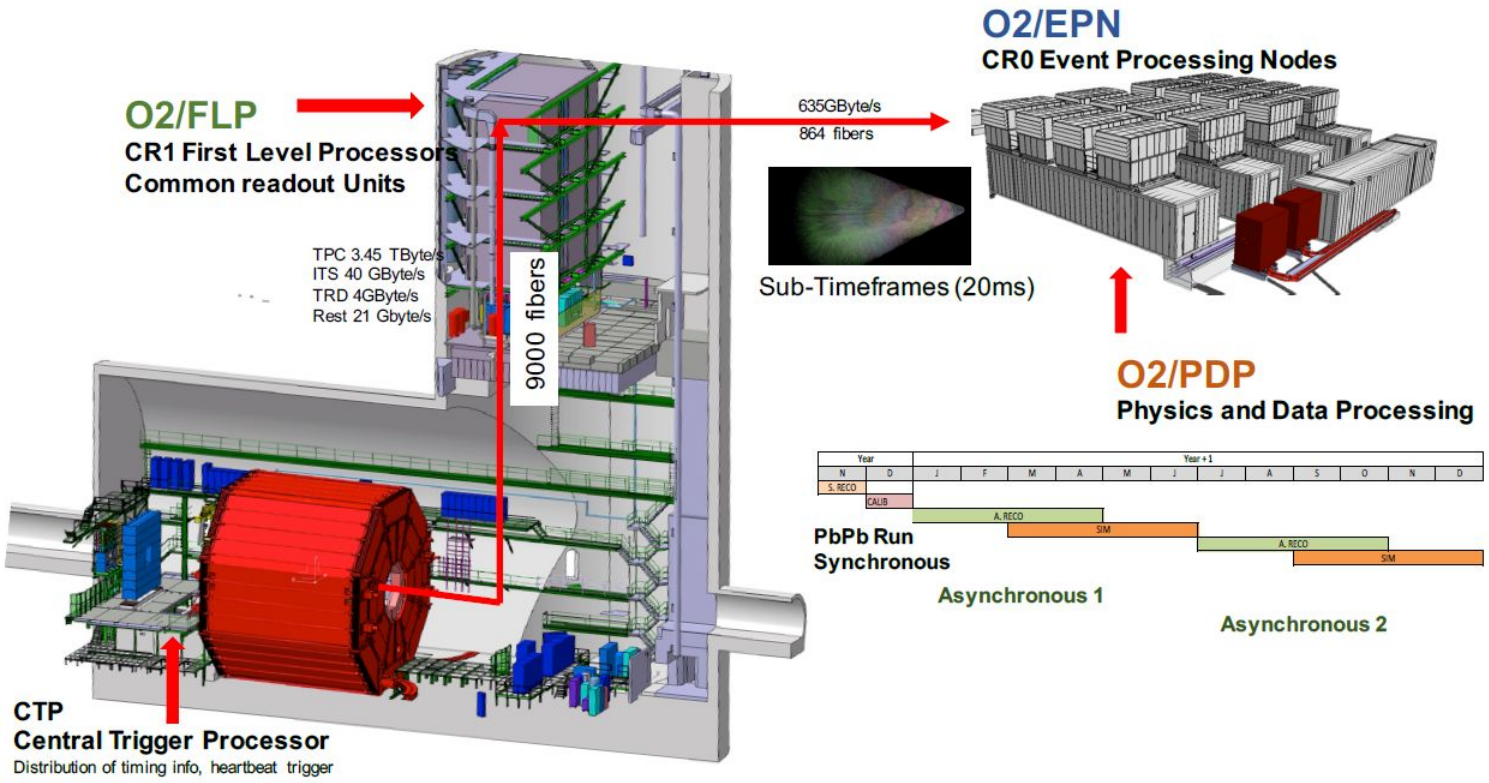
ALICE Upgrade report

L. Betev, M.Litmaath

Upgrade basics

- To be ready for Run 3 (2021)
  - The first year of Run3 will have p-p and Pb-Pb periods
- Entirely new detector readout and substantial modifications of the detector hardware
  - For example new TPC readout chambers with GEMs
- Focus on charm physics => continuous detector readout (no trigger)
  - x100 the event rate of Run1/Run2
  - No more event readout - the output is Time Frames (1000 events in one TF)
- Focus on online data compression
  - New O2 computing facility combining DAQ and Offline functions
- Reasonable rates after compression and new data processing model
  - Fit into a ‘flat budget’ resources growth scenario from the start

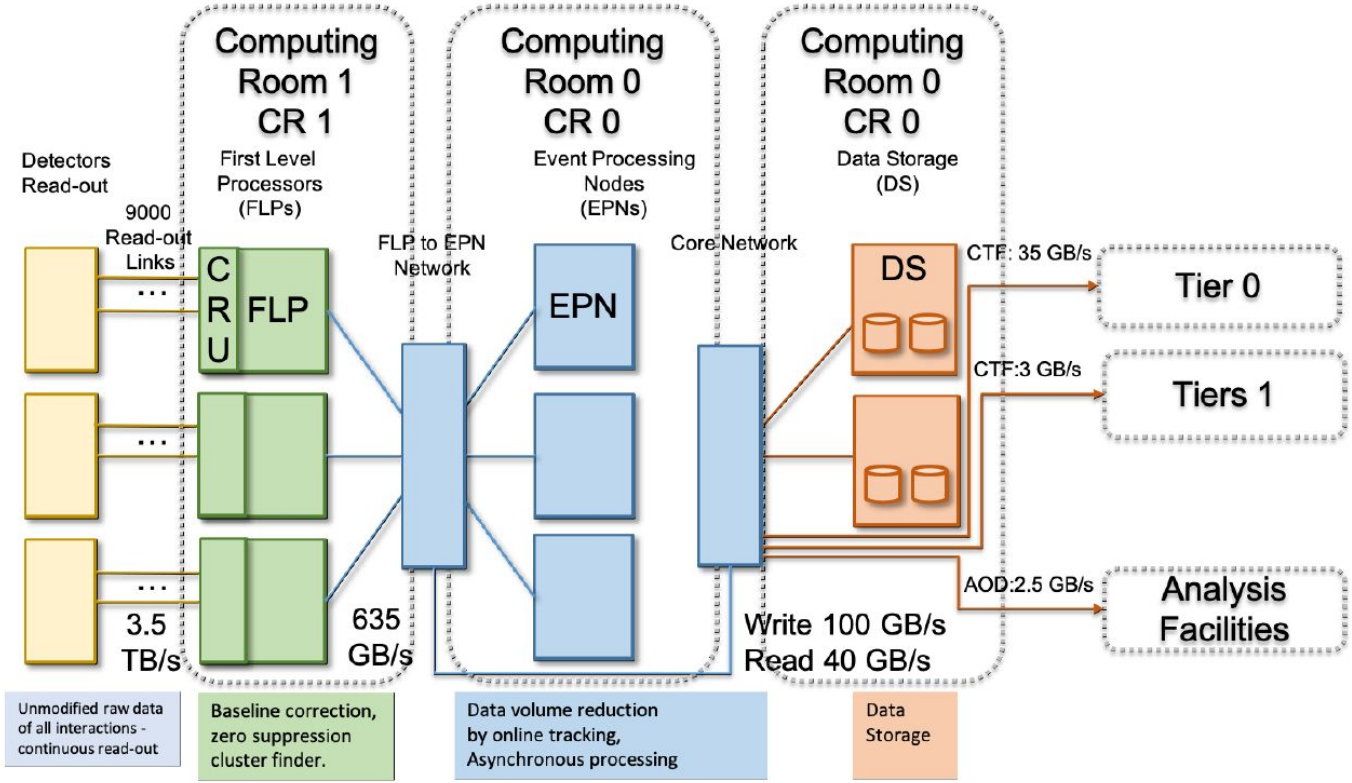
Elements and rates of the new ALICE readout



O2 elements abbreviations - synchronous processing

- Detector readout is connected to First Level Processors (**FLP**)
  - **FLP**s assemble the detector part of the continuous readout frames (**STF** - Sub-time Frames)
- **STF**s are passed on the Event Processing Nodes (**EPN**s)
  - **EPN**s apply calibration, run reconstruction and assemble the Compressed Time Frames (**CTF**s - immutable - equivalent of RAW data)
- **EPN**s record the **CTF**s on a large disk buffer
  - For subsequent asynchronous processing and writing to tape/transfers to T1s

O2 schema, location and links to the Grid



O2 - elements of the synchronous processing

- Primary O2 task - run synchronous reconstruction during data taking and assemble the Time Frames
- TPC track finding using an approximate calibration
  - 93% of the processing time
- Partial reconstruction of ITS and TRD to a level that allows precise calibration
- Removal of uninteresting portion of the event
  - Spurious signals, looper tracks
- Data compression and store to the O2 disk buffer

O2 compression factors and elements

Overall compression factor of O2 facility ~40x

Task name	CPU Time [s]	GPU Time [s]
TPC sector track finding	706	11
TPC track merging	40	2
TPC track fit	300	6
TPC looping track following	150	6
TPC data track-based compression	100	2
Sum	1296	27
ITS clustering	10	
TPC-ITS track matching	1	
Global track matching to TRD	1	
Global track matching to TOF	1	
ITS tracking	10	
ITS tracklet vertexer (seeding)	1	
ITS (MFT) data compression	3	
TPC data entropy compression	35	
TPC gain calibration	10	
TPC distortions calibration with residuals	20	
Sum	92	
Total	1388	

Emphasis on GPU algorithms for TPC reco: substantial reduction of overall processing time

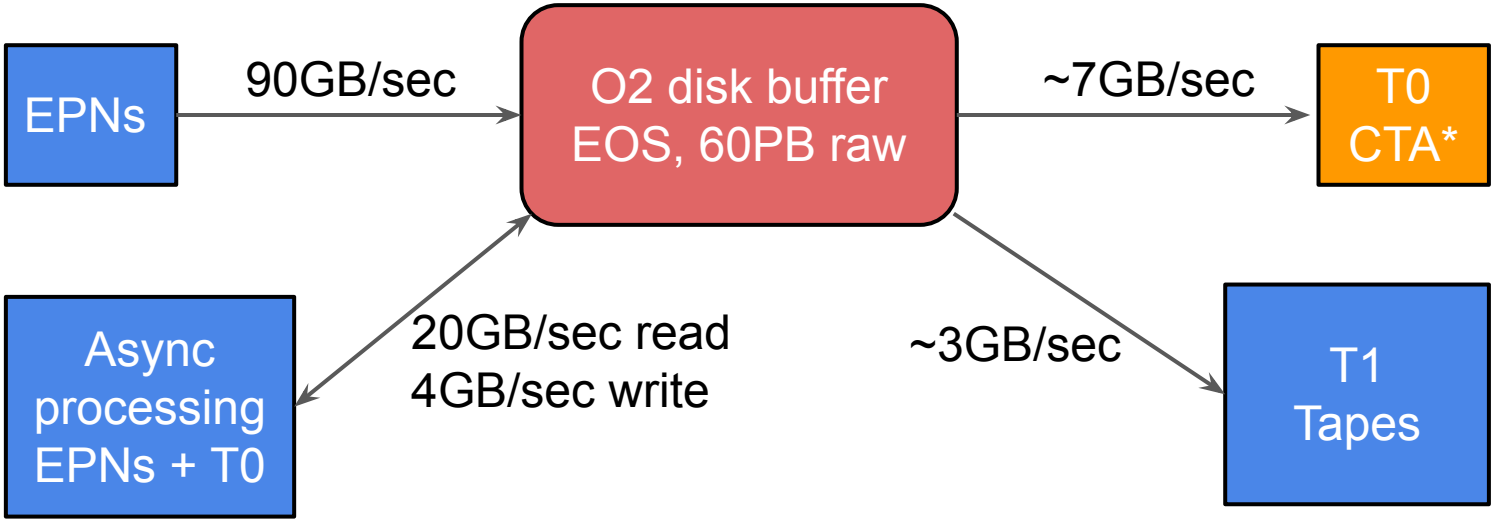
Total processing time: from timeframe to CTF < 30 sec

Timeframe content and O2 size

- Timeframe length: 20ms
  - Processing rate of 50Hz
- TF contains 1000 events @ collision rate of 50 kHz
- TF Average data volume 2GB
- O2 size @ the expected processing speed =>
  - 1500GPUs (917HS06/GPU) and 15000 CPUs (15HS06/core)
  - Processing power 1400 kHS06 (GPU) + 225 kHS06 (CPU)
  - Equivalent in power to a T1

Disk buffer

- 60PB raw capacity (some degree of safety to be included)
- Based on cheap JBODs, SATA drives, managed through EOS



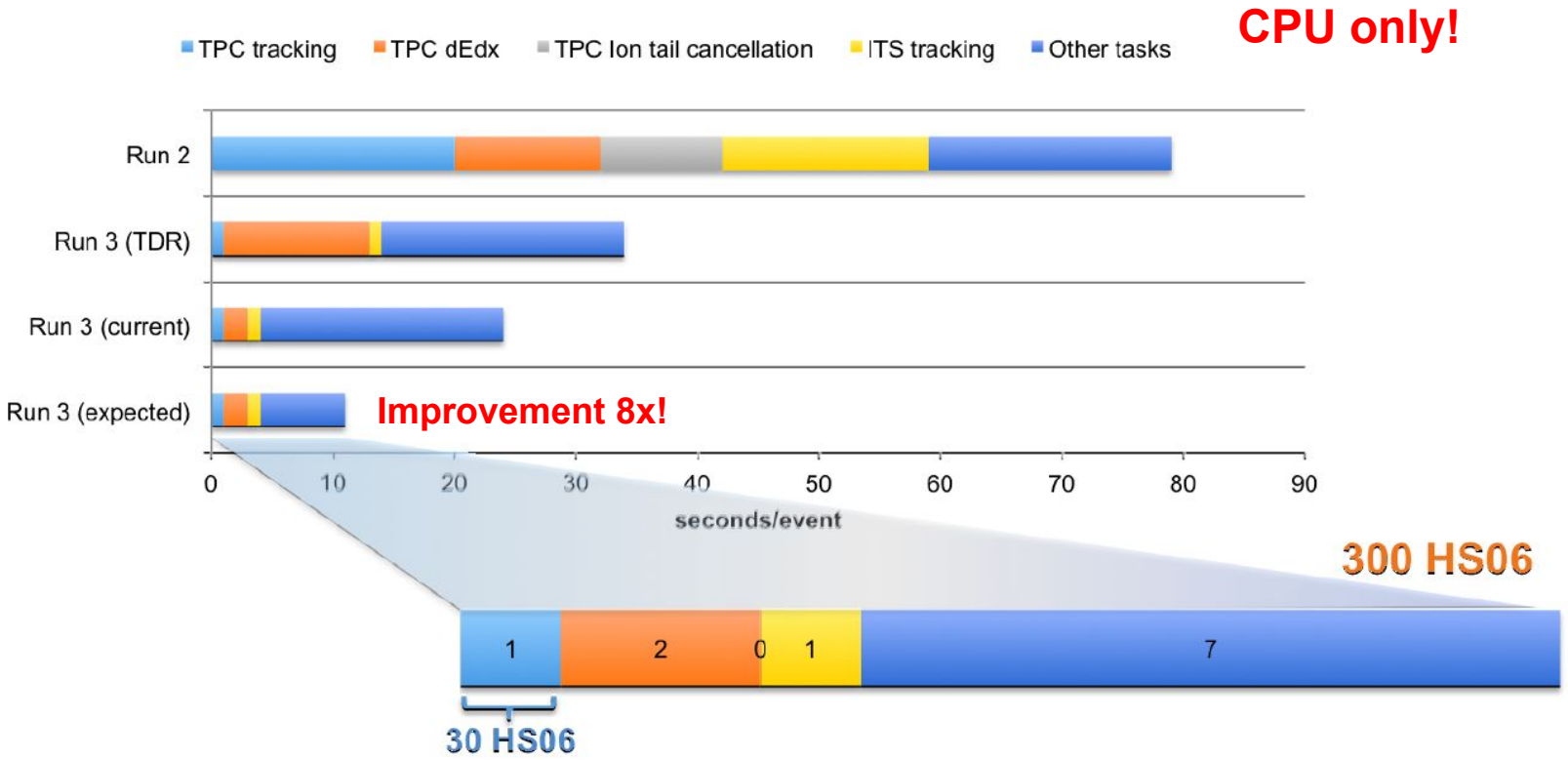
*CTA = CERN Tape Archive

ALICE

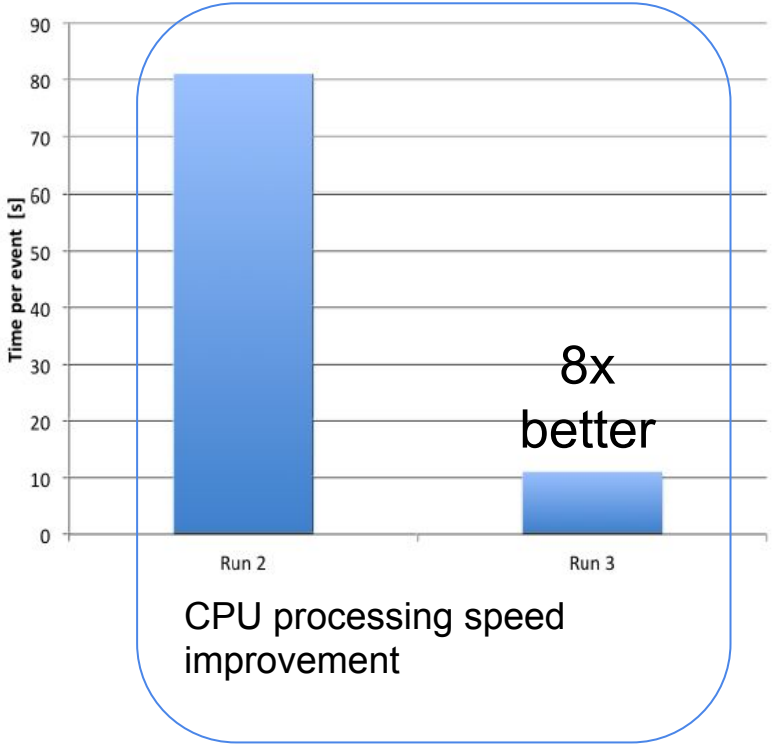
Asynchronous data processing

- Follows the data taking period
- 2 processing cycles per data taking year, with increasingly sophisticated calibration + improved reco software
- SINGLE persistent analysis object output - Analysis Object Data (AOD)
- Processing on O2+T0 (70% of CTF volume), T1s (30% of CTF volume)
- After 2-nd cycle, CTFs remain only on tape (removed from disk buffer) => any further cycle will happen only during LHC LS

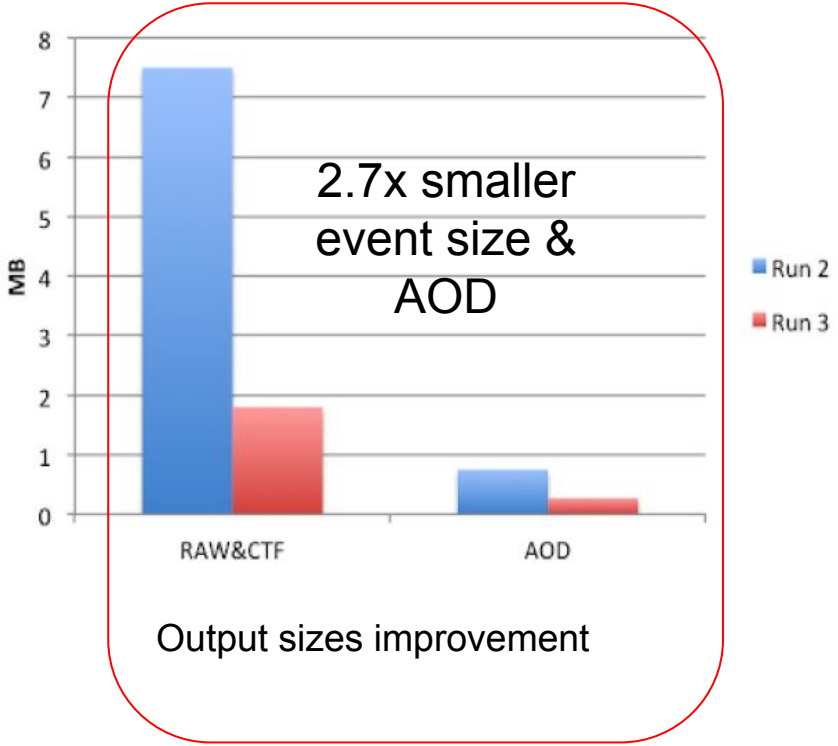
Comparison of processing algorithms (Run2-Run3)



Processing output and sizes comparison



CPU processing speed improvement



Output sizes improvement

ALICE

Software framework subdivisions

- Transport Layer
  - Uses FairMQ message passing toolkit (GSI development)
  - Abstracts the network fabric
  - Defines the core building blocks in terms of devices
  - Implements the communication between them
- O2 Data Model;
  - ALICE-specific description of the messages between devices
  - Computer language agnostic, extensible, efficient mapping of the data objects in shared memory or to the GPU memory
  - Supports multiple data formats and serialization methods

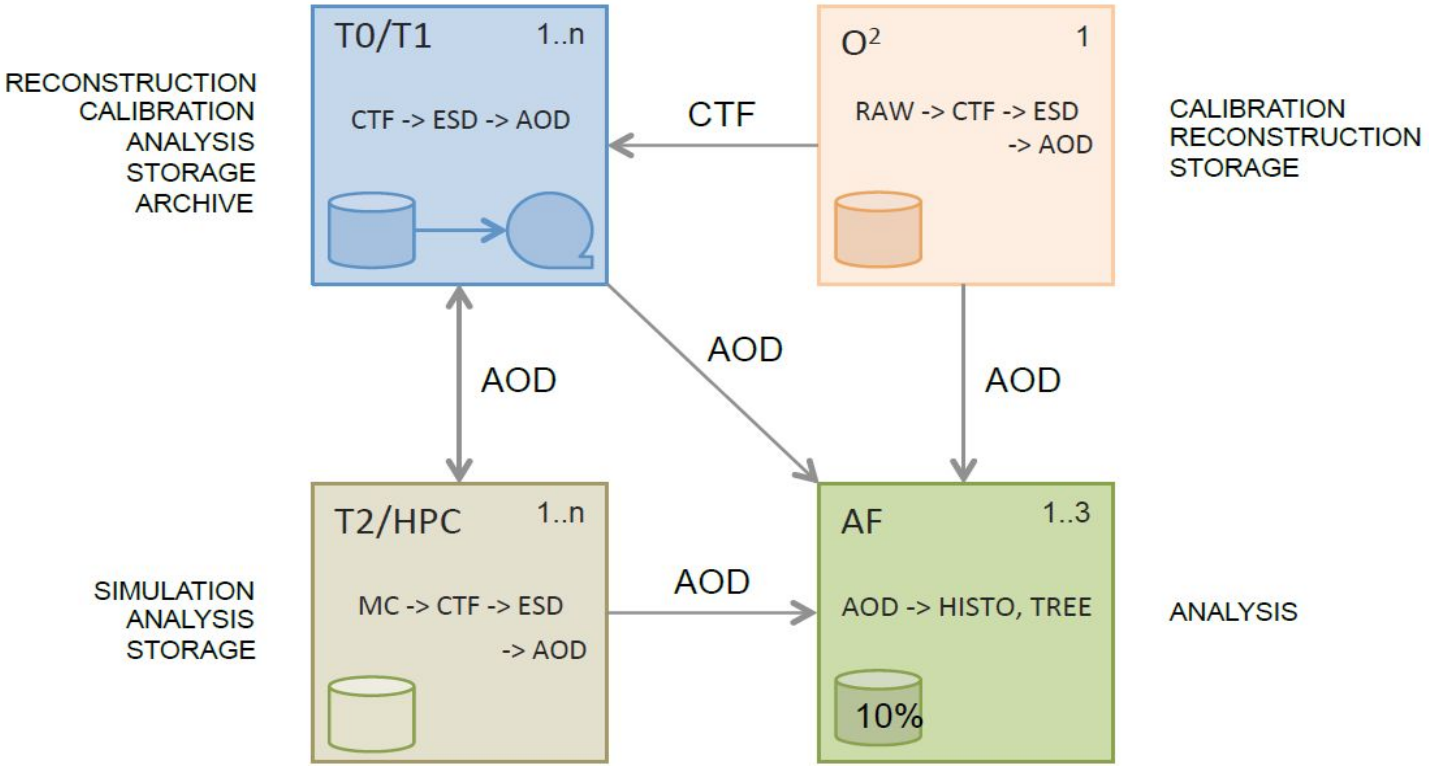
Software framework subdivisions (2)

- Data Processing Layer
  - Simplifies the life of the end user
  - Allows to describe computation as a set of data processors implicitly organized in a logical data flow transformation
  - A defined data flow is run by a single executable - the DPL driver
  - Includes a powerful GUI for logs/metrics and debugging
    - Especially helpful for individual users

Upgrades of Grid middleware: AliEn ⇨ jAliEn

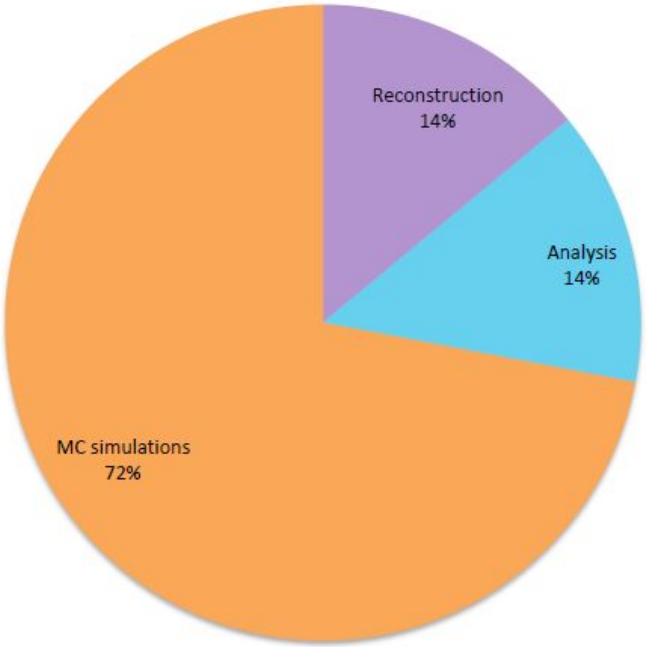
- Substantial rewrite of the system - all top-level and site-level (VO-box) parts are new, with new communication protocol
- More sophisticated data management services - easier to replicate data/reclaim storage
- JobAgent/Jobwrapper with user-switching and container-ready
- Entirely new and faster central catalogue
  - Uses Cassandra/Scylla backend
  - Tested to full speed demanded by the future workflow
- Complete ROOT integration
  - Allowing all interactions with the Grid from the ROOT shell
- Gradual replacement of the existing system - new services in operation as soon as ready

Computing model in a single figure

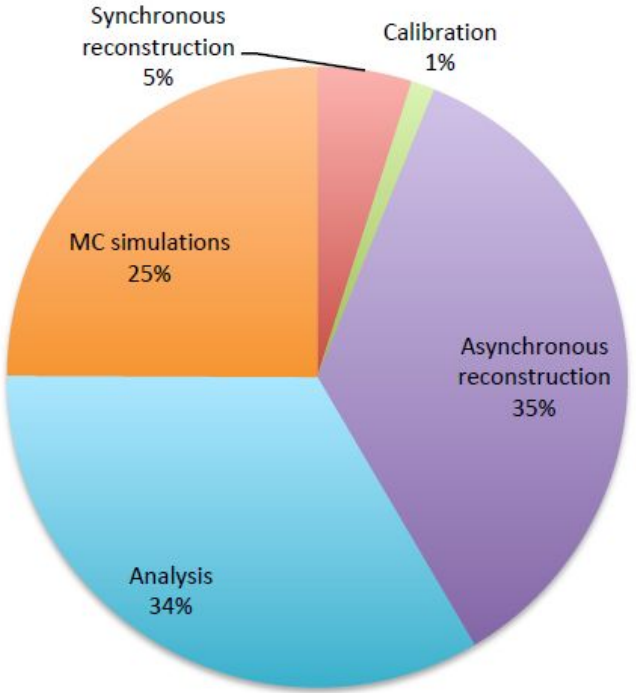


Resources share projection

Today



Run3+



Resources requirements projection

- Projections based on discrete resources simulation, including workflows, detector performance and LHC beam schedule show that all resources growth (without tapes) - compatible with ***flat budget*** scenario

Summary

- ALICE is in the critical phase of the Run3 upgrade preparation
- All building blocks of the upgraded system are defined and work is ongoing
- Substantial changes in the online and offline software, coalescing into a single framework and a new O2 compression facility
  - Re-written in large part
  - Time-critical algorithms ported to GPU to gain speed
  - Purpose-built facility with balanced CPU/GPU component and large storage
- New top-level Grid middleware adapted to the increased processing demands
- 1 ½ years remaining to complete the project
- Resources requirements are well understood, scrutinized and approved
- New software algorithms and computing model allow to fit into the standard Grid resource growth