## Introduction

The Large Hadron Collider(LHC) at CERN will go thorugh an upgrade (HL-LHC) to increase the rate of proton collisions, allowing experiments to collect one order of magnitude more data. This will demand a more efficient real-time event filter and this study shows how to perform jet classification on field-programmable gate arrays (FPGA) within O(100) ns. We compare Deep Sets and Interaction Network models, which are permutation-invariant, with MLP, which is not not permutation-invariant. Through quantization-aware training (QAT) and efficient FPGA implementations, we show that nanosecond inference using complex architectures like Deep Sets and Interaction Networks are feasible at low resource-cost.



ln tl consi gluoi N =with const

	Quantized Mod	els Perfo	rmand	ce							
odels are implemented using and Tensorflow libraries. For DS the input jet data is structured nconnected graph, while for it's structured as a fully ted graph. Each graph node is ited to a jet constituent and its es.	Quantization aware tra- implemented using the 50% sparsity is applied model so it can fit with constrains and for con- sparsity is applied to the models. Plots shows the terms of inverse averages a function of the bitwite constituents.	aining (QAT e QKeras lib to the 32-ce nin the FPG sistency, the he 32-consti ne models pe ge FPR (Fals dth and the	) of each rary. Pri onstitue A resour e same p tuent <i>M</i> erformat e Positiv number	n model is uning to nt IN ce oruning LP and DS nce in ve Rate) as			- MLP - DS - IN 12 14 16 bitwidth	4 0 16 const 16 const 0 0 10 5.188±0.152 7.263±0.408 10 5.294±0.106 6.301±0.622 0 0 0 0 0 0 0 0 0 0 0 0 0	32 const 12.114±1.155 12.091±0.822 5.641±0.101		
8 Jets	Results										
Pt     n       1 <sup>st</sup> constituent	The models described above are translated into firmware using HLS4ML library [3], then synthesized with Vivado HLS, targeting Virtex UltraScale+ VU13P FPGA with a clock frequency of 200 MHz. Number of jet constituents, reuse factor(RF), latency, initiali interval (II) and resource consumption for the models quantized to 8 bits are shown in the table. The cc next to the latency and II represents the number of clock cycles on the FPGA. '										
	Architecture	Constituen	ts RF L	atency [ns] (co	c) II [ns] (cc)	DSP	LUT	FF	BRAM18		
		8	1	105 (21)	5 (1)	262 (2.1%)	155,080 (9.0%)	25,714 (0.7%)	4 (0.1%)		
blishing the baseline performance	MLP	16	1	100 (20)	5 (1)	226 (1.8%)	146,515 (8.5%)	31,426 (0.9%)	4 (0.1%)		
e bellow shows the model		32.	1	105 (21)	5 (1)	262 (2.1%)	155,080 (7.2%)	25,714 (0.7%)	4 (0.1%)		
inties on the AUCs are all $\sim 0.001$		8	2	95 (19)	15 (3)	626 (5.1%)	386,294 (22.3%)	121,424 (3.5%)	4 (0.1%)		
	DS	16	4	115 (23)	15 (3)	555 (4.5%)	747,374 (43.2%)	238,798 (6.9%)	4 (0.1%)		
		32.	8	130 (26)	10 (2)	434 (3.5%)	903,284 (52.3%)	358,754 (10.4%)	4 (0.1%)		
$\frac{1}{2}$		8	2	160 (32)	15 (3)	2,191 (17.8%)	472,140 (27.3%)	191,802 (5.5%)	12 (0.2%)		
+0.1% 0.84 0.88 0.90 0.88 0.92	IN	16	4	180 (36)	15 (3)	5,362 (43.6%)	1,387,923 (80.3%)	594,039 (17.2%)	52 (1.9%)		
+ 0.3% 0.84 0.88 0.90 0.88 0.92		32.	8	205 (41)	15 (3)	2,120 (17.3%)	1,162,104 (67.3%)	/61,061 (22.0%)	132 (2.5%)		
+ 0.2% 0.84 0.88 0.91 0.89 0.92											
$\frac{20.2\%}{0.87}$ 0.87 0.89 0.91 0.90 0.94	<b>References:</b>										
$\pm 0.2\%$ 0.87 0.89 0.93 0.92 0.94											
± 0.2% 0.88 0.90 0.94 0.92 0.94	<ol> <li>1) HLS4ML LHC Jet Dataset , https://doi.org/10.5281/zenodo.3601443 , J.M.Duarte and others, 2020</li> <li>2) QKeras - https://github.com/google/qkeras , C.Coelho, 2019</li> <li>2) HLS4ML - https://feature.chingles.ming.org/bla4ml, 2019</li> </ol>										
± 0.2% 0.90 0.89 0.89 0.88 0.94											
± 0.1% 0.91 0.91 0.96 0.95 0.95	4) Ultrafast let Classification	on on FPGA at	the HL-I	HC - arXiv:2402.0	)1876v2 [hep-ex]						
±0.3% 0.91 0.91 0.96 0.95 0.95	,										

Neural Netw	ork Archit	echtures			Qua	ntized Mode	ls Perfor	mance	e					
Multilayer Perceptron MLP $(p_1^{j})$ <td colspan="4">Wulliver Perceptron MLP <math>\downarrow \downarrow </math></td> <td colspan="10">Quantization aware training (QAT) of each model is implemented using the QKeras library. Pruning to 50% sparsity is applied to the 32-constituent IN model so it can fit within the FPGA resource constrains and for consistency, the same pruning sparsity is applied to the 32-constituent MLP and DS models. Plots shows the models performance in terms of inverse average FPR (False Positive Rate) as a function of the bitwidth and the number of constituents</td>	Wulliver Perceptron MLP $\downarrow \downarrow $				Quantization aware training (QAT) of each model is implemented using the QKeras library. Pruning to 50% sparsity is applied to the 32-constituent IN model so it can fit within the FPGA resource constrains and for consistency, the same pruning sparsity is applied to the 32-constituent MLP and DS models. Plots shows the models performance in terms of inverse average FPR (False Positive Rate) as a function of the bitwidth and the number of constituents									
Jataset	1 .1 11		[4]											
isting of jets from five different origins: quark (q), n (g), W boson, Z boson, and top (t) jets, with up to 50 jet particle constituents. We use scenarios of jets N = 8, 16, 32 constituents. We define the tituents features as $P_t$ , $\eta$ and $\phi$ relative to the jet axis.			<b>Results</b> The models described above are translated into firmware using HLS4ML library [3], then synthesized with Vivado HLS, targetin Virtex UltraScale+ VU13P FPGA with a clock frequency of 200 MHz. Number of jet constituents, reuse factor(RF), latency, initial interval (II) and resource consumption for the models quantized to 8 bits are shown in the table. The cc next to the latency and represents the number of clock cycles on the FPGA. '											
					•	Architecture C	Constituent	s RF La	atency [ns] (co	c) II [ns] (cc)	DSP	LUT	FF	BRAM18
Full Precisio	n Models F	Performa	nce				8	1	105 (21)	5 (1)	262 (2.1%)	155,080 (9.0%)	25,714 (0.7%)	4 (0.1%)
ch model is first trained at floating-point precision, establishing the baseline performance at the quantised model on FPGA should match. The table bellow shows the model rformance for 8, 16, and 32 jet constituents. The uncertainties on the AUCs are all ~ 0.001 d thus not included for legibility.				MLP	16 32.	1 1	100 (20) 105 (21)	5 (1) 5 (1)	226 (1.8%) 262 (2.1%)	146,515 (8.5%) 155,080 (7.2%)	31,426 (0.9%) 25,714 (0.7%)	4 (0.1%) 4 (0.1%)		
				DS	8 16	2 4	95 (19) 115 (23)	15 (3) 15 (3)	626 (5.1%) 555 (4.5%)	386,294 (22.3%) 747,374 (43.2%)	121,424 (3.5%) 238,798 (6.9%)	4 (0.1%) 4 (0.1%)		
	0	1		AUC			32.	8	130 (26)	$\frac{10(2)}{15(2)}$	$\frac{434(3.5\%)}{2101(17.8\%)}$	903,284 (52.3%)	358,754 (10.4%)	$\frac{4(0.1\%)}{12(0.2\%)}$
Architecture Constituents Parameters FLOPs Accuracy $g = a = W = Z = t$			q W Z t		INI	δ 16	Ζ Λ	160(32)	15(3)	2,191(17.8%) 5 362 (43.6%)	4/2,140(2/.3%) 1 387 023 (80 397)	191,802(5.5%) 594 039 (17 29)	12(0.2%) 52(1.0%)	
MLP		26,826	53,162 $64.6 \pm 0.1\%$ 0.84	4 0.88 0.90 0.88 0.92		IIN	32.	4	205 (41)	15(3)	2.120 (17.3%)	1.162.104 (67.3%)	761.061 (22.0%)	132 (2.5%)
DS	8	3,461	$36,805  64.0 \pm 0.3\%  0.84$	4 0.88 0.90 0.88 0.92				0	200 (11)		_,	1,102,101 (07.070)	/ 01,001 (22.070)	
IN		3,347	$37,232  64.9 \pm 0.2\%  0.84$	4 0.88 0.91 0.89 0.92										
MLP		20,245	40,485 $68.4 \pm 0.3\%$ 0.8	7 0.89 0.91 0.90 0.94	Refe	rences:								
DS	16	3,461	71,109 $69.4 \pm 0.2\%$ 0.8	7 0.89 0.93 0.92 0.94	1) HLS4ML LHC Jet Dataset , https://doi.org/10.5281/zenodo.3601443 , J.M.Duarte and others, 2020									
		3,347	$\frac{140,432}{48,107}  \frac{10.8 \pm 0.2\%}{66.2 \pm 0.2\%}  0.8\%$	8 0.90 0.94 0.92 0.94	2) <i>QKeras</i> - https://github.com/google/qkeras , C.Coelho, 2019									
DS	32	3 461	$139.717 759 \pm 0.1\% 0.90$	1 0.91 0.96 0.95 0.94	3) <i>HLS4ML</i> - https://fastmachinelearning.org/hls4ml , 2018 4) Ultrafast Jet Classification on FPGA at the HL-LHC - arXiv:2402.01876v2 [hep-ex]									
IN	02	7,400	$109,556 \ 75.8 \pm 0.3\% \ 0.9$	1 0.91 0.96 0.95 0.95										
		,	,											

### Conclusions

Neural network based jet classification algorithms are synthesized on FPGA that mimic the environment within the hardware layers of the real-time data processing systems for a typical HL-LHC experiment. Using jet data with constituent level information, we show how one could synthesize machine learning algorithms pertaining to three different data representations on an FPGA by using the hls4ml library. Deep Sets network strikes a good balance between accuracy, latency, and resource consumption compared with the deployed and tested MLP and IN models. In conclusion, we have identified and shown the necessary ingredients to deploy a jet classifier in the level-1 trigger of the HL-LHC

# Andre Sznajder (UERJ)

in collaboration with: P.Odagiu (ETH), Z.Que(ICL), J.Duarte(UCSD), J.Haller(UH), G.Kasieczka(UH), A.Lobanov(UH), V.Loncar(MIT), W.Luk(ICL), J.Ngadiuba(FNAL), M.Pierini(CERN), P.Rincke(Uppsala), A.Seksaria(Uppsala), S.Summers(CERN), A.Tapper(ICL), T.K.Arrestad(ETH)



g a Xilinx ization

