The CMS Event Builder and Storage System

Gerry Bauer[7], Barbara Beccati[3], Ulf Behrens[2], Kurt Biery[6], Angela Brett[6], James Branson[5], Eric Cano[3], Harry Cheung[6], Marek Ciganek[3], Sergio Cittolin[3], Jose Antonio Coarasa[3,5], Christian Deldicque[3], Elizabeth Dusinberre[5], Samim Erhan[4], Fabiana Fortes Rodrigues[1], Dominique Gigi[3], Frank Glege[3], Robert Gomez-Reino[3], Johannes Gutleber[3], Derek Hatton[2], Markus Klute[7], Jean-Francois Laurens[3], Constantin Loizides[7], Juan Antonio Lopez Perez[2,6], Frans Meijers[3], Emilio Meschi[3], Andreas Meyer[2,3], *Remigius K Mommsen*[6], Roland Moser[3], Vivian O'Dell[6], Alexander Oh[3], Luciano Orsini[3], Vaios Patras[3], Christoph Paus[7], Andrea Petrucci[5], Marco Pieri[5], Attila Racz[3], Hannes Sakulin[3], Matteo Sani[5], Philipp Schieferdecker[3], Christoph Schwick[3], Josep Francesc Serrano Margalef[7], Dennis Shpakov[6], Sean Simon[5], Konstanty Sumorok[7], Marco Zanetti[3]

[1]CEFET, Rio de Janeiro, Brazil. [2]DESY, Hamburg, Germany [3]CERN, Geneva, Switzerland [4]UCLA, Los Angeles, California, USA [5]UCSD, San Diego, California, USA [6]FNAL, Chicago, Illinois, USA [7]MIT, Cambridge, Massachusetts, USA

presented at CHEP 2009 in Prague, Czech Republic

Introduction

The CMS event builder collects event fragments from approximately 500 detector front-end readout modules (FEDs) and assembles them into complete events at a maximum rate of 100 kHz. The complete events are handed to the High Level Trigger (HLT) processors running offline-style algorithms to select interesting events. Accepted events are transfered to the storage manager (SM) which temporarily stores the events on disk at a peak rate of 2 GB/s until they are permanently archived offline. In addition, events and data-quality histograms are served by the storage manager application to online monitoring clients.

Operational Experience

The readout system has been successfully used to commission the detector, to collect more than 370×10^6 cosmic ray events, and to record the first LHC beam events in September 2008.

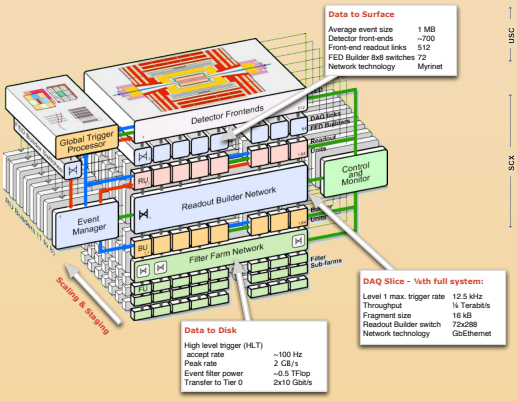
The event builder performs to the design specifications of 100 GB/s and the system has proven to be very flexible:

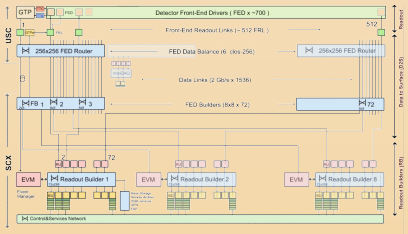
- The event builder can be subdivided into multiple independent readout systems that can be concurrently used to readout subsets of detector components.
- The system can be configured to focus on performance and reliability issues related to
  - high data throughput
  - large number of input channels
  - high CPU power in the high-level trigger
  - high data rate to local disk.
- Data payloads can be injected at various points in the system to test its functionality and performance.

The 3-month long running of the experiment in fall 2008 highlighted several areas of improvement:

- The event builder needs to be tolerant against missing or corrupted front-end data.
- The error and alarm handling should be based on a uniform and system-wide approach.
- The Storage Manager application should handle CPU-intensive tasks in independent threads to assure that they do not block the readout (see below).

These improvements are being deployed and tested in the full system and will be available for the first physics run in fall 2009.



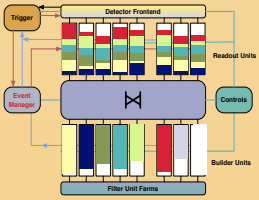


CMS Event Building – 2 Stages

CMS employs only 2 trigger levels: a first level trigger based on custom hardware, and a High Level Trigger (HLT) using commodity PCs. The assembly of event fragments into complete events takes place at a level 1 accept rate of 100 kHz. In order to cope with the aggregated bandwidth of 100 GB/s, and to provide a scalable system, a 2-tier approach for the event building was chosen.

FED Builder

Each FED-Builder assembles data from 8 front-end links into one super-fragment using redundant Myrinet switches. The super-fragment is delivered to one of up-to 8 independent readout slices, where it is buffered on commodity PCs, the readout units (RUs).



DAQ Slices

The DAQ slices are independent systems, which are in charge of building full events out of super-fragments, performing the physics selections, and forwarding the selected events to mass storage (SM). The independence of each DAQ slice has several advantages:

- Relaxed requirements:
  Each slide needs to handle a reduced event rate of 12.5 kHz instead of 100 kHz
- Staging:
  Slices can be added as needed
- Robustness:
  In case that one slice fails, data taking does not stall as other slices continue to process events
- Technology independence:
  Different technologies can be used for the FED-Builder and for the DAQ slice, or the technology can even differ between DAQ slices

The existing system uses the TCP/IP protocol over Gigabit Ethernet. Each readout unit (RU) is connected with two or three links ("rails") via a large switch to builder units (BUs). In order to achieve the necessary performance, a specific configuration of the networks with source to destination VLANs is required. The event building in each slice is controlled by one event manager (EVM), which receives the level 1 trigger information. The BUs store the complete events until the Filter Units (FUs) running the High Level Trigger (HLT) algorithms either rejected or accepted the events. Accepted events are sent to the Storage Manager (SM), which writes the events to disk.

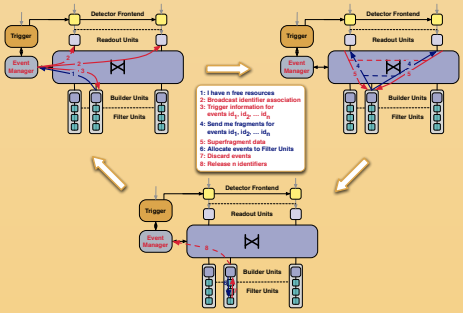
XDAQ Framework

The RU-Builder and the Storage Manager applications use the XDAQ framework. XDAQ is middleware that eases the development of distributed data acquisition systems. It provides services for data transport, configuration, monitoring, and error reporting. The framework has been developed at CERN and builds upon industrial standards, open protocols and libraries.

For details see the talk "CMS Data Acquisition System Software" by Johannes Gutleber.

RU-Builder Protocol – Token Passing

The RU-builder consists of 3 separate services: the readout unit (RU) buffers the event fragments during the assembly, the builder unit (BU) assembles the event, and the event manager (EVM) interfaces with the trigger and orchestrates the data flow. The applications exchange I2O binary messages (I2O stands for Intelligent Input Output) over the network and use First-In-First-Out queues (FIFOs) to keep track of requests, trigger data, and event data.

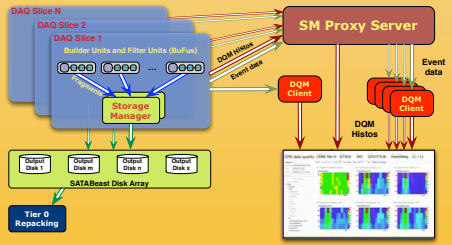


Storing Data on Disk – The Storage Manager

The Storage Manager (SM) is located at the end of the online data chain. It receives events accepted by the High Level Trigger (HLT) running on the Filter Units (FUs). The accepted events are split over several I2O binary messages (fragments) and send over a dedicated Gigabit-Ethernet network to the SM. Each DAQ slice has one or two SM applications. The SM reassembles the fragments into complete events and writes them to disk. Events are stored in one or multiple files, depending on the triggers that fired for the event (the trigger bits) and the HLT process that selected the event. These files are buffered on a local disk array (RAID-6) and subsequently copied to the central computing center (Tier 0), which repacks the events into larger files. These files are then fed to the offline event-reconstruction farms.

A subset of the events is sent to online consumers, using a HTTP request/response loop. Consumers either connect directly to one SM, or they connect to a proxy server which aggregates the data from all SMs in all DAQ slices. The consumers use this data for online data quality monitoring (DQM), for calibration purposes, or for displaying the events.

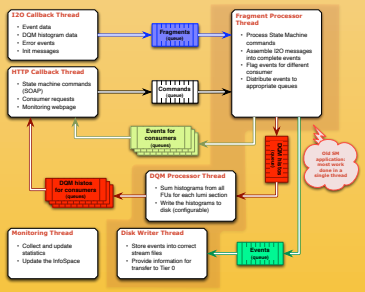
In addition, the FUs produce a set of histograms to monitor the data quality and trigger efficiencies. These histograms use events which are only available prior or during the event selection. These histograms are sent to the SM at each luminosity section boundary (every 93 seconds). The SM sums the histograms that he received from all FUs in the slice. It forwards the summed histograms to the proxy server. The proxy server receives the histograms from all SMs in all DAQ slices and sums them up again. The CMS data quality application retrieves the summed set of histograms for monitoring.



Redesign of the Storage Manager Application

The current implementation of the Storage Manager (SM) exposed some weaknesses during the global run. The main issue is that most of the work is done in a single thread. This caused dead-time in the system each time the SM was summing the DQM histograms for each luminosity section, i.e. every 93 seconds. In addition, the code evolved over several years and had been adapted to changing requirements. This resulted in a code base that was difficult to understand and hard to maintain. Therefore, the opportunity of the delayed LHC startup was taken for a redesign and re-factoring of the existing code.

The new design uses individual threads for the different tasks of the storage manager. The events are pushed into separate queues which are handled by dedicated threads. Care is taken that the main task of receiving events from the high level triggers and writing them to the appropriate files (streams) is not blocked by the CPU-intensive task of summing DQM histograms, or by rather unpredictable requests from consumers of events or DQM histograms.

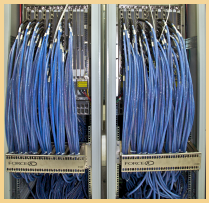


FED Builder

- The heart of the FED Builder is a set of six Myrinet switches installed underground, close to the detector.
- The data is transported to the surface using 1536 optical links operating at 2Gb/s.
- Another set of 72 8x8 Myrinet switches is located close to the event builder and High Level Trigger (HLT) filter farm (see right).

RU Builder Switch

- 8 Force-10 E1200 switches (one for each DAQ slice)
- 6 90-port linecards / switch
- 11 Gb/s duplex / port ~2 times oversubscribed
- Cabling is done to have full throughput on all ports using the fact that the traffic is basically uni-directional



RU Builder & HLT Farm

- 640 RU PCs (32 racks)
  Dell PE 2850 dual dual-core, 2 GHz, 4 GB memory
- 720 BU-FU PCs (24 racks)
  Dell PE 1950 dual quad-core, 2.6 GHz, 16 GB memory

Storage Manager

- A Force 10 Gigabit Ethernet switch connects the Storage Manager to the HLT processors.
- A separate Gigabit Ethernet switch is used for transfer to Tier 0.
- The Storage Manager's hardware provides a data buffer of 300 TB, which is equivalent of several days of data taking.
- The data is buffered on NexSan SATABeasts (RAID-6 disk arrays) connected through 2 Fibre Channel switches (QLogic SanBox 5600).