

Programmable and Configurable, Neuromorphically Inspired, Ultra-Low Power Computation

Professor Jennifer Hasler

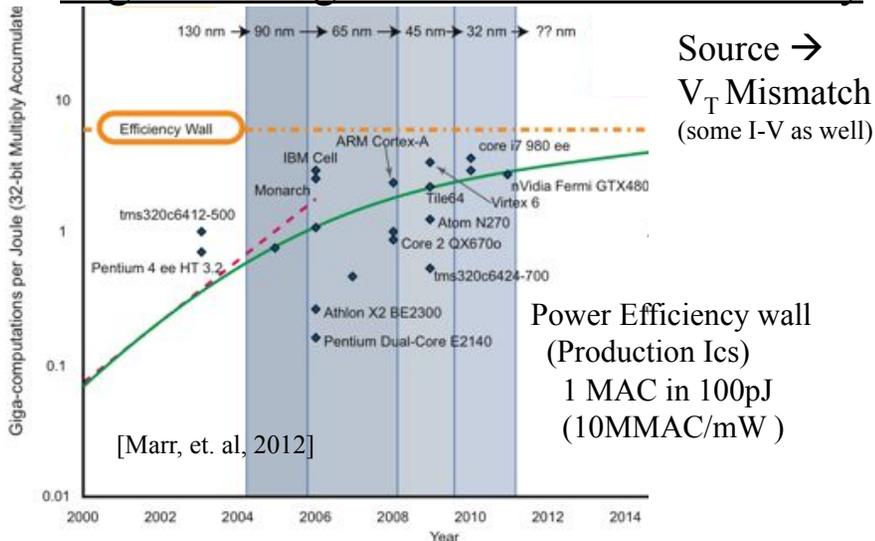
ph67@gatech.edu

Georgia Tech



Why Analog (Physical Based) Processing?

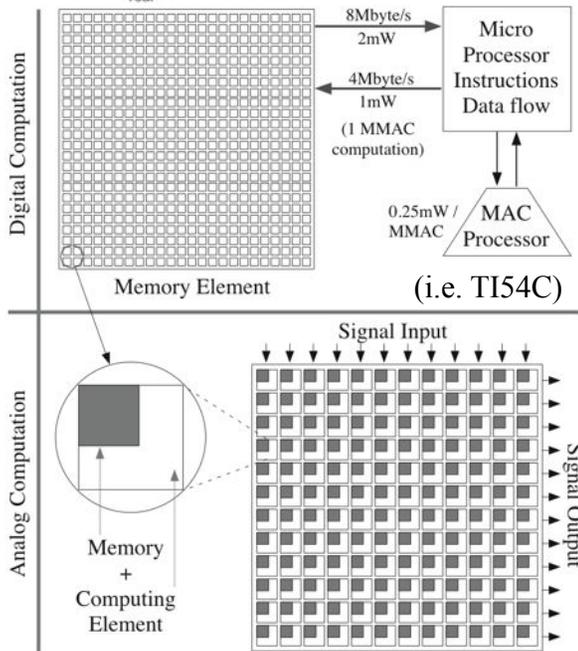
Digital Hitting Limits of Power Efficiency



Mead Hypothesis:
Analog x1000
efficiency improvement

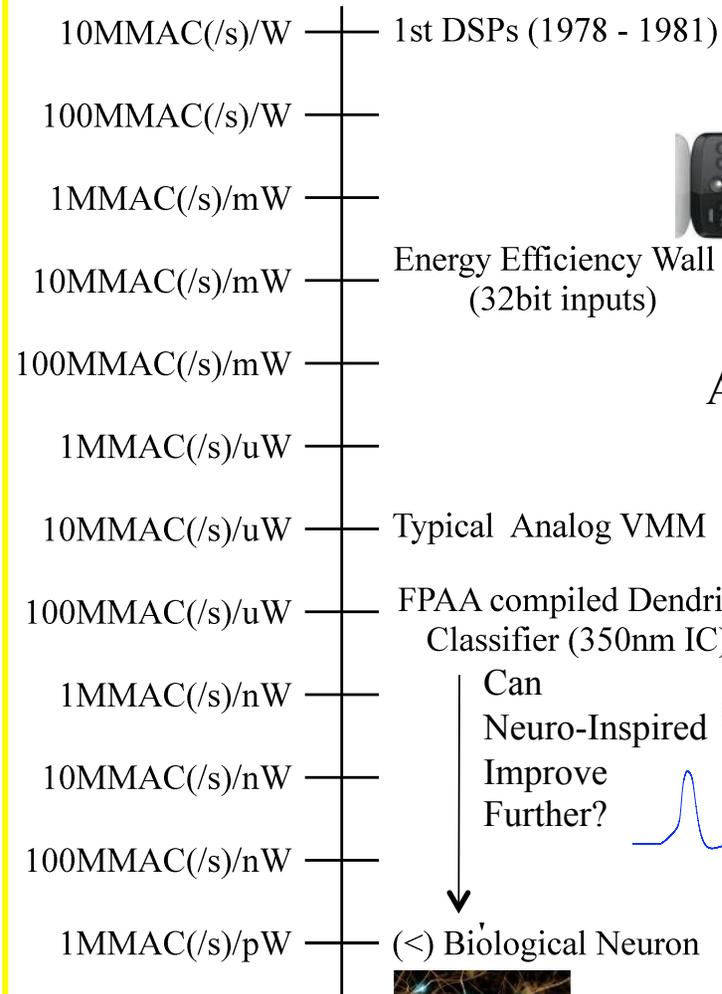
- Analog (VMM):
100 fJ / MAC
(10MMAC/ μ W)

- Other Analog SP similar:
Freq Decomp / Analog FT
VMM, GMM
Classifiers
Adaptive Filters

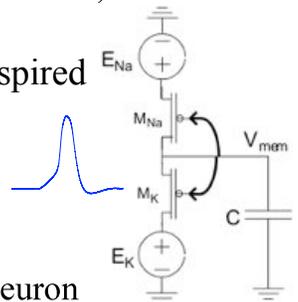
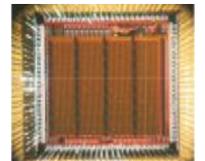


Performance Advantage for Emerging Architectures

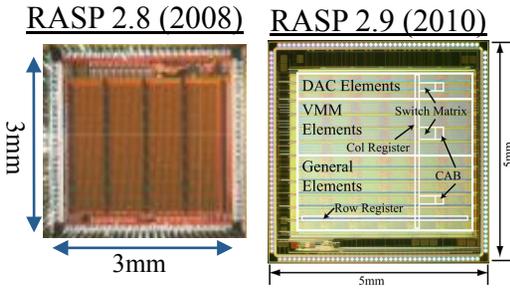
Power Efficiency Scaling



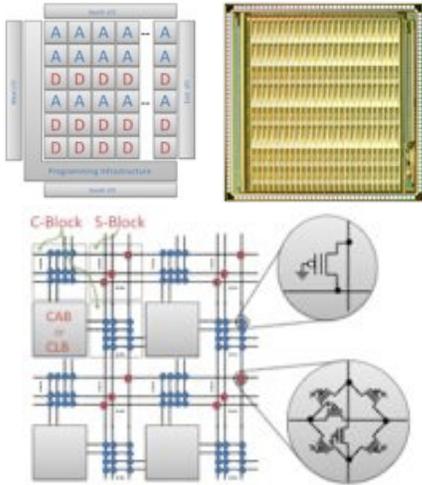
AnalogSP



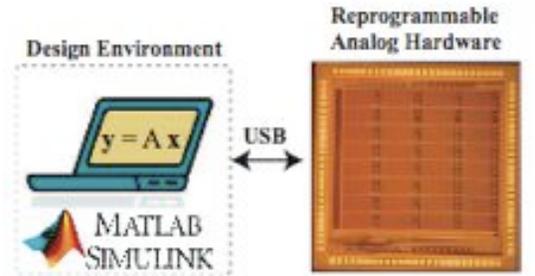
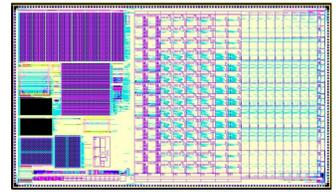
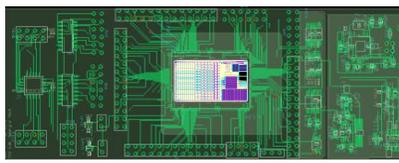
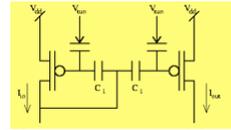
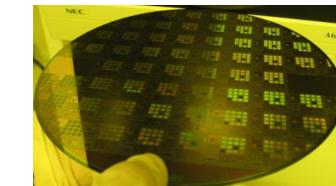
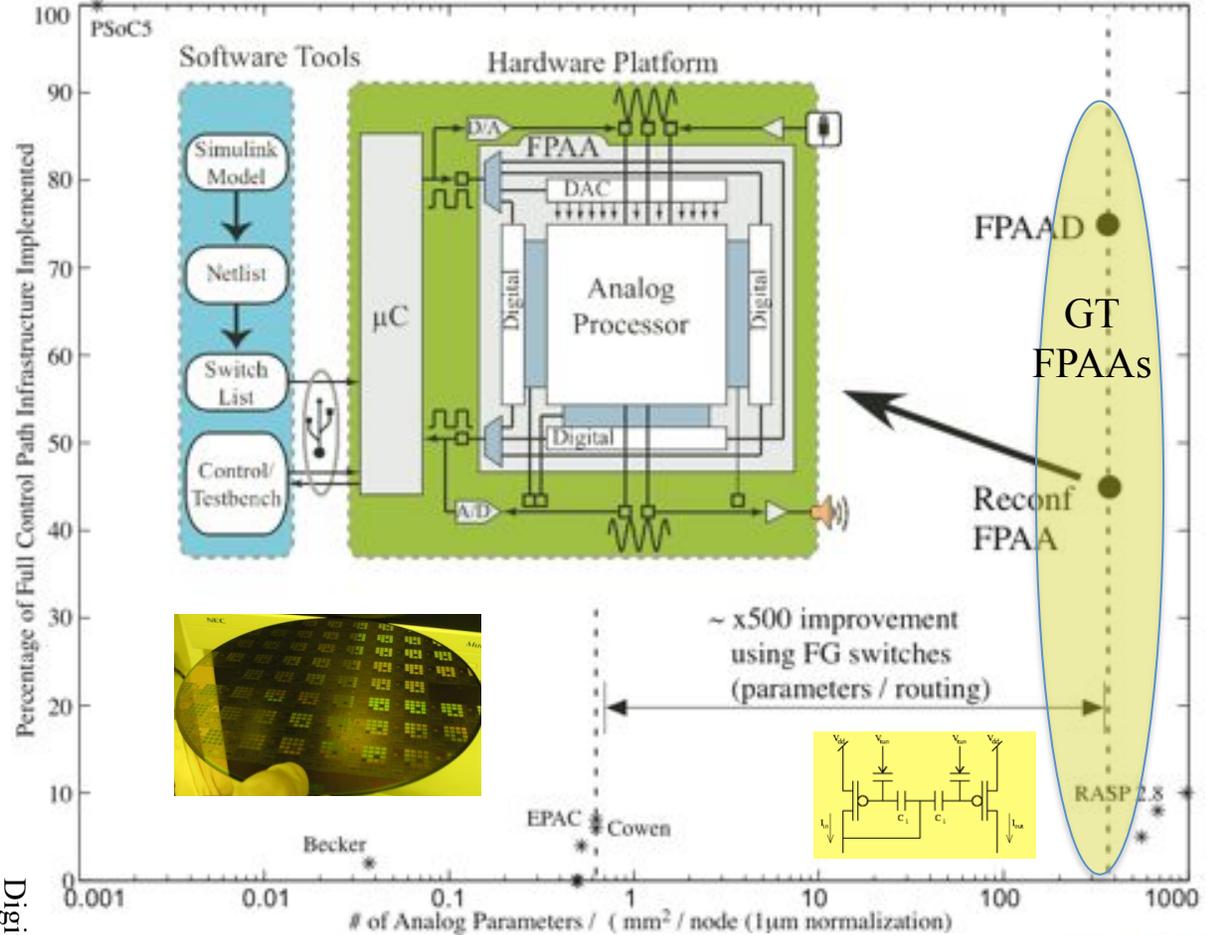
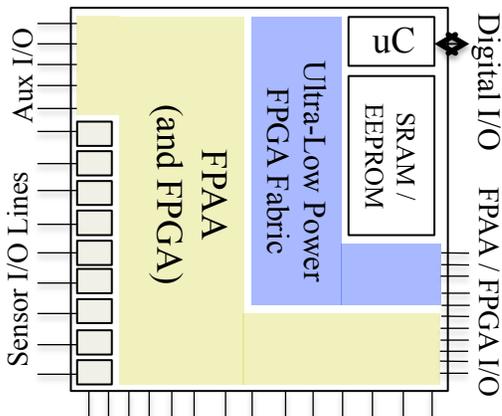
Large-Scale FPAA's → Practical Analog SP



Integrated FPAA + FPGA (2011)



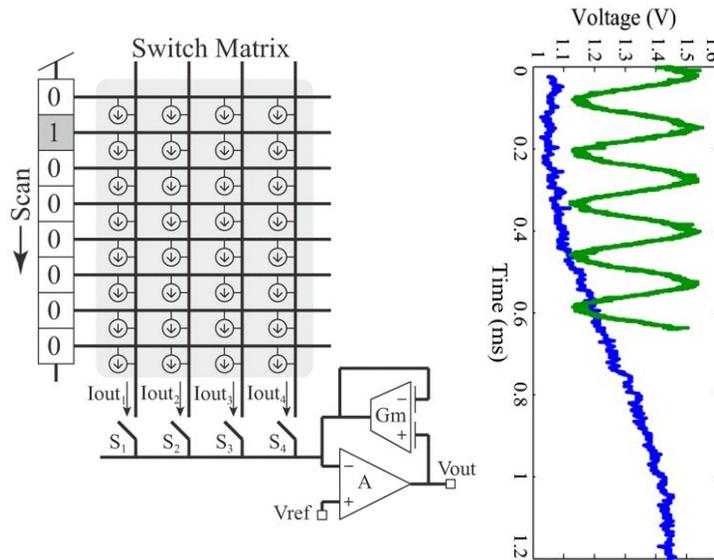
15 different working FPAA ICs since 2004



FPAA computing through routing fabric

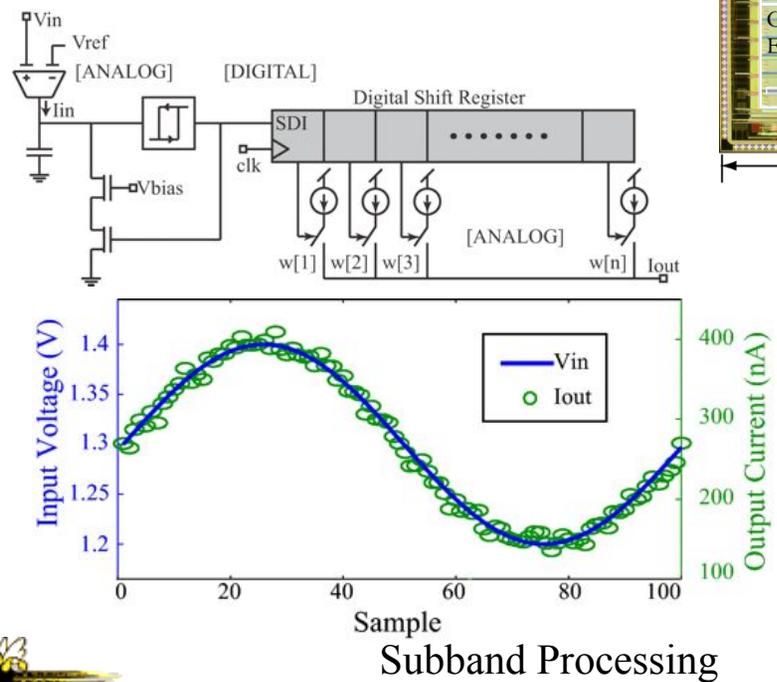
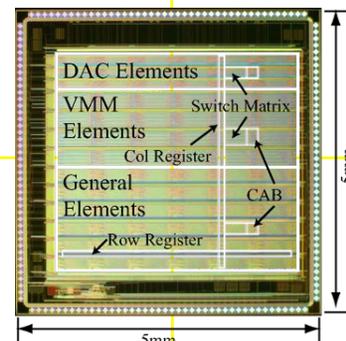
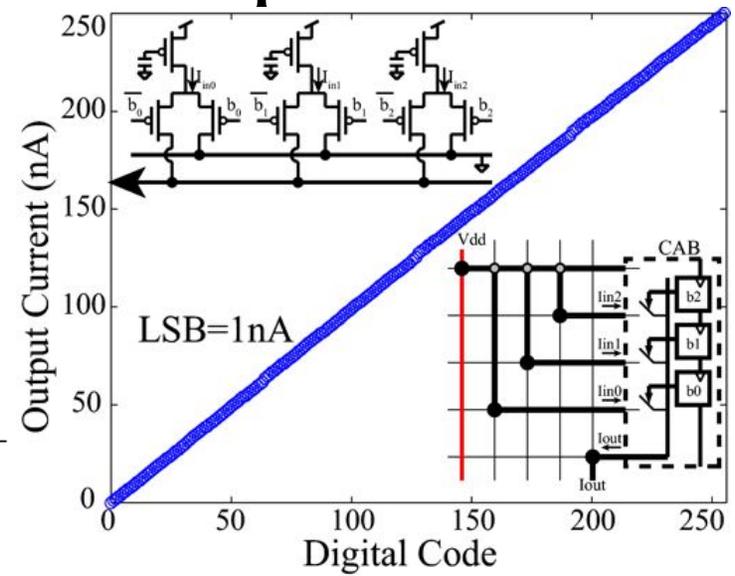


One FPAA, Four Examples



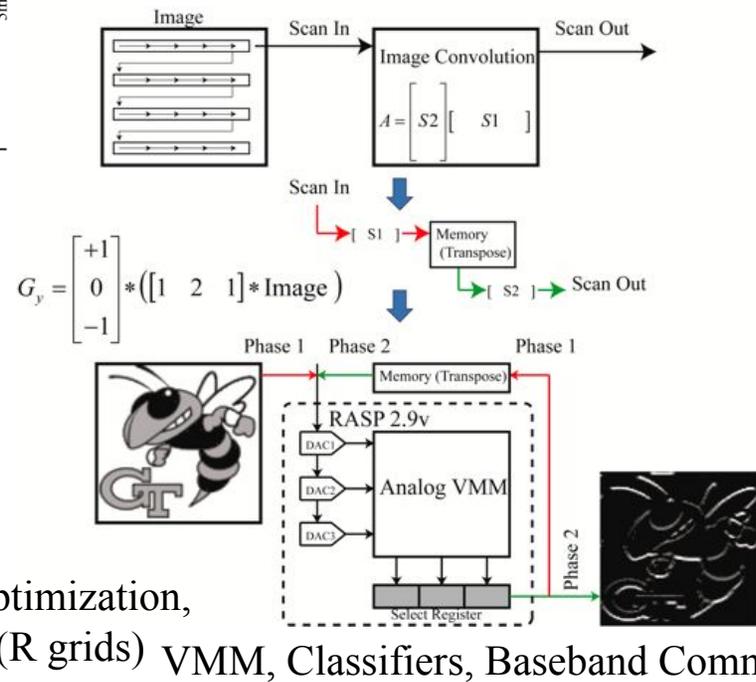
Arbitrary Waveform

DAC in Routing



Mixed-Signal FIR

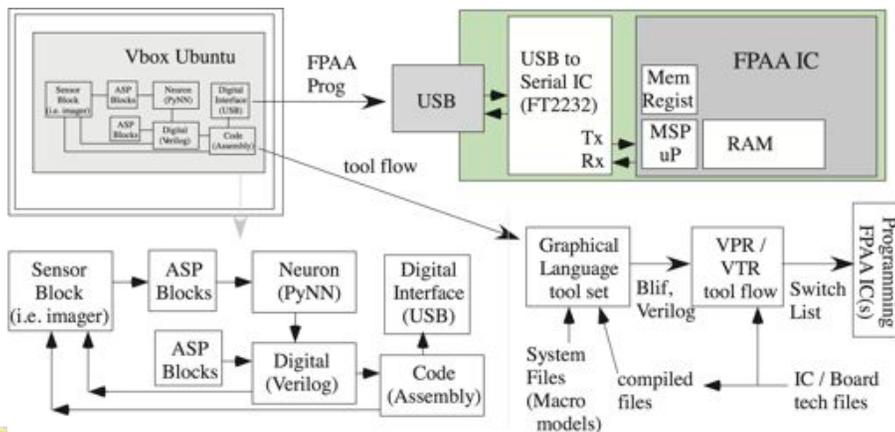
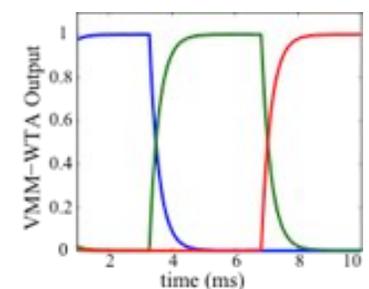
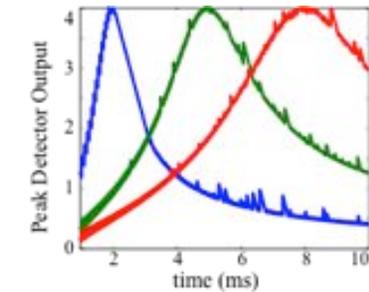
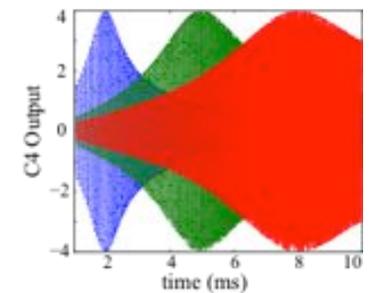
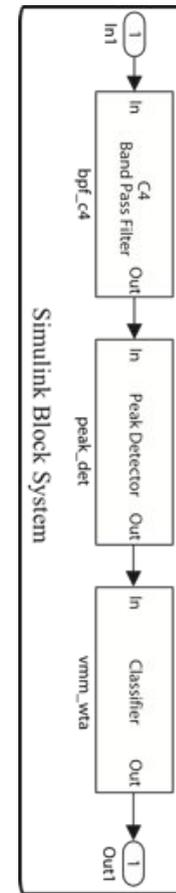
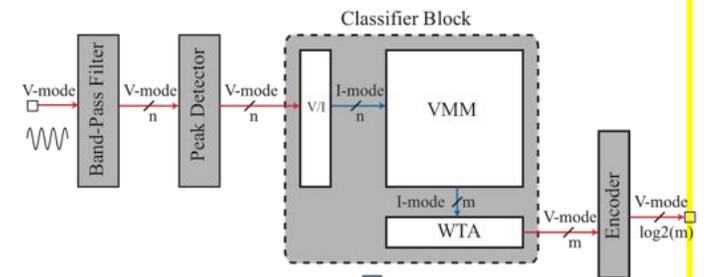
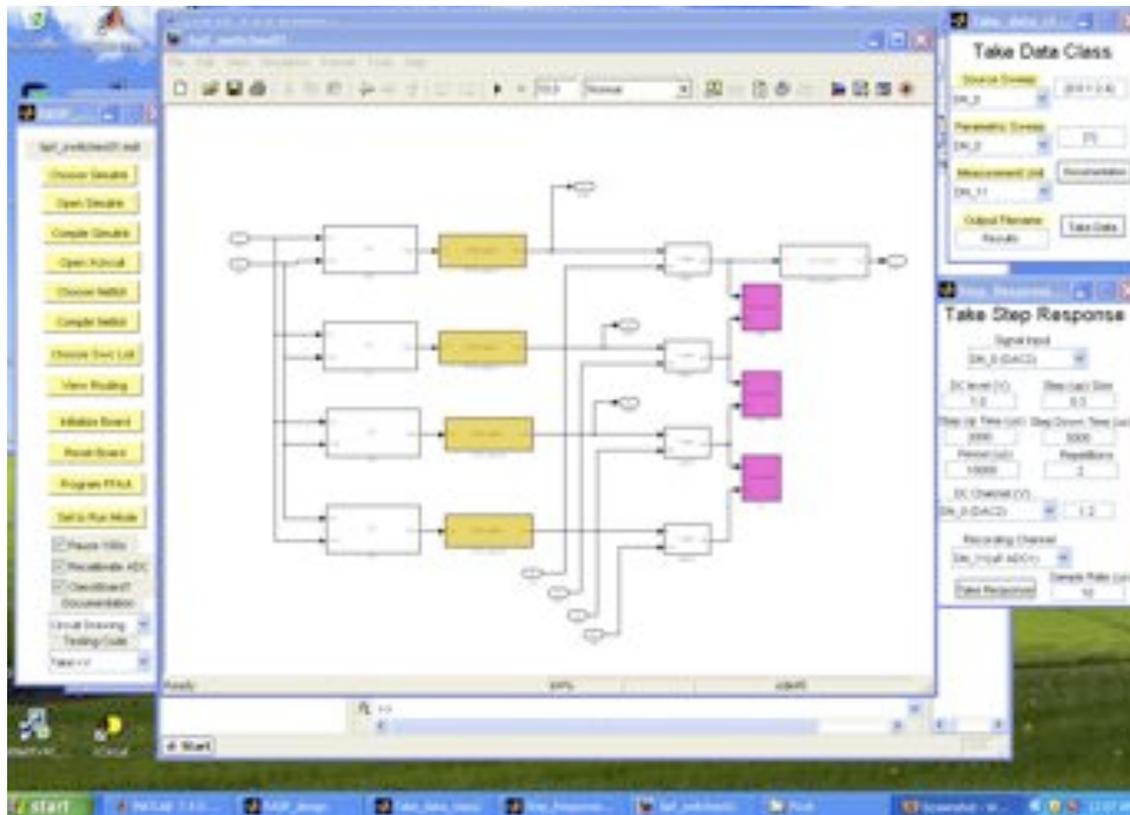
Image Convolution



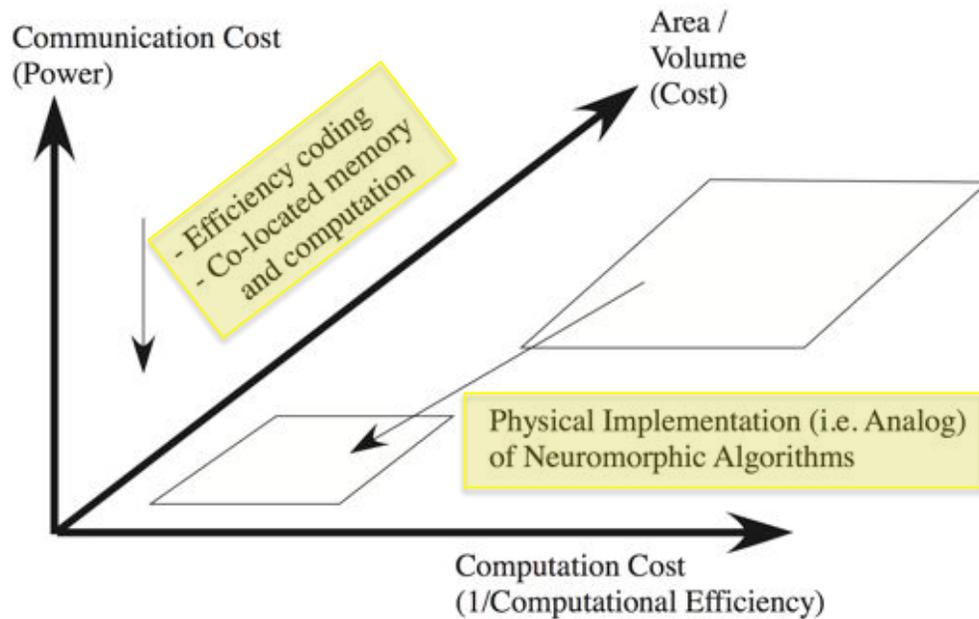
Constrained Optimization, Path-Planning (R grids) VMM, Classifiers, Baseband Comm



FPAA Tool Infrastructure



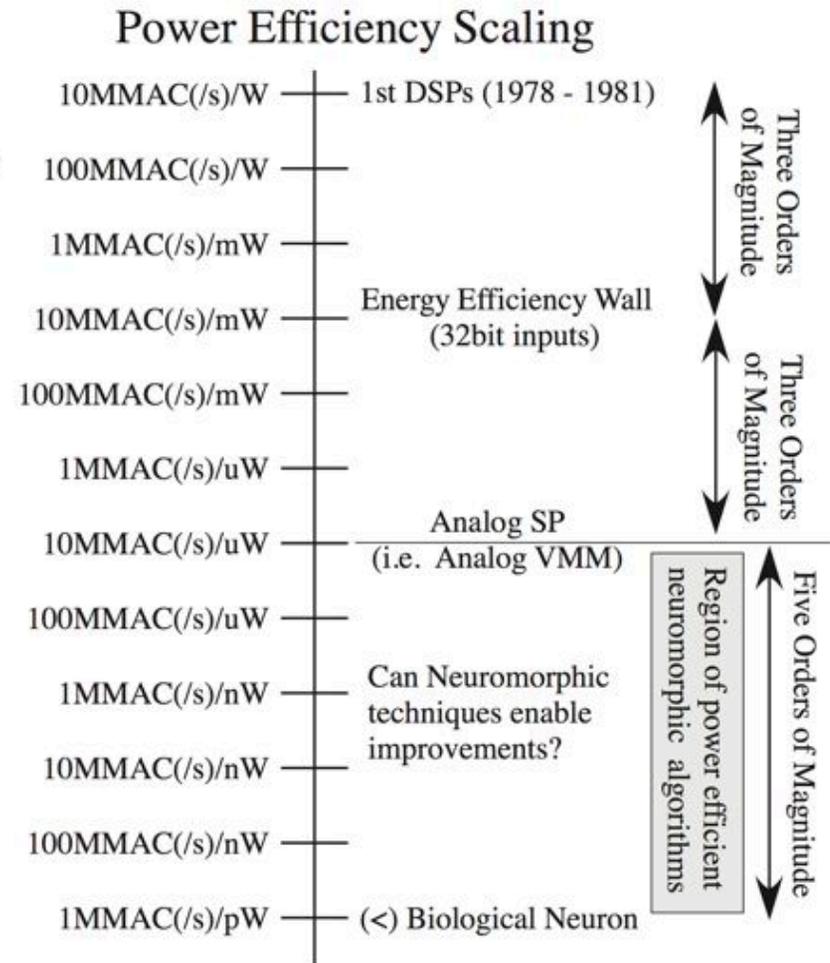
Neuromorphic Algorithms → Improved Apps



Brain is highly power efficient
- highly constrained by power available, key to its design (and for Si)

Building applications (i.e. robotics) makes power constraints real

Leverage Analog SP ICs for robust system development

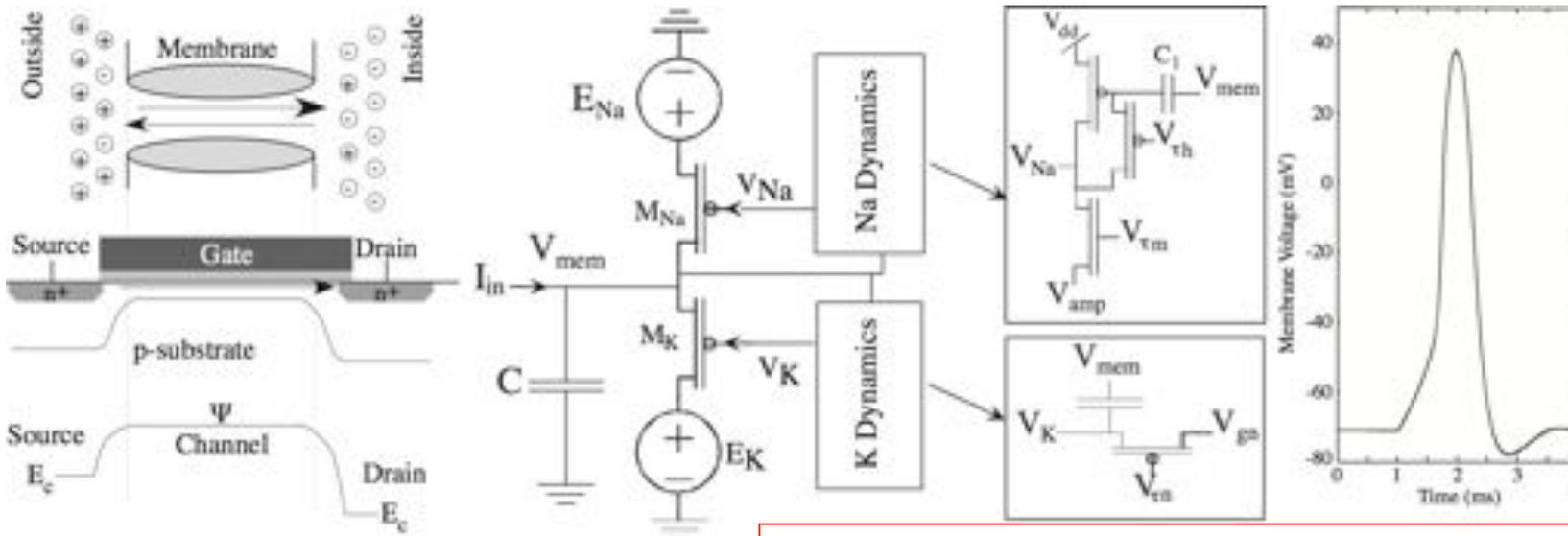


- Neuromorphic processing = event based processing uses power only when useful signals are present (“always on” in sensors or further processing)



Blocks for Large-Scale Neuromorphic Systems

Transistor Channel Models

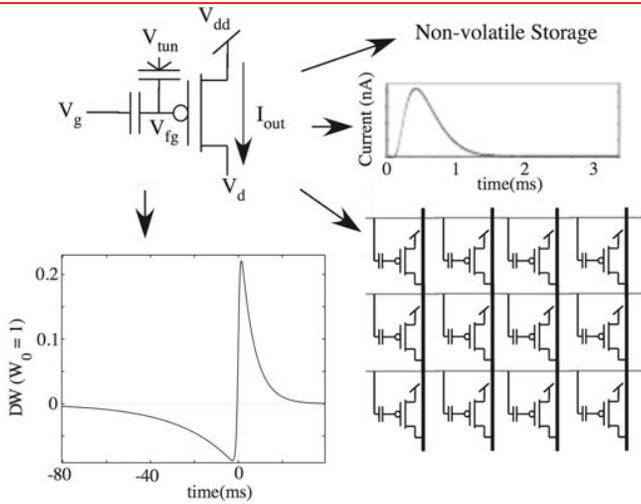
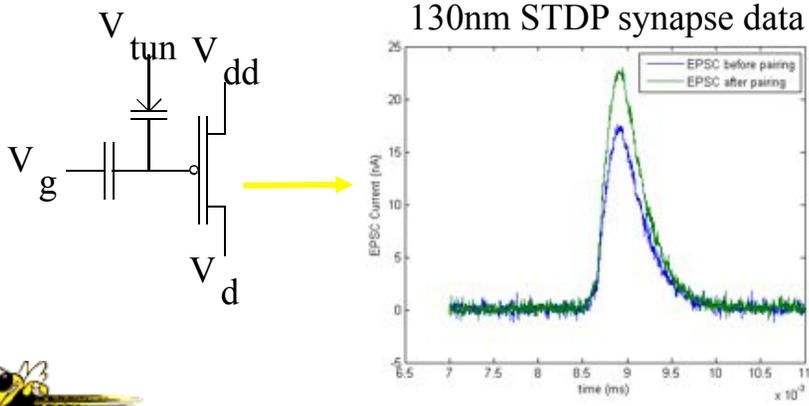


[Farquhar and Hasler, 2004]

Utilizing the physics of physical medium (Si) to efficiently implement computation

Single Transistor Learning Synapses

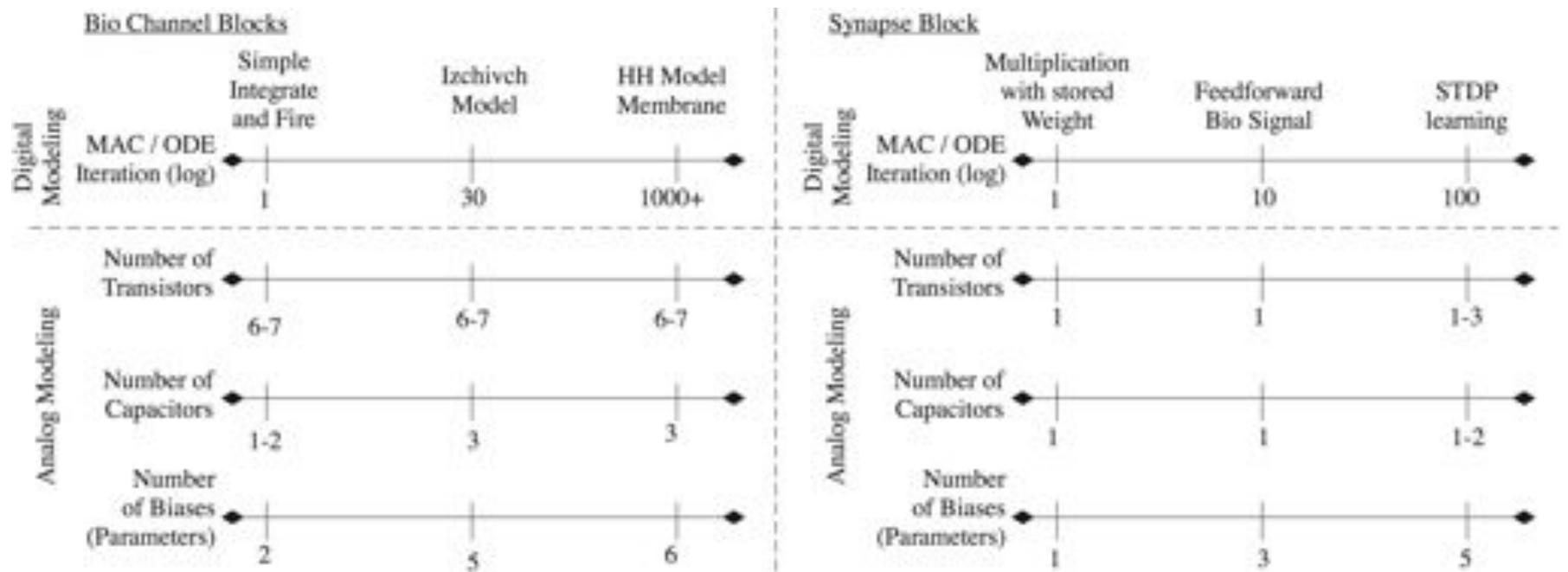
- Single Transistor Learning Synapses [Hasler, et. al, NIPS 1994, BMES 1994]



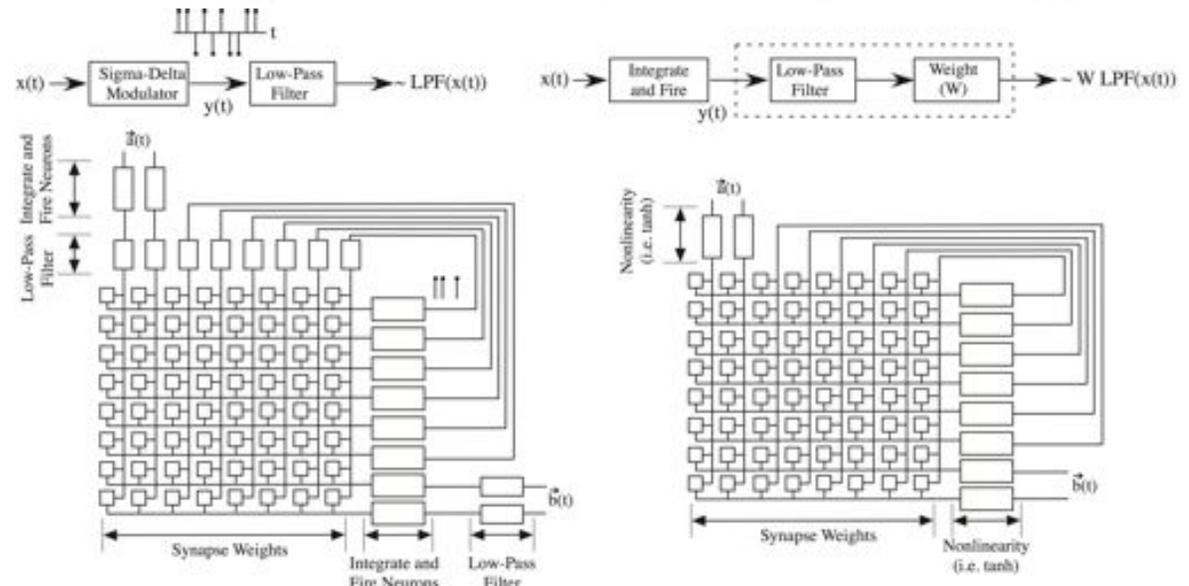
Si CMOS approach can achieve densities while avoiding issues with device integration with Si



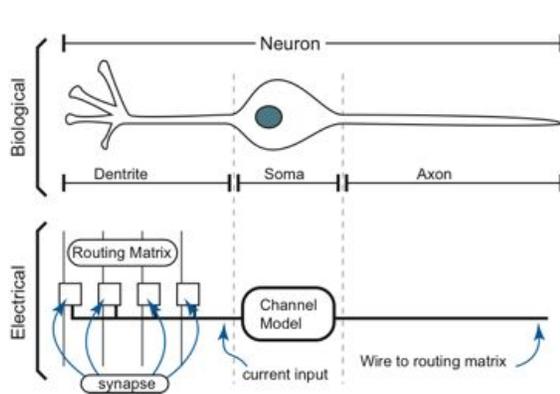
Comparing Physical to Digital Computations



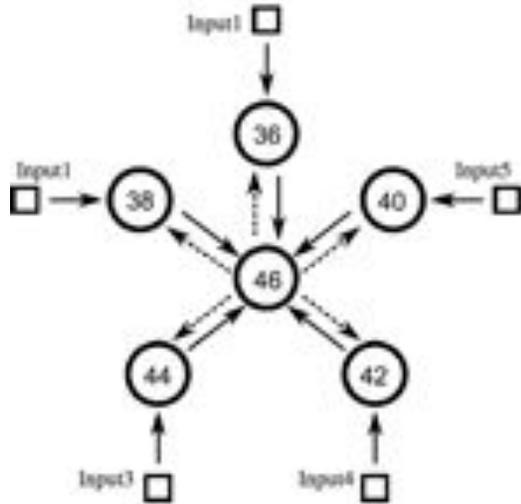
- FG enables analog precision
- not imprecise or overly noisy components
 - Power constrains digital to similar resolutions, worse ODE dynamics



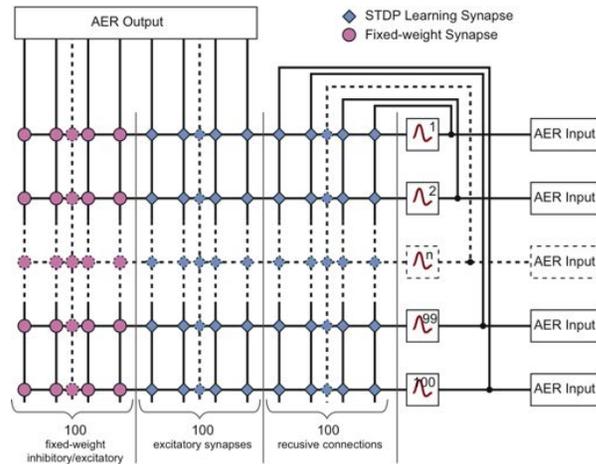
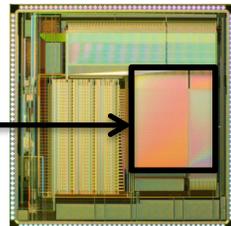
Dense Synapse + Neuron IC



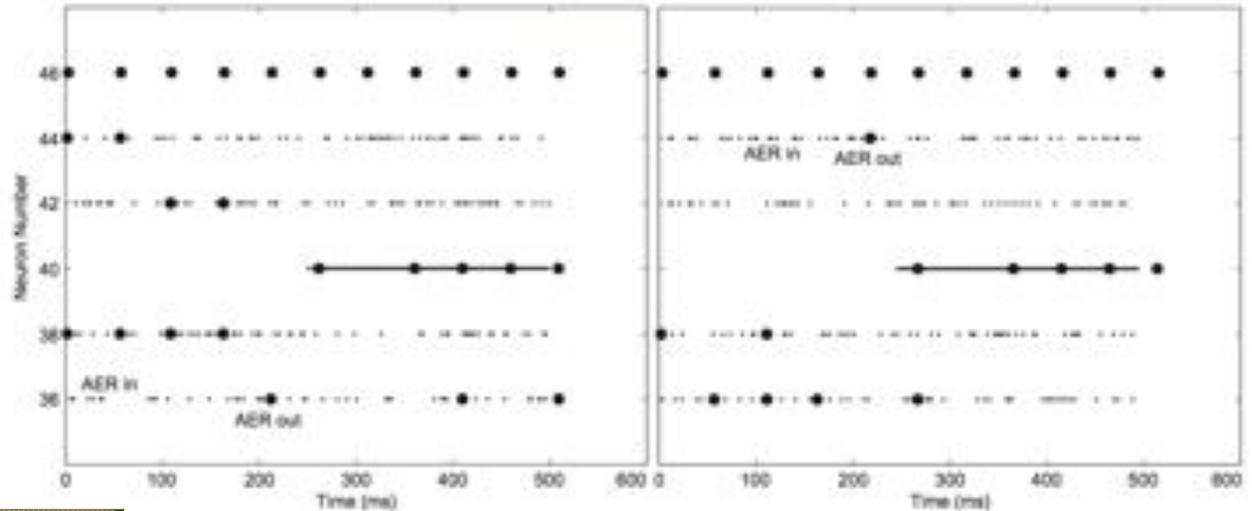
WTA network



Synapse Array
 (30k synapses
 ~ 3mm² space
 10k synapses/mm²)



- Synapse Array with configurable neuron blocks, STDP and programmable synapses
- Address Event I/O available
- Compiled from standard cells
- Mismatch Programming Essential

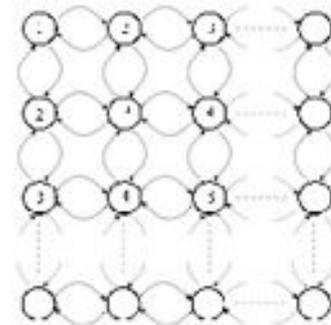
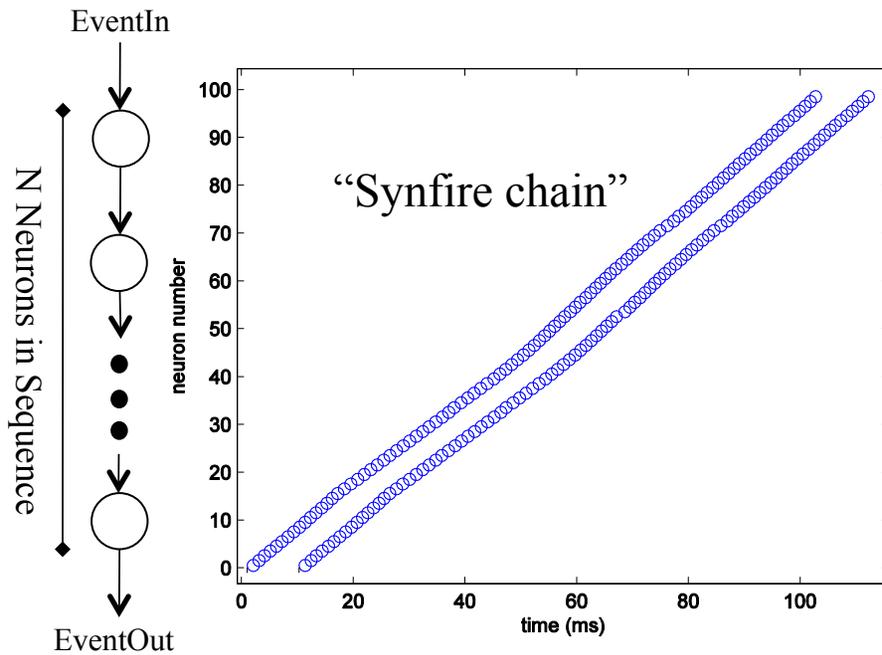
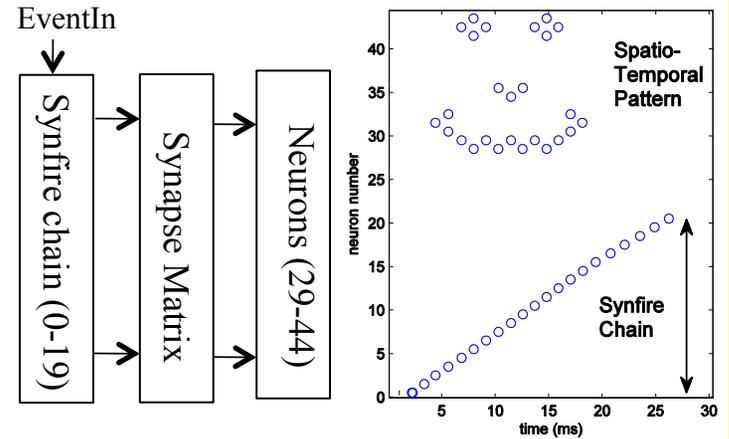
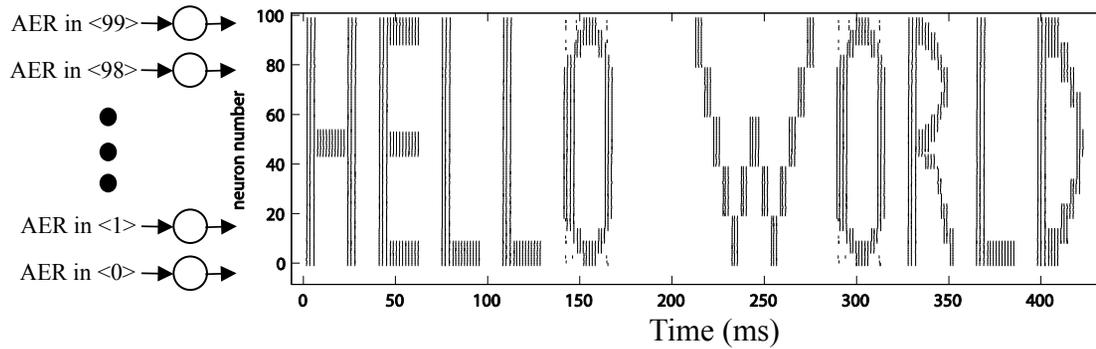


COMPARISON OF SYNAPSE DENSITY AND FUNCTION OF WORKING IMPLEMENTATIONS

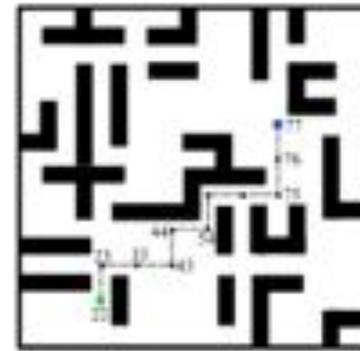
Chip Built	Process Node (nm)	Die Area (mm ²)	No of Synapses	Synapse Area (μm ²)	Syn density	Synapse Storage Resolution and Complexity
GT Neuron1d [25]	350	25	30000	133	1088	> 10bit, STDP
FACETs chip [69], [33]	180	25	98304	108	3338	4bit register
Stanford STDP	250	10.2	21504	238	3810	STDP, no storage
INI Chip [70]	800	1.6	256	4495	7023	1bit w/ learning dynam
ISS + INI Chip [71]	350	68.9	16384	3200	26122	2.5 w/ learning dynam



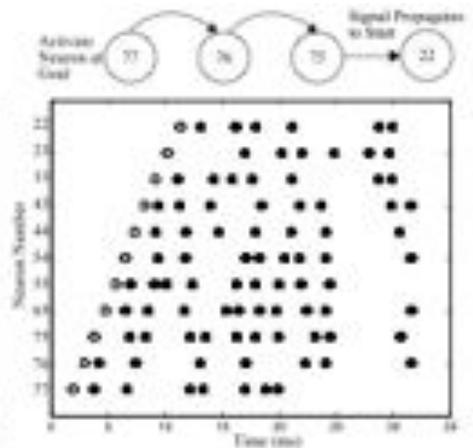
Larger Neuron Experiments



- Neuromorphic Path Planning

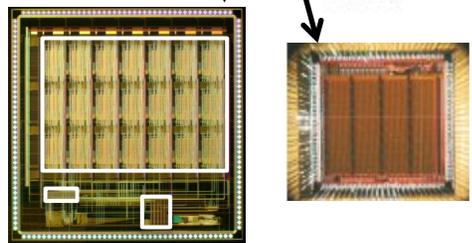
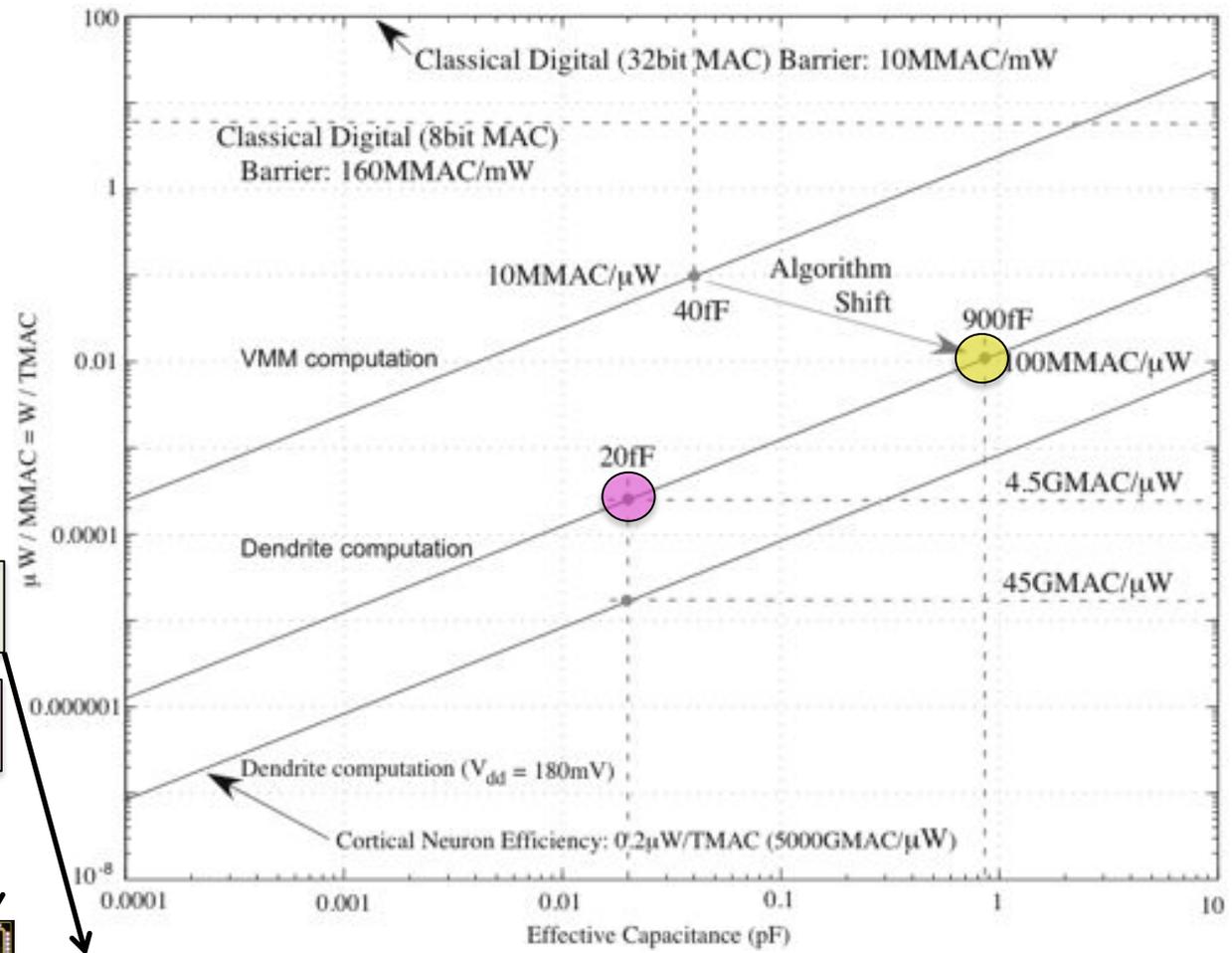
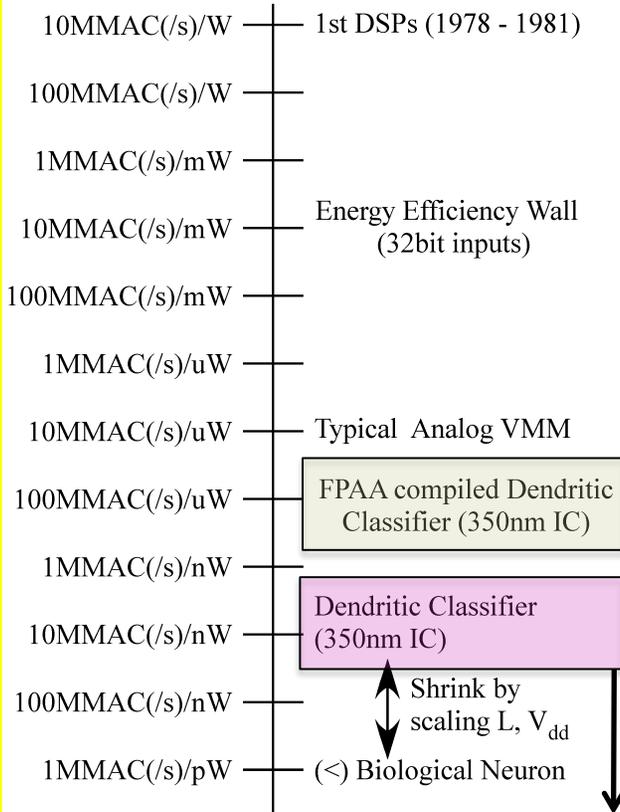


Solution Path (22, 23, 33, 43, 44, 54, 55, 65, 75, 76, 77)



Why Dendritic Computation?

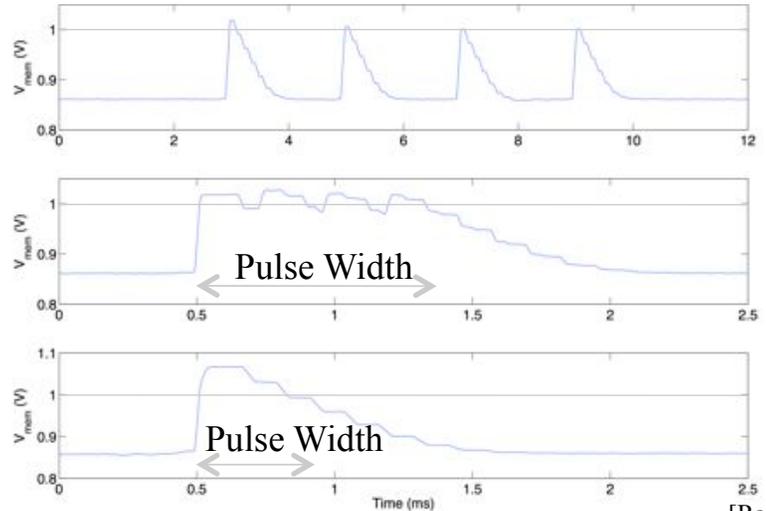
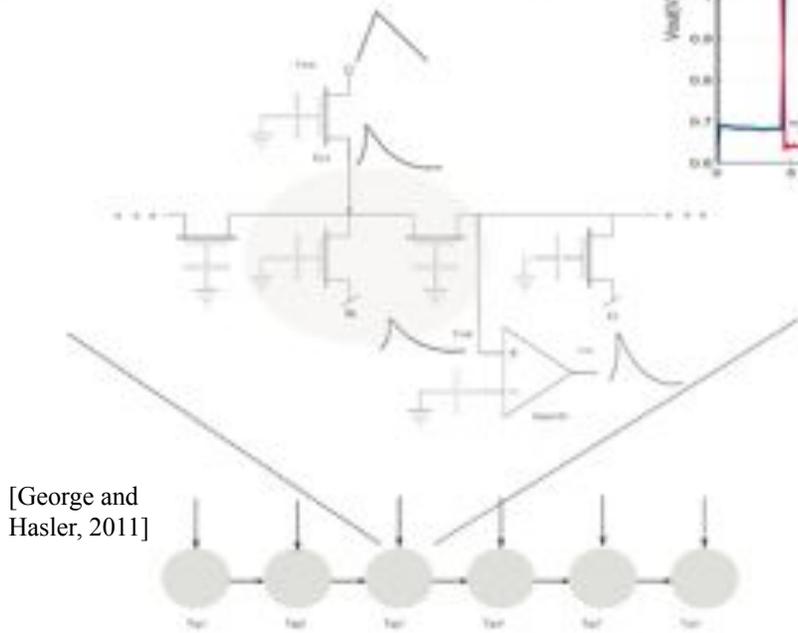
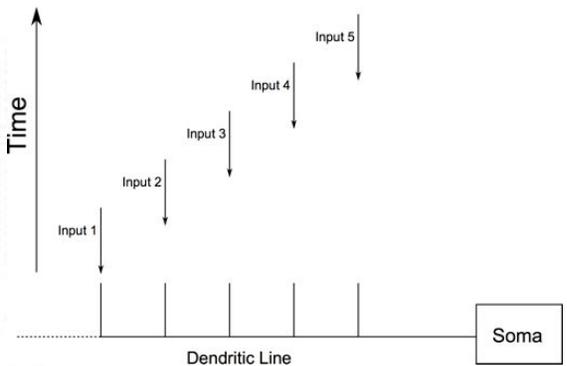
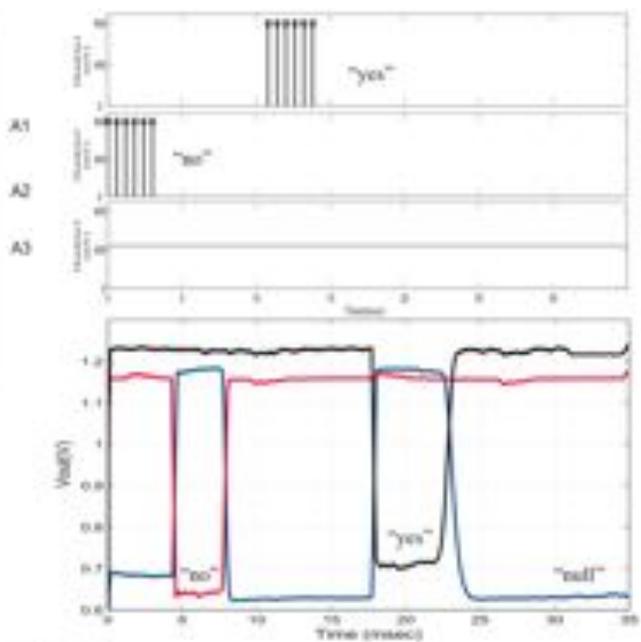
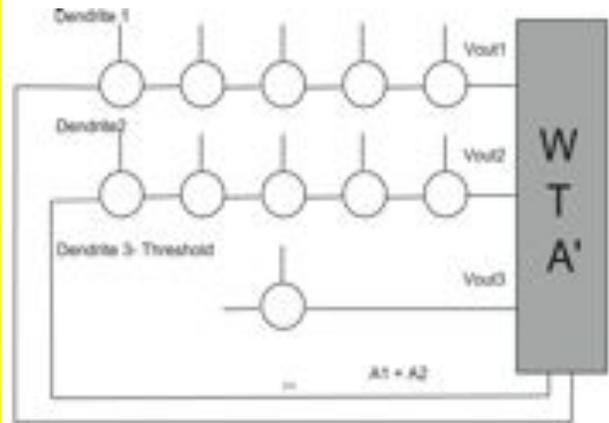
Power Efficiency Scaling



- Largest supercomputer (~ 3000TMAC) is 10^4 factor smaller than required for neural computation (~ 10^7 TMAC)



Dendrite-Model Wordspotting Classifier



Sensitivity to particular delay window
→ coincidence detection

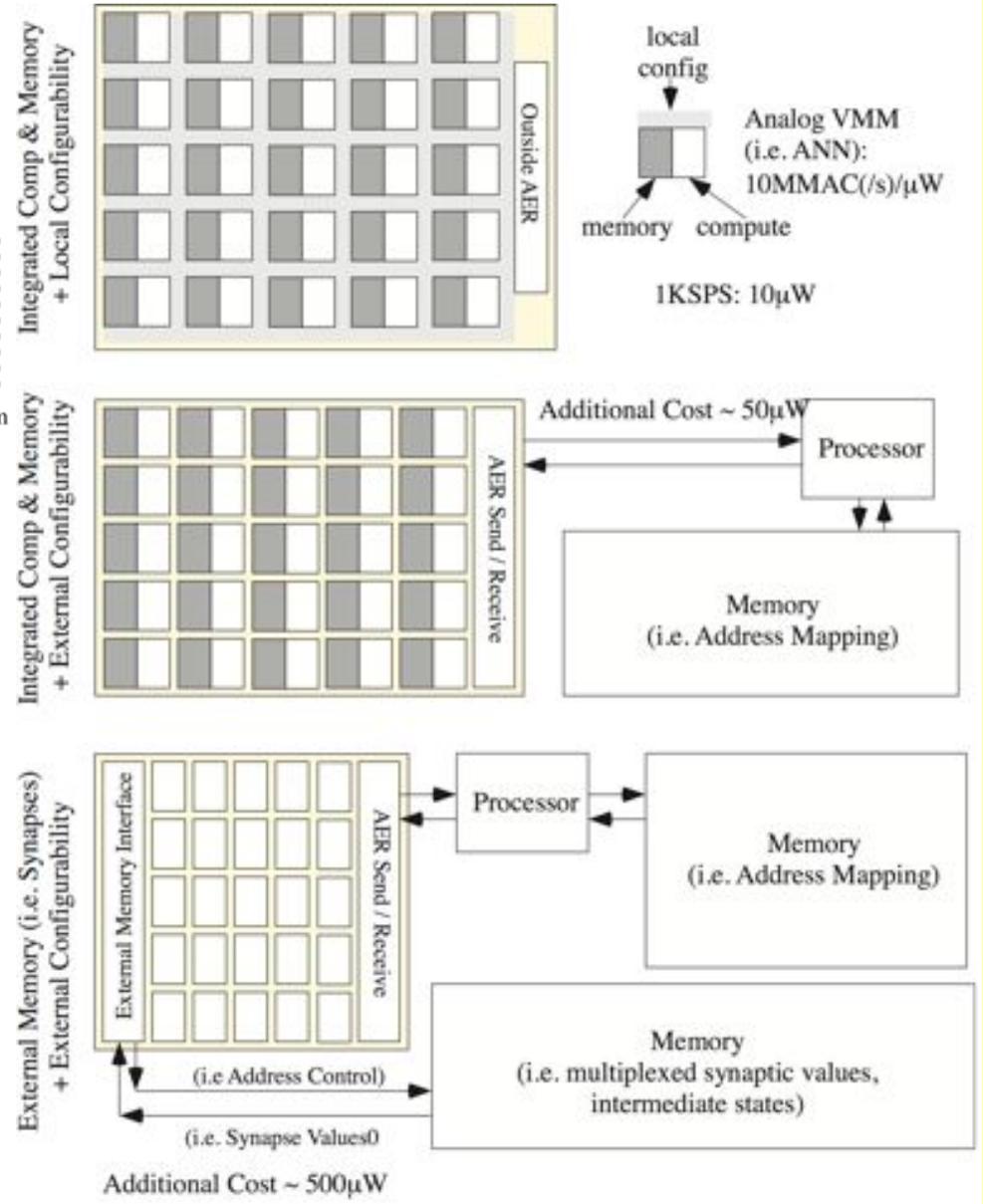
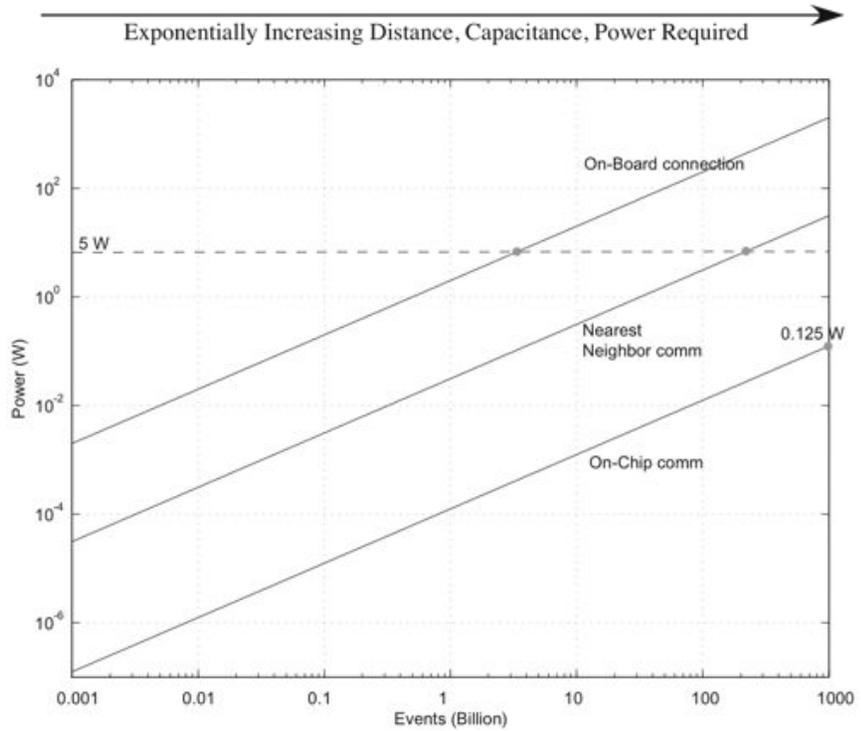
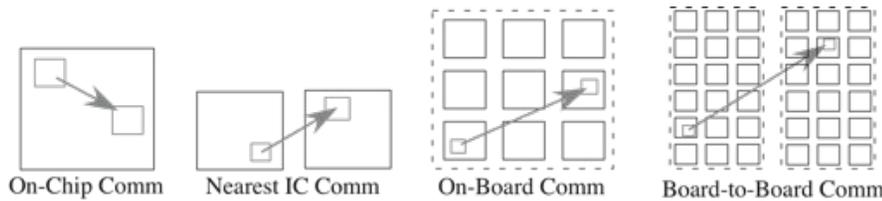
Decreasing event delay on resulting soma signal

[Ramakrishnan, et. al, 2012]



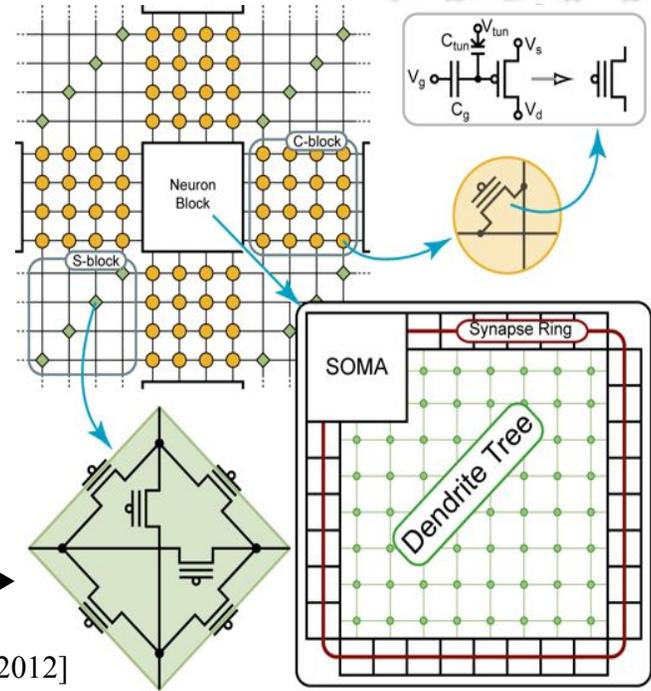
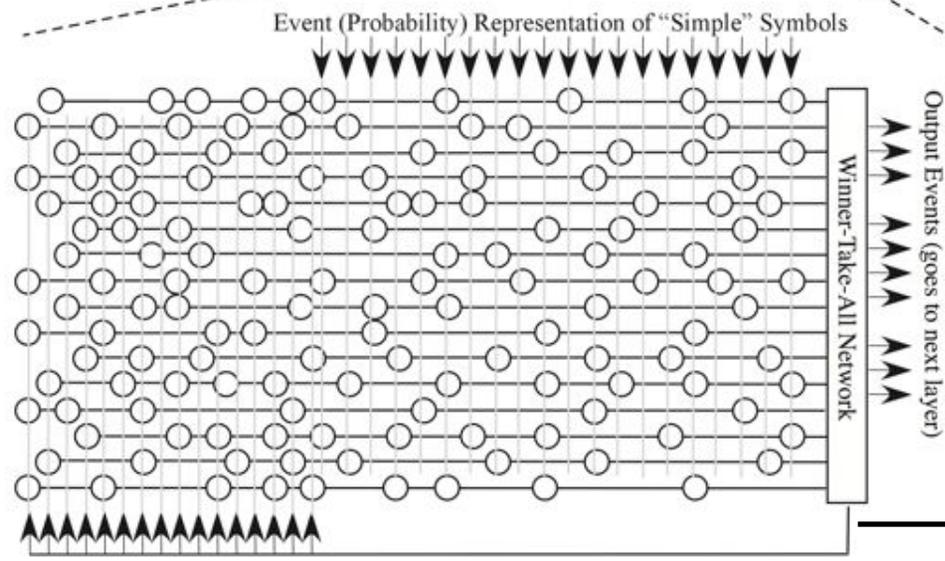
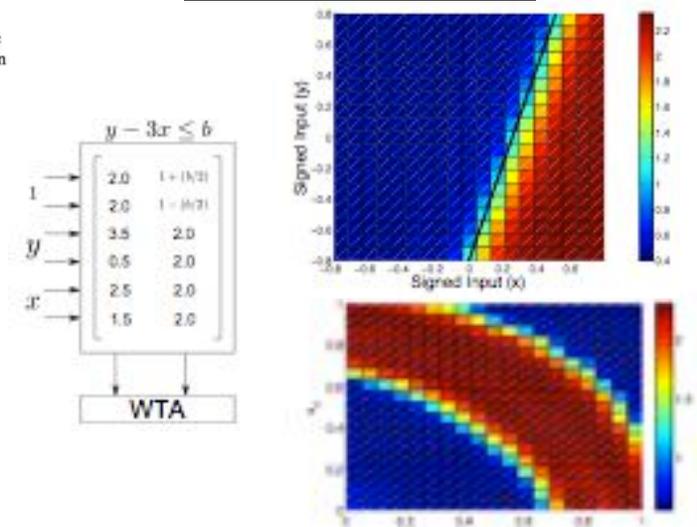
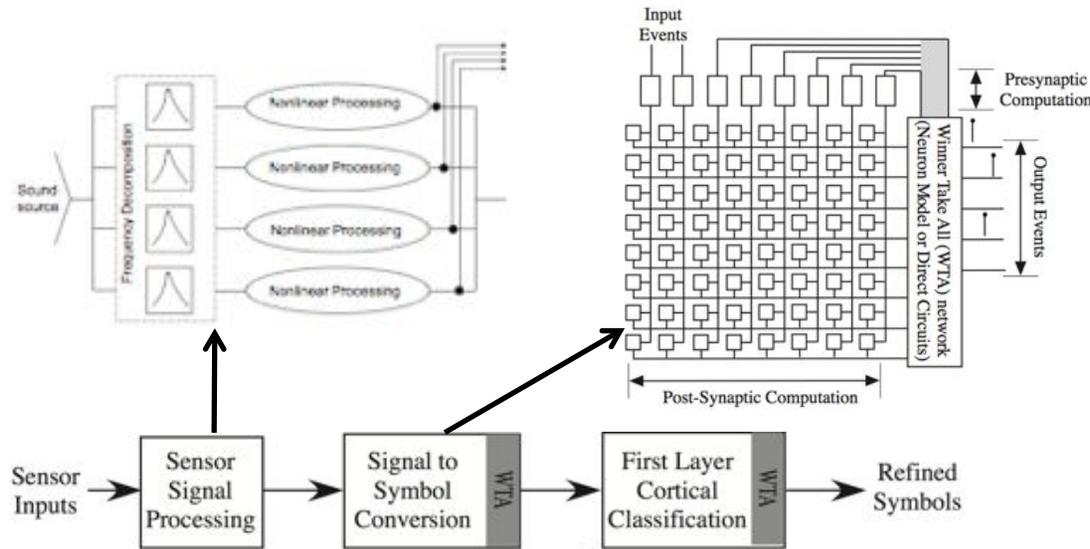
Scaling of Neuromorphic Computation

- Neurobiology: constrained by energy cost of communication (local comm is critical)
- Silicon systems also constrained by energy cost and complexity cost of communication.



Neural Classifier Approaches

VMM-WTA block



Neurons are configurable processors
(synapse → dendrites → soma)

[Ramakrishnan, et. al, 2012]



Summary of Learning & Adaptation

Neuron Implementations: FG based learning
 Synapse types: excitatory, inhibitory, NMDA
 Experimentally demonstrate

Early (unpublished) data on receptive field development, different event encodings

Application sets IC data flow speed
 (minimizing memory blocks)

Learning investigations: faster than real time
 - same structure, direct scaling of timescales

Why Learning? How about loading cortex...

Load time	15 minutes	1 day	10 days
Communication Rate	11.3Tbit/s	116Gbit/s	11Gbit/s
Power	10.4kW	109W	11W

Memristors: Multi-timescale Adaptation

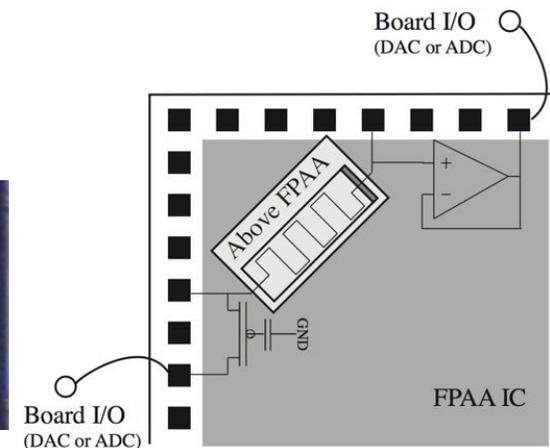
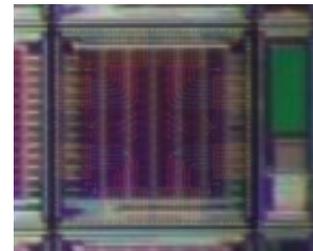
Synapses: Not competitive with 1T EEPROM cell for density
 (32nm cell ~ 50nm x 50nm device)
 Memristor arrays hard (Liu: 40 x 40)

Neurobiology: multi-timescale devices

- modulation timescales from 1s to hours
- adaptive FG allows approach
- memristors enable capability

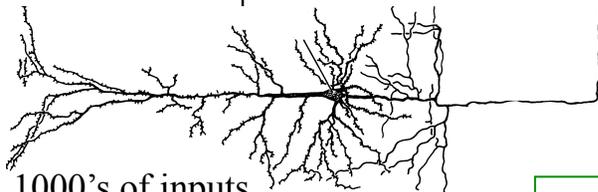
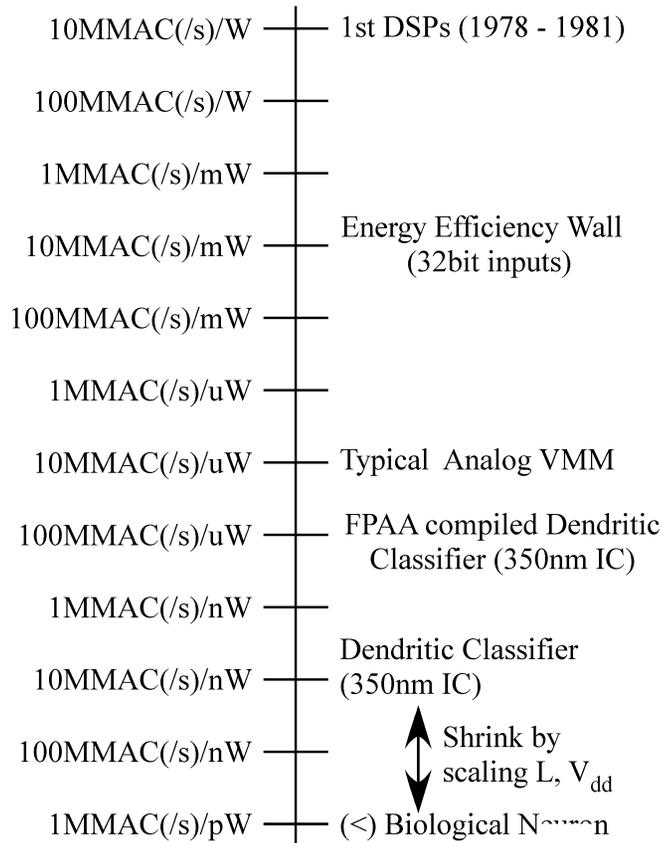
$$I = W R V, \quad \epsilon \tau \frac{dW}{dt} = f(V(t))$$

FPPA + Memristor (A)



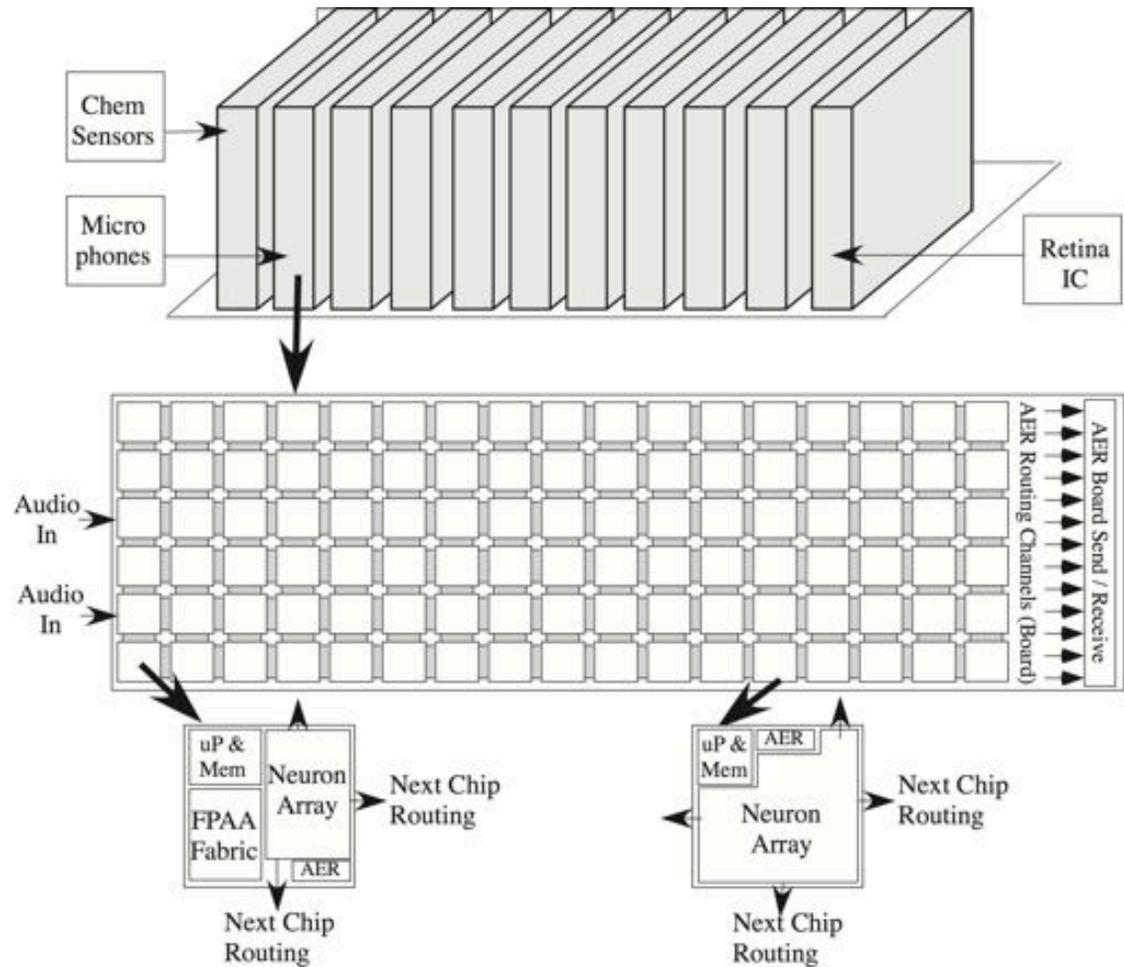
Neuromorphic → Higher Power Efficiency

Power Efficiency Scaling



1000's of inputs,
1000's of channel populations,
one output

Building Silicon Cortex



10nm: ~4M Pyramidal cell neurons → \$20M IC cost for Cortex
(digital: 1000 in parallel) (~100k chips)



