CHEP 2012
Computing in High Energy and Nuclear Physics 2012 • New York • United States

Report from CHEP 2012

Track report:

Distributed Processing and Analysis on Grids and Clouds

Armando Fella

SuperB contributions

- Computing for High Energy Physics contributions:
  - Oral presentation: “**Exploiting new CPU architectures in the SuperB software framework**”, M.Corvo
  - Oral presentation: “**SuperB R&D computing program: HTTP direct access to distributed resources**”, A.Fella
  - Poster: “**Testing and evaluating storage technology to build a distributed Tier1 for SuperB in Italy**”, S.Pardi
  - Poster: “**SuperB Simulation Production System**”, L.Tomassetti
  - Poster: “**DIRAC evaluation for the SuperB experiment**”, A.Fella

Report from CHEP, summary

- Contributions have been appreciated
  - Several questions and comments
  - Useful discussions and meetings arisen
  - Both orals cited in final track summaries
- Included in R.Pordes, “Open Science Grid in Adolescence: 2012-2016” oral presentation:
  - “Embrace future physics, nuclear physics, astrophysics experiments: Belle II, DES, EIC, LSST, SuperB”
- Many private meetings and discussions
  - PhEDEx system evaluation
  - Fermilab resource access
  - ROOT I/O optimization
  - Dirac system
  - GlideinWMS use in OSG

Meeting: PhEDEx system evaluation I

- **Participants**: D.Bonacorsi (CMS management), T.Wildish (developer leader), C.Grandi, A.Fella
- Since ~one year Phedex group is working for project generalization
- Integrated with long term project as FTS --> FTS3
- Proved to be an optimal data management framework
- Documentation will be available for non CERN experiments in few weeks
- Phedex backend:
  - Modeling pure data placement information
  - Adoption determines the SuperB information systems design:
    - bk-prod + bk-analysis + data-placement + file-catalogue
  - Isolation of data placement metadata seems to be a correct design choice, need to be verified
  - Difficult porting from Oracle to PostgreSQL IS tech

Meeting: PhEDEx system evaluation II

- CMS interest and declared support capacity is very high
- Testbed ready at CNAF and one day ready at CERN via CernVM
- Integration in a wider computing model scenario including Workload Manager has been discussed
  - "Compatible" with a federated storage environment
  - Simple integration divers file-catalogues ex: LFC/ng or DFC
- SuperB side:
  - Need to find a person for evaluation work coordination
  - Tentative next contact, end 2012

Fermilab resource access

- Participants: S.Timm (Data Management manager), A.Fella
- In the context of OSG support collaboration work (S.Timm introduced by G.Garzoglio)
  - SuperB requirements for official production use case
    - Disk resource access via dCache, amount, kind of services, per use case plan(production, analysis)
    - CPU availability, spare cycle
    - Plan on resource access at short/mid-term

ROOT I/O optimization

- P.Canal from fermilab (pcanal@fnal.gov)
  - http://root.cern.ch/drupal/content/root-presentation-chep-2012 and [*]
- Improvements in ROOT I/O span many dimensions including:
  - reduction and more control over the memory usage
  - drastic reduction in CPU usage
  - optimization of the file size and the hardware I/O utilization
- A certain level of support have been asked for SuperB developing the data access general library[**]
  - Email exchange has already started
  - We are proposing a discussion among computing group to agree on a ROOT version upgrade plan to better coordinate groups requirements and suggestions
  - [**] see G.Donvito presentation on distributed computing session: Sat 2nd, 18:00->19:30
  - [*]The ATLAS ROOT-based data formats: recent improvements and performance measurements: https://indico.cern.ch/contributionDisplay.py?sessionId=3&contribId=378&confId=149557
  - I/O Strategies for Multicore Processing in ATLAS: https://indico.cern.ch/contributionDisplay.py?sessionId=3&contribId=377&confId=149557

Dirac system

- A.Tsaregorodtsev: project leader and developer
- https://indico.cern.ch/search.py?p=dirac&confId=149557&collections=Contributions
- 17 contributions, posters and orals, about both Dirac itself and experiments are using it.
- Discussion about:
  - Coexistence of Dirac framework with other key elements of Data Model and Workload management
  - Integration between Dirac File catalogue and LFC/LFC-new-generation from EMI R&D works
  - Historical considerations around Dirac evolution and interactions with Ganga project

GlideinWMS use in OSG

- I.Sfiligoi (sfiligoi@fnal.gov)
  - https://indico.cern.ch/search.py?p=Glideinwms&confId=149557&collections=Contributions
- OSG resource exploitation via unique point of submission and brokering: GlideinWMS
- Collected information about procedures and setup to be applied to SuperB submission system to be compliant with GlideinWMS
- GlideinWMS group is available for supporting in such a task
  - http://tinyurl.com/glideinWMS
  - http://www.thinkmind.org/index.php?view=article&articleid=cloud_computing_2011_8_40_20068
  - http://iopscience.iop.org/1742-6596/331/7/072031
  - http://www.thinkmind.org/index.php?view=article&articleid=adaptive_2011_2_20_50040

Track report:
Distributed Processing and Analysis on Grids and Clouds

- 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
- ⇒ Largest Track - very difficult to make everybody happy
- Broad variety of Grid and Cloud related topics

Outline, macrosubject

- WM and DM evolutions for LHC exp
- WAN data access
- Clouds and virtualization
- EGI and OSG middlewares

Hot subjects, a catch all list

- Cloud computing and virtualization
- Non-relational databases
- Many core processors exploitation
- CERNVM File System for data access
- End to end network monitoring
- Event and file level caching
- Federated distributed storage systems
- WAN data access
- Http/WebDAV data interface
- Dynamic file catalogue
- FTS3
- Peer to peer data access solutions
- Three tiers memory stack including SSD

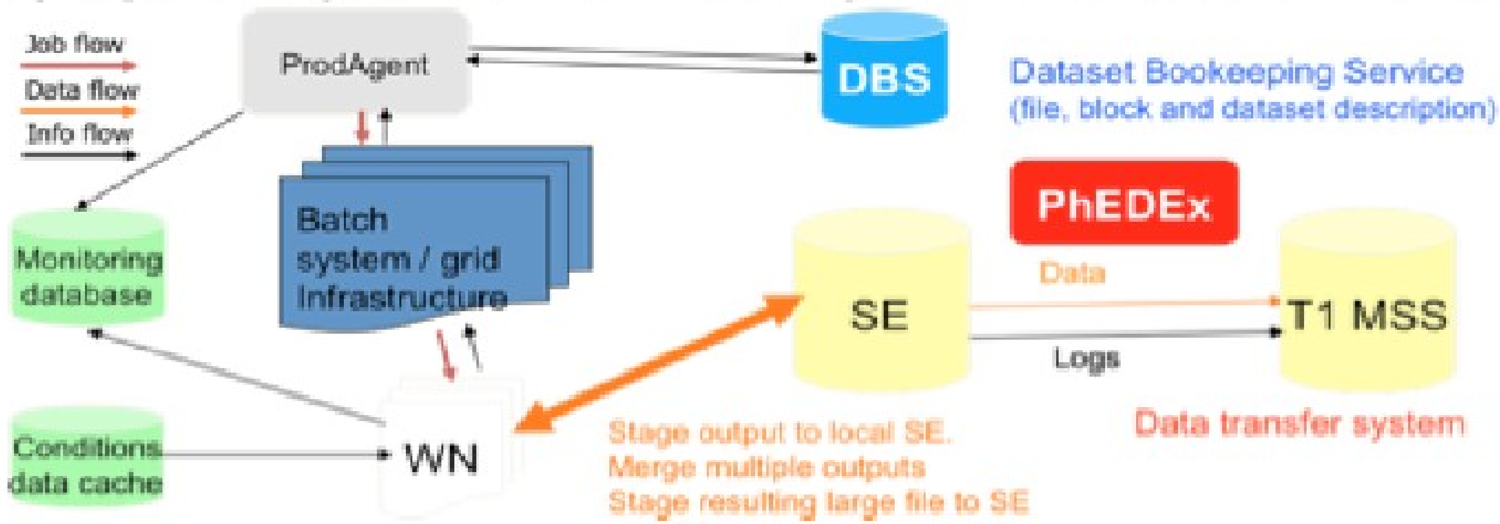
WM and DM evolutions for LHC exp

- All experiments have built their customized workload management systems for production and analysis and data management system on top of the existing grid middleware
  - Very successful in delivering physics results
- But experiments are trying to
  - streamline systems
  - remove unnecessary components
  - ease operations with limited person-power
  - find commonalities
  - scale to higher needs
  - adapt to new technologies
- The CMS workload management system https://indico.cern.ch/contributionDisplay.py?contribId=579&confId=149557
- The ATLAS Distributed Data Management Project, Past and Future https://indico.cern.ch/contributionDisplay.py?contribId=336&confId=149557

The CMS workload management system

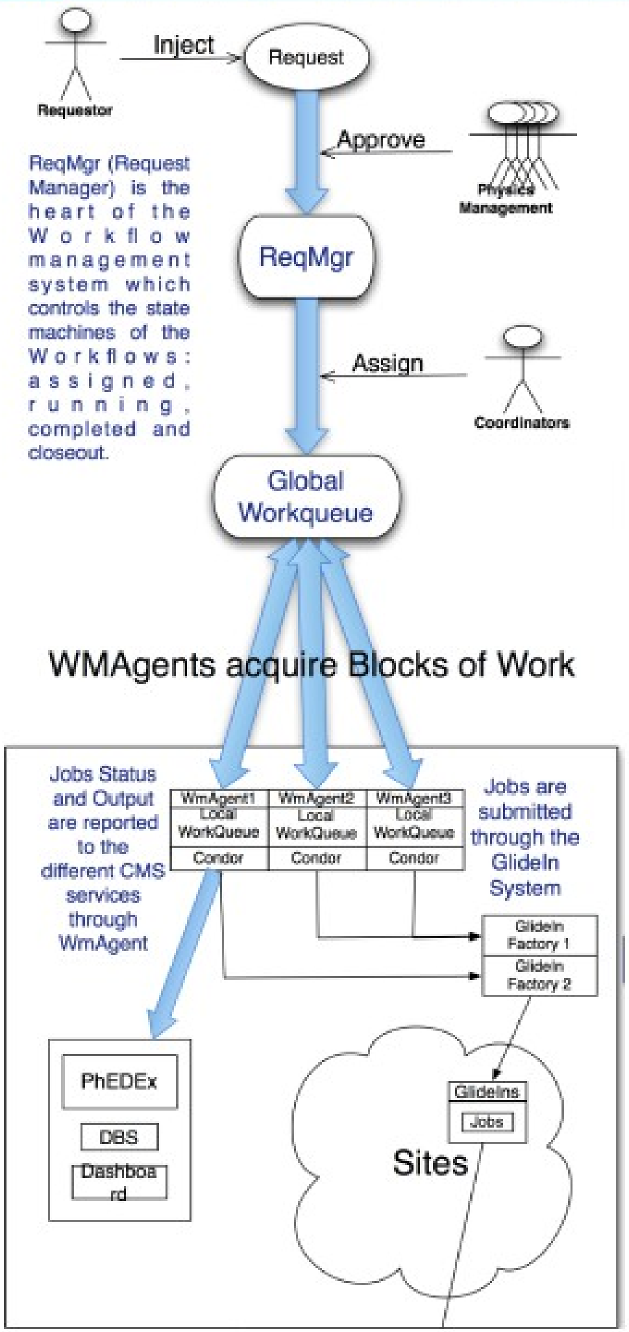
https://indico.cern.ch/contributionDisplay.py?contribId=579&confId=149557

Workload Management (old)



- No workflow repository or request bookkeeping.
- Agent scalability issues:
- Manpower intensive: feed workflow, monitor output, check for errors etc.
- Limit on number of jobs an instance can handle - speed of submission / tracking.
- Designed for producing simulation data rather than processing real data (loss of few % of events no longer acceptable).
- Analysis (users) used different system.
- Very little shared code / experience.

New WM system

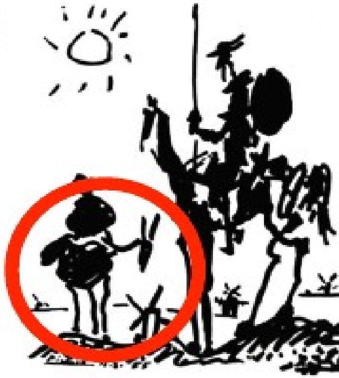


- Consolidate analysis and organized activity
  - Same components but different instances
    - Prevent interference
- Single entry point for requests.
  - Permanently recorded - reproducible
  - Requester can view status.
- Prioritization between requests.
- Approval chain.
- Work distributed automatically & 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.

The ATLAS Distributed Data Management Project Past and Future

https://indico.cern.ch/contributionDisplay.py?contribId=336&confId=149557

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
$\Rightarrow$ Conceptual model document

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

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

Open Protocols - $\mathcal{DQ}_2$Share

$\mathcal{DQ}_2$ via HTTP - `https://bourricot.cern.ch/dq2/share/`

- HDFS cluster as cache back-end
- HTTP redirection to webdav sites (in development)

DQ2 Share

ddo.000001.Atlas.Ideal.DBRelease.v0701010 | Get dataset and files

Search results: **ddo.000001.Atlas.Ideal.DBRelease.v07010104** Displaying 20 on 84 files

Type	Name	Size of all files (Bytes)	Number of files	Date	In the Cache
Dataset	ddo.000001.Atlas.Ideal.DBRelease.v07010104	1 754 037 441	84	2009-07-02 02:18:11	.tar: ✗

wget 19%...	Type	Name	Size (Bytes)	Checksum	In the Cache
☐	File	07010104._0121364.tar.gz	12 446 534	ad:b05542bb	✗
☐	File	07010104._0120760.tar.gz	12 183 186	ad:8bd6b6d3	✓
☐	File	07010104._0121457.tar.gz 100%	13 654 310	ad:2aee4fcc	✓
☐	File	07010104._0120884.tar.gz	11 927 416	ad:654e6578	✓
☐	File	07010104._0121416.tar.gz	18 179 853	ad:ed7f5d07	✗
☐	File	07010104._0121064.tar.gz	12 591 388	ad:a5464922	✓
☐	File	07010104._0120808.tar.gz	12 181 040	ad:203c2a15	✓
☑	File	07010104._0121198.tar.gz	13 648 670	ad:1afc7773	✓
☑	File	DBRelease-7.1.1.4.tar.gz	693 005 997	ad:3db1645e	✓
☑	File	07010104._0120713.tar.gz	11 958 239	ad:fa3b23a6	✗
☐	File	07010104._0120852.tar.gz	12 246 473	ad:a0d72443	✓
☐	File	07010104._0121412.tar.gz	12 455 380	ad:32a180f0	✗
☐	File	07010104._0121226.tar.gz	12 559 921	ad:d06936f5	✗
☐	File	07010104._0121414.tar.gz	12 461 831	ad:b8e44e10	✗

Rucio Base Technologies

Clients & Server

- RESTful APIs
- Web Service Gateway Interface (WSGI) Python server
- Service-Based authentication with token and support of different types of credentials: X509, GSS, etc.

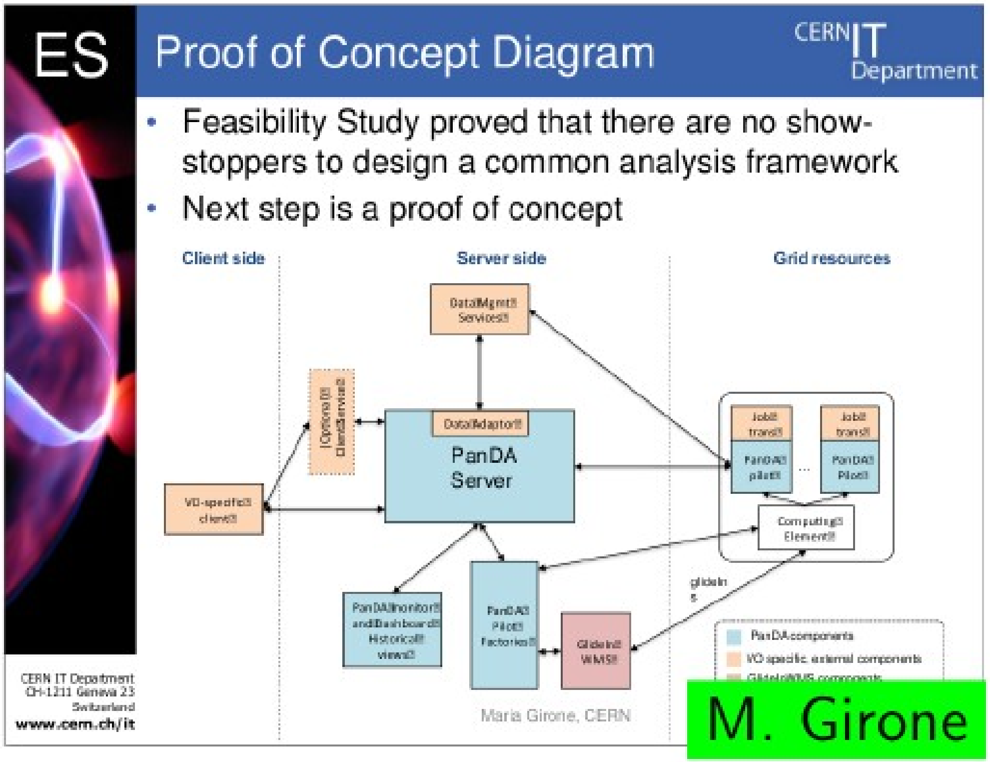
Backend baseline services

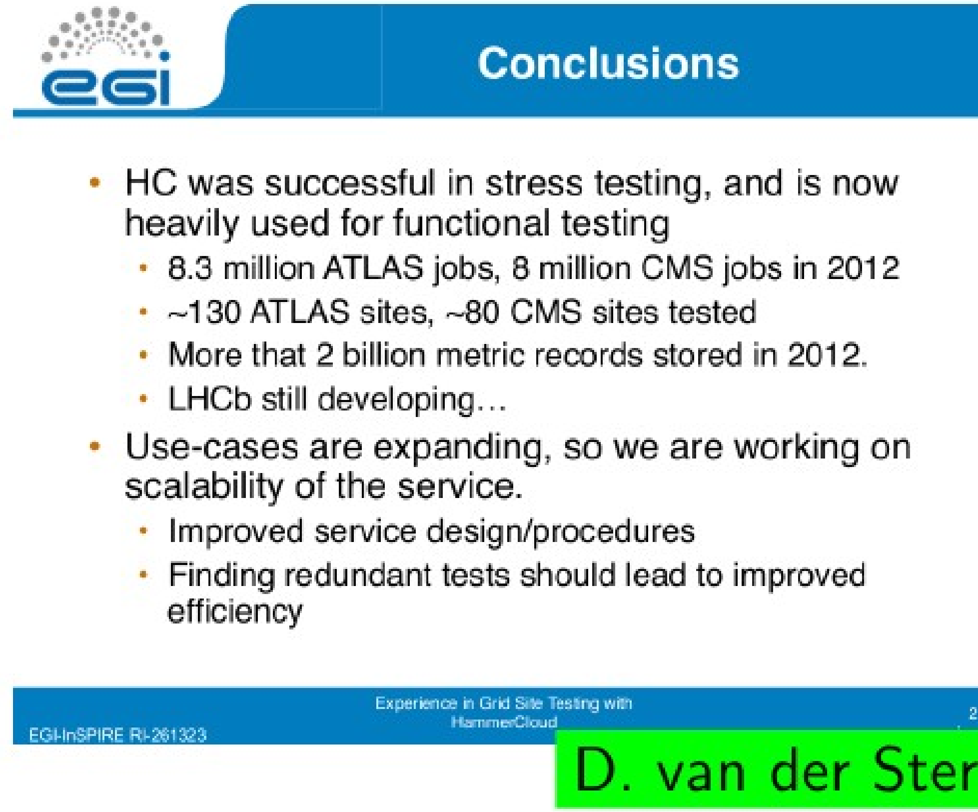
- Relational database management system (Oracle)
  - Use cases: Real-time data and transactional consistency
- Non relational structured storage (Hadoop)
  - Use cases: Search functionality, realtime stats, monitoring, meta-data, complex analytical reports over large volumes
  - See "The ATLAS DDM Tracer monitoring framework"

ATLAS DDM references

- ATLAS Distributed Data Management delivered a working $\mathcal{DDM}$ system to the collaboration in time for LHC data taking
- New services, to manage the complete data life cycle, have been introduced and tuned over the year
- The $\mathcal{DQ}_2$ system is scaling and manages the current load to date
- We continue to optimise and tune the system, but we need to adapt to a changing landscape of distributed computing services
- $\mathcal{DDM}$ team are currently developing a new version $\mathcal{R}$ucio , anticipated for 2013, in order to ensure system scalability, reduce operational overhead and support new ATLAS use cases

Experiment grid improvements, Commonalities

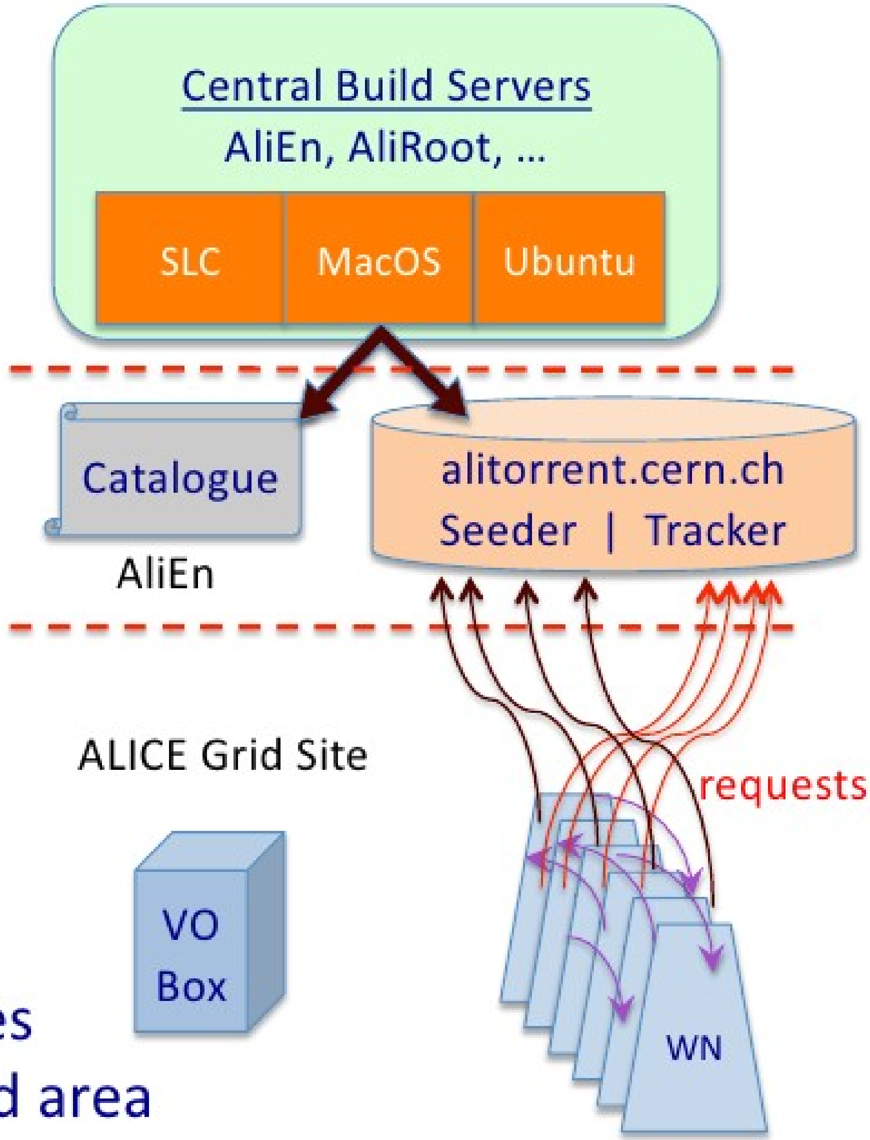




Study if ATLAS Panda is suitable for analysis in CMS - HammerCloud is used among 3 experiments for grid site validation

Employing peer-to-peer software distribution in ALICE Grid Services to enable opportunistic use of OSG resources

- 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



https://indico.cern.ch/contributionDisplay.py?contribId=499&confId=149557

Employing peer-to-peer software distribution in ALICE Grid Services to enable opportunistic use of OSG resources

- Software deployment on shared area
  - Bottleneck & site-level single point failure
  - site-level SW corruption requires admin intervention
- Torrent model → AliTorrent
  - Removes bottleneck & site-level single point of failure
  - Eliminates a site service & reduces site management
  - Performance capabilities meets typical ALICE workflow & site requirements
  - ➢ Eliminates requirement for site-specific VO box
- We have leveraged this capability to demonstrate AliEn workflow for opportunistic use of multiple OSG resources
- AliTorrent is a site-friendly tool for opportunistic (or general) use
  - don't ask the site to "do" something → install or manage a service
  - ask the site to "not do" something → block torrent use

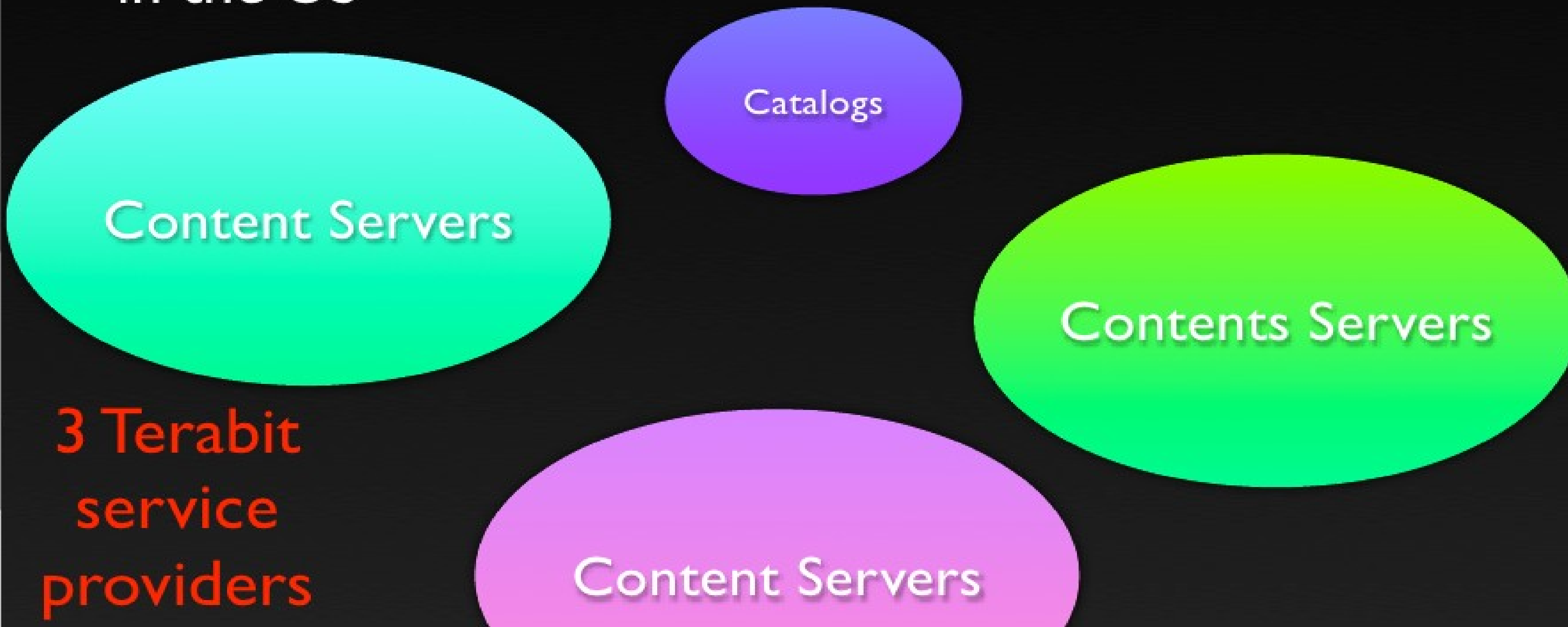
WAN data access

https://indico.cern.ch/contributionDisplay.py?contribId=591&confId=149557

Content Delivery

NETFLIX

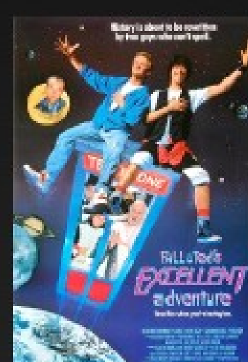
- Why is our problem harder than Netflix?
  - Netflix delivers streaming video content to about 20M subscribers
  - Routinely quoted as the single largest user of bandwidth in the US



By the numbers

- We have a smaller number of clients, less distribution, and higher bandwidth per client
- They have much less data

	NETFLIX	HEP
Bandwidth per client	1.5Mbit	1MB
Clients	1M*	100k cores
Serving	1.5Tbits	0.8Tbits
Total Data Distributed	12TB	20PB



Similar Problems
Not all files
are equally accessed

Forward Physics

LHCOne intro

- High Energy Physics has a lot of data in a highly distributed environment
  - Hard to make many multiple static copies
  - Need to be able to make dynamic replicas and clean up
  - Need to access data over long distances
- Trying to make networking more predictable
  - Enter LHCOne

LHCONE in a Nutshell

- LHCONE was born (out the 2010 transatlantic workshop at CERN) to address two main issues:
  - To ensure that the services to the science community maintain their quality and reliability
  - To protect existing R&E infrastructures against overuse by our traffic
- LHCONE is expected to
  - Provide some guarantees of performance
    - Large data flows across managed bandwidth that would provide better determinism than shared IP networks
    - Segregation from competing traffic flows
    - Use all available resources, especially transatlantic
    - Provide Traffic Engineering and flow management capability
  - Leverage investments being made in advanced networking

Using Xrootd to Federate Regional Storage

http://indico.cern.ch/contributionDisplay.py?contribId=381&confId=149557

Introducing Federations

- Remote access gives us data for *one* site. We need a federation to access all sites.
- Definition of a **federated storage system***:
  - A collection of disparate storage resources managed by cooperating but independent administrative domains transparently accessible via a common namespace.

* From the Lyon workshop on Federated Data Stores: http://indico.in2p3.fr/conferenceProgram.py?confId=5527

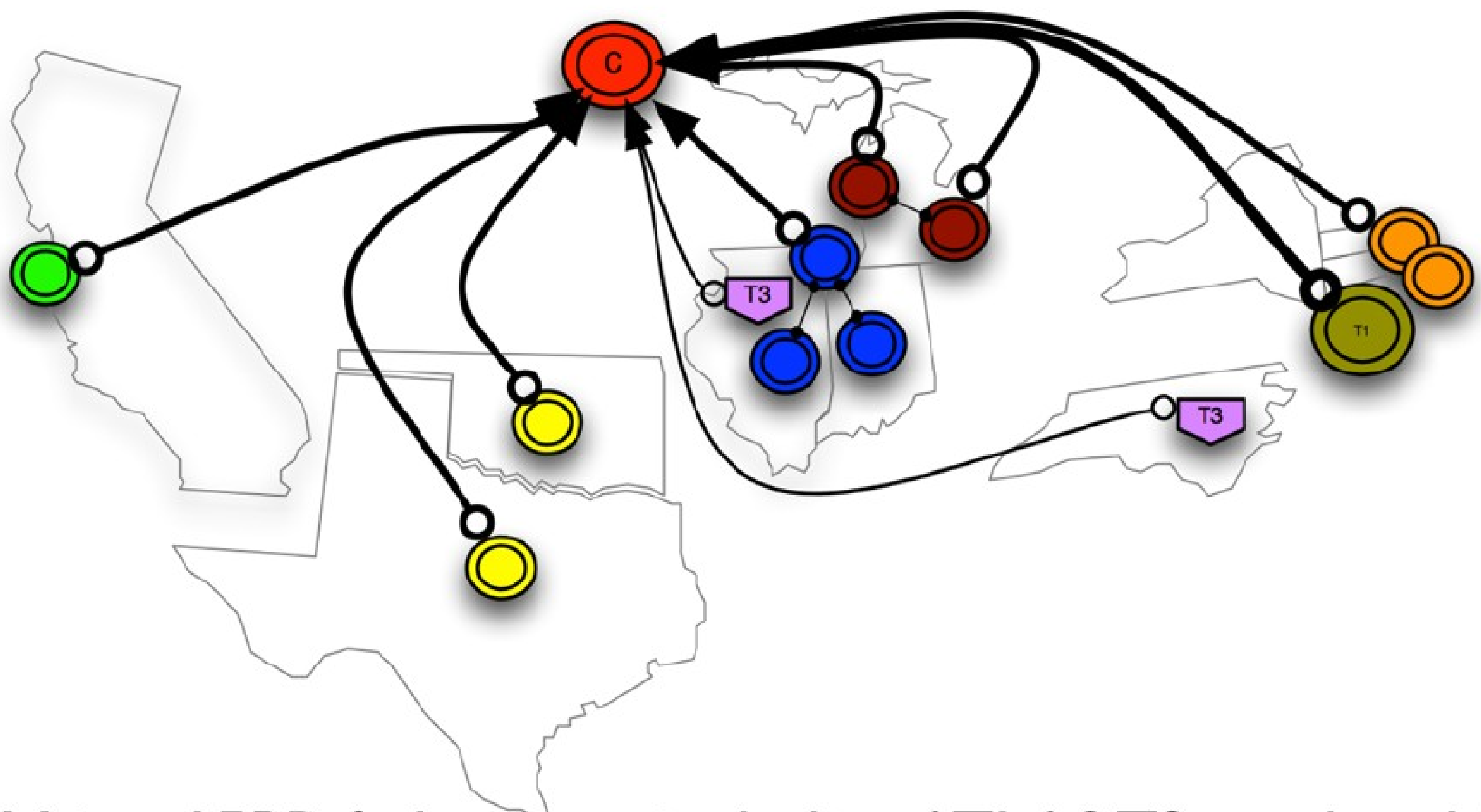
Federations, in practice

- The federation approach has been used by ALICE for many years; used ALIEN, not Xrootd to federate.
- USCMS started federating T2s in 2010; grew to all sites in 2011.
  - Project is named “Any Data, Any Time, Anywhere” or AAA.
- USATLAS started in 2011 and quickly grew to all sites.
  - Project named “Federated Atlas Xrootd”, or FAX.
- Equivalent projects in EU are being worked on.

AAA Deployment

- Currently, redirector at xrootd.unl.edu.
- Includes the FNAL T1 (dCache) and 8 T2s (5 HDFS, 1 dCache, 1 Lustre, 1 L-Store).
- During April, our monitoring recorded:
  - Over 300 unique users,
  - 900K file transfers
  - 300TB moved.

FAX Deployment



FAX is a 15PB federation, including ATLAS T3s and multiple layers of hierarchy.

Cloud Computing in ATLAS

ATLAS Cloud Computing R&D

EFFICIENCY, ELASTICITY

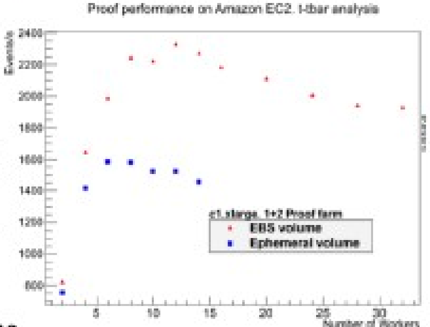
- ATLAS Cloud Computing R&D is a young initiative
  - Active participation, almost 10 persons working part time on various topics
  - Goal: How we can integrate cloud resources with our current grid resources?
- **Data processing and workload management**
  - PanDA queues in the cloud
    - Centrally 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

EGI-InSPIRE RI-261323 Fernando H. Barreiro Megino (CERN IT-ES) CHEP– New York May 2012 2 www.egi.eu

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

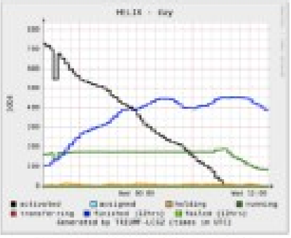
Proof performance on Amazon EC2. t-tbar analysis



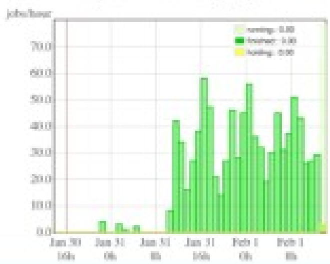
Results

- 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 ± 7038s	8136.6s ± 765.5s



The jobs/hour for all sites progress



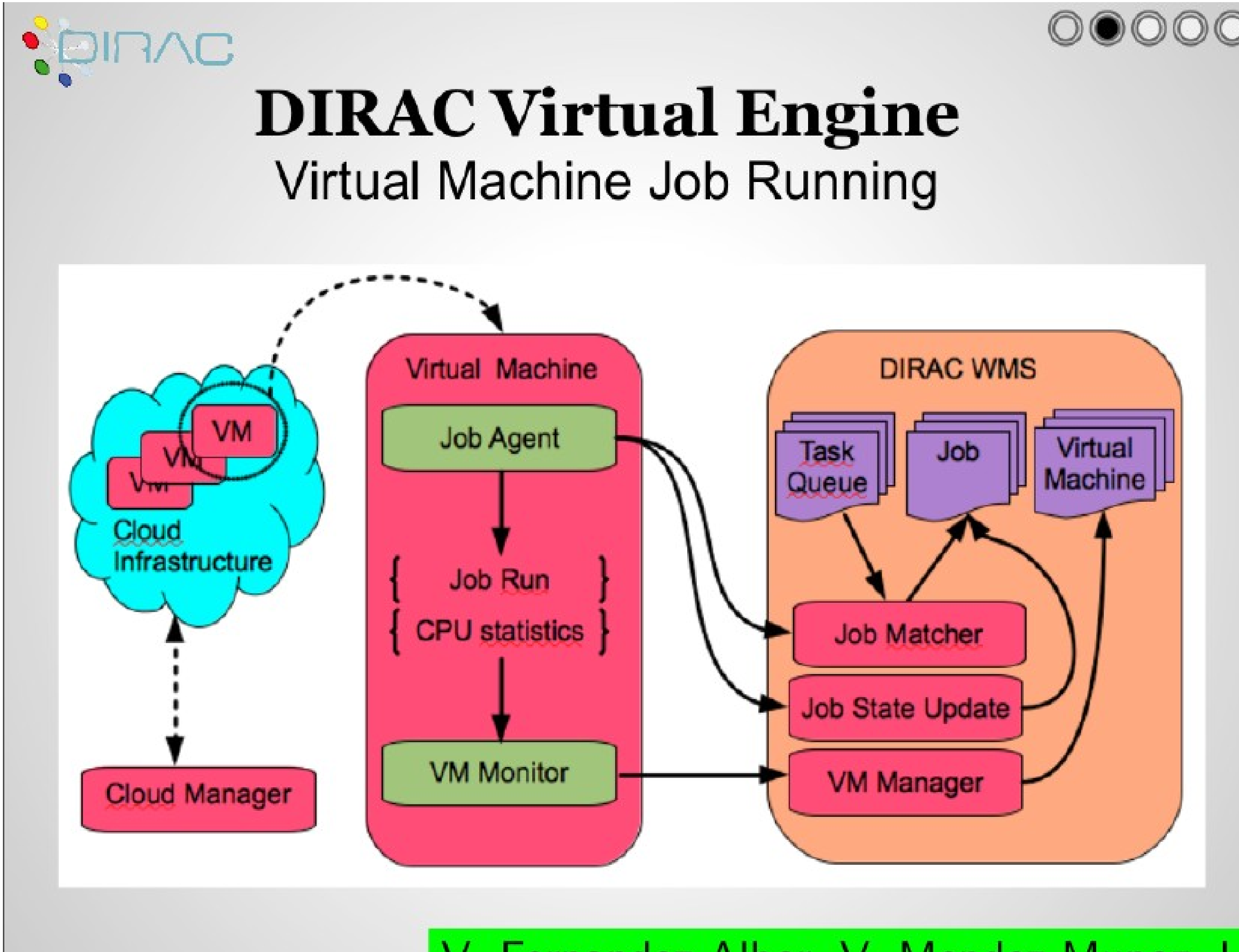
- Wall clock performance cannot be compared directly, since we don't have the same hardware on both sites
  - CloudSigma has ~1.5Ghz of AMD Opteron 6174 per jobslot, CERN has a ~2.3GHz Xeon L5640
- Best comparison would be CHF/event, which is presently unknown

EGI-InSPIRE RI-261323 Fernando H. Barreiro Megino (CERN IT-ES) CHEP– New York May 2012 6 www.egi.eu

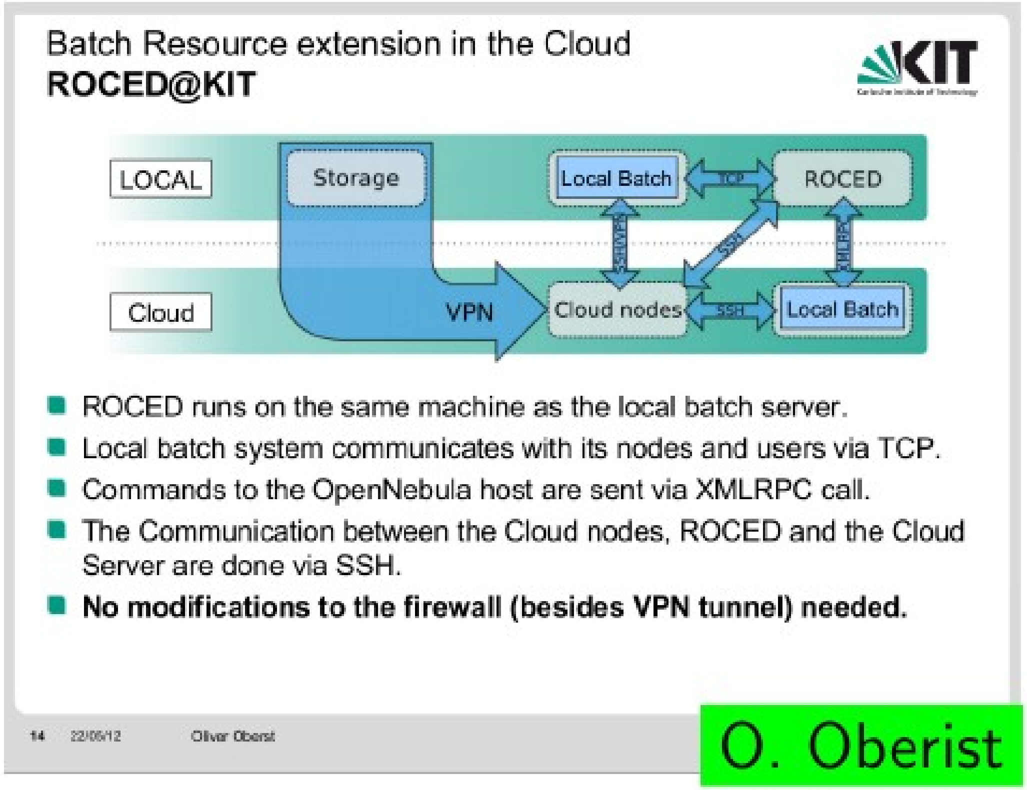
ATLAS with extensive Cloud R&D, tested production on commercial cloud

F. Barreiro Megino, ATLAS

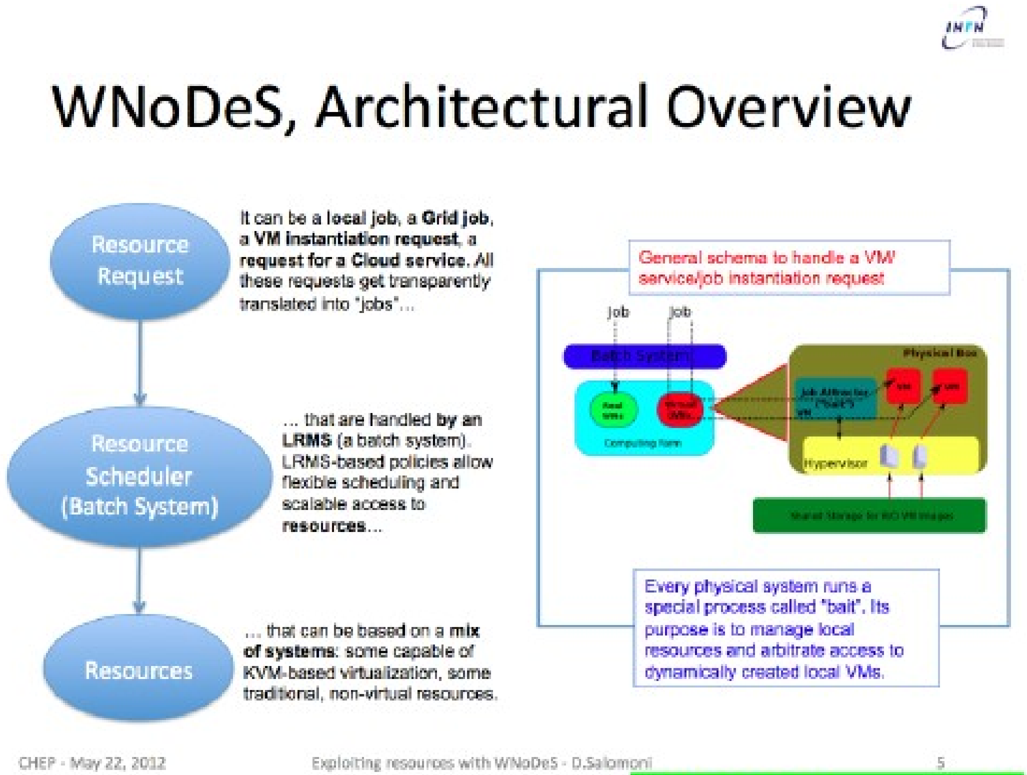
Further Cloud examples



V. Fernandez Albor, V. Mendez Munoz, LHCb



O. Oberist

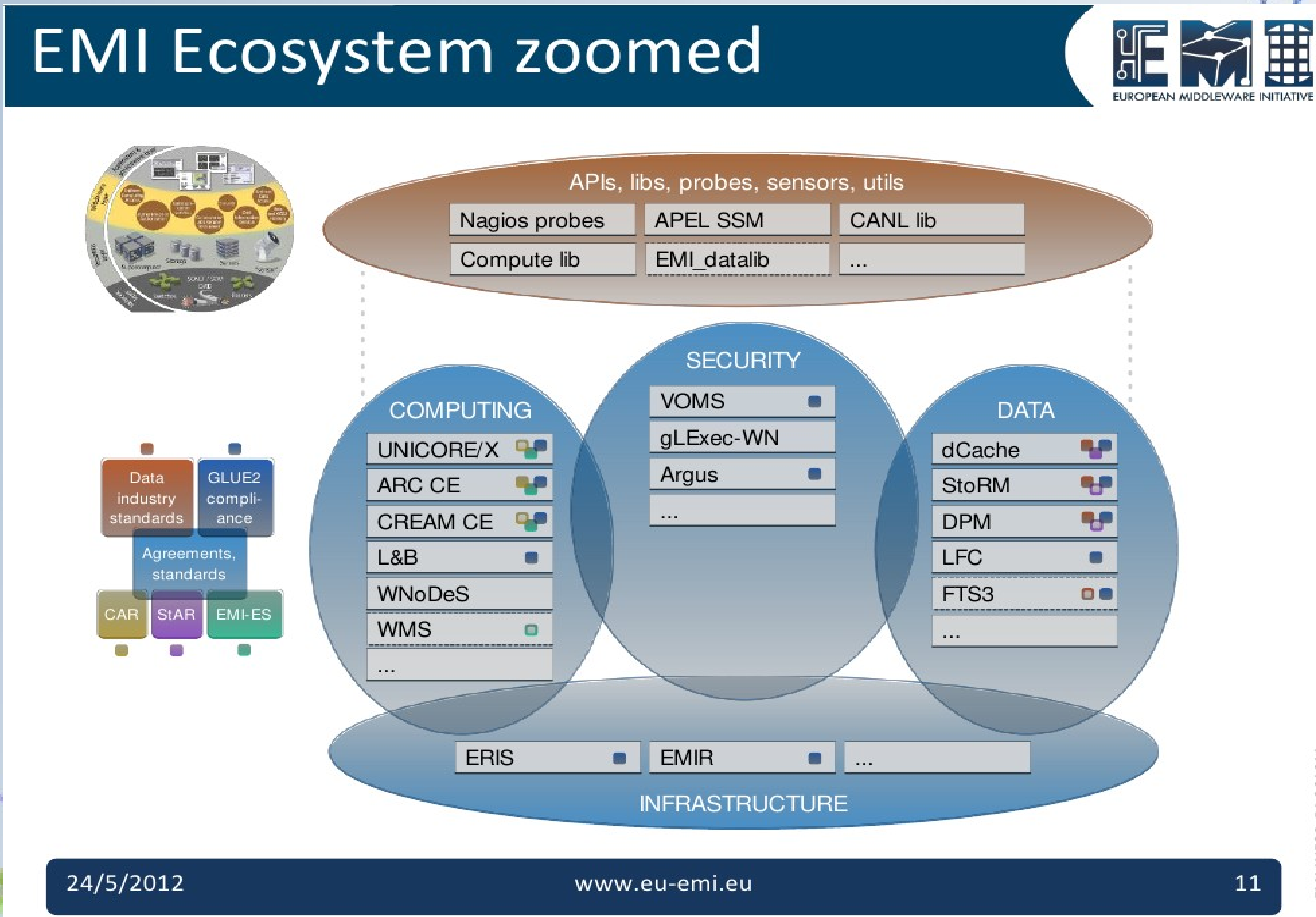


D. Salomoni

- 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

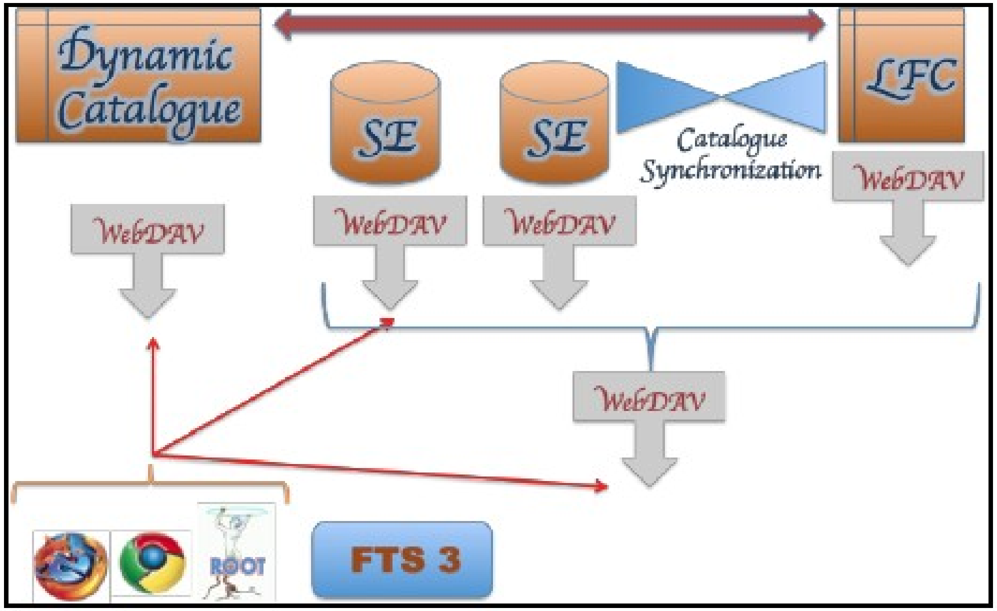
Development Roadmap of the EMI middleware

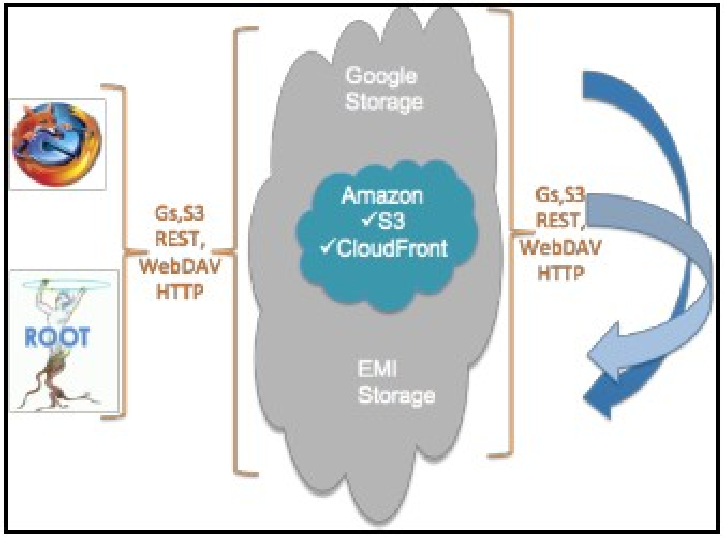
https://indico.cern.ch/contributionDisplay.py?contribId=273&confId=149557



Show-case 1: Data Industry Standards

- Industry standard protocols for accessing SEs and the catalog
  - DPM and dCache ready for NFS4.1
  - HTTPS offered by DPM, StoRM and dCache
  - WebDAV support in DPM and dCache
  - WebDAV support being developed in FTS3 and LFC
- Vital part of the greater vision for EMI Data





The last peak (Y3 development plans)

General strategy:

- Complete product developments:
  - FTS3, GFAL2
  - STS
  - EMI Datalib
- Product hardening, focus on usability
- Integration and adoption of common EMI solutions (EMIR, CANL)
- Migration plans, compatibility

The Open Science Grid – Support for Multi-Disciplinary Team Science – the Adolescent Years

https://indico.cern.ch/contributionDisplay.py?contribId=475&confId=149557

Maturing

- OSG is being supported for another 5 years.
  - Strong support from DOE and NSF.
- Endorsement to not only continue focus on physics but also continue broad engagement of other sciences making use of OSG.
  - Sustain services for LHC and make significant contributions to LHC upgrade.
  - Extend, simplify, and adapt services and technologies for other sciences.
  - Continue community partnerships and facilitation of peer infrastructures in Europe, South America, Asia, Africa.

Sustain Services & Expand Science Communities

- ✦ LHC
  - ★ Continued focus on LHC – support for ATLAS, CMS, ALICE USA distributed computing in the US.
  - ★ Active /proactive contributions on behalf of US LHC to WLCG – to TEG reports and implementation follow ons.
  - ★ Prepare for LHC shutdown and upgrade.

- ✦ Embrace future physics, nuclear physics, astrophysics experiments: Belle II, DES, EIC, LSST, SuperB…

(will explain these..)

Looking towards and beyond 2015 – Computer Science Research

- ✦ OSG's existing capabilities are effective but basic and primitive.
  - ★ Improvements will rely on external research, development and contributions.
- ✦ Integrate static resources with dynamically allocated resources (like clouds).
- ✦ New globally capable, usable, and integrated frameworks for collaborative environments : data, security, workflows, tools for transparency, diverse resource resources.

- ✦ http://osg-docdb.opensciencegrid.org/0011/001106/001/OSG-CSresearchNeeds.pdf

Posters of interest

- Hybrid C++/Python components for physics analysis and trigger
- Preparing for the new C++11 standard
- Improvements in ROOT I/O
- XRootD client improvements
- ROOT: High Quality, Systematically
- Computing On Demand: Dynamic Analysis Model
- The PhEDEx next-gen website
- From toolkit to framework - the past and future evolution of PhEDEx
- Belle II Data Handling System
- EMI-european Middleware Initiative
- Workload management in the EMI project
- A General Purpose Grid Portal for simplified access to Distributed Computing Infrastructures
- Improving Geant4 multi-core's performance and usability
- The Geant4 Virtual Monte Carlo
- GFAL 2.0 Evolutions & GFAL-File system introduction
- Multi-threaded Event Reconstruction with JANA
- The WLCG Messaging Service and its Future