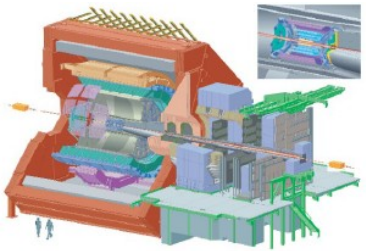
Commissioning of the ALICE readout software for LHC Run 3

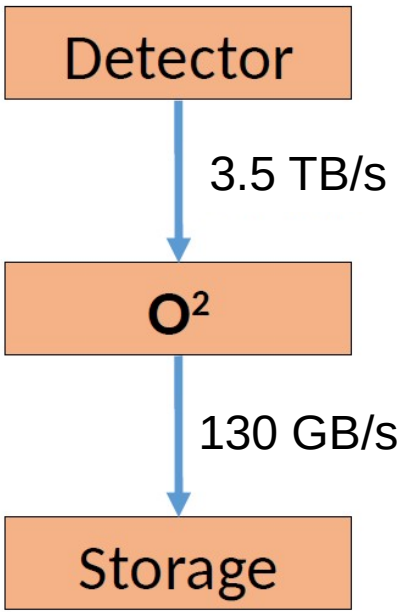
Sylvain Chapeland (CERN)
for the ALICE Collaboration

26th International Conference on Computing in High Energy and Nuclear Physics, CHEP 2023

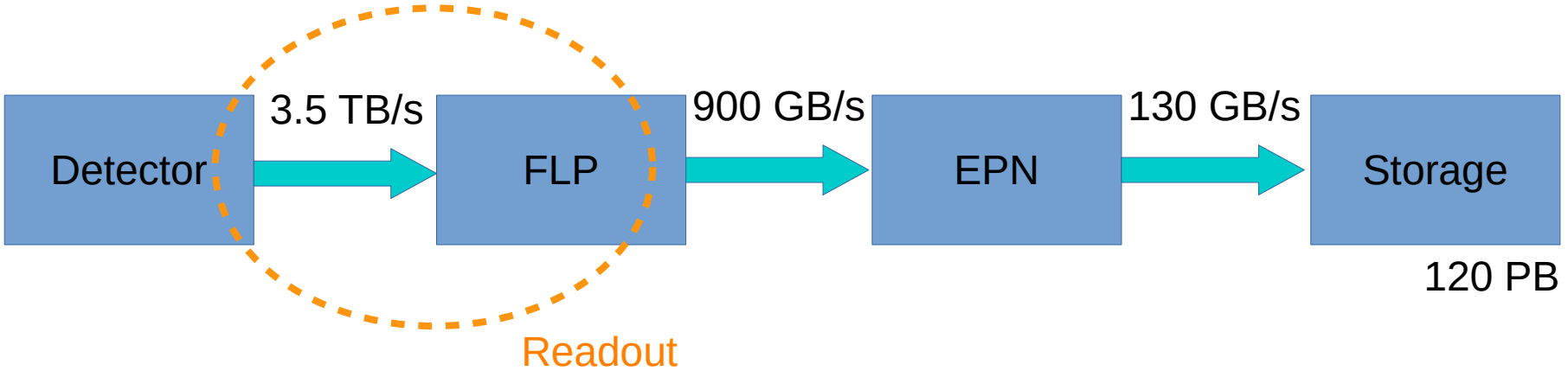
ALICE

- Upgrade highlights for LHC Run 3
  - Time Projection Chamber (TPC)
    36 GEM readout chambers, 500k channels, continuous readout
  - Inner Tracking System (ITS)
    CMOS chips in 7 layers, 12 Gpixels
  - Muon Forward Tracker (MFT), Fast Interaction Trigger (FIT)
  - 15 subdetectors in total
- Increased data throughput: *x100 vs LHC Run 2*
  - 3.5 TB/s from the detector
  - 8000 optical links
  - Streaming readout
- Demanding online processing and compression
  - O2: online – offline system
  - Synchronous and Asynchronous processing



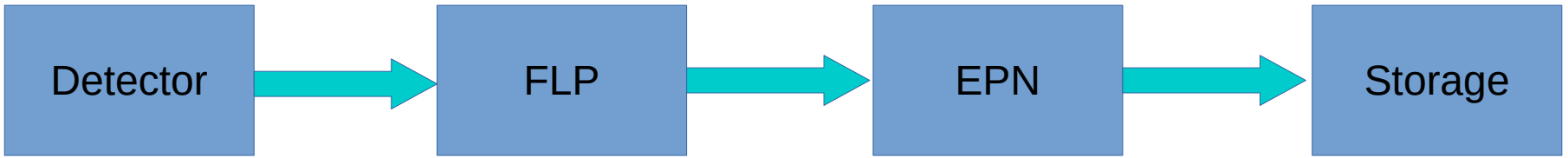


Data Acquisition and processing – online dataflow



- Readout: First Level Processors (FLP)
- Reconstruction: Event Processing Nodes (EPN)
- Data reduction in 2 steps

Data Acquisition and processing – online dataflow



Other related ALICE CHEP'23 talks (track)

11361 - The new ALICE Data Acquisition system (O2/FLP) for LHC Run 3 (2)
11373 - The ALICE Experiment Control System in LHC Run 3 (2)
11385 - The ALICE Data Quality Control (2)
11461 - Bookkeeping, a new logbook system for ALICE (Poster)
11527 - Security Models for ALICE Online Web-Based Applications (5)

DAQ, Control, QC, Bookkeeping, GUI

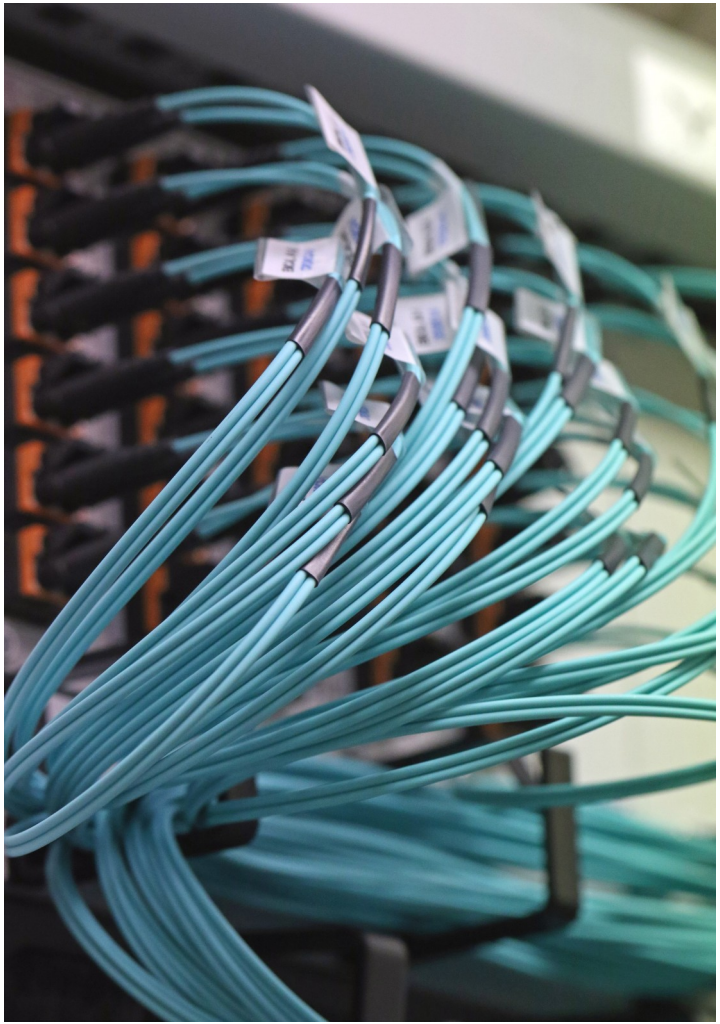
12432 - The O2 software framework and GPU usage in ALICE online and offline reconstruction in Run 3 (5)
11383 - A vendor-unlocked ITS GPU reconstruction in ALICE (2)
11317 - Calibration and Conditions Database of the ALICE experiment in Run 3 (1)
11299 - EPN2EOS Data Transfer System (1)

Reconstruction
GPU
Storage

Detector Data Links

- Mainly Versatile link / GBT
  - radiation-hard bi-directional 4.8 Gb/s optical fiber link between counting room and experiment
  - It delivers/receives : DATA, TRIGGER and SLOW CONTROL.
- Support for ALICE custom link DDL1 (used during Run 1 and Run 2)
  - 2.125 Gb/s
- ~8’000 links in total

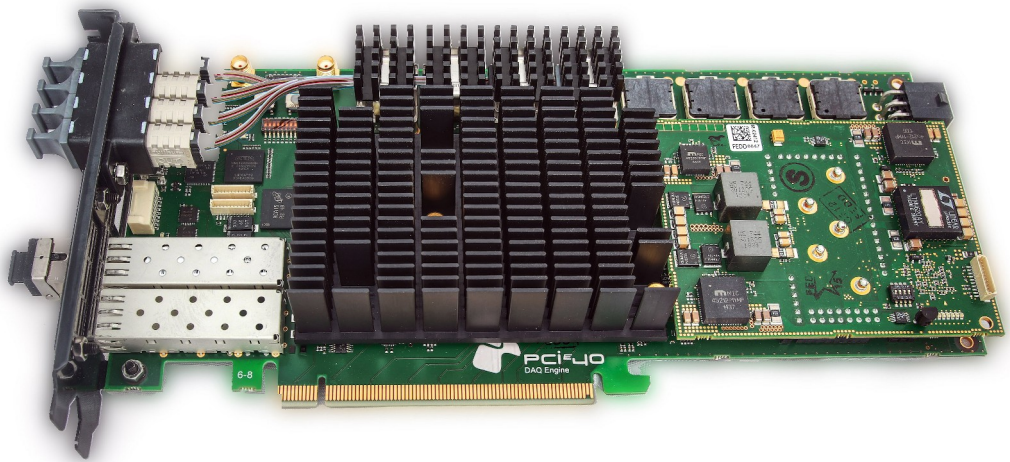
=> need for high-density readout system !



Detector links patch panel in counting room (CR1)

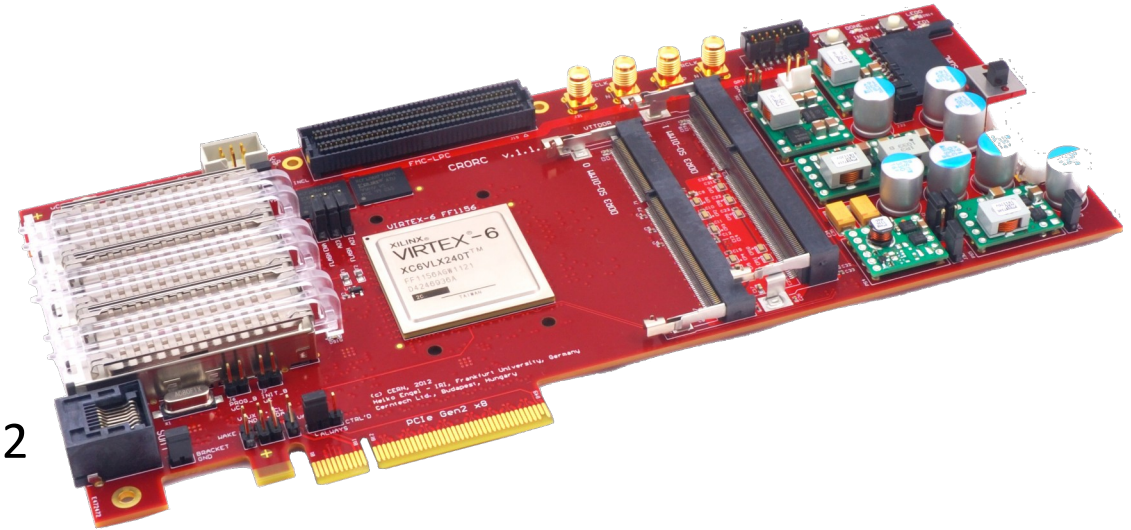
Readout hardware - GBT detector

- CRU (a.k.a. LHCb PCIE40)
  - FPGA: Intel Arria10
  - Using maximum 24 GBT links bidirectional
- PCIe gen.3 x16
  - Dual DMA engine Gen3 x8
  - Typical throughput: 110 Gb/s
- Firmware allows dedicated user logic for on-board compression

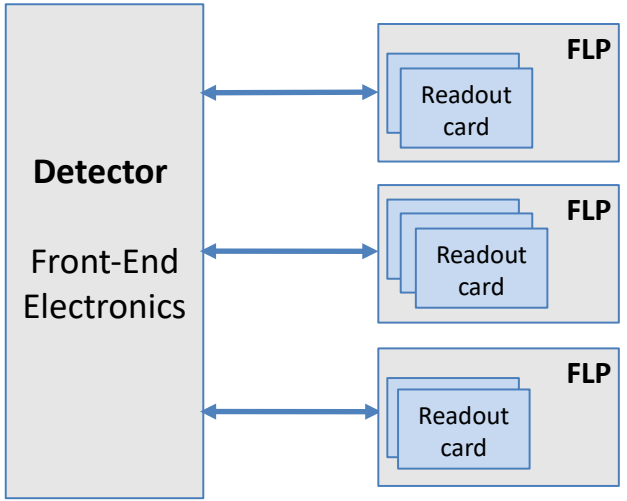


Readout hardware - DDL detector

- C-RORC
  - FPGA: Xilinx VIRTEX6
  - 6 DDL links
  - PCIe gen.2 x8
  - Used in ALICE Run 1 and Run 2



Readout hardware - servers



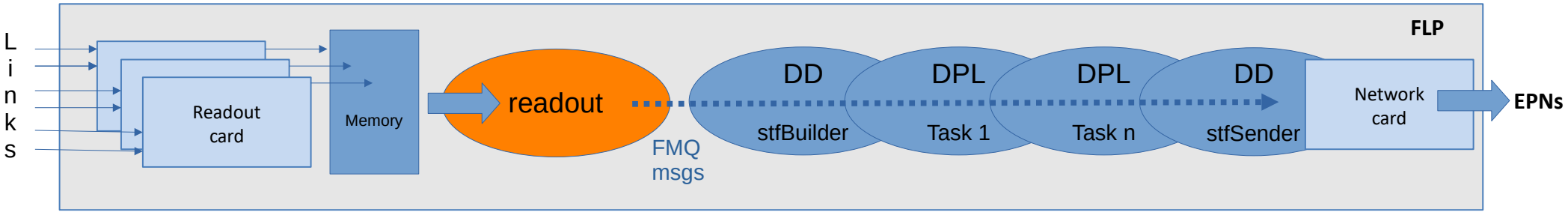
200 servers

500 readout cards (up to 3 per host)

Mainly:

2 x 10 cores CPUs Intel Xeon 4210
96GB RAM

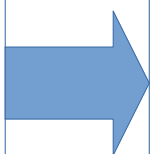
Readout software



- Input: readout hardware
  - 2 types of PCIe boards
  - Raw data written to memory
- Output: O2 Data Distribution software (DD)
  - Organizes network transport (Infiniband) to EPN processing nodes
  - O2 message-based format, using FairMQ and shared memory inter-process transport
  - Some processing tasks may run on the FLP readout nodes (Data Processing Layer, DPL framework)

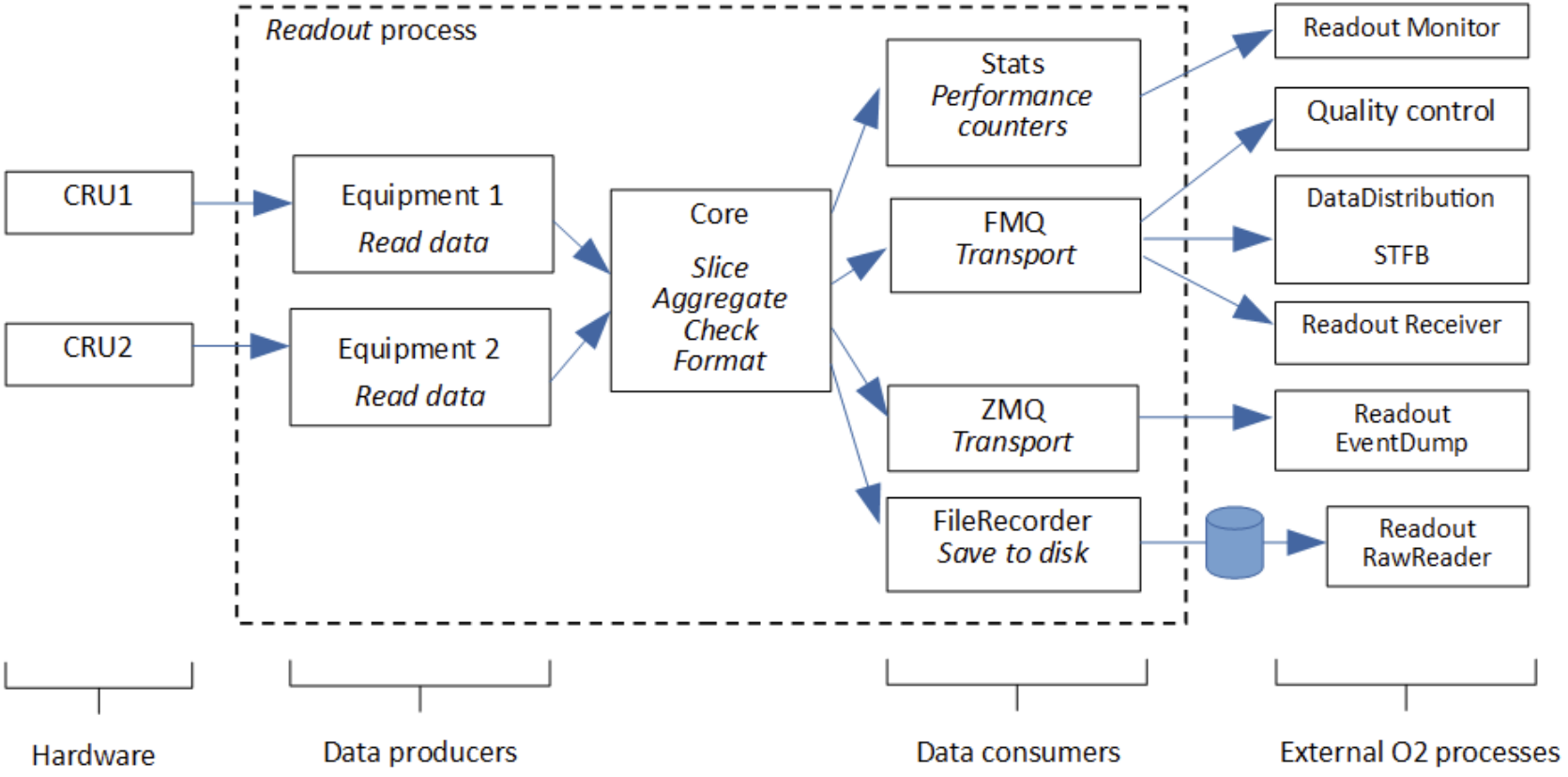
Readout software - requirements

- Adapt to input diversity
  - 15 subdetectors, 2 types of readout cards
  - Various data formatting flavors
    - packing, FPGA compression, etc
    - 1 common data format (RawDataHeader, RDH)
  - Various running modes – throughput and content
    - Physics (p-p, Pb-Pb, Cosmics), Calibration, Synthetic runs (data replayed from simulation files), Debugging
  - Detection and robustness against payload issues
- Adapt to output data format
  - Timeframe-based slicing
  - Shared memory
  - Messaging protocol
  - Metadata headers



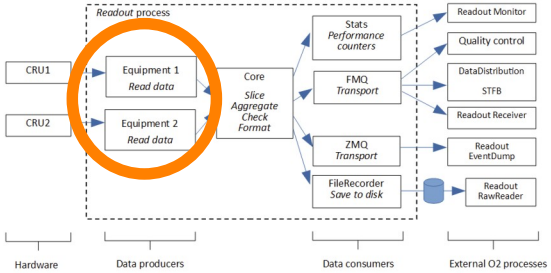
- Accommodate and find compromise to handle external constraints
  - Sometime conflicting hardware and data distribution / processing needs.
- Primary goal on performance
  - within fixed resources available (CPU, memory)
- Highly flexible configuration needed
  - Possibly at the cost of increased complexity
  - 200 nodes doing different things at different times

Readout software



Readout process – internal threads and data flow

Readout software – tasks



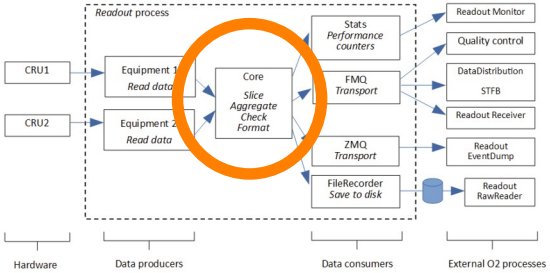
#1 – Read data

- Initialize hardware

  CRU, C-RORC using common ROC (Read-Out Card) driver interface

- Allocate memory buffers
- Provide data pages to be filled by PCIe device
  - Data are transferred into memory by DMA
  - Software provides / gets pointers to empty/ready blocks

Readout software – tasks

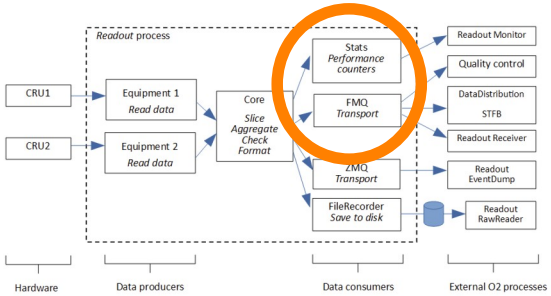


#2 – Aggregate and slice data

- Group data of different sources by time interval
  - Trigger and continuous detectors
  - Chunks of 32 LHC orbits = 1 timeframe
- Check data consistency

  Raw Data Headers from the incoming payload provide trigger counters, detector status bits, structure sizes, etc

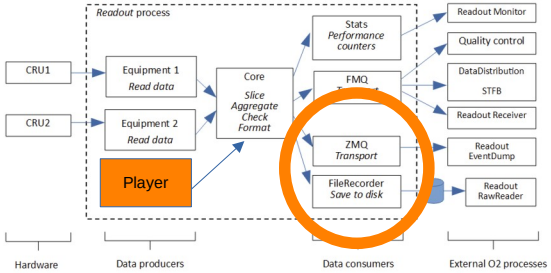
Readout software – tasks



#3 – Distribute data to consumers

- Formatting into O2 messages and adding top-level headers
- Forwarding to Data Distribution software in charge of pipelining local processing and sending to EPNs
- Report performance and errors

Readout software – tasks



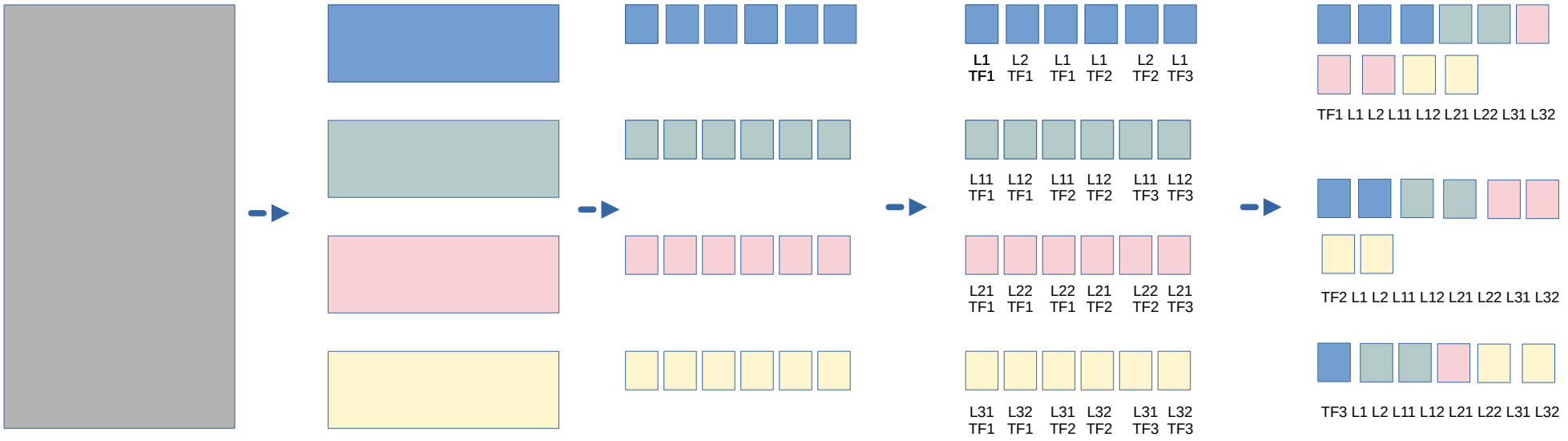
#4 – Side options

- Special features used for commissioning:
  - on-the-fly LZ4 compression
  - recording to disk
  - online data monitoring
  - simulated data file player
    - *exercising full online processing chain with realistic data*

Release cycle

- Software
  - 35 releases since beginning of 2022
    - No change in baseline concepts, but many on-demand features
  - Follows FLP weekly deployments
  - Sometime on-the-spot compilation for custom testing / debugging
- Configuration
  - Automated config generation
    - From limited source variables, CSV table: hosts, hardware serials, memory size
    - GIT repository to track reference of multiple configuration flavors
  - Support tools
    - NFS - based distribution of reference files
    - Consul instance for production

Memory layout



1 big block of shared memory ~32GB

1 block per DMA channel ~8 GB each

Data pages ~1MB each

CRU writes 1+ page per TF per link

Readout slices per TF per link

and further cuts in FMQ message parts of smaller size

Circular buffer

Memory layout

- Page size adjustment

  Not too big (loose space when not filled), not too small (inefficient)

- Fine-grain configuration needed

  Per detector, per host, per CRU… and possibly even per-link

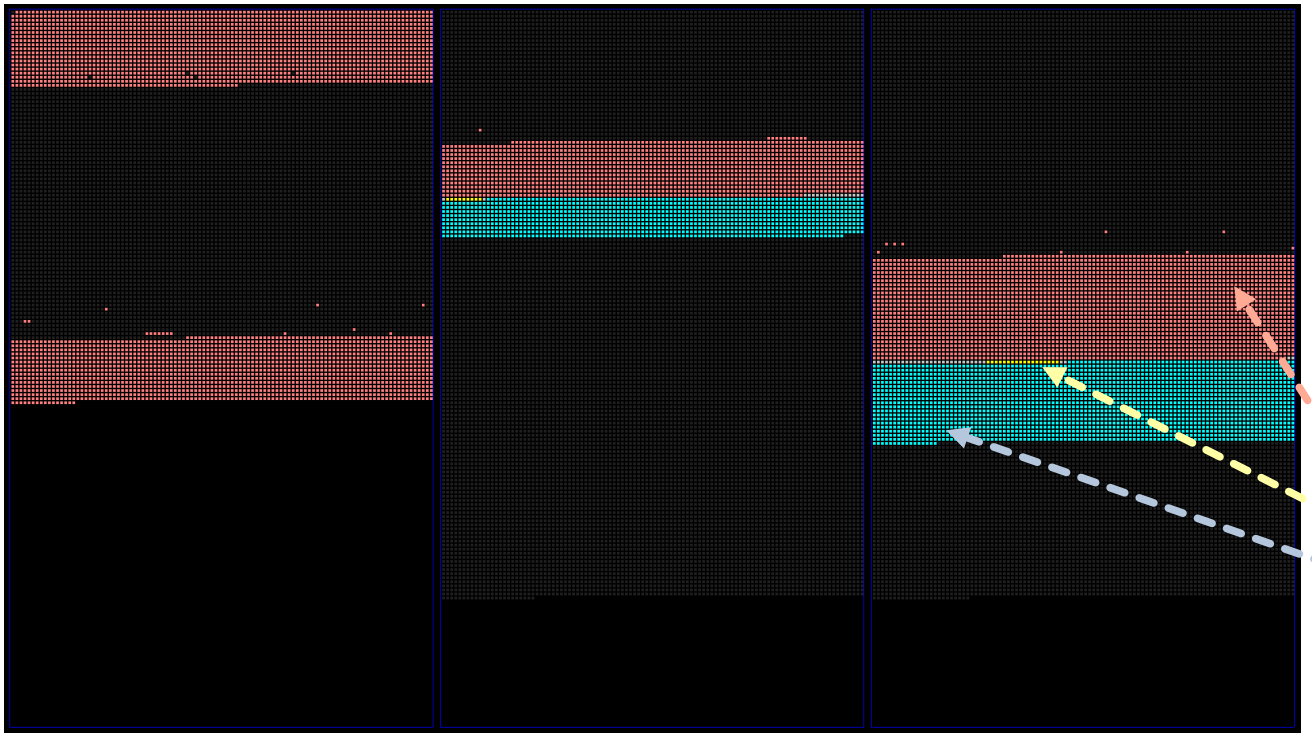
  - To adapt to individual link rates and events sizes

- Tuning requires detailed monitoring

Memory buffers monitoring

Time-based plots: host page release rate and latency, usage of each buffer.

Memory buffers monitoring



Realtime view

State of each data page is shown
Low-latency / high refresh rate

Pages being processed downstream
Pages being checked/formatted by readout
Pages in CRU buffer waiting to be written

Video stream example

Commissioning

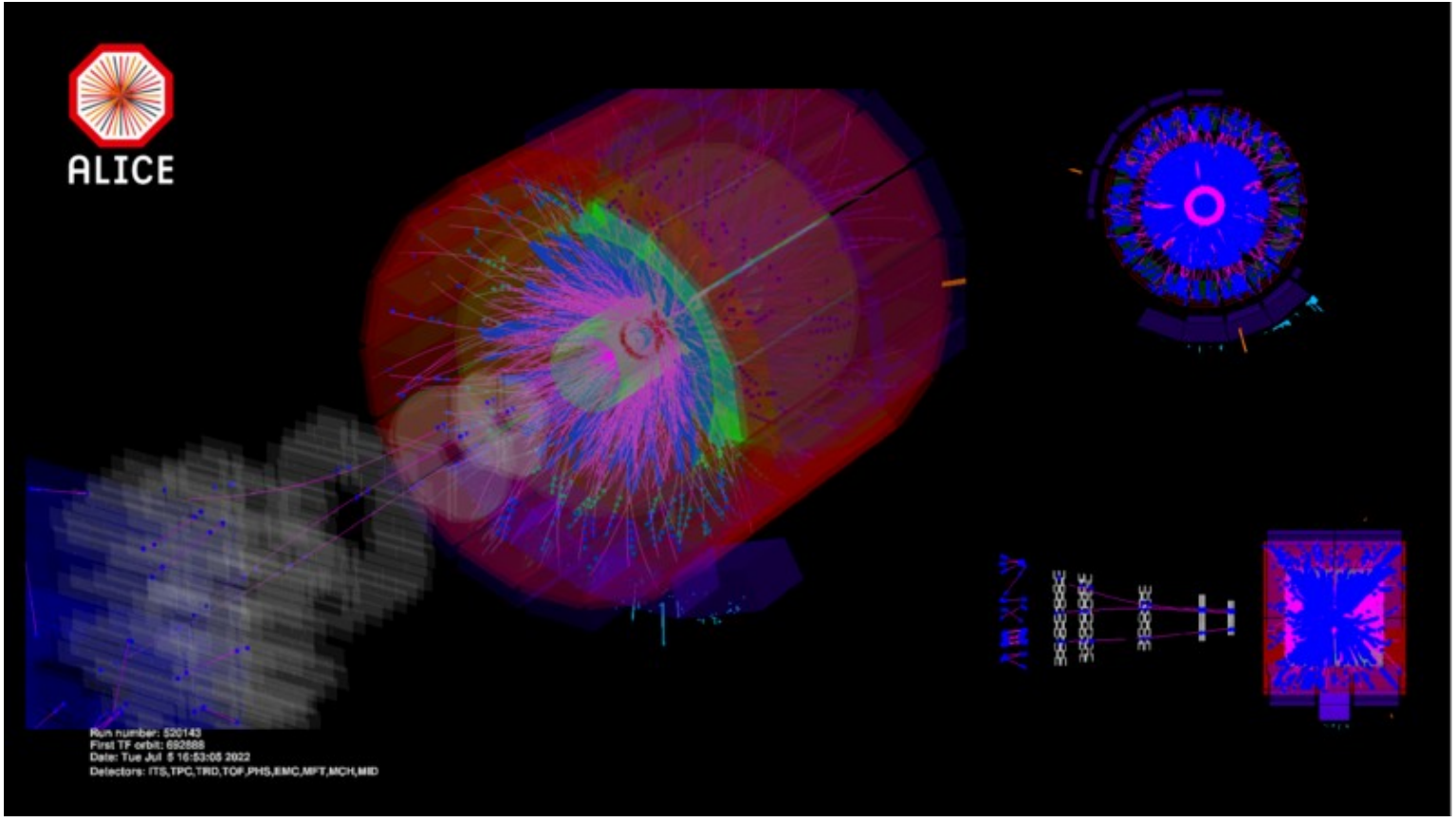
- Lots of work done in 2022 to get the system ready for LHC restart

  Over 2000 PB have been readout in 200 days (test + physics data)

  i.e. ~0.6 GB/s/FLP on a 6 months period

ALICE Control Room

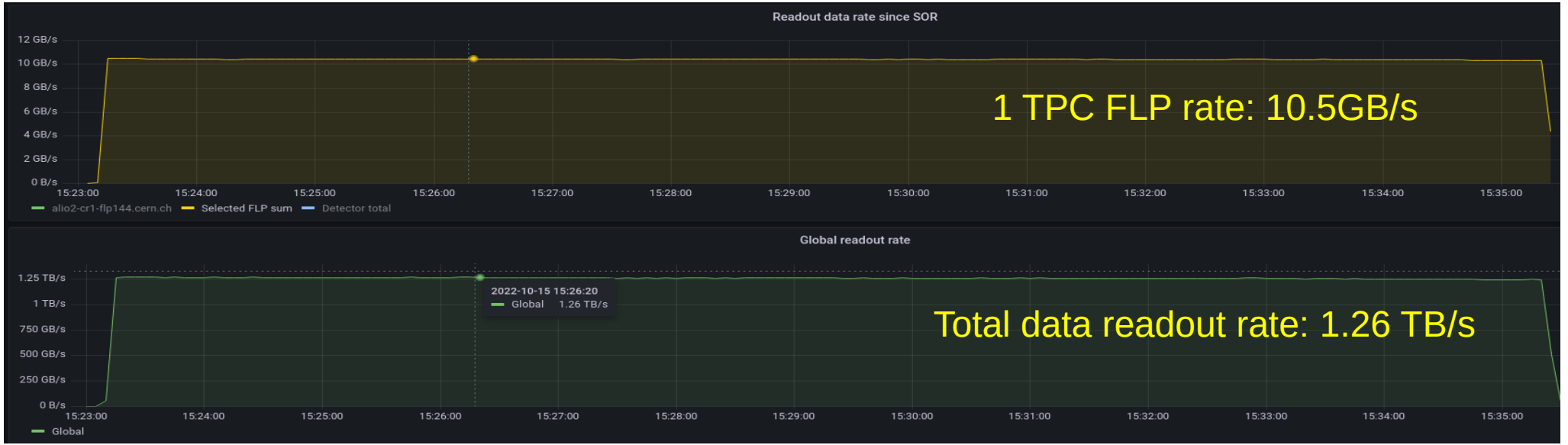
Commissioning



5 July 2022: ALICE first 13.6 TeV collisions of LHC Run 3

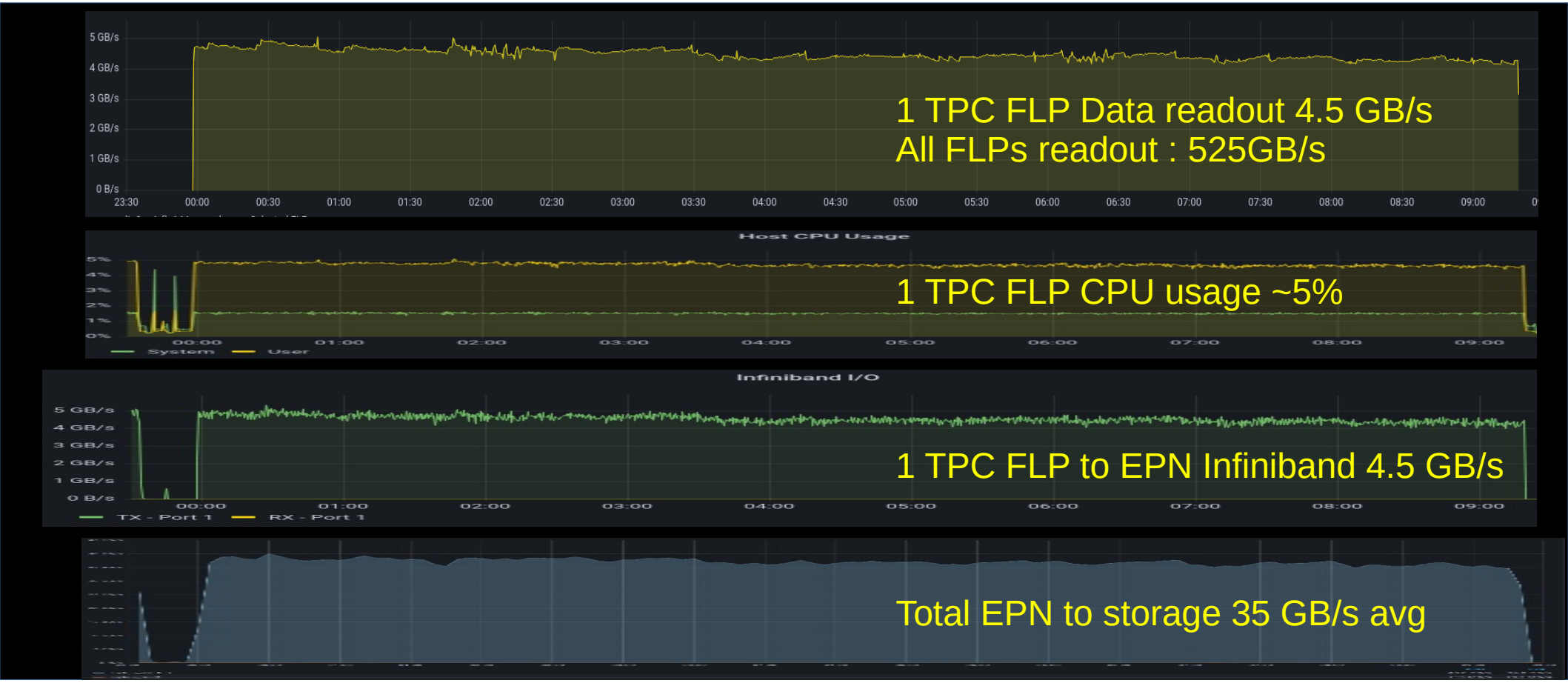
Readout and DAQ Performance

- Running routinely above the design validation criteria of 70Gb/s per FLP for testing
- Readout data flow exercised in demanding and changing conditions
  - Optimization of memory buffers (best page size depends on detector payload)
  - Little NUMA effects seen, plenty of headroom on QPI links



Run 527397 2MHz pp – no EPN – readout test at high rate

Readout and DAQ Performance



Physics run 527446 – duration 9h20m50s

Outlook

- Streaming readout system doing fine
  - Design margins find their use to accommodate detector evolving needs
  - Hardware does the job
  - Ongoing software developments: always more tools and features needed
    - Ease support and operations, facilitate performance optimization

- ALICE started Run 3 with success
  - Actively taking cosmics and p-p physics data
  - Exercising data flow and processing with p-p at high rates and simulated files
  - 2023 operations resumed
  - Waiting for HI collisions