Readout Strategies for future high-rate LHC experiments

Using Commodity Hardware and Data Center Technologies

Seminar: Marietta Blau Institute, Vienna

Niko Neufeld, CERN/EP

April 17$^{\text{th}}$, 2026

Introduction: The challenge of Data Acquisition (DAQ) in large experiments

- **High data rates:** ATLAS and CMS for Phase 2 will be around 50 Tbit/s, SKAO about 20 Tbit/s after local processing, LHCb Upgrade 2 will be around 300 Tbit/s
- **Custom links on detector electronics are slow:** 5 to 10 Gbit/s, planned 25 Gbit/s
  - Why custom links? Legacy designs, power, radiation, reliability ...
- **Processing requires large number of GPUs, CPUs, FPGAs**: example LHCb: 500 GPUs, 5000 (older) dual-socket servers
- Use **rapidly evolving AI/compute hardware** in experiments running for **decades**

Introduction: The Shift in DAQ Paradigms

- **Traditional approach:** Custom ASICs, specialized backplanes (VME, xTCA), proprietary point-to-point links.
- **The challenge:** High development cost, limited upgrade paths, niche expertise required.
- **The modern shift:** Leveraging Commercial Off-The-Shelf (COTS) components and standard data center IT.
- **Why now?**
  - Exponential growth in standard IT bandwidth (100/400/800 GbE).
  - PCIe Gen 4/5/6 throughput **out**-matching custom backplanes.
  - High-density computing and heterogeneous accelerators (GPUs, SmartNICs).

Network as the Interconnect

- Replacing custom backplanes with standard Ethernet switching.
- **Key enablers:**
  - Data center switches (shallow or deep buffer).
  - Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCEv2), Ultra Ethernet
  - Time-Sensitive Networking (TSN) for precision synchronization (e.g., White Rabbit / PTP).
- **Advantages:** Dynamic routing, load balancing, and immense aggregate bisection bandwidth without custom hardware.

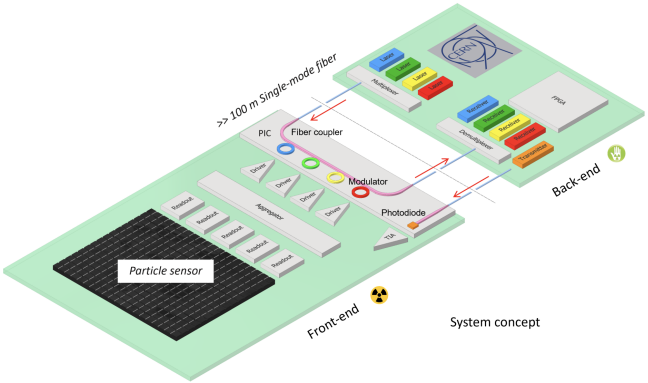
Precision Timing

- Even trigger-less readout-systems (e.g. LHCb) are **synchronous**
- High luminosity / occupancy can only be handled by using time in the reconstruction ("4D")
- LHC detectors require O(10) ps precision on the jitter and O(0.1) ns or better phase-stability and reproducibility
- Can be very challenging using FPGAs $\Rightarrow$: probably best to split clock-distribution and synchronous command distribution (White Rabbit)

The next logical step: replacing the front-end links

- **Key challenges**: comparatively complex problem, comparatively large minimal message (== frame size), protocol overhead, configuration
- **Two approaches:**
  - **Ultimate:** front-end ASICs or front-end FPGAs put data out on Ethernet (10, 25, 100 Gbit/s)
  - **Intermediate:** convert/translate front-end links to Ethernet
- Modern FPGAs can overcome a lot of the problems on the front-end, however FPGAs are always only part of the front-end and not suitable for high radiation environments (which is a problem in some areas of the LHC)

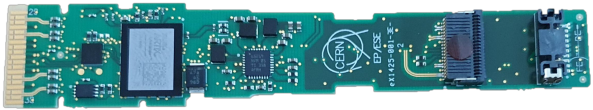
Low-mass / high bandwidth - Silicon Photonics (DART28)



- Developed by CERN EP/ESE
- 4 wave-lengths of 25 Gbit/s over a single fibre: coarse wave division multiplexing
- Photo-diode, driver, modulator etc... are radiation-hard, Laser is at the backend(!)
- **no copper** for FE-link
- Needs integration with front-end ASIC, high-speed (100 Gbit/s) $\Rightarrow$ ideal for high-granularity detectors, less convenient for large, low-occupancy technologies

Front-end to COTS: minimalistic approach

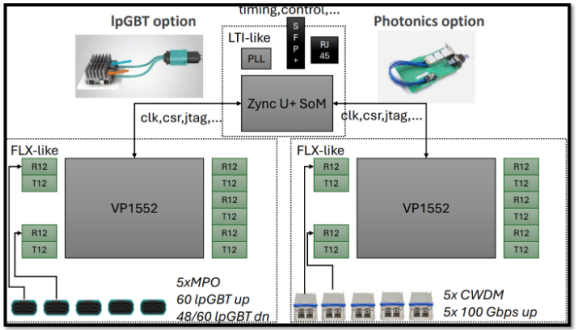
- **Bring Ethernet directly to the front-end**. Project in CERN EP/ESE by S. Baron and V. Stümpert
- Integrate all the conversion logic in a (small) FPGA in an SFP or QSFP transceiver module
- **Challenges:** space and power very constrained, **no additional processing of data**
- Can be plugged into standard NIC or Switch



Front-end to COTS: large-scale approach

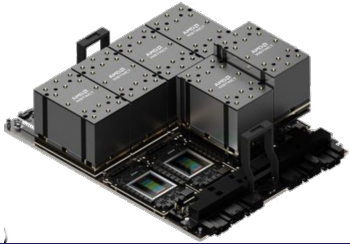
- **Addresses the following challenges:**
  - Aggregate a large number of slow custom links to a (reasonably) fast standard link
  - Facilitate in-line processing on FPGA
  - Fit into data-centre environment with standard form factor, powering, air-flow as an "appliance"
- **netFELIX** project will create such an appliance based on the experience on previous PCIe based plug-in modules.
  - Project supported by several important institutes in the DAQ community: CERN, BNL, NIKHEF, IN2P3, ...

the netFELIX project



- 1U 19-inch standard form factor, standard single-phase PSU, no back-plane
- 5/10/100 Gbit/s FE links, picosecond timing and clock distribution
- Requirements capture in progress, prototypes within 2 years
- GA 2030. Interest from major and smaller experiments: LHCb, ALICE, Aladdin, SHiP, ...

Formfactor changes in compute



- Higher power-density $\equiv$ lower cost
- Specifications for **1 MW**(!) racks are being discussed
- Requires liquid cooling
- . . . and novel form-factors
- higher speeds need cables / fibres etc...
- PCIe cards have uncertain future - contrary to PCIe as a protocol

Server Architecture for DAQ

- Standard 1U/2U rack servers now serve as readout and event-building nodes.
- **Critical hardware design considerations:**
  - **PCIe Topology:** Ensuring NICs, GPUs, and NVMe drives share root complexes efficiently to avoid CPU interconnect bottlenecks.
  - **NUMA Awareness:** Pinning memory buffers, interrupts, and processing threads to the correct CPU socket.
  - **Direct Cache Access (DDIO):** Allowing NICs to write directly to CPU cache, bypassing main memory latency and bandwidth limits.

Software-Defined Data Acquisition

- Transitioning hardware triggers to software-based event filtering.
- **Kernel Bypass Networking:**
  - Standard OS network stacks introduce too much jitter, context switching overhead, and copying.
  - Technologies like DPDK (Data Plane Development Kit) allow user-space applications to poll NICs directly.
  - Enables line-rate processing at $O(100)$ Gbps per core with microsecond latency variance.
- Architecture prioritizes lock-free queues, ring buffers, and thread-per-core models.

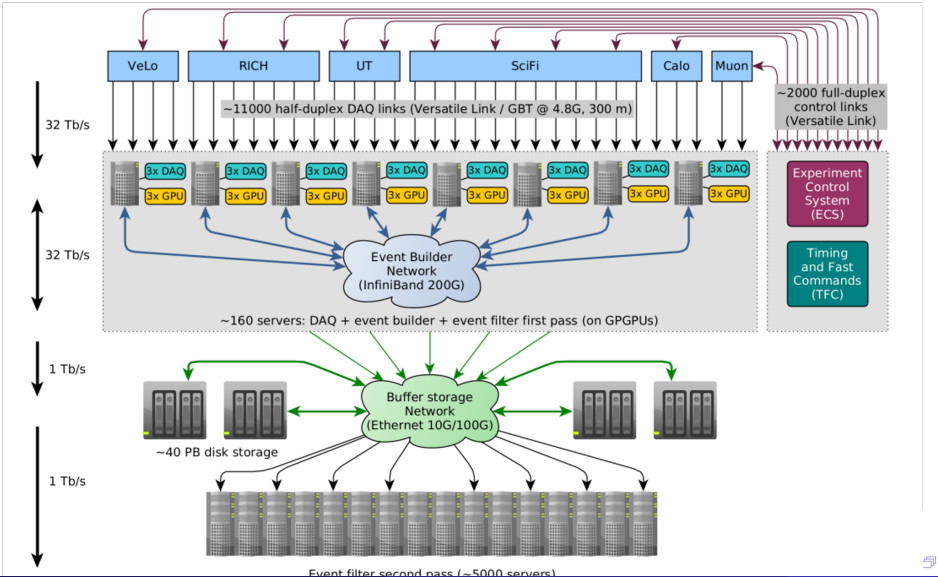
Inline Processing and Acceleration

- As CPU single-thread performance plateaus, heterogeneous computing is required to keep pace with I/O.
- **GPUs in the DAQ path:**
  - Massively parallel architectures are ideal for tracking, clustering, and pattern recognition.
  - *GPUDirect RDMA* enables NICs to push data directly into GPU memory, enabling zero-copy pipelines.
- **SmartNICs and FPGAs:**
  - Offloading packet parsing, decompression, and preliminary data reduction to the network interface itself.

Storage Tiering in the Data Center

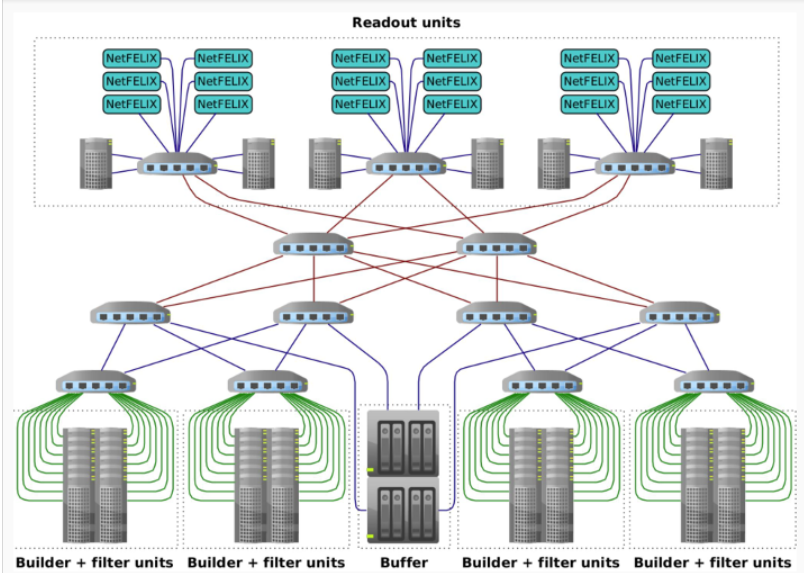
- DAQ systems must sink $O(\text{GB/s})$ to $O(\text{TB/s})$ of data continuously.
- Modern storage leverages standard modular tiers:
  1. **Burst Buffers:** Local NVMe PCIe drives on readout servers for immediate absorption of data spikes.
  2. **Distributed File Systems:** Systems like Ceph, DAOS, or Lustre providing a unified namespace over standard network fabrics (using NVMe-oF).
  3. **Cold Storage:** Automated tiering to high-density spinning disks or tape archives for long-term retention.

Tying it together: the LHCb DAQ for Run3 (today!)



- Data ingest: custom-made PCIe card with FPGA
- Network: InfiniBand HDR and Ethernet
- Nvidia GPUs, x86 servers, spinning drives

The future - Ethernet only



- All Ethernet
- Traffic shaping allows to avoid expensive deep-buffer switches
- Can accommodate any Ethernet attached compute and storage
- Only netFELIX modules needs to be close to the instrument, rest can be anywhere on the world or in the cloud

Challenges and Limitations of COTS and how to overcome them

- **Determinism:** IT gear is designed for high average throughput, not strict worst-case latency limits. Microbursts can cause switch buffer overflows → can be handled with traffic-shaping and improved hardware flow control in Ultra Ethernet
- **Lifecycle Management:** COTS hardware lifecycles (3-5 years) are much shorter than large-scale experiment lifecycles (10-20 years). Requires aggressive *rolling upgrade* strategies → the more COTS, the easier to seamlessly upgrade
- **Power and Cooling:** Standard enterprise rack density requires rigorous thermal management; data center power budgets become a significant fraction of experimental operational costs → need to work on this while we can hide behind accelerator and detector energy costs blush

What this means for SME (Small and Medium size Experiments)

- Small and Medium size experiments usually have limited resources to build new electronics / hardware → naturally want to reuse electronics developed for big experiments
- However don't have the specialists / teams for the firmware and backend electronics intended to go with these front-ends
- Ethernet based front-end or a standard gateway-appliance a la netFELIX can lead to a great simplification and democratization
- Finally we can get the experience of plugging together a couple of NIM and VME modules back to a real experiment smiley

What about AI?

What does the AI say about AI?

Google Gemini on the prompt "Give me some thoughts on data acquisition for large scientific instruments and artificial intelligence". Gemini's answer has been edited for brevity

1. AI Approach: Deploying Edge AI directly onto the detector's readout electronics.
   - Anomaly Detection: Instead of looking for specific, known particles, AI models (like autoencoders) can be trained to recognize "standard" events and flag anything that looks different as an anomaly.
   - Latency: The challenge is running these neural networks in nanoseconds. Tools like hls4ml (High Level Synthesis for Machine Learning) translate neural networks directly into FPGA firmware
2. "Smart" Data Compression
   - Intelligent Reduction: Rather than just truncating data, AI models can identify the "region of interest" in a detector (e.g., where a particle track actually exists) and preserve high-fidelity data only for those pixels, while discarding empty background noise.
   - On-Chip Processing: In theory FE-ASICs can perform this data reduction before the data is sent, reducing power consumption and bandwidth needs by orders of magnitude.

Reality check

- Anomaly detection is old stuff and not easy even offline
- Global or at least regional data are required for the inference $\Rightarrow$ extensive on-detector connectivity like on a classical trigger needed $\Rightarrow$ no reduction in complexity
- Reproducibility of trigger decisions?
- Noise-rejection works well already without AI
- Not clear if adding AI functionality saves more power than it costs
- For LHC at least everything needs to be "re-"designed with radiation in mind (granted, this makes for nice grant-requests and projects

Beyond the hype and the polemics - some real opportunities for AI

- Agentic AI to "fix" computer problems in large Online systems. Typical system administrator tasks can probably be automated - and complemented with a suggestion by the AI for physical interventions (until we have robots :-))
- Semi-agentic AI for controls of the experiment. Operating the detector is almost a sociological must in our community. However arguably the last 10% of lost efficiency in data-taking involve a large amount of human errors ($\Rightarrow$ **this should be measured!**. *"Semi"* means that we probably let the AI actions be confirmed by a human (accountability)
- Some more specialised applications are imaginable
  - finding optimal network schedules for event-building traffic
  - finding the root-cause of a problem out of unstructured text log-messages

Some Reflections

- The main idea of this talk - my main idea over the years - has been the commoditization of data acquisition and "trigger" / "real-time" / "online" processing
- **AI is mostly a data-center / cloud technology**, although there is some "AI on the edge ⇒ **favours commoditization and using industry standard** protocols, links
- AI technologies tend to need, power, cooling, latest and greatest process nodes ⇒ **disfavours implementation in hostile, power-, space-constrained environments**, implemented in old technologies as are the detector front-ends
- Opportunistic AI in available FPGA resources could make (financial) sense, anything else seems - to me - to be mostly a technical exercise with doubtful benefit
- Challenges resulting from limited determinacy, robustness and reproducibility of AI based processing should not be underestimated for trigger or control applications

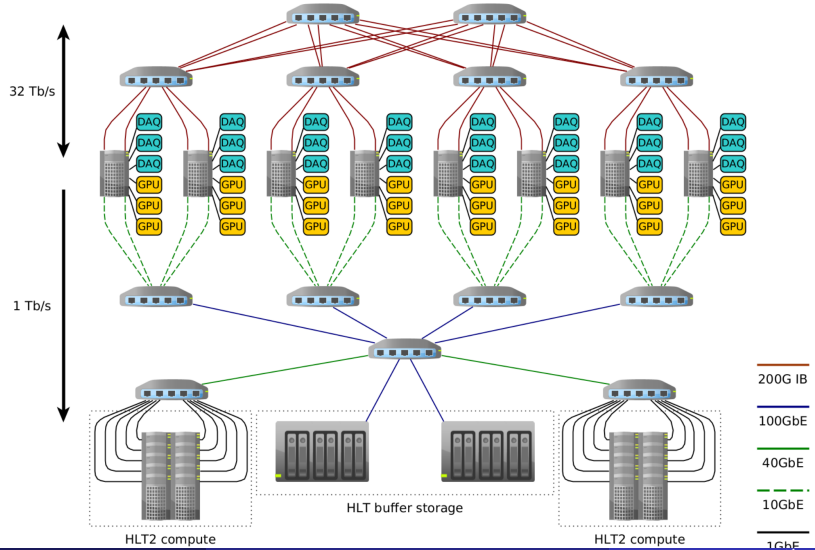
Conclusion

- The boundary between continuous Data Acquisition and generic High-Performance Computing (HPC) has largely dissolved.
- COTS-based DAQ provides unparalleled scaling capabilities and leverages hundreds of billions of dollars in industry R&D.
- The complexity shifts from custom hardware engineering to advanced software engineering (NUMA management, kernel bypass, lock-free parallel algorithms).

Questions?

Additional Material

Network-centric view on the LHCb Run3 DAQ



Gartner AI Hypecycle 2025

Figure 1: Hype Cycle for Artificial Intelligence 2025