Energy Efficiency in Adaptive Neural Circuits

Gert Cauwenberghs

Department of Bioengineering
Institute for Neural Computation
UC San Diego

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

Go World Champion vs. Google DeepMind

~ 100 W  ~ 100 kW
Energy Efficiency in Adaptive Neural Circuits
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.

Scaling of Task and Machine Complexity

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.**
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></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><strong># Neurons</strong></td>
<td>5216¹</td>
<td>1M²</td>
<td>512</td>
<td>65k</td>
<td>65k</td>
</tr>
<tr>
<td><strong>Neuron Area (μm²)</strong></td>
<td>N/A¹</td>
<td>3325 (14)²</td>
<td>1500</td>
<td>1800</td>
<td>140</td>
</tr>
<tr>
<td><strong>Peak Throughput</strong></td>
<td>5M</td>
<td>1G</td>
<td>65M</td>
<td>91M</td>
<td>73M</td>
</tr>
<tr>
<td><strong>Energy Efficiency</strong></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)


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 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
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.

Kuzum, Jeyasingh, Lee, and Wong (ACS Nano, 2011)
Spiking Synaptic Sampling Machine (S$^3$M)

Biophysical Synaptic Stochasticity in Inference and Learning

Synaptic stochasticity as biophysical model of continuous DropConnect

Time-varying Bernoulli random masking of weights

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

Silicon Learning Machines for Embedded Sensor Adaptive Intelligence

Large-Margin Kernel Regression

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

Class Identification

MAP Forward Decoding

Sub-microwatt speaker verification and phoneme recognition (NIPS ‘2004)

Sequence Identification

Analog

ASP

Digital

A/D
Energy Efficiency in Adaptive Neural Circuits

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

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

Classification results on MIT CBCL face detection data
Resonant Charge Energy Recovery

CID array

Resonance

Capacitive load

Dynamic

Adiabatic

Number of Active Inputs

Energy (fJ/MAC)

Data Probability Density

(capacitive load)
Adaptive Low-Power Sensory Systems

Charge-domain Analog Signal Processing

Low-dimensional, Low-resolution Digital Coding

Digital Adaptation

2pJ/MAC 14b 8 × 8 Linear Transform Mixed Signal Spatial Filter in 65nm CMOS with 84dB Interference Suppression
S. Joshi et al, ISSCC 2017
Linear Transform Analog and Mixed-Signal Sensory Processing

- Application Enabler
- Lower Power
- Analog *processing gain* lowers A/D requirements

Processing gain: Improvement in SNR/DR due to ASP

Spatial Processing Gain

Improvement in SNR/DR due to ASP

Conventional

Analog spatial processing

14-bit

ADC Dynamic Range

Amplification

Digitization

Signal

Interferer

8-bit information

22-bits to resolve both signal and interference

8-bits to resolve signal only

Analog Signal Conditioning

Digitization
System Measurements

![Nested Thermometer Multiplying DACs](image)

- Power (mW)
- Frequency (Hz)
- Gain (dB)
- 3dB Bandwidth (MHz)
- Above threshold
- 2 pJ/MAC
- Leakage dominated

$W_{1,j}$
$W_{8,j}$
$A_j$
$y_j$

$DPU$

$1.8\text{mm}$

- 550 $\mu$W
- 335 $\mu$W
- 91 $\mu$W

Gain (dB) vs. Frequency (Hz)

Power (mW) vs. 3dB Bandwidth (MHz)
Measurements: Angular Resolution

Finite gain of OTA affects performance below 10°

Expected suppression @ 14b
90% confidence bound @ 14b
Measured Suppression

Signal Source

Signal To Interferer Ratio = -18dB

Proposed System

Experimental setup.
Measurements: SIR

Performance maintained at 0dBm interferer power.

input switch nonlinearity limits performance.

$P_{\text{sig}}$: fixed, $P_{\text{int}}$: fixed

$P_{\text{int}} = +6\text{dBm}$
Application: MIMO Communication

Spatial filtering to separate signal mixture

Constellation

64-QAM resolved
RMS EVM 2.9%

16-QAM resolved
RMS EVM 3.1%
## Application: MIMO Communication

### Beamforming Performance (baseband only)

<table>
<thead>
<tr>
<th></th>
<th>Tseng et. al. JSSC 2010</th>
<th>Ghaffari et. al. JSSC 2014</th>
<th>Kim et. al. JSSC 2015</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Received EVM (dB)</td>
<td>-25</td>
<td>-</td>
<td>-28.8</td>
<td>-30.8</td>
</tr>
<tr>
<td>Effective number of bits</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Angular Resolution (°)</td>
<td>22.5</td>
<td>22.5</td>
<td>&lt;5&lt;sup&gt;a&lt;/sup&gt;</td>
<td>&lt;1&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Interferer Cancellation (dB)</td>
<td>30&lt;sup&gt;b&lt;/sup&gt;</td>
<td>15&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>48&lt;sup&gt;b&lt;/sup&gt;</td>
<td>&gt;80&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>CMOS Technology (nm)</td>
<td>90</td>
<td>65</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>Power at Baseband (mW)</td>
<td>10&lt;sup&gt;d&lt;/sup&gt;</td>
<td>68-195&lt;sup&gt;e&lt;/sup&gt;</td>
<td>1.3</td>
<td>0.396</td>
</tr>
<tr>
<td>Bandwidth at Baseband (MHz)</td>
<td>20</td>
<td>5</td>
<td>3</td>
<td>2.4</td>
</tr>
</tbody>
</table>

<sup>a</sup>Greater than 15 dB cancellation, <sup>b</sup>Cancellation at 45° angular separation, <sup>c</sup>Out of beam, <sup>d</sup>LO power only, <sup>e</sup>Total power reported baseband power not reported

CLOSING THE LOOP: INTERACTIVE NEURAL/ARTIFICIAL INTELLIGENCE

- **Neuro Bio**
  - Neurosystems Engineering
  - Learning & Adaptation
  - Micropower Mixed-Signal VLSI

- **Neuromorphic Engineering**
  - Adaptive Sensory Feature Extraction and Pattern Recognition

- **Biosensors, Neural Prostheses and Brain Interfaces**
Integrated Systems Neuroengineering

Neuromorphic/Neurosystems Engineering

Neural Systems

Learning & Adaptation

Environment

Silicon Microchips

Human/Bio Interaction

Sensors and Actuators