Outline

- Introduction – *"Is there a limit?"*
- Transistors – *"CMOS building blocks"*
- Parasitics – *"The [un]desirables"*
- The CMOS inverter – *"A masterpiece"*
- Gates – *"Just like LEGO"*
- Sequential circuits – *"Time also counts!"*
- Storage elements – *"A bit in memory"*
- Technology scaling – *"…, faster!"*
- Technology – *"Building an inverter"*

“A bit in memory”

Memory cells:

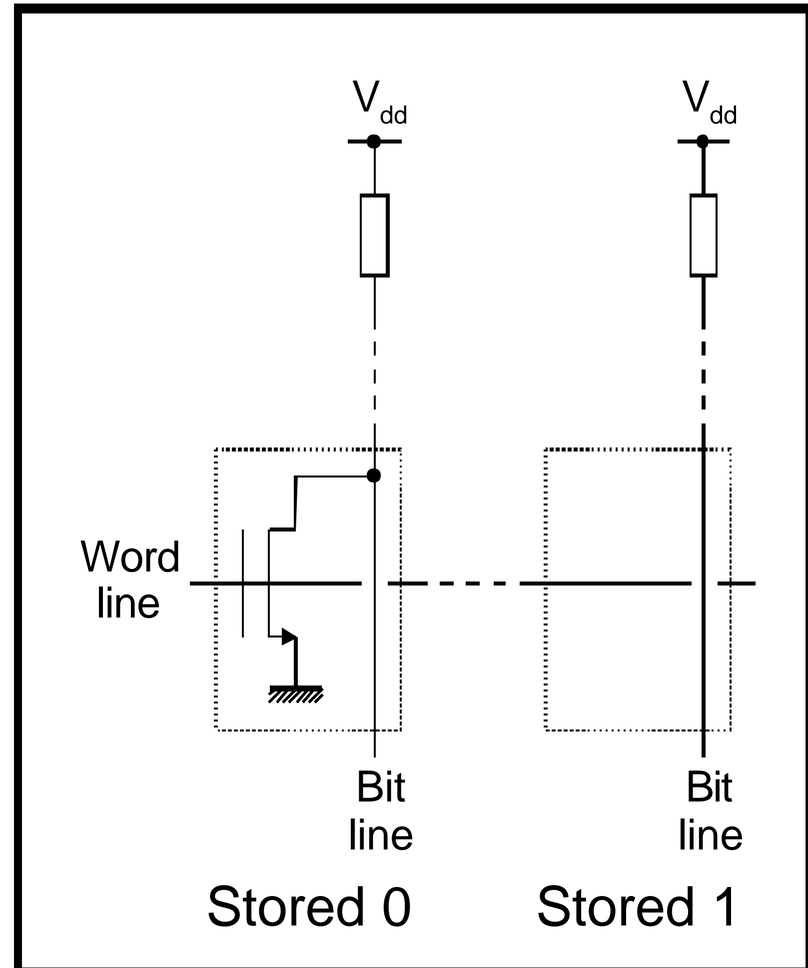
- Read-only
- Nonvolatile R/W
- Read-write
  - 6T SRAM
  - Resistive load SRAM
  - 3T dynamic
  - 1T dynamic

Storage elements

- The silicon area of large memory cells is dominated by the size of the memory core, it is thus crucial to <u>keep the size of the basic storage cell as small as possible</u>
- Reduced storage cell area results in:
  - reduced driving capability (small devices)
  - reduced logic swing and the noise margins
  - Consequently, sense amplifiers are necessary to restore full rail-to-rail amplitude

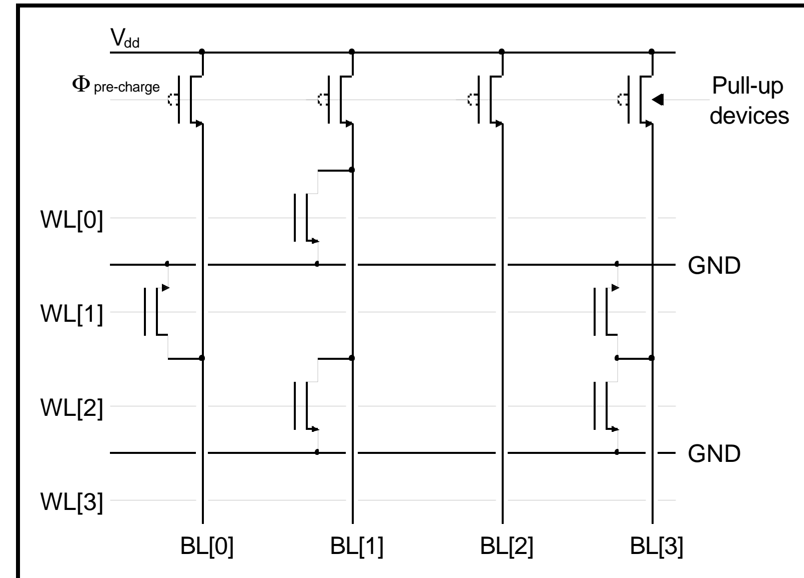
Read-only

- Because the contents is permanently fixed the cell design is simplified
- Upon activation of the word line a 0 or 1 is presented to the bit line:
  - If the NMOS is absent the word line has no influence on the bit line:
    - The word line is pulled-up by the resistor
    - A 1 is stored in the "cell
  - If the NMOS is present the word line activates the NMOS:
    - The word line is pulled-down by the NMOS
    - A 0 is stored in the cell



Read-only

- In practice a “always on” pull-up device is never used because:
  - V_{OL} would depend on the ratio of the pull-up/pull-down devices
  - A static current path would exists when the output is low causing high power dissipation in large memories
- In practice pre-charged logic is used:
  - Eliminates the static dissipation
  - Pull-up devices can be made wider



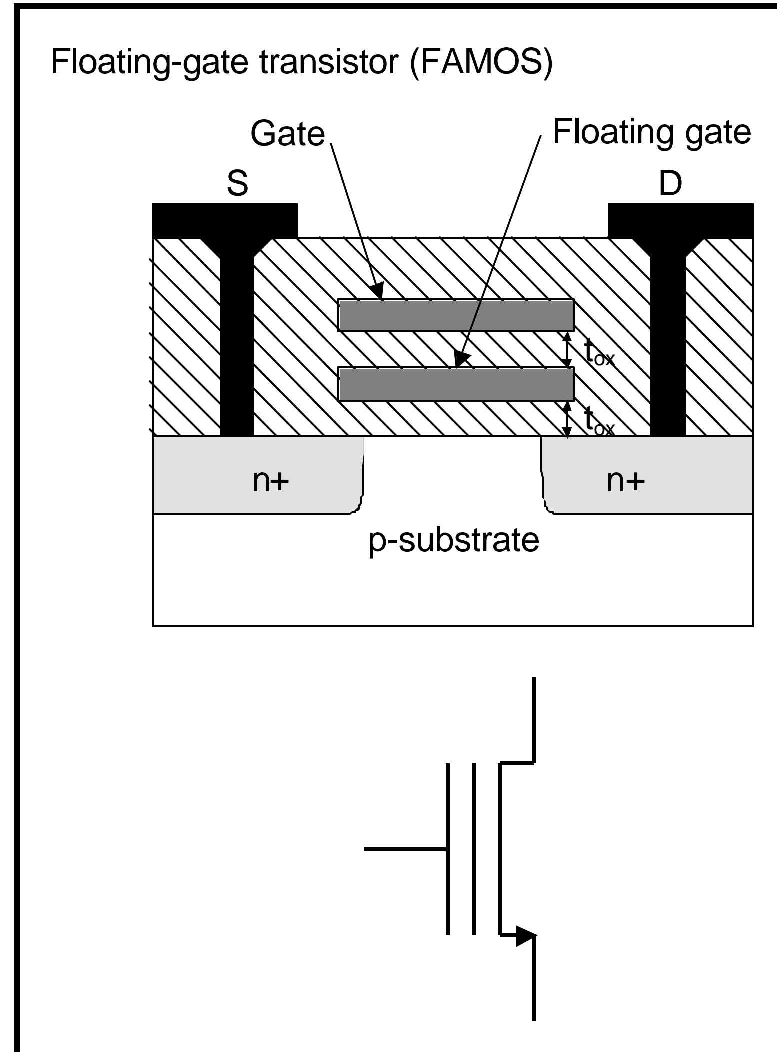
- The bit lines are first pre-charged by the pull-up devices
  - during this phase the word lines must be disabled
- Then, the word lines are activated (word evaluation)
  - during this phase the pull-up devices are off

Nonvolatile R/W

- The same architecture as a ROM memory
- The pull-down device is modified to allow control of the threshold voltage
- The modified threshold is retained "indefinitely":
  - The memory is nonvolatile
- To reprogram the memory the programmed values must be erased first
- The "hart" of NVRW memories is the Floating Gate Transistor (FAMOS)

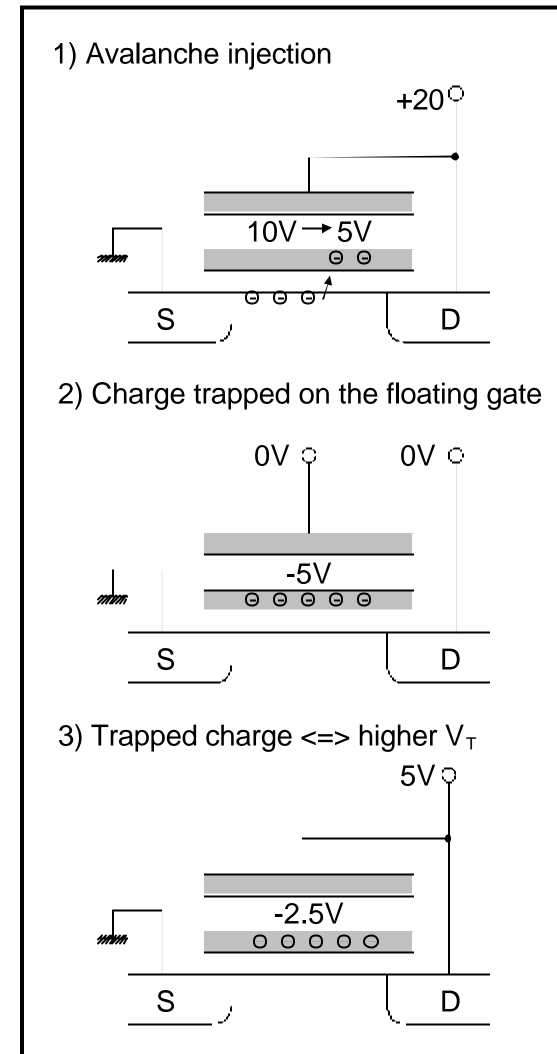
Nonvolatile R/W

- A floating gate is inserted between the gate and the channel
- The device acts as a normal transistor
- However, its threshold voltage is programmable
- Since the t_{ox} is doubled, the transconductance is reduced to half and the threshold voltage increased



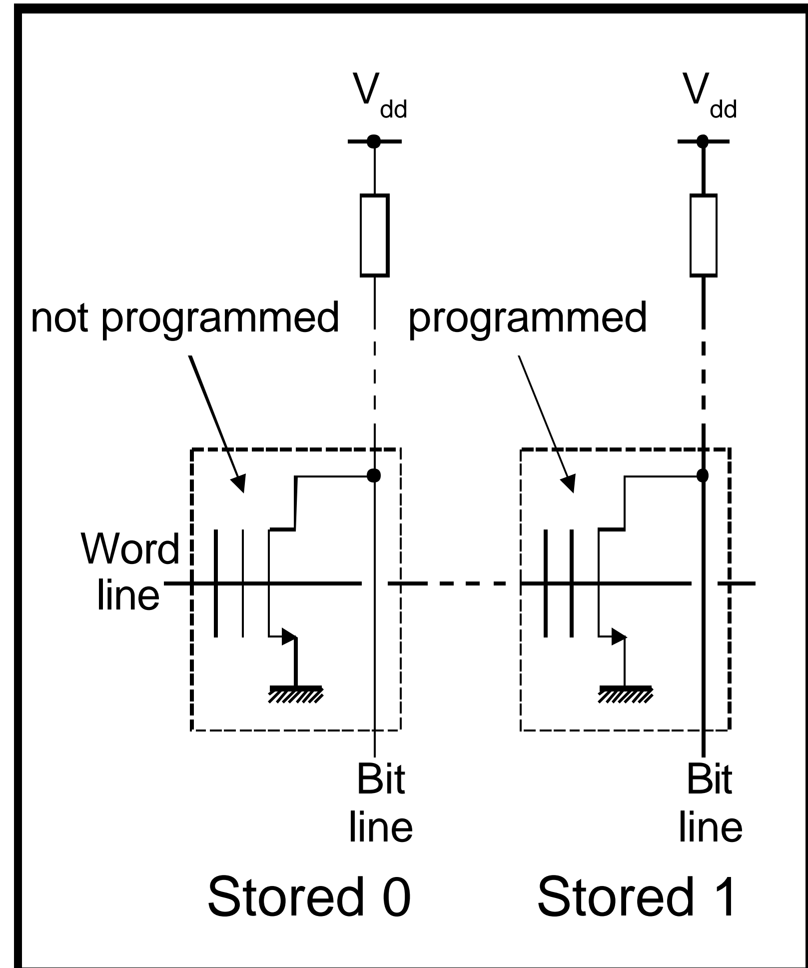
Nonvolatile R/W

- Programming the FAMOS:
  - A high voltage is applied between the source and the gate-drain
  - A high field is created that causes avalanche injection to occur
  - Electrons traverse the first oxide and get trapped on the floating gate (t_{ox} = 100nm)
  - Trapped electrons effectively drop the floating gate voltage
  - The process is self limiting: the building up of gate charge eventually stops avalanche injection
  - The FAMOS with a charged gate is equivalent to a higher V_T device
  - Normal circuit voltages can not turn a programmed device on



Nonvolatile R/W

- The non-programmed device can be turned on by the word line thus, it stores a “0”
- The word line high voltage can not turn on the programmed device thus, it stores a “1”
- Since the floating gate is surrounded by SiO_2, the charge can be stored for many years

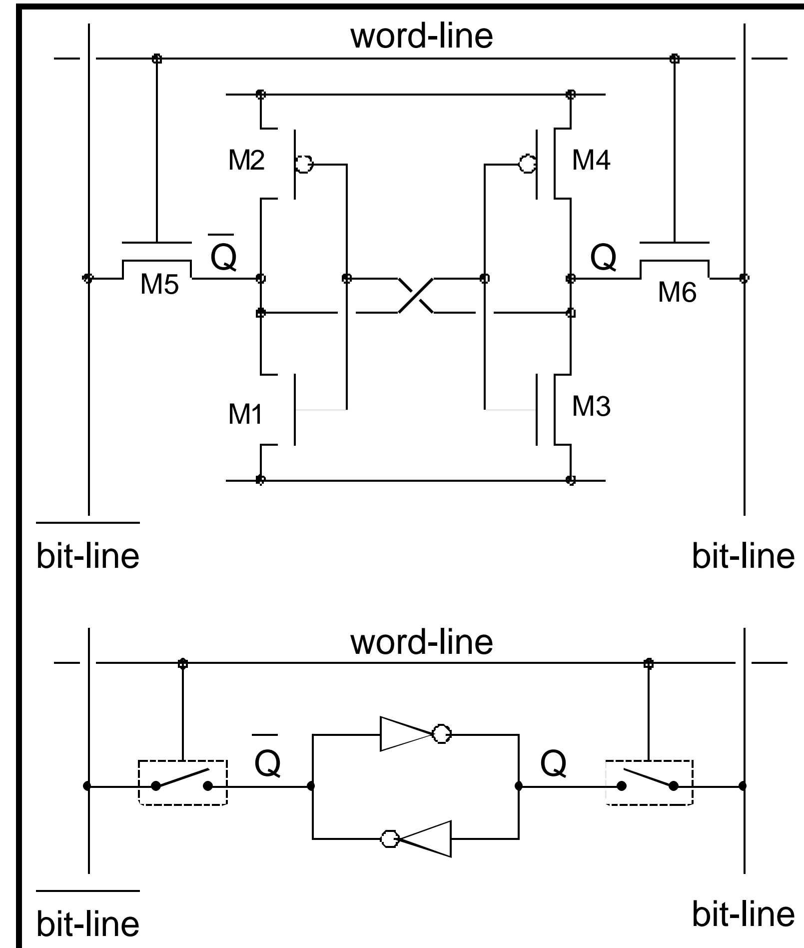


Nonvolatile R/W

- Erasing the memory contents (EPROM):
  - Strong UV light is used to erase the memory:
    - UV light renders the oxide slightly conductive by direct generation of electron-hole pairs in the SiO_2
  - The erasure process is slow (several minutes)
  - Programming takes 5-10μs/word
  - Number of erase/program cycles limited (<1000)
- Electrically-Erasable PROM (E^2PROM)
  - A reversible tunneling mechanism allows E^2PROM's to be both electrically programmed and erased

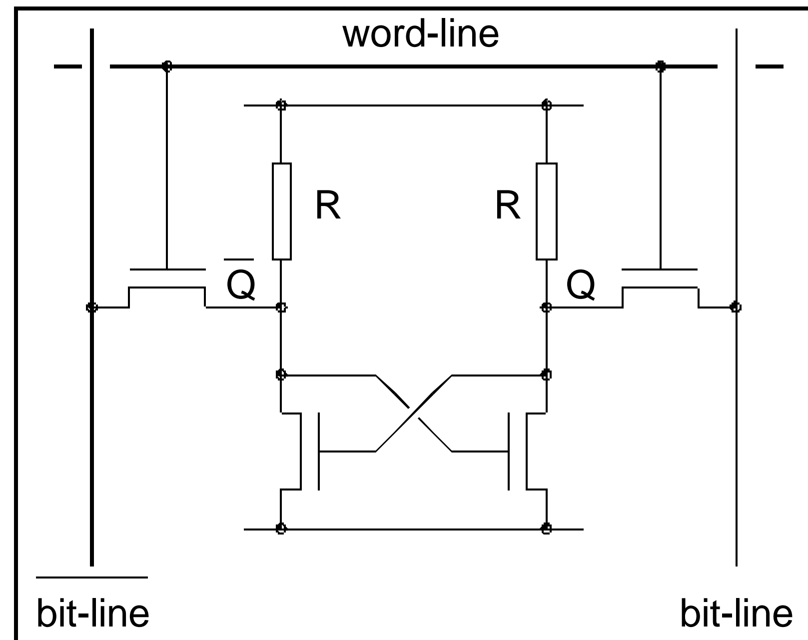
6T SRAM

- Static Read-Write Memories (SRAM):
  - data is stored by positive feedback
  - the memory is volatile
- The cell uses six transistors
- Read/write access is enabled by the word-line
- Two bit lines are used to improve the noise margin during the read/write operation
- During read the bit-lines are pre-charged to $V_{dd}/2$:
  - to speedup the read operation
  - to avoid erroneous toggling of the cell



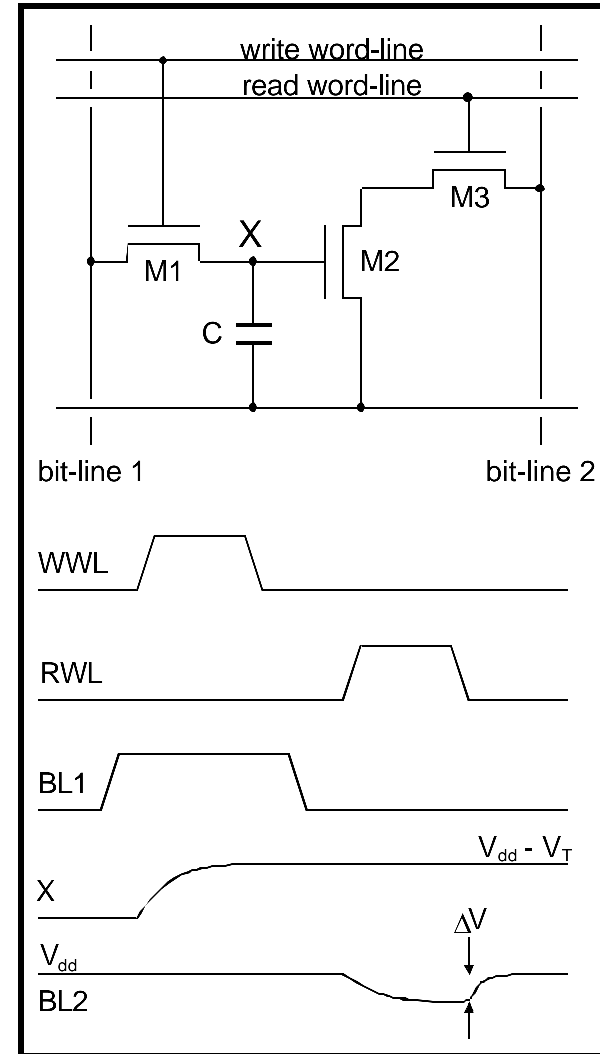
Resistive-load SRAM

- Resistive-load SRAM
  - employs resistors instead of PMOS's
  - The role of the resistors is only to maintain the state of the cell:
    - they compensate for leakage currents (10^{-15}A)
    - they must be made as high as possible to minimize static power dissipation
    - undoped polysilicon $10^{12}\Omega$/③
  - The bit-lines are pre-charged to V_{dd}:
    - the low-to-high transition occurs during precharge
    - the loads contribute "no" current during the transitions
  - The transistor sizes must be correctly chosen to avoid toggling the cell during read



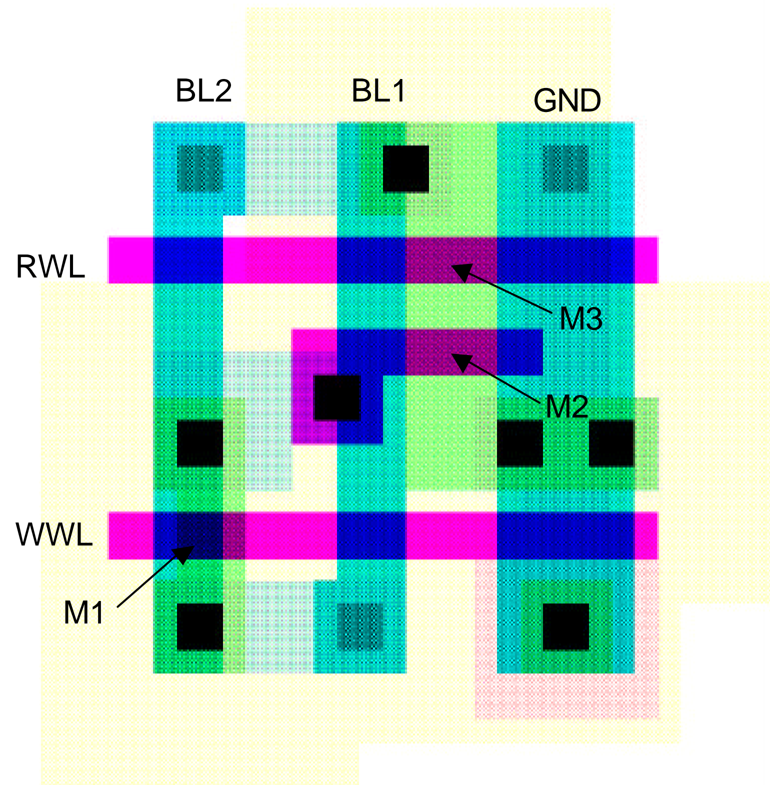
3T Dynamic

- Dynamic Random-Access Memory (DRAM)
  - In a dynamic memory the data is stored as charge in a capacitor
- Tree-Transistor Cell (3T DRAM):
  - Write operation:
    - Set the data value in bit-line 1
    - Assert the write word-line
    - Once the WWL is lowered the data is stored as charge in C
  - Read operation:
    - The bit-line BL2 is pre-charged to Vdd
    - Assert the read word-line
    - if a 1 is stored in C, M2 and M3 pull the bit-line 2 low
    - if a 0 is stored C, the bit-line 2 is left unchanged



3T Dynamic

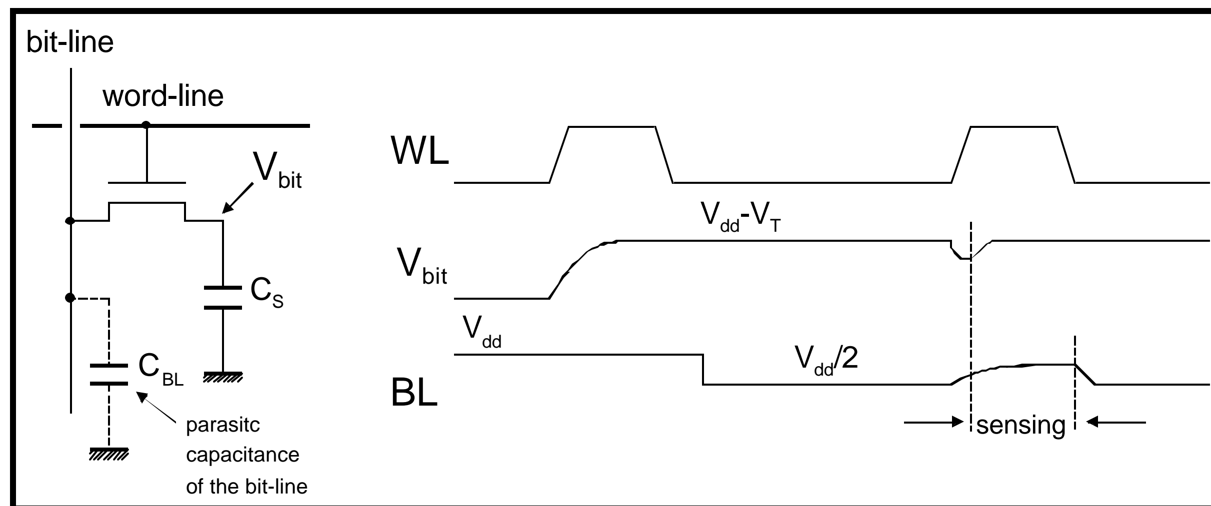
- The cell is inverting
- Due to leakage currents the cell needs to be periodically refreshed (every 1 to 4ms)
- Refresh operation:
  - read the stored data
  - put its complement in BL1
  - enable/disable the WWL
- Compared with an SRAM the area is greatly reduce:
  - SRAM ⇒ 1092 λ^2
  - DRAM ⇒ 576 λ^2
  - The area reduction is mainly due to the reduction of the number of devices and interlayer contacts



(from J. M. Rabaey 1996)

1T Dynamic

- One-Transistor dynamic cell (1T DRAM)
  - It uses a single transistor and a capacitor
  - It is the most widely used topology in commercial DRAM's
- Write operation:
  - Data is placed on the bit-line
  - The word-line is asserted
  - Depending on the data value the capacitance is charged or discharged

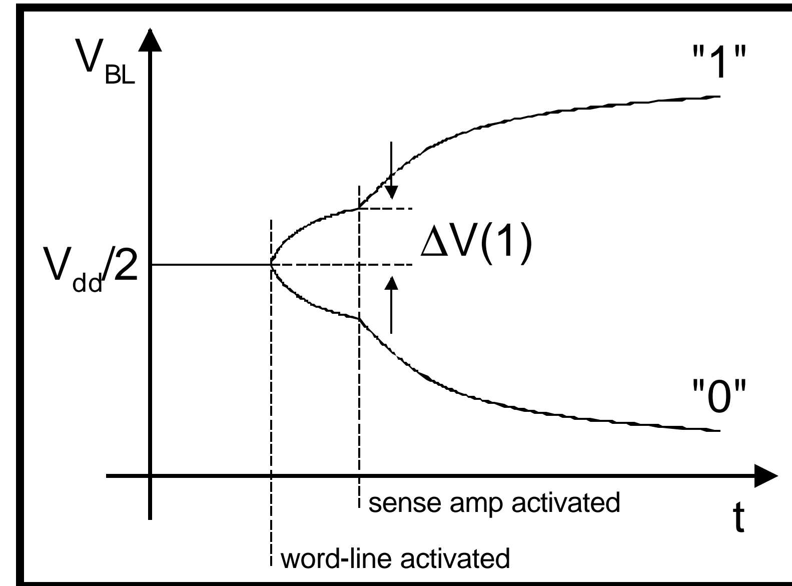


1T Dynamic

- Read operation:
  - The bit-line is pre-charged to $V_{dd}/2$
  - The word-line is activated and charge redistribution takes place between C_S and the bit-line
  - This gives origin to a voltage change in the bit-line, the sign of which determines the data stored:

$$\Delta V = \left(V_{BIT} - \frac{V_{dd}}{2}\right)\frac{C_S}{C_S + C_{BL}}$$

  - C_{BL} is 10 to 100 times bigger than $C_S \Rightarrow$ $\Delta V \cong 250mV$

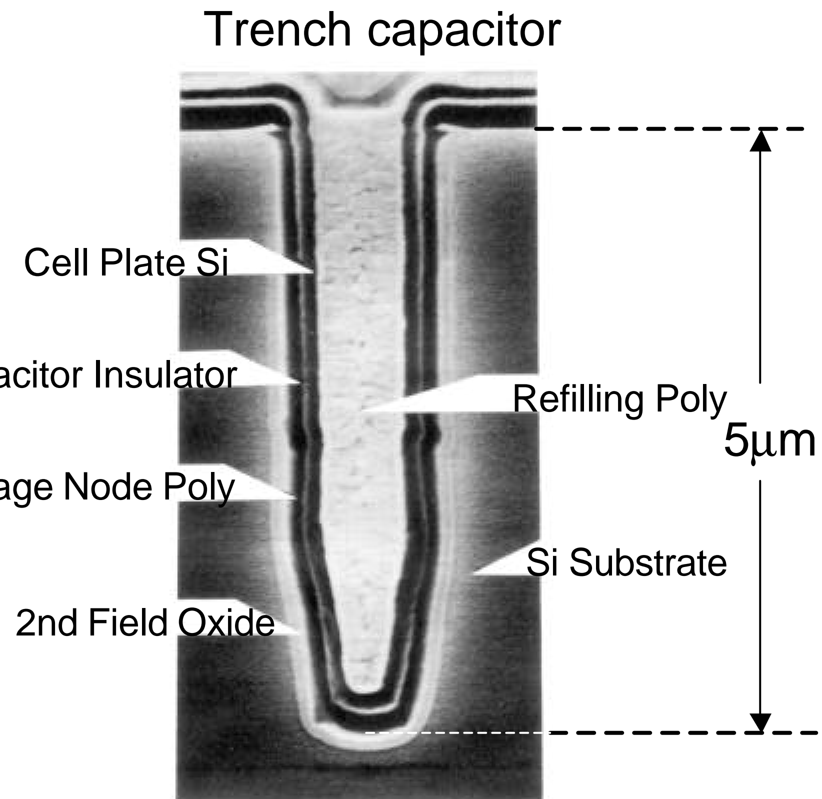


  - The amount of charge stored in the cell is modified during the read operation
  - However, during read, the output of the sense amplifier is imposed on the bit line restoring the stored charge

1T Dynamic

- Contrary to the previous cases a 1T cell requires a sense amplifier for correct operation
- Also, a relatively large storage capacitance is necessary for reliable operation
- A 1 is stored as V_{dd}-V_T. This reduces the available charge:
  - To avoid this problem the word-line can be bootstrapped to a value higher than Vdd

Trench capacitor



(from T. Mano et al., 1987)

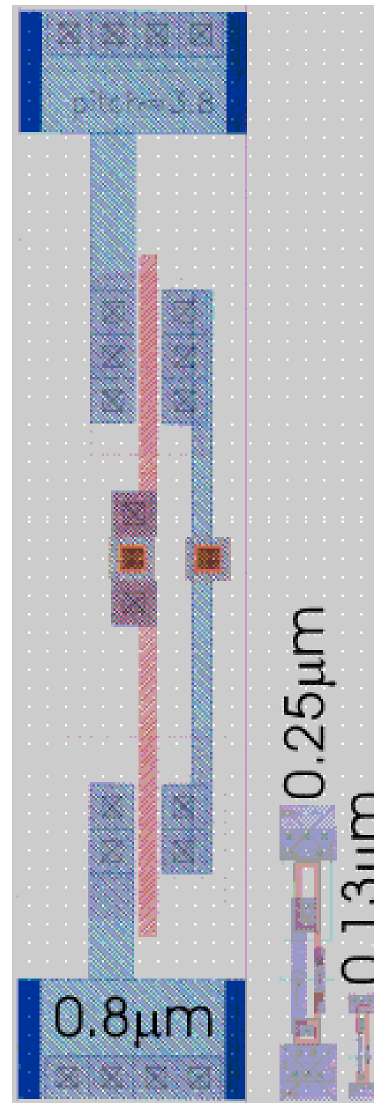
Outline

- Introduction – *“Is there a limit?”*
- Transistors – *“CMOS building blocks”*
- Parasitics – *“The [un]desirables”*
- The CMOS inverter – *“A masterpiece”*
- Gates – *“Just like LEGO”*
- Sequential circuits – *“Time also counts!”*
- Storage elements – *“A bit in memory”*
- Technology scaling – *“…, faster!”*
- Technology – *“Building an inverter”*

Technology scaling

- Scaling objectives
- Scaling variables
- Scaling consequences:
  - Device area
  - Transistor density
  - Gate capacitance
  - Drain current
  - Gate delay
  - Power
  - Power density
  - Interconnects

Scaling, why is it done?



Technology scaling

- Technology scaling has a threefold objective:
  - Increase the transistor density
  - Reduce the gate delay
  - Reduce the power consumption
- At present, between two technology generations, the objectives are:
  - Doubling of the transistor density;
  - Reduction of the gate delay by 30% (43% increase in frequency);
  - Reduction of the power by 50% (at 43% increase in frequency);

Technology scaling

- How is scaling achieved?
  - All the device dimensions (lateral and vertical) are reduced by $1/\alpha$
  - Concentration densities are increased by α
  - Device voltages reduced by $1/\alpha$ (not in all scaling methods)
  - Typically $1/\alpha = 0.7$ (30% reduction in the dimensions)

Technology scaling

- The **scaling variables** are:
  - Supply voltage: $V_{dd} \rightarrow V_{dd} / \alpha$
  - Gate length: $L \rightarrow L / \alpha$
  - Gate width: $W \rightarrow W / \alpha$
  - Gate-oxide thickness: $t_{ox} \rightarrow t_{ox} / \alpha$
  - Junction depth: $X_j \rightarrow X_j / \alpha$
  - Substrate doping: $N_A \rightarrow N_A \times \alpha$

  This is called **constant field** scaling because the electric field across the gate-oxide does not change when the technology is scaled

  If the power supply voltage is maintained constant the scaling is called **constant voltage**. In this case, the electric field across the gate-oxide increases as the technology is scaled down.

Due to gate-oxide breakdown, below 0.8µm only "constant field" scaling is used.

Scaling consequences

Some consequences of 30% scaling in the constant field regime ($\alpha = 1.43$, $1/\alpha = 0.7$):

- Device/die area:

$$W \times L \rightarrow (1/\alpha)^2 = 0.49$$

  - In practice, microprocessor <u>die size grows</u> about 25% per technology generation! This is a result of added functionality.

- Transistor density:

$$(\text{unit area}) / (W \times L) \rightarrow \alpha^2 = 2.04$$

  - In practice, <u>memory density</u> has been scaling as expected. (not true for microprocessors…)

Scaling consequences

- Gate capacitance:

$$W \times L / t_{ox} \rightarrow 1/\alpha = 0.7$$

- Drain current:

$$(W/L) \times (V^2/t_{ox}) \rightarrow 1/\alpha = 0.7$$

- Gate delay:

$$(C \times V) / I \rightarrow 1/\alpha = 0.7$$

$$\text{Frequency} \rightarrow \alpha = 1.43$$

  - In practice, microprocessor frequency has doubled every technology generation (2 to 3 years)! This faster increase rate is due to highly pipelined architectures (“less gates per clock cycle”)

Scaling consequences

- Power:

$$C \times V^2 \times f \rightarrow (1/\alpha)^2 = 0.49$$

  - In reality power consumption increases due to added functionality.
- Power density:

$$1/t_{ox} \times V^2 \times f \rightarrow 1$$

- Active capacitance/unit-area:
  Power dissipation is a function of the operation frequency, the power supply voltage and of the circuit size (number of devices).
  If we normalize the power density to $V^2 \times f$ we obtain the active capacitance per unit area for a given circuit. This parameter can be compared with the oxide capacitance per unit area:

$$1/t_{ox} \rightarrow \alpha = 1.43$$

  - In practice, for microprocessors, the active capacitance/unit-area only increases between 30% and 35%. Thus, the twofold improvement in logic density between technologies is not achieved.

Scaling consequences

- Interconnects scaling:
  - Higher densities are only possible if the interconnects also scale.
  - Reduced width $\rightarrow$ <u>increased resistance</u>
  - Denser interconnects $\rightarrow$ <u>higher capacitance</u>
  - To account for <u>increased parasitics</u> and <u>integration complexity</u> **more interconnection layers** are added:
    - thinner and tighter layers $\rightarrow$ local interconnections
    - thicker and sparser layers $\rightarrow$ global interconnections and power

Interconnects are scaling as expected

Scaling consequences

Parameter	Constant Field	Constant Voltage	
Supply voltage (V_{dd})	$1/\alpha$	1	Scaling Variables
Length (L)	$1/\alpha$	$1/\alpha$	
Width (W)	$1/\alpha$	$1/\alpha$	
Gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	
Junction depth (X_j)	$1/\alpha$	$1/\alpha$	
Substrate doping (N_A)	α	α	
Electric field across gate oxide (E)	1	α	Device Repercussion
Depletion layer thickness	$1/\alpha$	$1/\alpha$	
Gate area (Die area)	$1/\alpha^2$	$1/\alpha^2$	
Gate capacitance (load) (C)	$1/\alpha$	$1/\alpha$	
Drain-current (I_{dss})	$1/\alpha$	α	
Transconductance (g_m)	1	α	
Gate delay	$1/\alpha$	$1/\alpha^2$	Circuit Repercussion
Current density	α	α^3	
DC & Dynamic power dissipation	$1/\alpha^2$	α	
Power density	1	α^3	
Power-Delay product	$1/\alpha^3$	$1/\alpha$	