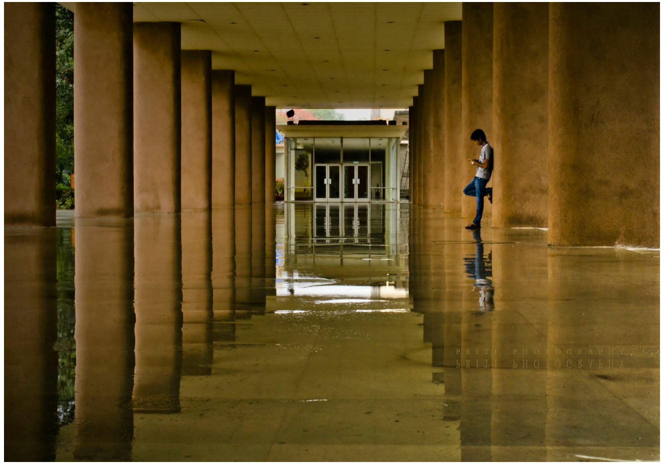
ALICE Computing Outlook for RUN3

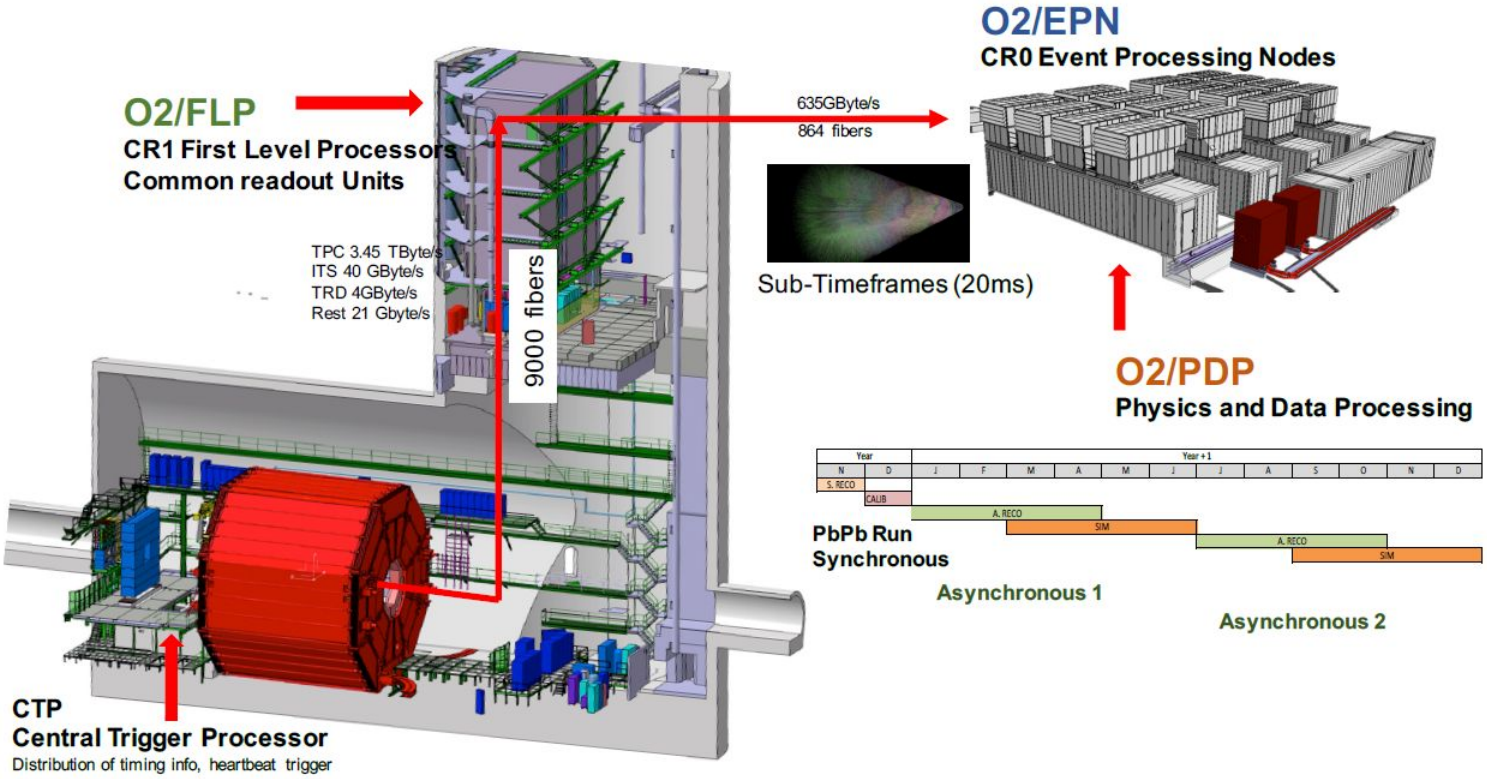
5th Asian Tier Center Forum
TIFR, Mumbai
25 October 2019



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/FLP

CR1 First Level Processors
Common readout Units

TPC 3.45 TByte/s
ITS 40 GByte/s
TRD 4GByte/s
Rest 21 Gbyte/s

9000 fibers

635GByte/s
864 fibers

O2/EPN

CR0 Event Processing Nodes

Sub-Timeframes (20ms)

O2/PDP

Physics and Data Processing

PbPb Run
Synchronous

Asynchronous 1

Asynchronous 2

CTP

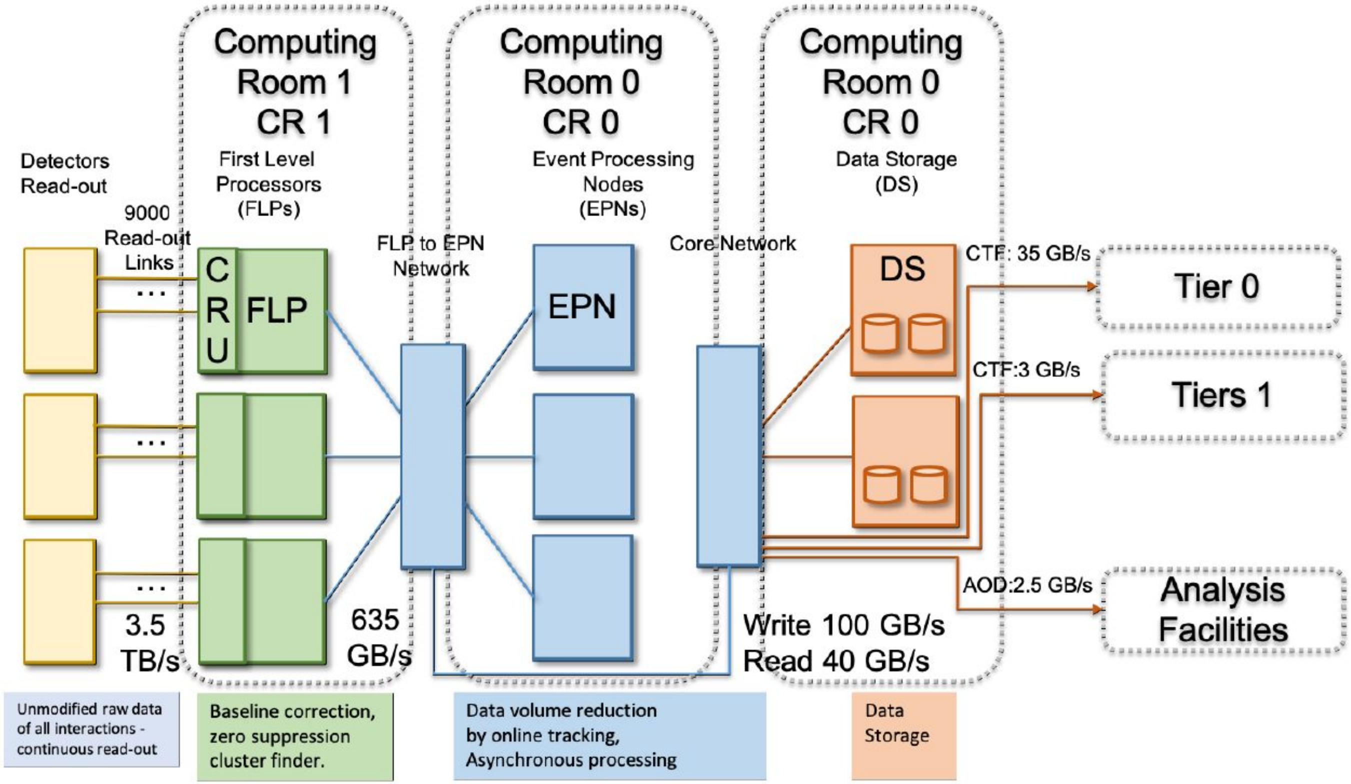
Central Trigger Processor

Distribution of timing info, heartbeat trigger

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 to RAW data)
- **EPN**s record the **CTF**s on a large disk buffer
  - For subsequent asynchronous processing and writing to tape/transfers to T0, T1s

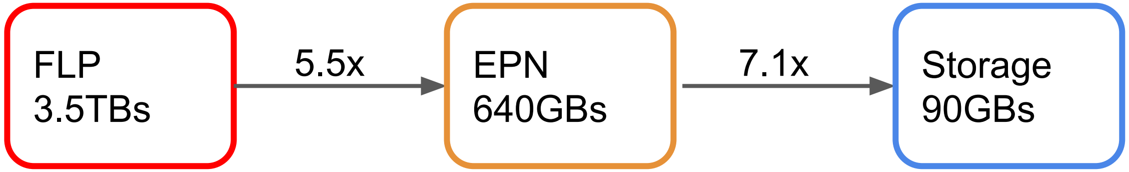
O2 schema, location and links to the Grid



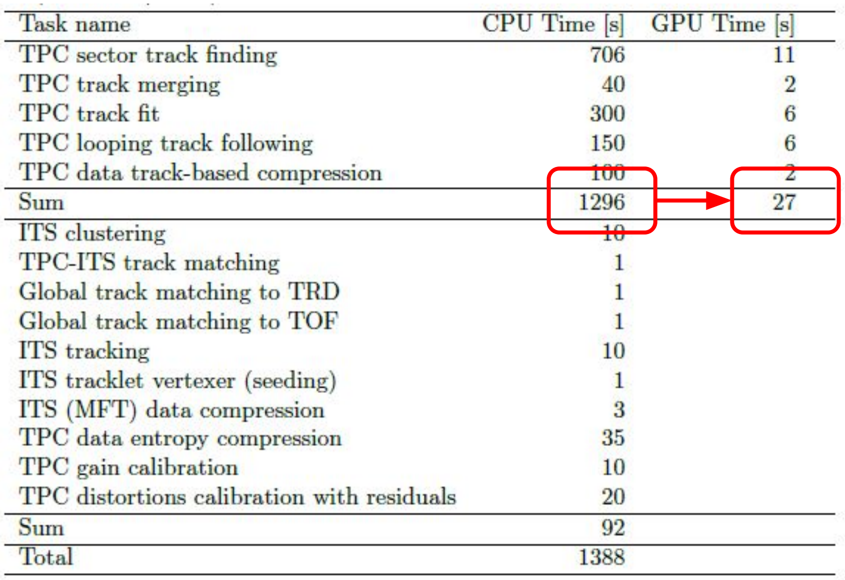
O2 - elements of the synchronous processing

- 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 for 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

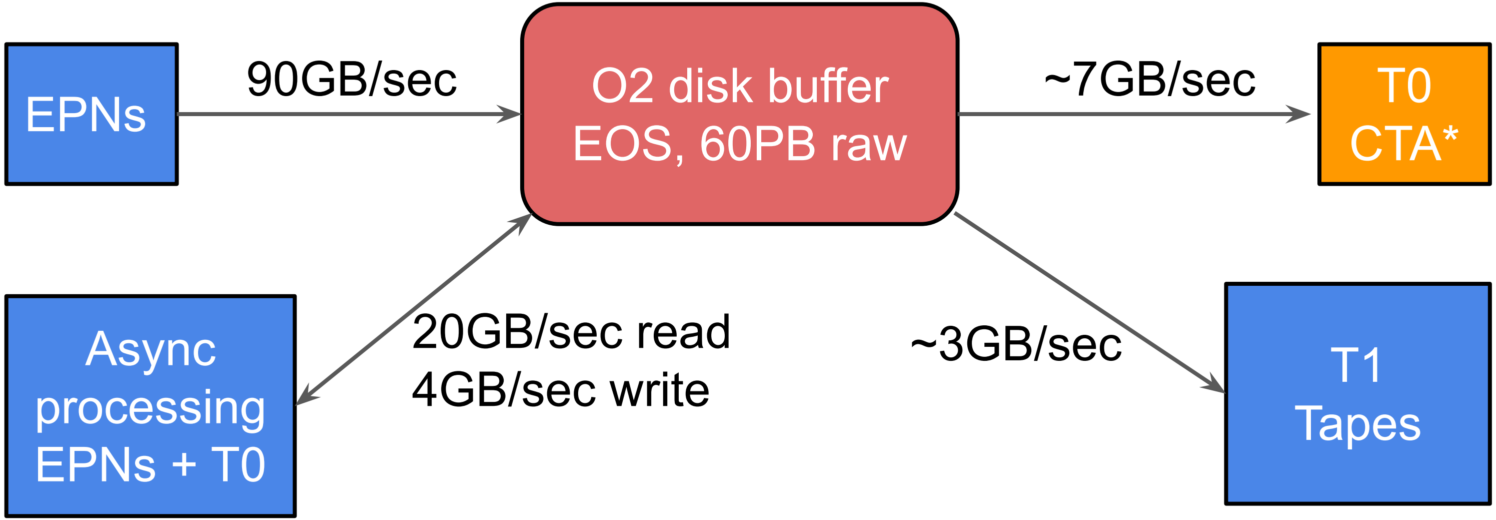
Total processing time: from timeframe to CTF < 30 sec

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 (numbers below are still being optimized) =>
  - 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



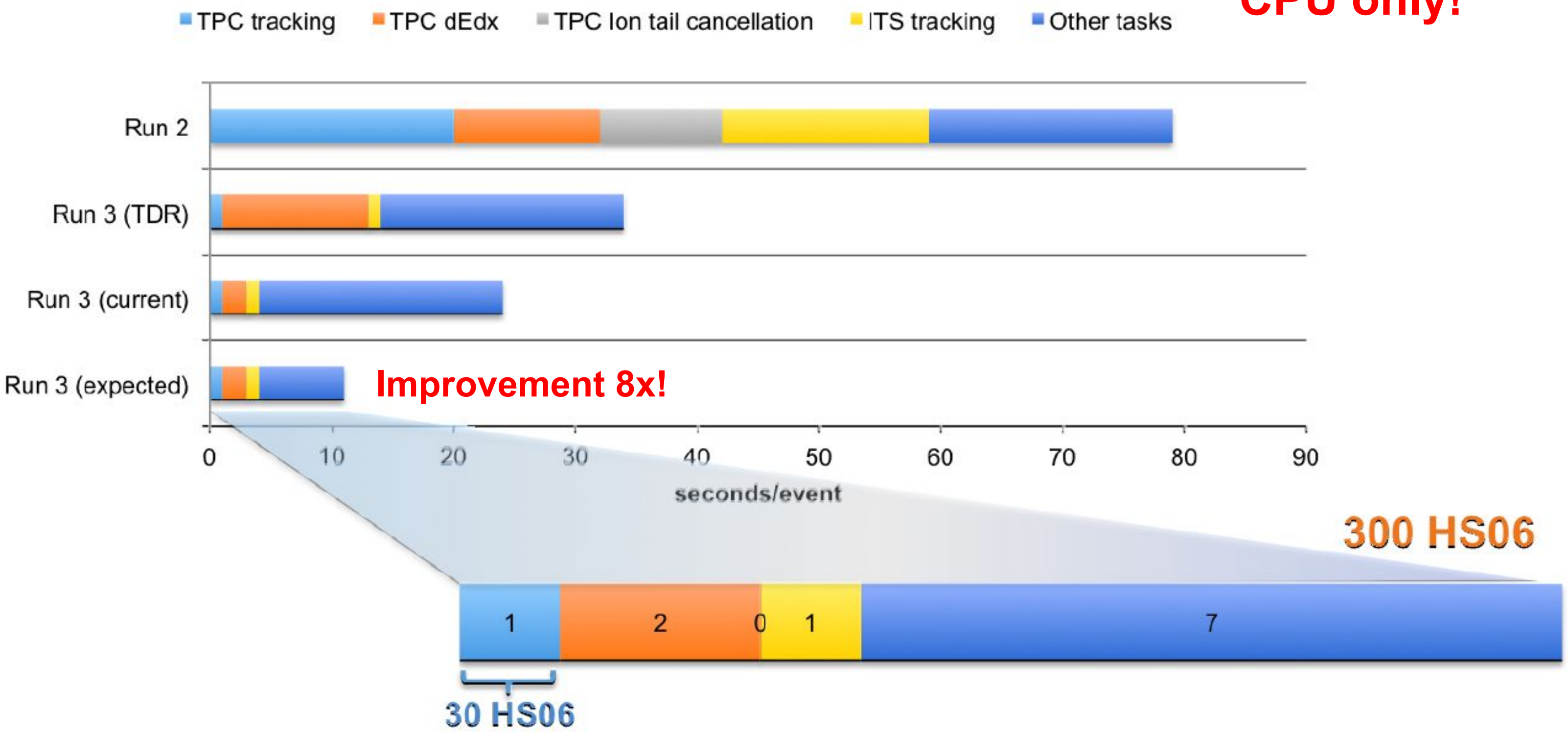
*CTA = CERN Tape Archive

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)

CPU only!

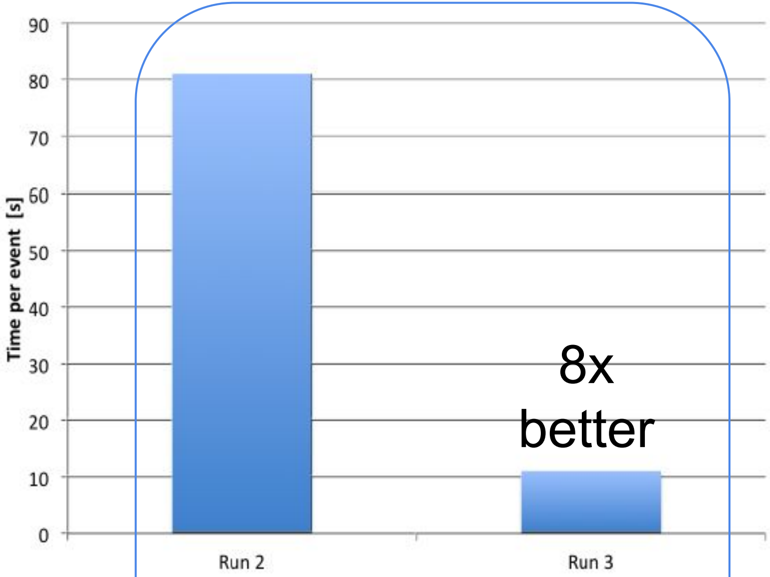


Improvement 8x!

300 HS06

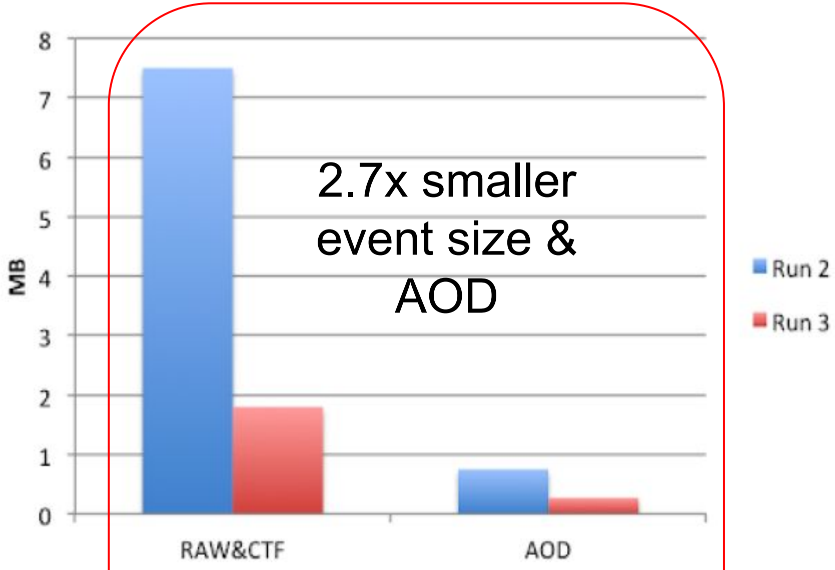
30 HS06

Processing output and sizes comparison



8x better

CPU processing speed improvement



2.7x smaller event size & AOD

Output sizes improvement

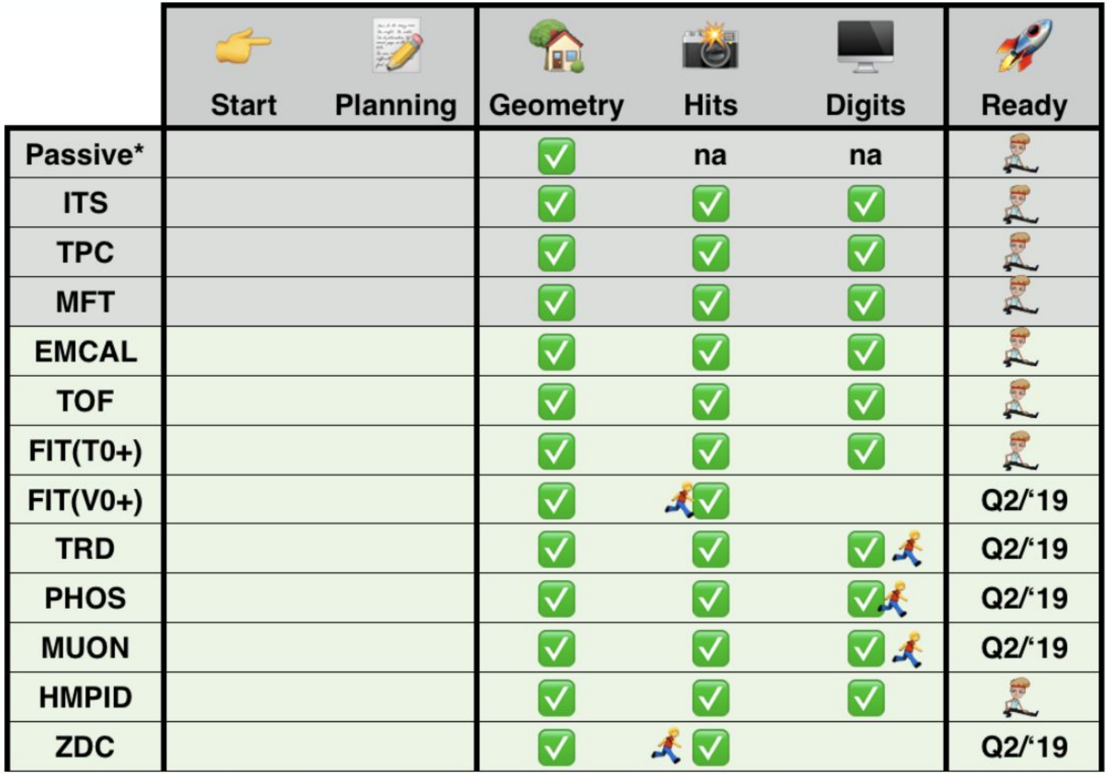
Focus on upgrade for Run 3 - Simulation

- O2 TDR assumption leading to estimate of
  - 5×10^8 central Pb-Pb
  - Using signal injection/embedding: > 2×10^{10} MB equivalent
  - Trivial scaling with number of events
- Large opportunity for non-trivial optimisation
  - In depth evaluation of the Physics Working Groups requests for simulation
  - Should lead to <1% MC errors
- 2018 Pb-Pb analysis will serve as benchmark and trigger more discussions

Simulation - list of major items

- Embedding techniques
  - Computing time reduction - digitisation and avoid redundant calculation steps
  - AOD size reduction
- Review of O2 TPC digitization code
  - Substantial improvement in CPU performance while keeping constant physics quality
  - Largely simplified GEM Amplification scheme
- Optimization of transport time (G3/G4)
  - Transport cuts and geometry configuration depending on physics
- Virtual MC supports simulation using several transport engines
  - Integrate fast and slow simulation
  - Physics models from different (or the same) transport engines
- More and specific details on the ongoing MC work - September report

Simulation - status of detector implementation

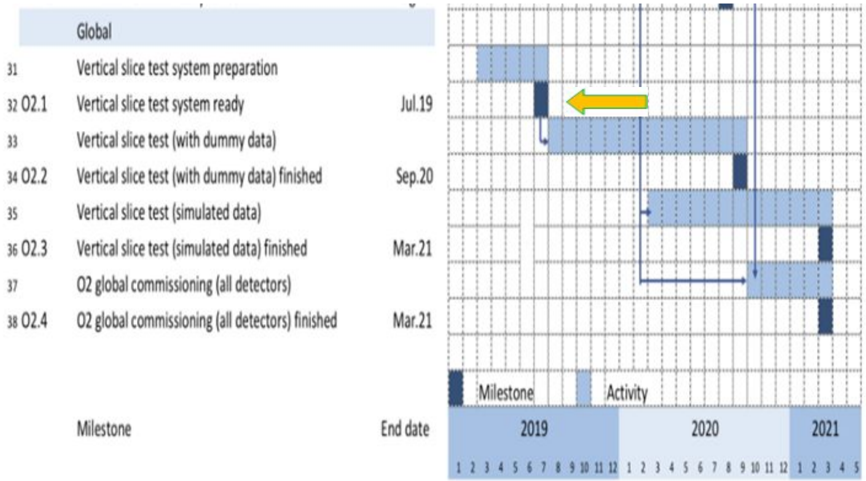


	👉 Start	📝 Planning	🏡 Geometry	📸 Hits	🖥 Digits	🚀 Ready
Passive*			✓	na	na	🧘
ITS			✓	✓	✓	🧘
TPC			✓	✓	✓	🧘
MFT			✓	✓	✓	🧘
EMCAL			✓	✓	✓	🧘
TOF			✓	✓	✓	🧘
FIT(T0+)			✓	✓	✓	🧘
FIT(V0+)			✓	🏃✓		Q2/‘19
TRD			✓	✓	✓ 🏃	Q2/‘19
PHOS			✓	✓	✓🏃	Q2/‘19
MUON			✓	✓	✓ 🏃	Q2/‘19
HMPID			✓	✓	✓	🧘
ZDC			✓	🏃 ✓		Q2/‘19

Excellent progress, main parts ready for Vertical Slice Test (next slides)

Next major exercise - the Vertical Slice Test

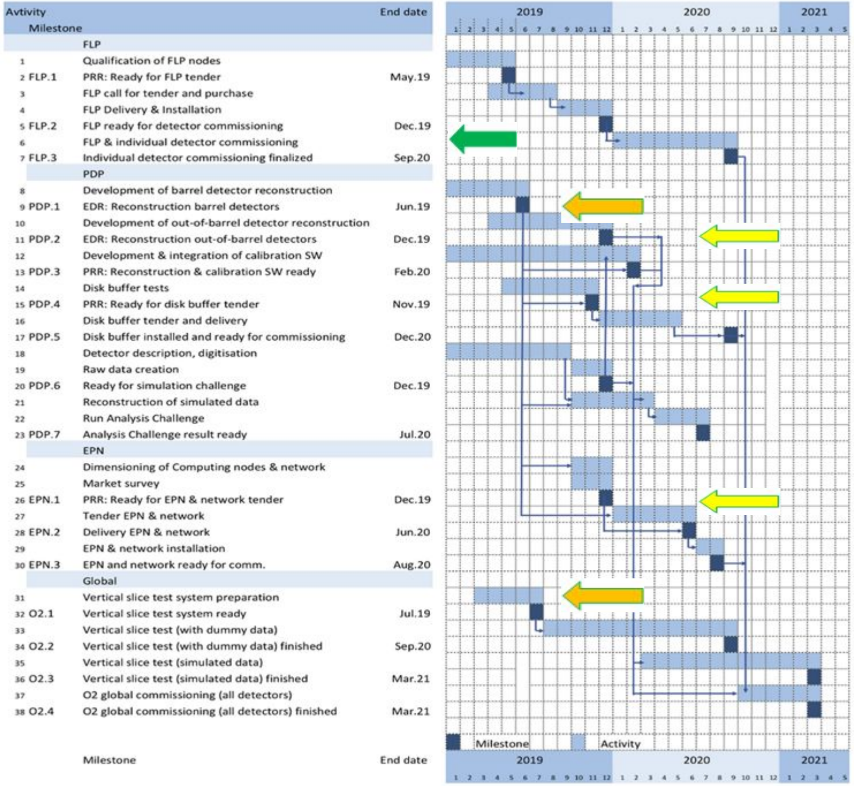
- New O2 project structure
  - FLP (Pierre Vande Vyvre), EPN (Volker Lindenstruth) and PDP (Andreas Morsch)
  - Computing Technical Coordination (Massimo Lamanna)
- Vertical Slice Test
  - Continuity tests
  - Initial core for the detectors commissioning
  - Initial core for the O2 final deployment
- Status of O2.1 milestone
  - New containers wiring advancing (EPN+FLP) + necessary connectivity to the Computer Centre
  - ~10% capacity, recycling Run2 computing equipment
  - On-track to have all the elements on the floor by July



		Milestone	End date
		Global	
31		Vertical slice test system preparation	
32	O2.1	Vertical slice test system ready	Jul.19
33		Vertical slice test (with dummy data)	
34	O2.2	Vertical slice test (with dummy data) finished	Sep.20
35		Vertical slice test (simulated data)	
36	O2.3	Vertical slice test (simulated data) finished	Mar.21
37		O2 global commissioning (all detectors)	
38	O2.4	O2 global commissioning (all detectors) finished	Mar.21

O^2 major milestones

- Done
  - FLP.1 (Ready for FLP tender) review D. Francis, F. Mejers, N.Neufeld + input from IT-CF/E. Bonfillou and IPT-PI/H. Gerster and F. Najeh
  - Concluded on May the 9th, Tender being validated
- Imminent
  - PDP.1 (Reco barrel detectors) review S. Ponce, F. Pantaleo and G. Stewart
- Highlights of future ones
  - PDP.2 Out-of-Barrel detectors (Dec 2019)
  - PDP.4 Ready for disk buffer tender" (Nov 2019) - Collaboration with IT restarted (based on the 2018 disk-buffer for HI daq)
  - EPN.1 Dimensioning of computing nodes and network (Dec 2019)



Avtivity	Milestone		End date
		FLP	
1		Qualification of FLP nodes	
2	FLP.1	PRR: Ready for FLP tender	May.19
3		FLP call for tender and purchase	
4		FLP Delivery & Installation	
5	FLP.2	FLP ready for detector commissioning	Dec.19
6		FLP & individual detector commissioning	
7	FLP.3	Individual detector commissioning finalized	Sep.20
		PDP	
8		Development of barrel detector reconstruction	
9	PDP.1	EDR: Reconstruction barrel detectors	Jun.19
10		Development of out-of-barrel detector reconstruction	
11	PDP.2	EDR: Reconstruction out-of-barrel detectors	Dec.19
12		Development & integration of calibration SW	
13	PDP.3	PRR: Reconstruction & calibration SW ready	Feb.20
14		Disk buffer tests	
15	PDP.4	PRR: Ready for disk buffer tender	Nov.19
16		Disk buffer tender and delivery	
17	PDP.5	Disk buffer installed and ready for commissioning	Dec.20
18		Detector description, digitisation	
19		Raw data creation	
20	PDP.6	Ready for simulation challenge	Dec.19
21		Reconstruction of simulated data	
22		Run Analysis Challenge	
23	PDP.7	Analysis Challenge result ready	Jul.20
		EPN	
24		Dimensioning of Computing nodes & network	
25		Market survey	
26	EPN.1	PRR: Ready for EPN & network tender	Dec.19
27		Tender EPN & network	
28	EPN.2	Delivery EPN & network	Jun.20
29		EPN & network installation	
30	EPN.3	EPN and network ready for comm.	Aug.20
		Global	
31		Vertical slice test system preparation	
32	O2.1	Vertical slice test system ready	Jul.19
33		Vertical slice test (with dummy data)	
34	O2.2	Vertical slice test (with dummy data) finished	Sep.20
35		Vertical slice test (simulated data)	
36	O2.3	Vertical slice test (simulated data) finished	Mar.21
37		O2 global commissioning (all detectors)	
38	O2.4	O2 global commissioning (all detectors) finished	Mar.21
		Milestone	End date

Figure 5.1: O2 milestones.

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

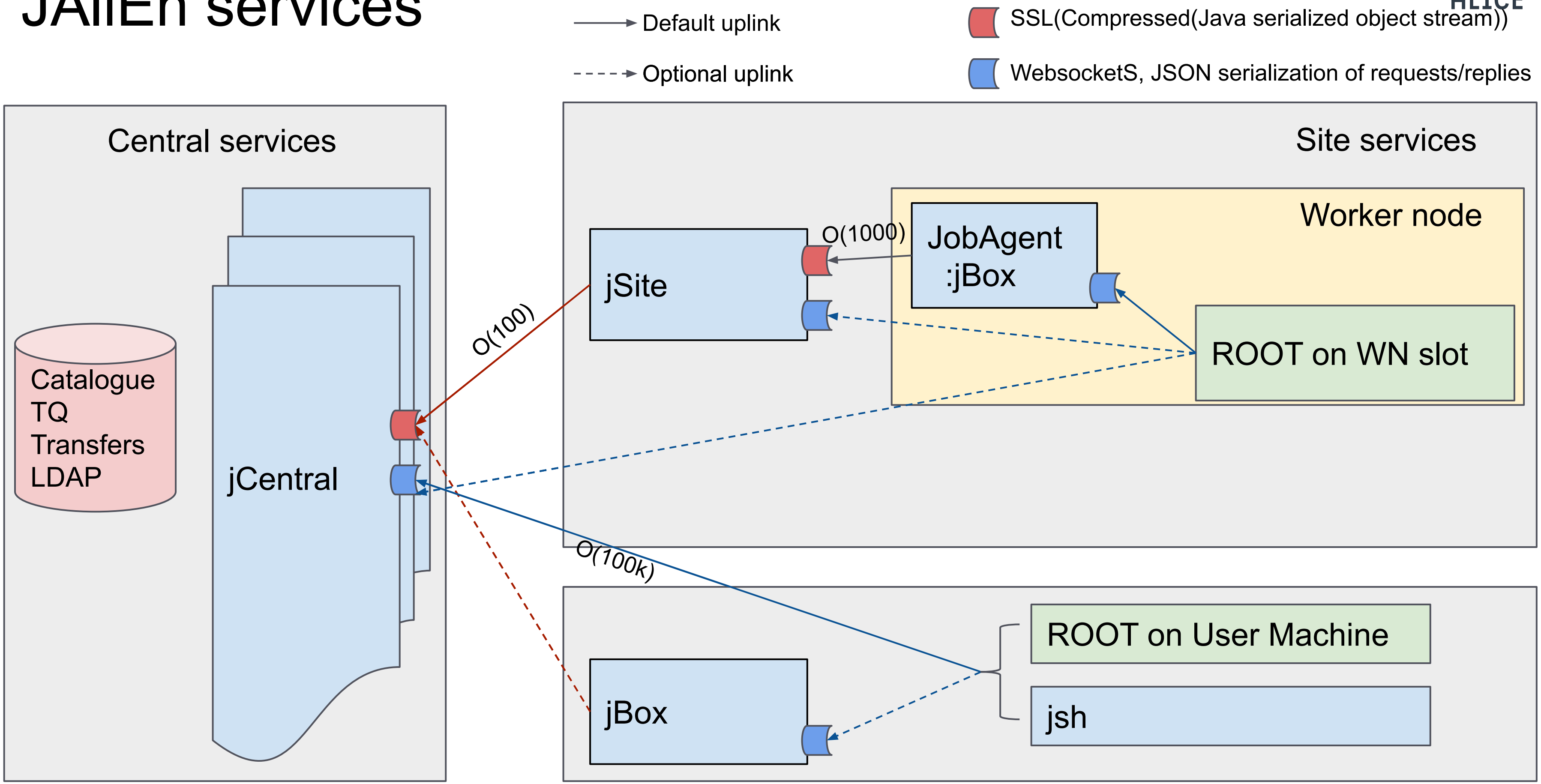
Software framework subdivisions (2)

- 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 (many already in production)

JAliEn services



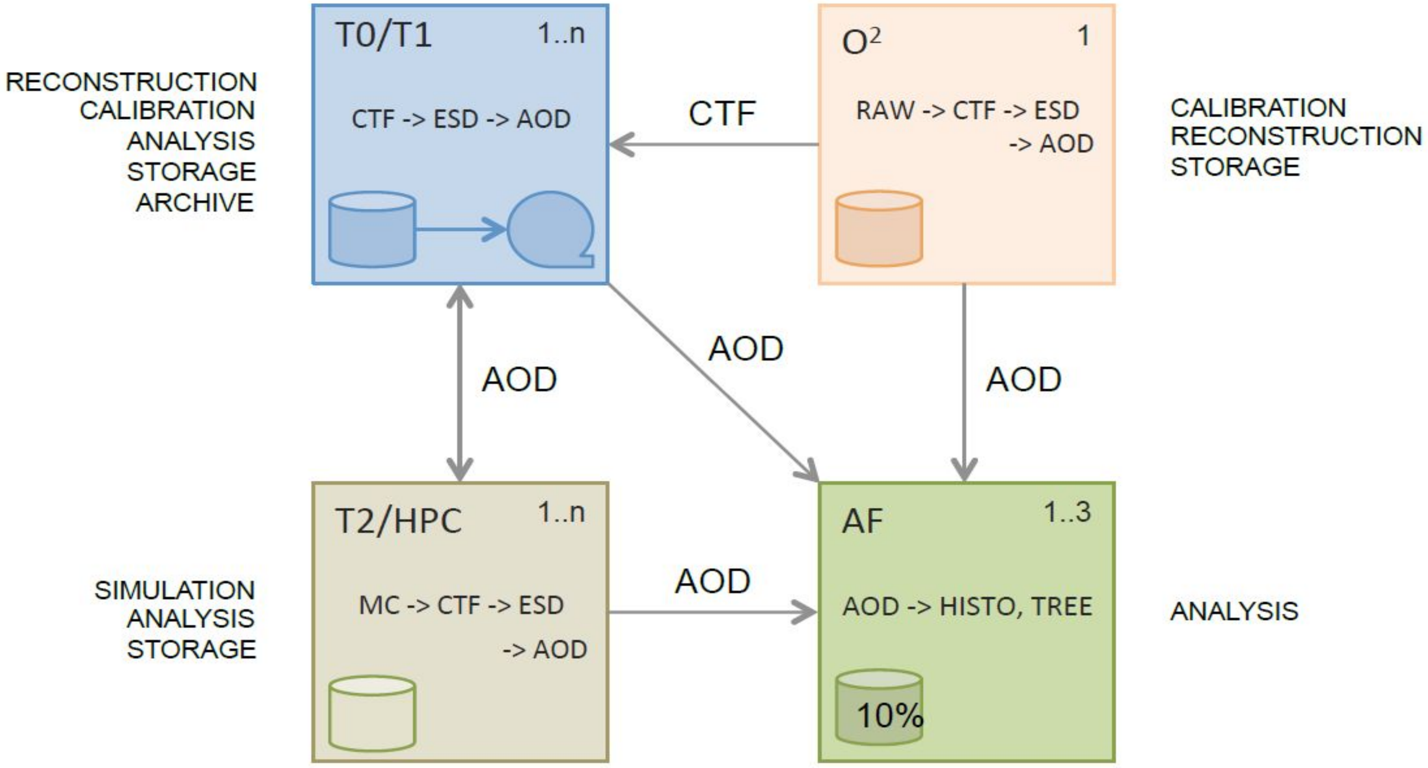
Websockets and deployment

- Websockets provide full-duplex communication channel over a single TCP connection
- Persistent channel, suitable for heavy load, low latency applications
- ROOT implementation: based on libwebsockets, an open source library available in all popular linux distributions
- Secure connections based on OpenSSL
- Embedded Tomcat server providing the websockets server endpoint
  - fixed port no. for central services
  - dynamic port no. for WN/user desktop instances
- ROOT plugin loads identity and server addr:
  - from environment (child process of JobAgent)
  - from *$TMPDIR/jalien_token_<uid>* (user desktop)
  - default locations (*~/.globus/user{cert,key}.pem* and *alice-jcentral.cern.ch:443*, for standalone ROOT instances)

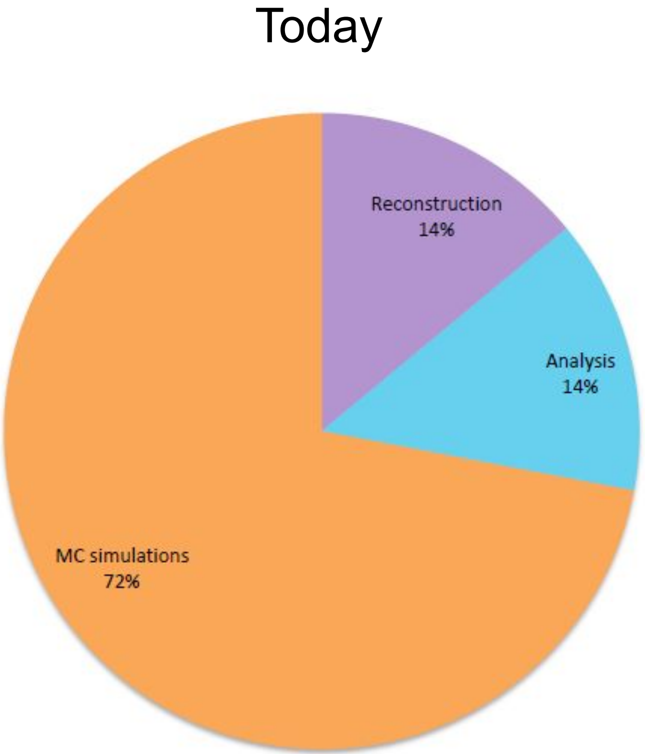
Flexible deployment schema

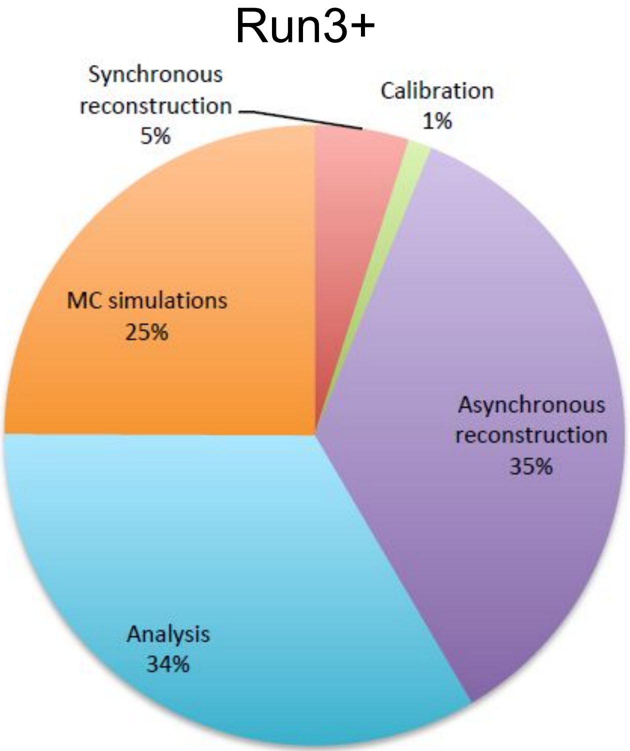
- Same codebase, each level multiplexes connections and caches objects
- Dual personality servers
  - Java binary serialization + SSL and compression
  - Efficient channel for inter-service communication
  - Asynchronous messages passed between endpoints
  - Websockets + SSL
  - End-clients (ROOT, custom clients)
- Both are long-lived, persistent connections
- Several sites are already running JAliEn in mixed mode with the old AliEnservices
- Central services fully deployed (without catalogue)
- All data management is now done through JAliEn

Computing model in a single figure



Resources share projection





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

		2019			**2020**			**2021**	
ALICE		Req.	C-RSG	Pledge	Req.	Pledge	2020 Pledge/Req.	Req.	2021/2020 Req
CPU	Tier-0	430	430	350	350	350	0.0%	471	34.6%
	Tier-1	365	365	331	365	353	-3.2%	498	36.4%
	Tier-2	376	376	370	376	410	8.9%	515	37.0%
	Total	**1171**	**1171**	**1051**	**1091**	**1113**	**2.0%**	**1484**	**36.0%**
Disk	Tier-0	34.3	34.3	31.2	31.2	31.2	0.0%	45.5	45.8%
	Tier-1	37.9	37.9	35.1	44	41.8	-5.0%	53.3	21.1%
	Tier-2	33.9	33.9	33.5	39	41.0	5.0%	44.8	14.9%
	Total	**106.1**	**106.1**	**99.8**	**114.2**	**114.0**	**-0.2%**	**143.6**	**25.7%**
Tape	Tier-0	44.2	44.2	44.2	44.2	44.2	0.0%	80.0	81.0%
	Tier-1	37.7	37.7	41.1	37.7	44.4	17.8%	55.0	45.9%
	Total	**81.9**	81.9	**85.3**	**81.9**	**88.6**	**8.2%**	**135.0**	**64.8%**

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
- This summer - Vertical Slice (~10%) comprehensive test of the entire data acquisition, simulation and processing chain
- 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