Why/what you should know about PPD

Tongguang Cheng, Zhen Hu

tongguang.cheng@cern.ch

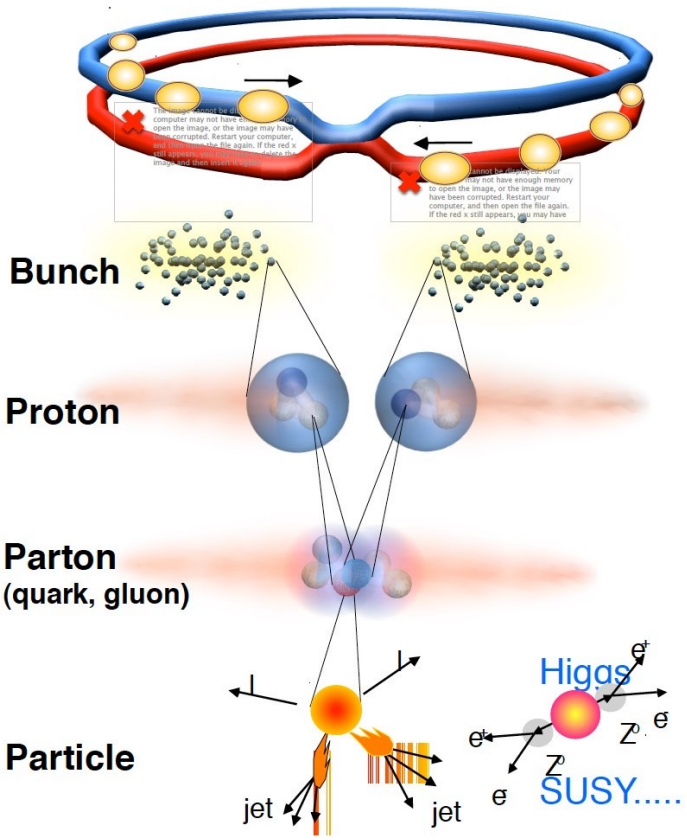
CMS China Winter Camp 2024

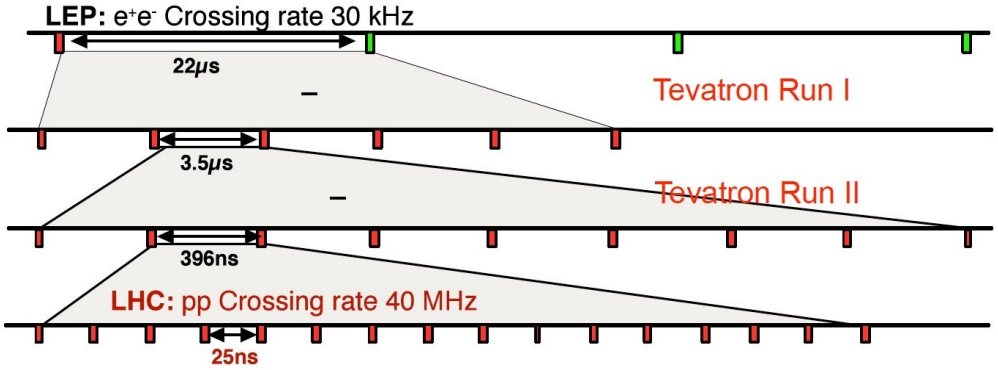
Caveat

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

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

Proton collisions at the CMS

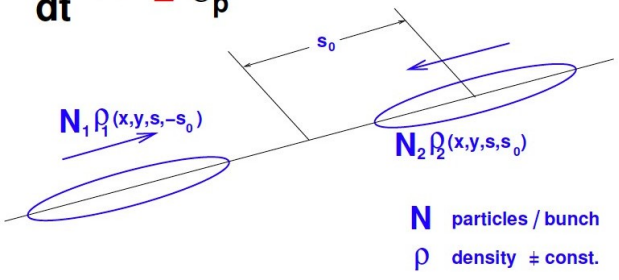




- Protons collide in bunches to increase the chance of rare processes
- Since 2015, LHC provides bunches with 25ns spacing

Proton collisions at the CMS

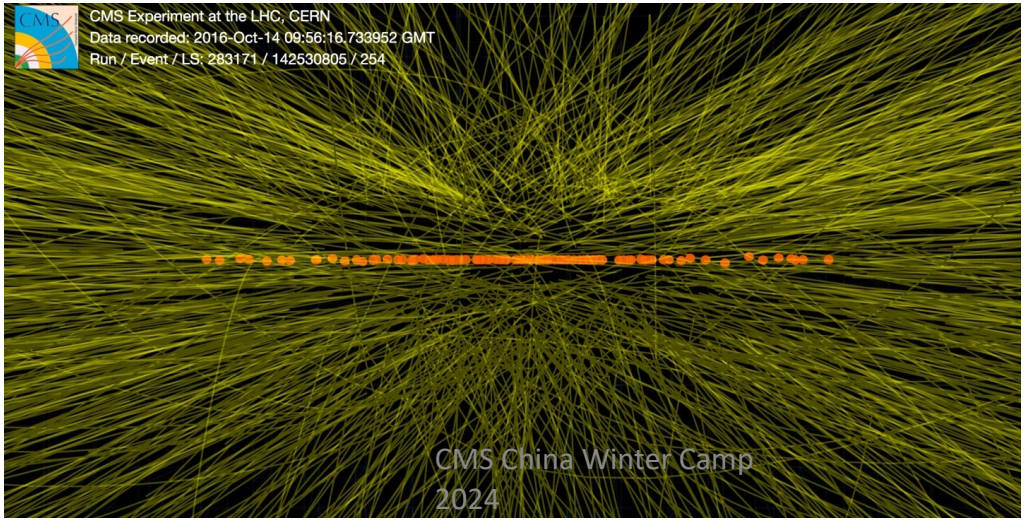
$$\frac{dR}{dt} = L\,\sigma_p$$



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

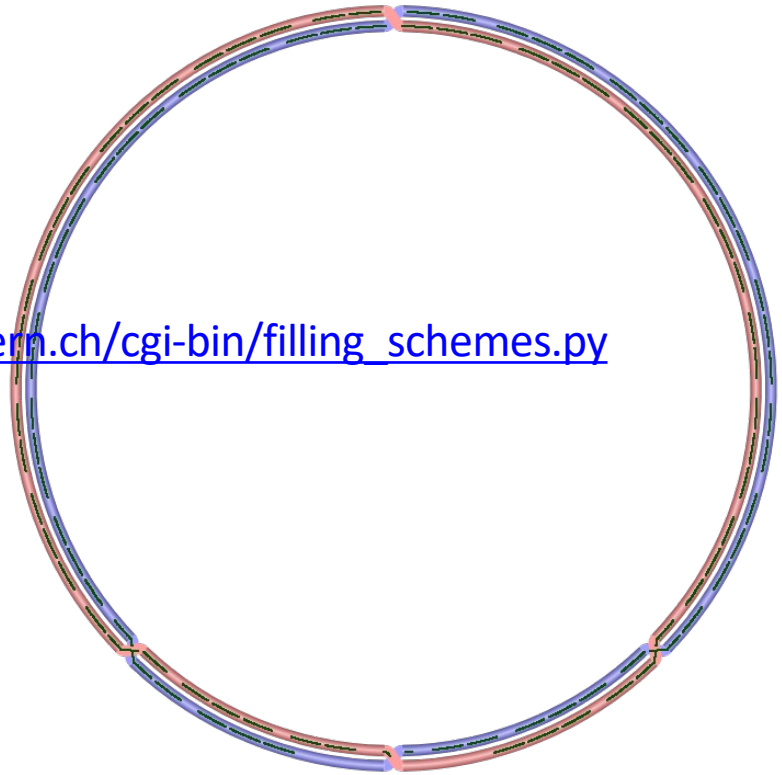
LHC parameters

- 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



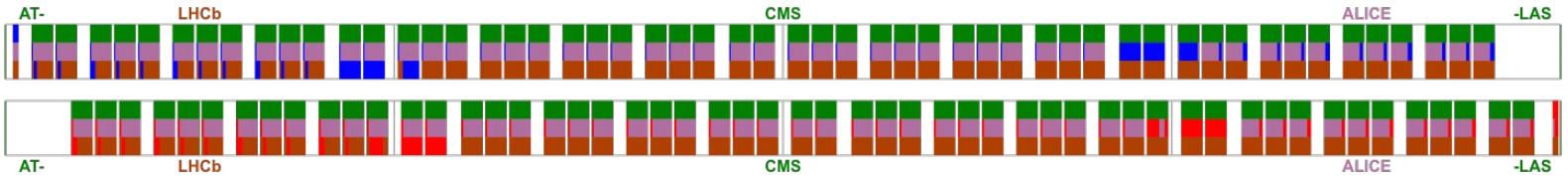
Proton collisions at the CMS

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



Bunch configuration

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



CMS Data Preparation and Coordination

CMS coordination

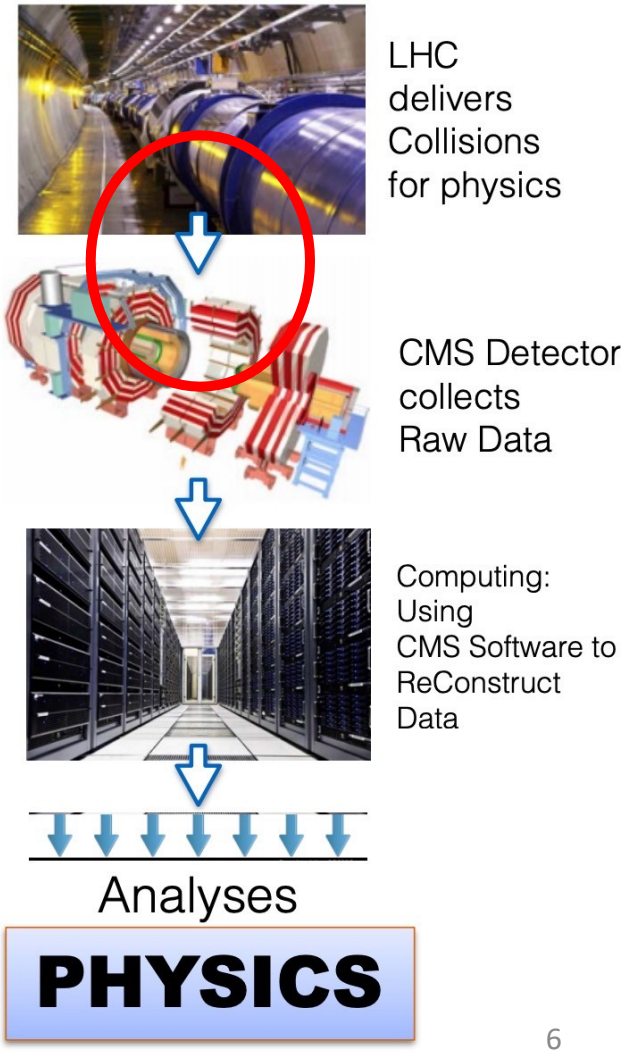
for Data Acquisition and Preparation

- 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

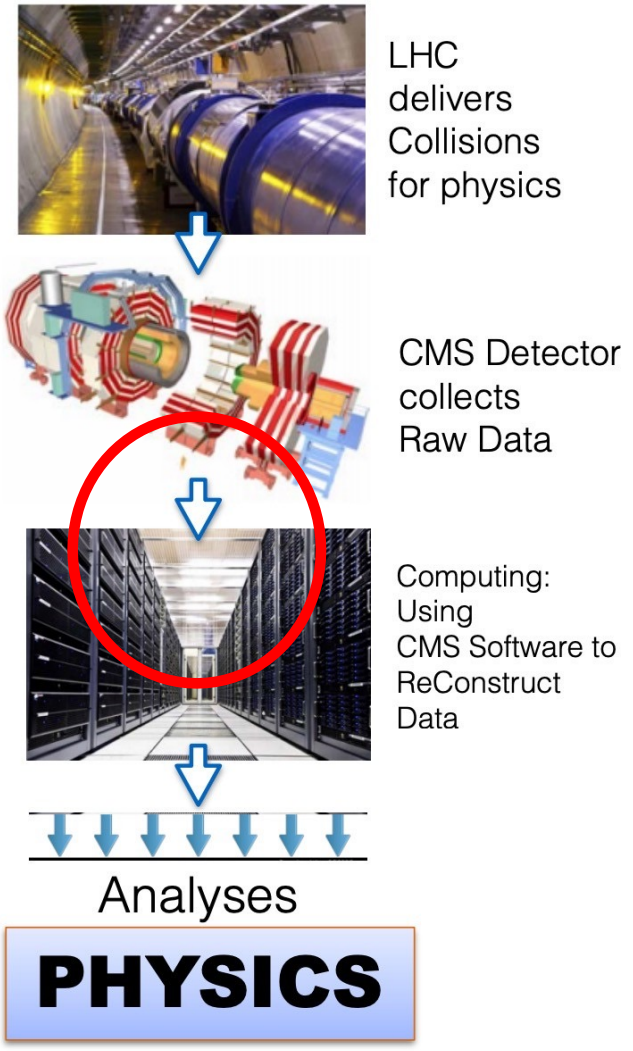
CMS coordination

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

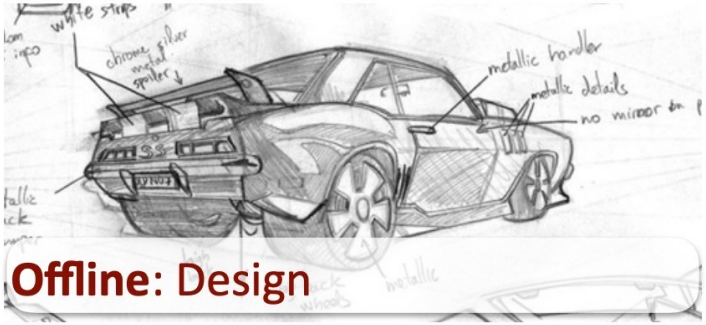


CMS Data Preparation and Coordination

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)
  - Data processing and MC generation (PdmV)



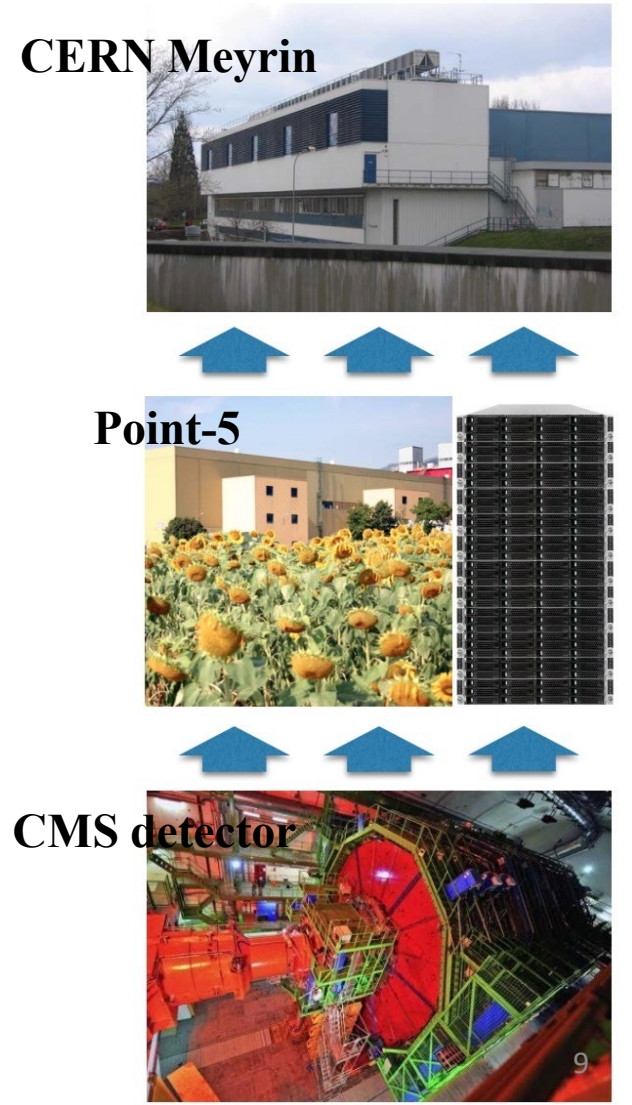
Data flow : from P5 to offline

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

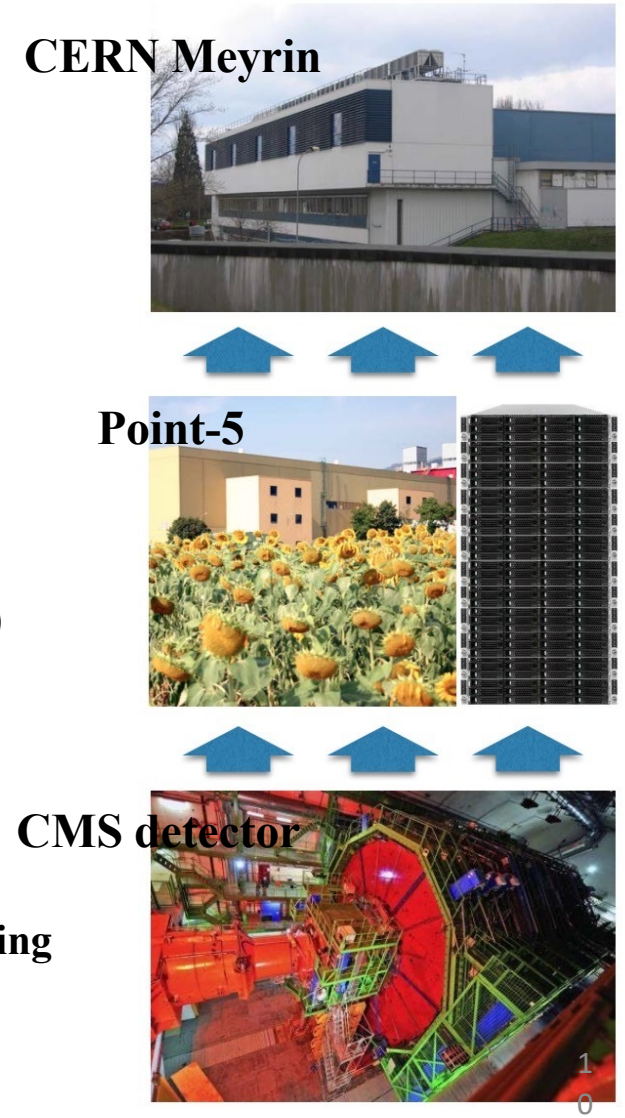
Events collected by CMS reach the Tier-0 at CERN for tape archival

(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

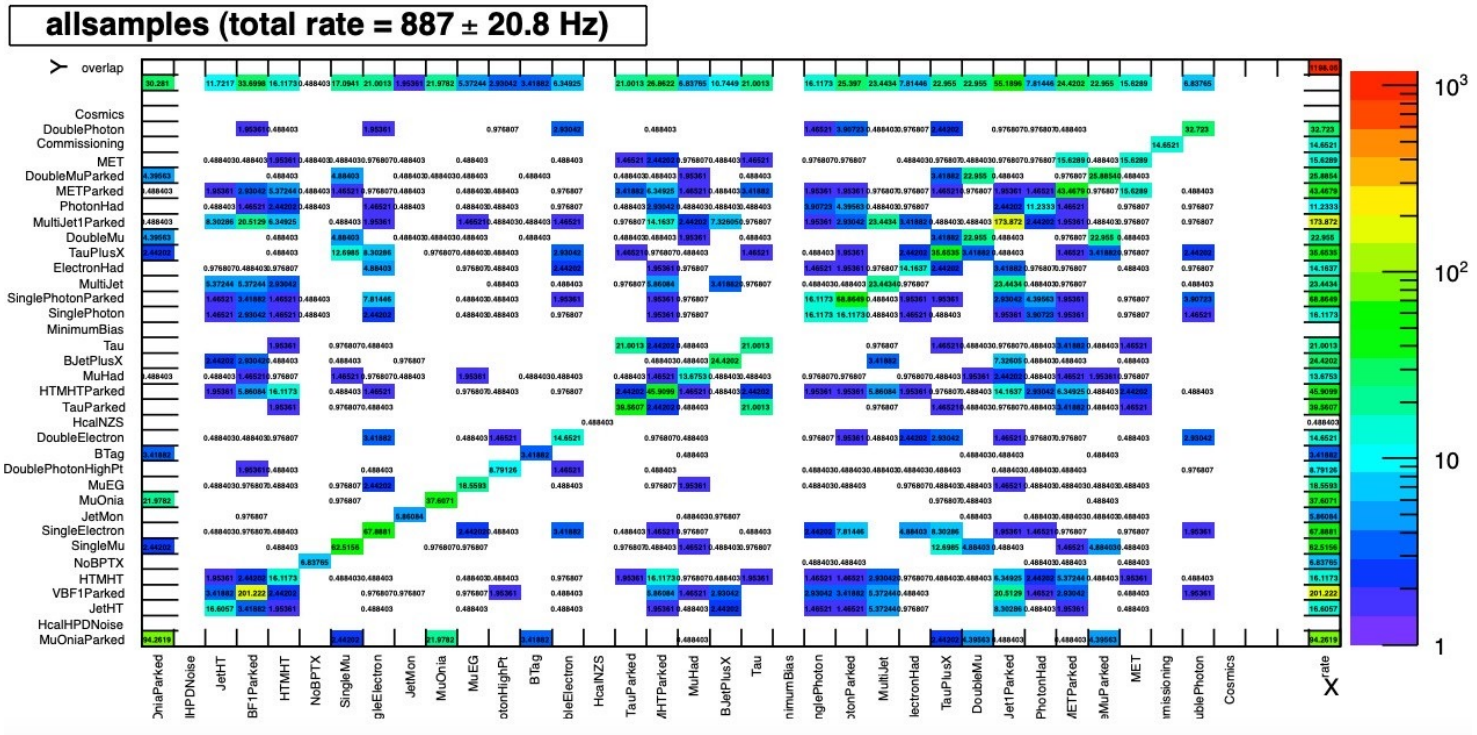
- **Offline calibration:**

  use alignment/calibration dataset and prompt-reco physics datasets

  to be used by End-of-Year (End-of-data taking period) re-reconstruction

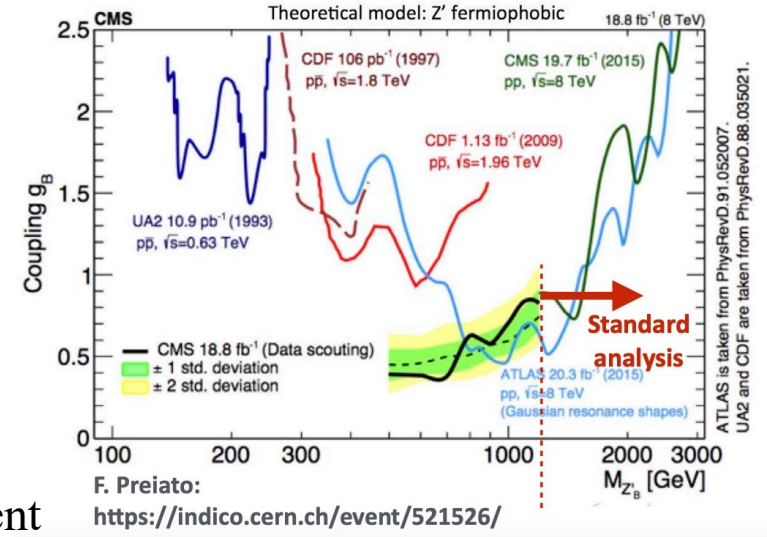
Primary datasets

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

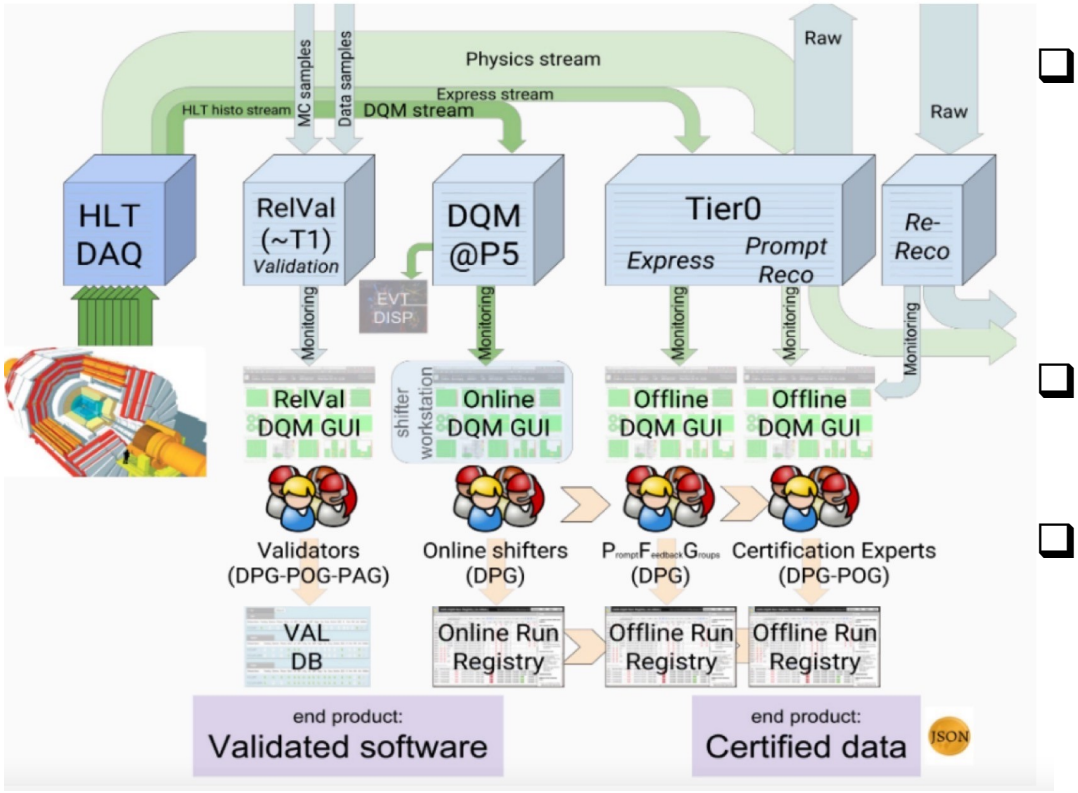


Data scouting and parking

- 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:

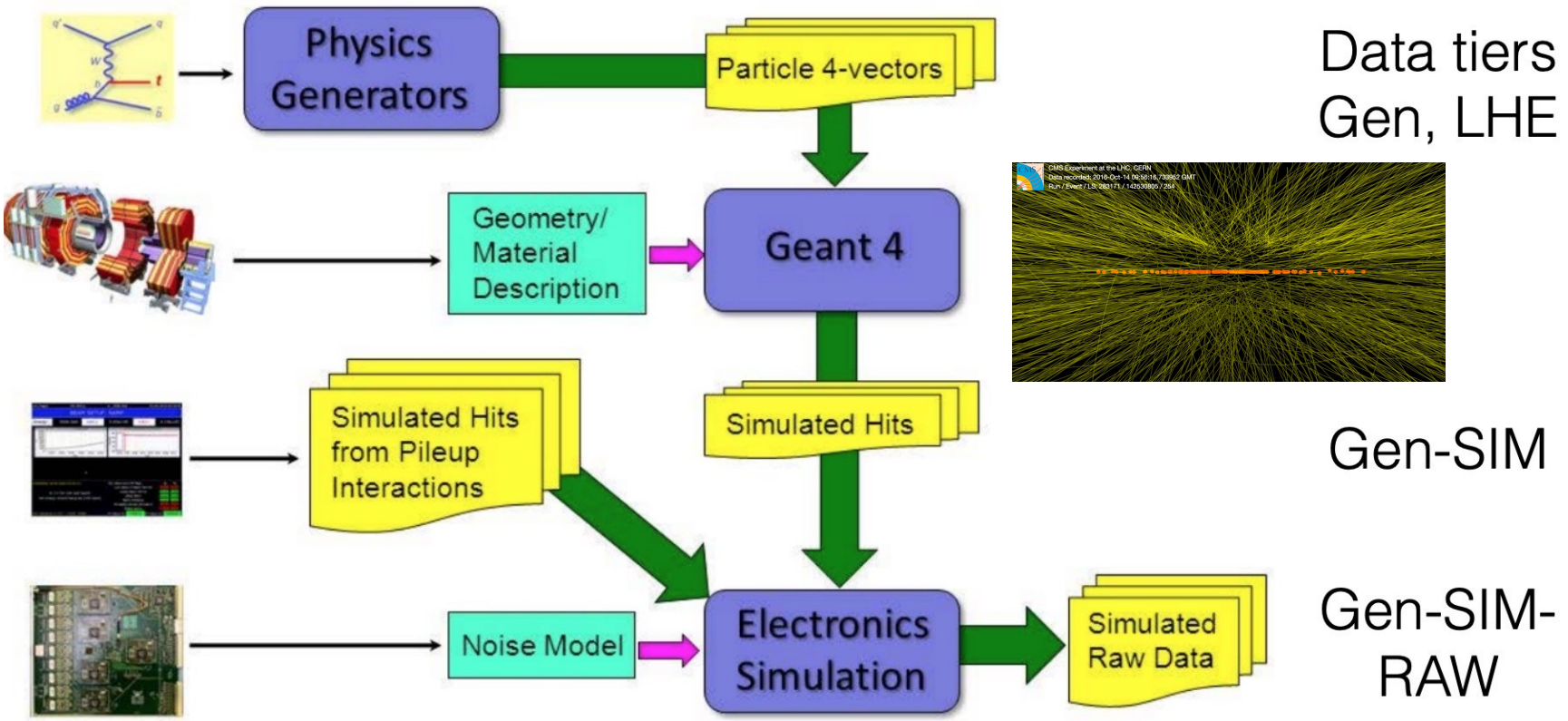
  provide central data certification –
  good runs/lumi sections to be used for most of the physics analyses

- 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”.

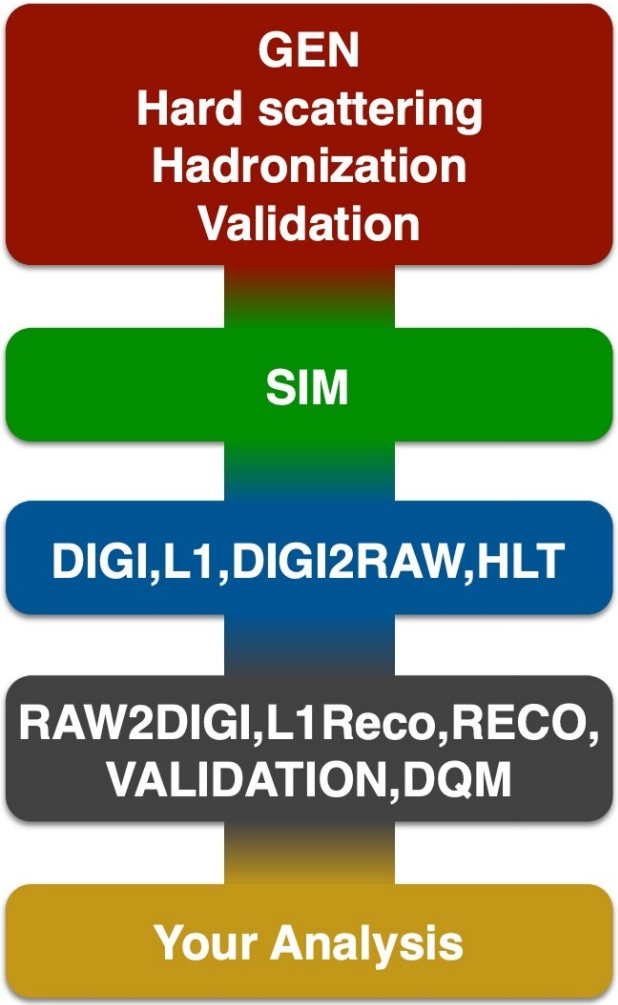
  https://cms-service-dqmdc.web.cern.ch/CAF/certification/

Event simulation (Monte Carlo)

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



From data/MC to your (physics) analyses



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

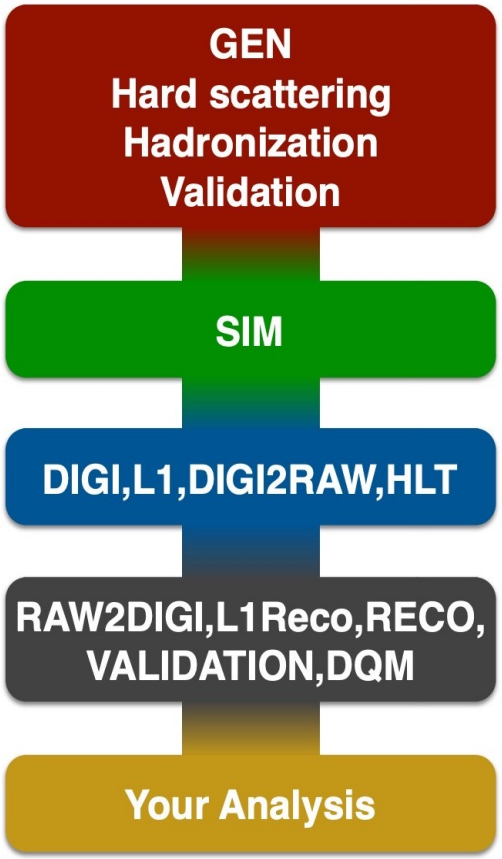
CMS Geometry, Magnetic field,...

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,...

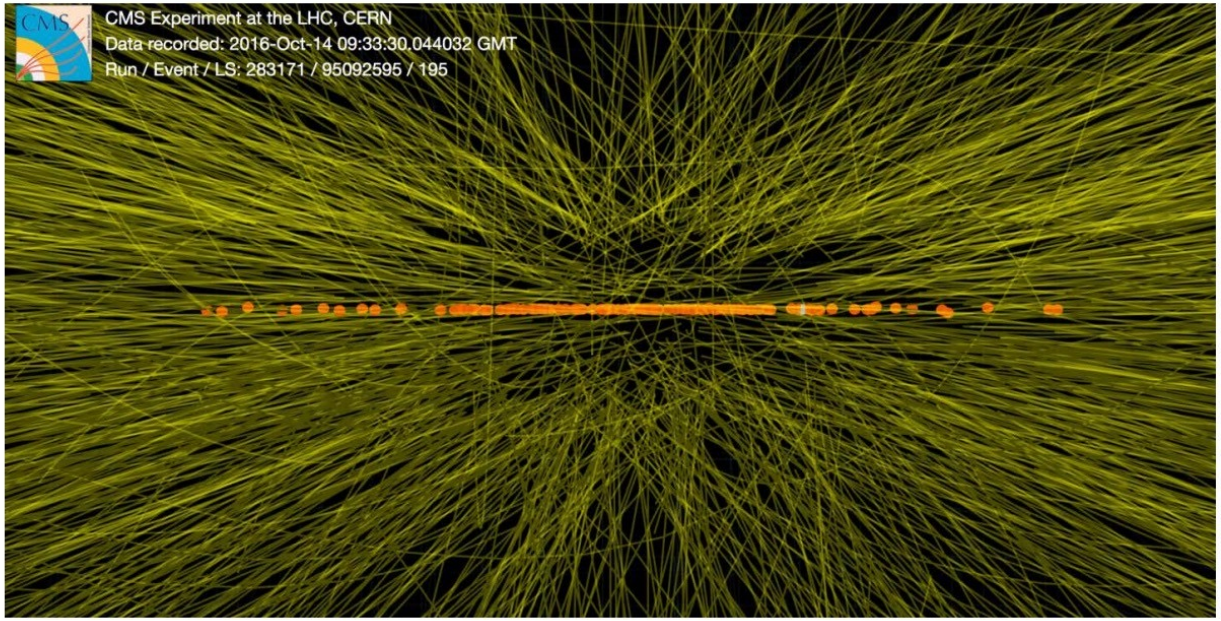
Detector electronics simulation, Pileup situation,
Alignment-Calibration, Trigger menu,...

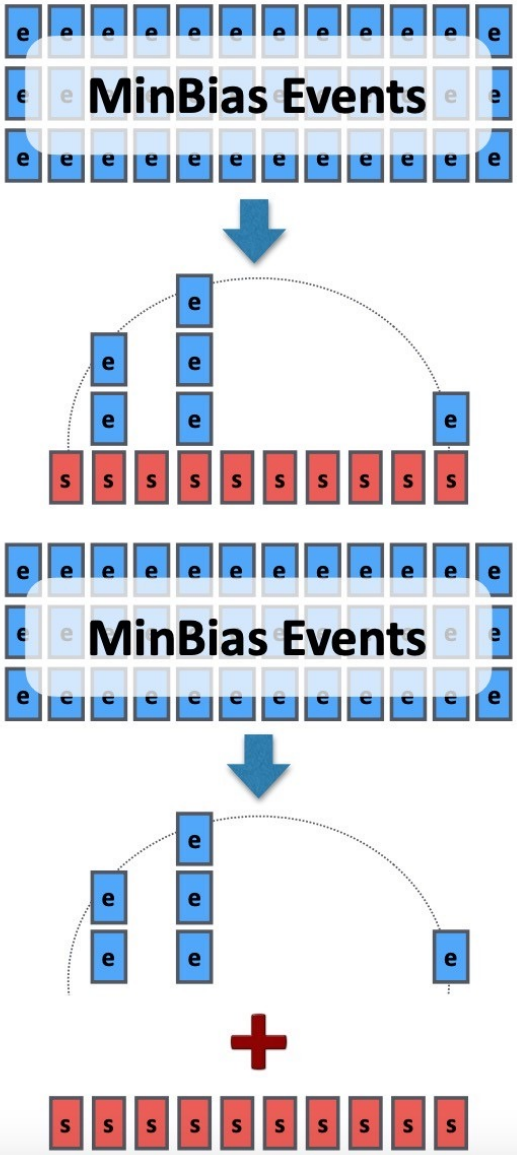
Reconstruction algorithms,
High level object creation

Your analysis code

Beam spot (profile) is an importance information for simulation :
its position and σ_x σ_y σ_z

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





Classic mixing

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

Pre-mixing

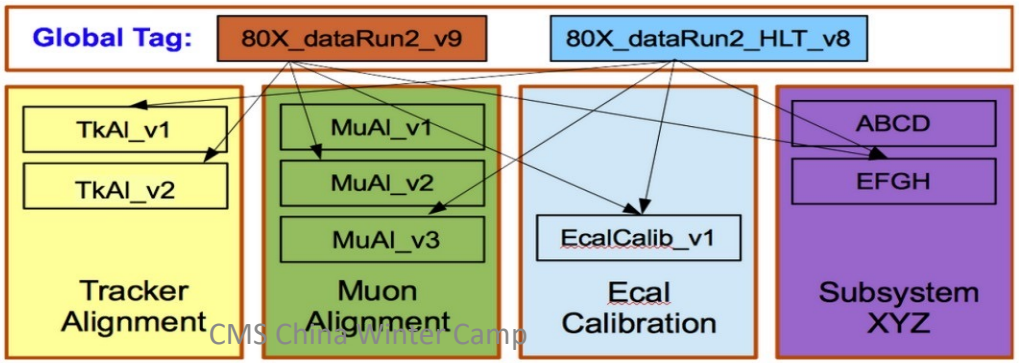
- MinBias events in RAWSIM format are overlaid on empty single neutrino events using a chosen pileup configuration. Digis made in this step are converted to RAW.
- 1-1 combination of PreMixed event - signal event. RawToDigi is done on-the-fly to premixed events before overlay.

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 validation relies on a small scale data/MC production 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 : PdmV info

https://twiki.cern.ch/twiki/bin/viewauth/CMS/PdmVRun3Analysis

2022 Analysis Summary Table

PPD suggests a quick summary for analyses in the following table: slide

DATA

Eras	Datasets	JIRAs	GTs	Comments
2022 FG Prompt	Prompt PDs: see section below	-	124X_dataRun3_PromptAnalysis_v2	New JECs (CMSTalk post)
2022 ABCDE ReReco	ReReco PDs: see section below	PDMVRERECO-55	124X_dataRun3_v15	GT for 2022ABCDE ReReco (CMSTalk post)
2022 ABCD Prompt	Prompt PDs: see section below	-	124X_dataRun3_PromptAnalysis_v1	GT identical to Prompt GT but with updated JEC/JR corrections for 2022ABCD (CMSTalk post)

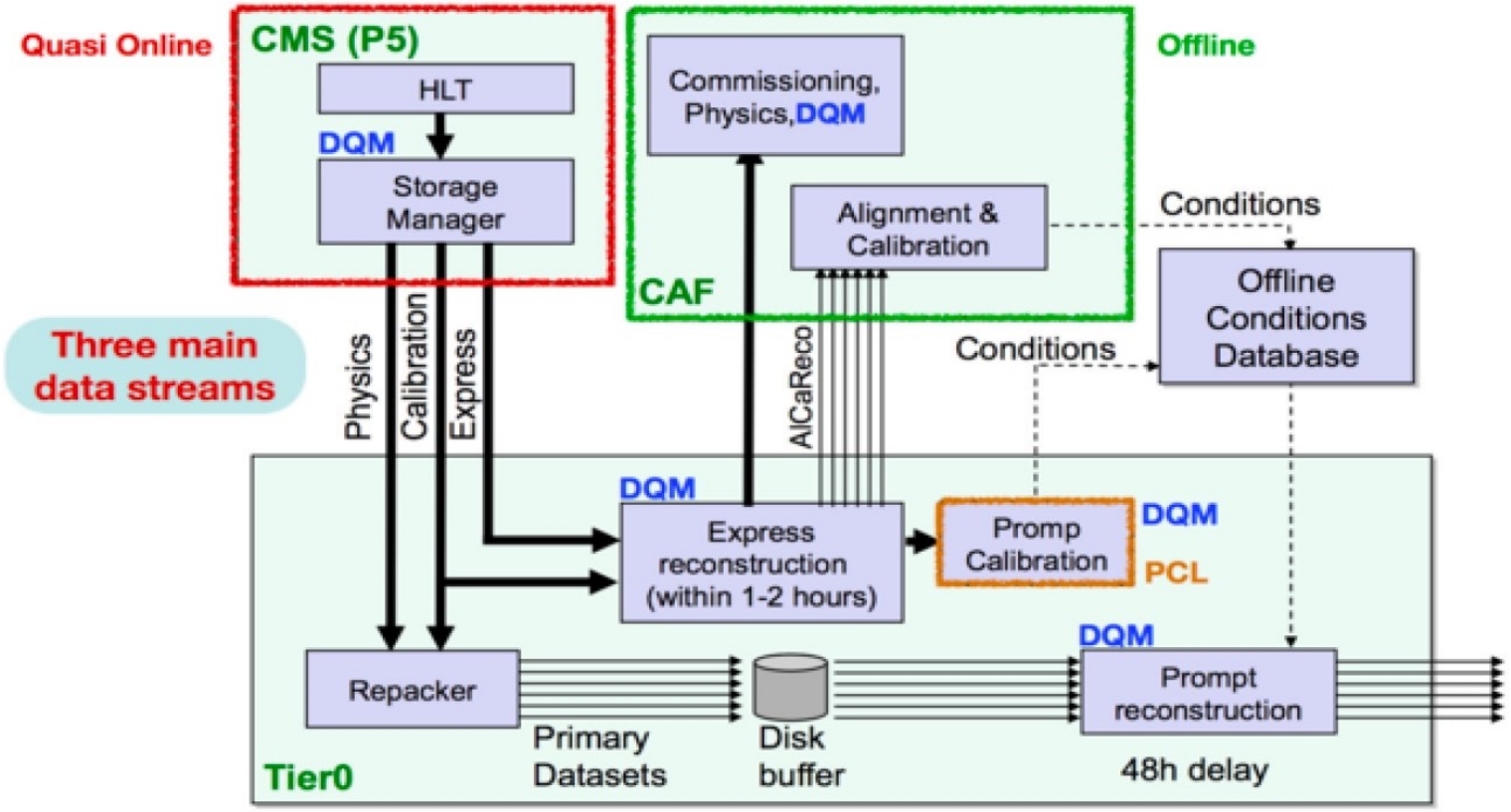
without normtag

ERA	Delivered by LHC [/fb]	Recorded by CMS [/fb]	Golden JSON [/fb]	Monte Carlo Campaign
B	0.1287	0.1154	0.0964	
C	6.924	6.301	4.953	Run3Summer22
D	3.745	3.323	2.922	Run3Summer22
E	6.592	6.117	5.672	Run3Summer22EE
F	19.963	18.423	17.610	Run3Summer22EE
G	3.588	3.247	3.055	Run3Summer22EE
Total	40.9407	37.5264	34.3084	

with normtag /cvmfs/cms-bril.cern.ch/cms-lumi-pog/Normtags/normtag_BRIL.json

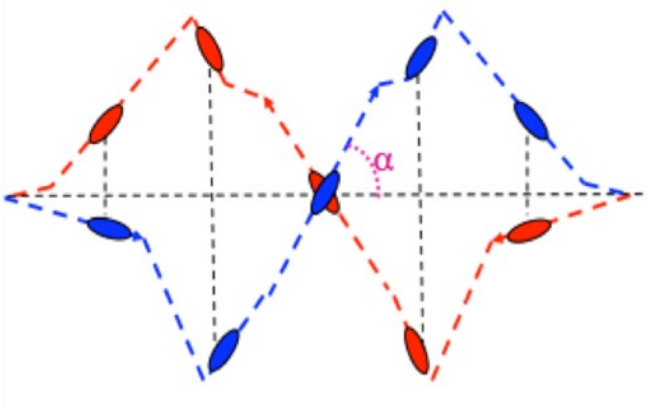
ERA	Delivered by LHC [/fb]	Recorded by CMS [/fb]	Golden JSON [/fb]	Monte Carlo Campaign
B	0.1292	0.1161	0.09768	
C	7.0909	6.4543	5.0707	Run3Summer22
D	3.850	3.4177	3.0063	Run3Summer22
E	6.8207	6.3304	5.8783	Run3Summer22EE
F	20.4134	18.8402	18.0070	Run3Summer22EE
G	3.6644	3.3163	3.1219	Run3Summer22EE
Total	41.9686	38.475	35.18188	

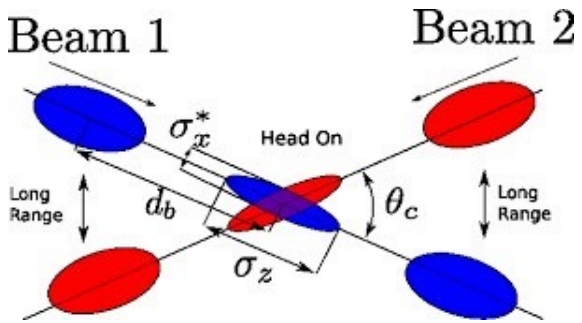
Summary

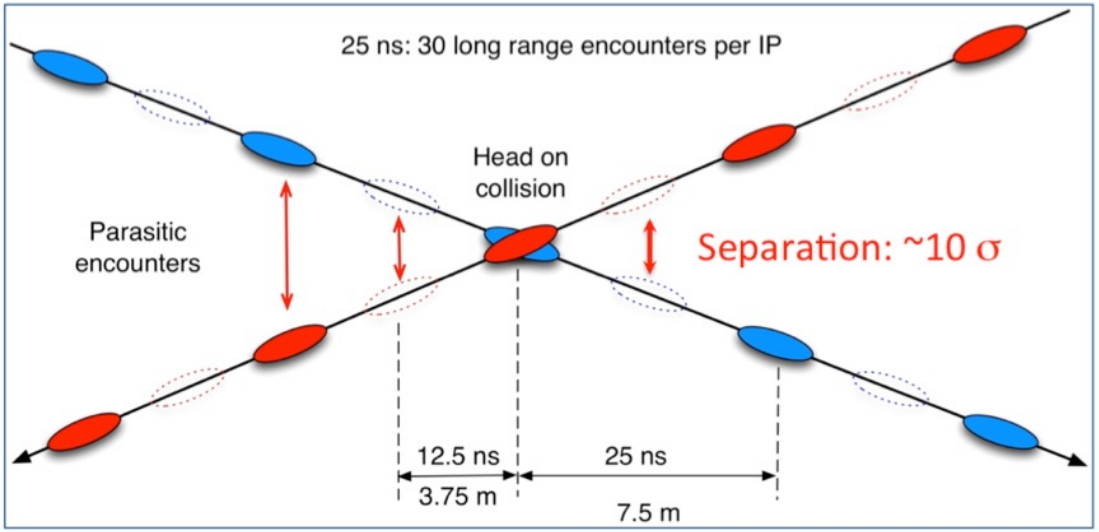

Quasi Online
CMS (P5)
HLT
DQM
Storage Manager
Commissioning, Physics, DQM
Offline
Alignment & Calibration
Conditions
CAF
Offline Conditions Database
Conditions
Three main data streams
Physics
Calibration
Express
AlCaReco
DQM
Express reconstruction (within 1-2 hours)
Promp Calibration
DQM
PCL
DQM
Repacker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

Backup

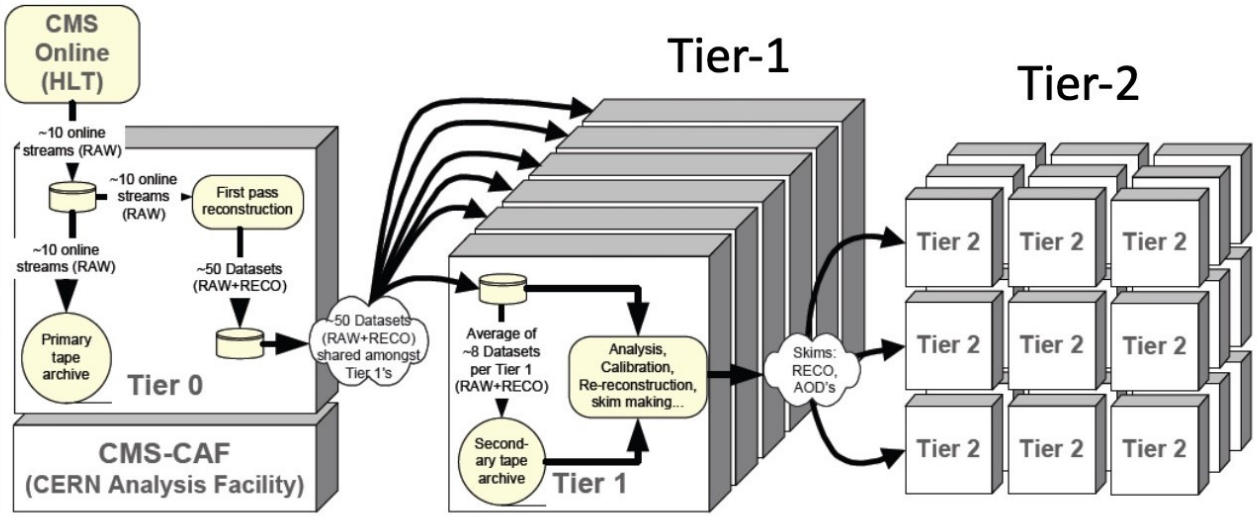
Proton collisions at the CMS

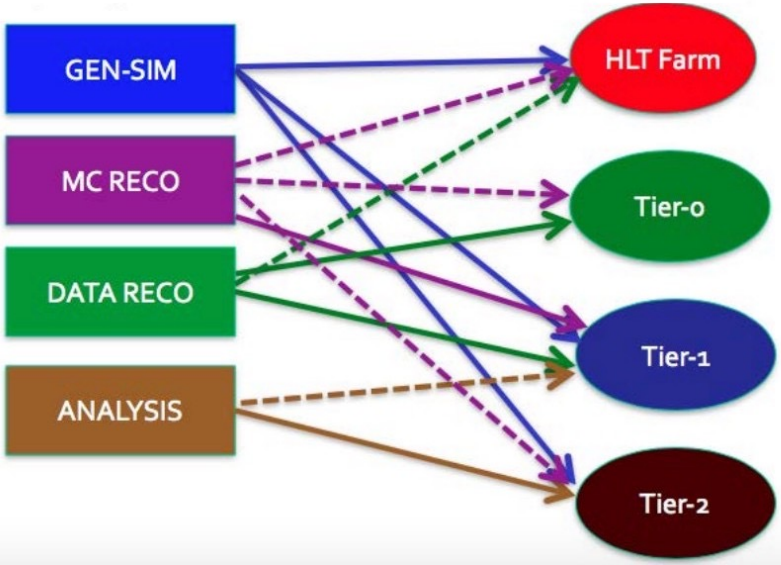






https://home.cern/news/news/accelerators/lhc-report-playing-angles





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

Run-1 Run-2

Examples of what are in RAW/RECO/AOD