GPU and FPGA as a Service for Machine Learning Inference Accelerations

Speaker: Yu Lou (University of Washington)
Authors:

Fermi National Accelerator Laboratory

Maria Acosta Flechas, Benjamin Hawks, Philip Harris, Burt Holzman, Thomas Klijnsma, Kyle Knoepfel, Mia Liu, Kevin Pedro, Nhan Tran, Michael Wang, Tingjun Yang

Massachusetts Institute of Technology

Jack Dinsmore, Philip Harris, Jeffrey Krupa, Dylan Rankin

University of California San Diego

Javier Duarte

University of Washington

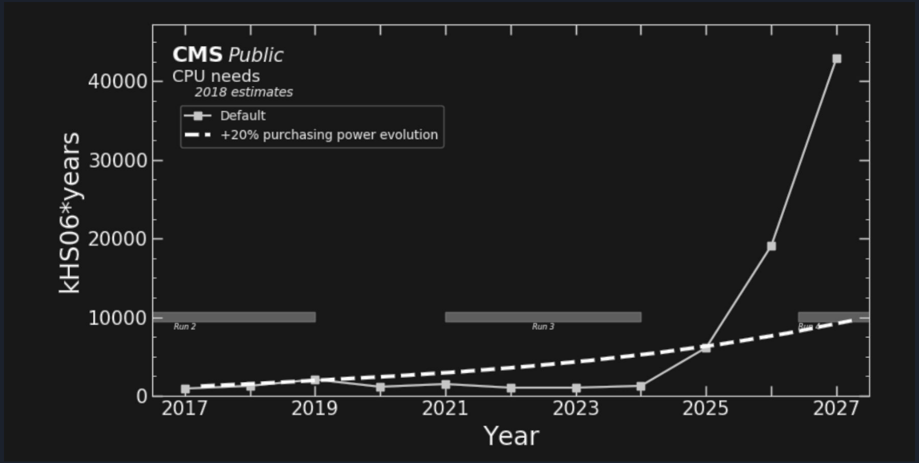
Scott Hauck, Ta-Wei Ho, Shih-Chieh Hsu, Kelvin Lin, Yu Lou, Natchanon Suaysom, Matthew Trahms

References:

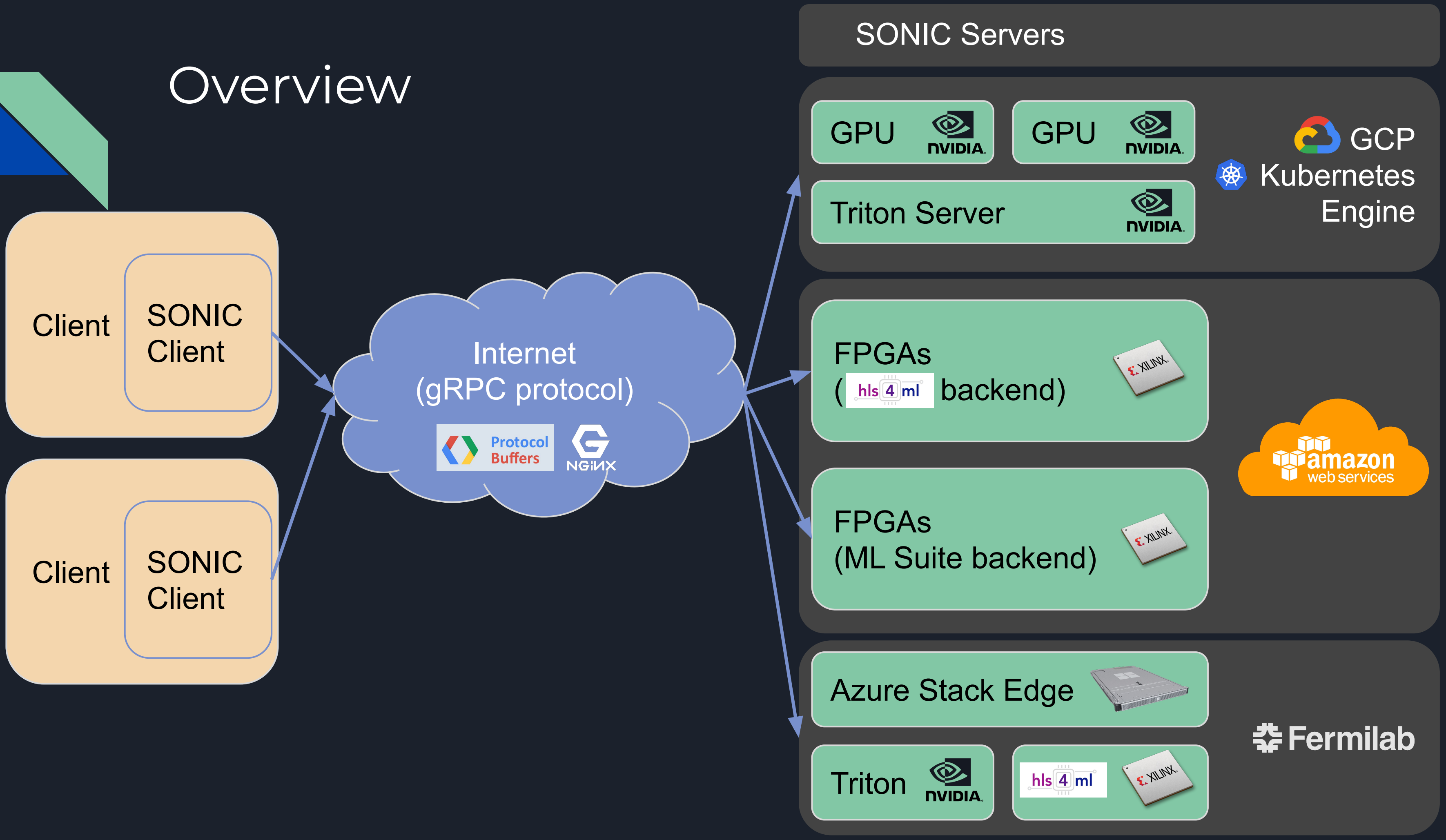
- *FPGAs-as-a-Service Toolkit (FaaST)*
- *GPU coprocessors as a service for deep learning inference in high energy physics*
- *GPU-accelerated machine learning inference as a service for computing in neutrino experiments*

Overview

- The demand for computer resources for LHC and neutrino experiments is going to surge after planned upgrades.
- We developed the Services for Optimized Network Inference on Coprocessors (SONIC) framework
  - Use hardware accelerators (GPU, FPGA etc.) to perform machine learning inferences.
  - Provide a uniform machine learning inference interface to the client



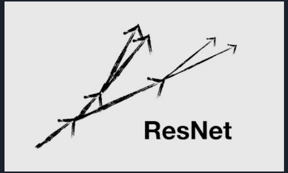
Overview


SONIC Servers
Client
SONIC Client
Client
SONIC Client
Internet
(gRPC protocol)
Protocol Buffers
NGINX
GPU
NVIDIA
GPU
NVIDIA
Triton Server
NVIDIA
GCP
Kubernetes Engine
FPGAs
(hls4ml backend)
XILINX
FPGAs
(ML Suite backend)
XILINX
amazon web services
Azure Stack Edge
Triton
NVIDIA
hls4ml
XILINX
Fermilab

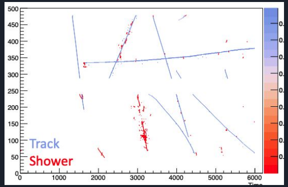
The Models

Fast Calorimeter Learning (FACILE): reconstruct the energy deposited by particles in the hadron calorimeter (HCAL) of the CMS experiment. (2000 parameters)

DeepCalo: electron energy regression for the ATLAS detector. (1.8 million parameters)

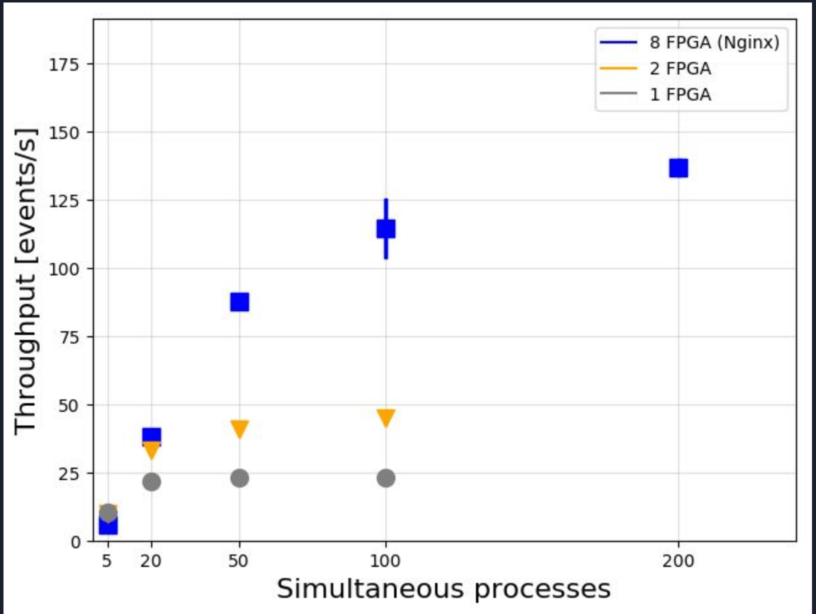


Top Quark Tagging: identify events containing top quarks. (23 million parameters)

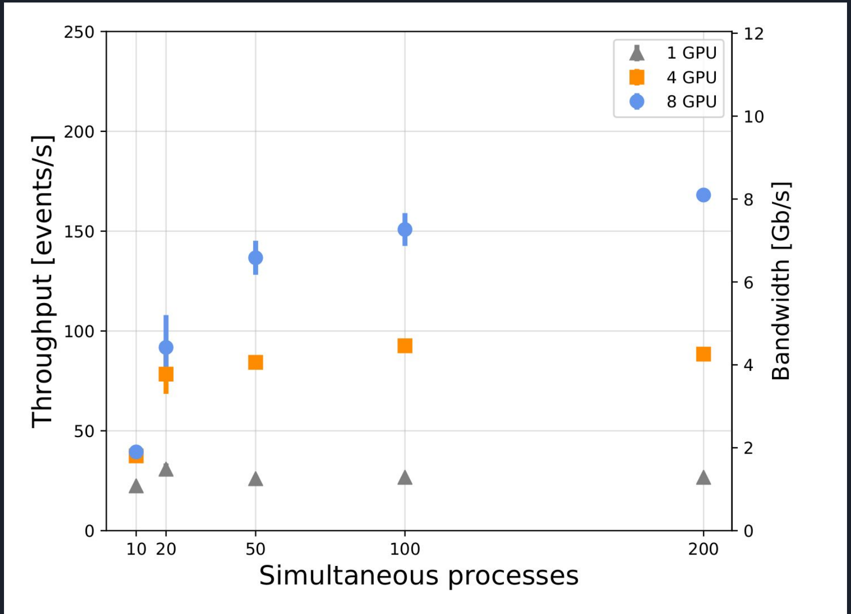


Neutrino classification: identify track and particle shower hits for ProtoDune single phase apparatus. (12 million parameters)

Benchmarking

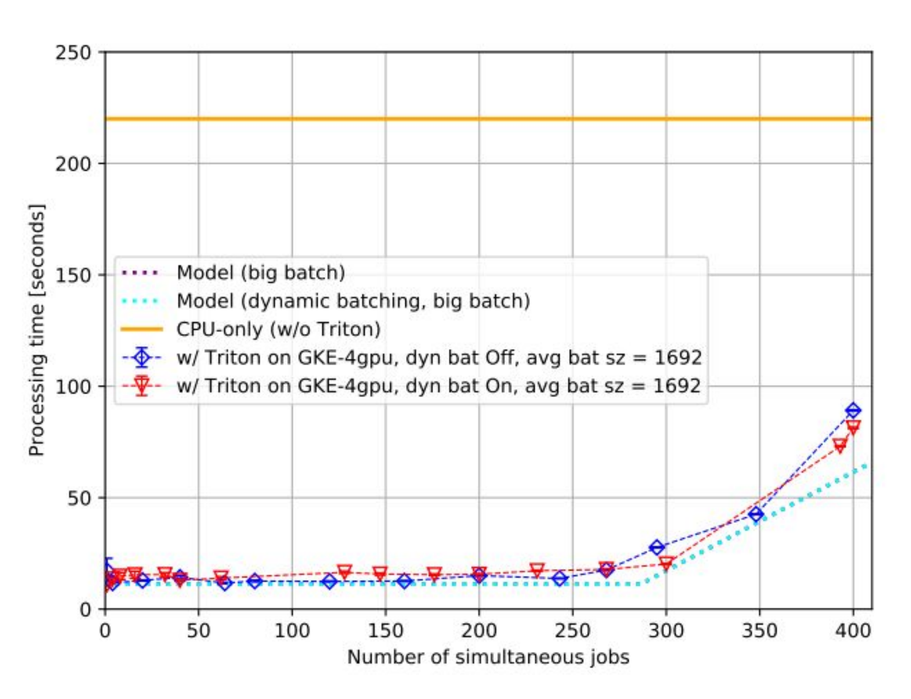


AWS FPGA, ResNet-50 (8 bit fixed point)



GCP Triton, ResNet-50

Benchmarking



GCP Triton, Neutrino classification

Q&A

Thanks!