New Memory-Centric Architecture Needed for AI

Sylvain DUBOIS – VP Business Development & Marketing

May 23rd , 2018

CROSSBAR

AI is moving to the Edge

- Uploading, processing and downloading from cloud takes time
- Transmitting data burns energy
- Some apps cannot rely on wireless connection
- Data less exposed if processed locally

BATTERY LIFE

PERFORMANCE

RELIABILITY

SECURITY & PRIVACY

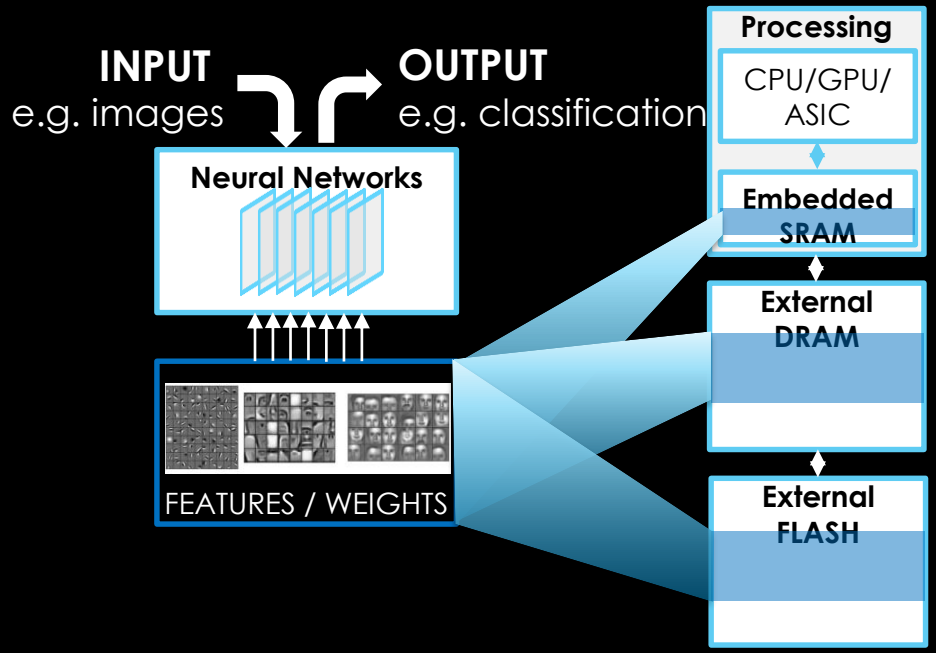
Opportunities to build Smarter Things

ReThink Machine Intelligence

- Train models in the cloud with massive amount of data
- Infer real-time decision at the edge
- >37B IoT semiconductor chips in 2018
- AI-IoT is about smart efficient designs - doing more with less

Distributed AI

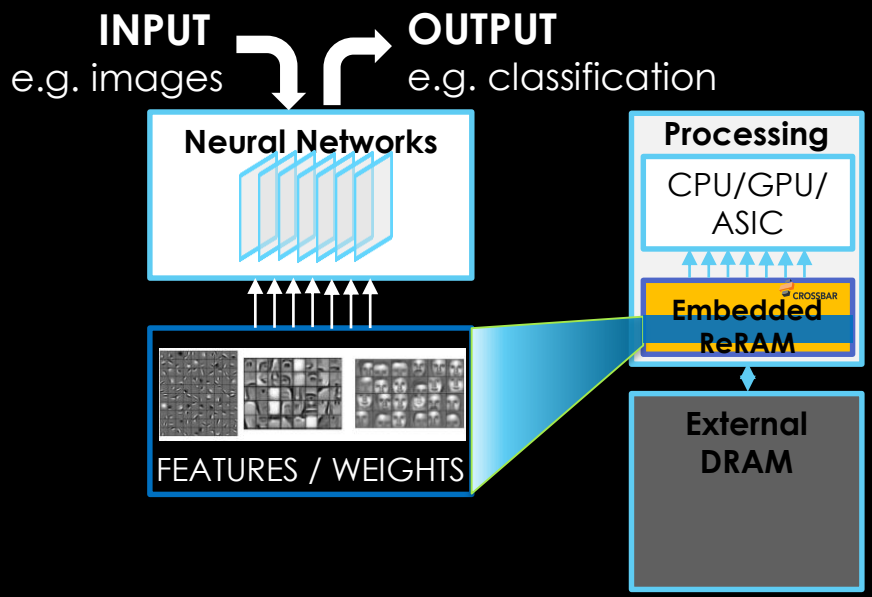
The Memory Bottleneck in AI



- Performance limited by memory bus bandwidth and latencies
- Energy wasted moving data from FLASH to DRAM to SRAM to processing cores
- Need to refresh / reload model in SRAM/DRAM at every wake-up

Bring data and algorithms on same chip

A Memory-centric Architecture for AI

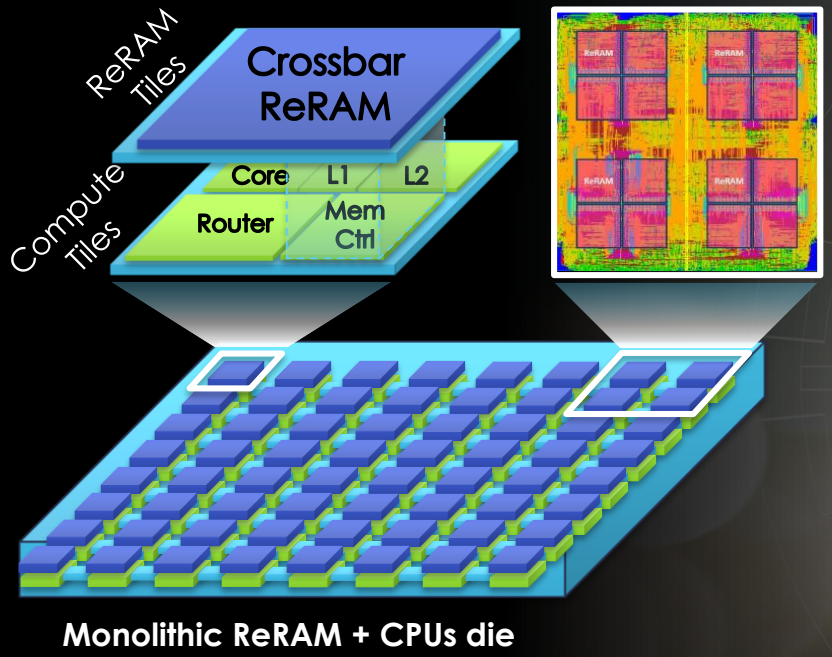


Crossbar embedded ReRAM:

- Highly parallel non-volatile memory
- Denser than SRAM
- Lower energy than DRAM
- DRAM equivalent reads

Bring data and algorithms on same chip

Building Monolithic Computers



ReRAM Tiles integrated with 256-bit RISC-V

Monolithic ReRAM + CPUs die

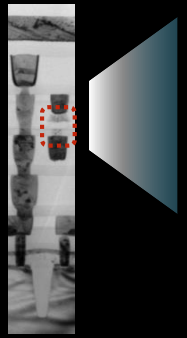
HIGHLY PARALLEL

DATA INTENSIVE

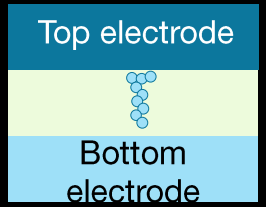
LOW ENERGY

ReThink computing architecture

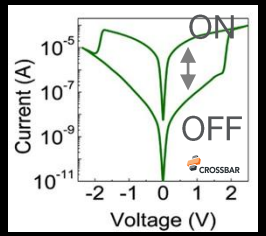
Crossbar ReRAM Addresses AI's Needs



ReRAM cell inserted between metal lines of standard CMOS



- Sub 10nm metallic nano-filament
- 2 masks, 8 process steps
- Re-using existing fab materials and tools



- 1000X ON/OFF ratio
- -40/+125C
- 1M+ write cycles
- 10 years retention
- 10ns read latency
- 1pJ/bit - <1uA/MHz/bit

EMBEDDED

LOW COST

RELIABLE

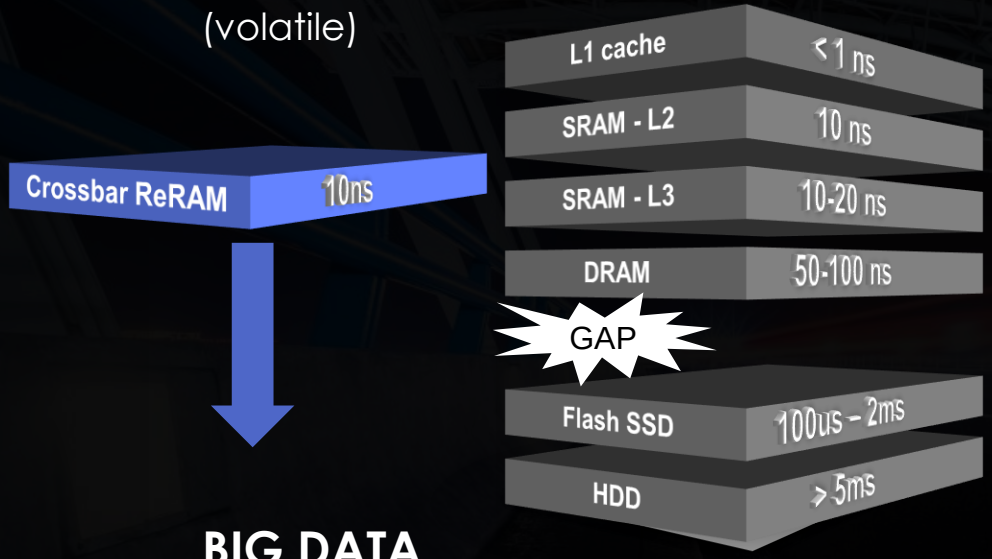
HIGH PERFORMANCE

LOW ENERGY

Low energy, high-performance embedded non-volatile memory

Leading Among Embedded Memory Technologies

COMPUTING
(volatile)



BIG DATA
(non-volatile)

Denser than SRAM
Crossing point of two lines

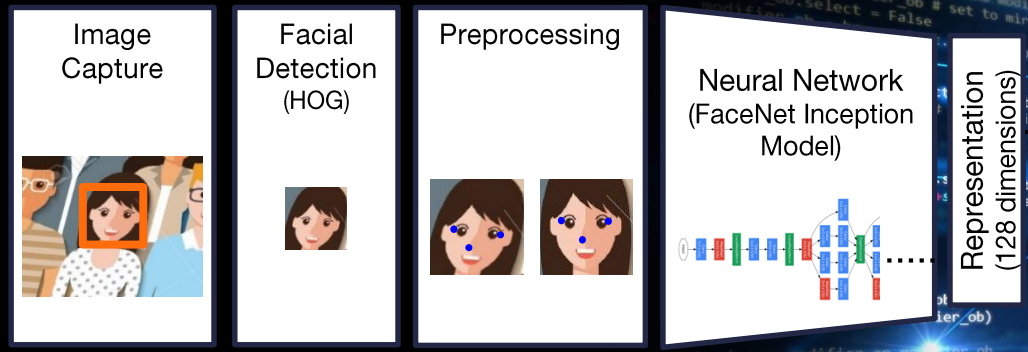
Lower cost than MRAM
2 masks – 3 films – existing tools

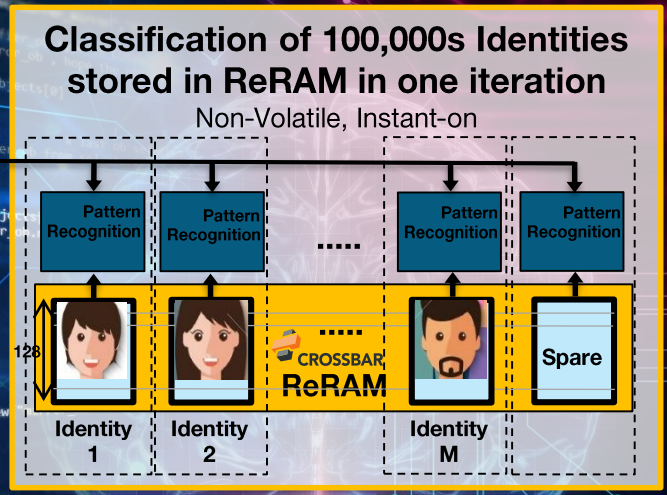
Scale better than FLASH
1x nm and below

Lower energy than DRAM
1pJ/bit vs 9pJ/bit for HBM2

Superior characteristics at advanced process nodes

Face Recognition with ReRAM





Simultaneous Processing with Deterministic Performance

- Parallel comparison against all identities
- If no match, new identity created (learning)
- Classification performed in one cycle independent of number of identities

Building Up an AI Ecosystem

- ReRAM currently evaluated and transferred to world's largest foundries at 1x nm, 2x nm and 40nm
- Several collaborations signed to design new architecture with strategic partners.
- Do you want to be part of it?

Crossbar joining Embedded Vision Alliance

ReRAM for AI Demo Available

Visit us at booth #606

Headquarters in Santa Clara, CA

www.crossbar-inc.com

CROSSBAR

ReThink with ReRAM

www.crossbar-inc.com