29.3 An 8.09TOPS/W Neural Engine Leveraging Bit-Sparsified Sign-Magnitude Multiplications and Dual Adder Trees

Hyochan An, Yu Chen, Zichen Fan, Qirui Zhang, Pierre Abillama, Hun-Seok Kim, David Blaauw, Dennis Sylvester

University of Michigan, Ann Arbor, MI

The computational complexity of neural networks (NNs) continues to increase, spurring the development of high-efficiency neural accelerator engines. Previous neural engines have relied on *two's-complement* (2C) arithmetic for their central MAC units (Fig. 29.3.1 top, left). However, gate-level simulations show that *sign-magnitude* (SM) multiplication is significantly more energy efficient; ranging from 35% (with uniformly distributed operands) to 67% (with normally distributed operands (μ=0, σ=25)). The drawback of sign-magnitude number representation is that SM *addition* incurs significant overhead in terms of energy consumption and area, requiring upfront comparison of the sign bits and muxing/control to appropriately select between addition and subtraction (Fig. 29.3.1 center, left). This SM addition overhead substantially offsets the gains from SM multiplication in general purpose computing. One recent effort [1] to employ SM representation in neural computation achieved modest energy improvement at the cost of 2.5× area increase due to full duplication of the MAC units, which would typically be unacceptable for area-/cost-sensitive IoT applications.

This paper experimentally demonstrates that neural engines provide a unique opportunity to leverage the efficiency of SM multiplication by performing an inner-product using *unsigned* dual adder trees to lower the overhead of SM additions. We propose an SM *inner-product unit* (IPU) that can be applied to accelerate convolutional, fully-connected, and transformer neural architectures. The idea can be applied orthogonally with other energy improvement techniques in most NN accelerators to further increase energy efficiency. The IPU performs eight multiplications in parallel and sums their results using two unsigned adder trees that combine their two results with a subtractor at their root. In this fashion, SM addition overhead is deferred to the root of the adder trees and incurred only once for the entire inner-product computation (Fig. 29.3.1, center right). Further, the SM IPU benefits from the reduced bit-switching activity of both neural weights and activations when represented as SM numbers since their values are concentrated near zero with fewer non-zero bits in the magnitude field. Finally, the SM IPU performs *intentional bit sparsification*, which can be efficiently implemented in SM representation, reducing energy consumption by an additional 16% with a very minor loss (~0.29%) in inference top-1 error rate.

The proposed SM inner-product unit was implemented in a neural engine (NE) using a systolic-array architecture with zig-zag input activation scanning and a weight circular buffer in 28nm CMOS and compared with an identical 2C companion implementation. The SM implementation achieves 50% energy efficiency improvement with 14% area increase compared to its 2C counterpart, with 8.09TOPS/W peak efficiency at 0.65V.

Figure 29.3.2 provides an overview of the test chip containing two NEs: an SM NE and a 2C NE. The two NEs operate with independent power domains and clocks, each containing a 4.3Mb global memory (MEM). Each NE consists of four processing lanes; a single lane has 3×8 processing elements (PEs). Each PE contains an (SM or 2C) inner-product unit and a register file for temporary accumulated data. The SM IPU consists of eight SM multipliers and two 8-input unsigned adder trees (one for positive and the other for negative additions only), whereas the 2C IPU has a single 8-input adder tree that accumulates 2C-represented signed numbers. The PE array is structured to form an output-stationary systolic array where PE data propagates only from the west and north. The accumulated (convolution or matrix multiplication) values are bit-shifted (scaled) by a post-processing unit (PPU), which also performs ReLU, and bit sparsification (for SM NE). The NE controller controls the datapath to support general convolution and fully connected layers with configurable parameters.

The SM inner-product unit exhibits excellent energy efficiency for a general NN workload. The 8b SM multiplier has 35% lower energy consumption than the 8b 2C multiplier when operands are uniformly distributed. This observed gain is mainly due to differences in the bit-toggle activity per operation (Fig. 29.3.3). Furthermore, it is well known that NN weights and activations typically have non-uniform distributions with high occurrences of small or zero values [2]. This fact further improves the energy efficiency of SM multipliers relative to 2C multipliers, which do not fully exploit these distributions since the sign change of a small value requires toggling most bits in a 2C representation. Post-synthesis gate-level simulations in Fig. 29.3.3 show that an 8-length SM IPU has 20-57% lower energy consumption than a 2C counterpart when input operands follow a zero-mean Gaussian distribution with standard deviations (σ) ranging from 127-16 for 8b fixed point SM or 2C operands.

Extending the structured sparsity concept to the number representation system, we propose a *reconfigurable bit-sparsification technique* to maximize the benefit of the SM IPU (Fig. 29.3.4). The scheme constrains the number of non-zero bits (NZB, excluding sign bit) in the magnitude of SM-represented weights and activations to reduce bit-toggle activity during computation, with relatively low quantization loss. The NZB parameter for bit-sparsification can be dynamically configured for each NN layer. To minimize accuracy degradation, each NN model is re-trained after imposing the bit-sparse quantization constraint on weights and activations. Fig. 29.3.4 (bottom right) shows that a popular NN (VGG-nagadomi [3]) exhibits only 0.29% TOP-1 error rate degradation for NZB=2 (or 1.16% for NZB=1). Considering the small overhead of on-the-fly bit-sparsification for activations, the technique reduces energy by 31% and 16% with NZB=1 and 2, respectively, compared to the original SM inner-product unit. The SM NE employs a configurable bit sparsifier implemented in the PPU to enable bit sparse quantization for activations with a layer-dependent reconfigurable (1–3, and 7) NZB. In the bit sparsifier (Fig. 29.3.4 top right), three fixed priority encoders are cascaded via an intermediate OR-chain working as a mask for the next encoder. The configurable bit sparsifier comprises < 0.05% of the SM inner-product unit area with negligible energy overhead since it is performed only once at the final activation output (weights are loaded in bit-sparsified form).

In the output-stationary systolic array, each output activation is sequentially updated in the horizontal dimension fully utilizing PEs that locally share/propagate data (Fig. 29.3.5). In this manner, input activations and weights are accessed from buffer memories in a predefined zig-zag order, which is implemented using three address pointers incremented by reconfigurable strides. Each activation is reused in the next PE column. For each convolution kernel, weights are read only once from the main memory and loaded to a weight circular buffer, which feeds necessary weights to all PEs until convolution completes. The NE can utilize PEs at 99.7% for 3×3 convolution layers with multiple (e.g., 256 input and 8 output) channels, while the PE utilization ratio is 33.3% for 1×1 convolution kernels and fully-connected layers.

Figures 29.3.5 (bottom) and 29.3.6 (top) show measurement results of the 28nm test chip. SM and 2C NE energy efficiencies are measured using uniformly or Gaussian-distributed random operands. The 2C NE does not benefit from operand distribution changes except when 50% zero activations are used (modeling ReLU function). On the other hand, SM NE energy efficiency is higher than 2C NE and it improves by 1.5× as the distribution changes towards lower σ and additional zeros. With NZB=2b-sparsification, the SM NE exhibits 50% higher energy efficiency than 2C NE when σ=25 with replacing negative values by zero (modeling ReLU function). A trained VGG-nagadomi NN was demonstrated on the fabricated NEs. The SM NE and bit-sparsification technique improve energy efficiency at each of the 8 VGG layers. Bit-sparsification provides larger improvements in layers having higher σ. Using NZB=2b-sparsification, the SM NE increases energy efficiency by 15-34% compared to the 2C NE and achieves 3-8TOPS/W depending on the layer characteristics.

The table in Fig. 29.3.6 compares the demonstrated SM NE and 2C NE to other state-of-the-art neural accelerators. The SM inner product unit with dual adder tree and bit-sparsified operation is orthogonal to most other NN acceleration techniques and can be implemented in other NN accelerators with small area overhead and negligible algorithm accuracy degradation. It is effective at reducing energy consumption of NN computation and can be applied to convolutional, fully-connected, and transformer neural architectures.

Acknowledgement:
The authors would like to thank Taiwan Semiconductor Manufacturing Company and Sony Semiconductor Solutions Corp./Sony Electronics Inc. for supporting this work.

References:

[1] P Whatmough et al., "A 28nm SoC with a 1.2GHz 568nJ/Prediction Sparse Deep-Neural-Network Engine with >0.1 Timing Error Rate Tolerance for IoT Applications," *ISSCC*, pp. 242-243, 2017.

[2] S. Han, H. Mao and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," arXiv preprint arXiv:1510.00149, 2015.

[3] Nagadomi, "Code for Kaggle-CIFAR10 Competition, 5th Place," https://github.com/nagadomi/kaggle-cifar10-torch7, 2014.

[4] S. Ryu et al., "BitBlade: Energy-Efficient Variable Bit-Precision Hardware Accelerator for Quantized Neural Networks," *IEEE JSSC*, vol. 57, no. 6, pp. 1924-1935, 2022.

[5] C.-H. Lin et al., "A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC", *ISSCC*, pp. 134-135, 2020.

2023 IEEE International Solid-State Circuits Conference (ISSCC) | 978-1-6654-9016-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISSCC42615.2023.10067269

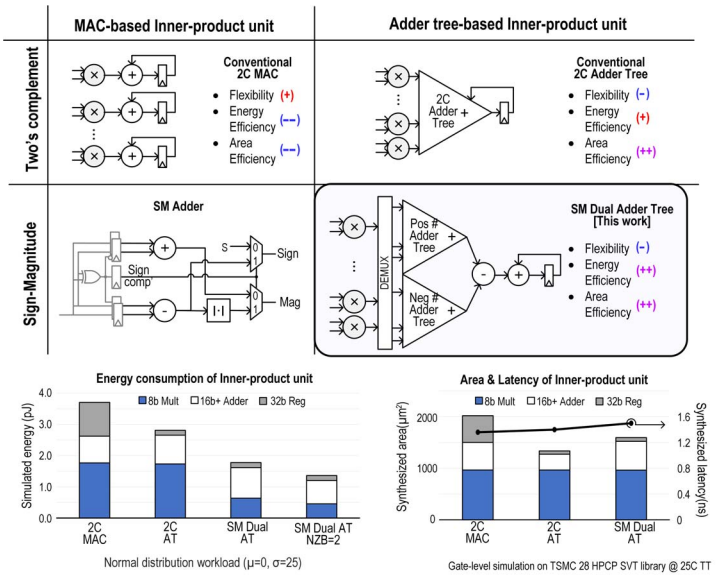


Figure 29.3.1: Comparison of the inner-product units (IPUs). (Bottom) Simulated energy, area, and latency comparison of three different IPUs.

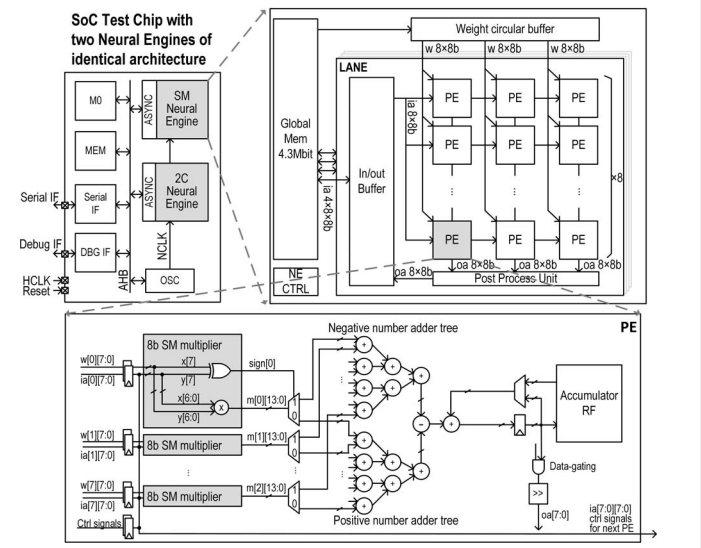


Figure 29.3.2: Architecture of the test chip with two neural engines: SM NE, and 2C NE. (Bottom) The PE architecture of SM NE. The PE of 2C NE contains 8b 2C multipliers and a conventional 2C adder tree.

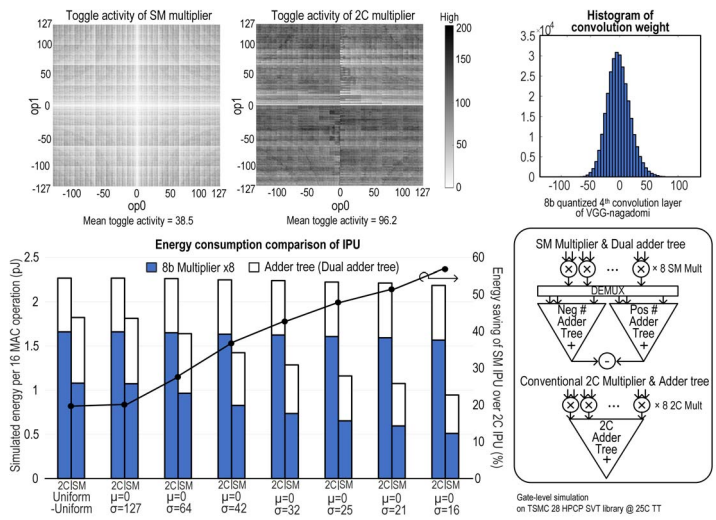


Figure 29.3.3: Comparison of arithmetic in SM and 2C. (Top, left) Toggle activity heat maps of SM and 2C multipliers. (Top, right) Weight distribution of the convolution kernel of the VGG-nagadomi. (Bottom) Energy consumption comparison of the IPUs.

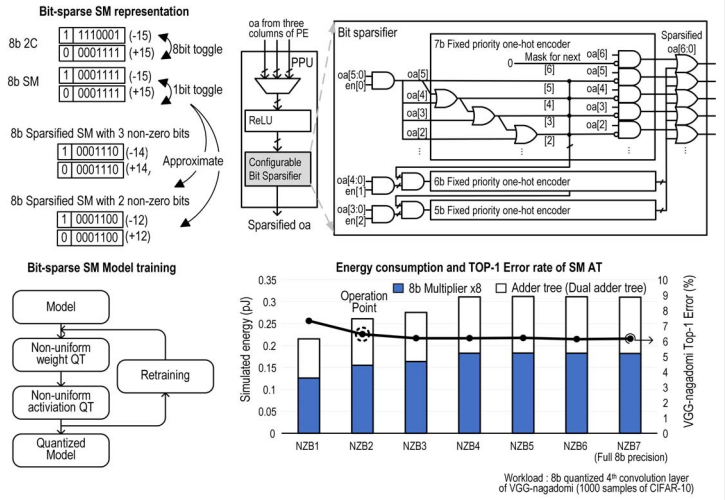


Figure 29.3.4: Proposed reconfigurable bit-sparsification. (Top, left) Bit sparse SM representation. (Top, right) Configurable bit sparsifier. (Bottom, left) Bit-sparse SM model training scheme. (Bottom, right) Energy consumption with bit-sparsification scheme.

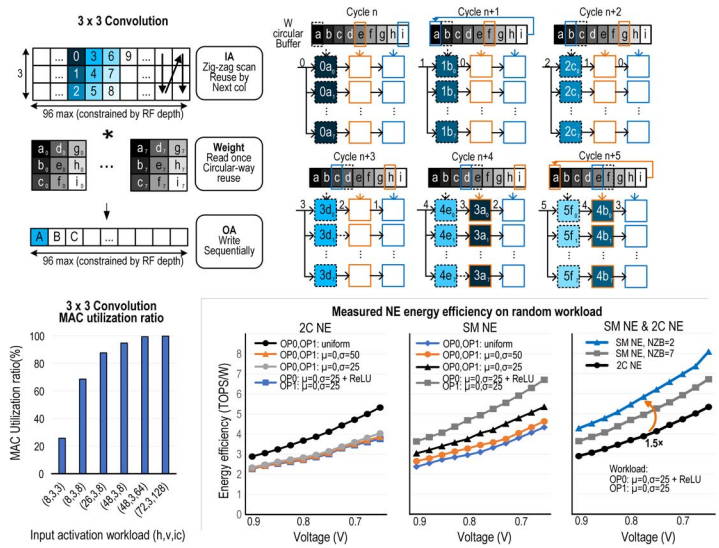
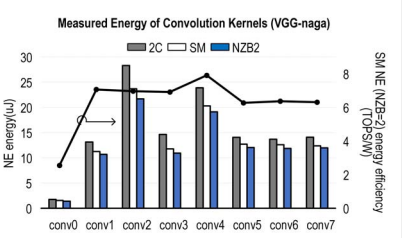


Figure 29.3.5: (Top) Output-stationary systolic array architecture. (Bottom) Measured NE energy efficiency on various random workloads.

VGG-nagadomi Convolution Kernel

	Weight						IA*
	h	v	ic	oc	Density(%)	σ	Density(%)
conv0	3	3	3	64	92.1	19.9	29.4
conv1	3	3	64	64	88.1	10.7	32.0
conv2	3	3	64	128	97.8	20.7	30.1
conv3	3	3	128	128	97.6	17.5	31.1
conv4	3	3	128	256	97.4	15.9	14.3
conv5	3	3	256	256	96.6	11.6	9.3
conv6	3	3	256	256	97.4	16.6	8.1
conv7	3	3	256	256	94.2	9.6	10.8

* Mapping overhead included



	[1] ISSCC 2017	[4] JSSC 2020	[5] ISSCC 2020	This work	
				2C NE*	SM NE*
Purpose	FL	FL, CL	FL, CL	FL, CL	FL, CL
Process (nm)	28 nm	28nm	12 nm	28 nm	28 nm
Area (mm)	5.76	0.71	709	2.18	2.18
Supply Voltage (V)	1.1 - 0.6	1.0 - 0.6	-	0.9 - 0.65	0.9 - 0.65
Frequency (MHz)	1200 @0.9V 667@0.715V	195 @1.0V 44 @ 0.6V	475 - 700	280 @0.9V 43 @0.65V	285 @0.9V 55 @0.65V
On-chip Memory (KB)	1024	144	196,608	537	537
Bit Precision	SM 16b/8b	2C 2,4,8b	2C 8b	2C 8b	bit-sparse SM 8b
Peak Performance (GOPS)	19.2 @ 0.9V 10.672 @ 0.715V	1420 @ 2bx2b, 1.0V 320 @ 2bx2b, 0.6V	825000	430 @0.9V 66 @0.65V	437 @0.9V 84 @0.65V
Peak Energy Efficiency (TOPS/W)	1.21 @ 0.9V 1.23 @ 0.715V	44.1 @ 2bx2b, 0.6V	2.95	2.8 @0.9V 5.3 @0.65V	4.2 @0.9V 8.09 @0.65V

* Randomized workload
OP0: μ=0,σ=25 + ReLU
OP1: μ=0,σ=25

Figure 29.3.6: (Top) Measured energy consumption of VGG-nagadomi. (Bottom) Comparison table.

Process	TSMC 28nm HPCP 1P10M	
Die Size	2610 x 2970 um	
NE	SM NE	2C NE
Die Area	1773 x 1233 um	
Total SRAM	4.3 Mbit for each NE	
ML Model	VGG-nagadomi	
Weight/IA Preicision	8b 2C	8b bit-sparse SM
Supply Voltage	0.9 - 0.65 V	0.9 - 0.65V
F max	280 - 43 MHz	285 - 55 MHz
Peak Performance	430 - 66 GOPS	437 - 84 GOPS
Peak Energy Efficiency	2.8 - 5.3 TOPS/W	4.20 - 8.09 TOPS/W (NZD=2)

* Measured with Randomized workload
OP0: μ=0,σ=25 + ReLU
OP1: μ=0,σ=25

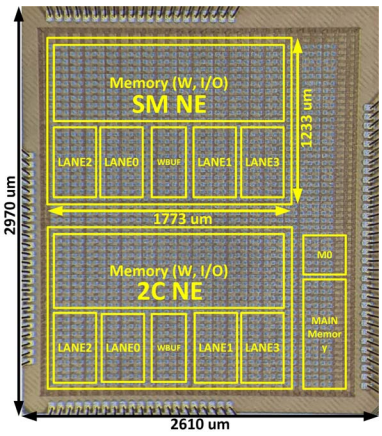


Figure 29.3.7: Die photo.

978-1-6654-9016-0/23/$31.00 ©2023 IEEE