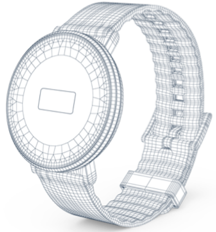
Flash Memory Summit

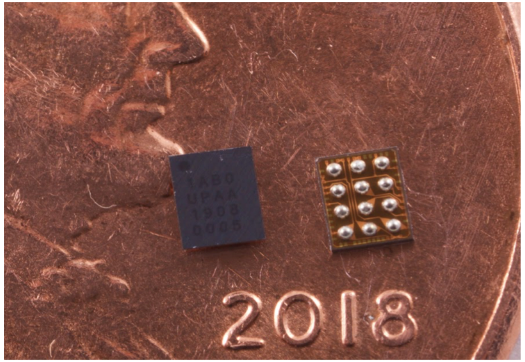
Flash-Based Analog Neural Networks: Possibilities and Tradeoffs

Jeremy Holleman
CTO, Syntiant Corp.

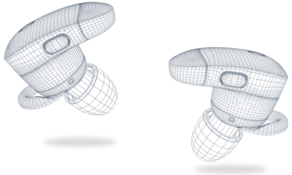
Syntiant Overview



Founded 2017 to combine machine learning and semiconductor expertise, deliver ML to the untethered edge.



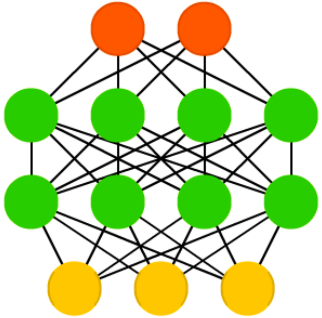
First product line NDP10x in production: Amazon AVS qualified wakeword solution at 140uW



SYNTIANT

Neural Network Basics

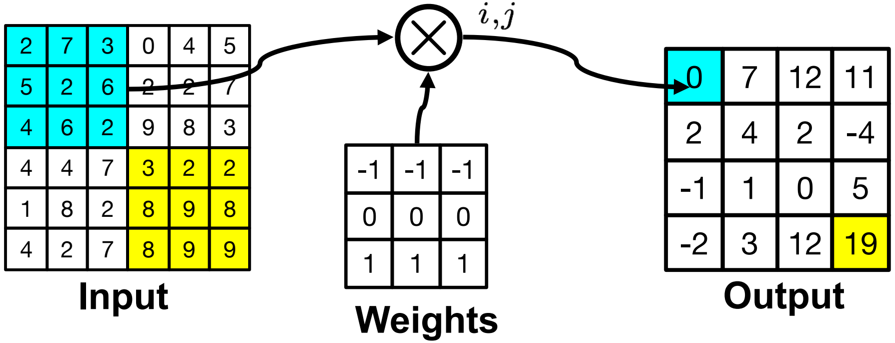
- NNs are mostly matrix-vector multiplications (MVMs) + nonlinearity
- Fully-connected layers are MVMs directly
- Convolutional layers easily built from MVMs



$$\vec{Y_N} = \mathrm{ReLU}(\mathbf{W_N}\vec{Y}_{N-1})$$

$$Y_{N,i} = \sum_j w_{i,j} Y_{N-1,j}$$

$$y_N[x,y] = \mathrm{ReLU}(\sum_{i,j} y_{N-1}[x+i,y+j]w[i,j]$$



Input

2	7	3	0	4	5
5	2	6	2	2	7
4	6	2	9	8	3
4	4	7	3	2	2
1	8	2	8	9	8
4	2	7	8	9	9

Weights

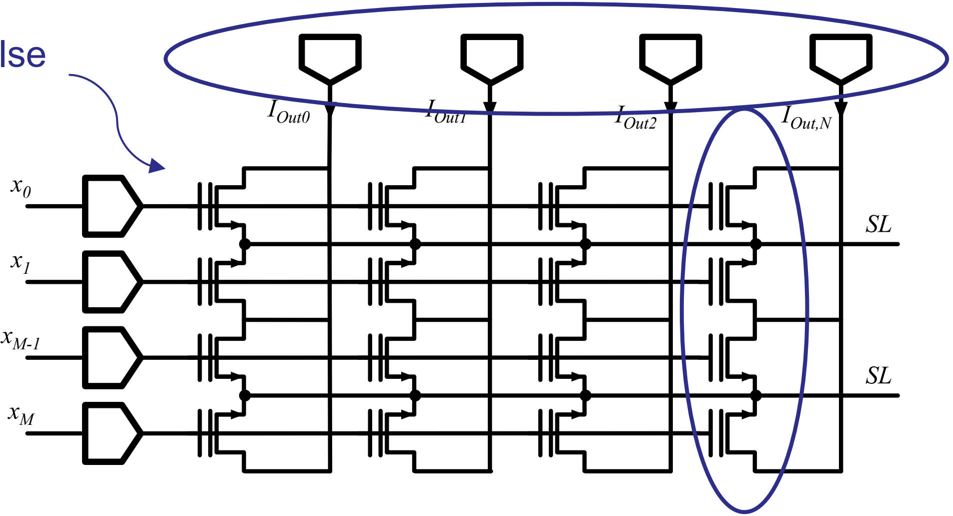
-1	-1	-1
0	0	0
1	1	1

Output

0	7	12	11
2	4	2	-4
-1	1	0	5
-2	3	12	19

Figure: van Veen, asimovinstitute.org

Basic Analog NN Block



Inputs encoded as pulse widths or voltages

DAC/Drivers broadcast input across rows

Current Sense/ADC collects combined column current

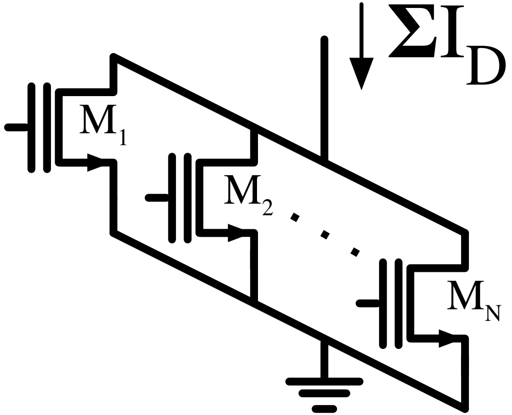
Neuron maps to a column (2 if differential)

NOR-flash directly implements matrix-vector multiply

Size/power advantage: 1 weight storage + MAC in 2T

Efficiency Limit (1)

- Resolution (SNR) dictates local power consumption
- Activation resolution related to summed current
- Strong dependence on resolution
- Modest precision possible with extreme efficiency



$$i_{n,RMS} = \sqrt{2qI_DB}(A^2/Hz)$$

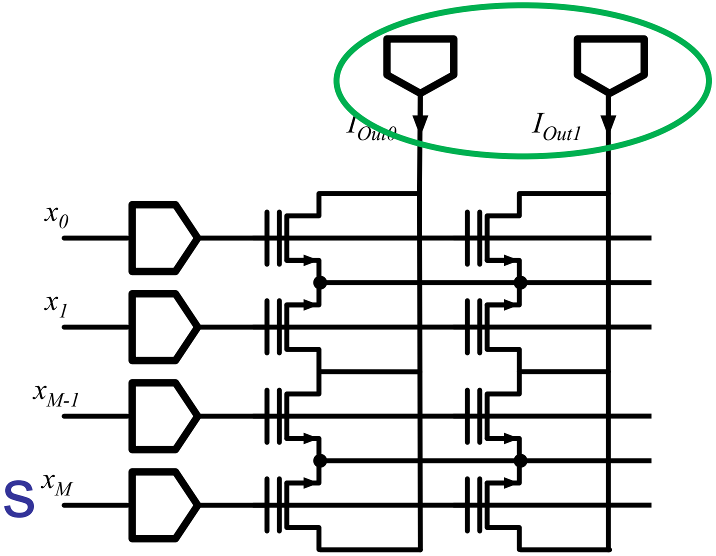
$$E = TI_DV_{DD} = \boxed{2q2^{2N_{bits}}V_{DD}}$$

~20 fJ @ 8b (N·100 TOPS/W)

~20 aJ @ 3b (N· 100 POPS/W

Efficiency Limit (2)

- Output sensing/ADC drives power
- SoA ADC[1] @ 10fJ/step,8b → 2.5 pJ/output
  - Amortized over column
  - More inputs/neuron improves efficiency.
    128→ 100 8b TOPS/W



SYNTIANT

[1] Murmann 2019, ADC Survey

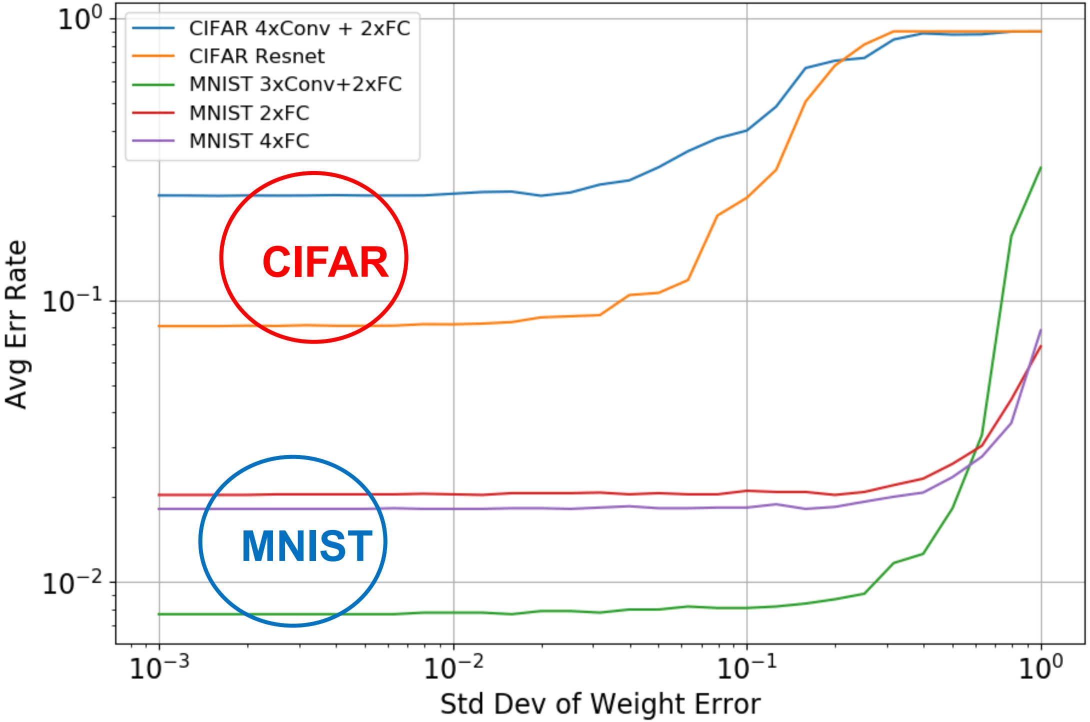
Analog NN Concerns

Offsets	Closed-loop training
Weight drift	Calibration
Retention.	
Temperature	
Noise	Noise << LSB discretized away. NN Robustness

- System requirements drive circuit specs
- Exploit NN robustness to approximate math

Weight Sensitivity

- Randomly vary weights
    - w'=w(1+e) e~N(0,σ)
- Sensitivity varies with application
- 4b-8b typically adequate

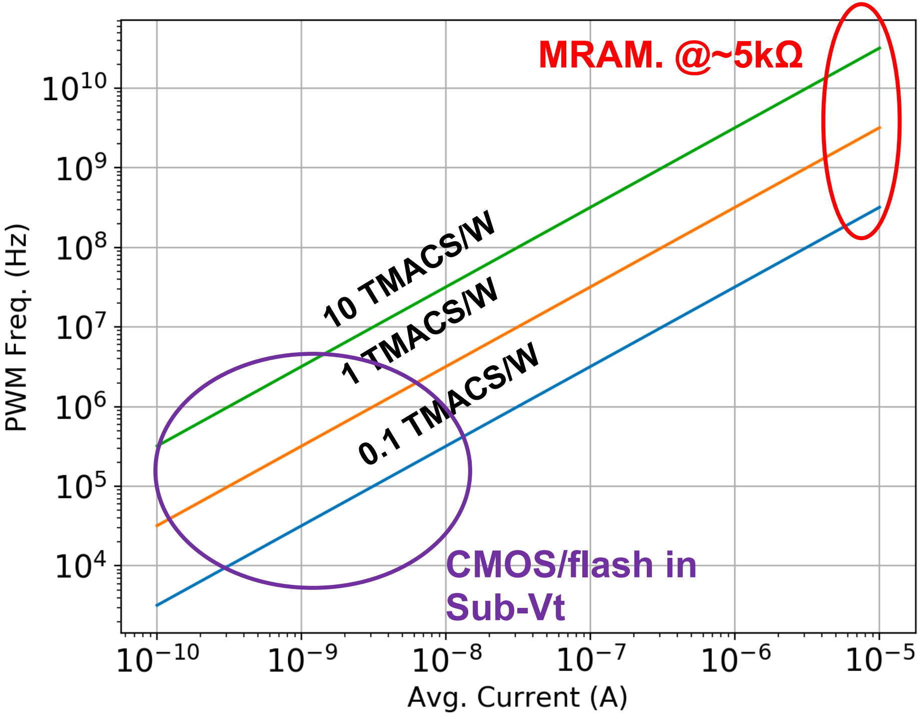


Emerging Material Considerations

- FET-based vs resistive
  - Exponential vs Linear
- On/Off current → power/speed tradeoff
- Noise
  - RTS – poorly modeled, difficult to calibrate

R_{on} Power Impact

- R_{Off}/R_{On} impacts dynamic range
- Low energy requires low curren or high speed
- R_{Avg} → frequency



Assume Icell/Itotal=10%,
Von=100mV

SYNTIANT

Temperature Stability

- Absolute temperature variation can be compensated, but adds complexity
- Mismatched temperature response corrupts weights

Matched Temperature Response

W	I_{ON} (T1)	I_{ON} (T2)	k	W'
1	1 nA	10 nA	.1	1
5	5 nA	50 nA		5
15	15 nA	150 nA		15

😀

Varied σ_{TC} = 10%

W	I_{ON} (25C)	I_{ON} (100C)	k	W'
1	1 nA	9.5 nA	.1	0.95
5	5 nA	51 nA		5.1
15	15 nA	118 nA		11.8

😬

Conclusions

- Analog neural networks show great promise for extreme efficiency and parallelism
- Implementation presents cross-disciplinary challenges
- Algorithm/Circuit/Device co-design key to overall performance

Flash Memory Summit

Questions?

SYNTIANT