PROCEEDINGS OF SCIENCE

Data Scouting in CMS

Dustin Anderson[*†]
California Institute of Technology
E-mail: dustin.james.anderson@cern.ch

Data scouting in collider experiments refers to the use of physics objects reconstructed online during data taking to perform searches and measurements. The technique, pioneered by the CMS experiment, allows events to be recorded for analyis at a rate of several additional kHz with negligible impact on total data volume. Dijet resonance searches have used data scouting to probe resonance masses far lower than those explorable with a standard offline physics analysis, and new developments for LHC Run II enable a wider range of analyses to take advantage of this new trigger paradigm. We describe the scouting technique, give an overview of its past use in CMS, and provide details on the implementation of the scouting streams being used in Run II. We also show results from the first scouting-based physics analysis with 13 TeV data.

38th International Conference on High Energy Physics
3-10 August 2016
Chicago, USA

[*]Speaker.
[†]On behalf of the CMS Collaboration.

PoS(ICHEP2016)190

1. Introduction: Data Rates and the CMS Trigger

The Large Hadron Collider (LHC) provides proton-proton bunch crossings with a center-of-mass energy $\sqrt{s} = 13$ TeV at a maximum rate of 40 MHz. The digital readout of the Compact Muon Solenoid (CMS) detector generates unprecedented amounts of raw collision data, much of which is potentially useful for physics analysis. However, full reconstruction of all collision events is not feasible with existing computing resources. CMS uses a two-level trigger strategy, consisting of a Level-1 Trigger (L1) [1] and a High-Level Trigger (HLT) [2], to reduce the data volume to approximately 1 kHz of full physics events.

The CMS L1 Trigger is implemented in hardware and performs a preliminary selection on physics events, reducing the 40 MHz rate of input events to approximately 100 kHz. The HLT is a software application running on a dedicated processor farm, consisting of more than 500 trigger paths, each selecting for a particular physics signature. Approximately 1 kHz of events are selected by the HLT and sent to the prompt reconstruction system.

To make a trigger decision for each event, the HLT performs physics object reconstruction and applies a selection based on the characteristics of the reconstructed objects. Trigger paths consist of sequences of producer modules, which build collections of objects; and filter modules, which reject events that do not fulfill certain criteria. Most trigger paths contain multiple phases of reconstruction and filtering, and generally the later phases of reconstruction yield physics objects whose performance is somewhat close to that of their offline counterparts. If an event reaches the end of any trigger path, it is accepted.

A number of factors restrict the trigger rates that the HLT can achieve:

- The amount of data storage space and the maximum throughput of the data acquisition system (DAQ)
- The capacity of the prompt reconstruction system
- HLT computing resources, which limit the complexity of the online reconstruction

Data parking, in which selected events are saved directly to tape with no prompt reconstruction, has been used in past LHC runs to increase the amount of data collected. While this strategy circumvents the restriction on rates imposed by the offline reconstruction system, it is limited by the other factors listed above.

In this note we describe data scouting, a technique that leverages online reconstruction of physics objects in order to attain extremely high trigger rates. Scouting complements data parking in a natural way and provides new opportunities for physics analysis outside the boundaries of the traditional trigger strategy.

2. Scouting: A New Trigger Paradigm

We seek to record CMS physics events at the highest possible rate while providing physics objects whose performance is suitable for offline analysis. To do this, we take advantage of the online reconstruction algorithms available at the HLT. HLT physics objects are less performant than their offline counterparts, but for many analyses (e.g. dijet resonance searches; see Section 3)

the difference does not significantly affect the sensitivity. Saving the objects reconstructed at the HLT, in lieu of those reconstructed offline, makes it much cheaper to record and store events.

This online reconstruction strategy is implemented via *data scouting* streams at the HLT. Each data scouting stream contains a number of trigger paths (scouting triggers), which perform event reconstruction and selection in the same way that ordinary HLT paths do. However, the selection criteria are much looser than for ordinary paths, and hence the rate of trigger firing is higher.

For events passing one or more scouting triggers, additional online reconstruction sequences are run in order to produce all physics objects necessary for an offline measurement or search. The produced objects are converted to a special compact event format and saved to disk. The data recorded by the scouting triggers is made available offline and can be used for physics analysis. For these events, no offline reconstruction is performed, and the raw data is not saved.

The scouting approach has the following advantages over the standard trigger strategy:

- The reduced, compact event format requires negligible space on disk and does not place any additional strain on the DAQ system. Events are 100 to 1000 times smaller on disk than the standard raw data format.
- No offline reconstruction is required; all reconstruction is performed online
- Scouting trigger paths can run 'in the shadow' of standard HLT paths, saving physics objects reconstructed by the standard HLT paths even for events that are rejected by those paths

Scouting has been used to increase the total number of CMS events recorded for physics by a factor of 2-6 beyond what the standard trigger strategy provides.

3. History of Scouting in CMS

The first scouting trigger was deployed at the CMS HLT during the last few pp fills of the 2011 LHC run. The trigger and associated stream collected data equal to 0.13 fb^{-1}. Events with H_T (defined as the scalar sum of jet transverse momenta) larger than 350 GeV were recorded and saved in a reduced format, containing only the set of jets reconstructed from particle-flow (PF) candidates by the HLT algorithm. The data were used to perform a search for heavy resonances decaying to dijets [3]. The search demonstrated sensitivity to resonances with masses between 0.6 and 0.9 TeV, a parameter region inaccessible to the standard CMS dijet resonance search.

After the 2011 test demonstrated the viability of the scouting technique, the strategy was repeated for the full 2012 CMS dataset. The scouting trigger selection cut was lowered to $H_T > 250$ GeV to accommodate an even larger rate of events. Due to CPU concerns related to the high rate, calorimeter jets were reconstructed and saved instead of PF jets.

The collected data, corresponding to 18.8 fb^{-1}, were used to perform a dijet resonance search analogous to the one carried out in 2011 [4]. The search results were interpreted as limits on the mass and coupling of a hypothetical leptophobic Z' resonance decaying to quarks. As indicated in Fig. 1, the limits were the strongest yet obtained for masses between 0.5 and 0.8 TeV, improving markedly on results from previous colliders.

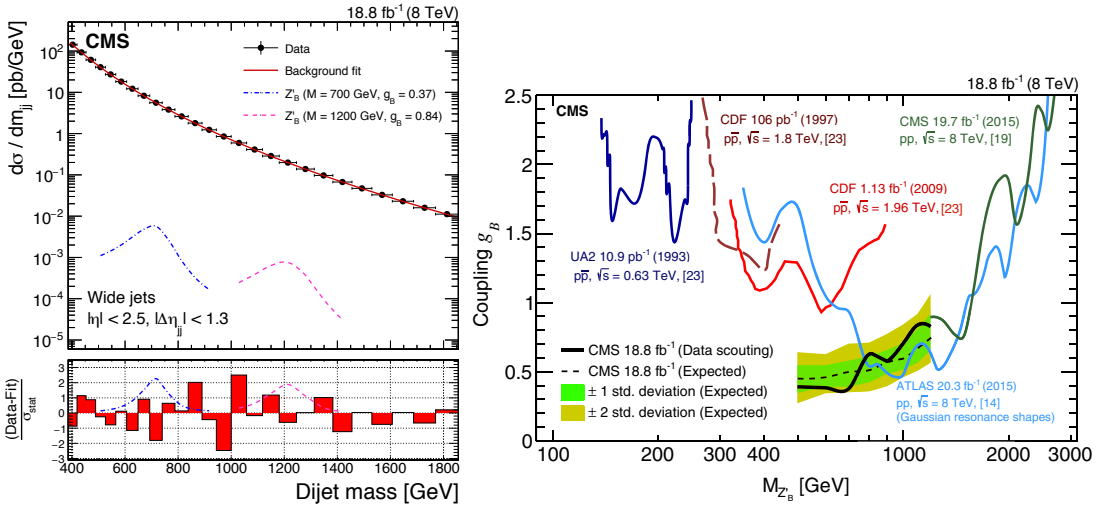

Figure 1: Left: dijet mass spectrum obtained from 18.8 fb^{-1} of scouting data collected in 2012, with fit to parameterized background function [4]. Right: limits from a variety of collider experiments on the mass and coupling of a hypothetical leptophobic Z'. The CMS search with 2012 scouting data (black solid line) is the best limit to date in the range from 0.5 to 0.8 TeV.

4. Scouting and Parking in LHC Run II

The success of data scouting in LHC Run I prompted an expansion of the strategy for Run II. The aim was to maintain the ability to search very low in H_T using calorimeter jets, as in 2012, while also providing an event format capable of supporting a broader range of scouting analyses. Two streams were deployed at the HLT for data taking in 2015 and 2016: one saving an event content based on calorimeter jets (the *calo-scouting* stream) and one saving an event content based on PF jets (the *PF-scouting* stream). Each stream features its own set of trigger paths and output datasets. Table 1 summarizes the rates and bandwidths measured for each stream in 2016.

New dedicated scouting event formats were designed and deployed for Run II as well. These use a set of minimalist C++ objects to store scouting physics objects as vectors of basic data types, ensuring low overhead and forward compatibility with future versions of the CMS software.

4.1 Calo-Scouting Stream

Triggers in the calo-scouting stream reconstruct jets from calorimeter deposits. The main signal trigger in this stream selects events with $H_T > 250$ GeV. Auxiliary prescaled trigger paths are also included in order to facilitate measurements of the signal trigger efficiency.

The event content for this stream includes the reconstructed calorimeter jets, the missing transverse momentum (MET), and ρ, a measure of the average energy density in the event. Local pixel track reconstruction provides b-tagging information for the jets. The size of this event content is about 1.5 kB on average for events passing the $H_T > 250$ GeV trigger.

4.2 PF-Scouting Stream

Triggers in the PF-scouting stream run the online version of the full PF sequence to reconstruct selected events. The main signal trigger in this stream selects events with $H_T > 410$ GeV. Auxiliary

triggers are also included to help measure the signal trigger efficiency. Additionally, the stream contains a trigger path selecting events with two muons having invariant mass above 3 GeV; this represents an effort to extend scouting to searches in final states other than fully-hadronic ones.

The event content for this stream includes the reconstructed PF jets, the PF MET, ρ, a collection of primary vertices, and all PF candidates with $p_T > 0.6$ GeV. It also contains electron, muon, and photon objects. The size of this event content is approximately 10 kB per event.

4.3 Monitoring the Quality of Scouting Data

To facilitate comparisons of the online physics objects to their offline reconstructed counterparts, and to use the CMS Data Quality Monitoring (DQM) framework to monitor the scouting data, a separate monitoring stream is deployed. This stream contains prescaled versions of all scouting triggers in the calo-scouting and PF-scouting streams. Events selected for this stream are saved in reduced scouting event format, but they are also sent to the CMS prompt reconstruction system for offline processing. This stream therefore contains online and offline versions of each event, which enables detailed object-by-object comparisons of the online and offline performance.

Data Stream	Rate at 10^{34} cm^{-2} s^{-1}	Bandwidth (MB/s)
Calo-Scouting H_T Signal	3700	11
PF-Scouting H_T Signal	720	9
PF-Scouting Dimuon Signal	480	6
Commissioning (PF + Calo)	30	< 1
Monitor	26	23

Table 1: Rates and bandwidths measured in 2016 data for various scouting data streams. The row marked 'Commissioninig (PF + Calo)' represents all auxiliary (non-signal) trigger paths in either the calo-scouting or PF-scouting data streams.

4.4 Data Parking

A main drawback of the scouting approach is the discarding of the raw data. In the event of a discovery (or a hint of one) in scouting data, it is useful to have the full raw data available so that the events can be analyzed in greater detail. This is accomplished using parked triggers, which send selected events directly to disk without reconstruction. In 2015 and 2016 the H_T events selected for the PF-scouting stream (corresponding to $H_T > 410$ GeV) were parked, to be reconstructed if needed in the future.

5. Dijet Resonance Search with Scouting Data in Run II

Data collected in the calo-scouting stream in 2016 were used to perform a search for dijet resonances with masses between 0.6 and 1.6 TeV [5]. The search was carried out on 12.9 fb^{-1} of data. The monitoring dataset described in Section 4.3 was used to perform detailed studies of the performance of the online jet objects and to re-calibrate the response of the HLT jets.

The results of the search were used to place limits on resonances decaying to gg, gq, and qq in a variety of models of new physics. The limits are shown in Fig. 2. The largest local significance observed was 2.6σ, for a gg resonance with mass 850 GeV.

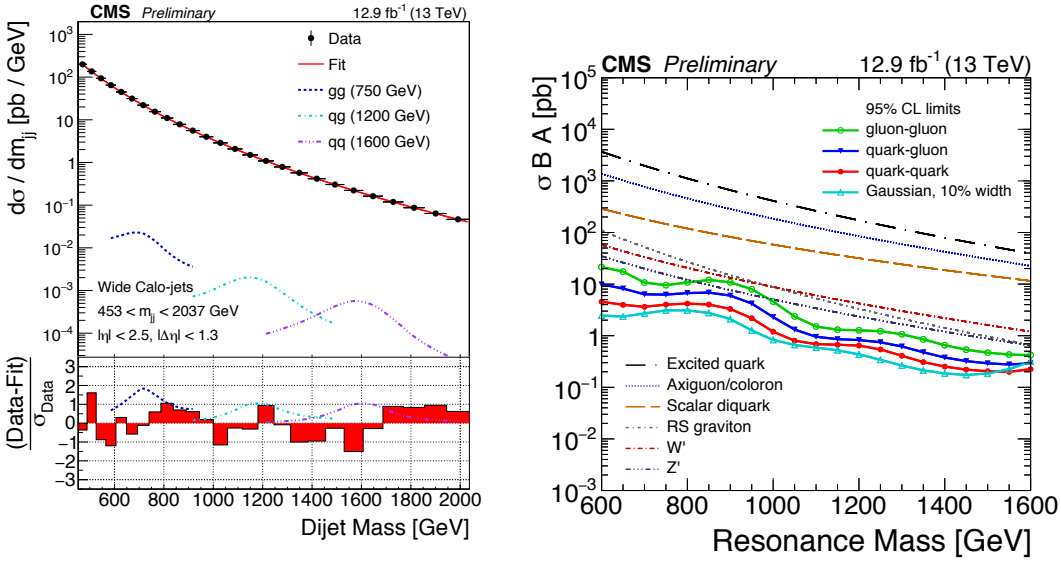

Figure 2: Left: dijet mass spectrum obtained from 12.9 fb^{-1} of scouting data collected in 2016, with fit to parameterized background function [5]. Right: limits placed on a variety of models of heavy resonances decaying to gg, qg, and qq using the 2016 scouting data.

6. Conclusion

Data scouting allows physics events to be collected at a rate dramatically higher than what is nominally achievable with the CMS trigger system. It is implemented with no changes needed to the basic HLT infrastructure and does not place a strain on the DAQ, disk resources, or the reconstruction system.

The two scouting streams deployed for Run II of the LHC strike a balance between specialization and versatility, with one stream oriented towards dijet resonance searches and the other designed to support arbitrary searches based on hadronic final states. The first physics result using the Run II scouting framework has been released, and it is hoped that analyzers will take full advantage of this tool to search in other regions previously unexplored at the LHC.

References

[1] J. Brooke [CMS Collaboration], PoS ICHEP **2012**, 508 (2013) [arXiv:1302.2469 [hep-ex]].

[2] V. Gori, Int. J. Mod. Phys. Conf. Ser. **31**, 1460297 (2014) doi:10.1142/S201019451460297X [arXiv:1403.1500 [physics.ins-det]].

[3] CMS Collaboration, CMS-PAS-EXO-11-094, http://cds.cern.ch/record/1461223?ln=en.

[4] V. Khachatryan *et al.* [CMS Collaboration], Phys. Rev. Lett. **117**, no. 3, 031802 (2016) doi:10.1103/PhysRevLett.117.031802 [arXiv:1604.08907 [hep-ex]].

[5] CMS Collaboration, CMS-PAS-EXO-16-032, http://cds.cern.ch/record/2205150?ln=en.