FAST INFERENCE ON FPGAs FOR HEP TRIGGER SYSTEMS

OPENLAB SUMMER STUDENTS LIGHTNING TALK

UMAR KHAYYAM

16 / 08 / 2018

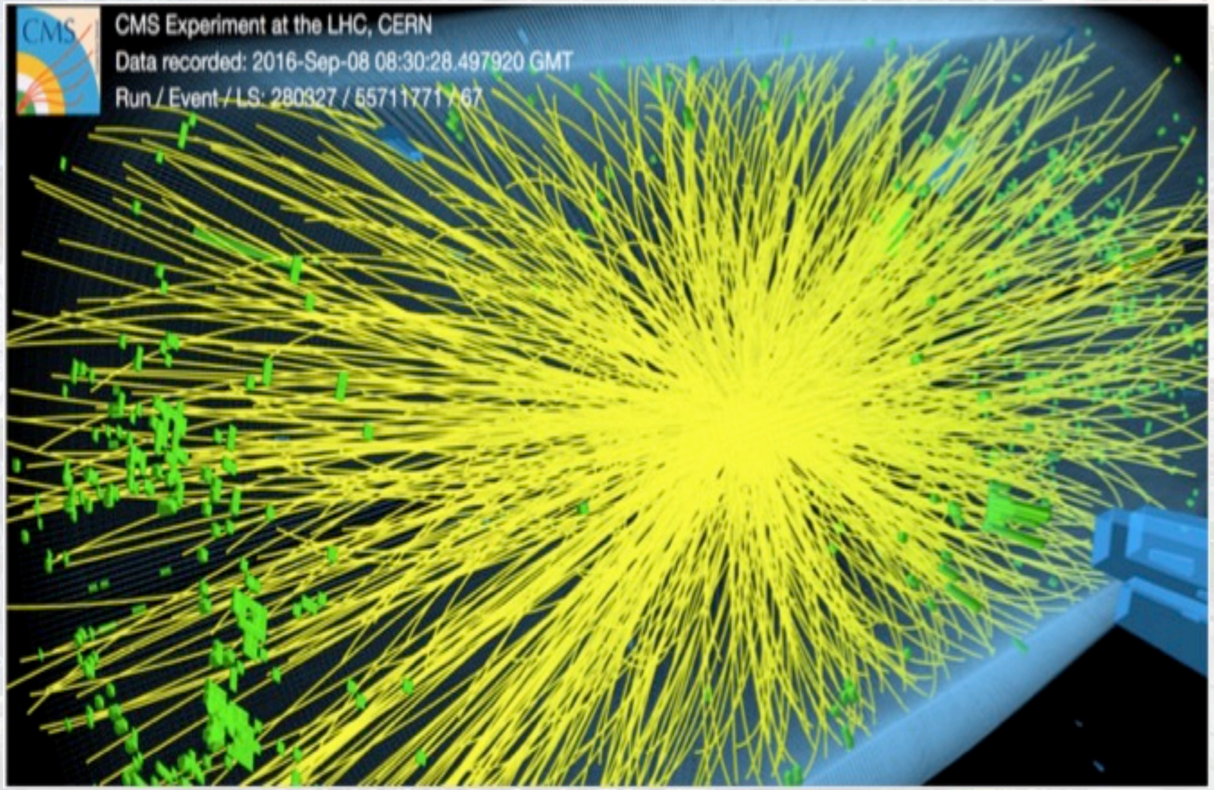
CHALLENGES OF TRIGGERING AT LHC

✓ Bunch Crossing frequency = 40MHz
✓ Data rates = 100 Tb/Sec

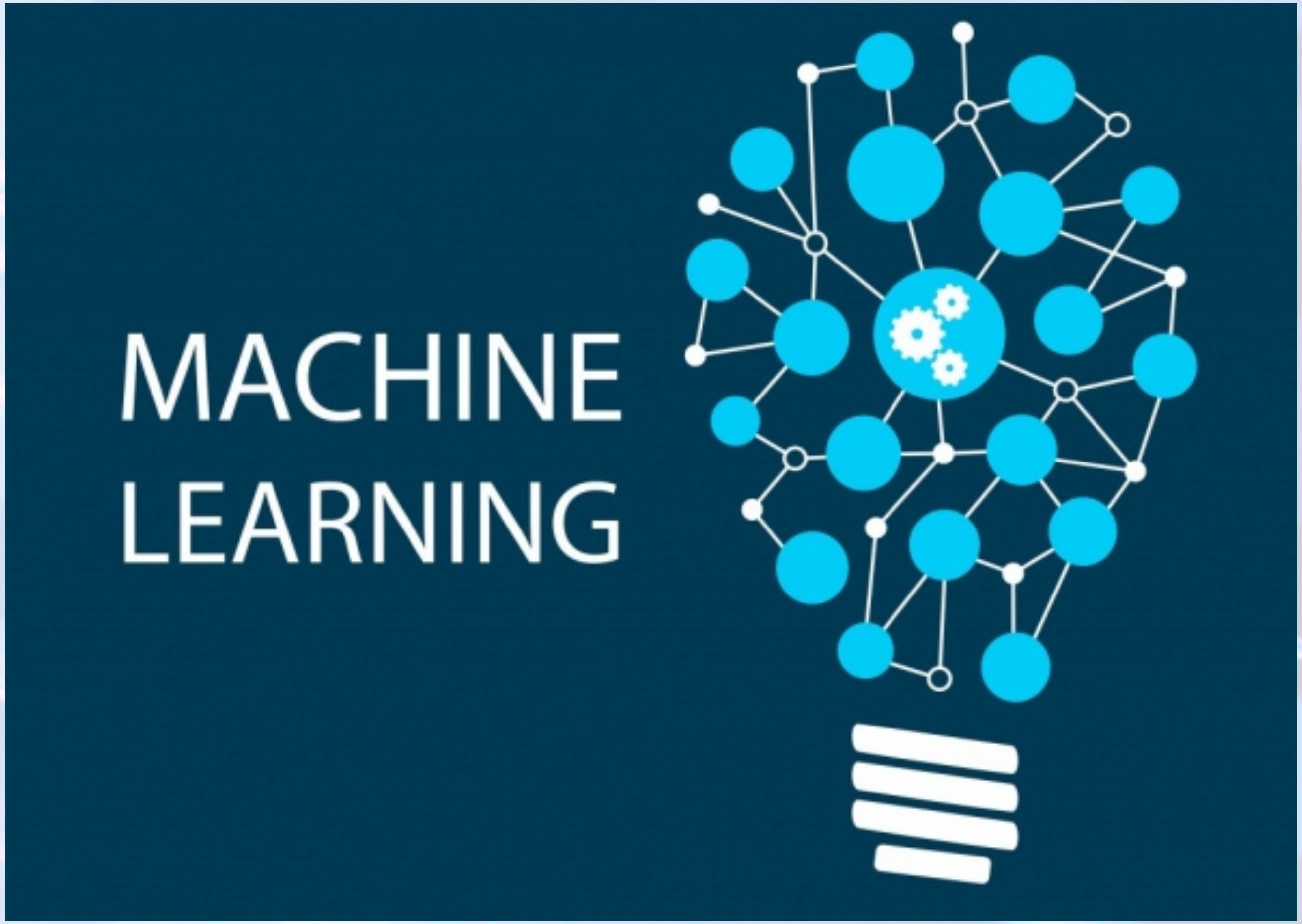
Triggering : Filter events to reduce data rates to manageable levels

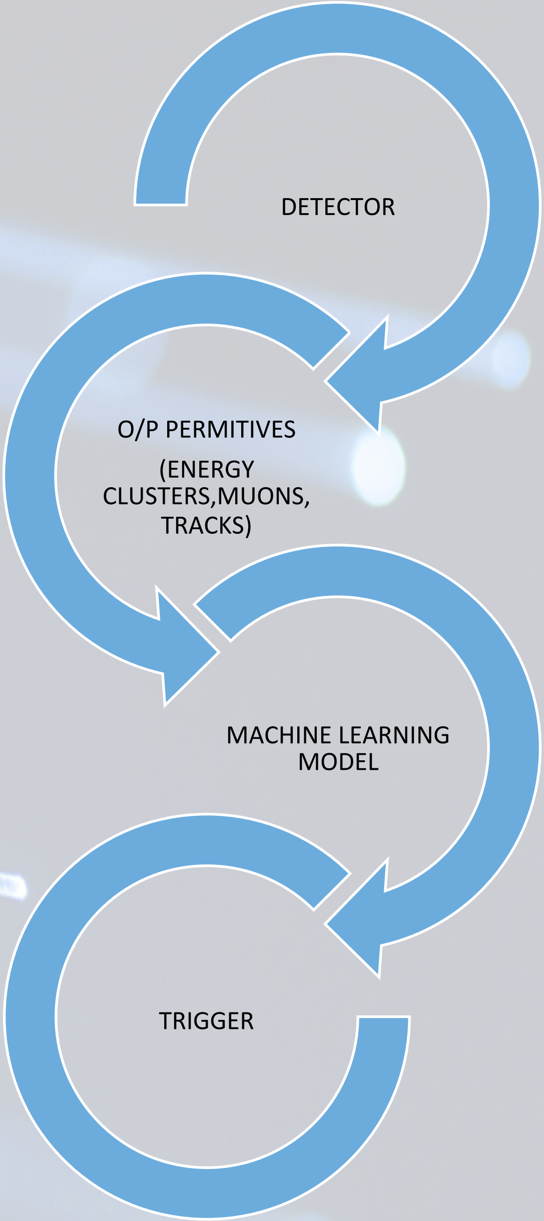
Challenge : To maintain physics in increasingly complex environment

- ➢ Un-triggered event will lost forever


CMS
CMS Experiment at the LHC, CERN
Data recorded: 2016-Sep-08 08:30:28.497920 GMT
Run / Event / LS: 280327 / 55711771 / 67

Solution?





DETECTOR

O/P PERMITIVES

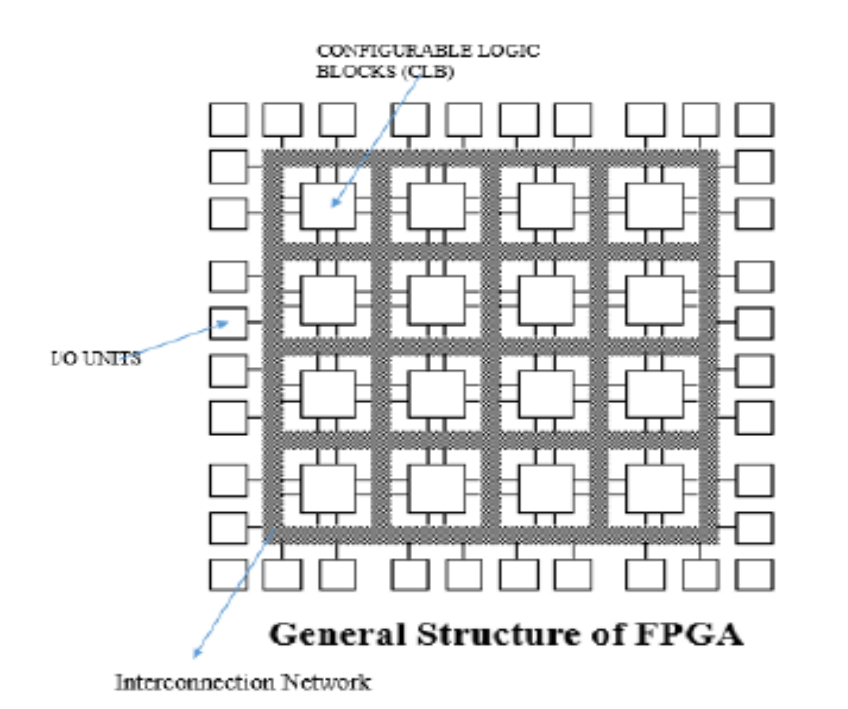
(ENERGY CLUSTERS,MUONS, TRACKS)

MACHINE LEARNING MODEL

TRIGGER

ML & FPGAs

- *All ML-Algorithms methods are typically deployed offline analysis.*
- *Low Latency – Real Time implementation just begun.*
- *This can be achieved by embedded devices called FPGA's*

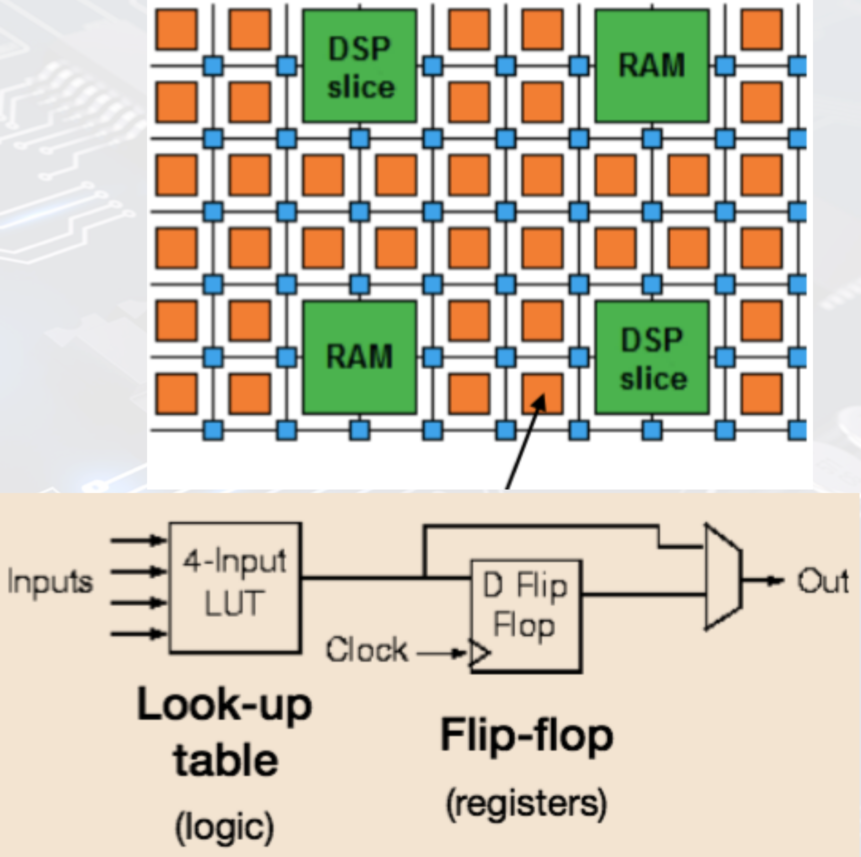


General Structure of FPGA

Field Programmable Gate Arrays (FPGA)

Contain array of logic cells used to configure low level operations (bit masking, shifting, addition)

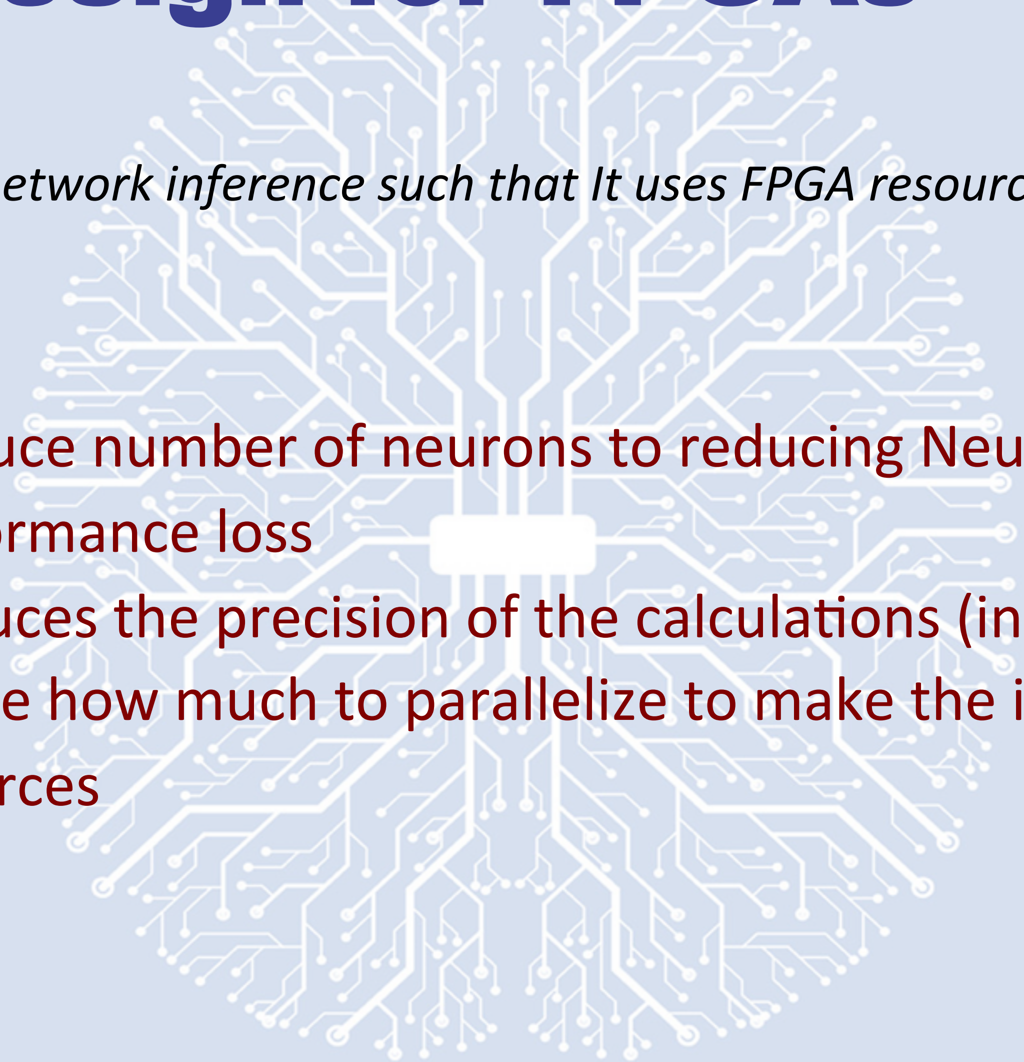
- *Parallelized & Pipelined Implementation.*
- *Low power consumption.*



How Do We Program them??

- *Typical way: Hardware Descriptive Language (HDL)*
- *New way: High Level Synthesis (HLS)*

Implementation of efficient neural network design for FPGAs



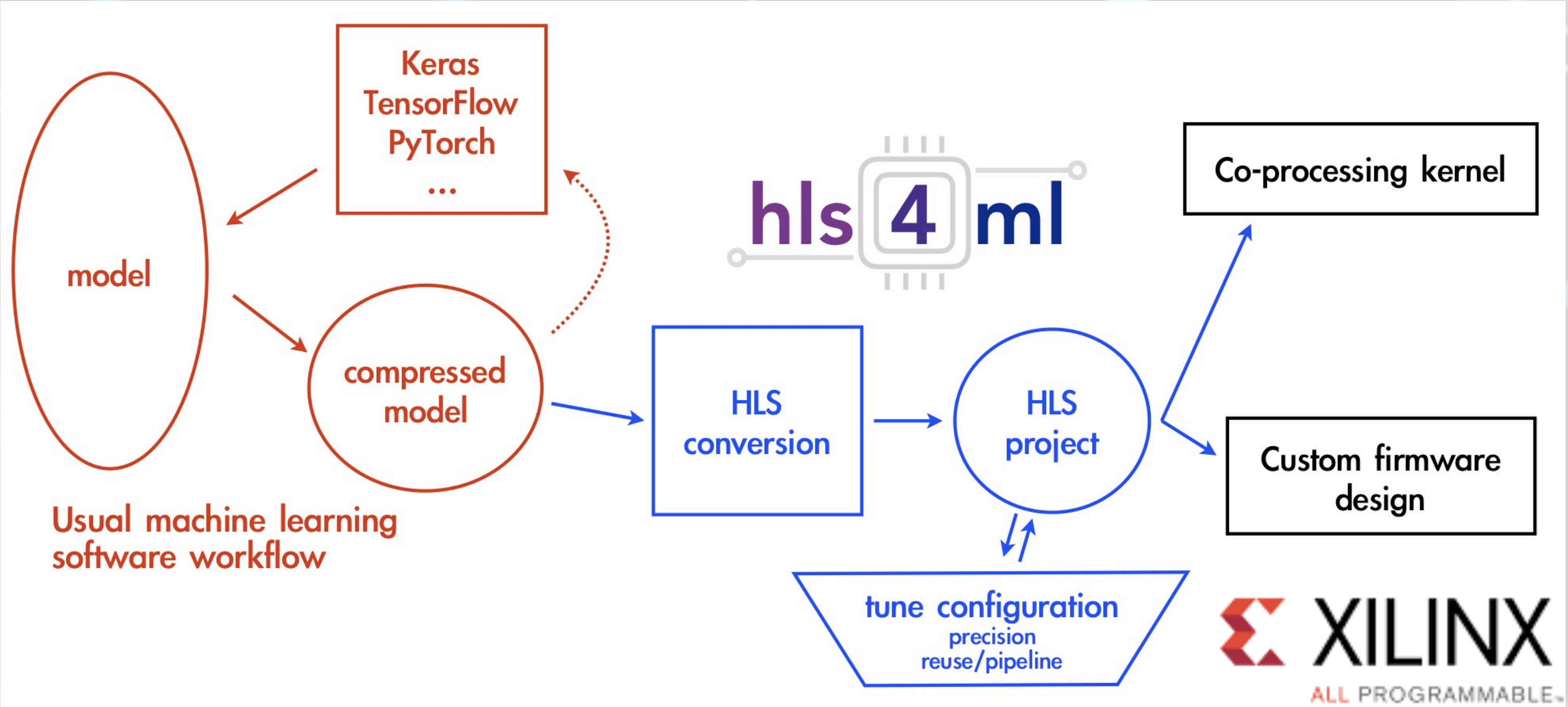
Focus is to tuning neural network inference such that It uses FPGA resources efficiently without having performance loss.

Three handling rules:

- Compression: Reduce number of neurons to reducing Neural multiplication w/o suffering any performance loss
- Quantization: Reduces the precision of the calculations (inputs, weights, biases)
- Parallelization: tune how much to parallelize to make the inference faster/slower versus FPGA resources

High Level Synthesis for Machine Learning

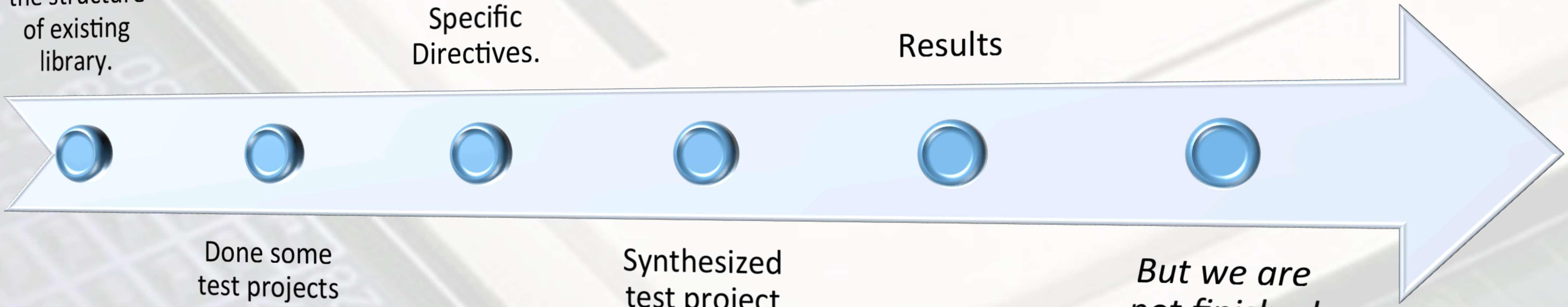
Compression, Quantization, and Parallelization made easy in hls4ml



Only comaptible with Xilinx Vivado HLS,

Intel Altera Quartus HLS

Task was to translating of HLS4ML to work with ALTERA QUARTUS HLS



Understood the structure of existing library.

Done some test projects on Vivado – HLS.

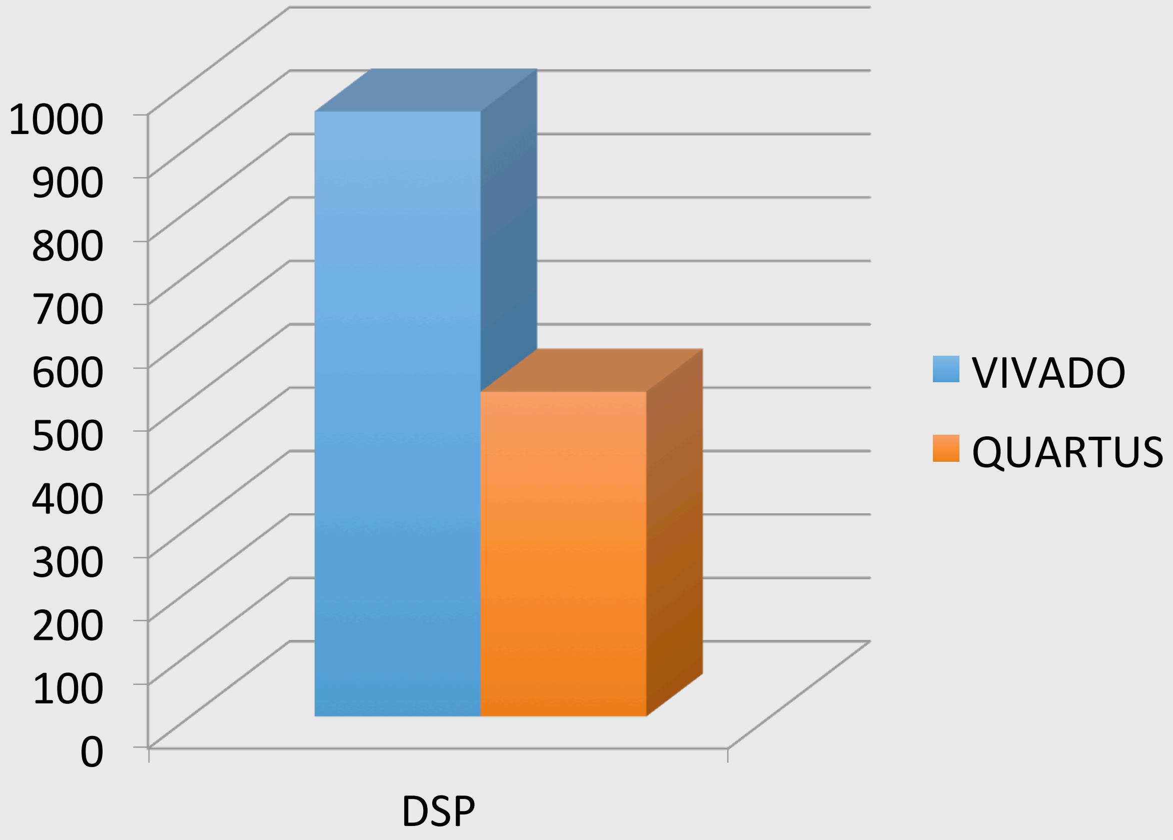
Changed Xilinix Specific Directives.

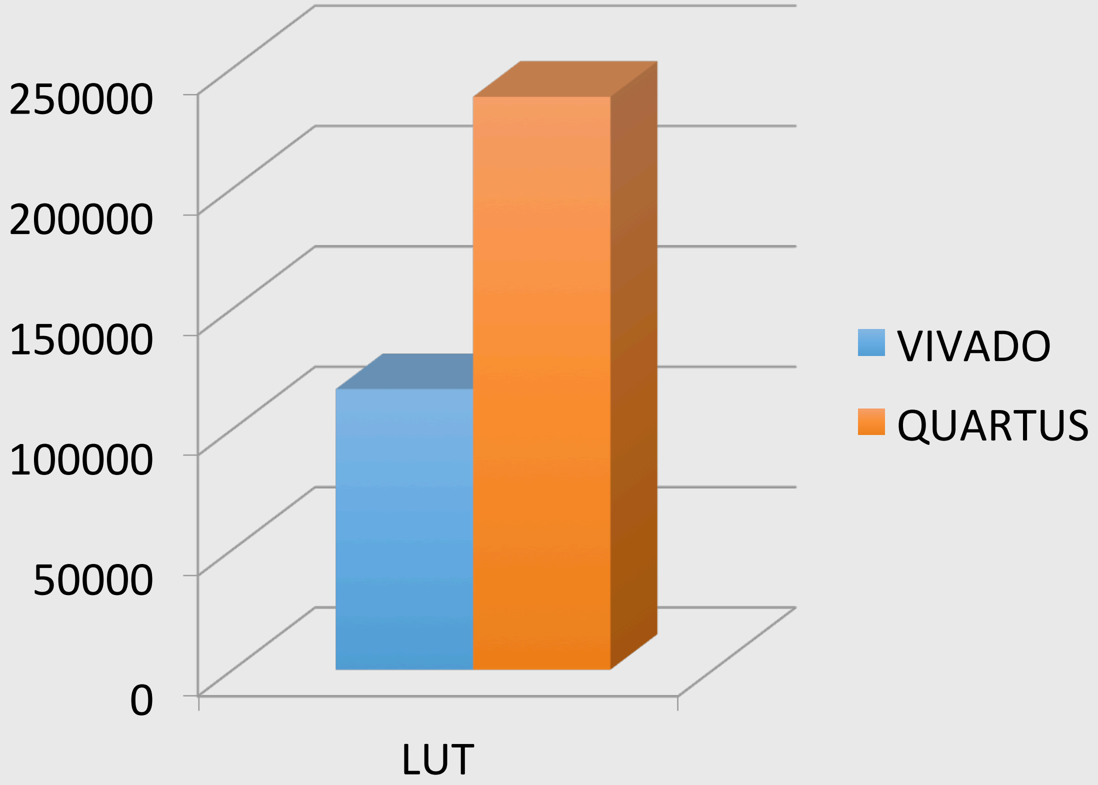
Synthesized test project on Quartus 14.

Results

But we are not finished yet, The project is still under progress.

Results up till now from SIMULATIONS





FUTURE WORK

- INTEL ARRIA 10 CARD JUST REACHED CERN
- IMPLEMENTATION OF NEW LIBRARY ON CARD
- RESULTS

THANK YOU

Special Thanks to My Supervisors

Jennifer Ngadiuba – Maurizio Pierini

CERN OPENLAB Organizers

Summer Fellows