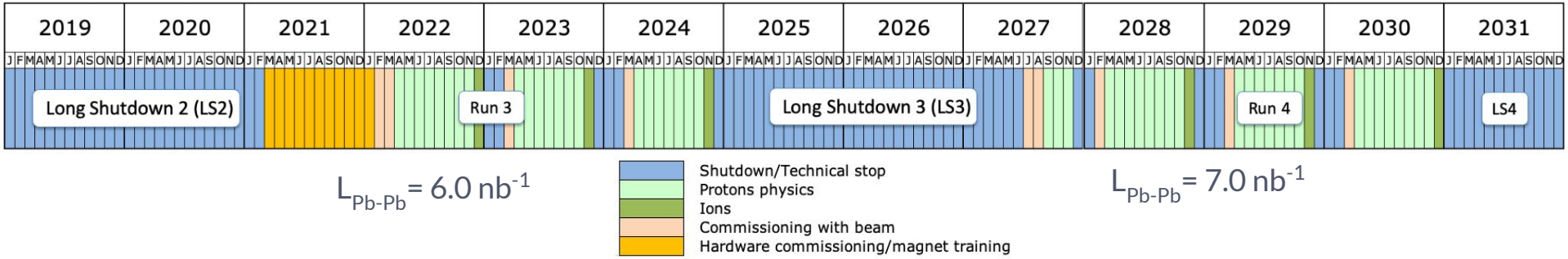
Workshop di CCR - May 28th, 2021

# Evolution of ALICE computing: model for and analysis framework for Run 3&4

M. Concas, N. Jacazio, F. Noferini,
D. Elia, M. Masera, S. Piano, S. Vallero

ALICE

# LHC schedule and ALICE focus

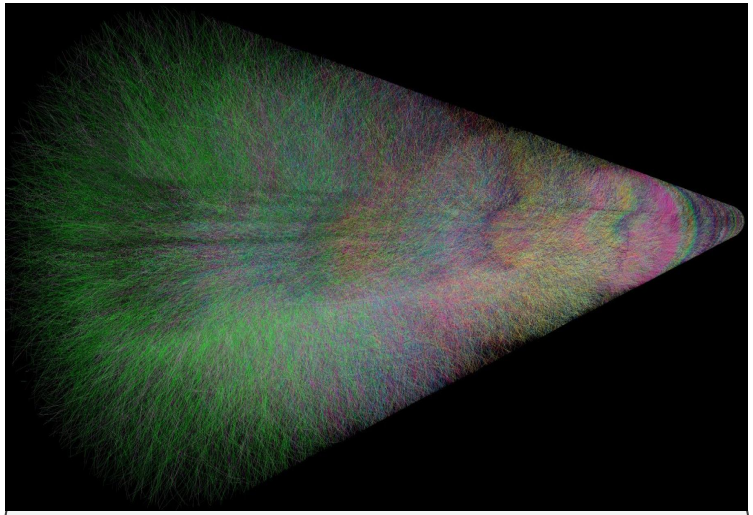



- Heavy-flavour mesons and baryons (down to very low $p_T$)
- Mechanism of quark-medium interaction
- Charmonium states
- Dissociation/regeneration as tool to study de-confinement and medium temperature
- Di-leptons from QGP radiation and low-mass vector mesons
- χ symmetry restoration, initial temperature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

CERN Yellow Report arXiv:1812.06772

# Towards an un-triggered data taking

- 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, principal tracking detector TPC
- Run 3: LHC will deliver **50 kHz in Pb-Pb** collisions
  - ALICE High-luminosity programme starts with Run 3
  - Goal: integrated luminosity **~13 $nb^{-1}$ at 5 TeV with minimum bias** in **Run 3&4**
- Triggerless data acquisition: **continuous readout**
  - Data in unfiltered readout buffers, **timeframe** (~11ms)
  - **100x higher data rate** wrt Run 2
- **Upgraded** and **new detectors** to enable new schema

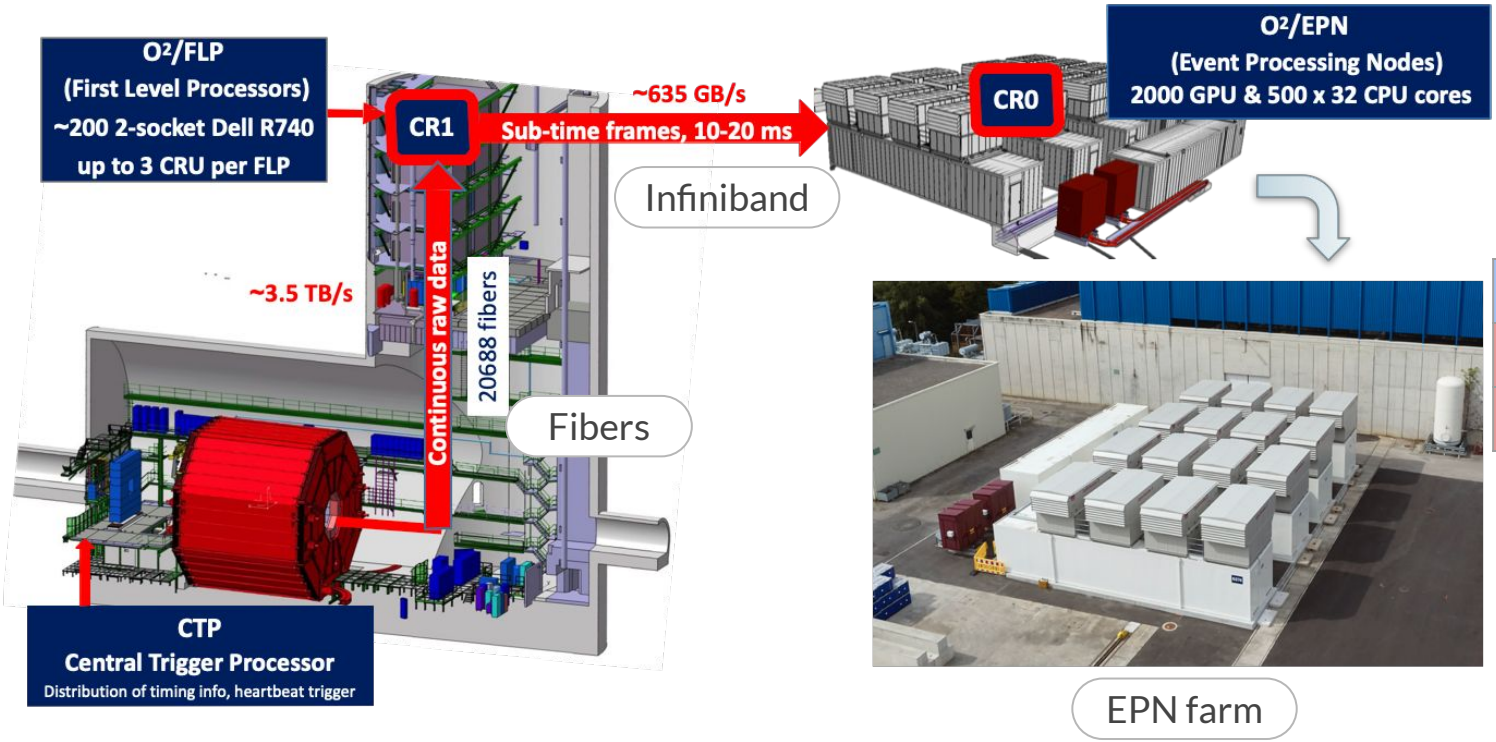


Overlapping events in TPC, @50 kHz Pb-Pb (2 ms)

# Run 3&4 inherent challenges

- Data reconstruction
  - Starts online: **synchronous reconstruction,** requires **online calibration**
  - Extremely demanding task → Dedicated farm Event Processing Nodes with **GPUs**
  - **Timeframes** rather than events **as fundamental unit** in reco algorithms
- Detector simulation and MC production
  - **New** geometries of the **detectors** with a **new digitization framework** to account for pileup
  - Timeframe reconstruction requires **larger memory resources**
  - **Multicore Grid operations** (until now single core jobs exclusively)
- Data analysis:
  - New analysis framework to work with increased regime of data to be processed
  - Flat data model: optimized for fast I/O and supports new data processing technologies on the market

# Raw data life cycle

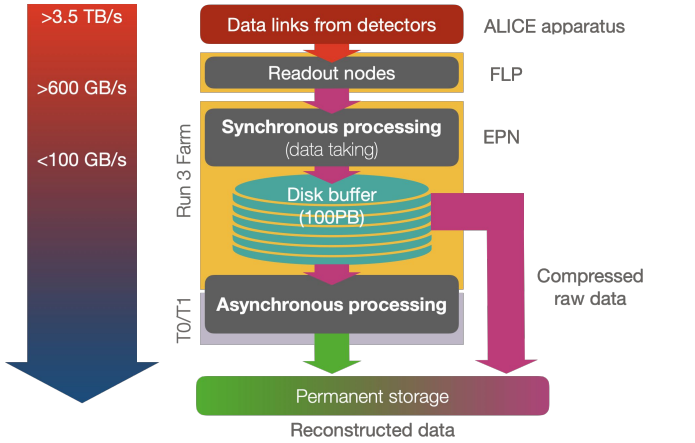


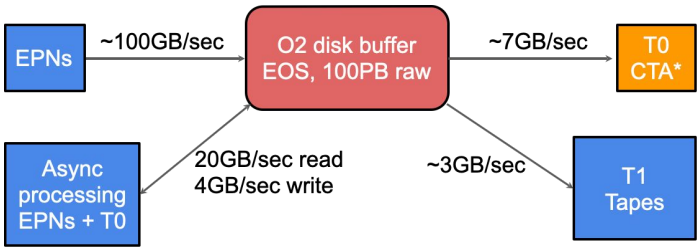

| Σ Raw data: 3465 GB/s | | | |
|---|---|---|---|
| TPC | ITS | TRD | Others |
| 3400 | 40 | 4 | 21 |

R. Shaoyan, CHEP 2019

# Data processing schema

- Data processing operates in two phases:
  - **Synchronous**: runs in “real-time”, data compression from >600GB/s → ~100GB/s, EPN-only, GPUs mission critical
  - **Asynchronous:** during no-beam or when resources are not fully deployed for synch reco → full reconstruction of all detectors and production of AODs
- Large pool of disks (~100 PB raw) to buffer data to be processed on asynchronous phase on both EPNs and Grid
  - Cheap JBODs, SATA drives, EOS managed

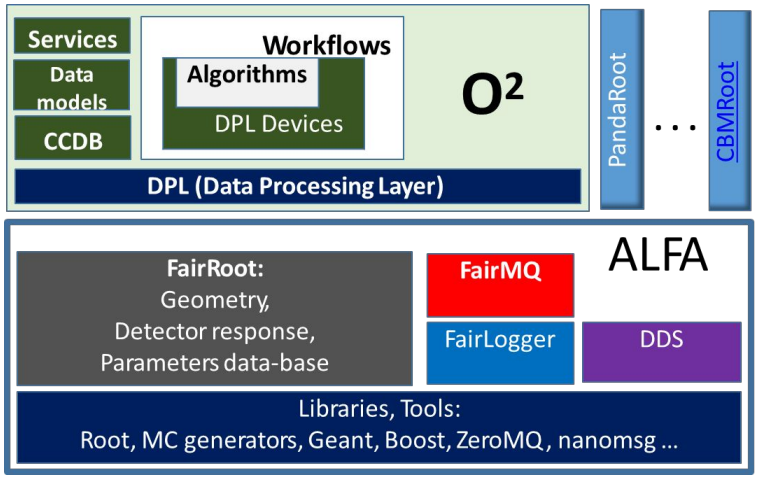



*CTA: CERN Tape Archive

# Online-Offline framework: $O^2$

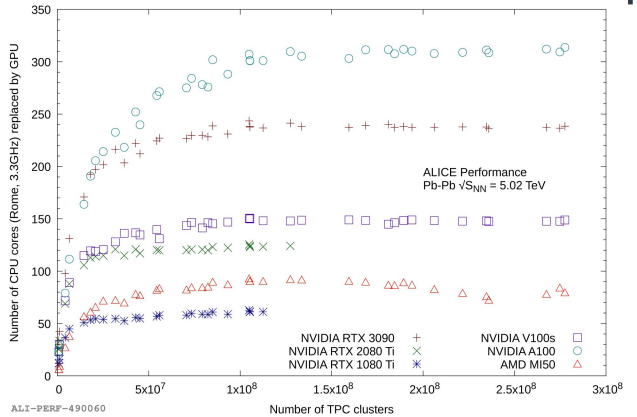
- Single stack for online and offline reconstruction, simulation and analysis
- ALFA as core engine and DPL as ALICE-specific layer for data reconstruction and processing
- Multi process approach with message queues-based interoperation: FairMQ supports from shared memory to Infiniband
- Workflows: groups of “devices” (processes) automatically matching their inputs and outputs
  - Generic description, can be deployed from local laptop to cluster of servers
- Support for heterogeneous architectures
  - GPU computing with HIP, OpenCL, CUDA

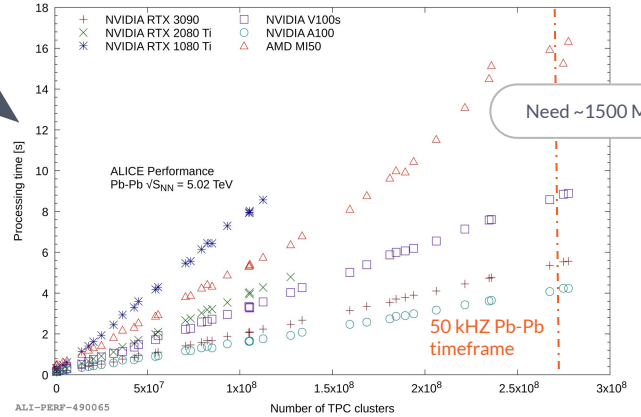



G. Eulisse, R. Shaoyan, CHEP 2019

# Synchronous reconstruction on GPU

- Runs **only EPN** servers:
  - 1x Supermicro server, 512GB of memory
  - 2x 32cores Rome CPUs
  - 8x **AMD MI50 GPUs**
- Standalone tests for GPU performance:
  - 1x GPU replaces **up to 80 Rome cores**
  - Time **linearly scales** with the size of processed data
- Full system test is started for synchronous
  - Replay of fraction of the rate expected in PbPb
  - Server sustains the rate
- Tested the expected rate for 230 servers
  - capable to process Pb-Pb data at 50 kHz with sufficient margin!




Need ~1500 MI50 GPUs



D. Rohr, vCHEP2021

# Asynchronous reconstruction on GPU

- Online reconstruction is possible thanks to GPU offloading, we aim for more
- Asynchronous reconstruction on GPU
  - TPC less relevant, lot of GPU resources available → offload larger set of algorithm on GPU
  - Some options already working, more to come. Different scenarios and priorities

**Synchronous processing**

| Processing step | % of time |
|---|---|
| TPC Processing | 99.37 % |
| EMCAL Processing | 0.20 % |
| ITS Processing | 0.10 % |
| TPC Entropy Coder | 0.10 % |
| ITS-TPC Matching | 0.09 % |
| MFT Processing | 0.02 % |
| TOF Processing | 0.01 % |
| TOF Global Matching | 0.01 % |
| PHOS / CPV Entropy Coder | 0.01 % |
| ITS Entropy Coder | 0.01 % |
| FIT Entropy Coder | 0.01 % |
| TOF Entropy Coder | 0.01 % |
| MFT Entropy Coder | 0.01 % |
| TPC Calibration residual extraction | 0.01 % |
| TOF Processing | 0.01 % |

**Asynchronous processing**

| Processing step | % of time |
|---|---|
| TPC Processing | 72.01 % |
| TRD Tracking | 12.69 % |
| TOF-TPC Matching | 9.94 % |
| MFT Tracking | 1.69 % |
| ITS Tracking | 0.78 % |
| TPC Entropy Decoder | 0.73 % |
| Secondary Vertexing | 0.69 % |
| ITS-TPC Matching | 0.56 % |
| Primary Vertexing | 0.14 % |
| TOF Global Matching | 0.11 % |
| PHOS / CPV Entropy Decoder | 0.10 % |
| FIT Entropy Decoder | 0.10 % |
| ITS Entropy Decoder | 0.06 % |
| TOF Entropy Decoder | 0.05 % |
| MFT Entropy Decoder | 0.05 % |

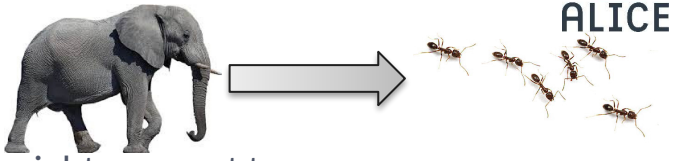
Running on GPU in baseline scenario

Running on GPU in optimistic scenario

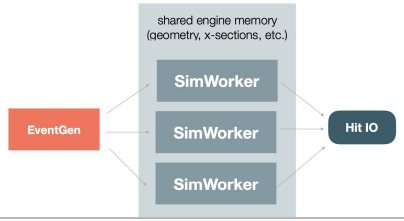
**Preliminary numbers: some algorithms not yet complete or not optimized!**

D. Rohr, vCHEP2021

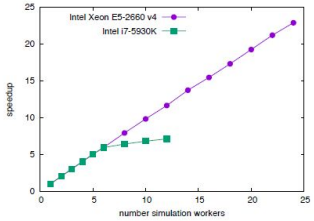
# $O^2$ simulation structure

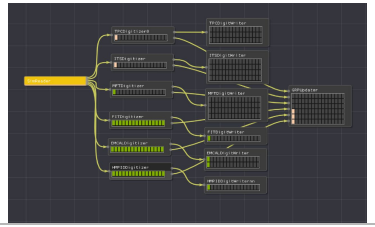


- Run 3: new detectors, readouts, dataformats and requirements, right moment to:
  - Move to *many-core* systems
  - Improve performance and efficiencies with revamped codebase layout for simulation
- Common effort with FAIR experiments, FairMQ backend to support VMC-based simulations
  - Complex and distributed systems with message passing
- Event splitting and collaborative simulation parallelism
  - Independent actors: message transfer across heterogeneous multi-processing transfer
  - Can process multiple events or split large events in parallel chunks (PbPb →O(h) time in single-core)


Sub-event simulation on multiple transport workers



Runs with (DPL): per- and in-detector parallelism

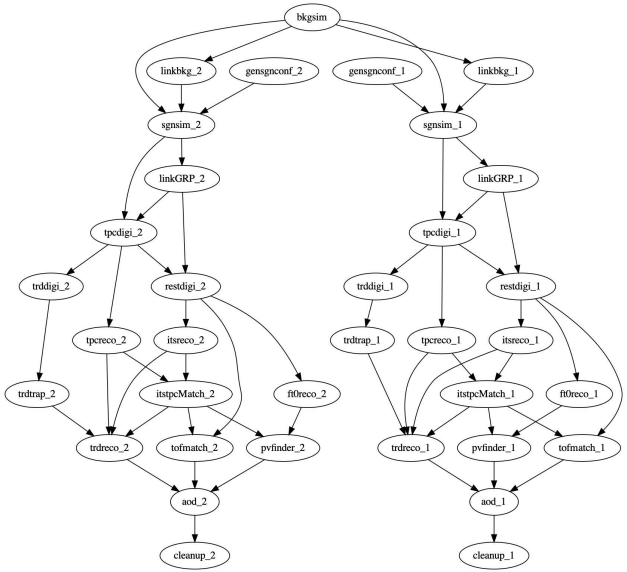
Now also on HPC!

# Novel multicore GRID executor “framework”

- ALICE globally based on multiprocessing rather than big multithreaded framework
  - sim, digi, reco different processes
- Previously (1-core ): simple shell script to call processes in order
- Move to 8-core GRID execution brings a new degree of freedom
  - more sophisticated approach to achieve high CPU efficiency

New **graph-oriented workflow modelling** and **resource-aware** GRID executor/scheduler

brings all good features of an (ETL) build system such as
- incremental execution
- restart at last failure, ...

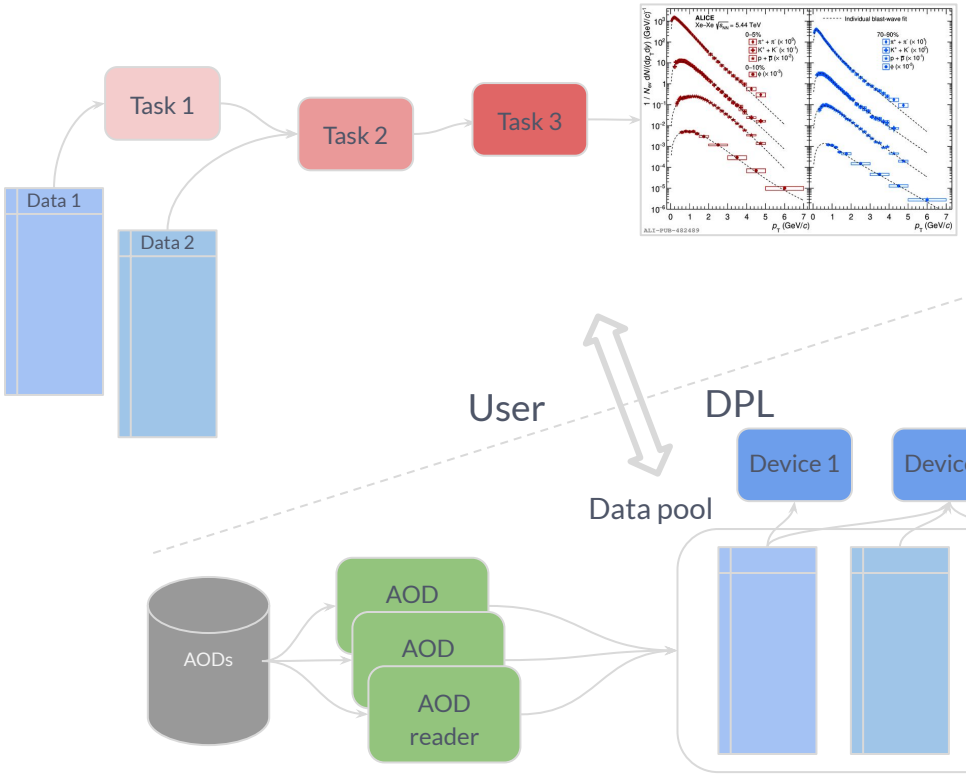



Toy graph pipeline for a 2-timeframe simulation with embedding

# Analysis model

- One month of Pb–Pb **~4 PB** of AODs
- Run 3&4 organized analysis: trains, with an improved 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) → 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 backend 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

# Analysis framework




Framework built with ***tables*** (collections of ***columns***) interconnected with indices in shared memory

- ROOT TTree when on disk
- Apache Arrow Table when in memory
- Dynamic columns (computed on the fly when requested) minimize disk usage
- Columns and Tables are represented by C++ types for the end user resulting in negligible performance overhead
- DPL translate the implicit workflow(s) to a FairMQ topology of devices

# AliHyperloop train system: current status

- New web-based interface for analysis workflow submission
- Provides:
  - Benchmarking configuration
  - Bookkeeping
  - Job submission to the GRID or Analysis Facility
  - Production of derived datasets
- Versatile platform: from Run 2 converted data to MC production for performance studies
- 30+ Physics cases already implemented to the $O^2$
- Analysis workflows tested locally, on Grid and AF (GSI)

| | Run 2 | Run 3 | Run 2 | Run 3 |
|---|---|---|---|---|
| Benchmark | AliPhysics | $O^2$ | AliPhysics | $O^2$ |
| Histograms | 3.9 | **38** | 4.5 | **9** |
| Correlations | 2 | **21** | 2.8 | **6** |
| Spectra | 4.6 | **14** | **5.4** | 4 |
| | events/s | | MB/s | |

- Multi-core capabilities of the $O^2$ workflow not accounted in comparisons
- More optimizations are foreseen, e.g. bulk reading in ROOT
- Higher throughput when enabling multi-core

# Analysis facilities

- Few dedicated sites capable to process large volumes of data in short times
  - Purpose: fast turnaround for cut tuning and task validation
- Aim: store and run on ~10% of AODs
- AF digests more than 5 PB of AODs in a 12-hour period
  - O(20000) cores and 5-10 PB of storage 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:
  - **2021 ~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
  - Wigner (former T0) coming soon

# Summary and conclusions

- In Run 3 is almost starting and ALICE is ready to face inherent computing challenges on multiple fronts to address its Physics programme
- Un-triggered data taking with online processing on GPUs
  - Working on extending their usage on asynchronous reconstruction
- Multi-core scalable simulations with event parallelism and splitting
  - HPC ready transport simulation, considerably faster digitization, efficient multicore GRID scheduler
- Brand new analysis framework with flat tables & dynamic columns: efficient data processing
  - Powered by DPL, data formats compatible with external data processing tools (ML, GPUs)
- Analysis facilities: testing and tuning of analyses to be run later on full datasets on Grid
  - GSI up and running,  Hyperloop system tested already
  - Wigner Centre (former T0) testing stage

# Backup

# 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) ) [3400 GB/s → ~100 GB/s]**
  - Full ITS+MFT clusterization [40 GB/s → ~5 GB/s]
  - 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 feed-back for reconstruction of following TFs
  - Entropy and data reduction, entropy compression (ANS), writing CTF

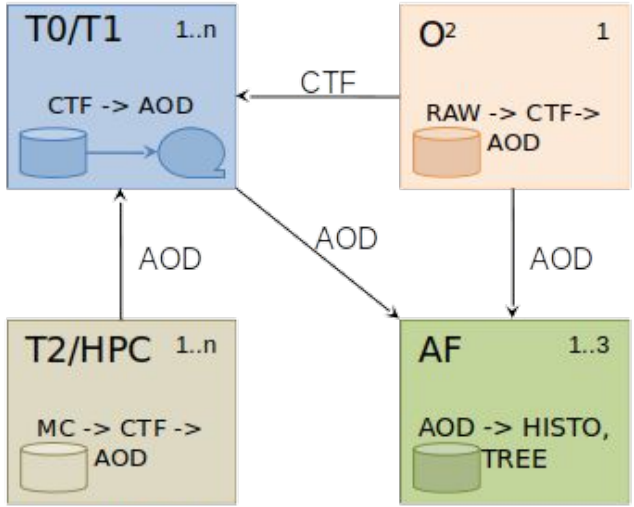
QC within EPN devices and on dedicated QC servers

D.Rohr, CHEP2019

# Computing model

**Grid Tiers will be mostly specialized for given role**

Reconstruction
Calibration
Archiving
Analysis




Calibration
Reconstruction
Compression

Simulation

Analysis

| CPU share | O2 | 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 O² + 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 movement 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 upon Physics Board approval.

O2 TDR and addendum

# Throughput Summary

Update

| | AliPhysics | | O2 | | | |
|---|---|---|---|---|---|---|
| | Local | Grid | Local | Grid | AF 2 core | AF 4 core |
| HistogramsFull | 22.7 MB/s | 4.5 MB/s | 15 MB/s | 9 MB/s | 15 MB/s | 29 MB/s |
| Transverse-mom. corr. | 12.7 MB/s | 1.8 MB/s | 5 MB/s | 5 MB/s | | |
| Two-particle correlations | 14 MB/s | 2.8 MB/s | 8 MB/s | 6 MB/s | 10 MB/s | 16 MB/s |
| J/ψ | 2.1 MB/s | 0.62 MB/s | | | | |
| Spectra | 22.7 MB/s | 5.4 MB/s | 8 MB/s | 4 MB/s | 6 MB/s | 9 MB/s |
| $\Lambda$ / $K^0$ | | | 6 MB/s | 6 MB/s | | 19 MB/s |

| | AliPhysics Grid | O2 Grid |
|---|---|---|
| HistogramsFull | 3.9 events/s | 38 events/s |
| Two-particle corr. | 2 events/s | 21 events/s |
| Spectra | 4.6 events/s | 14 events/s |

**AliPhysics → O2: Factor 3-10 higher event throughput**
**Number of reads already optimized**
**More O2 optimisation foreseen (bulk reading)**

# Retention

- 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 the tape
- The size of $O^2$ disk buffer should be sufficient accommodate CTF data from the entire period with contingency
- As soon as it is available, the CTF data will be archived to the tape and partially moved to the T1 disks

# Retention (2)

- Given the limited size of the disk buffers in O2 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 parked 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