PROCEEDINGS OF SCIENCE

Deployment and Operations of the CMS Prompt Skimming System

Si Xie[*] **for the CMS Collaboration**
Massachusetts Institute of Technology
E-mail: sixie@mit.edu

This article will present a system deployed at Tier-1 computing centers used to provide skimmed datasets for commissioning and analysis activities, and operational experience gained from the first period of data taking. CMS has many automated and time critical workflows that are used to monitor and commission the detector. Most of the automated workflows are run at the Tier-0 computing facility at CERN, but CMS has recently deployed an infrastructure for automated workflow submission at the Tier-1 centers. The Tier-1 skimming system automatically tracks and submits workflows from CERN to the Tier-1 centers as the data arrives through the grid interfaces. The Tier-1 facilities do not have the same low latency access to the data, but there is a larger pool of processing and storage resources at the remote sites than at CERN. The prompt skimming system is an interesting example of utilizing the Tier-1 centers as a natural extension of the data acquisition system to the remote facilities.

35th International Conference of High Energy Physics - ICHEP2010,
July 22-28, 2010
Paris France

PoS(ICHEP 2010)512

[*] Speaker.

1. Introduction

The CMS computing model is based on a distributed, multi-tiered computing infrastucture [1]. The data recorded by the detector is first processed at the Tier-0 facilities at CERN. Data is subsequently distributed to the 7 large Tier-1 computing centers located around the globe, each containing several PetaBytes of tape storage and a few thousand CPU cores for data processing, intended for central reprocessing of the data. The prompt skimming system is deployed at the Tier-1 sites and the creation and submission of new prompt skimming jobs is triggered by arrival of the relevant input datasets at the Tier-1 site. The processed data is then distributed to the large number of smaller Tier-2 sites dedicated for physics analysis.

2. Prompt Skimming System

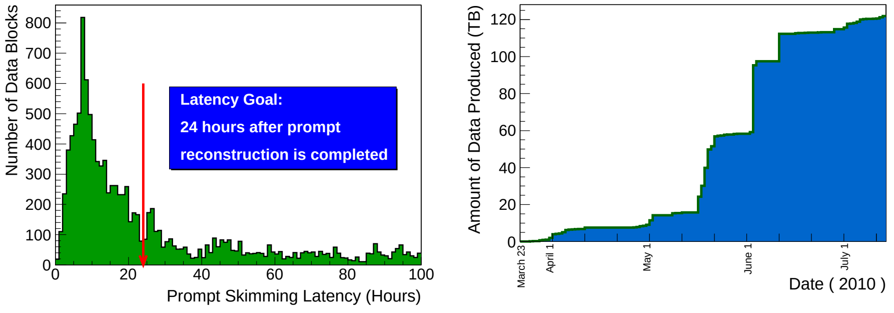
CMS categorizes data recorded by the detector and reconstructed at the Tier0 computing facilities into primary datasets immediately after the data is recorded. The categorization is based on trigger information, allowing datasets to be split based on physics interest. Even more specialized skims are produced, promptly, from the primary datasets in order to facilitate common analysis selections. This results in a significant reduction of the size of the data samples to be analyzed, and as a result significantly reduces the analysis latency. The prompt skimming system is currently implemented for various subdetector performance analyses, in particular prompt analyses of the electromagnetic calorimeter and muon systems, as well as various high priority physics analyses, such as jet energy scale correction measurements and momentum scale and tracking studies.

3. Deployment and Operations

The currently deployed prompt skimming system is a hybrid combination of components from CMS’ old workflow management system, based on message queues, and the new system based on a state machine. One of the main problems of the system based on message queues is the tendency to lose track of messages sent between various components during high load situations, resulting in loss of full accounting of all processing jobs. The state machine system is designed to address this deficiency by performing actions on workflows based only on well defined workflow states.

Prompt skimming jobs are triggered by the completion of the transfer of each data block, composed of a contiguous set of files, at the specified Tier-1 site. At the early stages of data taking, for safety of the data, we required that the data block has also been fully migrated to the tape storage system at the Tier-1 site before prompt skimming jobs are started. This requirement was later removed when it became clear that the integrity of the data files was not being compromised by the prompt skimming workflow.

All workflow management is currently executed remotely from dedicated servers at Fermilab, with some minimal access of the Tier0 database at CERN needed for synchronization. A latency goal of 24 hours after completion of the prompt reconstruction has been set. The average latency for the first half of data taking in 2010 is shown in Figure 1. The majority of promptly skimmed data has met the design goal. The long tail of the latency distribution is attributed to various operational issues in transferring the input data to the relevant Tier-1 site and migration of the data to the tape



(a) (b)

Figure 1: A histogram of the prompt skimming latency per file block is shown in (a). The long tails are mainly attributed to issues in data transfers and tape migration. The Volume of promptly skimmed data produced as a function of time is shown in (b).

based mass storage system. The volume of promptly skimmed data produced by the system as a function of time is also shown.

A number of operational problems were met and resolved during data taking in the first half of 2010. In the early stages of data taking, the prompt skim workflow output was too large due to a high skim efficiency, which caused overload of the disk and tape writing capabilities of the Tier-1 sites. This was resolved by redefinition of the primary datasets and modification of the skim workflows. The requirement for the data block to be completely migrated to the tape system before prompt skimming jobs are started was also found to be unnecessary, and was removed for later period of data taking.

4. Conclusion

The CMS automated prompt skimming system has been presented in this article. A number of issues in its deployment and operations have been discussed. The plan for the next year of data taking is to move to a fully operational state machine for job tracking, for which the prompt skimming workflows will be one of the first implementations in production.

References

[1] CMS Collaboration, "CMS Data Processing Workflows during an Extended Cosmic Ray Run", J. Instrum. 5 (2010) T03006 (arXiv:0911.4842)