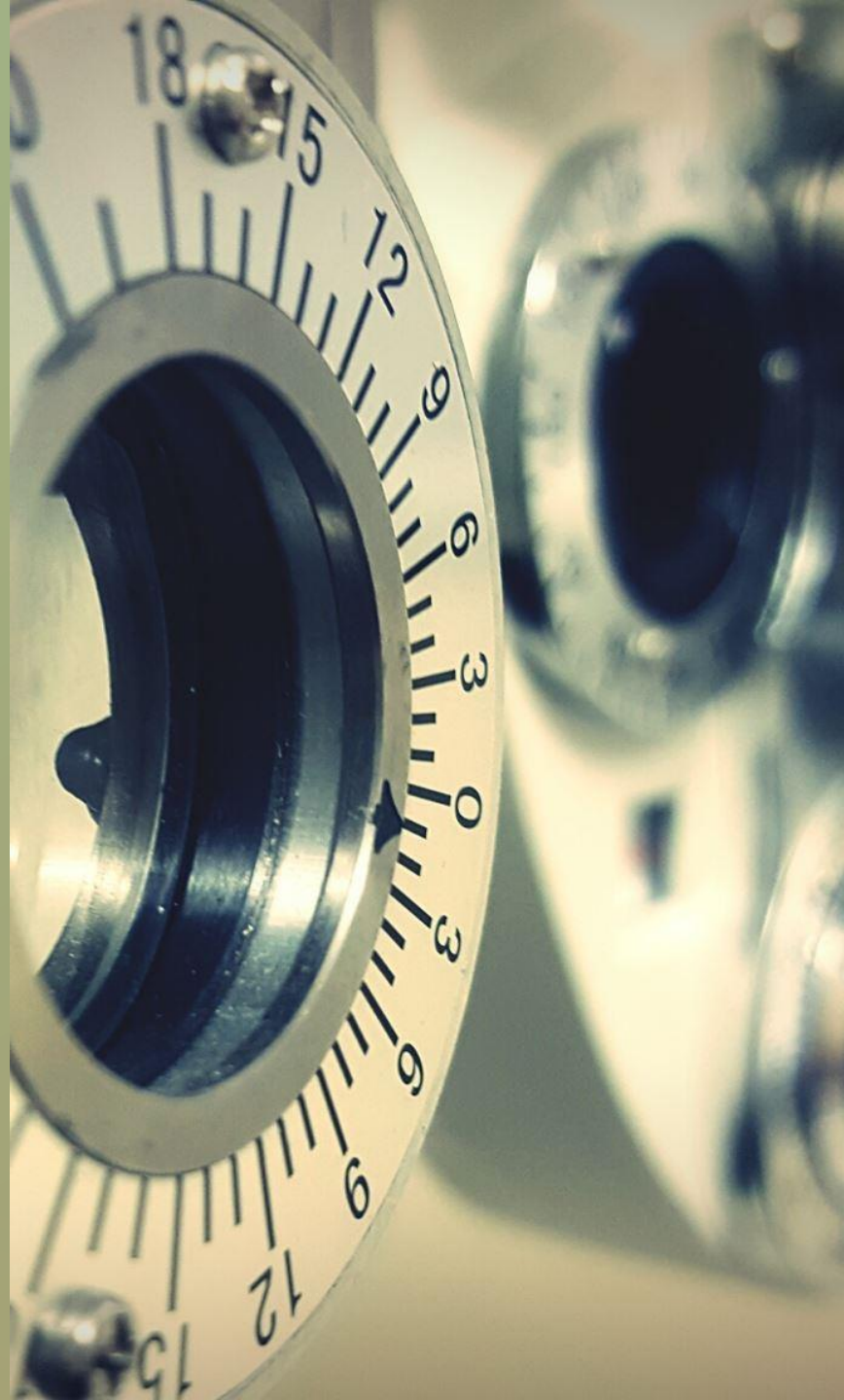
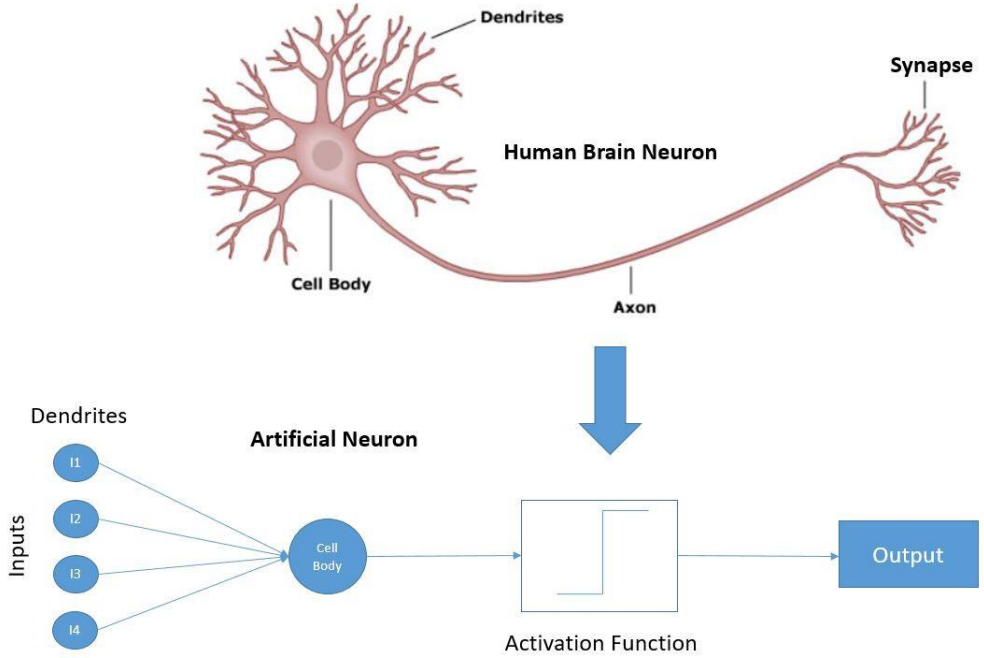
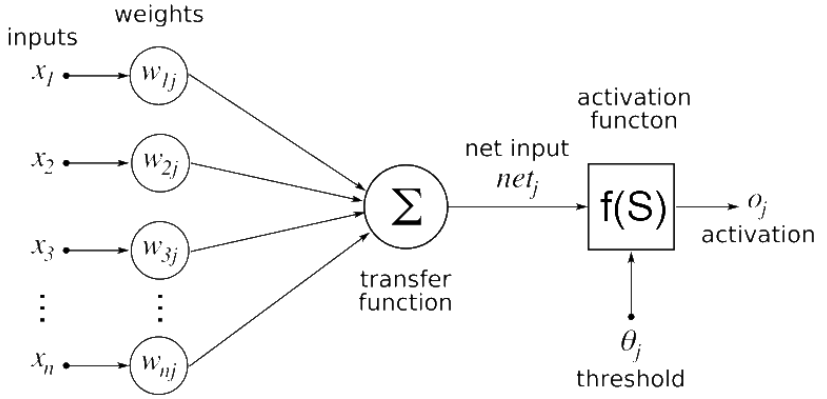
ANALOG DEEP LEARNING



NEURAL NETWORKS





Activation Functions

Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$

tanh
$\tanh(x)$

ReLU
$\max(0, x)$

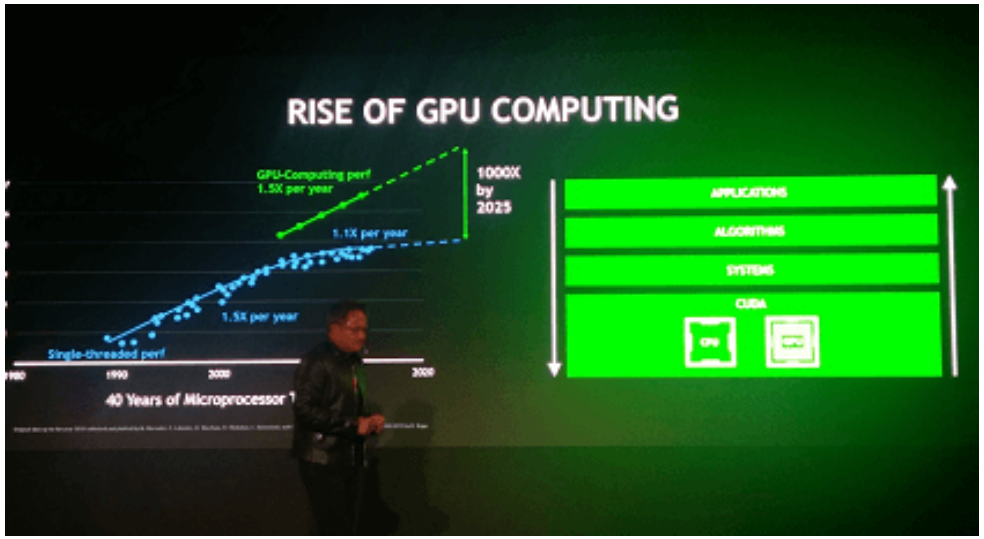
Leaky ReLU
$\max(0.1x, x)$

Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

THE END IS NIGH

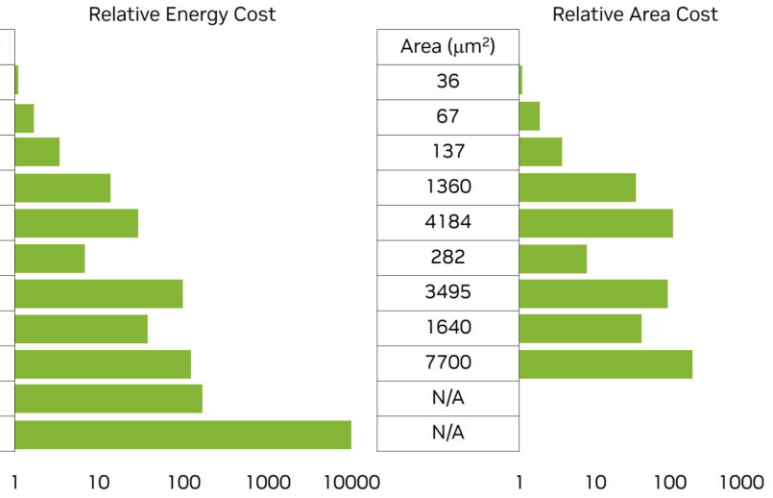
Physics Limits Transistor Size



Precision Can Only Go So Low

Cost of Operations

Operation:	Energy (pJ)	Area (μm²)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A



ANALOG COMPUTING

- The idea is to use electrical circuits to emulate arithmetic, differential equations, integration
- Replace Boolean math with continuous variables represented by currents, voltages, and charges
- Example: https://courses.engr.illinois.edu/ece486/fa2023/laboratory/docs/lab1/analog_computer_manual.pdf

Example 1.2: Determine the analog diagram and circuit to implement the equation

$$V_o = -0.35V_1 + 5.24V_2 + 2.6. \quad (1.17)$$

Solution: The analog diagram is given in Figure 1.12 and the electrical circuit in Figure 1.13. A GP-6 wiring diagram is supplied in Figure 1.14 to illustrate the actual connections needed.

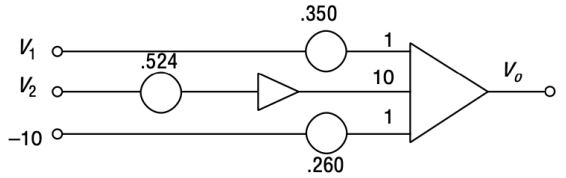


Figure 1.12: Analog Diagram for Example 1.2.

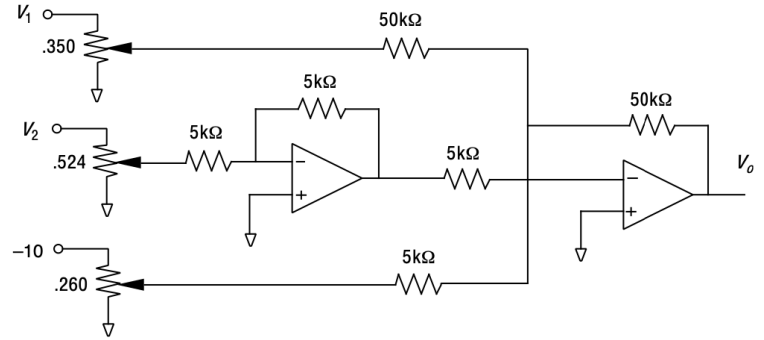
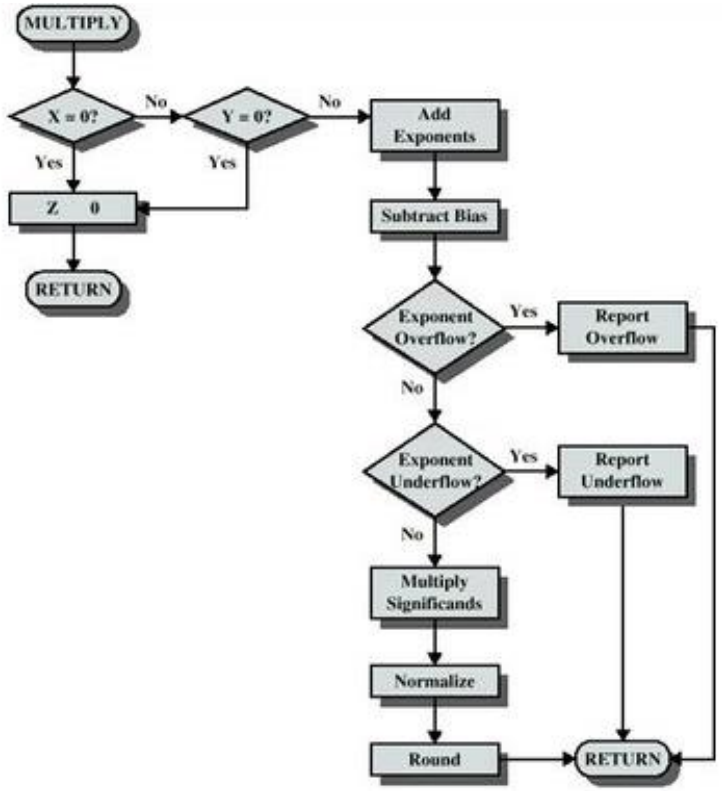
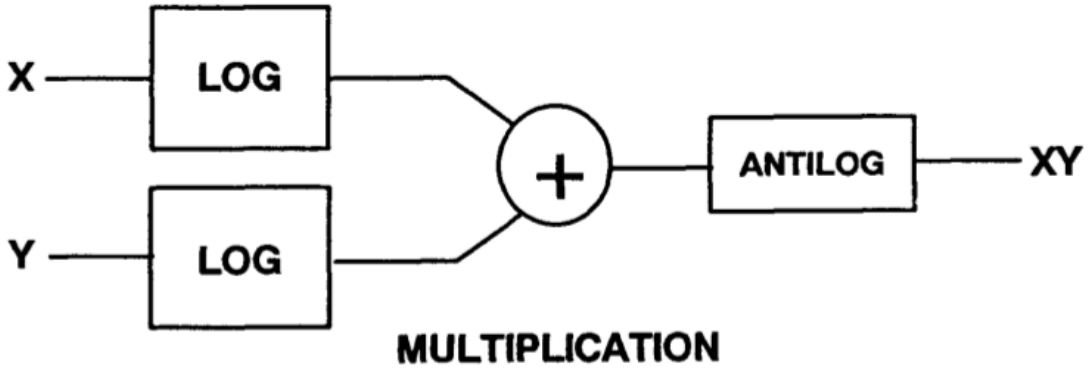


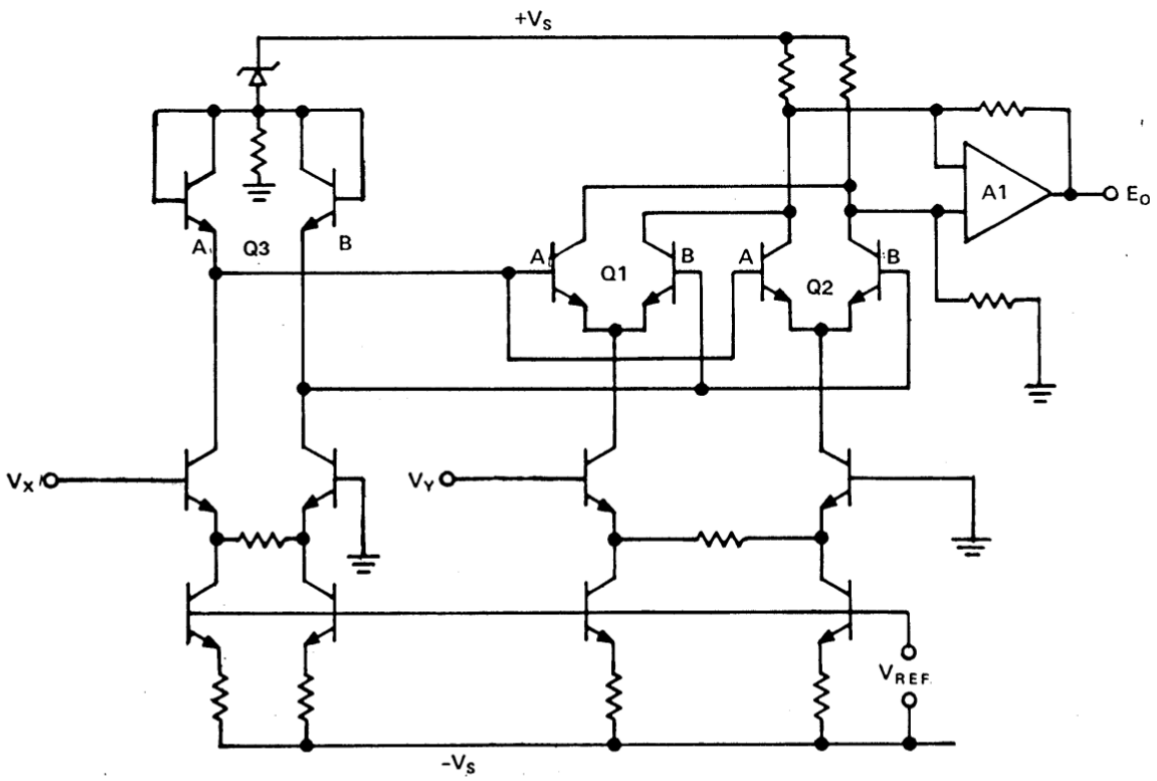
Figure 1.13: Circuit Diagram for Example 1.2.

DIGITAL VS ANALOG ARITHMETIC





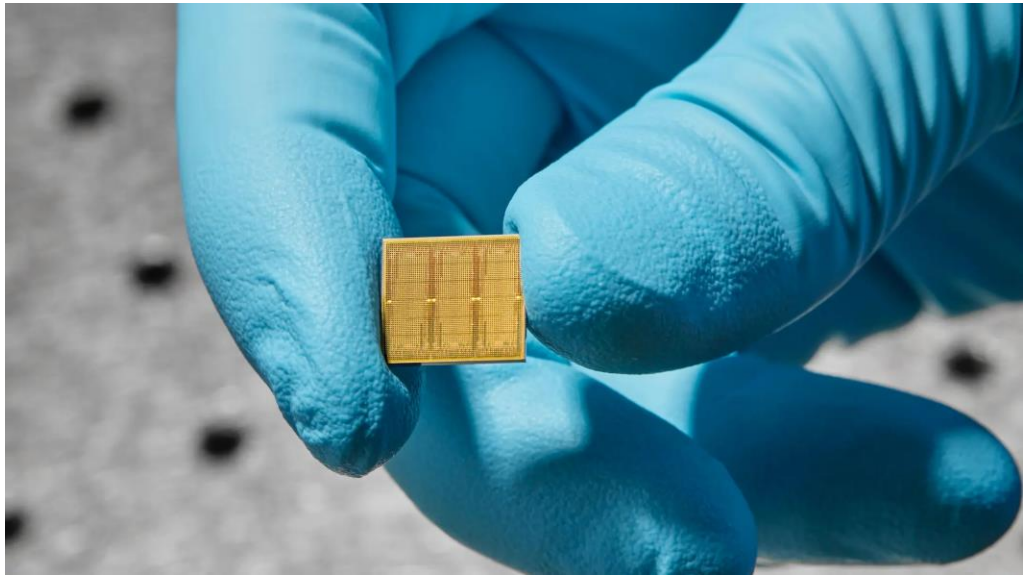
ANALOG MULTIPLICATION CIRCUIT



- The operand values are represented by the input and output voltages
- Transistors and operational amplifiers implement logarithmic and exponential functions
- Far fewer transistors than a digital circuit; result decided by physics

IBM ANALOG IN-MEMORY CHIP

- Nature article: submitted 12/2022, published 08/2023
  https://www.nature.com/articles/s41586-023-06337-5
- Summary:
  https://research.ibm.com/blog/analog-ai-chip-low-power



LOW-POWER, FAST AI

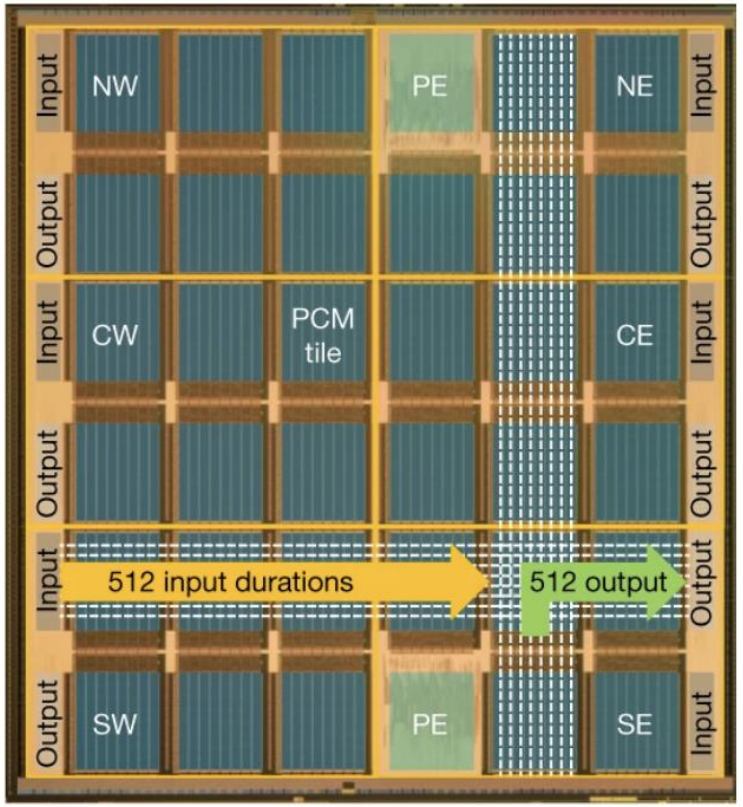
- Ideas:
  - Analog multiplication and addition = low energy
  - In-memory compute = log power
  - Data parallelism of fully-connected DNNs = fast
  - NVMe memory for weights, analog peripheral circuitry
- Preliminary results:
  - 12.4 TOPS/W (est. 14x better than digital circuitry)

PHASE CHANGE MEMORY

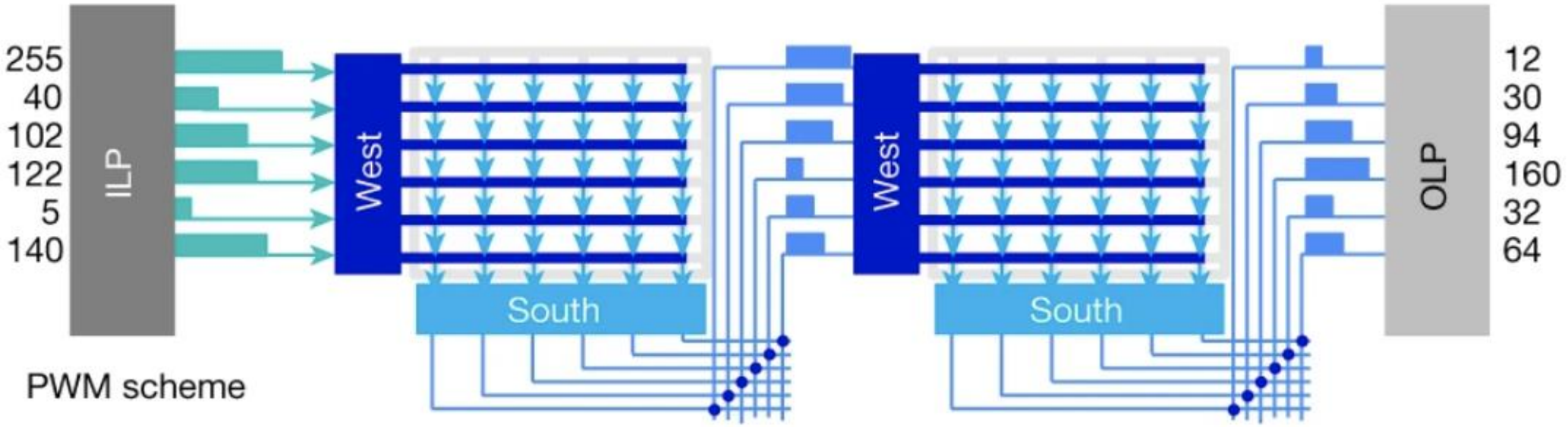
“Phase-change memory (PCM) works when an electrical pulse is applied to a material, which changes the conductance of the device. The material switches between amorphous and crystalline phases, where a lower electrical pulse will make the device more crystalline, providing less resistance, and a high enough electrical pulse makes the device amorphous, resulting in large resistance. Instead of recording the usual 0s or 1s you would see in digital systems, the PCM device records its state as a continuum of values between the amorphous and crystalline states... The memory is non-volatile, so the weights are retained when the power supply is switched off.”

IBM CHIP ARCHITECTURE

- Six tiles storing model weights
- Six ILP/OLP (input/output landing planes) pairs converting UINT8 inputs to elements of pulse-modulated durations
- A 512x2048 PCM crossbar in every tile
- After MAC, the charge on peripheral capacitors is converted into durations and sent either to other tiles or to the OLP

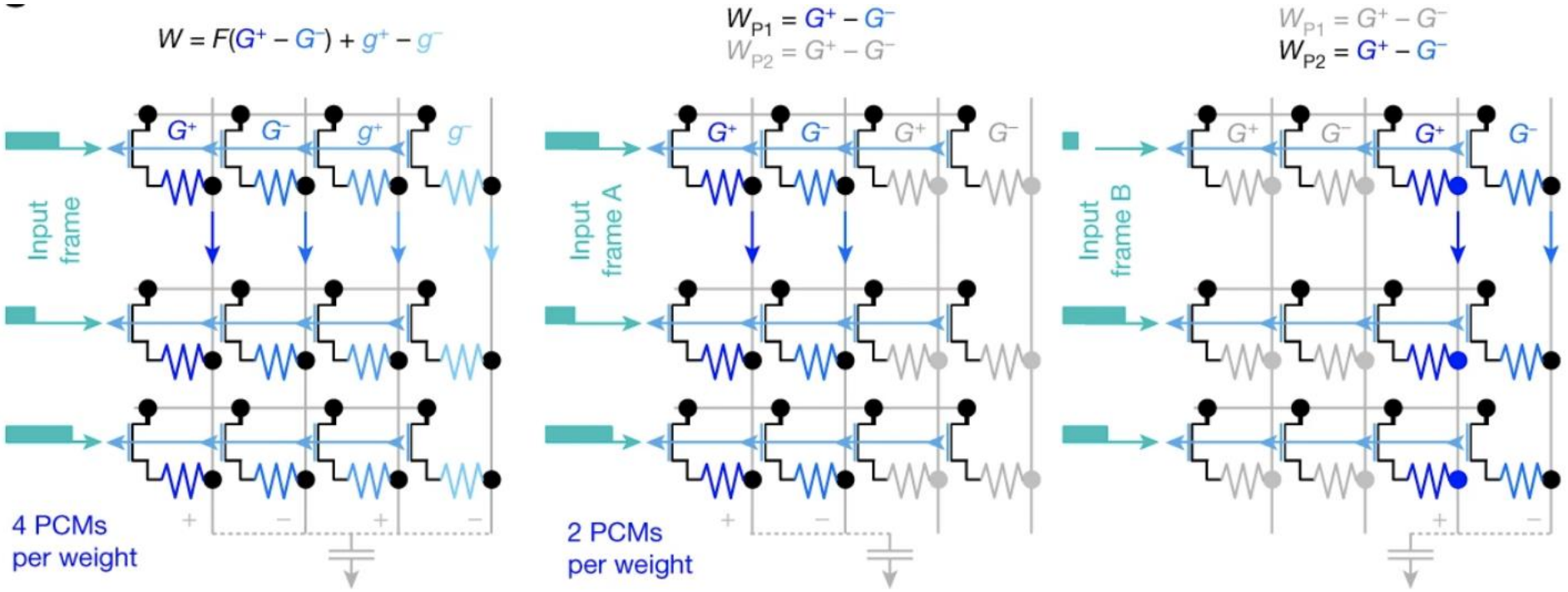


PULSE WIDTH MODULATION (PWM)



PWM scheme

WEIGHT ENCODING WITH PCM

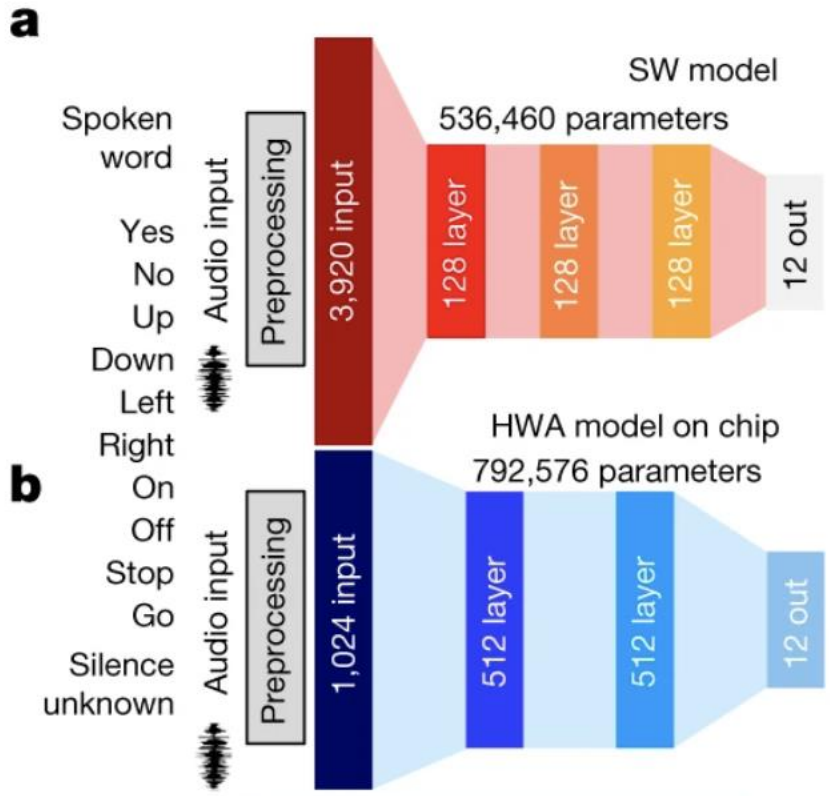


NOT A COMPLETE PRODUCT

- No on-chip computing cores or static random access memory for auxiliary operations, so conducting aux operations off-chip
- ReLU can be implemented in the analog domain

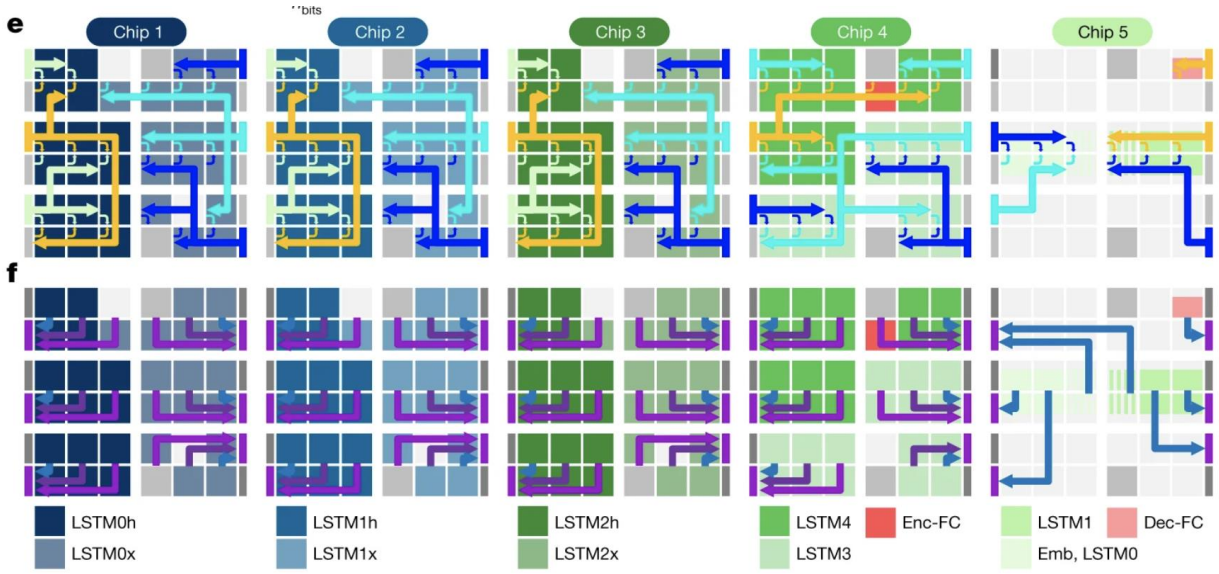
TEST 1: KEYWORD SPOTTING (KWS)

- Take a pre-trained fully-connected (FC) network
- Retrain using HWA techniques to make it more resilient to analog noise
- L2 regularize and bias remove
- Prune to 1024 inputs



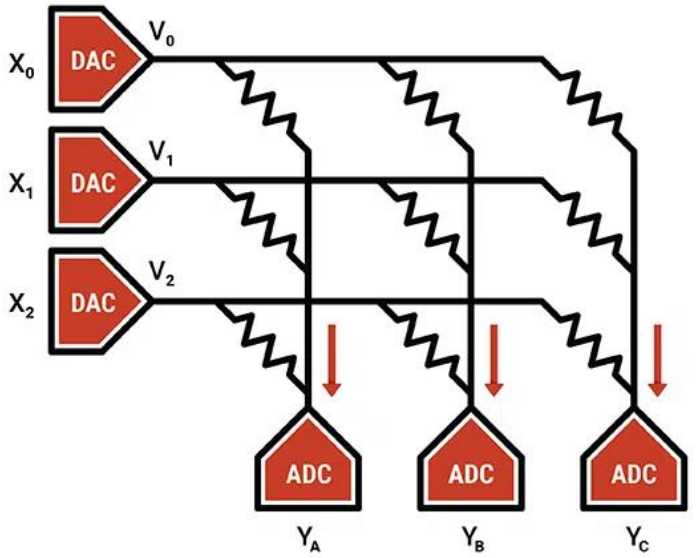
TEST 2: RNNT

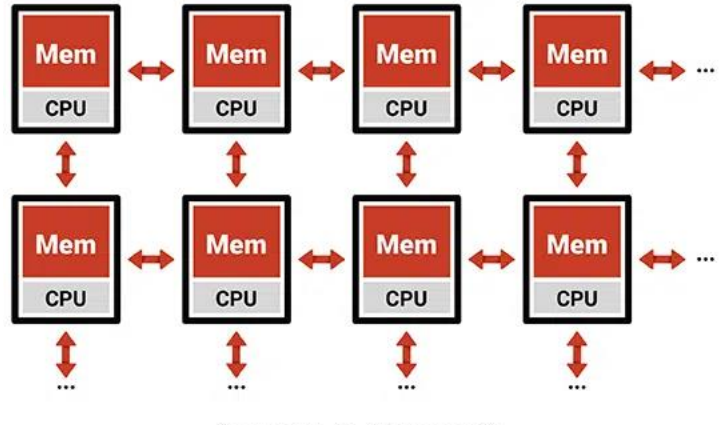
- Five chips combined
- 45 million weights using >140 million PCMs (avg 2.9 PCM per weight)
- 4h 20min of audio processed in 1.29s (omitting pre-processing)
- 14x (projected) system energy efficiency improvement over the best result submitted to MLperf

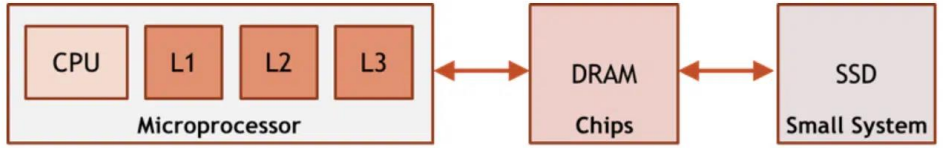


OTHER PLAYERS: MYTHIC

- https://mythic.ai
- Similar architecture to IBM's chip
- Out of Umich
- Conventional flash memory, stretched to store 8-bit resolution charge



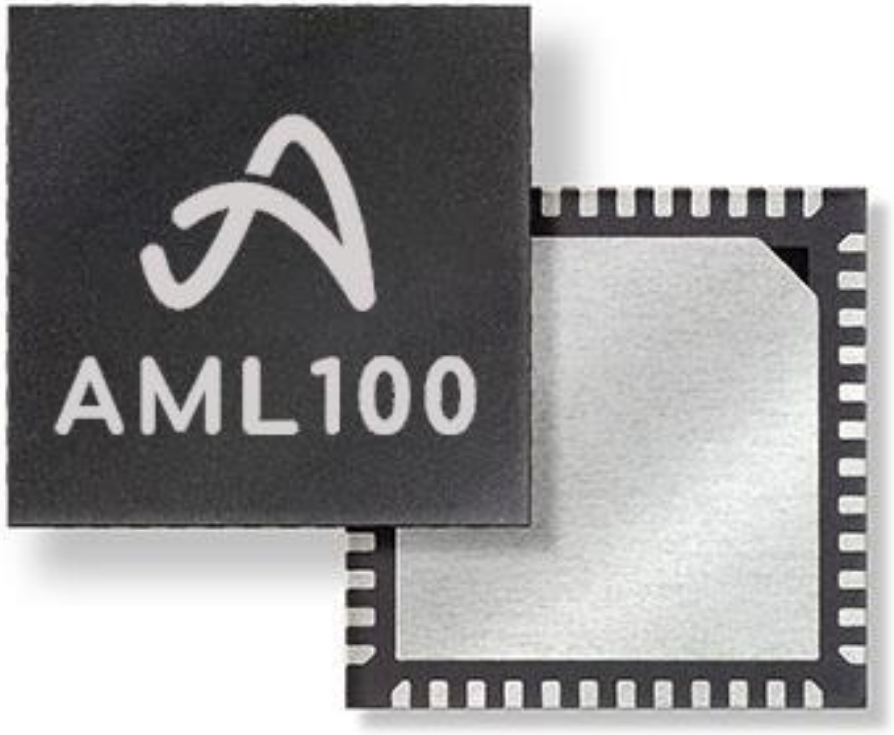




A Standard Computing Architecture

OTHER PLAYERS: ASPINITY

- https://www.aspinity.com/
- CMOS 10-bit NVM
- Analog circuits co-located with memory



PROGRAMMABLE RESISTORS (MIT)

- https://www.science.org/doi/10.1126/science.abp8064
- "The core idea behind analog training accelerators is to process information locally using physical device properties instead of conventional Boolean arithmetic—i.e., using Ohm's and Kirchhoff's laws for the matrix inner product and threshold-based updating for the outer product."

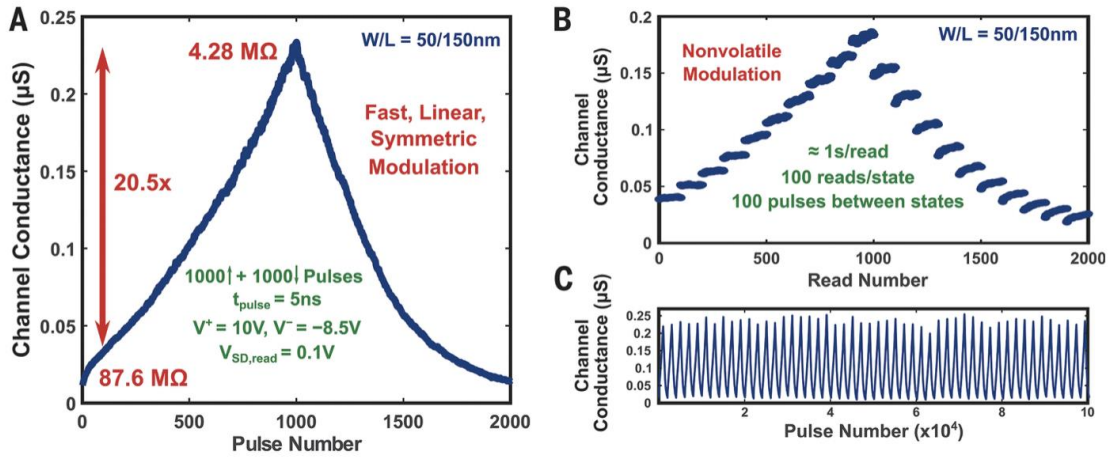


Fig. 2. Ultrafast and energy-efficient modulation characteristics of protonic programmable resistors.
(**A**) Modulation performance of a 50-nm–by–150-nm protonic device with 10-nm PSG, showing fast (5 ns per pulse), nearly linear, and symmetric characteristics. W, width; L, length. (**B**) Retention behavior of the protonic device for ≈100 s at different conductance levels over the full dynamic range. (**C**) Endurance characterization of the protonic device, displaying nondegrading modulation over 10^5 pulses conducted over 30 hours.

NEUROMORPHIC COMPUTING (INTEL)

- https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html
- “Spiking neural networks (SNNs), novel models that simulate natural learning by dynamically re-mapping neural networks, are used in neuromorphic computing to make decisions in response to learned patterns over time. Neuromorphic processors leverage these asynchronous, event-based SNNs to achieve orders of magnitude gains in power and performance over conventional architectures. ”