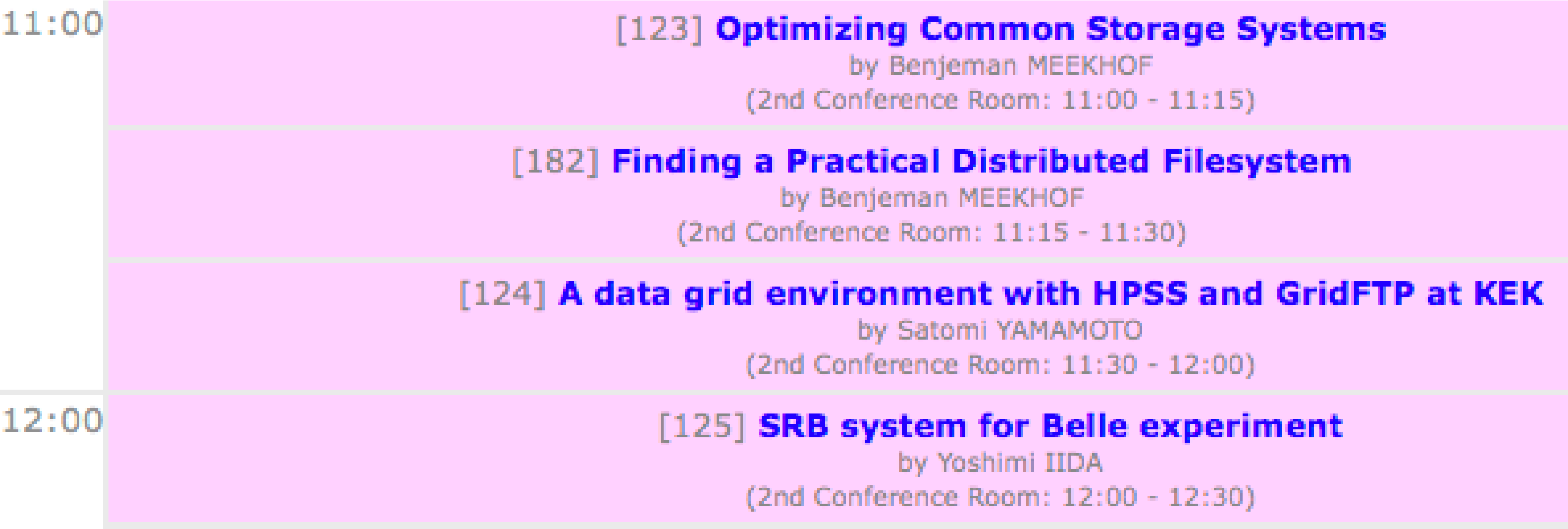


Summary

CERN IT Department

- Additional requirements for analysis at CERN have been collected and integrated into the development plan for DM components at CERN
  - a initial set-up has been opened rapidly to experiment users based on CASTOR/XROOTD
- Shared long term vision between analysis and data management teams
  - more de-coupled management of tape and disk to prepare for future disk based set-ups
  - moving as close as possible to POSIX semantics and a simple file system view for the end-user
  - plan to resolve remaining questions on MSS integration and end-user files systems in the analysis set-up at CERN and with other storage providers during 2009
- Short term development projects on track to remove remaining constraints - planned deployment in the 2009 LHC run
  - tape aggregation, low-latency I/O scheduling, end-to-end monitoring, meta-data performance & consistency, GC algorithms

Storage track – day two



11:00

[123] **Optimizing Common Storage Systems**
by Benjeman MEEKHOF
(2nd Conference Room: 11:00 - 11:15)

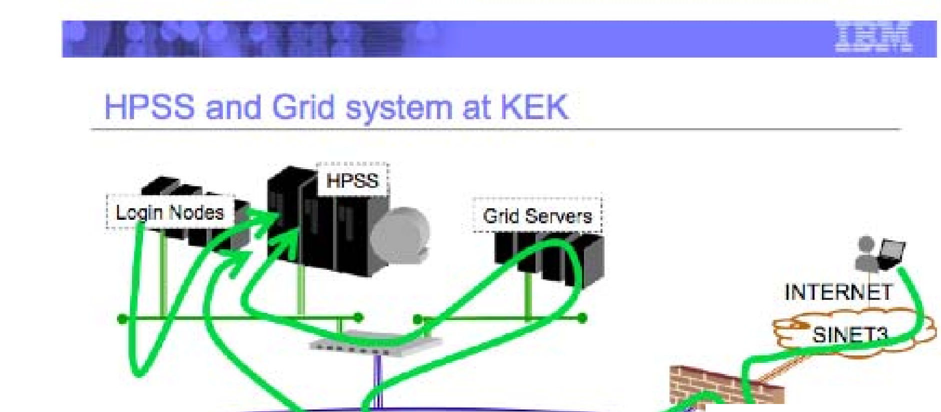
[182] **Finding a Practical Distributed Filesystem**
by Benjeman MEEKHOF
(2nd Conference Room: 11:15 - 11:30)

[124] **A data grid environment with HPSS and GridFTP at KEK**
by Satomi YAMAMOTO
(2nd Conference Room: 11:30 - 12:00)

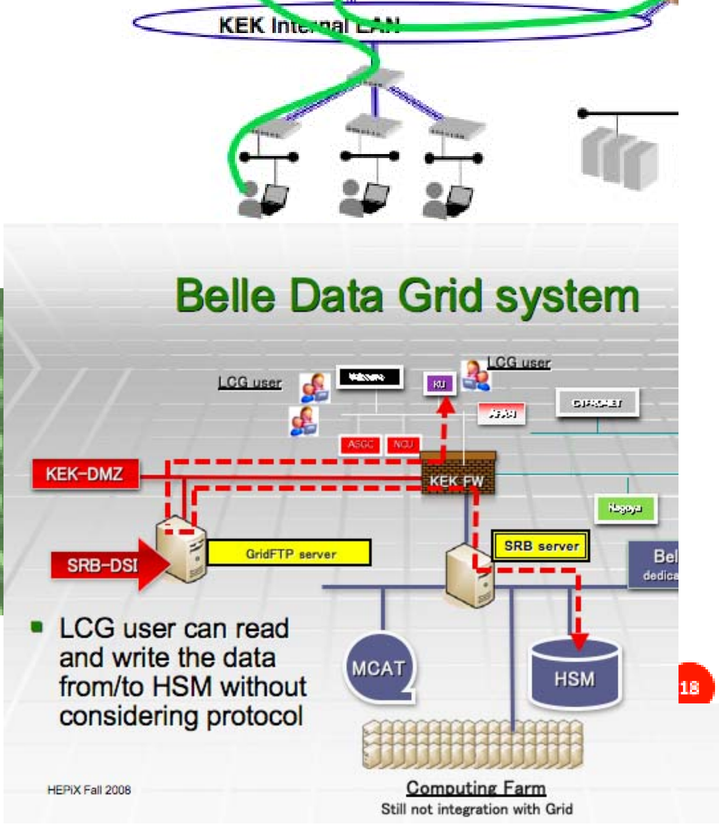
12:00

[125] **SRB system for Belle experiment**
by Yoshimi IIDA
(2nd Conference Room: 12:00 - 12:30)

Grid enabling



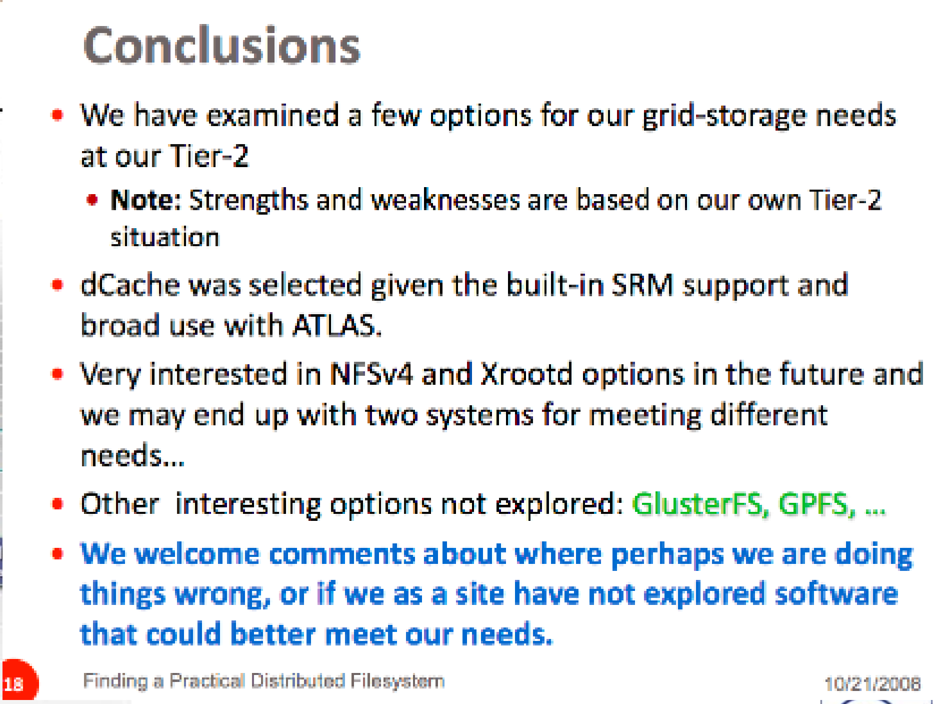
HPSS and Grid system at KEK



Belle Data Grid system

- LCG user can read and write the data from/to HSM without considering protocol

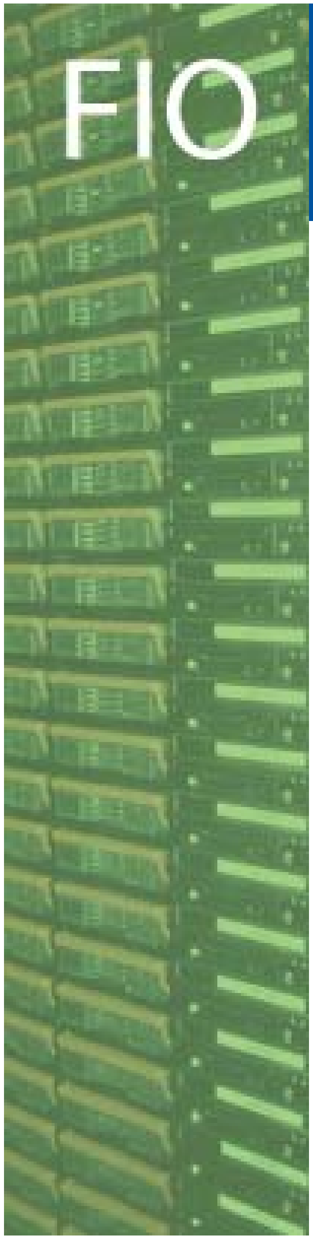
Uni Michigan tried AFS, NFS, dCache, Lustre and xrootd. Lustre may be revisited after important changes in 1.8 and 2.0



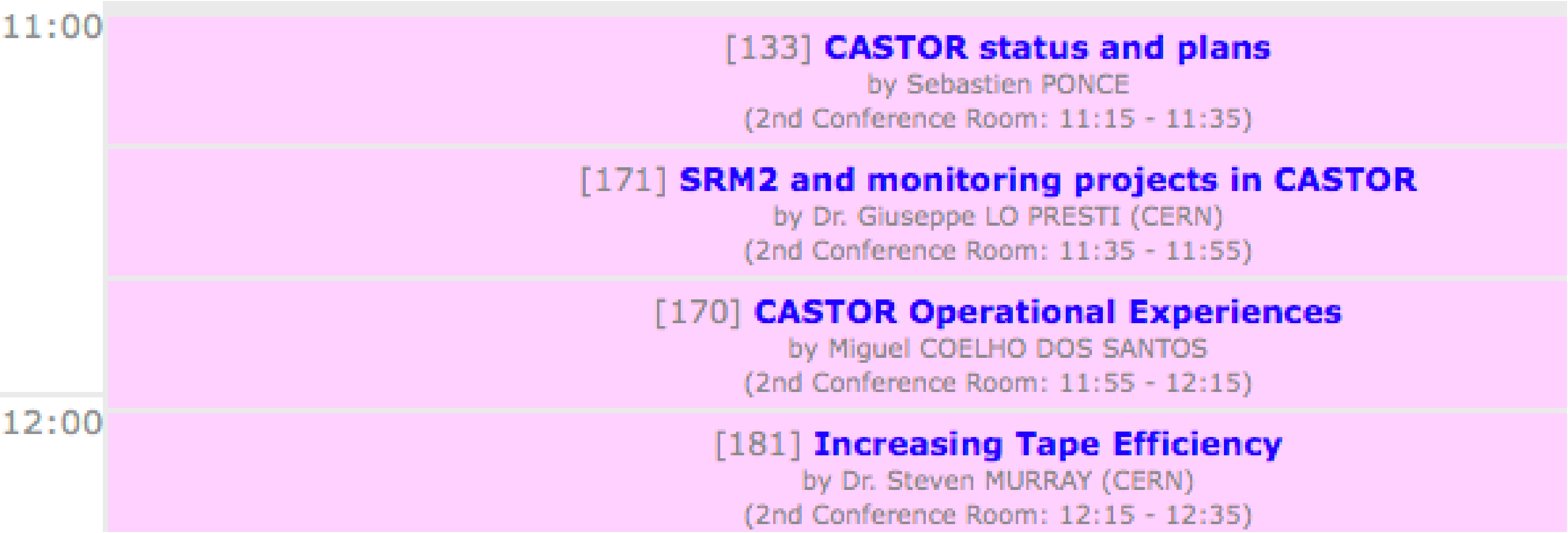
Conclusions

- We have examined a few options for our grid-storage needs at our Tier-2
  - **Note:** Strengths and weaknesses are based on our own Tier-2 situation
- dCache was selected given the built-in SRM support and broad use with ATLAS.
- Very interested in NFSv4 and Xrootd options in the future and we may end up with two systems for meeting different needs…
- Other interesting options not explored: GlusterFS, GPFS, …
- **We welcome comments about where perhaps we are doing things wrong, or if we as a site have not explored software that could better meet our needs.**

Finding a Practical Distributed Filesystem 10/21/2008



CASTOR track – day three



11:00

[133] **CASTOR status and plans**
by Sebastien PONCE
(2nd Conference Room: 11:15 - 11:35)

[171] **SRM2 and monitoring projects in CASTOR**
by Dr. Giuseppe LO PRESTI (CERN)
(2nd Conference Room: 11:35 - 11:55)

[170] **CASTOR Operational Experiences**
by Miguel COELHO DOS SANTOS
(2nd Conference Room: 11:55 - 12:15)

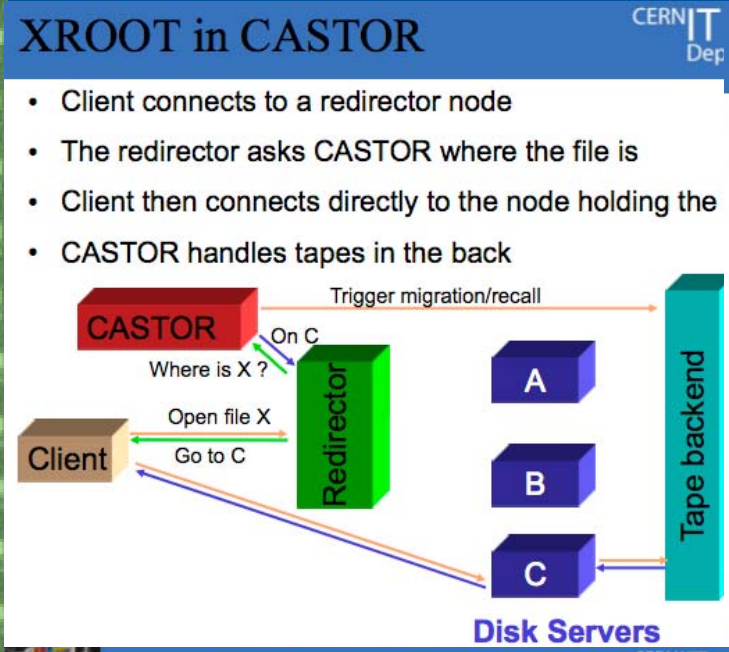
12:00

[181] **Increasing Tape Efficiency**
by Dr. Steven MURRAY (CERN)
(2nd Conference Room: 12:15 - 12:35)

FIO

CASTOR

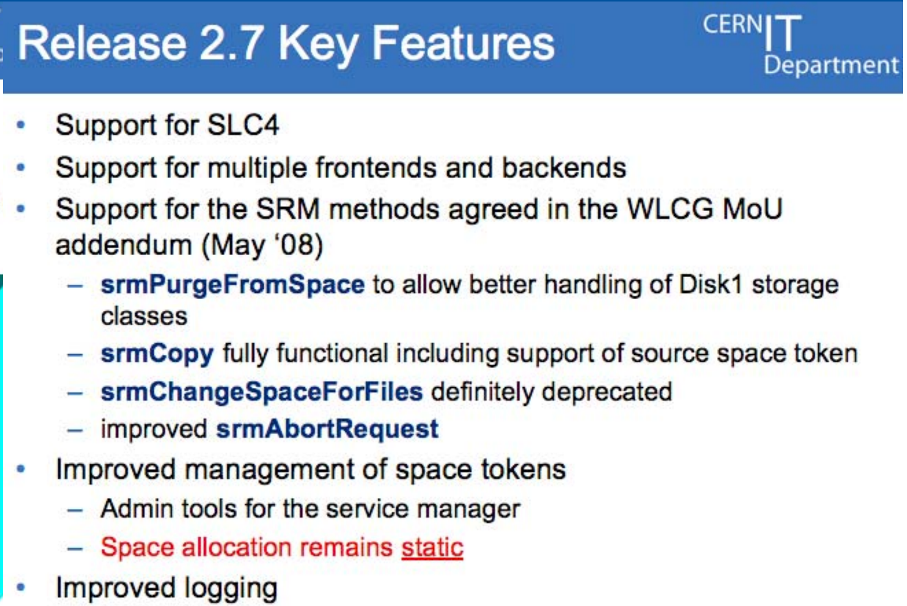
CERN IT Department



XROOT in CASTOR

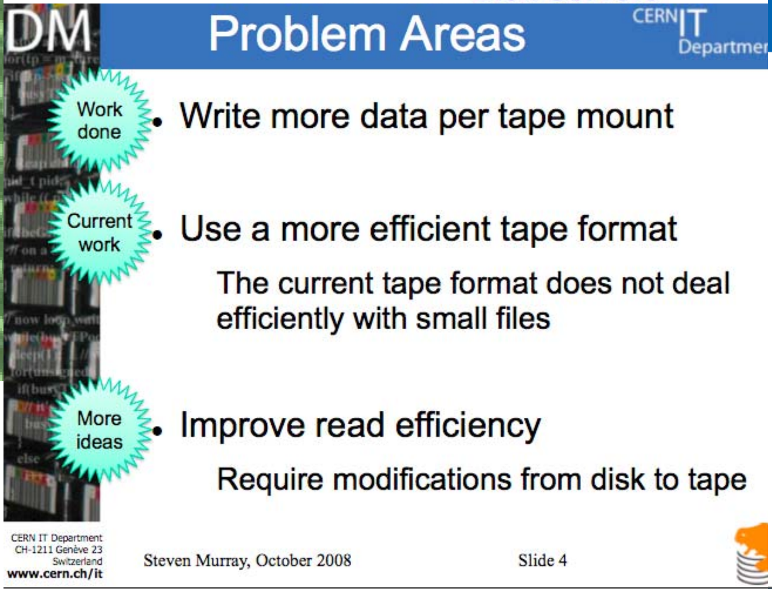
- Client connects to a redirector node
- The redirector asks CASTOR where the file is
- Client then connects directly to the node holding the
- CASTOR handles tapes in the back

CASTOR
Trigger migration/recall
On C
Where is X ?
Client
Open file X
Go to C
Redirector
A
B
C
Tape backend
Disk Servers



Release 2.7 Key Features

- Support for SLC4
- Support for multiple frontends and backends
- Support for the SRM methods agreed in the WLCG MoU addendum (May '08)
  - **srmPurgeFromSpace** to allow better handling of Disk1 storage classes
  - **srmCopy** fully functional including support of source space token
  - **srmChangeSpaceForFiles** definitely deprecated
  - improved **srmAbortRequest**
- Improved management of space tokens
  - Admin tools for the service manager
  - Space allocation remains static
- Improved logging



DM

Problem Areas

Work done
- Write more data per tape mount

Current work
- Use a more efficient tape format

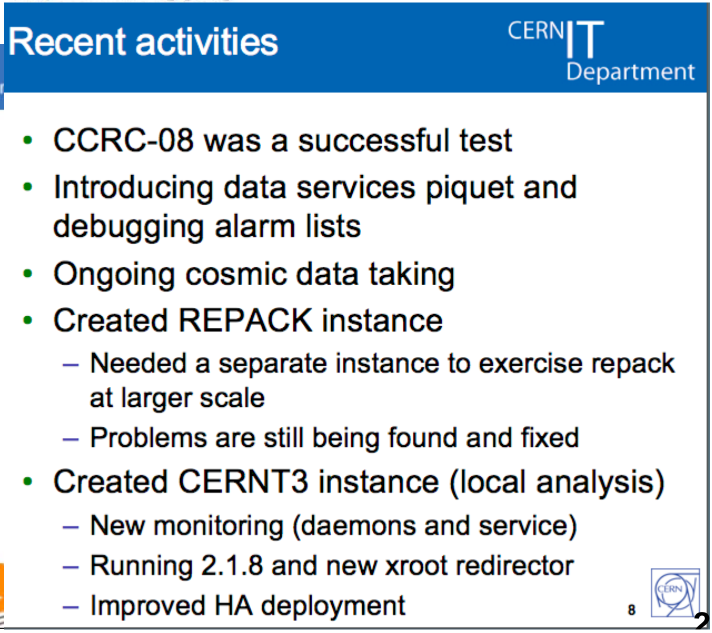
  The current tape format does not deal efficiently with small files

More ideas
- Improve read efficiency

  Require modifications from disk to tape

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

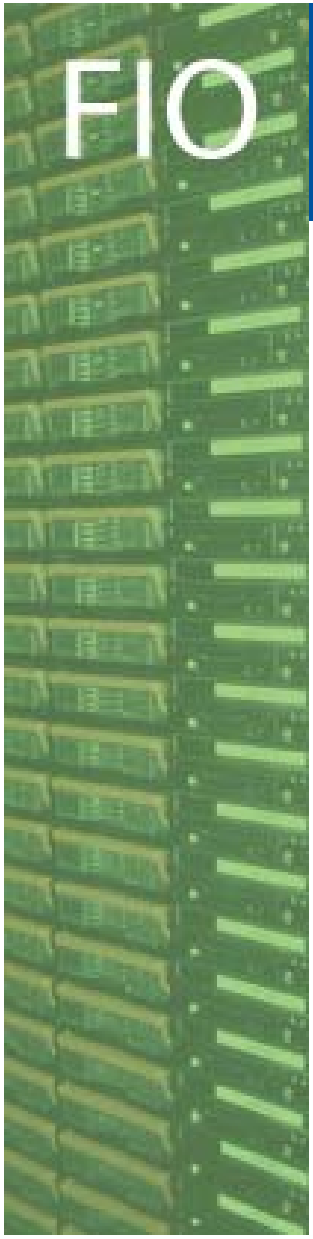
Steven Murray, October 2008 Slide 4



Recent activities

- CCRC-08 was a successful test
- Introducing data services piquet and debugging alarm lists
- Ongoing cosmic data taking
- Created REPACK instance
  - Needed a separate instance to exercise repack at larger scale
  - Problems are still being found and fixed
- Created CERNT3 instance (local analysis)
  - New monitoring (daemons and service)
  - Running 2.1.8 and new xroot redirector
  - Improved HA deployment

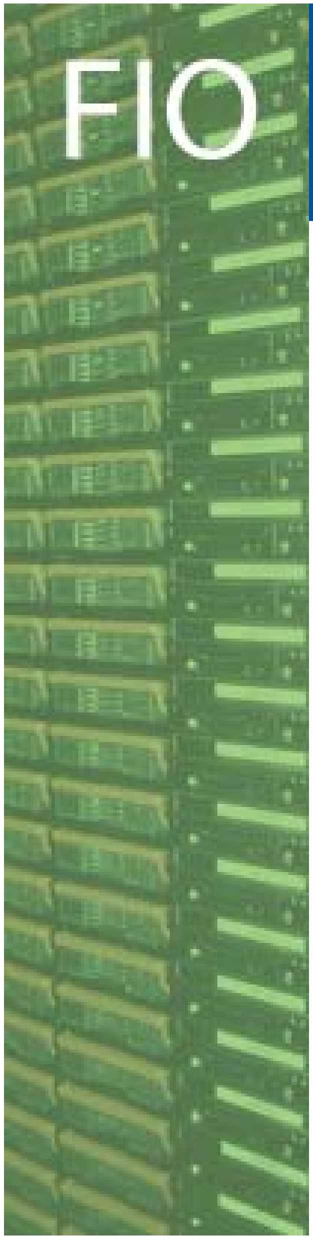
8



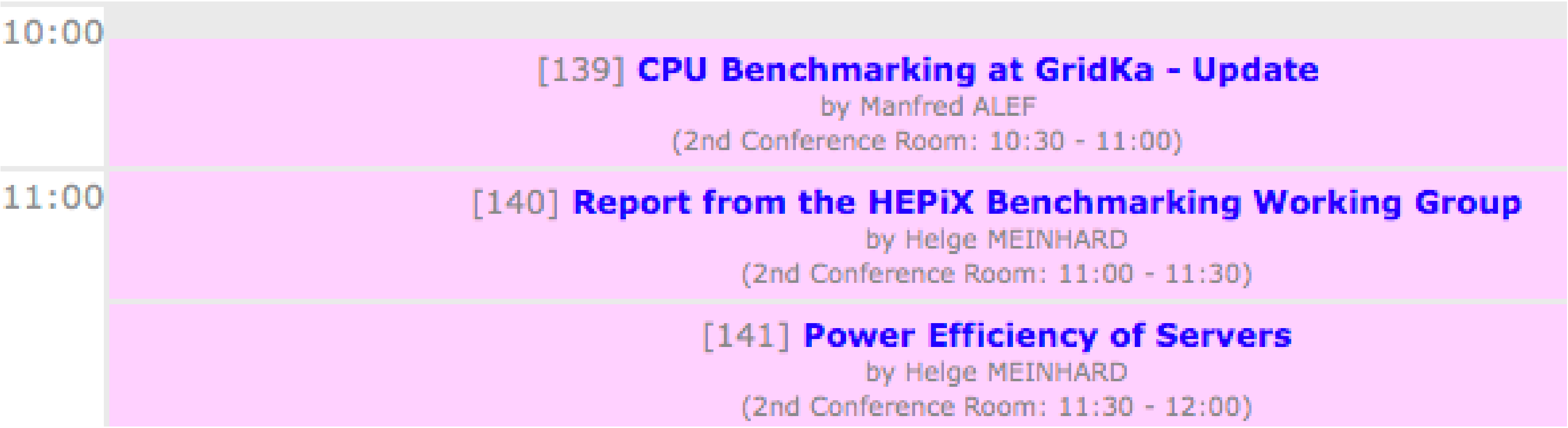
Storage Summary

CERN IT Department

- CASTOR
  - 5 Talks in total. Replay next week.
  - Analysis, xroot, scalability, SRM, tape performance, monitoring
- Lustre
  - 3 Talks
  - Used by 6 of top10 supercomputer sites, 40% of top100
  - Client cache and kerberos on the way
  - Showing good results, some important issues to be addressed (MDS scalability for example)
- Industry
  - 2+1 Talks (Data security & Performance + Lustre/SUN)
- Other
  - 4 Talks
  - Performance
  - Grid enabling (attaching gridftp servers to) existent solutions



Benchmarking Track – day 5



10:00

[139] **CPU Benchmarking at GridKa - Update**
by Manfred ALEF
(2nd Conference Room: 10:30 - 11:00)

11:00

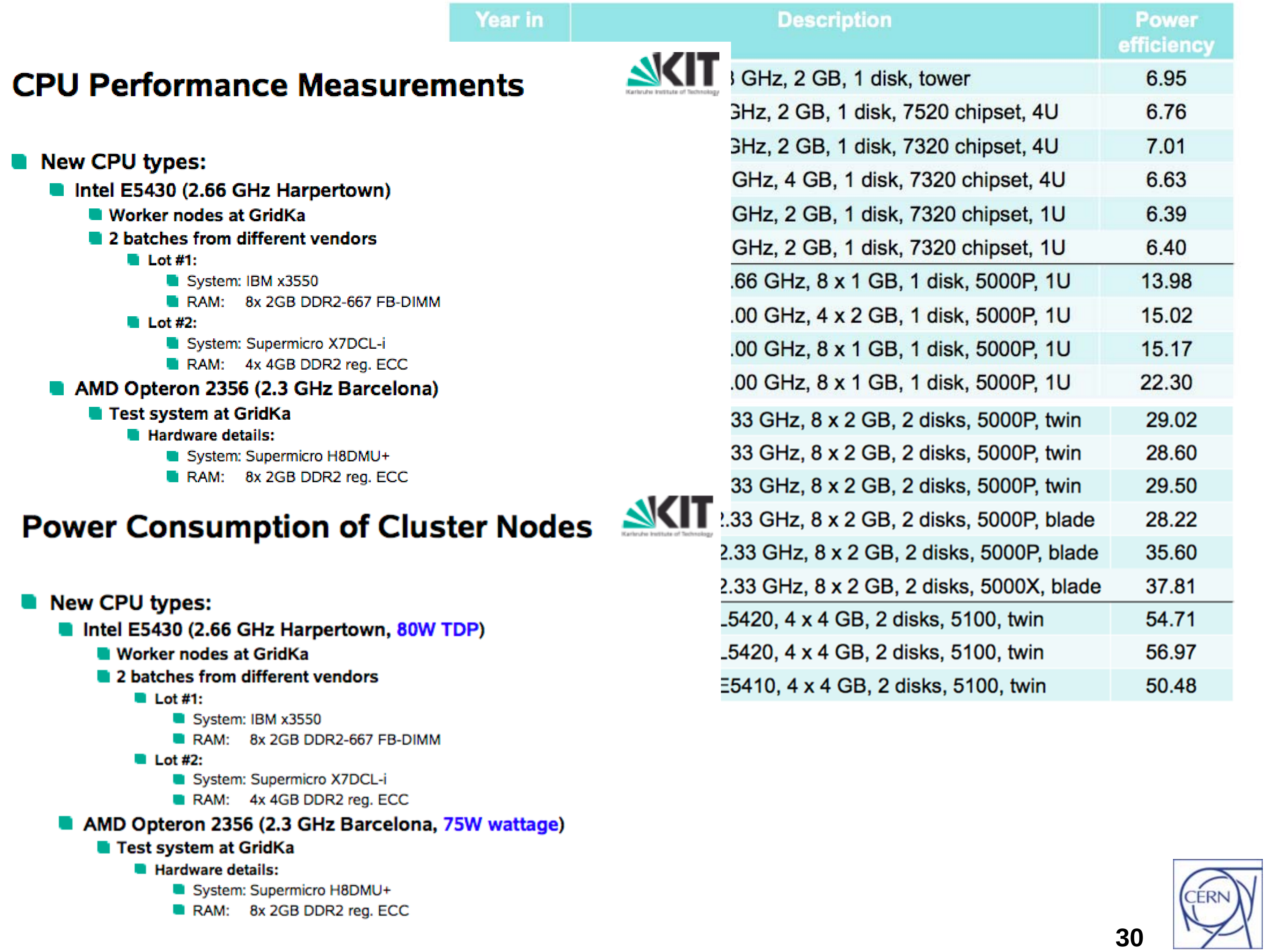
[140] **Report from the HEPiX Benchmarking Working Group**
by Helge MEINHARD
(2nd Conference Room: 11:00 - 11:30)

[141] **Power Efficiency of Servers**
by Helge MEINHARD
(2nd Conference Room: 11:30 - 12:00)

FIO

Benchmarking and power efficiency

CERN IT Department



CPU Performance Measurements

- New CPU types:
  - Intel E5430 (2.66 GHz Harpertown)
    - Worker nodes at GridKa
    - 2 batches from different vendors
      - Lot #1:
        - System: IBM x3550
        - RAM: 8x 2GB DDR2-667 FB-DIMM
      - Lot #2:
        - System: Supermicro X7DCL-i
        - RAM: 4x 4GB DDR2 reg. ECC
  - AMD Opteron 2356 (2.3 GHz Barcelona)
    - Test system at GridKa
      - Hardware details:
        - System: Supermicro H8DMU+
        - RAM: 8x 2GB DDR2 reg. ECC

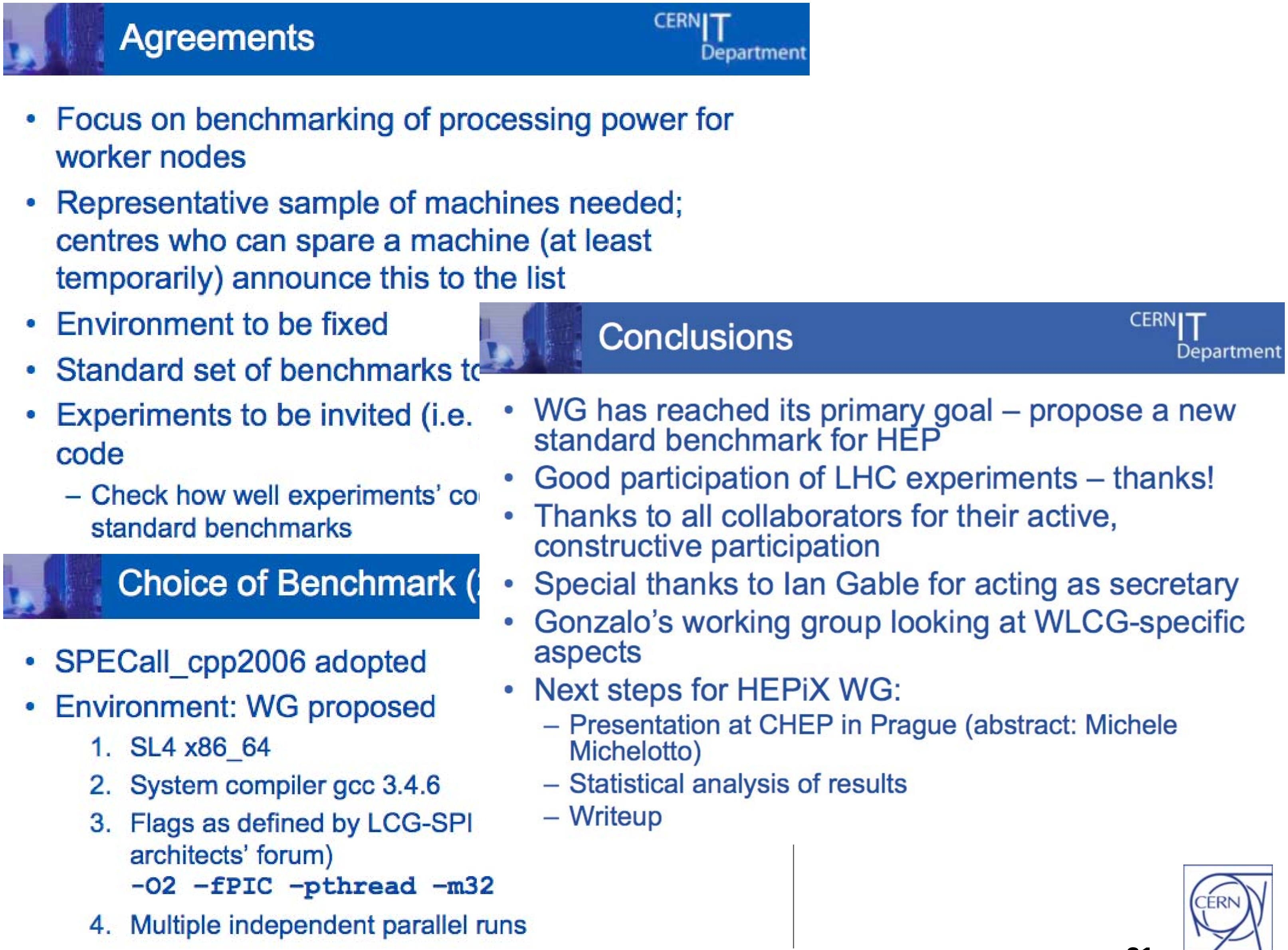
Power Consumption of Cluster Nodes

- New CPU types:
  - Intel E5430 (2.66 GHz Harpertown, 80W TDP)
    - Worker nodes at GridKa
    - 2 batches from different vendors
      - Lot #1:
        - System: IBM x3550
        - RAM: 8x 2GB DDR2-667 FB-DIMM
      - Lot #2:
        - System: Supermicro X7DCL-i
        - RAM: 4x 4GB DDR2 reg. ECC
  - AMD Opteron 2356 (2.3 GHz Barcelona, 75W wattage)
    - Test system at GridKa
      - Hardware details:
        - System: Supermicro H8DMU+
        - RAM: 8x 2GB DDR2 reg. ECC

Year in	Description	Power efficiency
	3 GHz, 2 GB, 1 disk, tower	6.95
	GHz, 2 GB, 1 disk, 7520 chipset, 4U	6.76
	GHz, 2 GB, 1 disk, 7320 chipset, 4U	7.01
	GHz, 4 GB, 1 disk, 7320 chipset, 4U	6.63
	GHz, 2 GB, 1 disk, 7320 chipset, 1U	6.39
	GHz, 2 GB, 1 disk, 7320 chipset, 1U	6.40
	.66 GHz, 8 x 1 GB, 1 disk, 5000P, 1U	13.98
	.00 GHz, 4 x 2 GB, 1 disk, 5000P, 1U	15.02
	.00 GHz, 8 x 1 GB, 1 disk, 5000P, 1U	15.17
	.00 GHz, 8 x 1 GB, 1 disk, 5000P, 1U	22.30
	33 GHz, 8 x 2 GB, 2 disks, 5000P, twin	29.02
	33 GHz, 8 x 2 GB, 2 disks, 5000P, twin	28.60
	33 GHz, 8 x 2 GB, 2 disks, 5000P, twin	29.50
	.33 GHz, 8 x 2 GB, 2 disks, 5000P, blade	28.22
	2.33 GHz, 8 x 2 GB, 2 disks, 5000P, blade	35.60
	2.33 GHz, 8 x 2 GB, 2 disks, 5000X, blade	37.81
	5420, 4 x 4 GB, 2 disks, 5100, twin	54.71
	5420, 4 x 4 GB, 2 disks, 5100, twin	56.97
	E5410, 4 x 4 GB, 2 disks, 5100, twin	50.48

FIO

Benchmarking WG



Agreements

- Focus on benchmarking of processing power for worker nodes
- Representative sample of machines needed; centres who can spare a machine (at least temporarily) announce this to the list
- Environment to be fixed
- Standard set of benchmarks t
- Experiments to be invited (i.e. code
  - Check how well experiments' co standard benchmarks

Choice of Benchmark (

- SPECall_cpp2006 adopted
- Environment: WG proposed
  1. SL4 x86_64
  2. System compiler gcc 3.4.6
  3. Flags as defined by LCG-SPI architects' forum)
     `-O2 -fPIC -pthread -m32`
  4. Multiple independent parallel runs

Conclusions

- WG has reached its primary goal – propose a new standard benchmark for HEP
- Good participation of LHC experiments – thanks!
- Thanks to all collaborators for their active, constructive participation
- Special thanks to Ian Gable for acting as secretary
- Gonzalo's working group looking at WLCG-specific aspects
- Next steps for HEPiX WG:
  - Presentation at CHEP in Prague (abstract: Michele Michelotto)
  - Statistical analysis of results
  - Writeup

Report on HEPiX Fall 2008 Taipei

Operating Systems and Applications

Virtualization

Network and Security

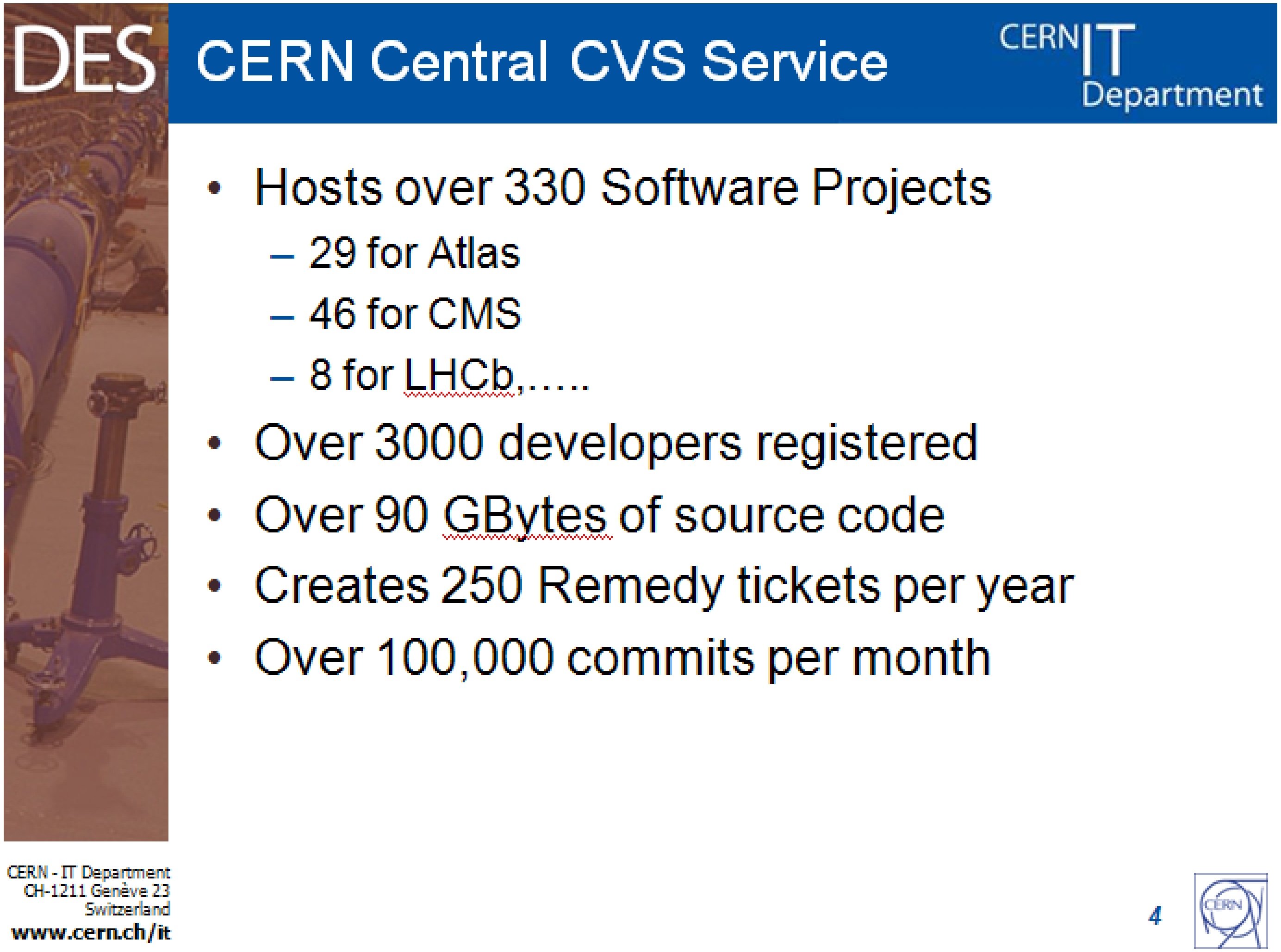
Miscellaneous

Operating Systems & Applications

CERN IT Department

- **Subversion(SVN) 2009**
- **Computation interface to NVidia graphics card**
- **Live CD and NTFS(read/write)**
- **Smooth transition to Exchange 2007**
- **Used by Atlas since their first data challenge**

CVS Replacement: Getting Subversive

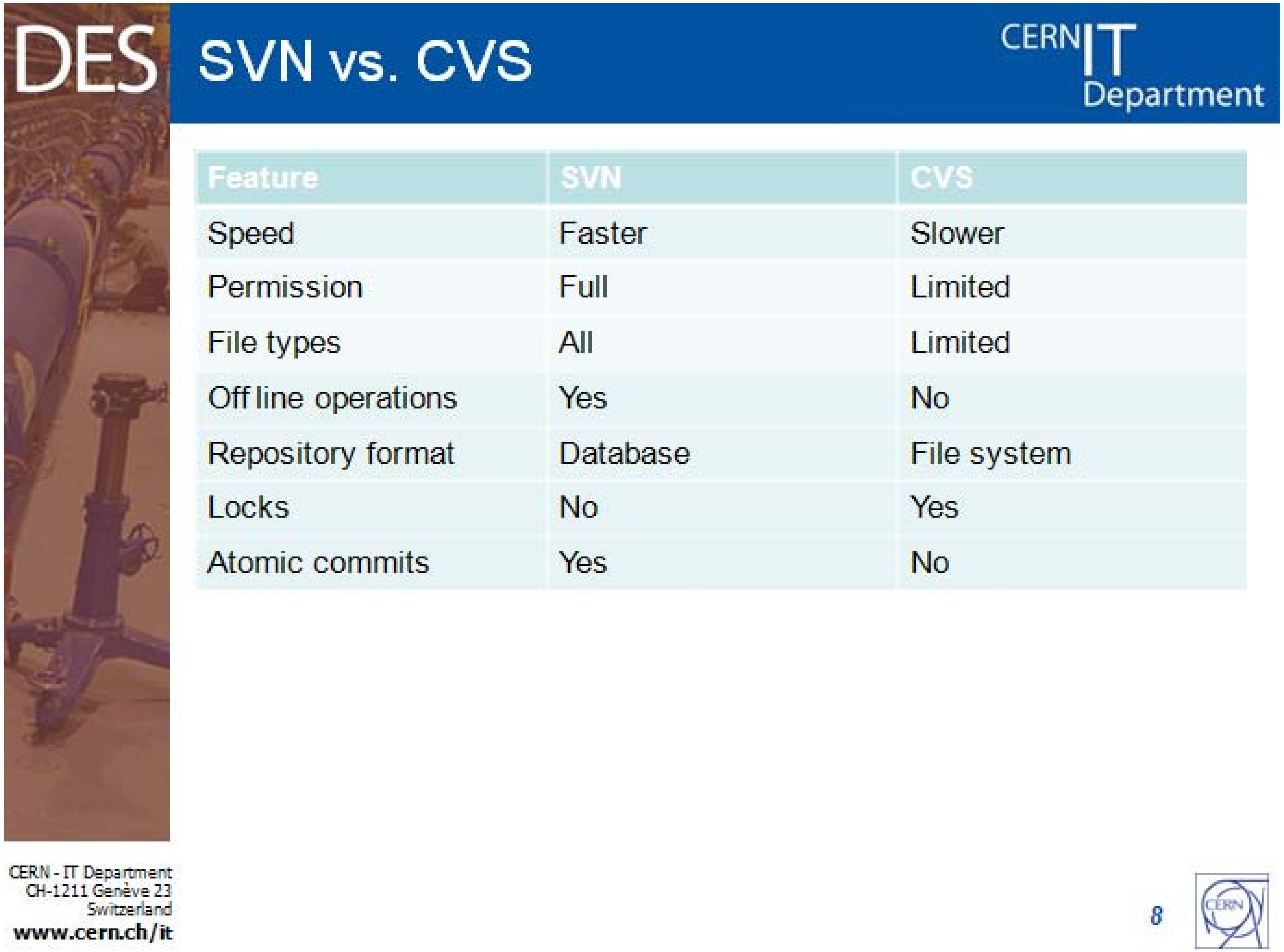


CERN Central CVS Service

- Hosts over 330 Software Projects
  - 29 for Atlas
  - 46 for CMS
  - 8 for LHCb,…..
- Over 3000 developers registered
- Over 90 GBytes of source code
- Creates 250 Remedy tickets per year
- Over 100,000 commits per month

CVS Replacement: Getting Subversive

CERN IT Department



DES

SVN vs. CVS

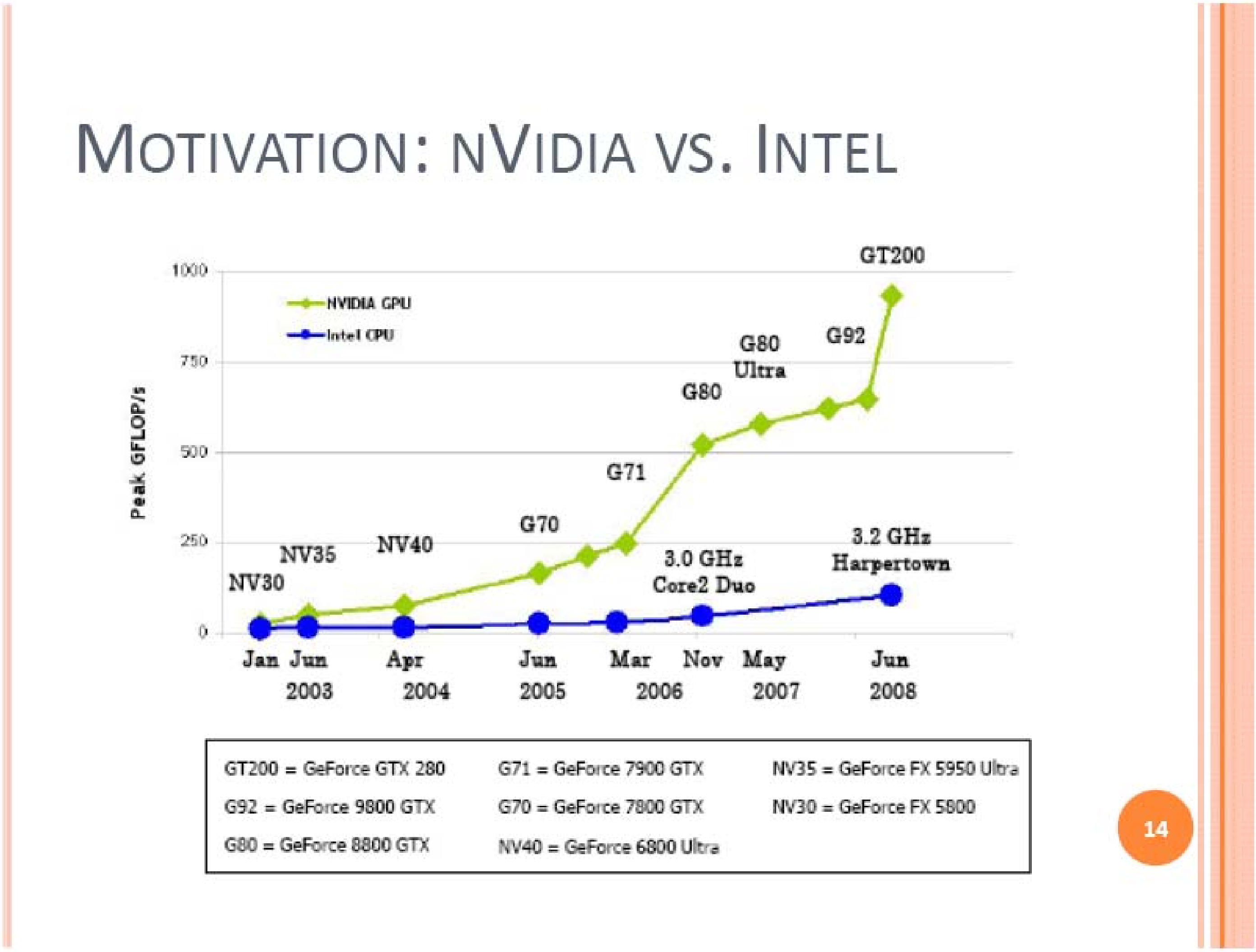
CERN IT Department

Feature	SVN	CVS
Speed	Faster	Slower
Permission	Full	Limited
File types	All	Limited
Off line operations	Yes	No
Repository format	Database	File system
Locks	No	Yes
Atomic commits	Yes	No

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

8

High Performance Cryptographic computing



Motivation: nVidia vs. Intel

GT200 = GeForce GTX 280	G71 = GeForce 7900 GTX	NV35 = GeForce FX 5950 Ultra
G92 = GeForce 9800 GTX	G70 = GeForce 7800 GTX	NV30 = GeForce FX 5800
G80 = GeForce 8800 GTX	NV40 = GeForce 6800 Ultra	

14

Scientific Linux Status Report

CERN IT Department

Scientific Linux

What we see in the future

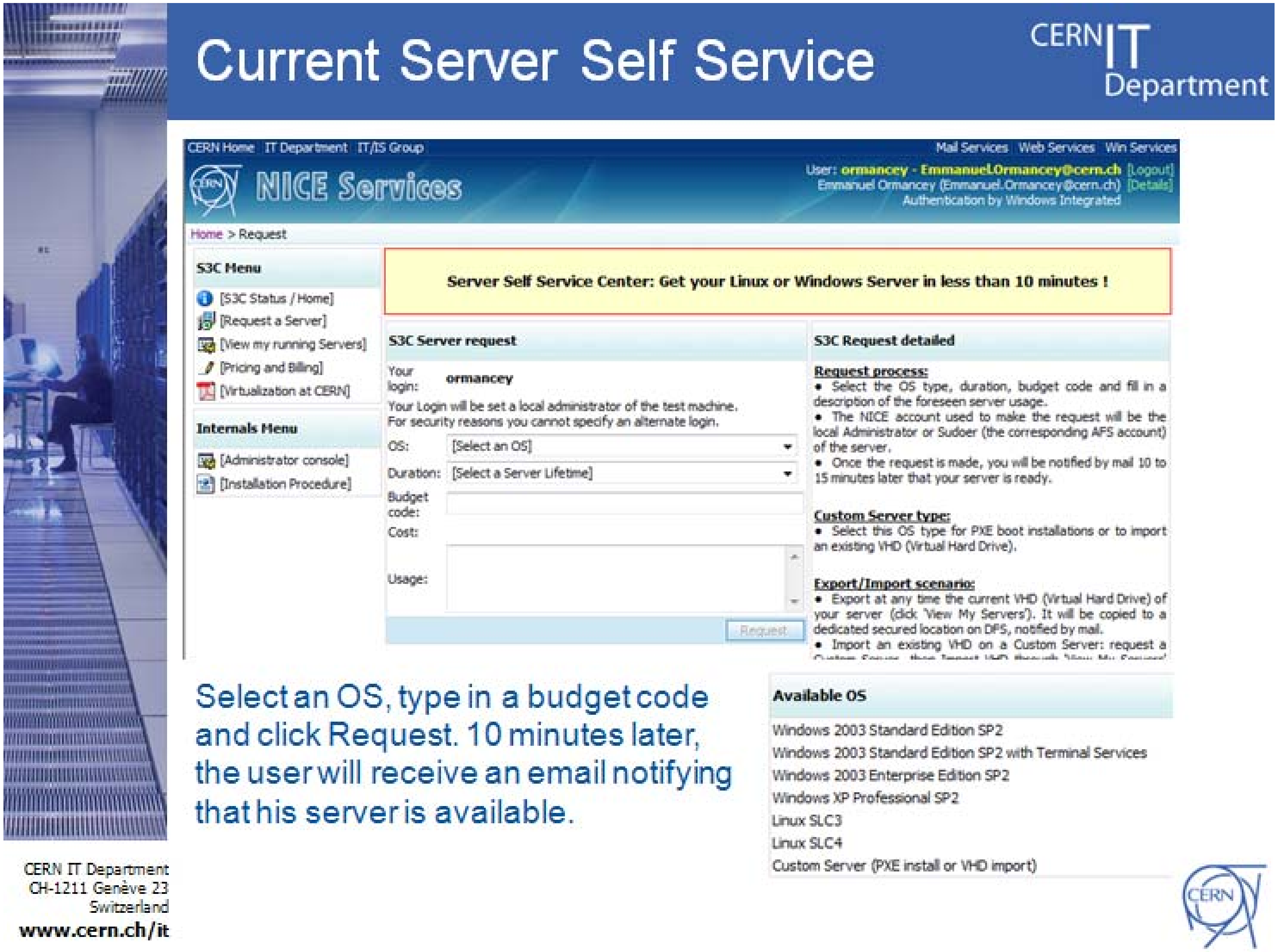
- When RHEL 5 Update 3 comes out, releasing S.L. 5.3
  - Estimate - March 2009
- When RHEL 4 Update 8 comes out, releasing S.L. 4.8
  - Estimate - June 2009
- When RHEL 6 comes out, releasing S.L. 6.0
  - Estimate - January 2010

Virtualization

- **Self service virtual machine in 10 minutes**
- **chroot-based VM still going strong since 2004**
- **Platform Enterprise Grid Orchestrator integrates virtualisation and the Grid**
- **Hyper-V and XenCenter can work together**
- **CISCO and F5 solutions for network virtual network services**

Virtualization with Hyper-V at CERN

CERN IT Department



Current Server Self Service

CERN IT Department

CERN Home IT Department IT/IS Group Mail Services Web Services Win Services

NICE Services

User: ormancey - Emmanuel.Ormancey@cern.ch [Logout]
Emmanuel Ormancey (Emmanuel.Ormancey@cern.ch) [Details]
Authentication by Windows Integrated

Home > Request

S3C Menu

- [S3C Status / Home]
- [Request a Server]
- [View my running Servers]
- [Pricing and Billing]
- [Virtualization at CERN]

Internals Menu

- [Administrator console]
- [Installation Procedure]

Server Self Service Center: Get your Linux or Windows Server in less than 10 minutes !

S3C Server request

Your login: **ormancey**

Your Login will be set a local administrator of the test machine.
For security reasons you cannot specify an alternate login.

OS: [Select an OS]

Duration: [Select a Server Lifetime]

Budget code:

Cost:

Usage:

Request

S3C Request detailed

Request process:

- Select the OS type, duration, budget code and fill in a description of the foreseen server usage.
- The NICE account used to make the request will be the local Administrator or Sudoer (the corresponding AFS account) of the server.
- Once the request is made, you will be notified by mail 10 to 15 minutes later that your server is ready.

Custom Server type:

- Select this OS type for PXE boot installations or to import an existing VHD (Virtual Hard Drive).

Export/Import scenario:

- Export at any time the current VHD (Virtual Hard Drive) of your server (click 'View My Servers'). It will be copied to a dedicated secured location on DFS, notified by mail.
- Import an existing VHD on a Custom Server: request a Custom Server, then Import VHD through 'View My Servers'

Select an OS, type in a budget code and click Request. 10 minutes later, the user will receive an email notifying that his server is available.

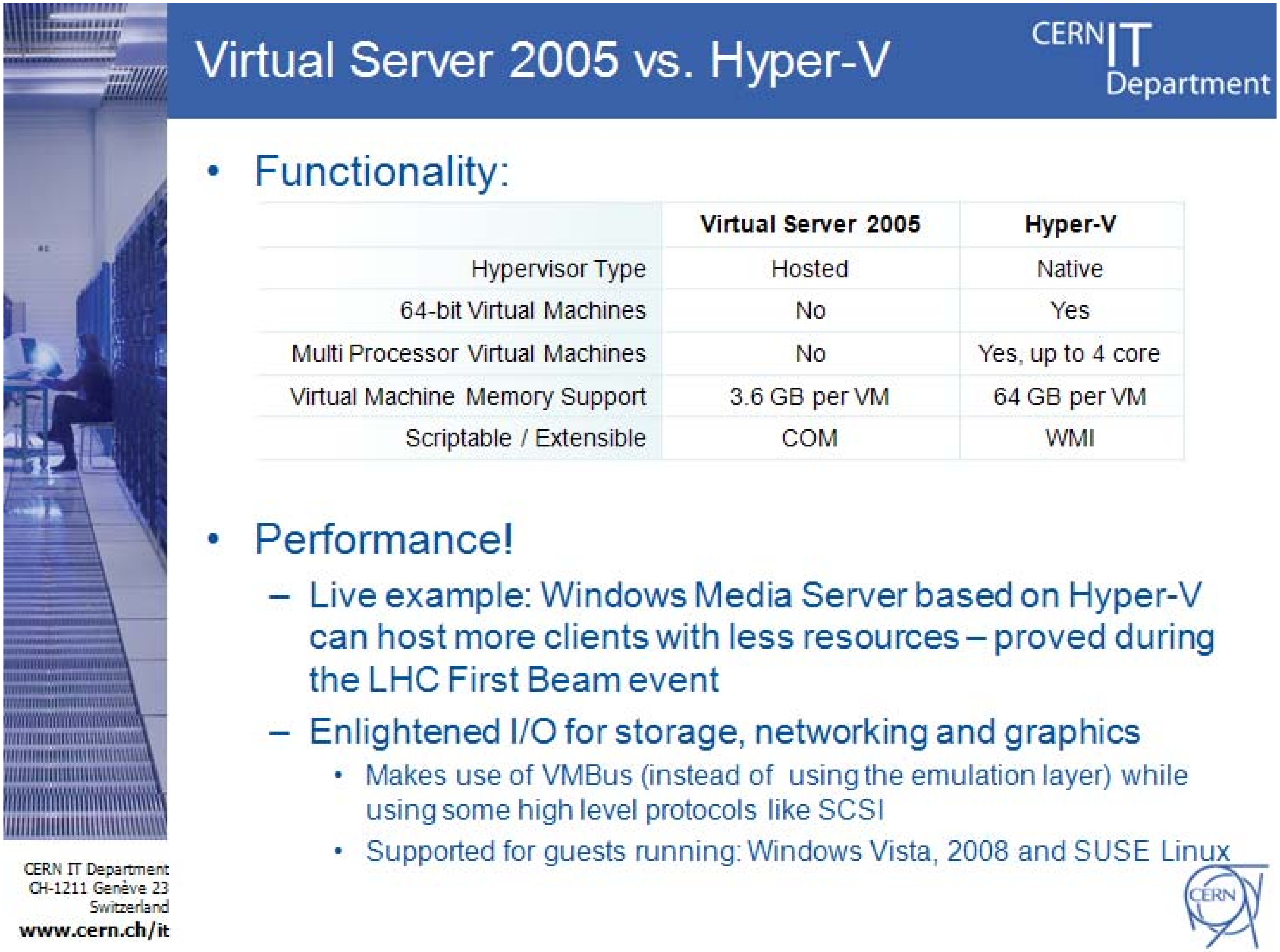
Available OS

Windows 2003 Standard Edition SP2
Windows 2003 Standard Edition SP2 with Terminal Services
Windows 2003 Enterprise Edition SP2
Windows XP Professional SP2
Linux SLC3
Linux SLC4
Custom Server (PXE install or VHD import)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Virtualization with Hyper-V at CERN

CERN IT Department



Virtual Server 2005 vs. Hyper-V

CERN IT Department

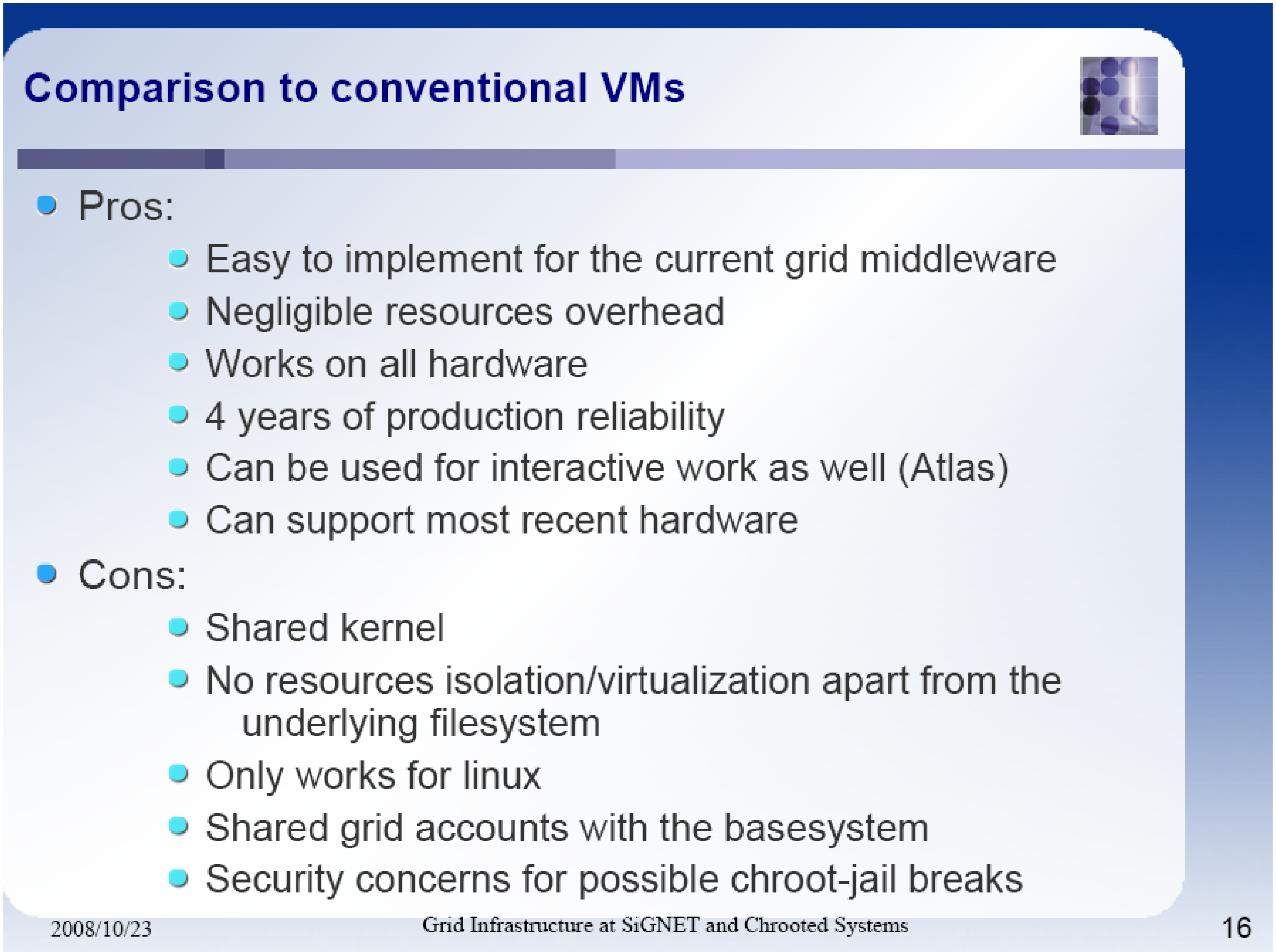
- Functionality:

	Virtual Server 2005	Hyper-V
Hypervisor Type	Hosted	Native
64-bit Virtual Machines	No	Yes
Multi Processor Virtual Machines	No	Yes, up to 4 core
Virtual Machine Memory Support	3.6 GB per VM	64 GB per VM
Scriptable / Extensible	COM	WMI

- Performance!
  - Live example: Windows Media Server based on Hyper-V can host more clients with less resources – proved during the LHC First Beam event
  - Enlightened I/O for storage, networking and graphics
    - Makes use of VMBus (instead of using the emulation layer) while using some high level protocols like SCSI
    - Supported for guests running: Windows Vista, 2008 and SUSE Linux

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Grid Infrastructure at SiGNET and Chrooted Systems



Comparison to conventional VMs

- Pros:
  - Easy to implement for the current grid middleware
  - Negligible resources overhead
  - Works on all hardware
  - 4 years of production reliability
  - Can be used for interactive work as well (Atlas)
  - Can support most recent hardware
- Cons:
  - Shared kernel
  - No resources isolation/virtualization apart from the underlying filesystem
  - Only works for linux
  - Shared grid accounts with the basesystem
  - Security concerns for possible chroot-jail breaks

Networking & Security

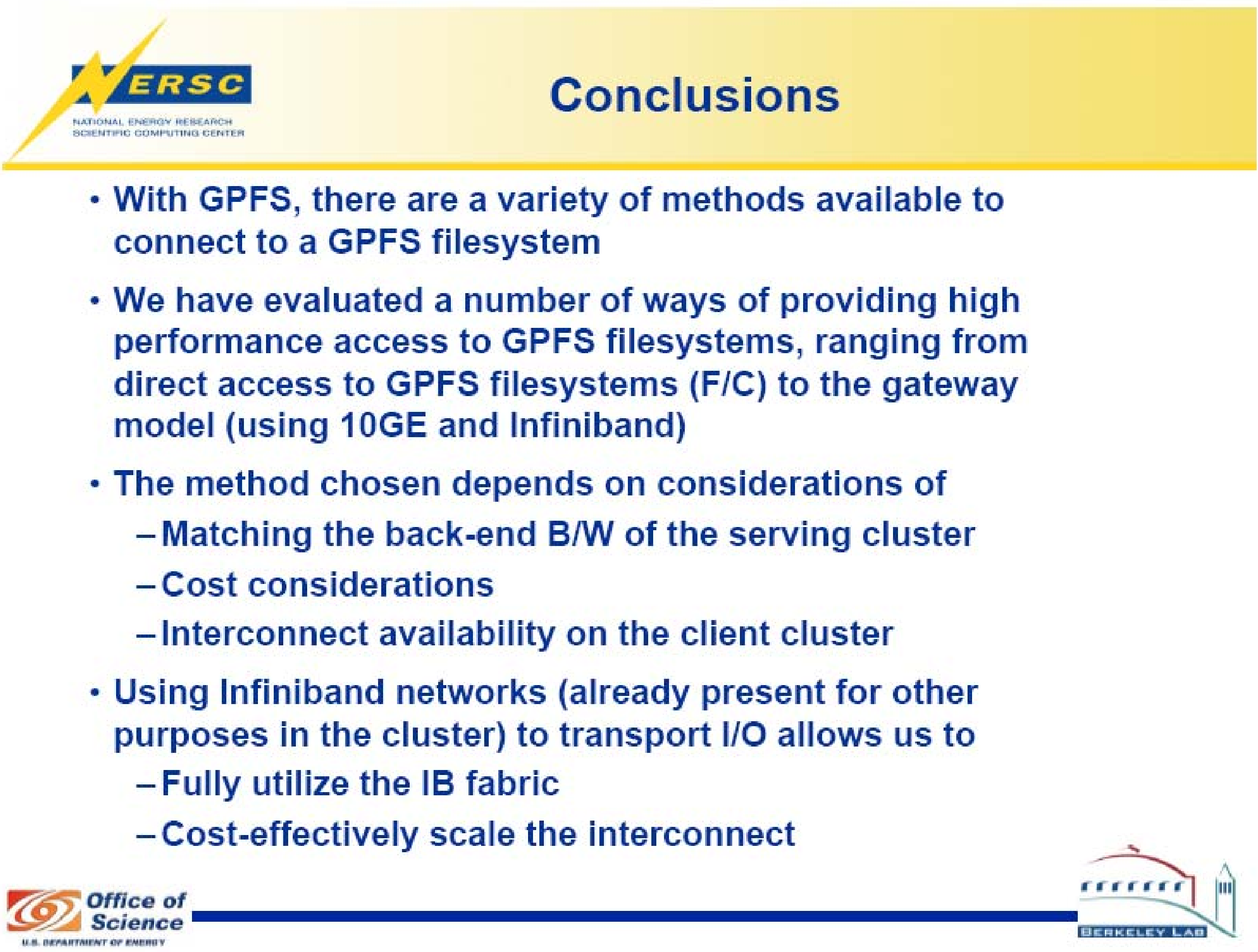
- 2 CAs, 1 RA and smart cards
- DBPowder will become open source
- Codian MCU will in production February 2009
- Good progress during last year
- Only one solution is long-term – the transition
- Cyber space isn't getting any safer

Miscellaneous

CERN IT Department

- VO communities detect problems first
- RT is good for out-of-the-box usage
- Tier-2 praguelcg2 will use cfengine everywhere
- GPFS usage recommendations made
- Standards are the only long-term solution

Evaluation of GPFS Connectivity over High-performance Network



Conclusions

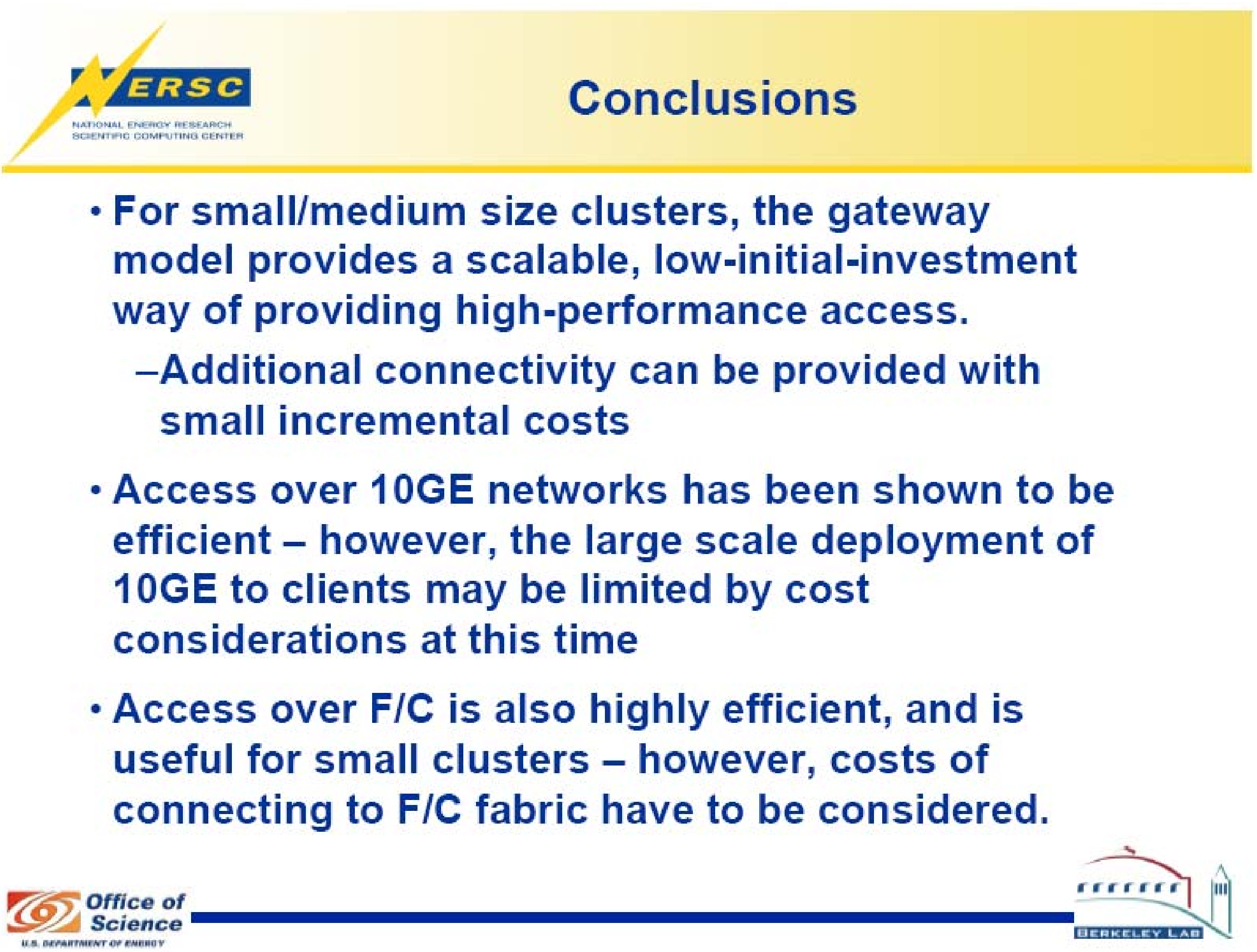
- With GPFS, there are a variety of methods available to connect to a GPFS filesystem
- We have evaluated a number of ways of providing high performance access to GPFS filesystems, ranging from direct access to GPFS filesystems (F/C) to the gateway model (using 10GE and Infiniband)
- The method chosen depends on considerations of
  - Matching the back-end B/W of the serving cluster
  - Cost considerations
  - Interconnect availability on the client cluster
- Using Infiniband networks (already present for other purposes in the cluster) to transport I/O allows us to
  - Fully utilize the IB fabric
  - Cost-effectively scale the interconnect

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER
Office of Science
U.S. DEPARTMENT OF ENERGY
BERKELEY LAB

DM

Evaluation of GPFS Connectivity over High-performance Network

CERN IT Department



Conclusions

- For small/medium size clusters, the gateway model provides a scalable, low-initial-investment way of providing high-performance access.
  - –Additional connectivity can be provided with small incremental costs
- Access over 10GE networks has been shown to be efficient – however, the large scale deployment of 10GE to clients may be limited by cost considerations at this time
- Access over F/C is also highly efficient, and is useful for small clusters – however, costs of connecting to F/C fabric have to be considered.

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER
Office of Science
U.S. DEPARTMENT OF ENERGY
BERKELEY LAB