

Ultra-Low Power Embedded IoT Machine Learning enabled through FPAAs

Professor Jennifer Hasler
Georgia Institute of Technology
<http://hasler.ece.gatech.edu>



Embedded Machine Learning

Machine Classification and Learning:
What we currently think

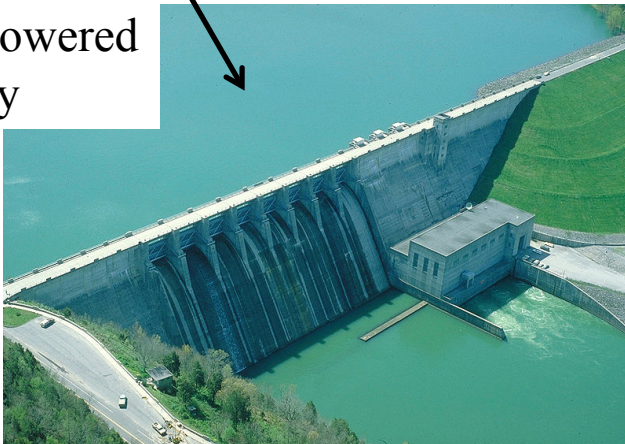


(IBM Sequoia)

~ 10MW

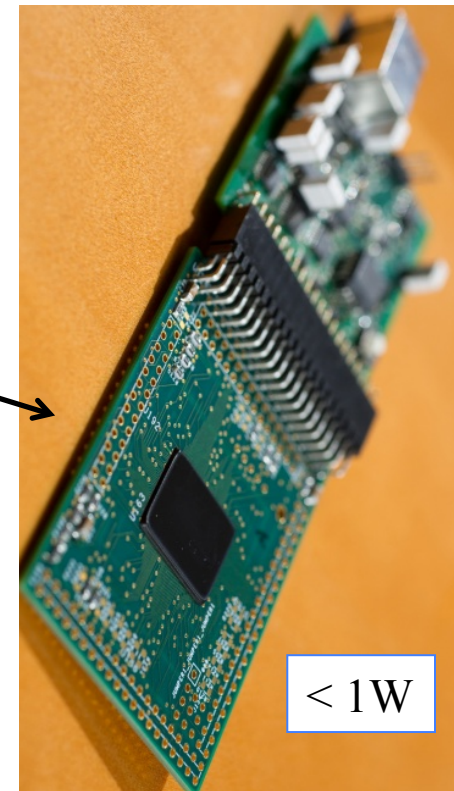
10^3 TMAC(/s)

Powered
by



Machine Classification and Learning:
What we want

Presentation begins
to build this
transformation



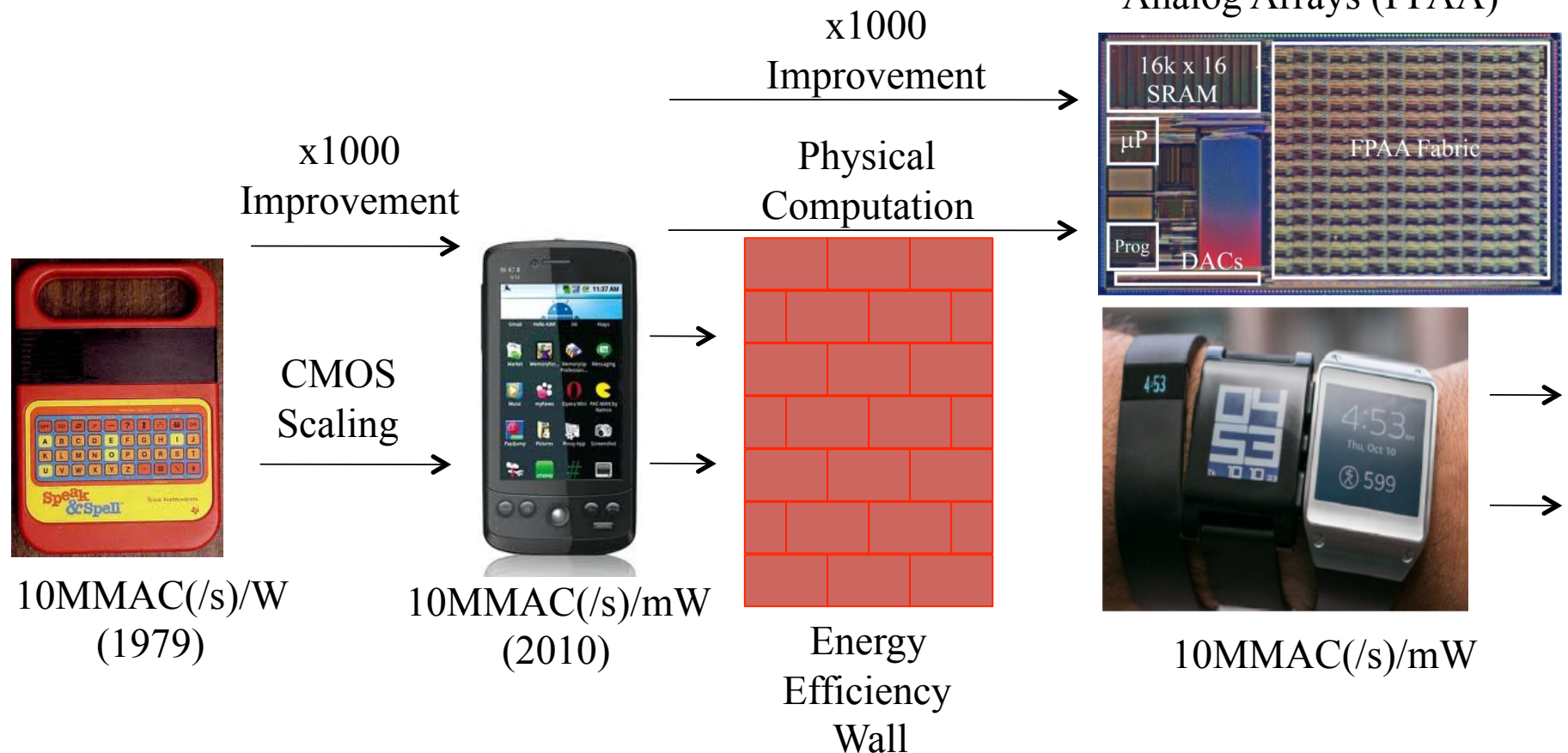
< 1W

10 TMAC(/s)



Physical Computing → Increased Computational Efficiency

Large-Scale Field Programmable Analog Arrays (FPAA)

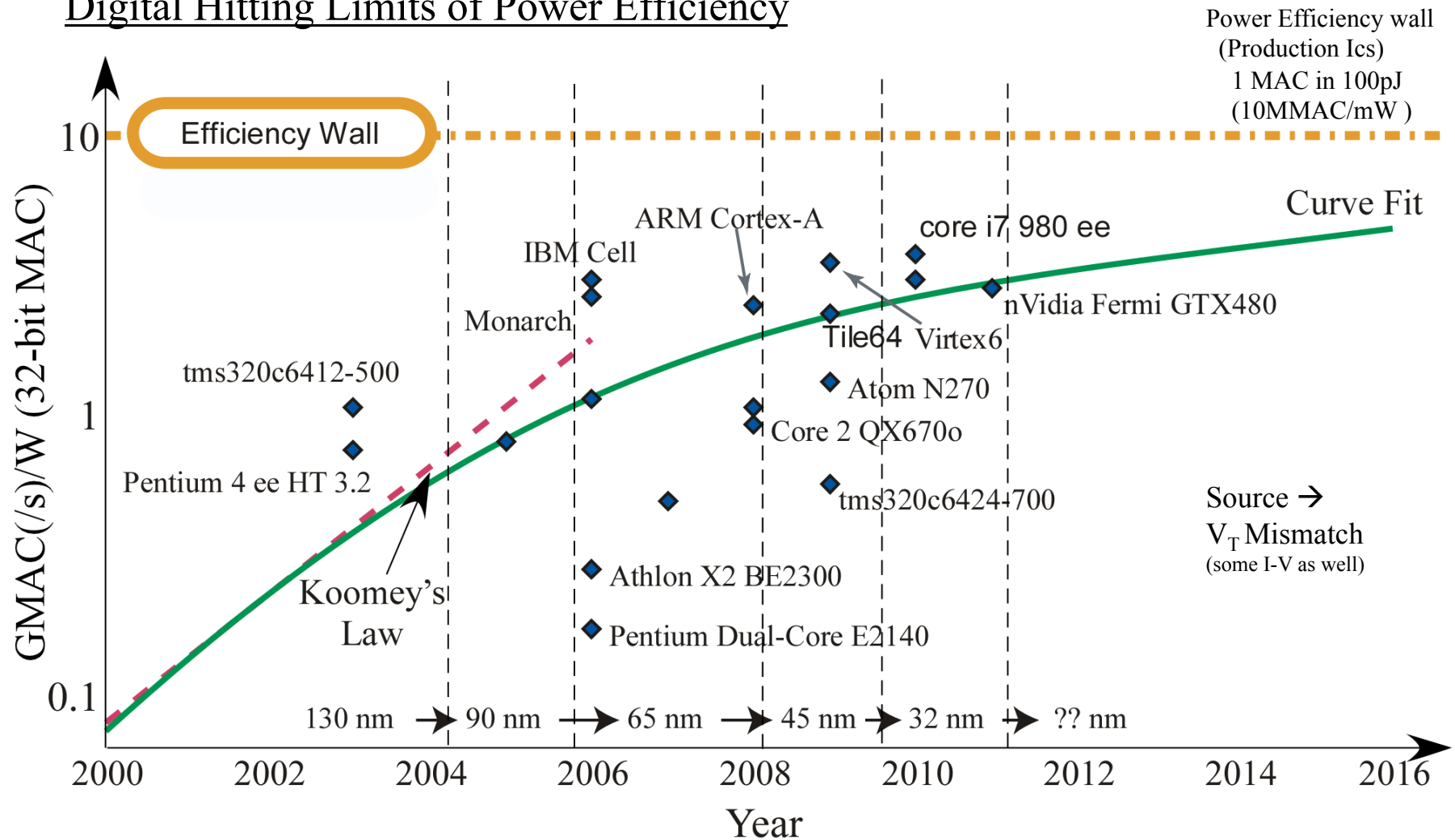


Wearable Devices Require more Efficiency



Why Analog (Physical Based) Processing?

Digital Hitting Limits of Power Efficiency



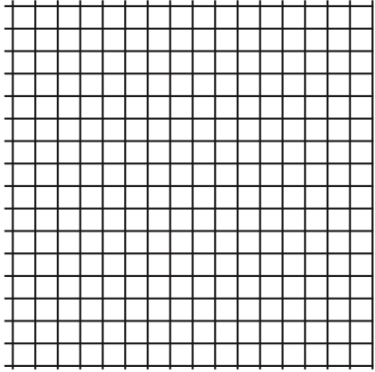

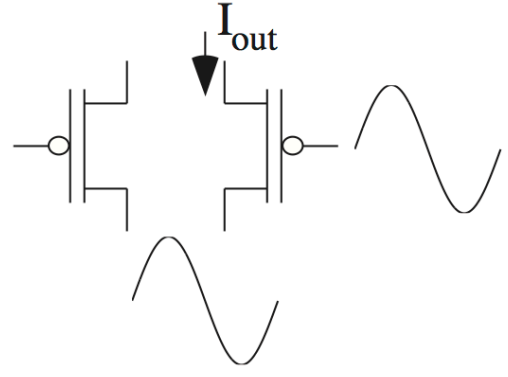
Results created its own DARPA program

Battery Energy Density: x10 over 40 years



Why Analog (Physical Based) Processing?

Mead Hypothesis (1990): Analog x1000 efficiency improvement

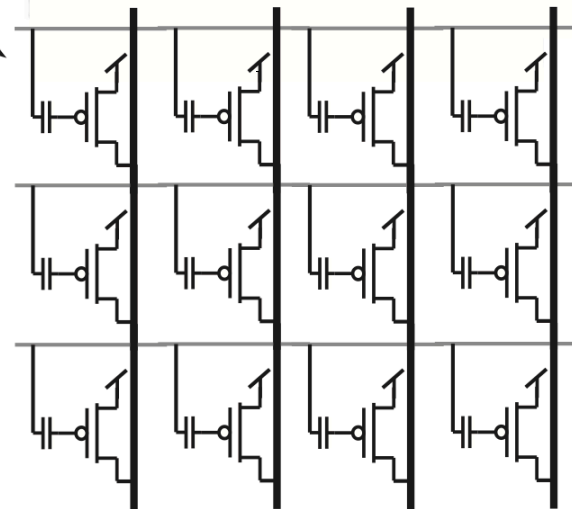
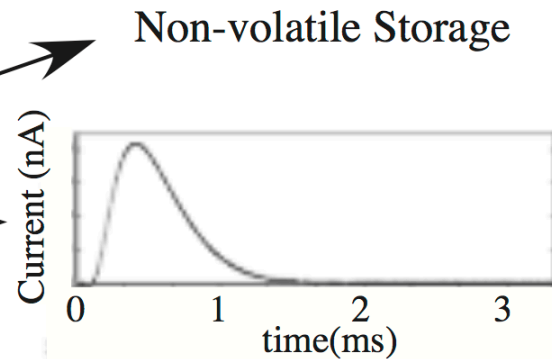
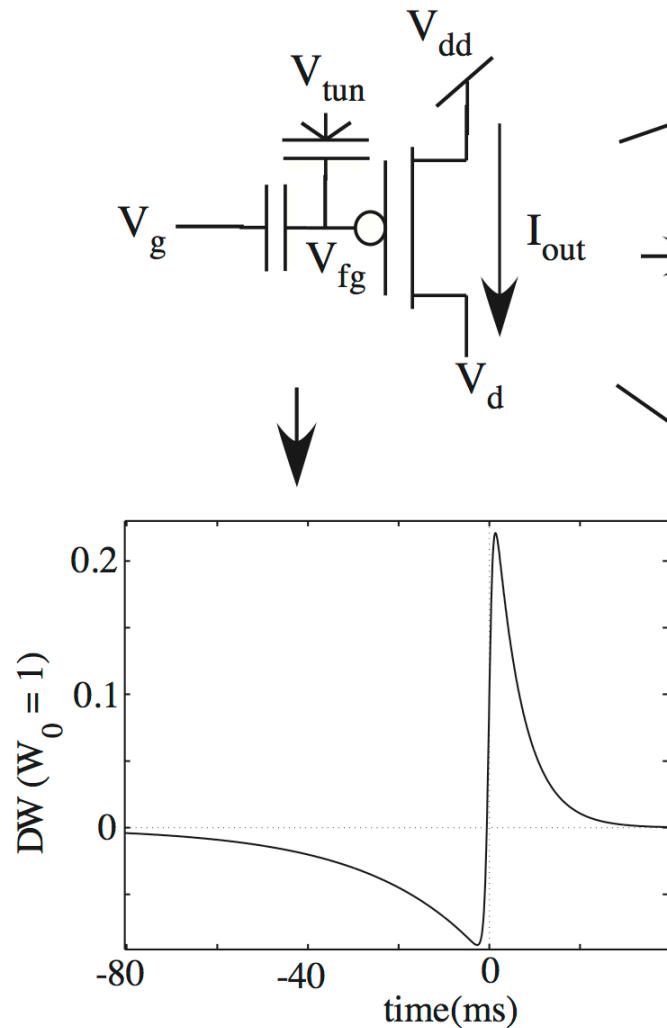
	<u>Digital</u>	<u>Analog</u>
Multiplication (digital: 16bit)	 20 transistors 	
Energy/ operation	x1000	x1
Size	x100	x1

- Analog (VMM): ~ 100 fJ / MAC (10MMAC/ μ W) @ yield
- Other Analog SP similar: Freq Decomp / Analog FT
VMM, GMM
Classifiers
Adaptive Filters

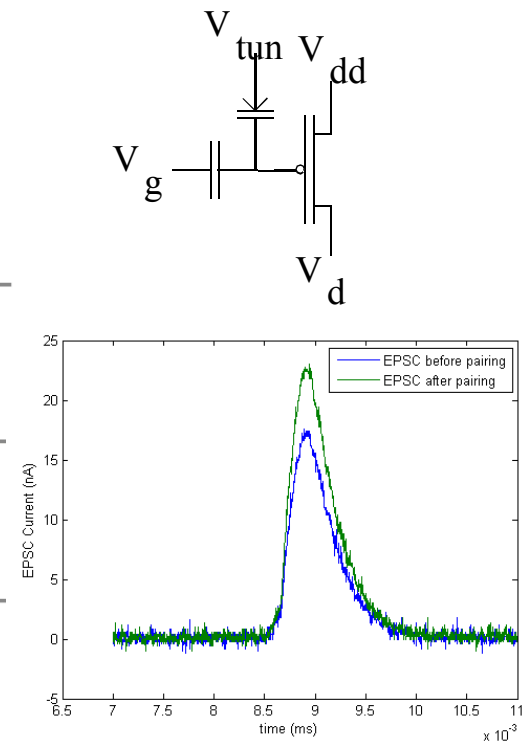


Single-Transistor Learning Synapse

Floating-Gate Circuits: Nonvolatile storage, computation, programmable, adaptable



130nm STDP synapse data

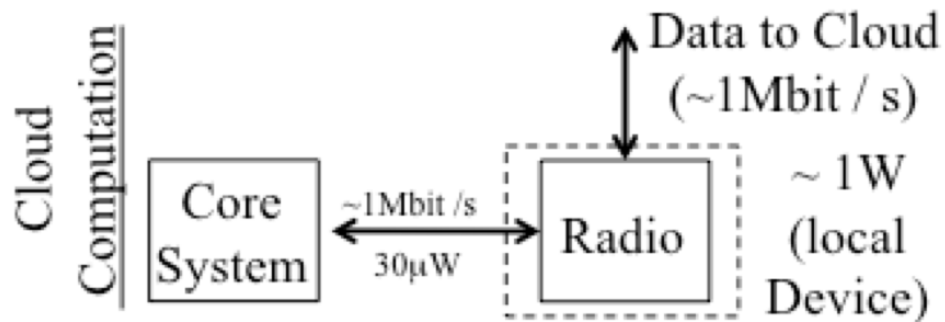


Si CMOS approach can achieve densities while avoiding issues with device integration with Si

[Hasler, et. al, NIPS 1994, BMES 1994, and later papers]

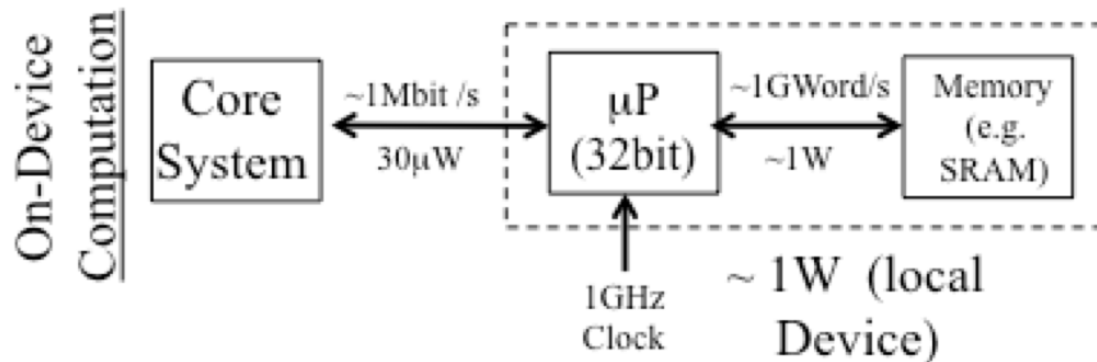


FPAA vs. Embedded / Cloud Computation



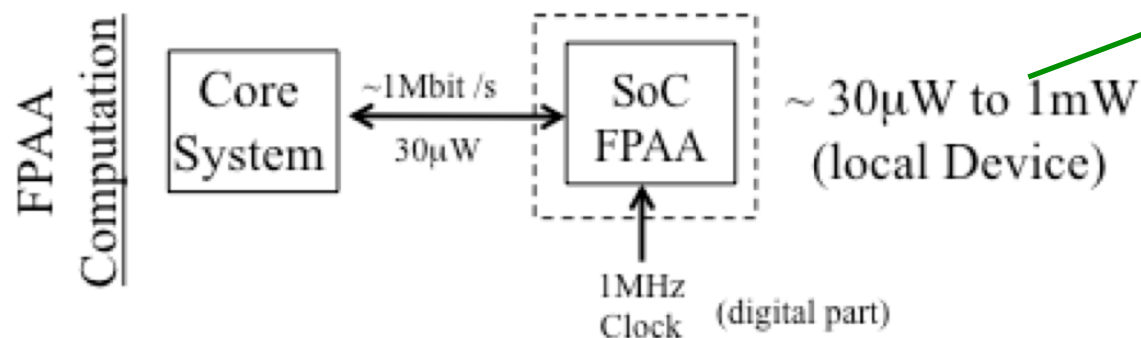
If cloud is “free”....

If cloud is “drops”,
then disaster



If both → more cost

SoC FPAA decreases energy
and resulting complexity



Physical Algorithms empowering
wearable devices

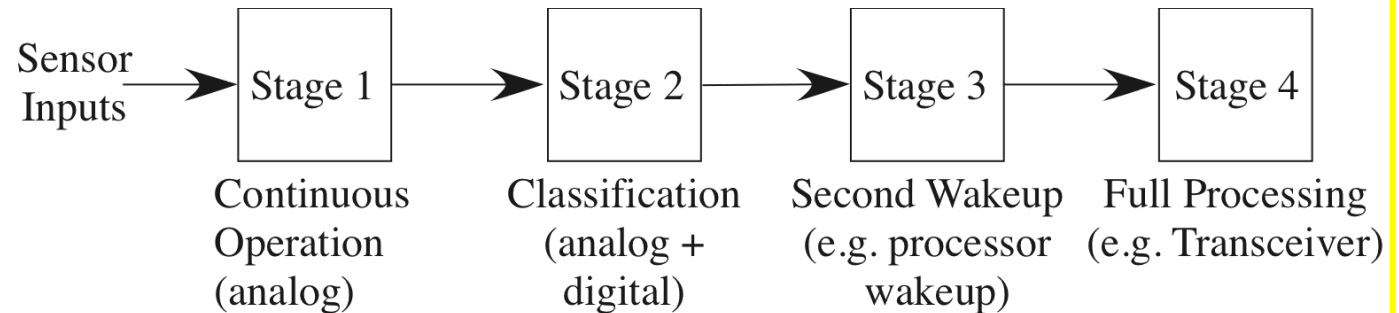


Energy
Harvesting:
 $10\mu\text{W} / \text{cm}^2$



Where to use ultra-low energy?

Sensor node < 100 μ W



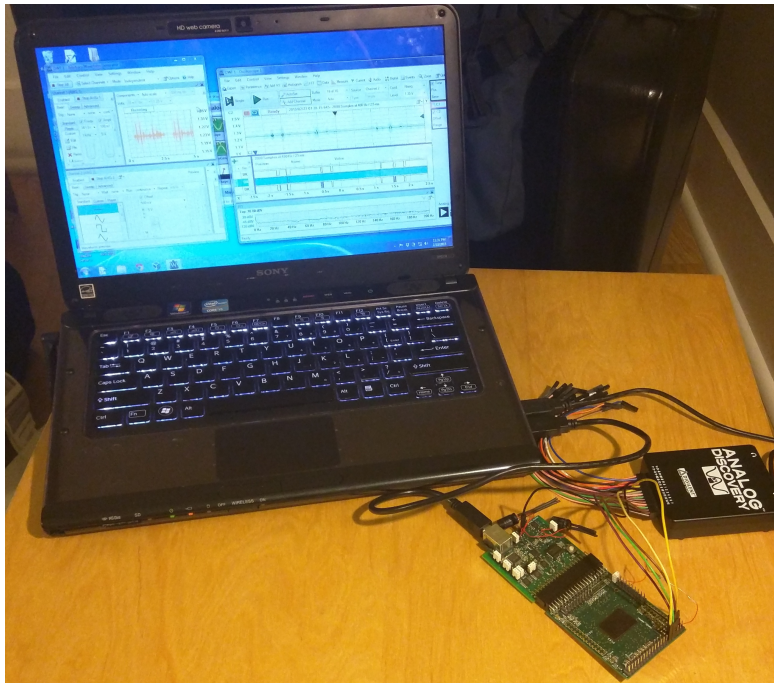
x1000 energy improvement utilizes context-aware physical computing to enable 100 μ W end-to-end sensor node.

Average on time	100%	1-3%	0.1-0.2%	0.01%
Operating power	1 to 10 μ W	\sim 100 μ W	1-5mW	30-100mW
Total(max) Power	10 μ W	3 μ W	10 μ W	10 μ W
Digital		<1MMAC/s	\sim 10-20MMAC/s or 20MHz clock	Transciever on
Analog	<u>10-100MMAC/s</u>	<u>1GMAC/s</u>	50GMAC/s	

← More computation near sensor
 Increasing Energy
 Decreasing Use →



Physical / Analog / Mixed-Signal Computing Exists



Command Word < $23\mu\text{W}$ power



Analog + Digital
FPAA

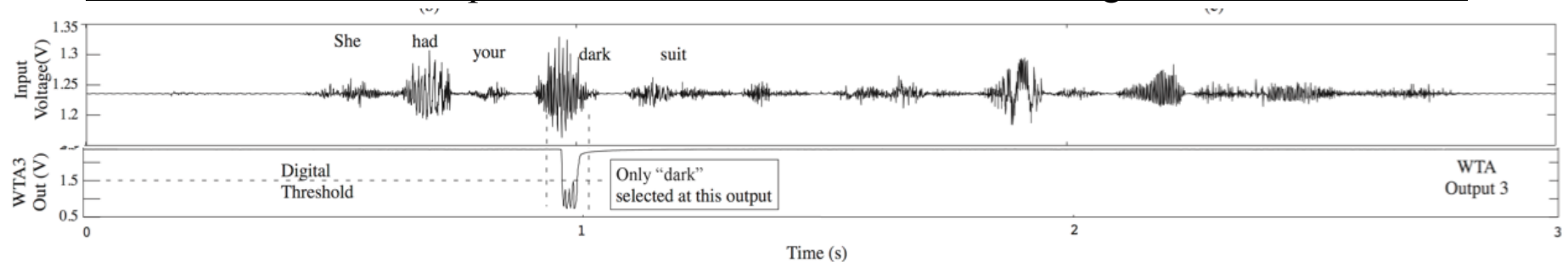
Applications in
sensors, acoustics,
imaging

On-chip Machine
learning shown
(VMM+WTA)

Capability over
multiple IC
processes

Knee-Joint Rehab < $15\mu\text{W}$

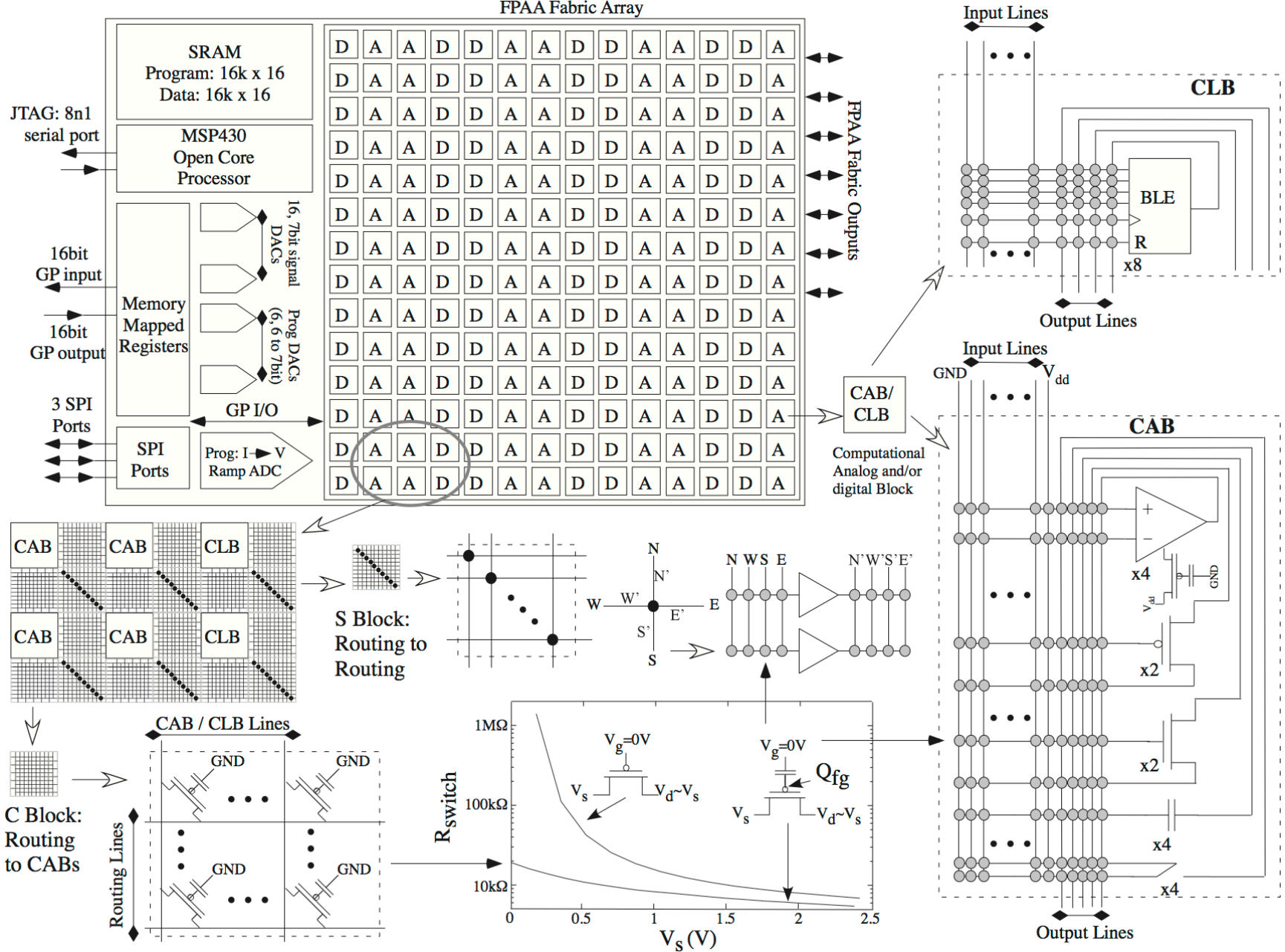
Measured Results for a phrase from the TIMIT database to recognize the word “Dark”



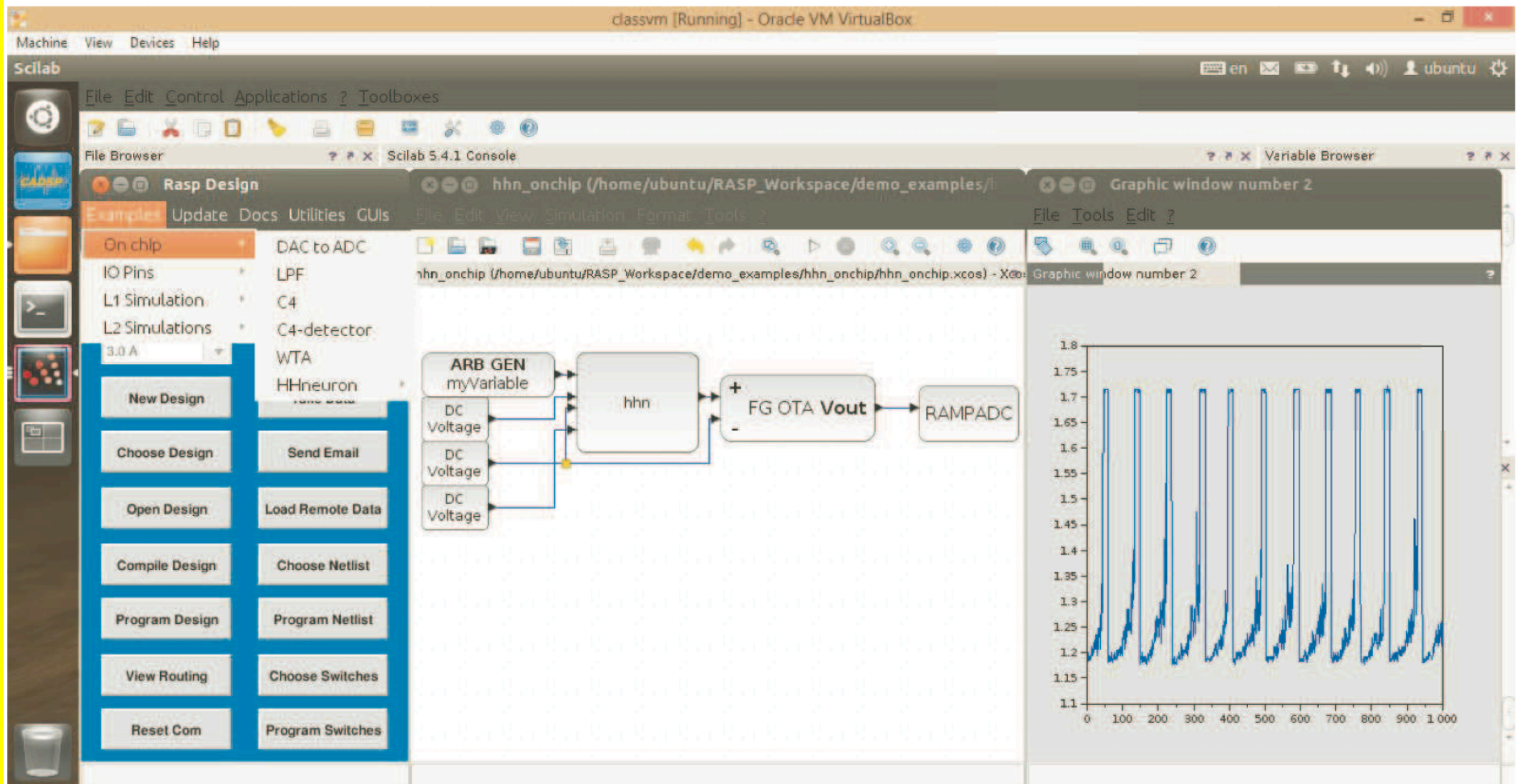
Analog SP Energy < 1000x Custom Digital SP



RASP 3.0: First SoC FPAA IC



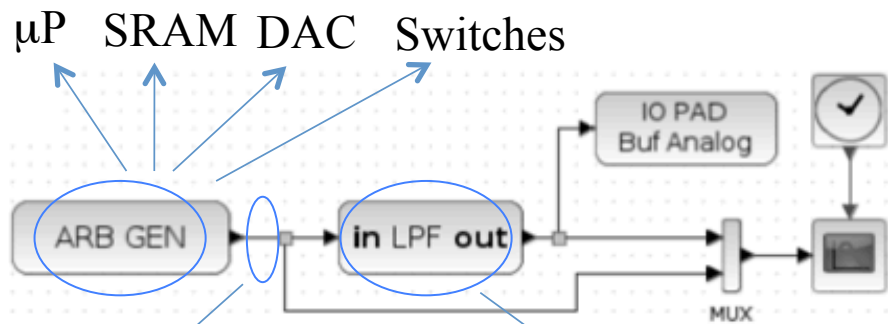
Scilab FPAA Synthesis & Modeling Tool



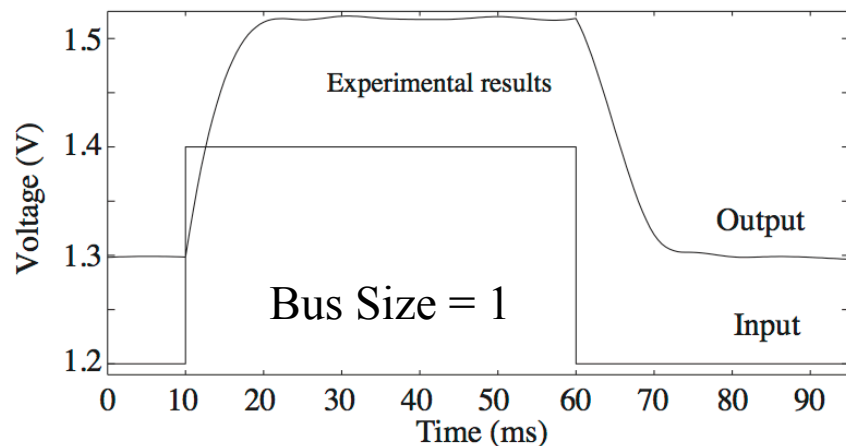
- Encapsulated in Ubuntu 12.04 VM
- Library of Components (low to high level)
- Measurement transistor channel model of HH neuron



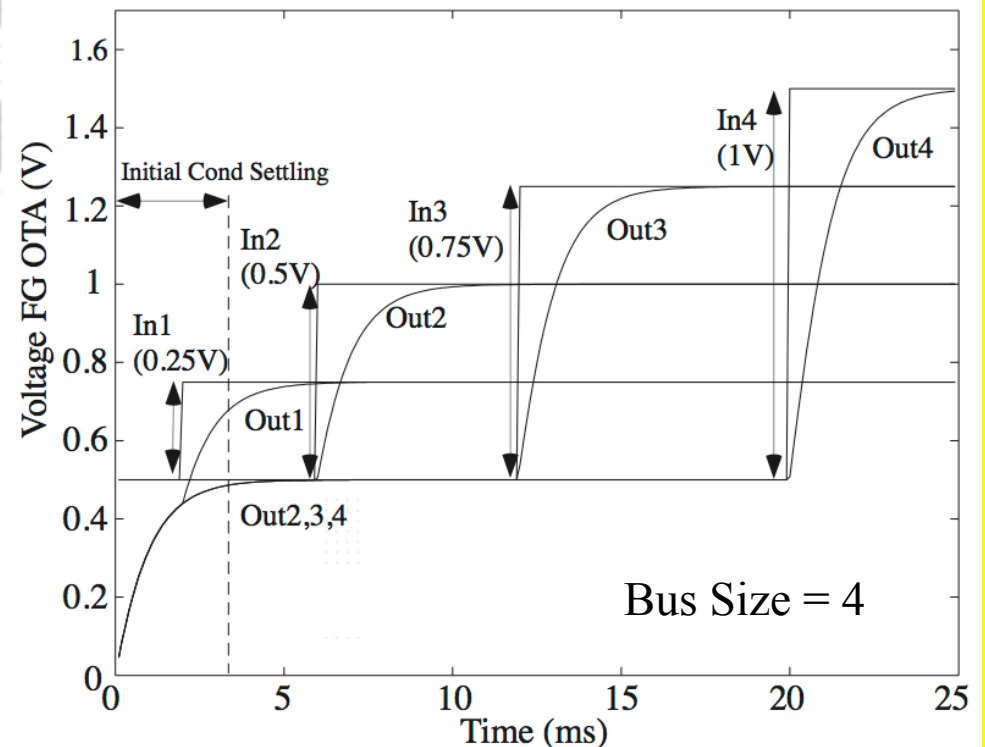
Tool: Measurement and Simulation (LPF)



Single or Bus of Wires 1 or many Filters

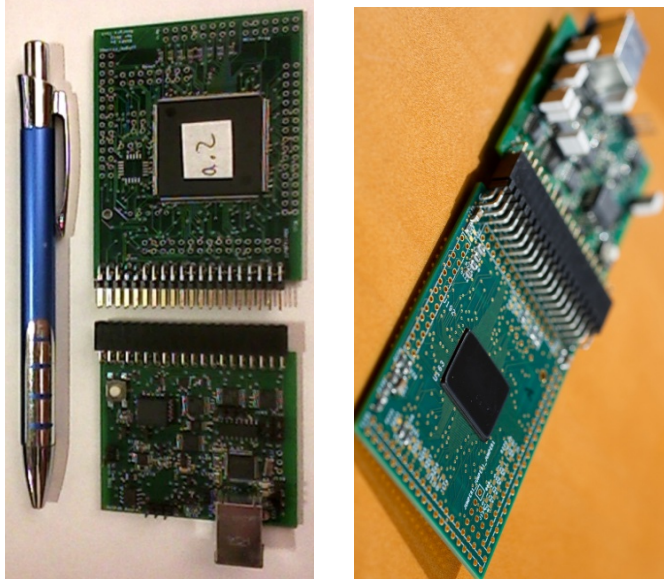


MacroModel Simulation (level = 1)



One toolset to design, to enable high level simulation,
and to compile to hardware

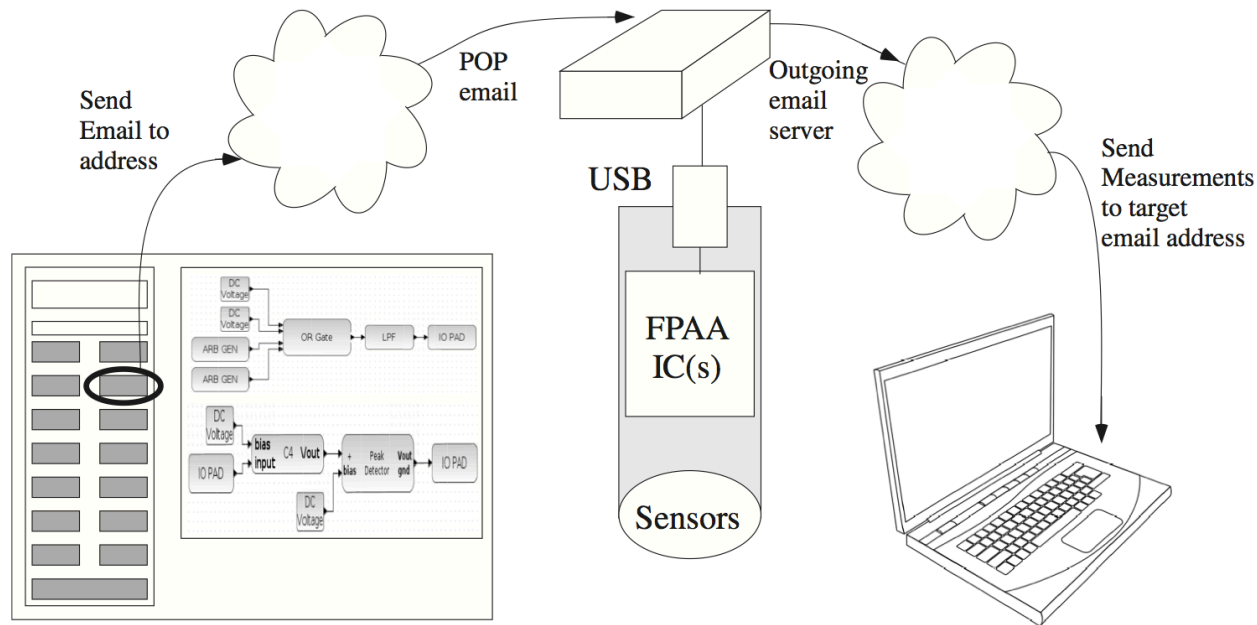




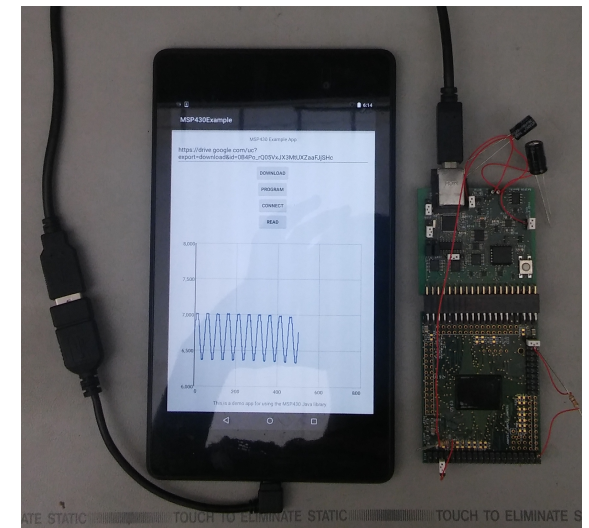
FPAA Infrastructure

- FG Programming looks like controlled download to μP device \rightarrow Straightforward to program a device (code in Scilab, Python, Java,)
- USB powered and controlled \rightarrow interfaces like a digital system

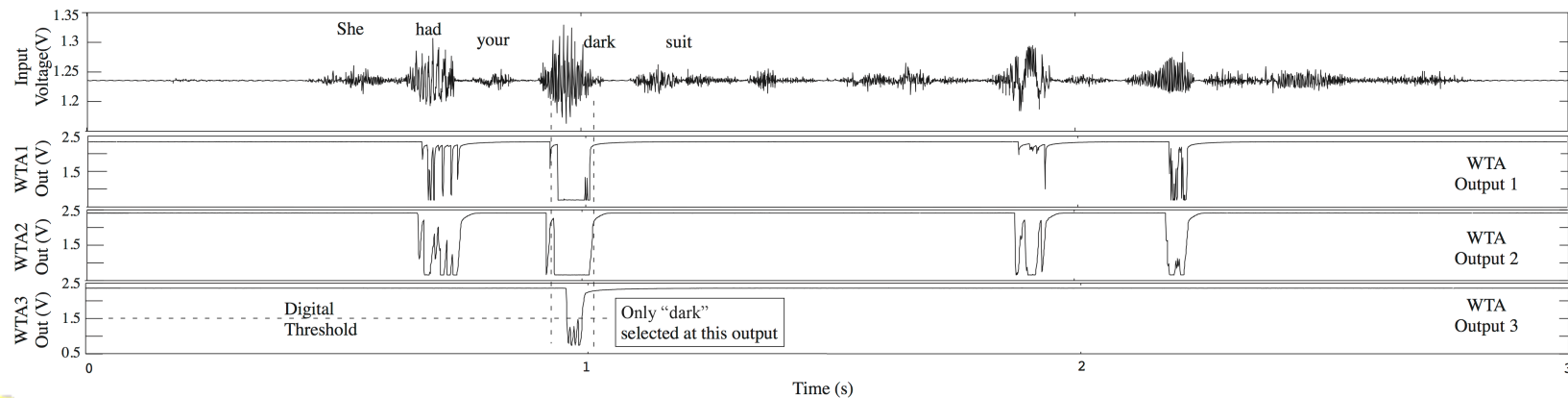
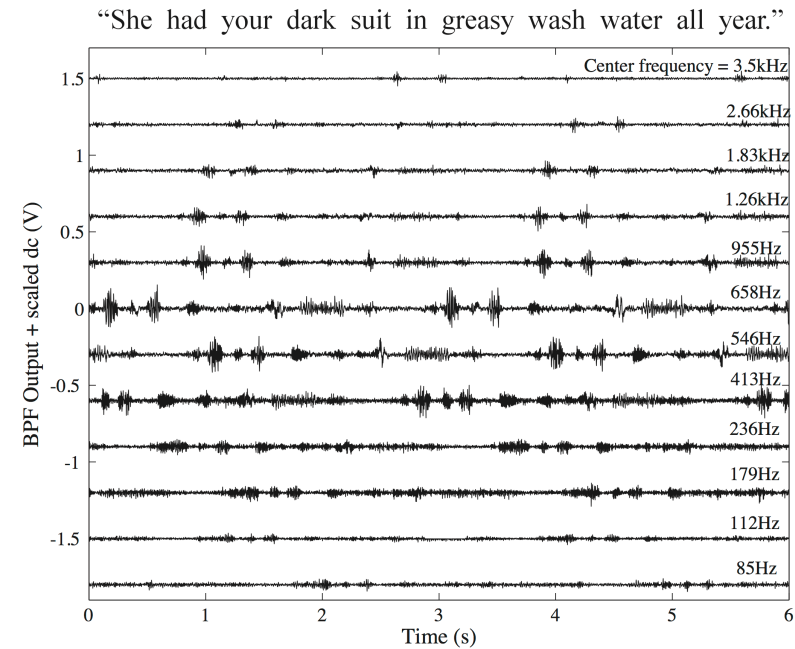
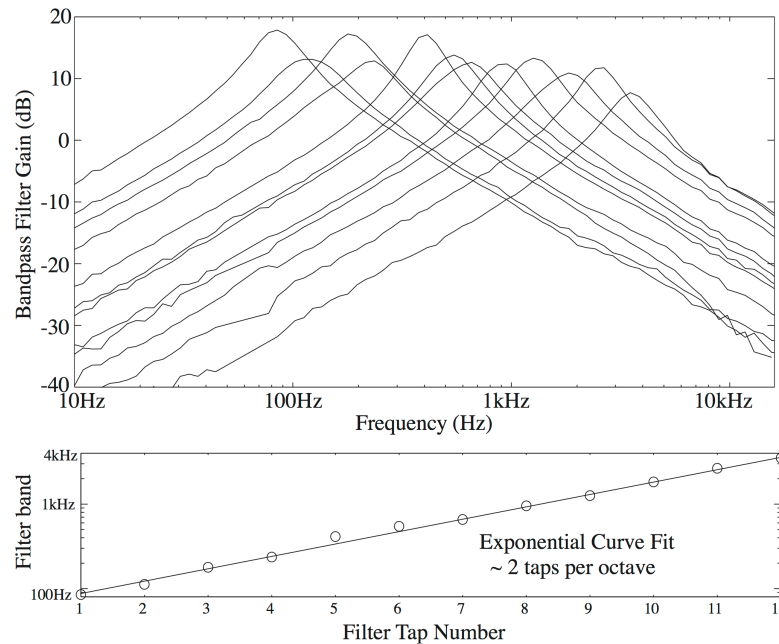
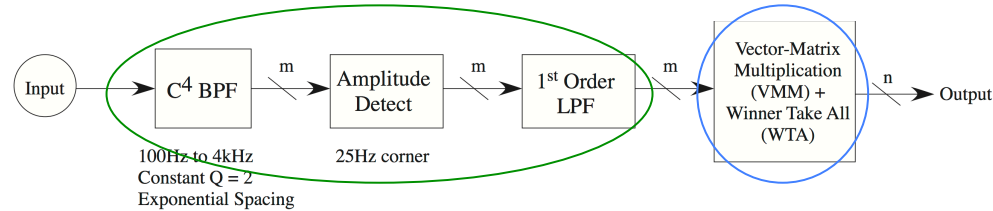
Remote FPAA System



Andriod Tablet FPAA's

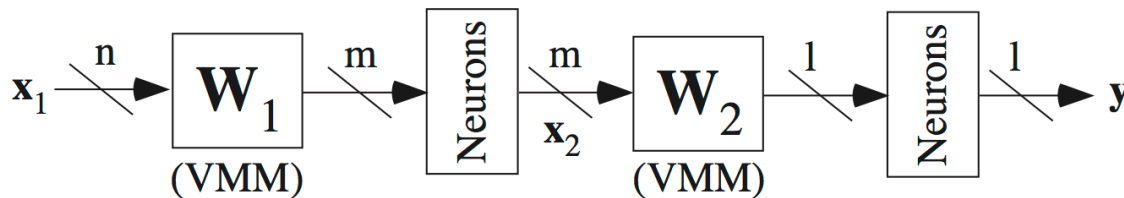


SoC FPAA Classifiers: VMM + WTA for Speech



Compiled VMM+WTA Classifier

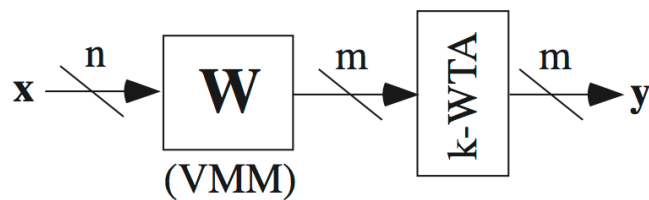
Two-Layer Neural Network (NN) Classifier



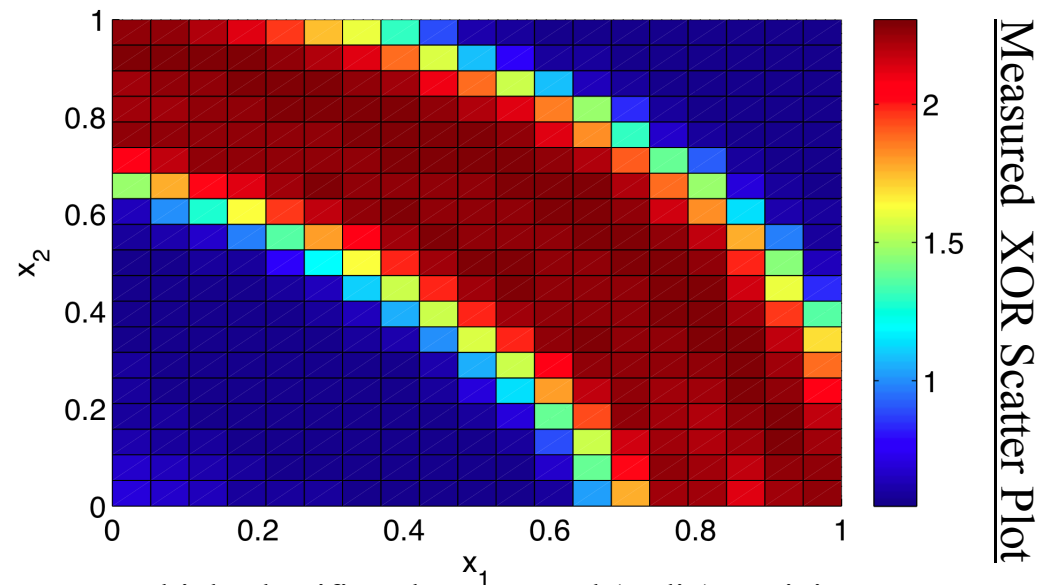
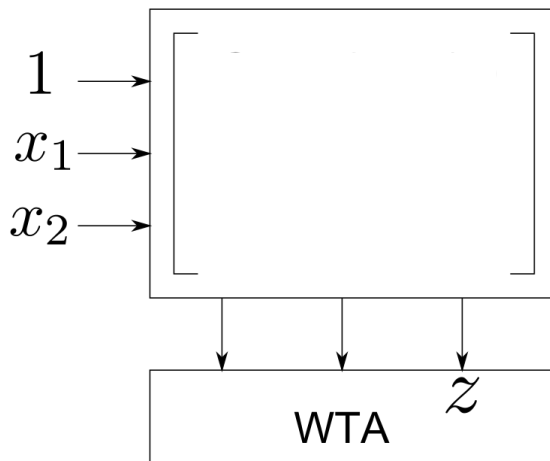
Minsky 1967: XOR classification requires more than one layer

NN was silenced for 15 years
Could solve in **two** layers

VMM+WTA Classifier



3 x 3 VMM



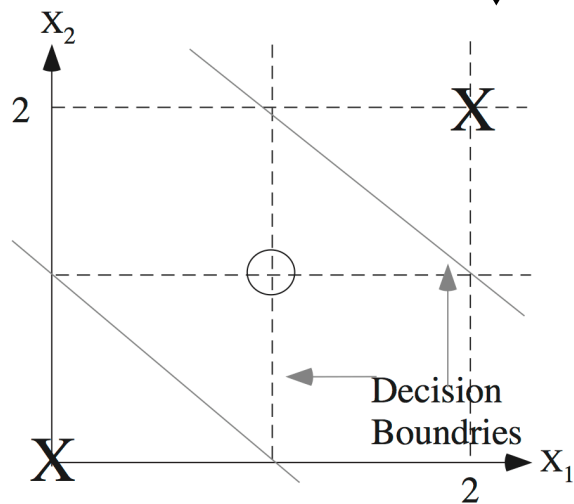
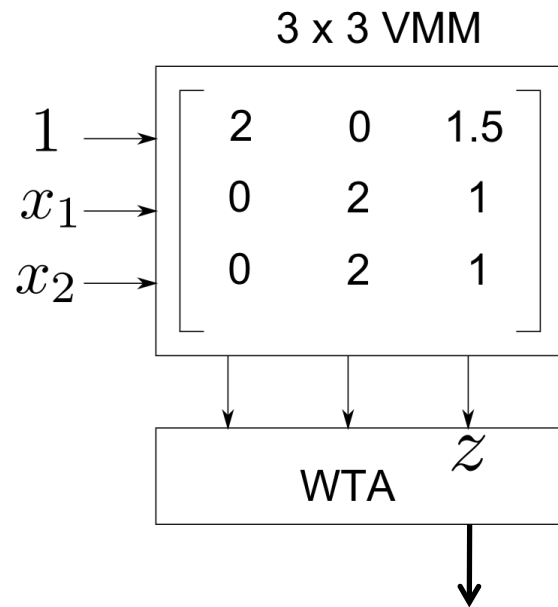
Multiple classifiers demonstrated (audio), Training

Analog, n -WTA **single** layer block can be a universal approximator (2 layer NN)

[Maass, et. al, 2000, Ramakrishnan, et. al, 2013]



Compiled VMM+WTA Classifier

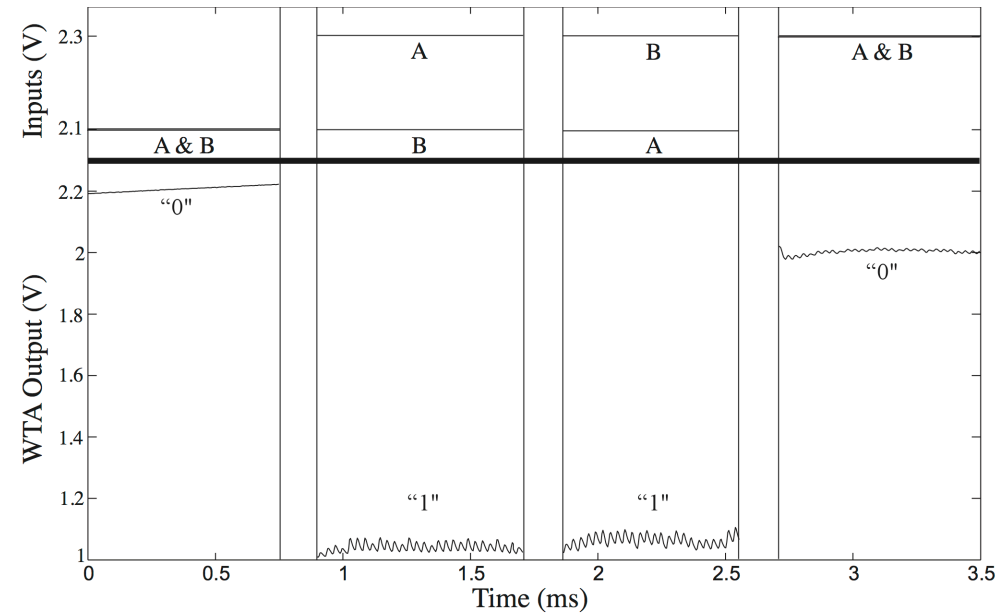


Offsets

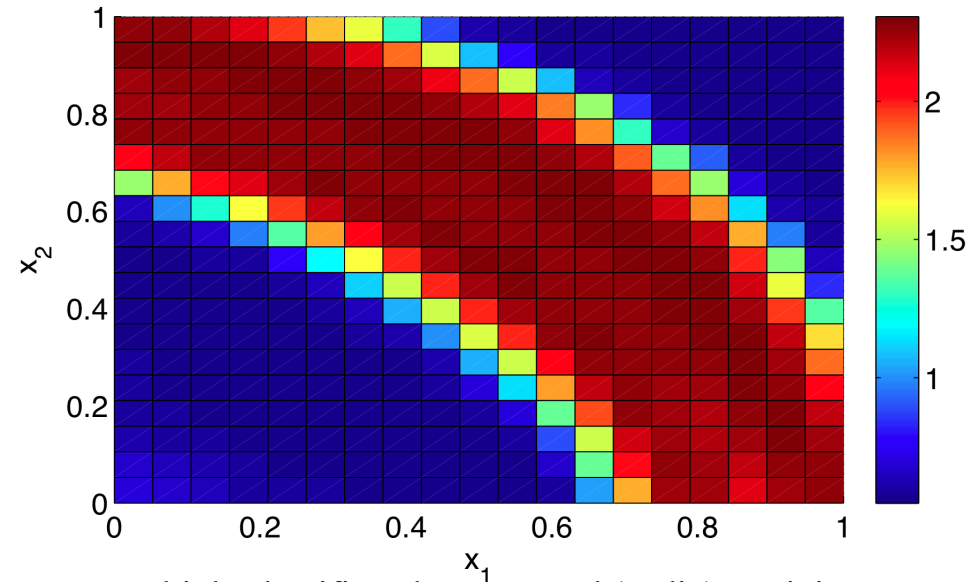
-2

-1/2

0



VMM+WTA on SoC FPA

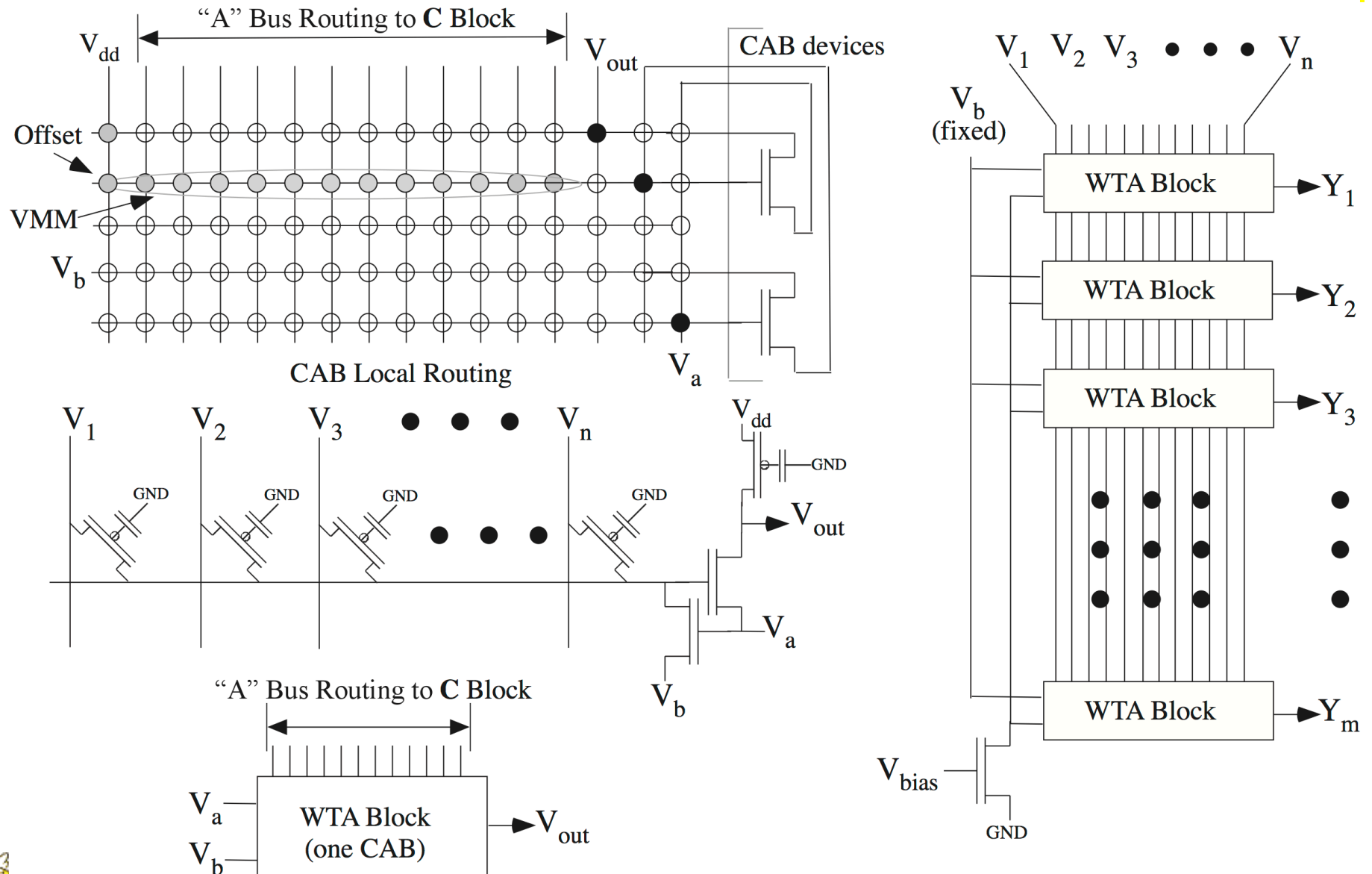


Multiple classifiers demonstrated (audio), Training

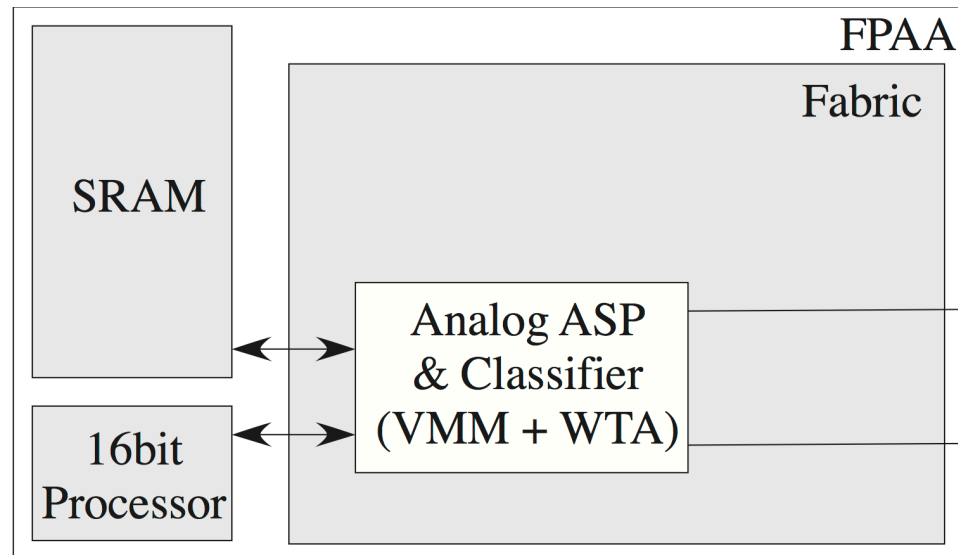




Physical FPAA Classifier Floorplan

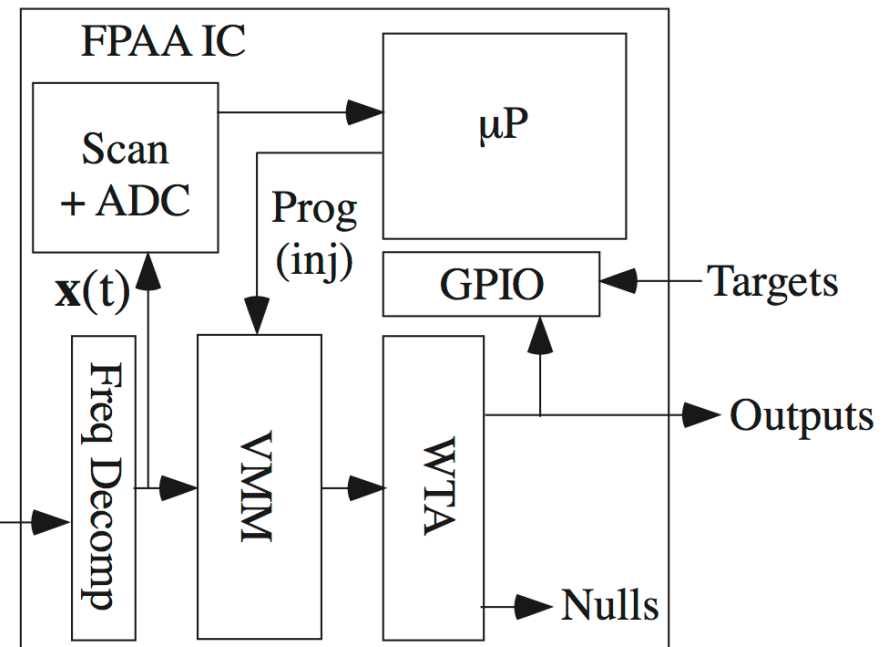
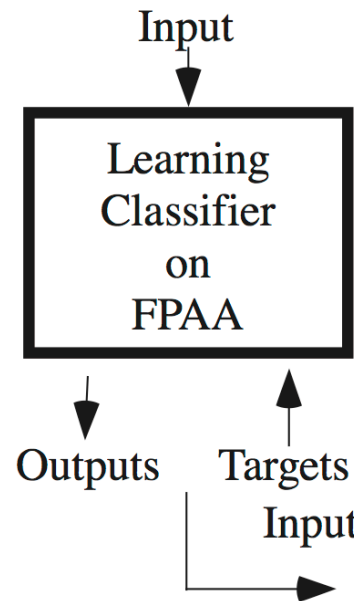


Developing FPAA Adaptation

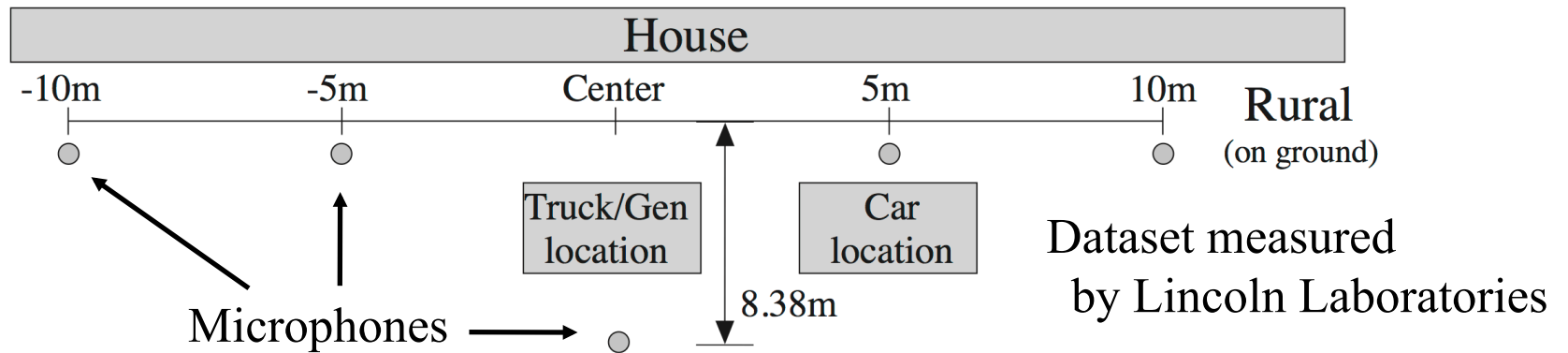


- Analog, Digital, and μP
- FG Programming uses μP device (Batch)
- Analog classifier, ultra-low energy datapath

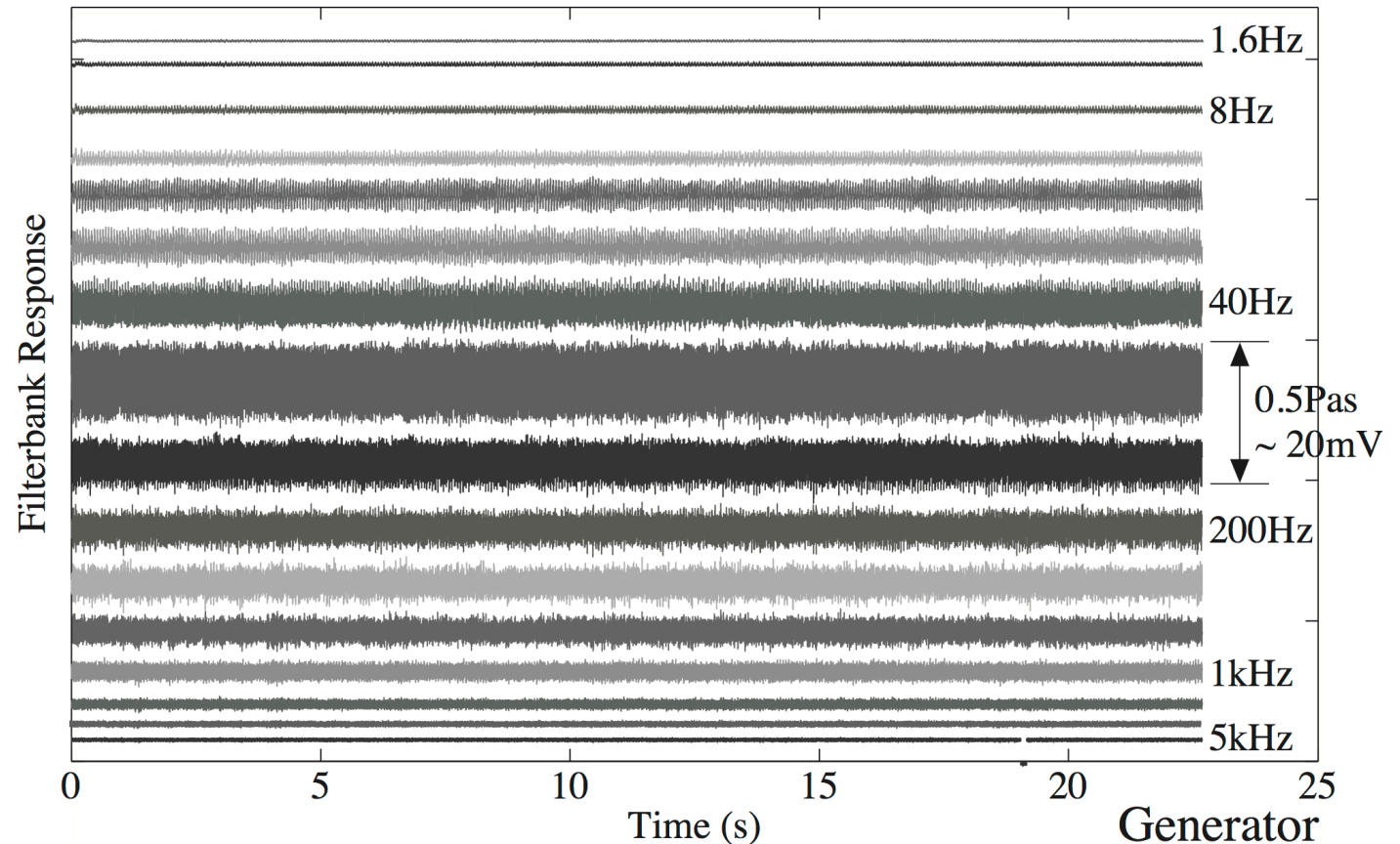
- Microphone input to classification
- statistics: analog or digital
- infrastructure and ADCs



Acoustic Classification Training Data

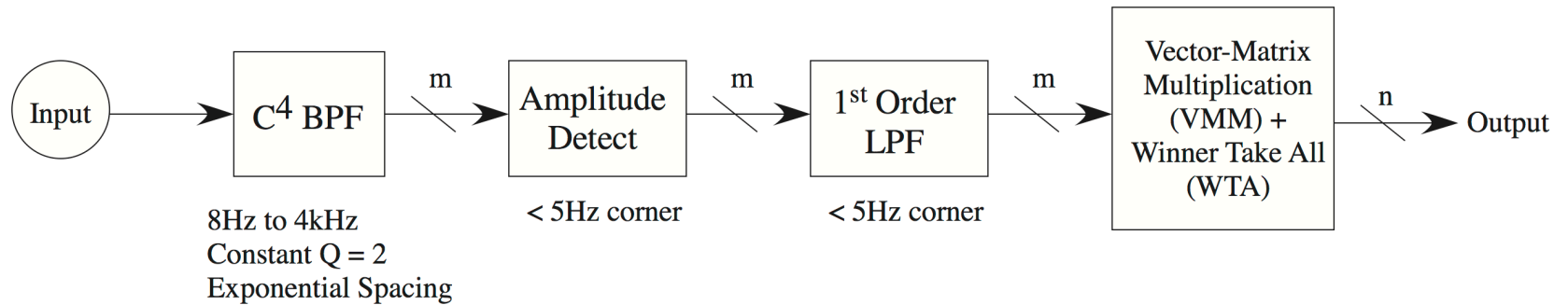


- Two minutes sensor recording (simultaneous)

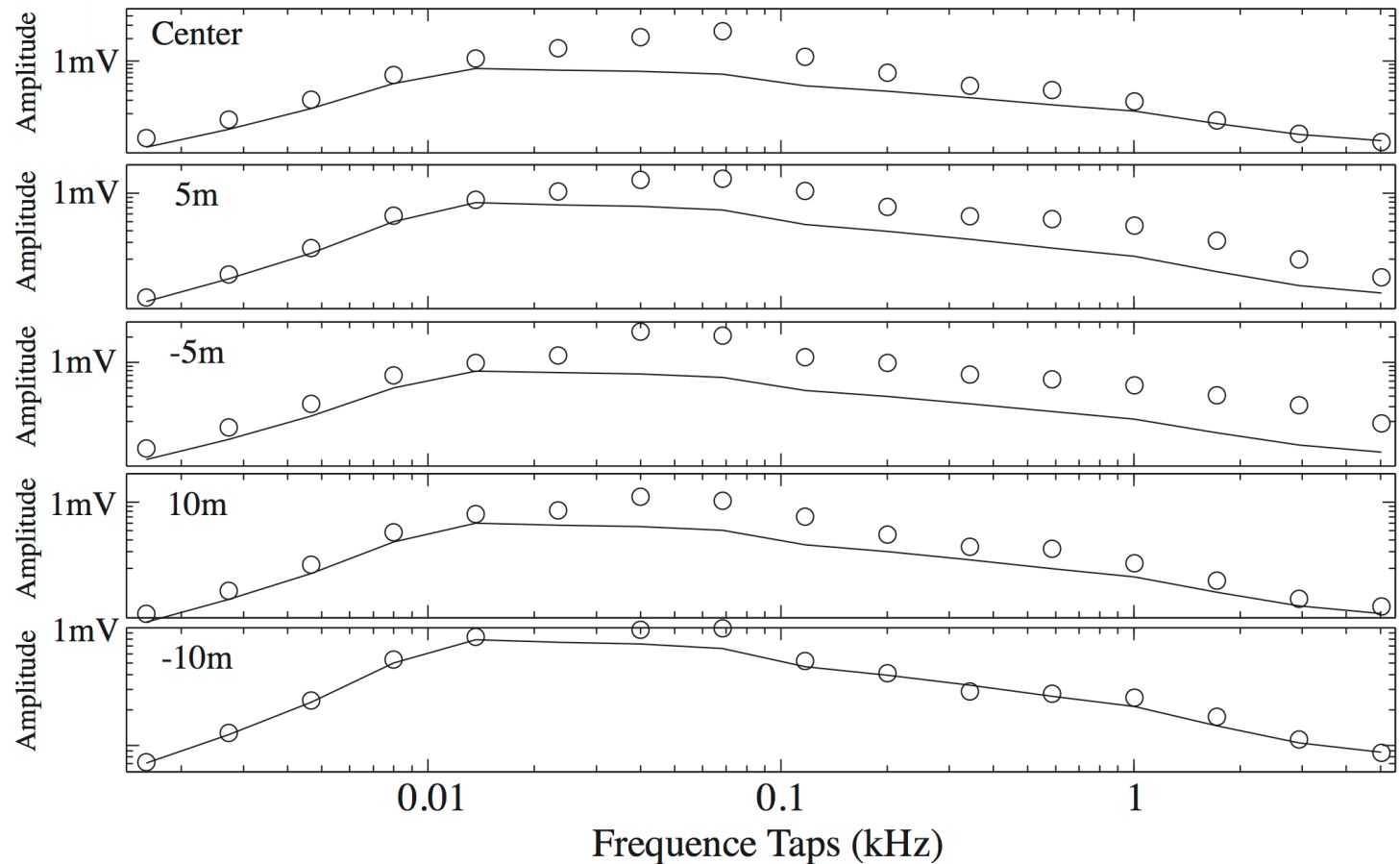


Acoustic Classification → FPAA

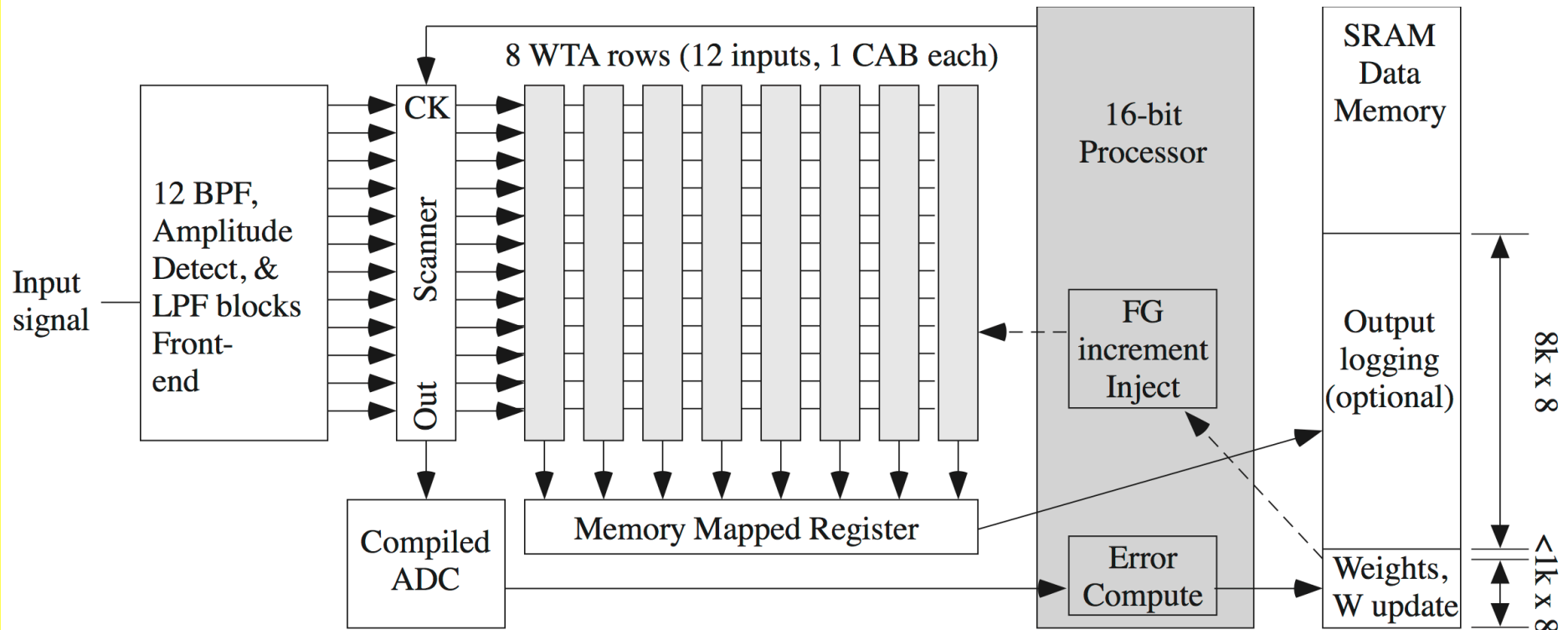
FPAA Datapath Architecture



- Amplitude vs. frequency for different distances
- Somewhat robust frequency features & complex background noise



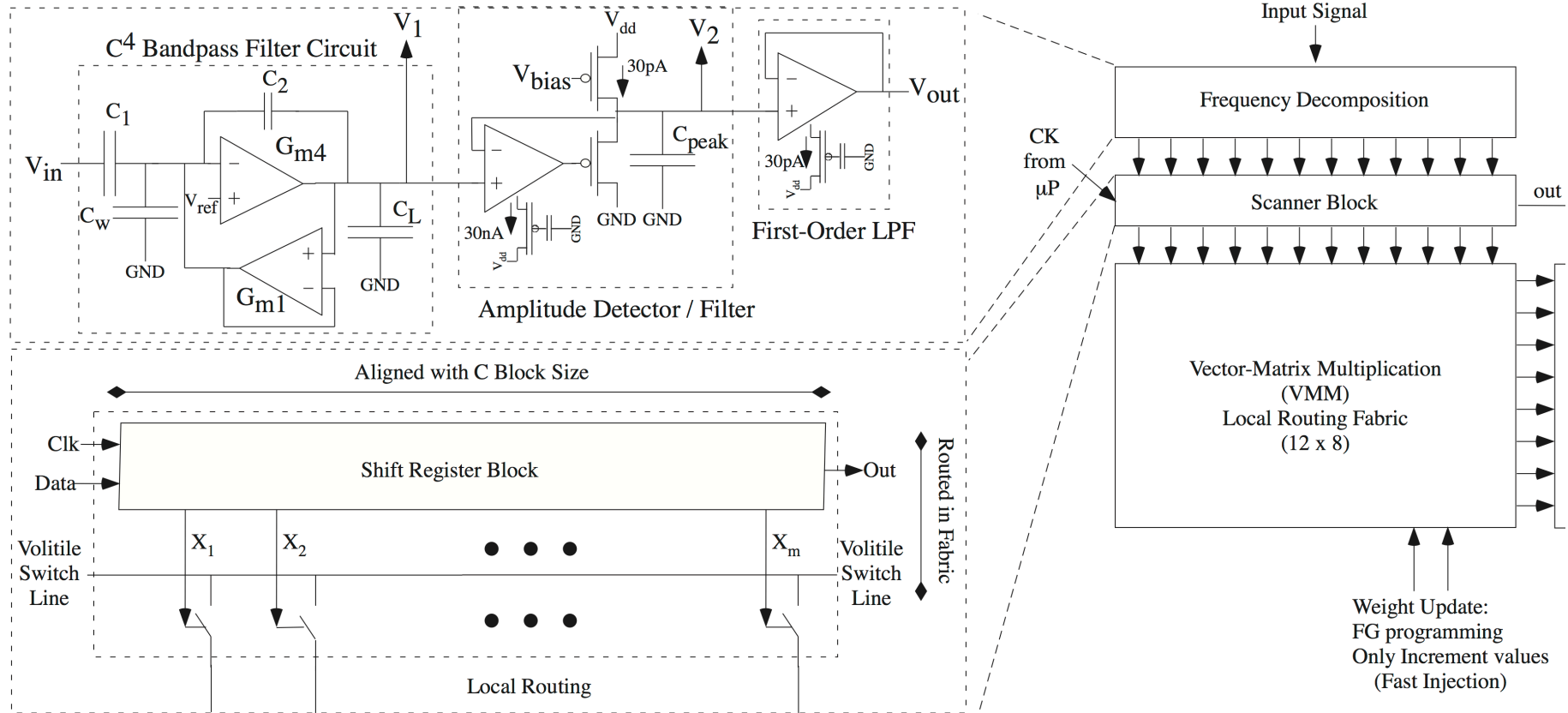
Developing FPAA Adaptation



- Combined Analog, Digital, and μ P Architectural Design
- FG Programming uses μ P device (Batch)
- Analog classifier, ultra-low energy classifier datapath



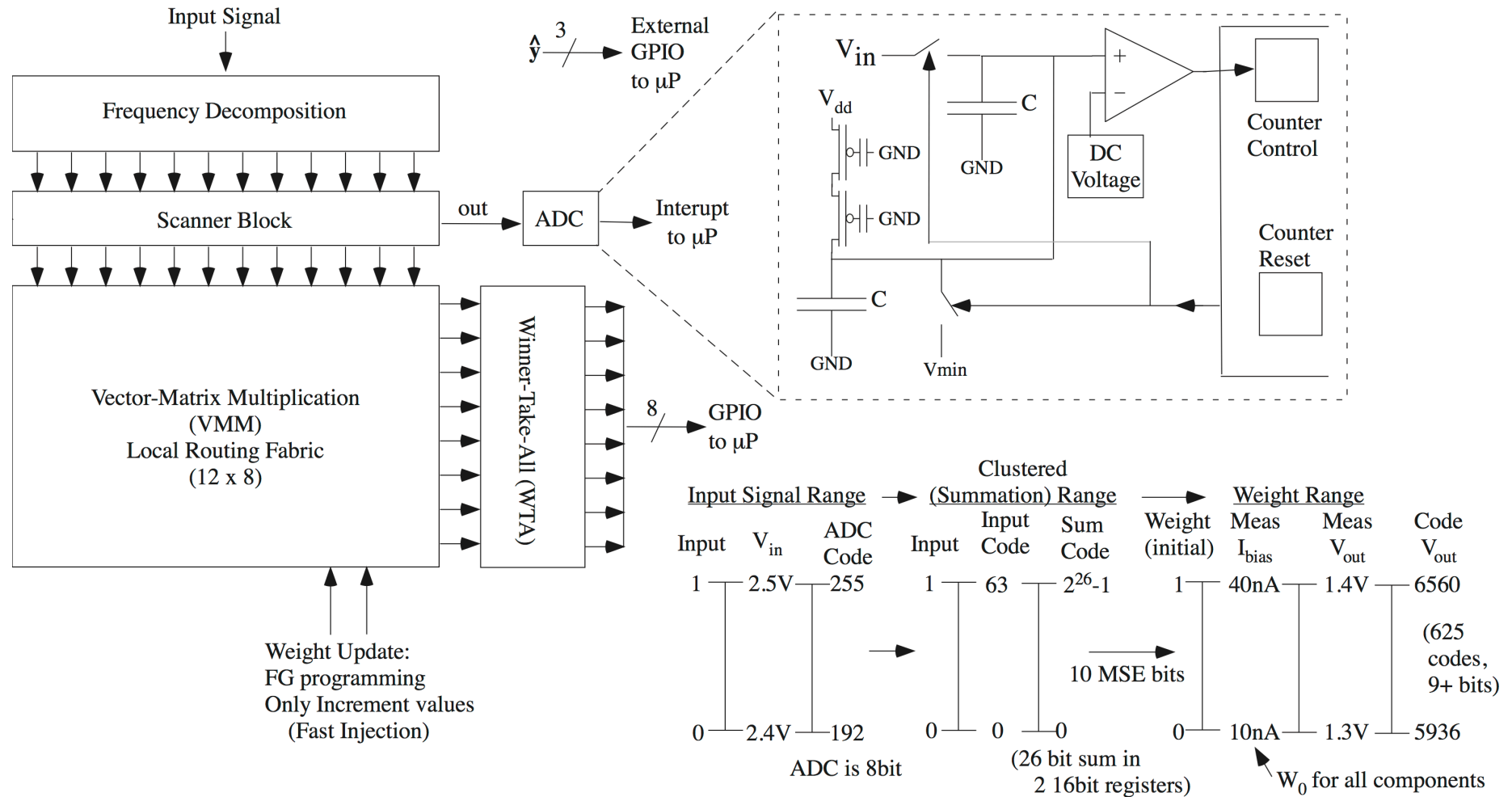
More Analog Classifier (VMM+WTA)



- Front-End Circuit blocks, “scanning” blocks to enable training
- VMM implemented in local-fabric routing



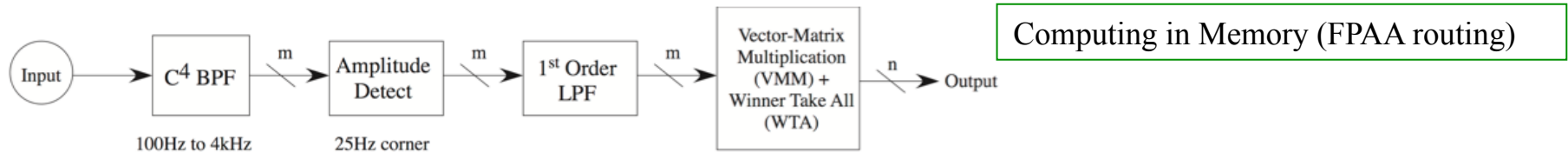
More Analog Classifier (VMM+WTA)



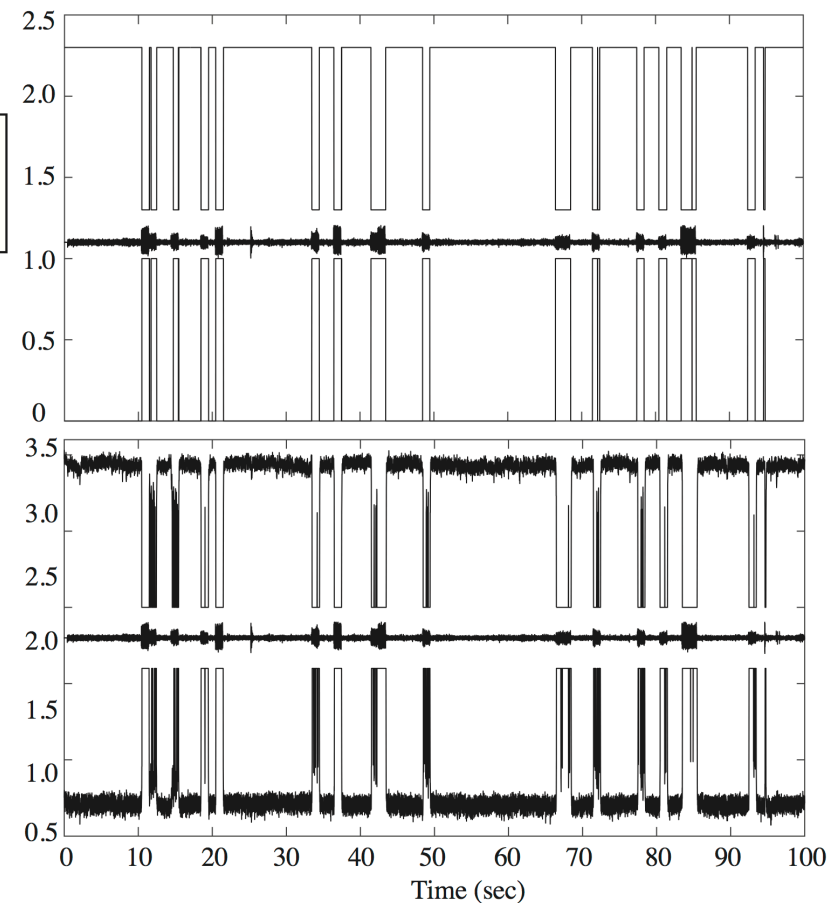
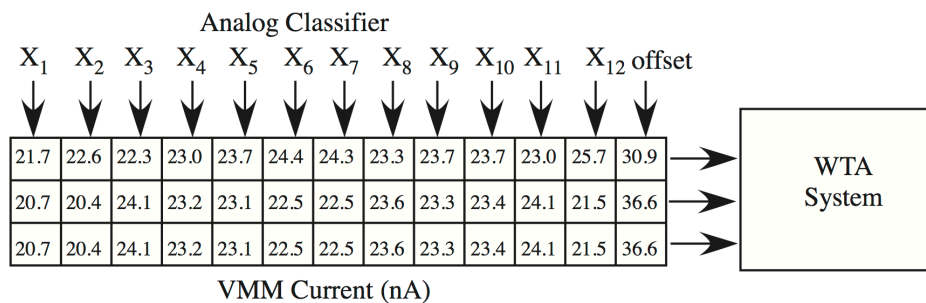
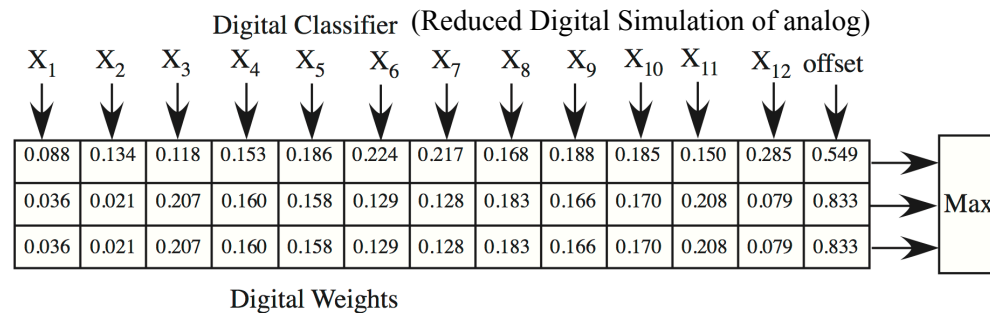
- Error metrics and FG update programming handled through μP
- Requires careful scaling of fixed-point arithmetic



Analog Learning Classifier (VMM+WTA)



Detect one acoustic signal
(1s signal bursts)

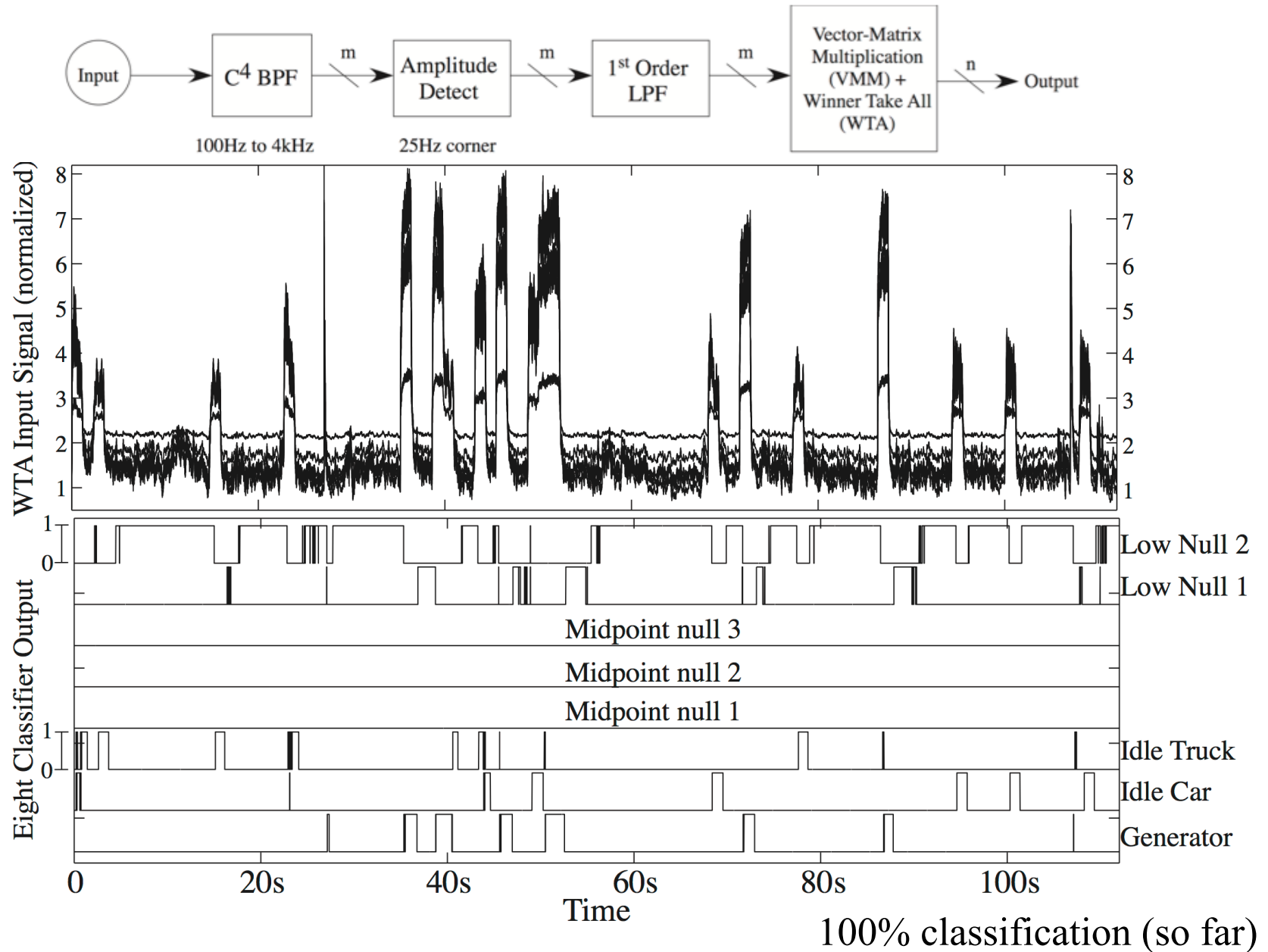


100% classification (so far)

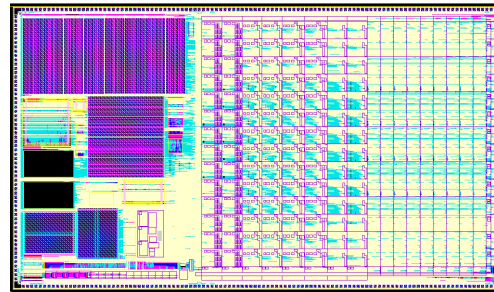


Analog Learning Classifier (VMM+WTA)

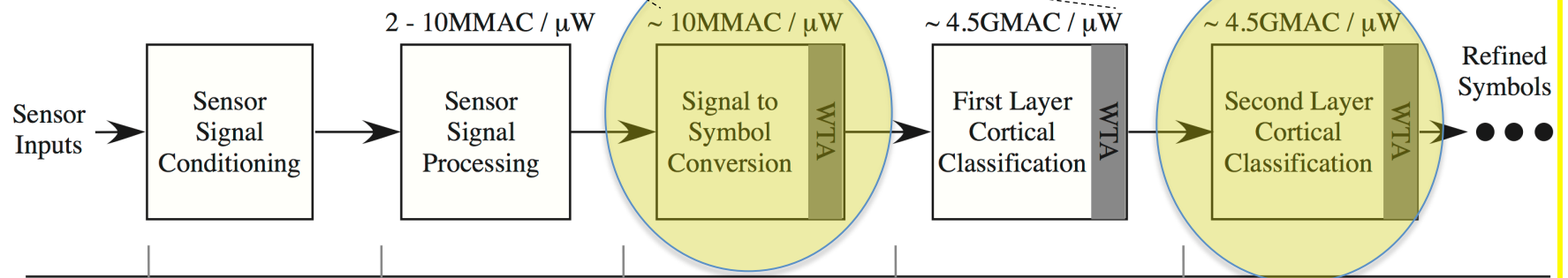
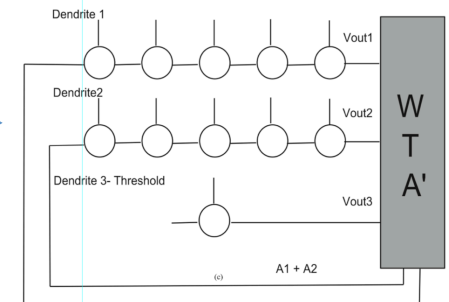
Detect one all three acoustic signals



Neuromorphic + Analog Computation



Dendrite Computation



Speech Recognition	Microphone Interface / filtering	Cepstrum	Basic Auditory Features (VQ, GMM)	Phoneme Classification	Low SNR Wordspotting
Image Processing	Image acquisition, color calculations	Retina (edge enhancement)	Edge / Corner Detection	Movement Sequence Classification	Gesture Recognition, etc.
Baseband Communications	Demodulation of desired band	Frequency Decomposition	Fundamental Comm Symbol Detection	Frequency Hopping Recognition	Complex Signal Detection

