Reduced Precision Strategies for Deep Learning: 3DGAN Use Case

4th IML Machine Learning Workshop

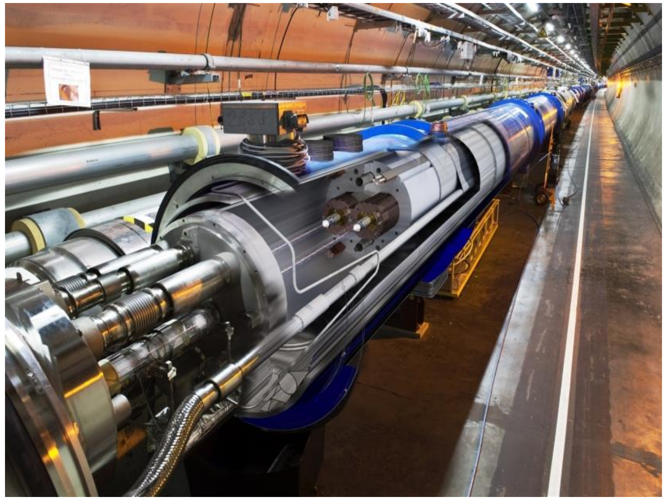
Florian Rehm [CERN openlab, RWTH Aachen University]

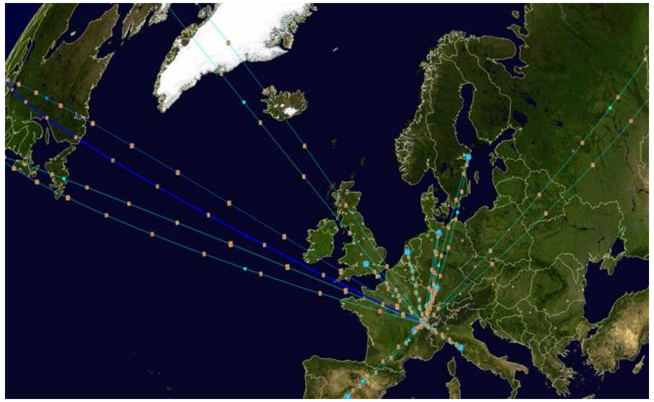
Sofia Vallecorsa [CERN openlab], Vikram Saletore [Intel], Hans Pabst [Intel], Adel Chaibi [Intel],
Kerstin Borras [DESY, RWTH Aachen University], Dirk Krücker [DESY]

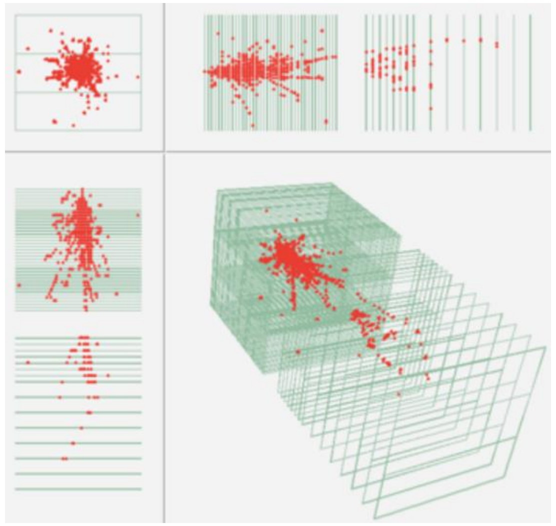
21.10.2020

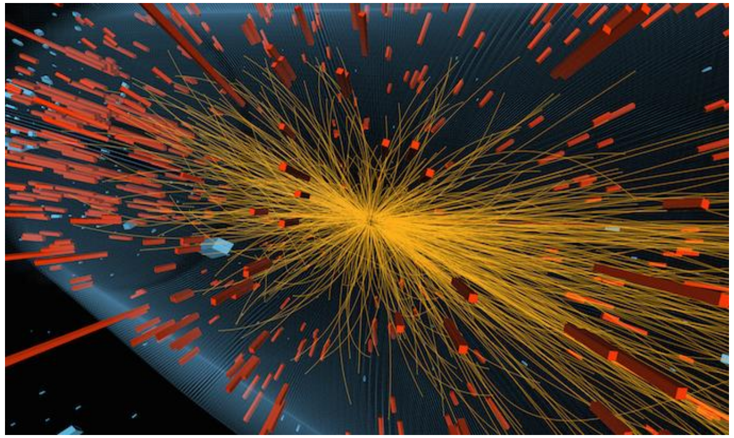
Electromagnetic Calorimeters

- Calorimeter detectors create particle showers and measure the energy of particles produced in the collisions
- Their simulations are resource intensive
  → replace them by a faster approach



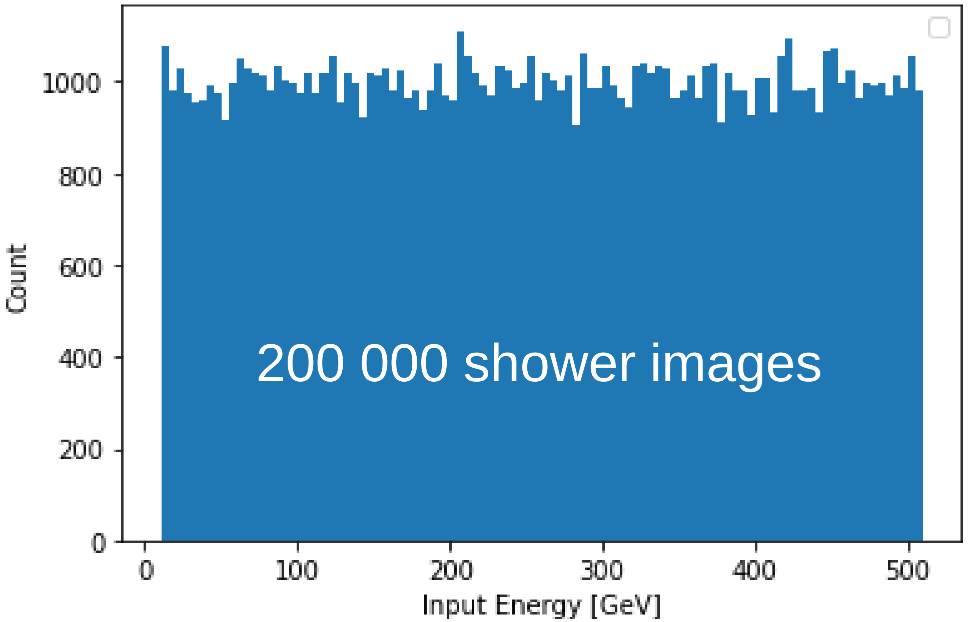


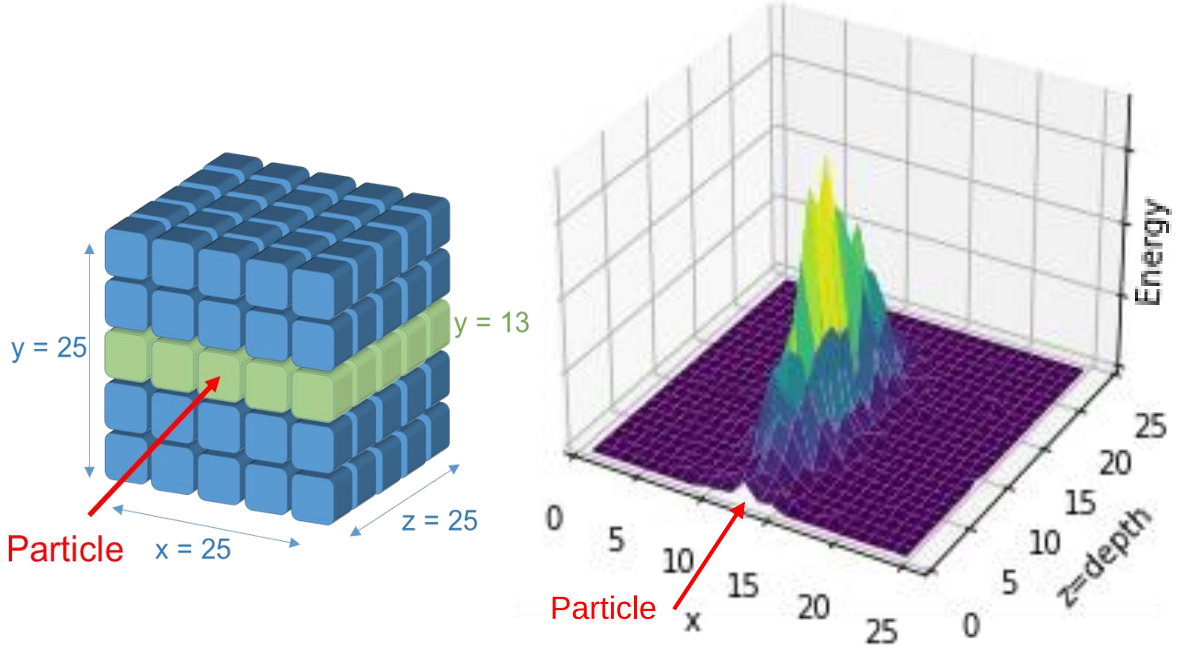




3D Training Data

3D shower image granularity: 25x25x25

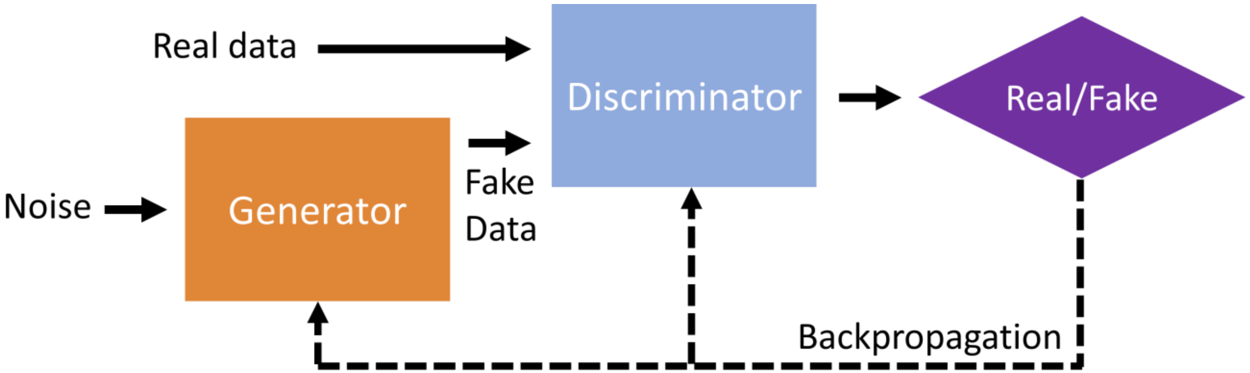




Generative Adversarial Networks

3DGAN

- Train two networks (Generator & Discriminator) in a minmax game

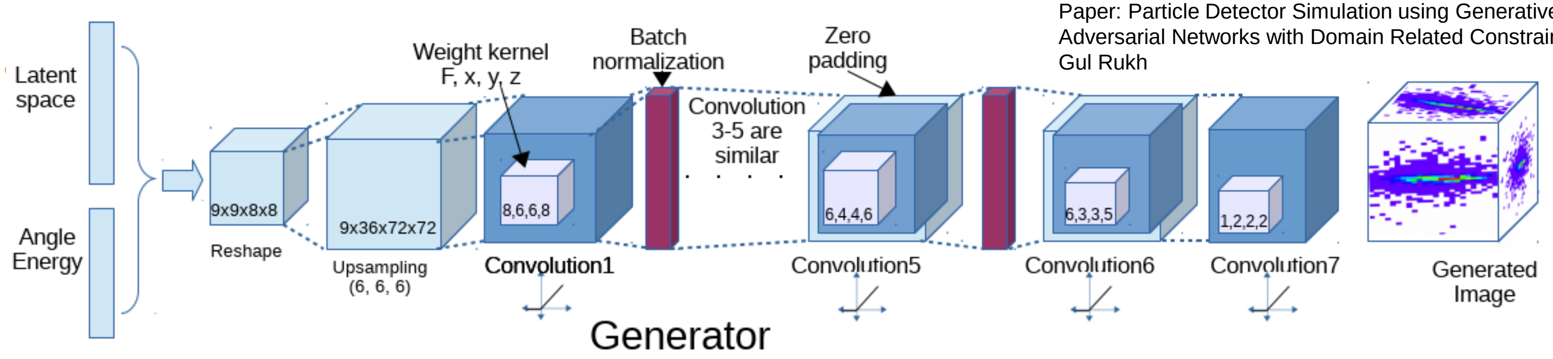


- GANs reach a good level of accuracy*
  - → We want to further decrease the computational speed with low precision computing

*Particle Detector Simulation using Generative Adversarial Networks with Domain Related Constraints, Gul Rukh

Old Conv3D Generator Network

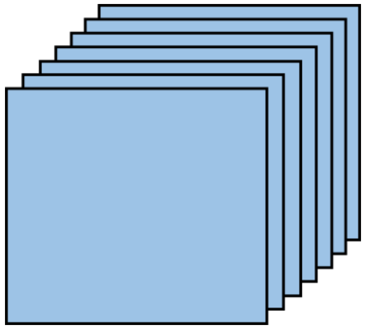
- Until now: Representing 3D images → Using 3D convolutional layers



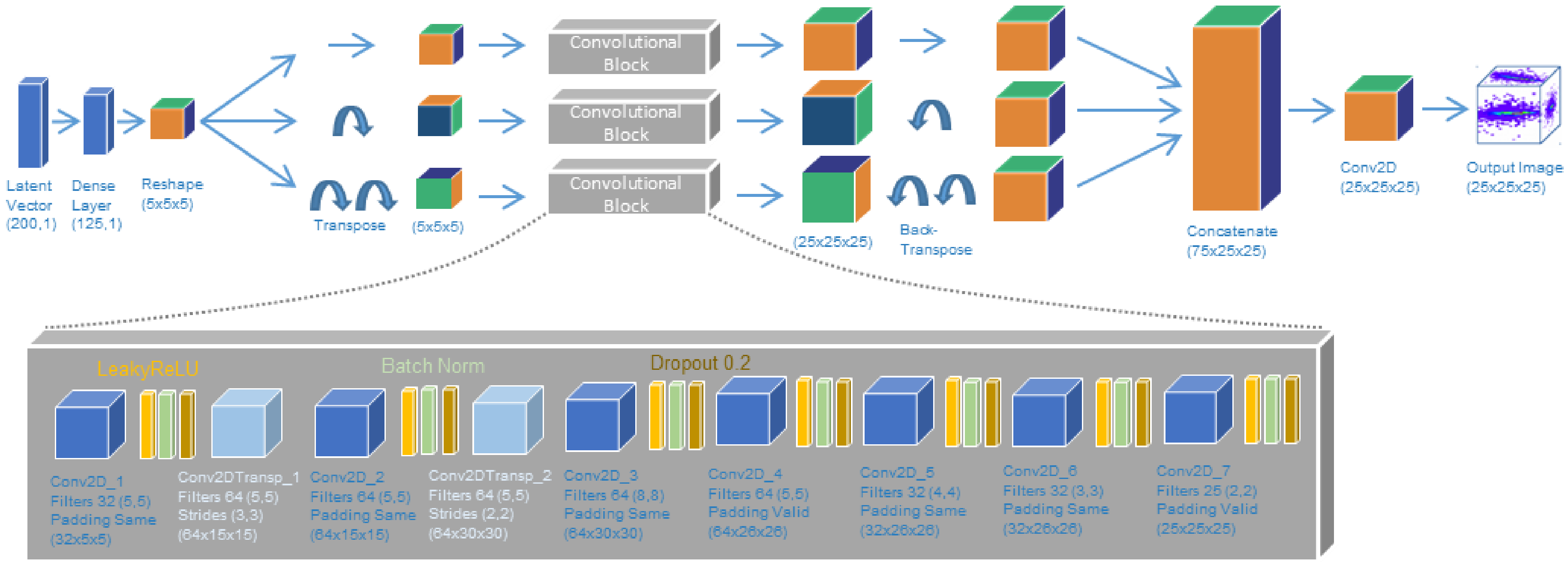
Paper: Particle Detector Simulation using Generative Adversarial Networks with Domain Related Constraints, Gul Rukh

- Conv3D layers are not supported in lower precision

  → Creating neural network consisting only of Conv2D layers
  - First approach: Channel dimension as 3rd dimension
    - Bad accuracy



New Conv2D Generator Network



- Solving a 3D problem with 2D layers
- **3x speedup** on GPU
- Increasing the number of parameters → more powerful network → higher accuracy

Why Reduced Precision?

- If we change to a new simulation approach, we want to make sure to use the hardware as efficient as possible

- Deep Learning training and inference are computational resource intensive
  - Models need a large amount of memory
  - Moving data to and from the compute engines strains the bandwidth

→ Reduced-precision computation reduces memory and bandwidth occupation

Reduced Precision Computing

- Quantization: Converting a number from a higher to a lower format
  - E.g. from float32 to int8

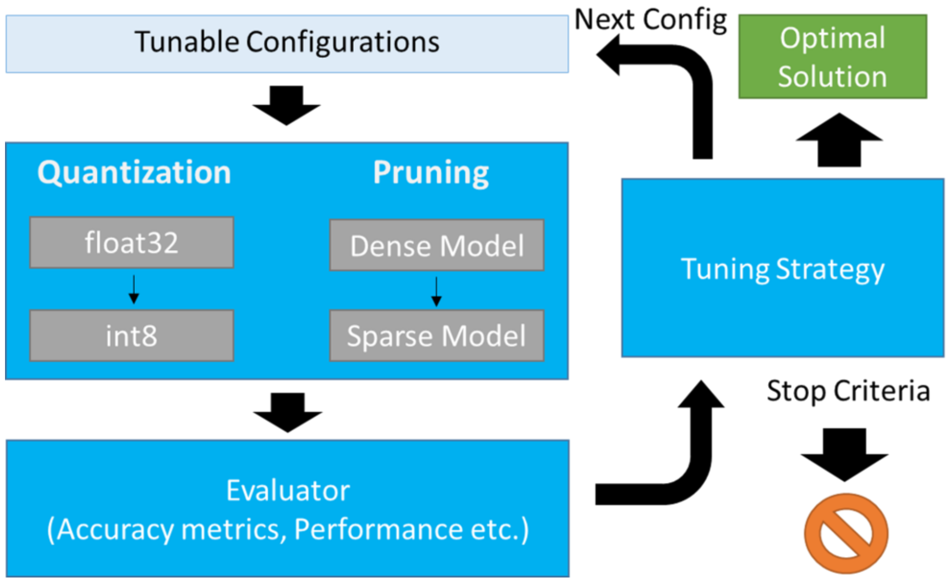
Float32
4 byte
Max Number: $3.4 * 10^{38}$

→

Int8
1 byte
Max Number: 255

- Quantization Tool:
  Intel Low Precision Optimization Tool (iLoT)
  https://github.com/intel/lp-opt-tool

- Reference Tool: TensorFlow Lite
  https://www.tensorflow.org/lite



Quantization Problems

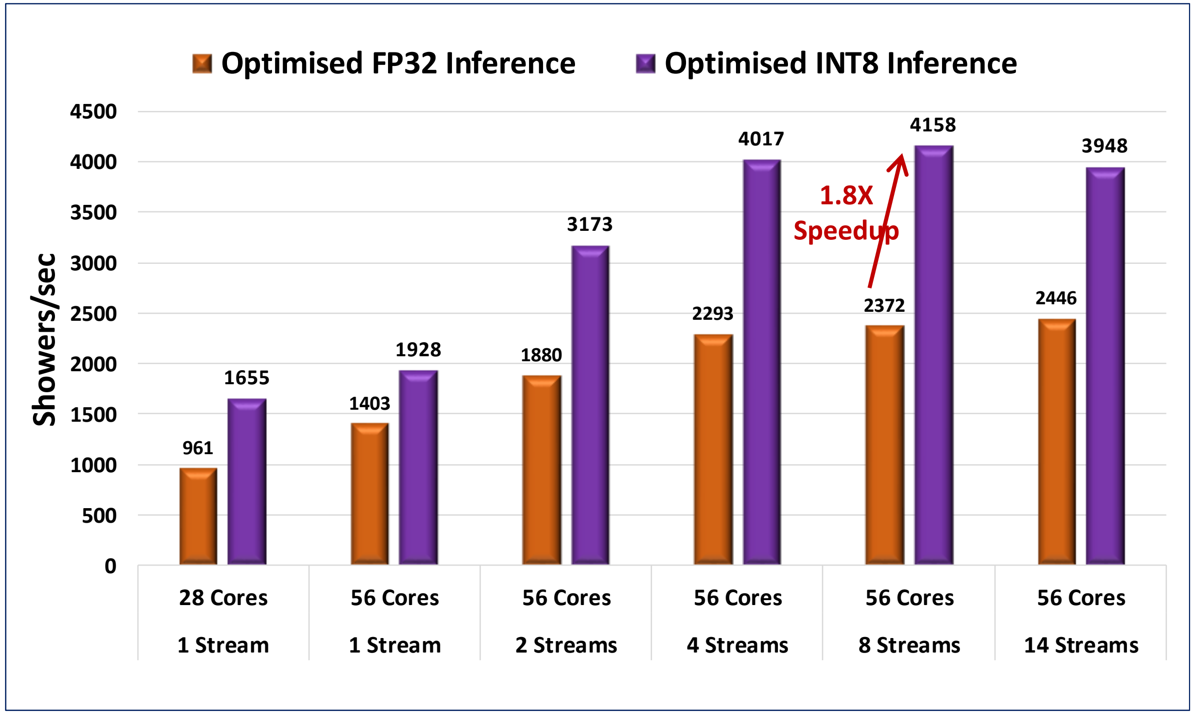
- TensorFlow supports no negative quantized values (signed int8)
- → All quantization tools do not support LeakyReLU function
- Implementation of those delayed our work

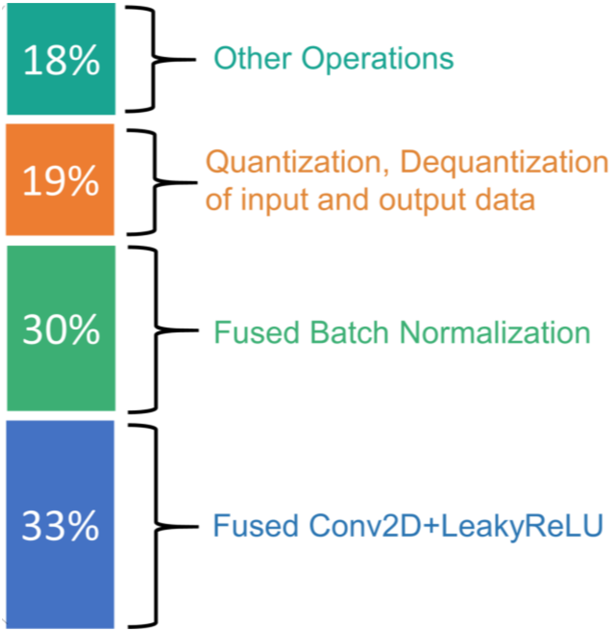
- TensorFlow Lite:
  - Does not support signed int8 → no LeakyReLU
  - Does not support transpose convolutional layers for up-sampling

Computational Evaluation

(of iLoT model)

- The quantized model reached a speedup of **1.8x**





- GAN simulation configuration: Oct, 2020: TensorFlow v2.3; Test by Intel; Test date: 11/10/2020; Platform: Intel(R) Xeon(R) Platinum 8280 CPU; Cascade Lake architecture; #Nodes: 1; #Sockets: 2; Cores/socket: 28; Threads/socket: 56; HT: On; Turbo: On; System DDR Mem Config: 12 slots / 16GB / 2933; OS: CentOS Linux 7; Kernel: 3.10.0-957.el7.x86_64
- Geant4 simulation time was taken from a previous measurement in 2018

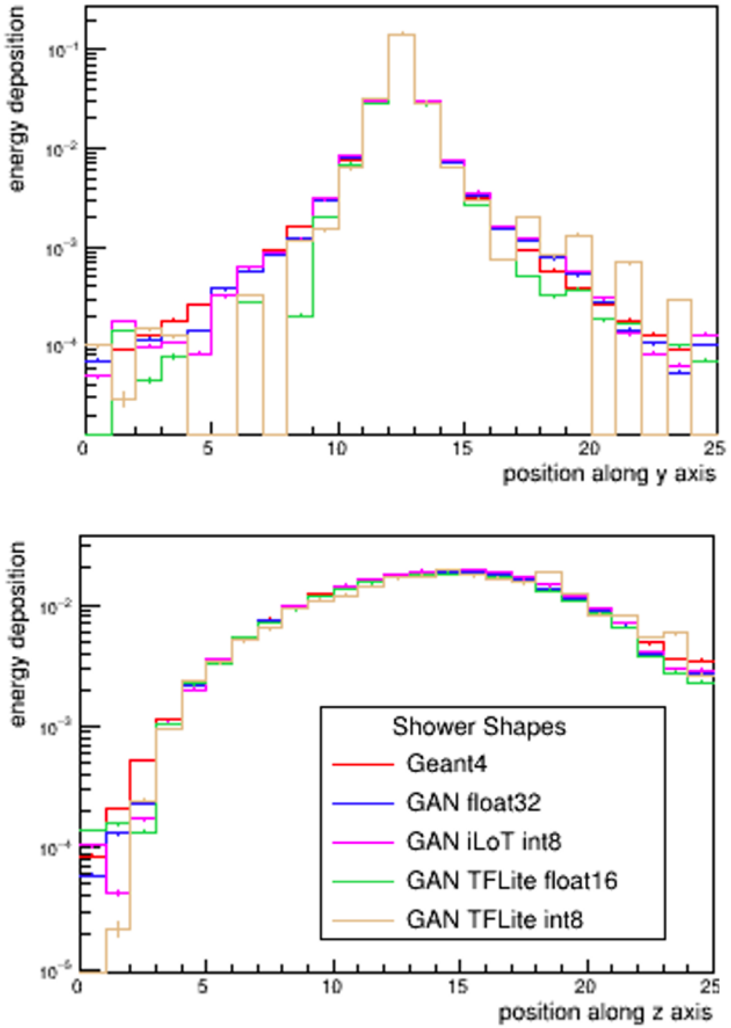
Physics Evaluation

- Validation metrics: Mean squared error (MSE) of 2D projections of the showers between GAN and validation data

Model	MSE (Lower is better)
float32	0.061
iLoT int8	**0.053**
TFLite float16	0.253
TFLite int8	0.340

➢ iLoT int8 shows a good accuracy

➢ TensorFlow Lite int8 performs worse

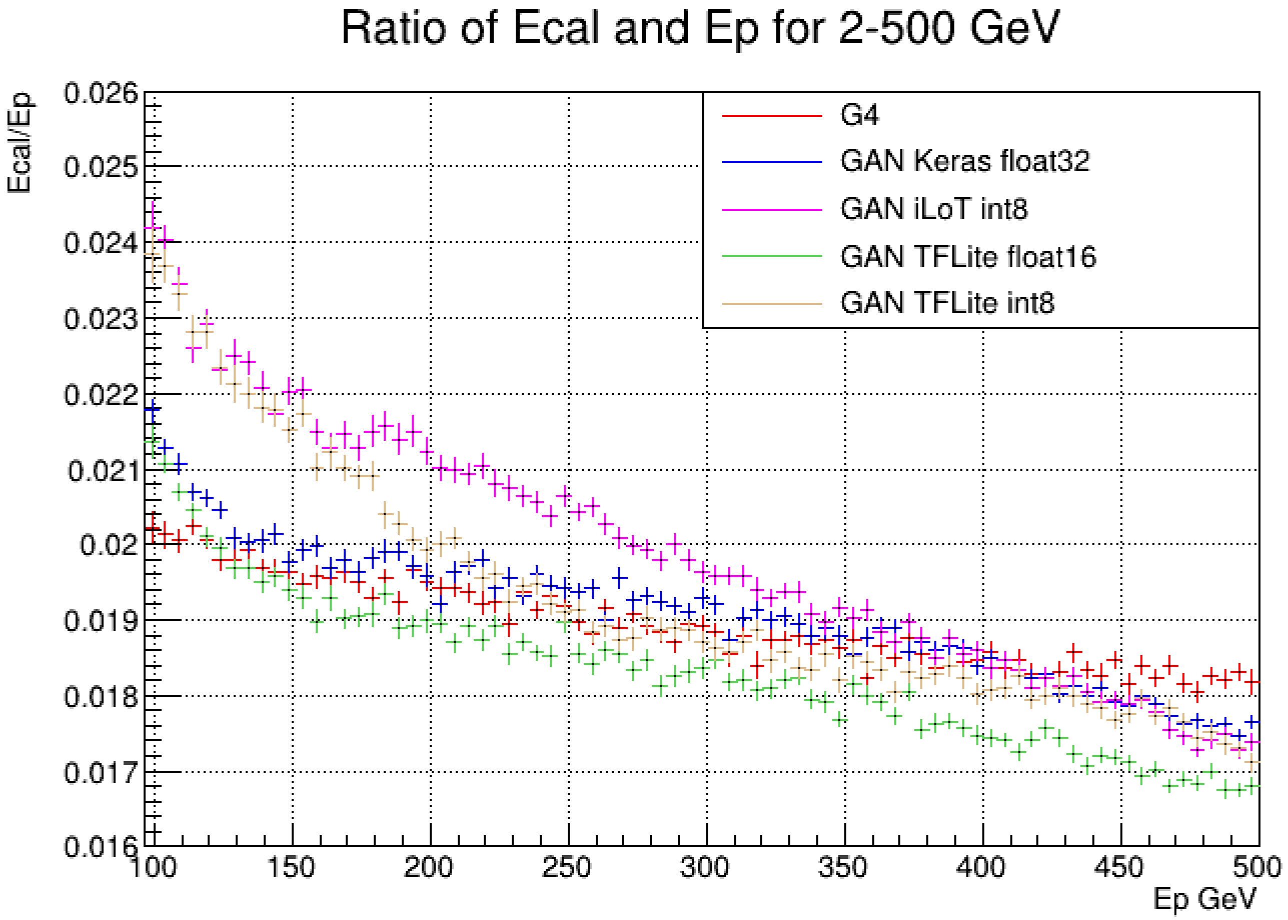


Physics Evaluation

- Ecal: Sum of all image cells ≙ total energy of the image
- Ep: Incident energy of the particle into the calorimeter or generator network

The optimization of the quantization does not take this metrics into account

→ more detailed studies needed



Summary

- **38000x** speedup of GAN to Geant4
- **2x** speedup due to conversion from Conv3D to Conv2D on CPU
  - **3x** speedup on GPU
- **1.8x** speedup due to quantization from float32 to int8

- **68000x** total speedup of quantized GAN versus Geant4 simulation

- Good physics accuracy for optimization metrics

QUESTIONS?

Reduced Precision Strategies for Deep Learning: 3DGAN Use Case

Florian Rehm [CERN openlab, RWTH Aachen]

Sofia Vallecorsa [CERN openlab], Vikram Saletore [Intel], Hans Pabst [Intel], Adel Chaibi [Intel],
Kerstin Borras [DESY, RWTH Aachen], Dirk Krücker [DESY]