CERN openlab

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

intel

Mixed-Precision Arithmetic for 3DGAN to Simulate High Energy Physics Detectors

John Osorio, Adrià Armejach, Gulrukh Khattak, Eric Petit, Sofia Vallecorsa, Marc Casas

13/10/2020

2020 IXPUG Annual Meeting

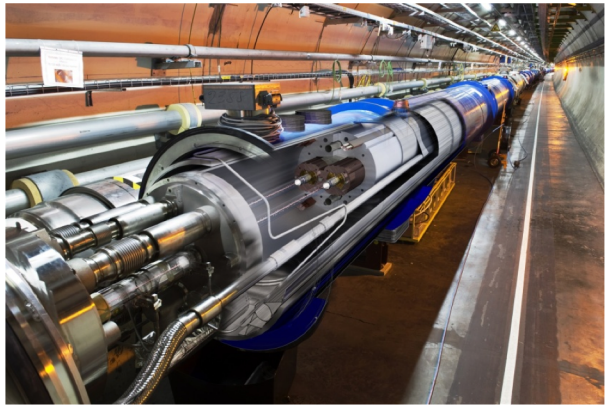
Overview

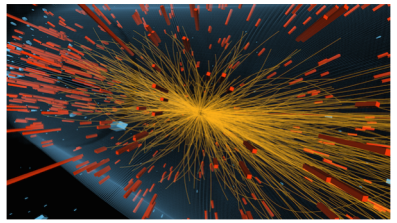
- Motivation
  - High Energy Physics (HEP)
  - Deep Neural Networks (DNN)
  - Mixed Precision Training (MP)
- Proposal
  - Emulation Pintool
  - Hardware+Software
- Results
- Future Plans

Motivation

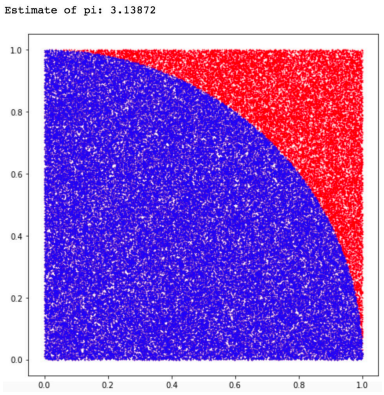
High Energy Physics Simulation

- Important
  - Essential for data analysis and detector design

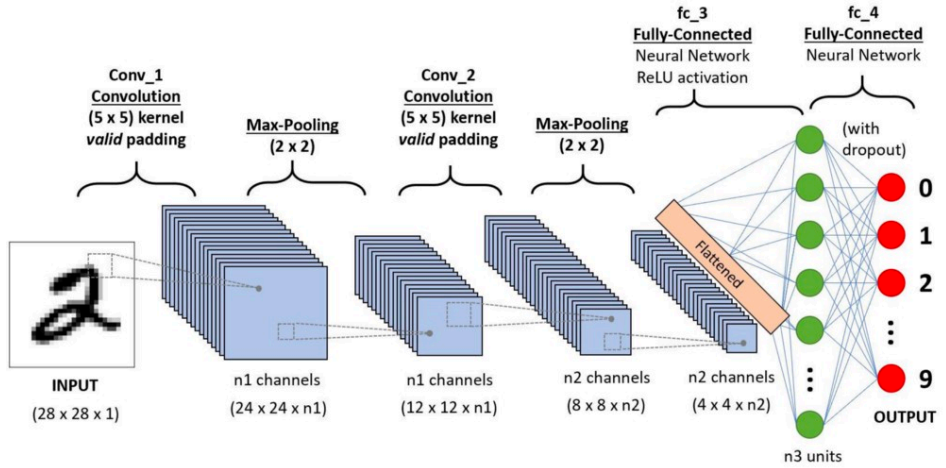




- Monte Carlo Simulation
  - Computationally expensive



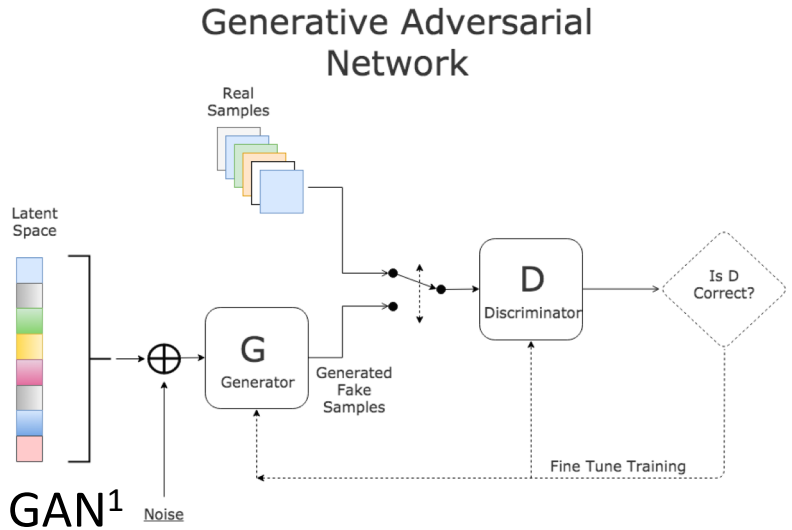
- DNN
  - Lots of data to train
  - Efficient Inference processes



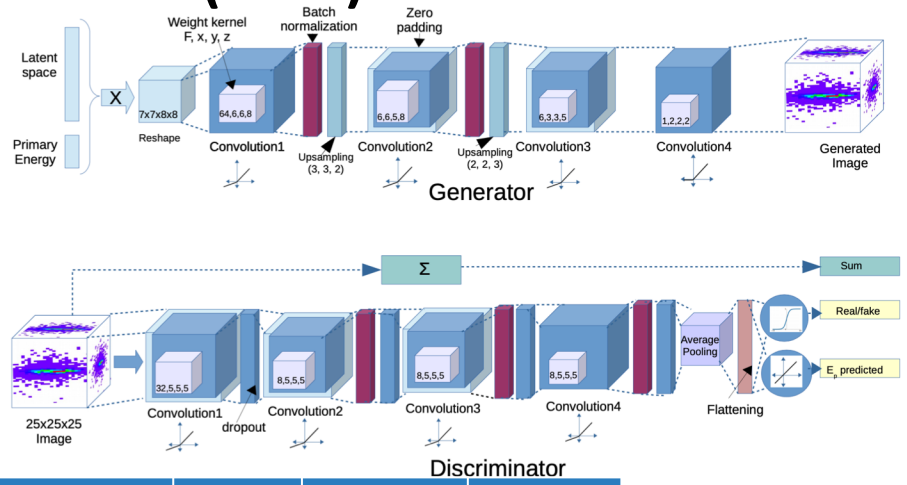
Motivation

Deep Neural Networks (DNN)

- Generative Adversarial Networks (GAN)



- HEP + GAN[1]
  - Training Process is expensive.
    - Train once and inference a lot
  - The 3DGAN approach generates the output with Speed-ups up-to 20000x compared to MC simulation.



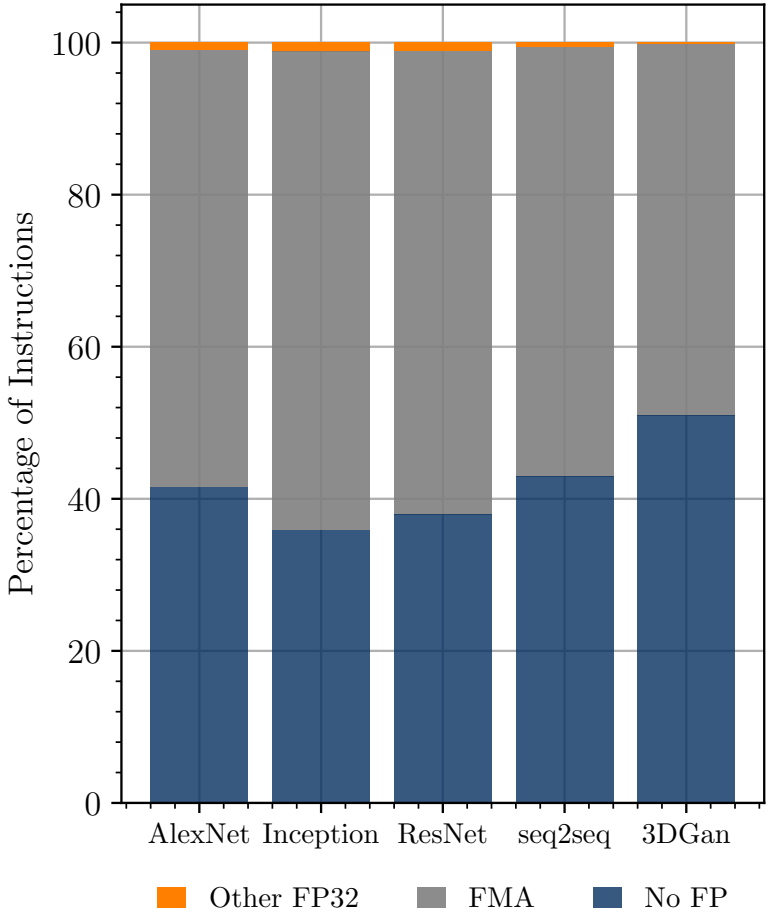
Method	Platform	Time/Shower (ms)	Speedup
Classical Monte Carlo (Geant4)	2S Intel Xeon Platinum 8180	17000	1.0
3DGAN (BS=128) 1-stream	2S Intel Xeon Platinum 8160	1.25	13600
3DGAN (BS=128) 2-stream		0.93	18279
3DGAN (BS=128) 4-stream		0.85	20000

1. Particle Detector Simulation

Motivation

Mixed Precision Training

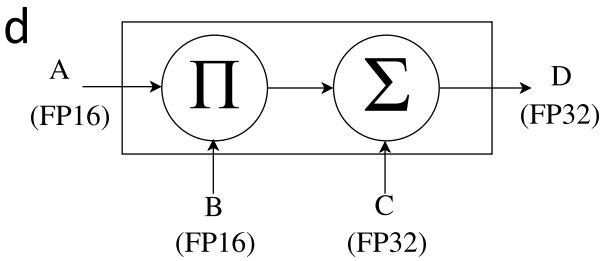
- Instruction Breakdown, a training batch:
  - AlexNet:
    - FMA – 57.42%
  - Inception V2
    - FMA - 60.93%
  - ResNet-50
    - FMA – 62.95%
  - *Seq2seq*
    - FMA - 56.44%
  - 3DGan
    - FMA - 48.80%
  - FMA instructions used in WU and BN
    - < 1.60 %

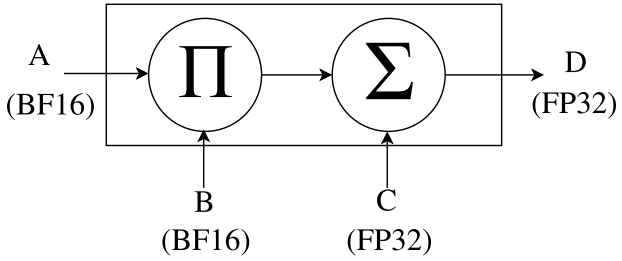


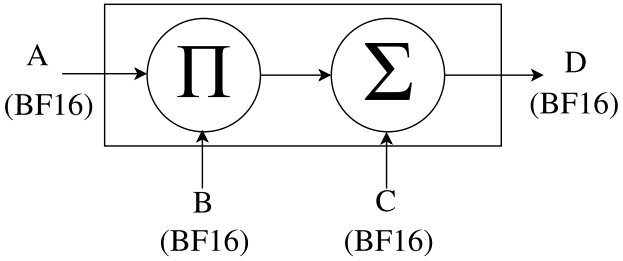
Motivation

Mixed Precision Training

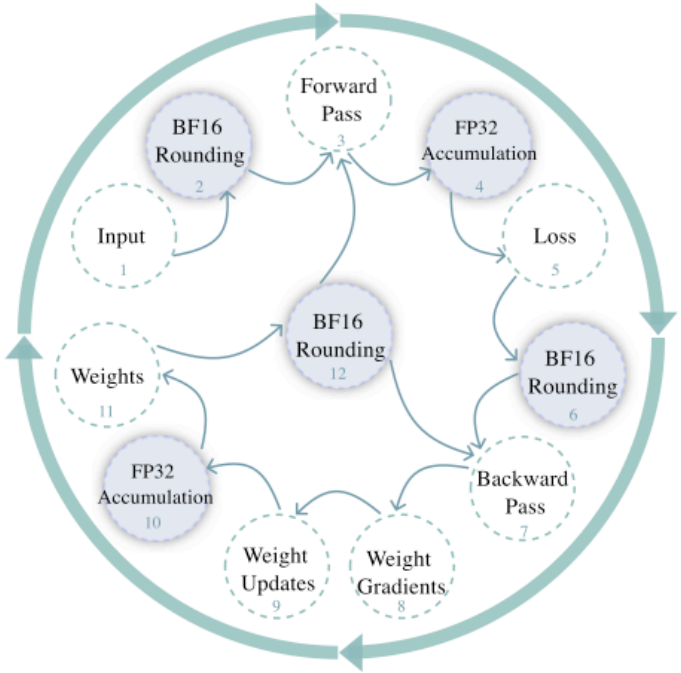
- Fused Multiply Add (FMA)



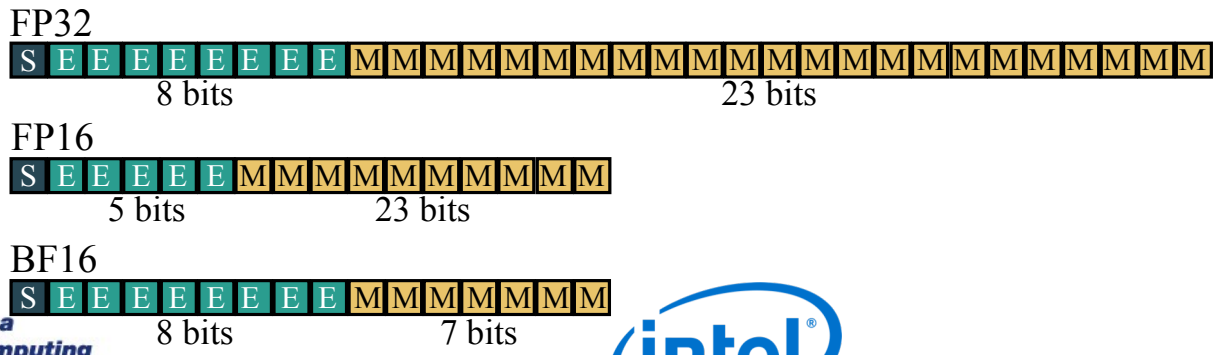




Training Process using MP



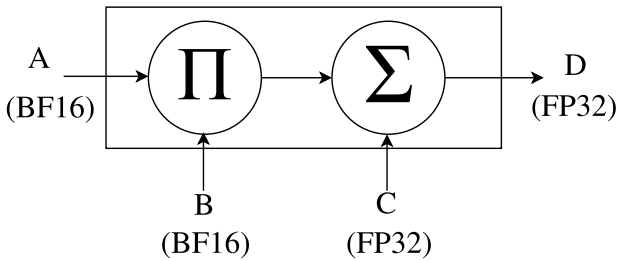
Numerical datatypes typically used to train DNN

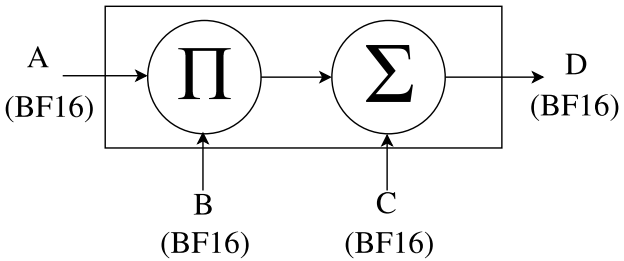


Proposal

Emulation PinTool

- The Binary Analysis Tool (Pintool) detailed:
  - Our PinTool implement a static scheme for the 3DGAN training.
    1. It checks the current FMA operation mode, which could be FP32, MP, or BF16.
    2. Routines that belong to WU calculations or BN layers, are calculated in FP32.
    3. For each FMA instruction, the binary analysis tool rounds off all operands that need to be converted to BF16 using a round to nearest algorithm.
  - The PinTool have support for dynamic precision approaches
    - BF16<->FP32
    - BF16<->MP





Proposal

Hardware+Software Used

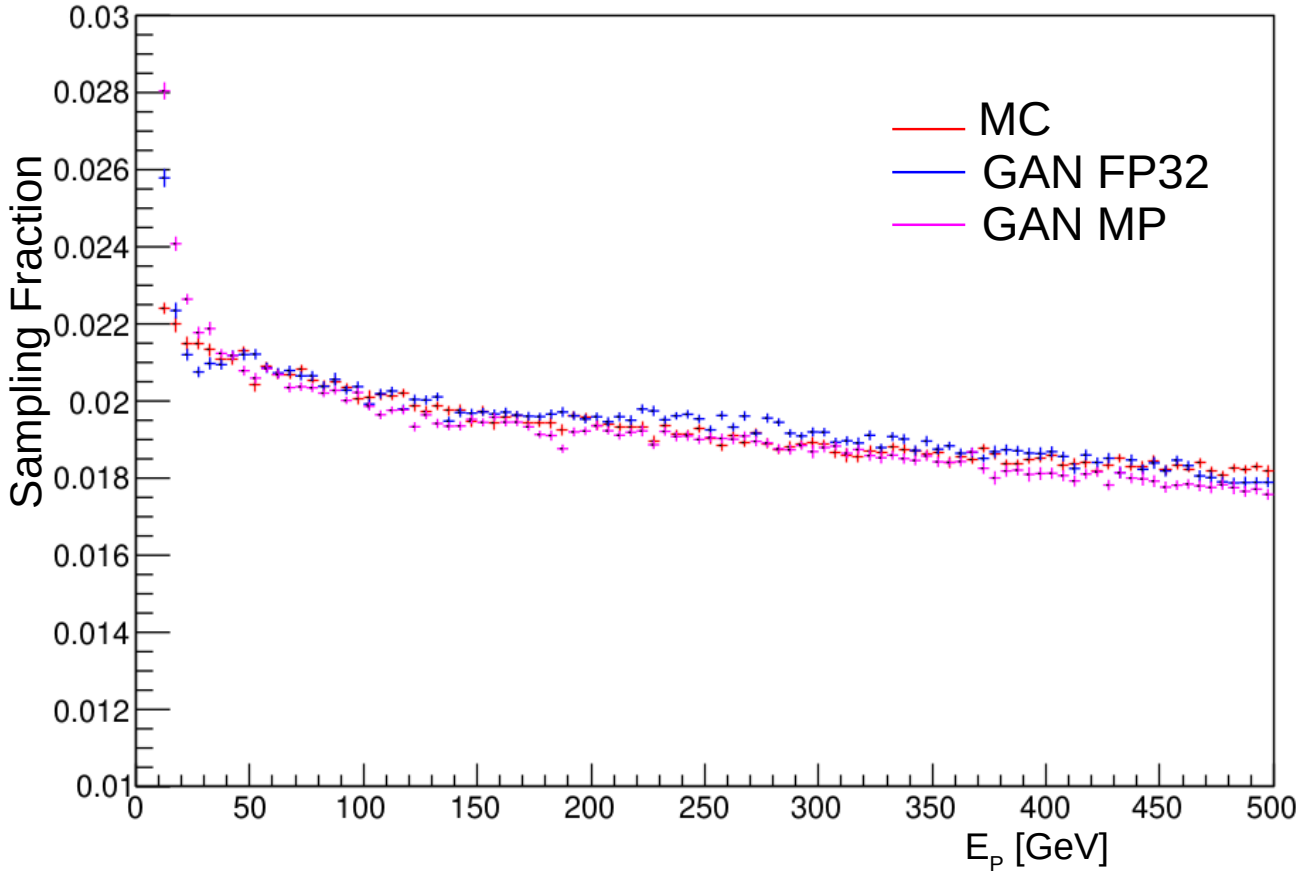
- Hardware used to run the tests

- Software used
  - Intel PIN 3.7
  - 3DGan – TensorFlow
  - The PinTool supports PyTorch and Caffe
  - Intel MKLDNN – DNN Primitives

Results

- Sampling Fraction

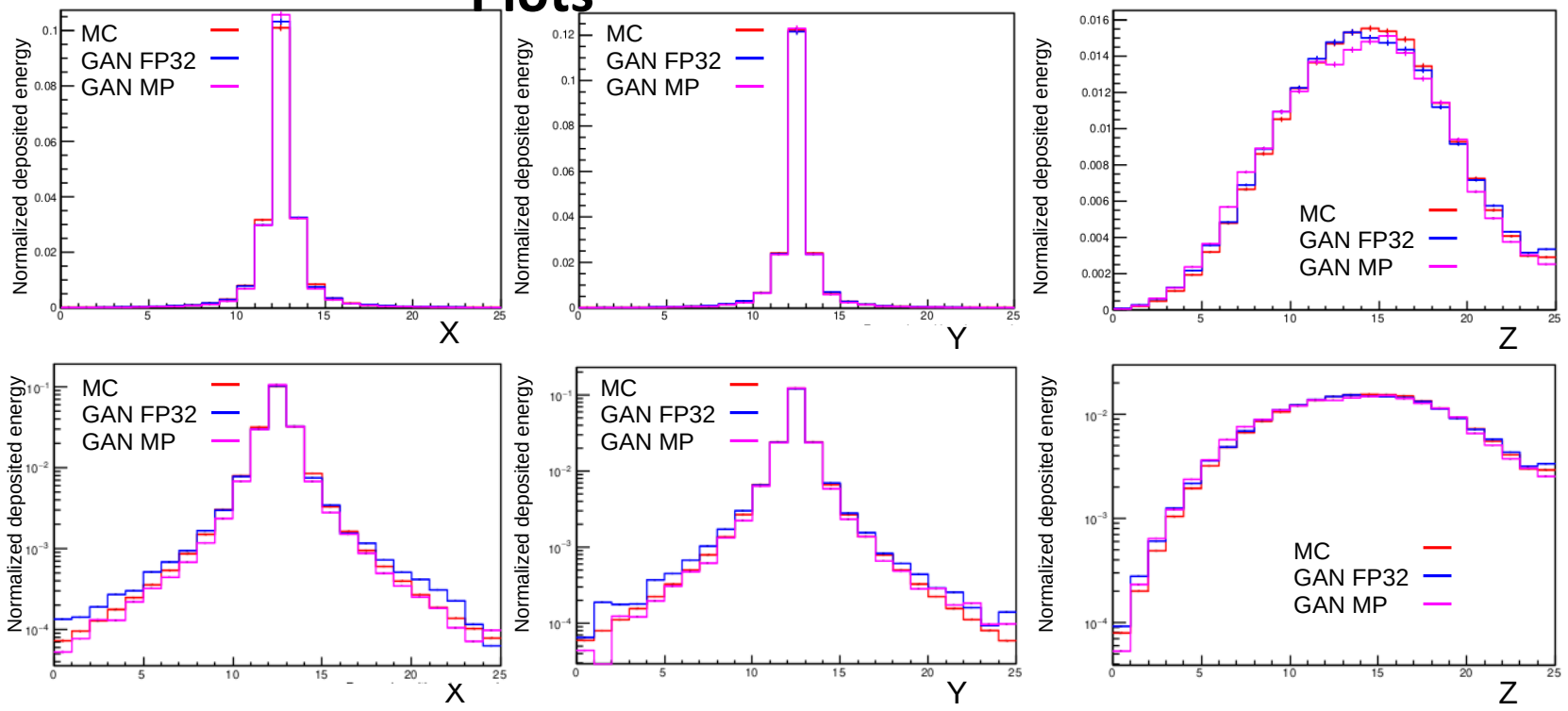
Plots



Results

- Shower Shapes

Plots



Future Plans

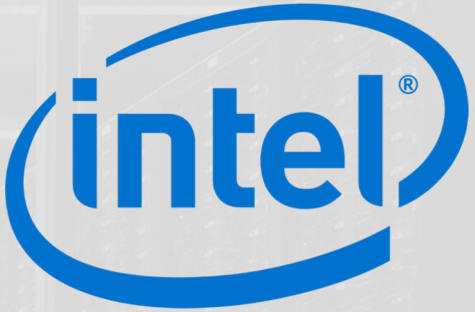
- PinTool
  - Enable using other datatypes:
    - FP8, FP4
    - FP16
    - Evaluate even with less bits
  - Evaluate the behavior of our approach using a Computer Architecture Simulator (Sniper – Pin Based)
- 3DGAN
  - Enable the training using full BF16 FMAs
  - Enable the training using a dynamic precision approach (BF16<->FP32, BF16<->MP)

CERN openlab

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

intel

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Thanks

john.osorio@bsc.es