Big Data solutions for the online processing of trigger-less detectors data

M. Migliorini, J. Pazzini, M. Zanetti

University of Padova
Introduction and Scope

Trigger-less (a.k.a. 40 MHz readout key to unbiased analysis for New Physics searches

- Typically trigger-less RO is applied to trigger information
- Here an example of RO of detector raw data
- Not just stored, but processed/analyzed online
Online processing setup

Detector → Detector’s electronics → Accelerated server → Broker → Distributed processing
Online processing setup

Detector

CMS Drift Tubes
muon chambers

Detector’s electronics

HL-LHC prototype
electronics

Accelerated server

Xilinx KCU
PCIe FPGA

Distributed processing

INFN CloudVeneto
data-center
Online processing setup

Detector

Detector’s electronics

Accelerated server

Distributed processing

CMS Drift Tubes
muon chambers

HL-LHC prototype electronics

Xilinx KCU PCIe FPGA

INFN CloudVeneto data-center

~20 km
• Chamber FrontEnd ASICs connected to Virtex 7 evaluation boards (VC707).
  – 138 TDC channels implemented in each VC707 (128 connected to detector’s channel)
  – Trigger logic (“mean timer”) computed and added to the TDC data stream
• Funnelling of TDC hits from VC707 to Kintex Ultrascale evaluation board (KCU1500)
  – Transmission based on GBTx protocol
  – VC707 use the 120MHz reference clock for serializing data at 4.8Gbps
  – KCU1500 recover the clock from received data
  – Gearbox in KCU1500 firmware for aggregating the two optical GBTx lanes
  – Data transferred to RAM by DMA transfers via Gen3 PCIe
• The data stream of TDC hits is read out continuously and as soon as the data is available
• With each 1kB data transfer via xDMA driver $\Rightarrow$ 128 TDC hits (8Bytes each)
• Stream of a-priori unordered hits $\Rightarrow$ Not “Event-like” structure
Analytic framework

- Non-sequential analysis of the hit stream needed in order to match coherent signals and extract higher-level features
- DataFrame-based analytics

Apache Spark
- Cluster computing system for distributed analytics
- Widely used for fast large-scale data processing
- Including API for streaming and ML applications

Apache Spark Streaming
- Discretized stream processing
- Records processed in batches with short tasks
- Each batch is a RDD (partitioned dataset)
Stream brokerage

- Data stream brokerage solution to serve the Apache Spark cluster ➔ Efficiently feed Spark cluster by streaming directly to the executors’ cores

Apache Kafka
- Distributed real-time stream processor platform
- Allow log-retention and replication
Stream brokerage

- Data stream brokerage solution to serve the Apache Spark cluster ➔ Efficiently feed Spark cluster by streaming directly to the executors’ cores

Apache Kafka
- Distributed real-time stream processor platform
- Allow log-retention and replication
Data stream processing

- Hits are broadcasted to the Kafka cluster via a Kafka producer. Each DMA transfer (1kB ↔ 128 hits) maps into a “message” sent by the producer to the brokers
  - At this point data from Legnaro to Padova
- Kafka brokers provides the data to the Spark cluster
  - Num of brokers matches number of spark executors
Data stream processing

- Hits are broadcasted to the Kafka cluster via a Kafka producer. Each DMA transfer (1kB ↔ 128 hits) maps into a “message” sent by the producer to the brokers
  - At this point data from Legnaro to Padova
- Kafka brokers provides the data to the Spark cluster
  - Num of brokers matches number of spark executors
Spark process the stream in batches via Spark Streaming API, producing 2 outputs
- chambers occupancy (all hits)
- Slimmed-down DataFrame (an “event”) consisting of hits compatible with the trigger tag "Event Building"

Both outputs are injected back to Kafka and processed by a python consumer for analysis and bokeh for visualization.
Cluster monitoring

- Monitoring of Spark cluster is provided out-of-the-box via the Spark WebUI
- Kafka brokers are configured to expose their metrics on a given port
  - Collected in a Prometheus database
  - Metrics visualized in a Grafana Dashboard
Scaling tests

- Aim at verifying scaling properties in view of dimensioning the system for HL-LHC operations
- Test scaling of block-by-block and overall
Scaling tests (1)

- Generate test pulse signals at high rate to test throughput to DMA
  - Currently implementing 1-lane 1kB DMA transfers
- Large improvements expected
  - Doubling the DMA channels (one DMA-lane per V7)
  - Increasing the DMA transfer size

![Graph showing throughput vs. input throughput. The graph shows a linear relationship with a slope indicating a throughput of >220 MBs at high input throughput.]
Scaling tests (2)

• Using Kafka Java producers from a Cloud node (i.e. decoupled form detector)
• Topic with 80 partitions on 3 brokers, testing sending data from:
  – 1 single Producer (equivalent to 1 KCU in current setup)
  – Multiple Producers

• All tests involving Spark are very processing-dependent
  – Efficient implementation of the Spark processing to contain processing time
  – Appropriate batch size to minimize scheduling overheads
• and depend on the output
  – Storing to DB/disk or sending output back to Kafka
Summary and plans

Status

• Unfiltered raw data from a mock-up CMS muon spectrometer sector are processed online at 40 MHz

• Scaling tests performed to cope with expected throughput at HL-LHC
  – 1 good μ / BX in CMS ➔ 1 μ / 60 BXs / sector ➔ 50 hits / 1.5μs ➔ ~250MB/s / sector

Plans

• Upgrade system to exceed 10Gbs

• Deploy it on a CMS sector for LHC run 3

• Add a processing step into PCIe card
  – Synergy with L1 local trigger reconstruction
  – Propagate downstream only high level information (muon tubs)

• Try out “new physics” cases, e.g. mimic slow muons (HSCP)
BACKUP
Geometry and nomenclature

- 4 layers of cells per super-layer (SL)
  - 50-80 cells per layer
- 3 SL per chamber
- 4 chambers per sector
- 5 sectors per wedge
- 12 wedges in CMS
  - 250k channels
Setup at Legnaro INFN Lab (LNL)

- A mock-up of a sector with 4 reduced-size chambers (mini-DT)
  - 2 full scale MB3 chambers also available
  - 2 additional mini-DT ready in April
- Pre-amplified signals from detectors processed by phase2 front-end electronics
  - “OBDT” boards being commissioned (not for production yet), evaluation boards (Virtex7) used in the meanwhile
  - Substituting phase1 “minicrate”
- Operate with cosmics (O(100) Hz) or test pulses
  - Possibly firing all channels at 40 MHz
- Internal trigger generated and added as additional word to the data flow
  - Phase1 “mean-timer” algorithm
Mini-DT

• A rescaled and refurbished version of the CMS Drift Tubes chamber
  – Overall transverse dimensions: ~70x70 cm²
  – Rather light, can be carried “by hand”

• Same cell geometry and electrodes configuration

• Same “superlayer” (quadruplet) configuration
  – Layers are staggered to allow mean-timer application

• At the test-beam:
  – 4 superlayers, 2 per line, all measuring the bending coordinate (x)
  – Structure to support them allows flexibility in x and y and place them ~50 cm apart in z
Setup at Legnaro INFN Lab
• The 4 mini-DT installed on an experiment along H2 (North Area)
  – 5x10^6 positron on target per spill to study the \( \mu^-\mu^+ \) production at threshold
• Data recorded with the 40 MHz system
  – No online processing, data simply stored for offline analysis (TB in hours)
• Sector 11, wheel +2 is being equipped with phase2 front-end electronics (OBDT)
  – Take cosmic data from around summer

• Under discussion, but likely to leave OBDTs together with FE electronics for 3\textsuperscript{rd} and 4\textsuperscript{th} stations (spilt chambers’ signals).