Available on CMS information server

CMS DP -2023/036

The Compact Muon Solenoid Experiment

CMS Performance Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

16 June 2023

End-to-end Deep Learning Inference in CMS software framework

CMS Collaboration

Abstract

Deep learning techniques have been proven to provide excellent performance for a variety of high energy physics applications, such as particle identification, event reconstruction and trigger operations. Using low-level detector information in end-to-end deep learning approach allows to probe the poorly explored regions for dark matter search. This note presents an implementation of the end-to-end deep learning inference framework in CMS Software framework (CMSSW) for various physics objects classifiers such as electron/photon, quark/gluon, top and tau. The inference is benchmarked on CPU and GPUs.

End-to-end Deep Learning Inference in CMS software framework

CMS Collaboration

May 4, 2023

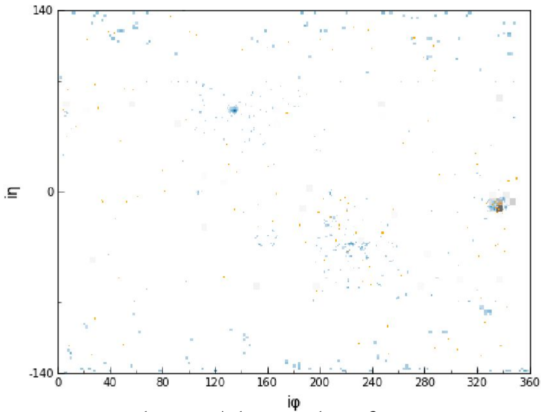
cms-conveners-ml-production@cern.ch

Abstract

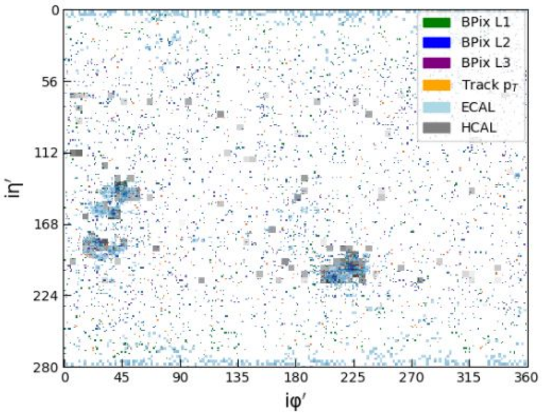
Deep learning techniques have been proven to provide excellent performance for a variety of high energy physics applications, such as particle identification, event reconstruction and trigger operations. Using low-level detector information in end-to-end deep learning approach allows to probe the poorly explored regions for dark matter search. This note presents an implementation of the end-to-end deep learning inference framework in CMS Software framework (CMSSW) for various physics objects classifiers such as electron/photon, quark/gluon, top and tau. The inference is benchmarked on CPU and GPUs.

Introduction : End-to-end deep learning

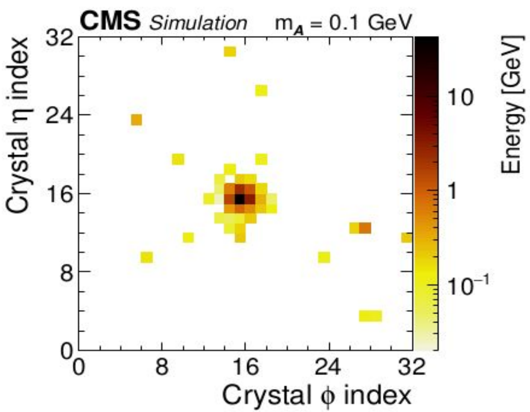
- Particle flow (PF) algorithm converts detector level information to physically intuitive objects however it comes with some information loss due to reduction in size and complexity.
- End-to-end deep learning algorithms can be trained from raw data before any particle processing performed.
- An end-to-end deep learning approach has been developed for
  - → Single particle reconstruction: electron, photon
  - → Jet Classification : quark, gluon, boosted top, tau
  - → Event reconstruction/classification: H→ AA→ 4γ



Electron/photon classification obtained from CMS open data simulations for pp collisions at √S = 8 TeV [1]



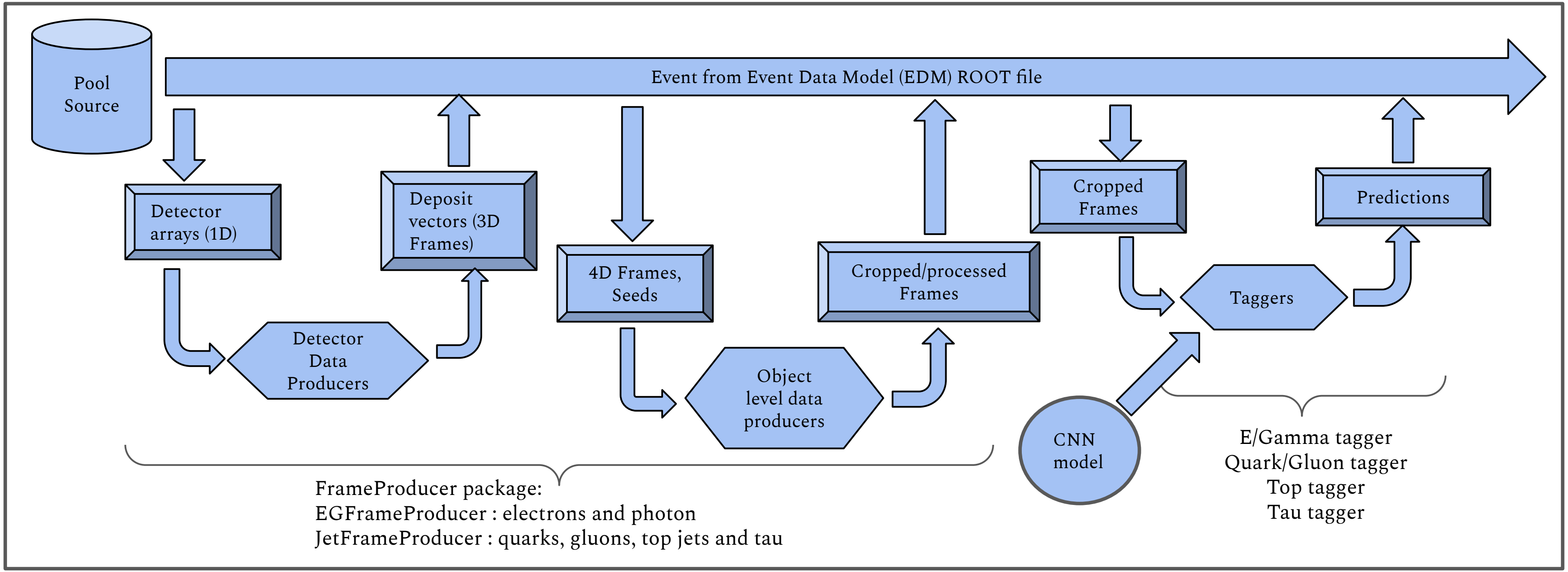
Boosted top/QCD jet classification obtained from CMS open data simulations for pp collisions at √S = 8 TeV [2]



Reconstruction of merged photons from H→ AA → 4γ process using end-to-end deep learning for pp collisions at √S = 13 TeV [3,4]

E2E inference within CMS software framework

The E2E inference framework is developed around the Event Data Model (EDM) in C++ based CMS software framework (CMSSW), it consists of three packages, namely, DataFormats, FrameProducer and Taggers.



1. Reading detector input → Storing the extracted vectors or graphs to EDM ROOT files
2. Extracting seed coordinates → Preparing the frames for inference
3. Running the inference on Convolutional Neural Network (CNN) model → Storing the predictions

Specifications of CNN models

- SimpleNet CNN pytorch model converted to ONNX [5].
- The inference of untrained CNN model is obtained using the ONNX C++ API present in the CMSSW framework with GPU support.

Tagger	No. of channels	Input tensor array size	Channels
E/Gamma	1	1×32×32	ECAL
Quark/Gluon	5	5×128×128	Track p_T, d_0, d_z, ECAL & HCAL
Top	8	8×128×128	Track p_T, d_0, d_z, BPIX layers, ECAL & HCAL
Tau	8	8×128×128	Track p_T, d_0, d_z, BPIX layers, ECAL & HCAL

ECAL: electromagnetic calorimeter, **Track p_T**: transverse momentum of the track,
d0 (**dz**): distance of minimum approach between the track and the primary vertex in transverse (longitudinal) plane.
HCAL: Hadronic calorimeter, **BPIX layers**: Barrel pixel layers.

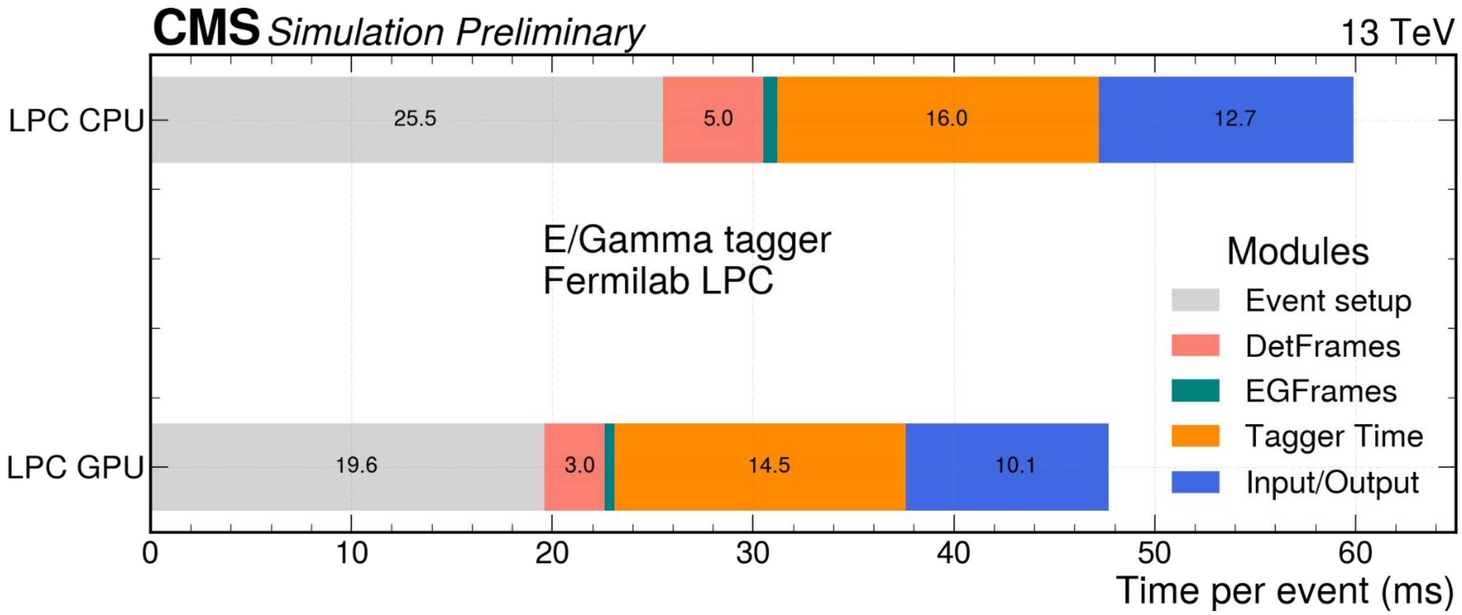
Specifications of CPU, GPU

Processor	GPU type	CPU @ GPU node	HBM
Fermilab LPC GPU	Tesla P100	Intel Xeon Silver 4110 16-cores	12 GB
NERSC Perlmutter GPU	Nvidia A100	AMD EPYC, 64-cores	40 GB
	CPU in analysis node		
Fermilab LPC CPU	AMD EPYC Processor, 8 CPUs, each with 1 core		

Details of inference studies

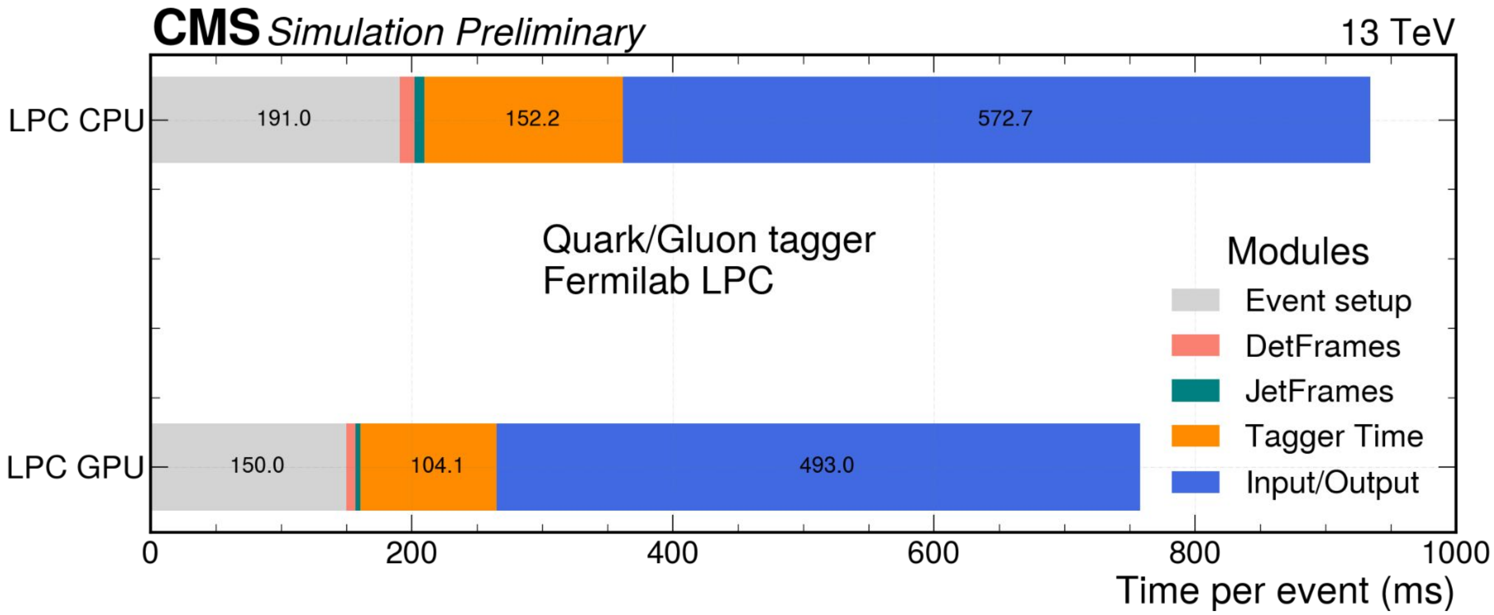
- CPU/GPU reserved for the benchmark studies.
- Inference obtained for **1000 events** with a **single thread.**
- Latency and throughputs are obtained through **FastTimeServices and Throughput services in CMS software framework.**
- A warm up run was performed.
- First 300 events are dropped from the calculation to stabilize the results.
- Measurement repeated 10 times.
- Used average of 10 measurements to benchmark latency and throughput.
- **Uncertainty of 0.5-3%** on measurements.

End-to-end E/Gamma tagger inference time breakdown per event



Time spent by end-to-end inference framework modules such as, Event setup (gray), DetFrames (Pink), EGFrames (Teal), Tagger time (orange), and input/output in blue for **E/Gamma tagger** per event in milliseconds. Timings are obtained from photon gun sample with transverse momentum 50 GeV reconstructed for Run2 2018 ultra-legacy conditions without pileup for proton-proton collisions at $\sqrt{S}$ = 13 TeV and compared for Fermilab LPC CPU and GPU. Total time per event is reduced by 20% with GPU usage compared to CPU.

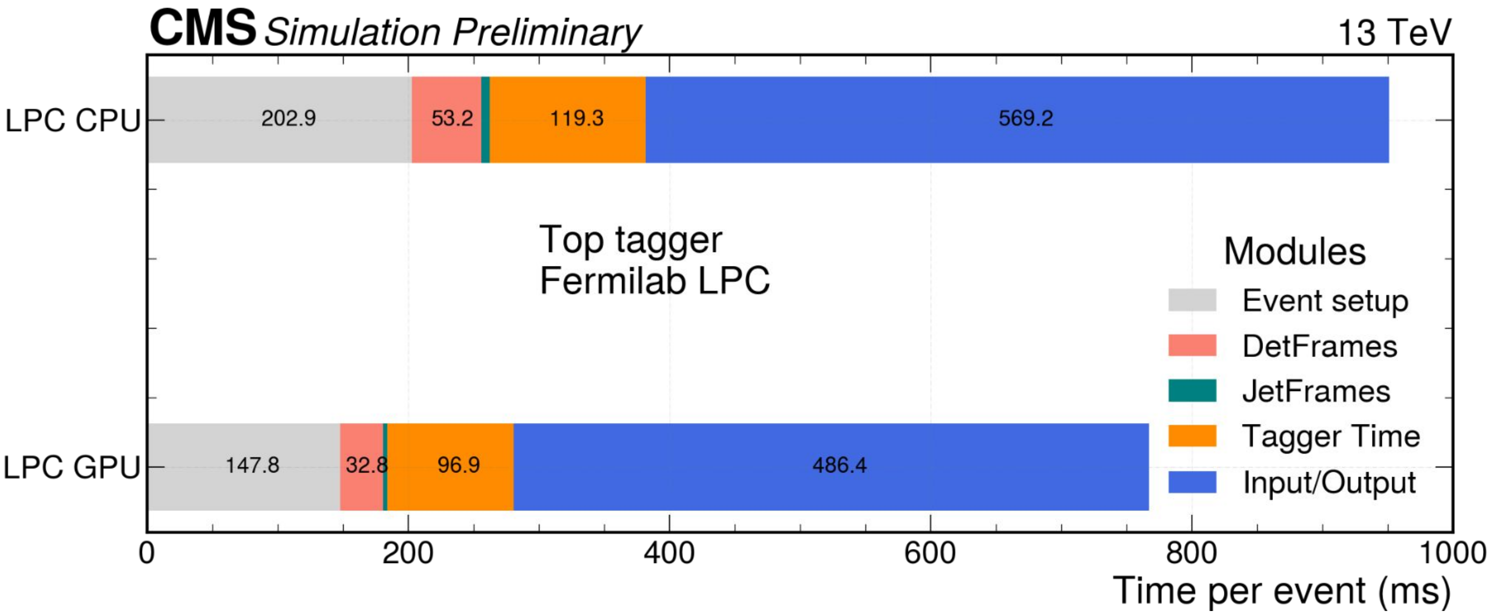
End-to-end Quark/Gluon tagger inference time breakdown per event



Time spent by end-to-end inference framework modules such as, Event setup (gray), DetFrames (Pink), JetFrames (Teal), Tagger time (orange), and input/output in blue for **Quark/Gluon tagger** per event in milliseconds. Timings are obtained from the quantum chromodynamic (QCD) multijets events with $\hat{p}_T$ between 300-470 GeV, reconstructed for Run2 2018 ultra-legacy conditions without pileup for proton-proton collisions at $\sqrt{S}$ = 13 TeV and compared for Fermilab LPC CPU and GPU. Total time per event is reduced by 19% with GPU usage compared to CPU.

* Input/output time can be speedup by 5 times for future studies.

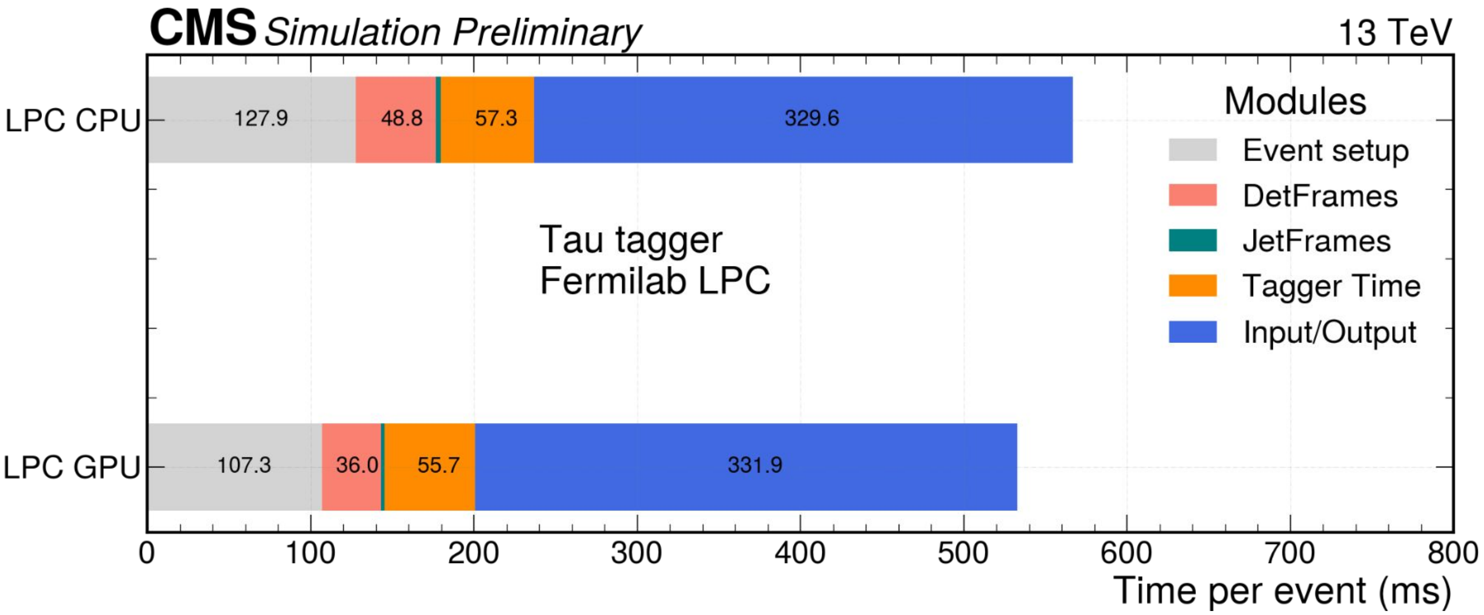
End-to-end Top tagger inference time breakdown per event



Time spent by end-to-end inference framework modules such as, Event setup (gray), DetFrames (Pink), JetFrames (Teal), Tagger time (orange), and input/output in blue for **Top tagger** per event in milliseconds. Timings are obtained from the top-antitop pair production events, reconstructed for Run2 2018 ultra-legacy conditions without pileup for proton-proton collisions at √S = 13 TeV and compared for Fermilab LPC CPU and GPU. Total time per event is reduced by 19% with GPU usage compared to CPU. Total time per event is reduced by 19% with GPU usage compared to CPU.

* Input/output time can be speedup by 5 times for future studies.

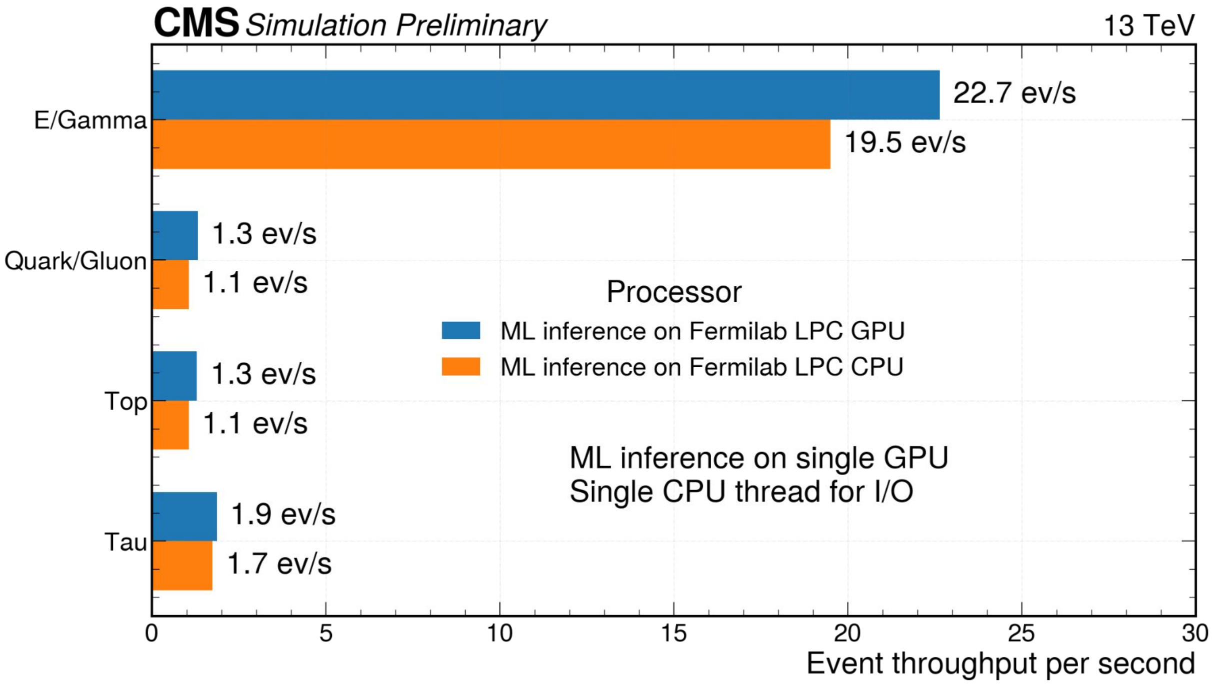
End-to-end Tau tagger inference time breakdown per event



Time spent by end-to-end inference framework modules such as, Event setup (gray), DetFrames (Pink), JetFrames (Teal), Tagger time (orange), and input/output in blue for **Tau tagger** per event in milliseconds. Timings are obtained from the higgs decaying to tau anti-tau events, reconstructed for Run2 2018 ultra-legacy conditions without pileup for proton-proton collisions at √S = 13 TeV and compared for Fermilab LPC CPU and GPU. Total time per event is reduced by 6% with GPU usage compared to CPU.

* Input/output time can be speedup by 5 times for future studies.

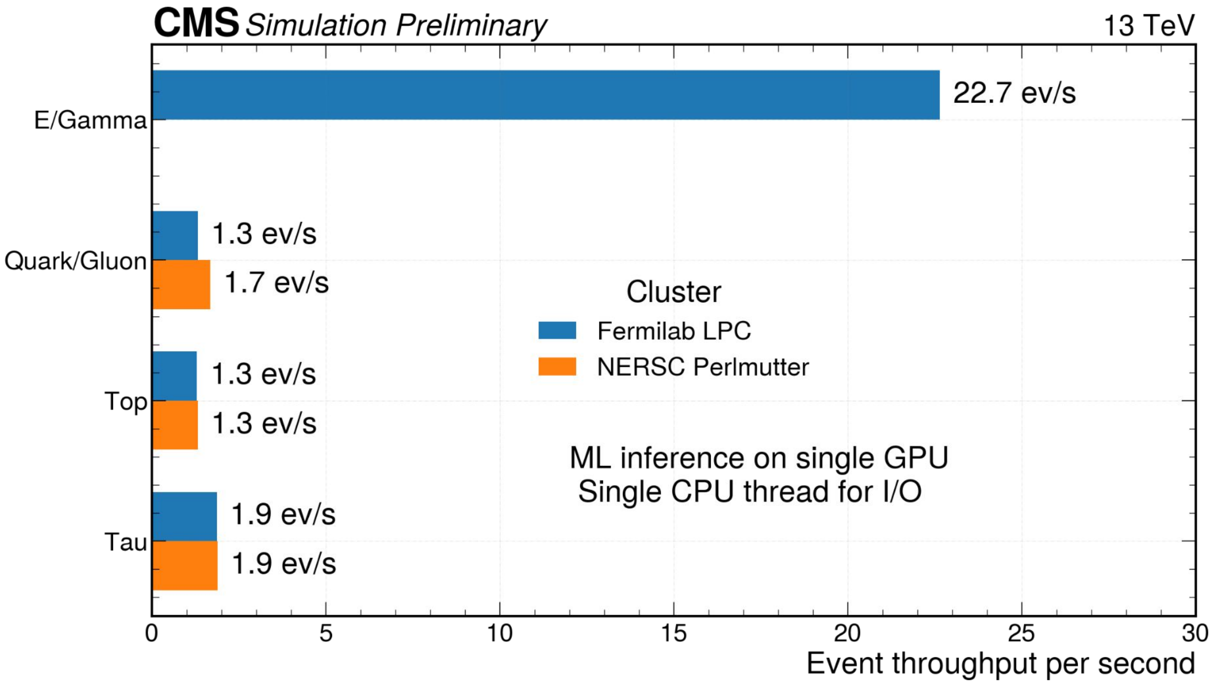
Benchmark E2E inference throughputs on LPC GPU and CPU



- E/Gamma frame is cropped in 32x32 matrix around Egamma seed, therefore larger throughput.
- JetFrame is cropped in 125x125 matrix around jet seed.
- **Uncertainty of 0.5, 2, 3, and 3%** on E/Gamma, Quark/Gluon, Top and Tau throughput measurements, respectively estimated by taking the average of 10 measurements.

End-to-end inference framework event throughput per second for E/Gamma, Quark/Gluon, Top, and Tau taggers compared for Fermilab LPC GPU and CPU. The ML inference is obtained on single GPU and single CPU with single thread for input/output. 16%,18%,18%,11% increased throughput with GPU compared to CPU for E/Gamma, Quark/Gluon, Top and Tau tagger, respectively.

Benchmark E2E inference throughputs on LPC and Perlmutter GPUs



- E/Gamma frame is cropped in 32x32 matrix around Egamma seed, therefore larger throughput.
- JetFrame is cropped in 125x125 matrix around jet seed.
- **Uncertainty of 0.5, 2, 3, and 3%** on E/Gamma, Quark/Gluon, Top and Tau throughput measurements, respectively estimated by taking the average of 10 measurements.

End-to-end inference framework event throughput per second for E/Gamma, Quark/Gluon, Top, and Tau taggers compared for Fermilab LPC GPU and NERSC Perlmutter GPU. The ML inference is obtained on a single GPU and single CPU thread is used for input/output.

References

1. M. Andrews, M. Paulini, S. Gleyzer, and B. Poczos,"End-to-End Physics Event Classification with CMS Open Data: Applying Image-Based Deep Learning to Detector Data for the Direct Classification of Collision Events at the LHC", 2018. Comput.Softw.Big Sci. 4 (2020) .
2. M. Andrews, B. Burkle, Y. Chen, D. DiCroce, S. Gleyzer, U. Heintz, M. Narain, M. Paulini, N. Pervan, Y. Shafi, W. Sun, E. Usai, and K. Yang, "End-to-end jet classification of boosted top quarks with the CMS open data" Phys. Rev. D 105, 052008.
3. CMS Collaboration, "Reconstruction of decays to merged photons using end-to-end deep learning with domain continuation in the CMS detector", arXiv: 2204.12313.
4. CMS Collaboration, "Search for exotic Higgs boson decays $H \rightarrow AA \rightarrow 4\gamma$ with events containing two merged diphotons in proton-proton collisions at $\sqrt{s}$ = 13 TeV", arXiv: 2209.06197.
5. J. Bai, F. Lu, K. Zhang, et al., Onnx: Open neural network exchange, https://github.com/onnx/onnx (2019).