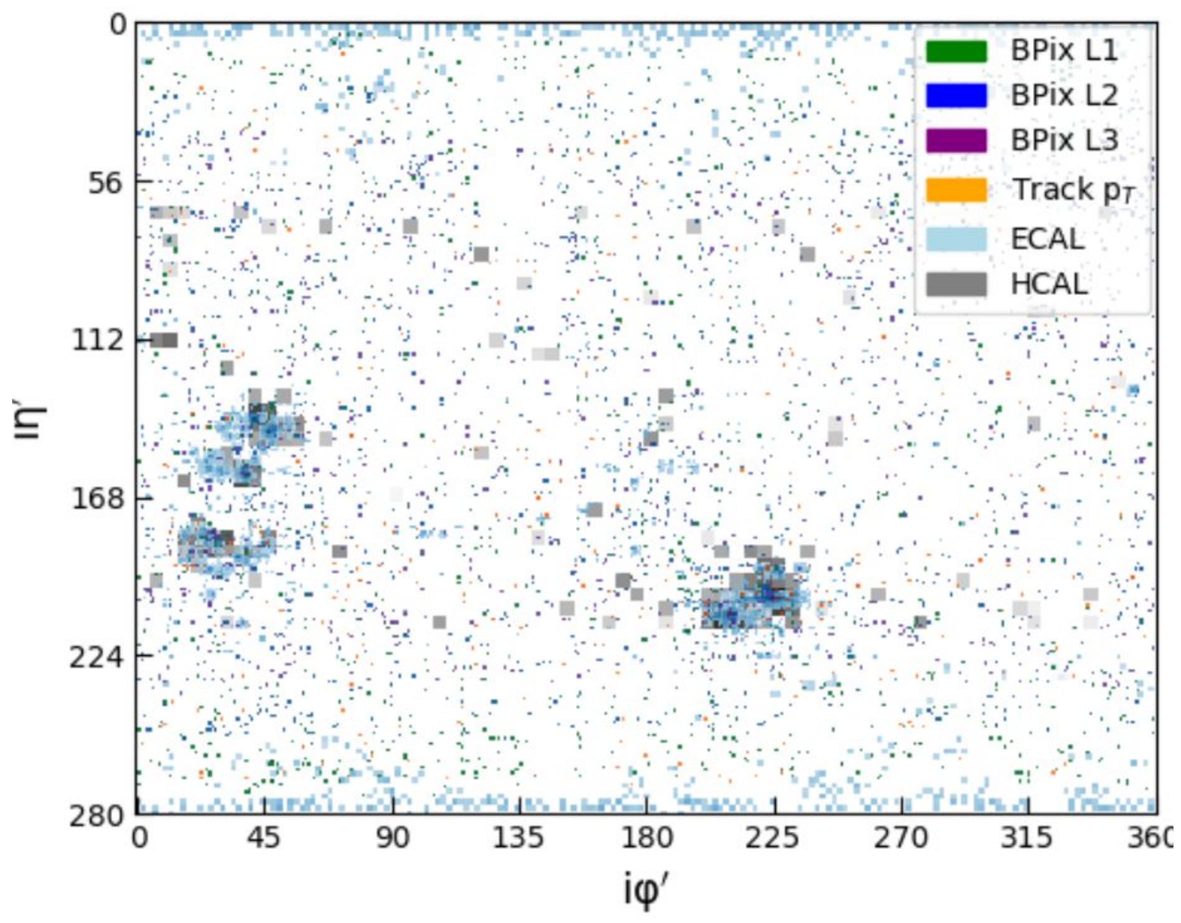
End-to-end deep learning inference with CMSSW via ONNX using docker

THE UNIVERSITY OF ALABAMA

Purva Chaudhari, Shravan Chaudhari, ***Ruchi Chudasama****, Sergei Gleyzer
on behalf of the CMS collaboration

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 (E2E) deep learning algorithms can be trained on raw data before any particle processing [1,2,3,4]



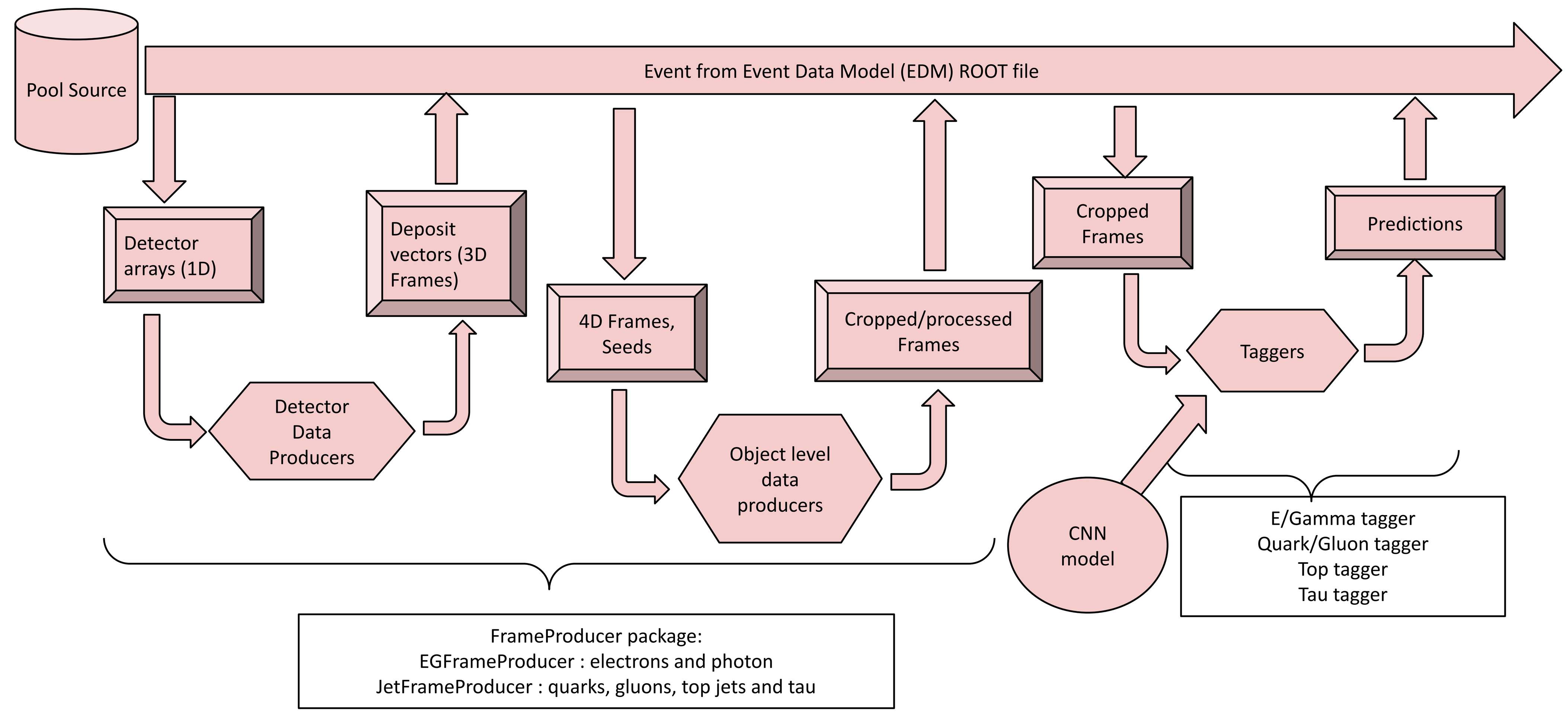
End-to-end image of a simulated top quark event for the full CMS detector. Image displays ECAL, HCAL,track p_T, and three pixel layers. [2]

E2E inference pipeline in CMS software framework

The E2E inference framework based on image-based approach, developed around the Event Data Model (EDM), it consists of three packages, namely, DataFormats, FrameProducer and Taggers.

1. Read the raw detector inputs and store the extracted vectors/graphs to EDM ROOT files.
2. The seed coordinates extracted and data frames prepared for inference.
3. Inference run using Convolutional Neural Network (CNN) model & predictions stored.

→ Four types of taggers developed: i) Electron/photon ii) Quark/Gluon iii) Top iv) Tau



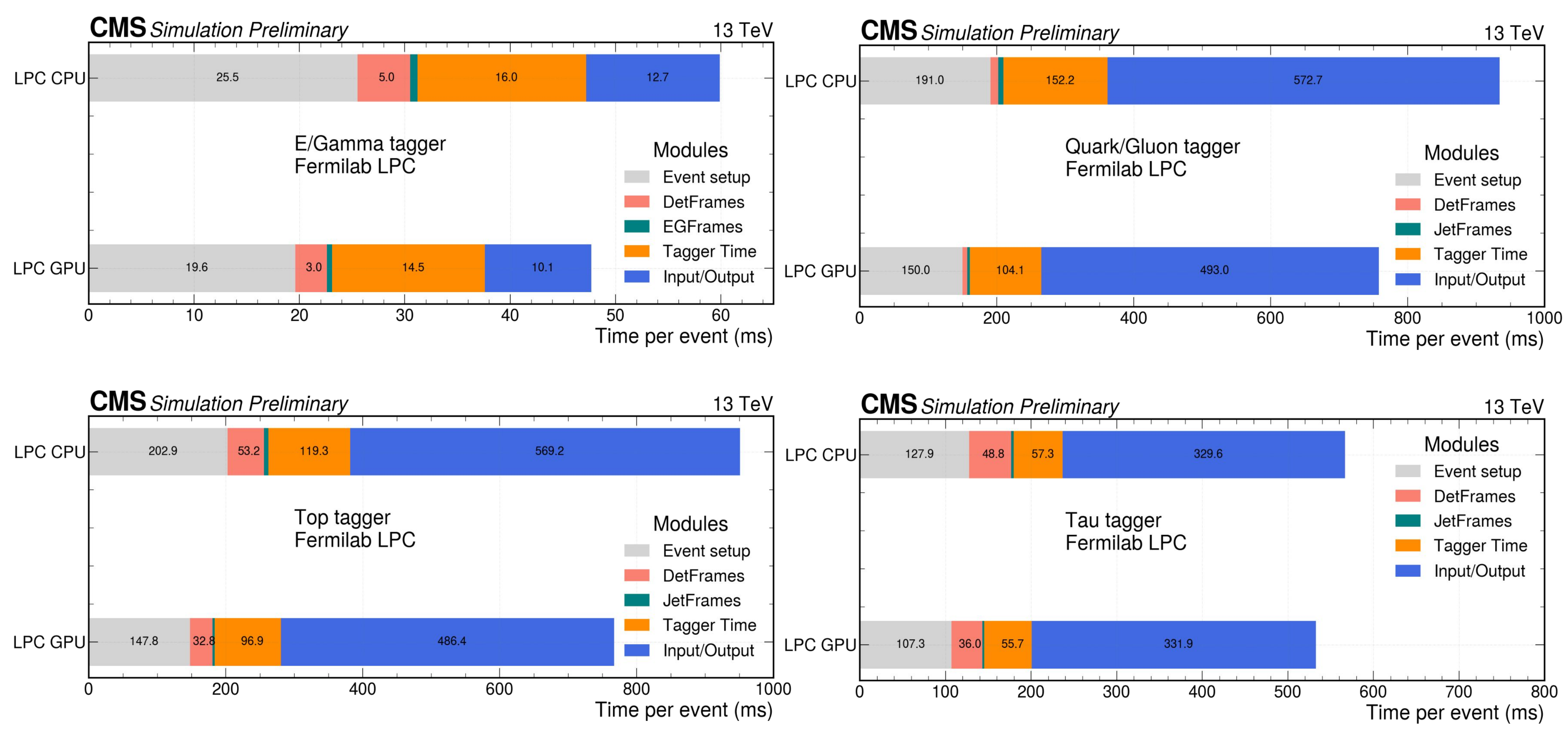
Specifications of the CNN models

- SimpleNet CNN pytorch model converted to ONNX.
- 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.

E2E tagger inference time breakdown for Fermilab LPC CPU & GPU

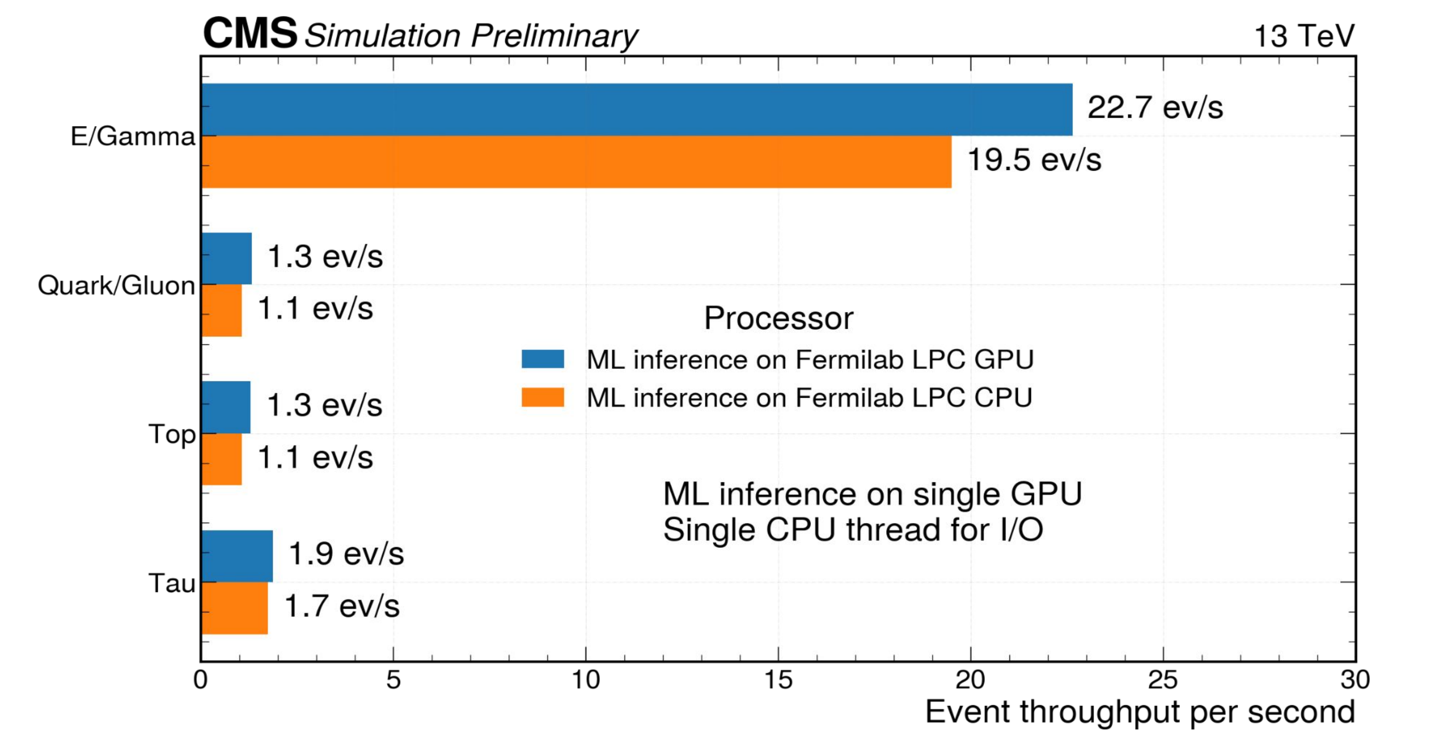


Time spent by end-to-end inference framework modules such as, Event setup (gray), DetFrames (Pink), EGFrames /JetFrames(Teal), Tagger time (orange), and input/output in blue for **E/Gamma (top left), Quark/Gluon (top right), Top (bottom left) and Tau tagger (bottom right)** per event in milliseconds. Timings are compared for Fermilab LPC CPU and GPU. Input/output timings can be speed up by more than 5 times for future studies..

Specifications of the GPU and CPU

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 at analysis node		
Fermilab LPC CPU	AMD EPYC Processor, 8 CPUs, each with 1 core		

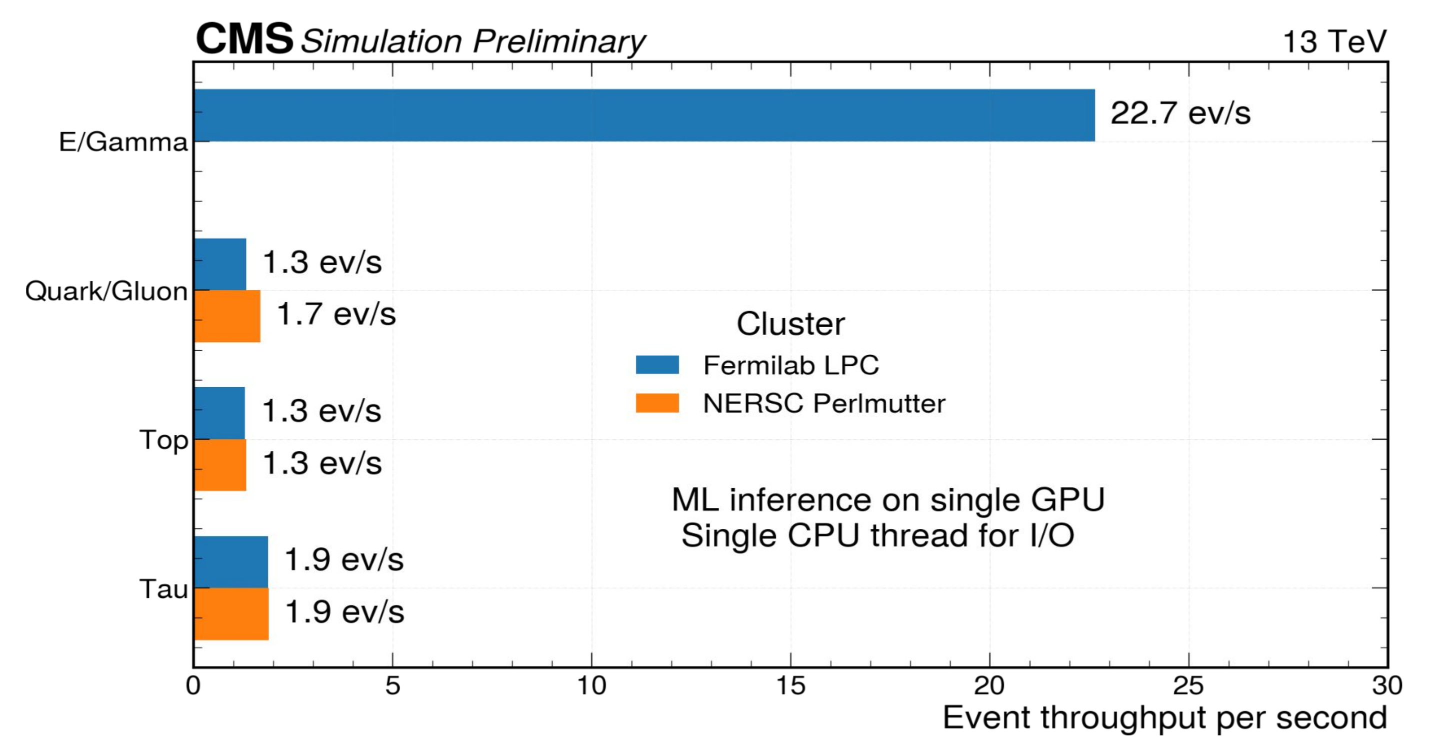
E2E throughput comparison for LPC GPU and CPU



More than ~15% speedup obtained with GPU compared to CPU.

E2E throughput comparison for LPC & Perlmutter GPU

The inference at NERSC Perlmutter cluster obtained by setting the CMS software framework using docker image



References

1. Sergei Gleyzer, et. al, End-to-End Physics Event Classification with CMS Open Data *Comput.Softw.Big Sci.* 4 (2020)
2. Sergei Gleyzer, et.al, End-to-end jet classification of boosted top quarks with the CMS open data, *Phys. Rev. D 105, 052008*,
3. The CMS Collaboration, Reconstruction of decays to merged photons using end-to-end deep learning in the CMS detector, *arXiv:2204.12313*
4. The CMS Collaboration, End-to-end Deep Learning Inference in CMS software framework, CMS-DP-23/XXX

* ruchi.chudasama@cern.ch

26th International conference on computing in High Energy and Nuclear Physics