BENG 207 Special Topics in Bioengineering

Neuromorphic Integrated Bioelectronics

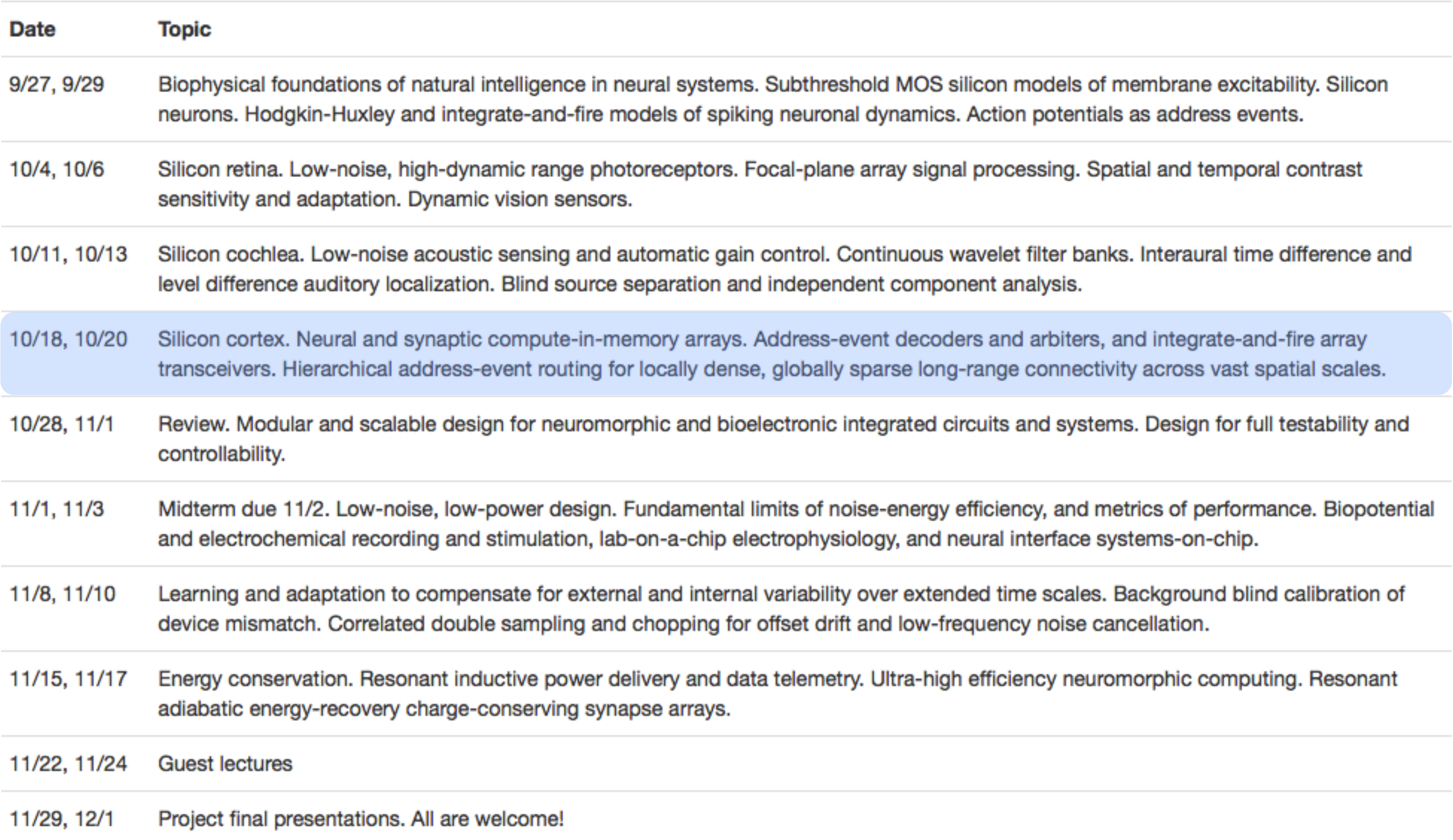
Week 4: Silicon Cortex

Gert Cauwenberghs

Department of Bioengineering
UC San Diego

http://isn.ucsd.edu/courses/beng207

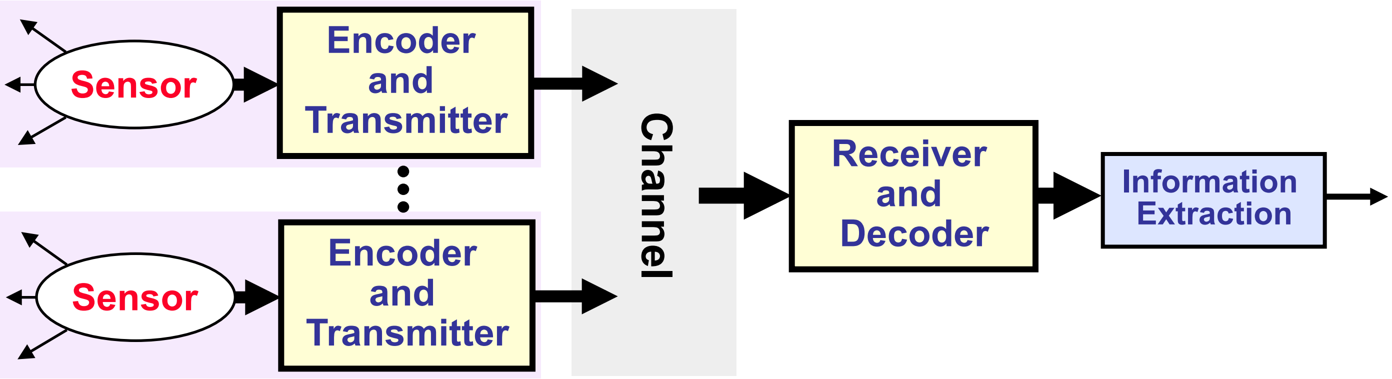
BENG 207 Neuromorphic Integrated Bioelectronics



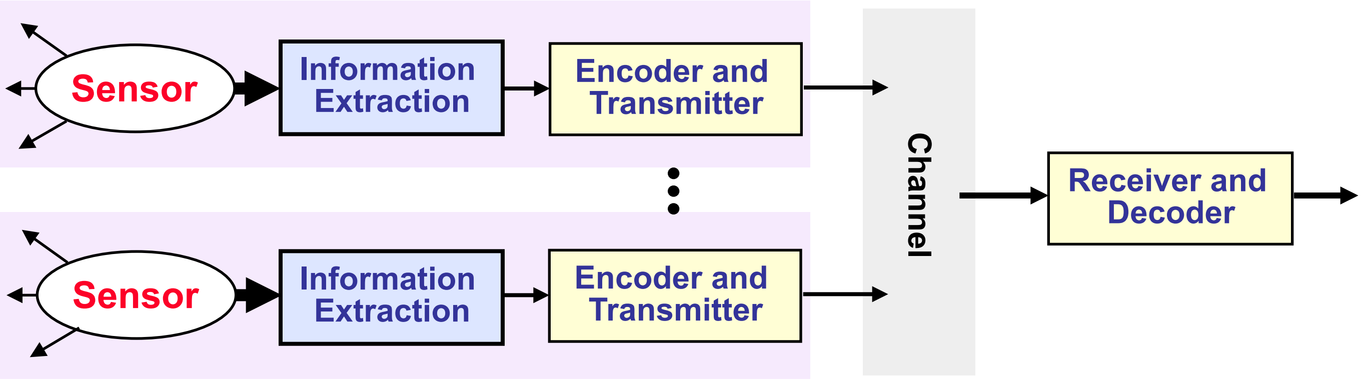
Date	Topic
9/27, 9/29	Biophysical foundations of natural intelligence in neural systems. Subthreshold MOS silicon models of membrane excitability. Silicon neurons. Hodgkin-Huxley and integrate-and-fire models of spiking neuronal dynamics. Action potentials as address events.
10/4, 10/6	Silicon retina. Low-noise, high-dynamic range photoreceptors. Focal-plane array signal processing. Spatial and temporal contrast sensitivity and adaptation. Dynamic vision sensors.
10/11, 10/13	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.
10/18, 10/20	Silicon cortex. Neural and synaptic compute-in-memory arrays. Address-event decoders and arbiters, and integrate-and-fire array transceivers. Hierarchical address-event routing for locally dense, globally sparse long-range connectivity across vast spatial scales.
10/28, 11/1	Review. Modular and scalable design for neuromorphic and bioelectronic integrated circuits and systems. Design for full testability and controllability.
11/1, 11/3	Midterm due 11/2. Low-noise, low-power design. Fundamental limits of noise-energy efficiency, and metrics of performance. Biopotential and electrochemical recording and stimulation, lab-on-a-chip electrophysiology, and neural interface systems-on-chip.
11/8, 11/10	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.
11/15, 11/17	Energy conservation. Resonant inductive power delivery and data telemetry. Ultra-high efficiency neuromorphic computing. Resonant adiabatic energy-recovery charge-conserving synapse arrays.
11/22, 11/24	Guest lectures
11/29, 12/1	Project final presentations. All are welcome!

Integrated “Smart” Sensors

- Sensor networks:



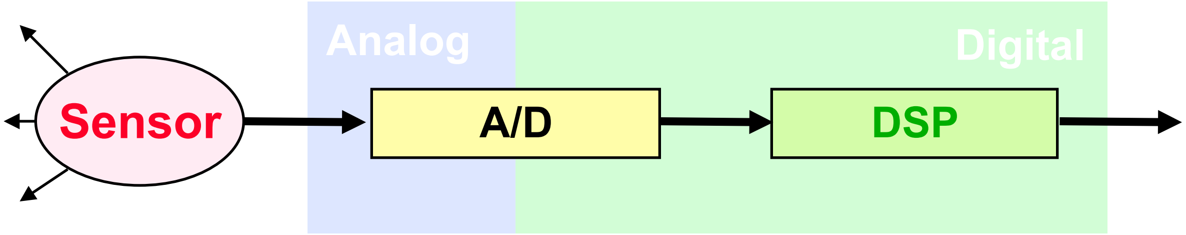
- Sensor Integration and Distributed Intelligence:



  - reduced bandwidth requirements
  - reduced power dissipation and form factor
  - trade-off between precision and complexity/power of integrated processing

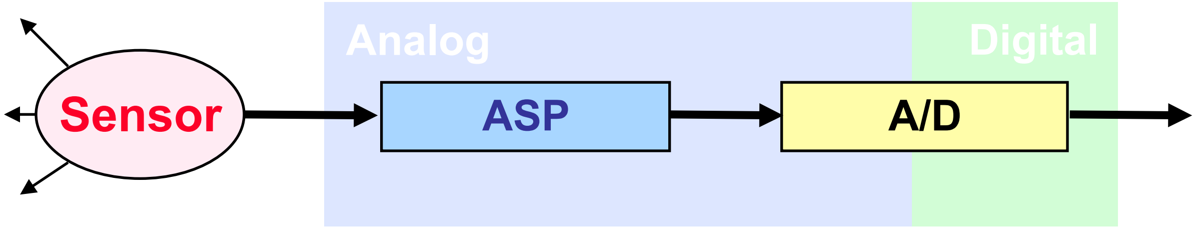
Pushing the Analog-Digital Boundary

- **Digital Sensory Processing:**



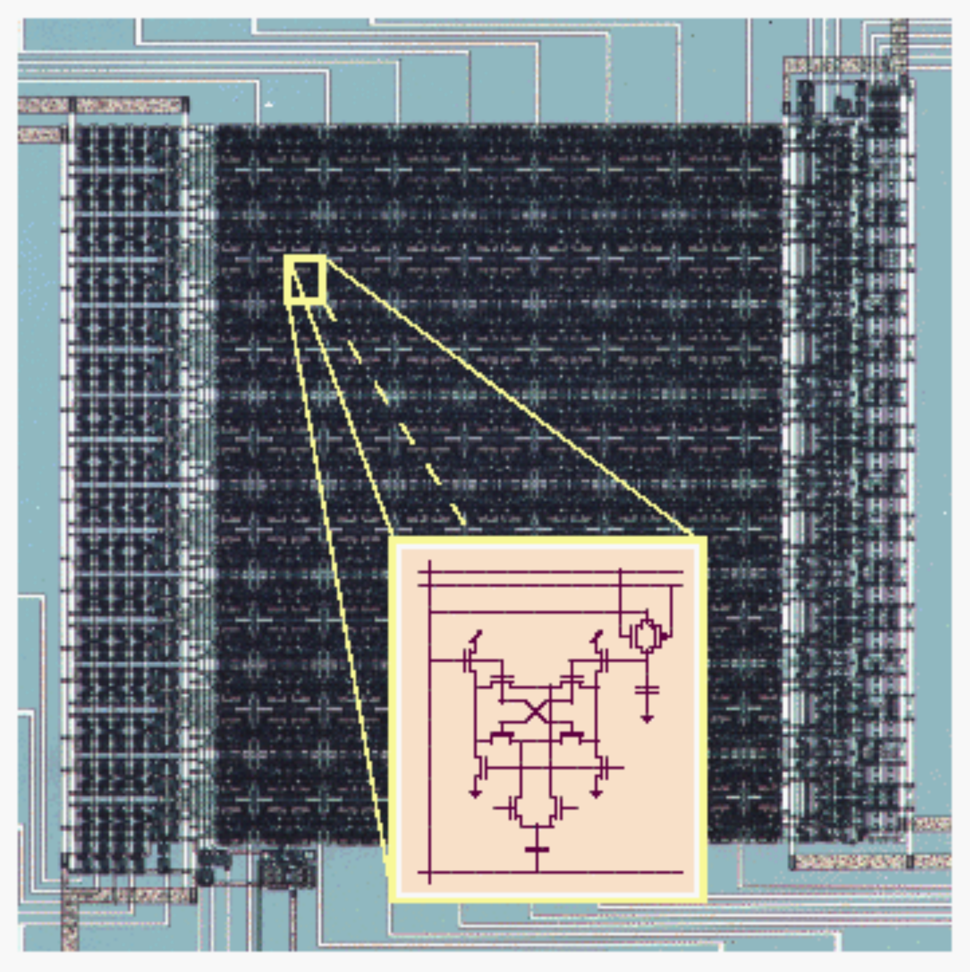
  - General-purpose
  - High precision (limited by A/D)

- **Analog and Mixed-Signal Sensory Processing:**



  - "Smart" A/D
  - Low power
  - Low complexity

Large-Scale Mixed-Signal Sensory Computation



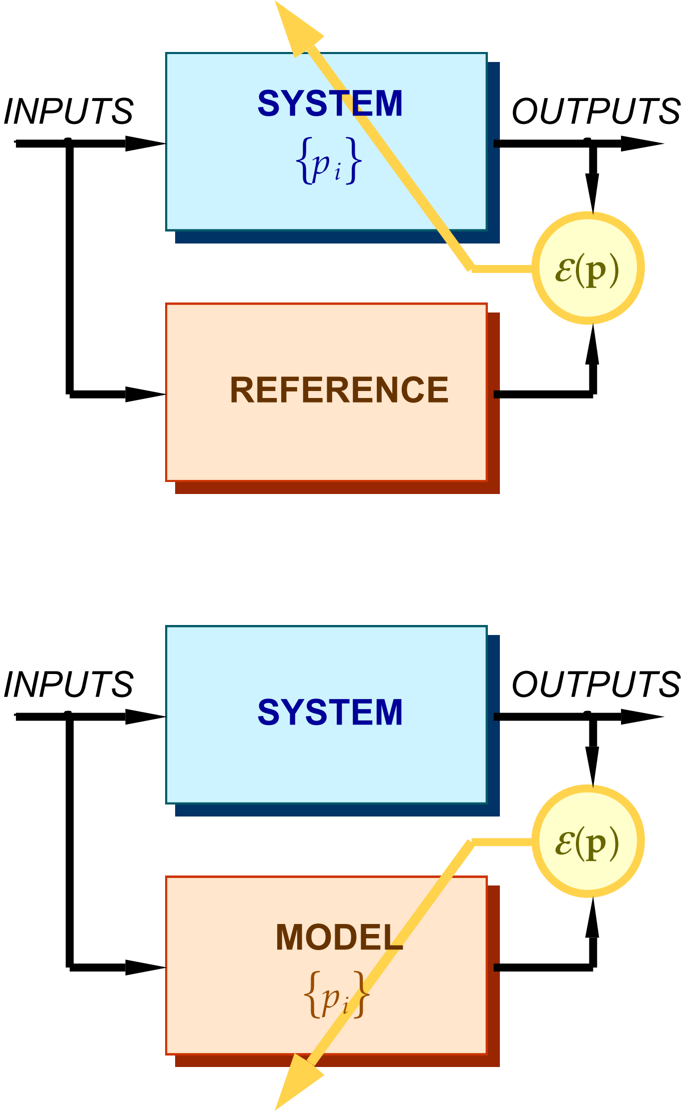
Example: VLSI Analog-to-digital vector quantizer (Cauwenberghs and Pedroni, 1997)

- **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 nonlineary (WYDINWYG)
  - fix: adaptation in redundancy

Learning on Silicon



Adaptation:

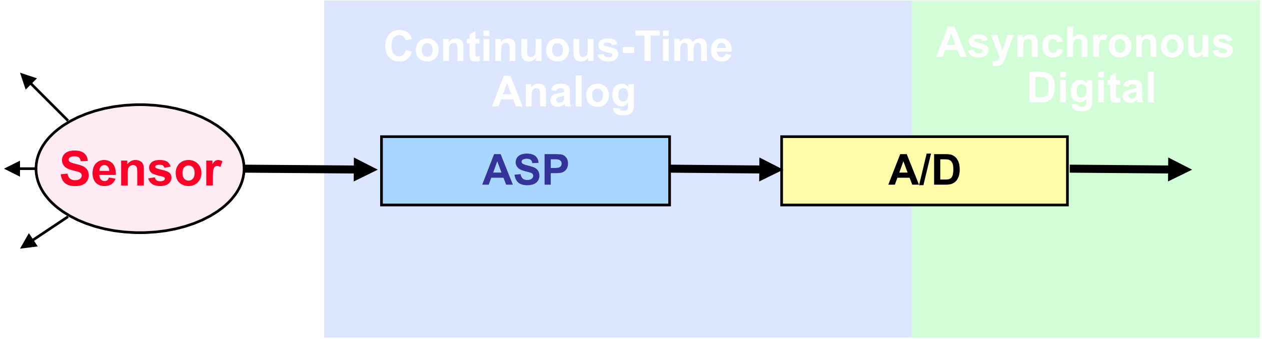
- – necessary for robust performance under variable conditions and in unpredictable environments
- – also compensates for imprecision in analog computation
- – avoids ad-hoc programming, tuning, and manual parameter adjustment

Learning:

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

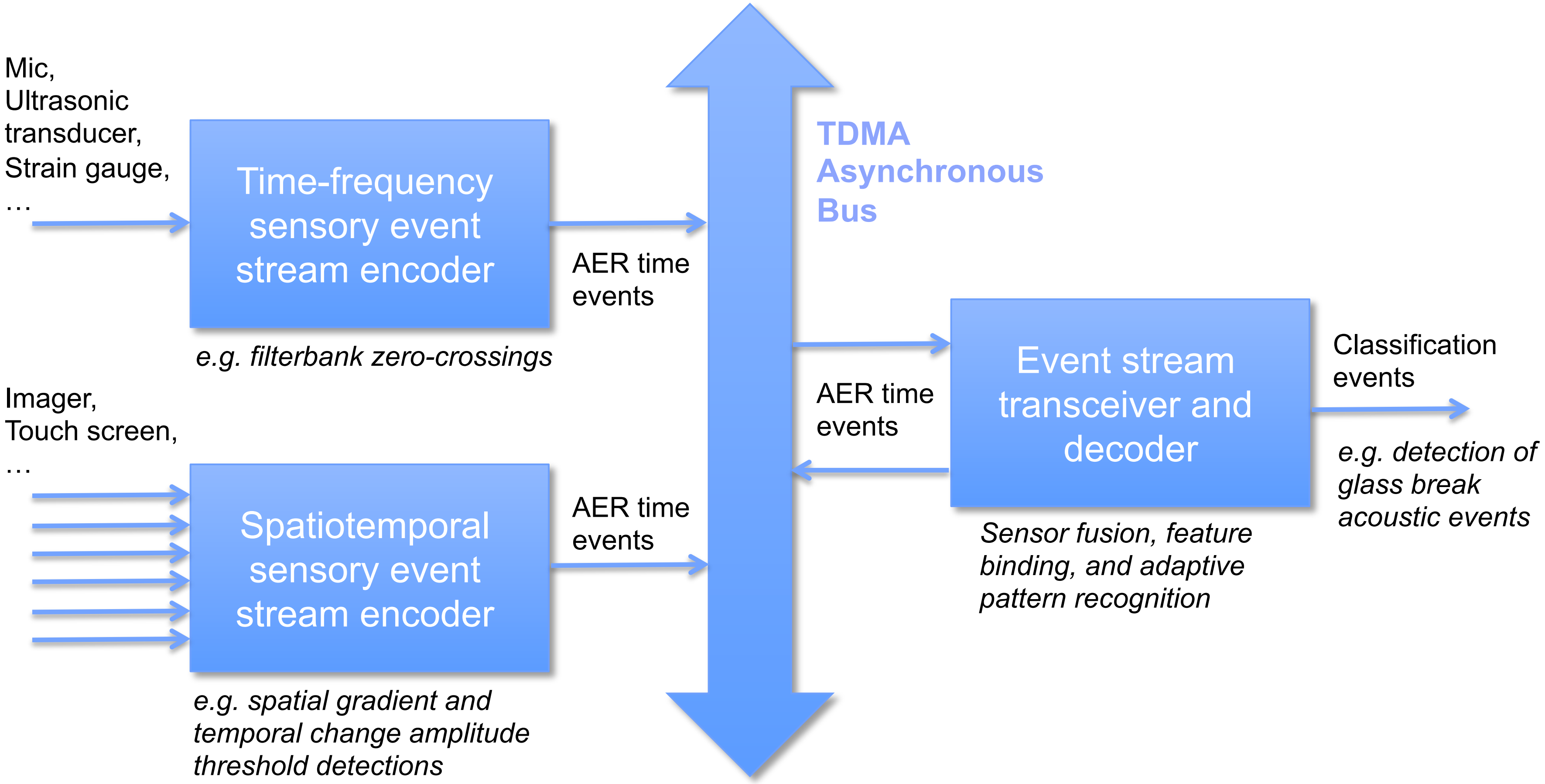
Cauwenberghs & Bayoumi, Eds., *Learning on Silicon*, Kluwer 1999.

Event-Driven Sensory Analog Processing



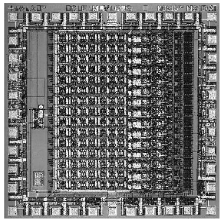
- **Data driven**
  - Communication bandwidth adjusts to information bandwidth in the signal
- **Asynchronous**
  - No quantization (binning) of time
  - No power-hungry clocks and synchronization across network nodes
- **Highly energy efficient**
  - Significant energy savings over Nyquist sampling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

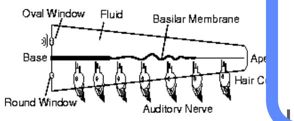
Multi-Modal Event-Driven Sensory Analog Processing

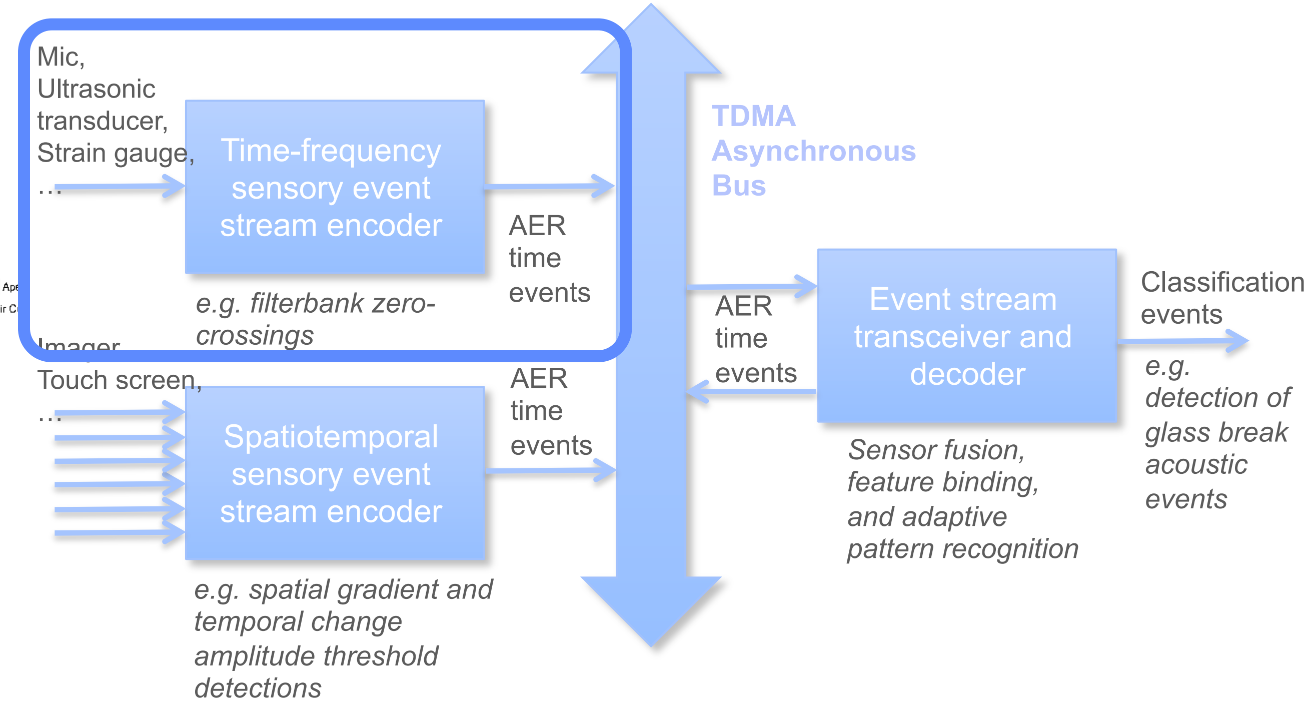

Mic,
Ultrasonic
transducer,
Strain gauge,
…
Time-frequency sensory event stream encoder
AER time events
e.g. filterbank zero-crossings
Imager,
Touch screen,
…
Spatiotemporal sensory event stream encoder
AER time events
e.g. spatial gradient and temporal change amplitude threshold detections
TDMA Asynchronous Bus
AER time events
Event stream transceiver and decoder
Sensor fusion, feature binding, and adaptive pattern recognition
Classification events
e.g. detection of glass break acoustic events

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

Multi-Modal Event-Driven Sensory Analog Processing







Mic, Ultrasonic transducer, Strain gauge, …

Time-frequency sensory event stream encoder

e.g. filterbank zero-crossings

AER time events

Imager, Touch screen, …

Spatiotemporal sensory event stream encoder

e.g. spatial gradient and temporal change amplitude threshold detections

AER time events

TDMA Asynchronous Bus

AER time events

Event stream transceiver and decoder

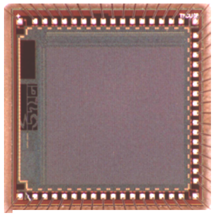
Sensor fusion, feature binding, and adaptive pattern recognition

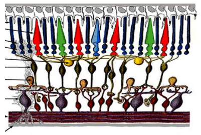
Classification events

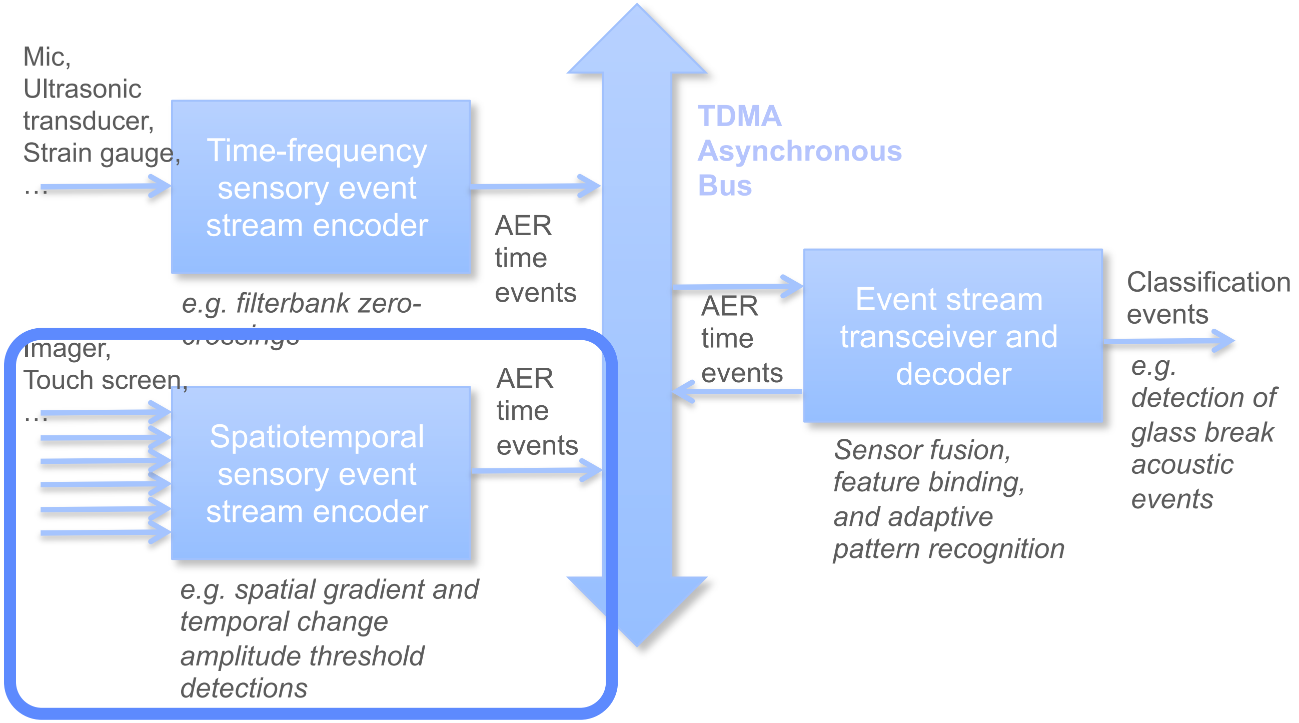
e.g. detection of glass break acoustic events

- **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

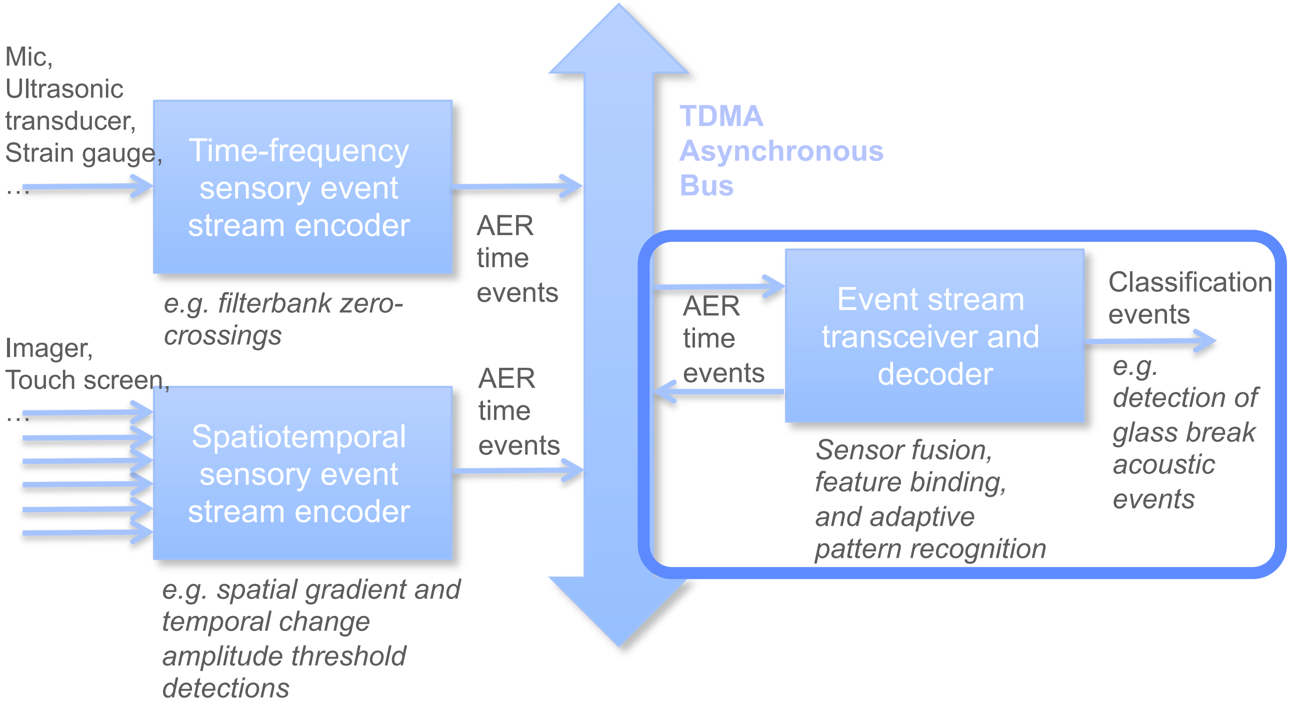





Mic,
Ultrasonic
transducer,
Strain gauge,
…
Time-frequency sensory event stream encoder
e.g. filterbank zero-crossings
AER time events
TDMA Asynchronous Bus
Imager,
Touch screen,
…
Spatiotemporal sensory event stream encoder
AER time events
e.g. spatial gradient and temporal change amplitude threshold detections
AER time events
Event stream transceiver and decoder
Sensor fusion, feature binding, and adaptive pattern recognition
Classification events
e.g. detection of glass break acoustic events

- **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



Mic,
Ultrasonic transducer,
Strain gauge,
…

Time-frequency sensory event stream encoder

e.g. filterbank zero-crossings

AER time events

Imager,
Touch screen,
…

Spatiotemporal sensory event stream encoder

e.g. spatial gradient and temporal change amplitude threshold detections

AER time events

TDMA Asynchronous Bus

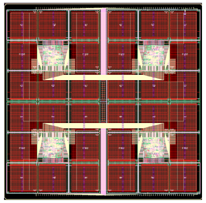
AER time events

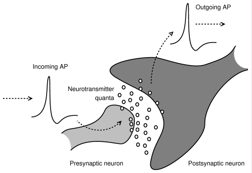
Event stream transceiver and decoder

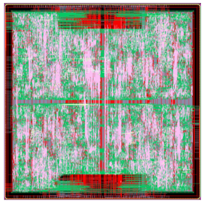
Sensor fusion, feature binding, and adaptive pattern recognition

Classification events

e.g. detection of glass break acoustic events

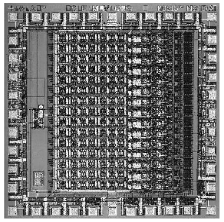


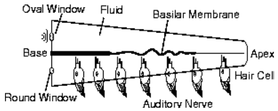


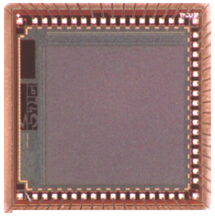


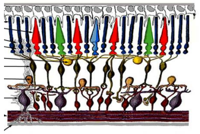
- **Event stream transceivers and decoders for adaptive pattern recognition:**
  - Offer a digitally programmable transformation in the timing relationships between events, converting an input event stream into a transformed output event stream.
  - Transformations operate on spike times of input event 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.

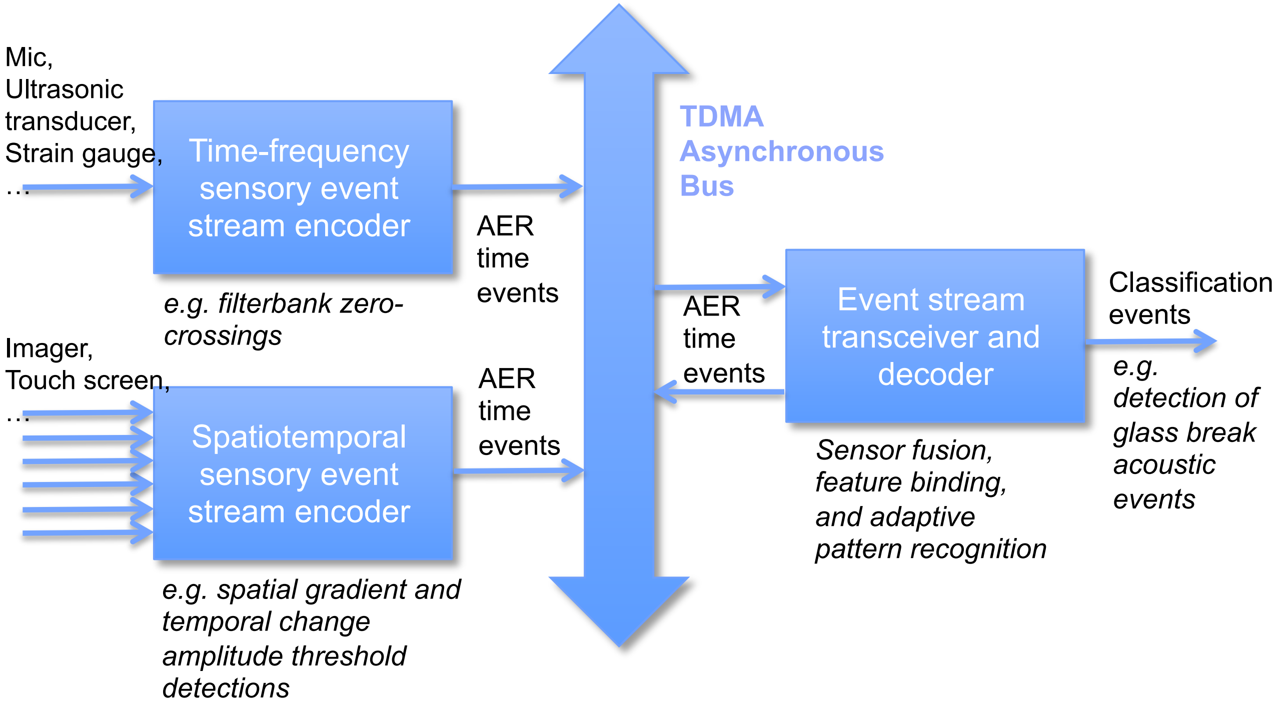
Multi-Modal Event-Driven Sensory Analog Processing











Mic, Ultrasonic transducer, Strain gauge, …

Time-frequency sensory event stream encoder

e.g. filterbank zero-crossings

AER time events

Imager, Touch screen, …

Spatiotemporal sensory event stream encoder

e.g. spatial gradient and temporal change amplitude threshold detections

AER time events

TDMA Asynchronous Bus

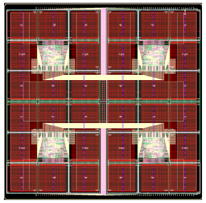
AER time events

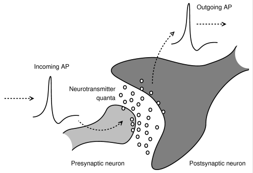
Event stream transceiver and decoder

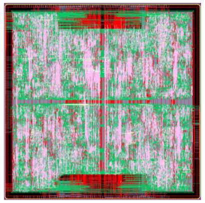
Sensor fusion, feature binding, and adaptive pattern recognition

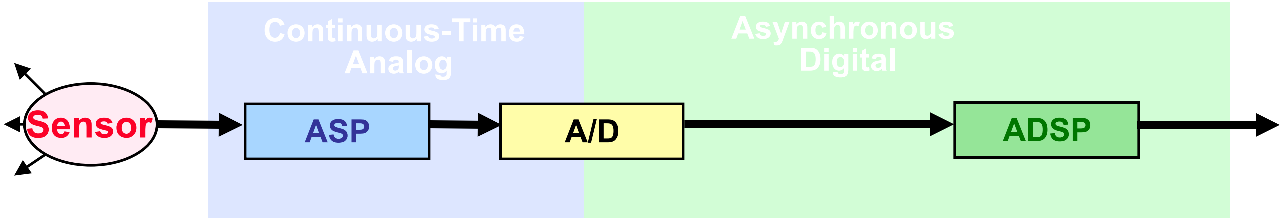
Classification events

e.g. detection of glass break acoustic events







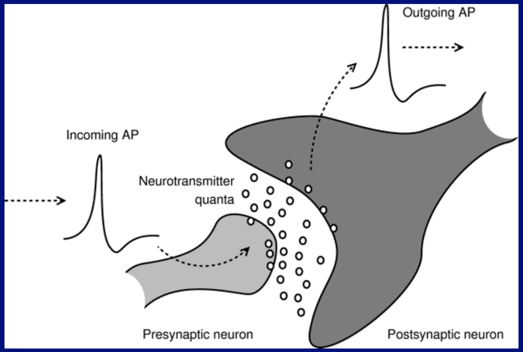


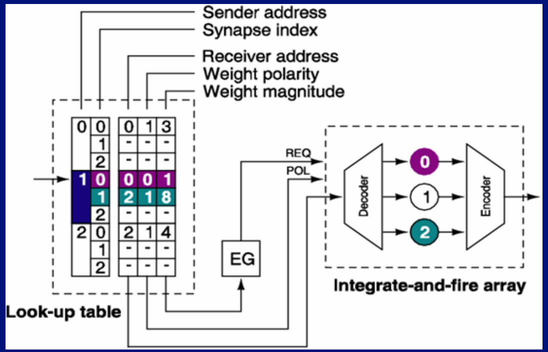
- Data driven
- Asynchronous
- Highly energy efficient
- Robust to additive noise in the signal

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

Reconfigurable Synaptic Connectivity and Plasticity

From Microchips to Large-Scale Neural Systems



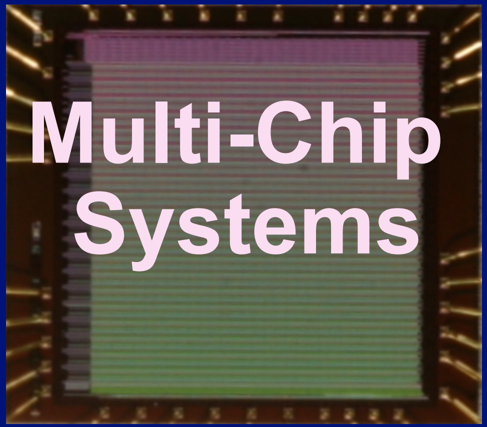


Sender address	Synapse index	Receiver address	Weight polarity	Weight magnitude
0	0	0	1	3
	1	-	-	-
	2	-	-	-
1	0	0	0	1
	1	2	1	8
	2	-	-	-
2	0	2	1	4
	1	-	-	-
	2	-	-	-

Address-Event Representation

Neural Systems

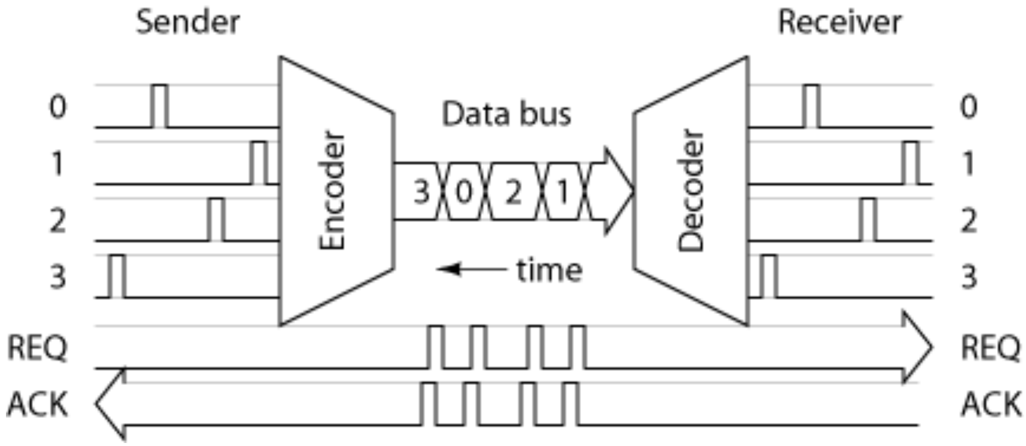
Synaptic Plasticity & Wiring



Multi-Chip Systems

Address-Event Representation (AER)

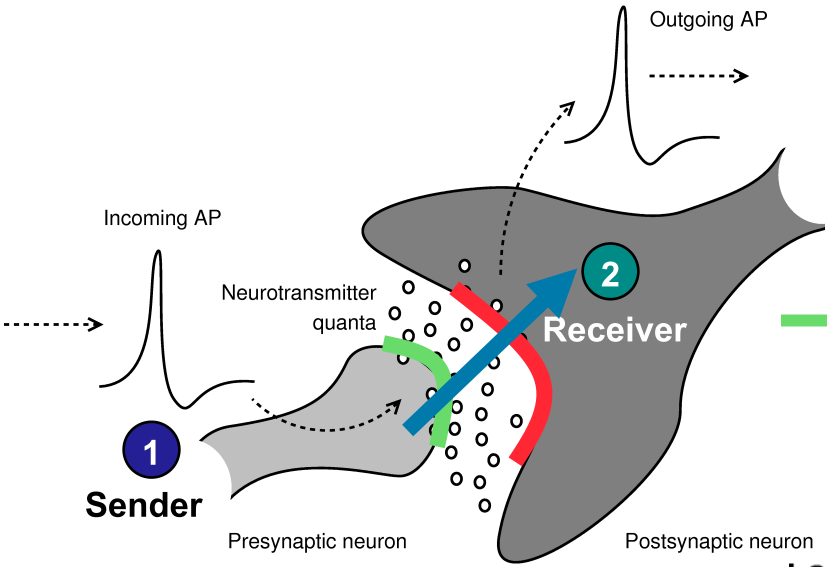
Lazzaro et al., 1993; Mahowald, 1994; Deiss 1994; Boahen 2000

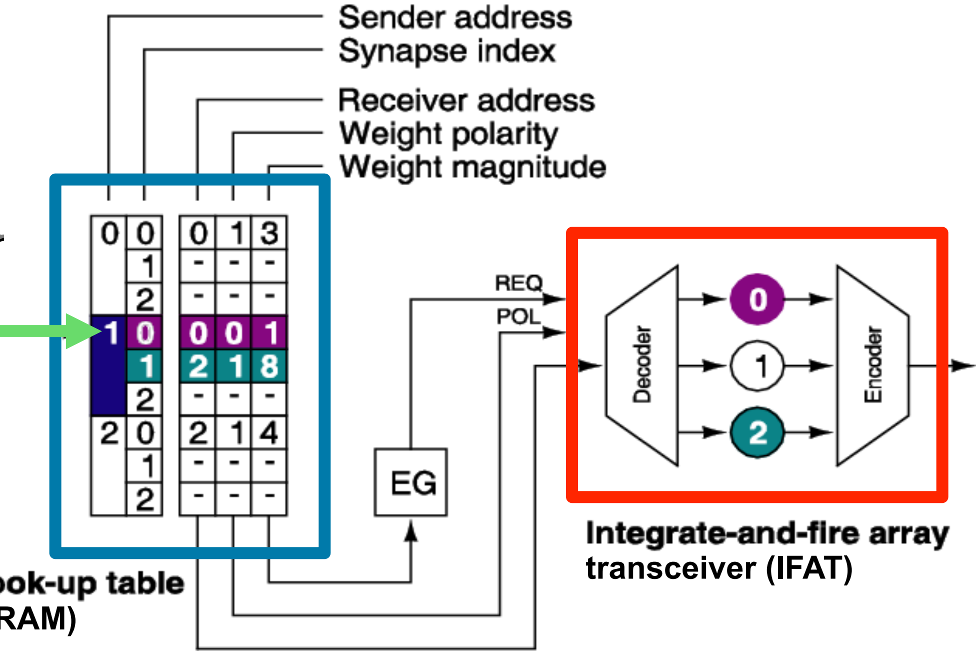


- AER emulates extensive connectivity between neurons by communicating spiking events time-multiplexed on a shared data bus.
- Spikes are represented by two values:
  - *Cell location (address)*
  - *Event time (implicit)*
- All events within Δt are "simultaneous"

Address-Event Synaptic Connectivity

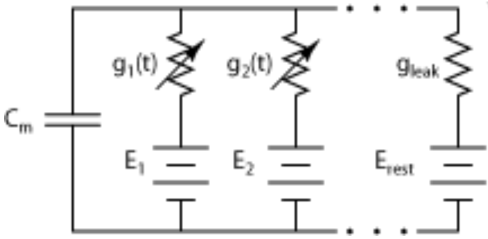
Goldberg, Cauwenberghs and Andreou, 2000

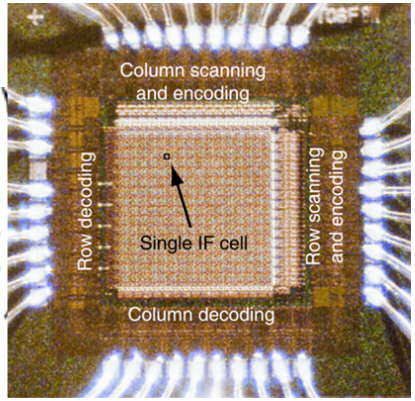




Sender address	Synapse index	Receiver address	Weight polarity	Weight magnitude
0	0	0	1	3
	1	-	-	-
	2	-	-	-
1	0	0	0	1
	1	2	1	8
	2	-	-	-
2	0	2	1	4
	1	-	-	-
	2	-	-	-

- ‘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)*



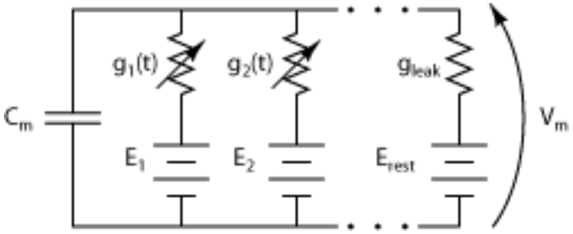


IFAT2 (2000)

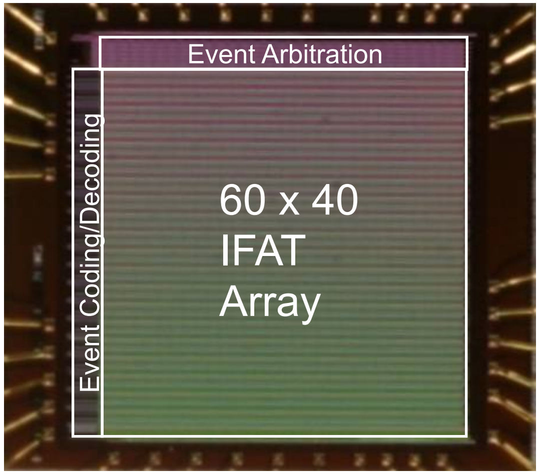
Silicon Membrane Array Transceiver

Vogelstein, Mallik and Cauwenberghs, 2004

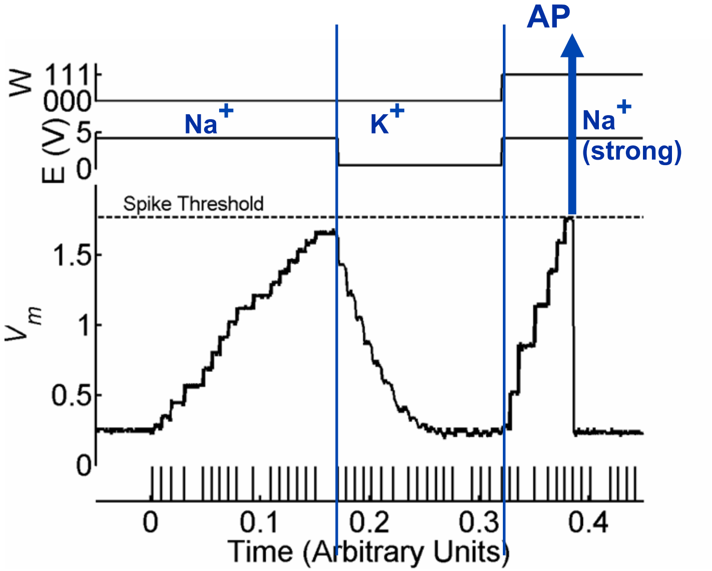
- Voltage-controlled membrane conductance
  - *Event-driven activation*
  - *Dynamically reconfigurable:*
    - *conductance g*
    - *driving potential E*



- Address-event encoding of pre-and post-synaptic action potentials

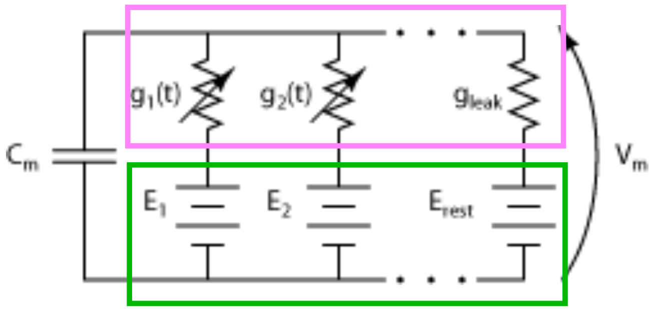


IFAT3 (2004)



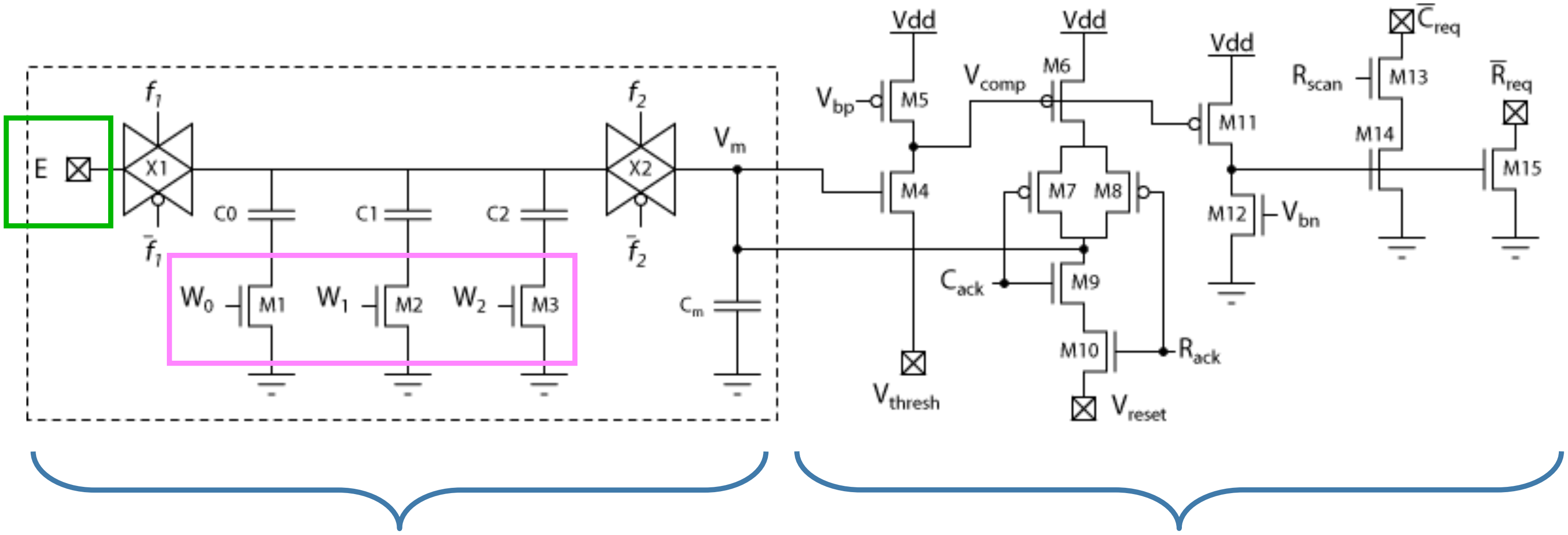
Silicon Membrane Circuit

Goldberg, Cauwenberghs and Andreou, 2000
Vogelstein, Mallik and Cauwenberghs, 2004



$g_i(t)$ ion-specific membrane conductance

E_i ion-specific reversal potential

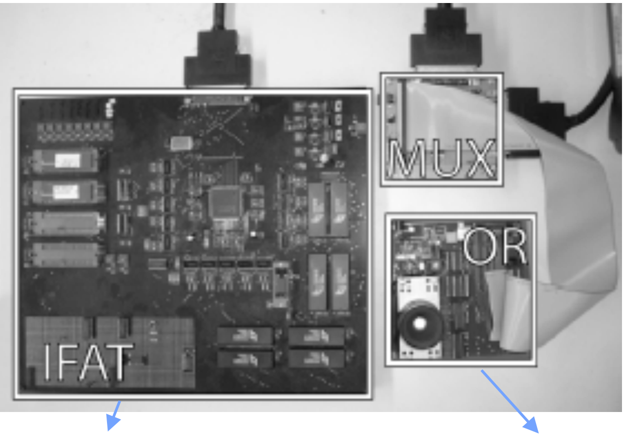


Synapse subcircuit

Action potential generation and AER handshaking

Hierarchical Vision and Saliency-Based Acuity Modulation

Vogelstein, Mallik, Culurciello, Cauwenberghs, and Etienne-Cummings, NECO 2007

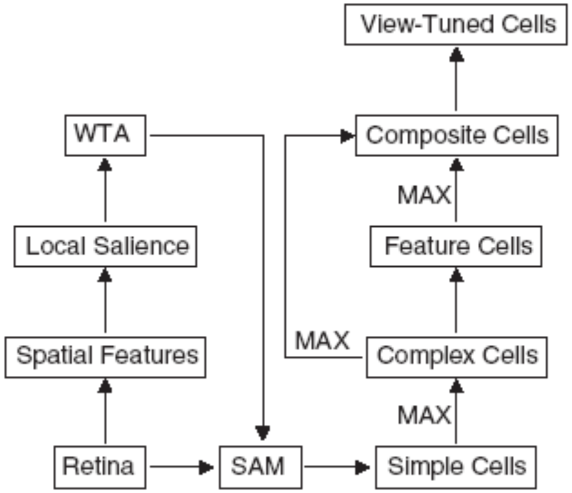


***IFAT* Cortical Model**

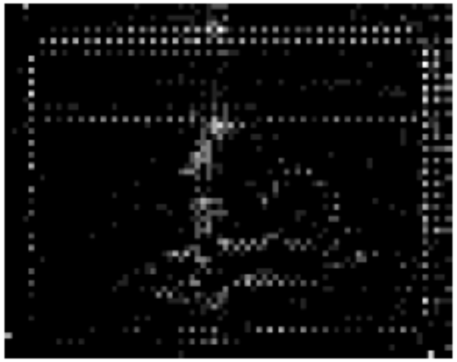
4800 silicon neurons
4,194,304 synapses

***Octopus* Silicon Retina**

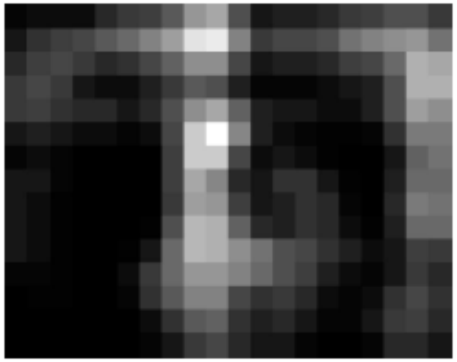
80 x 60 pixels
AER spiking output



OR image

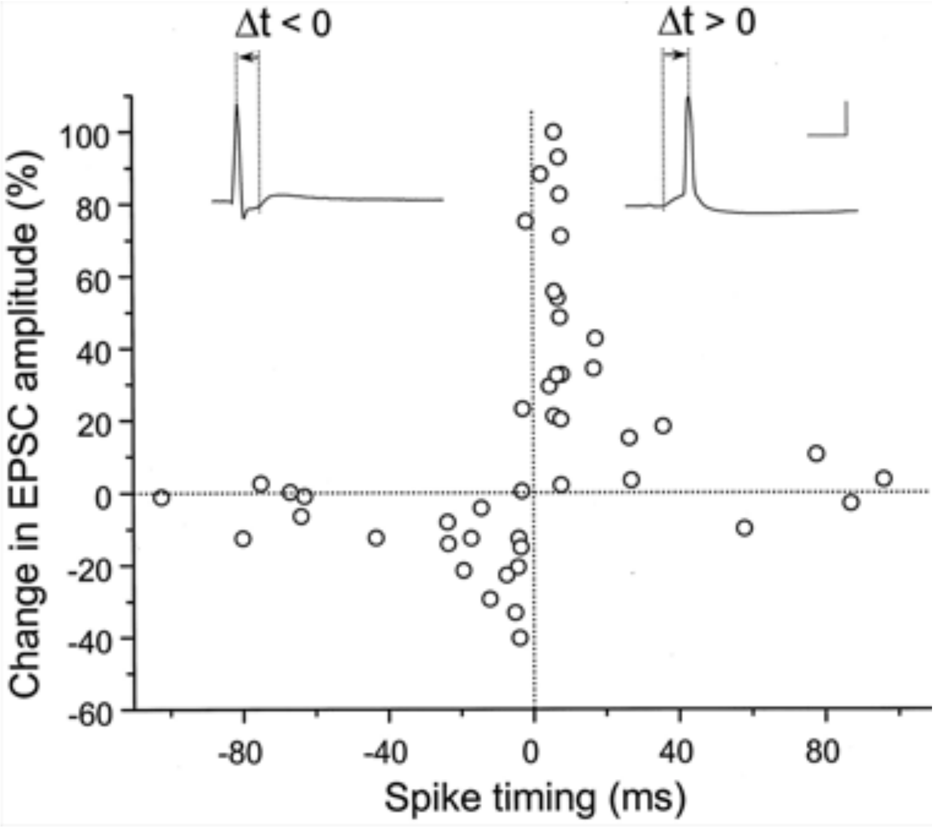


Simple cell response



Saliency map

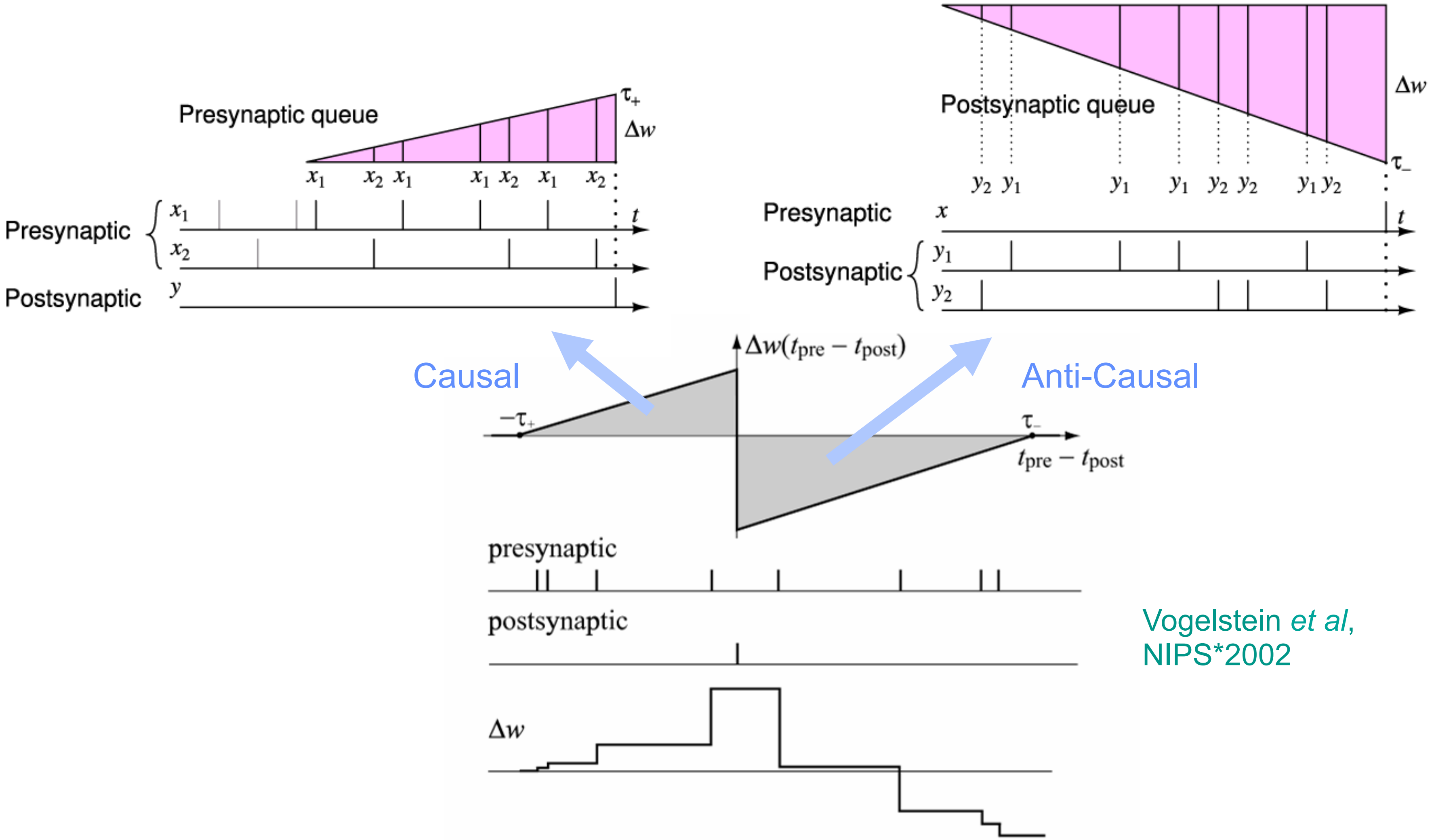
Spike Timing-Dependent Plasticity



Bi and Poo, 1998

Spike Timing-Dependent Plasticity

in the Address Domain



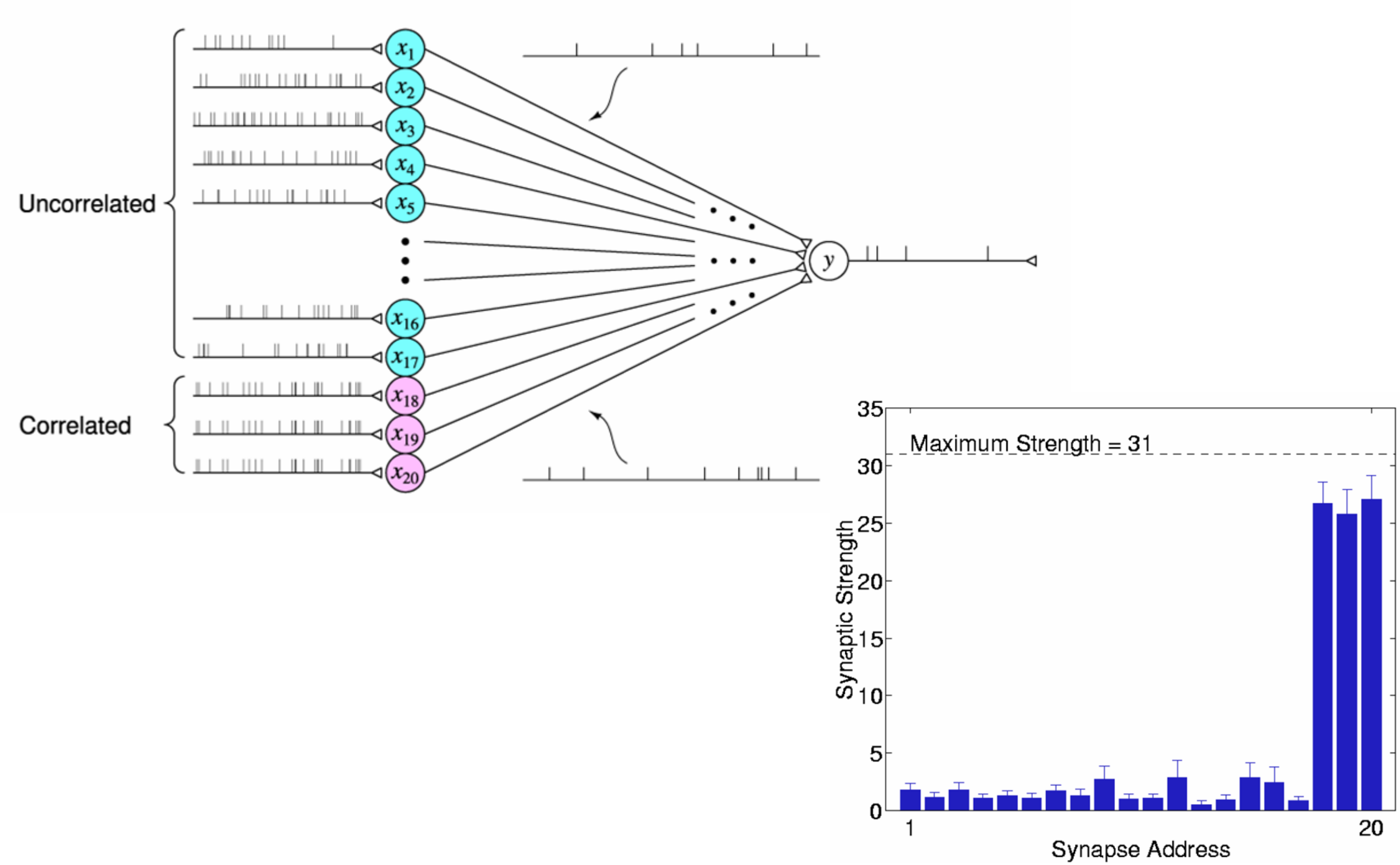
Causal

Anti-Causal

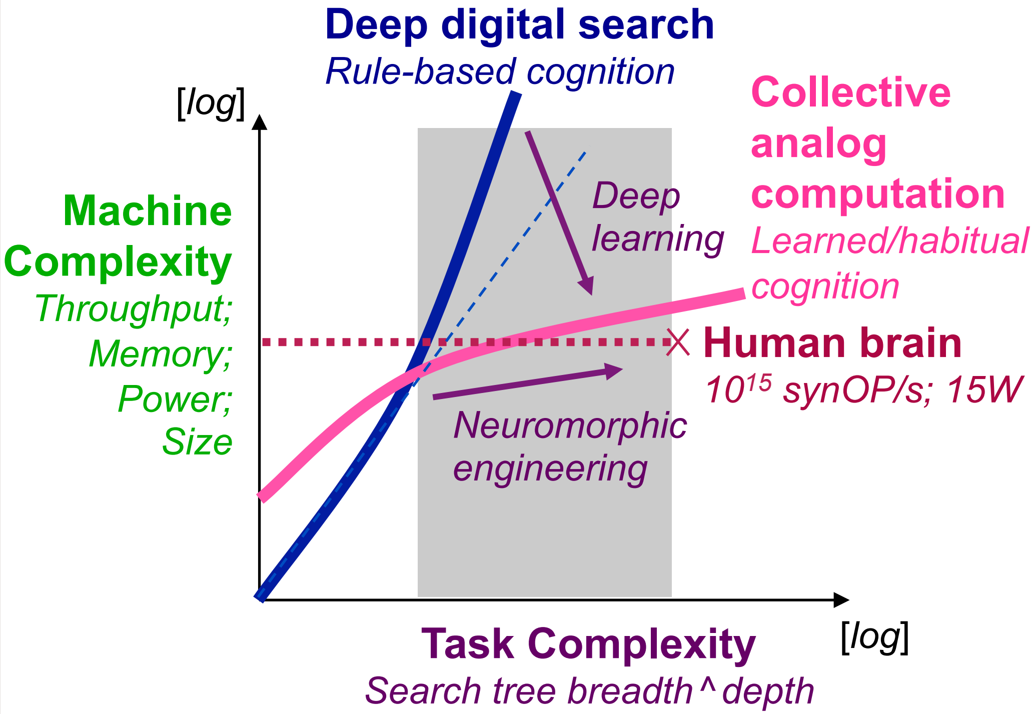
Vogelstein *et al*, NIPS*2002

Spike Timing-Dependent Plasticity on the IFAT

Vogelstein *et al*, NIPS*2002



Scaling of Task and Machine Complexity



Deep digital search

Rule-based cognition

Collective analog computation

Learned/habitual cognition

Machine Complexity

Throughput; Memory; Power; Size

Deep learning

Human brain

10^{15} synOP/s; 15W

Neuromorphic engineering

Task Complexity

Search tree breadth ^ depth

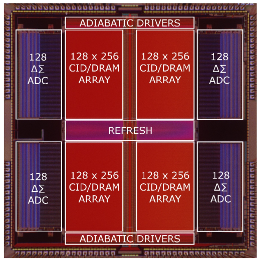
G. Cauwenberghs, "Reverse Engineering the Cognitive Brain," *PNAS*, 2013

Task Energy Efficiency:

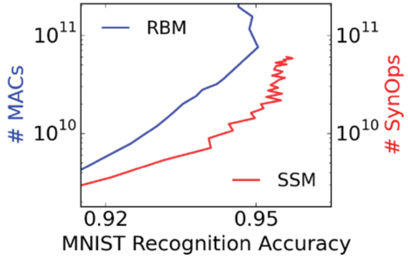
$$\frac{\text{Energy}}{\text{Task}} = \frac{\text{Energy}}{\text{Operation}} \times \frac{\text{Operations}}{\text{Task}}$$

$\frac{1\text{ fJ}}{\text{SynOp}}$ vs. $\frac{10\text{ pJ}}{\text{MAC}}$

$\frac{10^{10}\text{ SynOps vs. }10^{11}\text{ MACs}}{\text{MNIST @ 95\%}}$



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



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

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:**

EE
NanoE
Phys

- 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*

Neuro
CS
CogSci

- 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

	Stromatias 2013 **SpiNNaker** **Manchester**	**Davies 2018** **Loihi** **Intel**	**Merolla 2014** **TrueNorth** **IBM**	**Schemmel 2010** **FACETS/BrainScaleS** **Heidelberg**	**Benjamin 2014** **NeuroGrid** **Stanford**	**Park 2014** **IFAT** **UCSD**
Technology (nm)	130	14	28	180	180	90
Die Size (mm^2)	102	60	430	50	168	16
Neuron Type	Digital Arbitrary	Digital Conductance Integrate & Fire	Digital Accumulate & Fire	Analog Conductance Integrate & Fire	Analog Shared-Dendrite Conductance I&F	Analog 2-Compartment Conductance I&F
# Neurons	5216 [1]	128k [2]	1M [2]	512	65k	65k
Neuron Area (μm^2)	N/A [1]	240 (240k) [2]	14 (3325) [2]	1500	1800	140
Peak Throughput (Events/s)	5M	3.4G	1G	65M	91M	73M
Energy Efficiency (J/SynEvent)	8n	24p	26p	N/A	31p	22p

[1] Software-instantiated neuron model

[2] Time-multiplexed neuron processor

Benjamin, B., P. Gao, E. McQuinn, S. Choudhary, A. Chandrasekaran, J. Bussat, R. Alvarez-Icaza, J. Arthur, P. Merolla, and K. Boahen, "Neurogrid: A mixed analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, 102(5):699–716, 2014.

Davies, M. et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38 (1), pp. 82-99, 2018.

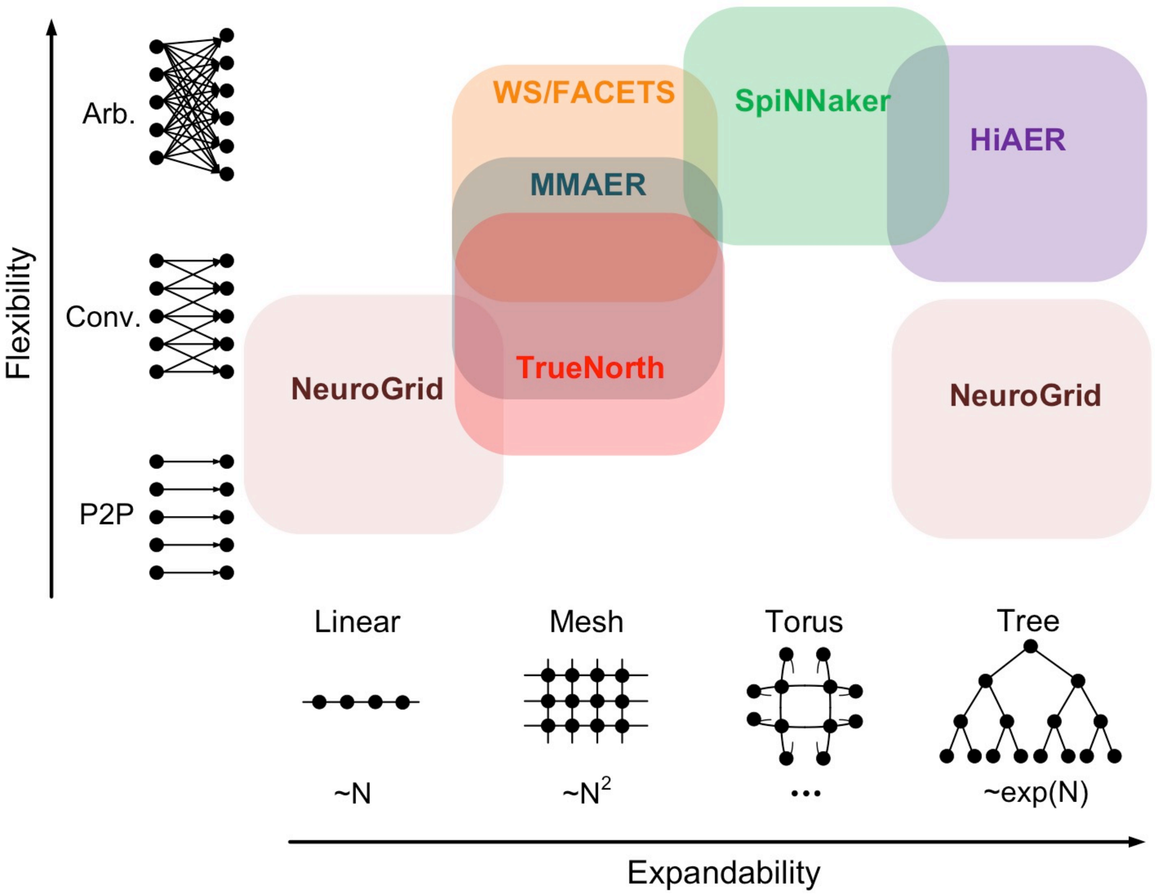
Merolla, P.A., J.V. Arthur, R. Alvarez-Icaza, A S. Cassidy, J. Sawada, F. Akopyan, B.L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S.K. Esser, R. Appuswamy, B. Taba, A. Amir, M.D. Flickner, W.P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, 345(6197):668–673, 2014.

Park, J., S. Ha, T. Yu, E. Neftci, and G. Cauwenberghs, "65k-neuron 73-Mevents/s 22-pJ/event asynchronous micro-pipelined integrate-and-fire array transceiver," *Proc. 2014 IEEE Biomedical Circuits and Systems Conf. (BioCAS)*, 2014.

Schemmel, J., D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner, "A waferscale neuromorphic hardware system for large-scale neural modeling," *Proc. 2010 IEEE Int. Symp. Circuits and Systems (ISCAS)*, 1947–1950, 2010.

Stromatias, E., F. Galluppi, C. Patterson, and S. Furber, "Power analysis of largescale, real-time neural networks on SpiNNaker," *Proc. 2013 Int. Joint Conf. Neural Networks (IJCNN)*, 2013.

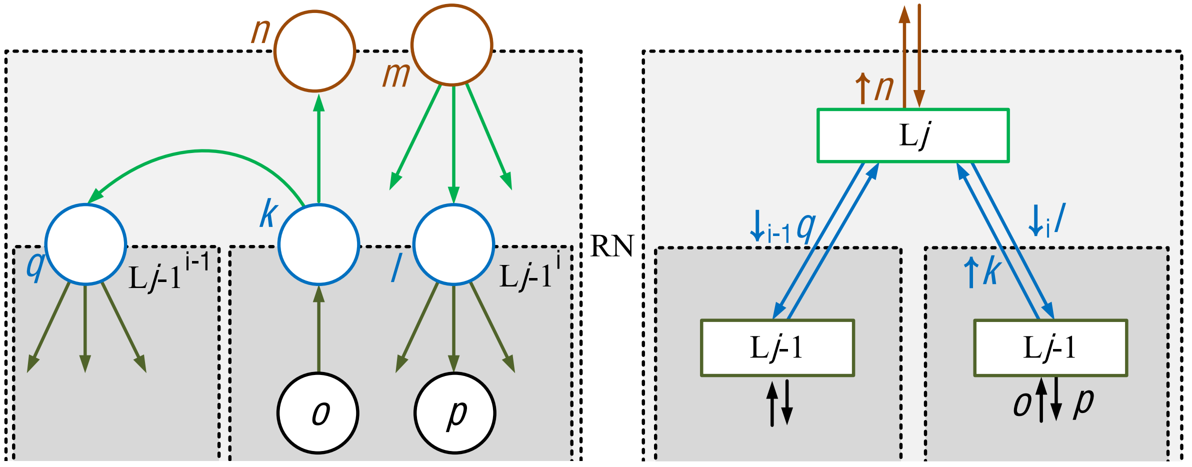
Long-Range Configurable Synaptic Connectivity



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.

Park et al, “Hierarchical Address Event Routing for Reconfigurable Large-Scale Neuromorphic Systems,” *IEEE TNNLS,* 2017

Hierarchical Address-Event Routing (HiAER)



(a)

(b)

L*j* SRT

PRE	POST	PARAMETER
m	$\downarrow_{i} l$	$d_{m,i:l}$
k	$\uparrow n$	$d_{k,n}$
k	$\downarrow_{i-1} q$	$d_{k,i-1:q}$

L*j*-1^{i} SRT

PRE	POST	PARAMETER
o	$\uparrow k$	$d_{o,k}$
l	p	$w_{l,p}$

(c)

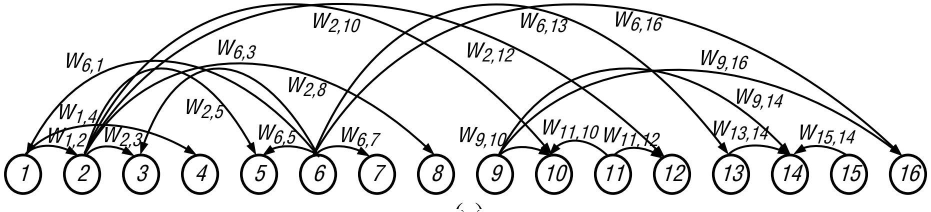
(a) Hierarchical neural network with ascending and descending neural projections. Physical neurons are represented by o and p, and inserted relay neurons (RN) interfacing across hierarchical partitions are denoted by *q*, *k*, *l*, *n*, *m*. Italic indices *j* and *j* – 1 represent levels in the hierarchy,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

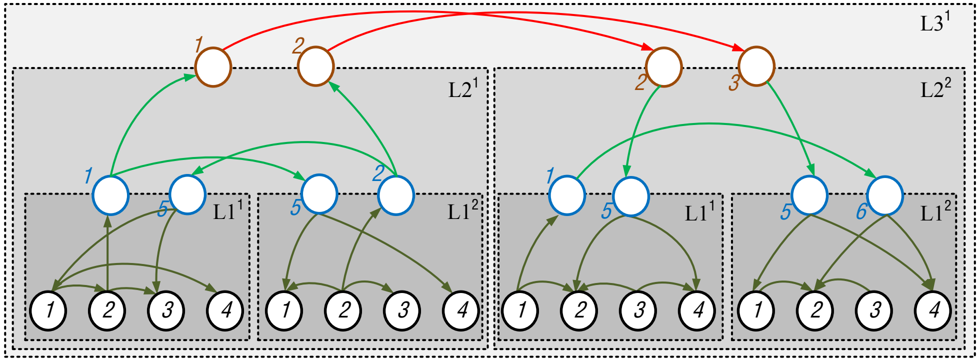
(c) Corresponding entries within the Synaptic Routing Table (SRT).

Joshi et al, 2010; Park et al, 2011, 2015

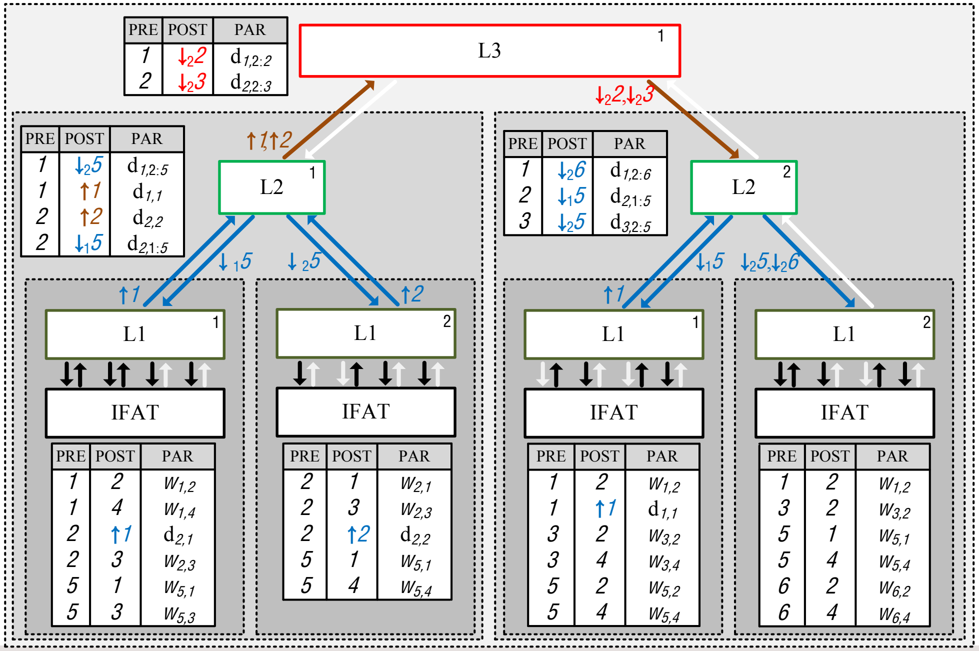
Hierarchical Address-Event Routing (HiAER)



Example network with 16 neurons and weighted synaptic connections.



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



PRE	POST	PAR
1	$\downarrow_2 2$	$d_{1,2:2}$
2	$\downarrow_2 3$	$d_{2,2:3}$

PRE	POST	PAR
1	$\downarrow_2 5$	$d_{1,2:5}$
1	$\uparrow 1$	$d_{1,1}$
2	$\uparrow 2$	$d_{2,2}$
2	$\downarrow_1 5$	$d_{2,1:5}$

PRE	POST	PAR
1	$\downarrow_2 6$	$d_{1,2:6}$
2	$\downarrow_1 5$	$d_{2,1:5}$
3	$\downarrow_2 5$	$d_{3,2:5}$

PRE	POST	PAR
1	2	$W_{1,2}$
1	4	$W_{1,4}$
2	$\uparrow 1$	$d_{2,1}$
2	3	$W_{2,3}$
5	1	$W_{5,1}$
5	3	$W_{5,3}$

PRE	POST	PAR
2	1	$W_{2,1}$
2	3	$W_{2,3}$
2	$\uparrow 2$	$d_{2,2}$
5	1	$W_{5,1}$
5	4	$W_{5,4}$

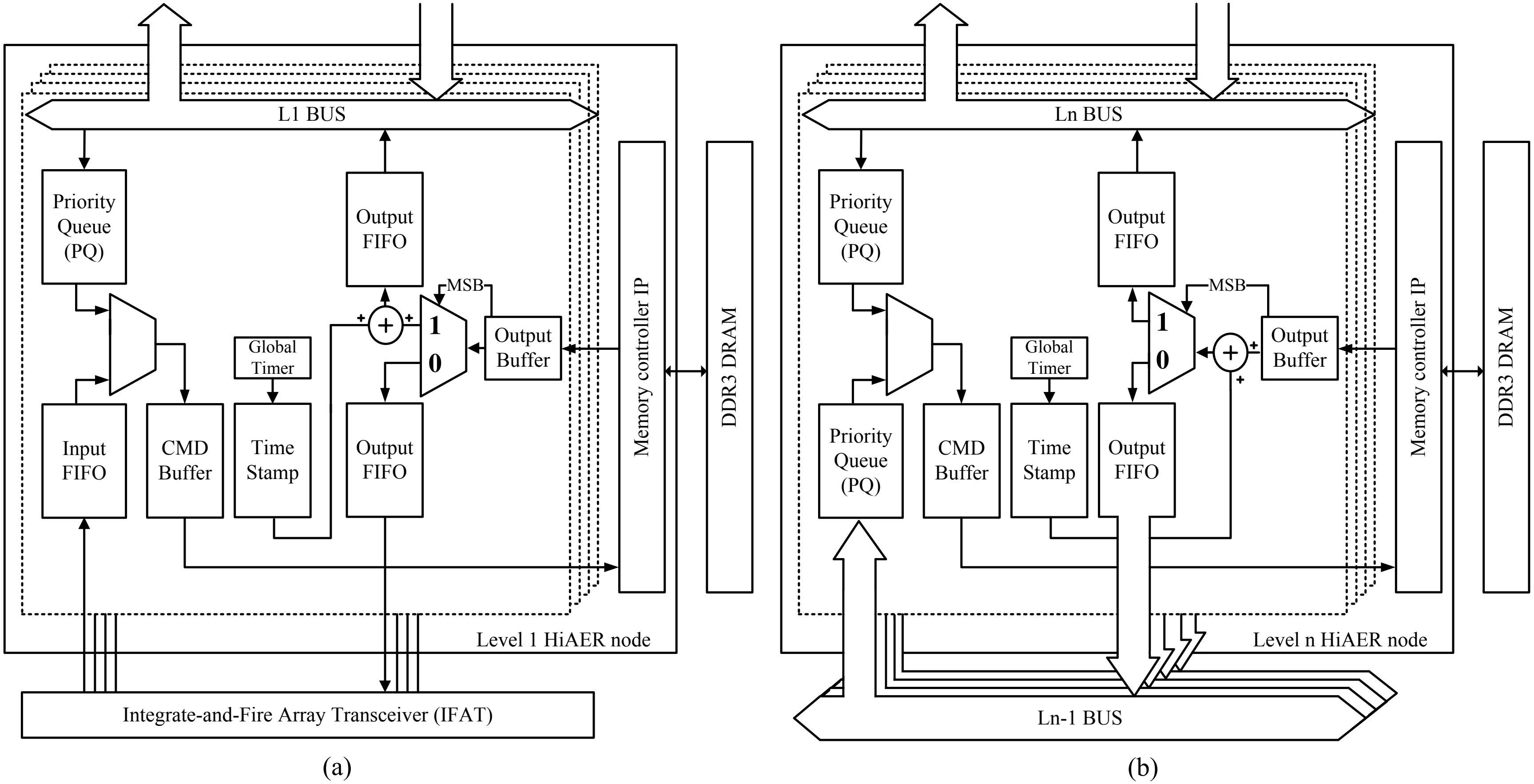
PRE	POST	PAR
1	2	$W_{1,2}$
1	$\uparrow 1$	$d_{1,1}$
3	2	$W_{3,2}$
3	4	$W_{3,4}$
5	2	$W_{5,2}$
5	4	$W_{5,4}$

PRE	POST	PAR
1	2	$W_{1,2}$
3	2	$W_{3,2}$
5	1	$W_{5,1}$
5	4	$W_{5,4}$
6	2	$W_{6,2}$
6	4	$W_{6,4}$

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

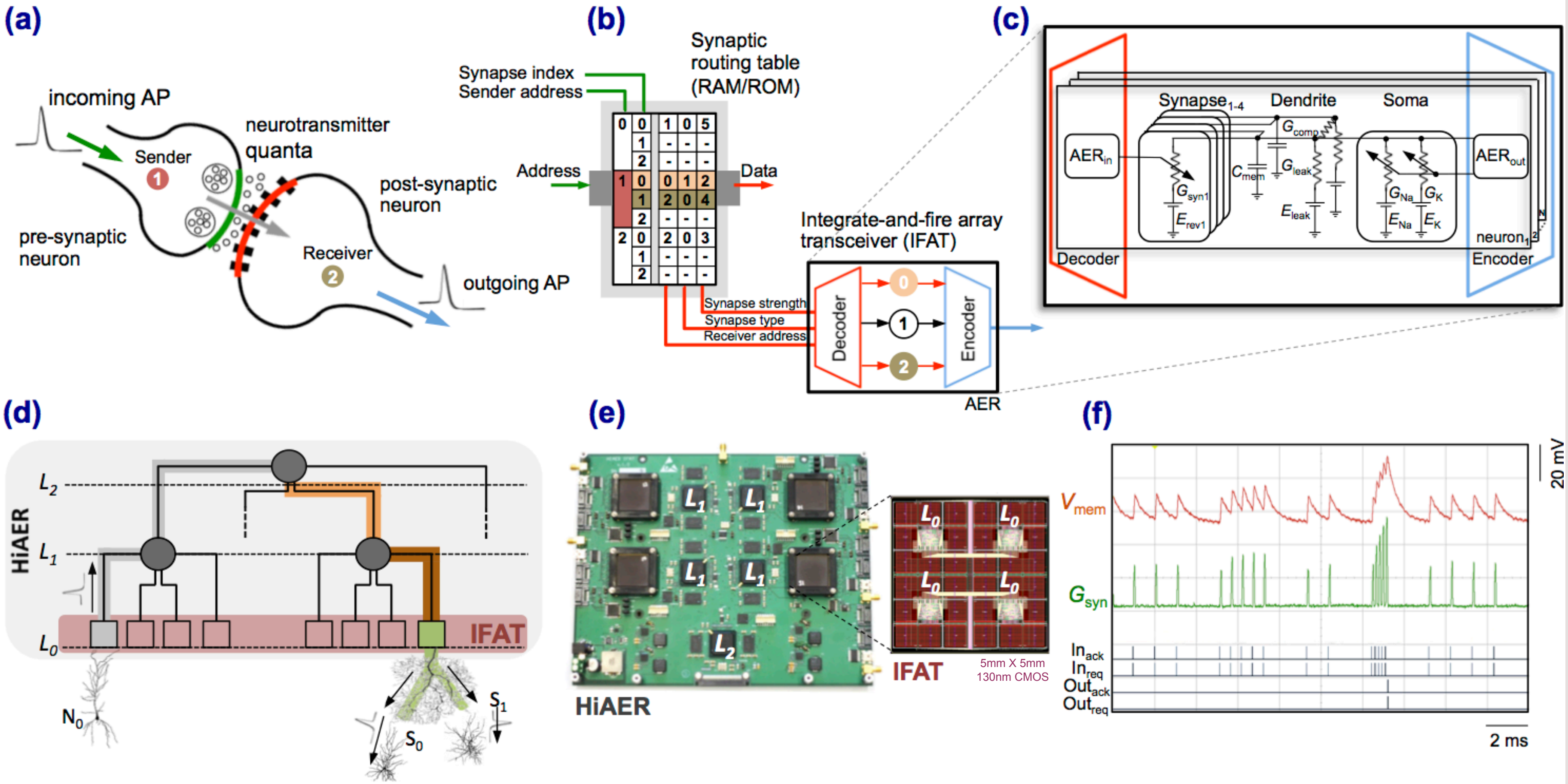
Hierarchical Address-Event Routing (HiAER)



(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*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 *Ln* – 1 bus, and of the *L*1 bus with a *Ln* 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*).

Large-Scale Reconfigurable Neuromorphic Computing



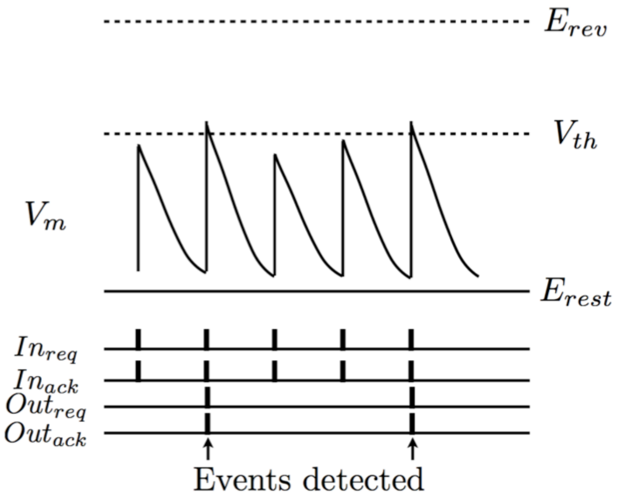
0	0	1	0	5
	1	-	-	-
	2	-	-	-
1	0	0	1	2
	1	2	0	4
	2	-	-	-
2	0	2	0	3
	1	-	-	-
	2	-	-	-

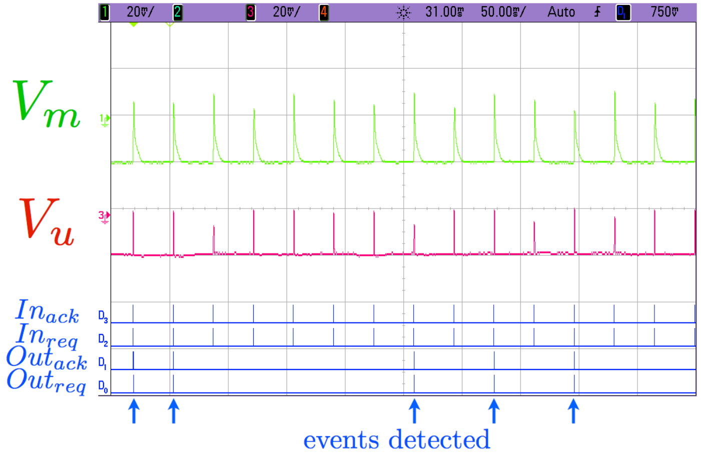
Hierarchical Address-Event Routing (HiAER) Integrate-and-Fire Array Transceiver (IFAT) for scalable and reconfigurable neuromorphic neocortical processing. (a) Biophysical model of neural and synaptic dynamics. (b) Dynamically reconfigurable synaptic connectivity is implemented across IFAT arrays of 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.

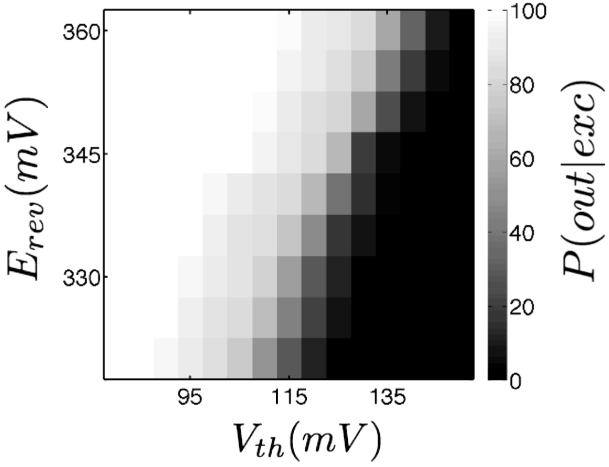
Yu et al, BioCAS 2012; Park et al, BioCAS 2014; Park et al, TNNLS 2017; Broccard et al, JNE 2017

IFAT Thermodynamics of Neural Excitability

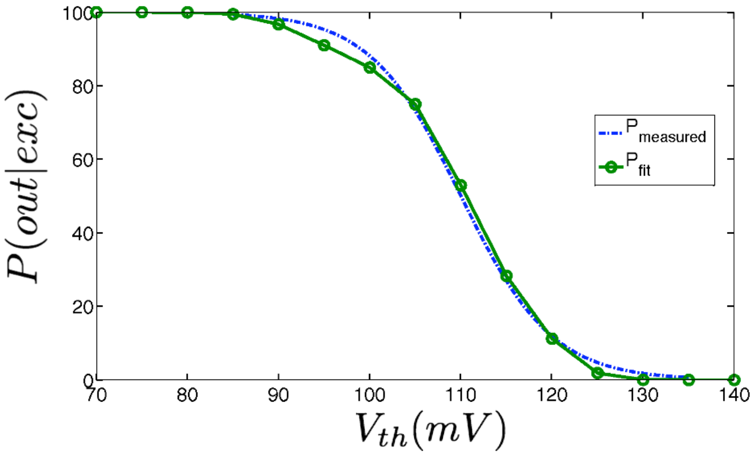
Yu, Park, Joshi, Maier, Cauwenberghs, BioCAS 2012





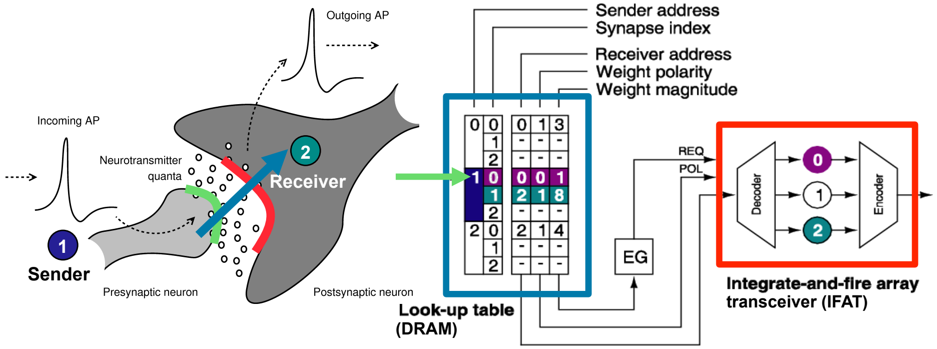


$$V_{gs} = \kappa E_{exc} - V_{th} - V_{TH,N}$$

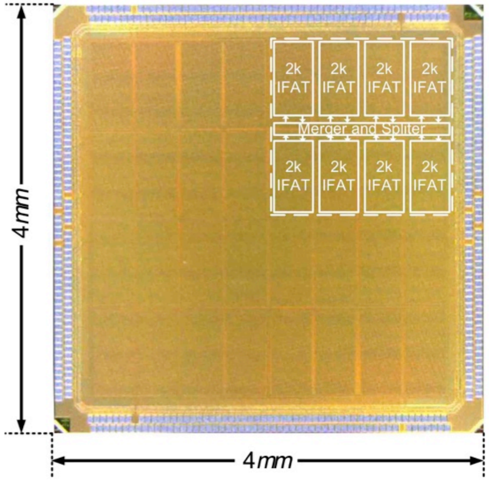


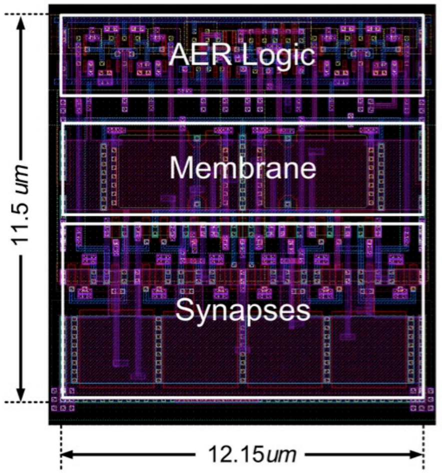
$$\begin{aligned} P(out|exc) &= f(V_{gs}) \\ &= 1/(1+e^{-\frac{V_{gs}}{U_T}}) \end{aligned}$$

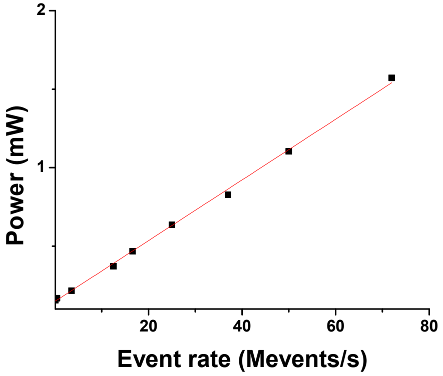
Large-Scale Reconfigurable Neuromorphic Computing

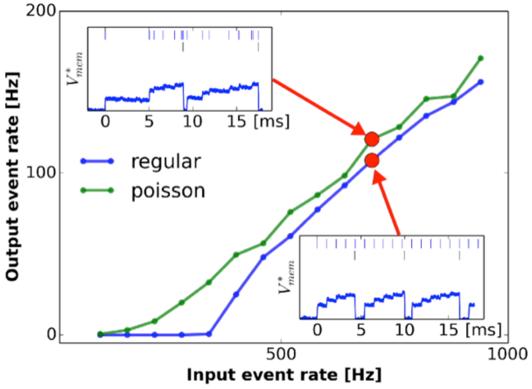


Sender address	Synapse index	Receiver address	Weight polarity	Weight magnitude
0	0	0	1	3
	1	-	-	-
	2	-	-	-
1	0	0	0	1
	1	2	1	8
	2	-	-	-
2	0	2	1	4
	1	-	-	-
	2	-	-	-





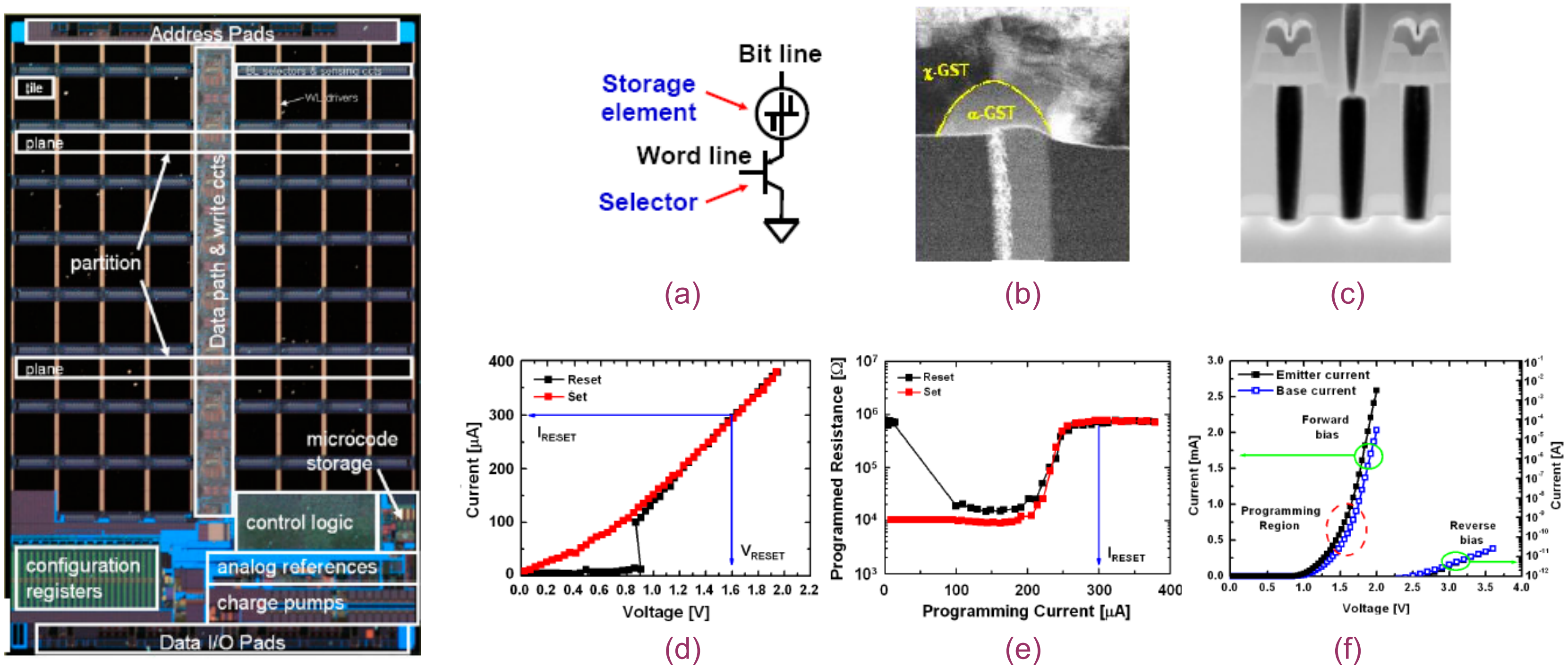




- *Integrate-and-fire array transceiver (IFAT) as digitally programmable analog neural supercomputer*
- *Biophysical detail in neural and synaptic continuous-time dynamics*
- *Record high density: 65k two-compartment neurons with 65M reconfigurable conductance-based synapses*
- *Record low energy: 22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

J. Park *et al*, “A 65k-Neuron 73-Mevents/s 22-pJ/event Asynchronous Micro-Pipelined Integrate-and-Fire Array Transceiver”, Proc. IEEE BioCAS 2014.

Memristive Synapse Arrays for Neuromorphic Processing-in-Memory



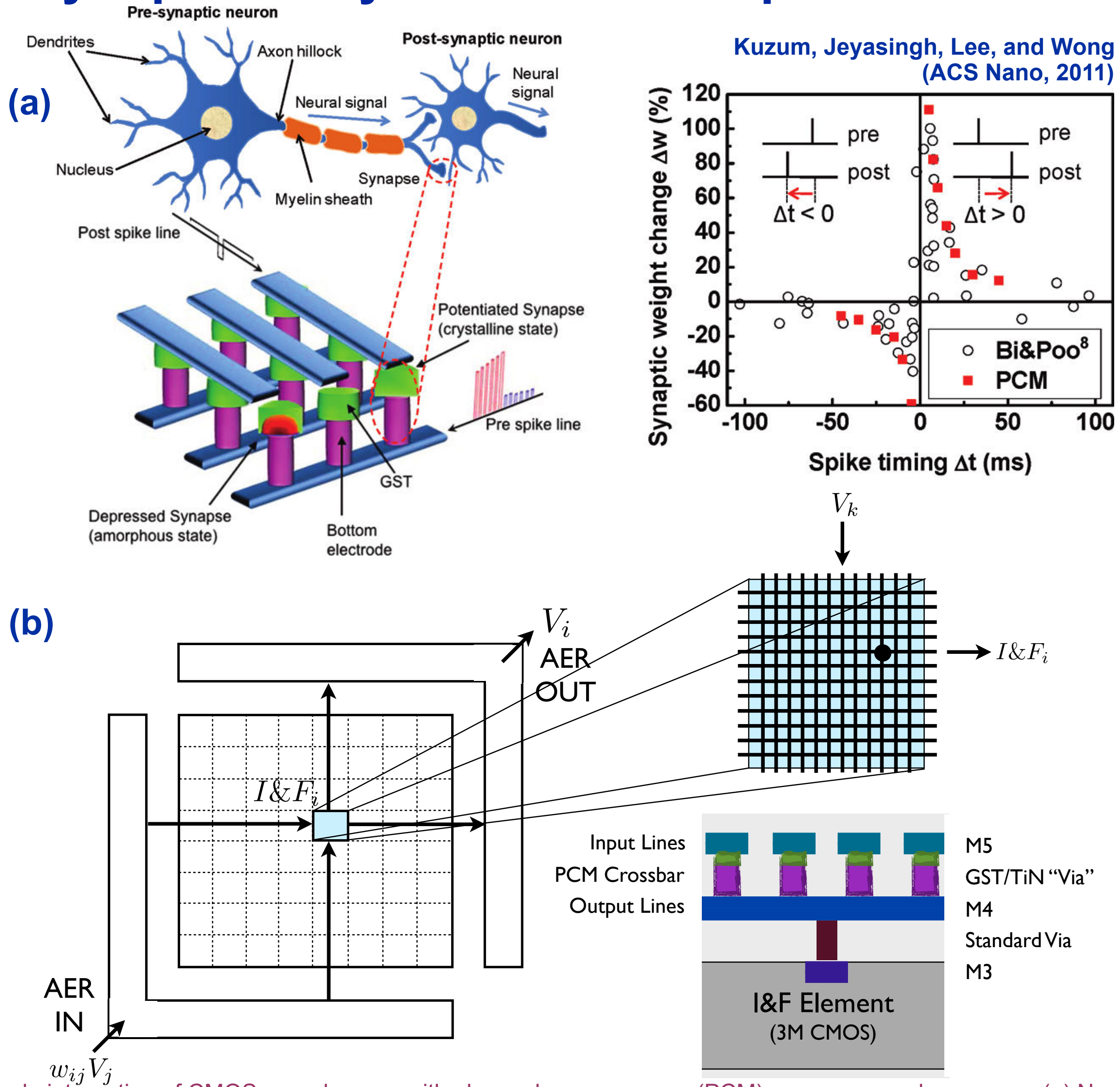
(a) (b) (c)

(d) (e) (f)

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



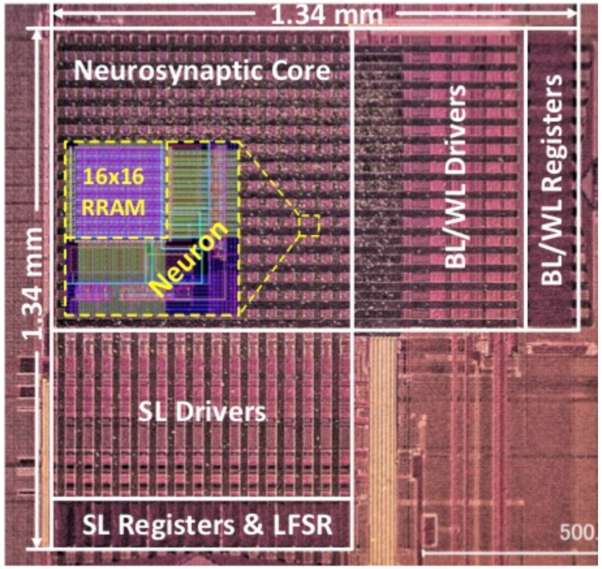
Kuzum, Jeyasingh, Lee, and Wong (ACS Nano, 2011)

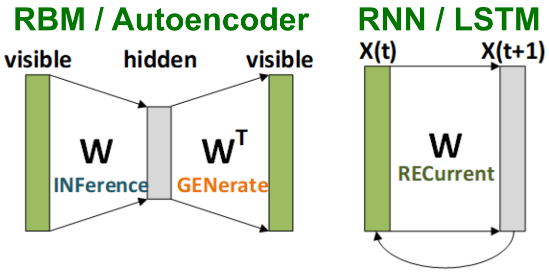
Hybridization and nanoscale integration of CMOS neural arrays with phase change memory (PCM) synapse crossbar arrays. (a) Nanoelectronic PCM synapse with spike-timing dependent plasticity (STDP) [Kuzum *et al*, 2011]. Each PCM element implements a synapse with conductance modulated through phase transition 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.

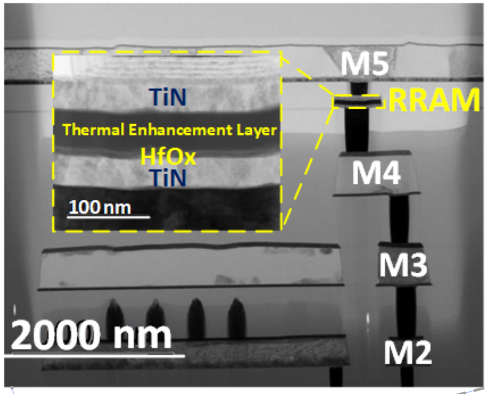
CMOS-RRAM Reconfigurable Neurosynaptic Array

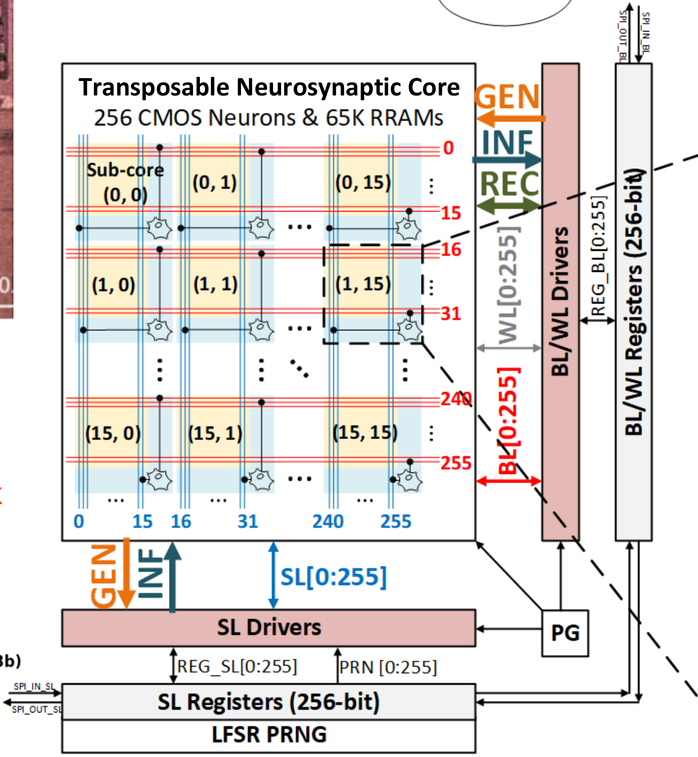
Wan et al, ISSCC 2020

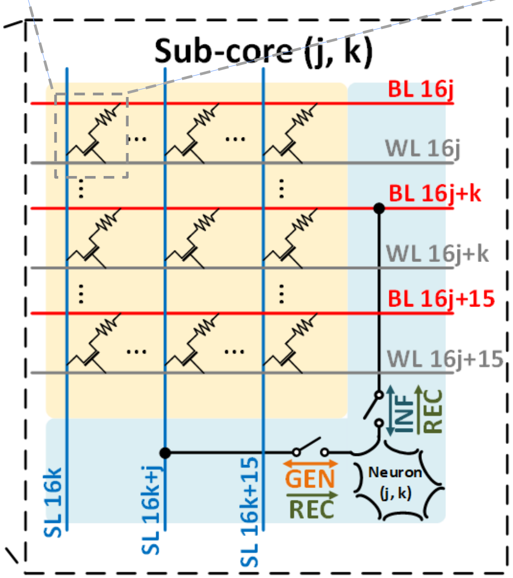
- Integration of CMOS neurons and resistive random-access memory (RRAM) memristive synapses with *in-situ* revertible dataflow at record efficiency

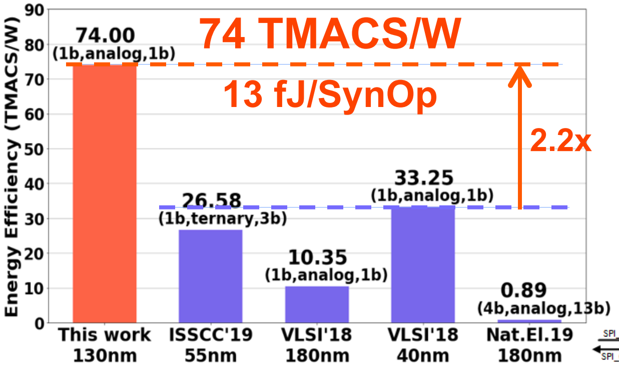












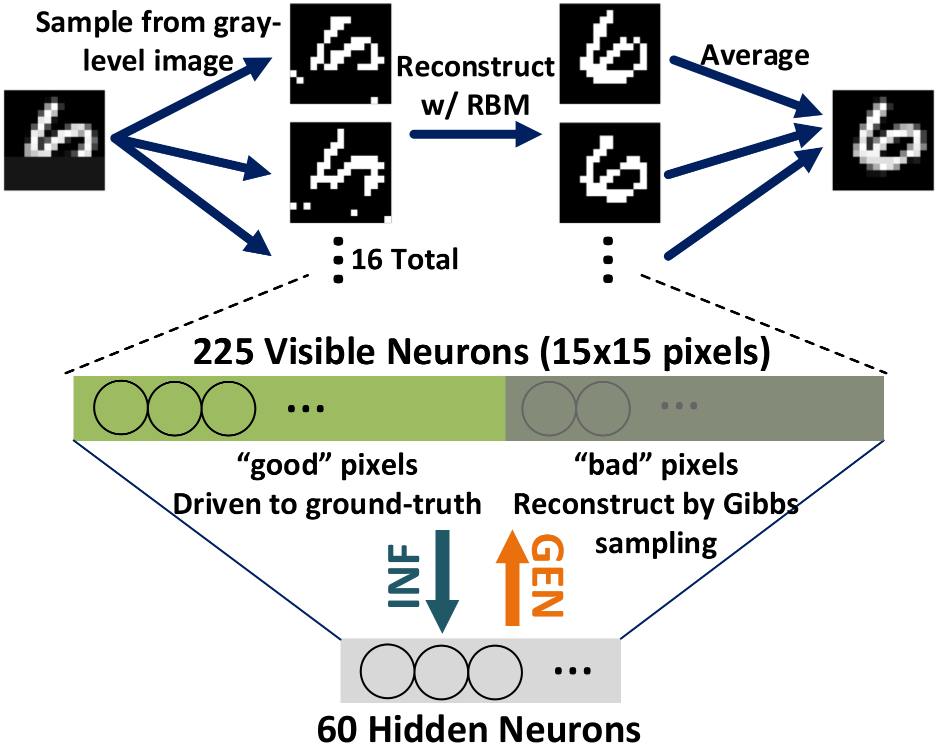
CMOS-RRAM Processor-in-Memory

W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, H-S. P. Wong, “33.1: A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models,” *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco CA, Febr. 15-19, 2020.

CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020

- Gibbs stochastic sampling for Bayesian generative inference
  - *Alternating between **INF**erence and **GEN**eration 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

16 Total

225 Visible Neurons (15x15 pixels)

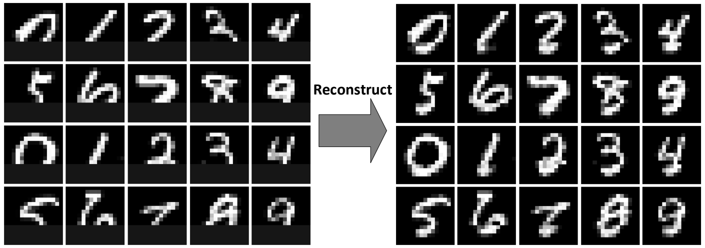
"good" pixels
Driven to ground-truth

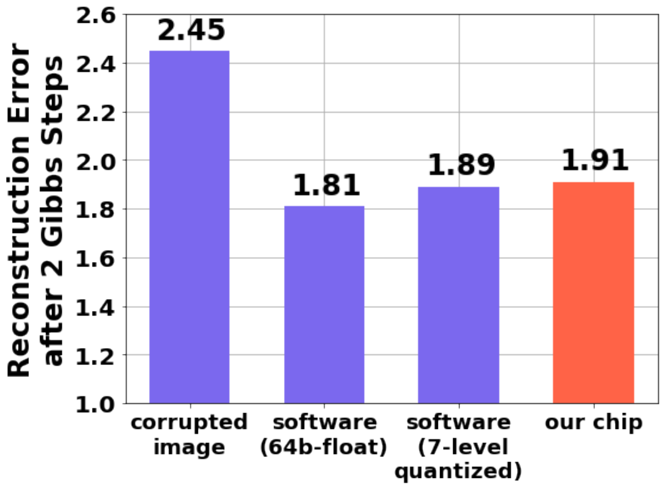
"bad" pixels
Reconstruct by Gibbs sampling

INF

GEN

60 Hidden Neurons

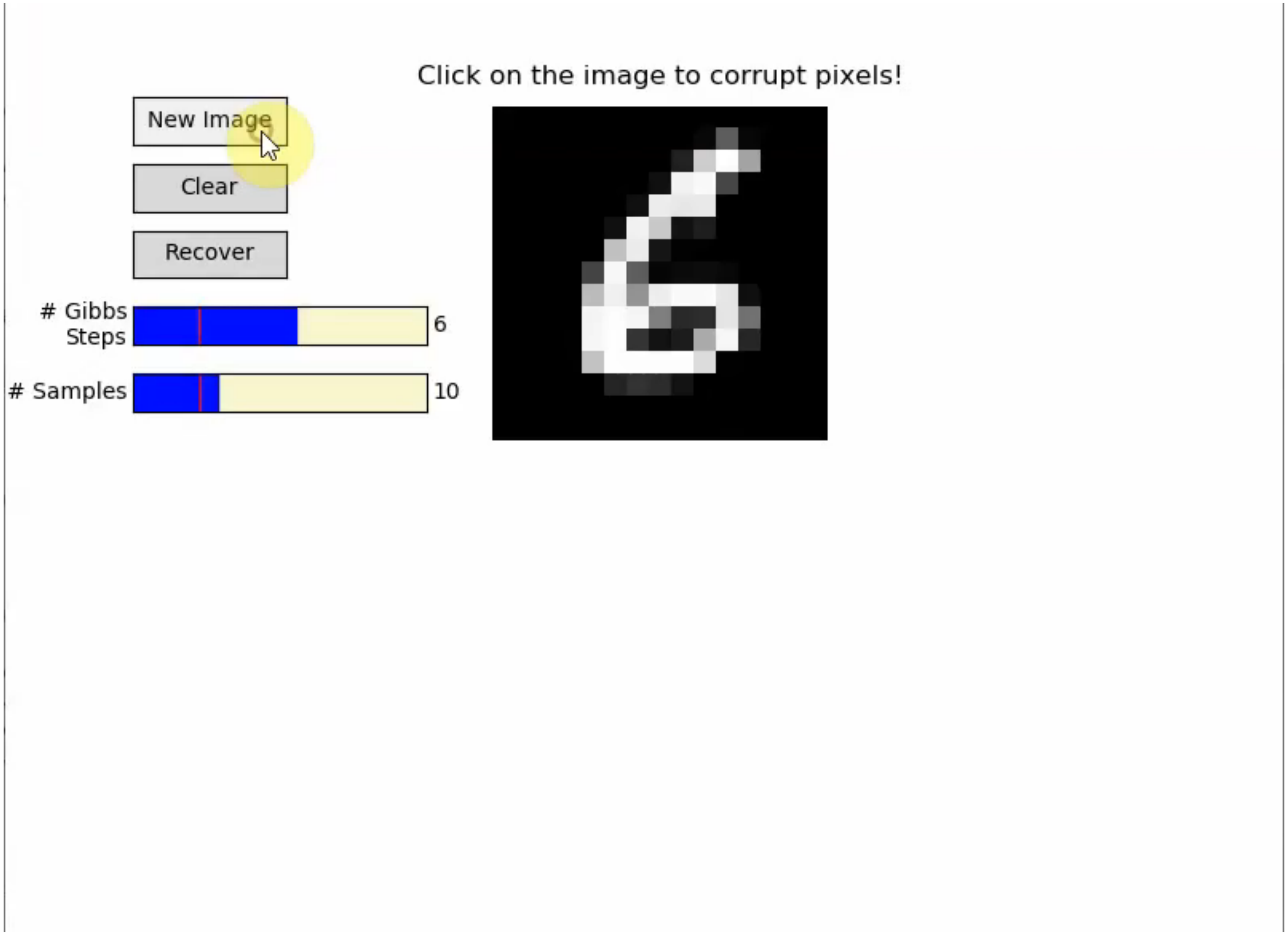




W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, H-S. P. Wong, "33.1: A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models," *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco CA, Febr. 15-19, 2020.

CMOS-RRAM Reconfigurable Neurosynaptic Array

Wan et al, ISSCC 2020



Click on the image to corrupt pixels!

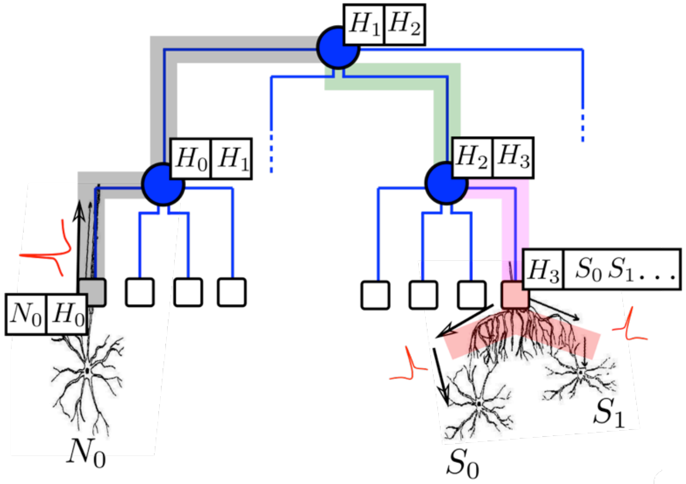
Gibbs Steps 6

Samples 10

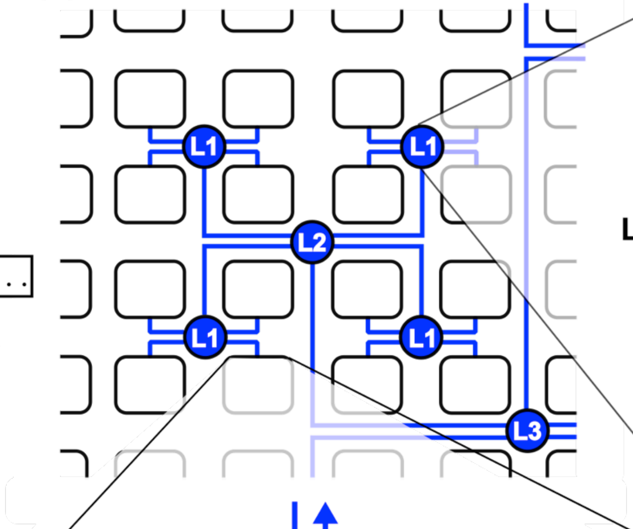
W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, H-S. P. Wong, “33.1: A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models,” *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco CA, Febr. 15-19, 2020.

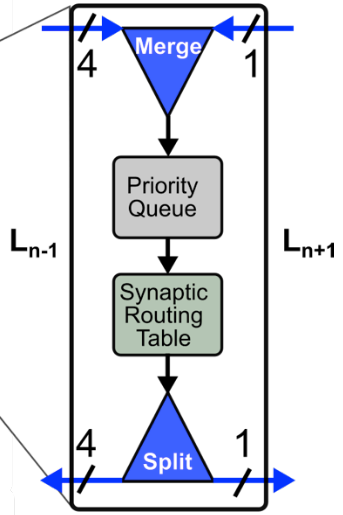
Hierarchical Address-Event Routing
Neural and Synaptic Array Transceiver
HiAER-NSAT

(b) Connectivity Model



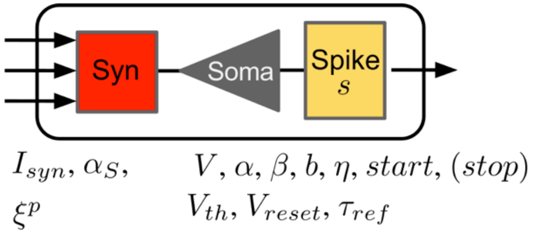
(c) HiAER Tree





(d) HiAER Node

(a) Neuron and Synapse Model

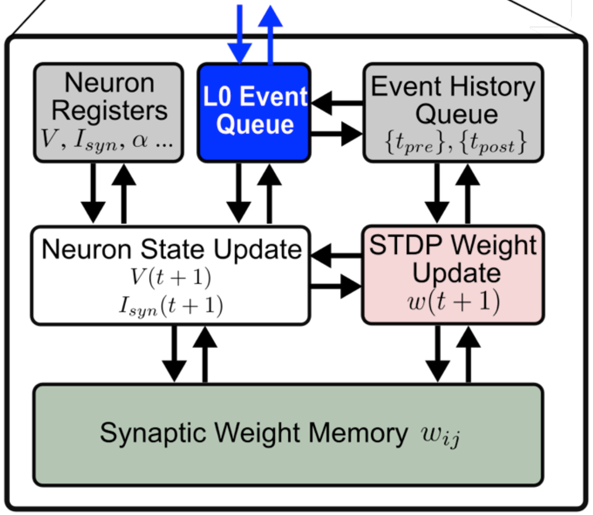


Syn
Soma
Spike s
I_{syn}, α_S, ξ^p
$V, \alpha, \beta, b, \eta, start, (stop)$
$V_{th}, V_{reset}, \tau_{ref}$

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

$$I_{syn,i}(t+1) = \alpha_S I_{syn,i}(t) + \textstyle\sum_j \xi^p_{ij} w_{ij} s_j\left(t - \theta_{ij}\right),$$

$$\Delta w_{ij}(t) \;=\; g(t)\left(s_j(t) e_i^{post}(t) + s_i(t) e_j^{pre}(t)\right)$$

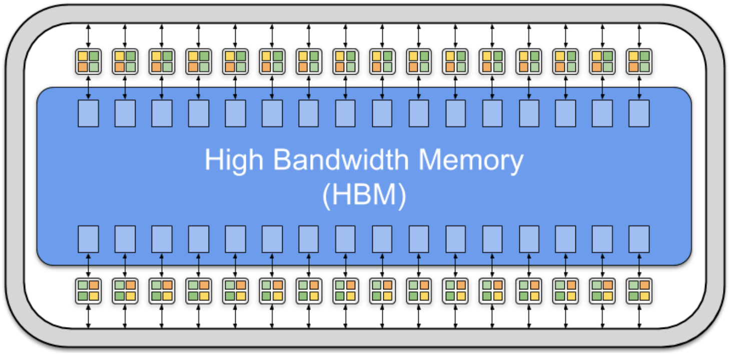


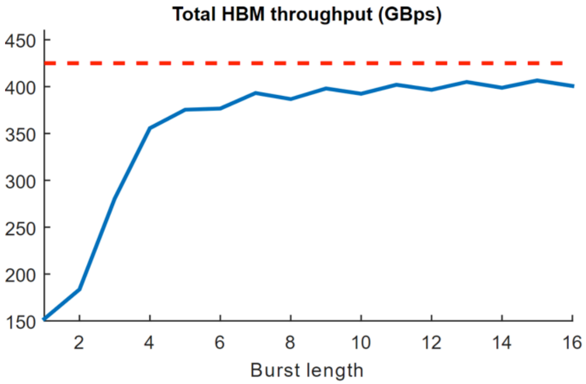
Neuron Registers V, I_{syn}, α ...
L0 Event Queue
Event History Queue $\{t_{pre}\}, \{t_{post}\}$
Neuron State Update $V(t+1)$ $I_{syn}(t+1)$
STDP Weight Update $w(t+1)$
Synaptic Weight Memory w_{ij}

(e) NSAT Core

G. Detorakis, S. Sheik,
C. Augustine, S. Paul,
B.U. Pedroni, N. Dutt,
J. Krichmar,
G. Cauwenberghs, and
E. Neftci, *Frontiers in Neuroscience*, 2018

HBM FPGA Reconfigurable Neuromorphic Computing





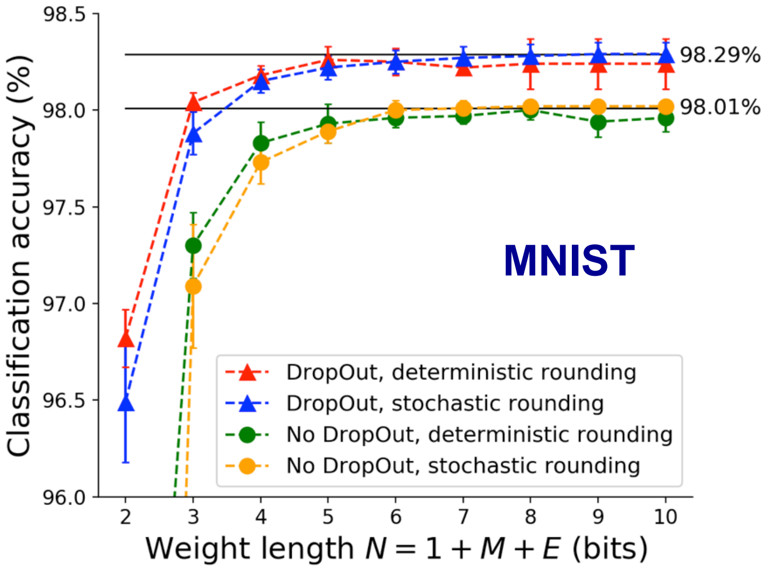
- Reconfigurable, high-throughput neuromorphic processing-in-memory (PIM)
  - *Xilinx UltraScale+ VU37P field-programmable gate array (FPGA)*
- High-Bandwidth Memory (HBM) for extreme PIM throughput
  - *Integrated 8GB HBM2.0 DDR4 SDRAM*
  - *Sustained > 400 GB/s random-access memory bandwidth at 25 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*

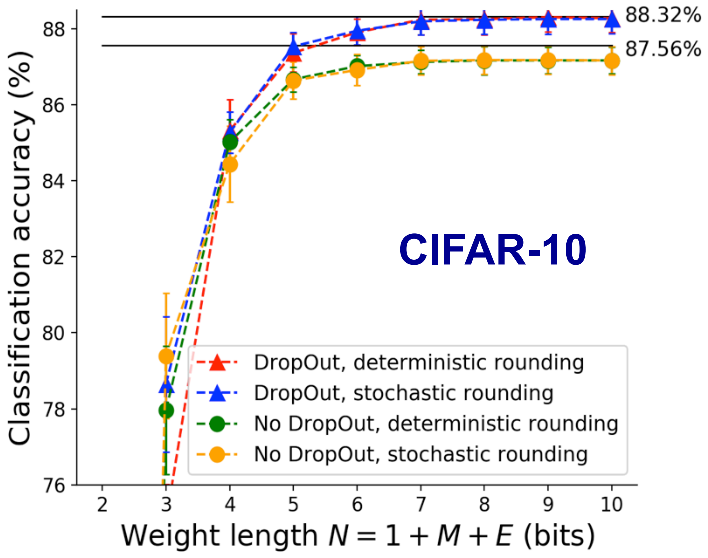
B. Pedroni, S. Deiss, N. Mysore, and G. Cauwenberghs (2020). "Design Principles of Large-Scale Neuromorphic Systems Centered on High Bandwidth Memory", *IEEE Int. Conf. on Rebooting Computing* (ICRC'2020), Nov. 2020.

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^{2^E}*

$$w = s\ 1.x_1 \ldots x_M\ 2^{e_1 \ldots e_E}$$
$$N = 1 + M + E$$



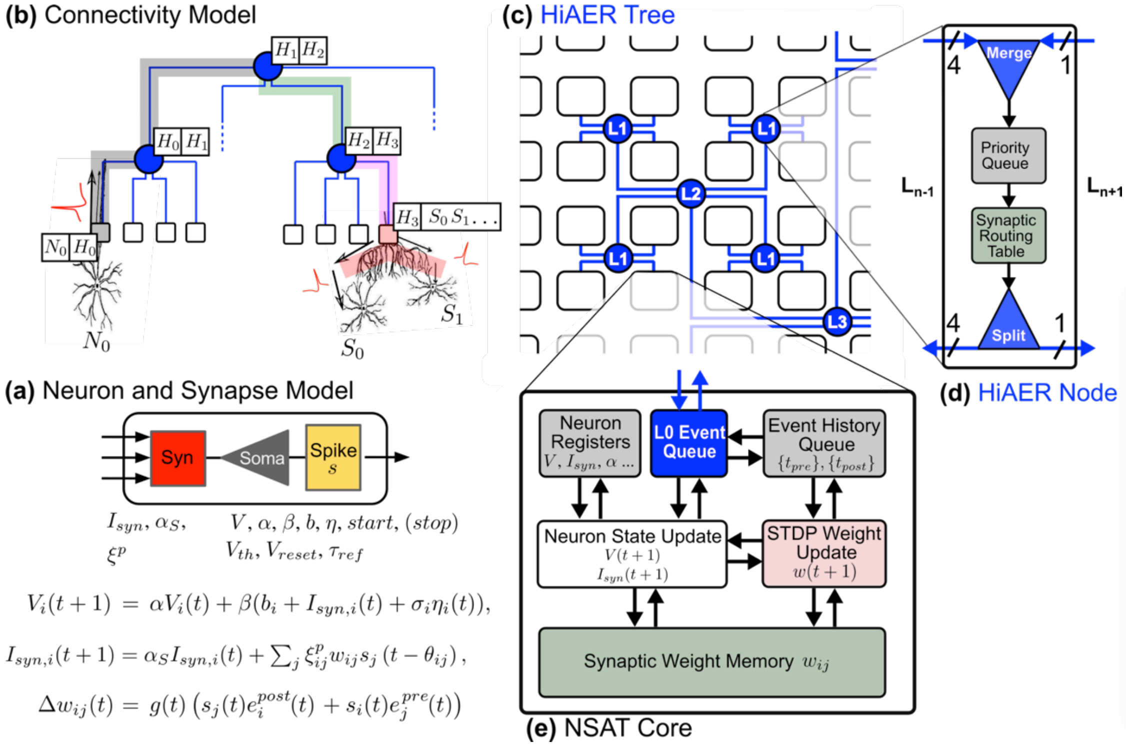


"Performance Trade-offs in Weight Quantization for Memory-Efficient Inference," P.M. Tostado, B.U. Pedroni and G. Cauwenberghs, *IEEE Int. Conf. Artificial Intelligence Circuits and Systems (AICAS'2019)*, Hsinchu Taiwan, March 18-20, 2019.

Large-Scale Reconfigurable Neuromorphic Computing

“CRI: CI-NEW: Trainable Reconfigurable Development Platform for Large-Scale Neuromorphic Cognitive Computing,” National Science Foundation CNS-1823366, G. Cauwenberghs (PI), E. Neftci, and A. Majumdar, 8/2018-7/2021.

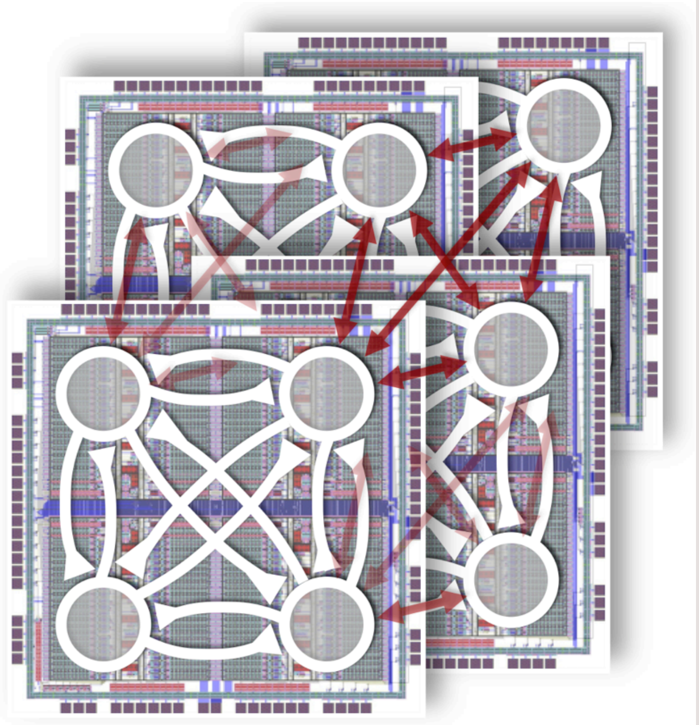
Provides open access to large-scale reconfigurable neuromorphic computing hardware and software as an experimental testbed and development platform with up to 128M neurons and 32B synapses for the research community at large.


(b) Connectivity Model
H_1 H_2
H_0 H_1
H_2 H_3
H_3 S_0 S_1...
N_0 H_0
N_0
S_0
S_1
(c) HiAER Tree
L1
L2
L3
(d) HiAER Node
Merge
Priority Queue
Synaptic Routing Table
Split
4
1
L_{n-1}
L_{n+1}
(a) Neuron and Synapse Model
Syn
Soma
Spike s
I_{syn}, α_S,
ξ^p
V, α, β, b, η, $start$, $(stop)$
V_{th}, V_{reset}, τ_{ref}
(e) NSAT Core
Neuron Registers V, I_{syn}, α ...
L0 Event Queue
Event History Queue $\{t_{pre}\}, \{t_{post}\}$
Neuron State Update $V(t+1)$ $I_{syn}(t+1)$
STDP Weight Update $w(t+1)$
Synaptic Weight Memory w_{ij}

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

$$I_{syn,i}(t+1) = \alpha_S I_{syn,i}(t) + \textstyle\sum_j \xi^p_{ij} w_{ij} s_j\left(t - \theta_{ij}\right),$$

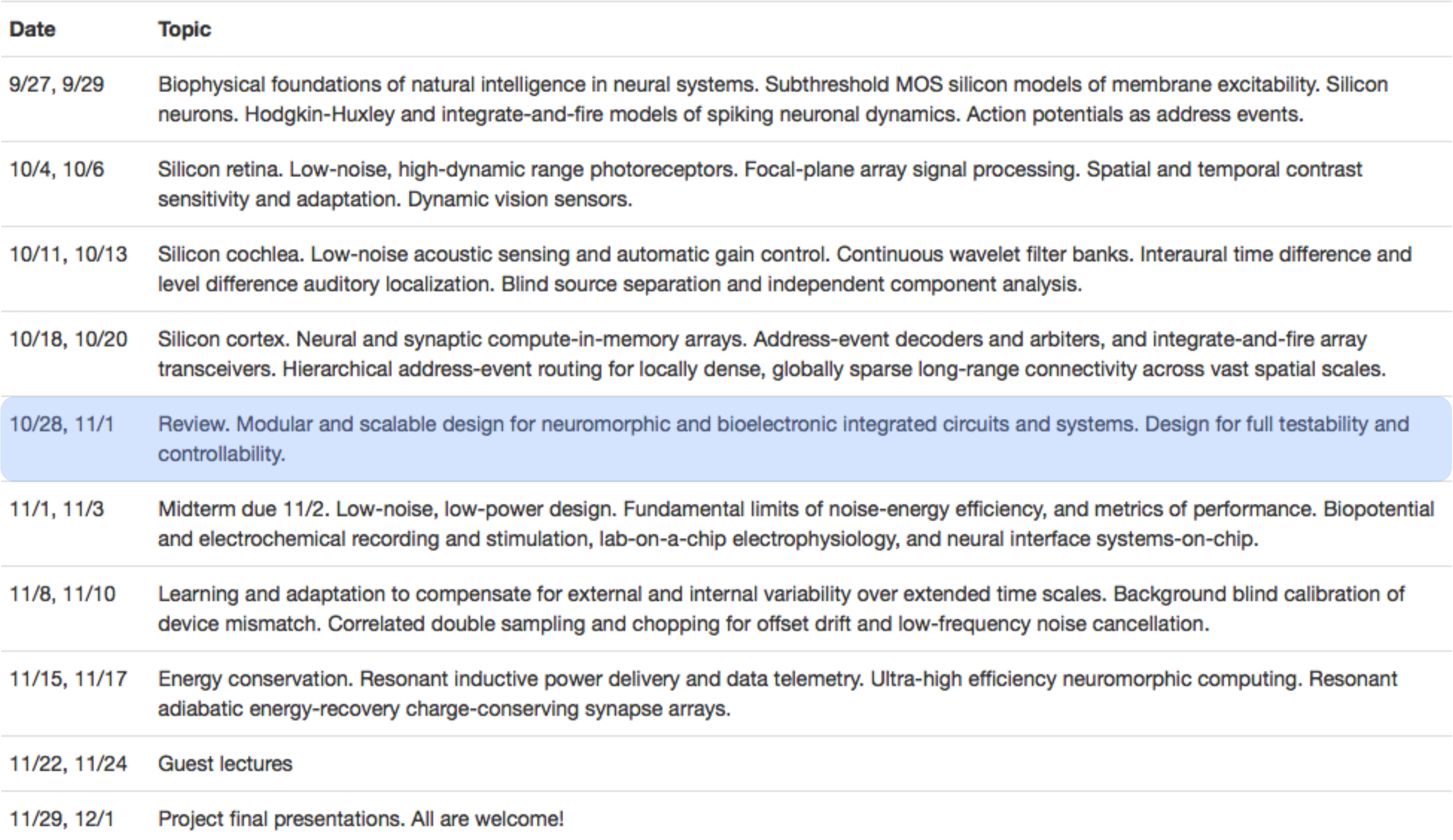
$$\Delta w_{ij}(t) = g(t)\left(s_j(t) e_i^{post}(t) + s_i(t) e_j^{pre}(t)\right)$$



Neural-Synaptic Array Transceiver (Detorakis et al, Frontiers in Neuroscience, 2018)

NeuroDyn (2021 Telluride Neuromorphic Workshop)

BENG 207 Neuromorphic Integrated Bioelectronics



Date	Topic
9/27, 9/29	Biophysical foundations of natural intelligence in neural systems. Subthreshold MOS silicon models of membrane excitability. Silicon neurons. Hodgkin-Huxley and integrate-and-fire models of spiking neuronal dynamics. Action potentials as address events.
10/4, 10/6	Silicon retina. Low-noise, high-dynamic range photoreceptors. Focal-plane array signal processing. Spatial and temporal contrast sensitivity and adaptation. Dynamic vision sensors.
10/11, 10/13	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.
10/18, 10/20	Silicon cortex. Neural and synaptic compute-in-memory arrays. Address-event decoders and arbiters, and integrate-and-fire array transceivers. Hierarchical address-event routing for locally dense, globally sparse long-range connectivity across vast spatial scales.
10/28, 11/1	Review. Modular and scalable design for neuromorphic and bioelectronic integrated circuits and systems. Design for full testability and controllability.
11/1, 11/3	Midterm due 11/2. Low-noise, low-power design. Fundamental limits of noise-energy efficiency, and metrics of performance. Biopotential and electrochemical recording and stimulation, lab-on-a-chip electrophysiology, and neural interface systems-on-chip.
11/8, 11/10	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.
11/15, 11/17	Energy conservation. Resonant inductive power delivery and data telemetry. Ultra-high efficiency neuromorphic computing. Resonant adiabatic energy-recovery charge-conserving synapse arrays.
11/22, 11/24	Guest lectures
11/29, 12/1	Project final presentations. All are welcome!