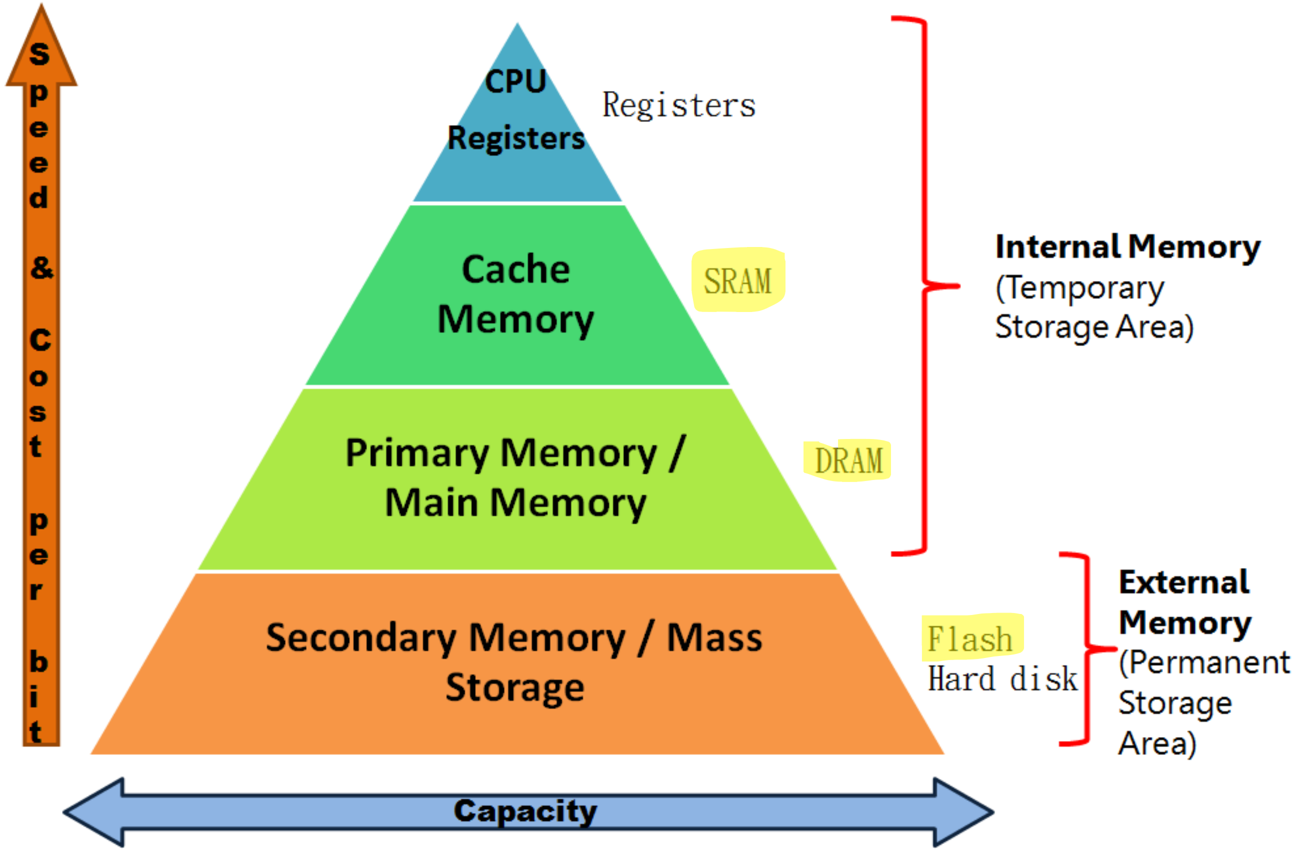
EITP25 2020

Lecture 2 – Current Memory Technologies

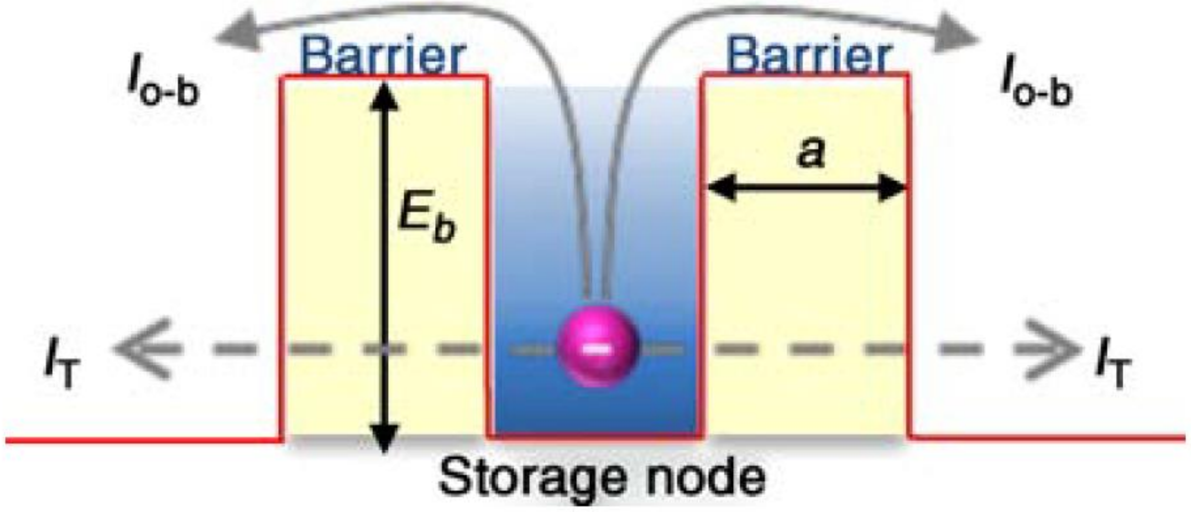
Outline

- Physics of charge-based memory
- DRAM
- Flash memory
- SRAM


Speed & Cost per bit
CPU Registers
Registers
Cache Memory
SRAM
Primary Memory / Main Memory
DRAM
Secondary Memory / Mass Storage
Flash
Hard disk
Internal Memory (Temporary Storage Area)
External Memory (Permanent Storage Area)
Capacity

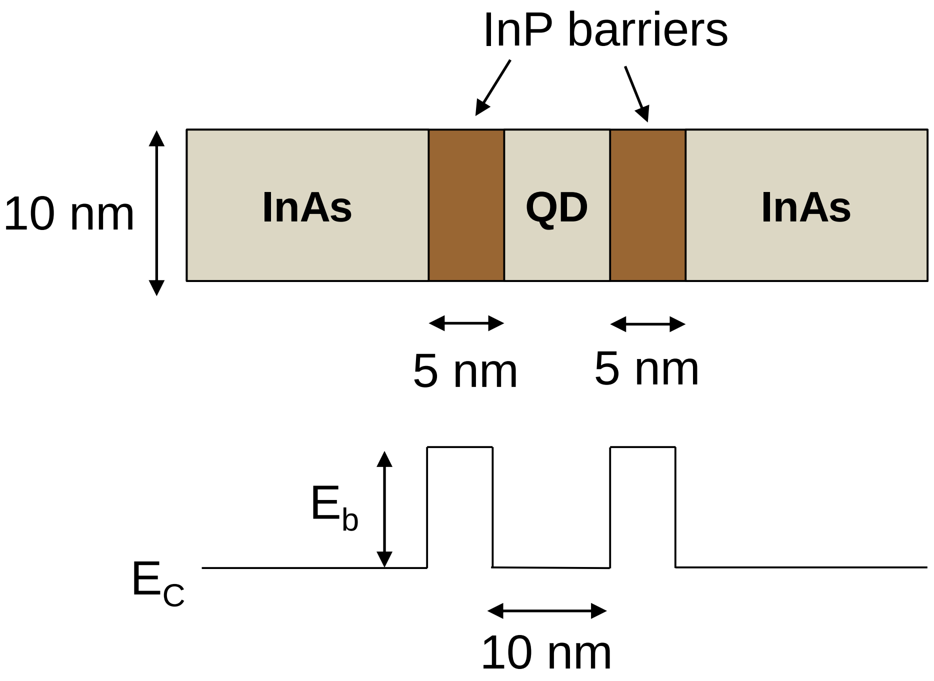
This Lecture is based on Emerging Nanoelectronic Devices Chapter 3

Charge-based storage



Emerging Nanoelectronic Devices chapter 3

2 min exercise – QD as 10 year storage?

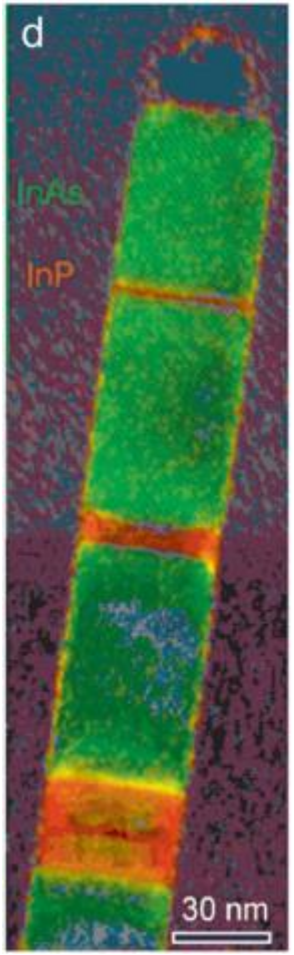


$m_e = 0.08m_0$
$E_b = 0.75$ eV
$L^2 = 100$ nm^2
kT = 1/40 eV

What is the emission retention time?

Emission: $t_r = \frac{1}{L^2 f_0} \exp\left(\frac{E_b}{kT}\right)$

(Tunneling: $t_T = \frac{1}{L^2 f_0^*} \exp\left(\frac{2\sqrt{2m_e E_b}}{\hbar} a\right)$)



Björk et al. Nano Lett. 2002

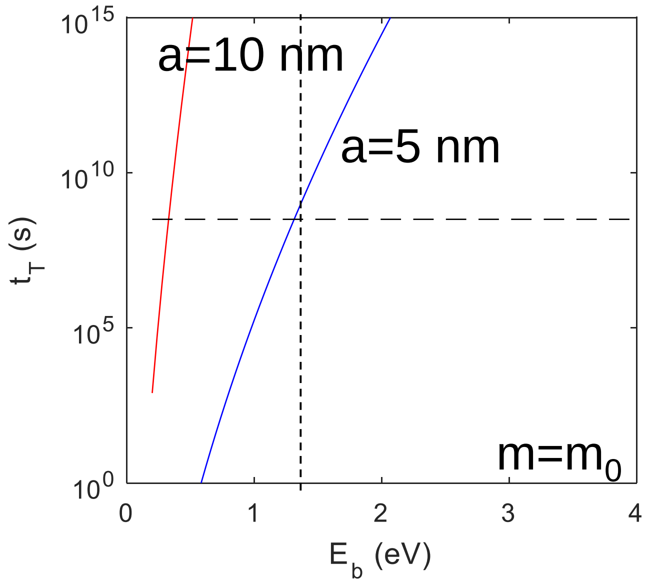
What kind of barrier is needed?

- For emission E_b > 1.4 eV for t_r > 10 years.
- Assuming $\underline{t_T}$ < t_r → a > 5 nm for oxide barrier

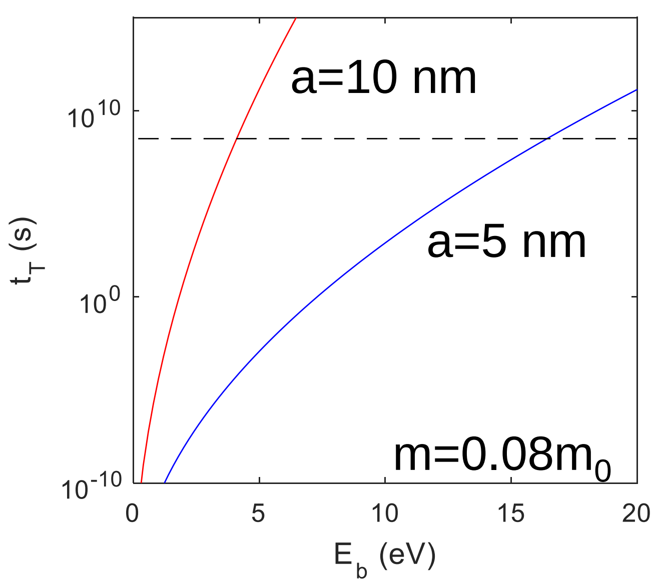
$$t_T = \frac{1}{L^2 f_0^*} \exp\left(\frac{2\sqrt{2m_e E_b}}{\hbar} a\right)$$

L = 10 nm

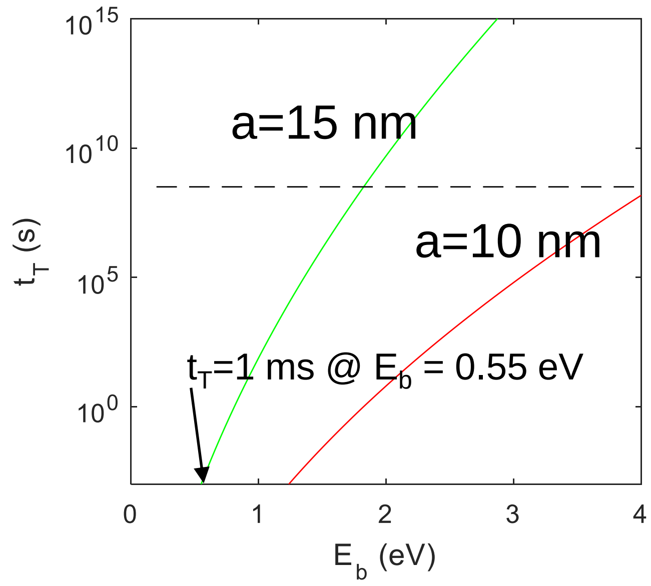
<u>”Oxide” barriers</u>



<u>InP barriers</u>

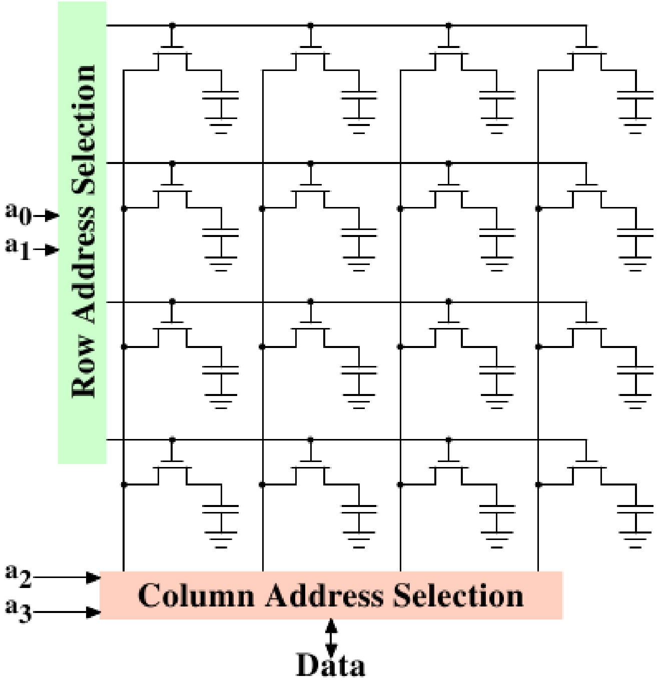


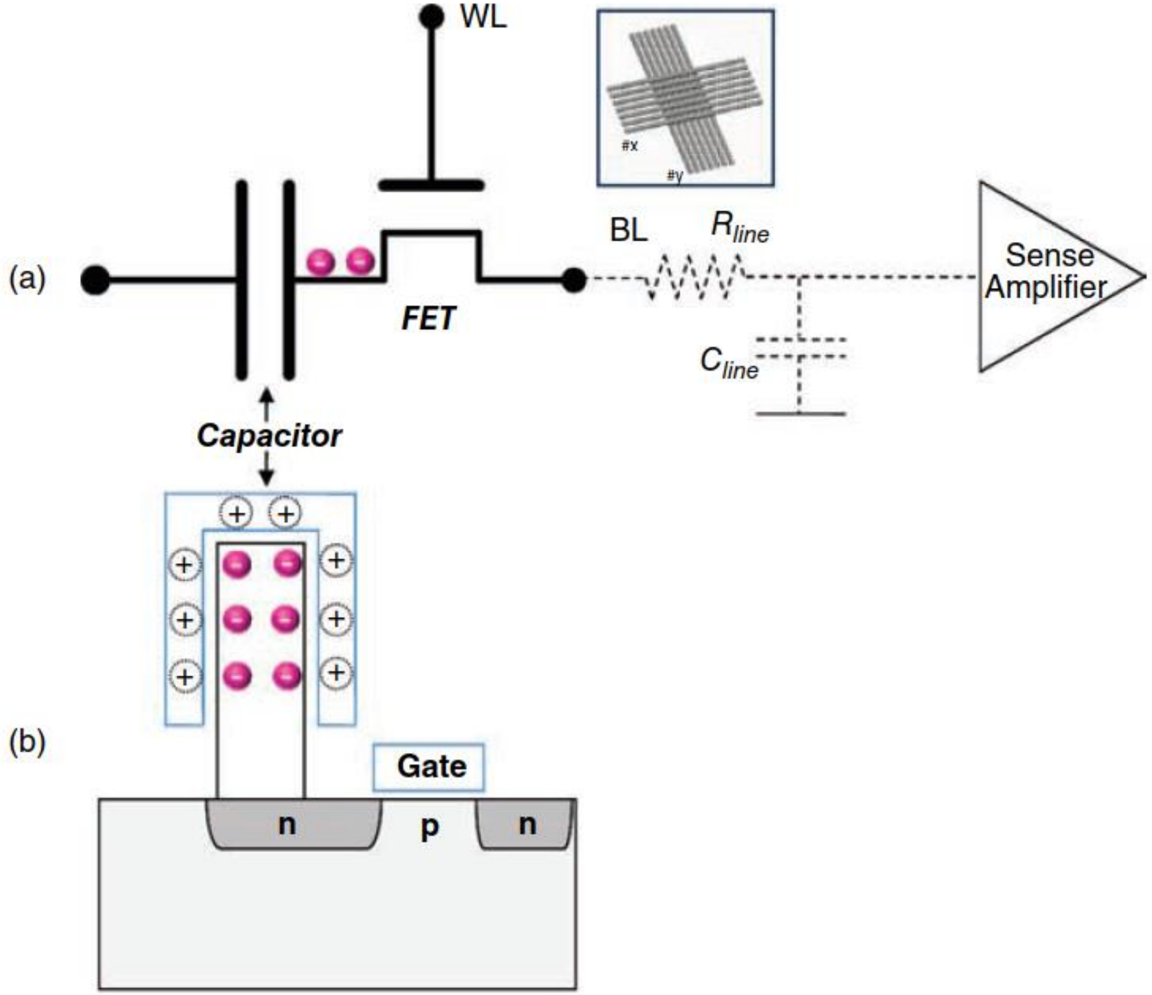
→ Hard to achieve 10 year retention with semiconductor barriers…but ms?



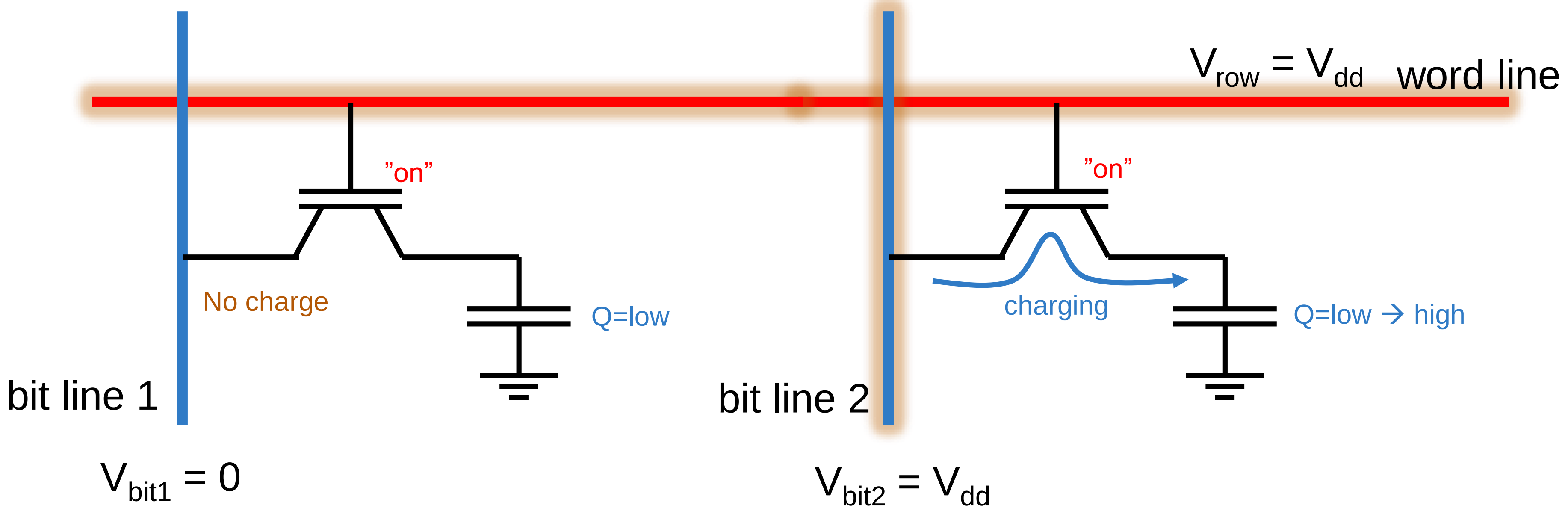
Dynamic Random Access Memory

- 1 transistor + 1 capacitor (1T1C)
- Charge stored on capacitor → memory bit
- Retention time limited by Si band gap (1.1 eV)
  ~ milliseconds



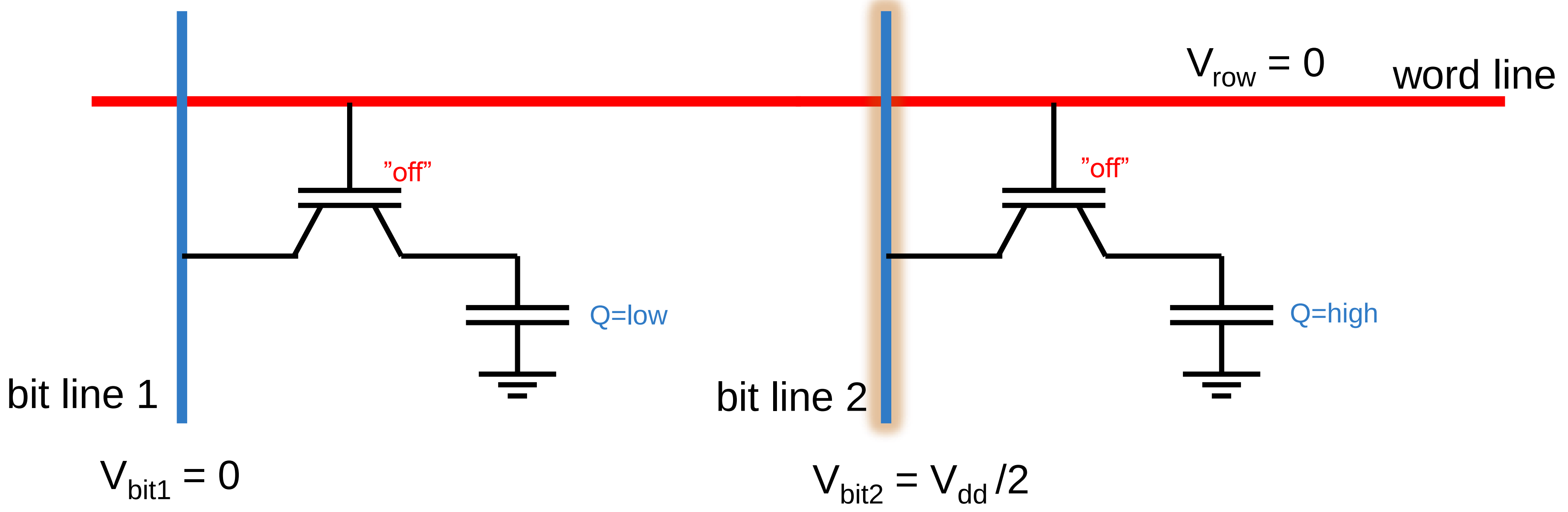


DRAM write a ”1”



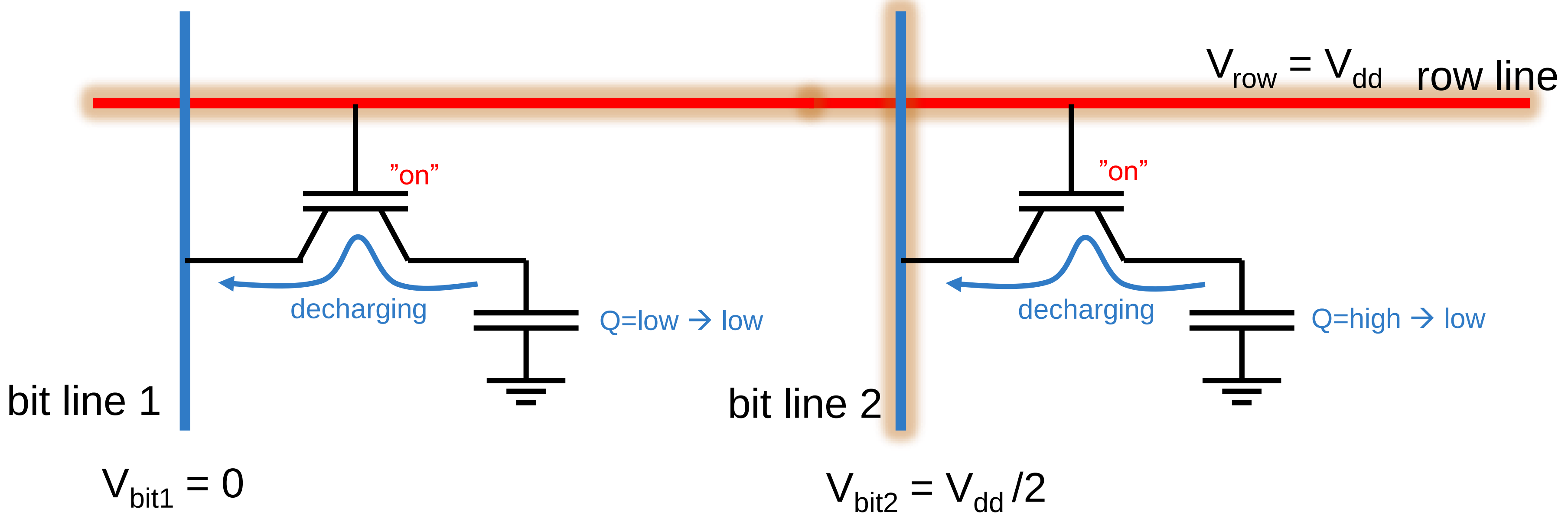
1. Drive bit line to V_{dd}
2. Select word line
3. Capacitor is charged and the state is saved.

DRAM read a bit – step 1



1. Precharge bit line to V_{dd}/2
   - Reduces read swing

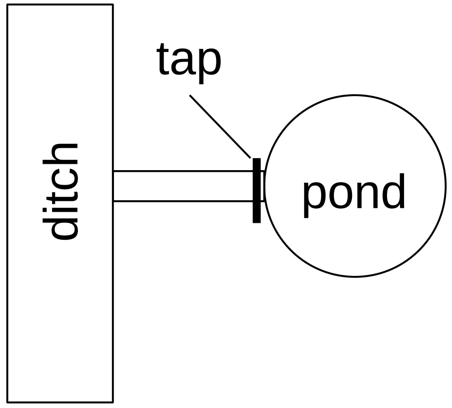
DRAM read a bit – step 2



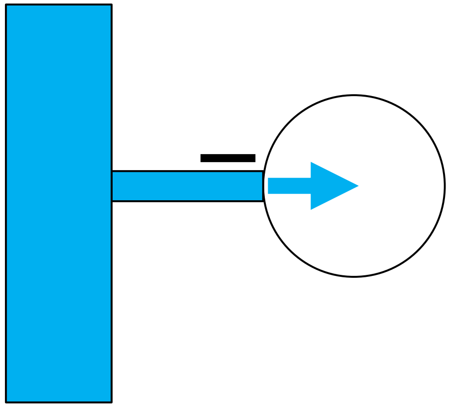
1. Precharge bit line to $V_{dd}/2$
2. Select the word line
   - Capacitors on whole row decharge (destructive read)
3. Finish by re-writing data on row

Water pond model

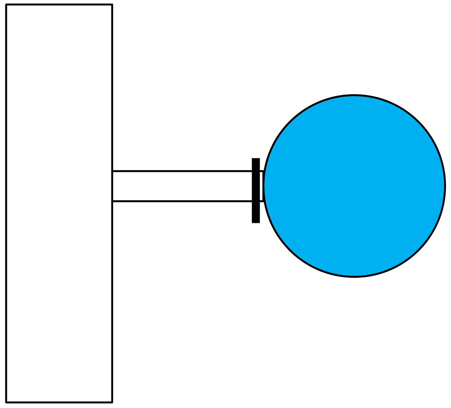
- (1) Fill the ditch and open the tap to fill pond.
- (2) Close the tap to store the water.
- (3) Water is lost by evaporation and must be refilled.
- (4) Opening tap to measure if there is water empties pond



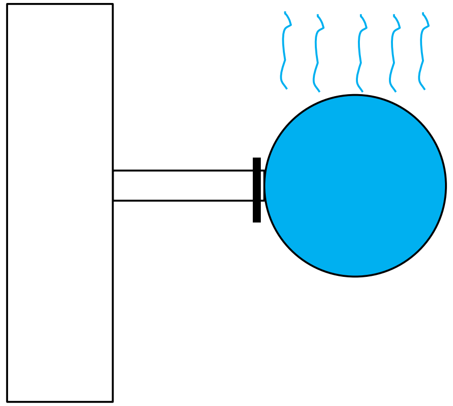
(1)



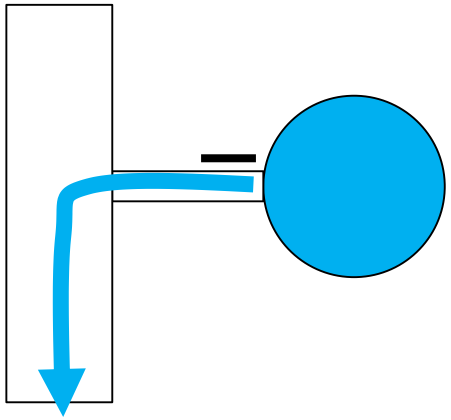
(2)



(3)



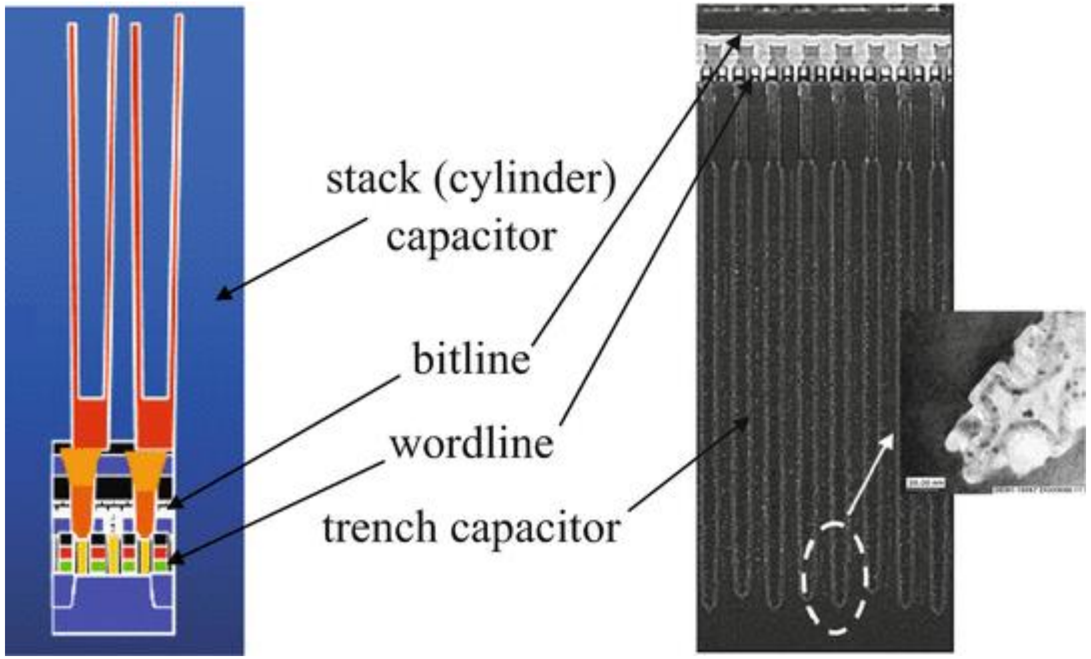
(4)



DRAM implementation

- Two main types of capacitor implementations:
  1) Stacked (above FET)
     - May interfere with metal routing
     - Done after logic
     - Two cells can share same BL → 6F^2 possible
  2) Trench (below FET)
     - Done prior to logic (must survive high T)
     - Hard to control depth by dry etching
     - No interference with interconnects
     - Not as scalable (8F^2)

Requires at least 25 fF capacitance for read-out
→ How high capacitors are needed?



Example – Capacitor height

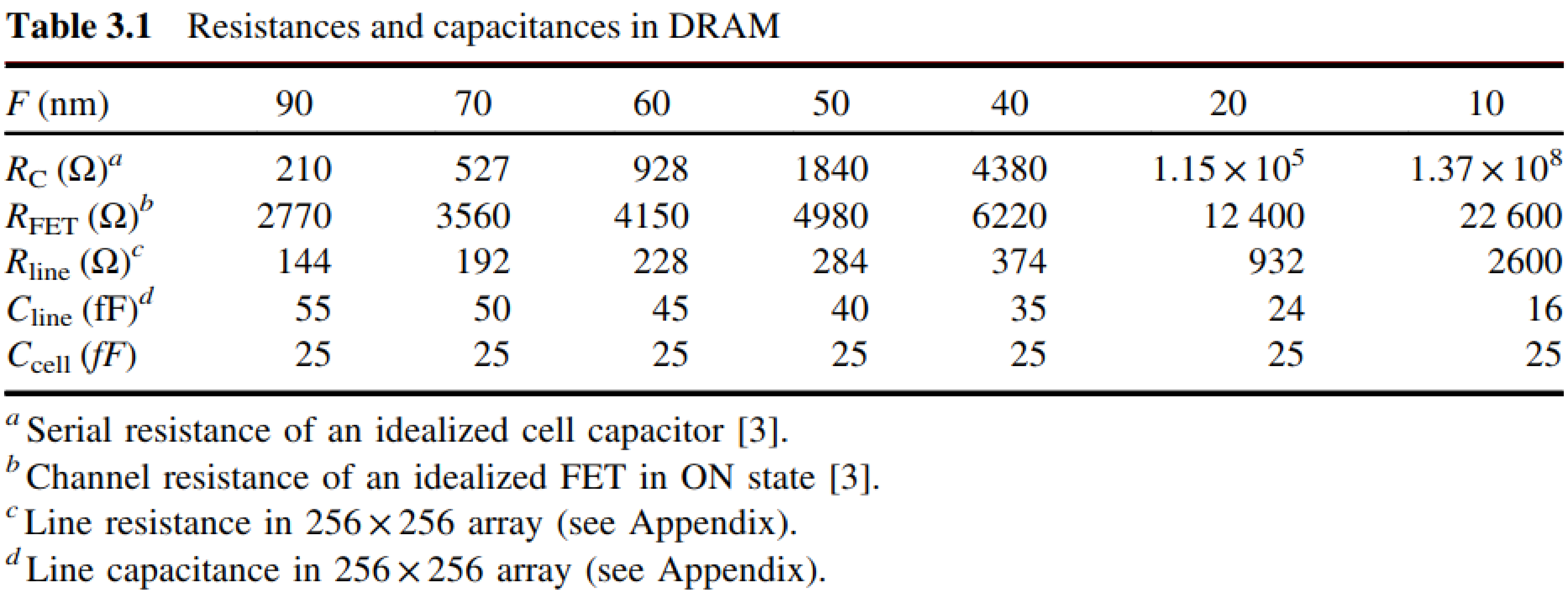


Table 3.1 Resistances and capacitances in DRAM

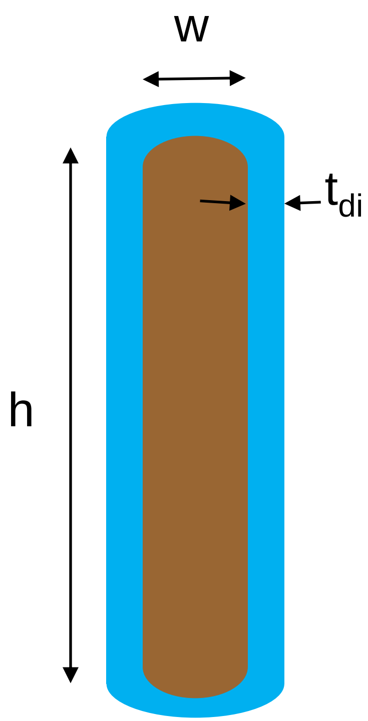
F (nm)	90	70	60	50	40	20	10
R_C (Ω)[a]	210	527	928	1840	4380	1.15×10^5	1.37×10^8
R_{FET} (Ω)[b]	2770	3560	4150	4980	6220	12 400	22 600
R_{line} (Ω)[c]	144	192	228	284	374	932	2600
C_{line} (fF)[d]	55	50	45	40	35	24	16
C_{cell} (fF)	25	25	25	25	25	25	25

[a] Serial resistance of an idealized cell capacitor [3].

[b] Channel resistance of an idealized FET in ON state [3].

[c] Line resistance in 256 × 256 array (see Appendix).

[d] Line capacitance in 256 × 256 array (see Appendix).



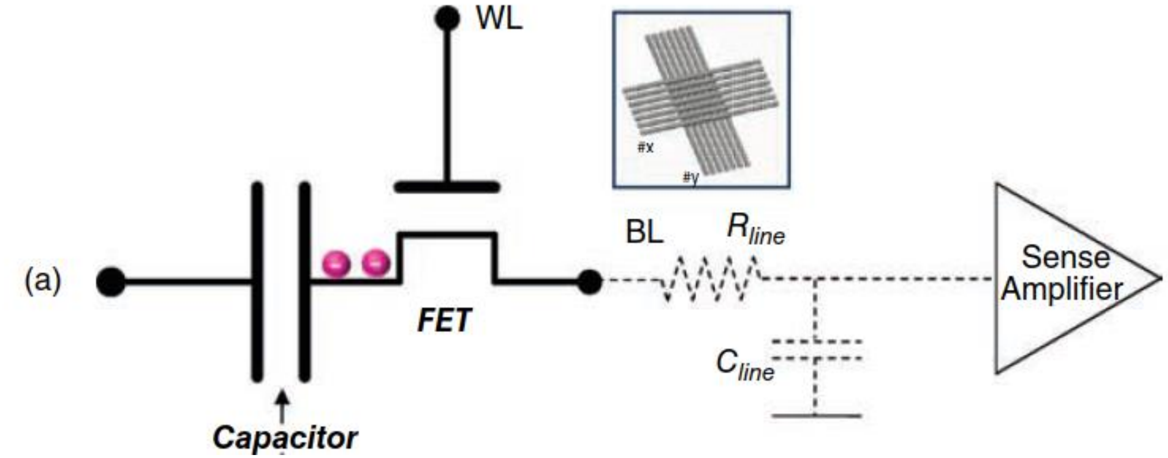
Energy usage of DRAM

- Write energy:

- $E_{DRAM} = \underbrace{(C_{line}+C_{cell})V_{cell}^2}_{} + \underbrace{C_{line}V_g^2}_{}$

”pump” N ~ 10^5 (25 fF) electrons through BL (C_{line}) to charge capacitor (C_{cell})

Energy for controlling the gate of FET via WL

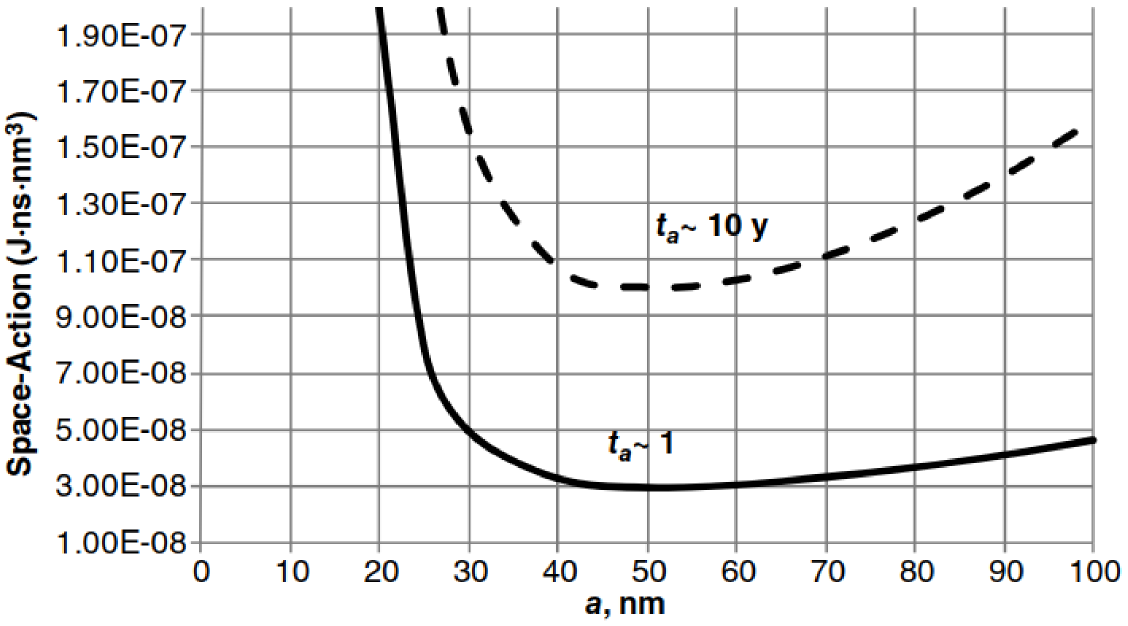


- Additional cost for refreshing memory (retention)
  - Average access interval $t_a \sim 1 - 10s$
  - Retention time, $t_r \sim 50 - 100\ ms$
- $E_{TOT} = E_{DRAM} + \frac{t_a}{t_r}E_{DRAM} = \left(1 + \frac{t_a}{t_r}\right)E_{DRAM}$ ~ <u>30-60 pJ</u>

DRAM access time

$$t_{DRAM} = (R_{cap} + R_{FET,on} + R_{line})(C_{cap} + C_{line})$$

Space-Action metric: energy x volume x access



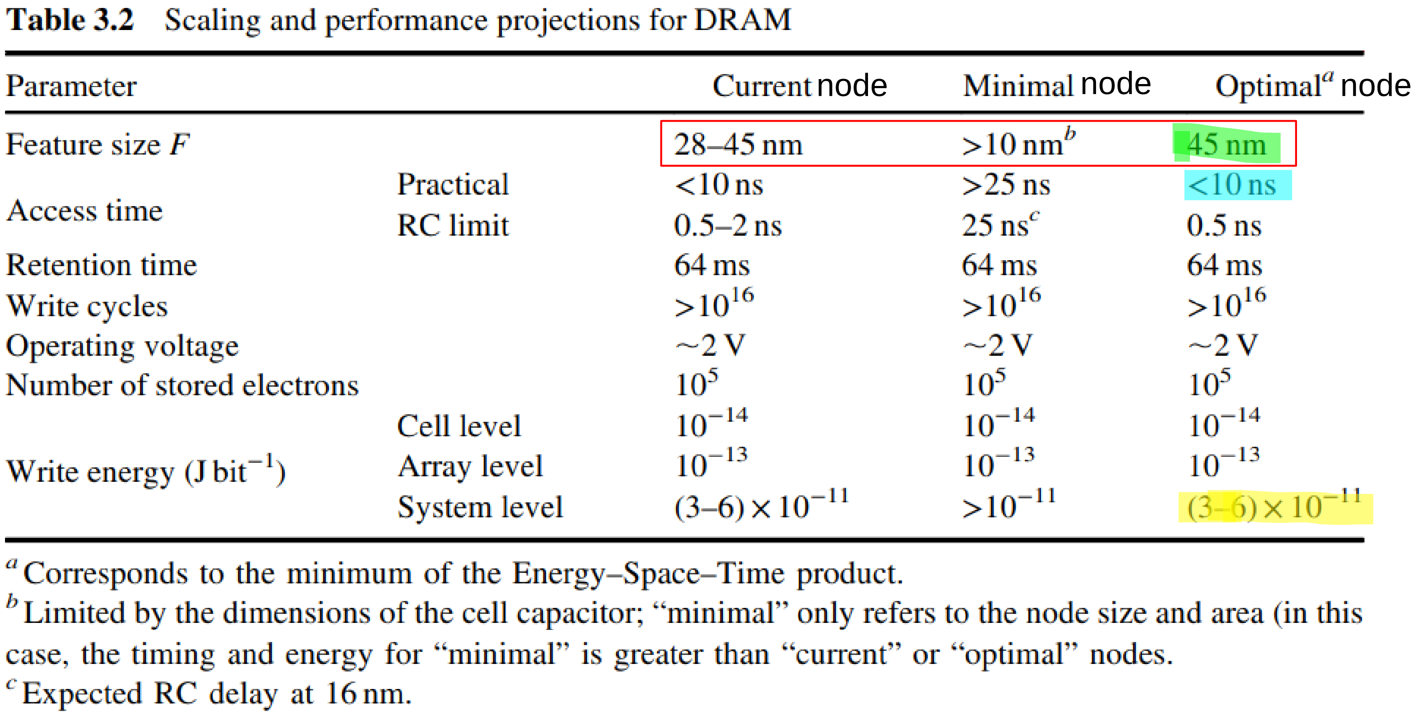


Table 3.2 Scaling and performance projections for DRAM

Parameter		Current node	Minimal node	Optimal[a] node
Feature size F		28–45 nm	>10 nm[b]	45 nm
Access time	Practical	<10 ns	>25 ns	<10 ns
	RC limit	0.5–2 ns	25 ns[c]	0.5 ns
Retention time		64 ms	64 ms	64 ms
Write cycles		$>10^{16}$	$>10^{16}$	$>10^{16}$
Operating voltage		~2 V	~2 V	~2 V
Number of stored electrons		10^5	10^5	10^5
Write energy (J bit^{-1})	Cell level	10^{-14}	10^{-14}	10^{-14}
	Array level	10^{-13}	10^{-13}	10^{-13}
	System level	$(3–6)\times10^{-11}$	$>10^{-11}$	$(3–6)\times10^{-11}$

[a] Corresponds to the minimum of the Energy–Space–Time product.
[b] Limited by the dimensions of the cell capacitor; "minimal" only refers to the node size and area (in this case, the timing and energy for "minimal" is greater than "current" or "optimal" nodes.
[c] Expected RC delay at 16 nm.

→ DRAM has a scaling limit due to growing series resistance of capacitor.

Emerging Nanoelectronic Devices Chapter 3

Flash memory

1 – 4 TB

SSD harddrives

USB sticks

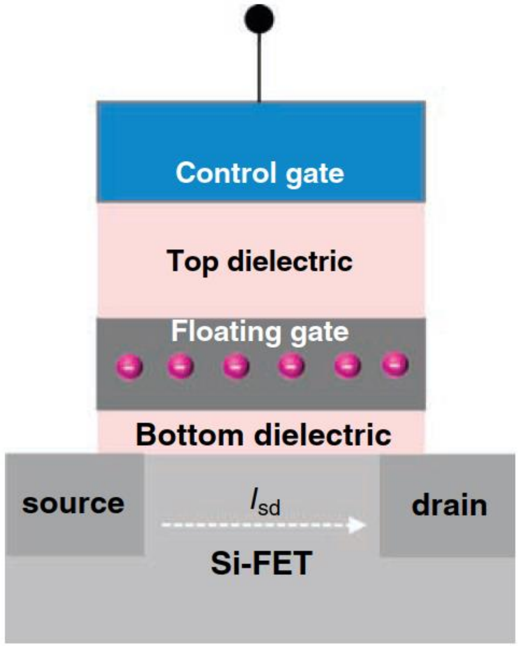
phones

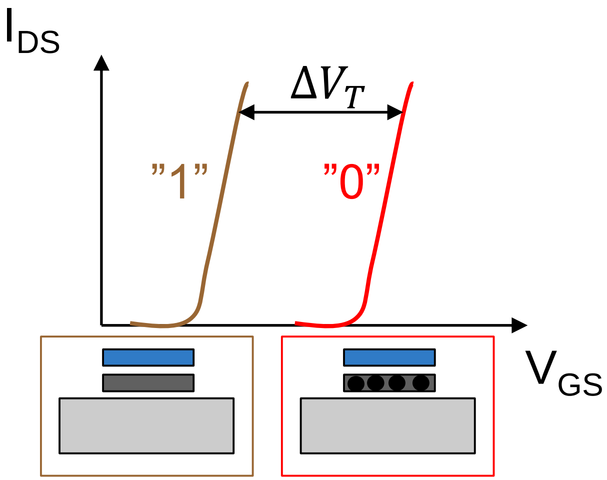
Flash memory cell

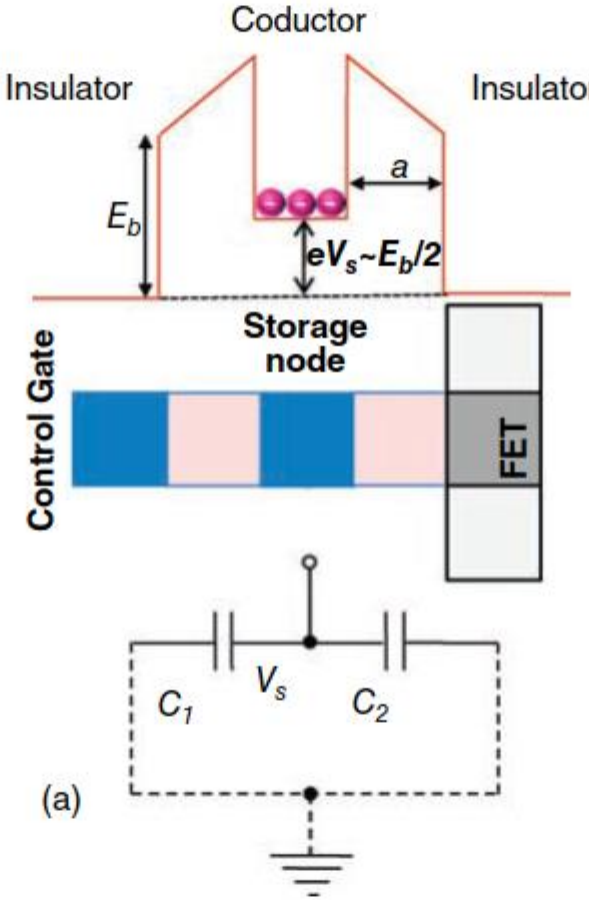
- Charges trapped in a floating gate → memory state
- Read out by V_T shift in n-MOSFET

- Barriers by dielectrics → E_b ~ 3 eV → non-volatile

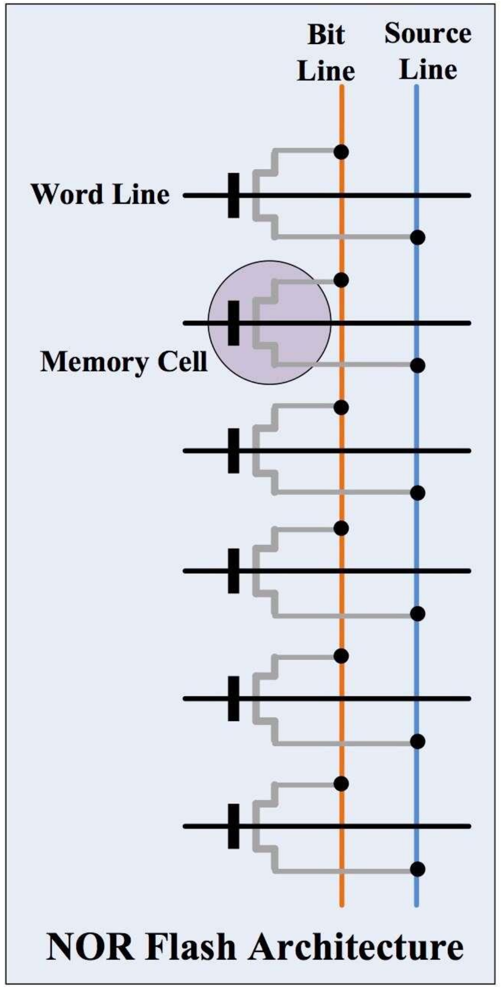
- Read/write performed by ”bending” barriers by biasing



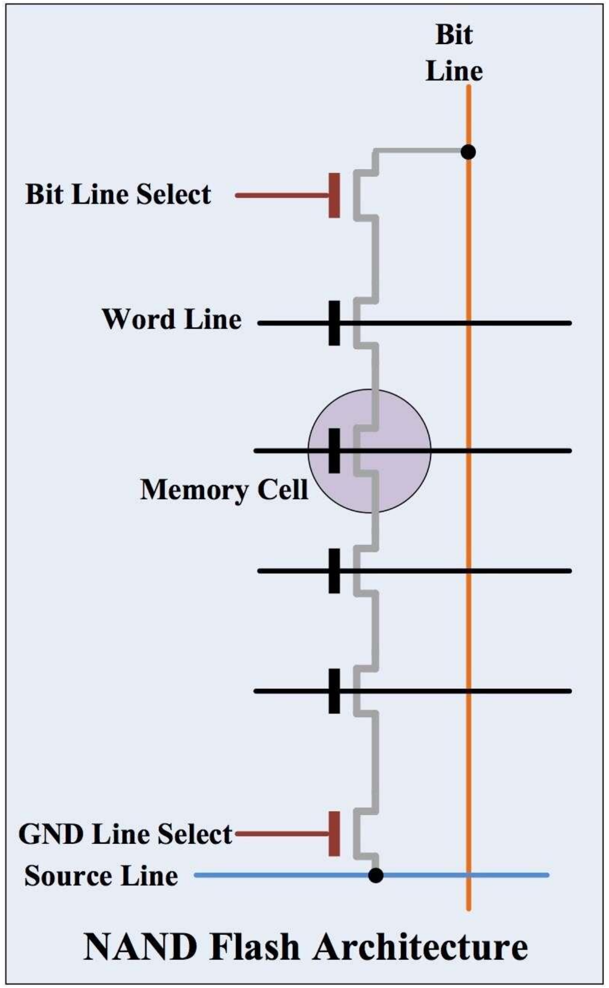




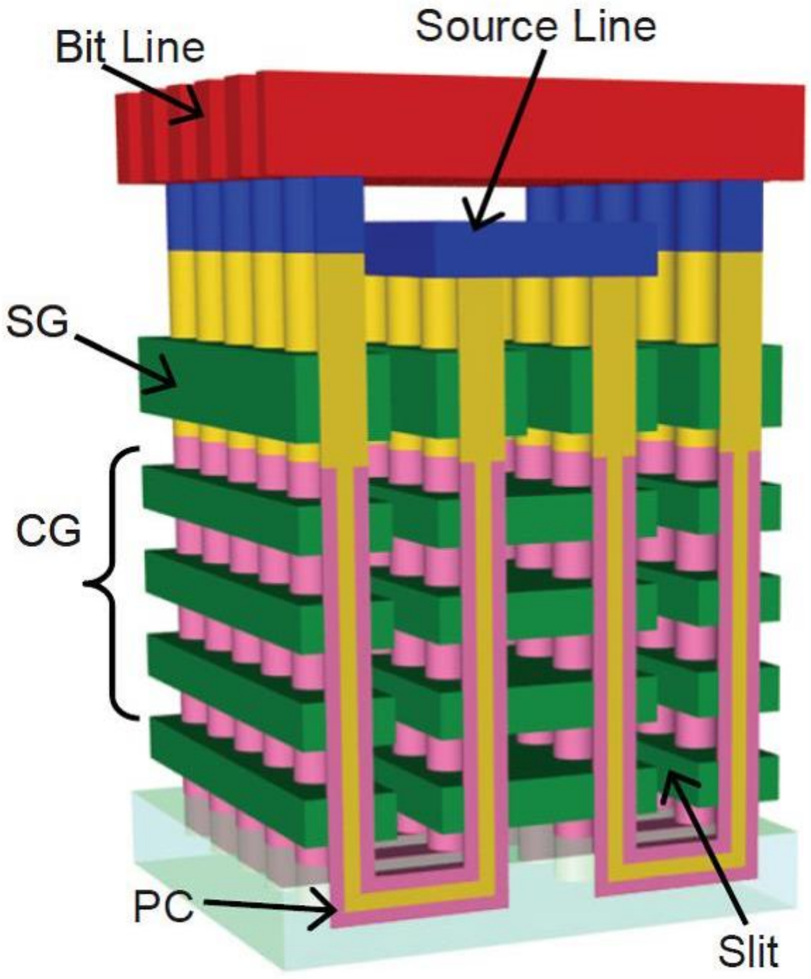
Different types of Flash



NOR Flash Architecture



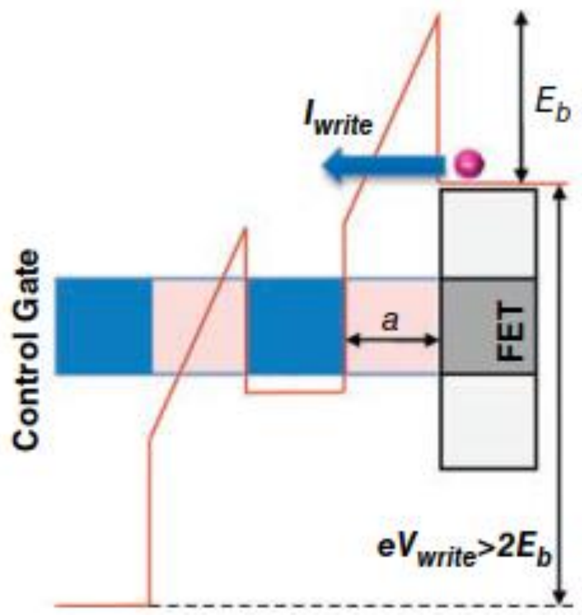
NAND Flash Architecture

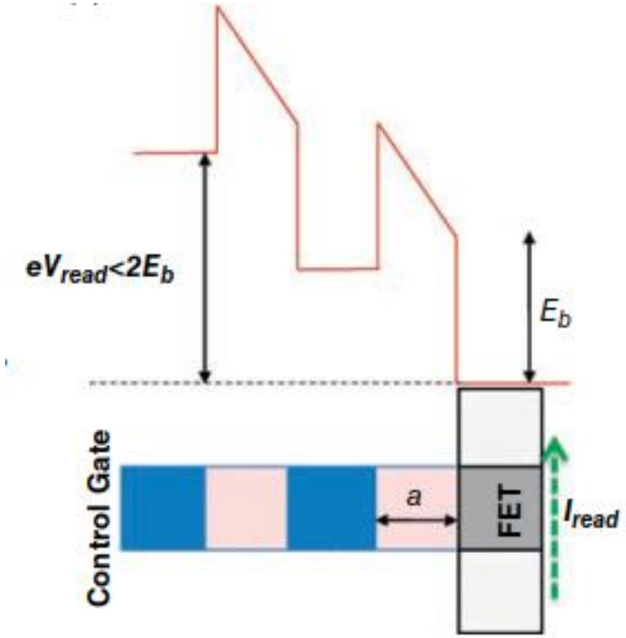


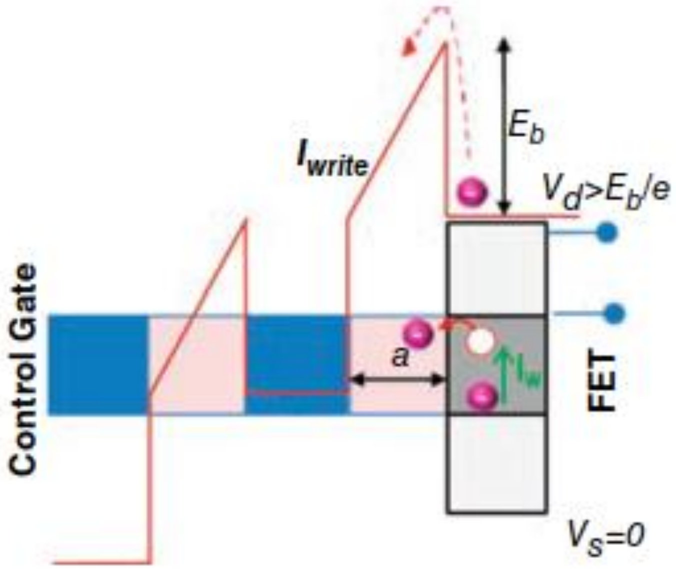
3D NAND Flash

https://www.embedded.com/flash-101-nand-flash-vs-nor-flash/, https://tech.nikkeibp.co.jp/dm/english/NEWS_EN/20090619/171977/?SS=imgview_en&FD=-925410382

Flash operation







Write by direct tunneling

- $V_{write} > 2E_b/q$ ~ **6-15 V**
- Barrier → triangular
- Tunneling into island
- Used by NAND

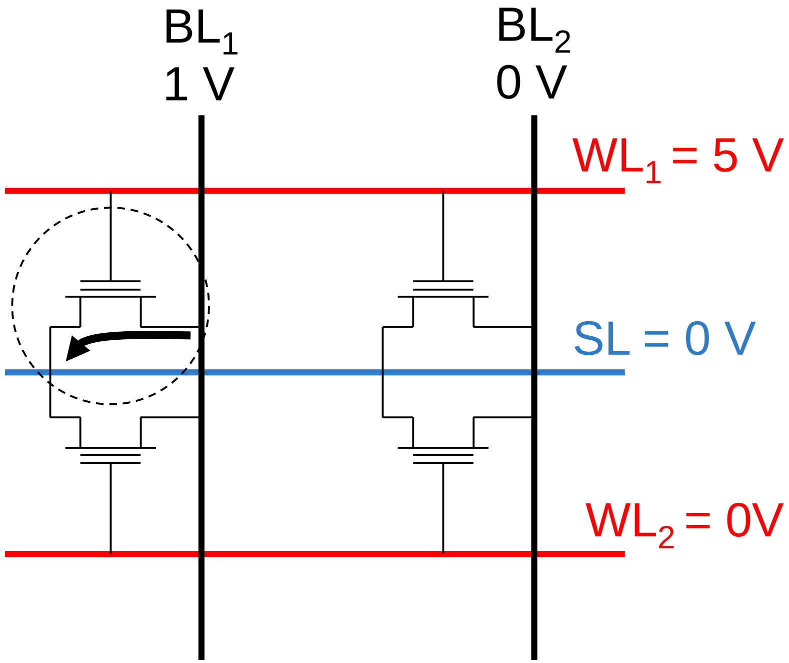
Read

- $V_{read} < 2E_b/q$ ~ **4-5 V**
- FET current senses charge state
- V_T shifted by charge

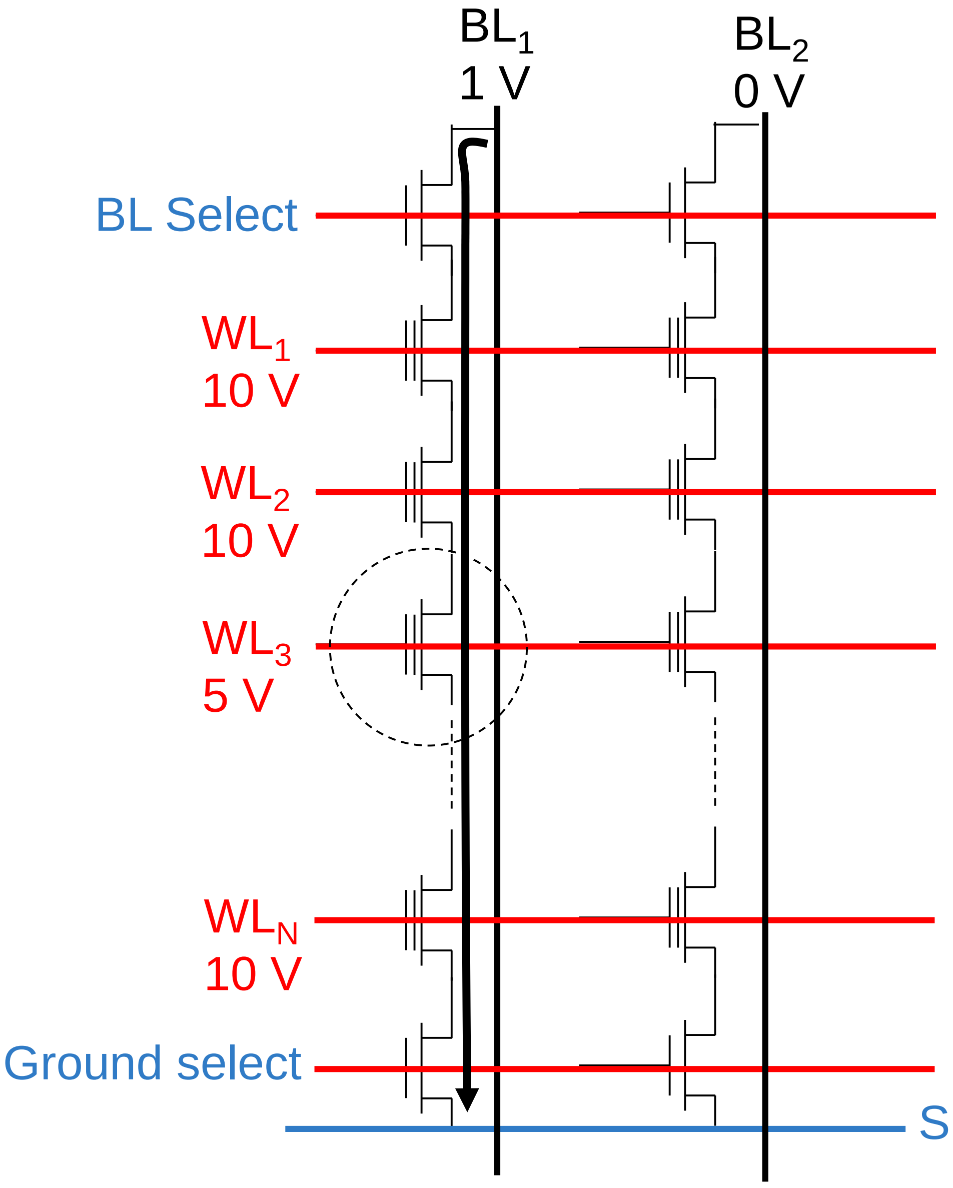
Write by hot electron injection

- High $V_{DS} > E_b/q$ ~ 3-4 V
- Injection above barrier
- Limits shortest L. **Why?**
- Faster than tunneling
- Used by NOR

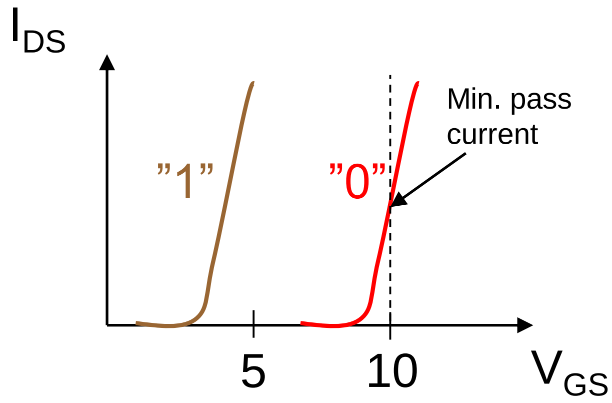
Reading Flash



NOR Flash



NAND Flash



Energy usage Flash

WRITE by
Direct tunneling

- **Bend barriers + pump charge onto island:**
- $\boldsymbol{E_{DT}} = \frac{\boldsymbol{C_1 C_2}}{\boldsymbol{C_1+C_2}} V_{write}^2 + qN_{el}V_{write}$

WRITE by
Hot electron injection

- **Inefficient: 1e$^-$ per 10^5-10^6 are injected (η=I$_{GS}$/I$_{DS}$)**
- $\boldsymbol{E_{HEI} = \frac{1}{\eta} N_{el} * qV_{ds}}$

Large V$_{write}$ → System energy consumption limited by line charging and peripheral circuitry

Summary Flash

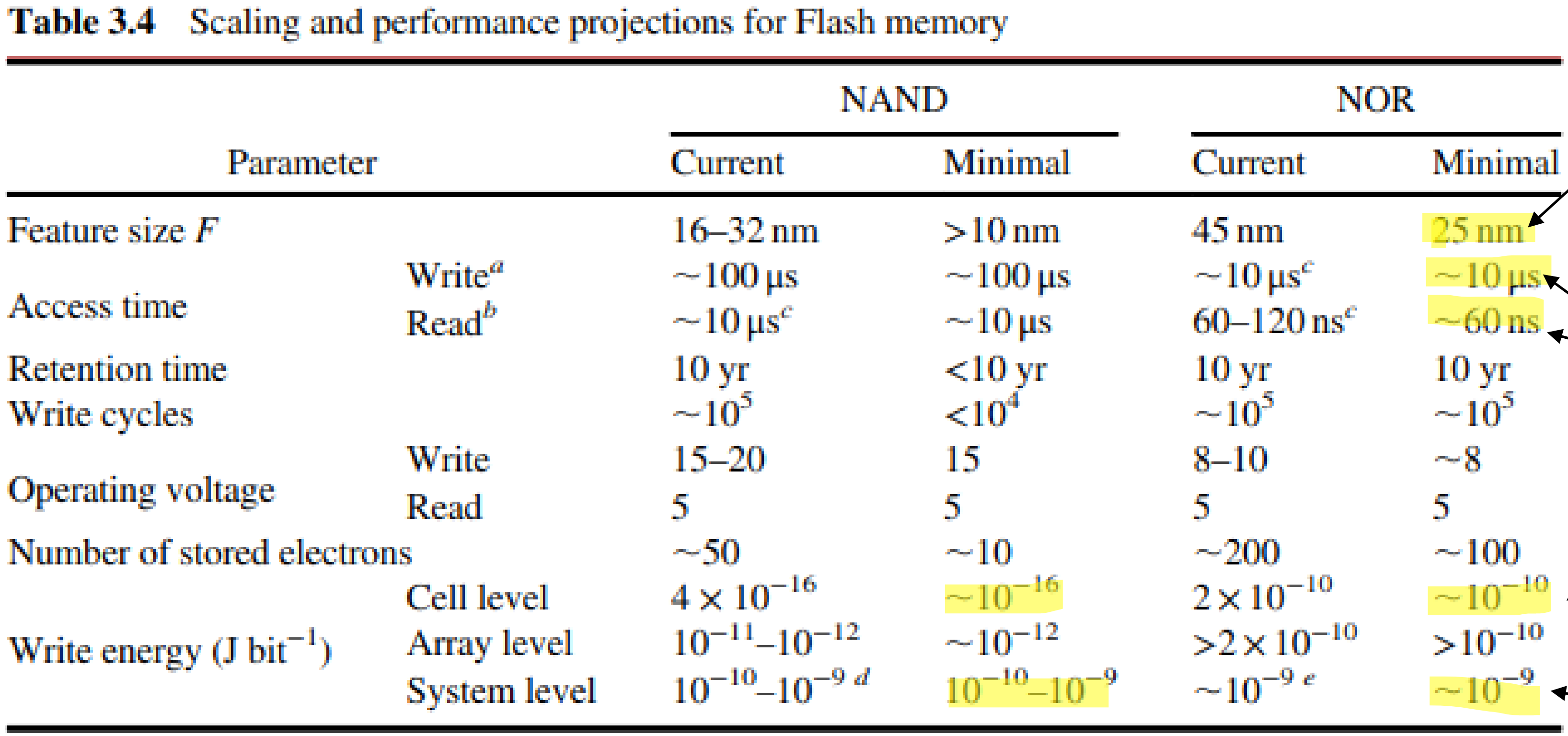


Table 3.4 Scaling and performance projections for Flash memory

		NAND		NOR	
Parameter		Current	Minimal	Current	Minimal
Feature size F		16–32 nm	>10 nm	45 nm	25 nm
Access time	Write[a]	~100 µs	~100 µs	~10 µs[c]	~10 µs
	Read[b]	~10 µs[c]	~10 µs	60–120 ns[c]	~60 ns
Retention time		10 yr	<10 yr	10 yr	10 yr
Write cycles		$\sim 10^5$	$<10^4$	$\sim 10^5$	$\sim 10^5$
Operating voltage	Write	15–20	15	8–10	~8
	Read	5	5	5	5
Number of stored electrons		~50	~10	~200	~100
Write energy (J bit^{-1})	Cell level	4×10^{-16}	$\sim 10^{-16}$	2×10^{-10}	$\sim 10^{-10}$
	Array level	10^{-11}–10^{-12}	$\sim 10^{-12}$	$>2 \times 10^{-10}$	$>10^{-10}$
	System level	10^{-10}–10^{-9} [d]	10^{-10}–10^{-9}	$\sim 10^{-9}$ [e]	$\sim 10^{-9}$

NOR scales worse $10F^2$ vs $4F^2$

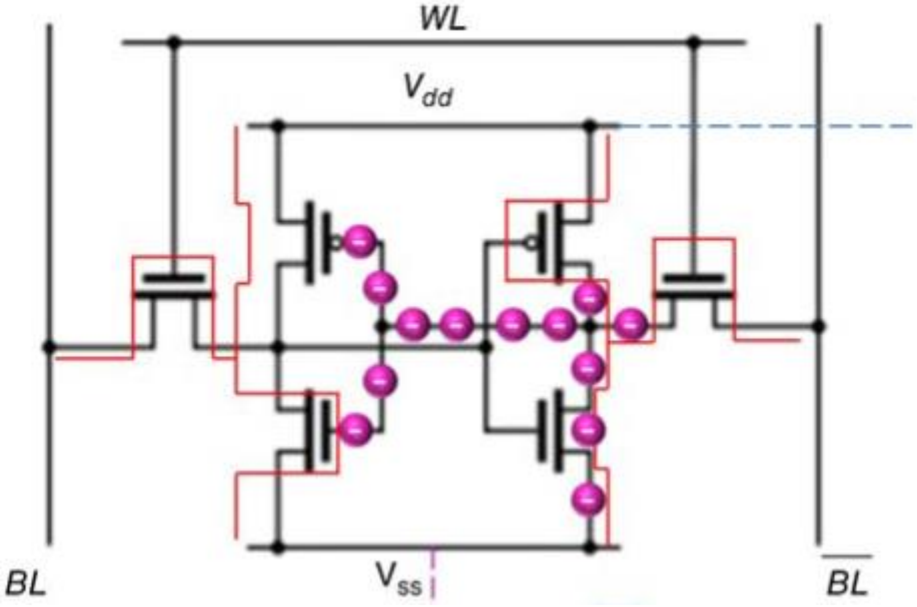
NOR is much faster

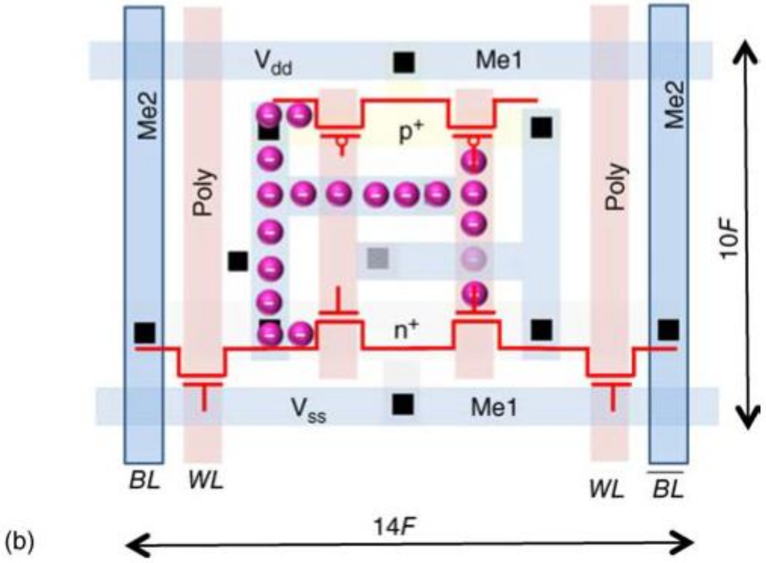
NAND cell write is more efficient

But on system level they are equal

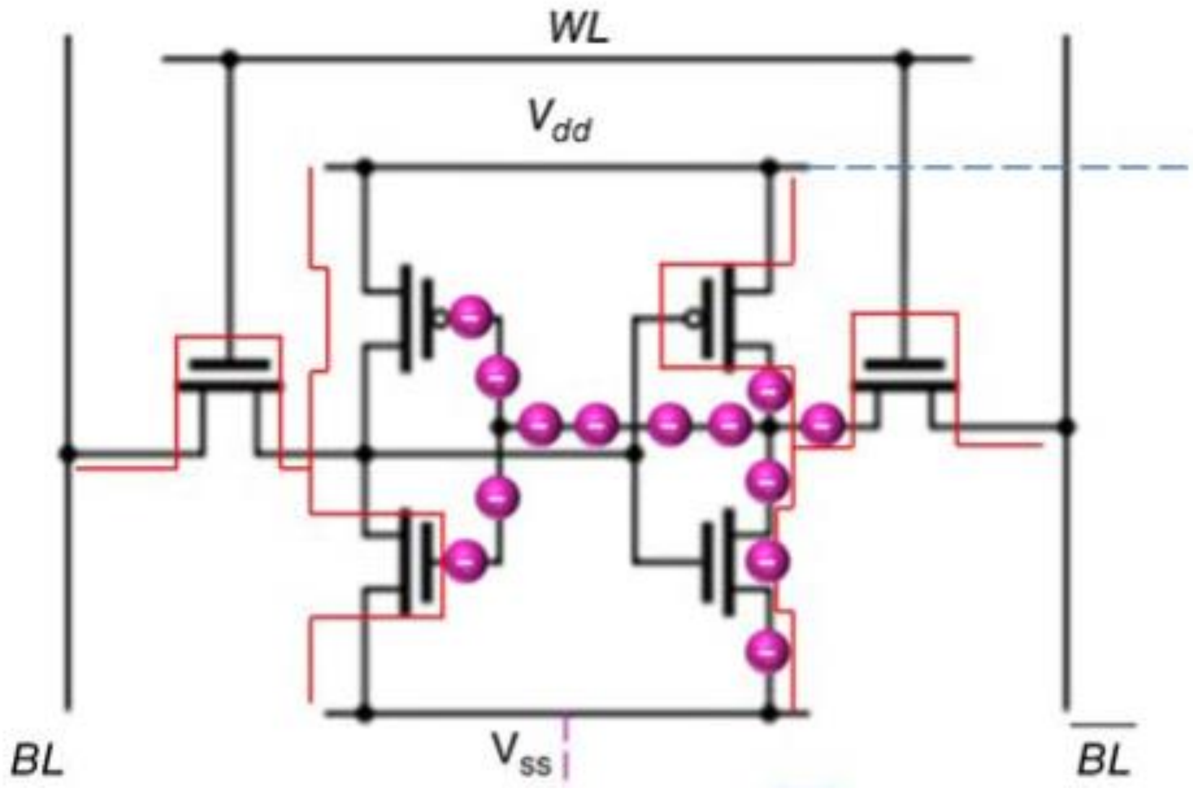
Static Random Access Memory

- Fastest memory there is
- Used for registers, caches
- Takes up ~ half of chip area
- Principle: 2 CMOS inverters connected back to back
- Transistor performance matching is crucial!
- 6 transistors per cell ~140F^2

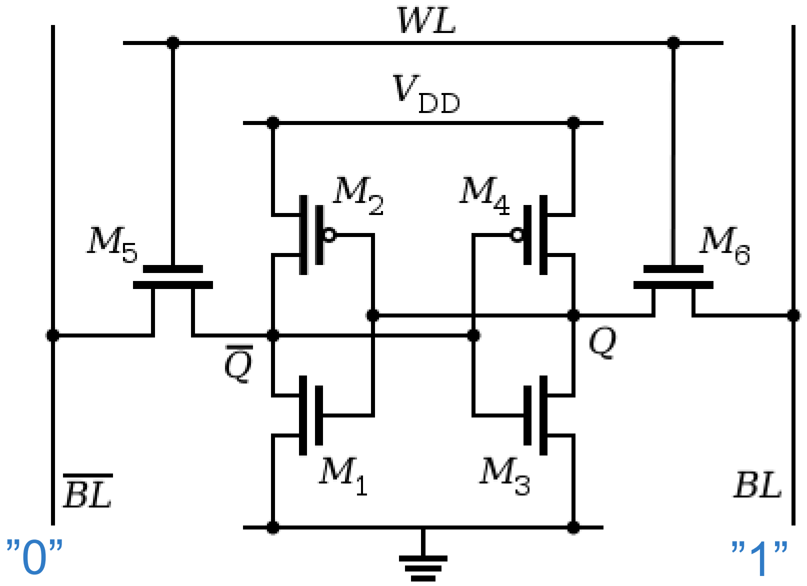




SRAM as charge based memory



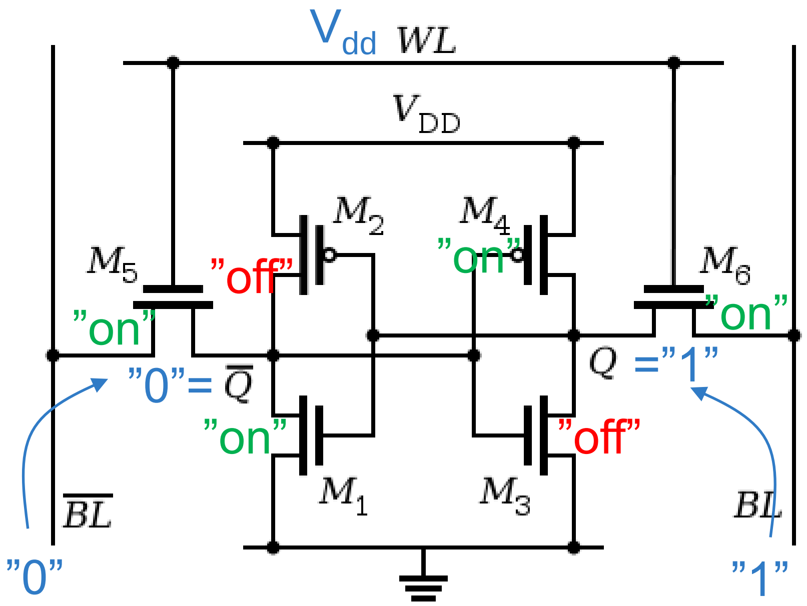
SRAM write



1. Apply ”bit” to bit line BL: ”1”

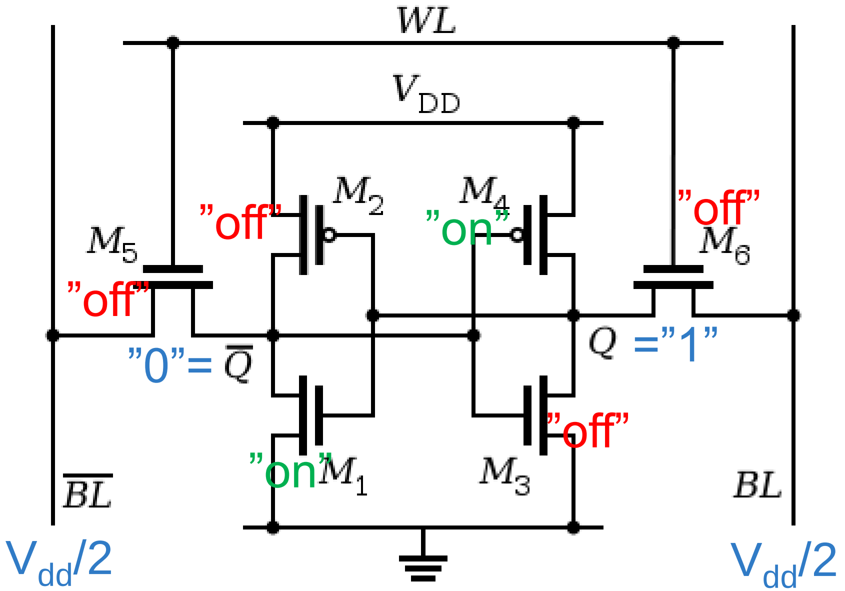
   - And opposite to conjugate BL

SRAM write



1. Apply ”bit” to bit line BL: ”1”
2. Turn on M5 and M6 via WL to save bit
   - M5 saves on right inverter
   - M6 saves on left inverter
   - M5/6 are stronger (larger) than M1-4

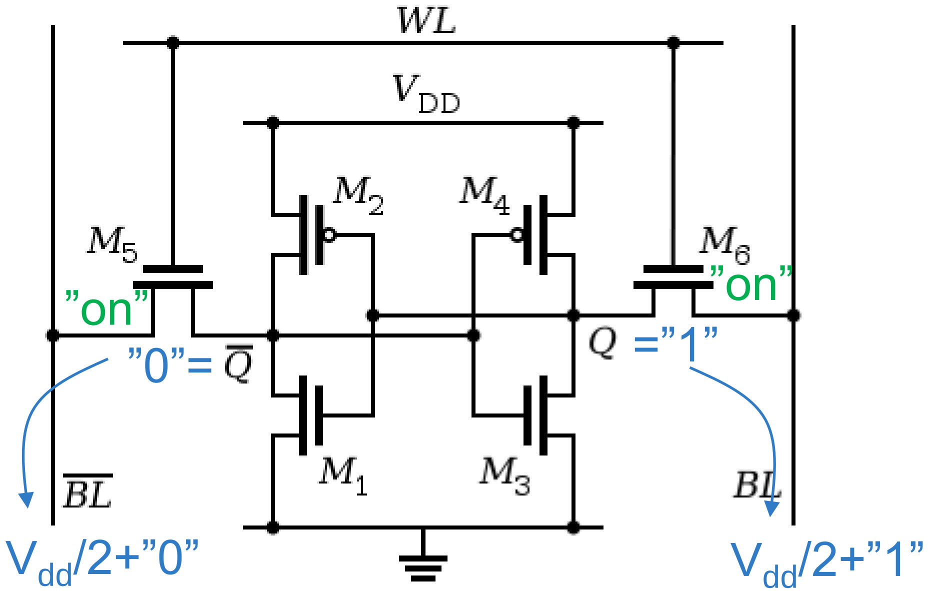
SRAM read



1. Precharge bit lines to $V_{dd}/2$

   - Saves time since lines are long

SRAM read

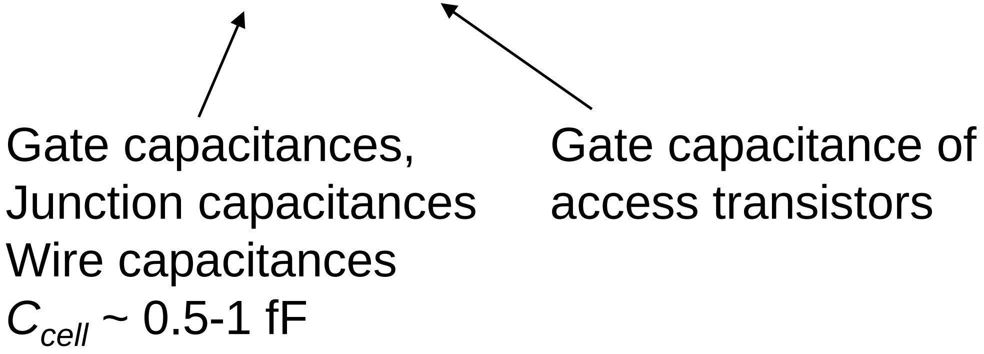


1. Precharge bit lines to $V_{dd}/2$
2. Turn on M5-6 to take out charge to bit lines
3. Voltage difference between BL and its conjugate is amplified and sensed. Sign → "0" or "1"

   - If done fast (small voltage change) then the SRAM state recovers (non-destructive read)

Energy usage and access time

- $E_{cell} = \left(C_{cell} + 2C_g\right)V_{dd}^2$ ~ 0.3-2 fJ/write



Gate capacitances,
Junction capacitances
Wire capacitances
C_{cell} ~ 0.5-1 fF

Gate capacitance of
access transistors

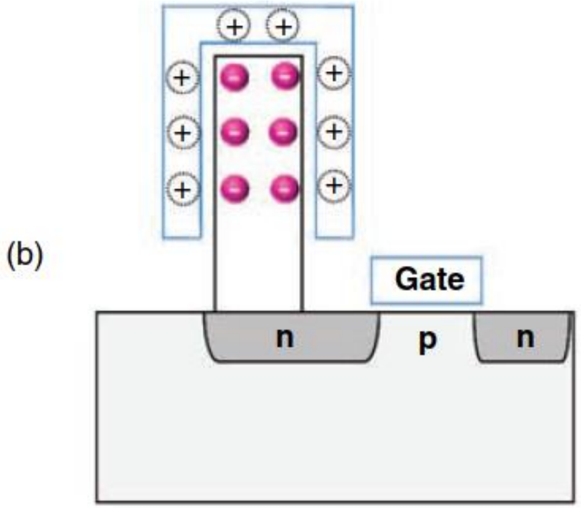
- Total energy dominated by capacitances of metal WL (n) and BL (m) lines

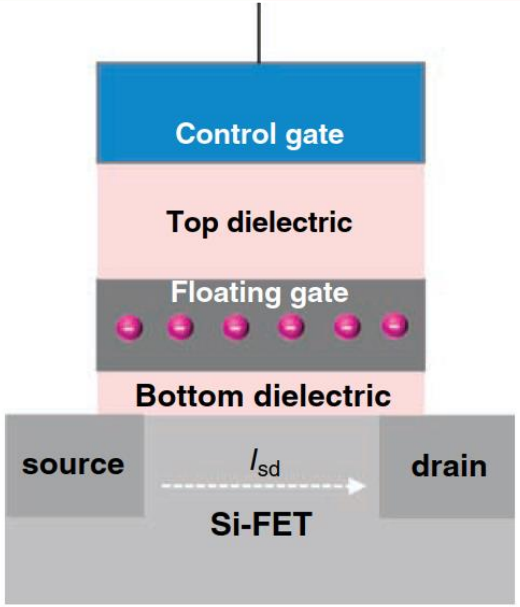
$$E_{write} \approx (n+m)C_{line}V_{dd}^2 \sim \ 100\ fJ$$

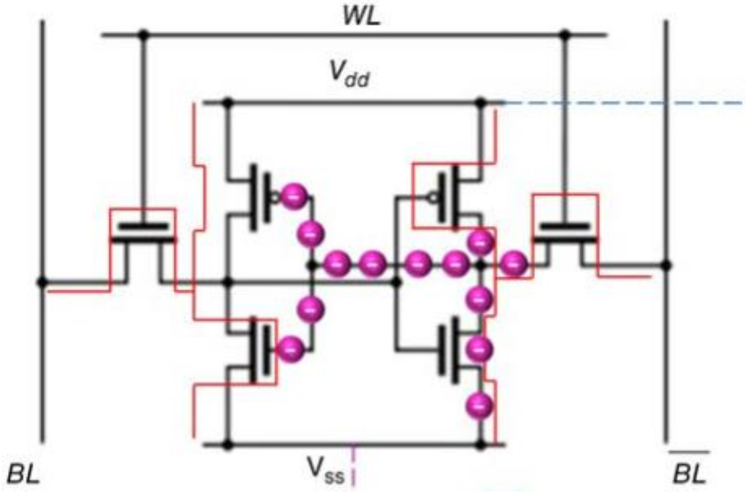
- Access time:
- $t_{SRAM} = (R_{FETon} + R_{line}) * (C_{cell} + C_{line}) \sim 1\ ns$

Summary

	DRAM	NAND FLASH	SRAM
Density	$6F^2$	$4F^2$ $(4/n\ F^2)$	$>100F^2$
Speed	10 ns	10 µs	1 ns
Storage	Volatile	Non-volatile	Volatile
Energy / write	10^{-11} J	10^{-9} J	10^{-13} J




Control gate
Top dielectric
Floating gate
Bottom dielectric
source
I_{sd}
drain
Si-FET


WL
V_{dd}
BL
V_{ss}
$\overline{BL}$