Digital Integrated Circuits
(83-313)

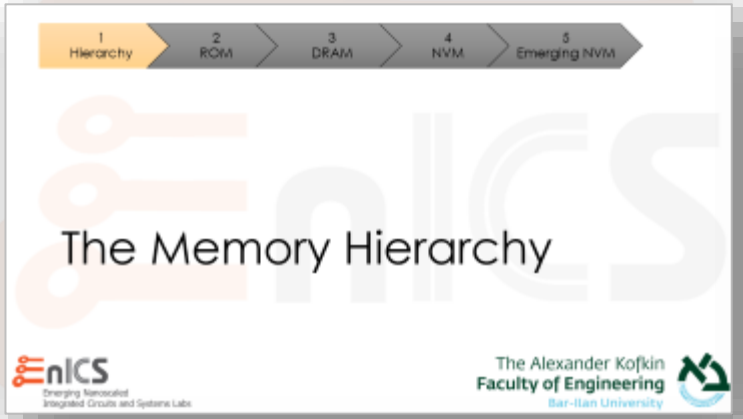
Lecture 11:
Other Memories

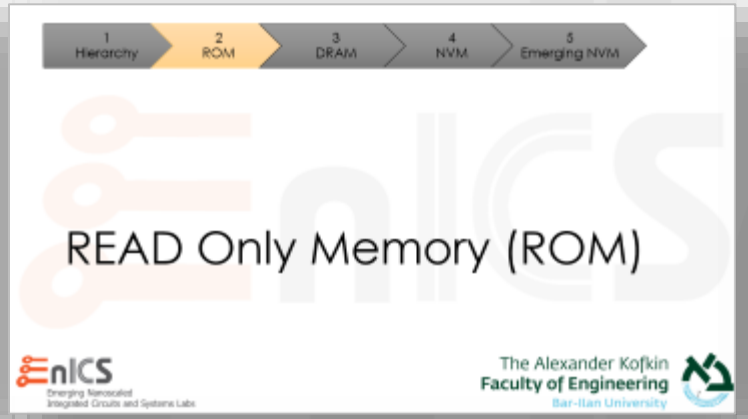
Prof. Adam Teman
27 May 2022

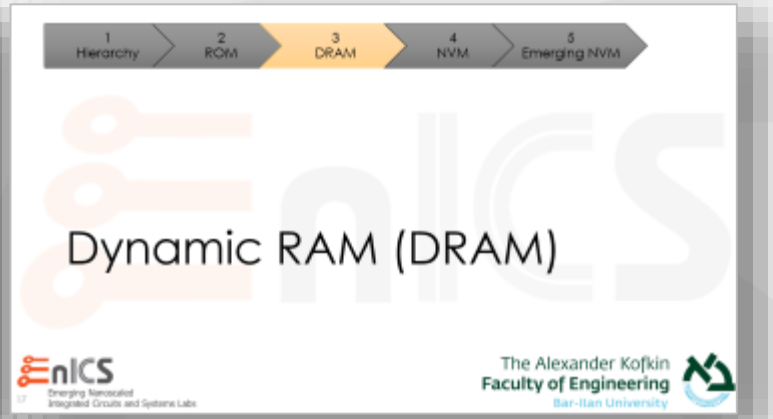
Emerging Nanoscaled
Integrated Circuits and Systems Labs

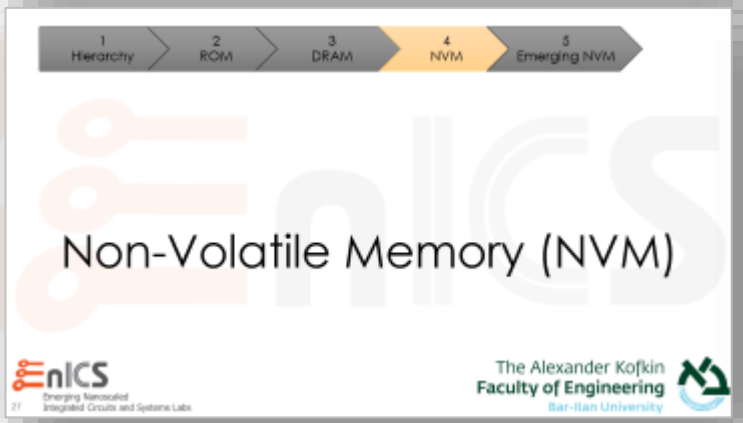
The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

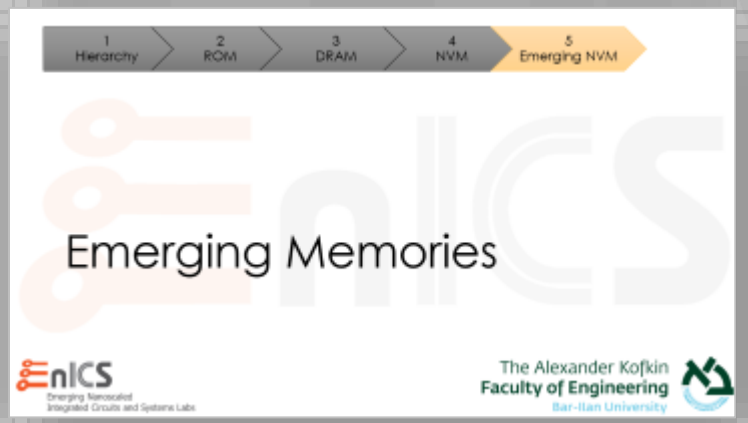
Lecture Content





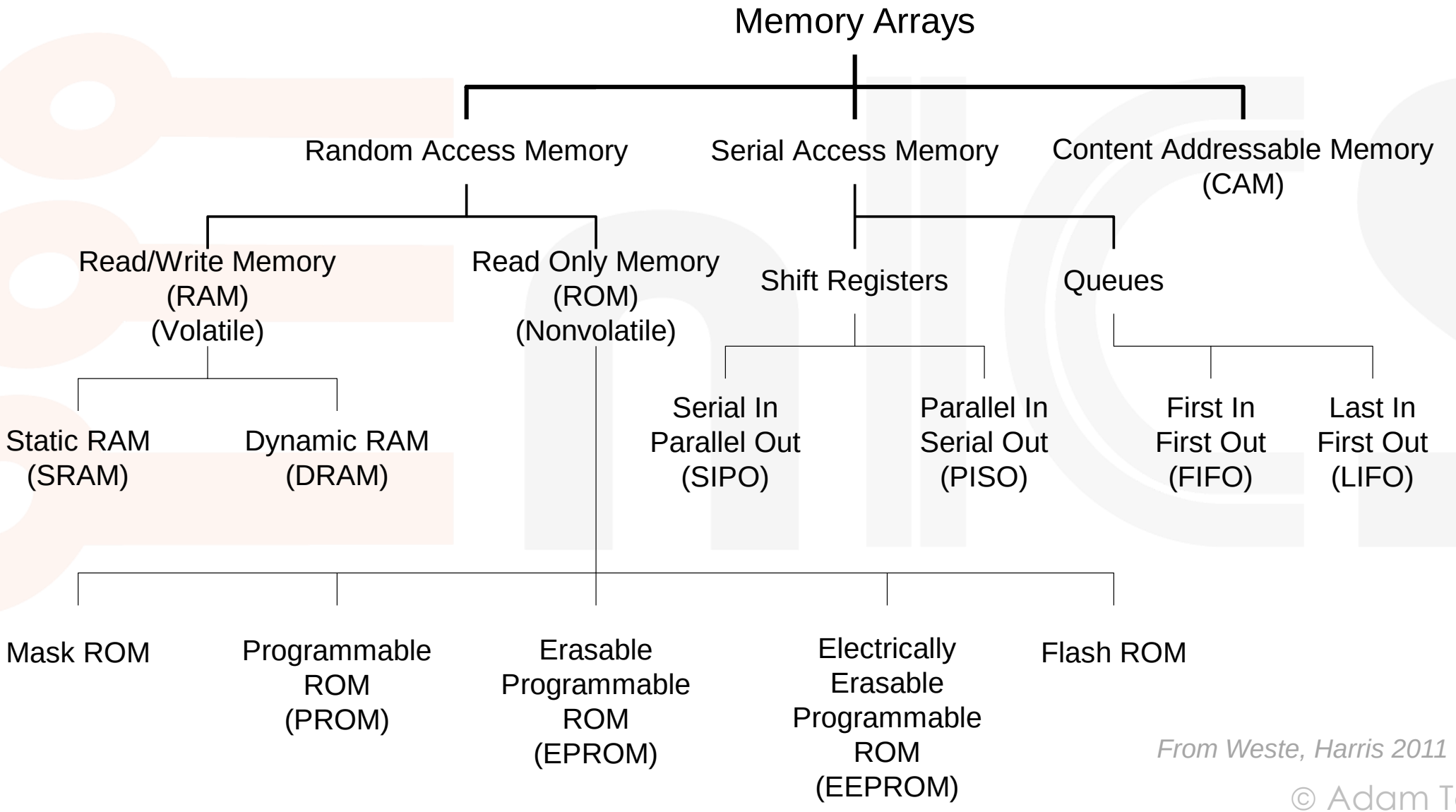






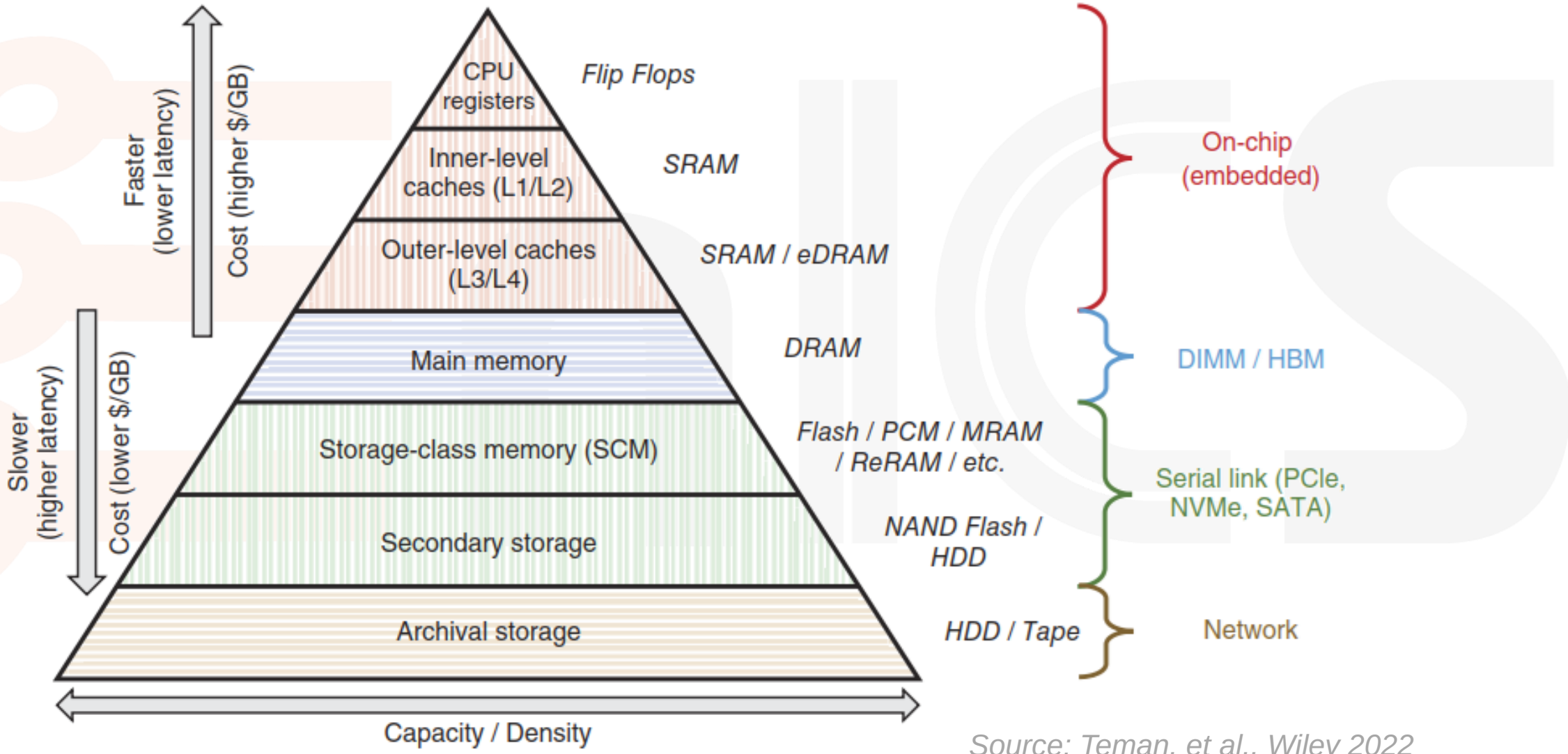
The Memory Hierarchy

Semiconductor Memory Classification



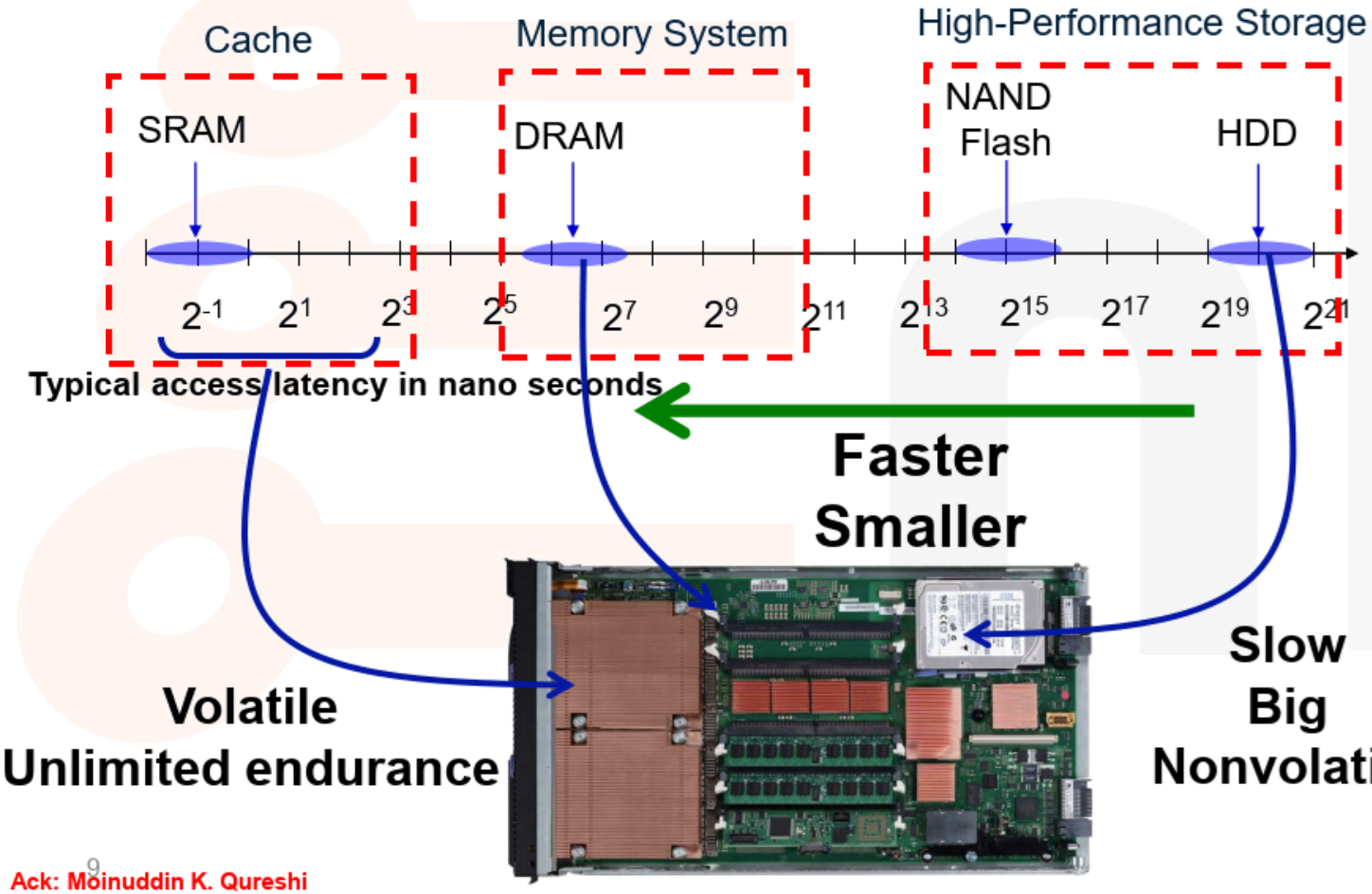
From Weste, Harris 2011

Memory Hierarchy of a Personal Computer

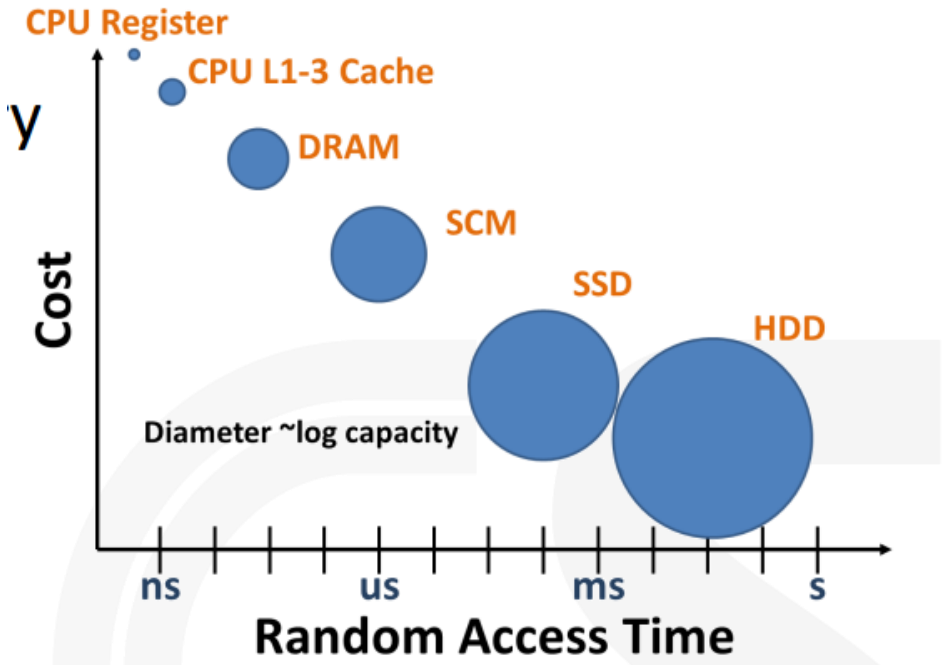


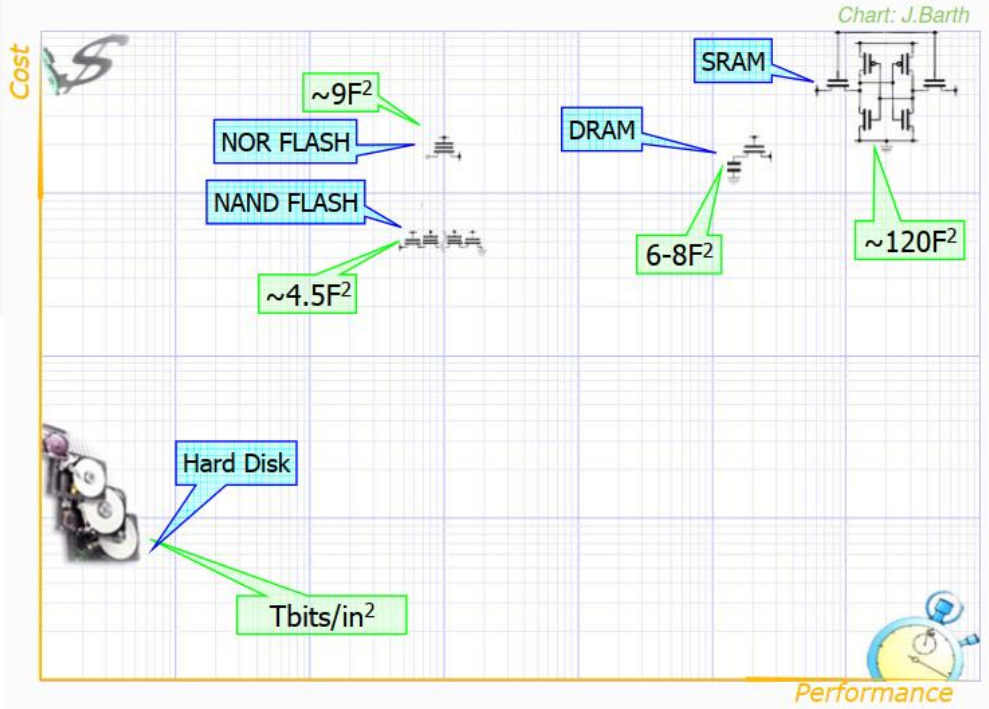
Source: Teman, et al., Wiley 2022

Another View



Ack: Moinuddin K. Qureshi

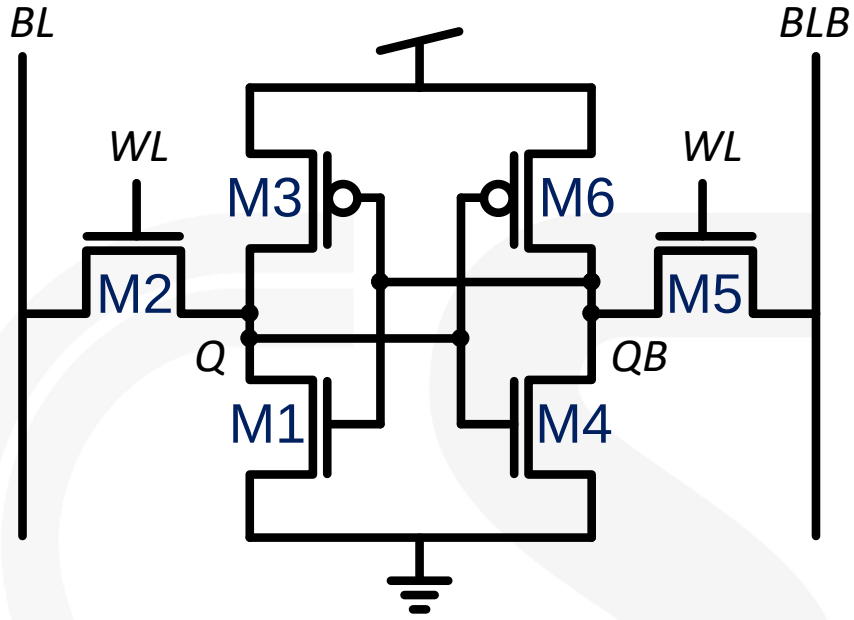


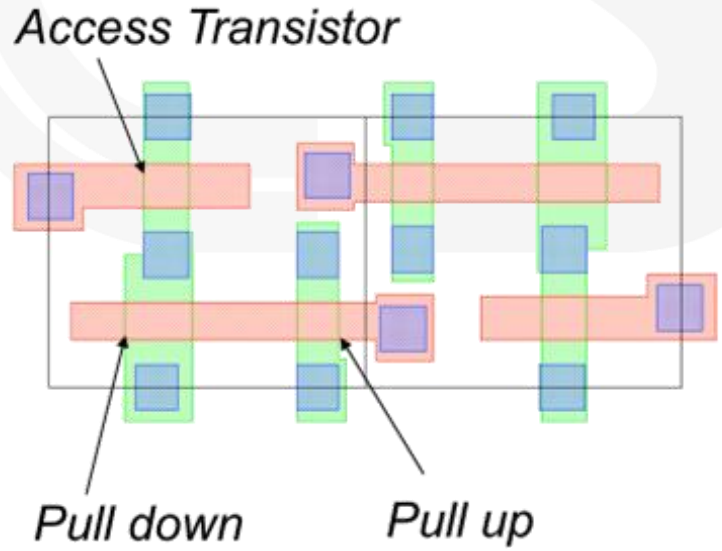


READ Only Memory (ROM)

SRAM Drawbacks

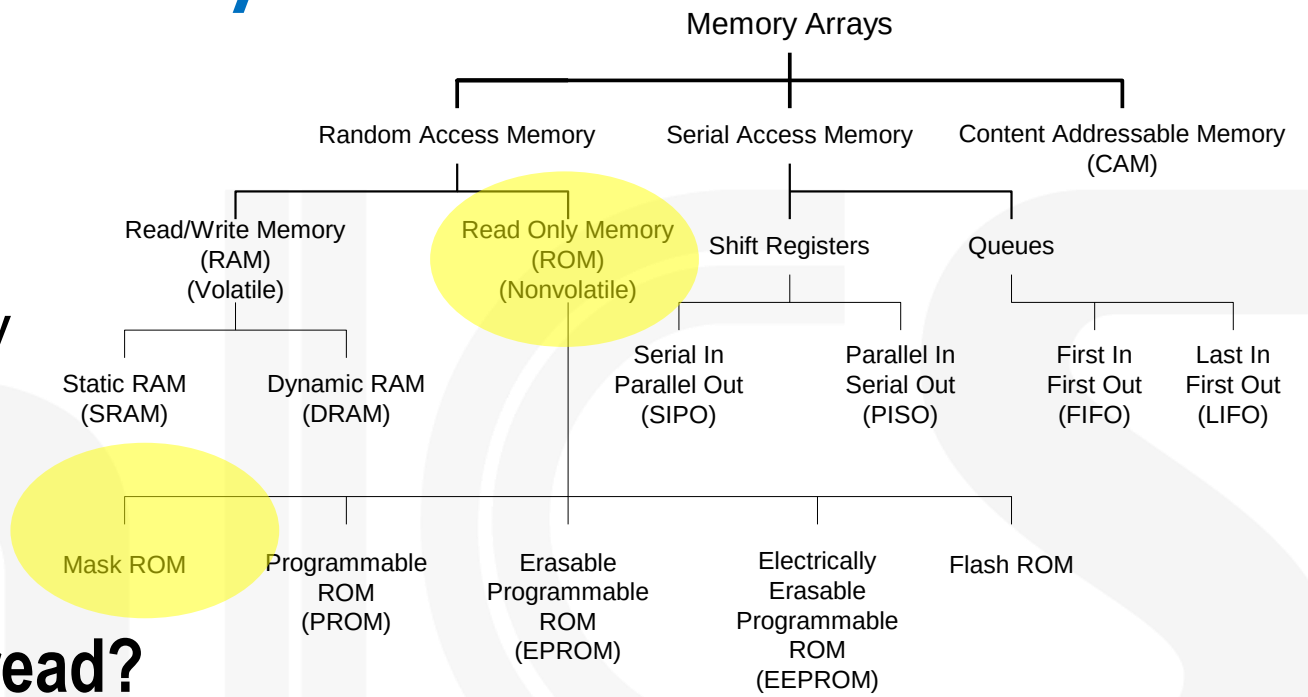
- **The 6T SRAM is by far the dominant on-chip (embedded) memory solution**
  - Static operation
  - High noise margins
  - Differential (→Fast) readout
  - Symmetric structure → dense layout
- **But six transistors is still a lot…**
  - On-chip SRAM is (usually) limited to a <30 MB
  - *Smallest* Linux distro requires >64 MB to run
  - Windows 10 requires >2 GB to run
- **Can we achieve higher density?**





Do all memories need R/W access?

- **We needed a lot of design to enable both read and write**
  - But some memories are Read Only

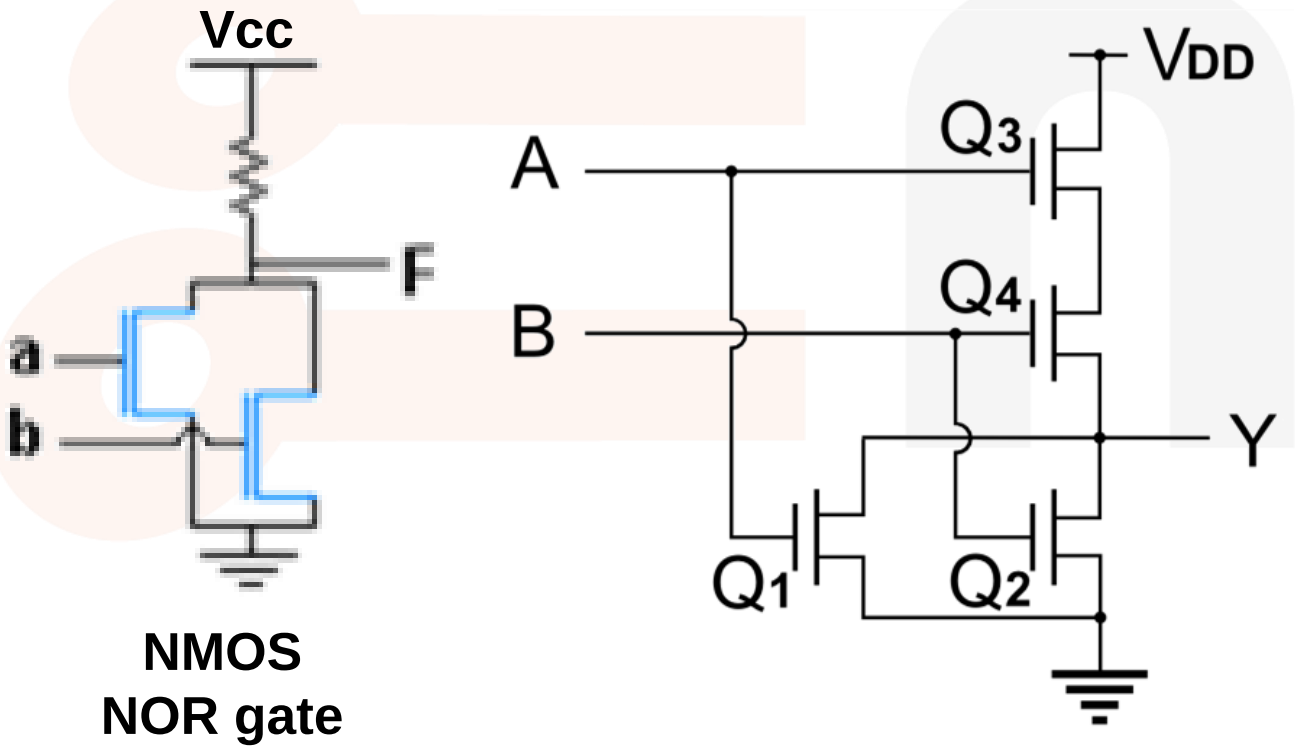


- **Can we manufacture a memory with predefined content that can only be read?**
  - One option is to define it in the RTL…
  - But what if we want it to be nicely structured and easily modified for bug fixes and future versions of our product?

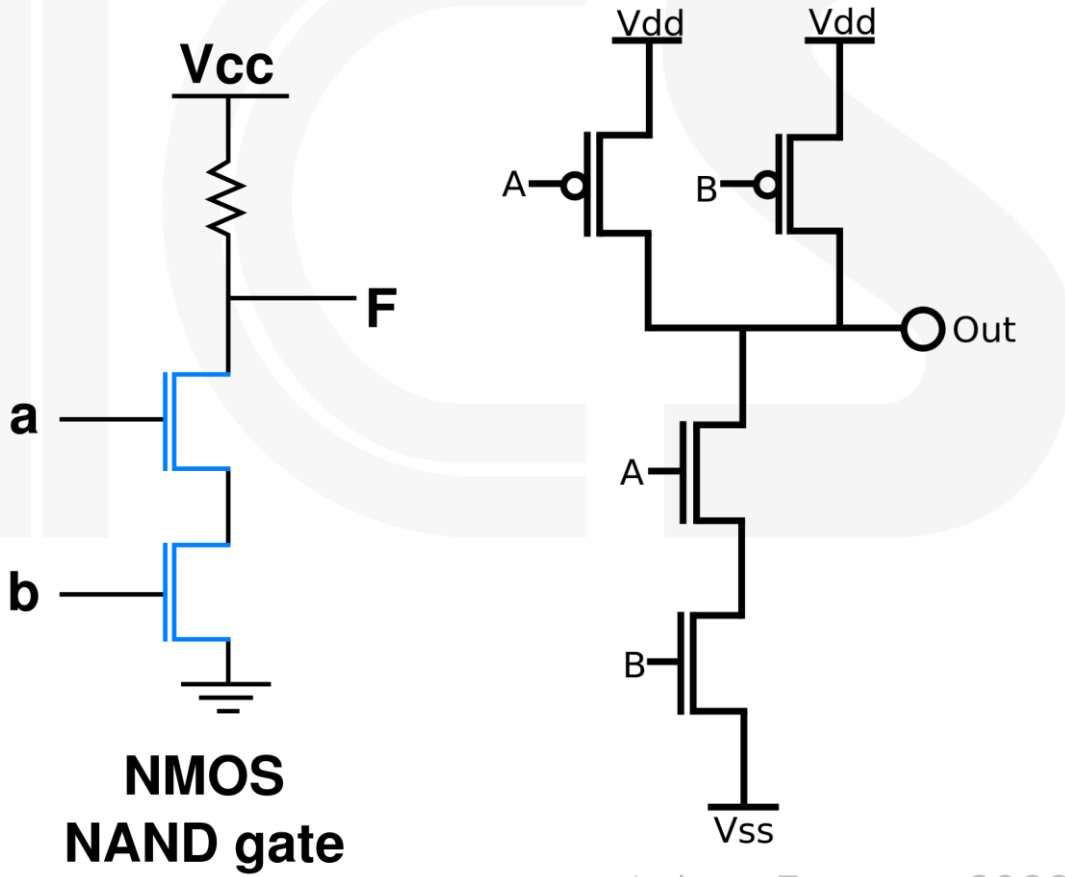
- **The solution: Mask ROM**

Reminder: NOR vs. NAND Pull-Down Networks

- In a **NOR** gate, we only need one conducting path to pull down the output
- In a **NAND** gate, we need all paths to be conducting to pull down the output



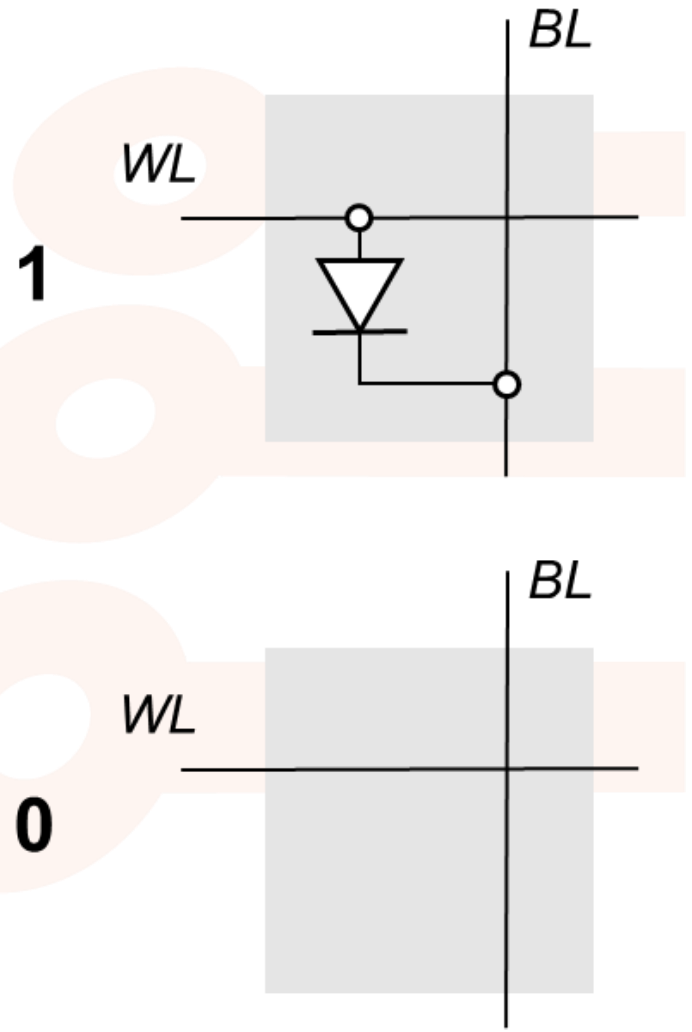
NMOS
NOR gate



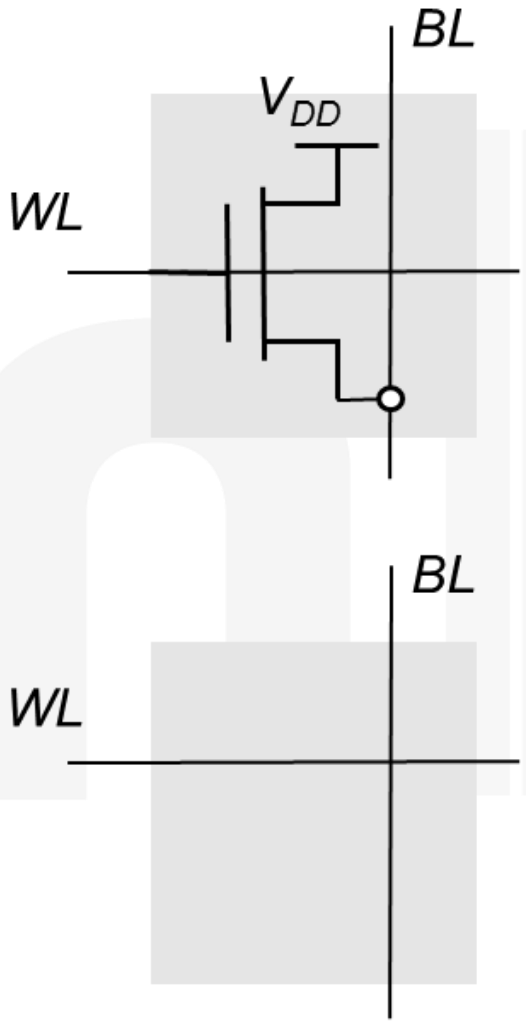
NMOS
NAND gate

Source: Wikipedia

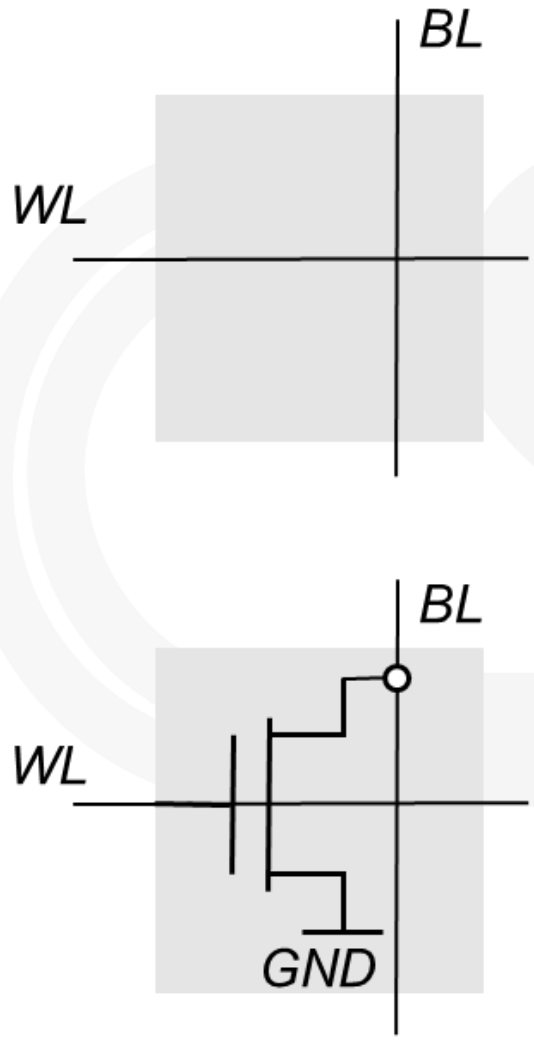
Read-Only Memory Cells



Diode ROM



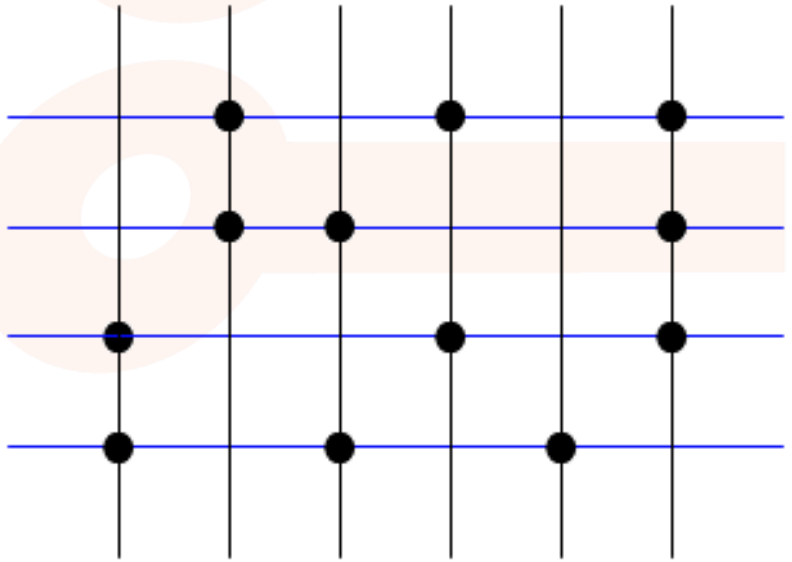
MOS ROM 1

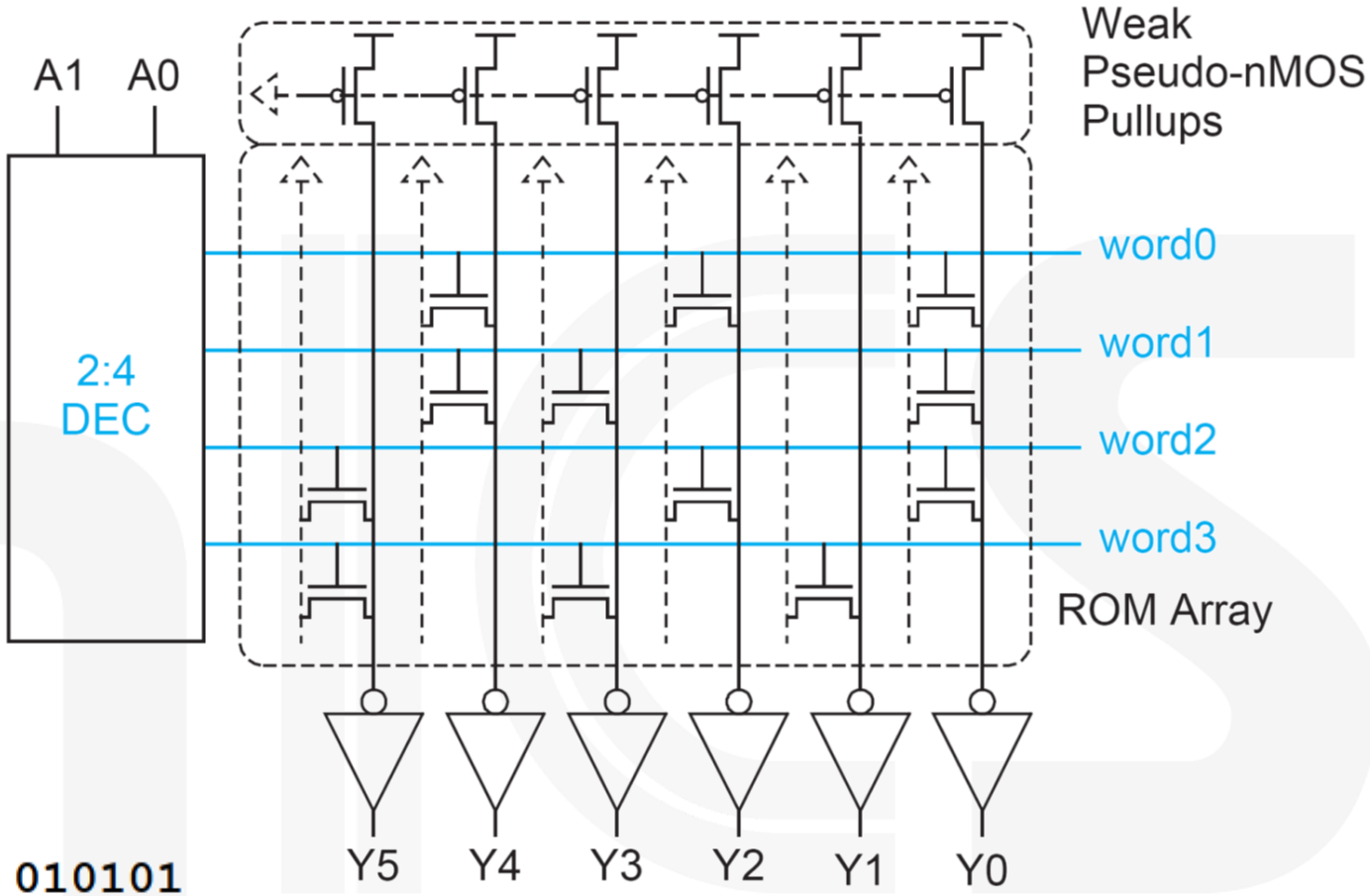


MOS ROM 2

NOR ROM

- **4-word x 6-bit NOR-ROM**
  - Selected word-line high
- **Represented with dot diagram**





From Weste, Harris 2011

Word 0: **010101**

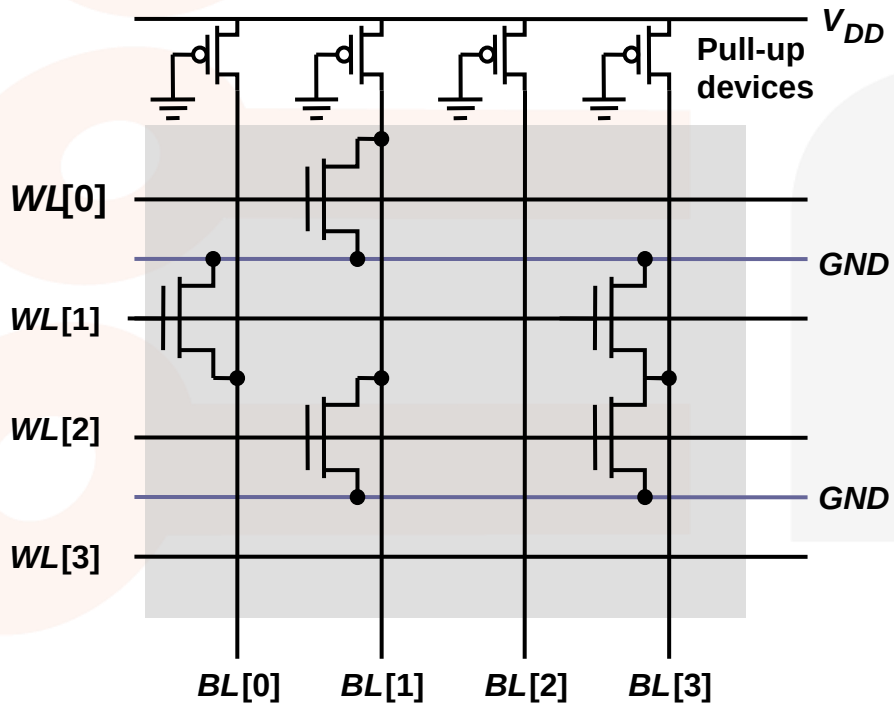
Word 1: **011001**

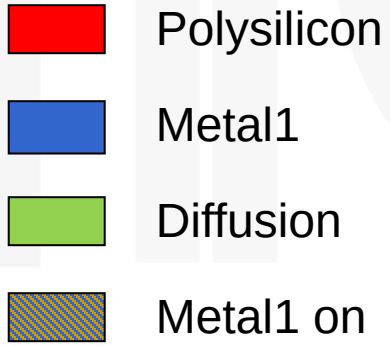
Word 2: **100101**

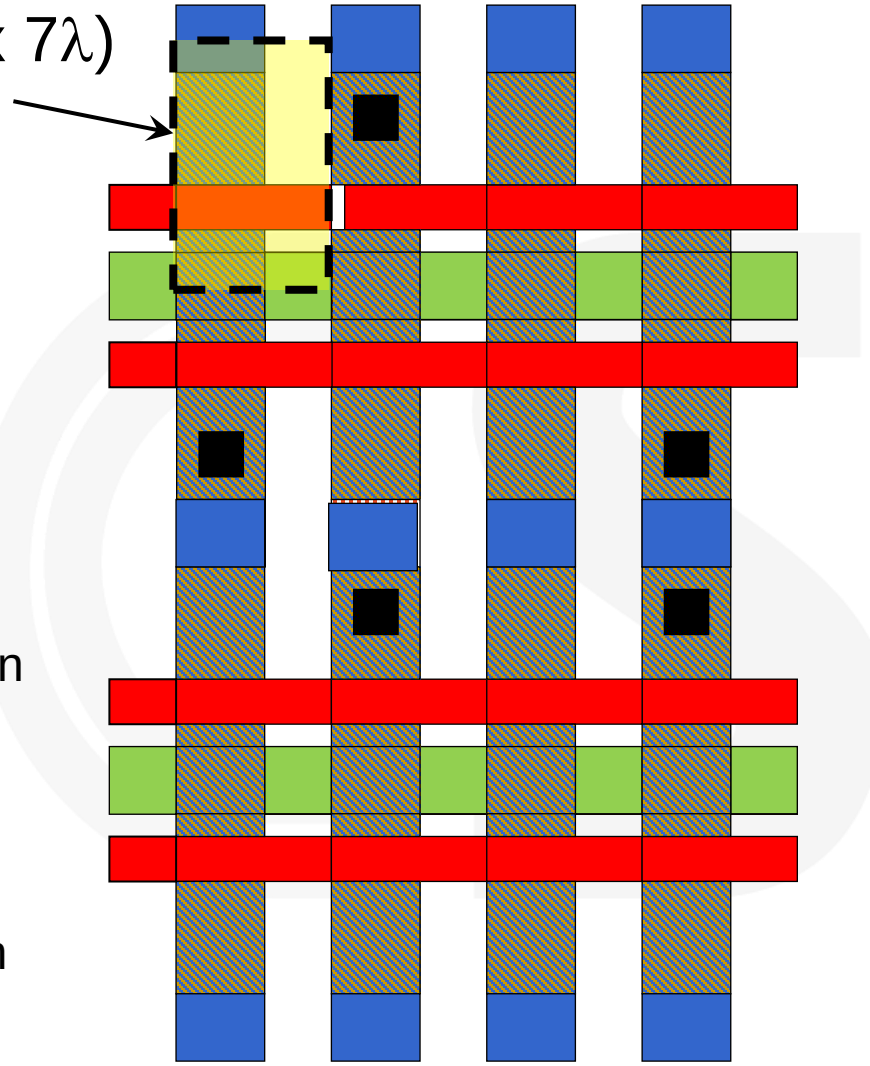
Word 3: **101010**

NOR ROM Layout

- **Programming using the Contact Layer Only**



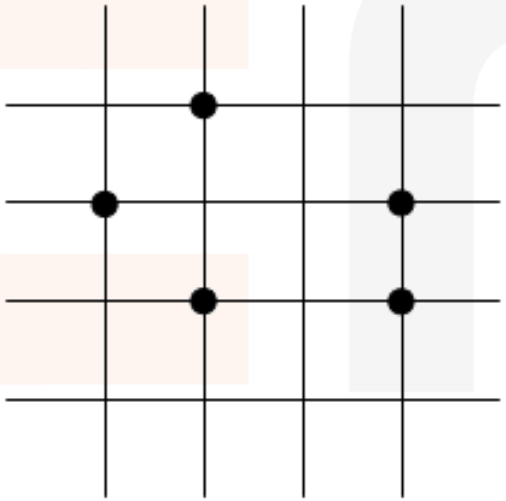


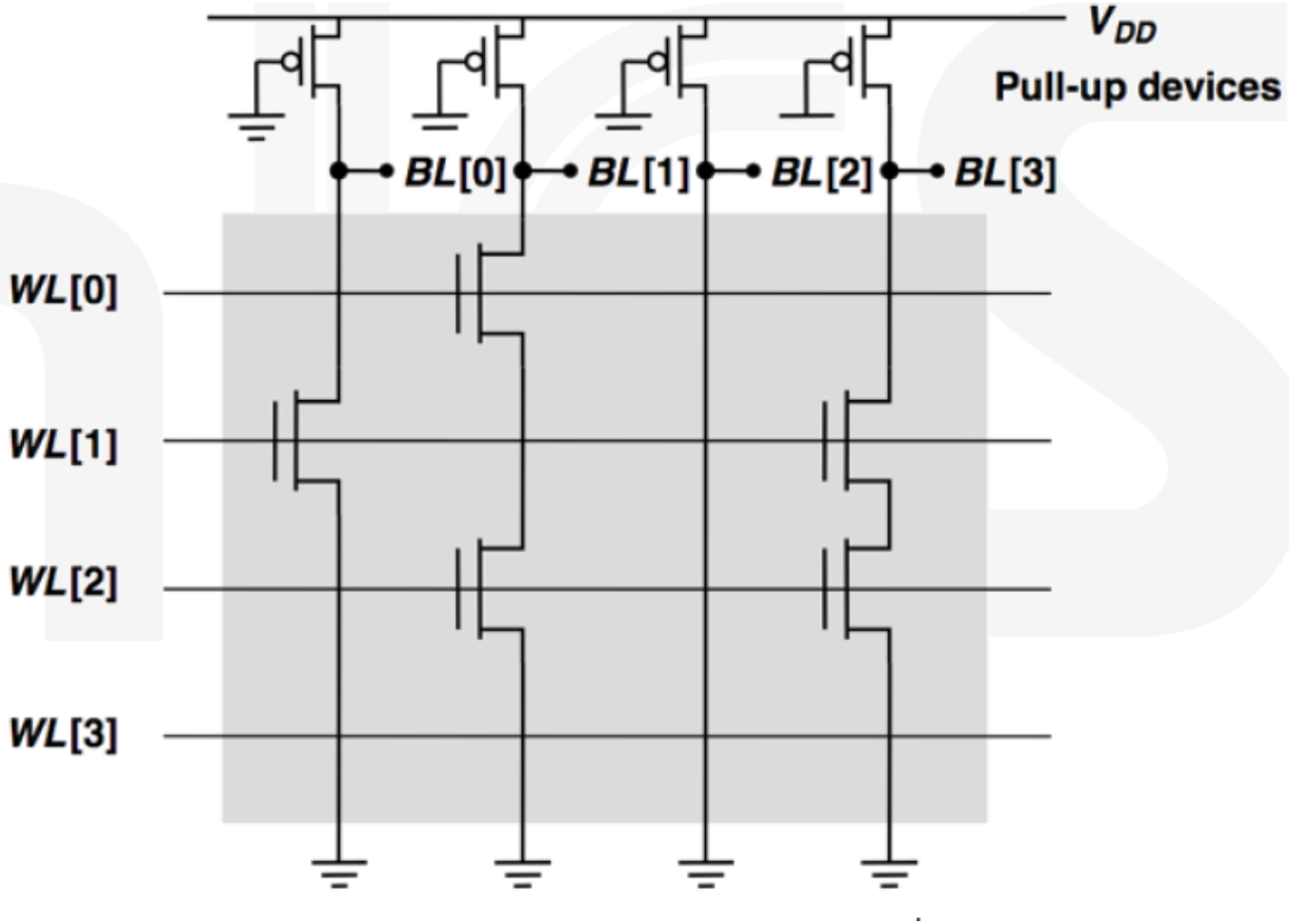


NAND ROM

- **4-word 4-bit NAND ROM**
  - All word-lines high by default with exception of selected row

WL[0]=0:**0100**
WL[1]=0:**1001**
WL[2]=0:**0101**
WL[3]=0:**0000**

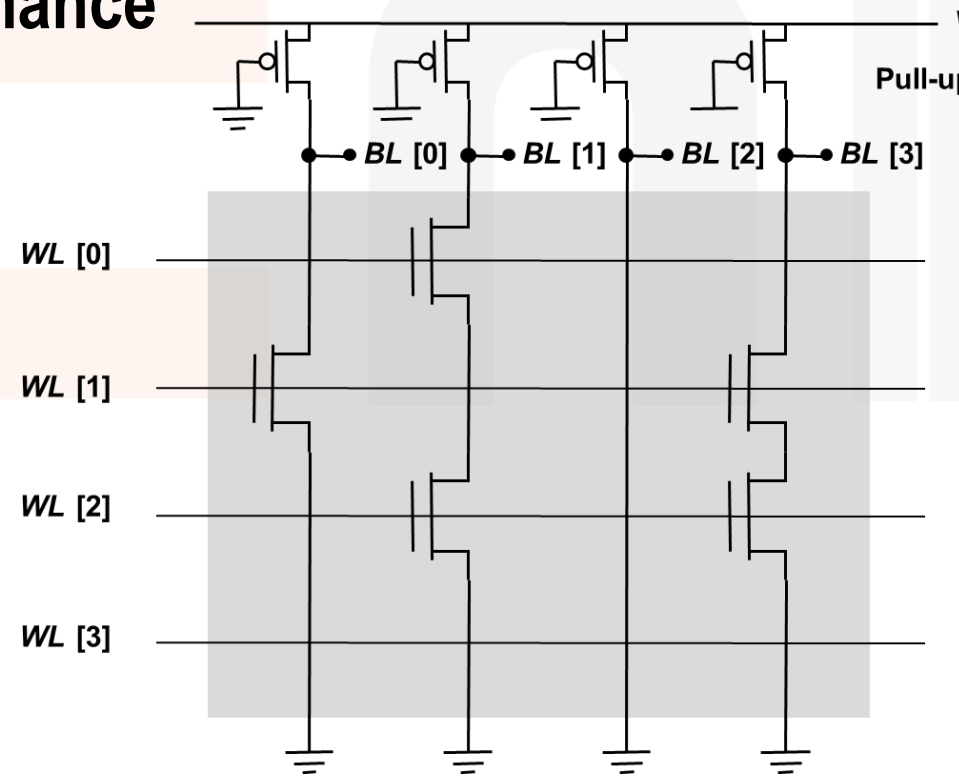


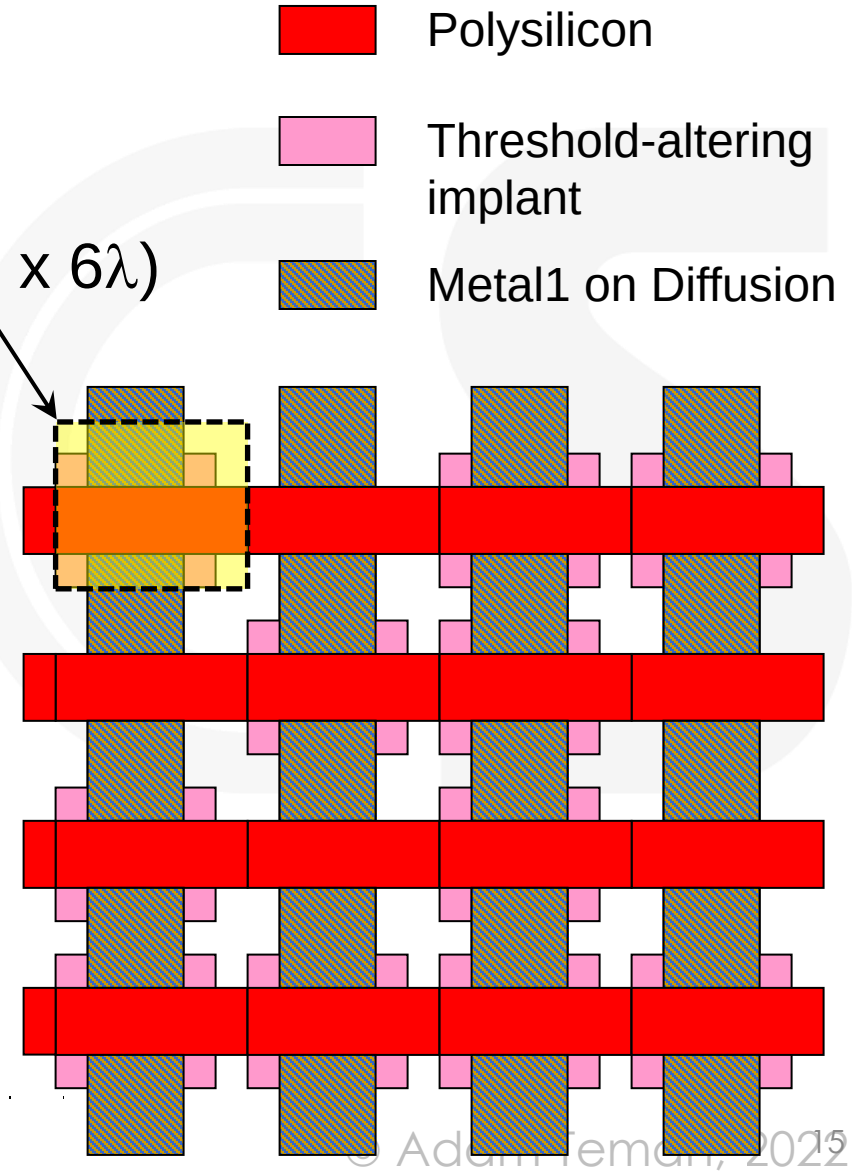


NAND ROM Layout

- Programming using Implants Only
- No contact to VDD or GND necessary

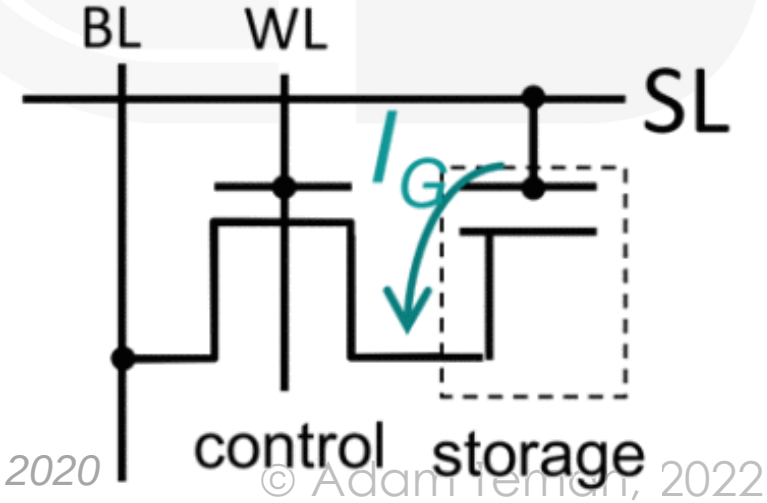
  → Drastically reduced cell size
- Loss in performance compared to NOR ROM





One-Time-Programmable Memory (OTP)

- **Mask ROM is set during fabrication**
  - An actual lithography mask encodes the ROM content.
  - A re-spin of the chip is needed in order to change the ROM.
- **Programmable Read-Only Memories (PROM) are an alternative**
  - Content of the PROM can be programmed (blown) after manufacturing
  - Generally, One-Time-Programmable (OTP) devices used today.
  - Often used to store boot code, encryption keys and configuration parameters.
- **Achieved by integrating an "antifuse" within the CMOS process**
  - Apply a high-voltage pulse (~6V) across the gate
    → oxide breakdown.



Source: Chung, 2020

Dynamic RAM (DRAM)

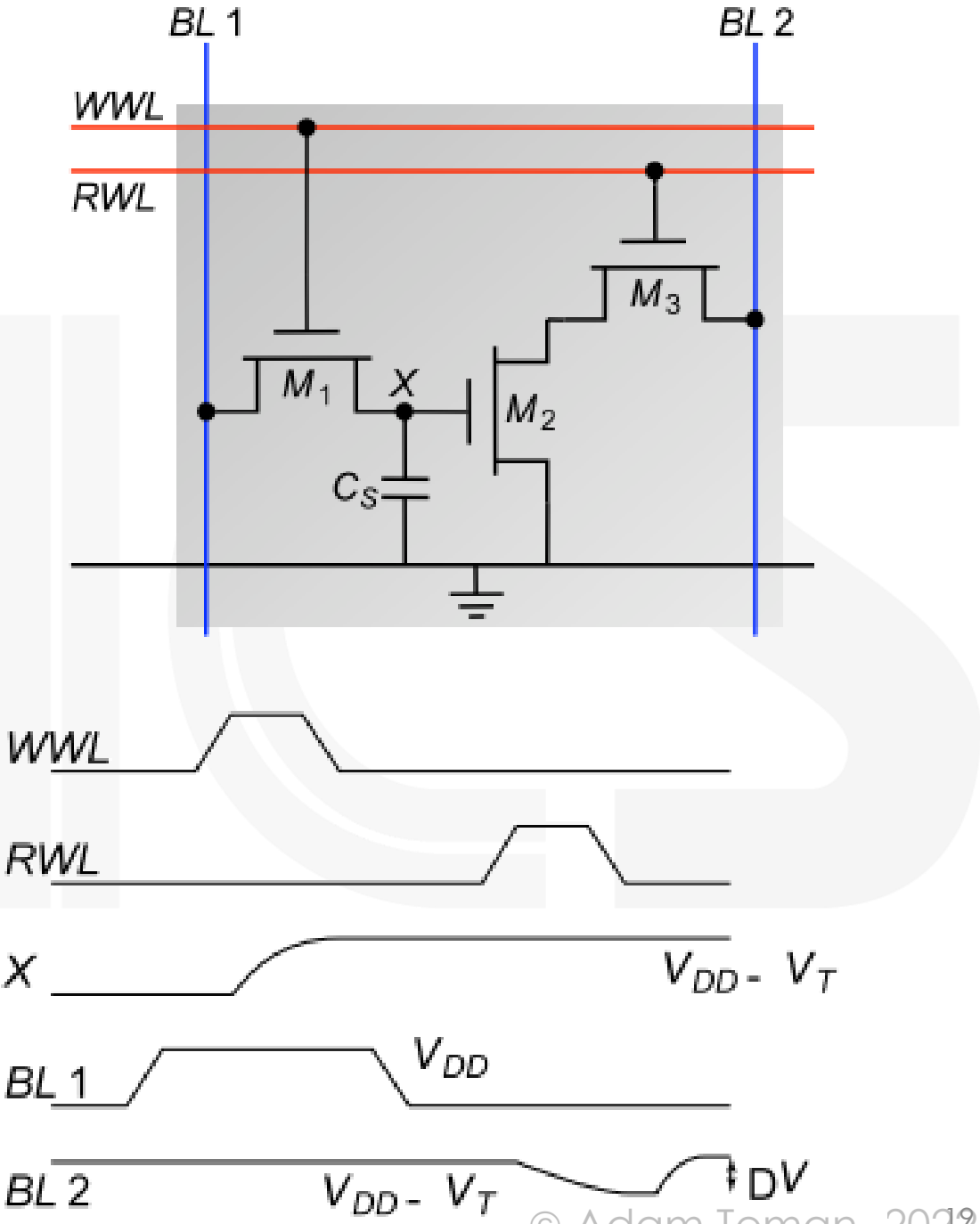
ROMs are very limited in usage

- **ROMs can be only used for data that is:**
  - Written during manufacturing
  - Can only be read.
- **OTPs:**
  - Can be configured after manufacturing
  - But still, only once.

- **How can we make a high-density (→ high-capacity) memory that can be used for both reading and writing?**

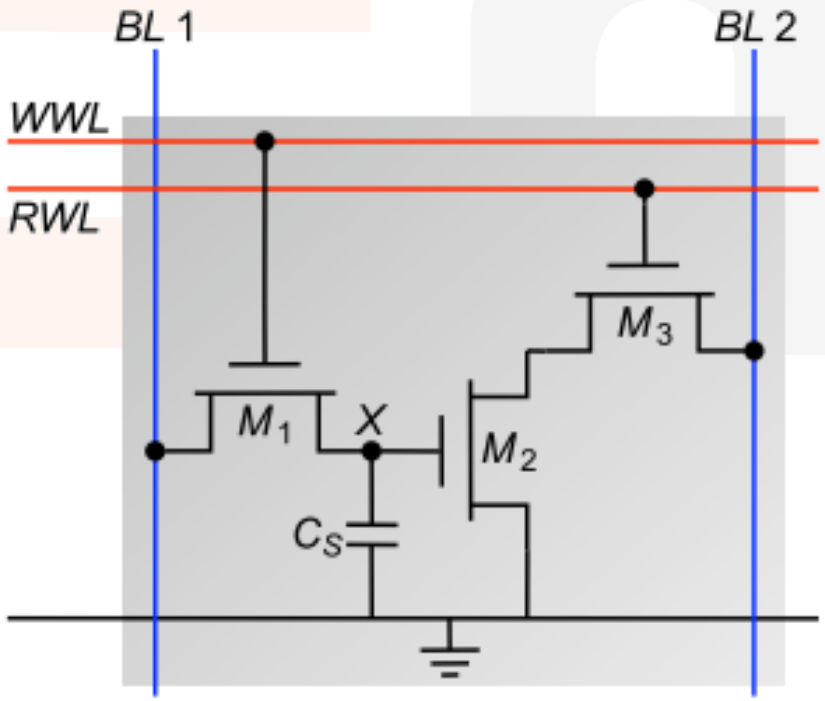
Let's go dynamic

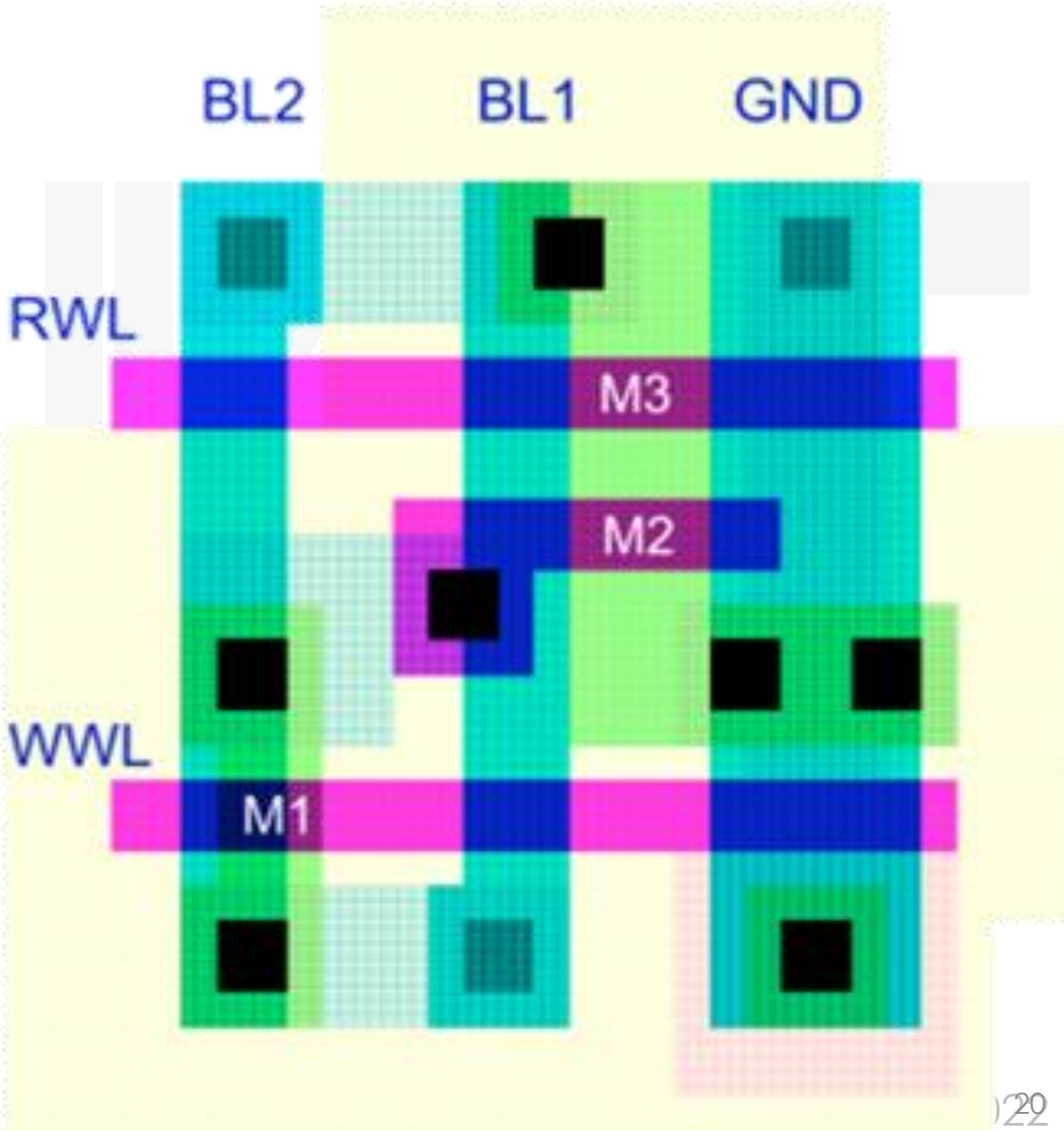
- **The Intel 1103 used a 3T DRAM cell**
  - Store data on capacitor (C_s)
  - Write through M_1
  - Read through M_2 and M_3
- **Advantages:**
  - No constraints on device ratios.
  - Reads are non-destructive.
- **Disadvantages**
  - Data leaks away over time
  - "Weak 1" stored on node X.
  - Single-ended read operation



3T DRAM Layout

- **Route word-lines (WWL, RWL) in Poly**
- **Route bit-lines (BL1, BL2) and GND in M2**
- **Disadvantage:**
  - Contact-dominated layout



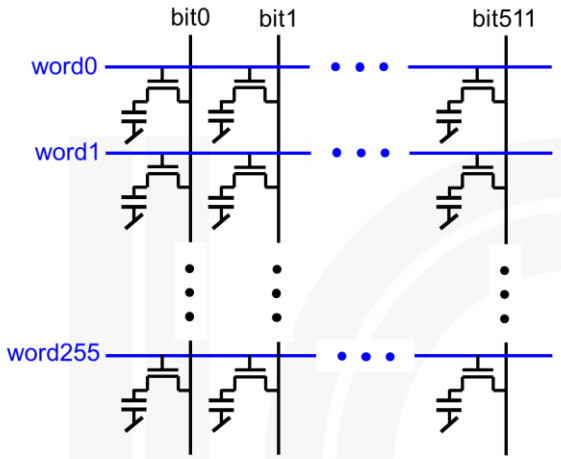


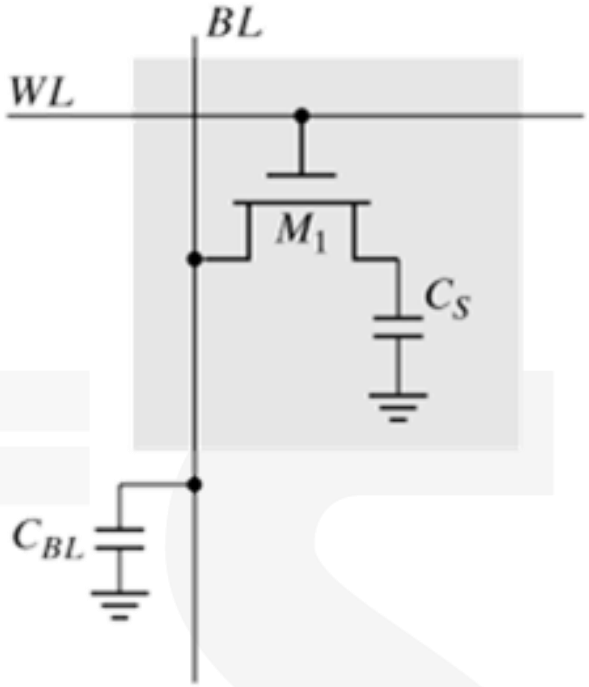
Can we go even smaller?

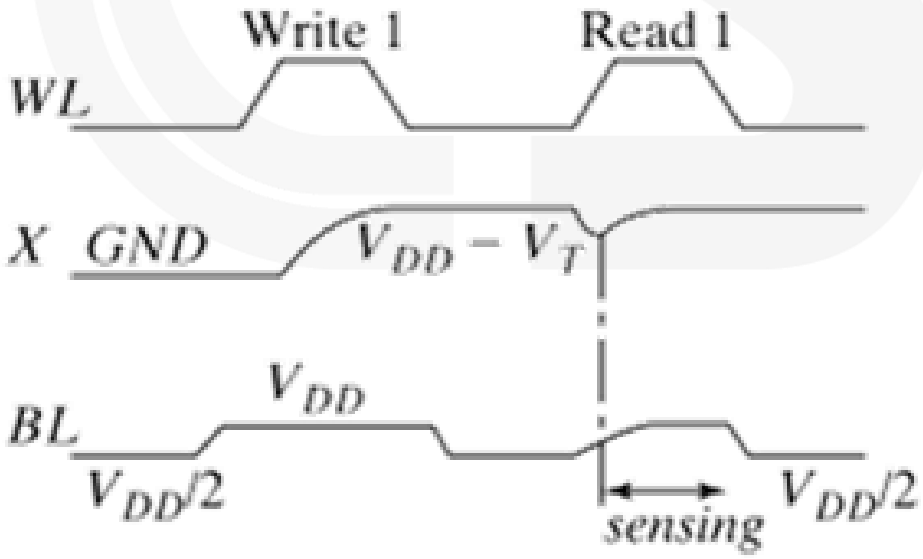
- **Dennard (IBM, 1967) patented a 1T cell**
  - Data is dynamically stored on C_s.
  - Write and read through M_1
- **Read achieved through charge sharing between C_s and C_{BL}:**

$$\Delta V = V_{\mathrm{BL}} - V_{\mathrm{pre}} = V_{\mathrm{BIT}} - V_{\mathrm{pre}} \frac{C_{\mathrm{S}}}{C_{\mathrm{S}} + C_{\mathrm{BL}}}$$

- **Drawbacks:**
  - Weak '1' stored on C_s
  - Charge sharing is slow, small delta
  - Read is destructive. Write back is necessary

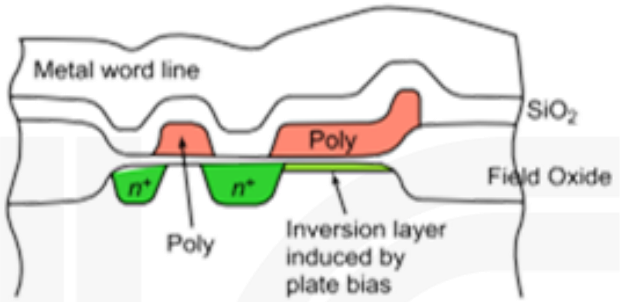




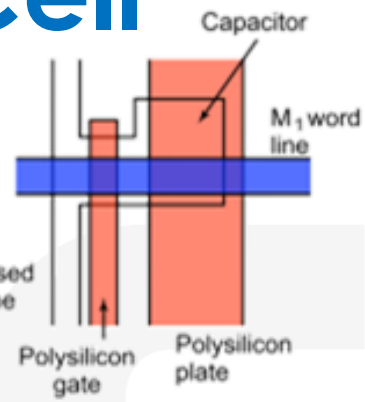


1-Transistor 1-Capacitor (1T-1C) DRAM Cell

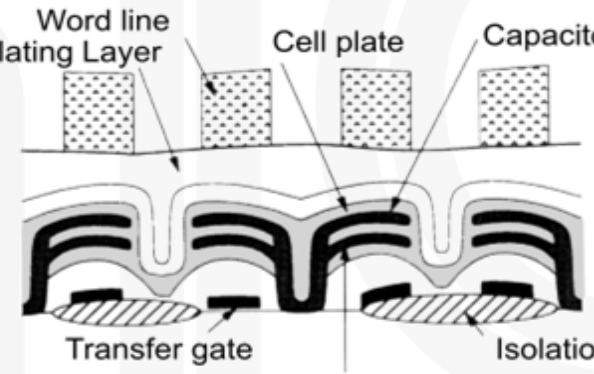
- **Standard process:**
  - Use polysilicon diffusion capacitance
  - Expensive in area (50 fF = 50 X $A_{bitcell}$)
  - Popular in 70s and 80s
- **Stacked Capacitor**
  - Make a “radiator” on top of the cell
- **Trenched Capacitor**
  - Dig a ditch under the cell
  - Fill it with conductor, insulator, conductor

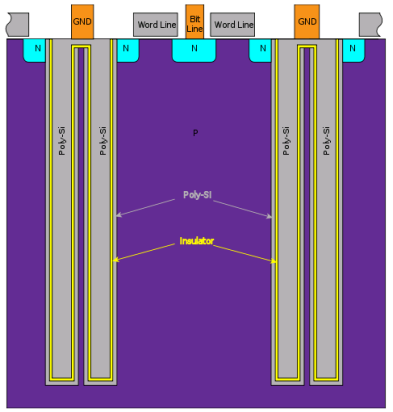


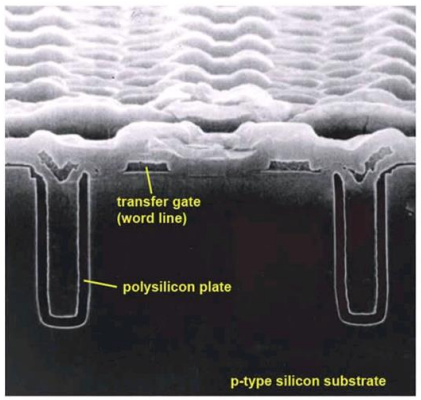
Cross-section

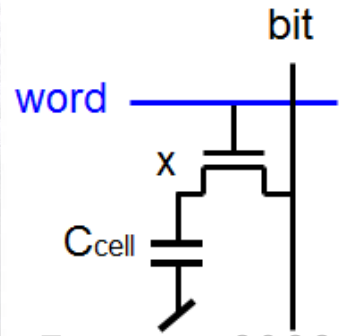


Layout



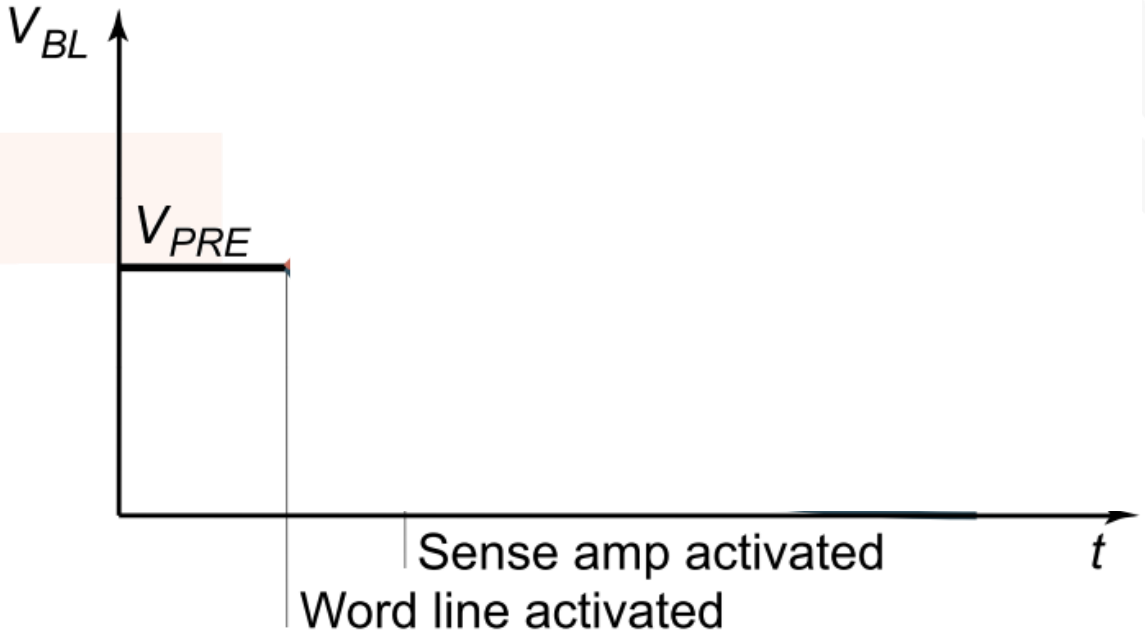


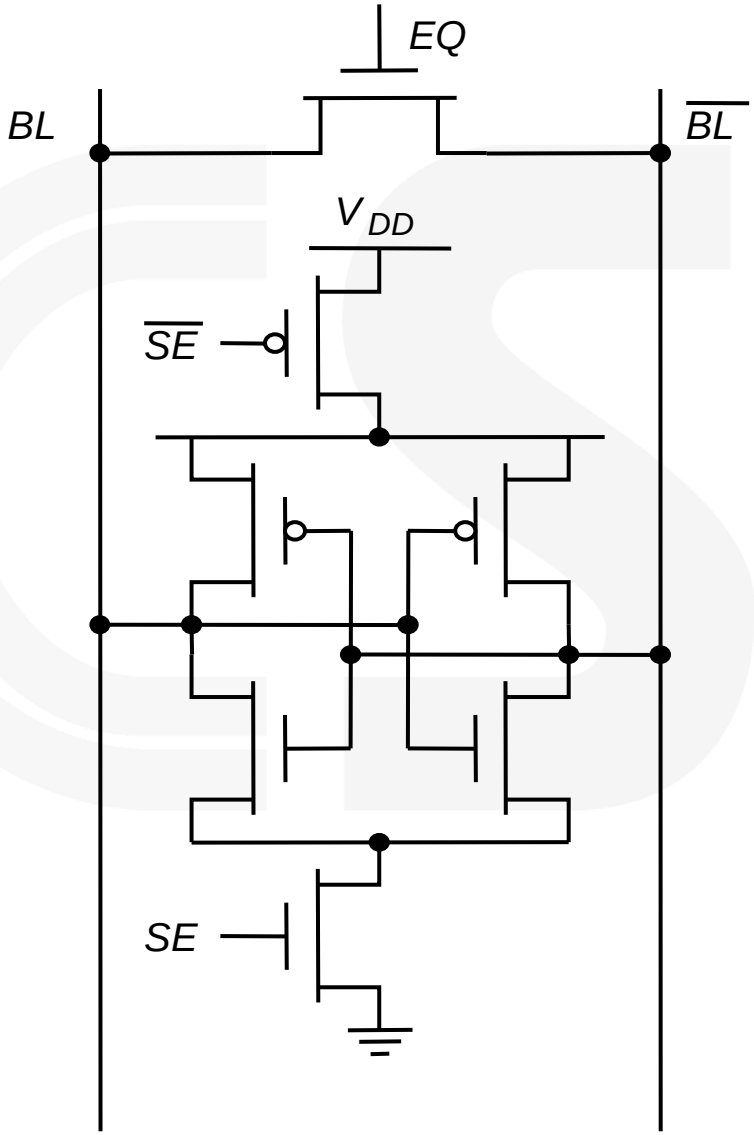




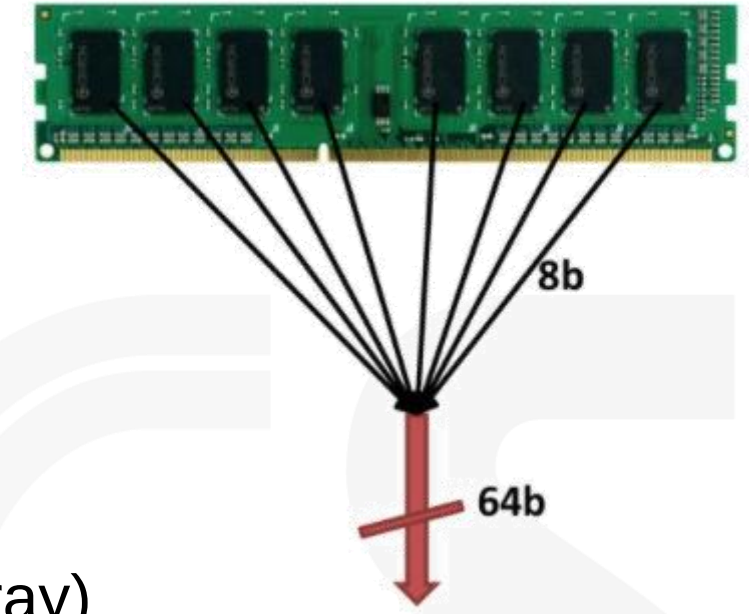
Latch-Based Sense Amplifier (DRAM)

- Initialized in its meta-stable point with EQ
- Once adequate voltage gap created, sense amp enabled with SE
- Positive feedback quickly forces output to a stable operating point.

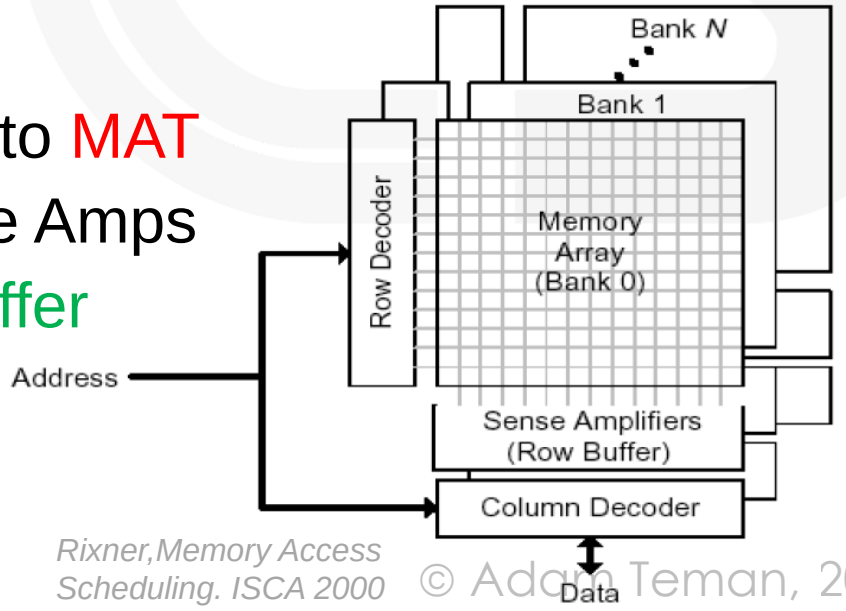




DRAM Architecture



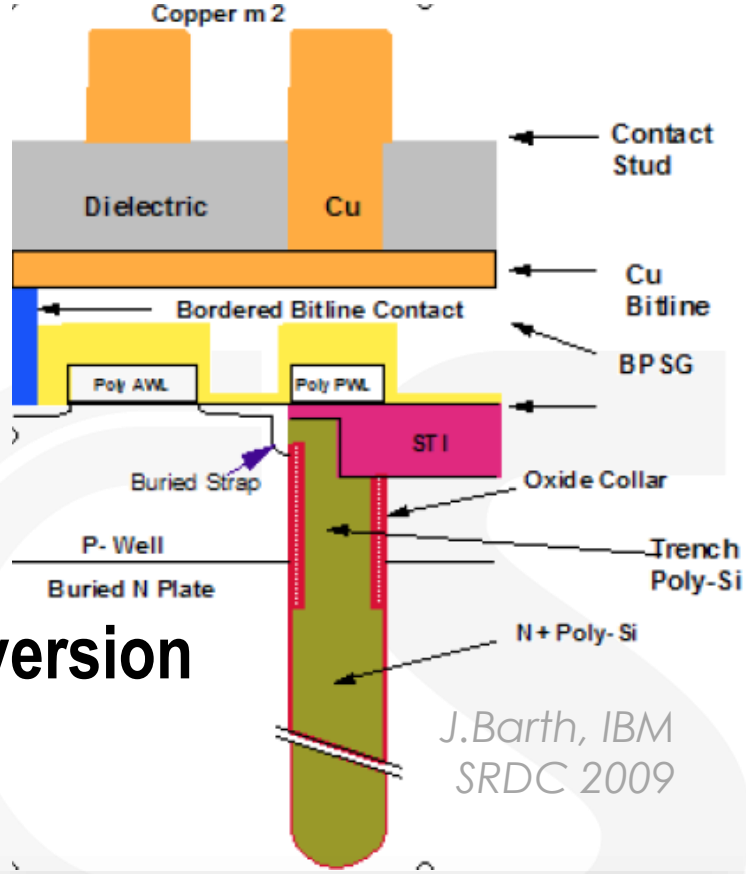
- **To get a 16GB DRAM DIMM (Dual-Inline Memory Module)**
  - DIMM has two ranks
  - Each rank (64-bit I/O) has 8 chips (1 GB/chip)
  - Each chip (8-bit I/O) has 8 banks (128 MB/bank)
  - Each bank (1-bit I/O) is divided into 8 arrays (16 MB/array)
  - Each array is divided into ~128 1024x1024 bit MATs (sub-arrays)
- **So, to access DRAM:**
  - First provide Row Access Strobe (RAS) – 10 bits to MAT
  - Readout latches rows (~128 x 1024 bits) in Sense Amps
  - ~7 bits of RAS select 1024 bits into bank Row Buffer
  - Then Column Access Strobe (CAS)
  - Select 1-bit from Row Buffer to bank output.



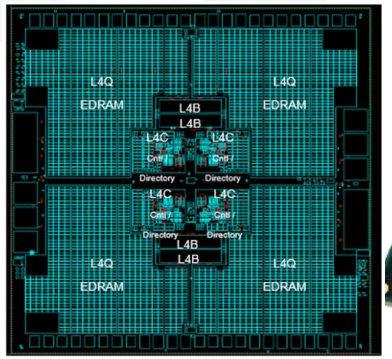
Rixner,Memory Access Scheduling. ISCA 2000

Embedded DRAM

- **DRAM requires a special (dedicated) process (Fab)**
  - Manufactured by only a handful of providers (predominantly Micron, Samsung and SK Hynix)
  - Is supplied on an external standalone chip.
- **However, it is possible to make an on-chip "embedded" version**
  - Known as "eDRAM"
  - Requires special process cost-adders to fabricate trench capacitors (IBM) or stacked MIM (Intel)
- **Recent chips with L4 eDRAM**
  - Intel Broadwell – 128 MB
  - IBM z15 SC – 960 MB

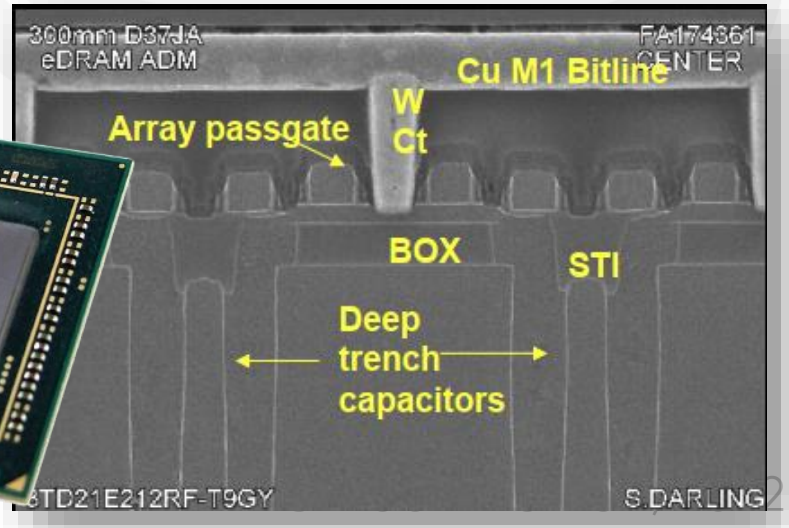


J.Barth, IBM SRDC 2009



Source: Wikichip

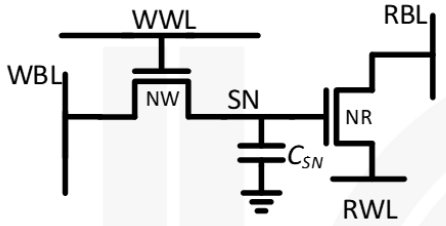
Source: AnandTech



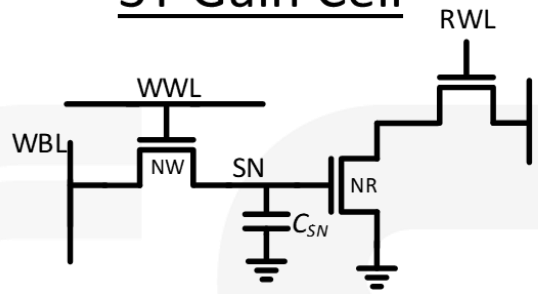
Gain-Cell Embedded DRAM

- **eDRAM is not "Logic Compatible"**
  - Process adders are expensive
  - Not available for many processes
  - Yet to be demonstrated under 14nm
- **Why not go back to the old Intel DRAM?**
  - Use parasitic capacitance for storage node
  - Manipulate current gain of read transistor
  - Known as "Gain-Cell embedded DRAM"
  - 2T, 3T, 4T, and others demonstrated
- **Most-scaled demonstrated GC-eDRAM**
  - "DynOR" – 28nm FD-SOI (Giterman, TCAS-I 2017)
  - "DAFNA" – 28nm (Giterman, JSSC 2018)
  - "Negev" – 16nm (Giterman, SSC-L 2020)

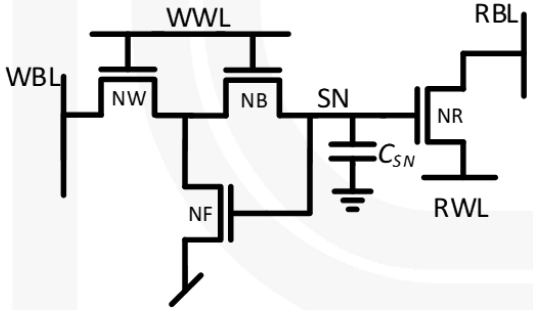
2T Gain Cell

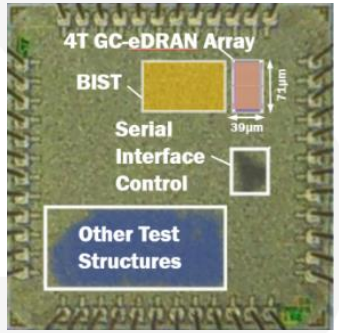


3T Gain Cell

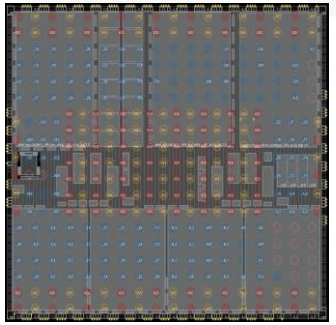


4T Gain Cell

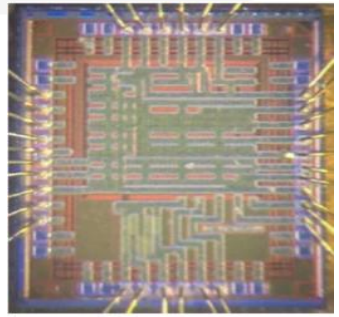




DynOR



Negev



DAFNA

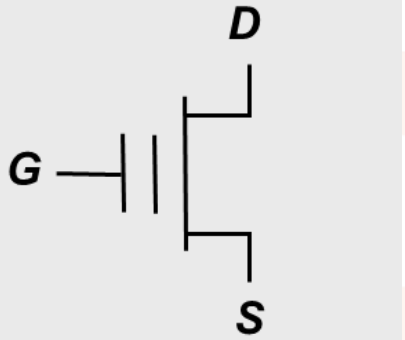
Non-Volatile Memory (NVM)

Memory Volatility

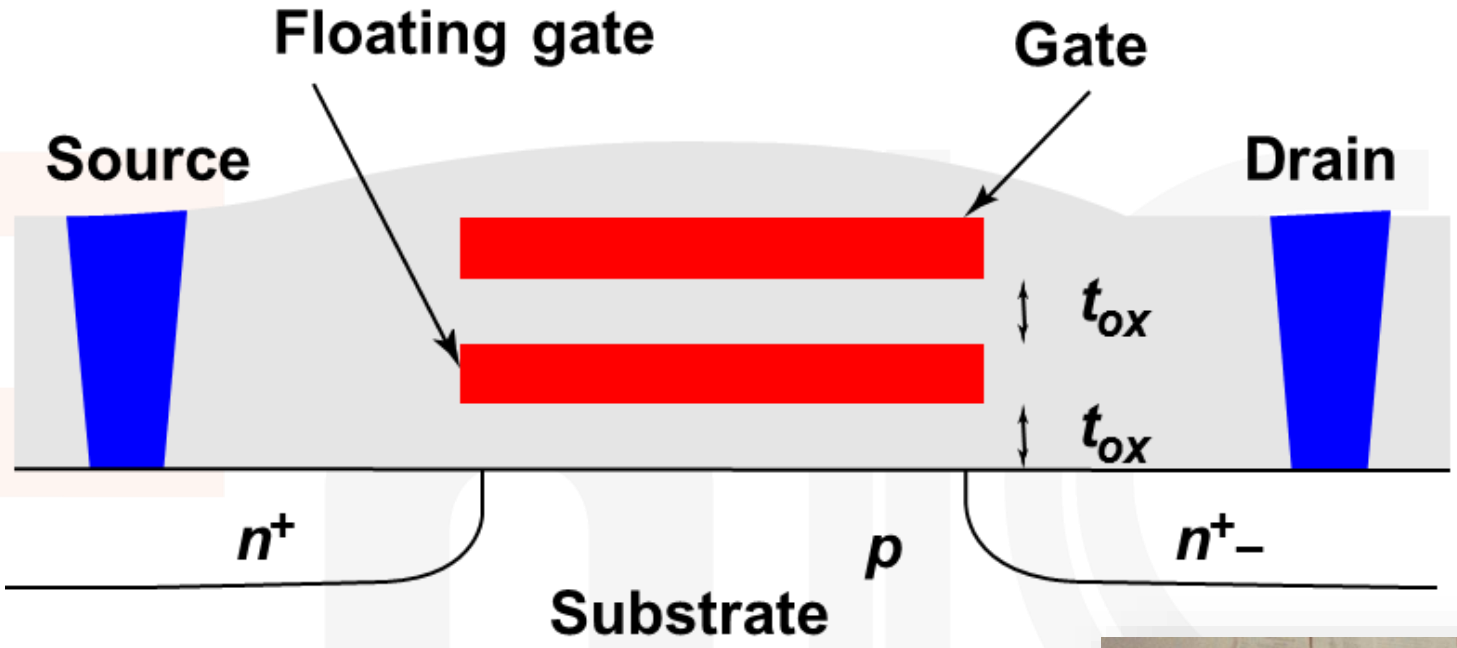
- **SRAM and DRAM are examples of *Volatile* memories**
  - They only store their data as long as a power source is connected
- **Mask ROM and OTP are *non-Volatile***
  - Their content is saved regardless of the state of the battery

- **What about non-Volatile Read/Write Memories?**

The Floating-gate transistor (EPROM)



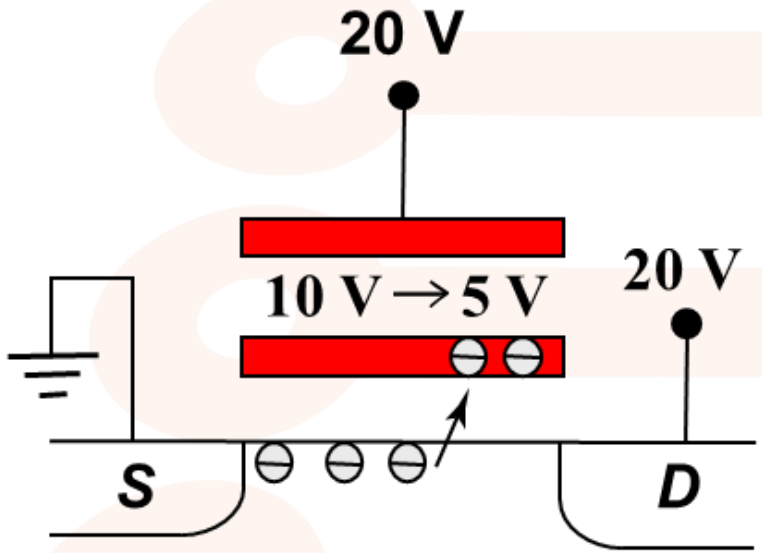
Schematic symbol



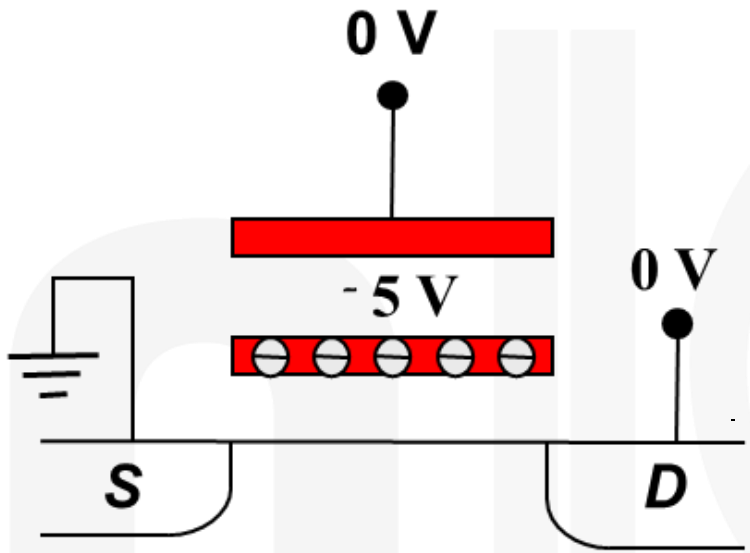
"as important in the development of the microcomputer industry as the microprocessor itself."

Gordon Moore on Dov Frohman's EPROM

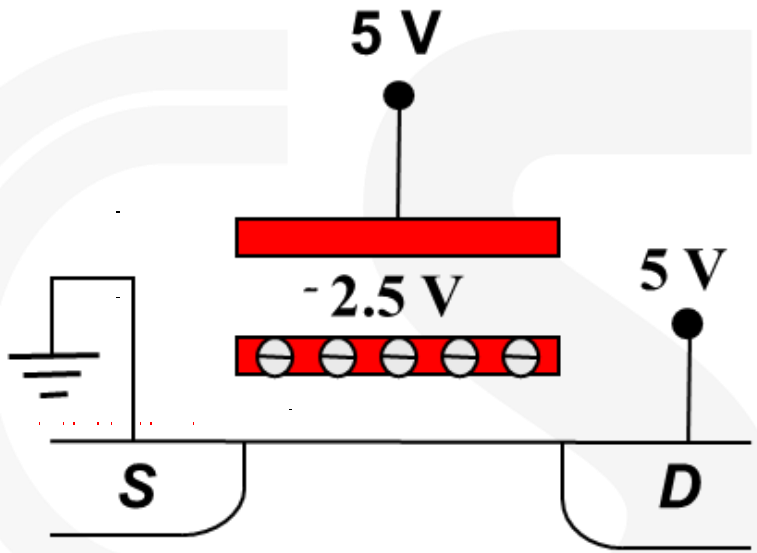
Floating-Gate Transistor Programming



Avalanche injection

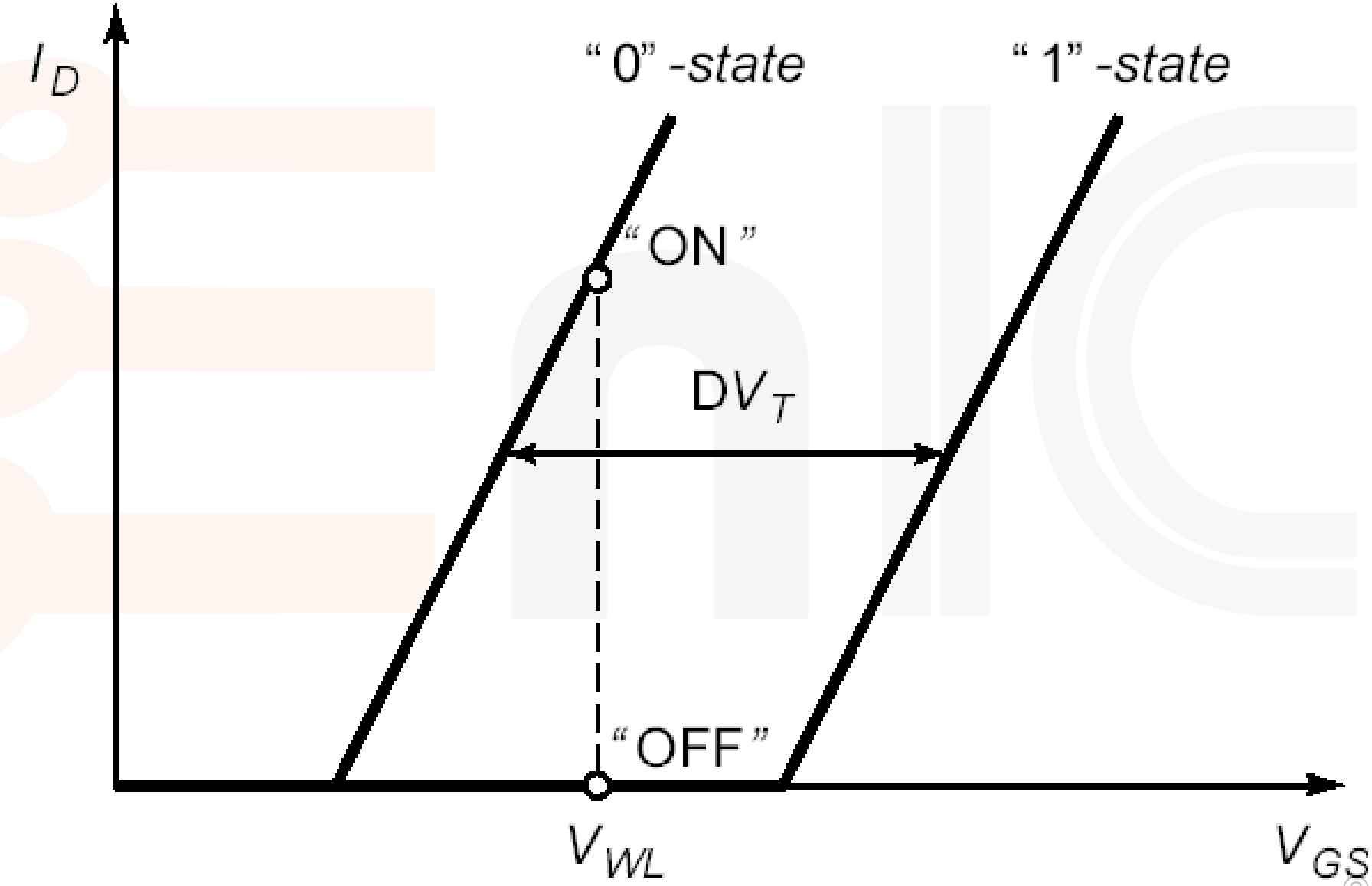


Removing programming voltage leaves charge trapped

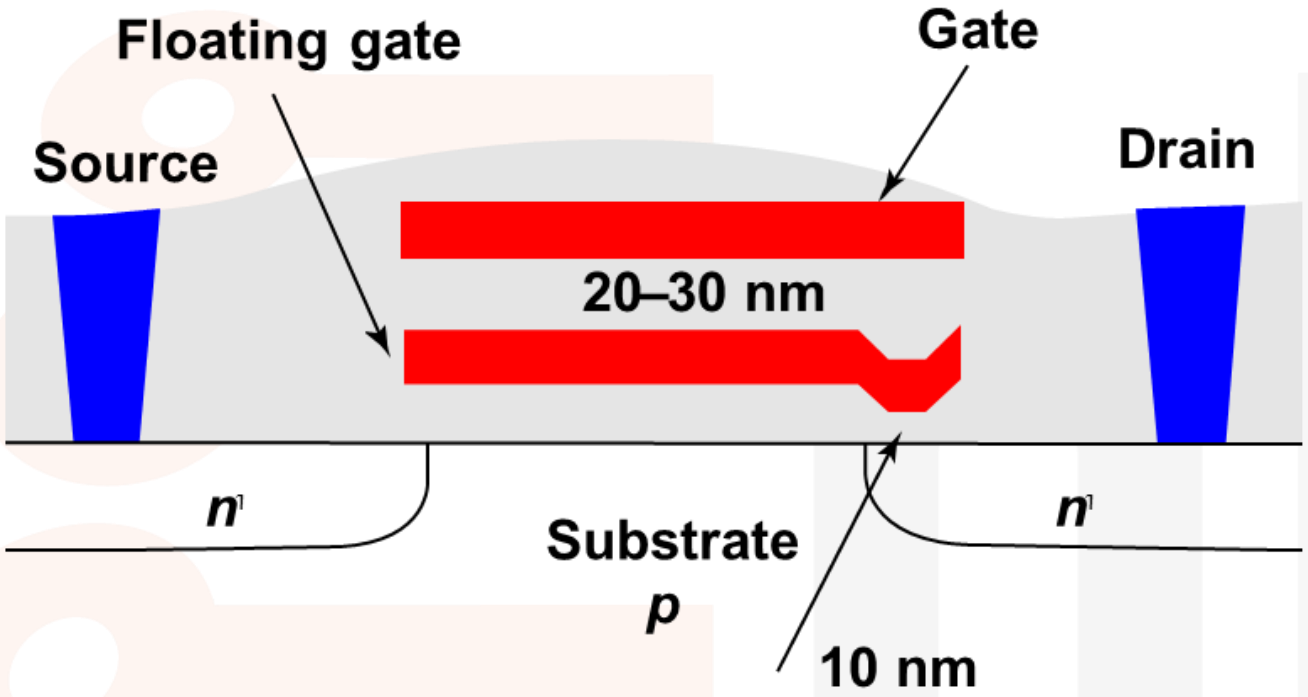


Programming results in higher V_T.

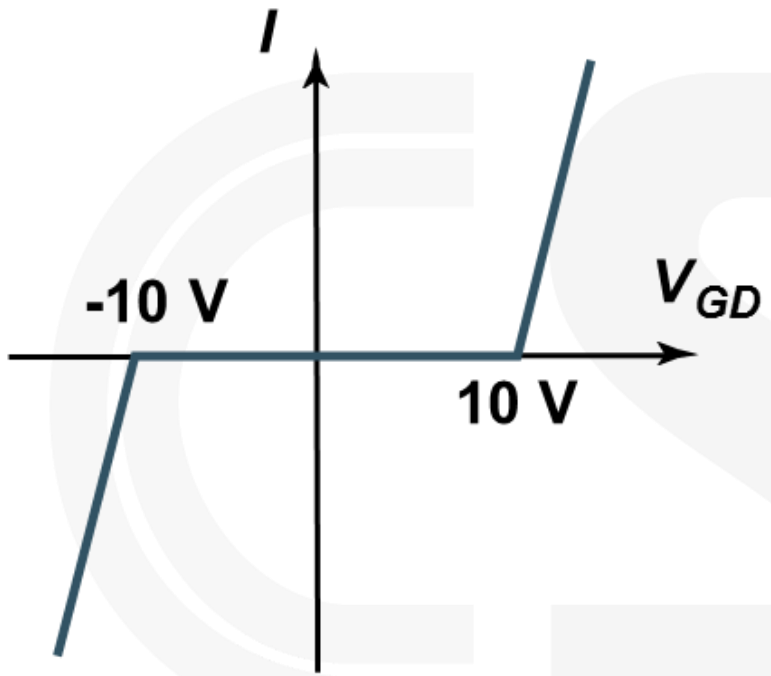
A “Programmable-Threshold” Transistor



FLOTOX EEPROM



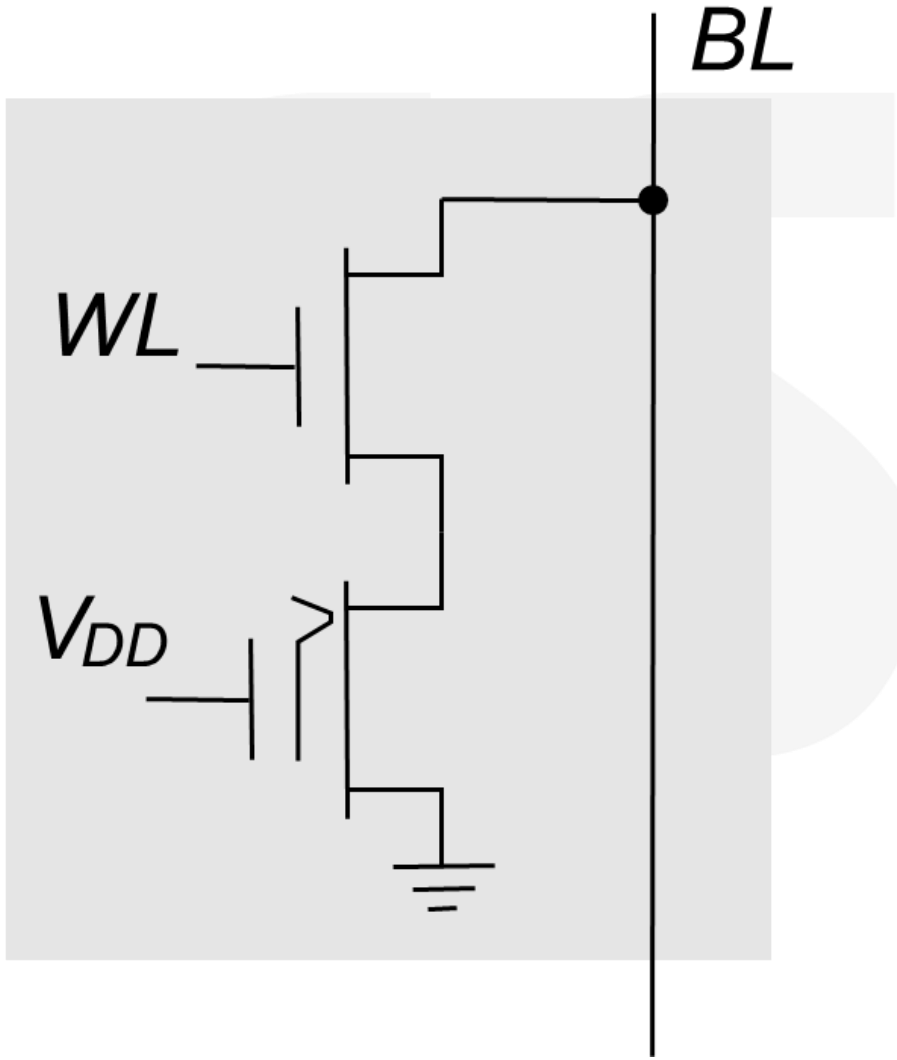
FLOTOX transistor



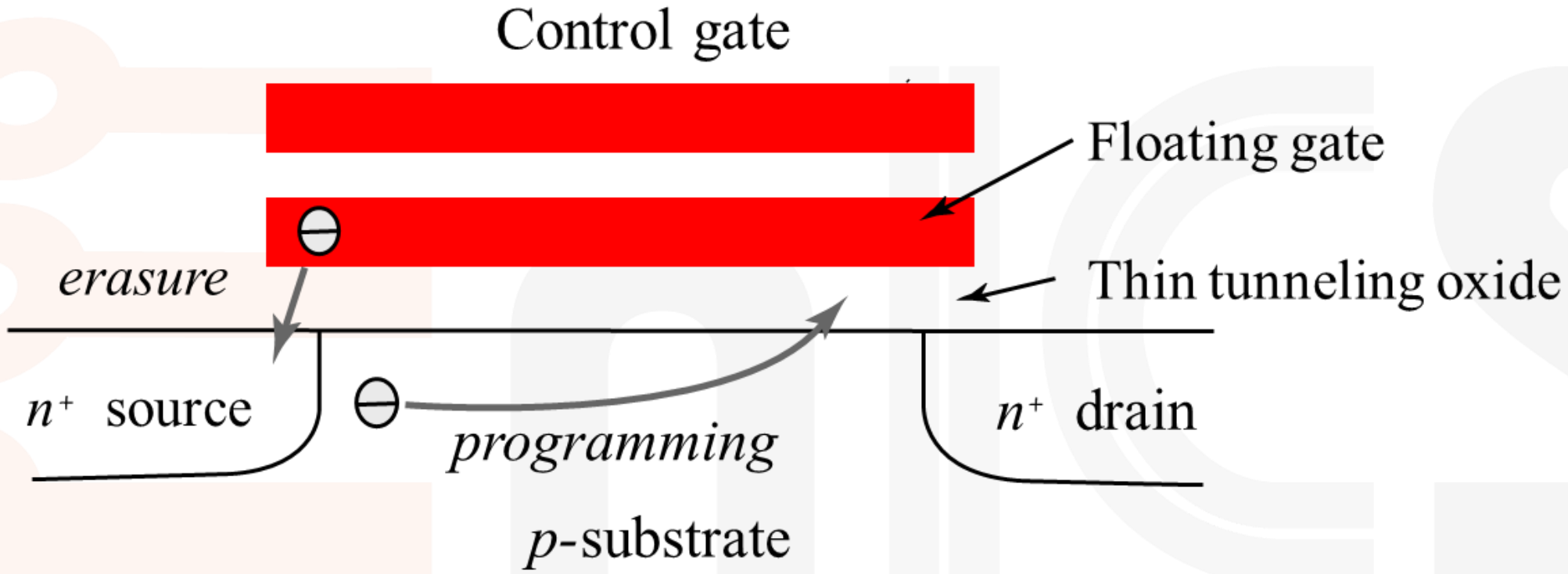
Fowler-Nordheim
I-V characteristic

EEPROM Cell

- **Absolute threshold control is hard**
  - Unprogrammed transistor might be depletion
- **Therefore, use a 2 transistor cell:**
  - Program it such that VT>VDD
  - Erase it, such that VT is really low (even VT<0)

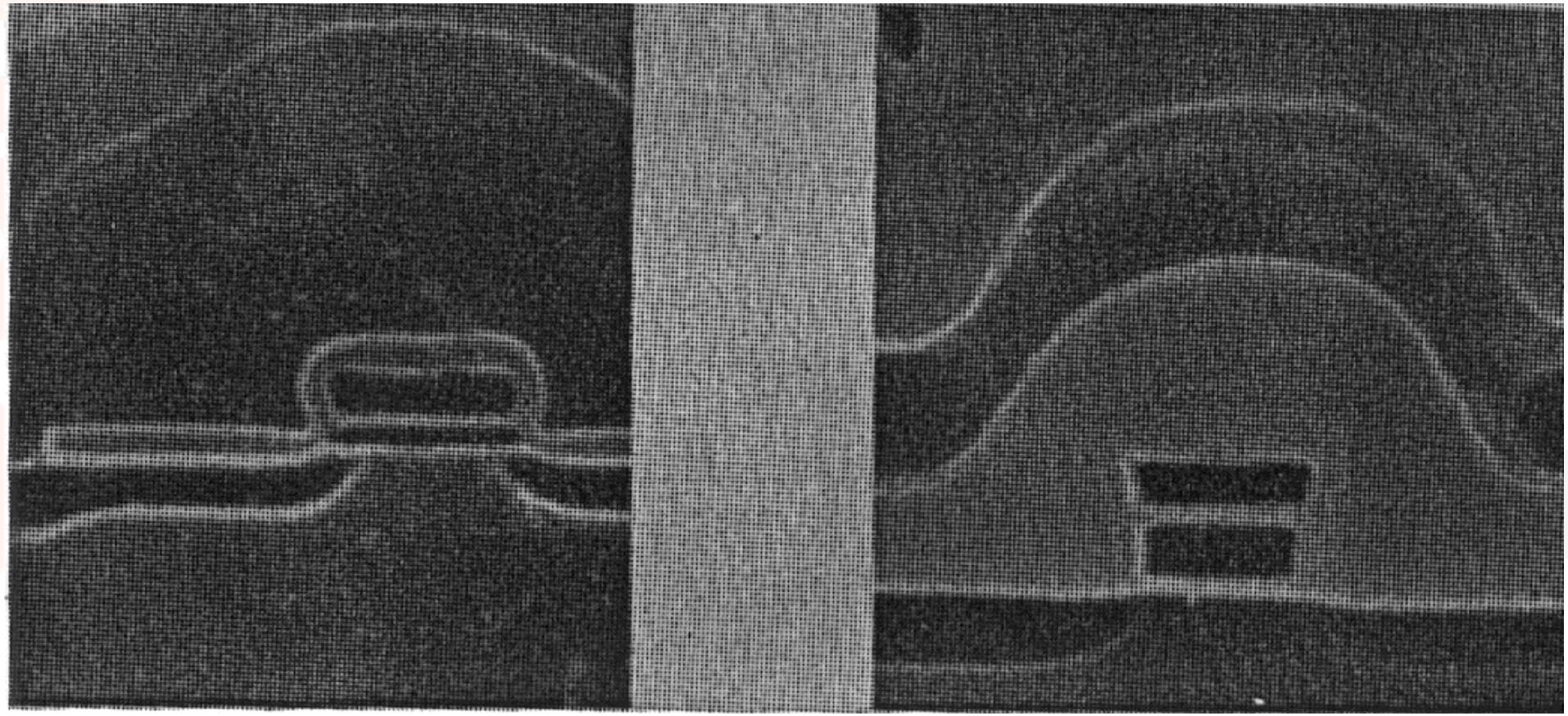


Flash EEPROM



Many other options …

Cross-sections of NVM cells



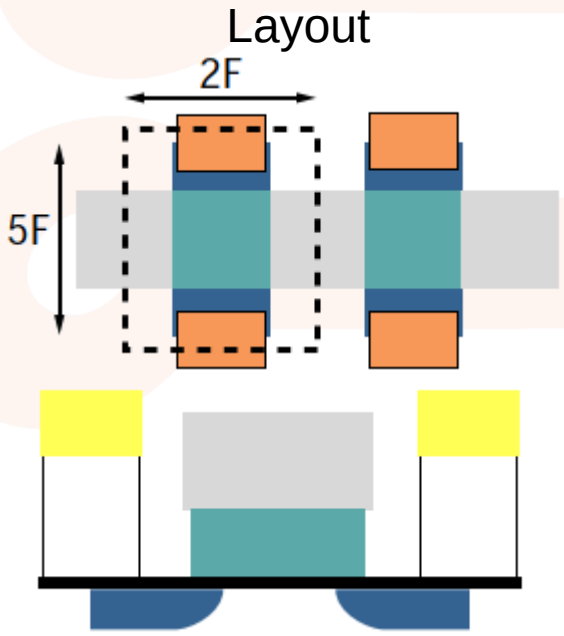
Flash **EPROM**

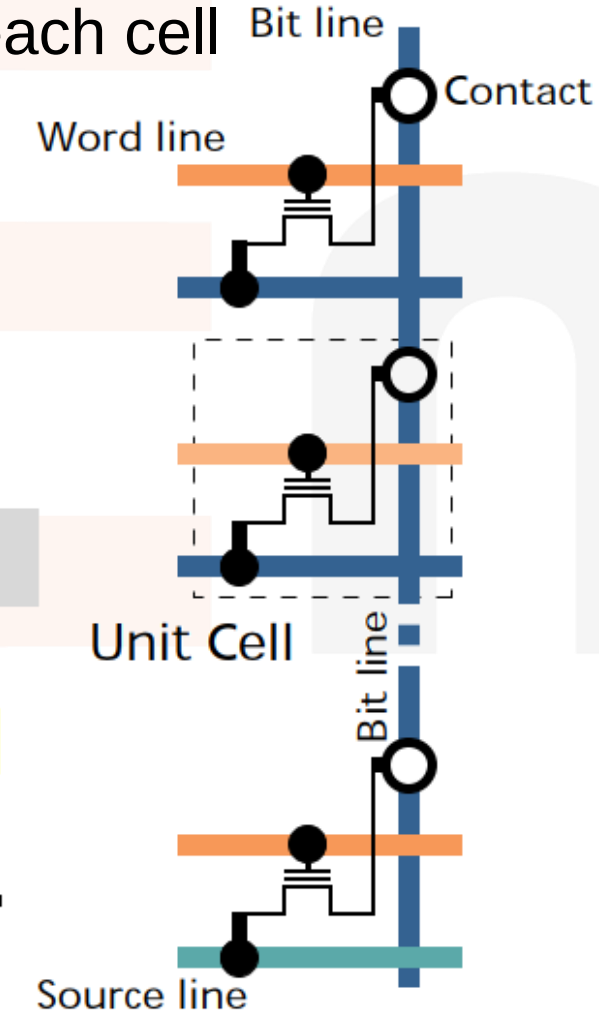
Courtesy Intel

Two major architectures: NOR and NAND

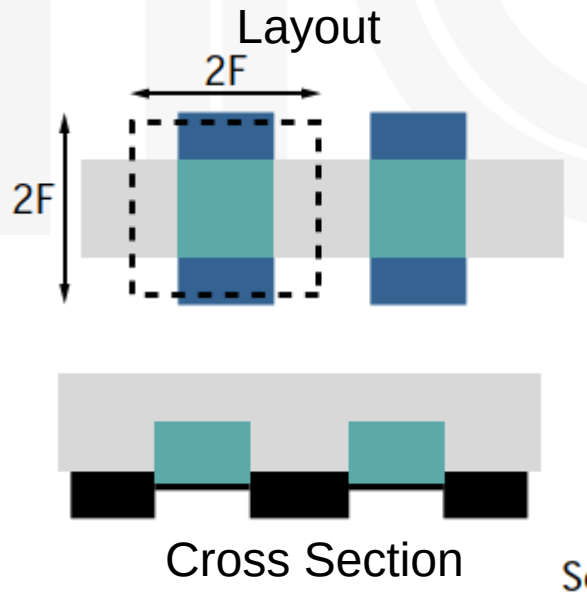
Source: Micron

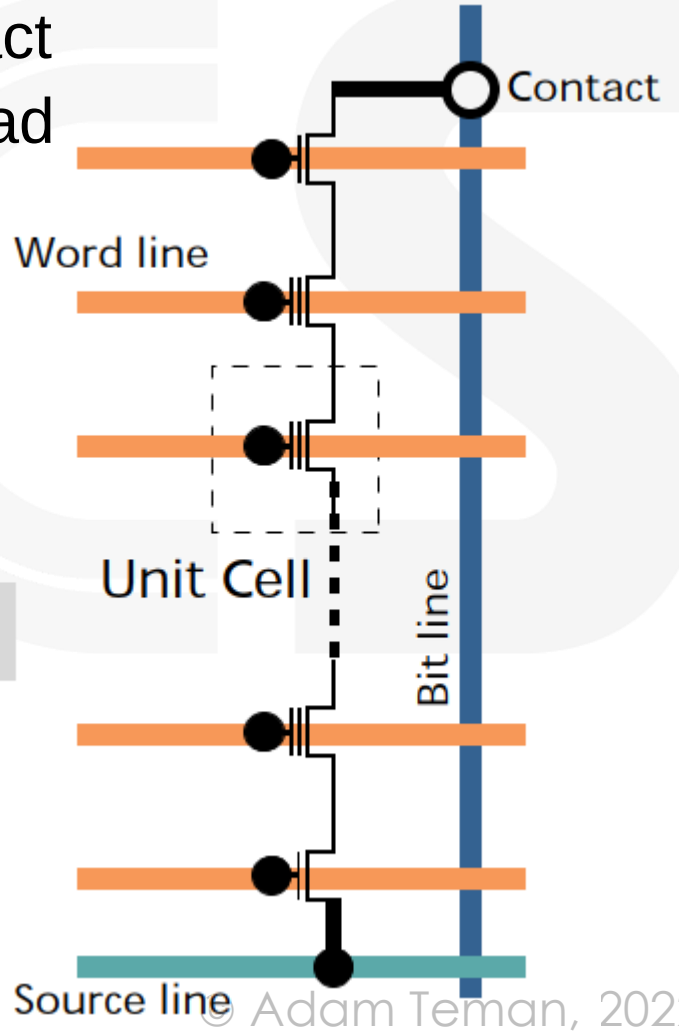
- **NOR Flash**
  - BL contact in each cell
  - Fast read
  - Cell size: $10F^2$





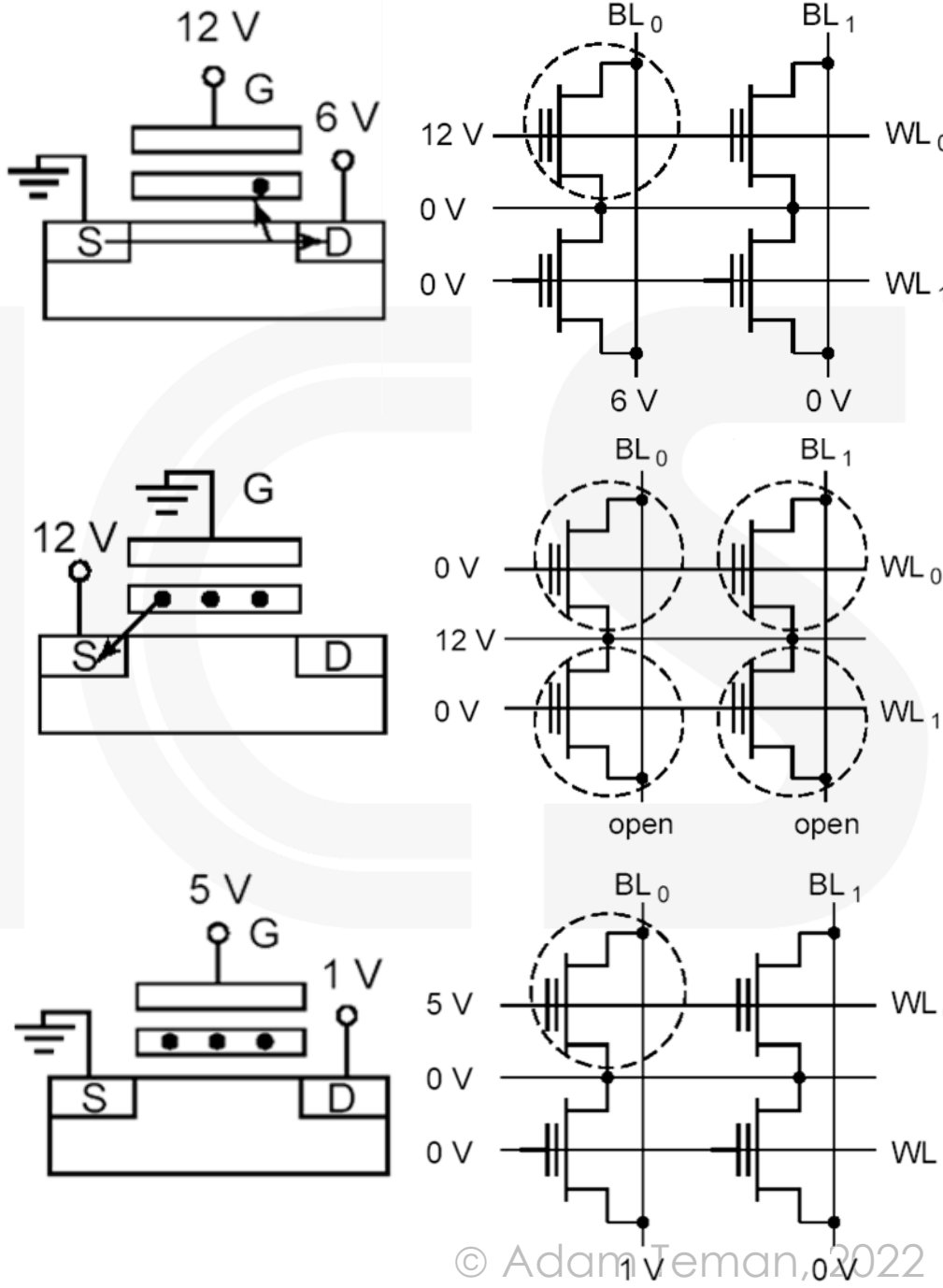
- **NAND Flash**
  - Single BL contact
  - Slow (serial) read
  - Cell size: $4F^2$





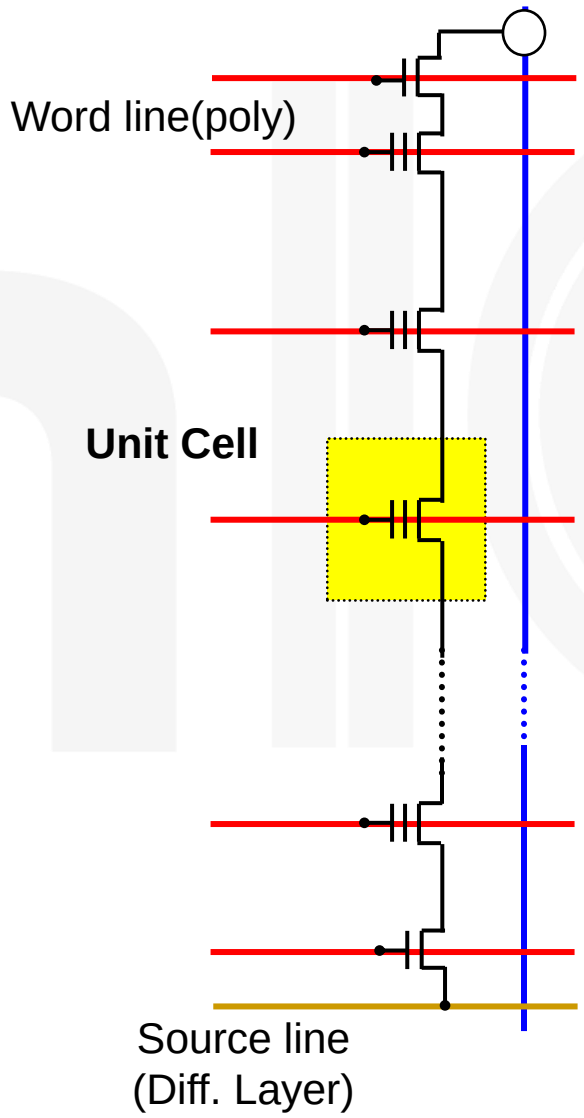
NOR Flash Operation

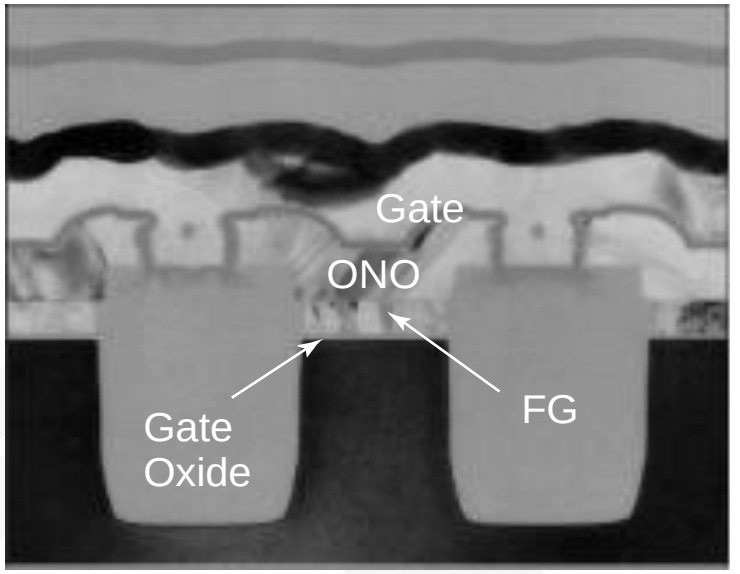
- **Program**
  - Run current through and apply vertical field
  - Takes a long time (µsecs)
  - Readout and see if programming successful.
- **Erase**
  - First program all cells in block
  - Then Erase entire block and verify
- **Read**
  - Apply median voltage on the gate (e.g., 5V)
  - ‘0’ State – BL will discharge
  - ‘1’ State – BL will stay charged



NAND Flash Operation

- **Erase:**
  - Source Line = 20V
  - BL=20V WL=0
  - All devices normally on
- **Program (Write '1'):**
  - Source Line disconnected
  - BL=0V
  - $WL_{selected}$=20V $WL_{unselected}$=5V
- **Read:**
  - Source Line connected
  - BL connected
  - $WL_{selected}$=0V $WL_{unselected}$=5V





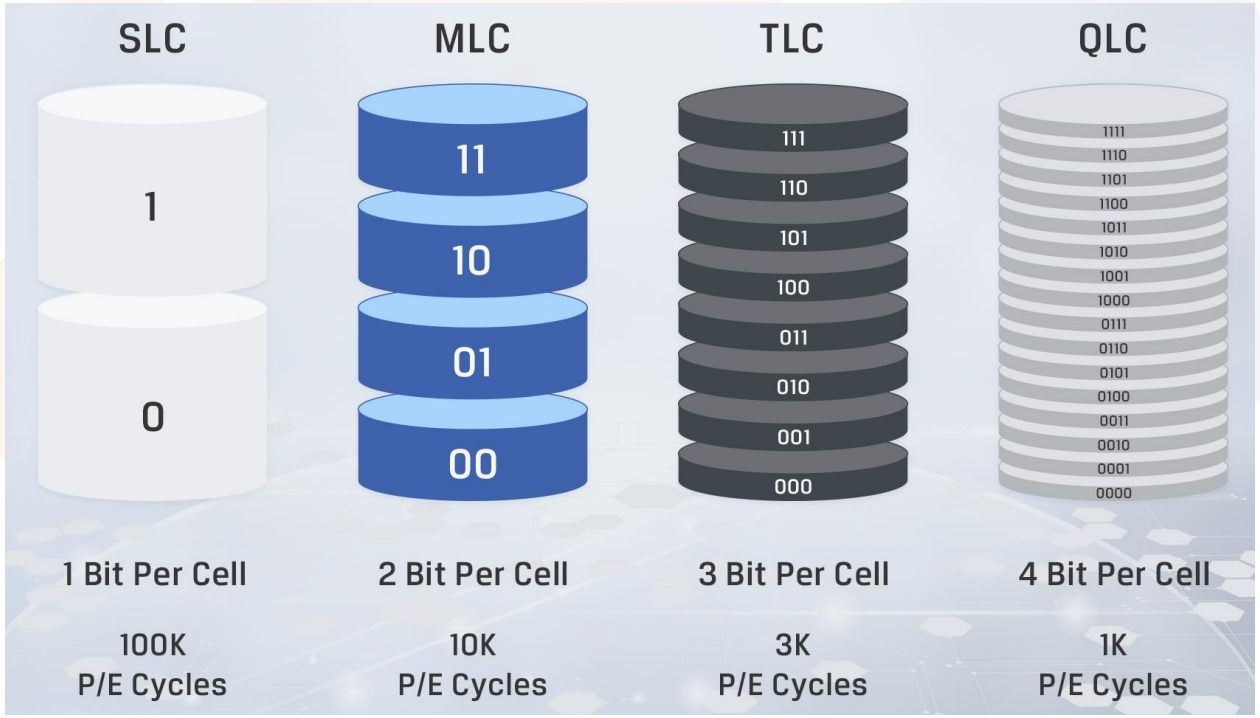
Source: Toshiba

NAND vs. NOR

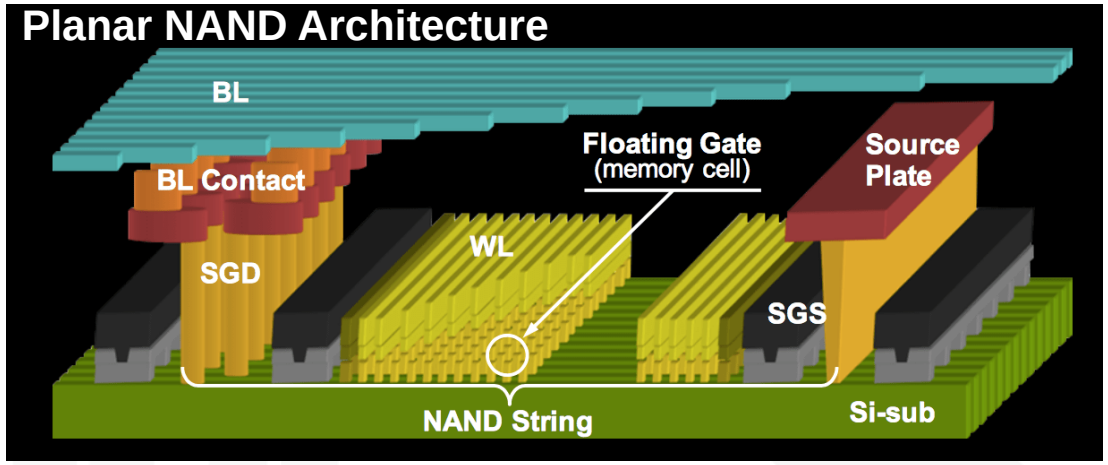
- **NOR architectures are large, but fast.**
  - Actually, NOR is fast for READs, but slow for WRITEs and ERASEs, as they need precise control over threshold voltages.
  - NORs are used for applications such as program code storage that are read a lot and at high speeds.
- **NAND architectures are slow, but small.**
  - Actually, just READs are slow. Programming and Erasure are relatively fast.
  - Used for video and audio file storage, where we need high density, as well as fast write/erase times.
  - NAND is about 40% smaller than NOR and generally uses Fowler Nordheim tunneling for program and erase.

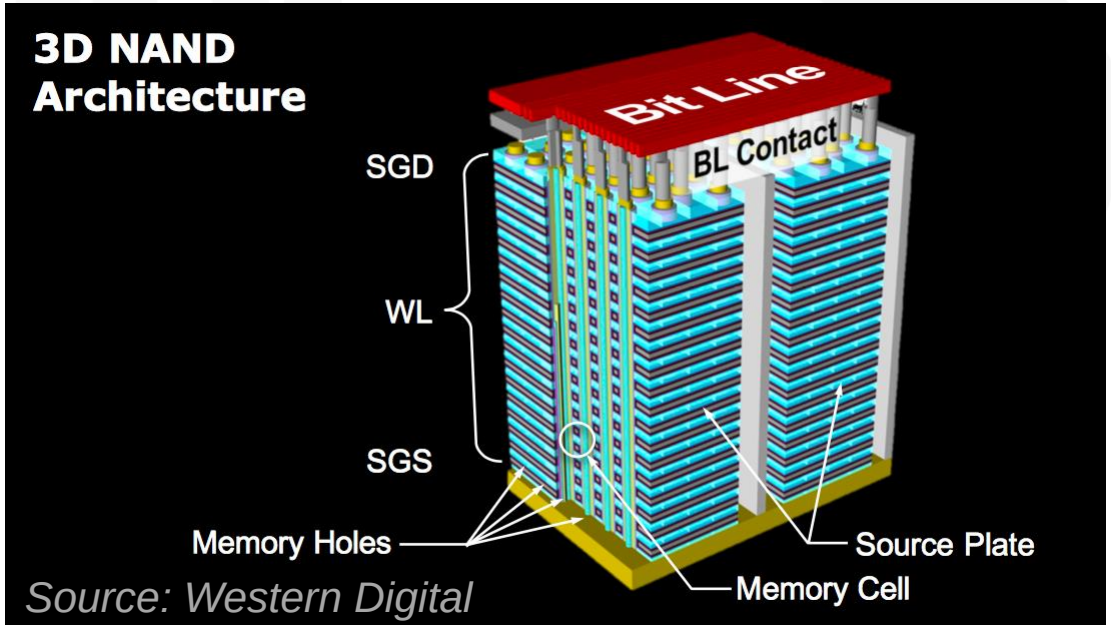
Increasing Flash Density

- **Add levels to every bit**
  - MLC = Multi-level Cell
- **Or perhaps go in the third dimension…**



Source: Kingston





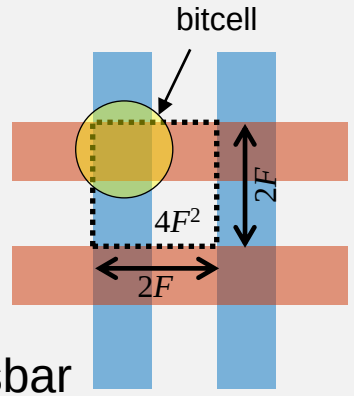
Emerging Memories

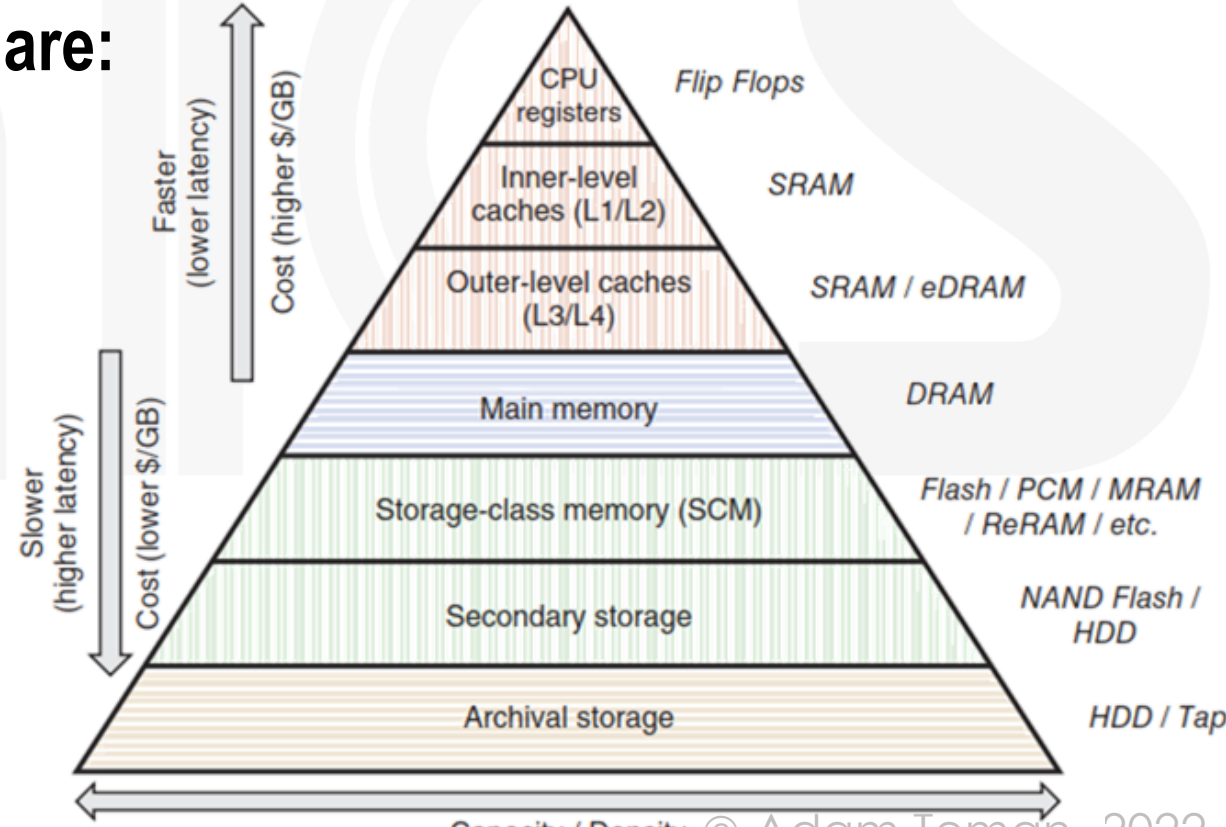
The "Universal Memory"

- **A single memory technology with the advantages of all existing technologies and without their limitations.**
- **The requirements of a Universal Memory are:**
  - High Density → Flash
  - Scalability → SRAM
  - Unlimited Retention → Flash/HDD
  - High Performance → SRAM
  - Unlimited endurance → SRAM/DRAM
  - Process Integration → SRAM/eDRAM
  - Low Power → NVM + Low Voltage

F^2 (F-squared)

Minimum bitcell size measured as F^2, where F is the minimum feature size.
The smallest possible structure is $4F^2$ achieved through a crossbar

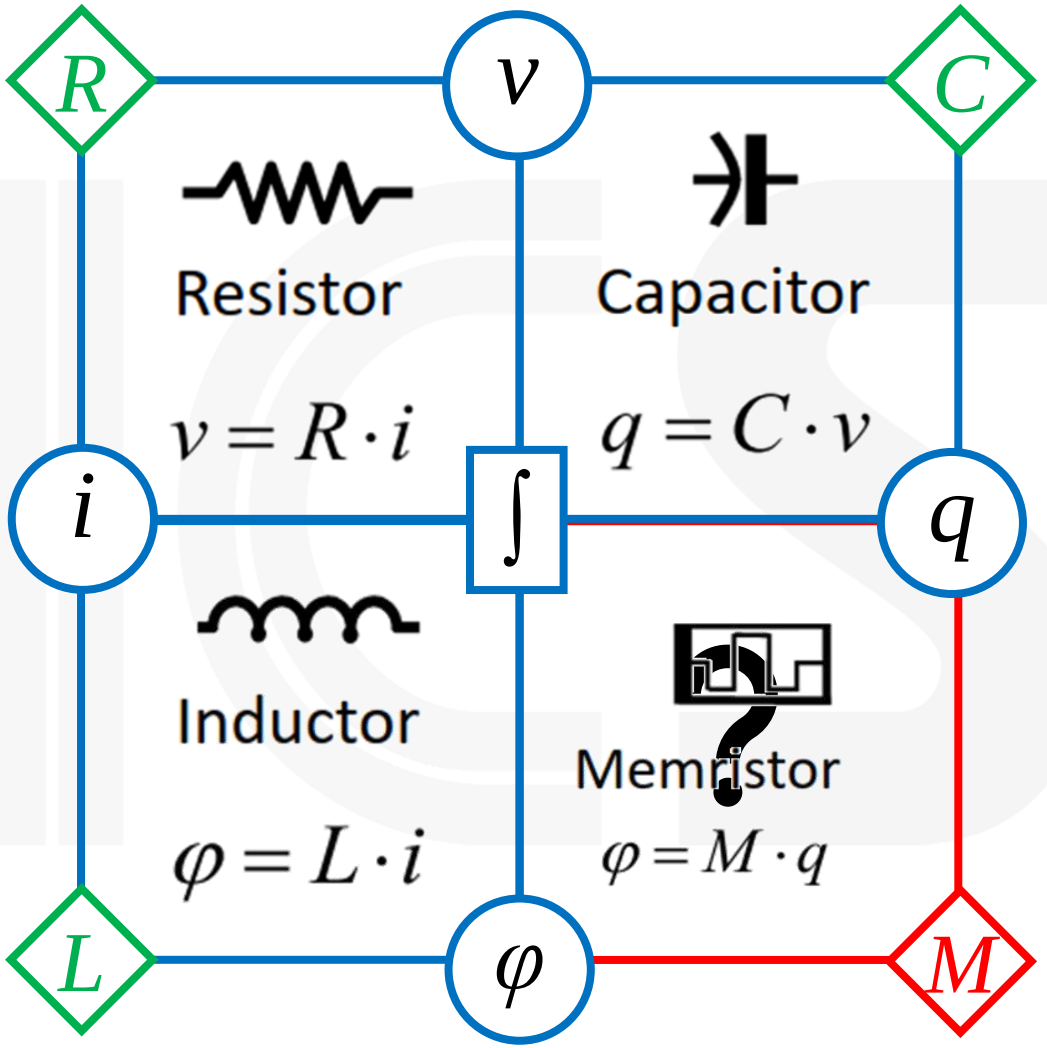




Fundamental Circuit Elements

Four basic circuit variables:

- **Voltage** $v(t)$
- **Current** $i(t)$
- **Charge** $q(t) \triangleq \int_{-\infty}^{t} i(\tau)\, d\tau$
- **Flux** $\varphi(t) \triangleq \int_{-\infty}^{t} v(\tau)\, d\tau$

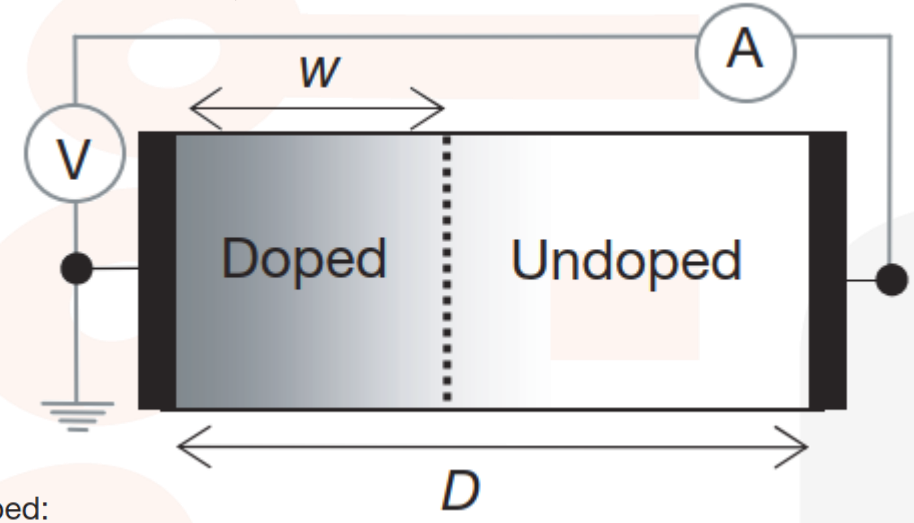


- **The "memristor" – the missing fourth element?**

L.O. Chua, "Memristor – The Missing Circuit Element," IEEE Trans., 1971

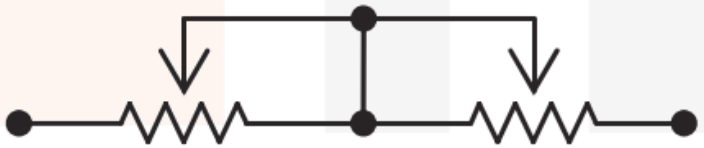
The Missing Memristor Found!

- **HP Labs, 2008**



Undoped:

Doped:

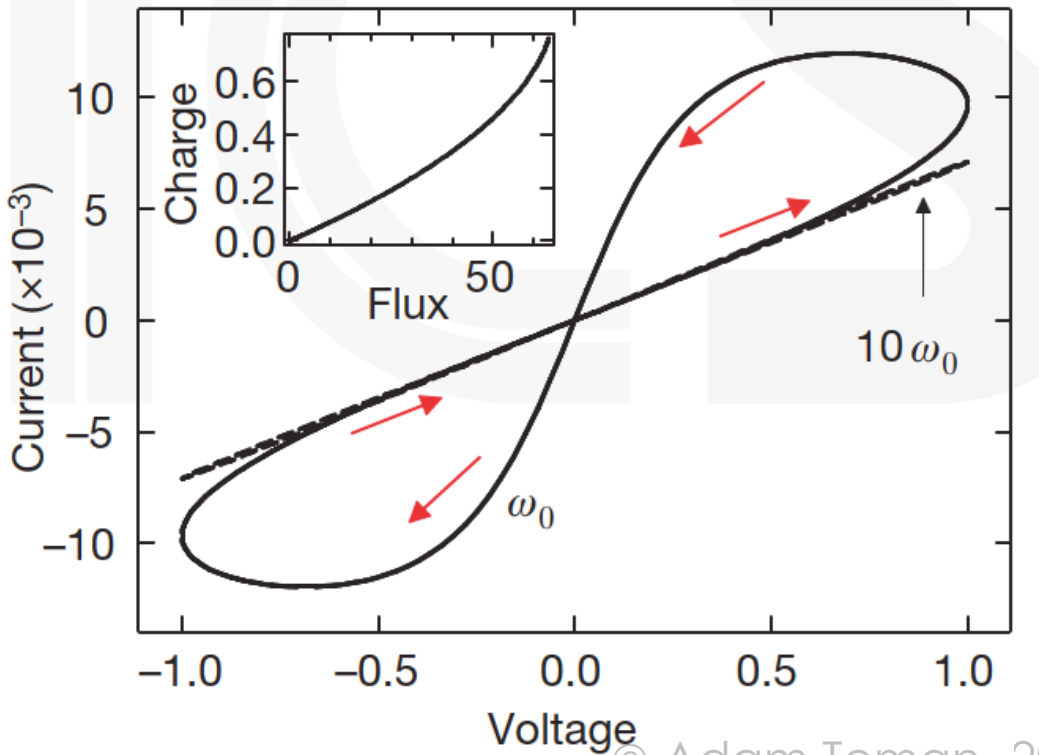


$$M(q) = R_{OFF}\left(1 - \frac{\mu_v R_{ON}}{D^2} q(t)\right)$$


nature
Vol 453|1 May 2008|doi:10.1038/nature06932
LETTERS

The missing memristor found

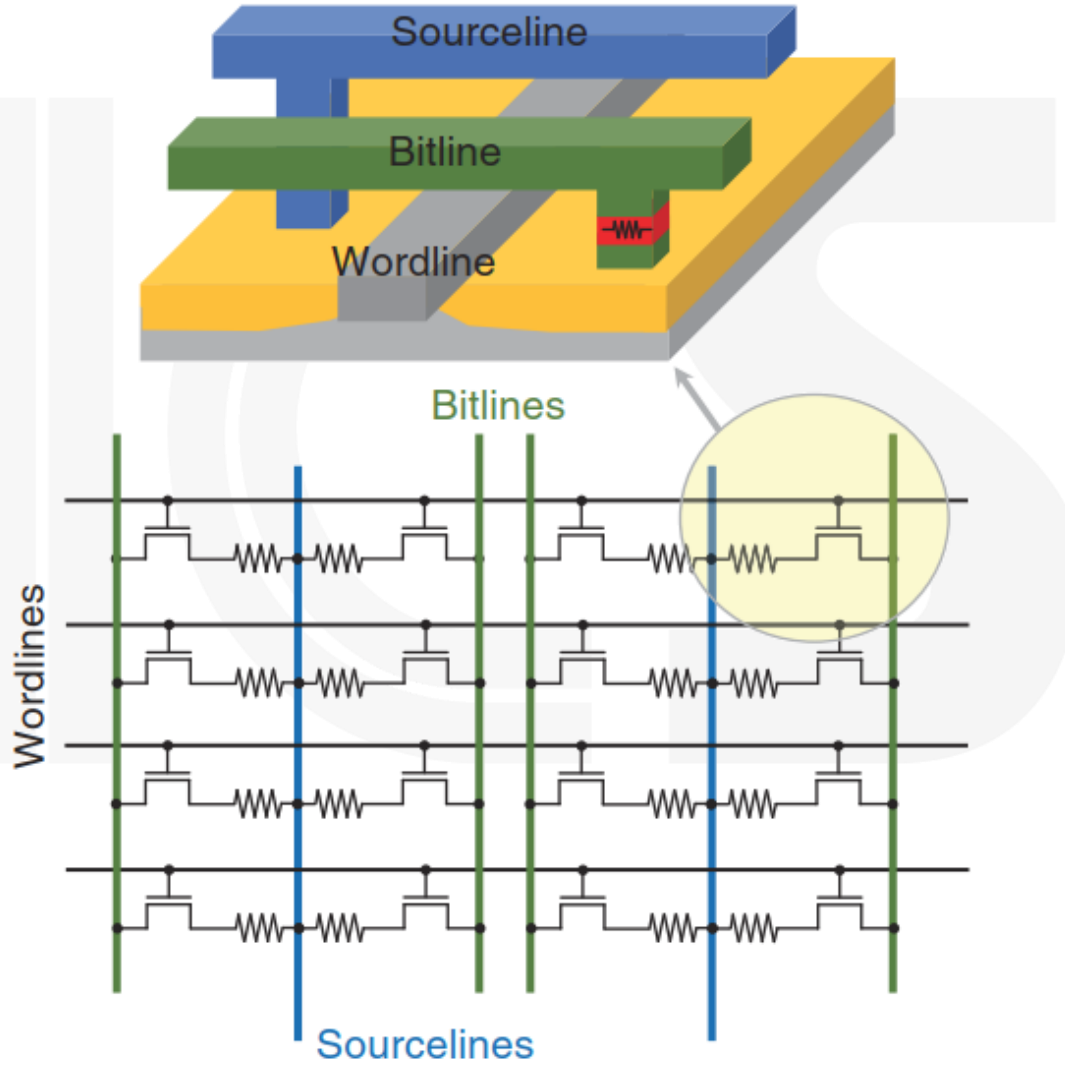
Dmitri B. Strukov[1], Gregory S. Snider[1], Duncan R. Stewart[1] & R. Stanley Williams[1]



D.B. Strukov et al, "The missing memristor found," Nature, 2008

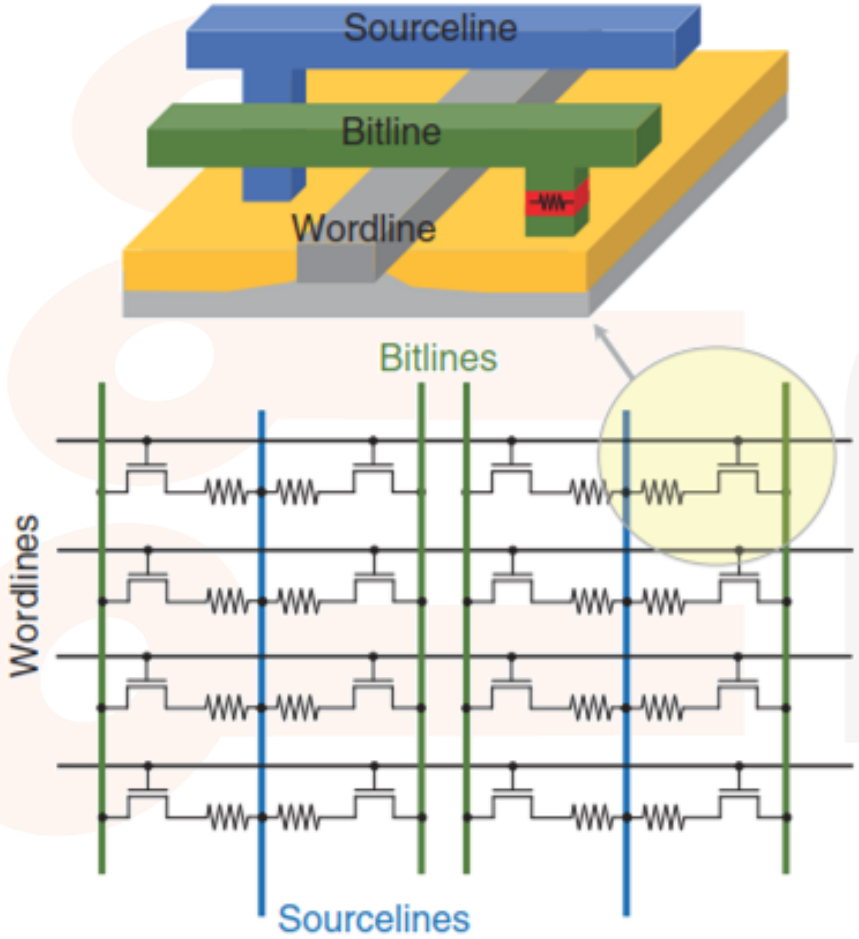
Charge vs. Resistive memories

- **Charge Memory (e.g., DRAM)**
  - Write data by capturing charge Q
  - Read data by detecting voltage V
- **Resistive Memory (e.g., PCM, STT-MRAM, memristors)**
  - Write data by pulsing current dQ/dt
  - Read data by detecting resistance R
  - Often constructed as a resistive crossbar.



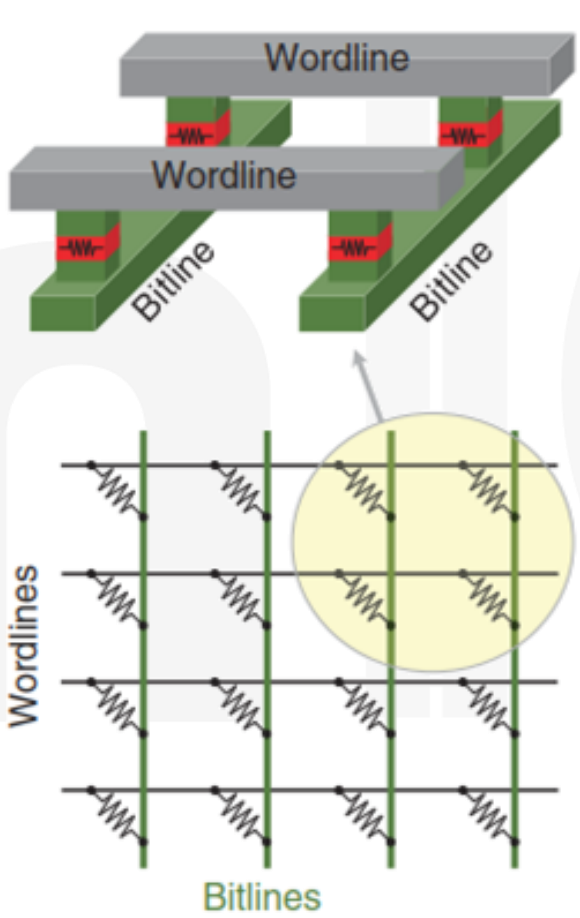
Source: Teman, et al., Wiley 2022

Resistive Memory Array Architectures



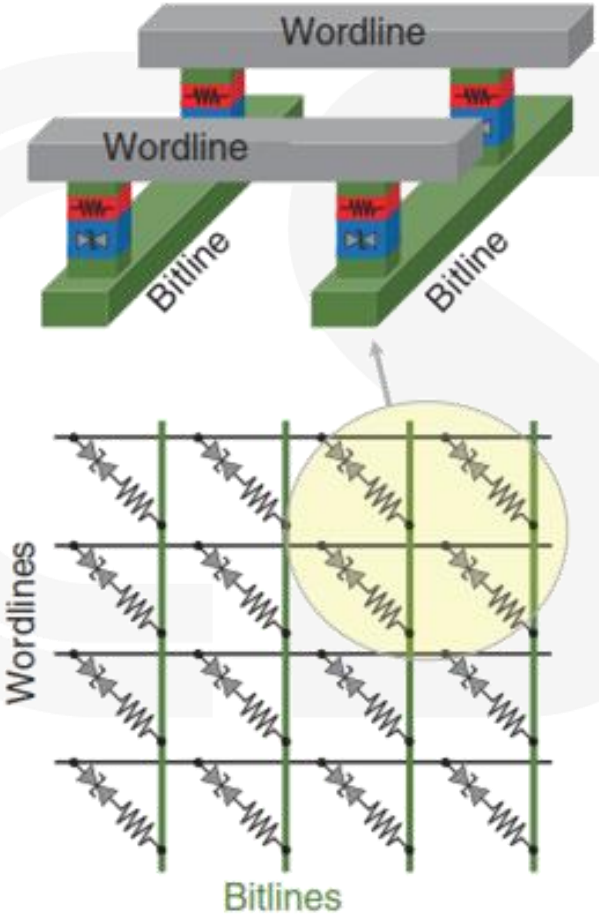
(a) 1-Transistor 1-resistor (1T1R)

> 8F²



(b) Cross-point or crossbar (1R)

4F²

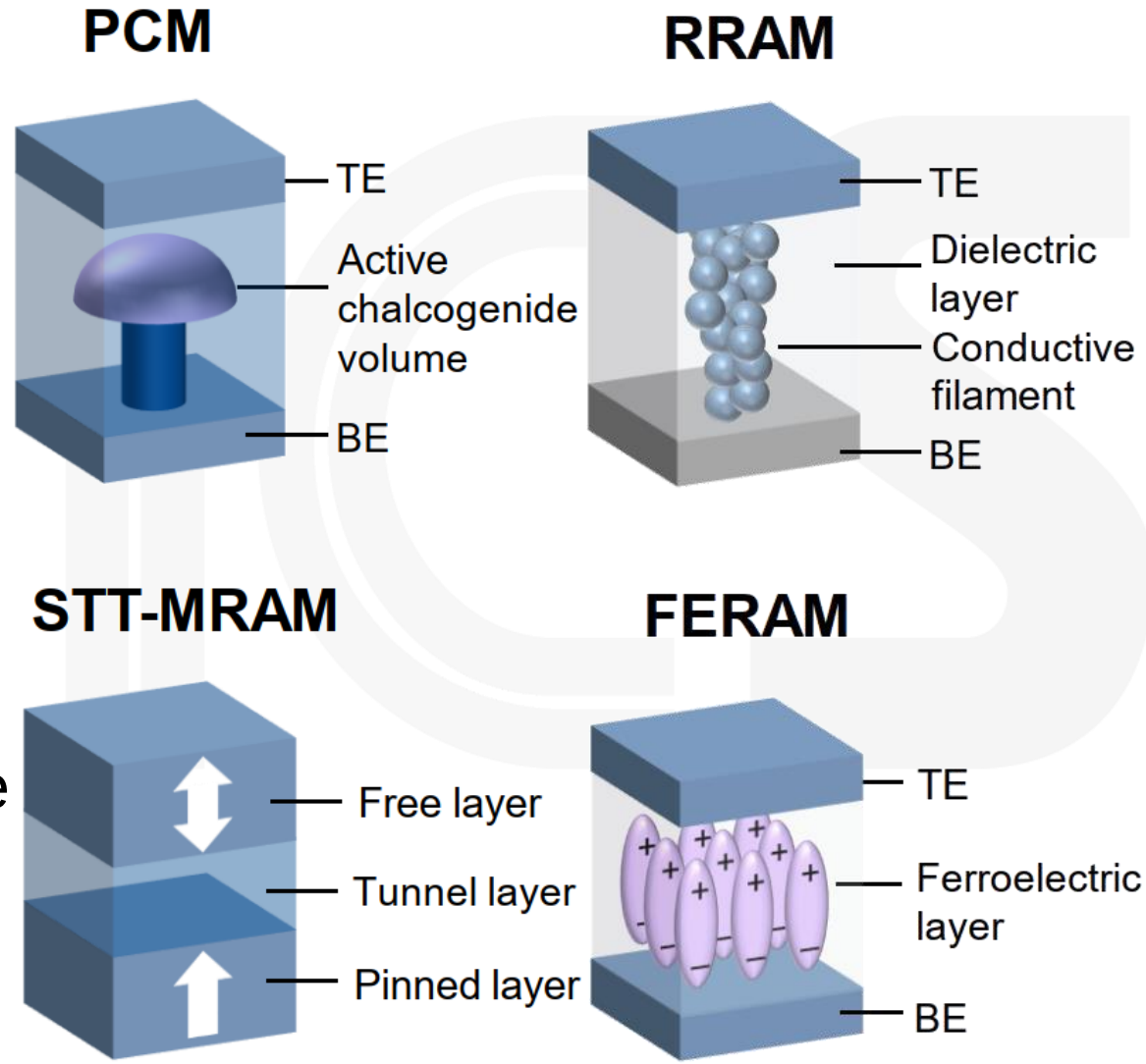


(c) 1-Transistor 1-selector (1T1S)

Source: Teman, et al., Wiley 2022

Emerging Nonvolatile Memory Technologies

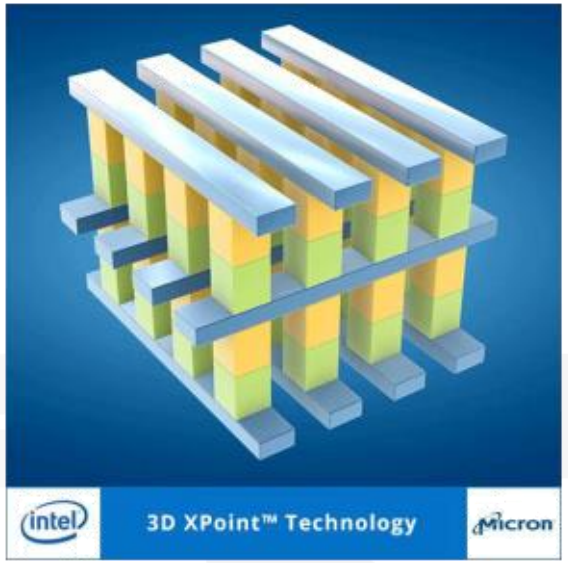
- **PCM**
  - Inject current to change material phase
  - Resistance determined by phase
- **STT-MRAM**
  - Inject current to change magnet polarity
  - Resistance determined by polarity
- **Memristors/RRAM/ReRAM**
  - Inject current to change atomic structure
  - Resistance determined by atom distance
- **FeRAM**
  - Utilize the ferroelectric effect
  - Resistance determined by dipole

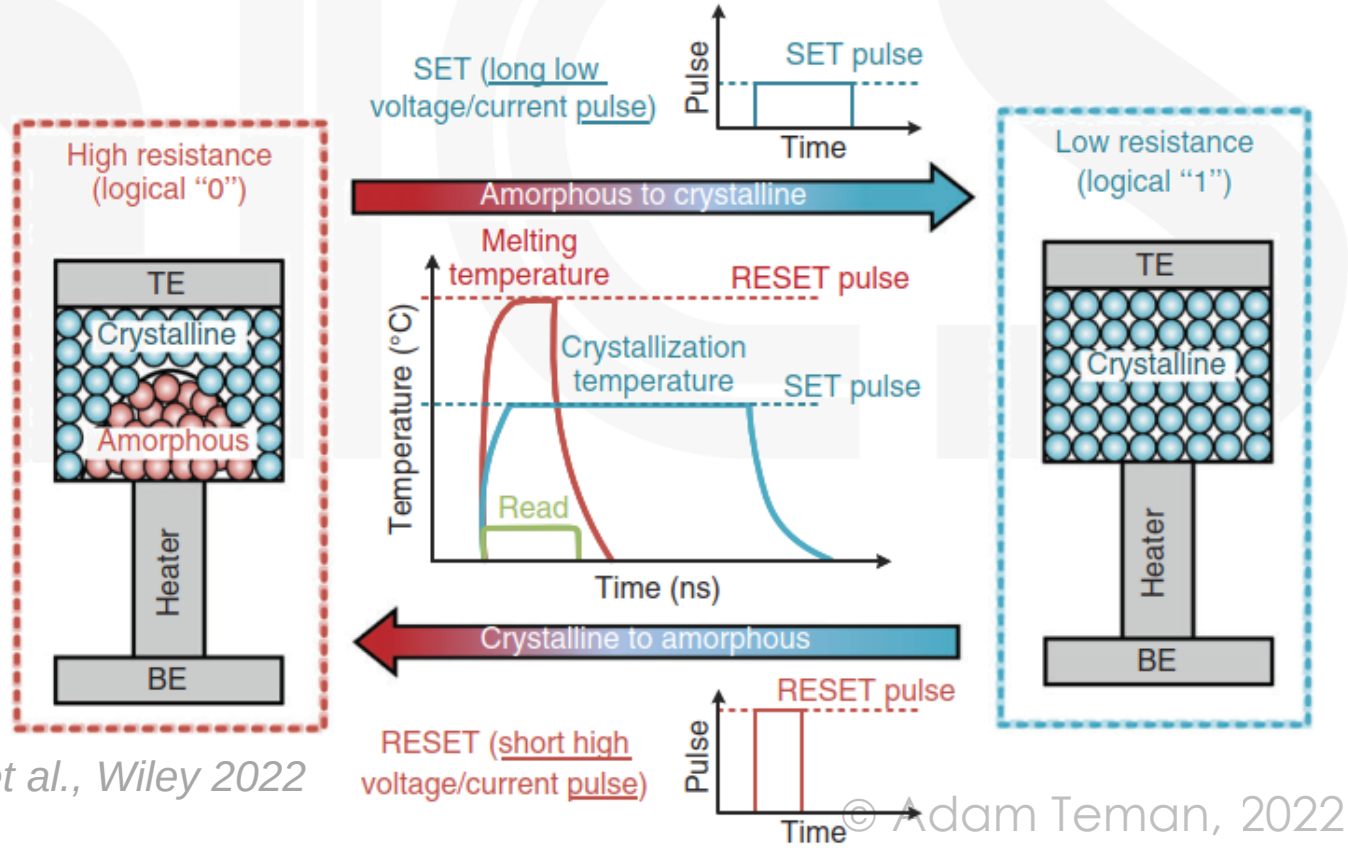


D. Ielmini and G. Pedretti, Adv. Intell. Syst. 2000040 (2020)

Phase Change Memory (PCM)

- **Chalcogenide glass (e.g., CD-ROM) exists in two states:**
  - Amorphous: High Resistance
  - Crystalline: Low Resistance
- **Advantages**
  - Scalable (sub 10nm)
  - Dense (4F^2, MLC, 3D Stacking)
  - Non-Volatile
  - Read/Write Performance (between DRAM and Flash)
- **Current availability:**
  - Intel Optane (3D Xpoint)
  - STMicro, Samsung

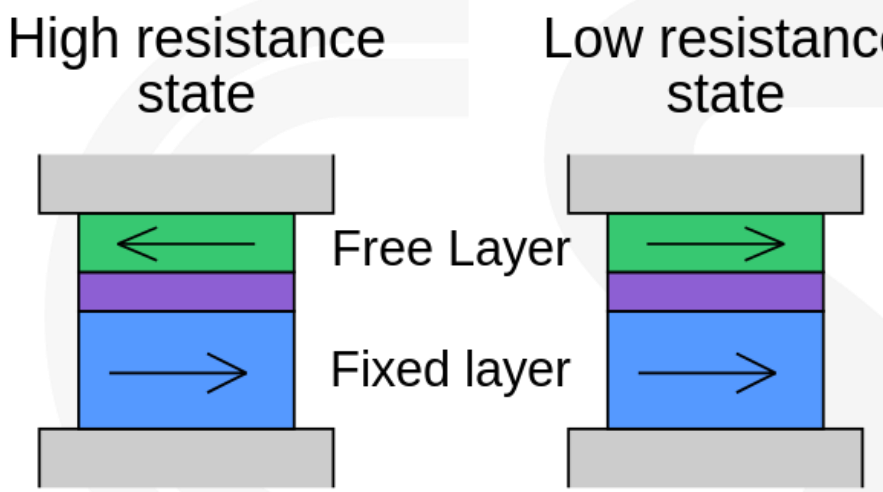




Source: Teman, et al., Wiley 2022

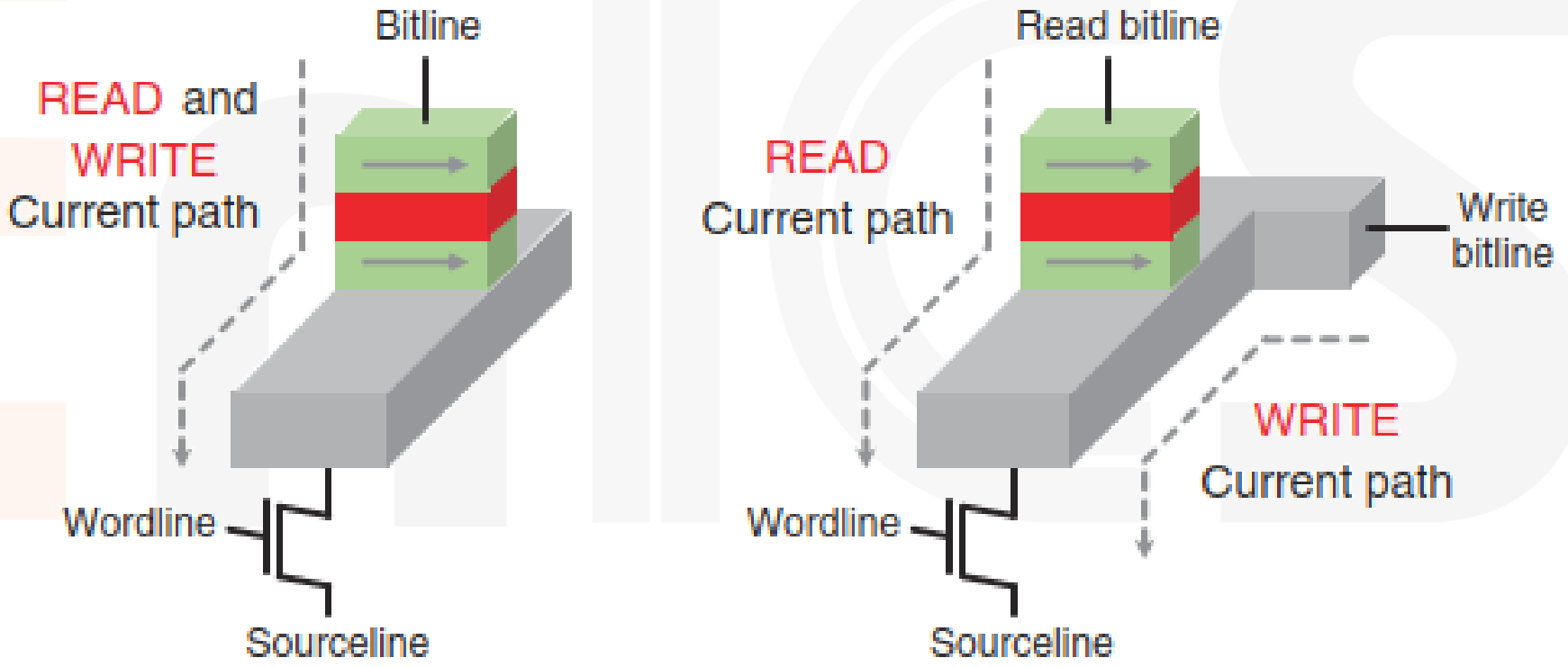
Magnetoresistance RAM (STT MRAM)

- **Resistance controlled by spin across Magnetic Tunnel Junction (MTJ)**
  - Parallel spin state: Low Resistance
  - Anti-parallel spin state: High Resistance
- **Advantages**
  - Scalable (sub 10nm)
  - Non-Volatile
  - Read/Write Performance (comparable to DRAM)
  - Endurance/Retention
- **Current availability:**
  - Everspin (1Gb, 28nm, standalone)
  - eMRAM: TSMC, GF, Samsung, Intel, UMC
  - Avalanche, Renesas, IBM



Spin-Transfer Orbit MRAM (SOT-MRAM)

- **Decouple read and write operations of MRAM**
  - Higher write speeds
  - Higher endurance
  - SRAM Competitor (Last-layer cache replacement)

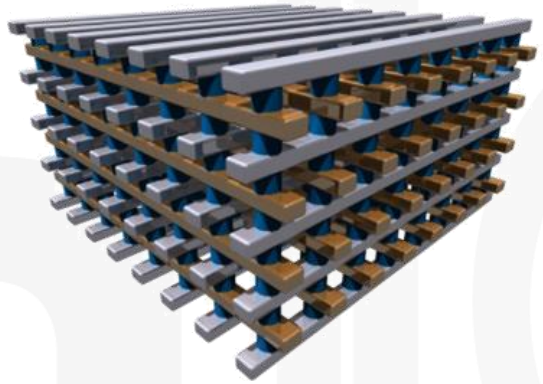


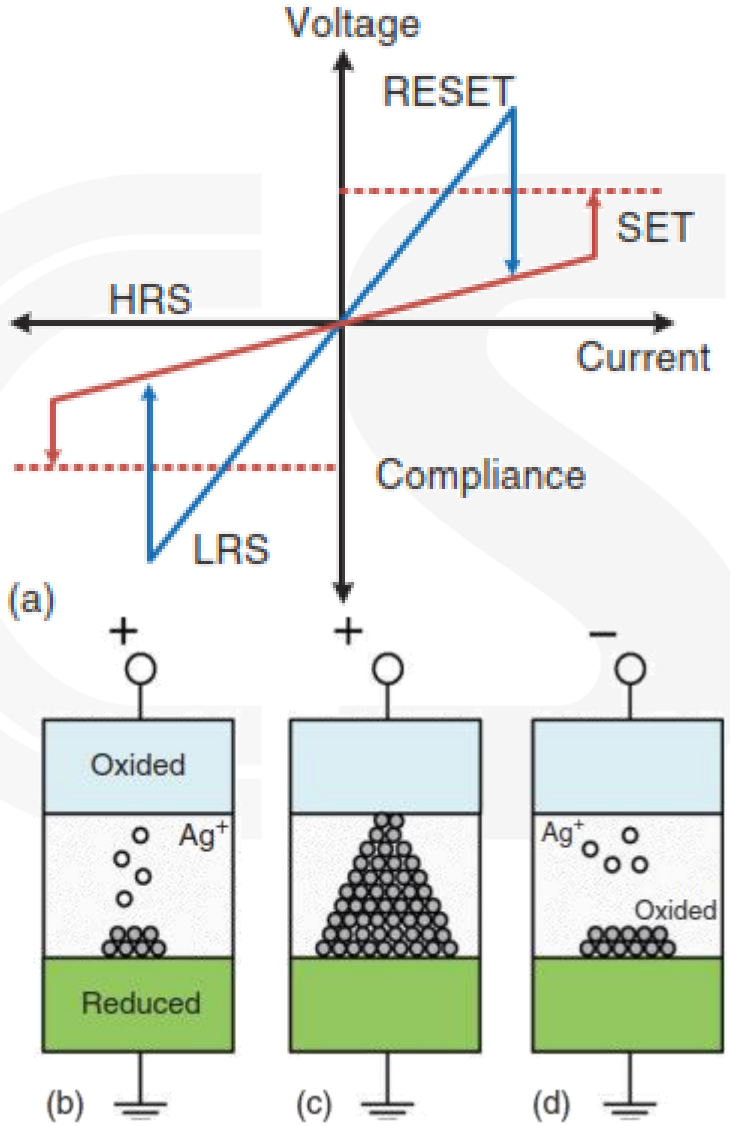
(a) STT-MRAM bitcell (b) SOT-MRAM bitcell

Source: Teman, et al., Wiley 2022

Resistive RAM (ReRAM/RRAM)

- **Change resistance by creating a conductive filament:**
  - Conductive Bridge (CBRAM)
  - Oxygen Vacancy (OxRAM)
- **Advantages**
  - Scalable (sub 10nm)
  - Dense (4F^2, MLC, 3D Stacking)
  - Non-Volatile
  - Read/Write Performance (between DRAM and Flash)
- **Current availability:**
  - Adesto (now Dialog) CBRAM licensed to GF
  - Cerfe Labs (ARM) CeRAM licensed from Symetrix
  - eReRAM offered by TSMC, GlobalFoundries
  - Mitsubishi, Fujitsu, Panasonic Winbond
  - WeeBit-Nano – promising Israeli start-up

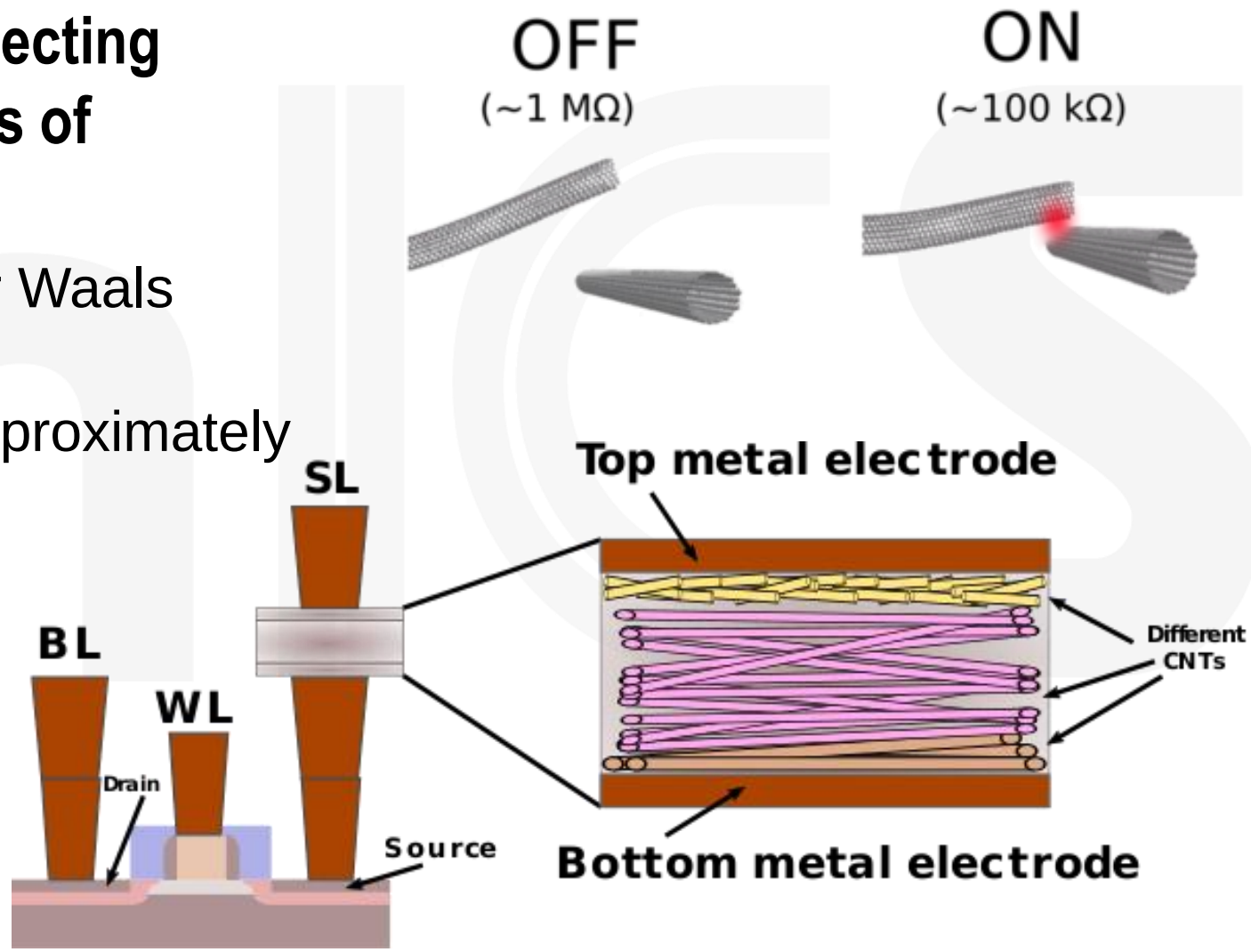




Source: Teman, et al., Wiley 2022

Carbon-Nanotube RAM (CNT-RAM/NRAM)

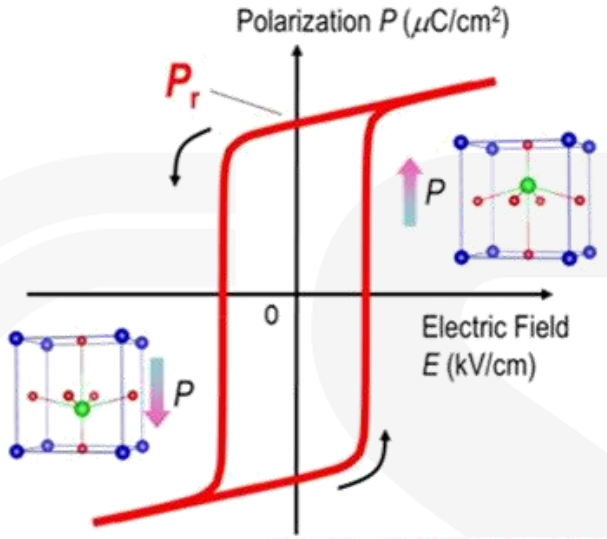
- **Create conductive bridge by connecting or disconnecting stochastic arrays of carbon nanotubes (CNTs)**
  - Held together by strong Van der Waals binding forces.
  - 5nm bitcell expected to have approximately 1000 switchable CNT junctions
  - Also demonstrated 3D Stacking
- **Commercialized by Nantero**

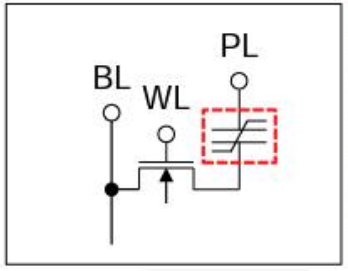


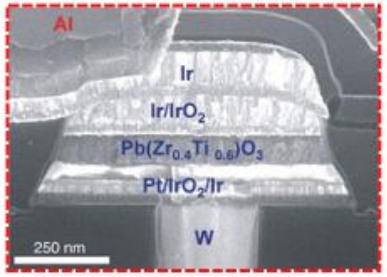
Source: Wikichip, Nantero

Ferroelectric Memory (FeRAM, FeFET)

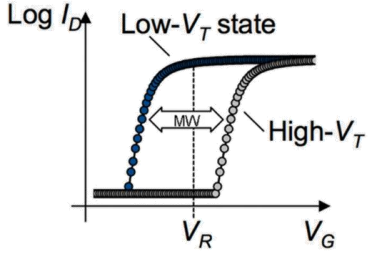
- **Ferroelectricity:**
  - A material with an electric polarization that can be reversed by application of an external electric field.
- **FRAM:**
  - A very old technology (invented 1952)
  - Classic bitcell: PZT-based 1T-1C (130nm)
- **FeFET:**
  - Based on ferroelectric HfO_2 (discovered in 2007)
  - Add HfO2 layer → 1T bitcell
  - Being developed by FMC

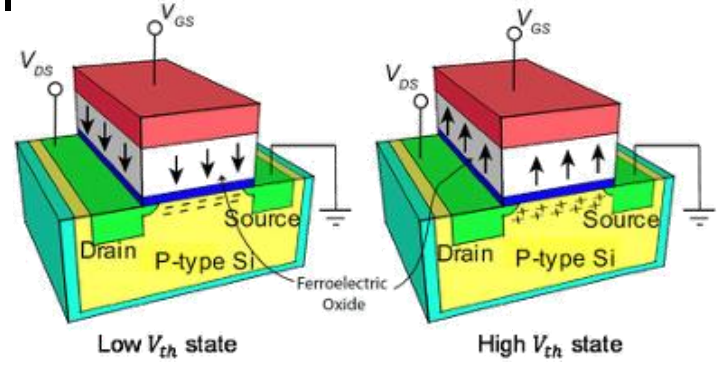


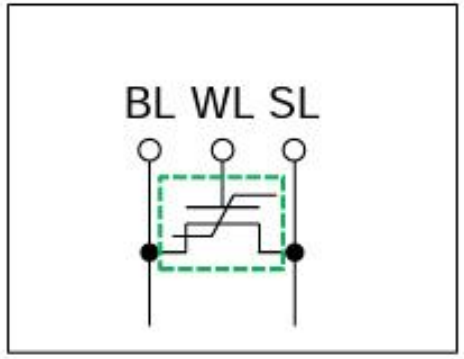




Source: Science, Vol. 315, pp. 954-959







Source: S. Mueller, EMF, 2015.

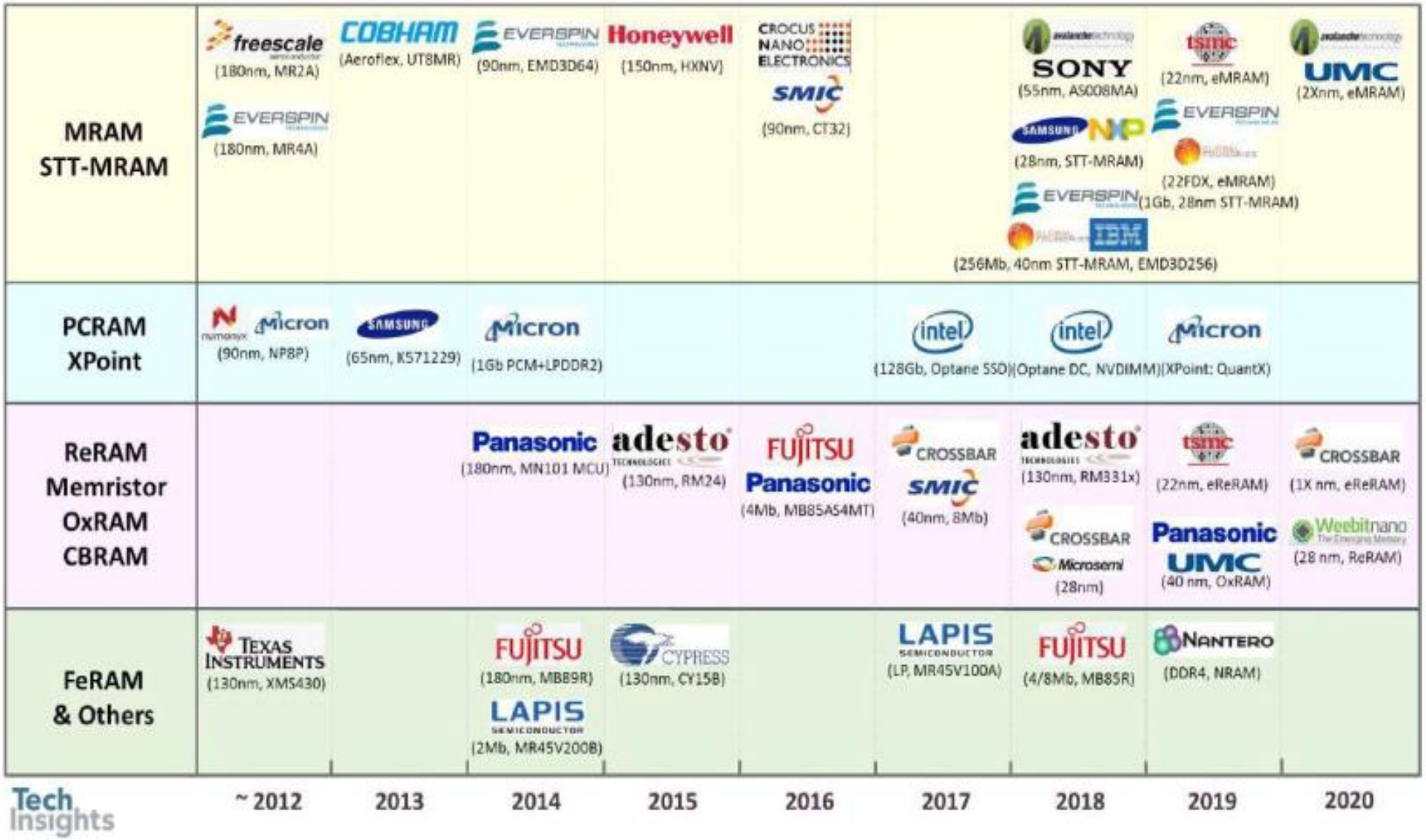
Source: FMC

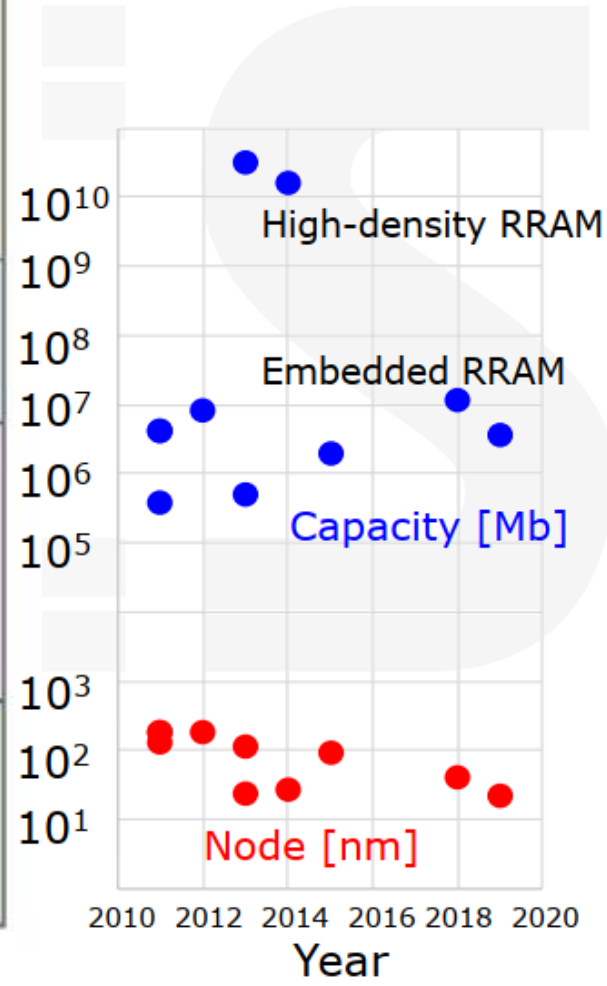
Comparison between technologies

	SRAM	DRAM	NAND Flash	NOR Flash	PCM	RRAM	STT-MRAM	SOT-MRAM	FeRAM	FeFET
Circuit	6T	1T1C	1T	1T	1T1R or 1D1R	1T1R or 1D1R	1T1MTJ	2T1MTJ	1T1C	1T
Cell Size	>150 F^2	6 F^2	<4 F^2 (3D)	10 F^2	4-12 F^2	4-10 F^2	6-50 F^2	12-100 F^2	6-50 F^2	6-50 F^2
Voltage	<1V	<1V	>10V	>10V	<3V	<3V	<1V	<1V	<2V	<3V
Stackable	No	Yes	Yes	No	Yes	Yes	No	No	No	No
MLC (bits/cell)	No	No	Yes (4)	Yes (2)	Yes (2)	Yes	No	No	No	Yes
Scaling	< 3 nm	~10nm	~14nm	~45nm	<10nm	<10nm	<10nm	<10nm	<10nm	<10nm
Read Latency	~1ns	~10ns	~10us	~50ns	<50ns	<50ns	~10ns	<10ns	<100ns	<50ns
Write Latency	~1ns	~10ns	>100us	>100us	<100ns	<100ns	<20ns	<3ns	<100ns	<100ns
Write Energy	~fJ	~10fJ	~10fJ	~100pJ	~10pJ	~pJ	~pJ	~pJ	~100fJ	~fJ
Endurance	> 10^{16}	>10^{16}	10^4-10^5	10^5	10^6-10^8	10^6-10^{11}	10^6-10^{15}	~10^{12}	10^9-10^{12}	10^5-10^9
Retention	Volatile	Volatile	>10yr	>10yr	>10yr	>10yr	>10yr	>10yr	>10yr	>10yr

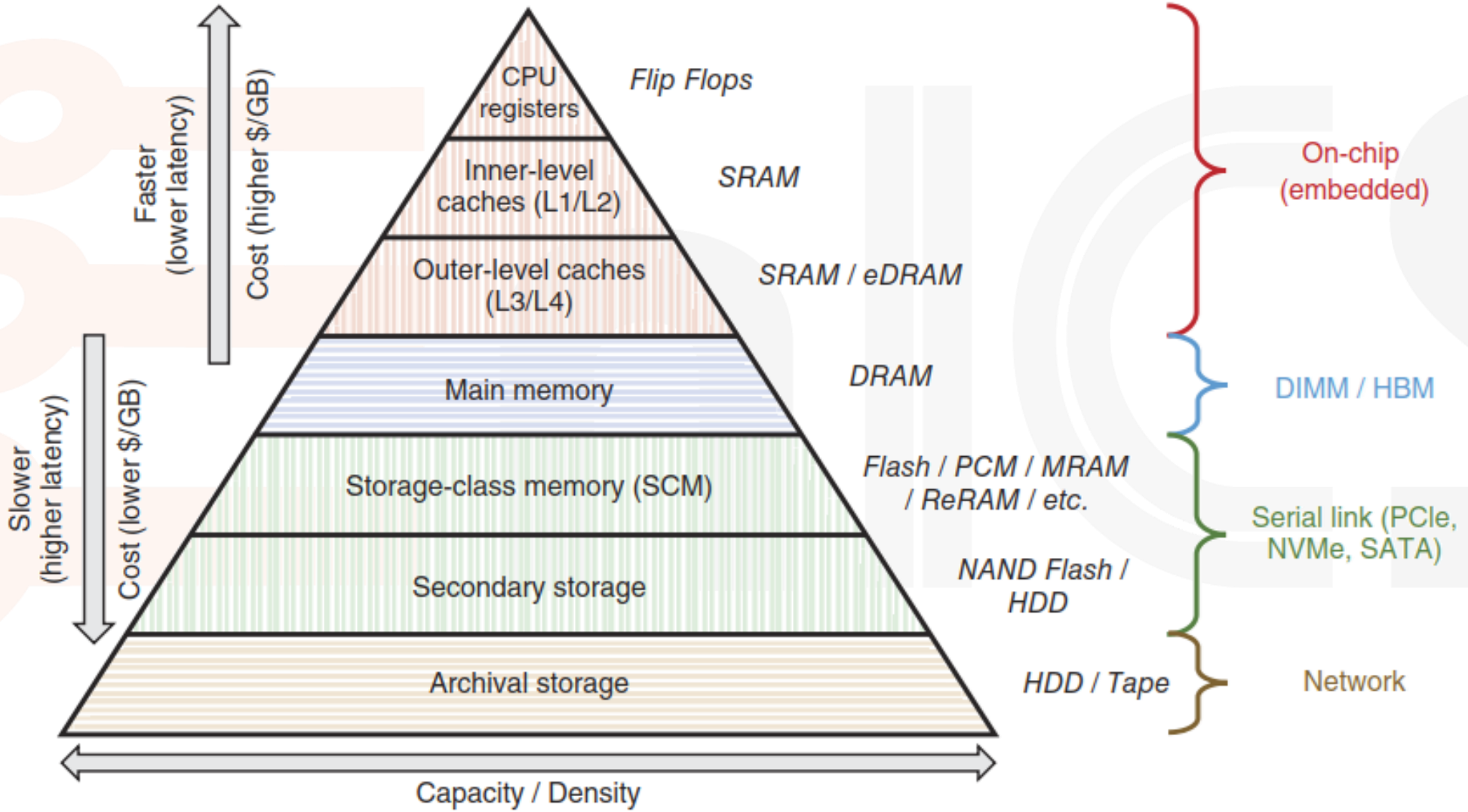
Source: Teman, et al., Wiley 2022

RRAM demos





Summary



Further Reading

- Rabaey, et al. “Digital Integrated Circuits” (2nd Edition)
- Elad Alon, Berkeley ee141 (online)
- Weste, Harris, “CMOS VLSI Design (4th Edition)” S. Kvatinsky “Memristors: Not Just Memory”, ChipEx 2013
- S. Kvatinsky “Emerging Non-volatile Memories: Opportunities and Challenges”, DevelopEx 2015
- Samira Khan, U. Virginia CS6354
- Daniele Ielmini “Resistive switching memory for in-memory computing applications”, ESSCIRC Tutorial 2020