
Medium term plans with FPGA co-processors

PREPARING FOR THE NEXT TARGET

<https://aws.amazon.com/summits/washington-dc/>



The banner features a dark red background with a yellow and orange graphic on the left. The graphic includes the text 'AWS PUBLIC SECTOR' in white, 'SUMMIT' in large white letters on a yellow background, and 'Washington DC' in white on an orange background. A small white box contains the text 'AWS_SummitPublicSector_Logo_RGB_Transp_Horiz_CityRight_WashingtonDC'. To the right of the graphic, the text 'AWS Public Sector Summit for government, education and nonprofits' is written in white. Below this, the dates 'June 12-14, 2017' and the location 'Walter E. Washington Convention Center' are also in white.

AWS PUBLIC SECTOR
AWS_SummitPublicSector_Logo_RGB_Transp_Horiz_CityRight_WashingtonDC

S U M M I T

Washington DC

AWS Public Sector Summit
for government, education and nonprofits

June 12-14, 2017

Walter E. Washington
Convention Center

STATUS OF THINGS

hls4ml (beta) is nearing primetime - few weeks

<https://hls-fpga-machine-learning.github.io/hls4ml/>

A generic package which does automatic translations of NN architectures into HLS and firmware

We made progress recently on “serial mode” - e.g. larger models without strict latency constraints

Fully validated for DNN

Not yet ready for CNN or LSTMs...

STATUS OF THINGS

Progress running an **HLS-based** project on the F1 instances

<https://indico.fnal.gov/event/16258/contribution/0/material/slides/0.pdf>

We have a straw-man for how to stream in data and drop directly in an hls4ml project

https://github.com/nhanvtran/SDAccel_Examples/tree/first-try/getting_started/host/hls4ml_1layer_hls

Not fully validated but have a roadmap of what to do

A (REASONABLE) WORK PLAN

Train a reasonably sized DNN for Higgs tagging (expert features)

Finish getting a small example running on F1 instance

→ Merge these into a working project

Benchmark this (throughput, timing) on a

CPU

GPU

vs. F1

Is this a substantial enough work plan to present on?

ABSTRACT

Deep learning acceleration to dig out the Higgs at the LHC

The discovery of the Higgs boson is one of the most profound physics achievements this century. As we continue detailed study of the Higgs as a potential window to other new physics, it requires unearthing it from overwhelming backgrounds. We develop deep learning algorithms to improve Higgs reconstruction efficiency and improve sensitivity using AWS GPU resources to build sophisticated neural network architectures for so-called “Higgs tagging”. Given exabyte-scale LHC datasets, improving the inference speed of these Higgs tagging networks, and other vital ML algorithms, will be very important for the future of the LHC data analysis. Therefore, we also benchmark the inference time of these networks using newly available accelerating CPU-FPGA co-processors and compare them to CPU and GPU performance using AWS resources.