Track 1 - Data and Metadata Organization, Management and Access

Martin Barisits (CERN), Diego Davila Foyo (UCSD), Mike Kirby (FNAL) , Mario Lassnig (CERN)
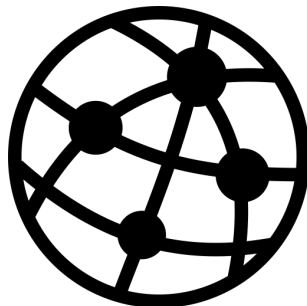
Thank you!

- 7 sessions
- 41 oral presentations
- 4 track conveners
  - Diego Davila (UCSD), Mike Kirby (FNAL), Mario Lassnig (CERN), Martin Barisits (CERN)
- Very interesting content
- Lots of discussions
  - Unfortunately, had to be cut short sometimes
- Very disciplined speakers - always on time

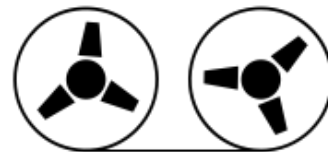The following slides are a personal & biased view, apologies!
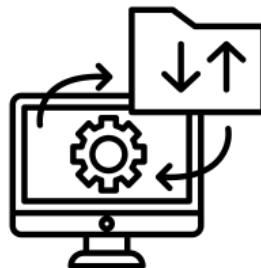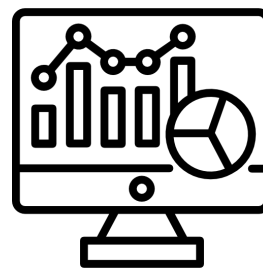
Sessions
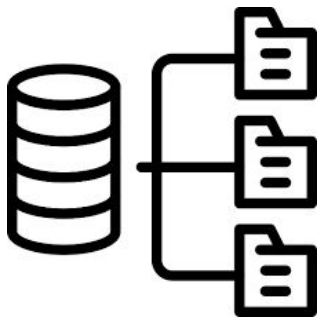
Storage

Networks

Clouds & Caches

Tapes

Databases & Metadata

Data Management

Analytics & Benchmarks

Storage

- **Storage overview talks: dCache, ECHO, EOS**
- **Specific features discussed**
  - Erasure coding for improved data protection in XRootD
  - Seamlessly accessing remote storage using classic POSIX
  - Exploration of continuity & recovery mechanisms in Ceph
- **Common themes emerge**
  - Storage production deployments are stable and efficient
  - Always ongoing performance improvements (developers) and tuning (sites)
  - Consolidation of access protocols and APIs
  - Tokens/AuthNZ support: Systems are slowly but surely converging
- **Object store backends**
  - Becoming strong contenders against classical file systems

Networks

- Common theme
  - Becoming more obvious that network is also a scarce resource
  - Impact to compute models is real
- Stop thinking of network as a black-box
  - Need better ways to monitor,
  - throttle,
  - and optimise;
- Multiple R&D projects were presented to address these issues
  - Caches/CDNs to reduce network traffic
  - Bandwidth reservations & tuning amount of data transfers
  - Identifying network flows through tagging

J.Balcas "Automated Network Services for Exascale Data Movement", CHEP 2023

How scitags work
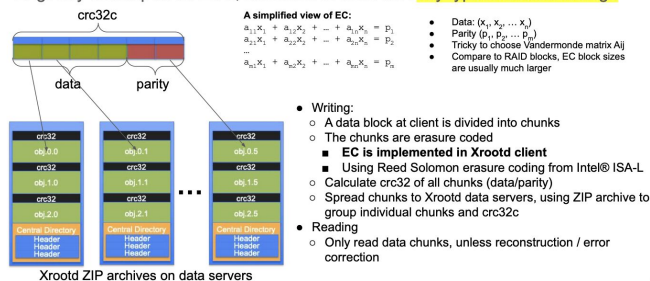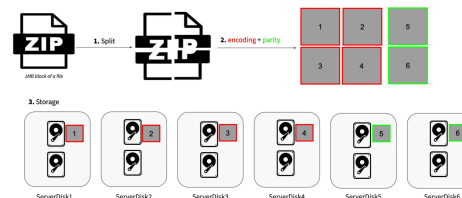
Clouds & Caches

- Cloud storage increasingly important, but challenging to include in HEP workflows
  - S3 Gateway to gap HEP & commercial world
  - Rucio extensions to AWS, GCS, SEAL
    - → CA conflicts are painful
  - Storing LHC data through RNTuples
- Critical to understand Caches
  - Understanding & predicting resource usage trends @ SOCal Cache
  - Understanding data access patterns for dCache
    - Predict popular datasets & pin them

S3 Gateway Architecture

- Based on **XrdCl** (xrootd client) http plug-in
  - Uses Davix, an HTTP SDK developed at CERN
    - Very reliable and supported
    - Performs better than most commercial SDK's
- Bridges HEP & commercial world security
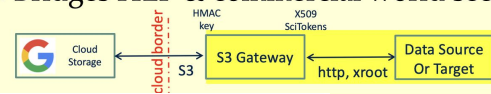
Predicted Volume of Cache Hits
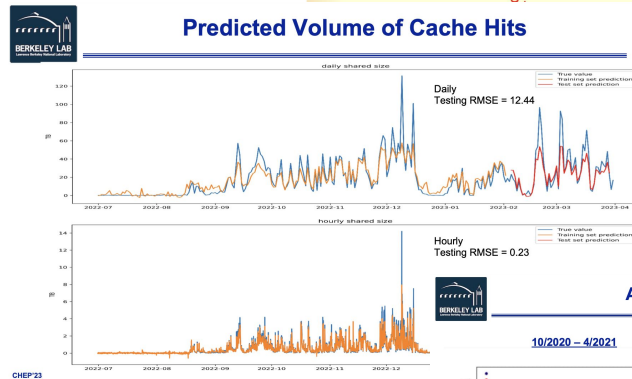
Access Forecasts
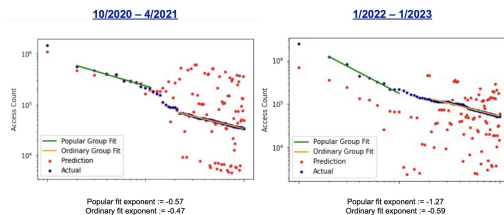
Tapes

- HTTP TAPE REST API
  - The end of SRM is within reach: first production deployments exist (EOSCTA, dCache/HPSS)
- DISK instead of TAPE
  - TCO/Comparison between a disk-based archive and tape-based archive
  - Interesting proposition (up to half-price) given current hardware estimates
- CTA
  - Free and open source community tape solution
  - Supports both EOS and dCache frontends
  - CBACK: User data backup to CTA @ CERN
- ATLAS Data Carousel
  - Data on demand & Smart Archive
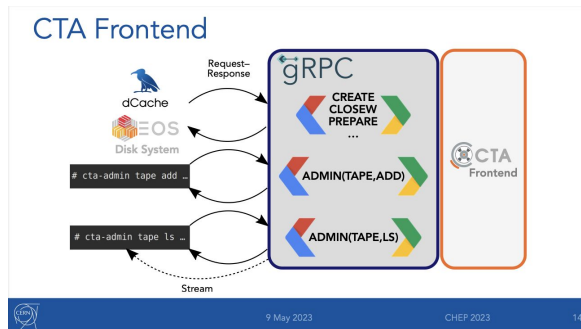
Databases & Metadata

- **Evaluating Rucio as a metadata service**
  - Rucio able to handle Belle II metadata requirements

- **Conditions DB**
  - ATLAS: Major shift coming for Run4
    - From COOL to CREST
  - ALICE: Interval of Validity queries in three flavors
    - Local / In-Memory / Grid SE-backed
  - HSF Conditions Database Reference Implementation
    - Implementations for sPHENIX and DUNE

Metadata tests

- A stress test was conducted using a snapshot of Belle II production instance (~100M files imported) and deployed on a test instance :
  - Similar DB backend as the production one, but only one Apache front-end
  - Test is querying a list of files in Rucio and set a few metadata for each file
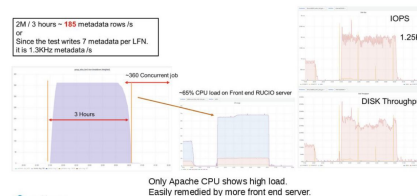 - Multiple tests are run in parallel using a batch system to increase the load on Rucio (up to 360 jobs)
- **No bottleneck observed on the DB side.**
- **Limitation comes from the single front-end used for the test, but can be scaled horizontally**

2M / 3 hours ~ **185** metadata rows /s
or
Since the test writes 7 metadata per LFN,
it is 1.3KHz metadata /s

~360 Concurrent job

IOPS
1.25K

~65% CPU load on Front end RUCIO server

3 Hours

DISK Throughput

Only Apache CPU shows high load.
Easily remedied by more front end server.

7

IoV queries, snapshots

Most recent object that covers t_o

| IoV start | Better calibration, based on the original data only | IoV end |
| IoV start | Async created object (manual calibration / CPass …) | IoV end |

"Time machine" / *If-Not-After* HTTP header

| IoV start | Sync created object | IoV end |

…

Object creation time

| IoV start | Sync created object | IoV end |

| IoV start | **Sync created object** | IoV end |

| IoV start | Sync created object | IoV end |

t_o

Time reference from data

ALICE Calibration and Conditions Database, CHEP 2023, May 8-12

6

Data Management

- **Nordic Data Lake**
  - Incorporation of geographically "distant" resources done successfully and performant
- **Data Management and Services for HEPS**
  - Full stack service with useful web interface
- **Image processing of LSST**
  - Using common HEP tools: Rucio, FTS
  - Expansion of raw data (50 PB/10 years) to much larger derived datasets (500 PB) is rather unique
- **FTS**
  - Evolution to real micro services is exciting
  - FENIX: Bridging different AAI systems (ESCAPE → FENIX)

Analytics & Benchmarks

- **Similar themes as in the networks session emerged**
  - Monitoring & prediction key to optimisation
- **Exploring Future Storage at the BNL/SDCC facility**
  - XrootD Lustre vs. dCache Lustre vs. dCache ZFS
- **Data popularity for Cache Eviction**
  - Surpassing LRU is challenging: Prediction & Integration
- **400Gbps benchmark of XRootD HTTP-TPC**
  - Number streams / Redirector overhead? / Transfer tool?
- **Scale Tests of the New DUNE Data Pipeline**
  - 3.6GB/s CERN → FNAL!

Comparing ML-based models with LRU

- When the high watermark is hit, prediction is run on all the files present in the cache. We then sort them based on the predicted reuse distance (or probability to be reused) in order to be evicted from the cache.
- Model with classification permofrms better, but still not able to beat LRU.
- When the cache size is 2-5 %, the difference between 1-2 days of reuse and 1-2 weeks becomes more crucial.

Data popularity for the Cache Eviction Algorithms using Random Forests

OSG

WLCG

CHEP2023 - New XRootD Monitoring implementation - Borja Garrido

Summary of summary (TL;DR)

- Clouds & object stores on the rise
- Better monitoring and prediction of our systems (networks, caches) is crucial to optimise them
- End of SRM is near!
- OIDC/oAuth2 tokens touch lots of systems
- Further push to common solutions

Thank you!