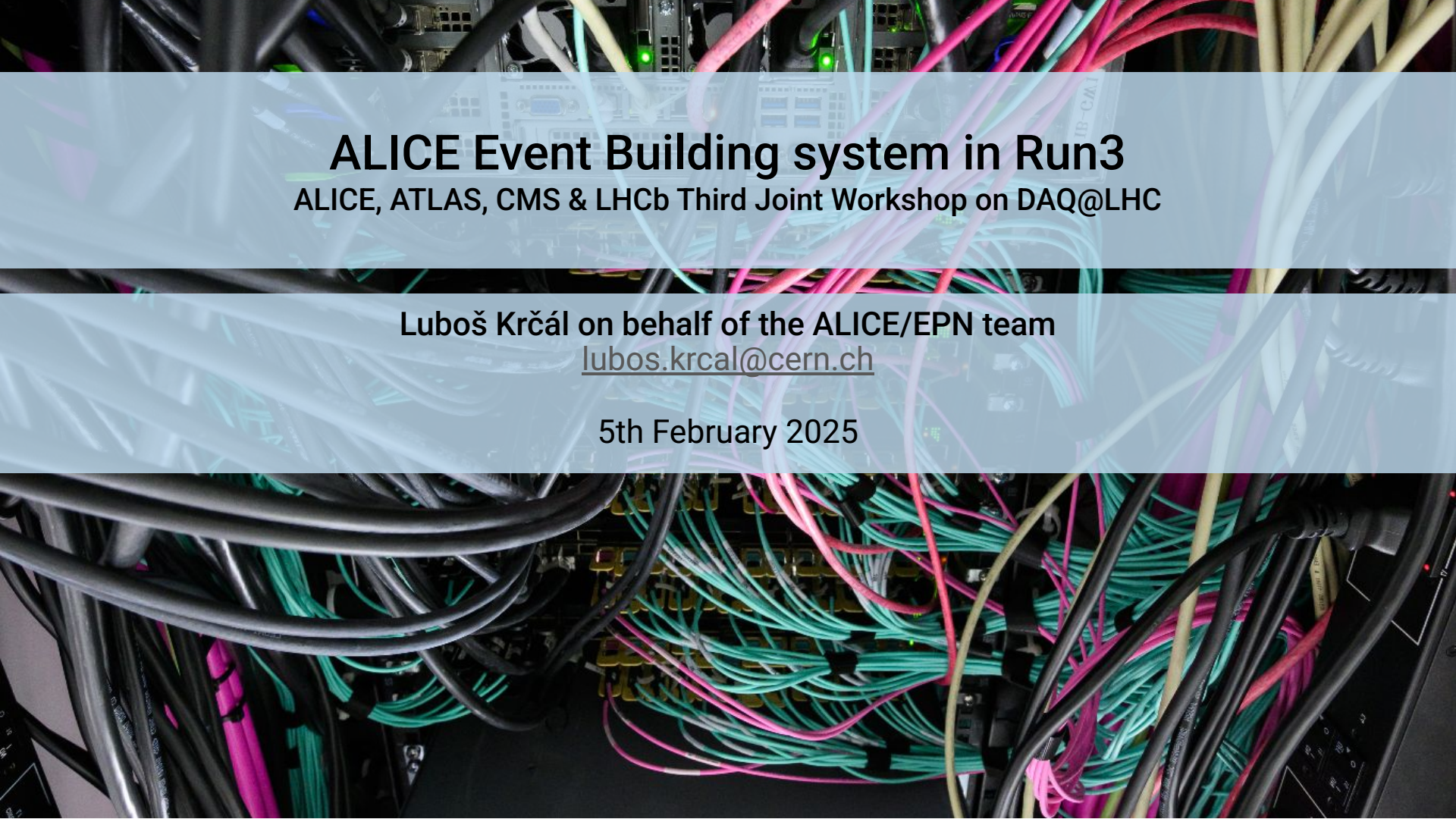


ALICE Event Building system in Run3

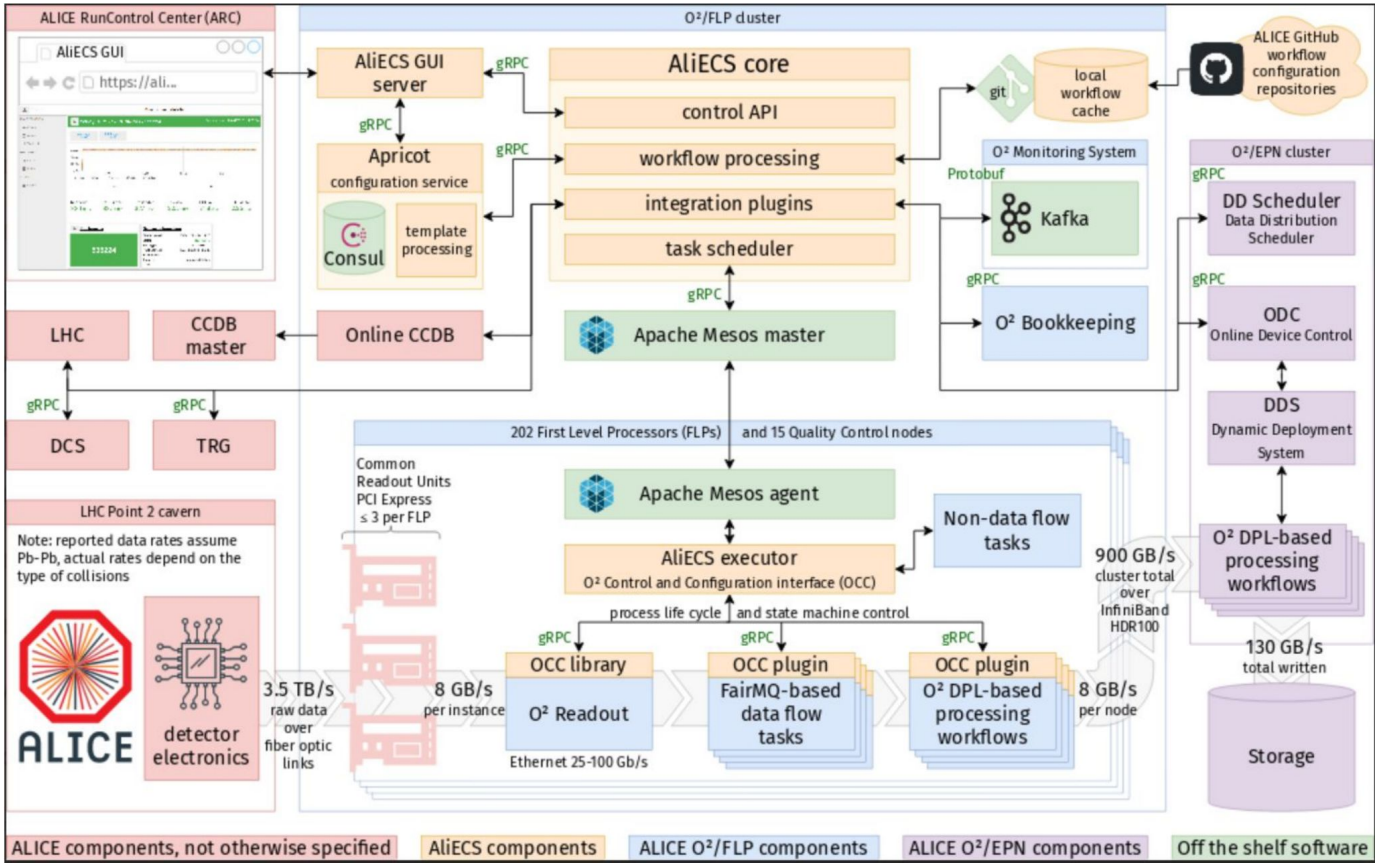
ALICE, ATLAS, CMS & LHCb Third Joint Workshop on DAQ@LHC

Luboš Krčál on behalf of the ALICE/EPN team
lubos.krcal@cern.ch

5th February 2025

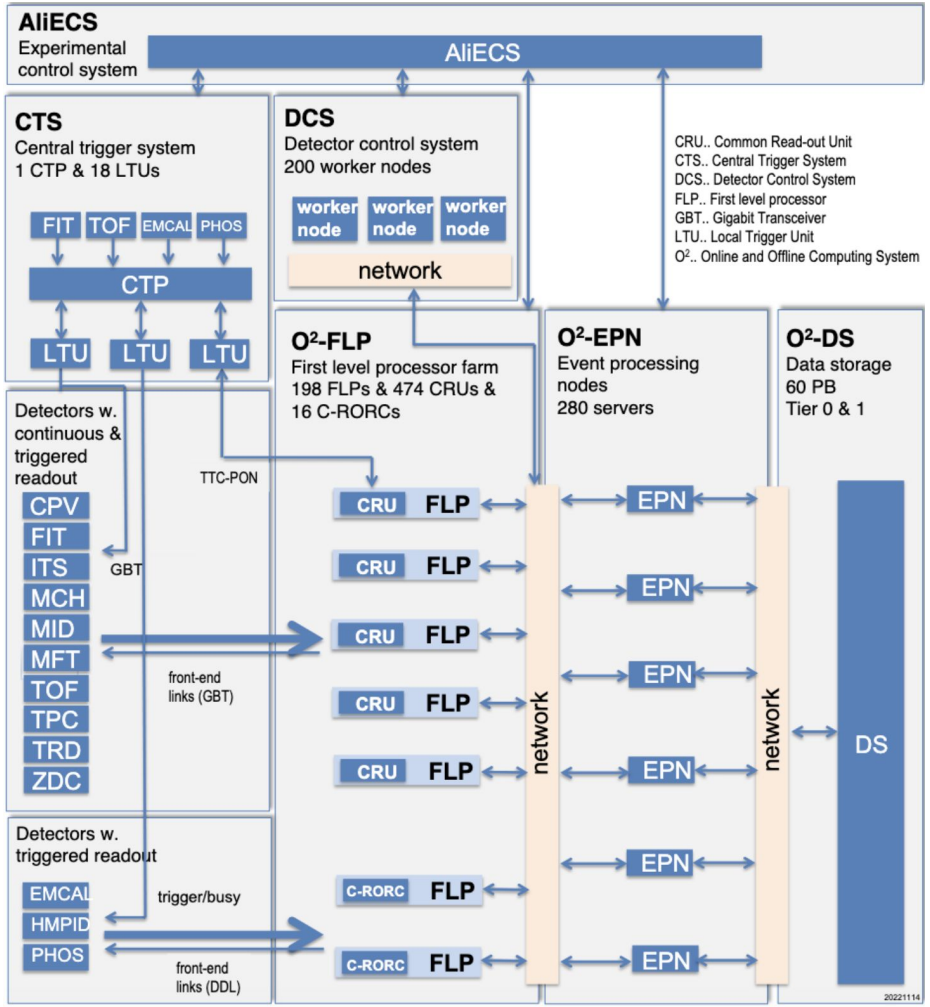
ALICE Architecture

Find event building…

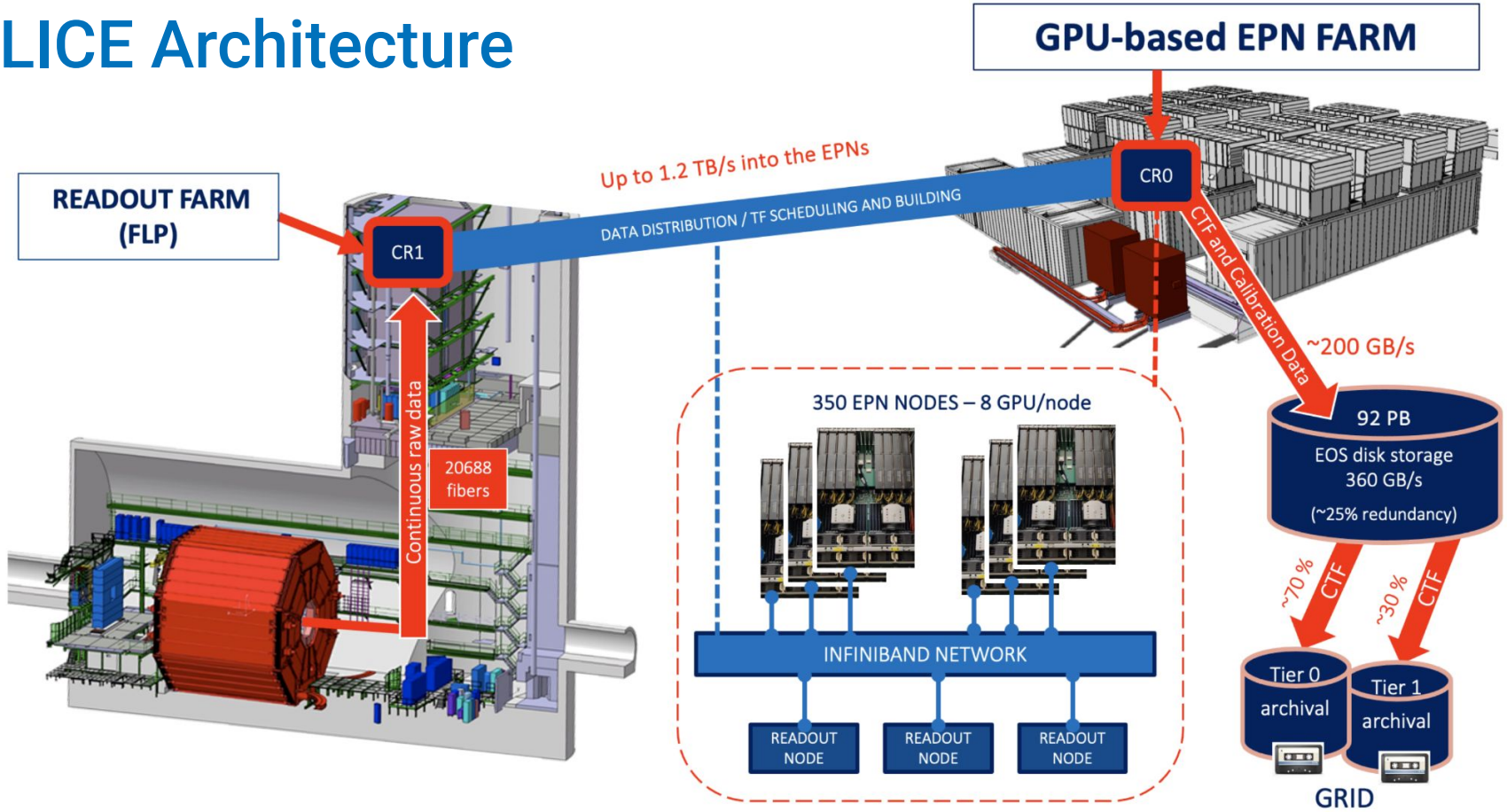


ALICE Architecture

Find event building…



ALICE Architecture

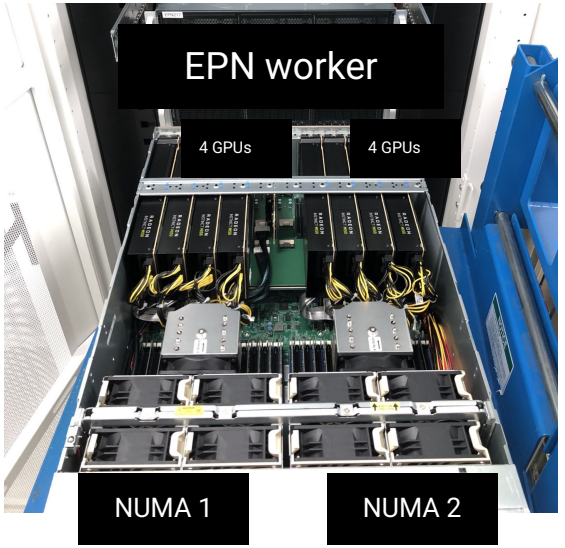


ALICE EPN Farm Recap

	70 MI100 EPNs	280 MI50 EPNs	16 Calib / Infra Nodes
GPU	8 AMD Instinct™ MI100 32 GB	8 AMD Instinct™ MI50 32 GB	
CPU	2 AMD EPYC™ 7552, 48 cores	2 AMD EPYC™ 7452, 32 cores	2 AMD EPYC™ 7452, 32 cores
MEMORY	1TB DDR4 3200 MHz	512GB DDR4 3200 MHz	512GB DDR4 3200 MHz
Networks	IB HDR 200/100 Gb/s, ETH 1 Gb/s		

Hardware specifications

- **280 nodes** equipped with 8 AMD MI50 32GB GPUs
- Additional **70 nodes** (installed in 2023) equipped 8 MI100 32GB GPUs
- Grand total of 350 (280 MI50 + 70 MI100) nodes and 2800 GPUs
  - Equivalent to ~373 MI50 nodes at MI100 = 4/3 MI50
- **Total of 43 PFLOPs FP32 from 2800 GPUs**



PERFORMANCE	MI-50
Compute Units	60
Stream Processors	3,840
Peak INT8	Up to 53.6 TOPS
Peak FP16	Up to 26.5 TFLOPS
Peak FP32	Up to 13.3 TFLOPS
Peak FP64	Up to 6.6 TFLOPS
Bus Interface	PCIe®Gen 3 and Gen 4 Supported[2]

PERFORMANCE	MI-100
Compute Units	120
Stream Processors	7,680
Peak BFLOAT16	Up to 92.3 TFLOPS
Peak INT4 | INT8	Up to 184.6 TOPS
Peak FP16	Up to 184.6 TFLOPS
Peak FP32 Matrix	Up to 46.1 TFLOPS
Peak FP32	Up to 23.1 TFLOPS
Peak FP64	Up to 11.5 TFLOPS
Bus Interface	PCIe® Gen 3 and Gen 4 Support[3]

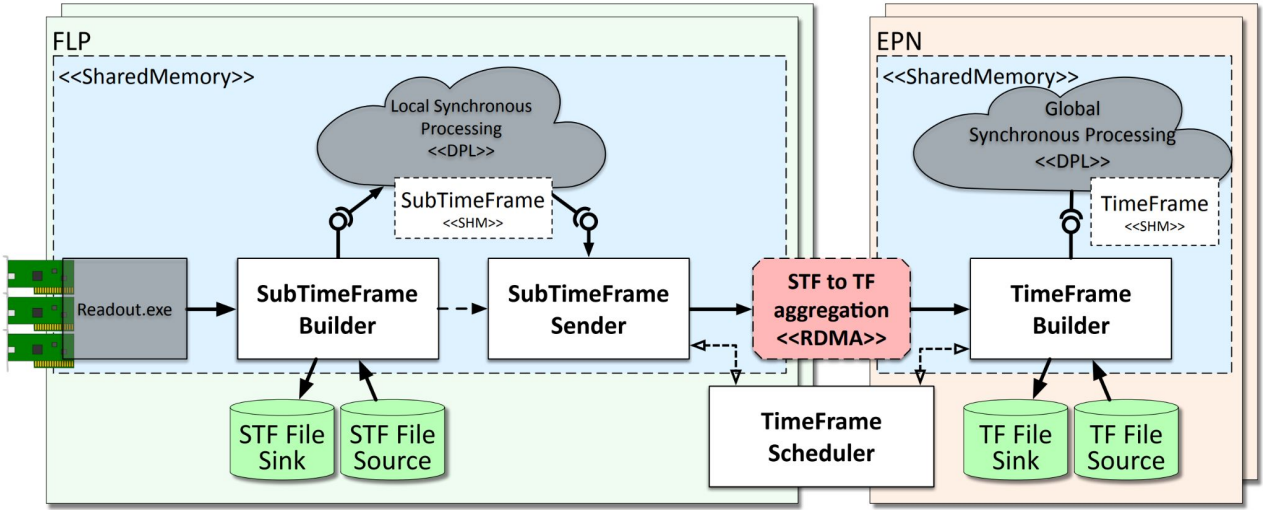
ALICE Architecture - Streamlined & Simplified

Input (13 Detectors)	**Map** (200 FLPs)	**Gather** (Infiniband)	**Map** (<=350 EPNs)	**Store** (EOS 150PB)
Trigger Heartbeat frames	Readout FPGA Zero suppression **STF (SubTimeFrame) building** Preprocessing / QC STF Buffer (32GB)	Data Distribution **RDMA of STFs over Infiniband** Software orchestrated	**TF (TimeFrame) aggregation** Event processing on GPUs TPC track clustering + reconstruction Calibration Compression	Warm storage EOS protocol over ETH
~4 TB/s	**~4 TB/s –> ~800 GB/s**	**~800 GB/s**	**~800 GB/s -> ~150 GB/s**	**~150 GB/s**
			Asynchronous processing GRID site	

Data Distribution Architecture

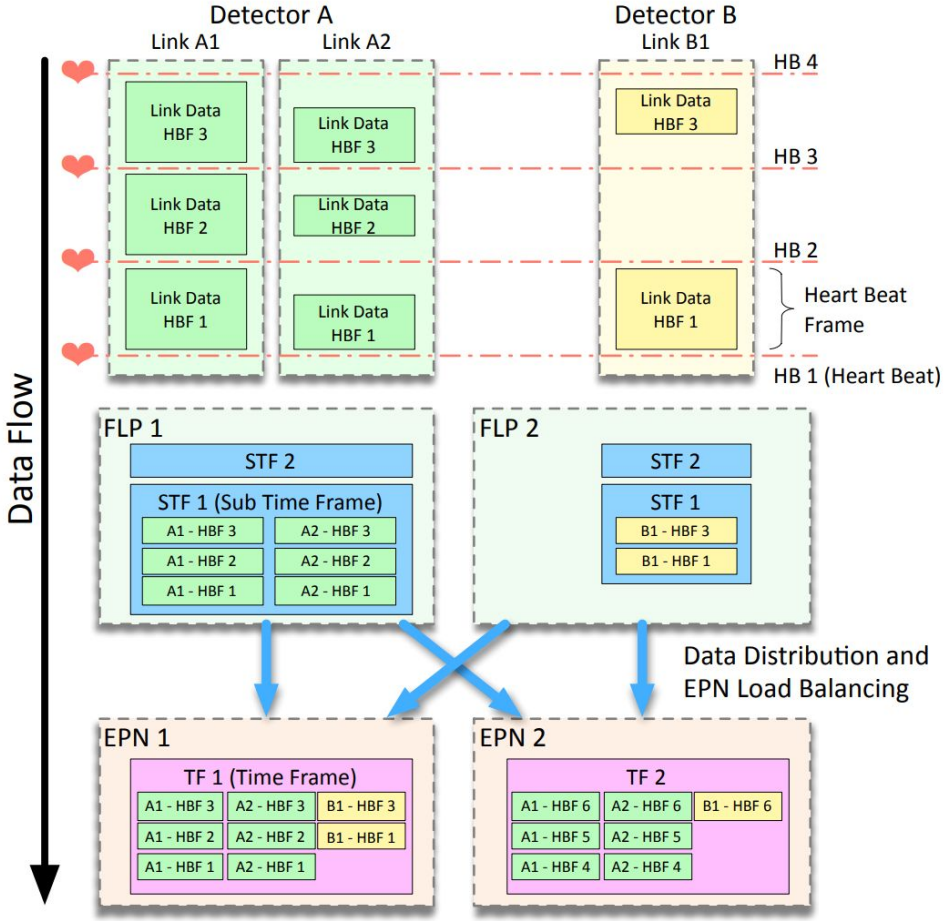
- **StfBuilder** receives data from readout (1x per FLP)
- **StfSender** receives STFs from StfBuilder or local processing and sends them to EPNs (1x per FLP)
- **TfBuilder** aggregates and builds TFs (1x per EPN)
- **TfScheduler** controls TF building, and regulates data flow (1x per run)
- **DD_ECS_Service** manages configuration and TfScheduler of based on AliECS requests

- DPL local (FLP) or global (EPN) processing is optional
- Data can be saved to disk at any step



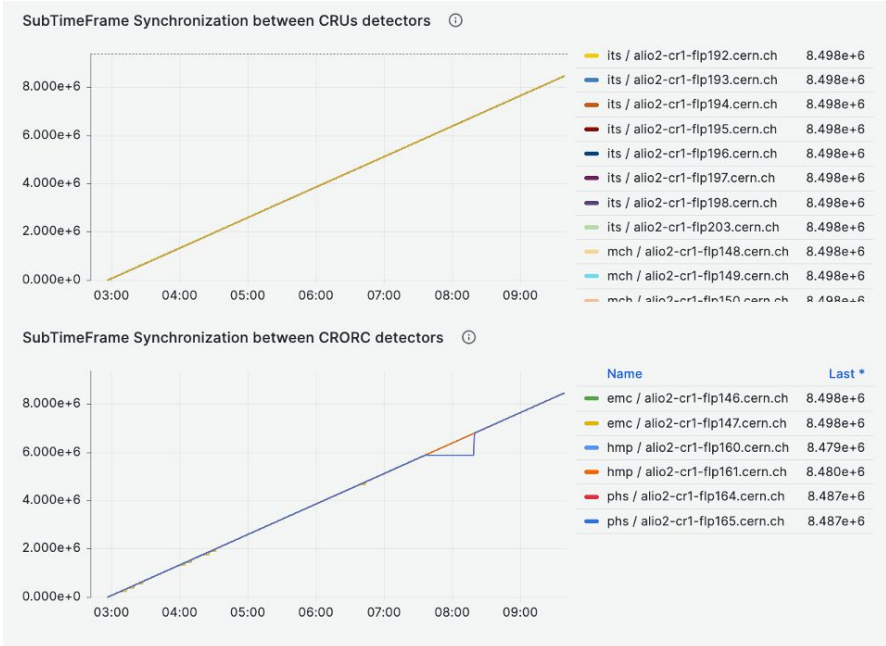
StfBuilder

- At every HB (HeartBeat) a raw data frame is captured
  - Depending on readout type may be sparser - not every HB
- Each frame consists of a header and payload
- Frames are aggregated to STF
  - Currently 32 orbits per STF
- Trigger information is used by StfBuilder to create STFs
- No halo/overlap of STFs in time



StfSender

- Component discovery and connections
- Each StfSender in the partition connects to each TfBuilder
- Metadata communication to TfScheduler
  - New STF information: TF_ID, size, detector, …
- Each StfSender publishes STFs strictly in order
  - Some variability in published times (e.g. some FLPs process STFs), order of 10 ms - OK
- Dedicated 32GB buffer for STFs
  - STF layout in memory is optimized for IB-RDMA transfers
  - Delaying building TFs increases pressure on FLP buffers
  - FLP DD buffers are 32 GB
  - At best ~3 seconds of pb-pb data

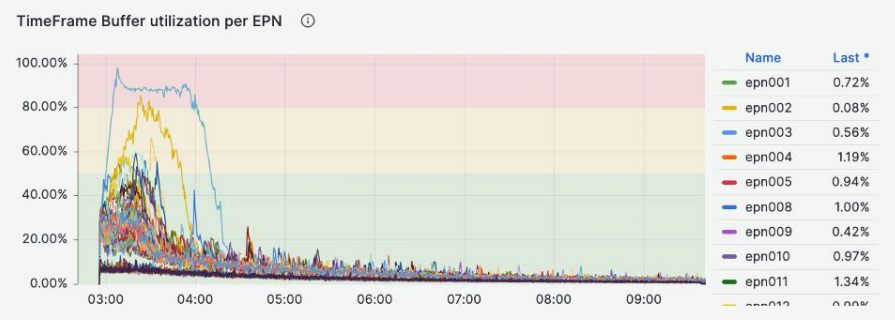


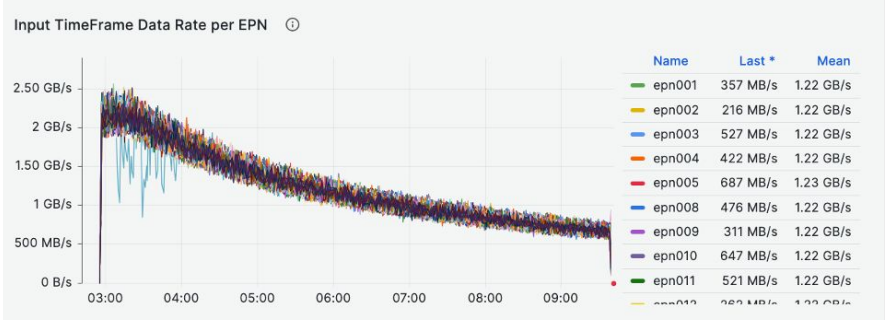
Sharp edge: *Data rate on FLPs must stay below the bandwidth limits of the IB card (100 G/s). This limits number of readout cards per FLP.*

Sharp edge: *Components creating too much jitter in StfSender. E.g. a QC task at high data rates. Reduces window of synchronization. Increases likelihood of timing our with incomplete timeframes*

TfBuilder

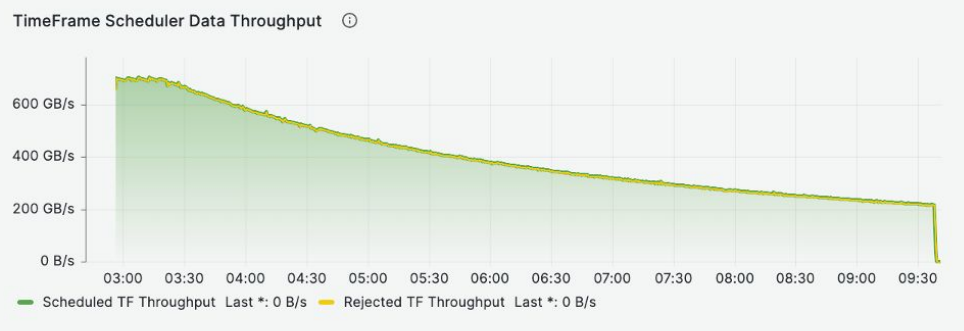
- Runs on each EPN
- Dedicated SHM buffer of 112GB (on nodes with 512 GB RAM) for TF building
  - Buffer utilization is the main feedback to indicate the resource utilization
- Each TF fits entirely into GPU memory
  - With 128 orbits per STF/TF the limit was 32 GB per TF
  - No RDMA all the way to GPU memory
- TfBuilders (EPNs) can join or leave during the run
  - Number of EPNs can be adjusted according to load of EPNs during the run
  - Never implemented in the rest of the O2 framework
- TfBuilders requests the STFs with the same ID from all StfSenders via RDMA





TfScheduler

- TfScheduler schedules each TF to new TfBuilder once complete TF metadata is received from STFSenders
- Round robin scheduling
  - Balanced scheduling in the backlog
- Allows each TFBuilder to aggregate seleval TFs in parallel
- Maintains real-time status of DD buffers:
  - Number and size of STFs in StfSender on FLP (in sending, or waiting for scheduling)
  - Utilization of each EPN (number of in-building, built TFs, and DD buffer utilization)
  - Has global view of “available” STFs on each FLP

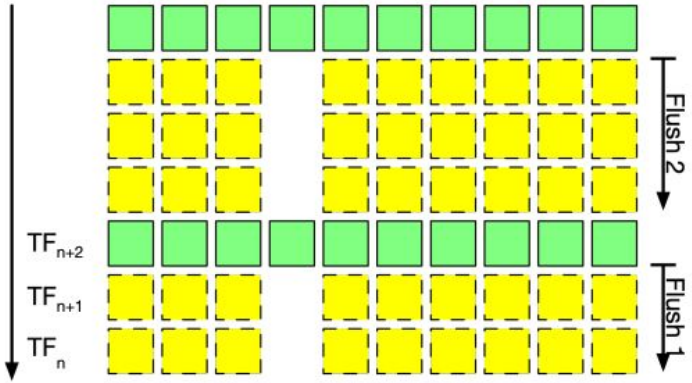


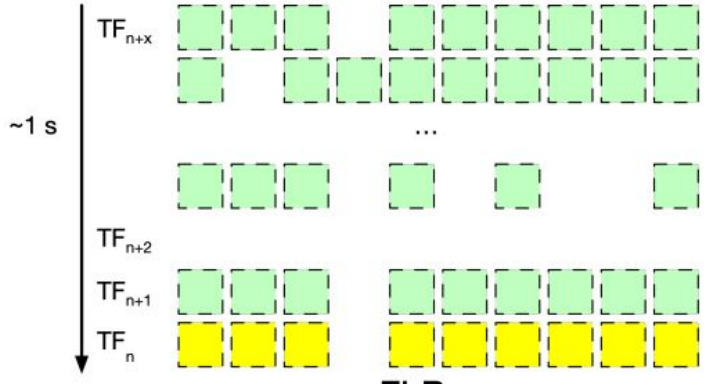
Sharp edge: *Round robin TF scheduling*

- *Unbalanced buffer utilization of MI50 and MI100 EPNs (these are ~33% faster)*
- *MI50 nodes fill up first→ Shifters panic!*

Incomplete TimeFrames

- **on-flush**: incomplete-TFs built because a TF with a higher ID is completed
  - There cannot be more data coming for incomplete TFs
- **on-timeout**: TFs built on timeout or other defined parameters
  - Build when last STF update of TFn was before predefined timeout (e.g. 1s)
  - Or, max number of incomplete TFs is exceeded (e.g. 100)
  - Or, buffer utilization on a FLP go over a threshold
- All incomplete TFs are either built or discarded based on a policy





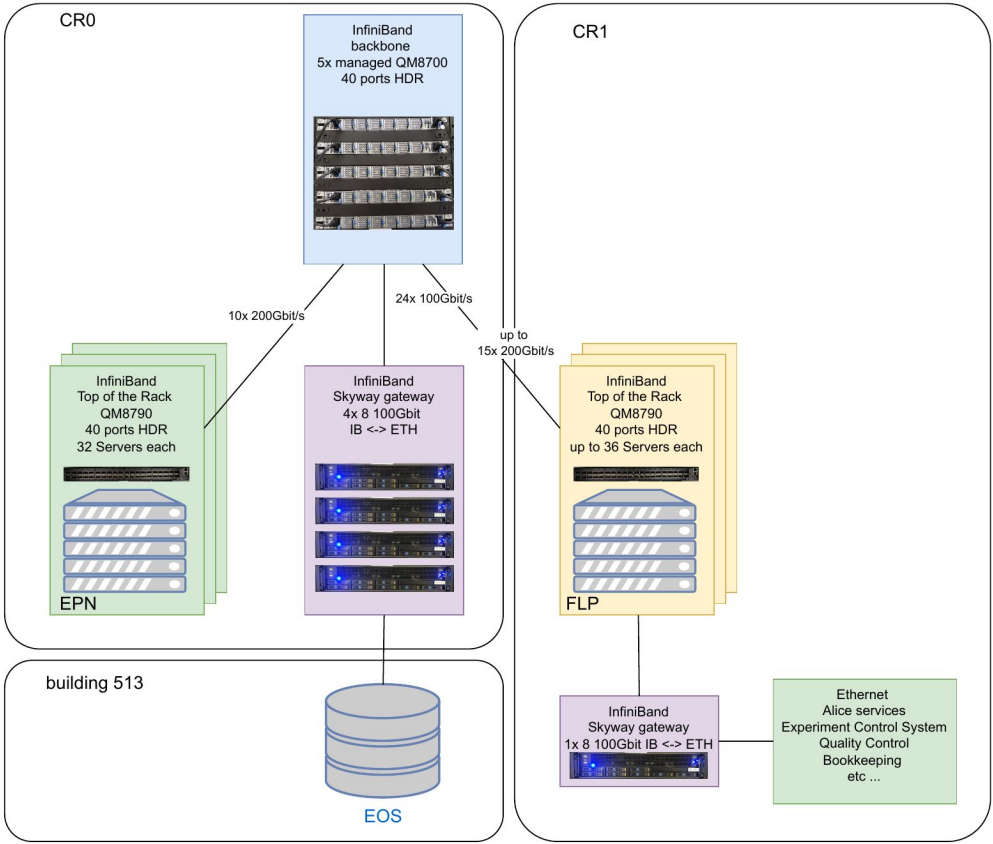
EPN-FLP-EOS Network - Infiniband HDR

FLP/CR1

- 72 links between CR1 and CR0, with a total capacity of 14.4 TBit/s or 1.8 TB/s
- Optical Mode 4 (OM4) multimode fibres
- Short range fibre connections @ 100 m

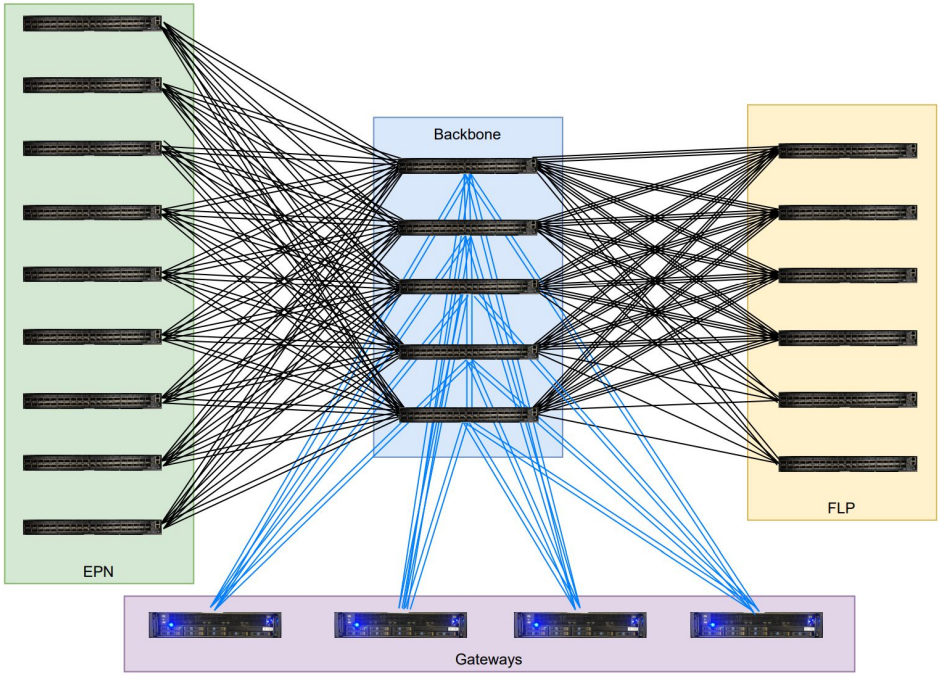
EOS

- ~6km distance to 513
- EPN machines do not act as gateways
  - Despite having both IB and ETH networks
- 4 Mellanox Skyway IB-ETH gateways
- Capacity 400 GB/s



Core Network Topology

- Core switches in CR0 (EPN farm)
- Fat tree
- Blocking
  - TPC FLPs have a blocking factor of 1.2
  - Other FLPs even more, up to 3.1
- A credit based flow control
  - Congestion management of the FLP to EPN many-to-one transmissions
- Adaptive Routing (AR)
  - Balance the traffic between switch links
  - Adds a bit of latency
- EPN building blocks
  - 1 TOR IB Switch, 1 ETH switch
  - Nodes connected via 200 or 100 GBit cables based on distance to TOR



InfiniBand network. Blue - 100 Gbit/s, black - 200 Gbit/s.

Sharp edge: *SPOFs, lead times, switch and gateway configurations*

Data Rates - During 2024 Pb-Pb

- FLP readout at ~4 TB/s
- FLP to EPN synchronous event building peaking at ~700-800 GB/s
- Scheduled order of 10^7 TFs per run
- Time to process a single TF on the EPNs was ~20s
  - Done in parallel in multiples of 8 on each node
- Had a compute resources margin of 10%
- EPN to EOS peaked at 150 GB/s

Both FLP to EPN and EPN to EOS rates were slightly higher in 2023 due to larger TPC data format

Tested limits

- Last tested synthetic throughput from FLP to EPN was 1.25 TB/s
- EPN to EOS tested successfully at sustained 200 GB/s writes
  - Higher transfer rates have balancing issues at EOS

	TDR 2015 (GB/s)	Update 2019 (GB/s)	Data Taking 2022 (GB/s)
CRU input rate	1095	3500	3500
FLP to EPN rate(total)	500	635	900*
FLP to EPN	400	570	830*
EPN to EOS	90	100	130

For reference: TDR data rate expectations