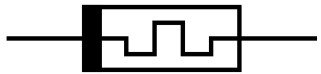
Memristors

Mohit L., Shreyas R.



The problem: Carbon footprint of computing

- “Training a single AI model can emit as much carbon as five cars in their lifetimes” (Hao, 2020)
- **AI, and computing in general, has a glaring hardware problem**.
- This is because of the traditional von Neumann architecture.
- Although this architecture has assisted us in achieving several strides in computing, it is simply **not efficient** in the way human brains are. (“von Neumann bottleneck”)

Fig. 1: The DeepMind Challenge Match between top Go player Lee Sedol and Google AlphaGo, famously won by the computer program.

Fig. 1: The DeepMind Challenge Match between top Go player Lee Sedol and Google AlphaGo, famously won by the computer program.

Despite AlphaGo being a major achievement, there is a key concern regarding power consumption in these kinds of *neural* computers.

- AlphaGo:
  - **170kW!**
  - 1200 CPUs
  - 176 GPUs
  - **Takes up an entire room!**
  - 30×10^{12} op/s
- Lee Sedol's Brain
  - **20W**
  - 1.5kg
  - **1250cm^3**
  - 10^{11} neurons
  - 10^{15} synapses

The solution: Compute-in-memory

- Compute-in-memory or in-memory processing makes **data operations directly available on the data memory**, as opposed to the traditional von Neumann architecture, where there is **significant power bottleneck** due to the **limited transfer rate** of data along the shared bus between the CPU and memory.
- The operating energy costs of traditional architecture, notwithstanding the cost of the complex systems themselves, make it increasingly **unfeasible and unaffordable** for the rapid pace of development in computing in recent years.

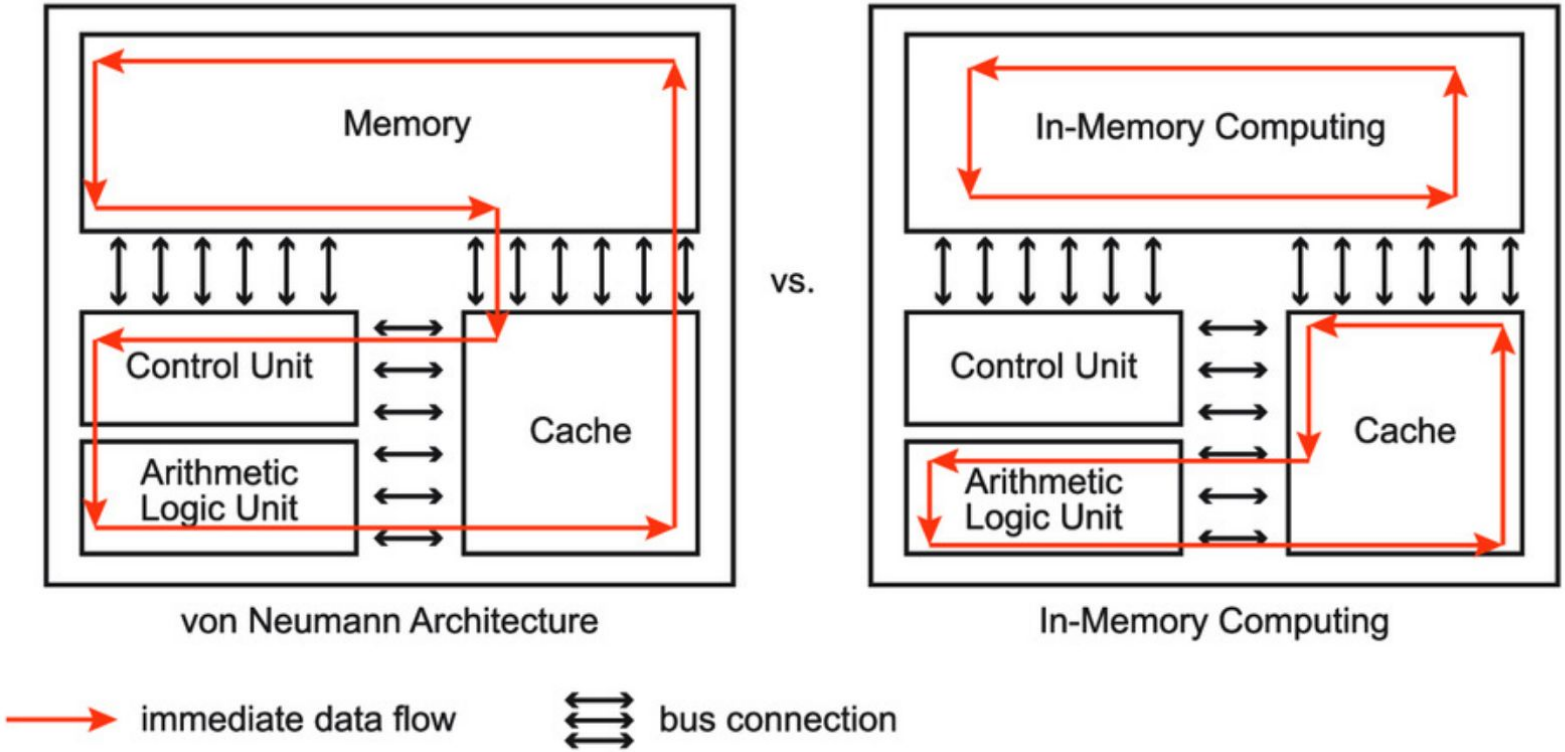


Fig. 2: Schemes of von Neumann architectures versus Compute-in-memory

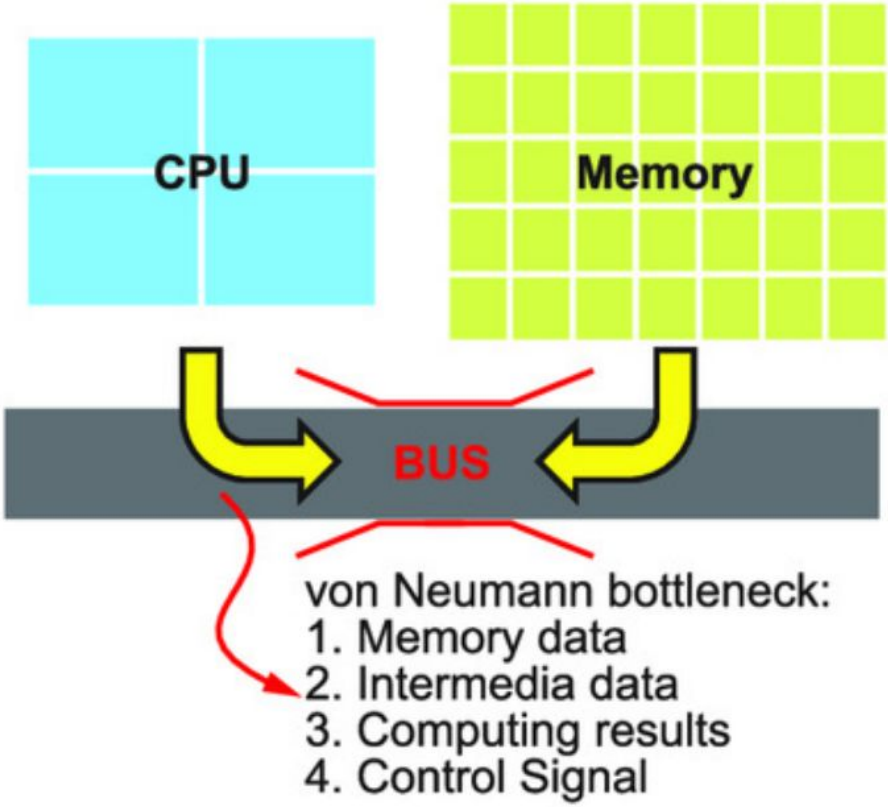


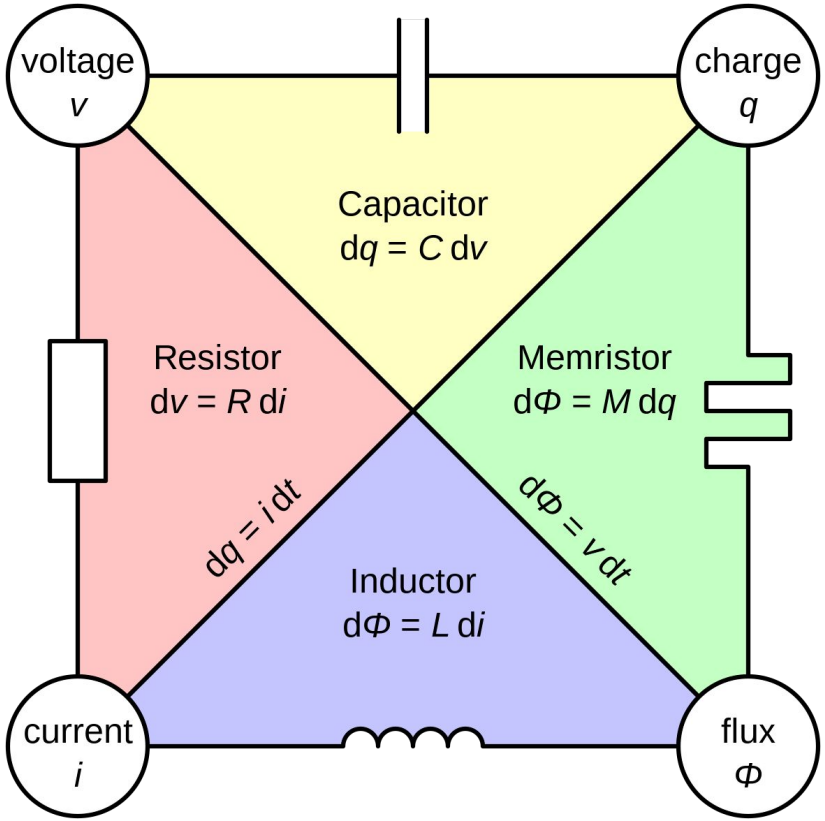
Fig. 3: Diagram of von Neumann bottleneck

An implementation: Resistive Random Access Memory

- Resistive Random Access Memory (RRAM) is an emerging technology that shows a lot of promise as a commercializable compute-in-memory solution due to its compatibility with available silicon CMOS architectures.
- Recent developments in RRAM technology keep up with the increasing computational demand for complex AI functionalities.
- By storing AI model weights in dense, analogue and non-volatile RRAM devices, and by performing AI computation directly within RRAM, **we can eliminate the bottleneck**.

The foundation: Memristors

- RRAM is based on memristors, or memory resistors, which are a unique type of electrical component which have the ability to **remember** or program the current that had previously flown through them.
- They are nonlinear two-terminal electrical components which relate electric charge q to magnetic flux linkage ϕ, as shown in **Fig. 4** (right).

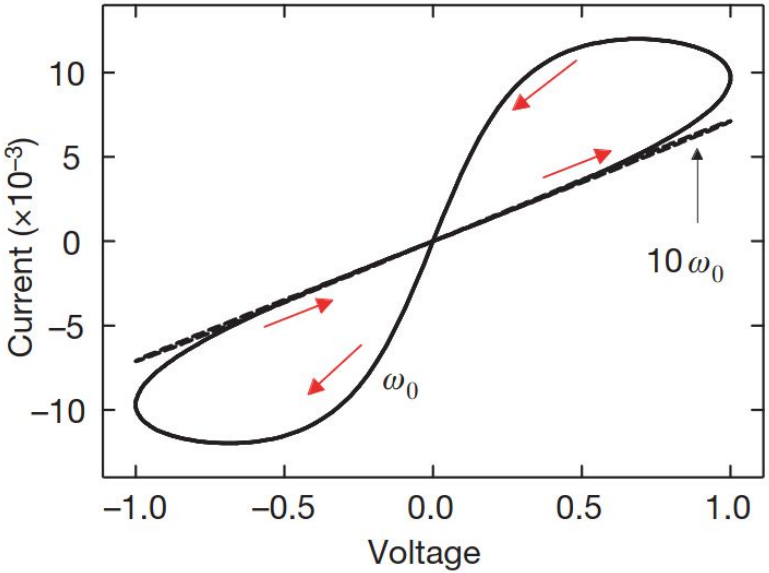


The theory: Memristors

- We can characterize the memristor by its memristance function describing the charge-dependent rate of change of magnetic flux linkage, or

$$M(q(t)) = \frac{d\phi/dt}{dq/dt} = \frac{V(t)}{I(t)}$$

- Or, $V(t) = M(q(t))I(t)$. If $M(q(t))$ is constant, then we obtain a linear relationship between current and voltage as Ohm's law. Else, it displays pinched hysteresis characteristics, hallmark of memristors, as shown in **Fig. 5** (right).



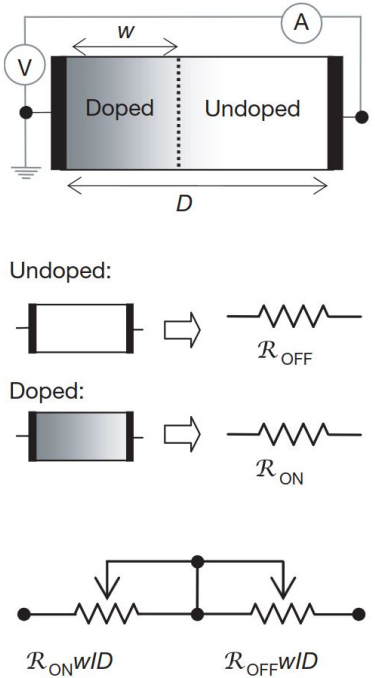
The slope of the pinched hysteresis curves represent the electrical resistance of the memristor.
The change in slope of the curves demonstrates switching between different resistance states
which is a phenomenon central to RRAM and other forms of two-terminal resistance memory. At
high frequencies, memristive theory predicts the pinched hysteresis effect will degenerate,
resulting in a straight line representative of a linear resistor

The working: Memristors

- **Fig. 6** (right): Diagram of doped metal oxide thin film and simple equivalent circuits. Electric current through the memristor shifts the doped oxygen vacancies in the metal oxide (generally TiO_2) thin film, causing a gradual and persisting change in electrical resistance.
- The memristance function can be mathematically modeled for TiO_2 thin films as

$$M(q(t)) = R_{\mathrm{OFF}} \left(1 - \frac{\mu R_{\mathrm{ON}}}{D^2} q(t)\right)$$

where R_{OFF} represents the RESET / high resistance state, R_{ON} represents the SET / low resistance state, μ represents the mobility of dopants in the film and D represents the thickness of the film.



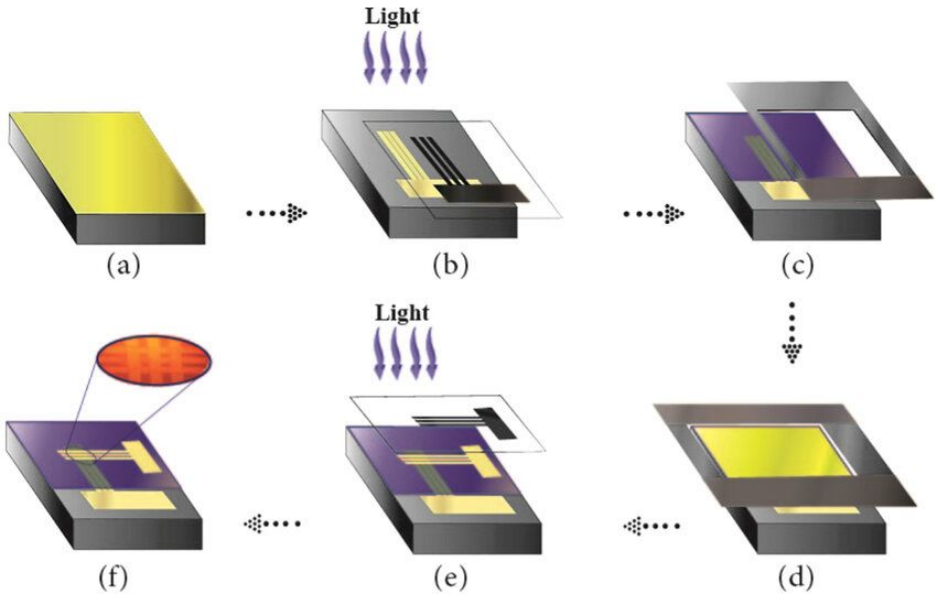
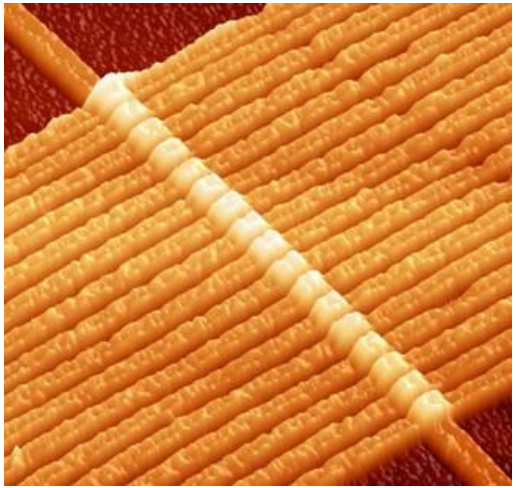


Fig. 7: Schematic of the fabrication process of the Au | TiO_2 | Au memristor device. (a) Deposition of gold thin film on the SiO_2 | Si (100) substrate, (b) pattering the gold thin film as BE by photolithography approach, (c) TiO_2 thin film deposition as active layer using a shadow mask, (d) deposition of gold thin film by second shadow mask, (e) photolithography patterning of the gold thin film as TE, (f) the final fabricated device.



“Since our brains are made of memristors, the flood gate is now open for commercialization of computers that would compute like human brains, which is totally different from the von Neumann architecture underpinning all digital computers” - Leon Chua