From detector building to physics publication: the real story of the data

PART II

Patrizia Azzi - INFN Padova

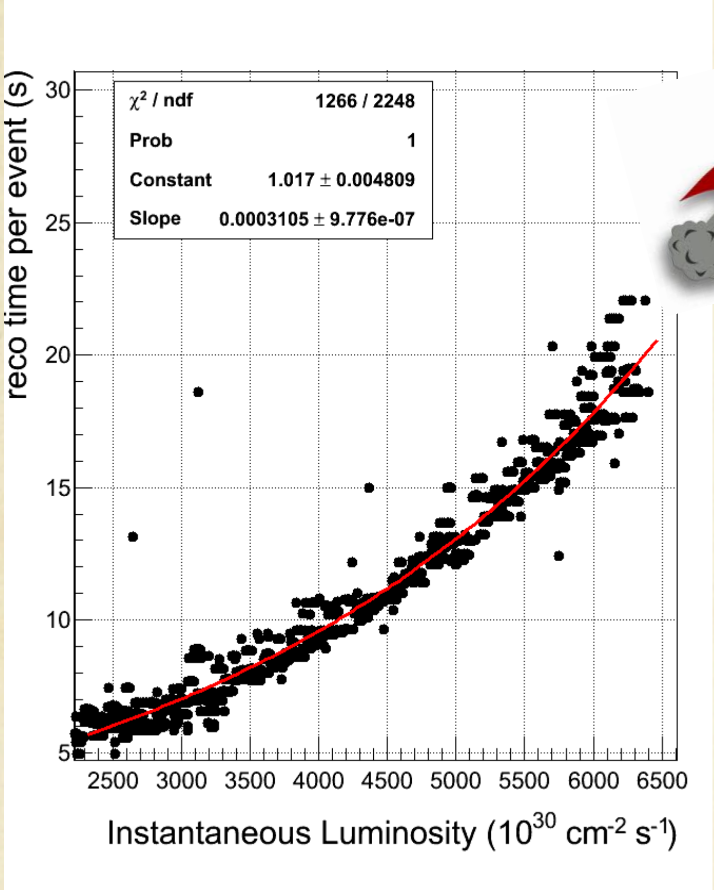
Outline

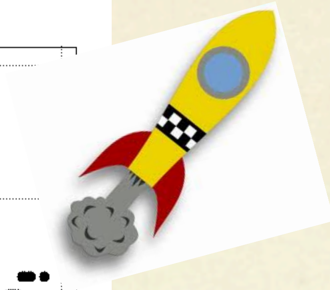
- Life during operations: get the physics out as fast as possible (really really fast)
  - Taking data
  - Calibrate your data
  - Certify your data
- Life during a shutdown: prepare for next data taking
  - Improve your simulation
  - Improve your reconstruction
  - Improve your computing

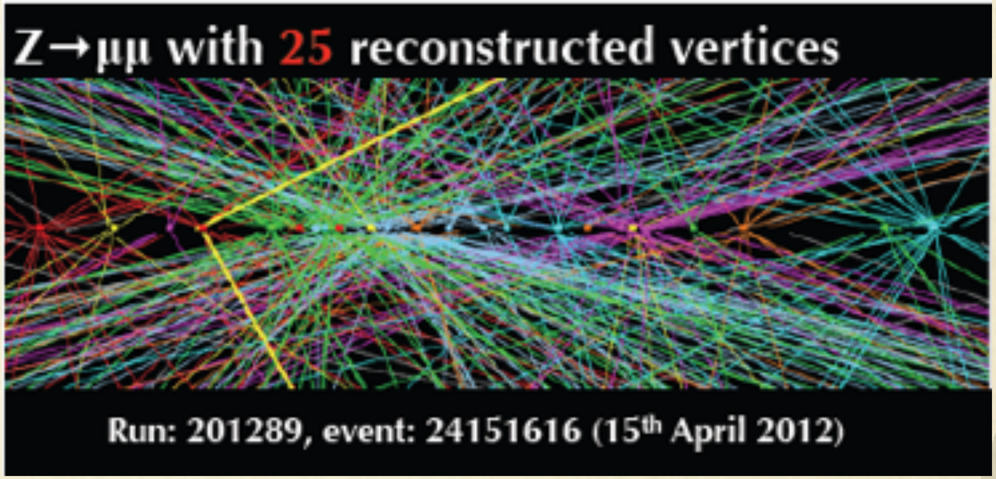
Life during a shutdown

Preparing for the next data taking

Challenges of the RunII

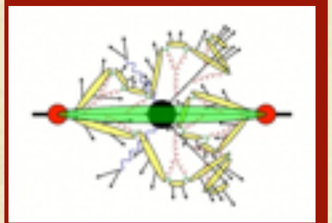






Higher Energy - Higher Luminosity - Higher Pile-Up

The data flow

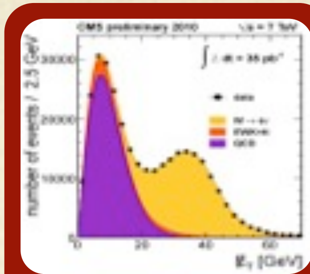


Monte Carlo Event Generator

Simulation of the Detector Response

Trigger

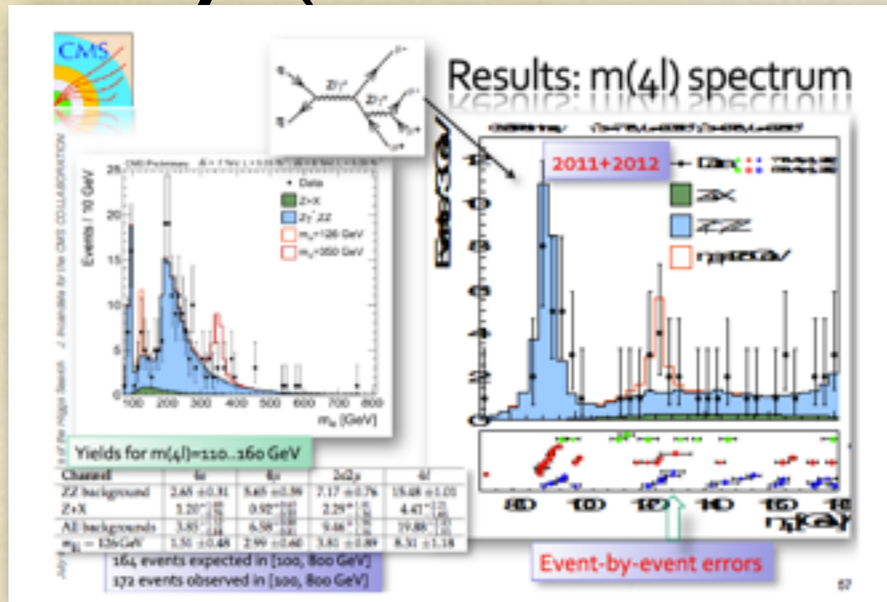
Event Reco.

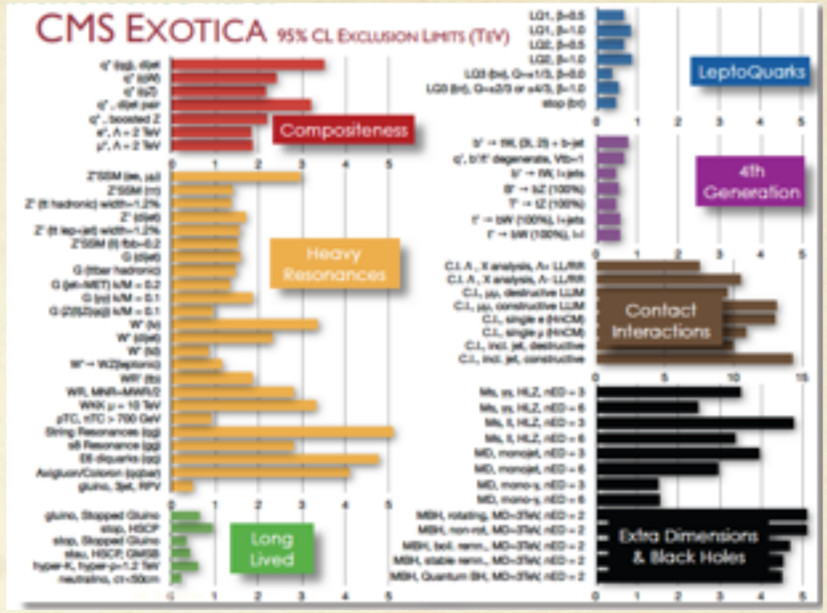


Physics Analysis

IMPROVING THE SIMULATION

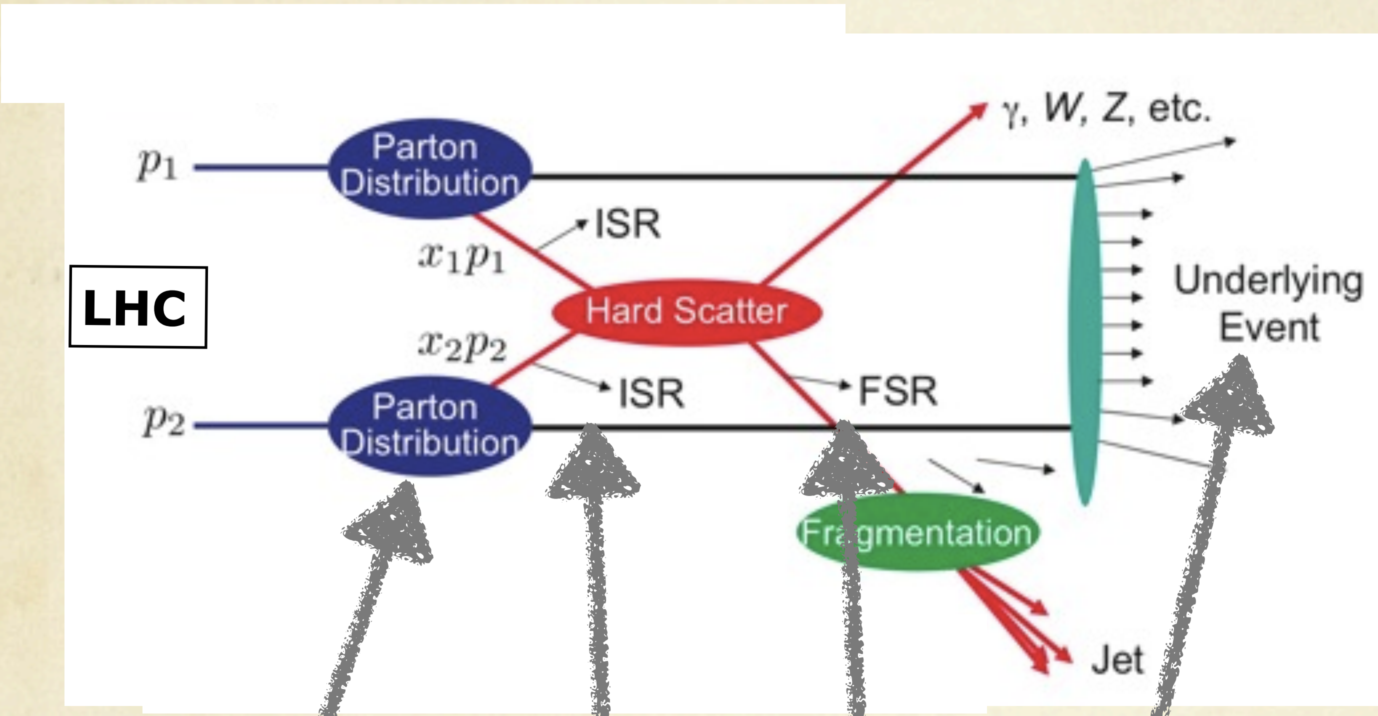
Why (do we need) MC generators





- **MOSTLY TO BE PREPARED!** To study signal that we have not seen yet. To develop strategies to cope with different energies/data taking conditions. To develop new detectors. Also to understand our data better...
  - We can: calculate **inclusive cross-sections**
  - We can: calculate **differential cross sections** as a function of variables of interest in the analysis
- They make the connection between a theoretical model and reality: **simulated events** can then be treated in the same way as real data.

Structure of LHC events



add the underlying event

let hadrons decay

hadronize partons

add the parton showers

add initial and final state radiations

generate hard process

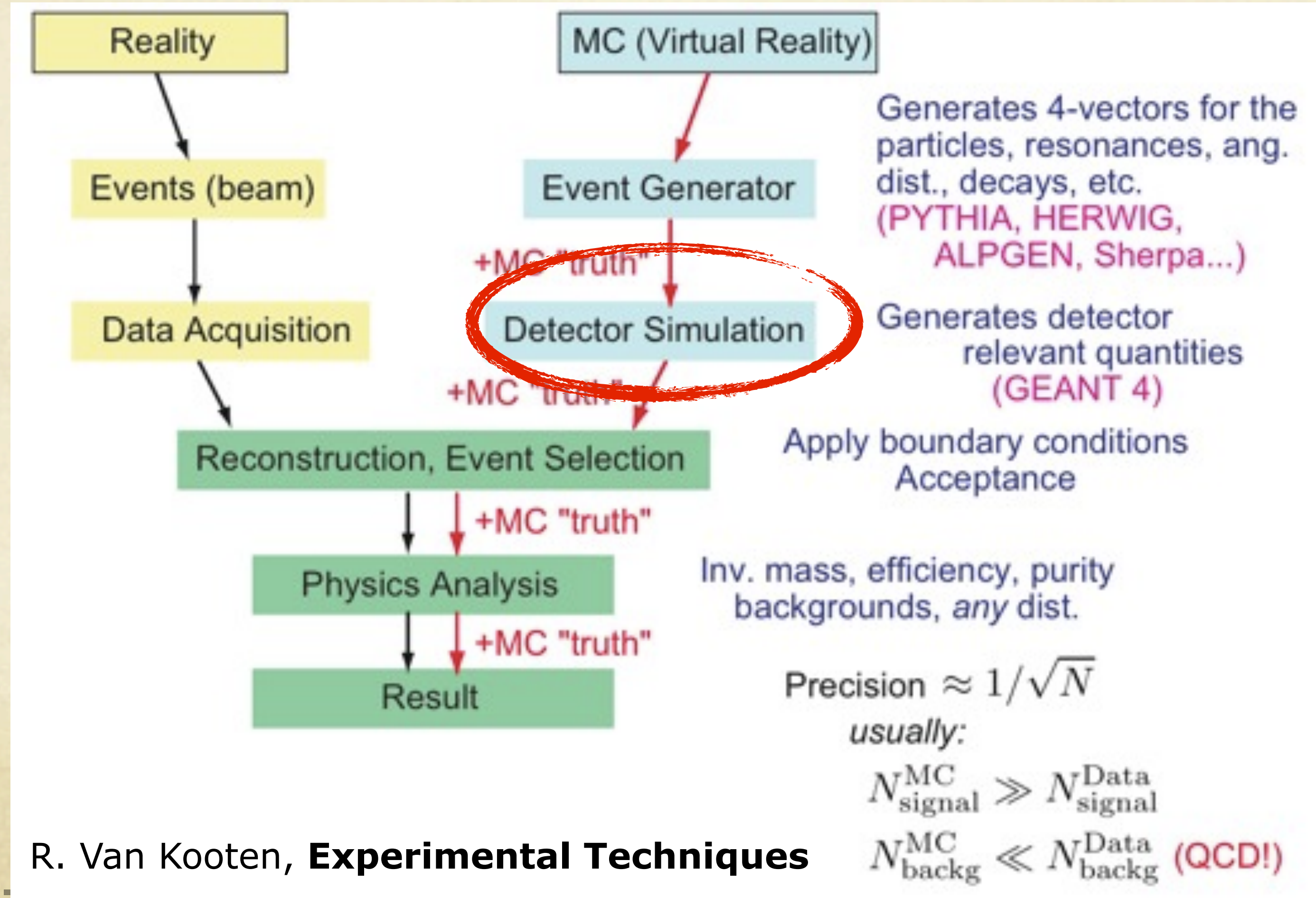
activity due to the proton remnants

parton probability distribution in the proton

radiation in the process

What happens after...

- the simulation is the next step in the production of MC samples: the generator level particles are fed to a program that will simulate the dector response
- The goal is that the simulated data can be treated exactly as the real data with the truth information available for use



R. Van Kooten, **Experimental Techniques**

How do we use the MonteCarlo datasets

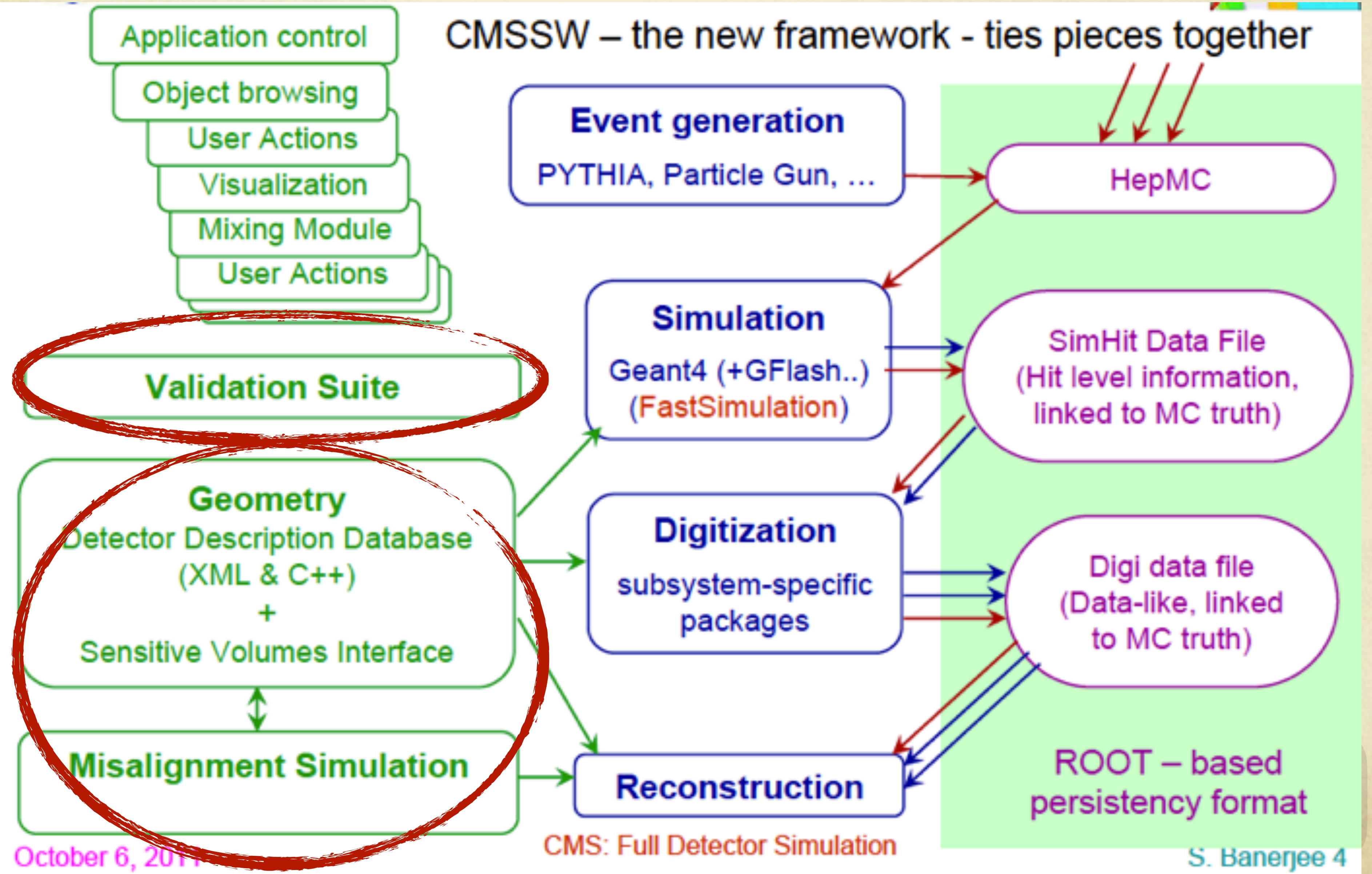
$$\sigma = \frac{N_{obs} - N_{bkg}}{\varepsilon \cdot \int \mathcal{L}dt}$$

The cross section of our favorite process

$$\varepsilon = \varepsilon_{tr} \cdot \varepsilon_{reco} \cdot \varepsilon_{ID} \cdot \varepsilon_{sel}$$

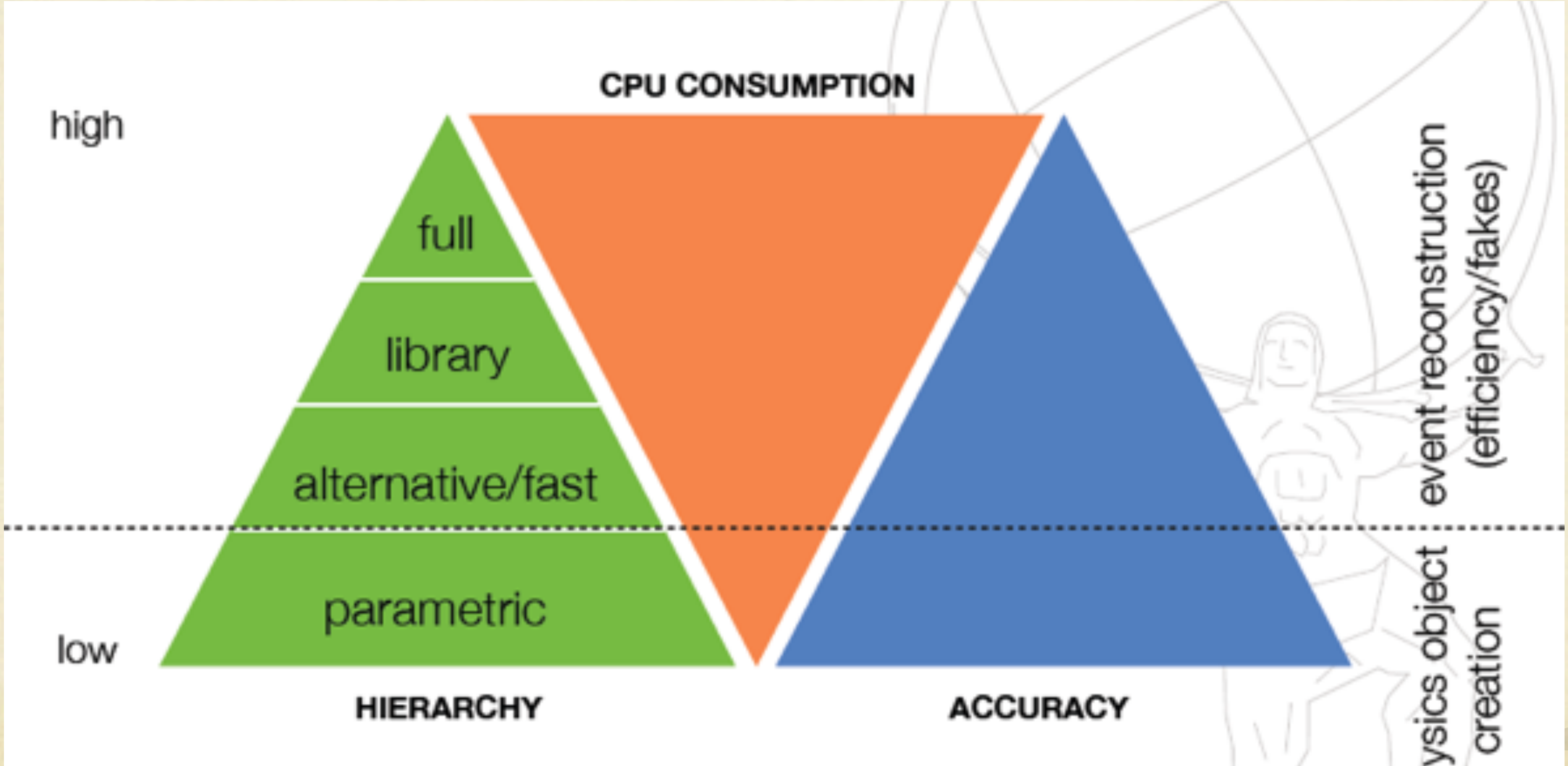
- (Assuming we have passed the events through the simulation of the detector)
- calculate what fraction of events from a given decay **falls within the detector acceptance and the selections** of the analysis
- need a **forecast of how the event develops** in space, after the interaction
- the **simulations** are necessary both for known physics objects (Z, W, top production) and even more to build searches for new physics
- the **uncertainties** in the input parameters of the model (Q^2, PDF, ISR, FSR...) are sources of systematics.

the CMS detailed implementation for FullSim



Choices, choices, choices

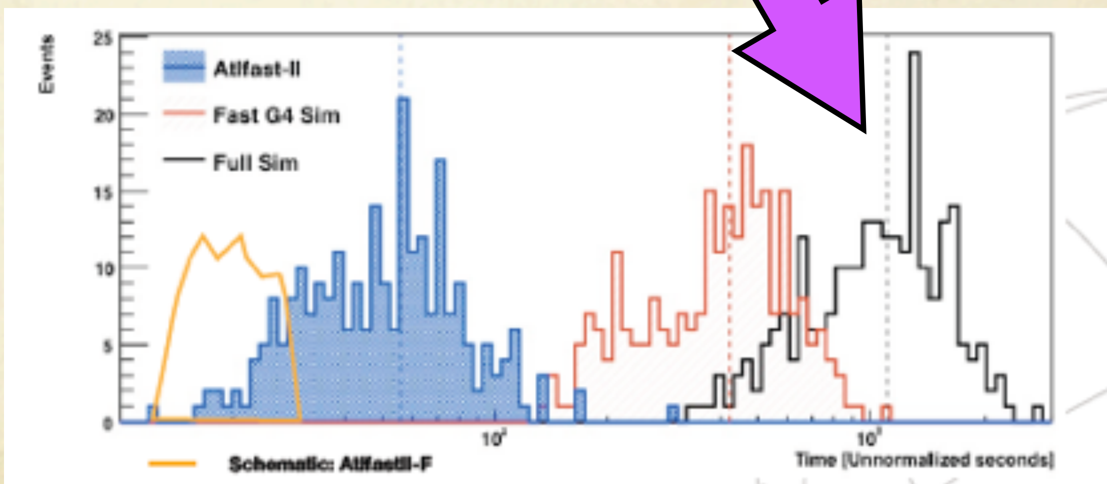
- Every experiment develops one (or more) simulations of their detector response.
- Optimization of which simulation to use for the various purposes is a complex strategy excercise that touches also Reconstruction, and Computing



NOTE: Here will not discuss toy or parametric simulation

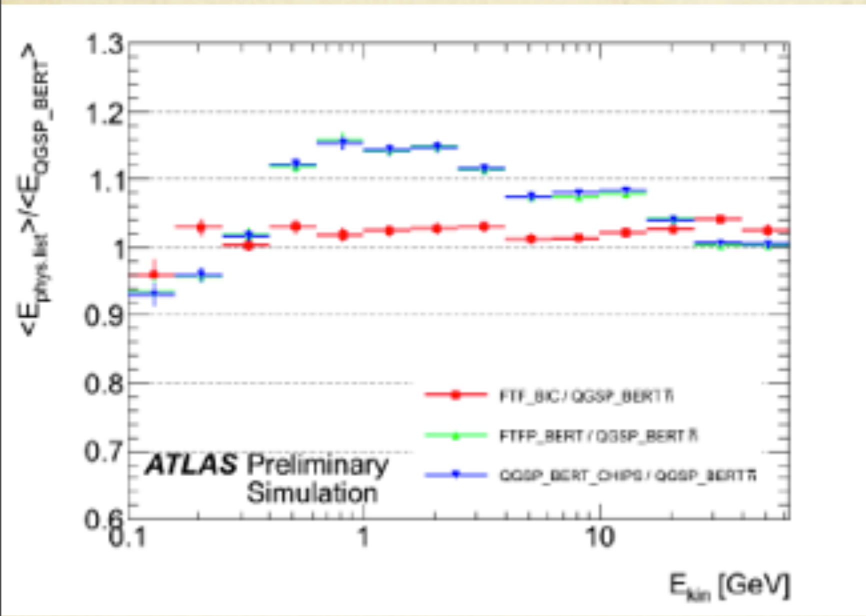
the levels of simulation: Full

- Full Simulation is based on GEANT4, a very detailed and sophisticated physics response engine.
- the GEANT program uses generator output (4-vectors) and simulates the **interaction of particles within the detector volume** (need a good description of the geometry):
  - particle ionization in trackers
  - energy deposition in calorimeters
  - intermediate particle decays/radiation
- Many handles to tune it:
  - geometry/material description of the detector
  - physics lists
  - step lenght
  - process cuts
- very slow, very precise

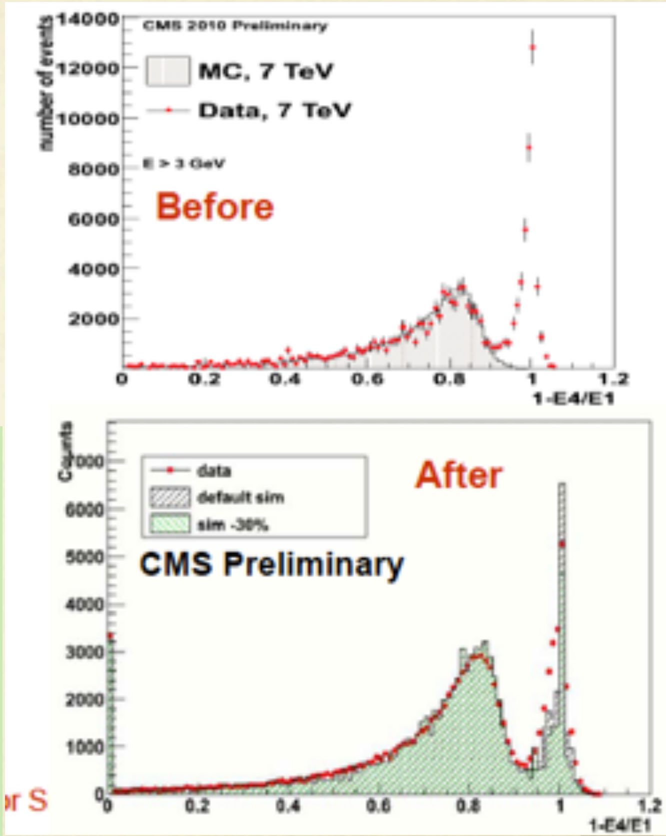


COMPARISON OF PROCESSING TIMES

Tuning GEANT simulation

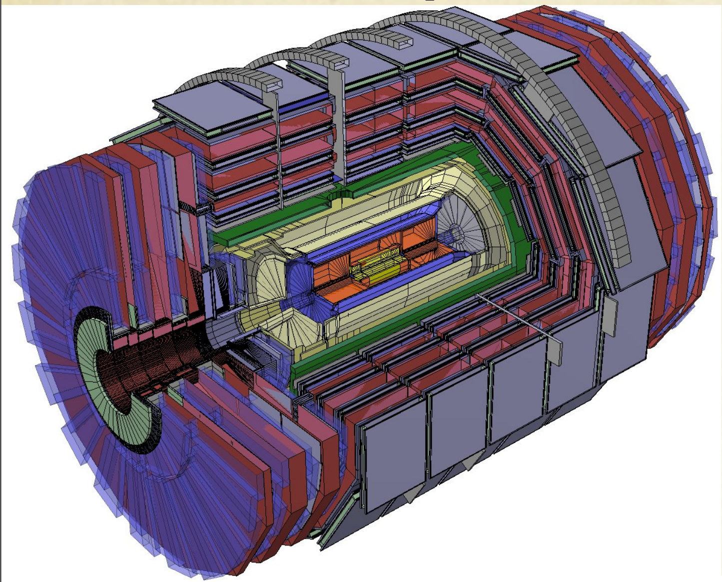


EFFECT OF CHANGING THE PHYSICS LIST
ATLAS calorimeter response to anti-neutrons
(0.1<pT<50GeV)

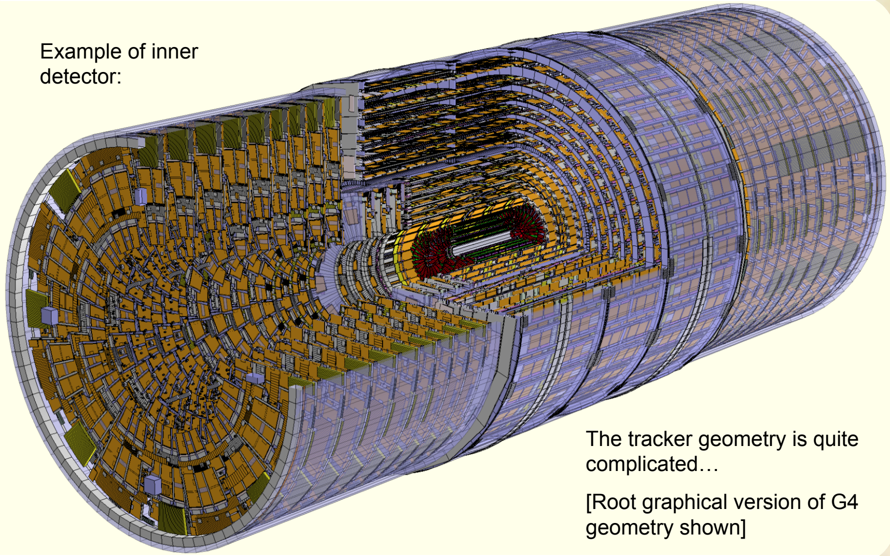


EFFECT OF CHANGING THE GEOMETRY/
MATERIAL
Anomalous hits observed in CMS Calorimeter data.
Origin traced to energy deposit in thin layer of
Silicon
in the APDs. «After» shows the result after
introducing
in the simulation geometry extra layers
of Silicon in the APDs as sensitive detectors

Extremely accurate geometry description needed



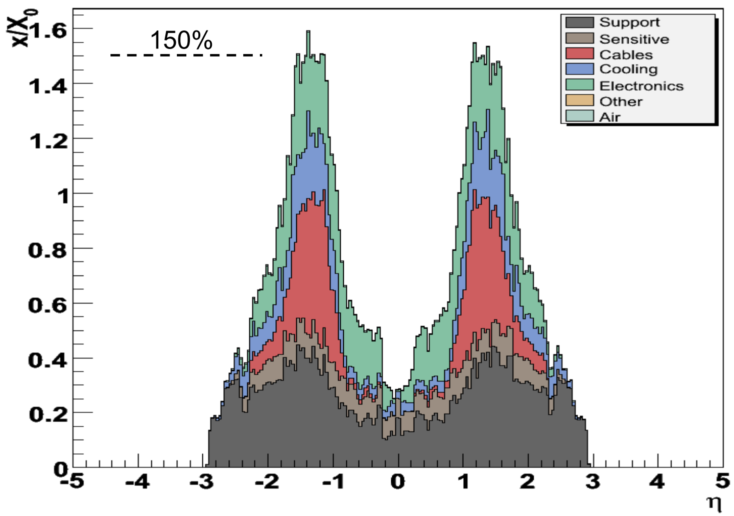
Example of inner detector:



The tracker geometry is quite complicated…

[Root graphical version of G4 geometry shown]

Material distribution in current CMS Tracker (estimated):



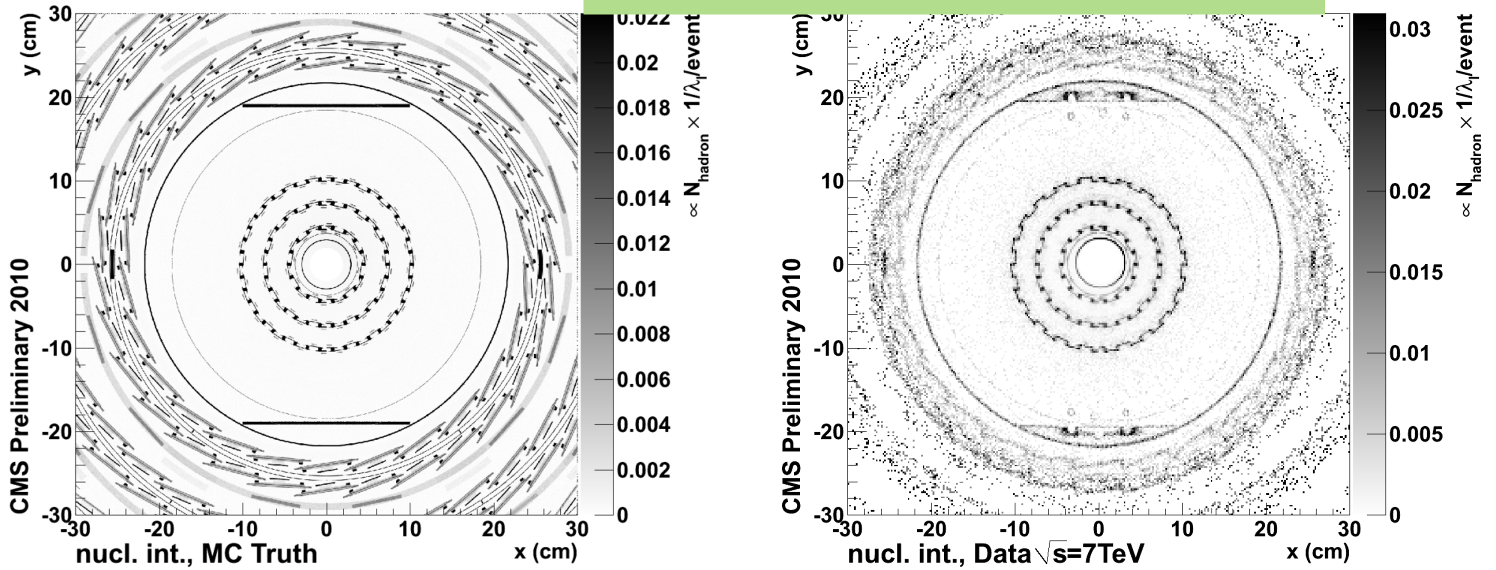
- Very large photon conversion probability
- large effects of multiple-scattering

} potentially large physics consequences

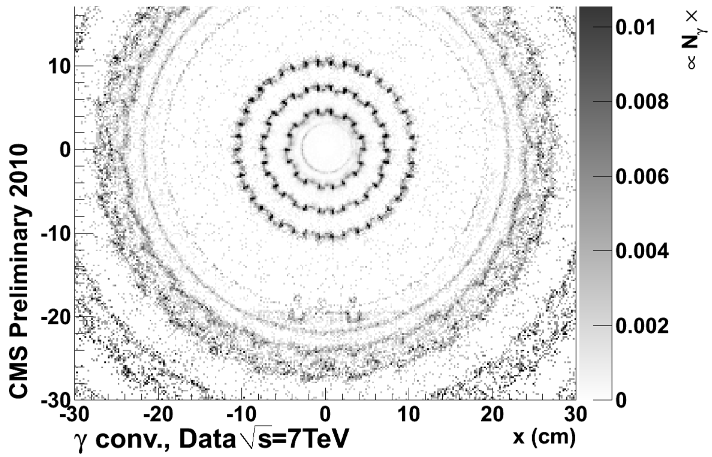
At the end of the day must always validate against Data!!!

Validate the tracker geometry with data

NUCLEAR INTERACTION DATA



Note the superior position resolution of the nuclear interaction data



The beampipe isn’t centered!

CMS PAS: TRK-10-003

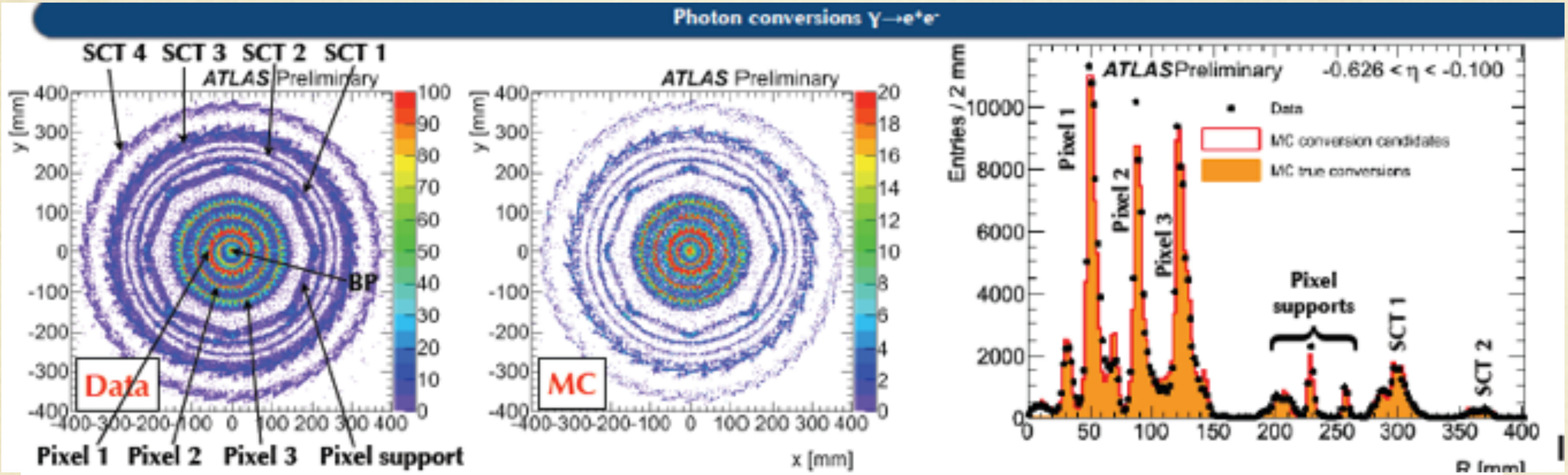
(8) *20 October, 2010*

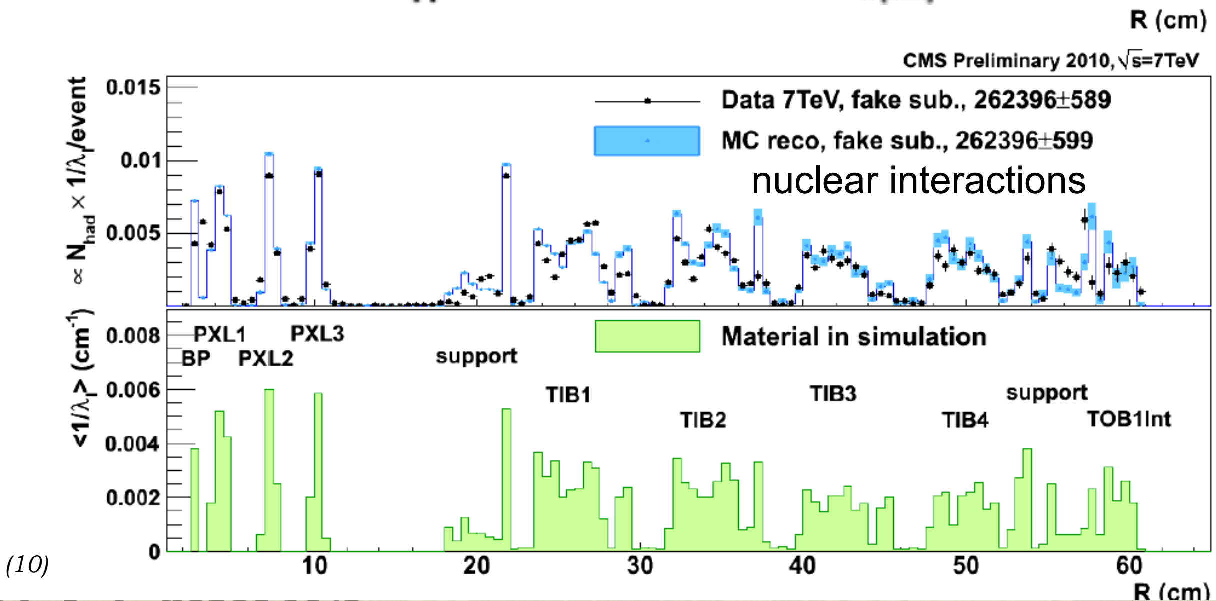
UNIVERSITY OF NOTRE DAME

Fermilab

PHOTON CONVERSION DATA

Validating the tracker material budget





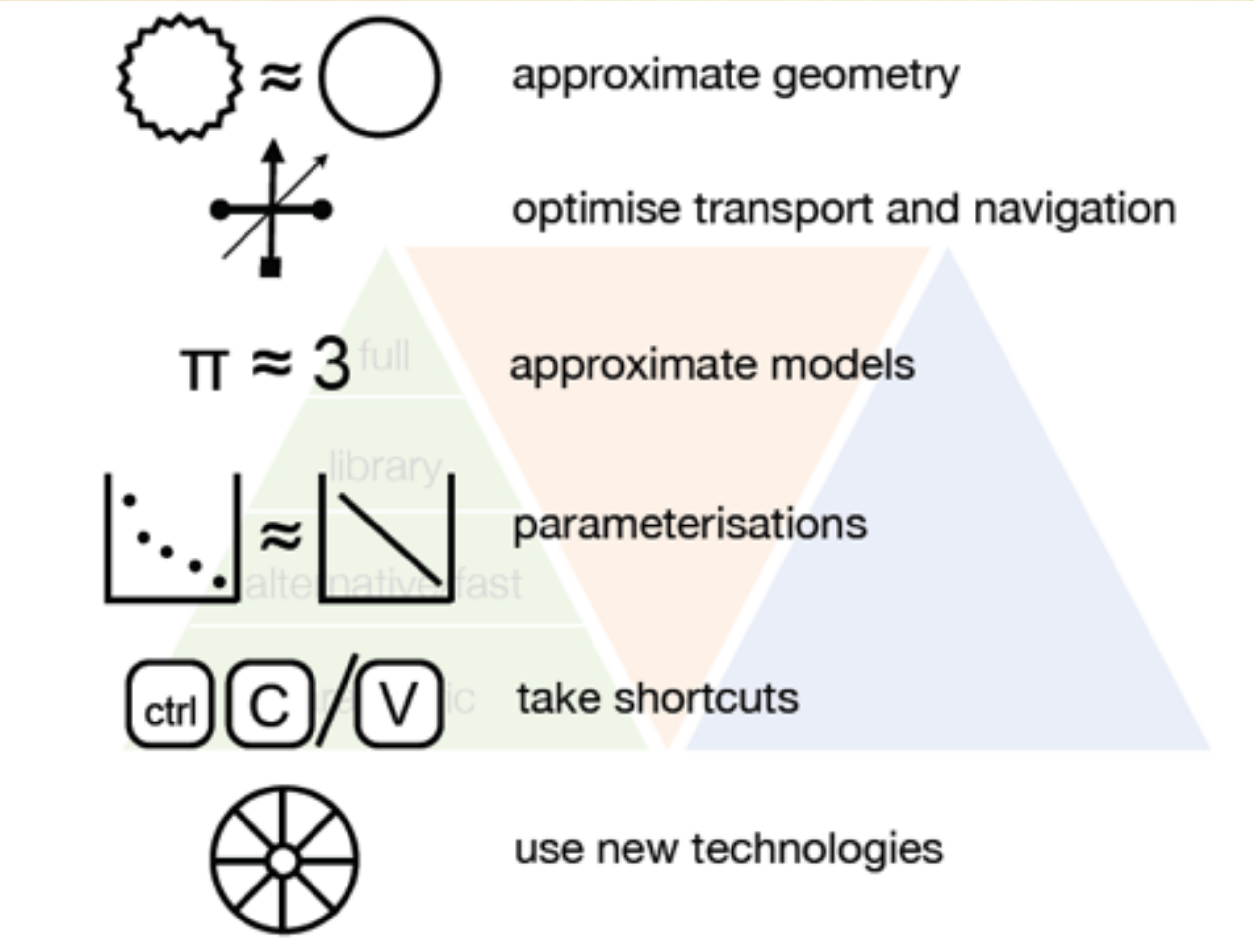
(10)

astonishingly good agreement between data and simulation

UNIVERSITY OF NOTRE DAME
Fermilab

levels of the simulation: Fast

- Several ways to speed up the simulation. All options tried

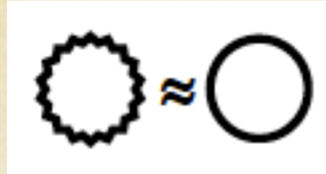


A variety of versions and names:
CMS: FastSim
ATLAS: AFII&AFIIF
ISF

why do we need a fast(er) simulation?

- Because we need very large amounts of MonteCarlo **(---> more later on MonteCarlo Production campaigns)**
  - to evaluate background with large cross section and small survival probability.
    - Filtering directly at RECO level is more efficient and less biased
  - To scan a model's parameter space of evaluate systematics
  - to train MVAs with sufficient statistics
  - to develop and test efficiently reconstruction and analysis algorithms
  - to study/test new geometries and conditions
- Some example in CMS (similar for ATLAS):
  - Top Mass extraction in 2l final states JHGEP 07(2011)049(mass templates)
  - Black Hole search,PLB 697(2011)434: used for signal samples scan
  - Most of Susy analyses: simplified model signatures parameter scan

Approximate Geometry

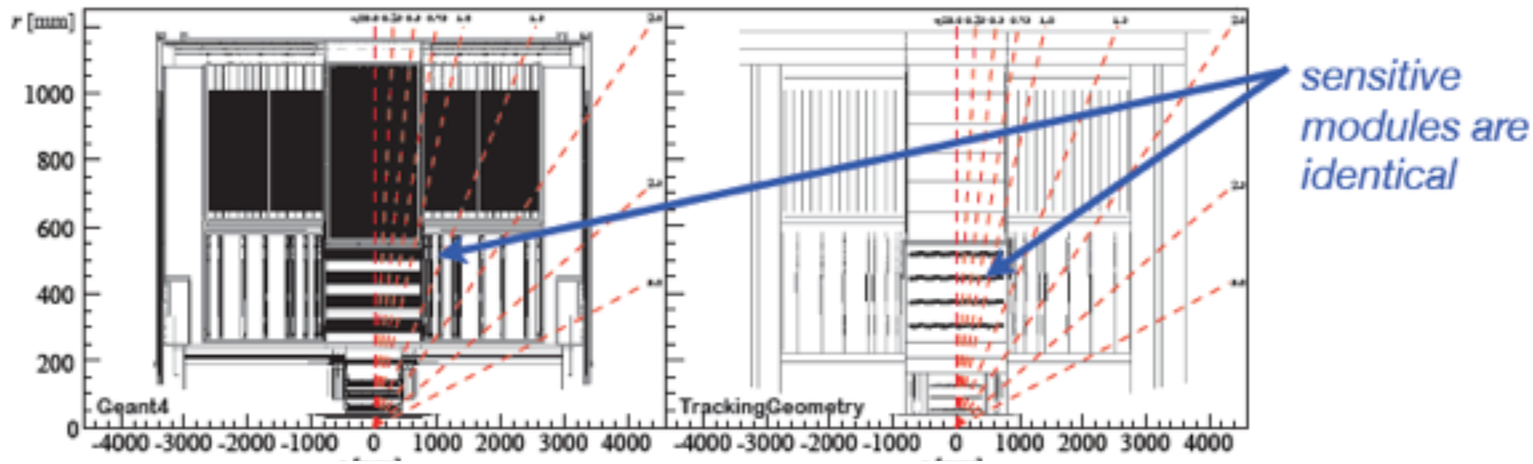


- The detectors as ATLAS and CMS have all been built with a similar philosophy: a «onion» with cylindrical simmetry and different components as a function of the radial distance from the interaction point.
- The approach of simplyfing the geometry description is then used by both: express the detailed Geant volumes in terms of layers and cylinders.
- The sensitive material is kept the same. A tuning of the inactive material (all clumped up in a few layers) is done on data.
- The navigation of the generated particles across this volumes is much faster! There is no Geant interaction to consider.

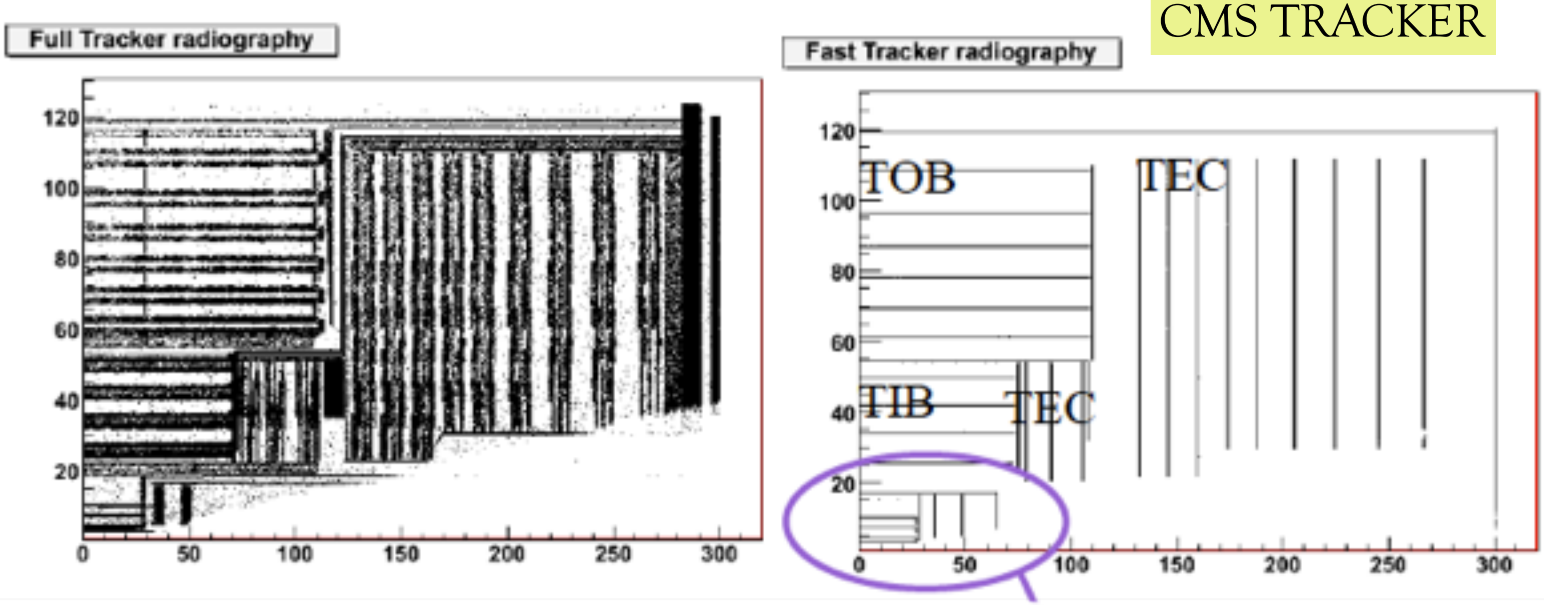
Comparing FULL and FAST geometries

- *Example Inner Detector:*
  O(100) layers and detector boundaries

ATLAS INNER DETECTOR



CMS TRACKER

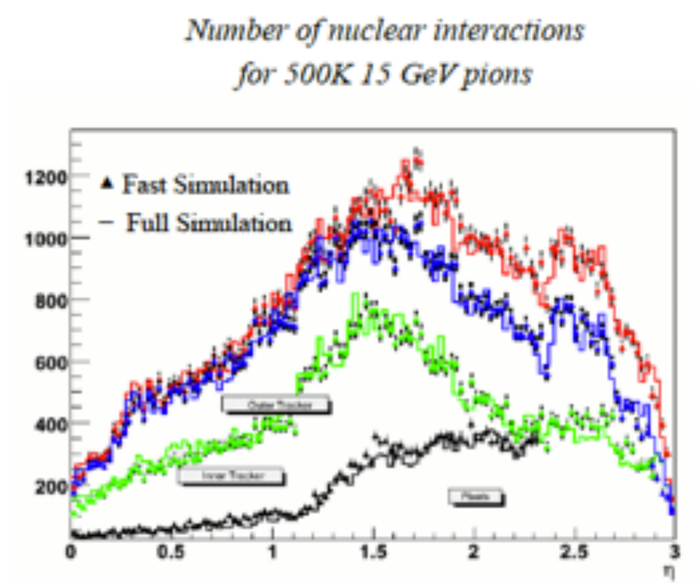


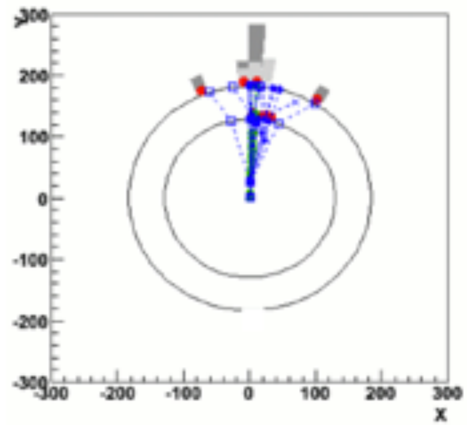
Some FastSimulation details (CMS)

- A FastSimulation is not necessarily a «simplistic» simulation. Several effects included.
- The simulated interactions are:
  1. electron Bremstrahlung
  2. Photon conversion
  3. charged particle energy loss by ionization
  4. charged particle multiple scattering
  5. nuclear interactions
  6. electron, photon and hadron showering
- The first 5 are applied to particles crossing the silicon tracker, while the latter is parameterized in the electromagnetic and hadron calorimeters. Muons propagate through the tracker, calorimeters and muon chambers with multiple scattering and energy loss by ionization taken into account in the propagation
- Very important note: all the calibrations, conditions, dead channels, noise, misalignement can be applied to the FastSim as in the FullSim.

Example: Nuclear Interactions in FastSim

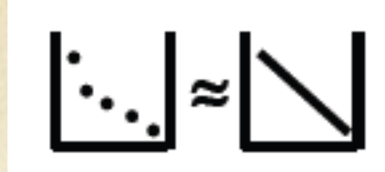
- Having the possibility to properly simulate the number of daughter particles, their angle of emission, and their momentum is very important for the accurate description of tracking efficiency for instance. (ATLAS has it too!)
- Data files of N.I. (2.5M) have been created for 9 different hadrons, 1<E<1000GeV
- when a N.I. occurs a particle is picked at random from these files in the relevant energy range





A single tau event where a pion undergoes a nuclear interaction in the tracker (fast simulation)

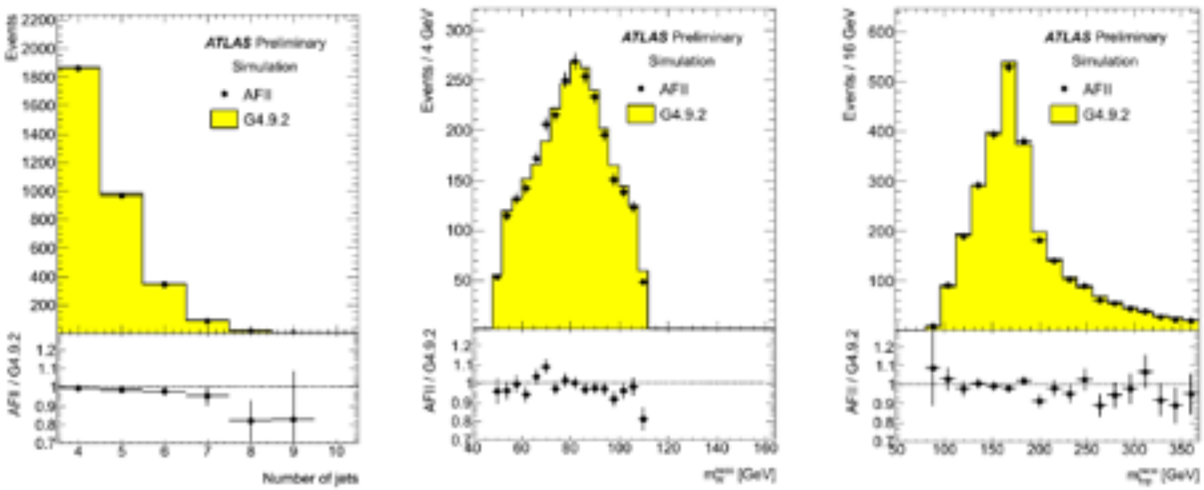
FastSimulation of the Calorimeters



- The faster simulation of the calorimeter response is based on shower parameterization and tuning (GFLASH or similar)
- FastSim heavily used in production of BSM samples with many scan points or for the production of systematic samples for precision measurement

ATLFAST2 validation in context of top physics analyses

- number of jets, jet properties well reproduced
- reproduce m_W and m_{top} within statistical uncertainties



	e channel		µ channel	
	$N_{jets} \geq 4$	$N_b \geq 1$	$N_{jets} \geq 4$	$N_b \geq 1$
Full sim.	4735 ± 13	3202 ± 10	7012 ± 14	4780 ± 11
ATLFAST	4617 ± 45	3104 ± 38	7033 ± 53	4793 ± 42
Fast / Full	0.975 ± 0.010	0.969 ± 0.013	1.003 ± 0.008	1.003 ± 0.009

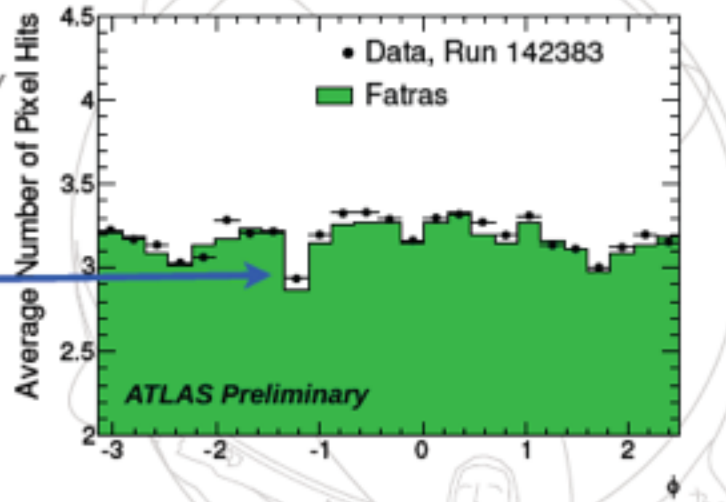
AFII jets being scaled by 1.0.1

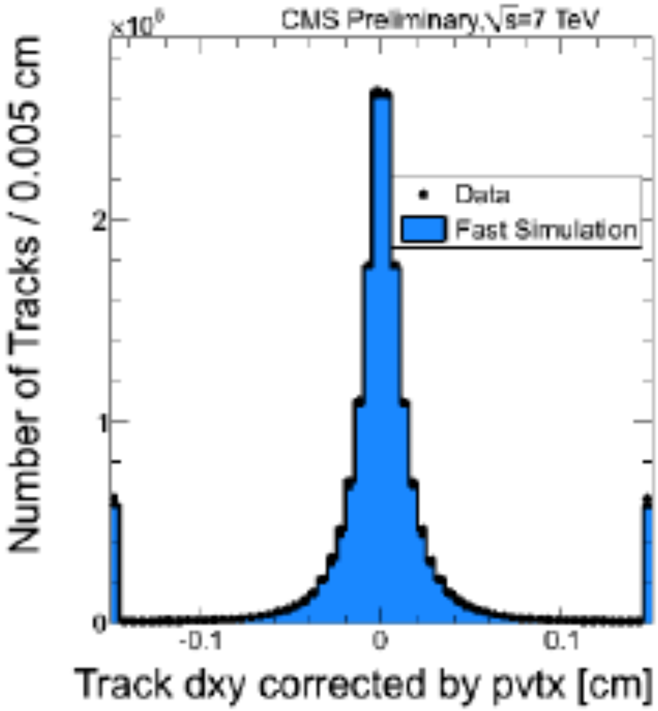
FastSimulation of the Tracking

- A better definition is «Tracking Emulation»
- Speed is achieved by skipping the slowest piece of track reconstruction code, the pattern recognition (i.e. finding the hits that belong to a track):
  - The hits are «assigned» to a track based on the MC Truth information. No possibility of «fakes tracks» exist
    - this approach works very well for high purity environement when fake rate is small. need to study what happens with large Pileup.
  - the efficiency is emulated applying the same selection cuts (seeding, quality etc) to the track parameters and checking if they are satisfied or not.
  - Performance validated with FullSim and data

FATRAS in comparison to data

- *ID reconstruction, tracks with p_T > 500 MeV*
- *using exact same sensitive detector elements:*
  conditions data being fully integrated

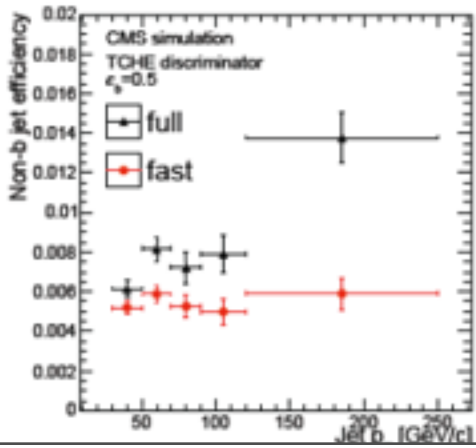




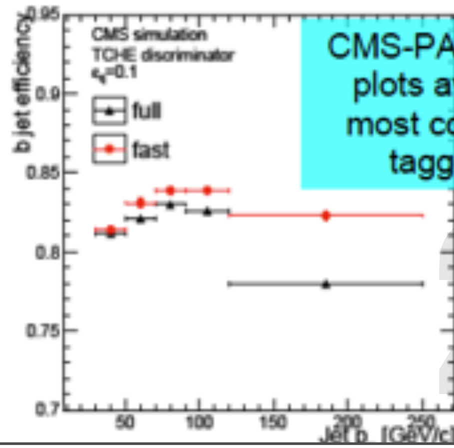
(Actually more recent plots even better)

FULL/FAST BTAGGING PERFORMANCE COMPARISON

Fake rate: generally lower



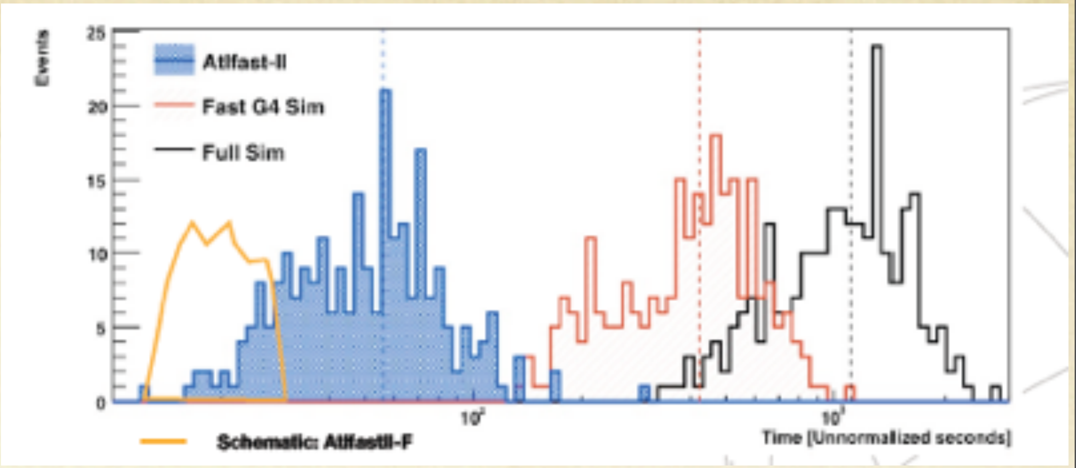
Efficiency: generally higher



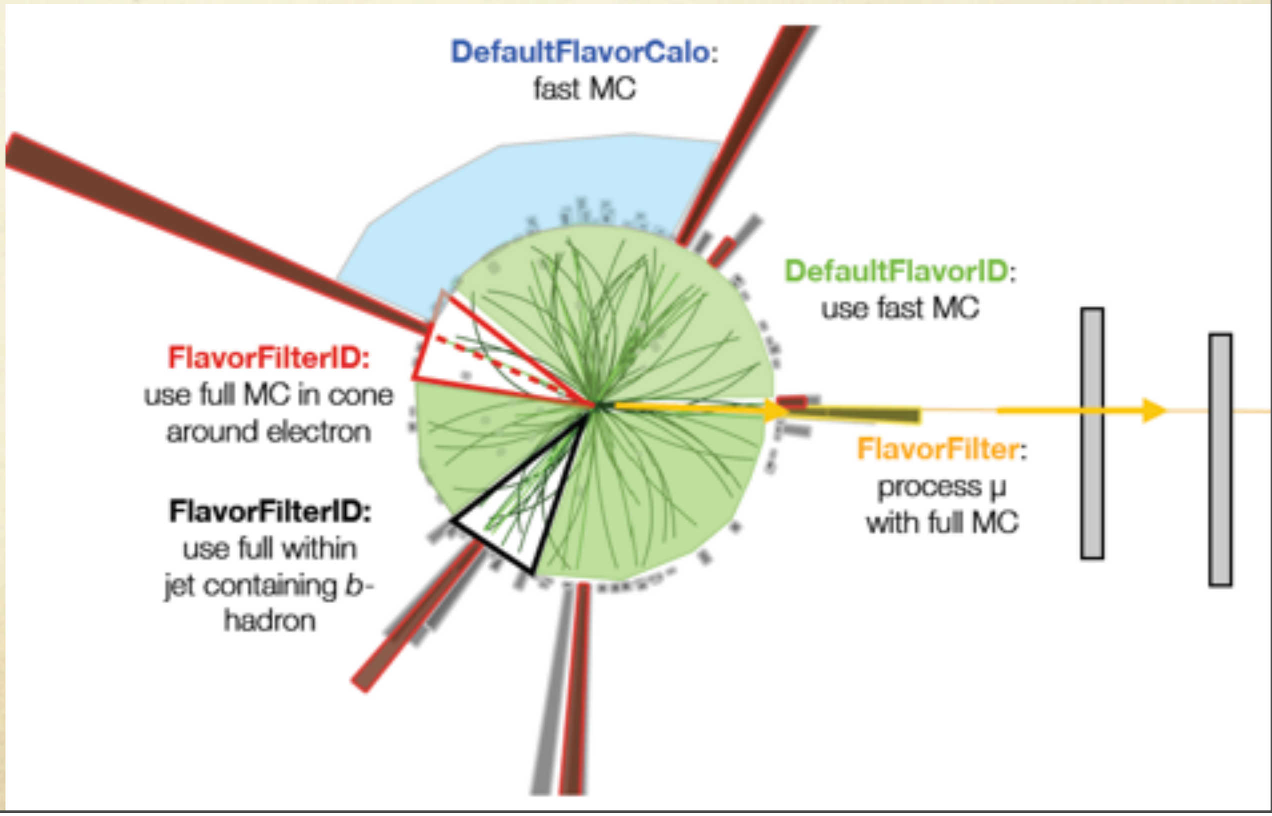
CMS-PAS-BTV-11-002 plots available for all most commonly used taggers in CMS

what's next?

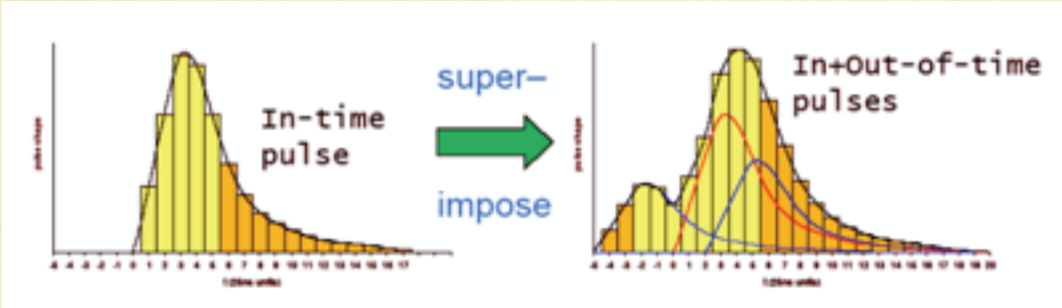
- big challenge for the future RunII. Ever bigger MonteCarlo productions ahead of us. One approach from ATLAS: ISF



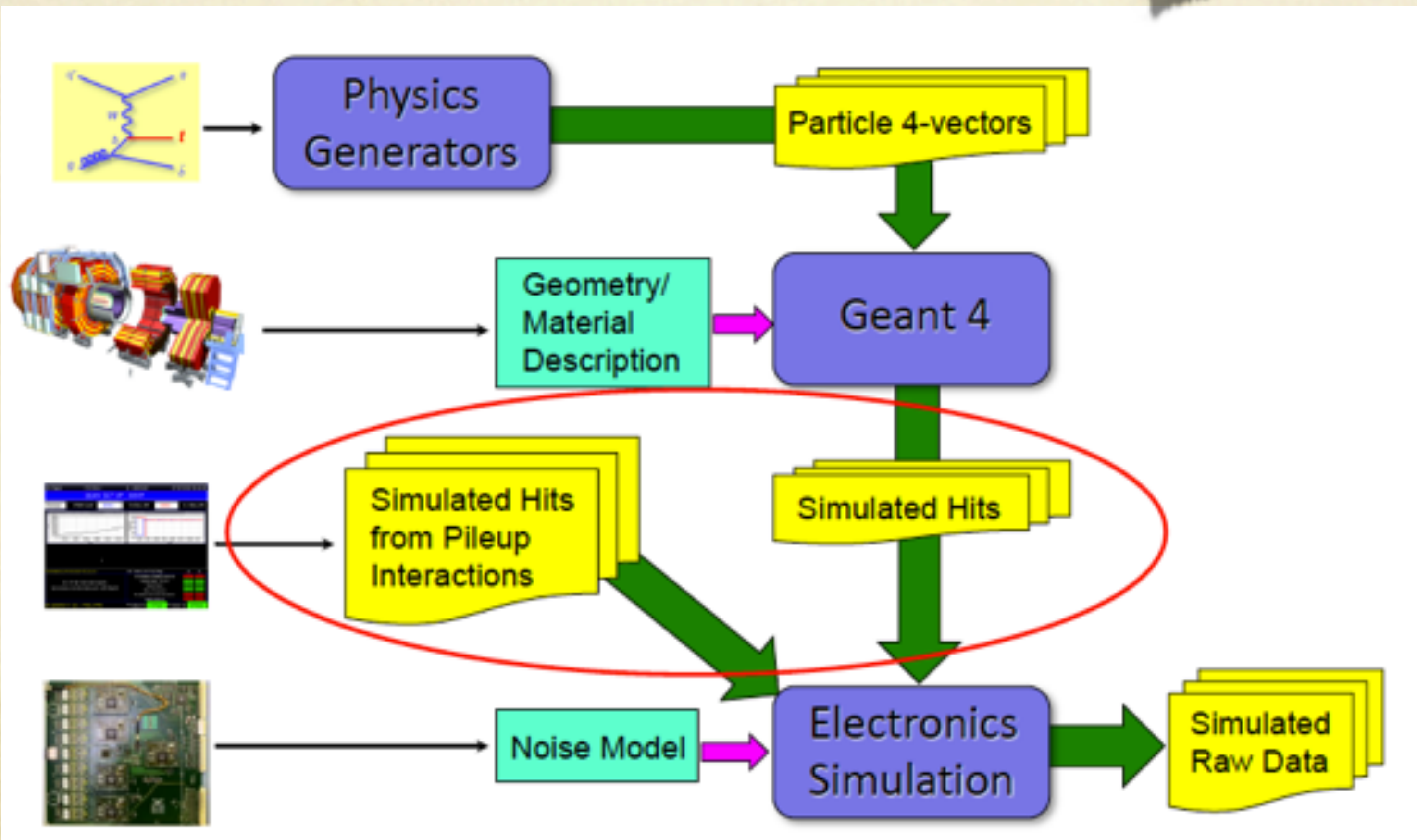
- a new framework where the different simulations approaches are fully integrated in a flexible manner.
- mix-and-match the choice of simulation depending on the physics you are interested into
- Choose at run time!
- (My thought: really cool but can become a validation nightmare...)



One more player in the game: the PileUp

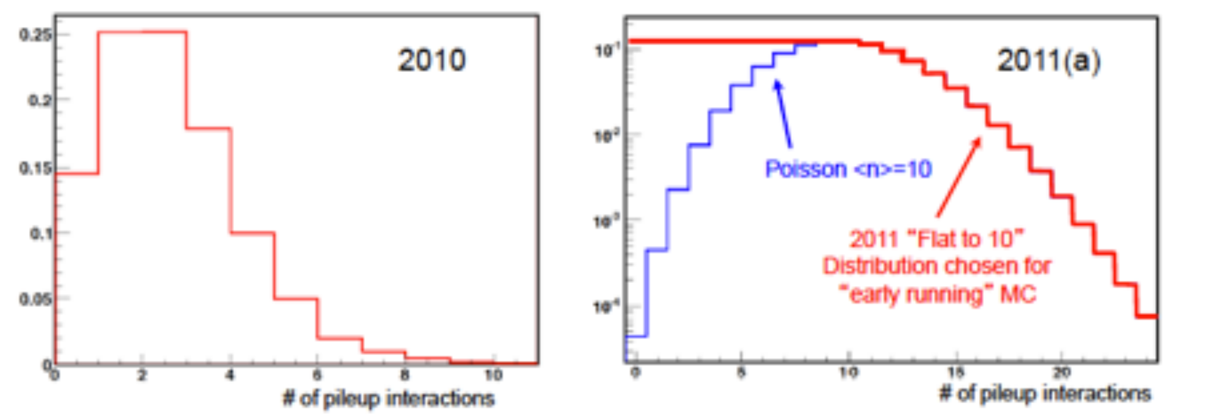


@Lumi = 1.0×10^{34}
if 50ns 50 PU
if 25ns 20PU + OOT



MC productions: guessing the PU ahead of time

- a good guess at the PU distribution to use in the MC campaign ahead of time can make a big difference in the efficiency of the production
  - for instance if the PU is underestimated when we reweight the distribution we might lose even a large fraction of the Monte Carlo statistics that was produced.
  - Sometimes it is better (more efficient) to resimulate from scratch!



2010 distribution chosen half-way through data taking. Extrapolation based on expected instantaneous luminosity. Matched data almost perfectly.

2011 distribution chosen well before running. Designed to generically represent data up to Summer Conferences. Reweighting necessary to match data luminosity distribution.

How to simulate it ?

- For each event, the instantaneous luminosity is chosen from the input distribution at random
  - the number of in- and out-of-time interaction to be overlaid are selected individually from a poisson distribution based on the chosen luminosity and the total inelastic cross section (CMS uses sigmatot=71.3mb)
  - Out-of-time interactions are simulated for each beam crossing that is «considered» for a given production configuration:
    - arbitrary number of bunches in 25ns steps
    - the times of the hits are shifted to match the bunch they belong to. the digitization simulation will consider the proper hit times for pulse shapes
    - typically simulate ±125ns of bunch crossing, but studies show might need more!
  - various ways to mix the hits from the collection of MinBias events and hard scatter «signal» before the processing
    - can mix generator particles, SimHits, digis, tracks, or pre-mix the proper number of events ahead of time and just overlay one «PU-event» with the signal
      - being investigated by CMS for future productions

Some timing/performance results

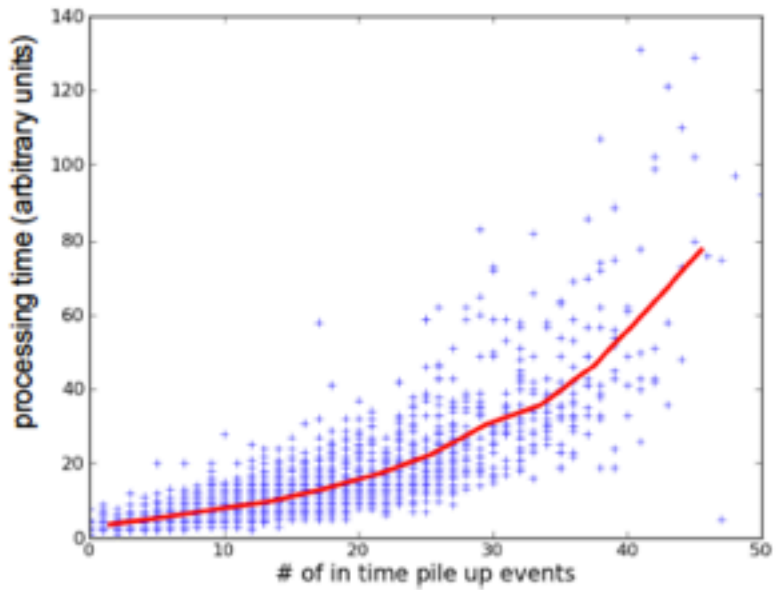
- here just note the «relative» increase with the PU, and consider that the RunII will have a much higher level

Scenario	Pileup + Digi time (a.u.)	Reco time (a.u.)
No pileup	2.3	3.0
Flat10+Tail (2011a)	8.9	7.1
Peak=14 (~3x10^{33})	6.6	8.7
Peak=20	9.6	14.5
Peak=32 (~5x10^{33})	12.3	26.1

x6! x10!

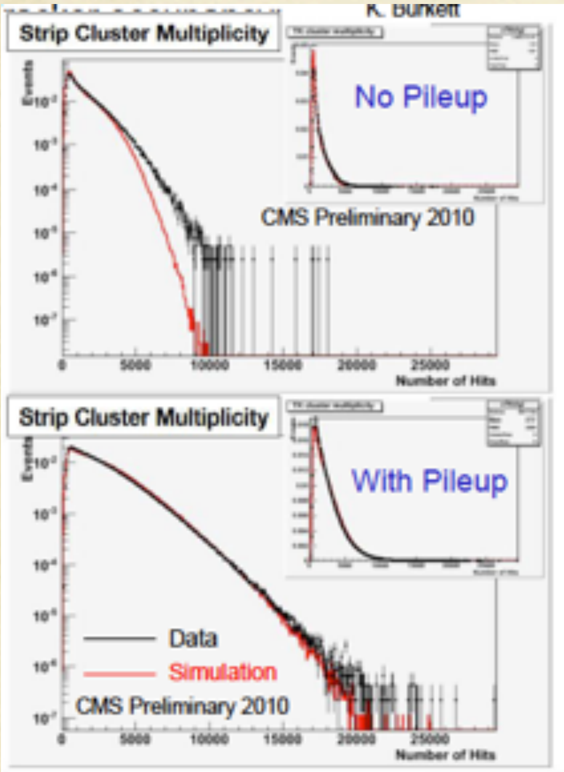
– increase in memory usage above no-pileup case:

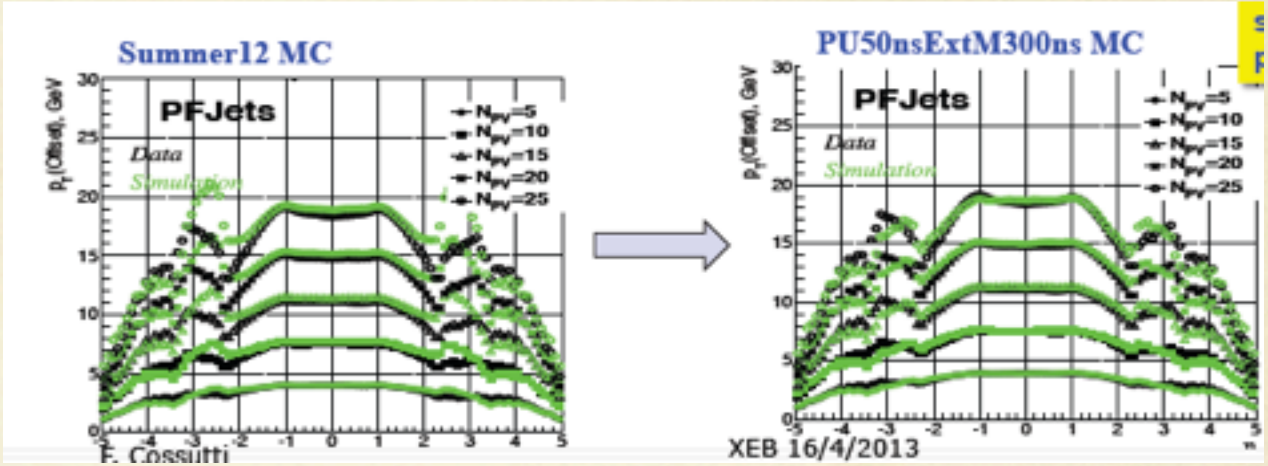
Scenario	ΔDigi Vsiz	ΔReco Vsiz
Peak=14 (~3x10^{33})	+510	+272
Peak=20	+468	+383
Peak=32 (~5x10^{33})	+628	+836



Validation of PU simulation

IN TIME PU ONLY HERE
Multiplicity of hits in Tracker

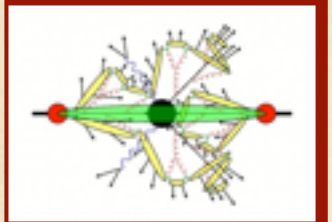




Scenario	Time (s/ev)
Summer12 <PU>~21, [-2,2], 50 ns	5
<PU>=20, [-12,2], 25 ns	12
<PU>=40, [-12,2], 25 ns	26
<PU>=140, [-2,2], 50 ns	56

- Studies have shown that the effect of the OOT PU on the calorimeters is significant.
- Need to simulate up to 300ns «before»
- This means mixing in even more events: consequences on production performance

The data flow

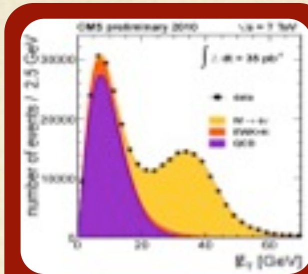


Monte Carlo Event Generator

Simulation of the Detector Response

Trigger

Event Reco.

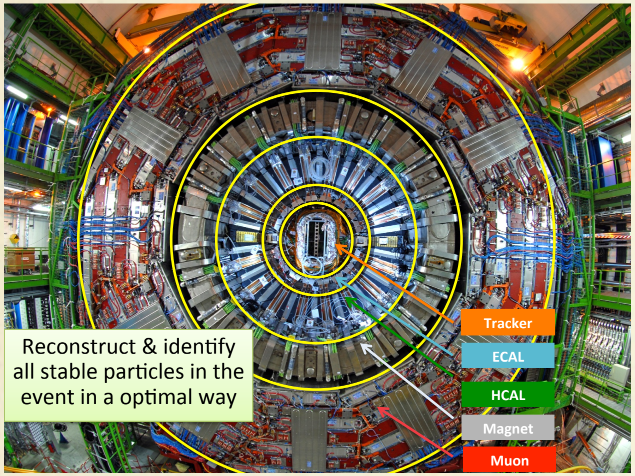


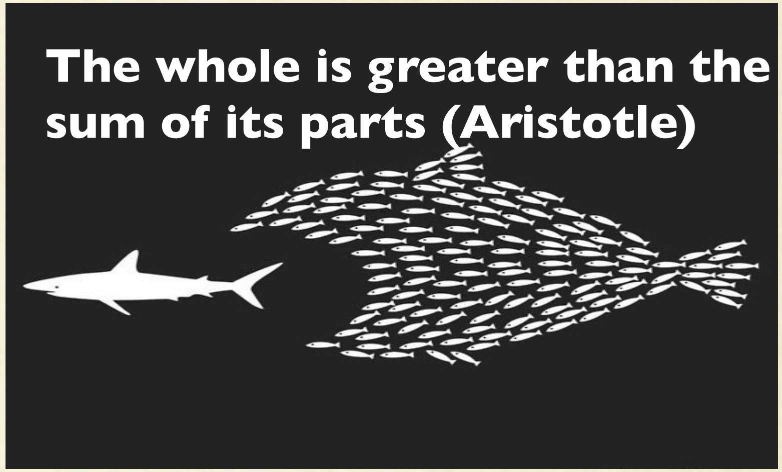
Physics Analysis

IMPROVING THE RECONSTRUCTION

Tying it all together

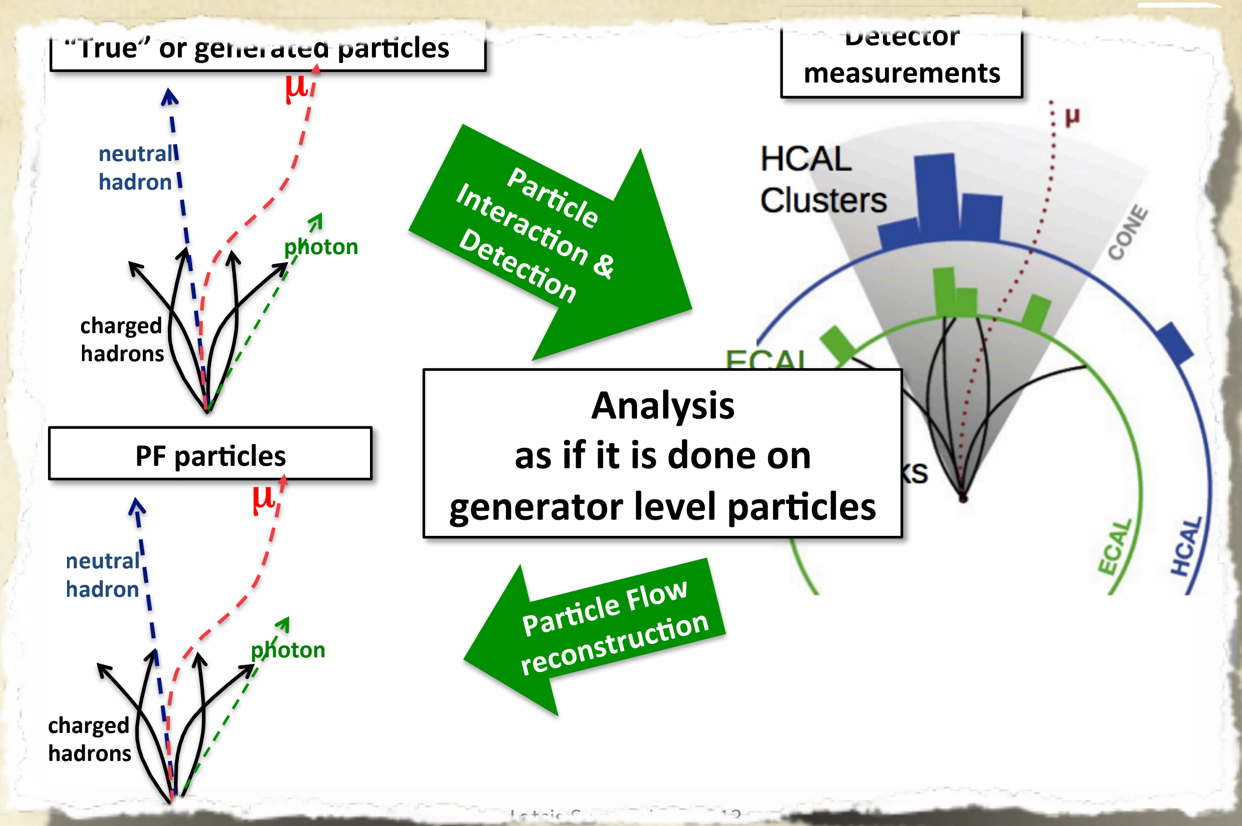
- Our detectors are very complex beasts, made up by many subdetectors with different purposes and characteristics.
- Sometime useful to consider higher level physics objects like separate entities (electrons, muons, jets…). This is the tradition for most of the collider experiments.





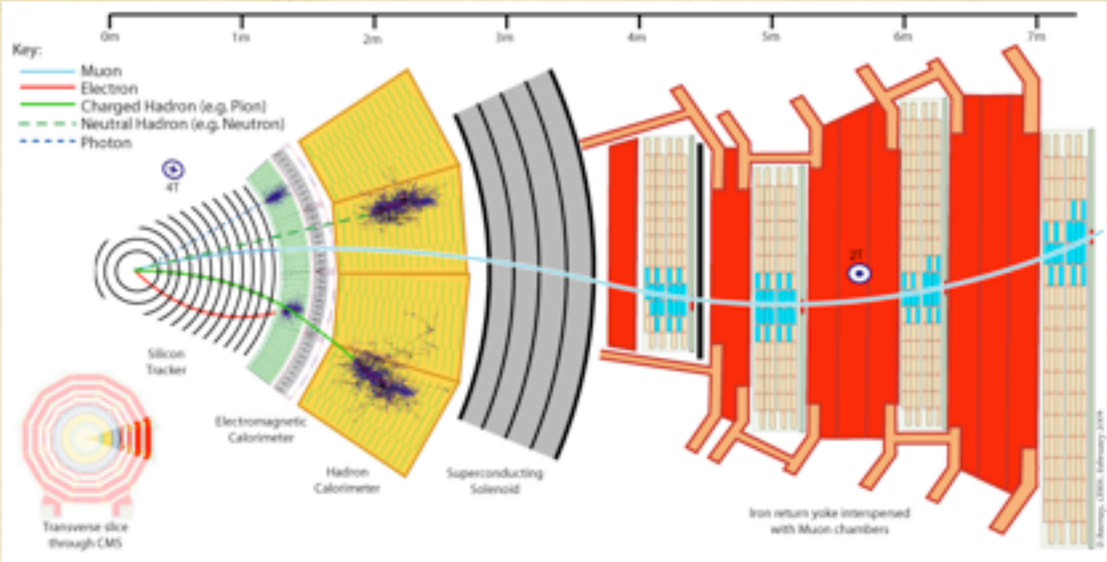
- However a collision event can also be looked at as a «whole». correlations and interconnections across the detectors taken into account
- The way of performing the reconstruction combining in an optimal way the information of all our detectors: Particle Flow approach (at CMS).
- Developed and used in CMS only (due to different detector characteristics). Might be considered also by ATLAS for the future to be used to fight the harsher PU conditions.

The Particle Flow



- moves the reference system from the «detector based» to the «particle based» using/combining the full information available into the event
- through combination of information it allows to maximally mitigate the PU effects
- List of individual particles is then used to build jets, determine missing transverse energy, to reconstruct and identify taus from their decay products, to tag b jets…

the particle flow



hits in the tracker → tracker tracks

cells in the calorimeter → calorimetric clusters

hits in muon detectors → muon tracks

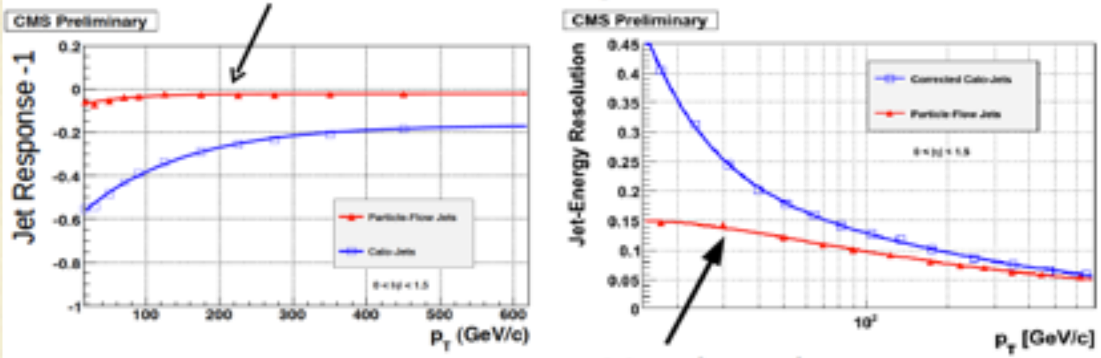
link the single objects with geometrical requirements on the extrapolated trajectories and create **blocks**

Comparison of Jet performance

- **Calorimeter jet:**
  - $E = E_{HCAL} + E_{ECAL}$
  - σ(E) ~ calo resolution to hadron energy: 120 % / √E
  - direction biased (B = 3.8 T)

- **Particle flow jet:**
  - charged hadrons
    - σ(pT)/pT ~ 1%
    - direction measured at verte
  - photons/electrons
    - σ(E)/E ~ 1% / √E
    - good direction resolution
  - neutral hadrons
    - σ(E)/E ~ 120 % / √E

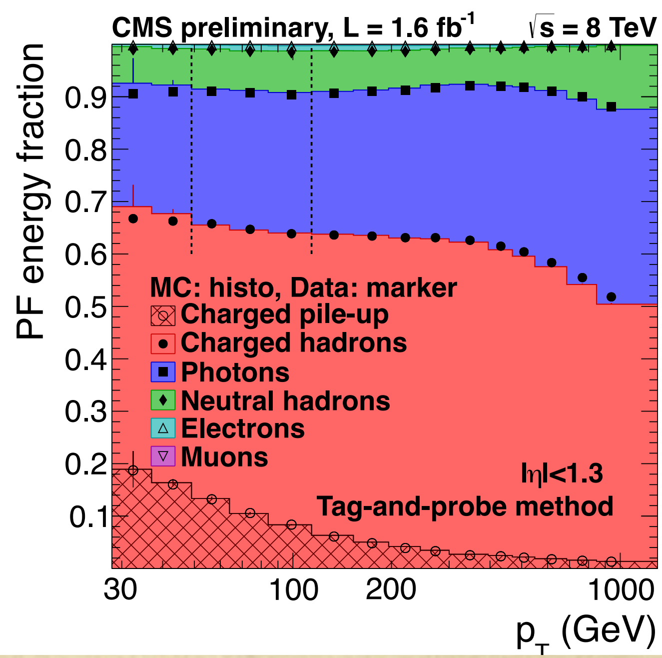
95-97% of the p_T reconstructed, over the whole range



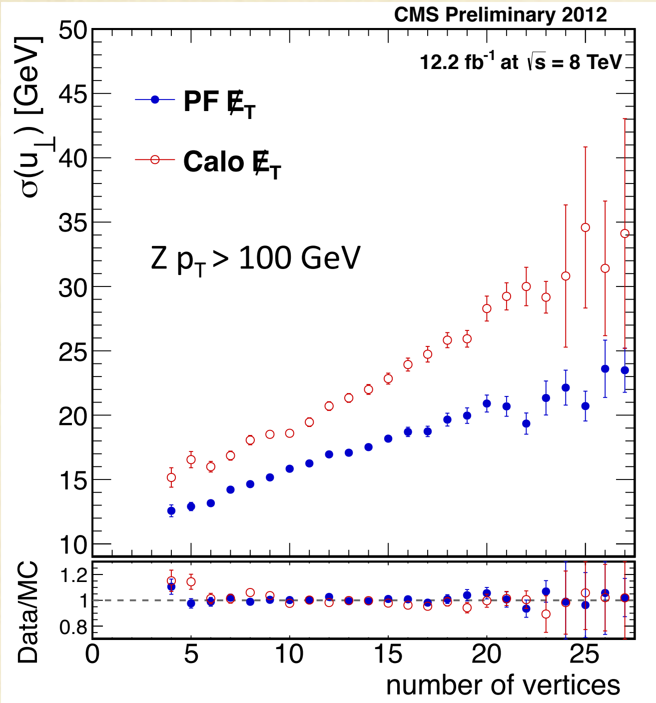
Very large improvement at low pT, thanks to the tracks

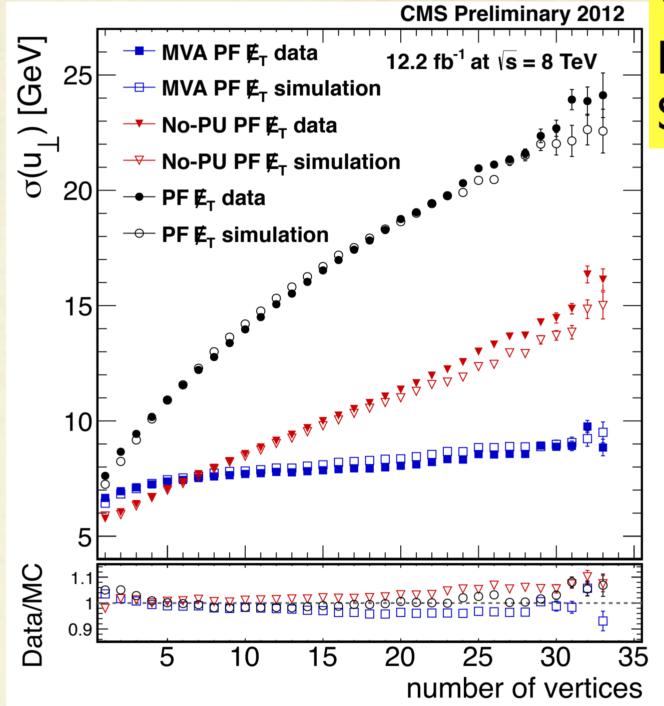
Still poor resolution, but neutral hadrons are the smallest component of the jet/event particles:

- **70% charged hadrons**
- **20% photons**
- **less than 10% neutral hadrons**



Missing Transverse Energy resolution





WILL SEE THIS PLOT AGAIN SOON!

$$\overrightarrow{E_T^{miss}} = -\sum_{i=1}^{N_{particles}} \overrightarrow{E_T}^i$$

Factor two improvement in the Missing E_T resolution compared to calorimeter based Missing E_T and more robust against PileUp.

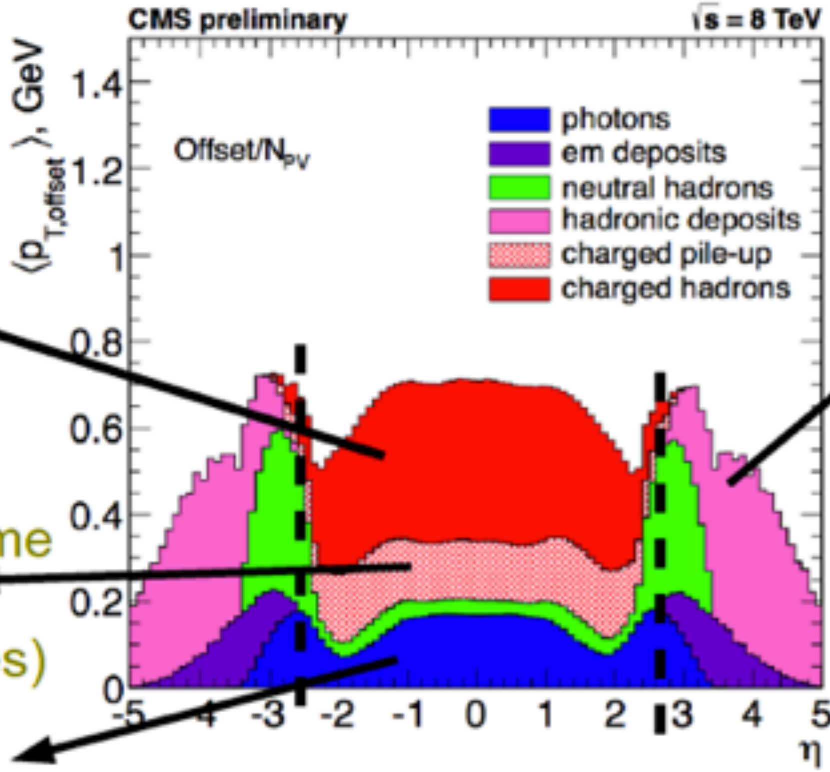
wasn’t it good for PU mitigation?

08/13/13 Philip Harris BOOST 9

Pileup Composition in CMS

Tracker is in the central region

Flag with Vertexing



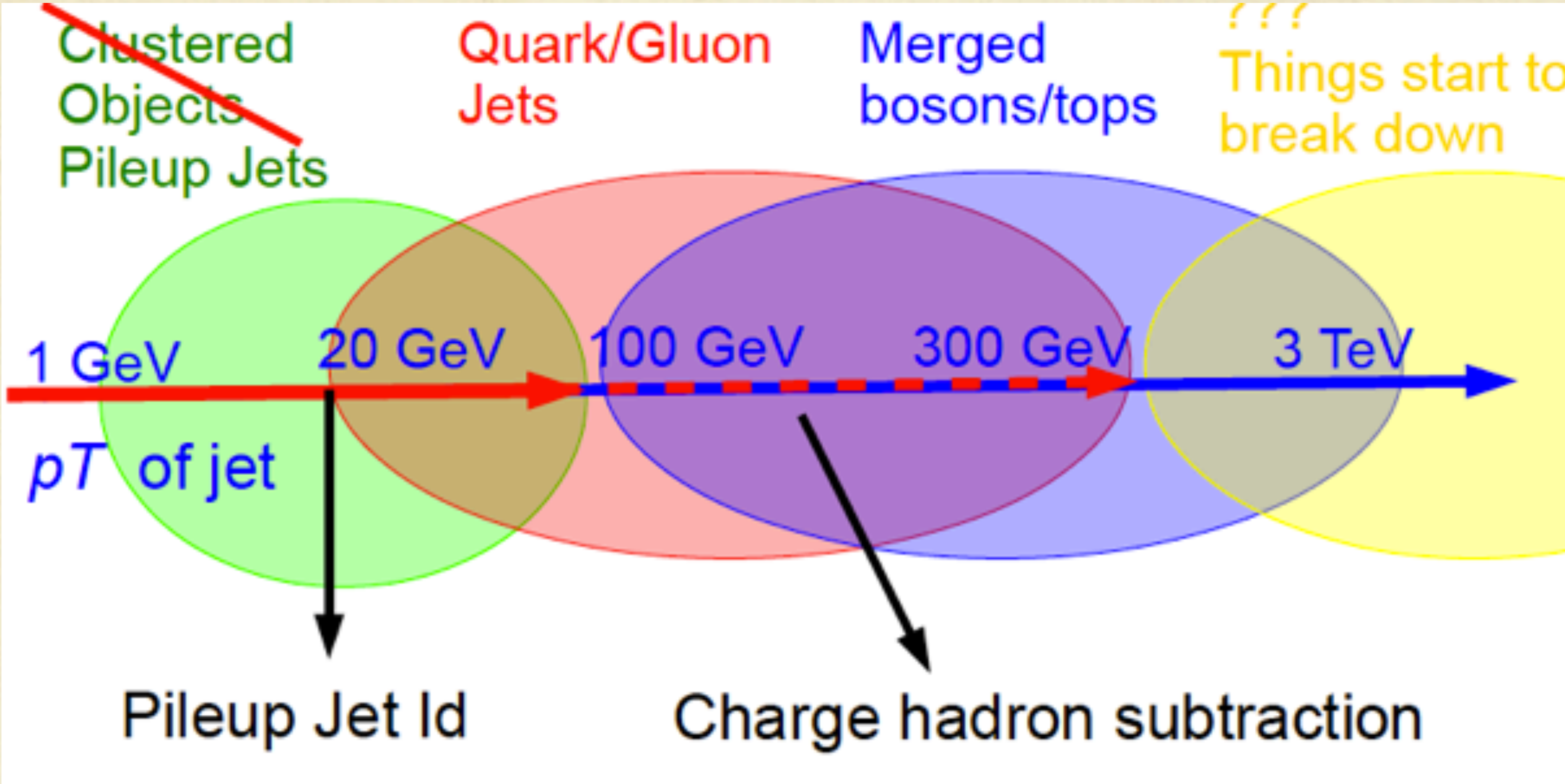
No tracker
Must rely on shapes to identify

Flag with some uncertainty (vtx + shapes)

Guilty by Association (rely on clustering)

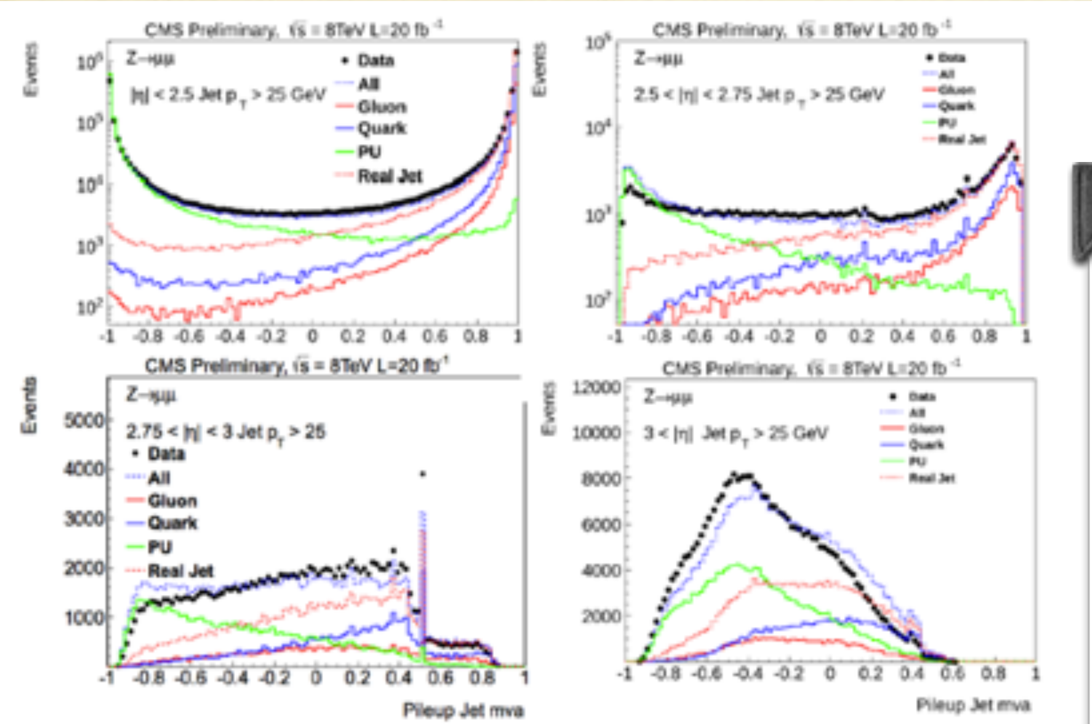
Charge Hadron Subtraction:
Flag from another vertex and remove it

Methods for PileUp mitigation

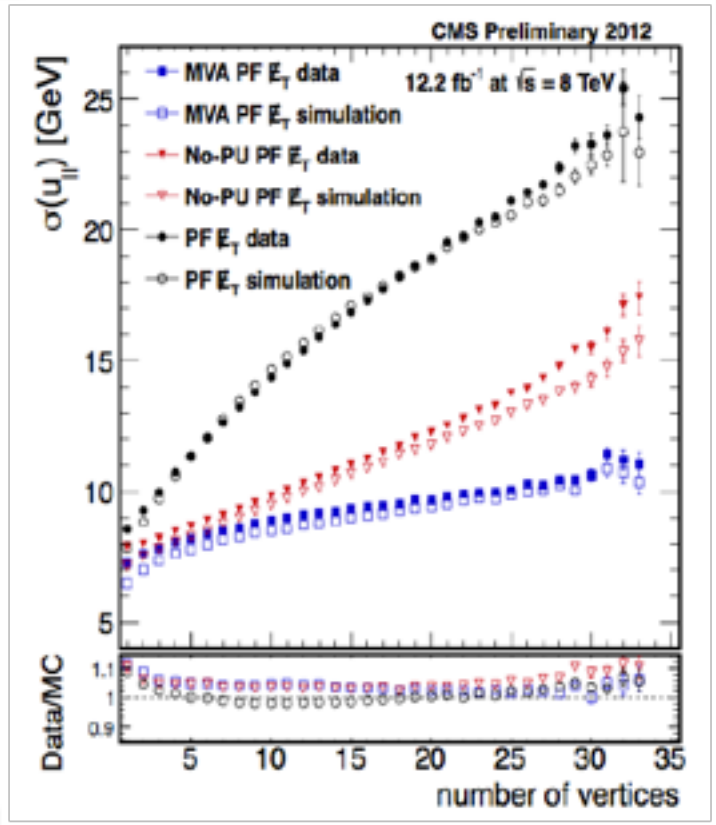


- Subtraction of tracks non associated with the Primary Vertex
- Subtraction of average energy deposit under the jet area from the PU event
- (On these «cleaned» jets) PileUp jet Identification based on Jet Shapes variables

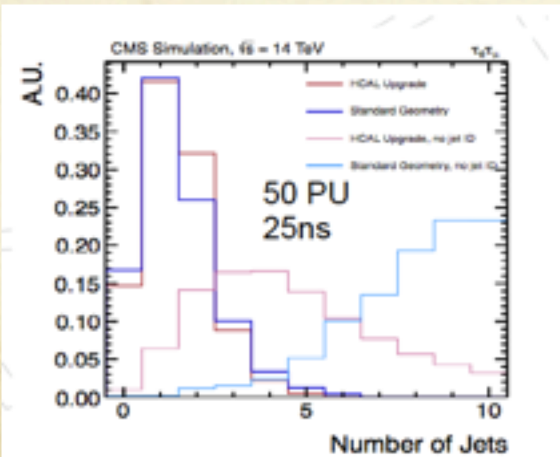
PileUp Jet ID in data & MVA MET



As expected PileUp jets fraction increases at larger eta



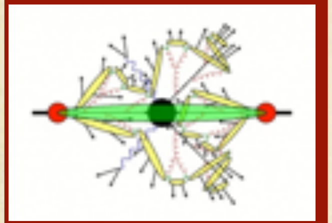
PileUp Jet-ID helps remove also the OOT PU



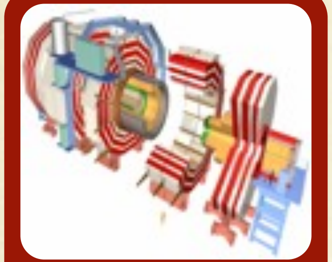
VBF selection (expect 1-2 jets)

Factor 4 improvement in PU dependence for MET resolution

The data flow



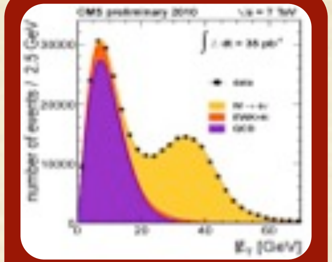
Monte Carlo Event Generator



Simulation of the Detector Response

Trigger

Event Reco.



Physics Analysis

COMPUTING CHALLENGES & MONTECARLO PRODUCTION

Offline and Computing challenges for 2015

- In 2015 there are 3 main factors that drive the need to increase computing resources.
  - we expect an increase in the number of pile-up events that with the current code would require a factor of 2.5 increase in reconstruction time to process.
  - the average trigger rate expected to grow a factor of 2.5 higher *(if we do not change the thresholds)*
  - currently the code reconstruction speed depends on out-of-time pile-up in the tracker *(might have a solution for this)*
- With no changes in the way the experiment works, we would require a factor of 6 increase in the processing resources to maintain the current activities. Need then to make changes since budget constrain will not allow that.
  - One different operationg mode will be to move PromptReconstructionprocessing from the T0 to the T1 as well.
    - procedure validated
    - less need for big reprocessing, more space at T1s
- Actually IF these assumptions hold the situation is not so gloomy, however...need always to plan for the worse.

The requests for CMS as a function of time

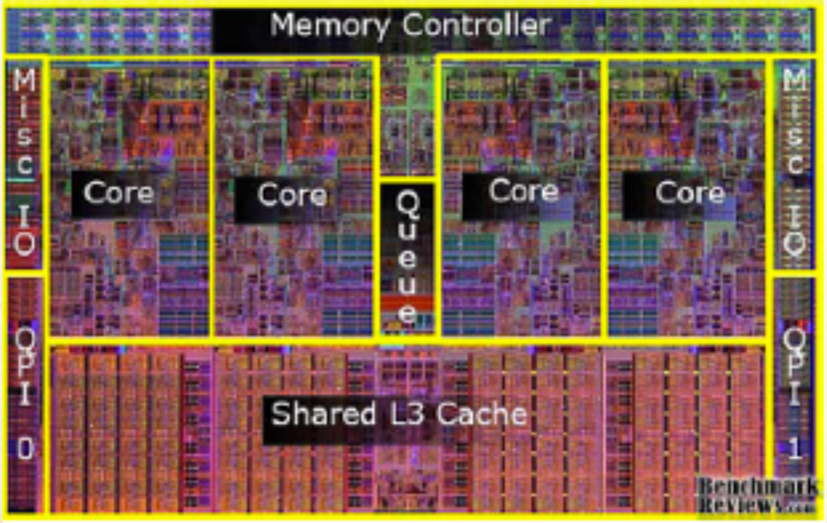
	2013	Increase from 2012	2014	Increase from 2013	2015	Increase from 2014
Tier-0 CPU	121	0%	121	0%	256	111%
Tier-0 Disk	7000	0%	7000	0%	3250	-53%
Tier-0 Tape	26000	0%	26000	0%	38000	46%
CAF CPU	0	0%	0	0%	12	
CAF Disk (TB)	0	0%	0	0%	12100	
T1 CPU	165	12%	175	0%	300	71%
T1 Disk (TB)	26000	0%	26000	0%	24000	-8%
T1 Tape (TB)	50000	11%	55000	11%	79500	43%
T2 CPU (kHS06)	350	8%	390	14%	500	25%
T2 Disk (TB)	26000	0%	27000	4%	31400	16%

The disk space needs scale because of moving to AOD and not using RECO for analysis even if running at a new energy.

Possible gains: multicore

- In any case:
  - Improved reconstruction algorithms in general (especially Tracking as we have seen)
  - Technical improvements on code size/layout, performance optimization. This always pays off.
    - <average> Physicist does not code very efficiently, maybe new generations better?

- **Multicore scheduling**: up to know we have been sending ~one job per core on multicores CPU.
- During LS1 will switch the scheduling to Multicore on all T1 sites.
- In first approximation the switch will not increase total troughput righ away. Still need to reduce memory requirements etc...
- In principle the number of jobs would drop from ~80-120K to 10-20K remaining constant.
  - production will run multicore, analysis will stay single-core but scheduled within a multi-core job



Possible gains: data management and computing model

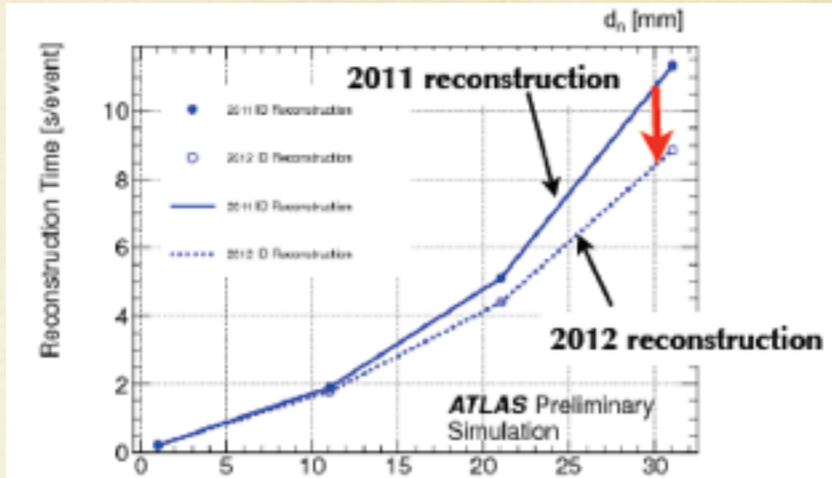
- Enough disk space to store all the data (twice!) however, static placement ensures safety but it is not the most efficient use of the resources.
- Several technologies ideas developed during LS1:
  - dynamic data placement and cleaning
  - remote access to data in other sites
  - separation of «archival» storage from disk storage at T1s
- Move away from the model of «pre-placing data» at a specific place and sending jobs needing those data only there.
  - Combine use of CRAB/Xrootd/CMSSW popularity + site readiness, Victor (data cleaner), PhEDEx --> triggering automatic transfer of data when needed
  - work in a more central(group) storage space for a more efficient use of the processing power.

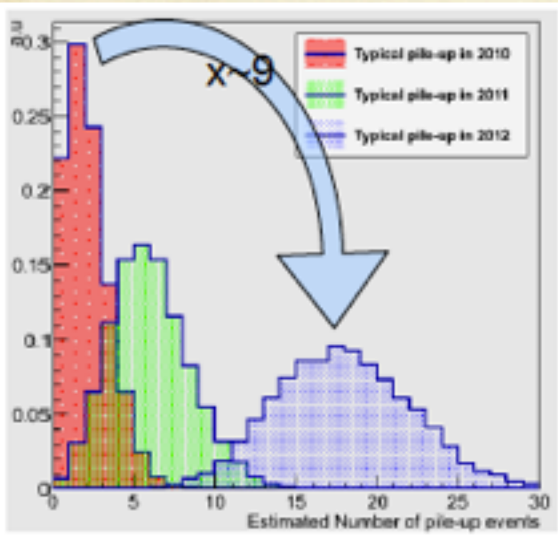
Something new: Opportunistic Computing

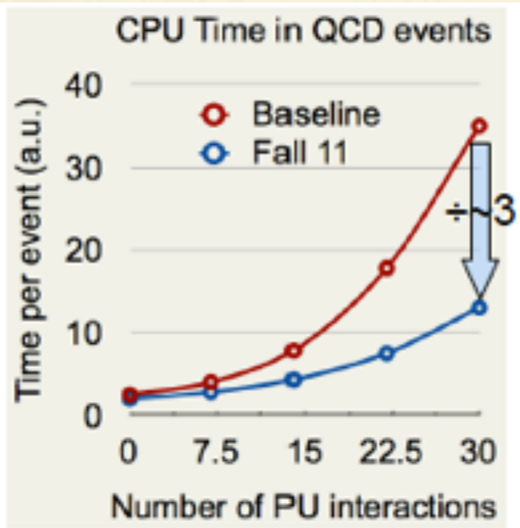
- During LS1 there will be a significant push into using additional «opportunistic» resources. These are resources usually processing capacity rather than storage, that do not «belong» to CMS/ATLAS, but to which we have access for various reasons.
- This will build on a number of techniologies we have adopted over the years to allow for a very «light» footprint at these sites: GlidInWMS, xrootd/AAA, CVMFS, Frontier, Parrot, remote stageout, etc...
- As we are not the primary users of these resources some sites may have «eviction» policies which require us to vacate the resources with little notice.
- Maximing our troughput on those resources might require additional developments.
- In addition, another avenue for increased «opportunistic» use is «volunteer computing», i.e. SETI@Home or BOINC-style processing.

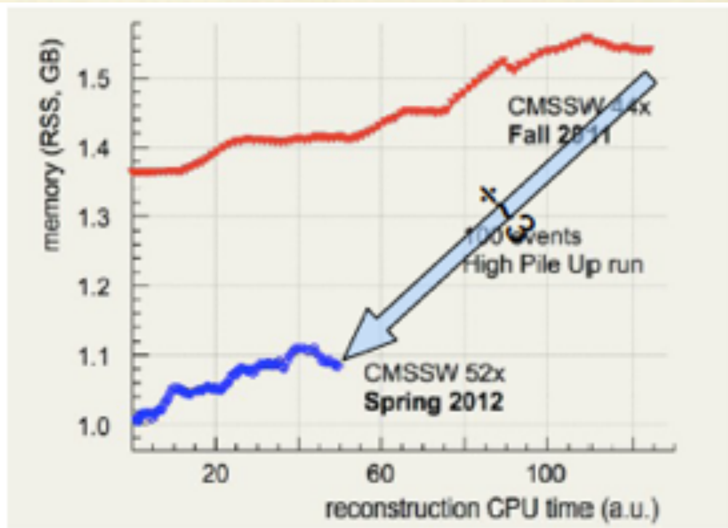
Lesson learned: Software developments from 2001 to 2012

- Going from 7 to 8TeV step had to cope with improved LHC operations: higher energy, luminosity and PileUp.
- Some improvement came from technical advancement in computing performance, orthogonal to physics event reconstruction
- Algorithmic development needed as well to cope with higher trigger rate without compromising physics
- Main gain from tracking algorithm optimization



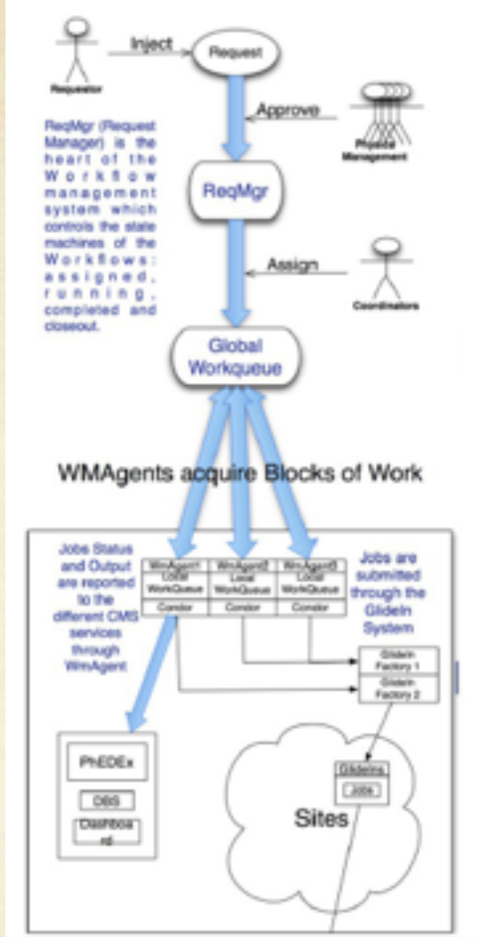






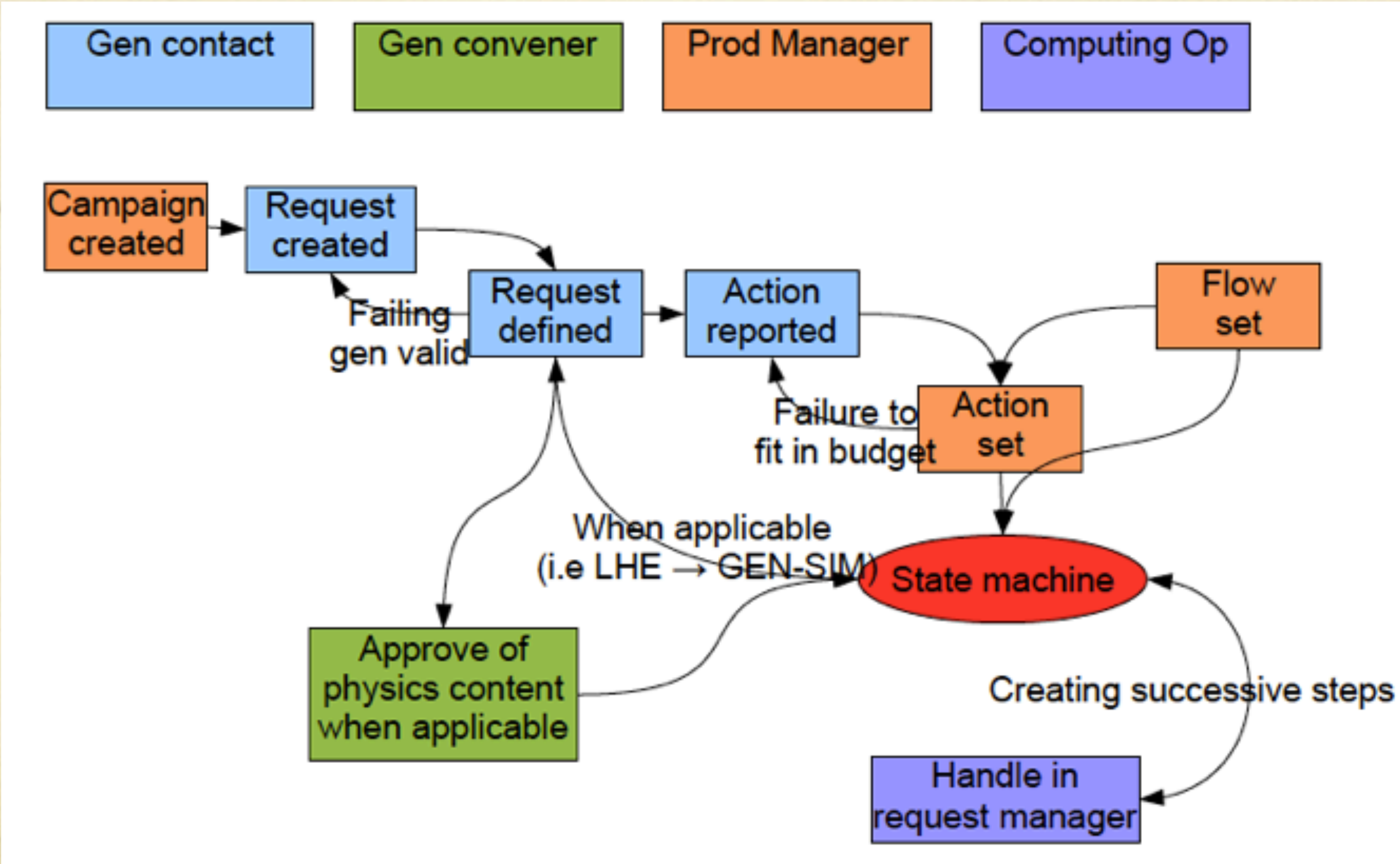
Maybe you want a sample? (Tools for MC production)

- Maybe you want ~100 of them (scan)
  - imputing/editing each of them by hand is tedious
- Maybe you need it "very fast",
  - It involves several steps and manual interventions
- Maybe you need it right
  - The generator configuration is subject to mistake
- Maybe you need it with a slight change of configuration
  - Which means manual intervention
- Maybe you need it for a given date
  - Which means everyone does and we need a schedule
- Maybe you **just** want it
  - And it can be lost in manual intervention, lack of book keeping, silent failures, ...



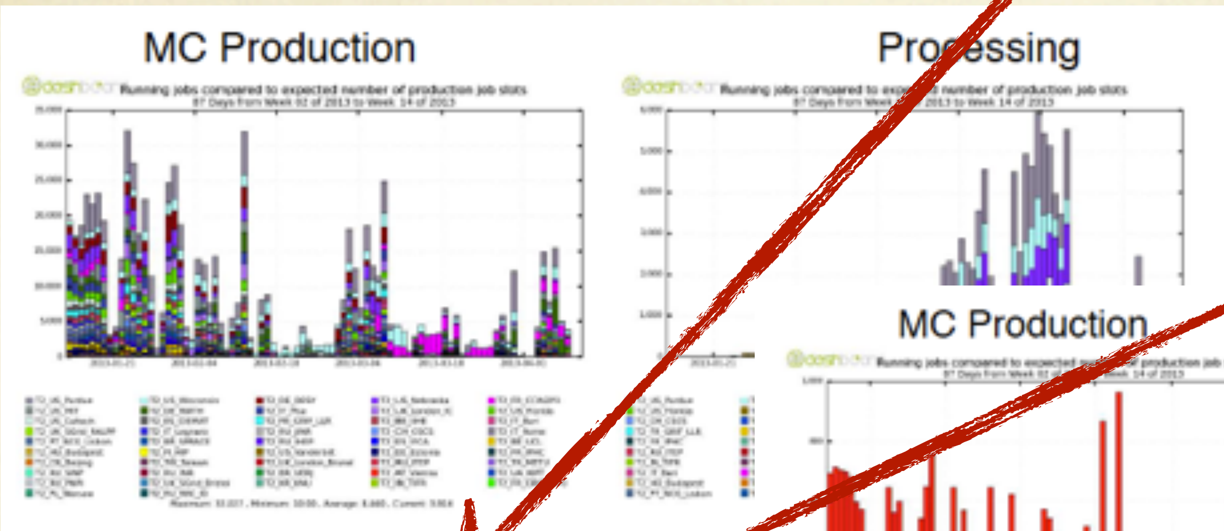
- Specialized tool to handle all the requests Bookeeping via DB. Chaining of similar request, campaign definitions.
- Most important (for the analyzer) improved the monitoring and traceability of the jobs

the complicated life of a MC request



Optimizing the resources: strategy!

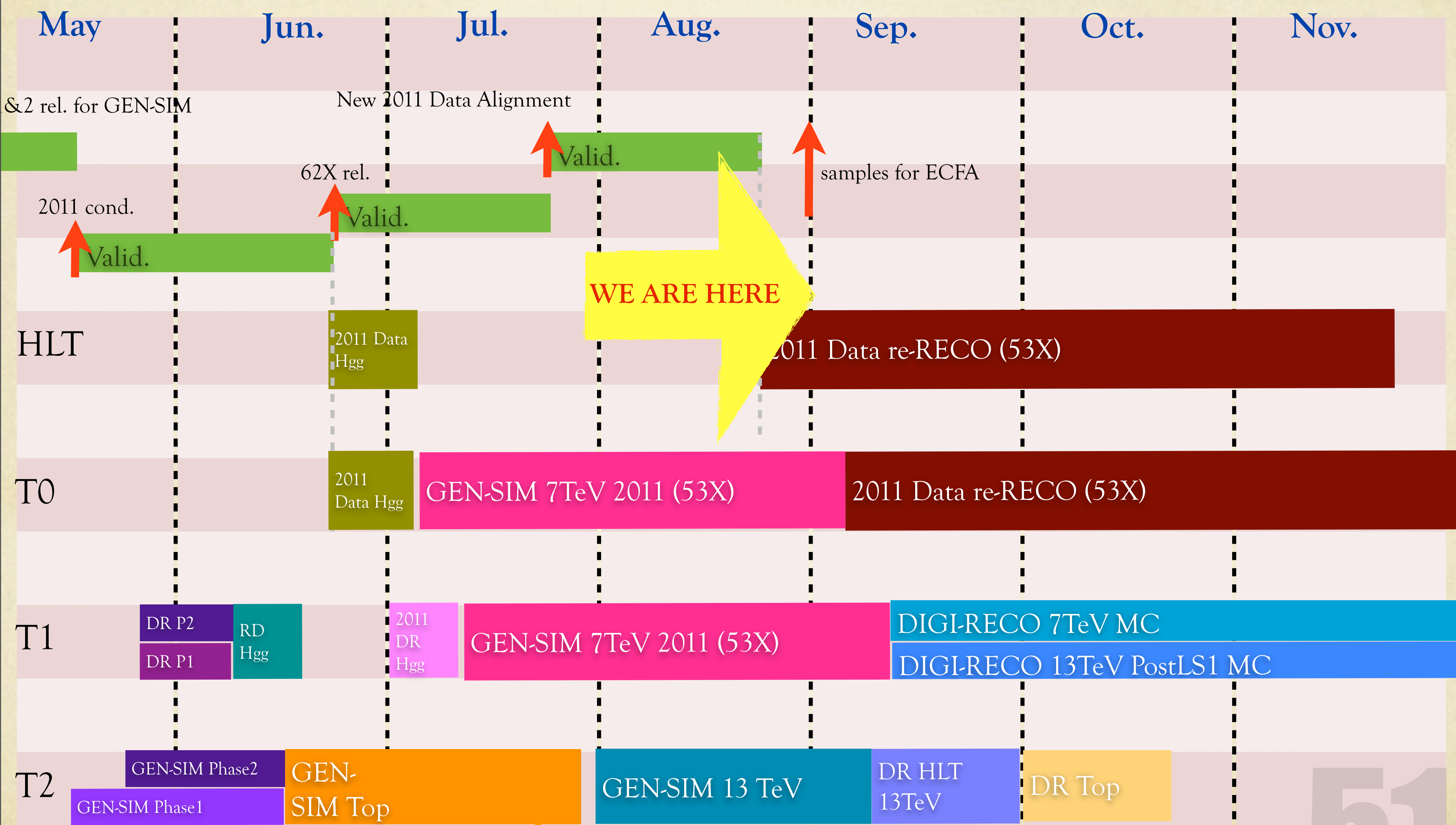
CMS Production since January 2013:
8TeV data reprocessing with optimal calibrations taking up most of the T1s



Or to opportunistic computing as San Diego Super Computing (SDSC)

Move the MC production to the T2s!

The plan keeps going...outlook for 2013!



The life of a Data Preparation Coordinator...

MonteCarlo Productions & Reprocessing

- Snapshot of what is going on now in CMS for the «Production» point of view:
  - reprocessing of 7 TeV data with legacy calibrations and reconstruction
  - production of corresponding 7TeV MC with legacy calibrations and reconstruction
  - production of 13TeV MC samples to start studies for RunII
  - production of 13TeV MC for Upgrades studies: detector degradation, new PhaseI configuration, new PhaseII configuration
- All these productions go along with a preliminary validation period before being launched.
- These special validation campaign add up on top of the regular release software validation for the new developments.
- A shutdown is not a vacation for everyone! ;-)

Planning the strategy for the future

- Based on the lessons learned during the past run we need to extrapolate:
  - The amount of MC that will need to be produced ahead of time
    - does not depend only on how much data to expect (x1.3), but which conditions as well
    - In the beginning small productions with «guesses» at beam spot, PU etc. Need to be trashed away and redone quickly.
  - the time/CPU it will take based on the expected perfomance of the code
    - need to work hard on the improvement to be able to sustain the HLT rate without reducing the Physics input (i.e. keeping same thresholds as RunI)
    - make sure processing time keeps up with the data flow
  - the computing model for processing and analysis
    - will need to evolve even more
- Readiness for Physics in the very beginning of RunII will be crucial. All the eyes will be on us.

Summary & Conclusions &

- During a shutdown there activity is fully projected into the future. Any improvement in technology, algorithms, or just good ideas in terms of strategy and models of operations are crucial.
  - It can make all the difference in times of tight budgets. Computing requests need to be reasonable but robust...not easy to predict the future.
  - A fast and accurate simulation of the detector, smart reconstructions algorithms and a well organized MC production ahead of collisions will allow analyzers to be ready to come out with physics results in a very short amount of time.
  - in between we will have also improved all those certification and calibration tools I talked to you about the other day
- **I want to thank the School Organizer for giving me the possbility of showing you these aspects of an experimental physicist life. Hope to see you all working with us very soon!**