



THE UNIVERSITY OF  
ALABAMA



BROWN

Carnegie  
Mellon  
University

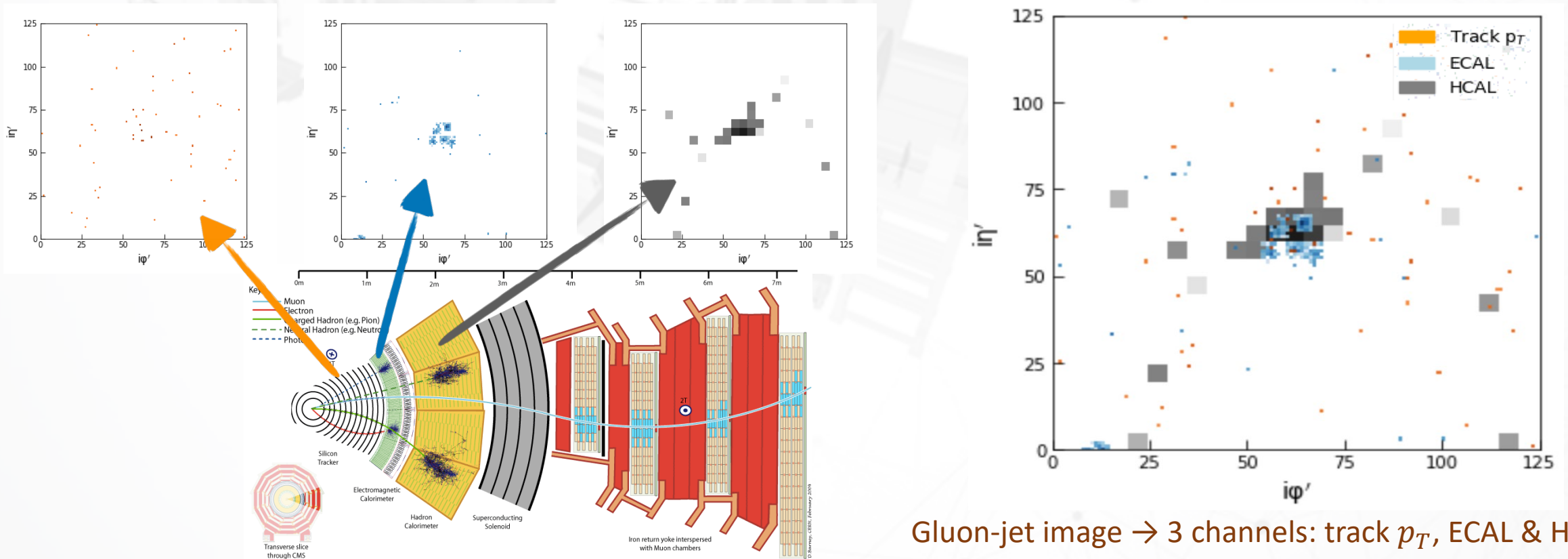


# ACCELERATING E2E DEEP LEARNING FOR PARTICLE RECONSTRUCTION USING CMS OPEN DATA

M. Andrews<sup>1</sup>, B. Burkle<sup>2</sup>, **D. Di Croce**<sup>3</sup>, S. Gleyzer<sup>3</sup>, U. Heintz<sup>2</sup>, M. Narain<sup>2</sup>, M. Paulini<sup>1</sup>, N. Pervan<sup>2</sup>,  
S. Chaudhari<sup>4</sup> and E. Usai<sup>2</sup>

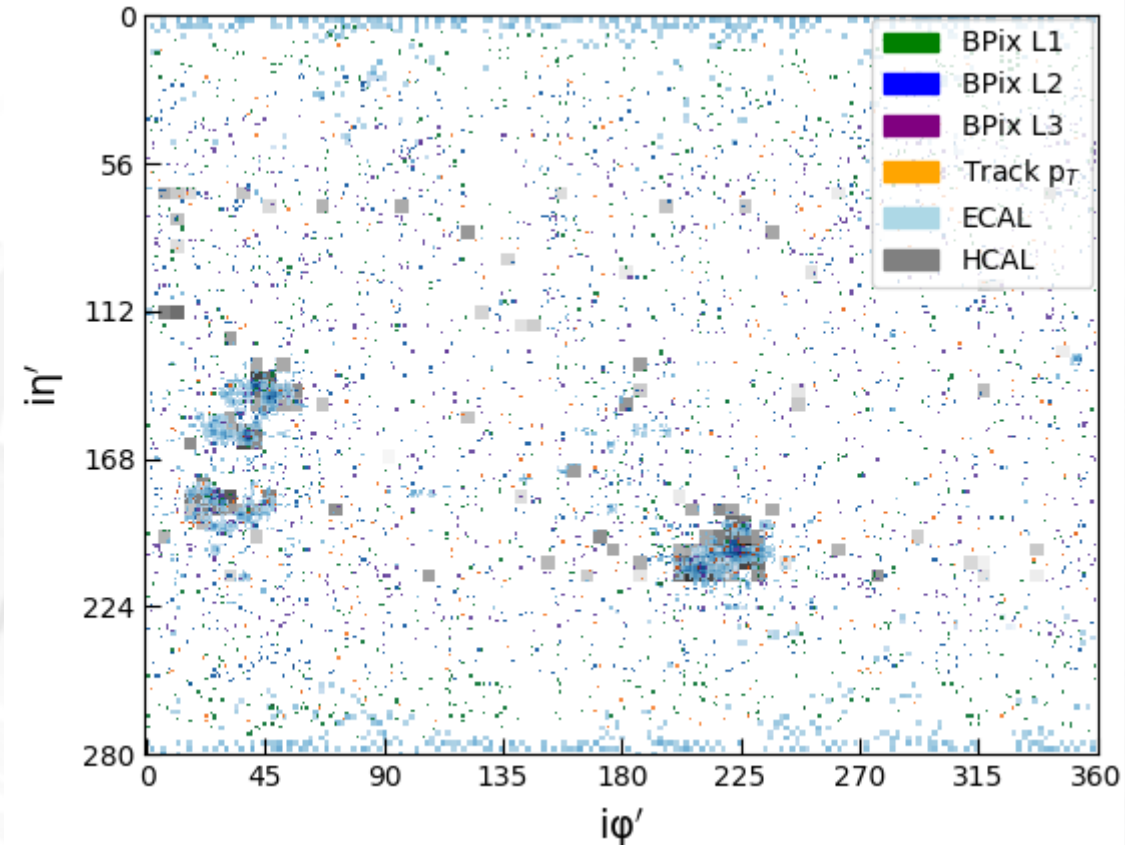
<sup>1</sup>Carnegie Mellon University, <sup>2</sup>Brown University, <sup>3</sup>University of Alabama and <sup>4</sup>BITS Pilani

- Particle Flow algorithms convert raw detector data into physically-motivated quantities until arriving at particle-level data. This method is dependent on the full understanding of particle decay phenomenology.
- Machine Learning algorithms can be trained directly from raw data and learning the pertinent features **unassisted**: the End-to-End Deep Learning approach.
- Low-level data from subdetectors are projected onto image layers.



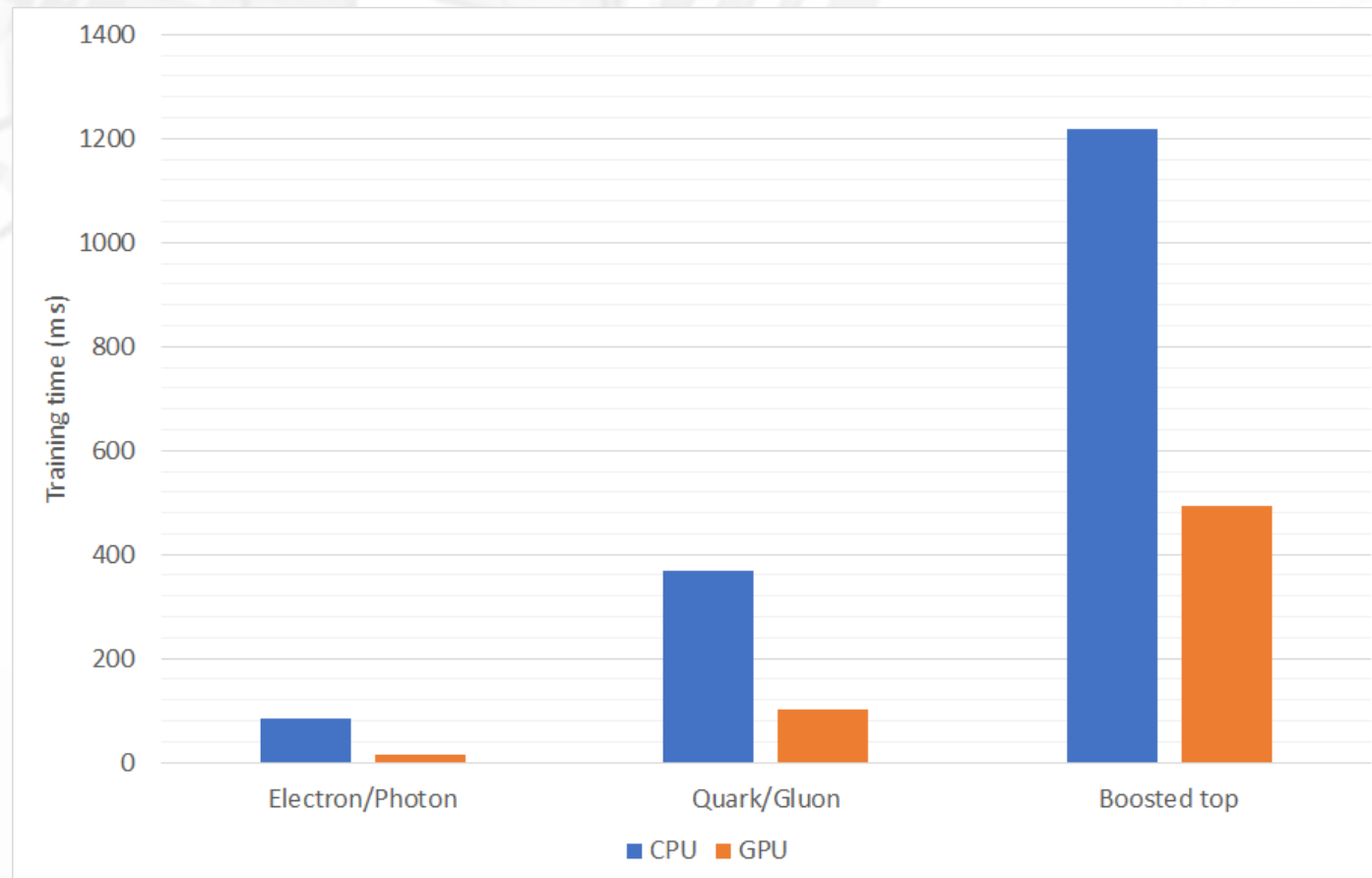
Gluon-jet image  $\rightarrow$  3 channels: track  $p_T$ , ECAL & HCAL

- End-to-End Deep Learning applications:
    - Single particle reconstruction: electron, photon
    - Jet classification: quark, gluon, boosted top, tau
    - Event classification
    - Simulation
  - We demonstrate the E2E implementation and perform studies on different E2E benchmarks
    - In this presentation, we will focus on the computing aspects of three different E2E benchmark:
      - Electron/photon classifier ([arXiv:1807.11916](https://arxiv.org/abs/1807.11916))
      - Quark/gluon classifier ([arXiv:1902.08276](https://arxiv.org/abs/1902.08276))
      - Boosted top classifier ([arXiv:2104.14659](https://arxiv.org/abs/2104.14659))
- See more details on [E2E boosted top classification talk](#)



$t\bar{t}$  event image - 6 channels:  
track  $p_T$ , 3 BPIX layers, ECAL & HCAL

- E2E electron/photon benchmark: 1 channel (ECAL)
- E2E quark/gluon benchmark: 3 channels (track  $p_T$ , ECAL & HCAL)
- E2E top quark jet benchmark: 8 channels (track  $p_T$ ,  $d_0$  and  $d_z$ , BPIX layers, ECAL & HCAL)



- We use the CNN E2E top quark benchmark to compare the training performance on single and multiple GPUs and TPU.

Cluster	Processor	CPU	Storage	HBM memory	Performance
Fermilab LPC	Tesla P100	Intel Xeon Silver 4110 8-core	HGST 1W10002 HD	16 GB	9.3 Single-Precision TeraFLOPS
NVIDIA Raplab	Tesla V100	4 Intel Xeon Gold 5118 12-core	SSD	32 GB	125 Mixed-Precision TeraFLOPS
Google Cloud	TPUv3-8	TPU	Google Cloud	128 GB	520 Mixed-Precision TeraFLOPS

- We used Fermilab LPC, Google Cloud and NVIDIA Raplab clusters to evaluate the performance of the ML models on different hardware architectures (CPUs, GPUs and TPU)

Comparison of I/O and training time for different computing architectures				
Config.	Processor	Tesla P100	Tesla V100	TPUv3-8
Config.	Batch size	32	64	64
	Bathes per epoch	80 k	40 k	40 k
x1 res	I/O time (1 batch)	0.119 s	0.018 s	0.018 s
	Train time (1 epoch)	321 min	19 min	14 min
x3 res	I/O time (1 batch)	0.833 s	0.063 s	0.189 s
	Train time (1 epoch)	1663 min	105 min	131 min

- Tesla V100 takes advantage of SSD storage which provides higher I/O speed
- Tesla P100: fewer CPU nodes when fetching batches and sending to GPU (data load bottleneck)
- TPUv3-8 spends 2.6 ms on forward and back propagation calculations, that is 4 x faster than the V100
- Tesla V100 and TPUv3-8 provide stronger data loading and training performance compared to Tesla P100

- Multi-GPU training on the standalone E2E boosted top jets benchmark using Horovod framework (<https://github.com/horovod/horovod>)
- Training performed on 2 Tesla V100 GPUs

Comparison of training time for different batch size configurations (2 GPUs)			
Batch size	<b>64*2</b>	<b>512*2</b>	<b>1024*2</b>
Train Time	11.8 min/epoch	7.5 min/epoch	7.4 min/epoch
ROC-AUC	0.981	0.979	0.976



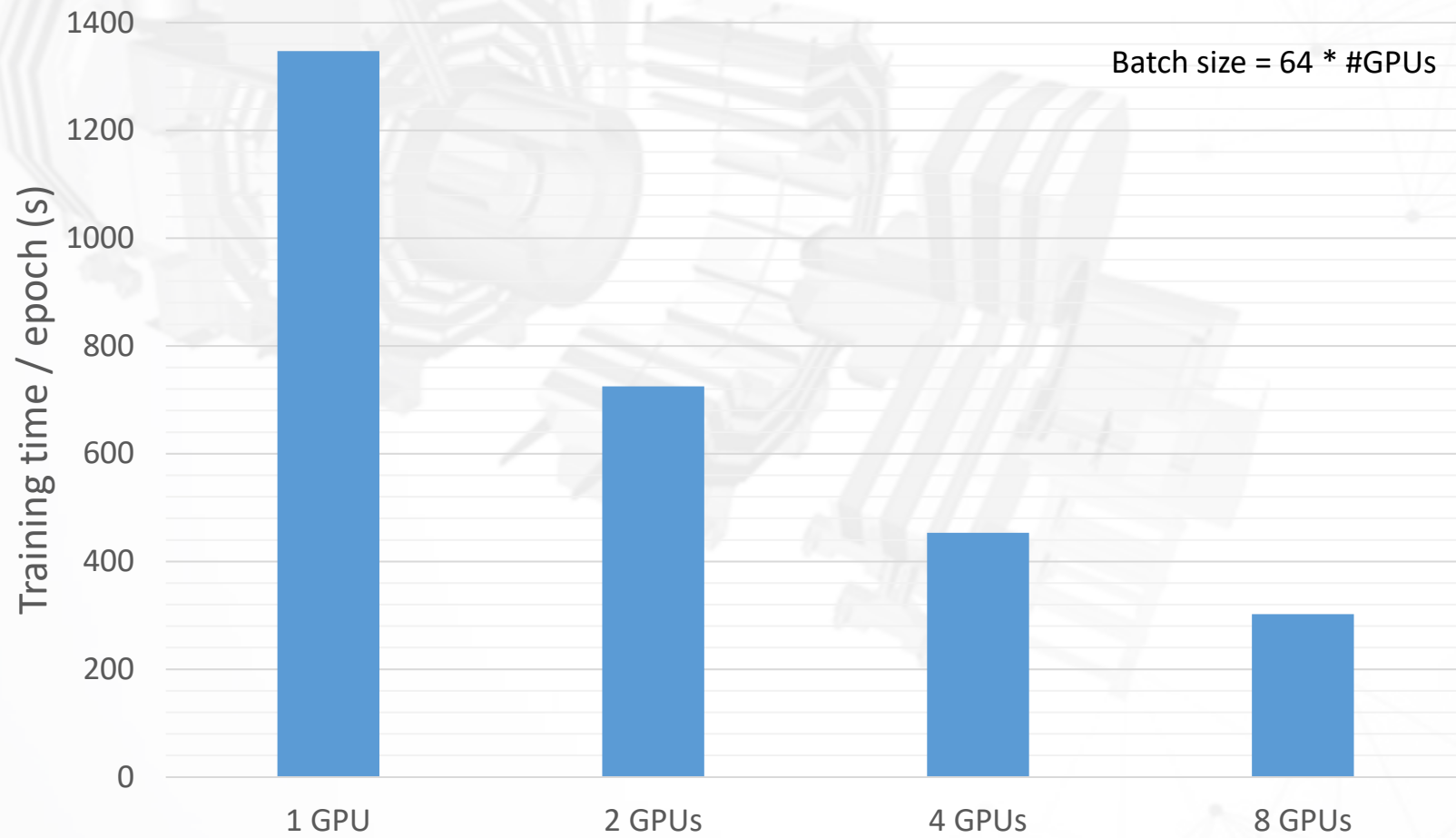
Computation time improved by increasing batch size with small performance deterioration

Performance of classifiers for different layer combination (2 GPUs, batch size 64)			
Layer combi.	<b>Track <math>p_T</math></b>	<b>Track <math>p_T+d_0+d_z</math></b>	<b>Track <math>p_T+d_0+d_z+ECAL+HCAL+BPIXhits</math></b>
ROC-AUC	0.953	0.972	0.981



Performance in agreement with previous results using TPU

- Scaling multi-GPU training on the standalone E2E boosted top jet benchmark performed on 8 Tesla V100 GPUs



- Scaling with more GPUs improves the training time up to 5 times



- End-to-end deep learning application was implemented and benchmarked
- E2E benchmark standalone models were trained on Tesla P100 GPU, Tesla V100 and TPUv3-8
- Tesla V100 and TPUv3-8 show significant (x17/x6) improvement in training and I/O compared to Tesla P100
  - This was achieved with the optimization of I/O infrastructure (CPUs and SSD)
- Scaling the deep learning training to multi-GPUs resulted in a significant speedup (x5)
  
- We would like to thank Fermilab, Google and NVIDIA for access to their servers which helped us speed up the training of machine learning models used in this study.

003250

×09

1-1

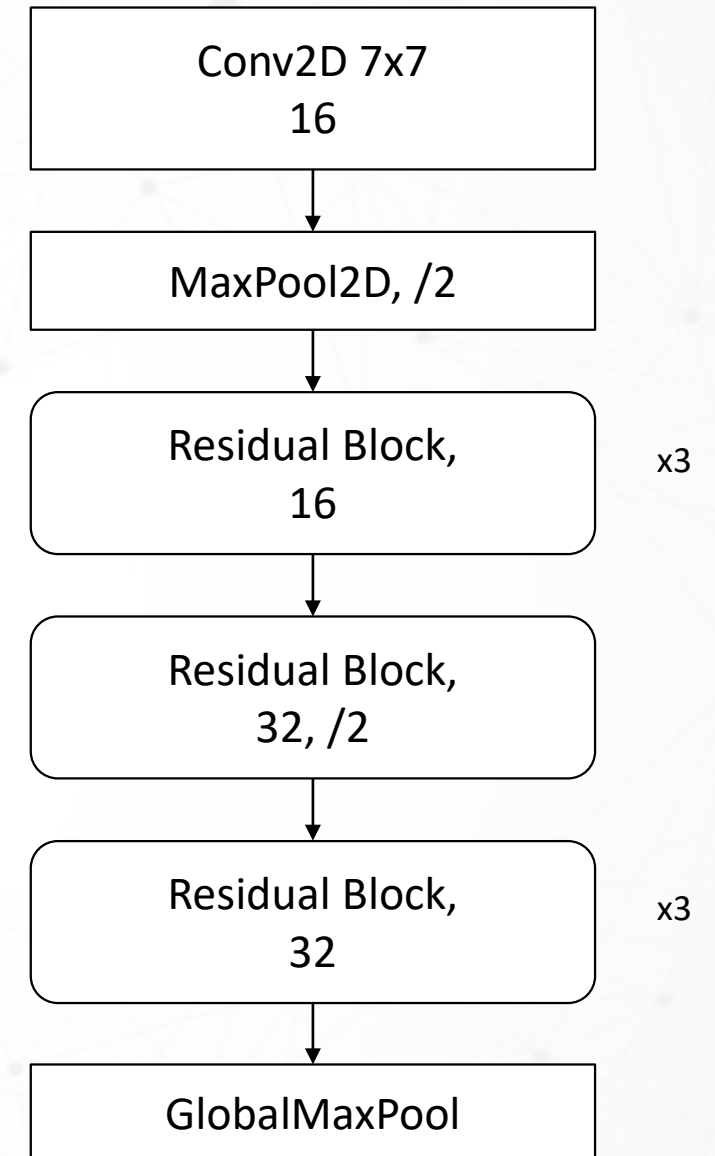
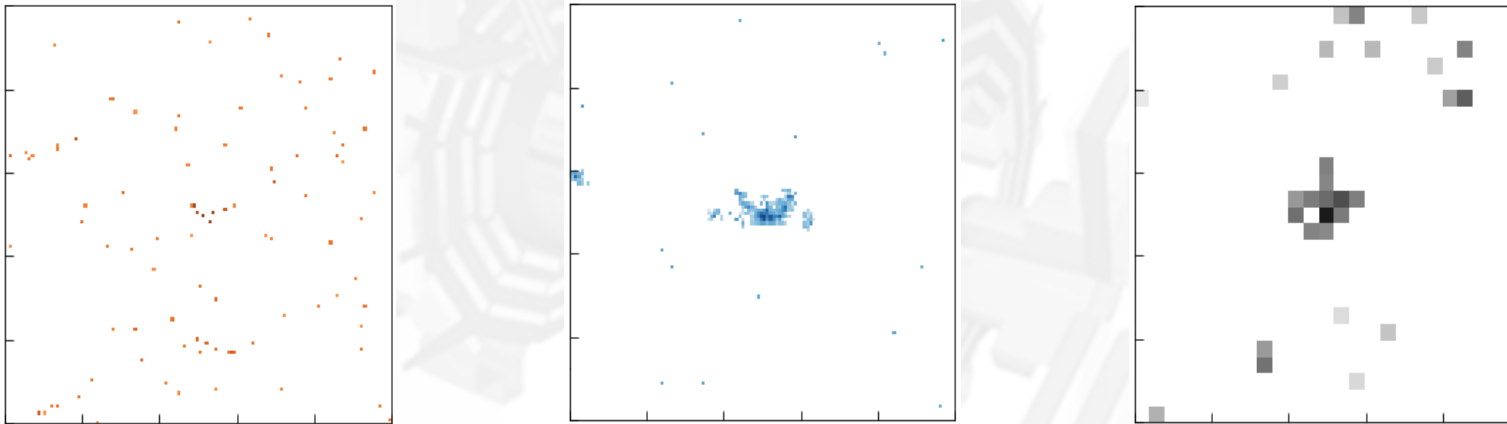
308

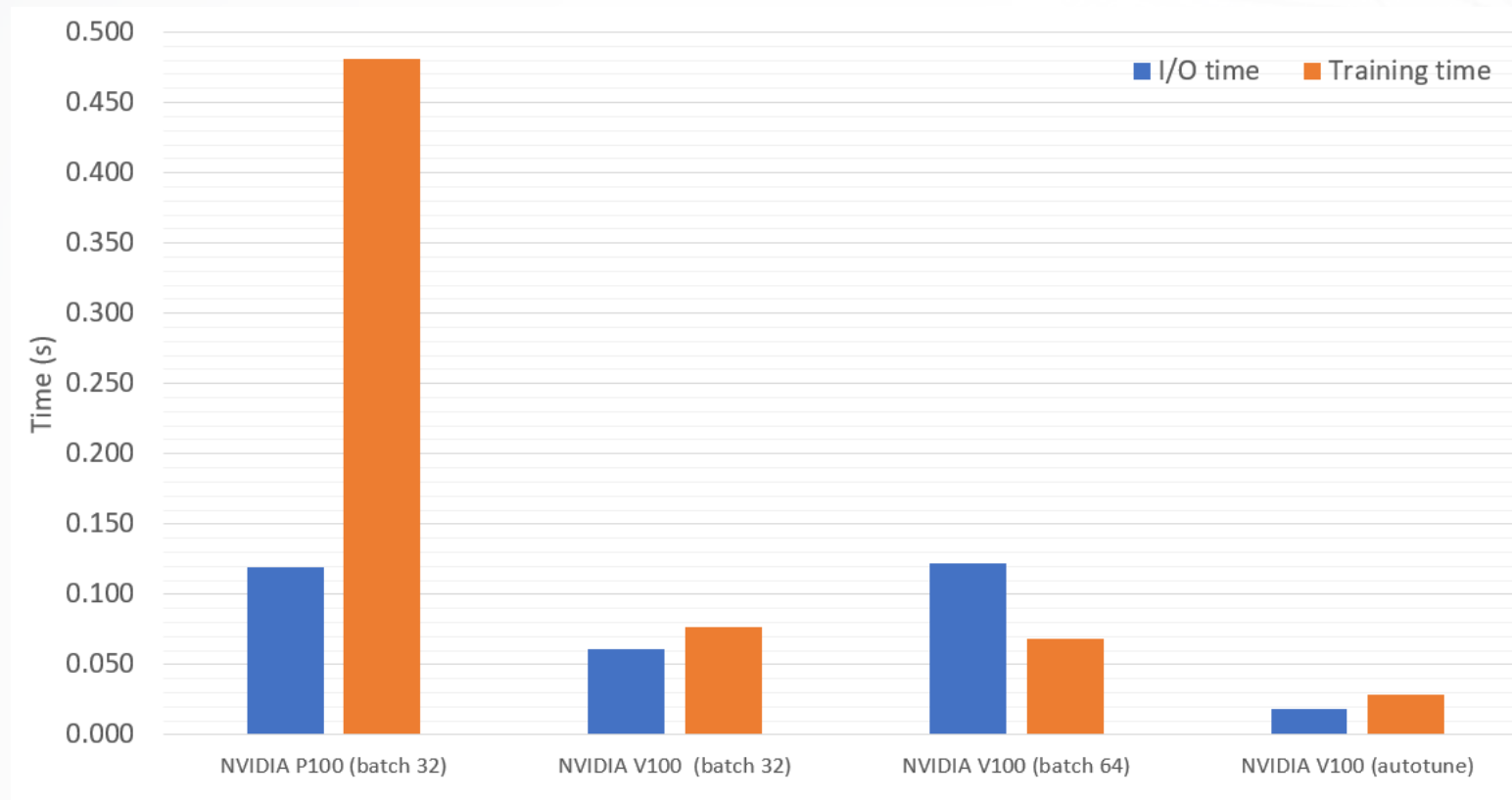
★BONUS★

SLIDES

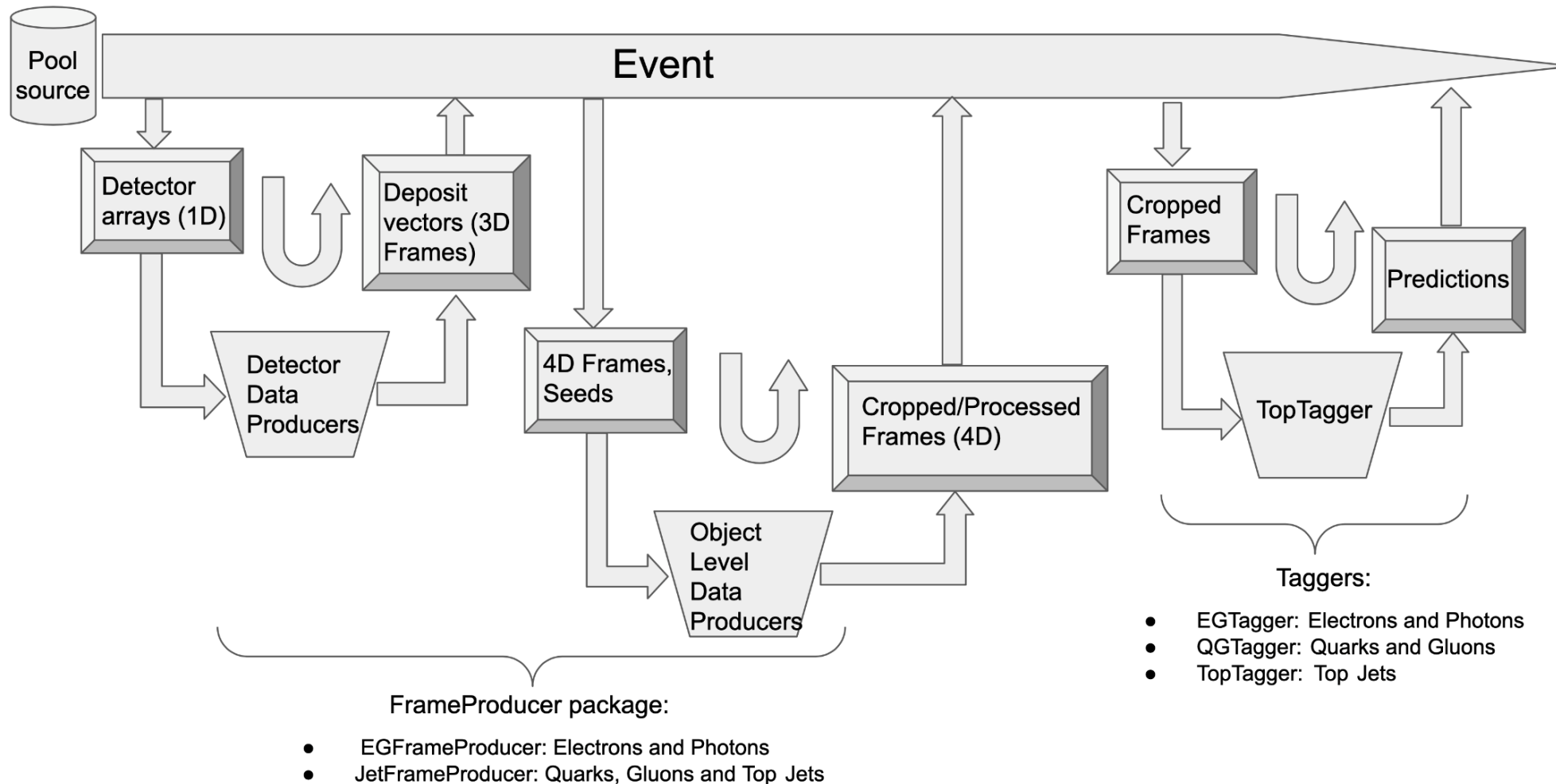


- Trained on single subdetectors, and combinations of subdetectors





- Compared to Tesla P100, Tesla V100 improvements come from:
  - Better hardware associated with reading, decompressing and pre-processing data.
  - 20/48 CPU cores, mixed precision and batch size optimisation were used to improve I/O speed
  - The number of parallel reads was set to autotune in order to mitigate bottleneck.



- Reading detector input → Storing the extracted vectors or graphs to EDM ROOT files → Extracting jet seed coordinates → Preparing the frames for inference → Running the inference → Storing the predictions