Weebit ReRAM
Enabling New Low-Power AI Architectures in Advanced Nodes

Marc Drouard,

Chief Architect, Weebit Nano

June 13th, 2025

Weebitnano
THE NEXT NVM IS HERE

cea list

LIST TECH DAYS

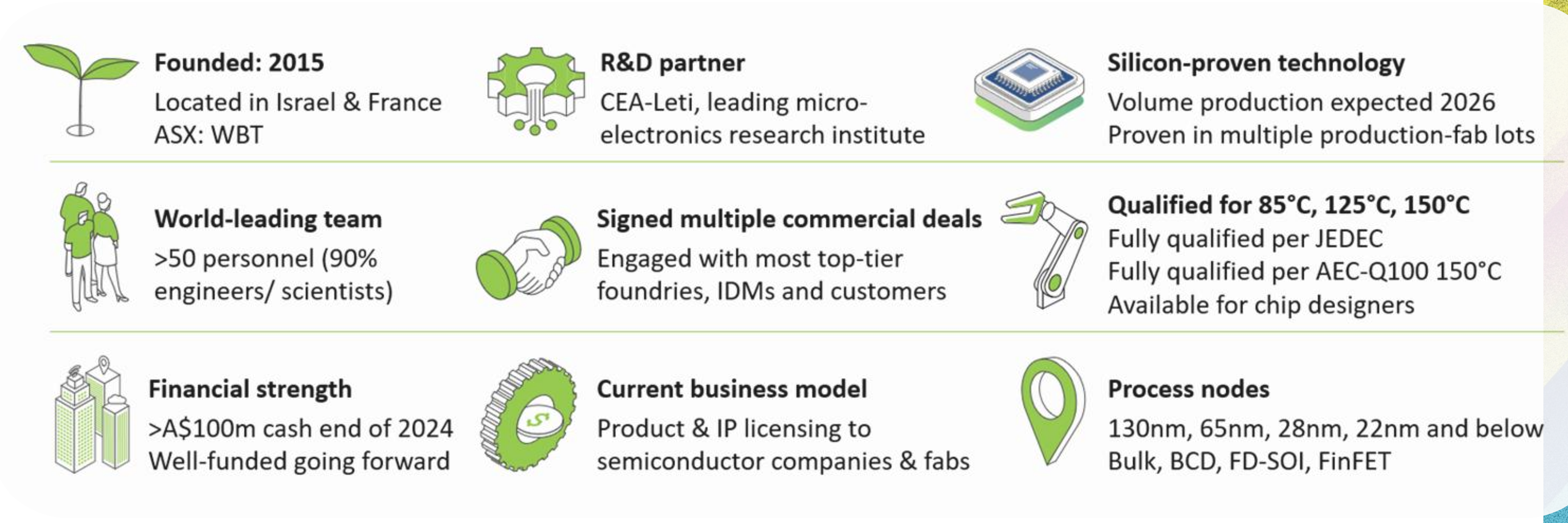
Contents

- Introduction
- Near Memory Computing using Weebit ReRAM
- Analog computing using Weebit ReRAM
- Conclusion

Weebit Nano - Leading Licensor of ReRAM IP

Next-generation memory technologies for the global semiconductor industry

We are enabling a leap forward in memory technology for a new era of connected devices



Founded: 2015
Located in Israel & France
ASX: WBT

R&D partner
CEA-Leti, leading micro-electronics research institute

Silicon-proven technology
Volume production expected 2026
Proven in multiple production-fab lots

World-leading team
>50 personnel (90% engineers/ scientists)

Signed multiple commercial deals
Engaged with most top-tier foundries, IDMs and customers

Qualified for 85°C, 125°C, 150°C
Fully qualified per JEDEC
Fully qualified per AEC-Q100 150°C
Available for chip designers

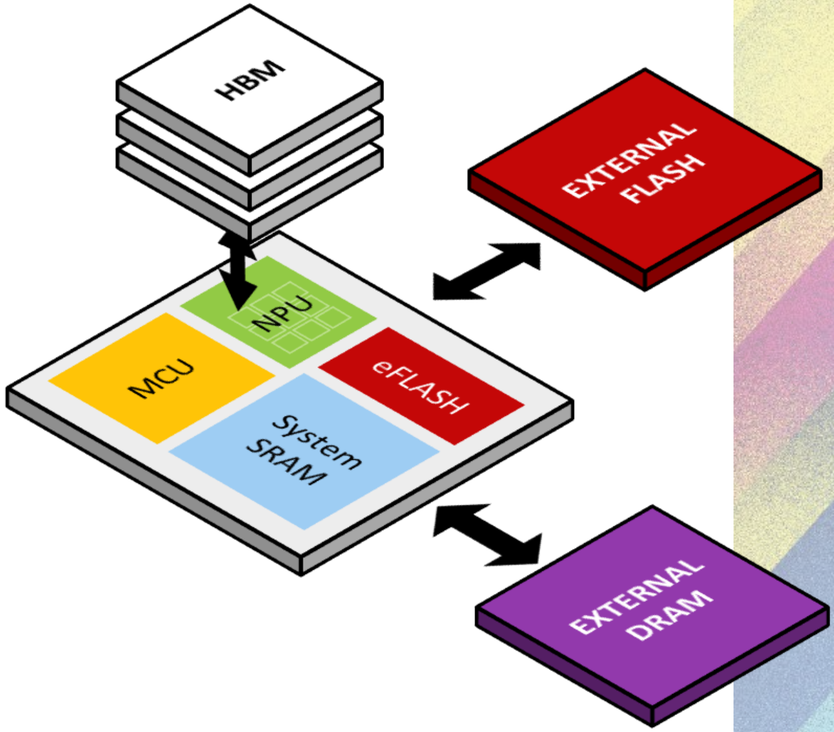
Financial strength
>A$100m cash end of 2024
Well-funded going forward

Current business model
Product & IP licensing to semiconductor companies & fabs

Process nodes
130nm, 65nm, 28nm, 22nm and below
Bulk, BCD, FD-SOI, FinFET

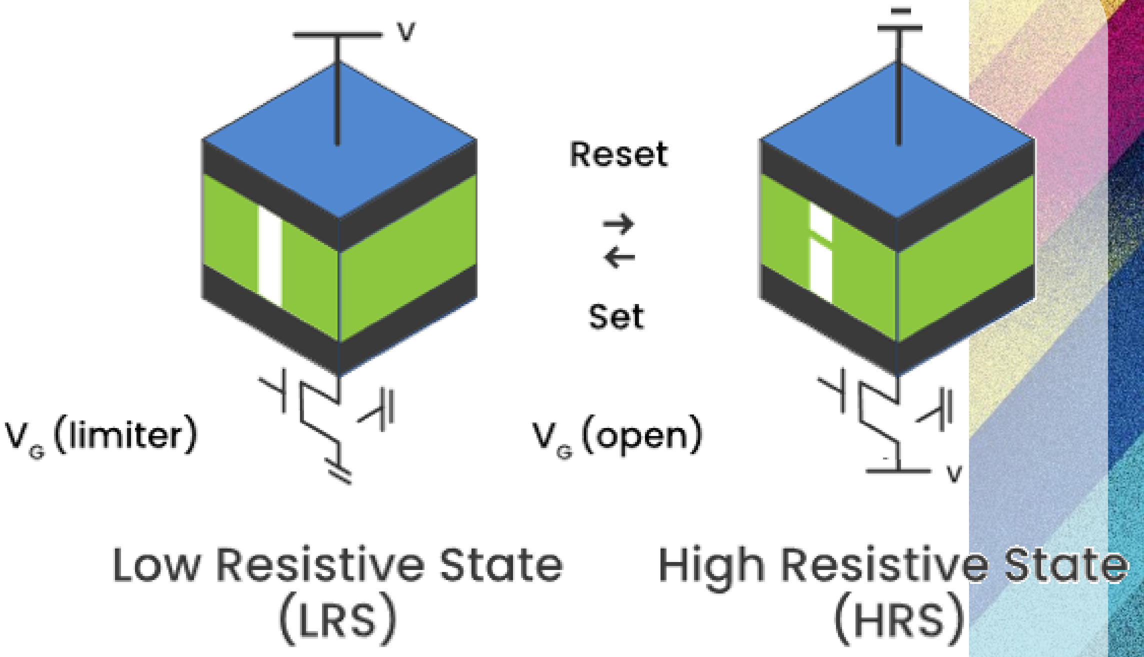
Edge AI – Memory overview

- Non-Volatile Memory (flash):
  - Stores model weights persistently
  - At startup, weights are loaded into faster memory
- On-Chip SRAM:
  - Staging buffer for frequently used weights
  - Reduces latency and power consumption during inference
  - Large SRAM capacity required
  - Large silicon cost and leakage power
- External DRAM:
  - NPU fetches weights directly from DRAM
  - Fetching is continuous to hide memory latency



What is Weebit ReRAM?

- Weebit ReRAM is based on oxygen vacancies filament
  - RESET (Erase) – Partial dissolution of the filament
  - SET (Program) – Re-creation of the conductive filament
- Data storage is resilient to environmental conditions
- Low power consumption
  - Low read voltage <1V
  - Low write voltage <3V
  - Low currents
  - Zero standby power
  - Fast operation



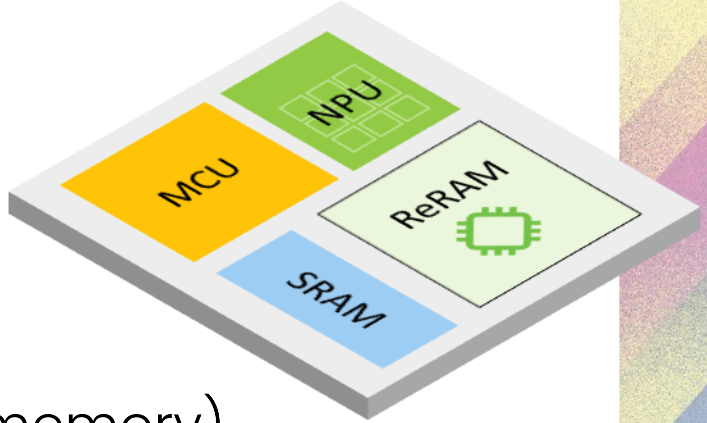
Near Memory Computing – Weebit ReRAM

Use Cases

- Replaces external flash/NVM (on-chip weights, firmware)
- Augments SRAM: reduces standby power in always-on or low-duty-cycle applications

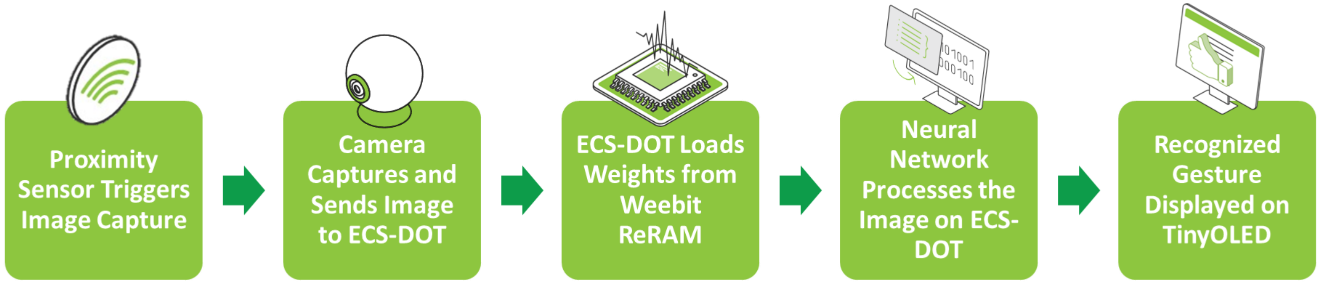
Key Benefits

- Instant-on: no need to reload weights at startup
- Higher performance
- Lower power consumption (extended battery life)
- Lower system cost (reduced SRAM requirements, no external memory)
- Enhanced security (fewer external attack endpoints)
- Better system integration (compact and efficient design)

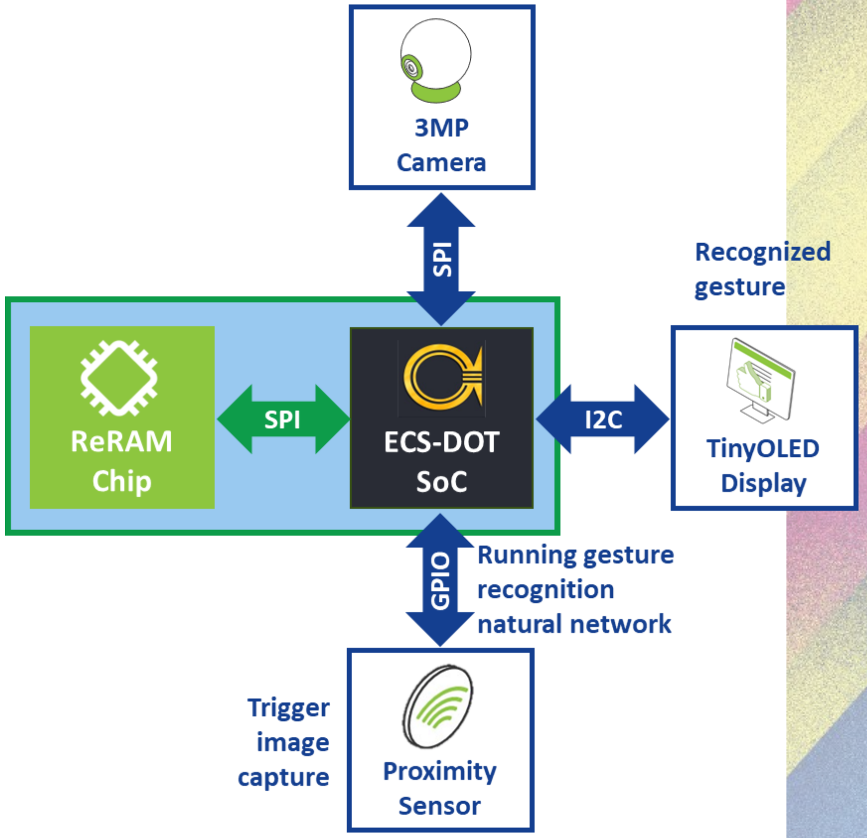


NMC Demonstration using Weebit ReRAM

- Combination of
  - Weebit ReRAM in 22nm technology
  - EMASS' ultra-low power AI SoC
- Demonstration of gesture recognition
- Designed for real-time inference with minimal power consumption
- Highly reduced energy consumption over flash
- Instant wake-up time



Proximity Sensor Triggers Image Capture → Camera Captures and Sends Image to ECS-DOT → ECS-DOT Loads Weights from Weebit ReRAM → Neural Network Processes the Image on ECS-DOT → Recognized Gesture Displayed on TinyOLED



3MP Camera

SPI

ReRAM Chip

SPI

ECS-DOT SoC

I2C

TinyOLED Display

Recognized gesture

GPIO

Running gesture recognition natural network

Trigger image capture

Proximity Sensor

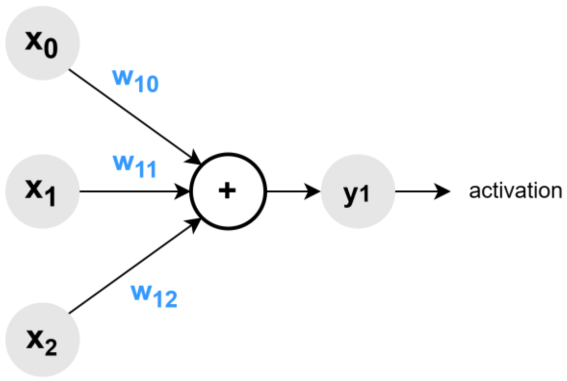
Weebitnano THE NEXT NVM IS HERE

EMASS

cea list

LIST TECH DAYS

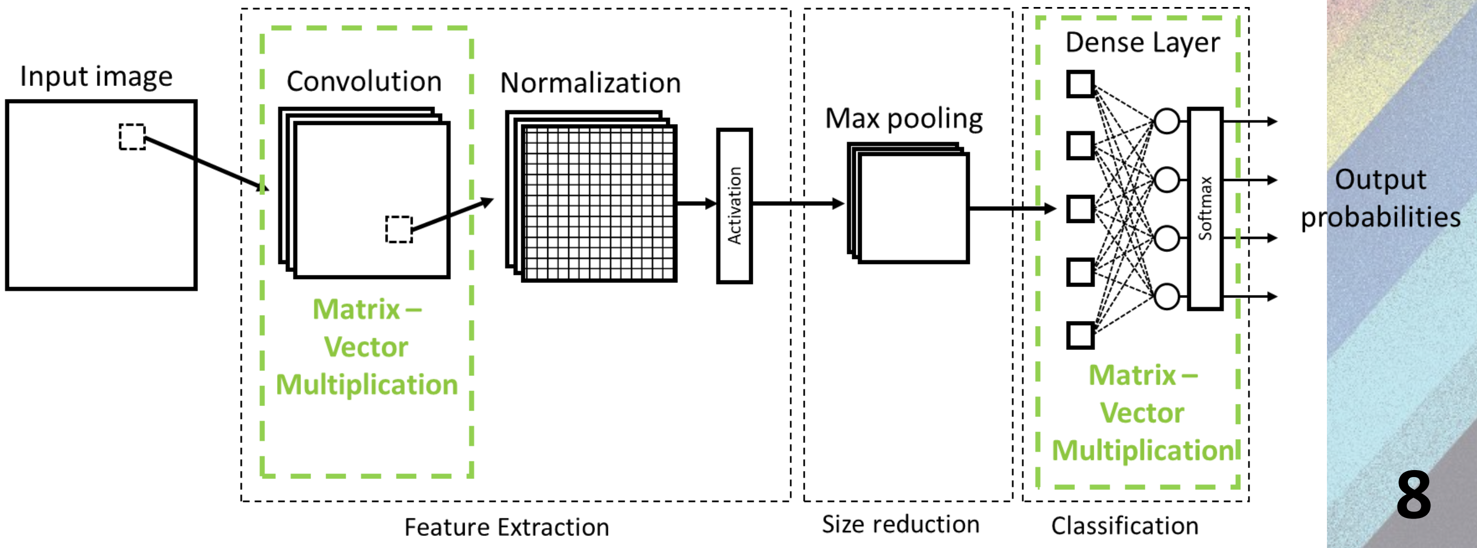
NPU Core Workload: Matrix-Vector Multiplication



- Scalar, Vector and Matrix operations
- Performed by Multiply-Accumulate units
- Weights matrices can be tens of MBs
- Weights must be updated periodically:
  - For algorithm improvements
  - Or more frequently for on-device learning

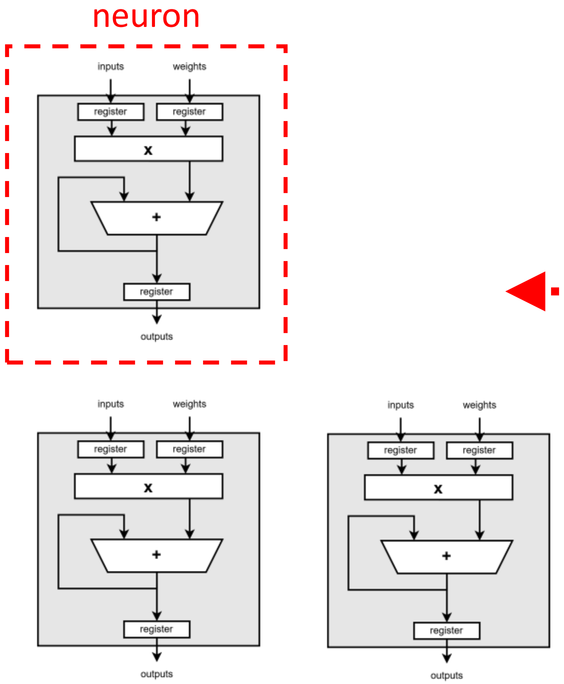
$$\begin{pmatrix} y_0 \\ \boxed{y_1} \\ y_2 \end{pmatrix} = \begin{pmatrix} w00 & w01 & w02 \\ w10 & w11 & w12 \\ w20 & w21 & w22 \end{pmatrix} * \boxed{\begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}}$$

Multiply-Accumulate



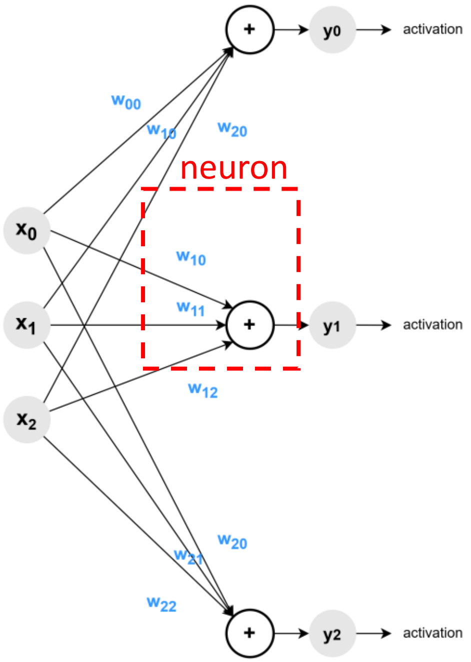
The Perspectives of In-Memory Computing

Array of MAC Structures (digital)

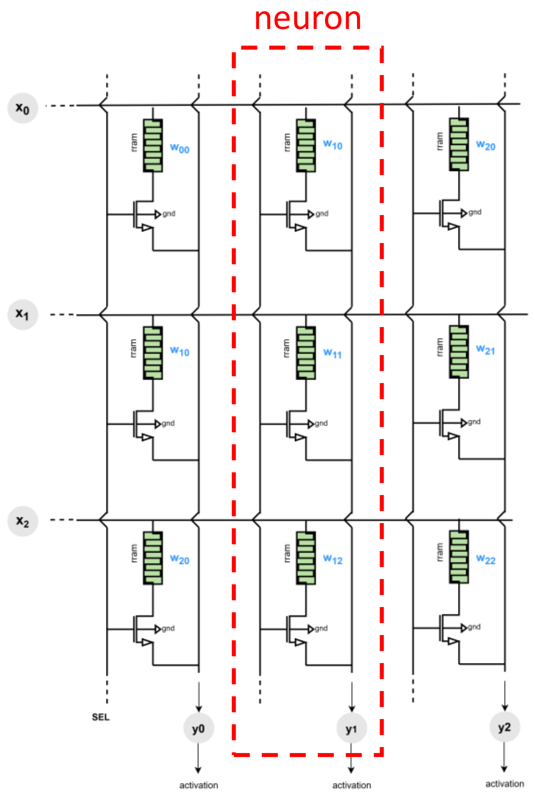


$$TOPS = \frac{2 * N_{MAC} * f_{clk}}{10^{12}}$$

Neural Layer (symbolic)



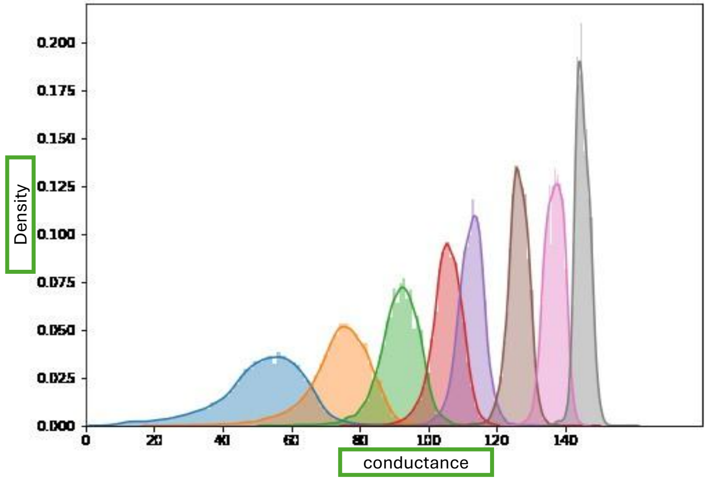
Crossbar Array (analog)



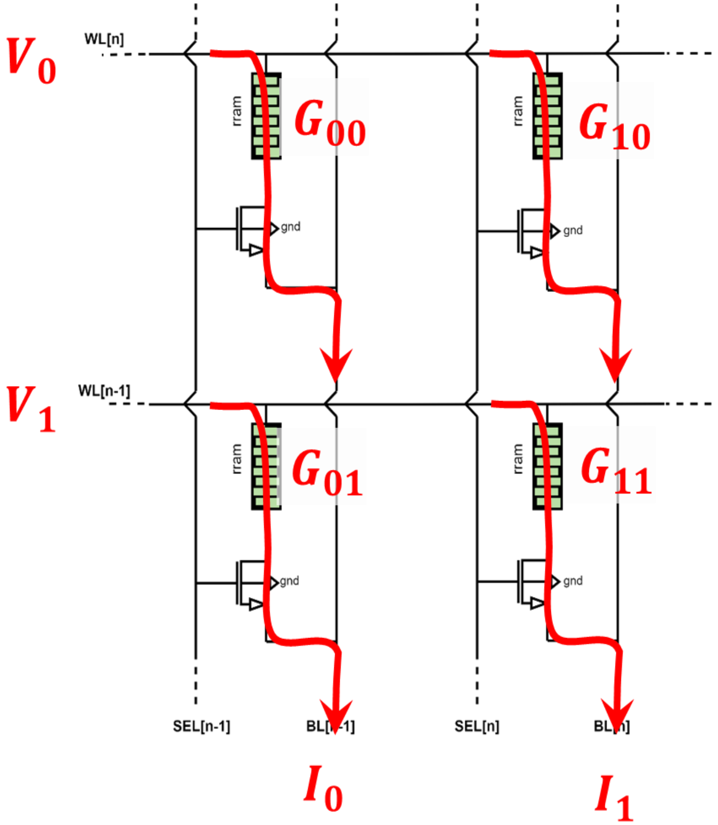
$$TOPS = \frac{2 * N * M * f_{clk}}{10^{12}}$$

Crossbar Array

- Parallel matrix-vector multiply using
  - Ohm’s law (Multiply)
  - Kirchoff’s law (Accumulate)
- ReRAM conductance states define the accuracy of the operation



📖 M. Pallo et al. (2025)



$$\begin{pmatrix} I_0 \\ I_1 \end{pmatrix} = \begin{pmatrix} G_{00} & G_{01} \\ G_{10} & G_{11} \end{pmatrix} \begin{pmatrix} V_0 \\ V_1 \end{pmatrix}$$

IMC Demonstration using Weebit ReRAM

- **Generalized model initialization**

Cloud pre-trained model (θ) obtained via learn-to-learn phase

- **Learning at the edge / user data adaptation**

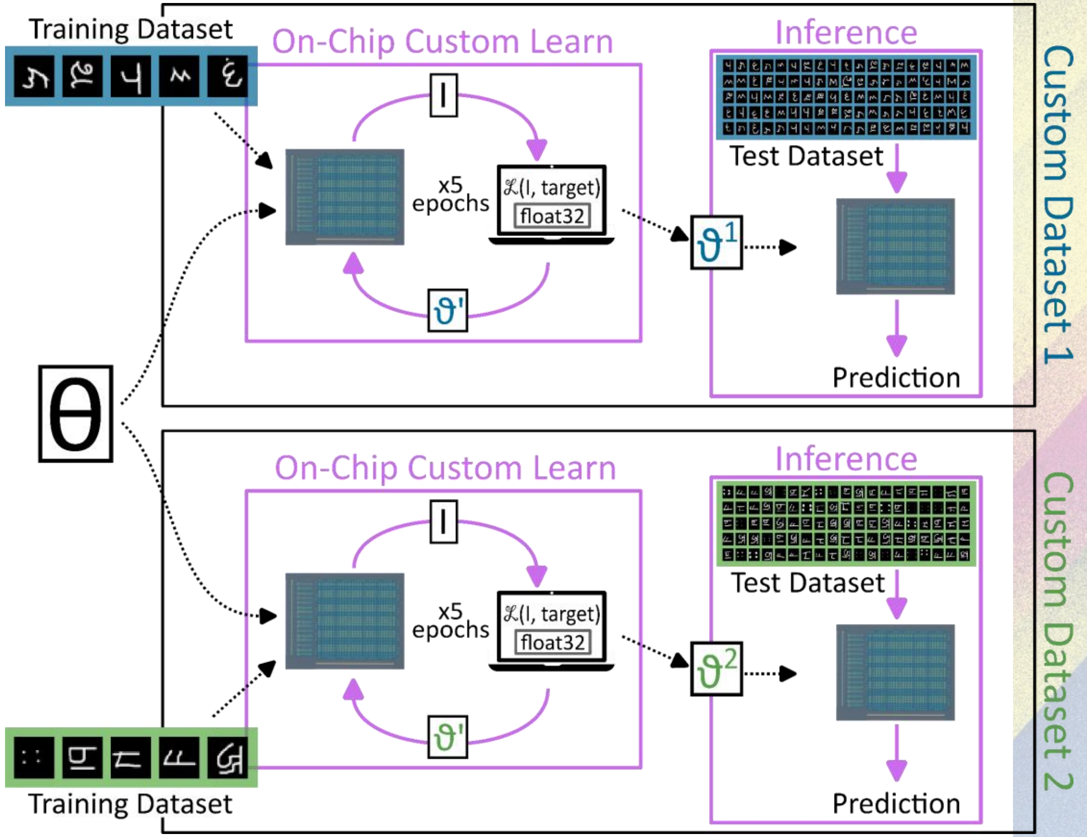
Customers can upload their dataset to personalize the model

- **Rapid fine-tuning**

Only 5 updates needed to adapt the model

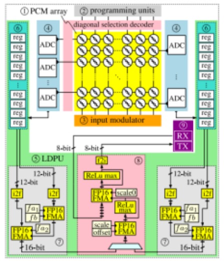
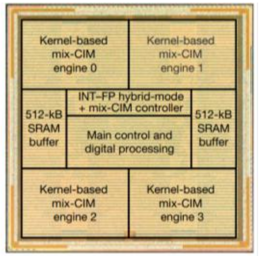
- **Efficient & accurate**

Good classification accuracy with minimal retraining effort



📖 M. Pallo et al. (2025)

IMC – State of the Art

	HERMES IBM, 2024	Mixed-Precision TSMC, 2025
Technology	14nm - PCM	22nm - ReRAM/SRAM
Cell Type	8T4R	1T1R/6T
Array Size	4 M	64 M
Energy Efficiency	~ 9.8 TOPS/W	> 25TOPS/W
Implementation	ResNet-9 LSTM for characters prediction	ResNet-20 MobilNet-v2
Image		

① PCM array
② programming units
diagonal selection decoder
③ input modulator
④
ADC
⑤ LDPU
⑥
reg
⑦
⑧
⑨
RX
TX
8-bit
12-bit
16-bit
i2f
f2i
FP16 FMA
ReLu max
scale0
scale
offset
fa1
fb
fa2

Kernel-based mix-CIM engine 0
Kernel-based mix-CIM engine 1
INT-FP hybrid-mode + mix-CIM controller
512-kB SRAM buffer
Main control and digital processing
512-kB SRAM buffer
Kernel-based mix-CIM engine 2
Kernel-based mix-CIM engine 3

The technology is becoming more mature and larger implementations come out every year.

NVIDIA JETSON ORIN NANO

~ 10 TOPS/W

AXELERA METIS M2

~ 20 TOPS/W

Conclusion – Why ReRAM is a Game Changer for AI

Unified Memory and Compute

- Simplified architecture
- Enables conventional digital processing near memory, reducing reliance on external memories
- Non-volatility ensures persistent data even in power and environment constrained systems
- Minimizes data movement, drastically reducing energy per operation
- Reduces memory-access latency, enabling faster processing and instant-on systems
- Straightforward integration into existing SoC/MCU/CPU/NPU designs
- Leaner system design: Eliminates need for complex flash controllers and charge pumps

In-Memory and Analog AI Computation

- Supports matrix-vector multiplication directly in ReRAM crossbar arrays
- Enables fast, local AI inference, especially valuable for low-power edge applications
- High endurance and write speed suitable for on-device learning, frequent updates, and adaptive AI tasks

Thank You!

Weebitnano

THE NEXT NVM IS HERE

cea list

LIST TECH DAYS