Computing Beyond the Trigger



Xavier Espinal

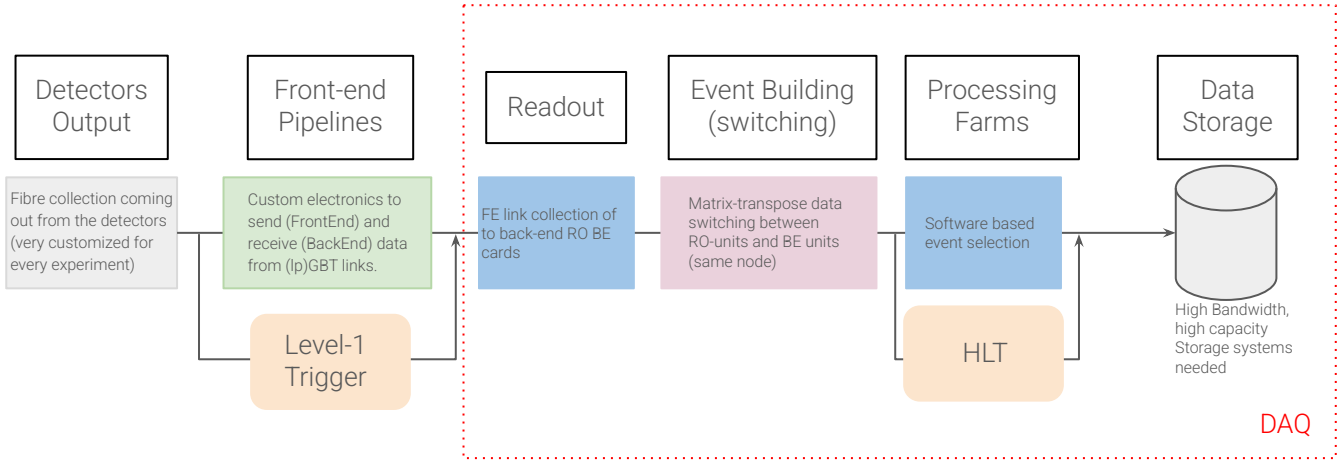
CTO, CERN-IT

** living a life of extreme simplicity, strict discipline, and self-denial without comfort or luxury*

with precious input from my DAQ colleagues: G. Avolio , V. Barroso, N. Neufeld and W. Vandelli

Computing for the next generation of DAQ

Higher readout rates and real-time processing are changing the computing requirements of DAQ.



What does this mean for computing infrastructures?

From online/offline to a computing continuum (1/3)



RUN 1 and 2 - Clear separation:

- Triggering and event building close to detector. Reconstruction predominantly offline.

  Detector → Trigger → HLT → Storage → Offline
- Dedicated online/offline infrastructures.
- DAQ location strongly coupled to functions.

From online/offline to a computing continuum (2/3)



RUN 3 and 4 - Blurring boundaries

- Trigger-less: CPU+GPU farms for real-time event selection. Pioneered merging online/offline architectures combining continuous readout with GPU processing and large remote storage pools.
- Trigger based: cases where there is no gain in physics-reach by moving to Streaming RO (lots of rejected events, eg. 1 MHz to 1 kHZ). Future combination of FPGA with ML algorithms.
- HLT shifting to "quasi-ofline": high-bandwidth networks and heterogeneous computing increasingly allow high level triggers, data selection and data reduction to be performed in real time and potentially far from the detector.

From online/offline to a computing continuum (3/3)



RUN 5+ DAQ location becoming secondary - distribute continuum?

- Evolution towards continuous real-time processing?

  Detector ⇄ High-bandwidth network ⇄ Distributed compute/storage

- Processing can move across a distributed computing infrastructure. Transient storage becomes a distributed service. Resuorces can potentially be shared dynamically (data taking vs. shutdowns).
- Location matters less, timing and processing workflows dependencies matter more.

The online/offline boundary evolves into a continuum of services? is the fundamental distinction becoming **Latency vs. Locality vs. Permanence**(Data)?

xavier.espinal@cern.ch

Different DAQ strategies: converging requirements

ALICE/LHCb: Streaming / triggerless

- Continuous readout
- CPU/GPU reconstruction
- Real-time selection, time-frames

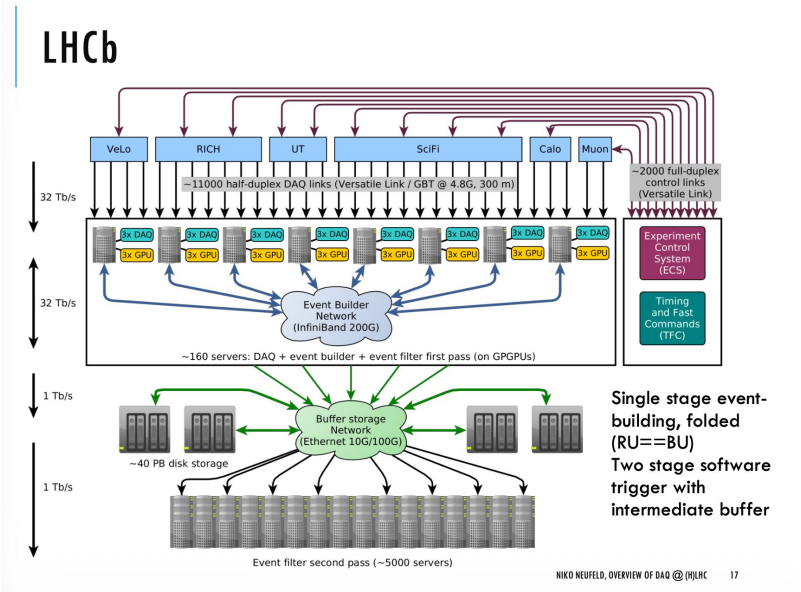
ATLAS/CMS: High-rate triggering

- Much higher processing rates
- FPGA/ML/hardware selection
- Sophisticated real-time processing

More drastic heterogeneous Compute:

- GPU usage extends far beyond ML/AI, core reconstruction tracking, clustering
- Emerging platforms combining tightly CPUs, GPUs and FPGAs.

	DAQ rate	Event size	EB B/W	LAN technology
ALICE	50 kHz	12 MB	0.7 TB/s	InfiniBand HDR
ATLAS Run3	90 kHz		55 GB/s	Ethernet / TCP/IP (100G)
ATLAS Run4	1 MHz	5 MB	5 TB/s	Ethernet / TCP/IP (400G)
CMS Run3	100 kHz	2 MB	0.2 TB/s	Ethernet / RoCE (100G)
CMS Run4	750 kHz	8.4 M	6.3 TB/s	Ethernet / RoCE (400G)
LHCb	40 MHz	60 kB	2.4 TB/s	InfiniBand HDR



xavier.espinal@cern.ch

Where Computing goes? *Latency vs. Locality vs. Data*

Traditional DAQ concepts had clear mappings:

- Low latency (µs to ms) → Local Processing (pits)→ Transient Storage (in-flight, buffered)
- High latency (s to m to h) → Remote (CERN-IT, Distributed/WLCG) → Persistent Storage (disk, tape)

New architectures could progressively break this model:

- HLT and real-time reconstruction: low latency, but potentially remote?
    - ATLAS Run 4 will operate half of HLT at CERN-IT (3 km away). This demonstrates real-time processing becoming integrated directly with large-scale, remote computing infrastructures.
- Data buffering: transient, but potentially geographically distributed?
- Calibration: traditionally offline, but increasingly part of real-time processing?
- Analysis: traditionally asynchronous, but can use the same resources as DAQ?
- Storage: thanks to network evolution, is no longer necessarily the boundary between online and offline?

xavier.espinal@cern.ch

The challenge is no longer only performance...

- Can we continue affording computing for physics experiments?

- Do we have the personpower to continue operating the infrastructures?

xavier.espinal@cern.ch

The challenge is no longer only performance

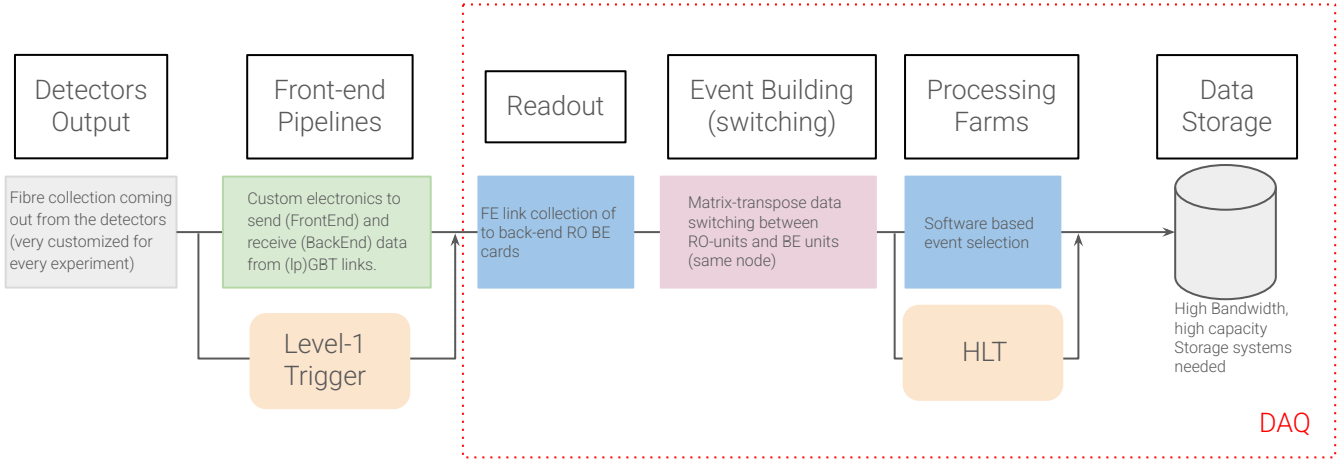
- Can we continue affording computing for physics experiments?
  - => *Commodity hardware, Resource sharing, Hardware, Lifecycle , Networking, Power and Cooling optimizations.*

- Do we have the personpower to continue operating the infrastructures?
  - => *Common frameworks, Common software, Monitoring & Reliability.*

DAQ is becoming a large-scale computing infrastructure challenge.

xavier.espinal@cern.ch

Can we run DAQ on generic computing ecosystems?



Challenge #1

- **Common** hardware across experiment for Read-Out.

- Enable use of **commodity hardware** for Read-Out nodes and Event Processing

- Possibility to increase #fibres/node and **transition** from PCI-express to **FPGA or SoC**.

- **Transition** from infiniband => **Ethernet, RoCE** architectures

Challenge #2

- **Transition** to **silicon-photonics** to diminish Copper-ball effects.

- Future potential to have direct links to EB (avoiding Read-Out)?

- **Radiation-hard** is still an issue, both economical (specs are niche) and technological (interference, noise).

Challenge #3

- Invest on **common software** activities to maximise economies of scale and increase off-the-shelf hardware usage.

- Active collaboration across experiments for common software solutions and frameworks, eg. **control systems homogeneity** based based on K8s, Common Tracking Software (ACTS)

Challenge #4

- AI/ML platforms to overcome personpower shortage in detector operations (agentic AI, LLMs, etc)

- Resource optimisation and sustainability for operations (personpower, expertise)

- Dynamic re-use of hardware: online vs offline activities.

DAQ meets Data Center challenges

- The technologies driving AI and hyperscale computing are reshaping the hardware market:
  - Computing capacity planning can no longer rely on flat-budget funding.
  - AI and hyperscalers driving the technical evolution of the computing market, and indirectly modelling costs.
  - Markets becoming crazy, Moore's-like laws defeated, and commodity hardware future is unclear.
- Infrastructure evolution costs also becoming first-order constraints: power densities increasing within fixed power envelopes and cooling constraints.
  - Key scalability enablers: network technology, interoperability, and services commonalities.
- **Optimizing the usage of our resources is a top priority.**

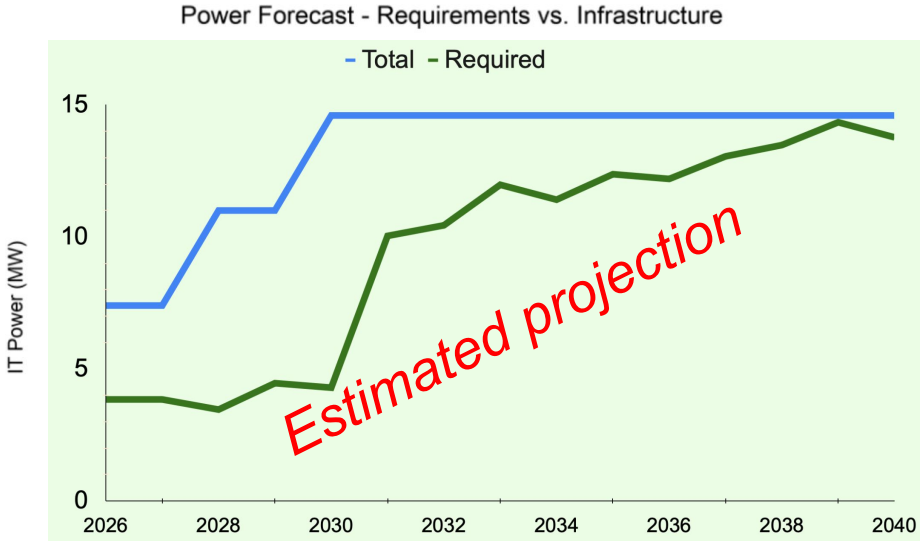
1) Can we adapt to a computing industry increasingly optimized for AI and hyperscalers?

2) Infrastructure lag: can we keep-up while capacity requirements grow faster than our economic and technological ability to expand capacity?

xavier.espinal@cern.ch

Power becoming an increasing constraint

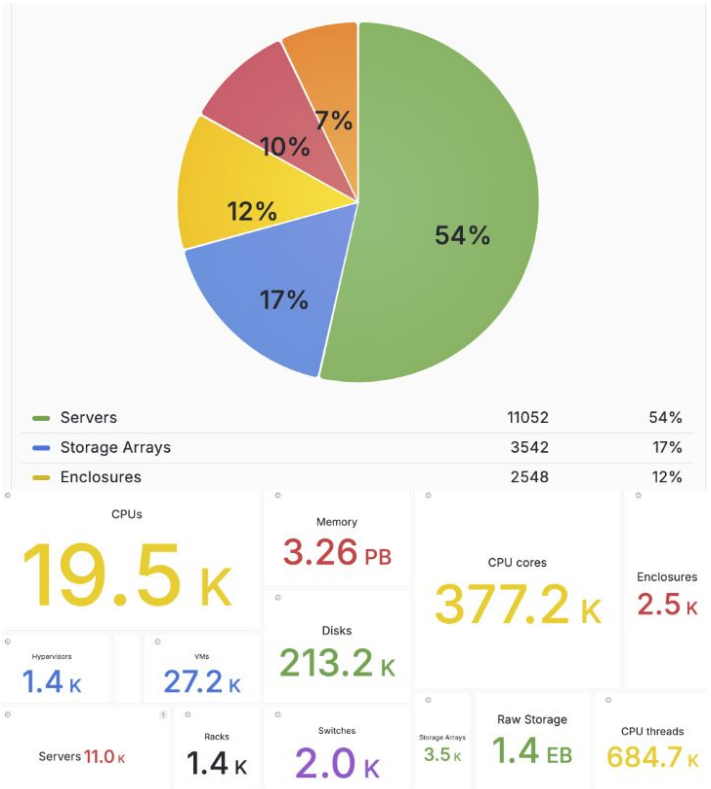
- Data-centre growth is increasingly limited by electrical and thermal dissipation capacity rather than simple floor space.
- Typical infrastructure is planned for peak consumptions, can leaving substantial power capacity unused.
- Optimizing around observed average loads can unlock significant power.
    - Overcommitment with power capping, CPU throttling, and workload-aware placement could maximize power usage.



Building/expanding data centers is lots of money and time, in some cases unlikely. Infrastructure capacity growth strongly coupled to maximise power extraction, allocation strategies and **hardware lifecycle/procurement policies**.

Hyperconvergence: Extracting more value from existing infrastructures

- Hyperconvergence direct impact on TCO costs: additional capacity with reduced rack footprint, power consumption and maintenance.
- HL-LHC required compute capacity correspond to roughly >10k nodes at the Tier-0.
- Tighter coupling between services: stateless services as a *remora* for nodes running stateful services.
- But this comes at the cost of **increased complexity running services**: co-existence (updates, maintenance), lifecycle, and performance management complexity.



Rethinking Hardware Lifecycles

- Infrastructure agility becoming an strategic asset nowadays.
- Scientific infrastructures operate (funded) on much longer timescales than technology markets waves.
- Storage and compute technologies are undergoing significant structural changes: power densities, AI-driven evolution of technology.
- Delaying large procurements benefit from future price vs. performance improvements and navigate(hope is true) through market cost peaks.
- Retaining selected hardware for longer increase reserve capacity translating into operational flexibility to absorb *surprises* in accelerator/experiment plans and computing models.
- Longer lifecycles require enhanced monitoring, risk management, and support strategies.

CERN-IT Hardware Lifecycle Strategy for LS3 and Run 4

Xavier Espinal, v6-17.11.2025

Introduction

The latest schedule for HL-LHC assumes that Run 4 will start June 2030, with limited resources (short run) in the first year, followed by a full luminosity run period from 2031 to 2033 (included).

The main goal of the LHC project is to achieve an integrated luminosity of **3/ab** by the end of the LHC accelerator operations in 2041. It should be noted that 83% of the total expected luminosity will be collected during the High-Lumi period 2031-2041.

Both offline and online infrastructures should cope with luminosities ranging between 200/fb to 330/fb per year, and a pile-up (mean number of interactions) between 130 and 170 per bunch crossing.

The expected rise in accelerator performance and data will require a significant expansion of computing resources. However, current market and technology projections suggest that fulfilling these requirements within the constraints of a flat budget will be challenging. As a mitigation strategy, we propose adopting a revised hardware lifecycle approach that offers greater flexibility to adapt to a variety of future scenarios.

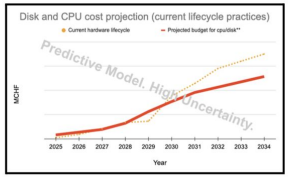


***Figure 1**. (dotted-orange) Run 4 purchase estimates for Disk and CPU vs (solid-red) flat budget . *Hardware cost estimation is based on the current lifecycle practice (with an average of 7 years). **The projected budget allocation for CPU and Disk is extrapolated taking into account the average over the total for the last 10 years (40%). [Disclaimer: The trends and projections shown in this figure are based on a set of assumptions and modeled scenarios that carry significant uncertainties. Actual outcomes may differ substantially due to factors beyond our control, including market dynamics, technological developments, and evolving computing models].*

It should be noted that the proposed lifecycle strategy is specifically designed for the functions and operations of CERN Tier-0, where services are provided to a defined user community and rely on highly specialized software stacks, such as our in-house storage system. Consequently, the strategy is closely aligned with the characteristics and usage patterns at CERN. Therefore, applying the same approach to other scientific computing facilities would require a detailed, case by case analysis, and should not be directly extrapolated.

Goal and scope: a guideline for a yearly rolling plan

During the long shutdown (LS3) no (or marginal) increase in computing resources is expected from the experiments. This introduces the possibility to review the hardware lifecycle strategy with the following following main goals:

- Reduce significantly large procurements during LS3, allowing to benefit from price/performance evolution.

Historically, hardware replacement was largely driven by age. Going forward, lifecycle decisions must be driven by market evolution, experiment requirements, operational risk, and economic opportunity.

xavier.espinal@cern.ch

Storage & Compute Lifecycles

DISK STORAGE LIFECYCLE

Degradation & Replacement

- O(100k) drives. Reliability degrades after 7-8 years (AFR doubles from 0.5% to 1%). **A 10-year replacement cycle balances cost and operational overhead.**

Operational Strategies

- Reactive: Swap on failure (0.75 FTE operational impact).
- Run to Fail (RtF): No swap; requires 12PB/year extra capacity.
- Lazy: Periodic batch replacement (accumulates ~1.0 PB/month).

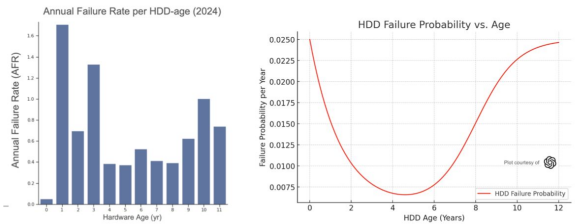


Figure 11: Left: HDD AFR 2024 in CERN IT, weighted by disk population in each year bin. Right: Typical bath-tube like shape (AI generated. It is difficult to get reliable world-wide absolute disk failure statistics over time).

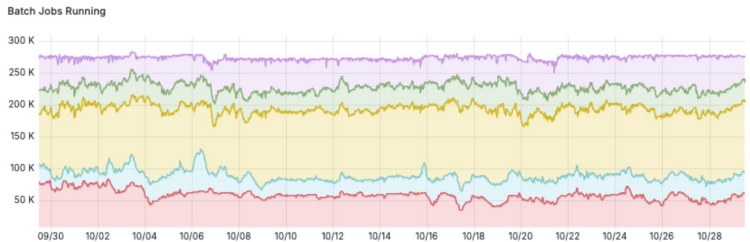
DATA PROCESSING LIFECYCLE

Stateless Ephemeral Compute

- Tens of thousands of batch servers. Systemic failures are dominated by platform components (motherboards, memory, PDUs). AFR is well under 0.5%.

Failure Mitigation & Lifespan

- Workload Rescheduling: Resolves transient server outages automatically.
- **10-Year Lifespan: Captures full value. Run to Fail (RtF) is viable.**
- Interventions: ~50/year, requiring ~3h each (~150 hours/year total).



xavier.espinal@cern.ch

Core & Network Lifecycles

CORE INFRASTRUCTURE LIFECYCLE

Private Cloud & Statefulness

- Hosts stateful platforms, heavily user facing. A 5 year lifecycle minimizes heterogeneity, simplifies maintenance, and leverages hardware evolution gains.

Key Strategies & Reliability

- Progressive Refresh: Deploy continuously to avoid large-scale upgrades.
- Platform Resiliency: Failures linked to platform parts; AFR remains well under 0.5%.
- **Stateless Shift:** Systems can transition to stateless setups after ~5 years.= (eg. batch systems)

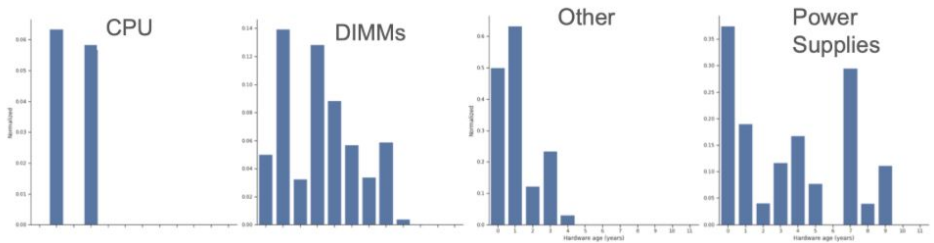
NETWORK INFRASTRUCTURE LIFECYCLE

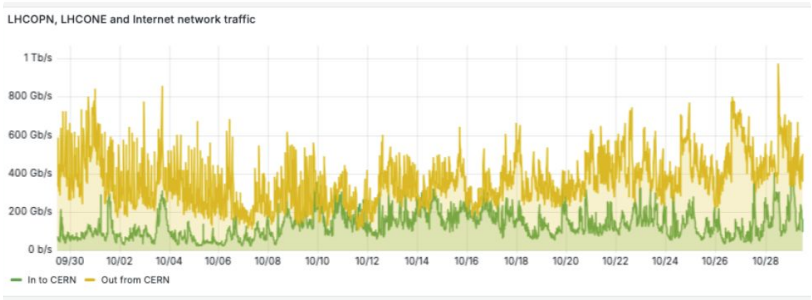
Accelerator Constraints & Longevity

- Constrained by accelerator run periods and Long Shutdowns (LS). Requires proactive, advance procurement rather than delayed purchasing.

Lifecycle & Operations

- Misaligned Cadence: Cycles cannot flexibly track compute capacity shifts.
- Long-Term Support: Decadal vendor contracts are essential for stability.
- High Readiness: Requires deep testing and validation prior to beam runs.





HL-LHC R&D: preparing for the next decade

Explore solutions for a (very likely) resource constrained scenario, build the evidence for decision making.

A common objective: maximise capacity delivered per unit of €, power, space and operational effort.

The objective is not to optimise today's infrastructure, but to establish the evidence and technologies needed to make the right infrastructure decisions for HL-LHC.

Activity	Scope
POWER	More IT capacity per MW
TAPE	Meet Run 4 data recording needs with an optimised tape technology & infrastructure strategy
DISK	More usable storage per €
ARCHIVE	Consolidate and optimise archive workflows, infrastructure and operations
STORAGE I/O	Meet IO performance requirements, optimise IOPS per €
IT SERVERS	More IT capacity per MW, € and rack

xavier.espinal@cern.ch

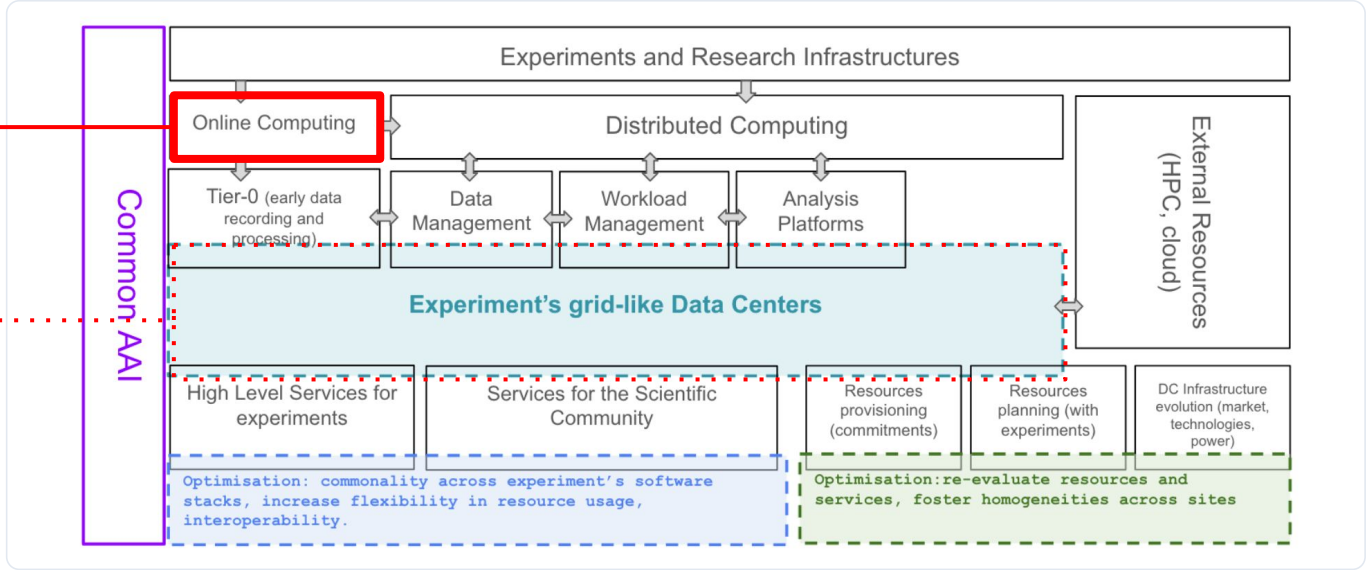
Summary & Outlook

- DAQ is moving towards heterogeneous, distributed computing. CPU, GPU and FPGA resources are increasingly part of the real-time data-processing chain.
- The online/offline boundary is becoming a computing continuum. Processing can increasingly be located according to latency, connectivity and data lifetime rather than proximity to the detector.
- Infrastructure efficiency becomes a physics-enabling requirement. Power, cooling, space, hardware cost and lifecycle increasingly constrain capacity.
- CERN needs optimisation across the complete data lifecycle. From real-time processing and transient storage to disk, tape and analysis. Capacity and performance must be matched to workload and data lifetime.
- Flexibility, efficiency and common infrastructure will be key to accommodating the next generation of DAQ.

xavier.espinal@cern.ch

Backup

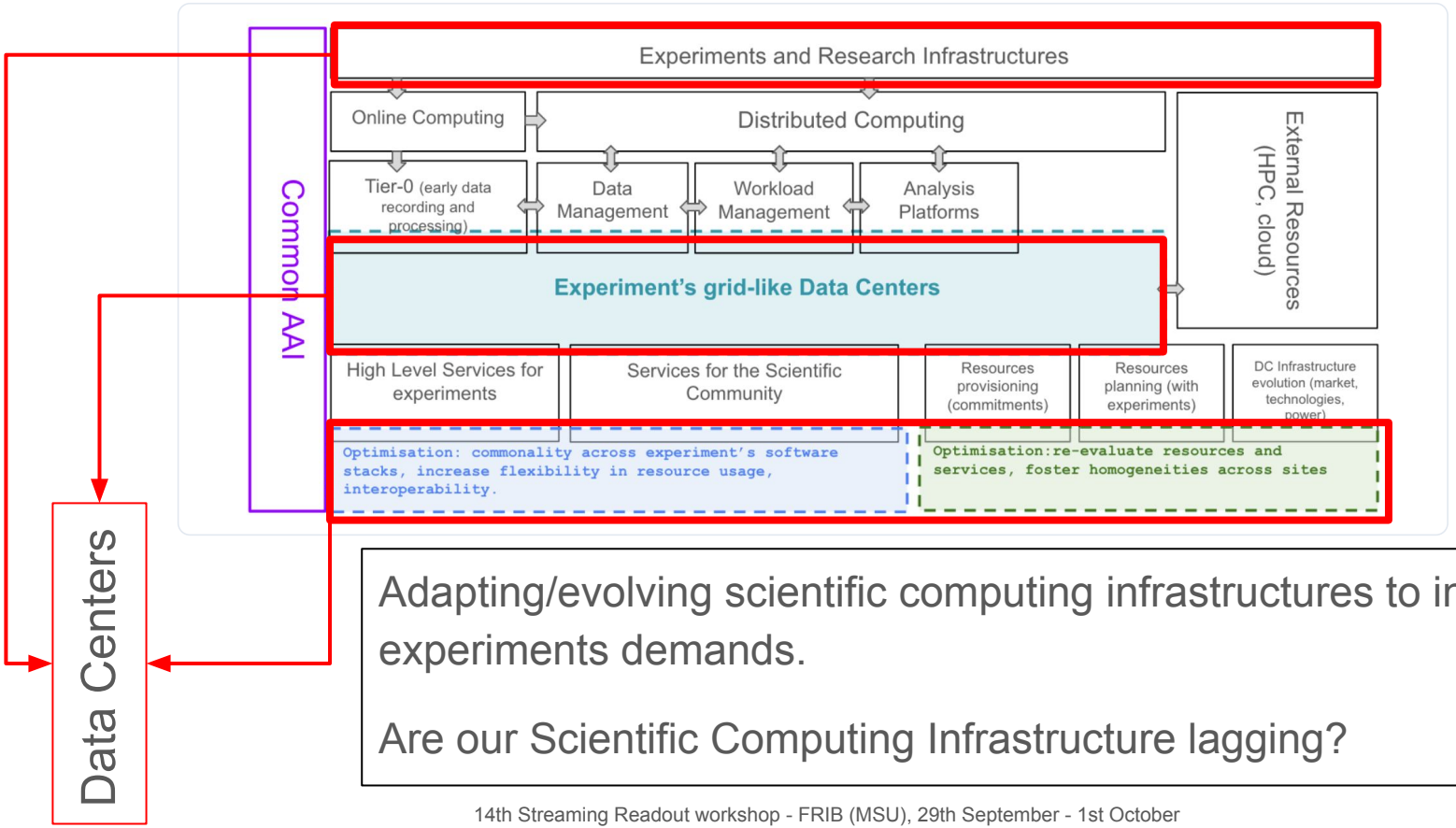
The shift in Scientific Computing Infrastructures



Event filtering and processing moving closer to DAQ.

DAQ increasingly moving closer to Data Centers. Where is the border between offline and distributed computing?

The shift in Scientific Computing Infrastructures



Adapting/evolving scientific computing infrastructures to increasing experiments demands.

Are our Scientific Computing Infrastructure lagging?

xavier.espinal@cern.ch

Online Storage: Capacity is not the only problem

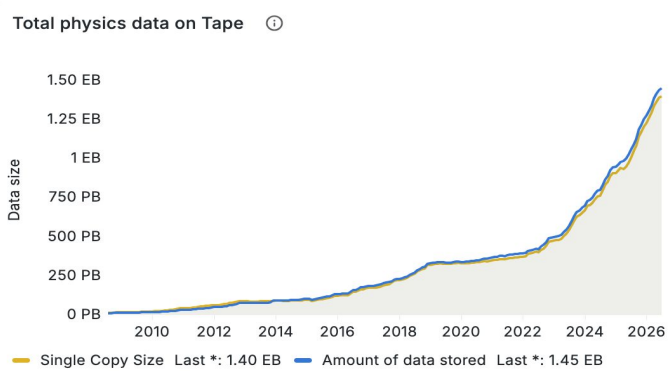
- Large storage systems are increasingly constrained by data access performance rather than raw capacity.
- HDD capacity continues to grow, but IOPS and JBoD bandwidth do not scale at the same pace.
- Erasure Coding configurations reduce overhead costs but increase pressure on LAN.
- SSD technology is rapidly improving in capacity and energy efficiency… but prices skyrocketing.
    - Hybrid SSD/HDD architectures could enable tiered storage and differentiated QoS models.

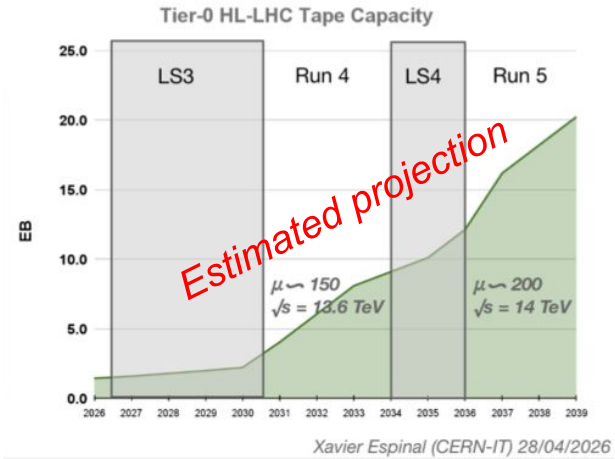
The storage bottleneck is shifting from storing data to efficiently serving data. Future storage systems will not be built around a single technology, but combine flash and disk to balance performance vs. capacity, energy consumption and space… (unfortunately we are at the mercy of market costs)

xavier.espinal@cern.ch

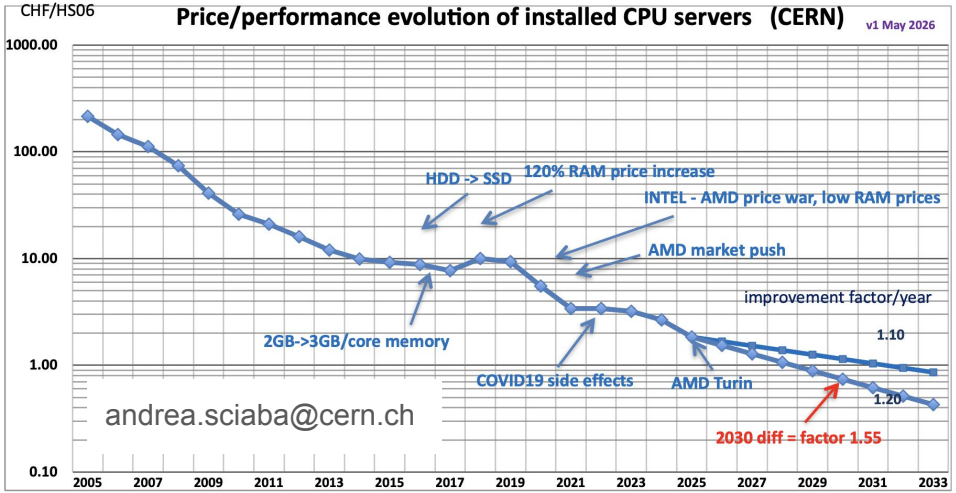
Tape storage remains strategic, but is evolving (disappearing?)

- HL-LHC-scale infrastructures will require tens of exabytes of archival storage.
- Traditional large tape libraries maximize throughput and operational efficiency, but future uncertain (?)
- New in-rack tape libraries offer attractive modular growth, improved rack integration, and enhanced cyber-resilience.
  - Different tape architectures (libary vs rack) can support differentiated archive (QoS) and nearline services.
  - But bring in several issues as well (tape drives fragmentation, cartridge management, etc.)
- Future tape infrastructures will likely combine technologies … with an increased use of *cheap* disk as archive (heavily EC'd)

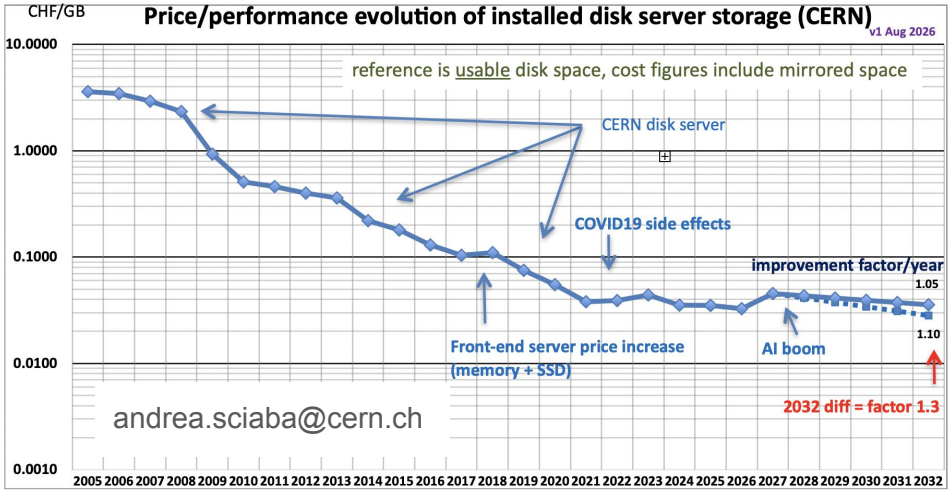




Price vs. Performance Evolution



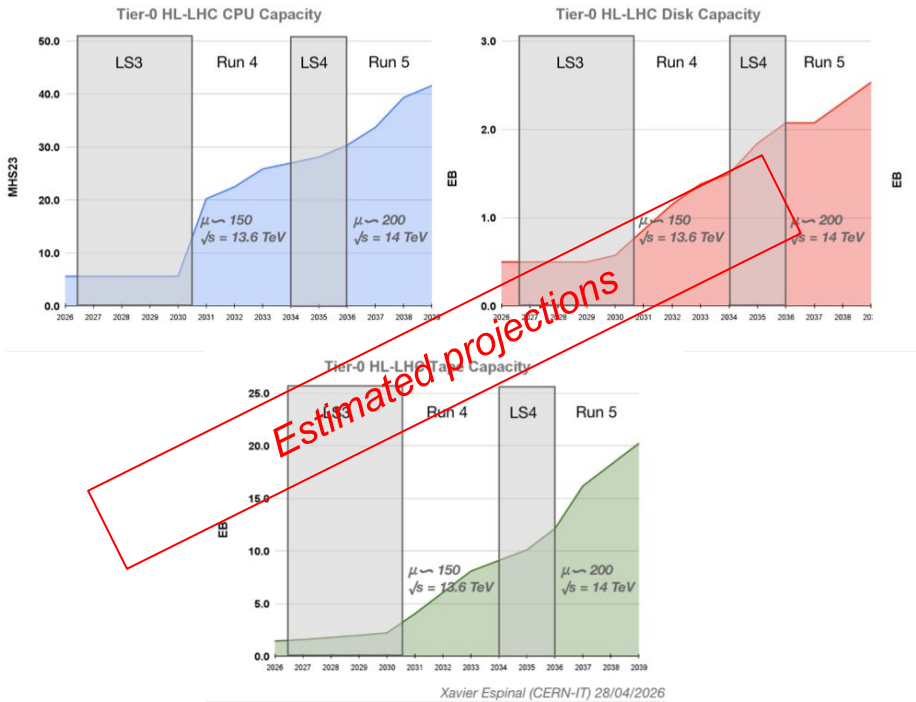
Last 5 year average improvement factor = 1.21



Last 5 year average worsening factor = 1.03

A Wave of Data-Intensive Experiments

- Multiple flagship experiments will enter operation over the next decade with increasingly larger data volumes and processing requirements.
- HL-LHC starts taking data in 2030, aiming for 3/ab integrated luminosity (83% of the total LHC program to be collected during HL).
- And besides HEP, experiments like SKAO, Rubin, CTAO, KM3NeT and Einstein Telescope will generate yearly EB-scale data streams and quasi real-time distributed processing.
- Computing challenges are similar across disciplines, creating a **common pressure on infrastructure providers**.

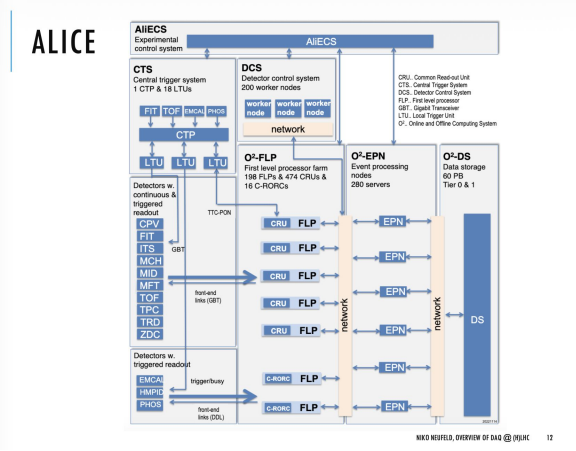


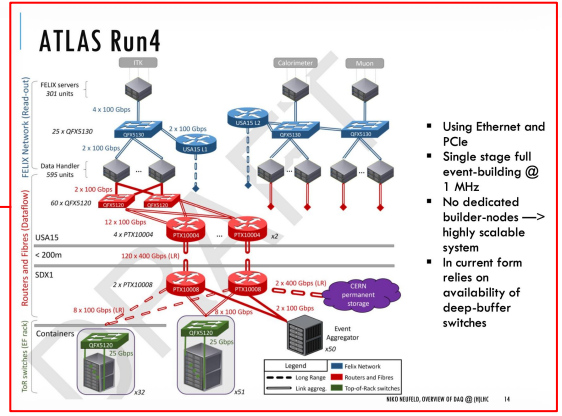
The challenge is not one experiment becoming larger. It is the simultaneous usage of the same facilities by multiple large-scale experiments with high computing demands.

Run 3 and Run 4 DAQ rates

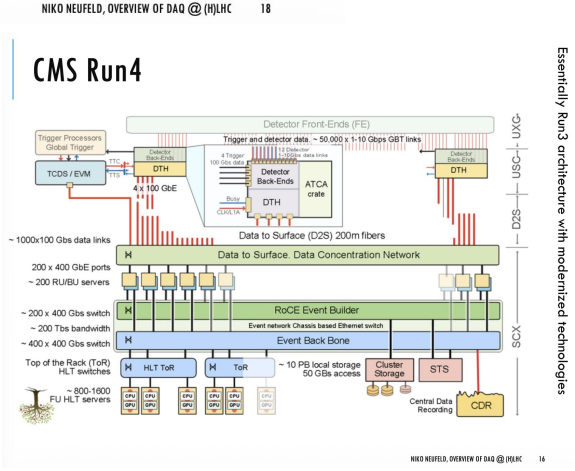
	DAQ rate	Event size	EB B/W	LAN technology
ALICE	50 kHz	12 MB	0.7 TB/s	InfiniBand HDR
ATLAS Run3	90 kHz		55 GB/s	Ethernet / TCP/IP (100G)
ATLAS Run4	1 MHz	5 MB	5 TB/s	Ethernet / TCP/IP (400G)
CMS Run3	100 kHz	2 MB	0.2 TB/s	Ethernet / RoCE (100G)
CMS Run4	750 kHz	8.4 M	6.3 TB/s	Ethernet / RoCE (400G)
LHCb	40 MHz	60 kB	2.4 TB/s	InfiniBand HDR

Numbers are approximate, actual installed network bandwidth is higher in all cases





50% of ATLAS HLT => CERN-IT Data Center



Essentially Run3 architecture with modernized technologies

