

Research Opportunities in Heterogeneous Computing for Machine Learning

Herman Lam
Site Director, NSF SHREC* Center
University of Florida
Gainesville, Florida, USA
hlam@ufl.edu

David Ojika
Dell EMC
University of Florida
Gainesville, Florida, USA
david_ojika@dell.com

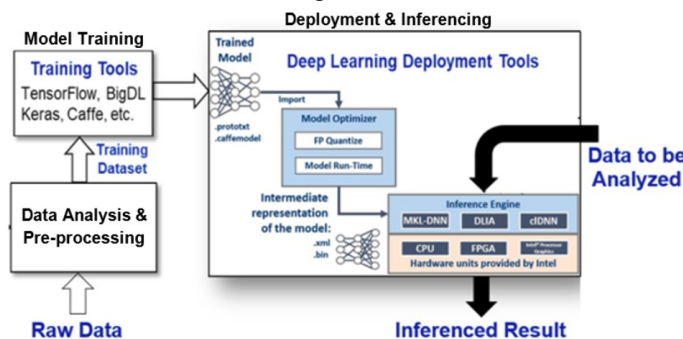
INVITED TALK EXTENDED ABSTRACT

A. Introduction

In recent times, AI and deep learning have witnessed explosive growth in almost every subject involving data. Complex data analyses problems that took prolonged periods, or required laborious manual effort, are now being tackled through AI and deep-learning techniques with unprecedented accuracy. Machine learning (ML) using Convolutional Neural Networks (CNNs) has shown great promise for such applications. However, traditional CPU-based sequential computing no longer can meet the requirements of mission-critical applications which are compute-intensive and require low latency. Heterogeneous computing (HGC), with CPUs integrated with accelerators such as GPUs and FPGAs, offers unique capabilities to accelerate CNNs. In this presentation, we will focus on using FPGA-based reconfigurable computing to accelerate various aspects of CNN. We will begin with the current state of the art in using FPGAs for CNN acceleration, followed by the related R&D activities (outlined below) in the SHREC* Center at the University of Florida, based on which we will discuss the opportunities in heterogeneous computing for machine learning.

B. R&D Activities @ University of Florida

In the SHREC* Center at the University of Florida, the R&D activities that explore the use of FPGAs in accelerating CNN are based on the concept diagram shown in Figure 1, which illustrates the general heterogeneous computing (HGC) workflow for machine learning.



The HGC workflow consists of three main stages: Data Analysis & Pre-processing, Model Training, and Deployment & Inferencing.

In the Data Analysis & Pre-processing stage, the input is the raw data collected/sensed from a mission under study. The function of this stage is to pre-process the data into a form that is usable by the Model Training Tools. For our current study, the raw data to be analyzed and pre-processed are two datasets collected in the Large Hadron Collider at CERN and made available to researchers: the ATLAS [1] dataset form CERN Openlab [2] and the TrackML Challenge dataset [3]. A key goal for this stage in SHREC* is to develop methods and tools to pre-process the enormous amount of CERN data in a form ready for model training.

The input to the Model Training stage is the pre-processed data in the form of a training dataset. The output is the trained models to be deployed for use in inference engines. We have explored the state-of-the-art training tools such as TensorFlow [4], BigDL [4], Keras [6], and Caffe [7]. A key goal for this stage in SHREC* is to explore ways to perform incremental learning on trained models to improve classification and to adapt to a changing operating environment.

For the Deployment & Inferencing stage, the two inputs are: the trained model from the Model Training stage and the data to be classified. The output is the inferred classification. We have explored and selected the use of the newly released OpenVINO Toolkit [8] from Intel Corporation. Within this toolkit is the Deep Learning Deployment Tools (as shown in Figure 1), which consists of the Model Optimizer and the Inference Engine. The Model Optimizer facilitates the transition between the training and deployment environment. It performs static model analysis and adjusts the deep learning models for optimal execution on a target inference engine. In OpenVINO, the target inference engine can be a CPU, GPU, FPGA, or some combination (HETERO). Thus, once the target inference engine has been selected, the trained model is translated into an intermediate representation suitable for the selected inference engine. The output of the inference engine is

* SHREC: NSF Center for Space, High-Performance, and Resilient Computing, formally CHREC (NSF Center for High-Performance Reconfigurable Computing)

a probability-based classification. A key goal for the Deployment & Inferencing stage in SHREC* is to explore the cost and benefits of using the FPGAs to accelerate the CNN inferencing process.

C. Collaboration

The R&D activities to be discussed in this presentation are in progress and made possible through collaboration among the partners shown in the following table. The University of Florida’s SHREC* Center is the research lead and provides the FPGA expertise. NERSC** from Lawrence Berkeley National Lab provides the machine learning expertise. The datasets being used are supplied by CERN. Dell and Intel provide the necessary computing equipment and tools.

Project Contributor	Role
Univ. of Florida	FPGA expertise (research lead)
NERSC	ML expertise
CERN OpenLab	Data source
Intel	AI hardware/tools
Dell	Solutions provider

learning; convolutional neural network

REFERENCES

- [1] ATLAS Open Data Platform, CERN, <https://atlas.cern/tags/open-data>.
- [2] CERN Openlab, <https://openlab.cern>.
- [3] TrackML: TrackML Particle Tracking Challenge, CERN, <https://sites.google.com/site/trackmlparticle>.
- [4] TensorFlow Open Source Machine Learning Framework, <https://www.tensorflow.org>.
- [5] BigDL: Distributed Deep Learning Library for Apache Spark, <https://github.com/intel-analytics/BigDL>.
- [6] Keras: The Python Deep Learning library, <https://keras.io>.
- [7] Caffe Deep Learning Framework by Berkeley AI Research, <http://caffe.berkeleyvision.org>.
- [8] OpenVINO Toolkit, Intel Corporation, <https://software.intel.com/en-us/openvino-toolkit>.

Acknowledgments

This research was funded by industry and government members of the NSF SHREC Center and the National Science Foundation (NSF) through its IUCRC Program under Grant No. CNS-1738420.

** NERSC: National Energy Research Scientific Computing Center, Lawrence Berkeley National Lab



Research Opportunities in Heterogeneous Computing for Machine Learning



Mission-Critical Computing

NSF CENTER FOR SPACE, HIGH-PERFORMANCE,
AND RESILIENT COMPUTING (SHREC)*



University of
Pittsburgh

BYU
BRIGHAM YOUNG
UNIVERSITY

Virginia
Tech

UF
UNIVERSITY OF
FLORIDA

Herman Lam

Site Director,
NSF SHREC* Center
University of Florida
Gainesville, Florida, USA
hlam@ufl.edu

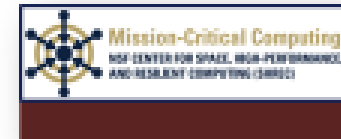
David Ojika

Dell EMC
University of Florida
Gainesville, Florida, USA
david_ojika@dell.com



Agenda

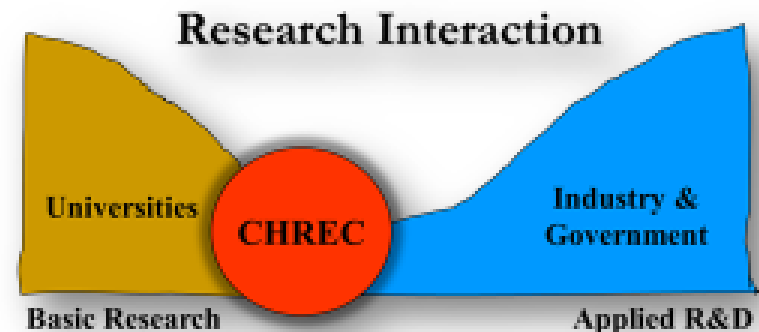
- What is CHREC/SHREC?
- Heterogeneous Computing (HGC) for Machine Learning (ML)
 - State of the art in using FPGAs for machine learning
- R&D activities in SHREC Center
 - Opportunities in HGC for ML
- Q & A



What is CHREC?



- NSF **C**enter for **H**igh-Performance **R**econfigurable **C**omputing
 - Operational since January **2007**
- Under auspices of **I/UCRC** Program at NSF
 - **I**ndustry/**U**niversity **C**ooperative **R**esearch **C**enter
- CHREC is both National **R**esearch **C**enter and **C**onsortium
 - **R**esearch **b**ase: 4 major universities
 - University of Pittsburgh (lead site)
 - University of Florida
 - Brigham Young University
 - Virginia Tech
 - **CHREC members:**
 - Industry & government organizations
(Later slide)



What is SHREC?



- CHREC **sunset** in Dec. 2017
- **Reborn** in Jan. 2018 as new NSF Center for **S**pace, **H**igh-performance, and **R**esilient **C**omputing (**SHREC**)
- Fundamental change - transition:
 - **From** a mature Center focused upon a **technology** (reconfigurable computing)
 - **To** a new Center focused upon a **purpose** (mission-critical computing)

SHREC focuses on 3 related domains of **mission-critical computing**

- *Space computing* for Earth science, space science, and defense
- *High-performance computing* for a broad range of grand challenge apps
- *Resilient computing* for dependability in harsh or critical environments

UF main focus

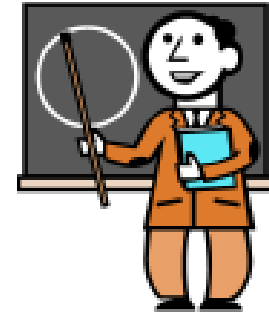


Center Members (2018)



1. AFRL Sensors Directorate
2. AFRL Space Vehicles Directorate
3. Army Research Laboratory
4. BAE Systems
5. Boeing
6. Draper Lab
7. Harris
8. Honeywell
9. Intel/Altera
10. L3 Space & Sensors
11. Laboratory for Physical Sciences
12. Lockheed Martin
13. Los Alamos National Laboratory
14. MIT Lincoln Laboratory
15. NASA Ames Research Center
16. NASA Goddard Space Flight Center
17. NASA IV&V Facility
18. NASA Johnson Space Center
19. NASA Langley Research Center
20. National Reconnaissance Office
21. National Security Agency
22. Naval Air Systems Command
23. Naval Research Laboratory
24. Office of Naval Research
25. Raytheon
26. Rockwell Collins
27. Sandia National Laboratories
28. Satlantis
29. Space Micro
30. Walt Disney Animation Studios

Center Faculty



- **University of Pittsburgh (lead)**
 - **Dr. Alan George**, Mickle Chair Professor of ECE – *Founder & Director*
 - **Dr. Alex Jones**, Professor of ECE – *Associate Director*
 - **Dr. Jun Yang**, Professor of ECE
 - **Dr. Ervin Sejdic**, Associate Professor of ECE
 - **Dr. Wei Gao**, Associate Professor of ECE
 - **Dr. Jingtong Hu**, Assistant Professor of ECE
- **Brigham Young University**
 - **Dr. Michael Wirthlin**, Professor of ECE – *Co-Director*
 - **Dr. Brent Nelson**, Professor of ECE
 - **Dr. Brad Hutchings**, Professor of ECE
 - **Dr. Jeff Goeders**, Assistant Professor of ECE
- **University of Florida**
 - **Dr. Herman Lam**, Associate Professor of ECE – *Co-Director*
 - **Dr. Greg Stitt**, Associate Professor of ECE
 - **Dr. Ann Gordon-Ross**, Associate Professor of ECE
 - **Dr. Janise McNair**, Associate Professor of ECE
 - **Dr. David Ojika**, Research Associate
- **Virginia Tech**
 - **Dr. Wu-Chun Feng**, Professor of ECE and CS – *Co-Director*
 - **Dr. Mark Gardner**, Network Research Manager and Faculty

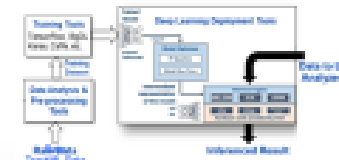
Most importantly,
SHREC features an
exceptional team of
students spanning our
university sites



SHREC Projects @ University of Florida Site

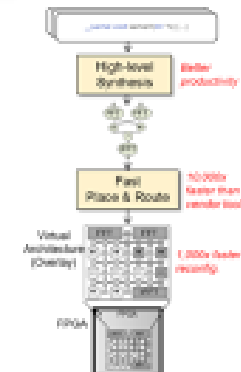
F1-18: Architecture Studies for New-Gen Systems

- 1) Custom Memory Cube (CMC) Research Platform
- 2) Heterogeneous Computing for Machine learning
- 3) Network Architecture Analysis for New-Gen Interconnects



F2-18: FPGA Virtualization and Application Case Studies

- 1) Overlay Strategies for Hardware Security
- 2) Overlay Strategies for Multi-FPGA Architectures
- 3) Ray Tracing on Emerging Architectures
- 4) DeepBench Evaluation on Xeon+FPGA



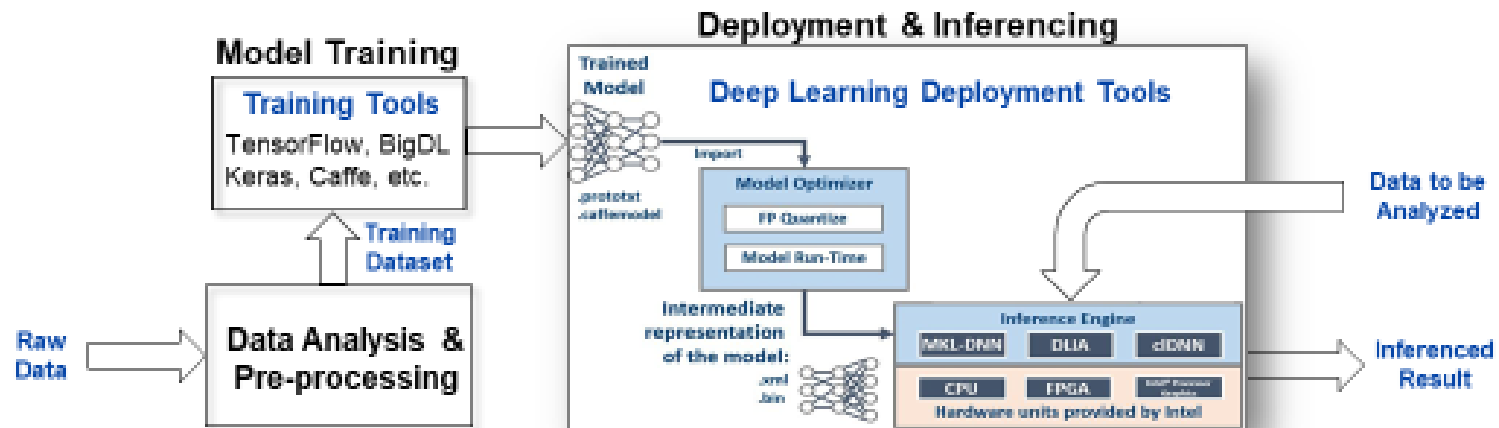
F1-P2: Heterogeneous Computing for Machine Learning

Motivation

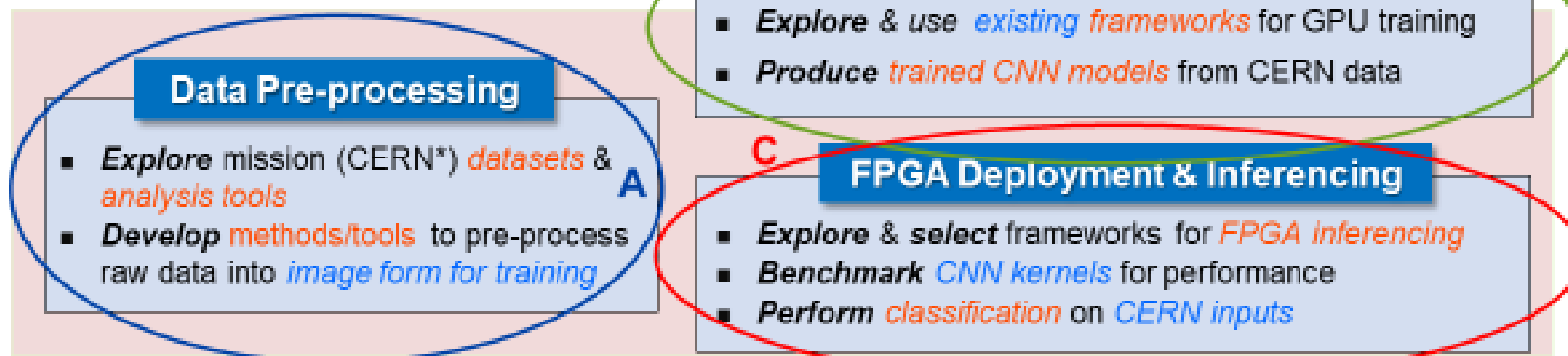
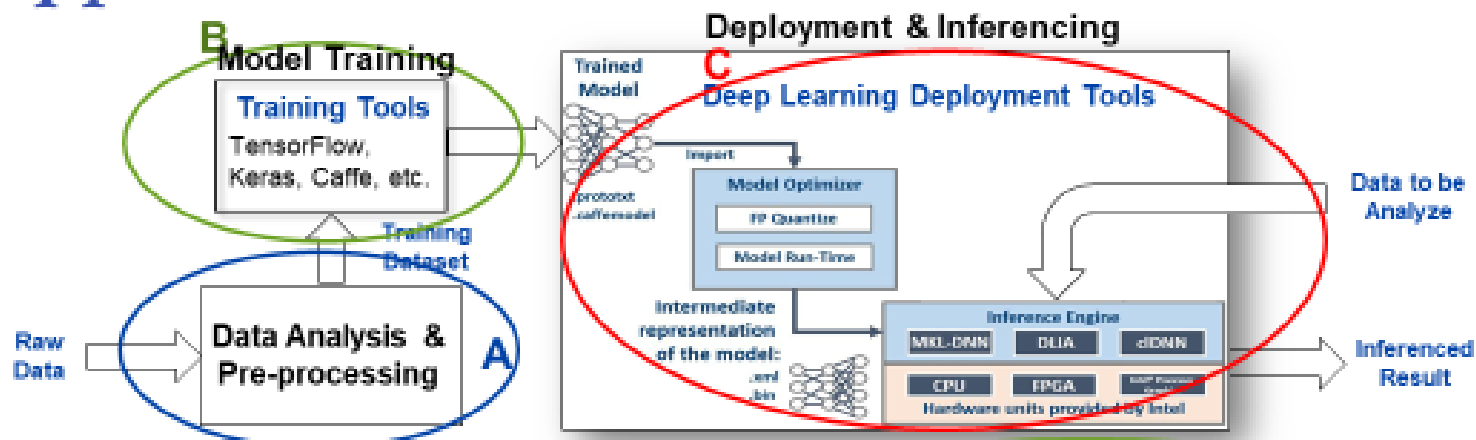
- CNN* holds great promise in **machine learning (ML)** for many **mission-critical** applications
- **Heterogeneous computing (HGC)** offers unique capabilities to accelerate CNN

Goal

- Explore state-of-the-art **HGC architectures** for **acceleration** of **ML algorithms** in selected mission-critical applications



Approach



Strength Through Collaboration



Project Contributor	Role
Univ. of Florida	Research lead
NERSC*	ML expertise
CERN OpenLab	Data source & domain expertise
Intel	ML primitives & tools
Dell	Solutions provider

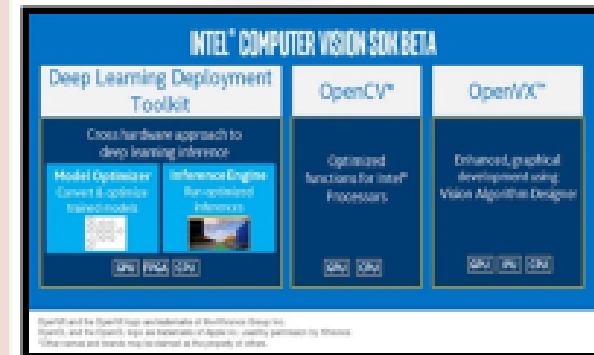
Dell & Intel



- Dell providing computing resources
- Early access to Intel Computer Vision SDK Beta under NDA (OpenVINO: recently announced product:)



Intel OpenVINO* Toolkit



NERSC at Lawrence Berkeley National Lab

- NERSC* has experience in pre-processing physics data into image form
- Interest in FPGA acceleration for machine learning
- One of UF students spending 10 weeks working with Berkeley group this summer.



Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE,
AND RESILIENT COMPUTING (SHREC)

* NERSC: National Energy Research Scientific Computing Center,
Lawrence Berkeley National Lab



University of
Pittsburgh

Virginia
Tech

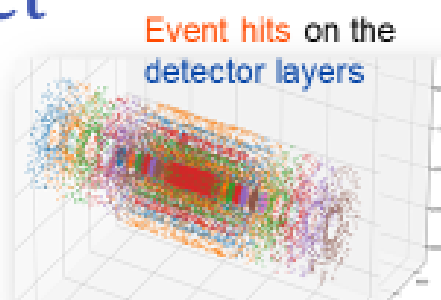
BYU
BRIGHAM YOUNG UNIVERSITY

UF
FLORIDA

TrackML Challenge* Dataset

Dataset

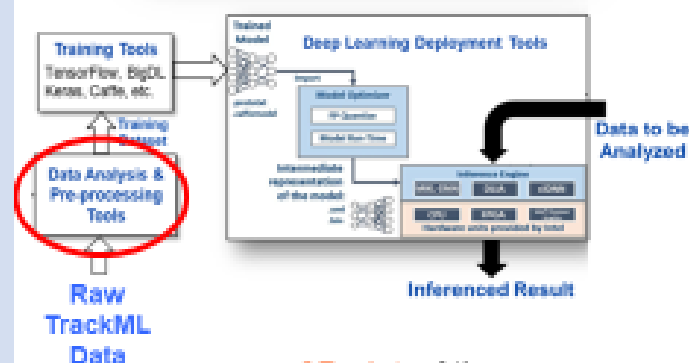
- Simulated dataset of real *HEP*** experiments
- Five sets of 1770 events each
- Total of 8850 events and 400GB of data



Pre-processing

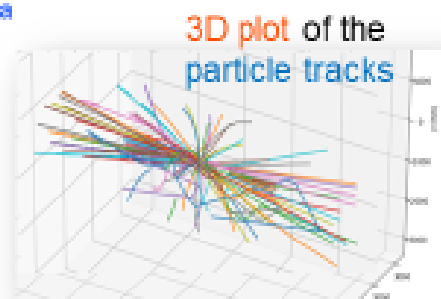
Developing *methods/tools* to pre-process data into *image form*

- Normalize raw data
- Plot sets of data into a 2D histogram
- Each *image* consists x, y and z coordinates of *event hits*



Machine-Learning Goal

- To perform classification on new incoming dataset to *identify particle tracks*
 - Using *FPGA-based inference engine*



Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE,
AND RESILIENT COMPUTING (SHREC)

* TrackML Challenge: <https://home.cern/about/updates/2018/05/are-you-trackml-challenge>

** High Energy Physics at Large Hadron Collider, CERN



ATLAS Dataset from CERN

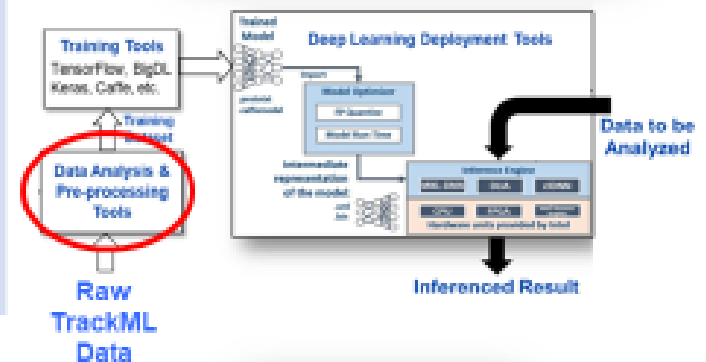
Dataset

- *ATLAS* – One of the key *detectors* at *LHC**
 - Six sets of *1000 events* each
 - Total of *6000 events* and *300GB* of data
- * LHC: Large Hadron Collider at CERN



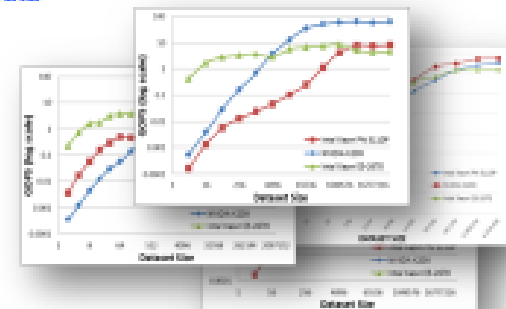
Pre-processing

- Explore *CERN analysis tools*
 - *CERN VM* – Linux environment with raw dataset and analysis software
 - *ROOT* – Framework for data pre-processing
- Develop methods/tools to *pre-process* the data into *image form*



Machine-Learning Goals **

- Identify *ML benchmarks*
 - Based on *trained ATLAS models*
- Benchmark *comparison*
 - Between *CPUs, GPUs, & FPGAs*
 - Latency, throughput, accuracy, energy, etc.

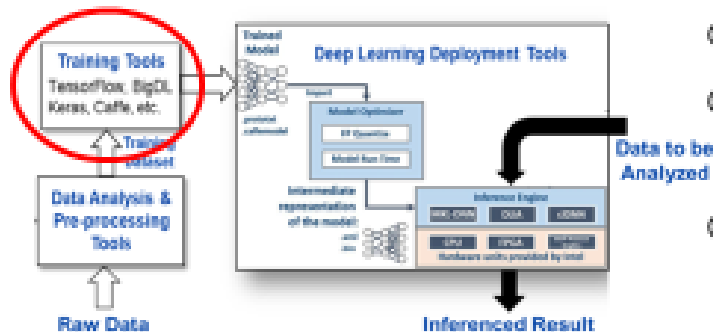


Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE,
AND RESILIENT COMPUTING (SHREC)

** In collaboration with NERSC: National Energy Research Scientific Computing Center, Lawrence Berkeley National Lab



CNN Model Training



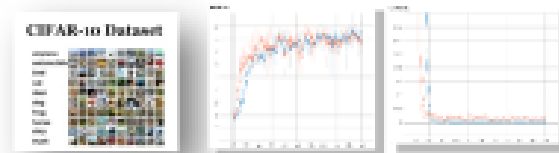
■ Explored existing ML frameworks

- TensorFlow, Keras, Caffe, and BigDL
- Implemented “vanilla” AlexNet model using TensorFlow
- Trained AlexNet (for image classification) using ImageNet & CIFAR-10 datasets

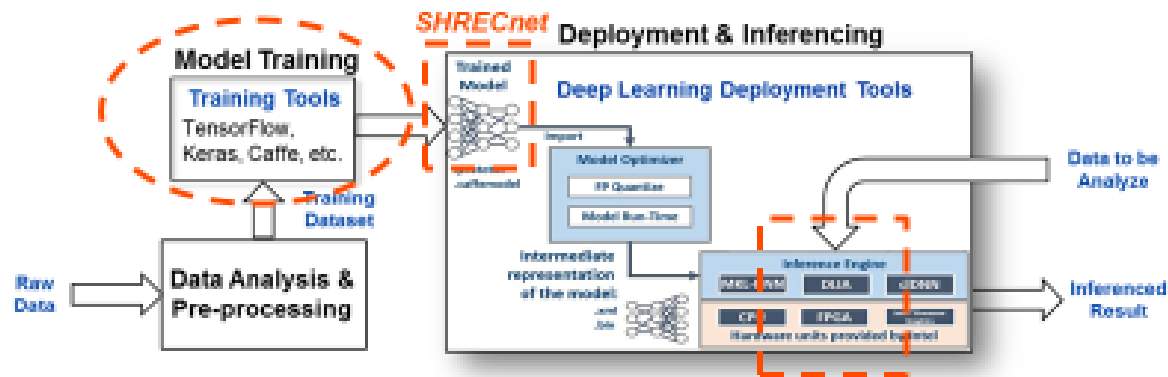


■ Going forward:

- Produce SHRECnet (trained CNN models) from pre-processed CERN datasets: TrackML and ATLAS
- Finalist in Dell EMC AI Challenge



Dell EMC AI Challenge



Winner to be announced
at *Supercomputing Conference*
Nov. 12, 2018, Dallas, TX

- **Focus** of our proposal for Dell EMC AI Challenge
 - To use the **provided Dell cluster** for the **Model Training stage**
- **At end of the challenge**
 - We will have **SHRECnet** based on the **ATLAS* & TrackML**** Challenge dataset from CERN
- SHRECnet will be **deployed**
 - On an **FPGA-based** inference engine
 - To perform **classification studies**



Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE,
AND RESILIENT COMPUTING (SHREC)

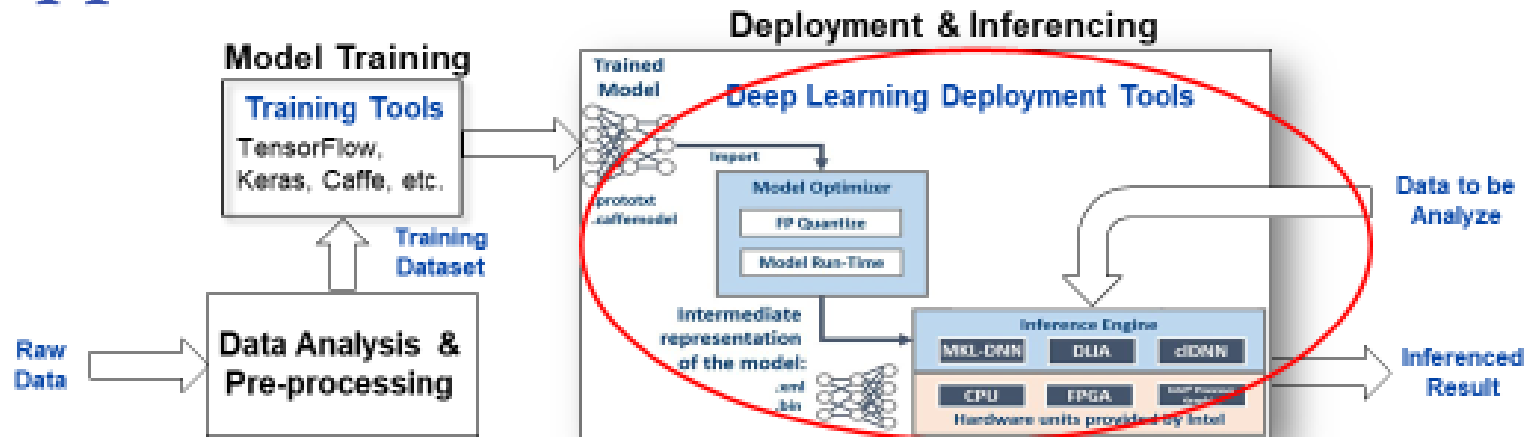
* ATLAS: a key detector in
Large Hadron Collider at CERN

** TrackML: TrackML Particle Tracking Challenge,
CERN, <https://sites.google.com/site/trackmlparticle/>

14



Approach



Data Pre-processing

- Explore mission (CERN*) **datasets** & **analysis tools**
- Develop **methods/tools** to pre-process raw data into **image form for training**

Model Training

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

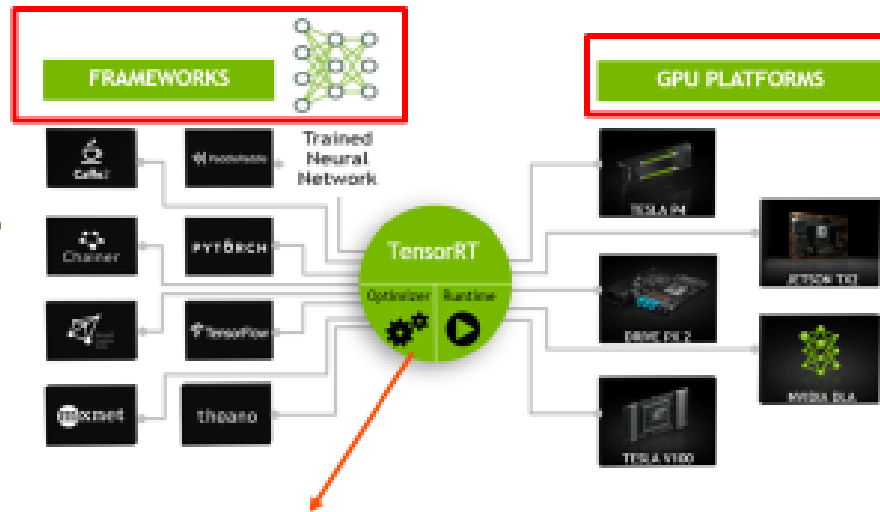
FPGA Deployment & Inference

- Explore & select frameworks for **FPGA inference**
- Benchmark **CNN kernels** for performance
- Perform **classification** on **CERN inputs**



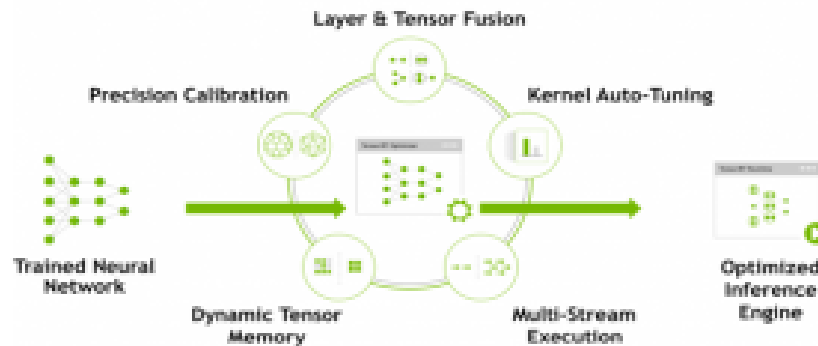
Inference Software & Tools for CPU/GPU

**NVIDIA
TensorRT**



- **Optimize** *neural network models*

- **Deploy** to platform for inferencing
- Tesla P4 • Tesla V100
- Tesla M40 • Drive PX2
- Tesla M4 • Jetson TX2



IBM PowerAI



- Software w/ complete lifecycle (end-to-end capability)
 - *Training > inferencing*
- PowerAI Inference Engine (AccDNN) supports
 - *CPU, GPU and (FPGA)*

FPGA Acceleration of CNN: Selected Research

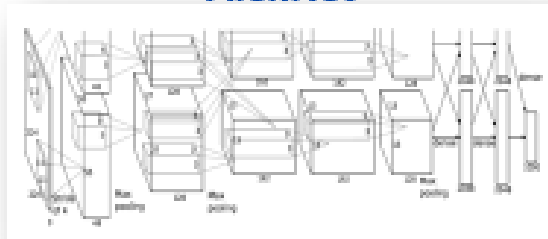
- Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., and Cong, J., "Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks," in Proceedings of ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2015.
- DiCecco, R., Lacey, G., Vasiljevic, J., Chow, P., Taylor, G. W., and Areibi, S., "Caffeinated FPGAs: FPGA Framework for Convolutional Neural Networks," CoRR, Vol. Abs/1609.09671, in Proceedings of International Conference on Field-Programmable Technology (FPT), 2016.
- Qiu, J. et al., "Going Deeper With Embedded FPGA Platform for Convolutional Neural Network," in Proceedings of ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2016.
- Suda, N. et al., "Throughput-Optimized OpenCL-based FPGA-based Accelerator for Large-Scale Convolutional Neural Networks," in Proceedings of ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2016.
- Aydonat, U., O'Connell, S., Capalija, D., Ling, A. C., and Chiu, G. R., "An OpenCL Deep Learning Accelerator on Arria 10," in Proceedings of ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2017.
- Wang, D, Xu, K, and Jiang, D, "PipeCNN: An OpenCL-Based Open-Source FPGA Accelerator for Convolution Neural Networks," in Proceedings of International Conference on Field-Programmable Technology (FPT), 2017.
- Zhao, R. et al., "Accelerating Binarized Convolutional Neural Networks with Software-Programmable FPGAs," in Proceedings of ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2017



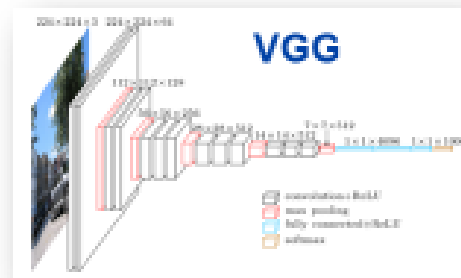
PipeCNN

- An OpenCL-Based FPGA accelerator
 - For Large-Scale Convolution Neuron Networks
- Implemented large scale CNNs on *Altera Stratix-V A7*
 - AlexNet* and VGG**

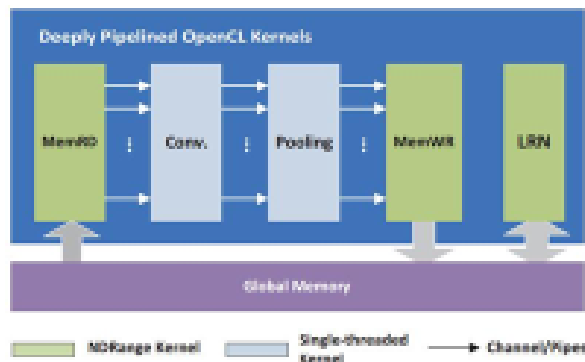
AlexNet



VGG






- Deeply pipelined architecture



- Free and Openly accessible architecture
- Highlights - *Data reuse* and *Task mapping* techniques
- Can be *reused* to explore new architectures for *neural network accelerators*

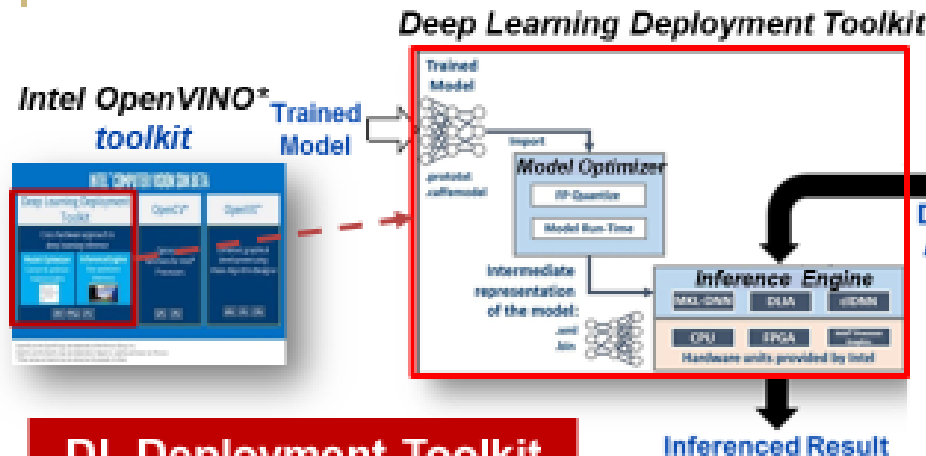


Tools for FPGA acceleration of CNN

- **NVIDIA Deep Learning Accelerator**  **NVIDIA**
 - Open architecture for design of deep learning inference accelerators
 - NVDLA is a *set of IP cores* for different platforms
 - *Verilog model* can be mapped onto **FPGA** for synthesis and simulation in **RTL form**
- **IBM PowerAI** 
 - Software w/complete lifecycle capability: **Training>inferencing**
 - PowerAI Inference Engine (AccDNN) supports
 - *CPU, GPU and (FPGA)*
- **Intel OpenVINO Toolkit** 
 - Cross Hardware Platform approach to deep learning inference
 - Optimized deep learning solution across multiple Intel platforms
 - *CPU, GPU and FPGA*



Intel OpenVINO: DL Deployment Toolkit



Model Optimizer

- Input: *trained model* from training tool (e.g., TensorFlow)
- Output: model files mapped to *targeted inference engine* (CPU, GPU, **FPGA**, heterogeneous)

Inference Engine

- Inputs: Trained model & *data to be analyzed*
- Output: Probability-based *classification*

DL Deployment Toolkit

- Explored toolkit for FPGA inferencing
 - Optimized/translated *AlexNet* ("vanilla" & modified models) and *GoogleNet* via *Model Optimizer*
 - Performing *classification* on images using both models using FPGAs



Going Forward

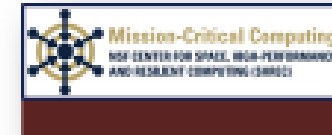
Classification studies on *FPGA inference engine*

- *Benchmark kernels* for CNN
- *Mission-critical datasets* from CERN: TrackML, ATLAS
- **FPGA** vs. CPU, GPU
- **Performance metrics:** latency, throughput, batch size vs. latency, FLOPS/Watts, #images/Watts, accuracy, varying bit precisions (8, 16, 32), floating-point vs. int.





Agenda



- What is CHREC/SHREC?
- Heterogeneous Computing (HCG) for Machine Learning (ML)
 - State of the art in using FPGAs for machine learning
- R&D activities in SHREC Center
 - Opportunities in HGC for ML
- Q & A



Q&A

