# Computer Memory History and Development Trend


Amir Mahdi Hosseini Monazzah

Room 332,

School of Computer Engineering,

Iran University of Science and Technology,

Tehran, Iran.

*monazzah@iust.ac.ir*

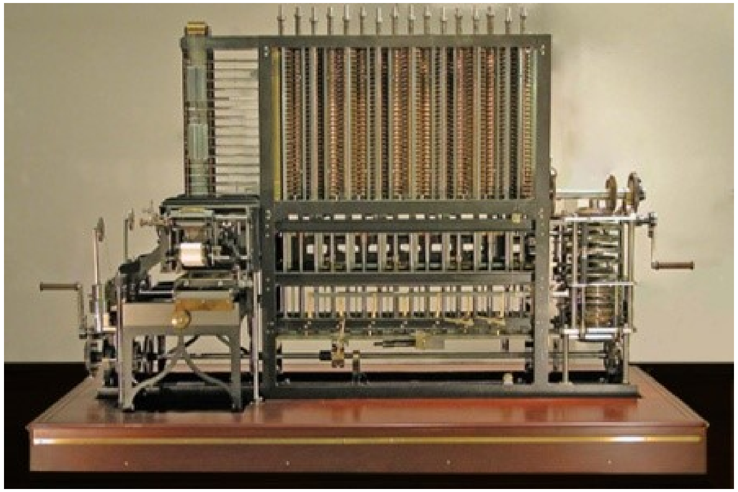

*Invited Talk at Shahid Rajaei University (1399/12/06)*

# Outline

- The emergence of memory
- Memory unit
  - Memory wall
  - Memory types2
  - Memory hierarchies
- Memory technologies
  - Volatile
  - Non-volatile
  - Comparison
- Conclusion

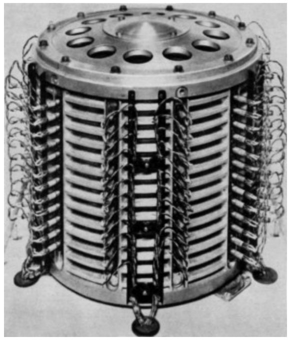
# How memories become matured during the time?

- 1800s
  - Charles Babbage constructs the first punched card machine with a memory store called the Analytical Engine (1833)

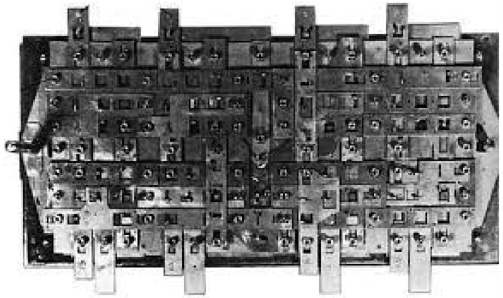


# How memories become matured during the time?

- 1930s
  - Gusta Tauschek builds drum memory (1932)

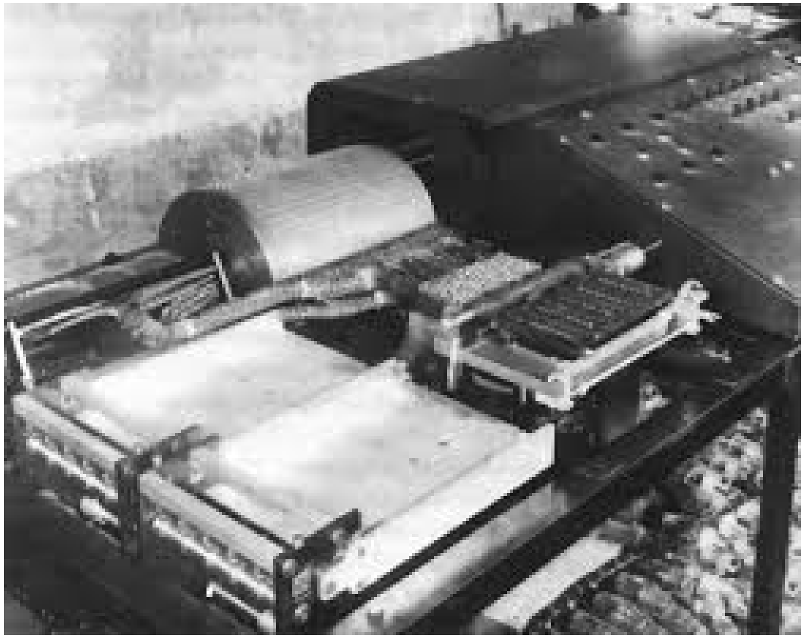


  - Konrad Zuse patents a mechanical combination memory (1936)



# How memories become matured during the time?

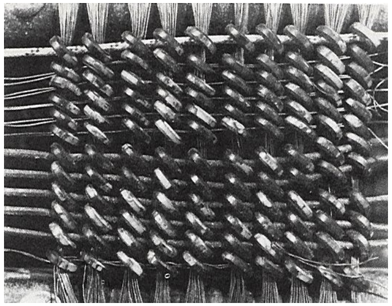
- 1930s
  - John Atanasoff and Clifford Berry build electronic 50-bit words digital memory (1939)



# How memories become matured during the time?

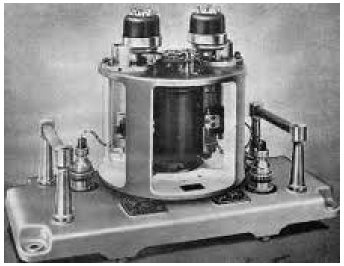
- 1940s
  - Helmut Schreyer develops neon lamp memory (1941)

  - The magnetic core memory with pulse transfer control was invented by Frederick Viehe, An Wang, and Kenneth Olsen independently (1947)



# How memories become matured during the time?

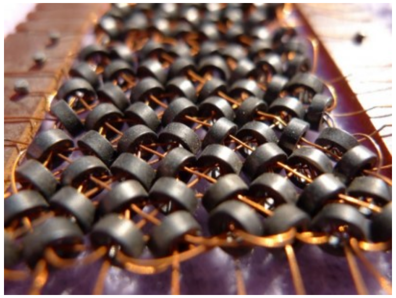
- 1940s
  - Jay Wright Forrester invents the magnetic random-access coincident-current drum



  - Maurice Vincent Wilkes builds the first practical stored-program computer with vacuum tube memory (1949)

# How memories become matured during the time?

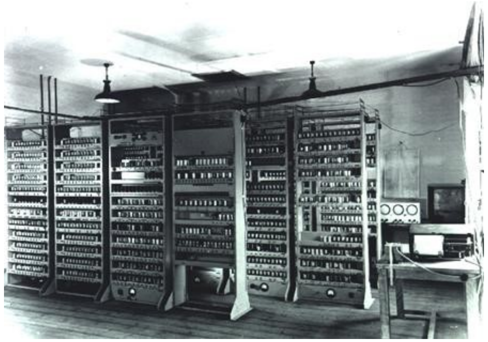
- 1950s
  - Jay Forrester patents the matrix core memory (1951)
    - Ram was born!



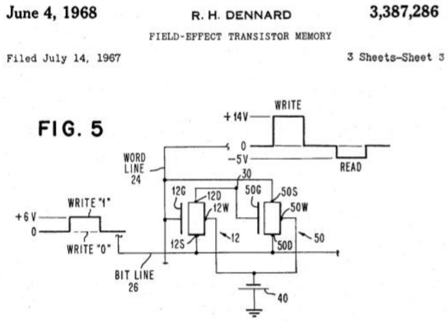
- 1960s
  - Atlas virtual memory is invented (1962)

# How memories become matured during the time?

- 1960s
  - Maurice Wilkes develops the idea of a cache computer memory (1965)

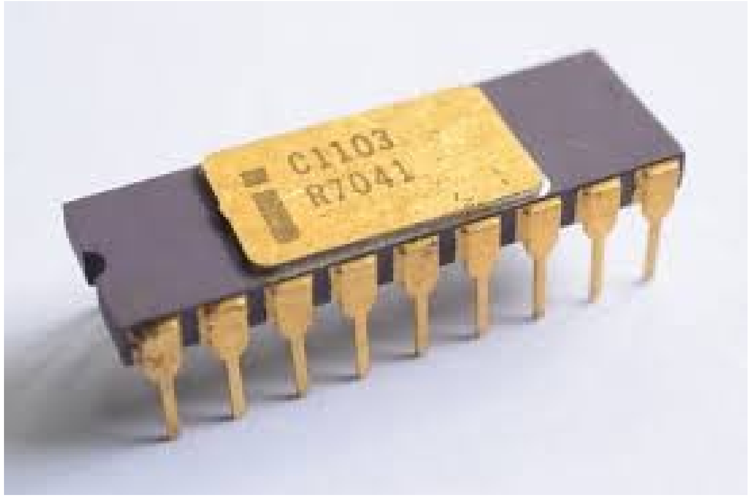


  - Robert Dennard invents and patents his Dynamic Random Access Memory, DRAM (1968)



# How memories become matured during the time?

- 1960s
  - Intel creates a 1 Kb RAM chip (1969)

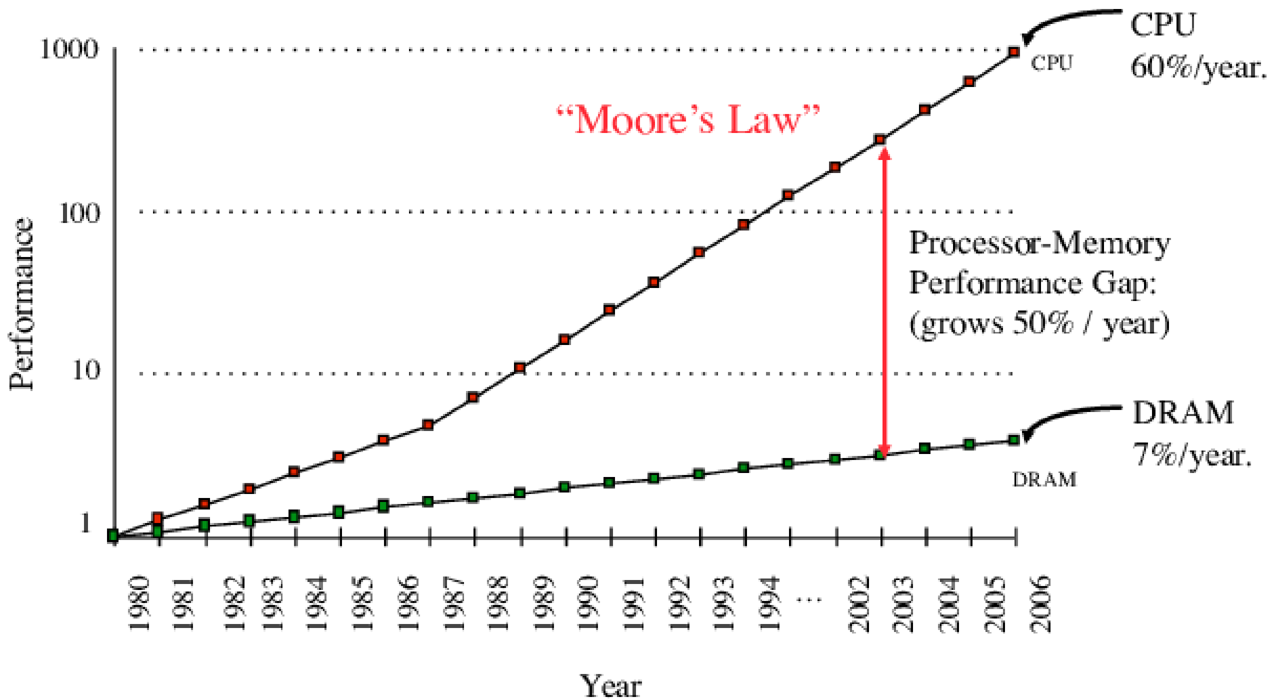



- And so on …

# Memories form computer architecture perspective

- Memory
  - An essential component in any general purpose computer since it is needed to store programs and data
- Main memory
  - Memory unit that communicates directly with the CPU
- Auxiliary memory
  - Devices that provide backup storage
  - Are used to store system programs, large data files and other backup information. Only programs and data currently needed by the processor reside in main memory

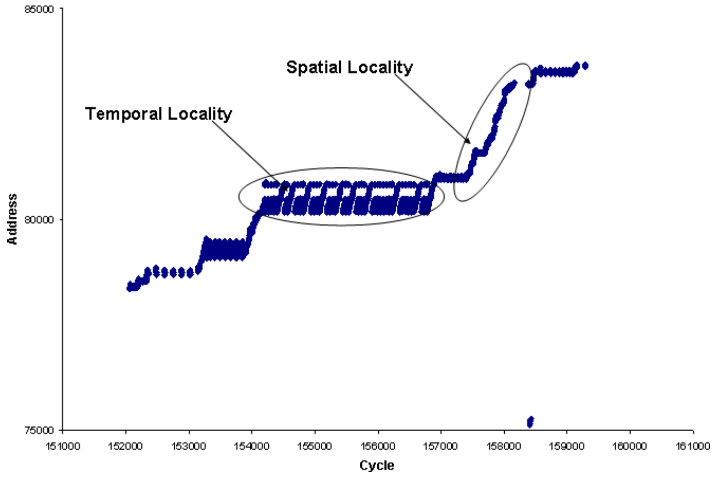
# A performance challenge!




M. Bahi and C. Eisenbeis, "High Performance by Exploiting Information Locality through Reverse Computing," *2011 23rd International Symposium on Computer Architecture and High Performance Computing*, Vitoria, Brazil, 2011, pp. 25-32, doi: 10.1109/SBAC-PAD.2011.10.

# Cache memory

- Memory that lies in between main memory and CPU
- Holds those parts of the program and data that are most heavily used
  - Increases the overall processing speed of the computer by providing frequently required data to the CPU at a faster speed



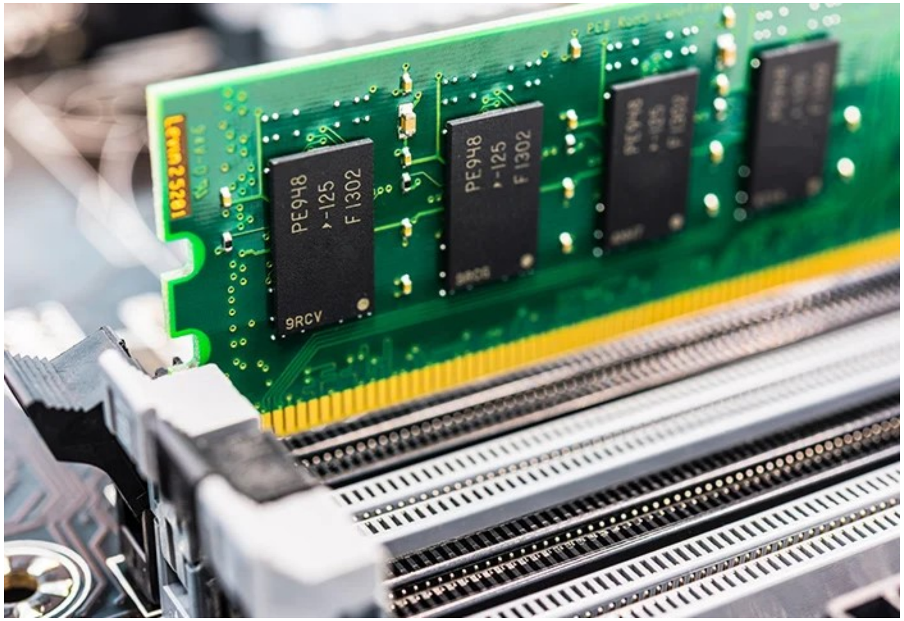

Obtained from Stackoverflow webpage

# From the access type perspective

- Sequential access memory
  - A class of data storage device that read their data in sequence
  - Are usually a form of magnetic memory
  - Typically used for secondary storage in general-purpose computers due to their higher density, resistance to wear and non-volatility
  - Example: hard disk, CD-ROMs, magnetic tapes, etc.

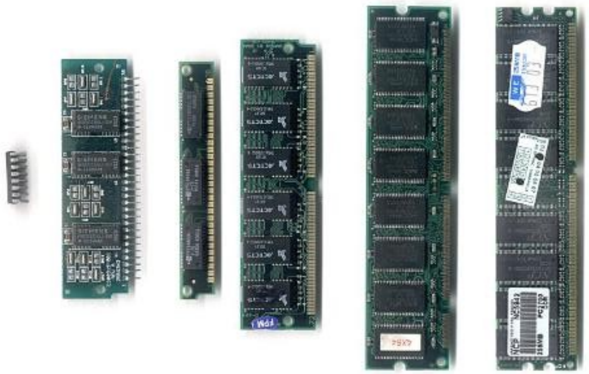
# From the access type perspective

- Random access memory
  - Is a form of computer data storage
  - Allows stored data to be accessed in any order
  - Type: SRAM and DRAM



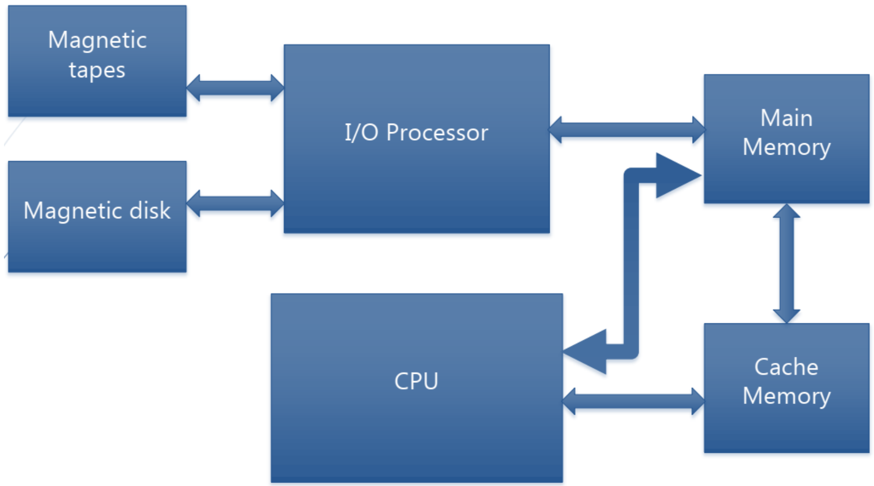
# From the storage permanence perspective

- Volatile memories
  - Require constant power to maintain the stored information
  - Holds data temporary
    - Eg.
- Non-volatile memories
  - Can store information even when there is no constant power
  - Holds data permanently

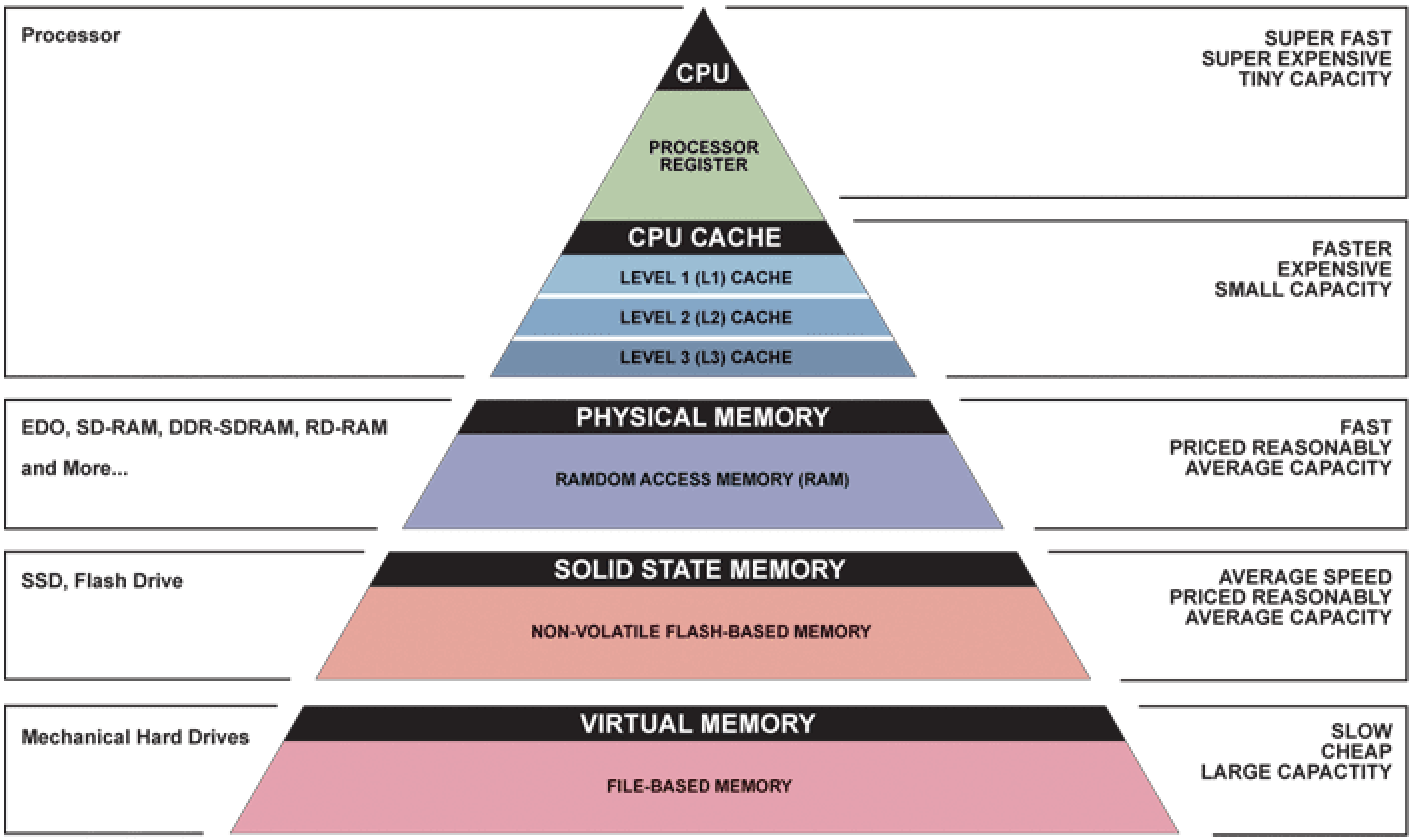


# An efficient and practical approach!

- To obtain the highest possible access speed while minimizing the total cost of the memory system
- Consists of all storage device in a computer system
  - Auxiliary
  - Cache
  - Main memory
  - High speed registers
  - And etc.

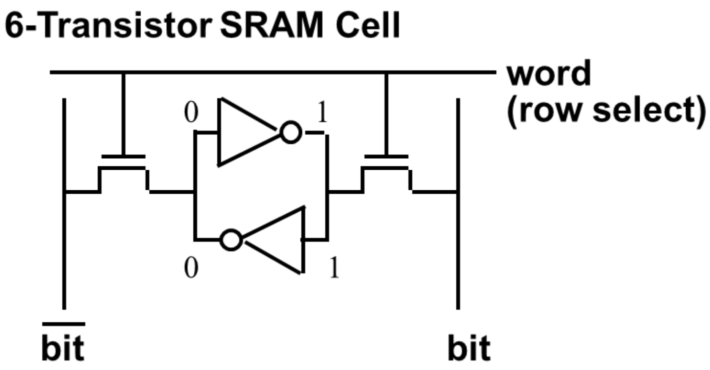


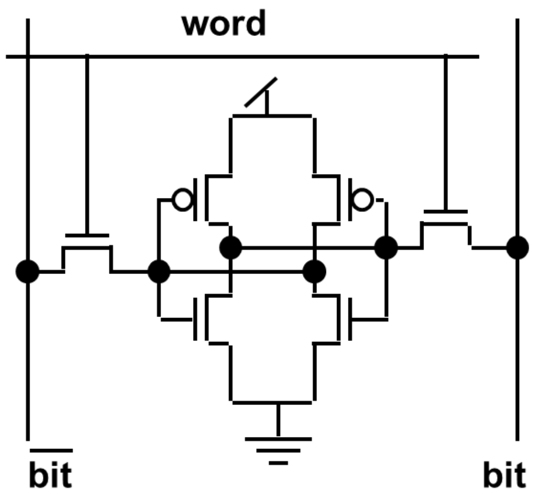
# An efficient and practical approach!




▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

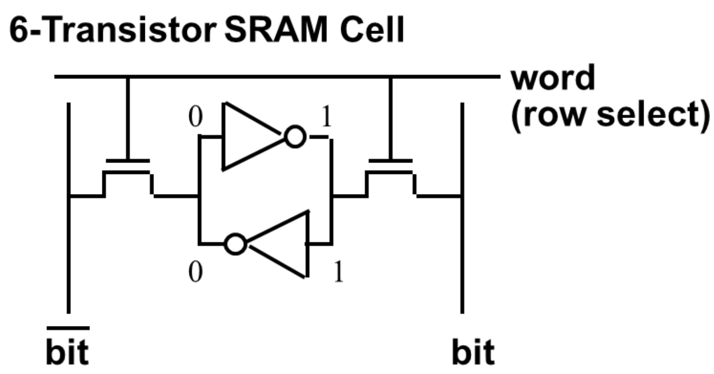
# Static RAM Cell: SRAM

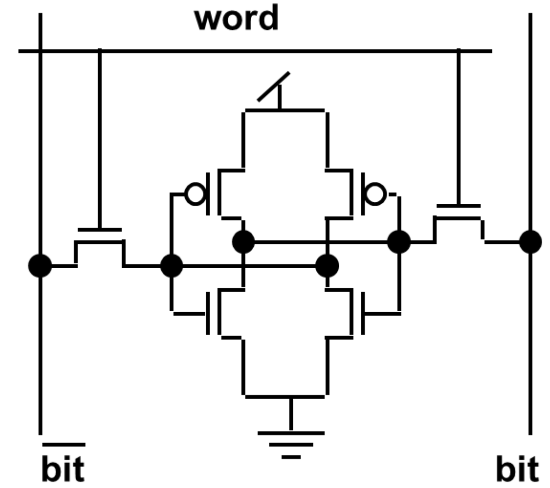



- Write
  - Drive bit lines (bit=1, bit=0)
  - Select row

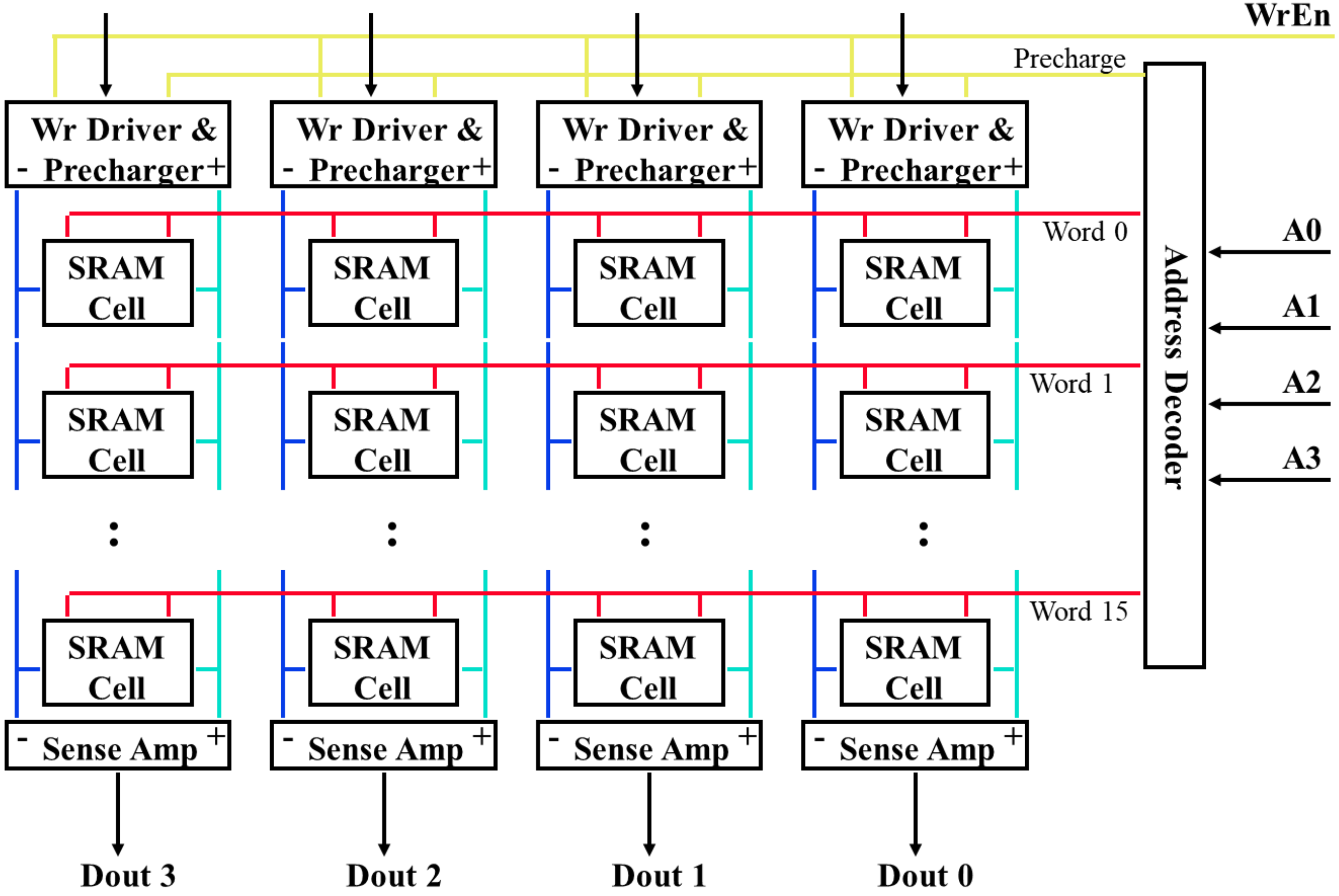
# Static RAM Cell: SRAM



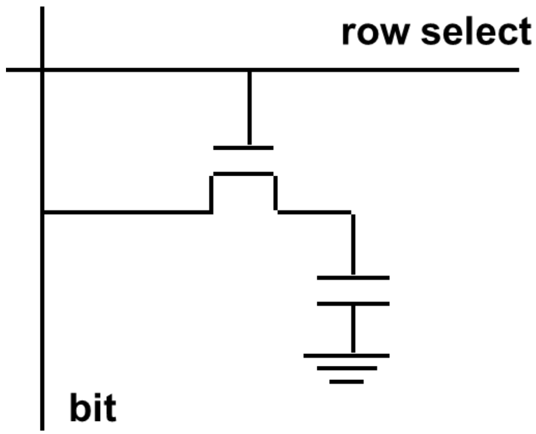




- Read
  - Precharge $\overline{\text{bit}}$ and bit to Vdd or Vdd/2 => make sure equal!
  - Select row
  - Cell pulls one line low
  - Sense amp on column detects difference between bit and bit

# Typical SRAM organization: 16-word x 4-bit

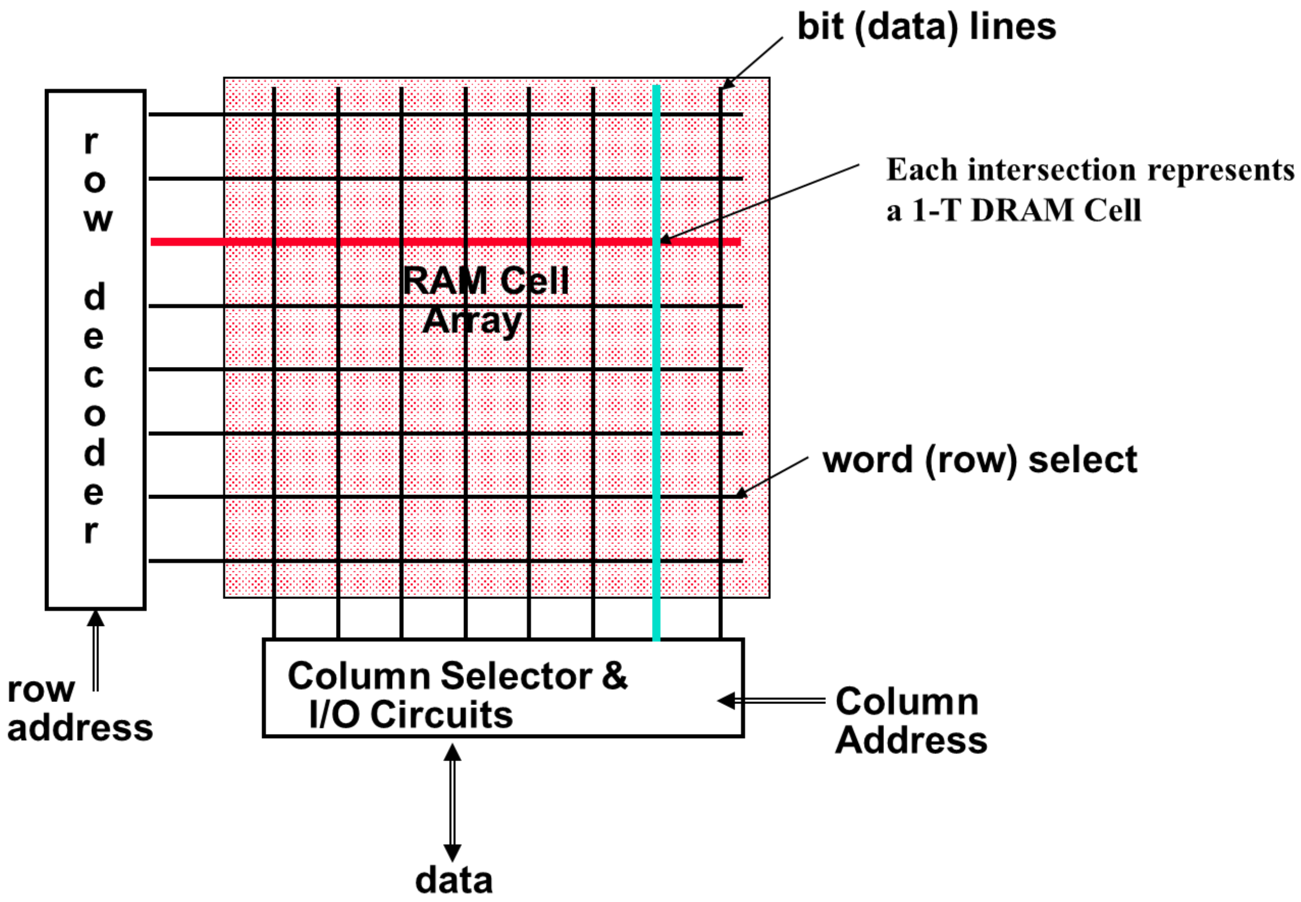


# 1-Transistor Memory Cell (DRAM)




- Write
  - Drive bit line
  - Select row
- Refresh
  - Just do a dummy read to every cell
- Read
  - Precharge bit line to Vdd
  - Select row
  - Cell and bit line share charges
  - Very small voltage changes on the bit line
  - Sense (fancy sense amp)
    - Can detect changes of ~1 million electrons
  - Write: restore the value

# Classical DRAM organization (square)



# Why non-volatile memories (NVMs)?

- Computer architecture perspective
  - Limitations of SRAM and DRAM
  - Power wall
  - Cost wall
  - Scaling wall
- Auxiliary storage perspective
  - Limitations of HDDs
  - Limitations of Flash-based SSDs

# Emerging NVM technologies

- Ferroelectric RAM (FRAM)
- Magnetic RAM (MRAM)
- Phase Change Memory (PCM)
- Spin Torque Transfer RAM (STT-RAM)
- Resistive Random Access Memory (ReRAM/RRAM)
- …

# NVM technologies major parameters

- Cell size ⇨ Unit Cost
- Read power
- Write power (set power / reset power)
- Leakage current (static power)
- Scalability
- Read latency
- Write latency (set vs. reset write latency)
- Read voltage level (and read current)
- Write voltage level (and write current)

# NVM technologies major parameters

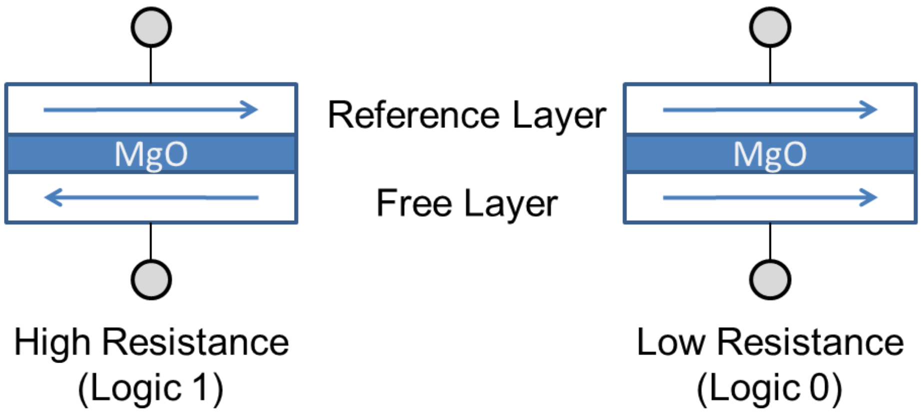
- In-place update
  - If not, erase latency?
- Yield
- Maturity
- Soft FIT rate
  - For various fault models (SEU, temperature, read disturbance, program disturbs)
- Hard FIT rate
- Endurance
- Mask count & fabrication complexity ⇨ NRE cost

# NVM technologies major parameters

- Data retention
  - Less than 1ms ⇨ Volatile
  - Greater than 1 year ⇨ Non-Volatile
  - Between?
- Bits per cell (MLC-2, MLC-3, …) ⇨ Unit Cost
- 3-D stacking capability
- Compatibility with CMOS technology
- NVM cell peripherals
  - For regular and non-regular structures
- Process variation

# Spin Transfer Torque Magnetic RAM (STT-MRAM)

- Cell




- Magnetic Tunnel Junction (MTJ)
- Relative magnetization direction
- Different resistances ⇨ Logic 0 or 1
- Write: spin-polarized current
  - Much less write current than conventional MRAM

# Some of the well-known memory technologies

| | SRAM | DRAM | Flash (NOR) | Flash (NAND) | FeRAM | MRAM | PRAM | RRAM | STT-MRAM |
|---|---|---|---|---|---|---|---|---|---|
| Non-volatile | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Cell Size [$F^2$] | 50-120 | 6-10 | 10 | 5 | 15-34 | 16-40 | 6-12 | 6-10 | 6-20 |
| Read Time [ns] | 1-100 | 30 | 10 | 50 | 20-80 | 3-20 | 20-50 | 10-50 | 2-20 |
| Write/Erase Time [ns] | 1-100 | 15 | 1μs/1ms | 1ms/0.1ms | 50/50 | 3-20 | 50/120 | 10-50 | 2-20 |
| Endurance | $10^{16}$ | $10^{16}$ | $10^{5}$ | $10^{5}$ | $10^{12}$ | $>10^{15}$ | $10^{8}$ | $10^{8}$ | $>10^{15}$ |
| Write Power | Low | Low | Very High | Very High | Low | High | Low | Low | Low |
| Other Power Consumption | Leakage | Refresh | None | None | None | None | None | None | None |
| High Voltage Required | No | 3V | 6-8V | 16-20V | 2-3V | 3V | 1.5-3V | 1.5-3V | <1.5V |
| | *Existing Products* | | | | | | *Prototypes* | | |

# Conclusion and future research trend

- The trend of memory design and technologies were explored
- Memory hierarchy is reviewed
- Two traditional memory technologies were introduced
- The main metrics for comparison and evaluation of NVMs are introduced
- As a case study, we introduced the most promising NVM technology (STT-MRAM) which could be a substitute for SRAM

# Conclusion and future research trend

- Many design challenges have been raised during the last decade that should be resolved
  - Performance
  - Power
  - Temperature
  - Capacity
    - SLC
    - MLC
  - …

# References


- David A. Patterson, “Memory system lectures”, UC Berkeley, Berkeley, California, 2000.
- Hossein Asadi, “Storage systems lectures”, Sharif University of Technology, Tehran, Iran, 2019.
- Sandesh Jonchhe et al., “Memory organization”, St. Lawrence College, Kingston, Canada, 2017.
- Ping Zhou et al., “Energy Reduction for STT-RAM Using Early Write Termination”, in ICCAD, 2019.

# The End