Cassia.ai

Cassia.ai: Better Power/Performance with Approximate Arithmetic for ML

Use **approximate arithmetic** from **Cassia.ai** to **increase performance and reduce power** in your ML accelerator **with no loss of accuracy**. Cassia.ai hyper-optimizes key operations (e.g. multiplication, GeLU, sigmoid, etc) with almost no loss in model-level accuracy.

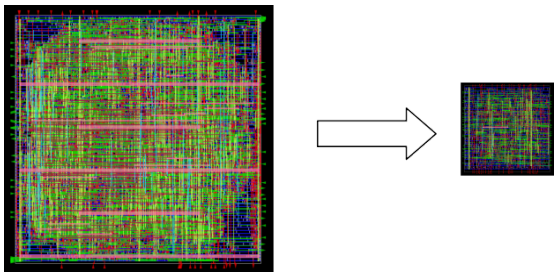

Figure 1: Layout comparison of standard FP32 multiplier with Cassia.ai CaFP32 multiplier in 45nm. We realize 6.85x improvement in TOPs/W and 18.6x improvement in TOPs/um2. Other quantization format (e.g. FP8, BF16, FP16, etc. are possible).

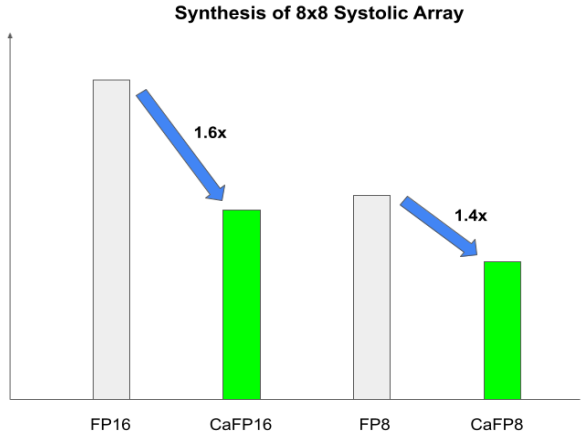

Figure 2: Size comparison of an 8x8 systolic array synthesized to different multiplier data types.

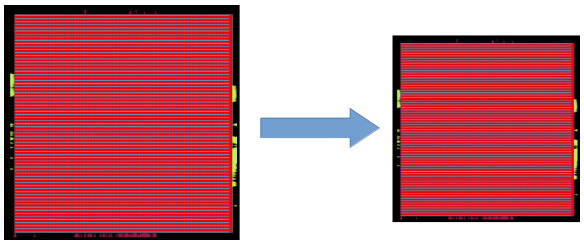

Figure 3: Fully placed 8x8 systolic arrays with standard multipliers on left versus Cassia multipliers on right.

	FP16	CaFP16	**Cassia**
Latency (ps)	*	*	**1.2x**
Power (W)	*	*	**2.1x**
Area (um2)	*	*	**1.5x**
GOPs/W	*	*	**2.5x**

Table 1a: Comparison of estimated latency, power, area, and efficiency for FP16 vs CaFP16 placed layouts. *available under NDA.

	FP8	CaFP8	**Cassia**
Latency (ps)	*	*	**1.2x**
Power (W)	*	*	**2.2x**
Area (um2)	*	*	**1.2x**
GOPs/W	*	*	**2.6x**

Table 1b: Comparison of estimated latency, power, area, and efficiency for FP8 vs CaFP8 layouts. *available under NDA.

Please visit www.cassia.ai or email info@cassia.ai to learn more.
Cassia.ai Inc.

Cassia.ai

Post Training Inference Model Accuracy

Model	Dataset	Float32 Accuracy	CFloat32 Accuracy	Accuracy Loss
FC+Sigmoid	MNIST	87.50%	87.40%	0.10%
LeNet	MNIST	98.60%	98.50%	0.10%
MobileNetV1	MNIST	98.46%	98.19%	0.27%
ResNet18	ImageNet	69.76% / 89.08%	69.22% / 88.79%	0.54%
ResNet50	ImageNet	76.13% / 92.86%	75.22% / 92.56%	0.91%

Table 2: Shows popular models and their errors with Cassia technology utilized. **Accuracy can be recovered by minor re-training with Cassia.ai arithmetic.**

Keep your Architecture with Cassia

With Cassia approximate arithmetic, your architecture does not change and you keep the standard integer and floating-point data formats. The only changes are to the core arithmetic operations: **multiplication**, division, exponentiation, and logarithms. Some key benefits:

- No need to transform the memory layout of your tensors.
- No need to increase the number of scatter-gather operations to and from memory.
- No special preparation of data either on-chip or off-chip.
- No constraints on tensor sizing or formatting (e.g. dimensions or channels).
- No special DMA requirements.

Engagement

1. Initial Cassia IP is delivered to customer with an NDA and Evaluation agreement.
2. Collaboration between Cassia and customer's engineers in Systems Architecture, VLSI design, Validation and Software departments to understand and integrate Cassia IP
3. Customer may want to pay consulting fees to adapt Cassia IP and minimally required Software modifications to their engineering flow.
4. Customer signs a Production agreement with a tape-out and licensing fees when they have validated their design with an emulated system and decided to proceed with usage of Cassia IP.