ALICE

A JOURNEY OF DISCOVERY

ALICE HLT Run2 performance overview

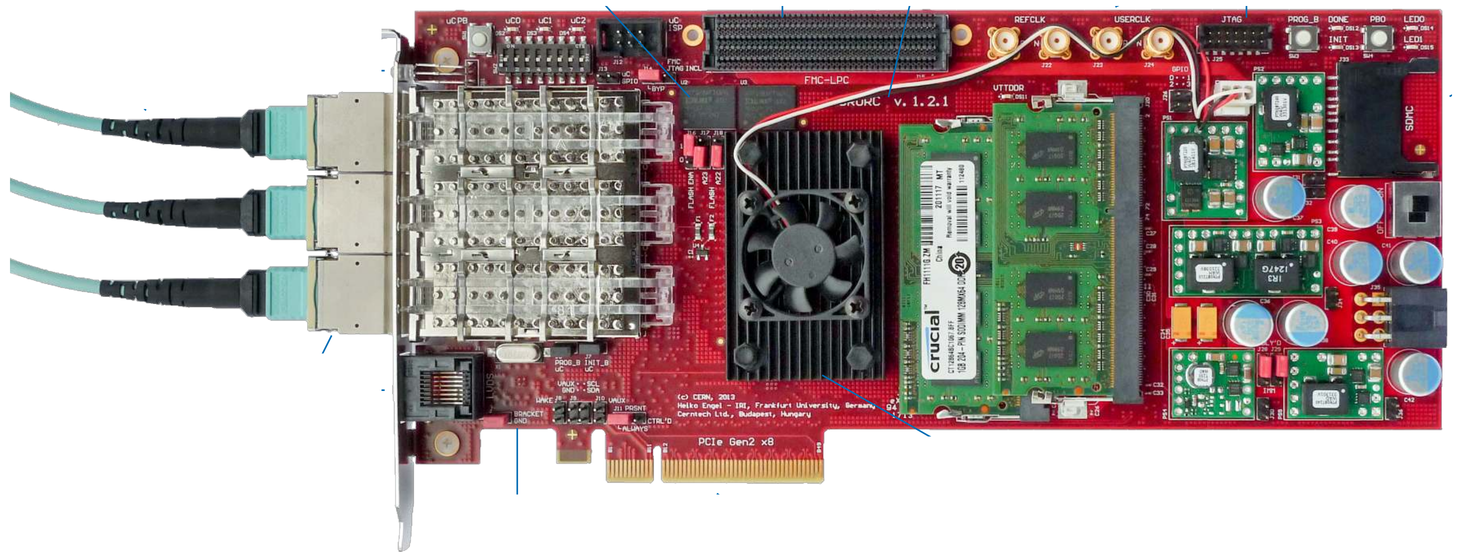
M.Krzewicki for the ALICE collaboration

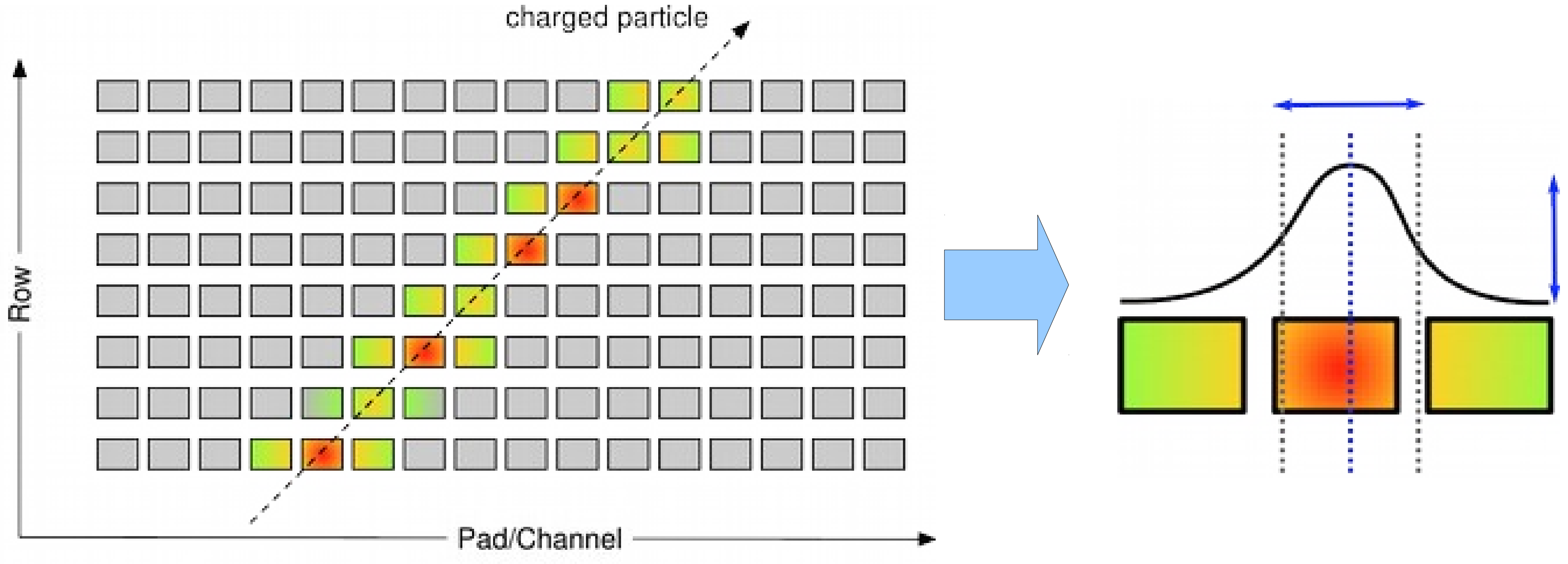
FIAS Frankfurt Institute for Advanced Studies

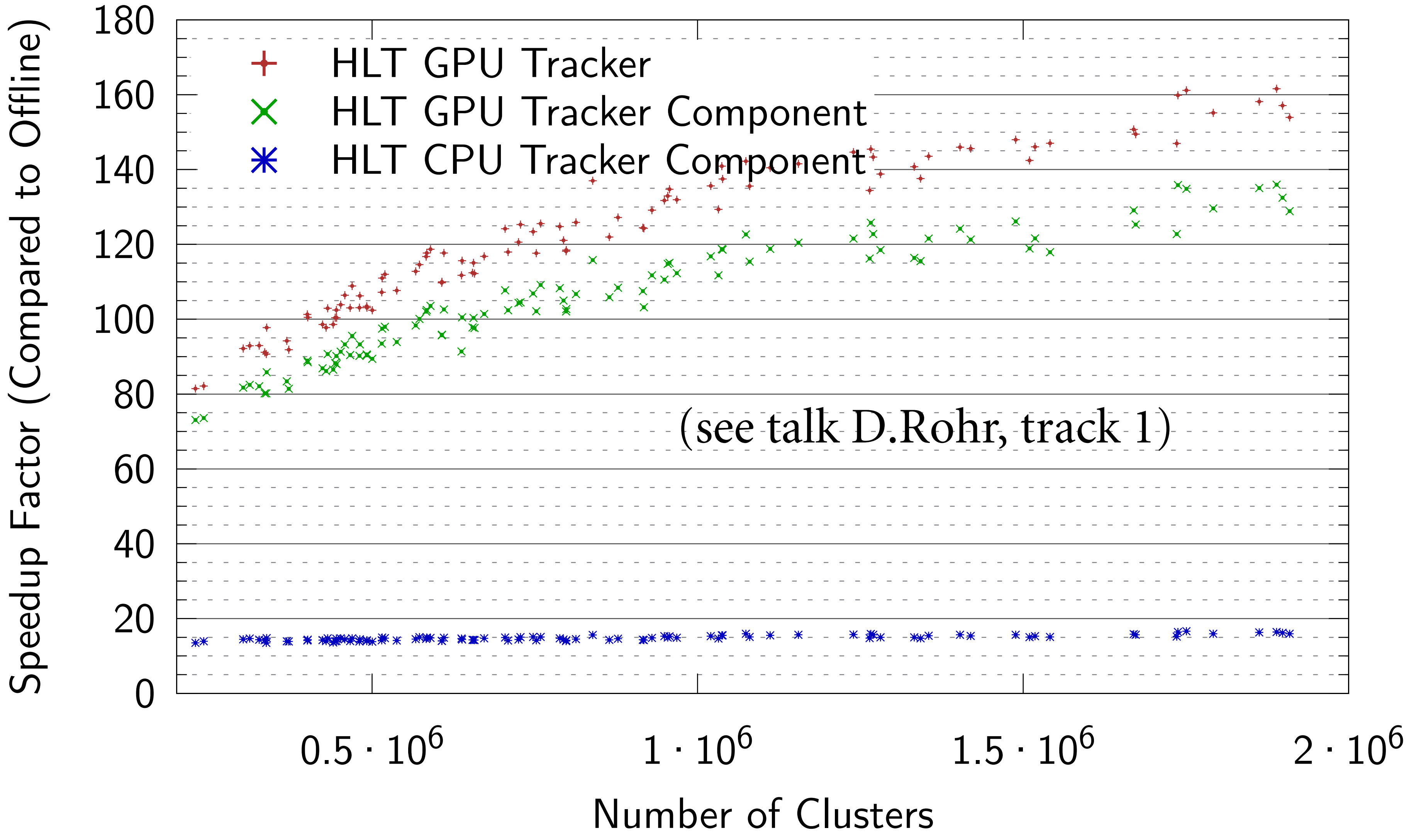
GOETHE UNIVERSITÄT FRANKFURT AM MAIN

ALICE High Level Trigger

- Online reconstruction and data compression facility.
- 180 worker nodes, 8640 HT cores.
- Efficiency through use of hardware acceleration.





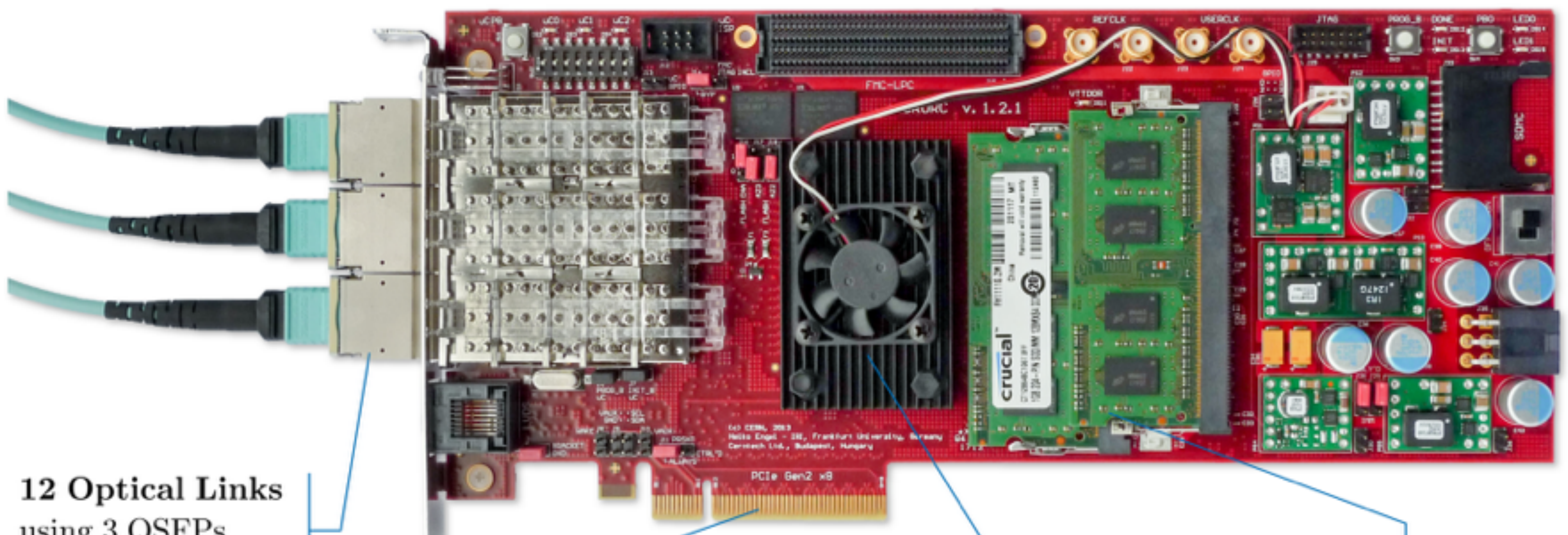


(see talk D.Rohr, track 1)

- FPGA clusterfinder.
  - 1 FPGA board ~ 125 XEON cores.

- GPU tracking.
  - cost savings: $0.5 + 1.0 million.

C-RORC overview



12 Optical Links using 3 QSFPs, connected to FPGA transceivers (GTX). Up to 6.6 Gbps per channel

PCIe Gen2, 8 Lanes 8x 5.0 Gbps, connected to Xilinx PCIe Hard Block

Xilinx Virtex-6 FPGA FF1156 package LX240T

2x DDR3 SO-DIMM Two independent SO-DIMMs up to 1066 Mbps with Xilinx MIG DDR3 softcore

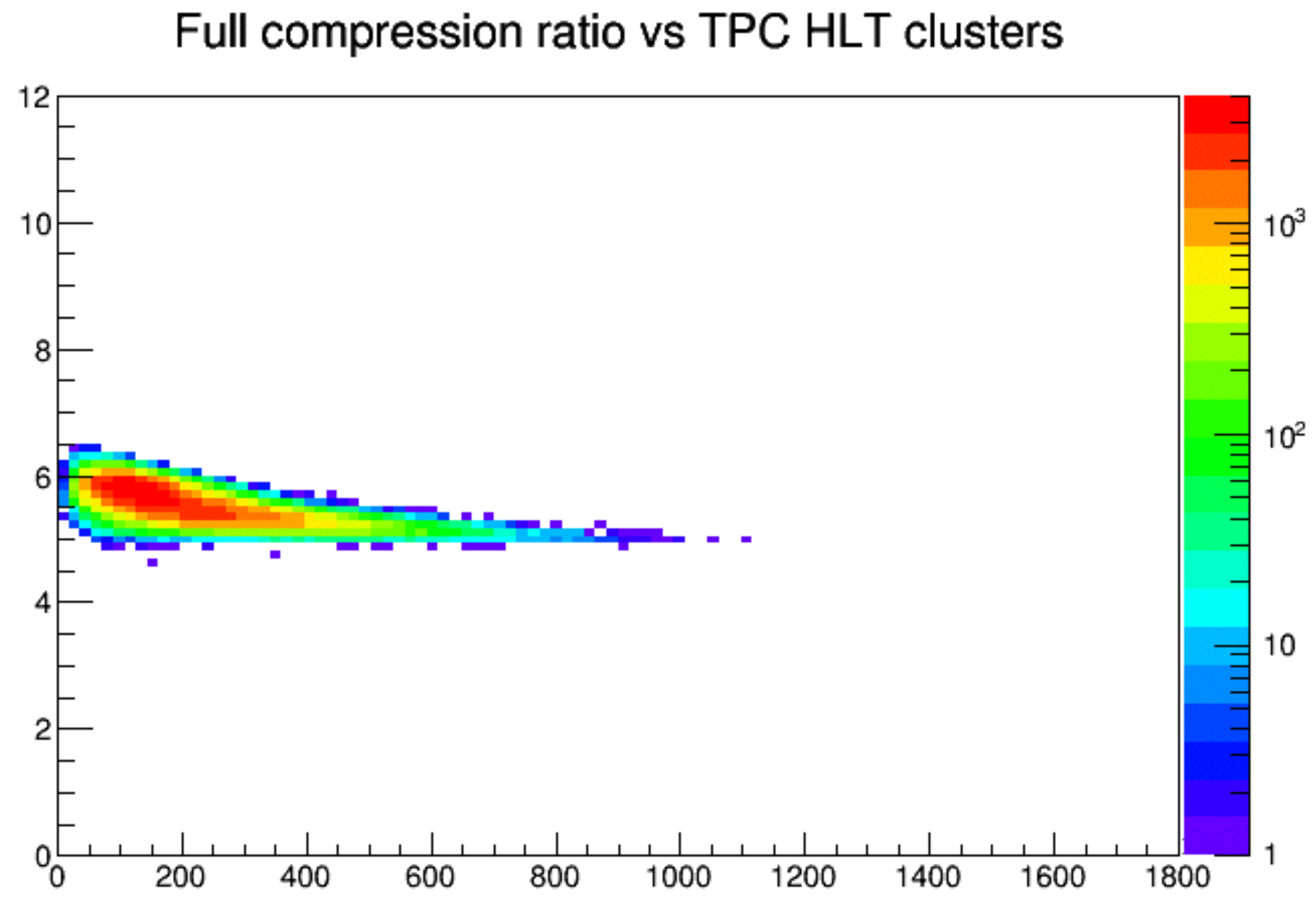
- Data input/output interface:
  @ALICE High-Level Trigger
  @ALICE Data Acquisition
  @ATLAS TDAQ ReadOut System
  @ATLAS Trigger RoI-Builder

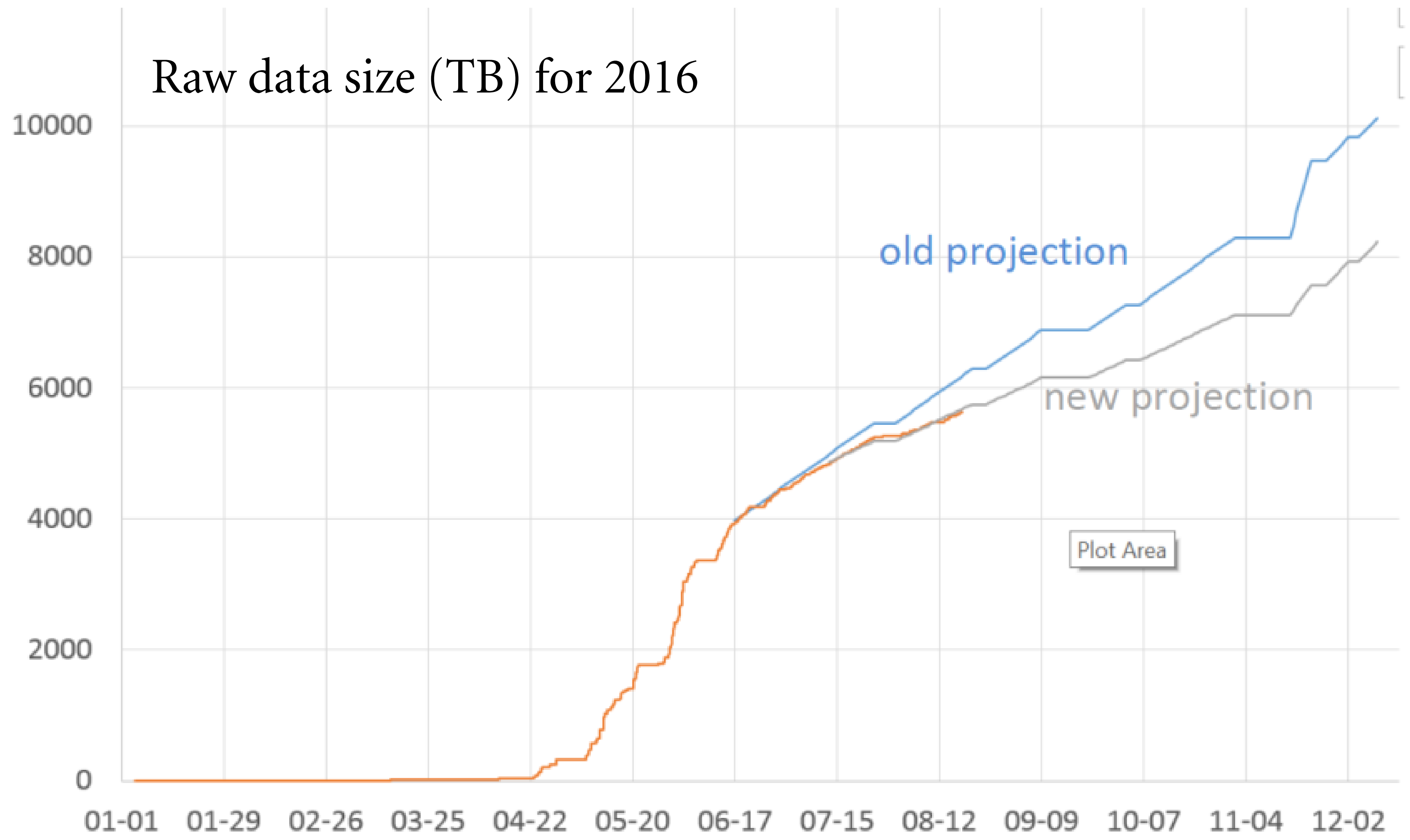
- TPC readout upgraded (RCU2 using DDL2 links - 2X more throughput).
- Firmware updated to comply with RCU2
- RCU1->RCU2: Changes in data layout + bandwidth
  - Over-proportional increase of clock frequency and buffer depth was necessary.
  - Combined RCU1 & RCU2 support.

	#DDLs	Link Speed [Gbps]	DMA Channels	Hardware Preprocessing
HLT_IN	Up to 12	up to 5.3125	12	-
HLT_IN_FCF	6	2.125 or 3.125	12	TPC Cluster finder
HLT_OUT	4	5.3125	4	-

(see talk H.Engel, track 1)

Compression





- Online TPC cluster compression: from factor ~4.3 to ~5.5.
  - Differential Huffman compression, tuned to 2016 data conditions.
  - 20% more efficient raw data storage.
- Ongoing effort, compression studies important for run 3 upgrade.
  - Under study (both run 2 and run 3): track model compression, smarter cluster charge encoding, junk removal. (see talk M.Richter, track 1)

HLT framework optimisations

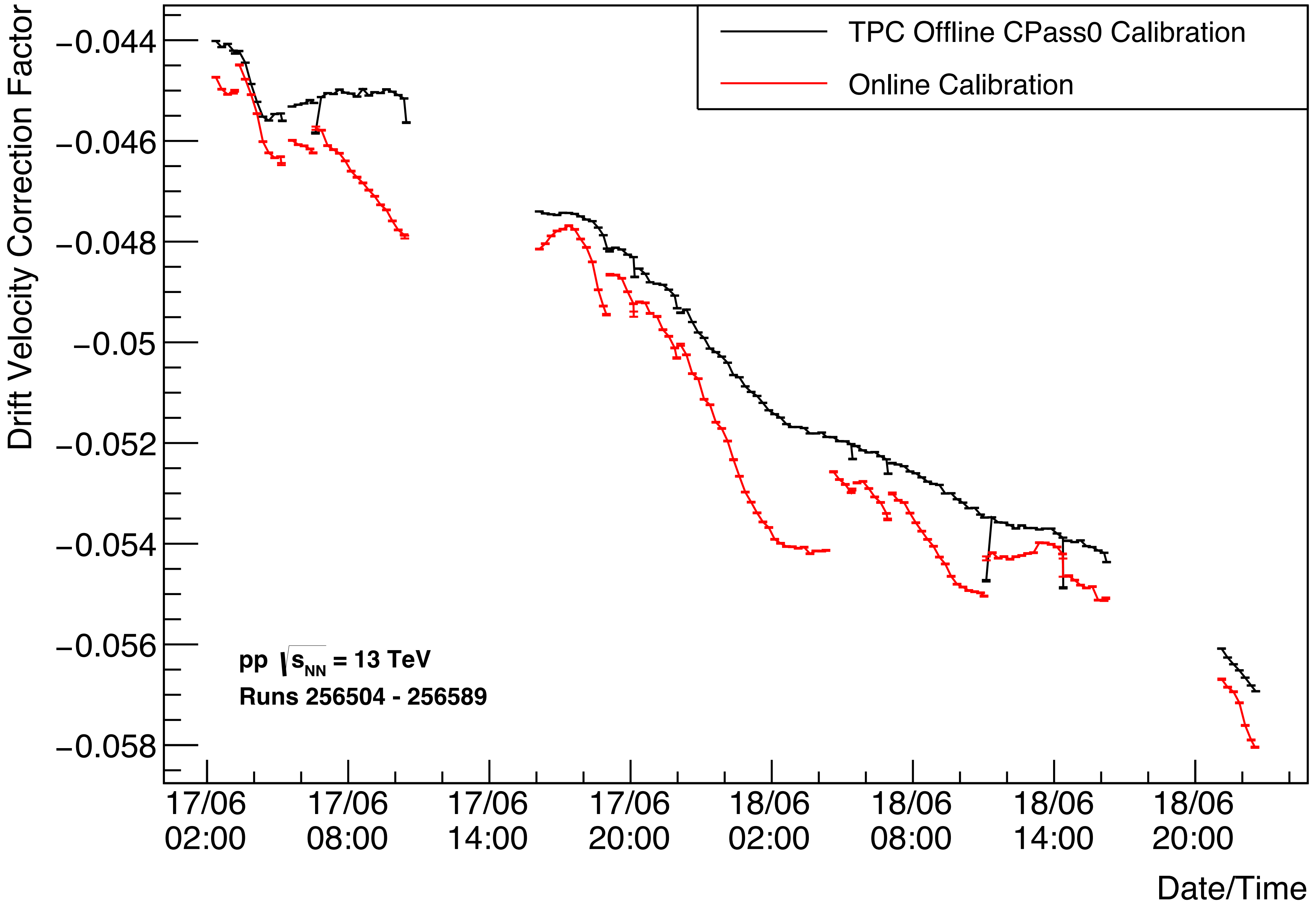
- Optimised GPU tracking.
- Optimised data flow.
- Optimised framework IPC: use shared memory + lower polling rates.
- HLT able to handle full DDL2 in/out bandwidth.
  - all planned run 2 triggering scenarios covered.

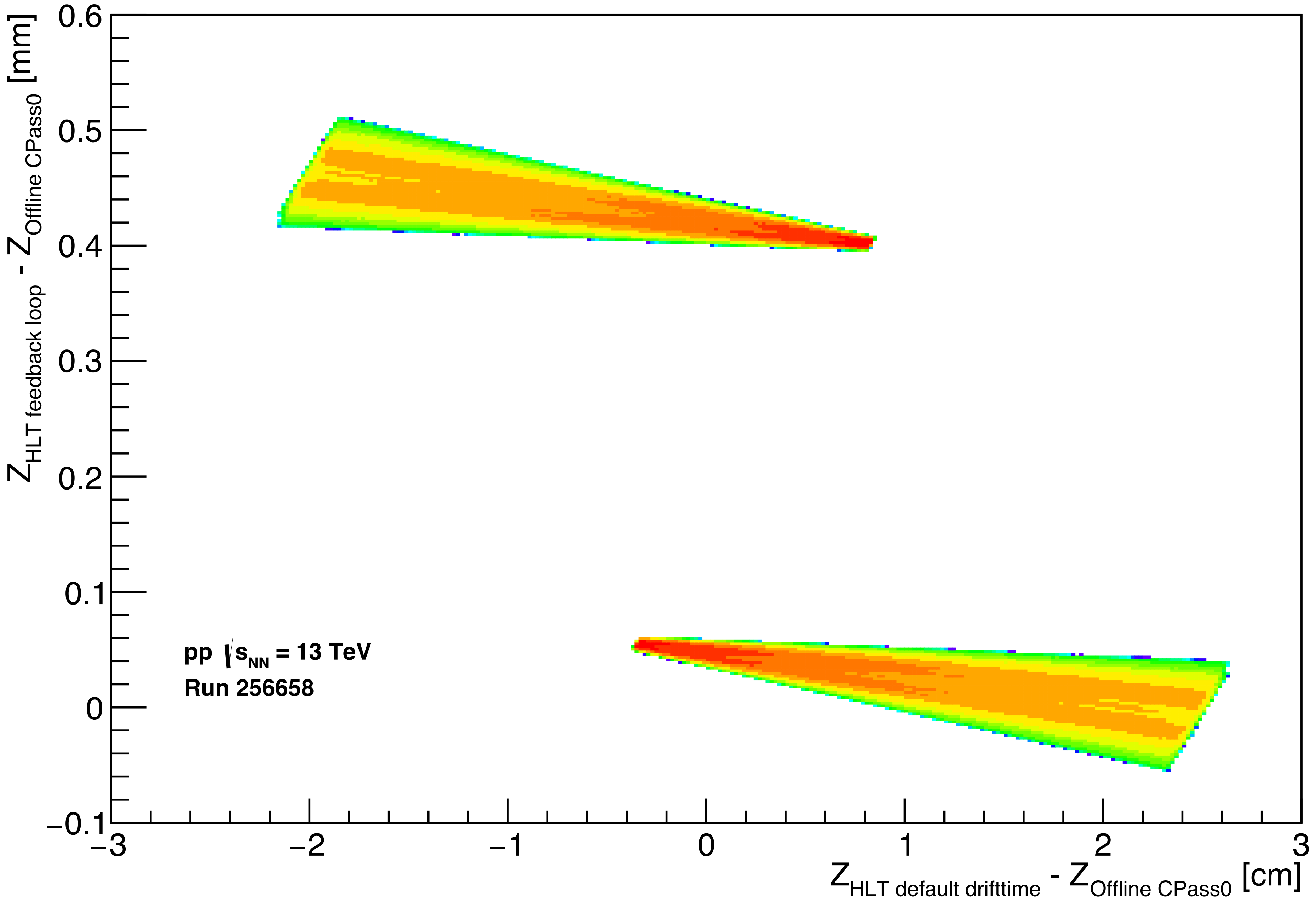
	running reconstruction (all events)	max rate	bottleneck
pp (22 interacting bunches)	TPC, ITS, EMCAL, V0, ZDC	4.5 kHz	CPU
pp (1495 interacting bunches)	TPC, ITS, EMCAL, V0, ZDC	2.4 kHz	RCU2 bandwidth
PbPb	TPC, ITS, EMCAL, V0, ZDC	0.95 kHz	RCU2 bandwidth
PbPb	ITS, EMCAL, V0, ZDC	6.0 kHz	Event merger
PbPb	TPC only, no framework overhead	2.5 kHz	CPU/GPU

- before: the limit was 500 Hz for highest luminosity PbPb and 3 kHz without TPC reconstruction.

(see talk D.Rohr, track 1)

Online calibration

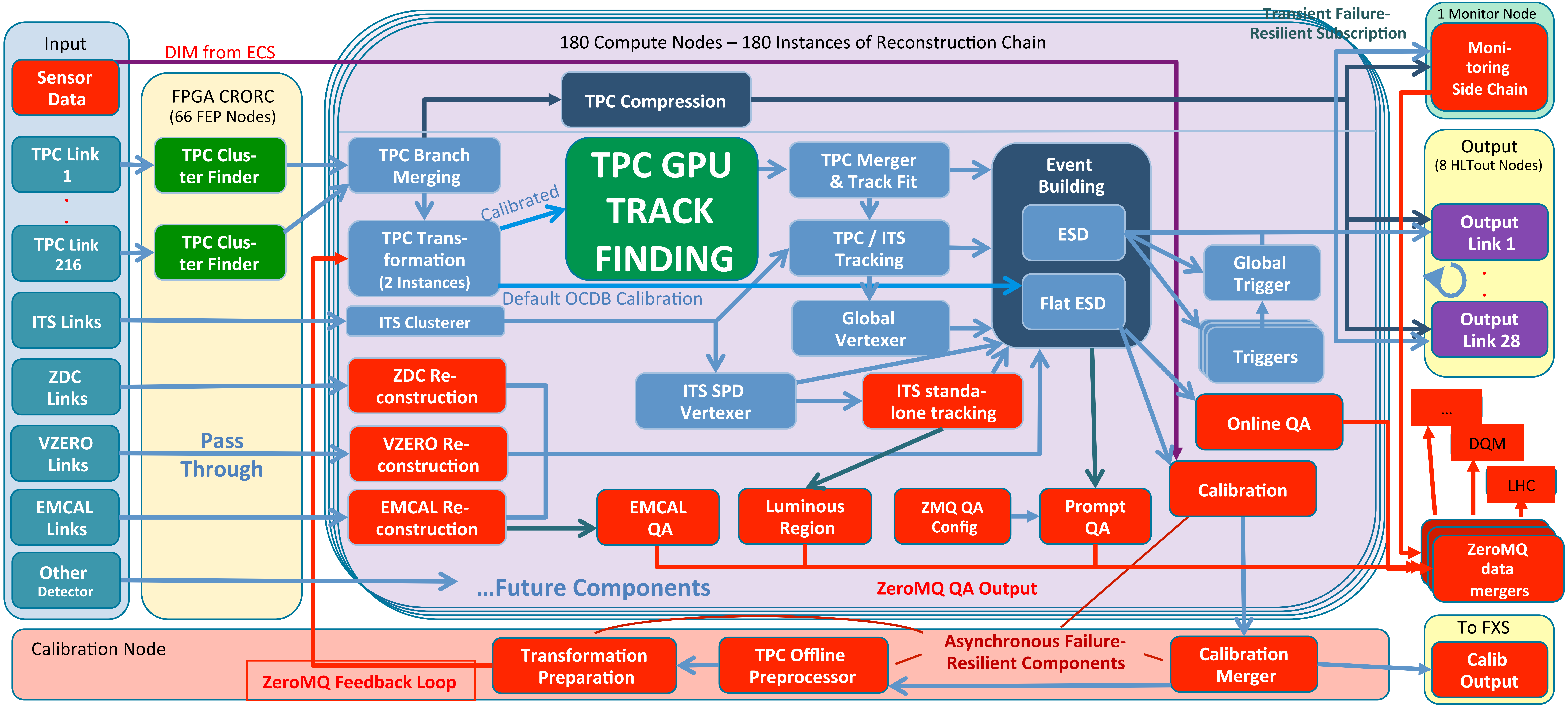




- Offline calibration code was adapted to run both online and offline using the new HLT analysis manager framework.
  - Runs asynchronously, does not slow down the data processing, failure resistant.
- Calibrated cluster positions compatible with offline calibration (to 0.5mm, within resolution).
- The performance of this schema is important to Run 3 related developments. Online calibration can, next to being an important exercise for Run 3, reduce the computing workload during the offline calibration and reconstruction cycle already in Run 2.

(see talk MK, track 1)

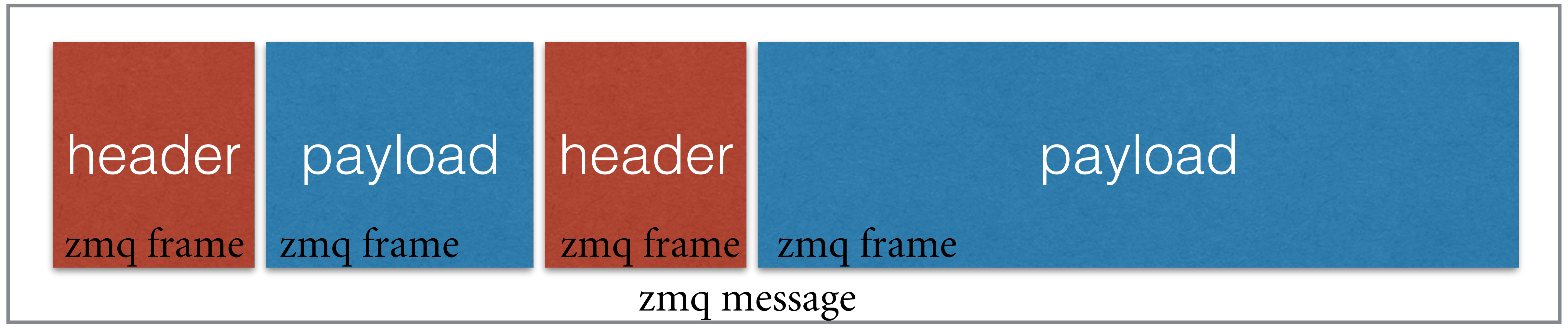
HLT functionality update



- HLT framework supports unidirectional, data synchronous flow only.
- New developments: asynchronous processing, feedback loop, out-of-chain processes (QA, LHC monitoring, etc.).
  - ZeroMQ used as additional transport supplementing the native HLT flow, providing the feedback loop and asynchronous capabilities.

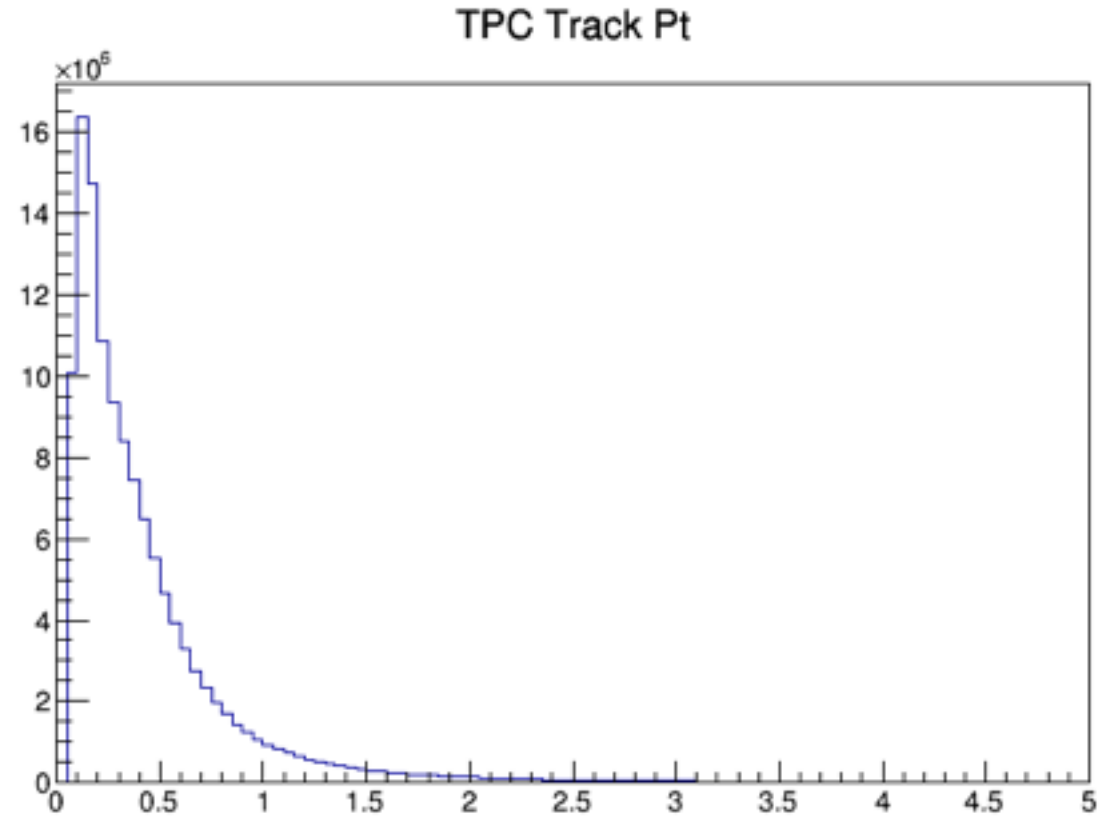
Message based transport

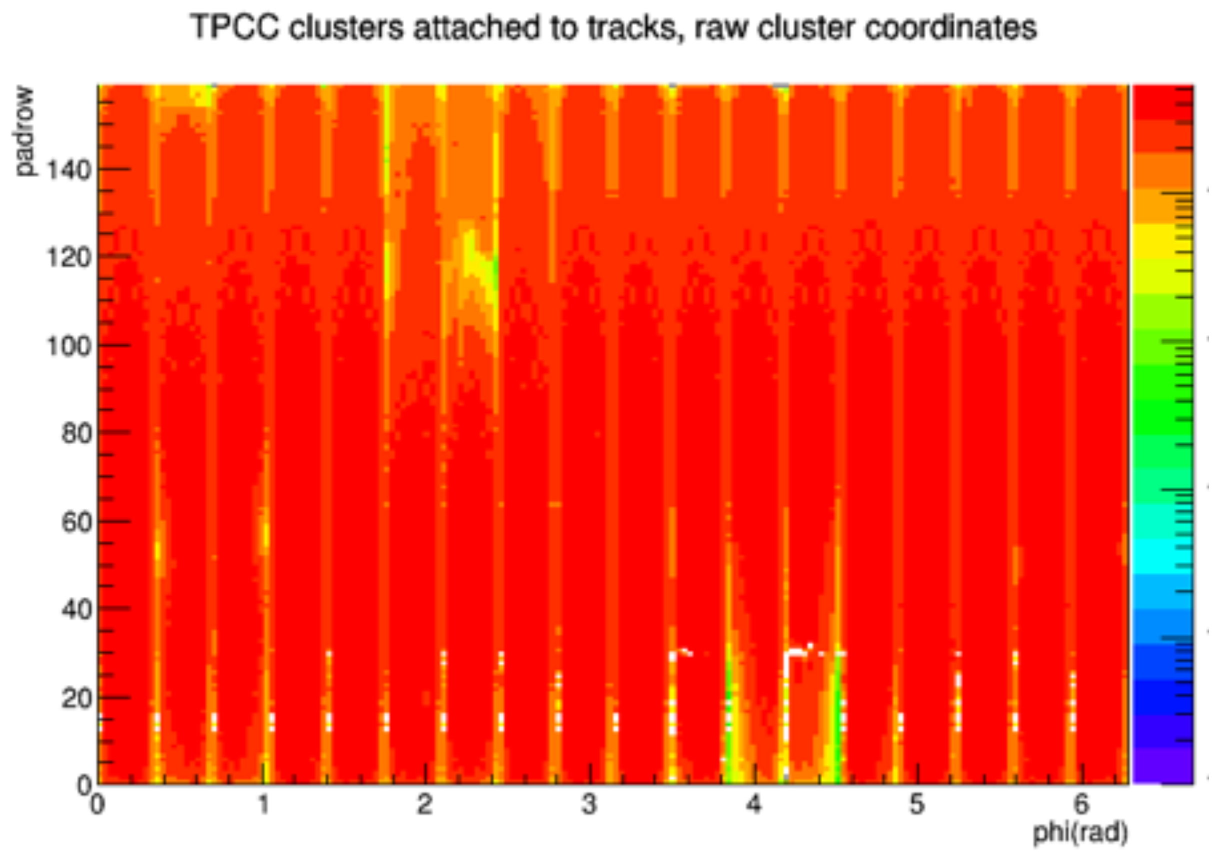
- Based on ZeroMQ multi-part messages.
  - Data and metadata already available in various places in memory (shmem, heap).
  - Keep metadata (header) and payload (e.g. serialised ROOT object) separate, combine many data types in a single zmq message.
  - "Zero" copy - or as close to zero as possible.
  - Easy navigation.

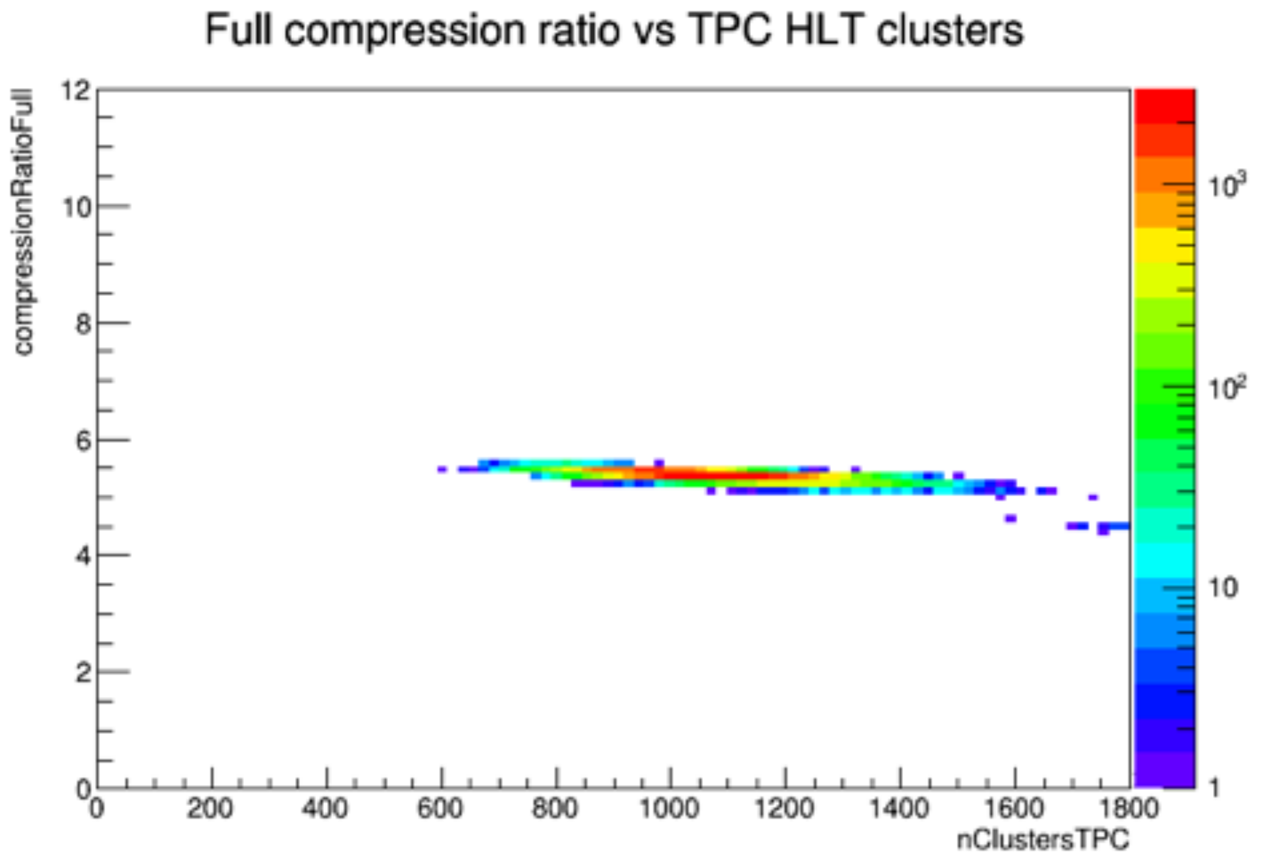


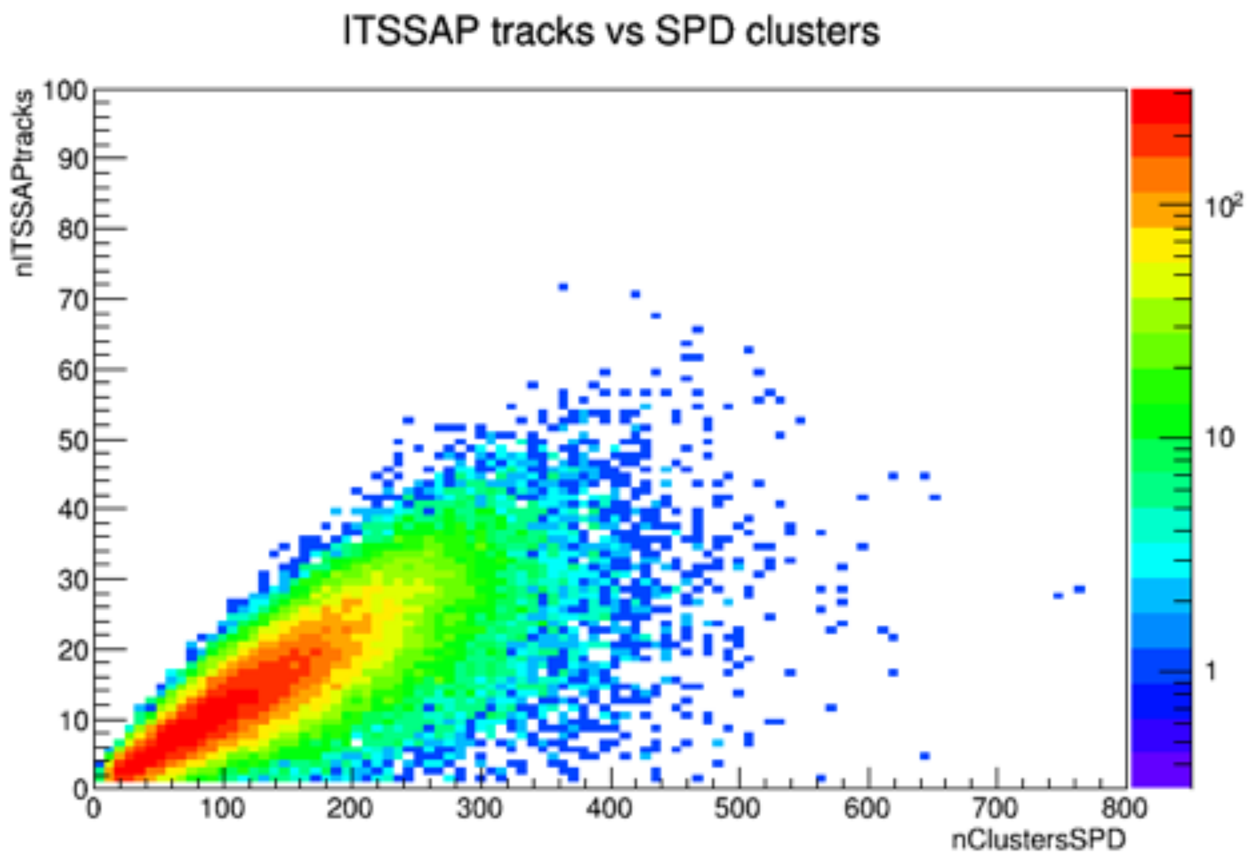
- A test bed for the new data flow model of the O2 system.
- Online calibration feedback loop.
- QA and monitoring framework.

Online physics QA and monitoring

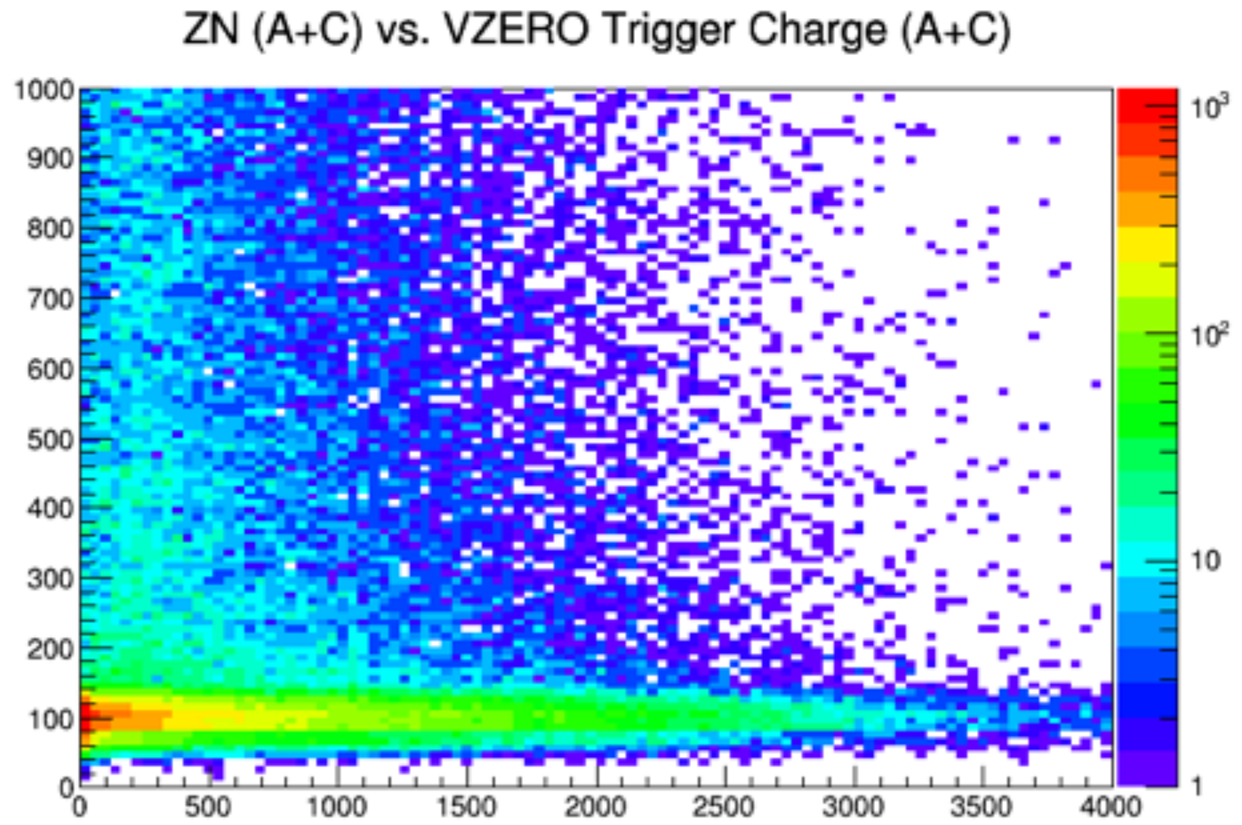

TPC Track Pt


TPCC clusters attached to tracks, raw cluster coordinates

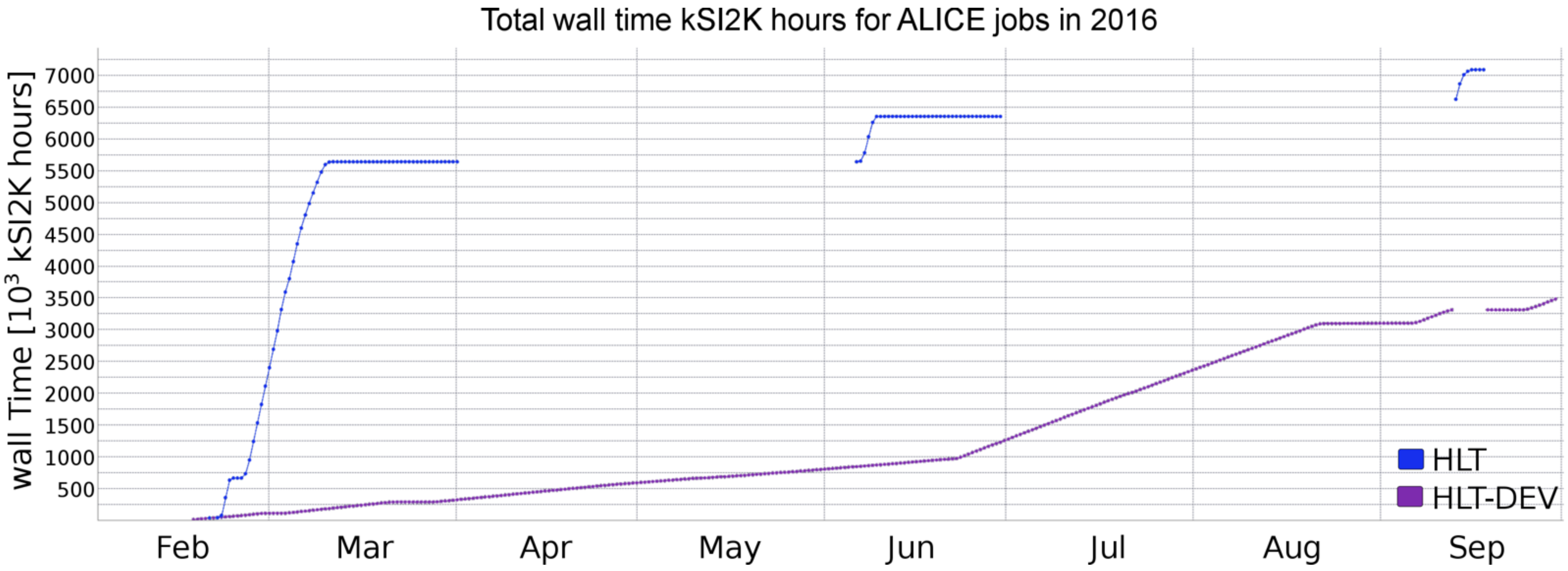

Full compression ratio vs TPC HLT clusters


ITSSAP tracks vs SPD clusters

- Utilising the online reconstruction of many detectors, a new monitoring scheme was developed to allow real-time monitoring of the physics performance of the ALICE detector.
- Includes slow out-of-chain, fast in-chain, and asynchronous running of offline QA and physics analysis code.
- (Re-)configuration on the fly.
- Simple external interface.
  - Data, metadata, ROOT streamers etc. easily added to a single ZeroMQ message efficiently for use e.g. off-site.


ZN (A+C) vs. VZERO Trigger Charge (A+C)

Offline use



- The spare compute resource, the development cluster consisting of older HLT infrastructure is run as a tier-2 GRID site using an Openstack-based setup, contributing as many resources as feasible depending on the data taking conditions.
- In periods of inactivity during shutdowns, also the production cluster is used.
- only MC jobs.
- 650k jobs done, ~2.5% of MC load (as of september).

(see talk J.Lehrbach, track 6)

Summary

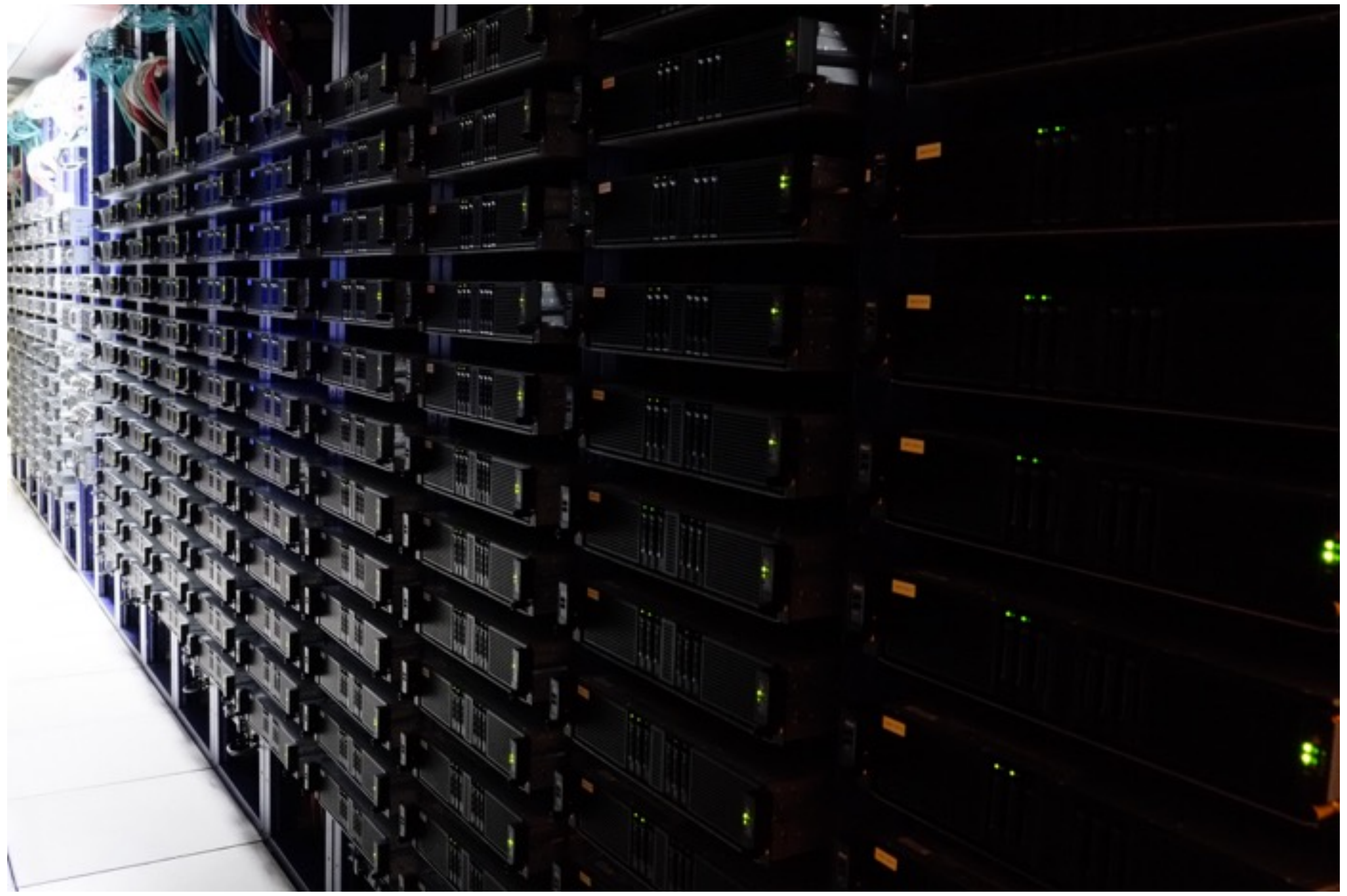
- Intergration with new TPC readout OK.
- Performance improvements to handle all foreseen workloads.
- Data compression improved by 20%.
- Online TPC calibration deployed using the HLT analysis manager framework.
- New online monitoring framework.
- Openstack opportunistic GRID site handling 2.5% of annual MC workload.
- Development goes on, some Run 3 ideas already in Run 2.

thanks

backup

The ALICE High Level Trigger

- 180 nodes - 4320 CPU cores:
  - 2x Intel Xeon E5-2697 CPUs (2.7 GHz, 12 Cores each).
  - 128 GB RAM.
  - 2x 240 GB SSD (used in Raid 1 - Mirroring).
  - 1 AMD FirePro S9000 GPU.
  - 1 C-RORC board (installed in 74 nodes).



- 6+ Infrastructure Nodes:
  - 2x Intel Xeon E5-2690, 3.0 GHz 10 Cores.
  - 128 GB RAM.
  - 2x 240 GB SSD (Raid 1 - mirroring).

- Network:
  - Data: Infiniband in IPoIB Mode (FDR with 56Gb/s, full bisection bandwidth).
  - Management: gigabit ethernet with sideband IPMI - one physical ethernet port per node.
    - 10Gbit backbone.