Large scale event generation in CMS with MG5

Alexis Kalogeropoulos
On behalf of the CMS GEN group

Annual MadGraph Meeting 2011

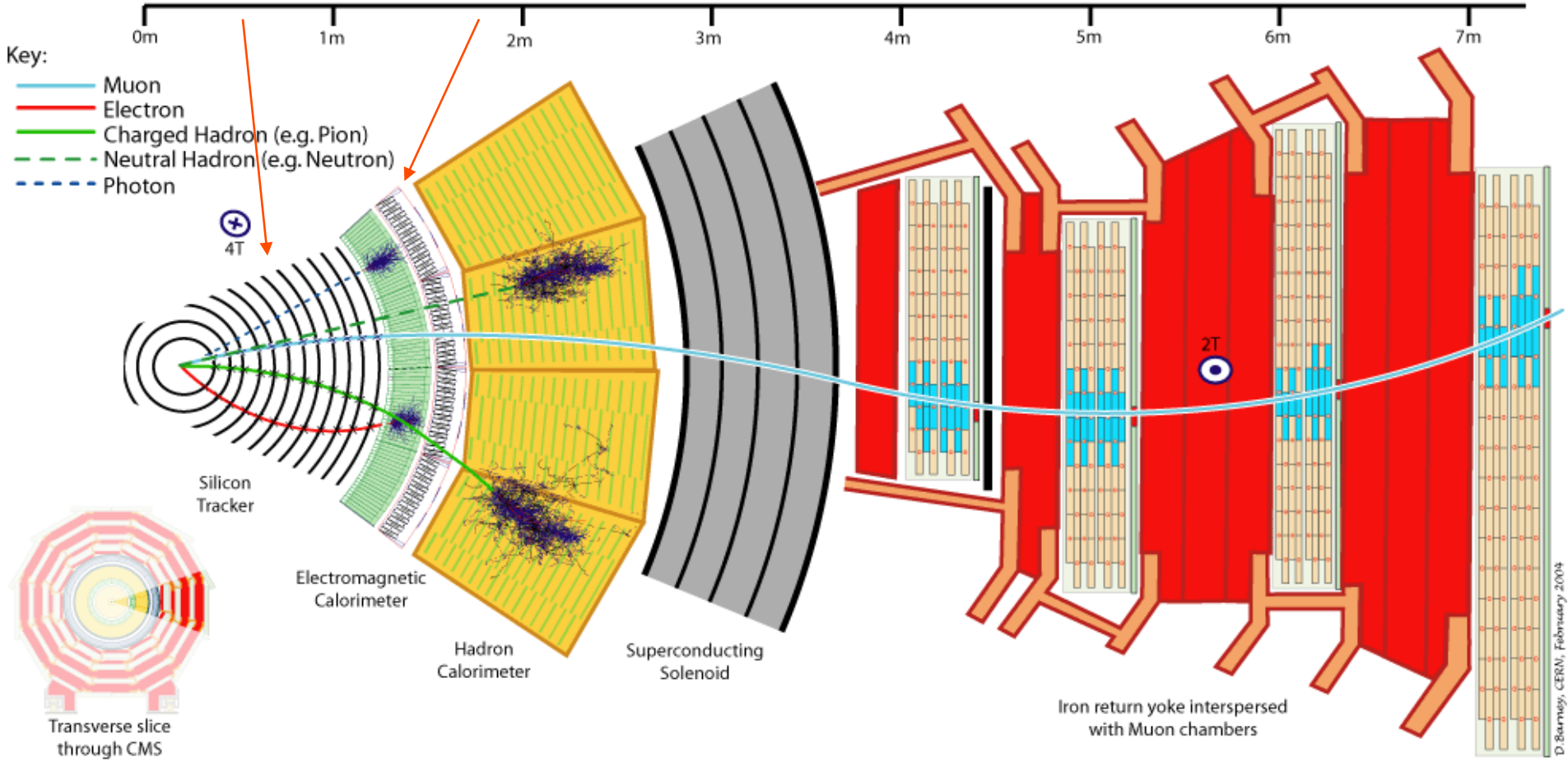
Rome , 20.9.11

Contents of this talk…

- Timeline of MG use in CMS during the last years and new-era use
- Validation plots mg4 vs mg5
- Description of our used workflow
- Open issues

$h^{\pm}$, $e^{\pm}$, γ, $\mu^{\pm}$ *inside* the barrel ($|\eta| < 2.5$)

Si Tracker + ECAL + muon-chambers



Si Tracker

Silicon micro-strips and pixels

Calorimeters

ECAL $PbWO_4$

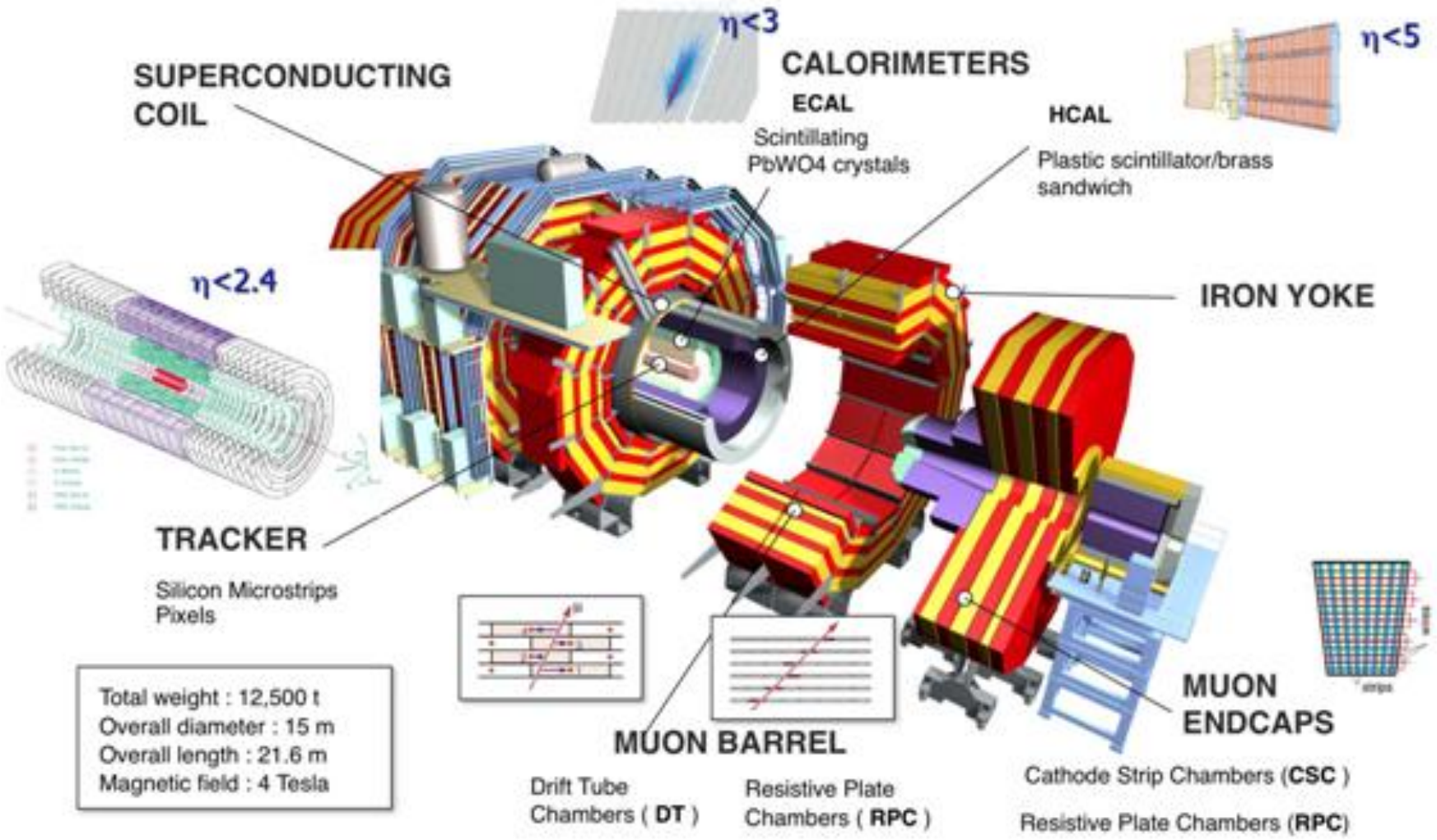
HCAL Plastic Sci/Steel sandwich

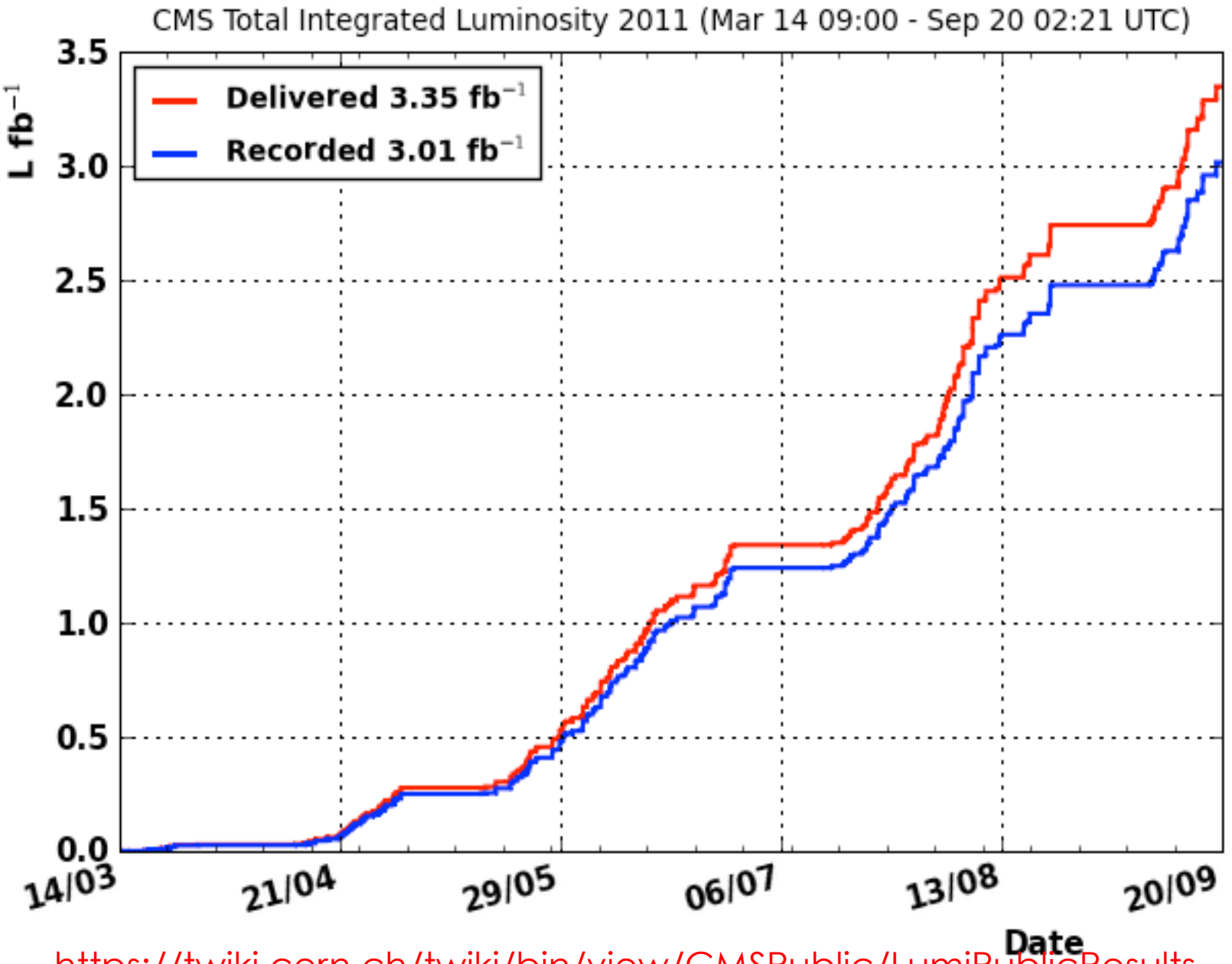
Muon Barrel

Drift Tube Chambers (**DT**)

Resistive Plate Chambers (**RPC**)

Main Parts of the **C**ompact Muon **S**olenoid (**CMS**)





https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults

Timeline

- Last year was a really busy one , as we had to cope with
  - CMS was supposed to run @ 8TeV for 2011 but that was changed after Chamonix ->7TeV – That means that we had to run...
  - The anticipated "new" MadGraph 5 version was still in beta
  - At this point, we were not sure if we should invest time and effort on the gridpacks mainly due to main different problems of "old" gridpacks
    - Old days workflow : After generation we wanted Pythia + DIGI + RECO in one go, BUT lots of failures
      - - T2 stucked unpacking thousands of gridpacks (1 per job)
      - - biased in phase space (too few events per gridpack problem)
  - Exercise:
    - What we now call *"standard wokflow"* include
      - Creation of proc_cards /run_cards + scripts to (auto)generate events+ post-processing+validation plots + RECO those samples
    - Nevertheless, we needed to test our scripts before we launch any large scale production
      - Efficiency varied in some of our farm clusters (CE)
      - Different ways of submission, PBS , gLite, CRAB , multicore, parallel…
      - Document all these at the end…
- Furthermore, we needed to do a 32 vs 64 bit comparison
  - Important for our new CMSSW releases
    - In principle, comparison only at Parton level is enough as the lhe card is the direct output of mg

Change to a new era

- After exhaustive discussions with MG authors, we concluded, that a closure test should ideally include the following processes
  - p p > W jets, w > l nu (up to 4 jets)
  - p p > Z jets, Z > l l, nu nu (up to 4 jets)
  - p p > t t~ jets (up to 3 jets)
  - p p > t j + jets $ W (up to 2 jets including one b)
  - p p > V V + jets (up to 2 jets)
  - p p > t t~ jets with decayed t/t~ (up to 1 jet)
  - p p > h + jets (up to 3 jets) with HIG=HIW=1 in HEFT
  - SPS1a SUSY p p > squark squark $ go and p p > go go + jets (up to 1 or 2 jets)
- In terms of
  - Physics : compare new mg5v1 vs mg4 Fall2010 vs data
  - Speed : Is the new mg5v1 faster ? Can we trust again the gridpacks? Thus we must
    - Produce tarballs + gridpacks for most important processes
    - Monitor tarballs & gridpacks outputs in terms of size and time

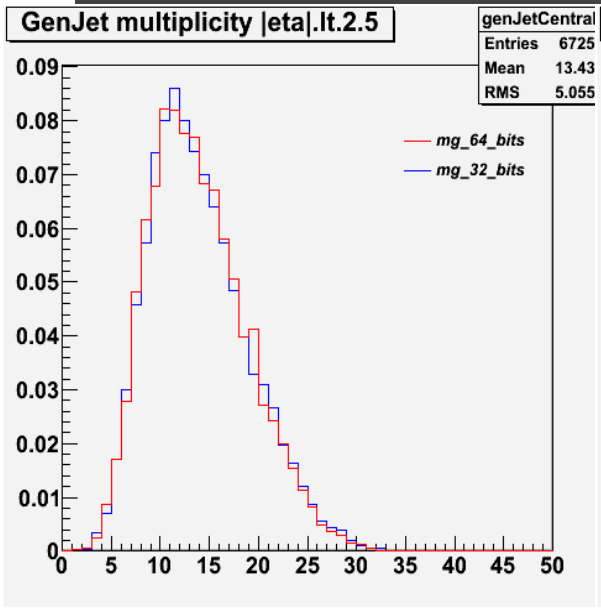
Example of processes used to verify Gridpacks robustness

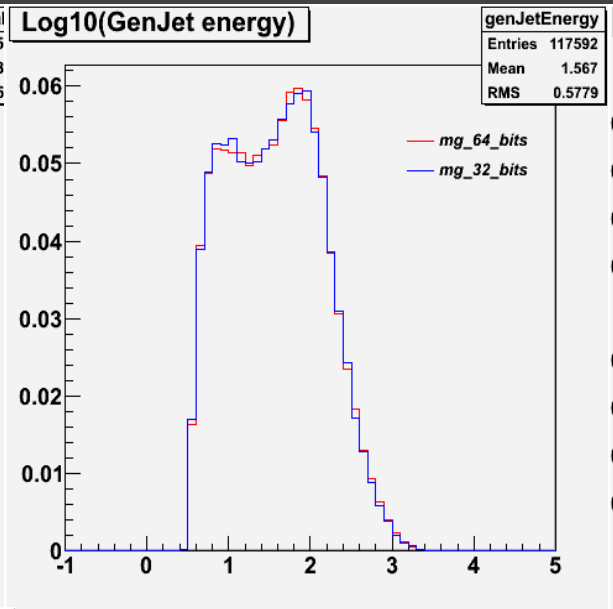
- MG4
  - @LYON
  - v.4.4.50
  - Cluster
  - slc5 gcc 4.34
- MG5
  - @Brussels
  - v0.51
  - Locally
  - slc5 gcc 4.34

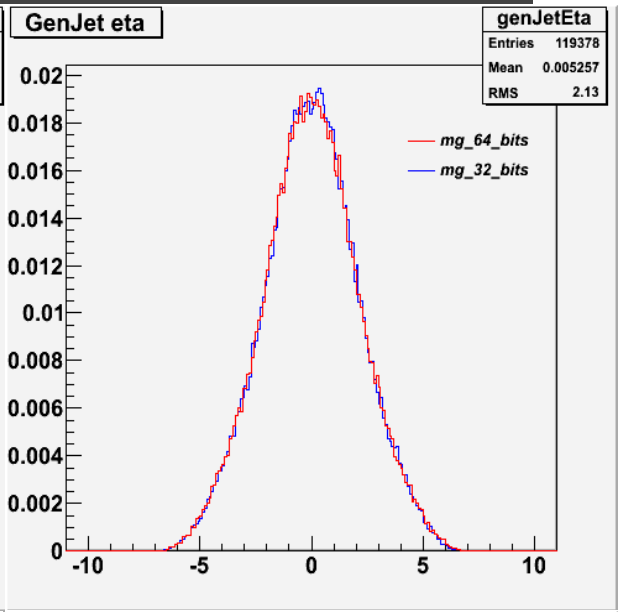
- Running time varies from minutes to..days!!
- On top, run the official validation workflow

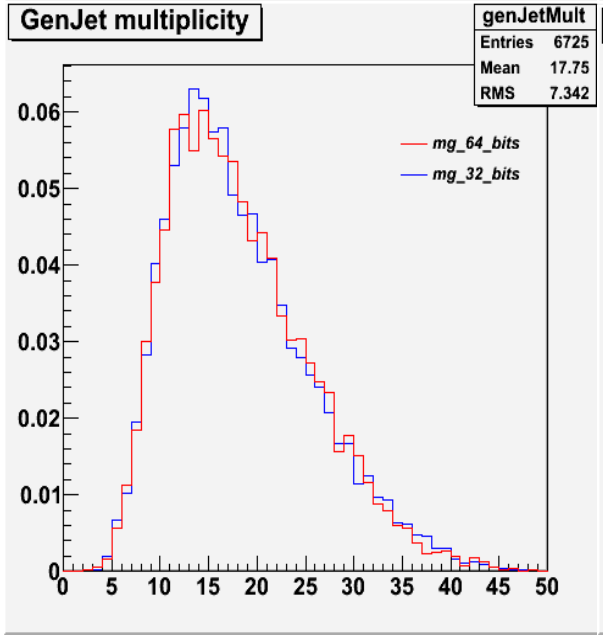
Process	Tarball (100k)	Validation
QCD	✓	✓
W (-> l ν)+ jets	✓	✓
Z / γ* (-> l+l-) + jets	✓	✓
Z / γ* (-> l+l-) + jets (smbmass)	✓	✓
Z (-> νν) + jets	✓	✓
V(-> l1l2)jj	✓	✓
V(-> l1l2)+ QQ~+jets	✓	✓
W(-> lν)+ c +jets	✓	✓
γ + jets	✓	✓
VV(-> 4l) + jets	✓	✓
γ V(-> 2l) + jets	✓	✓
γ γ + jets	✓	✓
W(->lν) γ γ	✓	✓
Z(->νν) γ γ	✓	✓
bb~ + jets	✓	✓
tt + jets	✓	✓
tt + Z + jets	✓	✓
tt + W + jets	✓	✓
single top – all channels	✓	✓
VQQ biased	✓	✓
W(ll same sign) + 2jets	✓	✓

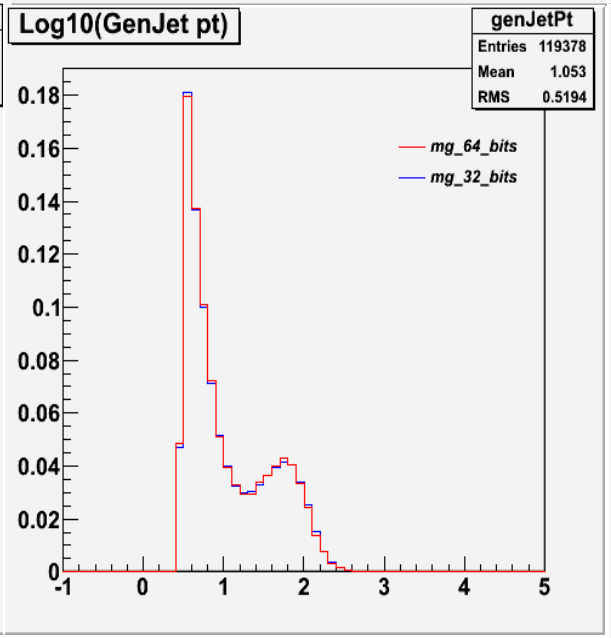
Validation Plots 32 vs 64 bit (ttbar) with MG5

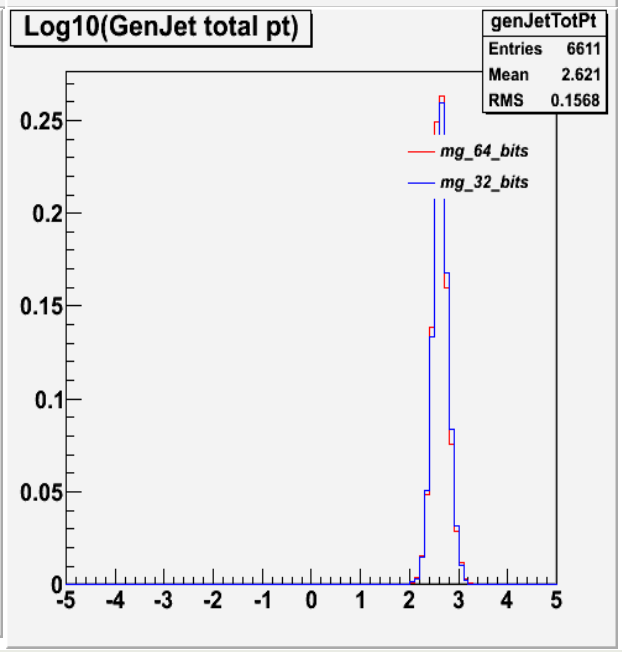












Step 1- Definition of processes

- Sometimes is easy,
  - p p > W jets, w > l nu (up to 4 jets)
  - p p > Z jets, Z > l l, nu nu (up to 4 jets)
  - p p > t t~ jets (up to 3 jets)
- sometimes it cakes weeks of discussion
  - p p > V V + jets (up to 2 jets)
  - p p > h + jets (up to 3 jets) with HIG=HIW=1 in HEFT
  - VBF processes
  - CKM matrix used in smzerobmass ?
  - Sm vs smzerobmass decision
- Up to now, MG has been mainly used for bkg samples – This probably will change in the near future – Already plans/ discussion for MG+SUSY signal / Simplified Models (SMS)

-- AlexisKalogeropoulos - 03-May-2011

MadGraph 7TeV Summer 2011 Production Page

- MadGraph 7TeV Summer 2011 Production Page
  - Foreword
  - How to read this page
    - Gridpack creation procedure
    - Event generation procedure
    - How to get the size/event and time/event
    - QCD Samples → >600M events
    - Single Vector Boson production → >2B events
    - Photon + X
    - Di-Photons
    - Vector Boson + heavy flavors
    - Multi Vector Boson production
    - Top production → >100M events
    - Exotica/4rth Generation production
    - Standard Model Samples for Theory/Modelling Systematics Studies
  - top
  - W/Z+jets
  - Wbb/Zbb

Step 2- Creation of Gridpacks

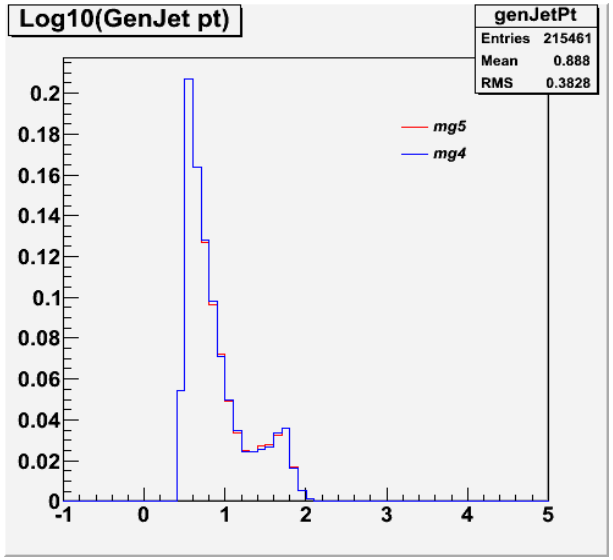
- After investing time validating gridpacks in terms of performance, we concluded that should be our workhorse
  - Define process and run cards
    - Make gridpacks
      - Mainly PBS or parallel running
      - Parallel running : In most of the cases unavoidable
        - Limitation : not much of a freedom, ie tighted with a specific setup :
        - Solution -> We have built our own LSF wrapper around PBS commands which clearly constraints us as in general, as assuming a given batch scheduler to be used, does not allow to export the tool(s) in other environments and can limit the interest in its massive usage.
        - Creation of gridpacks over the gird – **FORGET IT!**

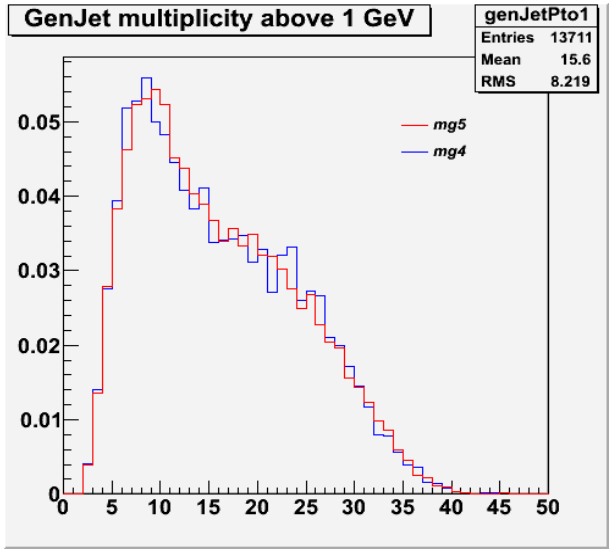
      Job failure is treated on a statistics basis in these environments, assuming all jobs are just a copy of what needed, while this is not true in Madgraph, where each job is unique and needed.

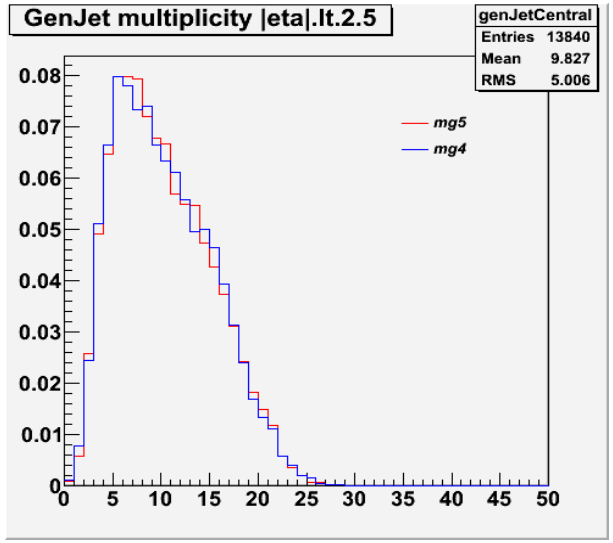
Step 2- Creation of Gridpacks

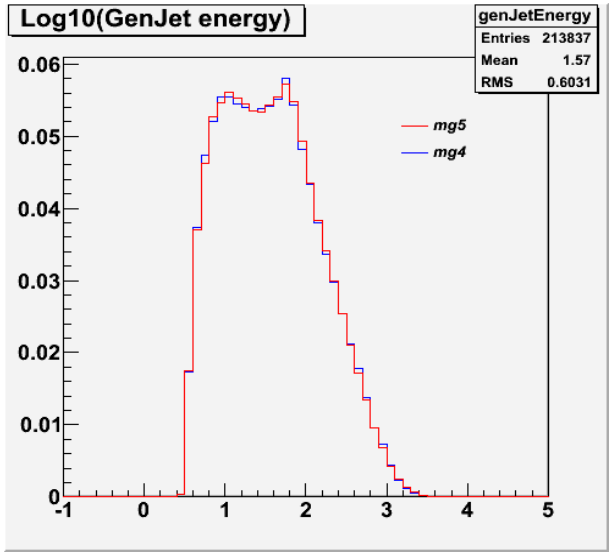
- Time to create a gridpack varies from hours (qcd for example) to several days (Z+up to 4 jets)
- The x-sec we get out of the gridpack is a very good estimation in order for us to finalize our request
- Once we set up our environment, no peculiar issues faced so far
- However, we would like to have a more clear picture what gridpacks does in every step (implement different debug levels ?)
- For a well planned generation of the order of few thousand events, we use plain "tarballs"

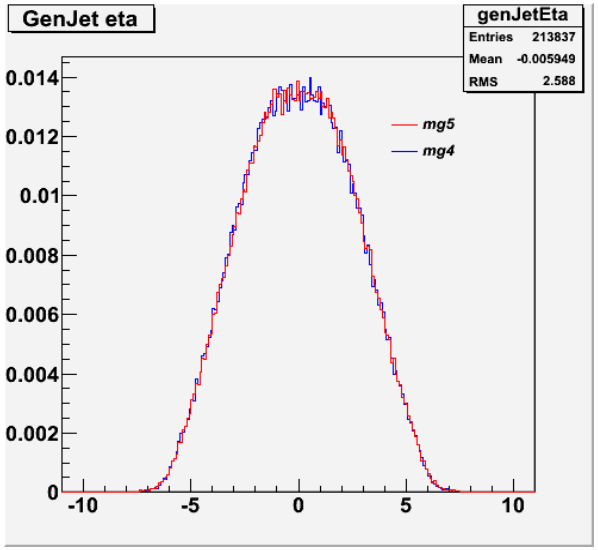
mg4 vs mg5 - QCD_{100_250} / Parton Level

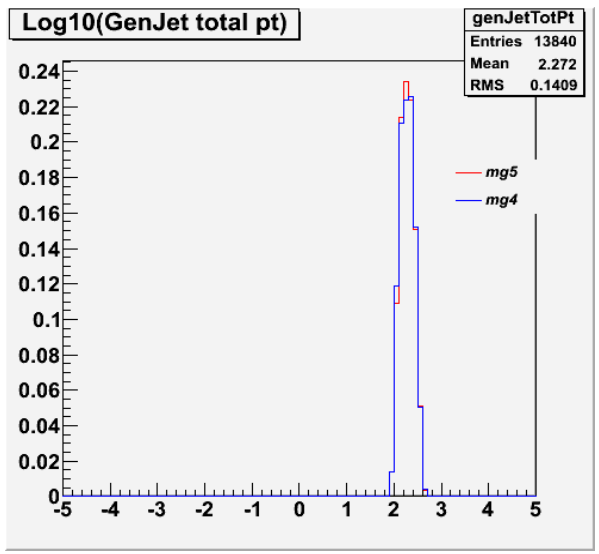




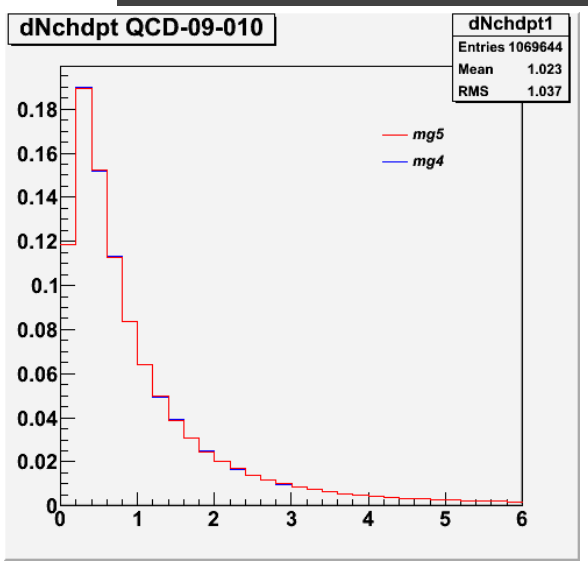


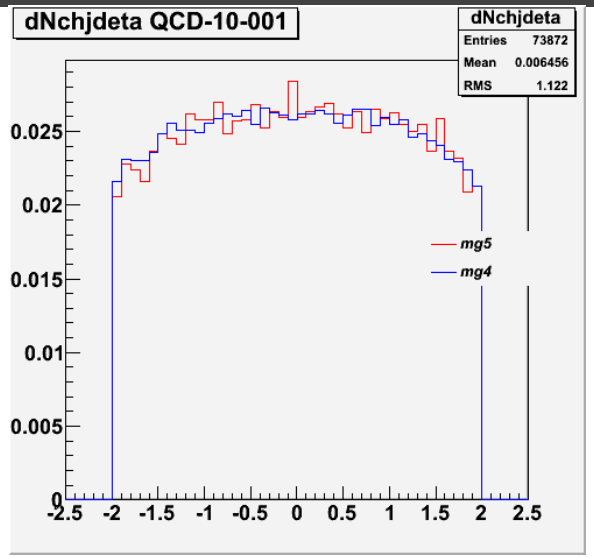


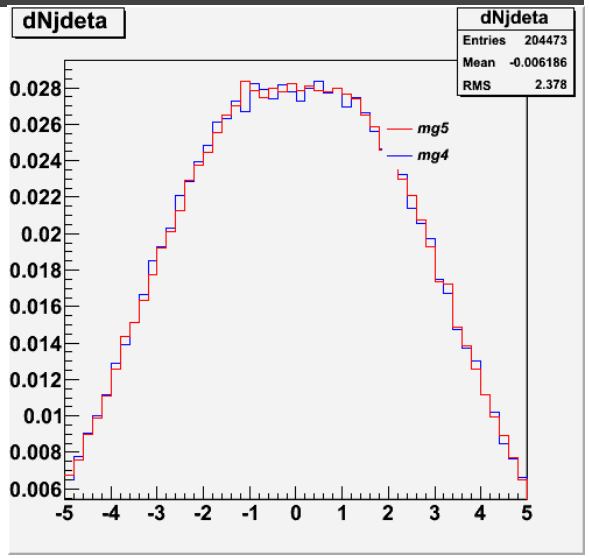


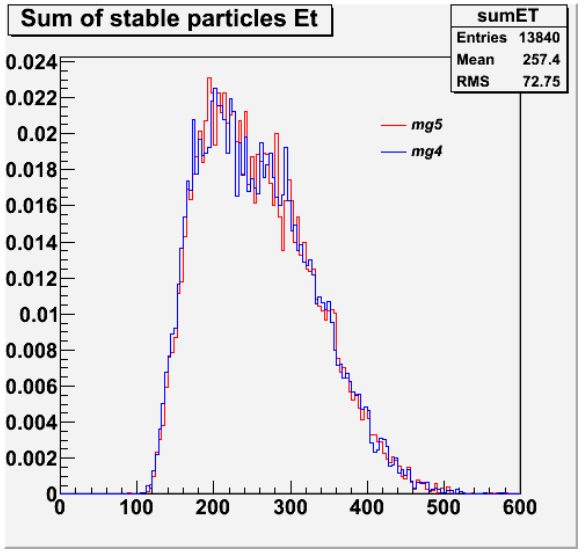


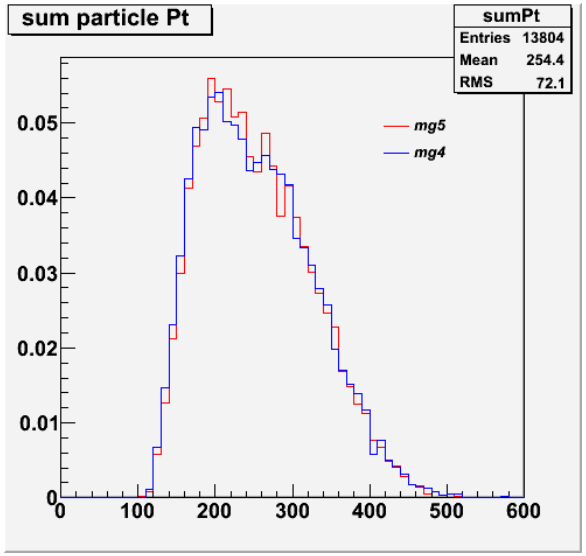
mg4 vs mg5 - QCD_{100_250} / Hadron Level

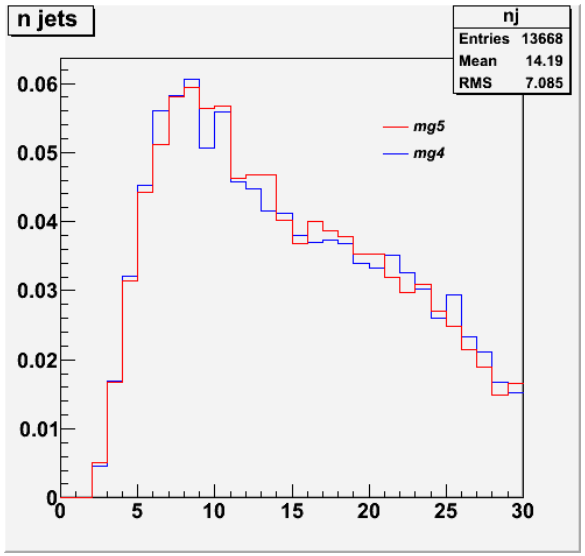




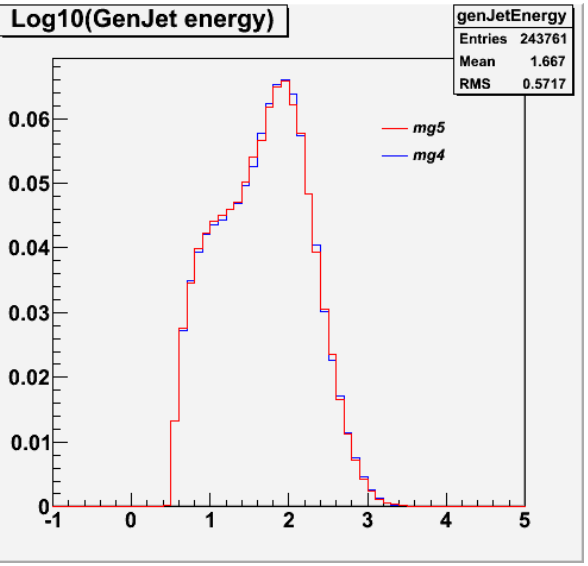


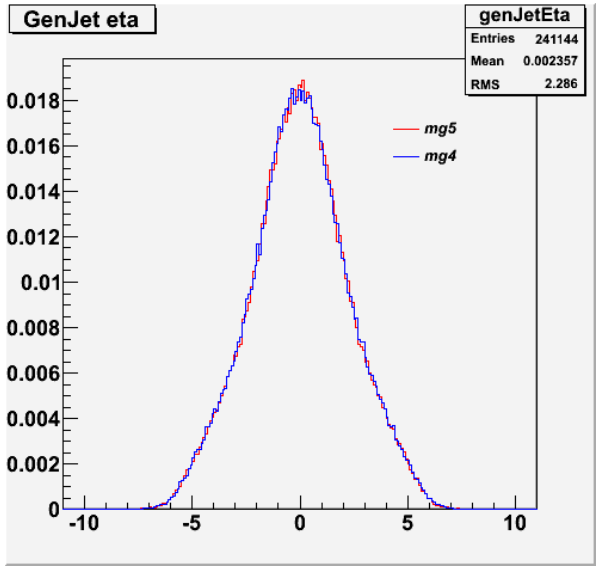


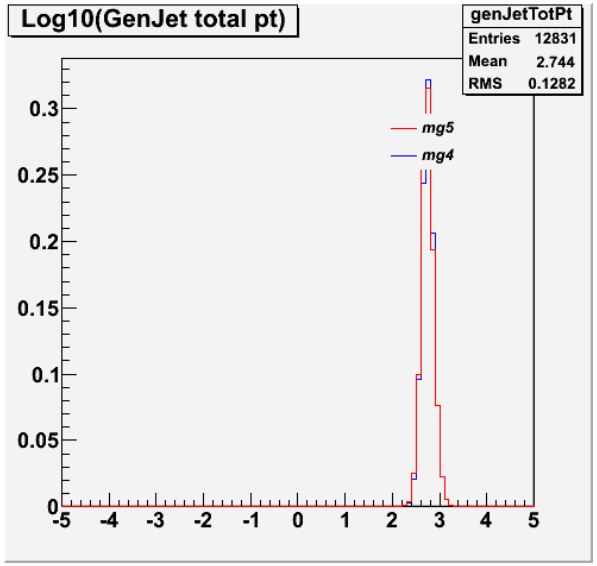


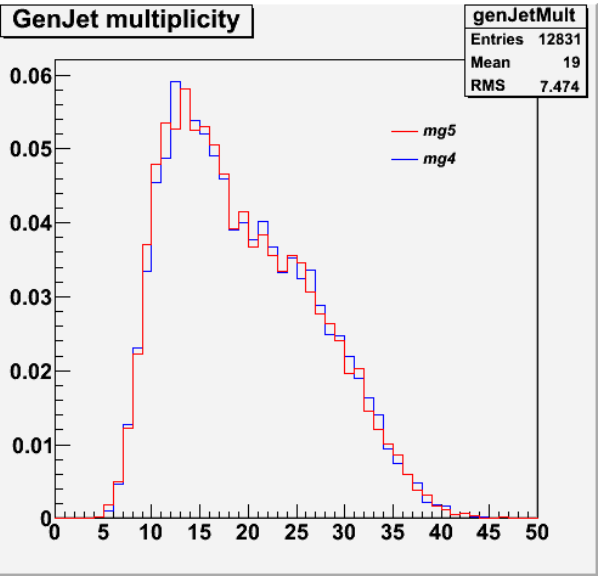


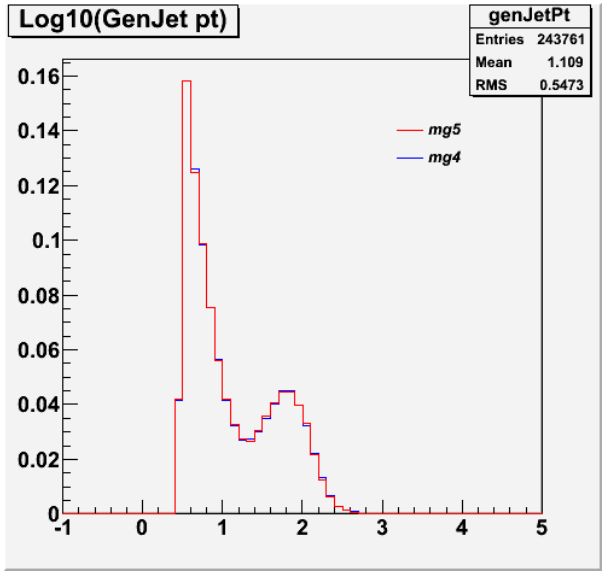
mg4 vs mg5 Zttbar / Parton Level

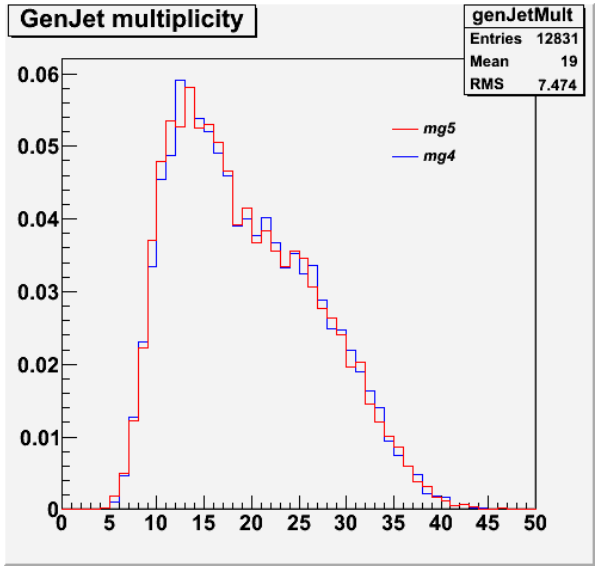




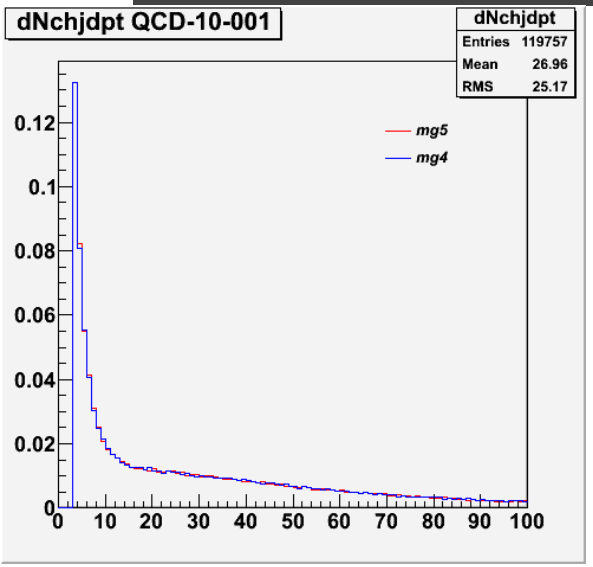


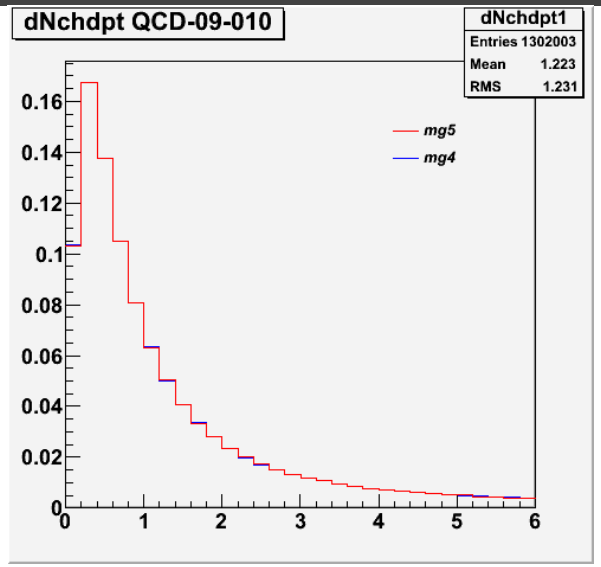


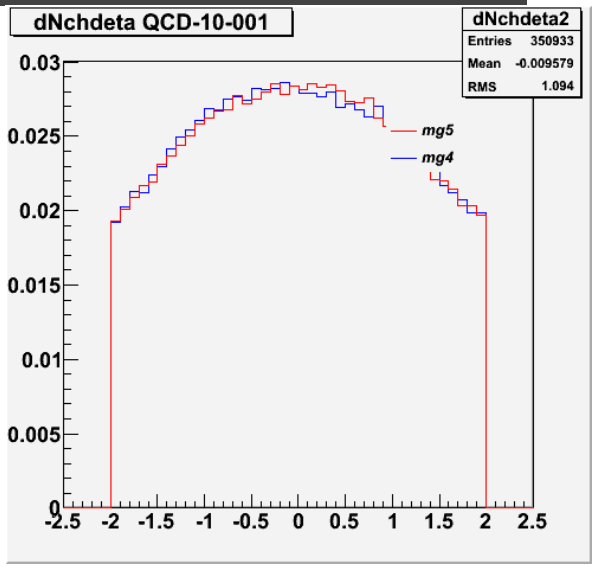


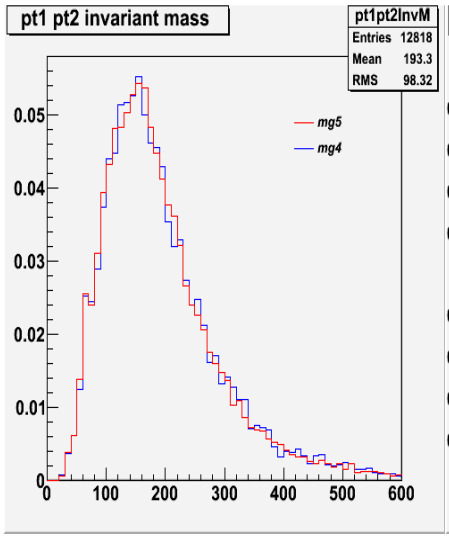


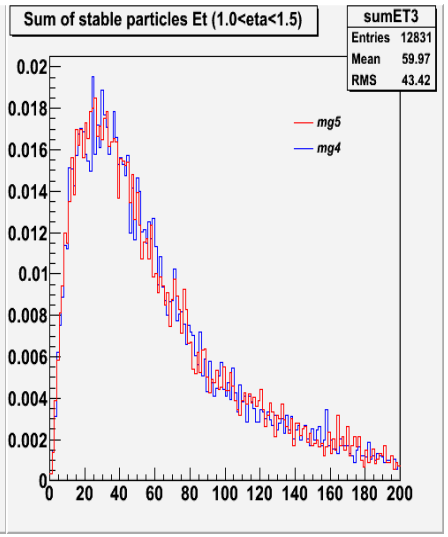
mg4 vs mg5 Zttbar / Hadron Level

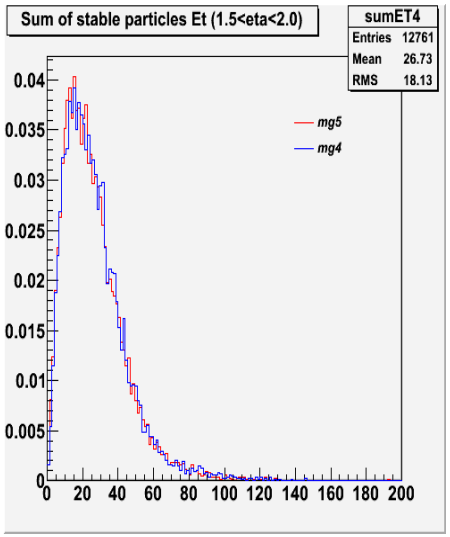


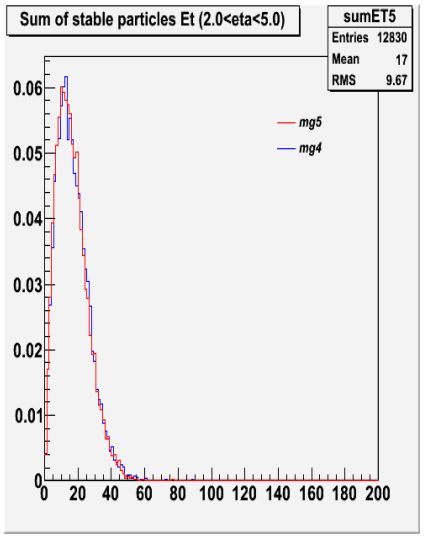




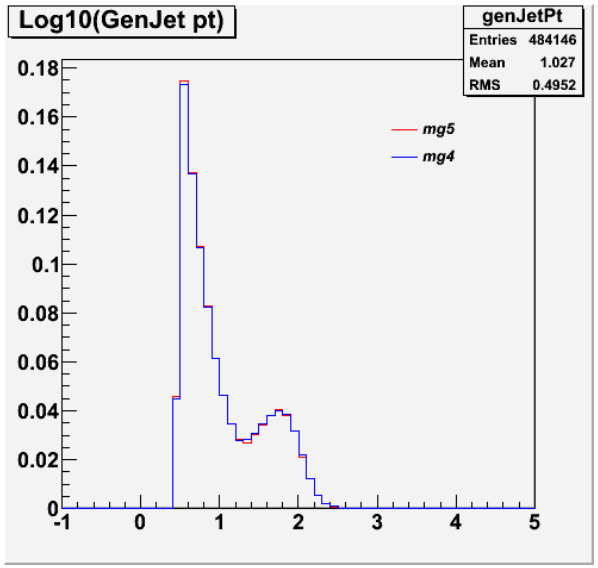


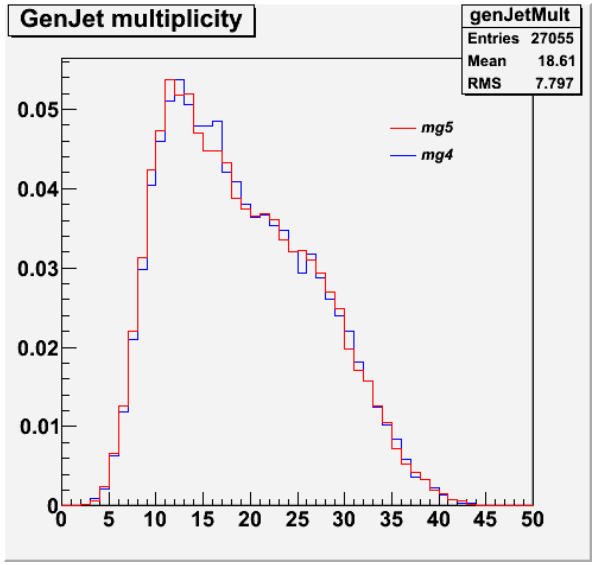


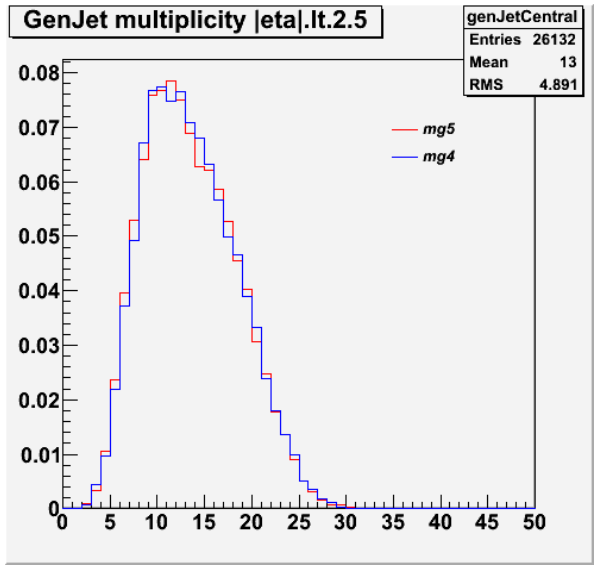


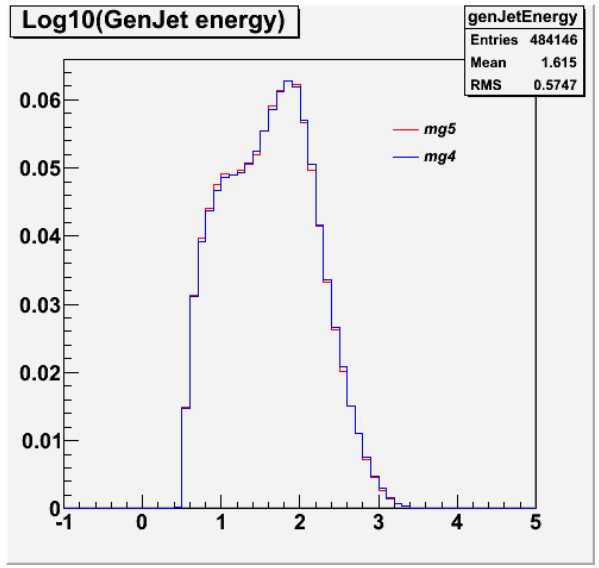


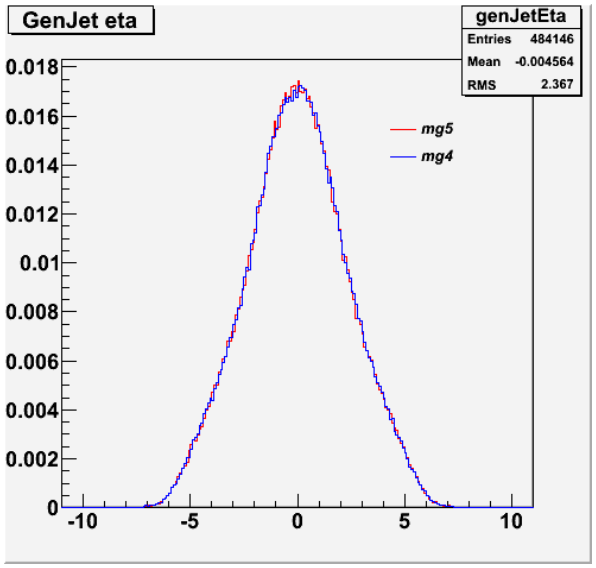
mg4 vs mg5 Ttbar / Parton Level

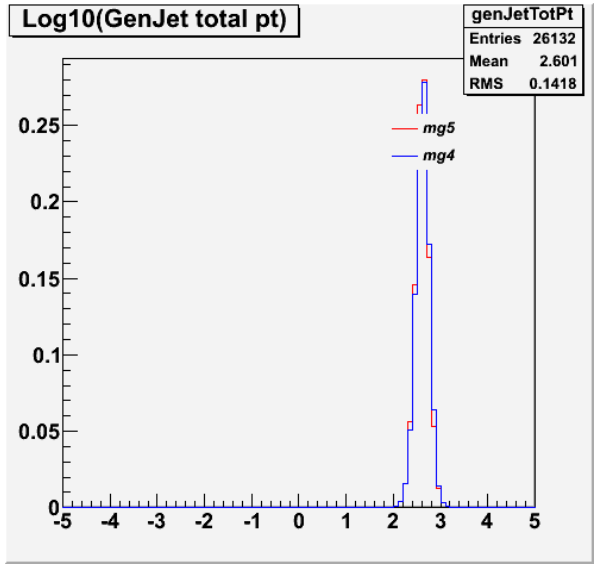




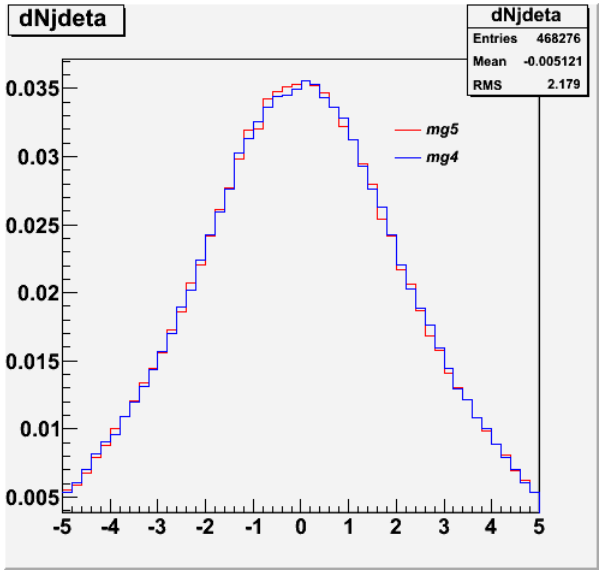


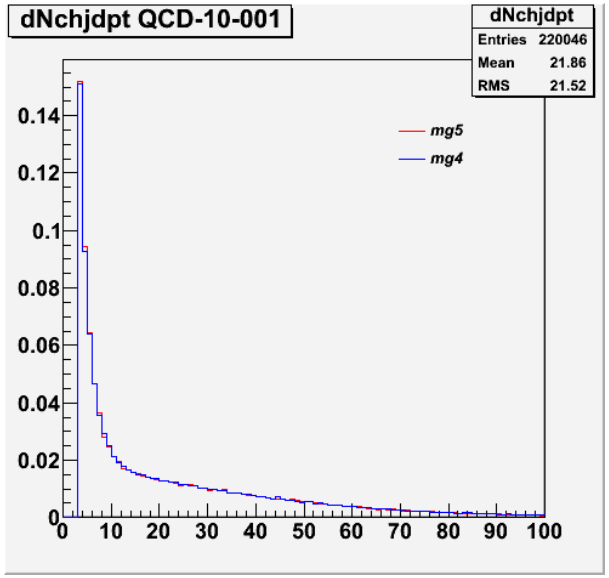


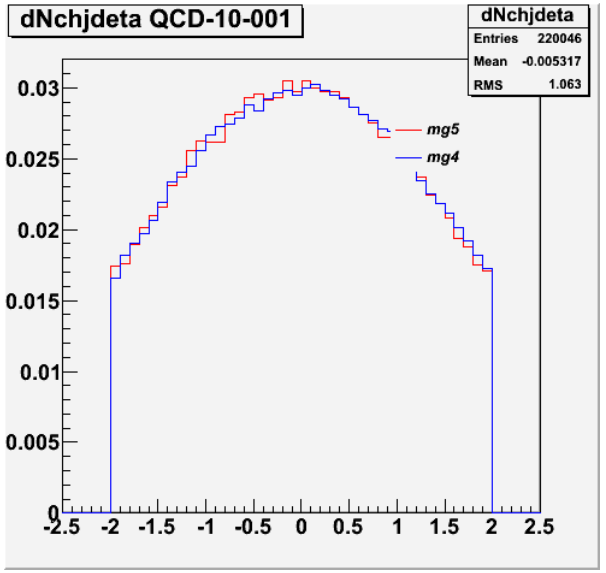


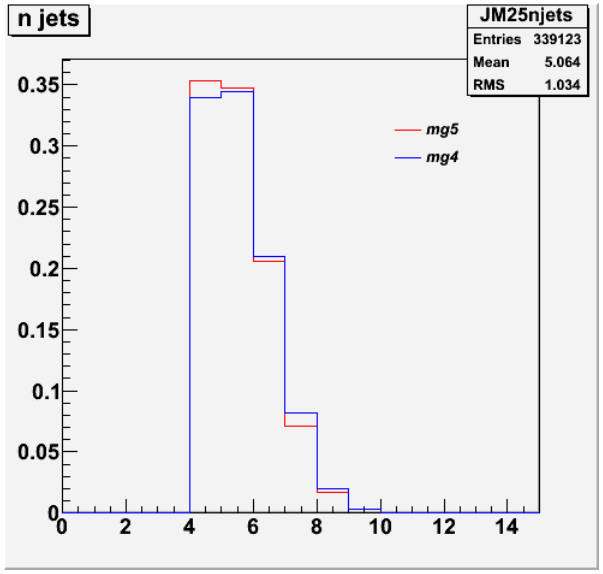


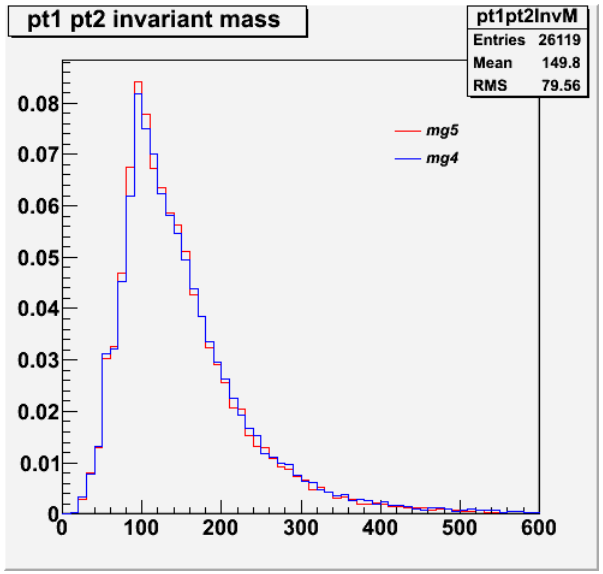
mg4 vs mg5 Ttbar / Hardon Level

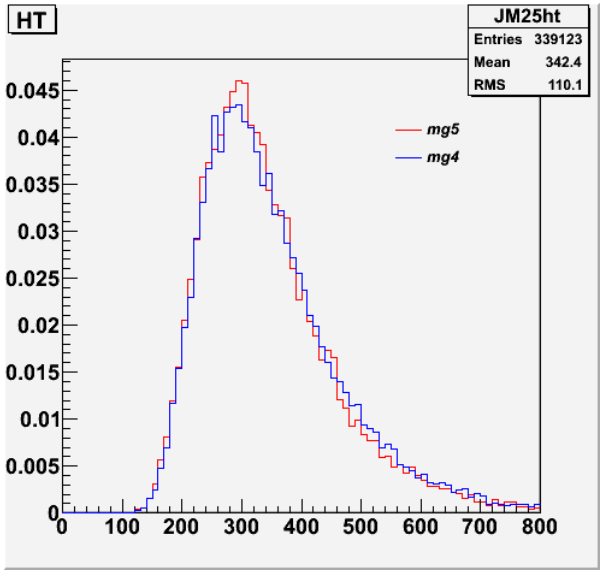












Step 3- Generation of Events

- Generation of events is one of the most important step…up to now we have generated **>3B** of events!!!
  - We used gLite in the beginning along with PBS submission, but we changed to CMS Remote Analysis Builder aka CRAB – Why ?
    - Is part of the official CMSSW software (support , upgrades etc)
    - Really easy to use – all you need is a .cfg file in which you define executables, (one can run ANY executable)parameters etc etcl
    - Our scripts load a CMSSW environment, so that we are sure about the environment in which we run the executable (more stable, trace-able)
    - Do not have to worry for the random seed number (thus duplicated events), as you just need to define just the first one
    - Output in a SE of your choice (local/CERN)
    - Different exit codes for different failure reasons
    - Traceback logs in case of failures

Step 3- Generation of Events

- But timescale is essential : From pressing the button, after how many hours/days should we get results ?
  - We faced problems for processes that need more than 3 days of running (CMS official policy kills jobs on the grid after 72h) –
    - Way to solve : run on fewer events (10-20k) usually does the trick
  - multicore NOT an option for us..
    - grid distributed batch systems where the CPU time limit of a single job is a relevant factor.

- *Fewer events* known issue
  - Although this has been observed and discussed in the past, we still experience some times outputs containing fewer events than ordered although the job was properly closed/finished..Ideas ?

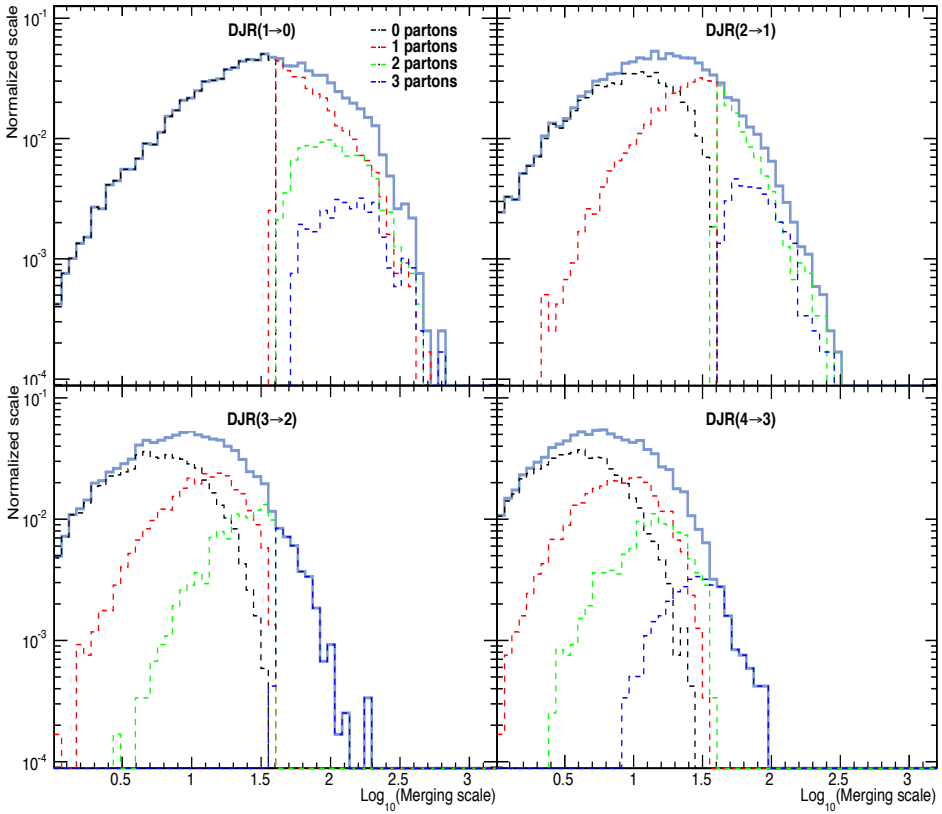
Step 4- PostProcessing of Events

- We need to verify the content of the LHE files and make them viable for CMSSW environment (ie SIM/DIGI/RECO steps)
  - We introduced several layer of post-processing tools like
    - Correcting the kinematics of leptons as usually we use massless leptons for simplicity while generating events
    - Putting the full particles-history like for the Z off-shell resonances
    - Check that the seed number reported IN the LHE file is actually the one that has been used while launching the job
    - Verifying the xsec of each LHE everytime and put different thresholds for this – Method of "relative ratios" wrt to the full xsec and #of <event> block for each given included process
    - One of the most important steps is to properly obtain the best-matching parameters (qcut) –via DifferentialJetRatio plots or *"xsec vs qcut stability"* – (details in next slide)

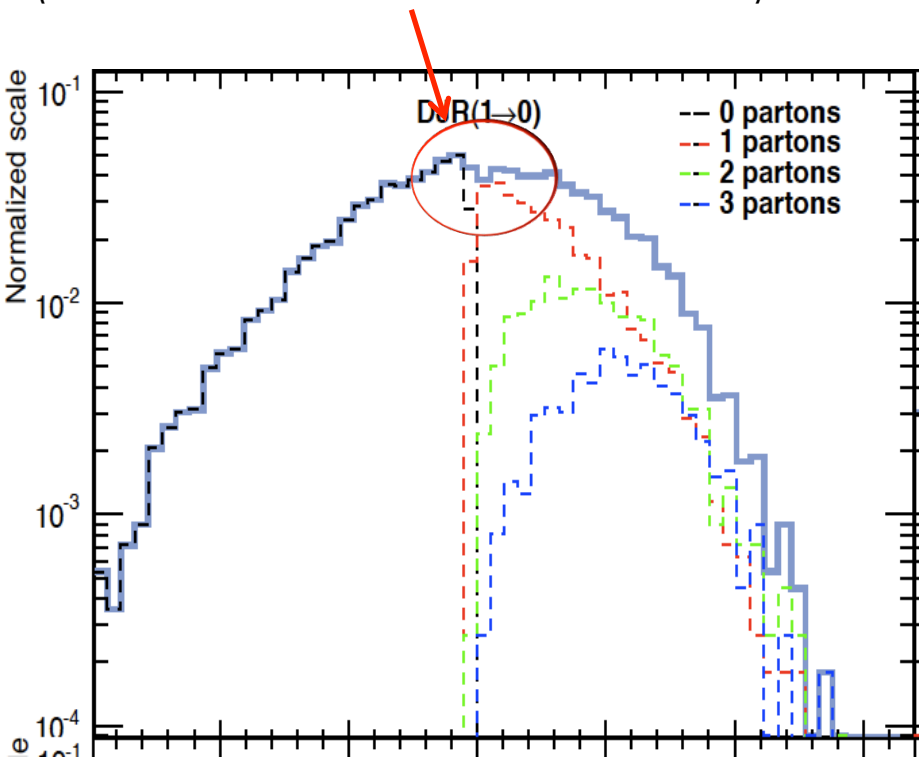
Finding the optimal qcut…

- Qcut vs xqcut
  - Obtain the DJR, ie how well a matching was performed – Transition from N+1 - >Nth jet clustering

Good DJR /ttbar qcut=40GeV, xqcut=20GeV

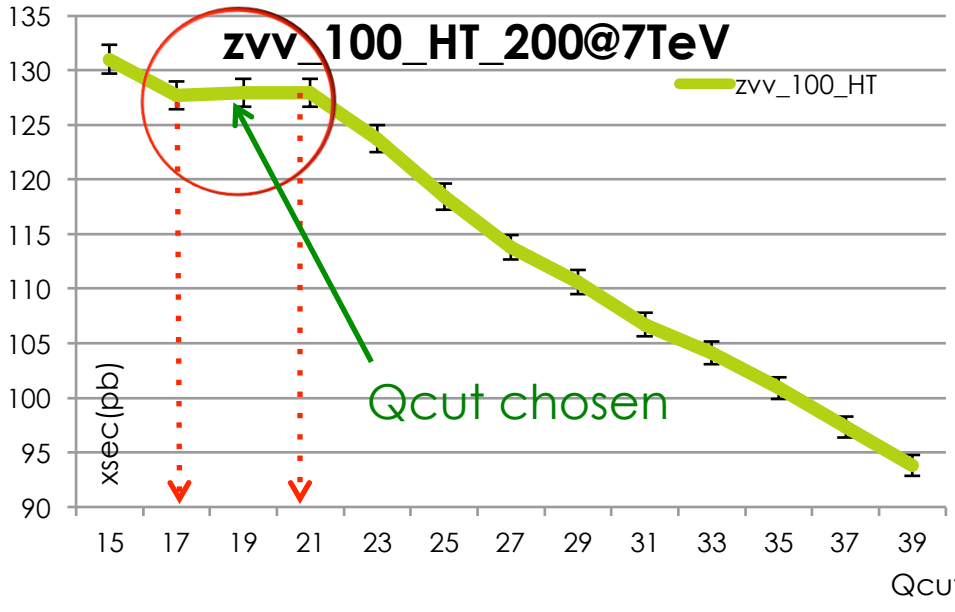


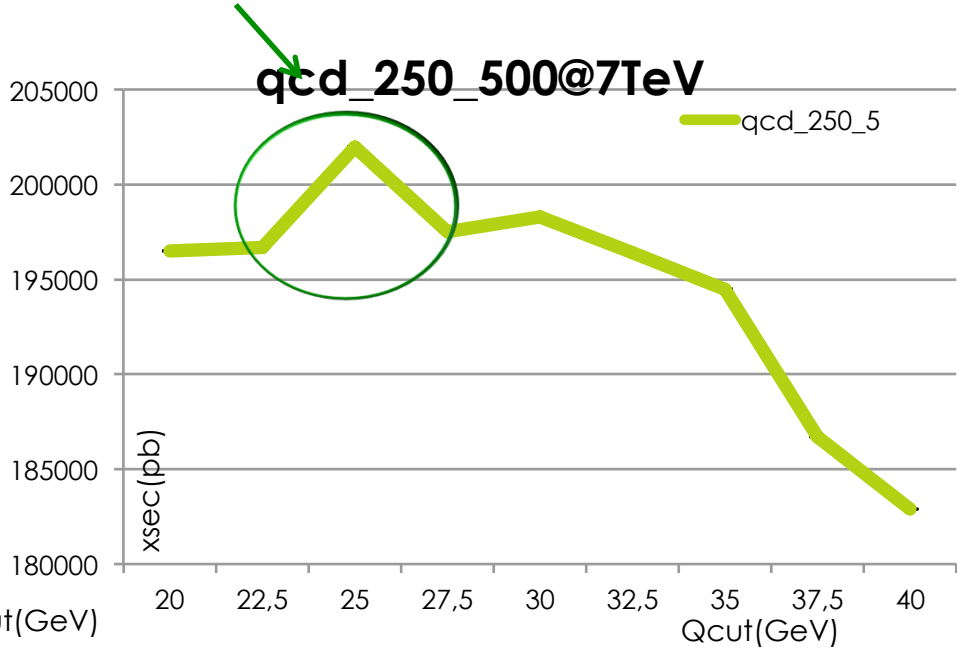
Bad DJR /ttbar qcut=30GeV, xqcut=20GeV
(not smooth transition from ME->PS area)



Finding the optimal qcut…

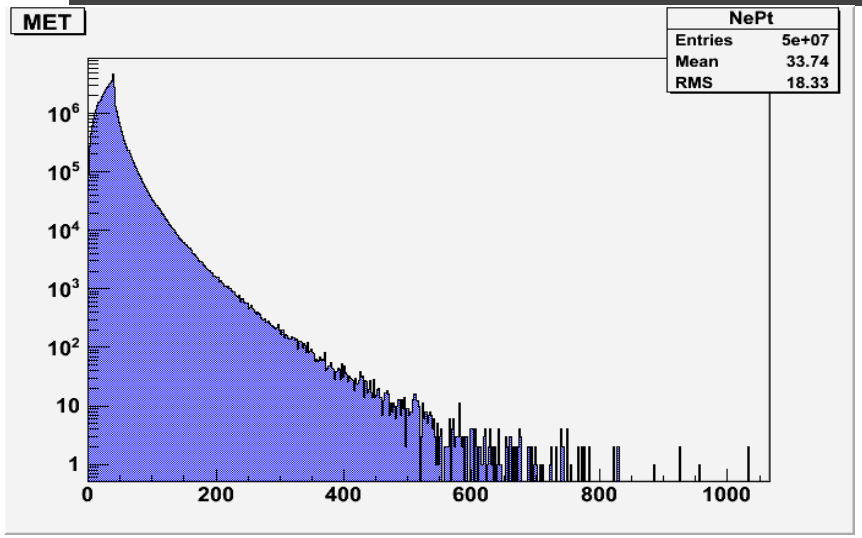
- Qcut vs xqcut method
  - Thumb rule used in the past qcut=1.5 * xqcut is clearly not safe at all due the new chosen tuning Z2, ordered showers, pdf reweighting etc etc
    - Difficulty to obtain the djr plots for samples not including the zero-multiplicityjets process – Example for the zvv_100_HT_200 the low HT cut kills the 0- jets events
      - More plots here http://mon.iihe.ac.be/~alkaloge/7TeV_Summer11/DJR/
      - Probably, we will make an internal note- which can serve as guide/workbook for future MC productions addressing all these issues
    - **Best plan is to draw xsec vs qcut and find the "plateau"** ….but not always trivial though….(*indication of wrong xqcut as well?*

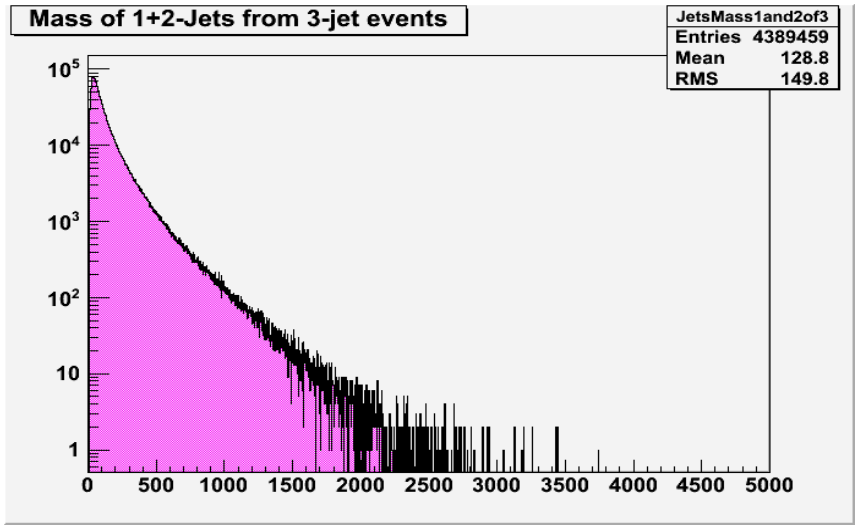


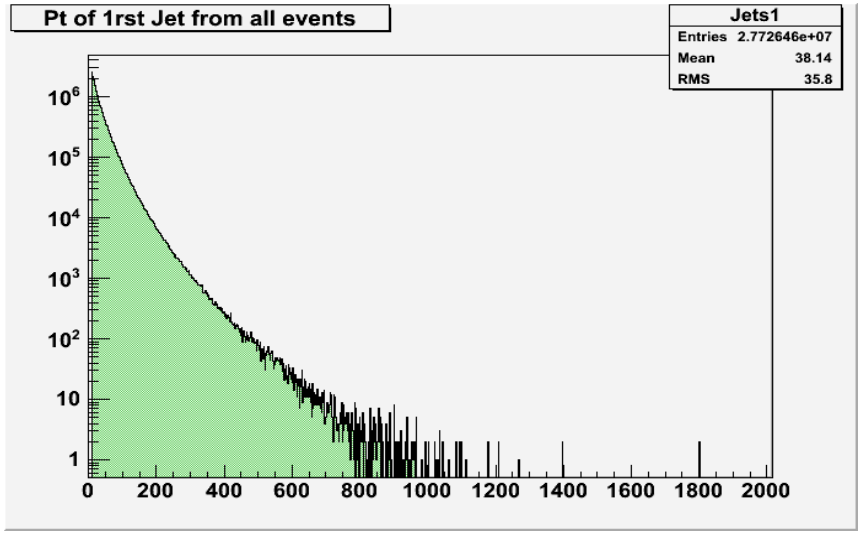


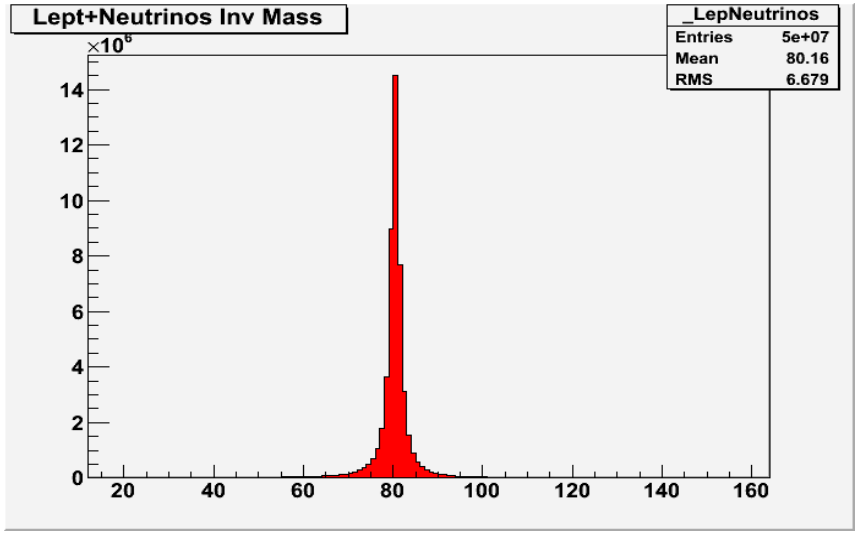
We need these plots for the systematics samples as well

How well do we perform?? Do we miss something on the tails ?





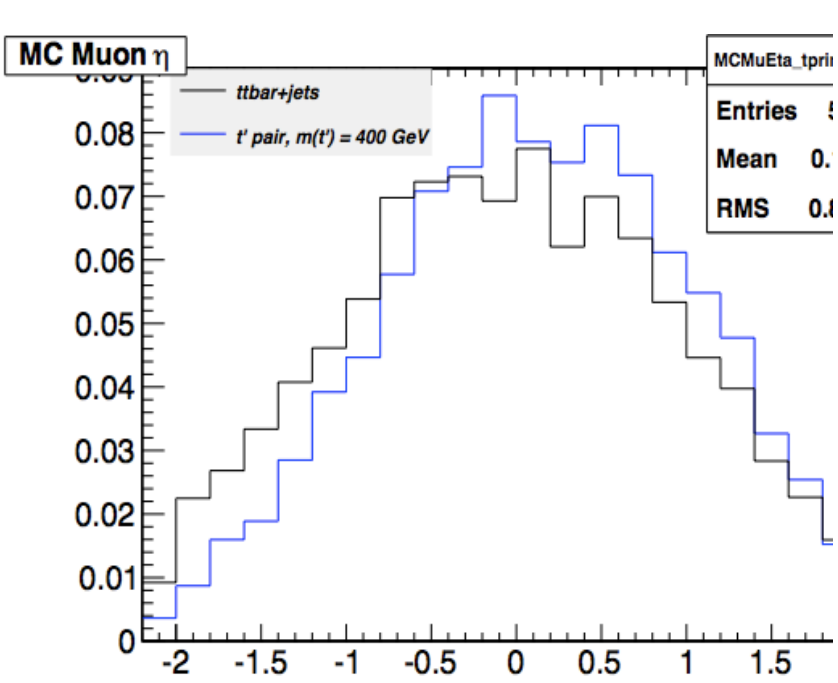


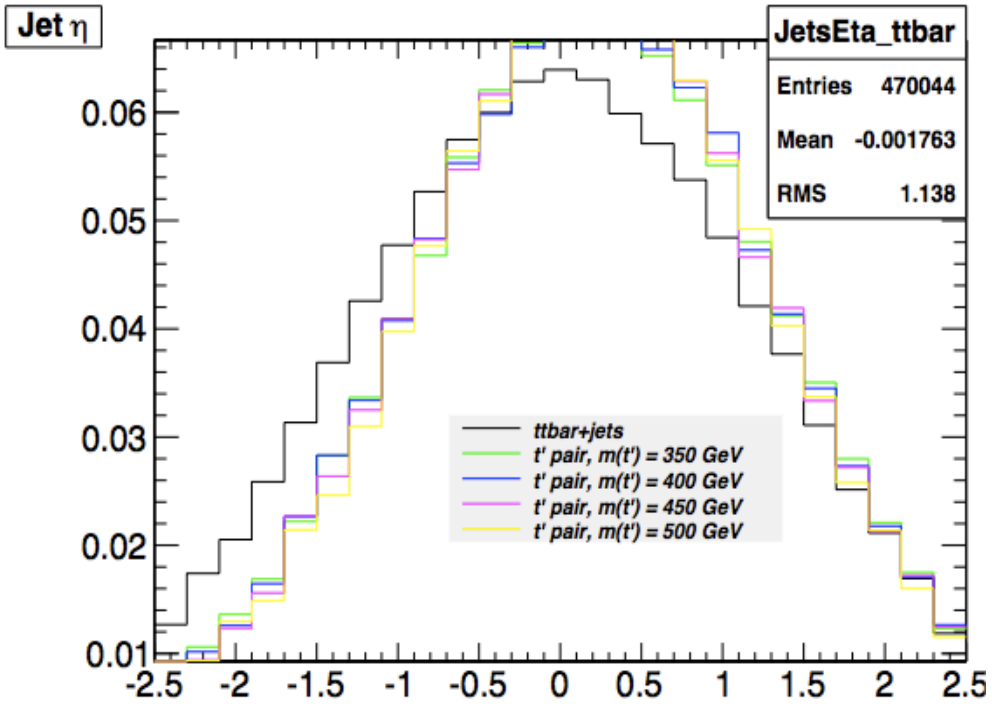


We have standardized more than 50different distributions (Jets invMasses, Kinematics These plots are actually a proof that MG5 works…

Some problems…

- We faced an eta asymmetry issue with some t't'~ samples:





- After some digging, turned out that the problem was due to the combination of use of massless leptons (sm) + pythia as LHE level reveled no problem at LHE level even with the massless model

- Use of the 4th_GEN model solved this, although looked to be a Pythia problem – So, can we trust massless models + Pythia ???

Further discussion

- Several "open" issues
- Matching scheme for singletop samples
  - Do we need both t_2_2 and t_2_3 processes ? Do we understand how to do the matching here ?
    - Use of a CDF like "custom" code which makes the "mixing" for the t_2_2 and t_2_3 as follows
      - Pt sectrum of these two processes are superimposed
      - Cut-off the common region, ie matching in "parton" level
- Matching issue with bb~ jets
  - This affects one of our heavier samples (4 HT bins) – We already got a patched version and now on the phase of testing
- Madweight
  - Starting discussions if we could use it to make systematics samples (right now, just variation of scales)

Summary

- MadGraph is our main generator and was chosen mainly due to
  - Simplicity to add new processes
  - aMC@NLO soon (we hope!!!)
- Extensive use of Gridpack
  - Patience pays us back– more than 3B events so far in < 6months
  - Several layers from point zero to RECO events – Sophisticated procedures ensure that we provide robust MC models to our community
- Given this opportunity, CMS would like to formally thank MG team and especially Johan, Fabio, Tim and Claude for their continuous support – We greatly appreciate this and counting on for our future plans

Summary

- MadGraph is our main generator and was chosen mainly due to
  - Simplicity to add new processes
  - aMC@NLO soon (we hope!!!)
- Extensive use of Gridpack
  - Patience pays us back– more than 3B events so far in < 6months
  - Several layers from point zero to RECO events – Sophisticated procedures ensure that we provide robust MC models to our community
- Given this opportunity, CMS would like to formally thank MG team and especially Johan, Fabio, Tim and Claude for their continuous support – We greatly appreciate this and counting on for our future plans

Never worry about theory as long as the machinery does what it's supposed to do.

Robert A. Heinlein