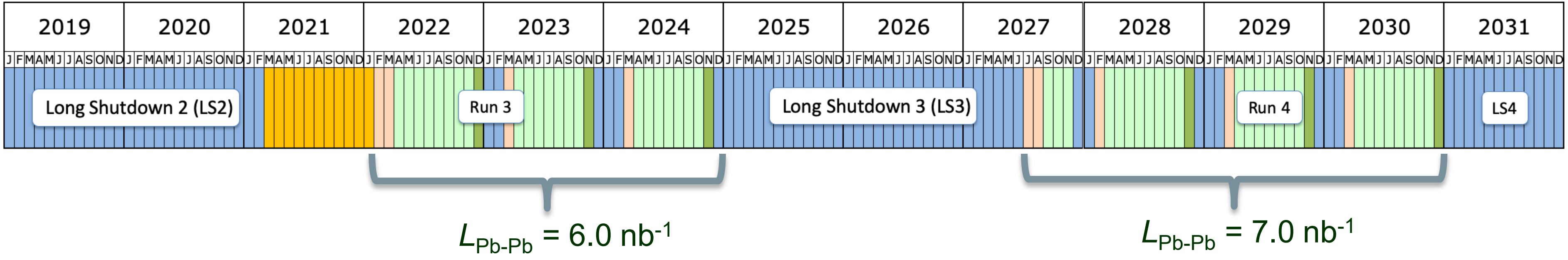
The ALICE O^2 computing model for Run 3 and 4

Stefano Piano
INFN sez. Trieste

ALICE

ALICE physics goals for Run 3 and 4



- Heavy-flavour mesons and baryons (down to very low p_T)
  - ⇒ mechanism of quark-medium interaction
- Charmonium states
  - ⇒ dissociation/regeneration as tool to study de-confinement and medium temperature
- Dileptons from QGP radiation and low-mass vector mesons
  - ⇒ χ-symmetry restoration, initial temperature and EOS
- High-precision measurement of light and hyper (anti-)nuclei
  - ⇒ production mechanism and degree of collectivity

CERN Yellow Report arXiv:1812.06772

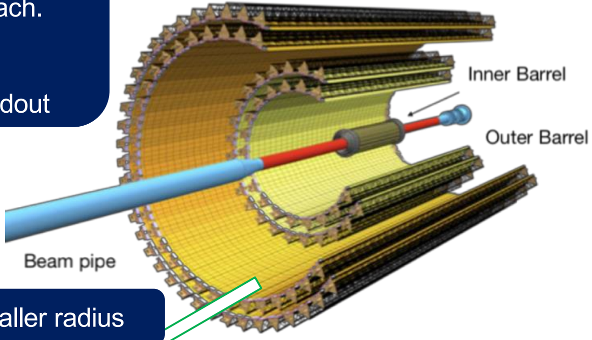
From triggered Run 2 to un-triggered Run 3 and 4

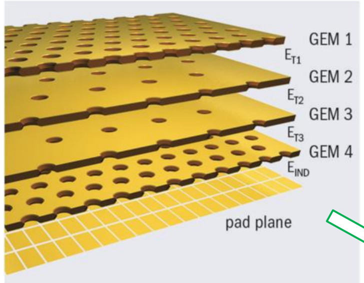
- In Run 2 ALICE operated at Pb-Pb interaction rates ~7-10 kHz (inspected ~1 nb^{-1}) with trigger rate < 1 kHz for Central Barrel
- **LHC plans to deliver 50 kHz Pb-Pb interaction rate after LS2**
- **ALICE plans for Run 3 & 4: collect 13 nb^{-1} of Pb-Pb collisions at 5 TeV** (of which 3 nb^{-1} with reduced magnetic field)
- Main limitations at these rates:
  - Principal tracking detector, TPC has ~90 μs drift time + at least ~200 μs gating grid to collect the ion backflow
  - Trigger rate limited to ~3 kHz (< 1 kHz accounting for bandwidth in Run 2)
  - At high multiplicities (dN/dη ~ 2000 + pile-up) very low S/B for rare probes:
    - ➢ Dedicated (HLT) trigger is not realistic
    - ➢ **Use continuous readout at least for TPC (no gating grid), increase bandwidth**
    - ➢ **Read out all events, store compressed data and inspect all events offline**

ALICE

ALICE HW upgrades

New Si Inner Tracker: 10 m^2 of MAPS with 29x27μm^2 pixel size
3 inner layers ~0.3% X0 each.
closer to the beam
+ 4 outer layers
50-500 kHz continuous readout

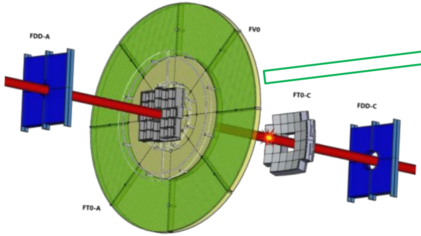


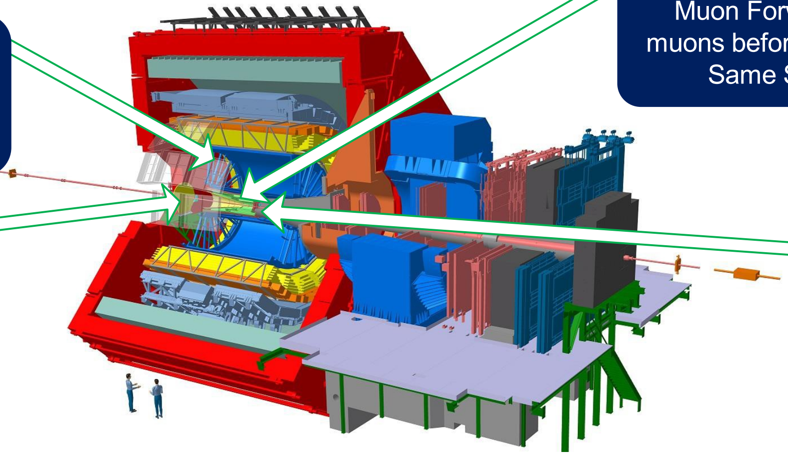


TPC readout → 4 layer GEM
(Intrinsic ion backflow ~99% blocking)
5 MHz continuous sampling

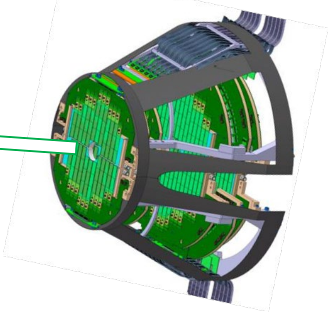
New beam pipe of smaller radius

Fast Interaction Trigger (FIT) detector
Scintillator (FV0, FDD) + Cerenkov (FT0) detectors to provide Min.Bias trigger for detectors with triggered R/O



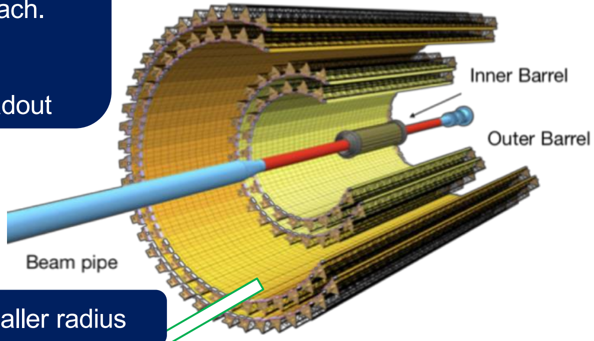


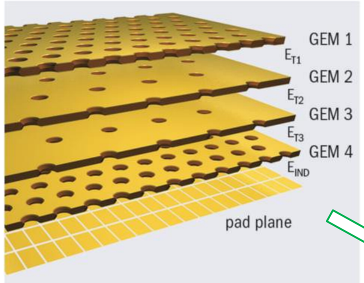
Muon Forward Tracker to match muons before and after the absorber.
Same Si chips as new ITS



ALICE HW upgrades

New Si Inner Tracker: 10 m^2 of MAPS with 29x27μm^2 pixel size
3 inner layers ~0.3% X0 each.
closer to the beam
+ 4 outer layers
50-500 kHz continuous readout



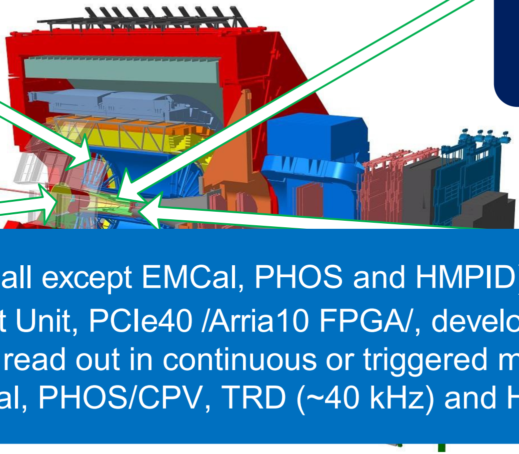


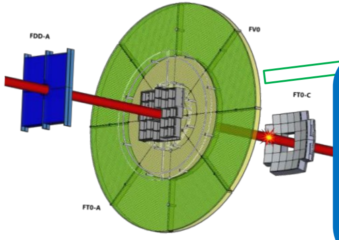
TPC readout → 4 layer GEM
(Intrinsic ion backflow ~99% blocking)
5 MHz continuous sampling

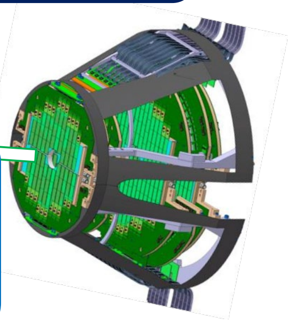
New beam pipe of smaller radius

Muon Forward Tracker to match muons before and after the absorber.
Same Si chips as new ITS

Fast Interaction Trigger (FIT) detector
Scintillator (FV0, FDD) + Cerenkov (FT0) detectors to provide Min.Bias trigger for detectors with triggered R/O

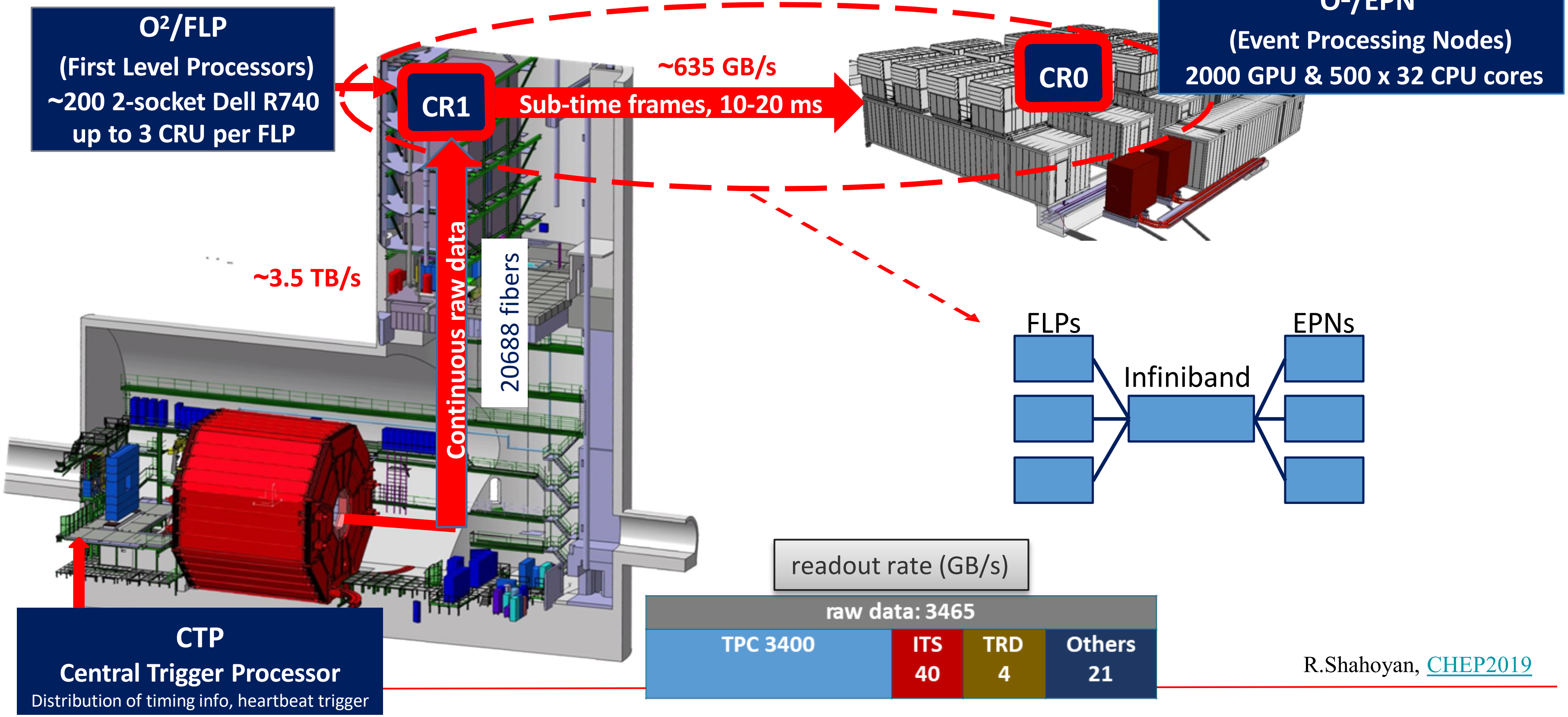






New readout (all except EMCal, PHOS and HMPID) via CRU
(Common Readout Unit, PCIe40 /Arria10 FPGA/, developed by LHCb)
Detectors can be read out in continuous or triggered modes, except triggered-only: EMCal, PHOS/CPV, TRD (~40 kHz) and HMPID (2.5 kHz)

ALICE Raw Data Flow in Run 3



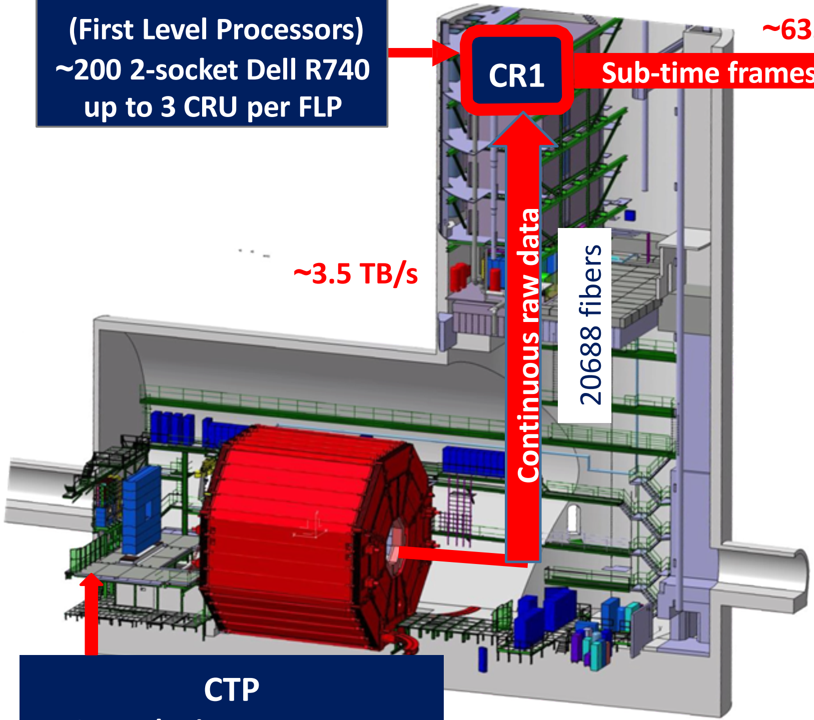
readout rate (GB/s)

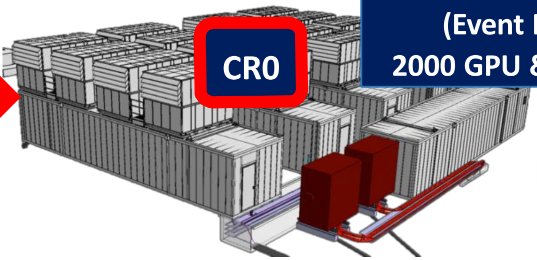
raw data: 3465			
TPC 3400	ITS 40	TRD 4	Others 21

R.Shahoyan, CHEP2019

ALICE Raw Data Flow in Run 3

O²/FLP
(First Level Processors)
~200 2-socket Dell R740
up to 3 CRU per FLP



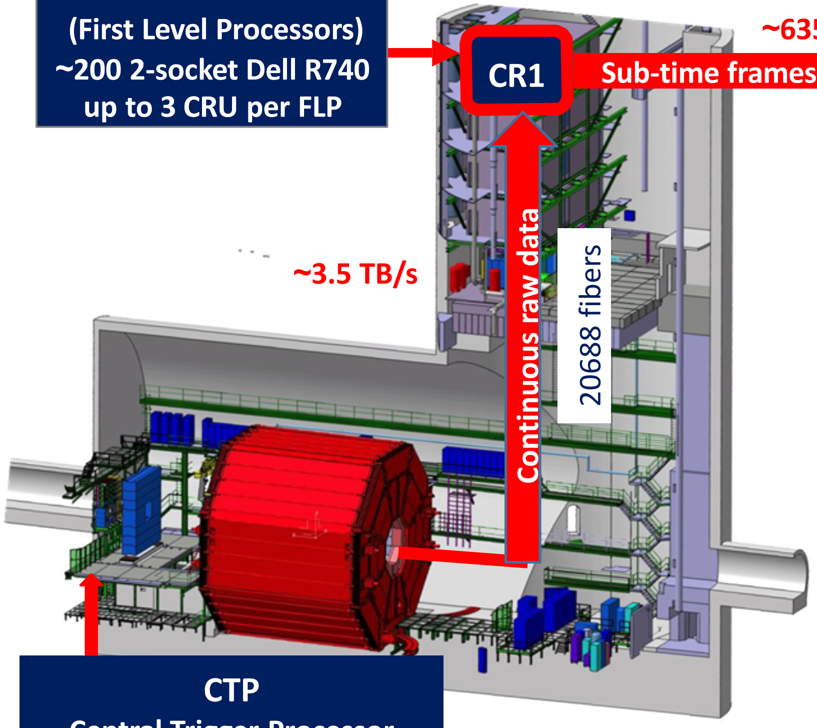


O²/EPN
(Event Processing Nodes)
2000 GPU & 500 x 32 CPU cores

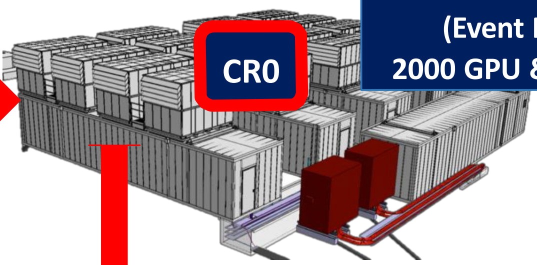
CTP
Central Trigger Processor
Distribution of timing info, heartbeat trigger

ALICE synchronous processing

O^2/FLP
(First Level Processors)
~200 2-socket Dell R740
up to 3 CRU per FLP



~635 GB/s
Sub-time frames, 10-20 ms



O^2/EPN
(Event Processing Nodes)
2000 GPU & 500 x 32 CPU cores

~100 GB/s
Compressed time frames
Calibration data

~80 PB
disk storage, 360 GB/s
JBOD, RAIN
(~25% redundancy)

~70 % CTF
~30 % CTF

Tier 0 archival

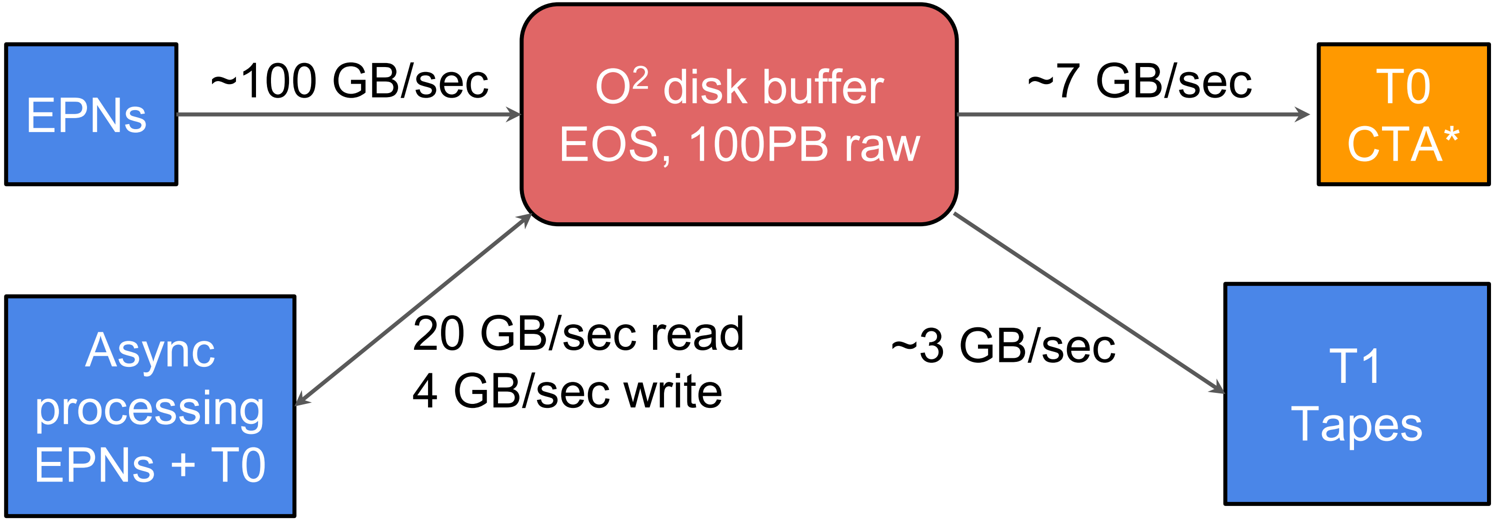
Tier 1 archival

Synchronous (real time processing) of the raw data recorded by the detector and compressed down to 100 GB/s:
Can only happen on-site EPN farm
Must keep step with the input data rate of 635 GB/s ⇒ 2000 GPUs

CTP
Central Trigger Processor
Distribution of timing info, heartbeat trigger

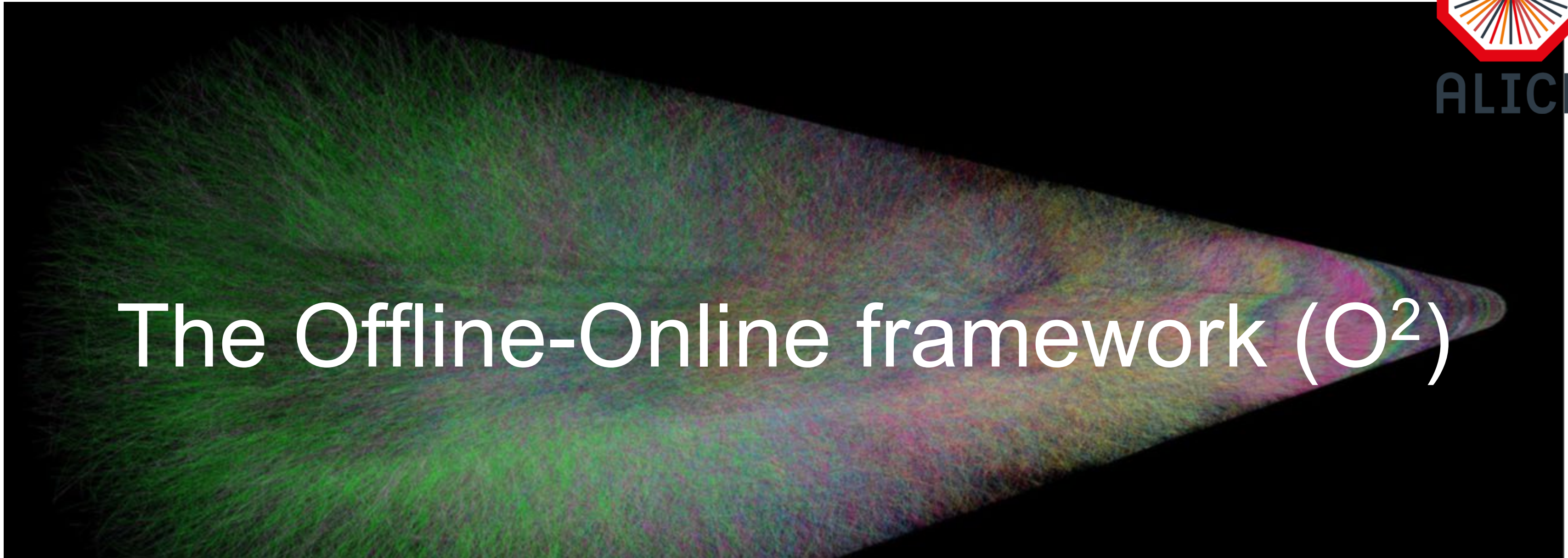
Data buffer for O^2 facility

- 100 PB raw capacity, RS erasure coded (level of security to be defined)
- Based on cheap JBODs, SATA drives, EOS managed



*CTA = CERN Tape Archive

L.Betev, EOS Workshop 2020



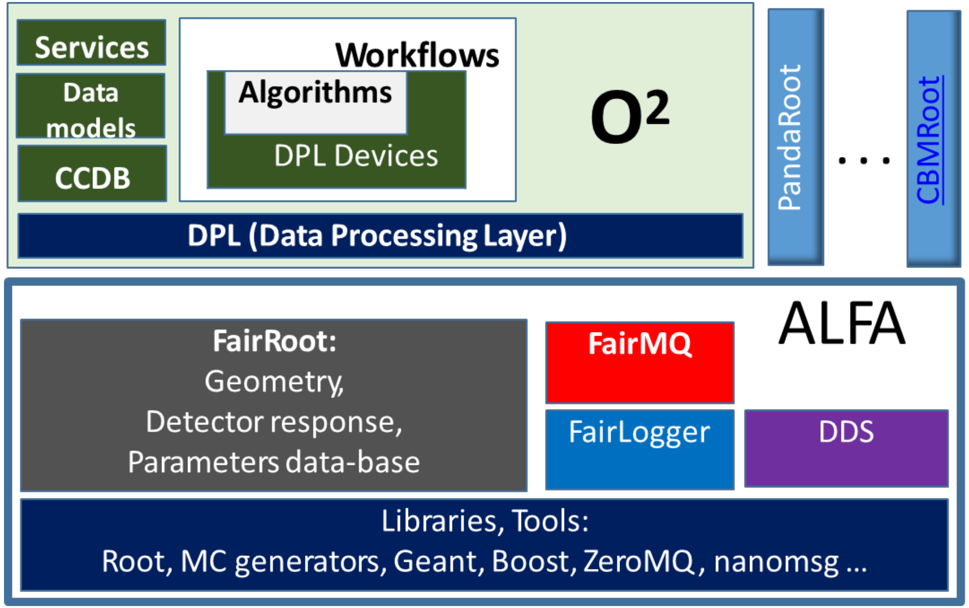
The Offline-Online framework (O^2)

Visualization of **2 ms of 50 kHZ Pb-Pb** data as expected in the ALICE TPC in LHC Run 3

- Basic processing unit: **Time Frame** (~10 ms of data ~500 collisions @ 50 kHZ Pb-Pb)
- Whole TF reconstructed in one shot
- In absence of triggers (reference for drift-time estimate) z position of clusters is not defined

O^2 software framework

Based on ALFA platform: common project of ALICE and FAIR



Key features:

- Message-queues based parallel processing done by separate devices (processes)
- FairMQ supports multiple transport engines (ZMQ, nanomsg) over different protocols (Ethernet, Infiniband, shared memory access)
- Technicalities of message exchange are hidden in the Data Processing Layer (DPL)
- DPL allows to wrap algorithms to Devices with particular Input and Output message type specifications.
- Workflows are built for group of Devices by automatic matching of their Inputs and Outputs

G.Eulisse, CHEP2019
R.Shahoyan, CHEP2019

ALICE

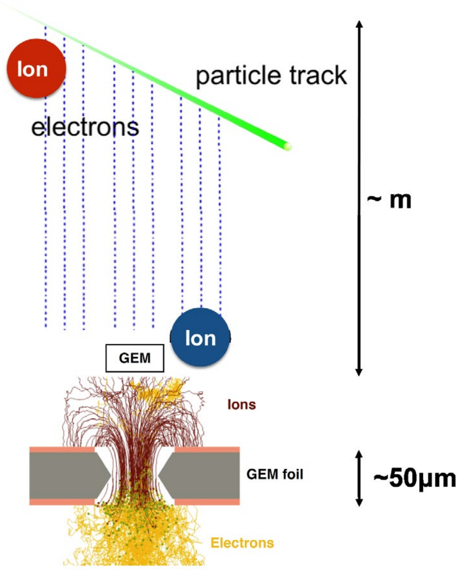
Outline of the TF synchronous stage processing

- The main aims:
  - Raw data reduction/compression to Compressed Time Frame (CTF), calibration data accumulation, QC
- The main tasks:
  - Full TPC clusterization and tracking **(GPU)**
  - Full ITS+MFT clusterization
  - Full FIT & ZDC reconstruction
  - Partial ITS tracking + ITS/TPC/TRD/TOF matching as much as needed for QC and calibration
  - Accumulation of data for offline calibrations (CCDB)
  - On-the-fly calibrations (e.g., TRD, TPC V-drift) with feedback for reconstruction of following TFs
  - Entropy and data reduction, entropy compression (ANS), writing CTF

QC within EPN devices and on dedicated QC servers

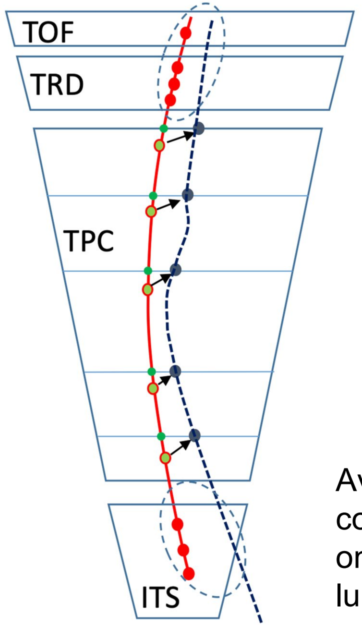
D.Rohr, CHEP2019

TPC average distortion calibration



- electron drift time ~ 100 µs
- ion drift time ~ 200 ms
- Ions belonging to **8000 different PbPb collisions!**

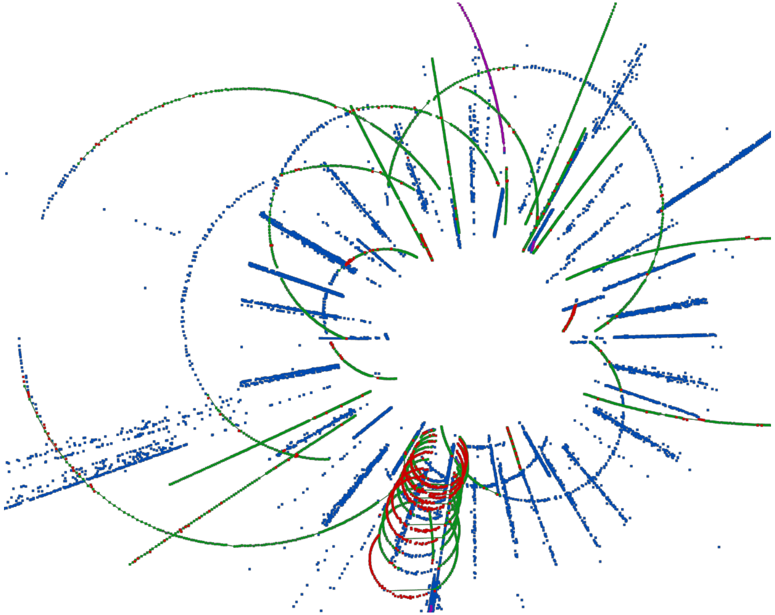
GEM detectors will release in the TPC active area **slow ions** that can **distort the electric field that guides the electrons**



average "shift" (distortion) and fluctuations in the positions of the reconstructed TPC clusters

Average distortions and distortion fluctuations need to be corrected **before the tracking is performed** (in Run 3 online only average corrections as a function of luminosity)

TPC data reduction



- TPC data rejection strategies:
  - A (conservative): reject only clusters of identified background / tracks loopers
  - B (aggressive): keep only clusters attached to or in proximity of identified signal tracks (p_T > 50 MeV/c)
- Estimates for Pb-Pb:
  - Compression factor depends on IR
  - Expected average IR ~22 kHz
  - We plan to compress data according to the more conservative strategy A only in 2022
  - CTF average event size @ 22 kHz:
    - A: 2.85 MB/event
    - B: 1.90 MB/event

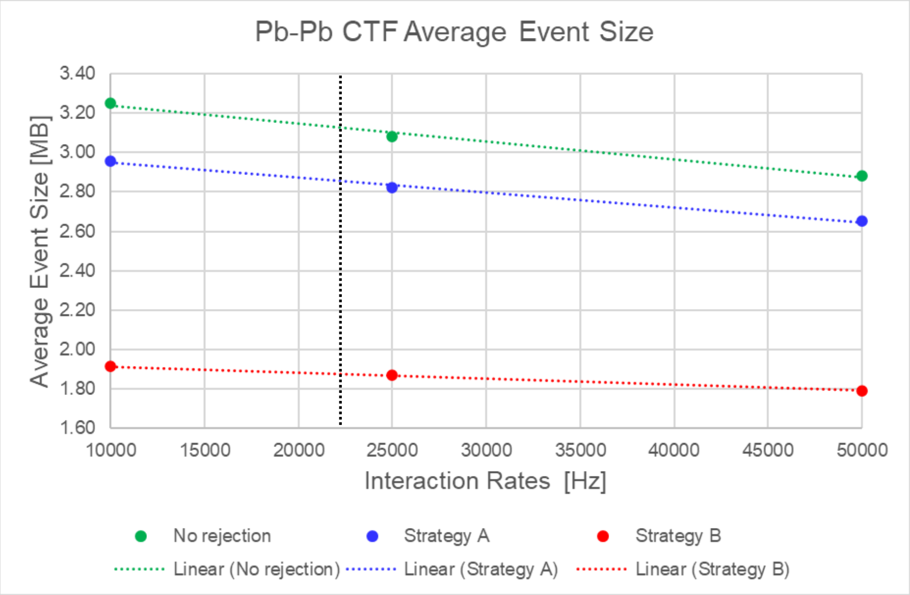
- In 2022 expected 2.7 nb^{-1}:
  - 2.0 x 10^{10} events x 2.85 MB/event ⇒ 57 PB

- Plan to install 100 PB raw disk space as O^2 buffer:
  - 57 PB + 20% contingency

D.Rohr, CHEP2019

CTF average event size

- TPC data rejection strategies:
  - A (conservative): reject only clusters of identified background / tracks loopers
  - B (aggressive): keep only clusters attached to or in proximity of identified signal tracks
- Estimates for Pb-Pb:
  - Compression factor depends on IR
  - Expected average IR ~22 kHz
  - We plan to compress data according to the more conservative strategy A only in 2022
  - CTF average event size @ 22 kHz:
    - A: 2.85 MB/event
    - B: 1.90 MB/event



- In 2022 expected 2.7 nb^{-1}:
  - 2.0 x 10^{10} events x 2.85 MB/event ⇒ 57 PB

- Plan to install 100 PB raw disk space as O^2 buffer:
  - 57 PB + 20% contingency

Generic software and GPU Benchmarks

- **Vendor- / architecture-independent software:**
  - All algorithms are written in **generic C++**, and can be dispatched to HIP, CUDA, OpenCL on GPUs or OpenMP on CPUs using small **wrappers** ⇒ good code maintainability
  - **GPU libraries linked dynamically on demand** ⇒ can distribute same binary software to CPU and GPU nodes

- Benchmarking of the synchronous software completed in August 2020:
  - **GPU performance @ 50kHz Pb-Pb**
    - ~1600 AMD MI50 and ~1100 NVIDIA Quadro RTX 6000
    - Compatible with our previous estimates <2000 GPU including 20% margin
  - **GPU Memory optimization**
    - 128 orbit TF needs 24 GB
  - **EPN Full System Tests performed with 70 orbit TF**
    - Validated processing rate of 1/230 of assumed rate at 50 kHz Pb-Pb (nominal 1/250)
    - Max. server memory consumption 280 GB and CPU load 44 cores (+20% in the final setup)

ALICE

Generic software and GPU Benchmarks

Specifications of EPN defined (EPN PRR on Aug. 21st 2020)

Server choice SuperMicro AS-4124GS-TNR with

- **Eight GPUs AMD MI50 – 32 GB** / Two CPUs 32 core AMD Rome 7452
- 512 GB 3200 Hz Main Memory, 200 GB SSD plus 4 TB SSD
- 100 Gb/s InfiniBand HCA

(EPN delivery on Dec. 18th 2020)

250 servers totalling 2000 GPUs and 16000 physical cores

- Benchmarking of the synchronous software completed in August:
  - **GPU performance @ 50kHz Pb-Pb**
    - ~1600 AMD MI50 and ~1100 NVIDIA Quadro RTX 6000
    - Compatible with our previous estimates <2000 GPU including 20% margin
  - **GPU Memory optimization**
    - 128 orbit TF needs 24 GB
  - **EPN Full System Tests**
    - Validated processing rate of 1/230 of assumed rate at 50 kHz Pb-Pb (nominal 1/250)
    - Max. server memory consumption 280 GB and CPU load 44 cores (+20% in the final setup)

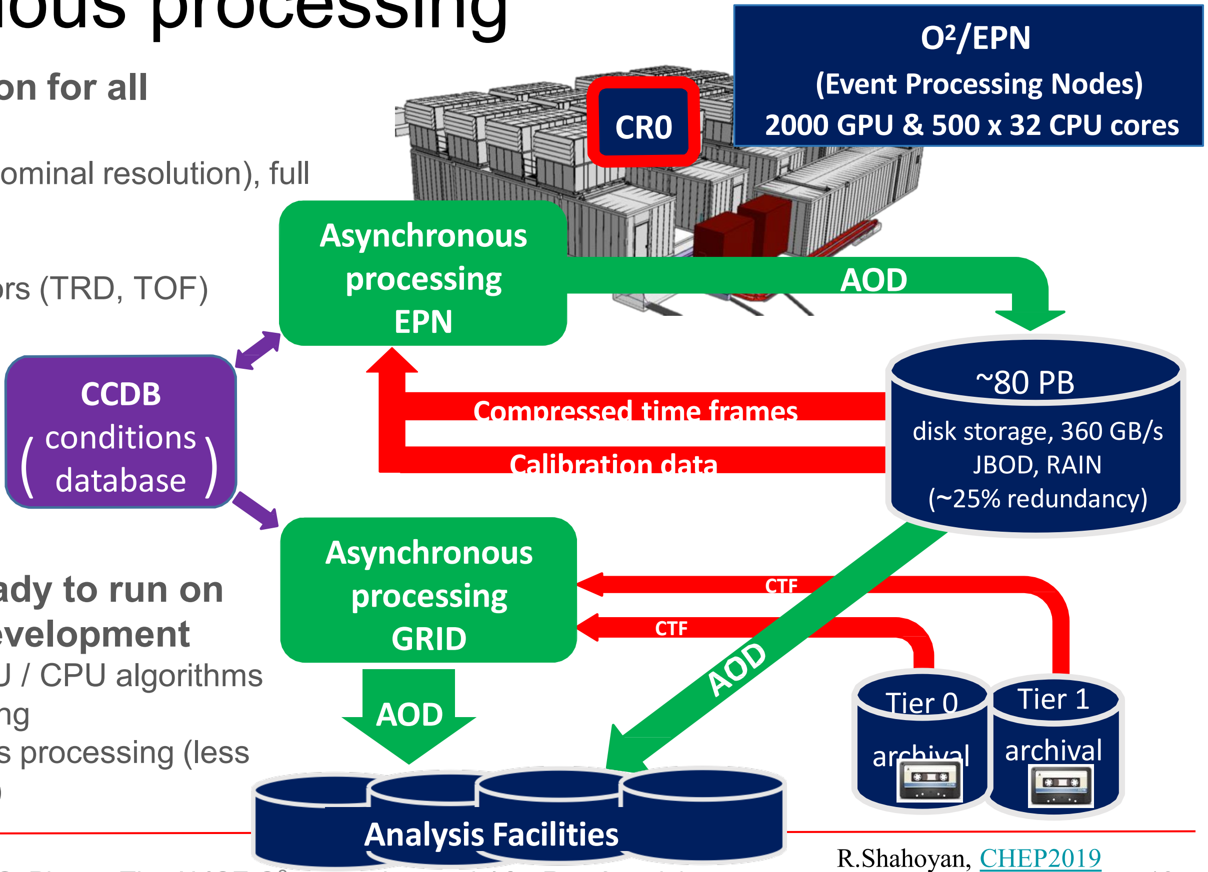
ALICE asynchronous processing

Full reconstruction, full calibration for all detectors:

- Full correction of TPC distortions (nominal resolution), full calibration for all detectors
- TPC – ITS tracks matching
- Tracks propagation to outer detectors (TRD, TOF)
- Global track fits
- Primary and secondary vertices
- PID hypothesis
- AOD as output

Asynchronous code (almost) ready to run on CPUs, porting to GPUs under development

- Different relative importance of GPU / CPU algorithms compared to synchronous processing
- TPC part faster than in synchronous processing (less hits, no clustering, no compression)



R.Shahoyan, CHEP2019

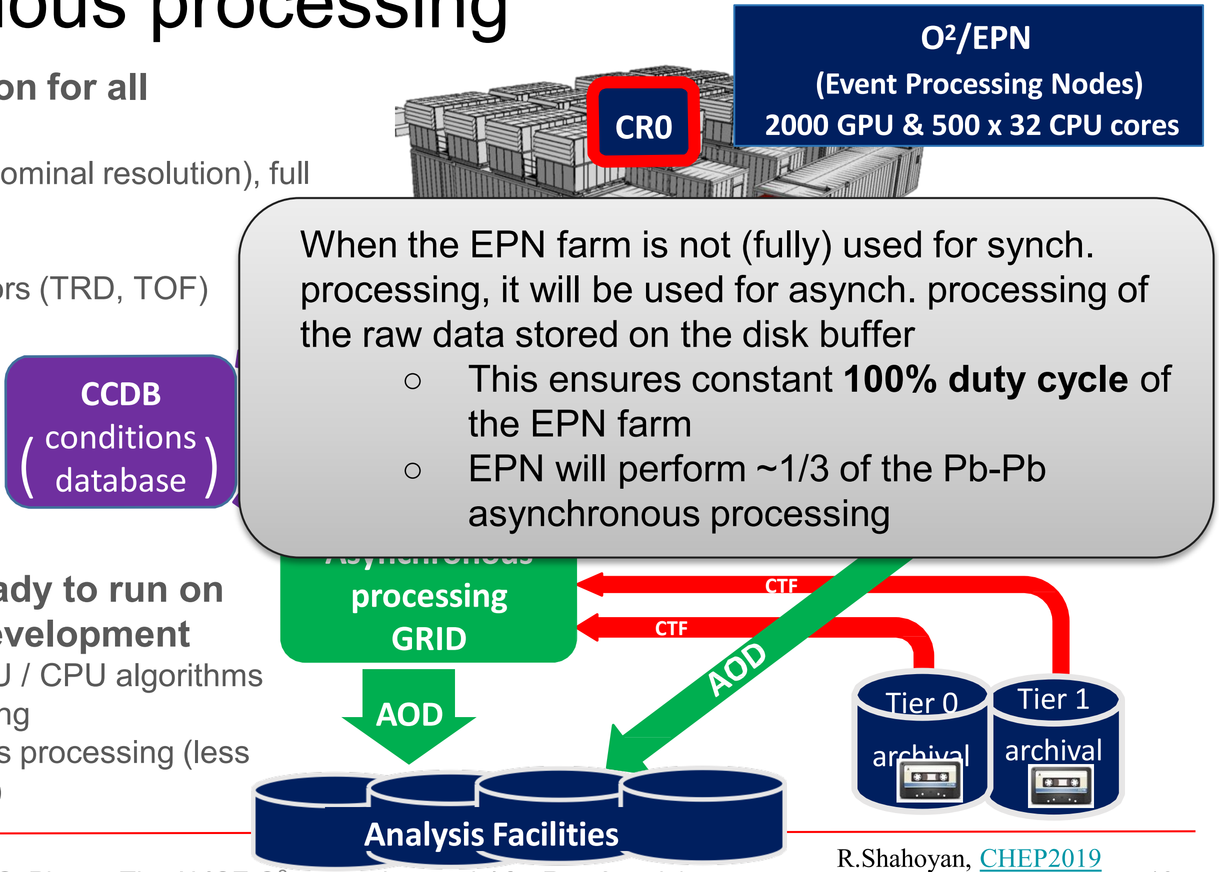
ALICE asynchronous processing

Full reconstruction, full calibration for all detectors:

- Full correction of TPC distortions (nominal resolution), full calibration for all detectors
- TPC – ITS tracks matching
- Tracks propagation to outer detectors (TRD, TOF)
- Global track fits
- Primary and secondary vertices
- PID hypothesis
- AOD as output

When the EPN farm is not (fully) used for synch. processing, it will be used for asynch. processing of the raw data stored on the disk buffer

- This ensures constant **100% duty cycle** of the EPN farm
- EPN will perform ~1/3 of the Pb-Pb asynchronous processing



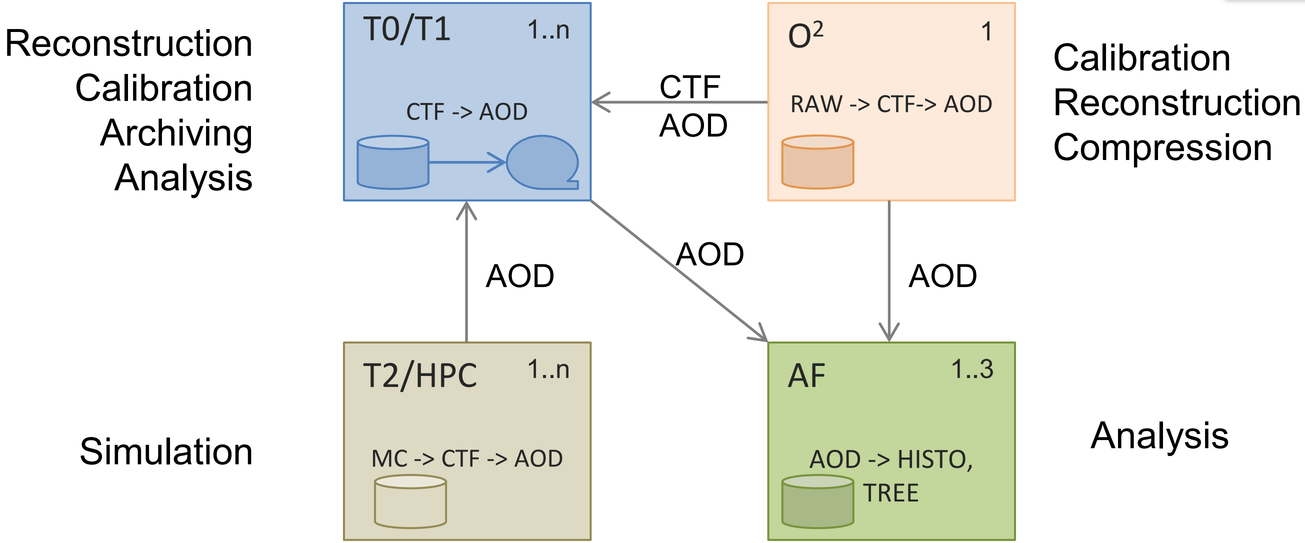
Asynchronous code (almost) ready to run on CPUs, porting to GPUs under development

- Different relative importance of GPU / CPU algorithms compared to synchronous processing
- TPC part faster than in synchronous processing (less hits, no clustering, no compression)

R.Shahoyan, CHEP2019

ALICE Run 3 Computing Model

Grid Tiers will be mostly specialized for given role



CPU share	O^2	T0	T1s	T2s	AF
Synch.	**100%**	**0%**	**0%**	**0%**	**0%**
Asynch.	**33%**	**33%**	**33%**	**0%**	**0%**
MC	**0%**	**0%**	**0%**	**100%**	**0%**
Analysis	**0%**	**30%**	**20%**	**0%**	**50%**

- Subject to fine tuning
- MC can be run as a backfill

2/3s of CTFs processed by O^2 + T0 and archived at T0;
1/3 of CTFs exported, archived and processed on T1s;

One calibration (sync.) and two reconstruction passes (async.) over raw data each year;

The goal is to minimize data transfer and optimize processing efficiency

10% of AODs sampled and sent to the Analysis Facility for quick analysis and cut tuning;
Analysis of full data sample across T0/T1s only performed upon Physics Board approval.

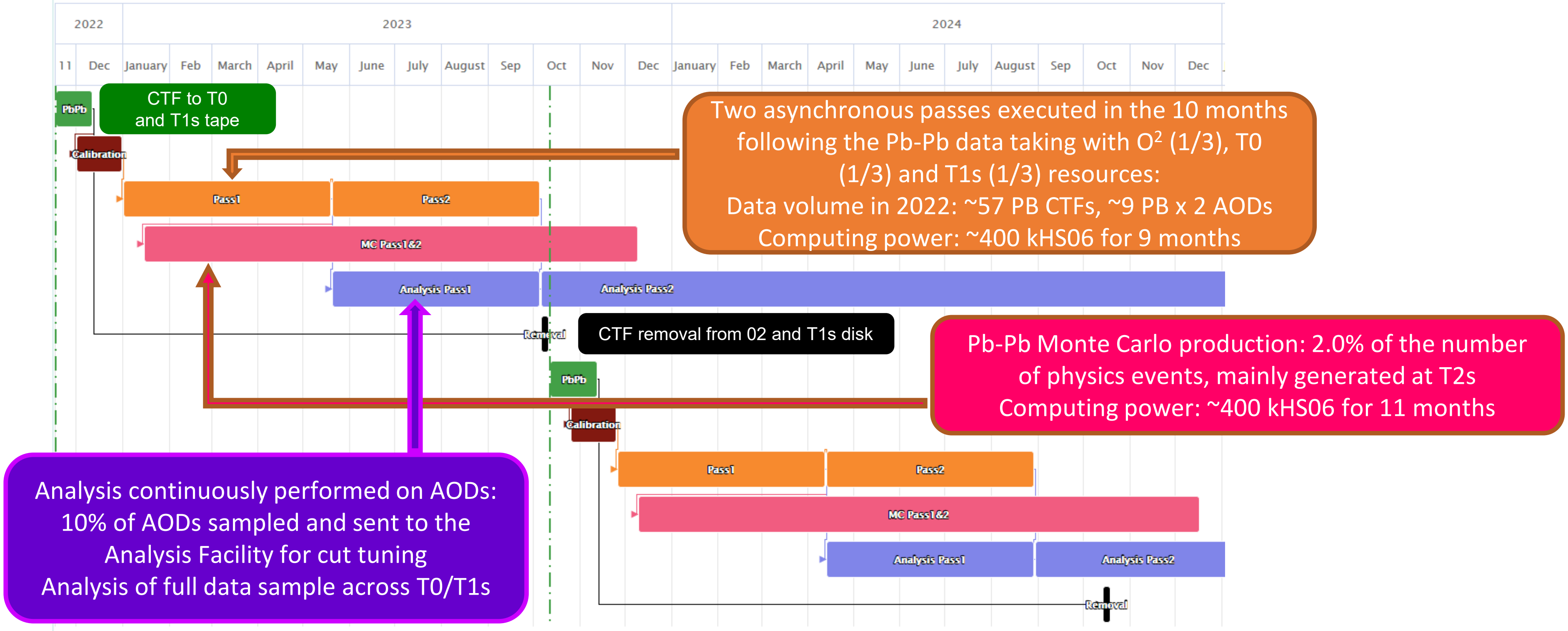
O^2 TDR and addendum

ALICE

Replication and deletion policy

- Due to a substantial increase in data volume, in Run 3, there will be only one instance of each raw data file (CTF) stored on disk with a backup on tape
  - EOS with erasure encoding provides 20% overhead in available storage space with high availability in large installations such as CERN
  - In case of data loss, we will restore lost files from tape
- The size of O^2 disk buffer should sufficiently accommodate CTF data from the entire period with contingency.
  - As soon as it is available, the CTF data will be archived on tape and partially moved to the T1 disks
- Given the limited size of the disk buffers in O^2 and Tier 1s, all CTF data collected in the previous year, will have to be removed before new data taking period starts.
- All data not finally processed during this period will remain stored on tapes until the next opportunity for re-processing arises: LS3
- The computing model foresees to have two copies of AOD data on disk from each asynchronous pass (load balancing and redundancy)
- One copy of the AOD per pass is planned to be archived to tape

Pb-Pb processing plan



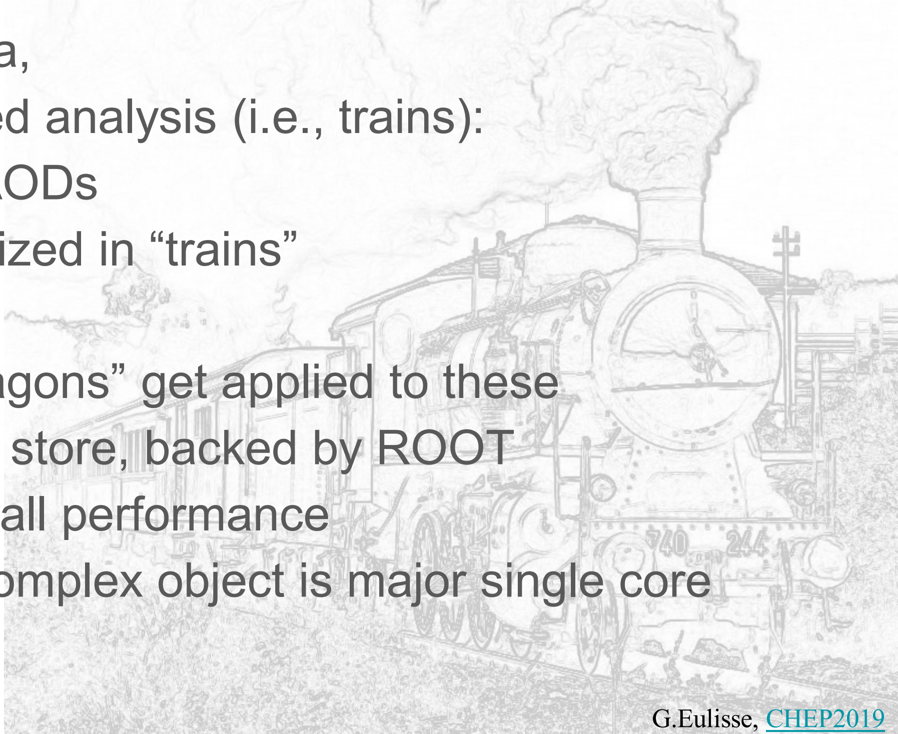
Analysis facilities

- Collect ~10% of AODs on a few dedicated sites (AFs) that are capable of quickly processing locally large data volumes
- Goal: fast turnaround for cut tuning and task validation
- The AF needs to be able to digest more than 5 PB of AODs in a 12-hour period
- ~20000 cores and 5→10 PB of disk on very-performing file system
- Run 2 analysis trains need on average 5 MB/s per job slot to be reasonably efficient
- (New O^2 analysis factor 3-10 higher event throughput)
- AF cluster file system able to serve 20,000 job slots at an aggregate throughput of 100 GB/s
- GSI is the first ALICE Analysis Facility (Wigner AF coming soon):
    - **2021 pledges 50 kHS06 and 4.7 PB**
    - All worker nodes have a direct mount of the shared Cluster file system Lustre
    - To optimize the I/O throughput of the analysis jobs self developed XRootD Plugin: the data can access directly without the need to communicate with the external Grid services (AliEn catalogue) (expected 115 GB/s)

Run 1 and 2 ALICE analysis model

In order to offset the costs of reading data,
ALICE has as strong tradition of organized analysis (i.e., trains):

- Analysis happens on both ESD and AODs
- Users provide tasks, “wagons”, organized in “trains”
- Trains run on the Grid
- Data are read only once per train, “wagons” get applied to these
- Data are kept in a generic C++ object store, backed by ROOT
- Slow sites / site issues dominate overall performance
- Data-access and de-serialization of complex object is major single core performance offender



G.Eulisse, CHEP2019

O^2 Analysis Model

The idea of organized analysis (trains) will stay. Improve on the implementation:

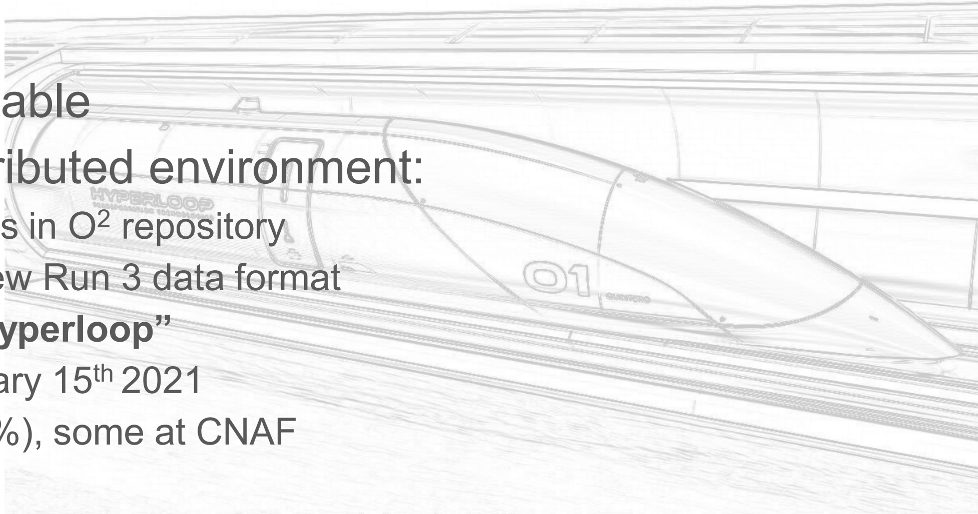
- Streamline data model, trade generality for speed: flatten data structures
- Recompute quantities on the fly rather than storing them: CPU cycles are cheaper
- Produce highly targeted ntuples (in terms of information needed and selected events of interest) to reduce turnaround for some key analysis: derived data sets

Key elements:

- FairMQ: Data processing in separate processes exchanging data via shared memory
- Apache Arrow as backing store for the messaging passing (interoperability)
- ALICE O^2 Data Processing Layer (DPL): translate the workflows to an actual FairMQ topology and taking care of parallelism

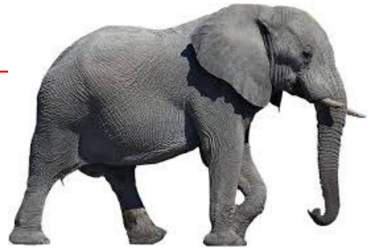
G.Eulisse, CHEP2019

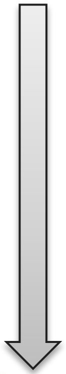
O^2 Analysis Challenge Outcome



- Novel O^2 analysis framework available
- Benchmark of real analysis in distributed environment:
  - So far 66 Physics Working Group tasks in O^2 repository
  - Converted Run 2 data samples into new Run 3 data format
  - New organized distributed analysis “**hyperloop**”
    - About 200 trains ran since January 15th 2021
    - Grid: jobs mostly at CERN (>80%), some at CNAF
    - GSI AF: 2 core and 4 core
- Wrt Run 2 analysis: factor 3-10 higher event throughput (still being optimized)
  - no optimization for optimal core usage done yet
  - but multi core shows speed up
- Expected further framework developments and optimizations in the next months

O^2 simulation ecosystem





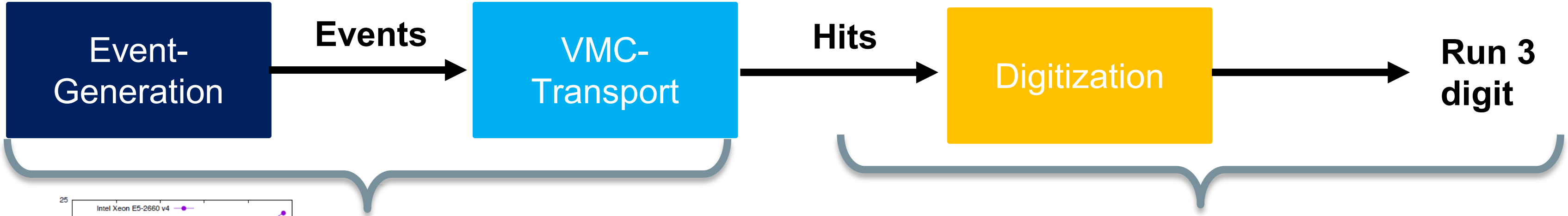
- Community collaboration: seek common solutions with FAIR experiments
  - Use FairRoot Monte Carlo classes as common blocks to build simulations in the VMC ecosystem
  - FairMQas the basis to build complex distributed systems based on messaging
- Novel requirements in Run 3 detector simulation:
  - Detector geometry changes; readout electronics changes (continuous readout); new data formats
  - Need to adapt to the availability of many-core systems and be able to make use of them efficiently
  - Chance of code cleanup; improvements in class and data layout; improvements in performance; etc.
- Simulating Pb-Pb collision is very demanding:
  - Up to 100k primaries in the collision to transport
  - Heavyweight resource utilization using standard single-core event-based simulation
  - ~O(GBs) of memory / event an ~O(h) of CPU time / event

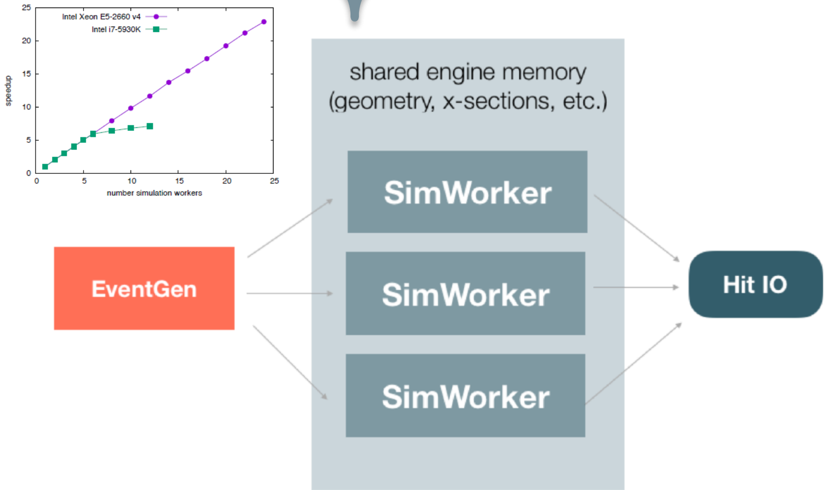
⇒ **event splitting and collaborative simulation parallelism**

⇒ **independent actors based on heterogenous multi-processing and message transfer**

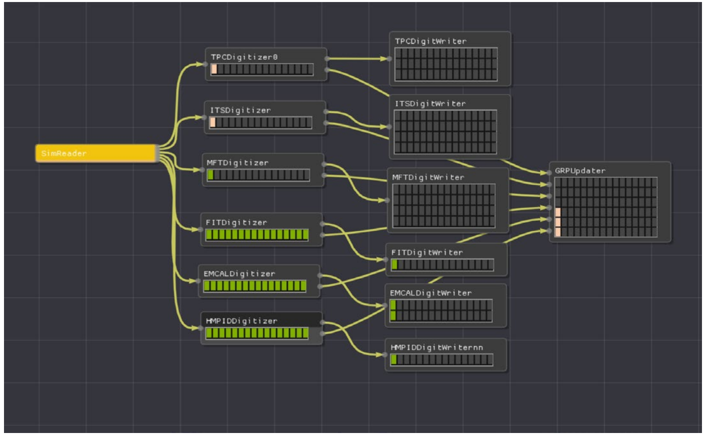
S.Wenzel, CHEP2018

O^2 simulation ecosystem: 2 independent workflows





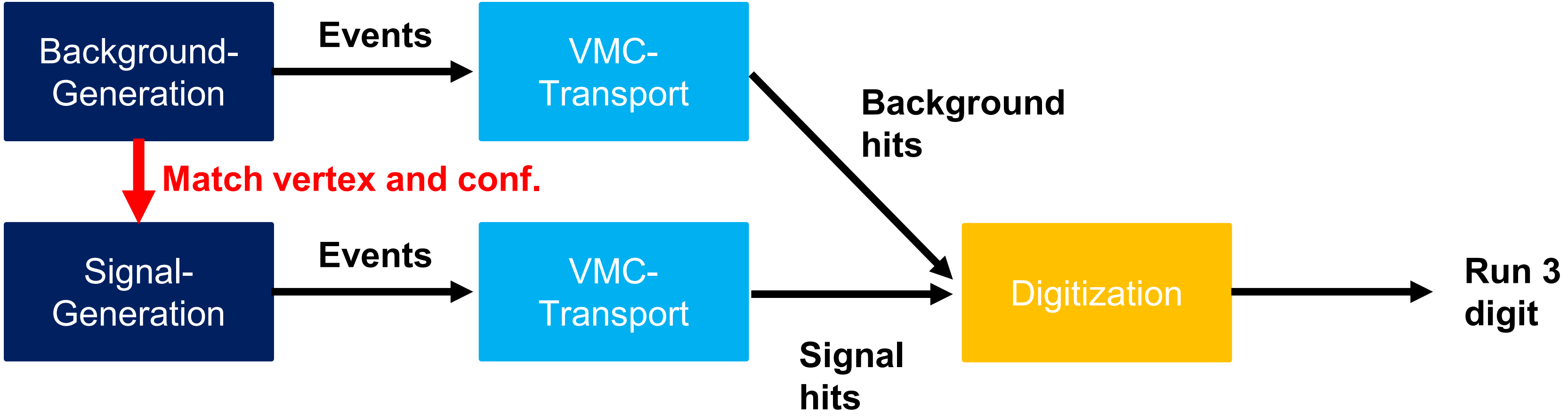
Sub-event parallel simulation across multiple transport workers



Implemented in the ALICE Run 3 framework (DPL)
Parallel across and within detector components

S.Wenzel, CHEP2018

O^2 simulation ecosystem: embedding in digitization



O^2 simulation challenge outcome:

Embedding of Heavy-Flavour events into Pb-Pb background TF

- With MC-to-MC embedding, Geant4 simulation is comparable in speed to Geant3:
  - Geant4 slower transport but faster digitization
  - Embedding gain: ~3.9 (~4.7) with Geant3 (Geant4)
  - Same number of timeframes as Geant3 in 24 hours

S.Wenzel, CHEP2018

ALICE

HPC and cloud resources

- Thanks to the new O^2 simulation and reconstruction code (Run 3) possible to fully exploit the multi process features
- Significant progress has been made to incorporate HPC and cloud resources in the standard ALICE Grid workflows
  - Multicore queues at CERN used to test and benchmark the O^2 MC code
  - Intel based HPCs (Marconi @ CINECA, Cori and Lawrencium @ LBNL) were used for the O^2 MC challenge
  - Cloud resources delivery at CERN - direct integration of Azure cloud as a Grid node

- Next steps:
  - porting the O^2 code to Power 9 and ARM platform

ALICE

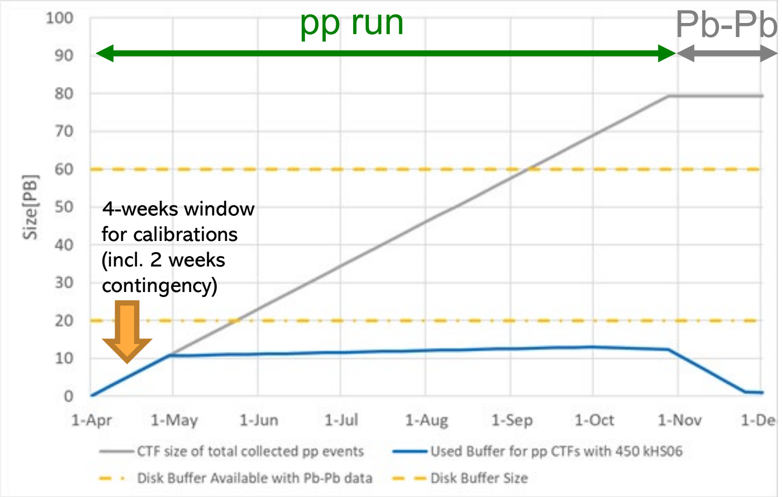
High-energy pp programme after LS2

- **Target data samples:**
  - L_{int} = 200/pb with continuous data-taking and processing followed event skimming with selectivity of ~10^{-3}
    - e.g., compared to Run 2 sample: x20 for high-multiplicity; x3000 for measurements that were based on minimum bias sample
  - L_{int} = 3/pb at low field (0.2 T) with continuous data-taking and no event selection
- Proposed running scenario: 500 kHz interaction rate
  - Requires 5 full pp years at 50% LHC efficiency (in physics periods)
- **Data processing strategy:**
  - Standard processing chain (synchronous pass → calibrations → asynchronous pass) on O^2 facility, with event selection during asynchronous pass
  - Compressed Time Frames reduced by keeping the clusters associated to tracks that point to the primary vertex of a selected collision within ±30 cm in *z* (strangeness decays)

ALICE-PUBLIC-2020-005

pp reconstruction requirements

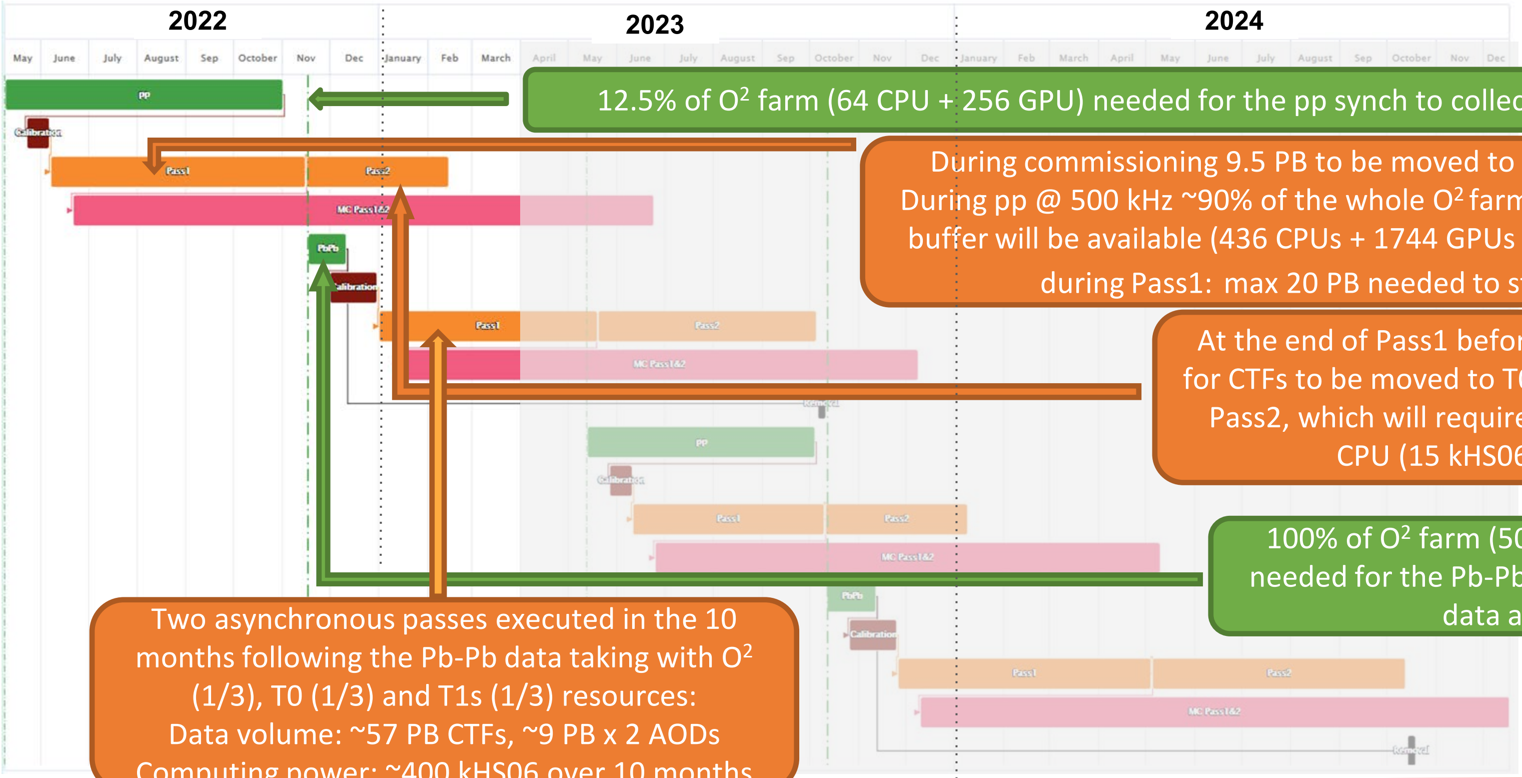
- Synchronous pass: 12.5% of O^2 facility nodes
  - pp 500 kHz ~ 12.5% of Pb-Pb 50 kHz particle rate, for which the farm is dimensioned
- Asynchronous pass: 450 kHS06, to consume pp data (async + event selection) with backlog < 20 PB (1/3 of O^2 facility buffer, available during Pb-Pb reco.)
- Aim to reduce CPU load reduction using GPUs and further software optimization



Yearly time profile of used disk on O^2 facility:
pp interaction rate: 500 kHz
Computing capacity used for pp asynch.: 450 kHS06

ALICE-PUBLIC-2020-005

Processing plan for 2022



12.5% of O^2 farm (64 CPU + 256 GPU) needed for the pp synch to collect pp data at 500 kHz

During commissioning 9.5 PB to be moved to T0 (2/3) and T1 (1/3)
During pp @ 500 kHz ~90% of the whole O^2 farm and the whole O^2 disk buffer will be available (436 CPUs + 1744 GPUs + 80 PB) to select data during Pass1: max 20 PB needed to store pp CTFs

At the end of Pass1 before Pb-Pb period: 1.2 PB for CTFs to be moved to T0 (2/3) and T1s (1/3) for Pass2, which will require negligible amount of CPU (15 kHS06 in 90 days)

100% of O^2 farm (500 CPUs + 2000 GPUs) needed for the Pb-Pb synch to collect Pb-Pb data at 50 kHz

Two asynchronous passes executed in the 10 months following the Pb-Pb data taking with O^2 (1/3), T0 (1/3) and T1s (1/3) resources:
Data volume: ~57 PB CTFs, ~9 PB x 2 AODs
Computing power: ~400 kHS06 over 10 months

ALICE

Baseline scenario for Run 3 and 2022

Baseline Run 3 schedule has been defined:

- Experimental caverns closure on February 1^{st} 2022
- Pilot beam in September 2021

2022 Running Conditions for Computing estimates including contingency

LPC

- ATLAS/CMS luminosity: <70/fb
- ATLAS/CMS average pile-up: 35
- LHCb luminosity: <10/fb
- ALICE luminosity: <90/pb
- Running time pp: 6×10^6 seconds
- Running time ions: 1.2×10^6 seconds

Numbers assume 50% stable beam time and fast ramp-up to 2018 conditions

2022 Running scenario:

- **Commissioning:**
  - pp collisions at 20-200 kHz during ramp-up (28d)
  - pp collisions at 1 MHz for pp ref. exercise (3d)
  - pp collisions at 5 MHz IR (eq. load of Pb-Pb@50 kHz)
- **High-energy pp physics programme:**
  - 135 days pp for physics in 2022
  - 500 kHz interaction rate @ 45% efficiency
  - 35 / pb ⇒ 2.6 10^{12} events
  - Strong event selection during asynch. pass on O^2 facility (1.2%) ⇒ 3.1 10^{10} events on disk
- **24 days of Pb-Pb in 2022:**
  - 24 days @ 50% LHC stable beams @ 95% ALICE eff
  - 2.56 / nb ⇒ 2.0 10^{10} events

ALICE

Resource Requirements for 2022

ALICE		2021					2022		
		Req.	Exp. Priority	C-RSG	Pledge	Pledge / Priority	Prev. Est.	Req.	Req. 2022 / Priority 2021
CPU [kHS06]	Tier-0	471	403	471	471	17%	471	471	17%
	Tier-1	498	420	498	412	-2%	498	498	19%
	Tier-2	515	432	515	481	11%	515	515	19%
	Total	**1484**	**1255**	**1484**	**1364**	**9%**	**1484**	**1484**	**18%**
Disk [PB]	Tier-0	45.5	36.3	45.5	45.5	25%	45.5	50.0	38%
	Tier-1	53.3	48.4	53.3	45.4	-6%	53.3	55.0	14%
	Tier-2	44.8	42.9	44.8	50.4	17%	47.0	49.0	14%
	Total	**143.6**	**127.6**	**143.6**	**141.3**	**11%**	**145.8**	**154.0**	**21%**
Tape [PB]	Tier-0	86.0	50.3	86.0	86.0	71%	86.0	95.0	89%
	Tier-1	57.0	41.2	57.0	59.8	45%	57.0	63.0	53%
	Total	**143.0**	**91.5**	**143.0**	**145.8**	**59%**	**143.0**	**158.0**	**73%**

Resource estimates	CPU (kHS06)	DISK (PB)	TAPE (PB)
Previous estimates for 2022	1484	145.8	143.0
Requests for 2022	1484	154.0	158.5
Difference	0%	5.6% (**8.2 PB**)	10.5% (**15.0 PB**)

Between previous estimates for 2022 and the present requests for 2022:

- No change in CPU
- 5.6% increase in T0/T1s/T2s disk
- 10.5% increase in tape space
- **1/3 of CTFs will be archived to Tiers-1 tape, while 2/3 of CTFs to Tier-0 tape, as foreseen in Run 3 computing model**

ALICE

Resource Estimates for 2023

ALICE		2022 Req.	2023 Est.	Est. 2023 / Req. 2022
CPU [kHS06]	Tier-0	471	541	15%
	Tier-1	498	572	15%
	Tier-2	515	592	15%
	Total	**1484**	**1705**	**15%**
Disk [PB]	Tier-0	50.0	58.0	16%
	Tier-1	55.0	63.0	15%
	Tier-2	49.0	56.0	14%
	Total	**154.0**	**177.0**	**15%**
Tape [PB]	Tier-0	95.0	126.0	33%
	Tier-1	63.0	79.0	25%
	Total	**158.0**	**205.0**	**30%**

2023 Running scenario:

- **Precise schedule still unknown**
- **Considered as upperlimit:**
  - **130 days of pp in 2023:**
    - pp collisions at 500kHz for high-energy pp physics programme
  - **28 days of HI in 2023:**
    - 28 HI days dedicated to Pb-Pb collisions
- **These assumptions accommodate with some margin different possible scenarios:**
  - pp collected at low field, additional short O-O and p-O runs
  - p-Pb, pp reference instead of Pb-Pb

Estimates for 2023:

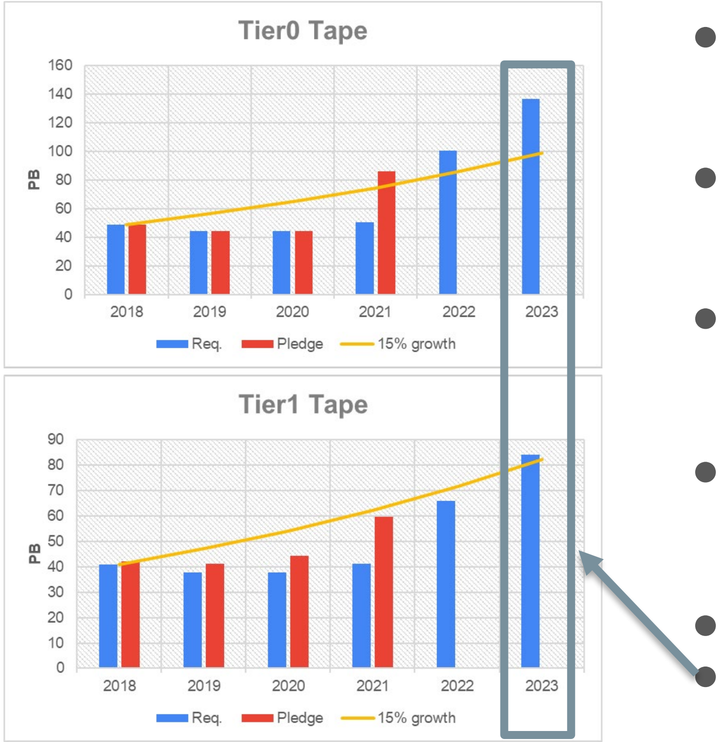
- Moderate growth of CPU (+15% per year) and disk
- Strategy B helps to mitigate the needed disk space in tape:
  - +31 PB at T0, +16 PB at T1s
- **A better assessment of the computing estimates for 2023 will be delineated for Autumn RRB**

ALICE

Expected growth of CPU and disk space in 2022-2023

------- 15% growth: annually compounded rate from 2018 along LS2

Expected growth of TAPE space in 2022-2023



- TAPE in 2022:
  - 66 PB @ T1s
  - 100.5 PB @ T0
- PLEDGED in 2021:
  - 59.8 PB @ T1s
  - 86.0 PB @ T0
- 75 PB needed in 2022:
  - 18.0 PB for commissioning
  - 57.0 PB for Pb-Pb
- Give flexibility to FA's to delay the deployment:
  - commissioning in February 2022
  - Pb-Pb in October 2022
- Crucial to adopt strategy B in the following years
- 2023 to be considered as an upperlimit

------- 15% growth: annually compounded rate from 2018 along LS2

Summary

- O^2 simulation and reconstruction framework can exploit all advantages
  - Parallelism and improved workload scheduling
  - Ready to run multicore GRID jobs and on Intel based HPCs
- Large effort in developing and realizing the Run 3 TPC synchronous reconstruction tasks for GPU platform
  - Vendor and architecture-independent software
  - 2000 GPUs are sufficient for continuous readout of Pb-Pb at 50 kHz (including a 20% margin)
- Porting of asynchronous reconstruction code to GPUs under development
- Software algorithms and updated computing model allow to fit into the standard Grid resource growth of CPU and disk space
- Steps of requests in 2022 for tapes, precise schedule for 2023 still unknown

Status of ALICE sites

ALICE