Heterogeneous Computing (HGC) for Deep Learning

[Chao Jiang, Nirali Patel, Chandana Prasad, Shreyas Sreedhara, Akash Vasishta
David Ojika, Herman Lam] SHREC@UF*, Bhavesh Patel[bhavesh_a_patel@dell.com]

Motivation

- AI and deep learning are experiencing explosive growth in domains involving analysis of big data
  - Deep learning using ***Deep Neural Networks (DNNs)*** has shown great promise for such data analysis applications
- Heterogeneous computing, with ***CPUs*** integrated with accelerators such as ***GPUs*** and ***FPGAs***, offers unique capabilities to accelerate DNNs

Overview & Project Goals

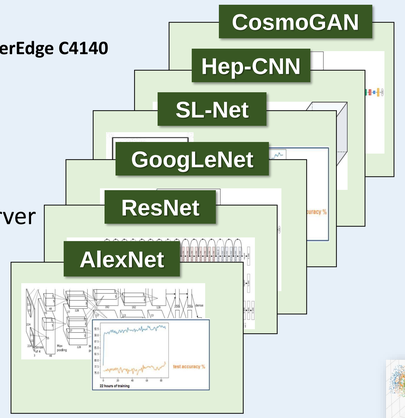
- Researchers at SHREC@UF* are developing such an HGC system to support a complete HGC workflow for deep learning
  - **Data Analysis & Pre-processing**
  - **Model Training**
  - **Deployment & Inferencing**
- This project focuses on the use of Intel ***FPGAs to accelerate*** the ***inferencing stage*** of the HGC workflow
  - Started at beginning of 2018 and is ongoing at the SHREC* Center

Collaboration

* SHREC@UF: NSF Center for Space, High-Performance, and Resilient Computing, at the University of Florida (UF)
in collaboration with:
- Dell EMC
- Intel
- NERSC @ Lawrence Berkeley Lab

Model Training

- ***Explore*** & *use existing frameworks* for GPU training
- ***Produce*** *trained DNN models* from CERN openlab data



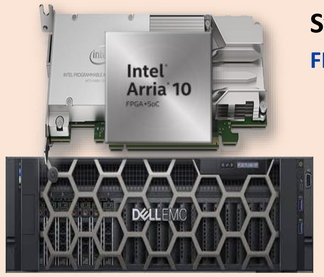
Hardware
- Dell PowerEdge C4140 server
- Nvidia Tesla V100 GPU

Training tools
- Tensorflow 1.7
- Keras 2.2
- Cuda 9.2

Nvidia Tesla V100 GPU

Deployment & Inferencing

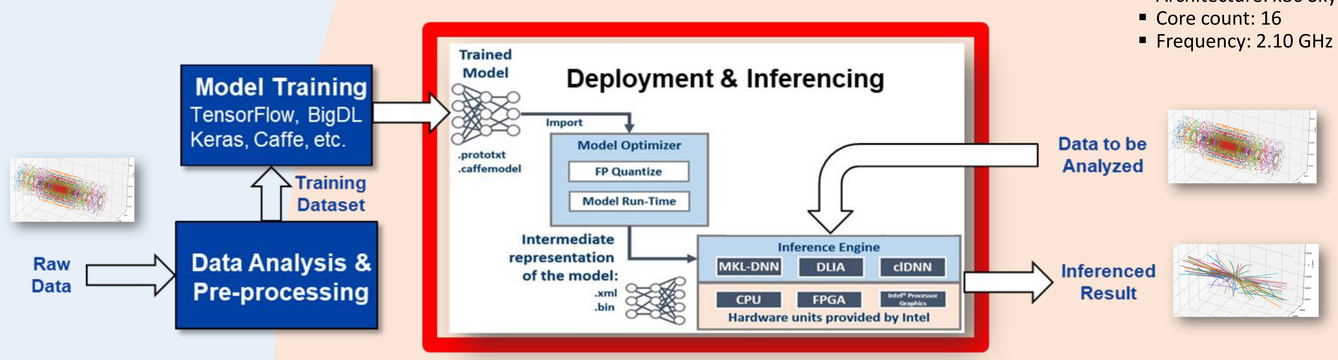
- ***Explore*** & ***select*** frameworks for *FPGA inferencing*
- *Intel OpenVINO*
  - Consists of *Model Optimizer* and *Inference Engine*
  - Enables Computer Vision workloads across *Intel* hardware
  - Exposes a *common API* across *CPU* and *FPGA*
  - Supports *library of CV functions* and *pre-optimized kernels* for ease of use and rapid app development

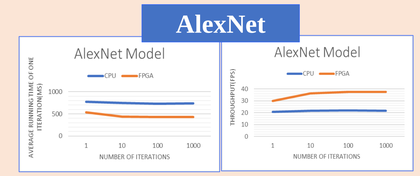


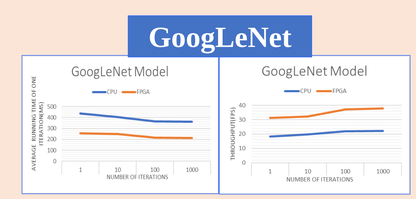
System specification

FPGA spec
- SKU: Intel PAC Arria-10 GX
- I/O: 96 full-duplex transceivers
- Data rate: 17.4 GBps chip to chip

CPU spec
- SKU: Intel Xeon Gold 6130
- Architecture: x86 Skylake
- Core count: 16
- Frequency: 2.10 GHz







Data Analysis & Pre-processing

- ***Explore*** mission-critical (e.g., CERN openlab) *datasets* ***&*** *analysis tools*
- Develop methods ***to*** *pre-process raw data* ***into*** *image form for training*

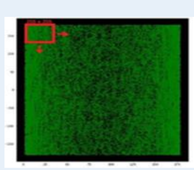
TrackML Challenge* Dataset

hit_id	x	y	z	volume_id	layer_id	module_id
1	-64.4099	-7.1637	-1502.5	7	2	1
2	-55.3361	0.635342	-1502.5	7	2	1
3	-83.8305	-1.14301	-1502.5	7	2	1
4	-96.1091	-8.24103	-1502.5	7	2	1
5	-62.6736	-9.3712	-1502.5	7	2	1

- *Simulated dataset* of real *HEP*[1] experiments
- Five sets of *1770 events* each
- Total of *8850 events (400 GB data)*

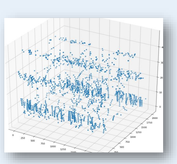
[1] High Energy Physics
* TrackML Challenge - https://www.kaggle.com/c/trackml-particle-identification

Pre-processing



Sliding window for data pre-processing

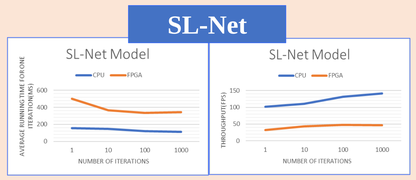
Based on SL-Net[2], implemented method to pre-process data into *image form*
- *Transposed x, y and z co-ordinates to spherical r, theta, and phi*
- Applied sliding window of 255 x 255 x 48
  - Where 48 is the hits on each sensing layer of detector
- Finally, two files were generated
  - Input data to the training stage and
  - Their respective labels

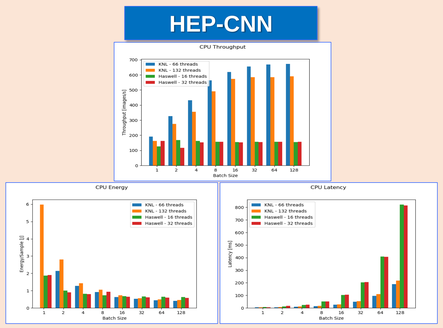


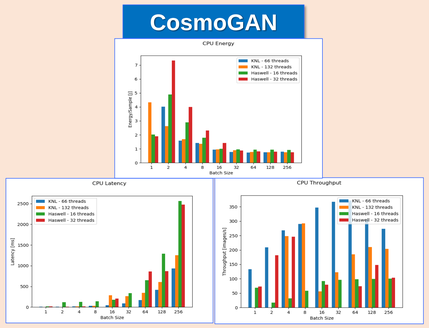
[2] SL-Net: D.Dmitriev. Computer Vision for TrackML, https://www.kaggle.com/denisdmitriev/computer-vision-for-trackml; customized to our needs

Early results & observations

- *Compared against Intel Skylake* CPU (single core)
- SL-Net: Intel Xeon *CPU* performs better than *FPGA*
- For GoogLeNet, ResNet, AlexNet, *FPGA performs better as complexity increases:*
  - *ResNet – 2.22x*
  - *GoogLeNet – 1.72x*
  - *AlexNet – 1.45x*







Conclusions & Going Forward

- Introduced & demonstrated a complete heterogeneous computing (HGC) workflow for deep learning
- Special focus on *FPGA-based acceleration* of inference
- Going forward
  - *Improve & optimize* OpenVino inferencing using Intel Deep Learning Acceleration (DLA) development suite
  - New deep-learning models of interest: e.g., *2D/3D GAN, 3D UNet*

Winner
DELL EMC AI Challenge
Push the boundARIes.