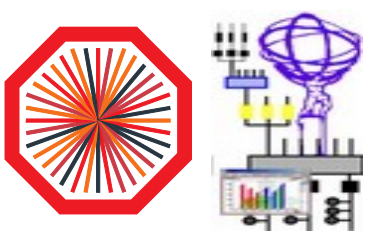


The C-RORC PCIe card and its application in the ALICE and ATLAS experiments

Heiko Engel
(IRI, Frankfurt University)
on behalf of

ALICE DAQ & HLT Team
ATLAS TDAQ Read-Out Team

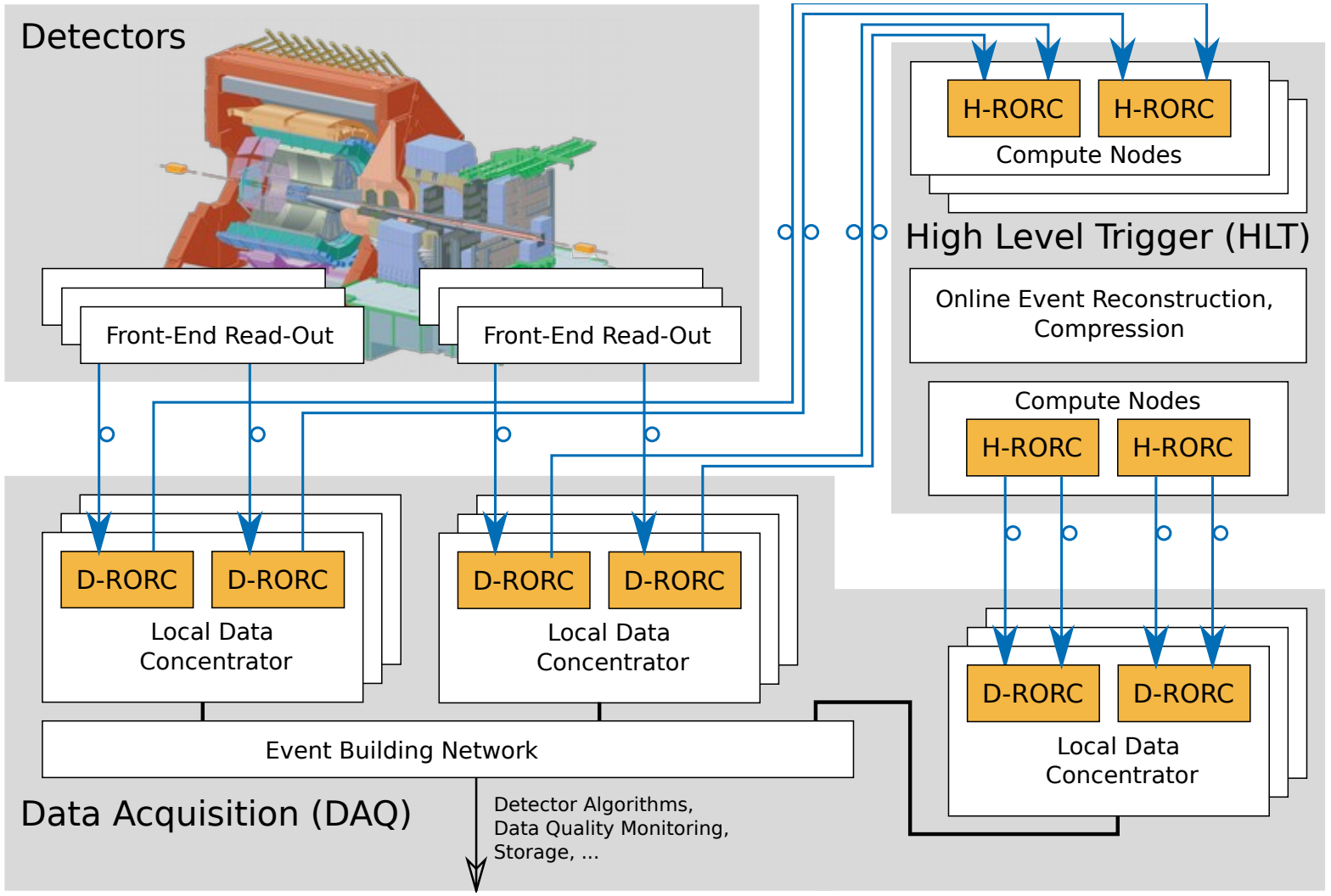
A. Borga (Nikhef), F. Costa (CERN), G. Crone (UCL), H. Engel (Frankfurt), D. Eschweiler (FIAS), D. Francis (CERN), B. Green (RHUL), M. Joos (CERN), U. Kebschull (Frankfurt), T. Kiss (MTA), A. Kugel (Heidelberg), W. Panduro Vazquez (RHUL), C. Soos (CERN), P. Teixeira-Dias (RHUL), L. Tremblet (CERN), P. Vande Vyvre (CERN), W. Vandelli (CERN), J. Vermeulen (Nikhef), P. Werner (CERN), F. Wickens (Rutherford Lab.)

TWEPP 2014 – Aix En Provence
25.09.2014

Outline

- Motivation
- Hardware platform
  - Hardware overview & performance
  - Common production: ALICE and ATLAS
- Applications
  - ALICE Data Acquisition
  - ALICE High-Level Trigger
  - ATLAS TDAQ Read-Out System

ALICE Online Read-Out Architecture



LHC Run1
- ~500 optical links from detectors

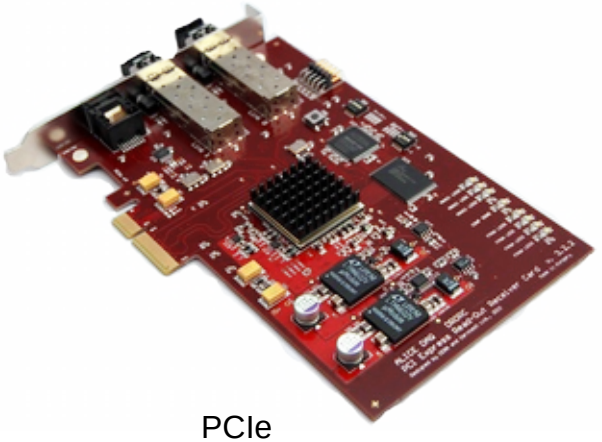
Data Acquisition:
- ~2500 CPU Cores
- ~400 D-RORCs

High-Level Trigger
- ~2500 CPU cores
- 240 H-RORCs
- 64 GPUs

Orange boxes: custom FPGA boards:
D-RORC: DAQ Read-Out Receiver Card
H-RORC: HLT Read-Out Receiver Card

ALICE Run1 Read-Out Receiver Cards

- ALICE Data Acquisition: D-RORCs

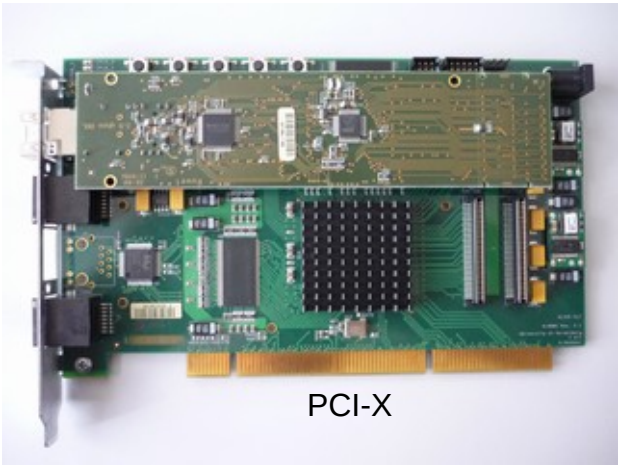


PCIe

Images: cerntech.hu

PCI-X

- ALICE High-Level Trigger: H-RORCs



PCI-X

Motivation

- ALICE: limitations of Run1 Read-Out Receiver Cards
  - host interfaces (PCI-X) become obsolete
  - optical link rate limited
- No suitable commercial replacement available

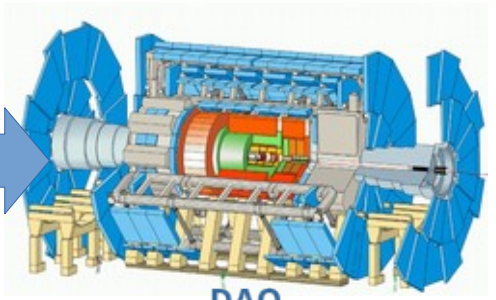
- Custom development
  - primarily for ALICE: common hardware for DAQ and HLT
  - replace Run1 D-RORC / H-RORC boards
  - compatible with Run1 detectors / interfaces / protocols
  - address Run2 upgrade needs for some detectors

- Still a quite general hardware platform

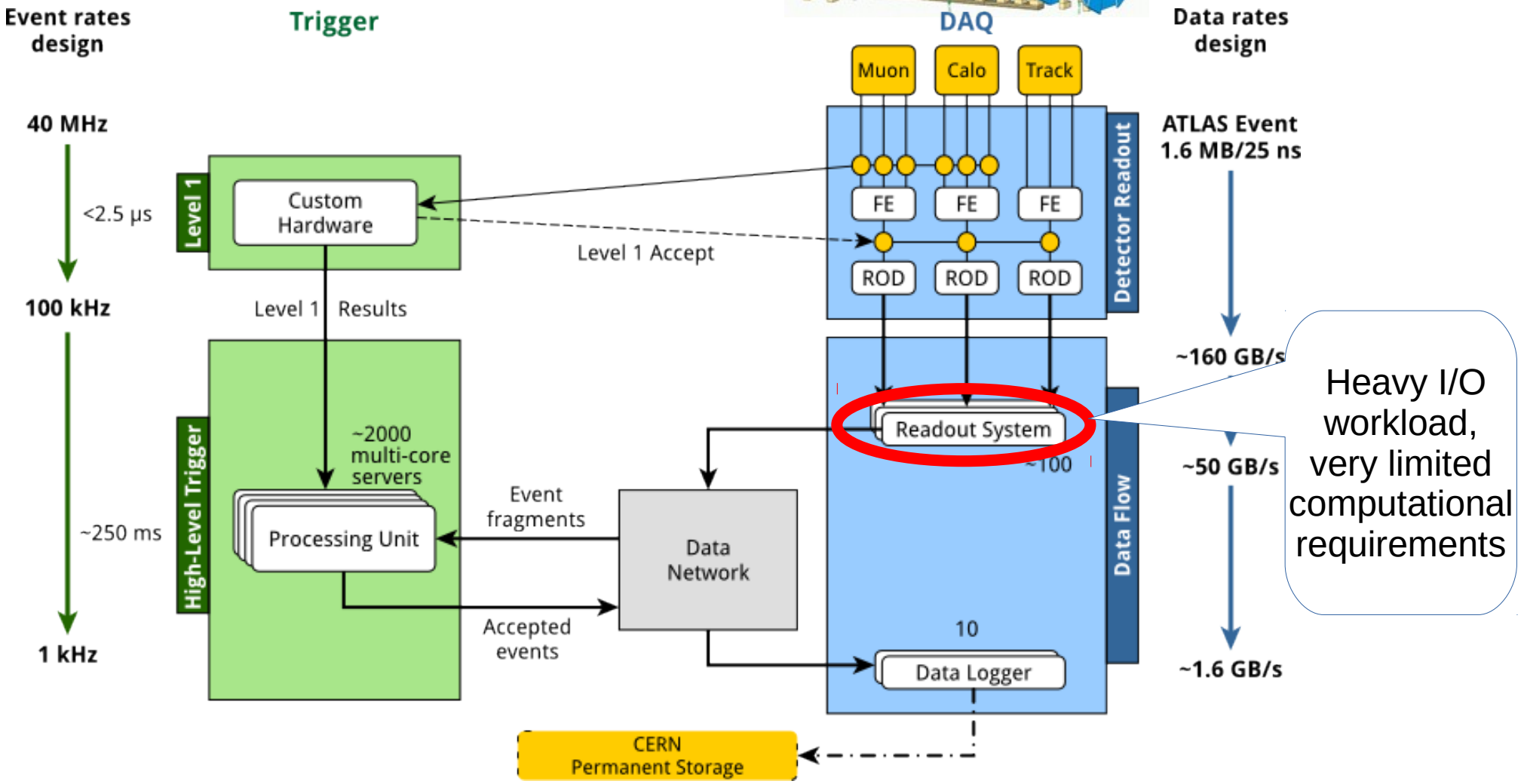
  → ATLAS joined the project:
  replace Robin with C-RORC a.k.a. RobinNP

ATLAS TDAQ in Run2

ATLAS@LHC - CERN
General purpose detector
Wide physics search goals
46m long, 22m high, 7000 tons
140M channels



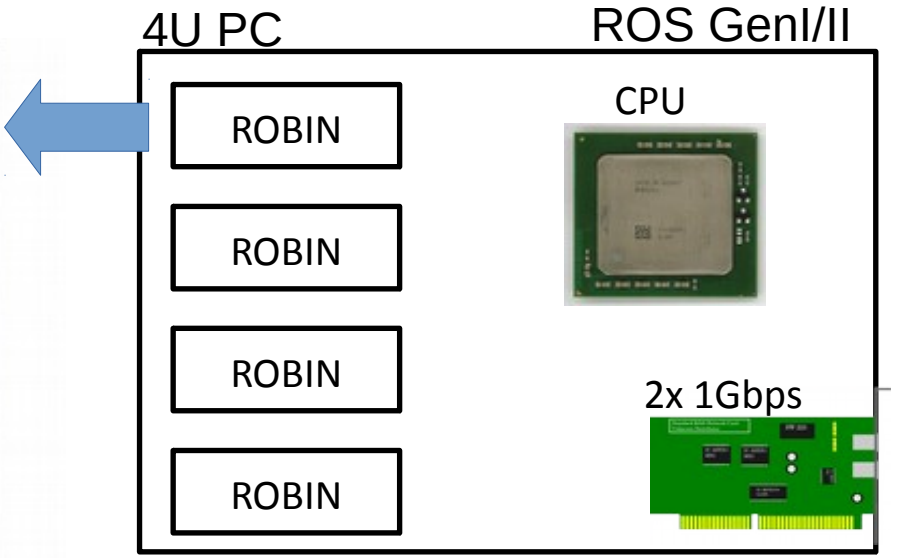
Run1 ReadOut System (ROS) Setup:
~1600 optical inputs
~150 PCs
~300 1GbE ports



Heavy I/O workload, very limited computational requirements

ATLAS ReadOut System in Run1

- 4U PCs equipped with
  - 2x 1GbE ports for data transfer
  - four (five) FPGA-based custom receiver and buffer cards (ROBIN)
- ROBIN
  - PCI interface (2.1 Gbps)
  - 3 optical inputs compatible with S-link (2 Gbps – nominal bandwidth 160 MB/s)
  - 64 MB/link buffer memory
  - on-board PPC processor for data and request management
  - 10-15% readout capability



ATLAS TDAQ evolution & ReadOut System

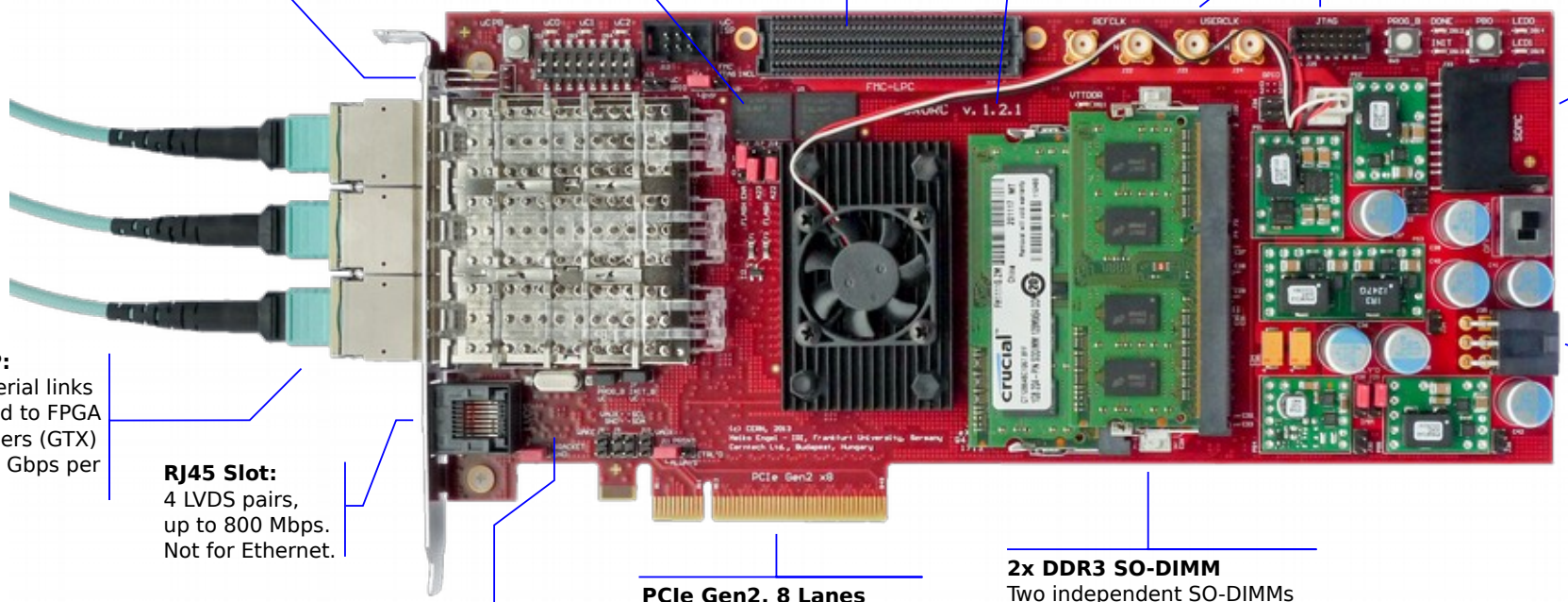
ReadOut System functions remain unchanged in Run2

- Changes on the detector side reflected by increased number of ROLs (+16%) → **denser solution** in terms of ROL/rack space
- Denser solution implies higher throughput per node → **move from 1 Gbps copper to 10 Gbps optical** Ethernet
- PCI is ageing technology, not very common in current COTS → **prefer a PCIe**-based solution
- Size of memory buffer limits the average processing time and ultimately the HLT farm size → **larger buffer memory**
- Higher luminosity and higher first-level trigger rate → **capable of 50% readout**
- Future compatibility with newer generations of faster ROL → **new, faster optical receivers**

ROS upgrade project → GenIII

The C-RORC Hardware Platform

Hardware Overview



Configuration Flash
2x 128Mbit Xilinx PlatformFlash XL

FMC LPC
VITA57.1 FMC LPC Connector

GTX RefClk
Configurable GTX Reference Clock

System Clock
200 MHz

SMA Connectors
for additional GTX and System Clocks

JTAG-Connection
for FPGA Programming Cables

Bi-Color **LED** in IO-Bracket

SDCard Socket

Power Connection
via 6-Pin PCIe GPU Power Connector

3x QSFP:
12 fast serial links connected to FPGA transceivers (GTX)
Up to 6.6 Gbps per channel

RJ45 Slot:
4 LVDS pairs, up to 800 Mbps.
Not for Ethernet.

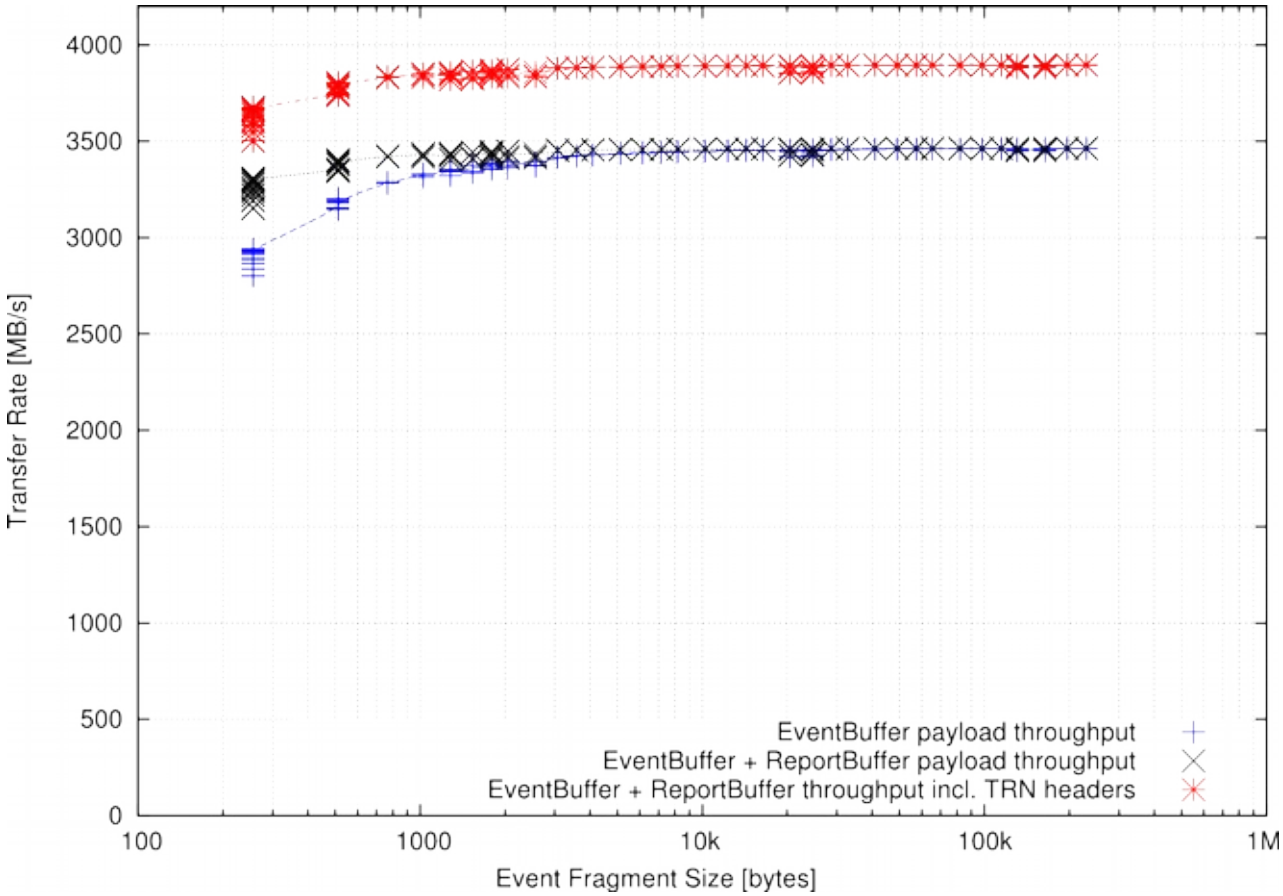
Microcontroller for Configuration monitoring and connection to Host via PCIe SMBus Lines.

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

2x DDR3 SO-DIMM
Two independent SO-DIMMs
Up to 1066 Mbps with Xilinx MIG DDR3 Softcore

C-RORC PCIe DMA Performance

- DMA to host RAM
  - PCIe Gen2 x8 with Xilinx Endpoint
  - custom DMA engine, two buffers
  - payload size: 256 byte
  - transaction level protocol overhead: 10-11%
  - very close to theoretical limit



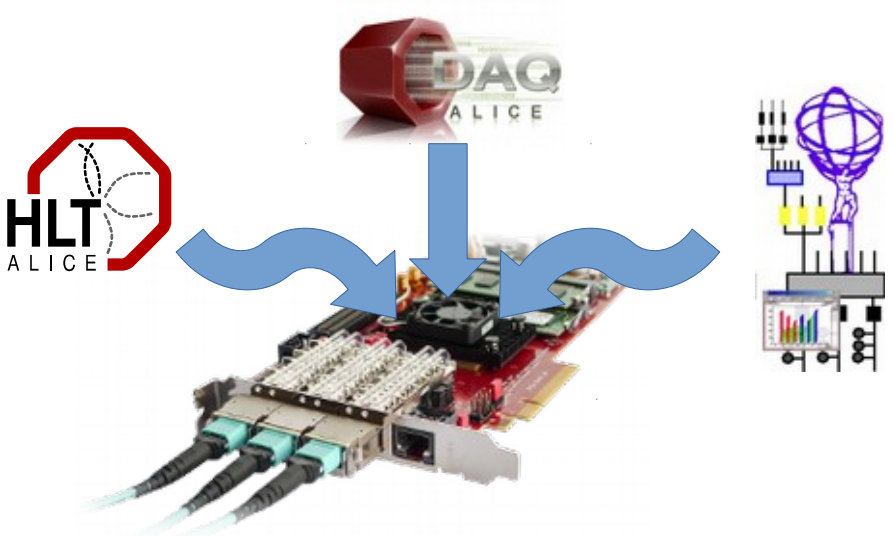
Production

- PCB layout and prototypes by Cerntech, HU
- Common hardware test suite to verify boards
  - PC, software, driver, firmware & documentation
  - tests cover a large subset of all board functionality
- ALICE & ATLAS common large scale production via international tender
  - approx. 370 boards have been produced, tested and delivered
  - installation at CERN ongoing

C-RORC Applications:

Common hardware platform -
still independent firmware development for different applications

ALICE Data Acquisition (DAQ)
ALICE High-Level Trigger (HLT)
ATLAS TDAQ Read-Out System (ROS)



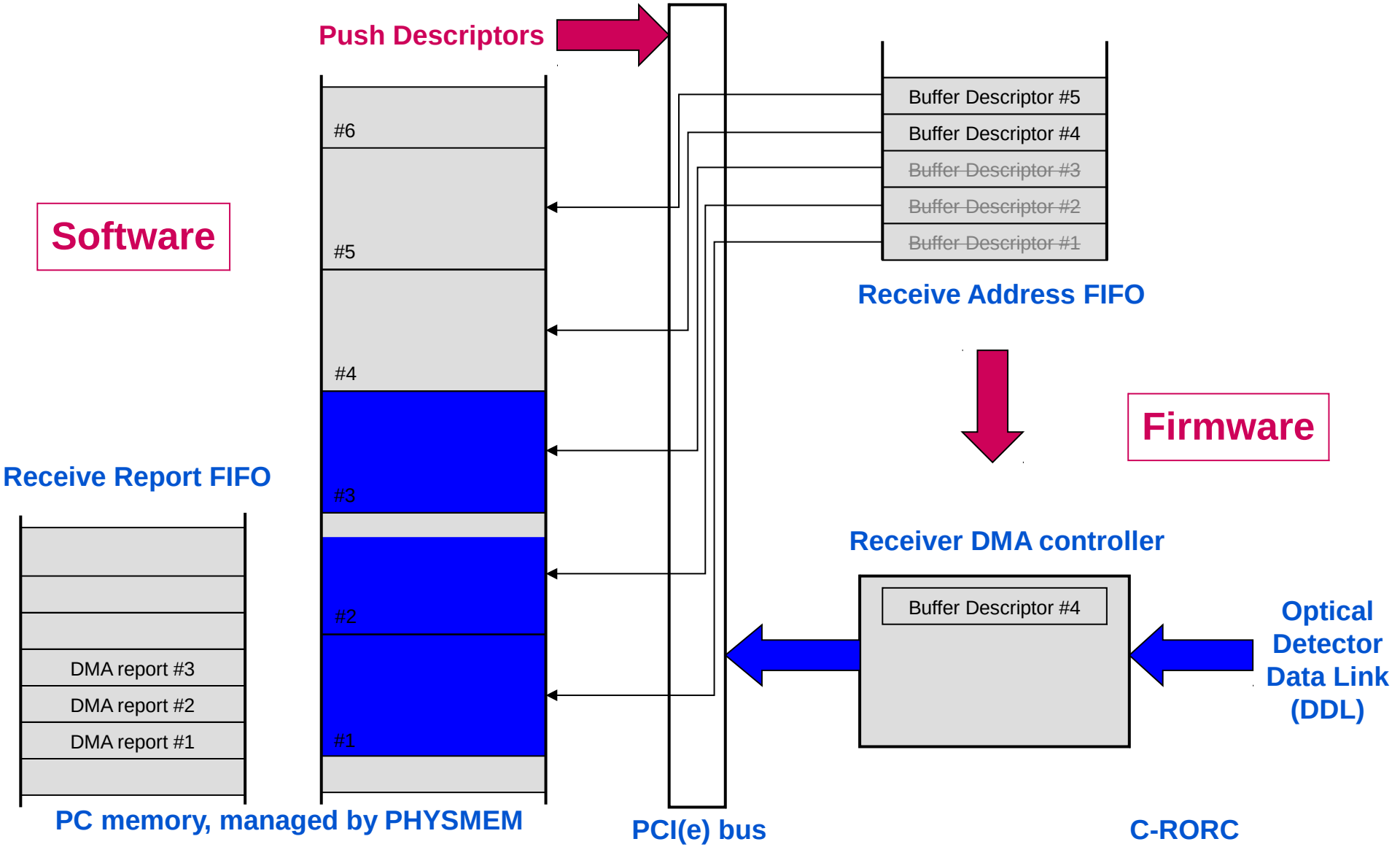
ALICE Data Acquisition Firmware

- Data Acquisition and Detector Configuration/Control protocols
  - DATA TAKING
    - RX channel used to download events from the detector electronics to the DAQ farm
    - TX channel used for flow control to avoid data overflow.
  - DETECTOR configuration:
    - TX channel used to send configuration data to the front end electronics
    - RX channel used to receive acknowledgement
- Replace TPC, TRD and HLT-to-DAQ D-RORCs
- Optics
  - Detector Data Link (DDL) Protocol, Gen1 at 2.125 Gbps
  - ported DDL protocol to higher link rates
    DDLv2 : 4.0 / 4.25 / 5.3125 Gbps
  - 6 links from detectors, 6 links to HLT
- DMA: Device to Host
  - Physmem buffer handling
  - firmware: PLDA DMA engine,
    6 data channels for data taking
    and for detector configuration

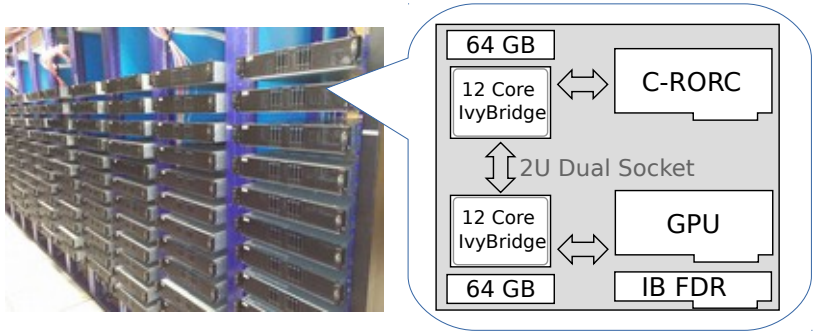
for details, see Poster *"DDL, the ALICE Data Transmission Protocol and its Evolution from 2 to 6 Gb/s"*

Run2 Setup
~130 Nodes / ~2900 cores
59 C-RORCs (36 TPC + 9 TRD + 14 HLT)
10G Ethernet Interconnect

ALICE DAQ C-RORC Dataflow



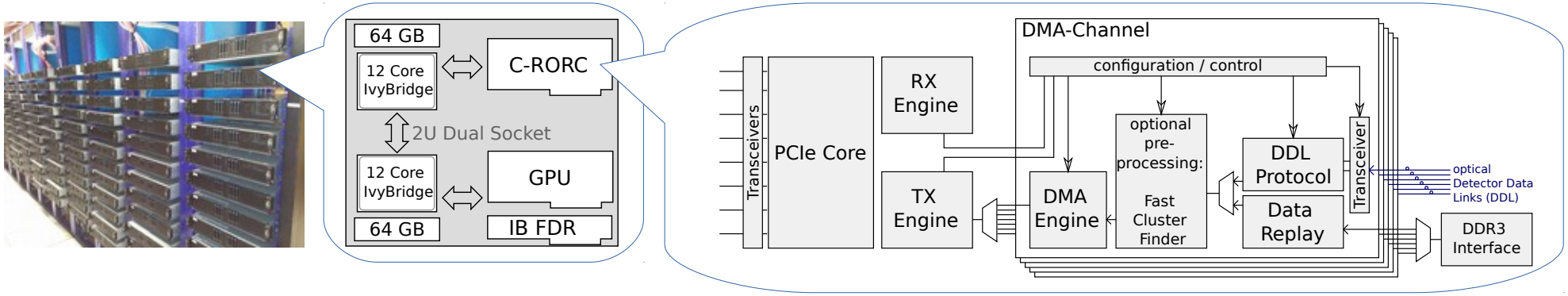
ALICE High-Level Trigger



Run2 Setup
~180 Nodes / ~4300 cores
~180 GPUs
~75 C-RORCs
Infiniband FDRx4 interconnect (56 Gbps)

- High-Level Trigger
  - receive event data from Data Acquisition
  - online pre-processing, reconstruction and compression
  - processed data and decision back to Data Acquisition
- Processing:
  - 1 C-RORC replaces 3-6 of the previous H-RORC boards
  - up to 12 input links per C-RORC
  - local pre-processing in C-RORC FPGA
  - global processing in CPU & GPU

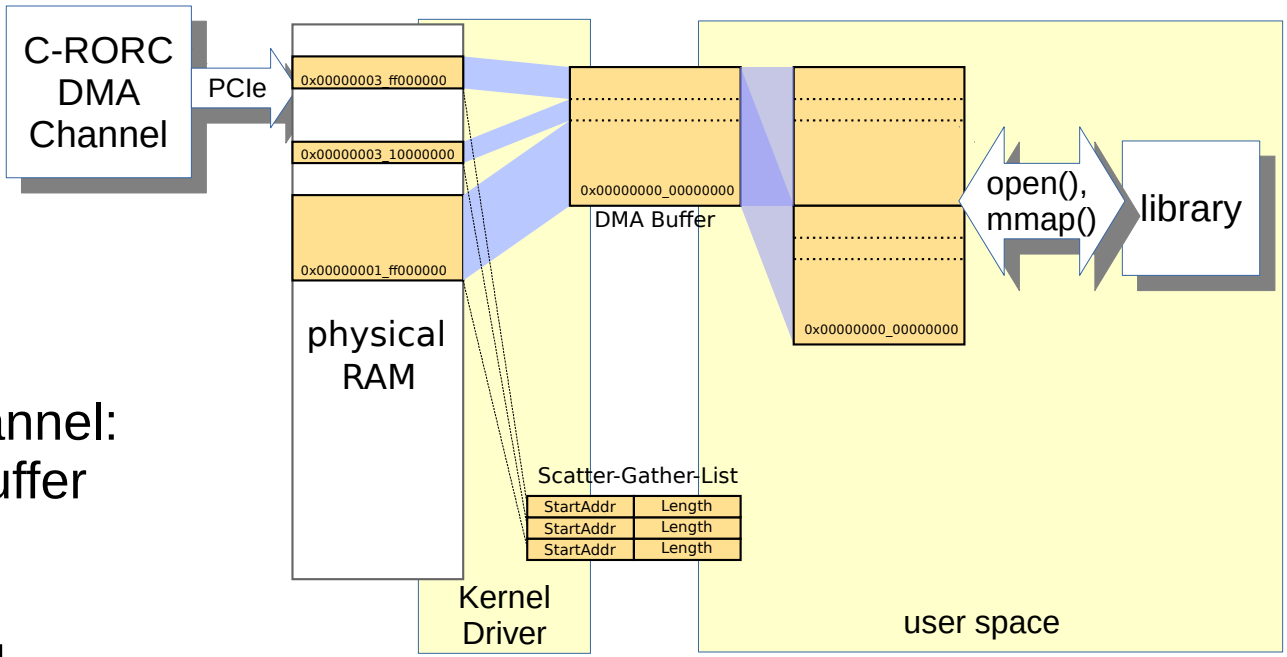
ALICE High-Level Trigger: Firmware



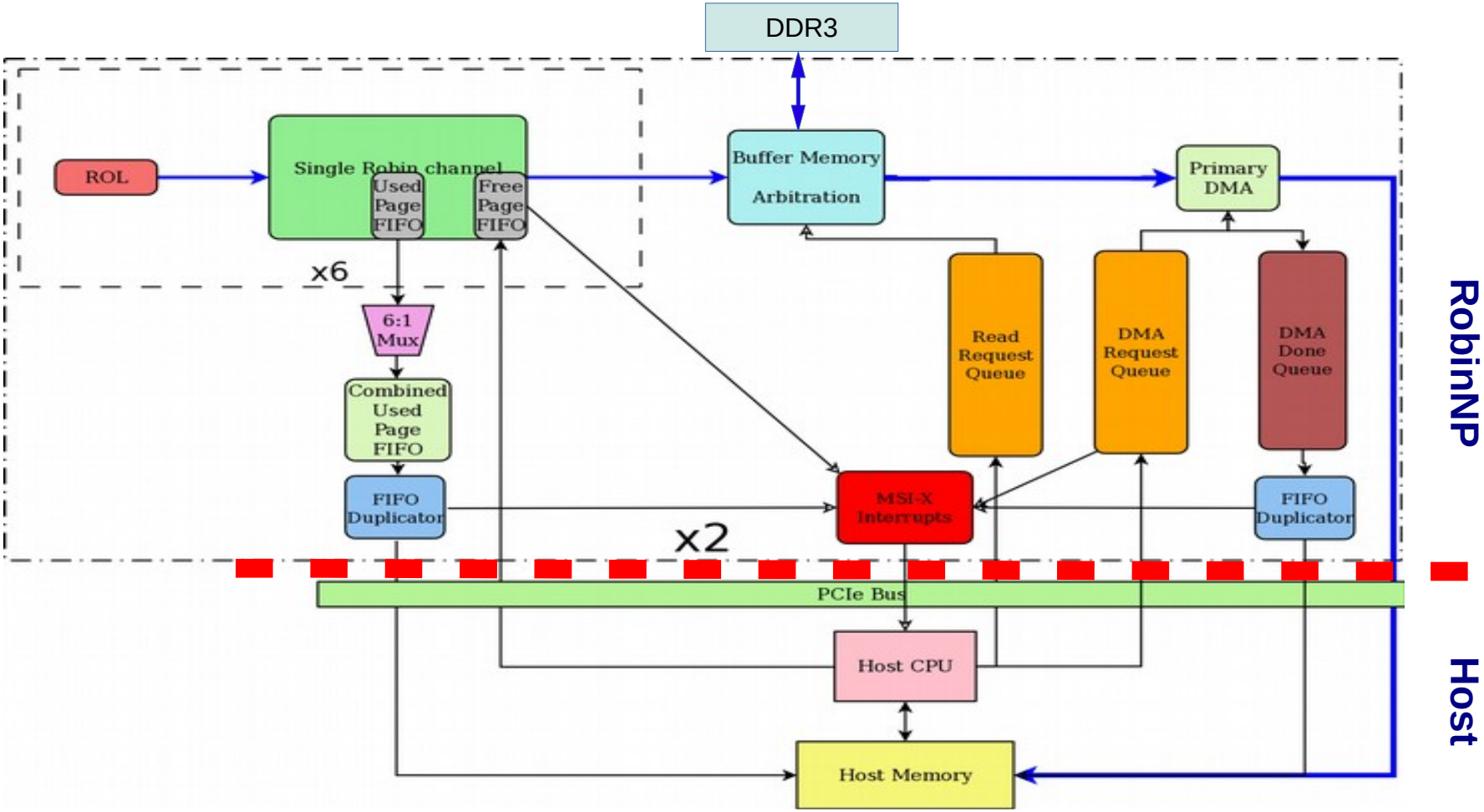
- C-RORC HLT Firmware
  - optical interface: same as DAQ – custom 2/4/5 Gbps, up to 12 links/board
  - online data pre-processing in hardware:
    - ClusterFinding for TPC raw data
    - saves ~25 CPU cores per optical link
  - custom DMA engine for scatter-gather DMA
  - DDR3 for DataReplay

ALICE High-Level Trigger: DMA

- Custom DMA firmware:
  - support scatter-gather DMA
  - configurable number of DMA channels
  - two ring buffers per channel: EventBuffer + ReportBuffer
- Host buffer allocation:
  - allocated from standard Linux memory subsystem
  - no kernel patches, minimal kernel module
  - physical memory can be scattered
  - virtually continuous
  - mapped twice in user space: transparent handling of wrap-around
  - all hardware access from user space



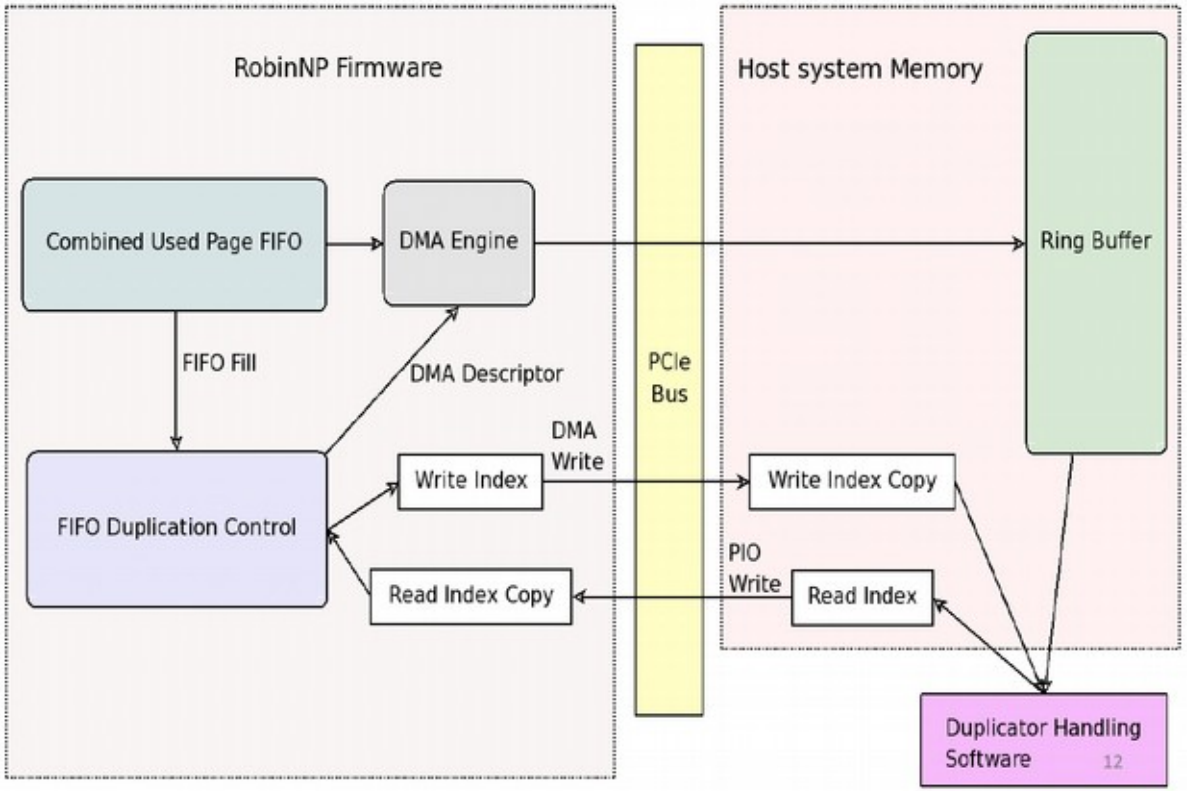
ATLAS RobinNP Firmware and Software



- C-RORC as RobinNP (No Processor)
  - based upon Robin firmware, but offload tasks to the host CPU
  - PLDA DMA engine
- Key innovations at the boundary between the RobinNP and the host

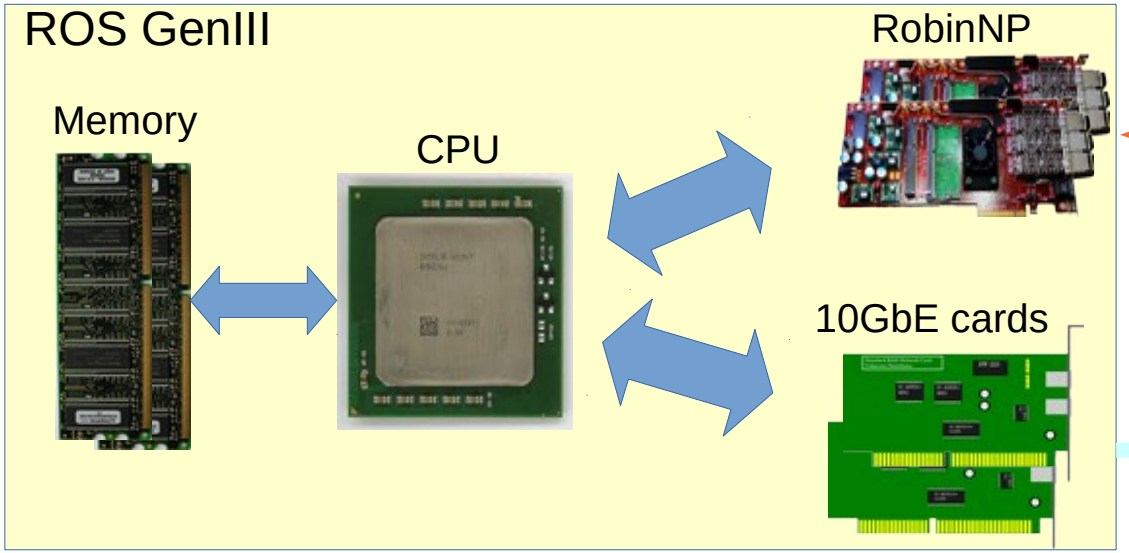
ATLAS RobinNP Firmware and Software

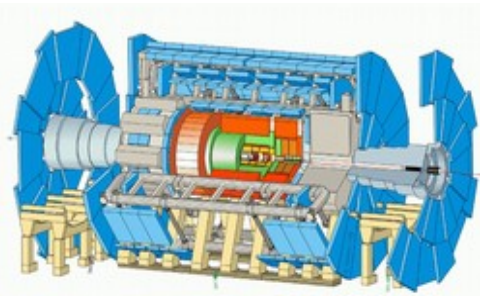
- FIFO Duplicator Control (Firmware)
  - compares write index and read index copy to establish if space available in ring buffer
  - compare available space with FIFO fill
  - transfer smaller of the two values with write index as starting point
  - stop at space before read index copy
  - update write index and write index copy (using DMA)
  - issue interrupt to wake software thread if FIFO was empty upon entry of data

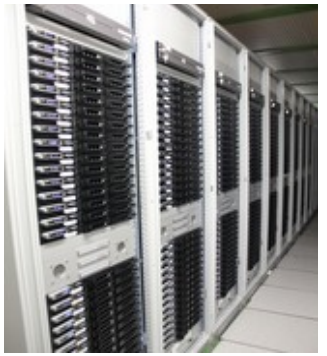


- Ring Buffer Control (Software)
  - dormant until interrupt received
  - compares write index copy and read index to establish fill of ring buffer
  - read data from ring buffer starting at read index (but stopping in space before write index copy)
  - update read index and read index copy (using PCIe write)
  - reset interrupter and await new signal

ATLAS ROS GenIII Architecture







- Baseline ROS GenIII configuration includes
  - 2x RobinNP → 24 input optical links
  - 4x 10GbE ports
    - redundancy
- Performance and development studies in lab setup
  - with/without optical data sources
    - RobinNP can internally generate test data
  - data-sink units equipped with 10 GbE connectivity

At 50% readout → up to **~16 Gbps**
- RobinNPs to memory
- memory to network cards

Run2 Setup:
- 98 ROS PCs
- 1860 ROLs
- 178 C-RORCs
- 392 10GbE links

Summary / Conclusion

- One common piece of hardware for 3 applications at ALICE and ATLAS
- All parties strongly profited from the collaboration
- Development, and esp. coordination of large scale production took much longer than anticipated
- Hardware tests successful
- Run2 installation ongoing