



Reconstruction of LHC events at CMS

Tommaso Boccali - INFN Pisa
Shahram Rahatlou - Roma University
Lucia Silvestris - INFN Bari
On behalf of CMS Offline group



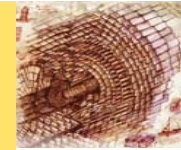
Outline



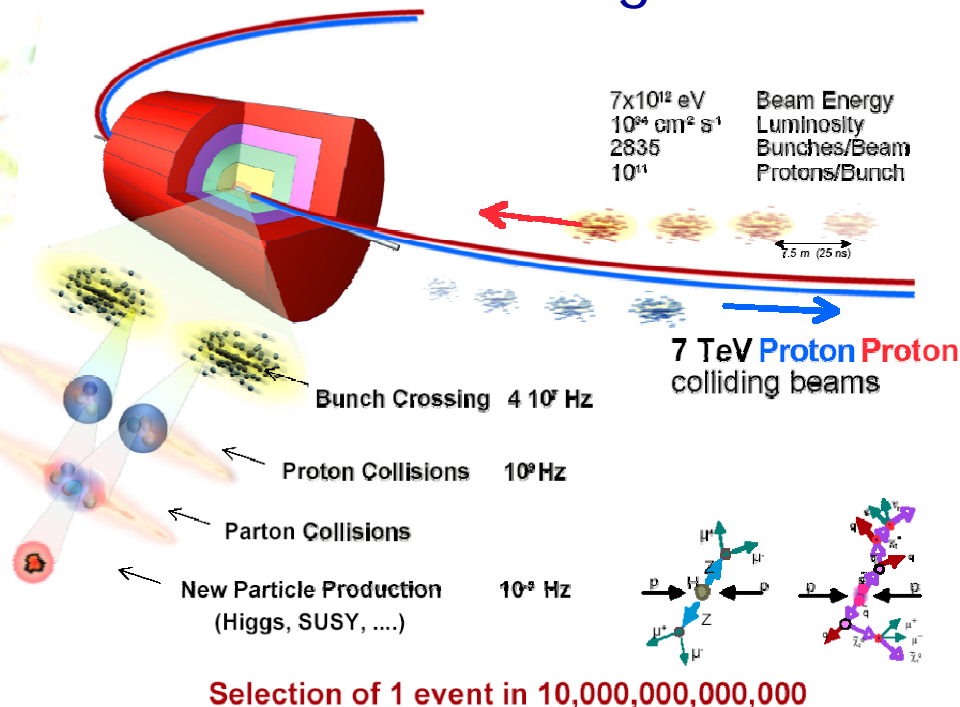
- Software for CMS reconstruction: use cases and needed performances
- The continued effort towards Data Taking
- Data Challenges and preparation for Physics



Requirements for a Reconstruction Software



1. Allow the processing of Raw Data to produce Data Objects useful for analyses
2. Allow the use of the full discriminating power of the CMS sub detectors in Physics measurements in offline analyses
3. ... but also something more @ CMS: HLT use case



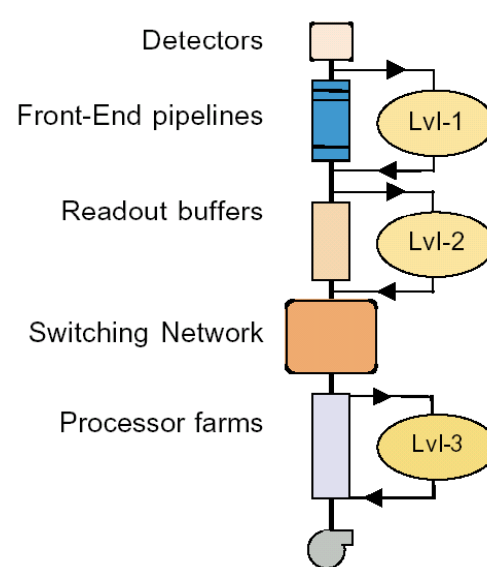
- The LHC collisions will occur @ 40 MHz, while the offline system can stream data to disk only at 150-300 Hz



CMS Trigger Strategy

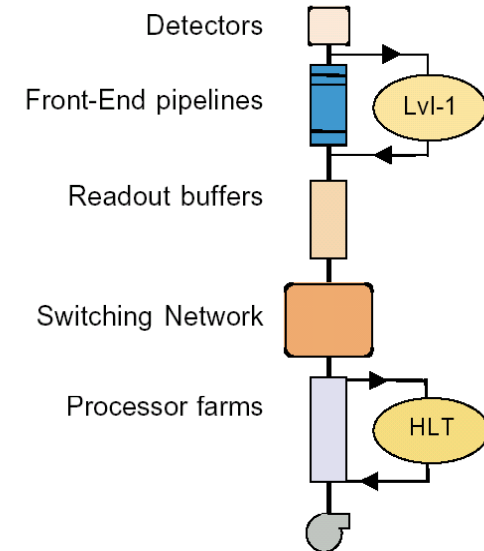


- CMS has chosen a trigger sequence in which, after a L1 (hardware based) response, reducing the events from 40 MHz to 100 kHz, the offline reconstruction code runs to provide the factor 1000 reduction to 150-300 Hz
- In this way:
 - We develop a single software for HLT and Offline
 - We can use the full complexity of Offline Software @ Trigger level
 - We can scale the system adding CPUs



Traditional

L2 functionality is the most challenging parameter. Does not benefit of full-granularity



CMS

HLT functionality depends on data rate and CPU resources available



Computing Model



■ CERN

- FilterFarm (HL Trigger)
- Cern Analysis Facility (small turn around time activities - typically calibrations)
- T0 activities (prompt reconstruction)
- Raw data custodial
 - ~ 7 MSi2k (2008)
 - ~ 5 PB on Tape; ~ 2 PB of disk (2008)

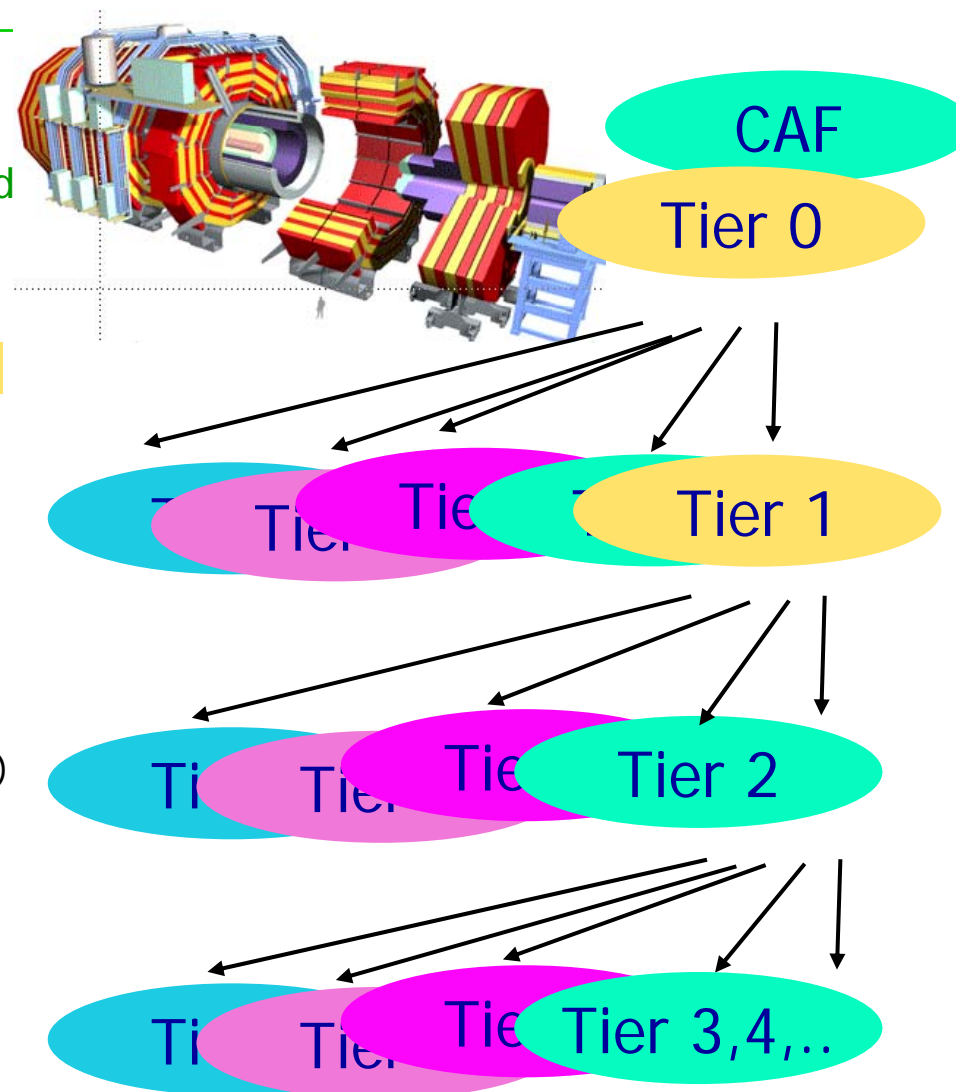
2008 CPU ~ 4Si2k

■ National / Super national level - Tier 1s

- Raw data custodial (shared)
- Re-Reconstruction
- Skimming / RECO/AOD production
 - ~ 10 MSi2k (2008)
 - ~ 10 PB on Tape; ~ 5 PB of Disk (2008)

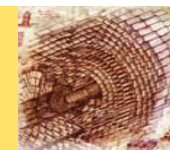
■ Community based - Tier 2s

- Simulation activities
- Analysis facilities
 - ~ 15 MSi2k (2008)
 - No tape; ~ 5 PB of Disk (2008)

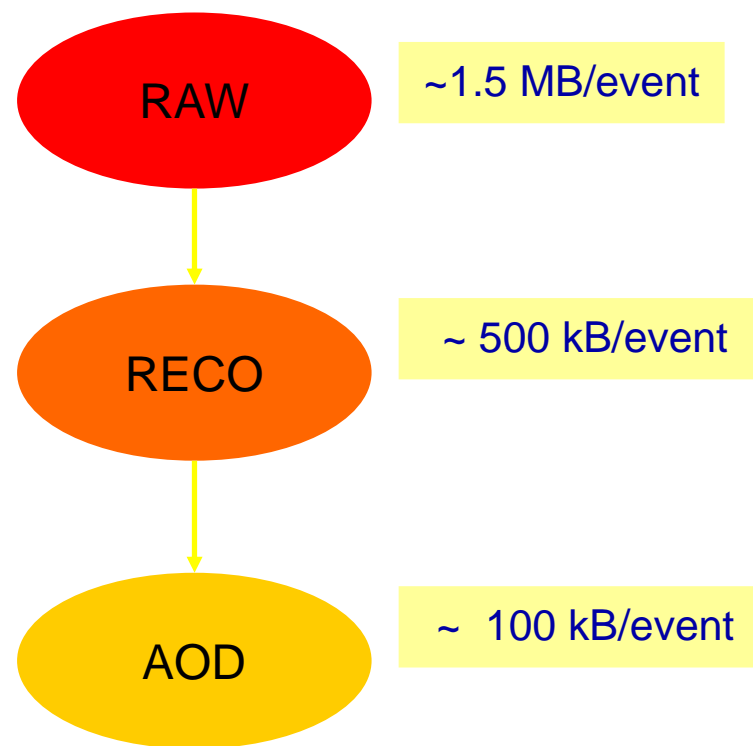




CMS Data Tiers



- CMS plans to implement a hierarchy of Data Tiers
 - **Raw Data:** as from the Detector
 - **Full Event:** contains Raw plus all the objects created by the Reconstruction pass
 - **RECO:** contains a subset of the Full Event, sufficient for reapplying calibrations after reprocessing
 - “Refitting but not re-tracking”
 - **AOD:** a subset of RECO, sufficient for the large majority of “standard” physics analyses
 - Contains tracks, vertices etc and in general enough info to (for example) apply a different b-tagging
 - Can contain very partial hit level information





Software environments



- The code must be able to run in three different scenarios:
 - Bare root: open the POOL file, and inspect the Data Objects _without any CMS infrastructure_
 - Portable on any machine which runs root, without _any_ effort
 - CMSSW-Lite: load a small number of libraries, don't allow access to any calibration, mag field map etc, but have full access to Physics Objects
 - Your laptop when you are not connected to internet
 - Full-CMSSW: full access to calibrations and full availability of libraries. Used mainly to produce reconstructed objects from RawData to lower Tiers
 - Your favourite T2

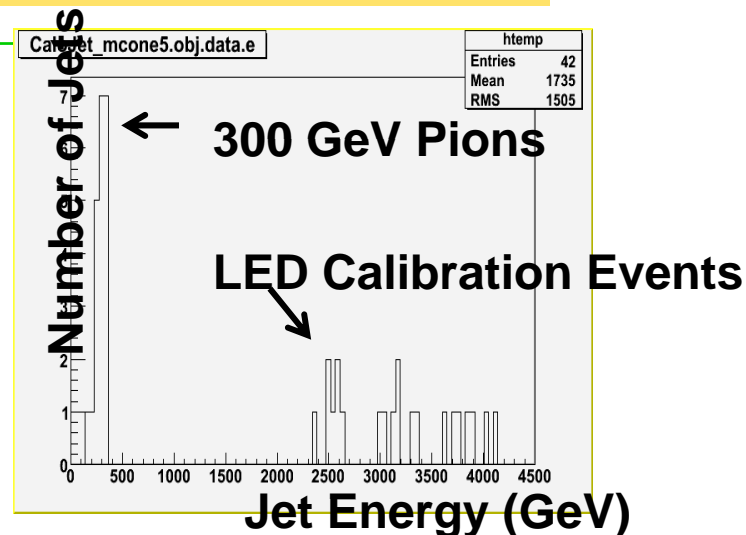


Status @ CHEP06



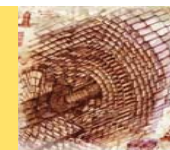
■ Last CHEP06 report

- New software being written (CMS transitioned to new framework in early 2005)
- Local reconstruction ok with simulated events (no infrastructure for calibrations and realistic conditions)
- Higher level reconstruction only sketched
 - Jets working (guinea pig)
 - Tracks "coming" (== under debug)
 - All the rest (higher algos) missing





Since then ...



- CSA06 (Sept/Oct 06):
Computing/Software and
Analysis challenge
 - Reconstruction enhanced with
 - Tracking
 - Electrons (initial version)
 - Photons (initial version)
 - B/tau tagging
 - Vertexing
 - Jets/MET
 - First definition of data Tiers
(FEVT, RECO, AOD)
 - Re-reconstruction, skimming
demonstrated

- Total events
processed > 100M
- Performance
 - < 25 sec/ev (on 1kSi2k
CPUs) even on ttbar
 - Memory tops at 500
MB/job after
hours/thousands of
events
 - Crash rate < 10^{-7} /event

Success!



Ramp up in code size / developers (Full CMSSW)



18 months = +900k lines of code
4x time the involved developers

Lines of code in the repository

Number of developer active
in a given month

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

CHEP06

04/09/2007

CHEP06

CMS Reconstruction

L. Silvestris 10



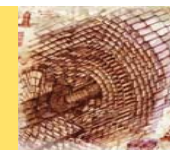
Now ...



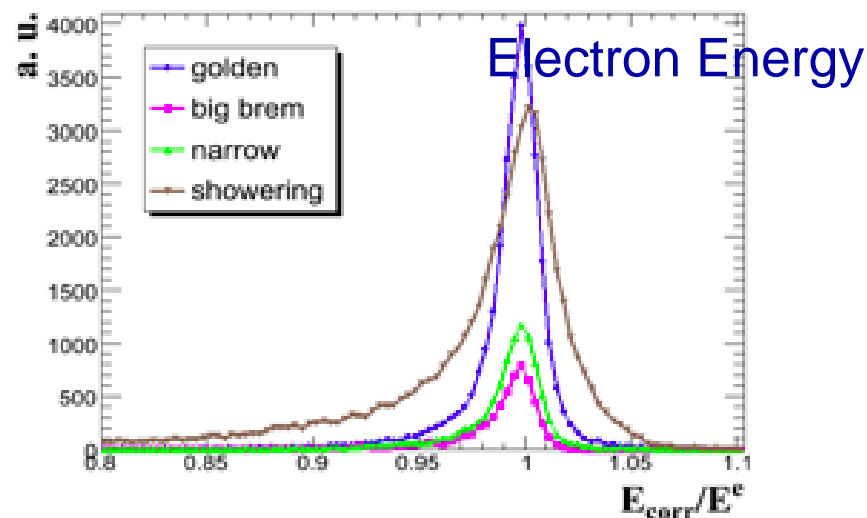
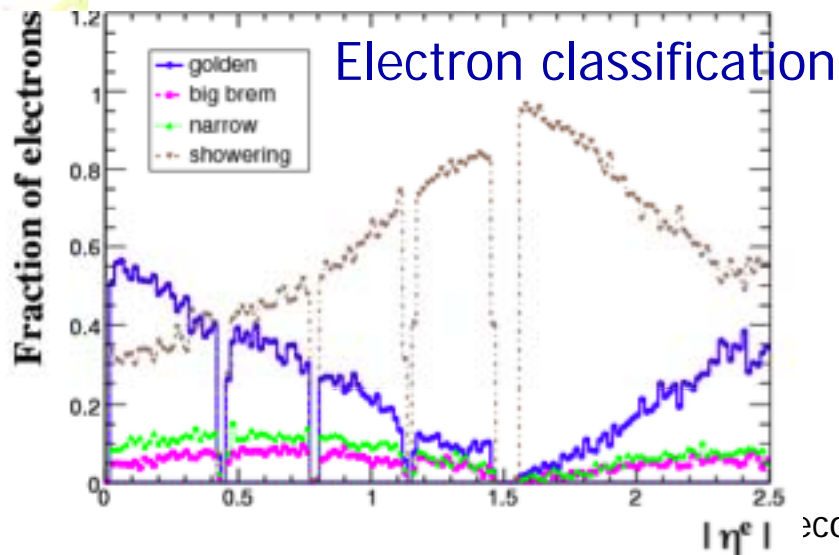
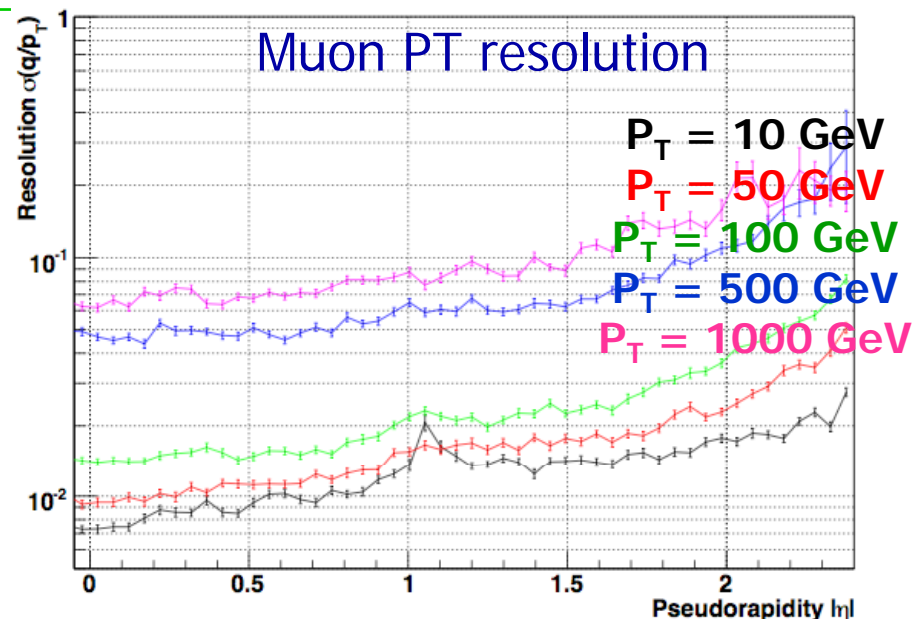
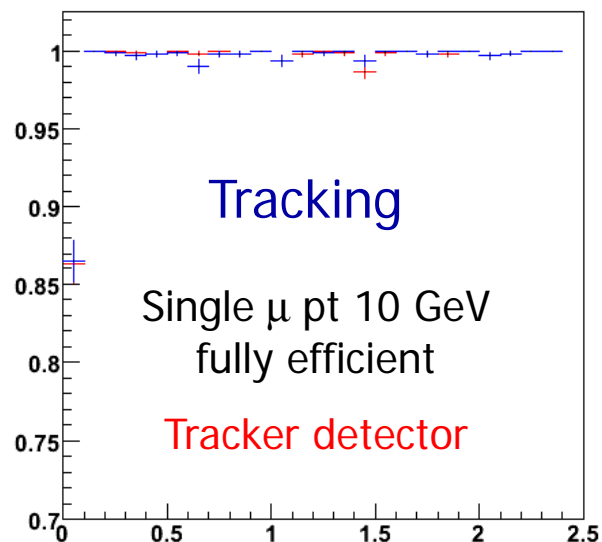
- CSA07 just taking off
 - Includes also HLT and analysis skim workflows working from raw data
- Extensive physics validation on Reconstruction code
 - Complete switch took 2 years
- Much more attention to calibrations
 - System “complete” - ready to handle real data (and being tested with commissioning data)
- Lots of new algorithms
 - Impossible to mention all of them
- **Tracking optimized**
- **Electrons/photons optimal**
- **Btagging, tau tagging**
- **Many more jet algos**
- **Lots of vertexing algos**
- **Muons optimal**
- **Particle Flow**
- ...



Few Results



efficiency vs η

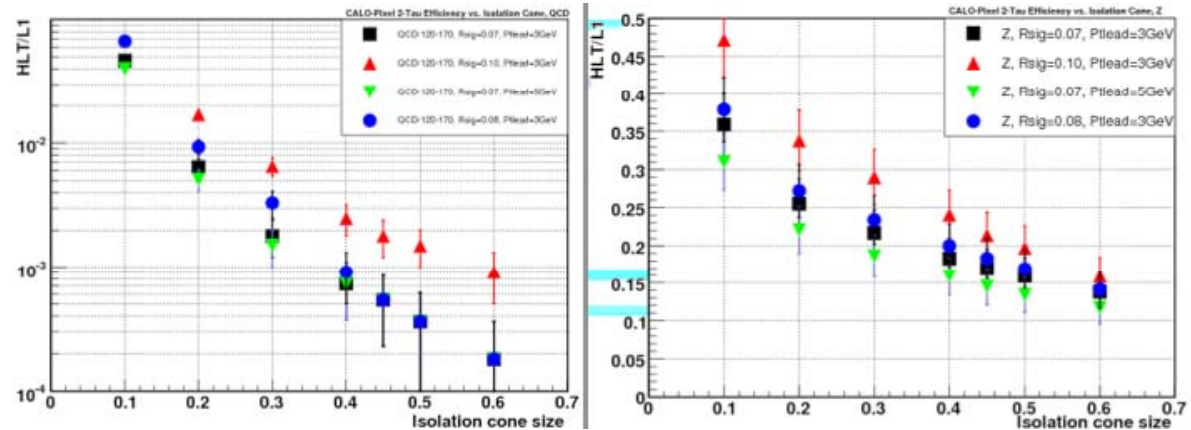




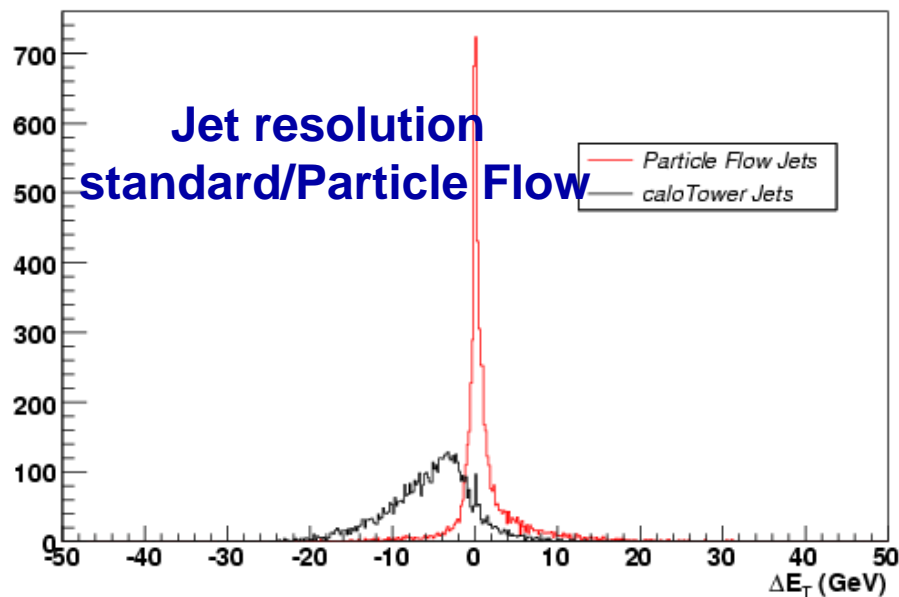
Few Results



Tau HLT Efficiencies (QCD and Z to tau tau)



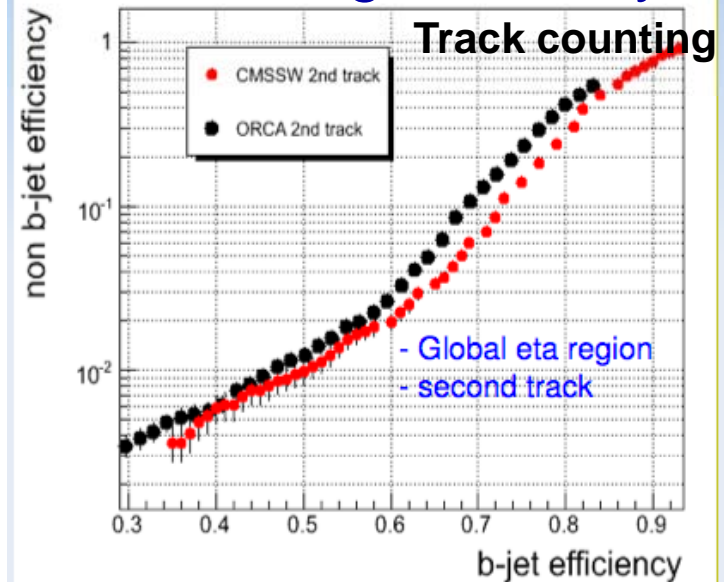
Jet Et difference ParticleFlow-MC



04/09/2007

CMS Reconstruction

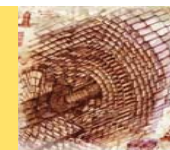
Btag efficiency



L. Silvestris 13



Data sizes

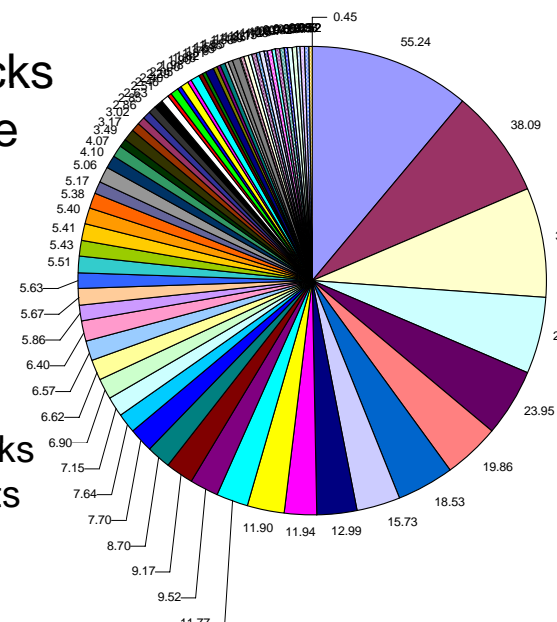


- Current figures show that we are not far from the estimated quota of 100 kB (AOD) and 500kB (RECO) (in the example, bb events w/o pile up)
 - Even including Simulated information, which accounts up to 20% of the payload
- Still to be reviewed when analysis use cases will be cleared
 - Expect some additions and a lot of cleaning

Data Tier	Size kB/event
RAWSIM	1420
RECOSIM	680
AODSIM	100

Validation and checks
release by release

RECO: Mostly tracks
and associated hits





Road from here to data taking



- CSA07 will probe reconstruction
 - On yet another $O(100M)$ events
 - With realistic calibrations loaded for all the subsystems
(== much higher RAM payload, ~ 1 GB/job)
 - With realistic misalignment scenarios
 - Using the complete range of reconstruction / re-reco /
skimming / reduced data sets production
- At the same time, reconstruction of data taken during commissioning (Global Runs) is certifying the correct use when starting from raw data



Conclusions



- CMS Reconstruction Software has jumped from being a development prototype to a working product in the period CHEP06-CHEP07
- It is actively debugged on
 - Hundreds of Million of Simulated Events (trying to operate in 2008 real data mode)
 - Real commissioning data (prepare/test the correct calibration/alignment infrastructure)
- We are eager to see the first pp data and confront with it - @ Next CHEP you will see Reconstruction code in Action on it!