Weebitnano
THE NEXT NVM IS HERE

Neuromorphic Computing: A Primer on the Path to Ultra-Efficient AI

Gideon Intrater, Strategic Advisor, Weebit Nano

November 5, 2025

embeddedworld North America
Exhibition&Conference

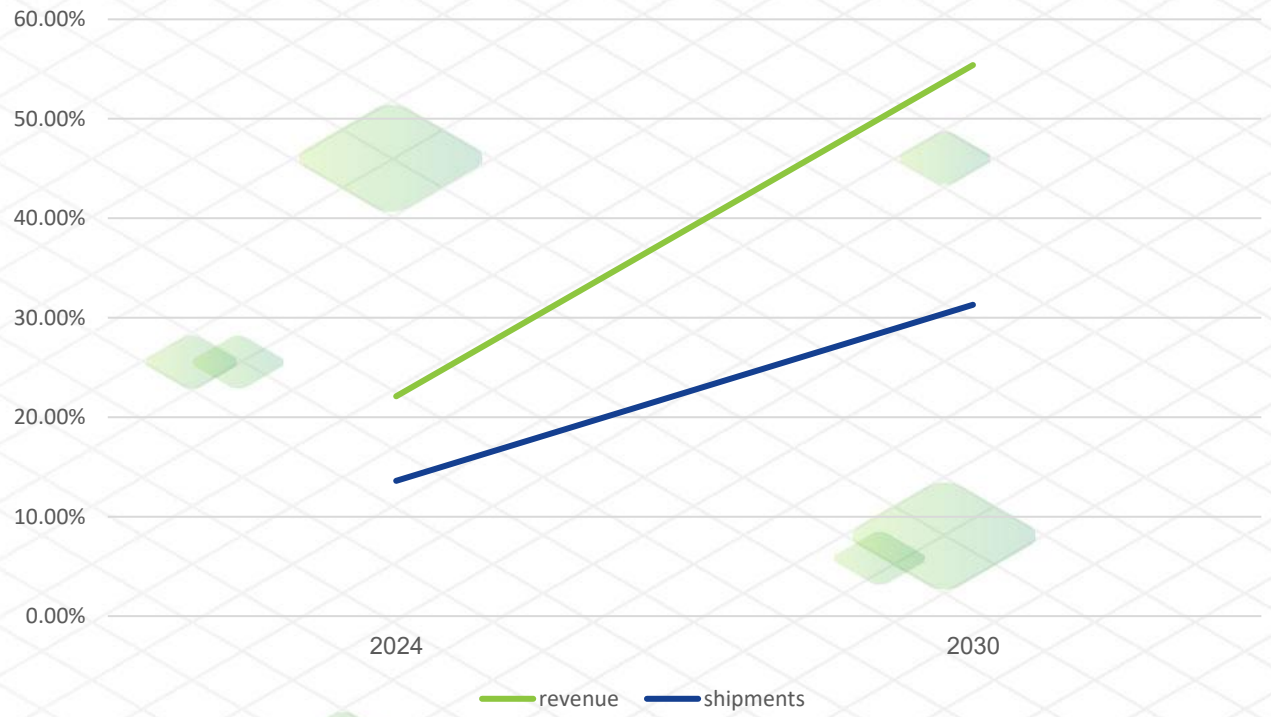
Agenda

- AI at the edge
- The need for new approaches
- The promise of neuromorphic
- Introduction to ReRAM
- NVM for In-Memory Compute
- The journey to new architectures for edge AI

AI at the Edge

- Traditionally, a significant portion of **AI inference** has been performed in the cloud
- AI inference is increasingly local (edge), driven by:
  - Real-time processing
  - Power efficiency
  - Ultra-low latency
  - Low bandwidth
  - Security/privacy
  - Smarter products; new applications

Edge AI SoC Market Penetration*



Edge-AI devices as a percentage of SoCs in the market

*Edge-AI Market Analysis: Applications, Processors and Ecosystem Guide, The SHD Group, April 2025

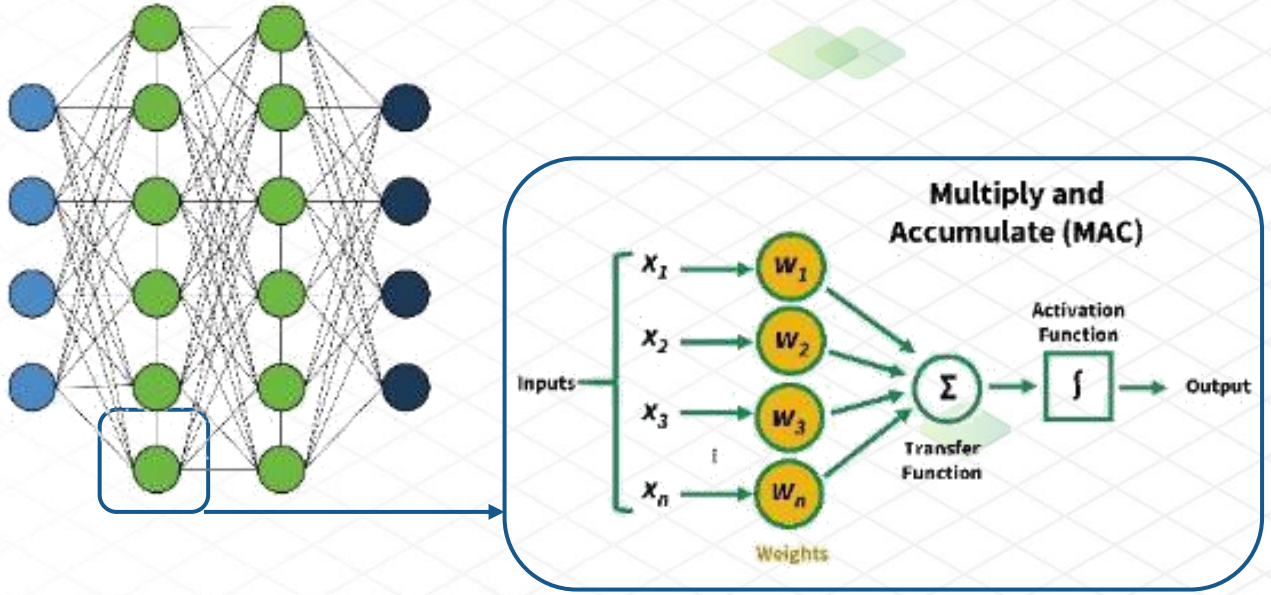
Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

AI Inference Basics

Most of the processing in AI are multiplications of vectors times a fixed weight matrix

- Weight matrix is large, typically up to 10s of MBs
- Weights need to be modified from time to time for updating the algorithms

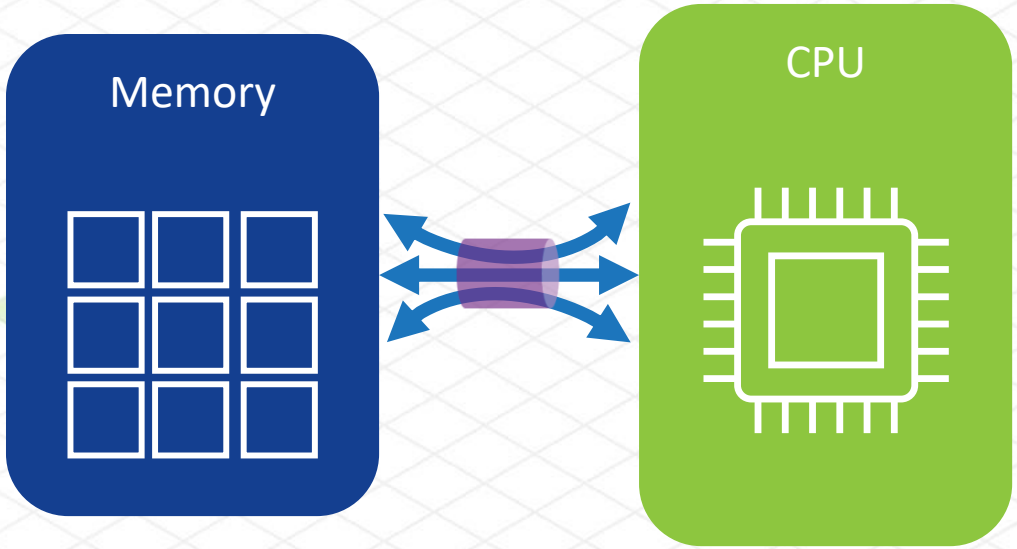


Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

von Neumann Bottleneck

- Traditional von Neumann architectures
  - Consume significant power due to continuous data movement between memory and CPU
  - Limited bandwidth slows down AI computation
- AI systems spend most of their energy moving and accessing data
  - Power consumption is a primary concern
  - Workloads are power and bandwidth-hungry
- Traditional architectures can’t scale
  - Can handle relatively simple algorithms
  - But the power consumption limits their scaling



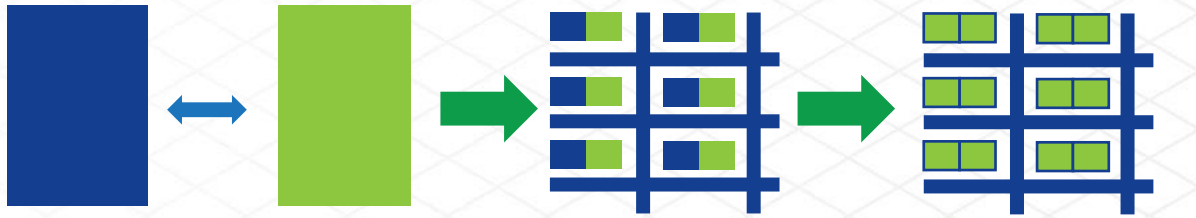
Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

Non - von Neumann Computing

- Non-von Neumann systems to bring memory closer to computation
  - Better efficiency and lower power consumption

- In Memory Compute (IMC): computation performed within memory arrays
  - Eliminates data traffic between memory and CPU → ultra-low power
  - Enables analog compute → store larger models efficiently



Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

The Promise of Neuromorphic Computing

- Inspired by biology
  - Mimics structure of neurons and synapses
  - Process vast amounts of data in **real time** at **ultra-low power**
- When coupled with IMC, overcomes limitations of von Neumann architectures
  - IMC: **no data movement** between neurons
  - **Highly parallel**: process data simultaneously across multiple neurons
  - Asynchronous: consuming power only when event triggers
  - **Scalable**: power and speed headroom offered by IMC enables compute-demanding algorithms
- At the edge: new capabilities in ultra-low-power devices
  - AI-driven predictive maintenance in smart IoT devices
  - Wearable AI for personal, real-time assistance
  - Self learning robots with real-time AI perception

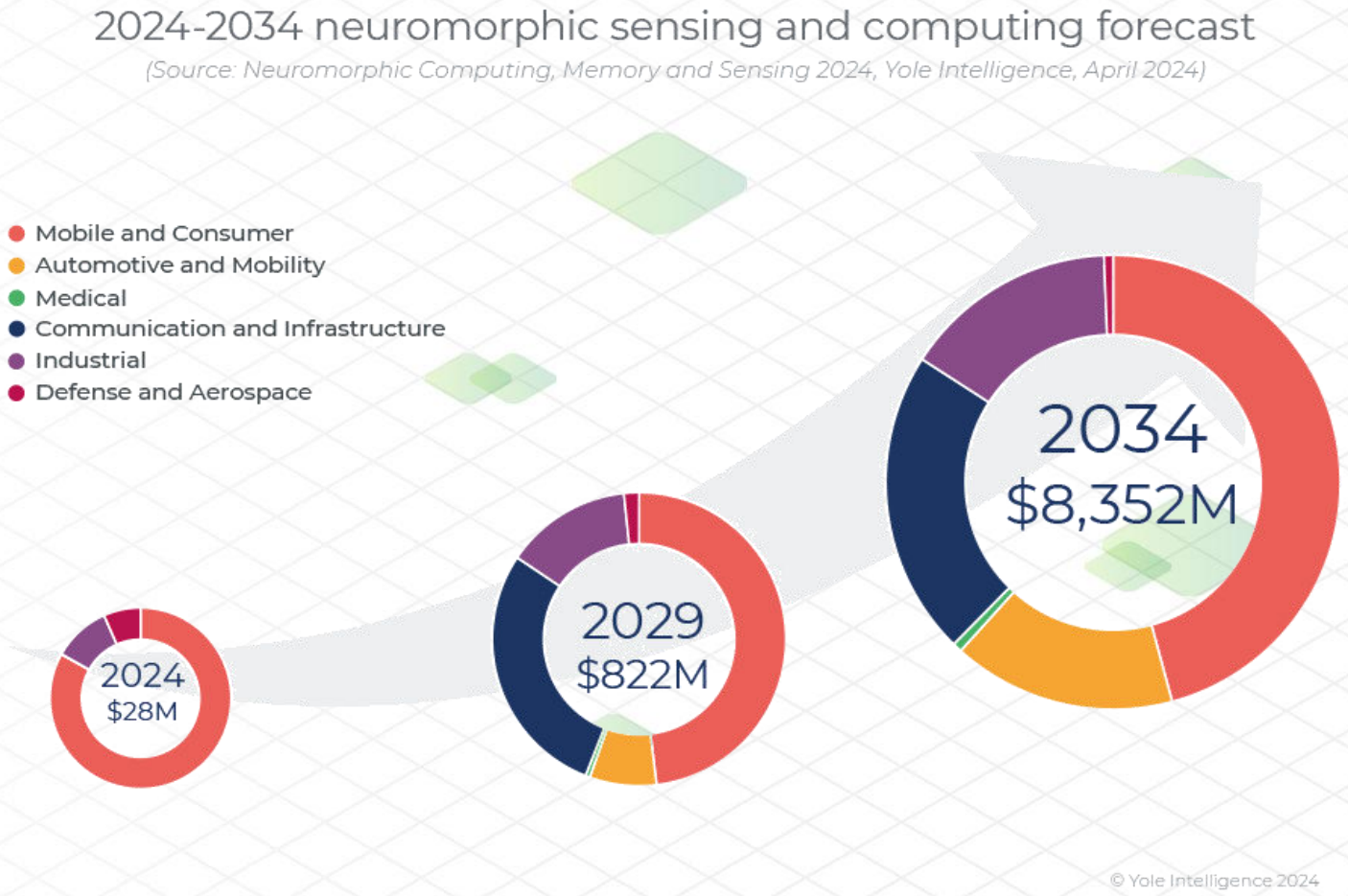
Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

The Neuromorphic Market is Set to Take Off

- Analog IMC and other IMC approaches with ReRAM/emerging memory will ramp up starting >2027 (Yole)

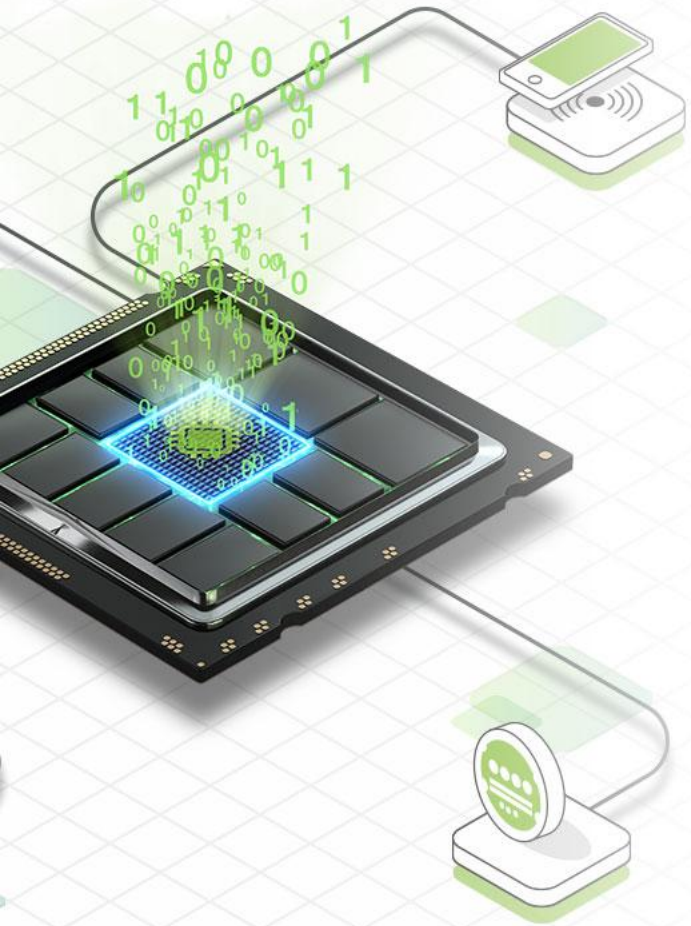
- “neuromorphic technologies show promise for **sustainable and efficient AI processing at the edge**.” (Yole)



Source: https://www.yolegroup.com/product/report/neuromorphic-computing-memory-and-sensing-2024

Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.



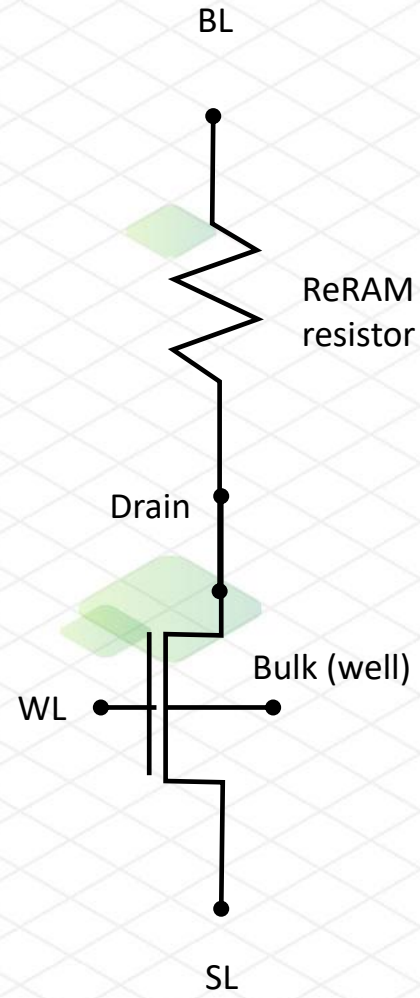
ReRAM for Edge AI

Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

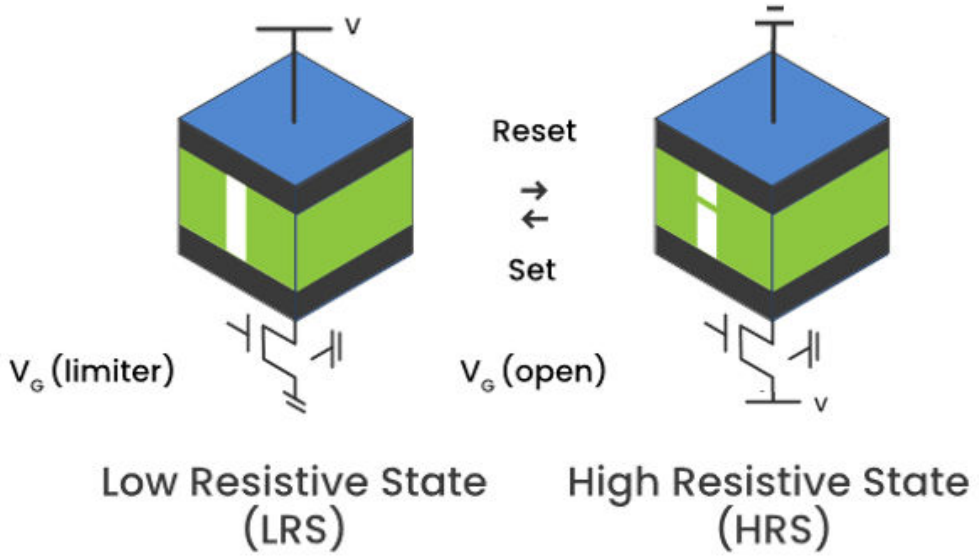
ReRAM Basic Operation

- ReRAM is based on oxygen vacancies filament (OxRAM)
  - By applying different voltage levels on the resistive layer, a filament is created or dissolved
    - RESET (Erase) – Partial dissolution of the Conductive Filament: LRS –> HRS
    - SET (Program) – Recreation of the Conductive Filament: HRS->LRS
  - Data retained within the stack is resilient to many environmental conditions



Low Power Consumption

- ✓ Low read voltage <1V
- ✓ Low write voltage <3V
- ✓ Low currents
- ✓ Zero standby power
- ✓ Fast operation

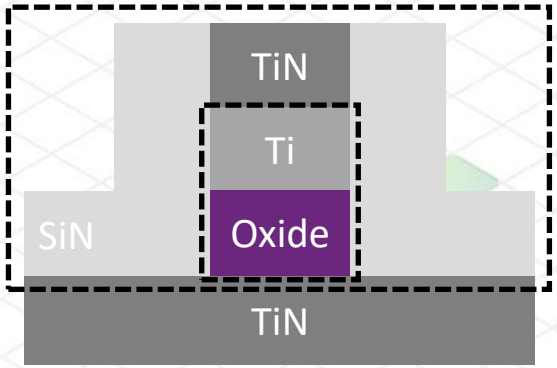


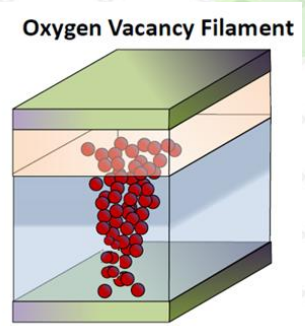
Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

The Most Cost-Effective NVM Solution

- 2-mask adder
  - Very few added steps compared to other NVM technologies
  - Lower wafer cost than competing NVM technologies
- Fab-friendly materials
  - No contamination risk, No special handling, etc.
- Using existing deposition techniques and tools
  - Easy to integrate into any CMOS fab
- BEOL technology
  - Stack between any 2 metal layers
  - No interference with FEOL – Easier to embed with existing Analog and RF circuits
  - Easy to scale from one process variation to another



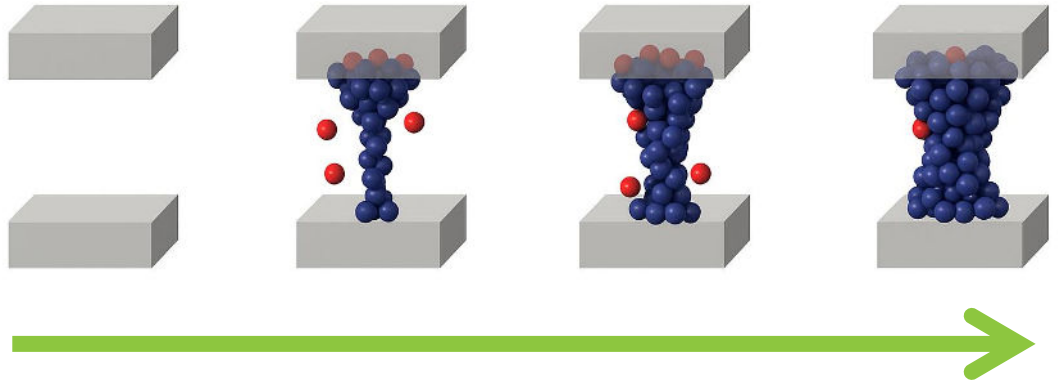


Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

Controlling the Resistance of ReRAM Cells

- The resistance of the ReRAM cell is a function of the size of the filament
- Programming of the cells is done through applying positive or negative pulses on the electrodes
- Control over the size of the filament is done through the amplitude and duration of the programming pulses



Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

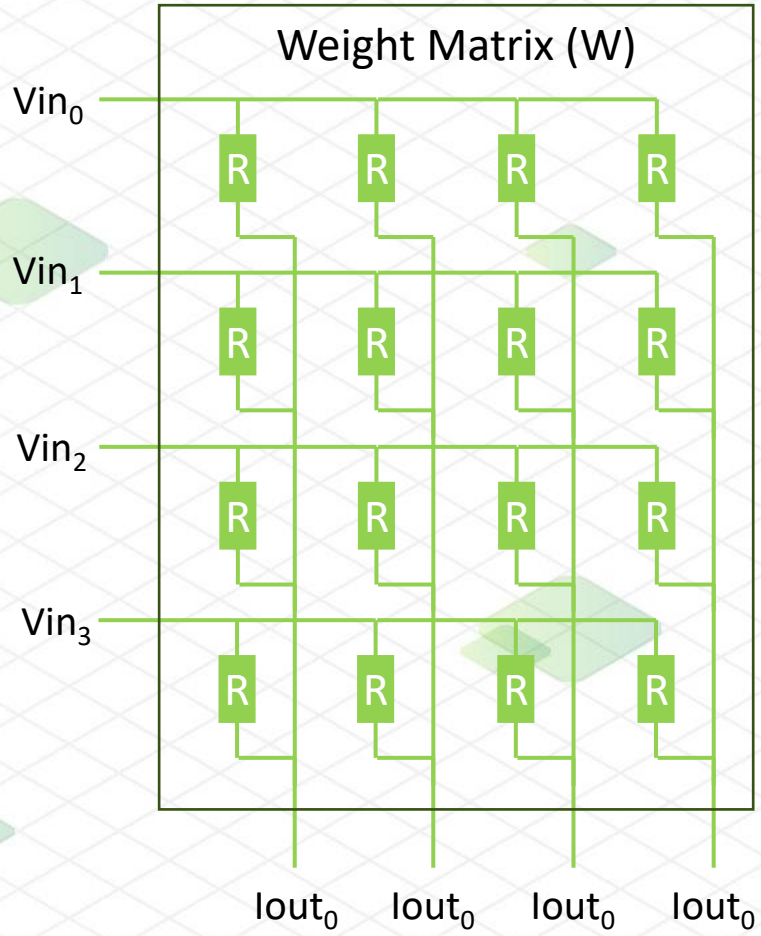
Analog In-Memory Compute with ReRAM

- Most of the processing in AI are multiplications of the input vector times a fixed weight matrix
- Resistor arrays built out of ReRAM elements perform these operations instantly:
  - The weights are represented by the cells' conductance G = $\frac{1}{R}$
  - Using Ohm's law each the current through each resistor is:

$$Iout = Vin \times \left(\frac{1}{R}\right) = Vin \times W$$

  - Using Kirchhoff's law, the current in each column is:

$$Iout_i = \sum_{j=0}^{j=n-1} W_{ij} Vin_j$$



© 2025 Weebit Nano Ltd. All rights reserved.

Other Approaches to Analog In-Memory Compute

- Fixed resistors
  - Provide great resolution, accuracy and stability
  - Work well for single-function SoC, not for multifunction ones
  - Impossible to update the weights in the field as any change requires a new chip
- Implement the resistors using flash cells
  - Flash offers a high dynamic range, thereby can provide weights with a high resolution
  - However embedded flash can scale only to 28nm
    - Limiting the size of the implementable AI networks to the simplest algorithms
- Implement the resistors using MRAM cells
  - MRAM cells are not ideal for analog IMC
    - Offering a very low dynamic range, and limited multilevel stability

© 2025 Weebit Nano Ltd. All rights reserved.

NVM for Edge AI

ReRAM is the NVM Best Positioned to Lead the Way

		External Flash	PCM	MRAM	ReRAM
Performance	Low power solution	🔴	🔴	🟢	🟢
	Fast programming	🔴	🟢	🟢	🟢
	High endurance	🟡	🟢	🟢	🟢
Cost	Low cost to manufacture	🟡	🟡	🔴	🟢
	Small die size	🟡	🟢	🟢	🟢
Reliability	High-temp reliability	🟢	🟢	🟢	🟢
	EMI immunity	🟢	🟢	🔴	🟢
Scalability	Scalable to advanced nodes	🔴	🟢	🟢	🟢
Integration	Ease of fab integration	🟡	🟡	🔴	🟢
Maturity	Mature technology	🟢	🟡	🟡	🟡

Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

Neuromorphic Computing with ReRAM

- ReRAM can be used to emulate biological synapses
  - Enabling efficient long-term memory retention
  - Adaptive learning
  - Ultra-low-power AI model storage
- Analog neuromorphic architecture with ReRAM
  - Exploits in-memory computing
  - Low-latency operations
  - Low-power operations
  - Suits most algorithms
  - A great deal of research is underway
  - Gaining maturity

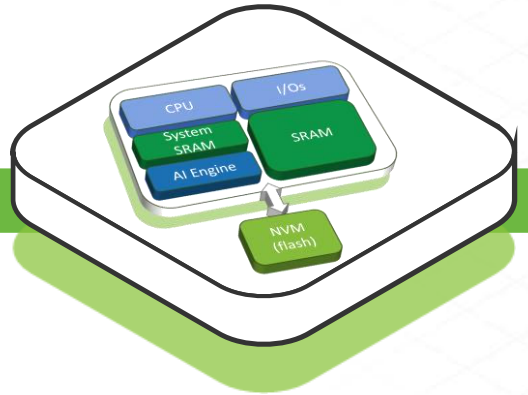
Advantages of ReRAM for IMC
Nonvolatile behavior
Good scaling
Integration at the back end of the line (BEOL)
Low current consumption
Ability of multilevel cell (MLC) operation

Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

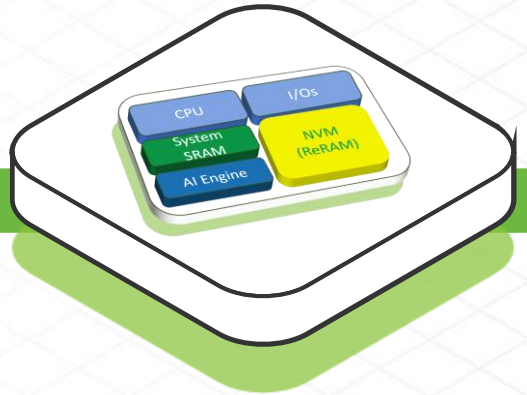
The Journey to New AI Architectures with ReRAM

TODAY'S TWO CHIP SOLUTION



- Wasteful in terms of size & cost
- Prohibitive power
- Constrained data bandwidth and performance
- Insecure, vulnerable to hacking

SINGLE CHIP NEAR-MEMORY COMPUTE



External NVM eliminated

- Eliminates most data movements → Low-power
- More efficient storage: **4X greater capacity** than SRAM

IN-MEMORY COMPUTE WITH RERAM

Computation performed *within* memory arrays

- **Completely eliminates** data traffic → Ultra-low power
- Enables **Analog Compute** → Store larger models efficiently

NEUROMORPHIC COMPUTE WITH RERAM

Future systems will mimic brain behavior

- Fast real-time processing on **massive amounts of data**
- Three orders of magnitude **(x1000) better energy efficiency**

ReRAM-based architectures are central to the transformation of AI,
bringing memory and compute together for faster, more brain-like intelligence

Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

Looking Ahead

- ReRAM represents an **ideal foundation for future IMC and neuromorphic architectures**
  - 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
  - Weebit is working with NeMo Consortium, NeAIxt, Edge AI Foundation, and universities on driving forward ReRAM development towards IMC and neuromorphic
  - With our partner cea leti Weebit has already demonstrated spiking neural networks with ReRAM

- IMC architectures hold a great deal of promise in the not-so-distant future; neuromorphic architectures are in the very early stages but advancing steadily
  - Challenges remain in limited precision, conductance drift over time and temperature, density scaling, manufacturing, interconnect complexity, efficient materials, programming models, software ecosystem (to name a few)

We welcome collaboration to advance research in these areas. Let's explore opportunities together!

Weebitnano
THE NEXT NVM IS HERE

© 2025 Weebit Nano Ltd. All rights reserved.

Thank You!

www.weebit-nano.com

Weebitnano

THE NEXT NVM IS HERE