A 4096-neuron 1M-synapse 3.8pJ/SOP Spiking Neural Network with On-chip STDP Learning and Sparse Weights in 10nm FinFET CMOS

Gregory K. Chen, Raghavan Kumar, H. Ekin Sumbul, Phil C. Knag, Ram K. Krishnamurthy
Circuits Research Lab, Intel Corporation, Hillsboro, OR USA
gregory.k.chen@intel.com

Abstract

A 4096-neuron, 1M-synapse SNN in 10nm FinFET CMOS achieves a peak throughput of 25.2GSOP/s at 0.9V, peak energy efficiency of 3.8pJ/SOP at 525mV, and 2.3μW/neuron operation at 450mV. The SNN skips zero-valued activations for up to 9.4× lower energy. Fine-grained sparse weights reduce memory by up to 16×. On-chip STDP trains RBMs to de-noise MNIST digits and to reconstruct natural scene images with RMSE of 0.036. A 50% sparse weight MLP classifies MNIST digits with 97.9% accuracy at 1.7μJ/classification.

Introduction

Spiking Neural Networks (SNNs) perform cognitive tasks using energy-efficient sparse spikes. Digital ASIC SNNs [1] have higher energy efficiency than CPU designs [2] and improved scalability and noise margin over mixed-signal designs [3]. Leaky Integrate and Fire (LIF) neurons process temporal spike sequences by integrating inputs over multiple time-steps. On-chip Spike Timing Dependent Plasticity (STDP) trains 7b weights to reinforce spike patterns generated by input data. Single time-step LIF operation implements Binary-activation Neural Networks (BNNs) [4] with {0, 1}-valued activations and 7b weights trained with deep learning. The 1.72mm^2 10nm FinFET [5] SNN (Fig. 1) combines sparse connectivity, stochastic operation, voltage scaling, and power gating to achieve 4.8× throughput and 6.8× energy improvements over previously reported SNNs (Table 1).

LIF and BNN Neuron Operation

The LIF neuron integrates temporally coded input spikes onto a 16b membrane potential (u) and generates an output spike when u exceeds threshold θ (Fig. 2). Between time-steps (ts) a leak (λ) is applied to u to prioritize recent events before primary inputs and trained bias values are added. BNN operation is a subset of LIF, with temporally concurrent input spikes multiplied by weights and accumulated in a partial sum (u). At each time-step, the binary neuron spikes if u exceeds 0 (θ) before being reset. Neuron state is split into two 64-entry RFs, with 20b for weight integration and 32b for time-step updates. Fine-grained clock gating reduces RF power by 2.4× compared to a single-RF design, and lowers datapath clock power by 35%.

On-chip Spike Timing Dependent Plasticity

Bio-plausible STDP performs on-chip training for SNNs, and deep learning implements training on BNNs (Fig. 3). BNN inference uses a unit step activation function, and offline training with error back-propagation uses a Straight Through Estimator (STE) of the step function and its derivative. For SNNs, on-chip STDP updates weights according to relative spike timing between connected neurons. Spiking neurons read the spike history of each of its neighbors (Δt) to index a LUT and read weight updates (Δw) (Fig. 4). Forward spikes read spike history in fan-out neurons to perform Long Term Depression (LTD). Back-spikes are sent to fan-in neurons and perform analogous Long Term Potentiation (LTP), eliminating duplicate spike history copies to reduce neuron state by 83%.

Stochastic Synapse Operation and Multicast Spikes

Stochastic synaptic operation circuits implement dropout for training, perform neural sampling during inference, and co-optimize SNN performance, energy, and noise for sparse spiking activity. Time-steps are allotted a user-defined period of time to complete, with shorter accelerated time-steps increasing SNN performance. PRNG circuits enforce dropping at a controlled rate to resolve collisions at select bandwidth bottlenecks. Within each time-step, algorithmically concurrent spike order and timing are randomized to improve noise characteristics. To reduce NoC dropping, a richly connected 2-ary 6-flattened butterfly delivers spikes with multicast load-balancing routing (Fig. 5). Multicast address encoding uses 2b symbols {0, 1, *}, with wildcard (*) indicating both 0 and 1. Each symbol corresponds to a NoC hop with branching at wildcards (Fig. 6). Routers process addresses bitwise in random order to leverage path diversity. Multicast operation reduces spike delivery energy by up to 3× over unicast routing.

Fine-grained Sparse Synaptic Connectivity

Sparse connectivity eliminates memory for omitted weights (Fig. 7). Prior to training, a pseudo-random sparse SNN topology is generated using a Galois field (GF(2^3)^2) multiply permutation function. At run-time, spikes access a contiguous range of weight SRAM, and GF multiplication circuits calculate connection pointers as the product of the synapse address and a user-defined seed (Fig. 8). A GF division circuit enables reverse mapping for STDP training and Restricted Boltzmann Machines (RBMs) with symmetric connections. Connection data at each core is stored in a 64×40b RF for 2% memory overhead over the 16k×7b weight SRAM. Sparse connectivity generation obviates storing per-synapse neuron pointers to reduce synapse memory by 2.7× for a 1024×1024 75% sparse SNN. It eliminates null memory entries for 3.9× less memory compared to a dense crossbar implementation. Power gating unused weights reduces leakage by up to 2.7×.

10nm FinFET Measurements

A 4096-neuron, 1M×7b-synapse SNN is fabricated in 10nm FinFET CMOS (Fig. 16). The SNN achieves a peak throughput of 25.2GSOP/s at 0.9V, 506MHz, 8.3pJ/SOP (pJ/Synaptic Operation) (Fig. 9). Voltage scaling to 525mV reduces energy by 2.2× to 3.8pJ/SOP at 105MHz, 2.6GSOP/s (Fig. 10). Ultra-low-voltage circuits enable 450mV, 2.3μW/neuron operation. Sparse connectivity circuits reduce computation per time-step, decreasing energy by 15.0× for 93.75% sparsity (Fig. 11). Stochastic time-step acceleration skips zero-valued activations for 32× higher performance with 9.4× lower energy (Fig. 12).

Neuromorphic and Deep Learning Workloads

Spiking RBMs trained with on-chip unsupervised STDP-based contrastive divergence extract features from MNIST digits and natural scene images. An RBM processes 8x8 image patches with a stride of 4 to reconstruct a landscape image with RMSE of 0.036 (Fig. 13). A 1024×1024, 1M-synapse RBM de-noises corrupted MNIST digits to help with real-world visual recognition problems with noise or occlusions (Fig. 14). A 784×1024×512×10 BNN Multi-Layer Perceptron (MLP) with 50% sparse weights in each layer is trained with error back-propagation to classify MNIST digits with 97.9% accuracy (Fig. 15). In this BNN, stochastic time-step acceleration increases throughput by 8× and reduces energy by 7.3× to 1.7μJ/classification at an average spike rate of 20%.

References

[1] P.A. Merolla et al., *Science*, vol. 345, no. 6197, pp 668-673, 2014.
[2] E. Painkras et al., *JSSC*, vol. 48, no. 8, pp. 1943-1953, 2013.
[3] B.V. Benjamin et al., *Proc. of the IEEE*, vol. 102, no. 5, pp. 699-716, 2014.
[4] K. Ando et al., *VLSI Circuits*, pp. C24-C25, 2017.
[5] C. Auth et al., *IEDM*, pp. 29.1.1-29.1.4, 2017.

978-1-5386-4214-6/18/$31.00 ©2018 IEEE

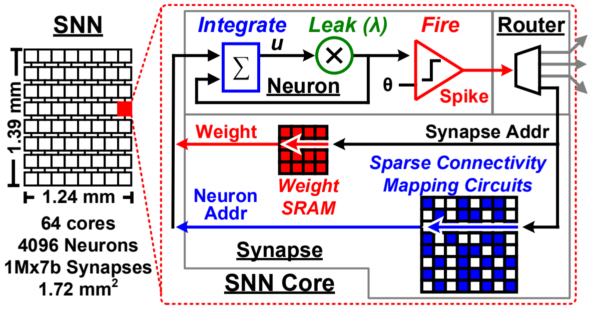


Fig. 1: Spiking Neural Network overview

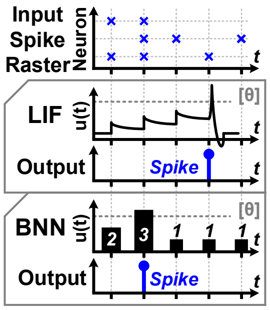


Fig. 2: LIF and BNN neurons

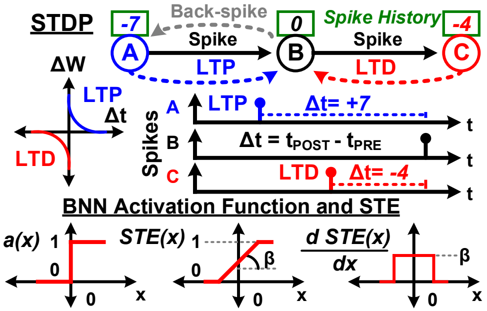


Fig. 3: STDP and binary deep learning

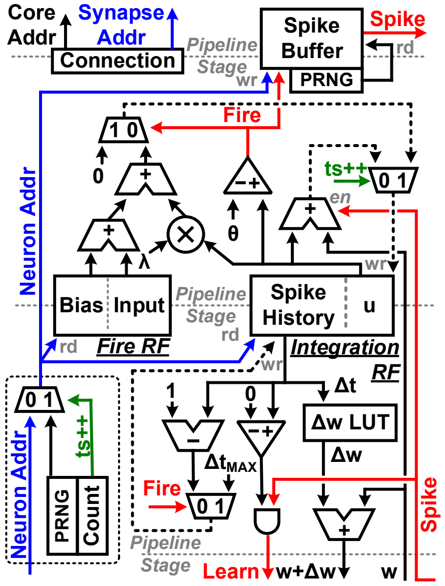


Fig. 4: Neuron LIF and STDP datapaths

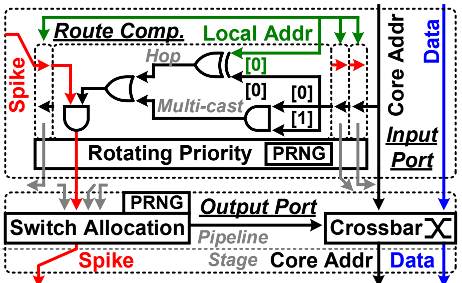


Fig. 5: NoC router circuits

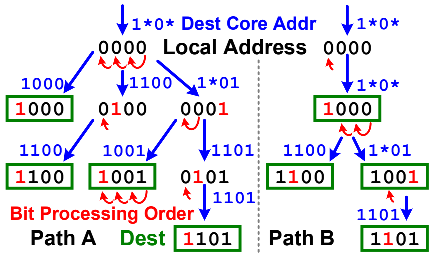


Fig. 6: Multicast spike delivery

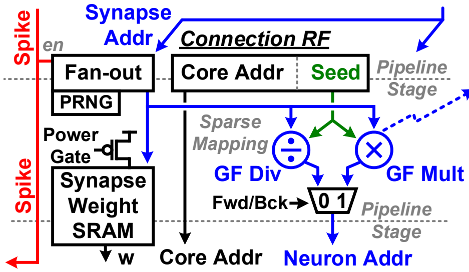


Fig. 7: Synapse datapath circuits

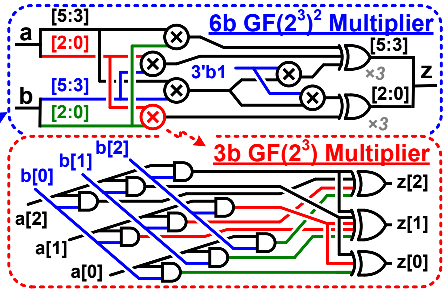


Fig. 8: Galois field multiply circuits

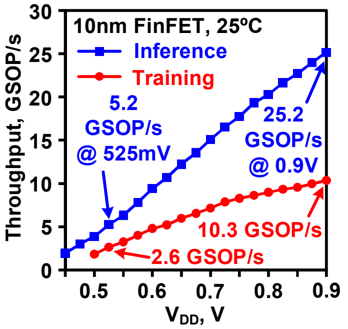


Fig. 9: Throughput measurements

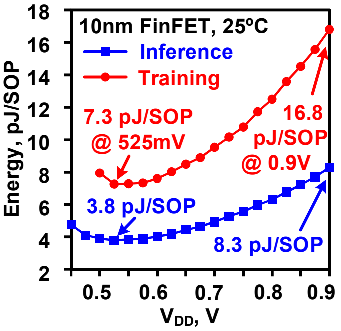


Fig. 10: Energy measurements

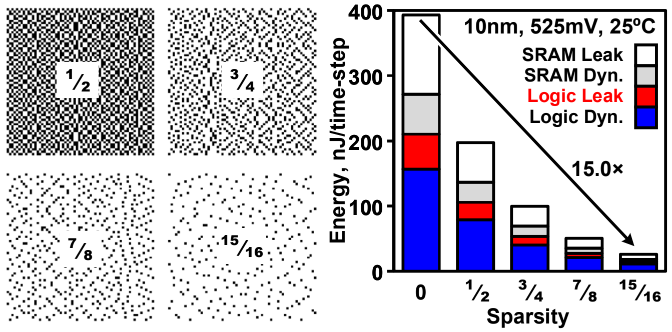


Fig. 11: Example sparse connectivity maps and energy

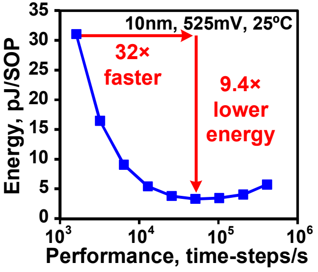
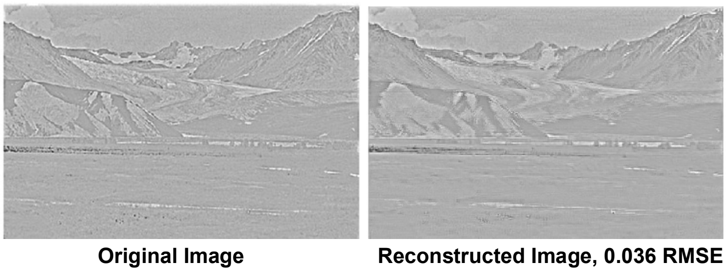
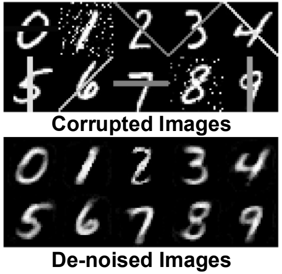


Fig. 12: Time-step acceleration


Original Image

Reconstructed Image, 0.036 RMSE

Fig. 13: Landscape feature detection and reconstruction


Corrupted Images

De-noised Images

Fig. 14: MNIST de-noising

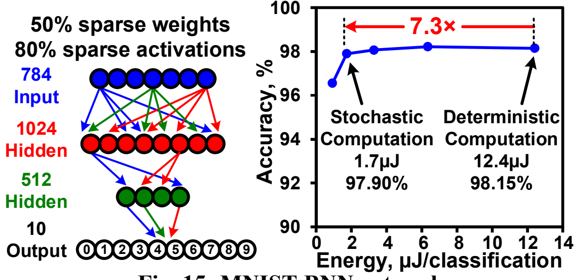


Fig. 15: MNIST BNN network

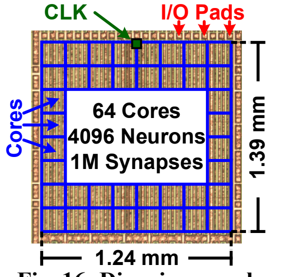


Fig. 16: Die micrograph

Table 1: Comparison with previous SNNs

Project	This Work		[1]	[2]	[3]
Process	10nm		28nm	130nm	180nm
Area, mm²	1.72		430	102	168
Synapse Bits	7b		1b	-	13b
Learning	Yes		No	Yes	No
Sparsity	Yes		No	Yes	No
Voltage	525mV	0.9V	0.775V	1.2V	1.8V
Freq., MHz	105	506	-	180	-
SOP/s/mm²	3.0G	14.6G	624M	784k	19.6M
pJ/SOP	3.8	8.3	26	4000	119