TrueNorth: a High-Performance, Low-Power Neurosynaptic Processor for Multi-Sensory Perception, Action, and Cognition

Andrew S. Cassidy*, Jun Sawada*, Paul A. Merolla, John V. Arthur, Rodrigo Alvarez-Icaza*, Filipp Akopyan, Bryan L. Jackson, and Dharmendra S. Modha

IBM Research—Almaden Research Center
650 Harry Road, San Jose, CA, USA 95120

*IBM Research—Austin Research Center
11501 Burnet Road, Austin, TX, USA, 78758

Abstract: *IBM's TrueNorth neurosynaptic processor is a radical departure from decades of traditional von Neumann computing. Containing 5.4 billion transistors and fabricated in a 28nm low-power CMOS process technology, TrueNorth contains 1 million neurons and 256 million synapses. With applications ranging from embedded and embodied intelligence to large-scale perceptual analysis of streaming multi-sensory data, this massively parallel processor consumes only 65mW typically.*

Keywords: low-power CMOS; neuromorphic computing.

Introduction

The human brain contains on the order of 10^{11} neurons and 10^{14} synapses, enabling it to perform a remarkable variety of visual, auditory, and other sensory perception tasks such as classification and recognition, as well as higher-level cognitive tasks such as abstraction, reasoning, and deduction. This immense amount of real-time biological information processing requires only approximately 20W of power. Neuromorphic computational architectures emulate the structure and principles of biological computation in artificial substrates such as silicon, in an effort to achieve similar computational performance and energy efficiency.

Based upon these principles, IBM's TrueNorth neurosynaptic processor [1, 2, 3] contains 1 million neurons and 256 million synapses, and is implemented in a 28nm low-power CMOS process technology. It consumes less than 100mW on typical visual classification tasks.

TrueNorth's non-traditional, non-von Neumann architecture embodies several strategies directly relevant to the next generation of computation. First, as an architectural innovation within standard CMOS, this architecture expands the performance potential of CMOS, by extracting more computation out of silicon for less energy. Second, due to the low-power nature, this neuromorphic architecture is readily amenable to 3D VLSI integration. And third, TrueNorth's architectural and computational principles extend to emerging materials as they become a viable reality.

Background

TrueNorth was born out of the multi-institutional DARPA SyNAPSE project beginning in 2008, inspired and deeply rooted in neuroscience [4]. We compiled the largest long-distance wiring diagram of the primate brain, giving insight into the communication architecture of the brain [5]. Our team delivered ground-breaking simulations [6] of large-scale neural systems on the BlueGene Supercomputers, achieving rat scale in 2007 [7], cat scale in 2009 [8], and culminating in human scale in 2012 [9]. Drawing on these neuroscience lessons, we distilled aspects of the architectural principles of biological computation, communication, and memory into the TrueNorth processor, which we introduced to the world in 2014 [1]. TrueNorth demonstrated orders of magnitude better energy efficiency over traditional processors on neural network algorithms [2]. Our efforts are now focused on both embedded and large-scale TrueNorth based systems, as well as biologically inspired algorithms and applications to run on those systems.

Architecture

Seven architectural principles summarize the guiding philosophy behind the TrueNorth architecture [3].

1. **Minimize active power** using event-driven asynchronous and synchronous circuits, eliminating global clock networks, collocating memory and computation, as well as implementing sparse memory access patterns.
2. **Minimize static power** using a low-power CMOS manufacturing process.
3. **Maximize parallelism** with 4096 cores, and using time-division multiplexing of computation in order to minimize core area.
4. **Real-time operation** via hierarchical communication (high-fanout crossbar and a network-on-chip), as well as global system (timestep) synchronization.
5. **Scalability** by tiling cores and chips, and locally generating core execution signals.
6. **Defect Tolerance** with circuit-level (memory) redundancies, and core-level redundancy at the architectural-level.
7. **Hardware-software one-to-one equivalence** using a fully digital implementation and deterministic global system synchronization.

The TrueNorth architecture is composed of programmable neurosynaptic cores (Fig. 1a). Each core contains 256 inputs (axons) and 256 computational elements (neurons). Axons are connected to neurons via the programmable 256x256 synaptic crossbar. Active synapses (denoted by a black circle) signify a connection from an axon to a neuron. Fig. 1b depicts a scaled-down 4x3 array of neurosynaptic cores, and the high-speed 2D mesh routing network that

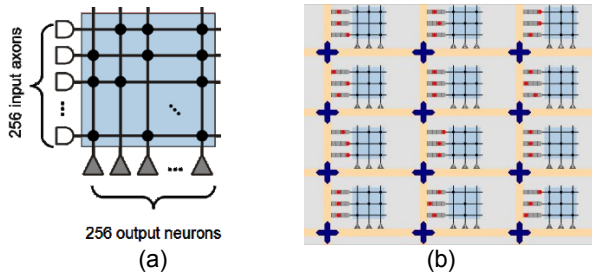

Figure 1. Logical view of TrueNorth: (a) a single neurosynaptic core. (b) 2D array of neurosynaptic cores, interconnected by a mesh routing network.

interconnects the cores. A TrueNorth processor contains an array of 4096 neurosynaptic cores. The data communication between cores as well as with external components is carried out using 'spikes,' analogous to the binary signals that propagate information between biological neurons.

Fig. 2 depicts the layout of the TrueNorth chip, a 64x64 2D array of neurosynaptic cores. At the edge of the array, the mesh network buses are merged into a single datapath for transmission over I/O pads. (Conversely, on the input side, the single input datapath is split into parallel buses.) Thus, TrueNorth processors seamlessly tile in two dimensions not only at the core level, but also at the chip level.

Operation of a single neurosynaptic core (Fig. 2, right) is as follows. Input data (spikes) asynchronously arrive at the scheduler input axon buffers (A) with an axon address and timeslot. As simulation time progresses (1ms timestep), buffered spikes are read out of the scheduler (B) and compared with the synaptic connections in the crossbar (C). The core SRAM also stores individual neuron parameters (D) for each neuron in the core, as well as the current neuron state (E). For every active synaptic connection with an active input, the neuron will integrate the synaptic input with its state (F). If the neuron state exceeds a programmable threshold, the neuron fires a spike, sending its output into the routing network towards its destination. Every timestep, all 256 neurons evaluate all of their 256 synaptic inputs and update their state.

Systems

We highlight two TrueNorth based boards shown in Fig. 3. The IBM Neurosynaptic System 4[14] Evaluation Platform (NS1e) is targeted at embedded applications (dimensions: 125mm x 69mm, and weight: 98g). It contains a single TrueNorth processor with 4092 user cores (four cores are reserved for manufacturing defect tolerance). The board also hosts a Xilinx Zynq Z-7020 providing two ARM Cortex-A9 cores and a configurable FPGA fabric. In terms of interfaces, the NS1e has an 10/100/1000 Ethernet port for system communication and data transmission, a microSD card slot for operating system and data storage, and a Micro USB port. There are 32-bits of bidirectional GPIO, including user selectable integrated pulse width modulation (PWM) signal generators for motor control. The NS1e also exposes to the user four input and one output TrueNorth-native bundled data asynchronous interfaces, for directly interfacing native event driven sensors to TrueNorth. For embedded applications, the NS1e contains a BMP-180 I2C-based pressure sensor for altitude measurements and temperature sensor, as well as a MPU-9150 I2C-based motion processing unit including gyroscope, accelerometer, magnetometer (compass), and motion coprocessor.

The NS16e board (Fig. 3c), with 16 TrueNorth neurosynaptic processors, has 16x larger computational capacity (16 million neurons and 4 billion synapses), for running larger scale neural networks. The TrueNorth processors are tiled in a 2D array, natively and seamlessly intercommunicating. Data enters and exits the array on the left side, via a set of interface FPGAs. The user interface to

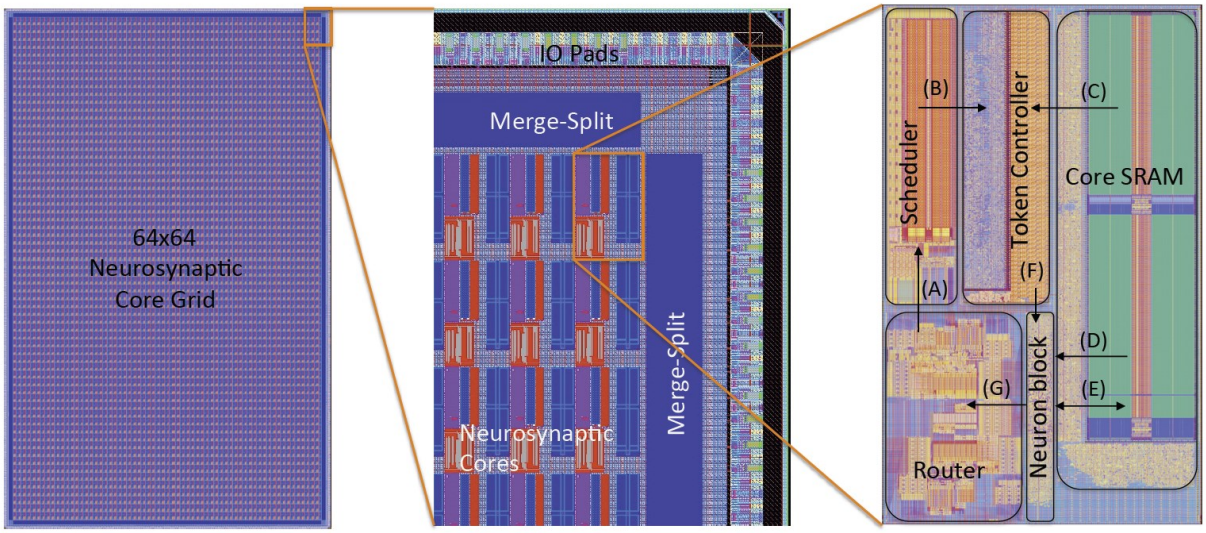

Figure 2. TrueNorth Layout: (left) 2D array of 64x64 neurosynaptic cores. (middle) top corner of array (3x3 cores), merge-split peripheral circuits, and I/O pads. (right) a single neurosynaptic core, labeled with its five major functional blocks.

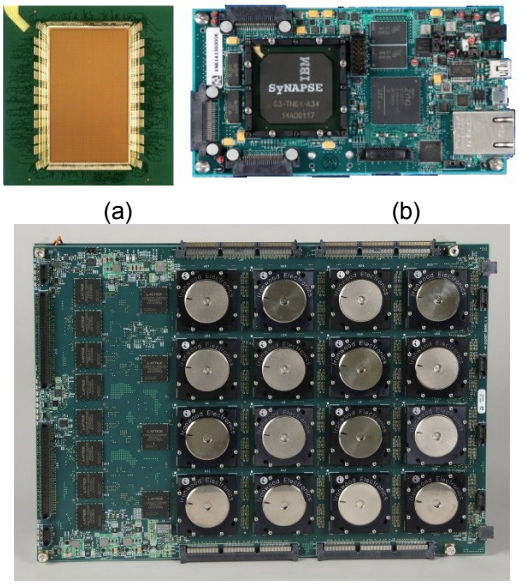

Figure 3. (a) TrueNorth chip die photo (5.4 billion transistors, 4.3 cm^2) with package substrate. (b) IBM Neurosynaptic System 4^{14} Evaluation Platform (NS1e). (c) NS16e with a 4x4 array of socketed TrueNorth processors.

the NS16e is through PCIe and/or Ethernet, and the form factor is 215mm x 285mm (approximately an 8.5" x 11" sheet of paper). The NS16e defines a modular building block for larger, scaled-up, multi-board and multi-rack TrueNorth installations. TrueNorth applications scale up in many orthogonal dimensions. All of the following examples scale as a function of the underlying system size: the scene size that can be covered, the number of input channels (spectral channels, frequencies, receiver dimensionality), the sensory modalities (sound, vision, RADAR, LIDAR, ultrasound, etc.), the network precision/accuracy, the number of objects to classify/recognize, the depth of the processing pipeline (feature extraction, saliency, recognition, association, prediction, action, etc.), as well as mixtures of systems with varying parameters.

Applications

Leveraging recent advances in machine learning and Deep Neural Networks, TrueNorth enables the capability to deploy high accuracy classifiers [10], [11] in very low-power applications. For example, the NeoVision [12] visual object recognition system finds, localizes, and classifies five classes of relevant objects (people, bikes, cars, trucks and buses). This system operates in real-time (30fps) across the full visual scene of a high-definition video stream (Fig. 4).

Deep neural networks have demonstrated state-of-the-art performance in a wide variety of tasks, including image recognition [13], speech recognition [14], and natural language processing [15]. These types of network can be run on TrueNorth for small amounts of energy per classification, to achieve both high-accuracy and low-power applications. One such network is shown in Fig. 5 [10] for classifying

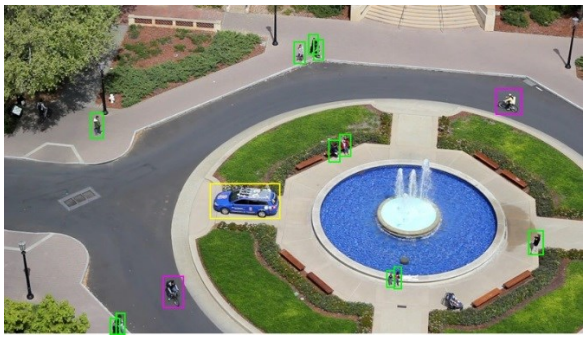

Figure 4. NeoVision Tower Dataset: detection, localization and classification of five types of objects; cars are labeled in yellow, cyclists in fuchsia, and pedestrians in green.

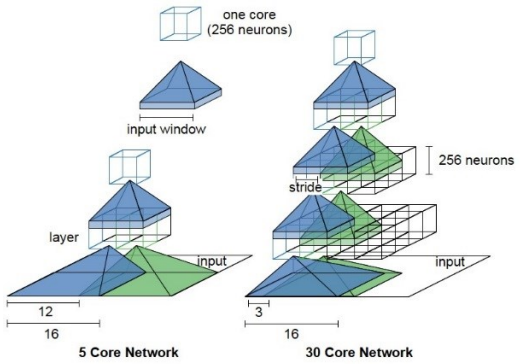

Figure 5. Deep Neural Networks for handwritten digit classification (MNIST) [10].

handwritten digits (MNIST dataset). The network input is 28x28 gray-scale pixels and the network output is the strength of prediction of each of the 10 digit classes.

Programming Model: The TrueNorth neurosynaptic processor is programmed using two basic approaches: 1) the corelet neurosynaptic programming language [16], and 2) a back-propagation based flow [10] for training classifiers. To build a TrueNorth executable, three things must be specified: 1) the neuron parameters, specifying the computation to be performed by each neuron, 2) the synaptic crossbar connections, specifying the connectivity between input axons and neurons, as well as 3) the inter-core connectivity, connecting neuron outputs to axonal inputs.

In the first programming approach, developers explicitly define all of the three parameters sets using the corelet programming language. A corelet is defined as a functional encapsulation of a network of neurosynaptic cores that collectively perform a specific task. The corelet programming language is hierarchical, modular, and composable, such that corelet building blocks can be readily combined and reused to design networks that perform the required computation. In order to expedite the design time for new functions, developers can use libraries of neuron functions [17] or pre-existing corelets.

The parameters for deep neural networks are not manually programmed, rather they are learned through a training

process using a labelled input dataset. In the second programming approach, as an alternative to manual programming, we have developed a training flow for learning TrueNorth programs from data. This training flow is natively compatible with TrueNorth's low-precision synapses and connectivity constraints, yet achieves near state-of-the-art classification accuracy, while consuming significantly less energy than traditional approaches. This flow leverages standard machine learning toolkits, such as Caffe and MatConvNet for dataset manipulation and network training. The resulting neural network is written out as a corelet network for execution on TrueNorth. These classifier corelets can be composed with other corelets to create larger systems.

Conclusion

Low-power, real-time, high-performance computing has the potential to dramatically change a wide range of defense and security systems, in addition to medical, automotive, mobile, consumer, and industrial applications. In both small- and large-scale systems, TrueNorth, acting as a cognitive and perceptual processor, plays a crucial role in filtering and transforming enormous amounts of raw data into small streams of refined information.

Consuming less than 100mW, TrueNorth is ideal for ultra-low power embedded applications where size, weight, and power are at a premium. This ranges from lightening the burden on the ground-based warfighter to embedding perceptual capabilities into self-driving sea, air, and ground vehicles. Equipped with numerous sensors: visual, audio, motion, ultrasound, LIDAR, and RADAR, these vehicles require systems to extract salient and relevant information from vast quantities of noisy and time/space varying data in order to perform even basic tasks, such as navigation.

At the large scale, TrueNorth's outstanding energy per classification performance and scalability, are ideal for real time analysis of massive streams of data from a range of sensors and sources. This is key in operational theatres where there is limited access to megawatts of power required to support large data centers. Finite bandwidth limitations and latency restrict the amount of data that can be relayed across the globe to supporting sites for data processing. Instead, edge analytics, processing the data at the sensor, enables fast, scalable extraction of intelligence.

TrueNorth's orders of magnitude improvement in computing efficiency will lead a technological paradigm shift for defense, intelligence, and homeland security systems, enabling a leap in overall system performance and the security of our collective global society.

Acknowledgements

This research was sponsored by DARPA 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.

References

1. P. A. Merolla, et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.
2. A. S. Cassidy, et al. "Real-time scalable cortical computing at 46 Giga-synaptic OPS/watt with ~100x Speedup in Time-to-Solution and ~100, 000x Reduction in Energy-to-Solution," in *SC*'2014.
3. F. Akopyan, et al. "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE TCAD*, vol. 34, no. 10, Oct. 2015.
4. D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, et al. "Cognitive computing," *Communications of the ACM*, vol. 54, no. 8, 2011.
5. D. S. Modha and R. Singh. "Network architecture of the long distance pathways in the macaque brain," *PNAS*, vol. 107, no. 30, 2010.
6. R. Preissl, et al. "Compass: A scalable simulator for an architecture for cognitive computing," *SC'*2012.
7. R. Ananthanarayanan and D. S. Modha. "Anatomy of a cortical simulator," *SC'*2007.
8. R. Ananthanarayanan, S. K. Esser, H. D. Simon, and D. S. Modha. "The cat is out of the bag: cortical simulations with 10^9 neurons, 10^{13} synapses" *SC'*2009.
9. T. M. Wong, et al. "10^{14}", *IBM Research Divsion, Research Report*, RJ10502, 2012.
10. S. Esser, R. Appuswamy, P. Merolla, J. Arthur, D. Modha. "Backpropagation for Energy-Efficient Neuromorphic Computing," *NIPS*, December, 2015.
11. D. Mart´ı, et al. "Energy-efficient neuromorphic classifiers." *arXiv*:1507.00235 (2015).
12. R. Kasturi, et al. "Performance evaluation of neuromorphic-vision object recognition algorithms." *IEEE ICPR*, 2014.
13. O. Russakovsky, et al. "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, 2015.
14. G. Hinton, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal Processing Magazine*, vol 29, no. 6, 2012.
15. R. Collobert, and J. Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." *ICML*, ACM, 2008.
16. A. Amir, et al. "Cognitive computing programming paradigm: A corelet language for composing networks of neuro-synaptic cores," *IJCNN*, IEEE, 2013.
17. A. S. Cassidy, et al. "Cognitive computing building block: A versatile and efficient digital neuron model for neurosynaptic cores," *IJCNN*, IEEE, 2013.