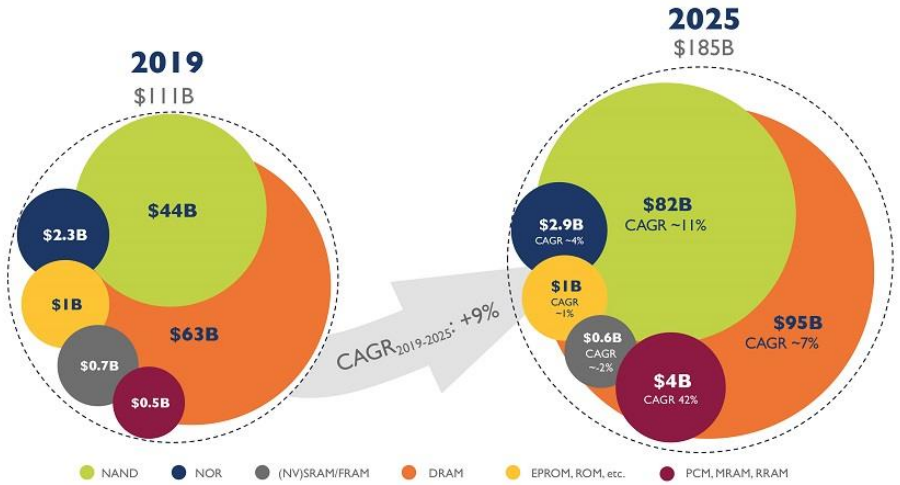
Lecture 8

Memory and Other Semiconductor Devices

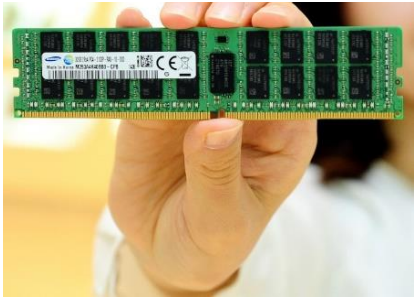
MOS-based memory devices

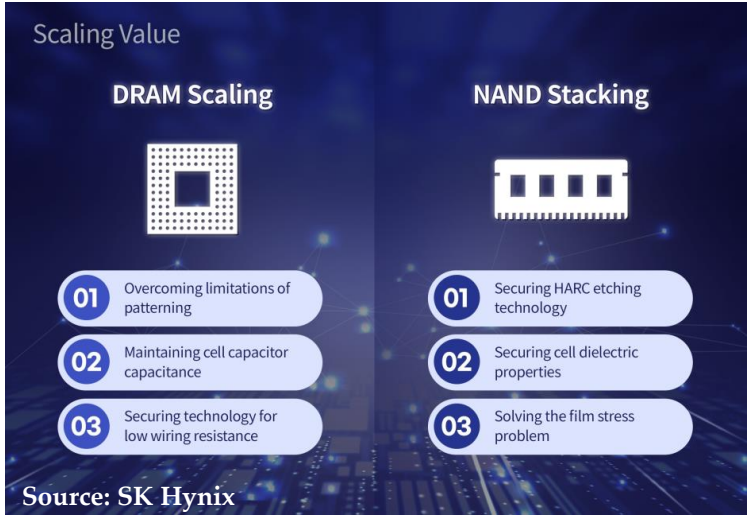
2019 – 2025 stand-alone memory market revenue forecast with breakdown by technologies

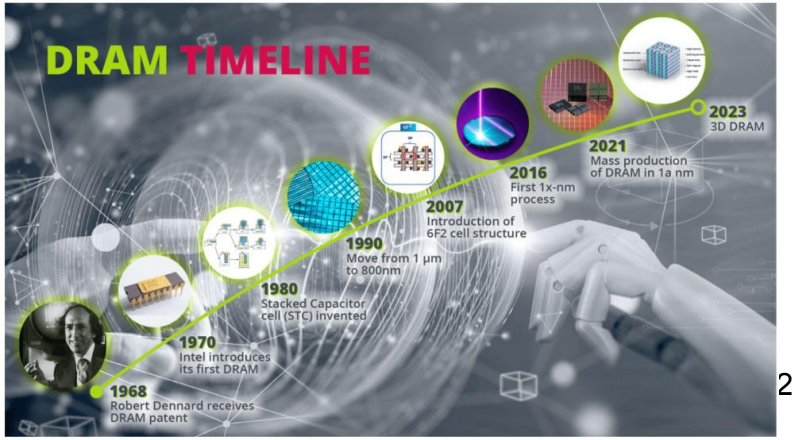
(Source: Status of the Memory Industry 2020 report, Yole Développement, 2020)



YOLE Développement © 2020 | www.yole.fr – www.i-micronews.com

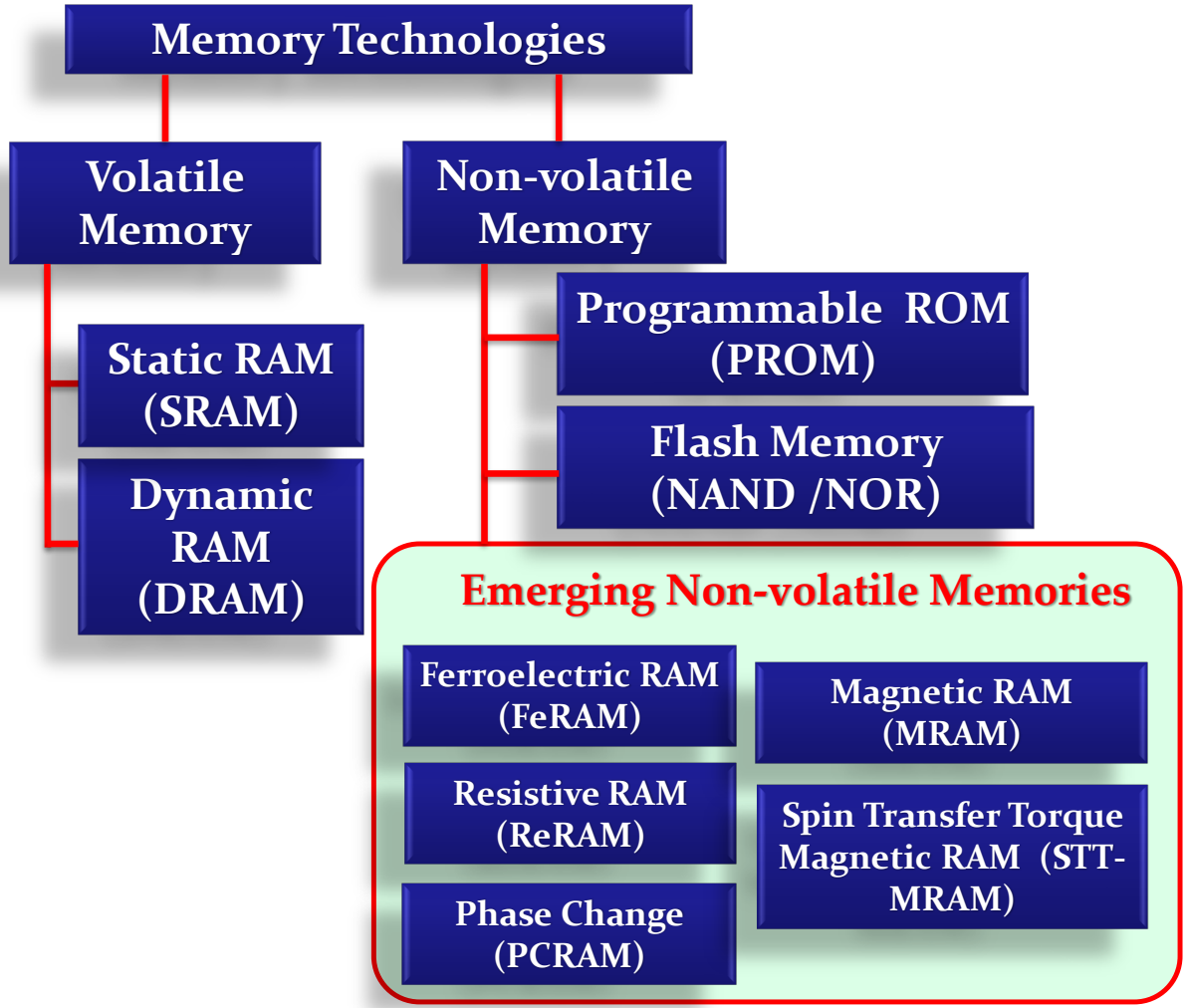






DRAM timeline. (Image: Neumonda)

Memory technologies – 1



Memory Terms	
Volatile Memory	Memory loss without power
Non-volatile Memory	Data is stored even without power
RAM (random access memory)	Memory with both read and write capabilities. Accessible to individual bit in an memory array.
ROM (read only memory)	Data is stored "permanently" after programed. Re-write require high voltage or UV light.
Static	Holds data as long as there is power.
Dynamic	Data must be refreshed periodically.

Memory technologies – 2

Memory Performance	Semiconductor memories				Emerging non-volatile memories			
	SRAM	DRAM	NOR	NAND	FeRAM	STT-MRAM	PCRAM	ReRAM
Non-volatile	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Cell Size (F^2)	>100	6	10	<4 (3D)	15-34	6-50	4-30	4-12
Read Time (ns)	~ 1	~ 10	~ 50	~ 10 µs	20-80	< 10	< 10	< 10
Write/Erase Time (ns)	~ 1	~10	100µs –1ms	100µs –1ms	50	< 10	~50	< 10
Endurance	> 10^{16}	> 10^{16}	> 10^5	>10^4	10^{12}	10^{15}	> 10^8	10^6
Retention (years)	NA	~64 ms	>10	>10	>10	>10	> 10	> 10
Write Energy (J/bit)	~ fJ	~ 10 fJ	~ 100 pJ	~ 10 fJ	~ 0.1 pJ	~ 0.1 pJ	~ 10 pJ	~ 0.1 pJ
Voltage	< 1V	< 1 V	> 10 V	>10V	2-3	< 1.5 V	< 3 V	< 3 V
	Existing Products						Developing	

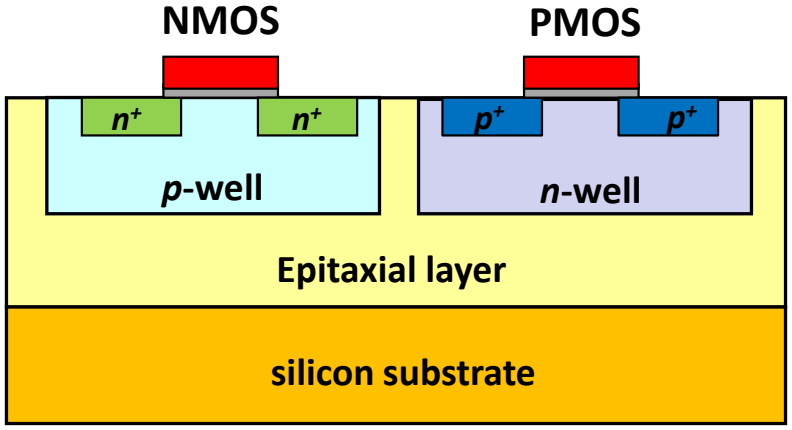
Complementary MOS (CMOS)

- Combining two types of MOS transistors to construct an inverter:

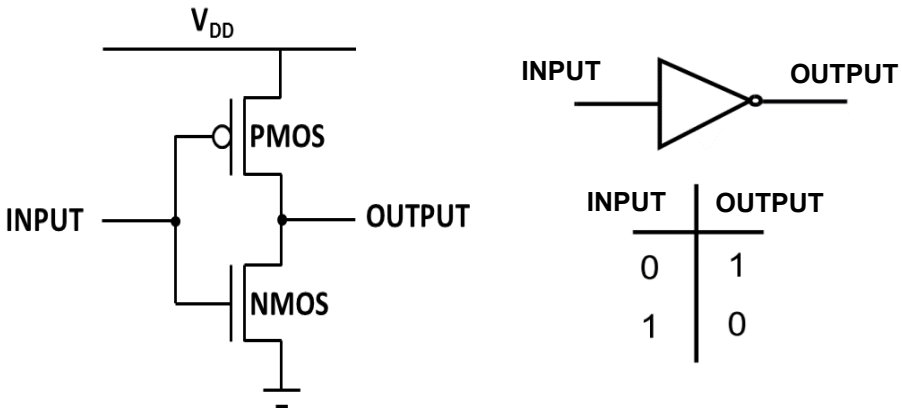
nMOS  *n*-channel (electron inversion layer) MOS transistor; turned on with positive gate voltage

pMOS  *p*-channel (hole inversion layer) MOS transistor; turned on with negative gate voltage

- When the Gate is High (Input is 1) the nMOS turns on pulling the output Low (Output 0). Likewise, when the Gate is Low (Input 0), the pMOS is turned on pulling the output High (Output 1).

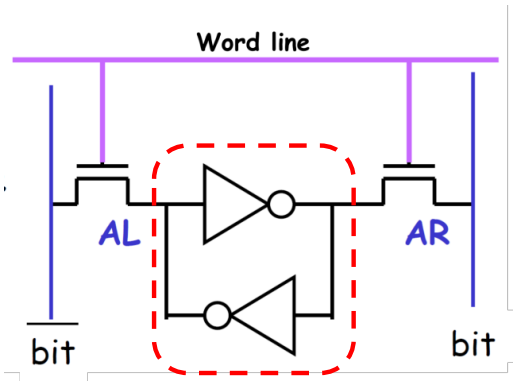


CMOS inverter (NOT gate)

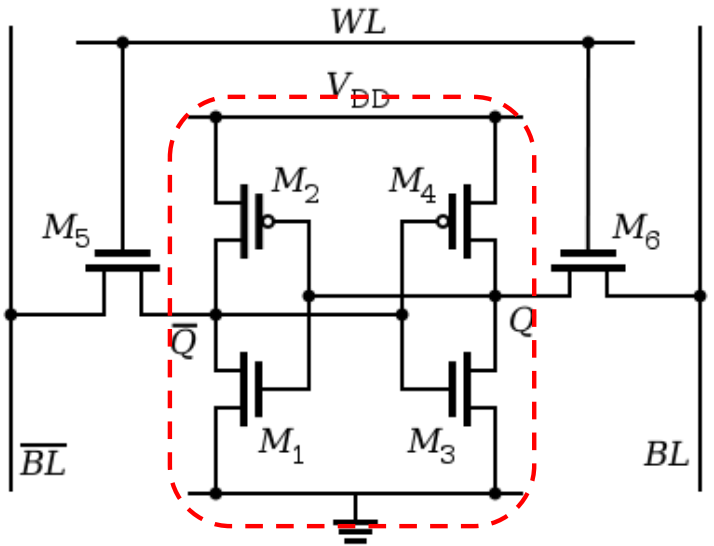


Static RAM (SRAM)

- A memory circuit is considered static if the stored data can be ***retained indefinitely***, as long as the ***power supply is on***, ***without any need for periodic refresh operation***.
- A typical SRAM cell is made up of six MOSFETs. The memory cell consists of simple CMOS inverters connected back to back, and two access transistors.
- Each bit in an SRAM is stored on four transistors (M1, M2, M3, M4) that form two cross-coupled (bistable) inverters.



A bistable inverters for storage, where AL and AR are access transistors



A six-transistor CMOS SRAM cell

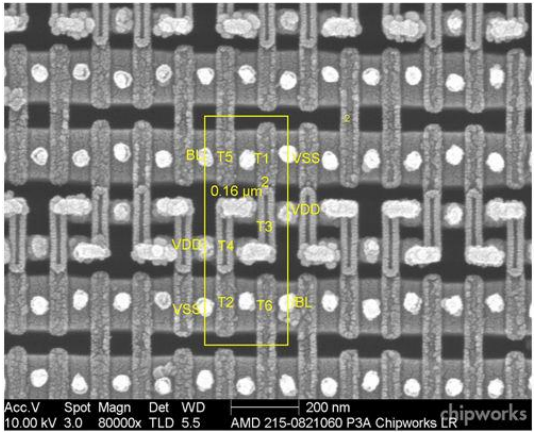
SRAM operation

- The data storage cell, *i.e.* the one-bit memory cell in the static RAM arrays, invariably consists of ***a simple latch circuit with two stable operating points***. Depending on the preserved state of the two inverter latch circuit, the data being held in the memory cell will be interpreted either as logic '**0**' or as logic '**1**'.

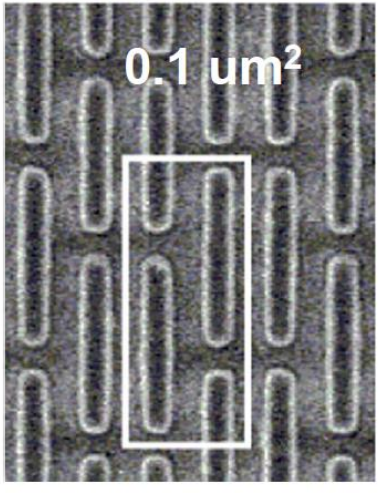
- Two additional ***access transistors*** serve to control the access to a storage cell, connecting the cell to the complementary bit line columns.

- ***Access to the cell during read or write operation is enabled by the word line*** which controls the two access transistors M5 and M6 which, in turn, control whether the cell should be connected to the bit lines.

- ***Bit lines are used to transfer data for both read and write operations***. It has two bit lines, both the signal and its inverse are typically provided.

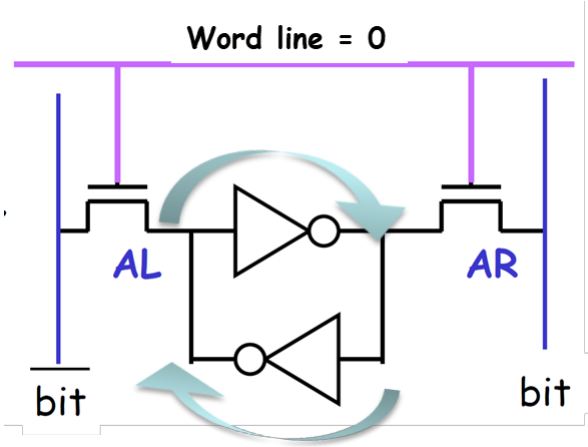


TSMC 20nm FinFET



SRAM writing and reading – 1

- A SRAM cell has three different states:
  - ***Standby*** or ***hold*** - the circuit is idle,
  - ***Reading*** - requesting the data condition,
  - ***Writing*** - updating the contents.



- WL = 0, hold operation
- WL = 1, read or write operation

SRAM Standby

- If the word line is not asserted, the access transistors M5 and M6 disconnect the cell from the bit lines. The two cross-coupled inverters formed by M1 – M4 will continue to reinforce each other as long as they are connected to the power supply.

SRAM Writing Operation

- The write cycle begins ***by applying the value to be written to the bit lines***.
- A "1" is written by inverting the values of the bit lines, *i.e.* setting BL to "1", and $\overline{BL}$ to "0". ***WL is then turned on and the new value that to be stored is latched in***.
- To write a "0", we would apply a "0" to the bit lines, *i.e.* setting BL to "0", and $\overline{BL}$ to "1".
- This works because the bit line is designed to easily override the previous state of the cross-coupled inverters.

SRAM reading and writing – 2

SRAM Reading Operation

- To read the stored SRAM data in the simplest operation: ***the word line is turned on, and the SRAM cell state can be retrieved*** by a single access transistor and bit line, *e.g*. M6, BL.

- If the cell is in state 1; the signal on BL line is high and the signal on $\overline{\text{BL}}$ line is low. The opposite is true if the cell is in state 0. Thus, the two bits lines are always complements of each other.

- In the reading mechanism, the read cycle is started by ***pre-charging*** both bit lines BL and $\overline{\text{BL}}$, *i.e*., driving the bit lines to a threshold voltage, which is about the midrange voltage between logical 1 and 0.

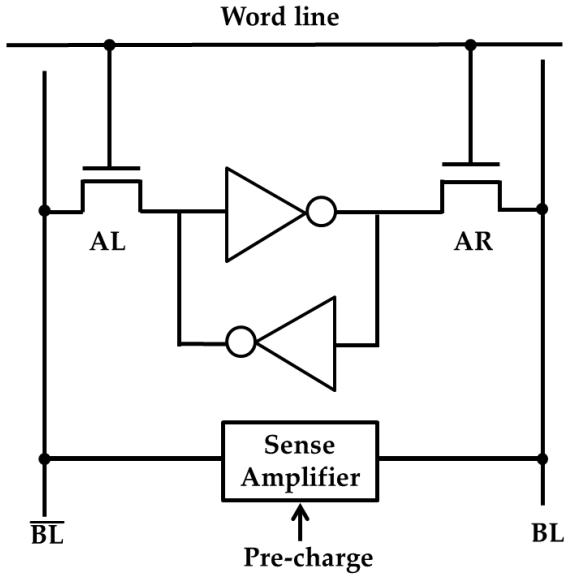
- Then, the word line is turned on to enable both the access transistors M5 and M6, which causes the bit line BL voltage to either slightly drop or rise.

- If BL voltage rises, the $\overline{\text{BL}}$ voltage drops, and vice versa. Then the BL and $\overline{\text{BL}}$ lines will have a small voltage difference between them.

SRAM reading and writing – 3

- A *sense amplifier* is used to sense which line has the higher voltage and thus determine whether 1 or 0 was stored. A sense amplifier is basically a simple differential amplifier to compare the difference between BL and $\overline{BL}$.

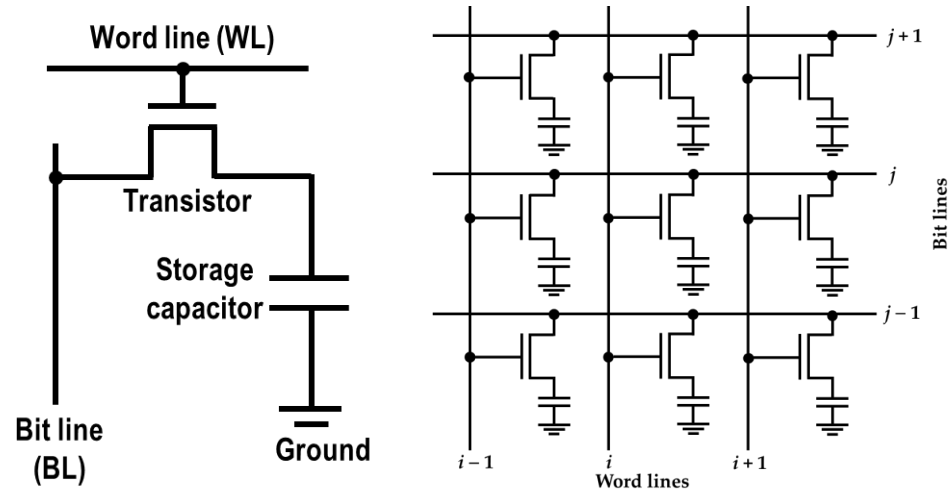
- If BL > $\overline{BL}$, output is 1, if BL < $\overline{BL}$, output is 0.



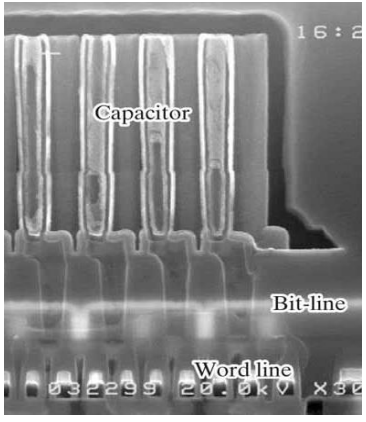
Read Operation with Sense Amplifier

Dynamic RAM (DRAM)

- Invented by Robert Dennard in 1966 at IBM, DRAM operation uses a single transistor and capacitor and its operation is based around the charge held on the capacitor.
- The ***transistor is used to access the data*** while the ***capacitor is used to stored the data***, *i.e.* the presence and absence of charge, determines the value of stored bit, "1" or "0".
- The drawback of storing charge in capacitor is, it cannot last for a long time because after a limited amount of time the charge gradually drains away. Hence, DRAM cells require a periodic refreshing of the stored data (msec).
- To write data to the cell the word line needs to be charged high while the bit line is charged either high or low, depending on the information to be stored in the cell.



Schematic of DRAM 1T1C (one-transistor-one-capacitor) cell and its memory cells architecture

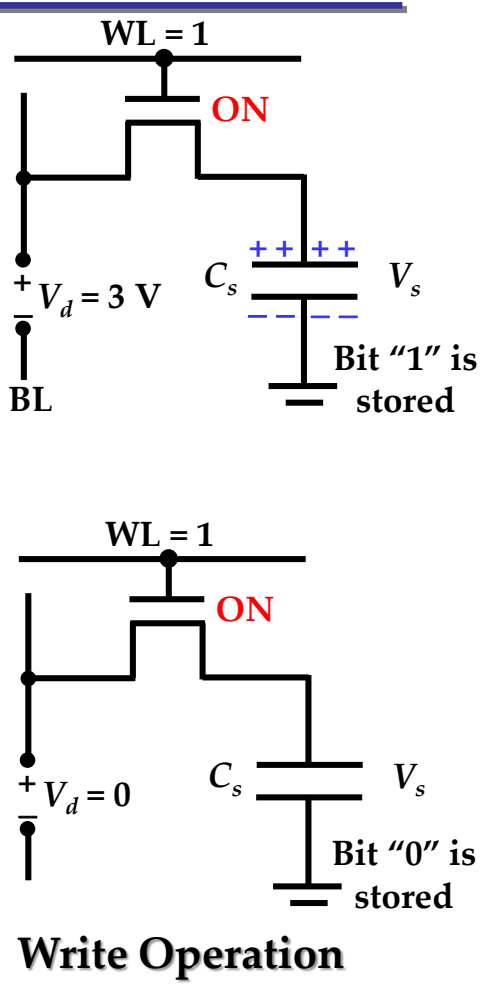


Cross-sectional SEM image of a DRAM cell.

DRAM cell write and read operation – 1

DRAM Write Operation

- To write a bit into the capacitor well, first the transistor is turned on by applying a gate potential (word line = 1). The data are written into the cell by placing the desired voltage level on the bit line, V_d.
- When the ***bit line is set at high***, *e.g.*, V_d = 3 V, it acts as the drain, while the capacitor acts as the source.
  - If V_s is originally at high level then no current flows and bit "1" stays.
  - If V_s is originally at low level then current will flow and the capacitor will be charged until the capacitance well is filled, *i.e.*, $Q_s = C_s V_s$, bit "1" is stored.
- When the ***bit line is set at low***, *e.g.* V_d = 0, it acts as the source, while the capacitor acts as the drain.
  - If V_s is originally at high level then current will flow and the capacitor is discharged completely, *i.e.*, $Q_s = C_s V_s = 0$, bit "0" is stored.
  - If V_s is originally at low level then no current flows and bit "0" stays.

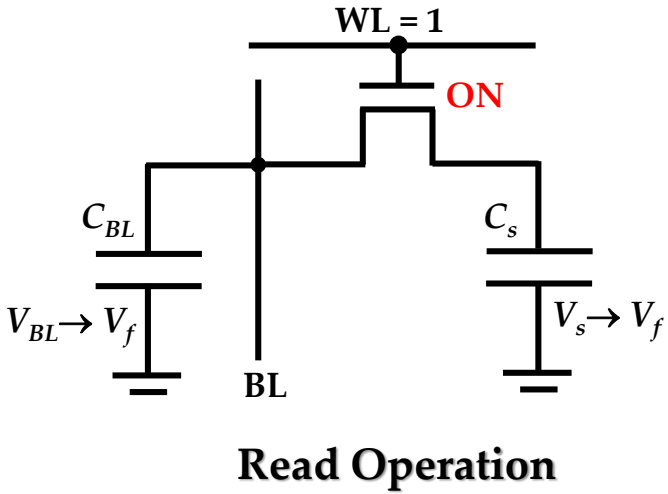


Write Operation

DRAM cell write and read operation – 2

DRAM Read Operation

- To read the DRAM cell, the ***bit line is pre-charged*** to some intermediate voltage, *e.g.*, ½V_d. Pre-charging ensures that the bit line is driven to a voltage midway between "0" and "1", so that when the cell is read out, the line need only be driven from the midway voltage to either "0" or "1".

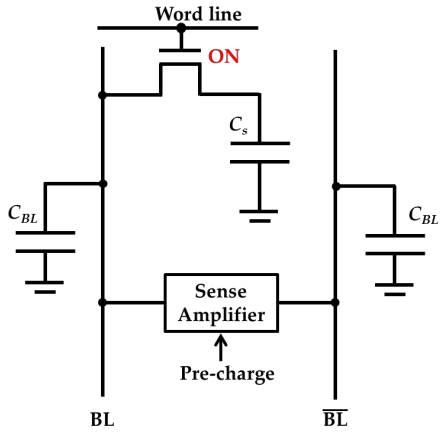


Read Operation

- Then the word line is activated, *i.e.* turn on the transistor, connecting the storage capacitor of the selected cell to the bit line ***causing the charge on the capacitor to be shared with the charge stored by the capacitance of the bit line***, C_{BL}.
- The charge redistribution continues until the voltages on both capacitors become equal to V_f.
- The voltage difference, ΔV between the pre-charged V_{BL} and V_f will be the "readout" signal. If the value stored by the cell capacitor is a "1", the bit line voltage will increase slightly, *e.g.*, a few mV. If the stored value is a "0", the bit line voltage will decrease slightly.
- The read operations wipe out the information stored in the bit cell, *i.e.*, destructive, which must then be rewritten with the detected value at the end of the read operation.

DRAM cell write and read operation – 3

- The charge stored on each capacitor is too small to be read directly and is instead measured by a circuit called a sense amplifier. Sense amplifiers are used to detect this small voltage change to produce a digital output value.



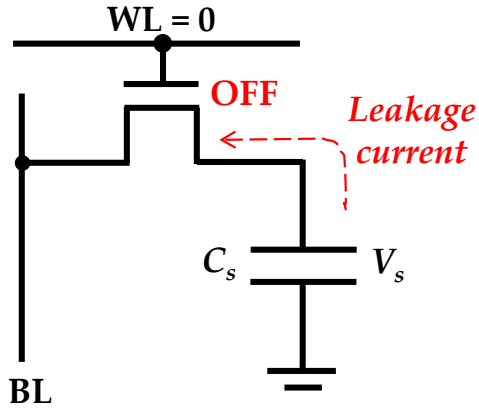
Sense Amplifier

DRAM Hold Operation

- In the hold operation, the gate potential is removed hence the channel disappears, and the charge remains in the capacitor potential well.
- However, during hold, leakage current, I_L slowly discharge C_s:

$$I_L = -\frac{\delta Q_s}{\delta t} = -C_s \frac{\delta V_s}{\delta t} \quad \Rightarrow \quad \text{hold time, } t_h \approx \left(\frac{C_s}{I_L}\right)\Delta V_s$$

- The time limit for voltage V_s to still be considered high enough as a logic "1" is defined as the hold time, $\boldsymbol{t_h}$.
- Hold time increases with larger C_s and lower I_L.



Hold Operation

Example 8a.7

QUESTION:

For DRAM operation it is assumed that a minimum of 10^5 electron for the MOS storage capacitor. If the capacitor has an area of 0.5 μm × 0.5 μm on the wafer surface, an oxide thickness of 5 nm, and is fully charged to 2.0 V, what is the required minimum depth of a rectangular-trench capacitor?

SOLUTION:

Capacitance per unit area is

$$C' = \frac{C}{A} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{3.9 \times 8.85 \times 10^{-14}}{50 \times 10^{-8}} = 6.9 \times 10^{-7} \text{ F/cm}^2$$

Note: Capacitance, $C = \frac{\varepsilon A}{d}$

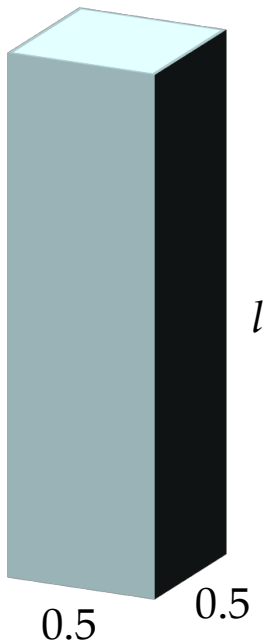
The total area is $A = (0.5 \times 0.5) + 4(0.5 \times l) = 0.25 + 2l \ \mu\text{m}^2$

As $Q = CV$, we have $C = \frac{Q}{V} = \frac{10^5 \times 1.6 \times 10^{-19}}{2.0} = 8 \times 10^{-15} \text{ F}$

Therefore, $C = C'A = 8 \times 10^{-15} = 6.9 \times 10^{-7} \left[(0.25 + 2l) \times 10^{-8} \right]$

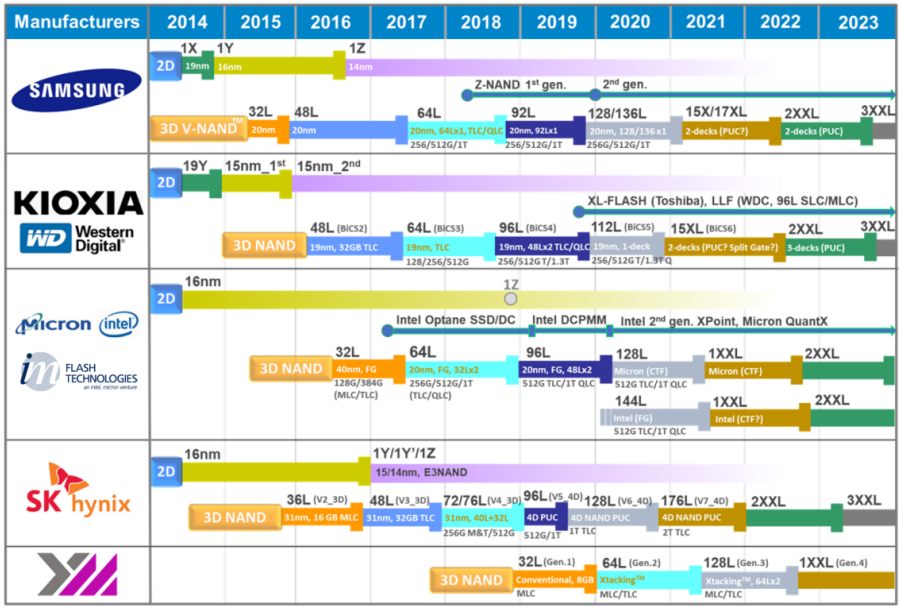
$$= 6.9 \times 10^{-15} (0.25 + 2l)$$

$$\therefore \ l = \frac{1.16 - 0.25}{2} = 0.46 \ \mu\text{m}$$



Floating gate memory device

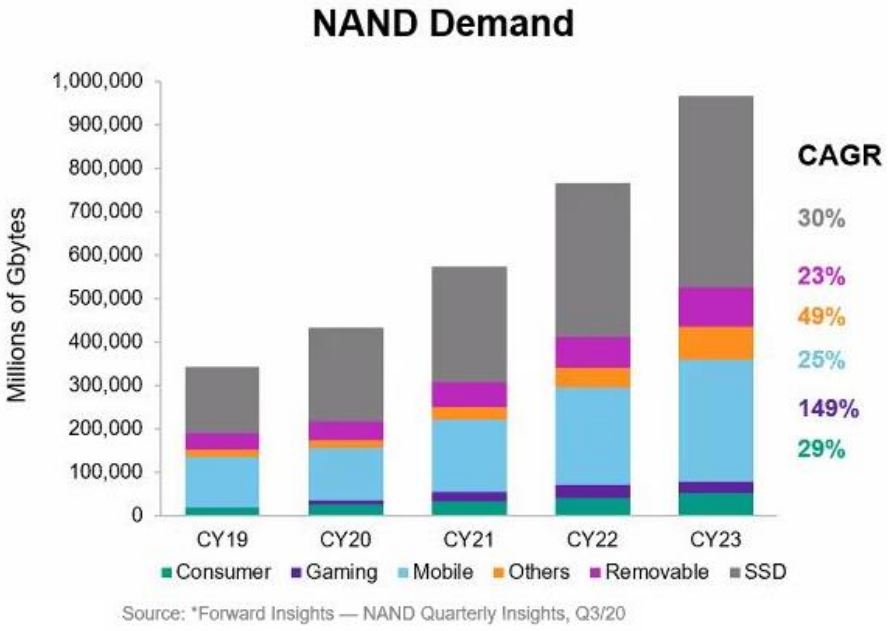
NAND Technology Industry Roadmap



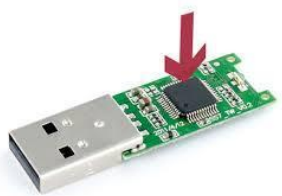
Q4/2019 updated

Tech Insights

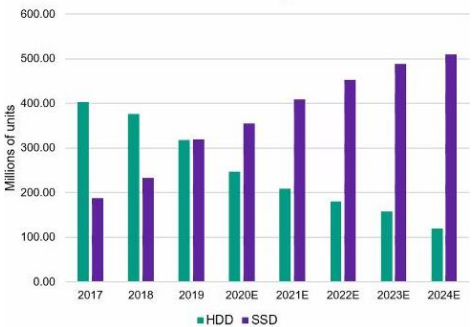
NAND Demand



Source: *Forward Insights — NAND Quarterly Insights, Q3/20



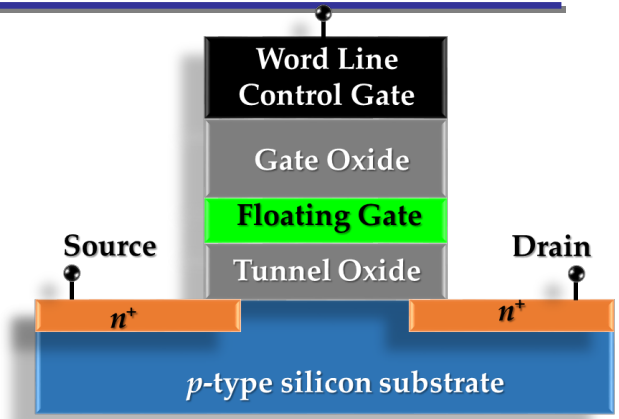
SSD vs HDD Shipments



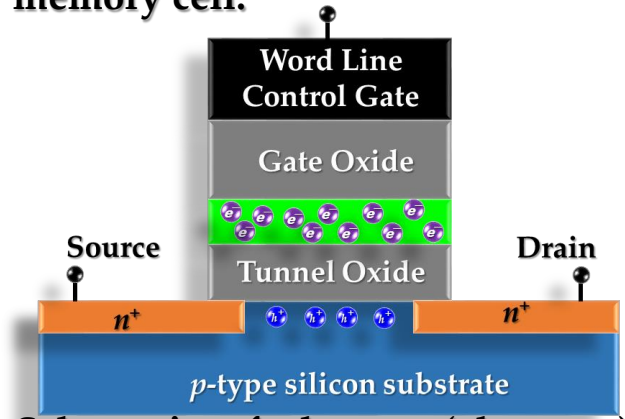
Source: Gartner — HDD vs. SSD Units, Sept. 2020

Floating gate memory operation – 1

- Toshiba invented floating gate memory structure in 1987. The memory cell consists of one MOS transistor with a floating gate, which is a charge storage layer (doped polysilicon) that is isolated from the control gate and semiconductor substrate.
- The floating gate transistor can have two threshold voltages corresponding to the presence of stored charges at the floating gate.
- When electrons are injected and accumulated at the floating gate, they attract the holes in the *p*-type channel and body to the channel region right below the floating gate. The increased density of holes changes the threshold voltage of the transistor, which becomes ***higher***. The memory cell cannot be turned on with a small read signal voltage applied to the control gate hence *the memory cell is considered to be in a "1" state* as a convention.
- *The threshold voltage of the transistor can be lowered by removing stored charges from the floating gate and the memory cell is regarded to be in a "0" state*.



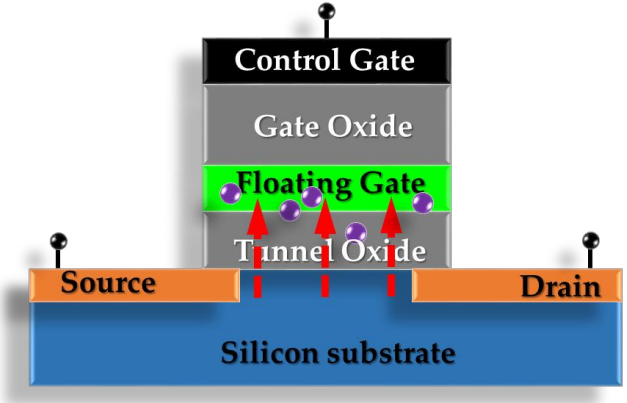
Schematic of a floating gate memory cell.



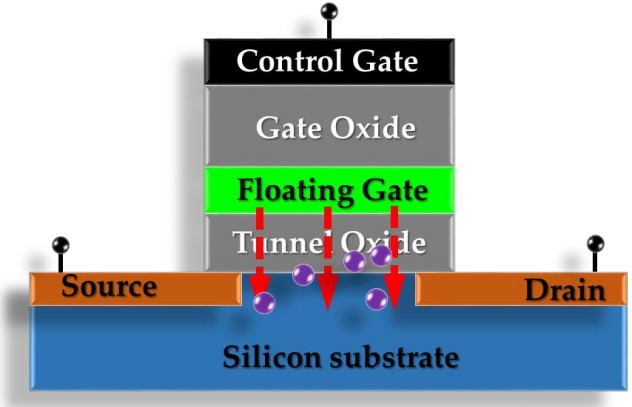
Schematic of charges (electrons) are trapped and stored at the floating gate memory cell.

Floating gate memory operation – 2

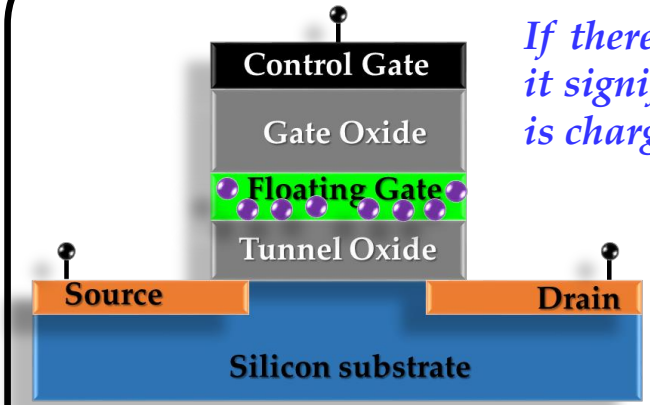
To write a cell



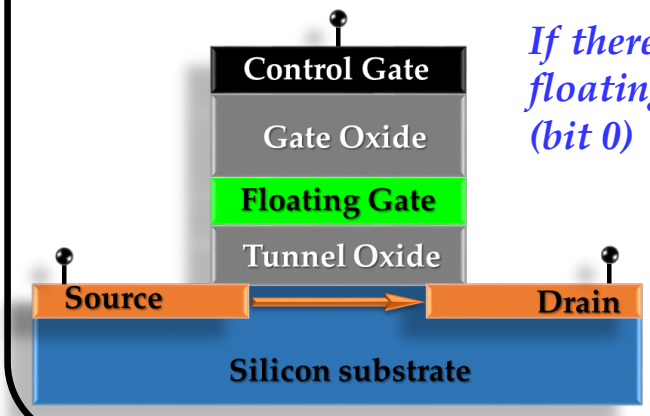
To erase a cell



To read a cell



If there is no current flow, it signifies the floating gate is charged (bit 1)

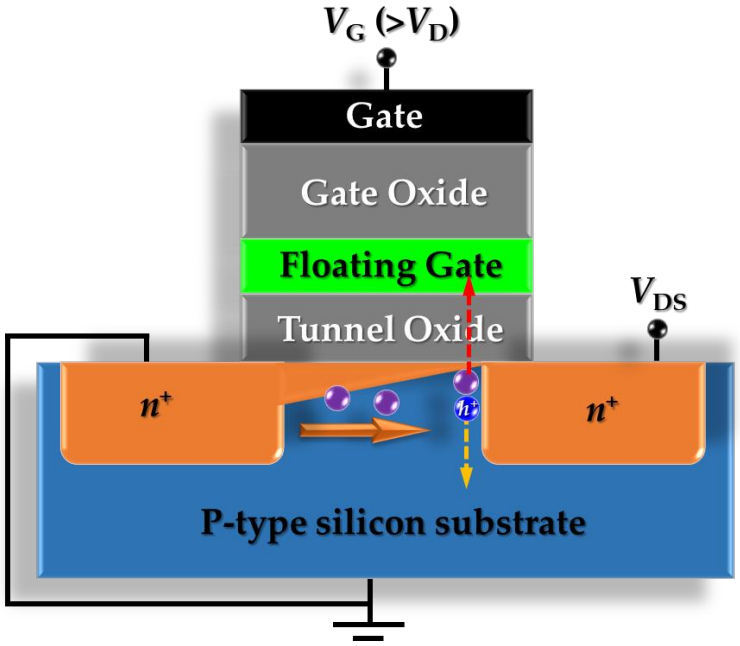


If there is current flow, the floating gate is not charged (bit 0)

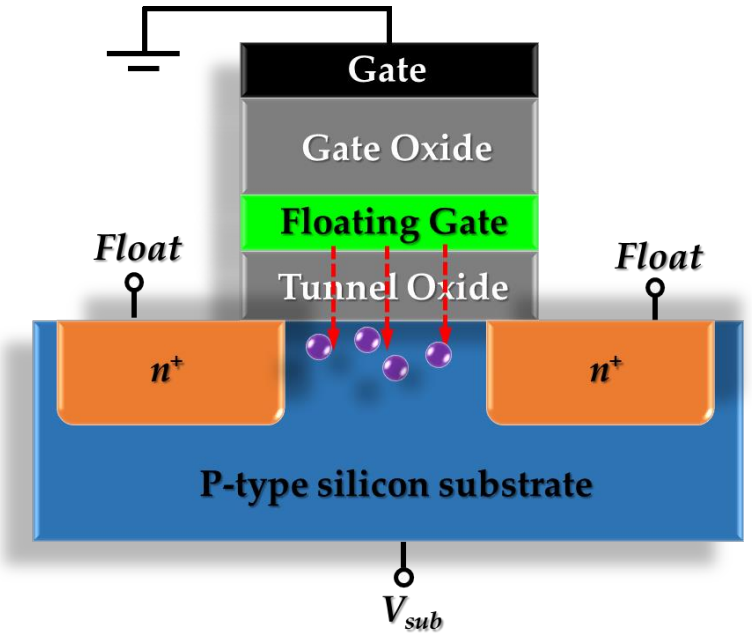
Floating gate memory operation – 3

- The ***information of flash memory is programmed by either storing or ejecting electrons in the floating gate of a MOS transistor***.
- The charge transfer into or out from the floating gate can be done via two modes : (1) ***hot-carrier injection*** and (2) ***Fowler-Nordheim tunnelling***.
- In the mechanism of hot-carrier injection the channel carriers acquire energy from the lateral field near the drain and become hot carriers. During the charging (programming) operation, a relatively ***high voltage (e.g., 12V) is applied to the control gate and across the drain*** to source (*e.g.*, 6V), electrons are heated by the high lateral electric field. ***Avalanche breakdown occurs at the near of the drain and electron-hole pairs are generated by the impact ionization***.
- The high voltage on the control gates attracts and injects electrons into floating gate through the oxide and the holes flow to the substrate as the substrate current. Secondary hot electrons generated can also be injected into the floating gate.
- During the erase operation the charge in the floating gate is removed by the tunneling current through the oxide with a high electric field (*e.g.*, >10 MV/cm).
- When ***the control gate is grounded*** (*i.e.* 0 V), and ***a high voltage is applied to the substrate***, while leaving the source and drain floated, electrons at the floating gate are ejected into the substrate by the tunneling effect.

Floating gate memory operation – 4



Charging/programming the floating gate via hot electron injection mechanism.



Erasing the charge in the floating gate via the Fowler-Nordheim tunnelling mechanism.

Threshold voltage shift

- After charging, the total stored charge Q is equal to the integrated injection current since the gate is floating. This causes a shift of the threshold voltage by the amount

Threshold voltage shift

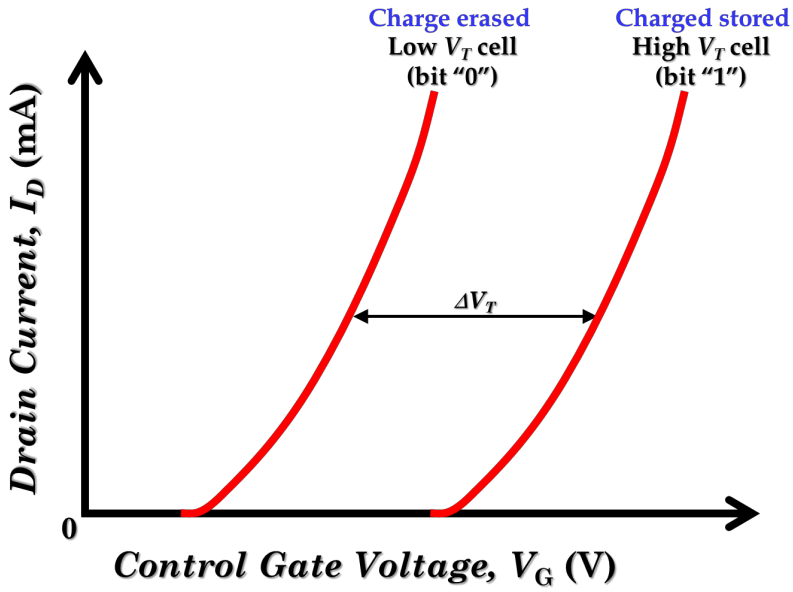
$$\Delta V_T = -\frac{d_2 \Delta Q}{\varepsilon_2} \qquad \textbf{(8a.1)}$$

Note: $CV = Q$

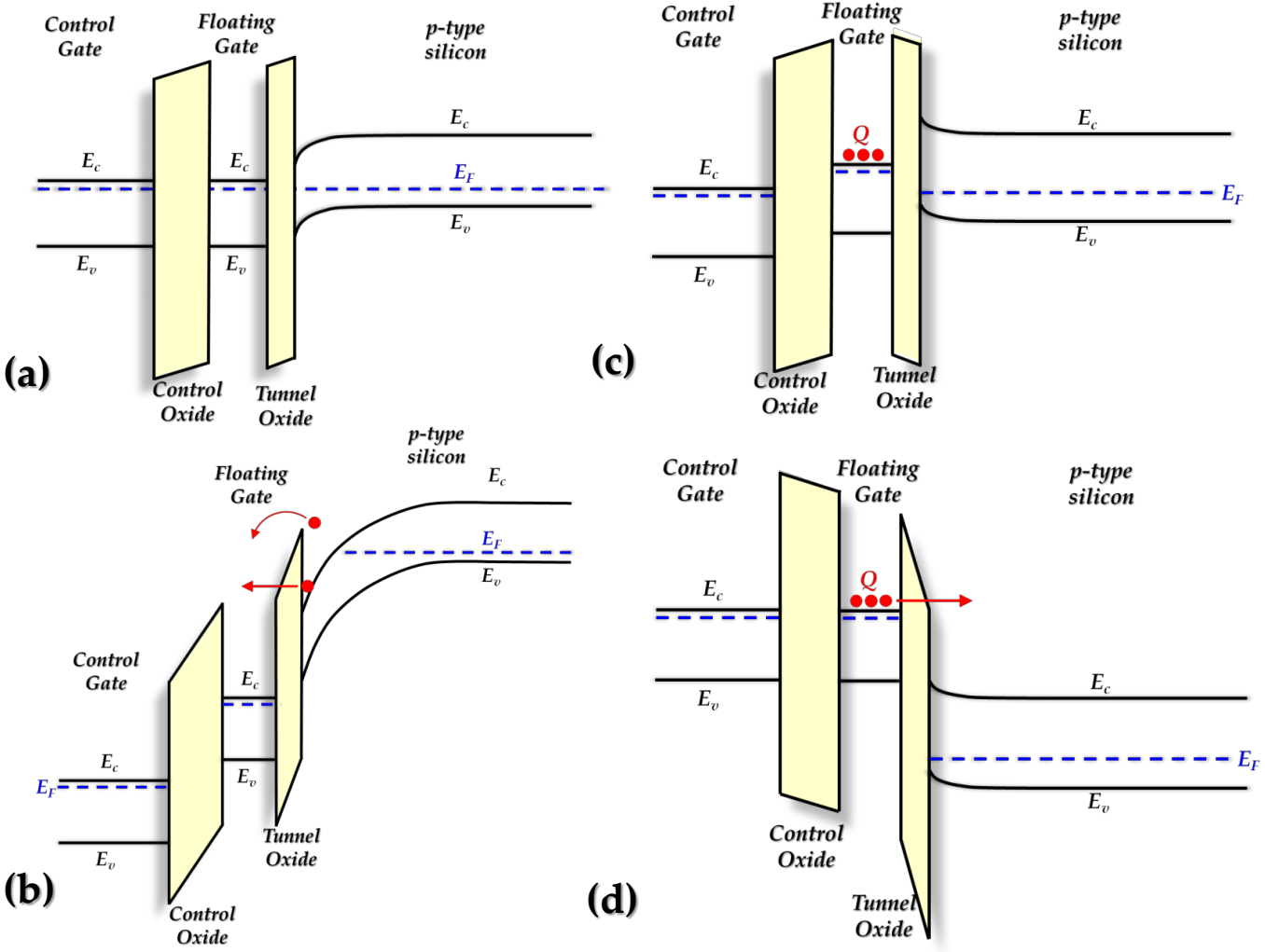
$$\Delta V = \frac{\Delta Q}{C} = \frac{t}{\varepsilon}\Delta Q$$

where ΔQ is the stored charge

- This threshold voltage shift can be directly measured from the ***I_D vs V_G plot***. After altering the charge on the floating gate by Q (negative charge), the plot shifts by ΔV_T.
- A control gate voltage is set to be large enough to turn on the low V_T transistor but insufficient to turn on the high V_T transistor.
- The high and low threshold voltage provide a mean to retrieve the stored information in the transistor.



Energy band diagram of floating gate memory operation – 1



Energy-band diagrams for a stacked-gate memory transistor at different stages of operation.

(a) Initial stage.

(b) Charging by hot electrons or electron tunnelling.

(c) After charging, the floating-gate having charge Q (negative) is at higher potential and V_T is increased.

(d) Erasing by electron tunnelling.

Energy band diagram of floating gate memory operation – 2

- The programming and erasing sequence of a floating-gate memory can be understood with the energy-band diagrams:
  - In (b), electron injection is possible can be due to hot carriers over the barrier, or tunnelling through the barrier.
  - In (c), the accumulated negative charge at the floating gate raises the threshold voltage compared to its initial condition in Fig.(a).
  - In (d), the erase is carried out by electron tunnelling from the floating gate back to the substrate.
- In both programming and erasing operations, it is important to modulate the floating-gate potential efficiently by the control-gate applied voltage.

Types of flash memory: NAND and NOR

- A flash memory chip comprises a large number of MOSFET memory cells with floating gates on a silicon wafer. The cells are linked by so-called word lines, bit lines, and source lines.
- Generally, there are two types of flash memory architecture, *i.e.*, the NOR type and NAND type, which employ similar floating gate cell structure but differ in line connection.
- In a NOR type chip, the ***source line and bit line are connected individually to each cell***, which makes it possible to perform write operations in 1-bit units.
- By contrast, in a NAND type chip, ***multiple cells are connected in series*** between the source line and bit line.
- The width of each connecting line is nanometers range and the cells are also very small. However, in the NOR type, the lines require a comparatively large amount of space in a where each cell has its own source line connection hence it has limitation in density.
- A NAND type chip on the other hand allows higher density because a source line is shared by multiple cells.