Heterogeneous online reconstruction at CMS

Andrea Bocci\textsuperscript{1}, Vincenzo Innocente\textsuperscript{1}, Matti Kortelainen\textsuperscript{2}, Felice Pantaleo\textsuperscript{1}, Marco Rovere\textsuperscript{1}

\textsuperscript{1} CERN, \textsuperscript{2} FNAL

for the \textit{Patatrack incubator} and the CMS Collaboration
the CMS High Level Trigger
the Compact Muon Solenoid

Muon detectors
- Inner Tracker
  - silicon pixel and silicon strip
- Electromagnetic calorimeter
  - lead tungstate crystals
- Hadronic calorimeter
  - brass and scintillators

4 Tesla superconducting magnet
the CMS Trigger & DAQ

L1 Trigger
100 kHz

raw data fragments
100 GB/s

Level 1 Trigger
- hardware based
- synchronous with LHC

Data Acquisition
- ADC converters
- event builder network

High Level Trigger farm
- software based
- multithreaded jobs

Storage Manager
- distributed filesystem
- transfer to Tier 0

event builder
20 TB RAM

on-demand reconstruction & event selection
> 30'000 CPU cores

5 GB/s to Tier-0

storage manager transfer system
the CMS Trigger & DAQ

the CMS High Level Trigger

- constraints
  - 300 ms average time to take a decision
  - 1 kHz average output rate (rejection 100:1)

- software event reconstruction and selection
  - runs on commercial servers
  - quasi-real-time, self-monitoring

- CMSSW: a modular C++ reconstruction software
  - over 4000 “modules”, written by hundreds of physicist
  - configured via a dedicated python library
  - exploit multi-threading to run multiple modules and reconstruct multiple events concurrently
Heterogeneous online reconstruction at CMS

- **particle flow reconstruction**: regional inside jets and around leptons
- **jet and MET**: reconstruction based on calorimetric and particle flow objects
- **ECAL**: local reconstruction and calibrations
- **HCAL**: local reconstruction and calibrations
- **muons**: detector local and regional global reconstruction
- **pixel tracking**: global and regional reconstruction
- **2018 data**
  - 50 average pileup
  - 2018 L1T and HLT
- **full tracking**: regional, partial reconstruction

264.5 ms
offloading to GPUs

**HCAL**: local reconstruction and calibrations

**ECAL**: local reconstruction and calibrations

see Monday’s talk in Track 9
“High Performance Computing for High Luminosity LHC”

pixel tracking: global reconstruction
details on the next slides

today we can offload ~24% of the online reconstruction!
the Patatrack demonstrator
• the overall approach
  • reconstruct pixel-based tracks and vertices on the GPU
  • leverage existing support for threads and on-demand reconstruction
  • minimise data transfer

• the full workflow
  • copy the raw data to the GPU
  • run multiple kernels to perform the various steps
    • decode the raw data
    • cluster the pixel hits
    • form hit doublets
    • form hit ntuplets (triplets or quadruplets) with a Cellular automaton algorithm
    • clean up duplicates
  • take advantage of the GPU computing power to improve the physics
    • fit the track parameters (Riemann fit, broken line fit) and apply quality cuts
    • reconstruct vertices
  • copy only the final results back to the host (optimised SoA format)
    • convert to legacy format if requested

see Monday’s talk in Track 5
“Bringing heterogeneity to the CMS software framework”
improved physics performance

single muon simulated dataset
- no pileup
- design detector conditions
- flat $p_T$ distribution from 0.5 GeV to 100 GeV
improved physics performance

\[ \text{Patatrack} \quad \text{CMS Open Data 2018} \quad 13 \text{ TeV} \]

- Tracking efficiency vs. \( p_T \) (GeV)
- Tracking fake rate from PV vs. \( p_T \) (GeV)
- Tracking duplicate rate vs. \( p_T \) (GeV)

\( \bar{t}t \) event tracks (\( \langle \text{PU} \rangle = 50 \)) for \( |\eta| < 2.5 \)

- \( \text{Patatrack} \) – triplets and quadruplets
- \( \text{Patatrack} \) – quadruplets
- CMS – 2018 online reconstruction

\( \bar{t}t \) simulated dataset:
- average of 50 pileup interactions
- \textit{design} detector conditions
improved physics performance

$t\bar{t}$ event tracks (⟨PU⟩=50)
$p_T > 0.9 \text{ GeV}$

- **Patatrack** – triplets and quadruplets
- **Patatrack** – quadruplets
- **CMS** – 2018 online reconstruction

$t\bar{t}$ simulated dataset
- average of 50 pileup interactions
- *design* detector conditions
GPU vs CPU validation

**Patatrack** CMS Open Data 2018 13 TeV

- **Tracking efficiency**
  - **t̅t̅ event tracks (⟨PU⟩=50)**
  - |η| < 2.5

- **Simulated track** $p_T$ (GeV)
  - $10^{-1}$ to $10^{2}$

- **Tracking efficiency**
  - **tt̅ event tracks (⟨PU⟩=50)**
  - $p_T > 0.9$ GeV

- **Tracking fake rate from PV**
  - $0$ to $0.2$

**Patatrack** – triplets and quadruplets on **CPU**

**Patatrack** – triplets and quadruplets on **GPU**

---

**Simulation dataset**
- average of 50 pileup interactions
- *design* detector conditions

---

November 7th, 2019
A. Bocci - Heterogeneous online reconstruction at CMS
**pixel tracks and vertices global reco**

**CPU**
- dual socket Xeon Gold 6130
- 2 × 16 cores (2 x 32 threads)
- throughput measured on a full node
- 4 jobs with 16 threads

---

**improved event throughput**

![Graph showing event throughput](#)
**pixel tracks and vertices global reco**

**CPU**
- dual socket Xeon Gold 6130
- 2 × 16 cores (2 x 32 threads)
- throughput measured on a full node
- 4 jobs with 16 threads

**GPU**
- single NVIDIA Tesla T4
- 2560 CUDA cores
- single job with 10-16 concurrent events
Improved event throughput

Pixel tracks and vertices global reco

CPU
- dual socket Xeon Gold 6130
- 2 x 16 cores (2 x 32 threads)
- throughput measured on a full node
- 4 jobs with 16 threads

GPU
- single NVIDIA Tesla T4
- 2560 CUDA cores
- single job with 10-16 concurrent events

See Monday’s talk in Track 5
“Heterogeneous reconstruction: combining an ARM processor with a GPU”
for a comparison of
- different NVIDIA GPUs
- ARM vs Intel CPUs
improved event throughput

pixel tracks and vertices global reco

CPU
- dual socket Xeon Gold 6130
- 2 × 16 cores (2 × 32 threads)
- throughput measured on a full node
- 4 jobs with 16 threads

GPU
- single NVIDIA Tesla T4
- 2560 CUDA cores
- single job with 10-16 concurrent events

transfer from GPU to CPU
- on demand
- small impact on event throughput

conversion to legacy data formats
- on demand, to be minimised
- small impact on event throughput
- high cost in CPU usage
improved event throughput

pixel tracks and vertices global reco

CPU
- dual socket Xeon Gold 6130
- 2 × 16 cores (2 x 32 threads)
- throughput measured on a full node
- 4 jobs with 16 threads

GPU
- single NVIDIA Tesla T4
- 2560 CUDA cores
- single job with 10-16 concurrent events

transfer from GPU to CPU
- on demand
- small impact on event throughput

conversion to legacy data formats
- on demand, to be minimised
- small impact on event throughput
- high cost in CPU usage

throughput (ev/s)

0 100 200 300 400 500 600 700 800 900 1000

legacy (on CPU) quadruplets (on GPU) triplets (on GPU) quadruplets (on CPU) triplets (on CPU)
conclusions
conclusions

- heterogeneous reconstruction in CMS is getting production ready
  - possible deployment in the HLT farm in Run 3
  - farm size reduced to ~80% of 2018
  - equip all nodes with a Tesla T4 GPU

- integrated in the experiment’s software
  - CUDA-based framework
  - pixel local reconstruction, track and vertices
  - calorimeters’ local reconstruction
  - more algorithms and applications to come …

- a new programming model
  - GPU/CPU code sharing and reuse
  - investigating “performance portability” frameworks (e.g. Alpaka, Kokkos, SYCL, …)