A 45nm CMOS Neuromorphic Chip with a Scalable Architecture for Learning in Networks of Spiking Neurons

Jae-sun Seo[1], Bernard Brezzo[1], Yong Liu[1], Benjamin D. Parker[1], Steven K. Esser[2], Robert K. Montoye[1],
Bipin Rajendran[1], José A. Tierno[1], Leland Chang[1], Dharmendra S. Modha[2], and Daniel J. Friedman[1]

[1]IBM T. J. Watson Research Center, [2]IBM Research - Almaden

Abstract- **Efforts to achieve the long-standing dream of realizing scalable learning algorithms for networks of spiking neurons in silicon have been hampered by (a) the limited scalability of analog neuron circuits; (b) the enormous area overhead of learning circuits, which *grows with the number of synapses*; and (c) the need to implement all inter-neuron communication via off-chip address-events. In this work, a new architecture is proposed to overcome these challenges by combining innovations in computation, memory, and communication, respectively, to leverage (a) robust digital neuron circuits; (b) novel transposable SRAM arrays that share learning circuits, which *grow only with the number of neurons*; and (c) crossbar fan-out for efficient on-chip inter-neuron communication. Through tight integration of memory (synapses) and computation (neurons), a highly configurable chip comprising 256 neurons and 64K binary synapses with on-chip learning based on spike-timing dependent plasticity is demonstrated in 45nm SOI-CMOS. Near-threshold, event-driven operation at 0.53V is demonstrated to maximize power efficiency for real-time pattern classification, recognition, and associative memory tasks. Future scalable systems built from the foundation provided by this work will open up possibilities for ubiquitous ultra-dense, ultra-low power brain-like cognitive computers.**

I. Introduction

In recent years, the development of large-scale networks with spiking neurons [1] and adaptive synapses based on spike-timing dependent plasticity (STDP) [2] has suggested new avenues of exploration for brain-like cognitive computing. A massively scaled example of this work is described in [3], which reports a cortical simulation at the scale of a cat cortex (10^9 neurons and 10^{13} synapses) using supercomputers. When compared to a brain that consumes 20W and occupies just 2L, the memory, computation, communication, power, and space resources needed for such a software simulation are tremendous. Consequently, there is an enormous opportunity for efficient, dedicated VLSI hardware that moves beyond prevalent von Neumann architectures to ultimately enable true large mammalian brain-scale networks (10^{10} neurons and 10^{14} synapses) with extremely low power consumption (< 1kW) and volume (< 2L) for ubiquitous deployment. Achieving this ambitious goal will demand a combination of advances in architectures designed to exploit spiking network properties coupled with exploitation of continued CMOS scaling.

A number of previous approaches implementing networks of spiking neurons in silicon [4][5][6] utilize analog circuit techniques (e.g. capacitor charge storage) and off-chip implementations of synapse plasticity, which limits portability to advanced technology nodes. On-chip synaptic learning was demonstrated with binary synapses in [7], but the number of synapses integrated on chip was limited as STDP circuitry was needed for *each* synapse element.

In this paper, we propose a scalable integrated circuit platform for large-scale networks of spiking neurons. We undertook a digital circuit approach, not only for scalability, inherent noise rejection, and robustness to variability, but also for reconfigurability, which enables our chip to serve as a general vehicle to support multiple learning algorithms and tasks. Tailored towards cognitive tasks such as pattern classification and object recognition, the architecture of the neuromorphic chip implemented here leverages a processor-in-memory approach to enable efficient communication between neurons and synapses. Efficient on-chip learning is enabled by a novel transposable synapse crossbar array to perform synapse weight updates in both row (post-synaptic) and column (pre-synaptic) directions in a single step. Such an organization allows sharing of the synapse learning circuitry across axons (rows) and dendrites (columns) at the crossbar periphery, which dramatically improves synaptic density. With an end target of a large-scale brain network operating in “real-time” (~ms time steps), low power consumption was achieved by the use of near-threshold circuits and an event-driven synchronous design.

Embodying the above unique solutions encompassing computation, memory, and communication, an integrated neuromorphic chip was fabricated in 45nm SOI-CMOS to demonstrate auto-associative learning and spatio-temporally sparse pattern recognition [8][9]. Additionally, three chip variants were designed and successfully demonstrated to precisely understand (1) density of reconfigurable versus customized neurons, (2) learning performance of chips with single-bit versus multi-bit synapses, and (3) power optimization by transistor threshold voltage adjustment.

II. Chip Design for Networks of Spiking Neurons

The essential dynamics of neurons and synapses in networks of spiking neurons is first provided, and we describe the chip design which efficiently implements those key features.

A. Networks of spiking neurons

A network of spiking neurons is comprised of neuron processing elements connected through synapse memory elements – both akin to the basic building blocks of the brain. Neurons communicate through binary messages called spikes, typically produced at 1-10 Hz. A spike is sent on a generating neuron’s axon to the dendrites of other neurons. The point of contact between an axon and dendrite is the synapse, which has a particular strength that determines the efficacy of a spike from a source, pre-synaptic neuron, on a target, post-synaptic neuron. Each neuron has a membrane potential that changes with incoming spikes and decays over time due to a leak. If the membrane potential exceeds a threshold, the neuron generates its own spike and the membrane potential is reset. Each neuron can be excitatory or inhibitory. Spikes from excitatory neurons increase the membrane potential of post-synaptic neurons, while spikes from inhibitory neurons do the opposite.

978-1-4577-0223-5/11/$26.00 ©2011 IEEE

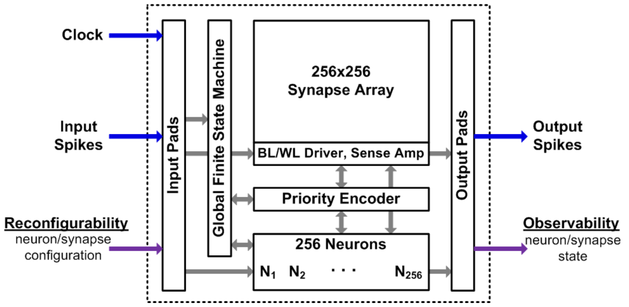


Fig. 1. Top-level block diagram of on-chip learning chip.

Networks of spiking neurons exhibit the ability to learn and adapt through the evolution of synapse weights. Such synaptic plasticity can be governed by time-dependent learning rules such as STDP, in which synapse weight changes depending on the order and distance in time between the spikes of source and target neurons. When a neuron fires, all incoming dendritic and outgoing axonal synapses must be updated.

The neuromorphic chip implemented in this work (Fig. 1), a scalable building block for large-scale networks of spiking neurons, integrates 256 neurons with 256x256 binary synapses to represent a fully recurrent network. In each time step, neurons that have exceeded their firing threshold generate spikes, which triggers synaptic integration in all target neurons as well as synapse weight updates based on the specified learning rule. Input spikes to the system can represent external stimuli (e.g. an input pattern) while generated spikes can represent output activity (e.g. recognition of a given pattern).

B. *Transposable SRAM Synapse Crossbar Array*

An *N*x*N* crossbar structure can effectively represent a system of *N* neurons and N^2 possible synaptic connections between them. Each row of the crossbar represents a neuron's axon and each column represents a neuron's dendrite. In stark contrast to traditional VLSI circuits, a large fan-in and fan-out can thus be efficiently realized as neuronal computations and synaptic updates can be parallelized, since an entire row or column of the crossbar can be operated upon simultaneously. To efficiently implement time-dependent learning rules, both row and column accesses are important, since pre-synaptic row and post-synaptic column updates are required. Conventional memory arrays are accessed only in rows, and column-based access would require inefficient serial row operations. Instead, this shortcoming is addressed in this work by a transposable SRAM cell (Fig. 2) to store synapse weights, which enables single-cycle write and read access in both row and column directions to significantly enhance on-chip learning performance. The transposable SRAM cell uses 8 transistors, adding two access transistors connected to word and bit lines in transposed orientations to a standard 6T design. The cell area is 1.6μm^2 in 45nm and is fully compatible with logic ground rules to enable excellent low voltage yield.

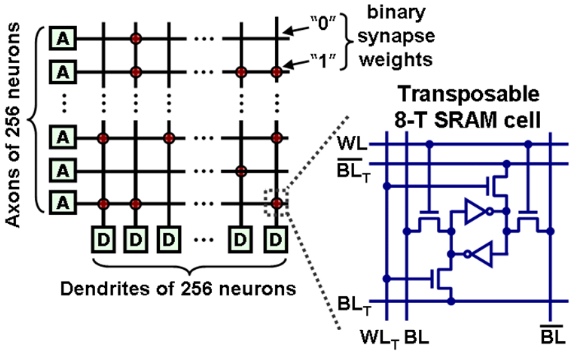


Fig. 2. Synapse array with transposable SRAM cells.

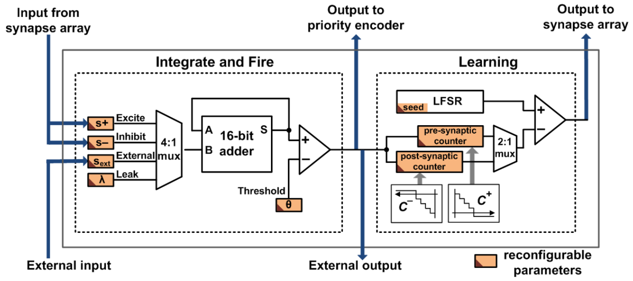


Fig. 3. Digital CMOS neuron with reconfigurability.

While much previous work focuses on analog synapse weights, this work utilizes binary synapses for optimal density, where each synapse is represented by a single memory cell. Instead of increasing or decreasing analog values, these binary weights are probabilistically set to '1' or '0' depending on the learning rule [10].

The synapse crossbar memory is implemented by a 256x256 array of transposable SRAM cells. To maximize array efficiency and timing margins, the 256 cells in either orientation are connected to a common bit line without any hierarchical structures. Read operations are performed by a conventional differential-pair sense amplifier while cross-coupled keeper PFET devices ensure noise tolerance and robustness to leakage effects during the slow cycle times needed for real-time system performance. Yield analysis considering process corners and statistical process variation using Monte Carlo methods was conducted to ensure design robustness.

C. *Digital CMOS Neuron Circuits*

Each neuron implements locally reconfigurable integrate-and-fire function [1] and learning rules (Fig. 3). At a functional level, the neuron follows the equation: $V(t)=V(t-1)+s_+n_+(t)-s_-n_-(t)-\lambda$ where *V* is the membrane potential, n_+ and n_- are the number of excitatory and inhibitory inputs received through "on" synapses, s_+ and s_- are input strength parameters and λ is a leak parameter. If *V(t)* exceeds a threshold *θ*, a spike is generated and *V(t)* is set to $V_{reset.}$. The membrane potential and parameters are 8-bit digital values. The integrate-and-fire module utilizes a 16-bit adder to handle overflow under worst-case conditions and a digital comparator to evaluate the spiking threshold. A 4:1 mux reuses the adder to integrate inputs and leak at each time step.

The circuits needed to perform probabilistic synapse weight update via time-dependent learning rules are implemented within the neuron, which thus shares learning circuitry across axonal rows and dendritic columns of synapses. 8-bit time-keeping counters C^+ and C^- enable independent control of pre- and post-synaptic updates, respectively, by tracking the time elapsed since the last spiking event of each neuron. When a neuron spikes these counters are set to 8-bit parameters C_{set}^+ and C_{set}^- and decay by 3-bit parameters C_{decay}^+ and C_{decay}^- each time step. During a pre- (post-) synaptic update, the neuron will activate the crossbar's column (row) word line associated

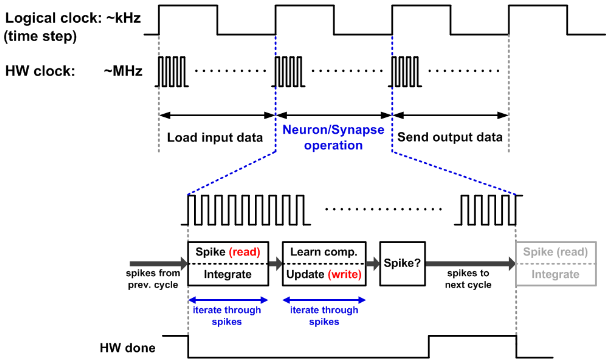


Fig. 4. Timing operation of overall chip.

with its dendrite (axon) and check the C^+ (C^-) counter of all source (target) neurons. To perform a probabilistic update, this counter value is compared with a pseudo-random number generated by a 15-bit linear feedback shift register (LFSR), and the synapse weights are either updated to 1 or 0 according to w_{set}^+ (w_{set}^-), a one bit parameter flag, or left unchanged. If C^+ (C^-) equals zero, an optional flag bit allows C_{zero}, a 6-bit parameter, to be used in place of zero for this probability calculation. In this case the polarity of w_{set}^+ (w_{set}^-) is reversed.

To maximize flexibility, each neuron includes 75 latches to store configuration information for integrate-and-fire function and learning rule parameters. Hebbian, anti-Hebbian, STDP, and anti-STDP learning could be configured with fine-tunable update probability by adjusting the learning parameters. Due in large to the general purpose nature of this neuron, the circuit area in 45nm is 2,500 μm^2.

D. Event-Driven Synchronous Operation

The chip operates in a synchronous fashion with a ~1kHz “logical clock” (corresponding to ~1ms biological time step) that enables “real-time” communication with the external world. Internally, the hardware control circuits run at a faster rate (~1MHz) to step through the integrate-and-spike phase and two learning phases (pre- and post-synaptic) for each neuron spiking event (Fig. 4). Management of these simultaneous (with respect to the logical time step) events is controlled by a finite state machine and a priority encoder (shown in Fig. 1) to regulate array access, maintain the proper order of operations, and to ensure that all spiking events are handled. Through the use of clock-gating techniques, the hardware clock is applied only when signals are queued for processing. In this way, an event-driven system with minimal power dissipation is realized; only those neurons that exceed the threshold in a time step participate in communication and only their associated synapses participate in weight updates.

E. Design Variants

Slim Neuron: Since configuration latches can consume significant area, one variant fixes all spiking and learning parameters for a two-layer learning network [8]. This achieves a 56% neuron area reduction relative to the base design.

4-bit Synapse: To compare the binary synapse design with conventional analog weight approaches, one variant implements synapses as 4-bit weights. The memory element combines four copies of the transposable SRAM cell with common word line terminals. Synapse updates occur in a three-step read-modify-write procedure in accordance with learning rules. The LFSR module is not needed in this variant as synapse weights are not updated probabilistically, which enables a 20% neuron area reduction.

Low Leakage Variant: In a system targeting real-time operation with a ~1MHz hardware clock, leakage power far outweighs active power. By leveraging ultra-high-Vt devices, a ~3X leakage current reduction is achieved compared to the base design with super-high-Vt devices, at the cost of increased σVt due to increased transistor channel doping, which raises the minimum operating voltage of the memory array.

III. Measurement Results

Prototype chips for each of the four design variants were fabricated in 45nm SOI-CMOS. As shown in Fig. 5, each variant occupies a pad-limited area of 4.2 mm^2. The physical design was completed using a combination of custom memory array design and automated synthesis and placement for the neuron and control logic.

The integrate-and-fire function, learning rule and possibility for fully recurrent connectivity implemented here capture the core elements found in networks of spiking neurons, with considerable functional reconfigurability provided by the parameters specified in Section II. To demonstrate the diverse capabilities of the chip, we implemented two possible configurations. The first is a system that relies on STDP to learn patterns delivered in a spatio-temporally sparse fashion [8] and the second is a classic system in the field of neural networks, a Hopfield-like network [9].

The first system learned patterns through auto-association. Patterns were chosen to include some subset of neurons in an input layer. During training, each pattern was presented to the system over 50 time steps, where each element in the pattern had only a small probability (0.05) of appearing each time step. To test recall, a small portion of each pattern was activated (units within the active portion had a 0.05 probability of appearing each time step). As shown in Fig. 6(a), after learning, the network could recall complete learned images when presented with the partial images. Weights in the recall pathway were initialized to zero, but Fig. 6(b) shows how the synapse weights evolved through on-chip learning.

The Hopfield-like network [9] performed an auto-associative task where the network was trained on multiple randomly produced patterns that were presented in full each

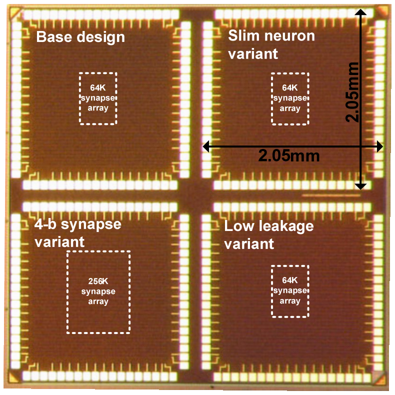
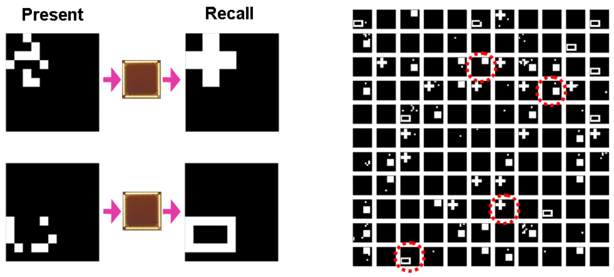


Fig. 5. Chip micrograph.



(a) present and recall of patterns (b) synapse weights after learning

Fig. 6. The effect of learning in the first system is shown. (a) After on-chip learning, the system is presented with an incomplete image (via input spikes) and a complete image is recalled (via output spikes). (b) The weights of synapses connecting different layers of excitatory neurons, after learning, are shown. Examples of the trained patterns are marked with circles. (For both spikes and synapses, white represents ‘1’, and black represents ‘0’.)

time step. Our hardware implementation of this network had a capacity of 0.047 patterns per neuron when the system was used with binary synapses, and this improved to 0.109 in the 4-bit synapse variant, which can be compared to the Hopfield networks theoretical capacity of 0.139.

Measured hardware successfully performed the learning tasks described above down to 0.53V for the base, slim neuron, and 4-bit synapse variants and 0.55V for the low leakage variant (due to increased σVt of the low leakage devices). For massively parallel systems, such near-threshold operation has been shown to provide optimal power efficiency [11], which will be critical to realizing an ultimate brain-scale system. As shown in Fig. 7, the low-leakage variant consumes a total leakage power of 0.5mW at 0.55V. When there is no spiking activity, this leakage power could be reduced by data-retention power-gating as long as the neuron potential and synapse values are maintained. The measured minimum retention voltage for the low-leakage variant is 0.25V, which achieves a leakage power of 26μW – a 20X reduction from 0.55V.

An overall comparison of the area, power, and learning performance of the four design variants is shown in Fig. 8. While application-specific designs provide optimal density, ultra-low-leakage devices provide optimal power, and multi-bit synapses provide optimal learning speed, each design carries significant tradeoffs and must be carefully considered when scaling to a large-scale system.

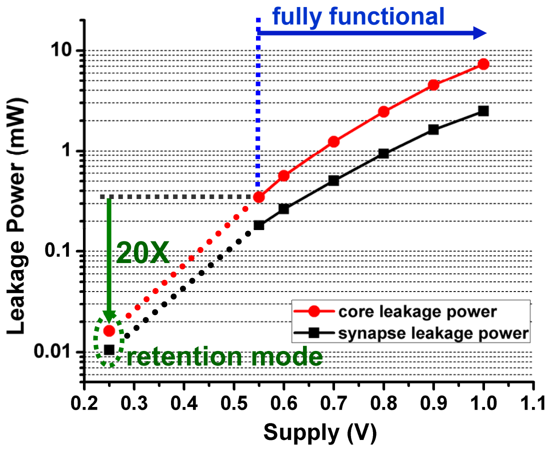
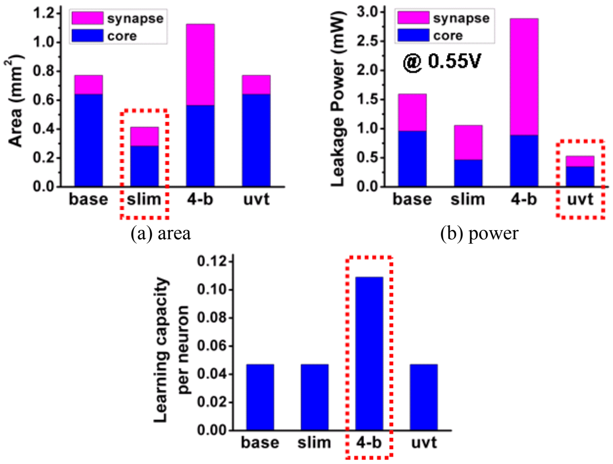


Fig. 7. Voltage scaling and power-gating of the low-leakage variant chip. ‘core’ includes 256 neurons, FSM, priority encoder, and clock distribution. ‘synapse’ includes 64K synapses and according peripheral circuits.



(a) area (b) power

(c) learning capacity

Fig. 8. Density, power, and learning performance (in Hopfield-like network) trade-off of 4 variants.

VI. Conclusion

In this paper, we demonstrated a highly configurable neuromorphic chip with integrated learning for use in pattern classification, recognition, and associative memory tasks. Through the use of digital neuron circuits and a novel transposable crossbar SRAM array, this basic building block addresses the computation, memory, and communication requirements for large-scale networks of spiking neurons and is scalable to advanced technology nodes. Future systems will build on this base to enable ubiquitously deployable ultra-dense, ultra-low power brain-like cognitive computers.

Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency under contract No. HR0011-09-C-0002. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of DARPA or the U.S. Government. Approved for public release, distribution unlimited.

References

[1] W. Gerstner and W. Kistler, *Spiking neuron models*, Cambridge, U.K.: Cambridge Univ. Press, 2002.

[2] S. Song, *et al.*, “Competitive Hebbian learning through spike-timing-dependent synaptic plasticity,” *Nature Neuroscience*, pp. 919-926, 2000.

[3] R. Ananthanarayanan, *et al.*, “The cat is out of the bag: cortical simulations with 10^9 neurons, 10^{13} synapses,” *Proc. Conf. High Perf. Computing Networking, Storage and Analysis (SC09)*, pp. 1-12, Nov. 2009.

[4] C. A. Mead, *Analog VLSI and Neural Systems*, Addison Wesley, 1989.

[5] G. Indiveri, *et al.*, “A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity,” *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 211-221, Jan. 2006.

[6] K. Boahen, “Neurogrid: emulating a million neurons in the cortex,” *Int. Conf. of the Engineering in Medicine and Biology Society*, 2006.

[7] J. Arthur, *et al.*, “Learning in silicon: timing is everything,” *Advances in Neural Information Processing Systems* 18, pp. 75-82, MIT Press, 2006.

[8] S. Esser, *et al.*, “Binding sparse spatiotemporal patterns in spiking computation,” *Int. Joint Conf. on Neural Networks*, pp. 1-9, Jul. 2010.

[9] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proc. Natl. Acad. Sci. USA*, vol. 79, no. 8, pp. 2554-2558, Apr. 1982.

[10] W. Senn and S. Fusi, “Convergence of stochastic learning in perceptrons with binary synapses,” *Phys. Rev. E 71*, 061907, 2005.

[11] L. Chang, *et al.*, “Practical strategies for power-efficient computing technologies,” *Proc. IEEE*, vol. 98, no. 2, pp. 215-236, Feb. 2010.