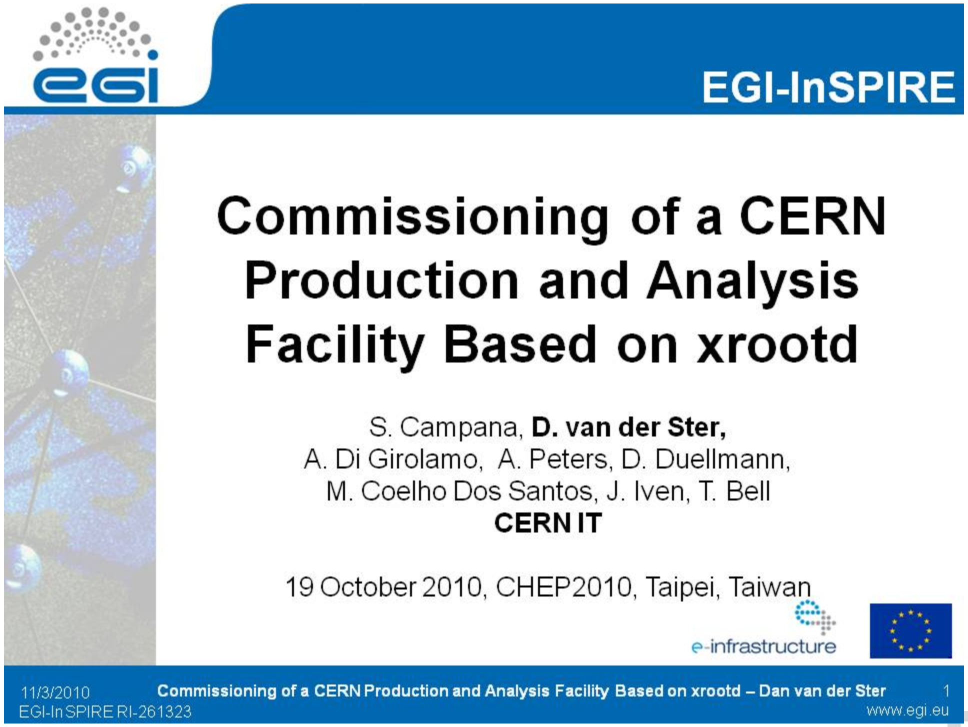
Atlas Tier 3

Doug Benjamin
Duke University
On Behalf of the Atlas Collaboration

Atlas XROOTD Demonstrator project

- Last June at WLGC Storage workshop
  - Atlas Tier 3 proposed alternative method for delivering data to Tier 3 using confederated XROOTD clusters
- Physicists can get the data that they actually use
- Alternative and simpler than ATLAS DDM
  - In testing now
  - Plan to connect Tier 3 sites and some Tier 2 sites
- CMS working on something similar (Their focus is between Tier 1/Tier 2 – complimentary – we are collaborating)



EGI-InSPIRE

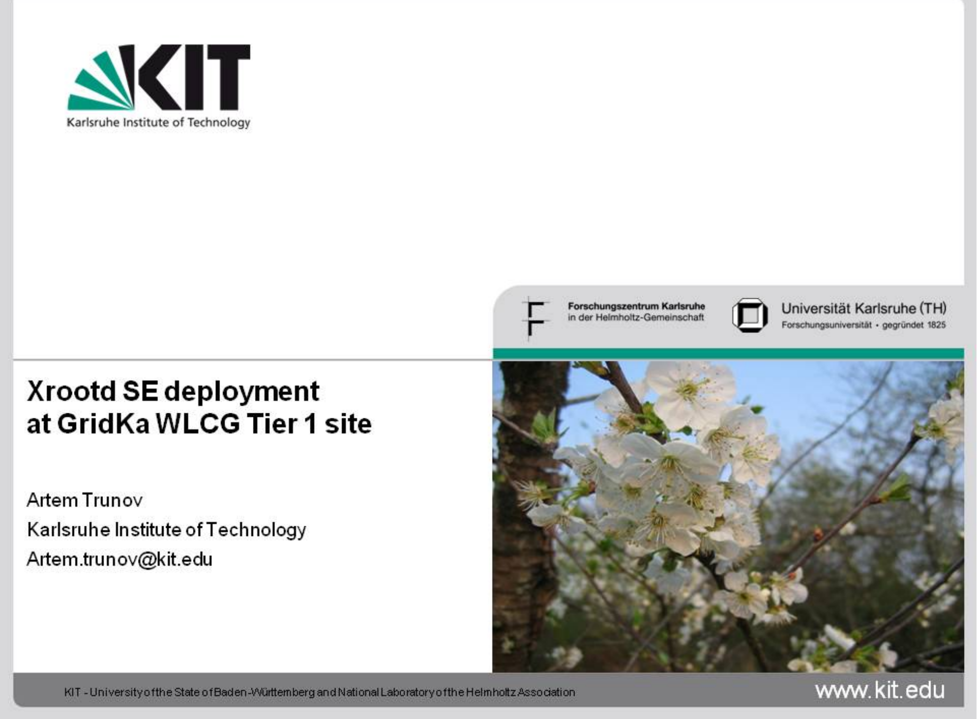
Commissioning of a CERN Production and Analysis Facility Based on xrootd

S. Campana, **D. van der Ster,**
A. Di Girolamo, A. Peters, D. Duellmann,
M. Coelho Dos Santos, J. Iven, T. Bell
CERN IT

19 October 2010, CHEP2010, Taipei, Taiwan

e-infrastructure

11/3/2010 **Commissioning of a CERN Production and Analysis Facility Based on xrootd – Dan van der Ster** 1
EGI-InSPIRE RI-261323 www.egi.eu



KIT Karlsruhe Institute of Technology

Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH) Forschungsuniversität • gegründet 1825

Xrootd SE deployment at GridKa WLCG Tier 1 site

Artem Trunov
Karlsruhe Institute of Technology
Artem.trunov@kit.edu

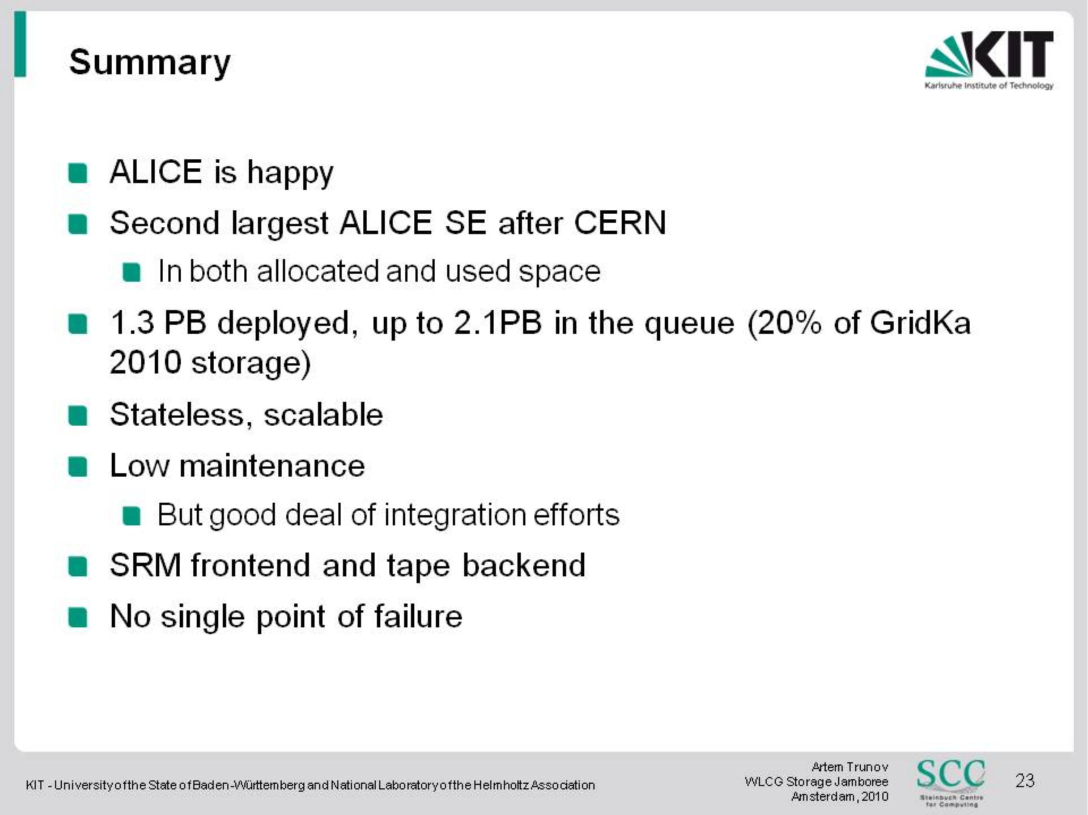
KIT - University of the State of Baden-Württemberg and National Laboratory of the Helmholtz Association

www.kit.edu

Conclusions

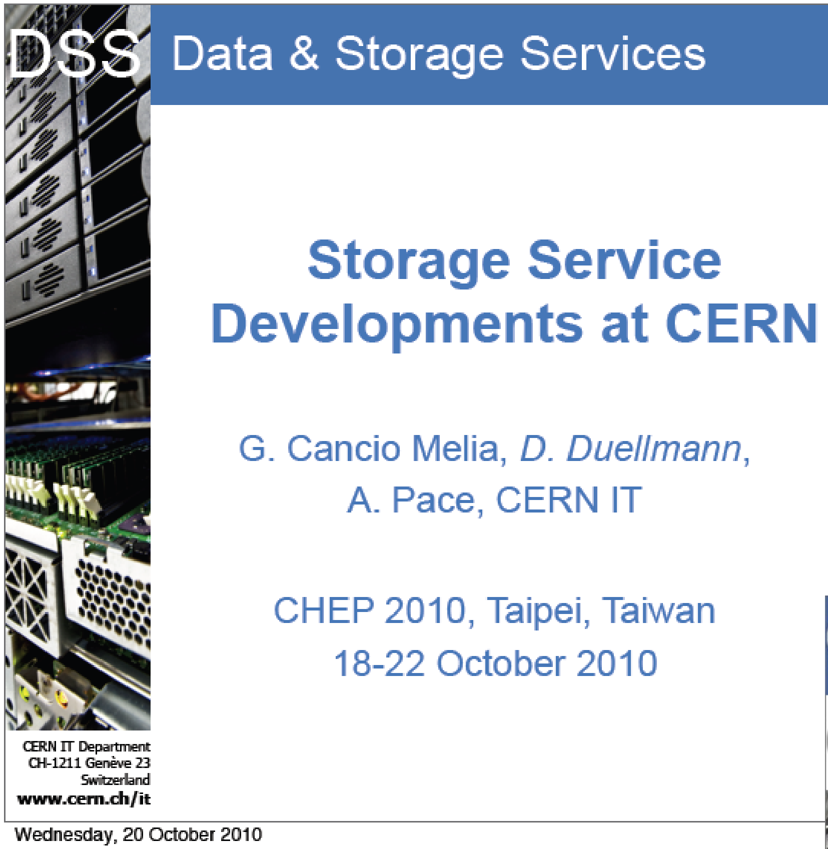
- CERN has been successfully handling its Tier 0 and CAF duties for many years
- Adapting to the chaotic access for data analysis has been a challenge.
- CERN IT and ATLAS have invested dedicated effort to commission the CERN grid analysis facility.
- The current setup can successfully cope with ATLAS grid analysis
  - And it is flexibly accommodating other use cases such as data reprocessing and local access.
- More improvements are expected from the new EOS prototype.

Commissioning of a CERN Production and Analysis Facility Based on xroo



Summary

- ALICE is happy
- Second largest ALICE SE after CERN
  - In both allocated and used space
- 1.3 PB deployed, up to 2.1PB in the queue (20% of GridKa 2010 storage)
- Stateless, scalable
- Low maintenance
  - But good deal of integration efforts
- SRM frontend and tape backend
- No single point of failure



DSS Data & Storage Services

CERN IT Department

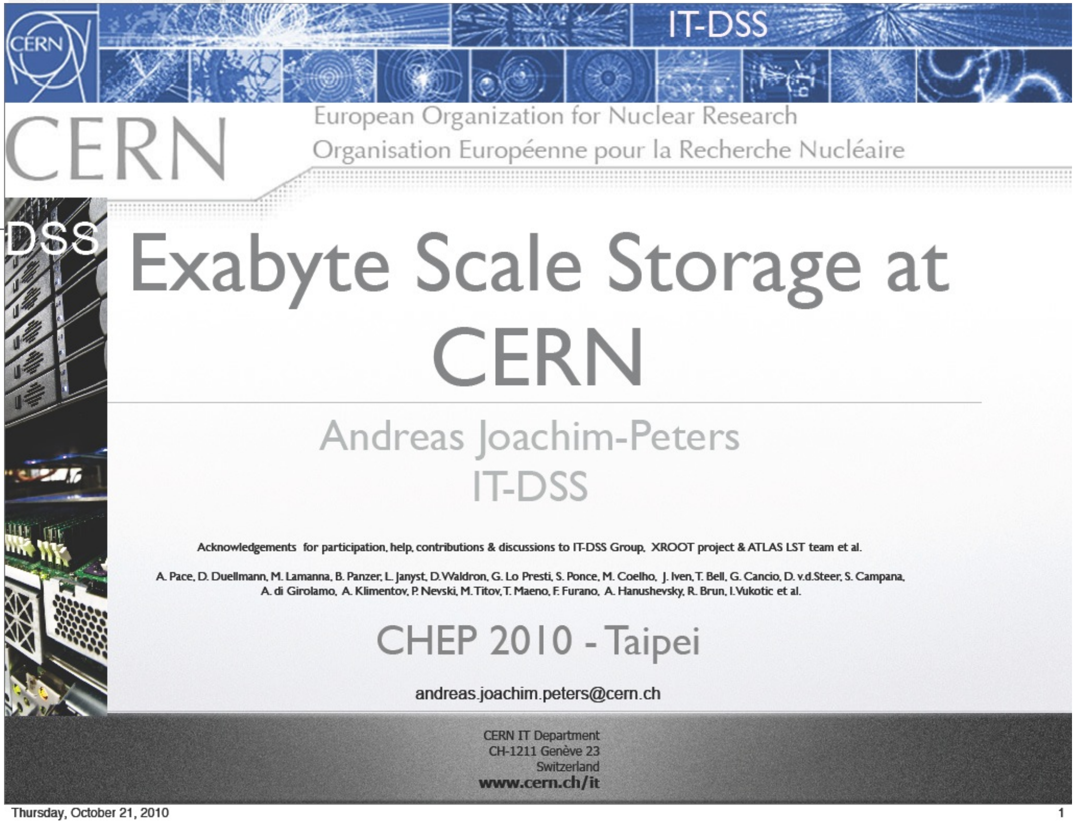
Storage Service Developments at CERN

G. Cancio Melia, *D. Duellmann*, A. Pace, CERN IT

CHEP 2010, Taipei, Taiwan
18-22 October 2010

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

Wednesday, 20 October 2010



IT-DSS

CERN European Organization for Nuclear Research
Organisation Européenne pour la Recherche Nucléaire

DSS

Exabyte Scale Storage at CERN

Andreas Joachim-Peters
IT-DSS

Acknowledgements for participation, help, contributions & discussions to IT-DSS Group, XROOT project & ATLAS LST team et al.

A. Pace, D. Duellmann, M. Lamanna, B. Panzer, L. Janyst, D.Waldron, G. Lo Presti, S. Ponce, M. Coelho, J. Iven, T. Bell, G. Cancio, D. v.d.Steer, S. Campana, A. di Girolamo, A. Klimentov, P. Nevski, M. Titov, T. Maeno, F. Furano, A. Hanushevsky, R. Brun, I.Vukotic et al.

CHEP 2010 - Taipei

andreas.joachim.peters@cern.ch

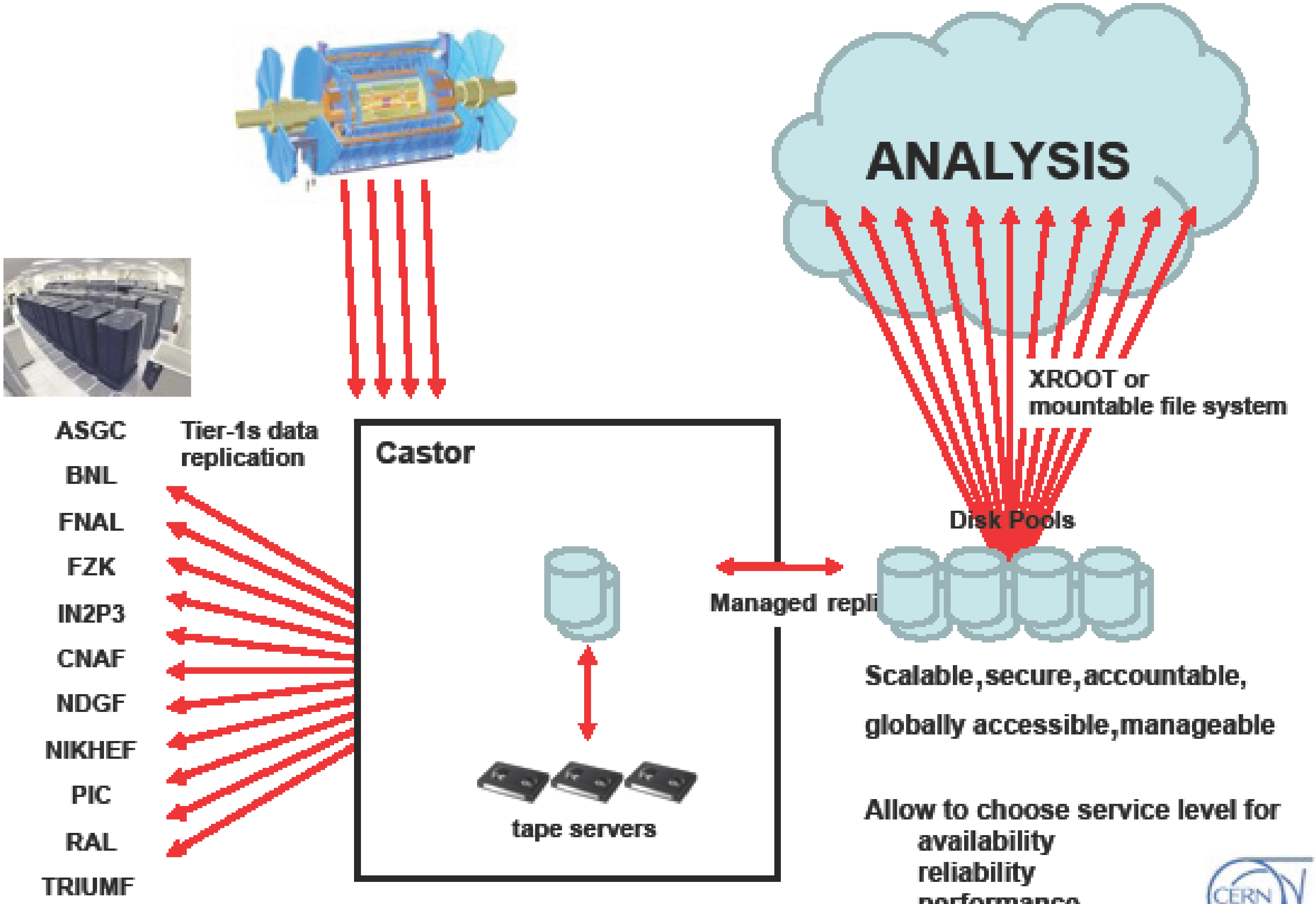
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Thursday, October 21, 2010

1

DSS First Step

LHC Experiments



Scalable, secure, accountable,

globally accessible, manageable

Allow to choose service level for
availability
reliability
performance
decoupled from HW

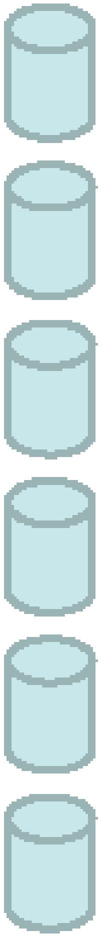
Strategy

- Use Castor for what it was designed for and for what it is good at
  - and continue current developments to improve the tape efficiency
  - larger scale re-engineering of Castor would take significant resources, time and create unacceptable risks on T0 operation
- Validate "simpler" alternate solutions to Castor disk pools to offer improved services necessary for analysis (EOS demonstrator)

DSS

EOS approach to redundancy

- EOS uses JBOD disk devices for storage
  - redundancy added on s/w layer
- Using “sets” of N *independent* disk devices
  - Current configuration uses N=6
- Each file / directory / pool can be configured to replicate files M times (with M < N)
  - For example, M=3 every file is written 3 times on 3 *independent* disks out of the 6 available



- On client reads:
  - any of the file replicas can be used
  - load is spread across many disks to achieve high throughput
  - more efficient than mirrored disks, and much better than RAID-5 or RAID-6

EOS Architecture

Management Server

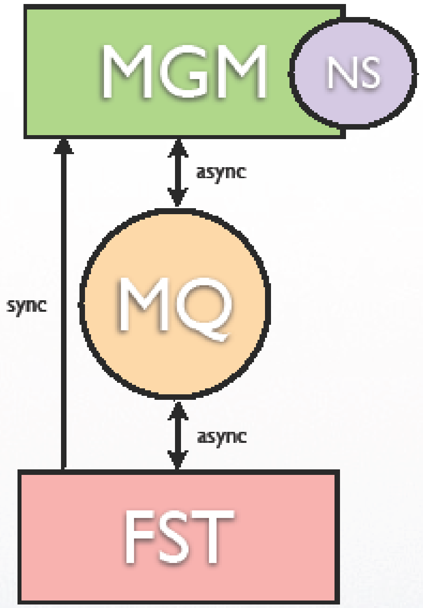
Pluggable Namespace, Quota
Strong Authentication
Capability Engine
File Placement
File Location

Message Queue

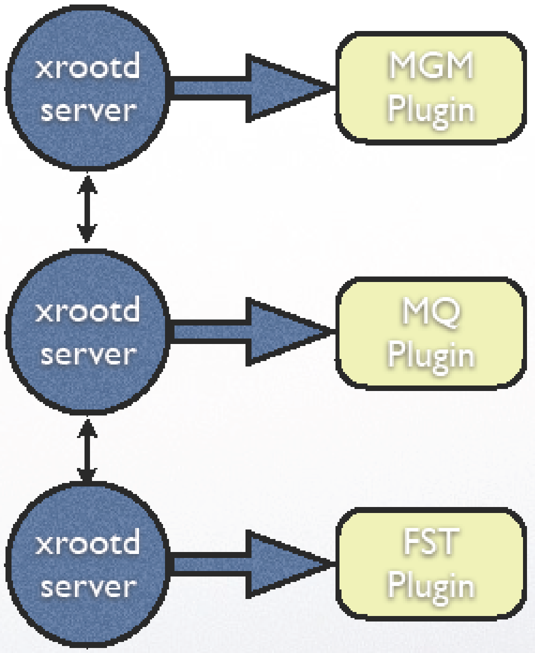
Service State Messages
File Transaction Reports

File Storage

File & File Meta Data Store
Capability Authorization
Checksumming & Verification (adler,crc32,md5,sha1)
Disk Error Detection (Scrubbing)



Implemented as plugins in **xrootd**



Current Plans

- Operation
  - 15th November opening of EOS Atlas to ATLAS users
    - operation by CERN operations team
      requires further integration & documentation
- Development
  - V2 namespace implementation & active active **ro** slave MGM server
- Larger Testbed (if available)
  - scale instance from today 600 disks to 2.000-8.000 disks (4-16 PB)

Remark: could also support NFS4.1 protocol - implementation is running inside xrootd servers but files can be accessed with any protocol supporting client stalling & redirection

CERN

European Organization for Nuclear Research
Organisation Européenne pour la Recherche Nucléaire

Tape archive challenges when approaching Exabyte-scale

CHEP 2010, Taipei

G. Cancio, V. Bahyl, G. Lo Re, S. Murray, E. Cano, G. Lee, V. Kotlyar

CERN IT-DSS

German.Cancio@cern.ch

Media migration

- Mass media migration or “repacking” required for
  - higher-density media generations, and / or
  - higher-density tape drives

- Completed last year migration from 500GB to 1TB tapes
  - Copying 45,000 tapes took around a year using 1.5 FTE and ~ 40 tape drives

- **Repack exercise is proportional to the total size of archive - and not to the fresh or active data**

- **Data rates for repack will soon exceed LHC data rates…**
  - Repack in 2012: ~55 PiB to migrate. **1.7 GB/s** sustained over a year
  - Repack in 2015: ~ 120 PiB “ **3.8 GB/s** “
  - Current LHC tape data rates : ~700MB/s

- **…. but we need to share the drives which become the bottleneck**

- Required improvements:
  - Tape write performance increases -> not sufficient drives otherwise
  - Remove network/disk server contention bottlenecks (10GiB Ethernet, disk spindles)
  - Turn media repacking into a **non-intrusive background activity** (opportunistic drive and media handling)

Conclusions

- Managing a near-Exabyte tape archive is an active task. The effort is proportional to the total archive size.
- A non-negligible fraction of resources need to be allocated for housekeeping such as migration and verification.
- Tape has a small *effective* lifetime requiring continuous media migration to new generations.
- Writing to tape scales OK - if you handle small files correctly.
- File-based HSM access will not scale for long. Move to what tape is built for: bulk archiving and streaming access.

Conclusions

- The advent of AJAX and HTML5 makes it possible to create dynamic, interactive web applications without requiring any third-party browser add ones
  - GWT is a toolkit which simplifies development of such web applications
    - Initial tests show that development is reasonably straightforward
    - We will use several GWT applications for EXO experiment and expect to use it more in future
    - Use of HTML5 features to improve plotting in future looks hopeful
      - Anyone interested in collaborating, warning us off?
  - GWT is successful enough that it is influencing other toolkits
    - Pyjamas – GWT "port" for python
    - Oracle: JavaFX – will compile to JavaScript in future (2012)
    - Many others…

Study Group for Data Preservation and
Long Term Analysis in High Energy Physics

Tests of Cloud Computing and Storage System features requested by H1 Collaboration Data Preservation model

Bogdan Lobodzinski

Session: Grid & Cloud Middleware

21 October 2010

CHEP
TAIPEI TAIWAN
2010

Cloud Computing concept for data analysis in data preservation model

- support of heterogeneous storage, OSs and VMs,
- idea of a private cloud computing – at some point the H1 software will be forced to use OS without valid security patches,
- easy base to maintain a possible migration of the H1 software to new platforms,
- common storage area for all virtual instances,
- scalability and centralized IT support,
- cost reduction – nodes can be shared with other projects,

Cloud Computing test configuration - summary

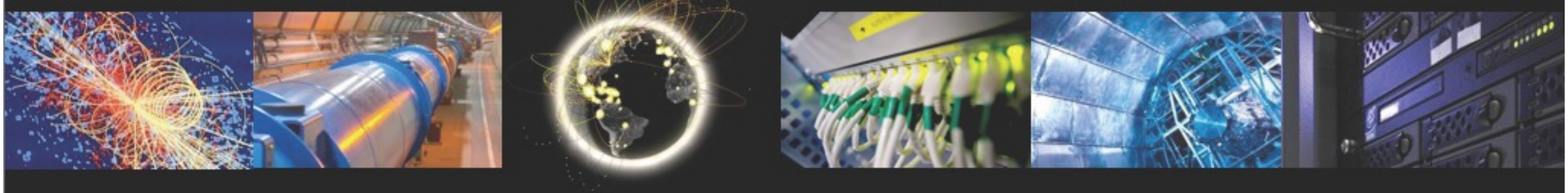
- General: Eucalyptus Cloud manager & CEPH look promising but not ready for production mode - too poor stability & reliability,
- CEPH Petabyte FS is truly experimental - all tests with bigger number of files failed,
- In both cases management is difficult,
- It is too early to present any benchmarks and performance,
- Results are encouraging, allow for optimistic anticipation of the future

Proxy Caches for File Access in a Multi-Tier Grid Environment

R.Brun, *D. Duellmann*, G. Ganis, A. Hanushevski, L.Janyst, AJ. Peters

CHEP 2010, Taipei, Taiwan

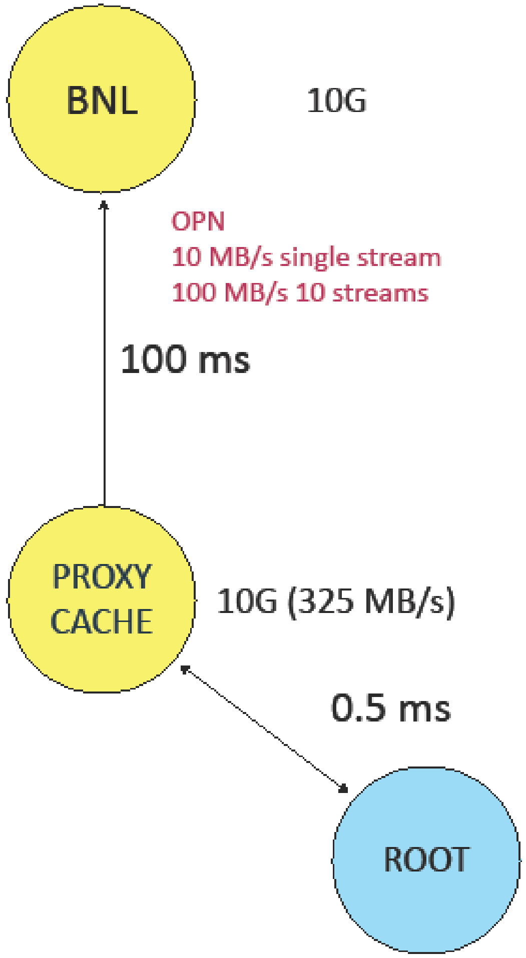
18-22 October 2010



WLCG
Worldwide LHC Computing Grid

First Tests with the Setup

atlasFlushed.root + TreeCache (R. Brun) ~1GB



WAN Performance (preliminary)

ATLAS AOD ROOT Native (no Athena/Pool)

Channel	IO rate
BNL	3-4 MB/s
Local	10 MB/s

Goal: reach 10 Mb/s over WAN

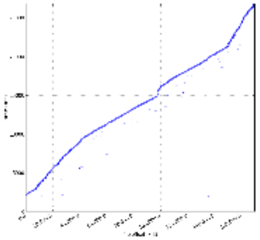
ALL branches	RT/s	MB/s	Cached	Performance wrt local access
1st read via proxy	288	3.7	0%	28%
2nd read via proxy	109	9.7	100%	97%
2 branches				
1st read via proxy*	104	0.323	0%	
2nd read via proxy	0.9	35	3.9%	
direct access @ BNL	49	0.68		

* the proxy currently cannot bridge readV requests

Summary

- Proxy-Cache demonstrator has been proposed at WLCG Data Management Jamboree
  - Aim: improved performance, lower service effort
- Progress in setting up a testbed between several sites
  - Measurements & Improvements for WAN access ongoing
  - Measurements for page cache started
    - preliminary results are promising
    - but systematic study is just starting...
  - ROOT tree cache plugin to be designed
- Sufficient resources and experiment contacts in place to finalise results for the demonstrator review in January

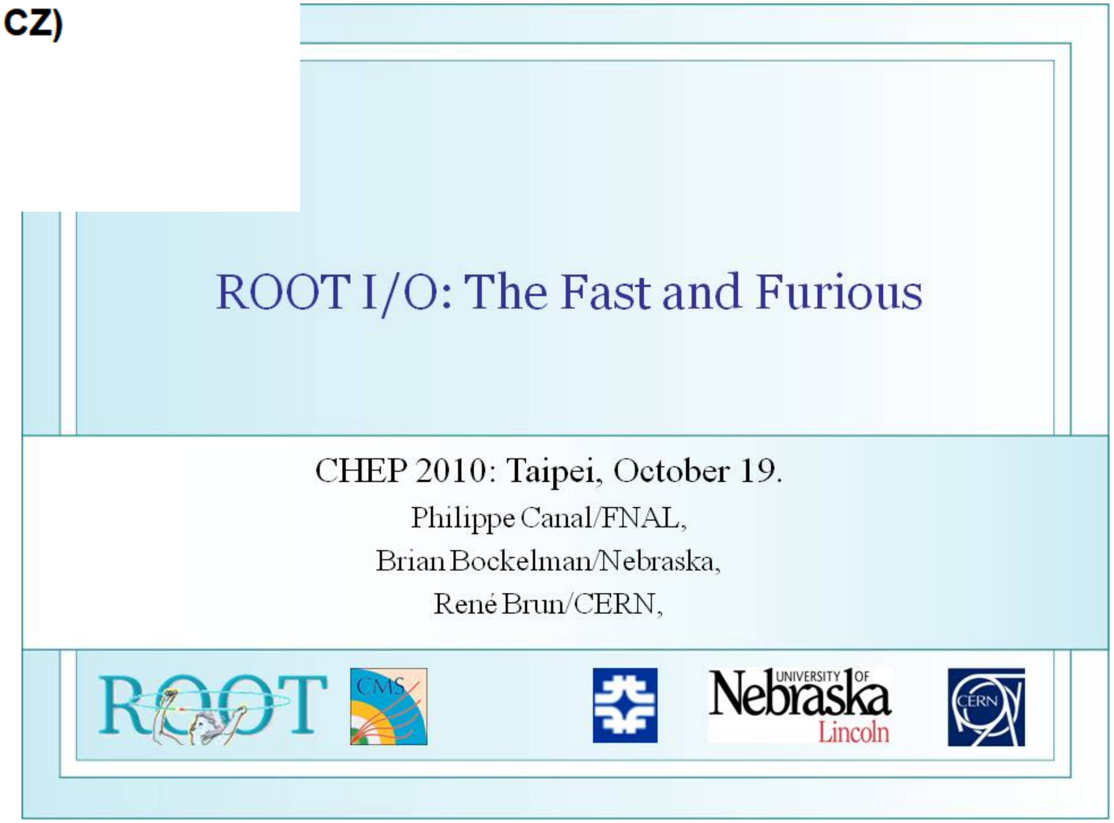
From detailed analysis of IO pattern of the HEP applications to benchmark of new storage solutions



Jiří Horký[1], Roberto Santinelli[2]

horky@fzu.cz Roberto.Santinelli@cern.ch

[1]**Institute of Physics, AS CR, Prague (FZU CZ)**
[2]CERN



ROOT I/O: The Fast and Furious

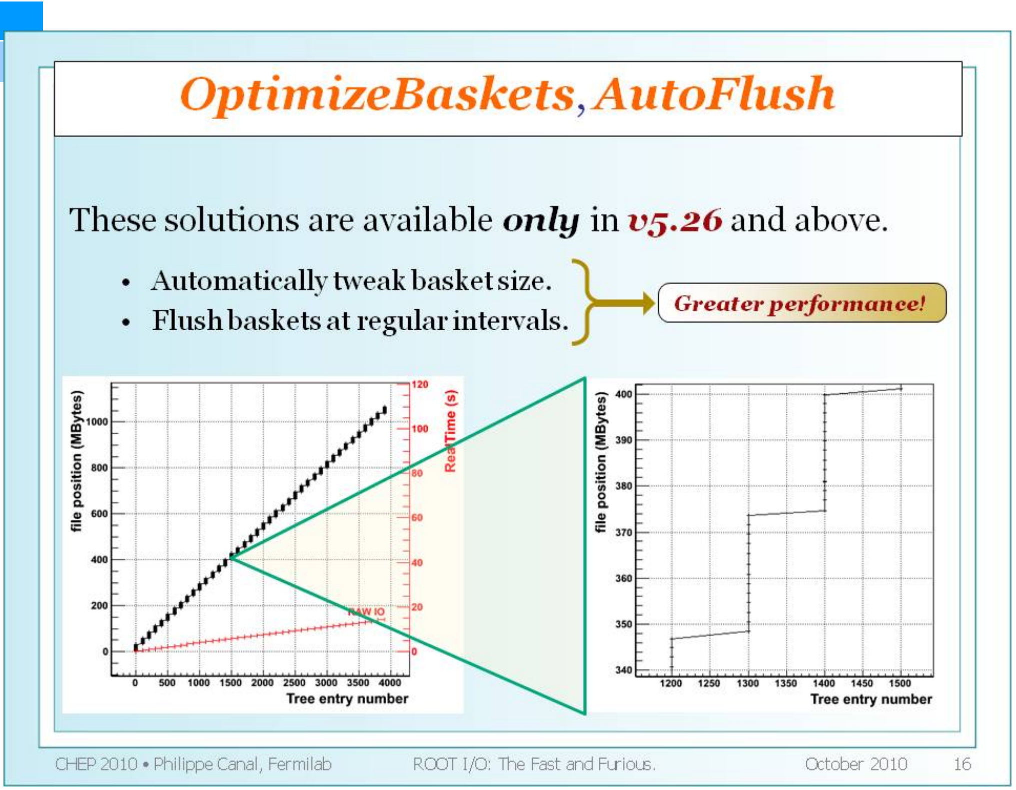
CHEP 2010: Taipei, October 19.

Philippe Canal/FNAL,
Brian Bockelman/Nebraska,
René Brun/CERN,

ROOT and its access pattern

- Changes in ROOT are really significant:
  - operations are now sequential --> overhead of caching much smaller
  - 5x improvement in reading for the ATLAS job (read 250MB from a 1GB file, see prev. slide)

ROOT version when saving + cache setting	#reads	time (ms)	kBytes read
5.22, cache off	191 209	25 895	247 203
5.22, cache on	3 205	7 297	541 572
5.26, cache off	14 658	5 338	212 428
5.26, cache on	797	4 348	236 759



OptimizeBaskets, AutoFlush

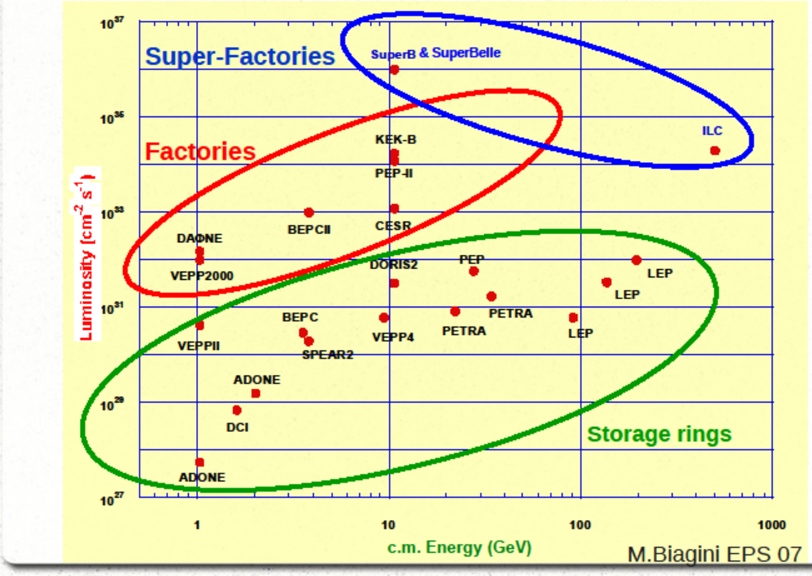
These solutions are available ***only*** in ***v5.26*** and above.

- Automatically tweak basket size.
- Flush baskets at regular intervals.

Greater performance!

CHEP 2010 • Philippe Canal, Fermilab ROOT I/O: The Fast and Furious. October 2010 16

e+e- colliders



M.Biagini EPS 07

Computing R&D topics

- Impact of new CPU architecture, software architecture and framework
- Code development: languages, tools, standards and QA
- Persistence, data handling models and Databases
- User tools and interfaces
- Distributed computing, GRID
- Performance and efficiency of large storage systems

INFN Istituto Nazionale di Fisica Nucleare

CHEP

SuperB

Computing for the Next Generation Flavour Factories

Armando Fella for the SuperB Computing Group
CHEP 2010, Taipei, 18-22 October

Future plan

- *Italian government experiment approval is expected within 2010*
- Our current commitment is to focus on support for the SuperB detector Technical Design Report
  - Concentrate on physics, detector, and background simulation studies
- We are planning for a SuperB computing TDR, describing the final computing model
  - to be released one year after the detector TDR (second half of 2012)
- The outline of an R&D program to be carried out in 2010 and 2011 has been defined
  - The first Computing R&D Workshop have taken place in Ferrara: http://www.fe.infn.it/superb/