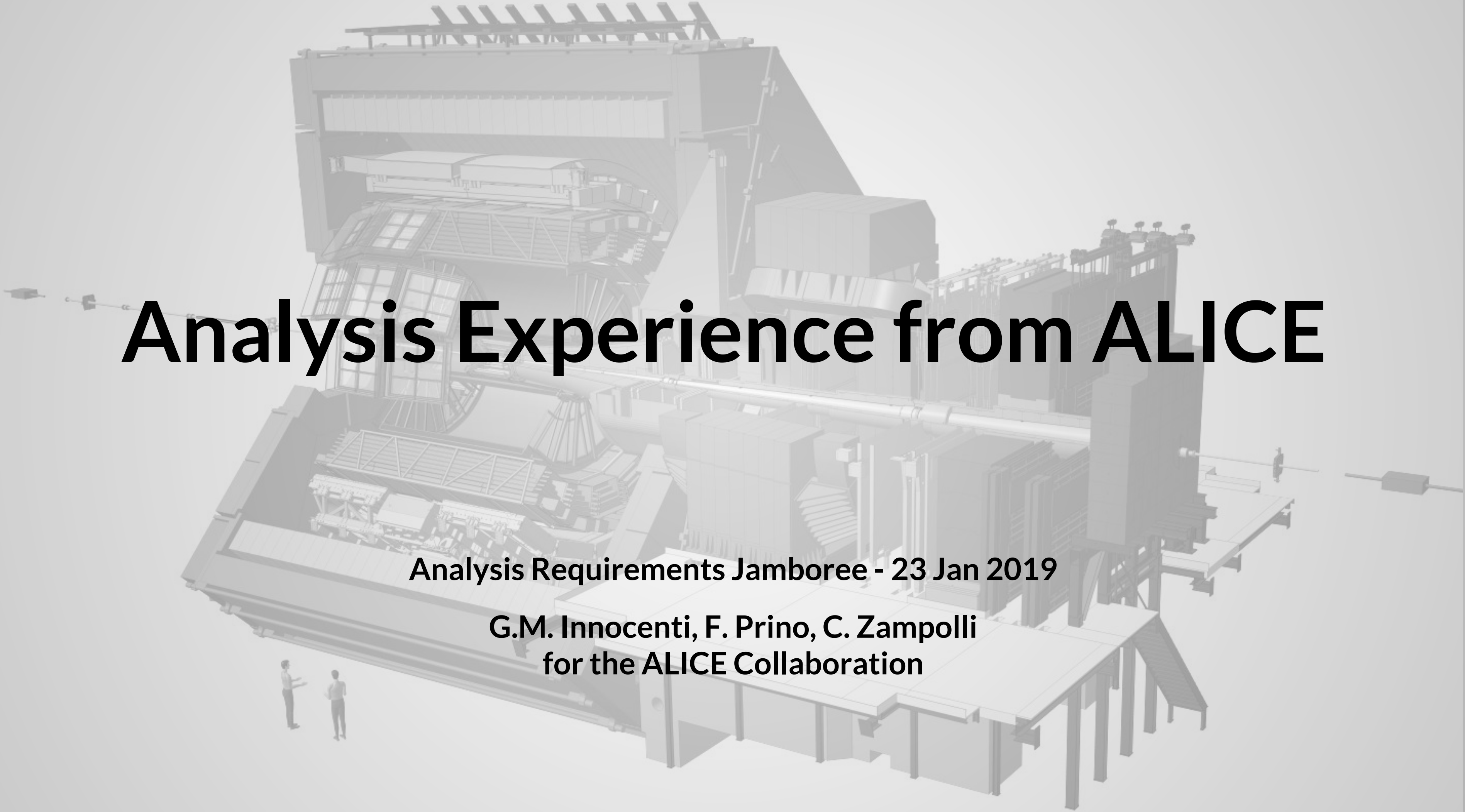


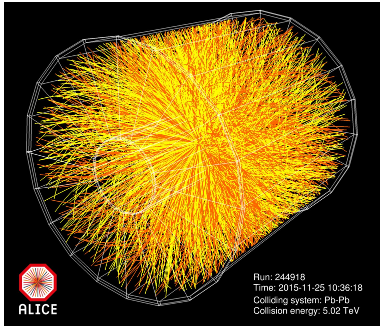
Analysis Experience from ALICE

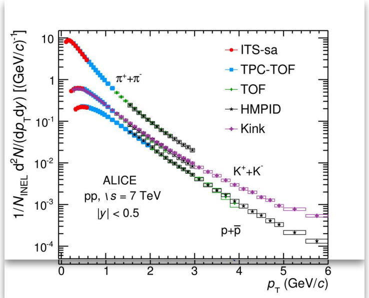
Analysis Requirements Jamboree - 23 Jan 2019

G.M. Innocenti, F. Prino, C. Zampolli
for the ALICE Collaboration

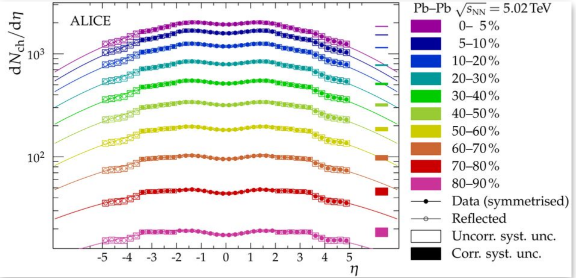
A Large Ion Collider Experiment - AKA “ALICE”

- ★ **Dedicated Heavy-Ion experiment**
- ★ Tracking detectors with geometrical acceptance |η| < 0.9 and full φ
- ★ Precision tracking capabilities in |η| < 0.9 down to very low momenta (100 MeV/*c*, low B field 0.5T)
- ★ Different particle identification detectors, some with limited geometrical acceptance
- ★ Excellent hadron identification from low to high momenta
- ★ Tracking detectors optimized for extremely high charged track multiplicities → low event rate capability
    - ○ Up to 159 track points in the TPC

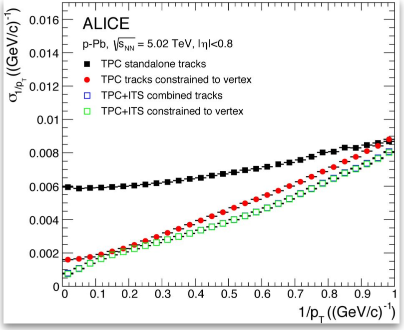




EPJC 75 (2015) 226



Phys.Lett. B 772 (2017) 567-577



Int. J. Mod. Phys. A 29 (2014) 1430044

In this talk...

Inclusive measurements

Soft probes

Hard probes

Differential measurements

Event characterization

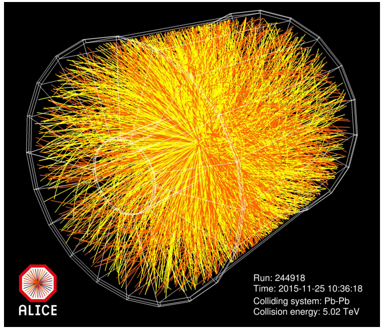
pp

PbPb

XeXe

pPb

$\sqrt{s_{NN}}$



Efficiency and acceptance corrections ...

Event mixing ...

Detector calibration

Background subtraction

Simulation

Machine Learning

...

Unfolding ...

Templates

Systematic uncertainty evaluation

Distributed computing ...

Cut variation

Reconstruction

Organized analysis ...

Embedding

Fitting

Multi-dimensional structures

Correlations

Vertexing ...

In this talk...

PbPb

pp

XeXe

pPb

$\sqrt{s_{NN}}$

Efficiency and acceptance corrections

...

Event mixing

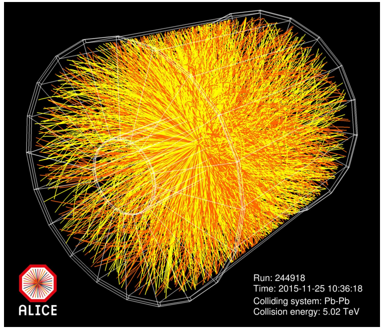
...

Inclusive measurements

Detector calibration

Background subtraction

Soft probes



Simulation

Machine Learning

Templates

...

Hard probes

Unfolding

...

Systematic uncertainty evaluation

Differential measurements

Distributed computing

...

Cut variation

Reconstruction

Organized analysis

...

Event characterization

Embedding

Fitting

Multi-dimensional structures

Correlations

Vertexing

...

What does (will) ALICE analyze?

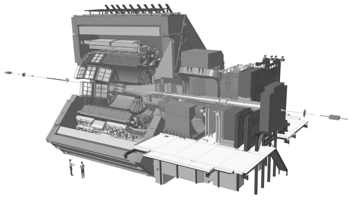
Run2 PbPb (2015 + 2018):

- ~300M MinBias events
- ~135M central (0-10%) events (~3-4 MinBias)
- ~120M semi-central (30-50%) events (~2 MB)
- ~3 PB ESDs, ~1 PB AODs
- 2015 sample (MinBias only): ~900 tracks/ev
- 2018 sample (MinBias + central + semi-central): ~2000 tracks/ev on average

Run3+4 PbPb:

- ~10^{11} MinBias events → factor 100x more statistics than Run2
- ~30 PB AODs
- ~900 tracks/ev

Workflow in ALICE



Data taking + calibration

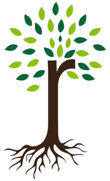
Reconstruction

Event **S**ummary **D**ata (ESD)

Tree containing:

➔ Reconstruction information of tracks (parameterizations, covariance matrices…)
➔ Vertices
➔ V0s & cascades
➔ PID
➔ Calorimeter information
➔ Forward detectors (e.g. Zero Degree Calorimeter)

AOD “filtering”

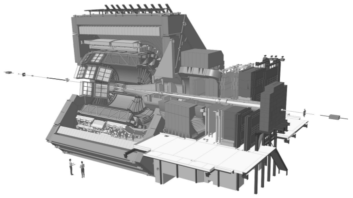


Analysis **O**bject **D**ata (AOD)

Tree containing

➔ Lighter version of ESDs (~1.5x of the size)
➔ Accompanied by “delta” AODs which contain extra information on reconstructed decays (e.g. charmed particles)

Workflow in ALICE



Data taking + calibration

Reconstruction

Heavy process (~160s/ev in MinBias PbPb events)

Event Summary Data (ESD)

AOD “filtering”

Light process (O(100) or more faster than reco

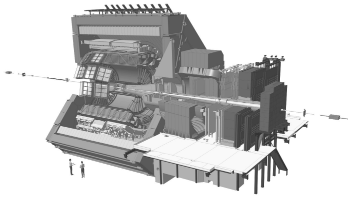
Analysis Object Data (AOD)

Stored on the grid with multiple replicas

Both ESDs and AODs are possible input for analysis even if

- ➢ Running on ESDs is strongly discouraged
- ➢ Despite the virtual inheritance from the same classes, the same analysis code cannot run simultaneously on both - some changes are needed

Workflow in ALICE



Reconstruction →

AOD "filtering" →

Data taking + calibration

Event **S**ummary **D**ata (ESD)

Analysis **O**bject **D**ata (AOD)

ALICE analysis run in tasks (deriving from ROOT TTask) that can be combined together in an analysis train → ALICE Analysis Framework

- ➢ Each event is read once, and each analysis task is executed sequentially on it
- ➢ "Service tasks" like event selection (to select on trigger, remove background…), centrality, PID handler are run in front of the analyses tasks, in a preparatory fashion
- ➢ Output is stored in a root file, the analyzers are responsible to define it
- ➢ Due to the statistics, local analysis is impossible → distributed computing
  - ○ Analysis run on grid nodes over groups of ESD/AOD files; merging over intermediate output done by the framework itself
- ➢ Only packages (ROOT+AliRoot+AliPhysics) that are centrally built and distributed can be used
  → Analysis code committed and versioned in ALICE git repositories

Distributed computing

Analysis Facility based on limited number of CPUs from a (local) cluster

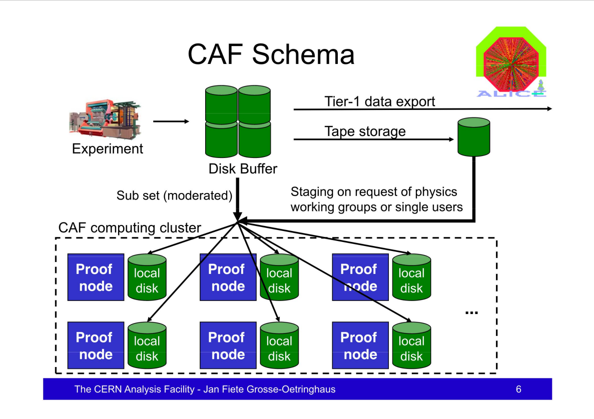
vs

WLCG Grid, worldwide distributed computing

➔ Pros: fast feedback; no waiting time overhead; results can be done and redone in a few minutes

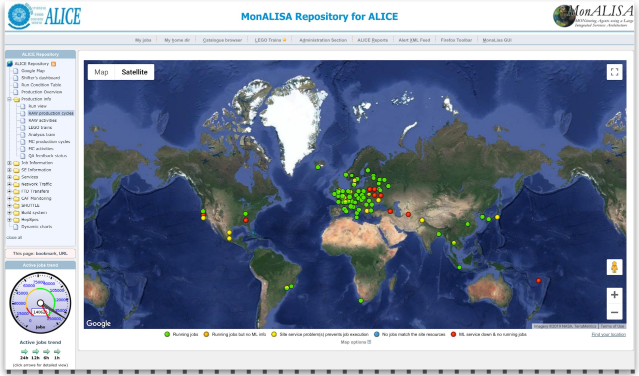
➔ Cons: cannot work on large datasets; human intervention on master and/or datasets (that got corrupted) often needed; high competition between users; some obscure aspects of the processing

J.F. Grosse-Oetringhaus, ROOT workshop, 27.03.07



➔ Pros: huge amount of resources to analyze the whole available statistics; large data and Monte Carlo productions possible; multiple users (and working groups) with almost no apparent competition; huge level of redundancy

➔ Cons: overhead of waiting time; several stages in the processing; some obscure aspects of the processing



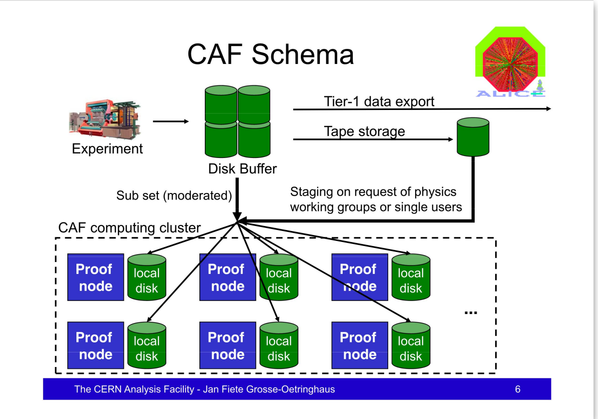
Active sites, real time, on 09.01.19

Distributed computing

Analysis Facility based on limited number of CPUs from a (local) cluster

vs

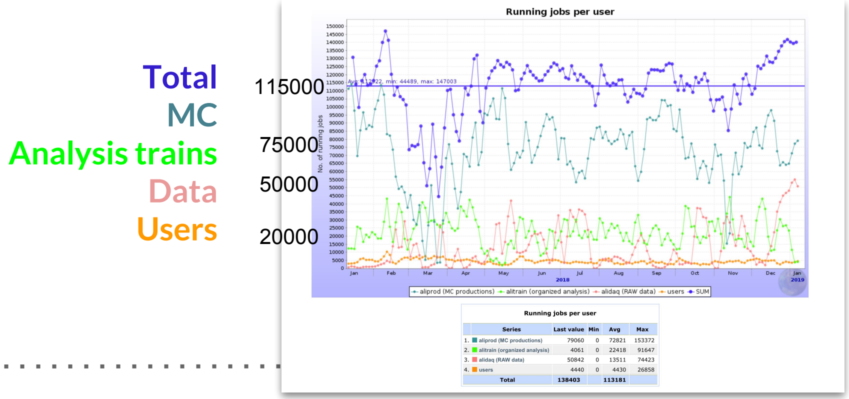
- ➔ Pros: fast feedback; no waiting time overhead; results can be done and redone in a few minutes
- ➔ Cons: cannot work on large datasets; human intervention on master and/or datasets (that got corrupted) often needed; high competition between users; some obscure aspects of the processing



J.F. Grosse-Oetringhaus, ROOT workshop, 27.03.07

WLCG Grid, worldwide distributed computing

- ➔ Pros: huge amount of resources to analyze the whole available statistics; large data and Monte Carlo productions possible; multiple users (and working groups) with almost no apparent competition; huge level of redundancy
- ➔ Cons: overhead of waiting time; several stages in the processing; some obscure aspects of the processing



Running jobs per user

	Series	Last value	Min	Avg	Max
1.	aliprod (MC productions)	79060	0	72821	153372
2.	alitrain (organized analysis)	4061	0	22418	91647
3.	alidaq (RAW data)	50842	0	13511	74423
4.	users	4440	0	4430	26858
	Total	138403		113181	

Number of running jobs in the last six months

Distributed computing

Analysis Facility based on limited number of CPUs from a (local) cluster

vs

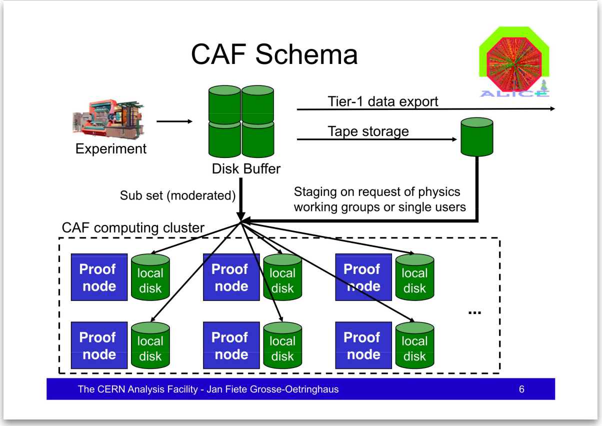
WLCG Grid, worldwide distributed computing

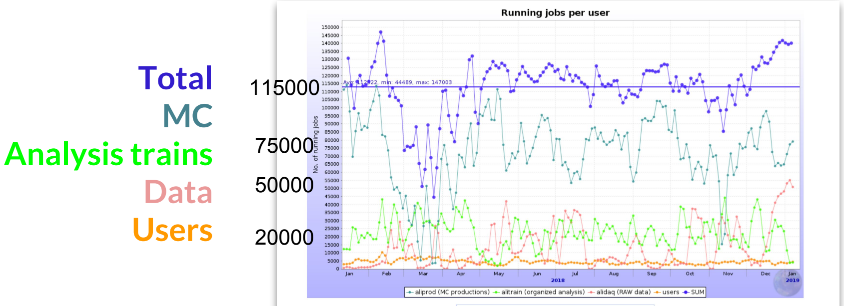
First ALICE paper was done using CAF; in the last years it was discontinued due to lack of power compared to the use cases, but it was always a good and useful tool for analyses requiring little statistics, for tests...

CAVEAT: Interface needed to run the analysis both on a CAF (or analogous) and on the grid (was the very useful case in ALICE), to limit the tools that analyzers need to use

processing; some obscure aspects of the processing

J.F. Grosse-Oetringhaus, ROOT workshop, 27.03.07





Running jobs per user

	Series	Last value	Min	Avg	Max
1.	aliprod (MC productions)	79060	0	72821	153372
2.	alitrain (organized analysis)	4061	0	22418	91647
3.	alidaq (RAW data)	50842	0	13511	74423
4.	users	4440	0	4430	26858
	Total	138403		113181	

Number of running jobs in the last six months

Distributed computing

Analysis Facility based on limited number of CPUs from a (local) cluster

vs

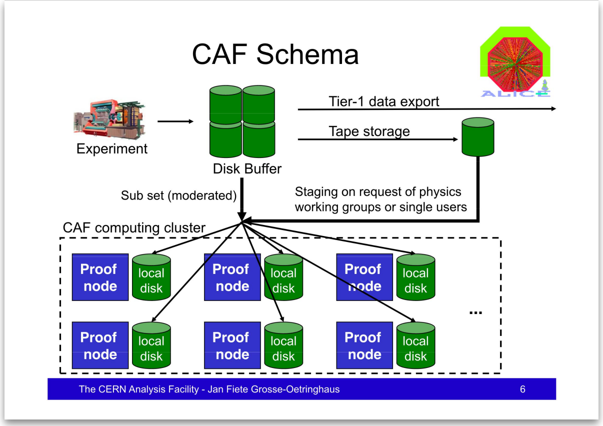
WLCG Grid, worldwide distributed computing

First ALICE paper was done using CAF; in the last years it was discontinued due to lack of power compared to the use cases, but it was always a good and useful tool for analyses requiring little statistics, for tests…

CAVEAT: Interface needed to run the analysis both on a CAF (or analogous) and on the grid (was the very useful case in ALICE), to limit the tools that analyzers need to use

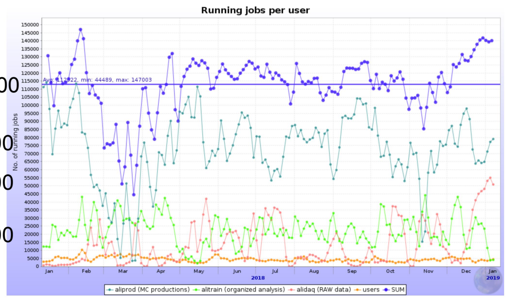
processing; some obscure aspects of the processing

J.F. Grosse-Oetringhaus, ROOT workshop, 27.03.07



Total
MC
Analysis trains
Data
Users

115000
75000
50000
20000



Running jobs per user

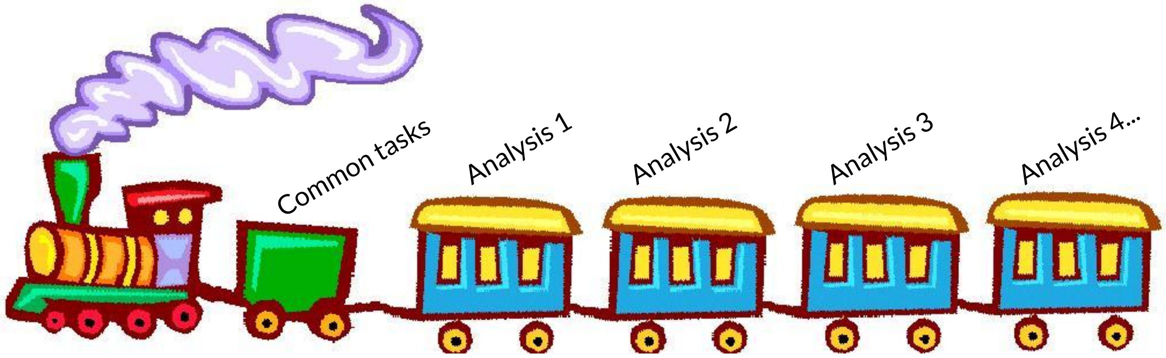
	Series	Last value	Min	Avg	Max
1.	aliprod (MC productions)	79060	0	72821	153372
2.	alitrain (organized analysis)	4061	0	22418	91647
3.	alidaq (RAW data)	50842	0	13511	74423
4.	users	4440	0	4430	26858
	Total	**138403**		**113181**	

Number of running jobs in the last six months

Organized distributed analysis

Central “analysis train” system in ALICE was developed to optimize the performance of the usage of computing resources in analysis

- ➢ Bottleneck is the I/O, reading the data several times (for each analysis)
- ➢ Physics Working Group group their analyses as much as possible to run together when needed over the same data



Adapted from link

- ➢ Each event is read once and used by all the attached analysis tasks
- ➢ Priority given to organized analysis over single user
- ➢ Limitations in quotas per user (CPU and number of jobs) not there for trains - but memory limits still exist
- ➢ Each user can add a wagon to a train
- ➢ Web interface with several features available: testing, performance feedback...

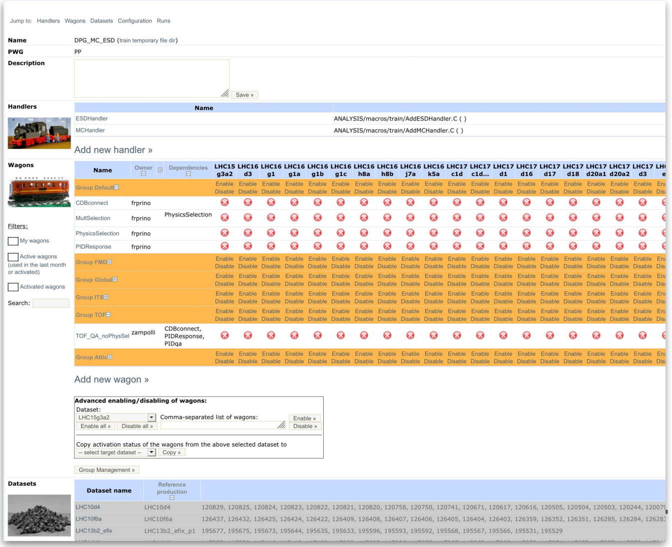
Organized distributed analysis

Central “analysis train” system in ALICE was developed to optimize the performance of the usage of computing resources in analysis

- ➢ Bottleneck is the I/O, reading the data several times (for each analysis)
- ➢ Physics Working Group group their analyses as much as possible to run together when needed over the same data

Database behind the system to keep track of the train configuration → the train number allows to retrieve all the information about some results

- ➢ Train number, analysis data sets documents in ALICE restricted detailed analysis note that are saved in a database



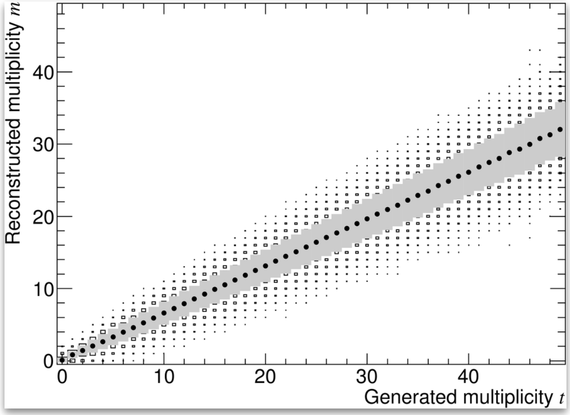
- ➢ Each event is read once and used by all the attached analysis tasks
- ➢ Priority given to organized analysis over single user
- ➢ Limitations in quotas per user (CPU and number of jobs) not there for trains - but memory limits still exist
- ➢ Each user can add a wagon to a train
- ➢ Web interface with several features available: testing, performance feedback...

Unfolding

Technique used to take into account detector/reconstruction resolution effects e.g.:

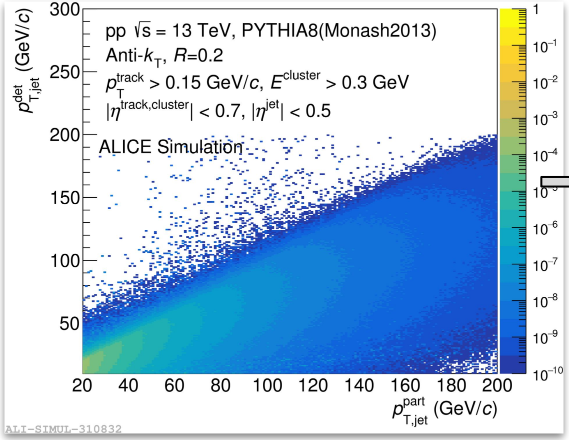
- ✓ To measure multiplicity, p_T distributions, electron spectra
- ✓ To measure jet spectra (correction for detector effects but also missing energy)
  - ▻ Response matrix built from simulation: measured/reconstructed vs true observable

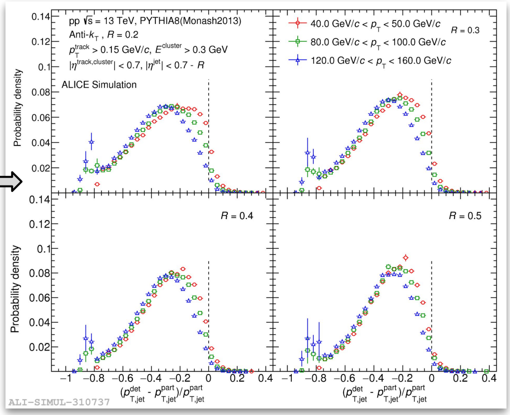
Multiplicity distribution



Eur. Phys. J. C 68 (2010) 89-108

Jet analysis





M.Fasel, HP 2018, hep-ex/1901.04304

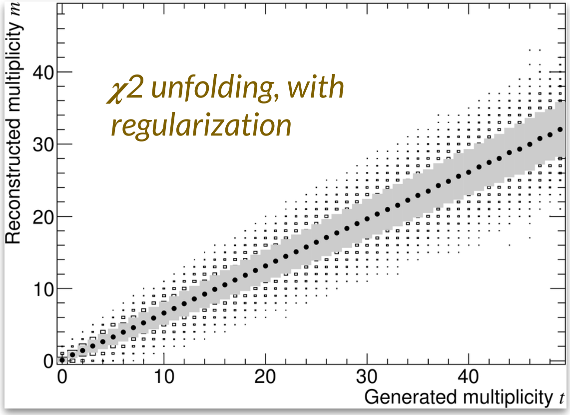
Unfolding

Technique used to take into account detector/reconstruction resolution effects e.g.:

- ✓ To measure multiplicity, p_T distributions, electron spectra
- ✓ To measure jet spectra (correction for detector effects but also missing energy)
  - ▹ Response matrix built from simulation: measured/reconstructed vs true observable

Multiplicity distribution

χ2 unfolding, with regularization

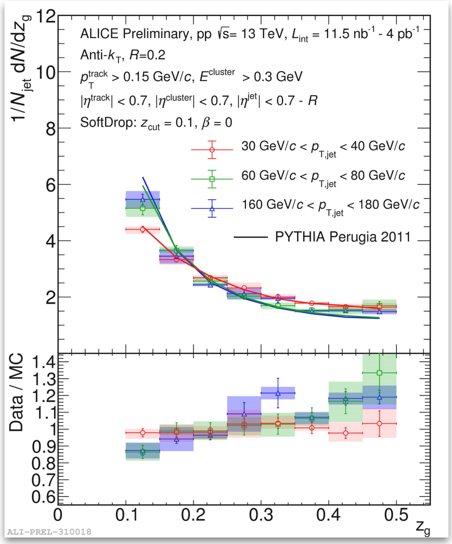


Eur. Phys. J. C 68 (2010) 89-108

Bayesian 2D unfolding for jet substructure measurements using RooUnfold

- *Package dedicated to unfolding (but including more than that - e.g. background subtraction), providing one single interface to different unfolding methods (Bayes, SVD - important for systematic studies) and implementations, some coming from ROOT*
- *Package maintenance? Why not part of ROOT? (several advantages - e.g. less dependencies)*
- *2D unfolding present only for Bayes at present to our knowledge*

Jet analysis



M.Fasel, HP 2018

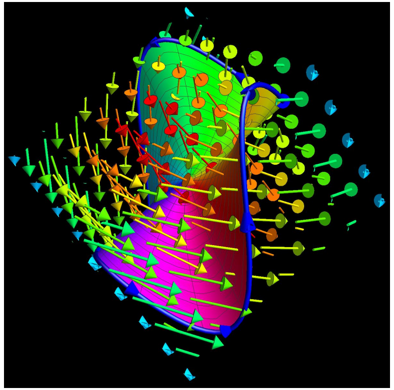
Multidimensional structures

To study corrections/correlations/observables that are **function of multiple variables**. E.g.:

- ➢ Cut variation for an “interactive” analysis - might be the output of an analysis train
- ➢ ALICE correction framework (based on ROOT THnSparse) allows to calculate the acceptance and efficiency correction as a function of multiple variables and at different stages of the analysis
- ➢ ALICE QA trending (using ROOT TTree) allows to correlate QA between different observables and detectors

Issues always related to **memory**

- ➢ Instantiation of many large objects may hit the limit of available resources, especially when running on “stricter” environment like the grid
- ➢ Merging of multi-dimensional (THnSparse) or non-scalable objects (trees) may be prohibitive
  - ○ Often needed not only in analysis (take the tree-based TMVA as one example), but also for calibration
  - ○ ThN is scalable but limited wrt THnSparse in terms of binning

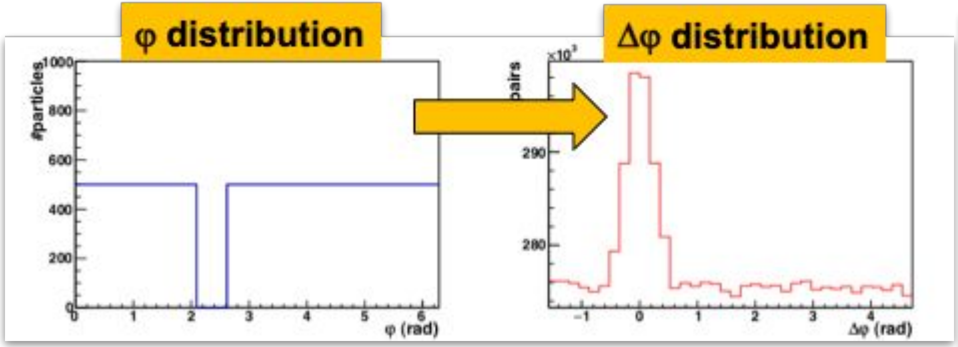


From link
I don’t know what it is, but it looks multi-dimensional

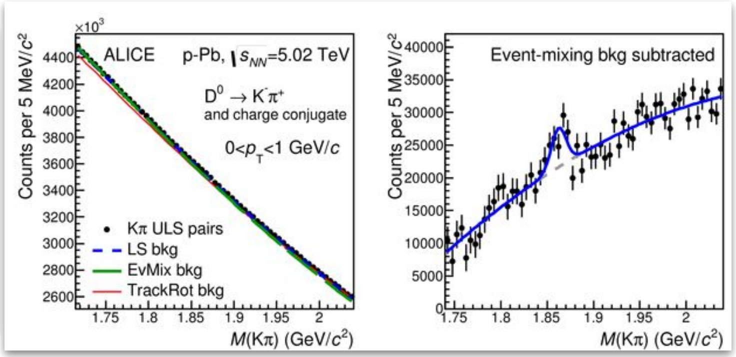
Event mixing

Technique commonly used in analysis to

- ➢ Remove detector acceptance effects in correlation analyses
- ➢ Evaluate background in signal extraction



E. Leogrande



Phys. Rev. C 94 (2016) 054908

Event mixing

Technique commonly used in analysis to

- ➢ Remove limited/incomplete detector **acceptance effects** (background) in correlation analyses
- ➢ Evaluate **background** in signal extraction

It consists in estimating the background building the correlation/signal candidate from tracks **from different events**

- ➢ Guaranteeing **compatibility between mixed events** (multiplicity, *z* of the vertex...)

Current framework in ALICE does not keep two files in memory, but **bufferizes** the tracks from previous events, categorized by the observables that allow them to be mixed (multiplicity, *z* of the vertex...)

Number of events to mix (in correlation analysis) determined by the statistical uncertainty (which should be smaller than the one for the signal)

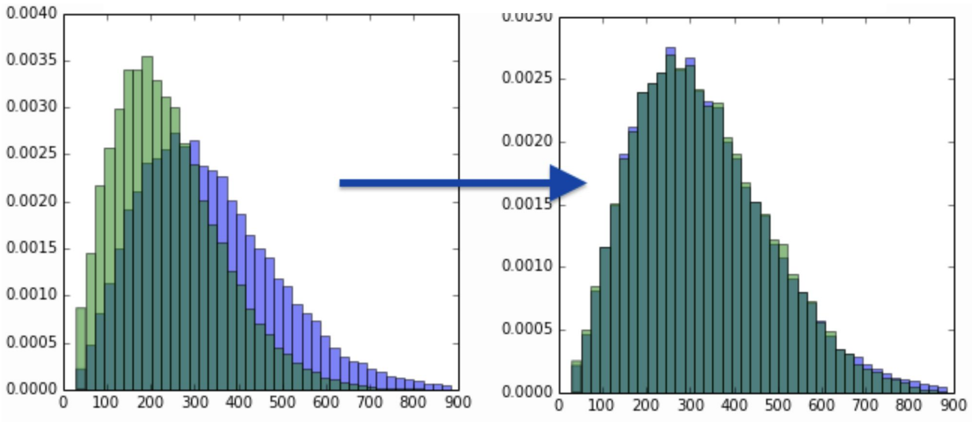
Closeness in time of mixed events is preferable, but so far not an issue

Events at the beginning of the processing are not mixed with “enough” events, those at the end of the processing are not used for mixing

Monte Carlo-Data reweighting techniques

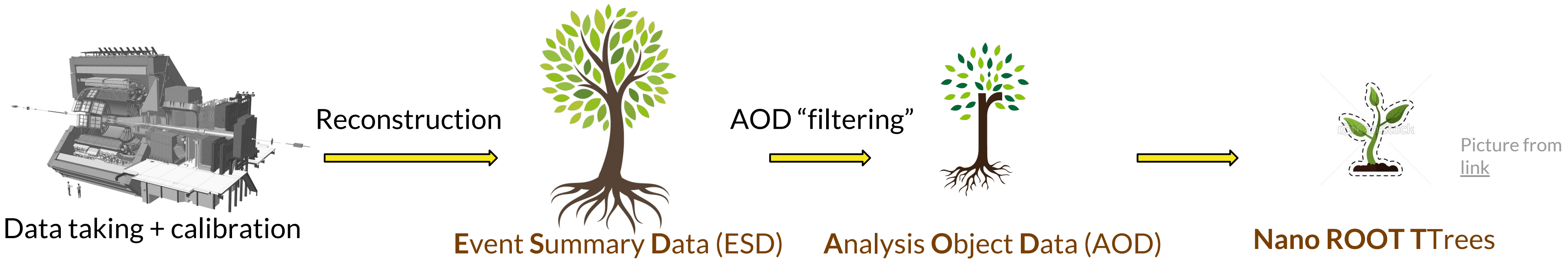
High luminosity 2018 and Run3 PbPb data will require more accurate Monte Carlo simulations with increased precision in describing detector effects, signal and background components:

✓ Crucial for proper efficiency estimations and for reliable cut optimisation procedures, which are very relevant for the measurements of rare signal (e.g. beauty decays) in presence of a very large combinatorial background



Strong push for developing new methods based on <u>multi-dimensional fits</u> and <u>machine learning</u> to improve the accuracies of our simulations

TTree production and skimming



New "layer"of data processing to produce ROOT trees with analysis-specific information:

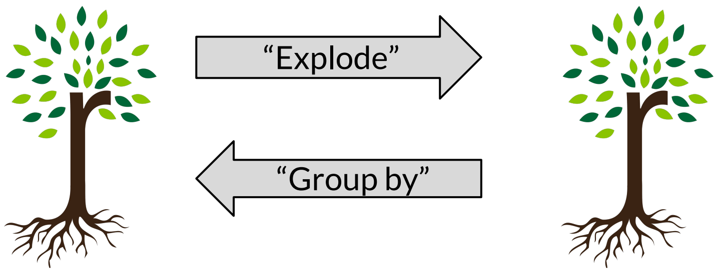
- Tight preselections to reduce data-size but loose enough to allow fine tuning of the cut parameters with traditional methods and with machine learning techniques
- To speed up the analysis cycle minimizing the time spent in job submission, I/O and processing
- TTrees can be downloaded and stored in local servers or farms for even faster processing

New ROOT data structures for analysis

Currently testing the use of **RDataFrame** for performing analysis on the Nano Trees:

Many advantages:

- ✓ Allows us to use flexible python interface, while preserving high-speed capacities of compiled C++ objects
- ✓ Friendly interface for event and candidate selection “Pandas-like”
- ✓ With few extra-functionality which are under development in collaboration with the ROOT team, we will be able to move from event-based TTrees to candidate-based TTree



Event based tree filled with std::vector for candidate variables

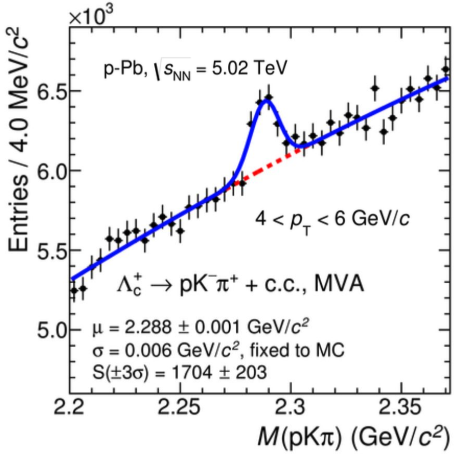
“Flat” candidate-based tree (easy conversion to Pandas DataFrames)

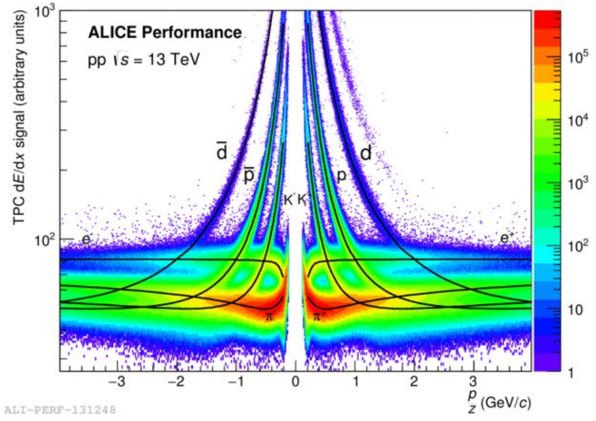
- ✓ “Explode” very useful for e.g. converting analysis tree to flat tree for ML optimisation
- ✓ “Group-by” will allows us to group objects from Run3 timeframes in “events” in Run3

Big thanks to Danilo Piparo for his availability to develop these new functionalities!

Machine Learning techniques for analysis and calibration

JHEP 04 (2018) 108





Machine Learning techniques *were used* (TMVA) and *are being developed* (Python) for:

- ✓ Improving the PID selection strategy
- ✓ Optimise the selection of rare signals like Λ_C or B meson decays
- ✓ Develop new techniques of underlying event subtraction e.g. for jets and HF-jets
- ✓ Develop selection strategies that minimise **both** systematic and statistical uncertainties
- ✓ ML techniques currently under study for calibration and QA for Run3

Conclusions and outlook

- ✓ Several analysis tools have been developed and used for Run1-Run2 analysis :
  - ➢ Focus on the analysis of low-p_T identified topologies in very high multiplicity environment
  - ➢ *only few selected items covered in this talk*

- ✓ The ALICE upgrade program for Run3 with:
  - ➢ x100 more heavy-ion statistics
  - ➢ Online reconstruction and calibration

  is driving the effort to define new:
  - ➢ Data format
  - ➢ Data processing and analysis workflow for real events for MC simulations
  - ➢ Fast-simulation techniques

Conclusions and outlook

- ✓ Several analysis tools have been developed and used for Run1-Run2 analysis :
  - ➢ Focus on the analysis of low-p_T identified topologies in very high multiplicity environment
  - ➢ *only few selected items covered in this talk*

- ✓ The ALICE upgrade program for Run3 with:
  - ➢ x100 more heavy-ion statistics
  - ➢ Online reconstruction and calibration

  is driving the effort to define new:
  - ➢ Data format
  - ➢ Data processing and analysis workflow for real events for MC simulations
  - ➢ Fast-simulation techniques

Looking forward to contributing to the newly born HSF analysis working group!