

Track 3 summary



21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Latchezar Betev for T3 (should not be confused with a T3)
17 April 2015

T3 keywords and people

- Databases/Data store and access
 - **A06** Data stores
 - **T04** Data handling/access
 - **T05** Databases
 - **T06** Storage systems
- Conveners
 - Gancho Dimitrov (CERN), Christopher Pinkenburg (BNL),
Alaistair Dewhurst (RAL), Giacomo Govi (CERN), Latchezar
Betev (CERN)
- Session chairs
 - Barthelemy von Haller, Laurent Aphenetche, Shaun De
Witt, Luca Magnoni, Latchezar Betev
- Track scientific secretary
 - Yasushi Watanabe (RIKEN)

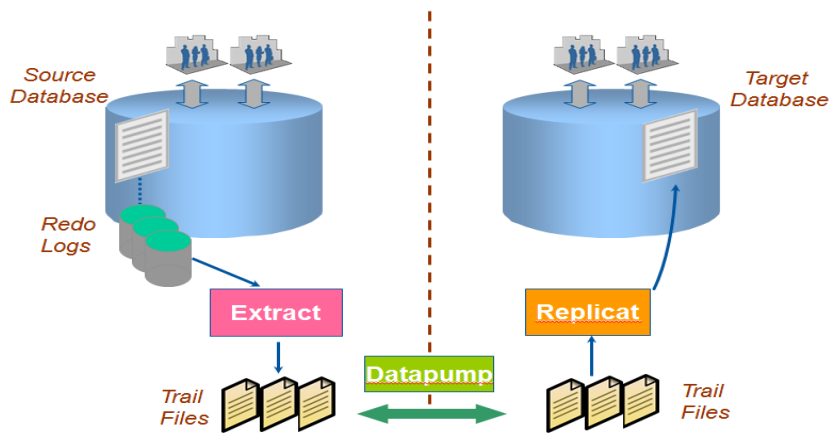
T3 stats

- 5 oral presentation sessions with 35 talks
- 2 poster sessions with 36 posters
- Room occupancy – *full* – about 55 people/session



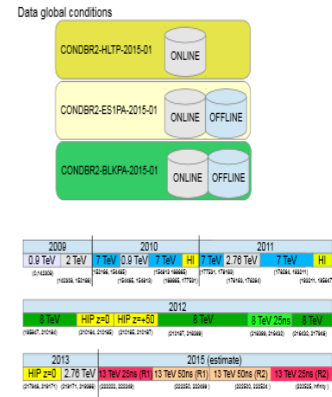
Databases – tuning for LHC Run2

- Evolution of ATLAS conditions data and its management for LHC run-2 (M.Boehler)
 - Logical group split – Run 1, Run 2, MC
 - Tags and global tags for processing



Summary:

- ▶ detector conditions needed both for data reconstruction and MC production
- ▶ 3 different global conditions tags need to be maintained - for data
- ▶ IOV dependent MC global conditions tag allow one single tag for different environments

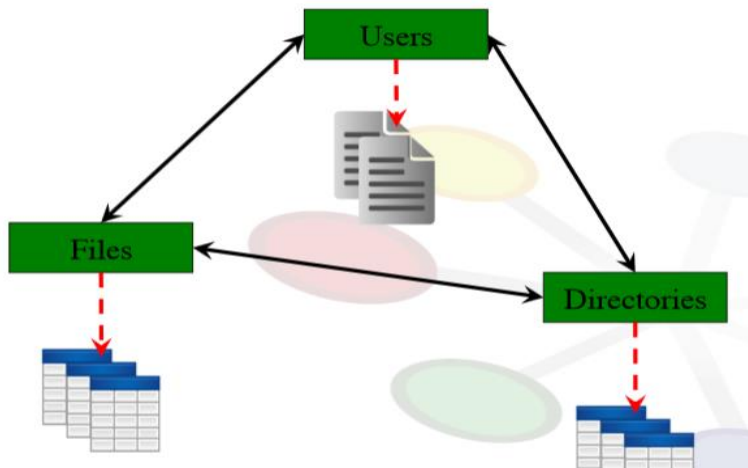


- Evolution of Database Replication Technologies for WLCG (S. Baranowski)

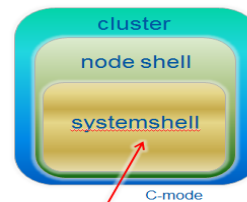
- From ORACLE streams to Dataguard/Golden Gate
- Improve replication granularity, stability and cost of replication

Databases – tuning for LHC Run2

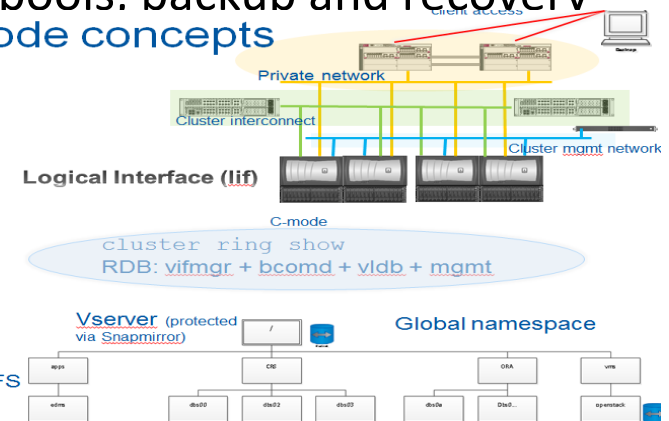
- Experience in running relational databases on clustered storage (R. Gaspar Aparicio)
 - Use of NetApp C-mode clusters for critical DB storage
 - Full setup, monitoring, FS movement. flash pools. backup and recovery



A few C-mode concepts



Logging files from the controller no longer accessible by simple NFS export



- Federating LHCb datasets using the Dirac File Catalog (C. Haen)
 - Logical representation of files and replicas
 - Used by all Grid application, top performance and Run2 ready
 - Replaces LFC, migration done

DB - New technologies evaluation

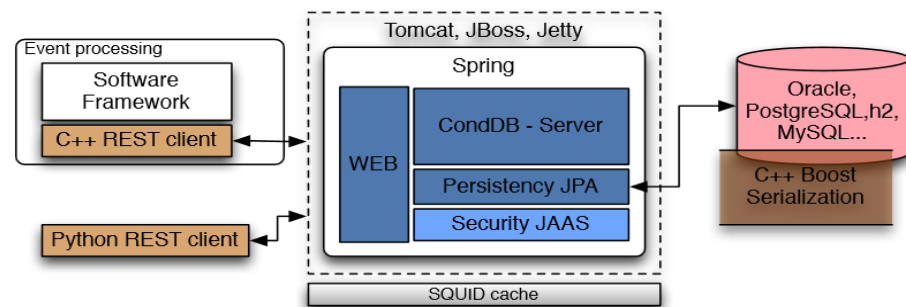
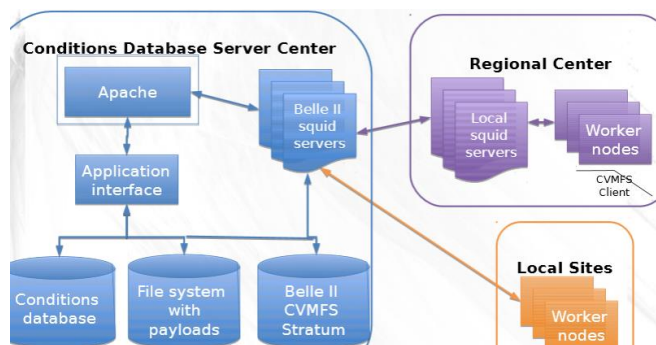
- NoSQL technologies for the CMS Conditions Database (R. Sipos)
- Evaluation of NoSQL databases for DIRAC monitoring and beyond (F. Stagni)
- Studies of Big Data meta-data segmentation between relational and non-relational databases (M.Golosova)
 - Performance evaluation of various products: InfluxDB, OpenTSDB, ElasticSearch, MongoDB, Cassandra, RIAK
 - Goal - optimize search, increase speed, aggregate (meta)data, time series
 - All evaluations show promising and better results, compared to standard SQL technologies
 - Some are about to be put in production
 - More work ahead...



Conditions databases for the future

- Designing a future Conditions Database based on LHC experience (A. Formica)
 - Common solution for ATLAS/CMS
 - Standard technologies, simple schema and opaque BLOBs
 - middleware in Java, REST API, same interface online and offline.

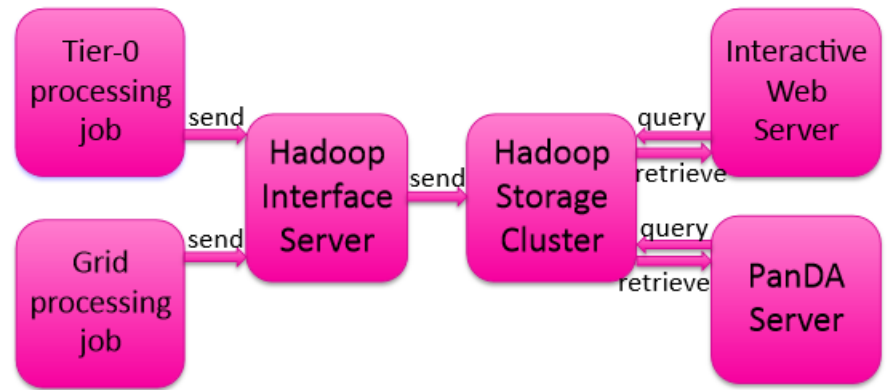
• Java based prototype



- The Belle II Conditions Database (M. Bračko)
 - Standard technologies – Apache, Squid, CVMFS
 - Payload stored in files
 - Simple GET/PUT/POST schema, fully integrated in processing framework

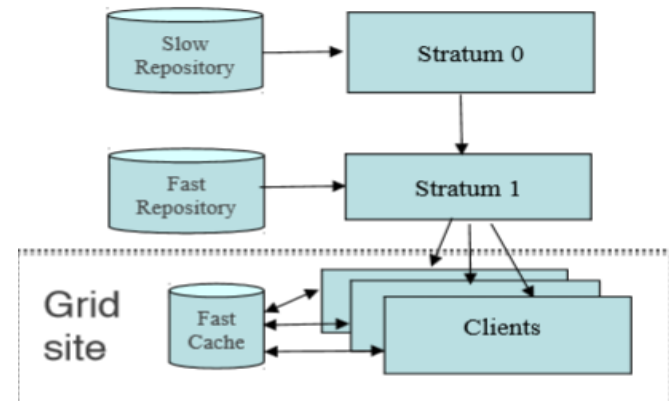
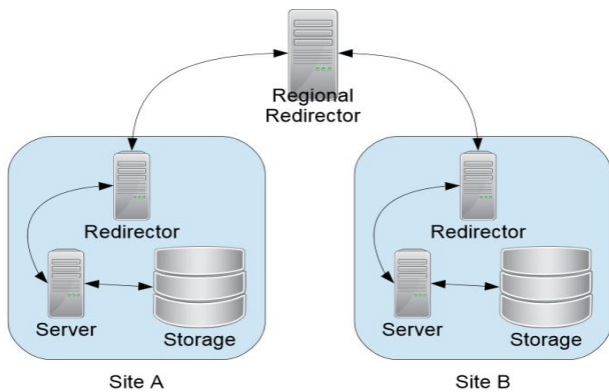
Event Indexation and storage

- The ATLAS EventIndex: architecture, design choices, deployment and first operation experience (D. Barberis)
- Distributed Data Collection for the ATLAS EventIndex (J. Sanchez)
 - Complete catalogue of *all* events in ATLAS
 - Billions of events amounting to 2TB of raw information for Run 1 (twice as much for Run20)
 - Web Service search - seconds for a search on a key, minutes for complex queries
 - Uses Hadoop at CERN
 - ActiveMQ as messaging protocol to transport info between producers and consumer, data encoded in JSON



Storage systems

- Operational Experience Running Hadoop XRootD Fallback (J. Dost)
 - allows a site to offload the redundancy from the local system onto the XRootD federation redundancy
 - Saves site storage capacity, in production and will be expanded

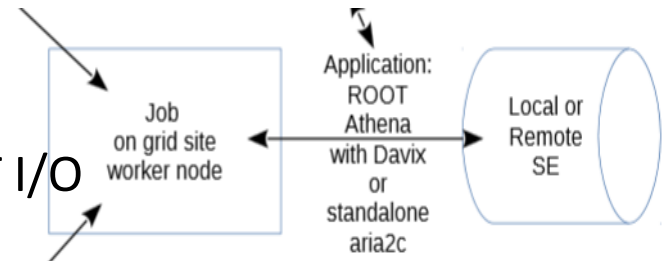


- Engineering the CernVM-FileSystem as a High Bandwidth Distributed Filesystem for Auxiliary Physics Data (J. Blomer)
 - Aimed at shared datasets of medium size, low hit ratio
 - Works well with 'alien' caches, for example site SEs

Data access and protocol performance

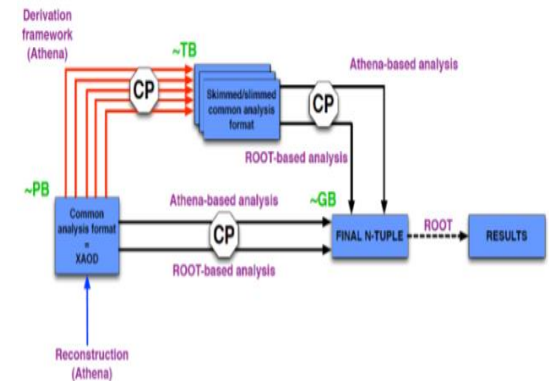
- New data access with HTTP/WebDAV in the ATLAS experiment (J. Elmsheuser)

- Goal – unique access protocol everywhere, aria2c for staging and Davix plugin for ROOT I/O
- Comparisons of data access speed (xrootd/dcap/webdav) show similar results



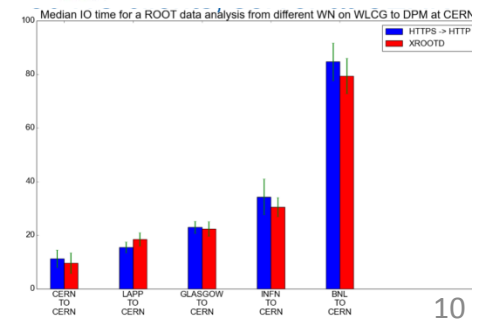
- ATLAS I/O Performance Optimization in As-Deployed Environments (T. Maier)

- New xAOD format with optimized structure for fast access
- Solves the problem of format proliferation



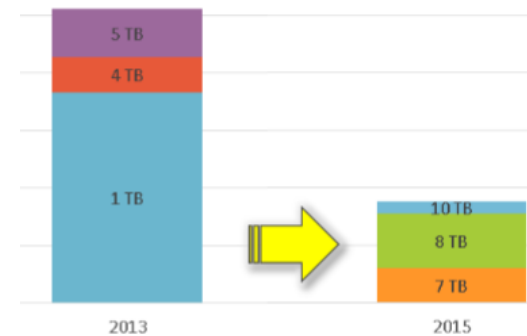
- Protocol benchmarking for HEP data access using HTTP and Xrootd (O. Keeble)

- DPM storage, tested xrootd and http over WAN and LAN
- Comparable performance in all conditions



Storage operations

- Mean PB to Failure -- Initial results from a long-term study of disk storage patterns at the RACF (C. Hollowell)
 - 12K drives in RACF, managed by xrood, dCache
 - Annual failure rate measured by R/W patterns, highest for extended write workload 0.94%, half of that for extended read
- Experiences and challenges running CERN's high-capacity tape archive (E. Cano)
 - Large-scale (100PB) media migration 51K to 17K tape cartridges
 - Keep data safe – new CASTOR architecture + automatic tape Incident Identification System
- Disk storage at CERN (L. Mascetti)
 - Multitude of managed storages – AFS, CASTOR+EOS, Ceph, CERNBox
 - More than 250PB, distributed over CERN and Wigner, ready for Run2



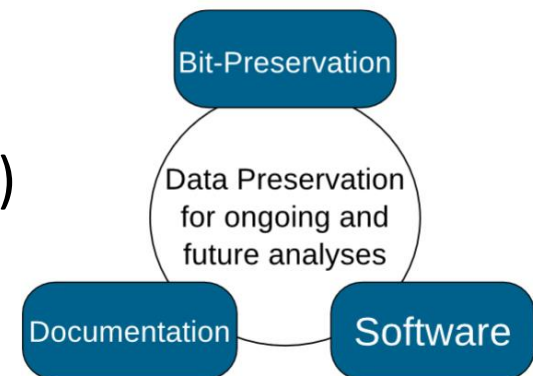
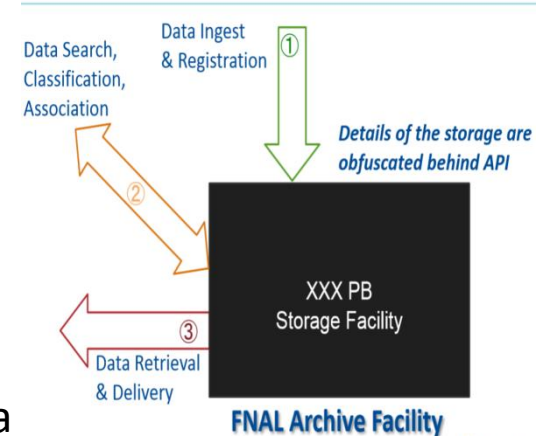
CASTOR
CERN Advanced STORage manager



EOS

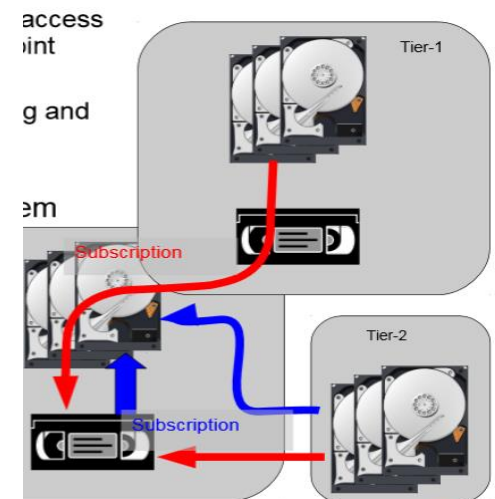
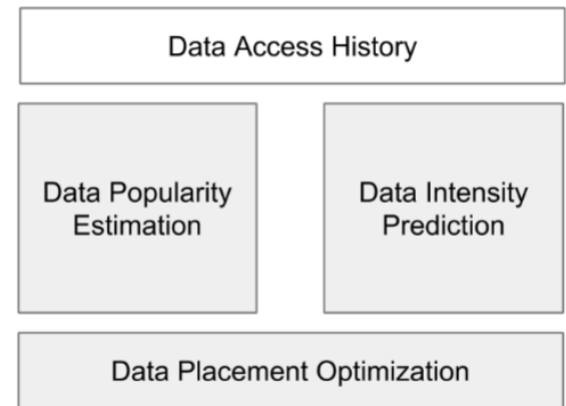
Archival and data preservation

- Archiving Scientific Data outside of the traditional High Energy Physics Domain, using the National Archive Facility at Fermilab (A. Norman)
 - Tools and strategy for archiving data with varying size, structure, format, complexity, provenance
 - A toolset comprising file transfer service for arbitrary data types, sequential access via metadata catalogue+search, retrieval and delivery protocol
- Data preservation for the HERA Experiments @ DESY using dCache technology (K. Schwank)
 - Assure the storage of unique documentation, software and data beyond the end of experiments
 - Provide the tools and methods to go from experiment-specific to institutional archiving solutions
- Deep Storage for Big Scientific Data (D. Yu)
 - Store multi-PB of non-compressible data into permanent storage
 - Efficient retrieval and monitoring, build on a 2-tier disk and tape combination



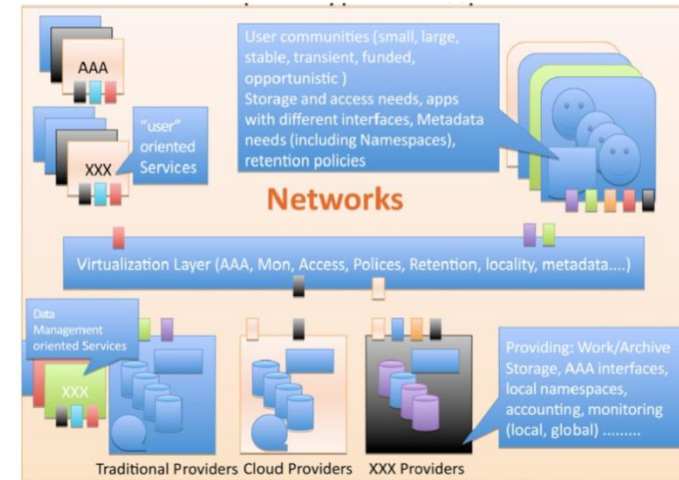
Resources use optimization

- Disk storage management for LHCb based on Data Popularity estimator (M. Hushchyn)
 - Based on complex analysis of data access patterns
 - Calculates the probability that a given data set will not be used in future., used for data placement optimization
 - Interesting results - method allows to save up to 40% of disk space and decrease downloading time up to 30%
- Pooling the resources of the CMS Tier-1 sites (C. Wissing)
 - Refinement of the roles of the storage at (predominantly) the T1s
 - Allows more flexible workload and more democratic access to resources

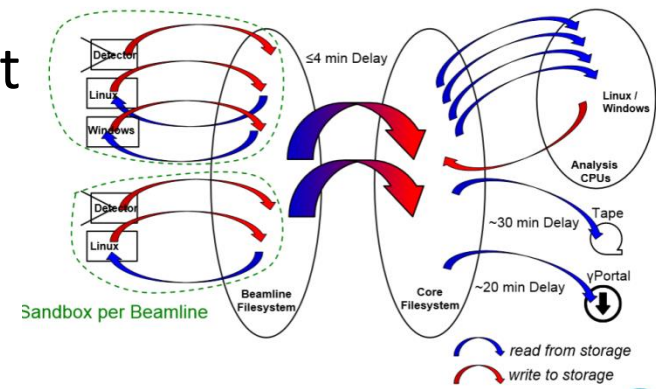


Storage architecture

- Architectures and methodologies for future deployment of multi-site Zettabyte-Exascale data handling platforms (M. Delfino)
 - Looking forward to reaching Exa (10^{18}) bytes in the next 10 years
 - The ZEPHYR project - architecture of the future data management

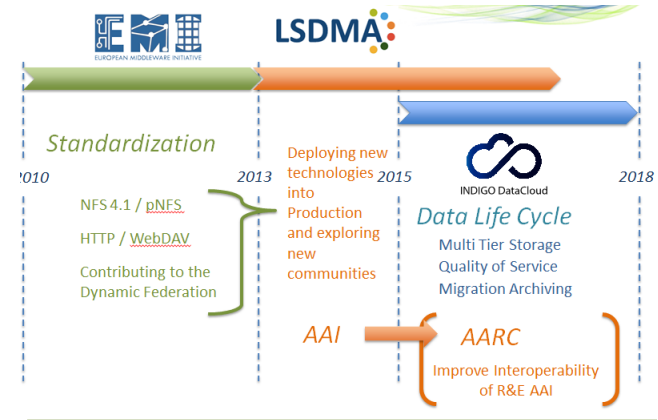


- Architecture of a new data taking and analysis infrastructure and services for the next generation detectors of Petra3 at DESY (M. Gasthuber)
 - How to store the results of 100s of experimental groups with no computational experience
 - From 'cradle to grave' system using NFSv3, SMB, ZeroMQ Tunnel and web portal

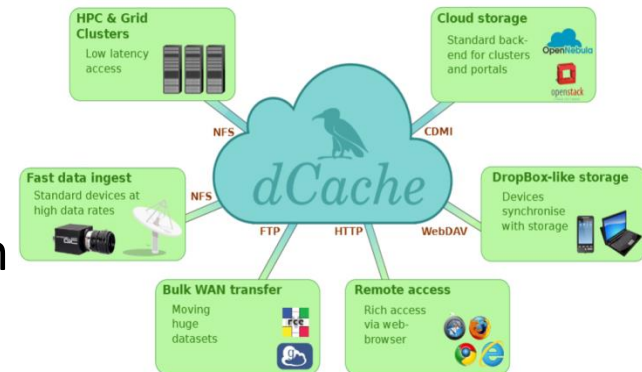


Storage management

- dCache, evolution by tackling new challenges (P. Fuhrmann)
 - Focus on efficiency and evolution of the product, site activities and standardization
 - NFS 4.1/pNFS, HTTP/WebDav, Dynamic federation
 - New funding – Data Life Cycle support and Software Defined storage
- dCache, Sync-and-Share for Big Data (P. Millar)
 - Dropbox alternative, integrated with the existing infrastructure at DESY
 - QoS, sharing, synching/non-synching.. with OwnCloud overlay



The scientific cloud vision



Storage management

- EOS as the present and future solution for data storage at CERN (A. J. Peters)
 - New project – started in 2010, in production since 2011
 - Uses cheap hardware, strong security, in-memory namespace, multiple protocol support
 - 3K users, 160KHz stat on namespace
 - Gemstone evolution – Beryl (current), Citrine (xrootd v4+infrastructure aware), Diamond (R&D)
 - WebDav access for Android, IOS, OSX, Windows; FUSE mount, sync&share, I/O proxy for server security
- Archiving tools for EOS (A. J. Peters)
 - Rationale – extend virtual available storage space, avoid ad-hoc user solutions
 - Integration of tape storage, with efficient data movement and rules

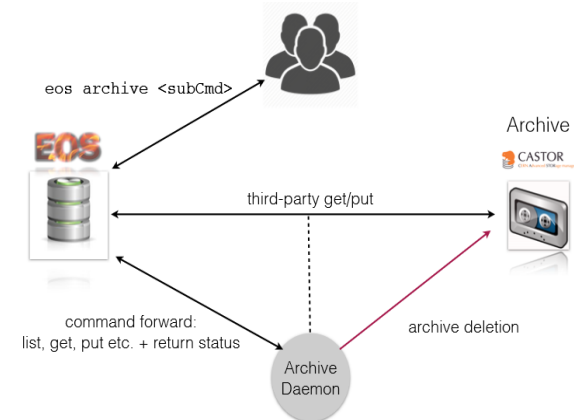
- Run 2 Production Version **BERYL Aquamarine** $\geq 0.3.107$
archiving tools: [CHEP talk 298](#)



- XRootD V4 based Version, infrastructure aware **CITRINE** 0.4.x



- Ceph based R&D bundle **DIAMOND**



CEPH-fest




- Enabling Object Storage via shims for Grid Middleware (S. Skipsey)
 - Direct Object store interface as alternative to “heavy” POSIX layers
 - Test interface for gfal -> RADOS -> Ceph
- POSIX and Object Distributed Storage system – performance Comparison studies and real-life usage in an experimental data taking context leveraging OpenStack/Ceph (M. Poat)
 - Rationale – re-use online nodes to build a reliable user files storage system
 - Tests with OpenStack Swift an Ceph, working better in I/O concurrency
 - Test with Ceph and CephFS under extreme loads – remarkable stability, versatility and recovery potential




...and more Ceph

- Current Status of the Ceph Based Storage Systems at the RACF (H. Ito)
 - Rationale – reliability as a direct object store without FS, installed on old, non-reliable storage
 - Use cases – RadosGW (object store), direct Block Device (dCache storage pool), CephFS (GridFTP, WebDav, xrood)
- Ceph-based storage services for Run2 and beyond (A. J. Peters)
 - 3PB in production for OpenStack Cinder volumes/Glance images, coming soon for CVMFS Stratum0
 - Significant performance boost with SSDs journals, critical for ceph-mon
 - Tuned to assure stability in all kind of (improbable) failure scenarios, further work needed for very large (tens of PB) sizes


**KEEP
CALM
AND
Read Your
CEPH Book**

...and even more Ceph

- Integrating Ceph in EOS (A. J. Peters) 
 - Comprehensive integration of all type of storage devices – KineticIO type, normal and SSD cached disks, for optimal performance for all types of applications
 - KineticIO storage testing – client talks directly to device (shingle disk technology), simple API with full object PUT/GET

36 Posters

No time to cover these



... which is a shame.

Winner of the T3 poster session



**Data Management
System of the DIRAC
Project (ID #325)**

Christophe Haen (CERN)

Congratulations!



Thanks to all presenters for the interesting and engaging talks and sticking to the time limit

Many thanks to our OIST hosts and the CHEP local organizing committee for the great facilities and support throughout the T3 sessions