ANAFLASH

Non-Volatile Neural Network Accelerator in Your SoC

Sang-Soo Lee, CEO and Co-Founder
Seung-Hwan Song, CTO and Co-Founder
ANAFLASH Inc.

Flash Memory Summit

Company Overview

ANAFLASH

WHAT WE DO?

Founded in 2017
we develop
Logic Compatible
NV-DNN and eFlash
IPs
for Edge Computing

TEAM

Executives have
combined 60+ years
of Engineering &
Management
Experience

TECHNOLOGY

Patent pending
NV-DNN and eFlash
IPs
in Standard CMOS
process

WHERE WE ARE?

Headquartered in
San Jose CA USA

IEDM 64th International Electron Devices Meeting

NSF

Department of Defense United States of America

University of Minnesota Driven to Discover®

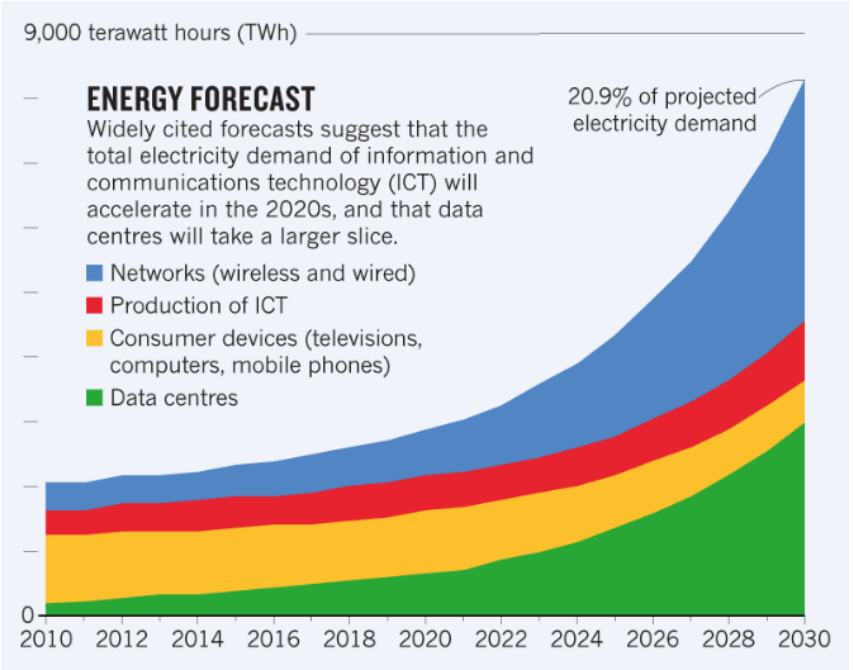
Berkeley SKYDECK

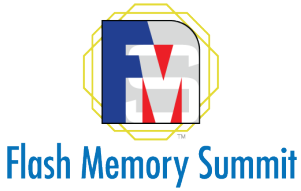
Flash Memory Summit

Artificial Intelligence in the Edge

Growing Cloud Energy Concern

N. Jones (Nature 2018)

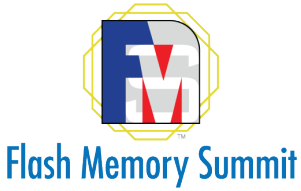




Let's move to Edge! However,…

- Challenges in the Edge Environment
  - Size, Weight, and Power (SWaP) limited
  - Compute and memory resource limited
  - Cost sensitive

- How to make it work under these challenges?



Approximate Computing

- Technique that allows approximate results in applications not requiring strict accuracy
- This can improve power efficiency a lot
- In case, such errors can be managed by system level techniques statistically (i.e. ECC and redundancy, etc.)
- Could be combined with Digital (However,…)

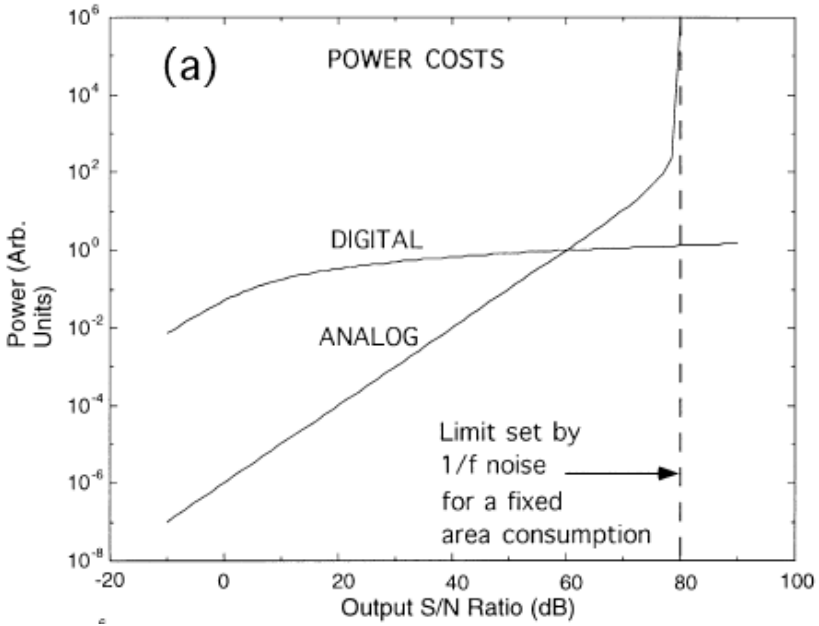
Analog vs. Digital Computation

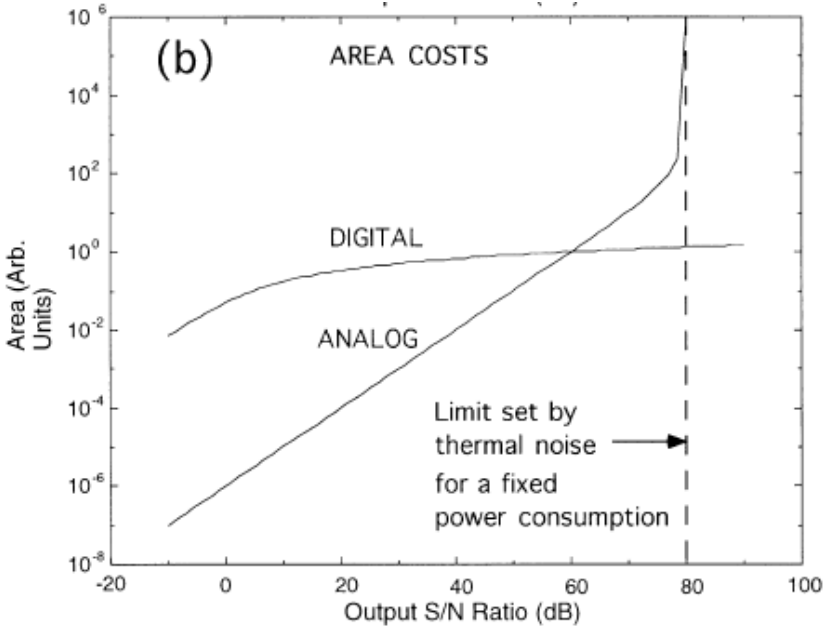
ANALOG	DIGITAL	Which is better for efficiency?
Narrow signal swing	Full VDD-GND swing	ANALOG
Information from single transistor (continuous)	Information from single transistor (1 or 0)	ANALOG
Multi-bit single wire	Single-bit single-wire	ANALOG
Result affected by noise and variation	High noise margin	DIGITAL

R. Sarpeshkar (Neural Computation 1998)

- Analog has more advantages for efficiency!

Let's do Analog Computing in Edge



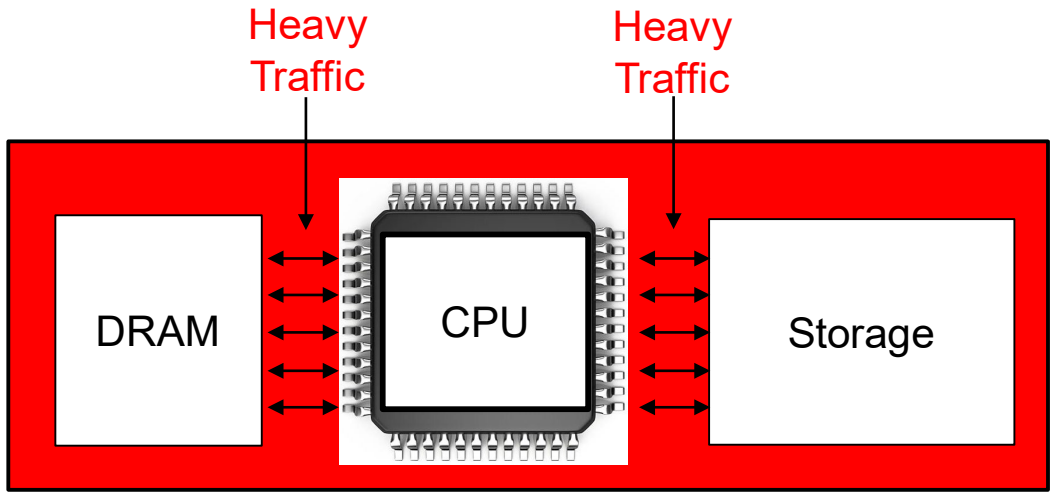


R. Sarpeshkar (Neural Computation 1998)

- Analog is significantly efficient at low-precision!

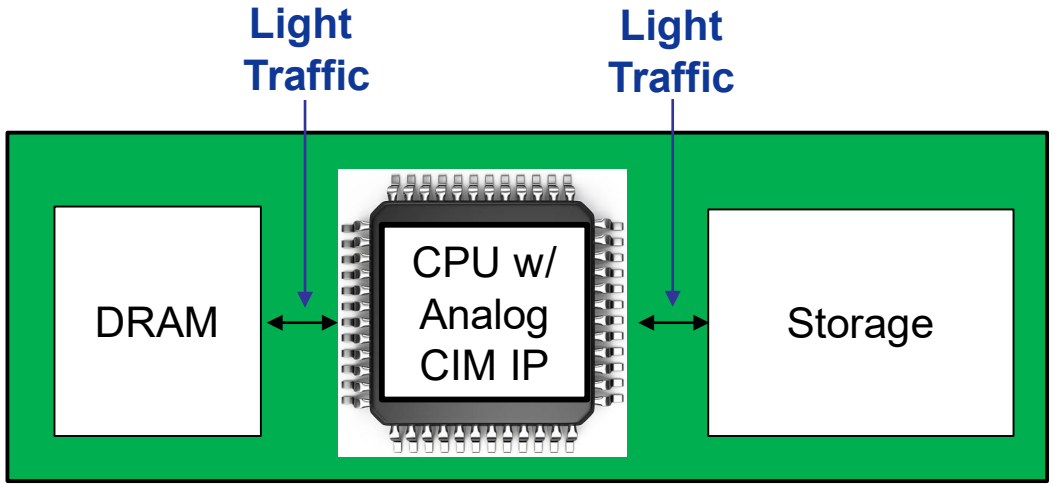
Memory Access Bottleneck

- Off-chip access from CPU to memory (storage) has long (and unpredictable) latency, and limited bandwidth



Analog CIM Architecture

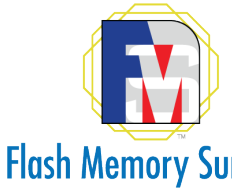
- Analog Compute-in-Memory IP integrated in CPU
- Reduce off-chip memory access



Lesson Learning from Human Brain

- Brain has much more efficiency with much small values of SWaP
  - $3.6x10^{15}$ synaptic operation with 12W → $3x10^{14}$
  - i9 CPU running 3GHz with 140W → $2x10^{7}$

- Biological neural network doesn't discriminate computational device and memory device

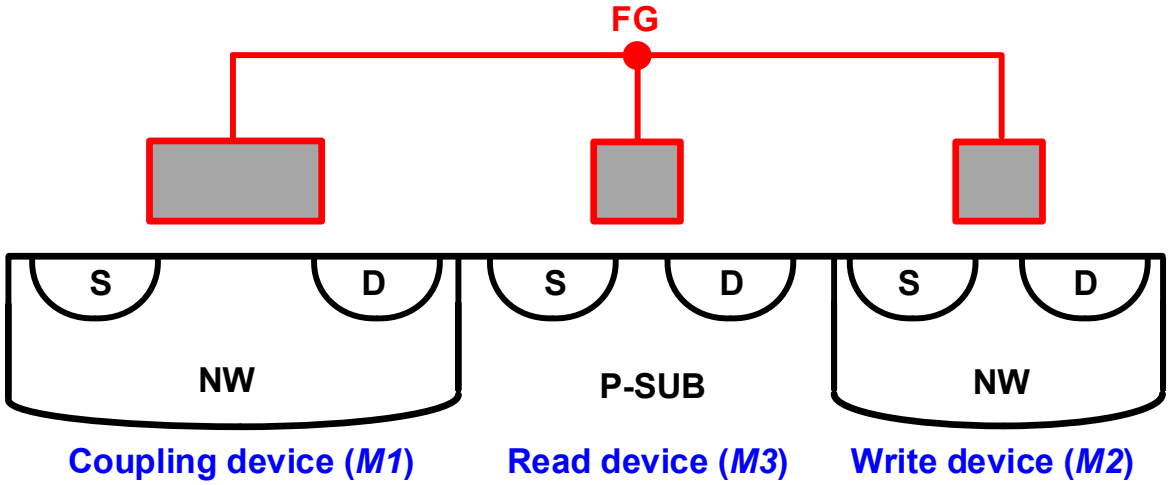


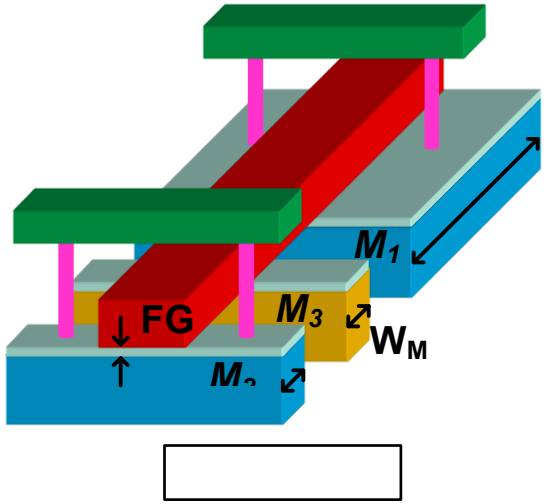
Candidates for Analog Computing

- Logic gates (e.g. NAND, XOR, etc.) → No
- Transistor, capacitor, inductor, etc. → Yes

- SRAM (Not able to store multi-bit, volatile) → No
- MRAM (Not able to store multi-bit, non-volatile) → No
- ReRAM (multi-bit, nonvolatile) → Yes
- Flash (multi-bit, nonvolatile) → Yes

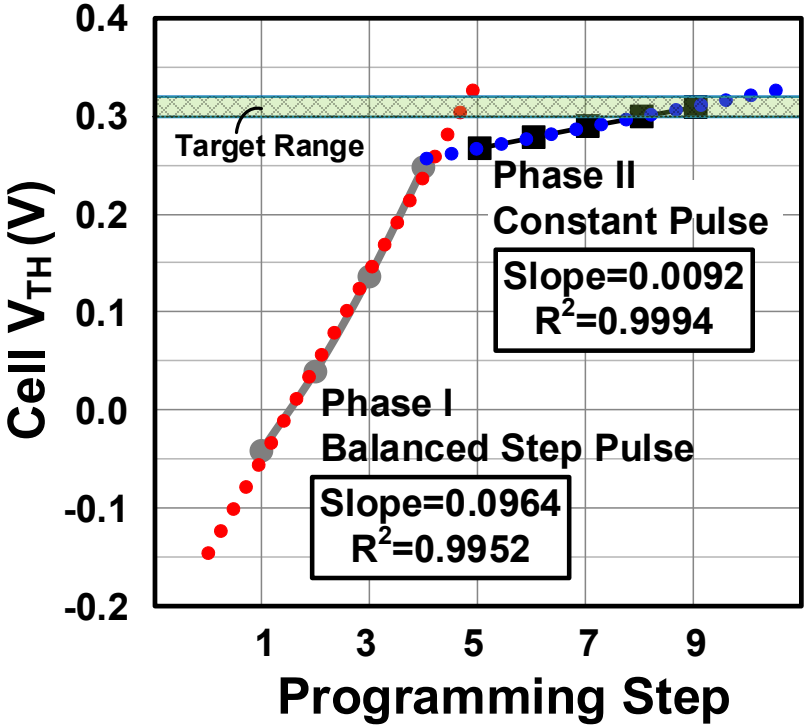
Logic Compatible Flash Memory IP

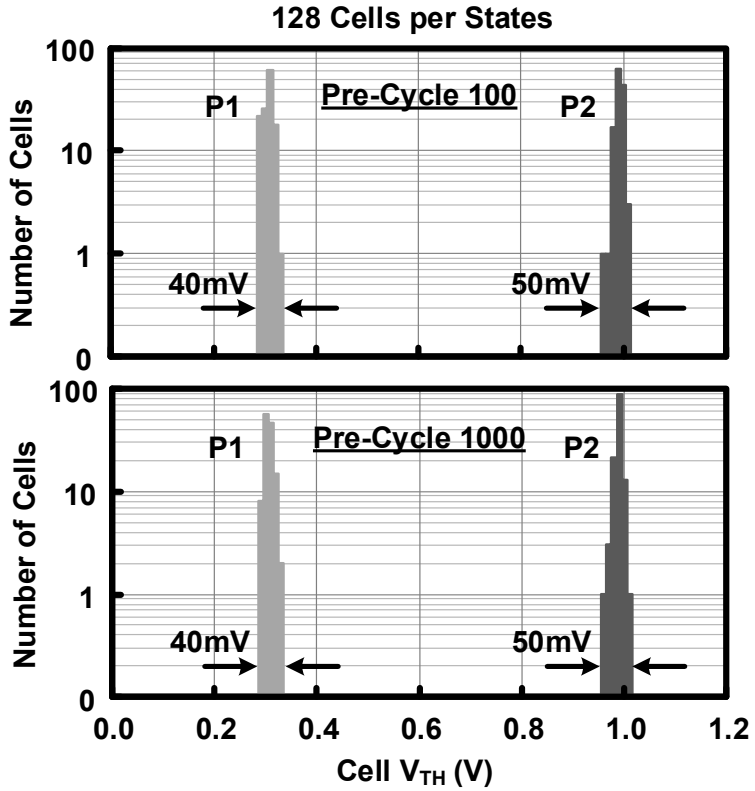




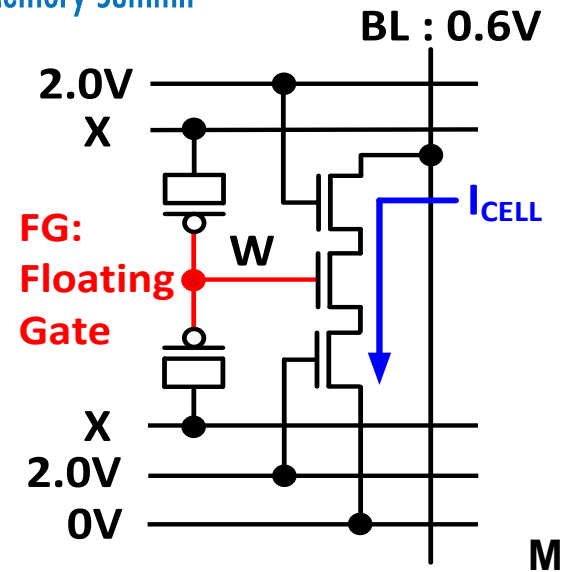
- No Process Overhead in Standard Logic Process
- Leverage High Performance Digital Logic (Scalable)

Precise Analog Programming Scheme



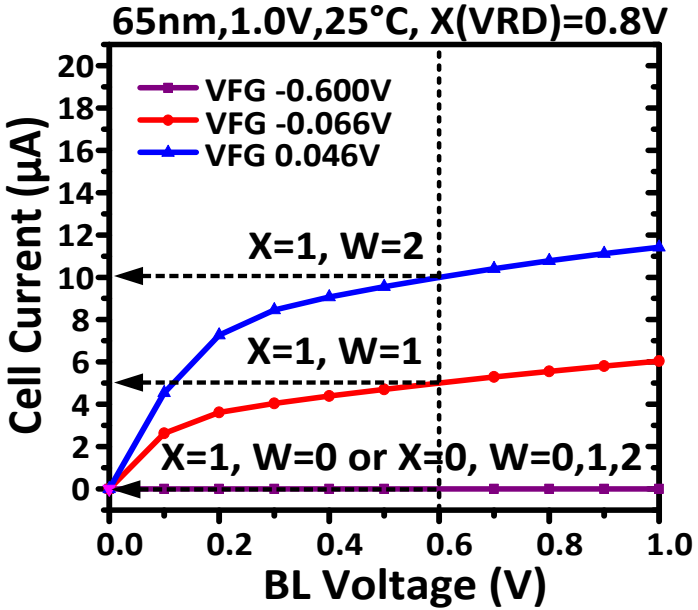


Logic Compatible Flash Based Synapse



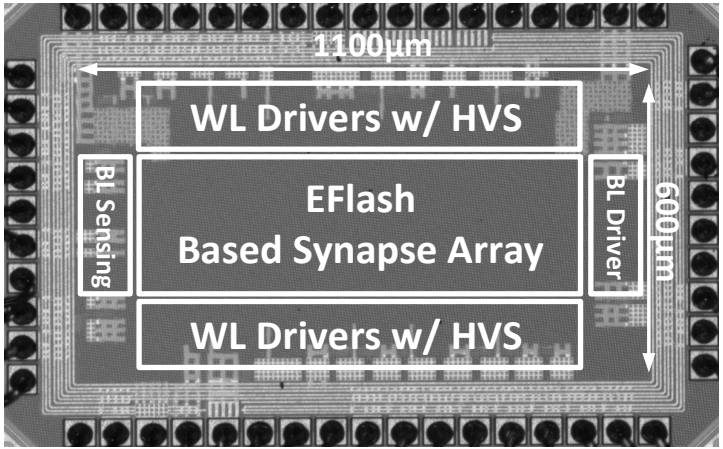
X	W	X·W	I_{CELL}
0	0	0	0μA
0	1	0	0μA
0	2	0	0μA
1	0	0	0μA
1	1	1	5μA
1	2	2	10μA

M. Kim et al., IEDM 2018



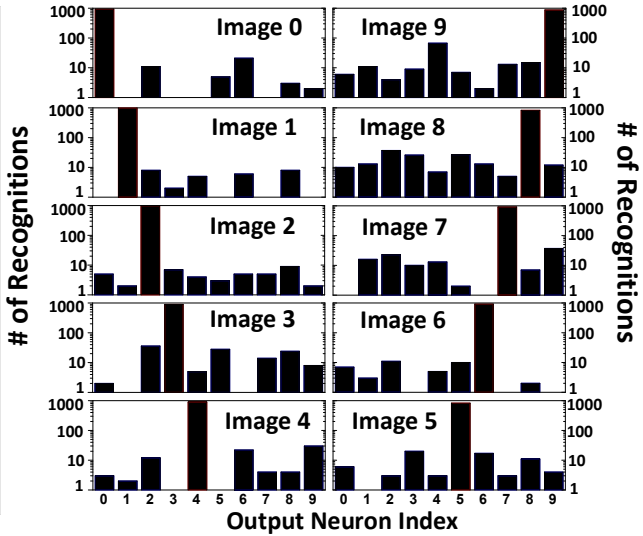
- Cell current proportional to X·W (=0μA,5μA,10μA)

65nm Test Chip Summary

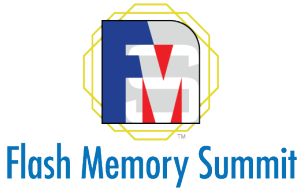


M. Kim et al., IEDM 2018

Technology	65nm CMOS
Circuit Area	1100 X 600 μm^2
VDD (Core, IO)	1.0V / 2.5V
# of Neurons	320
# of Synapses	22K (=68x320)
Throughput	1.28G pixels/s per core (tREAD : 50ns)
Power	15.9μW (per neuron)



- High efficiency (171.1 TOPS/W) by analog CIM arch.
- Recognition accuracy close to the SW model



Summary

- Growing need to move AI computing toward Edge
- Analog computing can improve power efficiency by approximately computing neural network
- Analog computing-in-memory using logic compatible embedded Flash memory is a strong candidate to overcome memory bottleneck
- Test chip result fabricated in 65nm logic process shows power efficiency of 171.1 TOPS/W

Flash Memory Summit

THANK YOU
FOR YOUR ATTENTION

info@anaflash.com

ANAFLASH

Always-on Local AI and NVM solution
For Battery-Powered Smart Devices

3003 N. First St. #221
San Jose, CA 95134