Feed-back from ALICE on potential mitigation measures for computing resources shortage

---

Thursday, February 16, 2017

Table of Contents

ALICE Run 2 Physics goals and collected statistics

The goal for Run 2 is to reach the integrated luminosity of 1 $nb^{-1}$of Pb-Pb collisions for which the ALICE scientific program was originally approved and the corresponding pp and p-Pb reference data. Targeting this objective for Run 2 will allow the experiment to extend the reach of several measurements crucial for the understanding of the basic properties of the QGP and consolidate preliminary observations from Run 1 data.

The objectives are as follows:

- For Pb-Pb collisions:
    - Reach the target of 1 nb^{-1} integrated luminosity in Pb-Pb for rare triggers.
    - Increase the statistics of the unbiased data sample, including minimum bias and centrality triggered events.
- For pp collisions:
    - Collect a reference rare triggers sample with an integrated luminosity of 40 pb^{-1}, which is equivalent to the 1 nb^{-1} sample in Pb-Pb collisions.
    - Enlarge the statistics of the unbiased data sample, including minimum bias collisions at top energy.
    - Collect a reference sample of 10^9 events at the reference energy of 5.02 TeV
- For p-Pb collisions:
    - Enlarge the existing data sample, in particular the unbiased events sample at 5.02 TeV.

To reach these objectives ALICE is exploiting the increase in instantaneous luminosity for Pb-Pb and benefits from the consolidation of the readout electronics, in particular of the TPC and TRD to increase the readout rate by a factor of 2. This has the effect of doubling the event rate and the data throughput of the entire dataflow including the migration of data to the Tier 0 computing centre, which now goes up to 10 GB/s.

Table 1: Data collected in 2015 and 2016

Year	System	Central barrel events	Average event size (MB)	Data volume (PB)
2015	pp	900 M	4.7	4.23
	Pb-Pb	210 M	12	2.52
2016	pp	1400 M	3.9	5.5
	p-Pb	880 M	1.7	1.5

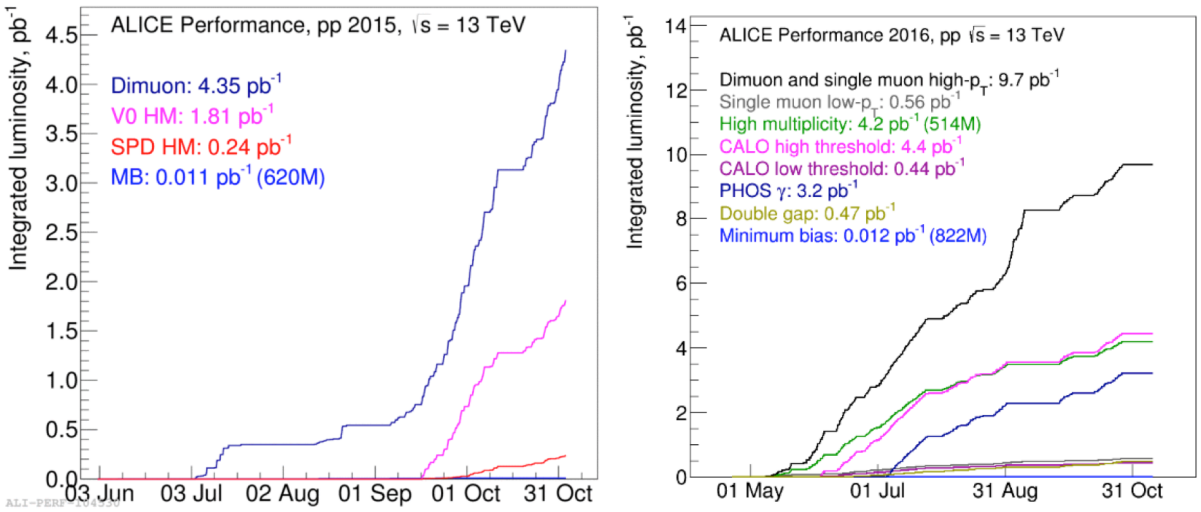

Figure 1: ALICE performance pp $\sqrt{s}$ = 13 TeV

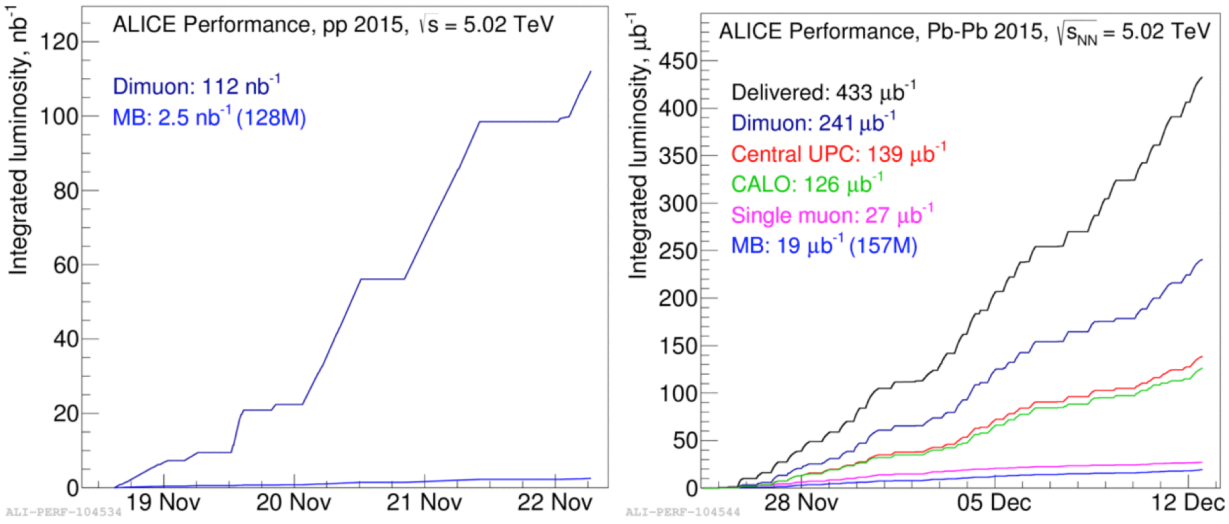

Figure 2: ALICE performance in pp and Pb-Pb at $\sqrt{s}$ = 5.02 TeV

ALICE Run 2 data taking summary

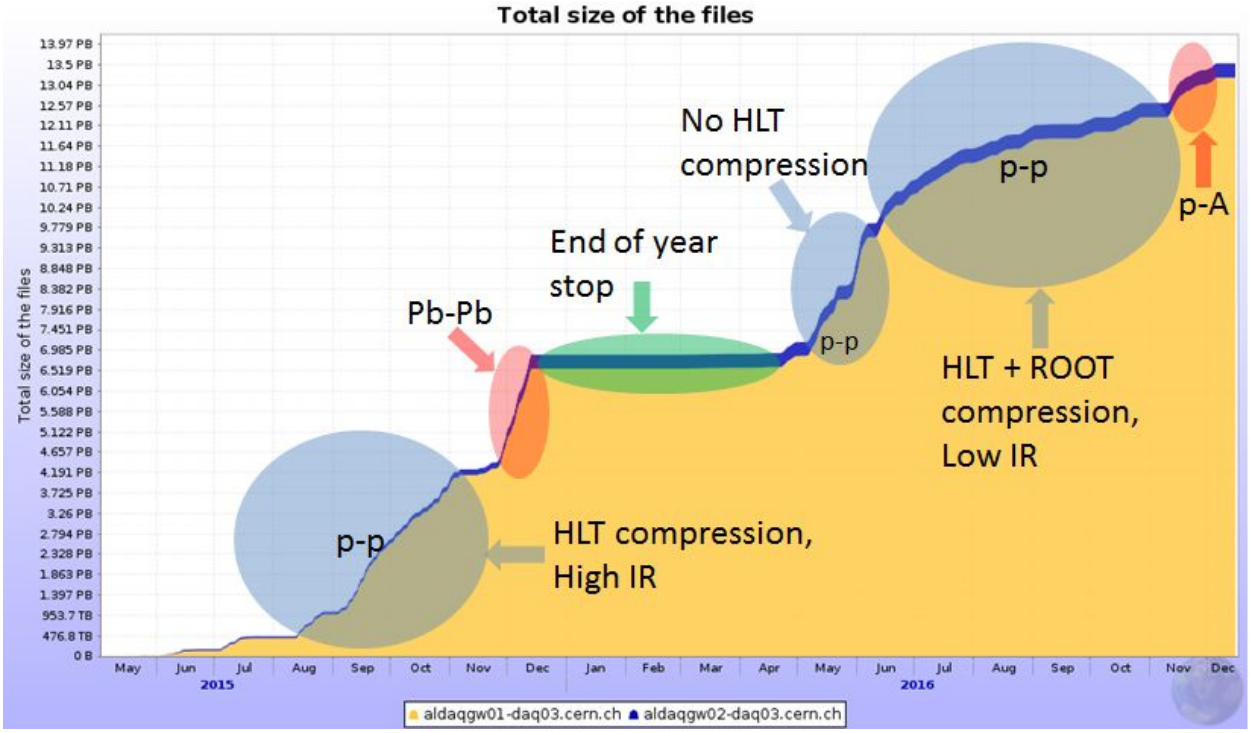

Figure 3: Raw data accumulated during Run 2.

After succesfull pp and Pb-Pb data taking in 2015, ALICE started the 2016 data taking in April 2016 using upgraded TPC readout and gas composition. The initial data taking was done without compression, in order to commission the new readout and to allow for tuning and validation of the HLT cluster finding algorithms. The HLT compression was switched on in June. Overall, ALICE has accumulated 7.5 PB of RAW data in 2016.

All RAW data has been passed through the calibration stages, including the newly developed track distortion calibration for the TPC and has been validated by the offline QA process. In addition, the di-muon and calorimeters triggers have been fully reconstructed and made ready for physics analysis.

The TPC track distortion correction algorithms have been finalized and fully validated with both pp and 2015 Pb-Pb data at various interaction rates. This has allowed us to start RAW data processing of the 2015 Pb-Pb period and also of the longest pp data taking periods, both from 2015 and from 2016. The data processing was completed in January 2017. The associated Monte-Carlo simulation, anchored to the conditions data and distortion corrections from the RAW data calibration cycles was also completed in the beginning of 2017.

Considering the higher LHC efficiency, ALICE can collect the event statistics as defined in the physics programme by running at a reduced interaction rate. This improves the running conditions and the data quality by reducing the distortion amplitude in the TPC and event pileup in the readout time of the ALICE detectors. This strategy was already applied to 2016 data taking during stable p-p period in weeks 24 – 35, 38 – 43 and resulted in a substantially lower average raw event size keeping cumulative tape usage at the end of 2016 was under the projected limit.

HLT compression

One example of technology improvement is the ongoing development of the HLT compression. The standard online TPC cluster compression results in a RAW data reduction by factor of ~4.3 to ~5.5 and is run throughout the data taking. To further reduce the RAW data volume during 2016 data taking, we developed and deployed a differential Huffman compression algorithm resulting in an additional 10-20% reduction of the TPC event size.

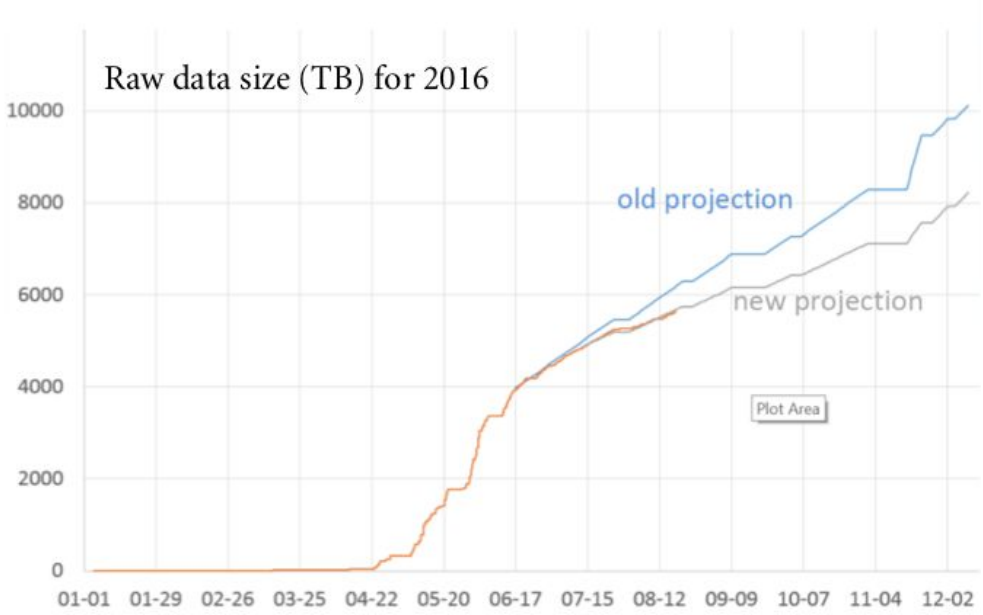

Figure 4: Impact of improved data compression in HLT on recorded data volume in 2016.

The HLT project continues to develop even more robust algorithms to filter out the noise. New algorithms are being investigated that might allow to reject 35% of fake clusters, i.e an additional 10% reduction of the RAW data size.

Projected requirements for 2017 and 2018

A 60% stable beam efficiency is assumed for pp and Pb-Pb together with an ALICE data taking efficiency of 95% leading to an effective running time of 7.4 and 7.5 Ms of pp in 2017 and 2018 respectively and 1.2 Ms of Pb-Pb in 2018.

In 2017 and 2018 the pp data taking mode will be set to limit the TPC readout rate to 400 Hz which will allow us to reach the statistics objective set for Run 2 in all trigger categories (minimum bias and several rare triggers) as well as at the reference energy of 5.02 TeV. The total amount of data recorded will be 17.5 PB.

During the Pb-Pb run in 2018 the TPC will be operated at the maximum RCU2 bandwidth of 48 GB/s. With an anticipated HLT compression of a factor of 6 and taking into account the data from other detectors, we anticipate a total readout rate of 10 GB/s and a total amount of data of 12 PB.

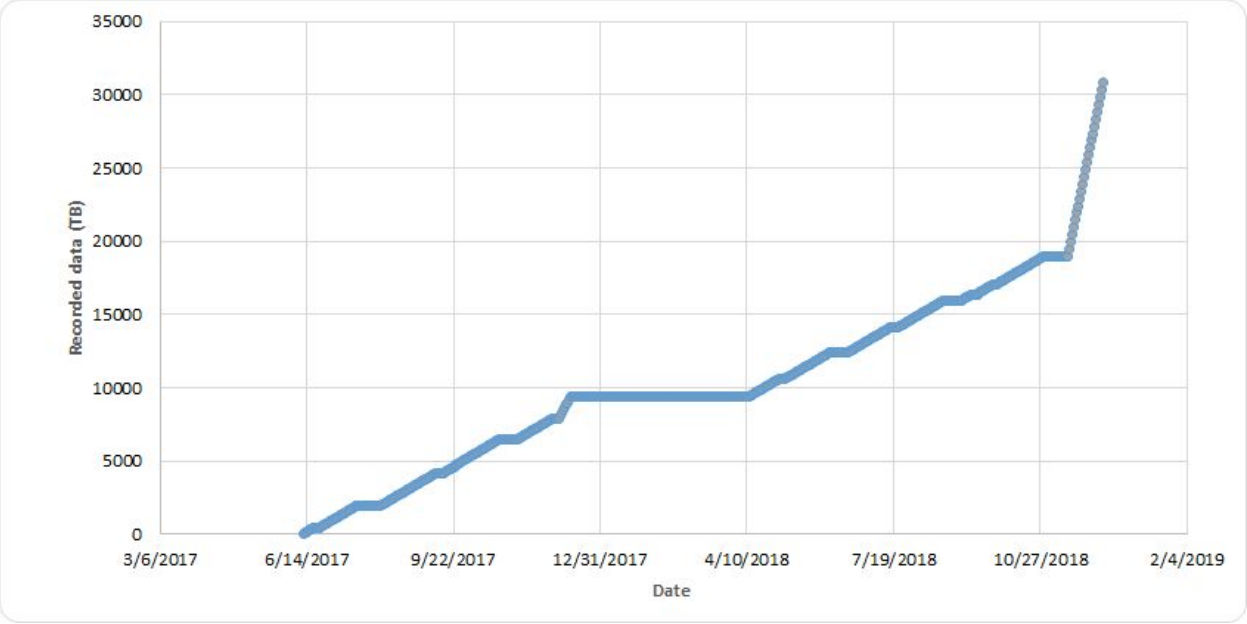

Figure 5: A graphical representation of 2017-2018 data taking scenario.

In addition to the runs mentioned above, we will have a 5 TeV pp run at the end of 2017, which is an important reference sample for our p-Pb and Pb-Pb data samples. ALICE will collect about 900M events (collecting mostly minimum bias triggers at high rate, 1.5-2.0 kHz, but a low interaction rate to reduce pile up and event size), giving a total data size of 1.3 PB.

Possible ways to reduce raw data volume

When it comes to reducing the amount of raw data collected, we are considering the following options:

- **Improved HLT compression**
  The TPC data are compressed by the HLT by running a cluster finder and storing only the cluster output (HLT mode C). This compression has been retuned in 2016, but a new effort to further compress the data by removing split clusters is under consideration. Early tests on existing data suggest that this may yield a 10-20% reduction of the data volume for pp. A full validation test will be performed at the start of the 2017 run. It is important that the data quality does not change from year to year, since the goal is to combine the full Run 2 pp sample for large statistics measurements of e.g. open charm production in pp collisions.
    - Impact on physics output: Minimal. If QA confirms that the new procedure works as expected we might still have a slight impact on measured quantities such as dEdx.

- **Change data taking scheme**
  In 2016, ALICE collected minimum bias pp events and rare triggers concurrently, running at an interaction rate of 100 kHz. It may be possible to reduce the total data rate by taking minimum bias events with a lower interaction rate, i.e. less pile-up in the TPC and a smaller event size. This however goes at the cost of some triggered data taking. In addition, since background event rates are approximately constant (5 kHz under good circumstances, depends on the machine conditions and filling scheme), the background event fraction increased when the interaction rate is reduced, leading to a reduced fraction of good events in the recorded sample. We worked out a scenario with minimum bias data taking at 50 kHz and triggered data at 150 kHz; the potential saving is approximately 10% of the total data size. We will monitor background rates in 2017 and may switch to this data taking scheme.
    - Impact on physics output: Minimal. If background conditions remain favorable.

- **Reduce replication**
  Currently we keep a distributed copy of our RAW data at the Tier-1 centers. If the tape resources are constrained, we would consider not to replicate the pp data immediately, but keep it only at CERN (Tier-0) and copy to Tier-1s during LS2.
    - Impact on physics output: Acceptable. This will inevitably delay pp data processing until we complete Pb-Pb reconstruction due to resource contention since data will be available only at CERN. There is increased risk of data loss associated with this scenario.

Optimization of Workflows

The average share of CPU utilisation between the three main ALICE workflows is as follows: simulation (70%), reconstruction (11%), analysis (19%). Possible optimizations of these workflows in terms of storage and CPU are discussed in the sections below.

Simulation

The ALICE simulation strategy is already highly optimised. We perform two types of Monte Carlo productions: *baseline productions* and *specialised productions* needed for specific physics analyses. The baseline productions correspond by their nature to minimum bias data. For pp and p-Pb collisions we produce MC samples corresponding approximately to the number of events in data, whereas for Pb-Pb they correspond to about 10%. For the specific productions we enhance the number of signals per event. This is achieved either by exploiting biasing techniques provided by the MC generators or by injecting (parameterised) signals on the event generator level. For rare signals selected from minimum bias events (e.g. charm hadrons) or for triggered events with a strong transverse momentum dependence (e.g. jets) one can obtain MC samples that correspond to multiple times the signal statistics in data using only a fraction of the resources needed for the baseline productions. All in all, it is unlikely that we can reduce the MC requirements for small systems (pp and p-Pb) significantly.

For Pb-Pb simulations, the baseline simulation corresponds to only 5-10% of the raw data volume. This gives a statistically uncertainty comparable to the systematic uncertainties over a limited p_T range for several common observables. For rare probes, we use the specialised productions. We currently re-generate background events together with the signal events. However, it is possible to reduce the CPU usage by simulating the underlying events (UE) only once. The different signal events are then merged with the UEs on the level of digits using each UE several times. The ALICE simulation framework already supports this strategy. The merging is performed using so called summable digits, which are digits before the addition of electronic noise and pedestal subtraction. Merging together with developments to reduce the digitization time, in particular for the TPC, have the potential to further optimise the MC requirements for Pb-Pb productions.

Fast parameterised simulation has been employed for performance studies of the ALICE detector upgrade: the estimation of the signal reconstruction performance and the background levels for rare signals. For MC productions corresponding to existing data samples the background is known from data and the primary scope for MC is the signal reconstruction efficiency determination. This is accurate enough for specific applications, in particular for upgrade performance, but the current parameterised simulation is not accurate enough to correct physics results. To obtain the required performance, a large scale development would be needed. In addition, the parameterised simulation would need to be calibrated against a sufficiently large (and differential) sample of full MC events. It is therefore unclear whether

significant gains can be achieved. It is also important to remark that in general our analyses are becoming more sophisticated and more sensitive to small effects, which also means that our simulations have to become more and more precise. This development rather suggests that we would increase our simulation efforts. In this context, ALICE has also been working on changing the particle transport code in simulation to Geant4 instead of the legacy Geant3. Geant4 is already used for specific simulation tasks: anti-nuclei, line-shape of quarkonia etc. This development is close to final, but incurs a performance penalty of a factor 1.6 for transport.

To summarise, we will investigate whether reusing the underlying event for Pb-Pb simulations is practical and whether it leads to significant reductions in CPU usage. However, to take full advantage of this requires some development, which means we cannot rely on this for 2017.

- Impact on physics output: The overall physics output has been largely optimised by dividing MC simulations into baseline productions, productions for specific physics analysis using injected signals or biasing and fast parameterised simulation for the ALICE detector upgrades. Manpower needed for improvement of the digitisation largely overlaps with reconstruction experts. A careful planning of the development tasks in both areas is needed for optimal gain in overall precision.

Reconstruction

Since start of Run 2 and after we observed a significant distortions in TPC that initially prevented us to reconstruct data with sufficient precision, we invested a lot of work in finding a solution to this problem. This required introduction of a new calibration step that produces time dependent correction that are used in subsequent reconstruction iteration. This problem is finally under control and we are back to our target precision in reconstruction. In the process, memory consumption was reduced while overall CPU needs were increased by 5% due to additional calibration step.

The software changes needed to reach the goal of using the TRD information not only for the PID but also for a more precise measurement of high p_T tracks were implemented but needs to be validated. While these changes are not expected to introduce performance penalty, they will not result in any gains except in better reconstruction quality for analysis.

- Impact on physics output: While the improvements in reconstruction and calibration had a small impact CPU resources they resulted in better data quality with additional margin for improvement that should not change our future CPU needs.

Analysis

The two types of analysis in ALICE are “Organised analysis” which is run in the so-called analysis trains, combining the analysis tasks of many users within a given physics working group and minimizes I/O, and “Individual user analysis” where the users run their own tasks. ALICE has a strict individual user quota, both in CPU and storage space. These quotas are

sufficient for users to develop and test their code, but not to run it over large data set, thus promoting the organised analysis.

In the past years, the individual analysis resource use has steadily declined and has levelled at ~4% of the total CPU. The remaining 15% is used by organised analysis and we see an upward increase of about 2% per year.

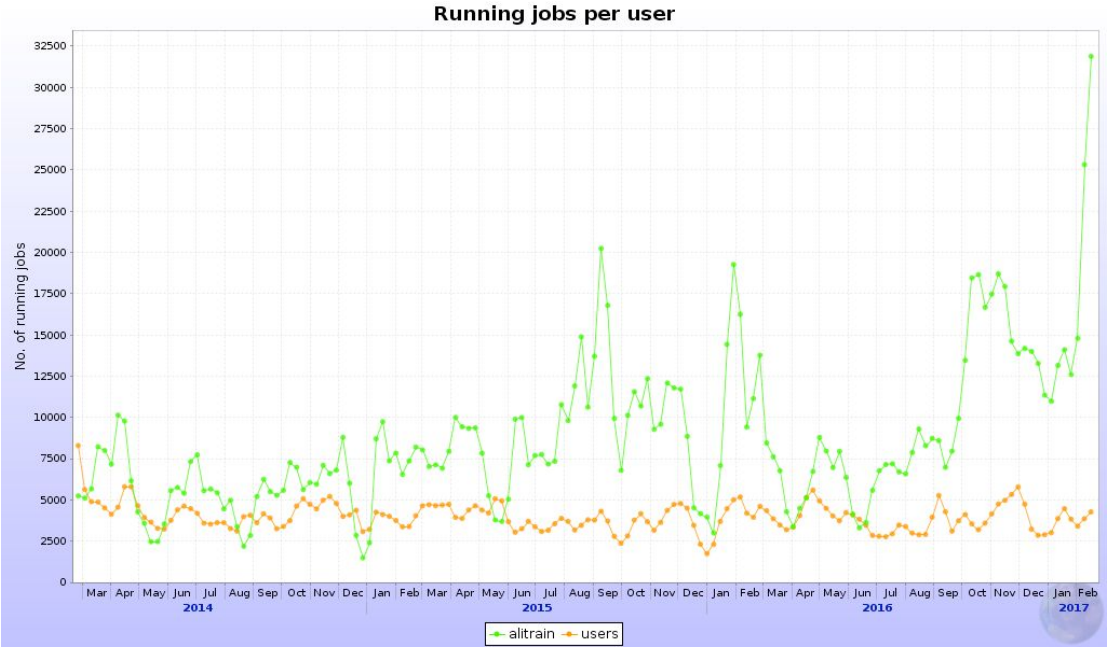

Figure 6: Number of analysis train and individual user jobs over past three years.

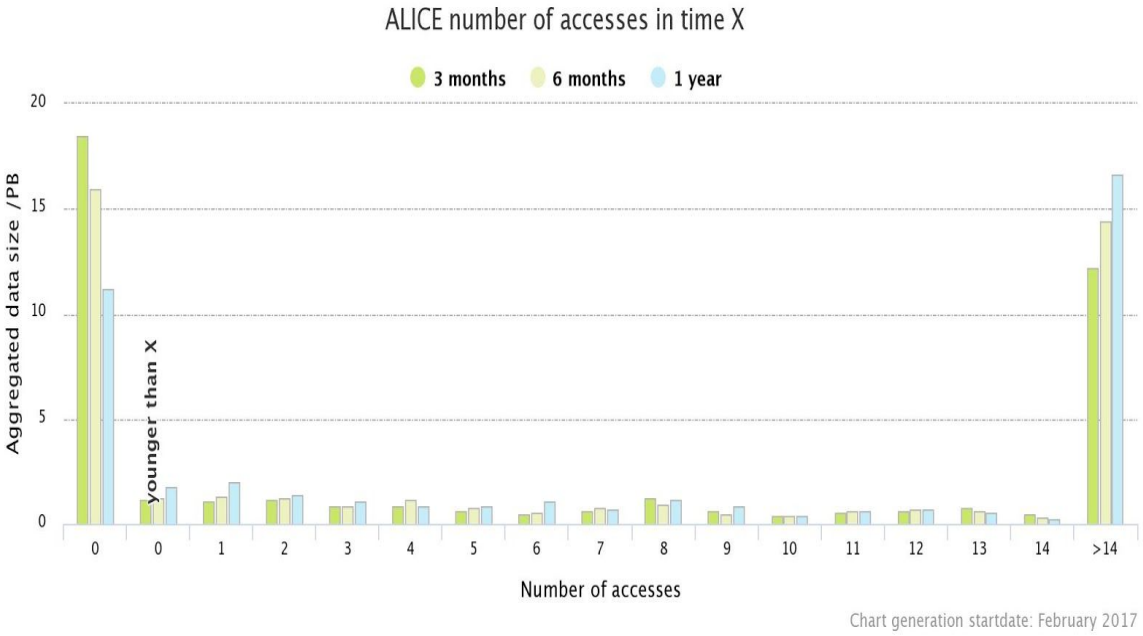

Figure 7: Current (as of February 2017) data popularity in ALICE

In addition to the optimization of number of replicas based on the dataset popularity, we continue to improve the AOD format with the goal to migrate all organised and individual user analysis to AOD. The ESDs will then be only used to produce new AODs, with improved cuts and content and thus ultimately the number of ESD copies can be reduced to 1. This is an ongoing process and the analysis optimization is a high priority task for ALICE. The economies of disk and CPU achieved by the methods outlined above are already taken into account in our requirements and we do not foresee any significant reduction of disk space needs in Run 2.

The standard data popularity plot for ALICE is shown in Fig.7. The high popularity of the current data is attributed to the preponderance of organized analysis, which tends to access complete datasets.

- Impact on physics output: Positive but not always doable. Moving away from ESDs speeds up the analysis in general. Certain types of analysis can be performed only on information stored in ESDs. Solving this problem is an ongoing task, as the emerging analyses always start from ESD and later could be converted to AODs. This always requires re-creation of a new version of AODs with increased size for new and the already existing datasets.

Number of copies, formats, versatility

The analysis relies on ESD and AOD input data. The ESD and AOD size relation to the RAW data are given in Table 3. Following the ALICE computing model, each RAW data file contains a completely unique information (there are no multiple data streams with same events) and each file is stored once at T0 and once more on one of our T1s. The ESD and AOD format resulting from RAW data reconstruction and MC simulation is the same. We keep 2 copies of the AODs from any active RAW reconstruction and MC simulation cycles and 1 or 2 copies of ESDs, depending on the projected popularity of the dataset. For example, the ESDs from Pb-Pb usually have two copies.

Table 2: Summary of data types used by ALICE

RAW	Raw data
ESD	Event Summary Data. All information after reconstruction, sufficient for any physics analysis, AOD re-stripping and re-calibration for subsequent processing cycles. ESDs are kept with a different replication factor depending on demand for the data and reconstruction cycle number.
AOD	Analysis Object Data. All events and all information needed to perform the majority of physics analysis. AODs are kept with a different replication factor, depending on demand and version number.

Table 3: Data replication across tiers

Data occupying T0/T1/T2 Storage					
	Event Size [MB]	# of copies on disk		# of versions	# of copies on tape
		minimal	typical		
RAW	3 (pp) 11 (Pb-Pb)			1	2 (one at T0 + one at one of the T1s)
ESD	10 to 30% of RAW, depending on type of collision system and luminosity	1	2	1-3	
AOD	10 to 15% of RAW, depending on type of collision system and luminosity	1	2.6	1-4 per ESD version	
MC ESD	0.37 (pp) 2.7 (Pb-Pb)	1	2	1	
MC AOD	30% of MC ESD	1	2.6	2	

The number of copies is reduced to 1 for both ESD and AOD if a given dataset RAW data reconstruction cycle is superseded by the next cycle, similarly for MC. In addition, we continuously use dataset popularity to remove replicas of unaccessed data and for very old MC or reconstruction sets, remove the ESDs/AODs altogether from disk storage.

We believe that we are already running on the lowest possible limit in terms of dataset redundancy and further reduction of copies will come with a substantial operational risks.

New analysis formats (nanoAODs) are being introduced for specific types of analysis. These have very small disk footprint and sometimes are compact enough to fit on individual user desktops.

The effect of nanoAOD use on global resources use still needs to be evaluated. The purpose of the nanoAODs is to speed up the analysis for which they are produced, but they cannot replace the ESDs and AODs both in content or versatility.

- Impact on physics output: Reduction below 2 ESD/AOD replicas for popular datasets could considerably slow the physics analysis by placing very high load on storage elements, thus reducing the efficiency of the sites, and/or induce statistical loss from

storage elements under maintenance, ultimately leading to non-reproducible results and repetition of analysis.. The number of replicas has been carefully studied over the years and tuned to correspond to the analysis practices of the physics community. We believe that any progress in this area will be gradual and there is no room left for immediate and considerable savings.

Data/CPU/tape management

ALICE keeps 2 copies of the collected RAW data - one full copy at T0 and one distributed copy at the T1s. Only verified good quality physics RAW data is replicated to T1 and in general, tape usage is dominated by RAW data. To save disk space, we put unpopular ESD sets on tape, but this amounts to a few percent of the total tape use only.

Taking into account the technology improvements in data compression, which are already applied on our RAW data stream, tape requirement reduction for ALICE can only be achieved by reduction of replication factor. For example not replicating the p-p data in 2018 would reduce our requirements for tape at the T1s. The non-replicated data will effectively be 'parked' at the T0 for later processing. The economy is for tape space only, as the production would only be delayed and the resources taken for RAW data reconstruction at the T0 will be offset by increase of the remaining activities at the T1s. In this scenario, the effective reduction of tape requirements in 2018 amounts to 9 PB, all at the T1s.

Additional small reduction of tape use at the T1s can be achieved by not replicating the RAW data during the HLT commissioning in 2017. The exact amount cannot be specified.

- Impact on physics output: Non replication of RAW data carries a small risk of partial of total data loss.

Parking/delayed processing

Parking data and delayed processing could be considered as an option for 2018 pp data reconstruction, see the above section for details. The drawback is keeping resources and manpower busy with Run 2 processing, while more human resources are needed for the Run 3 upgrade work.

- Impact on physics output: Parking pp data recorded in 2018 and delaying processing until after Pb-Pb dataset processing is completed might limit by 30% or more the pp statistics available for analysis before Quark Matter conference in autumn 2019. This could have an impact on ALICE competitiveness with the other LHC experiments.

Conclusions

ALICE continuously optimizes the use of computing resources by introducing improvements at all steps of data taking, processing and analysis workflows. In 2016, after switching to a new TPC gas mixture and readout (RCU2), the clusterization algorithms in HLT were optimized to achieve the maximum compression rate to date, resulting in smaller RAW event size than anticipated. This lead to less tape being used for RAW data storage than foreseen already in 2016.

The increased data taking and LHC efficiency allowed us to tune the trigger and interaction rate (IR) settings to achieve the ALICE physics goals and at the same time reduce the event pile-up and interaction rate related distortion in the TPC. Reducing the pile up results in direct reduction of RAW data size and also a small reduction of the reconstruction output. The trigger and interaction rate optimization tactic will be used in the remaining years of Run 2 and is taken into account in our new resource requirements. A small additional gain (~10%) can be obtained by taking minimum bias pp data in dedicated runs with low interaction rates. This is only worth considering if the background levels stay as low as they were during the best periods in 2016.

Following the larger than expected distortions in the TPC observed in 2015 and 2016, ALICE developed a comprehensive set of Offline iterative calibration. The software was tested and validated over wide range of interaction rate and TPC running conditions and has allowed us to process fully the 2015 Pb-Pb data, 2015 and 2016 long p-p periods and 2016 p-Pb data. At the same time, the reconstruction software was optimized for memory and CPU use and the improvements are sufficient to offset the additional CPU used by the new calibration.

Concerning disk usage, we are continually monitoring our usage and employing several methods to optimize the disk space. The main method is the reduction of ESD replicas to a single copy for data sets that are not actively being analysed and some that are analysed only occasionally. Same rules and methods are used to reduce the AOD replicas, although the size of the AODs are small compared to ESDs. We are also imposing strict user disk quotas, which keeps the user-occupied disk storage below 10% of the total. We also do a continuous scanning of all storage elements for dark data, thus assuring that it is less than 1% of the total space.

After careful consideration of the 2017-2018 running scenario we have come to the conclusion that the only direct way to save computing resources is the reduction of RAW data volume by continuing to develop HLT compression algorithms and reducing the file replication factor, thus reducing our tape requirements at the T1s. If adopted, this method could result in 9PB less tape required at the T1s in 2018. The drawbacks, apart from the obvious reduction in data safety, are the delayed processing of the non-replicated data, which will be pushed well into the LS2. This

will result in higher load on resources and personnel at the time when Run 3 preparations will be at their peak.