CMS Barrel Muon Workshop

Physics Department, 28 A 102, Aachen (Germany),

April, 29^{th} 2004

First experience on DST analysis at T1/T2 centers

Physics case study: $H \rightarrow WW \rightarrow 2\mu 2\nu$

Stefano Lacaprara, I.N.F.N. and Padova University

Outline

- Introduction,
- Analysis models on DC04 data,
- Future plan for data access,
- Physics case: $H \rightarrow WW \rightarrow 2\mu 2\nu$,
- Conclusion

Introduction

- During last two months Data Challenge $'04$
- Goal: demonstrate the capability of the CMS computing system to deal with events at 25 Hz in a distributed way
- Need of many millions of MC data, to be used also for physics study
- Large effort to setup best possible generation, detector simulation and response, reconstruction (from summer 2003)
- The MC production (aka PreDC04) was long, painful and not yet ended (only a fraction of event required by PRS have been fully processed)

- All MC data available at CERN (T0)
- Process MC data at 25 Hz to produce DST (at T0)
- Ship DST (plus streams) to "interested" T1 (eg CNAF, PIC, RAL, FNAL)
- Further move some DST from T1 to T2 (CNAF$\rightarrow$LNL)
- Lots of files to be moved around ($\mathcal{O}(10^6)$): hard bookkeeping, bottleneck, etc, ...
- Last step is **accessing data** for analysis
- Still missing/on going/painful !!
- Not yet clear the "CMS analysis model": ie how to access data by the final user
- "Fake" or "real" analysis use-cases

Fake Analysis

- Base idea: synchronous with data arrival
- As soon a new run fully arrive at Tn, run a "standard" executable on it
- Implementation is very tricky: not meant (at least at the beginning) to be real (name says much...)
- Every run is not attached to MetaData, so they are fully independent. Not possible to run on a full dataset (or on a fraction of it), only run per run. "Deep Winter Mode" (aka Ice Age mode)
- Need real expert help, usage of several COBRA tools, parsing results, setup proper `.orcarc`, etc...
- Set up for PRS b/τ and μ: running on several datasets

- Pros:
  - it works! (huge effort needed)
  - can run a defined analysis (== executable) as soon as new data arrive
  - May be a use case if you just need to increase statistics for your analysis
- Cons:
  - ⋆ definitively not for generic user!
  - ⋆ Very complex!
  - ⋆ not suited for developing analysis
  - ⋆ not easy (or even possible) to run many times on full dataset, only *incremental* analysis
  - ⋆ Large overhead for smallness of analyzed set of data

Real Analysis

- Base idea: access the data as you (generic, average user) always did
- No need of special expertise: just what the ORCA tutorial teach you
- Access dataset as such (not just a bunch of runs, but a well defined collection of them)
- Access full available data (or a fraction of them) as many time as you need (cpu power only limit!)
- Not necessarly real time analysis, but not too much delay either!
- Guarantee access to all users to all datasets (at least in principle)

Needs

- Data!
- Attached MetaData (collection of runs)
- Up-to-date catalog (where the files are)
- User access to farm where data are
- Needed information available to users: what data are available (dataset/owner/contents) and where they are (catalogs)

Ideas

- Tn should work with catalog, not user!
- As soon as data are available, Tn prepare a catalog with attached MetaData to be used by user
- User put this catalog and dataset/owner into `.orcarc`
- Run his/her analysis at Tn

Problems and status

- Many problems to have attached MetaData!! Large latency, very late availability (only very few datasets)
- Create a catalog for $\mathcal{O}(10^4)$ files (fraction of a dataset) not easy: long and fragile
- Data integrity is crucial! Found problems on transfer agent: not easy to check data integrity

IT WORKS

for a fraction of a dataset `mu03_DY2mu`, managed to have all stuff setup and to run on ~ 80 kev (then crash for missing file... data available $\sim 10^6$, $\frac{1}{3}$ of full dataset)

Off Site Access

- what about non LNL user (namely all but me?)
- two possibilities:
  - Give to each user login to LNL. Done for DAQ TDR, painful for user support!
  - Use `Grid` tool to gain access to LNL
- Grid access works!
- Need to know how to setup CMS software, where catalog are, etc, as usual!
- Need some training on GRID (edg) tools (such as submission, output retrieval, job status query, etc). Not really hard

Future

- Tn produces a catalog for each dataset it has
- Publish the catalog to RLS (*grid filesystem*)
- User does a query to the catalog to find it and its physical location
- Can have *standard* tool to perform such task: work on-going in Padova (N. Smirnov `Grape`)
- Use result of your query to decide where to run (in case a dataset is available in different Tn) use grid *data discovery* capability
- Use physical location of local catalog (of chosen site) in user `.orcarc`
- Run the jobs and get back the results

Status

- First dataset of DC04 data is becoming available at LNL for true analysis
- Only fraction of the dataset
- Many problems with data integrity: not possible (yet) to run on all available data, only small fraction
- Not clear when other dataset will become available (bottleneck is preparation of attached MetaData)
- Access to data via grid service is possible
- In principle anyone who can `edg-job-submit`, can look at the data
- Need to learn some grid tools, which are not as good as one would wish...
- Some ideas for short term improvement of grid access, with increasing exploiting of grid functionality

Physics case study:

$H \to WW \to 2\mu 2\nu$

S.L., M. Zanetti, V. Drollinger, I.N.F.N. and Padova University

Foreword

- All *results* presented based on DC04 data, running on DST only at LNL
- Run mainly to demonstrate data accessibility, "physics" is a by product (so far!)
- Very few result, due to lack of data
- Only two background dataset $t\bar{t} \to 2\mu$, $Z/\gamma^{\star} \to 2\mu$ available for running a *standard* executable, no signal available!
- Just a demonstration that data are (veeeeery slooowly) becoming available...

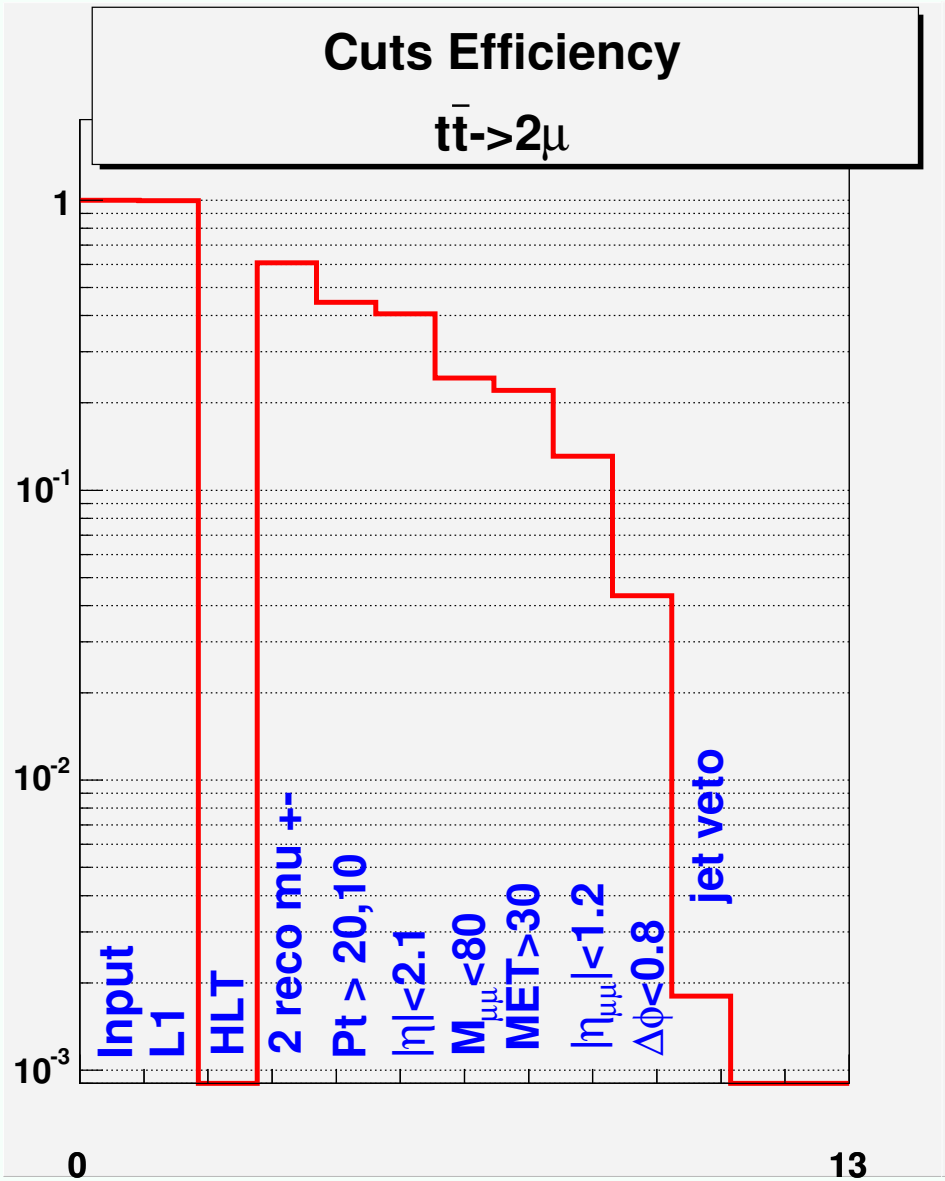
$H \rightarrow WW \rightarrow 2\mu 2\nu$

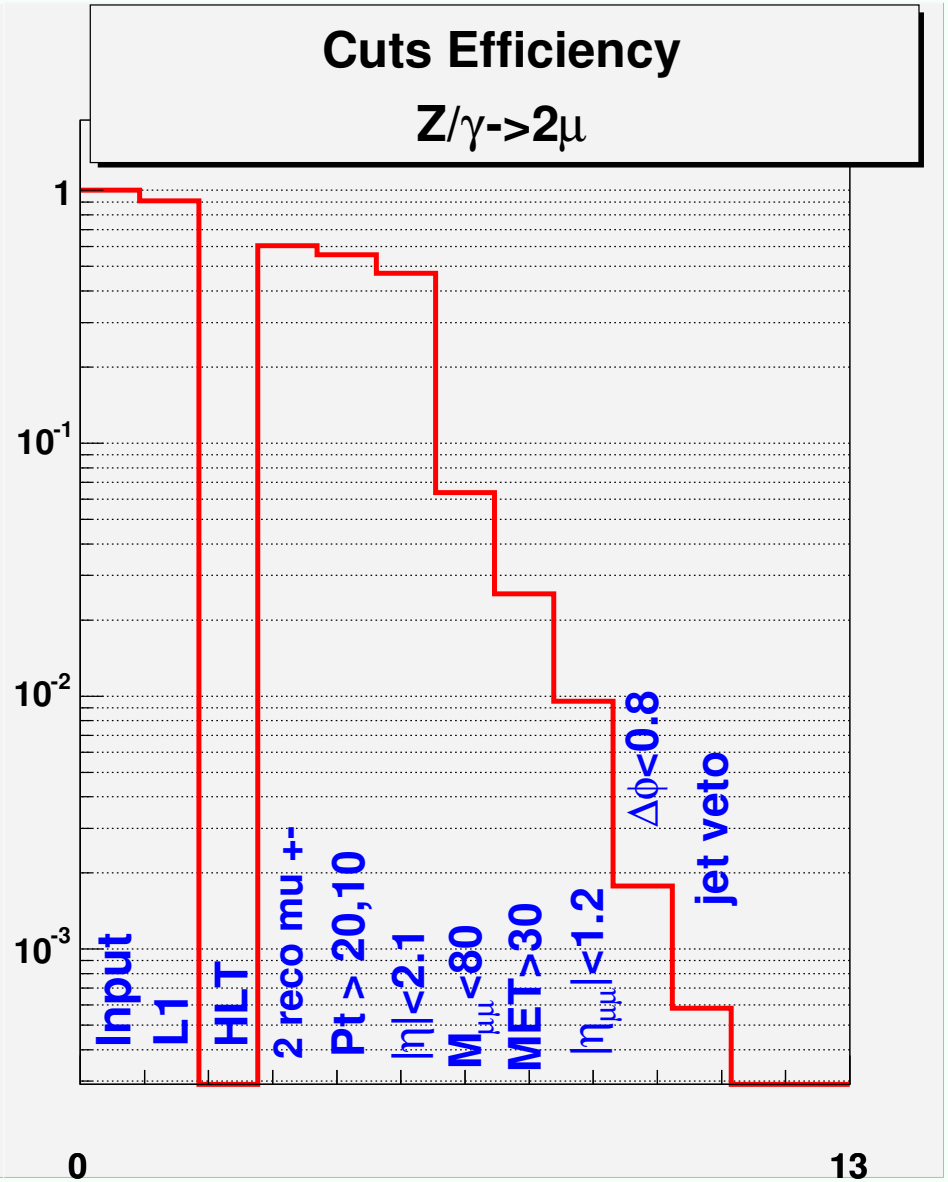
- Signal: 2 isolated μ, MET, no jets, $\mu\mu$ topology due to WW spin correlation ($\Delta\phi$), no peak in invariant mass plot, or similar
- Background: all topologies with 2μ in final state
- $WW,\ WZ,\ Z/\gamma^{\star},\ b\bar{b},\ t\bar{t},$ single top, ... $\rightarrow 2\mu + X$
- Set of cuts to reduce backgound
- Need large background reduction **and** reliable background estimation: counting experiment!
- Today: only two dataset partially available for analysis, and only few days ago! No signal!
- $t\bar{t} \rightarrow 2\mu \sim 430$ **kev**: **fake** analysis at LNL
- $Z/\gamma^{\star} \rightarrow 2\mu\ 82$ **kev**: **true** analysis with **grid access** to LNL from Padova

Cut Reminder

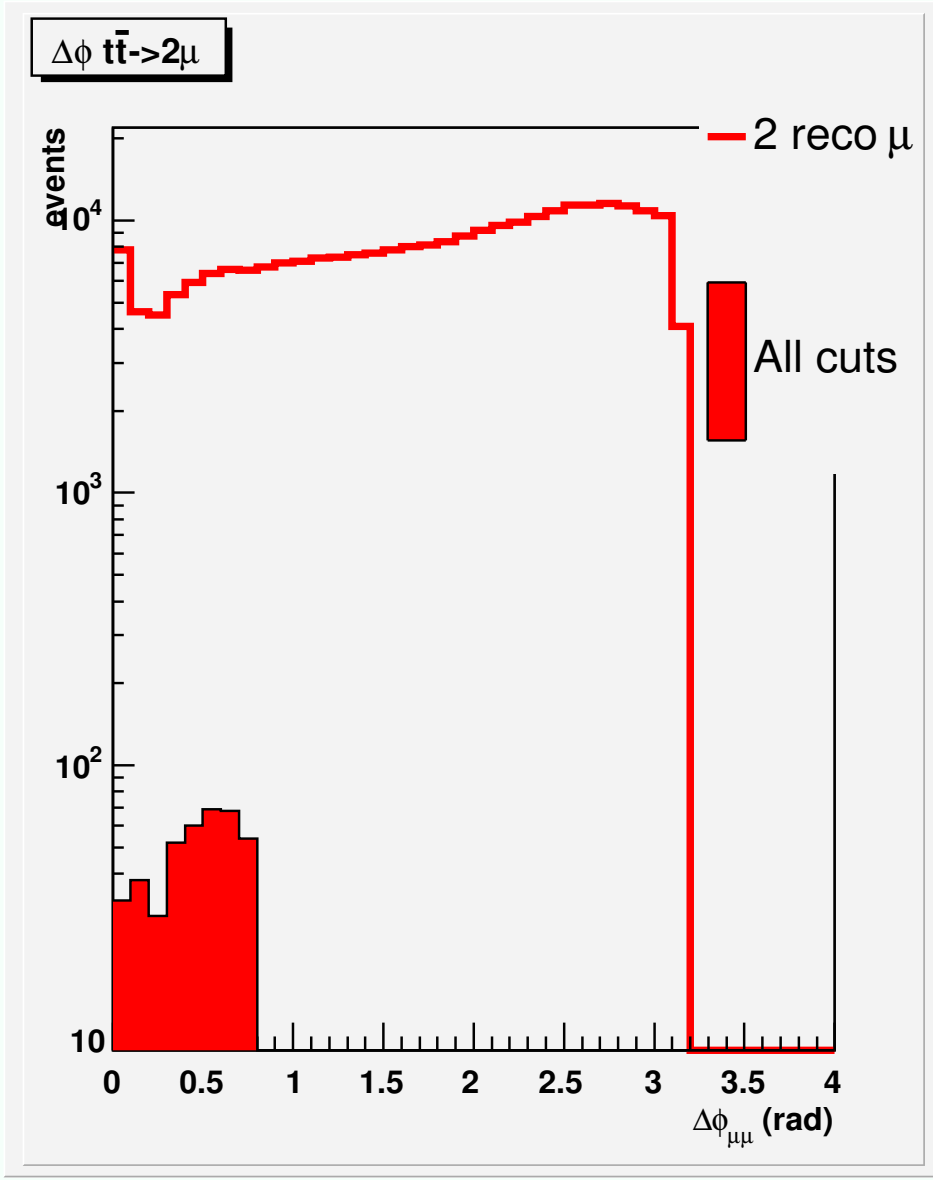
- L1 and HLT selection for single or di-muons,
- exactly 2 opposite sign muons reconstructed,
- $p_t^1 > 20$ GeV/c, $p_t^2 > 10$ GeV/c,
- $|\eta_\mu^{1,2}| < 2.1$ central muons,
- $M_{inv}^{\mu\mu} < 80$ GeV/c^2 Z peak cut,
- $MET > 20$ GeV neutrinos,
- $|\eta_{\mu\mu}| < 1.1$ central events,
- $\Delta\phi < 0.8$, $H \to WW$ spin correlation,
- No jet $p_t > 30$ in $|\eta| < 2.4$ (top events)

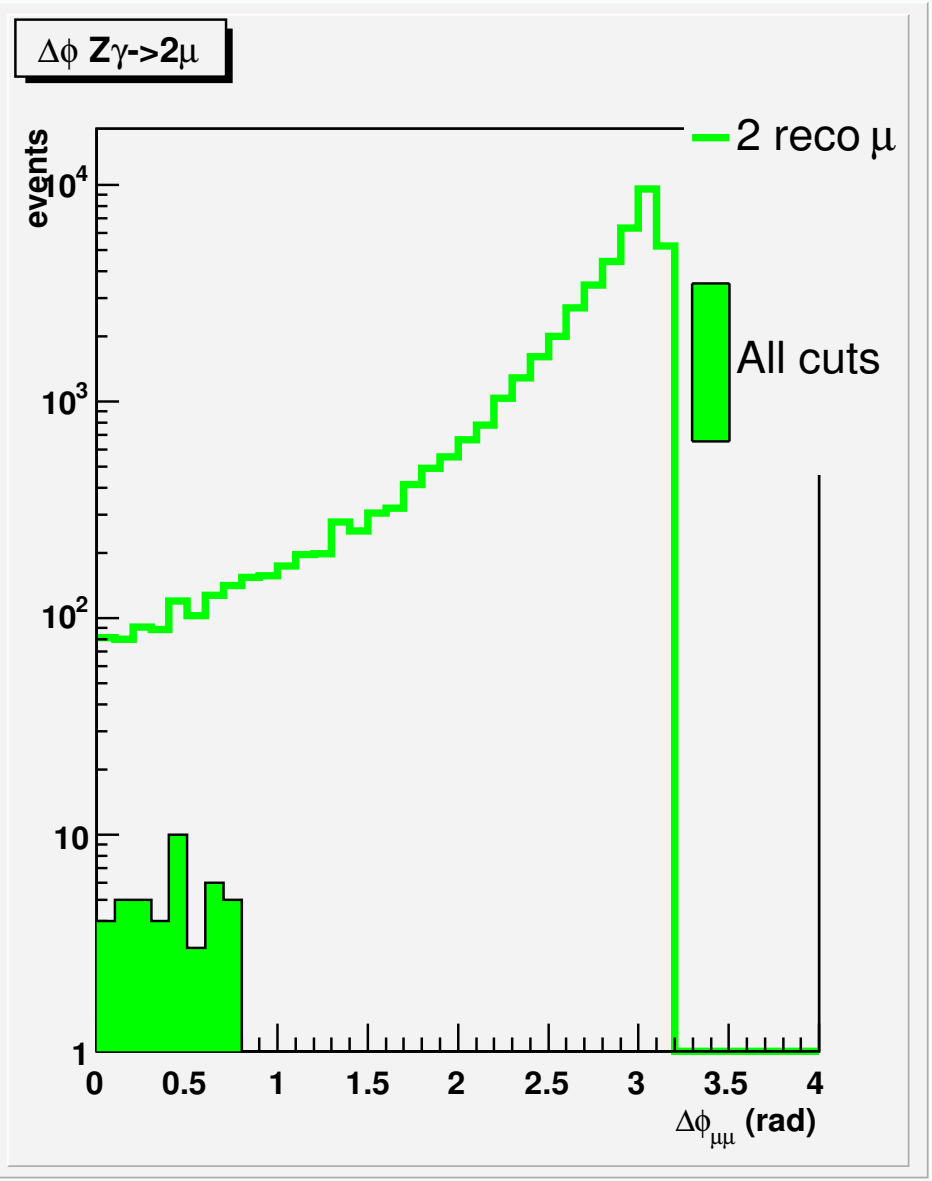
Background Rejection



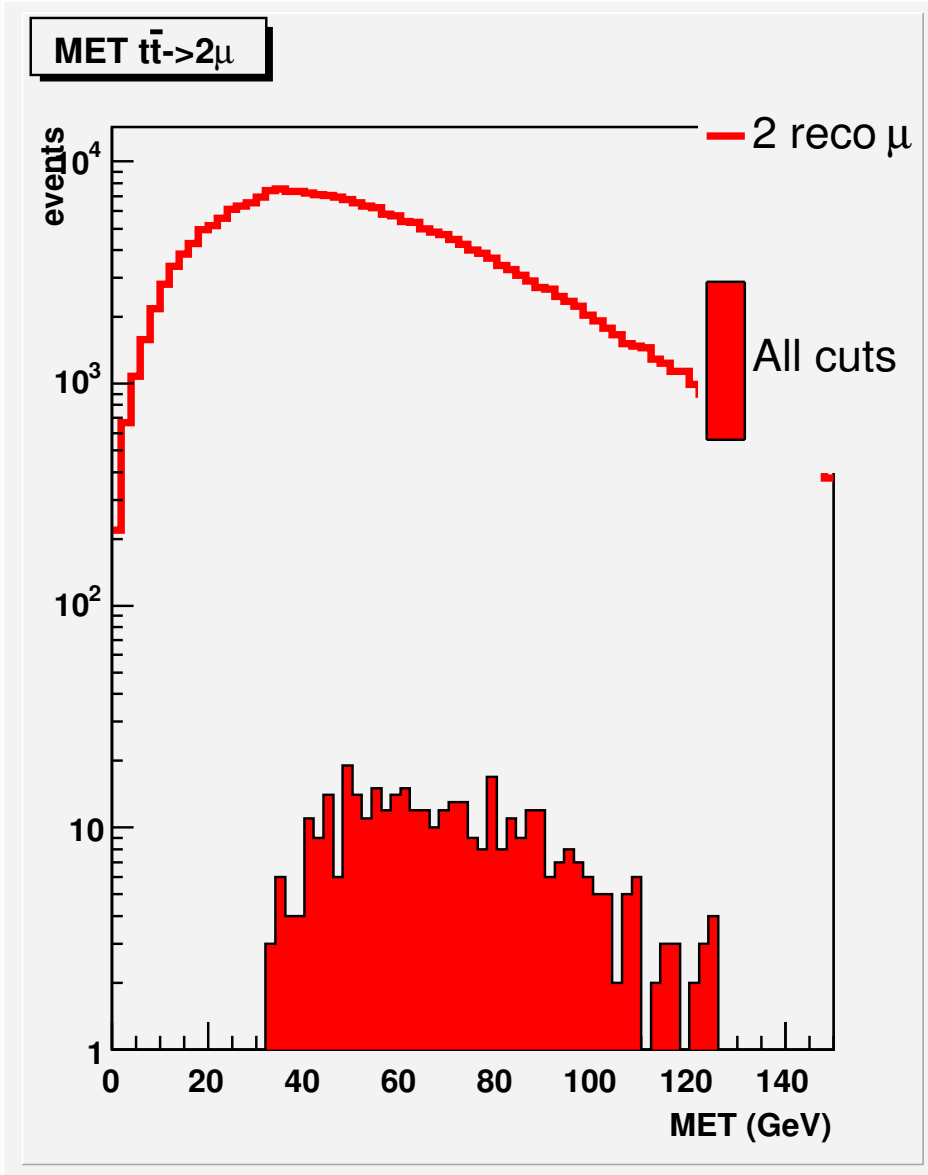


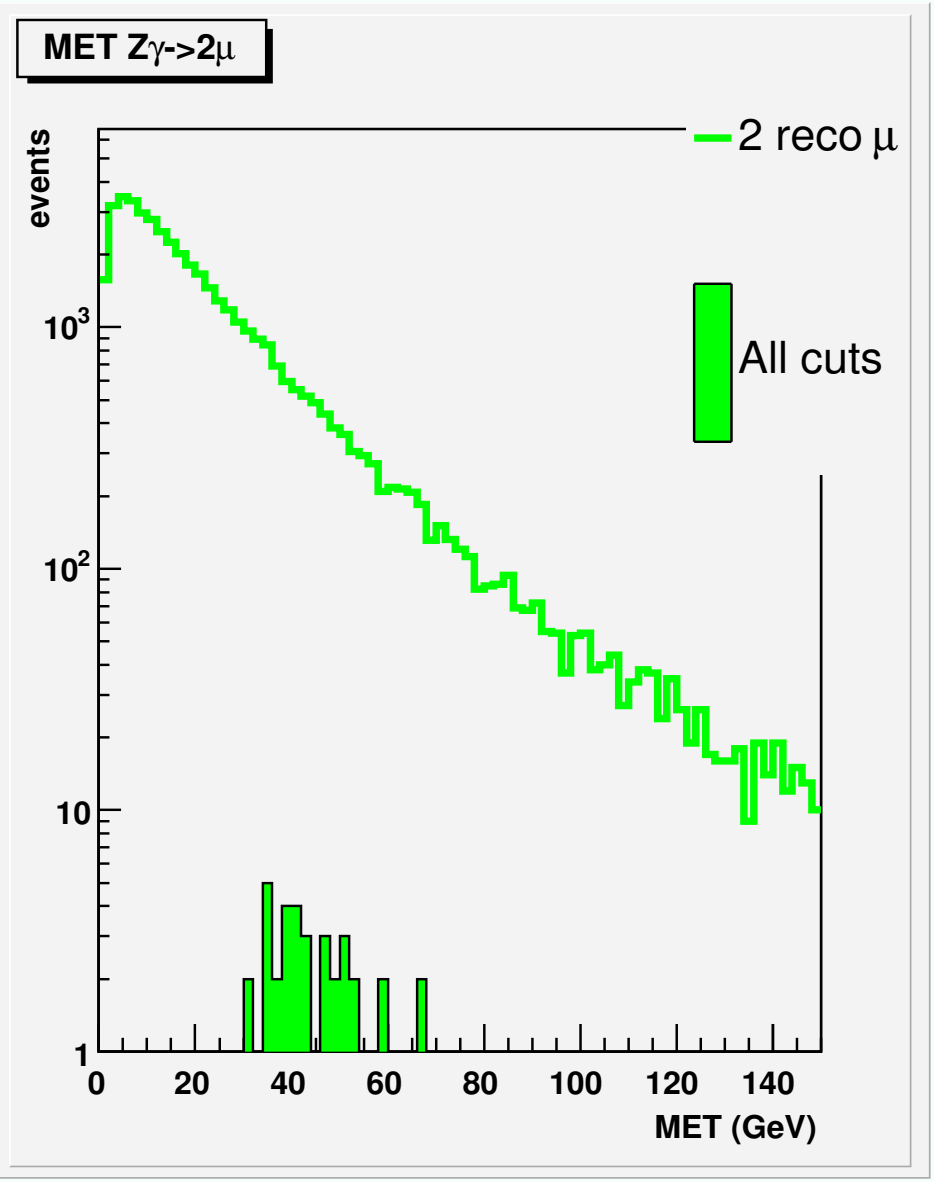
$\Delta\phi$


Δφ tt̄->2μ
2 reco μ
All cuts
events
10^4
10^3
10^2
10
0
0.5
1
1.5
2
2.5
3
3.5
4
Δφ_μμ (rad)

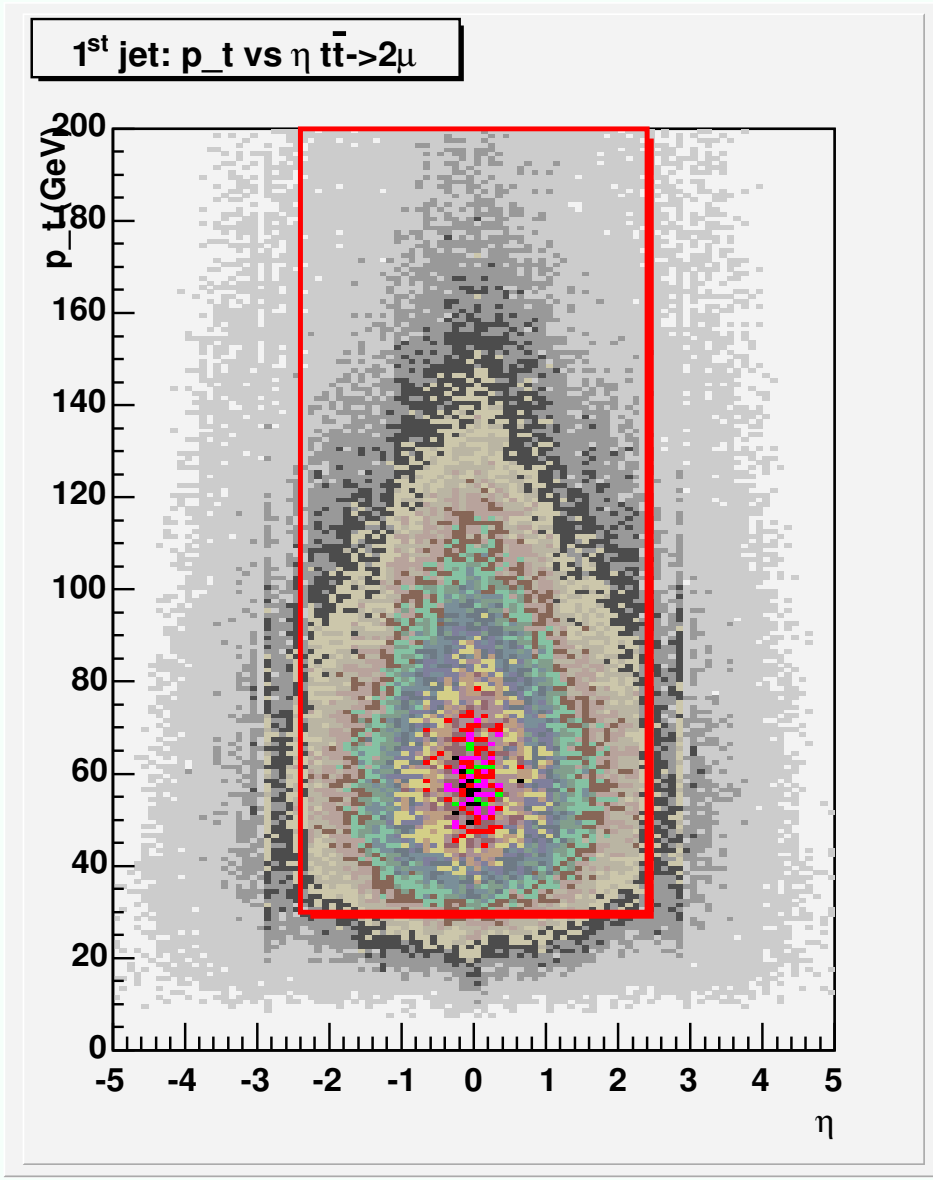

Δφ Zγ->2μ
2 reco μ
All cuts
events
10^4
10^3
10^2
10
1
0
0.5
1
1.5
2
2.5
3
3.5
4
Δφ_μμ (rad)

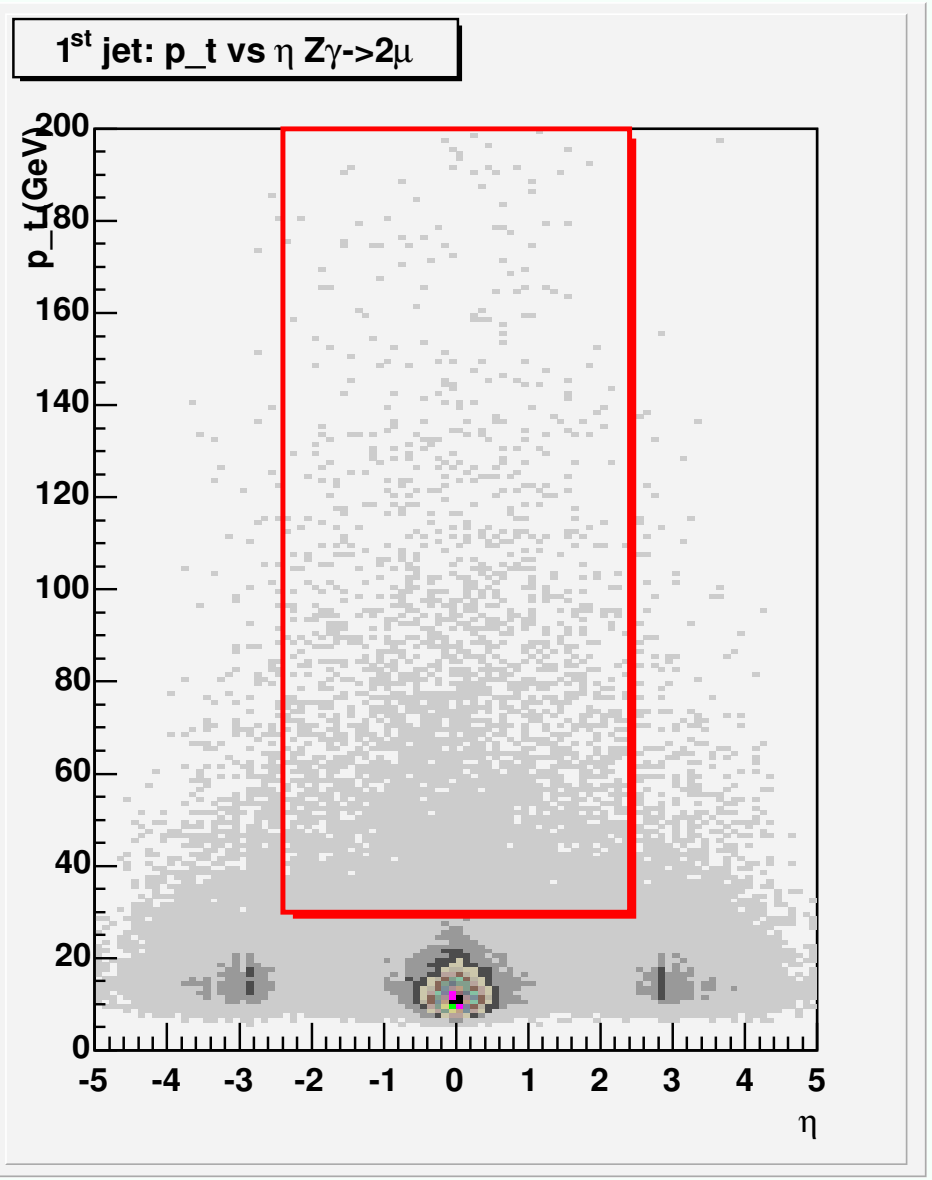
Missing E_t


MET tt̄->2μ
2 reco μ
All cuts
events
10^4
10^3
10^2
10
1
0
20
40
60
80
100
120
140
MET (GeV)

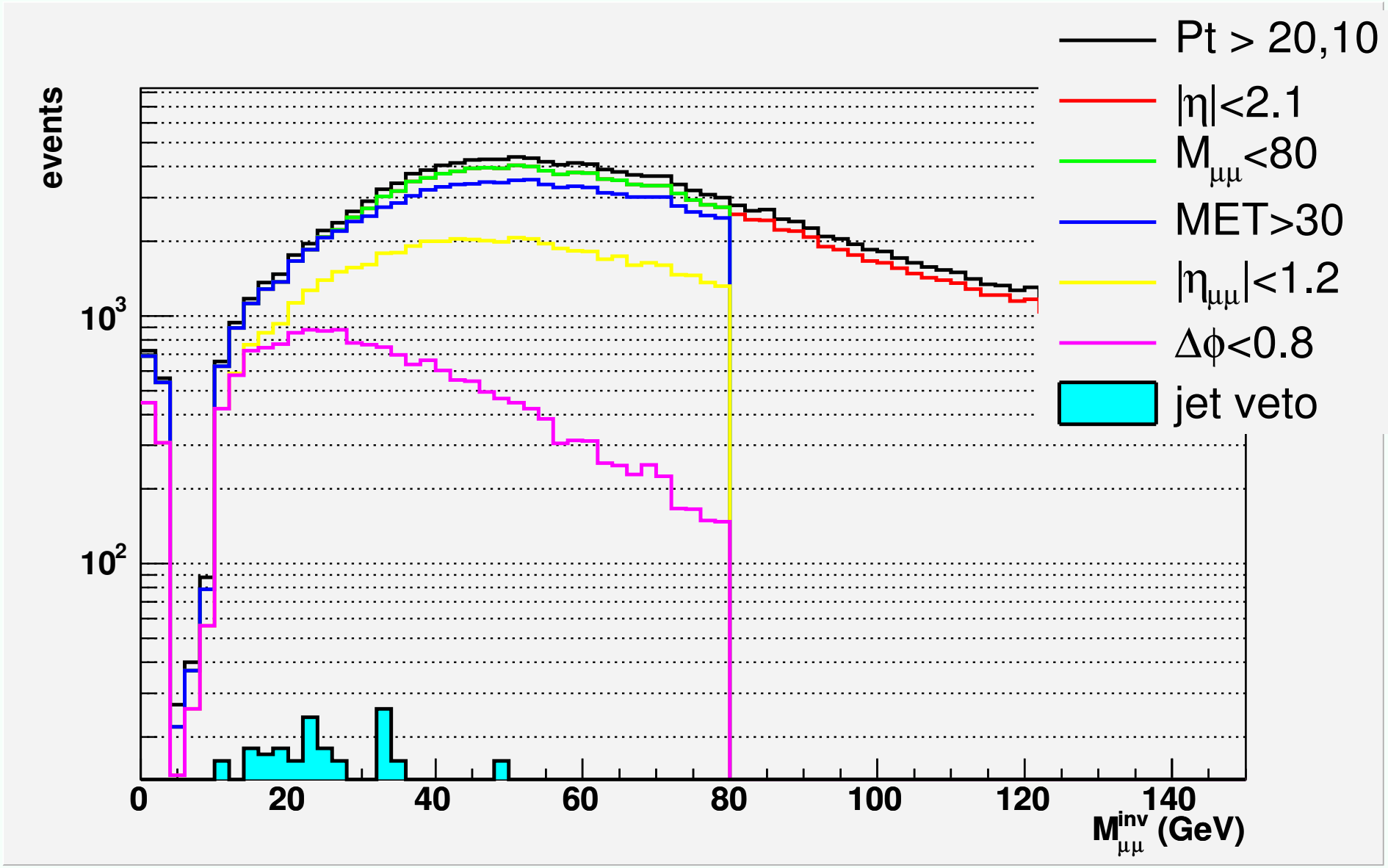

MET Zγ->2μ
2 reco μ
All cuts
events
10^3
10^2
10
1
0
20
40
60
80
100
120
140
MET (GeV)

Jet Veto

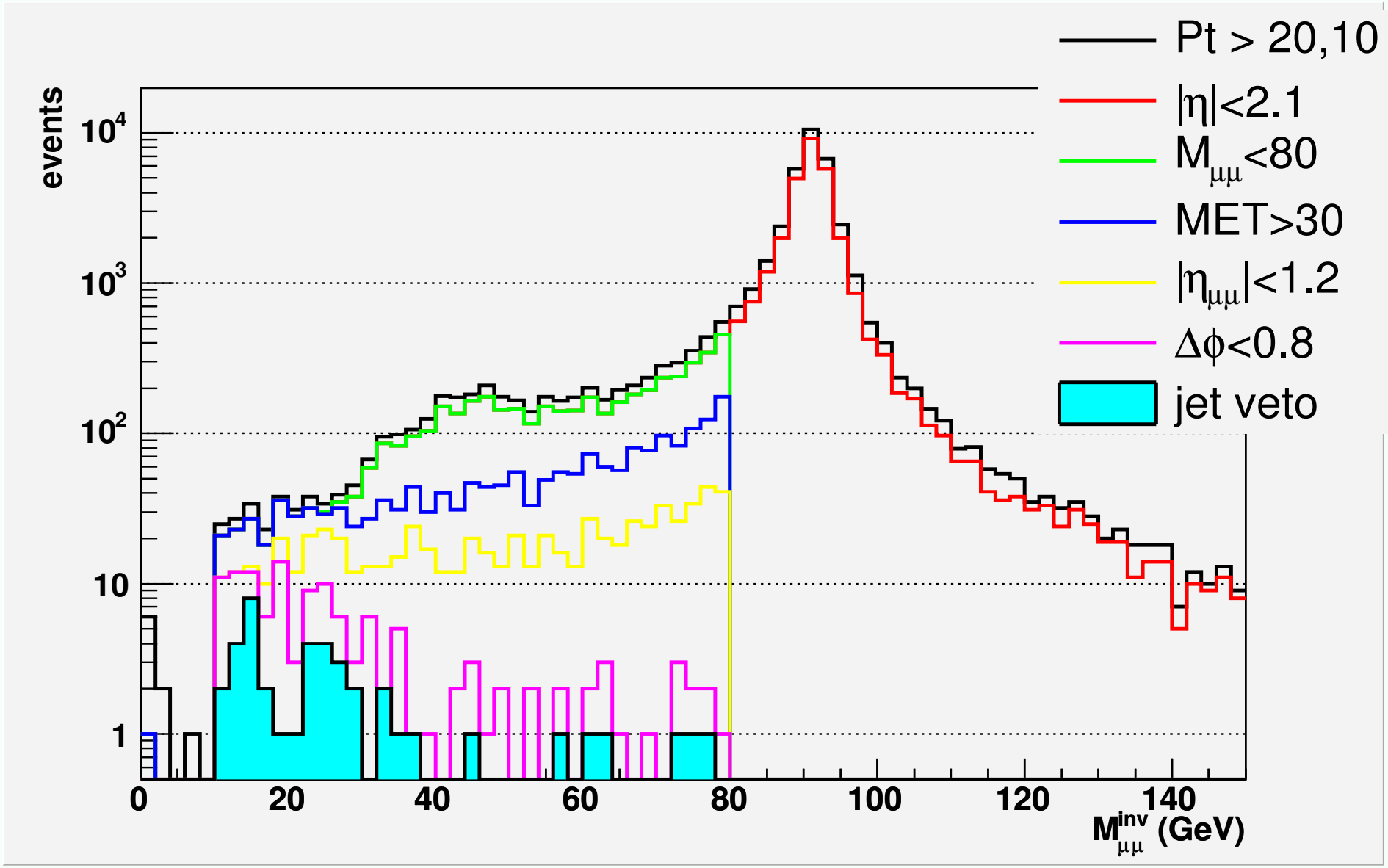

1st jet: p_t vs η tt̄->2μ
p_t (GeV)
200
180
160
140
120
100
80
60
40
20
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
η


1st jet: p_t vs η Zγ->2μ
p_t (GeV)
200
180
160
140
120
100
80
60
40
20
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
η

$M^{inv}_{\mu\mu}$ $t\bar{t} \rightarrow 2\mu$



$M^{inv}_{\mu\mu}$ $Z\gamma^{\star} \rightarrow 2\mu$



Residual Rate

- Very preliminary

Dataset	$t\bar{t} \to 2\mu$	$Z/\gamma^{\star} \to 2\mu$
#ev "analyzed" (kev)	430	82
full dataset (kev)	600	3000
fraction	70%	3%
Residual Rate (10^{-6} Hz)	32	17
#ev left	770	33
σ_{stat}	4%	17%
σ_{stat} full DS	3%	3%

$\star$ $H \to WW \to 2\mu 2\nu$ event rate (after cuts) from pre DC04 analysis (M.Zanetti thesis)

M_{Higgs} GeV/c^2	120	140	160	200
Residual Rate (10^{-6} Hz)	1.0	5.6	14	4.6

Conclusions

- Not a real analysis, yet: too few time to look at too few data!
- Test accessibility of DST data from DC04 shows first success
- "Results" presented here obtained running ORCA on DST (just reading, very fast!) from DC04 production, submitting jobs over the grid
- Need more effort to allow full PRS community to be able to do the same: timescale can be short
- Need much more data as soon as possible!!!
- A lot already fully processed, need to be made accessible

- Have a lot to learn how to access and use efficiently DST (feedback)
- Same for all different reconstruction algorithms (muons, jets, MET, vertexes, tracks, etc): problems, performances, tuning...
- Very first look at two backgrounds for $H \to WW \to 2\mu 2\nu$: need lot of work but promising!!
- Statistics (will be) available from DC04 enough for small statistical error (real work will be to estimate systematics)
- Is it finally time to start analysis??