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
  - CERN and 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

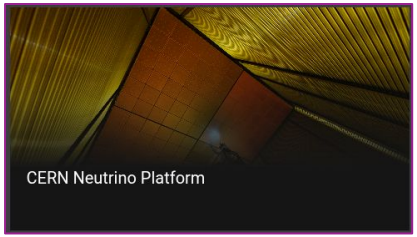
The **Large Hadron Collider (LHC)** is a 27km ring of superconducting magnets, accelerating particles to near-light speed for high-energy collisions.

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**)



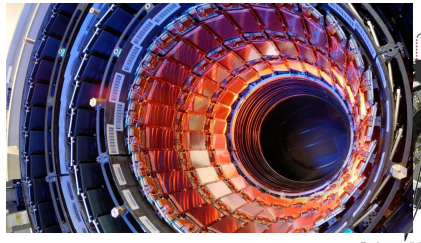
DPDK SUMMIT
Powering the Future of Networking Software

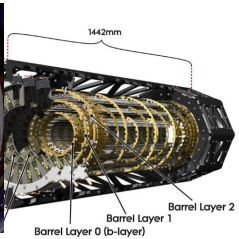
Data Acquisition & Selection

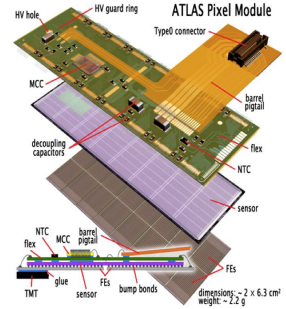
Data AcQuisition - DAQ

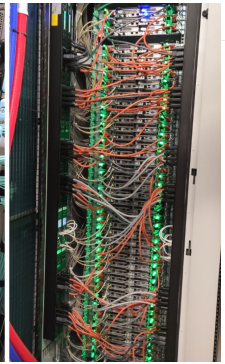
Getting data from millions of sensors and their channels to permanent storage safely and efficiently is the core mission of Data Acquisition Systems.

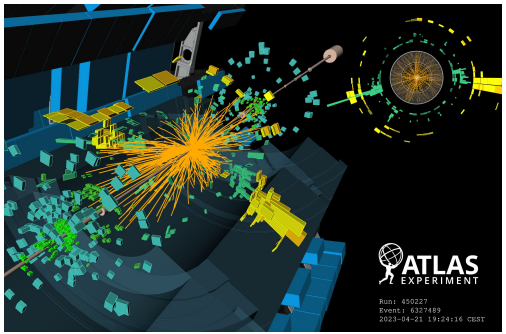
- **DAQ is easy:** Just get the data out from electronics.
- **Zero Loss:** Don't lose any on the way.
- **Controlled Reduction:** Take into account that data usually gets reduced in a controlled way during transport.
- **Distribution:** Route data to various processing elements and monitoring systems.
- **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

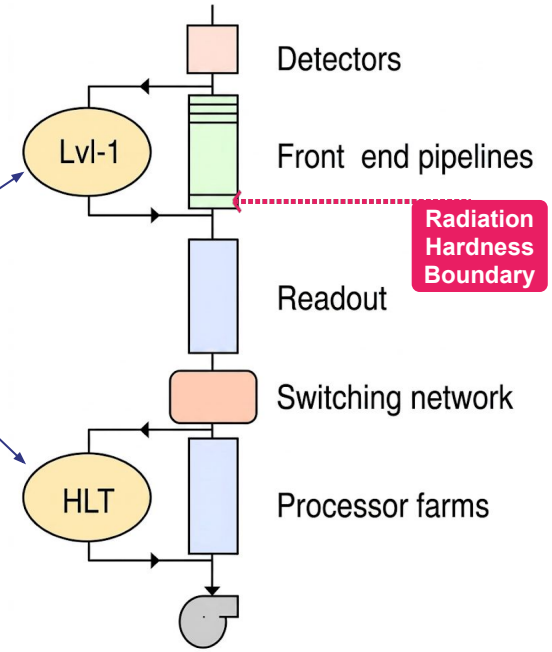
Trigger system(s) act as a filter, deciding in microseconds which collision events are worth keeping from millions of candidates per second.

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 (like Higgs decay) are not lost amidst the overwhelming background "noise".



DAQ is Post-Trigger Focused

We normally care about data after the noise filtered data (post L1 trigger).

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, which is particularly true for the LHC experiments.

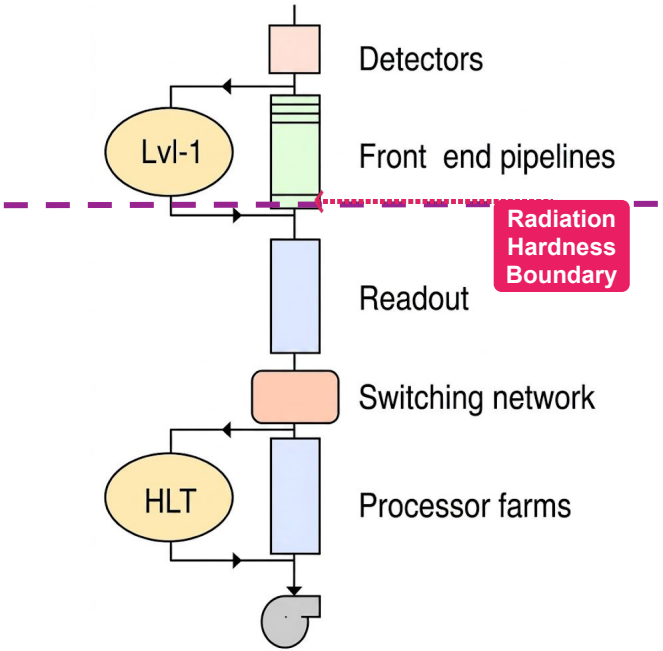
Event Building & HLT

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

Custom hardware vs. standards

Need for radiation hard and custom versatile links and protocols

- With strong error correction (FEC)
- Should transport fast synchronous timing signals for synchronization
- Operate efficiently with a very small frame size (e.g.: 128 bits)
- Should be simple: no addresses, switching or aggregation



Need for custom electronics to send and receive data

- Usually receivers merge and interface with industry-standard protocols, like PCIe
- Many “hybrid” solutions of different levels of aggregation, principles and strategies
- Meanwhile maintaining strict programming logic capacities and constraints in hardware

DAQ @ CERN & DPDK use-cases

DAQ @ LHC Workshop

Technological evolution is shifting DAQ towards streaming data to COTS compute farms, reducing the reliance on hardware-heavy triggering.

- Blurring lines between the ultra-fast and high-level stages in the DAQ pipeline.
- 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 creates prime opportunities for 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

Interest

Exploring DPDK at 10/40Gbps line rates with testing tools

2018-2019

Prototyping

Evaluating DPDK for UDP detector data stream readout

2020-2023

Development

Demonstrators at NA62, and neutrino detector prototypes

2024-2026

Production & scaling

Multi-100 Gbps UDP per node in production at CERN Neutrino Platform; DUNE baseline

Future

R&D

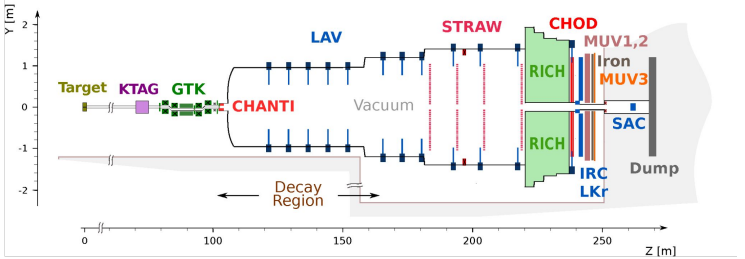
LHCb studies for Event Building, SmartNICs & SmartSFPs (Detector Systems R&D)

DPDK SUMMIT
Powering the Future of Networking Software

CERN


Demonstrator to Production

NA62 Experiment



Physics & Detector

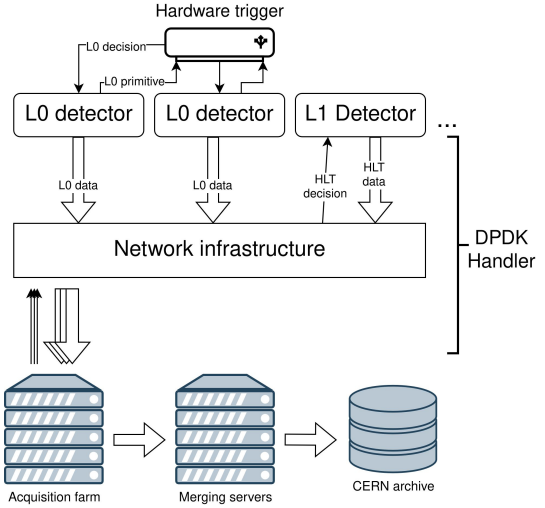
Measures ultra-rare charged Kaon decay in a 270m detector at CERN North Area. Uses SPS beam for precision tracking.

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**.

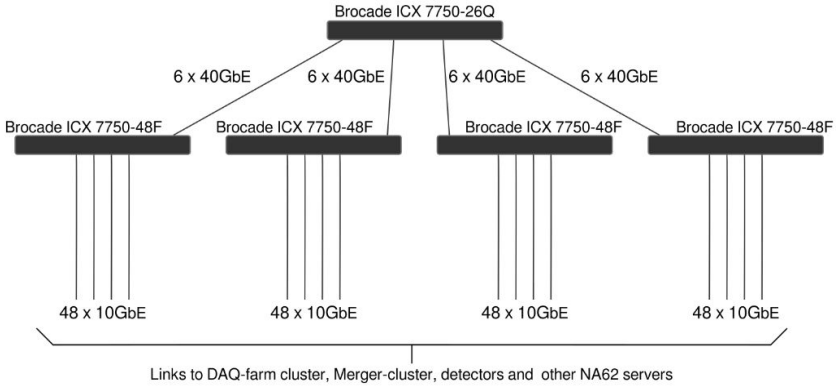


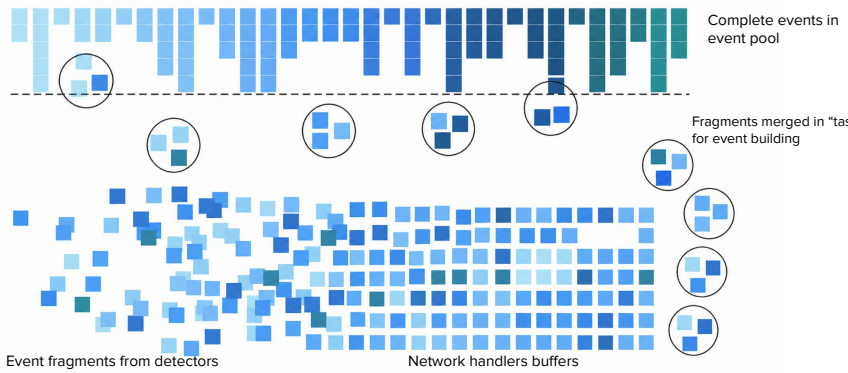
DPDK SUMMIT
Powering the Future of Networking Software

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 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.



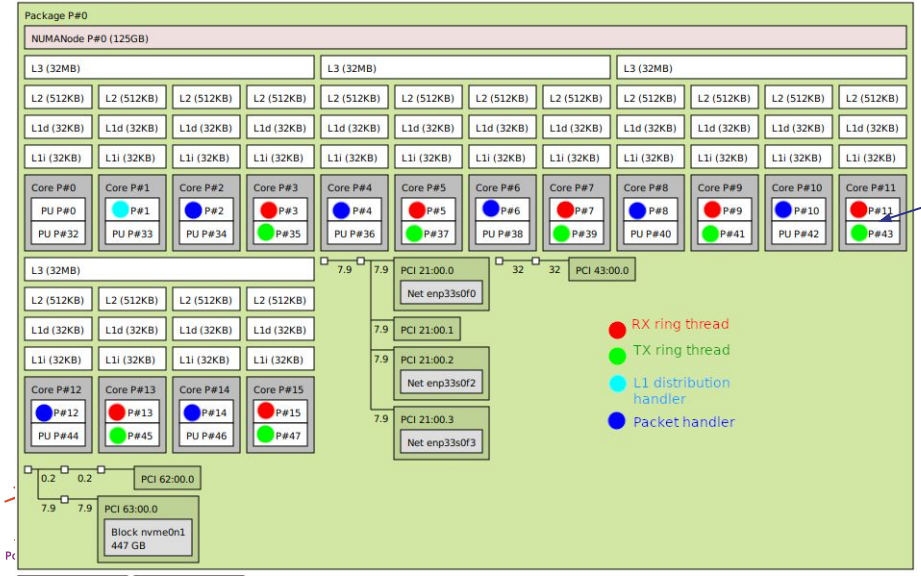


DPDK SUMMIT
Powering the Future of Networking Software

CERN

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**:
  - In use: 7 packet handlers with own TX-RX rings.
  - NUMA-aware thread spawning, all on NUMA 0.
  - 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.

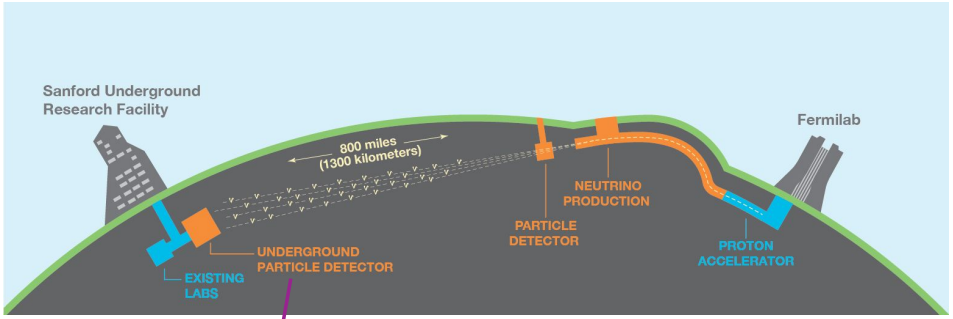
Completed events are held in an event pool until they are dispatched to the mergers.

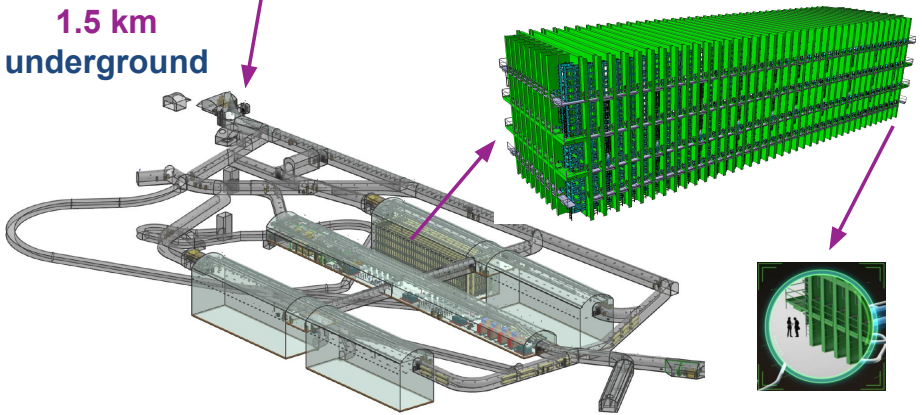

Use-case at Scale

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 common DAQ technical design and its readout & trigger systems.





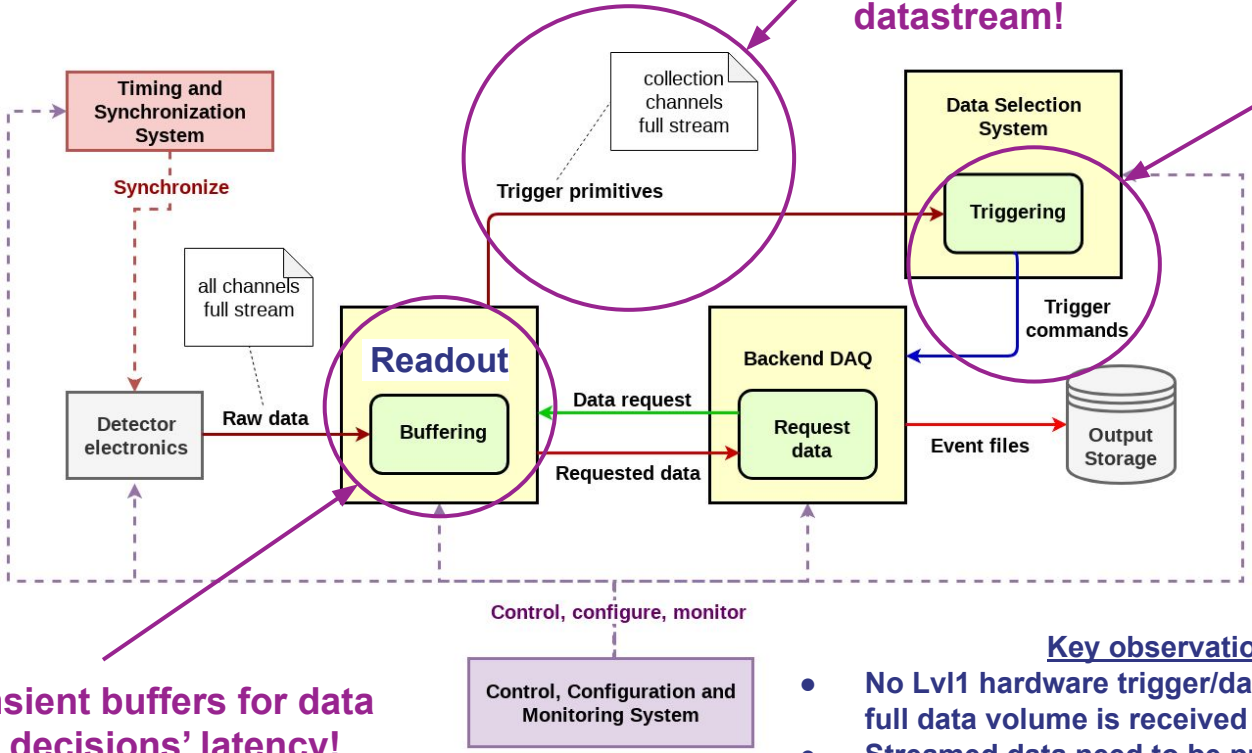
DPDK SUMMIT
Powering the Future of Networking Software

CERN

DUNE-DAQ principles

Quasi real-time processed datastream!

Self-triggered system!



Deep transient buffers for data request decisions' latency!

Key observations

- No Lvl1 hardware trigger/data reduction: full data volume is received from electronics.
- Streamed data need to be processed for constructing decisions about what data to keep.
- Request data chunks from several seconds deep "latency" buffers.

DPDK SUMMIT
Powering the Future of Networking Software

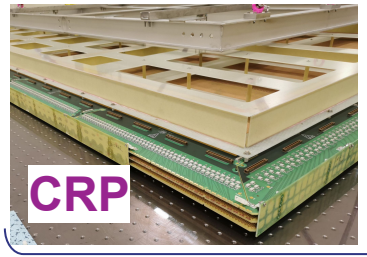
CERN

DUNE Front-End to DAQ

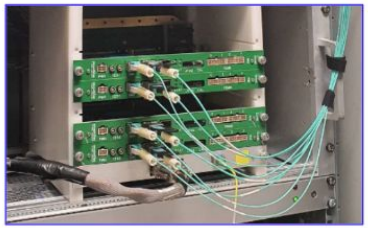
PAYLOAD CHARACTERISTICS OF DETECTOR ELEMENTS

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

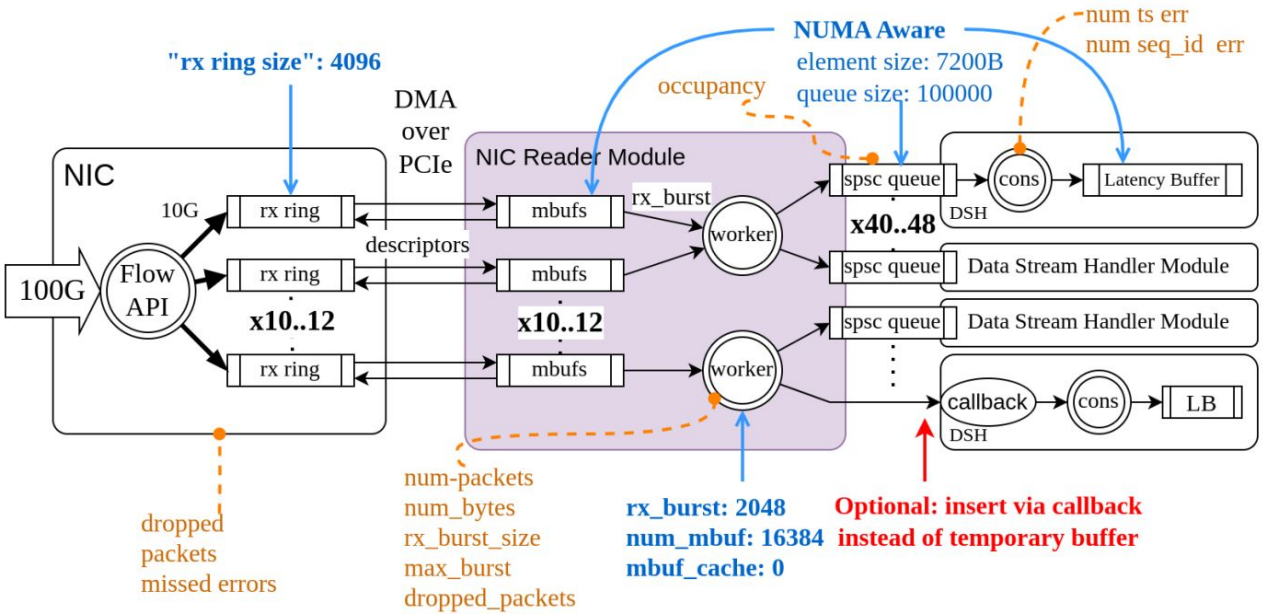
- Detector electronics transmits data over **10 Gbps links**
- Those are **aggregated into 100 Gbps uplinks** via switches
- 100Gbps links are fed to **Readout Units with COTS NICs**
- Total throughput: **~15 Tbps / module**

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

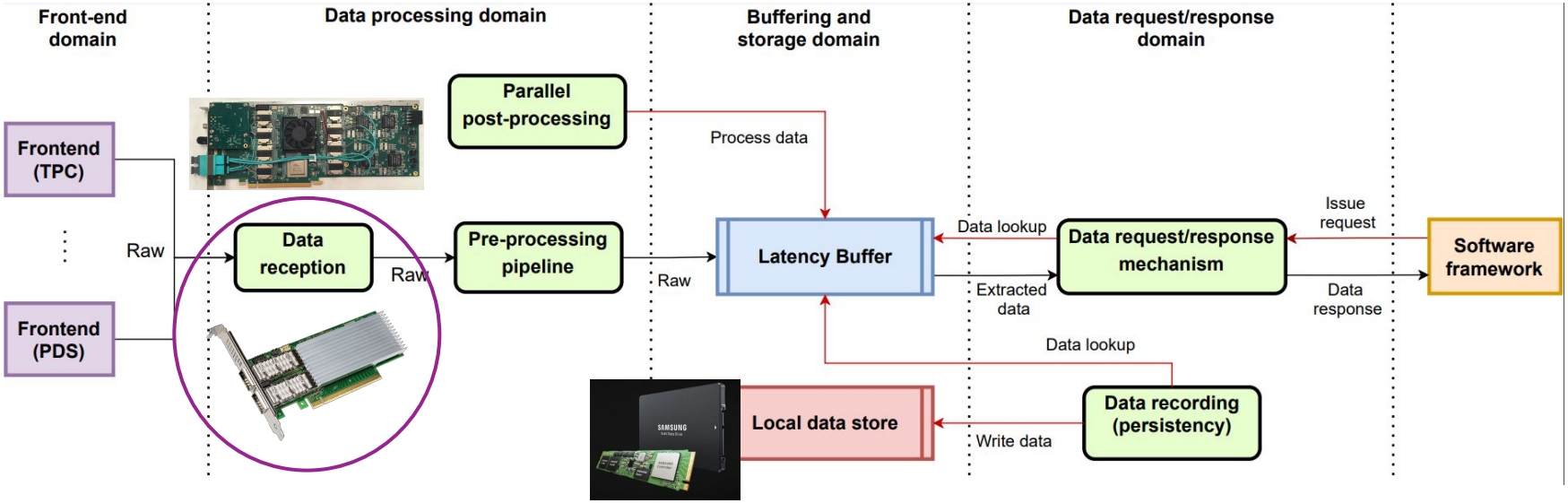
```
iface ← confIfaceId                          ▷ Configured parameters
coreid ← confCpuCore
mbsize ← confMaxBurstSize
queues ← rxCoreMap[coreid]
mbufs                                        ▷ Assigned buffers available in scope
while !stopSignal.load() do
    for q : queues do                        ▷ 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
```

DPDK SUMMIT
Powering the Future of Networking Software

CERN

So far it's quite simple, but …

DUNE-DAQ Readout System



Meanwhile the talk focuses on UDP data reception, the readout system does much more:

- Process every data frame for finding interesting activity, and produce information for the data selection subsystem
- 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

Scalability and Optimization

Modular Design: The apparatus enables splitting the readout into 150 units (~10 GB/s each).

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

Optimization & Tuning

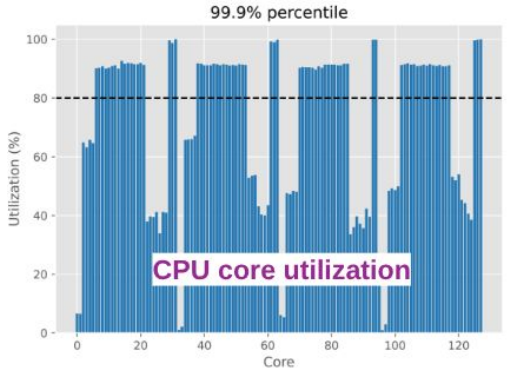
Critical Component Isolation:

- CPU affinity masks for majority of threads
- Sensitive components are **kernel isolated**
- Minimizing socket cross-talk & allow single-copy to buffers
- Allocate resources based on device locality & PCIe root complex
- Cache topology & attention to features: DCA vs. DMA

Vendor Agnosticity

Extensive Testing Since 2017:

- **Intel & AMD Support:** Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, Zen3-5
- **NIC Agnostic:** Evaluated with the 3 main vendors
- Attention to feature set differences both for NICs and CPUs
- **Baseline change to 400Gbps aggregation** in scale-up testing



DPDK SUMMIT
Powering the Future of Networking Software

CERN

DAQ Operations @ 400Gbps

- **DUNE prototype detectors at the Neutrino Platform facility are read out with single high-performance servers**
  - 4 x CRPs & 4 x APAs, close to line rate at 4x100Gbps inputs
  - Validated every readout component operating in parallel (including NVMe persistency and line rate data processing)

- Motivation is the **power-draw reduction and price optimization** with this configuration
  - Factor 2 reduction of needed servers with this topology

- **Load balancing and resource isolation techniques are essential** to reach deterministic performance on quasi-real time COTS hardware and software
  - Balancing O(100) processing threads per 100G input
  - From which, couple of cores are designated to networking

- Performance testing activity includes **analysis of performance reports that lead to further optimization options** that are continuously fed into the readout system

Neutrino PLATFORM

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

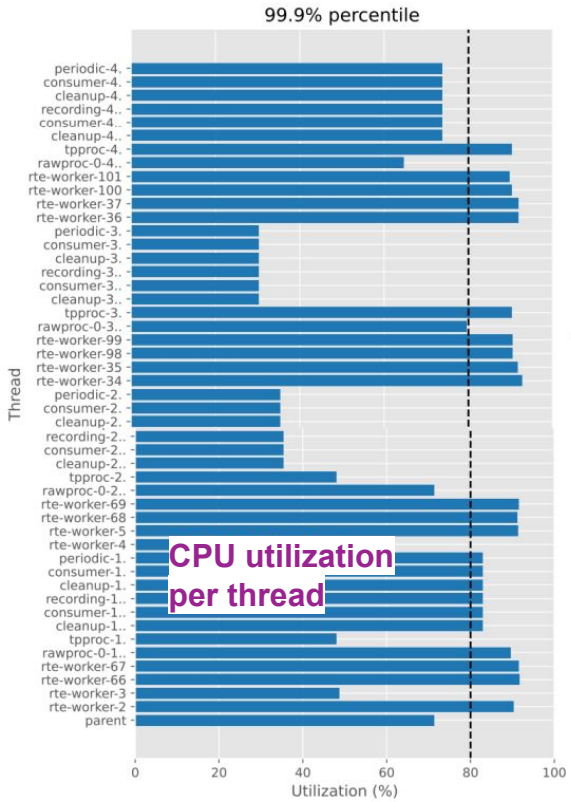
Established DAQ performance with real detector apparatus and conditions: **Quasi DUNE FD conditions achieved in Q4 2024!**

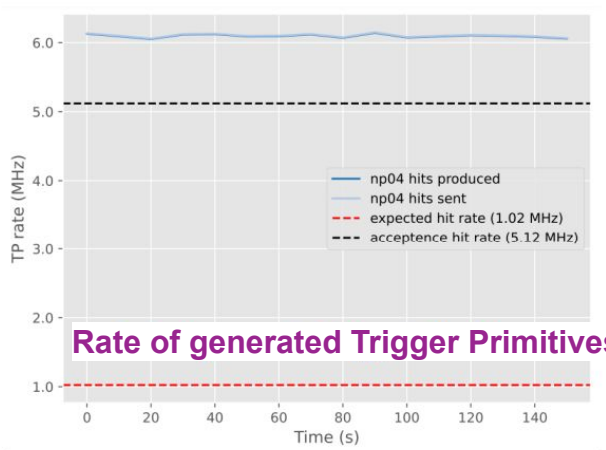
- **6+ MHz of Trigger Primitives (TPs)**
- **Data requests @ downstream network limit ~ 7 Gbps + ~1.5 Gbps (TPs) sustained / 10G down**
- **Most complex feature extraction algorithm used**
- **SNB on NVMe recording criteria satisfied**
- **No packet loss with stable resource utilization**

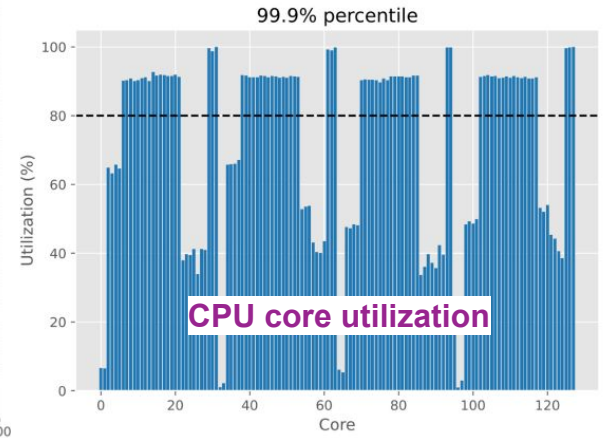
DPDK SUMMIT
Powering the Future of Networking Software

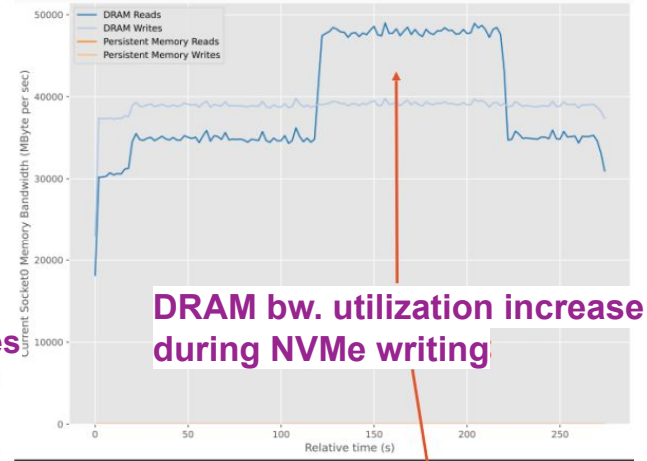
CERN

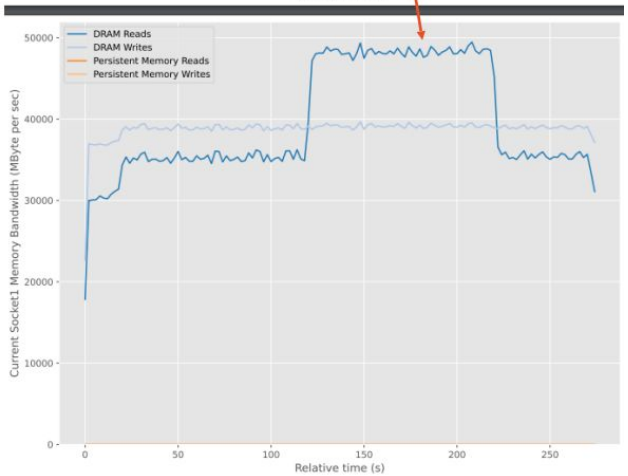
Performance











Metrics successfully meet KPI requirements!

DPDK SUMMIT
Powering the Future of Networking Software

CERN


Frontier & Beyond

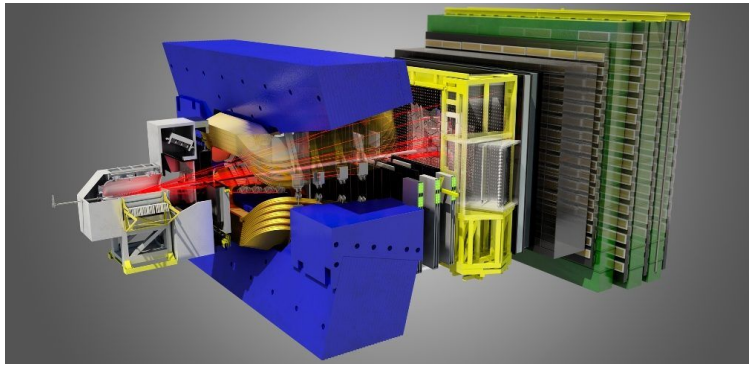
LHCb

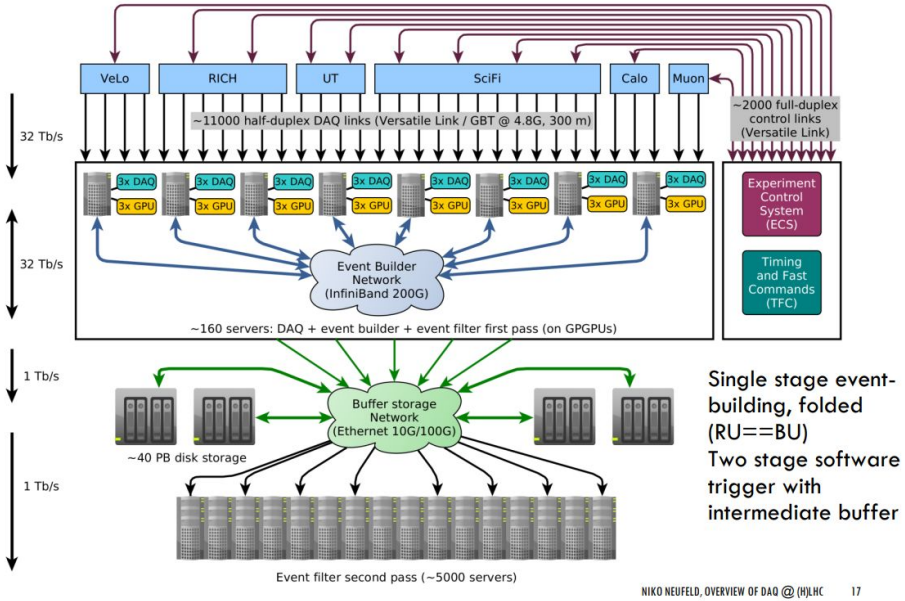
Physics & DAQ

- 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 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.
- Motivation to transition from proprietary InfiniBand 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.

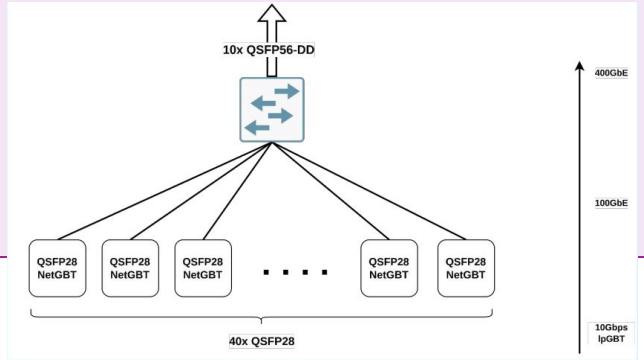
The radiation challenge is maintained and increases:

- **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

DPDK - Software Foundation for DAQ R&D

Standardizing the software solutions

As DPDK serves already as a critical high-performance backend layer, enabling near-line-rate UDP reception for multiple experiments and test-beds, there is a motivation to:

- **Unify best practices:** Develop or use libraries that encapsulate hardware resources to logical DAQ components.
- **Maintain scalability and performance:** Support reception of multi-100Gbps aggregated inputs, without compromising other mission-critical parts (e.g: processing)
- **Ease of use:** Hide networking complexity and provide the ability of configure the overall topology and balancing at ease
- **Support for SMEs** (Small and Medium scale Experiments): Experiments that have limited expertise and personnel should be able to reuse already developed ecosystems.

DPDK-DAQ

Development is already ongoing that aims for a common software architecture inspired by the **DUNE** experiment architecture.

- Exercised through **SmartNIC R&D** for FPGA offloading of parts of the DUNE-DAQ data processing pipeline that is not CPU friendly.
- A customizable detector-emulation demonstrator is integrated into the **DAQling** common framework to build and deploy DAQ topologies on COTS hardware.

- Extensible for various detector R&D projects, and upcoming mid-sized experiments like SHiP.

Summary and Outlook

Current Achievements

- DPDK successfully integrated as standard high-performance backend for physics experiments
- Using DPDK PMDs with various commercial NICs achieved near-line-rate UDP reception.
- Scalable architecture developed and in operation, supporting up to 400 Gb/s throughput.

Future Outlook

- Expansion of the Unified Software Stack for diverse R&D projects.
- SmartNIC integration for advanced FPGA offloading capabilities.
- Addressing HL-LHC scaling demands (300 Tbps target).

References

- The DAQ of NA62
- Ethernet readout of DUNE DAQ

Thank you for your attention!

Your comments, questions, and suggestion are more than welcome!

Many thanks for the work and contributions of my colleagues 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**

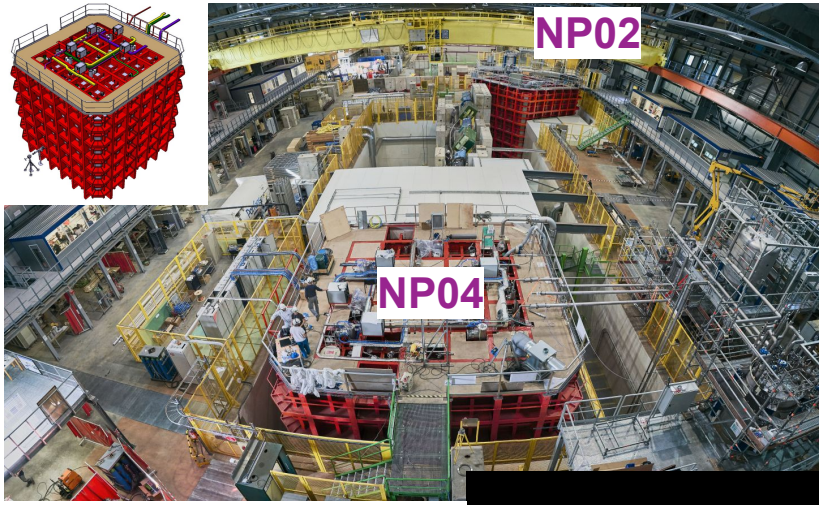
Shoutout to the following collaborators:

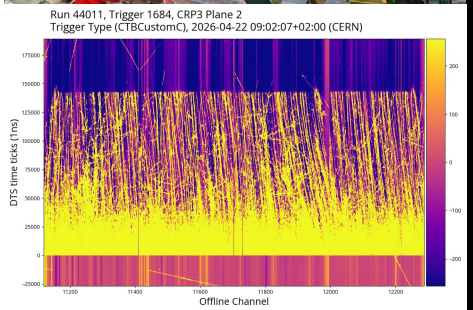
- **DAQ community at CERN**
- **CERN Neutrino Platform**
- **DUNE Collaboration, especially:**
  - **UK institutes,**
  - **Rutherford Appleton Laboratory (RAL),**
  - **University of Toronto**

Backup Slides

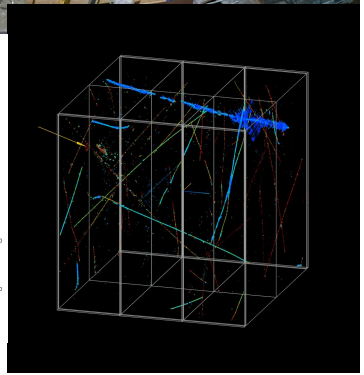
CERN Neutrino PLATFORM & DUNE prototypes

- 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

Ethernet TX firmware block

The DAQ team 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 the User Datagram Protocol (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.

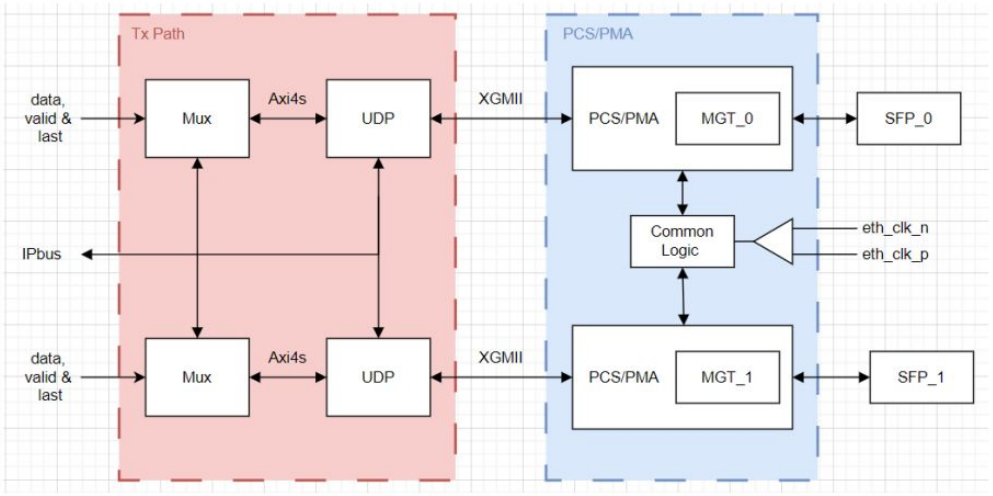


Fig. 2. Architecture of the transmitter block provided for the front-end electronics. The PCS/PMA area contains modified Xilinx IP[6] components, and the Tx Path is a custom firmware block developed by engineers at RAL Technical Division. The overall block is responsible for equipping the detector data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


DPDK
SUMMIT
Powering the Future of Networking Software