CMS Run 2 Event Building

Remigius K Mommsen
Fermilab
Overview

- Overview of CMS event builder for LHC run II
- Event-building protocol
- Measurements
CMS Event Builder

Detector front-end (custom electronics)

Front-End Readout Optical Link (FEROL)
- Optical 10 GbE TCP/IP

Data Concentrator switches
- Data to Surface
- Aggregate into 40 GbE links

Up to 108 Readout Units (RUs)
- Combine FEROL fragments into super-fragment

Event Builder switch
- Infiniband FDR 56 Gbps CLOS network

72 Builder Units (BUs)
- Event building
- Temporary recording to RAM disk

Filter Units (FUs)
- Run HLT selection using files from RAM disk
EvB Protocol

FEROLs

EVM

RU 1

RU 2

Event Request

Event

Super-fragment

Assign event to BU1

Super-fragment

Super-fragment

BU 1

BU 2
Achieving Performance

Avoid high rate of small messages
- Request multiple events at the same time
- Pack data of multiple events into one message

Avoid copying data
- Operate on pointers to data in receiving buffers
- Copy data directly into RDMA buffers of Infiniband NICs
- Stay in kernel space when writing data

Parallelize the work
- Use multiple threads for data transmission and event handling
- Write events concurrently into multiple files

Bind everything to CPU cores and memory (NUMA)
- Each thread bound to a core
- Memory structures allocated on pre-defined CPU
- Interrupts from NICs restricted to certain cores
- Tune Linux TCP stack for maximum performance
Computers

Readout Unit (RU)
- Dell PowerEdge R620
- Dual 8 core Xeon CPU E5-2670 @ 2.60GHz
- 32 GB of memory

Builder Unit (BU)
- Dell PowerEdge R720
- Dual 8 core Xeon CPU E5-2670 0 @ 2.60GHz
- 32+256GB of memory
  (240 GB for Ramdisk on CPU 1)
Data Network

40/56Gb NICs (Infiniband or Ethernet)
  - Mellanox Technologies
    MT27500 Family
    [ConnectX-3]
10/40 GbE switches
  - Mellanox
    SX1024 & SX1036
Infiniband switches
  - Mellanox SX6036

Infiniband CLOS network
Measurement Technique

Semi-automatic scanning using python scripts (w/o run control)

- Generate fake data on the FEROL in free-running mode
  - EVM gets always 1024 Bytes fragments & is not counted as RU
- Full event building, but events not written to disk unless noted
- 20 measurements every 5s done at each point after waiting >60s

Measurement schemes

- Data-concentrator measurement
  - Vary number of FEDs (TCP streams) sent to 1 RU
- Canonical setup uses 8 FEDs (TCP streams) per RU
  - All streams use the same fragment size
  - Scan different fragment sizes
  - Measure various sizes of the event-building system
- Real FED builder setup with different number of FEDs
  - 1-18 FEDs per RU
  - Set fragment sizes to roughly expected size for each FED
  - Scale fragment sizes linearly for different event sizes
Data Concentrator

4 – 24 Streams
(1 stream / FEROL)

1 kB

EVM

1 kB

RU

256B - 16 kB

1 – 256 kB

BU 1

BU 2

1 RU

2 BUs
Data Concentrator

N streams per RU - 1 RU → 2 BUs
- 4 streams
- 8 streams
- 16 streams
- 24 streams

Event Rate at EVM (kHz)

Fragment Size (bytes)

12 Mar 2016
Scalability of the Event Builder

- **1 kB**
- **256B - 16kB**
- **3 kB – 8 MB**
- **8 Streams per RU**
  (1 or 2 streams / FEROL)
- **1 – 101 RUs**
- **1 – 72 BUs**

- **EVM**
- **1 kB**
- **2 - 128 kB**
Scalability of the Event Builder

8 streams per RU
- 4 RUs ⇒ 4 BUs
- 22 RUs ⇒ 22 BUs
- 48 RUs ⇒ 48 BUs
- 72 RUs ⇒ 72 BUs
- 101 RUs ⇒ 72 BUs

CMS Preliminary

Throughput on RU (MB/s)

Event Rate at EVM (KHz)

Fragment Size (bytes)

14 Mar 2016
Effect of NUMA Settings

'`/usr/bin/numactl --physcpubind=10,12,14,26,28,30 --membind=1'` used to start executives

Threads and memory structures are bound to cores/memory using XDAQ policies
Custom IB Routing

- Optimize the routing to prefer traffic from RU to BU
- See next talk from Andre about IB routing
BU Scalability

1 kB

256B - 16kB

1 – 101 RUs

3 BUs

EVM

1 kB

2 - 128 kB

8 Streams per RU
(1 or 2 streams / FEROL)
BU Scaling vs Number of RUs

CMS Preliminary

N RUs + EVM to 3 BU
- 4 RUs
- 9 RUs
- 22 RUs
- 48 RUs
- 72 RUs
- 101 RUs

Throughput on BU (MB/s)

Event Size (kB)

16 Mar 2016
BU Performance

72 RUs + EVM to 3 BUs

- building only
- + CRC32c+CRC
- + write & delete

CMS Preliminary

17 Mar 2016
Production FED Builders

2-18 Streams per RU
(1 or 2 streams / FEROL)

60 or 82 RUs

72 BUs

1 kB

4B – 12kB

1 - 118kB

100kB – 5MB

1 kB

EVM

1 kB

4B – 12kB

1 - 118kB

100kB – 5MB
Standard pp FED Builder

pp FED builder
60 RUs ⇒ 72 BUs

- building only
- + checksums
- + write & delete

CMS Preliminary

Total EvB throughput (GB/s)

Event Rate at EVM (KHz)

Relative event size

Event Size (kB)

17 Mar 2016
Prod. FED Builder vs Canonical

Production FED builders
- 60 RUs $\Rightarrow$ 72 BUs
- 82 RUs $\Rightarrow$ 72 BUs
- 8 streams per RU

CMS Preliminary

17 Mar 2016

Remi Mommsen – CMS Run 2 Event Builder
Summary

CMS has a complete new event-building system for LHC run 2

- State-of-the-art technology
- Order of magnitude smaller than run-1 DAQ system

Optimal use of high-end hardware

- New event-building protocol
- New software to exploit hardware capabilities
- A lot of fine-tuning to get full performance

Infiniband performance sensible to traffic pattern

- Requires custom routing
- Performance diminishes the more uneven the system becomes
RU Scalability

- 8 FEROLs per RU
- 1 RU

- 1kB

- 256B - 16kB

- 1 – 256 kB

- 1 – 72 BUs
RU Scaling vs Number of BUs

1 RU + EVM to N BU

- 4 BUs
- 9 BUs
- 22 BUs
- 48 BUs
- 72 BUs

CMS Preliminary

Throughput on RU (MB/s)

Event Rate at EVM (kHz)

Fragment Size (bytes)

16 Mar 2016
Checksums

17 Mar 2016

- RUs and/or BUs can verify the CRC16 calculate by the FED on the payload
- BUs calculate a CRC32c on the complete event which is rechecked by the HLT
Number of Builder Threads

72 RUs + EVM to 3 BUs, write & delete

- 1 thread
- 2 threads
- 3 threads
- 4 threads
- 5 threads (default)
- 6 threads

Event Rate at BU (kHz)

Throughput on BU (MB/s)

Event Size (kB)

08 Feb 2016