Reverse Engineering the Cognitive Brain in Silicon

Gert Cauwenberghs

Department of Bioengineering
Institute for Neural Computation
UC San Diego

http://inc.ucsd.edu
BRAIN Initiative
Brain Research through Advancing Innovative Neurotechnologies

- Interdisciplinary collaborative initiative engaging broad participation in the sciences, engineering, arts, and humanities alongside neuroscientists in advancing our understanding of the circuits and function of the healthy and diseased brain
- Initial investments with contributions from industry (NIH, NSF, Salk Institute, Allen Brain Institute, etc.)
A Nanotechnology-Inspired Grand Challenge for Future Computing

OCTOBER 20, 2015 AT 6:00 AM ET BY LLOYD WHITMAN, RANDY BRYANT, AND TOM KALIL

Summary: Today, the White House is announcing a grand challenge to develop transformational computing capabilities by combining innovations in multiple scientific disciplines.

In June, the Office of Science and Technology Policy issued a Request for Information seeking suggestions for Nanotechnology-Inspired Grand Challenges for the Next Decade. After considering over 100 responses, OSTP is excited to announce the following grand challenge that addresses three Administration priorities—the National Nanotechnology Initiative, the National Strategic Computing Initiative (NSCI), and the BRAIN initiative:

Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain.

While it continues to be a national priority to advance conventional digital computing—which has been the engine of the information technology revolution—current technology falls far short of the human brain in terms of both the brain’s sensing and problem-solving abilities and its low power consumption. Many experts predict that fundamental physical limitations will prevent transistor technology from ever matching these twin characteristics. We are therefore challenging the nanotechnology and computer science communities to look beyond the decades-old approach to computing based on the Von Neumann architecture as implemented with transistor-based processors, and chart a new path that will continue the rapid pace of innovation beyond the next decade.

https://www.whitehouse.gov/blog/2015/10/15/nanotechnology-inspired-grand-challenge-future-computing
Today’s Hottest Microchip

*Intel’s Itanium 2*

**The numbers …**
- 0.5 billion transistors in 120nm CMOS
- 1.6GHz clock, 64-bit instruction, 9MB L3 cache, 6.4GB/s I/O
- 2553 SPECfp_base2000 (30% faster than 2.8GHz P4)
- 130 Watts

**… and what they mean**

Faster/cooler:
- *Scientific computing*
- *Database search*
- *Web surfing*
- *Video games*

What about intelligence?
Chips and Brains

• **Itanium:**
  - $3 \times 10^9$ floating op/s
    • $5 \times 10^8$ transistors
    • $2 \times 10^9$ Hz clock
  - $10^{10}$ Hz memory I/O
    • 128-b data bus @ 400MHz
  - 130 Watts

• **Human brain:**
  - $10^{15}$ synaptic op/s
    • $10^{15}$ synapses
    • 1 Hz average firing rate
  - $10^{10}$ Hz sensory/motor I/O
    • $10^8$ nerve fibers
  - 25 Watts

• Silicon technology is approaching the raw computational power and bandwidth of the human brain.

• However, to emulate brain intelligence with chips requires a radical paradigm shift in computation:
  - Distributed representation in massively parallel architecture
    • *Local adaptation and memory*
    • *Sensor and motor interfaces*
  - Physical foundations of computing
Analysis by Synthesis

Richard Feynman

Carver Mead
Multi-scale levels of investigation in analysis of the central nervous system (adapted from Churchland and Sejnowski 1992) and corresponding neuromorphic synthesis of highly efficient silicon cognitive microsystems. Boltzmann statistics of ionic and electronic channel transport provide isomorphic physical foundations.

Physics of Neural Computation

Silicon and Lipid Membranes

Mead, 1989

Voltage-dependent $p$-channel
- Hole transport between source and drain
- Gate controls energy barrier for holes across the channel
- Boltzmann distribution of hole energy produces exponential decrease in channel conductance with gate voltage

Voltage-dependent conductance
- $K^+/Na^+$ transport across lipid bilayer
- Membrane voltage controls energy barrier for opening of ion-selective channels
- Boltzmann distribution of channel energy produces exponential increase in $K^+/Na^+$ conductance with membrane voltage

Squid giant axon (Hodgkin and Huxley, 1952)
Event-Coding Silicon Retina

Zaghloul and Boahen, 2006

- Models coding and communication of visual events in the mammalian retina and optic nerve
  - Integrated photosensors (rods)
  - On and off transient and sustained ganglia cell outputs
    - Spatiotemporal compressed coding and communication in optic nerve
    - Address-event coding of spikes
Change Threshold Detection APS CMOS Imager

Chi, Mallik, Clapp, Choi, Cauwenberghs and Etienne-Cummings (2007)

- Event-driven video compression
  - Change detection and threshold encoding on the focal plane
- 6T pixel combines APS and change event coding
- 4.3mW power at 3V and 30fps
Reconfigurable Synaptic Connectivity and Plasticity

From Microchips to Large-Scale Neural Systems

Address-Event Representation

Neural Systems

Synaptic Plasticity & Wiring

Multi-Chip Systems
Address-Event Representation (AER)
Lazzaro et al., 1993; Mahowald, 1994; Deiss 1994; Boahen 2000

- AER emulates extensive connectivity between neurons by communicating spiking events time-multiplexed on a shared data bus.
- Spikes are represented by two values:
  - Cell location (address)
  - Event time (implicit)
- All events within $\Delta t$ are “simultaneous”
Address-Event Synaptic Connectivity
Goldberg, Cauwenberghs and Andreou, 2000

- ‘Virtual’ synapses
  - Dynamically reconfigurable
  - Wide-ranging connectivity
  - Rewiring and synaptic plasticity

- Quantal release: \( R = n \cdot p \cdot q \)
  - \( n \): multiplicity (repeat event)
  - \( p \): probability of release (toss a coin)
  - \( q \): quantity released (set amplitude)
Hierarchical Vision and Saliency-Based Acuity Modulation
Vogelstein, Mallik, Culurciello, Cauwenberghs, and Etienne-Cummings, NECO 2007

IFAT Cortical Model
4800 silicon neurons
4,194,304 synapses

Octopus Silicon Retina
80 x 60 pixels
AER spiking output

View-Tuned Cells
Composite Cells
Feature Cells
Complex Cells
Simple Cells

WTA
Local Saliency
Spatial Features
Retina
SAM

OR image
Simple cell response
Saliency map
Spike Timing-Dependent Plasticity

Bi and Poo, 1998
Spike Timing-Dependent Plasticity

*in the Address Domain*

Vogelstein et al, NIPS*2002
Spike Timing-Dependent Plasticity on the IFAT

Vogelstein et al, NIPS*2002
Achieving (or surpassing) human-level machine intelligence will require a convergence between:

- **Advances in computing resources approaching connectivity and energy efficiency levels of computing and communication in the brain**;
- **Advances in deep learning methods, and supporting data, to adaptively reduce algorithmic complexity**.

Scaling and Complexity Challenges

- Scaling the event-based neural systems to performance and efficiency approaching that of the human brain will require:
  - Scalable advances in silicon integration and architecture
    - Scalable, locally dense and globally sparse interconnectivity
      - Hierarchical address-event routing
    - High density \((10^{12}\) neurons, \(10^{15}\) synapses within 5L volume)
      - Silicon nanotechnology and 3-D integration
    - High energy efficiency \((10^{15}\) synOPS/s at 15W power)
      - Adiabatic switching in event routing and synaptic drivers
  - Scalable models of neural computation and synaptic plasticity
    - Convergence between cognitive and neuroscience modeling
    - Modular, neuromorphic design methodology
    - Data-rich, environment driven evolution of machine complexity
Long-Range Configurable Synaptic Connectivity

Comparison of synaptic connection topologies for several recent large-scale event-driven neuromorphic systems and the proposed hierarchical address-event routing (HiAER), represented diagrammatically in two characteristic dimensions of connectivity: expandability (or extent of global reach), and flexibility (or degrees of freedom in configurability). Expandability, measured as distance traveled across the network for a given number of hops $N$, varies from linear and polynomial in $N$ for linear and mesh grid topologies to exponential in $N$ for hierarchical tree-based topologies. Flexibility, measured as the number of target destinations reachable from any source in the network, ranges from unity for point-to-point (P2P) connectivity and constant for convolutional kernel (Conv.) connectivity to the entire network for arbitrary (Arb.) connectivity.

MMAER: Multicasting Mesh AER; WS: Wafer-Scale.
Hierarchical Address-Event Routing (HiAER)

(a) Hierarchical neural network with ascending and descending neural projections. Physical neurons are represented by o and p, and inserted relay neurons (RN) interfacing across hierarchical partitions are denoted by q, k, l, n, m. Italic indices j and j – 1 represent levels in the hierarchy, while boldface indices i and i – 1 represent individual blocks within one level in the hierarchy.

(b) The edge-vertex-dual of the hierarchical routing network.

(c) Corresponding entries within the Synaptic Routing Table (SRT).

Joshi et al, 2010; Park et al, 2011, 2015
(d) Example network with 16 neurons and weighted synaptic connections.

(e) Example partitioning into hierarchical neural network with ascending and descending projections through inserted relay neurons.

(f) Corresponding edge-vertex-dual HiAER implementation with synaptic routing tables (SRT) at each level in the hierarchy.

Joshi et al, 2010; Park et al, 2011, 2015
Hierarchical Address-Event Routing (HiAER)

(a) Simplified system architecture of a HiAER node at Level 1 (leaf in the hierarchy), routing synaptic events through the Synaptic Routing Table (SRT) between physical neurons in the local Integrate-and-Fire Array Transceiver (IFAT) and relay neurons on the $L_1$ bus. The SRT maps incoming events from any neuron onto outgoing events either to the final synaptic destination on the IFAT (along with synaptic strength $w$), or up the hierarchy through the $L_1$ bus (along with timing information for axonal delay $d$).

(b) Digital system architecture of a HiAER node at Level $n > 1$ (higher in the hierarchy), largely identical to Level 1 except for substitution of the IFAT with a $L_n - 1$ bus, and of the $L_1$ bus with a $L_n$ bus. In the absence of physical neurons, events are transmitted only between relay neurons higher and/or lower in the hierarchy (along with timing information for axonal delay $d$).
Hierarchical Address-Event Routing (HiAER) Integrate-and-Fire Array Transceiver (IFAT) for scalable and reconfigurable neuromorphic neocortical processing [Park et al, 2012; Yu et al, 2012].  (a) Dynamic reconfigurable synaptic connectivity across IFAT arrays of addressable neurons is implemented by routing neural spike events through DRAM synaptic routing tables (SRT).  (b) The IFAT neural array multiplexes and integrates (top traces) incoming spike synaptic events to produce outgoing spike neural events (bottom traces).  (c) Full-size HiAER-IFAT network with 4 boards, each with 4 IFAT modules, serving 1M neurons and 1G synapses, and spanning 4 levels in connection hierarchy.  (d) Each IFAT chip module comprises a 65k-neuron Tezzaron 130nm CMOS IFAT microchip, Xilinx Spartan-6 FPGA (Level 1 HiAER), and two 2Gb DDR3 SDRAM SRTs serving 65M synapses.  (e) Each neural cell models conductance based membrane dynamics in proximal and distal compartments for synaptic input with programmable axonal delay, conductance, and reversal potential.  IFAT chip measured energy consumption is 48 pJ per spike event, several orders of magnitude more efficient than emulation on CPU/GPU platforms.
IFAT Thermodynamics of Neural Excitability
Yu, Park, Joshi, Maier, Cauwenberghs, BioCAS 2012

\[ V_{gs} = \kappa E_{exc} - V_{th} - V_{TH,N} \]

\[ P(out|exc) = f(V_{gs}) = \frac{1}{1 + e^{-\frac{V_{gs}}{V_T}}} \]
Large-Scale Reconfigurable Neuromorphic Computing

- Integrate-and-fire array transceiver (IFAT) as digitally programmable analog neural supercomputer
- Biophysical detail in neural and synaptic continuous-time dynamics
- Record high density: 65k two-compartment neurons with 65M reconfigurable conductance-based synapses
- Record low energy: 22 pJ per synaptic event
- Real-time at 73M spikes per second

## Large-Scale Reconfigurable Neuromorphic Computing

### Technology and Performance Metrics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology (nm)</strong></td>
<td>130</td>
<td>28</td>
<td>180</td>
<td>180</td>
<td>90</td>
</tr>
<tr>
<td><strong>Die Size (mm²)</strong></td>
<td>102</td>
<td>430</td>
<td>50</td>
<td>168</td>
<td>16</td>
</tr>
<tr>
<td><strong>Neuron Type</strong></td>
<td>Digital Arbitrary</td>
<td>Digital Accumulate &amp; Fire</td>
<td>Analog Conductance Integrate &amp; Fire</td>
<td>Analog Shared-Dendrite Conductance I&amp;F</td>
<td>Analog 2-Compartment Conductance I&amp;F</td>
</tr>
<tr>
<td># Neurons</td>
<td>5216 (^1)</td>
<td>1M (^2)</td>
<td>512</td>
<td>65k</td>
<td>65k</td>
</tr>
<tr>
<td>Neuron Area (µm²)</td>
<td>N/A (^1)</td>
<td>3325 (14) (^2)</td>
<td>1500</td>
<td>1800</td>
<td>140</td>
</tr>
<tr>
<td>Peak Throughput (Events/s)</td>
<td>5M</td>
<td>1G</td>
<td>65M</td>
<td>91M</td>
<td>73M</td>
</tr>
<tr>
<td>Energy Efficiency (J/SynEvent)</td>
<td>8n</td>
<td>26p</td>
<td>N/A</td>
<td>31p</td>
<td>22p</td>
</tr>
</tbody>
</table>

1 Software-instantiated neuron model  
2 Time-multiplexed neuron (256x)


Phase Change Memory (PCM) Nanotechnology

Intel/STmicroelectronics (Numonyx) 256Mb multi-level phase-change memory (PCM) [Bedeschi et al, 2008]. Die size is 36mm² in 90nm CMOS/Ge2Sb2Te5, and cell size is 0.097µm². (a) Basic storage element schematic, (b) active region of cell showing crystalline and amorphous GST, (c) SEM photograph of array along the wordline direction after GST etch, (d) I-V characteristic of storage element, in set and reset states, (e) programming characteristic, (f) I-V characteristic of pnp bipolar selector.

- Scalable to high density and energy efficiency
  - < 100nm cell size in 32nm CMOS
  - < pJ energy per synapse operation
Hybridization and nanoscale integration of CMOS neural arrays with phase change memory (PCM) synapse crossbar arrays. (a) Nanoelectronic PCM synapse with spike-timing dependent plasticity (STDP) [Kuzum et al., 2011]. Each PCM element implements a synapse with conductance modulated through phase transition as controlled by timing of voltage pulses. (b) CMOS IFAT array vertically interfacing with nanoscale PCM synapse crossbar array by interleaving via contacts to crossbar rows. The integration of IFAT neural and PCM synapse arrays externally interfacing with HiAER neural event communication combines the advantages of highly flexible and reconfigurable HiAER-IFAT neural computation and long-range connectivity with highly efficient (fJ/synOP range energy cost) local synaptic transmission.
Deep Learning in Spike-Based Neuromorphic Systems

- **Neural Sampling**: Integrate & Fire (I&F) neurons can perform MCMC sampling of a Boltzmann distribution.
- Restricted Boltzmann Machines can be trained using STDP.

- 92% accuracy on MNIST hand-written digit recognition task.

We identified conditions under which spike trains from general integrate-and-fire neurons in the presence of noise generate Monte-Carlo Markov Chain (MCMC) samples from a Boltzmann distribution.

This framework provides the foundation for event-driven on-line stochastic learning using contrastive divergence in Boltzmann machines.

Event-Driven Contrastive Divergence

On-line Training of Boltzmann Machines Using STDP

- Emulates contrastive divergence (CD) for training standard Restricted Boltzmann Machines (RBMs) using neural sampling with integrate-and-fire neurons.
- On-line spike event-driven training using spike-timing dependent plasticity (STDP)
  - Temporally symmetric form produces the correlations $\langle vh \rangle$ in on-line form
  - Modulation $g(t)$ controls wake-sleep phases (data vs. reconstruction)

Event-Driven Contrastive Divergence
Learning a Model of MNIST Hand-Written Digits

- MNIST hand-written digit recognition accuracy:
  - \textit{CD with standard RBM}: 93.6%
  - \textit{eCD with neural sampling}: 91.9%
- Extends to deep learning across multiple RBM layers for greater accuracy

**Event-Driven Contrastive Divergence**

*Inference, Generation, and Cue Integration*

- Generative power of the Boltzmann machine model:
  - **Bottom-up**: Classification of incoming data
  - **Top-down**: Generation of prototypical data for a class label
  - **Hybrid**: Cue integration with missing data based on class label priors

Spiking Synaptic Sampling Machine (S$^3$M)

Biophysical Synaptic Stochasticity in Inference and Learning

- Stochastic synapses for spike-based Monte Carlo sampling
  - Models biophysical origins of noise in neural systems
  - Activity dependent noise: multiplicative synaptic sampling rather than additive neural sampling

- Online unsupervised learning with STDP
  - Biophysical model of spike-based learning
  - Event-driven contrastive divergence

![Diagram showing S$^3$M with synaptic stochasticity as biophysical model of continuous DropConnect and comparison with RBM](image)

The S$^3$M requires fewer synaptic operations (SynOps) than the equivalent Restricted Boltzmann Machine (RBM) requires multiply-accumulate (MAC) operations at the same accuracy.

Spike-Timing Dependent Eligibility

Reinforcement Learning by Reward Modulation of STDP

Spike timing-dependent eligibility (STDE):
- Variant on biologically inspired spike timing-dependent plasticity (STDP)
- Quantifies the sensitivity of post-synaptic spiking probability, conditioned on timed pre-synaptic spike input, to synaptic strength
- Direct replacement for input activity term in Hebb-type incremental outerproduct update rules for gradient-based learning in rate-based ANNs

Temporal-difference reinforcement learning
- STDE-based Dopamine modulation of reward

P. Frady et al, SfN 2009
Large-Margin Kernel Regression

MAP Forward Decoding

MAP Forward Decoding

Class Identification

Sequence Identification

Kerneltron: massively parallel support vector “machine” (SVM) in silicon (JSSC 2007)

Sub-microwatt speaker verification and phoneme recognition (NIPS’2004)
**Kerneltron: Adiabatic Support Vector “Machine”**

Karakiewicz, Genov and Cauwenberghs, 2007

\[ y = \text{sign} \left( \sum_{i \in S} \lambda_i y_i K(x, x_i) + b \right) \]

- **1.2 TMACS / mW**
  - adiabatic resonant clocking conserves charge energy
  - energy efficiency on par with human brain (10^{15} SynOP/S at 15W)

Classification results on MIT CBCL face detection data

Karakiewicz, Genov, and Cauwenberghs, VLSI 2006; CICC 2007

---

Gert Cauwenberghs  
Reverse Engineering the Cognitive Brain in Silicon  
gert@ucsd.edu
Silicon support vector machine (SVM) and forward decoding kernel machine (FDKM)

Forward decoding MAP sequence estimation

- Subthreshold translinear MOS circuits
- Programmable with floating-gate non-volatile analog storage

Biometric verification

840 nW power
Closing the Loop: Interactive Neural/Artificial Intelligence

Neuromorphic Engineering

Learning & Adaptation

Micropower Mixed-Signal VLSI

Neuro Bio

Neurosystems Engineering

Adaptive Sensory Feature Extraction and Pattern Recognition

Biosensors, Neural Prostheses and Brain Interfaces
Distributed Brain Dynamics of Human Motor Control


NSF EFRI-1137279: Mind, Machines and Motor Control (M3C)
Real-Time Estimation and 3D Visualization of Source Dynamics and Connectivity Using Wearable EEG

Mobile Dry EEG Brain-Body Activity Imaging

Cognionics 64-channel wireless dry EEG system
http://www.cognionics.com

UCSD SCCN SIFT and ASR
http://sccn.ucsd.edu/wiki/SIFT

Mobile dry EEG ASR and SIFT video demonstration
CMOS Imaging in Awake Behaving Rats
Murari, Etienne-Cummings, Cauwenberghs, and Thakor (2010)

- First simultaneous behavioral and cortical imaging from untethered, freely-moving rats.
Integrated Systems Neuroengineering

Neural Systems

Learning & Adaptation

Silicon Microchips

Human/Bio Interaction

Environment

Sensors and Actuators

Neuromorphic/Neurosystems Engineering