DPDK SUMMIT
Powering the Future of Networking Software

12–13 May 2026
Stockholm, Sweden

DPDK Powered Data Acquisition Systems at CERN

Roland Sipos
Computing Engineer @ CERN

@DPDKSummit

Agenda

- Introduction
  - Data AcQuisition (DAQ) systems
- Use-cases
  - DPDK in action at various physics experiments
- Scaling and vision
  - Other interesting R&D activities
- Summary and outlook

European Organization for Nuclear Research

Main Goal: Understanding the Universe

CERN's core mission is to uncover the fundamental laws of nature using the world's most complex scientific instruments

The Accelerator Complex

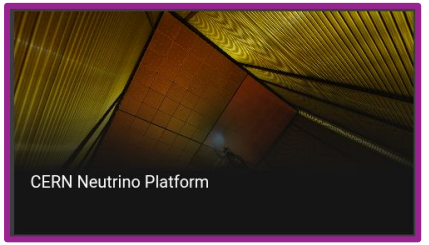
The Large Hadron Collider (LHC) is a 27km ring of superconducting magnets, accelerating particles to near-light speed for high-energy collisions

Major LHC Experiments

ATLAS, CMS, ALICE, and **LHCb** use sophisticated detectors to study the fundamental building blocks of our universe

Beyond the Collider

CERN has a diverse physics program, many experiments and R&D facilities (E.g.: **NA62, Neutrino Platform**)



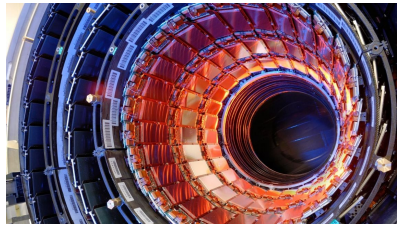
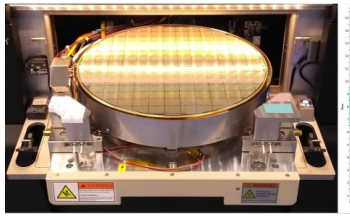
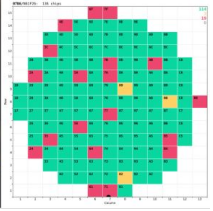
DPDK SUMMIT
Powering the Future of Networking Software

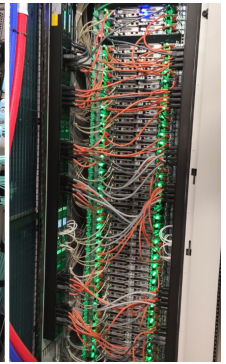
Data Acquisition & Selection

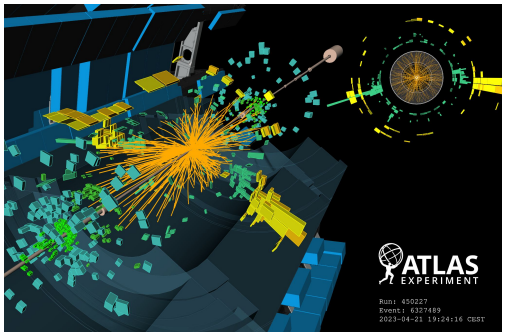
Data AcQuisition - DAQ

Getting data from millions of sensors and their channels to permanent storage safely and efficiently is the core mission of DAQ systems.

- **It's easy:** "Just" get the data out from electronics
- **Zero Loss:** Don't lose any on the way
- **Controlled Reduction:** Take into account that data usually gets reduced in a controlled way during transport
- **Distribution:** Route data to various processing elements and monitoring systems
- **Reliability:** All this 24/7, without error, no human intervention, and ideally fitting into stringent power, cooling and budget requirements





DPDK SUMMIT
Powering the Future of Networking Software

CERN

Trigger and DAQ system pipeline

Extreme Data Reduction

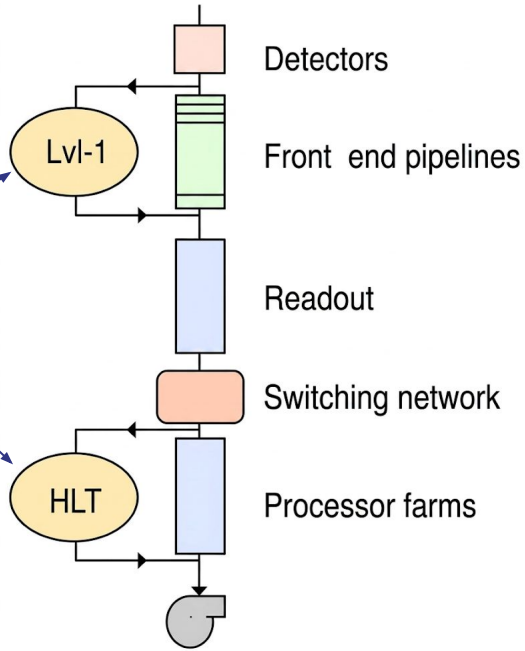
Trigger system(s) act as a filter, deciding in microseconds which collision events are worth keeping from millions of candidates per second.

Multi-Level Architecture

Typically consists of a hardware Level-1 (L1) for ultrafast filtering and a software High-Level Trigger (HLT) for detailed analysis.

Mission: Identifying Rare Physics

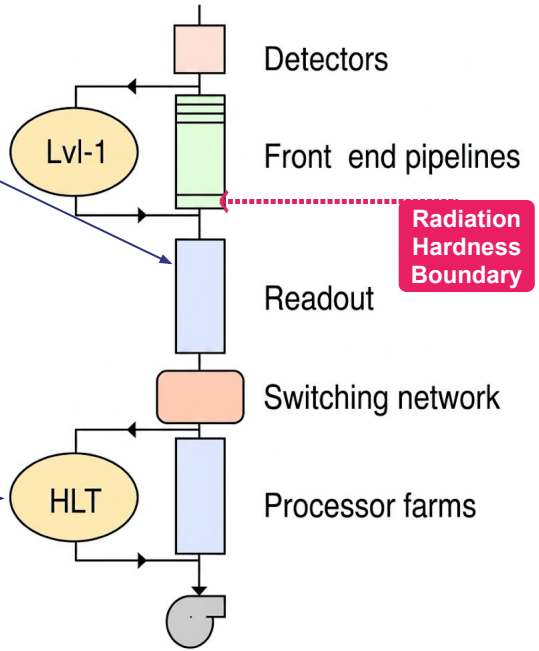
Its primary mission is to ensure rare and interesting physical phenomena are not lost amidst the overwhelming background "noise".



Trigger and DAQ system pipeline

1. DAQ is Post-Trigger Focused

Mainly cares about data after noise filtering (post L1 trigger).



2. Front-End vs. Back-End

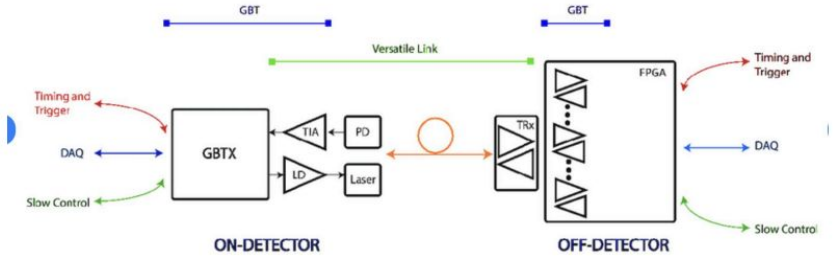
Anything before L1 is "front-end" or "detector". Readout consists of links and "back-end".

Radiation levels are usually very high close to the detectors. Front-end usually has strict radiation-hard requirements.

3. Event Building & HLT

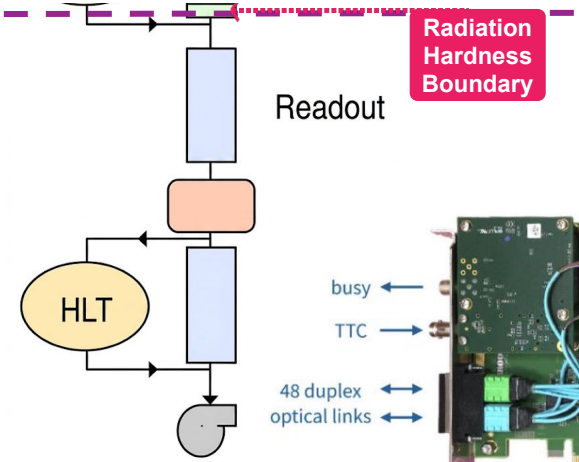
Everything between readout and processor farms is "event-building". Usually computer farms run the "High Level Trigger" (HLT).

Custom links and protocols



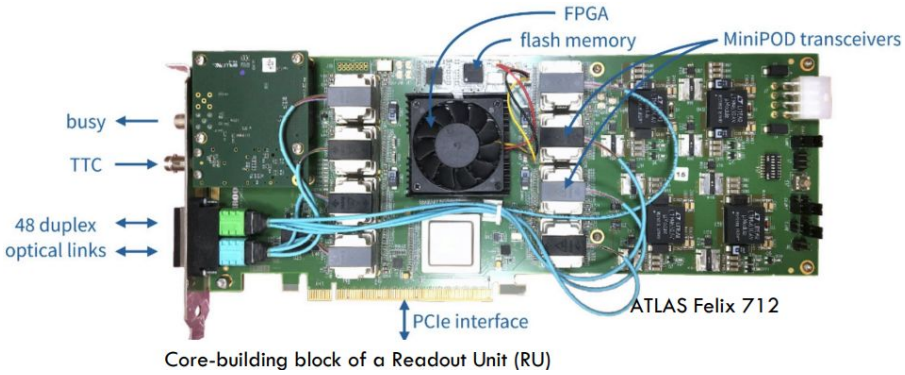
Need for radiation hard, custom versatile links and protocols

- With strong error correction (FEC)
- Transport fast timing signals for synchronization of electronics
- Operate efficiently with very small frame sizes (e.g.: 128 bits)
- Should be simple: no addresses, switching or aggregation: Maintaining strict logic capacities



Need for custom electronics to send and receive data

- Receivers merge and interface with industry-standard protocols, like PCIe
- Many “hybrid” solutions of different levels of aggregation, principles and strategies



Core-building block of a Readout Unit (RU)

DPDK SUMMIT
Powering the Future of Networking Software

CERN

DAQ @ CERN & DPDK use-cases

DAQ @ LHC Workshop

Evolutionary shift towards streaming data to COTS apparatus, reducing the reliance on hardware-heavy triggering.

- Blurring lines between ultra-fast and high-level data selection and triggering.
- Increased use of standard protocols and 3rd party toolkits in order to rely on industry and engineering standards.
  (LHC experiments use RDMA technologies since many years in their event builder infrastructures.)

DPDK in subsystems

The transition to commercial components and high-level software techniques created opportunities to use DPDK.

Key Adoption Areas:

- **Readout & Aggregation:** Managing high number of 1-10G detector links with Ethernet.
- **Event Building:** High-throughput networking for (hybrid) compute farms.
- **R&D areas:** SmartNIC evaluations, support for next generation detector interconnects, and more.

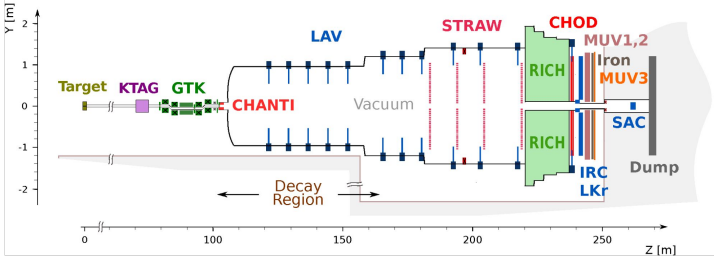
2016-2017	2018-2019	2020-2023	2024-2026	Future
Interest	**Prototyping**	**Development**	**Production & scaling**	**R&D**
Exploring DPDK at 10/40Gbps line rates with testing tools	Evaluating DPDK for UDP detector data stream readout	Demonstrators at NA62, and neutrino detector prototypes	Multi-100 Gbps UDP per node at CERN Neutrino Platform; DUNE baseline	LHCb benchmarks, future links, SmartNICs

DPDK SUMMIT
Powering the Future of Networking Software

CERN


Demonstrator to Production

NA62 Experiment



Physics & Detector

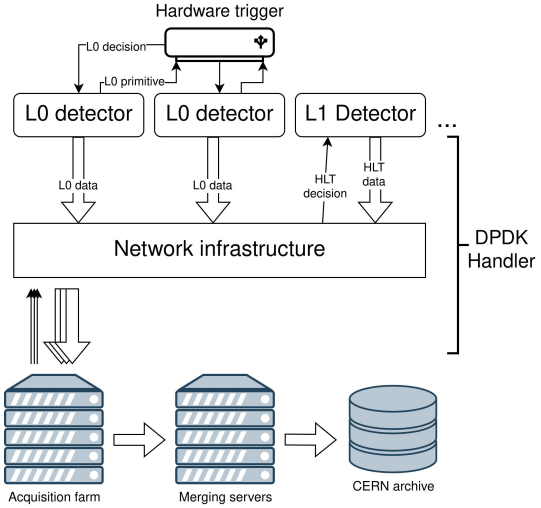
Measures ultra-rare charged Kaon decay in a 270m detector at CERN North Area. Uses SPS beam for precision tracking.

SPS Duty Cycle: ~6s on / ~12s off burst-based data taking.

DAQ Overview & Key Components

Handles 16 sub-detector data streams with real-time hardware triggering.

- Radiation-protected readout electronics.
- Acquisition farm: Collects event fragments & runs HLT software.
- Merger cluster: Manages file cataloging & transfer to persistent storage.
- Readout utilizes DPDK-based **Network Handler**.

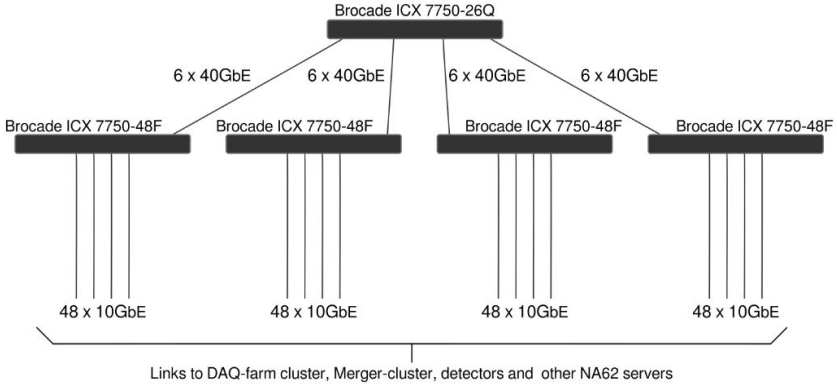


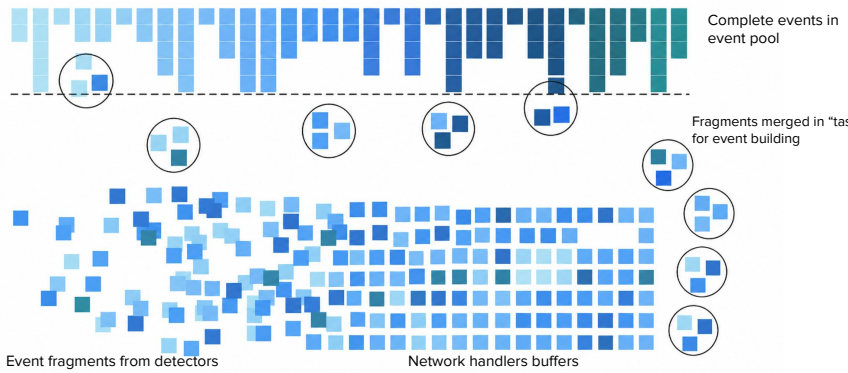
DPDK SUMMIT
Powering the Future of Networking Software

CERN

NA62 DAQ

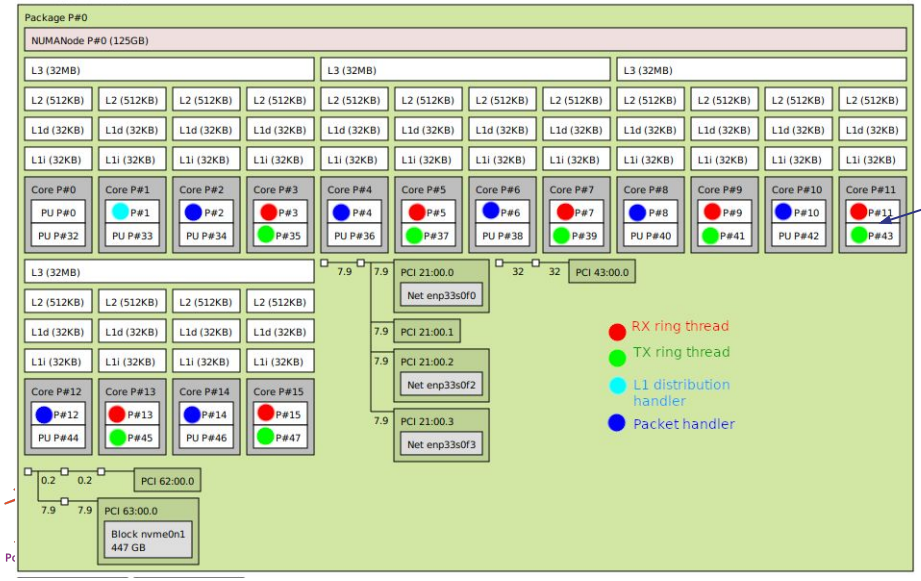
- Detector **Front-End Electronics (FEE) use FPGA-based UDP TX** at 4x1Gbps.
- Data is sent to the acquisition farm in a Round-Robin fashion:
  - 12 × 104 Pkt/s during bursts (~2Gbps total)
  - ~2M L0 events per burst
  - **~1.2M UDP packets per node per burst**
- Aggregator switches connect to core infrastructure:
  - The **core network (10-40Gbps) routes traffic between detectors and nodes**
  - Asymmetric, **bursty traffic pattern** (primarily FEE to DAQ).
- The Acquisition Farm (16 nodes) handles:
  - Event fragment collection (readout) from FEEs
  - High-Level Trigger (HLT) software filtering
  - Event building and forwarding to mergers for data file writing.
- Data requests from the acquisition farm follow HLT decisions with O(1s) latency.





NA62 Network Handler

- Acquisition farm originally commissioned in **2016** with proprietary zero-copy network handler. (Issues: licensing tied to NIC, rigid internal structure)
- Network handler reimplemented using **DPDK-22.11**:
  - Per node, 7 packet handlers with own TX-RX rings.
  - NUMA-aware resource allocation, and affinity control
  - Kernel isolation of critical threads
- In continuous operation since **2024**



- **L1 distribution handler:**
  It enqueues high-level trigger requests that packet handlers then send to the detectors.
- **Network handler:**
  This module manages the packet handlers, processes ARP requests, and provides a monitoring interface.
- **Packet handlers:**
  They use a single interface port to read received frames, send data, and manage the task queue.
- **Port abstraction:**
  Each port is linked to a physical NIC and features configurable rings for TX/RX operations.
- **Ring abstraction:**
  It manages memory buffers and coordinates low-level **send** and **receive** ring threads.

Data requests from the L1 distribution handler are distributed to packet handlers using a Round-Robin mechanism.

Packet handlers push received frames into local queues before enqueuing them into a shared MPMC task queue for event building on NUMA 1.

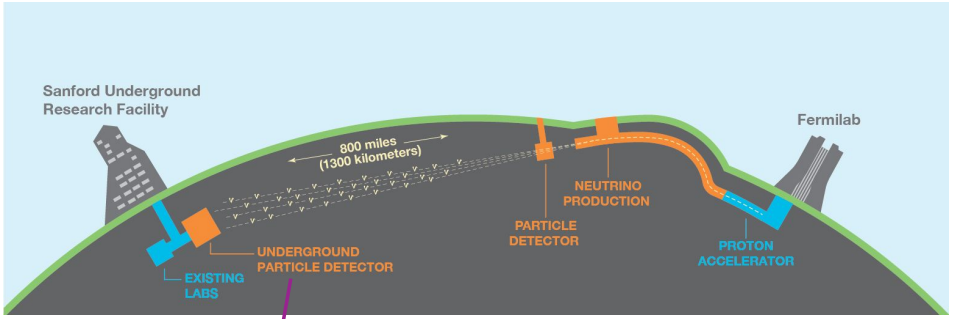
Completed events are held in an event pool until they are dispatched to the mergers.

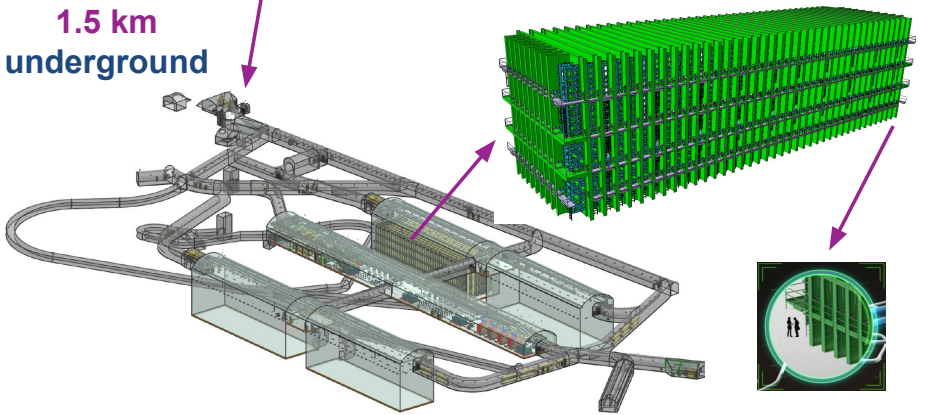
Use-case at Scale

Deep Underground Neutrino Experiment

DEEP UNDERGROUND NEUTRINO EXPERIMENT

- Future experiment in the US due to take data in late 2020s with a varied neutrino physics program
- DUNE "Far Detectors":
  - 2 super-modules with different detector technologies: Vertical and Horizontal Drift Modules
  - 17.000 ton LAr (87K / -186 °C)
  - Shielded, underground environment in a former gold mine
- Remote location: very strict power and cooling budget (~120kW underground)
- Many CERN contributions, including DAQ technical design and its readout & trigger systems.

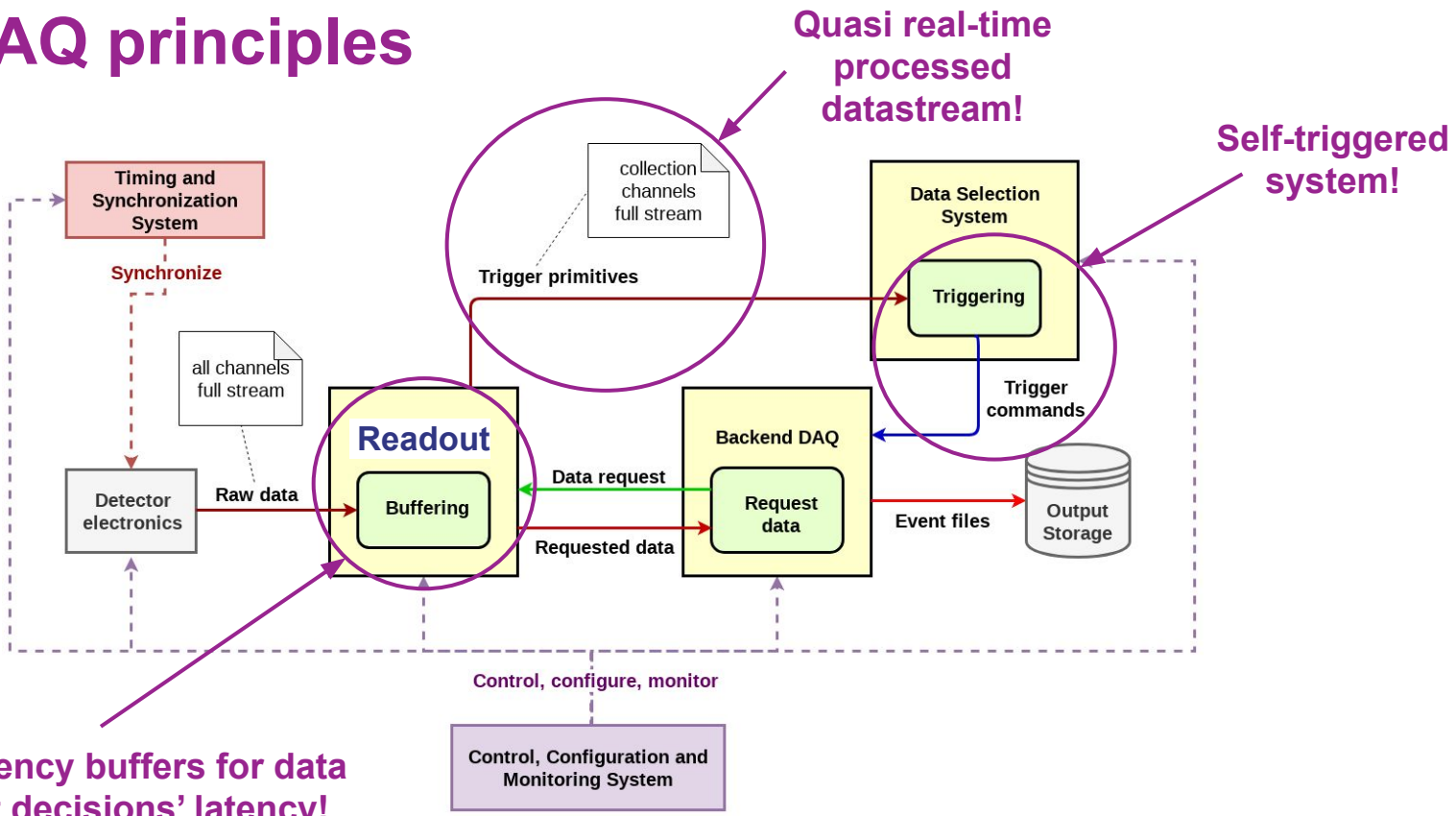




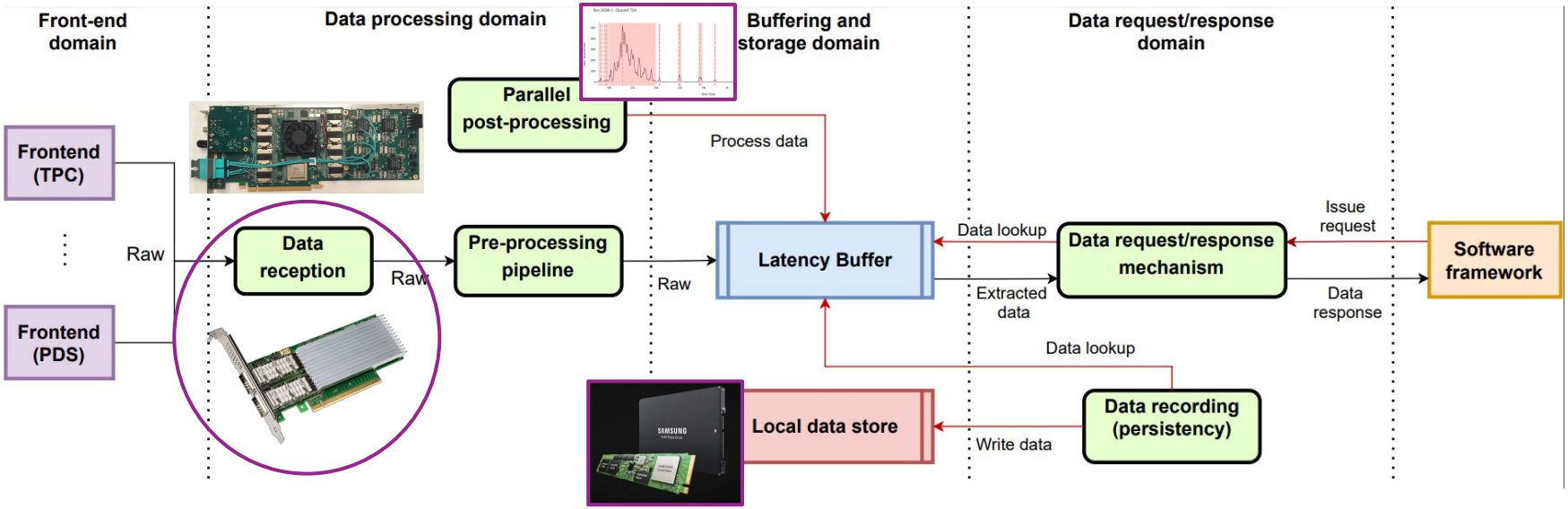
DPDK SUMMIT
Powering the Future of Networking Software

CERN

DUNE-DAQ principles



DUNE-DAQ Readout System



- Process every data frame for finding interesting activity, and produce information for the data selection subsystem
  - Most processing demanding component!
- Buffer data in DRAM for ~10 seconds for requests
- Persist data on NVMe up to 100 seconds

DPDK SUMMIT
Powering the Future of Networking Software

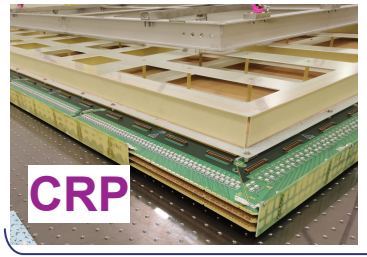
CERN

DUNE Front-End to DAQ

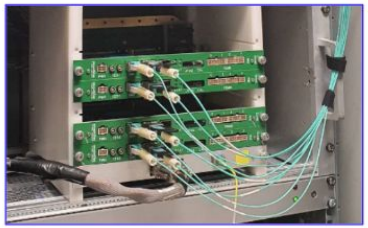
PAYLOAD CHARACTERISTICS OF DETECTOR ELEMENTS

Detector component for charge readout	Links and Data Streams	Payload size and arrival rate	Total throughput (incl. protocol headers)
Anode Plane Assembly (APA)	10 links 40 streams	7200 Bytes @ 30.5 kHz x 40 streams	~70.1 Gbit/s
Charge Readout Plane (CRP)	12 links 48 streams	7200 Bytes @ 30.5 kHz x 48 streams	~84.8 Gbit/s

Detector Elements



Front-End Electronics



- 5/6 WIBs per detector element (Warm Interface Board)
- Each WIB with 2 physical links
- Each link with 4 data streams
- **Simple UDP TX in FPGA**

Readout Network

Readout Units

Vertical Drift: x160 CRPs
Horizontal Drift: x150 APAs

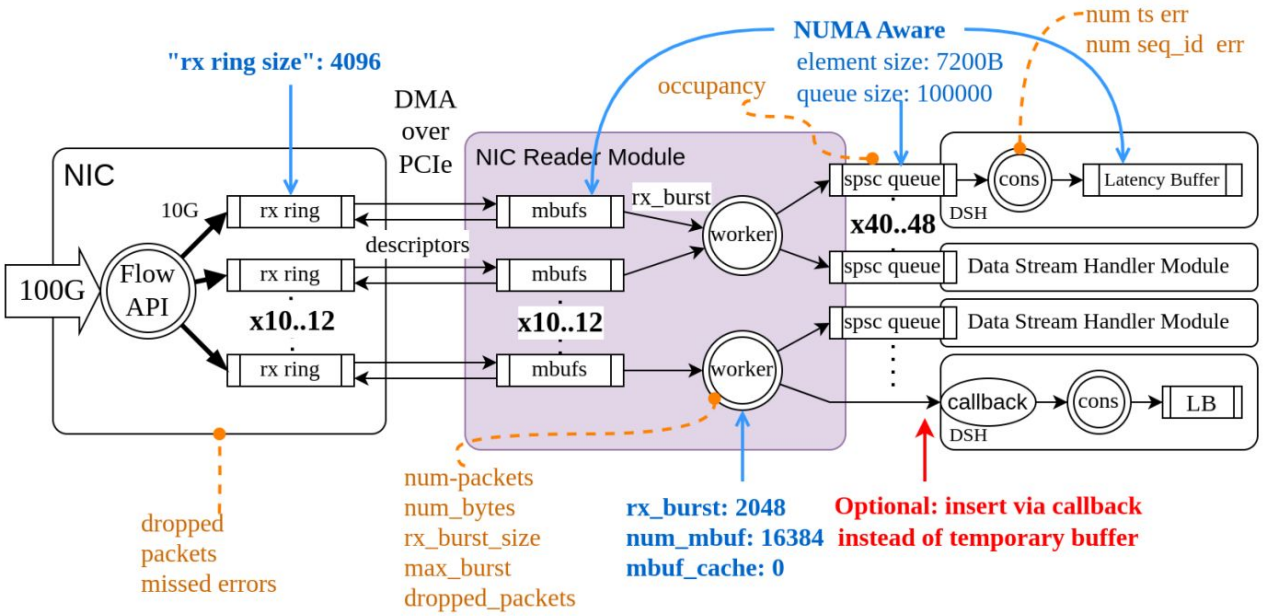
- Detector electronics transmits data over **10 Gbps links**
- Those are **aggregated into 100 Gbps uplinks** via switches
- 100Gbps links are fed to **Readout Units with COTS NICs**
- Total throughput: **~15 Tbps / module**

DPDK SUMMIT
Powering the Future of Networking Software

CERN

DUNE-DAQ Data Reception

DPDK based readout system diagram highlighting configuration parameters, monitoring metrics, and the packet processor (worker) function.



Algorithm 1 Packet processor function

```
iface ← confIfaceId                          ▷ Configured parameters
coreid ← confCpuCore
mbsize ← confMaxBurstSize
queues ← rxCoreMap[coreid]
mbufs                                        ▷ Assigned buffers available in scope
while !stopSignal.load() do
    for q : queues do                        ▷ Loop and RX burst queues
        qMbuf* ← mbufs[q.Id]
        nbRx ← rxBurst(iface, q.Id, qMbuf, mbize)
        if nbRx! = 0 then
            for buf : qMbuf do               ▷ Loop on burst results
                if isValidFrame(buf) then
                    payload ← getUdpPayload(buf)
                    handlePayload(payload)
                end if
            end for
        end if
        rxFreeBulk(qMbuf, nbRx)              ▷ Free processed
    end for
    if noFullBurst then                      ▷ Opportunistic sleep
        nanosleep(confSleepUs)
    end if
end while
```

Only a couple of workers are designated to fan-out and copy data to their destination deep buffers

Scalability and Optimization

Modularity: The detector enables splitting the readout into 150 units (~10 GB/s each)

Scale-Up: This allows a **single computer to read out multiple 100G uplinks**.

Optimization & Tuning

Critical Component Isolation:

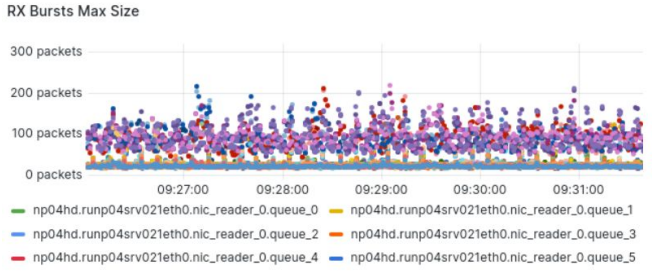
- CPU affinity masks for majority of threads (Hundreds of them!)
- Sensitive components are **kernel isolated**
- Minimizing socket cross-talk & allow single-copy to buffers
- Allocate resources based on device locality and even PCIe root complex

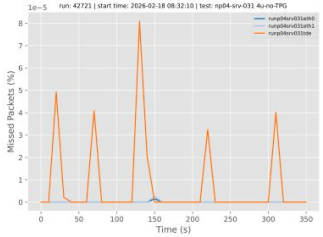
Vendor Agnosticity

Extensive Testing Since 2017:

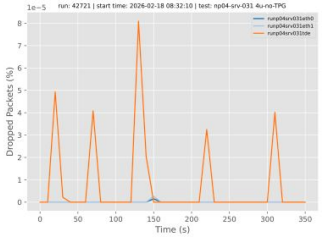
- **Intel & AMD families:** Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, Zen3-5
- **NIC Agnostic:** Evaluated with the 3 main vendors
- Attention to differences both for NICs and CPUs: cache topology & features like DCA vs. DMA

Baseline change to 400Gbps aggregation in scale-up evaluation

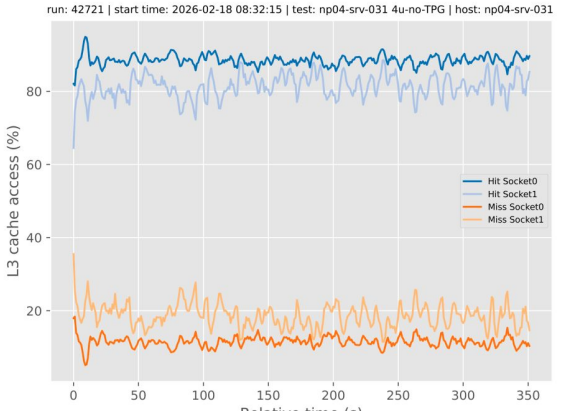




Percentage of packets missed by the receivers on the NIC.



Percentage of packets dropped by the worker threads in the data reception.



DPDK SUMMIT
Powering the Future of Networking Software

CERN

DAQ Operations @ 400Gbps

- **Prototype detectors at the Neutrino Platform facility are read out with single high-performance servers**
  4 x CRPs & 4 x APAs
  4 x 100 Gbps input per Readout Unit

- Motivation is the **power-draw reduction and price optimization** with this configuration with this topology

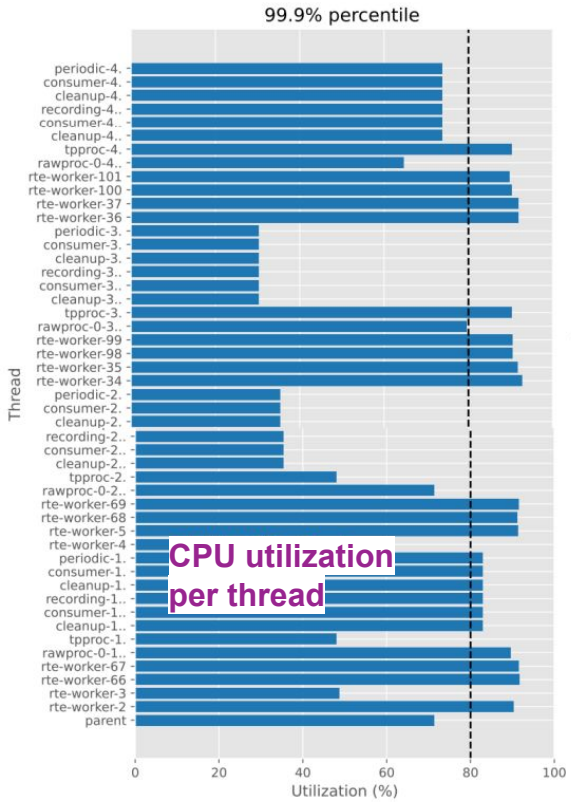
- **Load balancing and resource isolation** techniques are essential to reach deterministic performance on COTS hardware and software
  Balancing O(100) processing threads per 100G input

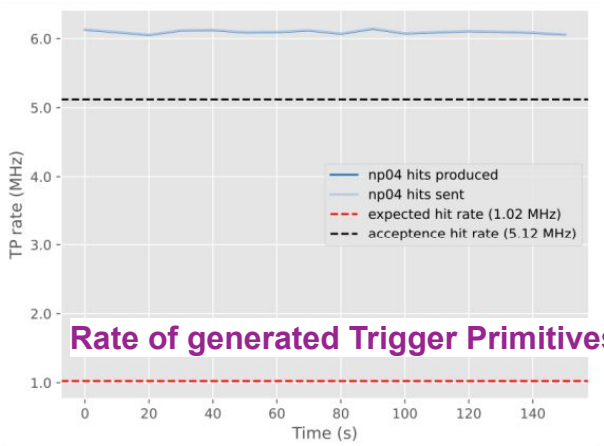
Intel and AMD Readout Units
(2S Sapphire Rapids and Turin platforms)

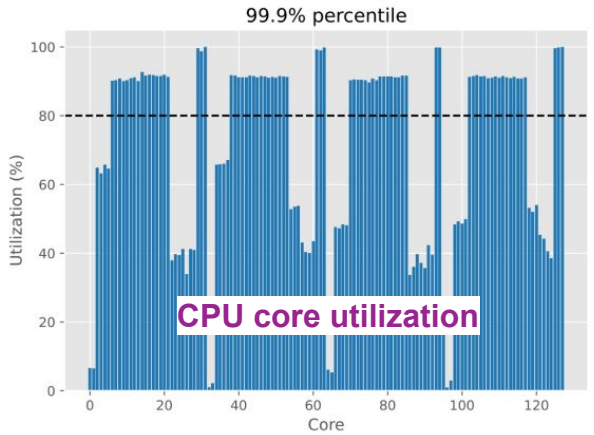
Established DAQ performance with real detector apparatus
Quasi DUNE FD conditions achieved in Q4 2024!

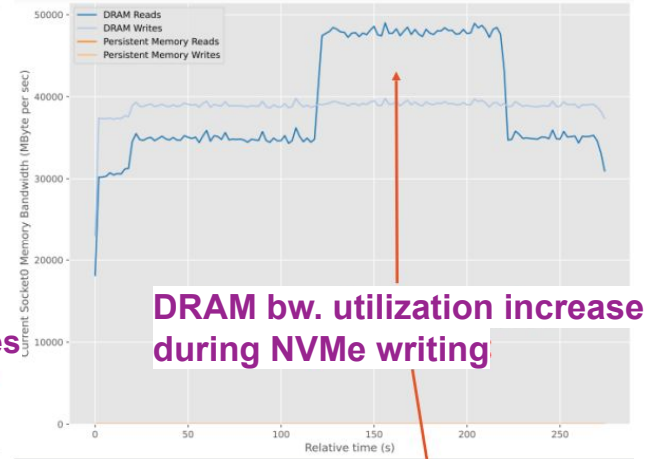
- **Data request rates maintained at downstream network limit**
- **Most complex feature extraction algorithm used**
- **SNB on NVMe recording criteria satisfied**
- **No packet loss with stable resource utilization**

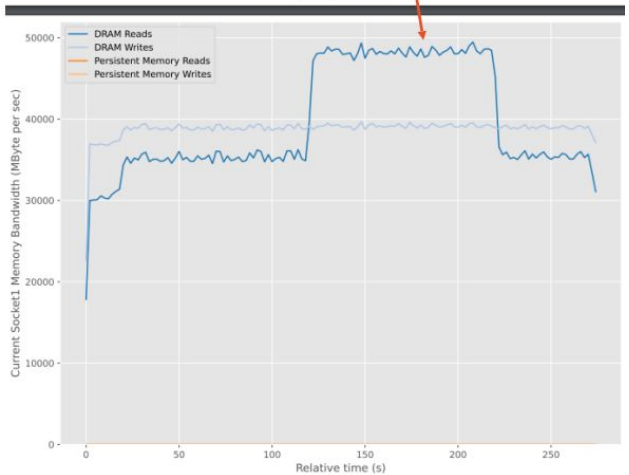
Performance











DPDK SUMMIT
Powering the Future of Networking Software

CERN

Metrics successfully meet KPI requirements!


Frontier & Beyond

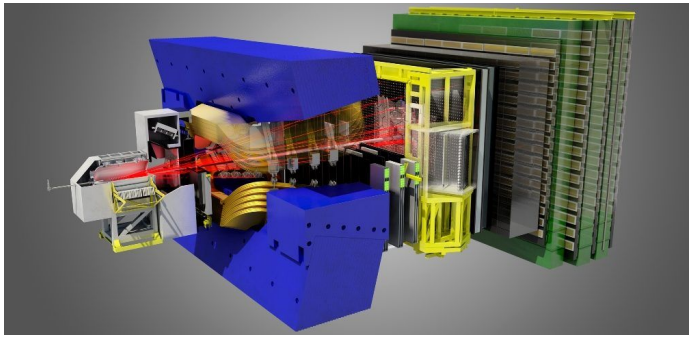
LHCb

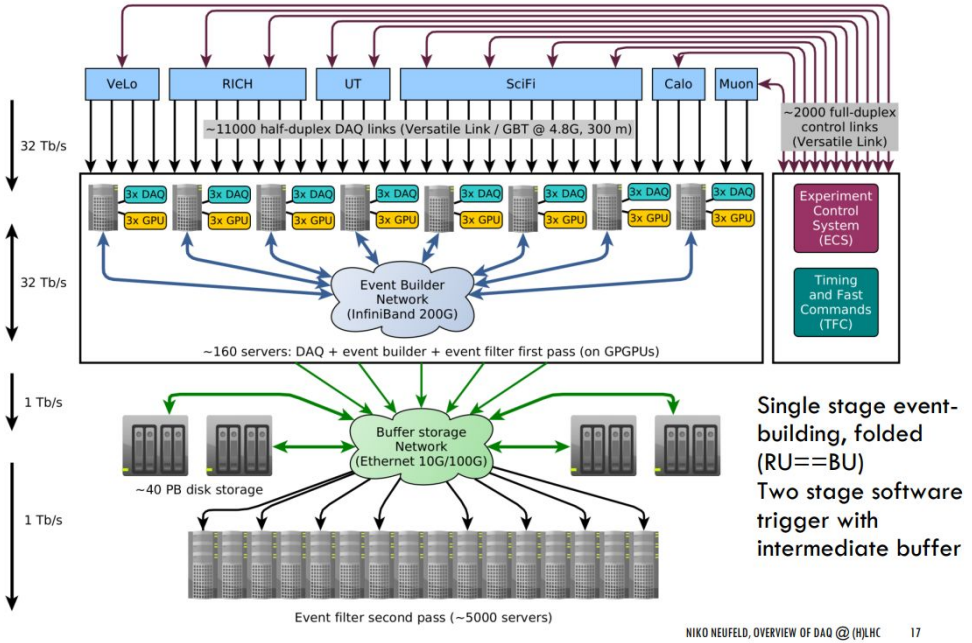
Physics & DAQ

- LHCb is a flavor-physics experiment studying heavy-flavor hadrons and rare decays
- The **triggerless DAQ** continuously reads out detector data at Tb/s scale using FPGA-based readout and a large GPU based event-builder farm
- Current DAQ infrastructure relies on PCIe and InfiniBand technologies to sustain **~40 Tb/s aggregate throughput**.

Future scaling challenges

- LHCb foresees up to a **5× increase in throughput requirements** over the next decade during the High-Luminosity LHC upgrade
- **Motivation to transition from proprietary fabrics** toward standardized Ethernet-based solutions
- DAQ R&D evaluated multiple Ethernet transport technologies for FPGA-based readout systems: UDP, RoCE, TCP





DPDK SUMMIT
Powering the Future of Networking Software

CERN

LHCb's DPDK benchmark

Motivation & Methodology

As UDP is simple and FPGA-friendly and vanilla Linux networking does not scale to 400G, a custom benchmark was developed for unidirectional line-rate traffic evaluation.

Traffic Generator

Configurable packet sizes, with 128-bit monotonic counters embedded in packets.

Packet Checker

Detects dropped and out-of-order packets from embedded counters. Computes real-time throughput from received payload.

Testbench

- Direct 400GbE link using NVIDIA ConnectX-8 NICs.
- TX: AMD EPYC 9334, RX: Intel Xeon Gold 6454S

Software Stack

- RHEL 9.5 (Kernel 5.14)
- DPDK v22.11.2, OFED v25.10, 1024 × 2 MB hugepages

Performance Results

- **398.82 Gb/s sustained throughput (near line rate).**
- 5 TB transferred successfully.
- Zero packet loss over 610 million packets.

Outcome

The UDP + DPDK testbed sustained near-line-rate 400 Gb/s using a single-thread software receiver. It reduces software/GPU overhead compared to more complex protocols. Became candidate technology for future LHCb DAQ upgrades.

DPDK SUMMIT
Powering the Future of Networking Software

CERN

DAQ & Electronics R&D

Addressing HL-LHC Scaling Demands

As LHCb transitions to the **HL-LHC era**, throughput requirements surge from **32 Tbps to 300 Tbps**, necessitating 30k optical links

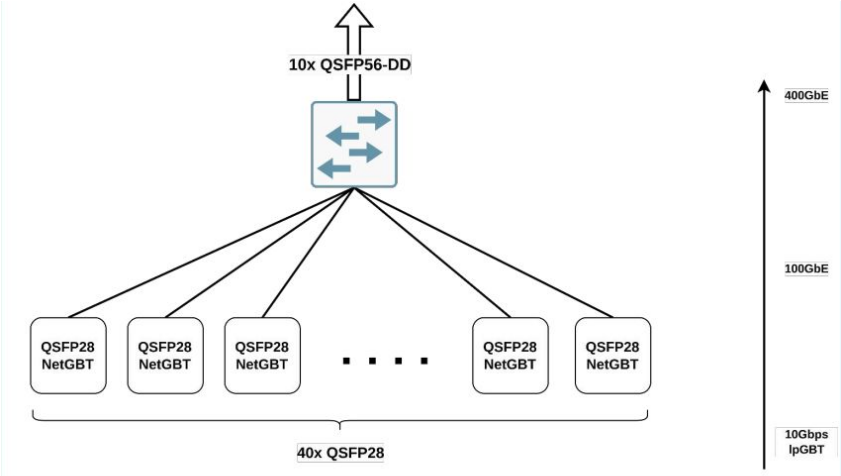
The radiation challenge is maintained and increases:

- **lpGBT Reliability:** Radiation-hard protocols are mandatory for front-end data transmission in high-radiation zones
- **Idea on Protocol Exchanges:** The **NetGBT** solution bridges these specialized links to standard **UDP/IP Ethernet**
- **Cost Efficiency:** Leveraging mid-end FPGAs for conversion reduces costs drastically compared to expensive PCIe-based systems

Proof-of-Concept

The Proof of Concept presented showed promising results:

- **lpGBT to UDP/IP:** Successful conversion with no back pressure
- **Efficiency:** Low resource utilization, leaving space for data processing offload
- **Flexibility & modularity:** Enabled by the standard Ethernet uplink. Links can be aggregated using COTS Ethernet switches



A first draft of a DAQ system using 100GbE-capable NetGBT.

DPDK SUMMIT
Powering the Future of Networking Software

CERN

Software Foundation for DAQ R&D

Standardizing the software solutions

DPDK serves already a high-performance backend layer for UDP reception for multiple experiments and test-beds.

This motivates to:

- **Unify best practices:** Develop or use libraries that encapsulate network resources to logical DAQ components
- **Maintain scalability and performance:** Support reception of multi-100Gbps aggregated inputs, without compromising other mission-critical data paths and processing pipelines
- **Ease of use:** Hide networking complexity and provide the ability of configure the overall topology and balancing at ease
- **Support Small and Medium Sized Experiments** (SMEs): Users with limited resources could benefit from experience

DPDK-DAQ

Development is already ongoing for a common software library inspired by the **DUNE** use-case:

- Exercised through **SmartNIC R&D** for FPGA offloading of some parts of the DUNE-DAQ data processing pipeline
- A customizable detector-emulation package is being integrated into the **DAQling** framework to deploy test-beds without the need of detector electronics

- **Extensible for various detector R&D projects (e.g.: SmartQSFPs), and upcoming experiments like SHiP.**

DPDK SUMMIT
Powering the Future of Networking Software

CERN

Summary and Outlook

Achievements

- DPDK used in operations by various experiments' DAQ
- Using DPDK with various commercial NICs and CPU families since 2017
- Scalable architecture developed supporting multi 100 Gbps inputs managed on single node

Outlook

- Expansion of a unified software stack for diverse R&D projects and upcoming experiments
- Addressing HL-LHC scaling demands (Multi 100 Tbps targets)

References

- Overview of DAQ at LHC
- The DAQ of NA62
- Ethernet readout of DUNE DAQ
- Evaluation of SmartNIC Devices for Use in Trigger and DAQ Systems
- A Low-Cost, Low-Power Media Converter Solution for Next-Generation Detector Readout Systems
- ESE: Options for future readout links

Thank you for your attention!

Your comments, questions, and suggestion are welcome!

Acknowledgement

With sincere thanks to colleagues and collaborators at CERN:

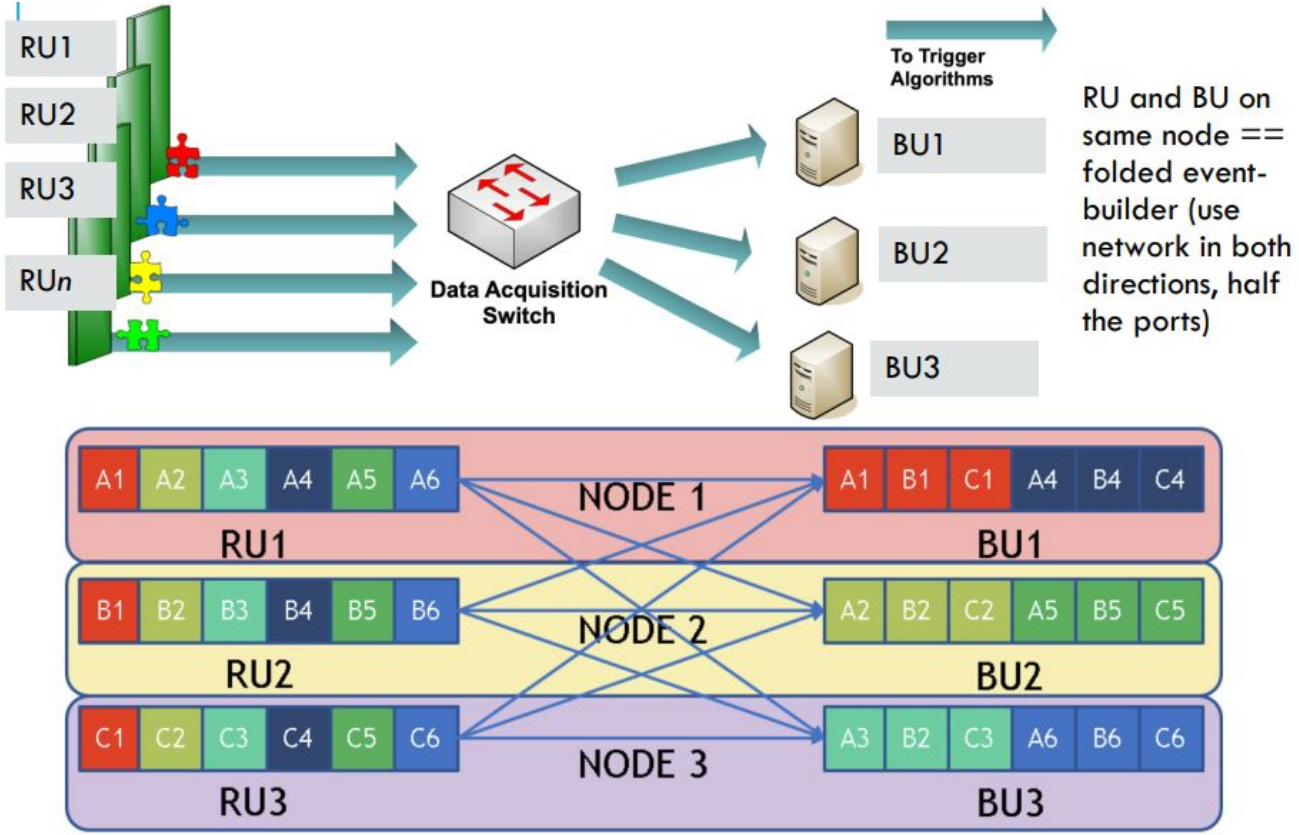
- LHCb: **Alberto Perro, Niko Neufeld, Tommaso Colombo**
- EP-DT: **Enrico Gamberini, Marco Ceoletta, Deniz Tuana Ergonul, Adam Cervenka**
  - Our Alumni: **Andreas Klavenes Berg, Marco Boretto**
- EP-NU: **Alessandro Thea**

With appreciation to our collaborators and institutes:

- CERN: DAQ community,
  - ESE (Electronics Group)
  - Neutrino Platform,
  - IT Department
- DUNE Collaboration, especially:
  - UK DUNE DAQ Project,
  - Rutherford Appleton Laboratory (RAL),
  - University of Oxford
  - University of Toronto

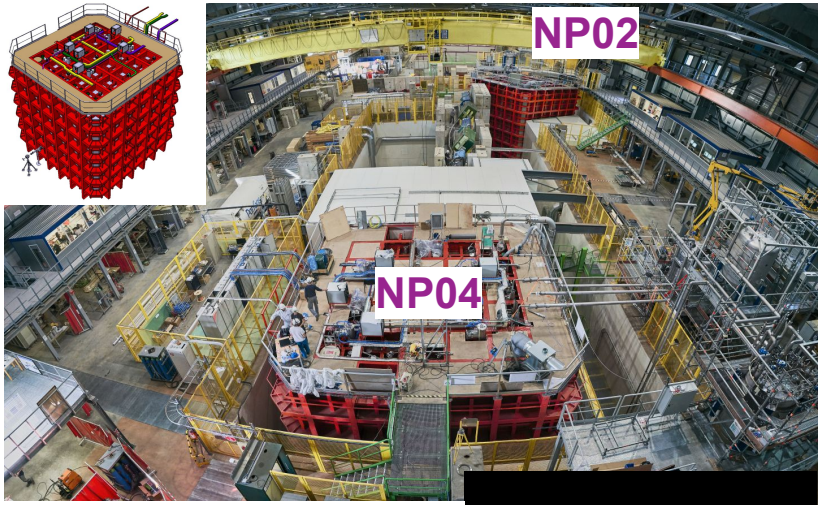

Backup Slides

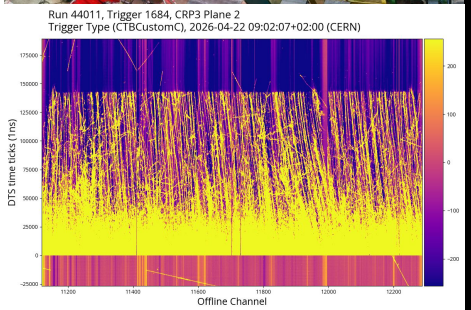
Event building / matrix-transpose with network



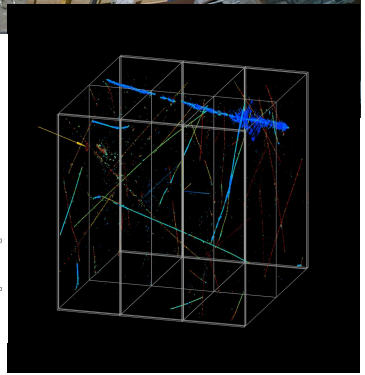
CERN Neutrino PLATFORM & DUNE prototypes

- NP02 - Vertical Drift (VD), and NP04 - Horizontal Drift (HD) DUNE prototype detectors
  - Demonstrate the design, construction and operation of the detector
  - Common DAQ system
- Charged particle beam from SPS
  - Opportunities and challenges to demonstrate triggering based on multiple signals, while being exposed to high cosmics background.
- Common DAQ system
  - Integration point of DPDK based Ethernet readout system, which is in continuous operation since 2023





Extended cosmic shower



Beam event

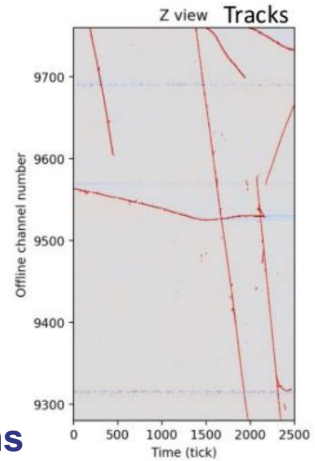
DPDK SUMMIT
Powering the Future of Networking Software

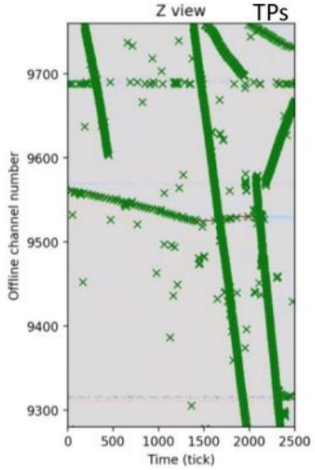
CERN

Processing - feature extraction

Key readout requirement: Real-time processing and streaming of interesting data regions for trigger decision.

- Processing task for each data streams:
  - Extracts ADC waveforms of channels
  - Does pedestal, filtering, hit-finding (ADC above threshold)
- Algorithms are implemented using **AVX2 registers and instructions** at different levels of complexity and resource needs
  - The most CPU intensive component of the readout



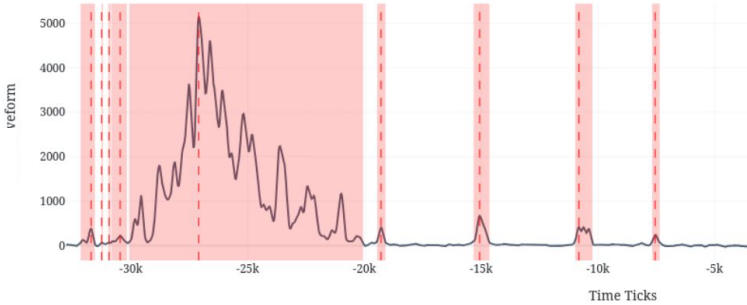


OVERVIEW OF COMPONENTS AND THEIR RESOURCE NEEDS FOR 2 CRPS

Component	Number of threads	CPU cores assigned	Maximum CPU core utilization (%)
Data reception (Packet processors)	8	4 phys. and 4 HT [a]	~48.2
Data processing (TPG)	96	10 phys. and 10 HT	~55.8

Simplest algo. case!

For ~3000 channels per APA/CRP:



DPDK SUMMIT
Powering the Future of Networking Software

CERN

UDP TX firmware block

DUNE DAQ provides a firmware block developed at the Rutherford Appleton Laboratory (RAL) Technical Division that may be integrated into the front-end electronics FPGA boards.

This transmitter block is responsible for sending Ethernet frames following UDP where the carried payloads are the front-end electronics data frames.

Every data frame also carries a unified and versioned DAQ header that contains geographic and physical location information about the source of the data stream. It also contains the timestamp from the timing system of the detector, and a sequence identifier for data integrity and continuity checks.

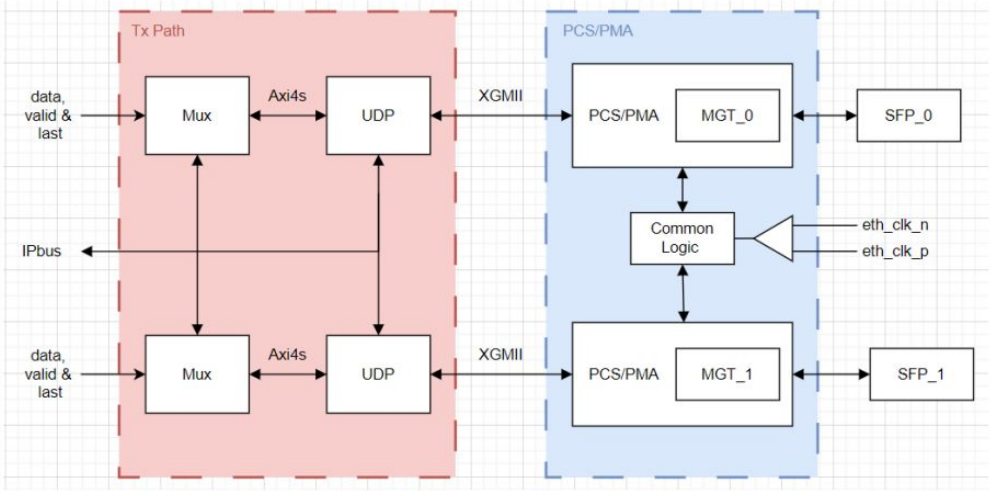


Fig. 2. Architecture of the transmitter block provided for the front-end electronics. The PCS/PMA area contains modified Xilinx IP[6] components, and the Tx Path is a custom firmware block developed by engineers at RAL Technical Division. The overall block is responsible for equipping the detector data frames with IPv4 and UDP headers following the communication protocol.

R&D on Readout Links for DAQ

ESE Strategy for Future Readout

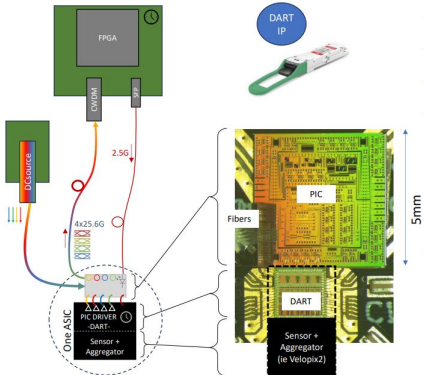
The ESE group targets **Long Shutdown 4 (LS4)** and beyond with highly scalable link options:

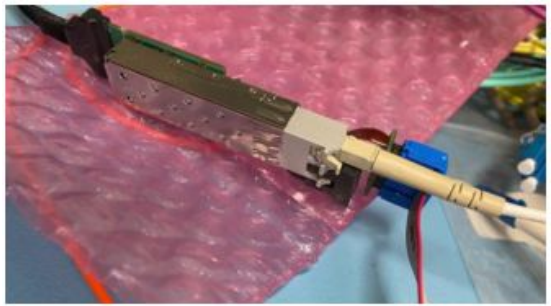
- **Versatile Link+ Evolution:** Core option offering **40 Gb/s upstream** and 2.5 Gb/s downstream, building on proven lpGBT tech.
- **Silicon Photonics (SiPh):** Early R&D for extreme radiation hardness, aiming for **100 Gb/s** per link.
- **DART IP:** 4x25G demonstrator currently targeting the **LHCb Velo upgrade** as a candidate technology.

The "No-Backend" Interconnect

The shift toward industry-standard protocols bridges custom hardware and DAQ:

- **Translated Links:** Using "Smart" modules to convert custom protocols directly into **standard Ethernet frames**.
- **Direct Aggregation:** Eliminates complex backends by connecting detector links directly to **COTS Ethernet switches**.
- **Simplified DAQ:** Standardizing on UDP/IP allows for uniform, software-driven data processing pipelines.





DPDK SUMMIT
Powering the Future of Networking Software

CERN


DPDK
SUMMIT
Powering the Future of Networking Software