Scalable Neuromorphic Spike-Based Learning Systems

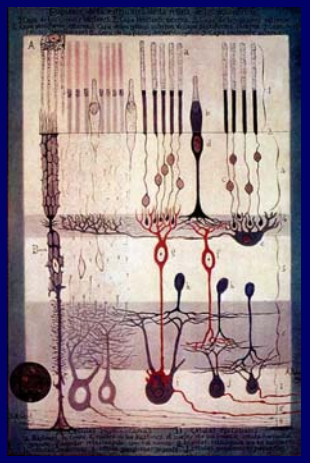
Gert Cauwenberghs

University of California San Diego

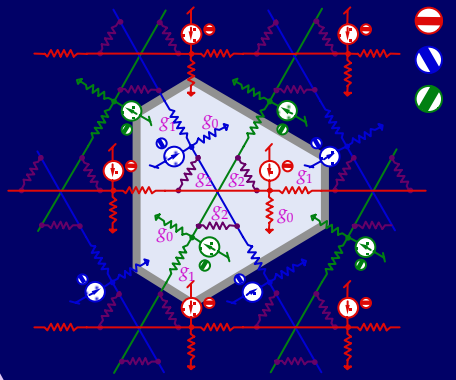
gert@ucsd.edu

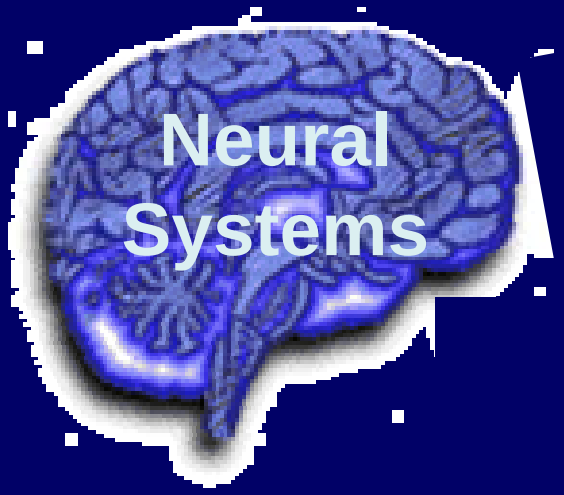
Neuromorphic Engineering

"in silico" neural systems design

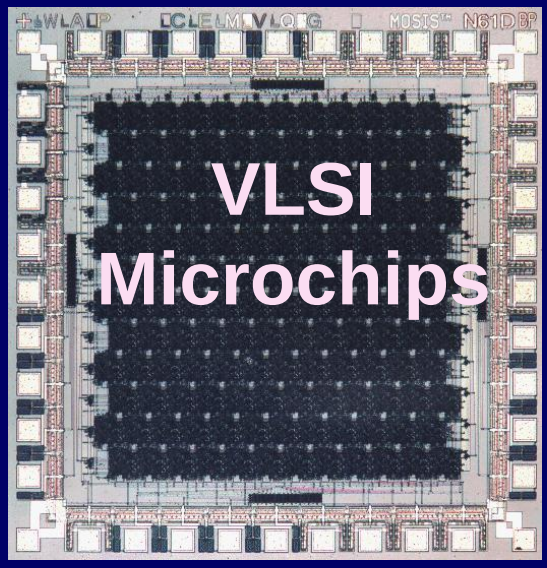


Neuromorphic Engineering



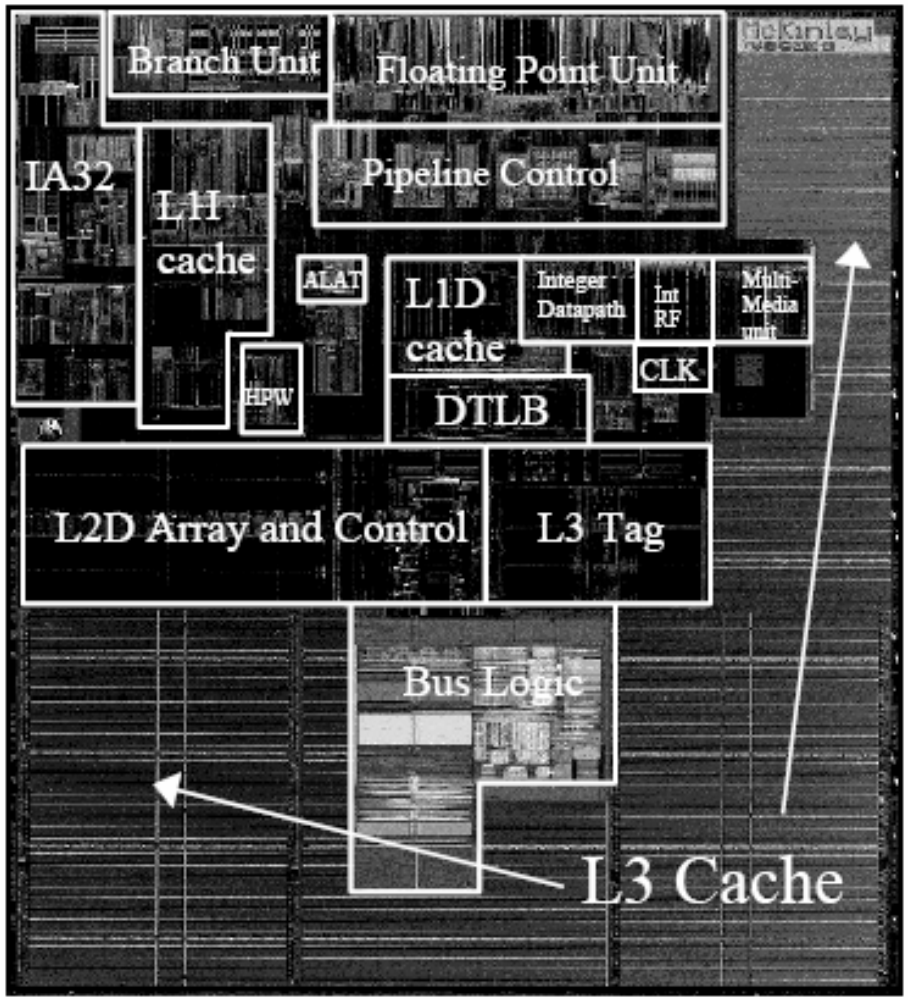


Learning & Adaptation



Today's Hottest Microchip

Intel's Itanium 2



Source: IEEE ISSCC'2002

The numbers …

- 0.5 billion transistors in 120nm CMOS
- 1.6GHz clock, 64-bit instruction, 9MB L3 cache, 6.4GB/s I/O
- 2553 SPECfp_base2000 (30% faster than 2.8GHz P4)
- 130 Watts

… and what they mean

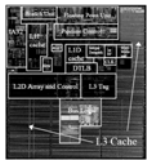
Faster/cooler:

- *Scientific computing*
- *Database search*
- *Web surfing*
- *Video games*

What about intelligence?

Chips and Brains

- **Itanium:**



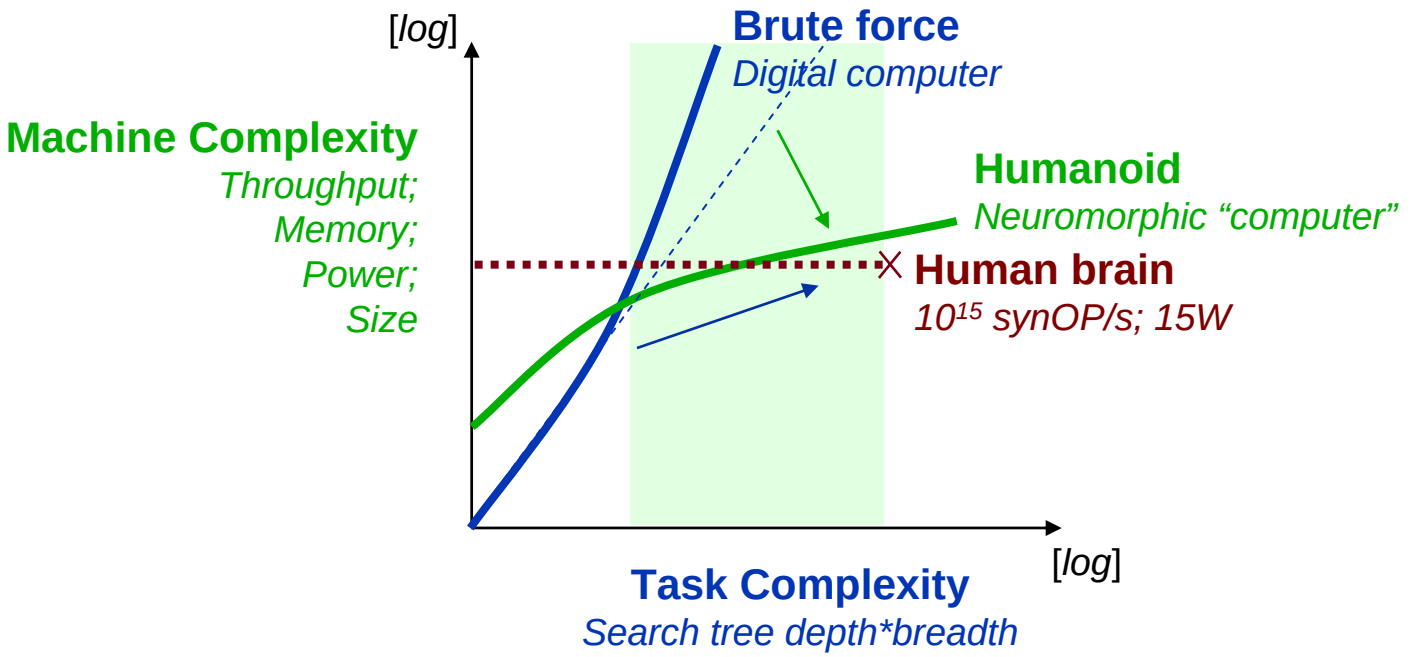
  - 3 10^9 floating op/s
    - *5 10^8 transistors*
    - *2 10^9 Hz clock*
  - 10^{10} Hz memory I/O
    - *128-b data bus @ 400MHz*
  - 130 Watts

- **Human brain:**

  - 10^{15} synaptic op/s
    - *10^{15} synapses*
    - *1 Hz average firing rate*
  - 10^{10} Hz sensory/motor I/O
    - *10^8 nerve fibers*
  - 25 Watts

- **Silicon technology is approaching the *raw* computational power and bandwidth of the human brain.**
- **However, to emulate brain intelligence with chips requires a radical paradigm shift in computation:**
  - Distributed representation in massively parallel architecture
    - *Local adaptation and memory*
    - *Sensor and motor interfaces*
  - Physical foundations of computing

Scaling of Task and Machine Complexity



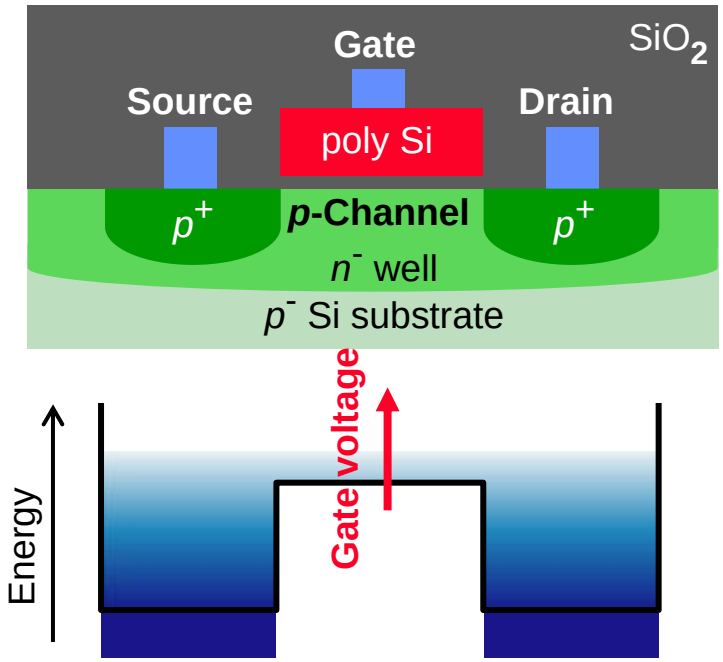
Achieving (or surpassing) human-level machine intelligence will require a convergence between:

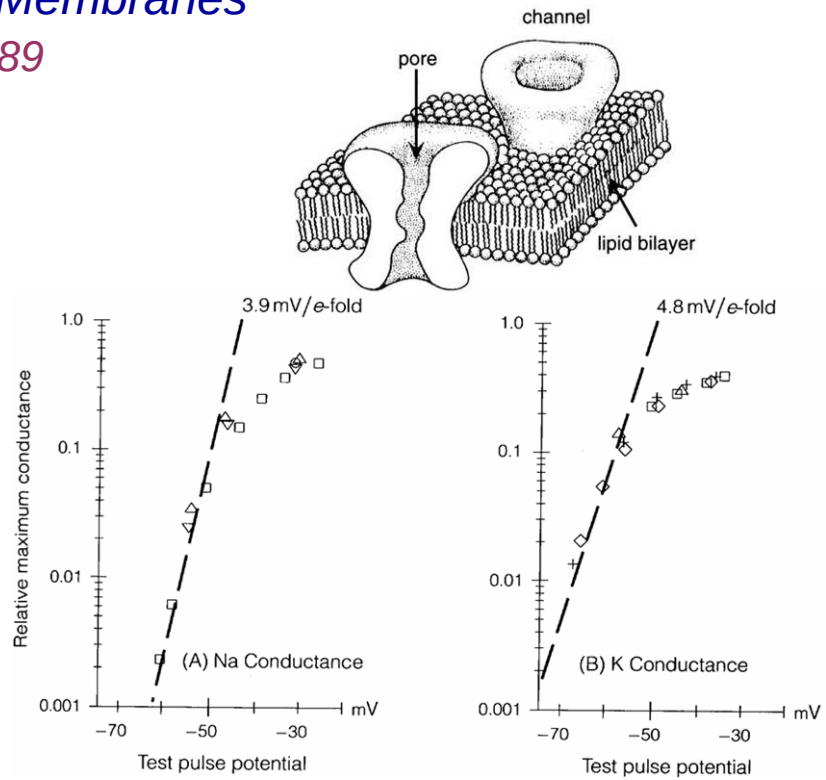
- *Advances in computing hardware yielding connectivity and energy efficiency levels of computing and communication in the brain;*
- *Advances in training methods, and supporting data, to adaptively reduce algorithmic complexity.*

Physics of Neural Computation

Silicon and Lipid Membranes

Mead, 1989





Squid giant axon (Hodgkin and Huxley, 1952)

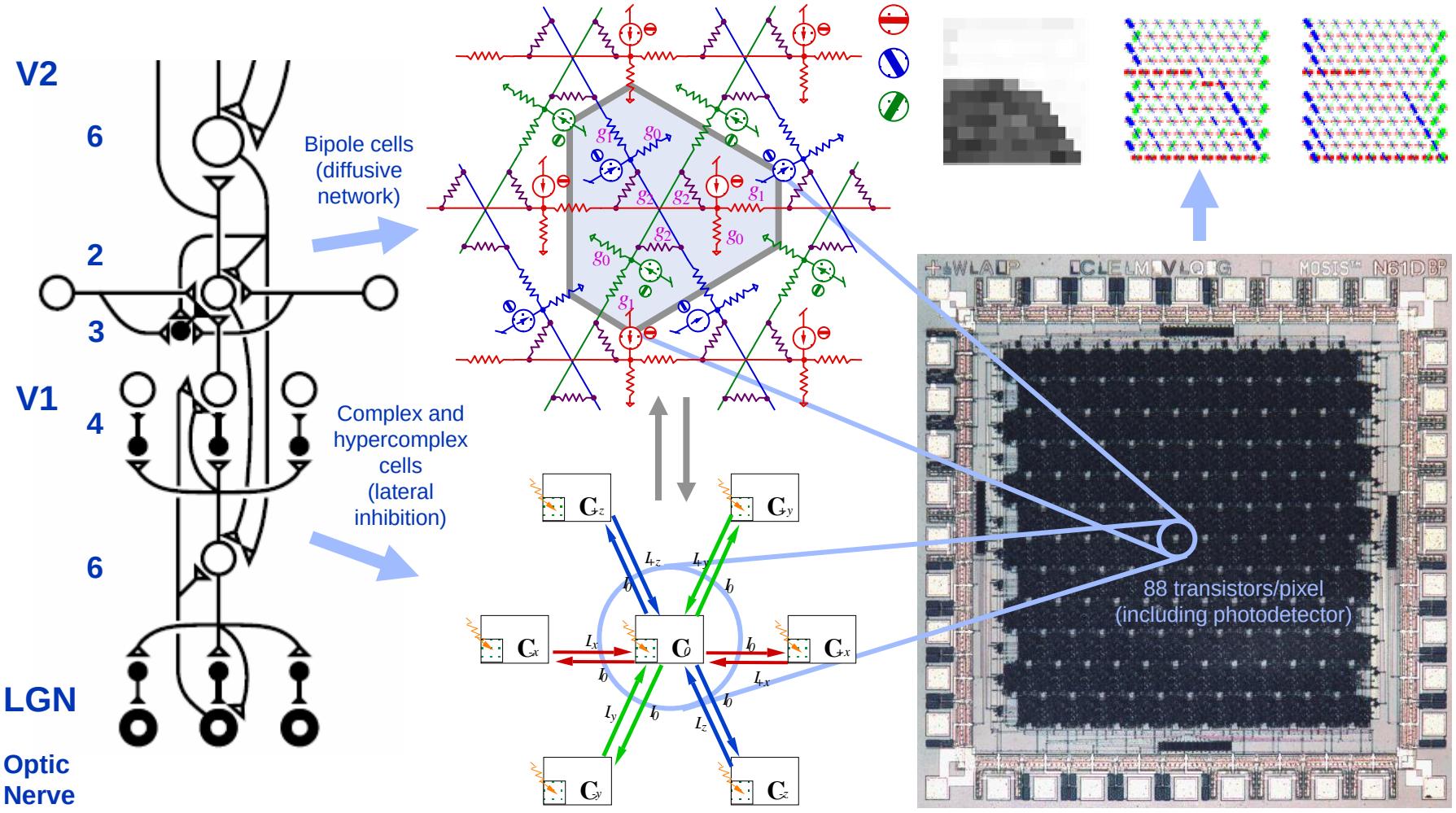
Voltage-dependent *p*-channel

- *Hole* transport between source and drain
- Gate controls energy barrier for holes across the channel
- Boltzmann distribution of *hole energy* produces exponential *decrease* in channel conductance with gate voltage

Voltage-dependent conductance

- K^+/Na^+ transport across lipid bilayer
- Membrane voltage controls energy barrier for opening of ion-selective channels
- Boltzmann distribution of *channel energy* produces exponential increase in K^+/Na^+ conductance with membrane voltage

Example: Silicon Model of Visual Cortical Processing

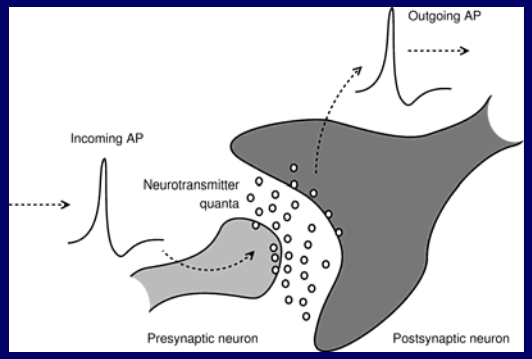


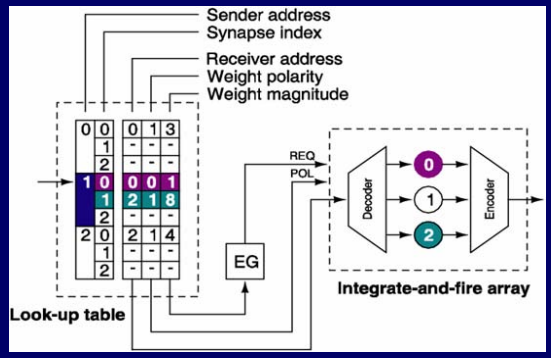
Neural model of boundary contour representation in V1, one orientation shown (Grossberg, Mingolla, and Williamson, 1997)

Single-chip focal-plane implementation (Cauwenberghs and Waskiewicz, 1999)

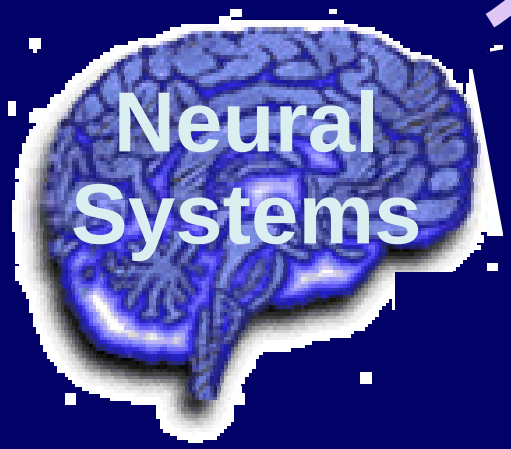
Reconfigurable Synaptic Connectivity and Plasticity

From Microchips to Large-Scale Neural Systems

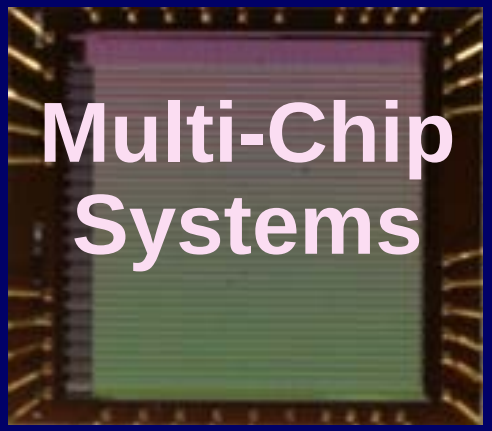




Address-Event Representation

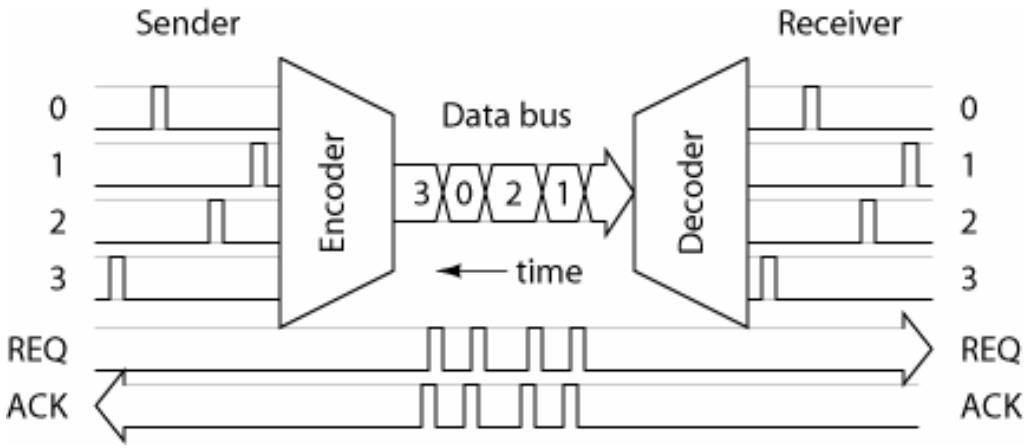


Synaptic Plasticity & Wiring



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"

Biochemical Synapse Mechanisms

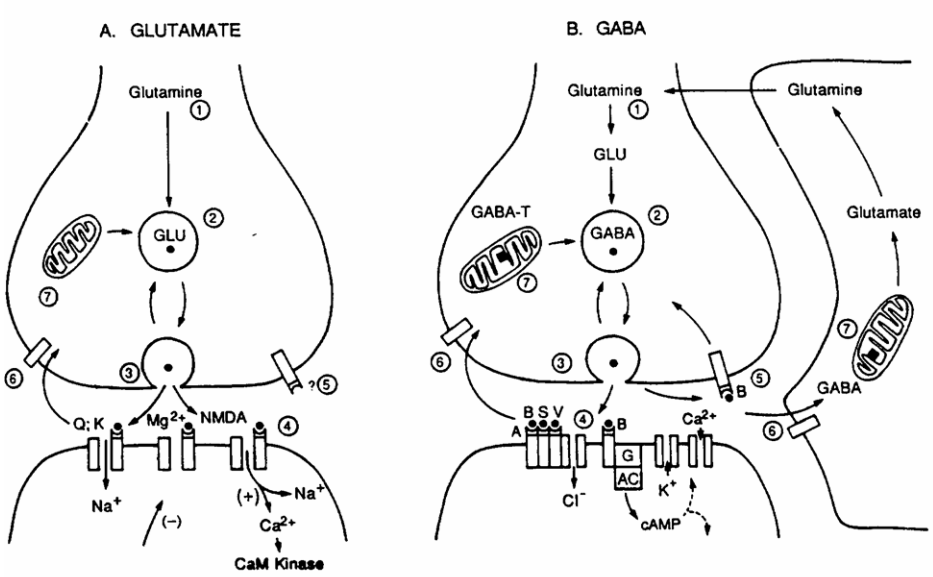
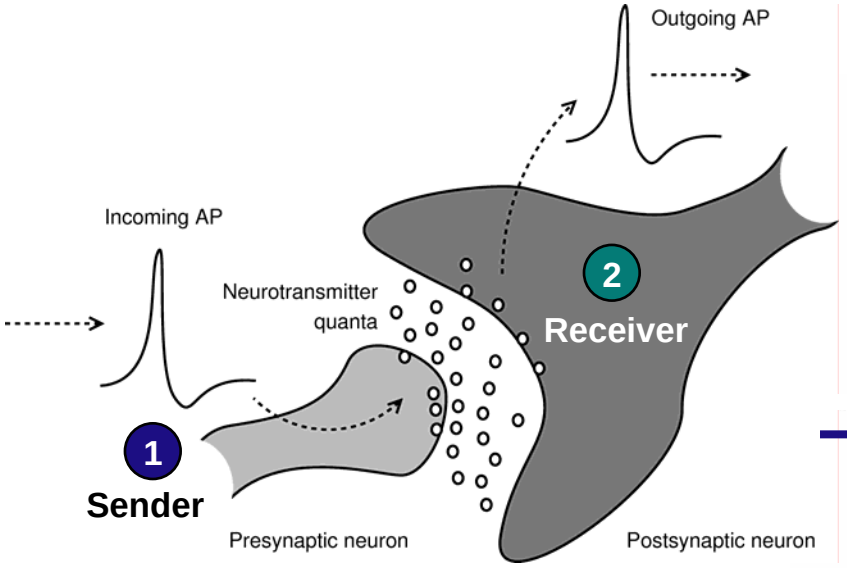


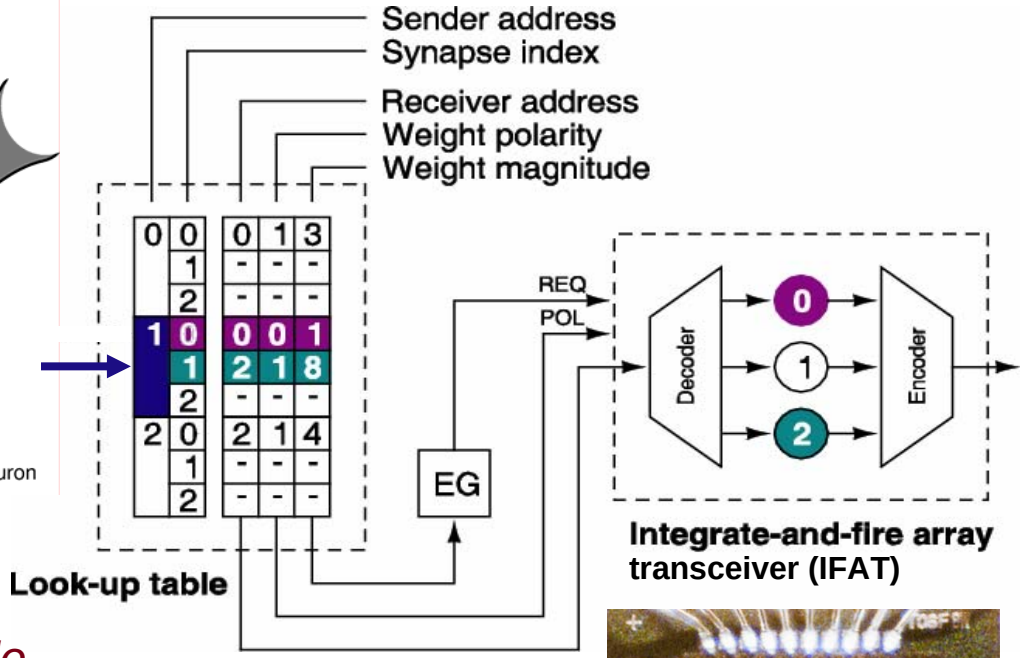
FIG. 2.8. Molecular mechanisms of amino acid synapses. **A.** Glutaminergic synapses: (1) synthesis of glutamate (GLU) from glutamine; (2) transport and storage; (3) release of GLU by exocytosis; (4) binding of GLU to quisqualate (Q), kainate (K), and NMDA receptors. The Q and K receptors gate Na^+ and K^+ flux; the NMDA receptor also allows Ca^{2+} entry when the membrane potential is depolarized (+). When the membrane potential is hyperpolarized (−), Mg^{2+} blocks this channel. The release of GLU may be regulated by presynaptic receptors (?5). Once GLU is released, it is removed from the synaptic cleft by reuptake (6) and processed intracellularly (7). **B.** GABAergic synapse: (1) synthesis of GABA from glutamine; (2) transport and storage of GABA; (3) release of GABA by exocytosis; (4) binding to a $GABA_A$ receptor which can be blocked by bicuculline (B), picrotoxin, or strychnine (S) and can also be modified by benzodiazepines, such as valium (V); $GABA_B$ receptors, by contrast, are linked via a G-protein to K^+ and Ca^{2+} channels which are blocked by GABA; (5) release of GABA is under the control of presynaptic $GABA_B$ receptors; GABA is removed from the synaptic cleft by uptake into terminals or glia (6); (7) processing of GABA back to glutamine. (A from Shepherd, 1988, based upon Cooper et al., 1987; Jahr and Stevens, 1987; Cull-Candy and Usowicz, 1987. B from Shepherd, 1988a; modified from Cooper et al., 1987; Aghajanian and Rasmussen, 1988; Nicoll, 1982.)

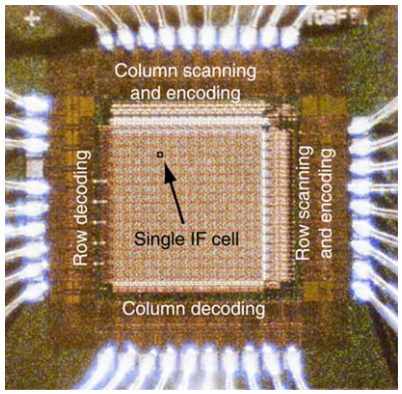
- **It is infeasible to implement networks of detailed synaptic models on a single silicon chip.**
- **Hybrid approach:**
  - Analog silicon chips model continuous-time membrane dynamics
  - Action potentials are encoded as 'address-events'.
  - A look-up table indexed by address-events implements synaptic connectivity and plasticity in the address domain

Address-Event Synaptic Connectivity

Goldberg, Cauwenberghs and Andreou, 2000







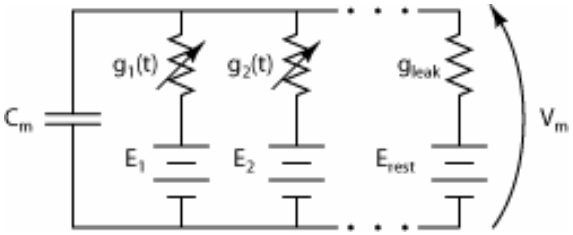
IFAT2 (2000)

- 'Virtual' synapses
  - *Dynamically reconfigurable*
  - *Arbitrary connectivity*
- Quantal release: $R = n\,p\,q$
  - *n: multiplicity* *(repeat event)*
  - *p: probability of release* *(toss a coin)*
  - *q: quantity released* *(set amplitude)*

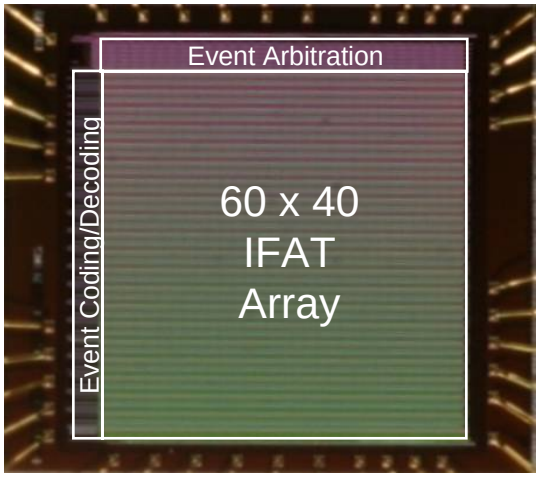
Silicon Membrane Array Transceiver

Vogelstein, Mallik and Cauwenberghs, 2004

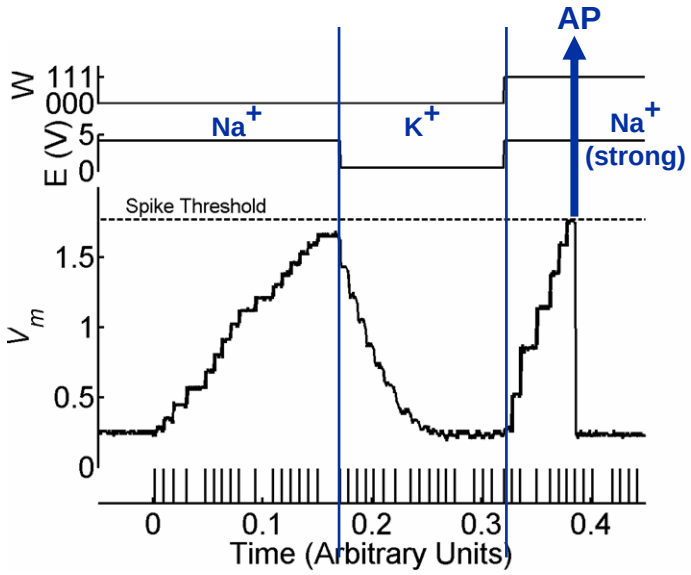
- Voltage-controlled membrane ion conductance
  - *Event-driven activation*
  - *Dynamically reconfigurable:*
    - *conductance g*
    - *driving potential E*



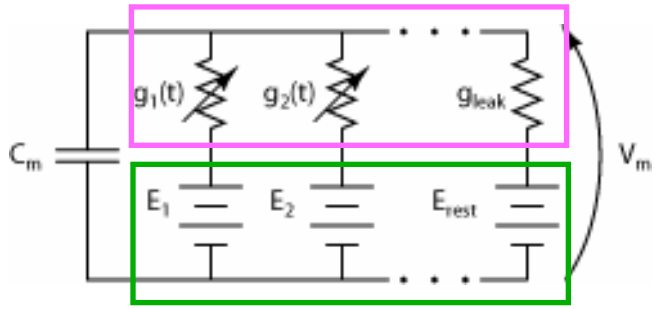
- Address-event encoding of pre-and post-synaptic action potentials



IFAT3 (2004)

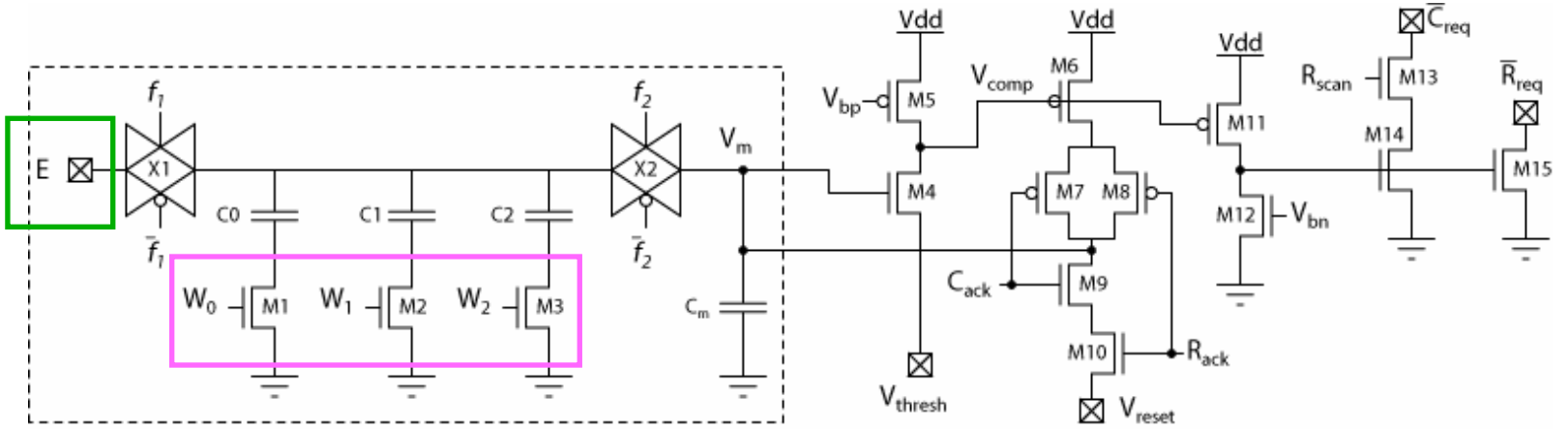


Silicon Membrane Circuit



$g_i(t)$ *ion-specific* membrane conductance

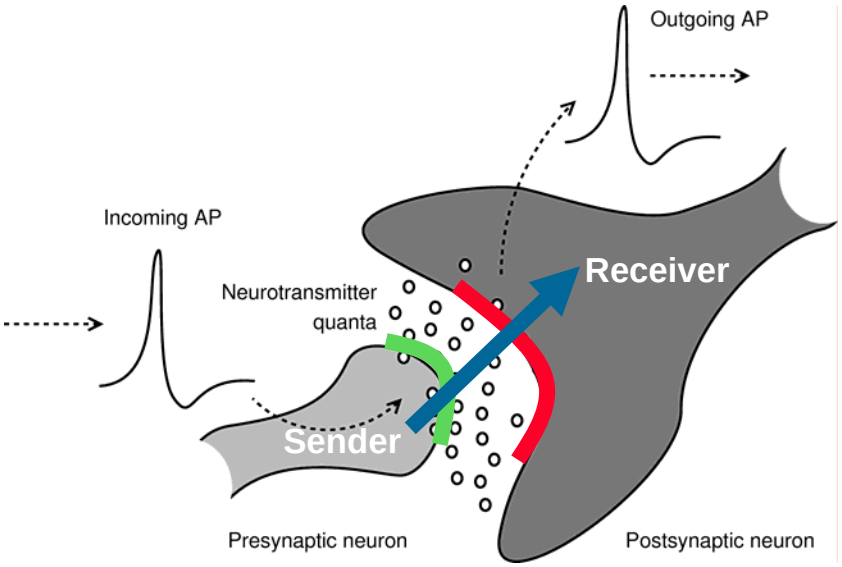
E_i *ion-specific* driving potential

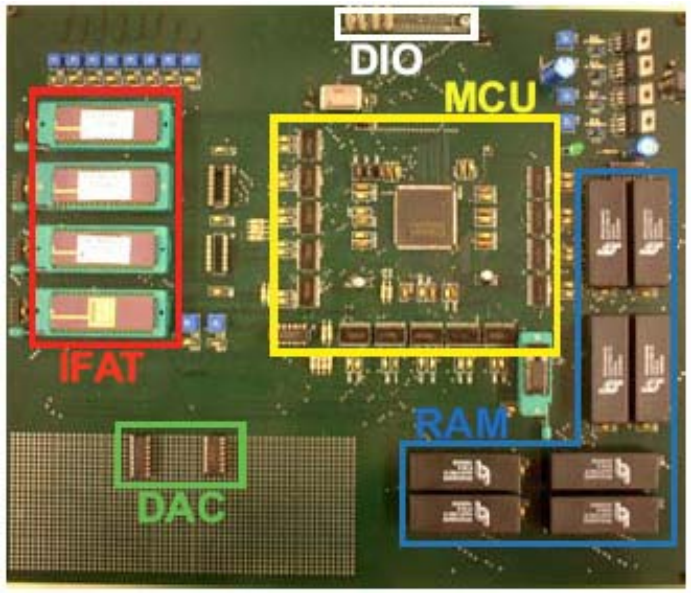


Synapse subcircuit

Action potential generation and AER handshaking

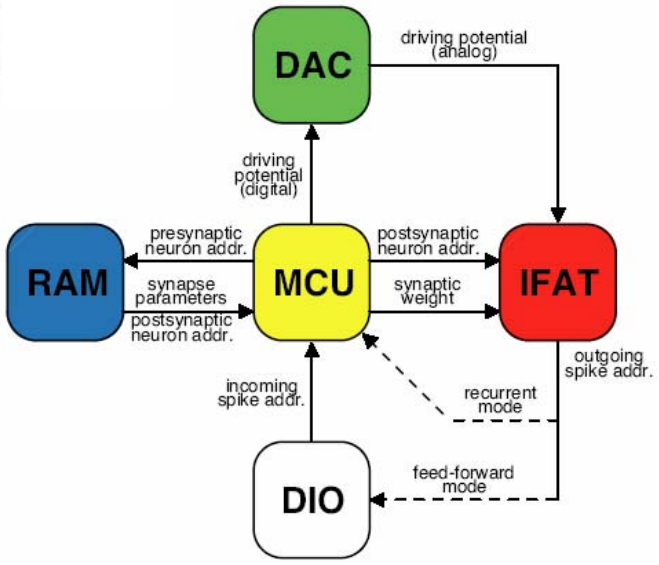
Reconfigurable Silicon Large-Scale Neural Emulator





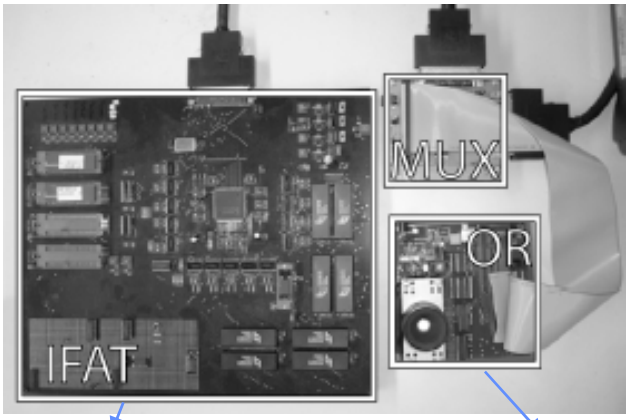
IFAT3 (Vogelstein, Mallik and Cauwenberghs, 2004)

- **9,600 neurons**
  - 4 silicon membrane chips (**IFAT**)
- **4 million, 8-bit "virtual" synapses**
  - 128MB (32bX4M) non-volatile **RAM**
- **1 million synaptic updates per second**
  - 200MHz Spartan II Xilinx FPGA "**MCU**"
- **Dynamically reconfigurable**
  - Rewiring and synaptic plasticity (STDP etc.)
  - Driving potential (**DAC**) and conductance (**IFAT**)



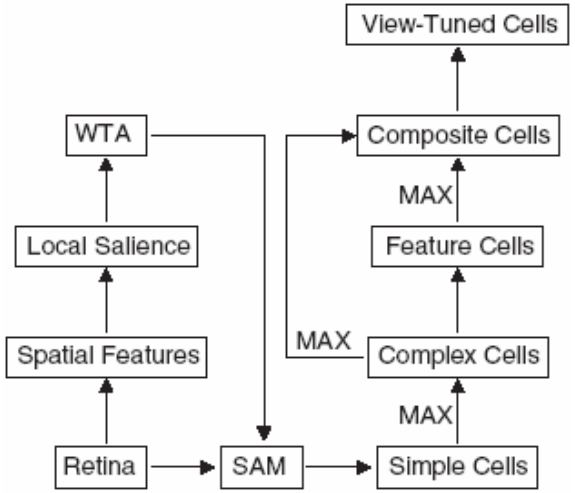
Hierarchical Vision and Saliency-Based Acuity Modulation

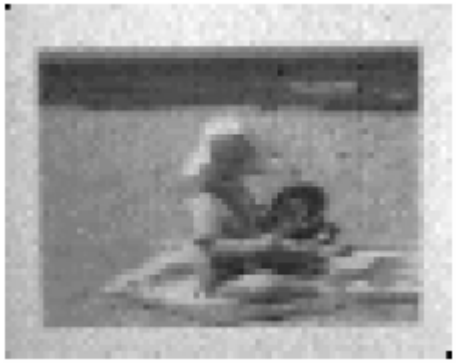
Vogelstein et al, NECO 2007



IFAT Cortical Model
4800 silicon neurons
4,194,304 synapses

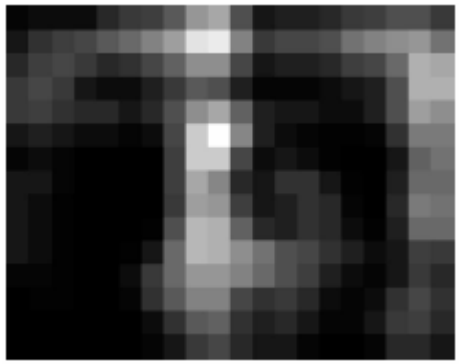
Octopus Retina
80 x 60 pixels
AER spiking output





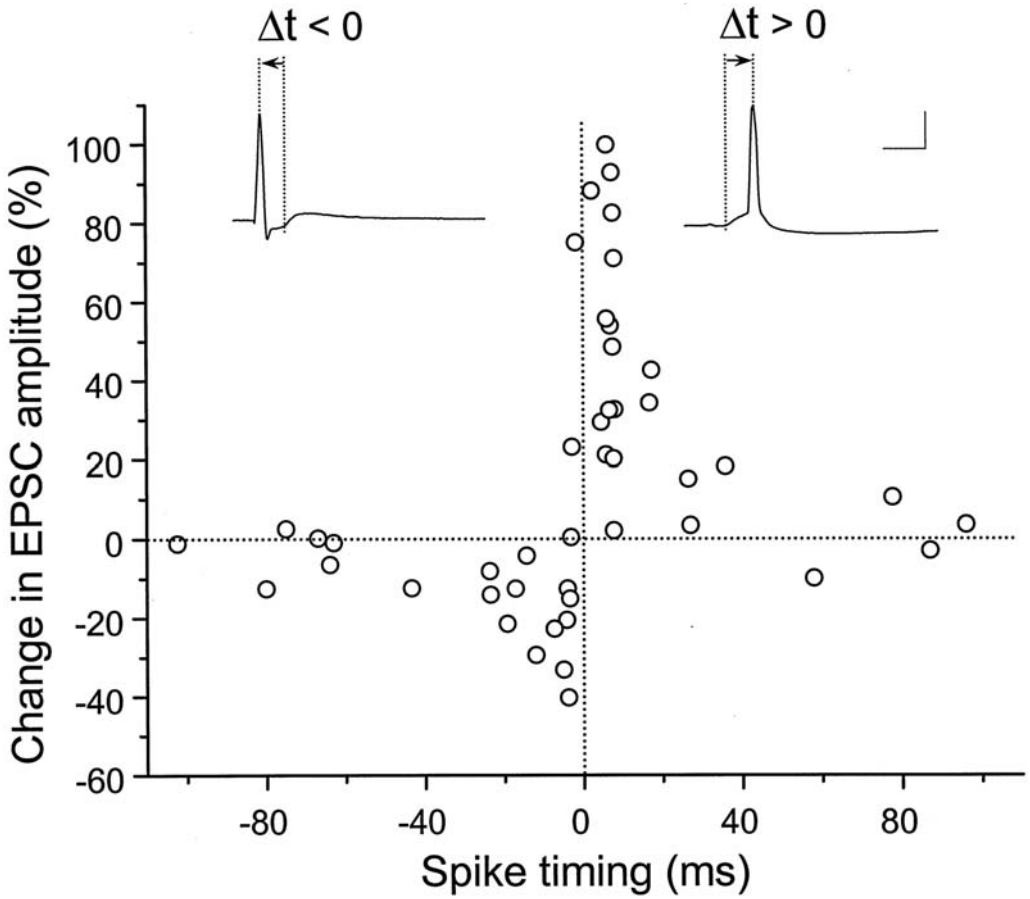
OR image

Simple cell response



Saliency map

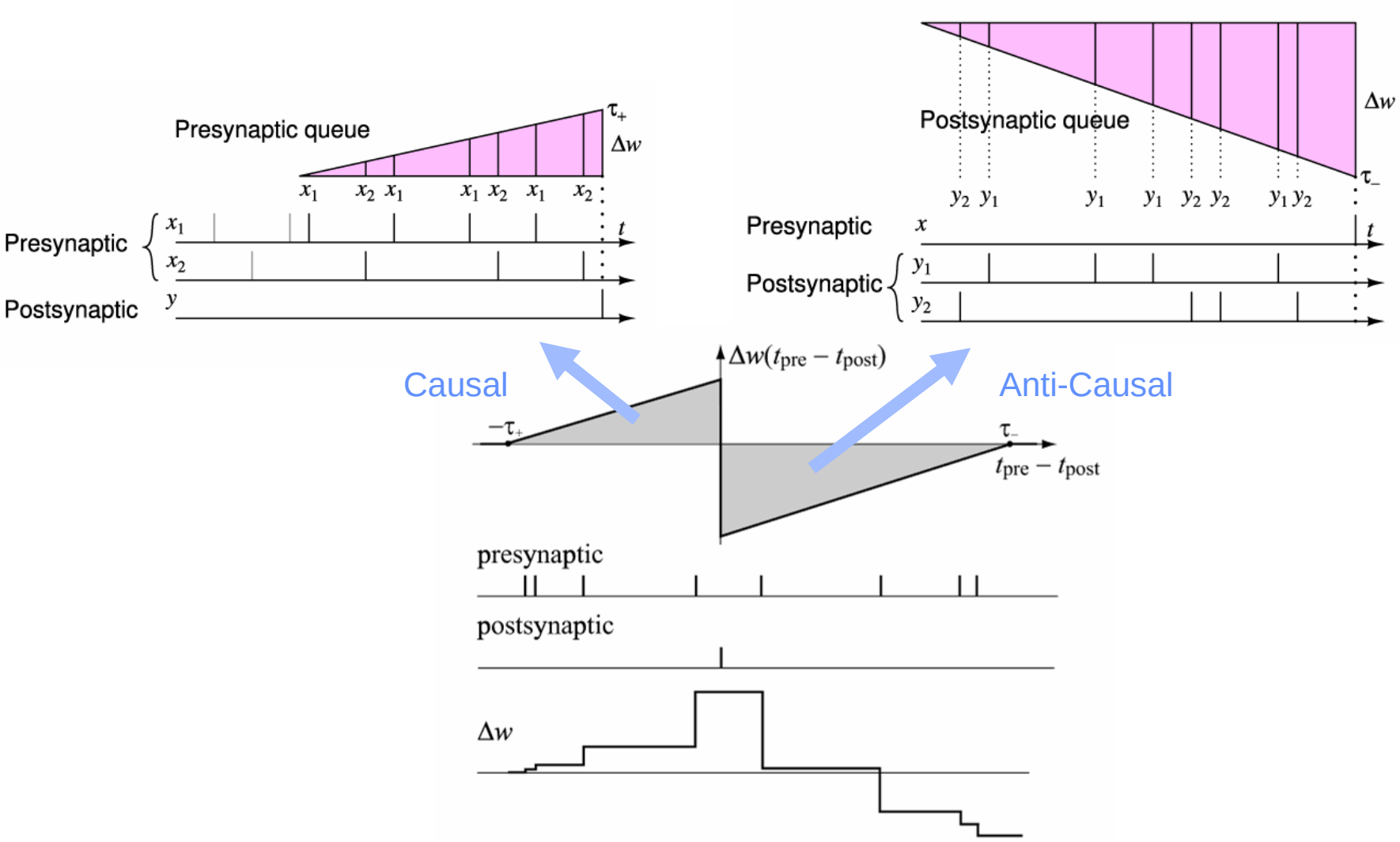
Spike Timing-Dependent Plasticity



Bi and Poo, 1998

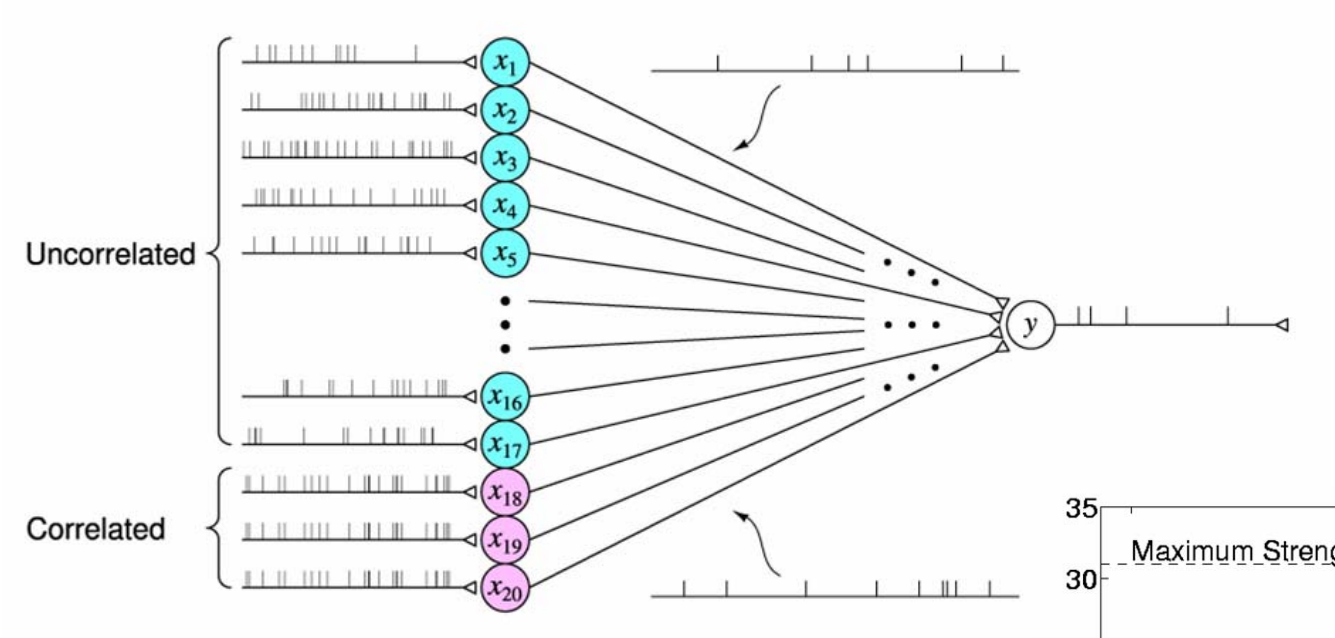
Spike Timing-Dependent Plasticity

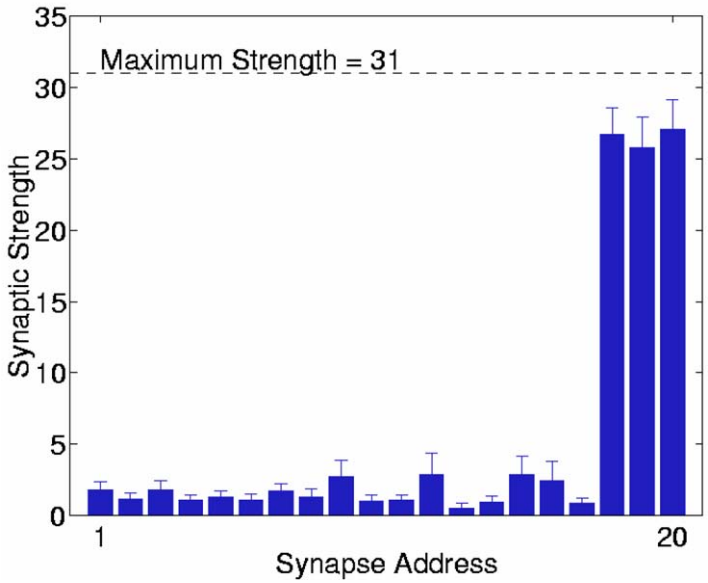
in the Address Domain



Spike Timing-Dependent Plasticity on the IFAT

Vogelstein *et al*, NIPS*2002

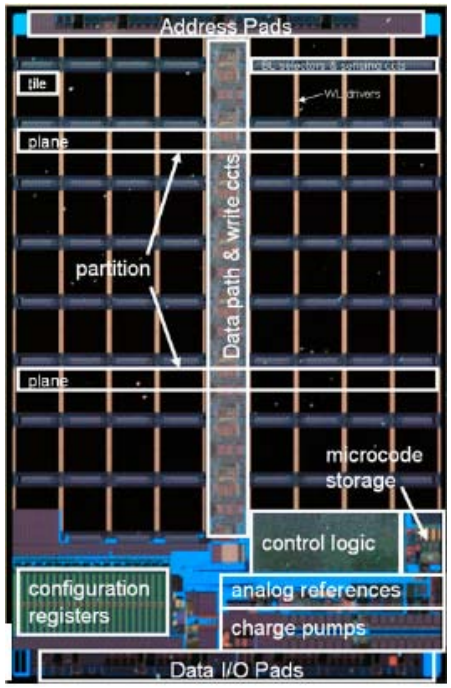


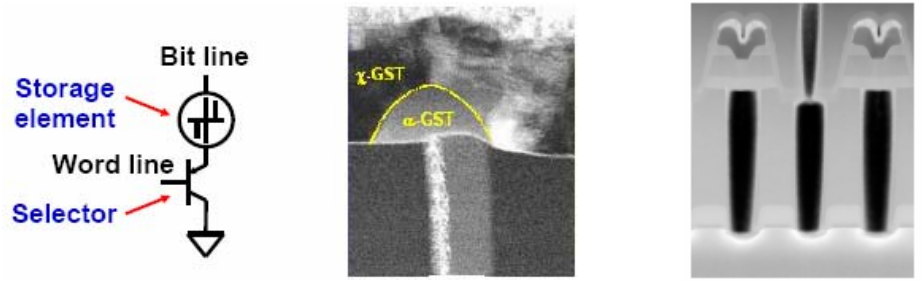


Scaling and Complexity Challenges

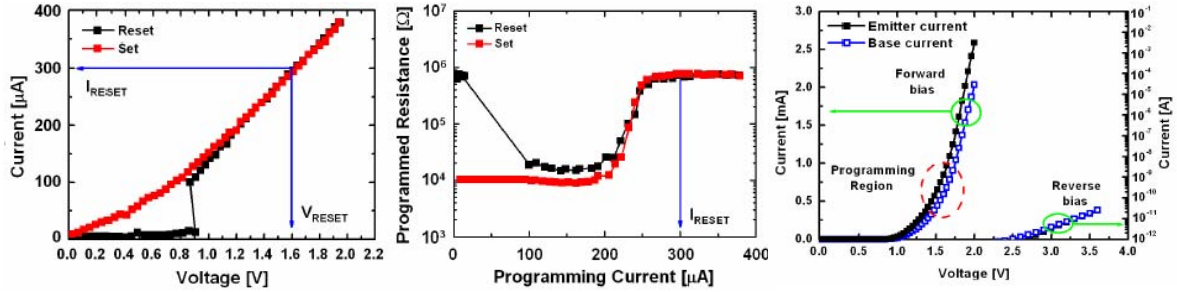
- **Scaling the event-based neural systems to performance and efficiency approaching that of the human brain will require:**
  - Integration of synaptic arrays with neural event transceivers
    - *High density (10^{12} neurons, 10^{15} synapses within 5L volume)*
      - *Non-volatile memory technology (Flash, PCM, MRAM, …)*
    - *High energy efficiency (10^{15} synOPS/s at 15W power)*
      - *Adiabatic switching in event routing and synaptic drivers*
    - *Dynamic resource allocation*
      - *Efficient mapping of sparse brain architecture onto tiled synaptic arrays*
  - 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*

Phase Change Memory (PCM) Nanotechnology





(a) (b) (c)

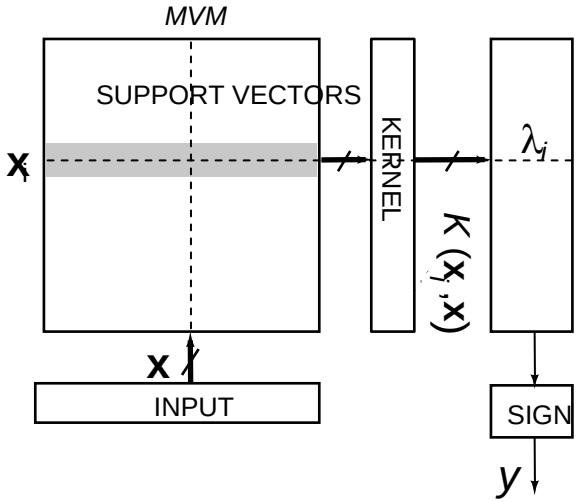


(d) (e) (f)

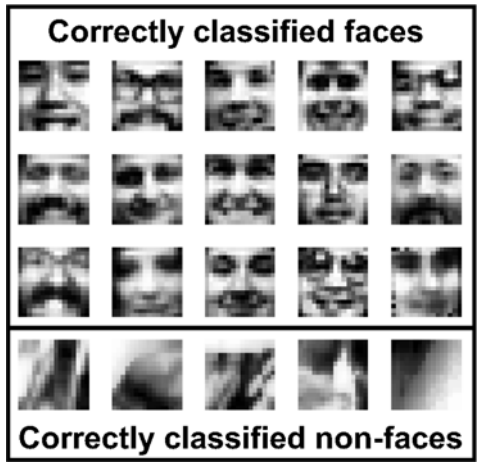
Intel/STmicroelectronics (Numonyx) 256Mb multi-level phase-change memory (PCM) [Bedeschi et al, 2008]. Die size is 36mm2 in 90nm CMOS/Ge2Sb2Te5, and cell size is 0.097μm2. (a) Basic storage element schematic, (b) active region of cell showing crystalline and amorphous GST, (c) SEM photograph of array 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*

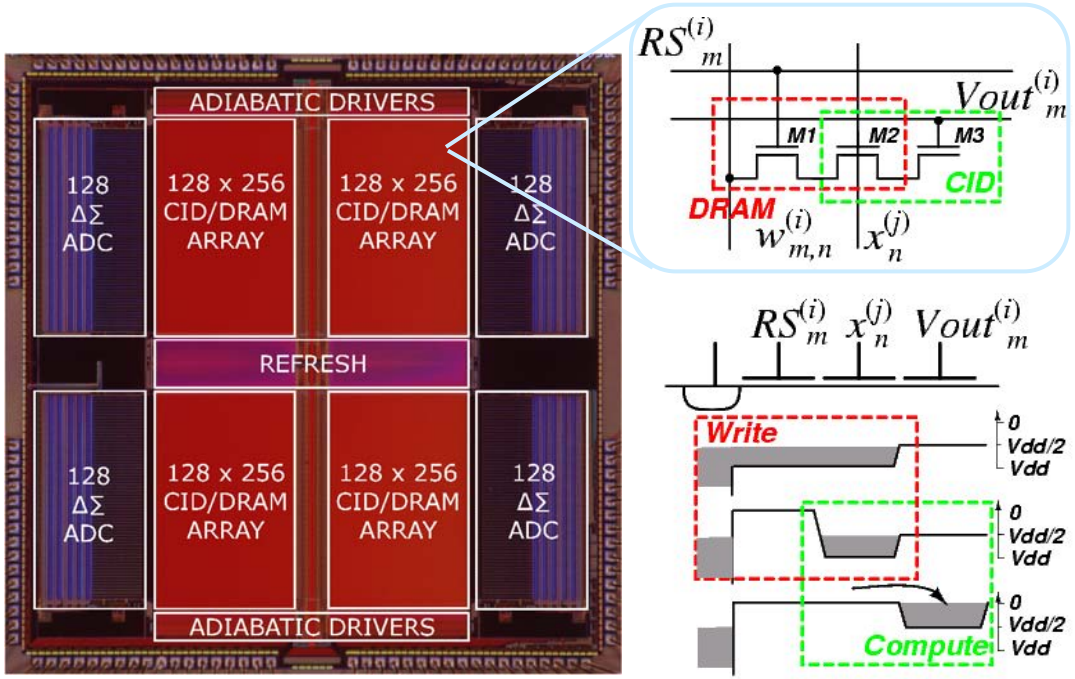
Kerneltron III: Adiabatic Support Vector "Machine"



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



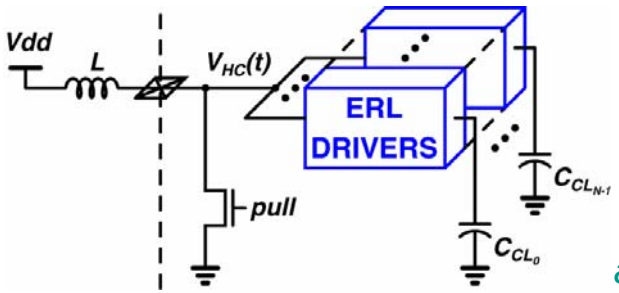
Classification results on MIT CBCL face detection data



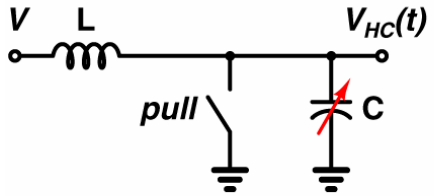
Karakiewicz, Genov, and Cauwenberghs, VLSI'2006; CICC'2007

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

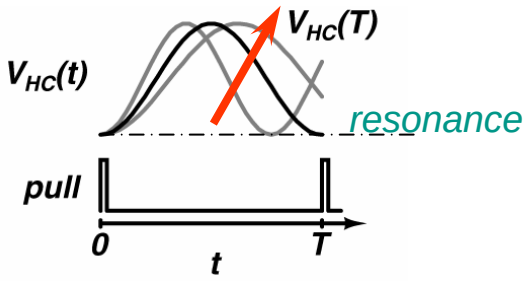
Resonant Charge Energy Recovery

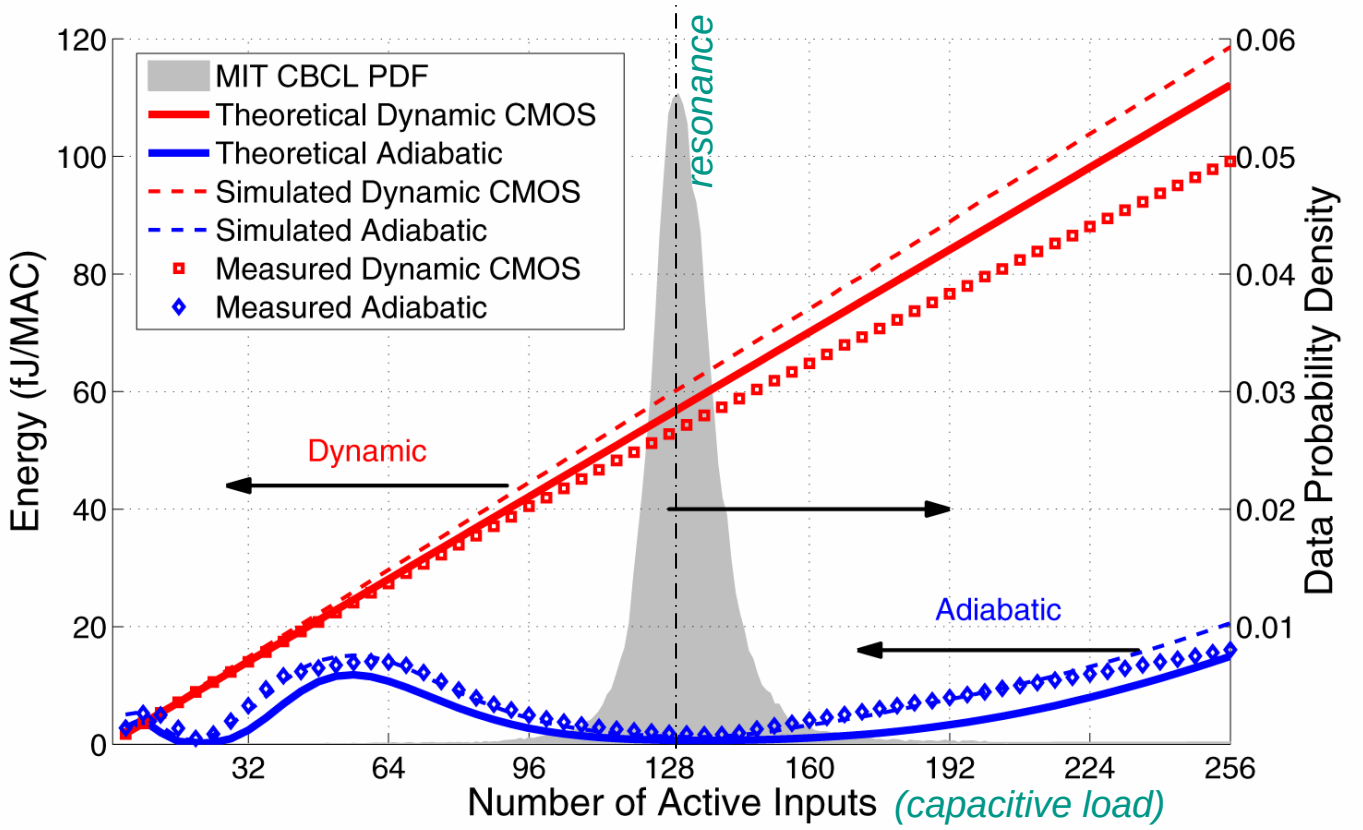


CID array



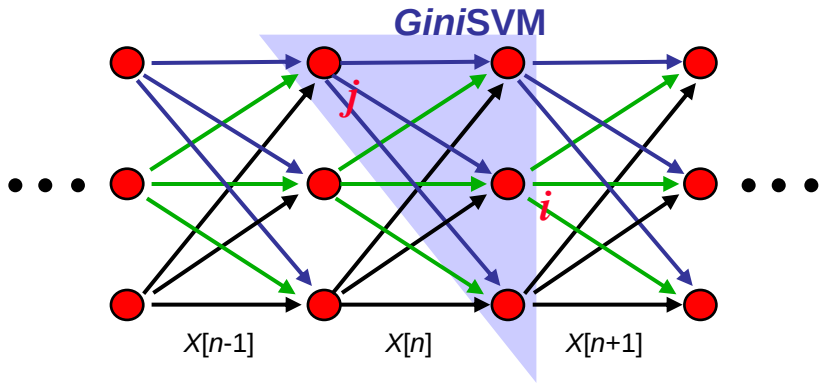
capacitive load





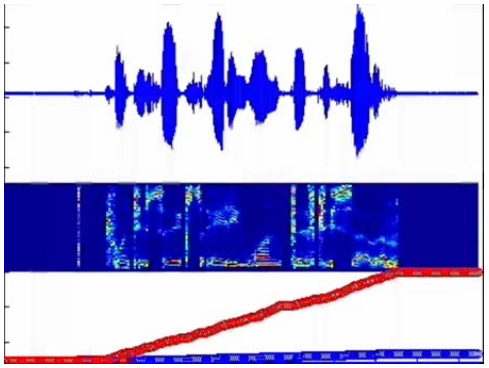
Sub-Micropower Analog VLSI Adaptive Sequence Decoding

Chakrabartty and Cauwenberghs (NIPS'2004)

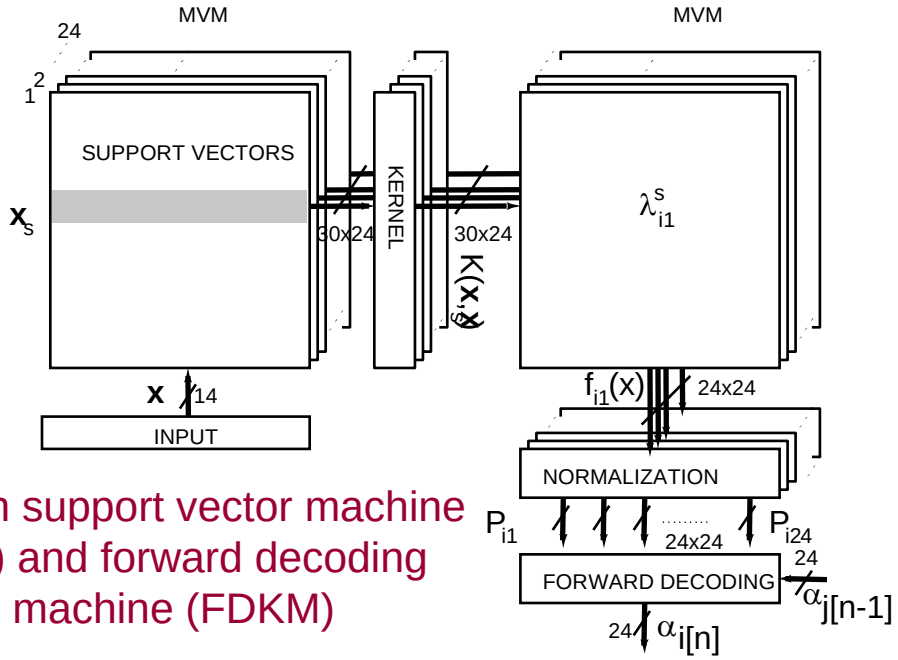


Forward decoding MAP sequence estimation

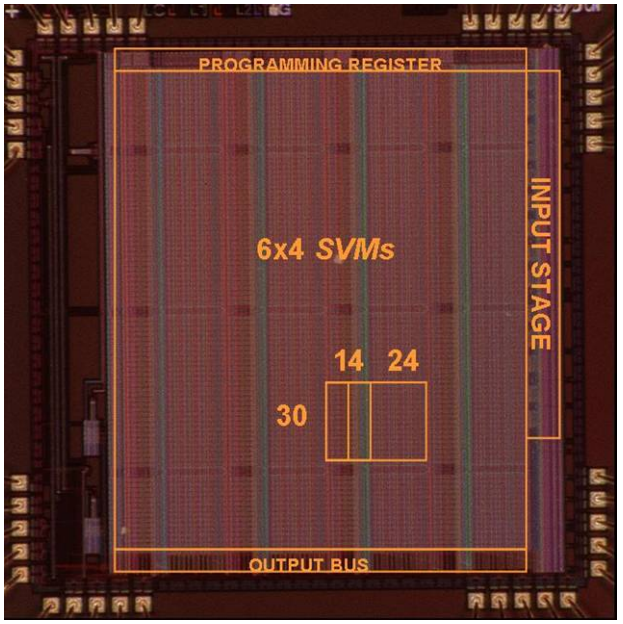
840 nW power



Biometric verification

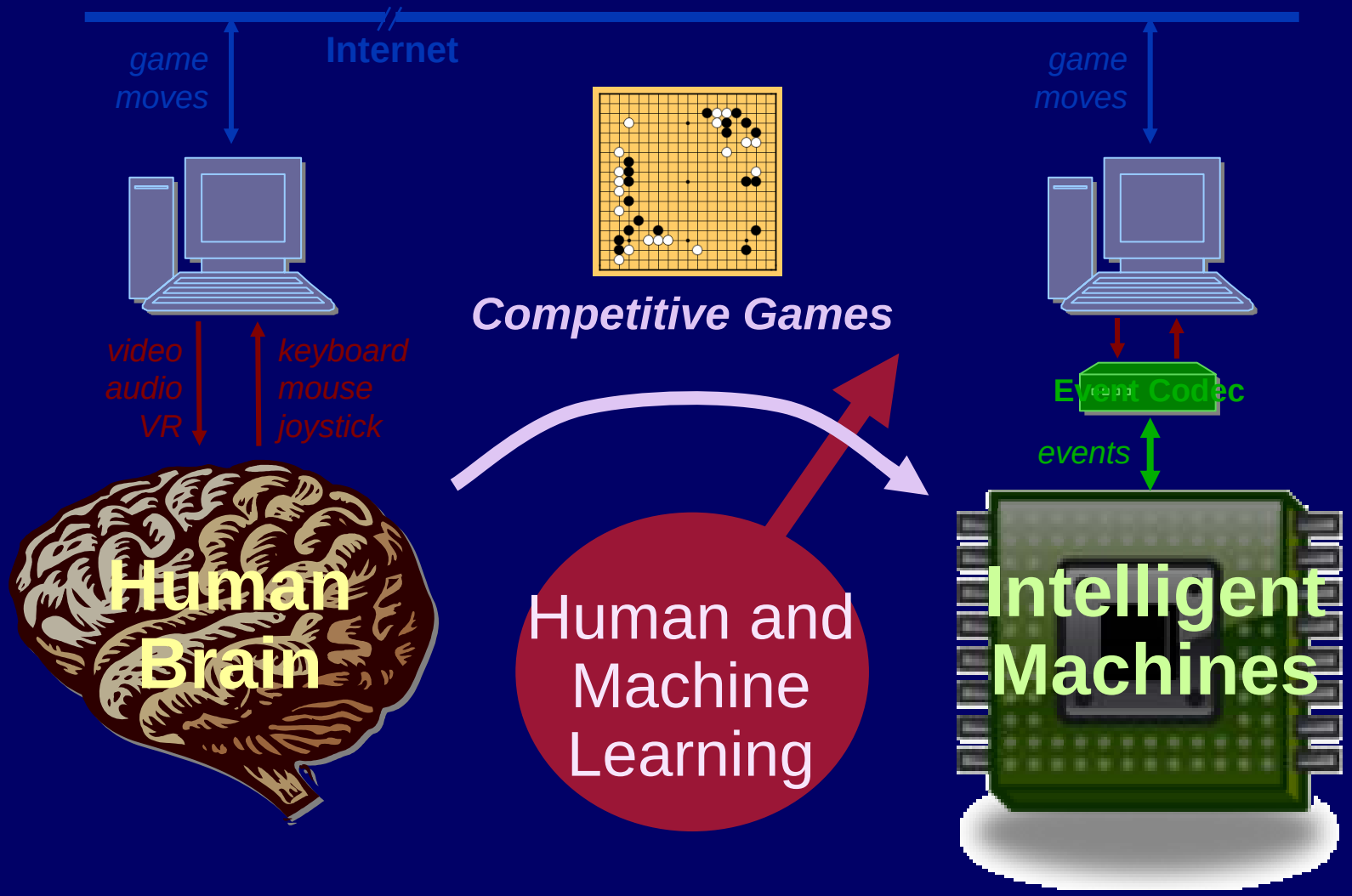


Silicon support vector machine (SVM) and forward decoding kernel machine (FDKM)

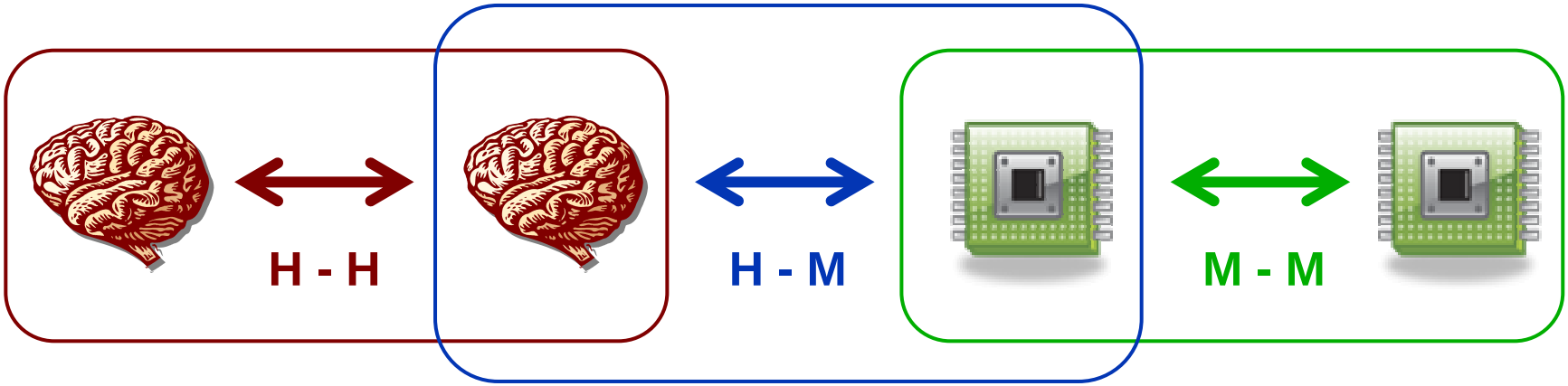


Adaptive Machine Intelligence

Training Machines towards Human Performance through Games



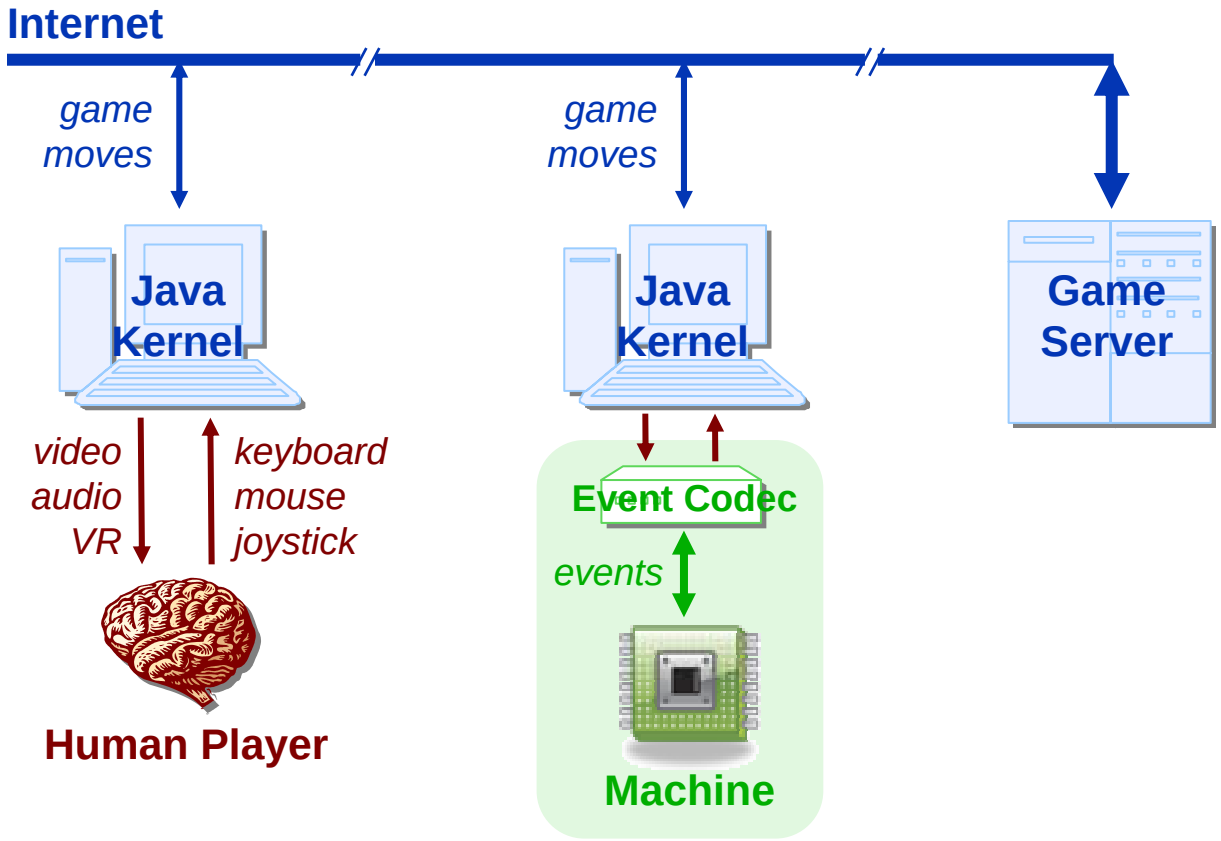
Competitive Games: Humans and Machines



- Learning through experience in two-player zero-sum games:
  - Humans to humans: *Novices learn from experts to become experts.*
  - Humans to machines: *Towards human-level machine performance.*
  - Machines to machines: *Beyond human-level machine performance.*
- Heterogeneous competitive ranking:
  - *ELO score ranks humans and machines alike.*
  - *Turing test.*

Web-Based Competitive Games

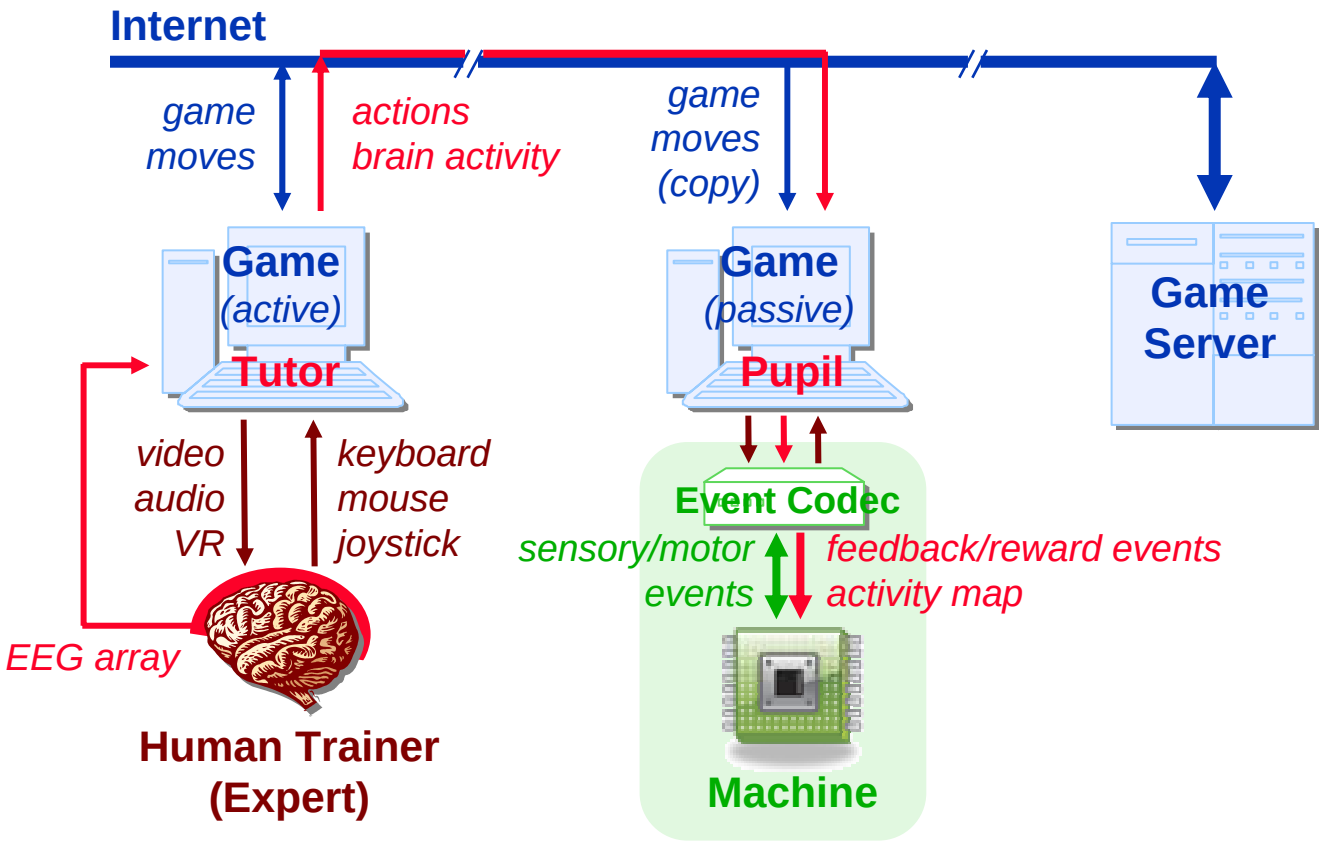
Humans and Machines



- Event codec adapter and machine interface
- Central logging, ranking, and matchmaking at external game server

Web-Based Competitive Games

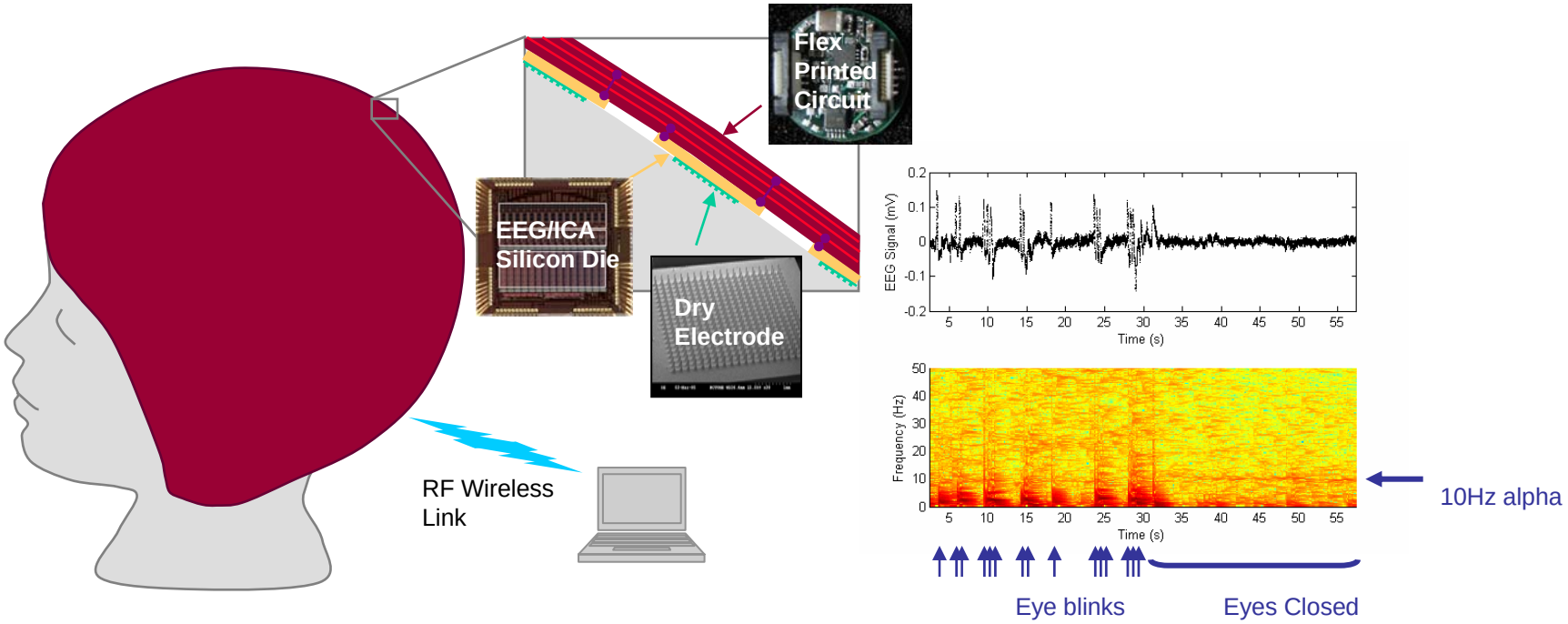
Humans Tutoring Machines



- Machine learns by observing actions *and* internal representation (EEG brain activity) of human expert.
- Neuromorphic: trained machine approaches human brain function *and* form.

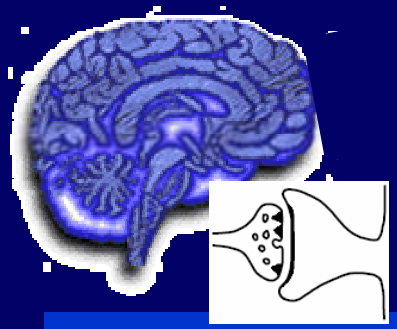
Wireless EEG/ICA Neurotechnology

with Tom Sullivan, Steve Deiss, Tzyy-Ping Jung and Scott Makeig

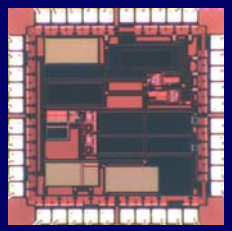


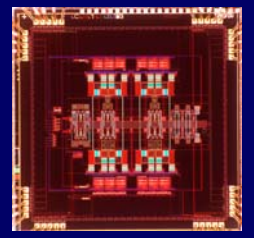
- **Integrated EEG/ICA wireless EEG recording system**
  - Scalable towards 1000+ channels
  - Dry contact electrodes
  - Wireless, lightweight
  - Integrated, distributed independent component analysis (ICA)

Integrated Systems Neurobiology



Neuromorphic/ Neurosystems Engineering





Neural Systems

Learning & Adaptation

Silicon Microchips

Human/Bio Interaction

Environment

Sensors and Actuators