Reverse Engineering the Cognitive Brain in Silicon and Memristive Neuromorphic Systems

Gert Cauwenberghs

Integrated Systems Neuroengineering Laboratory
Department of Bioengineering, Jacobs School of Engineering
Institute for Neural Computation
UC San Diego

http://isn.ucsd.edu
Lee Sedol vs. AlphaGo

Go World Champion vs. Google DeepMind

~ 100 W  ~ 100 kW
Steep Costs of Today’s Deep Learning

Energy Efficiency Spectrum

<table>
<thead>
<tr>
<th>Technology</th>
<th>Energy Efficiency Spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU + DDR3</td>
<td>2.10^{10}</td>
</tr>
<tr>
<td>FPGA/GPU + HBM2.0</td>
<td>2.10^{11}</td>
</tr>
<tr>
<td>Compute-in-memory (CIM)</td>
<td>10^{12}</td>
</tr>
<tr>
<td>Memristive CIM</td>
<td>7.10^{13}</td>
</tr>
<tr>
<td>Human brain</td>
<td>10^{15}</td>
</tr>
<tr>
<td>Resonant adiabatic CIM</td>
<td>10^{17}</td>
</tr>
<tr>
<td>Landauer limit (Op/s)/W</td>
<td>2.10^{20}</td>
</tr>
</tbody>
</table>

- **J/op [log]**
  - 5.10^{-10}
  - 5.10^{-11}
  - 10^{-12}
  - 10^{-13}
  - 10^{-15}
  - 10^{-17}
  - 5.10^{-21}

- **Energy Consumption**
  - 70 pJ/b
  - 0.5 nJ/MAC
  - 7 pJ/b
  - 50 pJ/MAC
  - 1 pJ/MAC
  - 13 fJ/MAC
  - 1 fJ/SynOp
  - 10 aJ/MAC
  - kT/bit

- **Performance Metrics**
  - 1.3 10^{10} MB/s
  - 8 W power
  - 5 10^{11} MB/s
  - 30 W power
  - 65nm CMOS
  - 130nm
  - 10 Hz spike rate
  - 10^{15} synapses
  - 22nm FDSOI
  - 10^{16} SynOp/s
  - 10 W metabolic power

- **Temperature**
  - Room T
  - 300 K
Neuromorphic Engineering
“in silico” neural systems design

Neuromorphic Engineering

Neural Systems

Learning & Adaptation

VLSI Microchips

Gert Cauwenberghs
Reverse Engineering the Cognitive Brain in Silicon and Memristive Neuromorphic Systems
gcauwenberghs@ucsd.edu
The Computer and the Brain

John von Neumann

John von Neumann

The Neumann Computer & the Brain

John von Neumann

Foreword by Ray Kurzweil
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.

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
Achieving (or surpassing) human-level machine intelligence requires 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.

Machine Complexity

- Throughput
- Memory
- Power
- Size

Task Complexity

- Search tree breadth
- Depth

Deep digital search

- Rule-based cognition

Collective analog computation

- Learned/habitual cognition

Neuromorphic engineering

- Human brain
  - 10^{15} synOP/s; 15W

Task Energy Efficiency:

\[
\frac{\text{Energy}}{\text{Task}} = \frac{\text{Energy}}{\text{Operation}} \times \frac{\text{Operations}}{\text{Task}}
\]

- 1 fJ vs. 10 pJ
- 10^10 SynOps vs. 10^{11} MACs
- MNIST @ 95%

Synaptic Sampling Machine (SSM)
E. Neftci et al, 2016

Adiabatic CID-DRAM SVM (Kerneltron)
R. Karakiewicz et al, 2013

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
## Large-Scale Reconfigurable Neuromorphic Computing

### Technology and Performance Metrics

<table>
<thead>
<tr>
<th>Technology (nm)</th>
<th>Die Size (mm²)</th>
<th>Neuron Type</th>
<th># Neurons</th>
<th>Neuron Area (µm²)</th>
<th>Peak Throughput (Events/s)</th>
<th>Energy Efficiency (J/SynEvent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>14</td>
<td>28</td>
<td>180</td>
<td>180</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>60</td>
<td>430</td>
<td>50</td>
<td>168</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>5216</td>
<td>128k</td>
<td>1M</td>
<td>512</td>
<td>65k</td>
<td>65k</td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td>240 (240k)</td>
<td>14 (3325)</td>
<td>1500</td>
<td>1800</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>5M</td>
<td>3.4G</td>
<td>1G</td>
<td>65M</td>
<td>91M</td>
<td>73M</td>
<td></td>
</tr>
<tr>
<td>8n</td>
<td>24p</td>
<td>26p</td>
<td>N/A</td>
<td>31p</td>
<td>22p</td>
<td></td>
</tr>
</tbody>
</table>

1. Software-instantiated neuron model
2. Time-multiplexed neuron processor


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) Integrate-and-Fire Array Transceiver (IFAT) for scalable and reconfigurable neuromorphic neocortical processing. (a) Biophysical model of neural and synaptic dynamics. (b) Dynamically reconfigurable synaptic connectivity is implemented across IFAT arrays of addressable neurons by routing neural spike events locally through DRAM synaptic routing tables. (c) Each neural cell models conductance based membrane dynamics in proximal and distal compartments for synaptic input with programmable axonal delay, conductance, and reversal potential. (d) Multiscale global connectivity through a hierarchical network of HiAER routing nodes. (e) HiAER-IFAT board with 4 IFAT custom silicon microchips, serving 256k neurons and 256M synapses, and spanning 3 HiAER levels (L0-L2) in connectivity hierarchy. (f) The IFAT neural array multiplexes and integrates (top traces) incoming spike synaptic events to produce outgoing spike neural events (bottom traces). The latest IFAT microchip measured energy consumption is 22 pJ per spike event, several orders of magnitude more efficient than emulation on CPU/GPU platforms.

Yu et al, BioCAS 2012; Park et al, BioCAS 2014; Park et al, TNNLS 2017; Broccard et al, JNE 2017
Memristive Synapse Arrays for Neuromorphic Processing-in-Memory

- Scalable to high density and energy efficiency
  - < 100nm cell size in 12nm CMOS
  - < pJ energy per synapse operation
  - Vertically stacked integration in Intel-Micron 3D Xpoint/Optene SSD persistent memory

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.
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.
CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020

- Integration of CMOS neurons and resistive random-access memory (RRAM) memristive synapses with \textit{in-situ} revertible dataflow at record efficiency
CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020

- Gibbs stochastic sampling for Bayesian generative inference
  - Alternating between INFerence and GENeration in transpose datapaths
  - Restricted Boltzmann Machine (RBM) / Variational Autoencoder (VAE)
- Real-time image reconstruction from corrupted/noisy MNIST input

```
Sample from gray-level image

Reconstruct w/ RBM

Average

225 Visible Neurons (15x15 pixels)

60 Hidden Neurons

“good” pixels
Driven to ground-truth

“bad” pixels
Reconstruct by Gibbs sampling

Reconstruct

```
CMOS-RRAM Reconfigurable Neurosynaptic Array

*Wan et al, ISSCC 2020*

<table>
<thead>
<tr>
<th>New Image</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear</td>
<td></td>
</tr>
<tr>
<td>Recover</td>
<td></td>
</tr>
</tbody>
</table>

Click on the image to corrupt pixels!

<table>
<thead>
<tr>
<th># Gibbs Steps</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td># Samples</td>
<td>10</td>
</tr>
</tbody>
</table>

---

CMOS-RRAM Integration for AI on the Edge

- First fully integrated RRAM-based compute-in-memory chip to combine:
  - High versatility in configuring cores for diverse model architectures;
  - Superior (>2x) energy efficiency;
  - High inference accuracy comparable to software.

Applications to reconfigurable and heterogeneous AI on the edge.

CMOS-RRAM Integration for AI on the Edge

CMOS-RRAM Integration for AI on the Edge

Model-Driven Chip Calibration
- Initialize chip operating conditions
- Perform MVM using training-set data
- Output fully utilizing & not saturating ADC input swing?
  - Yes: Find offset of each row/column for compensation during testing
  - No: Adjust operating conditions

Noise-Resilient NN Training & Analog Weight Programming
- Inject noises w/ characterized distribution into weights during training
- Quantize (STA) 

Chip-in-the-loop Progressive Model Fine-Tuning
- Step n
  - Weights already programmed on chip
  - Program the weights of layer-n, and run inference
  - Use the measured outputs from layer-n to finetune the weights of the rest of layers (not on chip yet)
- Step n+1

CIFAR-10 Classification Using ResNet-20

Spiking Synaptic Sampling Machine (S\textsuperscript{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*
  - *Sparsity in neural activity and in synaptic connectivity*

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

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

---

**Kerneltron: Adiabatic Support Vector “Machine”**

Karakiewicz, Genov and Cauwenberghs

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

- **0.8 fJ/MAC in 0.25μm CMOS**
  - adiabatic resonant clocking conserves charge energy
  - energy efficiency on par with human brain \((10^{15} \text{ SynOP/S at 15W})\)

- **10 aJ/MAC in 22nm FDSOI CMOS**
  - 100x super-human

Classification results on MIT CBCL face detection data

Karakiewicz, Genov, and Cauwenberghs, VLSI’2006; CICC’2007; JSSC 2007; SJ 2013
Resonant Adiabatic Energy Recovery for Computing
Karakiewicz, Genov and Cauwenberghs

- **Kerneltron** support vector machine for visual pattern recognition with resonant hot clock adiabatic energy recovery in charge-domain processing-in-memory computing at 1 fJ of energy per multiply-accumulate

- Energy recovery logic (ERL) CMOS adiabatic line drivers recover 98% of the \( CV^2 \) electrostatic energy in the charge-mode array

Karakiewicz, Genov, and Cauwenberghs, VLSI’2006; CICC’2007; JSSC 2007; SJ 2013
Large-Scale Reconfigurable Neuromorphic Computing


Provides open access to large-scale reconfigurable neuromorphic computing hardware and software as an experimental testbed and development platform with up to 160M neurons and 40B synapses for the research community at large.

Neural-Synaptic Array Transceiver (Detorakis et al, Frontiers in Neuroscience, 2018)
Lifelong Learning at Scale (L2S, 2022 Telluride Neuromorphic Workshop)
Neuromodulatory Control (NMC, 2021 Telluride Neuromorphic Workshop)
Neuromorphic Systems Engineering

Integrated Systems Neuroengineering

Neural Systems

Learning & Adaptation

Environment

Silicon Microchips

Sensors and Actuators

Human/Bio Interaction

Neuromorphic/Neurosystems Engineering
# Acknowledgments

**Integrated Systems Neuroengineering Laboratory**

Department of Bioengineering, Institute for Neural Computation, UC San Diego

## Trainees and Alumni:

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abraham Akinin</td>
<td>Nanovision</td>
</tr>
<tr>
<td>Maruan Al-Shedivat</td>
<td>CMU</td>
</tr>
<tr>
<td>Fred Broccard</td>
<td>UCSD</td>
</tr>
<tr>
<td>Shantanu Chakrabarty</td>
<td>WUSTL</td>
</tr>
<tr>
<td>Steve Deiss</td>
<td></td>
</tr>
<tr>
<td>Yu Mike Chi</td>
<td>Cognionics</td>
</tr>
<tr>
<td>Roman Genov</td>
<td>U Toronto</td>
</tr>
<tr>
<td>Soumil Jain</td>
<td></td>
</tr>
<tr>
<td>Siddharth Joshi</td>
<td>U Notre Dame</td>
</tr>
<tr>
<td>Raf Karakiewicz</td>
<td>Intel</td>
</tr>
<tr>
<td>Rajkumar Kubendran</td>
<td></td>
</tr>
<tr>
<td>Christoph Maier</td>
<td></td>
</tr>
<tr>
<td>Hesham Mostafa</td>
<td>Intel AI</td>
</tr>
<tr>
<td>Rawan Naous</td>
<td>UC Berkeley</td>
</tr>
<tr>
<td>Emre Neftci</td>
<td>UC Irvine</td>
</tr>
<tr>
<td>Jongkil Park</td>
<td>ETRI</td>
</tr>
<tr>
<td>Akshay Paul</td>
<td></td>
</tr>
<tr>
<td>Bruno Pedroni</td>
<td></td>
</tr>
<tr>
<td>Ivan Rajen</td>
<td></td>
</tr>
<tr>
<td>Sadique Sheik</td>
<td>aiCorTeX</td>
</tr>
<tr>
<td>Pablo Tostado</td>
<td></td>
</tr>
<tr>
<td>Alice Yepremyan</td>
<td>JPL</td>
</tr>
<tr>
<td>Theodore Yu</td>
<td>Apple</td>
</tr>
<tr>
<td>Dejan Vucinic</td>
<td>Western Digital</td>
</tr>
<tr>
<td>Weier Wan</td>
<td>Stanford</td>
</tr>
<tr>
<td>Jun Wang</td>
<td>Harvard</td>
</tr>
<tr>
<td>Margot Wagner</td>
<td></td>
</tr>
<tr>
<td>Jiajia Wu</td>
<td></td>
</tr>
</tbody>
</table>

## Collaborators:

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Arthur</td>
<td>IBM</td>
</tr>
<tr>
<td>Henry Abarbanel</td>
<td>UCSD</td>
</tr>
<tr>
<td>Charles Augustine</td>
<td>Intel</td>
</tr>
<tr>
<td>Suman Datta</td>
<td>UND</td>
</tr>
<tr>
<td>Ken Kreutz-Delgado</td>
<td>UCSD</td>
</tr>
<tr>
<td>Duygu Kuzum</td>
<td>UCSD</td>
</tr>
<tr>
<td>Vijay Narayanan</td>
<td>PSU</td>
</tr>
<tr>
<td>Justin Mauger</td>
<td>NIWC Pacific</td>
</tr>
<tr>
<td>Paul Merolla</td>
<td>Neuralink</td>
</tr>
<tr>
<td>Dharmendra Modha</td>
<td>IBM</td>
</tr>
<tr>
<td>Somnath Paul</td>
<td>Intel</td>
</tr>
<tr>
<td>Yasufumi Sakai</td>
<td>Fujitsu</td>
</tr>
<tr>
<td>Khaled Salama</td>
<td>KAUST</td>
</tr>
<tr>
<td>Terrence Sejnowski</td>
<td>Salk</td>
</tr>
<tr>
<td>H.S. Philip Wong</td>
<td>Stanford</td>
</tr>
</tbody>
</table>

## Sponsors:

- Office of Naval Research (ONR)
- National Institutes of Health (NIH)
- Defense Advanced Research Projects Agency (DARPA)
- Intel
- Texas Instruments
- Western Digital
- Fujitsu
- Naval Information Warfare Center (formerly SPAWAR)