FPGA implementation of a histogram-based parent bunch-crossing identification for the Drift Tubes chambers of the CMS experiment

Fabio Montecassiano, Nicola Pozzobon, Pierluigi Zotto
Università degli Studi di Padova, INFN sezione di Padova
on behalf of the CMS Collaboration
a new trigger primitive generator for the CMS Drift Tubes

- the **barrel region** of the CMS experiment is equipped with **Drift Tubes** chambers for muon tracking and trigger

  - the **primary signal** is a charge collection time, measured by a **Time-to-Digital Converter (TDC)**

  - a conversion to drift time allows to reconstruct **local track segments at trigger level** exploiting the staggered layout of DTs

    - 3 super-layers: 4+4 layers for the transverse projection, 4 layers for the longitudinal projection

- this is **currently done in the front-end electronics**, which also embeds a **trigger, synchronous with the LHC bunch timing**

- the **front-end is being replaced** in view of Phase II upgrades: **TDC signals** from active wires will be sent asynchronously to the back-end using the CERN GBT

- a new trigger will be needed in the back-end, which will process data received asynchronously from high-speed serial optical links
from an “on-detector” to an “off-detector” local trigger

Minicrate: legacy front-end
- TDC
- local trigger
- copper to fibre

read-out server

concentrator

muon trigger

DAQ

Minicrate 2: Phase II front-end
- TDC on FPGA + GBT

backend:
- local trigger
- muon trigger read-out pipeline

current

Phase II

- L1 trigger
- DAQ

see: The Phase-2 Upgrade of the CMS Muon Detectors Technical Design Report [CERN LHCC-2017-012] [cds.cern.ch/record/2283189]
a histogram-based DT trigger in brief

- we proposed a Hough Transform-based algorithm
  - advantages: a histogram-based statistical approach, robust against spurious hits and outliers, which removes the need of explicit left-right ambiguity solving...
  - three parameters must be identified: muon crossing time $t_0$, track slope $m$ and track intercept $x_0$
    - we exploit parameter reduction to work only with 1D histograms to solve for one parameter at a time, processing multiplets of hits in parallel
    - qualities are assigned by applying thresholds on histogram content
  - everything happens in the parameter space: the output is not a mere "I found a track", but "I found a track with these parameters" → in principle there is no need to refit

- the full algorithm is emulated in C in a HW-oriented style: algebra, fixed length integer types, “thermometer counters” ...

- the BX ID is already running on FPGA

finding the muon crossing time

- **mean-timing using the staggered layout** of DT chambers is one of the key features of the DT project since RD5

- starting from drift time in 3 wires in 3 different layers, the reduction of slope and intercept leads to a **linear combination of** such drift times $T_j$ that is **proportional to the maximum drift time** $T_{\text{max}}$

- the specific layout of CMS DT chambers makes all the coefficients to be integer

\[
2 T_A + 3 T_B + T_C = 2 T_{\text{max}}
\]

- but drift time can be written as $T_j = t_j$ (charge collection time) $- t_0$

- $t_0$ can be computed directly

\[
6 t_0 = 2 t_A + 3 t_B + t_C - 2 T_{\text{max}}
\]

- so far, mean timing has been used in calibration procedures by CMS

- combining 4 layers, 26 different patterns have physical meaning to compute $t_0$

- using only 3 layers, meaningful patterns are 48

\[
\text{exploiting symmetries, only 10 equations describe all possibilities}
\]

- the result needs correction as **the drift is not uniform along a DT cell**

\[
\text{we apply an average correction depending on an estimate of the crossing angle of the candidate muon based on the assumed same-side pair of TDCs for a given pattern}
\]

\[
\text{exploiting symmetries, only 4 angle-estimating equations}
\]

\[
\text{the correction is tabulated}
\]

see: EPJ Web of Conferences 127 (2016) 00012 [doi:10.1051/epjconf/201612700012]
- only few neighbouring wires should be processed to propose a candidate
  - we define a **macro-cell** of 18 channels

- large enough to **accept any possible track** in a DT chamber with a static geographical assignment of wires to macro-cells

- a full macro-cell is represented by 176 equations

- partial superposition 10:18 to ensure **redundancy**: at least one macro-cell must see all hits from the same muon

- \( t_i \) precision in computation: \( \text{BX}/8 = 3.125 \text{ ns} \)

- \( t_0 \) precision at output: 1 BX = 25 ns
what is expected from simulations

- acceptance, efficiency and resolution have been evaluated with simulations within the CMS framework
- the full parameter space has been systematically explored to test (almost) all possible cases

Efficiency of the BX ID as a function of the local track angle $\phi_{\text{local}}$, shown for low and high BX ID qualities, respectively 3/4 and 4/4, in each $\phi$ super-layer, and for two different quality thresholds after combining super-layer 1 and super-layer 3 with a logical “or”. Denominator: all muons associated via MC truth to at least one collected DT signal collected. Numerator: all candidate muons associated by the algorithm to the correct BX.

Candidate muon time distribution for local track angle $\phi_{\text{local}}$ smaller than $15^\circ$, for two different quality thresholds after combining super-layer 1 and super-layer 3 with a logical “or”. Both histograms are normalised to the area of the most inclusive threshold of a candidate muon compatible with a minimum of 3 hits out of 4 super-layers. Duplicate candidate muons within each $\phi$ super-layer are purged.
designing the algorithm with Xilinx Vivado High Level Synthesis

- the algorithm was implemented from a C simulation using HLS tools
  - HLS enables a fast and affordable exploration of design architecture even for big projects with few designers
  - the algorithm was originally drafted in C++, both standalone and embedded in the CMS framework
  - we moved to plain C as we found it more HLS-friendly (C++ was hiding the true algorithm complexity)
- the full firmware includes HDL modules where needed
  - no Multipliers-And-Accumulators (MAC) used: this FPGA resource is limited in amount and should be saved for the implementation of the Compact Hough Transform
  - multiple use of same information to avoid re-computing
  - direct calculations instead of nested loops
  - minimised idle time while waiting for data processing
  - includes a collection module from GBT streaming
I/O modules (not shown) emulated for bench tests

hit collector: HLS-based, 1 clock latency, 1.5 K FF, 0.8 K LUT [see next page]

hit driver [provisional block]: HLS-based, 26 clocks latency, 16 K FF, 15 K LUT
  - can be run in parallel to the SL processor

synch module: RTL-VHDL

super-layer processor: HLS-based, 90 clocks latency, 10 K FF/macro-cell, 16 K LUT/macro-cell
the hit collector module

- collects TDC hits over a **chosen time window**, large enough to include all hits generated by the same muon
- retains also the TDC hits from the previous time window, to **handle muons with hits split over consecutive windows**
  - with the clock running at the target frequency of 320 MHz, the net latency would be 284 ns, and hit collection over 128 clock ticks, i.e. 400 ns, would safely span the maximum drift time ~380 ns
- it also filters the streaming retaining only a maximum number of hits per layer, currently the most recent 3 hits, others are lost, in view of high occupancy at HL-LHC
first implementation on Xilinx Virtex Ultrascale

- the **first prototype** was built on a XCVU440 FPGA speed grade -3: Xilinx SSI technology, 3 SLR
  - resource budget: 2533 K LUTs, 5065 M FFs, 2880 DSP slices
- I/O is emulated: input is read from a pre-filled BRAM and a fictitious streaming is generated within the FPGA, output sent to PC with a serial line
  - timing has been closed at 200 MHz
  - the net latency of 91 clocks with hit driver module running in parallel is 460 ns
  - the static instantiation of 21 macro-cells (half super-layer) results in 380 K LUT and 278 K FF required
    - 17% of FPGA LUTs

I/O emulator
hit collector
super-layer processor
(21 macro-cells)
bench tests on Xilinx Virtex Ultrascale

- TDC hits produced from CMS simulation
- **C emulator featuring a scheduling scheme** for the hit collector module analogous to the firmware
  - input to the hit collector is ensured to be always the same in these tests, so the **inclusive output is compared 1:1**
  
  **(A)** a small sample of well isolated muons shot right in front of a DT chamber [only SL1 shown]
  - **99.8 %** agreement

  **(B)** a large sample of $O(10^6)$ muons shot from the interaction region over the whole CMS barrel, with an occupancy equivalent to an overall muon rate of 5 MHz, including possible spatial and temporal superposition of muons, obtained by a Poisson-arrival time distribution, the algorithm is processing roughly half of a single DT chamber [only SL1 shown]
  - **98.9 %** agreement

- **minor differences due to different rounding and to the C emulation of the scheduling scheme and hit collector filter**
looking at real muons with a continuous DAQ setup

- INFN Legnaro laboratories host a DAQ setup for tests of the CMS 40 MHz “triggerless” scouting project

- 4 reduced area DT super-layers, called mini-chambers, built with the same technology of CMS DT chambers

- two Xilinx Virtex VC707 boards, used to implement prototypes of the new TDC Phase II DT boards (OBDT), provide unfiltered TDC signal streaming on fibre using the GBT protocol, collecting data from 64+64 channels each

- a Xilinx Kintex Ultrascale KCU1500 board hosting the back-end: a FW implementing two GBT de-serializer modules and a merger module, which also stores data on PC HDD via PCIe using Direct Memory Access

- cosmic muon rate is well within the DAQ throughput

---

porting the algorithm on Xilinx Kintex Ultrascale

- **XCKU115** FPGA speed grade -2: Xilinx SSI technology, 2 SLR
  - resource budget: 663 K LUTs, 1326 M FFs, 5520 DSP slices

- the first deployment of the algorithm on the XCKU115 showed **no bit mismatch against the previous implementation**

- the algorithm input module was then modified to **read the TDC stream from GBTs for trigger processing**

- the trigger processing is done with the very same FPGA implementing the back-end

- both **the original TDC stream and the trigger output are merged** and stored via PCIe, generating a **single dataset**

- timing was closed at 160 MHz by choice, the net latency is ~569 ns
  - enough to cope with cosmic muon rates, allowing fast iterations to test several flavours of the firmware

- the footprint of each macro-cell is unchanged, only 7 macro-cells are statically implemented to process a full mini-chamber (64 channels)
  - 22% of FPGA LUTs
cosmic muon test on Xilinx Kintex Ultrascale

- performance is evaluated exploiting data from all mini-chambers

- cleaning of duplicate triggers is done at software level
  - cleaning is necessary to be independent of synchronisation of the scheduling time windows in the hit collector module

- emulated triggers and actual hardware triggers are cleaned with the same procedure
cosmic muon test on Xilinx Kintex Ultrascale

- a candidate muon is tagged by the coincidence of two emulated triggers in mini-chambers (0 or 1) and 3
- only high quality triggers in both mini-chambers are retained
- in simulations, \(96.95 \pm 0.02\%)\ of such coincidences are in-time with the parent muon

- the hardware trigger processes only data from mini-chamber 2
  - mini-chamber 0 has 8/64 dead channels and cannot provide an efficient event tag if the test were carried on mini-chamber 1
- both emulated and hardware-produced triggers are evaluated
performance evaluation with cosmic muons

- flawless non-stop data taking: 25 h 45 m
- trigger payload written 8,705,737 times
- continuously monitored with chip scope, no errors from both GBT and algorithm

after removing duplicates:

(A) $t_0$ from hardware and emulator agree $98.348 \pm 0.006$ % of times

(B) trigger quality is the same $99.444 \pm 0.003$ % of times

(C) the selected trigger is from the same macro-cell $99.778 \pm 0.002$ % of times
performance evaluation with cosmic muons

- *after removing duplicates:*

  - **(A) trigger time residual distributions** w.r.t. the event tag fairly compares to each other
    - the relative population of in-time triggers is \( 58.70 \pm 0.07 \% \) of total (\( 85.40 \pm 0.05 \% \) of high-quality)

  - **(B) the drift time distributions** obtained from hardware and emulated triggered events are superposing
    - the shape of both trigger time residual and drift time distributions is affected by the asynchronous nature of cosmic muons w.r.t. LHC bunch scheme that guided the design
outlook

- the trigger BX ID, collecting muon data for the first time, perform consistently with expectations
  - successfully interfaced to a CMS-like Phase II front-end
  - robust against the scheduling synchronisation
- planned improvements are already in design phase
  - hit collector module: dynamical geographical assignment of input data to macro-cells will reduce area and latency
    - contextual downscaling of macro-cells to 10 wires
  - hit driver module: already removed in the latest emulator thanks to improved TDC hit encoding
  - sub-BX output precision can be obtained with finer grain histogram thanks to freed resources
- the porting to FPGA of the CHT-based track identification has already started
  - with the R&D advantage of sharing the same underlying architecture already developed for the BX ID with MMT
additional material
... about the DAQ setup

dead channels in mini-chamber 2 affect macro-cells 0 and 1

rate per channel [Hz]

BX count

rate per channel [Hz]

BX count

TDC count [BX/30]

TDC count in orbit [BX/30]
... about the trigger

known feature in addressing macro-cell 6

cosmic muon run, INFN LNL, Aug. 2019

rate per macrocell [Hz]

trigger macrocell

0 1 2 3 4 5 6 7

0 5 10 15 20 25 30 35

rate per macrocell [Hz]

after duplicate removal

trigger macrocell

0 1 2 3 4 5 6 7

0 5 10 15 20 25 30 35

rate per macrocell [Hz]

after duplicate removal

trigger quality

0 1 2 3 4 5 6 7

0 500 1000 1500 2000 2500 3000 3500 4000

rate per macrocell [Hz]

after duplicate removal

trigger macrocell

0 1 2 3 4

0 5 10 15 20 25 30 35

rate per macrocell [Hz]

after duplicate removal

trigger quality

0 1 2 3 4

0 500 1000 1500 2000 2500 3000 3500 4000

rate per macrocell [Hz]

known feature in addressing macro-cell 6

mini-chamber
3 emulated
2 emulated
1 emulated
0 emulated
2 HW

rate per macrocell [Hz]

after duplicate removal

mini-chamber
3 emulated
2 emulated
1 emulated
0 emulated
2 HW

rate per macrocell [Hz]

after duplicate removal
... about the trigger

TWEPP 2019 — N. Pozzobon — FPGA histogram-based BX ID for CMS DT chambers
trigger output monitor