Energy Efficiency in Adaptive Neural Circuits

Gert Cauwenberghs

Department of Bioengineering
Institute for Neural Computation
UC San Diego

http://inc.ucsd.edu

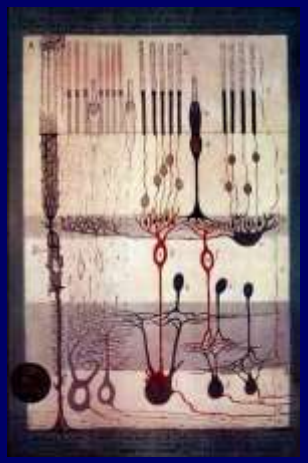
Lee Sedol vs. AlphaGo

Go World Champion vs. Google DeepMind

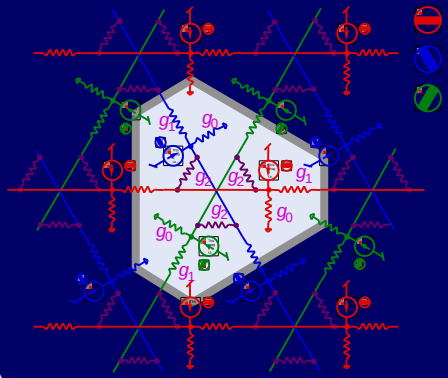
~ 100 W **~ 100 kW**

Neuromorphic Engineering

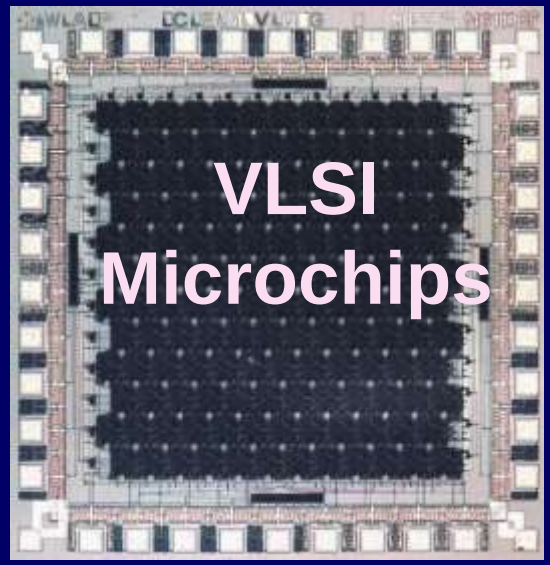
"in silico" neural systems design



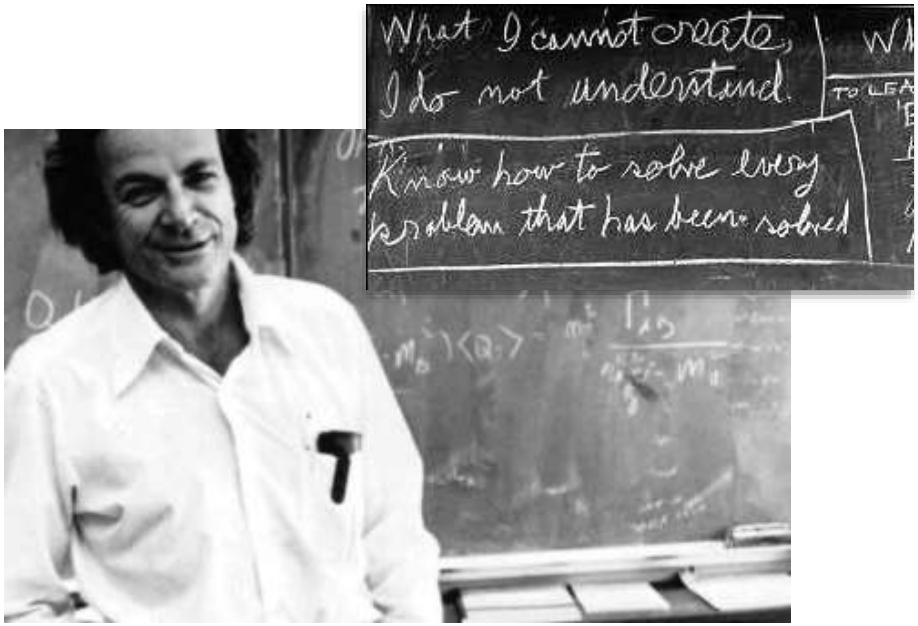
Neuromorphic Engineering



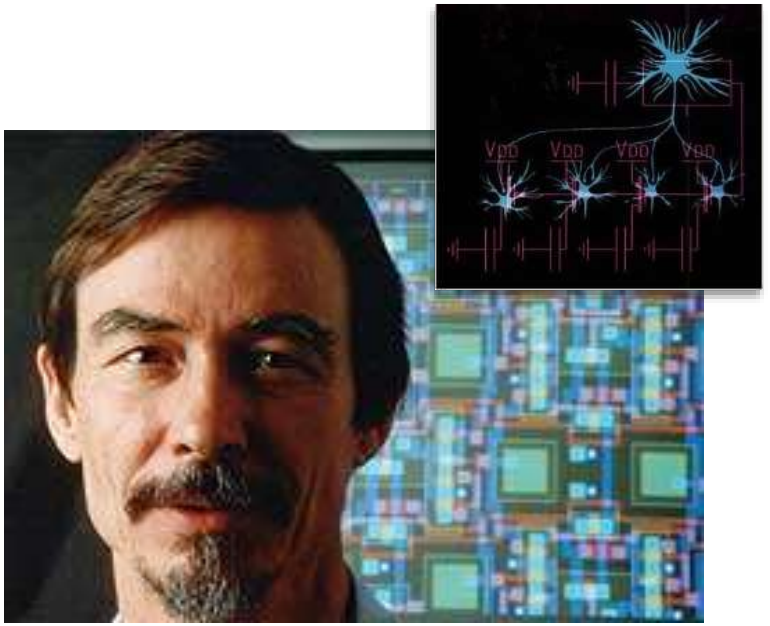
Learning & Adaptation



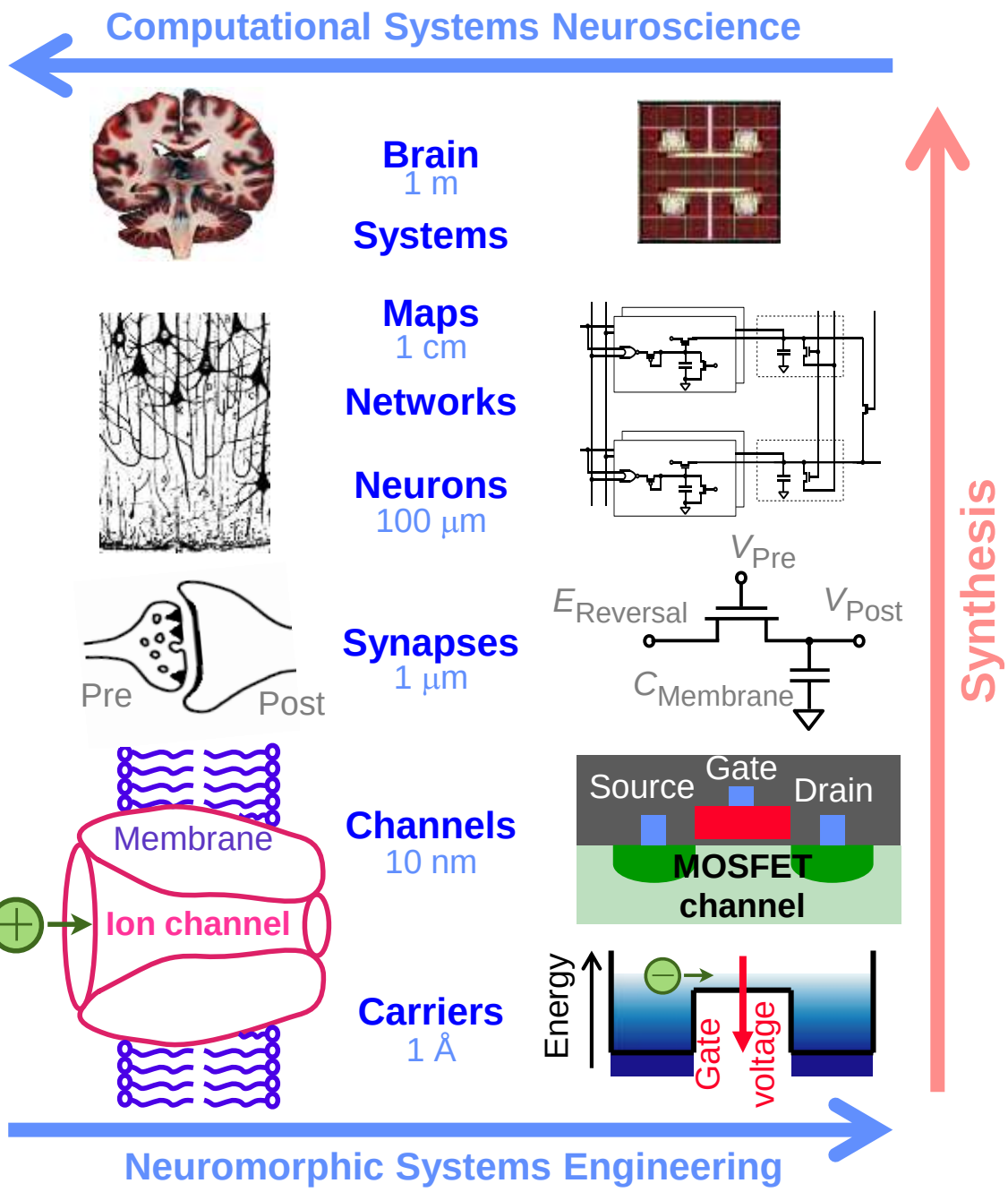
Analysis by Synthesis



Richard Feynman



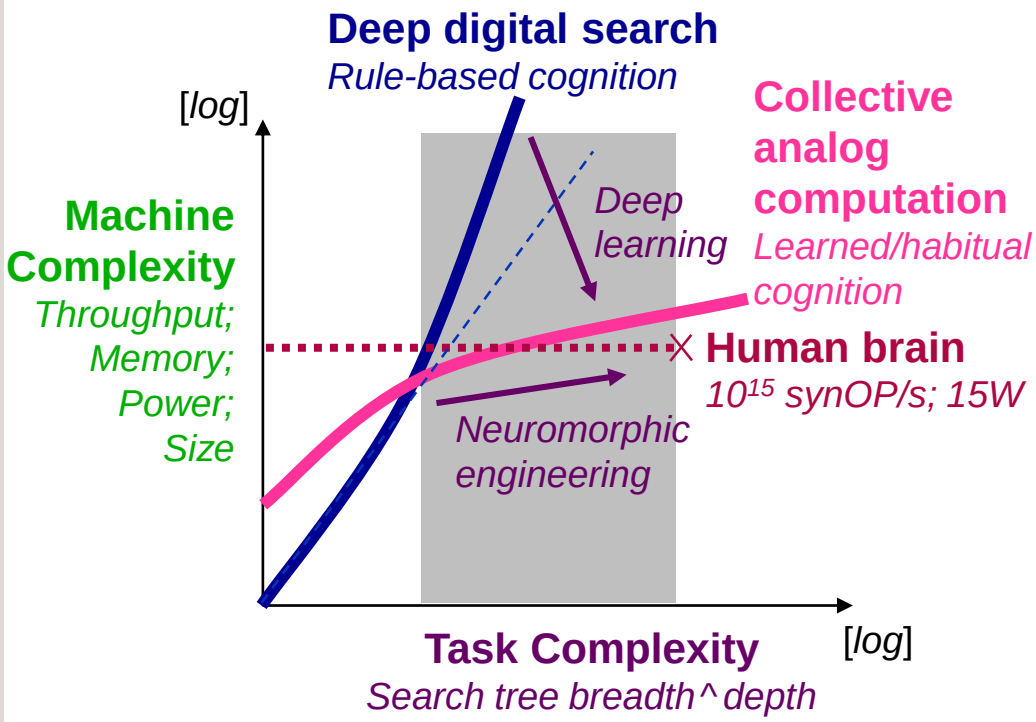
Carver Mead



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.

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

Scaling of Task and Machine Complexity

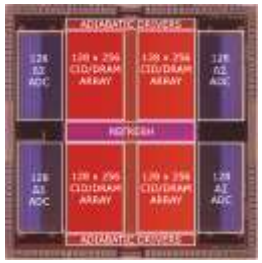


G. Cauwenberghs, "Reverse Engineering the Cognitive Brain," *PNAS*, 2013

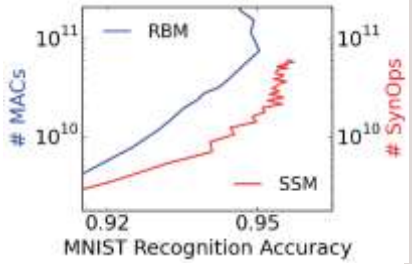
Task Energy Efficiency:

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

$$\frac{1\text{ fJ}}{\text{SynOp}} \text{ vs. } \frac{10\text{ pJ}}{\text{MAC}} \qquad \frac{10^{10}\text{ SynOps vs. } 10^{11}\text{ MACs}}{\text{MNIST @ 95\%}}$$



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:**
  - 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*

EE NanoE Phys

  - 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*

Neuro CS CogSci

Large-Scale Reconfigurable Neuromorphic Computing

Technology and Performance Metrics

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

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

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.

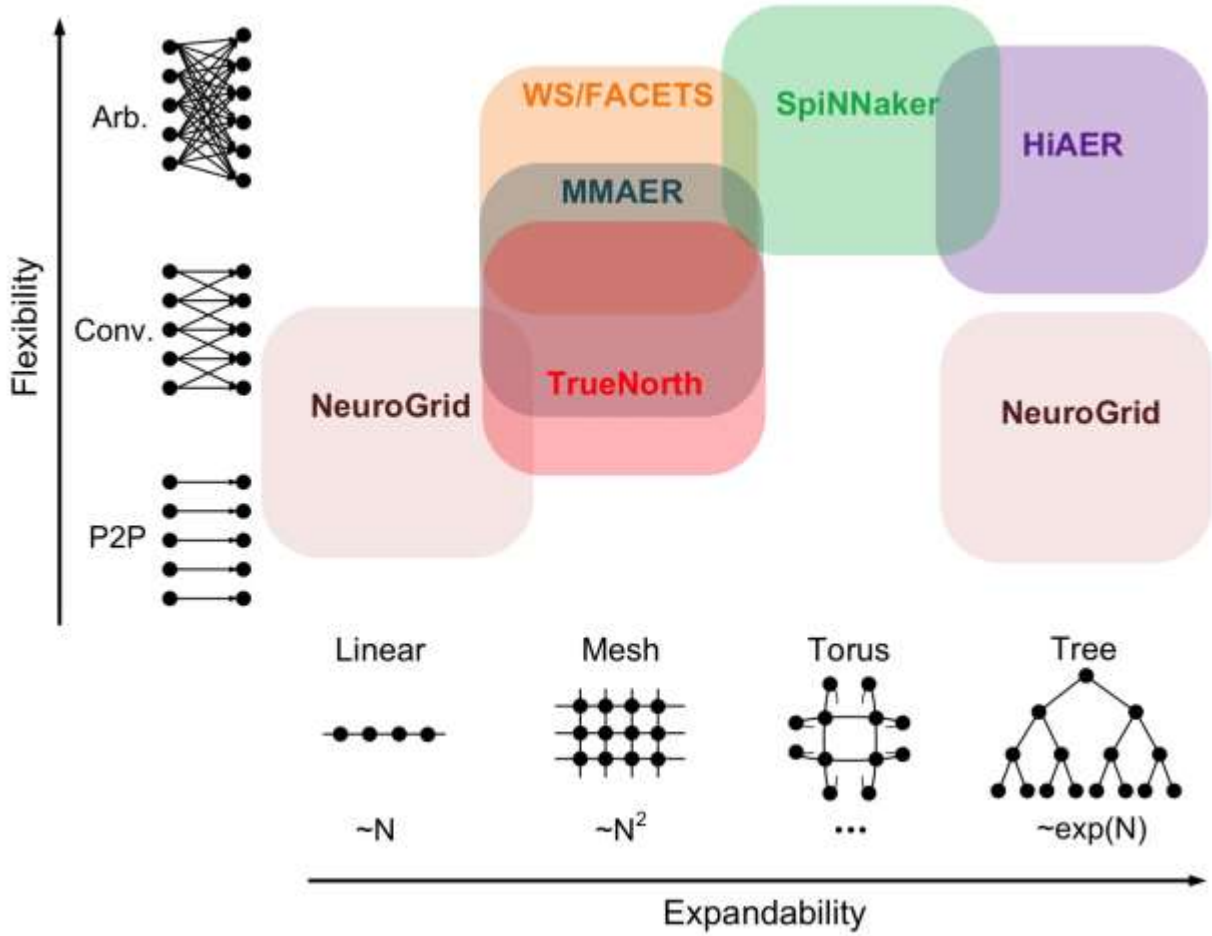
Merolla, P.A., J.V. Arthur, R. Alvarez-Icaza, A S. Cassidy, J. Sawada, F. Akopyan, B.L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S.K. Esser, R. Appuswamy, B. Taba, A. Amir, M.D. Flickner, W.P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable 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.

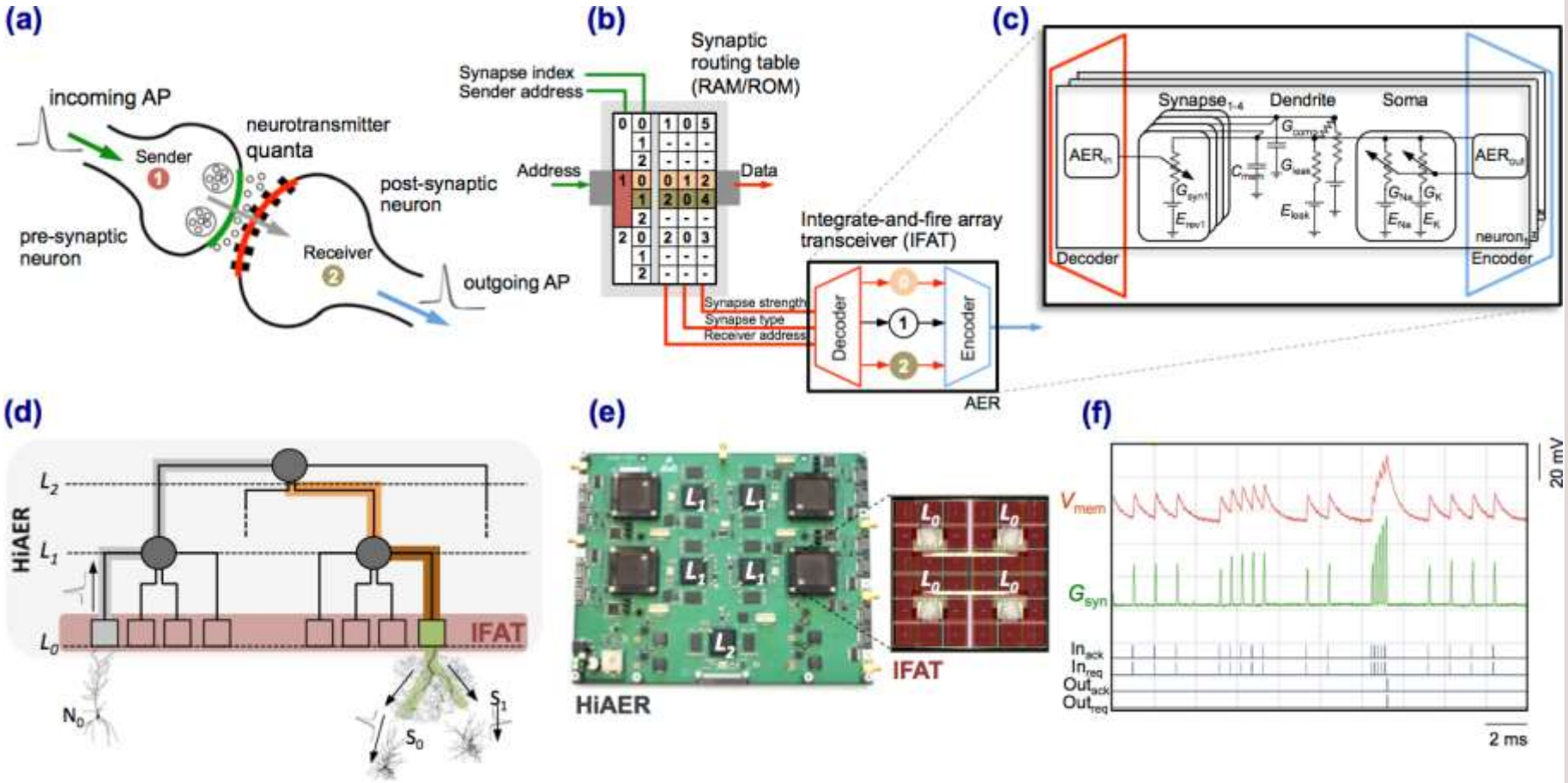
Schemmel, J., D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner, "A waferscale neuromorphic hardware system for large-scale neural modeling," *Proc. 2010 IEEE Int. Symp. Circuits and Systems (ISCAS)*, 1947–1950, 2010.

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



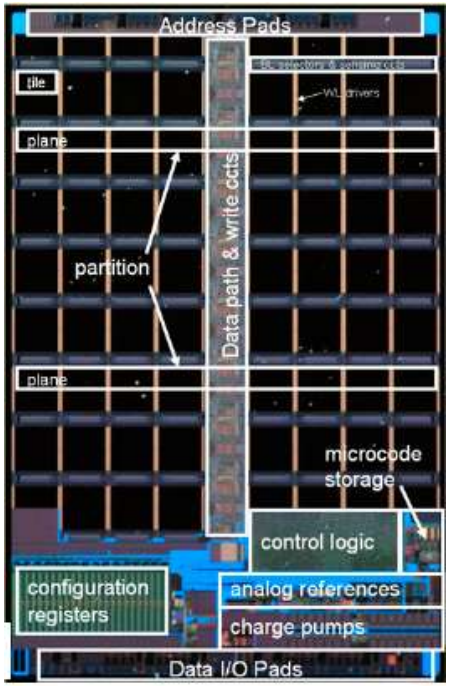
Comparison of synaptic connection topologies for several recent large-scale event-driven neuromorphic systems and the proposed hierarchical address-event routing (HiAER), represented diagrammatically in two characteristic dimensions of connectivity: expandability (or extent of global reach), and flexibility (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.

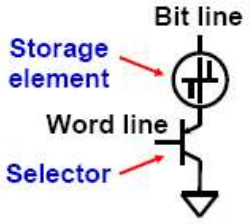


Hierarchical Address-Event Routing (HiAER) Integrate-and-Fire Array Transceiver (IFAT) for scalable and reconfigurable neuromorphic neocortical processing. (a) Biophysical model of neural and synaptic dynamics. (b) Dynamically reconfigurable synaptic connectivity is implemented across IFAT arrays of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

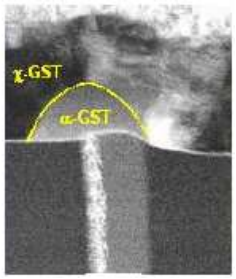
Yu et al, BioCAS 2012; Park et al, BioCAS 2014; Park et al, TNNLS 2017; Broccard et al, JNE 2017

Phase Change Memory (PCM) Nanotechnology

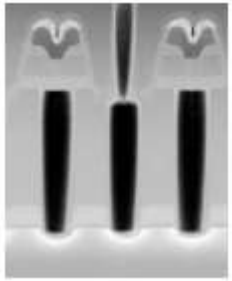




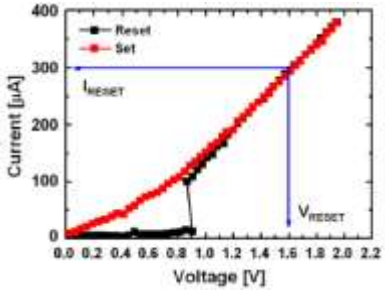
(a)



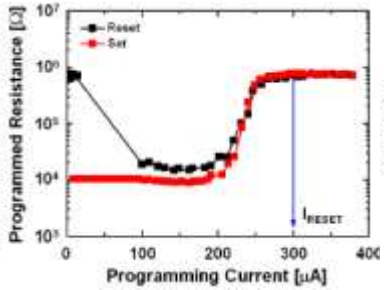
(b)



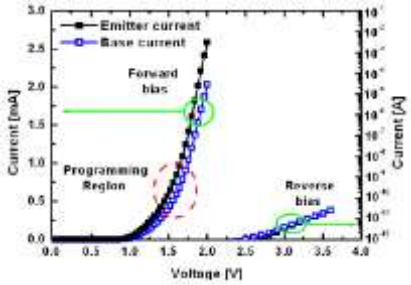
(c)



(d)



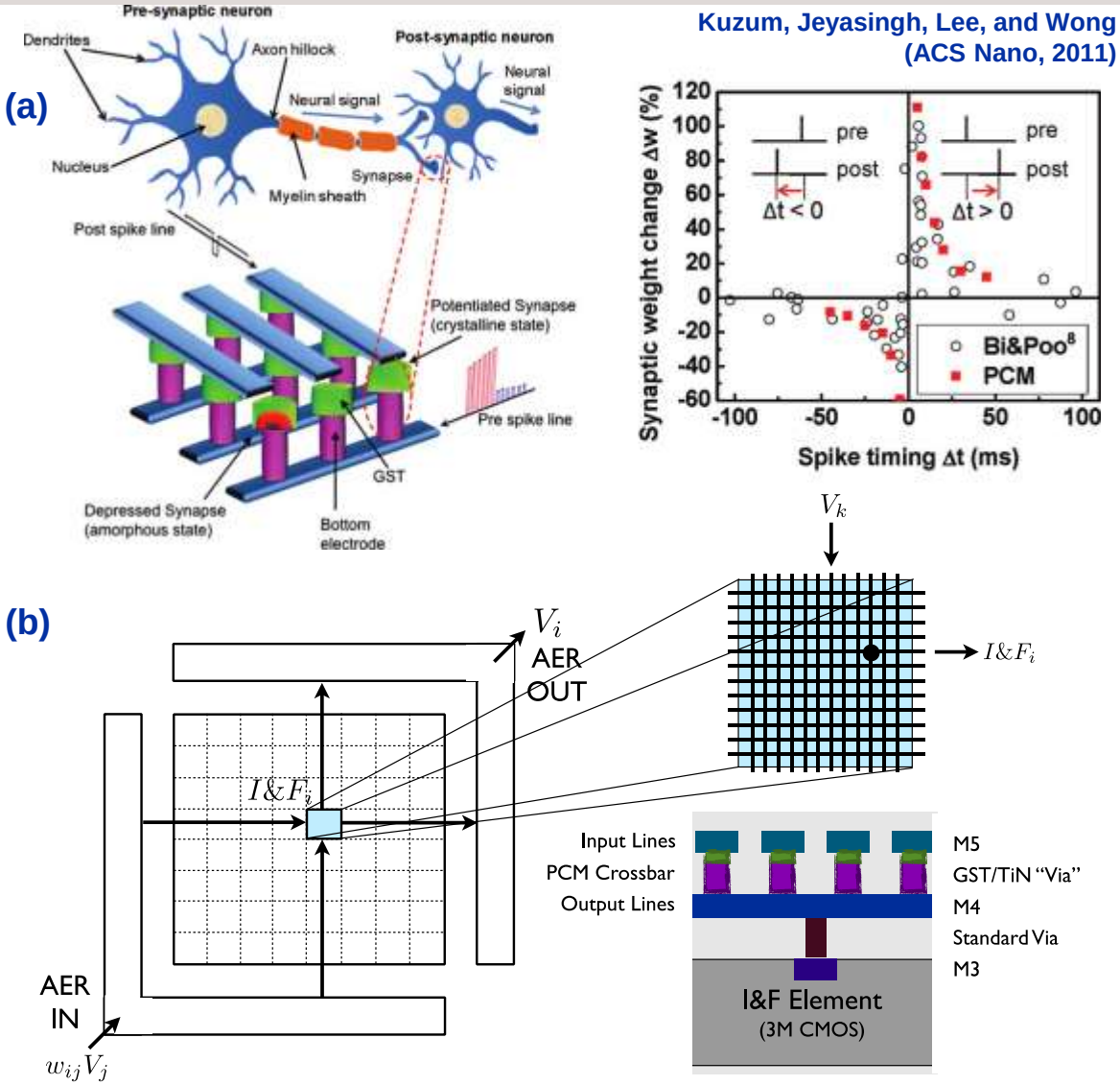
(e)



(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 32nm CMOS*
  - *< pJ energy per synapse operation*

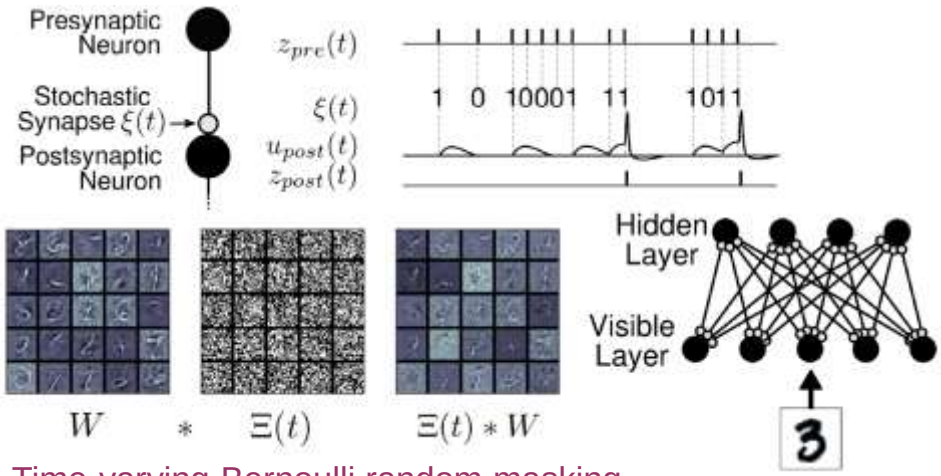


Hybridization and nanoscale integration of CMOS neural arrays with phase change memory (PCM) synapse crossbar arrays. (a) Nanoelectronic PCM synapse with spike-timing dependent plasticity (STDP) [Kuzum *et al*, 2011]. Each PCM element implements a synapse with conductance modulated through phase transition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

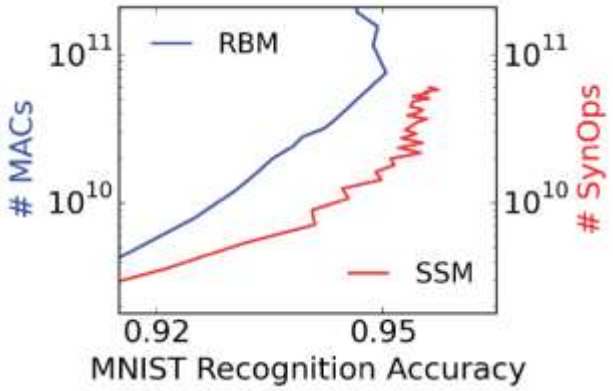
Spiking Synaptic Sampling Machine (S^3M)

Biophysical Synaptic Stochasticity in Inference and Learning

Synaptic stochasticity as biophysical model of continuous *DropConnect*



Time-varying Bernoulli random masking of weights

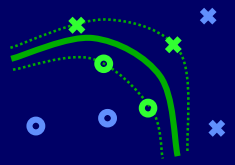


The S^3M requires fewer synaptic operations (SynOps) than the equivalent Restricted Boltzmann Machine (RBM) requires multiply-accumulate (MAC) operations at the same accuracy.

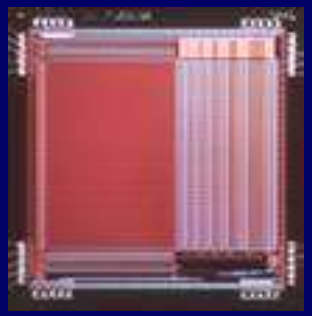
- Stochastic synapses for spike-based Monte Carlo sampling
  - *Models biophysical origins of noise in neural systems*
  - *Activity dependent noise: multiplicative synaptic sampling rather than additive neural sampling*
  - *Sparsity in neural activity and in synaptic connectivity*
- Online unsupervised learning with STDP
  - *Biophysical model of spike-based learning*
  - *Event-driven contrastive divergence*

Emre O. Neftci, Bruno U. Pedroni, Siddharth Joshi, Maruan Al-Shedivat, Gert Cauwenberghs, "Stochastic Synapses Enable Efficient Brain-Inspired Learning Machines," *Frontiers in Neuroscience*, vol. 10, pp. 3389:1-16 (DOI: 10.3389/fnins.2016.00241), 2016.

Silicon Learning Machines for Embedded Sensor Adaptive Intelligence

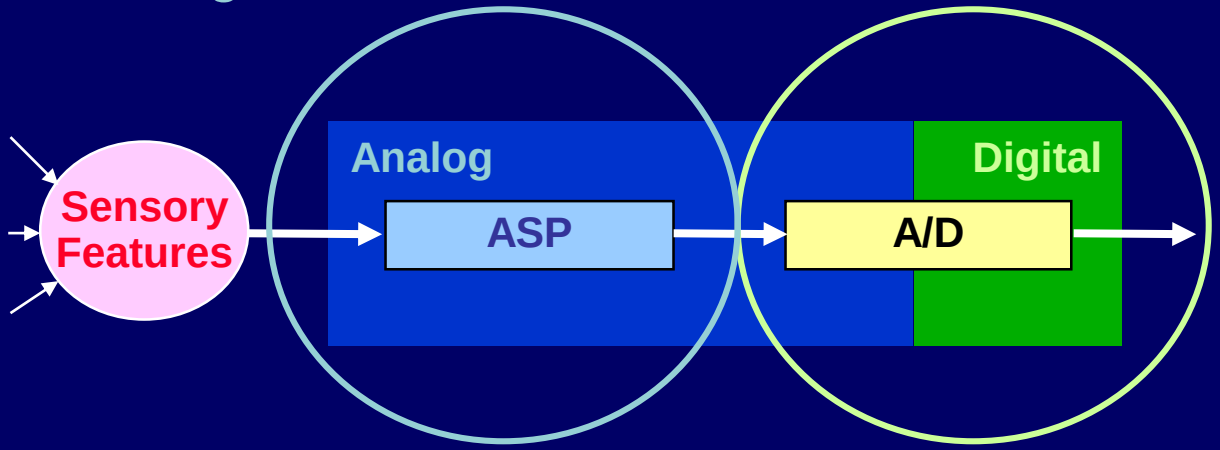


Large-Margin Kernel Regression

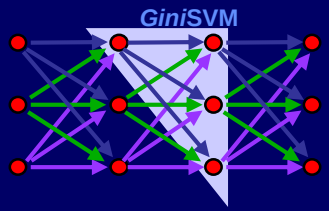


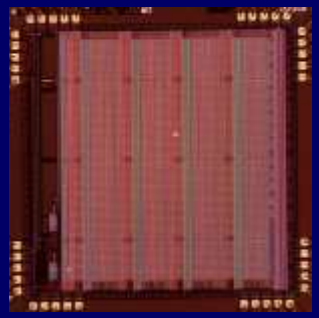
Kerneltron: massively parallel support vector "machine" (SVM) in silicon (*JSSC 2007*)

Class Identification



MAP Forward Decoding



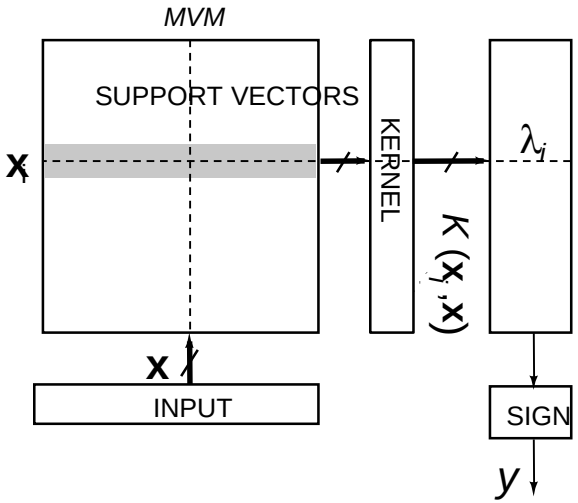


Sequence Identification

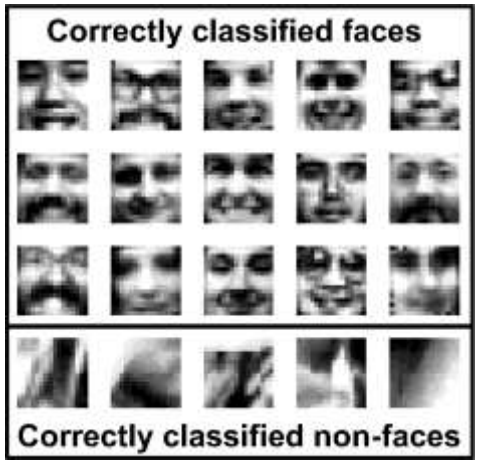
Sub-microwatt speaker verification and phoneme recognition (*NIPS'2004*)

Kerneltron: Adiabatic Support Vector "Machine"

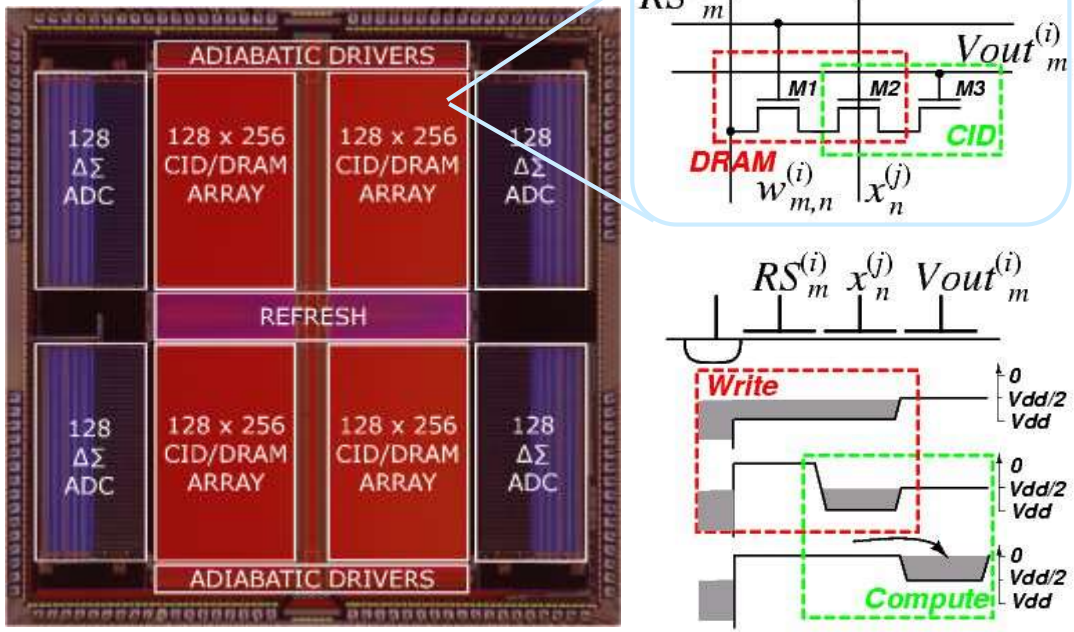
Karakiewicz, Genov and Cauwenberghs , 2007



$$y = \text{sign}\left(\sum_{i \in S} \lambda_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right)$$

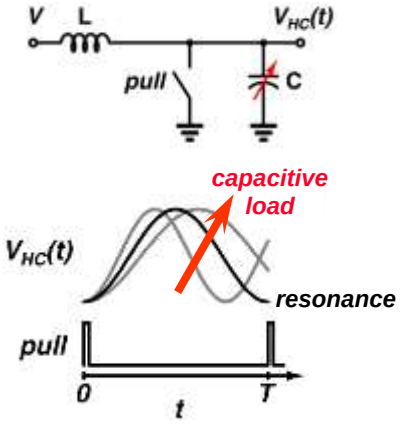


Classification results on MIT CBCL face detection data

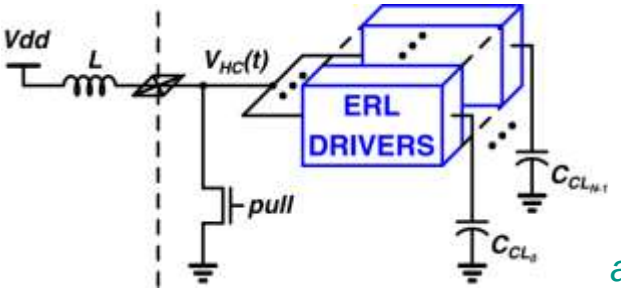


Karakiewicz, Genov, and Cauwenberghs, VLSI' 2006; CICC' 2007

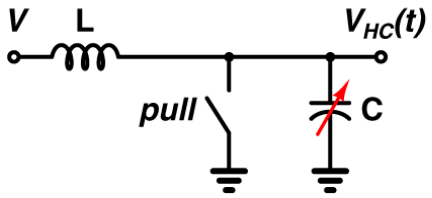
- ***1.2 TMACS / mW***
  - *adiabatic resonant clocking conserves charge energy*
  - *energy efficiency on par with human brain (10^{15} SynOP/S at 15W)*



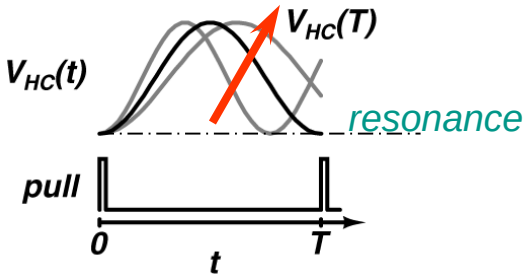
Resonant Charge Energy Recovery

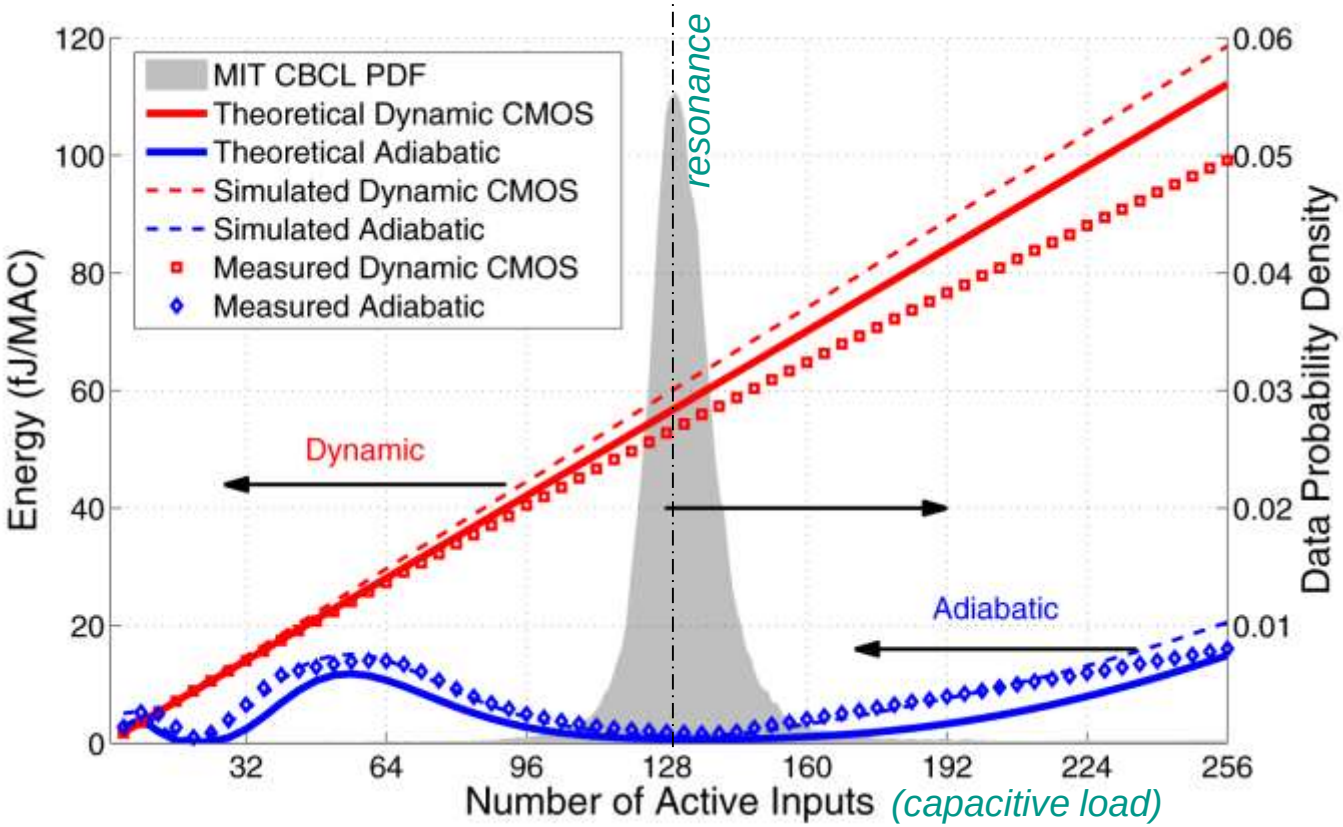


CID array

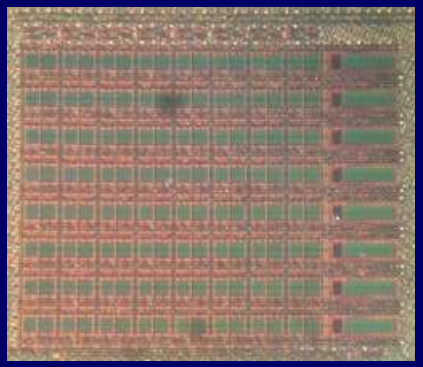


capacitive load



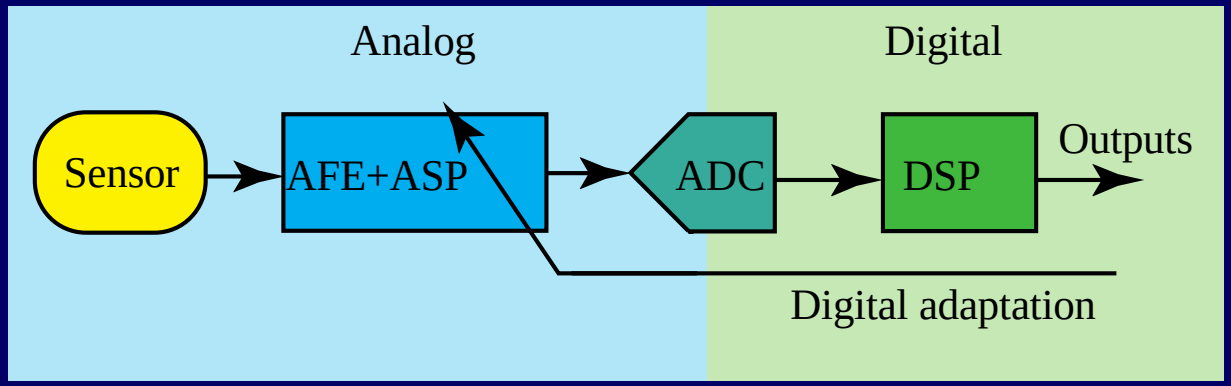


Adaptive Low-Power Sensory Systems



2pJ/MAC 14b 8×8 Linear Transform MixedSignal Spatial Filter in 65nm CMOS with 84dB Interference Suppression

S. Joshi et al, ISSCC 2017

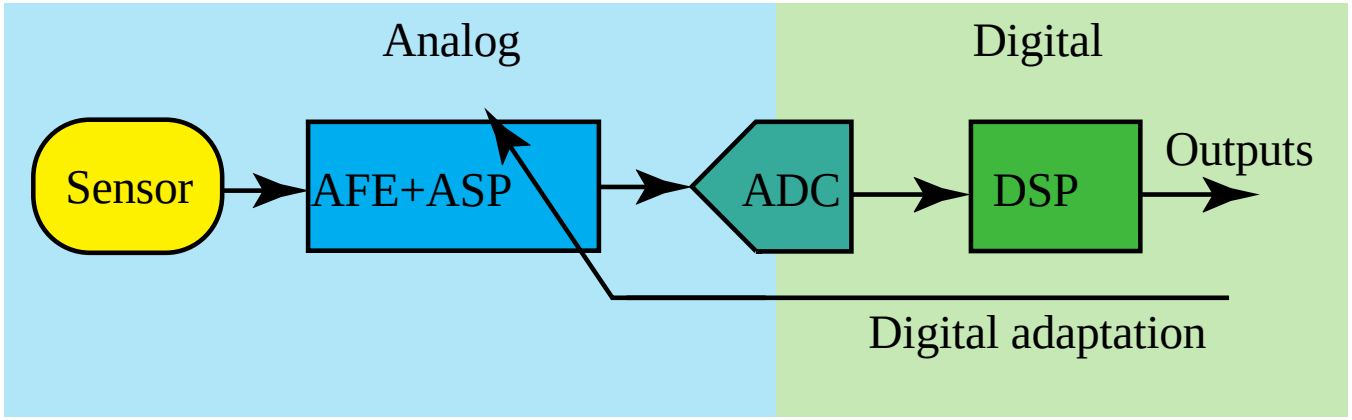


Charge-domain Analog Signal Processing

Low-dimensional, Low-resolution Digital Coding

Digital Adaptation

Linear Transform Analog and Mixed-Signal Sensory Processing



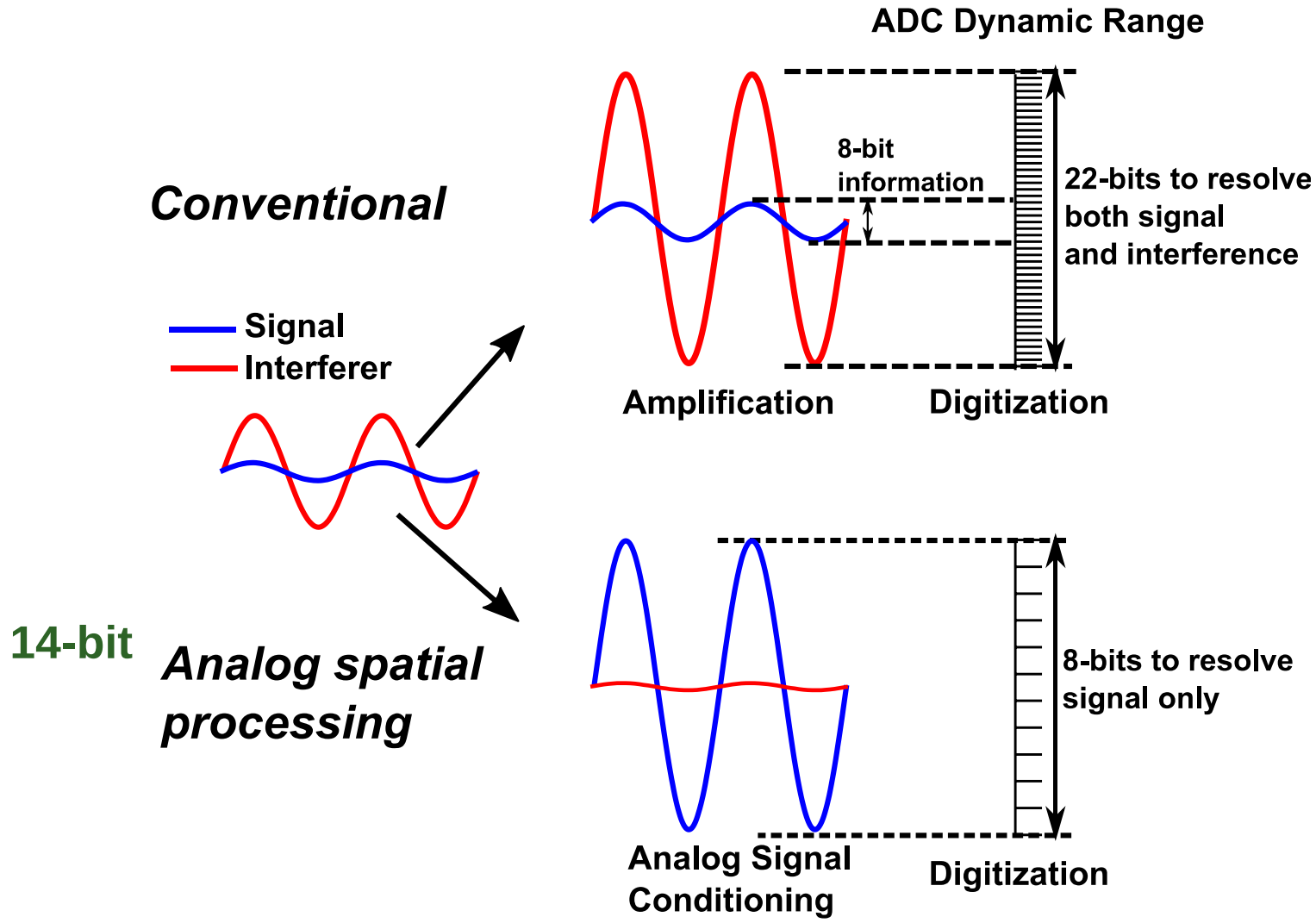
- Application Enabler
- Lower Power
- Analog ***processing gain*** lowers A/D requirements

Processing gain: Improvement in SNR/DR due to ASP

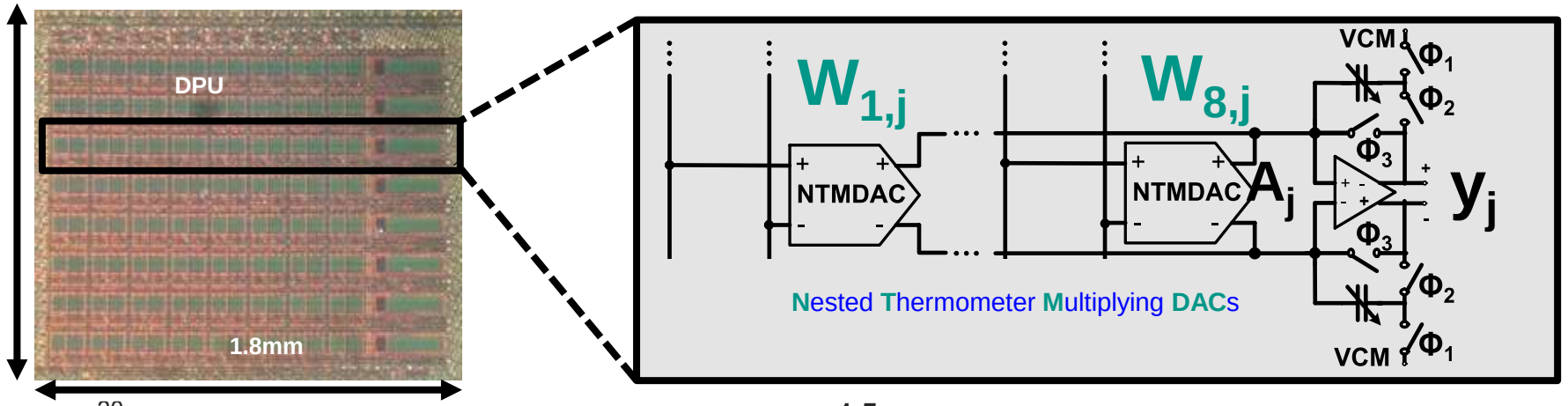
S. Joshi et al, "2pJ/MAC 14b 8 × 8 Linear Transform MixedSignal Spatial Filter in 65nm CMOS with 84dB Interference Suppression," ISSCC 2017

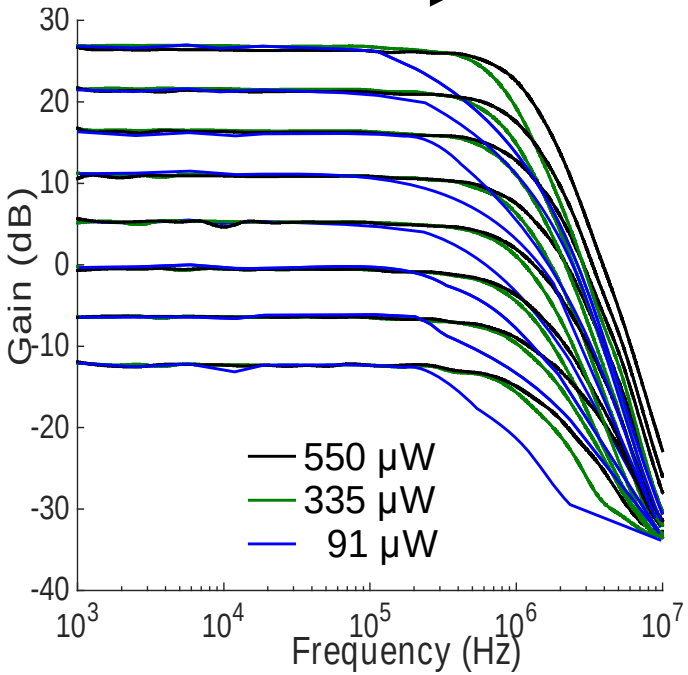
Spatial Processing Gain

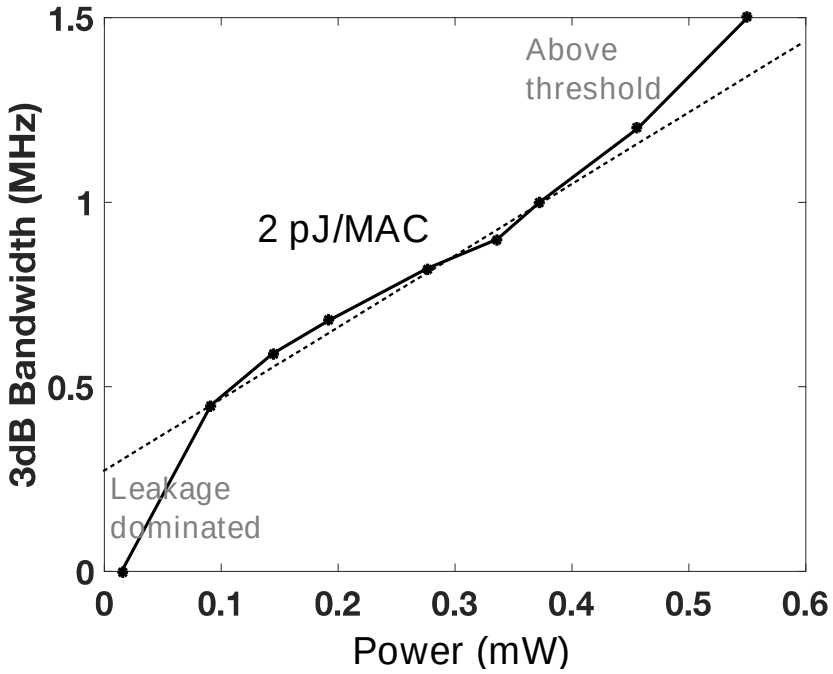
Improvement in SNR/DR due to ASP



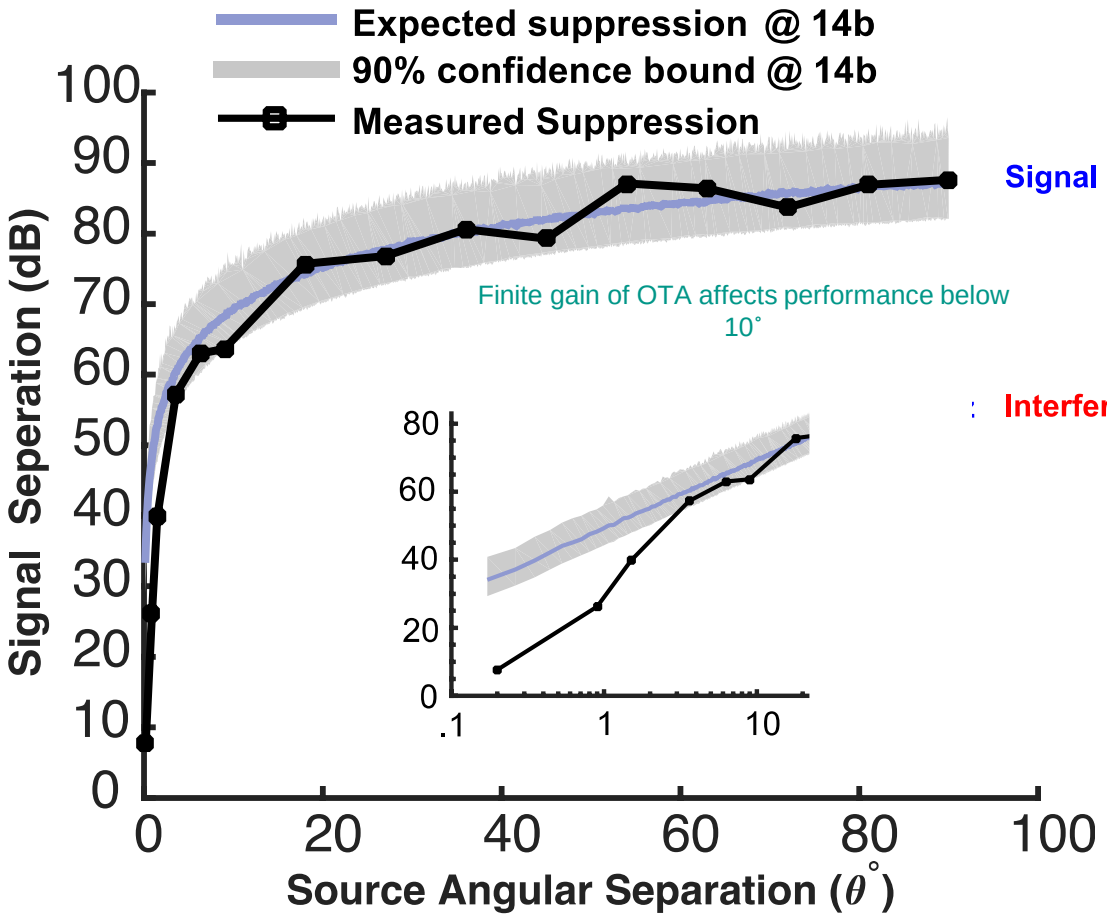
System Measurements



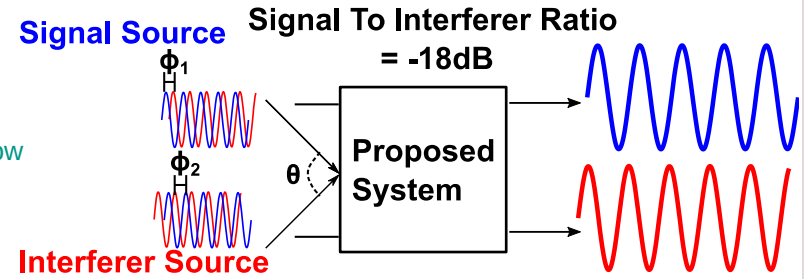




Measurements: Angular Resolution

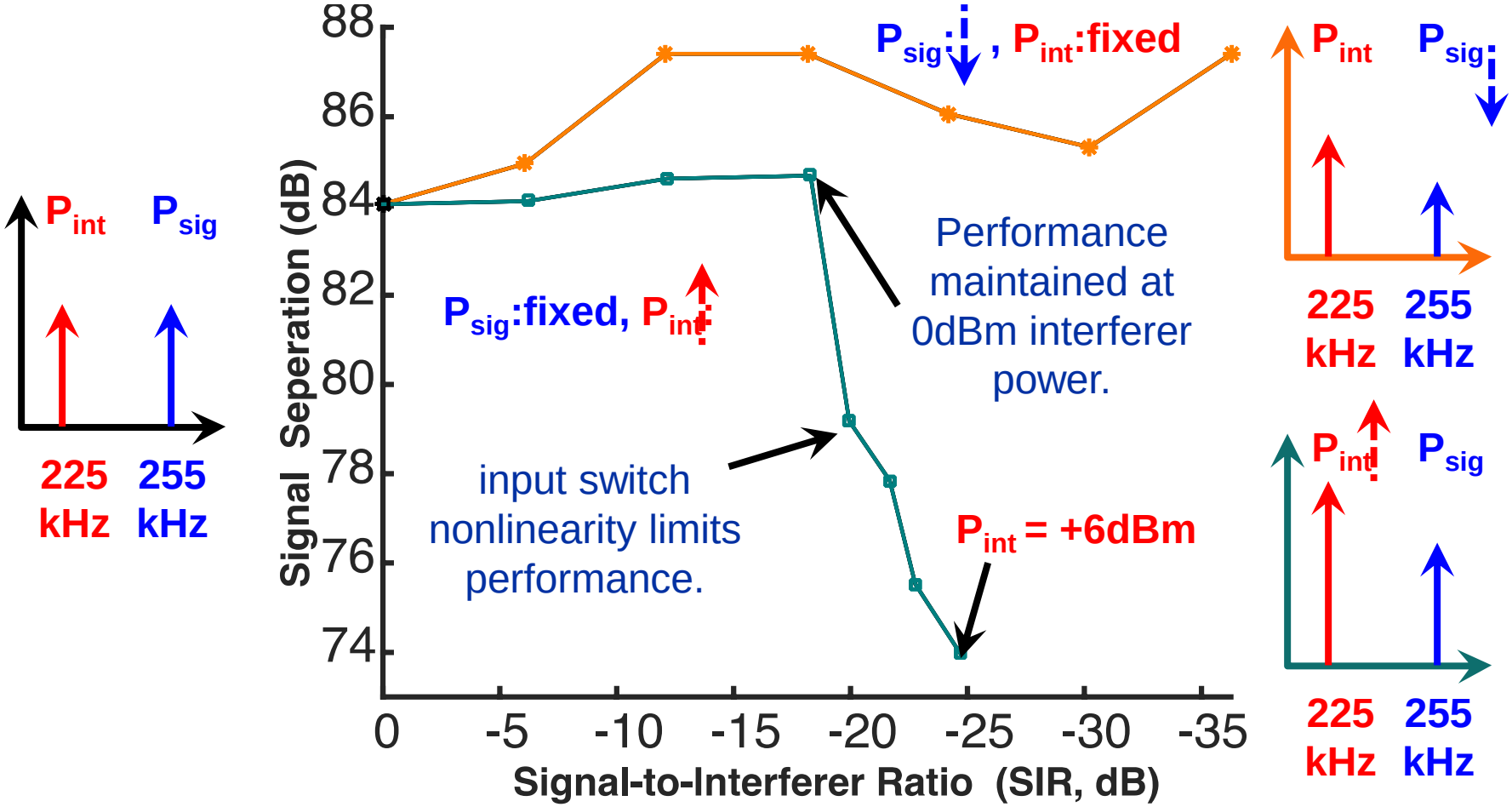


Finite gain of OTA affects performance below 10˚



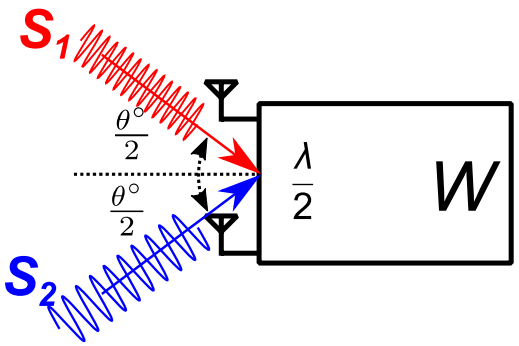
Experimental setup.

Measurements: SIR

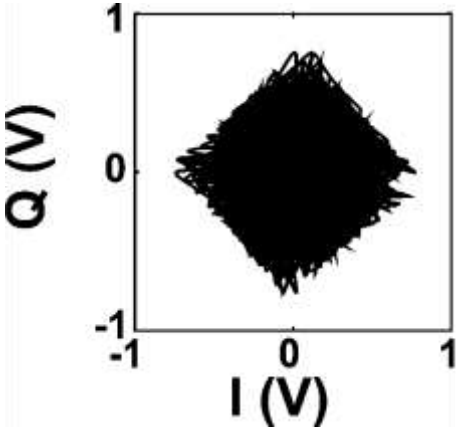


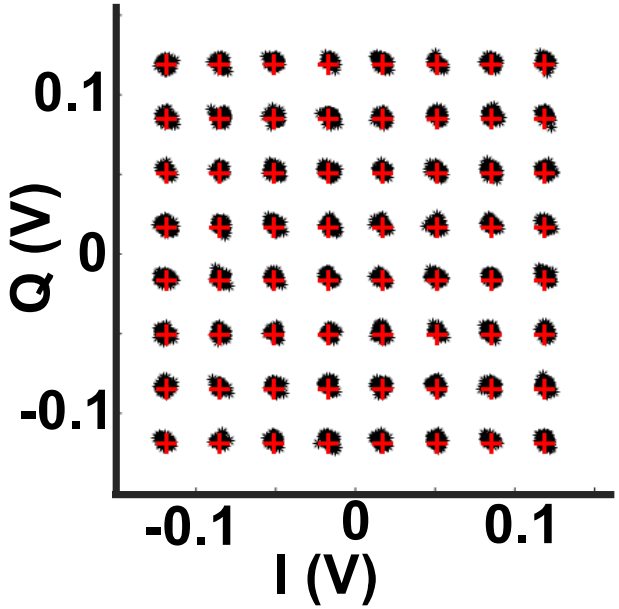
Application: MIMO Communication

Spatial filtering to separate signal mixture

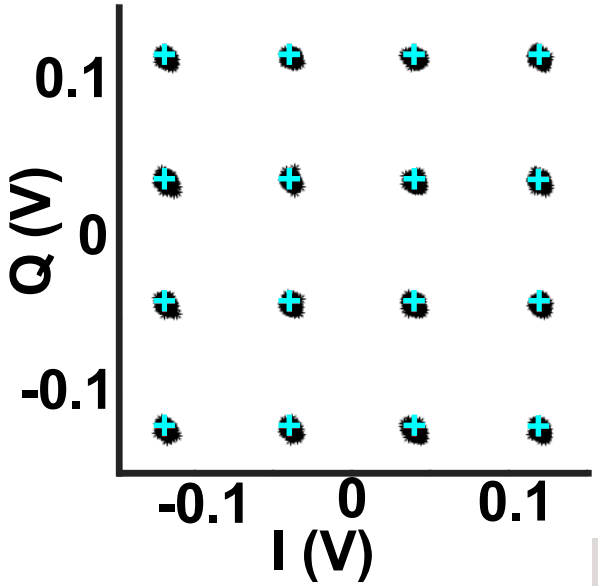


Constellation





64-QAM resolved
RMS EVM 2.9%



16-QAM resolved
RMS EVM 3.1%

Application: MIMO Communication

Beamforming Performance (baseband only)

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

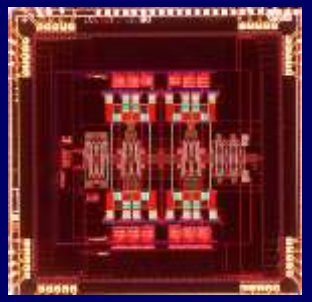
[a]Greater than 15 dB cancellation, [b]Cancellation at 45° angular separation, [c]Out of beam, [d]LO power only, [e]Total power reported baseband power not reported

S. Joshi et al, "2pJ/MAC 14b 8×8 Linear Transform MixedSignal Spatial Filter in 65nm CMOS with 84dB Interference Suppression," ISSCC 2017

Closing the Loop: Interactive Neural/Artificial Intelligence

Neuromorphic Engineering

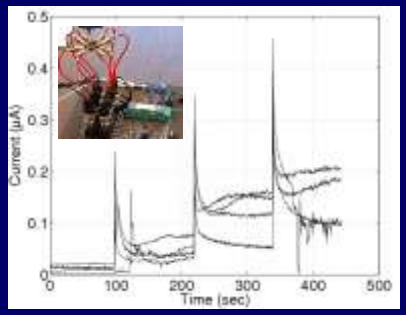
Adaptive Sensory Feature Extraction and Pattern Recognition



Neuro Bio

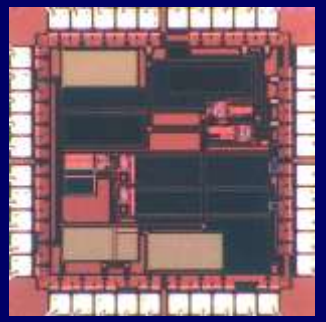
Learning & Adaptation

Micropower Mixed-Signal VLSI



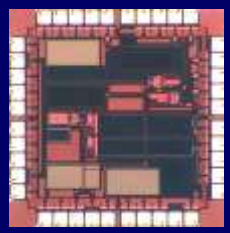
Neurosystems Engineering

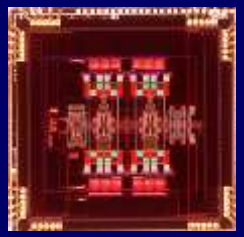
Biosensors, Neural Prostheses and Brain Interfaces



Integrated Systems Neuroengineering

Neuromorphic/ Neurosystems Engineering





Neural Systems

Learning & Adaptation

Silicon Microchips

Human/Bio Interaction

Environment

Sensors and Actuators