Fast Inference of **ML** on **FPGAs** for **HEP** Trigger Systems

Hamza Javed

Supervisors : Maurizio Perini, Jennifer Ngadiuba, Vladimir Loncar

CERN openlab

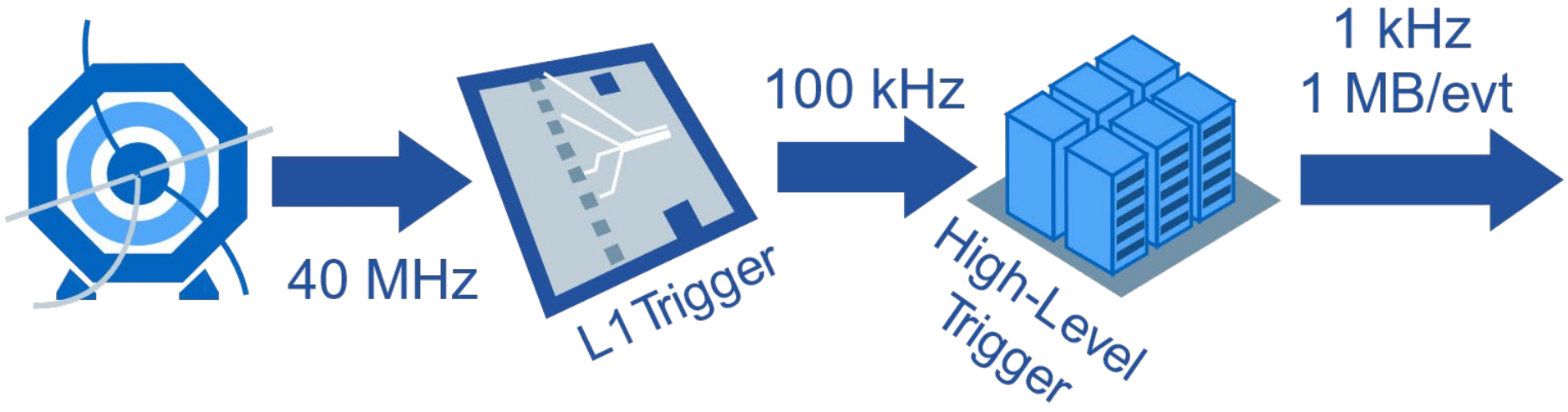
CERN

CMS

intel

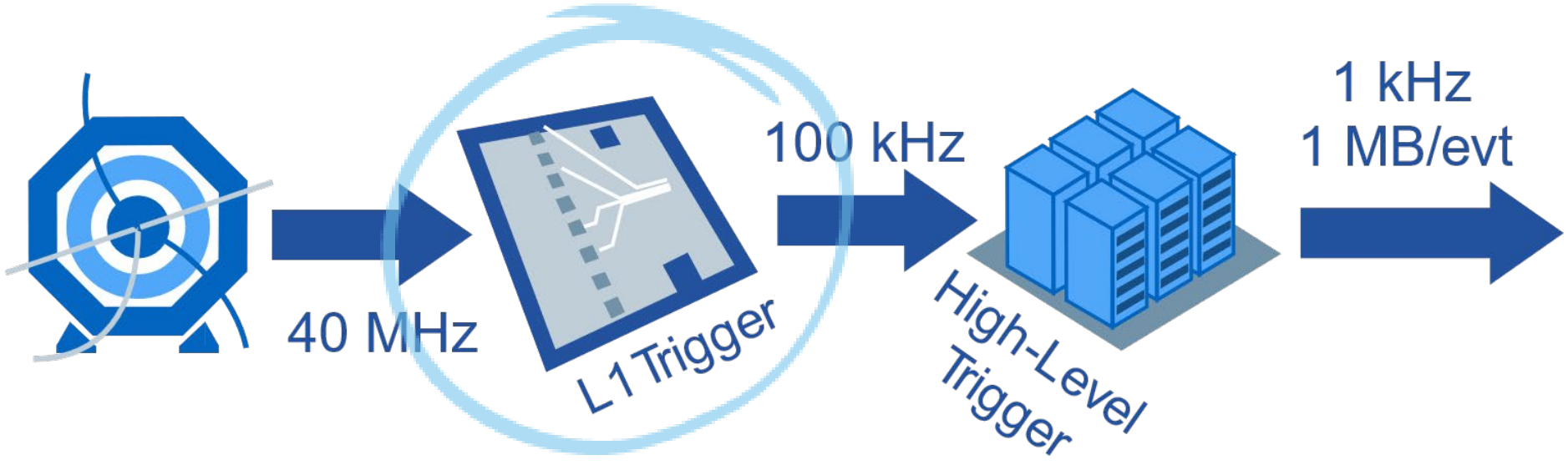
hls4ml

Problem



- 1 collision every 25 nanoseconds
- 99.99975% of data has to be rejected

Why FPGAs?

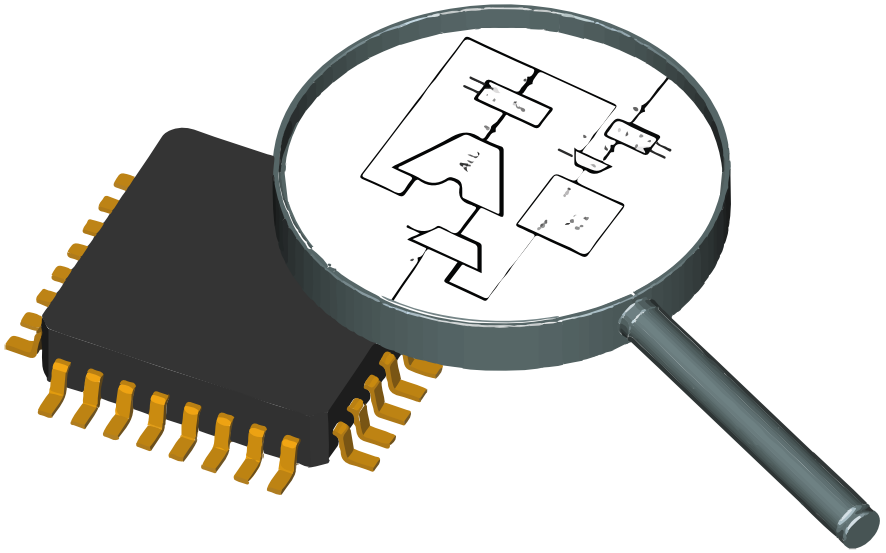


- During First Trigger
  - 99.75% rejected
  - Decision in ~**10 µs**

FPGA's are fast and massively parallelizable

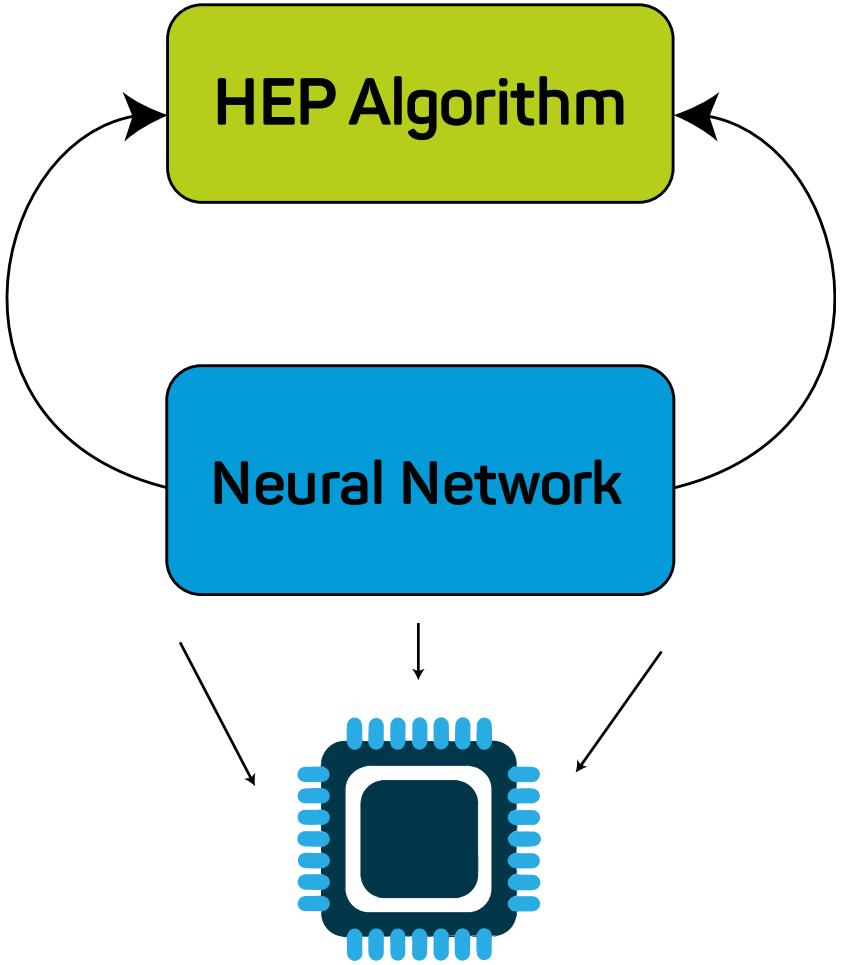
What are FPGAs?

- It’s a whiteboard for hardware
- Programmable Circuits
- Can emulate any logic



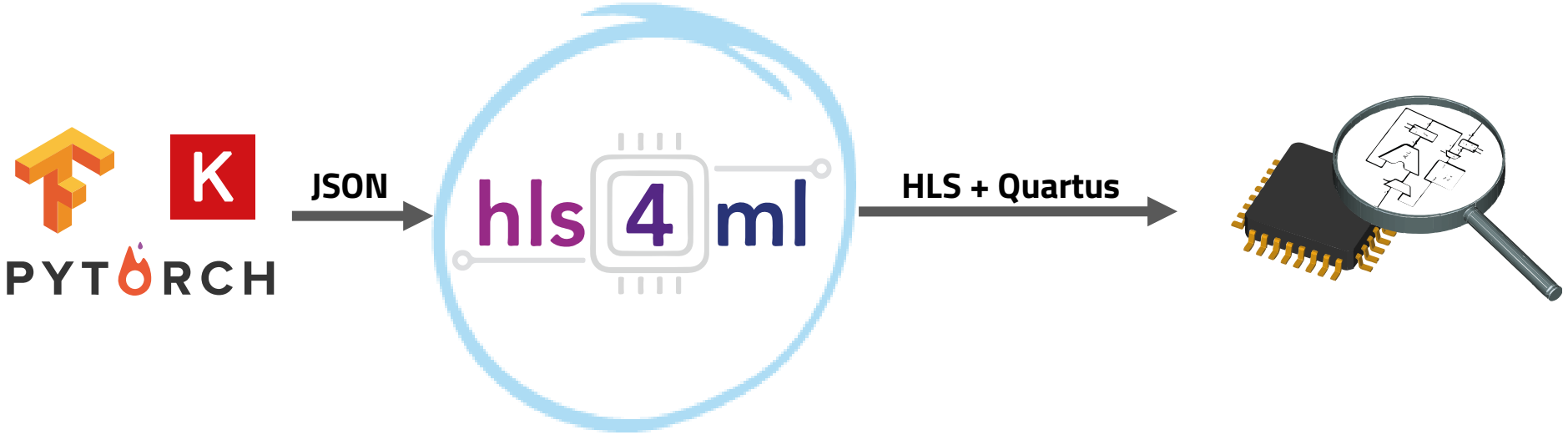
Why Machine Learning?

- Classical HEP algorithms for triggers are accurate.
- But, Machine Learning algorithms are faster and can be parallelized
- Is there a way to combine the two?



Fast inference of DNN on FPGAs for L1 Systems

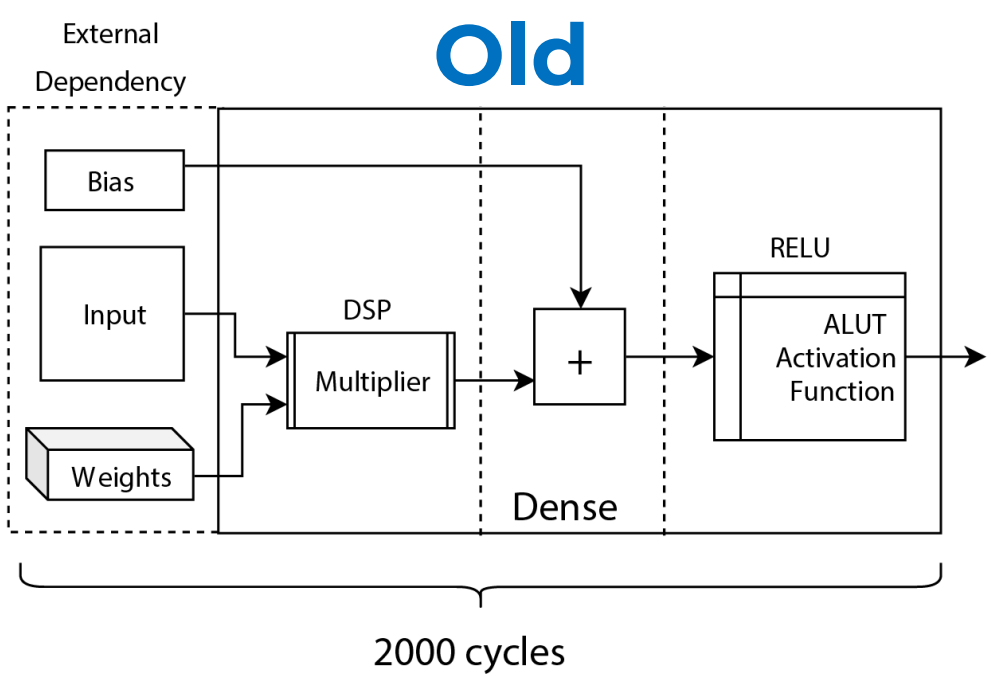
Approach



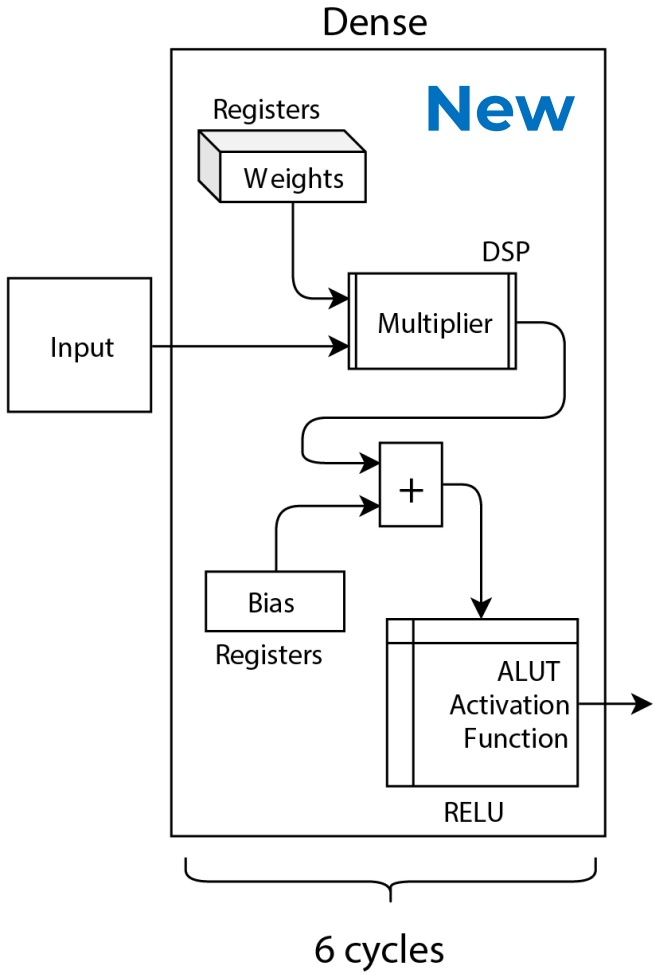
- Model can be trained in any library
- Converted and Optimzied by **hls4ml**
- Deployed on an FPGA

[1] Javier Duarte, Song Han, Philip Harris, Fast inference of deep neural networks in FPGAs for particle physics, 2018.

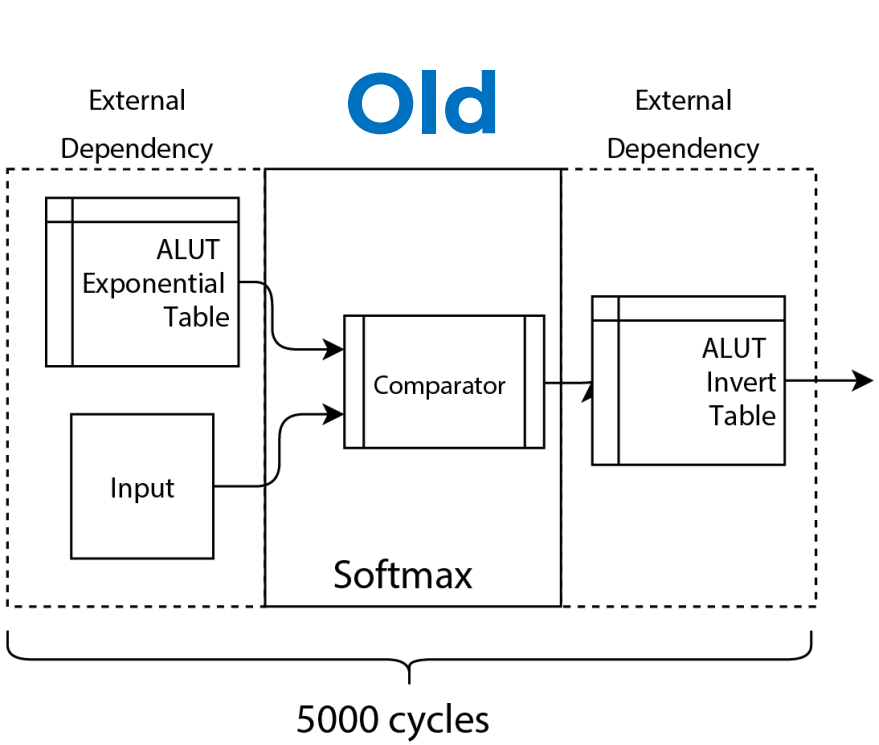
Architecture

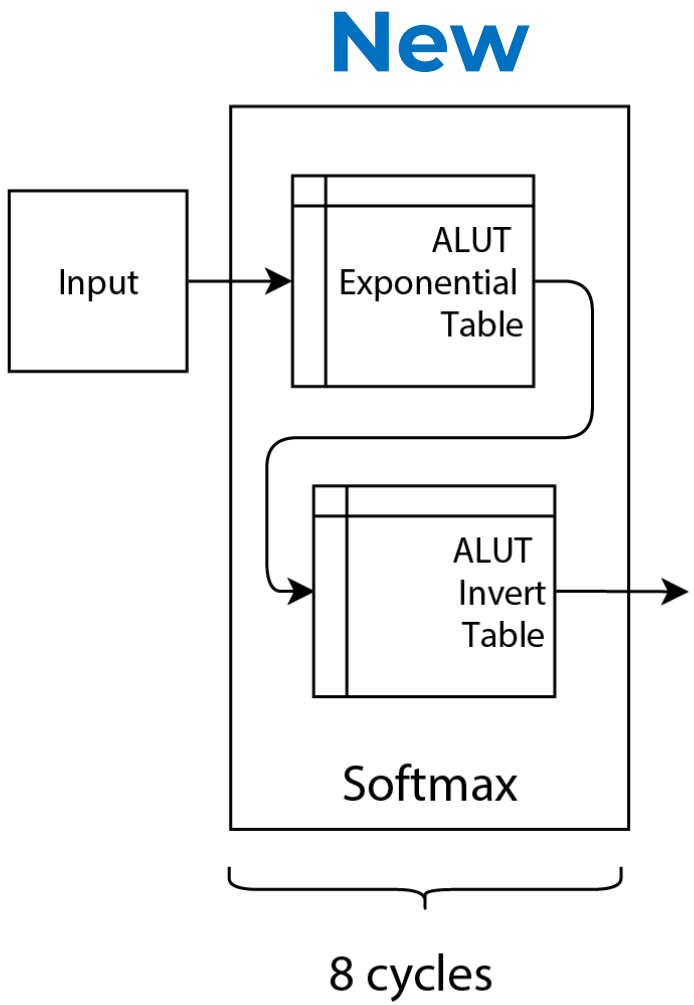


300x Faster



Architecture

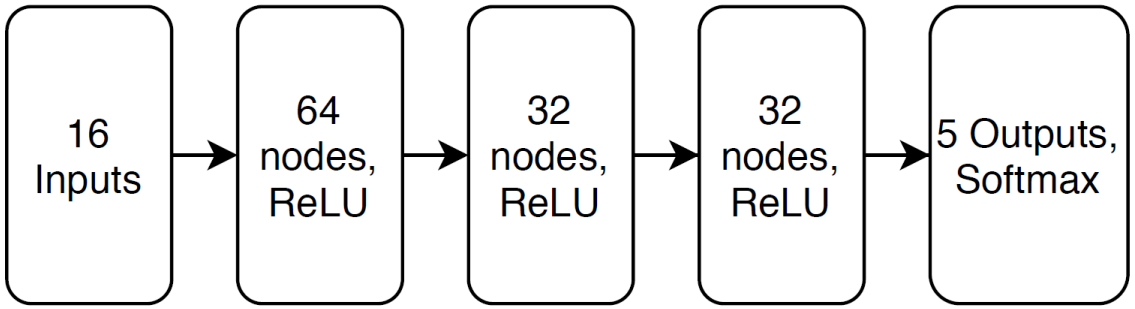


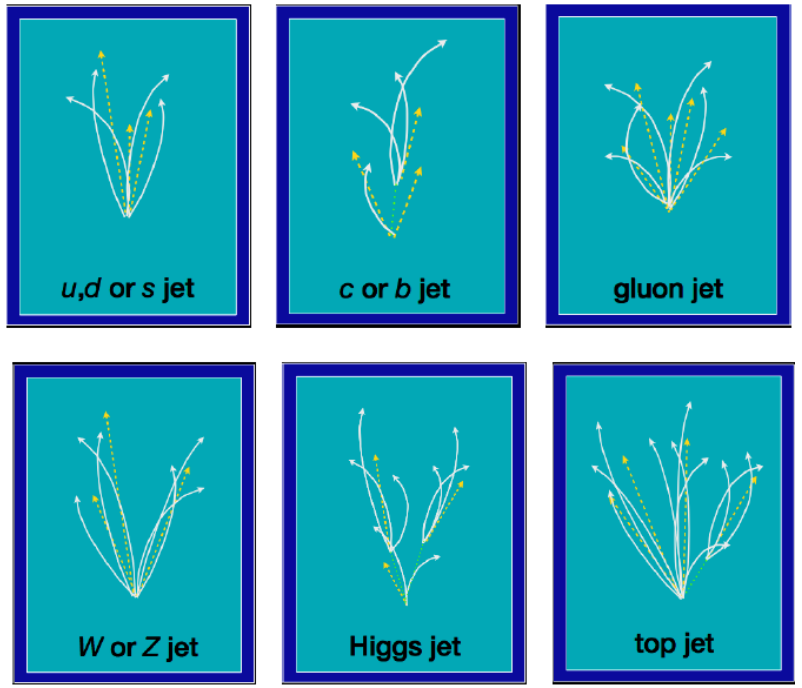


600x Faster

Test Case

- A Deep Neural Network trained for jet classification
- Five outputs corresponding to each of the **g**, **q**, **w**, **z** & **t** jet.





[1] Javier Duarte, Song Han, Philip Harris, Fast inference of deep neural networks in FPGAs for particle physics, 2018.

Results

Verification Statistics

	Invocations	Latency (min,max,avg)	II (min,max,avg)	Details
myproject	10	54,54,54	1,1,1	Click for details
Explicit component invocations	0	n/a,n/a,n/a	n/a,n/a,n/a	Click for details
Enqueued component invocations	10	54,54,54	1,1,1	Click for details

54 x 4ns = **220ns** latency (50 times lower)

Throughput of **250M events/sec**

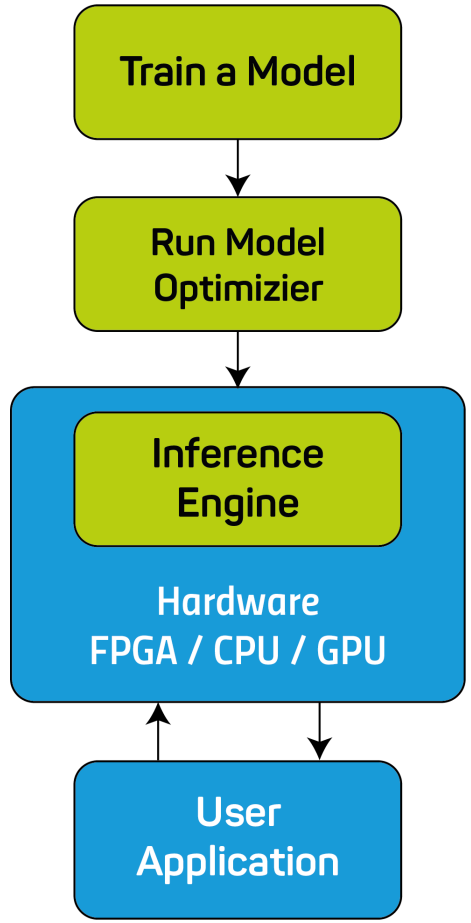
Accelerated ML for HLT Triggers using Openvino

What is Openvino?

- It’s a toolkit for deploying neural networks across multiple intel platforms.
- Highly scalable and cross platform compatible.

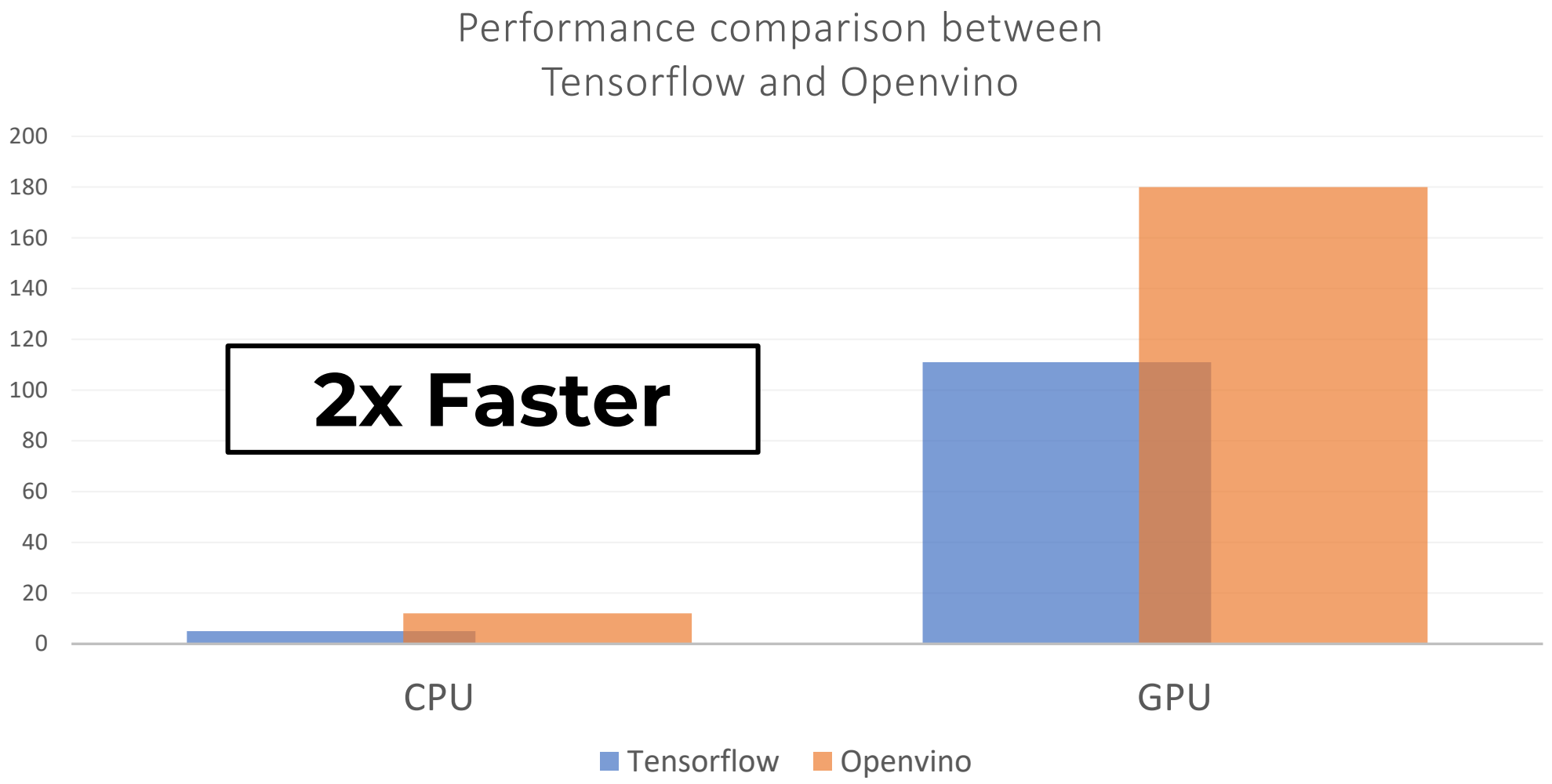
Accelerated ML for HLT Trigger

- Custom Resnet-50 model used for identifying **top quark jets**
- Optimized using Openvino.
- Tested on FPGA, GPU and CPU



[1] Javier Duarte, Philip Harris, Scott Hauck, FPGA-accelerated mache learning inference as a service for particle physics computing, 2019.

Accelerated ML for HLT Trigger



Accelerated ML for HLT Trigger

Type	Hardware	Accuracy	Inference Time	Max Throughput	Power	Throughput/W
CPU	2 x Xeon 2.1Ghz, 8 core	0.91	88 ms	11.36 img/s	~30W	0.38
FPGA	Arria 10 PAC - FP16	0.91	14 ms	84.44 img/s	~31W	2.72
	Arria 10 PAC - FP11	**0.88**	**7.5 ms**	**187.21 img/s**	**~31W**	**6.03**
GPU	GTX 1080	0.90	7.5 ms	192 img/s	~180W	1.06

- FPGA is 3-6x more efficient than a GPU

Acknowledgements

- Maurizio Perini, Jennifer Ngadiuba, Vladimir Loncar

Thank You!

hjaved@cern.ch

Questions?