# Readout software
## for the ALICE integrated Online-Offline (O2) system

Filippo Costa, Sylvain Chapeland (CERN)

for the ALICE Collaboration

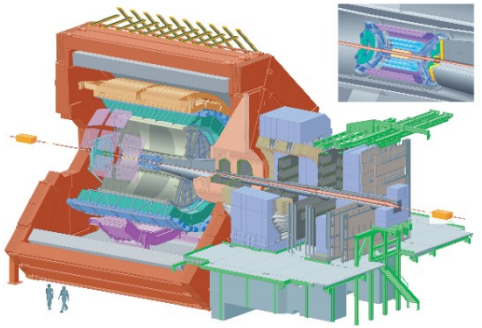
23nd International Conference on Computing in High Energy and Nuclear Physics, CHEP 2018



# Outline

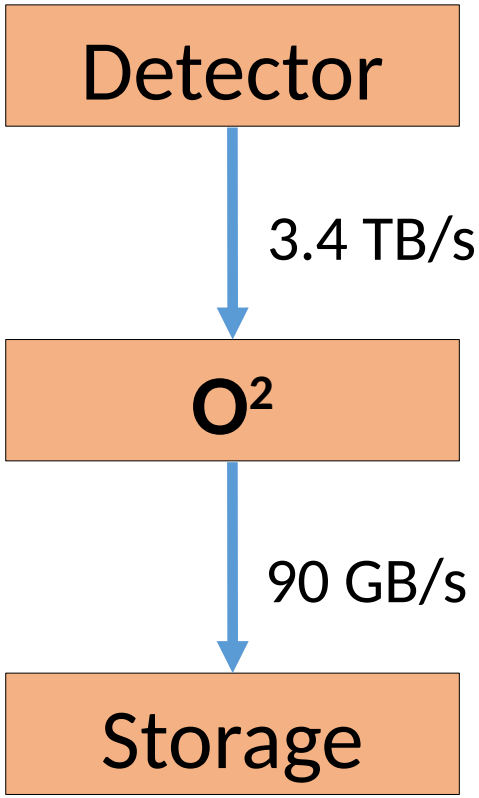
- ALICE $O^2$ project
- A word about hardware
- Readout, first link in the software chain
- Architecture and implementation
- Performance
- Summary

# ALICE Online-Offline project



- ALICE detector to be upgraded
  - LHC long shutdown 2019-2020
- Increased data throughput
  - Demanding processing and compression
  - Estimated farm size:
    - Readout: ~250 nodes
    - Online reconstruction: ~1500 nodes
- See other ALICE $O^2$ talks at CHEP'18

  *TODO: # ids to be listed here*



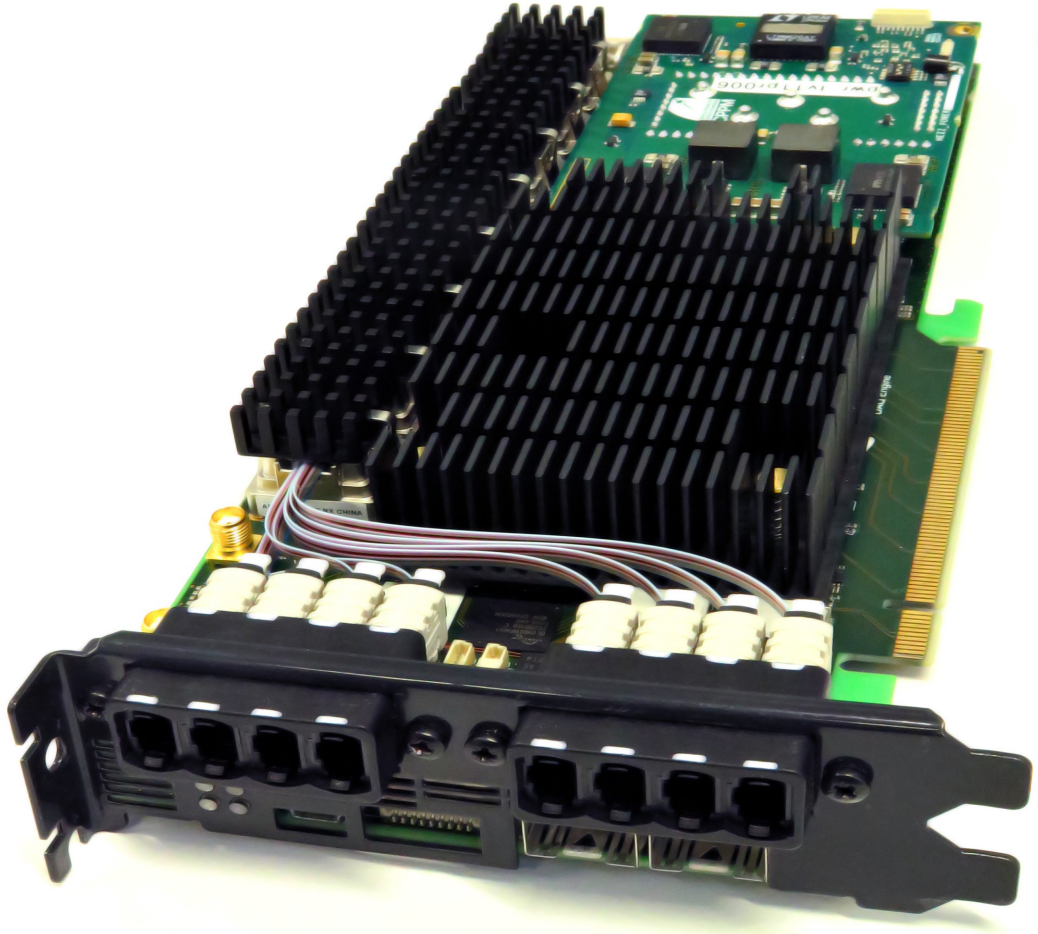
# Detector data links

- Mainly GBT
  - radiation-hard bi-directional 4.8 Gb/s optical fiber link between counting room and experiment
- Also support for DDL
  - ALICE custom link used in run 1 & 2
  - for legacy detectors participating in run 3
- ~8000 links in total

  => need for high-density readout system !

# Readout hardware

- CRU
  - Maximum 48 GBT links input
  - PCIe x16 board
    - FPGA: Intel Arria10
    - Dual DMA engine Gen3 x8
  - Typical throughput: 110 Gb/s
- C-RORC
  - 6 DDL links input, 5.3 Gb/s each
  - Legacy device from run 1 &2

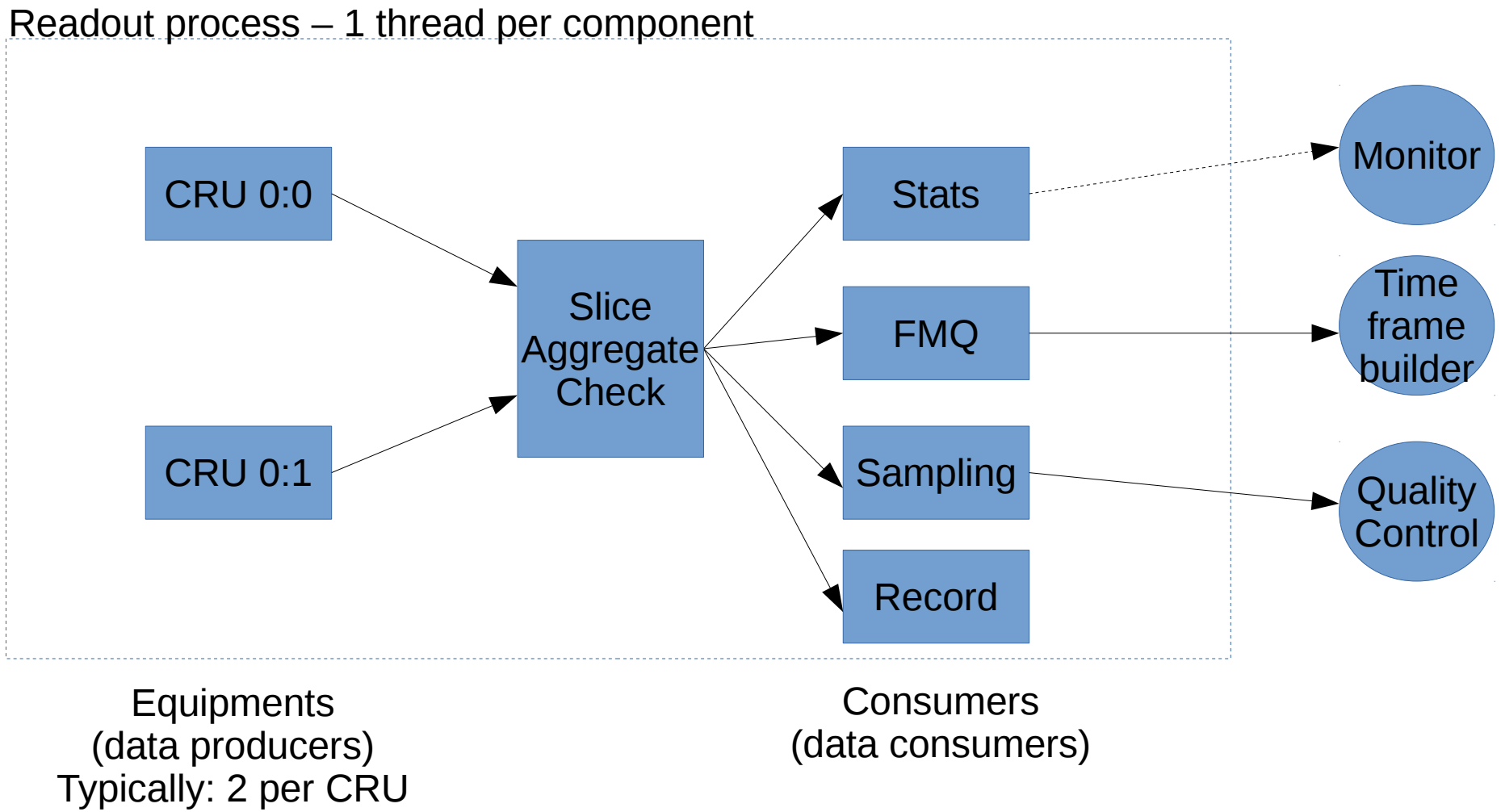


# Readout software tasks

- Move the data from detector electronics into memory of PC
  - Initialize hardware: CRU, C-RORC
  - Allocate memory buffers
  - Provide data pages to be filled by PCIe device
  - Aggregate and slice data input
  - Check data consistency
  - Distribute data to consumers
  - Report performance and errors

=> A process able to readout and record data from multiple devices, with connections to monitoring, data sampling, timeframe building

# Readout architecture



# Readout features

- Multi-threaded application
  - 1 thread per component
  - connected by lockless 1-to-1 FIFO buffers
- Polymorphic definition of producers and consumers for easy extension
  - active components configurable at runtime
- Multiple memory types supported
  - malloc, memoryMappedFile (ROC library, based on HugeTLBFS), shmem (FMQ)
- Minimize CPU needed for readout
  - Controlled polling, leaves CPU for other local tasks
  - Configurable memory banks NUMA pining
  - CPU deep-sleep disabled for best DMA performance (otherwise latency may increase)
- Integrated with other O2 components:
  - configuration, logging, monitoring, data sampling, transport
- Helper components
  - CRU software emulator, local file recording

# Memory layout

- To be done
  - Readout prepares free data pages
  - Give them to CRU
  - Waits them back ready

  - (add CRU  /driver in previous slide?)

# Performance


- Standard setup: 2 CRUs (i.e. 4 equipments), 26 GB/s sustained for 5 days, 6.5GB/s per end-point, using 12% of one CPU core
- CRUv2, ASUS G3 motherboard @ aidrefflp02

- Tried with up to 8 CRUs (i.e. 16 equipments), 40 GB/s (half-duplex PCIe, no optimization)

# Summary

- Readout is a versatile process
  - Initiates DMA transfert between readout cards and PC memory
  - First software component in the O2 online pipeline
  - Connected to the O2 subsystems
  - Lightweight
  - Extendable
- Ready to be used in detector commissioning starting this summer