Track Summary: Distributed Processing and Analysis on Grids and Clouds

Oliver Gutsche
Philippe Canal
Johannes Elmsheuser

FNAL
Ludwig-Maximilians-Universität München

25 May 2012/CHEP 2012, New York

Outline

Track Statistics

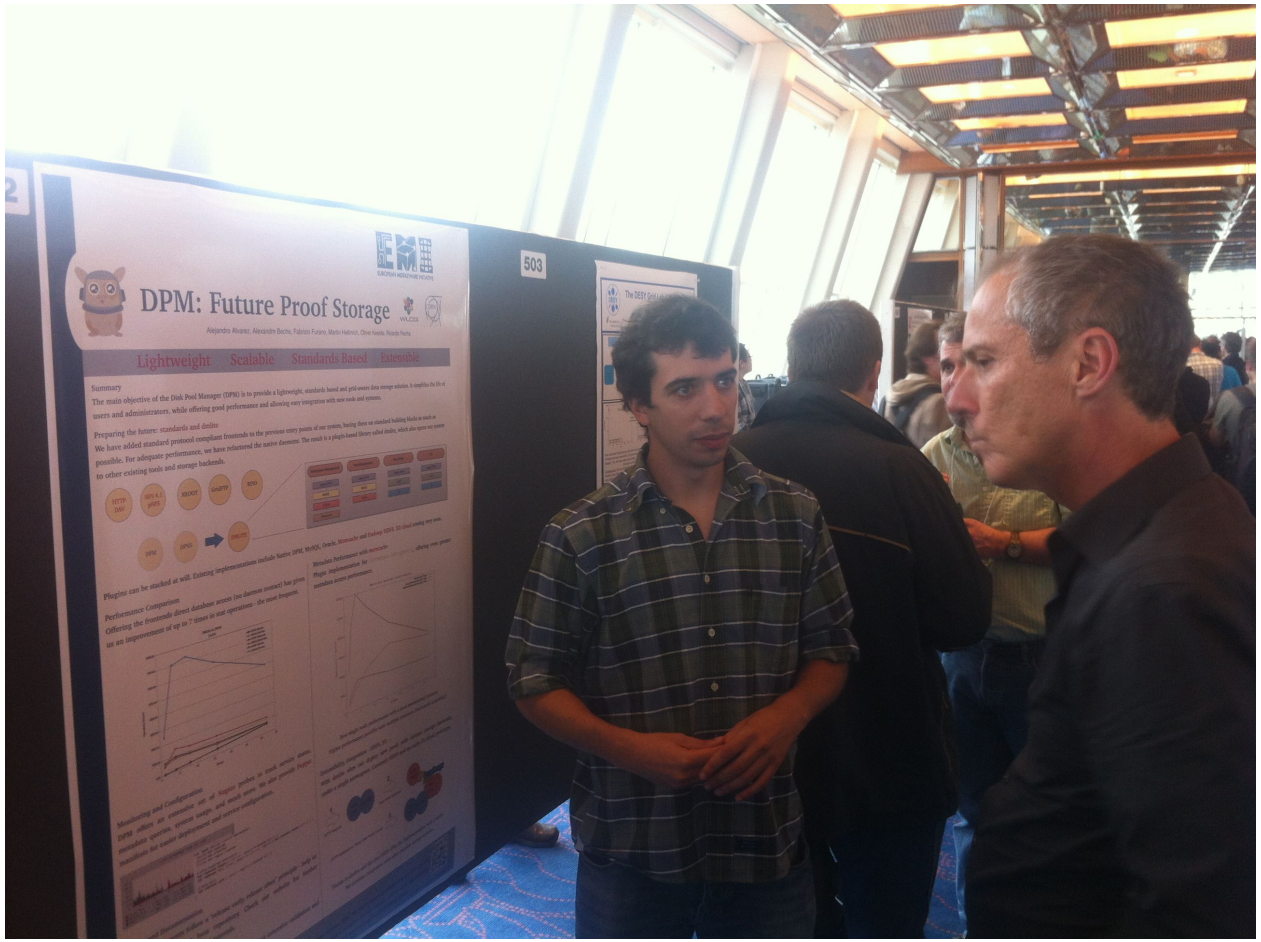
- Merged track from previous CHEPs:
  Grid and Cloud Middleware and Distributed Processing and Analysis
- 174 abstracts after merging and reassignments to/from other tracks
- 31 talks in 7 parallel sessions - 2 no-shows
- 143 posters accepted
- 27 papers already submitted to the journal
- $\Rightarrow$ Largest Track - very difficult to make everybody happy
- Broad variety of Grid and Cloud related topics

Track Topics

Topics of the different session:

- LHC experiment DDM and WM overviews
- Multi-Core and parallel scheduling
- Virtualization and Clouds
- Monitoring and Security
- Storage
- Grid middleware
- Dynamic data placement
- Experiment computing models

Poster Sessions



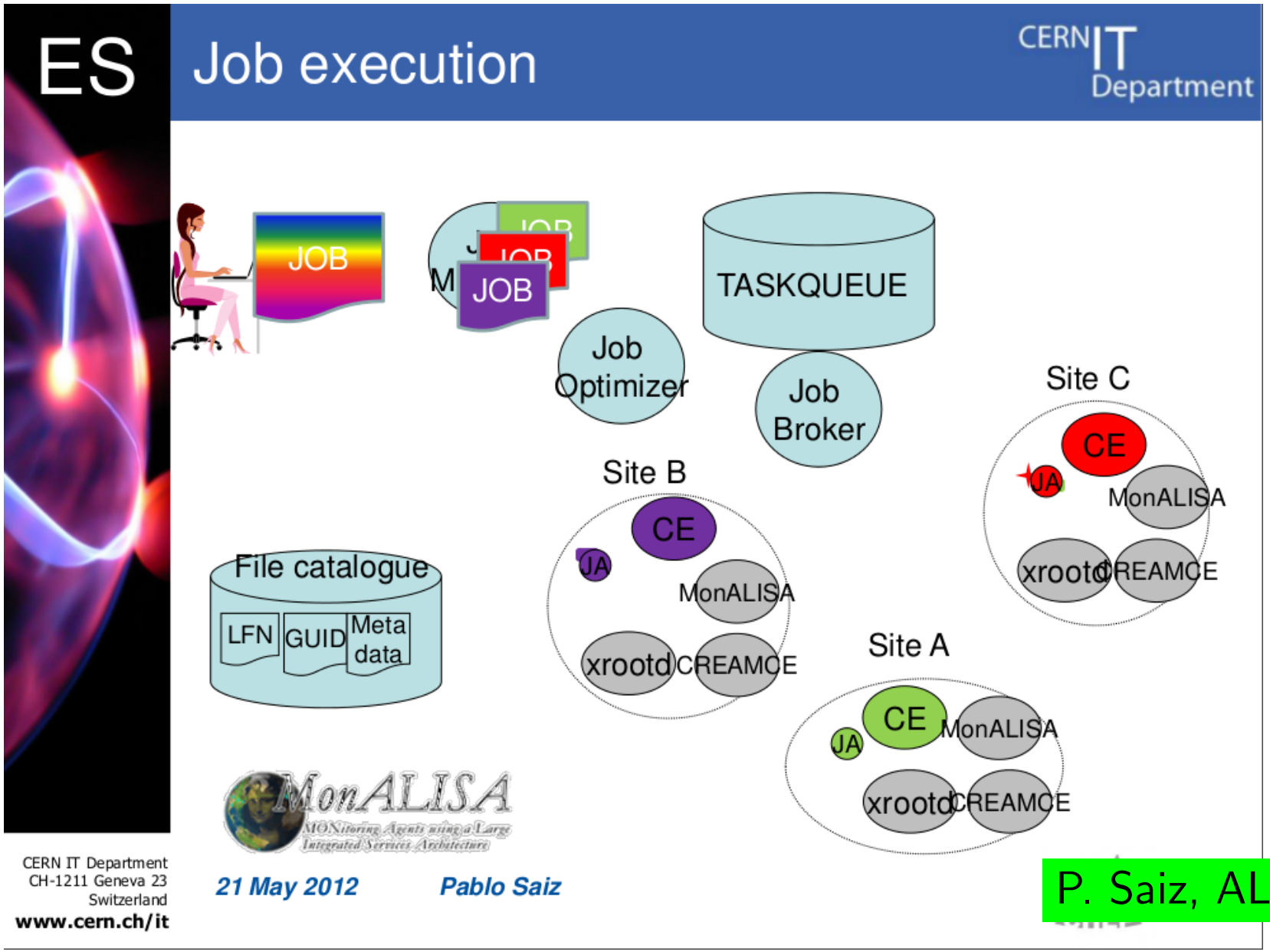
Lively discussions during the Poster sessions
142 posters are close to impossible to mention during the summary, apologies for being so brief about the posters

Outline

LHC experiment status

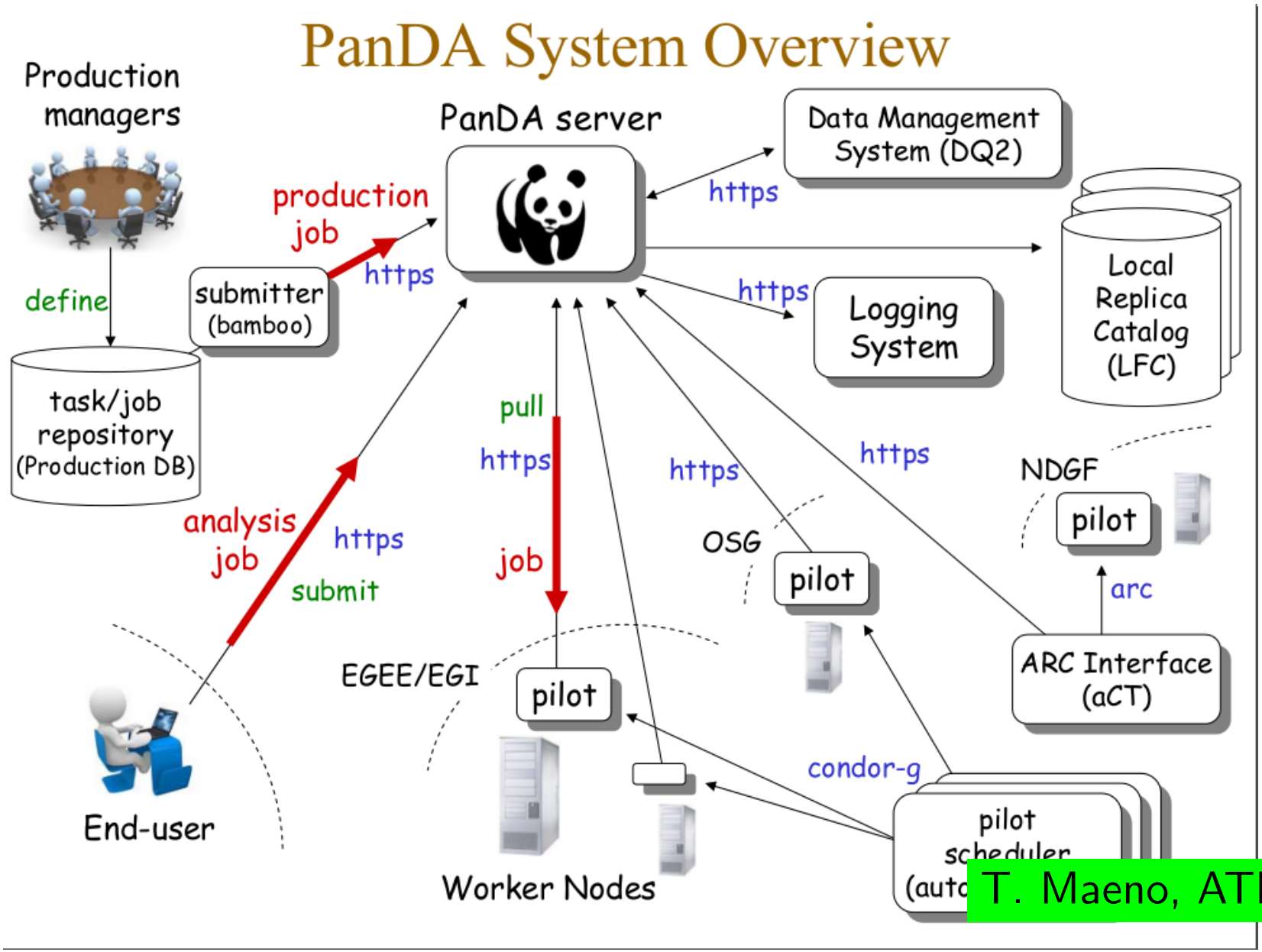
- All experiments have built their customized workload management systems for production and analysis and data management system on top of the existing grid middleware
  $\Rightarrow$ Very successful in delivering physics results
- But experiments are trying to streamline systems, remove unneccessary components, ease operations with limited person-power, find commonalities, scale to higher needs, adapt to new technologies

Workload Management Systems

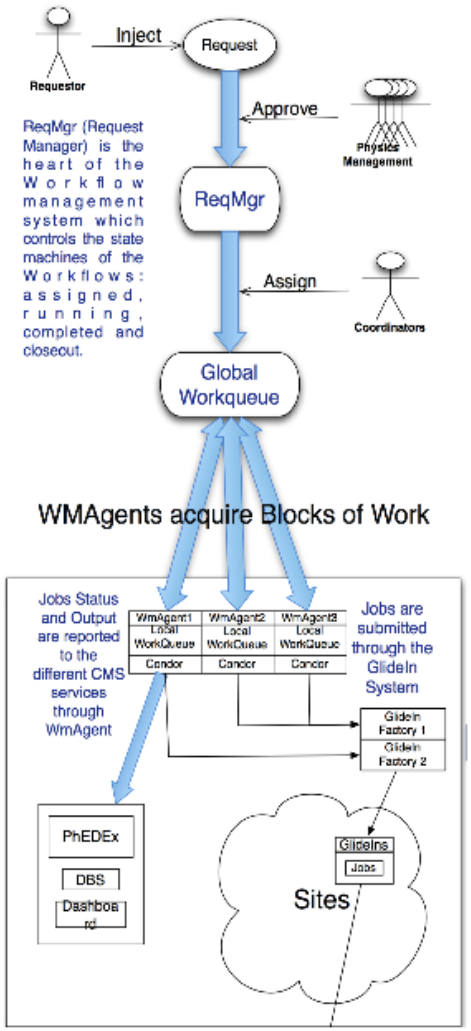


P. Saiz, ALICE

Workload Management Systems



New WM system

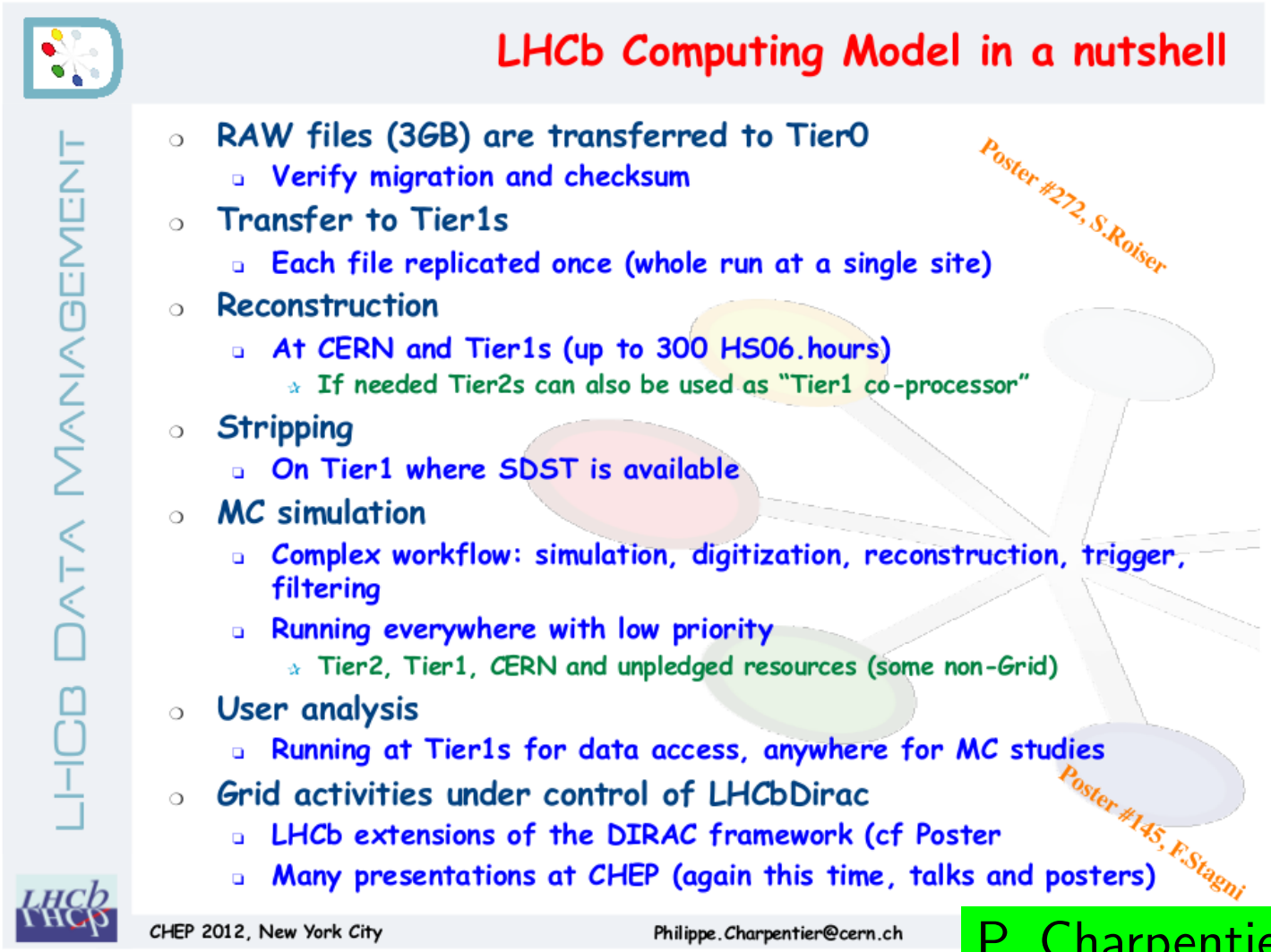


- Consolidate analysis and organized activity
  - Same components but different instances
    - Prevent interference
- Single entry point for requests.
  - Permanently recorded - reproducible
  - Requester can view status.
- Prioritization between requests.
- Approval chain.
- Work distributed automatically & optimally to resources.
- Reduced manpower needs.
- Adapt to new features / requirements:
  - Pilot jobs.
  - Multi-core processes.
- Some use cases require all events to be processed.
  - Cope with intermittent problems.

S. Wakefield, CMS

Stuart Wakefield — Imperial College London

Data Management Systems

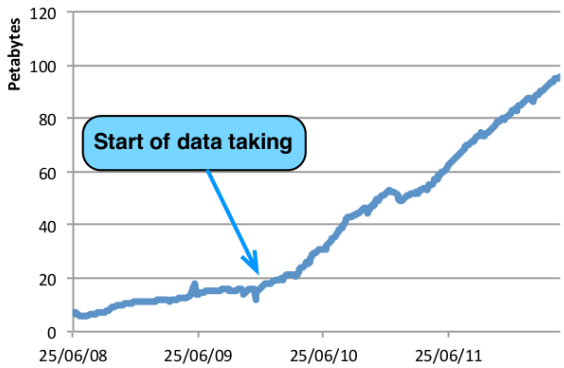


P. Charpentier, LHCb

Data Management Systems

$\mathcal{DQ}_2$ Status

Total Grid Space Usage



Central Services

- 12M read queries/day
- 0.6M write requests/day

Scale

- 95 PB and 300 million files managed
- 800 users
- 130 sites with 700 storage endpoints

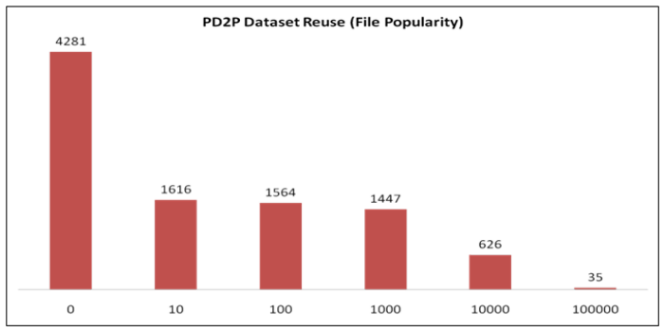
Grid File Accesses

- 5M/day
- Linear increase (factor 2/year)

7

V. Garonne, ATLAS

Frequency of Reuse for PD2P Data



- 45% of data were never reused while others were well reused
- Users' interest is largely unpredictable and volatile
  - The request based model of PD2P suits user behaviour

21

T.Maeno

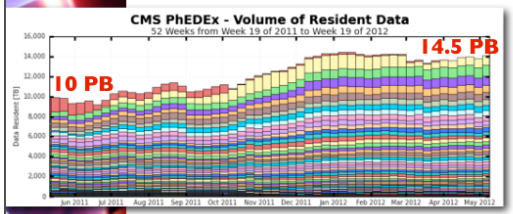
Large amounts of data replicated - but sometimes never touched - ATLAS pioneered popularity based automatic data replication with Panda/DQ2

Data Management Systems

ES

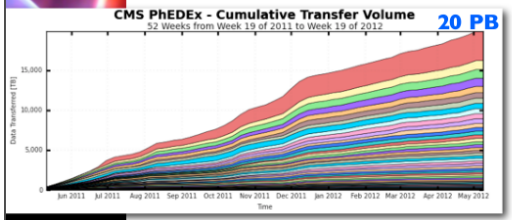
Many data in many places

CERN IT Department



Official datasets needed for analysis (real data, sim data) are distributed among the ~50 CMS T2 sites

- ~14.5 PB of data are resident @ T2 site
- ~4.5 PB added just in the last year of LHC operation
  - NB: copies of a datasets are stored in different places



O(20PB) : volume of data transferred to the CMS T2 sites in the last year

- Transfers ~ 5 x Resident

Need to know how we are using our data

- Data used / reused /unused by physics groups in different sites

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it
D. Giordano

ES

Dataset time evolution

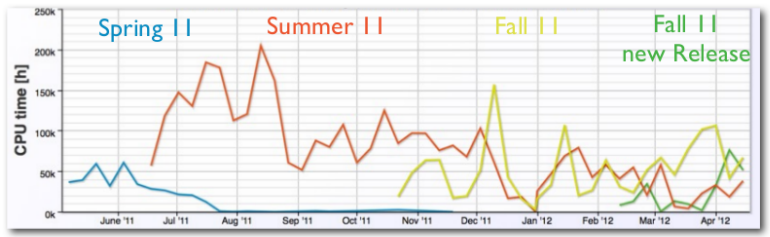
CERN IT Department

Allow to monitor the trend of usage of the same kind of data (same MC process, same data acquisition period) along the different processing campaigns

- New reconstruction release, new detector conditions, new simulated scenarios, etc

Users follow the processing campaigns

At some point becomes useful to remove old datasets



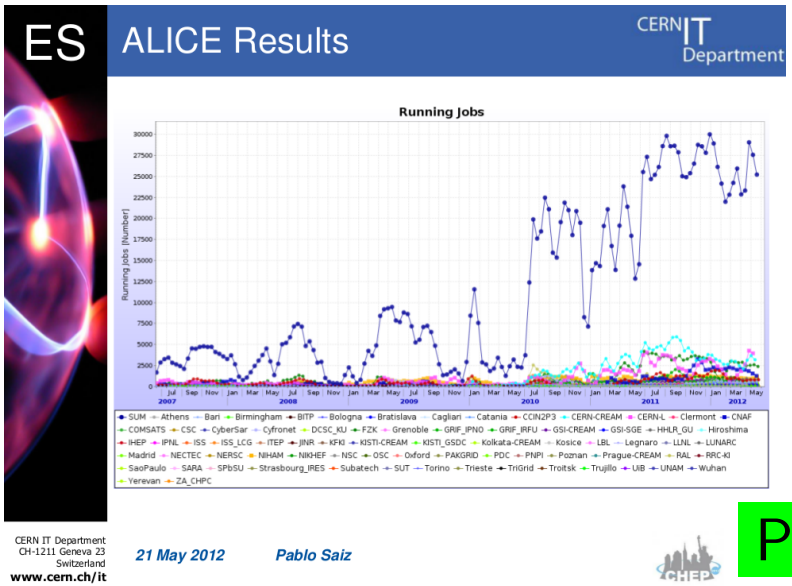
Time evolution of W+jet datasets

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it
D. Giordano

D.Giordano, CMS

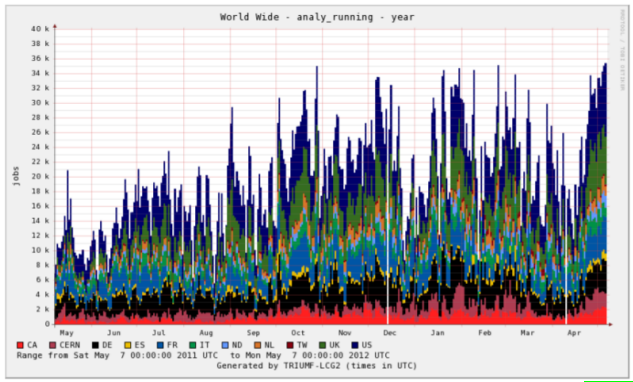
Similar large amounts of data replicated - now also moving to popularity system and replica deletion

Experiments Statistics



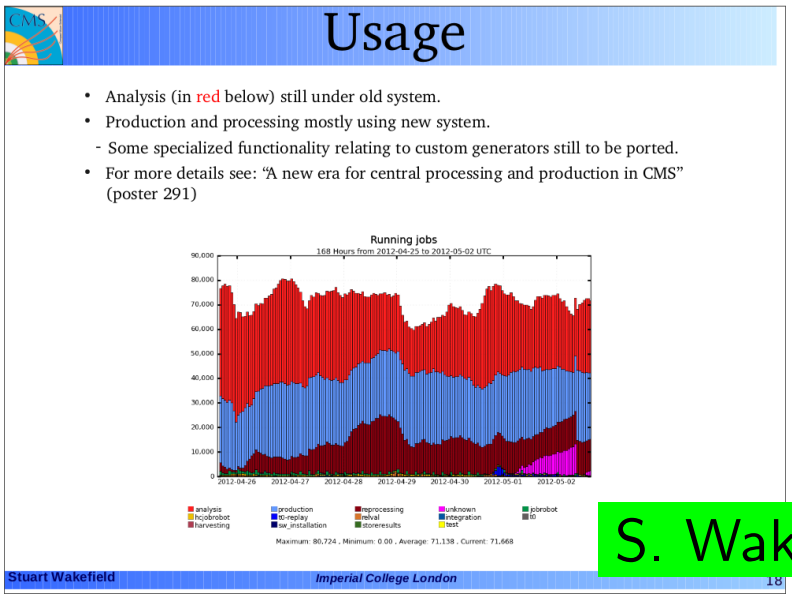
P. Saiz

Analysis Jobs Concurrently Running
'11 –'12



- Steady increase of analysis activities

T. Maeno



S. Wakefield

Several 20k-100k concurrent running production and analysis jobs for each experiment

Experiment grid improvements, Commonalities

The next $\mathcal{DDM}$ version: $\mathcal{R}$ucio

Why a new major version ?

- New high-level use cases and workflows
- New technologies, paradigms and middleware
- Difficult to extend the existing system with new concepts
- Old design (2006) with some conceptual limitations and heavy operational burden

High Level Roadmap

2011: Technical meetings with other LHC experiments, user surveys, collection of use cases
⇒ Conceptual model document

2012: Parallel and incremental development track, incubator projects, preparatory steps

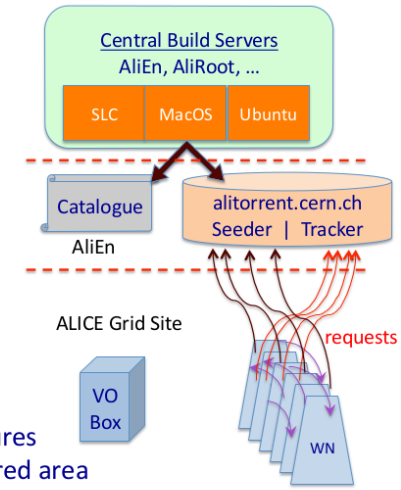
2013: $\mathcal{R}$ucio in production

V.Garonne

AliTorrent Software Deployment

ALICE

- Managed Central Software
  - Additional AliEn torrent store
  - Catalogue, seeder & tracker
- Grid site SW deployment
  - VO Box is not involved
  - Jobs pull SW from:
    - alitorrent.cern.ch seeder
    - local peers
    - other sites as available
      - though typically behind a FW
- Resolves:
  - Bottleneck & single point failures
  - Site level maintenance of shared area



R. Jeff Porter LBNL 6

J.Porter, ALICE

ATLAS is building new DDM system - ALICE tries out bit-torrent based SW distribution

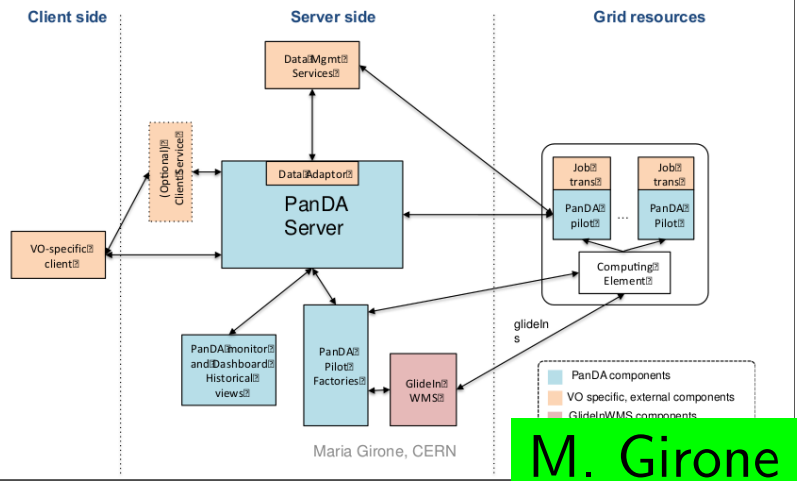
Experiment grid improvements, Commonalities

ES

Proof of Concept Diagram

CERN IT Department

- Feasibility Study proved that there are no show-stoppers to design a common analysis framework
- Next step is a proof of concept



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

Maria Girone, CERN

M. Girone

Conclusions

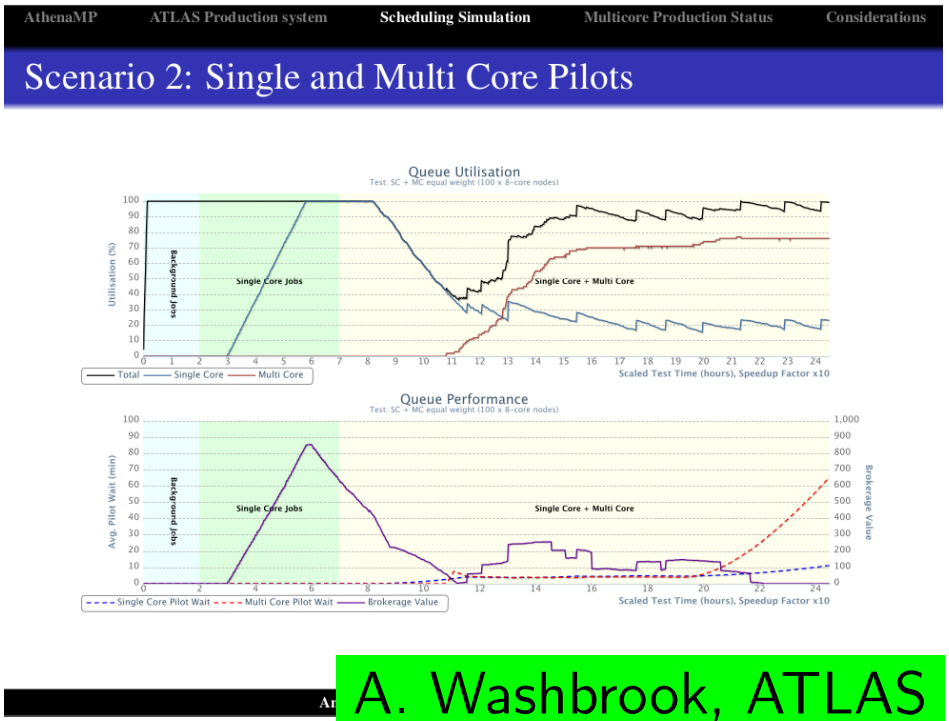
- HC was successful in stress testing, and is now heavily used for functional testing
  - 8.3 million ATLAS jobs, 8 million CMS jobs in 2012
  - ~130 ATLAS sites, ~80 CMS sites tested
  - More that 2 billion metric records stored in 2012.
  - LHCb still developing…
- Use-cases are expanding, so we are working on scalability of the service.
  - Improved service design/procedures
  - Finding redundant tests should lead to improved efficiency

D. van der Ster

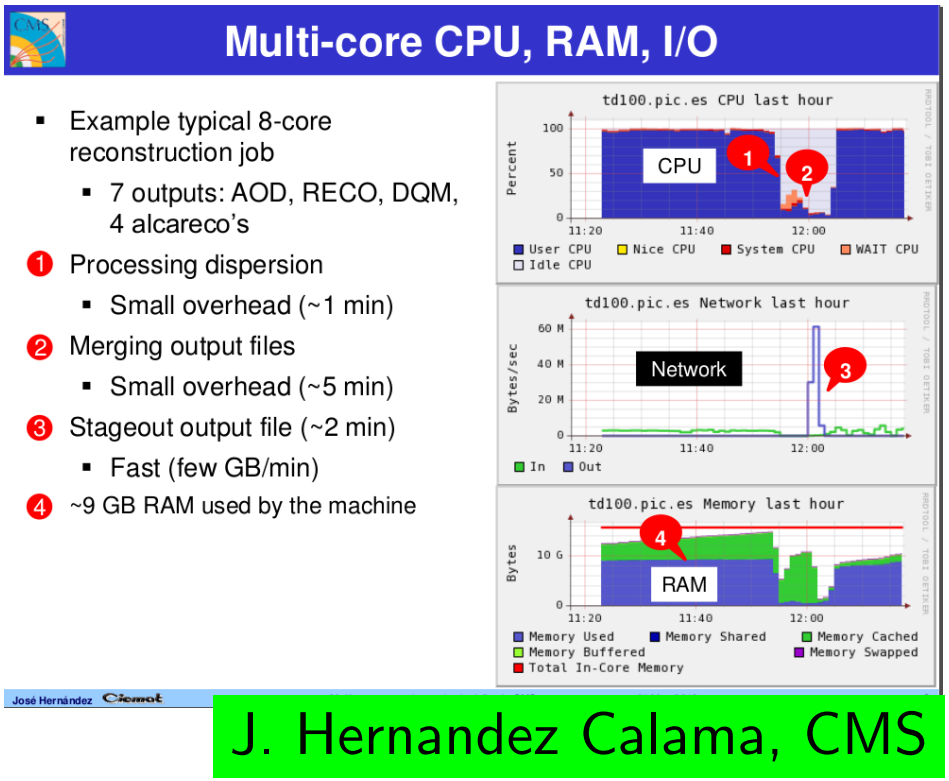
Study if ATLAS Panda is suitable for analysis in CMS - HammerCloud is used among 3 experiments for grid site validation

Outline

MultiCore and Job Scheduling



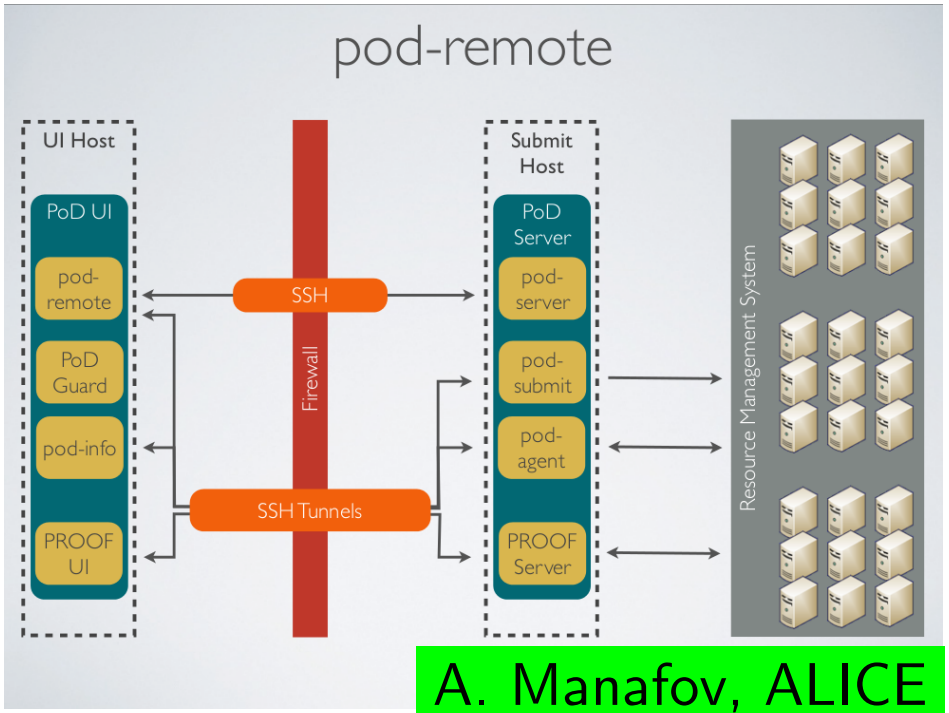
A. Washbrook, ATLAS



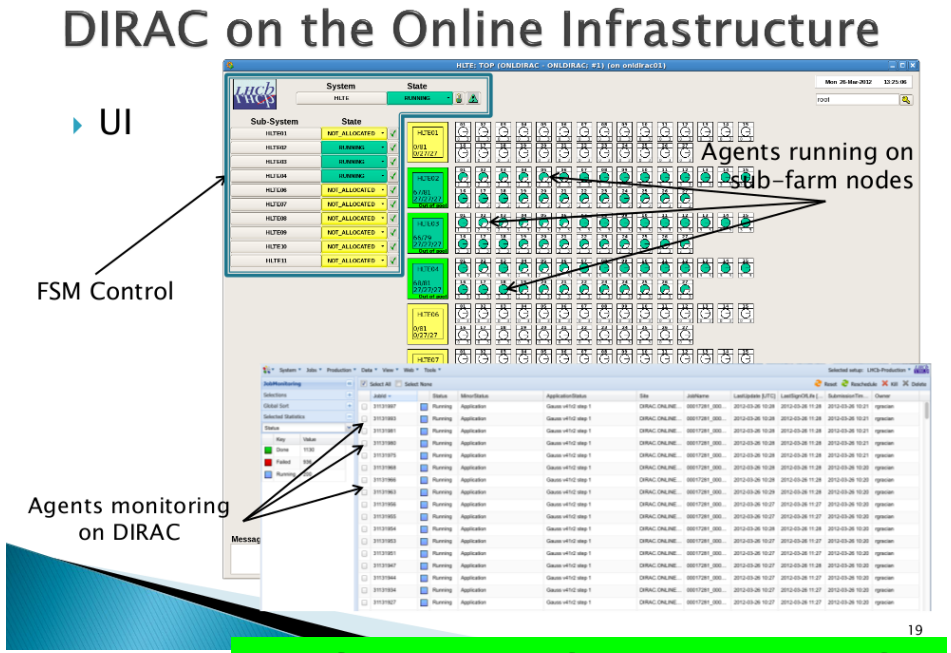
J. Hernandez Calama, CMS

- Number of cores per node is increasing
- Physical memory per core is decreasing
- Different multi-core job scheduling options possible - full node scheduling with slightly different resource usage

MultiCore and Job Scheduling



A. Manafov, ALICE



L. Granado Cardoso, LHCb

Alternative methods for parallel job scheduling on local cluster or online system

Outline

Cloud Computing in ATLAS

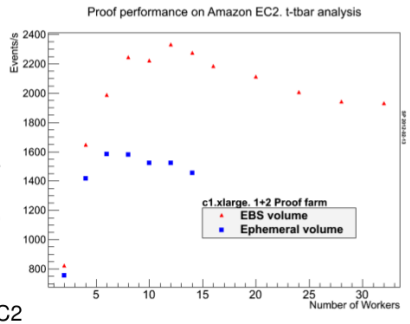
ATLAS Cloud Computing R&D

EFFICIENCY, ELASTICITY

- ATLAS Cloud Computing R&D is a young initiative
  - Active participation, almost 10 persons working part time on various topics
  - Goal: How we can integrate cloud resources with our current grid resources?
- **Data processing and workload management**
  - PanDA queues in the cloud
    - Centrally managed, non-trivial deployment but scalable
    - Benefits ATLAS & sites, transparent to users
  - Tier3 analysis clusters: instant cloud sites
    - Institute managed, low/medium complexity
  - Personal analysis queue: one click, run my jobs
    - User managed, low complexity (almost transparent)
- **Data storage**
  - Short term data caching to accelerate above data processing use cases
    - Transient data
  - Object storage and archival in the cloud
    - Integrate with DDM

Data Access Tests

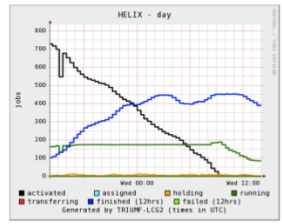
- Evaluate the different storage abstraction implementations that cloud platforms provide
- Amazon EC2 provides at least three storage options
  - Simple Storage Service (S3)
  - Elastic Block Store (EBS)
  - Ephemeral store associated with a VM
  - Different cost-performance benefits for each layout that need to be analyzed
- Cloud storage performance on 3-node PROOF farm
  - EBS volume performs better than ephemeral disk
  - But ephemeral disk comes free with EC2 instances
    - Scaling of storage space and performance with the size of the analysis farm



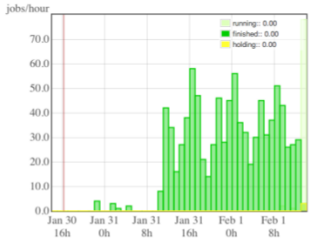
Results

- 100 nodes/200 CPUs at Cloud Sigma used for production tasks
- Smooth running with very few failures
- Finished 6 x 1000-job MC tasks over ~2 weeks
- We ran 1 identical task at CERN to get reference numbers

	HELIX	CERN
Success Rates	265 failed, 6000 succeeded	36 failed, 1000 succeeded
Mean Running Times	$16267s \pm 7038s$	$8136.6s \pm 765.5s$



The jobs/hour for all sites progress

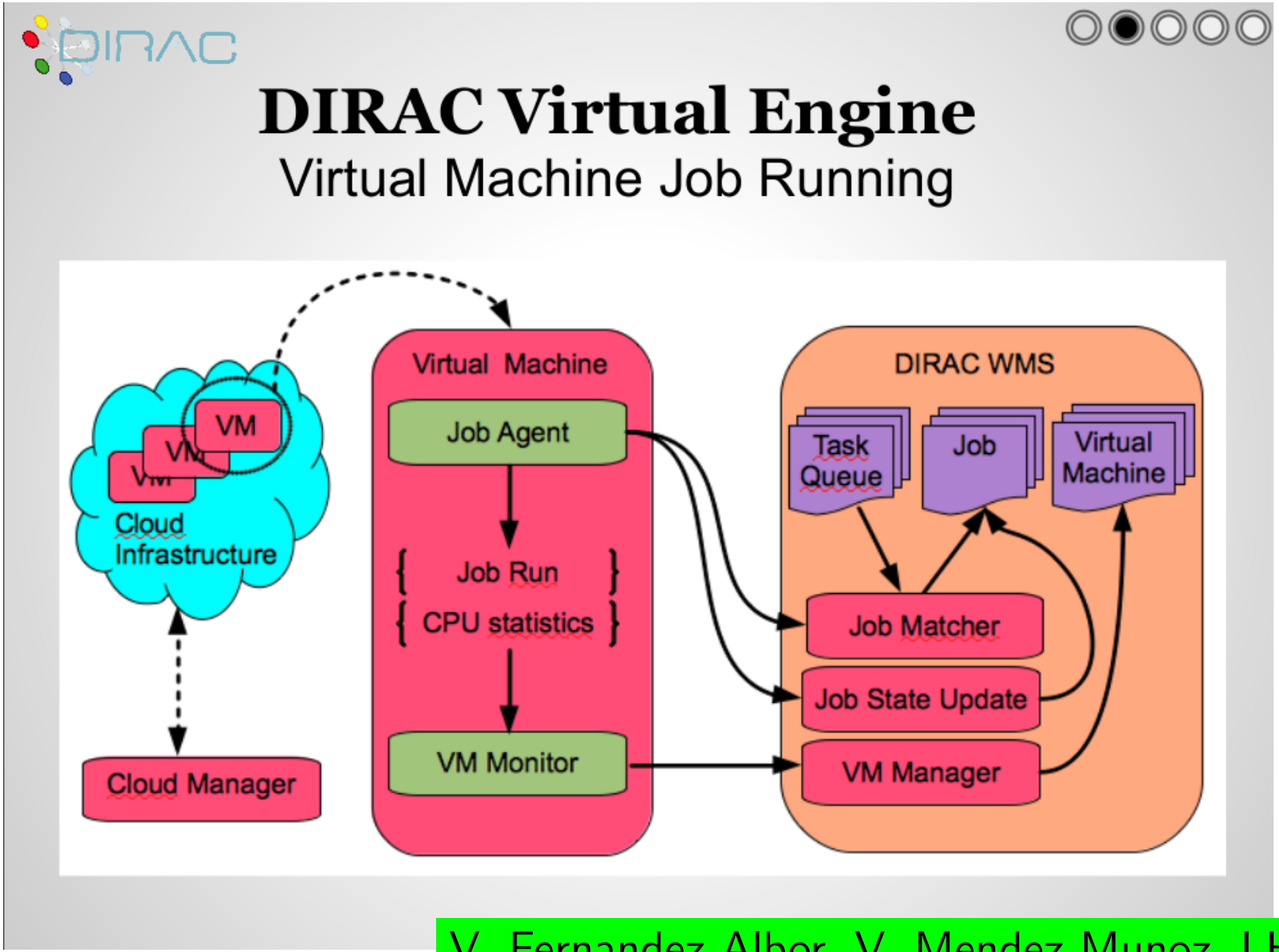


- Wall clock performance cannot be compared directly, since we don't have the same hardware on both sites
  CloudSigma has ~1.5Ghz of AMD Opteron 6174 per jobslot, CERN has a ~2.3GHz Xeon L5640
- Best comparison would be CHF/event, which is presently unknown

ATLAS with extensive Cloud R&D, tested production on commercial cloud

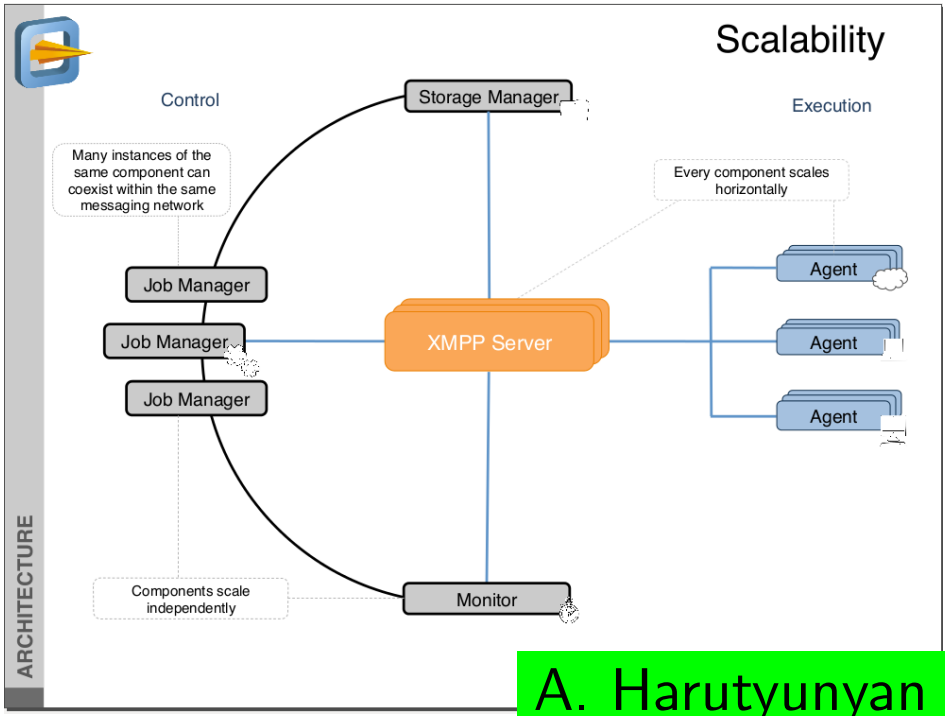
F. Barreiro Megino, ATLAS

Further Cloud examples



V. Fernandez Albor, V. Mendez Munoz, LHCb

Further Cloud examples



A. Harutyunyan

Summary

- The HEPiX Virtualisation Working Group has shown how trusted user generated images can be safely instantiated at Grid sites
  - compatible with site security policies and obligations, and
  - with a guaranteed environment for the experiments.
- Three options now for following this up
  - Ignore and hope it goes away
  - Integrate trusted virtual images into a traditional batch environment
  - **Exploit the proposal** to deliver a "Grid of Clouds" that enables sites to easily and simply present to experiments a seamless—but dynamically changing—set of resources that the experiments can exploit to schedule work according to physics priorities.

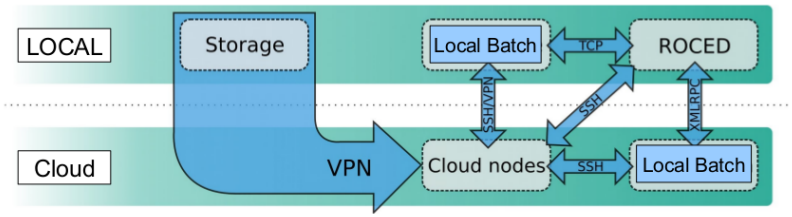
Tony.Cass@CERN.ch 11

T.Cass

CERNVM co-pilot on opportunistic resources - Virtualization offers coherent environment

Further Cloud examples

Batch Resource extension in the Cloud
ROCED@KIT

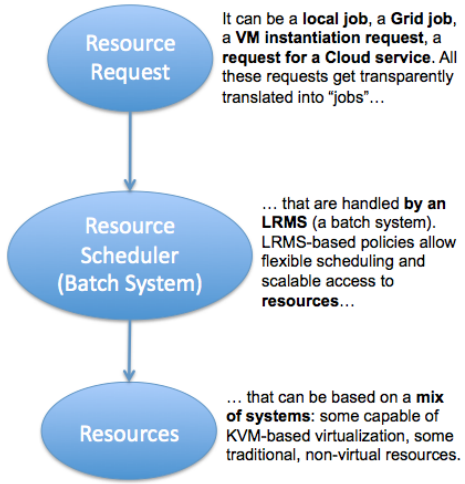


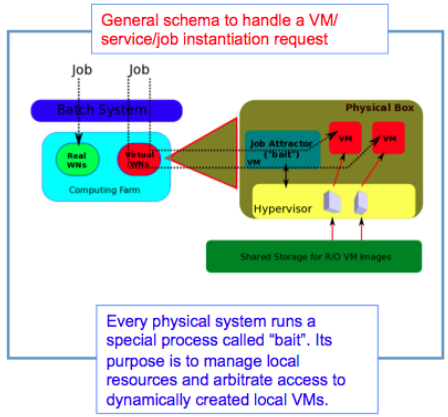
- ROCED runs on the same machine as the local batch server.
- Local batch system communicates with its nodes and users via TCP.
- Commands to the OpenNebula host are sent via XMLRPC call.
- The Communication between the Cloud nodes, ROCED and the Cloud Server are done via SSH.
- **No modifications to the firewall (besides VPN tunnel) needed.**

14 22/05/12 Oliver Oberst

O. Oberist

WNoDeS, Architectural Overview





CHEP - May 22, 2012 Exploiting resources with WNoDeS - D.Salomoni 5

D. Salomoni

- Setup local Virtualization or Cloud cluster with ROCED
- WNoDeS Mixed Mode lets a resource center to progressively introduce virtualized services without disrupting existing setups and maximizing resource utilization

Outline

Computing at Belle II, SuperB

Estimated Data Rates

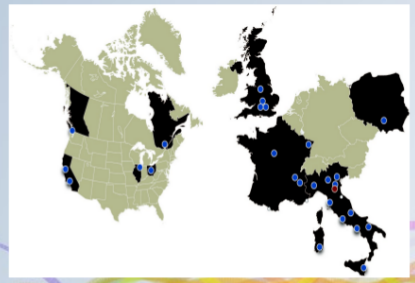
Experiment	Event Size [kB]	Rate [Hz]	Rate [MB/s]
High rate scenario for Belle II DAQ:			
Belle II	300	6,000	1,800
LCG TDR (2005):			
ALICE (HI)	12,500	100	1,250
ALICE (pp)	1,000	100	100
ATLAS	1,600	200	320
CMS	1,500	150	225
LHCb	25	2,000	50

KIT Thomas Kuhr CHEP 22.05.2

T. Kuhr, Belle II

SuperB distributed resources

- The distributed computing infrastructure, as of May 2012, includes several sites in Europe and North America
- EGI and OSG Grid flavours have been enabled
- The LHC Computing Grid architecture was adopted to provide the minimum set of services and applications upon which the SuperB distributed model could be built.
- Computing resources needed in a typical year of SuperB data taking are of the same order as corresponding ATLAS and CMS estimation for 2011



Parameter	typical Year
Luminosity (ab^{-1})	15
Storage (PB)	
Tape	113
Disk	52
CPU (KHep-Spec06)	
Event data reconstruction	210
Skimming	250
Monte Carlo	670
Physics analysis	570
Total	1700

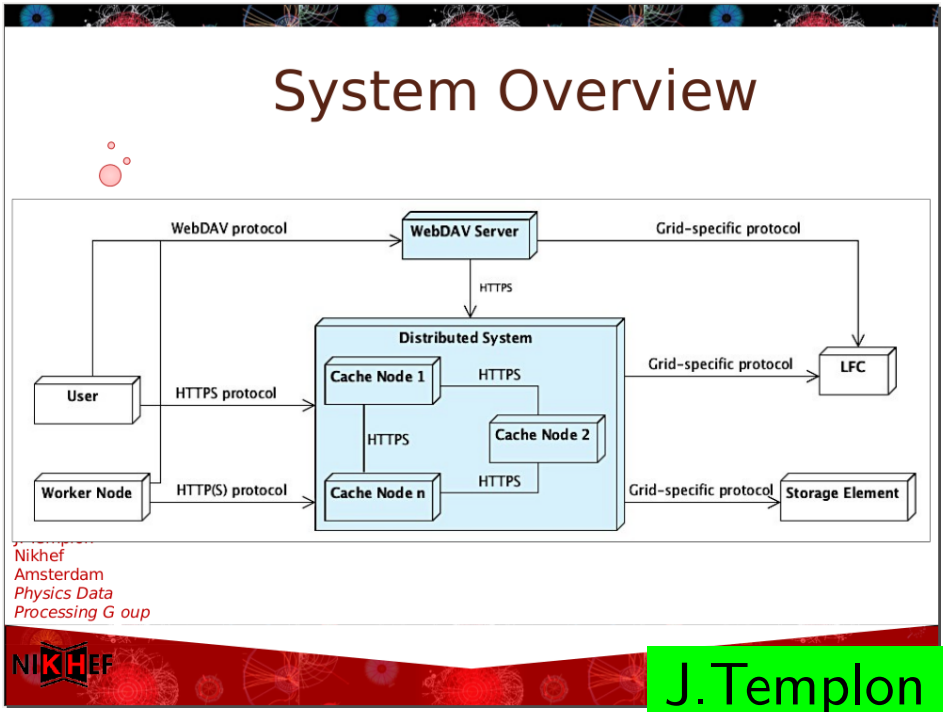
CHEP12, New York, May 24th, 2012 4

A.Fella, SuperB

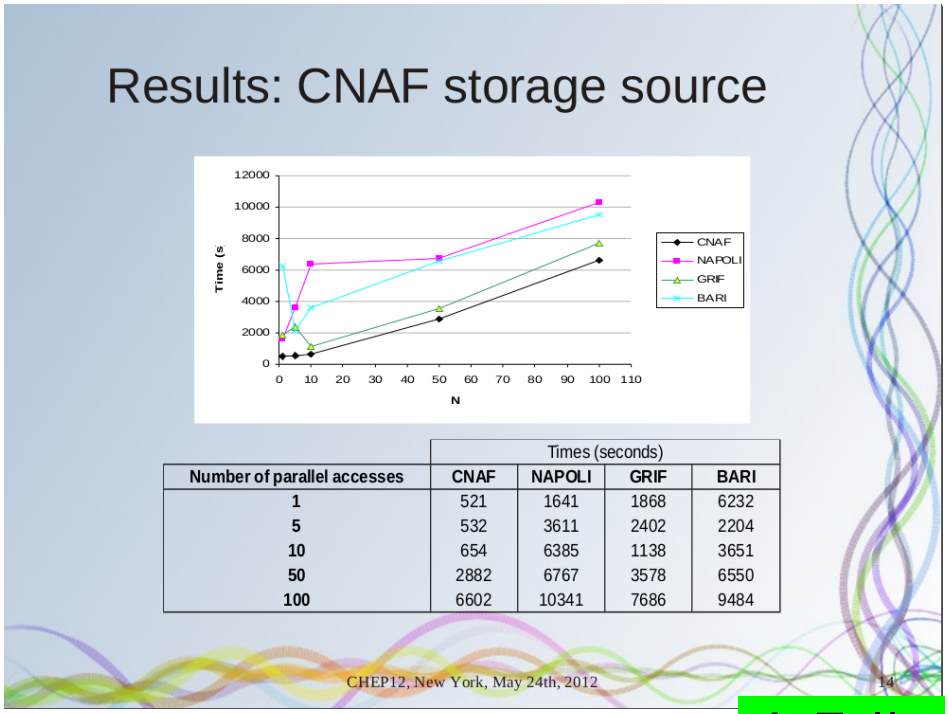
New experiments adopting LHC grid software: DIRAC, AMGA, HappyFace, CVMFS

Outline

Software and Middleware



J.Templon

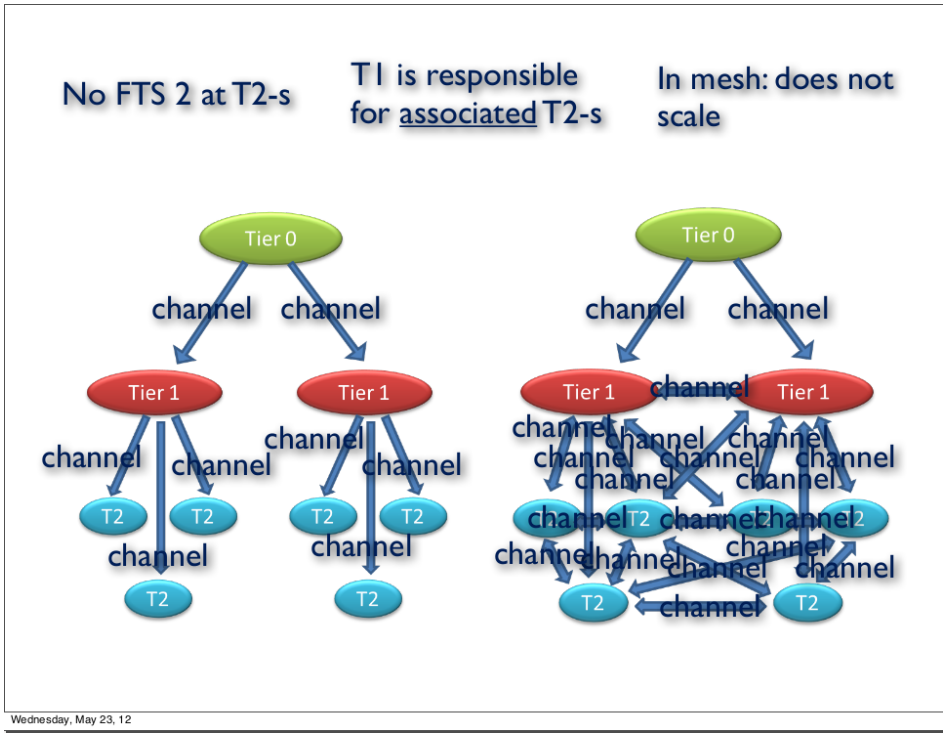


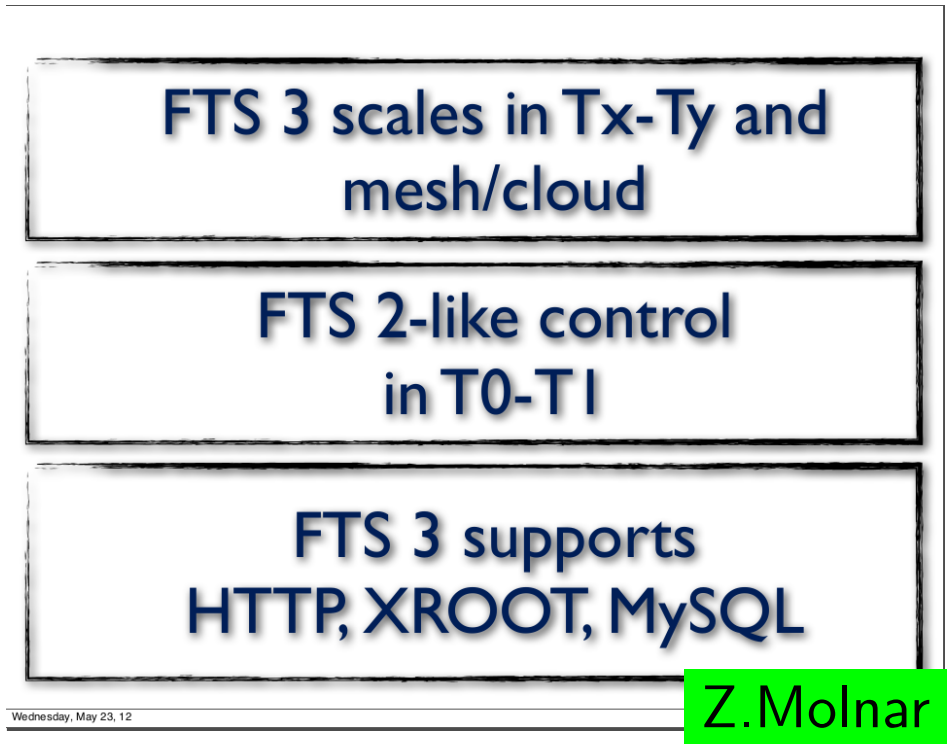
Number of parallel accesses	Times (seconds) CNAF	NAPOLI	GRIF	BARI
1	521	1641	1868	6232
5	532	3611	2402	2204
10	654	6385	1138	3651
50	2882	6767	3578	6550
100	6602	10341	7686	9484

A.Fella

Local and Remote HTTP data access

Software and Middleware





Computing Models have evolved from hierarchical system to mesh - FTS3 will adapt to these changes

Software and Middleware

News: dCache & pNFS

dCache.org

- NFS v4.1 / pNFS has been supported since 2009.
- Deployed **in production** (at DESY) for over a year.
- Fermilab's REX dept. evaluated dCache NFSv4.1 for their Intensity Frontier experiments:

  "**Results look promising**, throughput scales well with number of pool nodes"
- Supports:
  - authn: trusted-host and Kerberos
  - all three GSS security modes.

dCache, agile adoption of storage technology | Paul M

D.Millar

DESY Grid Lab

Summary

- Storage systems based on traditional spinning disks have a limited I/O performance.
- Adding SSDs, as an additional cache layer in currently deployed storage solutions (DPM, dCache..) can boost I/O performance significantly.
- An analysis of CMS user activities has shown that only 10% of cache, results in 70% of random reads from the cache.
- A storage system with 3 Tiers is transparent for the user and cost effective for sites.

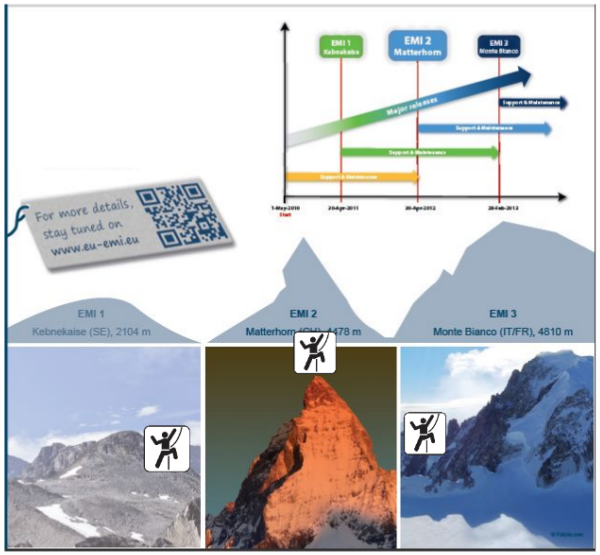
5/22/12 CHEP 2012, The need of 3 Tier storage 17

D.Ozerov

News in several areas from large scale storage system dCache - 3 tiered system with SSD cache allow better performance

Projects

EMI Releases



24/5/2012 www.eu-emi.eu

B.Konyan

Transition of some support from External Projects

- Bestman support now under the OSG Software Area.
- ESNET DOEGrid PKI infrastructure/Certificate Authority -> OSG
  - Receiving additional support from ESNET to implement.
  - Need to have no disruption in production service.
  - Deploy OSG PKI infrastructure for User and Host/Service certificates based on DigiCert commercial backend by Oct. 2012.
  - US ATLAS and US CMS contributing to the 1st 2 year license to a defined threshold. It is not yet known if this is sufficient.
  - Feb. 2013 DOEGrids ceases to issue new certificates.
  - Feb. 2014 DOEGrids ceases to publish CRL lists.
- Understanding of support needed for Globus Toolkit
  - GRAM, GridFTP

Ruth Pordes OSG@CHEP, New York, 2012

R.Pordes

Both EMI and OSG are looking into the future, with EMI planning to hand over software projects to software teams and OSG looking to continue support and integrate more non-HEP users

Conclusions

- Many thanks to the Organizers, Speakers and Poster Presenters for their interesting contributions
- We think the sessions were a big success and well received
- In general the session were on time although we had lively discussions
- Proceedings:
  - Due to the large number of possible proceeding contributions we are asking all track participant to help
  - Please volunteer to review 1-2 papers
  - Don't worry: we would not expect detailed technical comments about the content

Thanks a lot for an interesting and fun time in NYC