MITSUBISHI ELECTRIC
Changes for the Better

EPFL

Hardware-Efficient Quantization for Green Custom Foundation Models

Toshiaki Koike-Akino[(1,2)], Chang Meng[(2)], Volkan Cevher[(2)], Giovanni De Micheli[(2)]
[(1)]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA
[(2)]École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

Introduction

- We show the energy efficiency of floating-point (FP) multipliers over integer multipliers when synthesized on custom hardware chips.
- We propose **hardware-efficient quantization (HEQ)**, enabling hardware profiles differentiable to optimize the weight quantization for power reduction.
- Our HEQ framework achieves **25%** power reduction, and our custom multipliers provide up to **20-fold** power reduction altogether.

Floating-Point vs. Integer Multiplier

- FP multipliers are more energy efficient than integer multipliers.
- `bfloat16` is **2-fold** efficient than `int16` multipliers (fewer bits in mantissa).

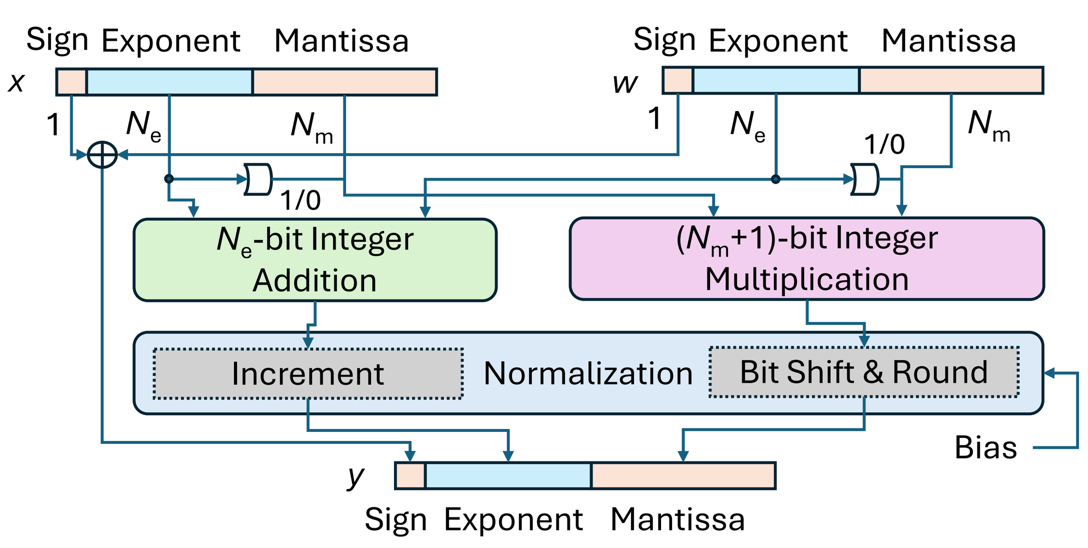


Figure 1: General FP multiplier diagram: exponent adder; mantissa multiplier; normalization. Hardware complexity is dominated by $(N_\mathrm{m}+1)$-bit integer multiplier block.

Table 1: Power/delay/area profiles of general multipliers designed through Yosys[2]/ABC[3] logic synthesis and Synopsys Design Compiler[4] on 45nm CMOS technology standard cell library[1]. Power consumption is at 0.2GHz clock frequency.

Multipliers	int32	float32$_{e8m23}$	int16	float16$_{e5m10}$	bfloat16$_{e8m7}$	int8	float8$_{e5m2}$	float8$_{e4m3}$	int4	float4$_{e3m0b6}$
Power (μW)	5,883.5	**4,886.3**	1,054.6	814.6	**435.6**	170.5	**63.3**	101.3	15.6	**8.4**
Delay (ns)	4.99	5.00	2.67	3.76	3.25	1.58	1.25	1.65	0.45	0.29
Area (μm^2)	5,412.8	4,063.9	1,157.6	828.9	508.6	231.2	95.2	144.7	29.5	16.0

Green Custom Foundation Models

- We design full-custom AI chip with constant quantized weights.
- Constant multipliers are lower power than general multipliers (5–20 folds).
- Power consumption depends on weight distributions.

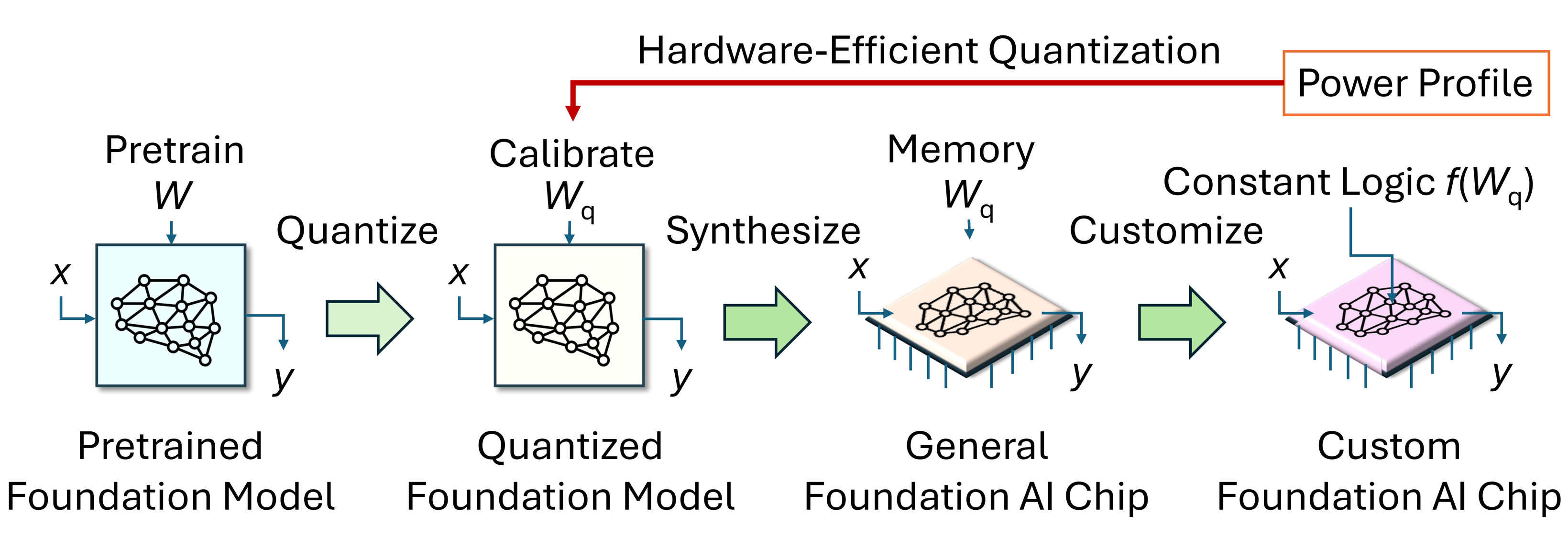


Figure 2: Design of green custom foundation models.

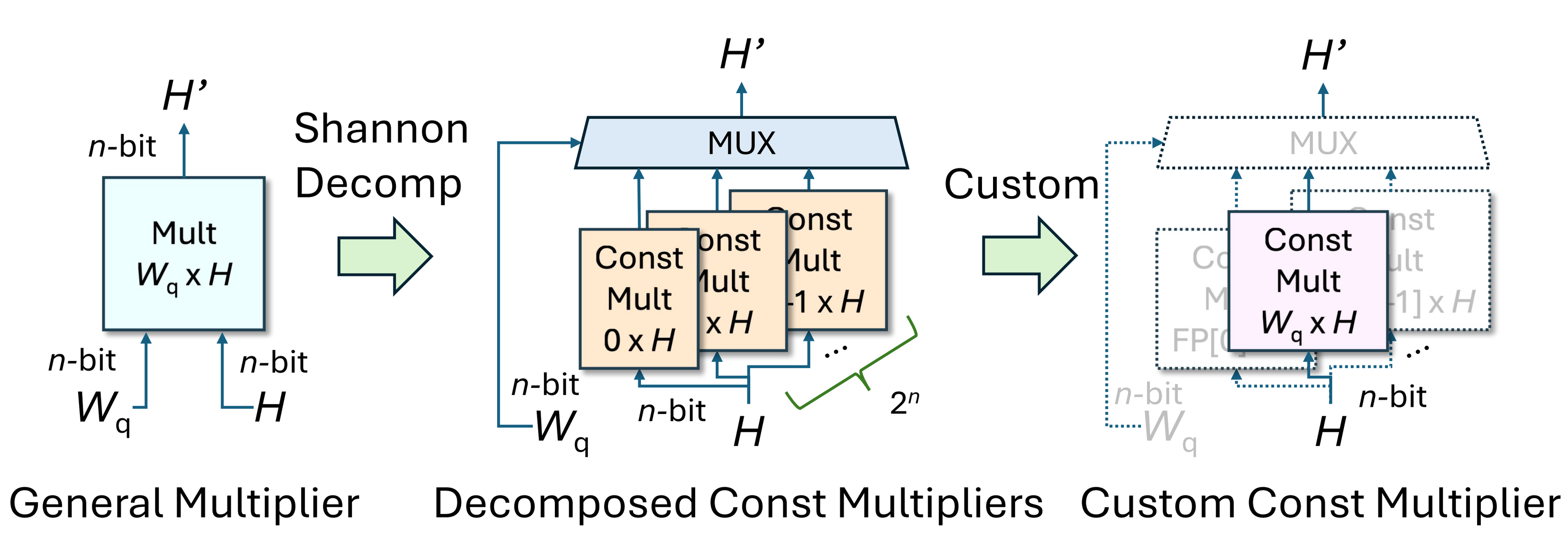


Figure 3: Shannon decomposition of general multiplier towards custom constant-weight multiplier.

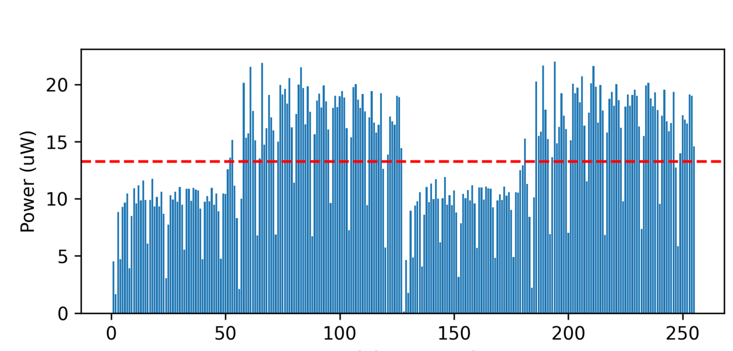


Figure 4: Power profile across quantized weight value for custom FP8 e4m3 multipliers. Average power is 13.3μW, average delay is 0.48ns, and average area is $28.7\mu m^2$. **8-fold** power efficient than general FP8 multipliers.

HEQ

- HEQ optimizes weight quantization distribution to jointly minimize cross entropy and power consumption for custom multipliers.

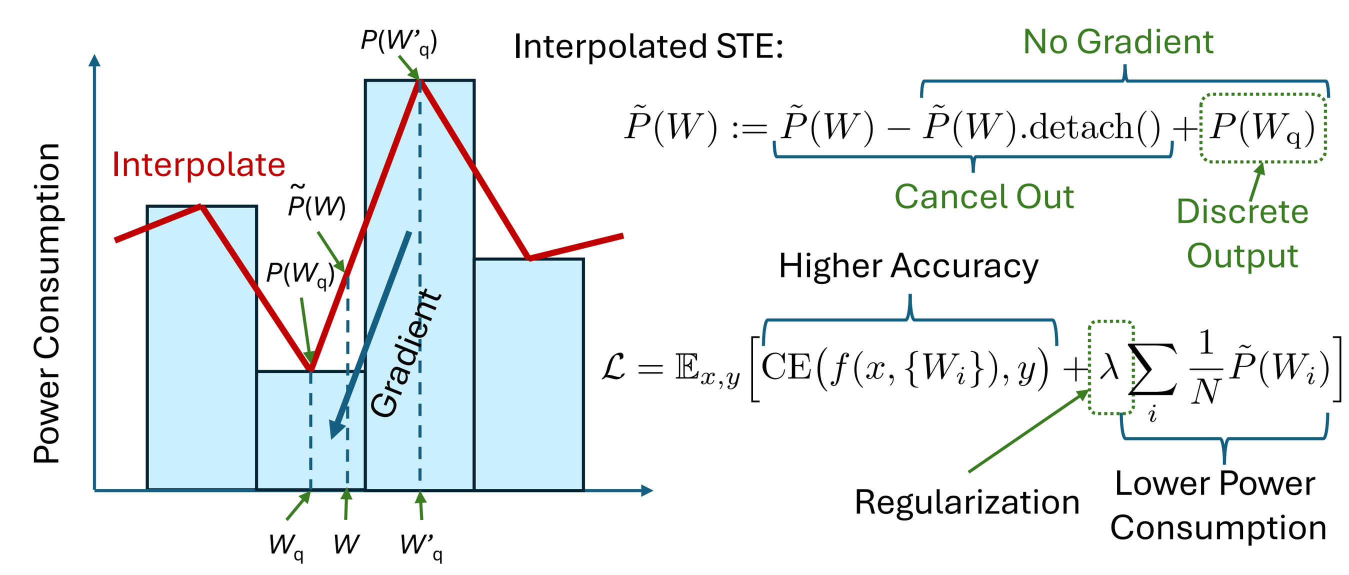
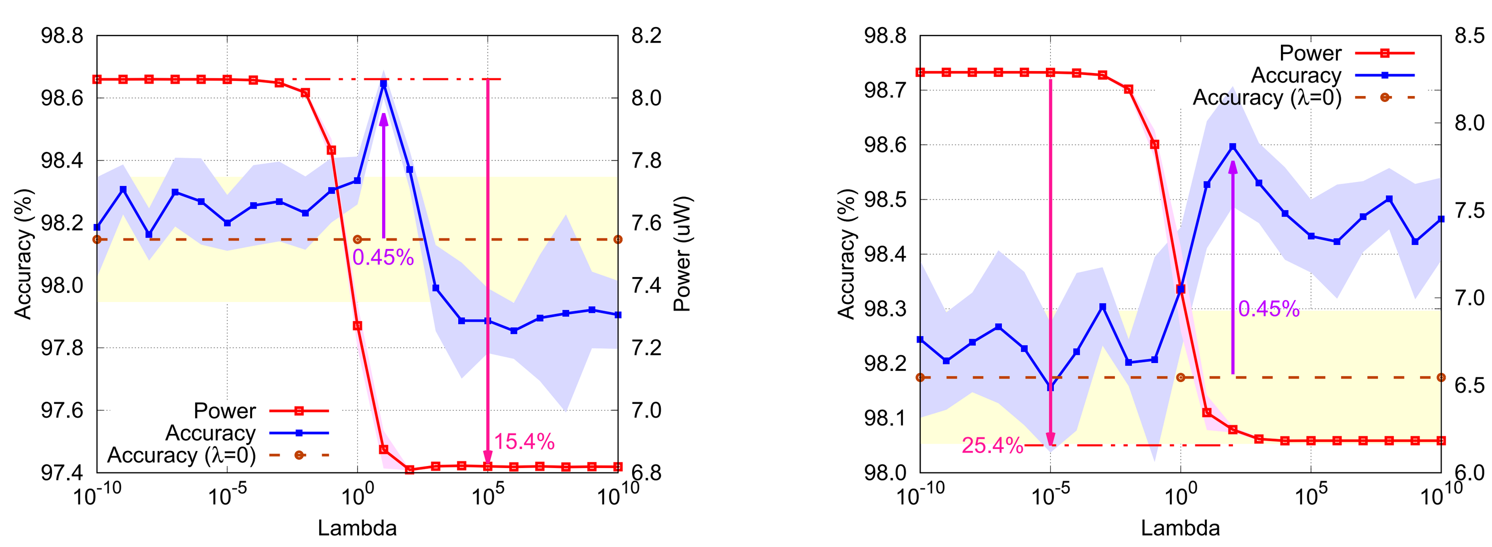


Figure 5: Interpolated STE[5] for differentiable hardware profile, enabling quantization-aware training (QAT). Regularized loss to minimize cross entropy and power consumption.

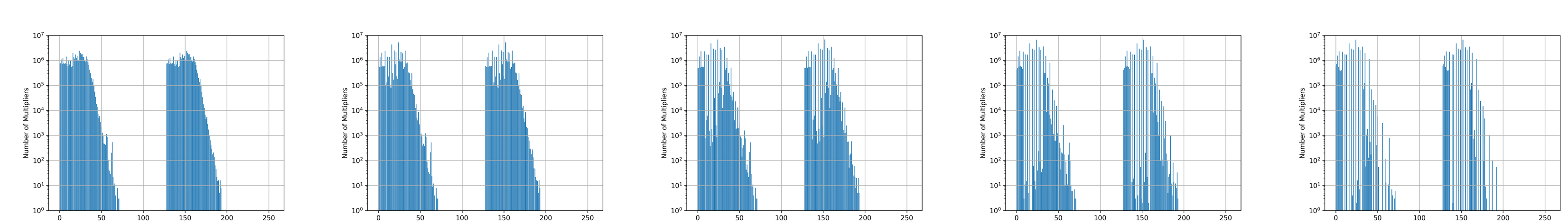
Experiments & Results

- HEQ regularization improves **both** performance and energy efficiency.
- FP3 HEQ achieves $\mathbf{7\times 10^4}$ **greener** than FP32 within 1% loss.



(a) FP8 e5m2 (b) FP8 e4m3

Figure 6: Power-aware quantization results across regularization factor λ. Error band shows a confidence interval under one standard deviation over 7 random seeds.



(a) $\lambda=10^{-1}$ (b) $\lambda=10^{0}$ (c) $\lambda=10^{1}$ (d) $\lambda=10^{2}$ (e) $\lambda=10^{3}$

Figure 7: Quantized weight histogram for custom FP8 multiplier e4m3, across regularization λ.

Table 2: Comparison of quantization methods for implementing custom ViT model[6].

	PTQ on General Multiplier									
Precision	FP32$_{e8m23}$	FP16$_{e5m10}$	BF16$_{e8m7}$	FP8$_{e5m2}$	FP8$_{e4m3}$	FP6$_{e3m2b7}$	FP5$_{e3m1b7}$	FP4$_{e3m0b6}$	INT4$_{e0m3b4}$	FP3$_{e2m0b5}$
Accuracy (%)	98.29$_{\pm 0.11}$	98.03$_{\pm 0.21}$	98.04$_{\pm 0.22}$	98.01$_{\pm 0.25}$	97.81$_{\pm 0.25}$	97.83$_{\pm 0.00}$	97.45$_{\pm 0.00}$	92.49$_{\pm 0.90}$	10.69$_{\pm 1.25}$	14.25$_{\pm 2.06}$
Power (μW)	4,886.3	814.6	435.6	63.3	101.3	46.62	24.04	8.4	15.6	1.2
	HEQ on Custom Multiplier									
Precision	FP32$_{e8m23}$	FP16$_{e5m10}$	BF16$_{e8m7}$	FP8$_{e5m2}$	FP8$_{e4m3}$	FP6$_{e3m2b7}$	FP5$_{e3m1b7}$	FP4$_{e3m0b6}$	INT4$_{e0m3b4}$	FP3$_{e2m0b5}$
Accuracy (%)	—	98.70$_{\pm 0.09}$	—	98.65$_{\pm 0.05}$	98.60$_{\pm 0.11}$	**98.78**$_{\pm 0.05}$	98.67$_{\pm 0.09}$	97.99$_{\pm 0.08}$	55.91$_{\pm 6.74}$	97.35$_{\pm 0.14}$
Power (μW)	—	179.09$_{\pm 0.82}$	—	6.87$_{\pm 0.06}$	6.25$_{\pm 0.02}$	2.35$_{\pm 0.00}$	1.19$_{\pm 0.01}$	0.60$_{\pm 0.00}$	0.13$_{\pm 0.00}$	**0.07**$_{\pm 0.00}$

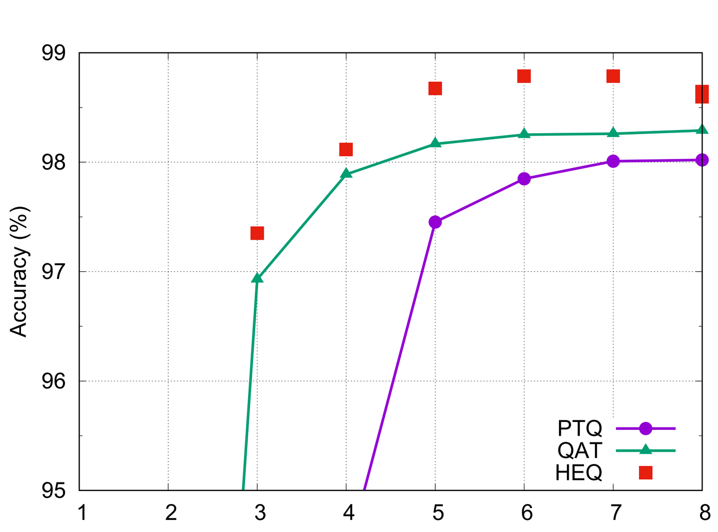


Figure 8: Accuracy vs. quantization bits.

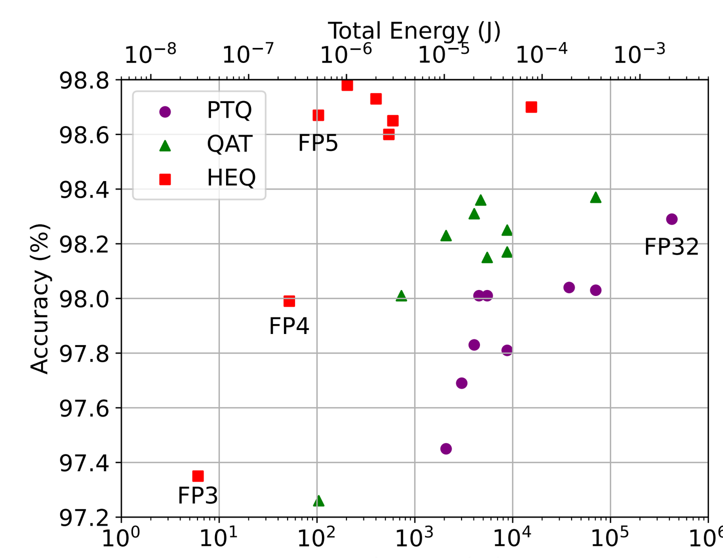


Figure 9: Accuracy vs. power tradeoff.

References

[1] Nangate. "The Nangate 45nm Open Cell Library," https://si2.org/open-cell-library/, 2011.

[2] C. Wolf, et al., "Yosys - A free Verilog synthesis suite," Austrochip, vol. 97, 2013.

[3] R. Brayton and A. Mishchenko, "ABC: An academic industrial-strength verification tool," CAV 2010, pp. 24–40. Springer, 2010.

[4] Synopsys, "Design Compiler," https://www.synopsys.com/, 2024.

[5] A. Gholami, et al., "A survey of quantization methods for efficient neural network inference," In Low-Power Computer Vision, pp. 291–326. Chapman and Hall/CRC, 2022.

[6] A. Dosovitskiy, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2020.

ICML Workshop ES-FoMo-II 2024: koike@merl.com; {chang.meng, volkan.cevher, giovanni.demicheli}@epfl.ch