Operational Experience with the CMS Data Acquisition System

Hannes Sakulin, CERN/PH
on behalf of the CMS DAQ group
Compact Muon Solenoid

- General purpose detector at the LHC
- 55 million readout channels
  - Event size of 1MB
- Proton physics
  - At 7 TeV in 2010/11
  - At 8 TeV in 2012
- Heavy Ion physics
  - In 2010 & 2011
Two-level trigger concept

Collisions rate 40 MHz

First Level 100 kHz

Rate on tape few 100 Hz

High Level Trigger (HLT) ~ s

First Level: Clock driven Synchronous Triggers

Higher Levels: Event Driven Asynchronous Triggers
CMS DAQ requirements

- Read out 700 detector front-ends (max. average fragment size 2 kB)
- Build complete events at 100 kHz (L1 trigger rate)
- Make them available to a filter farm of O(10000) cores
- Store 100’s of Hz to disk (10’s of TB/day)
- Scalable system employing commercial components wherever possible

Proprietary / Commercial: Front-Ends, VME, PCI, PC servers, networks, Protocols, OS
CMS two-stage event builder

- **Front-end Readout Links**
  - **Super-fragment builder (Myrinet)**: 100 kHz
  - **Readout Builders (Gigabit Ethernet)**: 8 x 12.5 kHz

- **Custom Commercial hardware**
  - 2k 4k fragment
  - 16k super fragment

- **Front-end Output Links**
  - **Detector Front-End Drivers (FED x ~700)**
  - **Trigger Throttle System (TTS), Fast Merging Module (FMM)**

- **DAQ Slices**
  - **DAQ Slice 1**
  - **DAQ Slice 2**
  - **DAQ Slice 8**

- **Event Manager**
  - 12.5 kHz

- **Scale readout bandwidth**: No. DAQ systems (1 to 8 x 12.5 kHz)
CMS two-stage event builder

Operational Experience with the CMS DAQ System, CHEP 2012, 21 May 2012
H. Sakulin / CERN PH

Superfragment builder (Myrinet)

100 kHz

Frontend Readout Links

16k Readout Builders (Gigabit Ethernet) 8x 12.5 kHz

DAQ Slice 1

DAQ Slice 2

DAQ Slice 8

Control & Services Network

Scale readout bandwidth: No. DAQ systems (1 to 8 x 12.5 kHz)
CMS two-stage event builder

Front-end Readout Links

Super-fragment builder (Myrinet) 100 kHz

Readout Builders (Gigabit Ethernet) 8x 12.5 kHz

DAQ Slice 1  DAQ Slice 2  DAQ Slice 8
CMS two-stage event builder

Trigger Throttling System

Loss-less DAQ always builds complete events back-pressure in case of congestion
CMS two-stage event builder

Front-end Readout Links

Emulator Mode: FRLs generating data

Super-fragment builder (Myrinet) 100 kHz

Readout Builders (Gigabit Ethernet) 8x 12.5 kHz

Scale readout bandwidth: No. DAQ systems (1 to 8 x 12.5 kHz)
Installed hardware

- Custom compact PCI Modules
  - 512 Frontend Readout Links
  - 60 Fast merging modules (trigger throttling)
- Myrinet Switches
  - 12 clos-256 enclosures
  - 1536 2.5 Gb/s links underground to surface
- “Readout Unit” PC nodes
  - 640 times dual 4-core E5130 (2007)
  - Each node has 3 links to GbE switch
- Gbe Switches
  - 8 times F10 E1200 routers
  - In total ~4000 ports (1 Gb/s)
- Event builder–output + HLT nodes (“BU-FU”)
  - Currently ~13000 cores, 26 TB RAM
  - Extensible – see later
- Storage Manager
  - 16 PCs
  - Storage Area Network (NexSan SataBeasts), 300 TB
  - 2.1 GB/s write speed (2.6 GB/s w/o Tier0-Transfers)
CMS DAQ Software

Run Control System – Java, Web Technologies

- Run Control Web Application
  - Apache Tomcat Servlet Container
  - Java Server Pages, Tag Libraries, Web Services (WSDL, Axis, SOAP)

- Function Manager
  - Node in the Run Control Tree
  - defines a State Machine & parameters
  - User function managers dynamically loaded into the web application

GUI in a web browser
- HTML, CSS, JavaScript, AJAX

XDAQ Framework – C++, XML, SOAP

- XDAQ applications control hardware and data flow

XDAQ is the framework of CMS online software
- It provides Hardware Access, Transport Protocols, Services etc.

XDAQ Application

~20000 applications to control
Top level control Web - GUI

- GUI is a web-page
- Top level is Global state machine, aware of LHC states, eg stable beams
- Trigger configuration and clock source (LHC/local)
- Control of individual sub-systems for fast recovery
- Cross-checks and warnings to help the DAQ shifter
Monitoring

- Monitoring tuples and error messages
  - O(2000) PCs
  - O(20000) applications

- Collect and aggregate
  - Hierarchy of collectors
  - Load balancing
  - Latency ~seconds

- Access service for
  - Error reporting GUI
  - Visualization applications
  - DAQ Doctor (“expert system”)

Poster #139 / session 2: Distributed error and alarm processing in the CMS data acquisition system
The DAQ doctor

- Constantly analyzes monitoring information
- Detects abnormal situations
  - Warns the shift crew with Text & Audio alerts
  - Gives recovery instructions
  - Now also creates new DAQ configurations
  - Dumps diagnostic info for post-mortem analysis
- All diagnostic information is archived & categorized by sub-system
System status display

700 Front-end readout sub-systems
500 x 4 Gb/s data channels

8x (80 x 158) DAQ systems

8 x 1600 cores HLT farms
16 x 20 Terabyte local mass storage
Data acquisition in operation

Operational Experience with the CMS DAQ System, CHEP 2012, 21 May 2012

H. Sakulin / CERN PH

30/04/11 PROTON PHYSICS DAQ state Running Run Number 163758
Sat 18:04:30

Lv1 rate 30.929 kHz

Ev. <Size> kB DeadTime(AB) Acc. Hz (%) HLT <CPU>
401.0 [224.8] 0.673 % 30991.1 (100.0) % 21.96 %

SM streams

Data to Surface

Sub-System State FRL FED IN
TRG Running 3 3 3
CSC Running 6 6 6
DAQ Running 0 0 0
DQM Running 0 0 0
DT Running 6 6 6
ECAL Running 54 54 54
ES Running 39 39 39
HCAL Running 26 26 26
HFLUMI Running 6 6 6
PIXEL Running 40 40 40
RPC Running 3 3 3
SCAL Running 1 1 1
TRACKER Running 250 438 438
CASTOR Running 3 3 3

Data Flow

LHC mode: PROTON PHYSICS, STABLE BEAMS
Run# 163758 history time window [2.6 h]

UTC time 30/04/11 16:04:30
Local time: Geneva 18:04, Los Angeles 09:04, Chicago 11:04, Moscow 20:04, Beijing 01:04
Operational Experience with the CMS DAQ System, CHEP 2012, 21 May 2012

Operational Efficiency

CMS control room, Cessy, France …
CMS Central DAQ efficiency, 2011 - pp

CMS central DAQ availability during stable beams: 99.7 %
CDAQ down time : < 4 hours
Luminosity lost: ~ 0.5% of delivered
Central DAQ Down times

- Software (24 down times, 3 hours)
  - Due to surfacing and newly introduced bugs
    - Often related to features that were added to the original design
  - Usually fixed as soon as identified

- Hardware (8 down times, 1 hour)
  - 1 Broken Myrinet link
  - 1 Broken Gigabit-ethernet switch line card
  - 1 Broken control network switch
  - 203 PC failures

Only 1 hour of down time due to HW? => Resilience
Resilience features of CMS DAQ

- **Automatic restart** of crashed Event Filter processes during an ongoing run

- **Tolerance** against crashed processes & machines
  - Data flow applications / machines
    - Builder & Filter Units, Storage Manager
    - Run continues with reduced throughput
  - Applications controlling custom hardware
    - Run continues with degraded monitoring

- **Slice Masking**: fast workaround for single points of failure in a DAQ Slice (Readout Units, GbE switches, etc.)
  - Mask the slice and continue with 7/8th of capacity
  - Requires stop/start of the run
Resilience features of CMS DAQ (2)

- Fast Configuration Change
  - Mask a broken machine
    (except those controlling custom hardware)
  - Mask a rail in one leg of the Myrinet Super-Fragment Builder
  - Use only 1 out of 2 racks of Storage Managers

- Tool: CMS DAQ Configurator
  - Until mid 2010: Several tools needed, manual bookkeeping
    new configuration in ~10 minutes
  - mid 2010 - 2011: One-Step tool with blacklist database
    new configuration in ~2 minutes
  - Since 2012: One-Step tool automatically launched by DAQ Doctor
    new configuration in ~40 seconds

- Configuration change requires a run stop/start
Over-all CMS data taking efficiency 2011

Overall CMS data taking efficiency 2011

pp: 91.2 %

PbPb: 94.4 %

In part due to Single-Event upsets

Lumi lost in down times (pp)
New: Automatic Recovery from Single Event Upsets

- Frequent sub-detector DAQ failures due to Single-Event upsets observed towards the end of 2011 with increasing instantaneous luminosity

- Recovery typically needed re-configuration of the system

New in 2012: Automatic Single-event-upset Recovery Mechanism

- Coordinated by top-level run control
- Sub-detector detects SEU problem and notifies top-level run control
- Top-level Run Control
  - Invokes a recovery transition
    - On the requesting sub-system
    - Other sub-systems may do preventive actions in the shadow
Impacting over-all efficiency: startup time

- Start of data taking session (starts all software): < 3 minutes
- Run stop & start: 1 min 15 seconds

During stable beams, Apr 13 – May 2, 2012
Evolution of operating conditions
Evolution of operating conditions

- **Design**
  - $L = 10^{34} / \text{cm}^2\text{s}$, 25 ns bunch spacing, 14 TeV
  - Pile-up of 20
  - DAQ at 100 kHz

- **2012**
  - $L = 7 \times 10^{33} / \text{cm}^2\text{s}$ (expected),
    - 50 ns bunch spacing, 8 TeV
  - Pile-up of 35 (~2x design)
  - DAQ at 100 kHz
Event with 30 reconstructed vertices
Can we handle the event size?
Can we handle the event size?

Globally yes, but have to look at individual Inputs & Super-fragment builders.

CMS design event size: 1 MB.

Expected 2012: 30 primary vertices corresponding to pile-up of 35.
Bandwidth at various stages

SLINK: **400 MB/s** (64b @ 50 MHz)  
✔ No problem

Myrinet link: **500 MB/s** (2 rails of 2.5 Gbit/s)  
✔ No problem

Myrinet Cross-bar switch: ~**260 MB/s**  
Wormhole-routed  
No buffering in switch  
Head-of-line blocking reduces throughput by up to 50% when no traffic-shaping applied  
Some Super-Fragment Builders critical

Gigabit Ethernet: 3 rails: **375 MB/s**  
Ethernet switches have internal buffer shared memory – no HOL blocking  
✔ No problem
32 inputs (Pixel sub-system) may exceed available throughput at pile-up of 35

Solution: super-fragment builders with fewer than 8 inputs for pixel combine some smaller super-fragment builders,
Throughput in Heavy-Ion Operation
CMS Experiment at LHC, CERN
Data recorded: Mon Nov 8 11:30:53 2010 CEST
Run/Event: 150431 / 630470
Lumi section: 173
# Proton physics – Ion physics

<table>
<thead>
<tr>
<th></th>
<th>Proton physics</th>
<th>Ion Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero suppression for Si-strip tracker</td>
<td>In FED (hardware)</td>
<td>In HLT farm (software)</td>
</tr>
<tr>
<td>Fragment size</td>
<td>2 kB</td>
<td>50 kB (100 kB after merging)</td>
</tr>
<tr>
<td>Event size</td>
<td>1 MB</td>
<td>20 MB</td>
</tr>
<tr>
<td>Max trigger rate</td>
<td>100 kHz</td>
<td>3.5 kHz</td>
</tr>
<tr>
<td>Max. DAQ throughput per input (8x8 super-fragment building)*</td>
<td>260 MB/s</td>
<td>350 MB/s (DAQ settings tuned for large fragments)</td>
</tr>
</tbody>
</table>

*Log-normal distributed event size
std-dev = average
DAQ performance at start of 2011 HI fill

- **2.7 kHz L1 rate**
- **20 MB / event**
- **Zero-suppression in HLT farm -> 1MB**
- **560 MB/s to disk**

**2010 HI run: ZS offline / ROOT compression in HLT**

11 MB / event, 1.8 GB/s to disk
High-Level Trigger
Filter Farm deployment strategy

- High-level trigger based entirely on commodity hardware
- Buy the processing power just in time
  - Better value for money
- Computing requirements evolve with LHC luminosity
  - Higher luminosity requires higher selectivity
    ➔ more complex algorithms
  - Higher luminosity ➔ more pile-up ➔ more time consuming tracking
- Challenge: increasing number of cores per machine
High-Level Trigger Software

- Trigger algorithms are processed with CMS offline software framework CMSSW
- 1 Process per core / per hyperthread but limited memory available
- Copy On Write: 1) Prototype process loads configuration and conditions 2) Child processes are forked
- Coupling between XDAQ and CMSSW very tight
  - same compiler, same process

Poster #219 / session 2: The CMS High Level Trigger System: Experience and Future Development
# HLT farm evolution

### 2009:

<table>
<thead>
<tr>
<th>Original HLT System</th>
<th>2011 extension</th>
<th>2012 extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dell Power Edge 1950</td>
<td>Dell Power Edge c6100</td>
<td>Dell Power Edge c6220</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Form factor</th>
<th>1 motherboard in 1U box</th>
<th>4 motherboards in 2U box</th>
<th>4 motherboards in 2U box</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPUs per mother-board</td>
<td>2x 4-core Intel Xeon E5430 Harpertown, 2.66 GHz, 16GB RAM</td>
<td>2x 6-core Intel Xeon X5650 Westmere, 2.66 GHz, hyper-threading, 24 GB RAM</td>
<td>2x 8-core Intel Xeon E5-2670 Sandy Bridge, 2.6 GHz, hyper-threading, 32 GB RAM</td>
</tr>
<tr>
<td>#boxes</td>
<td>720</td>
<td>72 (=288 motherboards)</td>
<td>64 (=256 motherboards)</td>
</tr>
<tr>
<td>#cores</td>
<td>5760</td>
<td>3456 (+ hyper-threading)</td>
<td>4096 (+ hyper-threading)</td>
</tr>
<tr>
<td>cumulative #cores</td>
<td>5.6k</td>
<td>9.1k</td>
<td>13.2k</td>
</tr>
<tr>
<td>cumulative #CMSSW</td>
<td>5k</td>
<td>11k</td>
<td>20k</td>
</tr>
</tbody>
</table>
HLT machine performance with HLT playback

~30% gain from hyper-threading

HLT menu for $5 \times 10^{33}/(\text{cm}^2\text{s})$, recent data sample & software
## HLT farm evolution

<table>
<thead>
<tr>
<th></th>
<th>Original HLT System Dell Power Edge 1950</th>
<th>2011 extension Dell Power Edge c6100</th>
<th>2012 extension Dell Power Edge c6220</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Form factor</strong></td>
<td>1 motherboard in 1U box</td>
<td>4 motherboards in 2U box</td>
<td>4 motherboards in 2U box</td>
</tr>
<tr>
<td><strong>CPUs per motherboard</strong></td>
<td>2x 4-core Intel Xeon E5430 Harpertown, 2.66 GHz, 16GB RAM</td>
<td>2x 6-core Intel Xeon X5650 Westmere, 2.66 GHz, hyper-threading, 24 GB RAM</td>
<td>2x 8-core Intel Xeon E5-2670 Sandy Bridge, 2.6 GHz, hyper threading, 32 GB RAM</td>
</tr>
<tr>
<td><strong>#boxes</strong></td>
<td>720</td>
<td>72 (=288 motherboards)</td>
<td>64 (=256 motherboards)</td>
</tr>
<tr>
<td><strong>#cores</strong></td>
<td>5760</td>
<td>3456 (+ hyper-threading)</td>
<td>4096 (+ hyper-threading)</td>
</tr>
<tr>
<td><strong>cumulative #cores</strong></td>
<td>5.6k</td>
<td>9.1k</td>
<td>13.2k</td>
</tr>
<tr>
<td><strong>cumulative #CMSSW</strong></td>
<td>5k</td>
<td>11k</td>
<td>20k</td>
</tr>
</tbody>
</table>

### Per-event CPU budget @ 100 kHz:

- **2009:** ~50 ms / evt
- **2011:** ~100 ms / evt
- **2012:** ~150 ms / evt

(CPU budgets are on 1 core of an Intel Harpertown)
States of HLT nodes at start of a pp fill before extension 2

HLT farm almost fully utilized at start of fill (since September 2011)

Algorithms are tuned for available computing power

Fill 2536, 20 Apr 2012

$L_{\text{peak}} = 6.1 \times 10^{33}/(\text{cm}^2\text{s})$
HLT states with HLT extension 2

- Ready for higher instantaneous luminosity and more complex algorithms

HLT extension-2 in 5 out of 8 DAQ slices

Fill 2645, 19 May 2012

\[ L_{\text{peak}} = 6.1 \times 10^{33} / (\text{cm}^2\text{s}) \]
Summary

- CMS DAQ system building events at 100 kHz in 2 stages
  - 1MB event size, 100 GB/s throughput

- Central DAQ availability 2011: 99.7%

- Continuous effort to improve CMS over-all efficiency

- Increased data volume due to higher pile-up with 50 ns LHC bunch spacing can be handled

- HLT farm being extended as required
  - reached 13000 cores this month. Ready for higher luminosity.

Beyond 2012: see next talk …
Thank You
Bonus track
## Comparison of HLT machines

<table>
<thead>
<tr>
<th></th>
<th>Harpertown</th>
<th>Westmere</th>
<th>Sandy Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Xeon E5430, 2.66 GHz</td>
<td>Xeon X5650, 2.66 GHz</td>
<td>Xeon E5-2670 2.6 GHz</td>
</tr>
<tr>
<td>#cores</td>
<td>8 (2x4)</td>
<td>12 (2x6) + HT</td>
<td>16 (2x8) + HT</td>
</tr>
<tr>
<td>SPEC int (max)</td>
<td>25</td>
<td>37 (= 25 * 1.5)</td>
<td>52 (= 25 * 2.1)</td>
</tr>
<tr>
<td>HEP Spec</td>
<td>73</td>
<td>208</td>
<td>386</td>
</tr>
<tr>
<td>CPU burner test*</td>
<td>1.0</td>
<td>3.6</td>
<td>5.4</td>
</tr>
<tr>
<td>Eg Action 11 test (CPU + memory)</td>
<td>1.0</td>
<td>2.2</td>
<td>3.3</td>
</tr>
<tr>
<td>HLT 2011</td>
<td>1.0</td>
<td>2.4</td>
<td>-</td>
</tr>
<tr>
<td>HLT playback*</td>
<td>1.0</td>
<td>2.8</td>
<td>3.9</td>
</tr>
</tbody>
</table>

* Does not include event building
CMS DAQ installation

Data to surface
- 12 Myrinet switches (clos-256)
- 1536 D2S links (2.5 Gb/s)
- 2 Tb/s total

Detector readout.
- 650 Slink/FMM cables
- 500 FRL + 60 FMM modules
- 60 FRL/FMM crates
- 200 DAQ/DCS PCs

High Level Trigger Farm
- 110 water cooled racks
- Extensible design
- Currently 13000 cores, 26 TB RAM
- Storage to disk at 2.1 GB/s
- 300 TB Mass storage

Readout Builders
- 640 Readout Unit PCs (4-core)
- 8 Force-10 GBE switches
  4000 ports in total
A general and expandable architecture has been deployed for the experiments’ Run control and monitoring largely based on the emerging Internet technology developed in the field of WWW services.
Two-stage event building architecture

- Level-1 Trigger: 100 kHz
- EVM Readout
- Front-End Drivers: ~650, ~500
- Frontend-Readout Links...
- SLINK-64 Cable: up to 10m, 400 MB/s
- Frontend Readout Link reads 1 or 2 links
- Interfaces to Myrinet NIC

Sender Mezzanine

500x
200 MB/s
Two-stage event building architecture

Superfragment builder technology: Myrinet

- Wormhole-routed cross-bar switch
  - 2.5 GB/s / link
  - Low latency
  - No buffering
  - Link level flow control
  - Head-of-line blocking

- NICs
  - Programmable RISC processor
  - Custom protocol
  - LUT-based destination assignment
Two-stage event building architecture

Event builder technology: Gigabit ethernet

- Standard 1 Gb/s Ethernet
- 8 switches (by Force-10)
  - 1 per slice
  - 4000 ports in total
- 3 rails per Readout Unit PC
- 1 or 2 rails per Builder/Filter PC according to performance
Two-stage event building architecture

Storage Managers
- 2 Storage Manager PCs per slice
- NexSan SataBeasts (RAID-6 disk array) connected through redundant Fibre Channel switches
- Max write speed 2.1 GB/s with simultaneous transfer to Tier-0 2.6 GB/s w/o transfer
- Local storage 300 TB (several days)
Storage Manager Performance

- Total capacity: 300 TB (several days of data talking)
- HLT compresses event data (root); reduction by factor ~2
- Event data to disk
  - pp; ~200 MB/s, design 600 MB/s
  - Heavy Ions: ~1.4 GB/s (up to 2.8 GB/s w/o transfer)
Super-fragment size in pp runs \( (n_{\text{vertex}}) \)

At 100 kHz can take 2.5 kB per FED or 20 kB per super-fragment.

- Some super-fragment builders at the limit with 2011 configuration
- Fixed by re-arrangement of super-fragment composition

Super-fragment size at 30 vertices / kB

| BPIX s1d06-30 (0.687 kB/vertex) | 22.3 |
| BPIX s1d06-29 (0.686 kB/vertex) | 22.3 |
| EB- s2d10-03 (0.382 kB/vertex) | 20.5 |
| EB+ s2d10-10 (0.373 kB/vertex) | 20.4 |
| EB+ s2d10-02 (0.377 kB/vertex) | 20.4 |
| EB- s2d10-11 (0.376 kB/vertex) | 20.3 |
| BPIX s1d06-29 (0.596 kB/vertex) | 19.3 |
| BPIX s1d06-30 (0.576 kB/vertex) | 18.9 |
| HF s2d10-07 (0.182 kB/vertex) | 16.6 |
DAQ throughput per input / pp and HI (generated events)

- DAQ optimized for large fragment sizes: reach 350 MB/s (limited by GBe)
- Max rate at 100 kB/FRL: 3.5 kHz
- Max aggregate EVB throughput: ~150 Gbyte/s (436 x 350 MB/s)

Fragment size distribution
Log-normal, Std-dev = mean
HLT states during 2011 Heavy Ion run

- In 2011, Tracker zero-suppression done in HLT farm