



SRO@ALICE

-

Streaming readout Workshop SRO-XI

Filippo Costa
for the ALICE O²/FLP

Filippo Costa



- Responsible for the detector readout activities in ALICE.
- ALICE Software Release Coordinator.
- Software and firmware developer for the readout card (PCIe card with FPGA).

INTRODUCTION

The presentation will describe the operations of the ALICE experiment during the first HI data taking in 2023.

(Description of the new ALICE readout concept was presented during past version of the workshop in 2021

<https://indico.phy.ornl.gov/event/112/contributions/478/>)

The first part describes the major ALICE O² components and how the experiment is controlled from a single central point.

The central part of the presentation gives details concerning dataflow, different processes running and system performance.

The last part of the presentation is dedicated to describe the major challenges in operating such a large system running at high data rate.



ALICE

MOTIVATION

ALICE

Technical Design Report

CERN-LHCC-2015-006

ALICE-TDR-019

June 2, 2015

Upgrade of the Online - Offline computing system

Technical Design Report

ALICE upgrade is based on the LHC running conditions after **LS2** which will deliver Pb–Pb collisions at up to $L = 6 \cdot 10^{27} \text{ cm}^{-2} \text{ s}^{-1}$, corresponding to an interaction rate of **50kHz**.

In order to keep up with the 50kHz interaction rate, the TPC will also require the implementation of a continuous read-out process to deal with event pile-up and avoid trigger-generated dead time.

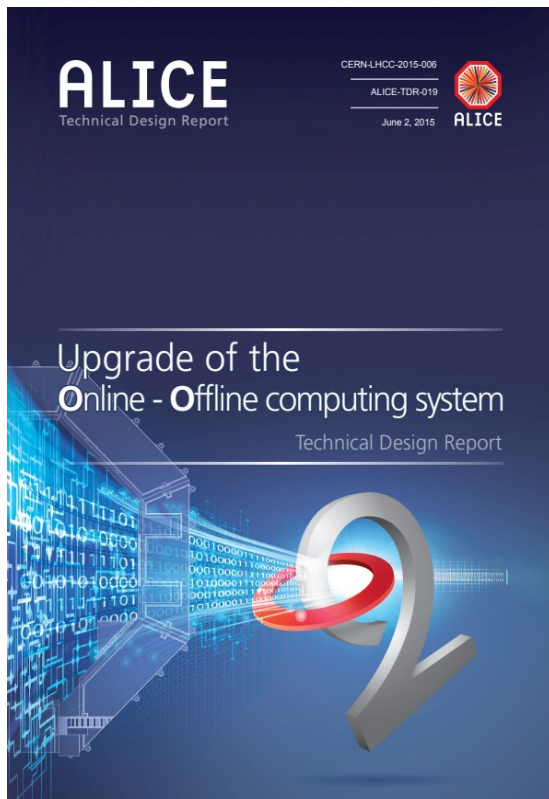
The resulting data throughput from the detector has been estimated to be greater than **3TB/s for Pb–Pb** events, several orders of magnitude more than in Run 1/2.



Last update: April 2023



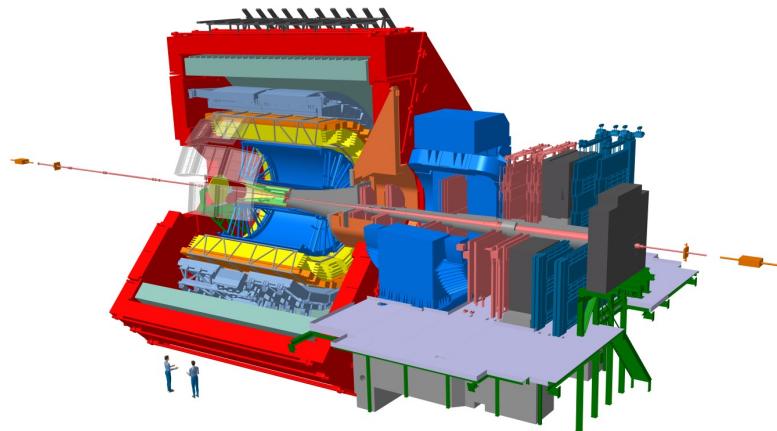
ALICE O² UPGRADE for the LHC RUN 3



A new **ALICE Online and Offline (O²)Computing** has been developed, the **ALICE O²**.

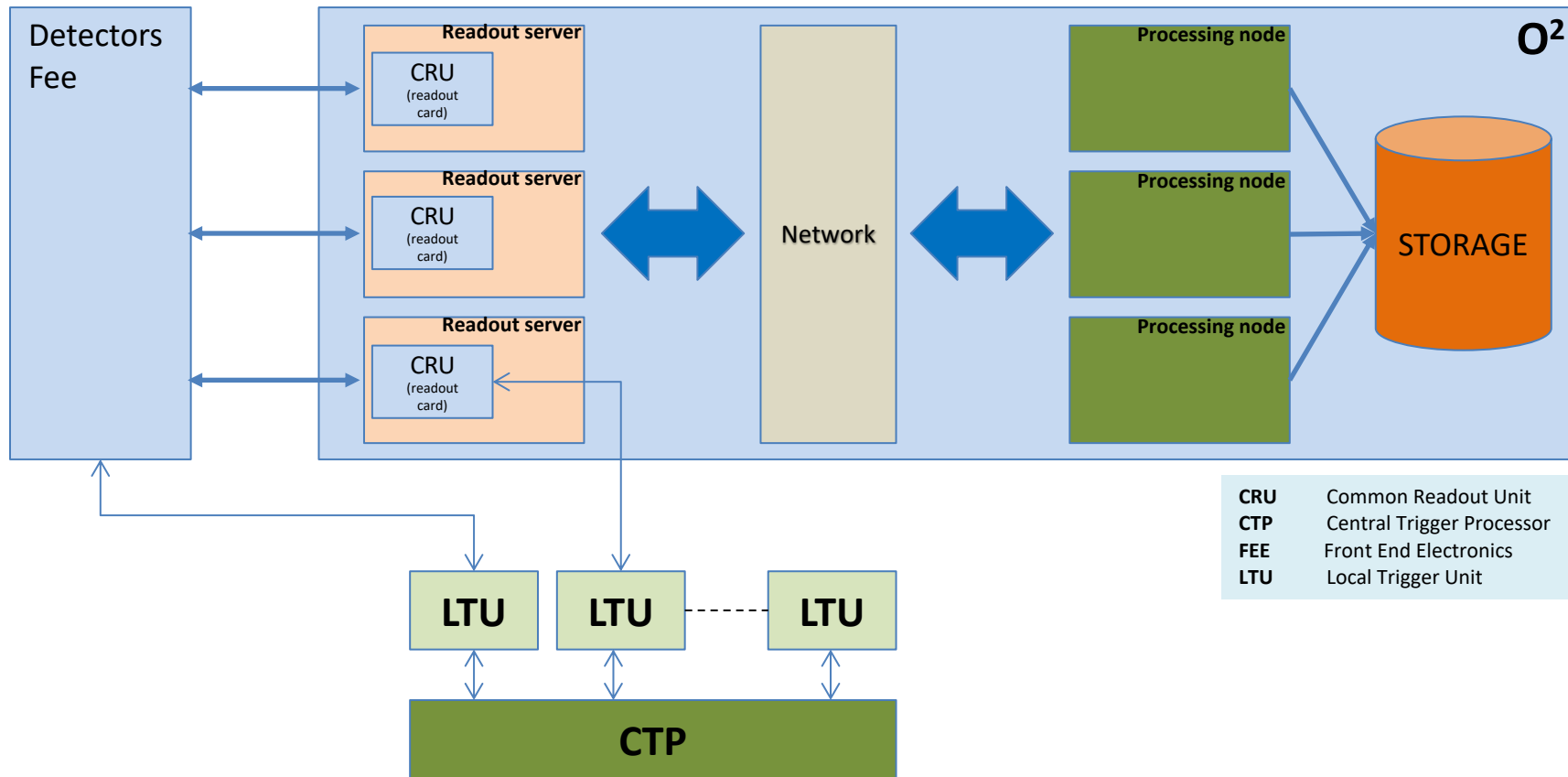
The new O² facility provides:

- Continuous readout.
- Synchronous and asynchronous reconstruction.
- Two different categories of computing nodes, corresponding to the two data aggregation steps.
- GPU data processing.



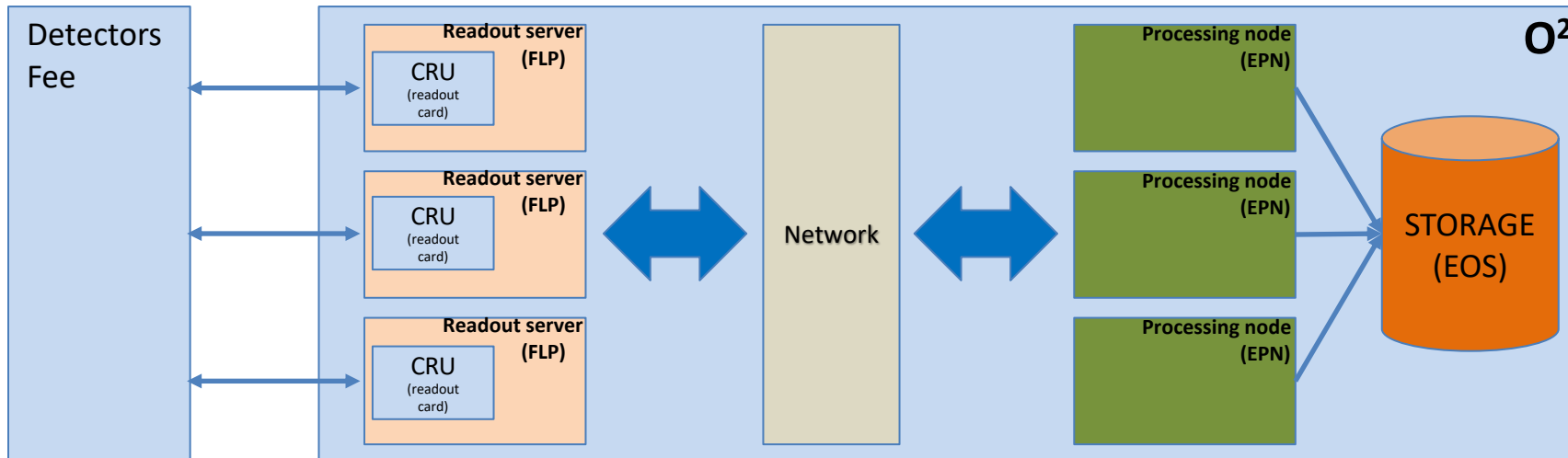


ALICE O² FARM





ALICE O² FARM



3.5 TB/s

data compression (770 GB/s)

170 GB/s

- 8000 links connect the detectors to O² farm. Total data rate > 3 TB/s
- 200 FLPs receive data from the detectors FEE. Total data rate ~770 GB/s
 - 500 readout cards.
- 350 EPN collect and store data on EOS. Total data rate 170 GB/s

CRU Common Readout Unit
FLP First Level Processors
EPN Event Processing Nodes
EOS EOS Open Storage



ALICE

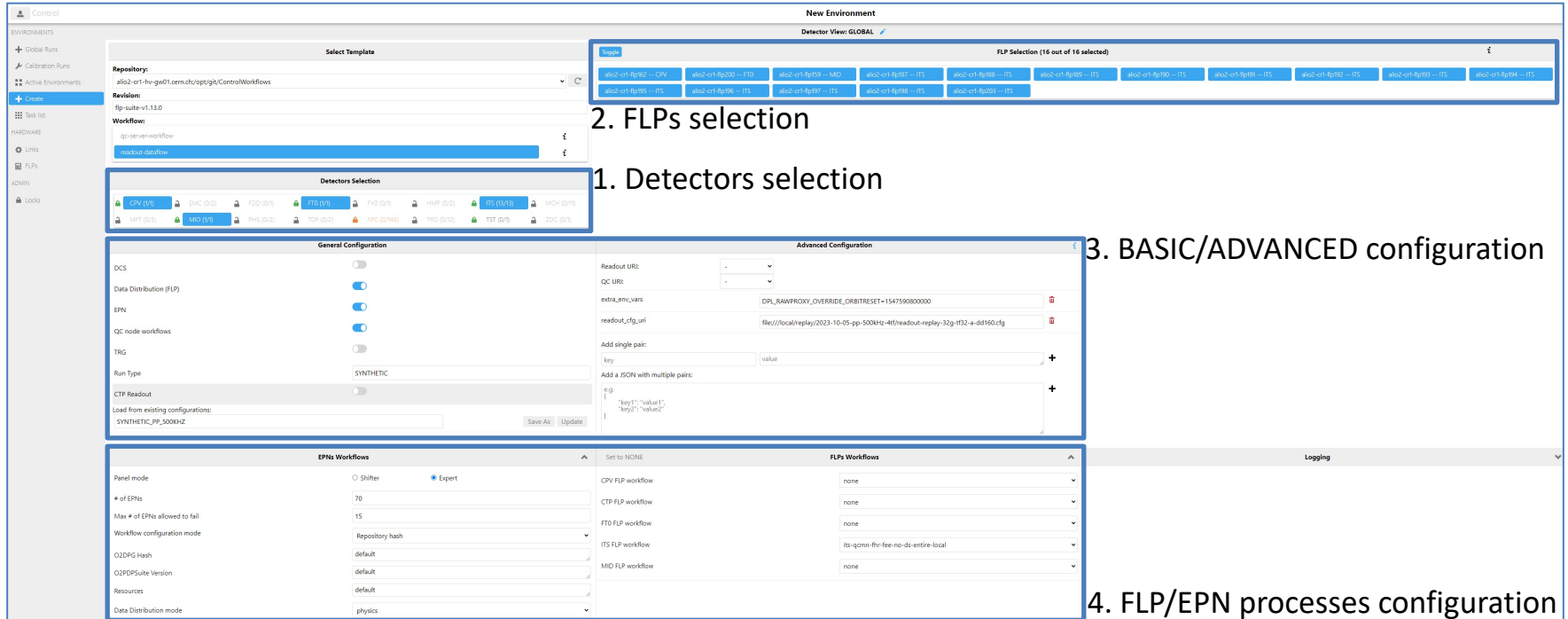
CHALLENGE ACCEPTED



CHALLENGES:

- Control from a central system
- DATA processing (no backpressure)
- DATA quality (must be good)
- High efficiency

AliECS global overview



The screenshot displays the AliECS Control interface for configuring a new environment. The interface is divided into several sections:

- Left Sidebar:** Contains navigation options such as "Global Runs", "Calibration Runs", "Active Environments", "Create", "Data View", "Hardware", "Links", "FLPs", "Admin", and "Locks".
- Top Bar:** Shows "New Environment" and "Detector View: GLOBAL".
- Repository Section:** Displays the repository path "alic2-crl-1v-gm01.cern.ch/opt/glu/ControlWorkflows" and the revision "fip-suite-v1.13.0".
- Workflow Section:** Shows the workflow "qc-server-workflow" and "readout-standalone".
- Detectors Selection:** A grid of detector icons including CPV, EMC, FDD, ITD, FV0, HMP, ITS, MCH, SFT, MID, PHD, TPC, TRD, and ZDC.
- FLP Selection:** A row of 16 FLP selection buttons, all of which are highlighted in blue, indicating they are selected.
- General Configuration:** Includes settings for DCS, Data Distribution (FLP), EPN, QC node workflows, TRG, Run Type (set to SYNTHETIC), and CTP Readout.
- Advanced Configuration:** Contains fields for Readout URI, QC URI, extra_env_vars, and readout_cfg_uri, along with a JSON editor for key-value pairs.
- EPNs Workflows:** Shows configuration for EPN mode (Shifter or Expert), number of EPNs (70), and repository hash.
- FLPs Workflows:** A list of workflow assignments for various FLPs, such as "none" for CPV, CTP, and MID, and "its-qcmn-flr-fee-no-ds-entire-local" for ITS.
- Bottom Right:** A "Logging" section.

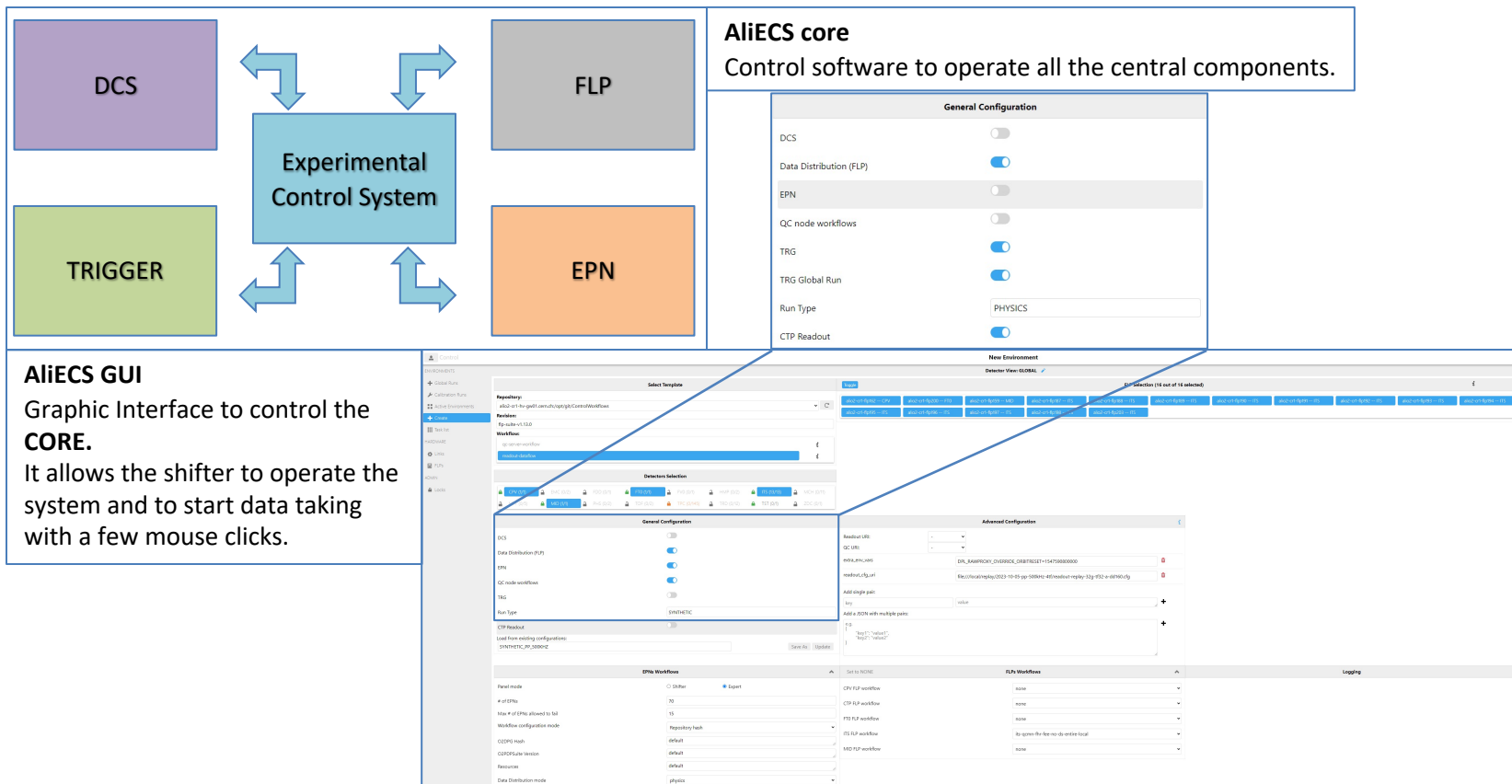
2. FLPs selection

1. Detectors selection

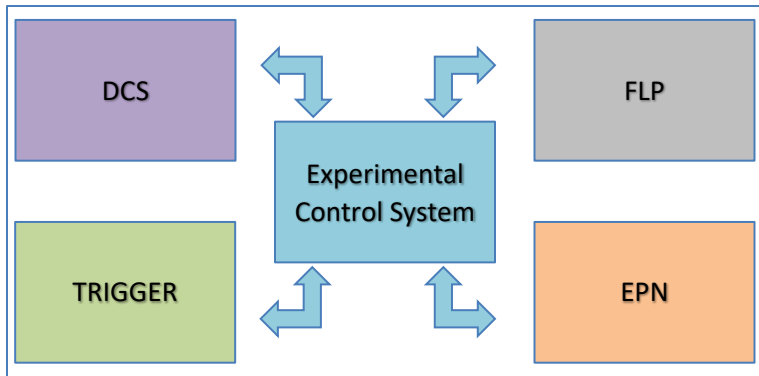
3. BASIC/ADVANCED configuration

4. FLP/EPN processes configuration

ALICE CONTROL SYSTEM (AliECS)



ENVIRONMENT CRATE/CONFIGURE



The ENVIRONMENT creation sets:

- list of detectors included in data taking,
- list of FLPs,
- number of EPNs,
- configuration of processes running on FLP/EPN.

All the operations are executed in parallel.

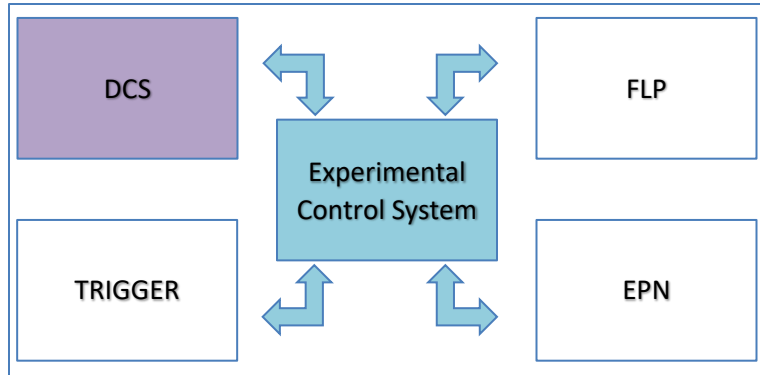
DCS : Prepare For Run. Detectors are configured via DCS.

TRIGGER : the list of detectors is sent to CTP to prepare the trigger configuration.

FLP : memory is allocated, readout cards are configured for data taking.

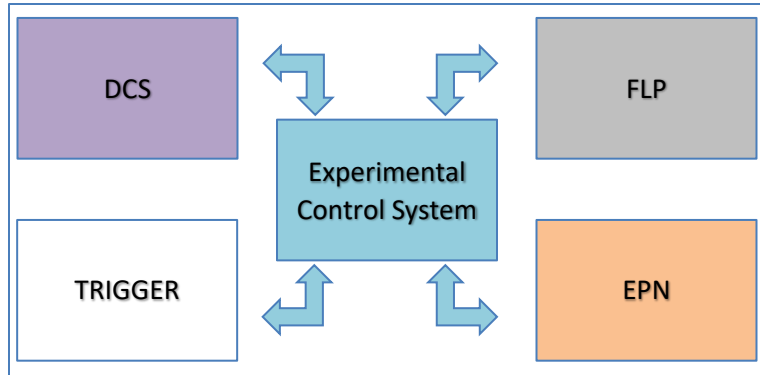
EPN : memory is allocated, detectors processes are started.

START (DCS SOR)



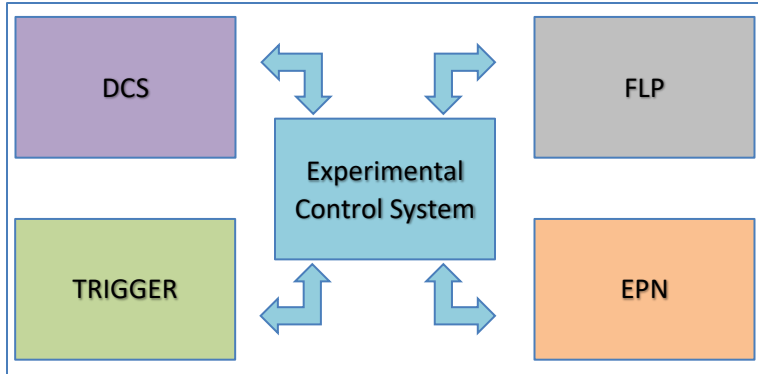
1. **DCS** : DCS SOR. Detectors that don't have PFR are configured now

START (FLP EPN READY)



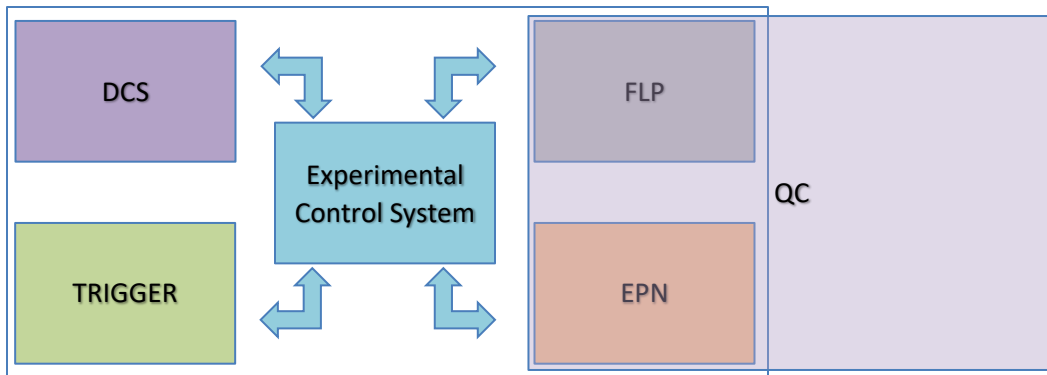
1. **DCS** : DCS SOR. Detectors that don't have PFR are configured now
2. **FLP** : DMA is enabled. **EPN** : ready to receive data from FLP

START (TRIGGER SOR)



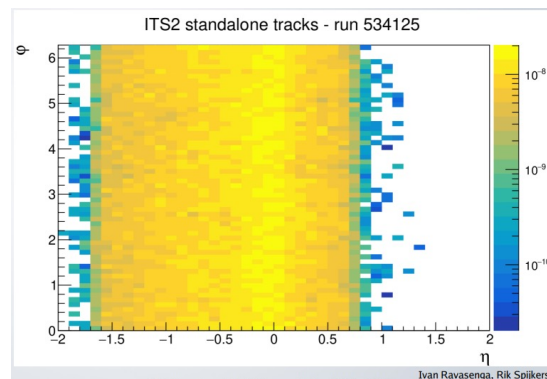
1. **DCS** : DCS SOR. Detectors that don't have PFR are configured now
2. **FLP** : DMA is enabled. **EPN** : ready to receive data from FLP
3. **TRIGGER** : Start of RUN trigger is sent to all the detectors

DATA QUALITY CONTROL?



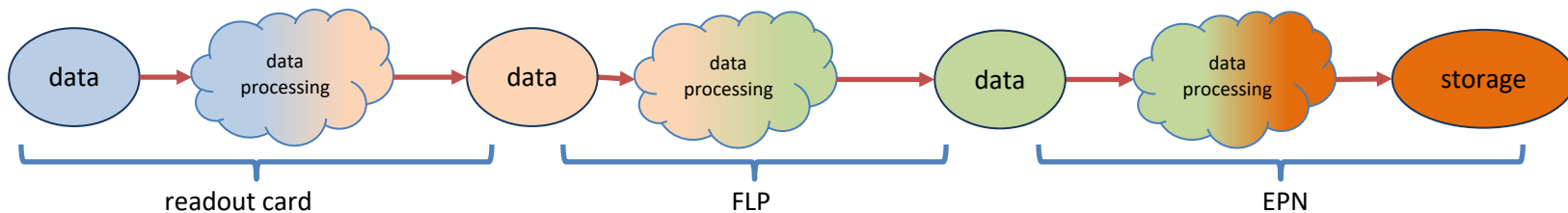
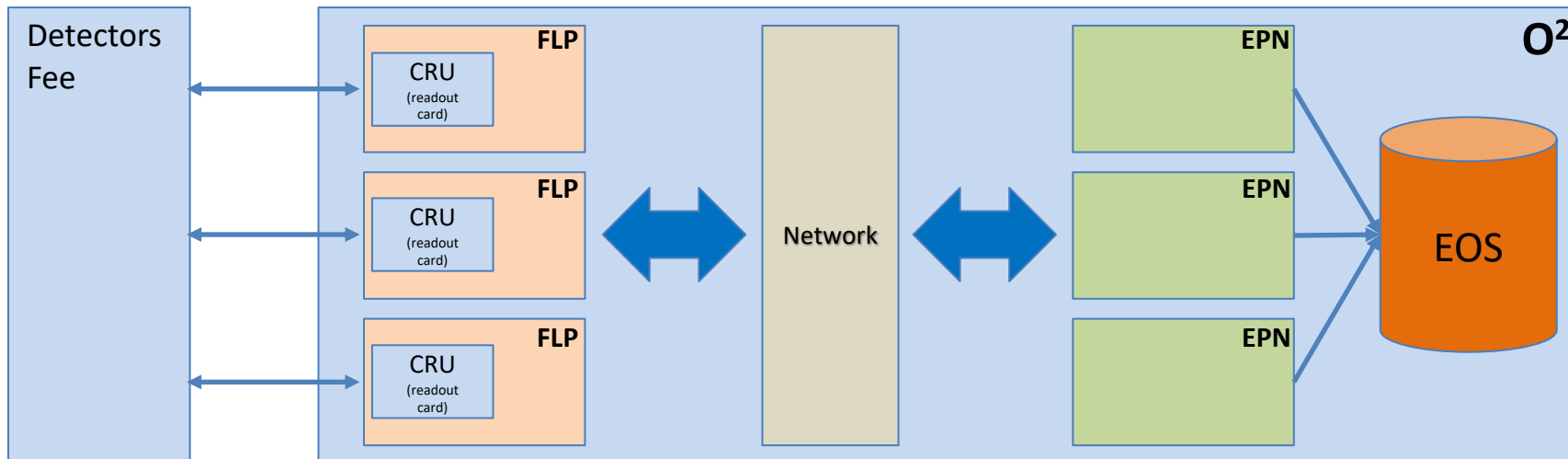
QC is controlled by ECS and it is connected to both FLP and EPN.

It receives data whenever a new RUN is started and it provides detector specific information on the data quality.



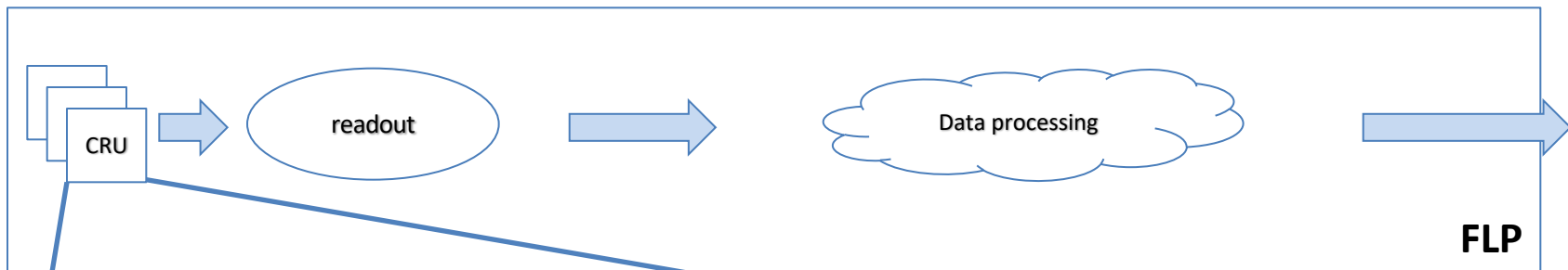


DATA FLOW- from detector to disk

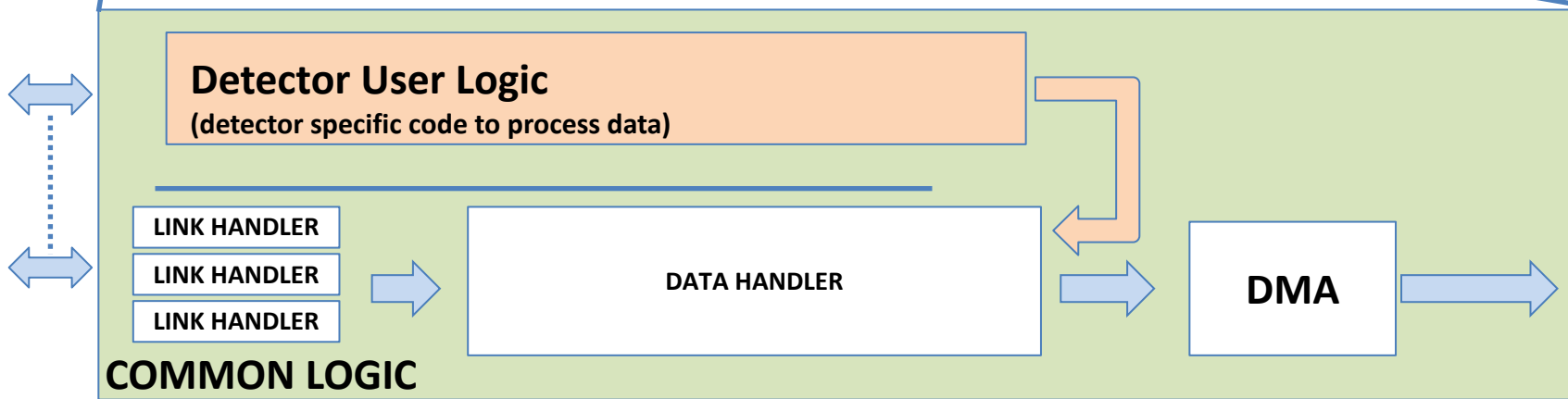




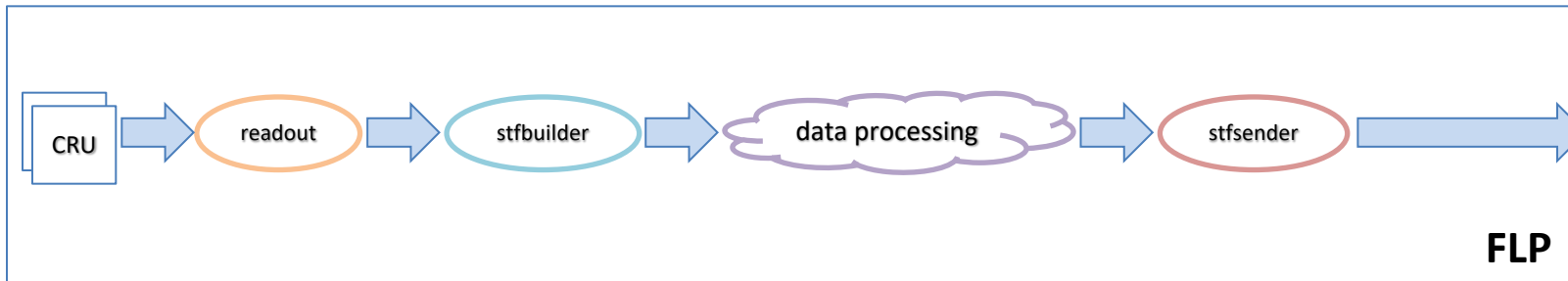
DATA PROCESSING (CRU)



CRU FPGA



DATA PROCESSING (FLP)



alfo2-cr1-flp178

Name	PID	Locked	Status	State
readout	4449	🔒	ACTIVE	CONFIGURED
stfbuilder	4450	🔒	ACTIVE	CONFIGURED
stfsender	4451	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-internal-dpl-clock	4457	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-readout-proxy	4463	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_10	4470	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_11	4477	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_12	4487	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_13	4495	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_14	4509	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_15	4508	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_16	4517	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_17	4522	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_18	4527	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_19	4557	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_110	4565	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_111	4572	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_112	4580	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_113	4586	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-tof-compressor-0_114	4603	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-Dispatcher	4614	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-qq-task-TOF-TaskRaw	4655	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-TOF-TaskRaw-proxy	4680	🔒	ACTIVE	CONFIGURED
jit-a6368d3a765462d9c2db1085fb5e5a128dfd9e1b-dpl-output-proxy	4711	🔒	ACTIVE	CONFIGURED

List of processes varies from detector to detector

DATA PROCESSING (EPN)

2j89yzFY7j - CONFIGURED

InfoLogger FLP InfoLogger EPN

General Information

ENV Created: 11/28/2023, 10:18:17 AM
 Transitioning: CONFIGURED
 State: SYNTHETIC
 Run Type: readout-dataflow
 RUN Started: Global:
 RUN Ended:
 Template:
 Global:

Components

FLPs: 199
 Detectors: CPU EMC FDD FV0 FV0 HMP ITS MCH MFT MID PHS TOF TPC TRD ZDC
 DCSS: OFF
 Data Distribution (FLP): ON
 EPNs: 50
 TRG: OFF
 CTP Readout: OFF
 CDC: -
 DDS: RUNNING
 DDS Session ID: e642bc6-288e-406f-8348-93a8793686d0

EPN

READY

hosts: 49

19425/19425

Tasks Summary

FLP	# hosts: 199	QC Nodes	# hosts: 15	CTP Readout	# hosts: 0	EPN	# hosts: 49
CONFIGURED	2249/2249	CONFIGURED	360/360			READY	19425/19425
ACTIVE	2249/2249	ACTIVE	360/360				

FLP Tasks by Detector(s) Summary

CPV	# hosts: 5	MCH	# hosts: 55	ZDC	# hosts: 12	MFT	# hosts: 80	EMC	# hosts: 10	FDD	# hosts: 5	MID	# hosts: 5	PHS	# hosts: 10	ET0
CONFIGURED	5/5	CONFIGURED	55/55	CONFIGURED	12/12	CONFIGURED	80/80	CONFIGURED	10/10	CONFIGURED	5/5	CONFIGURED	5/5	CONFIGURED	10/10	CONFIGURED
ACTIVE	5/5	ACTIVE	55/55	ACTIVE	12/12	ACTIVE	80/80	ACTIVE	10/10	ACTIVE	5/5	ACTIVE	5/5	ACTIVE	10/10	ACTIVE

TOF	# hosts: 52	HMP	# hosts: 22	TPC	# hosts: 1728	ITS	# hosts: 195	TRD	# hosts: 60
CONFIGURED	52/52	CONFIGURED	22/22	CONFIGURED	1728/1728	CONFIGURED	195/195	CONFIGURED	60/60
ACTIVE	52/52	ACTIVE	22/22	ACTIVE	1728/1728	ACTIVE	195/195	ACTIVE	60/60

Path

main/RecoGroupM50/RecoCollectionM50_0/qc-task-MFT-MFTClusterTask_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-MFT-MFTAgncTask-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-MCH-Errors_reco_0

main/RecoGroupM50/RecoCollectionM50_0/trid-entropy-encoder_t1_reco_0

main/RecoGroupM50/RecoCollectionM50_0/muon-track-matcher_reco_0

main/RecoGroupM50/RecoCollectionM50_0/PHOSClusterizerSpec_reco_0

main/RecoGroupM50/RecoCollectionM50_0/TOF-TaskDigits_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-track-finder_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-TOF-TaskDigits_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-ITS-ITSDECODING_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-Errors-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-MCH-Digits_reco_0

main/RecoGroupM50/RecoCollectionM50_0/zdc-tdscalb-epn_reco_0

main/RecoGroupM50/RecoCollectionM50_0/trid-entropy-encoder_t2_reco_0

main/RecoGroupM50/RecoCollectionM50_0/CPV-PhysicsOnEPNs-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/internal-dp-injected-dummy-sink_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-Digits-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/h0-reconstructor_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mft-tracker_t1_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-MID-QcTaskMIDDigits_reco_0

main/RecoGroupM50/RecoCollectionM50_0/Dispatcher_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-data-decoder_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-tracker_t5_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mid-QcTaskMIDClust-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/zdc-digi_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-TOF-TaskDigits_reco_0

main/RecoGroupM50/RecoCollectionM50_0/CPVClusterizerSpec_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-tracker_t3_reco_0

main/RecoGroupM50/RecoCollectionM50_0/h0-reconstructor_reco_0

main/RecoGroupM50/RecoCollectionM50_0/phos-entropy-encoder_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-Fits-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/EMCALRawToCellConverterSpec_t2_reco_0

main/RecoGroupM50/RecoCollectionM50_0/MIDTracker_reco_0

main/RecoGroupM50/RecoCollectionM50_0/its-stf-decoder_t4_reco_0

main/RecoGroupM50/RecoCollectionM50_0/qc-task-MCH-Decoding_reco_0

main/RecoGroupM50/RecoCollectionM50_0/cpv-entropy-encoder_reco_0

main/RecoGroupM50/RecoCollectionM50_0/HMP-RawStreamDecoder_reco_0

main/RecoGroupM50/RecoCollectionM50_0/emcal-entropy-encoder_reco_0

main/RecoGroupM50/RecoCollectionM50_0/TRDTRACKLETTRANSFORMER_reco_0

main/RecoGroupM50/RecoCollectionM50_0/its-tracker_t2_reco_0

main/RecoGroupM50/RecoCollectionM50_0/hdd-reconstructor_reco_0

main/RecoGroupM50/RecoCollectionM50_0/TOFIntegrateClusters_reco_0

main/RecoGroupM50/RecoCollectionM50_0/TOFCluster_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mft-stf-decoder_t0_reco_0

main/RecoGroupM50/RecoCollectionM50_0/primary-vertexing_reco_0

main/RecoGroupM50/RecoCollectionM50_0/EMCALRawToCellConverterSpec_t1_reco_0

main/RecoGroupM50/RecoCollectionM50_0/MFT-MFTClusterTask-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-Decoding-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/its-stf-decoder_t3_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mch-Rofs-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/mft-stf-decoder_t1_reco_0

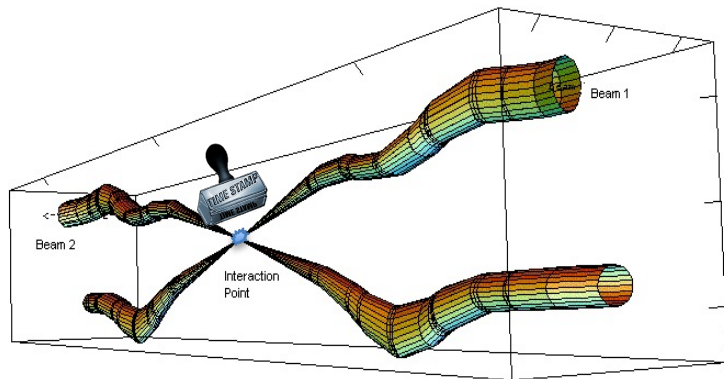
main/RecoGroupM50/RecoCollectionM50_0/MFT-MFTClusterTask-prony_reco_0

main/RecoGroupM50/RecoCollectionM50_0/h0-data-reader-dpl_reco_0

To be able to process data fast enough we had a processing farm consisting of a total of 350 EPN nodes and 2800 GPUs (Without GPUs, more than **2000 64-core servers** would be needed for online processing!). All the EPNs have the same list of processes as there is no DETECTOR-EPN specific, but there are some dedicated to CALIBRATION.

- PHYSICS
- CALIBRATION
- SYNTHETIC/REPLAY

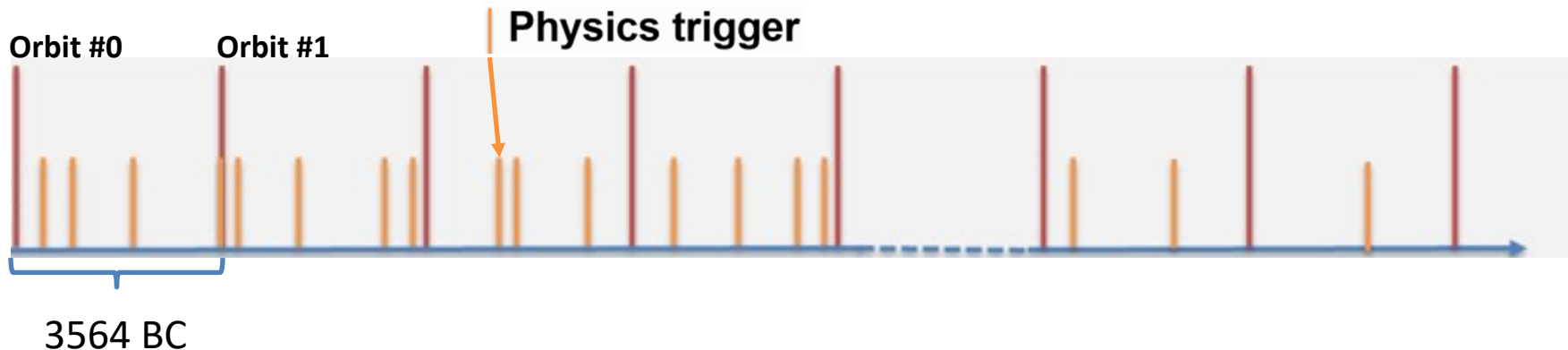
ALICE EVENT IDENTIFICATION



LHC has a global **MACHINE clock** that runs @ 40 MHz.
2 numbers are used to identify each interaction:

- **Bunch Crossing (BC)** : a number from 0 to 3563.
- **ORBIT** : a 32 bit number.

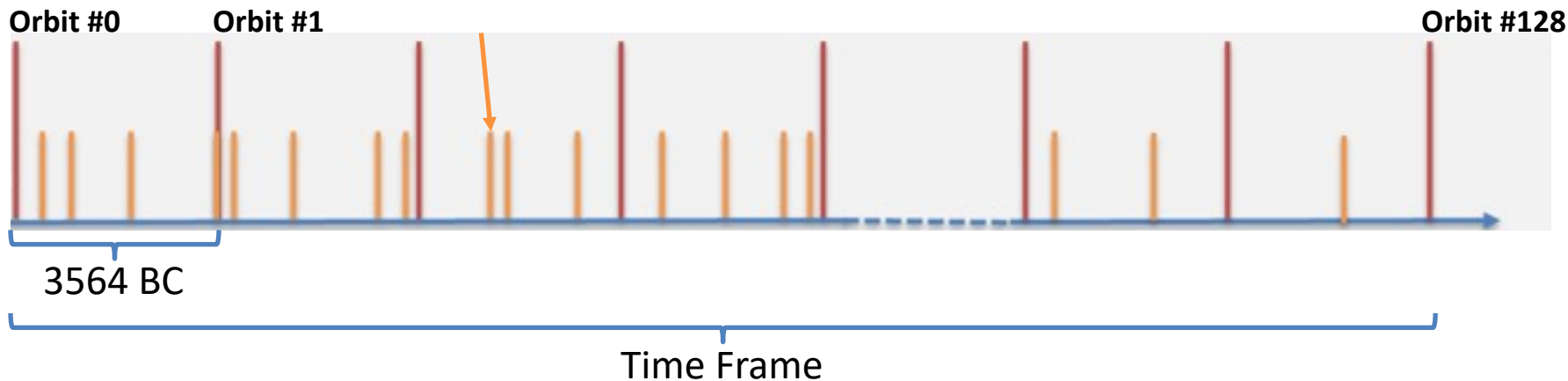
Every time BC overlaps, ORBIT+1.
The combination of the 2 gives a unique time stamp every ~4days





HB TRIGGER and TIME FRAME

LHC clock 40 MHz
3564 Bunch Crossing in 1 ORBIT
ORBIT rate ~ 10 KHz
Time Frame = 128 Orbits



TF length, why it is important

**LONGER TF, less statistic we lose,
SHORTER TF, less memory we need.**

- **Time Frame = 256 orbits**
 - This was done to optimize the data transfer and reduce the percentage of data at the boundary of each TF. Every TF reaches a different EPN.
- **Time Frame = 128 orbits**
 - EPN has 512 GB of RAM. We reduced the size as 256 required more memory than available. Even in this configuration we were on the edge of the memory (with only 30 GB of 512 GB left).
- **Time Frame = 32 orbits**
 - GPUs are processing individual time frames, using a TF or 32 ORBIT was the best choice for performance and memory usage.
 - Synchronous processing: EPN farm build for synchronous processing!
 - Asynchronous reprocessing : More detectors with significant computing contribution (more memory requirements)
 - Even with this configuration on the GRID sites we run out of memory (under investigation).

PERFORMANCE – how well did ALICE operate



A challenging year

We were getting ready for our HI commissioning:

- IP8 Inner Triplet incident on 17/07
- 7 weeks without beam during a crucial period to prepare the HI run
- No testing of detectors stability
- SW upgrades every week, but only partial validation with synthetic and cosmic runs

A challenge

We were getting

- IP8 Inner Trip
- 7 weeks with
- No testing of
- SW upgrades runs



ind cosmics



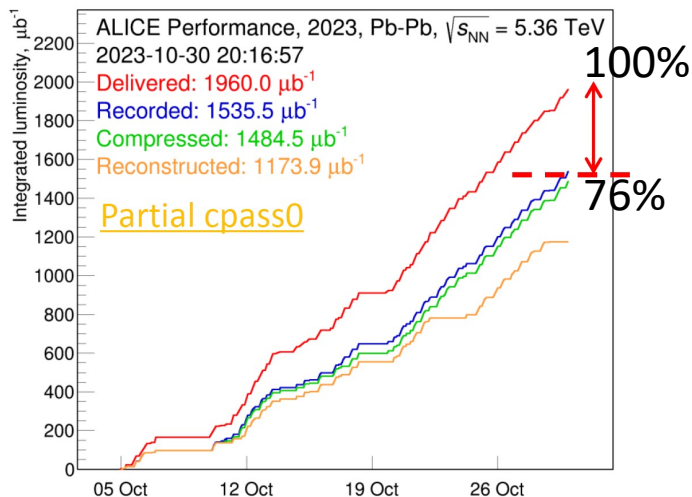
MAJOR CHALLENGE

Running an experiment with 10'000 links, ~500 servers, thousands of processes (more than 1000 in each EPN) comes with some technical challenges.

ALICE adopted a readout system without BUSY, so you can't slow down 3.5 TB/s and you don't want to lose data.

ALICE efficiency 76%

- HI run max IR 47 kHz.
- Data rate into EPN = 770 GB/s
- Data rate into STORAGE = 170 GB/s



24% of inefficiency is caused by runs stopped due to :

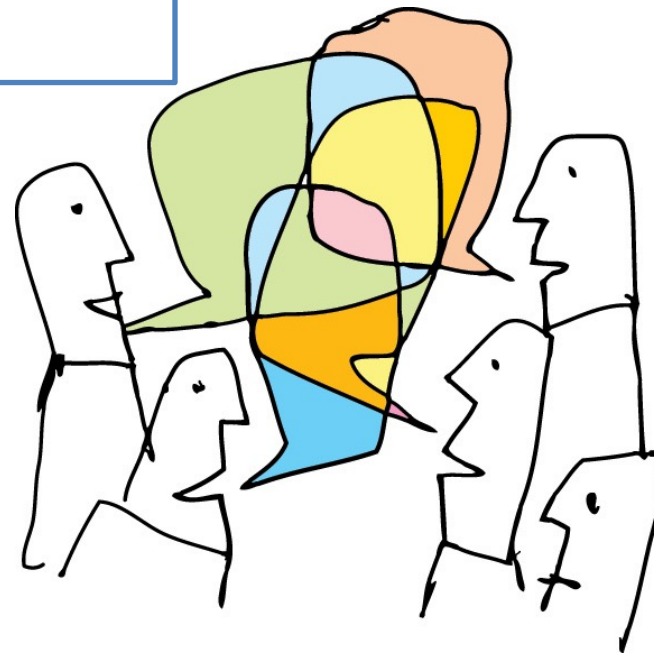
- user error,
- Detector problem (misconfiguration, SEU, ...),
- Process crashed,
- Clock phase changes,
- ... and more.

detector	% data stored on disk
CPV	0.00515878866767245
CTP	0.00882093634283997
EMC	0.0965574200869967
FDD	0.0127348827326441
FT0	0.0401928297161023
FV0	0.0149288995582678
HMP	0.019687999041066
ITS	3.94027096898005
MCH	1.5866397777917
MFT	1.24892675680083
MID	0.0404086086670641
PHS	0.0180259765773141
TOF	0.328620040146893
TPC	90.9539527948842
TRD	1.33096730644395
ZDC	0.354106013562415

Simple task, complex infrastructure

To achieve a simple task like “data taking”, ALICE O² runs many components (hardware and software), developed by different teams, that have to communicate with each other at the same time, in a well coordinated way.

This makes the system rather complex.





User eXperience

From the moment you click START, the user is flooded with information. Every process prints messages in the logging software, generating several hundreds of information per seconds.

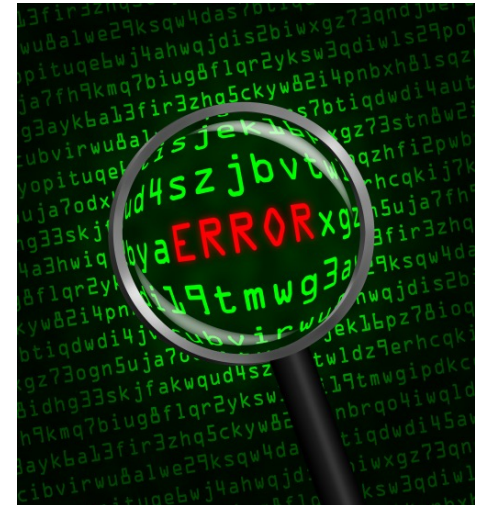
Currently it is not easy to have clear overview of the status of the system.



Error detection

There are several components that run together in a defined sequence. Interference, connection timeout, wrong software configuration result in an ENVIRONMENT in ERROR.

While this is clearly notified to the shifter, to identify the source of the problem is one of the major challenge even for the experts.



Error detection, fast reaction

The creation time of the ENVIRONMENT last several minutes (~7 minutes).

In 2023 was very important to identify possible source of errors as early as possible to reduce the number of environment creation to the minimum.

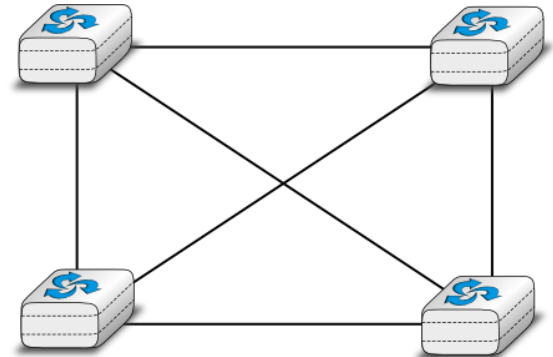
All the time spent in creating a new ENVIRONMENT is beam time lost.



Resource sharing

EPNs are equipped with powerful hardware for data processing (GPU). When they are not used in online for data taking, they are used in offline to process the data collected.

Currently the move of EPN from online to offline is a lengthy operation reducing the optimization of resource usage.

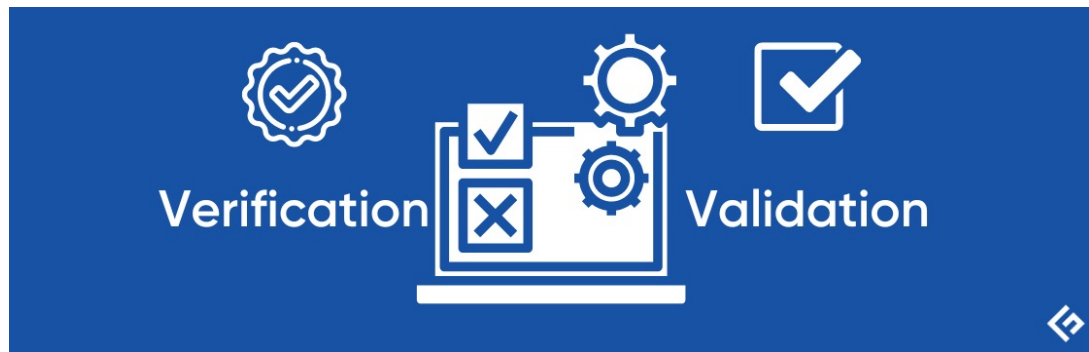


SW verification

Every software deployment brings in many new commits. Although we have different systems to test the software, the final verification is done in **PRODUCTION**.

During Stable Beam period it makes difficult to deploy new software and it requires coordination between different experts.

During HI in 2023 we spent a lot of effort to deploy only selected commits in time window where the BEAMs were not circulating.



CONCLUSIONS

- The O² framework worked well. ALICE could process data coming from the detectors during HI without introducing backpressure.
- ALICE was writing data into the storage at 170 GB/s.
- We are currently working to improve the software to have a system
 - Faster.
 - Better user experience.
 - More stable.