



Norfolk, Virginia, USA • May 8-12, 2023

CHEP 2023

Computing in High Energy & Nuclear Physics

A summary of CHEP 2023

Caterina Marcon (INFN Milano) & Edward Karavakis (Brookhaven National Laboratory)

ATLAS Weekly
23 May 2023



A bit of context

- CHEP conference series addresses the computing, networking and software issues for world's leading data-intensive science experiments that currently (or will) analyze hundreds of PBs using worldwide computing resources
- Rotates between the Americas, Asia and Europe, every 18 months
- Previous physical CHEP conference (24th) was held in Adelaide, Australia in November 2019
- Next one was supposed to be at Jefferson Lab in May 2021 and then the pandemic came... converted into a purely virtual conference: vCHEP 2021 (25th)
- CHEP 2023 (26th) was hosted by Jefferson Lab in Norfolk, Virginia
- An optional pre-conference WLCG & HSF workshop held the weekend prior the conference with 154 participants, while not covered in this presentation, you can find a link to the agenda and to many interesting talks [here](#)
 - First day focused on [Analysis Facilities](#)
 - Second day focused on [Non-x86 and Heterogeneous Computing](#)

12 Parallel Tracks!

Track 1 - Data and Metadata Organization, Management and Access

Storage management frameworks; data access protocols; object, metadata and event store systems; content delivery and caching; data analytics; FAIR data principles; non-event data; data classification; online and offline databases.

Track 2 - Online Computing

Data acquisition; high-level triggers; streaming and trigger-less data acquisition; online data calibration; online reconstruction; real-time analysis; event building; configuration and access controls; detector control systems; real-time analytics and monitoring; trigger techniques and algorithms; hardware trigger algorithms.

Track 3 - Offline Computing

MC event generation; detector simulation; fast simulation; offline reconstruction; detector calibration; detector geometries; data quality systems; data preparation; physics performance.

Track 4 - Distributed Computing

Grid middleware; monitoring and accounting frameworks; security models and tools; distributed workload management; federated authentication and authorisation infrastructures; middleware databases; software distribution and containers; heterogeneous resource brokerage.

Track 5 - Sustainable and Collaborative Software Engineering

Software frameworks; collaborative software; sustainable software; software management, continuous integration; software building; testing and quality assurance; software distribution; programming techniques and tools; integration of external toolkits.

Track 6 - Physics Analysis Tools

Analysis algorithms; object identification; object calibration; analysis workflows; lattice QCD; theory calculations; high performance analysis frameworks.

Track 7 - Facilities and Virtualization

Cloud resources; HPC and supercomputers; deployment of virtual machines and container technologies; anything-as-a-service; private and commercial clouds; dynamic provisioning; networking; computing centre infrastructure; management and monitoring; analysis facilities.

Track 8 - Collaboration, Reinterpretation, Outreach and Education

Collaborative tools; reinterpretation tools; analysis preservation and reuse; data preservation for collaboration; outreach activities; open data for outreach; training initiatives; event displays; open science cloud initiatives.

Track 9 - Artificial Intelligence and Machine Learning

Machine learning algorithms; machine learning for online; machine learning for simulation and reconstruction; machine learning tools and techniques for analysis; machine learning for reinterpretation; massive scale machine learning; hyperparameter optimization.

Track 10 - Exascale Science

Algorithm scaling; exascale computing models; exabyte-scale datasets; exaflop computing power; generic algorithms; weak scaling.

Track 11 - Heterogeneous Computing and Accelerators

Compute accelerators; concurrency in software frameworks; accelerator-as-a-service; FPGA programming; software design and implementation for heterogeneous architectures; heterogeneous resource usage for online and offline.

Track 12 - Quantum Computing

Quantum computing for theory calculations; quantum computing for event generation, simulation and reconstruction; quantum computing for analysis; quantum computing applications.

Merged into **Track X - Exascale Science**

582 participants - 582 contributions!

- 10 parallel tracks
- 138 posters
- 33 plenary talks (out of them 10 plenary summaries & 1 wrap-up)
- 411 parallel talks



Disclaimer

- This **is** quite a **biased** overview of CHEP
- We could only attend just **one** talk at a time while there were **ten!**
 - Luckily, there was no big overlap in sessions between the two of us
- Not enough time to cover everything in just 15 minutes..
- Check the [agenda](#) yourself for more!

Networking: Software Defined Networks

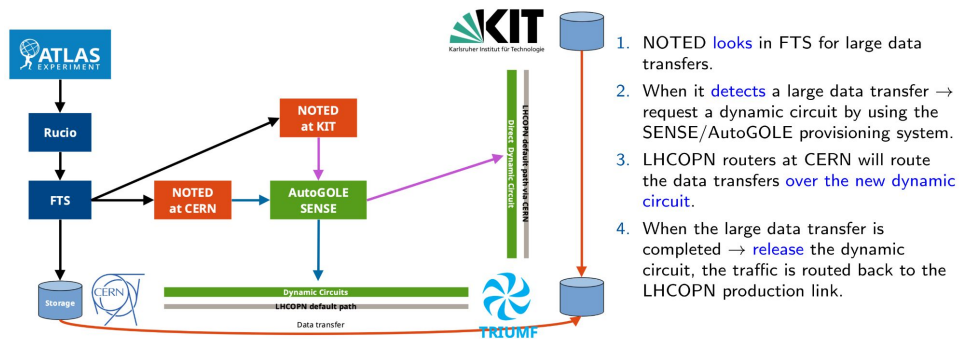
Becoming more obvious that network is also a scarce resource - impact to compute models is real

NOTED: An intelligent network controller to improve the throughput of large data transfers in File Transfer Services by handling dynamic circuits - C. Misa Moreira

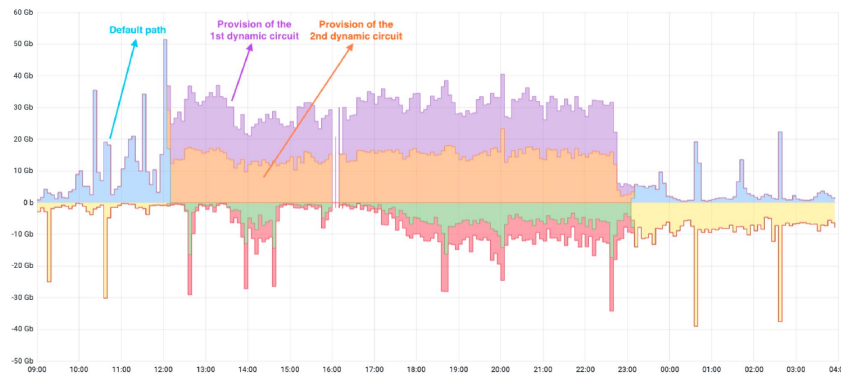
Large data transfers can **saturate** network links while alternative paths may be left **idle**

Goal is to **reduce** duration of large data transfers and improve the **efficient** use of network resources by monitoring FTS production transfers

NOTED demo for SC22



NOTED demo for SC22



Super Computing 22 Conference Demo

Identifying and Understanding Scientific Network Flows - S. McKee

Identifying network flows through tagging with Scientific Network Tags (SciTags) - <https://scitags.org>

scitags is an initiative promoting identification of science domains and their high-level activities at the network level

You can find HEP (tools) even in Astrophysics!

[Global Data Management in Astronomy and the link to HEP: are we collaborators, consumers or competitors? - R. Bolton \(SKA\)](#)

The Data Management legacy of the ESCAPE project

Escape helped extend knowledge of **communities** with shared concerns

It was a springboard for us in SKA, CTAO and LSST Rubin to better understand HEP solutions

Enabled community-led data replication and data access challenges

Led to self-managed Rucio instances at CTAO, Rubin and SKA



See: <https://projectescape.eu/sites/default/files/ESCAPE-D2.2-v1.0.pdf>

Intended take-home messages

Global Data Management is a shared concern – CERN / WLCG have paved the way but several upcoming astronomy observatories will face the challenge alongside HL-LHC

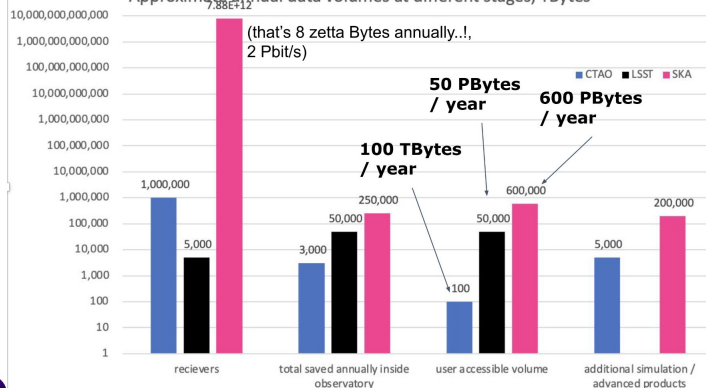
Collaborators? Yes indeed! The ESCAPE project was a brilliant nursery slope for me at SKA. In spite of lockdowns we built a genuine collaboration – personal connections made will ensure cooperation in future developments

Consumers? Yes, but mindfully so – we need to recognise our differences and be prepared to adapt tools to fit our needs, not wait for Astro-needs to automatically emerge from HEP-focused tooling. **It is not a "one size fits all" situation:** differences in dataset size, file size, governance & control, user access pattern, protection, data lifecycle

Competitors? I hope not! We will not be competing for storage (separate pledges / facilities) or network, but do need to be mindful of **pressure on shared sites** and on the **development of tools** we are exploring. Continued ESCAPE collaboration will help ensure we have forum to discuss how to solve technical challenges.

Annual Data volumes through the systems

Approximate annual data volumes at different stages, TBytes



Astronomy world moving towards **exascale** observatories and experiments, with data distribution and access challenges comparable with - and on similar timescales to - those of **HL-LHC**

More Astro stuff

Overview of the distributed image processing infrastructure to produce the Legacy Survey of Space and Time (LSST) - F. Hernandez



Legacy Survey of Space and Time (cont.)

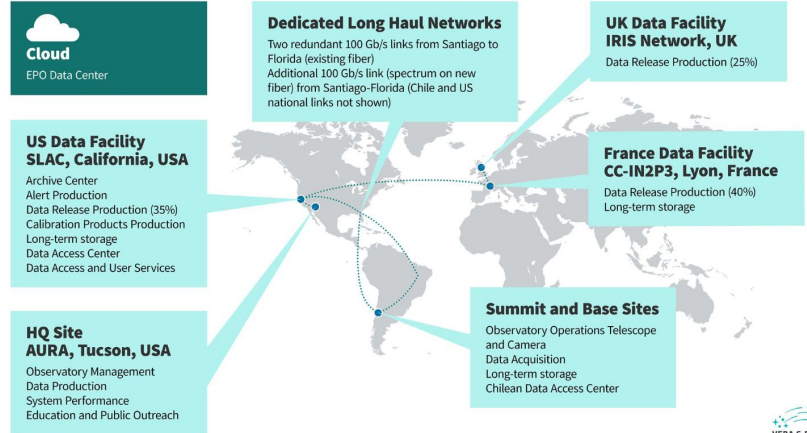
Raw data
6.4 GB per exposure (compressed)
2000 science + 500 calibration images per night
20 TB per night, ~5 PB per year

Aggregated data over 10 years of operations
image collection: ~6 million exposures
derived data set: ~0.5 EB
final astronomical catalog database: 15 PB

Operations to start early 2025

Source: Rubin Observatory System & LSST Survey Key Numbers

Vera C. Rubin Observatory | CHEP 2023
Acronyms & Glossary
7

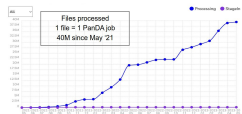


...and more HEP stuff

Distributed image processing

- Batch Production Service (BPS)
 - Generates the workflow to be executed at each facility: a directed acyclic graph of independent units of work
 - Takes into account data dependencies and data location
- PanDA
 - Creates pilot jobs and coordinates the execution of the workflow
 - Each job executes one or several science algorithms over a set of input data, stores output data in the butler repository local to the facility

Vera C. Rubin Observatory | CHEP 2023
Acronyms & Glossary
15



Inter-site data replication

- Data replication will be achieved with open-source software:
 - Rucio and FTS3
 - Proven to work at scale by the ATLAS and CMS collaborations, among others
 - Rucio
 - Replica catalog: Where does my data live?
 - Data policy enforcement: How many copies of the data, and where?
 - Transfer scheduling: Arranges to satisfy your policies with external services!
 - FTS3
 - Executes transfers scheduled externally on behalf of Rucio
 - Highly configurable for tuning handling of many transfers to many sites
 - Rubin-specific tools
 - To identify data which needs replication among the facilities (e.g. exclude intermediates) and ask Rucio to replicate it
 - To trigger actions at each facility to timely ingest replicated data into the local data butler repository

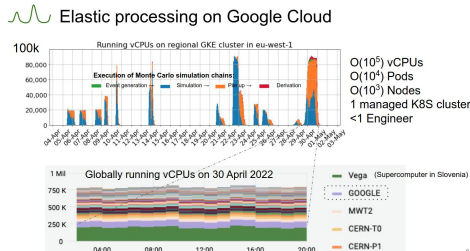
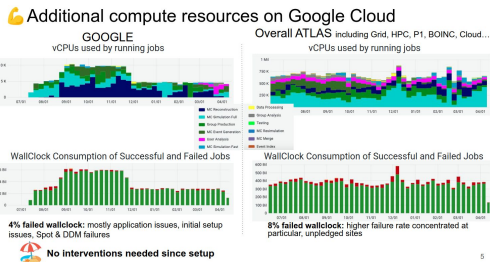
Data replication over high-latency network links

Vera C. Rubin Observatory | CHEP 2023
Acronyms & Glossary
16

Clouds

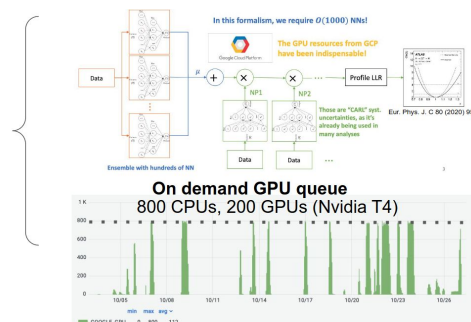
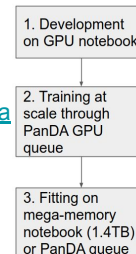
Lots of R&D projects on clouds: Access to resources not available on prem!

- [Extending Rucio with modern cloud storage support: Experiences from ATLAS, SKA and ESCAPE - M. Lassnig](#) - Significant effort to support Commercial Clouds (Google, AWS and SEAL)
- [Accelerating science: the usage of commercial clouds in ATLAS distributed computing - F. Barreiro](#)



- Low maintenance cost
- Additional compute resources
- Elastic processing
- Running “à la Grid” technically possible but not desirable due to **very high ingress costs**

- [ATLAS data analysis using a parallelized workflow on distributed cloud-based services with GPUs - J. Sandesara](#)
- [Financial Case Study on the Use of Cloud Resources in HEP Computing - S. Misawa](#)
 - Multiple factors complicate calculation of costs in the cloud: overhead, electricity, and cooling
 - Showing up to 6x advantage of using resources on premises rather on the cloud



Summary

Cloud hosting is significantly more expensive than hosting on premises for just the major resources required by NP/HEP

Cloud based resources will incur additional expenses, beyond just compute and storage, most notably network egress fees if all jobs run on premises are moved to the cloud

Calculation of costs in the cloud is complicated and are highly dependent on the services being moved, how the services are moved, and the data center that is making the move.

Mistakes in the enumeration of requirements can result in significant increases in costs in the cloud

- Inadvertent use of metered services
- Underestimation of usage of metered services

Power Consumption

[A holistic study of the WLCG energy needs for the LHC scientific program - S. Campana](#)

Attempt of a summary of energy/power needs of WLCG with extrapolations

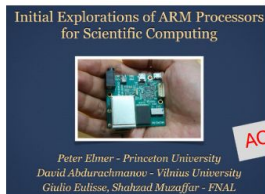
“ARM looks like a potential step-changing technology”

ARM as next Technology Step ?



ARM (Advanced RISC Machine) chips have low power consumption and heat generation and used extensively in portable, battery-powered devices, such as smartphones, laptops.

LHC experiments have kept an eye on this technology for over a decade. But ...



| Benchmarks - Simulation | | | | | |
|---------------------------|-------|-------|---------------------|---------------------|--|
| Type | Cores | Power | Events/ min/core | Events/ min/Watt | |
| Esynos4412 Prime @ 704MHz | 4 | 4W? | 1.14 | 1.14 | |
| Xeon E5-2620 @ 2.27GHz | 2x4 | 120W? | 3.50 | 0.23 | |
| Xeon E5-2630L @ 2.06GHz | 2x6 | 190W? | 3.33 | 0.21 | |

ACAT 2013

☹ Low power but slow in 2013

☹ Porting HEP software on non-X86 arch. non trivial

Conclusions



- This study shows the trends and does not pretend to make predictions
- The energy needs in HEP computing can be kept under control leveraging four pillars
 - The modernization of the facilities, going in the direction of more energy efficiency. Major capital investment
 - The improvements in the software and computing models. A gradual process bringing early benefits
 - The improvement in the hardware technologies and the optimization of the hardware lifecycle strategy. We need to invest in software portability
 - Turning off the air conditioning at CHEP
- I focused on the first three. Each pillar is important, but the improvements in software and computing models are an area where everyone in the WLCG community can contribute and where the largest gains should be expected
- In all scenarios, GWh/fb⁻¹ decreases over time: more physics per kW

😊 Recent huge increase in performance of mobile devices over last 10 years

😊 LHC experiments have software releases for ARM (used in some HPC) and have done some (but not all) physics validation



[The ATLAS experiment software on ARM - J. Elmsheuser](#)

Storage

[An HTTP REST API for Tape-backed Storage - J. Afonso](#)

End of SRM, collaboration between EOSCTA, dCache, StoRM and FTS - [spec document](#)

First production deployments already exist: EOSCTA, dCache/HPSS

[Challenging the economy of tape storage with the disk-based one - S. Ahn](#)

TCO/Comparison between a disk-based archive and tape-based archive (up to half-price) given current hardware estimates

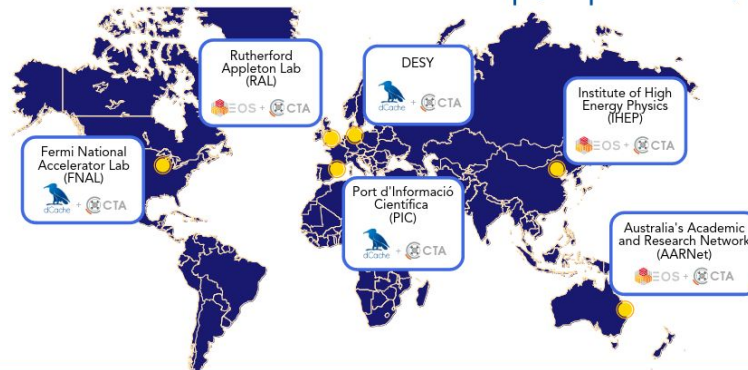
[CTA beyond CERN - M. Davis](#) - Supports both EOS and dCache frontends

The tape software landscape is consolidating:

- Changing license model/costs for commercial solutions

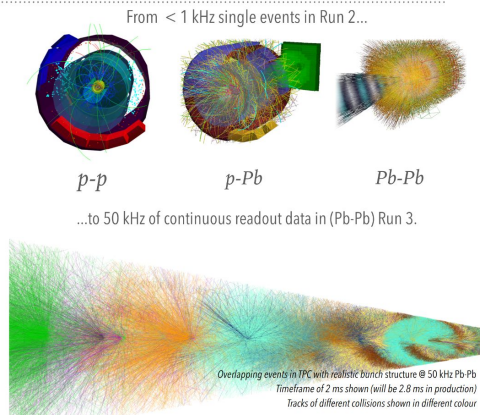
- Some free software solutions at end-of-life (CASTOR, Enstore, ...)

CTA Sites (7th EOS Workshop, Apr. 2023)



CHALLENGES FOR ALICE IN RUN 3

- **Completely new detector readout and substantial detector upgrades:** new ITS, MFT, FIT. New GEM for TPC readout.
- Reconstruct TPC data in **continuous readout** in combination with triggered detectors.
- Reconstruct **O(100x) more events online**.
- **Store O(100x) more events** (needs factor 36x for TPC compression). Cannot store all raw data, use **GPUs to do compression online**.
- WLCG **"flat budget"** scenario (4x more resources over 10 years, for 100x more events). **Use online GPU farm offline to speedup processing**.



- One integrated system, from data taking to final reconstruction (and beyond)
- No trigger, all Pb-Pb collisions recorded
- Continuous readout recording time frames instead of events
- 100x more collisions, much more data
- Cannot store all raw data → online compression
- Use GPUs to speed up online (and offline) processing
- Synchronous processing during data taking in the Event Processing Node (EPN) online computing farm
- When no beam in the LHC, EPNs are used for asynchronous (offline) processing. Asynchronous processing also on the GRID.

GPU usage in ALICE in the past

- ALICE has a long history of GPU usage in the online systems, and since 2023 also for offline:



Conclusions

- ALICE employs GPUs heavily to speed up online and offline processing.
 - 99% of **synchronous reconstruction** on the GPU (no reason at all to port the rest).
 - Today ~60% of full **asynchronous processing** (for 650 kHz pp) on GPU (if offline jobs on the EPN farm).
 - Will increase to 80% with full barrel tracking (**optimistic scenario**).
- **Synchronous processing successful in 2021 - 2023.**
 - pp data taking and **low-IR Pb-Pb** went **smooth** and as expected, but not causing full compute load.
 - **Full rate** will come with Pb-Pb in **October 2023**.
 - 50 kHz Pb-Pb processing **validated** with data replay of MC data (~ 30% margin).
- **Asynchronous reconstruction** has started, processing the TPC reconstruction on the GPUs in the EPN farm, and in CPU-only style on the CERN GRID site.
 - EPN nodes are **2.51x** faster when using GPUs.

BigPanDAmon system

- The BigPanDAmon system is an essential part of the monitoring infrastructure for ATLAS providing a wide range of views from the top-level summaries to a single computational job and its logs;
- Over the past few years of the PanDA WMS advancement in the ATLAS experiment several new components: Harvester, iDDS, Data Carousel, and Global Shares;
- Relevant data from all PanDA WMS components and accompanying services are accumulated and displayed in the form of interactive charts and tables.

Project structure

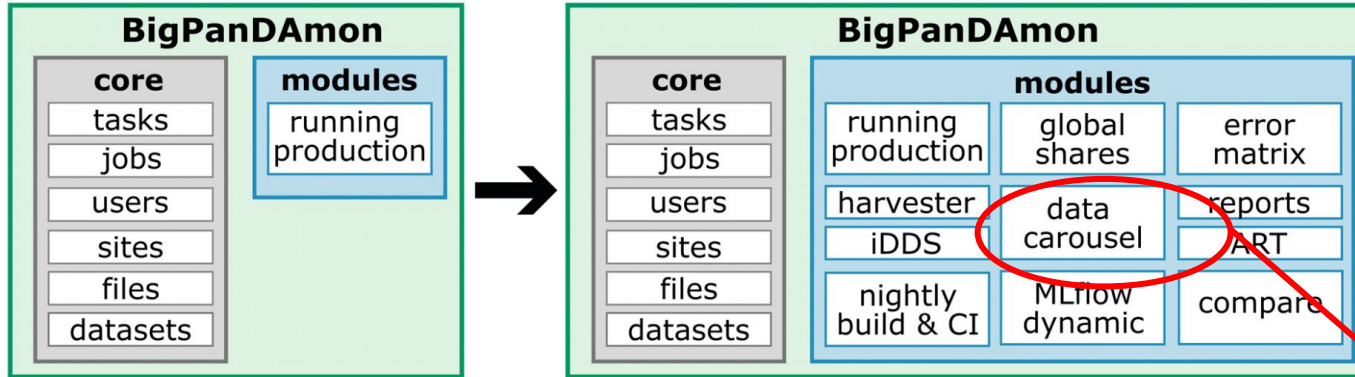


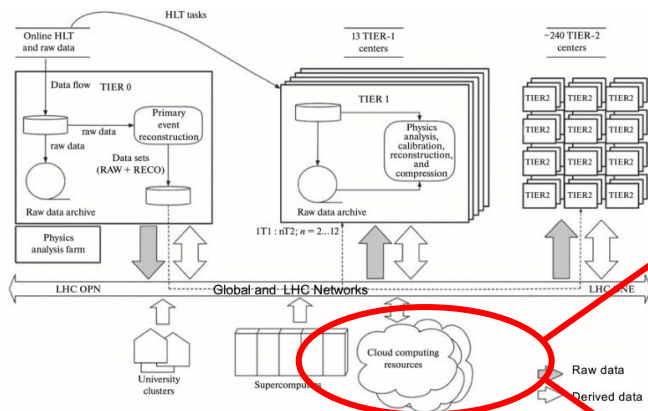
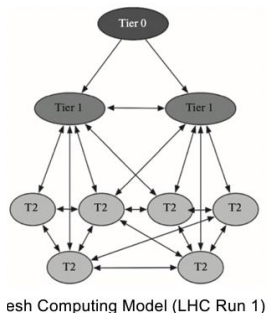
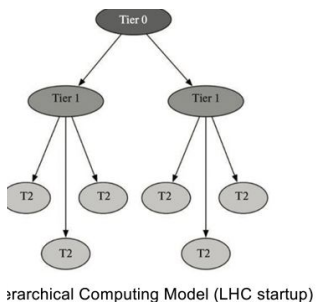
Figure 3 – BigPanDAmon structure evolution from 2017 to 2023

<https://indico.jlab.org/event/459/contributions/11474/>

<https://indico.jlab.org/event/459/contributions/11307/>

ATLAS Computing Model Evolution

- Heterogeneity of computing resources increased dramatically after/during Run2:
 - 1M+ payloads per day are executed via supercomputers, grids, and clouds;
 - Workflow complexity is growing.



Computing model implemented for the LHC Run 2 and Run 3

LHC Run 1 : 2010-2012; Run2 : 2015-2018; Run3 started at 2022

- Solutions developed in ATLAS with the PanDA system:

<https://indico.jlab.org/event/459/contributions/11482/>

- In the past few years ATLAS evaluated **commercial clouds** as an additional part of their computing resources:

- Amazon;
- Google.

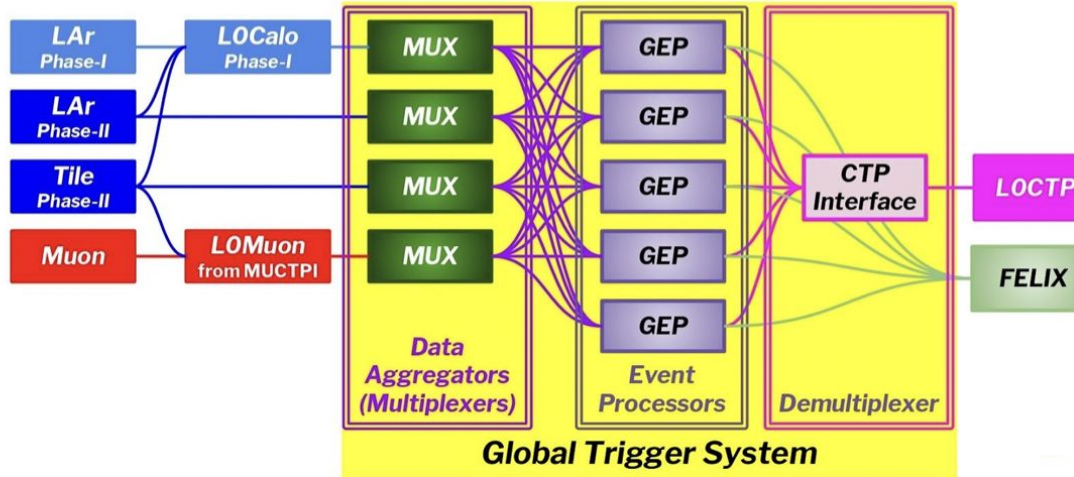
<https://indico.jlab.org/event/459/contributions/11636/>

- cloud-native deployments on K8s:

<https://indico.jlab.org/event/459/contributions/11626/>

<https://indico.jlab.org/event/459/contributions/11465/>

ATLAS Trigger & DAQ upgrades



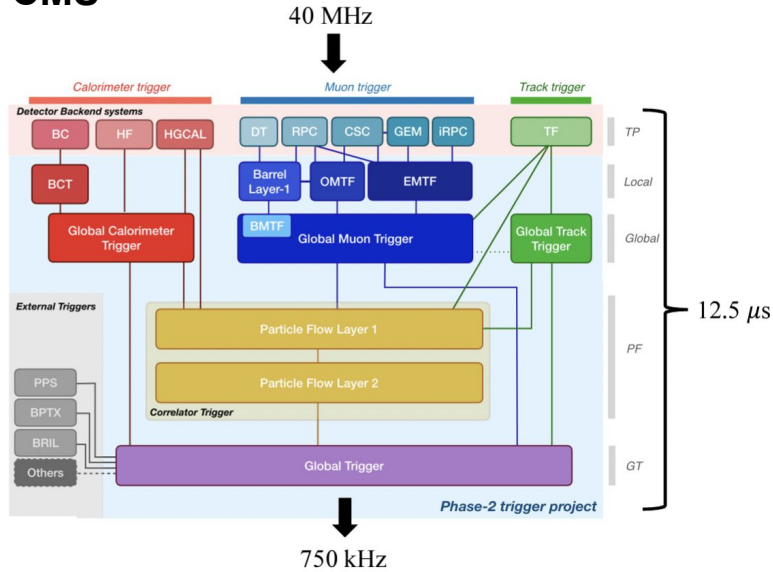
<https://indico.jlab.org/event/459/contributions/11368/>

- The detector upgrades themselves also present new requirements and opportunities for the trigger and data acquisition system;
- The design of the TDAQ upgrade comprises several different aspects (Level-0 Global Trigger, Central Trigger, High Level Trigger, Readout, Event Filter, etc);
- Upgrades based on a mix of commodity and custom solutions:
 - Most projects already passed many reviews;
 - Prototypes available for many projects.

Trigger and/or DAQ: an overview

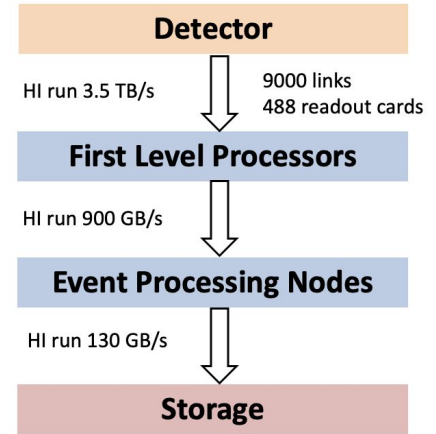
<https://indico.jlab.org/event/459/contributions/11361/>

CMS

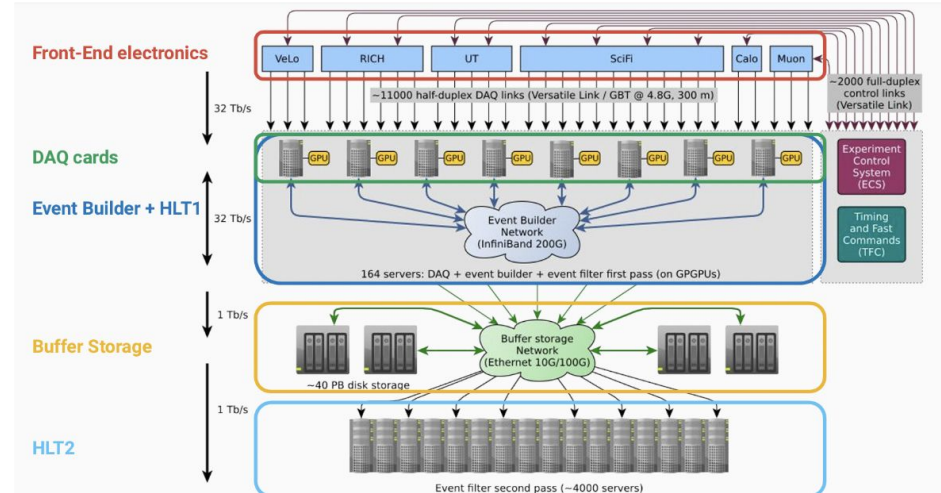


<https://indico.jlab.org/event/459/contributions/11401/>

ALICE



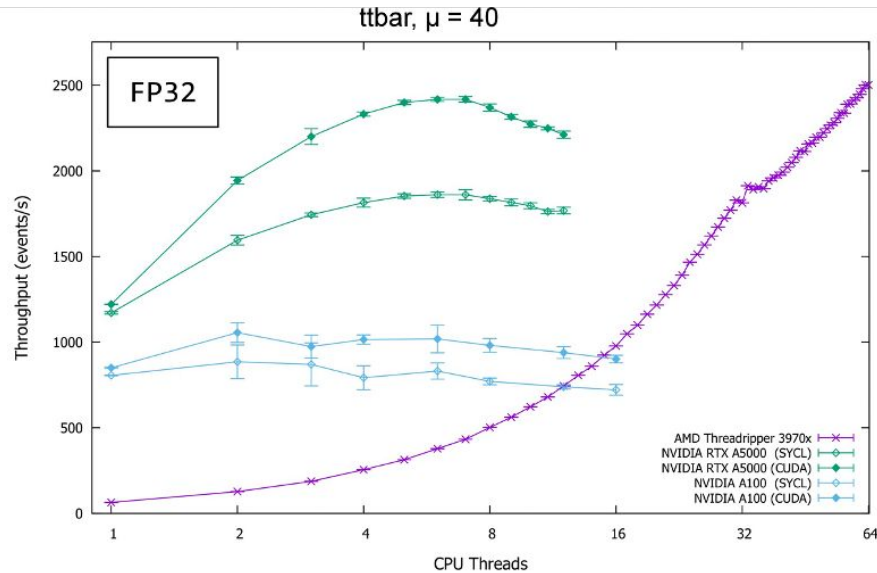
LHCb



<https://indico.jlab.org/event/459/contributions/11394/>

Using GPUs for Tracking

ATLAS tracc project: a close to single-source track reconstruction demonstrator for CPU and GPU.



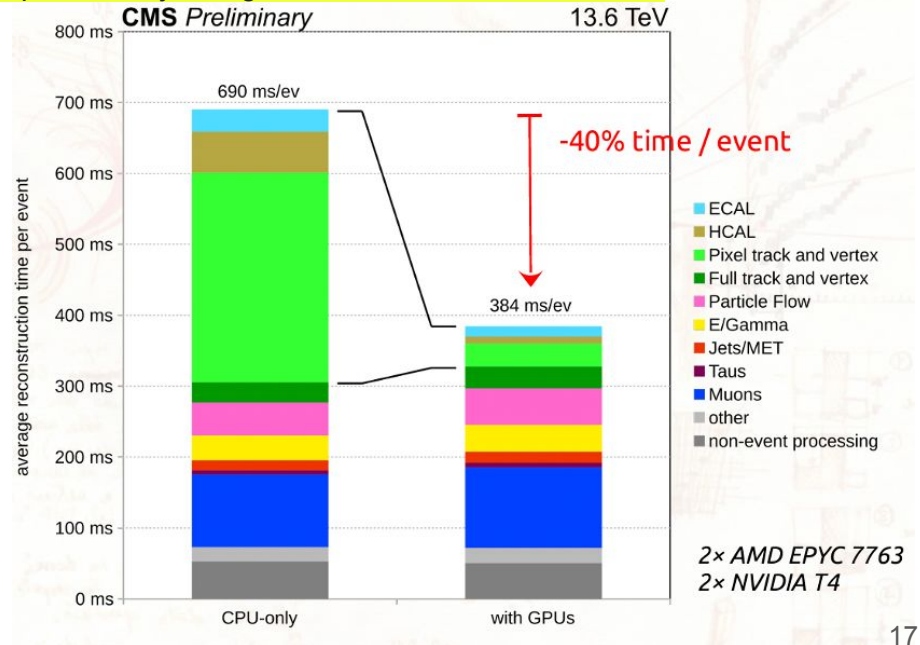
In **CMS** an intense activity in this field is ongoing:

<https://indico.jlab.org/event/459/contributions/11402/>

<https://indico.jlab.org/event/459/contributions/11825/>

<https://indico.jlab.org/event/459/contributions/11822/>

<https://indico.jlab.org/event/459/contributions/11412/>



ATLAS data formats for the future Runs

File sizes ([kB per event], using the current Run 3 prototype):

| Actual size | Run 2 MC $t\bar{t}$ | Run 3 MC $t\bar{t}$ | data17 |
|-------------|---------------------|---------------------|--------|
| PHYS | 34.2 | 40.7 | 21.7 |
| PHYSLITE | 13.6 | 16.3 | 6.2 |

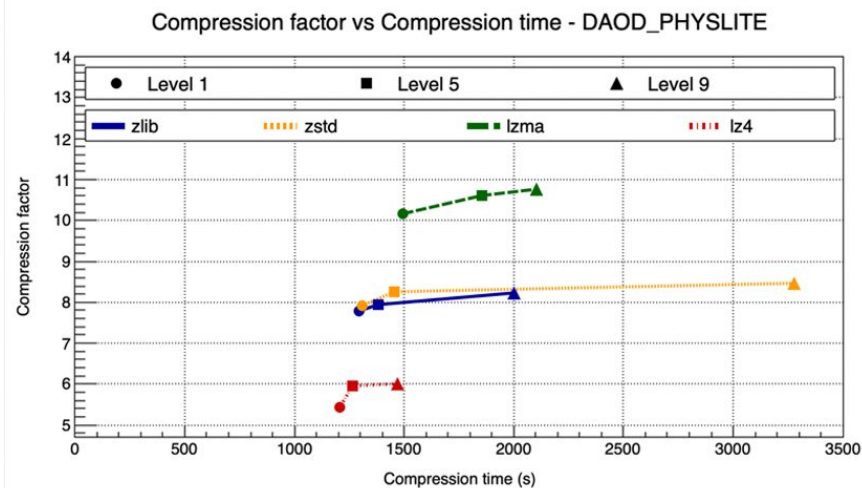
| Target size | MC | Data |
|-------------|----|------|
| PHYS | 50 | 30 |
| PHYSLITE | 12 | 10 |

Work is ongoing to further reduce PHYSLITE size

<https://indico.jlab.org/event/459/contributions/11586/>

- AUGMENTATION: It is possible to add event augmentations to ATLAS standard data products (such as DAOD-PHYS or PHYSLITE) avoiding duplication and limiting their size increase.

<https://indico.jlab.org/event/459/contributions/11422/>



<https://indico.jlab.org/event/459/contributions/11429/>

Some (other) suggestions


- **Track n. 6** -> Physics Analysis Tools Summary

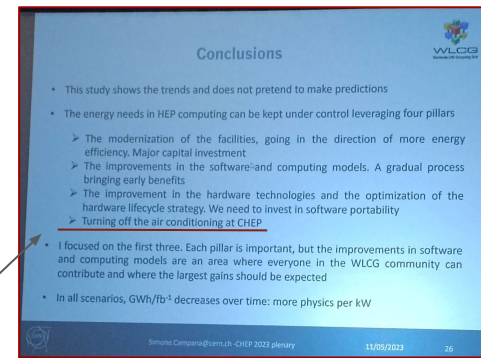
<https://indico.jlab.org/event/459/contributions/12627/>

- **Track n. 3** -> Offline computing (Simulation, Reconstruction, Data Preparation and Physics Performance)

<https://indico.jlab.org/event/459/contributions/12624/>

General feeling / Various Issues

- Air conditioning in the rooms - too cold!
 - at some point it was set to 16.8 degrees Celsius!
- No power plugs in the rooms 
- Strict 12 minutes for talk, 3 for questions - no time for actual discussions
- Rooms for parallel tracks were distributed in 3 floors, difficult to jump sessions
- VISA for Russian colleagues many of whom are affiliated with an American uni/lab - many talks/posters had to be given by their colleagues
- No lunch provided, for this reason lunch breaks were long (1.5hrs)
 - Norfolk's restaurants were definitely **NOT** ready for ~600 people storming the streets at the same time trying to have lunch *quickly* in order to return to parallel sessions on time



Summary

Encourage you to check the [agenda](#) yourself for more!



See you at CHEP 2024 : Krakow, Poland
(Oct 19-25, 2024)