Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

CHEP 2024 Highlights NPSS/SDCC meeting

Carlos Fernando Gamboa

11/13/2024

@BrookhavenLab

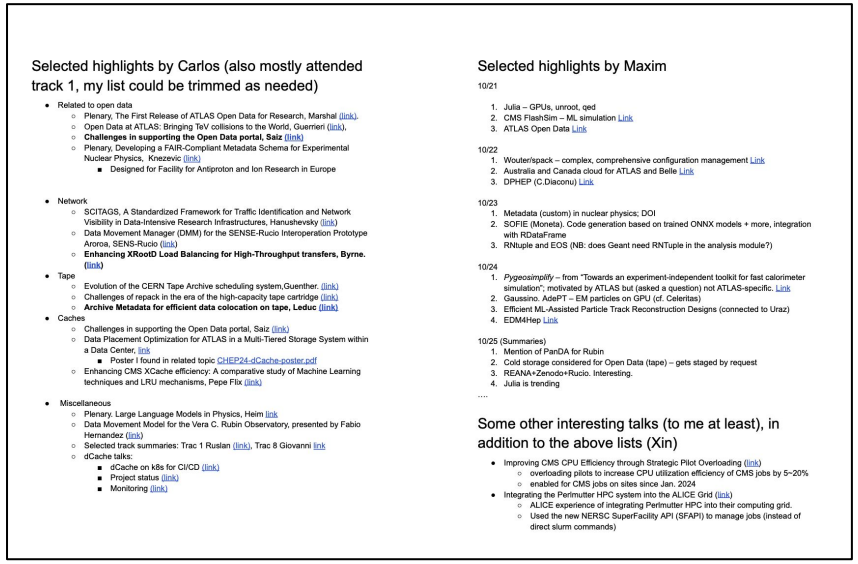
CHEP 2024 Highlights 2/2

The first series of highlighted talks covered last week in the NPSS meeting included a selection from BNL colleagues attending CHEP2024.

CHEP24 - selected highlights by JE:

- Overall timetable: https://indico.cern.ch/event/1338689/timetable/?view=standard
- Good conference, good logistics, good coffee chats, great city, easy travel from CERN - definitely worth attending !
- Plenary talks Mon-Fri in the morning, 9 parallel tracks Mon-Thu afternoons
- 470 participants, 351 talks & 143 posters
- Plenaries (selected highlights, personal taste):
  - Julia in HEP, Graeme Stewart (CERN), link
  - Large Language Models in Physics, Sarah Heim (DESY), link
  - Panel Discussion: Quantum Computing and LLMs. Will they bring disruptive changes to HEP computing? - good discussion - no real disruption expected by panel experts since AI/ML is used in HEP since many years, quantum was hardly discussed
  - Total Cost of Ownership and Evaluation of Google Cloud Resources for the ATLAS Experiment at the LHC, David South (DESY), link
  - ROOT RNTuple and EOS: The Next Generation of Event Data I/O, Andreas Joachim Peters (CERN), Jakob Blomer (CERN), link
  - The WLCG Data Challenge, Katy Ellis (RAL, STFC), link
  - Hardware technology trends in HEP computing, Andrea Sciabà (CERN), link
    - Bernd Panzer who did these kind of talks for years is about to retire
- Disappointment of the week: 3 Quantum Computing plenary talks that were not particularly good
- Mainly went to parallel Track 1: Data and Metadata Organization, Management and Access
  - Sessions on data challenge topics (link), tape usage (link), dCache, Storm, FTS (link), data management in various experiments (link), RNTuple and other file storage topics (link), data caching (link), databases and filesystems (link)
  - Scitags: A Standardized Framework for Traffic Identification and Network Visibility in Data-Intensive Research Infrastructures, presented by Andy Hanushevsky (link)
  - Data Movement Model for the Vera C. Rubin Observatory, presented by Fabio Hernandez (link)
- Parallel talks, Posters or session summaries by BNL speakers that I found in the agenda (did not attend them all): Scott (link), Kirby (link), Fang-Yin (link), Xin (link), Sakib (link), Johannes (link), Kirby (link), Marcin (link), Carlos (link), Haider (link), Jacob (link), David Park (link), Maxim (link), Ruslan (link) - sorry if I overlooked someone
- Reminder: ATLAS internal deadline for proceedings is 13 Dec
- Next CHEP will be 25-29 May 2026 in Bangkok, Thailand

[Johannes E. document on CHEP 2024 for NPSS meeting 110624]



Selected highlights by Carlos (also mostly attended track 1, my list could be trimmed as needed)

- Related to open data
  - Plenary, The First Release of ATLAS Open Data for Research, Marshal (link).
  - Open Data at ATLAS: Bringing TeV collisions to the World, Guerrieri (link),
  - **Challenges in supporting the Open Data portal, Saiz** (link)
  - Plenary, Developing a FAIR-Compliant Metadata Schema for Experimental Nuclear Physics, Knezevic (link)
    - Designed for Facility for Antiproton and Ion Research in Europe
- Network
  - SCITAGS, A Standardized Framework for Traffic Identification and Network Visibility in Data-Intensive Research Infrastructures, Hanushevsky (link)
  - Data Movement Manager (DMM) for the SENSE-Rucio Interoperation Prototype Aroroa, SENS-Rucio (link)
  - **Enhancing XRootD Load Balancing for High-Throughput transfers, Byrne.** (link)
- Tape
  - Evolution of the CERN Tape Archive scheduling system,Guenther. (link)
  - Challenges of repack in the era of the high-capacity tape cartridge (link)
  - **Archive Metadata for efficient data colocation on tape, Leduc** (link)
- Caches
  - Challenges in supporting the Open Data portal, Saiz (link)
  - Data Placement Optimization for ATLAS in a Multi-Tiered Storage System within a Data Center, link
    - Poster I found in related topic CHEP24-dCache-poster.pdf
  - Enhancing CMS XCache efficiency: A comparative study of Machine Learning techniques and LRU mechanisms, Pepe Flix (link)
- Miscellaneous
  - Plenary. Large Language Models in Physics, Heim link
  - Data Movement Model for the Vera C. Rubin Observatory, presented by Fabio Hernandez (link)
  - Selected track summaries: Trac 1 Ruslan (link), Trac 8 Giovanni link
  - dCache talks:
    - dCache on k8s for CI/CD (link)
    - Project status (link)
    - Monitoring (link)

Selected highlights by Maxim

10/21

1. Julia – GPUs, unroot, qed
2. CMS FlashSim – ML simulation Link
3. ATLAS Open Data Link

10/22

1. Wouter/spack – complex, comprehensive configuration management Link
2. Australia and Canada cloud for ATLAS and Belle Link
3. DPHEP (C.Diaconu) Link

10/23

1. Metadata (custom) in nuclear physics; DOI
2. SOFIE (Moneta). Code generation based on trained ONNX models + more, integration with RDataFrame
3. RNtuple and EOS (NB: does Geant need RNTuple in the analysis module?)

10/24

1. *Pygeosimplify* – from "Towards an experiment-independent toolkit for fast calorimeter simulation"; motivated by ATLAS but (asked a question) not ATLAS-specific. Link
2. Gaussino. AdePT – EM particles on GPU (cf. Celeritas)
3. Efficient ML-Assisted Particle Track Reconstruction Designs (connected to Uraz)
4. EDM4Hep Link

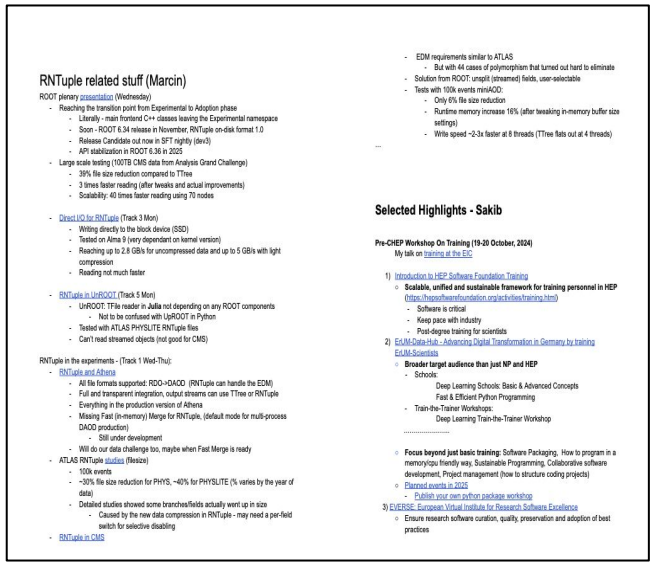
10/25 (Summaries)

1. Mention of PanDA for Rubin
2. Cold storage considered for Open Data (tape) – gets staged by request
3. REANA+Zenodo+Rucio. Interesting.
4. Julia is trending

....

Some other interesting talks (to me at least), in addition to the above lists (Xin)

- Improving CMS CPU Efficiency through Strategic Pilot Overloading (link)
  - overloading pilots to increase CPU utilization efficiency of CMS jobs by 5~20%
  - enabled for CMS jobs on sites since Jan, 2024
- Integrating the Perlmutter HPC system into the ALICE Grid (link)
  - ALICE experience of integrating Perlmutter HPC into their computing grid.
  - Used the new NERSC SuperFacility API (SFAPI) to manage jobs (instead of direct slurm commands)



RNTuple related stuff (Marcin)

ROOT plenary presentation (Wednesday)
- Reaching the transition point from Experimental to Adoption phase
  - Literally - main frontend C++ classes leaving the Experimental namespace
  - Soon - ROOT 6.34 release in November, RNTuple on-disk format 1.0
  - Release Candidate out now in SFT nightly (dev3)
  - API stabilization in ROOT 6.36 in 2025
- Large scale testing (100TB CMS data from Analysis Grand Challenge)
  - 39% file size reduction compared to TTree
  - 3 times faster reading (after tweaks and actual improvements)
  - Scalability: 40 times faster reading using 70 nodes
- Direct I/O for RNTuple (Track 3 Mon)
  - Writing directly to the block device (SSD)
  - Tested on Alma 9 (very dependant on kernel version)
  - Reaching up to 2.8 GB/s for uncompressed data and up to 5 GB/s with light compression
  - Reading not much faster
- RNTuple in UnROOT (Track 5 Mon)
  - UnROOT: TFile reader in **Julia** not depending on any ROOT components
    - Not to be confused with UpROOT in Python
  - Tested with ATLAS PHYSLITE RNTuple files
  - Can't read streamed objects (not good for CMS)

RNTuple in the experiments - (Track 1 Wed-Thu):
- RNTuple and Athena
  - All file formats supported: RDO->DAOD (RNTuple can handle the EDM)
  - Full and transparent integration, output streams can use TTree or RNTuple
  - Everything in the production version of Athena
  - Missing Fast (in-memory) Merge for RNTuple, (default mode for multi-process DAOD production)
    - Still under development
  - Will do our data challenge too, maybe when Fast Merge is ready
- ATLAS RNTuple studies (filesize)
  - 100k events
  - ~30% file size reduction for PHYS, ~40% for PHYSLITE (% varies by the year of data)
  - Detailed studies showed some branches/fields actually went up in size
    - Caused by the new data compression in RNTuple - may need a per-field switch for selective disabling
- RNTuple in CMS
  - EDM requirements similar to ATLAS
    - But with 44 cases of polymorphism that turned out hard to eliminate
  - Solution from ROOT: unsplit (streamed) fields, user-selectable
  - Tests with 100k events miniAOD:
    - Only 6% file size reduction
    - Runtime memory increase 16% (after tweaking in-memory buffer size settings)
    - Write speed ~2-3x faster at 8 threads (TTree flats out at 4 threads)

...

Selected Highlights - Sakib

Pre-CHEP Workshop On Training (19-20 October, 2024)
My talk on training at the EIC

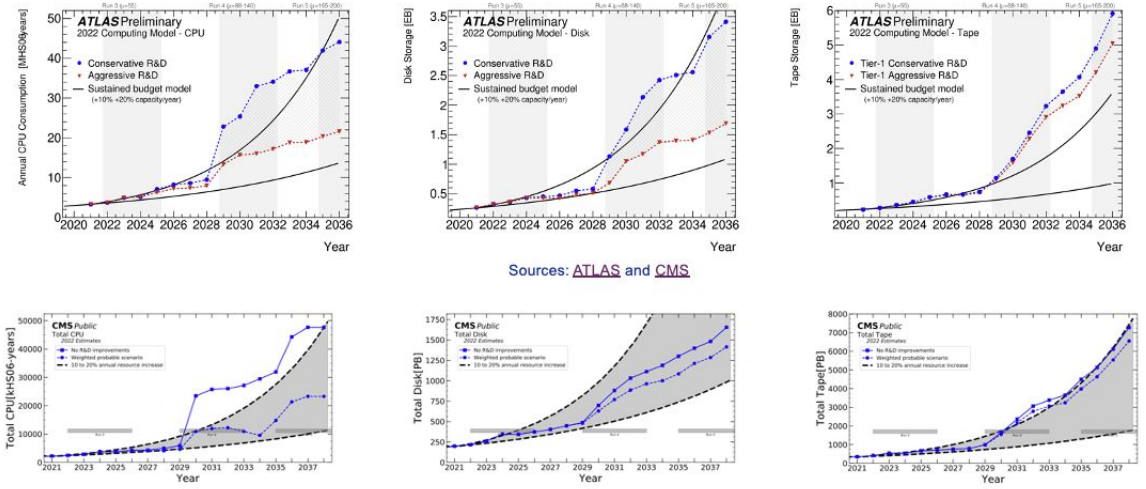
1) Introduction to HEP Software Foundation Training
   - **Scalable, unified and sustainable framework for training personnel in HEP** (https://hepsoftwarefoundation.org/activities/training.html)
     - Software is critical
     - Keep pace with industry
     - Post-degree training for scientists
2) EVUM-Data-Hub - Advancing Digital Transformation in Germany by training EvUM-Scientists
   - **Broader target audience than just NP and HEP**
     - Schools:
       Deep Learning Schools: Basic & Advanced Concepts
       Fast & Efficient Python Programming
     - Train-the-Trainer Workshops:
       Deep Learning Train-the-Trainer Workshop
   - **Focus beyond just basic training:** Software Packaging, How to program in a memory/cpu friendly way, Sustainable Programming, Collaborative software development, Project management (how to structure coding projects)
   - Planned events in 2025
     - Publish your own python package workshop
3) EVERSE: European Virtual Institute for Research Software Excellence
   - Ensure research software curation, quality, preservation and adoption of best practices

A selection of highlights will be covered in this short talk

Hardware trends: disk, network, compute

1. Hardware technology trends in HEP computing, Andrea Sciabà (CERN)

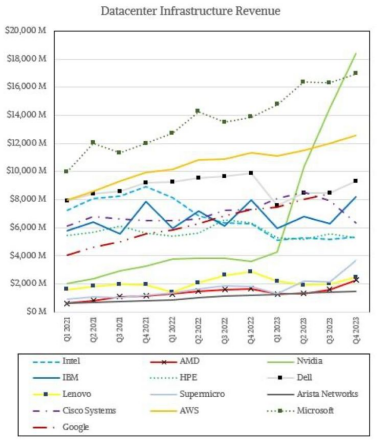
ATLAS and CMS resource needs up to HL-LHC



Sources: ATLAS and CMS

Comparing major players at scale

- **Revenues steadily increasing in the last few years, with a few exceptions**
  - Intel dropped quite a bit and lost most of its value, future uncertain
  - Nvidia skyrocketed in 2023
- **AI is the main driver and Nvidia has a practical monopoly**
  - AMD might increase their share, as demand is very high and have competitive products



Source: The Next Platform

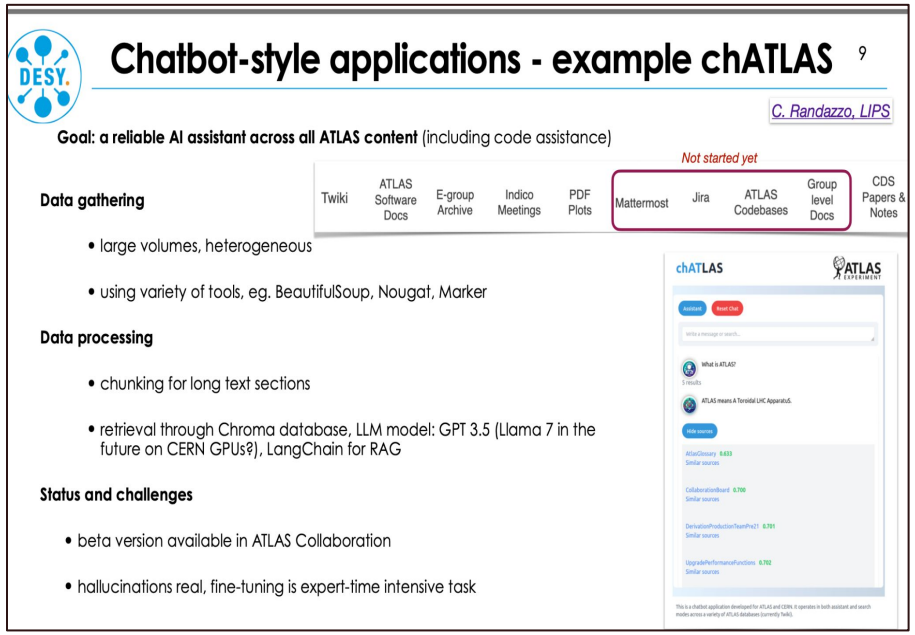
Conclusions

- **Technology tracking essential to make cost-efficient choices for HEP computing**
  - Done in different contexts in our community
- **Many server hardware components are rising in price due to the AI boom**
  - Memory, GPUs, flash, HDD are all affected
- **AMD, Arm and Intel show healthy competition**
  - A lot of attention to performance/Watt for many reasons
- **Evolution of GPUs is not going in a direction very useful for us**
  - FP32/64 performance not increasing in the short/medium term, to maximize AI performance
- **Shipped storage capacity increasingly driven by the global trend**
  - SSDs, HDDs and tape all still relevant and making technological progress
- **Network bandwidth correspondingly increasing on LAN and WAN**
  - To cope with increase in cores and storage/server
  - For LHC, driven by HL-LHC data rates
- **Sustainability is more important than ever**
  - CO2 emissions, liquid cooling, electricity costs and distribution

AI driving market and technological trends towards AI like workloads

What about AI/ML

Large Language Models in Physics by Sara Heim

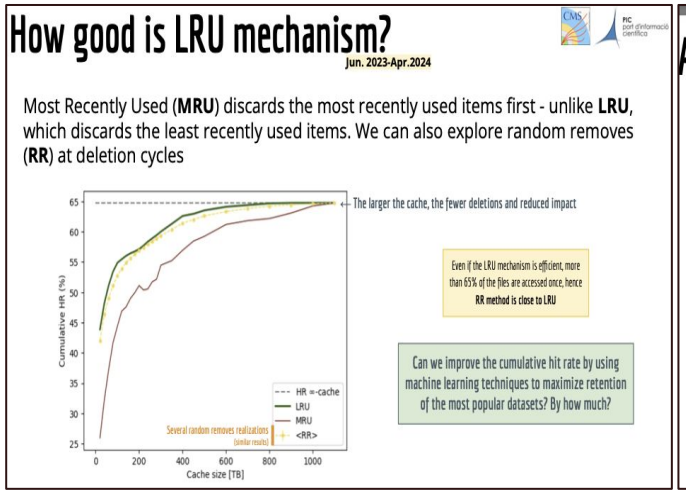


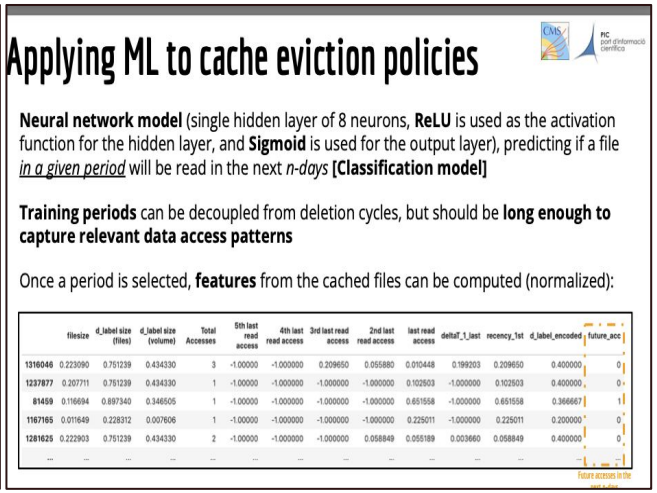
Usage within facilities seems to be in an R&D stage, despite its application in high-energy physics (HEP).

AI + CACHE studies

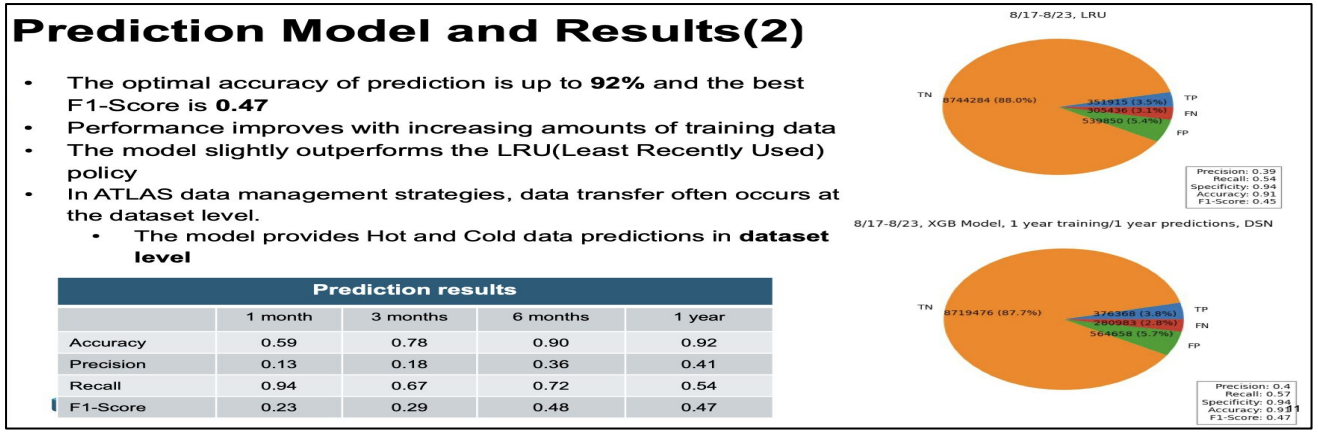
Enhancing CMS XCache efficiency: A comparative study of Machine Learning techniques and LRU mechanisms, Pepe Flix SPAIN, CMS

Leverage the use of data caches as part Content Delivery Network





Data Placement Optimization for ATLAS in a Multi-Tiered Storage System within a Data Center, Q. Huang BNL , US, ATLAS



TAPE

CERN reported on improvements to their TAPE backed

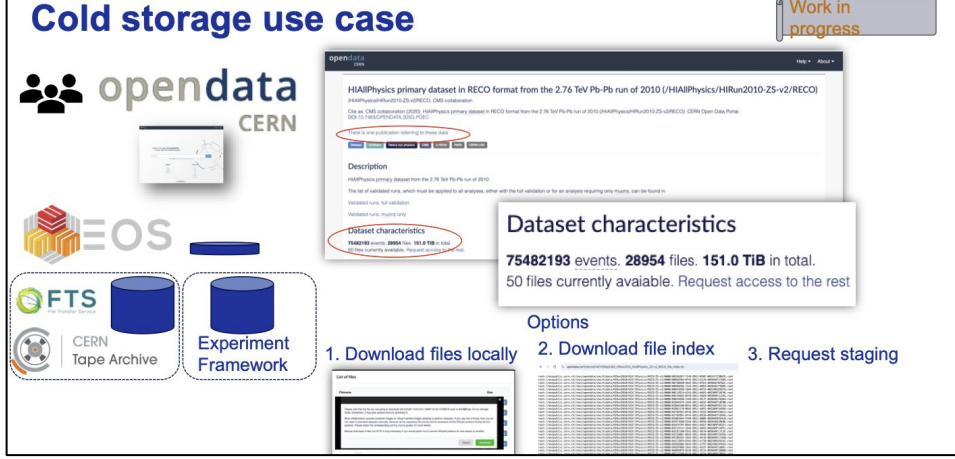
- CTA Scheduler DB implementation, based on relational database technology.Guenther
- CTA Challenges of tape repack Afonso

Tape: Metadata

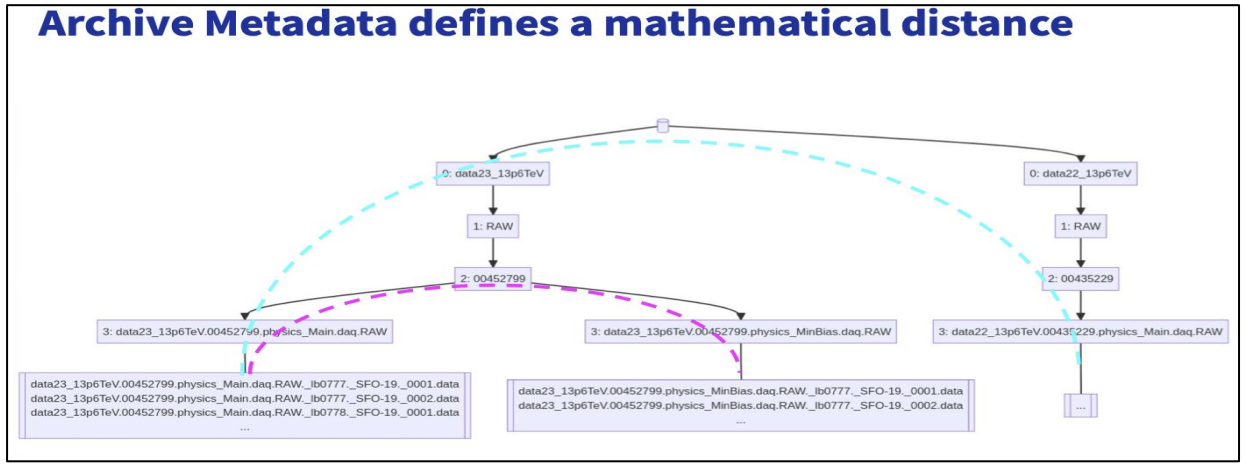
Leverage the use of metadata to optimize tape access, reads and writes

- Optimize writes to tape by colocating files using dCache extended atributes DORIN, dCache+HPSS

- Archive Metadata for efficient data collocation on tape,Leduc
  - EOS+CTA



Challenges in supporting the Open Data portal, Saiz (link)



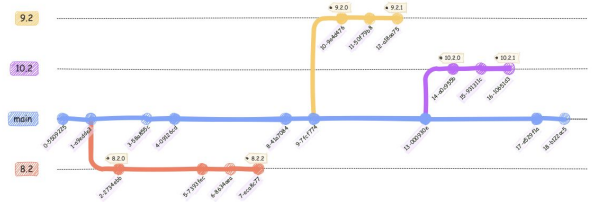
Brookhaven National Laboratory

Storage

We are here

dCache versions

- 8.2 out of support
  - Security fixes only.
- 9.2 Current
  - Support till fall 2025.
- 10.2 the next **Golden Release** (LTS)
  - Release last week.
  - Support till fall 2026.



2024-10-22 dCache project update 19/21

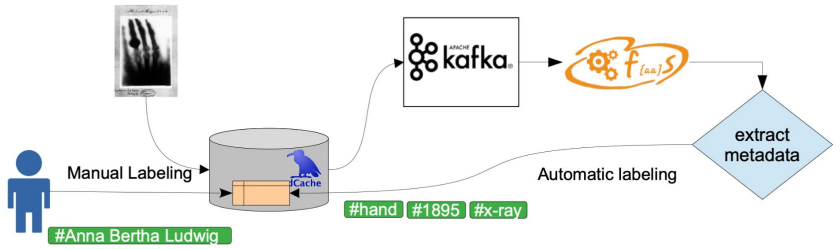
Labeling

- Add custom labels to existing file
- Labels dynamically create file collections
- Collections exposed as virtual-directories
- Files can be accesses by regular path
- Files can be accessed by collections

#Anna Bertha Ludwig #hand #1895 #x-ray

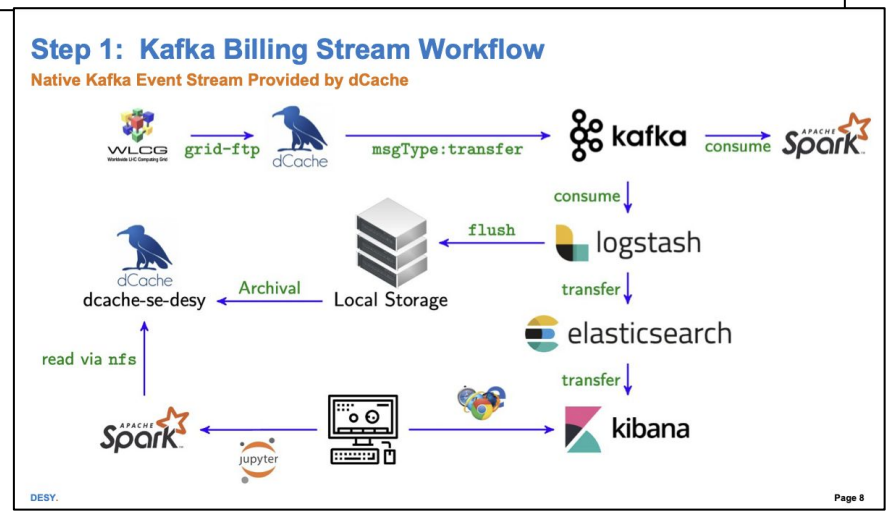
2024-10-22 dCache project update 9/21

Metadata Population



For details, visit our Poster: label-based virtual directories in dCache, *Oct 24*

2024-10-22 dCache project update 10/21



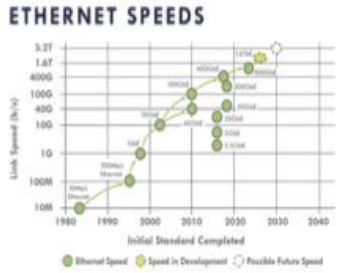
dCache talks:

- dCache on k8s for CI/CD (link)
- Project status (link)
- Monitoring (link)

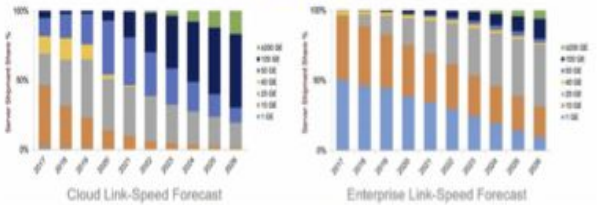
DATA+Network

LAN and interconnect technologies

- **InfiniBand provides high throughput/low latency networking**
  - Useful for AI and HPC simulations
  - Can provide Remote Direct Memory Access (RDMA)
  - Now controlled by Nvidia, cost may amount to up to 20% of an HPC cluster
  - RoCE (RDMA over Converged Ethernet) is a much cheaper alternative that works over Ethernet
- **Ultra Ethernet consortium aims at producing an alternative to InfiniBand**
  - Open standard supported by AMD, Broadcom, Cisco, HPE, Meta, Microsoft, Oracle, Linux Foundation, and many others (> 60 companies so far, even Nvidia!)
  - Improves the Ethernet protocol to allow for high bandwidth/low latency, would replace RoCE
- **Omni-Path is a competing standard originally from Intel**
  - It will be made compatible with Ultra Ethernet to stay relevant



Network interface port speeds



Cloud Link-Speed Forecast

Enterprise Link-Speed Forecast

Cloud: All about 100G+

Enterprise: Mix of 10G, 25G, 100G

Paving The Way For 800 Gb/sec Ethernet In The Enterprise (nextplatform.com)

- The 200 Gpbs SerDes allows to send 200 Gbps on a single wavelength, reducing power and cost
- Co-packaged optics embed the lasers in the motherboard, potentially reducing costs

Trends on WAN connectivity

- **LHC network traffic exponentially increasing, will need Tb/s links on major routes by 2029**
  - Aggregate network traffic from ATLAS + CMS will be O(10 Tb/s)
- **R&D effort focusing on**
  - Better estimates of the required scale
  - Better models and well-defined metrics for success
  - ML for system optimization
  - Better automation (monitoring, intelligence, network OSes and tools, controllability)

The WLCG Data Challenge, Katy Ellis (RAL, STFC), link

SCITAGS, A Standardized Framework for Traffic Identification and Network Visibility in Data-Intensive Research Infrastructures, Hanushevsky (link)

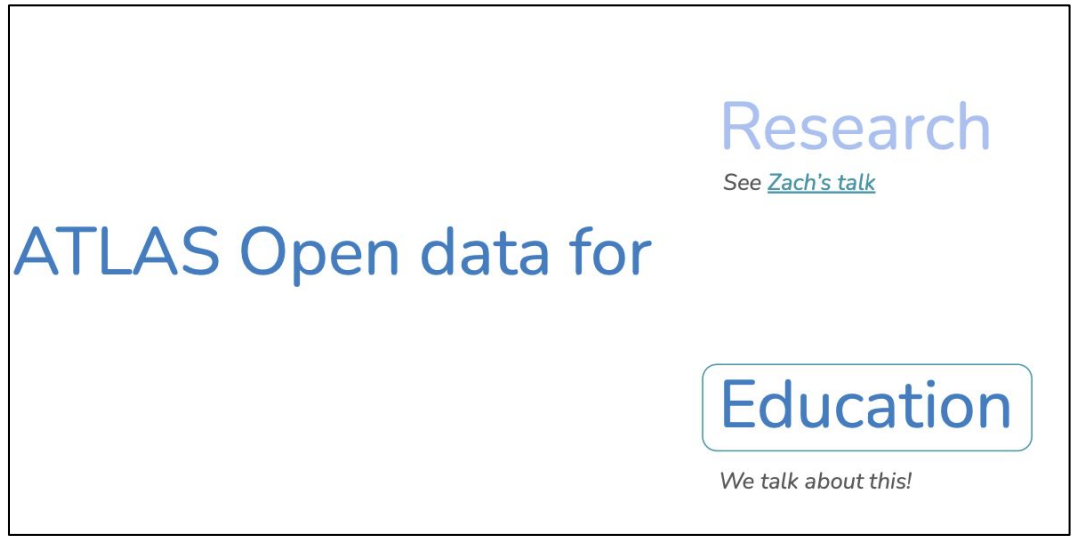
Data Movement Manager (DMM) for the SENSE-Rucio Interoperation Prototype Aroroa, SENS-Rucio (link)

Enhancing XRootD Load Balancing for High-Throughput transfers, Byrne. (link)

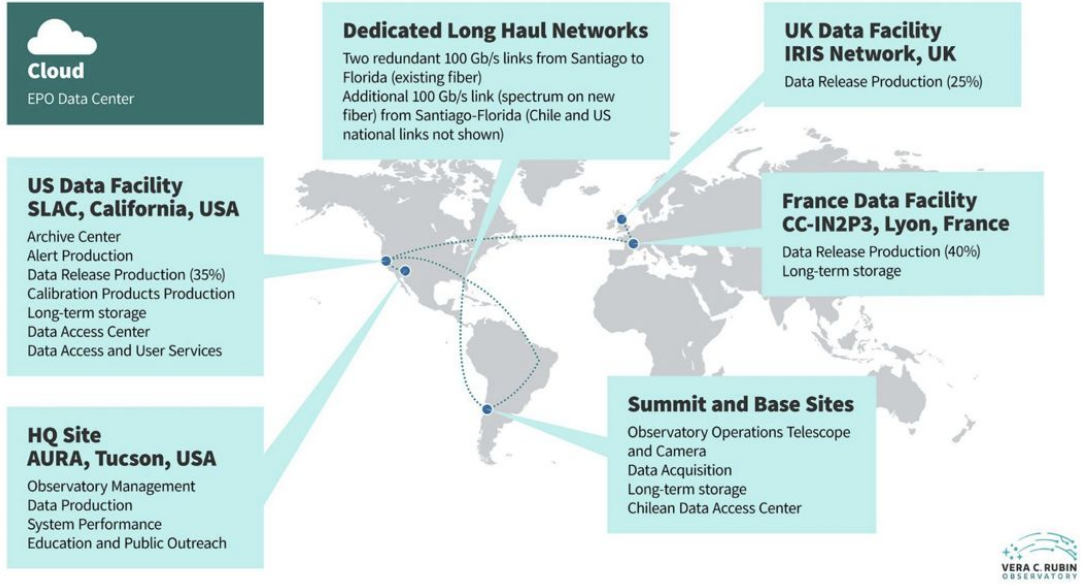
Other interesting talks

Selected track summaries: Trac 1 Ruslan (link), Trac 8 Giovanni link

ROOT RNTuple and EOS: The Next Generation of Event Data I/O, Andreas Joachim Peters (CERN), Jakob Blomer (CERN), link



Plenary, The First Release of ATLAS Open Data for Research, Marshal (link).
Open Data at ATLAS: Bringing TeV collisions to the World, Guerrieri (link),



Data Movement Model for the Vera C. Rubin Observatory, presented by Fabio Hernandez (link)

Plenary, Developing a FAIR-Compliant Metadata Schema for Experimental Nuclear Physics, Knezevic (link)

- Designed for Facility for Antiproton and Ion Research in Europe

Brookhaven National Laboratory

Summary of CHEP 2024 Highlights:

- Key insights from CHEP 2024 were shared across multiple formats and meetings.
- BNL colleagues actively participated.
- An ATLAS-wide overview was presented, with special thanks to Tatiana.
- A curated selection of talks was covered, encouraging further discussion.