AI Use Cases on HPC Environments

Intel-CERN-SURF Research Collaboration

Adel Chaibi – AI TSS

intel®

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Current Contributors

Vikram Saletore, Adel Chaibi, Emma Call, Hans Pabst

Valeriu Codreanu, Damien Podareanu, Maxwell Cai
Casper van Leeuwen, Joris Mollinga

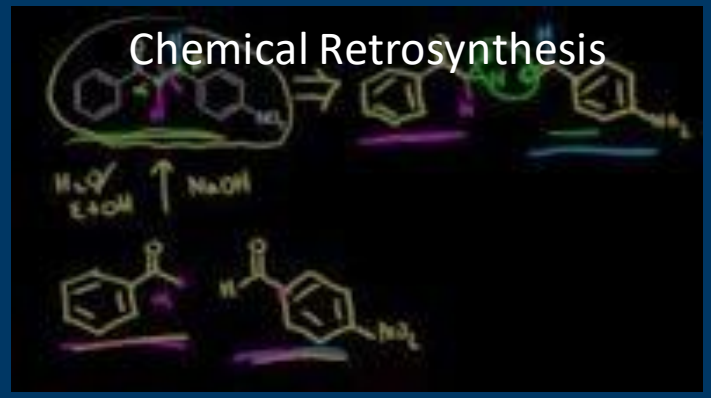
Sofia Vallecorsa, Gul Rukh Khattak, Florian Matthias

HPC + AI

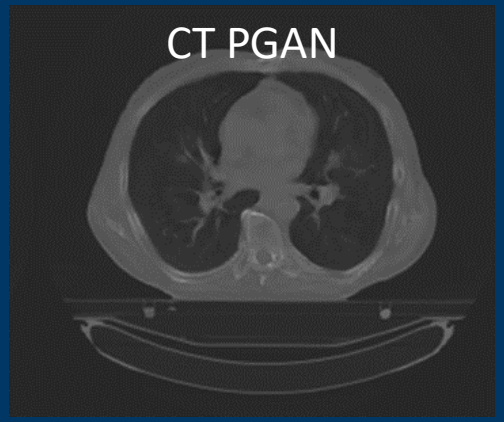
Histopathology

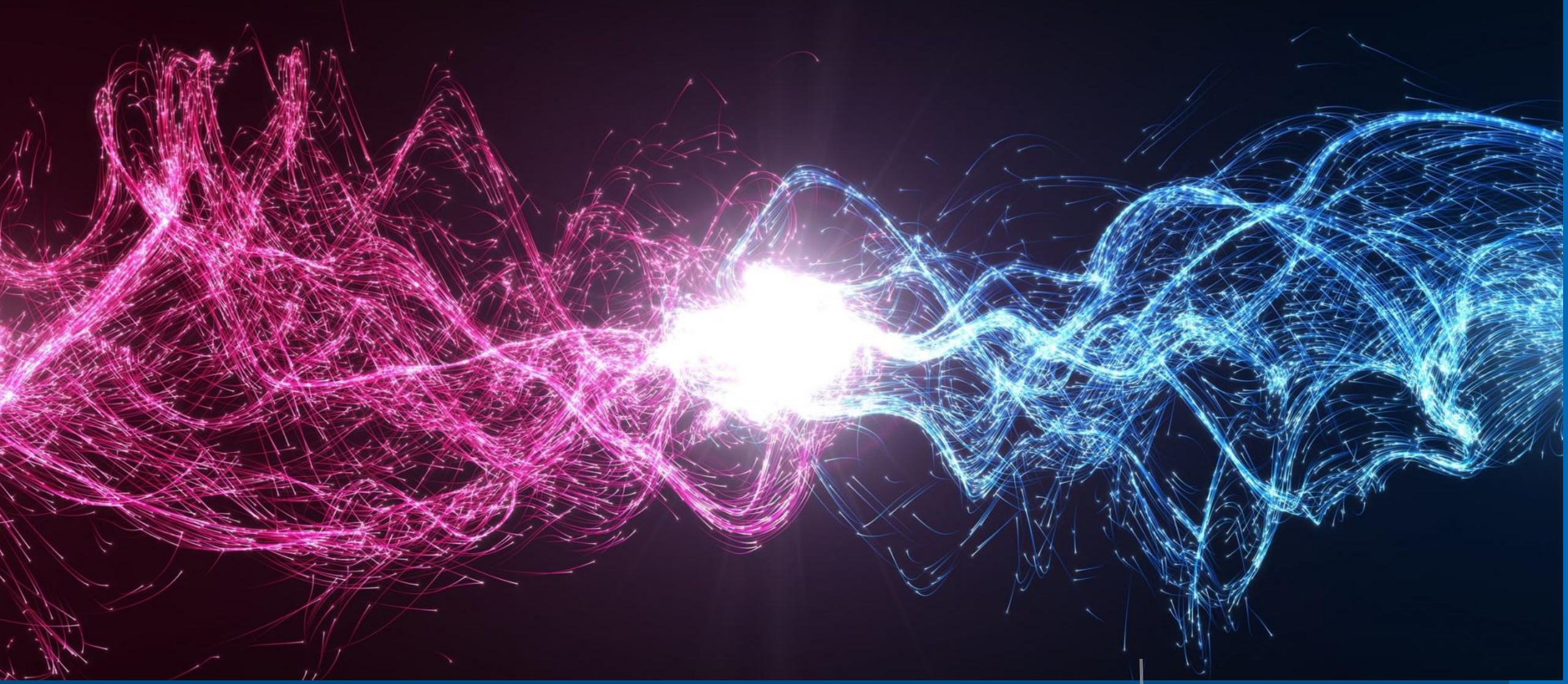
Deep Galaxy

N Body Simulation


Chemical Retrosynthesis

CERN GAN Simulation


CT PGAN



3D Convolutional GAN for Fast Simulation

Sofia Vallecorsa, Florian Rehm, Vikram Saletore, Emma Call, Adel Chaibi, Valeriu Codreanu, Hans Pabst

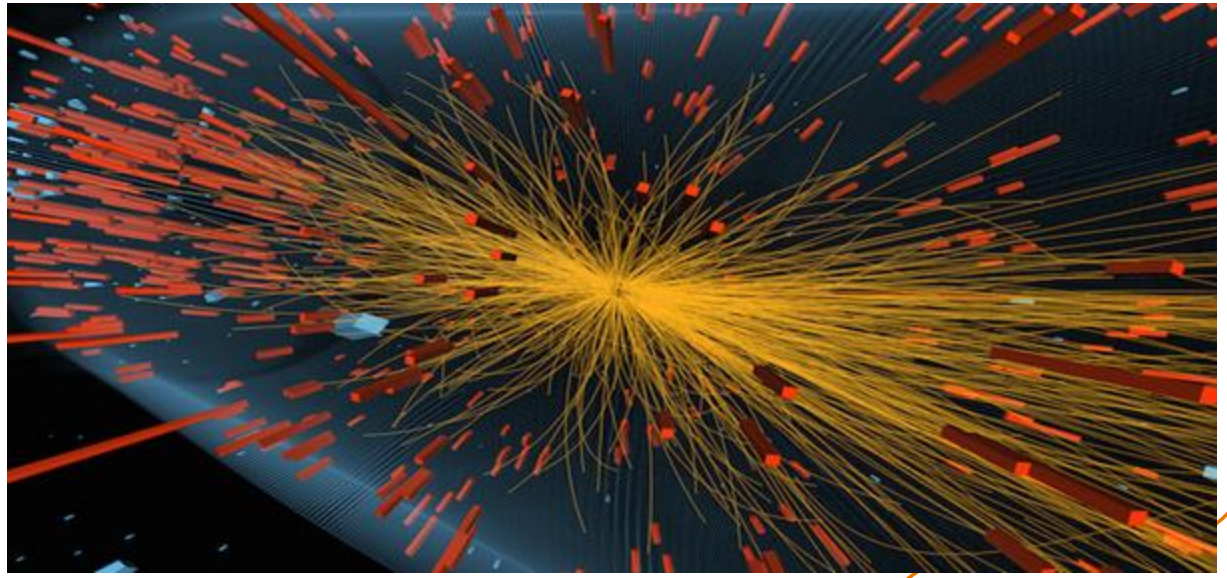
Abstract

With the Hight Luminosity LHC run at CERN, the amount of data-generation will increase even further than in its current iteration. In order to keep the statistical significance of new measurement, the amount of simulated data will have to increase accordingly . The 3DGAN model performs fast simulation considering particle energy and trajectory angle achieving state-of-the-art results in terms of physics accuracy.

We used the Intel Neural Compressor (INC) software tool for graph optimization and quantization of the 3D-Conv Generative Adversarial Network (GAN) with INT8. We show that INT8 can be used without the loss of accuracy for the complex 3DGAN. We show a 20% improvement with INC graph optimization. We demonstrate that 4th Gen Xeon Sapphire Rapids with HBM and AMX/TMUL achieves improvement of 4.8X over 3rd Gen Xeon® Ice Lake. We also show that the quantized INT8 model accuracy matches baseline FP32 accuracy.

Summary

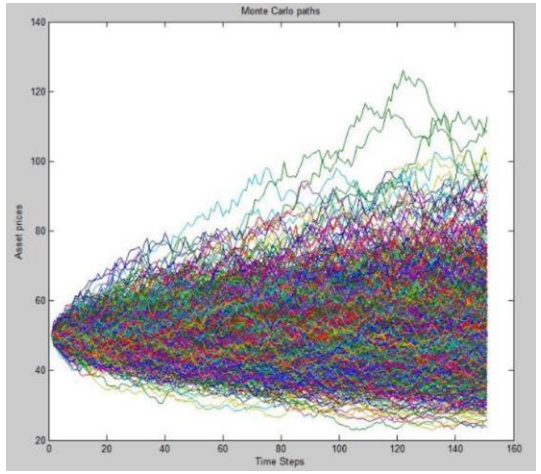
- Geant4 is a toolkit for the simulation of the passage of particles through matter for CERN's High Energy Physics
- Uses compute intensive Monte Carlo extensively in the simulation
- 3D-GAN AI Model replaces Monte Carlo with high degree of accuracy
- **The reduced precision quantized AI model on 4th Gen Intel® Xeon® Processor achieves > 8500X improvement over Geant4**

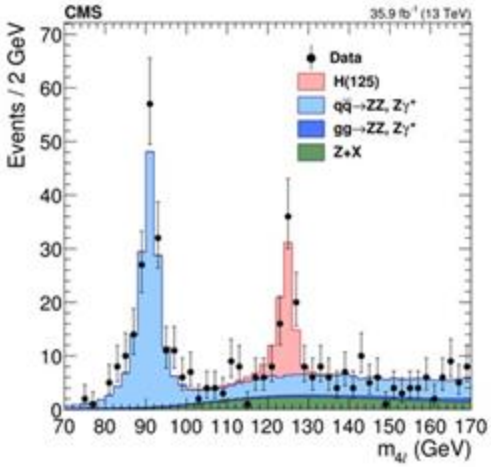


Introduction

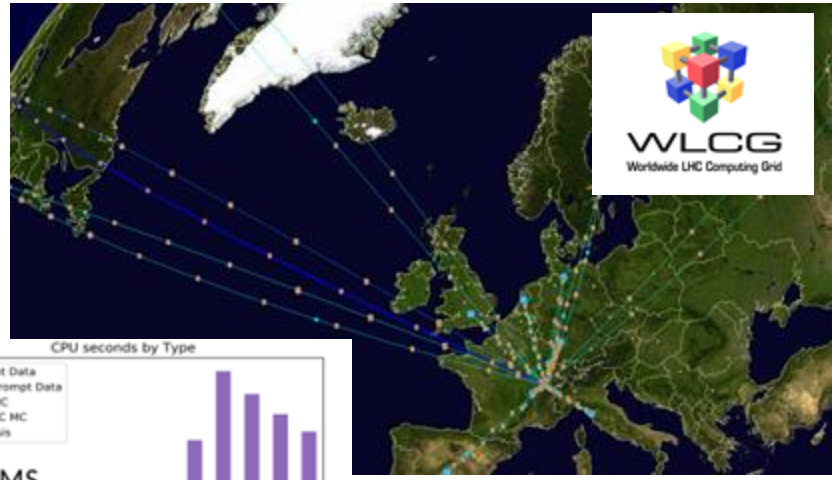
- HPC-AI Workload at Large Hadron Collider
- 3D-Generative Adversarial Networks (3D-GANs)

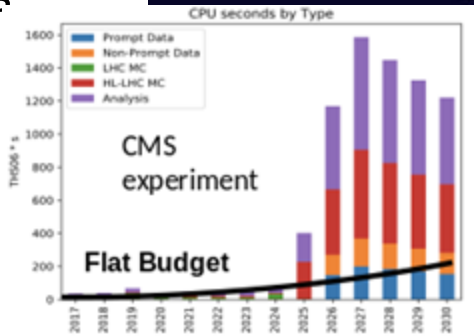
High Energy Physics Simulation



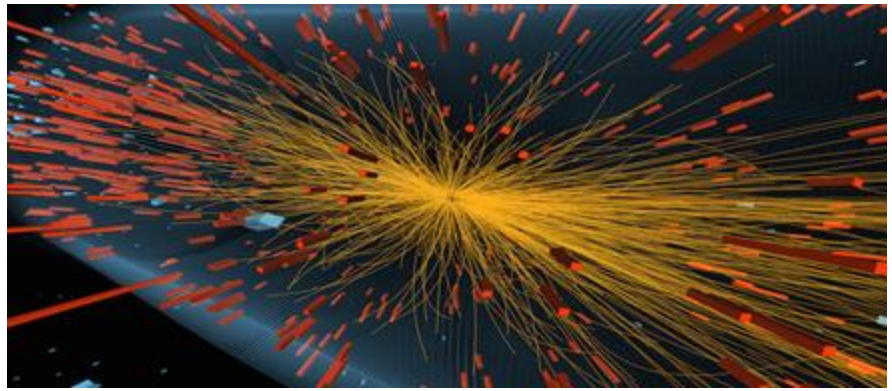


- Understand how detector design affects measurements and physics
  - Correct for inefficiencies, inaccuracies, unknowns
  - Compare theory models to data
- Complex physics and geometry modeling
  - Monte Carlo Simulations
  - >50% of Worldwide LHC Computing Grid (WLCG) power today
  - Increase by 100x by 2028!

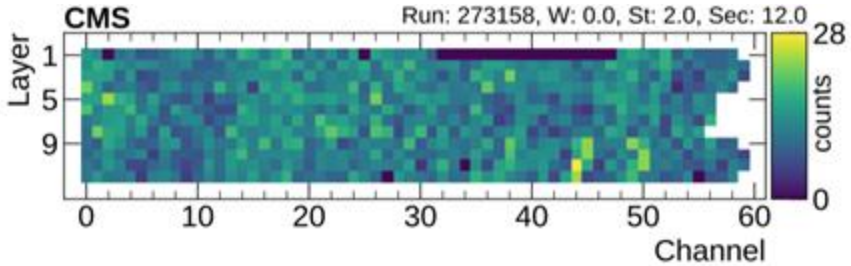




AI Models for High Energy Physics

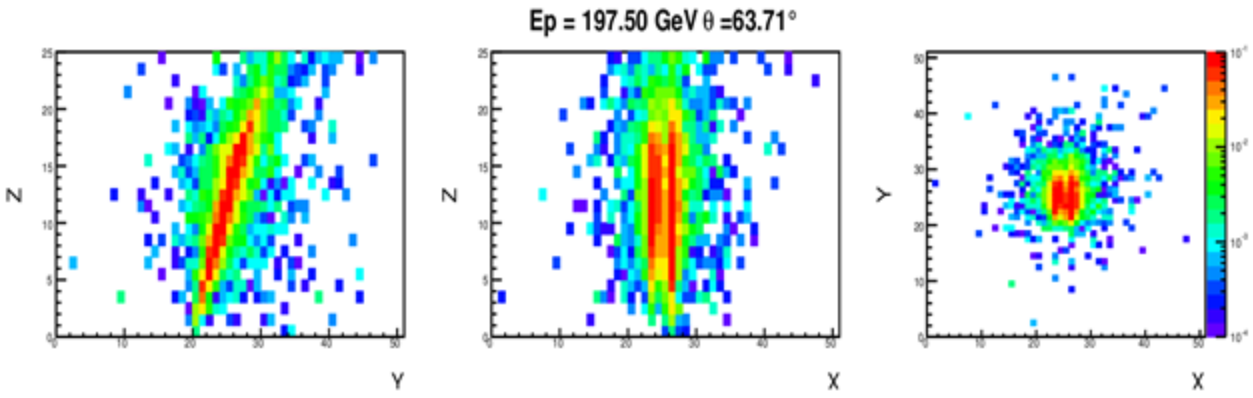


- All venues of science are benefitting from AI for problems where:
  - Underlying processes are difficult to model
  - Require high computational sources
  - Time consuming
  - Noisy data
- High Energy Physics
  - Applications
    - Data Processing and Analysis
    - Real time selection
    - Simulation
- AI crucial for HEP experiments
  - HPC hardware
    - Maximize performance
    - Fast time-to-model

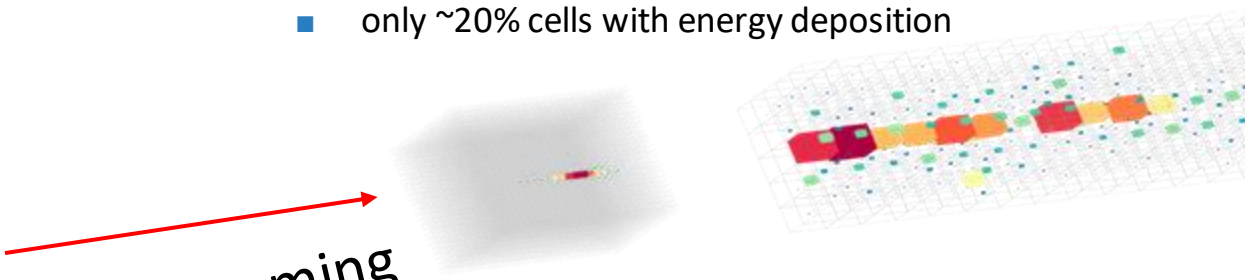


Dataset

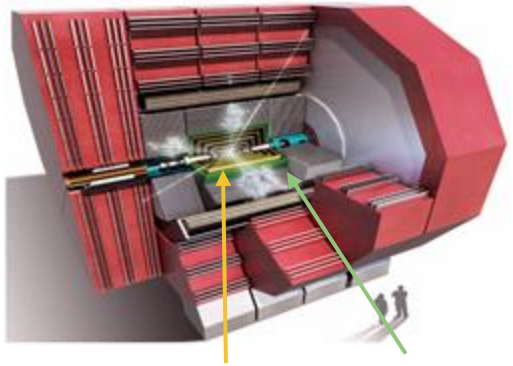
Single particles depositing energy in a detector via electromagnetic interactions (calorimeter detector)



- **Monte Carlo (Geant4) generated data set** developed for ML applications
  - Future detector geometry demonstrator
  - Highly granular sensors → critical for particle identification and energy determination
- Single electrons with energies in the range 2-500 GeV
  - Realistic trajectories (parametrized using one angle)
- Detector response interpreted as 3D grid images
  - 400,000 events → each a 51 x 51 x 25 image → 65,025 cells
  - Highly sparse
    - only ~20% cells with energy deposition

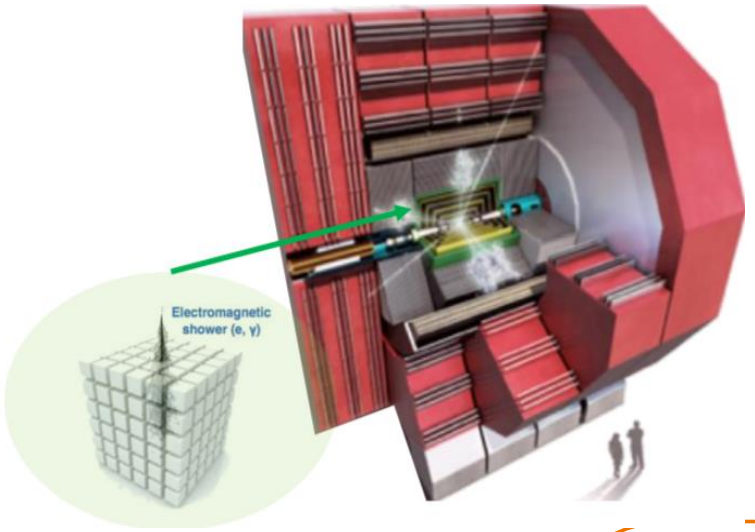


http://cds.cern.ch/record/2254048



http://clicdp.web.cern.ch/

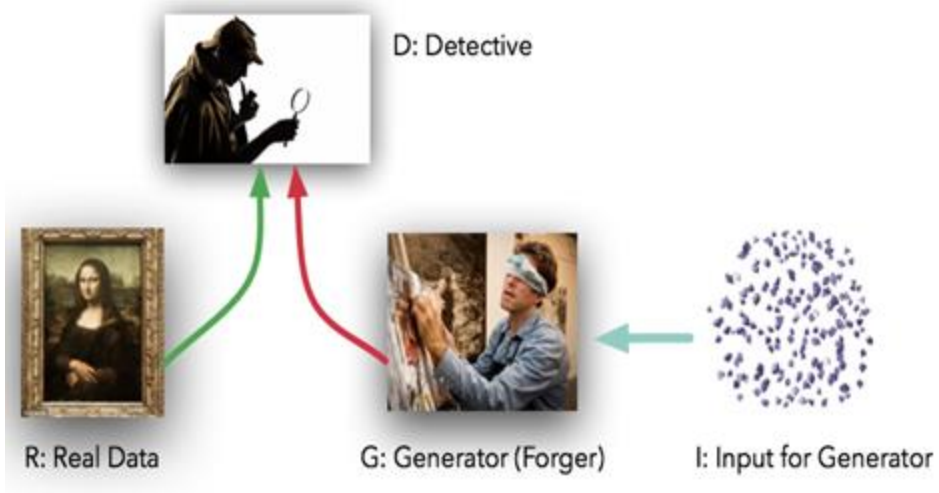
Why Use GANs for HEP Simulations

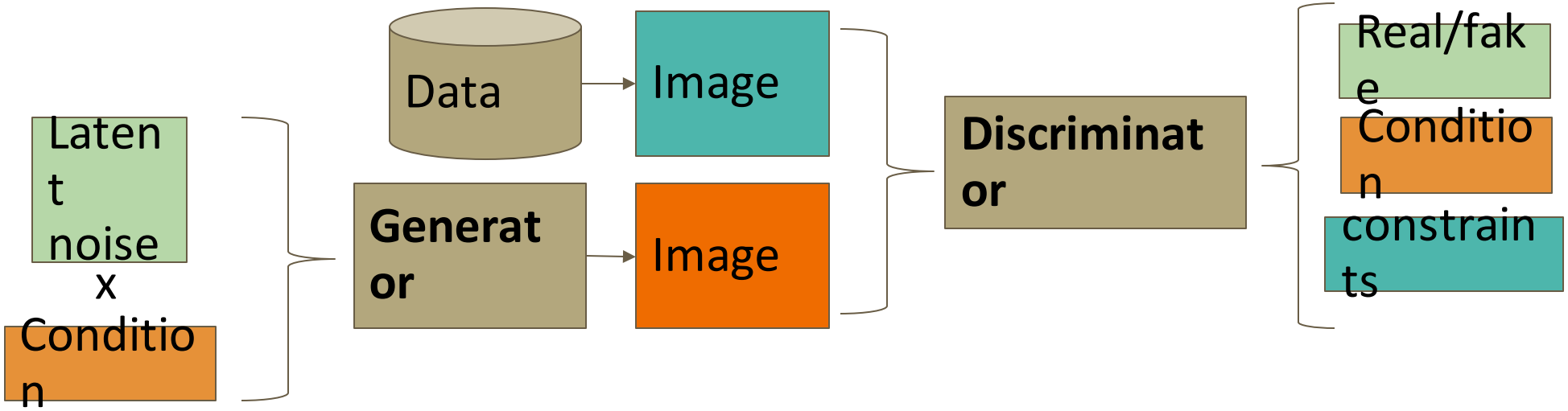


- HEP detectors take 3D pictures of energy showers
  - Each pixel = the energy value
- GANs excel at image generation
- GANs can simulate new data
- Motivation
  - Significantly faster than Monte Carlo during **INFERENCE** mode
  - Highly accurate
  - Flexible
    - Architecture optimization
    - High level of transfer

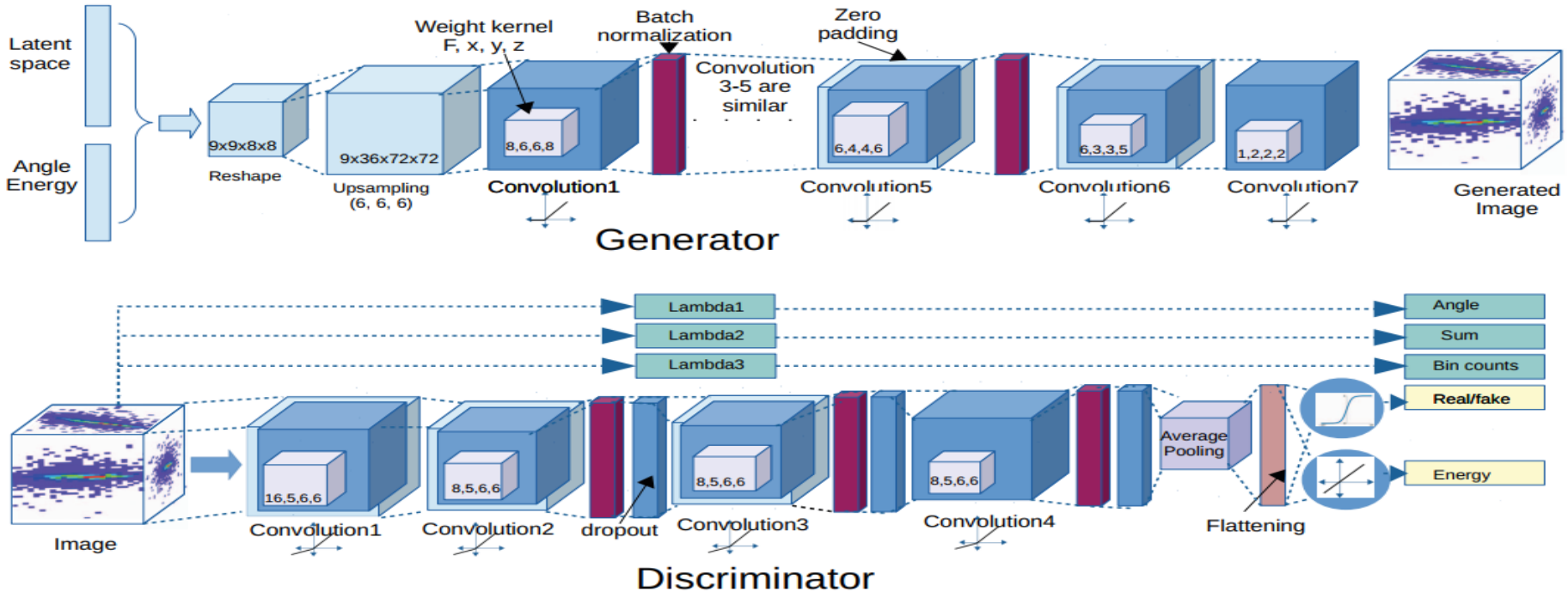
Generative Adversarial Networks

- Simultaneously train two networks that compete and cooperate with each other
  - Discriminator
  - Generator





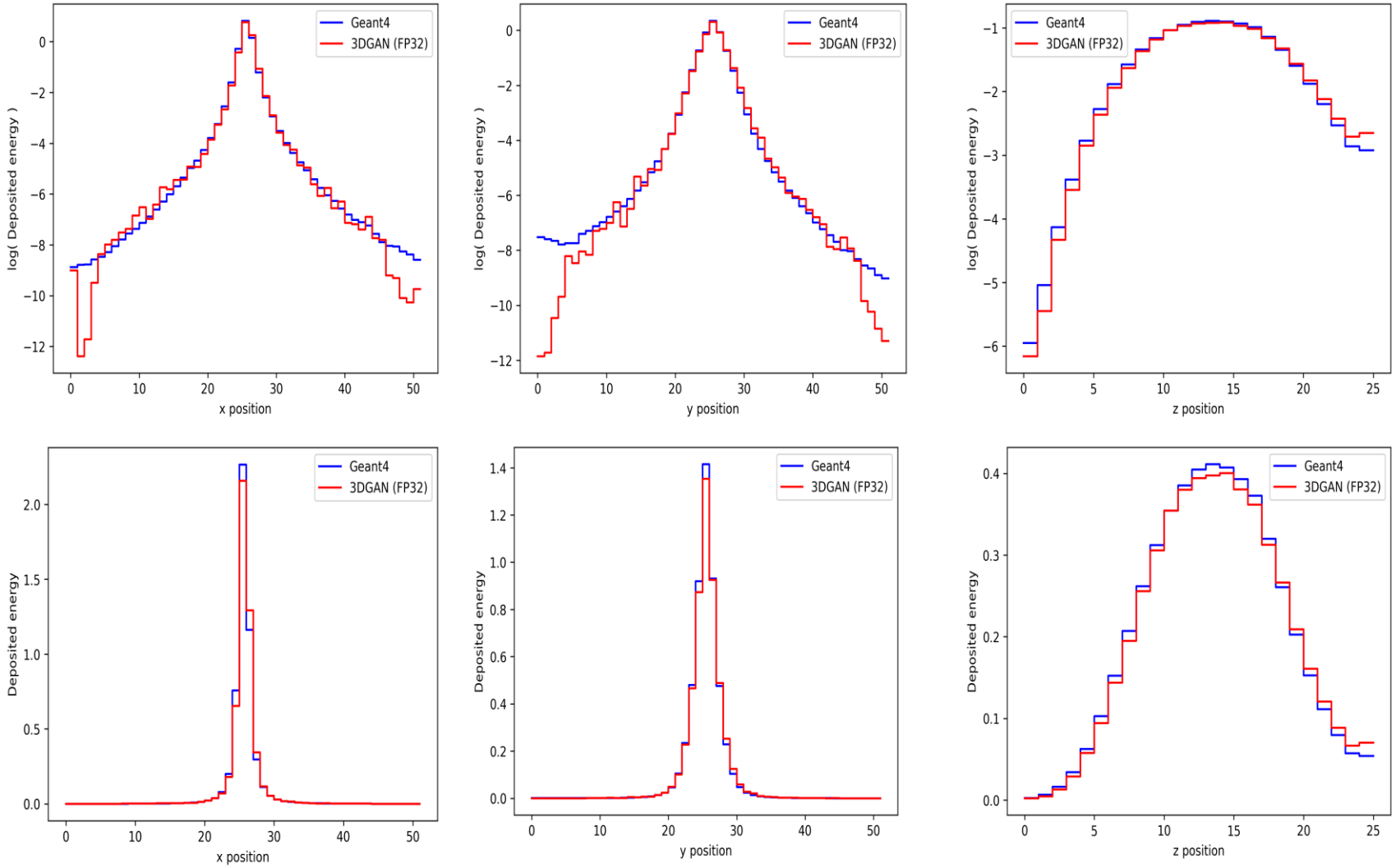
3D-GANs Architecture

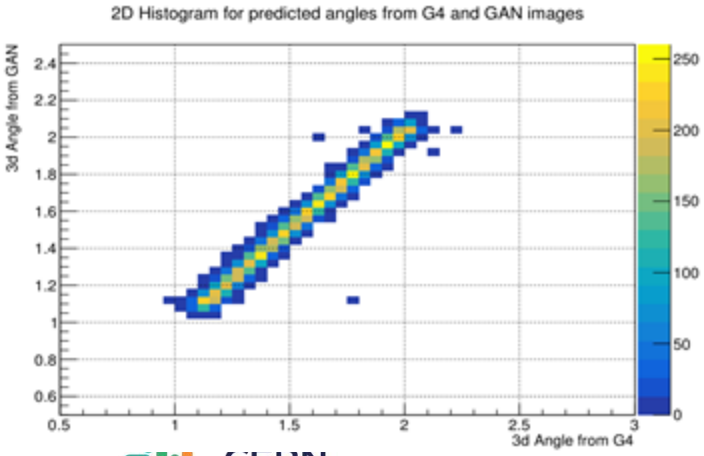


- **Generator**
  - **7 Layer Network**
  - **Concatenate random vector with energy and angle data**
  - **Up-sampling operations**
- **Discriminator**
  - **Outputs a sigmoid neuron and a linear neuron**
  - **Lambda functions**
- https://github.com/svalleco/3Dgan/

Physics Performance: Geant4 vs 3D-GAN (FP32)

- Shower Shapes
- Energy deposited Along x, y and z axis





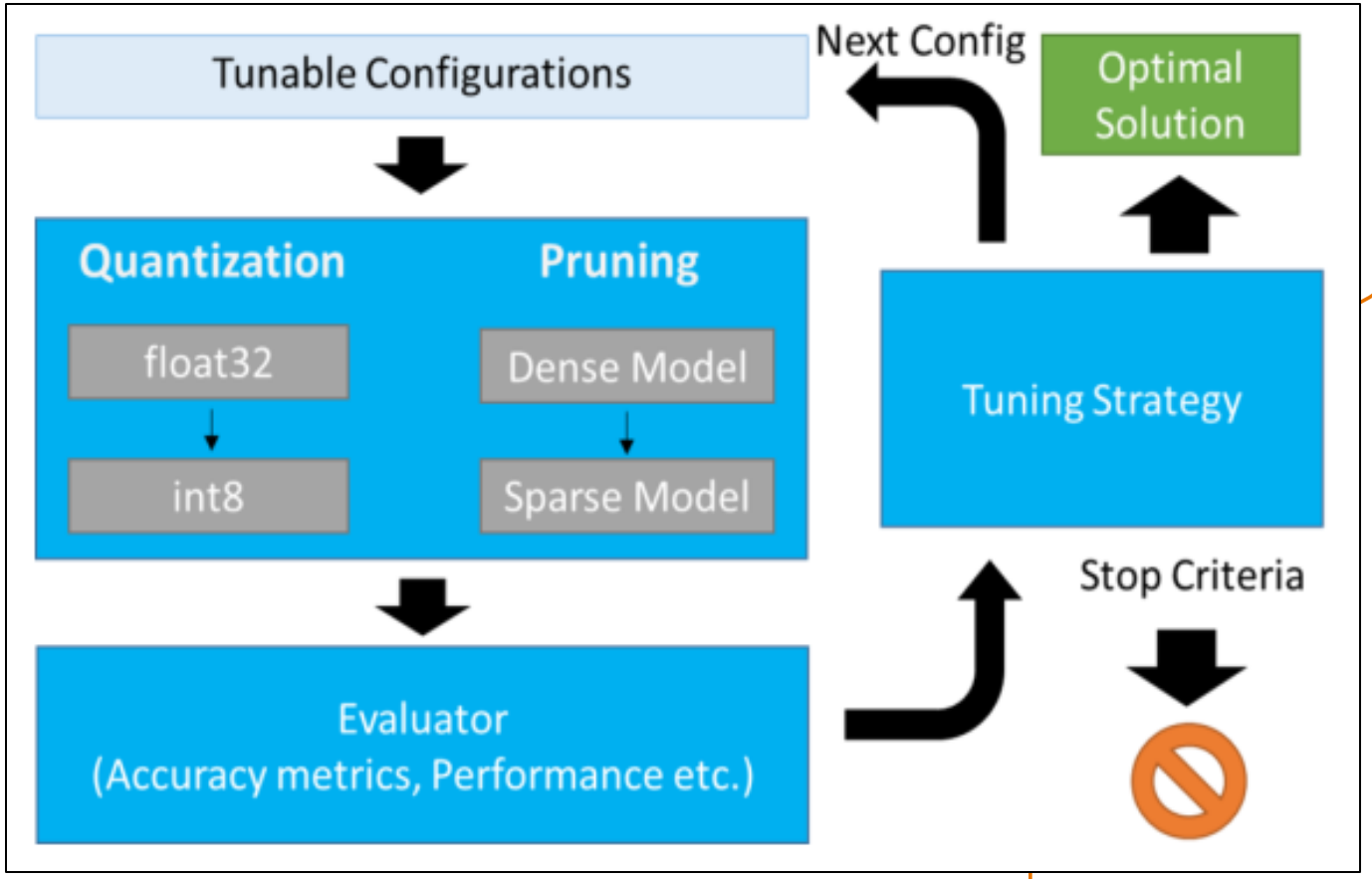
Inference Performance

Time to Create an Electron Shower on Intel Xeon®		
Method	**msecs/Shower**	**Speedup**
Classical Monte Carlo (Geant4) Intel Xeon®	5000	1.0
3D-GAN (BS=256) 2-stream, Intel® Xeon® 8360 (FP32 Précision)	2.3	**1988X**

CAN WE DO BETTER WITH REDUCED PRECISION?

Quantization with Intel Neural Compressor

- Quantization Tool: Intel Neural Compressor
  - Intel® Neural Compressor performs model compression to reduce the model size and increase the speed of deep learning inference for deployment on CPUs or GPUs
  - Converge quickly on quantized models though automatic accuracy-driven tuning strategies.
  - Prune model weights by specifying predefined sparsity goals that drive pruning algorithms.
  - Distill knowledge from a larger network (“teacher”) to train a smaller network (“student”) to mimic its performance with minimal precision loss.



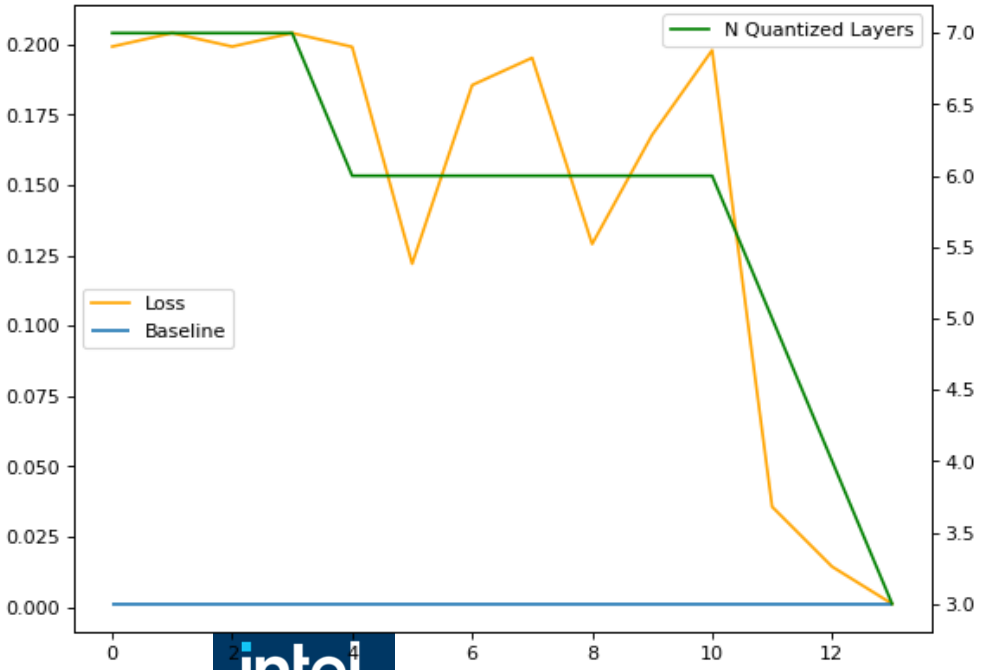
https://www.intel.com/content/www/us/en/developer/tools/oneapi/neural-compressor.html
https://github.com/intel/neural-compressor/releases/tag/v1.14.2

Meeting Accuracy with Intel Neural Compressor

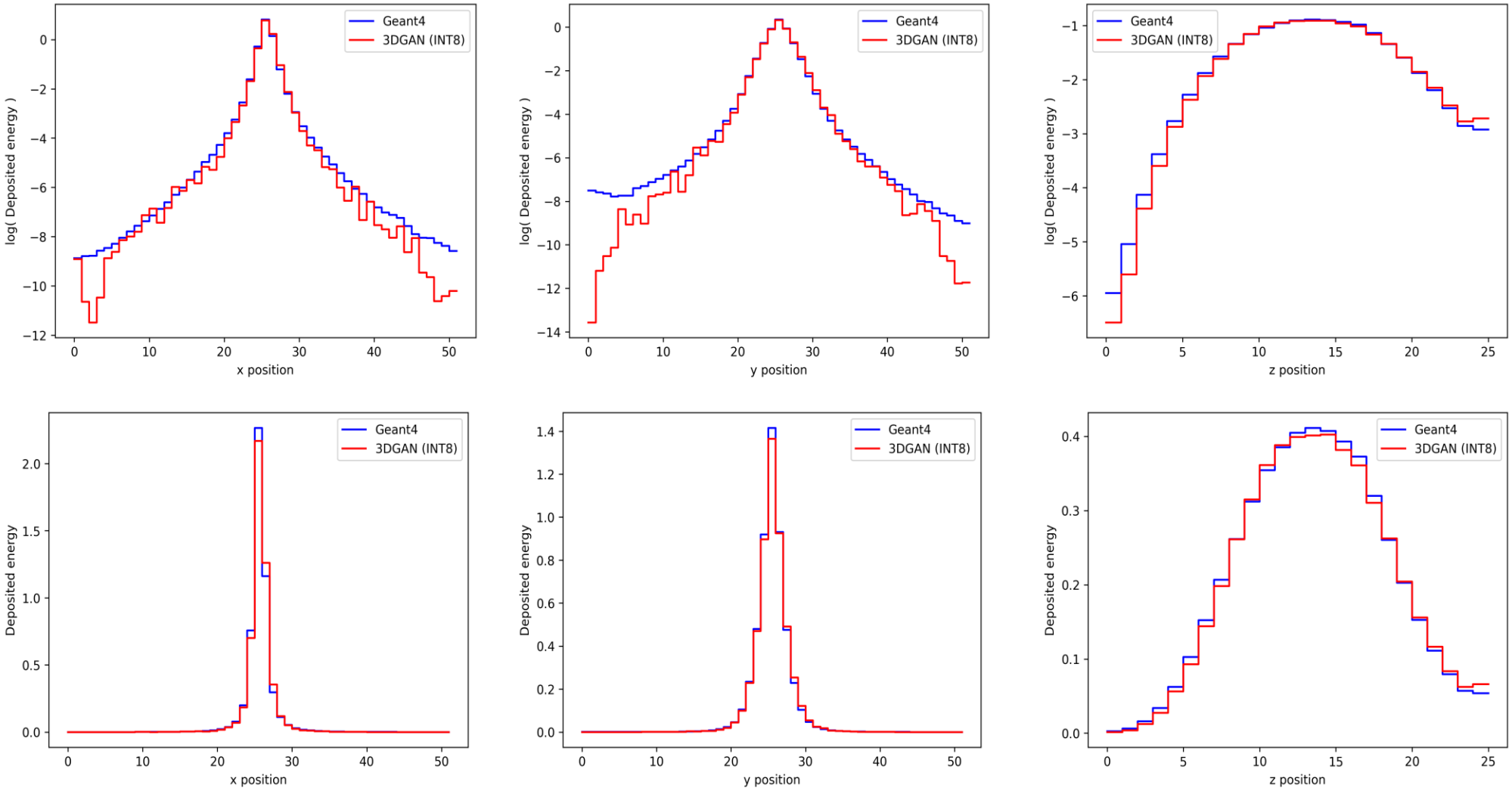
- Loss Function for INT8 Quantization

$$
\begin{aligned}
D &= \text{dataset} \\
I &= \text{prepare}(D) \\
\\
O &= \text{network}(I) \\
\mathbf{E_x} &= \text{mean}(O, \text{axis} = x) \\
\mathbf{E_y} &= \text{mean}(O, \text{axis} = y) \\
\mathbf{E_z} &= \text{mean}(O, \text{axis} = z) \\
\\
\mathbf{A_x} &= \text{mean}(D, \text{axis} = x) \\
\mathbf{A_y} &= \text{mean}(D, \text{axis} = y) \\
\mathbf{A_z} &= \text{mean}(D, \text{axis} = z) \\
\\
\text{meanError} &= \sum_{i=0}^{N}(E_{x_i} - A_{x_i})^2 + \sum_{i=0}^{N}(E_{y_i} - A_{y_i})^2 + \sum_{i=0}^{N}(E_{z_i} - A_{z_i})^2 \\
\text{varError} &= \ldots \\
\text{error} &= \text{varError} + \text{meanError}
\end{aligned}
$$

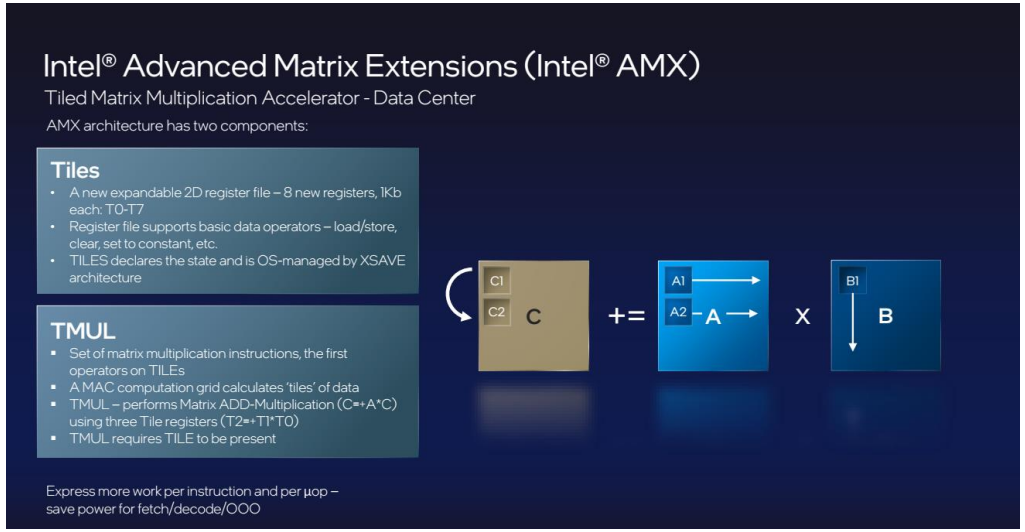
step_id	Num Quantized Layers	Loss
Baseline	**0**	**0.0011**
INC Steps		
0	7	0.1992
1	7	0.2040
2	7	0.1992
3	7	0.2040
4	6	0.1991
5	6	0.1220
6	6	0.1855
7	6	0.1952
8	6	0.1290
9	6	0.1678
10	6	0.1979
11	5	0.0356
12	4	0.0143
13	**3**	**0.0011**

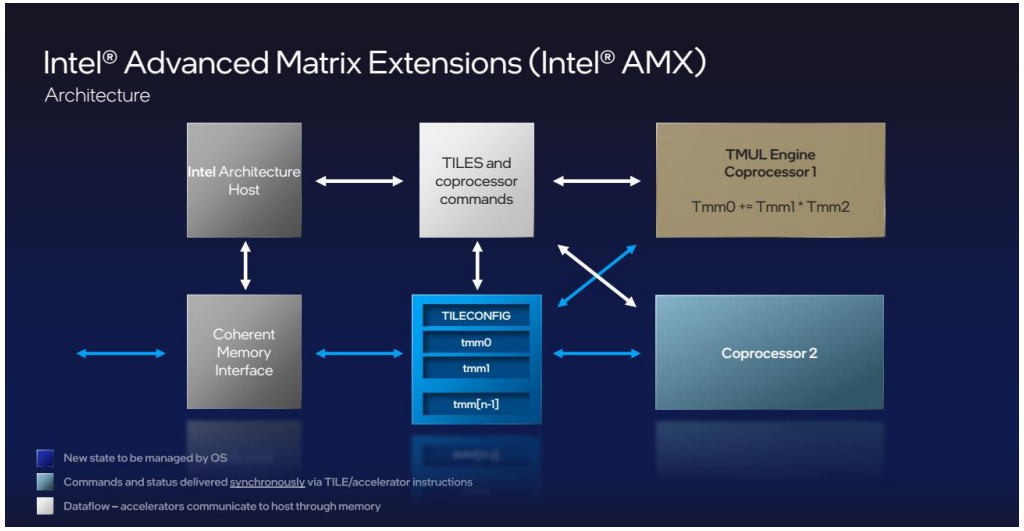


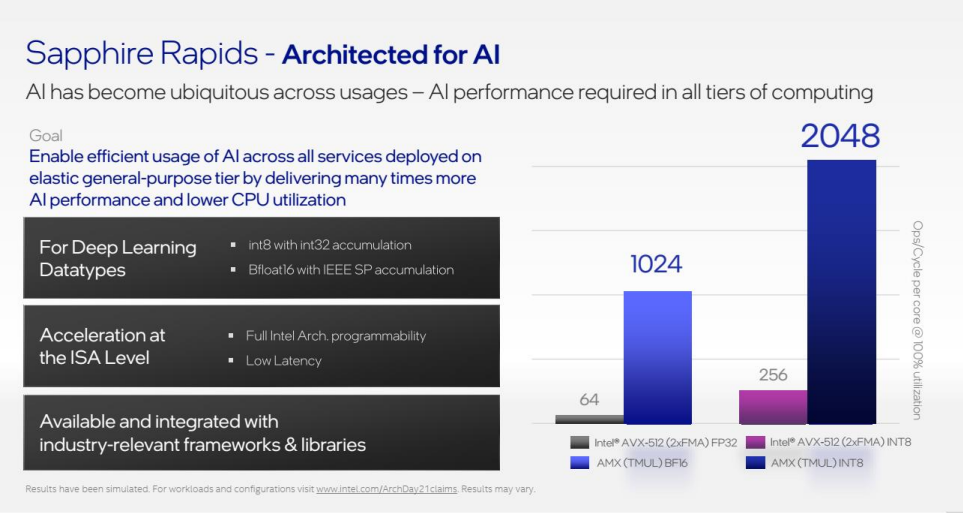
Physics Performance: Geant4 vs Quantized 3D-GAN INT8 Inference



Perf Benefits from 4th Gen Intel® Xeon®: AMX and TMUL



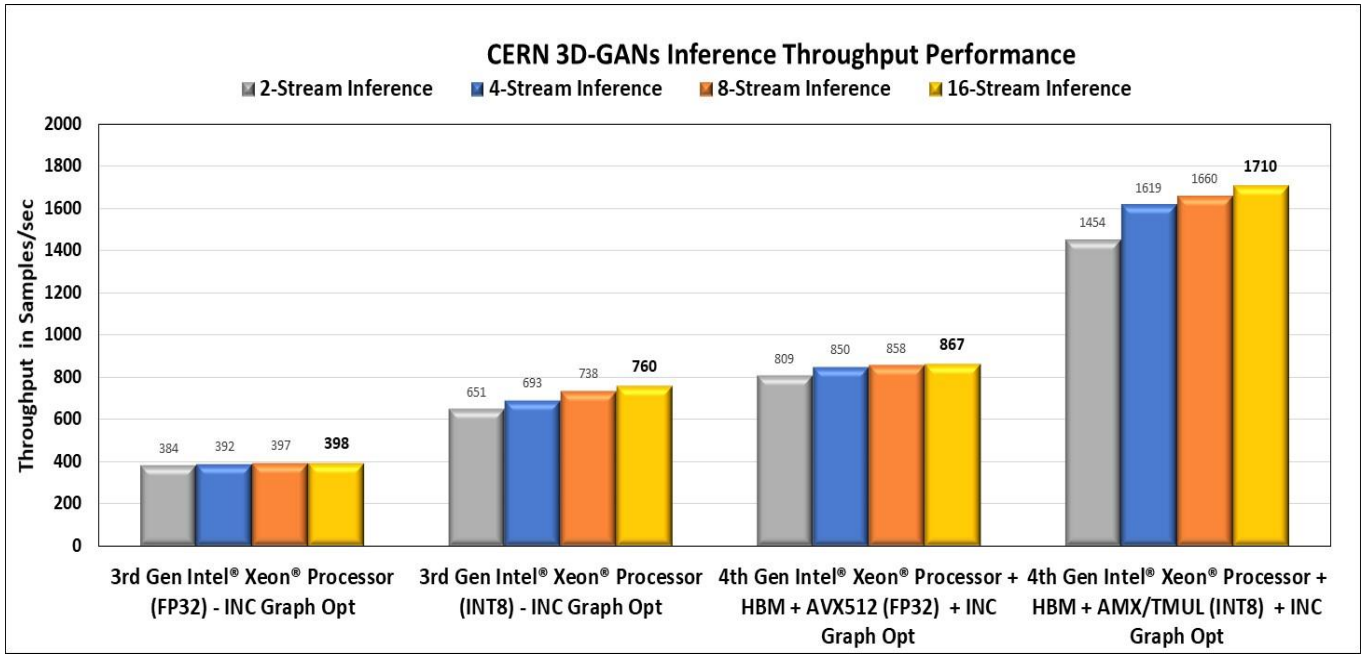


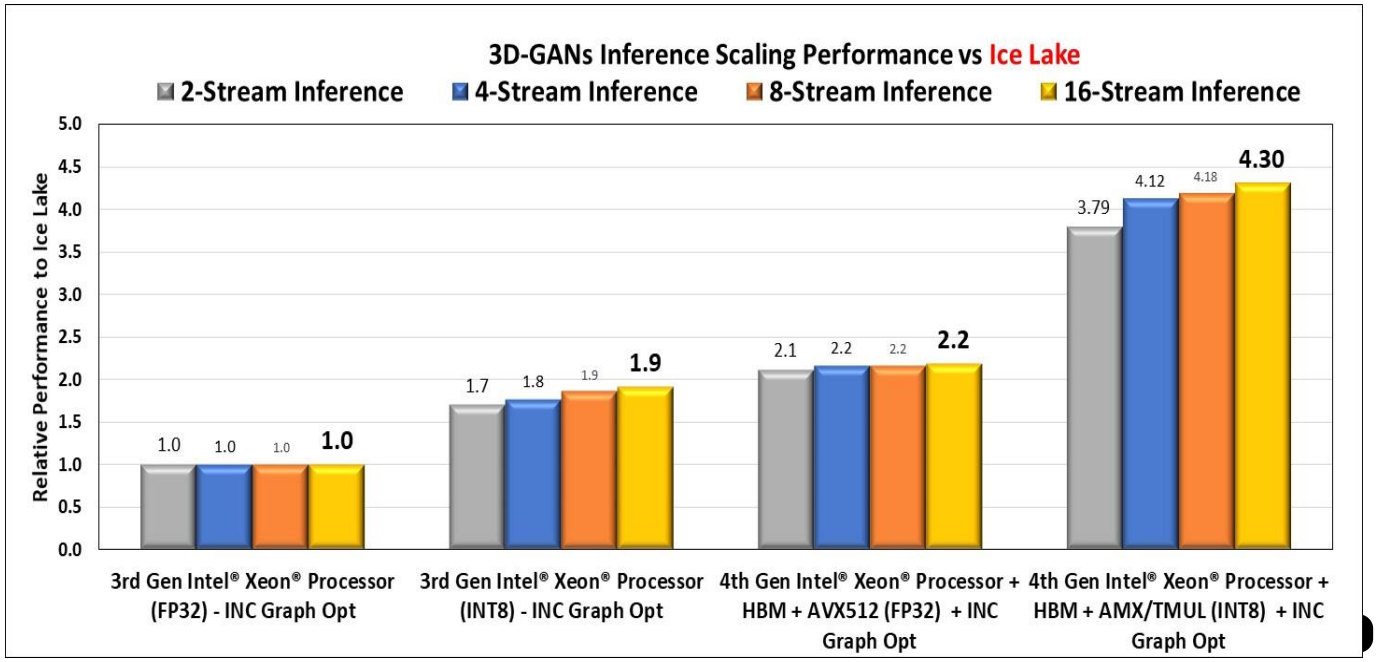


Inference Performance: Gen-Gen

- Used Intel Neural Compressor Graph Optimizations
- With Intel optimized TensorFlow 2.10
  - **Baseline**: Intel 3rd Gen Xeon – FP32
  - Intel 3rd Gen Xeon – INT8
  - Intel 4th Gen Xeon with HBM – FP32
  - Intel 4th Gen Xeon with HBM – AMX/TMUL INT8

UP TO 4.3X IMPROVEMENT

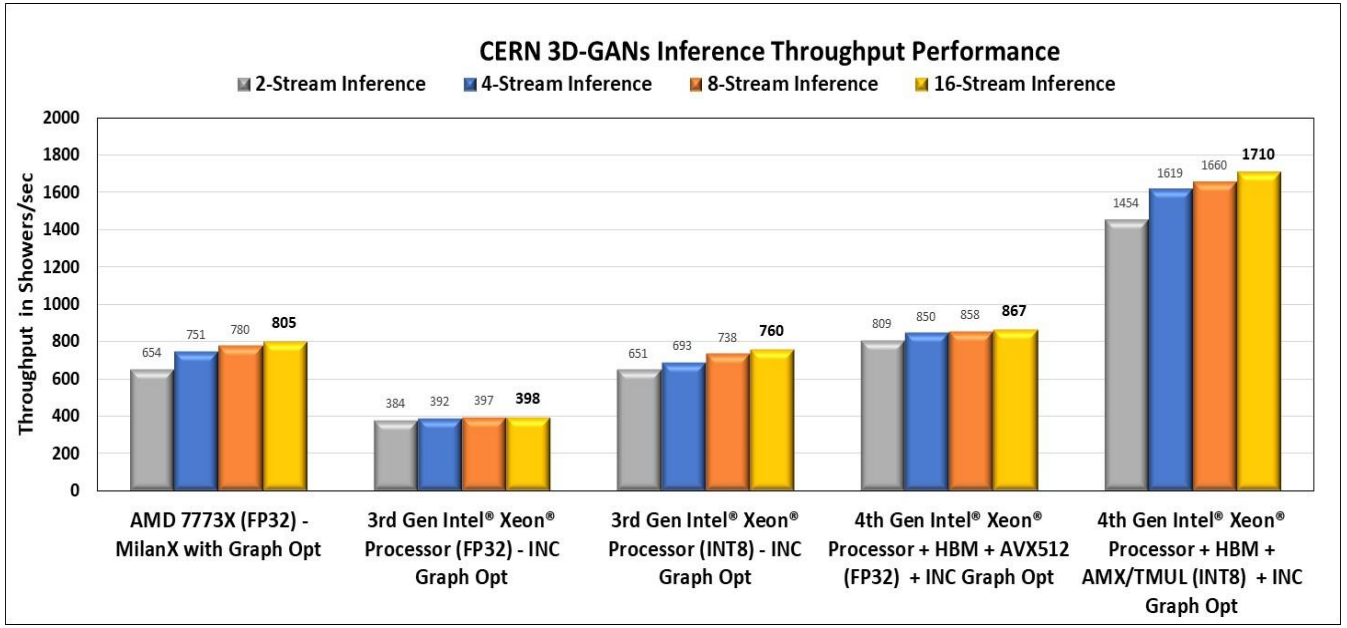


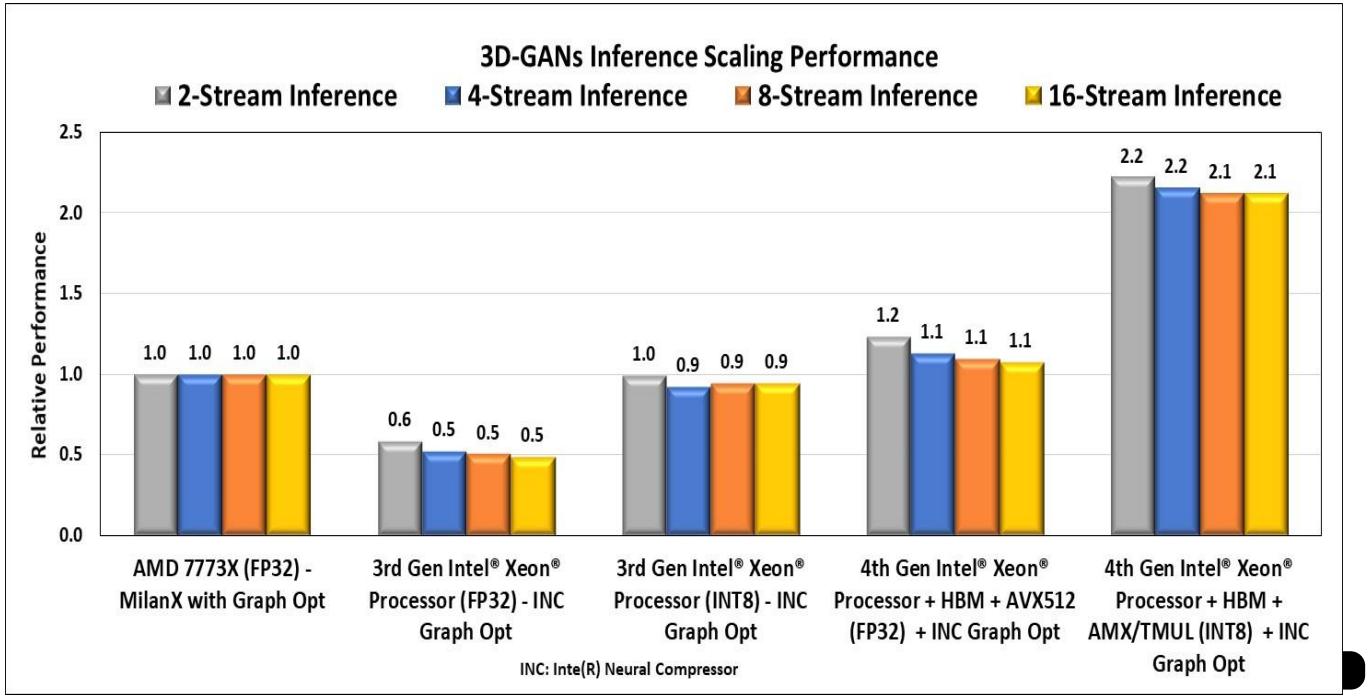


Inference Performance: Relative to Competition

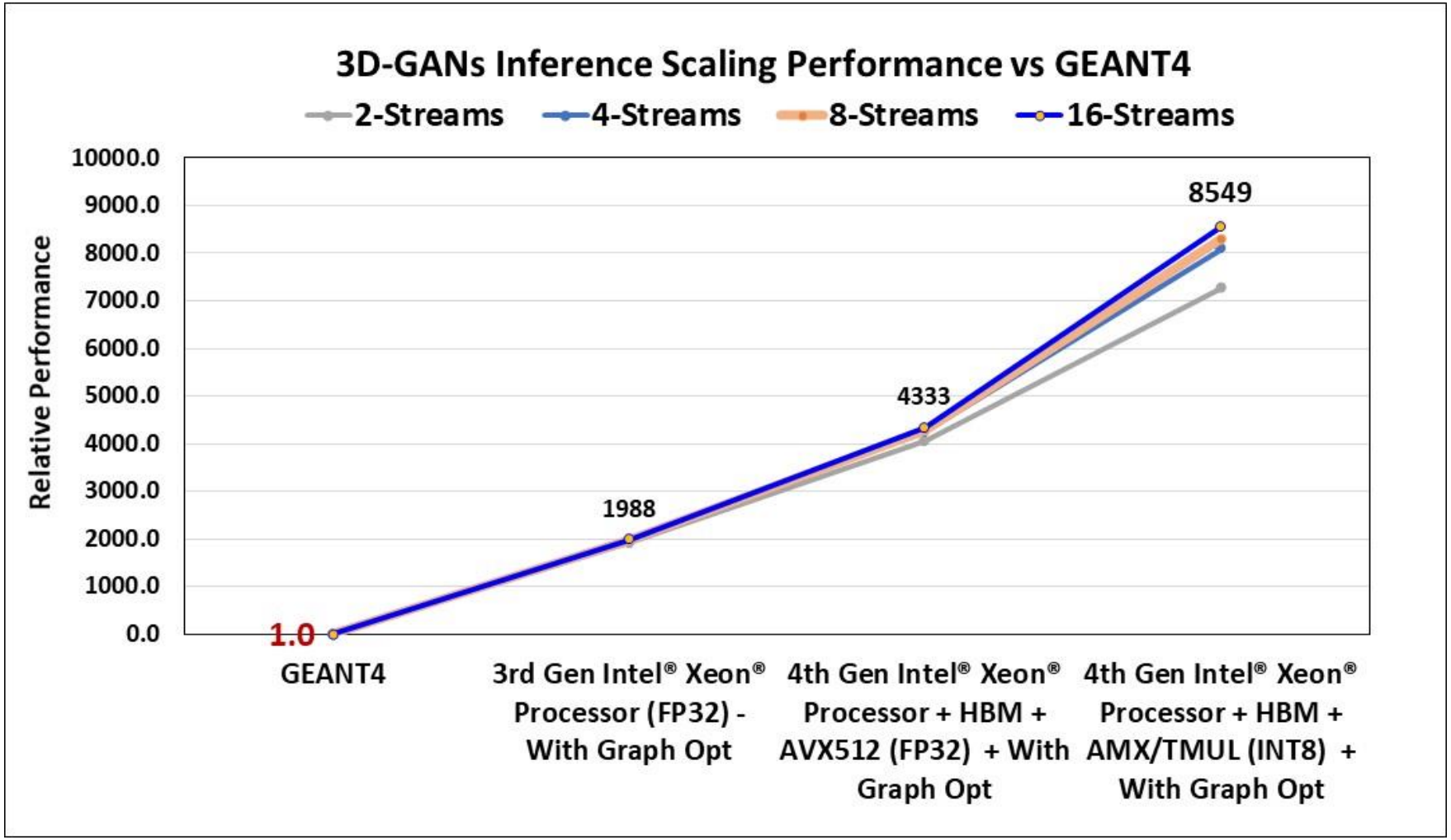
- Used Intel Neural Compressor Graph Optimization for Competitive and Intel® Xeon®
- With Intel optimized TensorFlow 2.10
  - **Baseline**: AMD Milan 7773X – FP32
  - Intel 3rd Gen Xeon – FP32
  - Intel 4th Gen Xeon with HBM – FP32
  - Intel 4th Gen Xeon with HBM – AMX/TMUL INT8

UP TO 2.1X IMPROVEMENT





Inference Performance Relative to Geant4



4TH GEN INTEL® XEON® ACHIEVES > 8500X IMPROVEMENT OVER GEANT4

Conclusions

- Results on 3D-GANs optimization on Intel® Xeon® are very promising
- Intel Neural Compressor Tool enables quantize FP32 model to INT8 without much effort
- Showed that INT8 quantized Model meets **FP32 accuracy** and also **Physics Accuracy**
- **Demonstrated that 4th Gen Intel® Xeon® accelerates > 8500X Improvement over CERN's Geant4 simulation**

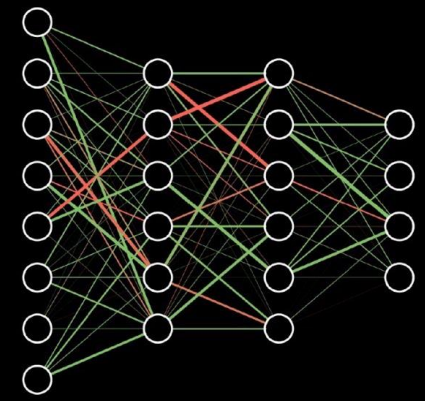
DeepGalaxy: Deducing the Physical Properties of Galaxy Mergers using DNNs

Maxwell X. Cai
Jeroen Bédorf, Vikram Saletore, Valeriu Codreanu, Damian Podareanu,
Adel Chaibi, Penny X. Qian

Simulation

Observation

Neural Networks

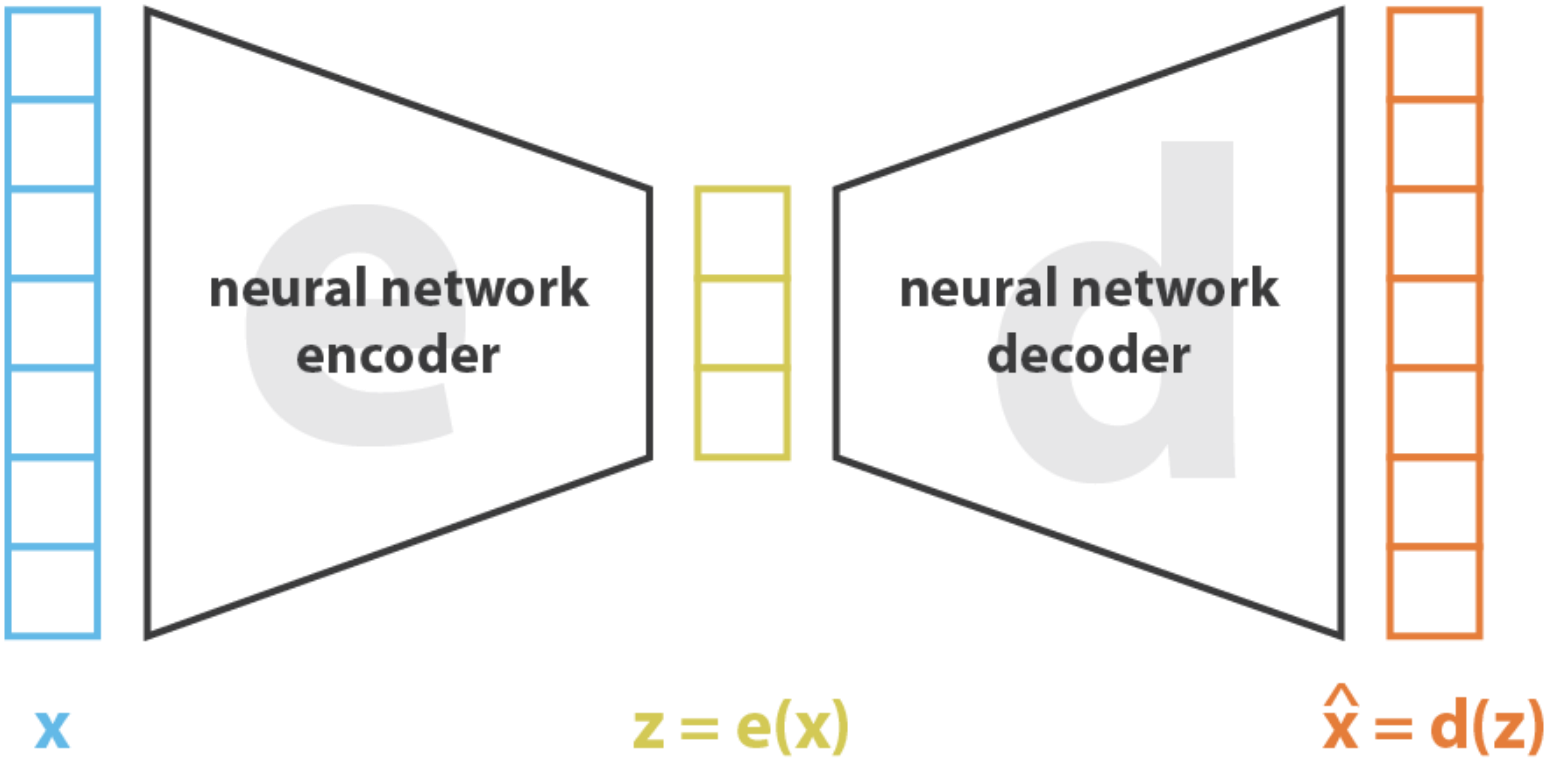


Train

Predict

Astrophysical problem → Pattern recognition problem

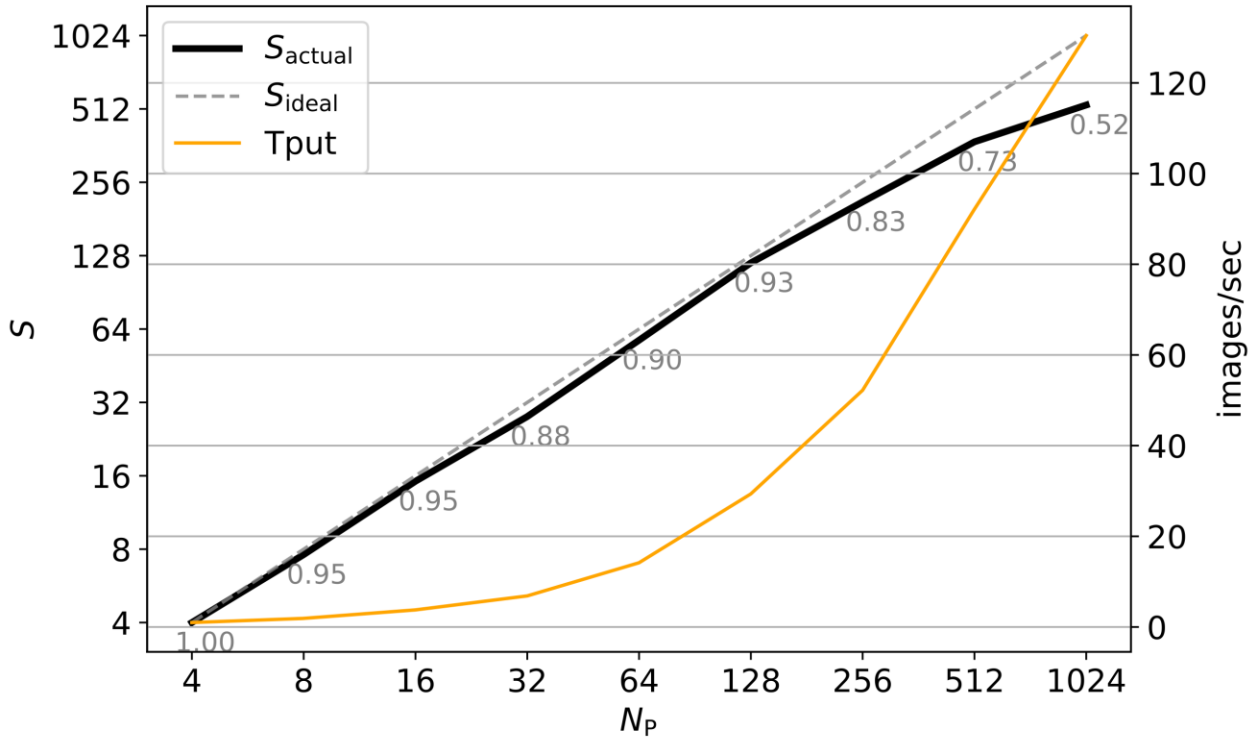
Variational Auto Encoders (VAE)



$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

DeepGalaxy HPC Results

Resulting Speedup over simulations ~ 100,000 X



- Bridges the gap between observational and theoretical astrophysics
- Processing Images and running simulations take about ~ 2 Days
- Replacing simulations with deep learning Model takes < 2 seconds

- **1024x1024 res**
- **29K Images; 192GB**
- **256 Nodes (4W/node)**

Thanks to System Team for Intel Endeavor Cluster

See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to https://software.intel.com/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

Thank You

intel®