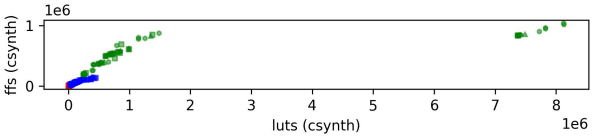
Comments About the Paper

From circulation Aug 3

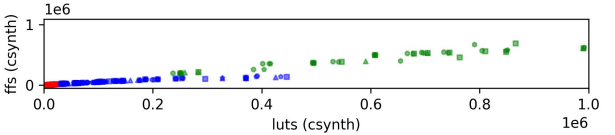
Minor Comments

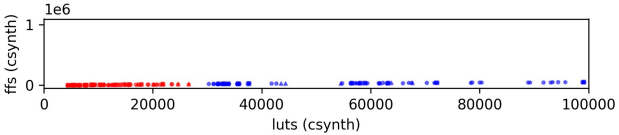
- Typos to fix
- Some changes to figures to make more readable
- Increase text size and make figure 5 easier to read
- For figure 7, Jennet wants us to comment if model 2 is similar -> we can say so once we have done model 2 vsynth

Major Comment - FFs + LUTs



think it's fine to use ff+luts because it looks like they increase at the same r





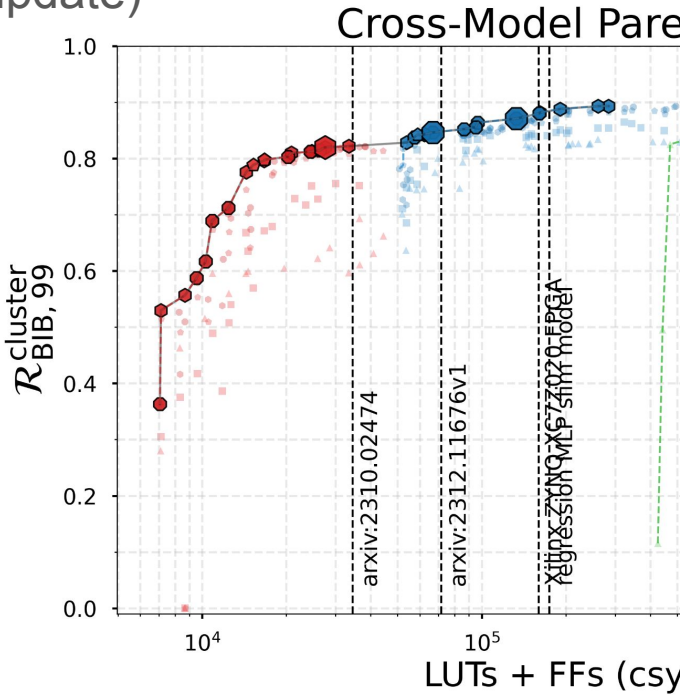
- Concern about adding FF and LUTs together
- “(2) It is questionable to combine LUTs and FFs into a single hardware-resource metric, as these resources are not interchangeable.” - Zhiqiang Que
  - Also he thinks we should use HGQ and da4ml. That’s not changing for this submission
- My response is that figure 8 (cross pareto front) is anyway an **estimate** to select feasible models
- Perhaps justify with paper from Lino?
- Give version of cross pareto where x axis is just LUTs or just FFs in appendix?
- Giuseppe thinks can justify with https://docs.amd.com/r/en-US/ug474_7Series_CLB/CLB-Overview --> AMD 7-series and UltraScale logic blocks provide approximately two FFs per LUT
- we could have a = 1, b = 0.5 and so 1 * LUTs + 0.5 * FFs

Major Comment/Change: new regression model numbers

- Giuseppe gave FF, LUT for the updated regression models, and they higher than original arxiv , currently by a factor of 3 (will update)
- Should we select a bigger model?
- But our ASIC area estimates are way higher than what's in their paper…

	LUT	6,505	31,842	111,948	492,029
FPGA	FF	2,208	13,495	47,202	201,191
Resource	DSP	0	0	0	0
Usage	BRAM	0	0	0	0
(vsynth)	Clock Period [ns]	9.275	9.275	9.275	8.745
	Target Clock [ns]	10	10 ns	10	10
	Latency [ns]	160	160	200	3130*
ASIC	Area Score [mm^2]	0.140	0.586	1.900	8.394
Resources	Clock Period [ns]	10	10	10	10
Estimates	Latency [ns]	200	250	260	19,460

Table 2: Comparison including quantization, background rejection, FPGA vsynth resources, and ASIC metrics of four representative configurations.



Old vs. new regression model

Old regression model estimates

QKeras Model Analysis		Alveo U250 Synthesis		ASIC Synthesis (45nm)	
4-bit ops.	50,140	Clock Period	5 ns	Clock Period	5 ns
8-bit ops.	3,040	BRAM_18K	12.5	Area Estimate	1.4 mm^2
1st Layer N_0	17	DSP48E	0	Buffer Area	0.0017 mm^2
2nd Layer N_0	9	FF	14,289	Inverter Area	0.055 mm^2
3rd Layer N_0	2	LUT	57,398	Logic Area	0.80 mm^2
4th Layer N_0	117	URAM	0	Sequential Area	0.52 mm^2
5th Layer N_0	15	Latency	1.46 μs	Latency	27 μs
6th Layer N_0	51	Interval (II)	1.38 μs	-	-

New regression models:

Model Architecture		Model 1	Model 2		Model 3
Weight Quantization		8-bit	10-bit	10-bit	10-bit
Number of Parameters		265	711	2,637	14,241
Cluster Rejection Rate	$\mathcal{R}^{cluster}_{BIB,95}$	86.9 %	88.9 %	91.8 %	93.8 %
	$\mathcal{R}^{cluster}_{BIB,98}$	84.7 %	86.8 %	89.5 %	92.1 %
	$\mathcal{R}^{cluster}_{BIB,99}$	82.0 %	84.6 %	87.1 %	90.5 %
Data Reduction	$\mathcal{R}^{data}_{BIB,99}$	88.4 %	90.4 %	91.8 %	94.4 %
	$\mathcal{R}^{data}_{BIB,99} + [3\sigma_t, 5\sigma_t]$	98.5 %	98.7 %	98.9 %	99.3 %
FPGA Resource Usage (vsynth)	LUT	6,505	31,842	111,948	492,029
	FF	2,208	13,495	47,202	201,191
	DSP	0	0	0	0
	BRAM	0	0	0	0
	Clock Period [ns]	9.275	9.275	9.275	8.745
	Target Clock [ns]	10	10 ns	10	10
	Latency [ns]	160	160	200	3130*
ASIC Resources Estimates	Area Score [mm^2]	0.140	0.586	1.900	8.394
	Clock Period [ns]	10	10	10	10
	Latency [ns]	200	250	260	19,460

able 2: Comparison including quantization, background rejection, FPGA vsynth re-
urces and ASIC metrics of four representative configurations.

	Max models	Full models			Slim models		
	Conv2D	Conv2D	Conv1D	MLP	Conv2D	Conv1D	MLP
NN Parameters	1898	1767	2734	2264	1682	2649	2179
Latency / II [clk]	2/1	2/1	2/1	2/1	2/1	2/1	2/1
Slack [ns]	14.33	14.33	14.33	16.36	16.62	15.72	15.72
HLS area [mm^2]	0.6353	0.6197	0.6067	0.3004	0.2880	0.3487	0.3350

TABLE I: Model sizes, area, and timing results from HLS targeting TSMC 28 nm and 25 ns clock period.

BRAM	DSP	FF	LUT	URAM
0	1106	42503	131921	0

Giuseppe is updating these FF, LUT numbers targeting our same fpga we used for figure 8

Comments from Zhiqiang Que

Zhiqiang Que (ZQ) [1:55 PM]

Hi Daniel, some comments for your paper. Hope these help.

(1) The C-synthesis resource estimates can be quite inaccurate. V-synthesis would provide a more reliable estimate and may yield an LUT count approximately 30% lower.

(2) It is questionable to combine LUTs and FFs into a single hardware-resource metric, as these resources are not interchangeable. LUT utilisation is generally more likely to be the limiting factor on an FPGA. For Figure 8, reporting LUT usage alone would be more meaningful, although ideally LUTs and FFs should be reported separately.

(3) The paper should report the initiation interval (II) in addition to latency. I suspect that the II is 1, but this should be stated explicitly because latency alone does not indicate the achievable throughput.

(4) The ASIC-area statement in the conclusion is inconsistent with Table 2. The conclusion states that the areas of Models 1 and 2 are below (10^5 um^2), but none of the corresponding values in Table 2 is below this threshold.

(5) The role of QKeras is described inaccurately in the abstract. QKeras is used for quantisation-aware training, whereas the hardware estimates are obtained using hls4ml, Vitis HLS, and Catapult.

(6) The reported FPGA latency of 3130 ns for Model 3 is marked with an asterisk, but no corresponding footnote or explanation is provided. (edited)

Daniel Abadjiev [2:10 PM]