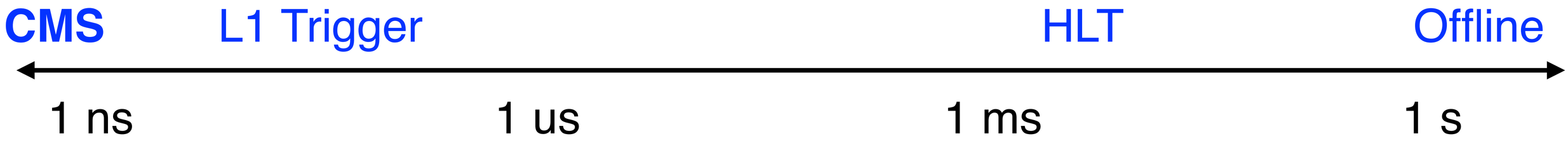
Machine learning in FPGAs

Javier Duarte, Burt Holzmann, Sergo Jindariani, Benjamin Kreis,
Kevin Pedro, Ryan Rivera, Nhan Tran
Fermilab

+ collaborators from UIC, MIT, CERN and industry

LATENCY LANDSCAPE



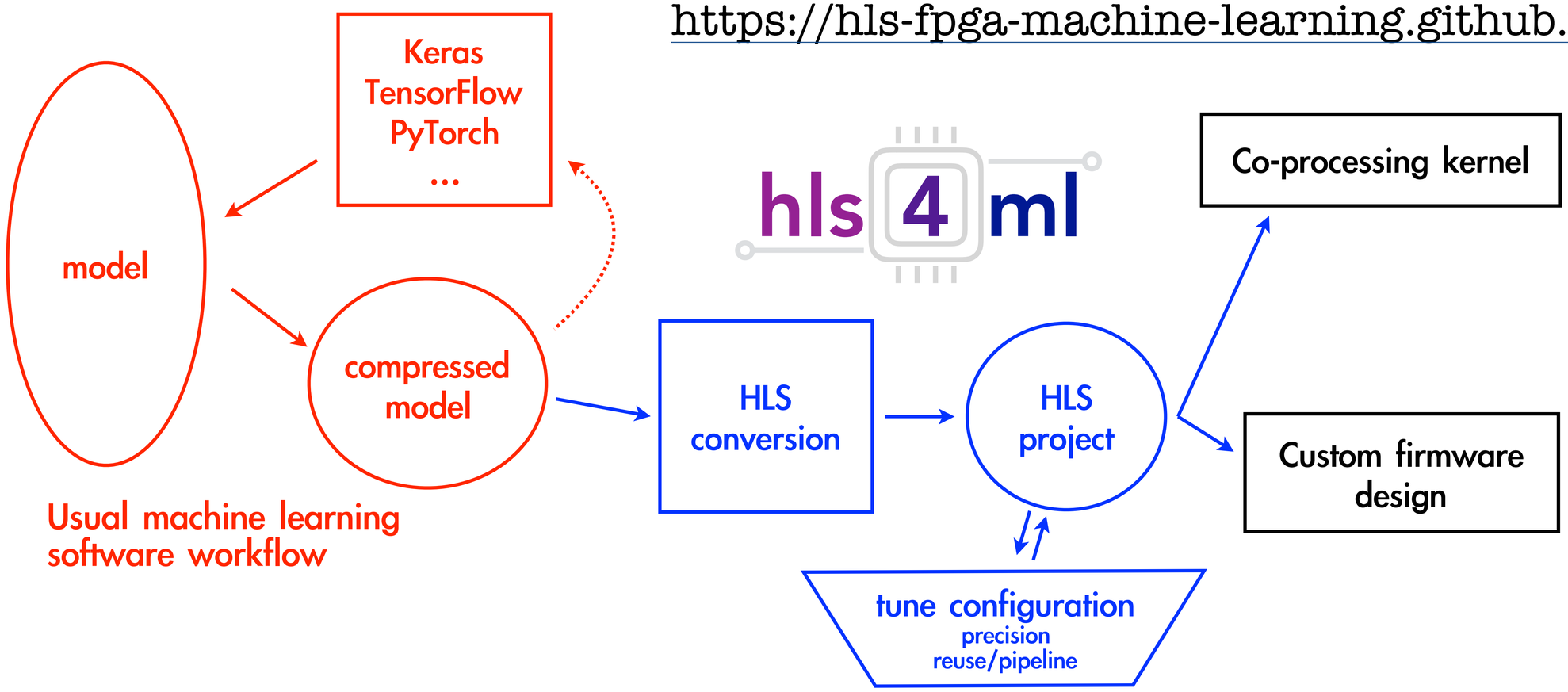
Machine learning is being used to solve a wide array of problems across a large range of latency constraints

Assumptions

most algorithms can be formulated as machine learning problems
new specialized hardware will be optimized for machine learning

First complete, general trigger study

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



Fast inference of deep neural networks in FPGAs for particle physics

Javier Duarte[a] , Song Han[b] , Philip Harris[b] , Sergo Jindariani[a] , Edward Kreinar[c] , Benjamin Kreis[a] , Jennifer Ngadiuba[d] , Maurizio Pierini[d] , Ryan Rivera[a] , Nhan Tran[a] , Zhenbin Wu[e]

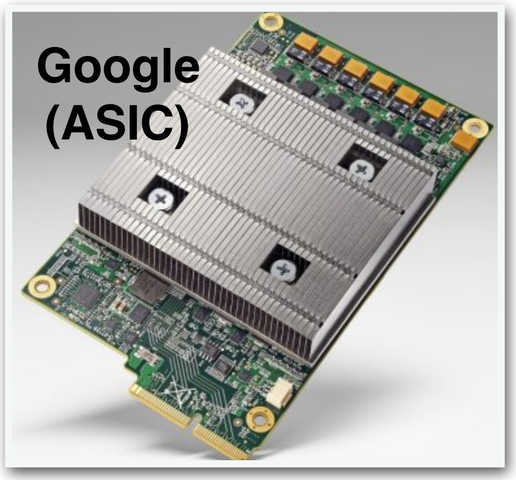
[a] *Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*
[b] *Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[c] *HawkEye360, Herndon, VA 20170, USA*
[d] *CERN, CH-1211 Geneva 23, Switzerland*
[e] *University of Illinois at Chicago, Chicago, IL 60607, USA*

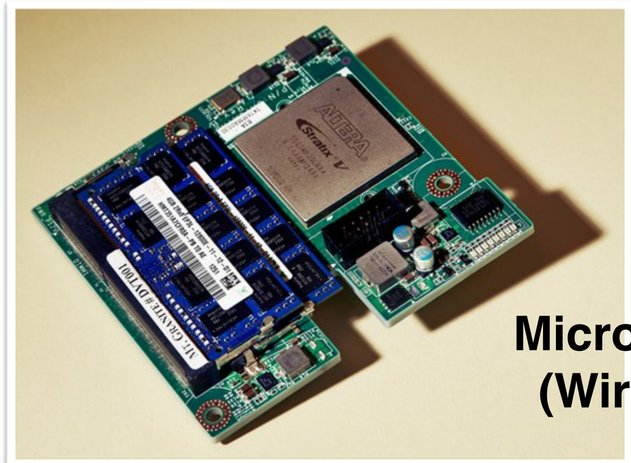
https://arxiv.org/abs/1804.06913

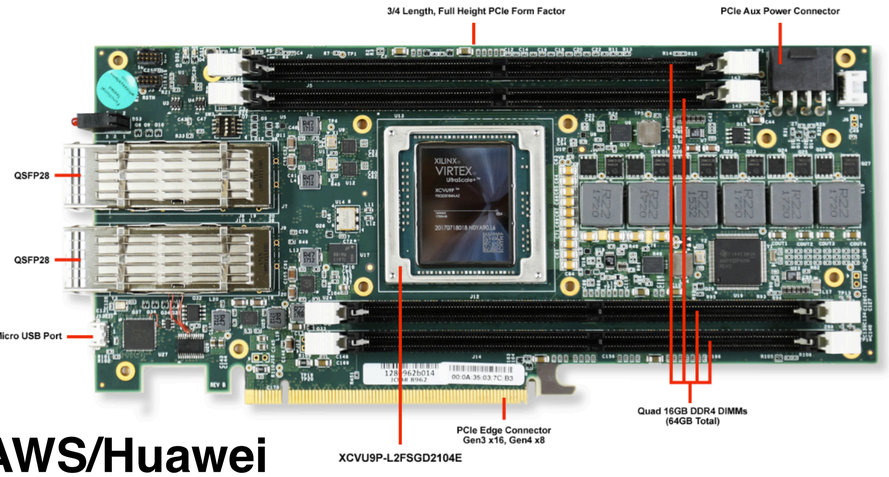
See Research Techniques Seminar
By Javier Duarte,
April 24 @ 1030 AM

CO-PROCESSORS

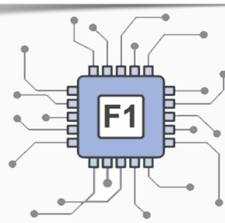
LARGE SPEED/ENERGY GAINS OVER CPU/GPU!







Xilinx/AWS/Huawei



amazon web services

Efficiently interface between CMSSW and accelerator hardware?

- Some pilot projects in progress with: Microsoft Azure + MSR , Amazon Web Services
- Exploring connections through CERN OpenLab with: Intel
- Academia?
- ***These systems will only improve (higher throughput, more DSPs/memory)***

Goals, benchmark performance of co-processor hardware against other computing architectures (CPU. GPU. etc.)

SEE MACHINE LEARNING FORUM TALK, ANDREW PUTNAM (MICROSOFT RESEARCH), MAY 14 @ 1PM