The LHCb Online system for Run 3:
trigger-free readout
with (almost exclusively) off-the-shelf hardware

Tommaso Colombo
on behalf of the LHCb Online Group

ACAT, 21 Aug 2017, Seattle
The LHCb experiment

One of the 4 major experiments at CERN’s Large Hadron Collider

Single-arm forward spectrometer, built for precision measurements of B-meson decays
The LHCb Run 3 upgrade

Now

- **Motivation:**
  - Cope with higher luminosity
  - Increase trigger efficiency
    (see Rosen Matev’s talk)
  - No more hardware trigger
  - Full readout of the detector
    at the 30 MHz rate of inelastic collisions delivered by the LHC
  - All-new readout electronics
  - All-new event builder
  - Upgraded event-filter farm

---

2020

**LHCb Run 2 Trigger Diagram**

- **40 MHz bunch crossing rate**

**L0 Hardware Trigger:**

- 1 MHz readout, high $E_T/P_T$ signatures

- Partial event reconstruction, select displaced tracks/vertices and dimuons

- Buffer events to disk, perform online detector calibration and alignment

- Full offline-like event selection, mixture of inclusive and exclusive triggers

- **12.5 kHz (0.6 GB/s) to storage**

---

**LHCb Run 3 Trigger Diagram**

- **30 MHz inelastic event rate (full rate event building)**

**Software High Level Trigger**

- Full event reconstruction, inclusive and exclusive kinematic/geomeric selections

- Buffer events to disk, perform online detector calibration and alignment

- Add offline precision particle identification and track quality information to selections

- Output full event information for inclusive triggers, trigger candidates and related primary vertices for exclusive triggers

- **2-5 GB/s to storage**
LHCb Run 3 DAQ & event filter

- Detector front-end electronics
- Event builder network
- Event builder PCs
- Event filter farm (~ 4000 dual socket nodes)
- Online storage
  - 300 GB/s
  - ~ 100 PB
  - 2 – 4 k streams
- ~500
- ~10000 Point-to-point links
- ~500
- 100 Gbit/s
- 100 Gbit/s
- Clock & fast commands
- Sub-farm switch
- TFC
- DAQ cluster
- Filter cluster
- Underground
- Surface
Readout

- Front-end / DAQ interface:
  - GBT (link layer) + Versatile Link (physical layer)
  - Radiation-hard optical link interface
  - Up to 4.48 Gb/s per link

- DAQ readout: TELL40
  - PCIe card in event builder PC
  - Receives data from GBT links
  - Buffers the data in the main PC memory via DMA

- Even with no low-level trigger, still need timing & synchronous command generation + distribution: SODIN + SOL40
PCIe40: one card, many uses

• One card: PCIe Gen 3.0 x16 add-in card
  - Arria10 FPGA
    • Custom 100 Gb/s DMA engine
  - High-density optical I/O:
    • up to 48 bidirectional GBT ports
    • dedicated fast control port

• Three firmwares:
  - Readout (TELL40)
  - Timing & DAQ supervisor (SODIN)
  - Fast & slow control fan-out (SOL40)
  - Or the three combined: Mini-DAQ for development/testbed

• Only one type of custom hardware in the system
  - Easier maintenance, lower costs
  - Pre-series manufactured, series production to start this year
Event builder: hardware

- ~10000 input links
- Event size up to 150 kB
  → **36 Tbs total event building bandwidth** (40 Tbs with margin)
- Node: TELL40 + off-the-shelf hardware
  - 1 TELL40 (up to 48 inputs)
  - 1 “DAQ” 100 Gb/s NIC (event builder network)
  - 1 “FILTER” 100 Gb/s NIC (output network)
- Need ~500 nodes:
  - Assuming 80% network utilization
Event builder: architecture

- **3 software units:**
  - Readout unit (RU): read and buffer data from TELL40
  - Builder unit (BU): collect event fragments from RUs, send out built events
  - Event manager (EM): decide which BU builds an event

- **Network considerations:**
  - Traffic pattern is **all-to-all** gather:
    - For each event, one BU receives fragments from all RUs
    - Many events → All BUs receive fragments from all RUs
  - Need network with full bisection bandwidth: fat-tree topology
Event builder: scalability

• DAQPIPE: an event-builder benchmark
  - Supports different network technologies:
    • InfiniBand, OmniPath, Ethernet (WiP)
  - Implements RU, BU, EM

• Large parameter space to play with:
  - Communication scheduling
    (linear shift, random)
  - Communication size
  - Number of in-flight communications

• Goal: maximize network usage
  - Not an easy task on fat-trees and similar networks
  - Scheduling and routing are key
  - Collisions (two or more senders using the same network path at the same time) must be avoided

• Reassuring results so far
  - Tested on various 100 Gb/s fat-trees (HPC clusters)
  - Good scalability on InfiniBand with up to 64 nodes
  - Each node gets at least 70% of its maximum
  - Larger scale tests already in the works

![Graph showing per-node throughput (Gb/s) vs. nodes]

- Linear shift
- Random

Per-node throughput (Gb/s)

Nodes
**Event builder: communication scheduling**

- **Idea:**
  - use the local clock of EB nodes to precisely schedule communications
  - avoid network conflicts
- **Implementation:**
  - Standard linear-shift all-to-all:
    - $N$ servers $\rightarrow$ $N$ phases
    - In phase $i$, server $n$ sends data to server $m = (n + i) \mod N$
  - Standard fat-tree modulo routing
  - If all servers start each phase at the same time $\rightarrow$ no conflicts on the network links
- **Small scale test:**
  - 32 nodes with NTP-synchronized clocks
  - 1 Gb/s Ethernet fat-tree
- **Promising results:**
  - Nodes get 90% of max throughput with 200 ms phases
  - Should be tested at larger scale
Event filter: architecture

- Basic strategy remains the same as Run 2:
  - First filter: HLT1
    - Fast reconstruction and selection
    - Synchronous with DAQ at 30 MHz
    - Output: ~1 MHz
  - Disk buffer for HLT1-accepted events
  - Second filter: HLT2
    - Full reconstruction and selection
    - Asynchronous (events from disk)
    - Output: ~100 kHz
Event filter: buffer storage

- The disk buffer allows exploiting LHC downtime
  - Maximize event filter farm utilization
  - Need large buffer to absorb long LHC runs:
    ~100 PB for a week’s worth of data

- Currently investigating both centralized and distributed solutions

- Requirements:
  - Must sustain a total of: ~150 GB/s input + ~150 GB/s output
  - I/O pattern:
    1 sequential read stream + 1 sequential write stream per filter node
  - No need for a file-system: an object store is enough
  - Minimal redundancy: some data loss is acceptable
  - Non-uniform data access costs is acceptable:
    filter nodes should process “local” data first
  - A global name-space is desirable for ease of operation and monitoring

Online storage
- 300 GB/s
- ~ 100 PB
- 2–4 k streams
Slow control system

- **Experiment Control System**
  - Based on the same architecture and tools used successfully in Run 1 and 2
  - CERN JCOP framework
  - WinCC-OA SCADA system
  - DIM middleware
Conclusion and outlook

- The LHCb Online system upgrade for Run 3 is an ambitious plan:
  - 30 MHz read-out
  - 40 Tb/s event building and filtering
  - Up to 100 PB buffer storage
  - In 2020!

- The plan execution is proceeding well:
  - Read-out boards, firmware, and associated control software are already well advanced in development
  - The event builder benchmarks present no show-stoppers
  - Implementation evaluations are underway for:
    - Event builder nodes and network
    - Event filter nodes, storage, and network

Big challenges remain: interesting times ahead!