



O² Project Status Report

Pierre Vande Vyvre for the O2 Project



Outline

- Detector read-out
 - Design and simulation
 - Detector read-out development using the C-RORC
 - Data aggregation and formatting
- TPC software
- Core software development
 - Condition and Calibration Data Base (CCDB)
 - Data Quality Control (QC)
 - Control, Configuration and Monitoring (CCM)
 - FLP Prototype
- Software process
- Data storage performance tests
- CR0 procurement
- Milestones and schedule



ALICE O² in a nutshell

Requirements

1. LHC min bias Pb-Pb at 50 kHz
~100 x more data than during Run 1
2. Physics topics addressed by ALICE upgrade
 - Rare processes
 - Very small signal over background ratio
 - Needs large statistics of reconstructed events
 - Triggering techniques very inefficient if not impossible
3. 50 kHz > TPC inherent rate (drift time ~100 μs)
Support for continuous read-out (TPC)
 - Detector read-out triggered or continuous

New computing system

- Read-out the data of all interactions
- ➔ Compress these data intelligently by online reconstruction
- ➔ One common online-offline computing system: O²
- Paradigm shift compared to approach for Run 1 and 2

Unmodified raw data of all interactions shipped from detector to online farm in triggerless continuous mode

HI run 3.4 TByte/s ↓

Baseline correction and zero suppression
Data volume reduction by zero cluster finder.
No event discarded.
Average compression factor 6.6

500 GByte/s ↓

Data volume reduction by online tracking.
Only reconstructed data to data storage.
Average compression factor 5.5

90 GByte/s ↓

Data Storage: 1 year of compressed data

- Bandwidth: Write 170 GB/s Read 270 GB/s
- Capacity: 60 PB

20 GByte/s ↔

Tier 0, Tiers 1
and
Analysis Facilities

↕

Asynchronous (few hours)
event reconstruction with
final calibration

Hardware facility



270 FLPs
First Level
Processors
(FLPs)

1500 EPNs
Event Processing
Nodes
(EPNs)

34 Storage
Servers

68 Storage
Arrays

Detectors

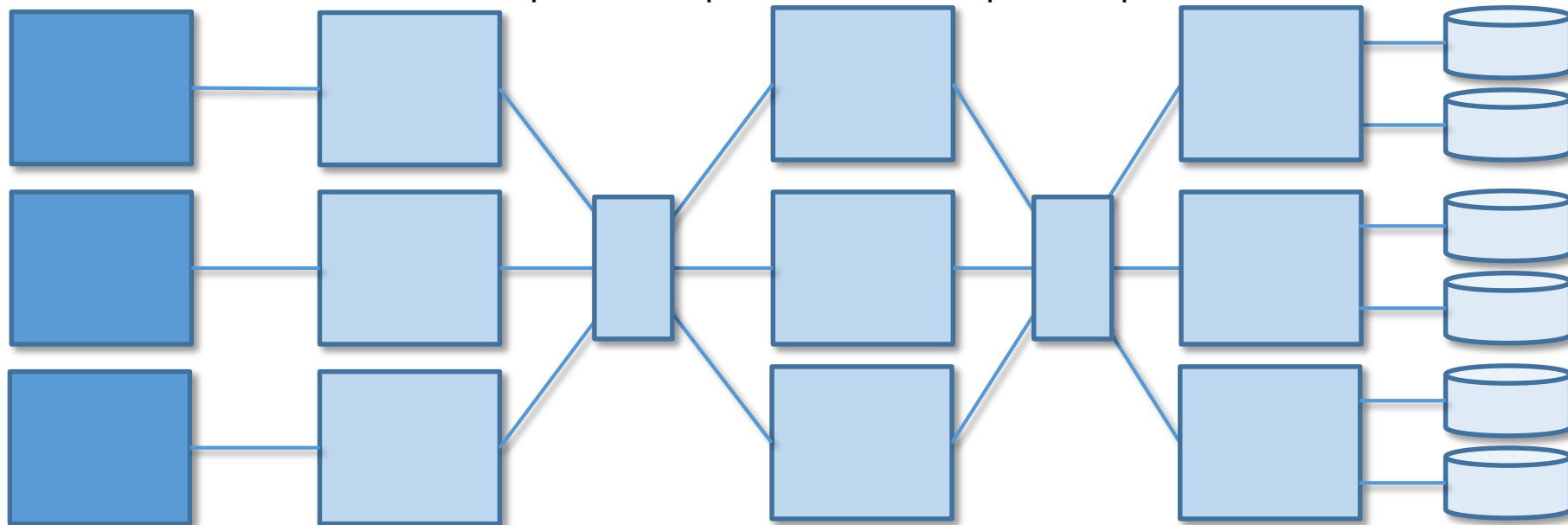
Switching
Network

Storage
Network

9000 Read-out
Links

Input: 270 ports
Output : 1500 ports

Input: 1500 ports
Output : 34 ports



3.4 TB/s

500 GB/s

90 GB/s

Outline

➤ **Detector read-out**

- **Design and simulation**
- **Detector read-out development using the C-RORC**
- **Data aggregation and formatting**

• TPC software

• Core software development

- Condition and Calibration Data Base (CCDB)
- Data Quality Control (QC)
- Control, Configuration and Monitoring (CCM)
- FLP Prototype

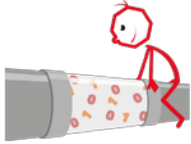
• Software process

• Data storage performance tests

• CR0 procurement

• Milestones and schedule

CWG3: Detector read-out architecture



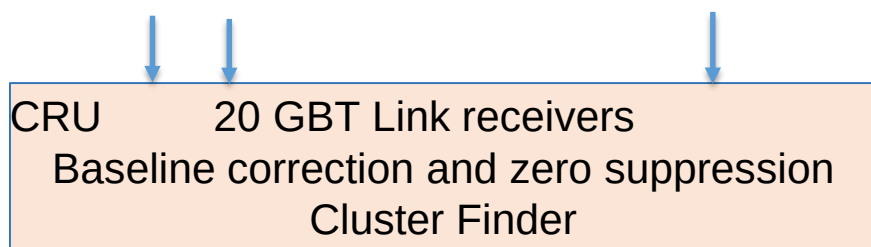
- Modifications of the TPC electronics: data volume reduction functionality moved off detector to the read-out electronics (CRU)
- Refinement to and modifications of the read-out architecture
- In particular introduction of a collective throttling mechanism in the continuous read-out
- **See presentation “Readout and Trigger system ”**



O² with new TPC read-out



GBTs in Wide bus mode
Up to $20 \times 4.48 = 89.6$ Gb/s (5 MHz sampling)



PCIe Gen3
~ 90-100 Gb/s Peak



Network
~ 20 Gb/s Avr

Unmodified raw data of all interactions shipped from detector to online farm in triggerless continuous mode

HI run 3.4 TByte/s

Baseline correction and zero suppression
Data volume reduction by zero cluster finder.
No event discarded.
Average compression factor 6.6

500 GByte/s

Data volume reduction by online tracking.
Only reconstructed data to data storage.

Average compression factor 5.5

90 GByte/s

Data Storage: 1 year of compressed data

- Bandwidth: Write 170 GB/s Read 270 GB/s
- Capacity: 60 PB

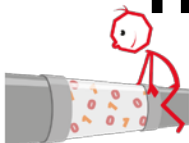
20 GByte/s

Tier 0, Tiers 1
and
Analysis Facilities

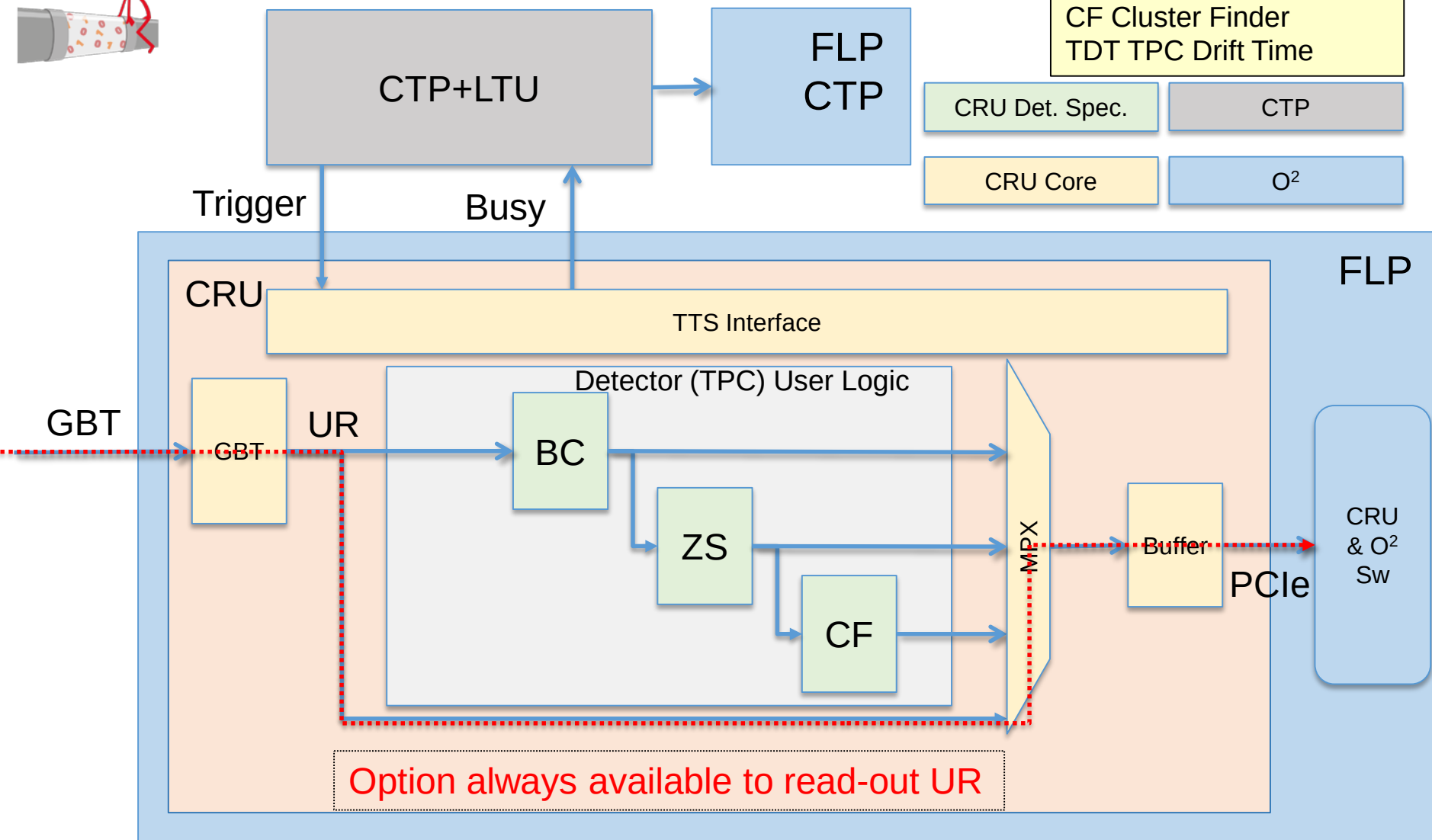
Asynchronous (few hours)
event reconstruction with
final calibration



Triggering and Read-Out System



UR Unmodified Raw data
BC Baseline Correction
ZS Zero Suppression
CF Cluster Finder
TDT TPC Drift Time



TPC requirements

Mode	Run type	Period	Trigger		Acq. Win.	Data type			
			TRG	Rate (Hz)		UR	BC	ZS	CF
Physics	Physics		Cont.	N.A.	Cont.				•
Calibration	Pedestal	1 / Fill	Trig.	10	TDT	•			
	Pulser	1 / Fill	Trig.	10	TDT	O	O	O	O
	Laser	Demand	Trig.	10	TDT	O	O	O	O
	Laser	Physics	Trig.	10	TDT				•
	Krypton	1 / year	Cont.	N.A.	Cont.			O	O
Debug	Technic.	Demand	Trig.	10	TDT	O	O	O	O
	Physics	Demand	Trig.	10	TDT			&	&

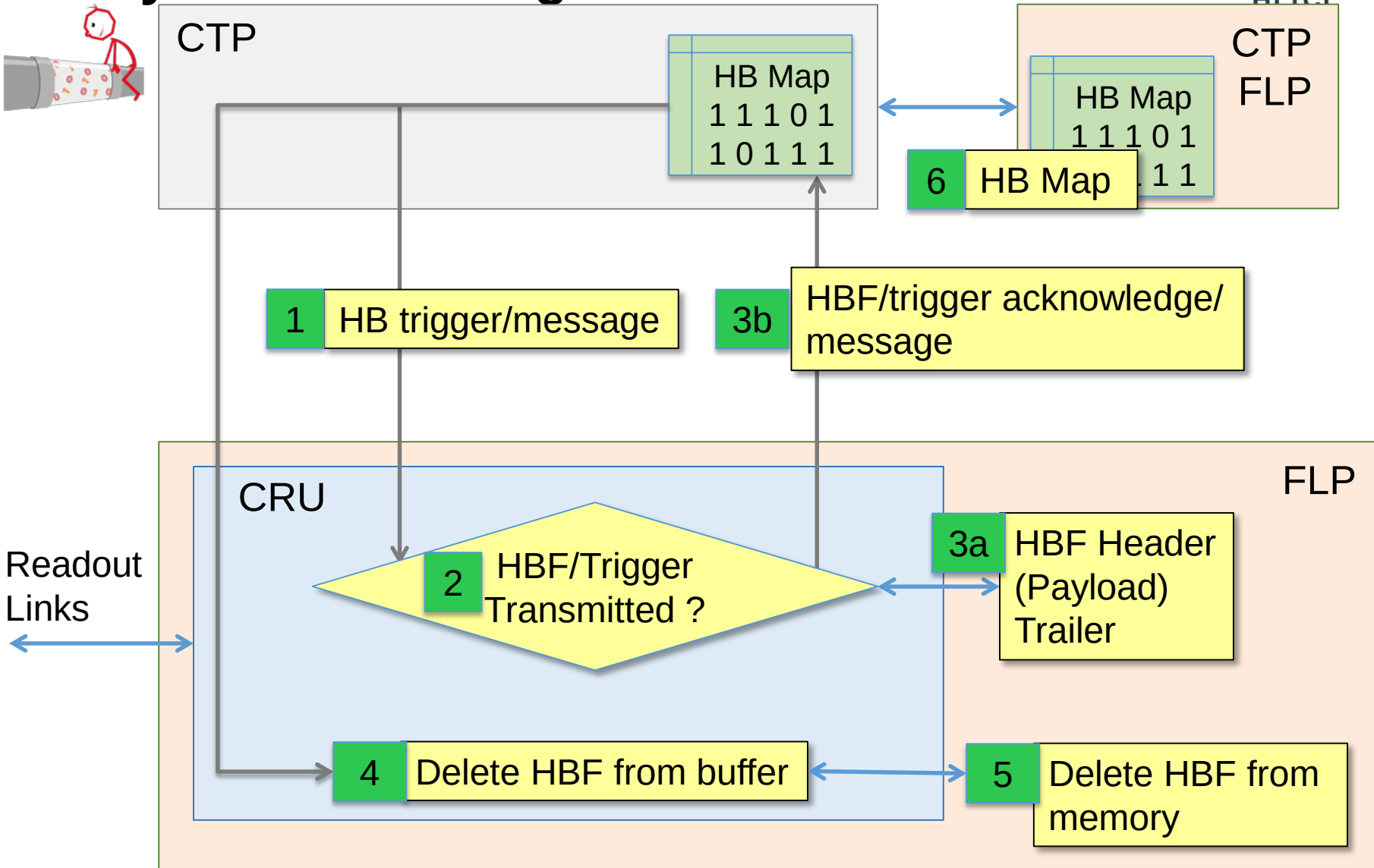
UR Unmodified Raw data
 BC Baseline Correction
 ZS Zero Suppression
 CF Cluster Finder
 TDT TPC Drift Time

• : data type selected
 O: one of the data types will be selected at run time
 &: data types selected and taken simultaneously in the same run

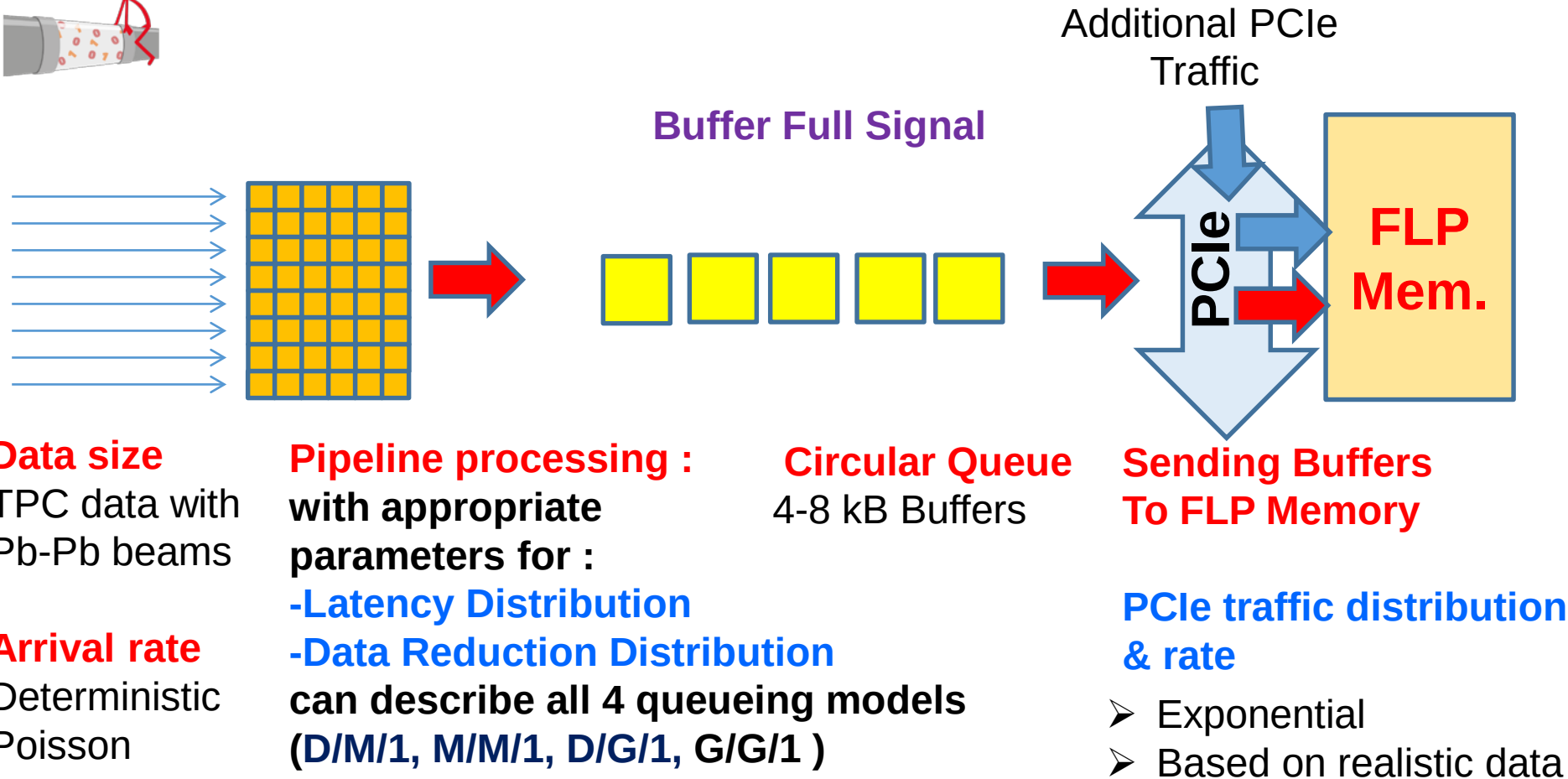


ALICE

System throttling



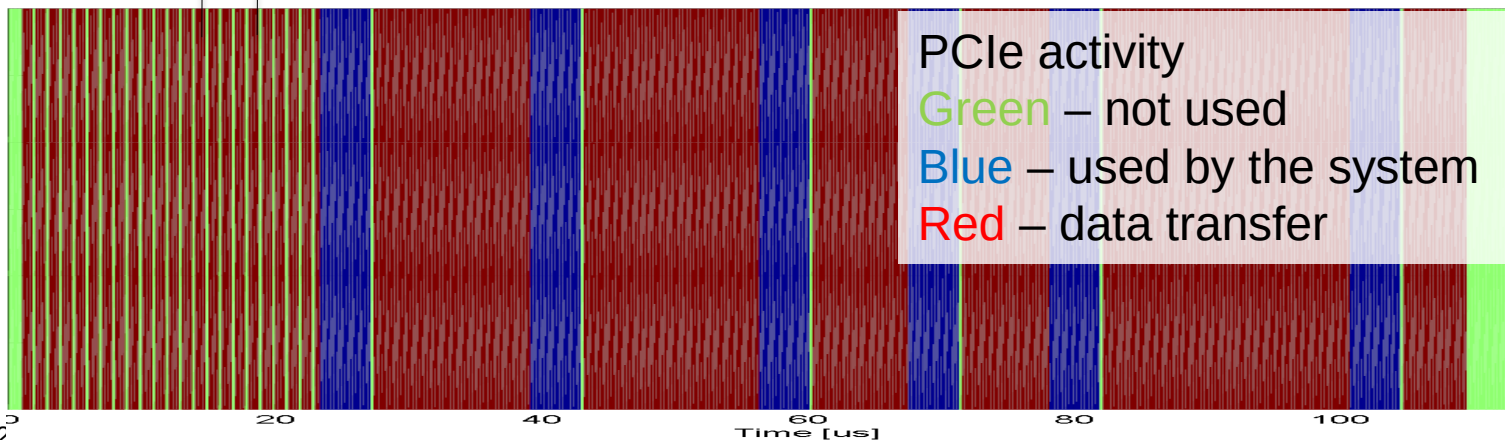
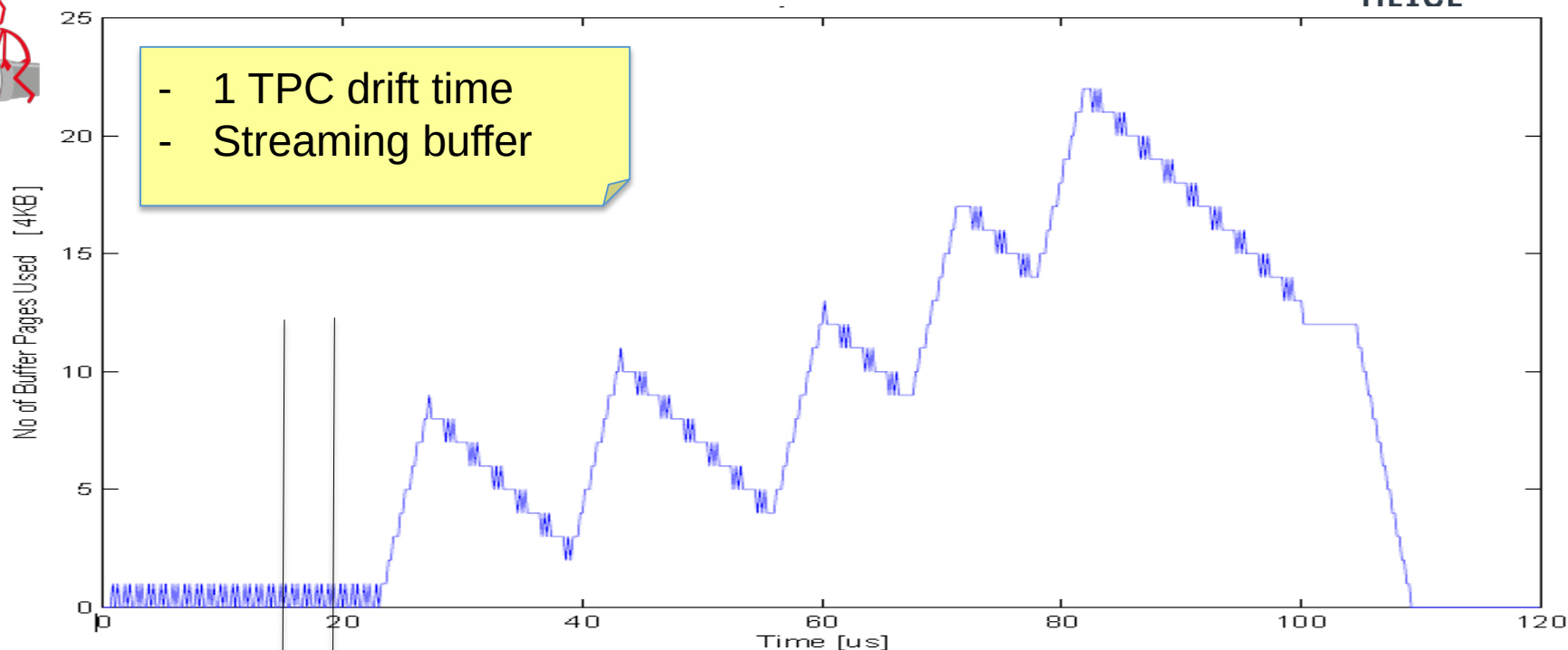
CWG3: Read-out simulation model



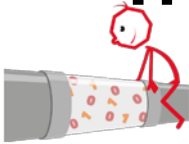
Estimate the probability of data loss as input to the full simulation

Dataflow simulation: impact of the system activity

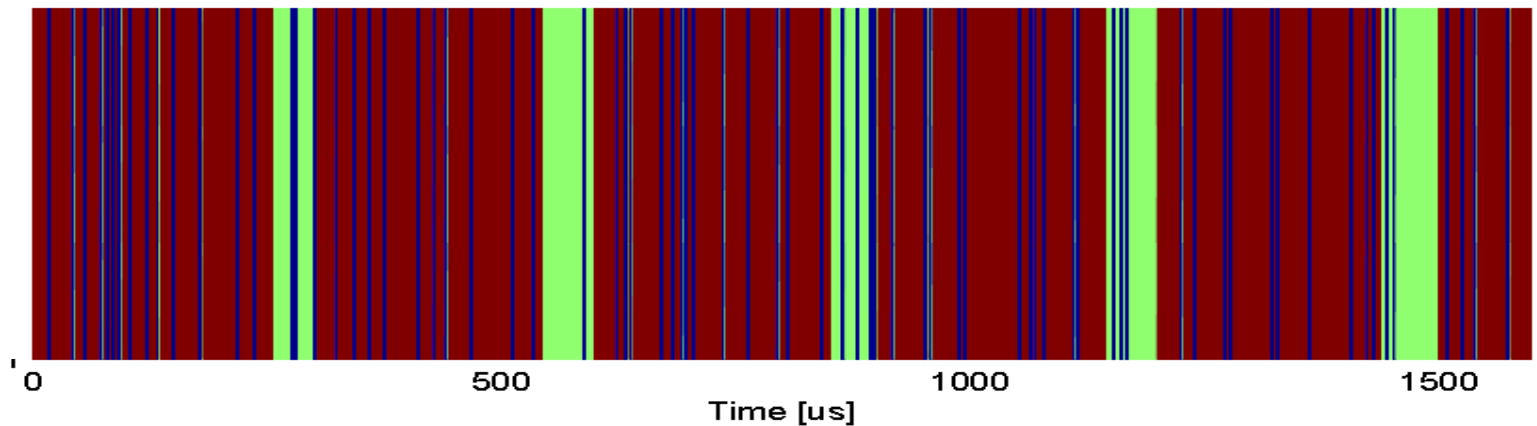
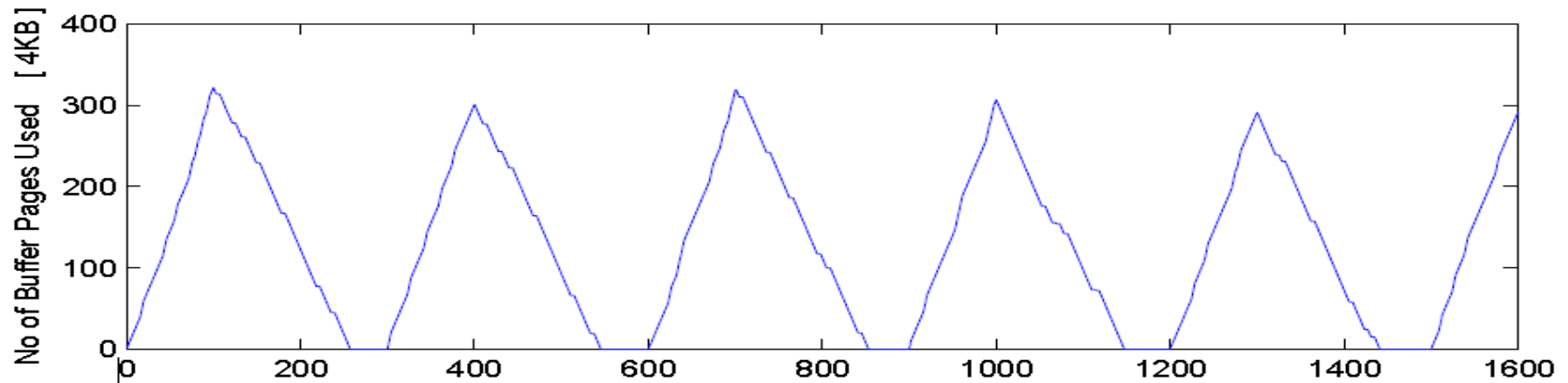
- 1 TPC drift time
- Streaming buffer



Triggered read-out mode

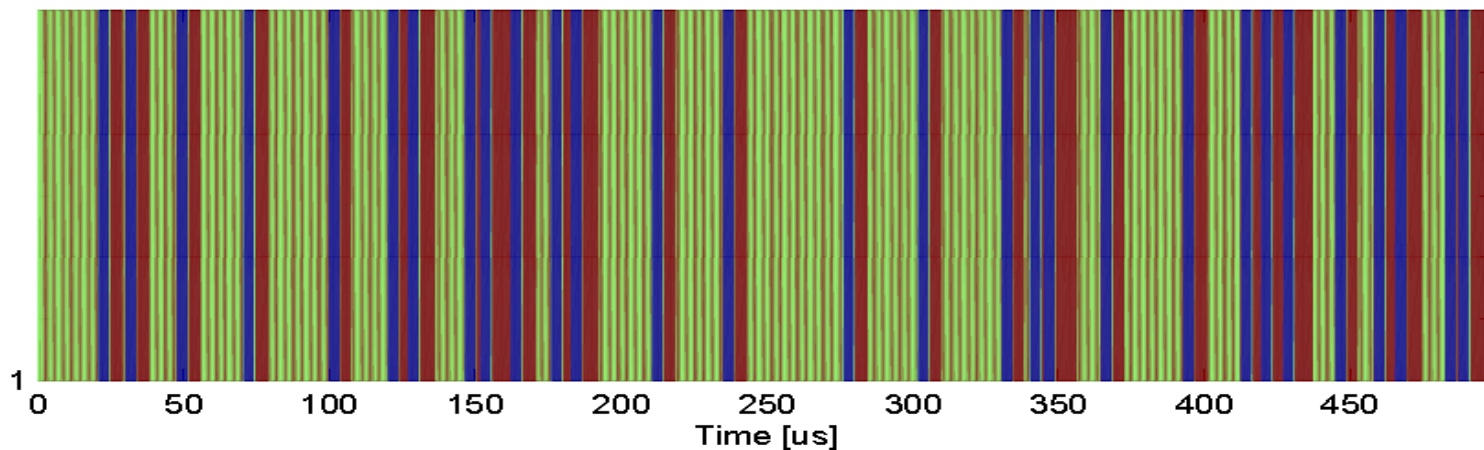
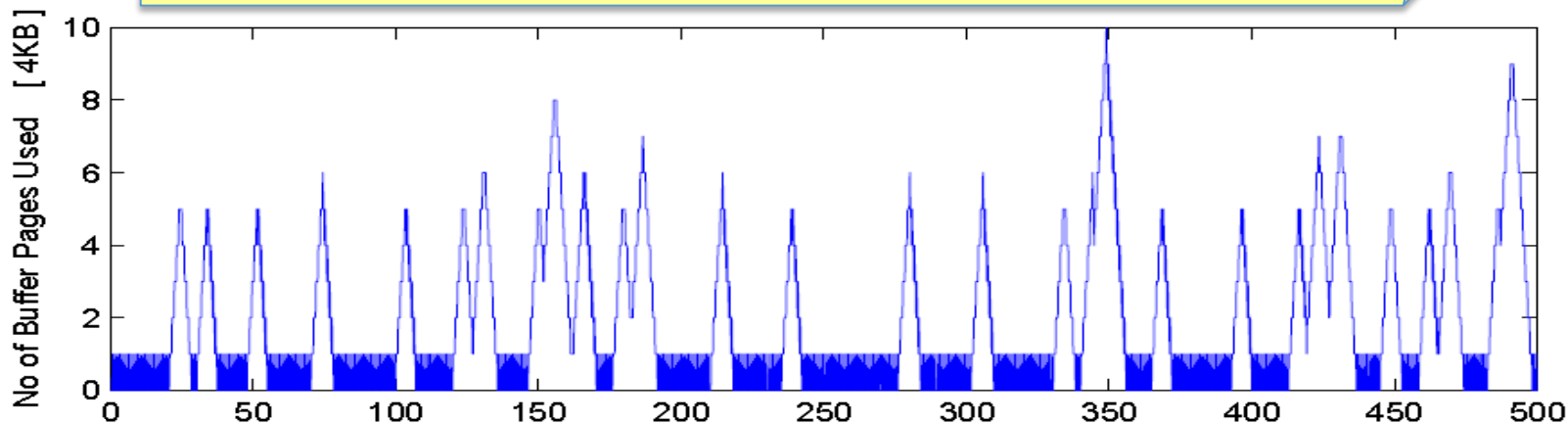


10 kHz trigger (100 μ s TPC drift time) of unmodified raw data
Simulating fixed read-out: 1 out of 3 triggers is read-out



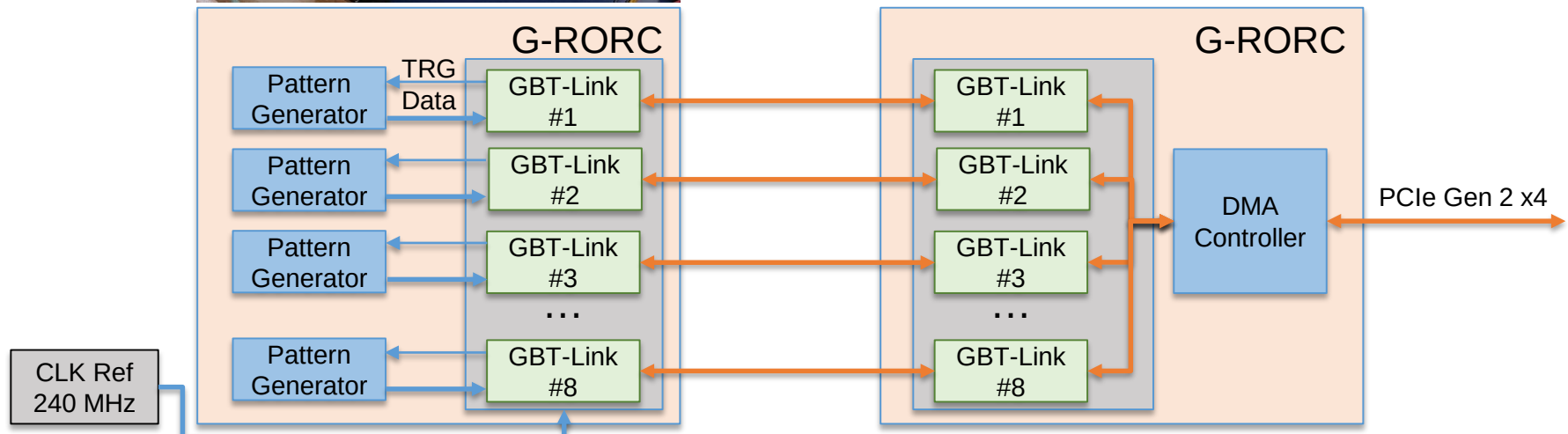
Continuous read-out mode

10 kHz trigger (100 μ s TPC drift time) of cluster data



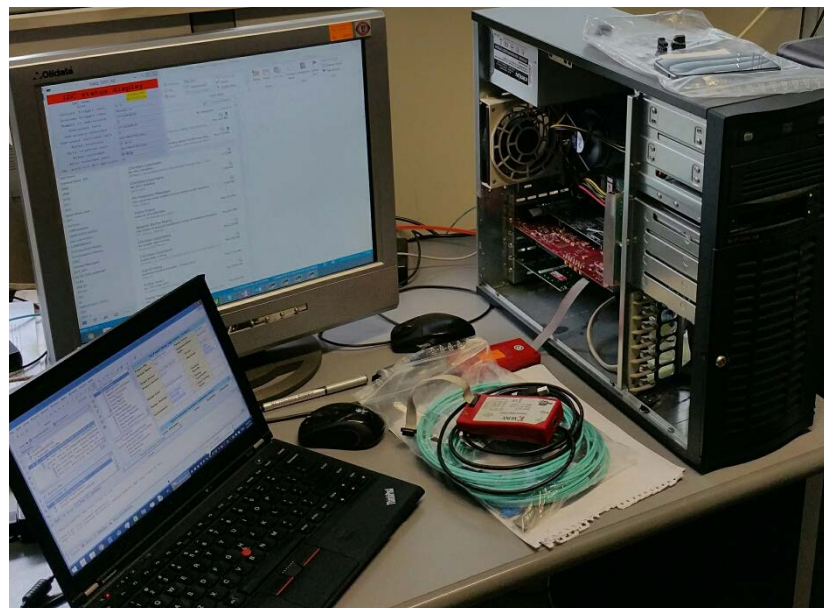
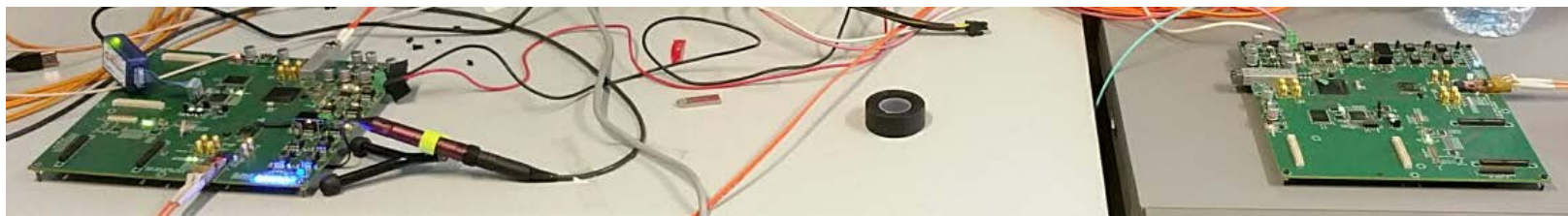
G-RORC: GBT data generator and receiver

- Use of the hw developed for Run 2 (C-RORC) during the transition period
- C-RORC + GBT receiver firmware => G-RORC
- Detector Data Generator (DDG): 8 GBT links each one with an internal data generator.
- Read-out: triggered mode implemented and continuous mode under development.



G-RORC for detector read-out

- GBT packet readout (triggered)
- Tested in the lab using 8 GBT links
- Used for reading out
 - ITS read-out prototype card
 - 2 TOF prototype cards at up to 340 kHz



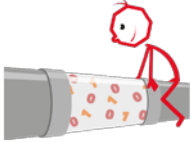
G-RORC development



	# GBT links	Readout mode	VHDL Simulation	Status (System)
Data Generator	8 ch	Triggered	Yes	Yes (DDG)
Read-out	1 ch	Triggered	Yes	Yes (TOF, ITS)
Read-out	2 ch	Triggered	Yes	Yes (TOF)
Read-out	1 ch	Continuous	Yes	In progress (MID)
Read-out	4 – 8 ch	Continuous	In progress	In progress (TPC)
Control	1 ch			In progress (DCS, MCH)

- Several detectors have a G-RORC and are building/using a test system for their FEE using GBT links: FIT, ITS, MCH, MID, TOF, TPC
- The DCS is also working with a G-RORC to develop the new DCS interface

CRU driver status



- CRU driver: based on PDA, tested on CENTOS 7.
- CRU DAQ API
 - Higher-level C++ API
 - Serverless architecture: channel ownership handled by a client library
 - Uses file locks and shared memory for synchronization and persistent state
 - Uses Huge Pages (hugetlbfs) for DMA buffer to minimize SG list
 - Currently in usable prototype state for G-RORC
 - CRU support being worked on
- CRU DCS API: work in progress.

Continuous read-out: Triggers, Heart Beats, Timeframe (1)

Heart Beat (HB)

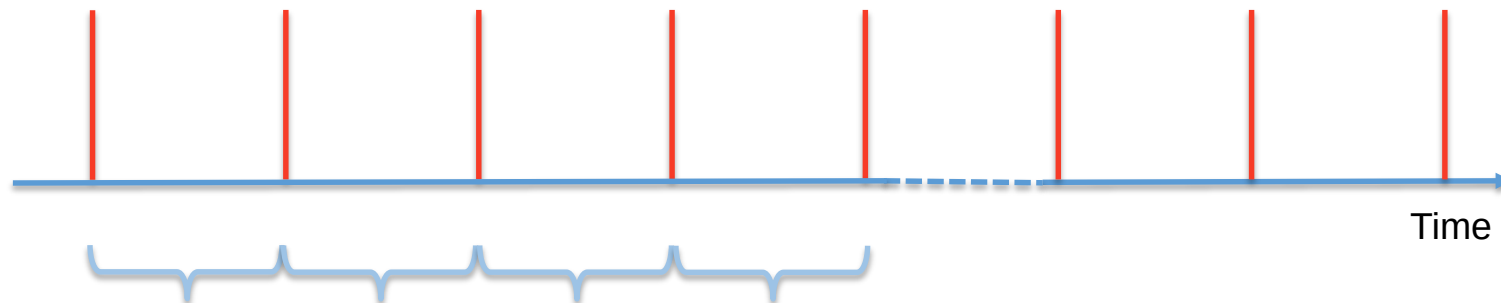
Issued in continuous & triggered modes to all detectors

HB and TF rates programmable

Typical values:

- HB: 1 per orbit ~ 10 kHz
- TF: 1 every 20 ms ~ 50 Hz

Continuous read-out



Heart Beat Frames (HBF): stream of data delimited by two HBs



Sub-Time Frame (STF): grouping of all consecutive HBFs during a time period



Timeframe (TF): software slicing of continuous data

Time Frame grouping of all STFs from all FLPs for the same time period

Continuous read-out: Triggers, Heart Beats, Timeframe (2)

Heart Beat (HB)

Issued in continuous & triggered modes to all detectors

Physics trigger

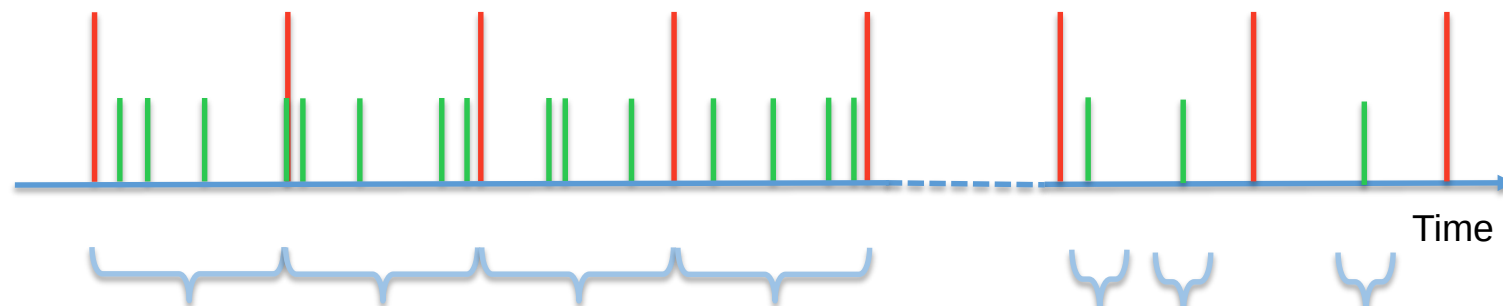
can be sent to upgraded detectors
will be sent to non-upgraded detectors

HB and TF rates programmable

Typical values:

- HB: 1 per orbit ~10 kHz
- TF: 1 every 20 ms ~50 Hz

Triggered read-out



Heart Beat Frames (HBF): stream of data delimited by two HBs

Trigger data fragments

FLP1

FLP2

FLP N



Sub-Time Frame (STF): grouping of all consecutive HBFs during a time period

STF: encapsulating subevents

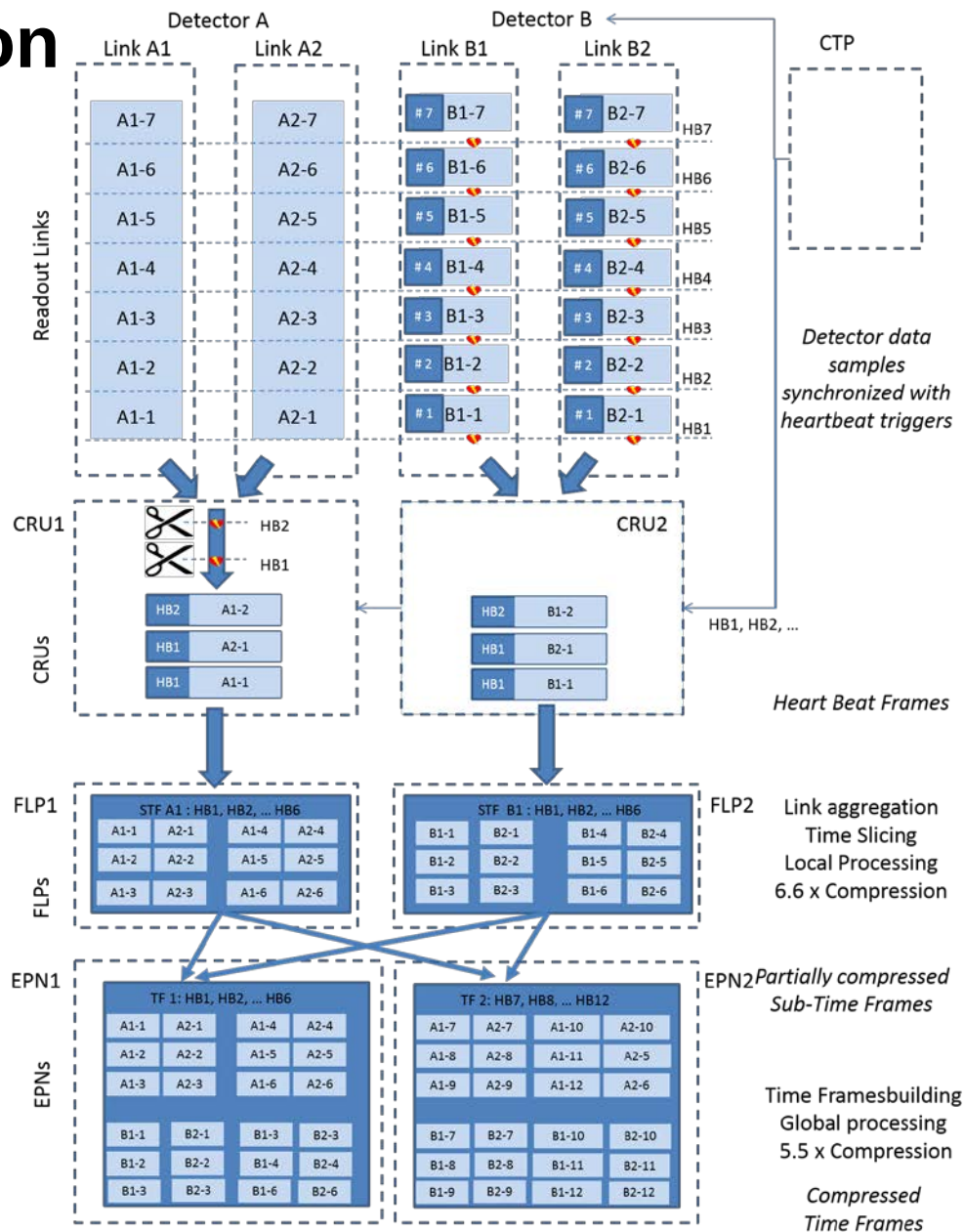
EPN

Timeframe (TF): grouping of all STFs from all FLPs for the same time period from detectors triggered or read out continuously

Data Aggregation

Unformatted data (e.g. TPC)
Insertion of the Standard Data Header (SDH) by the CRU

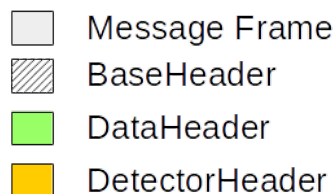
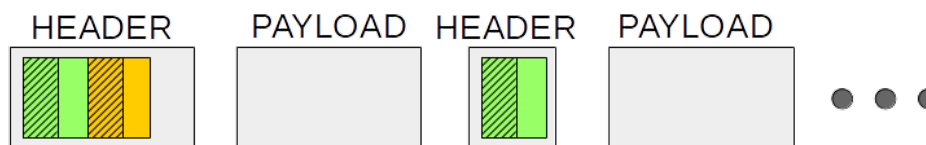
Formatted data (e.g. ITS)
Creation of the SDH by the detector electronics





CWG4: Raw data format

- Aggregation of HBFs each with one SDH into STFs and TFs
- STFs and TFs consist of multiple parts (separate header and payload message parts).
- The header message part consists of one or multiple headers
- Each header is derived from a BasicHeader with the general information (e.g. unique number, type, flag for next header)
- The payload is described by the DataHeader derived from BaseHeader
- Optional headers follow, e.g. detector specific header with trigger information, time frame specific header



Outline

- Detector read-out
 - Design and simulation
 - Detector read-out development using the C-RORC
 - Data aggregation and formatting

➤ **TPC software**

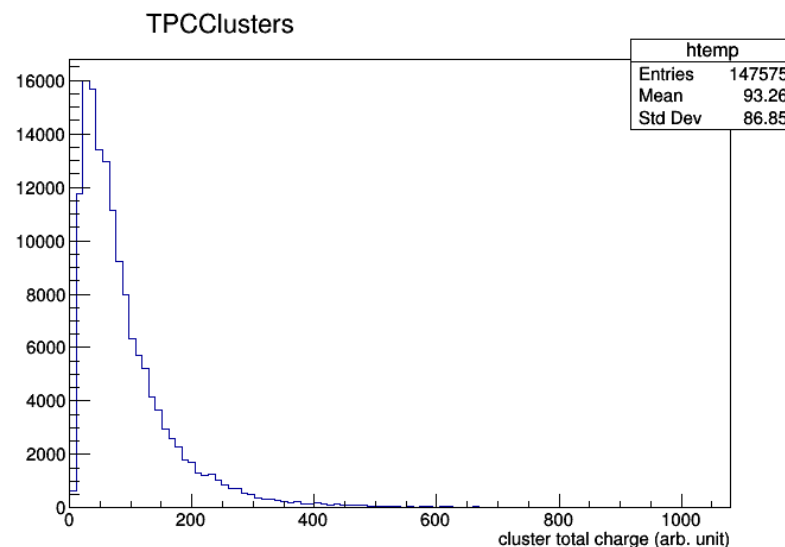
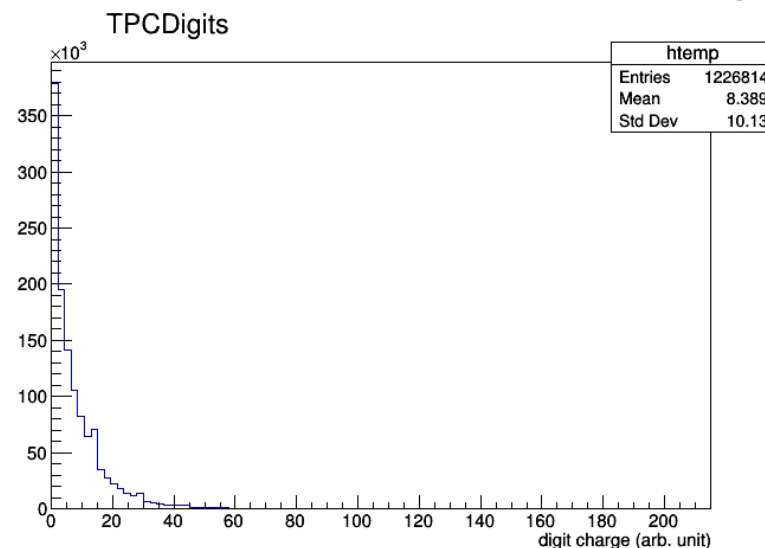
- Core software development
 - Condition and Calibration Data Base (CCDB)
 - Data Quality Control (QC)
 - Control, Configuration and Monitoring (CCM)
 - FLP Prototype
- Software process
- Data storage performance tests
- CR0 procurement
- Milestones and schedule

TPC Software Status

Jens Wiechula



- Basic simulation implemented:
 - Geometry / hit creation
 - Digitisation
 - Clusterisation
 - New pad plane layout mapping
- Still event based



TPC software: next TPC milestones

Jens Wiechula

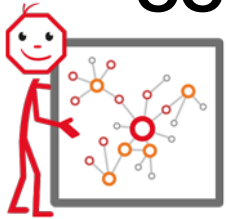


- Implement present HLT tracker
 - As standalone version (Q4 2016)
 - Natively in O2 Q2 2017
- More realistic energy loss and GEM response
- Use O2 framework to analyse and simulate test beam data planned for Q2/3 2017
- Implement static distortions
- Simulating continuous readout
- Start implementing distortion calibration procedures using ITS + TRD in O²

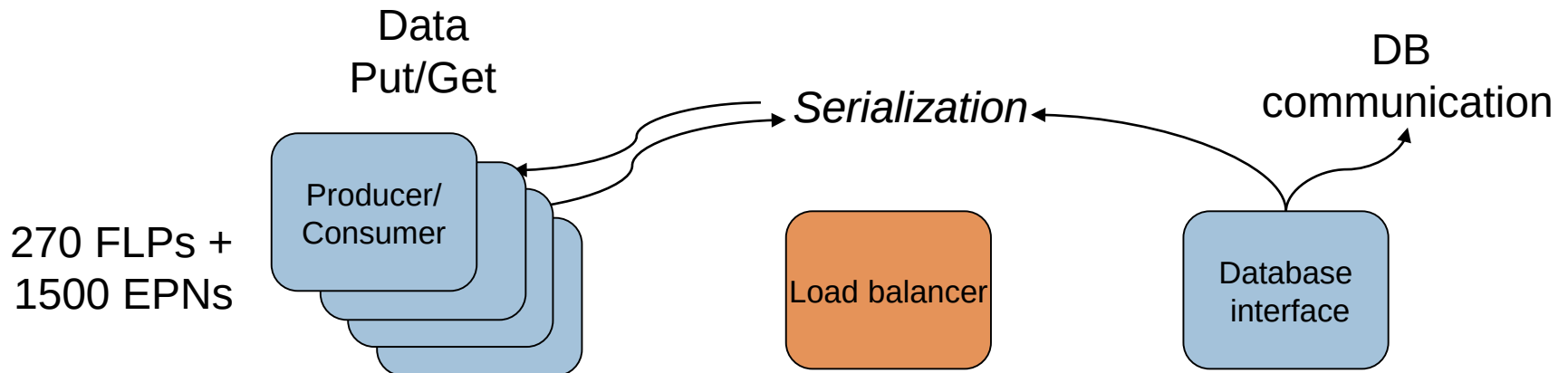
Outline

- Detector read-out
 - Design and simulation
 - Detector read-out development using the C-RORC
 - Data aggregation and formatting
- TPC software
- **Core software development**
 - **Condition and Calibration Data Base (CCDB)**
 - **Data Quality Control (QC)**
 - **Control, Configuration and Monitoring (CCM)**
 - **FLP Prototype**
- Software process
- Data storage performance tests
- CR0 procurement
- Milestones and schedule

CCDB Development

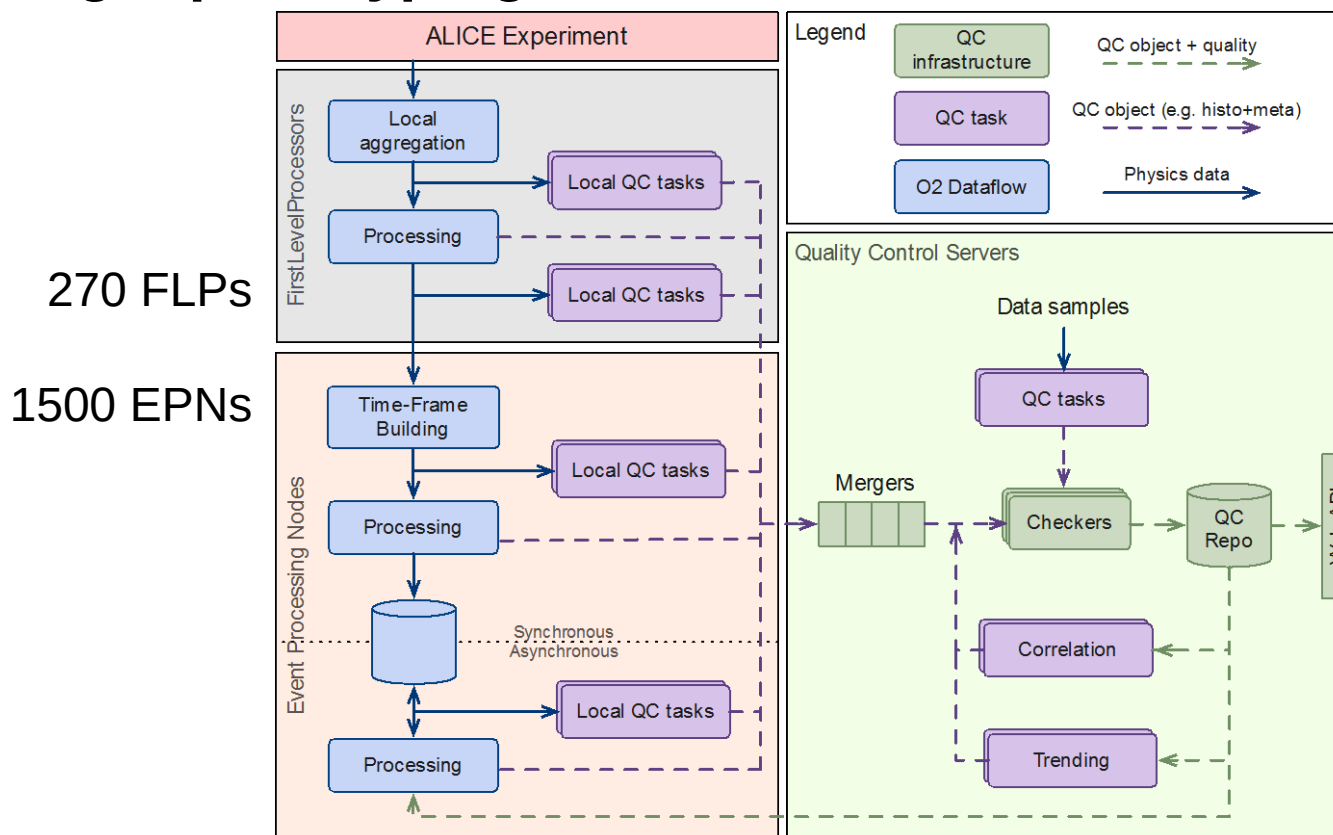


- New Parameter manager under development : Condition and Calibration Data Base (CCDB)
- The Parameter manager consists of different processes that offer the following functionality



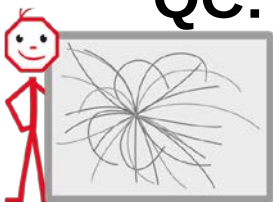
- Producer/Consumer common functionality using Riak (distributed highly available, scalableNoSQL database)

QC: merger prototyping



- Major challenge of the QC consists of merging results from many nodes
- General purpose merger to merge efficiently ROOT objects (calibration, reconstruction and analysis objects)
- Integrated in the ALICE O² framework based on FairROOT
- Scaling and monitoring via Dynamic Deployment System (DDS)
- Systematic benchmarking to characterize the different software elements

QC: merger performance tests

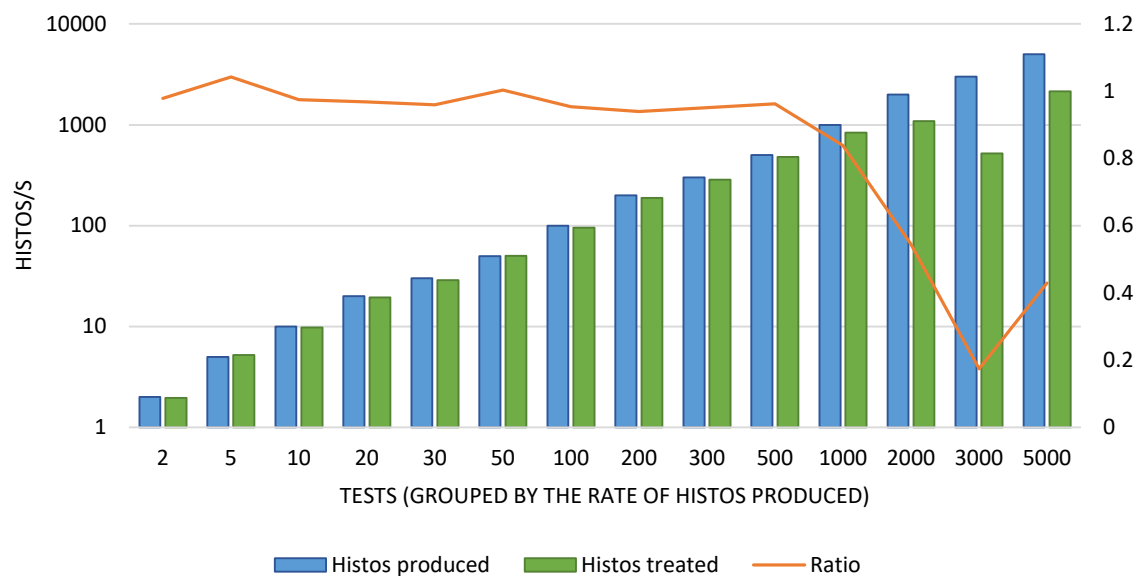


- First implementation of the QC framework
 - Tasks, checkers, storage and clients
 - QC tasks, reconstruction and calibration processes, are highly parallel (1500 EPNs)
 - Merging their output synchronously is challenging

- Benchmark:

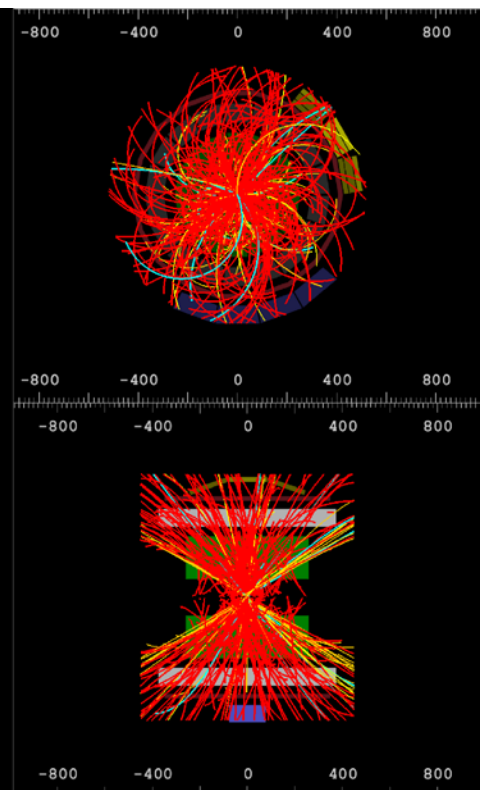
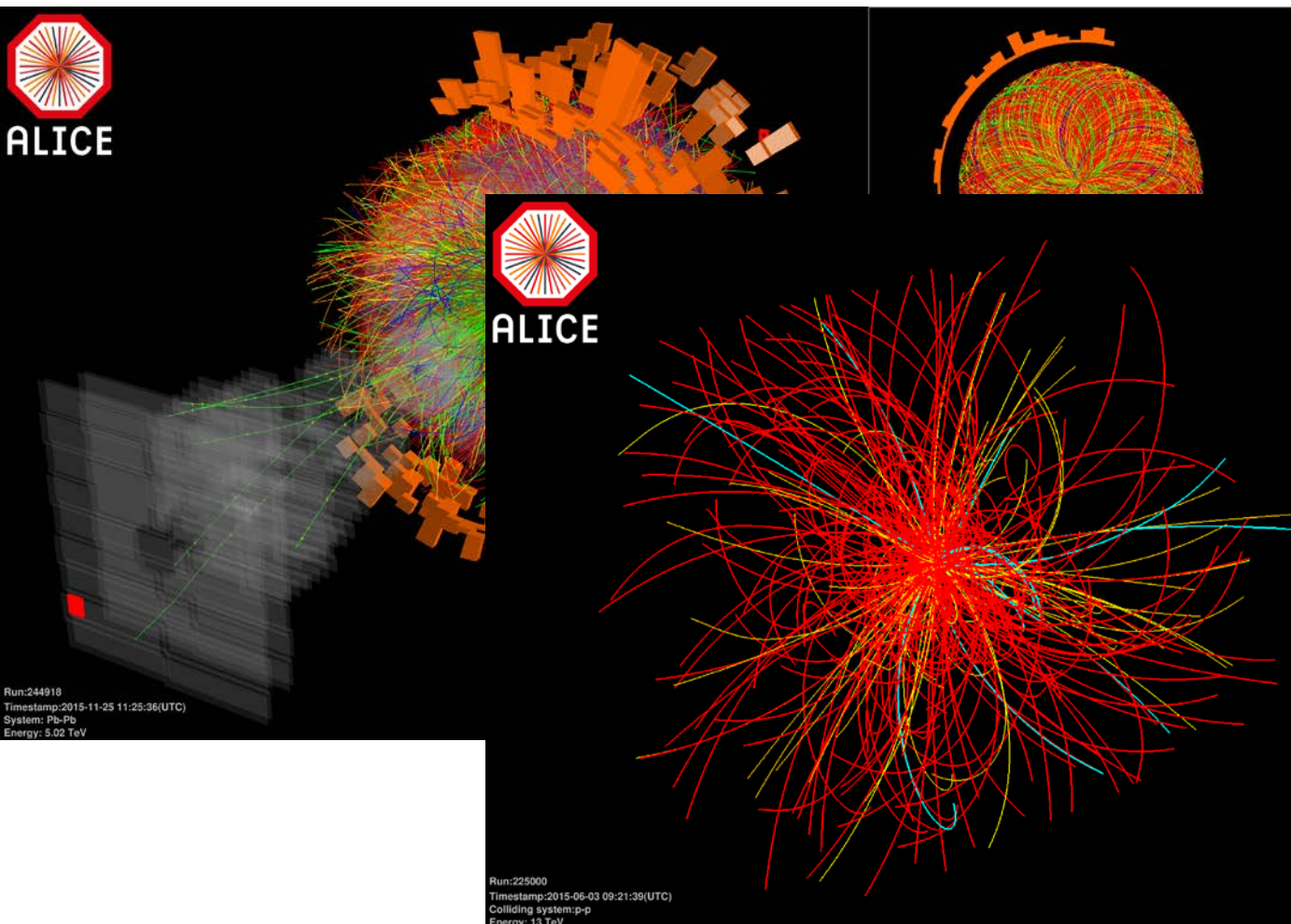
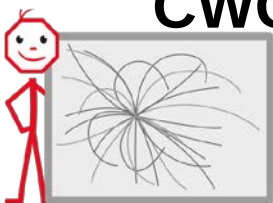
- Checkers can cope with more than the required 400 histos/s in production
- CPU&RAM usage is reasonable and mostly due to serialization and network
- Special care should be given to the database to avoid creating a bottleneck

Comparison of the rate of histograms produced and treated in the system (all tests aggregated)

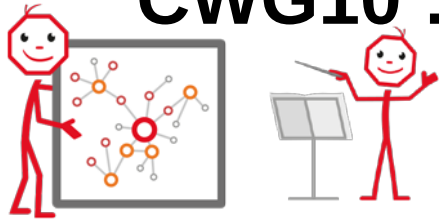


CWG9: Event Display

- Successful development of the ED for Run 2
- Evolutionary strategy from Run 2 to Run 3



CWG10 : Control status



- Dynamic Deployment System (DDS) V1.2 released on 06-07-2016
 - dds_intercom_lib
 - SLURM and Mesos plugins
- Ongoing tests with
 - DDS, FairMQ, State Machine, Zookeeper
 - Provide feedback to developers, identify possible limitations
 - Use QC as test bench: Multiple processes, multiple hosts
- Ongoing evaluation of [Apache Mesos](#) (Operating system of the computing Center):
 - Small (250 cores) ALICE Offline cluster uses Mesos in production. Experiences and evaluations as input to create a larger Mesos deployments.
 - Initial setup deployed on O² development cluster (44 machines, 704 cores). Work ongoing to prepare a demonstrator:
 - Show how Mesos can be used to manage O² workflows
 - Use cases with the goal of describing a cluster architecture which can scale to ALICE O² needs.

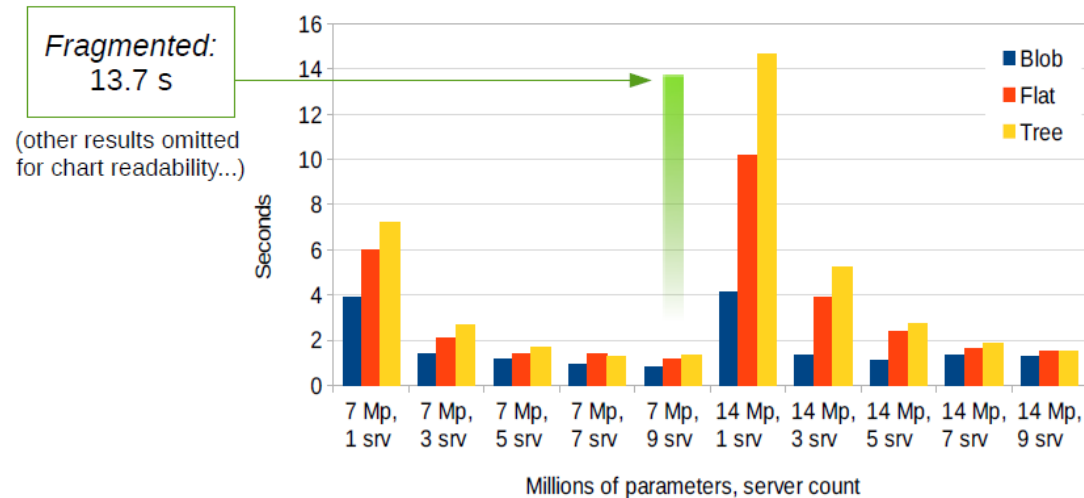
CWG10 : Configuration status

- Configuration library

- Simple put/get interface
- Allow processes to read/write configuration from repository
- Supports multiple backends
 - From file
 - From etcd

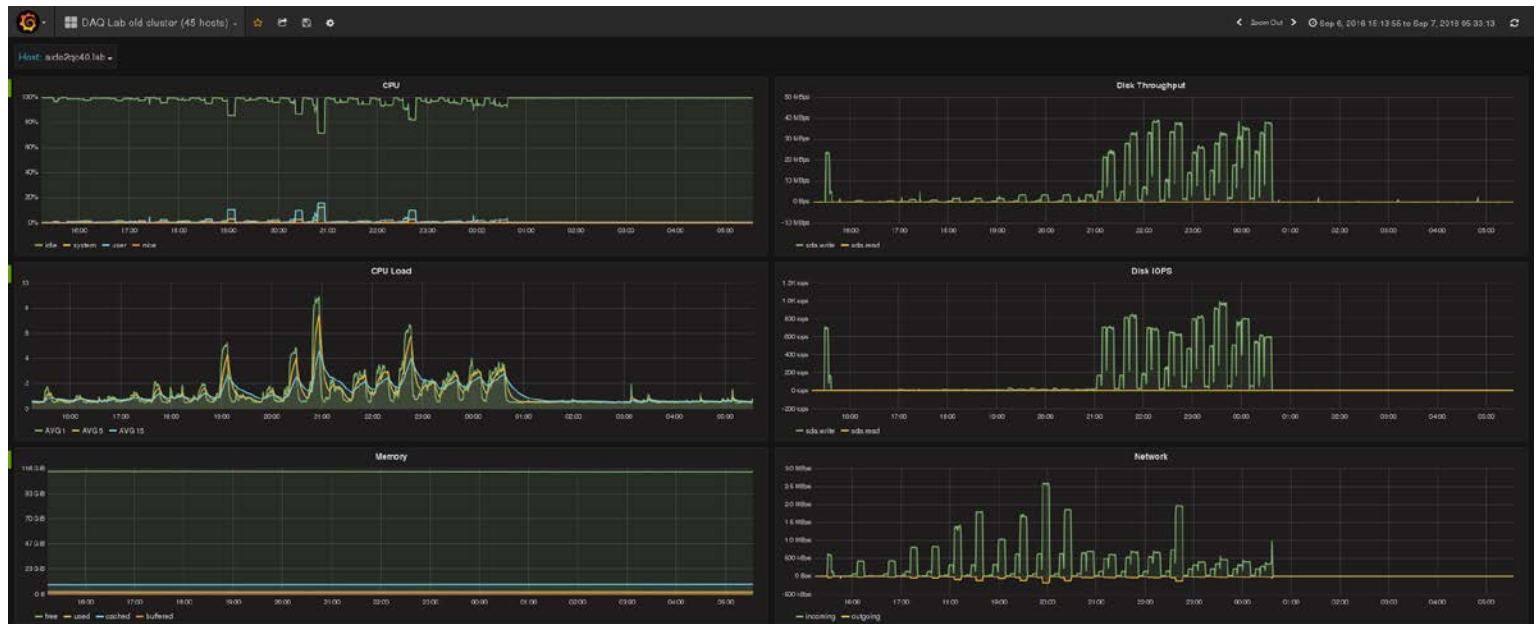
- Etcd

- Distributed key-value store
 - Performance tests look promising
 - 14M parameters retrieved in less than 2s
 - Coming: more tests – authentication, more data structures, proxy, linearized read mode, write performance
- Coming: similar tests with Consul, MySQL, shared file system



CWG10 : Monitoring status

- Monitoring library
 - Send application specific values to monitoring system
 - Perform self-monitoring of processes (CPU, mem)
 - Generate derived metrics (rate, average)
 - Support multiple backends (cumulatively)
- Evaluating many tools, some mature (MonALISA), some emerging (influxdb)
 - Monitoring sensors: MonALISA, collectd
 - Monitoring backends: Logging, MonALISA, InfluxDB
 - Monitoring GUIs: MonALISA, Grafana
- Coming: performance tests, more tools (Apache Flume, statsd, Apache Spark)

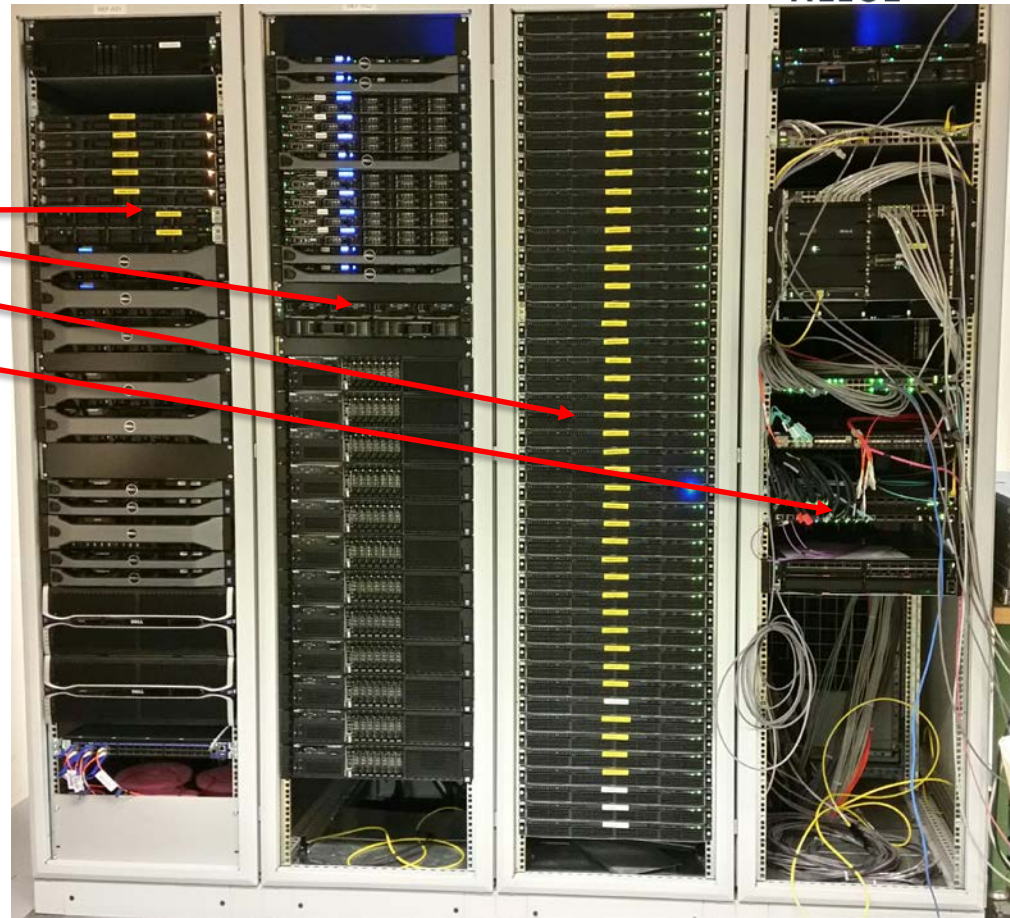


Grafana
dashboard
in O²
development
cluster

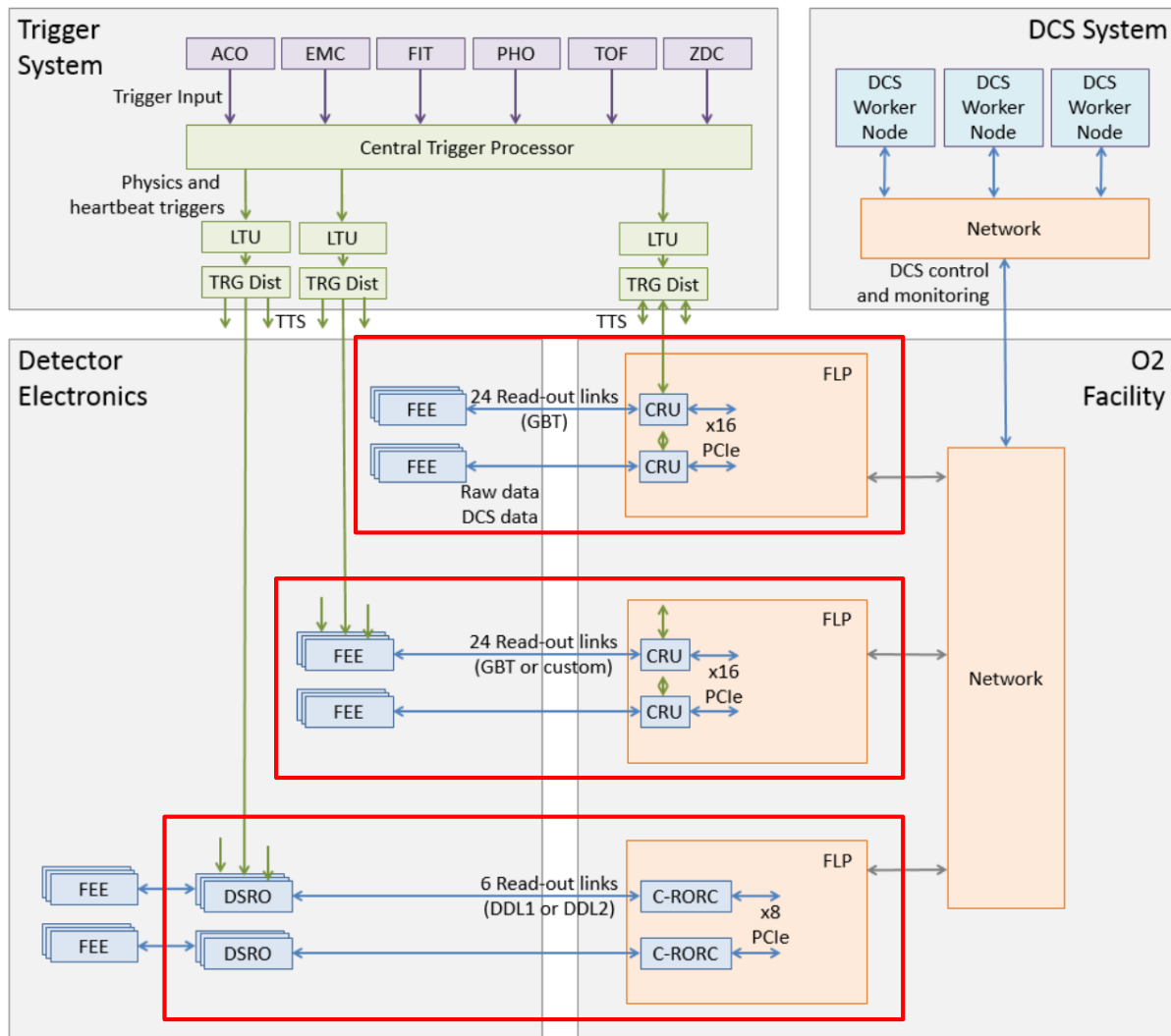


Services running in O2 development cluster

- Cluster
 - Servers for all services
 - 44 worker nodes, 704 cores
 - Network
- Monitoring: MonALISA
 - sensors installed on all nodes
 - service and repository running on monitoring (mon) server
- Monitoring:
collectd + influxdb + grafana
 - collectd running on all nodes (cpu + mem + network)
 - influxdb running on mon server
 - grafana running on web server
- Configuration: etcd
 - etcd server running on 10G server (will be moved to conf server)
 - etcd-browser running on web server



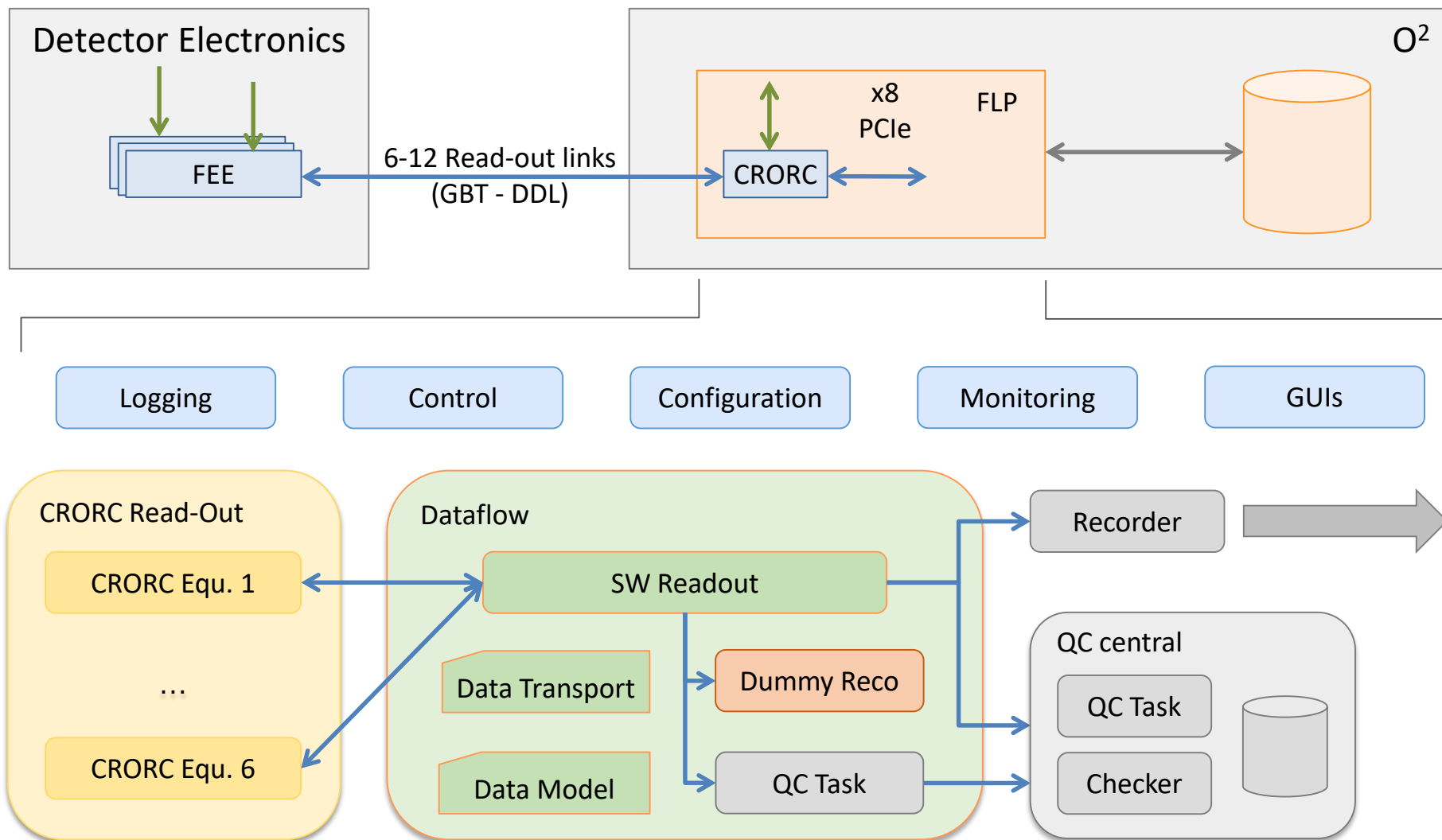
FLP Prototype



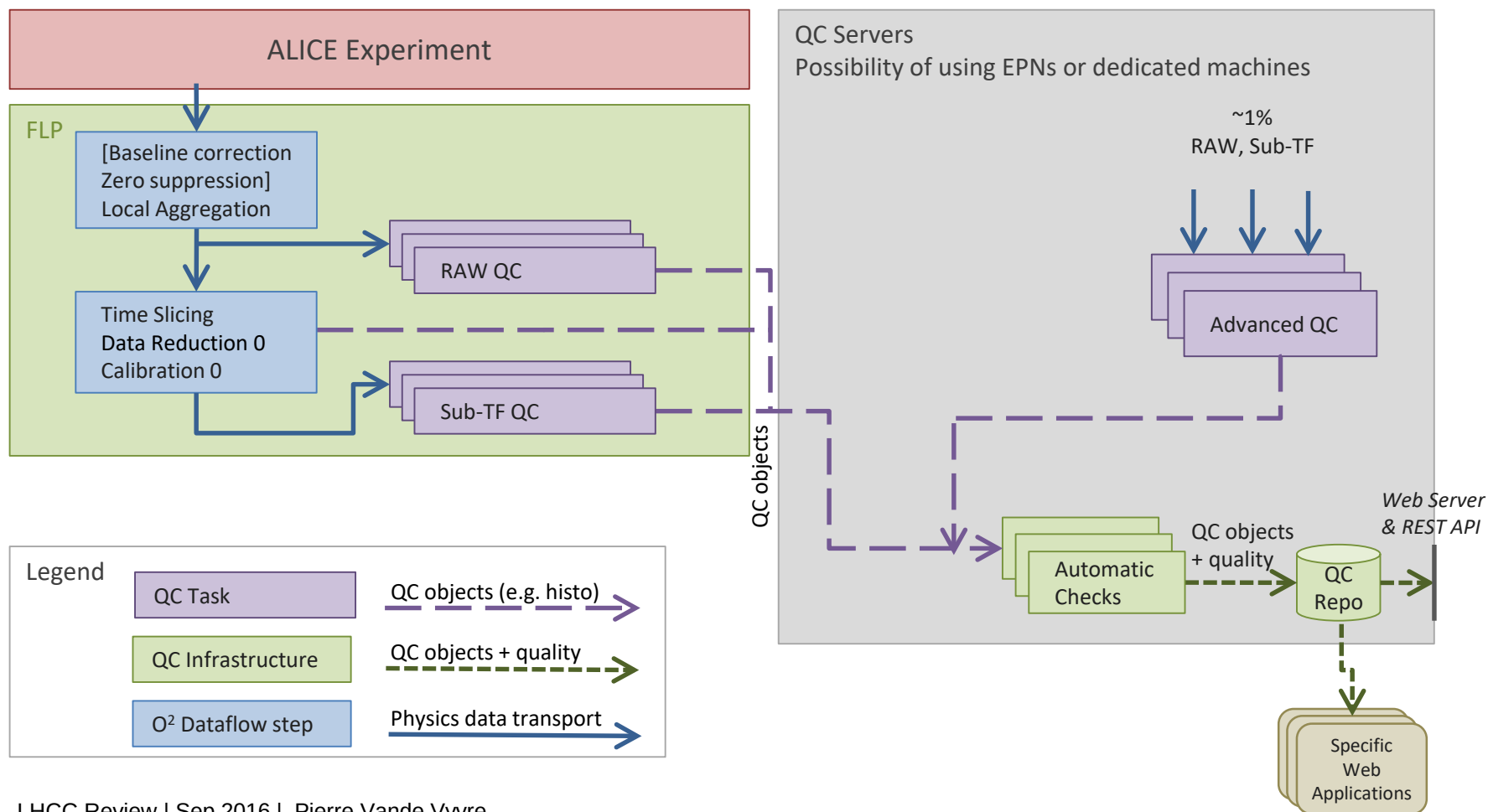
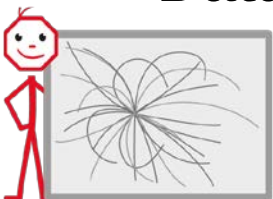
Prototyping and short term needs (2017)

- Kick off development
- Practice and evaluate dev. guidelines
- Meet detector needs

FLP Prototype : software modules



Data QC for the FLP prototype



Outline

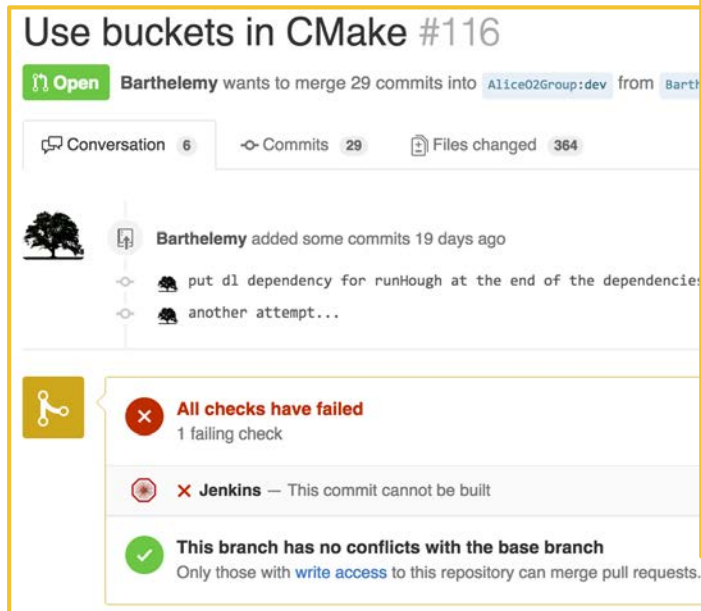
- Detector read-out
 - Design and simulation
 - Detector read-out development using the C-RORC
 - Data aggregation and formatting
- TPC software
- Core software development
 - Condition and Calibration Data Base (CCDB)
 - Data Quality Control (QC)
 - Control, Configuration and Monitoring (CCM)
 - FLP Prototype
- **Software process**
 - Data storage performance tests
 - CR0 procurement
 - Milestones and schedule

CWG11: Software process (1)

Focus on enabling O2 developers to start working now
More refinements added as we go.



- Infrastructure for testing all GitHub pull requests to O2 with Jenkins
- Interface to add new O2 projects to JIRA
- Defined a CMake example directory organization with tests too



Use buckets in CMake #116

Open Barthelemy wants to merge 29 commits into AliceO2Group:dev from Barth

Conversation 6 Commits 29 Files changed 364

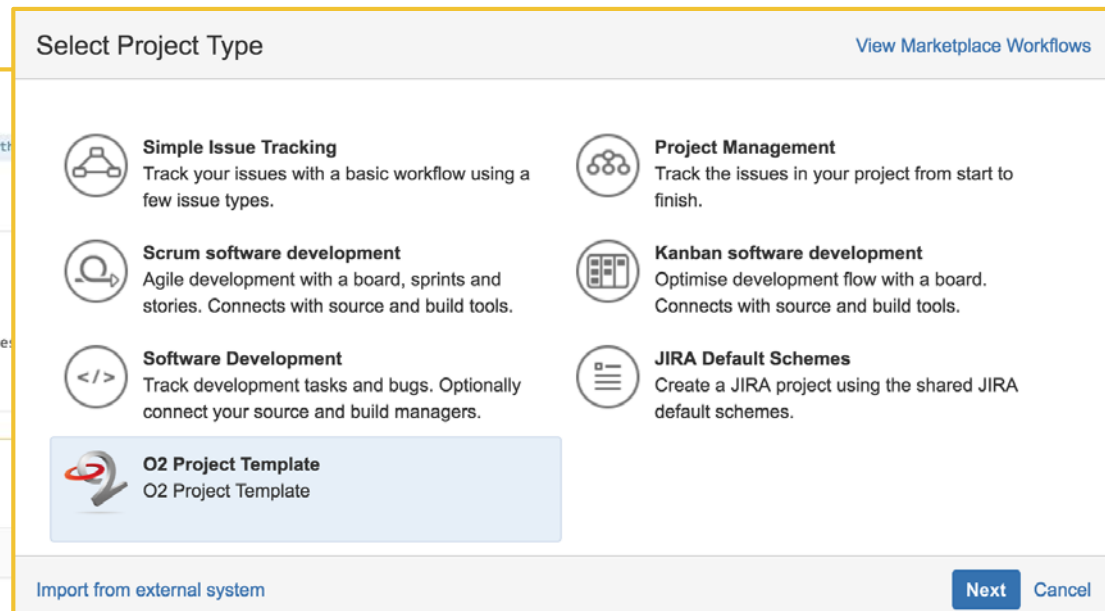
Barthelemy added some commits 19 days ago

- put dl dependency for runHough at the end of the dependencies
- another attempt...

All checks have failed
1 failing check

Jenkins — This commit cannot be built

This branch has no conflicts with the base branch
Only those with [write access](#) to this repository can merge pull requests.



Select Project Type [View Marketplace Workflows](#)

- Simple Issue Tracking**
Track your issues with a basic workflow using a few issue types.
- Project Management**
Track the issues in your project from start to finish.
- Scrum software development**
Agile development with a board, sprints and stories. Connects with source and build tools.
- Kanban software development**
Optimise development flow with a board. Connects with source and build tools.
- Software Development**
Track development tasks and bugs. Optionally connect your source and build managers.
- JIRA Default Schemes**
Create a JIRA project using the shared JIRA default schemes.

O2 Project Template
O2 Project Template

[Import from external system](#) **Next** Cancel

CWG11: Software process (2)



- **Definition of the policies of the Git repositories: done**

Automatic tool where CERN users can map their GitHub account to their CERN account by simply clicking on a link.

- **Project management and progress report in JIRA: done**

Currently being used. Progress report visualization via Gantt charts.

- **Automatic module testing with Jenkins: done**

Project CMake structure enable developers to add tests.

- **Enforcing code quality and formatting rules: in progress**

Evaluation of our first tool, SAS (github.com/dpiparo/SAS) completed: it does not look either mature or easy to maintain yet. Currently evaluating CodeChecker (github.com/Ericsson/codechecker), considered by GSI for FairRoot too.

Outline

- Detector read-out
 - Design and simulation
 - Detector read-out development using the C-RORC
 - Data aggregation and formatting
- TPC software
- Core software development
 - Condition and Calibration Data Base (CCDB)
 - Data Quality Control (QC)
 - Control, Configuration and Monitoring (CCM)
 - FLP Prototype
- Software process
- **Data storage performance tests**
- CR0 procurement
- Milestones and schedule

CWG1 : Data storage requirements

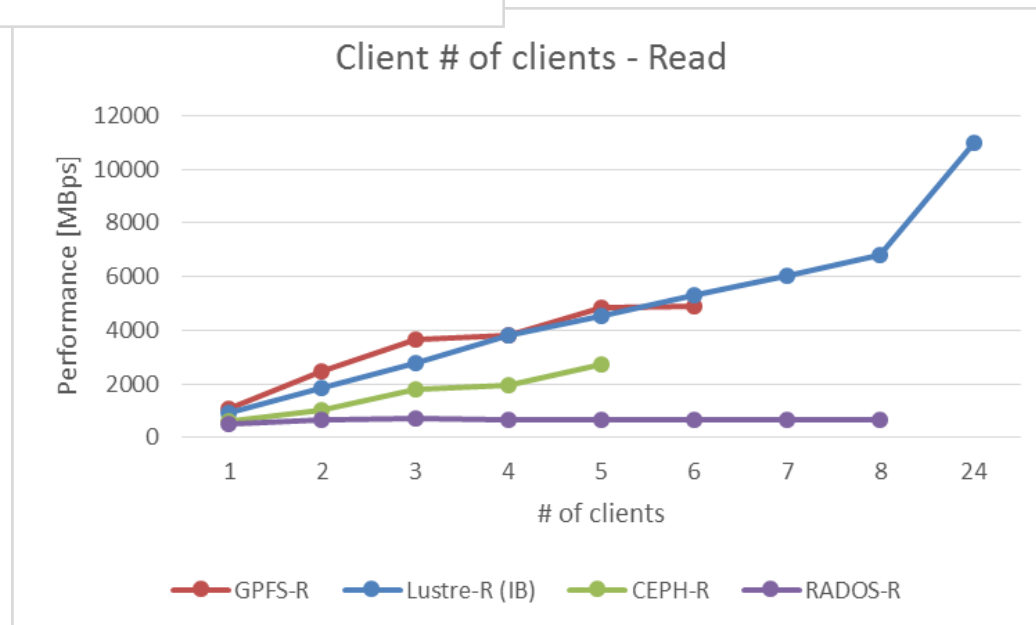
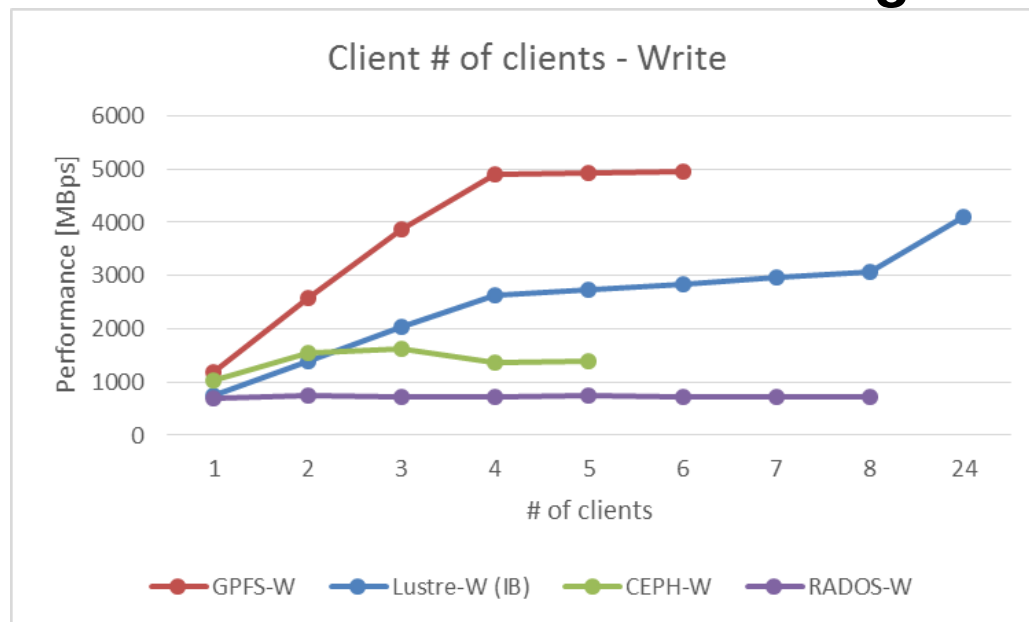


- Review of the requirements:
 - High-Capacity, High-Throughput file system
 - ~60 PB, ~440 GB/s, 10^{10} files
 - Number of clients: ~2000
- Usage Pattern:

File Type	Read Rate (files/s)	Write Rate (files/s)	Read b/w (GB/s)	Write b/w (GB/s)
CTF	100	50	200	100
ESD	100	100	30	30
AOD	200	200	40	40
Total	400	350	270	170

- Several solutions (EOS, Lustre, GPFS) exist and performance tests are going on

Performance tests of data storage solutions



Outline

- Detector read-out
 - Design and simulation
 - Detector read-out development using the C-RORC
 - Data aggregation and formatting
- TPC software
- Core software development
 - Condition and Calibration Data Base (CCDB)
 - Data Quality Control (QC)
 - Control, Configuration and Monitoring (CCM)
 - FLP Prototype
- Software process
- Data storage performance tests
- **CR0 procurement**
 - Milestones and schedule



CWG12 : CR0 procurement

- Following the standard CERN procedure
 - Market Survey (MS): questionnaire received from 18 firms out of which 8 were qualified
 - Invitation to Tender (IT) : Supply, Installation and Commissioning of Container Data Centres (CDC)
- Technical specifications prepared with LHCb and the Neutrino Platform (NP) who have similar needs for the same period.

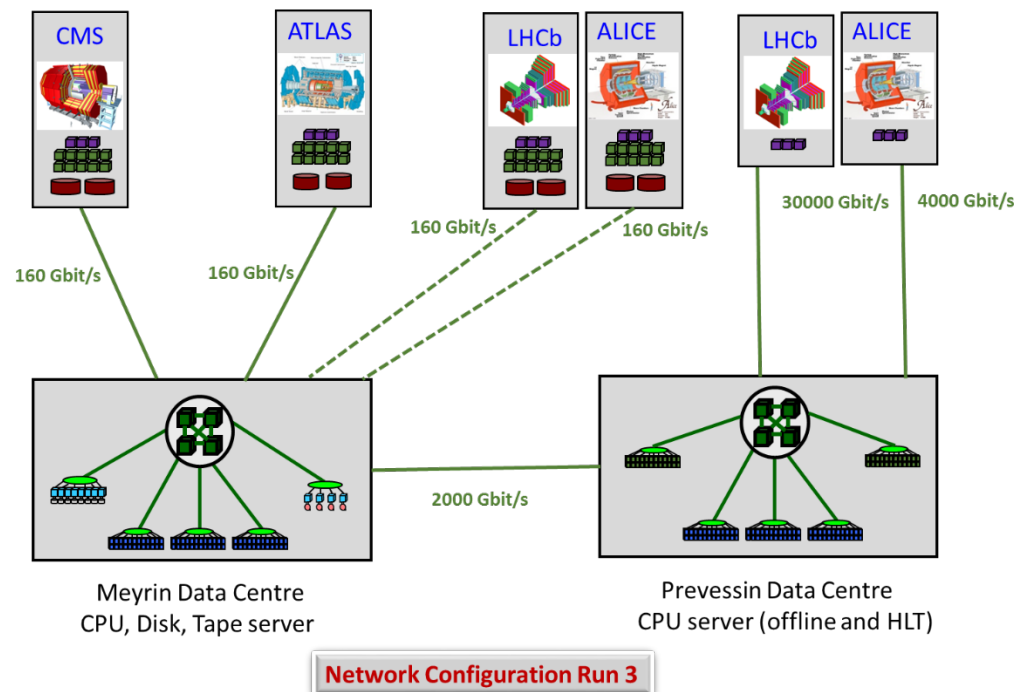
	Number of racks units	Total Power	Footprint	Delivery
	(U)	(MVA)	(m ²)	
ALICE	2300	2.1	27 x 24	Q1 17 - Q3 19
LHCb	4000	2.0	35 x 16	Q3 18 – Q1 21
NP	650	0.4	22 x 9	Q1 17

- Schedule
 - 09-Sep: Publication of the IT. Publication time: 5 weeks
 - 14-Oct: papers have to be ready for the Finance Committee
 - 14-Dec: CERN Finance Committee



Common Data Center in Preveessin

- An alternative to the CR0 at P2 has been proposed by CERN
- New common data center in Preveessin regrouping some IT services and possibly the HLT farms of ALICE and LHCb
- More detailed proposal expected at CERN by mid-October



Outline

- Detector read-out
 - Design and simulation
 - Detector read-out development using the C-RORC
 - Data aggregation and formatting
- TPC software
- Core software development
 - Condition and Calibration Data Base (CCDB)
 - Data Quality Control (QC)
 - Control, Configuration and Monitoring (CCM)
 - FLP Prototype
- Software process
- Data storage performance tests
- CR0 procurement

➤ Milestones and schedule

Milestones (1/3)

V1.4 (Sep-16)	2015	2016	2017	2018	2019	2020
O2 TDR approved						
Software process (6/16)						
DDS release (6/16)						
CCDB prototype (12/16)						
DQM prototype (12/16)						
CCM prototype (12/16)						
First ALFA release (6/17)						
Detector read-out DDL+GBT CRORC (12/16)						
Detector read-out GBT CRU (6/17)						

CCM: Control, Configuration & Monitoring
CCDB: Condition & Calibration Data Base

DDS: Dynamic Distribution System
DQM: Data Quality Monitoring

Milestones (2/3)

V1.4 (Sep-16)	2015	2016	2017	2018	2019	2020
Test time frame (6=>12/16)			😊			
ITS and TPC in simulation framework (6/17)			😊			
Reconstruction for ITS and TPC (6/17)			😊			
Simulation for all detectors in O2 (12/17)				😊		
Calibration framework prototype (12/17)				😊		
Reconstruction for all detectors in O2 (6/18)				😊		
Simulation challenge (6/19)					😊	
10% data challenge (12/19)						😊

Milestones (3/3)

V1.4 (Sep-16)	2015	2016	2017	2018	2019	2020
Definition of cooling system + EDR (3/17)			😊			
CR0 infrastructure ready (9/18)				😊		
Detector read-out (FLP) procurement (3/18)				😊		
Detector Read-out (FLP) installation (1/19)					😊	
Procurement 10% data processing and storage (12/18)				😊		15 months contingency
Installation 10% data processing and storage (9/19)					😊	pp data taking →
Procurement 90% data processing and storage (3/20)						😊 10 months contingency
Installation 90% data processing and storage (11/20)						😊 PbPb data taking →
LS2					◆	◆

O² Schedule

- CWG1 - WBS 7 ● CWG8 - WBS 15
- CWG2 - WBS 8 ● CWG9 - WBS 16
- CWG3 - WBS 10,11 ● CWG10 - WBS 17
- CWG4 - WBS 12 ● CWG11 - WBS 9
- CWG5 - WBS 18 ● CWG12 - WBS 19,20
- CWG6 - WBS 14 ● CWG13 - WBS 12
- CWG7 - WBS 13

