Why/what you should know about PPD

Tongguang Cheng, Zhen Hu

tongguang.cheng@cern.ch

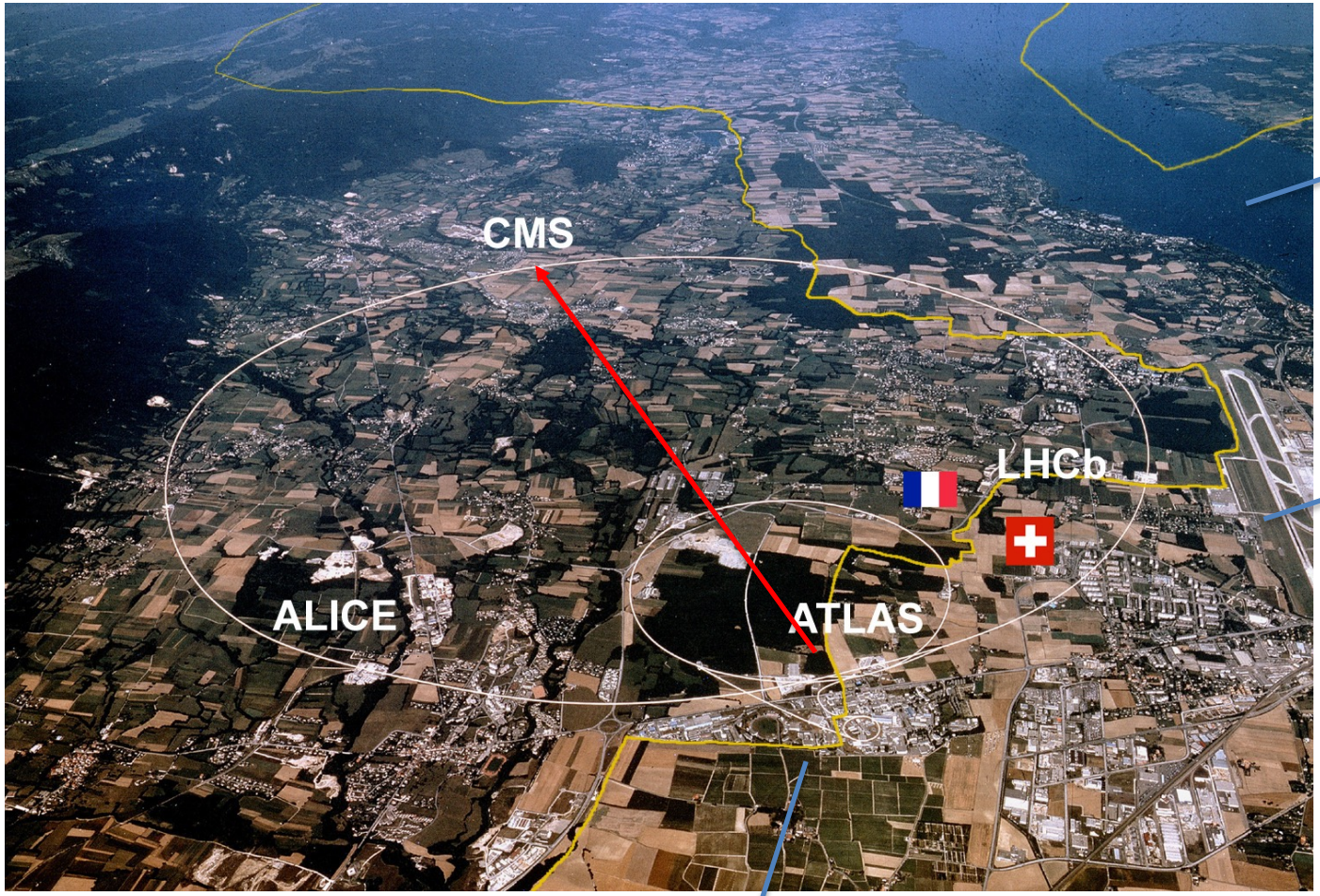
1st China CMS Winter Camp

Caveat

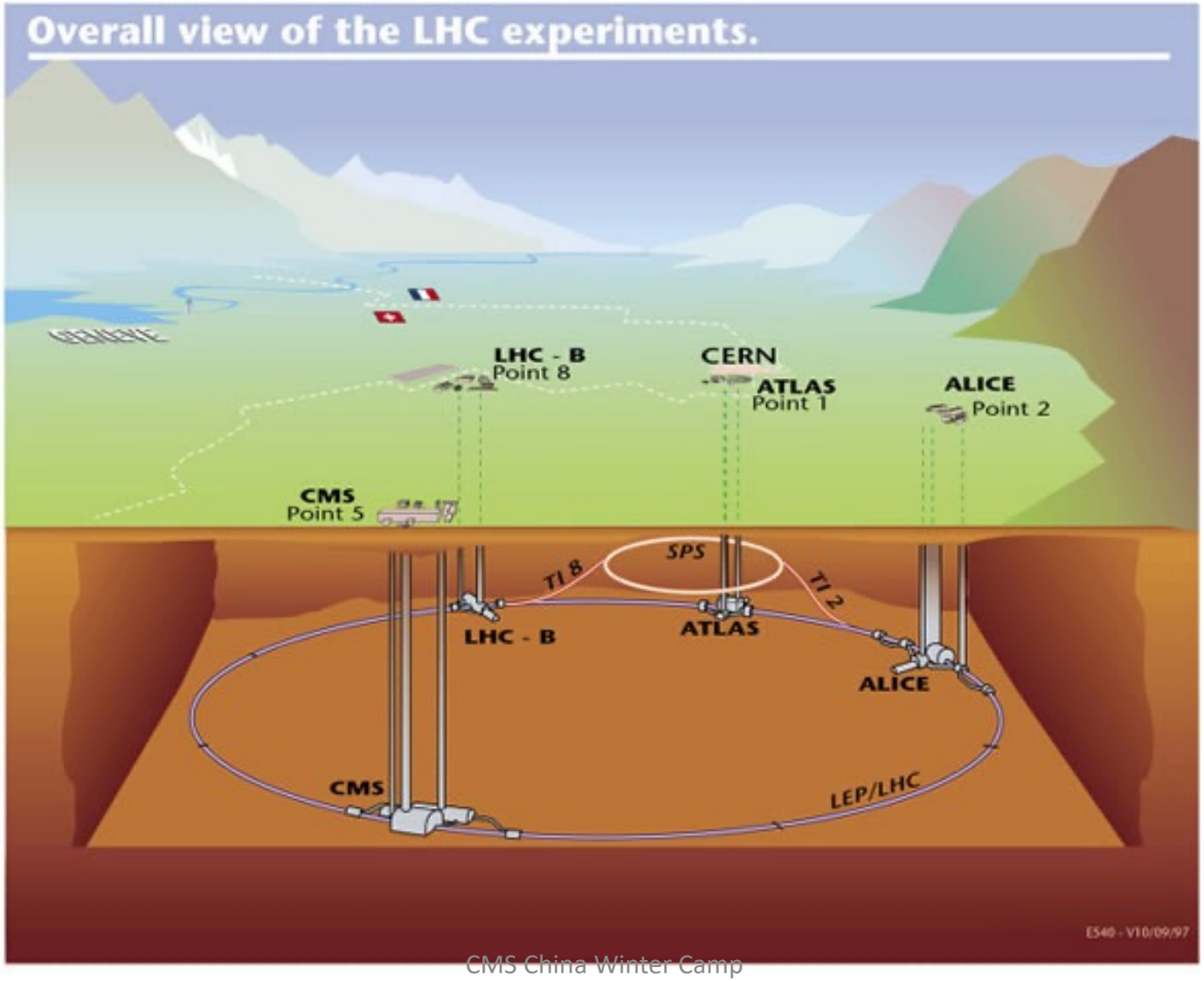
These slides may be not (directly) helpful for the exercise.

The slides try to give some feelings how PPD (and offline computing) gets involved in (offline) data processing and physics analyses.

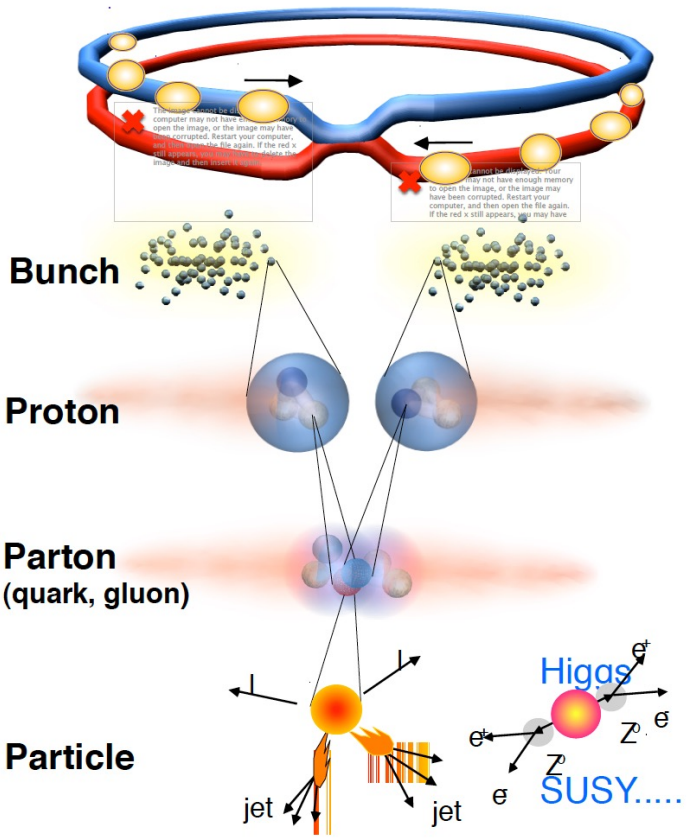
CMS and the Large Hadron Collider (LHC)

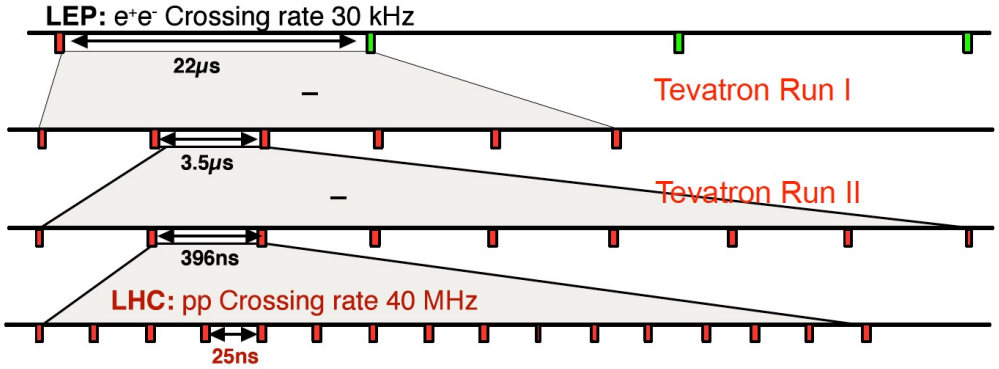


CMS and Point-5



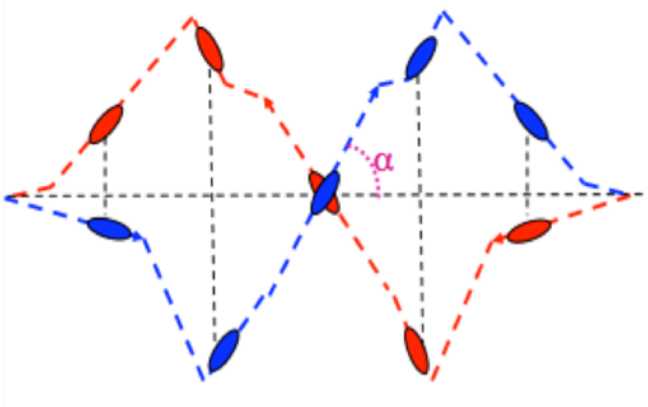
Proton collisions at the CMS

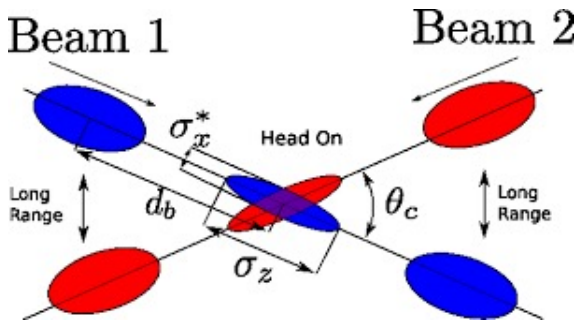


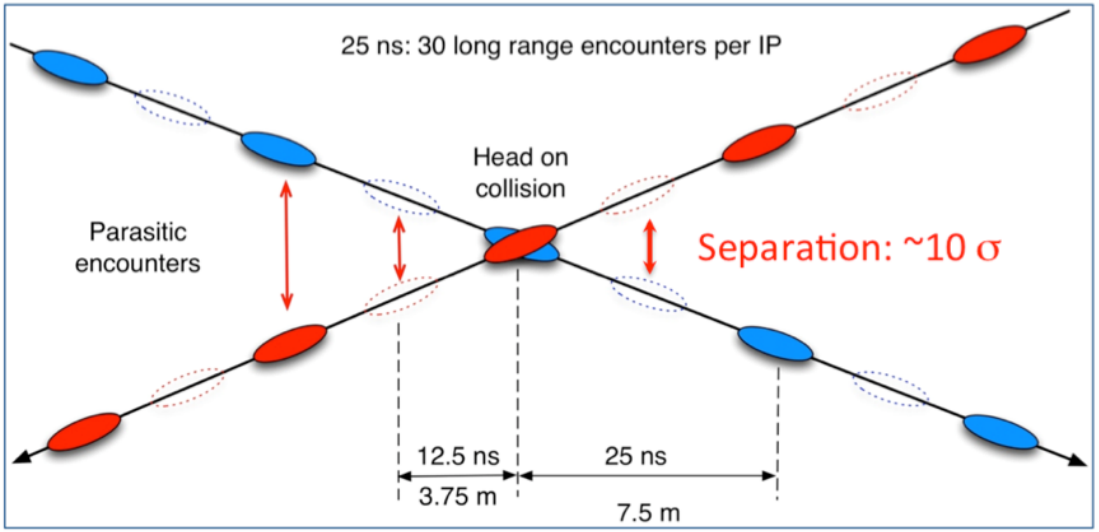


- Protons collide in bunches to increase the chance of rare processes
- Since 2015, LHC provides bunches with 25ns spacing

Proton collisions at the CMS



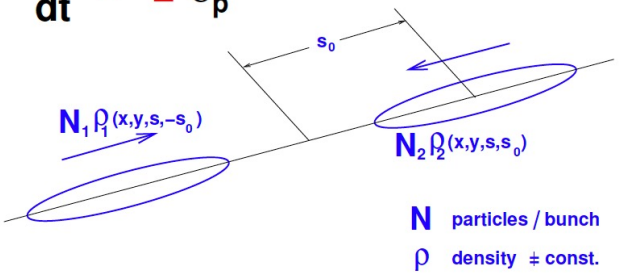




https://home.cern/news/news/accelerators/lhc-report-playing-angles

Proton collisions at the CMS

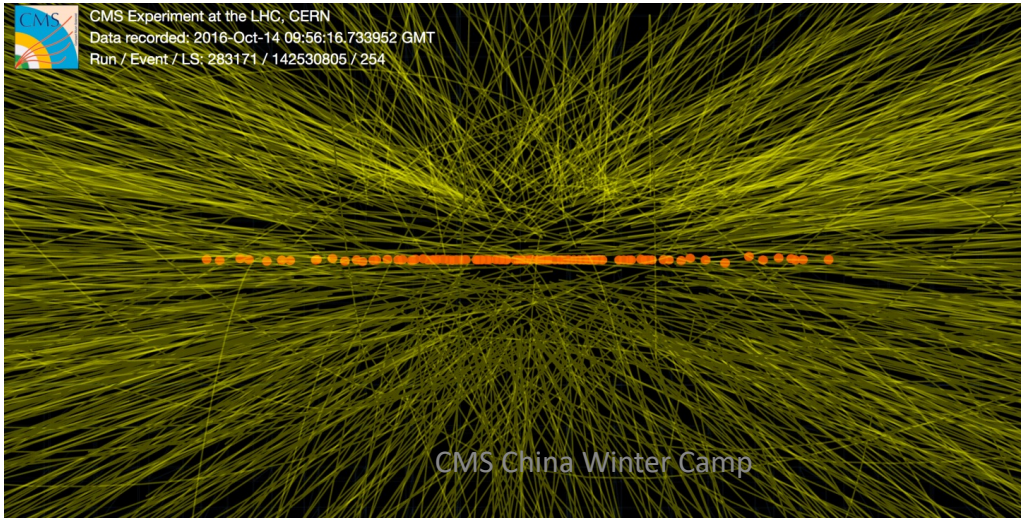
$$\frac{dR}{dt} = L\,\sigma_p$$



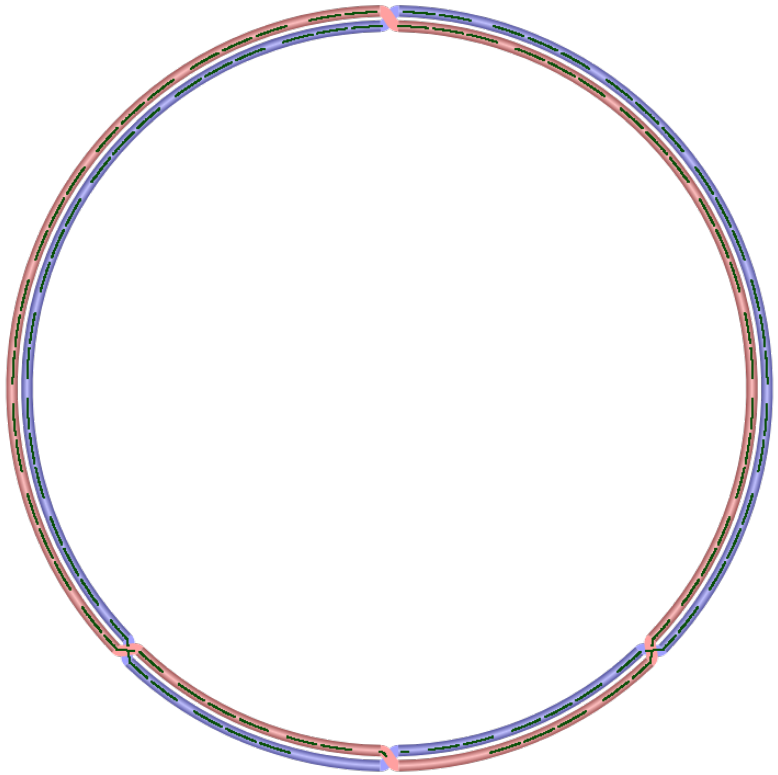
$$\mathcal{L} = \frac{N_1 N_2 f N_b}{2\pi\sqrt{\sigma_{1x}^2+\sigma_{2x}^2}\sqrt{\sigma_{2y}^2+\sigma_{2y}^2}}$$

LHC parameters

- Protons/bunch : ~10^{11}
- Bunch spacing : 25ns
- Max # of bunches : 27km/(c*25ns) ~ 3600
- Luminosity : L=2x10^{34} $cm^{-2}s^{-1}$
- Average number of interactions per bunch crossing (**in-time pileup**) :
  n = L x $\sigma_{minbias}$ x 25ns x (3600/2556) ~ 50-60
  - **out-of-time pileup** : contribution from different(previous) bunch crossings

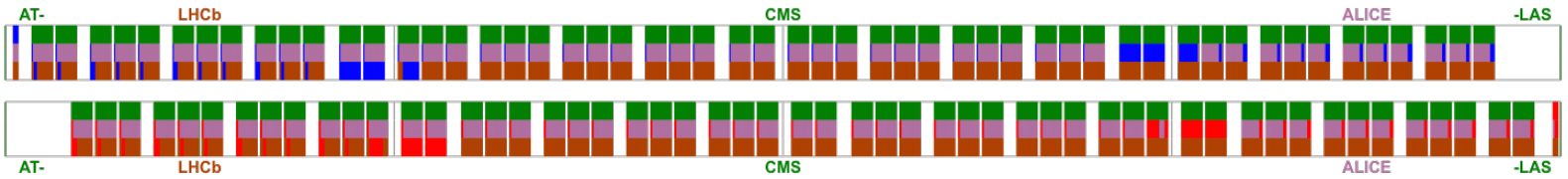


Proton collisions at the CMS



Bunch configuration

- Not all bunches are filled
- Pileup depends on the filling schemes



https://lpc.web.cern.ch/cgi-bin/filling_schemes.py

CMS Data Preparation and Coordination

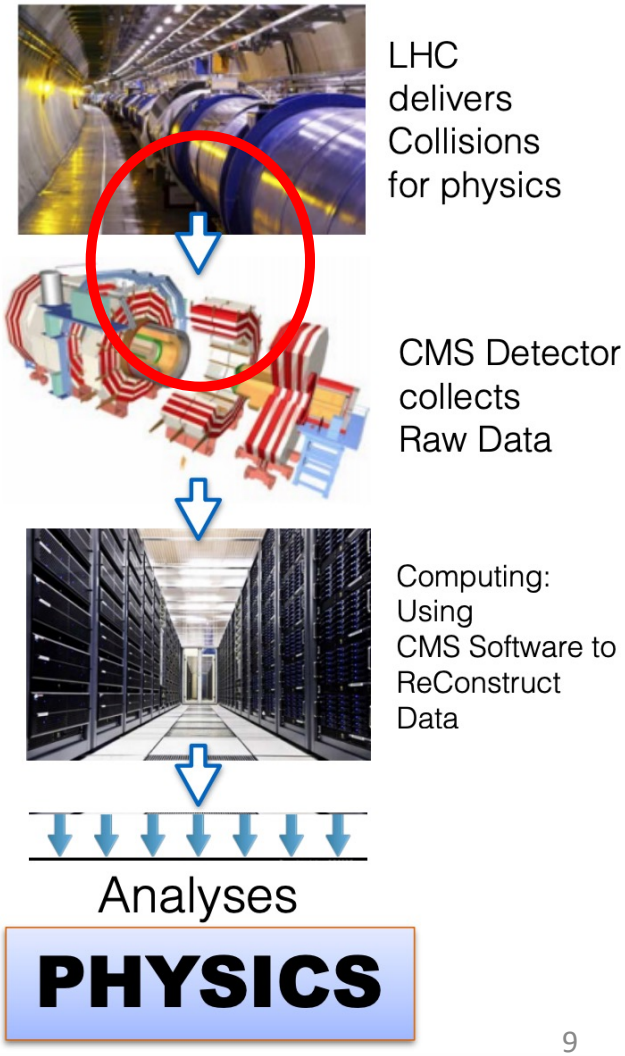
CMS coordination

for Data Acquisition and Preparation

- Run coordination

  Online "Real Data" Collection in **RAW** data format at **Point5**

  - communicate with LHC
  - coordinate CMS detector subsystems, Trigger, Data acquisition, Online monitoring etc.
  - communicate with Technical Coordination for the infrastructure status such as magnets, power, cooling, gas systems, etc.
- Trigger coordination : L1 and HLT trigger



CMS Data Preparation and Coordination

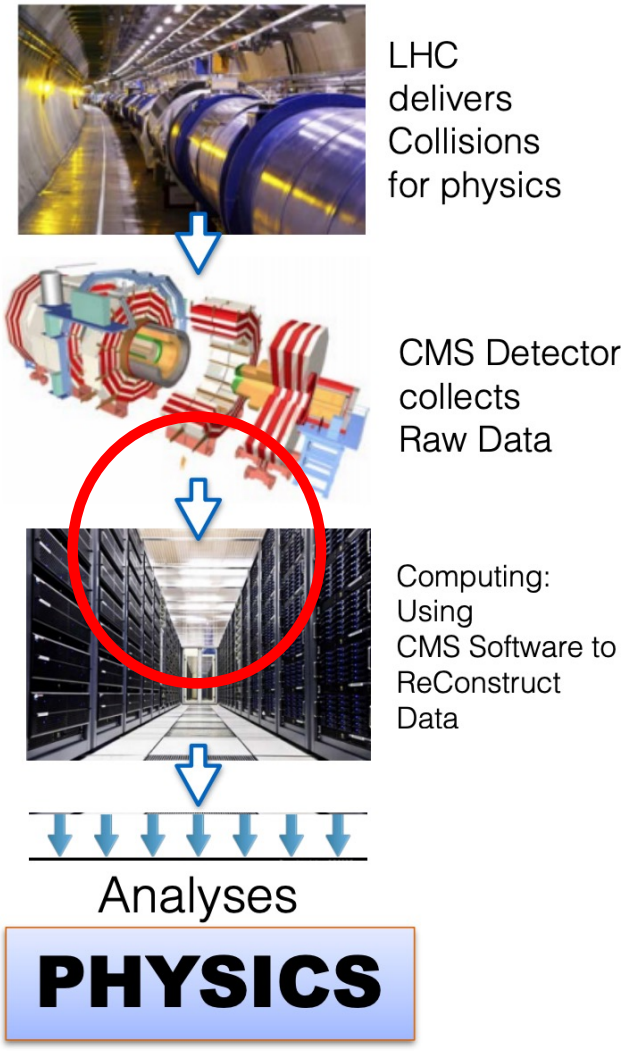
CMS coordination

for Data Acquisition and Preparation

- **Offline & Computing (O&C)**

  offline data/Monte Carlo(MC) events

  - CMSSW software development, event reconstruction and simulation
  - data processing and simulated events(MC) generation (This is mainly what your exercise is about.)
  - data/MC events storage and management



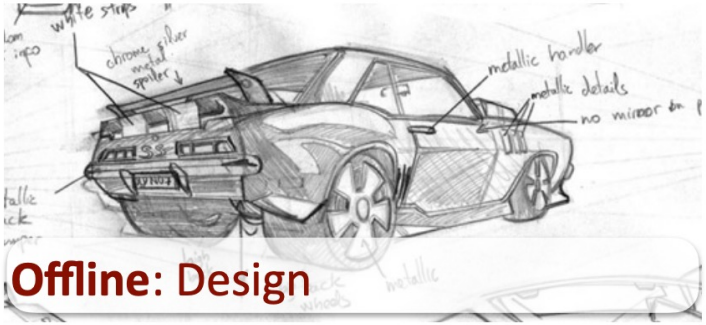
CMS Data Preparation and Coordination

CMS coordination

for Data Acquisition and Preparation

- **Physics Performance and Datasets (PPD)**
  - Data quality & Data certification (DQM-DC)
  - Alignment, calibrations and database (AlCaDB)
  - Physics Data and MC validation (PdmV)
    (This is mainly what your exercise is about.)

If you don't know or not familiar with PPD, it is just because it works well so far.



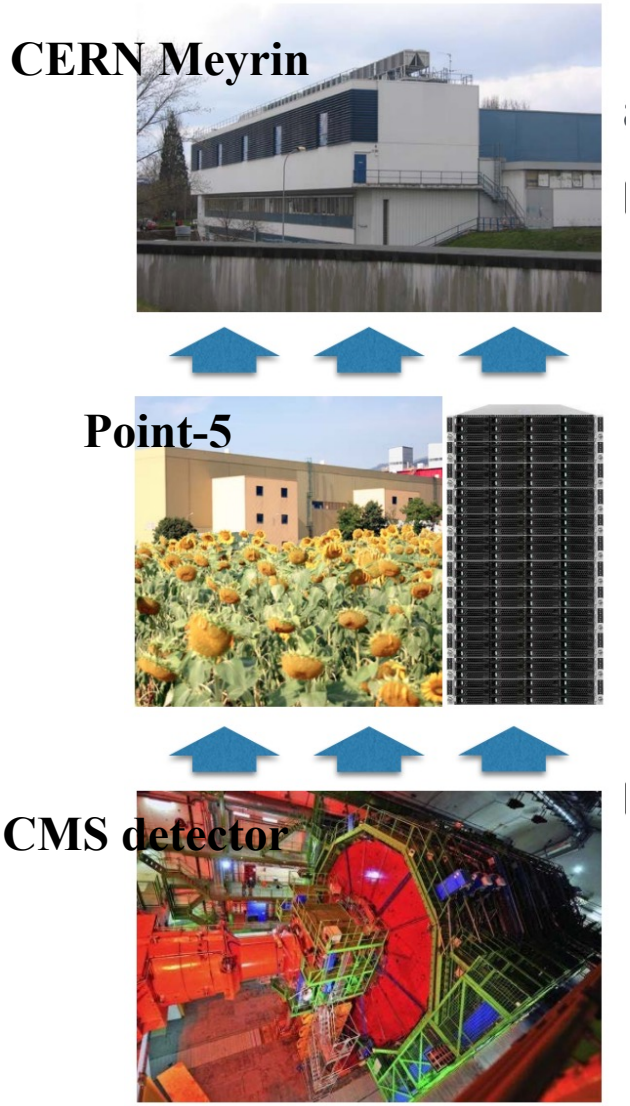
Data flow : from P5 to offline

Events collected by CMS reach the Tier-0 at CERN for tape archival

(Tape is the final destination for RAW data)

Data streams:

- **Express:**
  available ~2h after data collection.
  bandwidth shared by alignment/calibrations, detector/physics monitoring
- **Alignment/Calibration:**
  dedicated event selection/event content devised for calibration process



Data flow : from P5 to offline

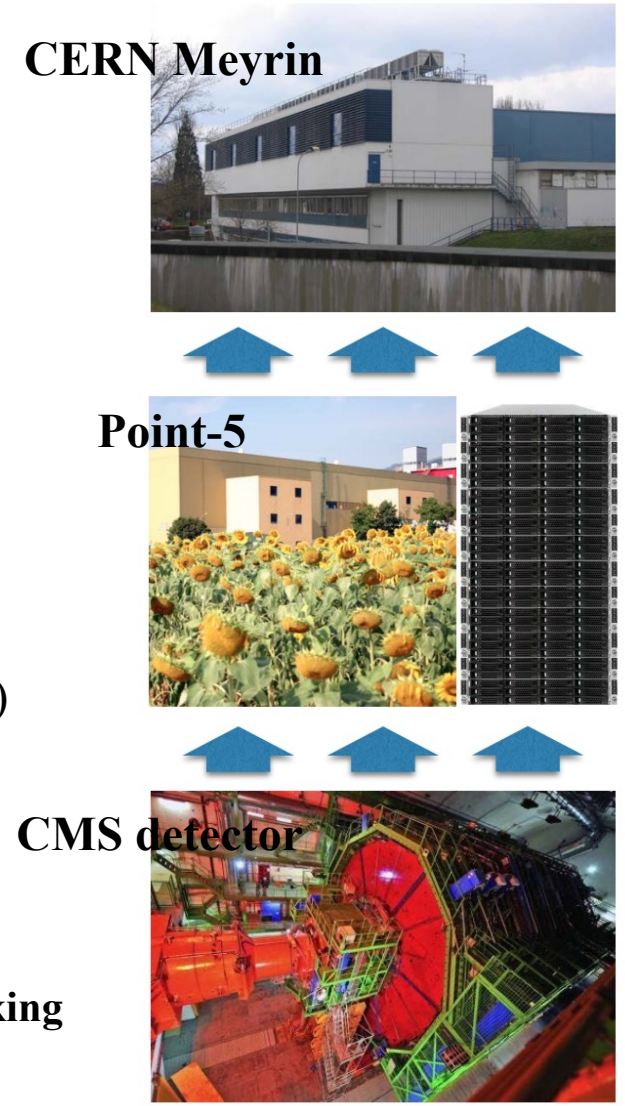
Events collected by CMS reach the Tier-0 at CERN for tape archival

(Tape is the final destination for RAW data)

Data streams:

- **Physics:**
  split into primary datasets and promptly reconstructed for physics analyses (**Prompt-Reco**)
- Other specialized streams:
  Scouting/Parking

Data rates in Run II: **1 kHz of Prompt-Reco**
+ **high rate of scouting data with reduced event content** + **parking**



Express data and Prompt reconstruction

- **Express:**

  data processed for monitor,

  calibration, beamspot and alignment

- **Prompt calibration Loop (PCL)**

  Express data is used as input to automated calibration workflows running at Tier-0 :

  strip gain, pixel large structure alignment, beamspot, etc.

- **Prompt Reco**

  Physics streams (datasets from physics analyses) reconstructed consuming calibrations from PCL. Normally start prompt reconstruction within 48 hours

  (not a hard limit but has limited extension)

Interlude : alignment/calibration workflows

Workflows for different time scales of updates

(sometimes means speed to deliver the calibration,

sometimes means the statistics need to derive the calibration)

- **Quasi-online calibrations for HLT and express :**

  example : O2O (online to offline)

- **Prompt calibration (Loop) :**

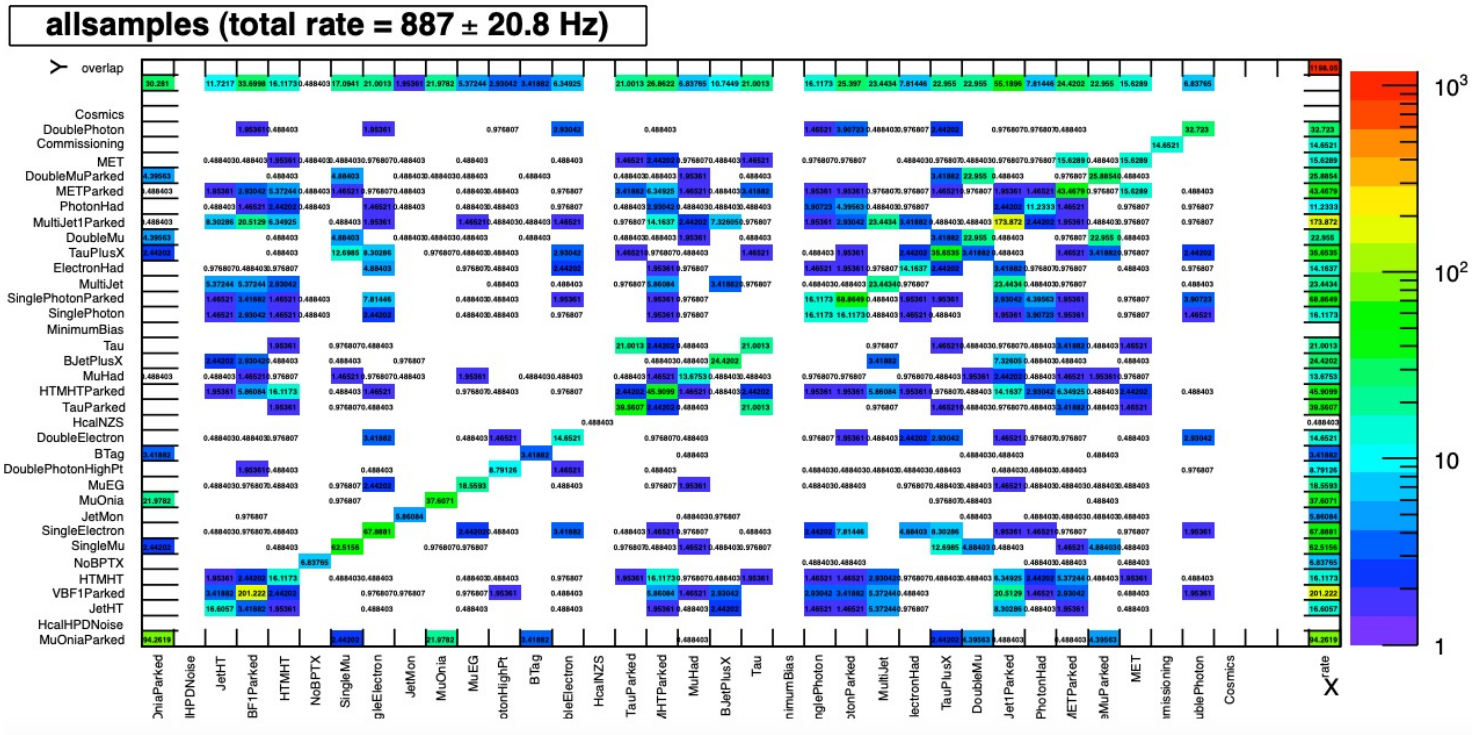
  monitor and update conditions expected to vary run-by-run, or per lumi-section to guarantee performance of prompt reco

- **Offline calibration:**

  use alignment/calibration dataset and prompt-reco physics datasets to be used by End-of-Year (End-of-data taking period) re-reconstruction

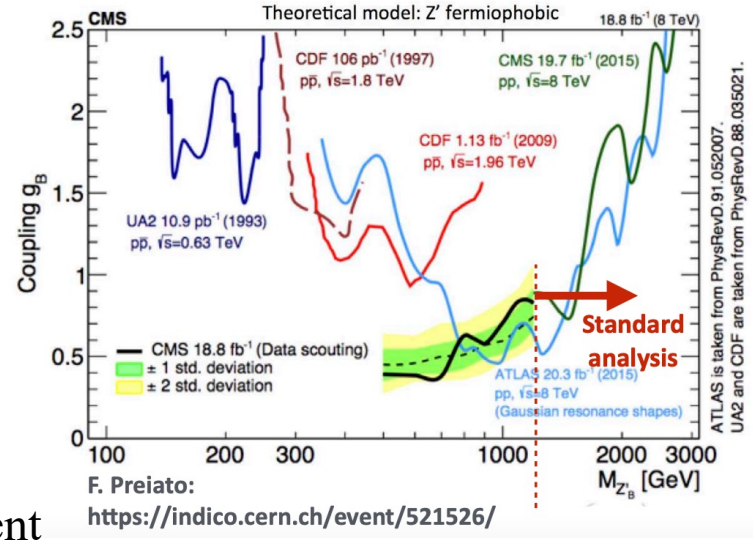
Primary datasets

The physics steams from P5 are split to Primary Datasets (PD) on the basis of HLT results in order to group events with related topology and limited overlap among different PDs

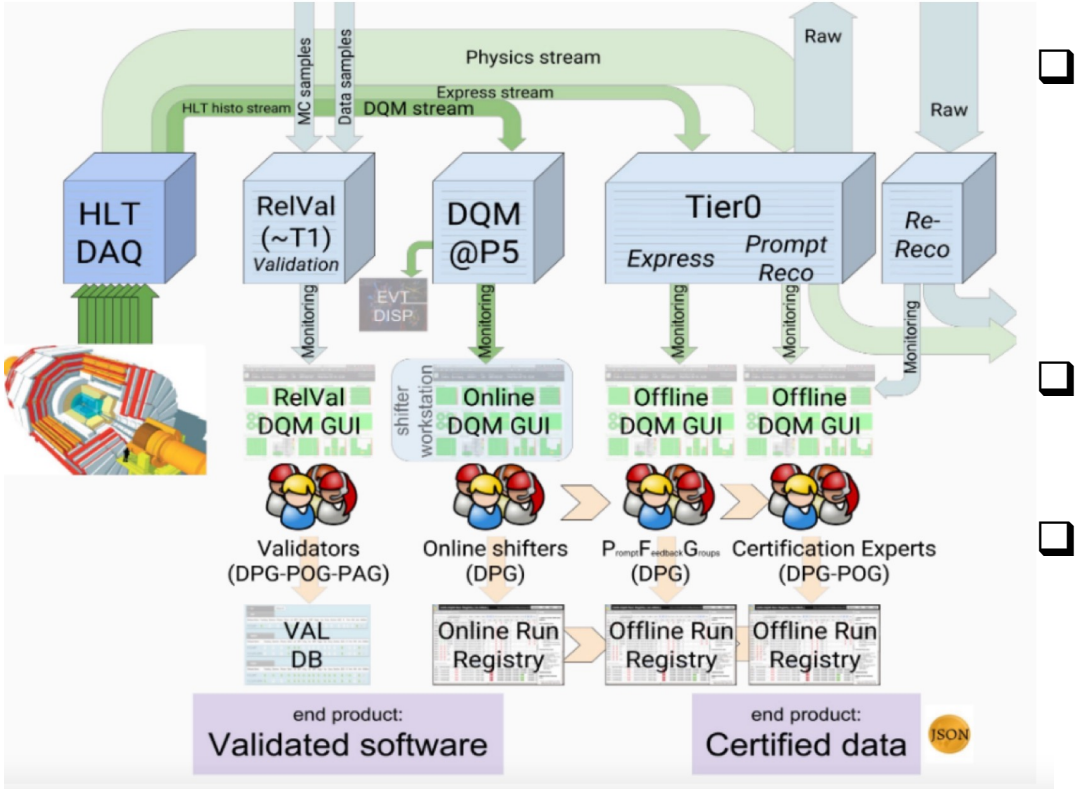


Data scouting and parking

- Trigger rates are constrained by the CMS prompt reconstruction system, which cannot process much more than 1 kHz of events.
  - cannot get more events by simply adding non-overlapping selection paths
- **To by-pass the computing limit**
  - Data parking: send events from the HLT to tape without reconstruction
  - Data scouting: save only a small subset of the event content (e.g., only the HLT-level jet objects)
  - Use in physics analyses searching for physics beyond the Standard Model for e.g., Z', dark photon



PPD: DQM and DC



- DQM :
  DQM packages run in CMSSW, create histograms/plots for monitoring
- DQM GUI :
  to display the DQM Histograms/plots
- Run Registry:
  to keep track of monitoring/ certification results

PPD: DQM and DC

- Data certification:

  provide central data certification –

  good runs/lumi sections to be used for most of the physics analyses

- Central data certification information at

  https://twiki.cern.ch/twiki/bin/viewauth/CMS/DataQuality

- Golden JSON require all sub-detectors/POGs to be “GOOD”.

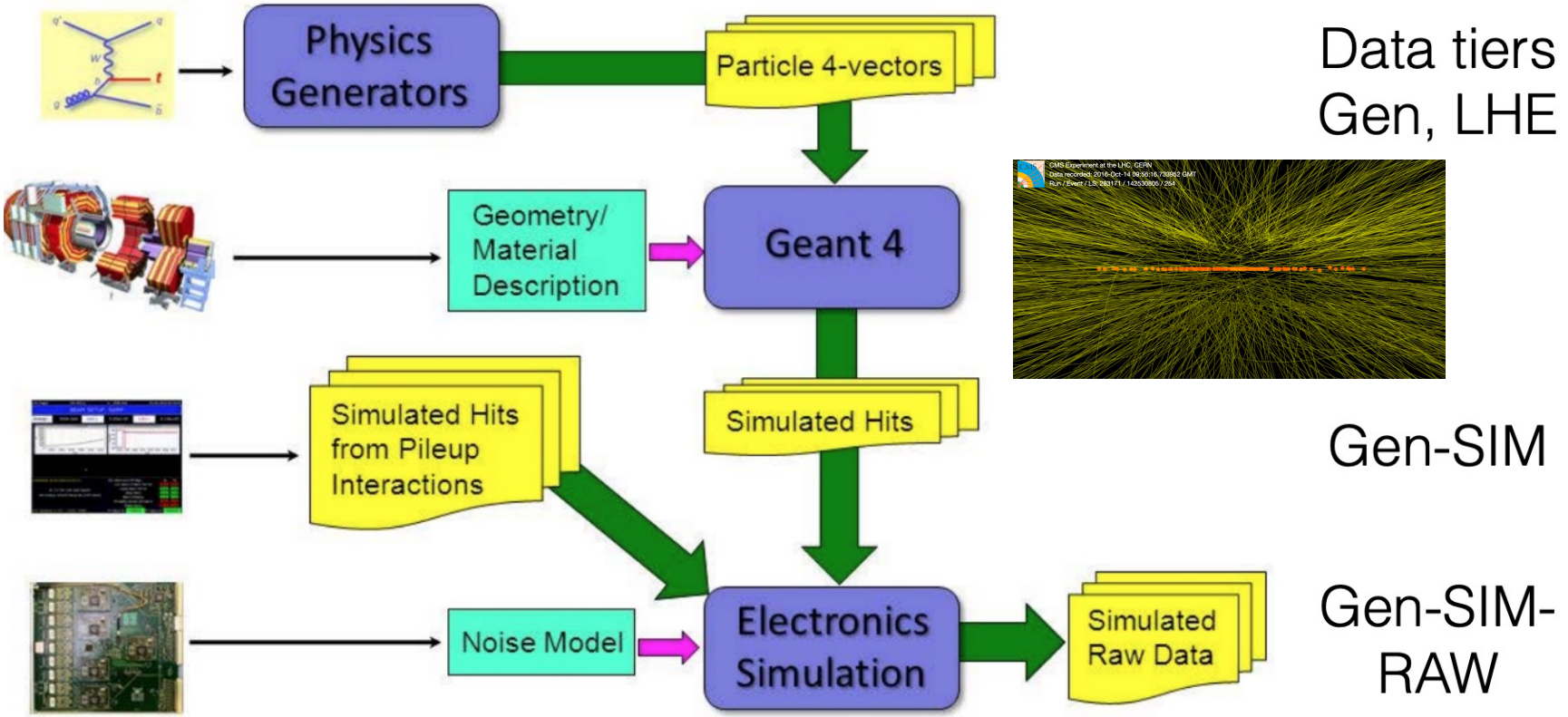
  File information are announced in Physics Validation “HyperNews”.

  https://cms-service-dqmdc.web.cern.ch/CAF/certification/

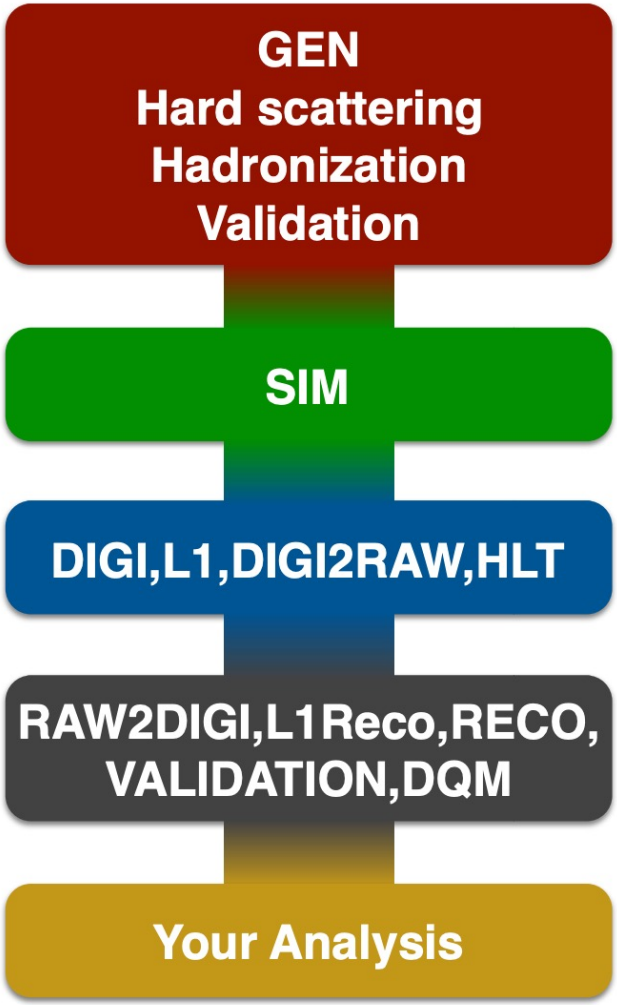
Event simulation (Monte Carlo)

The simulation sequence aims at producing MC truth and Raw data as it comes from point 5.

(I will leave the treatment of **pileup** to you when you work on the exercise.)



From data/MC to your (physics) analyses



What physics are we studying?
What generator should be used?
Which parameters should be modified?

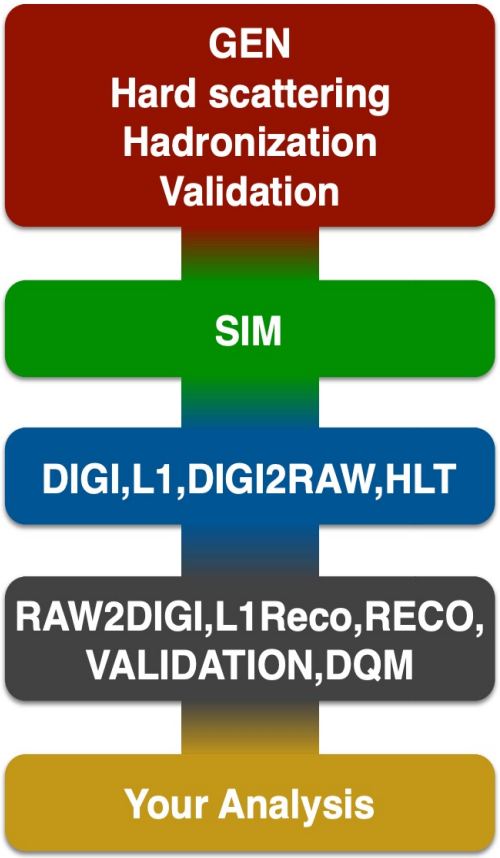
CMS Geometry, Magnetic field,...

Detector electronics simulation, Pileup situation,
Alignment-Calibration, Trigger menu,...

Reconstruction algorithms,
High level object creation

Your analysis code

Why you should know about AlCaDB (PPD)



What physics are we studying?
What generator should be used?
Which parameters should be modified?

CMS Geometry, Magnetic field,...

Detector electronics simulation, Pileup situation,
Alignment-Calibration, Trigger menu,...

Reconstruction algorithms,
High level object creation

Your analysis code

AlCa mainly comes into the game in digitization and reconstruction，also crucial for HLT (for e.g., JEC at HLT)

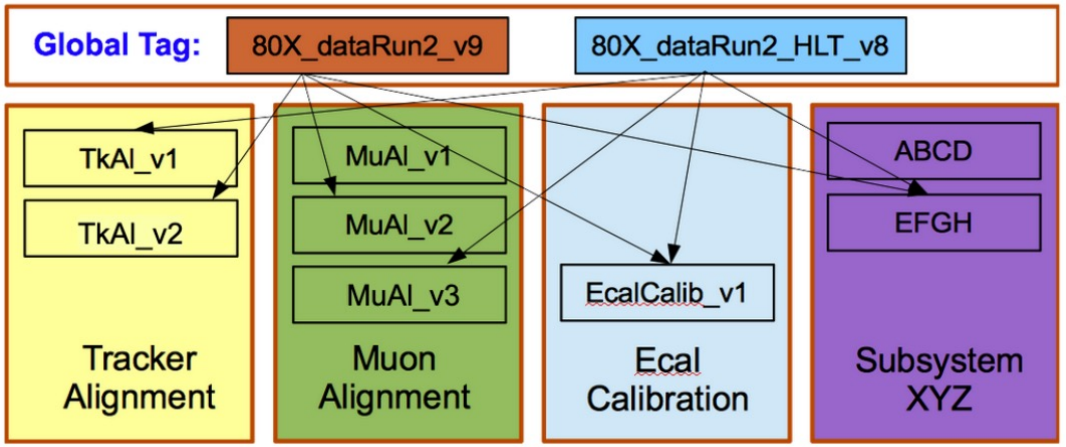
AlCa terminology : condition, payload, Tag

- The "atom" of condition data is the **Payload**, it
  - represents the set of parameters consumed in the data/MC processing
  - associated to a C++ class in CMSSW (condition interface to CMSSW)
- **The time information for the validity of the Payloads** is specified with a parameter called Interval Of Validity (IOV)
  - Time is represented by a Run number, luminosity section id or an universal timestamp
- **Tag :**

  a fully qualified set of conditions consists of a set of Payloads and their associate IOVs covering the time span required by the workload

AlCa terminology : global Tag

- A collective label called **Global Tag** identifies **the set of Tags assigned to the Records (condition entry toDB)** involved in a given data/MC processing flow
- Global Tags provides the full set of AlCa content
  - for a Monte Carlo production scenario (campaign)
  - for a data reprocessing scenario (campaign)
- AlCaDB has strategy to validate Tags (condition update)
- Campaign v ı
  by PdmV



AlCa terminology : global Tag customization

- Conditions sometimes need update when analysing data/MC
  - usually related to high level object
  - for e.g., JEC, E/Gamma energy regression

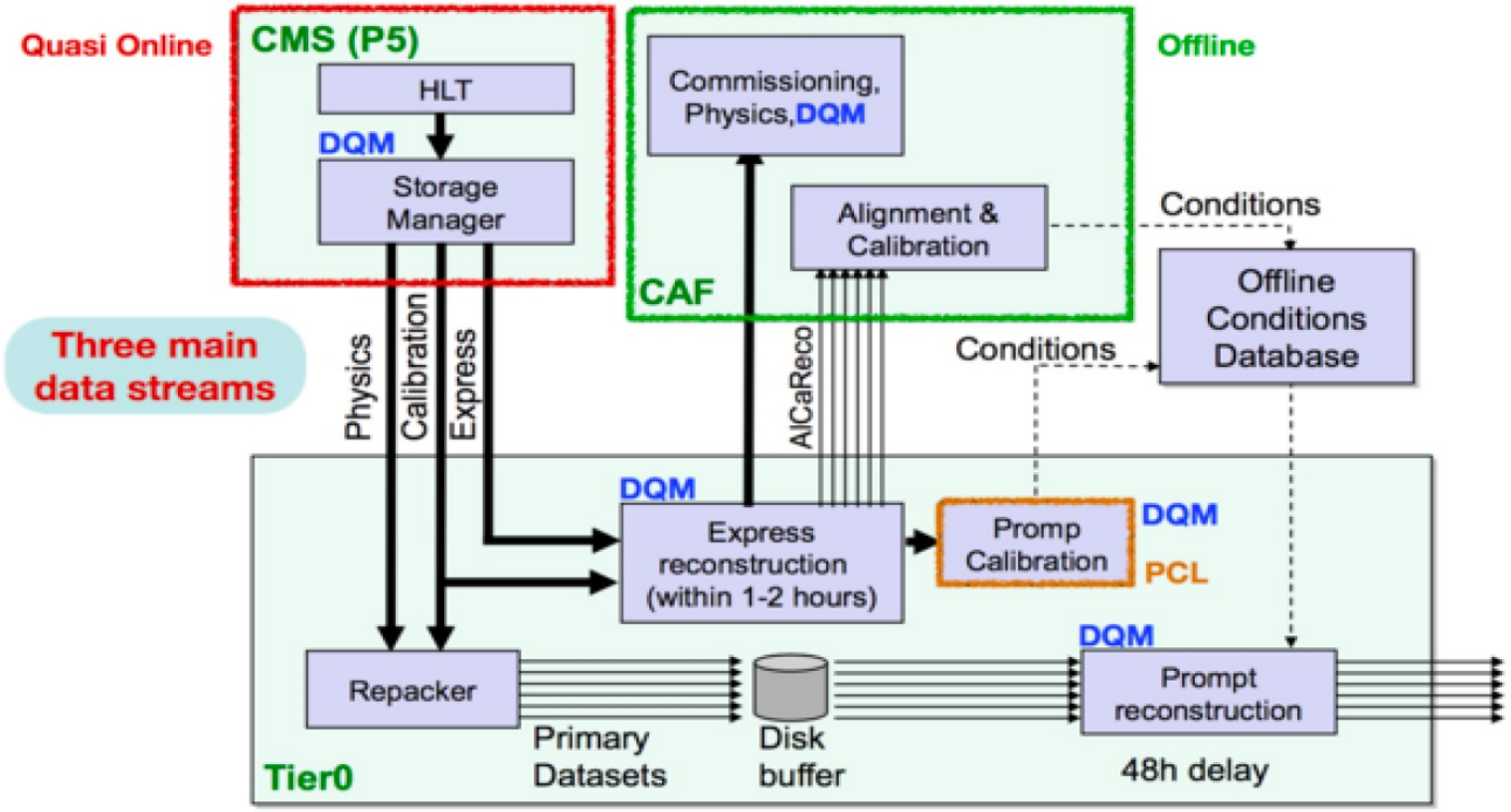
```
process.GlobalTag.toGet.append(
  cms.PSet(
  record = cms.string(“RECORD_NAME”),
  Label = cms.string(“RECORD_LABEL”),
  tag = cms.string(“TAG_NAME"),
  connect = cms.string("frontier://FrontierProd/CMS_CONDITIONS")
  )
)
```

Last but not least important

	Legacy 2016 preVFP (APV)	Legacy 2016 postVFP	Legacy 2017	Legacy 2018
CMSSW	CMSSW_10_6_20 and GT for data analysis 106X_dataRun2_v32 (for all NanoAOD)			
Data	For information about the data, please refer to this page: https://twiki.cern.ch/twiki/bin/view/CMS/PdmVRun2LegacyAnalysis			
Int. lumi.	~19.5 /fb	~16.8 /fb	41.48 /fb	59.83 /fb
MC AOD	RunIISummer19UL16APV(*) RunIISummer20UL16APV	RunIISummer19UL16(*) RunIISummer20UL16	RunIISummer19UL17 RunIISummer20UL17	RunIISummer19UL18 RunIISummer20UL18
MC MiniAODv1	RunIISummer19UL16APV(*) RunIISummer20UL16APV	RunIISummer19UL16(*) RunIISummer20UL16	RunIISummer19UL17 RunIISummer20UL17	RunIISummer19UL18 RunIISummer20UL18
MC NanoAODv8	RunIISummer19UL16APV(*) RunIISummer20UL16APV	RunIISummer19UL16(*) RunIISummer20UL16	RunIISummer19UL17 RunIISummer20UL17	RunIISummer19UL18 RunIISummer20UL18
	The contents of NanoAOD is described in the NanoAOD doc			
GT for MC Analysis	106X_mcRun2 _asymptotic_preVFP_v9	106X_mcRun2 _asymptotic_v15	106X_mc2017 _realistic_v8	106X_upgrade2018 _realistic_v15_L1v1
Recipe	https://twiki.cern.ch/twiki/bin/view/CMS/PdmV#Analysis_Recipe			

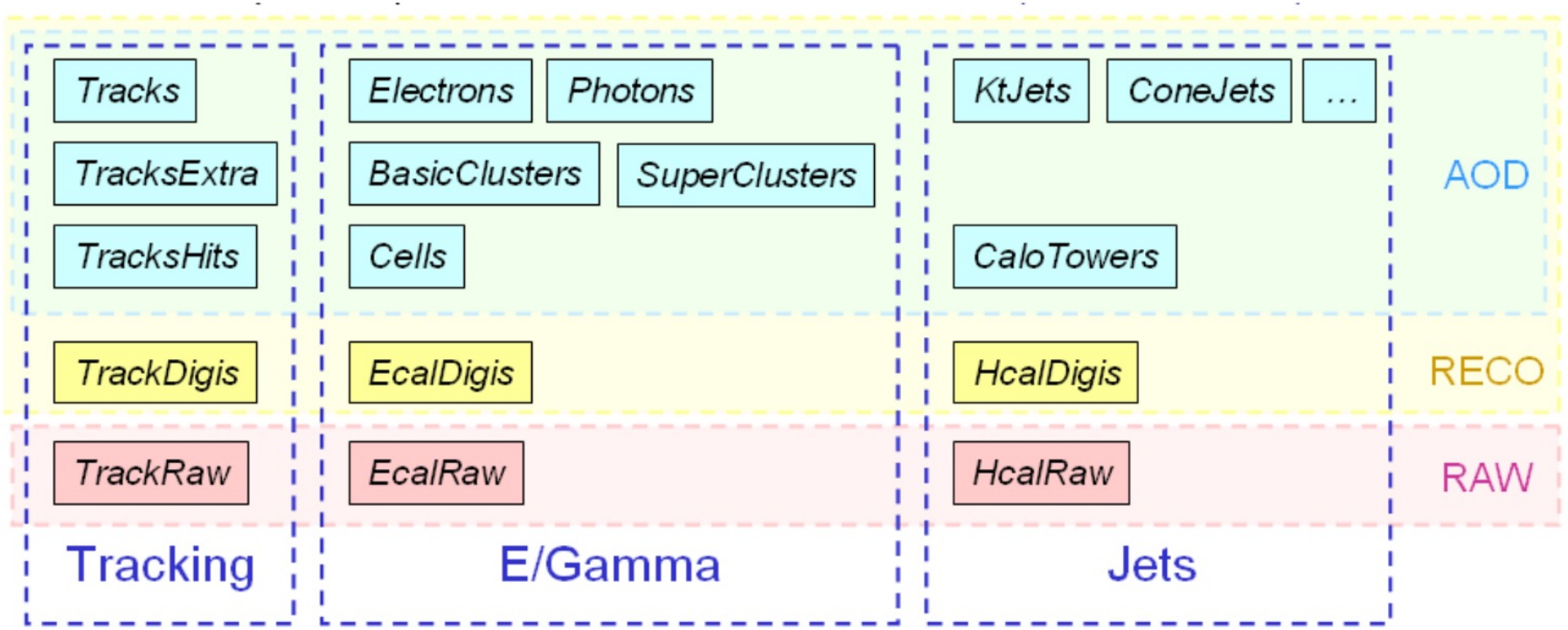
https://twiki.cern.ch/twiki/bin/view/CMS/PdmVRun2LegacyAnalysis

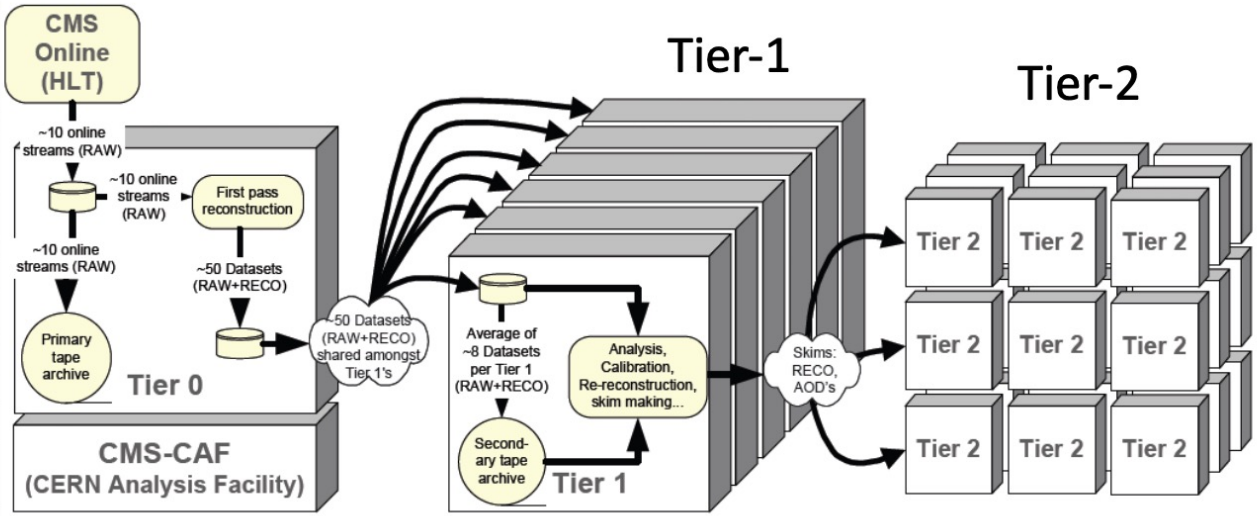
Summary

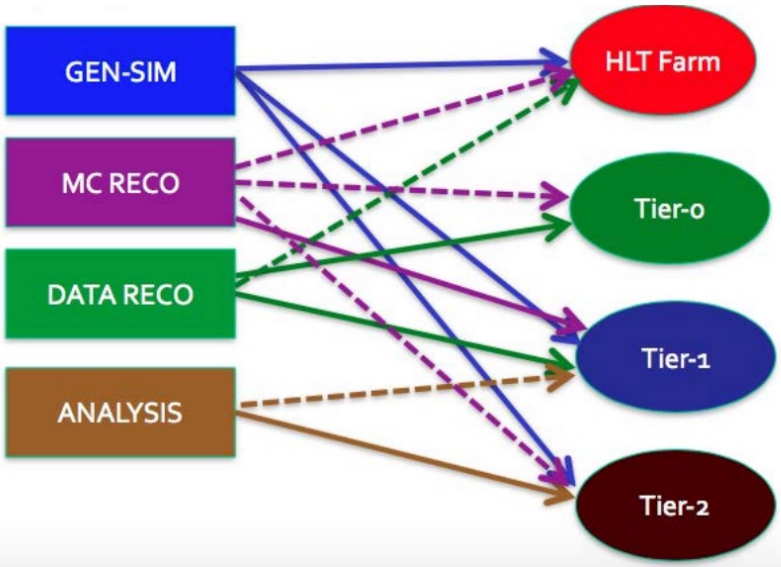

Quasi Online
CMS (P5)
HLT
DQM
Storage Manager
Commissioning, Physics,DQM
Offline
Alignment & Calibration
Conditions
CAF
Offline Conditions Database
Conditions
Three main data streams
Physics
Calibration
Express
AlCaReco
DQM
Express reconstruction (within 1-2 hours)
Promp Calibration
DQM
PCL
DQM
Repacker
Prompt reconstruction
Primary Datasets
Disk buffer
48h delay
Tier0

Backup

Examples of what are in RAW/RECO/AOD

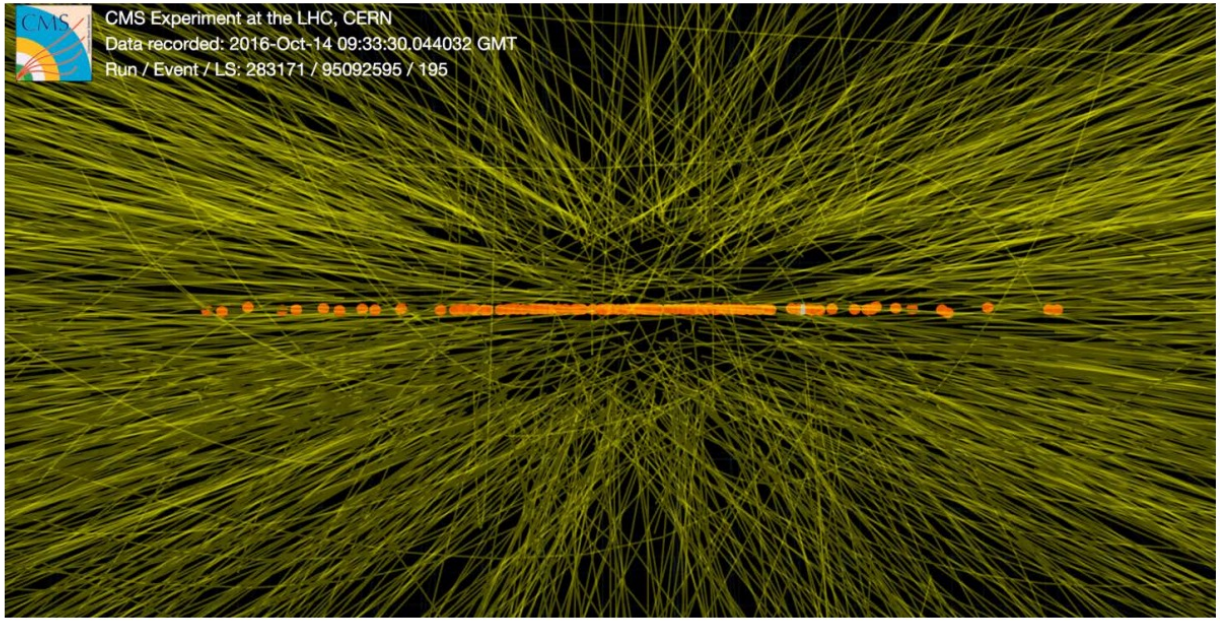


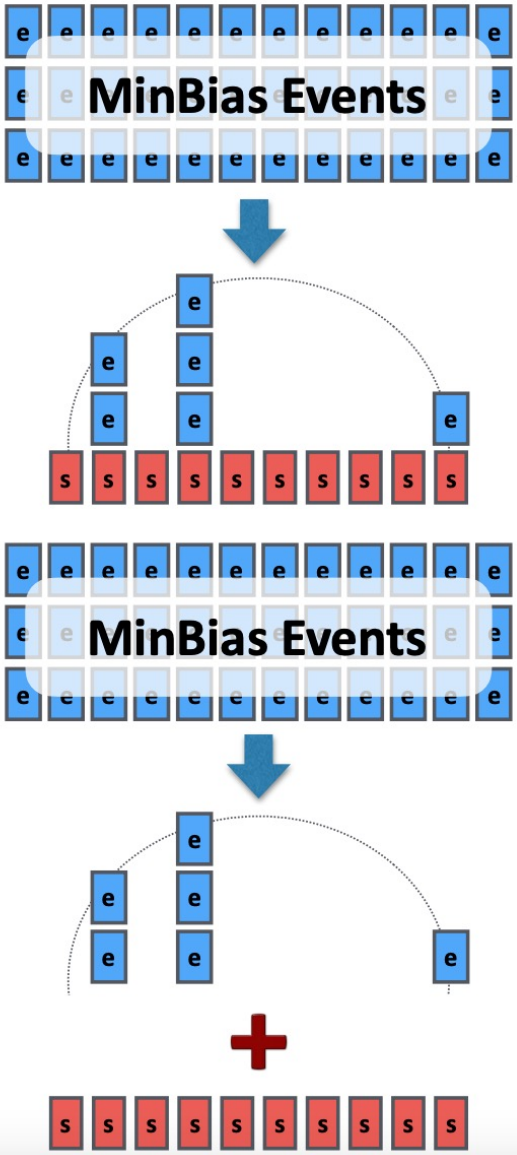




- **Increasing the flexibility for facilities and workflows**
- More places that jobs can run

Run-1 Run-2





Classic mixing

- GENSIM Signal (MC Hard-scatter event) is overlaid with GENSIM MinBias with chosen pileup configuration.

Pre-mixing

- MinBias events in RAWSIM format are overlaid on empty single neutrino events using a chosen pileup configuration. Digis made in this step are converted to RAW.
- 1-1 combination of PreMixed event - signal event. RawToDigi is done on-the-fly to premixed events before overlay.