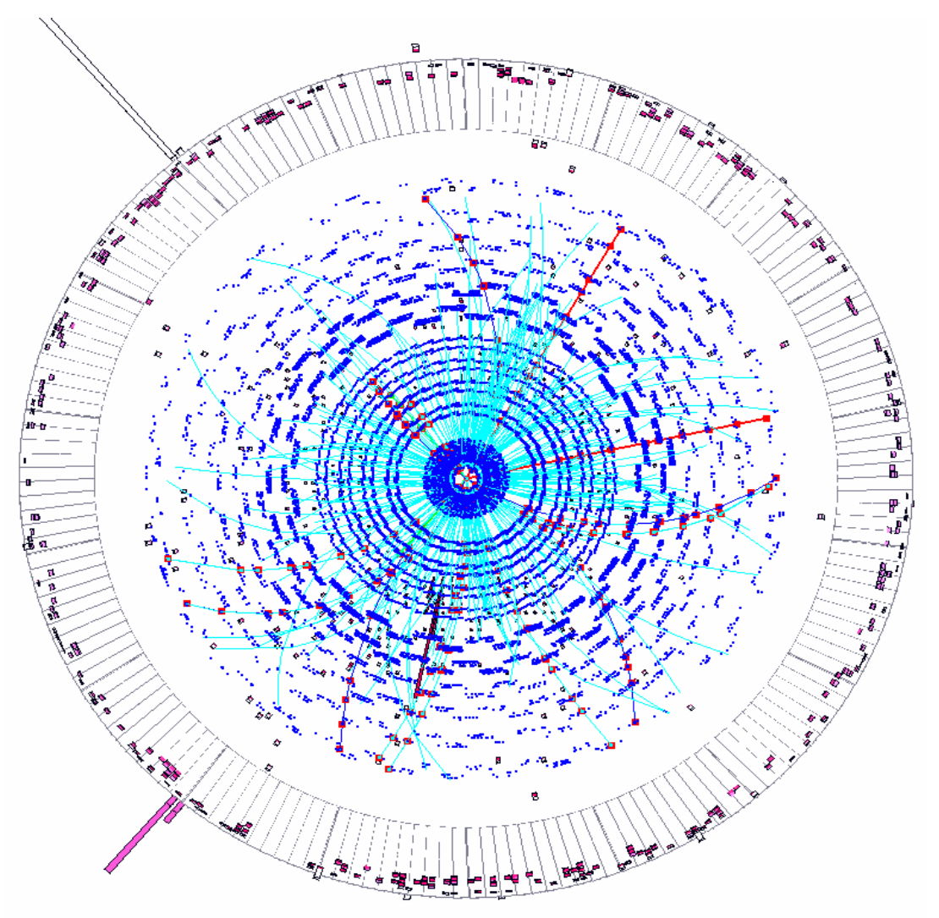
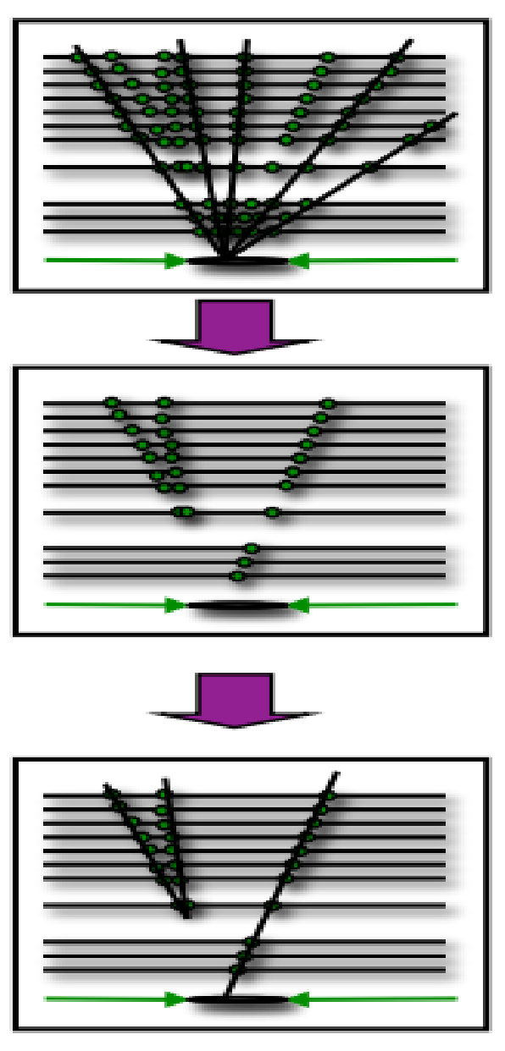
# Prediction of event processing time at the CMS experiment of the Large Hadron Collider

Samir Cury[1], Oliver Gutsche[2], Dorian Kcira[1]

[1]California Institute of Technology, [2]Fermi National Accelerator Laboratory

One of the biggest challenges of the CMS experiment is the precise reconstruction of charged particle tracks in the detectors as well as the combination of information from the different sub-detectors. This is done through carefully designed, elaborate algorithms, which translate into CPU-intensive tasks. At the Large Hadron Collider, understanding the details of the algorithm performance and its relation to event complexity is one of the key factors to facilitate the processing of workflows in a more uniform and efficient way. The analysis presented here aims at estimating the event reconstruction time for the future LHC data based on observations on data already acquired.

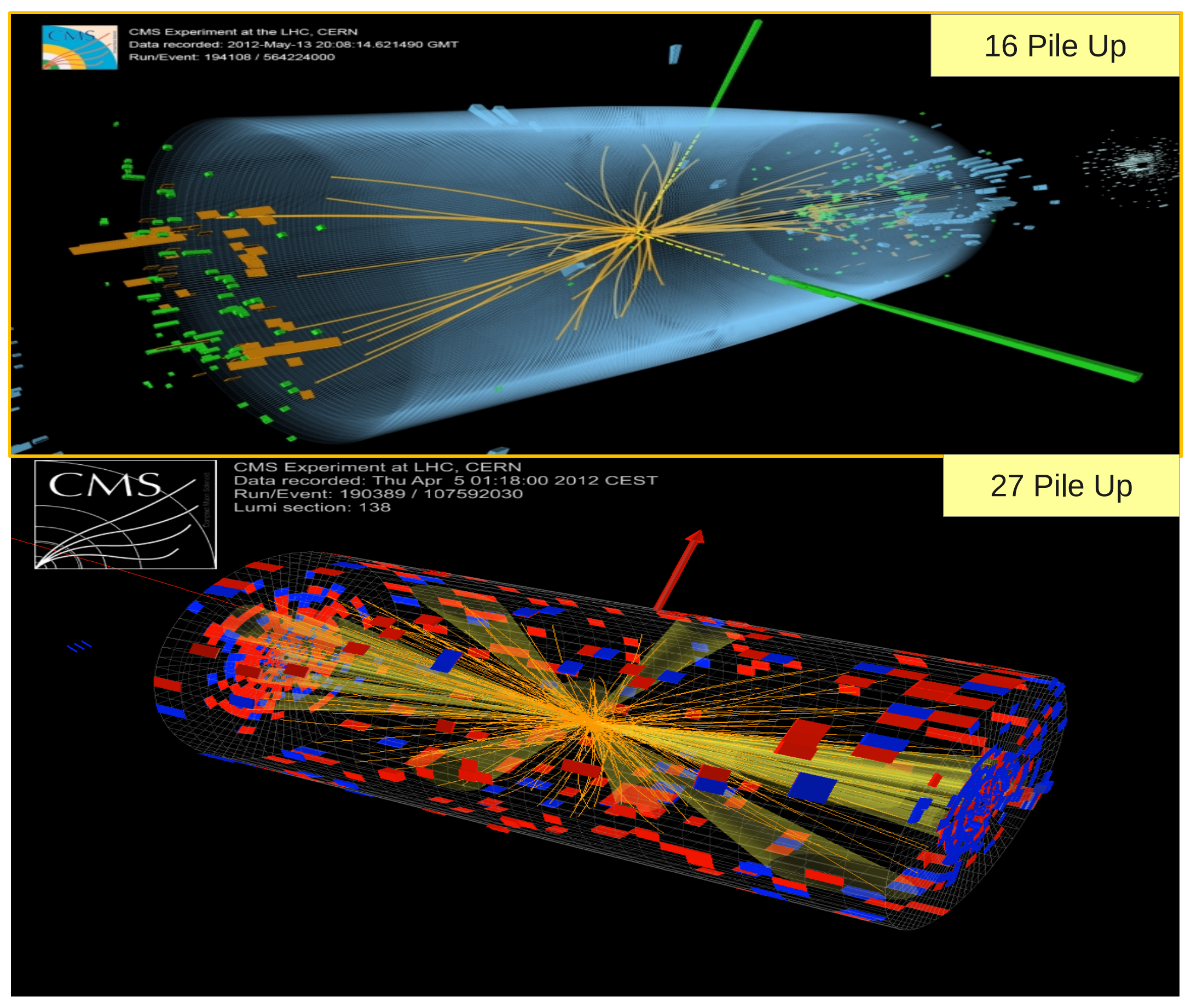




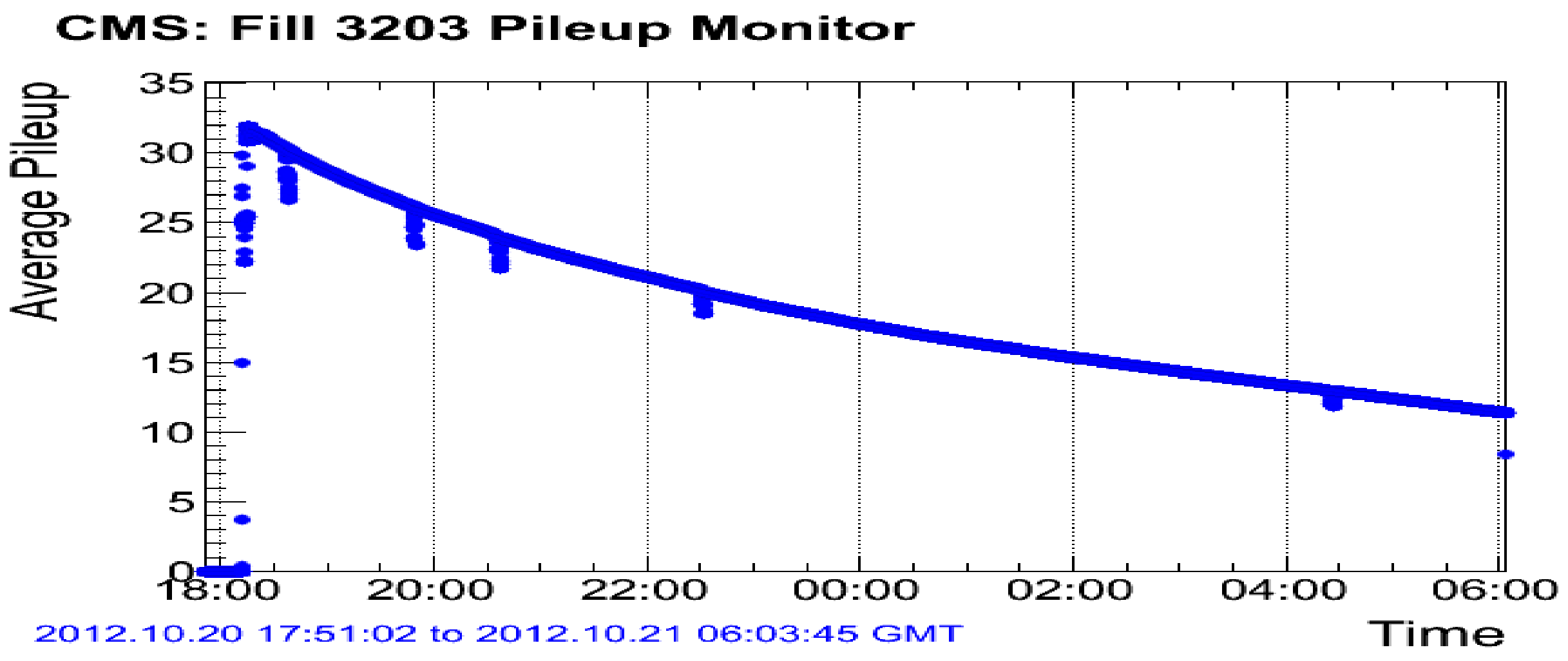
CMS Tracker – event visualization (ORCA)

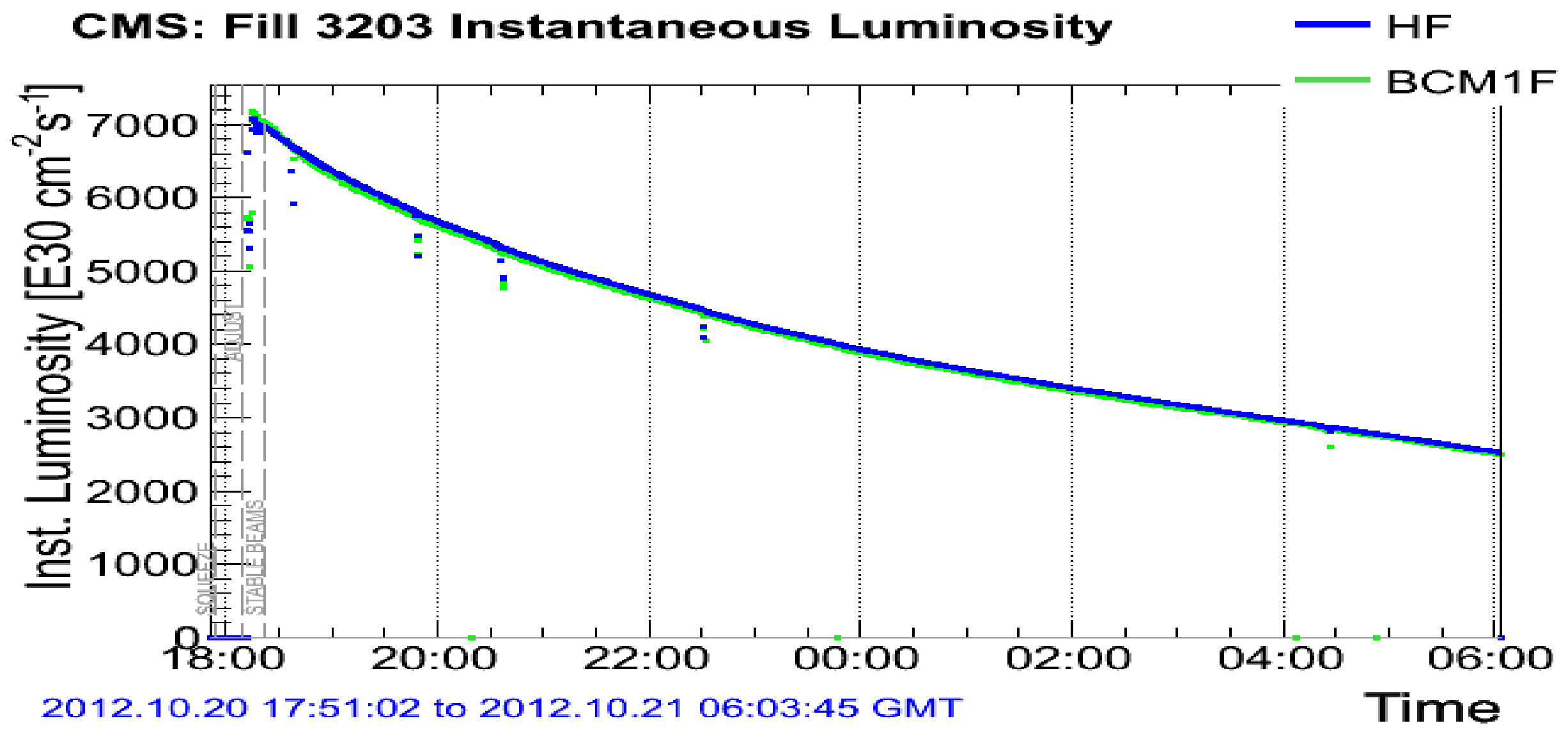
The complexity of track reconstruction is due to the large number of charged particle tracks from the collisions as well as the overlap among them. Iterations become thus necessary in order not only to fit hits in the tracking detectors but also to distinguish the possible combinations resulting from combinatorics. The number of hits used for reconstructing tracks depends strongly on instantaneous luminosity and the number of collisions that happen simultaneously per beam bunch crossing (pile-up interactions). Pile-up itself is a function of the accelerator running conditions and instantaneous luminosity. The focus of this study is, therefore, on instantaneous luminosity.

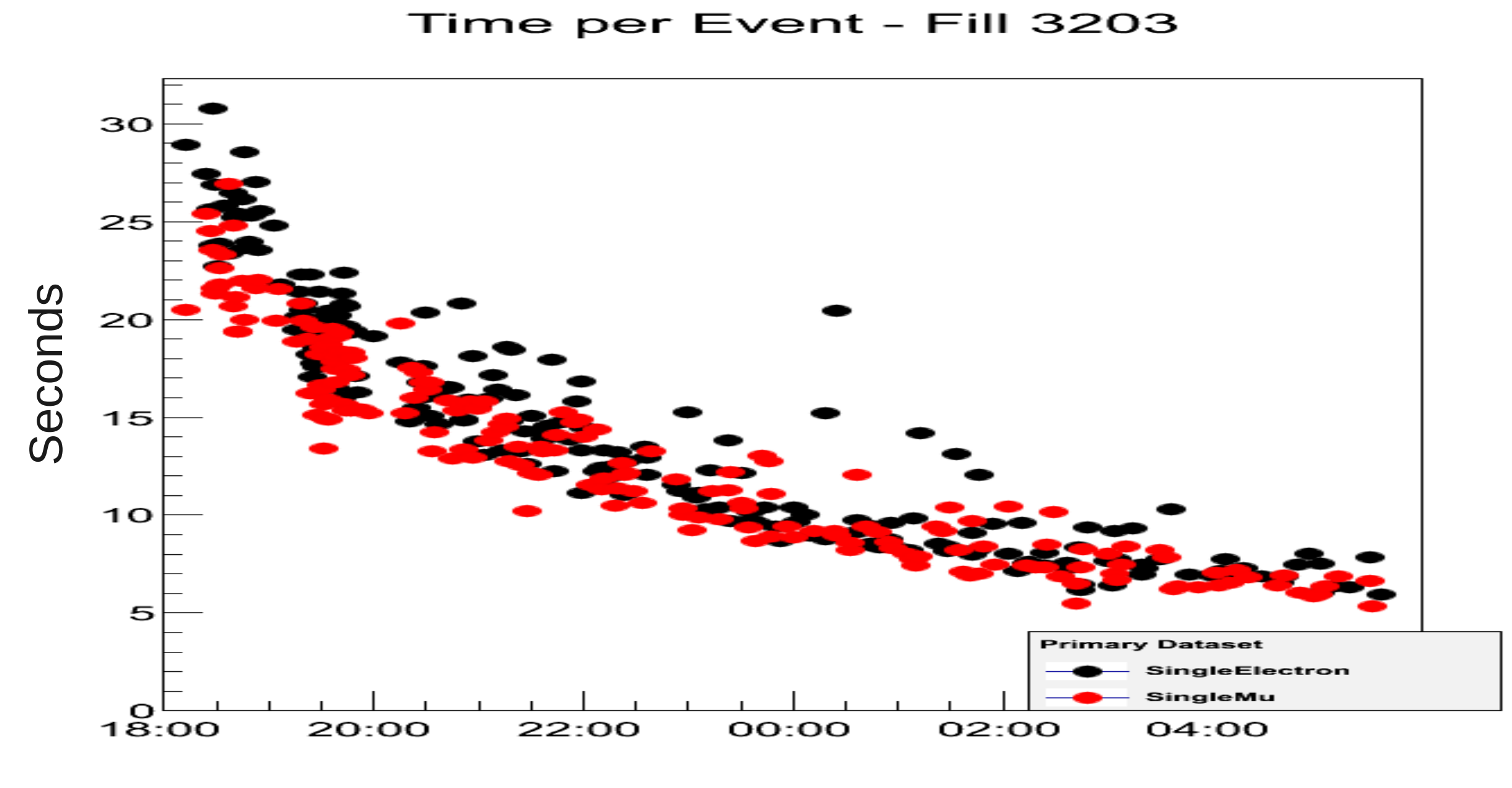
The event displays below show tracks coming out of collisions with lower number of tracks (top, 16 pile-up events per bunch-crossing) or many tracks (bottom) as in events with high instantaneous luminosity and large event pile-up (27 PU/bunch-crossing).


CMS Experiment at the LHC, CERN
Data recorded: 2012-May-13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000
16 Pile Up
CMS
CMS Experiment at LHC, CERN
Data recorded: Thu Apr 5 01:18:00 2012 CEST
Run/Event: 190389 / 107592030
Lumi section: 138
27 Pile Up

The CMS Fill Report provides plots of instantaneous luminosity and pile up over time. Here we can compare those with the observed reconstruction time per event of these data observed at the CMS Tier-0. The comparison shows that time per event has a direct relation with pileup.

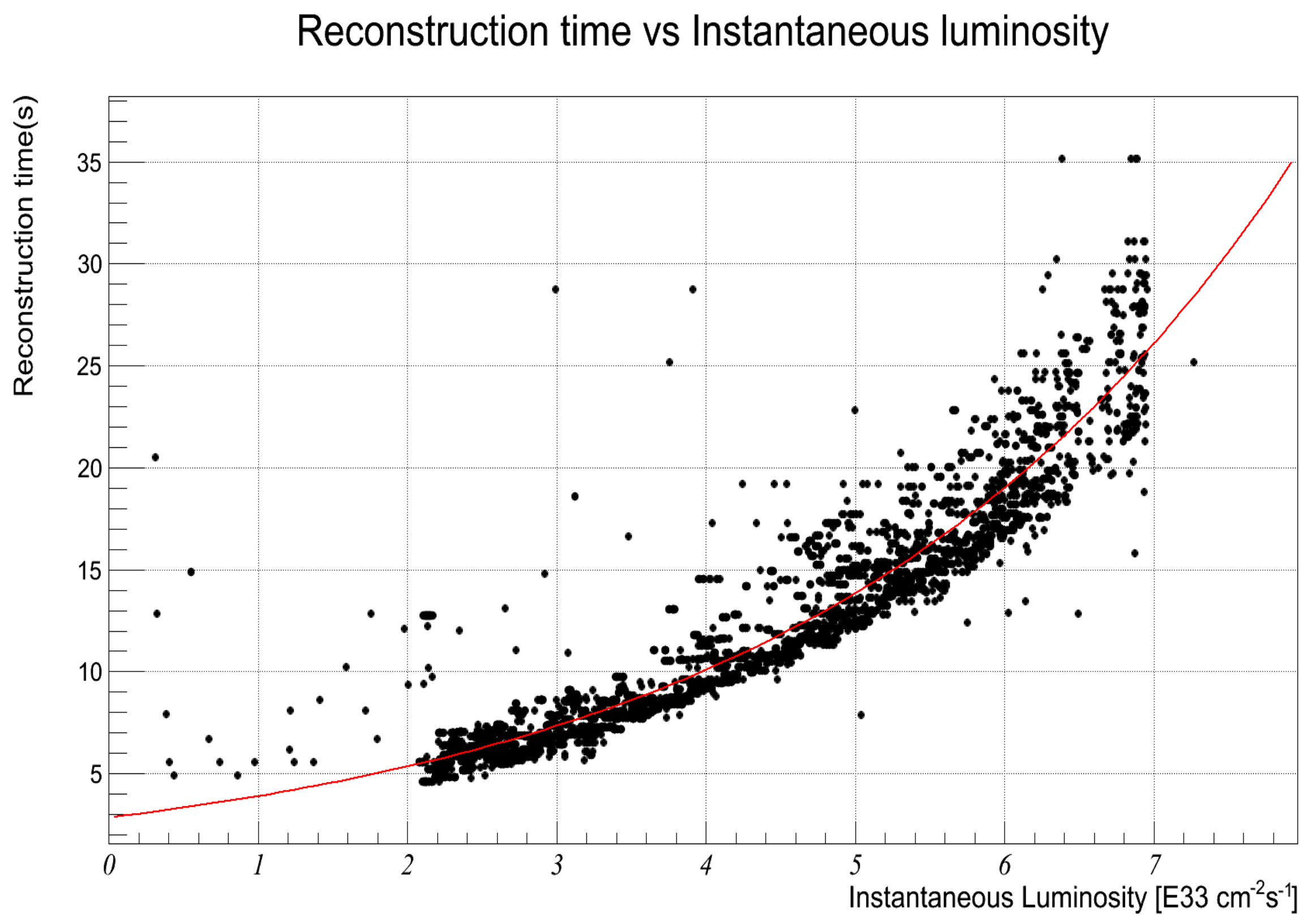






The following is a measurement of the CMS software (CMSSW) performance for a given software release and type of events (primary dataset). The performance varies significantly according to the type of physics. Different physics signatures naturally produce more, or less tracks.

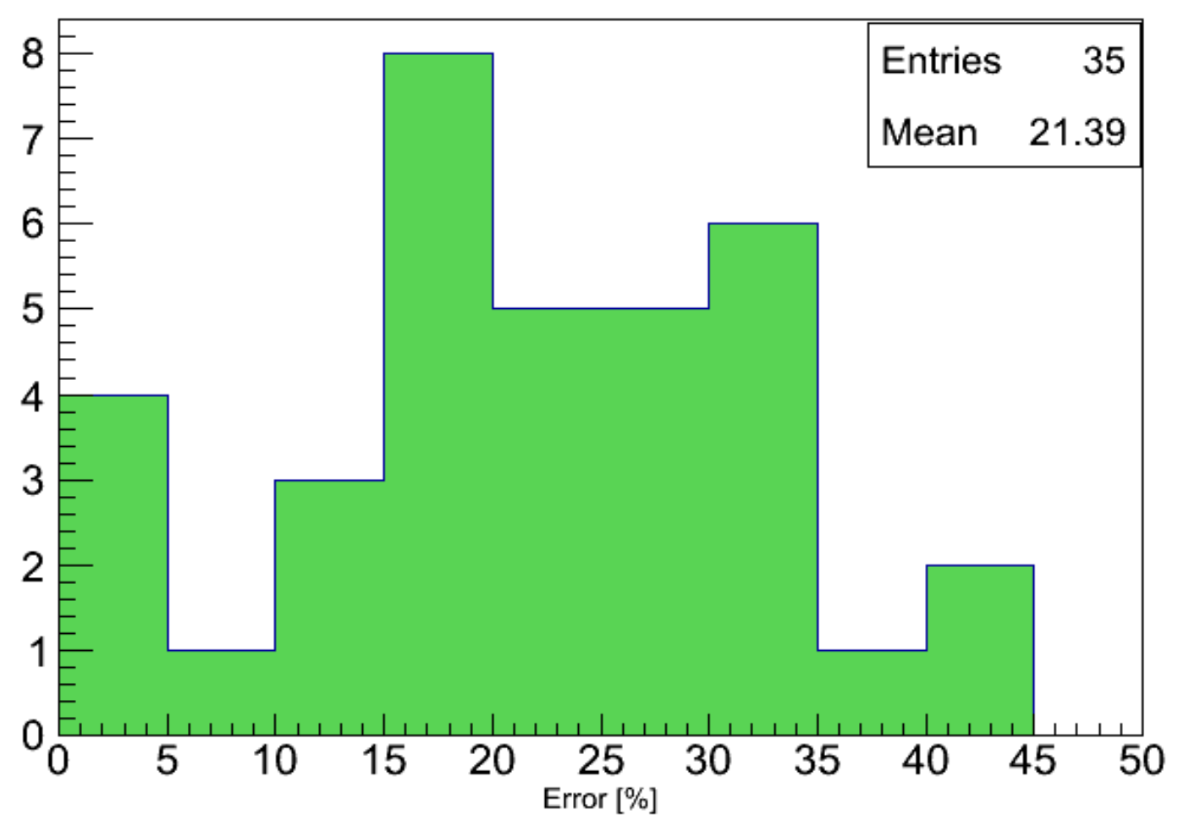
These measurements on existing processed data are used to estimate processing time of future LHC data. One important factor to consider in the estimate are systematic shifts in these measurements caused by the heterogeneity of the processing farms. Different CPU models will result in different processing time for the same collision type. Our measurements have been done over different CPU models as we believe that the resulting average is a representative value that will be the most useful as an estimate for the CMS central operations.



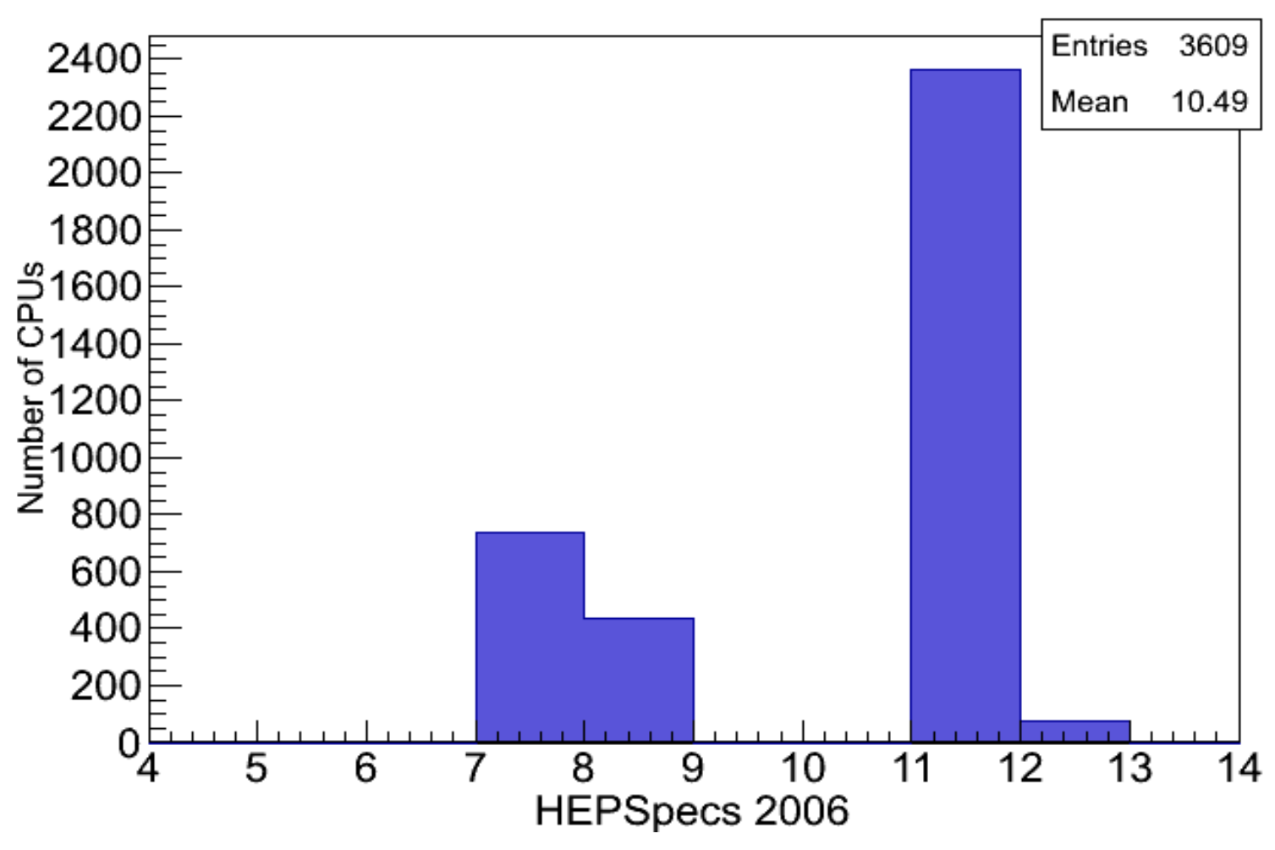
Generic Single Muon performance curve for CMS Reconstruction

Measurements were done on 35 PromptReco workflows to observe how close to the real value our estimation gets. The error introduced by the CPU speed fluctuation in the Tier-0 farm can be up to 37.75%. This comes from the difference of HEPSpecs 2006 (Benchmark unit) between the fastest and slowest CPU models. The green histogram shows the distribution of error values for all workflows. The blue is a histogram of the number of cores in the farm per HS06 values, showing how they contribute to the error. In the table below, some specific measurements, from smaller to bigger error.

| Run # | Estimated job length | Observed job length | Error [%] |
|---|---|---|---|
| 202075 | 6:15:00 | 6:14:35 | 0.1 |
| 202088 | 3:12:44 | 2:59:39 | 7.2 |
| 202014 | 6:50:15 | 5:39:11 | 20.9 |
| 201707 | 5:20:26 | 7:53:23 | 32.3 |

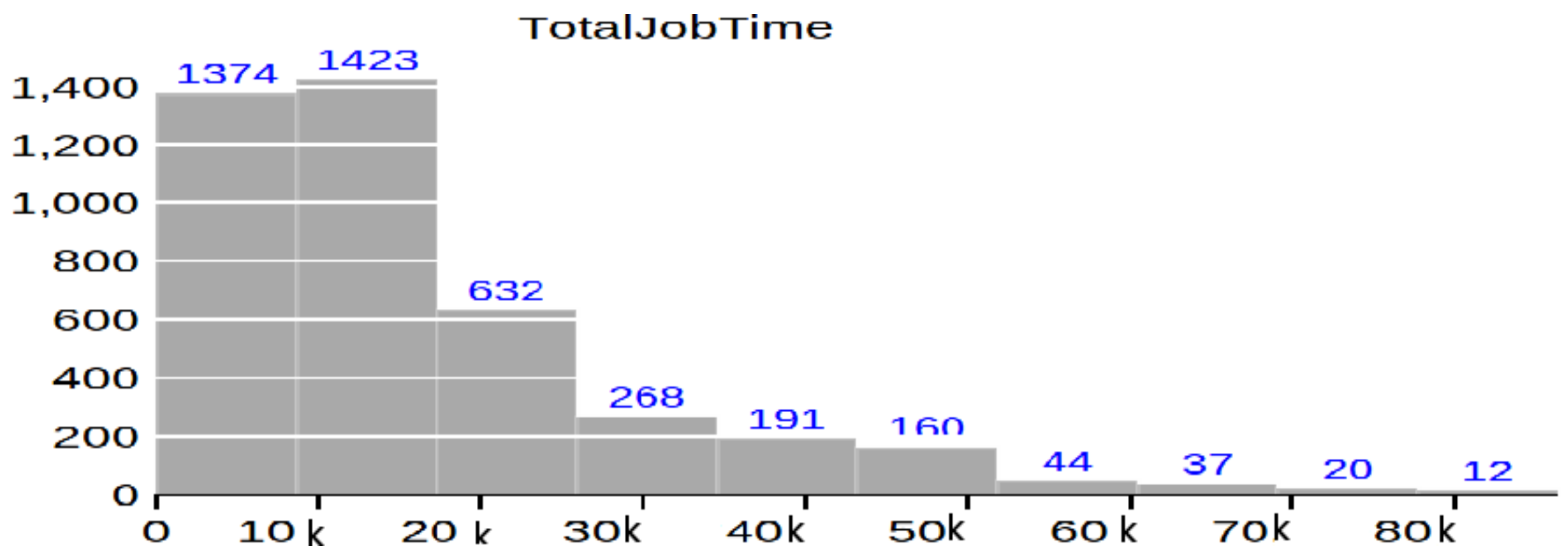


Error distribution – expected x observed average job length

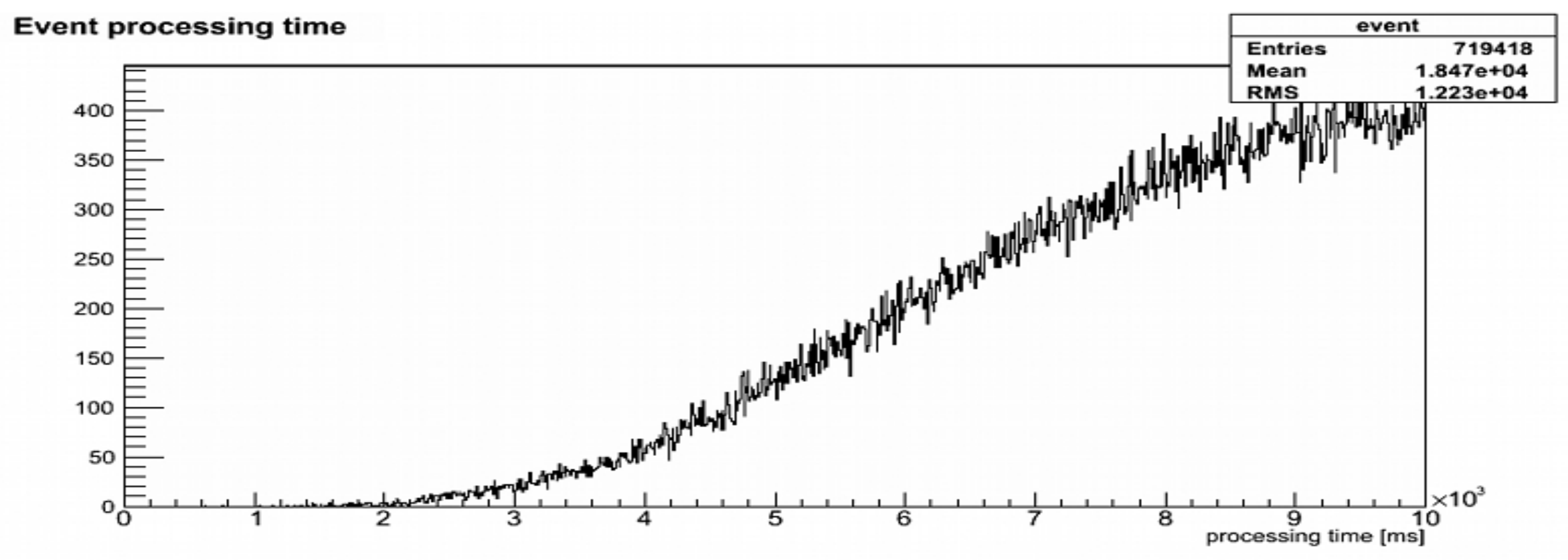


Tier-0 Farm CPU speed distribution

Due to the wide range of luminosity values, a wide distribution of job lengths in a multi-run reconstruction workflow is observed. As a consequence so-called tail effects are observed that consist in a little number of jobs take from 3 to 7 days to finish after 90% of the workflow is already finished. This delay is caused by jobs processing high-luminosity data, where a single job can take up to 48h to finish and more in the case of retries needed due to job failures.



Wide job length distribution – consequence of wide event processing time range



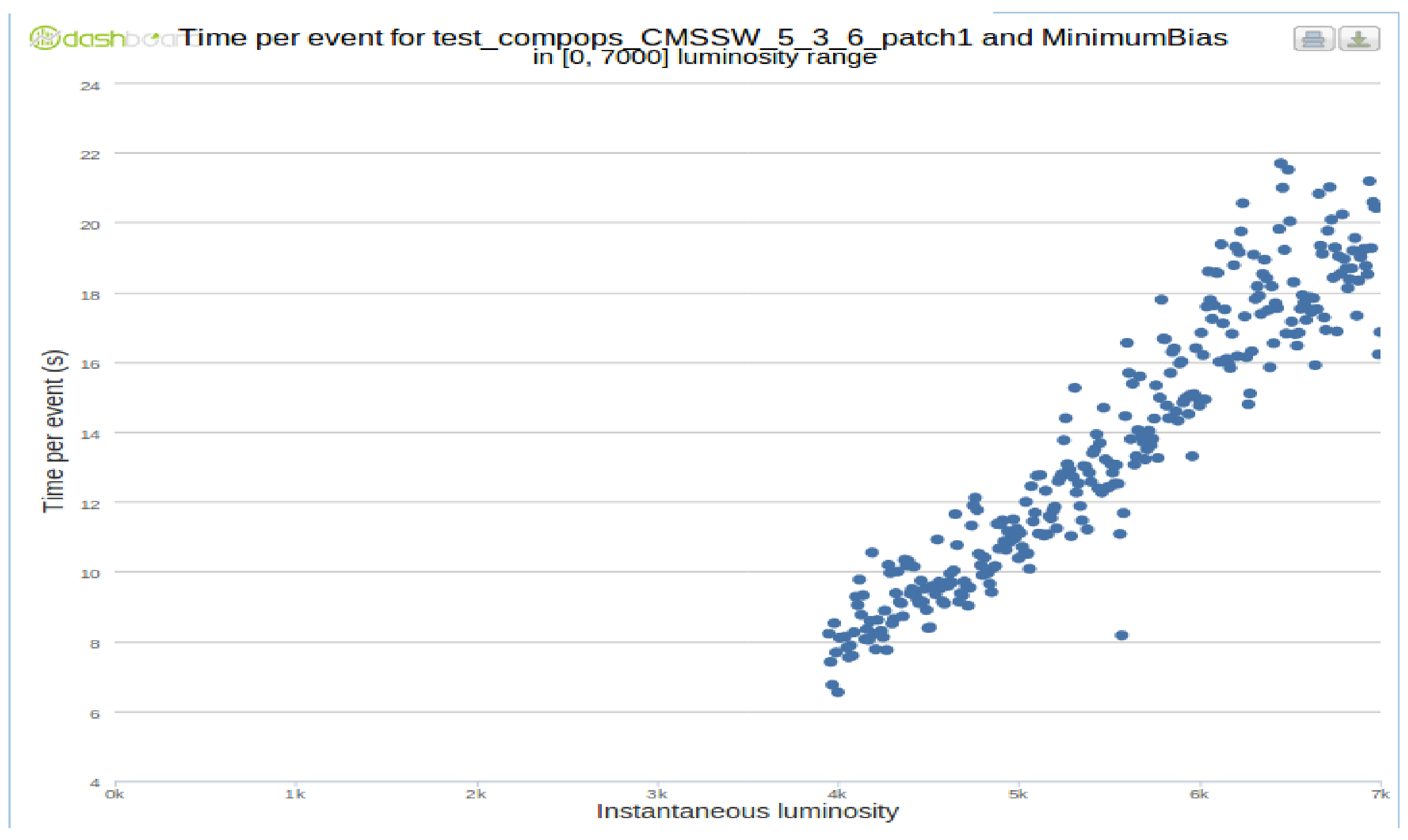
Wide processing time range - consequence of wide instantaneous luminosity range

This study motivated a solution to diminish the long tail effects in CMS data processing. As the relation between instantaneous luminosity and reconstruction time is well known, we are able to predict the time-per-event by using the luminosity value from the data. Different CMS web services exist that provide access to this kind of information. A job-splitting algorithm was developed for the Workload Management Agent that uses this information to estimate a processing time per event. In addition the number of events per processing job is chosen dynamically such that the processing times become more uniform. The ideal processing time per job is approximately 8 hours.

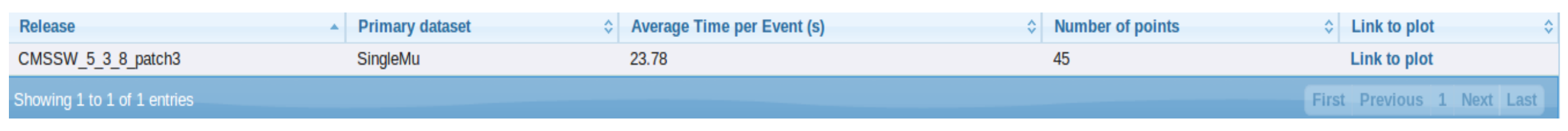
```
2013-10-05 09:47:05,686:DEBUG:LuminosityBased:This file has average instantaneous luminosity 5823.701863 average time per event 18.329333 and is getting 1178 events per job
2013-10-05 09:47:05,702:DEBUG:LuminosityBased:This file has average instantaneous luminosity 6957.163043 average time per event 26.156875 and is getting 825 events per job
```

Example of real log messages to demonstrate how the algorithm works

The job-splitting algorithm uses performance information collected at the end of each workflow by the WMAgents. This information is reported to a specific service maintained by the CMS Dashboard. The information is not only used from the data service for automated systems but also for web-based interfaces to be used by CMS members to visualize performance curves or average processing times per release and dataset.


dashboard
Time per event for test_compops_CMSSW_5_3_6_patch1 and MinimumBias
in [0, 7000] luminosity range
Time per event (s)
24
22
20
18
16
14
12
10
8
6
4
0k
1k
2k
3k
4k
5k
6k
7k
Instantaneous luminosity

DashBoard – Performance curve from reported data

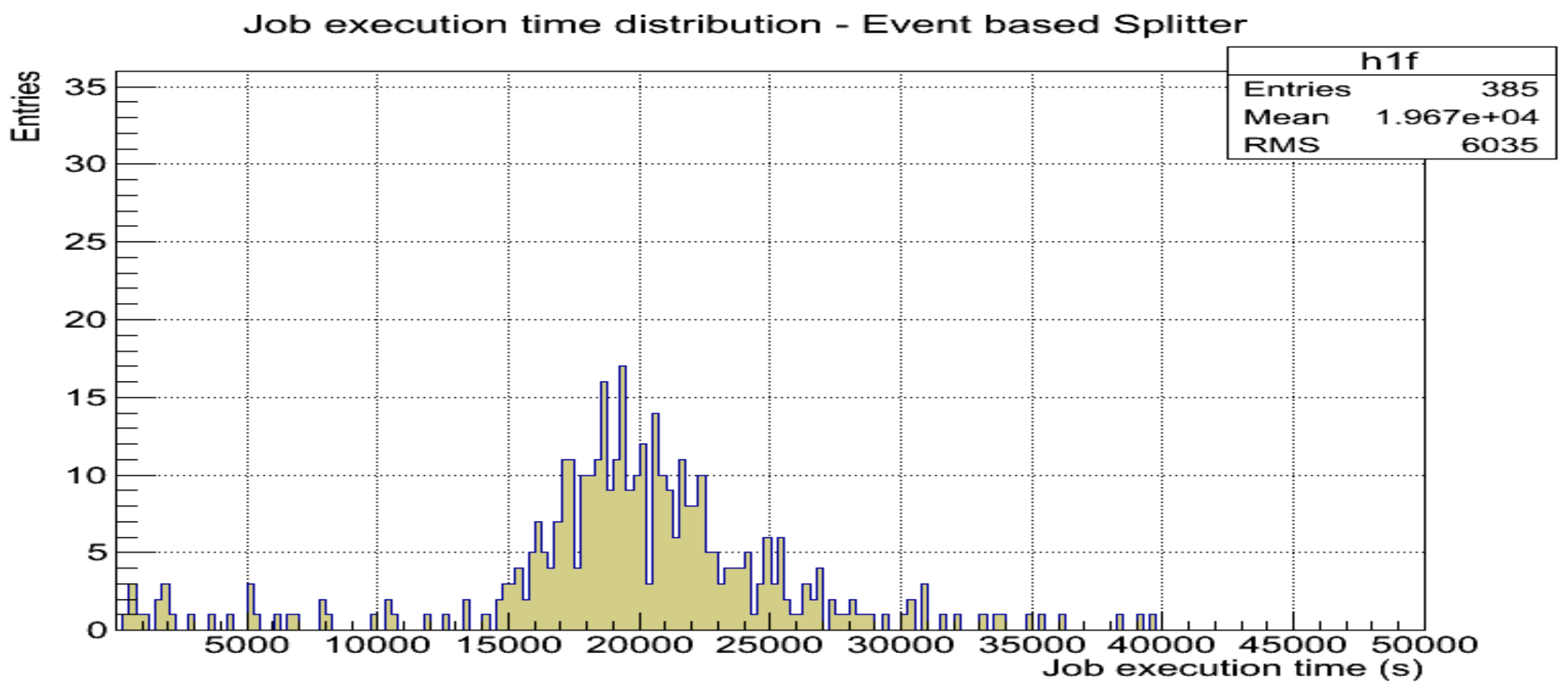


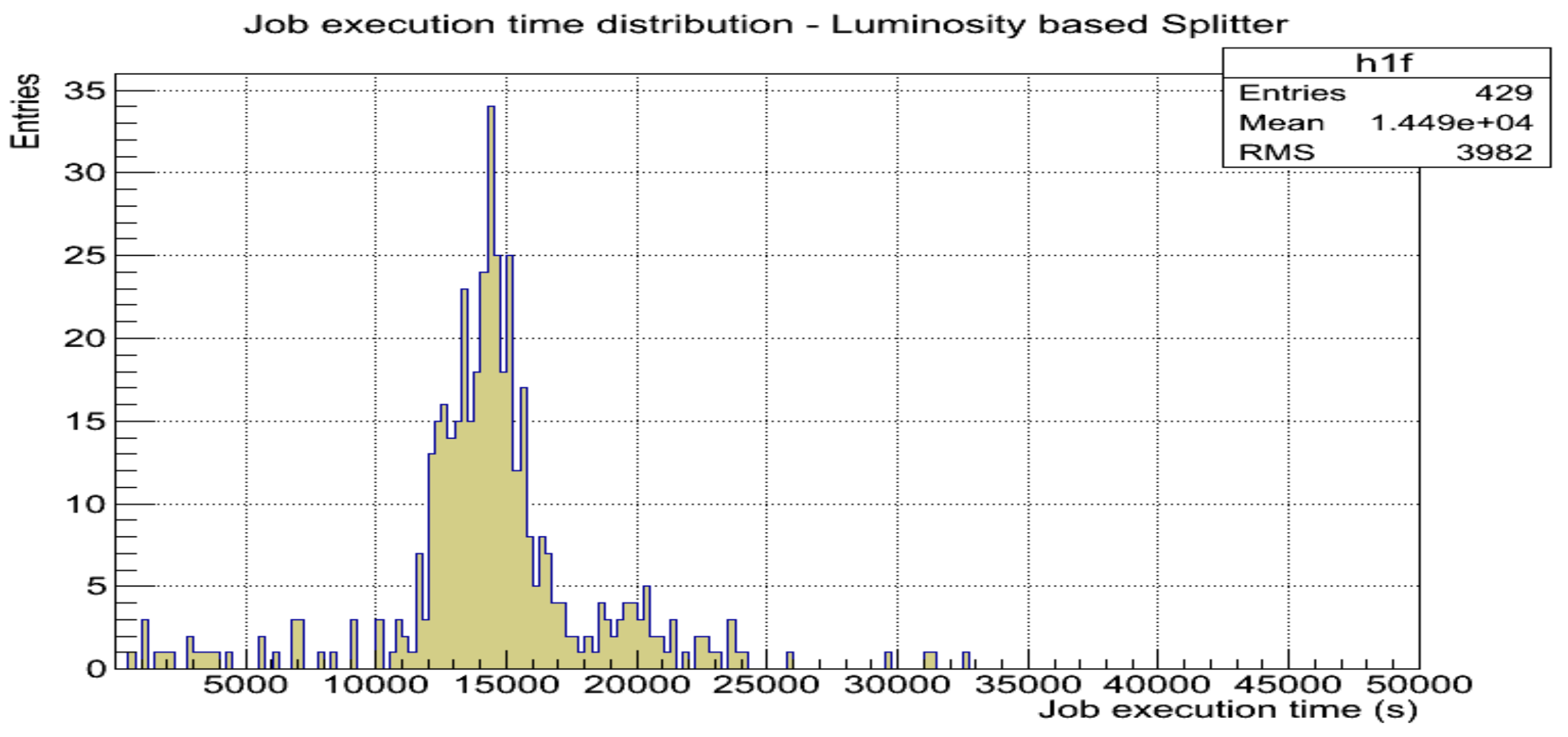
| Release | Primary dataset | Average Time per Event (s) | Number of points | Link to plot |
|---|---|---|---|---|
| CMSSW_5_3_8_patch3 | SingleMu | 23.78 | 45 | Link to plot |

Showing 1 to 1 of 1 entries

DashBoard - Time per Event for workflow requestors

The observed effect is demonstrated in the figures below. Both figures are result of the same Reconstruction Workflow. The first figure shows the effect for jobs splitted by the common EventBased algorithm, where the number of events is fixed for all jobs. The second figure shows the case where the splitting is done by the algorithm LuminosityBased algorithm, described above. It is expected from the Luminosity splitter, a more narrow distribution, as it makes the job execution time more uniform in a workflow. In cases where performance information is not yet available, it will split jobs exactly as EventBased would. The improvements shown here are expected to be even larger in future production systems as the performance information gathered is expected to increase.





## Conclusion

This initial study shows that it is feasible to predict the time-per-event behavior for reconstruction workflows of CMS. It was observed that heterogeneous farms introduce considerable systematic variations into the mechanism, and should be taken into account. We demonstrated that this information can be used in order to reduce the data processing tails, which have been until now a time-consuming problem impacting many time-critical prompt reconstruction workflows in the CMS Tier0. Furthermore, a job splitting algorithm has been developed that uses performance data dynamically according to the data-taking conditions.

## Acknowledgements

CMS Tier-0 Team, CMS Workload Management Development Team, WLCG DashBoard Team, UERJ Department of High Energy Physics, US-CMS Groups at the California Institute of Technology and Fermi National Accelerator Laboratory.

## References

[1] D Giordano and G Sguazzoni, CMS reconstruction improvements for the tracking in large pile-up events, doi:10.1088/1742-6596/396/2/022044

[2] The CMS Collaboration, Performance of CMS muon reconstruction in pp collision events at sqrt(s) = 7 TeV, arXiv:1206.4071

[3] T Hauth, V Innocente and D Piparo, Development and Evaluation of Vectorised and Multi-Core Event Reconstruction Algorithms within the CMS Software Framework, doi:10.1088/1742-6596/396/5/052065

## CHEP 2013 Conference

**Poster presenter : gutsche@fnal.gov,**
**Authors : samir@hep.caltech.edu, dkcira@caltech.edu, gutsche@fnal.gov**