“On-chip” Computation

FPGA frameworks for edge and near-edge computing:
examples and strategy towards HL-LHC

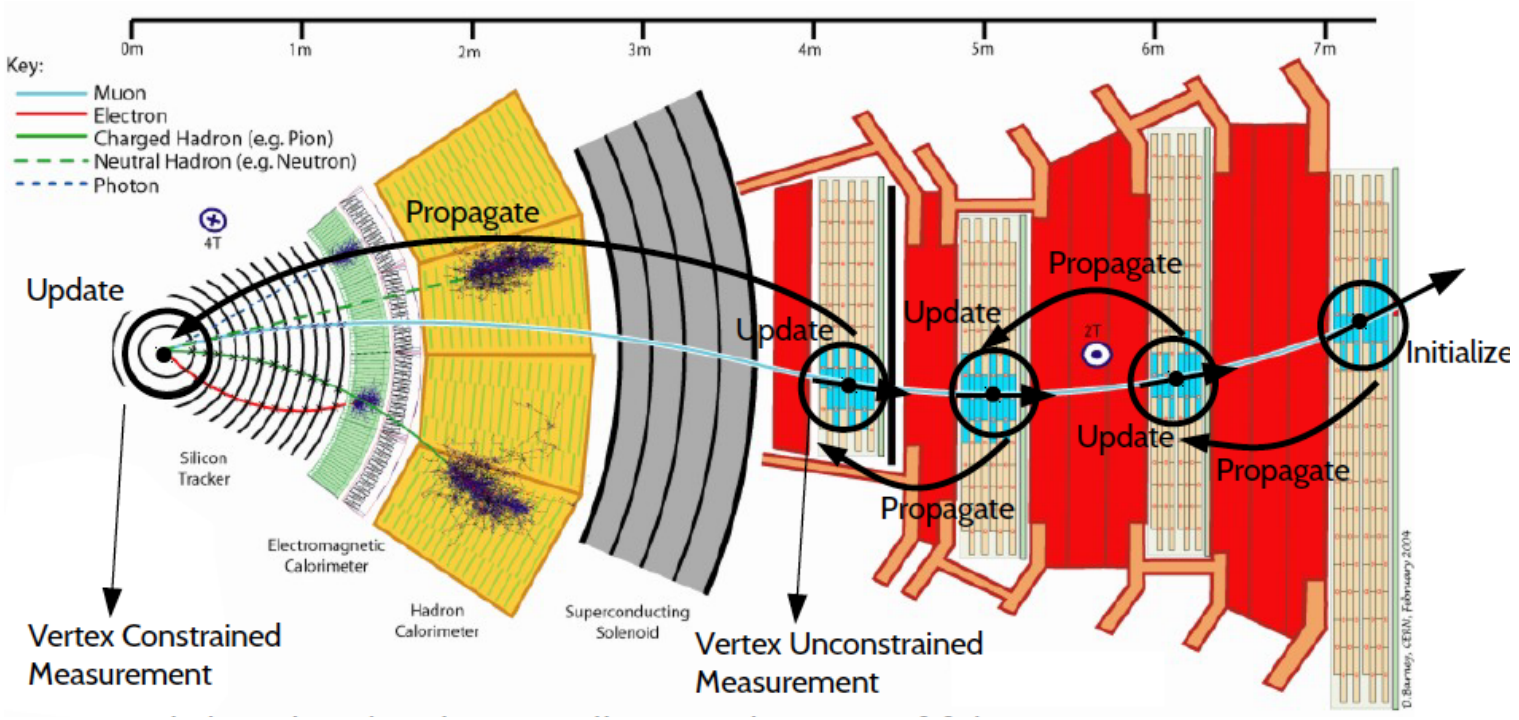
Michalis Bachtis (UCLA), Javier Duarte (FNAL)

Fermilab

US CMS

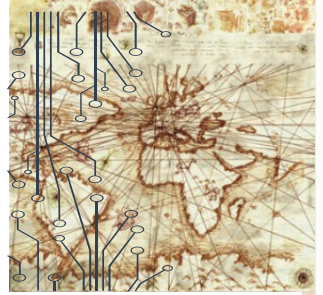
Compute on hardware

- Track Reconstruction: CPU expensive algorithm at HL-LHC
- Recent application of "offline" track reconstruction in FPGAs with a **Kalman filter** in the L1 Muon Trigger in current CMS data taking
  - opens doors towards accelerating track reconstruction algorithms with FPGAs also offline at HL-LHC



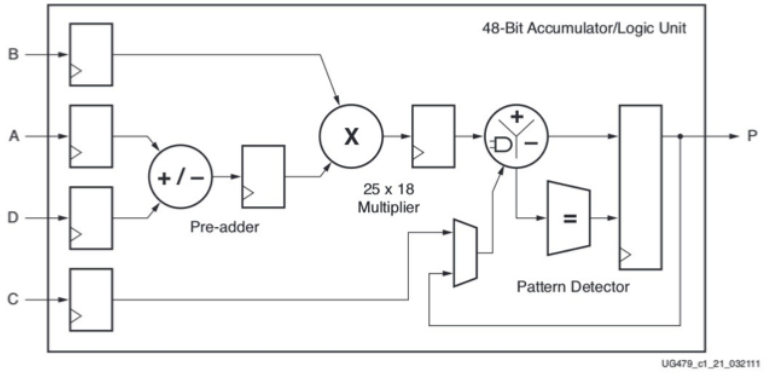
M. Bachtis et al

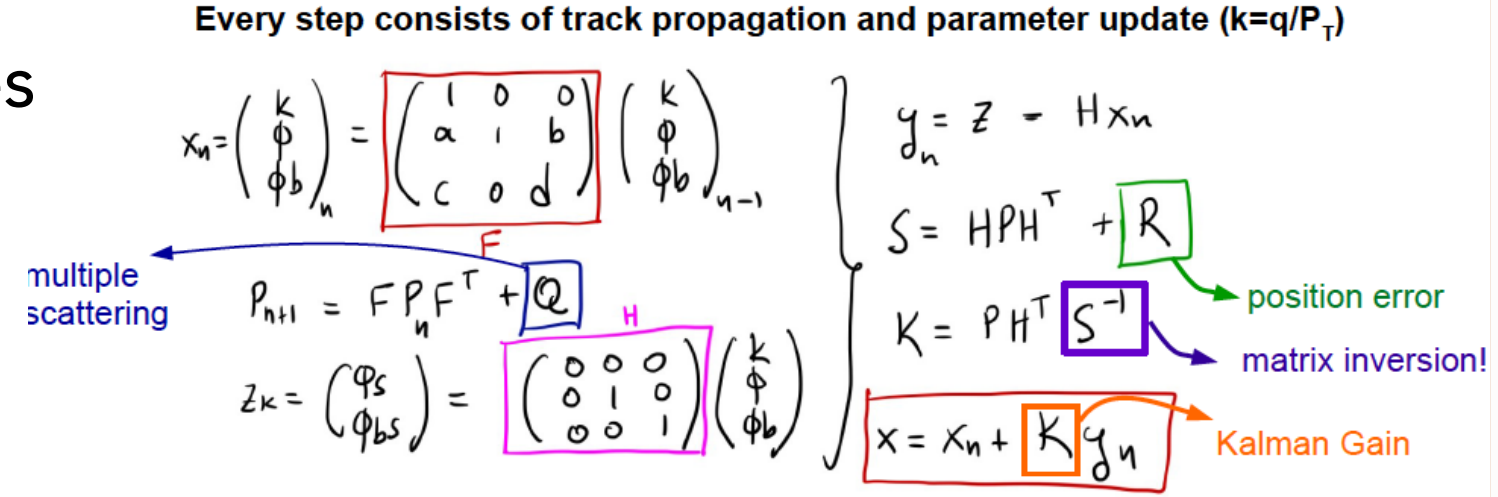
Implementing a Kalman Filter in an FPGA



Matrix algebra including matrix inversion!

Map algebra to DSP cores





Modern FPGAs: 1000s of DSP cores

- Exist for filtering, AI, and military applications
  - ASIC cores in the FPGA that contain wide multipliers and adders
- Exploiting this commercially available resource reduced required FPGA resources by x5

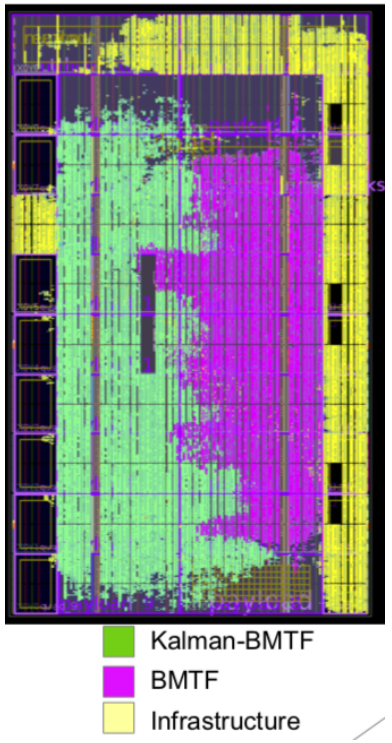
Implementation results

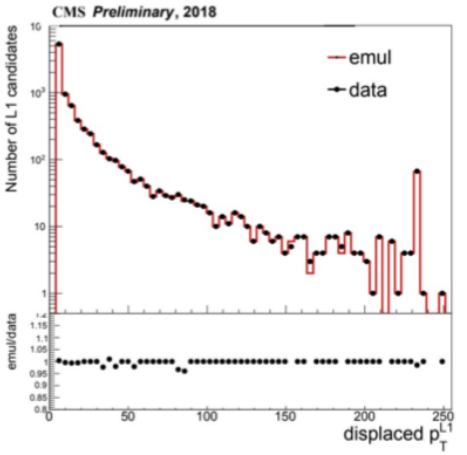
FPGA firmware written with the latest High Level Synthesis (HLS) tools in C

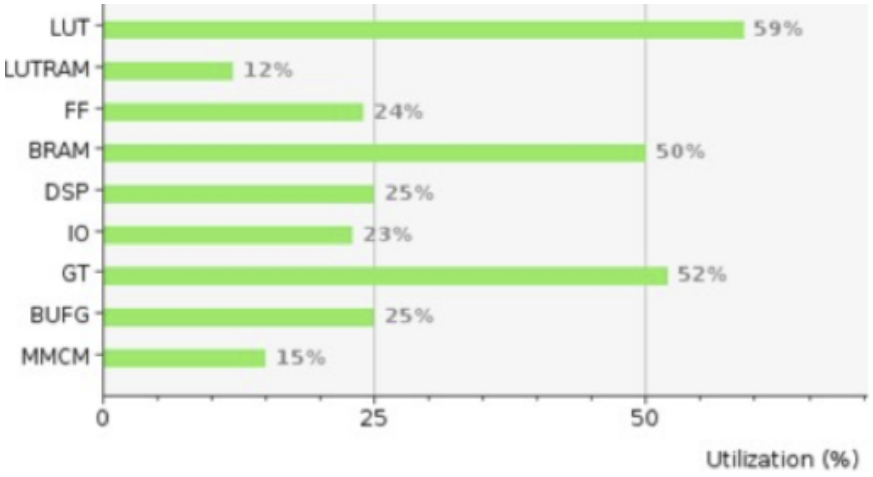
- Deployed in CMS L1 data taking in Run I
- Reconstructs all muon tracks in 150ns!

Proving that C code can run efficiently on an FPGA

- Fundamental step towards accelerating current C/C++ offline algorithms with FPGAs in HL-LHC







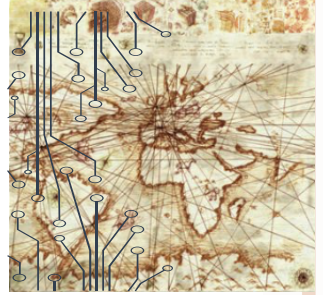
Fermilab

Interface between L1 Trigger and computing towards HL-LHC

US CMS performing R&D with cutting edge technology FPGAs for the L1 Trigger

FPGA vendors moving towards combining many different technologies in a single chip

- Future generation FPGAs [to arrive in the market in 2020] will combine FPGA logic with CPUs and specific AI cores towards an adaptable computing engine
- While the L1 Trigger (due to latency limits of ~μs) will mostly benefit from the FPGA logic, the same device can be re-configured for a computing application accelerating algorithmic parts using the FPGA logic



HL-LHC strategy from Muon trigger implementation

1. Advanced/Clever Programming of Modern FPGAs

- Exploit DSP cores to reduce resource usage: More algorithms in a chip
- Running C algorithms in an FPGA: Enables acceleration of offline algorithms
- Bigger and faster FPGAs
  - Faster clocks make algorithms faster
  - Embedded computing elements inside chip perform co-processing

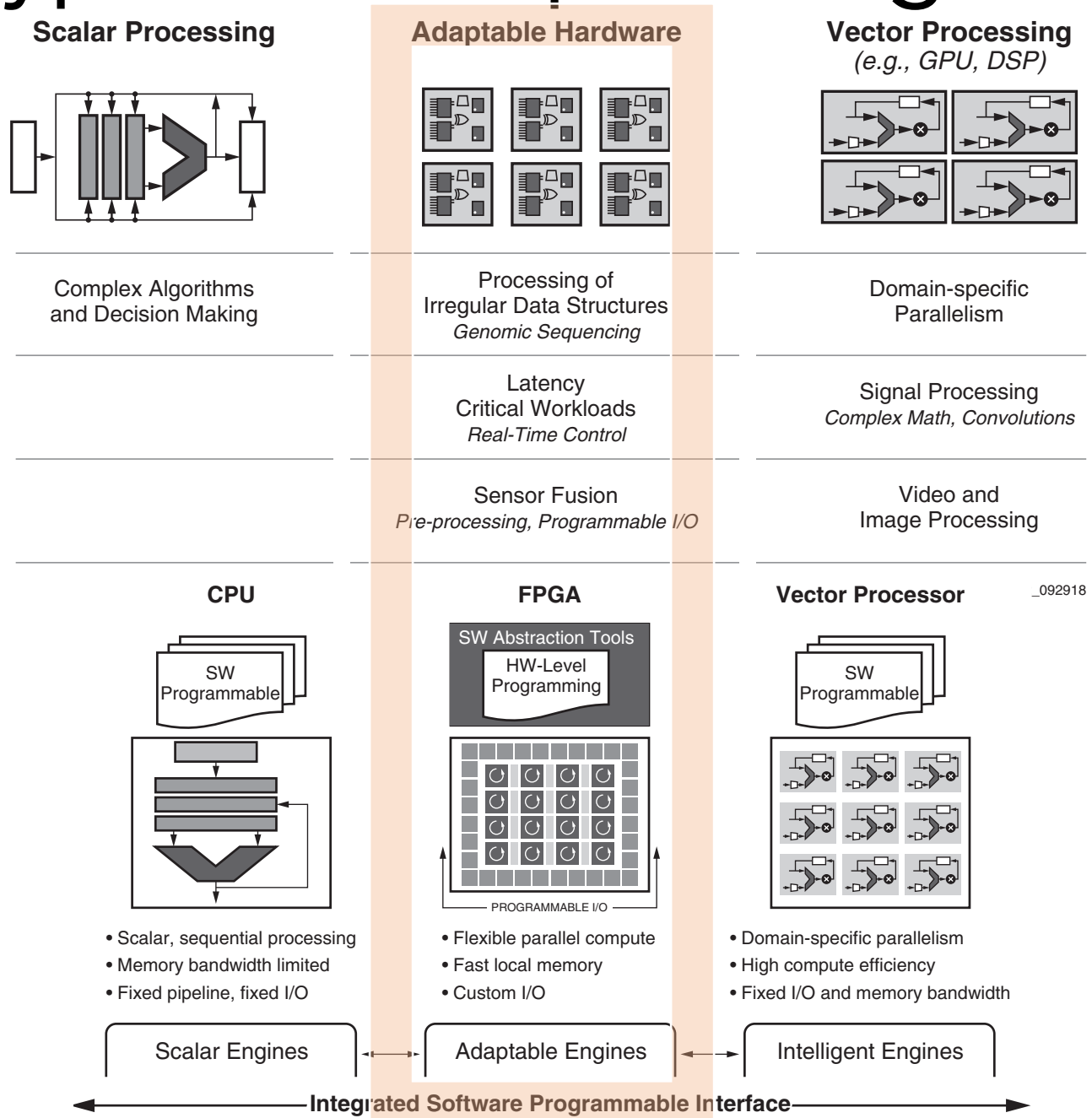
2. High Speed on-board data links & Hybrid On board (or on Chip) computing

More and/or higher speed links (~25 100Gbps Ethernet connections/FPGA!)

Connect multiple devices together

Many devices in the same Chip: Adaptable computing optimized for the application

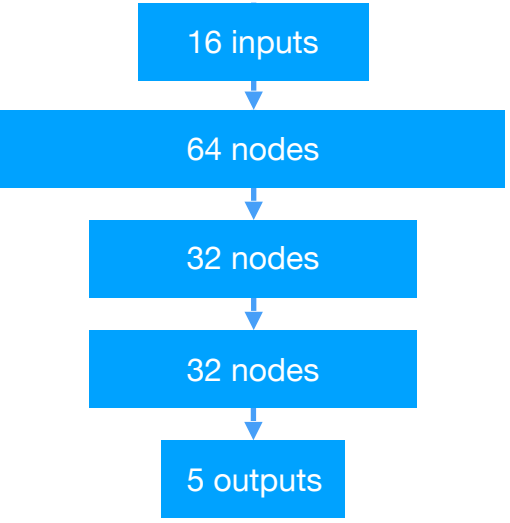
Reminder: types of compute engines

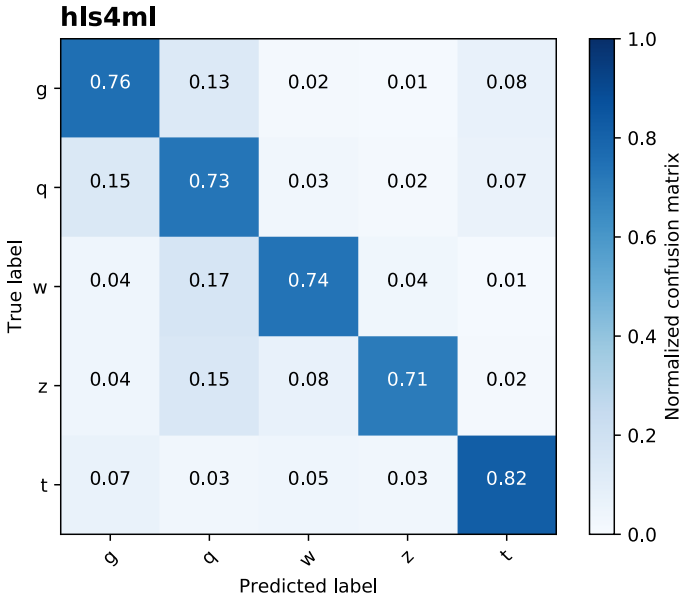


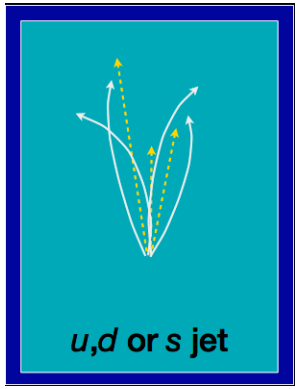
Fermilab

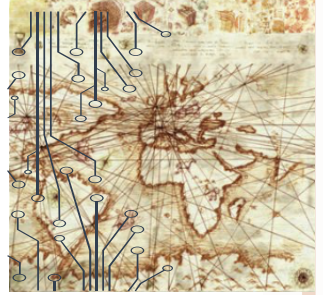
US CMS

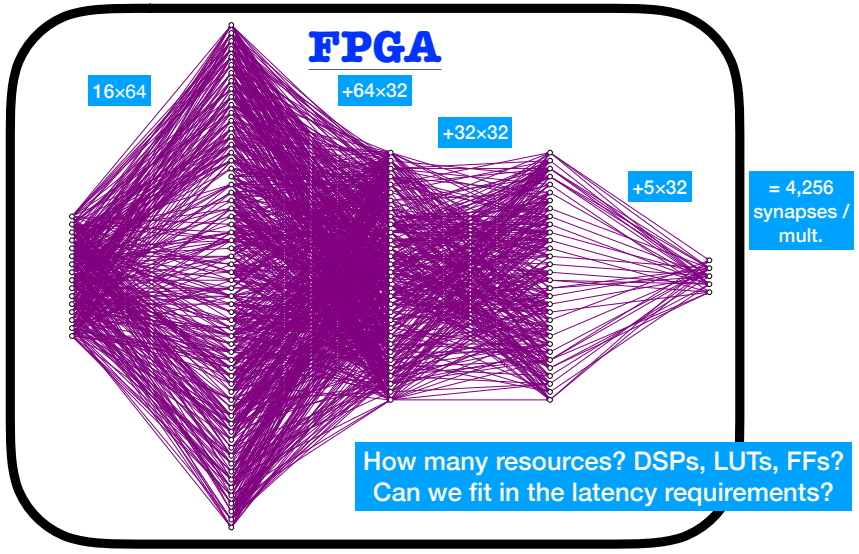
NN correctly identifies jets 70-80% of the time











OPTIMIZE NN's for FPGAs resource

Compress: Maintain high performance while removing redundant synapses and neurons

Quantize: Reduce precision from 32-bit floating point to 20-bit, 8-bit, …

Parallelize/Reuse: Balance: parallelization (how fast) with FPGA resources needed (how costly)

J. Duarte et al.

Fermilab

US CMS

JINST 13 P07027

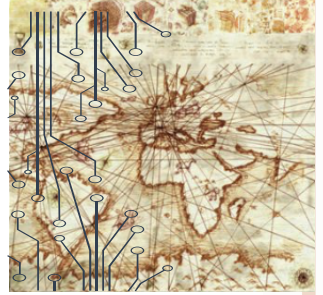
ics.ins-det] 28 Jun 2018

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]

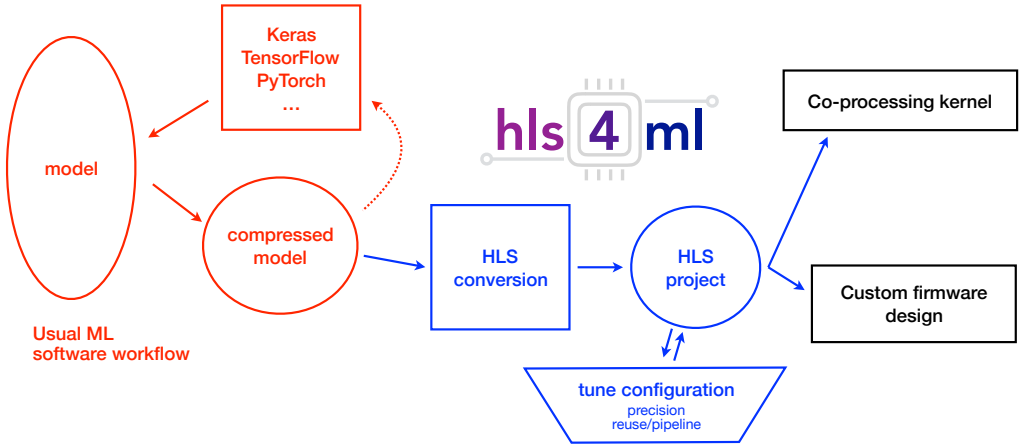
[a] Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
[b] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] HawkEye360, Herndon, VA 20170, USA
[d] CERN, CH-1211 Geneva 23, Switzerland
[e] University of Illinois at Chicago, Chicago, IL 60607, USA

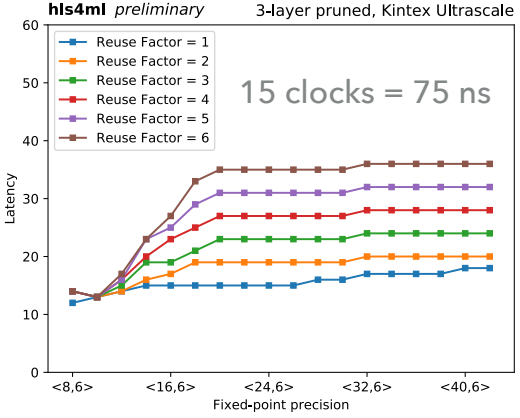
E-mail: hls4ml.help@gmail.com

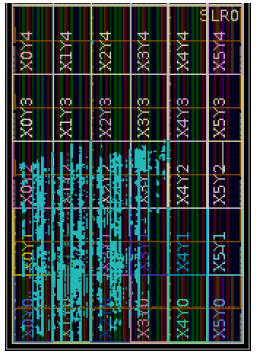


Tool: hls4ml

- **hls4ml** for physicists or ML experts to translate **ML algorithms** into **FPGA firmware**







- DNN inference in **< 100 ns**
- with **30% of Kintex Ultrascale FPGA DSPs**!

Citation

If you are using the package please cite:

- DOI 10.5281/zenodo.1204445
- J. Duarte *et al.*, "Fast inference of deep neural networks in FPGAs for particle physics", JINST 13 P07027 (2018), arXiv:1804.06913.

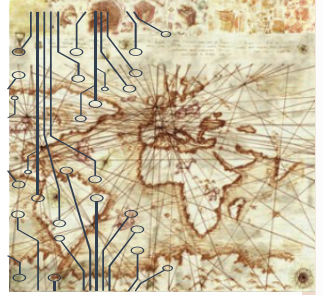
Contributors

- Vladimir Loncar, Jennifer Ngadiuba, Maurizio Pierini, Sioni Summers [CERN]
- Javier Duarte, Sergo Jindariani, Benjamin Kreis, Ryan Rivera, Nhan Tran [Fermilab]
- Edward Kreinar [Hawkeye360]
- Song Han, Philip Harris, Dylan Rankin [MIT]
- Zhenbin Wu [University of Illinois at Chicago]
- Mark Neubauer [University of Illinois Urbana-Champaign]
- Shih-Chieh Hsu [University of Washington]
- Giuseppe Di Guglielmo [Columbia University]

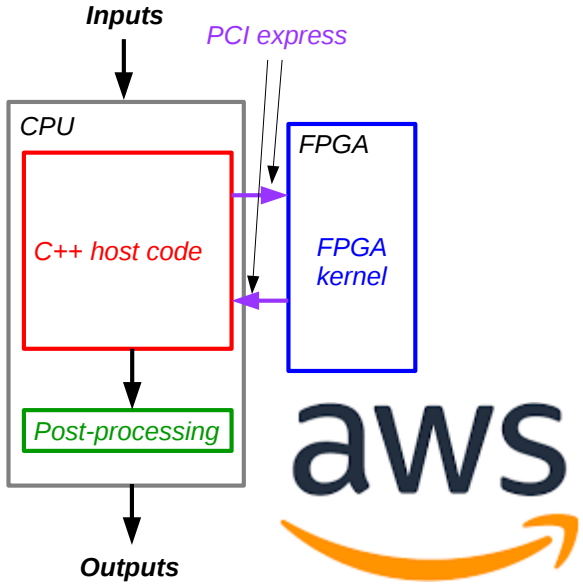
DUNE, ATLAS, Accel. Division interested

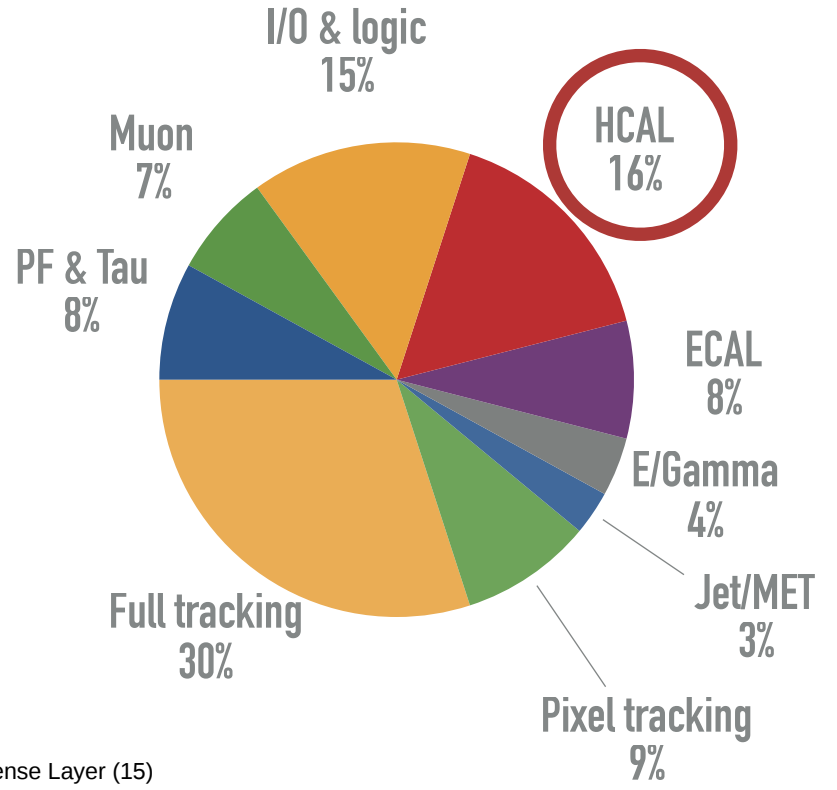
Fermilab

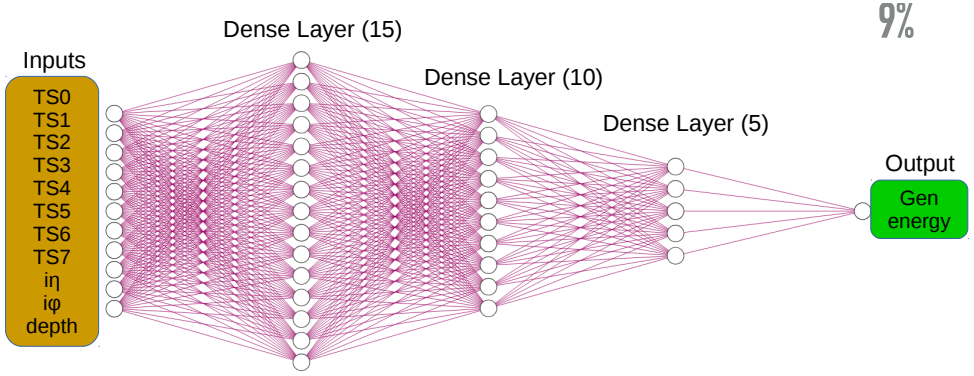
Accelerating High-Level Trigger with FPGAs



- HCAL local reconstruction contributes significantly to HLT compute time
- ML+FPGA as co-processor can reduce HCAL local reco. compute time by up to ×16
- Tested using AWS FPGAs







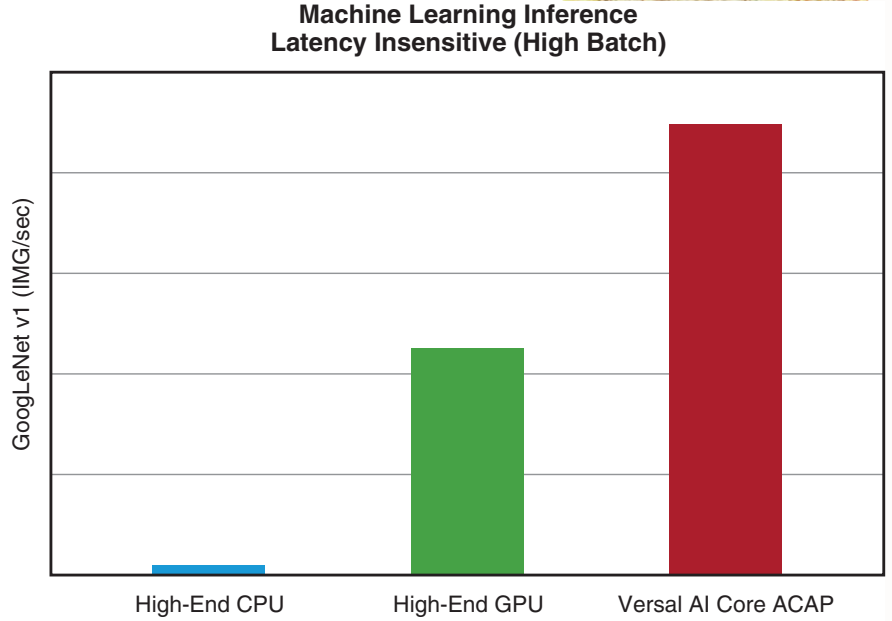
Fermilab

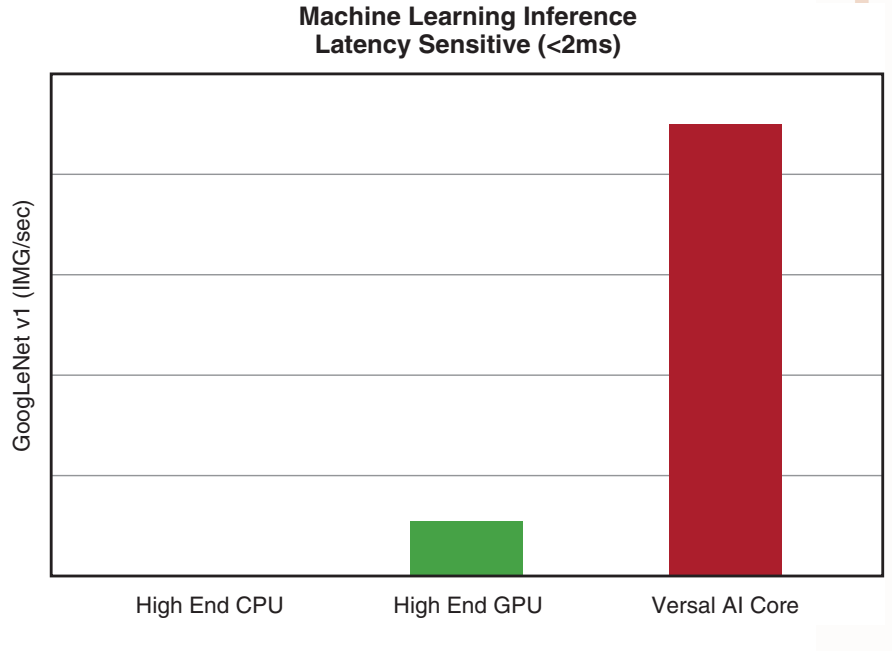
XILINX

WP505 (v1.0) October 2, 2018

Summary

- Exploiting new paradigms to improve and accelerate HL-LHC trigger algorithms with applications for the future computing model in HL-LHC
  - Algorithm acceleration with FPGAs programmed in C
  - High speed interconnect of computing elements
  - New adaptable hardware
- Strengthening connections and familiarity with new industry tools and technologies
- Developed techniques applicable in ATLAS, DUNE, Accelerator controls, and more (with interested collaborators)





XILINX announced (end of 2018) their 7nm technology (Versal) which does parallel, sequential processing and AI. White Paper: Versal ACAPs (https://bit.ly/2IZf1BS)