Week 4: Silicon Cortex

Gert Cauwenberghs

Department of Bioengineering
UC San Diego

http://isn.ucsd.edu/courses/beng207
# BENG 207 Neuromorphic Integrated Bioelectronics

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/11, 10/13</td>
<td>Silicon cochlea. Low-noise acoustic sensing and automatic gain control. Continuous wavelet filter banks. Interaural time difference and level difference auditory localization. Blind source separation and independent component analysis.</td>
</tr>
<tr>
<td>10/28, 11/1</td>
<td>Review. Modular and scalable design for neuromorphic and bioelectronic integrated circuits and systems. Design for full testability and controllability.</td>
</tr>
<tr>
<td>11/8, 11/10</td>
<td>Learning and adaptation to compensate for external and internal variability over extended time scales. Background blind calibration of device mismatch. Correlated double sampling and chopping for offset drift and low-frequency noise cancellation.</td>
</tr>
<tr>
<td>11/22, 11/24</td>
<td>Guest lectures</td>
</tr>
<tr>
<td>11/29, 12/1</td>
<td>Project final presentations. All are welcome!</td>
</tr>
</tbody>
</table>
Integrated “Smart” Sensors

• Sensor networks:

![Diagram of sensor networks]

• Sensor Integration and Distributed Intelligence:

- reduced bandwidth requirements
- reduced power dissipation and form factor
- trade-off between precision and complexity/power of integrated processing
Pushing the Analog-Digital Boundary

• Digital Sensory Processing:
  - General-purpose
  - High precision (limited by A/D)

• Analog and Mixed-Signal Sensory Processing:
  - “Smart” A/D
  - Low power
  - Low complexity
Large-Scale Mixed-Signal Sensory Computation

• **Massive Parallelism**
  - distributed representation
  - local memory and adaptation
  - analog sensory interface
  - physical computation
  - analog accumulation on single wire

• **Scalable**
  silicon area and power scale linearly with throughput

• **Highly Efficient**
  factor 100 to 10,000 less energy/operation than DSP

• **Limited Precision**
  - analog mismatch and nonlinearity (WYDINWYG)
  - fix: adaptation in redundancy

Example: VLSI Analog-to-digital vector quantizer (Cauwenberghs and Pedroni, 1997)
Learning on Silicon

**Adaptation:**
- necessary for robust performance under variable conditions and in unpredictable environments
- also compensates for imprecision in analog computation
- avoids ad-hoc programming, tuning, and manual parameter adjustment

**Learning:**
- generalization of output to previously unknown, although similar, stimuli
- system identification to extract relevant environmental parameters

Event-Driven Sensory Analog Processing

- **Data driven**
  - Communication bandwidth adjusts to information bandwidth in the signal
- **Asynchronous**
  - No quantization (binning) of time
  - No power-hungry clocks and synchronization across network nodes
- **Highly energy efficient**
  - Significant energy savings over Nyquist sampling for signals of sparse activity and medium amplitude resolution
- **Robust to additive noise in the signal**
Multi-Modal Event-Driven Sensory Analog Processing

- Asynchronous routing of sensory address events
- Expandable dimensionality and integration of multiple sensory modalities
- Reconfigurable and adaptive general-purpose signal processing and identification

Mic, Ultrasonic transducer, Strain gauge, ...

Spatiotemporal sensory event stream encoder
- e.g. spatial gradient and temporal change amplitude threshold detections

Time-frequency sensory event stream encoder
- e.g. filterbank zero-crossings

Event stream transceiver and decoder
- Sensor fusion, feature binding, and adaptive pattern recognition
- e.g. detection of glass break acoustic events

AER time events

TDMA Asynchronous Bus

Classification events

Imager, Touch screen, ...

AER time events

- Mic, Ultrasonic transducer, Strain gauge, ...
- Spatiotemporal sensory event stream encoder:
  - e.g. spatial gradient and temporal change amplitude threshold detections
- Time-frequency sensory event stream encoder:
  - e.g. filterbank zero-crossings
**Time-frequency sensory event stream encoders:**
- Convert a continuous-time analog sensory input, such as an acoustic signal, into an output stream of spike time events.
- Time events correspond to time instances of zero-crossings of bandpass filtered versions of the signal.
- Each bandpass filter with different center frequency is coded as a frequency address in the zero-crossing time event stream for distributed time-frequency encoding.
Multi-Modal Event-Driven Sensory Analog Processing

- **Spatiotemporal sensory event stream encoders:**
  - Convert multidimensional input signals from a spatial array of sensors, such as an image sensor or a touch-sensitive display, into an output stream of spike time events.
  - Time events correspond to combinations of spatial gradient and temporal change amplitude threshold detections across pixels (sensor locations).
  - Each pixel has an identical set of various spatial gradient and temporal change event feature types, each coded as a unique address in the event stream for distributed spatiotemporal encoding.
**Multi-Modal Event-Driven Sensory Analog Processing**

- **Event stream transceivers and decoders for adaptive pattern recognition:**
  - Offer a digitally programmable transformation in the timing relationships between events, converting an input event stream into a transformed output event stream.
  - Transformations operate on spike times of input event streams through a network of integrate-and-fire neurons with digitally adjustable strength and delay in the synaptic interconnections between neurons.
  - Universal computing capabilities in a Turing formalism with continuous-time state transitions.
  - Greater tolerance to sensor noise in classification performance than features computed using conventional digital signal processing.
Multi-Modal Event-Driven Sensory Analog Processing

- Data driven
- Asynchronous
- Highly energy efficient
- Robust to additive noise in the signal
- Asynchronous routing of sensory address events
- Expandable integration of sensory modalities
- Reconfigurable and adaptive general-purpose signal processing and identification

Time-frequency sensory event stream encoder
Spatiotemporal sensory event stream encoder

Mic, Ultrasonic transducer, Strain gauge, ...
Imager, Touch screen, ...

e.g. filterbank zero-crossings
e.g. spatial gradient and temporal change amplitude threshold detections

TDMA Asynchronous Bus
Event stream transceiver and decoder
Classification events
e.g. detection of glass break acoustic events

Sensor fusion, feature binding, and adaptive pattern recognition

Continuous-Time Analog
Asynchronous Digital

Sensor
ASP
A/D
ADSP
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: \( R = n \ p \ q \)
  - \( n \): multiplicity  (repeat event)
  - \( p \): probability of release  (toss a coin)
  - \( q \): quantity released  (set amplitude)

\[ R = n \ p \ q \]

\begin{align*}
R &= n \ p \ q \\
n &\text{multiplicity} \\
p &\text{probability of release} \\
q &\text{quantity released}
\end{align*}
Silicon Membrane Array Transceiver
Vogelstein, Mallik and Cauwenberghs, 2004

- Voltage-controlled membrane conductance
  - *Event-driven activation*
  - *Dynamically reconfigurable:*
    - conductance $g$
    - driving potential $E$

- Address-event encoding of pre-and post-synaptic action potentials
Silicon Membrane Circuit
Goldberg, Cauwenberghs and Andreou, 2000
Vogelstein, Mallik and Cauwenberghs, 2004

$g_i(t)$ ion-specific membrane conductance

$E_i$ ion-specific reversal potential

Synapse subcircuit

Action potential generation and AER handshaking
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 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

EE
NanoE
Phys

Neuro
CS
CogSci
# Large-Scale Reconfigurable Neuromorphic Computing

## Technology and Performance Metrics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Die Size (mm²)</td>
<td>130</td>
<td>14</td>
<td>28</td>
<td>180</td>
<td>180</td>
<td>90</td>
</tr>
<tr>
<td>Neuron Type</td>
<td>Digital Arbitrary</td>
<td>Digital Conductance</td>
<td>Digital Accumulate &amp; Fire</td>
<td>Analog Conductance</td>
<td>Analog Shared-Dendrite</td>
<td>Analog 2-Compartment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integrate &amp; Fire</td>
<td></td>
<td>Integrate &amp; Fire</td>
<td>Conductance I&amp;F</td>
<td>Conductance I&amp;F</td>
</tr>
<tr>
<td># Neurons</td>
<td>5216 ¹</td>
<td>128k ²</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>240 (240k)²</td>
<td>14 (3325)²</td>
<td>1500</td>
<td>1800</td>
<td>140</td>
</tr>
<tr>
<td>Peak Throughput (Events/s)</td>
<td>5M</td>
<td>3.4G</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>24p</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 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)

(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
Hierarchical Address-Event Routing (HiAER)

Example network with 16 neurons and weighted synaptic connections.

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

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, 2017
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\textsubscript{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\textsubscript{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\textsubscript{n} – 1 bus, and of the L\textsubscript{1} bus with a L\textsubscript{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. (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
IFAT Thermodynamics of Neural Excitability

Yu, Park, Joshi, Maier, Cauwenberghs, BioCAS 2012

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

Memristive Synapse Arrays for Neuromorphic Processing-in-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.

- 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
Memristive Synapse Arrays for Neuromorphic Processing-in-Memory

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 *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)

“good” pixels
Driven to ground-truth
Inf

“bad” pixels
Reconstruct by Gibbs sampling
Gen

60 Hidden Neurons

Reconstruction Error after 2 Gibbs Steps

CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020

Click on the image to corrupt pixels!

New Image
Clear
Recover

# Gibbs Steps: 6
# Samples: 10

Hierarchical Address-Event Routing
Neural and Synaptic Array Transceiver
HiAER-NSAT

(a) Neuron and Synapse Model

\[
\begin{align*}
I_{syn, i} + \alpha_S, & \quad V, \alpha, \beta, b, \eta, \text{start, (stop)} \\
\xi^p & \quad V_{th}, V_{reset}, \tau_{ref}
\end{align*}
\]

\[
V_i(t + 1) = \alpha V_i(t) + \beta(b_i + I_{syn, i}(t) + \sigma_i \eta_i(t)),
\]

\[
I_{syn, i}(t + 1) = \alpha S I_{syn, i}(t) + \sum_j \xi^p_{ij} s_j(t - \theta_{ij}),
\]

\[
\Delta w_{ij}(t) = g(t) (s_j(t) c_i^{post}(t) + s_i(t) c_j^{pre}(t))
\]

(b) Connectivity Model

(c) HiAER Tree

(d) HiAER Node

(e) NSAT Core

HBM FPGA Reconfigurable Neuromorphic Computing

- Reconfigurable, high-throughput neuromorphic processing-in-memory (PIM)
  - Xilinx UltraScale+ VU37P field-programmable gate array (FPGA)
- High-Bandwidth Memory (HBM) for extreme PIM throughput
  - Integrated 8GB HBM2.0 DDR4 SDRAM
  - Sustained > 400 GB/s random-access memory bandwidth at 25 ns latency, delivering > 100 GSynOp/s throughput at 32b/Syn weight resolution
  - 32 independent HBM ports aligned with 32 neurosynaptic cores on the FPGA
- Demonstrated record low-latency, high-throughput MNIST image classification
  - 10,000-image MNIST dataset classified, at 94% accuracy, in 720 ms, or 72 µs/image
  - Single FPGA core implementing 784 x 500 x 10 DNN with binary threshold units

Weight Quantization for Memory-Efficient Inference

- **Dropout** during training improves not only generalization performance, but also resilience to round-off in weight quantization for memory-efficient inference.

- Stochastic rounding of binary mantissa $M$ and radix-2 exponent $E$ retains nearly full performance with just a total of $N = 6$ bits per signed weight $w$:
  
  - **Resolution given by mantissa $M$**
  - **Dynamic range given by exponent $E$**: $2^{E}$

\[
 w = s \, 1.x_1...x_M \, 2^{e_1...e_E}
\]

\[
 N = 1 + M + E
\]

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 128M neurons and 32B synapses for the research community at large.
# BENG 207 Neuromorphic Integrated Bioelectronics

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/11, 10/13</td>
<td>Silicon cochlea. Low-noise acoustic sensing and automatic gain control. Continuous wavelet filter banks. Interaural time difference and level difference auditory localization. Blind source separation and independent component analysis.</td>
</tr>
<tr>
<td>10/28, 11/1</td>
<td>Review. Modular and scalable design for neuromorphic and bioelectronic integrated circuits and systems. Design for full testability and controllability.</td>
</tr>
<tr>
<td>11/8, 11/10</td>
<td>Learning and adaptation to compensate for external and internal variability over extended time scales. Background blind calibration of device mismatch. Correlated double sampling and chopping for offset drift and low-frequency noise cancellation.</td>
</tr>
<tr>
<td>11/22, 11/24</td>
<td>Guest lectures</td>
</tr>
<tr>
<td>11/29, 12/1</td>
<td>Project final presentations. All are welcome!</td>
</tr>
</tbody>
</table>