The readout system upgrade for the LHCb experiment

on behalf of the LHCb collaboration

Paolo Durante
paolo.durante@cern.ch
Trigger from Run2 to Run3

“The LHCb Trigger in Run-II” (10 Jun 2016, 09:50)
Run3 upgrade

- Filter farm will need to handle:
  - Larger **event rate**: 30MHz (+ 10 MHz empty crossings)
  - Larger **event size**: ~130 kB (@ 30 MHz)

- New challenges for DAQ & High-Level Trigger

Data Network - Throughput

- Detector
- Hardware trigger
- Data acquisition
- High Level Trigger farm
- CERN long term storage
- Offline physics analysis
Run3 online system

- Dimensioning the system:
  - ~10000 versatile links
  - ~500 readout nodes
  - ~40 MHz event-building rate
  - ~130 kB event size

- High bisection bandwidth in event builder network
  - ~40 Tb/s aggregate bandwidth
  - Use industry leading 100 Gbit/s LAN technologies

- Global configuration and control via ECS subsystem

- Global synchronization via TFC subsystem
Slow & Fast Control Systems

SLOW CONTROL (ECS)

- Controls and monitors all subsystems:
  - DAQ, TFC, HLT, farm...
- Upgrade will continue to use same software stack as today...
  - JCOP / DIM / WinCCOA / SMI++ / Recipes
- ...but will also evolve to interface with new hardware
  - GBT-SCA

"Controlling DAQ Electronics using a SCADA Framework" (10 Jun 2016, 10:55)

FAST CONTROL (TFC)

- Distributes synchronous commands and reference clock
- Drives all detector frontends ("fast commands")
- Integration of PON technology (Passive Optical Network) for upgraded TFC

"Timing and Readout Control in the LHCb Upgraded Readout System" (7 Jun 2016, 15:00)
Long-distance optics

- Counting room on surface
  - Power, cooling, space constraints in underground area
  - ~350 meter distance

- Based on CERN technology
  - Rad-hard Versatile Link on frontends
  - Initially qualified for ~100m

- Loopback tests in 2015
  - ~12 months, ~700 meters
  - OM3 and OM4
  - Avago MiniPOD transceivers
  - Bit Error Rate < 10^{-18}
  - Full system equivalent (on 10000 links): < 5 errors/day
Readout board hardware (PCIe40)

- PCI Express add-in card
  - Altera Arria10 FPGA
  - High-density optical IO, up to 48 transceivers
  - 2 PCI-Express Gen3 interfaces (x8x8)

- At the heart of several subsystems
  - Data Acquisition (DAQ)
  - Experiment Control System (ECS)
  - Timing & Fast Commands (TFC)

- Decouple FPGA from network
  - Maximum flexibility in network technology

- Exploit commercial technologies
  - PCI Express Gen3 interconnect
  - COTS servers designed for GPU accelerators

- Also adopted by ALICE (called CRU)
PCIe bifurcation option

- Reduces complexity, power, cost
- Requires BIOS support

SWITCHED ROOT COMPLEX

I/O board

CPU

RC

Switch

x16 slot

x8 HIP

FPGA

Readout board

BIFURCATED ROOT COMPLEX

I/O board

CPU

RC

RC

x16 slot

x8 HIP

FPGA

Logic

Readout board
Readout board firmware
DMA architecture

Event Fragments

DMA sink (x2) -> Stream parser ->
DMA controller (meta data) -> MSI generator

DMA controller (main data) -> Descriptor scheduler

DMA controller (odin data) optional

TFC banks

PCle Hard IP

main/meta = ~100/1

250 MHz
40 MHz

Gen3 x8
DMA controller

Data stream ➔ On-chip memory ➔ Host memory

Input stream ➔ M

Virtual memory map

Streaming
Memory-mapped

DMA controller

M

DMA buffer

S

Virtual circular buffer

Driver (pcie40.ko)

mmap()

BAR
Commands
Status
Control

DMA data (write)

DMA data (read)

Input stream

Memory descriptors (read)
PCIe DMA performance

Continuous DMA performance histogram over 3 days

- Average 111.55 Gbps
- Stdev < 0.1 Gbps
Readout unit dataflow

- A single Readout unit must sustain ~400 Gbps I/O bandwidth
- Precompute fragment boundaries in FPGA (meta data)
- Optimize memory bandwidth
- Can be realized with mid-range modern server
Event-building performance

1 process, size

- “Large-scale DAQ tests for the LHCb upgrade” (7 Jun 2016, 11:40)
- “Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade” (7 Jun 2016, 15:00)
Conclusion: current status

- **Rad-hard optical links**: validated for long-distance operation
- **FPGA throughput**: compatible with 100G event-builder network
- **PCIe40 hardware**: initial production currently ongoing
- **Event-builder**: successfully tested on small clusters, full scale test imminent
- **Data-centre**: design being finalized, compact layout + fast interconnects
- Continuing close collaboration with industry partners to maximize performance of upcoming technologies (networking, but also **computation**)
  - e.g: “Particle identification on an FPGA accelerated compute platform for the LHCb Upgrade” (7 Jun 2016, 15:00)

For a full software trigger in LHCb RUN3, the online system is on track to deliver 40Tbit/s of frontend data to the filter farm, leveraging commercial technologies wherever possible.
Thank you
PCIe MPS parameter

MPS = 128 bytes  MPS = 256 bytes

On Linux: `pci=pcie_bus_perf` in kernel command line
FPGA occupancy of firmware

![Graph showing FPGA occupancy for different configurations.](image-url)
DMA + ECS performance

- Emulate register accesses by ECS to stress system with concurrent DMA
- Evaluate different reads/writes ratio
- Performance still consistently over 54 Gbps!

![DMA throughput (Gbps) bar chart]

**DAQ + ECS performance (~30e6 regs/sec)**

- DAQ only
- DAQ + 100%W
- DAQ + 25%R + 75%W
- DAQ + 50%R + 50%W
- DAQ + 75%R + 25%W
- DAQ + 100%R
Dimensioning the system

- Event-size (@ 2x10^33) ~ 130 kB
- Eventbuilding-rate 40 MHz (of which 30 MHz contain collisions and 10 MHz are empty)
- 500 event-builder nodes
- Between 1000 and 4000 event-filter nodes
  - Dual-socket, accelerator to be decided
- 500 port minimum event-building network
  - TDB: Intel OmniPath, InfiniBand, Ethernet
- 1500 – 4500 port filter network
  - Ethernet?
- New data-centre
  - 4000 rack-units max
  - 2 MW max
- 50 to 100 nodes for “slow” and “fast” control
  - Using PCIe40 cards
- Rest of control-system on virtual machines as today
- Local storage on each filter-unit at least 20 TB → will depend on disk-technology
- Central buffer storage ~ 1 to 2 PB
- ~ 10000 uni-directional fibres for DAQ (4.8 Gbit/s)
- ~2000 fibre-pairs for ECS/TFC (GBT)