united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

301/1246-2

Microprocessor Laboratory
Third Regional Course on
Advanced VLSI Design Techniques
13 November - 1 December 2000

Lima - Peru

INTRODUCTION TO VLSI ASIC DESIGN AND TECHNOLOGY

available also on

http://pcvlsi5.cern.ch/MicDig/VLSI_Trieste/VLSI_Trieste.htm

Paulo Rodrigues S. MOREIRA
CERN
EP/MIC
CH-1211 Geneva 23
SWITZERLAND

Introduction to VLSI ASIC Design and Technology

P. Moreira, CERN-EP/MIC
Geneva
Switzerland

Outline

- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

"The world is digital..."

- Analogue loses terrain:
  - Computing
  - Instrumentation
  - Control systems
  - Telecommunications
  - Consumer electronics

"...analogue will survive"

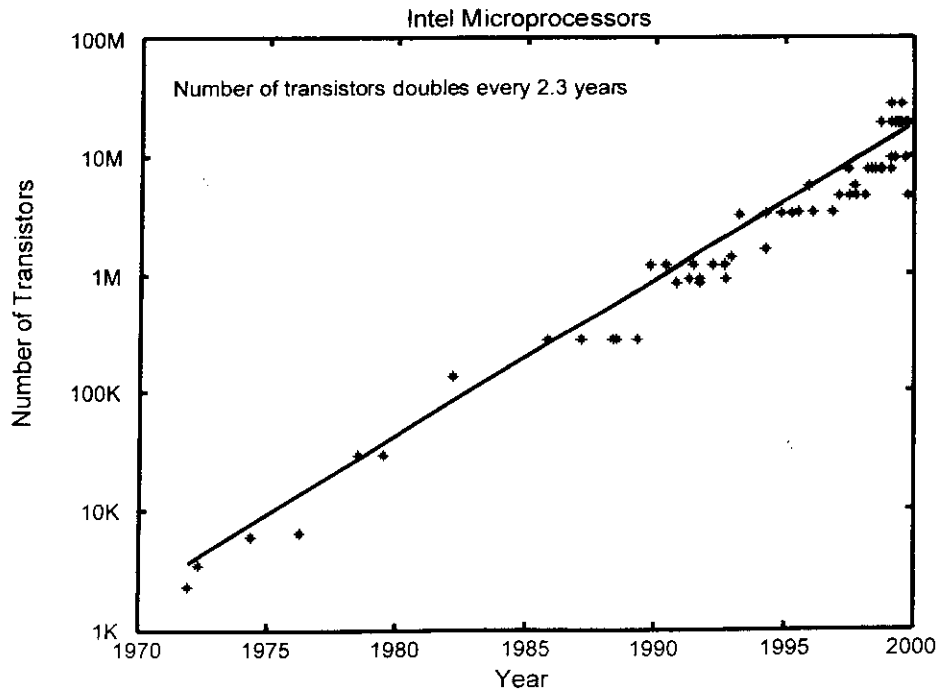
- Amplification of very week signals
- A/D and D/A conversion
- Very high frequency amplification
- Very high frequency signal processing
- As digital systems become faster and faster and circuits densities increase:
  - Analogue phenomena are becoming important in digital systems

"Moore's Law"

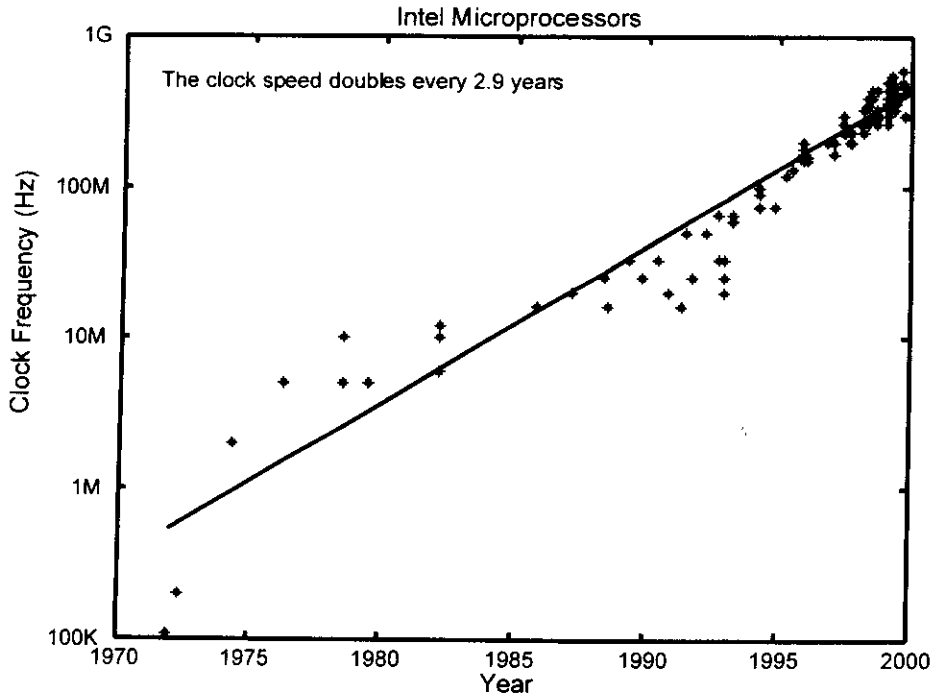
The number of transistors that can be integrated on a single IC grows exponentially with time.

"Integration complexity doubles every three years", Gordon Moore - 1965

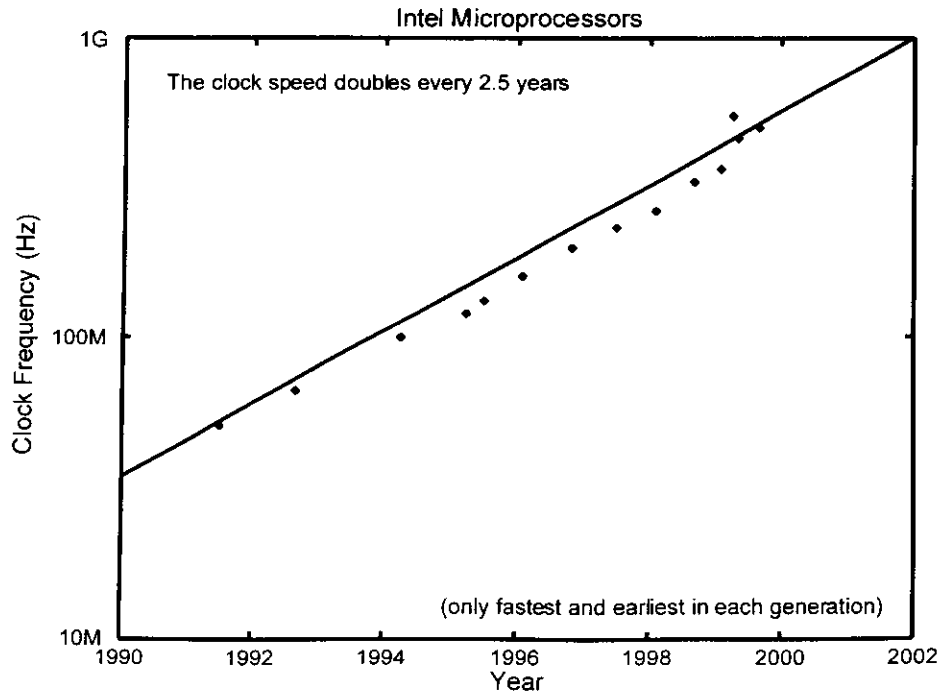
Trends in transistor count



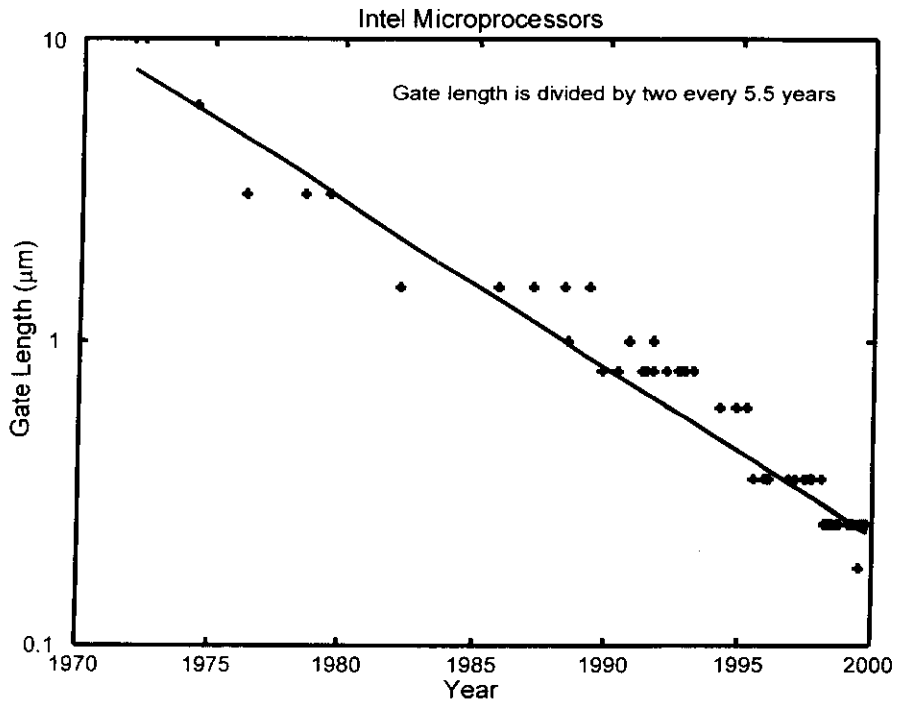
Trends in clock frequency (1)



Trends in clock frequency (2)



Trends in feature size



Driving force: Economics (1)

- Traditionally, the cost/function in an IC is reduced by 25% to 30% a year.
- To achieve this the number of functions/IC has to be increased. This demands for:
  - Increase of the transistor count
  - Decrease of the feature size (*contains the area increase and improves performance*)
  - Increase of the clock speed

Driving force: Economics (2)

- Increase productivity:
  - Increase equipment throughput
  - Increase manufacturing yields
  - Increase the number of chips on a wafer:
    - reduce the are of the chip: smaller feature size & redesign
  - Use the largest wafer size available

Example of a cost effective product (typically DRAM): the initial IC area is reduced to 50% after 3 years and to 35% after 6 years.

2001 and beyond ?

Semiconductor Industry Association (SIA) Road Map, 1998 Update

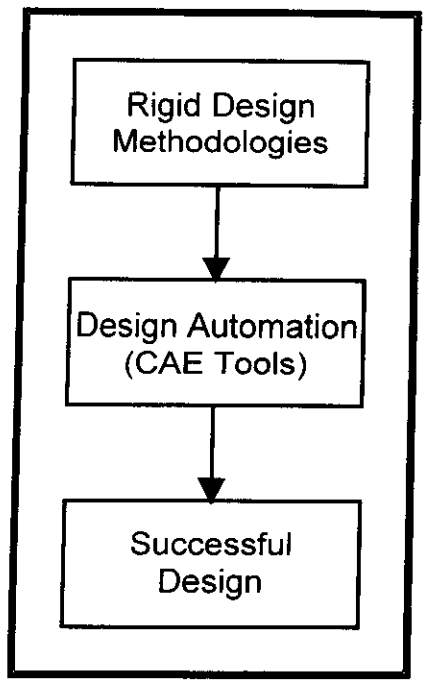
	1999	***2002***	***2014***
Technology (nm)	180	130	35
Minimum mask count	22/24	24	29/30
Wafer diameter (mm)	300	300	450
Memory-samples (bits)	1G	4G	1T
Transistors/cm² (μP)	6.2M	18M	390M
Wiring levels (maximum)	*6-7*	*7*	*10*
Clock, local (MHz)	1250	2100	10000
Chip size: DRAM (mm²)	400	560	2240
Chip size: μP (mm²)	340	430	901
Power supply (V)	1.5-1.8	1.2-1.5	0.37-0.42
Maximum Power (W)	*90*	*130*	*183*
Number of pins (μP)	700	957	3350

IEEE Spectrum, July 1999

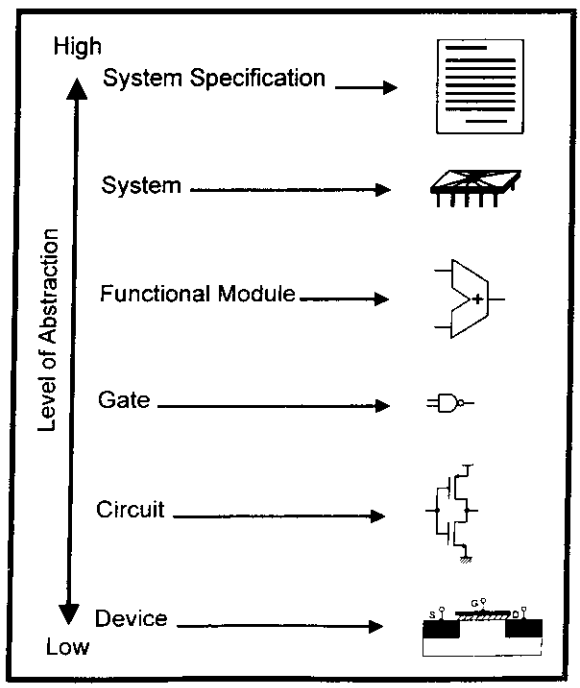
Special report: "The 100-million transistor IC"

How to cope with complexity?

- By applying:
  - Rigid design methodologies
  - Design automation



Design abstraction levels



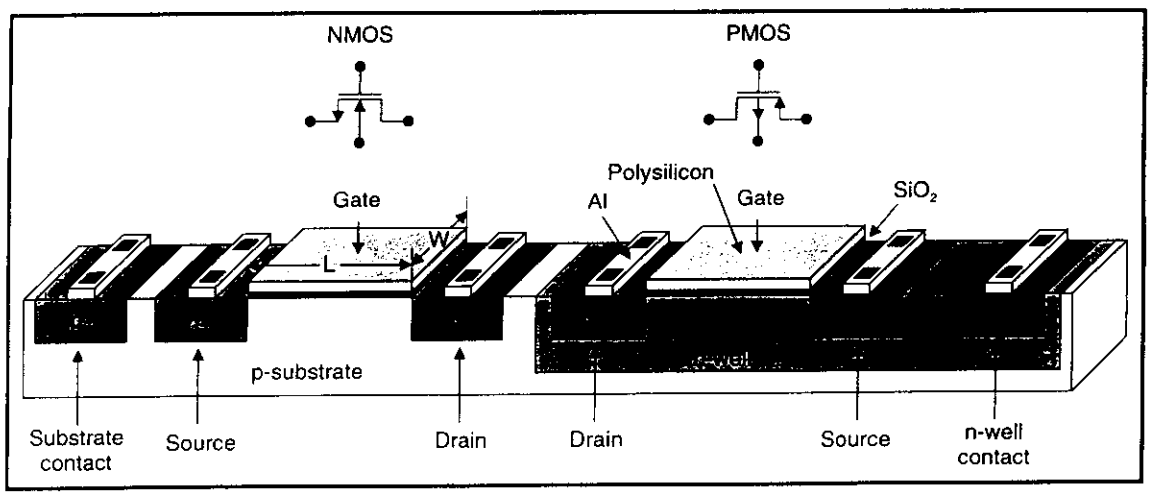
Outline

- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS devices

- CMOS devices
- pn-Junction diodes
- MOSFET equations
- What causes delay?
- MOSFET capacitances
- CMOS device hazards

CMOS devices



CMOS devices

In a CMOS process the devices are:

- PMOS FET's
- NMOS FET's

+ unwanted (but ubiquitous):

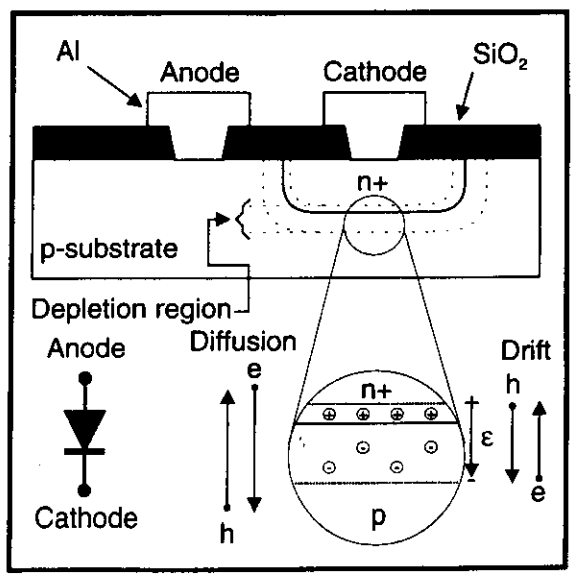
- pn-Junction diodes
- parasitic capacitance

and

- parasitic bipolars
- *parasitic inductance*

pn-Junctions diodes

- Any pn-junction in the IC forms a diode
- Majority carriers diffuse from regions of high to regions of low concentration
- The electric field of the depletion region counteracts diffusion
- In equilibrium there is no net flow of carriers in the diode



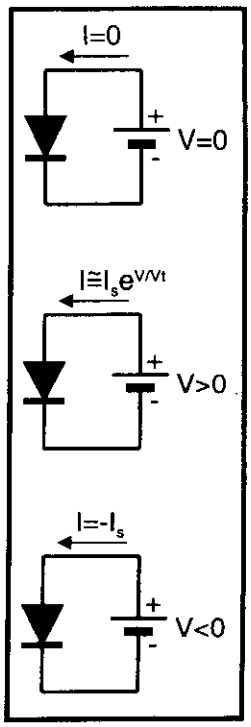
pn-Junction diodes

- Under zero bias there is a built-in potential across the junction
- The built-in potential is:

$$\phi_0 = \phi_T \cdot \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right)$$

$$\phi_T = \frac{k \cdot T}{q} \cong 26 \text{ mV } @ \text{ } 300^\circ \text{K}$$

$$n_i = 1.5 \times 10^{10} \text{ cm}^{-3} \text{ for silicon } @ \text{ } 300^\circ \text{K}$$



pn-Junction diodes

- Ideal diode equation

$$I_D = I_s \cdot \left(e^{V/\phi_T} - 1\right)$$

- For V>ϕ_T (forward bias)

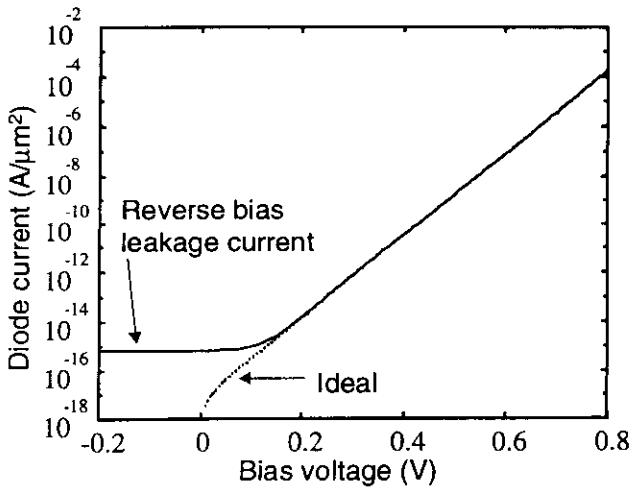
$$I_F \cong I_s \cdot e^{V/\phi_T}$$

- For V<0 (reversed bias)

$$I_R \cong -I_s$$

- In practical diodes due to thermal generation

$$I_R \cong 100 \text{ to } 1000 \times (-I_s)$$



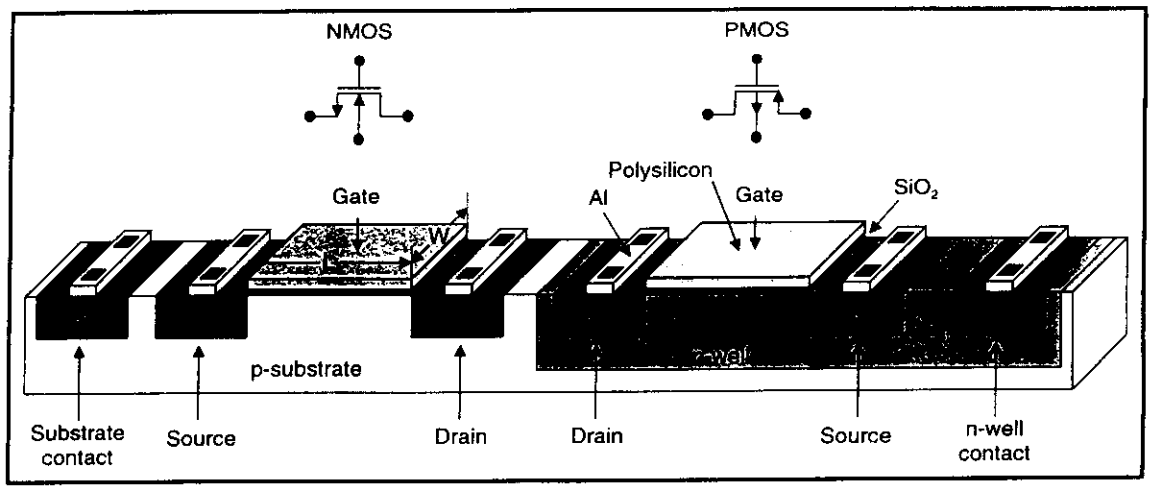
Outline

- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS devices

- CMOS devices
- pn-Junction diodes
- MOSFET equations
- What causes delay?
- MOSFET capacitances
- CMOS device hazards

CMOS devices



CMOS devices

In a CMOS process the devices are:

- PMOS FET's
- NMOS FET's

+ unwanted (but ubiquitous):

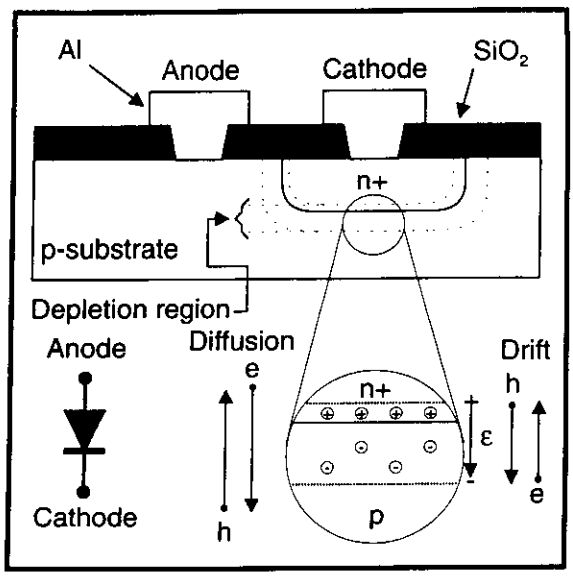
- pn-Junction diodes
- parasitic capacitance

and

- parasitic bipolars
- *parasitic inductance*

pn-Junctions diodes

- Any pn-junction in the IC forms a diode
- Majority carriers diffuse from regions of high to regions of low concentration
- The electric field of the depletion region counteracts diffusion
- In equilibrium there is no net flow of carriers in the diode



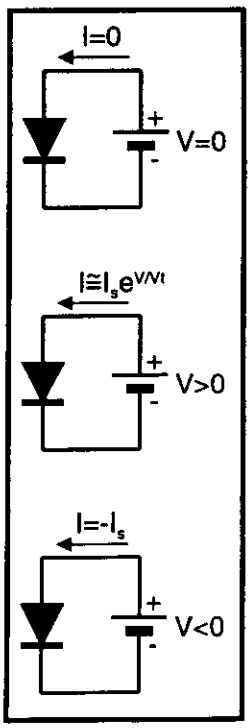
pn-Junction diodes

- Under zero bias there is a built-in potential across the junction
- The built-in potential is:

$$\phi_0 = \phi_T \cdot \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right)$$

$$\phi_T = \frac{k \cdot T}{q} \cong 26\ \text{mV} \ @\ 300^\circ\text{K}$$

$$n_i = 1.5 \times 10^{10}\ \text{cm}^{-3}\ \text{for silicon} \ @\ 300^\circ\text{K}$$



pn-Junction diodes

- Ideal diode equation

$$I_D = I_s \cdot \left(e^{V/\phi_T} - 1\right)$$

- For V>ϕ_T (forward bias)

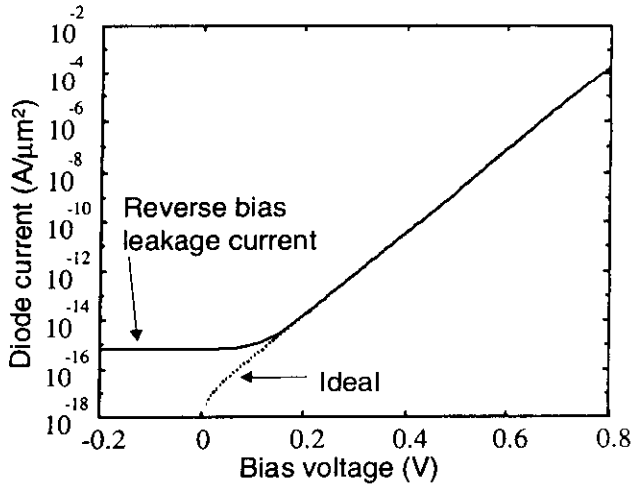
$$I_F \cong I_s \cdot e^{V/\phi_T}$$

- For V<0 (reversed bias)

$$I_R \cong -I_s$$

- In practical diodes due to thermal generation

$$I_R \cong 100 \text{ to } 1000 \times (-I_s)$$



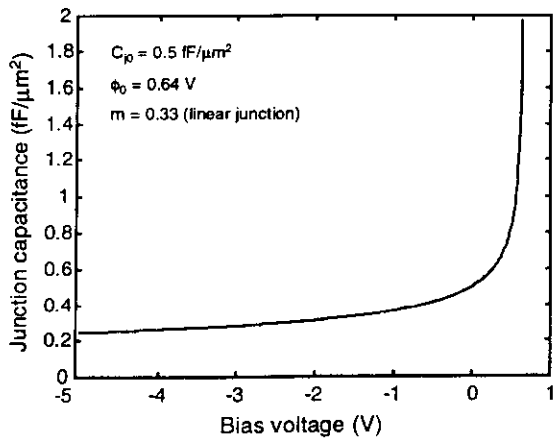
Depletion capacitance

- The depletion, the n- and the p-type regions form a capacitor
- This capacitor is bias dependent:

$$C_j = \frac{C_{j0}}{\left(1 - \frac{V}{\phi_0}\right)^m}$$

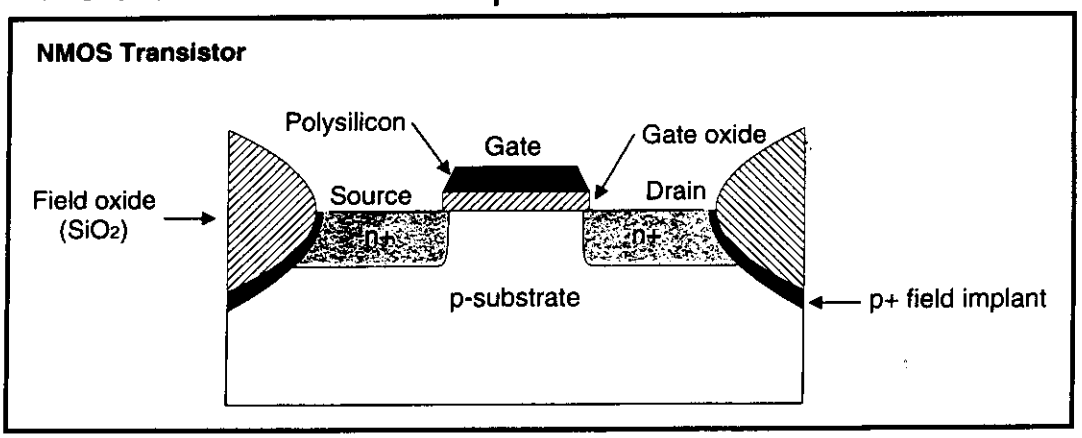
- Simplification: for V<0

$$C_j = k \cdot C_{j0}$$



The NMOS

- Substrate: lightly doped (p-)
- Source and drain: heavily doped (n+)
- Gate: polysilicon
- Thin oxide separates the gate and the "channel"
- Field oxide and field implant isolate the devices



MOSFET equations

- Cut-off region

$$I_{ds} = 0 \quad \text{for} \quad V_{gs} - V_T < 0$$

- Linear region

$$I_{ds} = \mu \cdot C_{ox} \cdot \frac{W}{L} \cdot \left[\left(V_{gs} - V_T \right) \cdot V_{ds} - \frac{V_{ds}^2}{2} \right] \cdot \left(1 + \lambda \cdot V_{ds} \right) \text{ for } 0 < V_{ds} < V_{gs} - V_T$$

- Saturation

$$I_{ds} = \frac{\mu \cdot C_{ox}}{2} \cdot \frac{W}{L} \cdot \left(V_{gs} - V_T \right)^2 \cdot \left(1 + \lambda \cdot V_{ds} \right) \text{ for } V_{ds} > V_{gs} - V_T$$

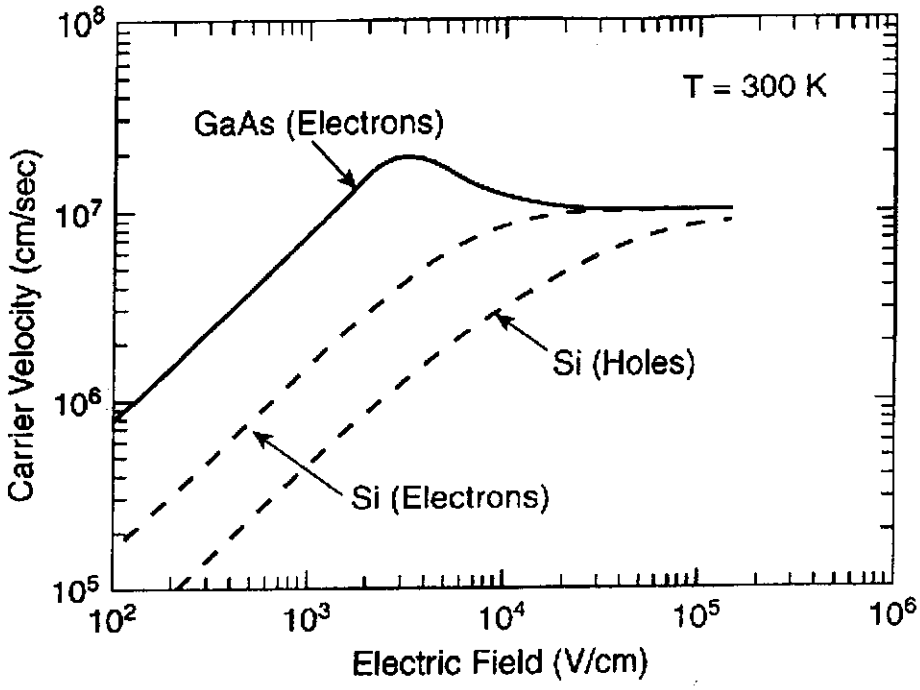
- Oxide capacitance

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \quad \left(\mathrm{F/m^2} \right)$$

- Process "transconductance"

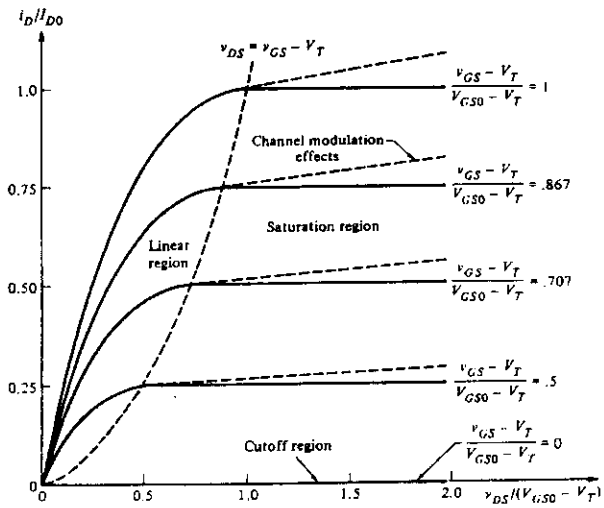
$$\mu \cdot C_{ox} = \frac{\mu \cdot \varepsilon_{ox}}{t_{ox}} \quad \left(\mathrm{A/V^2} \right)$$

Mobility



MOS output characteristics

- Linear region: $V_{ds}<V_{gs}-V_T$
  - Voltage controlled resistor
- Saturation region: $V_{ds}>V_{gs}-V_T$
  - Voltage controlled current source
- Curves deviate from the ideal current source behavior due to:
  - Channel modulation effects



Bulk effect

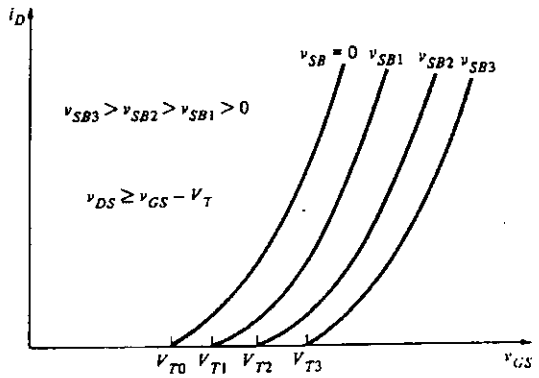
- The threshold depends on the:
  - Doping levels
  - Source-to-bulk voltage
  - Gate oxide thickness

$$V_T = V_{T0} + \gamma \cdot \left(\sqrt{|2\phi_F + V_{sb}|} - \sqrt{|2\phi_F|} \right)$$

$$V_{T0} = \phi_{ms} - 2\phi_F - \frac{1}{C_{ox}} \left[Q_{b0} + Q_{ox} + Q_I \right]$$

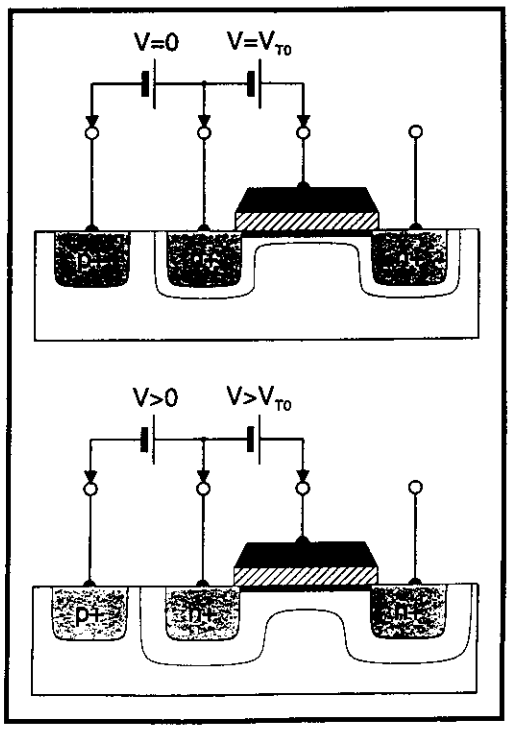
$$\gamma = \frac{\sqrt{2 q \varepsilon_{si} N_A}}{C_{ox}}$$

$$\phi_F = \phi_T \ln \left[\frac{N_A}{n_i} \right] \quad \text{for p - substrate}$$



Bulk effect

- When the semiconductor surface inverts to n-type the channel is in "strong inversion"
- $V_{sb} = 0 \Rightarrow$ strong inversion for:
  - surface potential > $-2\phi_F$
- $V_{sb} > 0 \Rightarrow$ strong inversion for:
  - surface potential > $-2\phi_F + V_{sb}$

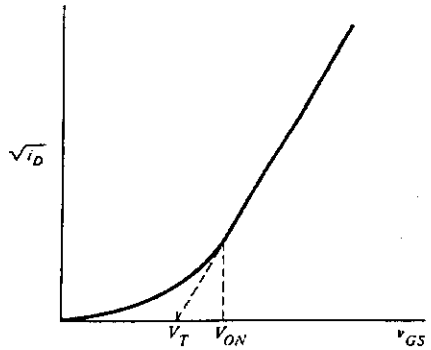


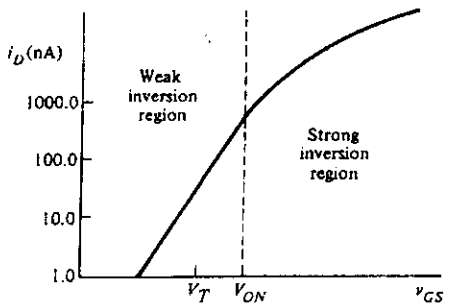
Weak inversion

- Is I_d=0 when $V_{gs}<V_T$?
- For $V_{gs}<V_T$ the drain current depends exponentially on V_{gs}
- In week inversion and saturation:

$$I_d \cong \frac{W}{L} \cdot I_{do} \cdot e^{\frac{q \cdot V_{gs}}{n \cdot k \cdot T}}$$

- Used in very low power designs
- Slow operation

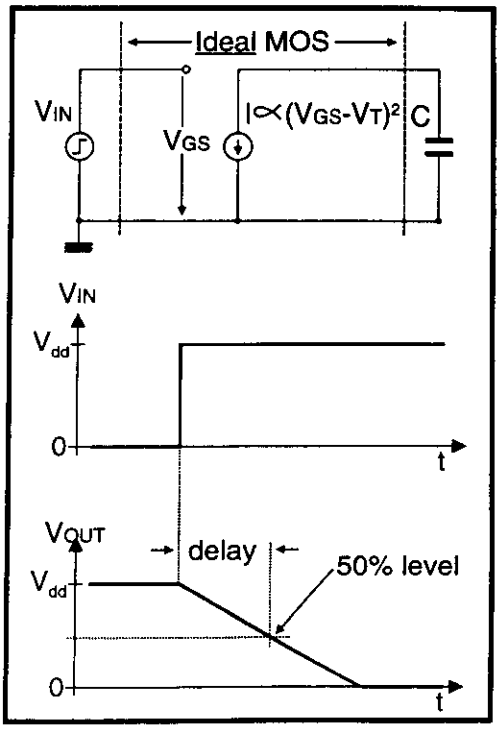




What causes delay?

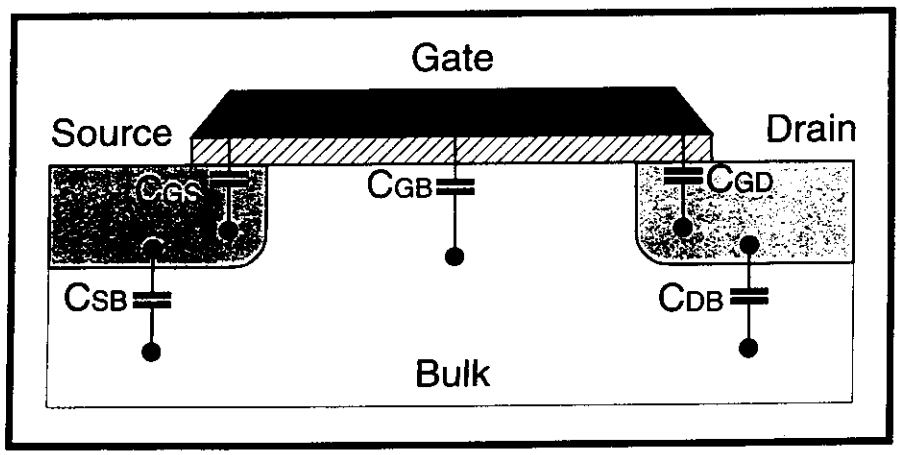
- In MOS circuits capacitive loading is the main cause
- Due to:
  - Device capacitance
  - Interconnect capacitance

$$\Delta t = C \cdot \frac{\Delta V}{I} \approx \frac{C}{2 \cdot \mu \cdot C_{ox} \cdot V_{dd}} \cdot \frac{L}{W}$$



MOSFET capacitances

- MOS capacitances have three origins:
  - The basic MOS structure
  - The channel charge
  - The pn-junctions depletion regions



MOS structure capacitances

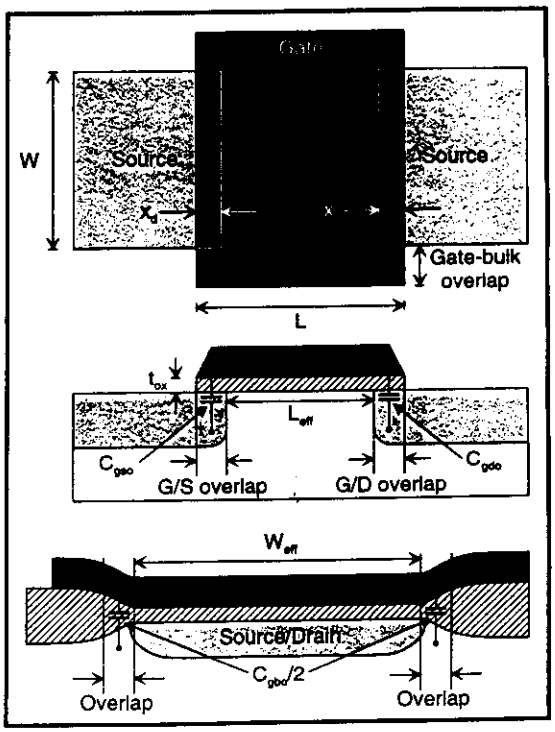
- Source/drain diffusion extend below the gate oxide by:

  x_d - the lateral diffusion
- This gives origin to the source/drain overlap capacitances:

$$C_{gso} = C_{gdo} = C_o \times W$$

$$C_o \ (\mathrm{F/m})$$
- Gate-bulk overlap capacitance:

$$C_{gbo} = C_o' \times L, \quad C_o' \ (\mathrm{F/m})$$



Channel capacitance

- The channel capacitance is nonlinear
- Its value depends on the operation region
- Its formed of three components:

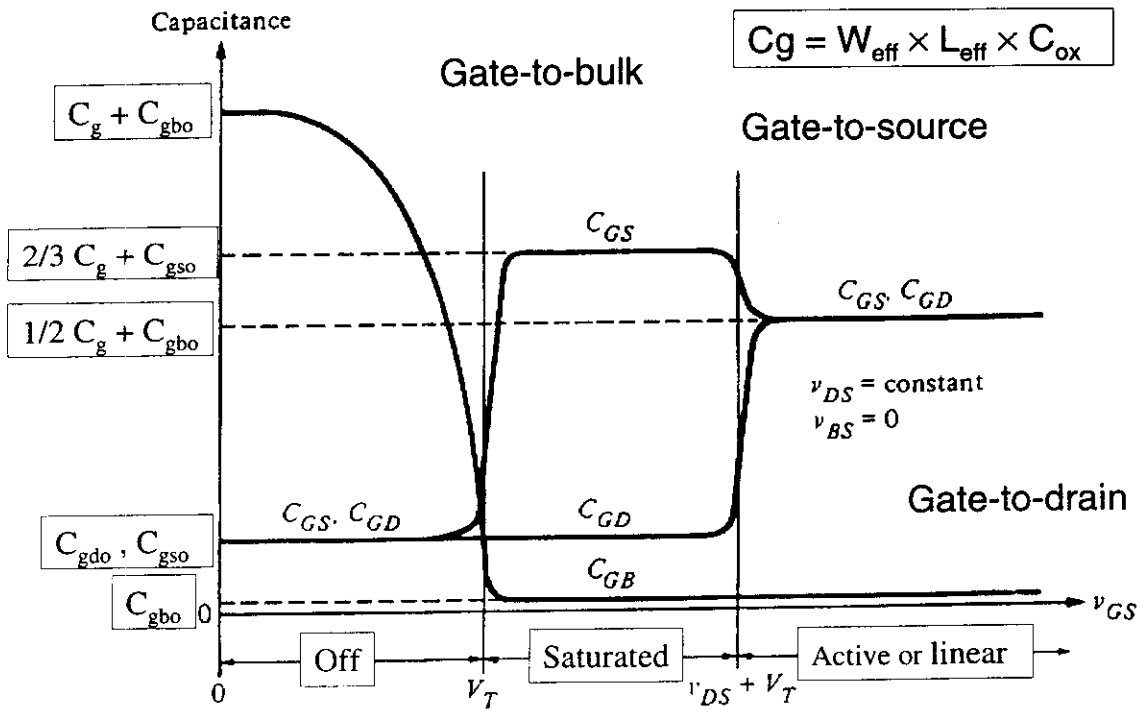
  C_{gb} - gate-to-bulk capacitance

  C_{gs} - gate-to-source capacitance

  C_{gd} - gate-to-drain capacitance

Operation region	C_{gb}	C_{gs}	C_{gd}
Cutoff	C_{ox} W L	0	0
Linear	0	(1/2) C_{ox} W L	(1/2) C_{ox} W L
Saturation	0	(2/3) C_{ox} W L	0

Channel capacitance



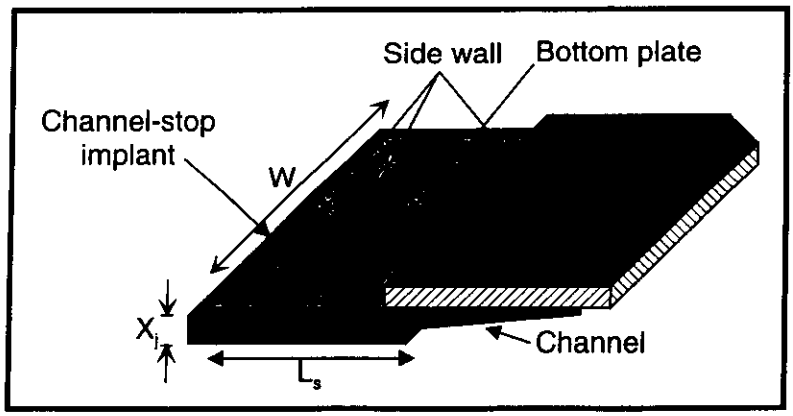
Junction capacitances

- C_{sb} and C_{db} and diffusion capacitances composed of:
  - Bottom-plate capacitance:

$$C_{bottom} = C_j \cdot W \cdot L_s$$

  - Side-wall capacitance:

$$C_{sw} = C_{jsw} \cdot (2\,L_s + W)$$

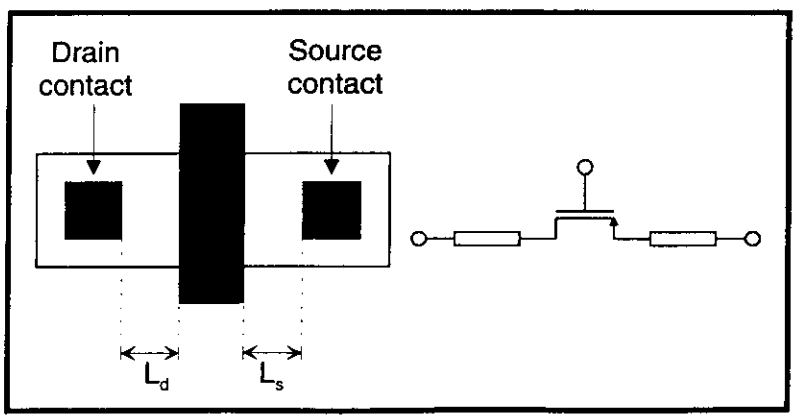


Source/drain resistance

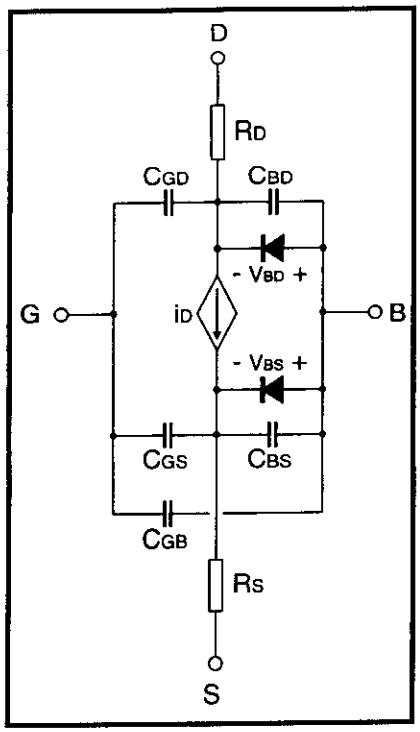
- Scaled down devices ⇒ higher source/drain resistance:

$$R_{s,d} = \frac{L_{s,d}}{W} \cdot R_{sq} + R_c$$

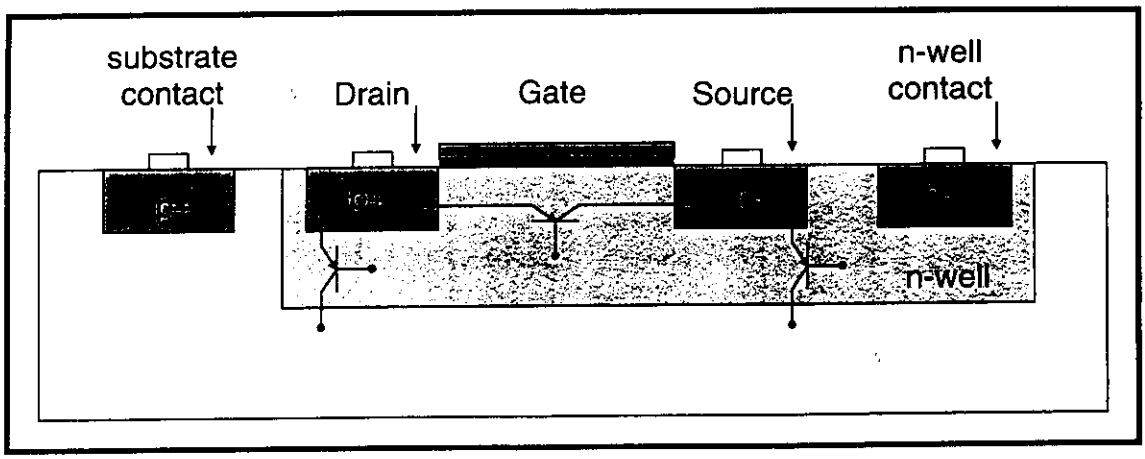
- In sub-μ processes *silicidation* is used to reduce the source, drain and gate parasitic resistance



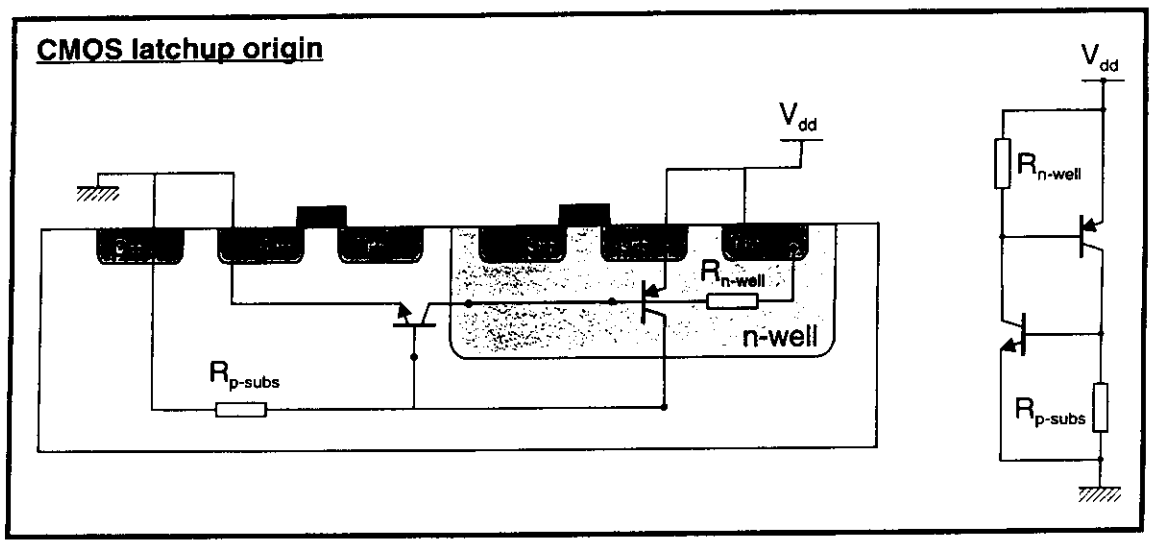
MOSFET model



CMOS parasitic bipolar



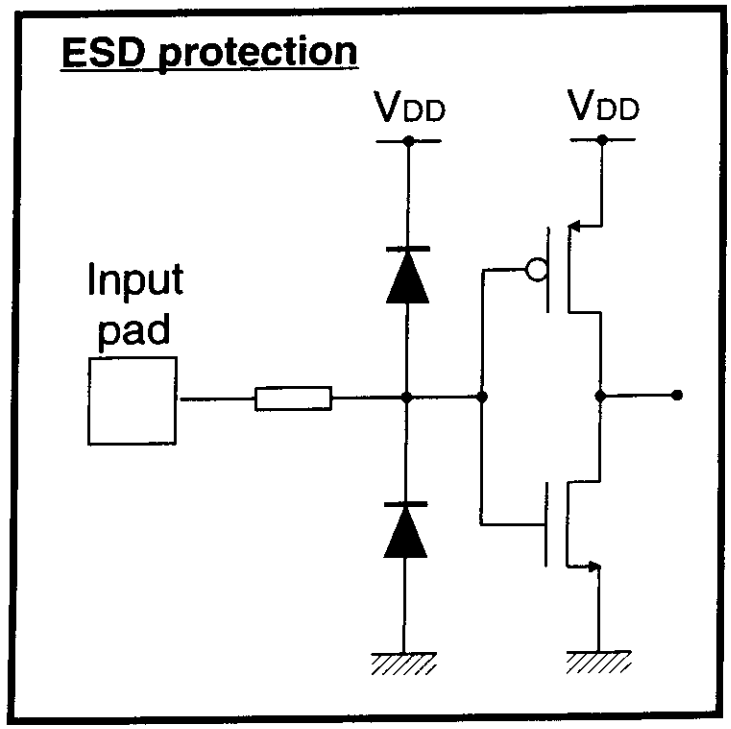
CMOS device hazards



CMOS device hazards

- Sources of latchup:
  - Electrical disturbance
  - Transient on power and ground buses
  - Improper power sequencing
  - Radiation
  - ESD
- How to avoid it:
  - Technological methods (beta reduction, substrate resistance reduction, trench isolation)
  - Layout rules:
    - Spacing rules
    - Contact distribution
    - Guard rings

CMOS device hazards



Outline

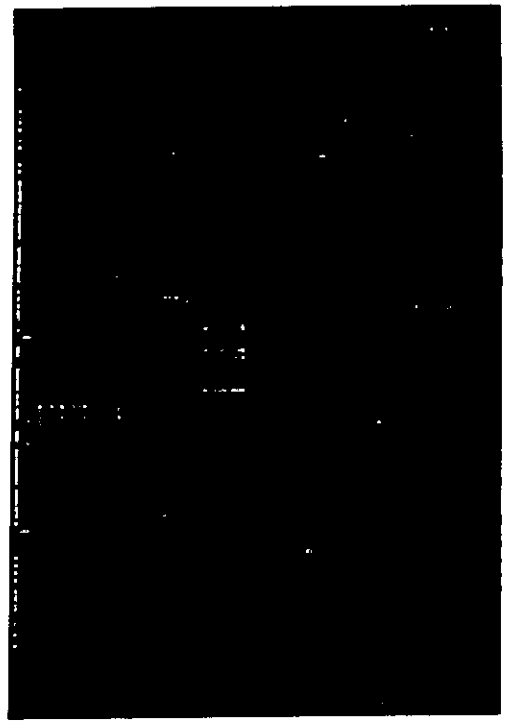
- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS technology

- Lithography
- Physical structure
- CMOS fabrication sequence
- Yield
- Design rules
- Other processes
- Advanced CMOS process
- Process enhancements
- Technology scaling

CMOS technology

- An *Integrated Circuit* is an electronic network fabricated in a single piece of a semiconductor material
- The semiconductor surface is subjected to various processing steps in which impurities and other materials are added with specific geometrical patterns
- The fabrication steps are sequenced to form three dimensional regions that act as transistors and interconnects that form the switching or amplification network



Lithography

Lithography: process used to transfer patterns to each layer of the IC

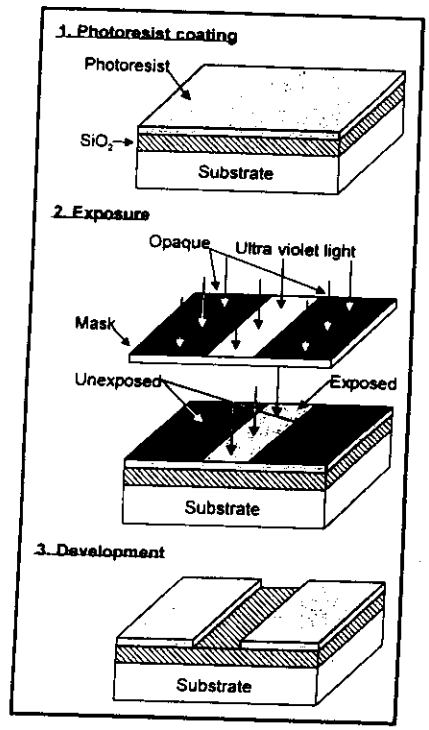
Lithography sequence steps:

- Designer:
  - Drawing the layer patterns on a layout editor
- Silicon Foundry:
  - Masks generation from the layer patterns in the design data base
  - Printing: transfer the mask pattern to the wafer surface
  - Process the wafer to physically pattern each layer of the IC

Lithography

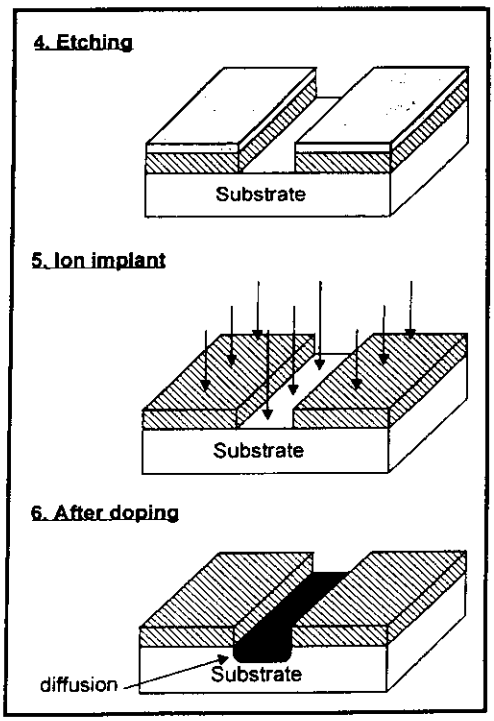
Basic sequence

- The surface to be patterned is:
  - spin-coated with photoresist
  - the photoresist is dehydrated in an oven (photo resist: light-sensitive organic polymer)
- The photoresist is exposed to ultra violet light:
  - For a positive photoresist exposed areas become soluble and non exposed areas remain hard
- The soluble photoresist is chemically removed (development).
  - The patterned photoresist will now serve as an etching mask for the SiO_2



Lithography

- The SiO_2 is etched away leaving the substrate exposed:
  - the patterned resist is used as the etching mask
- Ion Implantation:
  - the substrate is subjected to highly energized donor or acceptor atoms
  - The atoms impinge on the surface and travel below it
  - The patterned silicon SiO_2 serves as an implantation mask
- The doping is further driven into the bulk by a thermal cycle

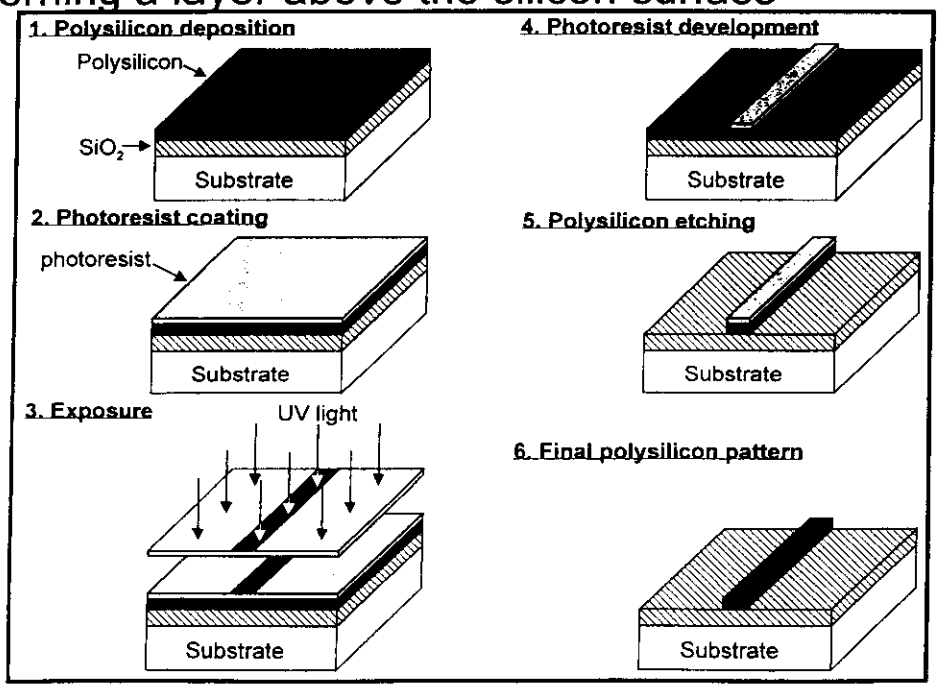


Lithography

- The lithographic sequence is repeated for each physical layer used to construct the IC. The sequence is always the same:
  - Photoresist application
  - Printing (exposure)
  - Development
  - Etching

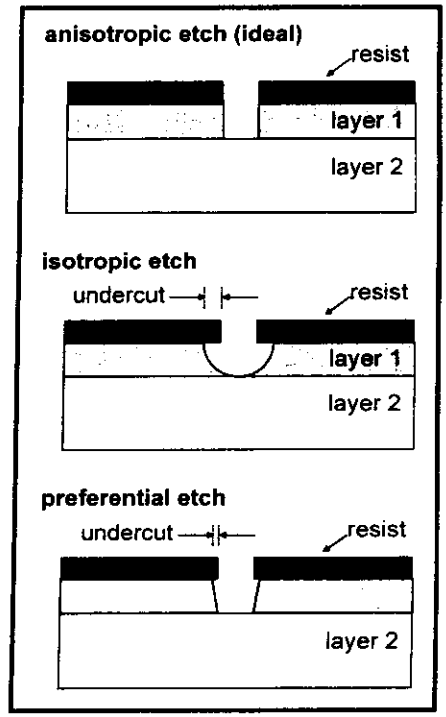
Lithography

Patterning a layer above the silicon surface

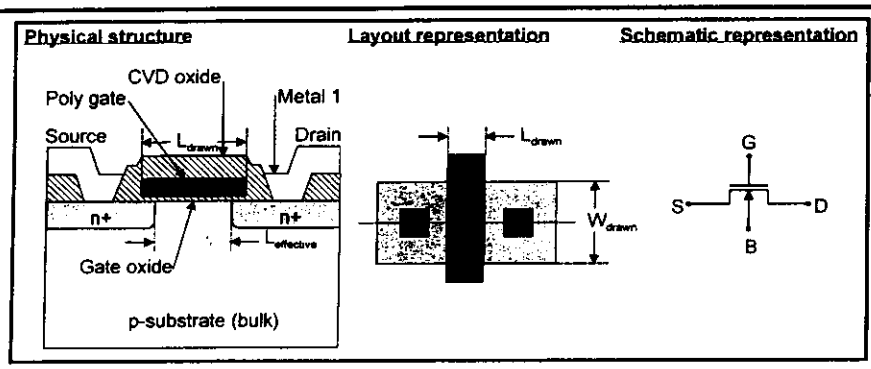


Lithography

- Etching:
  - Process of removing unprotected material
  - Etching occurs in all directions
  - Horizontal etching causes an under cut
  - "preferential" etching can be used to minimize the undercut
- Etching techniques:
  - Wet etching: uses chemicals to remove the unprotected materials
  - Dry or plasma etching: uses ionized gases rendered chemically active by an rf-generated plasma



Physical structure



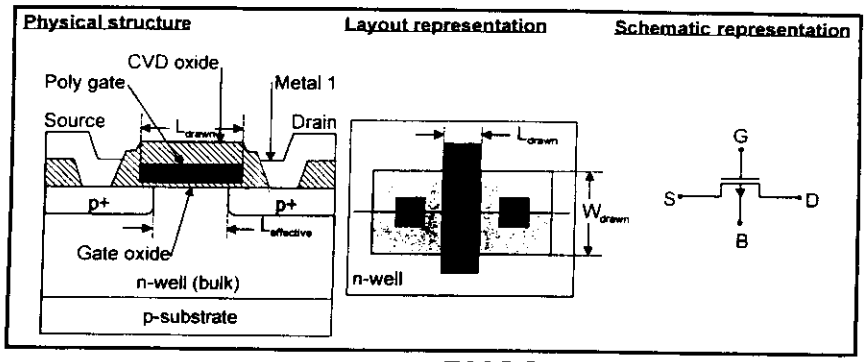
NMOS physical structure:

- p-substrate
- n+ source/drain
- gate oxide (SiO_2)
- polysilicon gate
- CVD oxide
- metal 1
- $L_{eff}<L_{drawn}$ (lateral doping effects)

NMOS layout representation:

- Implicit layers:
  - oxide layers
  - substrate (bulk)
- Drawn layers:
  - n+ regions
  - polysilicon gate
  - oxide contact cuts
  - metal layers

Physical structure



PMOS physical structure:

- p-substrate
- n-well (bulk)
- p+ source/drain
- gate oxide (SiO_2)
- polysilicon gate
- CVD oxide
- metal 1

PMOS layout representation:

- Implicit layers:
  - oxide layers
- Drawn layers:
  - n-well (bulk)
  - n+ regions
  - polysilicon gate
  - oxide contact cuts
  - metal layers

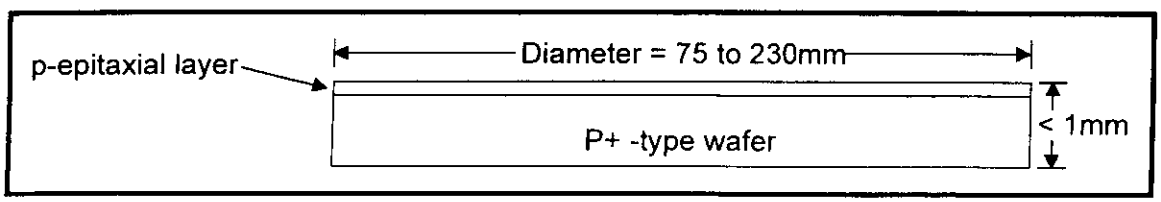
CMOS fabrication sequence

0. Start:

- For an n-well process the starting point is a p-type silicon wafer:
- wafer: typically 75 to 230mm in diameter and less than 1mm thick

1. Epitaxial growth:

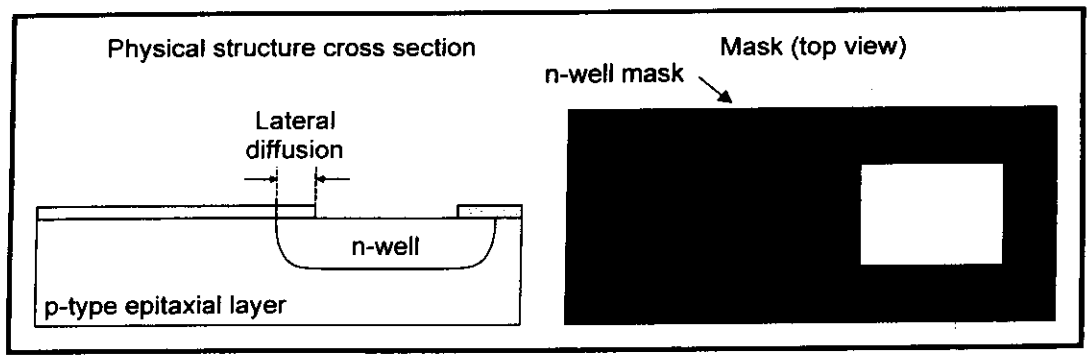
- A single p-type single crystal film is grown on the surface of the wafer by:
  - subjecting the wafer to high temperature and a source of dopant material
- The epi layer is used as the base layer to build the devices



CMOS fabrication sequence

2. N-well Formation:

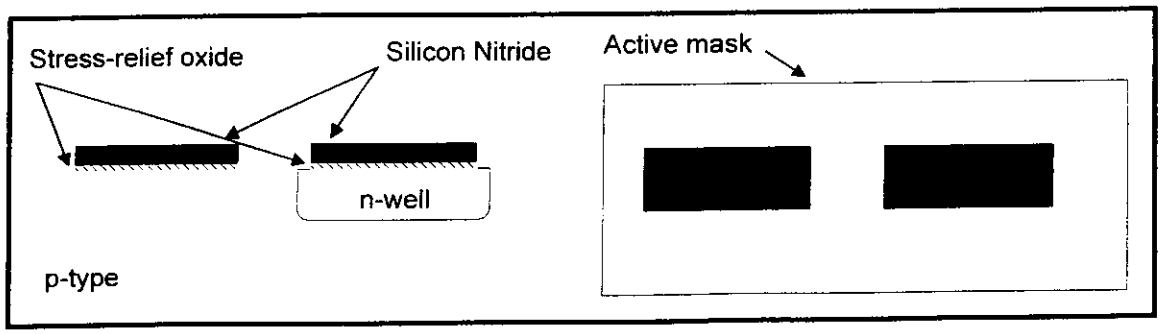
- PMOS transistors are fabricated in n-well regions
- The first mask defines the n-well regions
- N-well's are formed by ion implantation or deposition and diffusion
- Lateral diffusion limits the proximity between structures
- Ion implantation results in shallower wells compatible with today's fine-line processes



CMOS fabrication sequence

3. Active area definition:

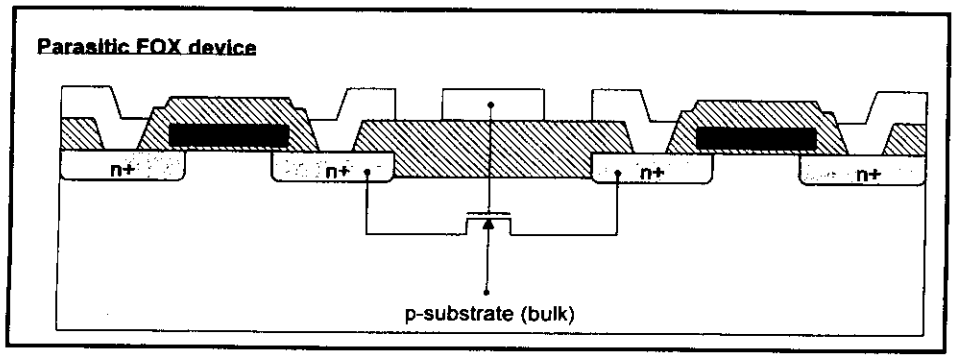
- Active area:
  - planar section of the surface where transistors are build
  - defines the gate region (thin oxide)
  - defines the n+ or p+ regions
- A thin layer of SiO_2 is grown over the active region and covered with silicon nitride



CMOS fabrication sequence

4. Isolation:

- Parasitic (unwanted) FET's exist between unrelated transistors (Field Oxide FET's)
- Source and drains are existing source and drains of wanted devices
- Gates are metal and polysilicon interconnects
- The threshold voltage of FOX FET's are higher than for normal FET's

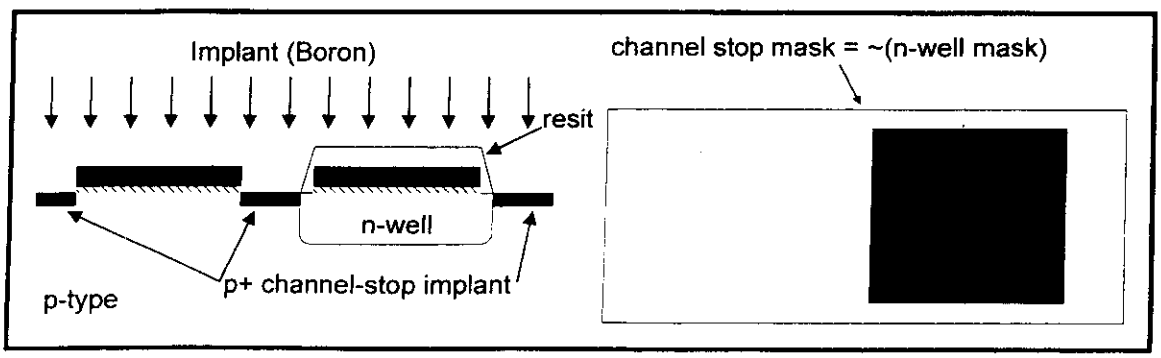


CMOS fabrication sequence

- FOX FET's threshold is made high by:
  - introducing a channel-stop diffusion that raises the impurity concentration in the substrate in areas where transistors are not required
  - making the FOX thick

4.1 Channel-stop implant

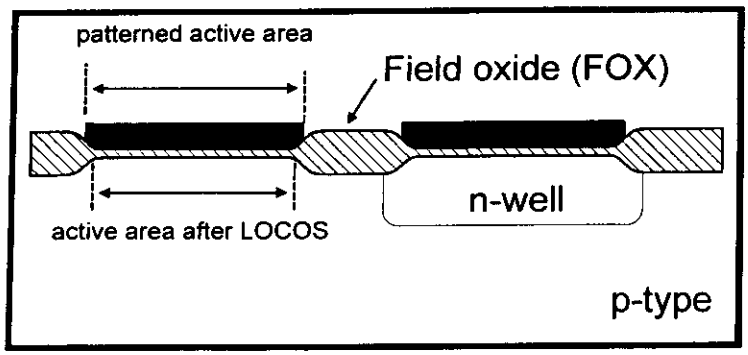
- The silicon nitride (over n-active) and the photoresist (over n-well) act as masks for the channel-stop implant



CMOS fabrication sequence

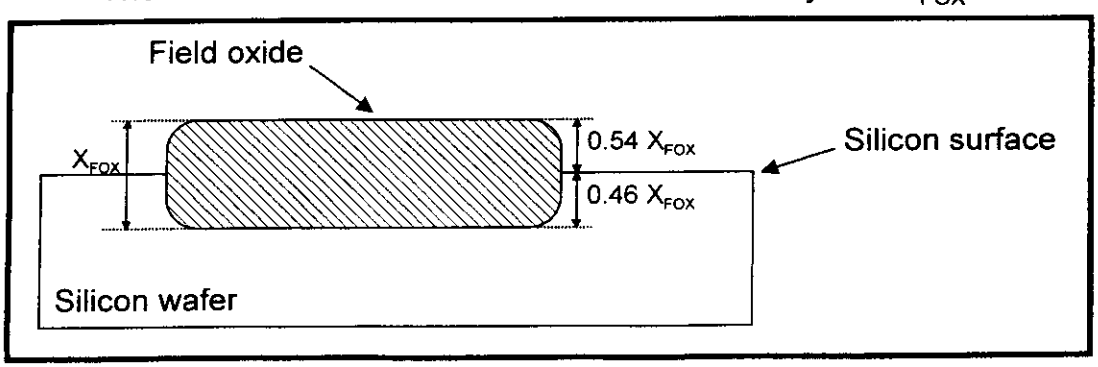
4.2 Local oxidation of silicon (LOCOS)

- The photoresist mask is removed
- The SiO_2/SiN layers will now act as a masks
- The thick field oxide is then grown by:
  - exposing the surface of the wafer to a flow of oxygen-rich gas
- The oxide grows in both the vertical and lateral directions
- This results in a active area smaller than patterned



CMOS fabrication sequence

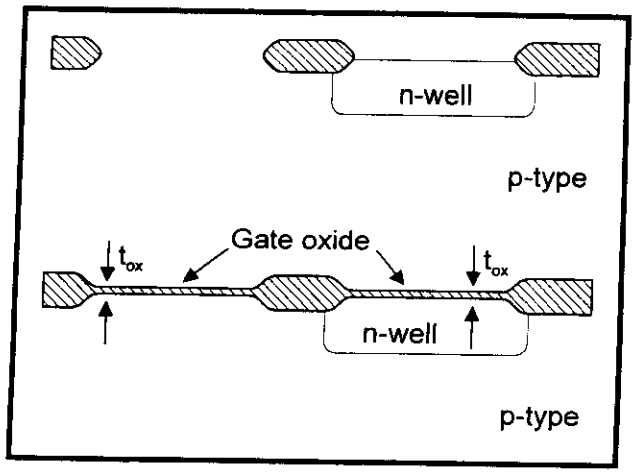
- Silicon oxidation is obtained by:
  - Heating the wafer in a oxidizing atmosphere:
    - Wet oxidation: water vapor, T = 900 to 1000°C (rapid process)
    - Dry oxidation: Pure oxygen, T = 1200°C (high temperature required to achieve an acceptable growth rate)
- Oxidation consumes silicon
  - SiO_2 has approximately twice the volume of silicon
  - The FOX is recedes below the silicon surface by $0.46X_{FOX}$



CMOS fabrication sequence

5. Gate oxide growth

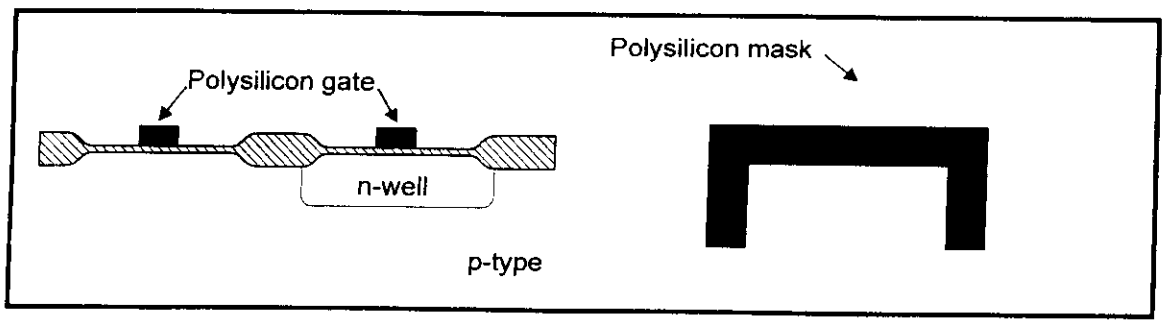
- The nitride and stress-relief oxide are removed
- The devices threshold voltage is adjusted by
  - adding charge at the silicon/oxide interface
- The well controlled gate oxide is grown with thickness t_{ox}



CMOS fabrication sequence

6. Polysilicon deposition and patterning

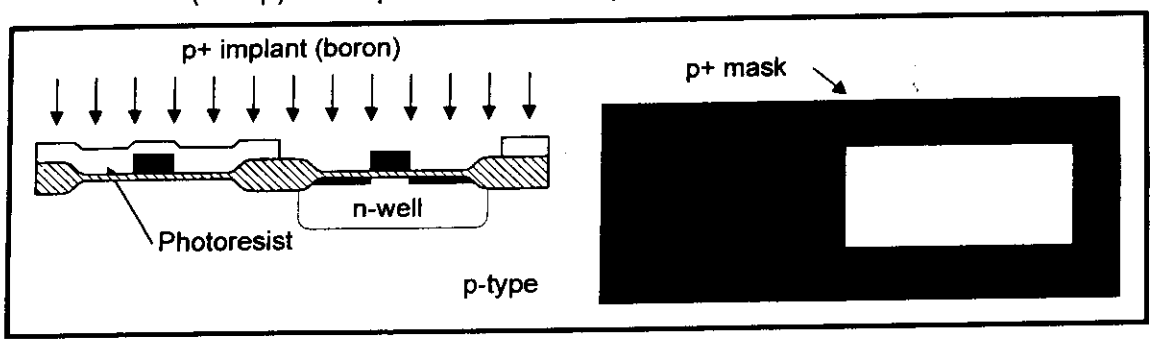
- A layer of polysilicon is deposited over the entire wafer surface
- The polysilicon is then patterned by a lithography sequence
- All the MOSFET gates are defined in a single step
- The polysilicon gate can be doped (n+) while is being deposited to lower its parasitic resistance (important in high speed fine line processes)



CMOS fabrication sequence

7. PMOS formation

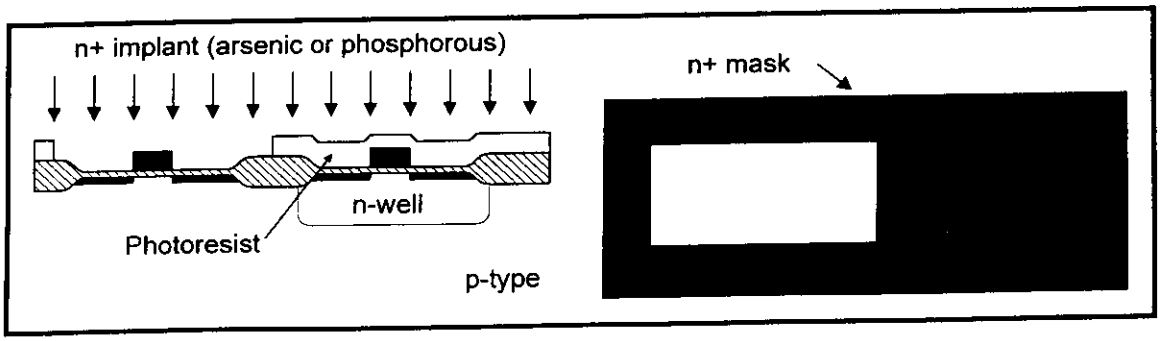
- Photoresist is patterned to cover all but the p+ regions
- A boron ion beam creates the p+ source and drain regions
- The polysilicon serves as a mask to the underlying channel
  - This is called a self-aligned process
  - It allows precise placement of the source and drain regions
- During this process the gate gets doped with p-type impurities
  - Since the gate had been doped n-type during deposition, the final type (n or p) will depend on which dopant is dominant



CMOS fabrication sequence

8. NMOS formation

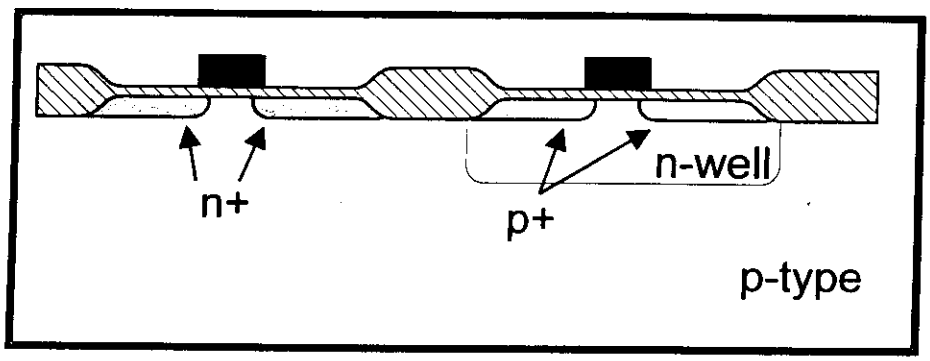
- Photoresist is patterned to define the n+ regions
- Donors (arsenic or phosphorous) are ion-implanted to dope the n+ source and drain regions
- The process is self-aligned
- The gate is n-type doped



CMOS fabrication sequence

9. Annealing

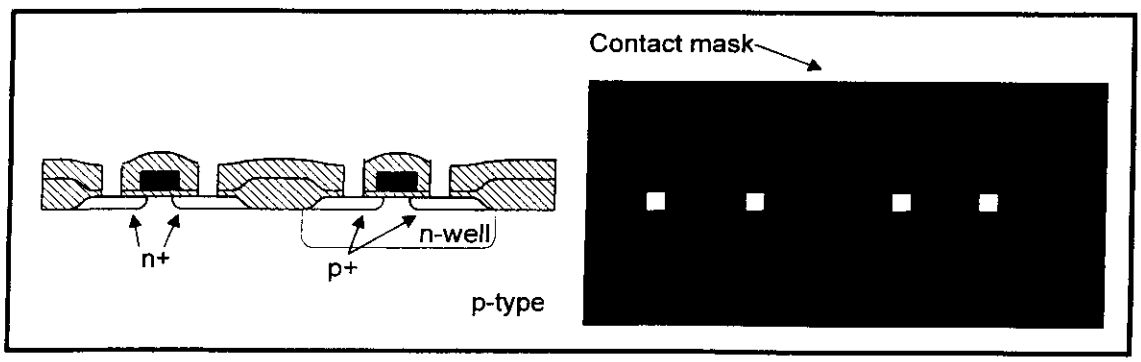
- After the implants are completed a thermal annealing cycle is executed
- This allows the impurities to diffuse further into the bulk
- After thermal annealing, it is important to keep the remaining process steps at as low temperature as possible



CMOS fabrication sequence

10. Contact cuts

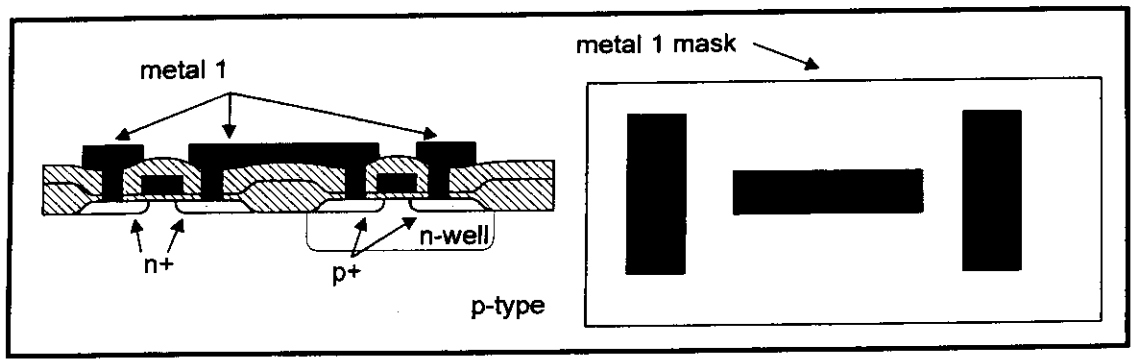
- The surface of the IC is covered by a layer of CVD oxide
  - The oxide is deposited at low temperature (LTO) to avoid that underlying doped regions will undergo diffusive spreading
- Contact cuts are defined by etching SiO_2 down to the surface to be contacted
- These allow metal to contact diffusion and/or polysilicon regions



CMOS fabrication sequence

11. Metal 1

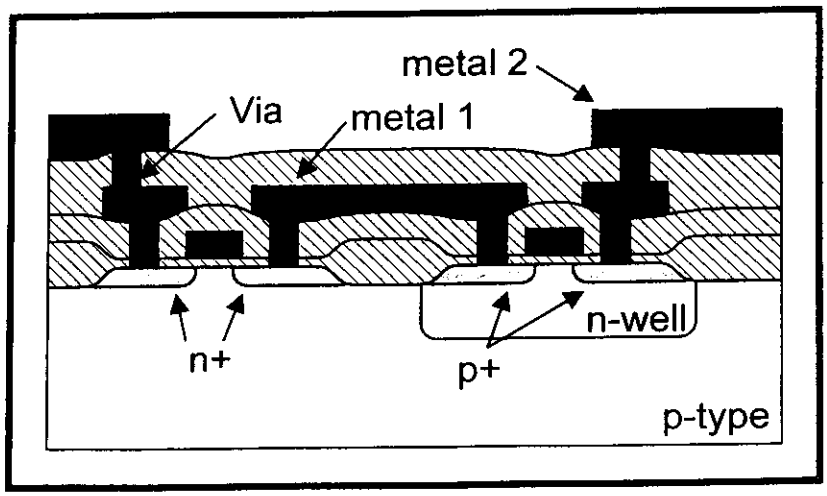
- A first level of metallization is applied to the wafer surface and selectively etched to produce the interconnects



CMOS fabrication sequence

12. Metal 2

- Another layer of LTO CVD oxide is added
- Via openings are created
- Metal 2 is deposited and patterned



CMOS fabrication sequence

13. Over glass and pad openings

- A protective layer is added over the surface:
- The protective layer consists of:
  - A layer of SiO_2
  - Followed by a layer of silicon nitride
- The SiN layer acts as a diffusion barrier against contaminants (passivation)
- Finally, contact cuts are etched, over metal 2, on the passivation to allow for wire bonding.

Yield

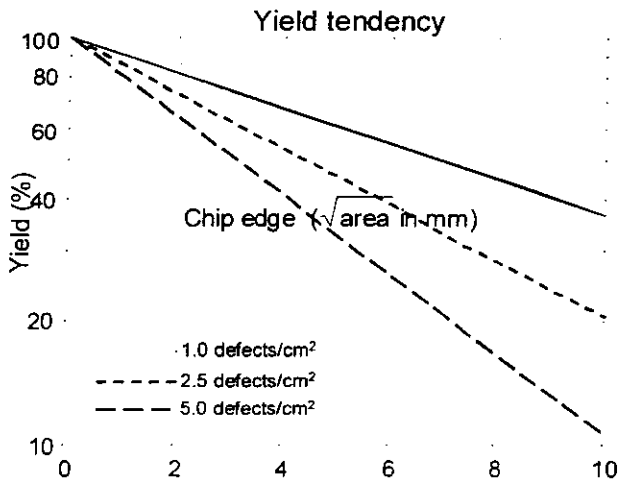
- Yield

$$Y = \frac{number\ \text{of good chips on wafer}}{\text{total number of chips}}$$

- The yield is influenced by:
  - the technology
  - the chip area
  - the layout
- Scribe cut and packaging also contribute to the final yield
- Yield can be approximated by: $Y = e^{-\sqrt{A}D}$

A - chip area (cm^2)

D - defect density (defects/cm^2)

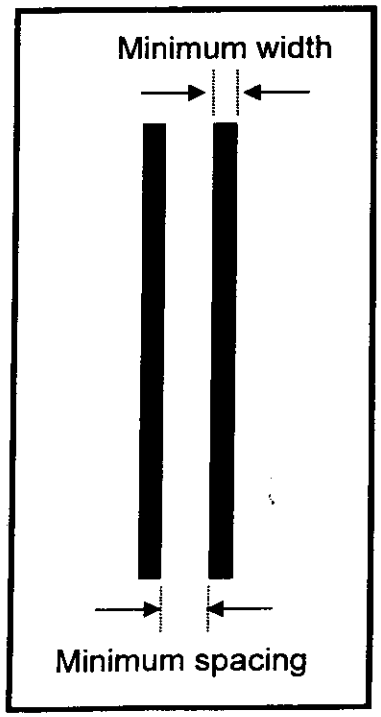


Design rules

- The limitations of the patterning process give rise to a set of mask design guidelines called design rules
- Design rules are a set of guidelines that specify the minimum dimensions and spacings allowed in a layout drawing
- Violating a design rule might result in a non-functional circuit or in a highly reduced yield
- The design rules can be expressed as:
  - A list of minimum feature sizes and spacings for all the masks required in a given process
  - Based on single parameter λ that characterize the linear feature (e.g. the minimum grid dimension). λ base rules allow simple scaling

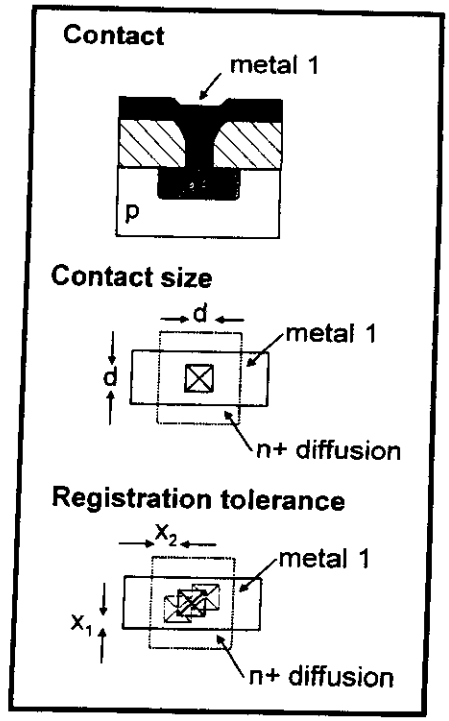
Design rules

- Minimum line-width:
  - smallest dimension permitted for any object in the layout drawing (minimum feature size)
- Minimum spacing:
  - smallest distance permitted between the edges of two objects
- This rules originate from the resolution of the optical printing system, the etching process, or the surface roughness



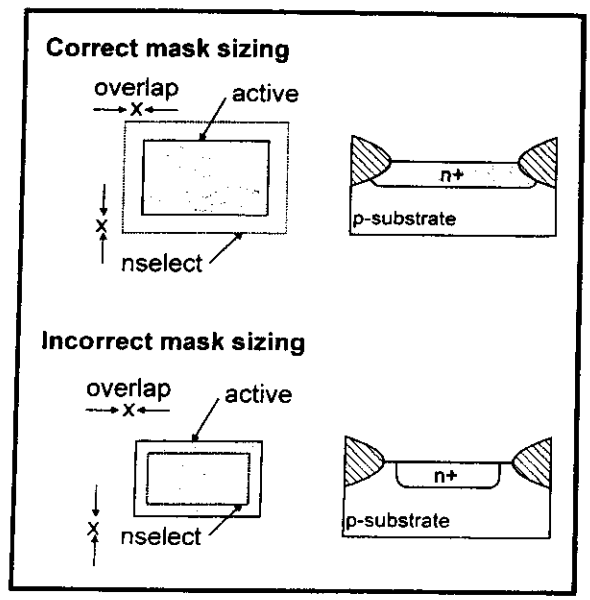
Design rules

- Contacts and vias:
  - minimum size limited by the lithography process
  - large contacts can result in cracks and voids
  - Dimensions of contact cuts are restricted to values that can be reliably manufactured
  - A minimum distance between the edge of the oxide cut and the edge of the patterned region must be specified to allow for misalignment tolerances (registration errors)



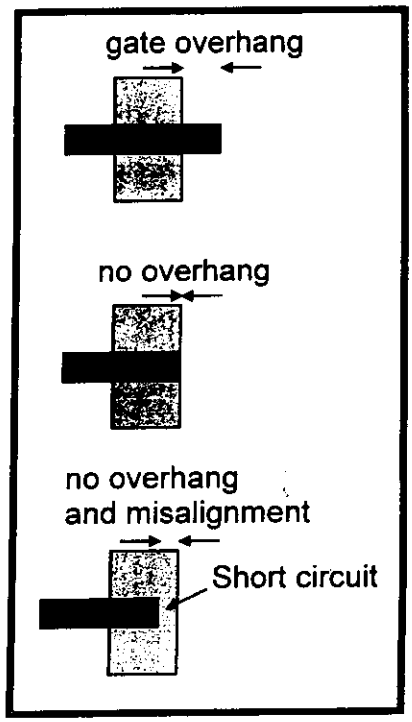
Design rules

- MOSFET rules
  - n+ and p+ regions are formed in two steps:
    - the active area openings allow the implants to penetrate into the silicon substrate
    - the nselect or pselect provide photoresist openings over the active areas to be implanted
  - Since the formation of the diffusions depend on the overlap of two masks, the nselect and pselect regions must be larger than the corresponding active areas to allow for misalignments



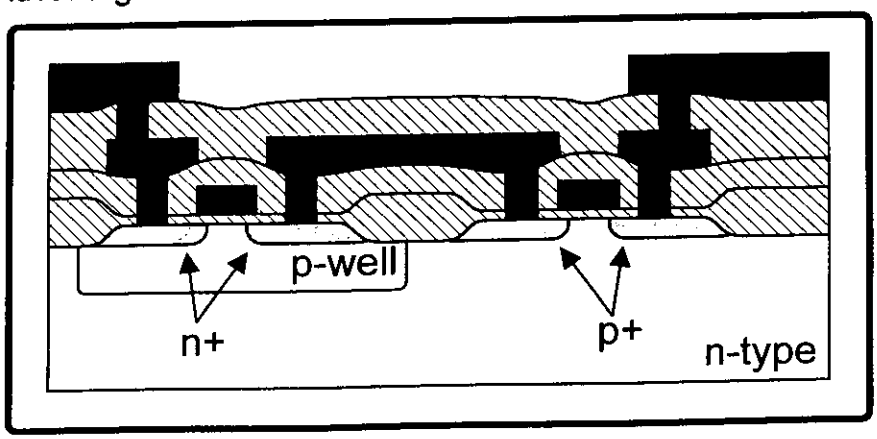
Design rules

- Gate overhang:
  - The gate must overlap the active area by a minimum amount
  - This is done to ensure that a misaligned gate will still yield a structure with separated drain and source regions
- A modern process has may hundreds of rules to be verified
  - Programs called Design Rule Checkers assist the designer in that task



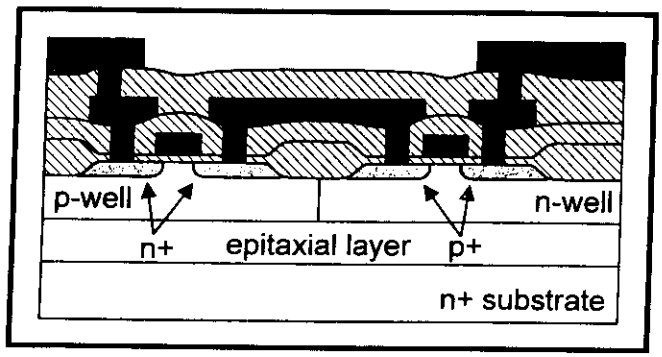
Other processes

- **P-well process**
  - NMOS devices are build on a implanted p-well
  - PMOS devices are build on the substrate
  - P-well process moderates the difference between the p- and the n-transistors since the P devices reside in the native substrate
  - Advantages: better balance between p- and n-transistors



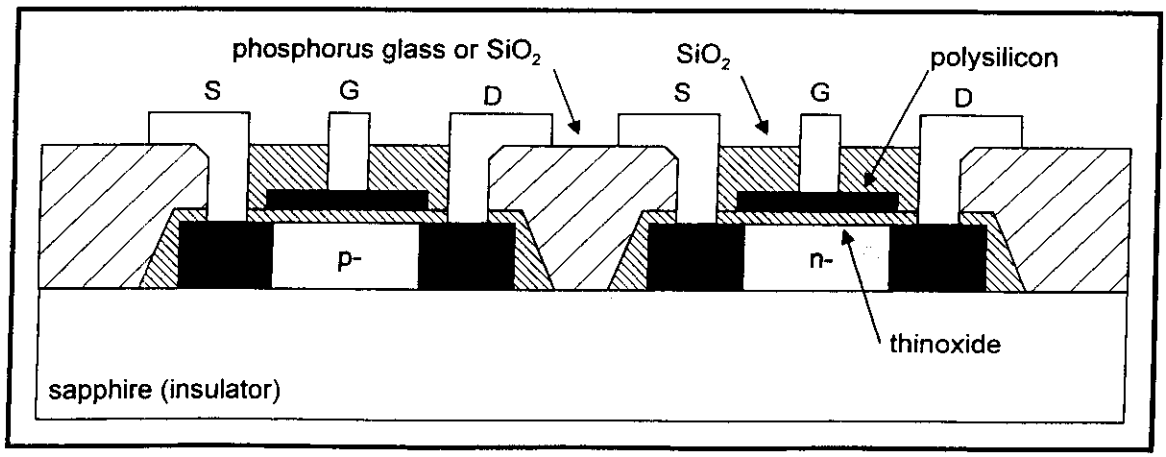
Other processes

- **Twin-well process**
  - n+ or p+ substrate plus a lightly doped epi-layer (latchup prevention)
  - wells for the n- and p-transistors
  - Advantages, simultaneous optimization of p- and n-transistors:
    - threshold voltages
    - body effect
    - gain



Other processes

- **Silicon On Insulator (SOI)**
  - Islands of silicon on an insulator form the transistors
- Advantages:
  - No wells ⇒ denser transistor structures
  - Lower substrate capacitances



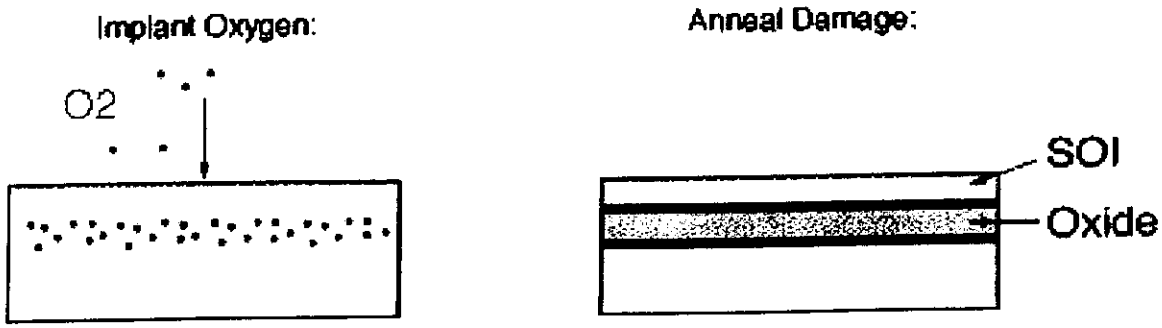
Other processes

- Very low leakage currents
- No FOX FET exists between unrelated devices
- No latchup
- No body-effect:
  - However, the absence of a backside substrate can give origin to the "kink effect"
- Radiation tolerance

- Disadvantages:
  - Absence of substrate diodes (hard to implement protection circuits)
  - Higher number of substrate defects ⇒ lower gain devices
  - More expensive processing

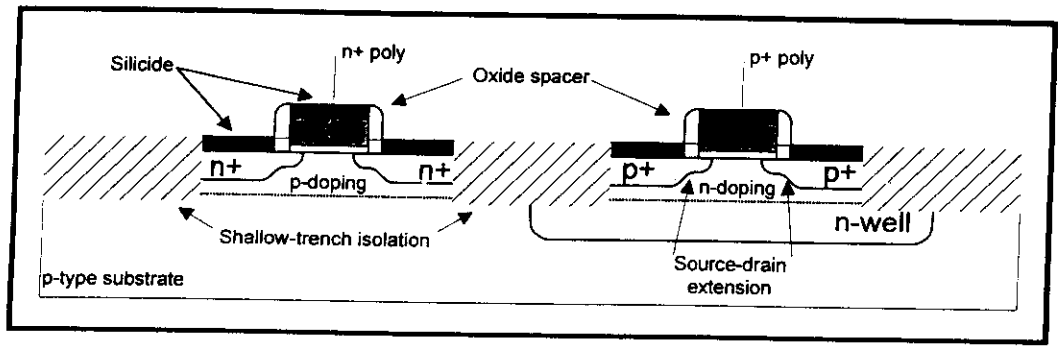
Other processes

- SOI wafers can also be manufactured by a method called: Separation by Implantation of Oxygen (SIMOX)
- The starting material is a silicon wafer where heavy doses of oxygen are implanted
- The wafer is annealed until a thin layer of SOI film is formed
- Once the SOI film is made, the fabrication steps are similar to those of a bulk CMOS process



Advanced CMOS processes

- Shallow trench isolation
- n+ and p+-doped polysilicon gates (low threshold)
- source-drain extensions LDD (hot-electron effects)
- Self-aligned silicide (spacers)
- Non-uniform channel doping (short-channel effects)



Process enhancements

- Up to six metal levels in modern processes
- Copper for metal levels 2 and higher
- Stacked contacts and vias
- Chemical Metal Polishing for technologies with several metal levels
- For analogue applications some processes offer:
  - capacitors
  - resistors
  - bipolar transistors (BiCMOS)

Technology scaling

- Currently, technology scaling has a threefold objective:
  - Reduce the gate delay by 30% (43% increase in frequency)
  - Double the transistor density
  - Saving 50% of power (at 43% increase in frequency)
- 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.

Technology scaling

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

- Device/die area:

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

  - In practice, microprocessor die size grows 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, memory density has been scaling as expected. (not true for microprocessors...)

Technology scaling

- 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 two factors:
    - the number of gate delays in a clock cycle decreases with time (the designs become highly pipelined)
    - advanced circuit techniques reduce the **average gate delay beyond 30%** per generation.

Technology scaling

- Power:

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

- 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.

Technology scaling

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

Interconnects are scaling as expected

Technology scaling

Parameter	Constant Field	Constant Voltage	
Supply voltage (V_{dd})	$1/\alpha$	1	↑
Length (L)	$1/\alpha$	$1/\alpha$	
Width (W)	$1/\alpha$	$1/\alpha$	**Scaling Variables**
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	↑
Depletion layer thickness	$1/\alpha$	$1/\alpha$	
Gate area (Die area)	$1/\alpha^2$	$1/\alpha^2$	**Device Repercussion**
Gate capacitance (load) (C)	$1/\alpha$	$1/\alpha$	
Drain-current (I_{dss})	$1/\alpha$	α	↓
Transconductance (g_m)	1	α	
Gate delay	$1/\alpha$	$1/\alpha^2$	↑
Current density	α	α^3	
DC & Dynamic power dissipation	$1/\alpha^2$	α	**Circuit Repercussion**
Power density	1	α^3	
Power-Delay product	$1/\alpha^3$	$1/\alpha$	↓

Technology scaling

Lithography:

Optics technology	Technology node
248nm mercury-xenon lamp	180 - 250nm
248nm krypton-fluoride laser	130 - 180nm
193nm argon-fluoride laser	100 - 130nm
157nm fluorine laser	70 - 100nm
13.4nm extreme UV	50 - 70nm

Technology scaling

Lithography:

- Electron Beam Lithography (EBL)
  - Patterns are derived directly from digital data
  - The process can be direct: no masks
  - Pattern changes can be implemented quickly
  - However:
    - Equipment cost is high
    - Large amount of time required to access all the points on the wafer

Outline

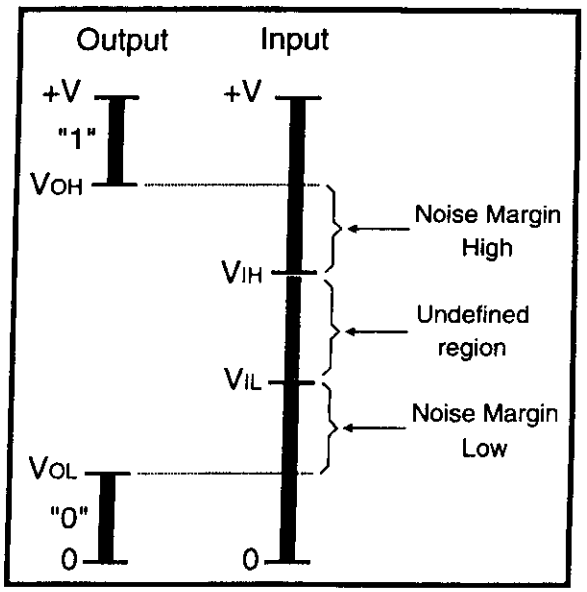
- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS logic structures

- CMOS logic: "0" and "1"
- The MOST - a simple switch
- The CMOS inverter
- The CMOS pass gate
- Simple CMOS gates
- Complex CMOS gates

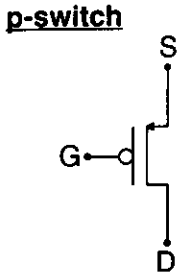
CMOS logic: "0" and "1"

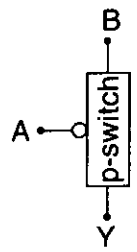
- Logic circuits process Boolean variables
- Logic values are associated with voltage levels:
  - $V_{IN} > V_{IH} \Rightarrow$ "0"
  - $V_{IN} < V_{IL} \Rightarrow$ "0"
- Noise margin:
  - $NM_H = V_{OH} - V_{IH}$
  - $NM_L = V_{IL} - V_{OL}$



The MOST - a simple switch

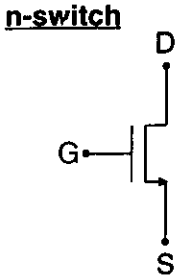
p-switch

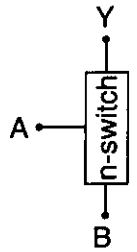




A	B	Y	
0	0	bad 0	(source follower)
0	1	**good 1**	
1	0	?	(high Z)
1	1	?	(high Z)

n-switch



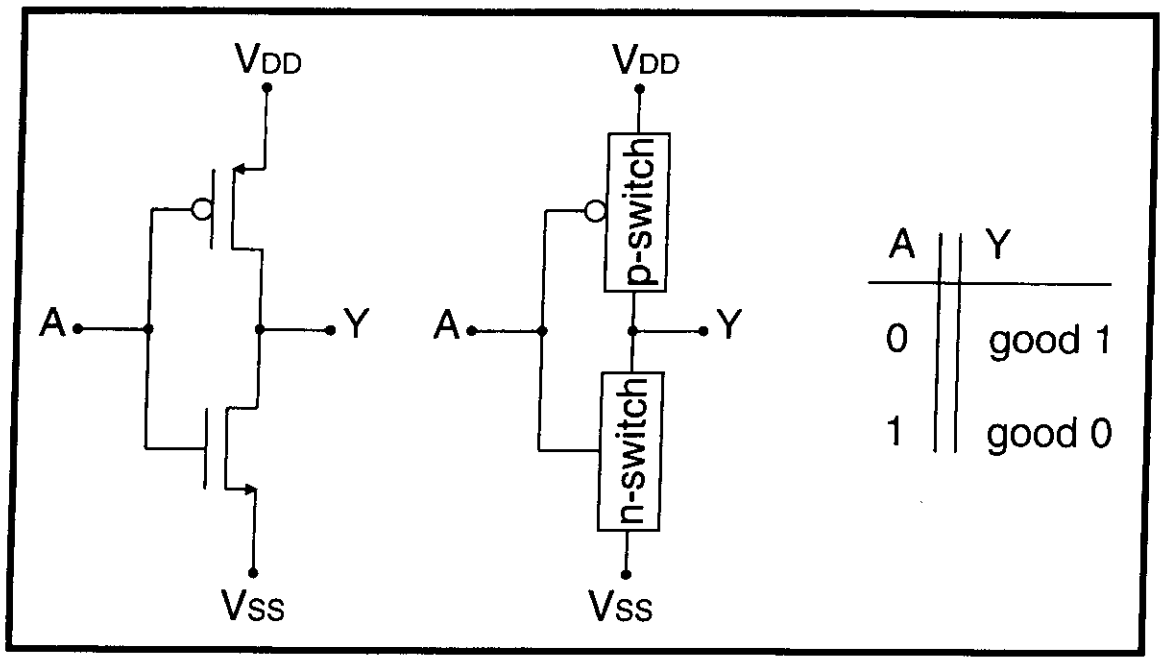


A	B	Y	
0	0	?	(high Z)
0	1	?	(high Z)
1	0	**good 0**	
1	1	bad 1	(source follower)

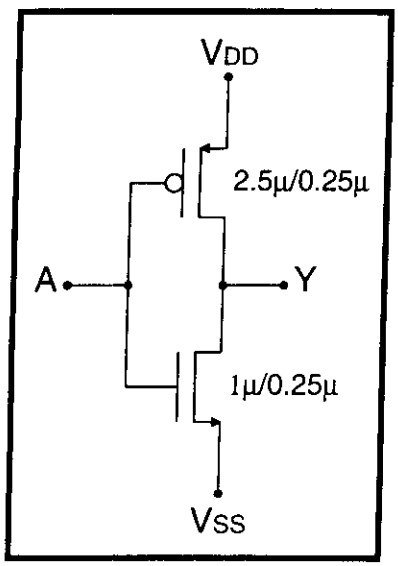
MOSFET's in digital design

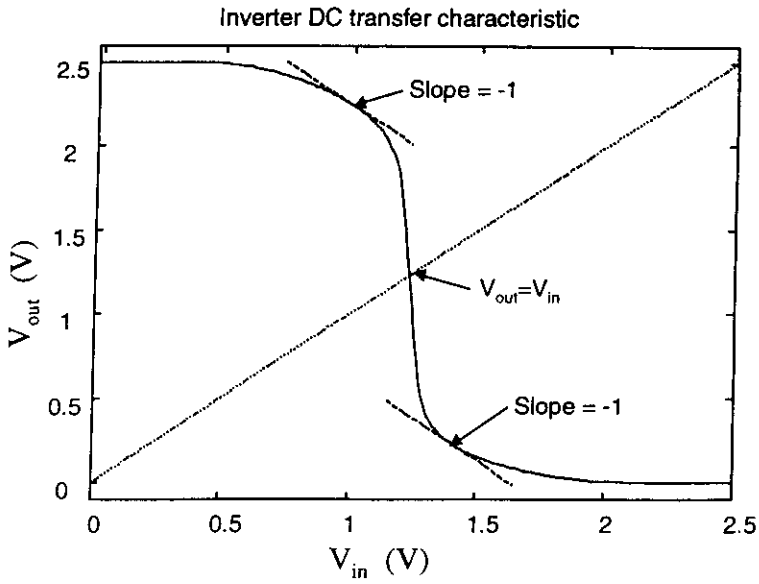
- Important characteristics:
  - It is an unipolar device
    - NMOS - charge carrier: electrons
    - PMOS - charge carrier: holes
  - It is a symmetrical device
    - Source = drain
  - High input impedance (Ig=0)
    - Low standby current in CMOS configuration
  - Voltage controlled device with high fan-out

The CMOS inverter



The CMOS inverter



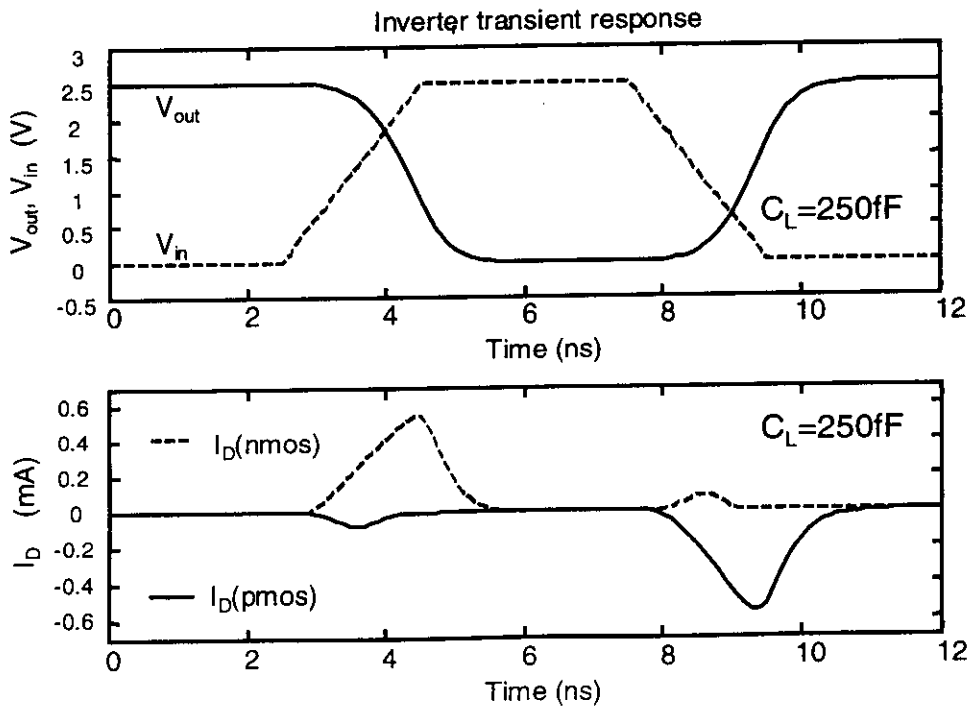


The CMOS inverter

Regions of operation (balanced inverter):

V_{in}	n-MOS	p-MOS	V_{out}
0	cut-off	linear	V_{dd}
$V_{TN}<V_{in}<V_{dd}/2$	saturation	linear	$\sim V_{dd}$
$V_{dd}/2$	saturation	saturation	$V_{dd}/2$
$V_{dd}-\|V_{TP}\|>V_{in}>V_{dd}/2$	saturation	linear	~0
V_{dd}	linear	cut-off	0

The CMOS inverter



The CMOS inverter

- Propagation delay
  - Main origin: load capacitance

$$t_{pLH} = \frac{C_L \cdot V_{dd}}{k_p (V_{dd} - |V_{TP}|)^2} \approx \frac{C_L}{k_p \cdot V_{dd}}$$

$$t_{pHL} = \frac{C_L \cdot V_{dd}}{k_n (V_{dd} - |V_{TN}|)^2} \approx \frac{C_L}{k_n \cdot V_{dd}}$$

$$t_p \approx \frac{1}{2}(t_{pLH} + t_{pLH}) = \frac{C_L}{2 \cdot V_{dd}} \left(\frac{1}{k_n} + \frac{1}{k_p} \right)$$

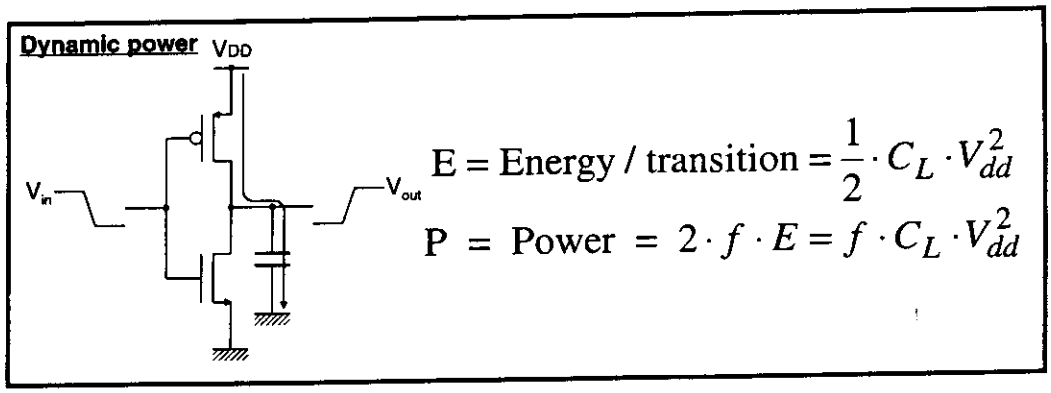
  - To reduce the delay:
    - Reduce C_L
    - Increase k_n and k_p. That is, increase W/L

The CMOS inverter

- CMOS power budget:
  - Dynamic power consumption:
    - Charging and discharging of capacitors
  - Short circuit currents:
    - Short circuit path between power rails during switching
  - Leakage
    - Leaking diodes and transistors

The CMOS inverter

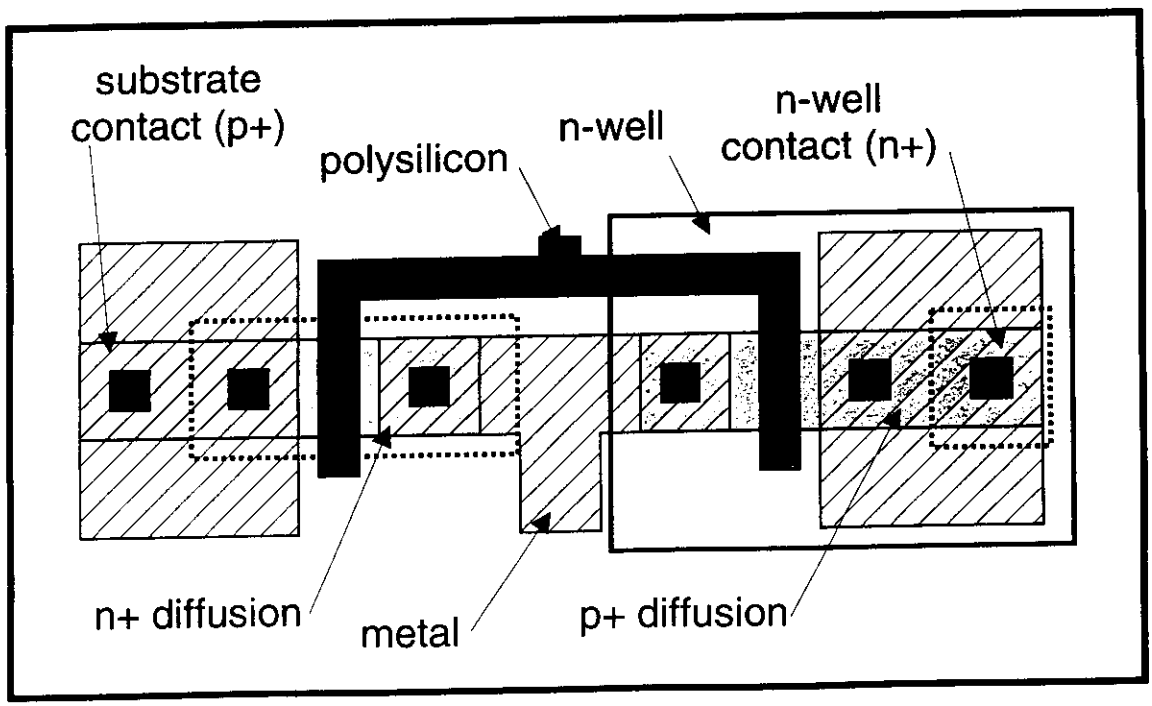
- Dynamic power dissipation
  - Function of the transistors size
    - Gate and parasitic capacitances
  - To reduce dynamic power dissipation
    - Reduce: C_L
    - Reduce: V_{dd} ⇐ The most effective action
    - Reduce: f



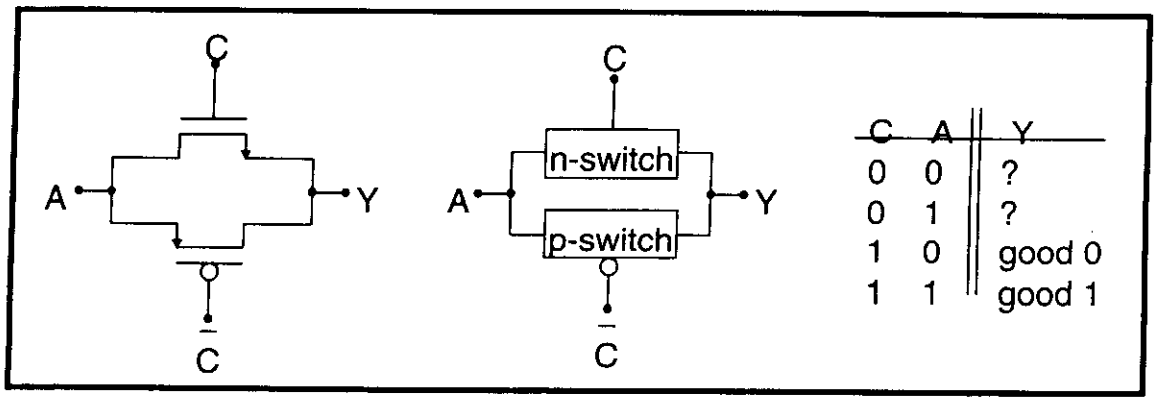
$$E = \text{Energy / transition} = \frac{1}{2} \cdot C_L \cdot V_{dd}^2$$

$$P = \text{Power} = 2 \cdot f \cdot E = f \cdot C_L \cdot V_{dd}^2$$

The CMOS inverter



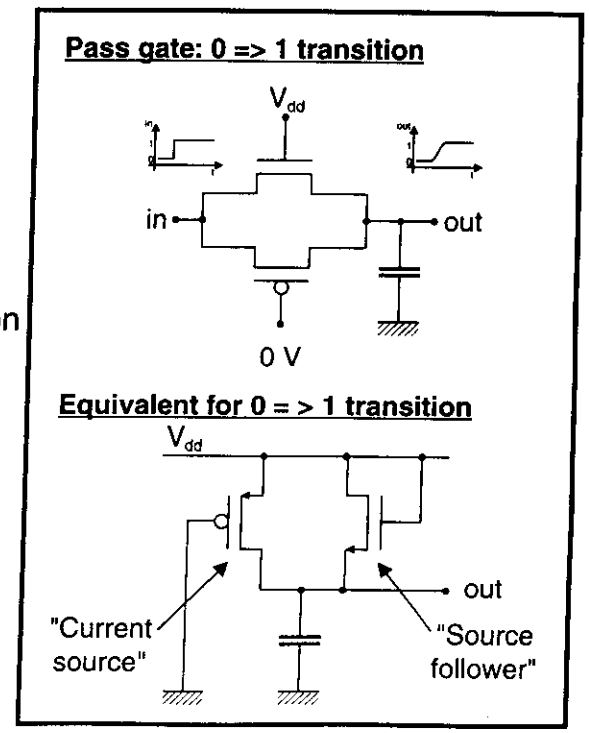
The CMOS pass gate



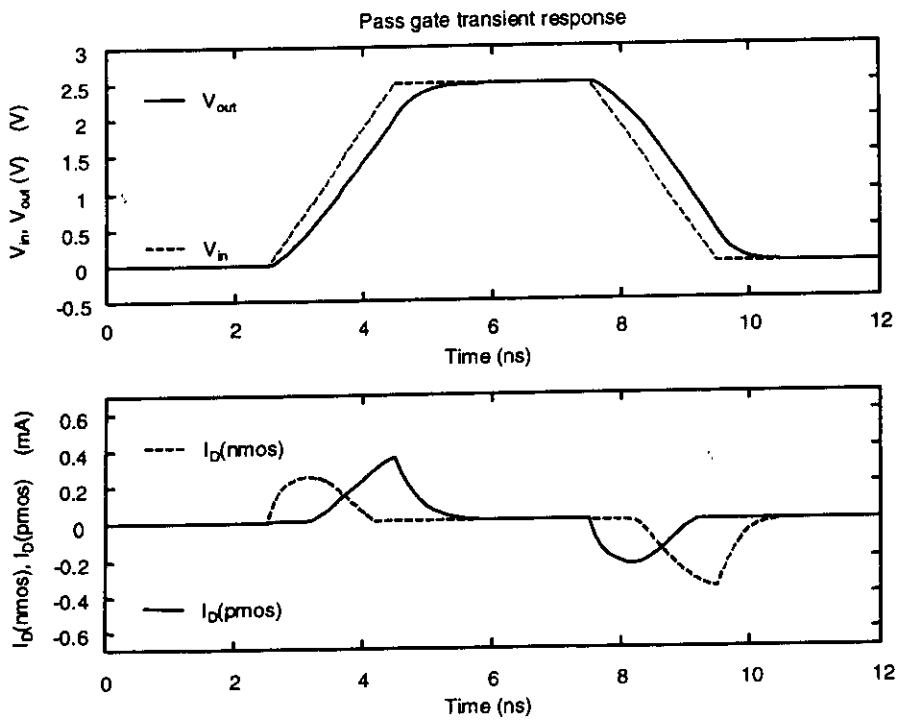
The CMOS pass gate

Regions of operation:
"0" to "1" transition

- NMOS:
  - source follower
  - $V_{gs} = V_{ds}$ always:
    - $V_{out} < V_{dd}-V_{TN} \Rightarrow$ saturation
    - $V_{out} > V_{dd}-V_{TN} \Rightarrow$ cutoff
  - $V_{TN} > V_{TN0}$ (bulk effect)
- PMOS:
  - current source
  - $V_{out} < |V_{TP}| \Rightarrow$ saturation
  - $V_{out} > V_{TP} \Rightarrow$ linear



The CMOS pass gate



The CMOS pass gate

- Regions of operation: "0" to "1" transition

$V_{out} < \|V_{TP}\|$	NMOS and PMOS saturated
$\|V_{TP}\| < V_{out} < V_{dd} - V_{TN}$	NMOS saturated, PMOS linear
$V_{out} > V_{dd} - V_{TN}$	NMOS cutoff, PMOS linear

- Regions of operation: "1" to "0" transition

$V_{out} > Vdd - V_{TN}$	NMOS and PMOS saturated
$V_{dd} - V_{TN} > V_{out} > \|V_{TP}\|$	NMOS linear, PMOS saturated
$V_{TP} > V_{out}$	NMOS linear, PMOS cutoff

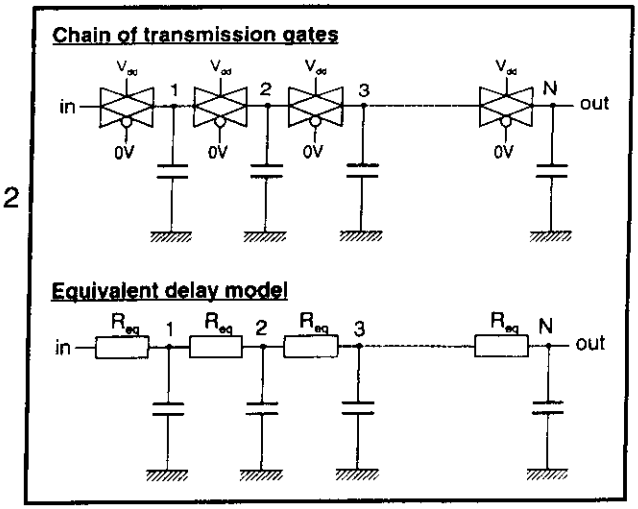
- Both devices combine to form a good switch

The CMOS pass gate

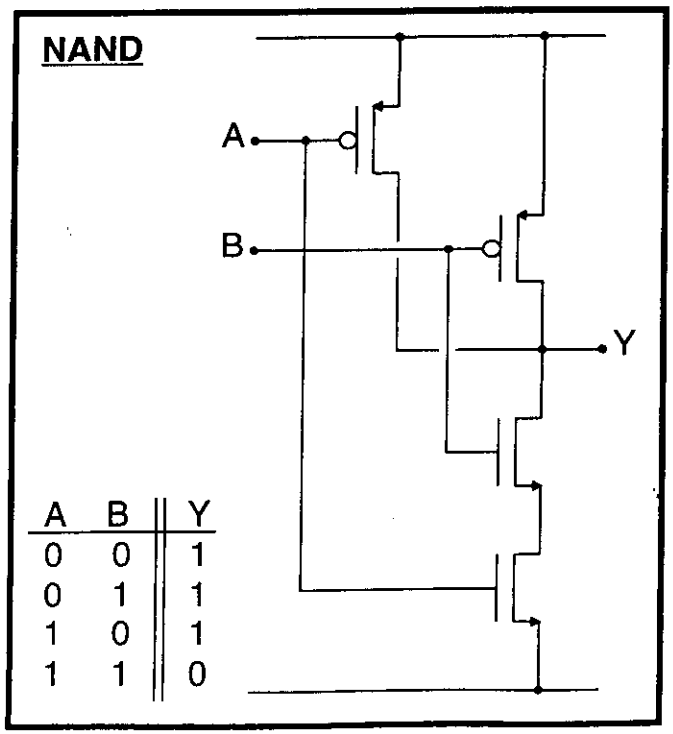
- Delay of a chain of pass gates:

$$t_d \propto C \cdot R_{eq} \cdot \frac{N \cdot (N+1)}{2}$$

- Delay proportional to N^2
- Avoid N large:
  - Break the chain by inserting buffers

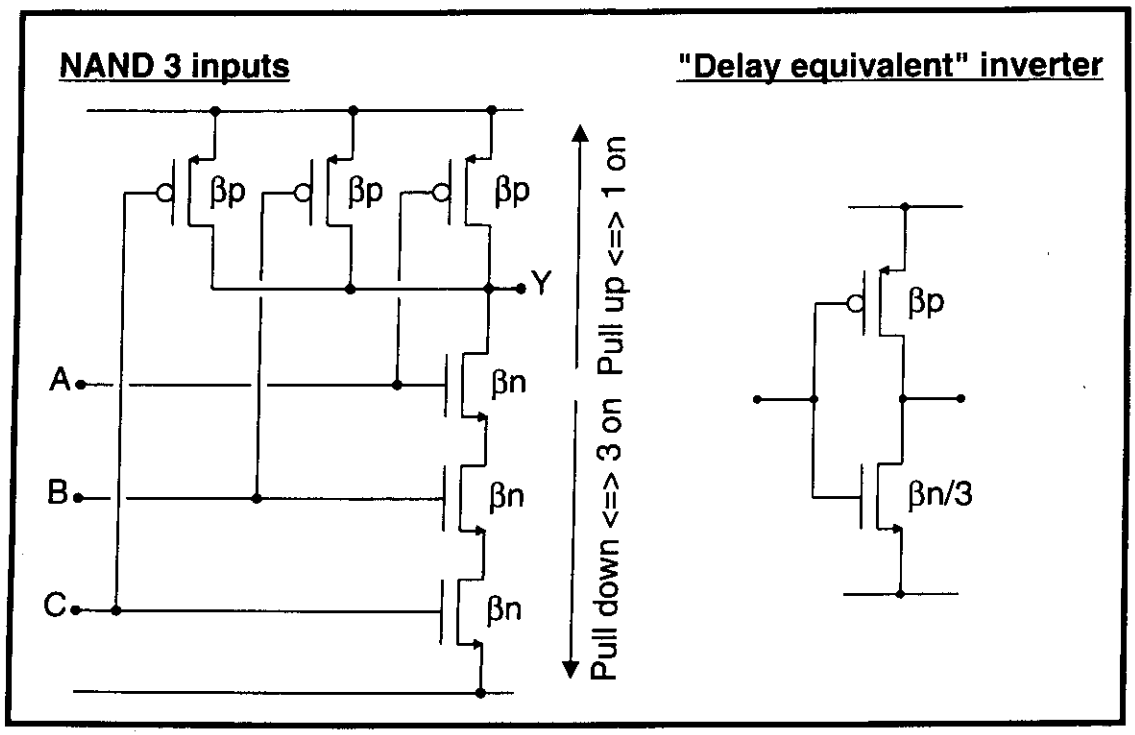


Simple CMOS gates

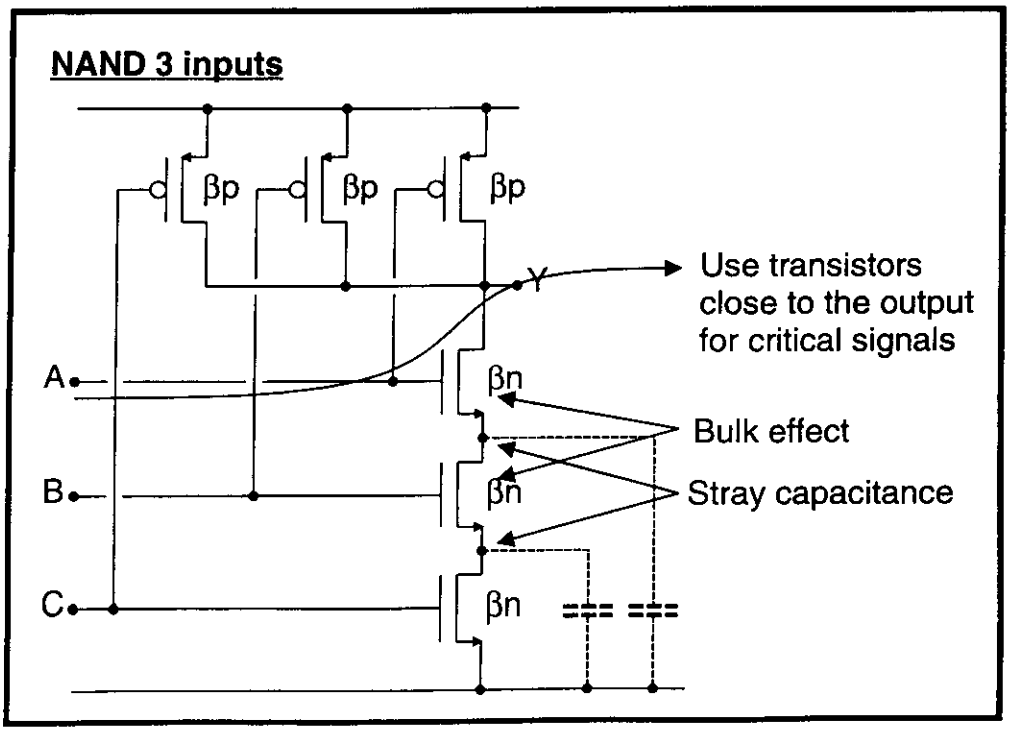


A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

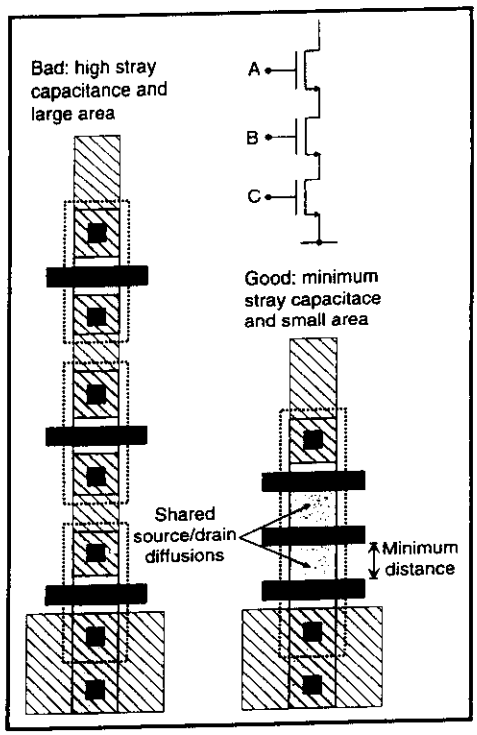
Simple CMOS gates



Simple CMOS gates



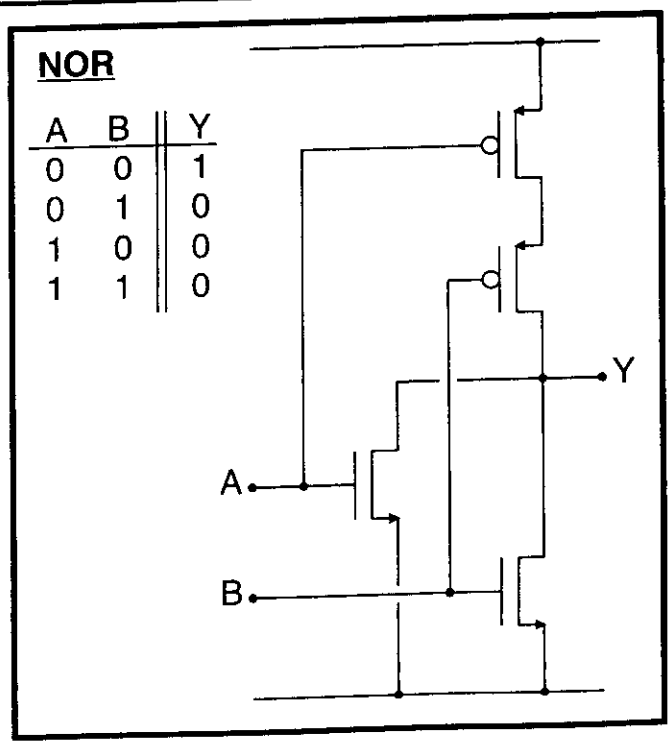
Simple CMOS gates



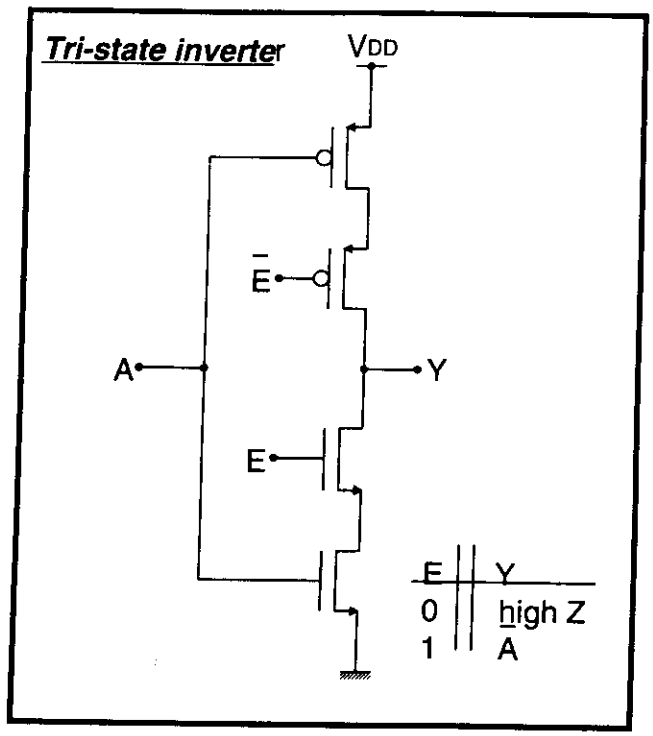
Simple CMOS gates

NOR

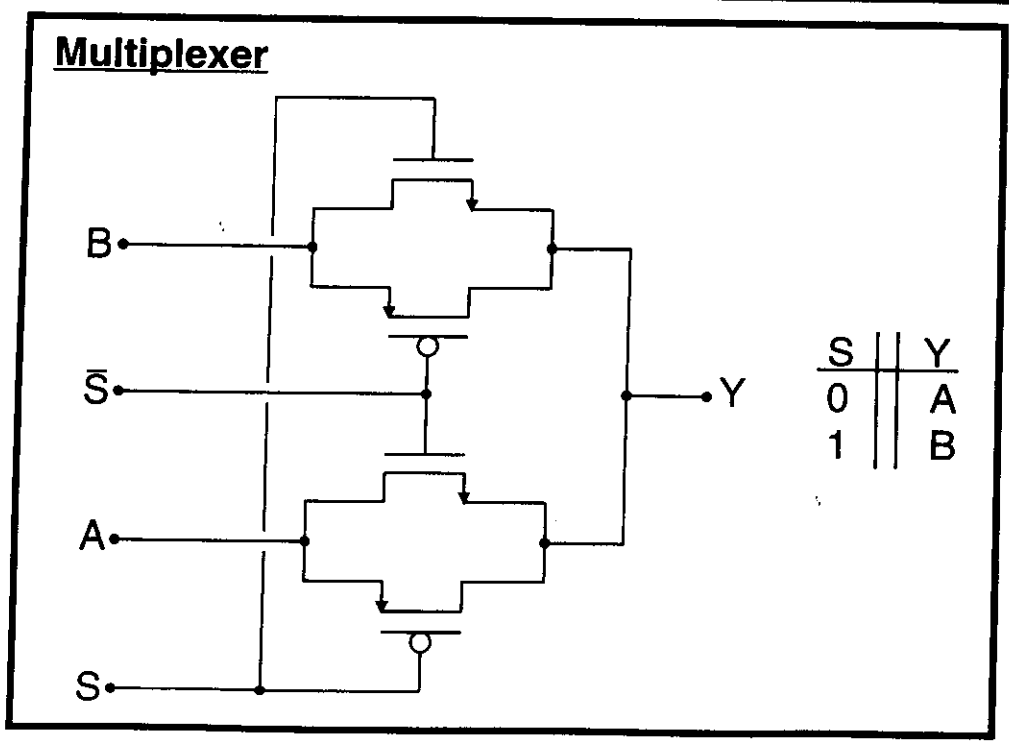
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0



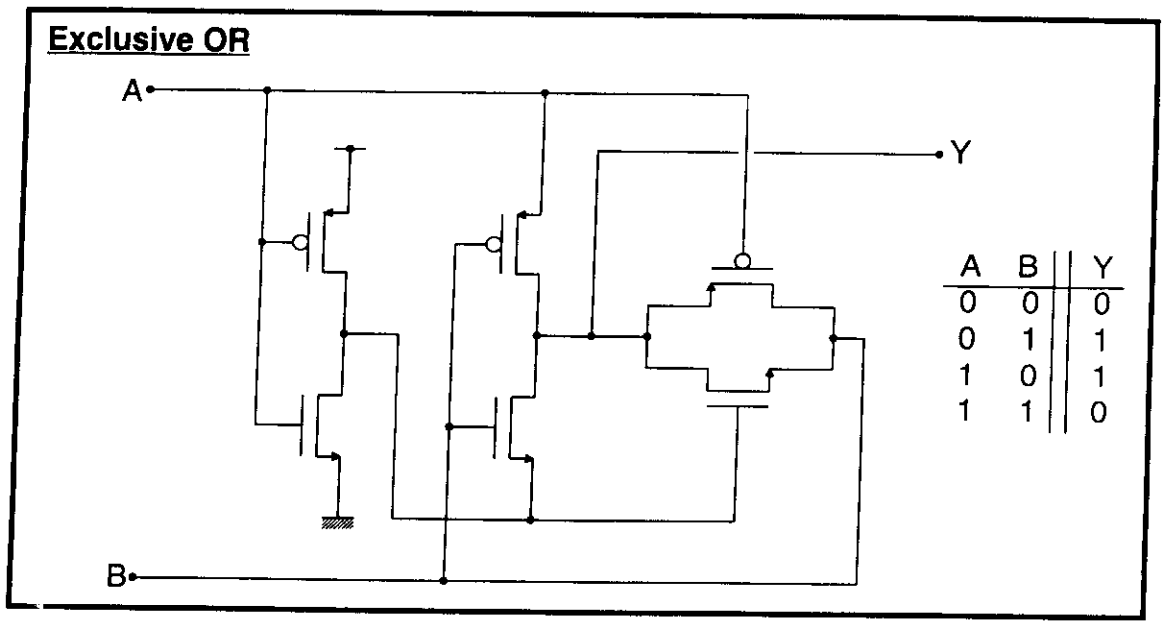
Simple CMOS gates



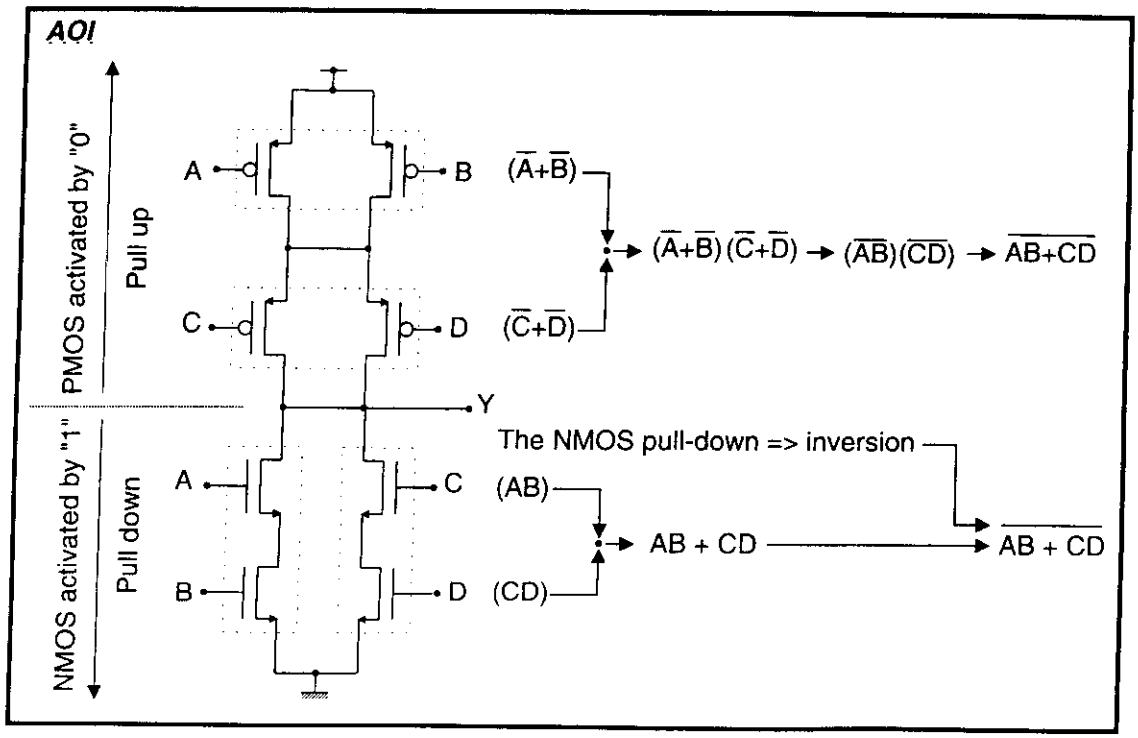
Complex CMOS gates



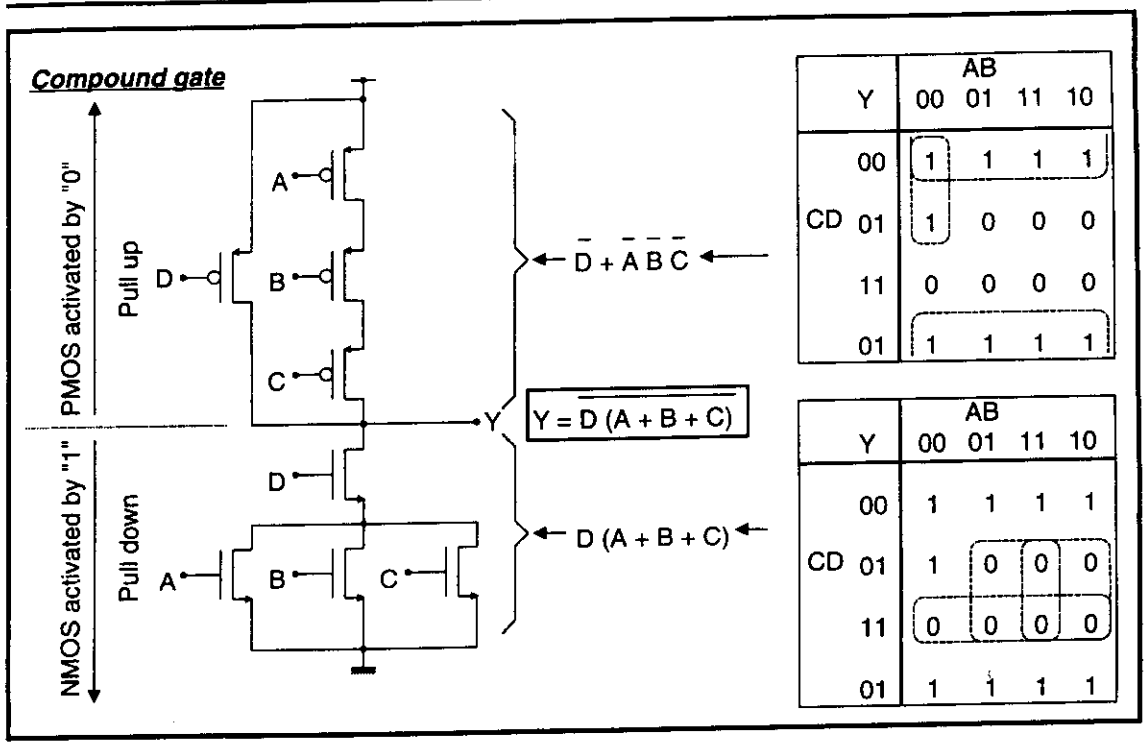
Complex CMOS gates



Complex CMOS gates



Complex CMOS gates



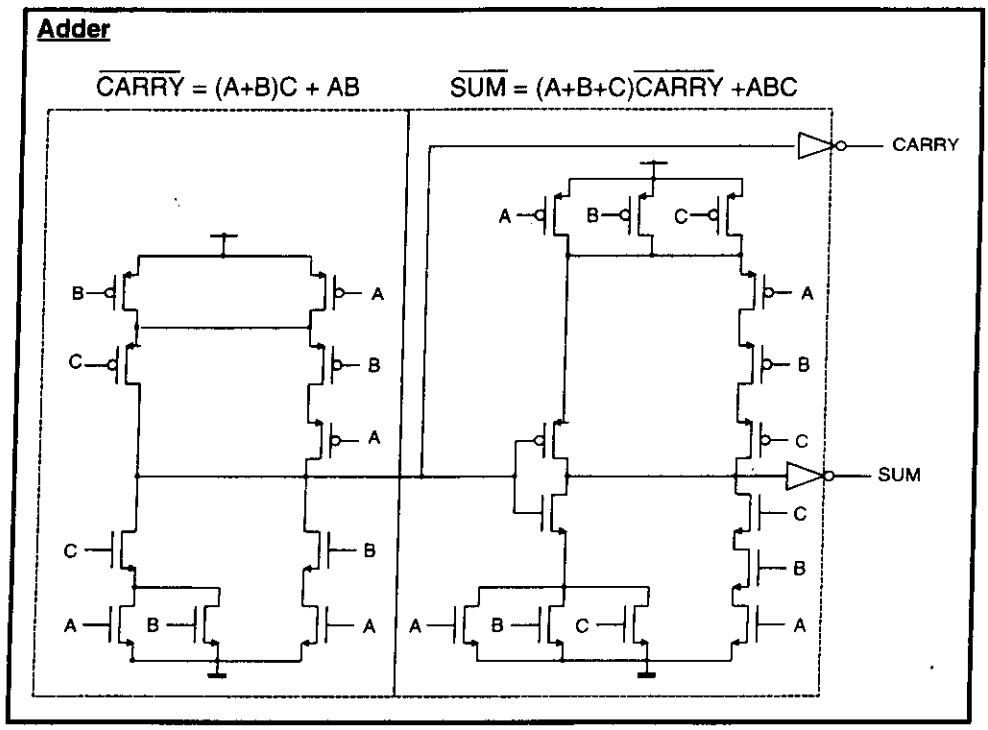
Complex CMOS gates

- Can a compound gate be arbitrarily complex?
  - NO, propagation delay is a strong function of fan-in:

$$t_p = a_0 \cdot FO + a_1 \cdot FI + a_2 \cdot (FI)^2$$

  - FO ⇒ Fan-out, number of loads connected to the gate:
    - 2 gate capacitances per FO + interconnect
  - FI ⇒ Fan-in, Number of inputs in the gate:
    - Quadratic dependency on FI due to:
      - Resistance increase
      - Capacitance increase
  - Avoid large FI gates (Typically FI ≤ 4)

Complex CMOS gates



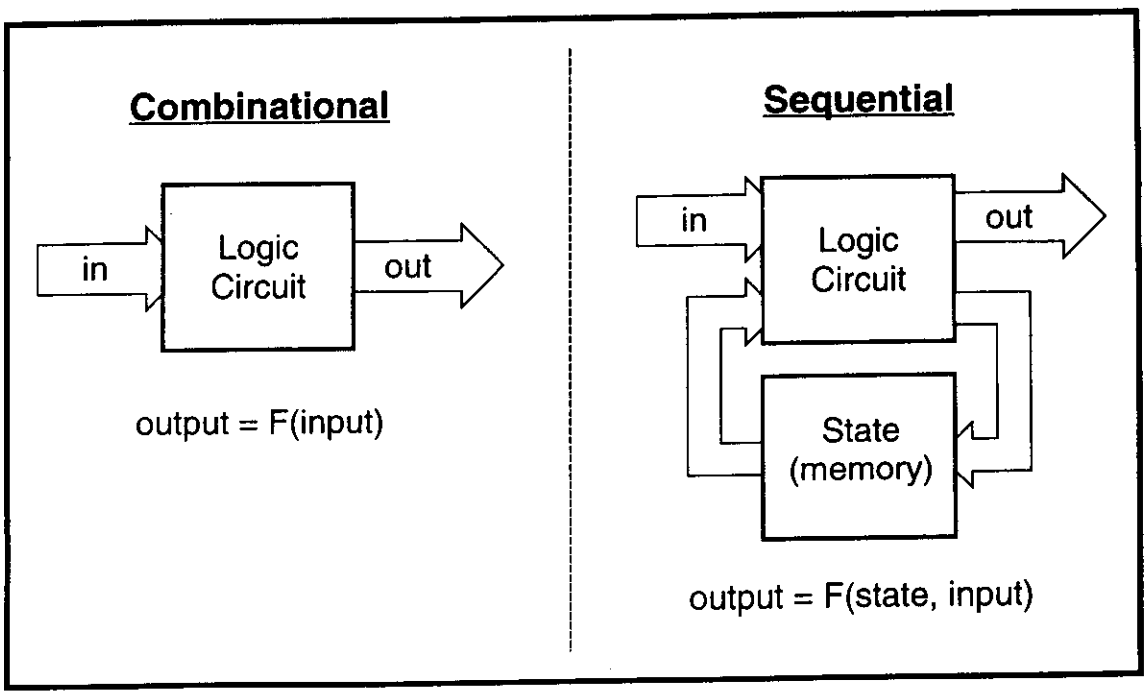
Outline

- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS sequential circuits

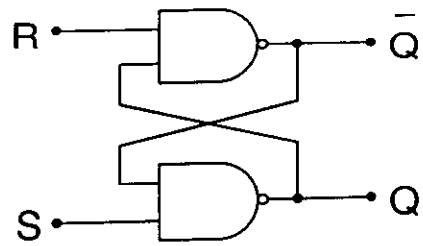
- Sequential circuits
- Interconnects
- Clock distribution
- DLL's and PLL's

Sequential circuits



Sequential circuits

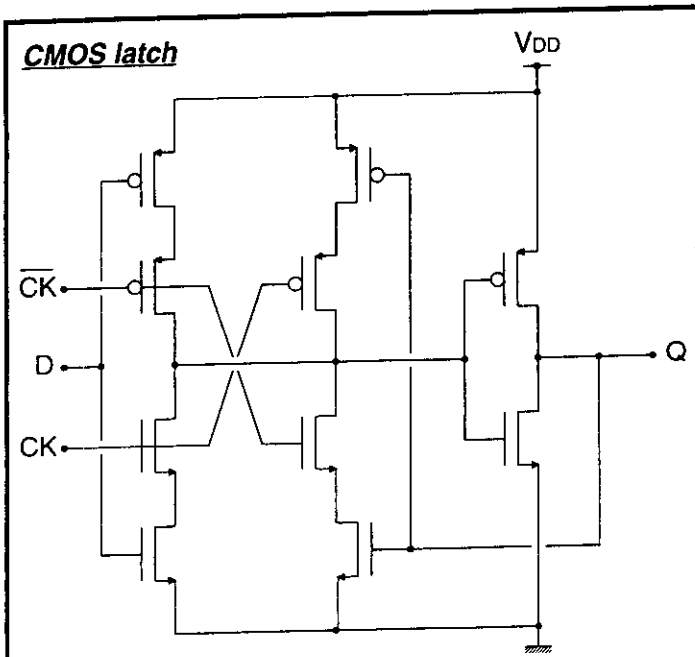
RS latch

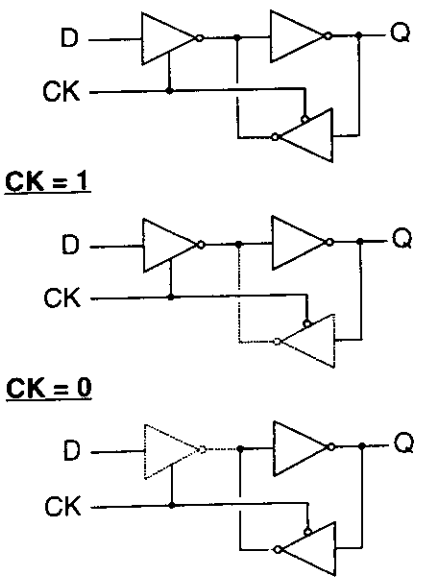


R	S	Q_{n+1}	$\overline{Q}_{n+1}$	
0	0	1	1	(ilegal)
0	1	0	1	
1	0	1	0	
1	1	Q_n	$\overline{Q}_n$	(memory)

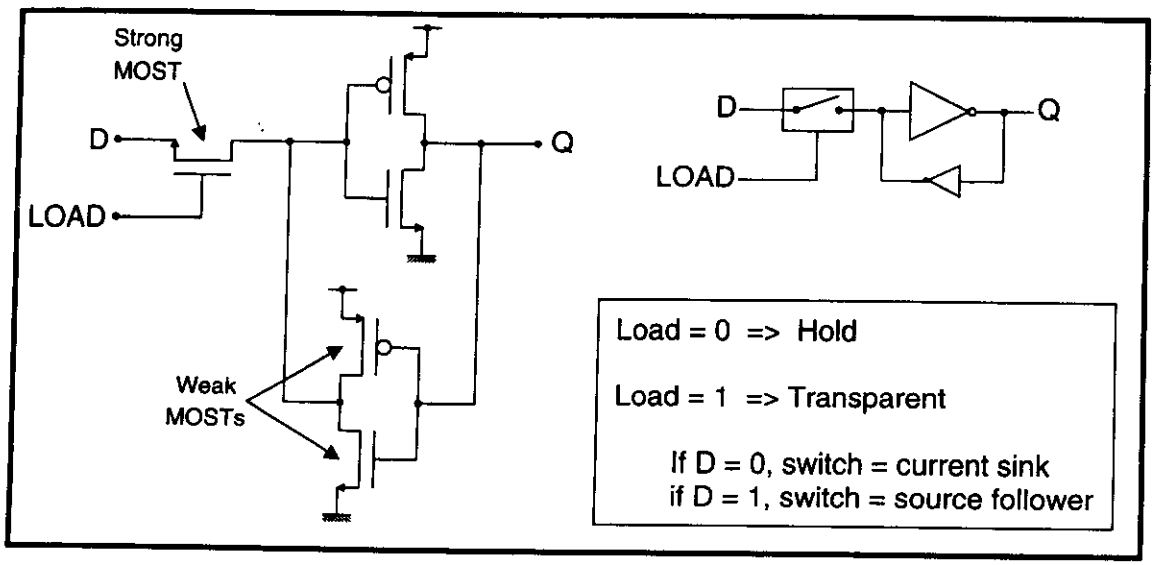
Sequential circuits

CMOS latch

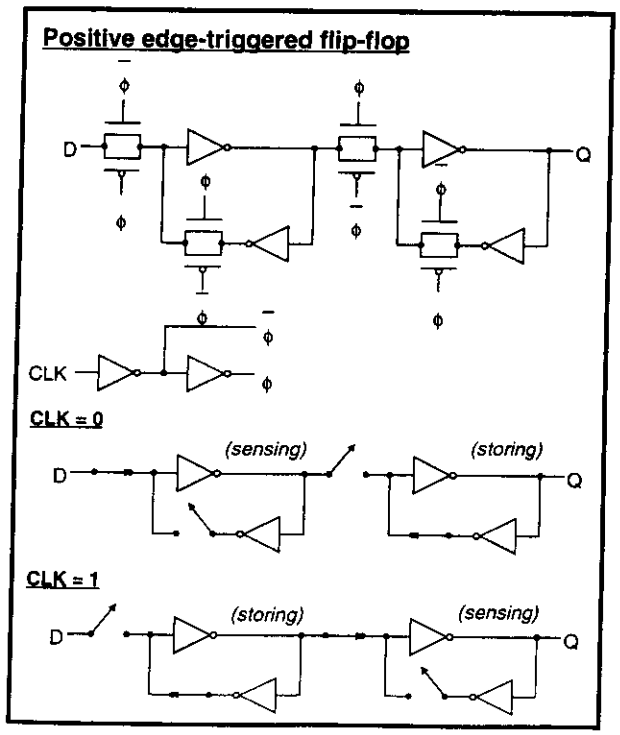




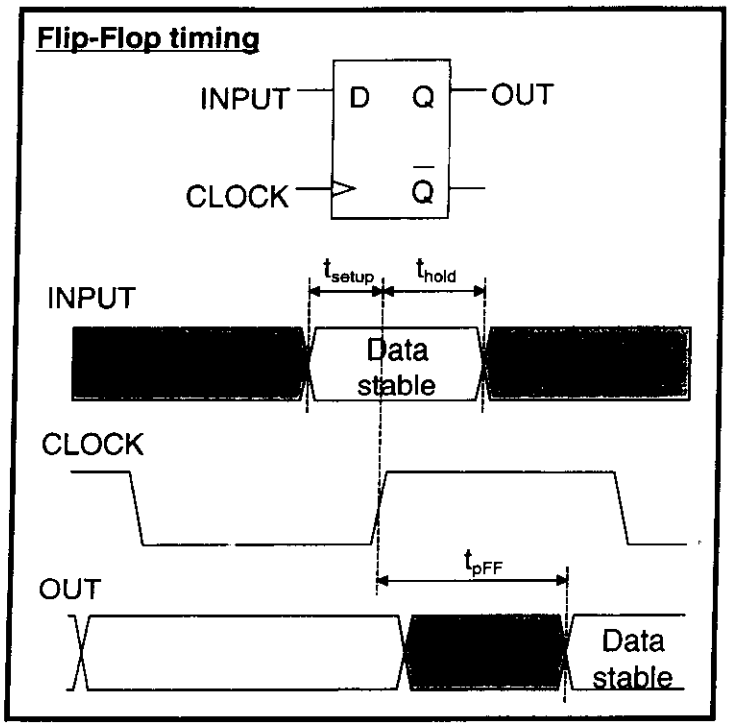
Sequential circuits



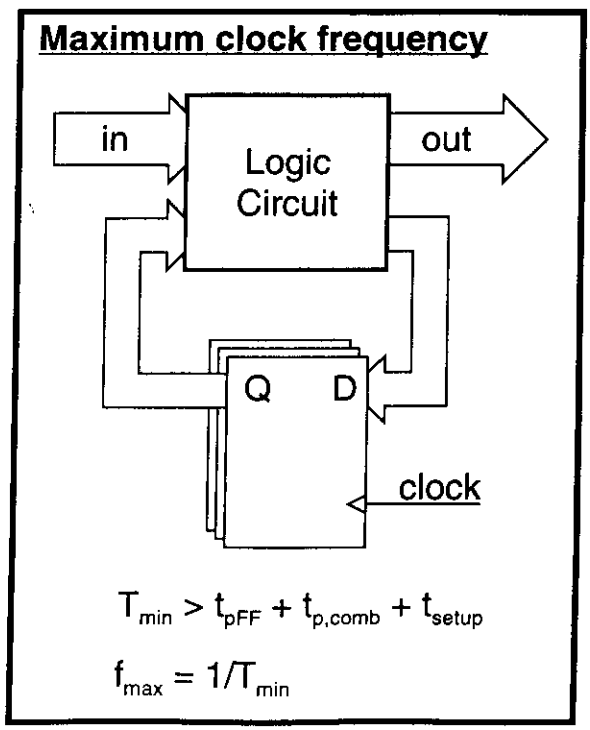
Sequential circuits



Sequential circuits



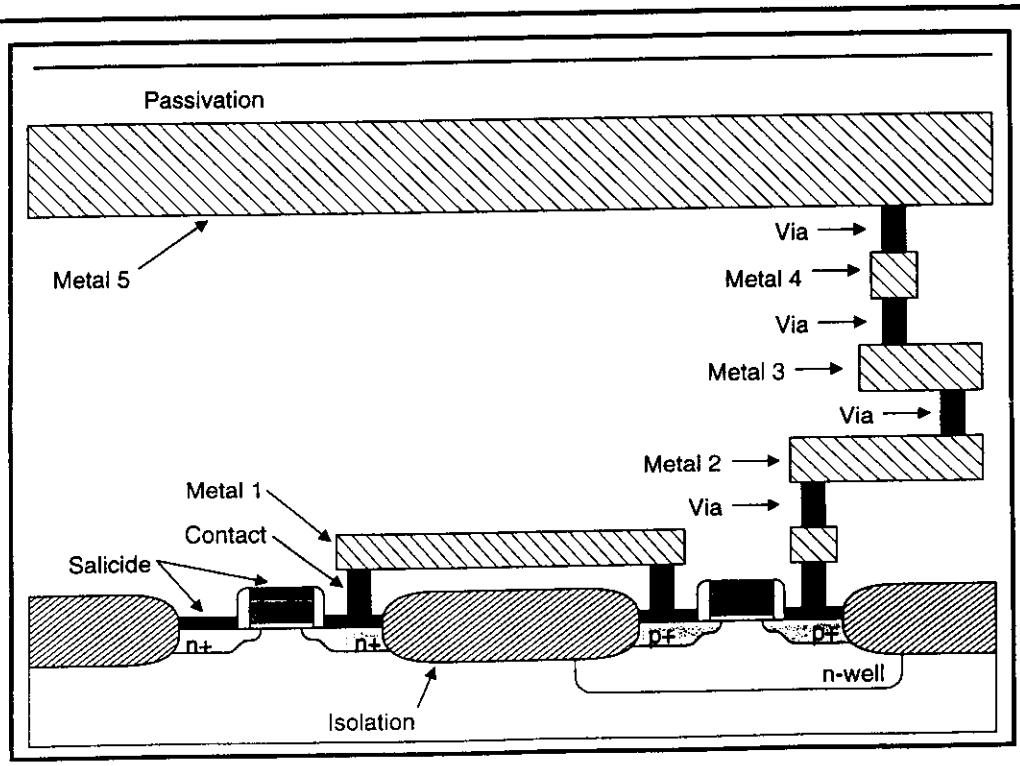
Sequential circuits



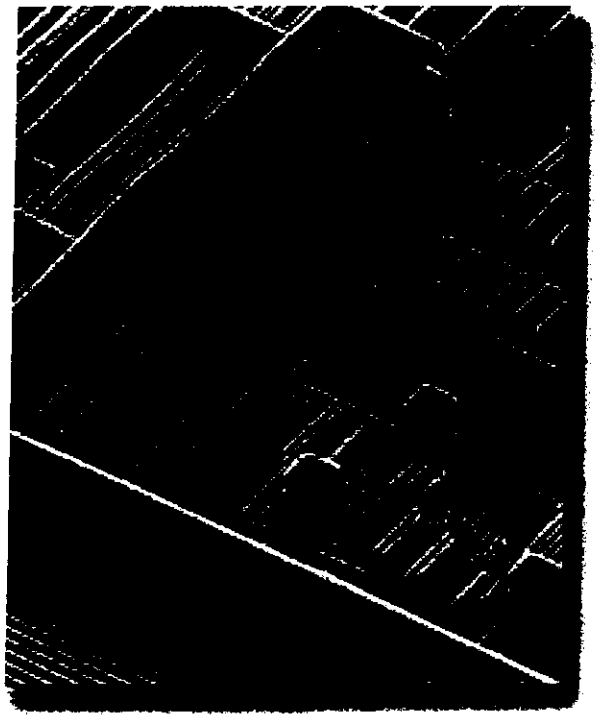
Interconnects

- The previous result assumes that signals can propagate instantaneously across interconnects
- In reality interconnects are metal or polysilicon structures with associated resistance and capacitance.
- That, introduces signal propagation delay that has to be taken into account for reliable operation of the circuit

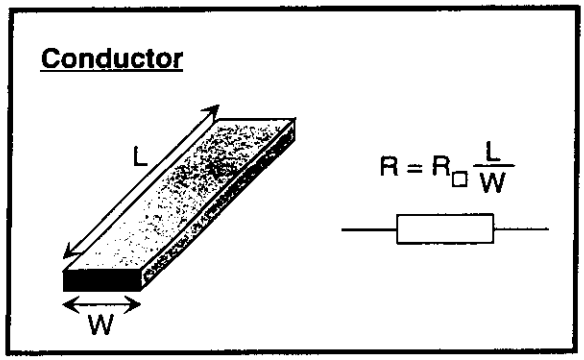
Interconnects



Interconnects



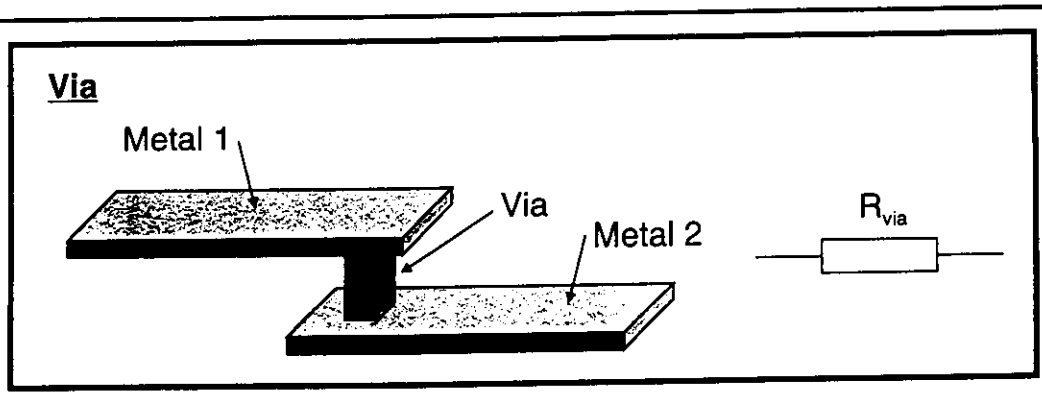
Interconnects



Film	Sheet resistance (Ω/square)
n-well	310
p+, n+ diffusion (salicided)	4
polysilicon (salicided)	4
Metal 1	0.12
Metal 2, 3 and 4	0.09
Metal 5	0.05

(Typical values for an advanced process)

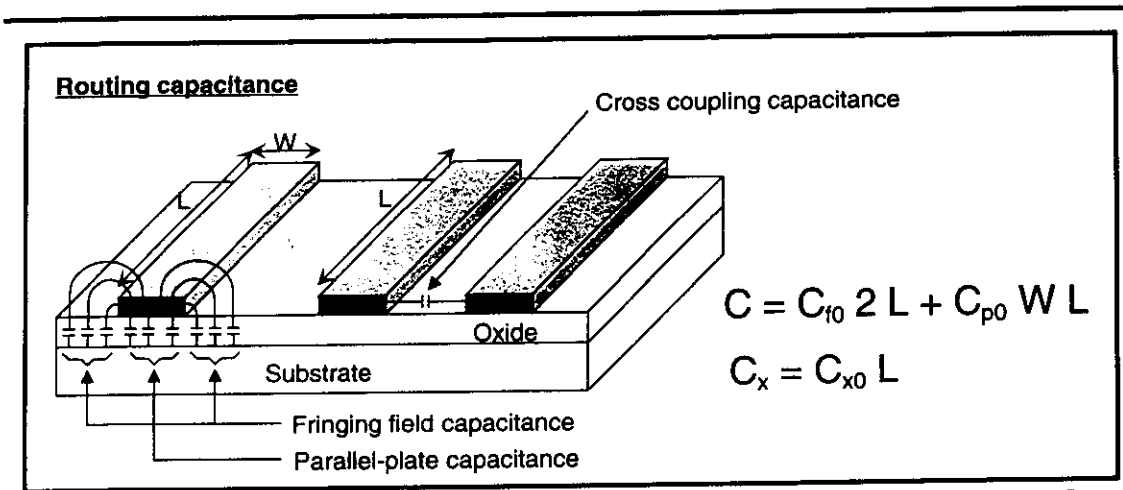
Interconnects



- Via or contact resistance depends on:
  - The contacted materials
  - The contact area

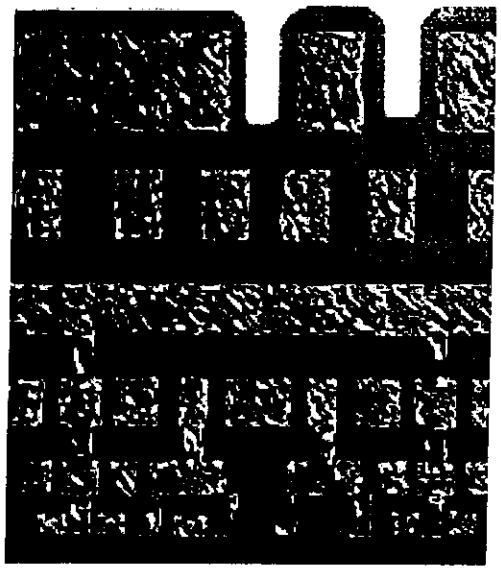
Via/contact	Resistance (Ω)
M1 to n+ or p+	10
M1 to Polysilicon	10
V1, 2, 3 and 4	7

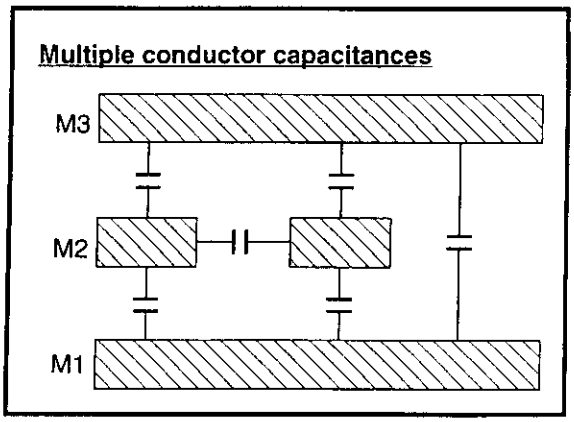
Interconnects



Interconnect layer	Parallel-plate (fF/μm²)	Fringing (fF/μm)
Polysilicon to sub.	0.058	0.043
Metal 1 to sub.	0.031	0.044
Metal 2 to sub.	0.015	0.035
Metal 3 to sub.	0.010	0.033

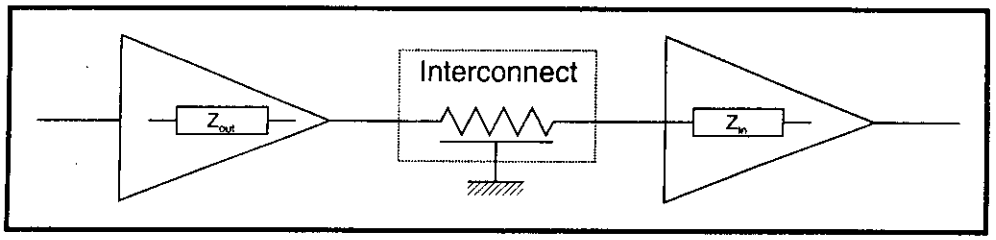
Interconnects





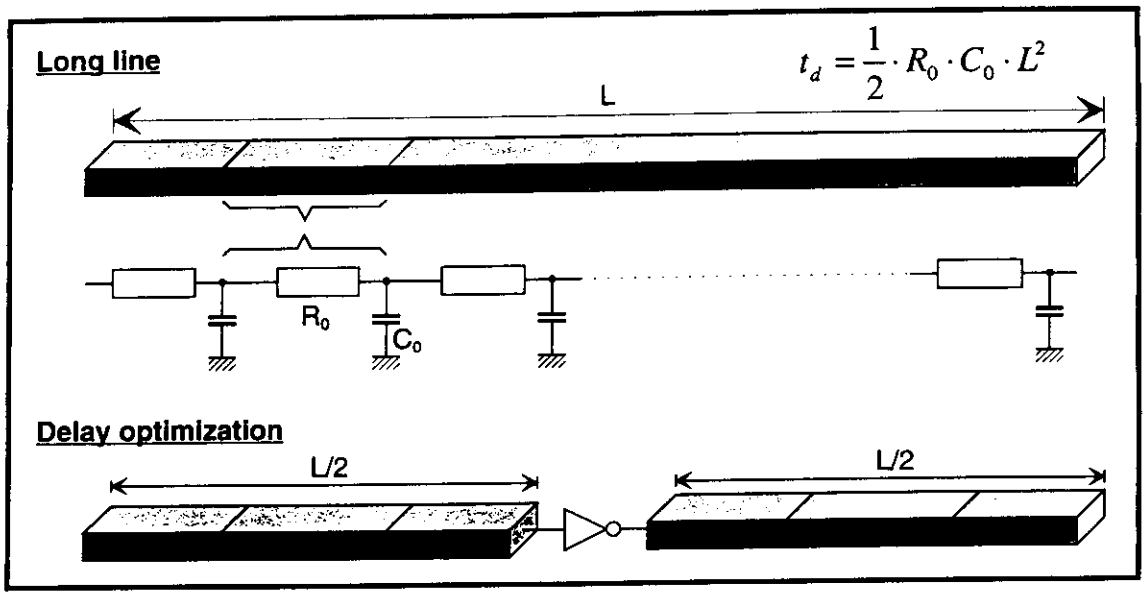
- Three dimensional field simulators are required to accurately compute the capacitance of a multi-wire structure

Interconnects



- Delay depends on:
  - Impedance of the driving source
  - Distributed resistance/capacitance of the wire
  - Load impedance
- Distributed RC delay:
  - Can be dominant in long wires
  - Important in polysilicon wires (relatively high resistance)
  - Important in salicided wires
  - Important in heavily loaded wires

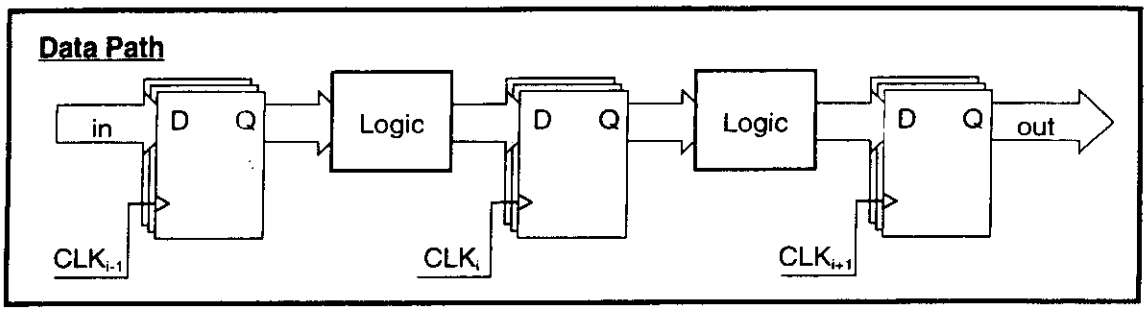
Interconnects



Clock distribution

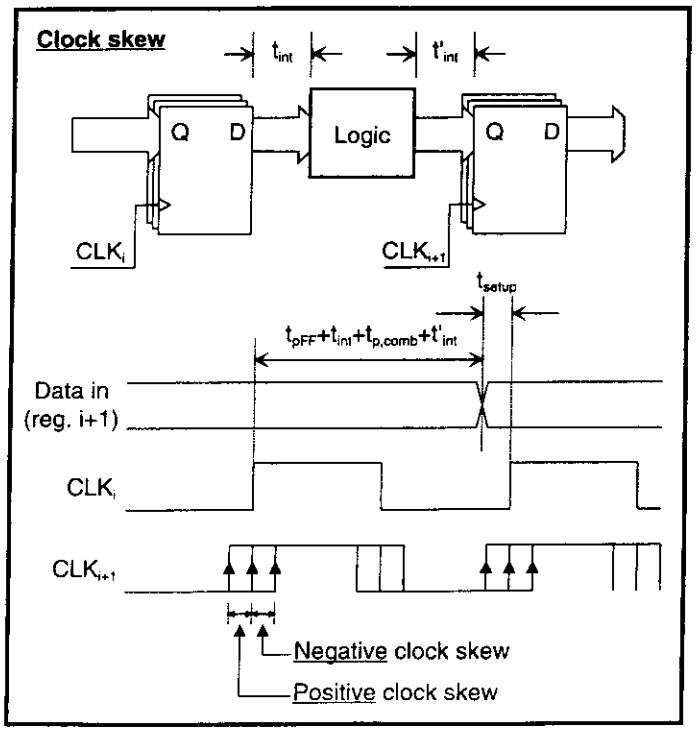
- Clock signals are "special signals"
- Every data movement in a synchronous system is referenced to the clock signal
- Clock signals:
  - Are typically loaded with high fanout
  - Travel over the longest distances in the IC
  - Operate at the highest frequencies

Clock distribution



- "Equipotential" clocking:
  - In a synchronous system all clock signals are derived from a single clock source ("clock reference")
  - Ideally: clocking events should occur at all registers simultaneously ... = t(clk$_{i-1}$) = t(clk$_i$) = t(clk$_{i+1}$) = ...
  - In practice: clocking events will occur at slightly different instants among the different registers in the data path

Clock distribution



Clock distribution

- Skew: difference between the clocking instants of two "sequential" registers:
  Skew = $t(CLK_i) - t(CLK_{i+1})$
- Maximum operation frequency:

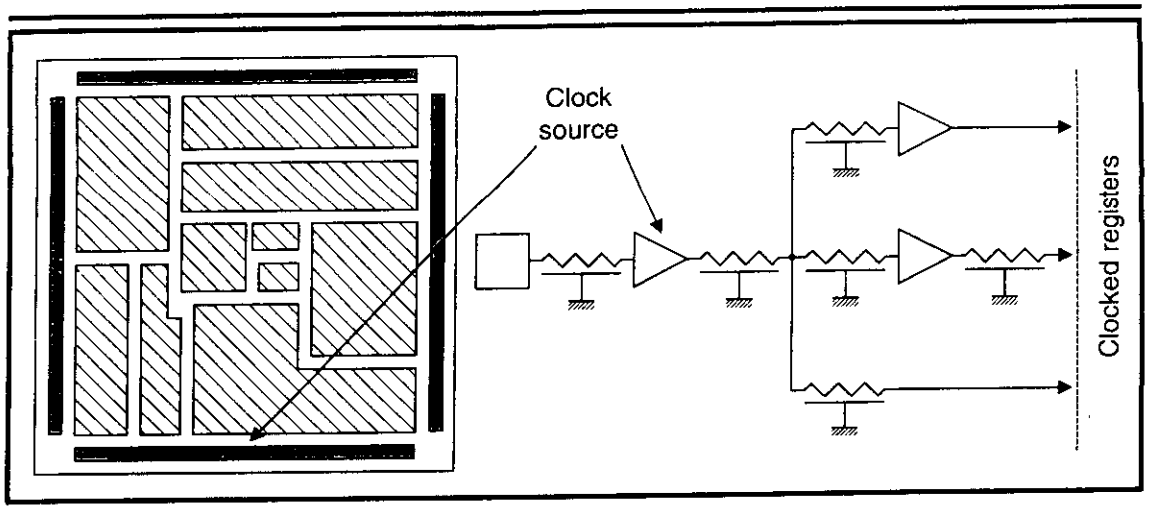
$$T_{\min} = \frac{1}{f_{\max}} = t_{dFF} + t_{\text{int}} + t_{p,comb} + t'_{\text{int}} + t_{setup} + t_{skew}$$

- Skew > 0, decreases the operation frequency
- Skew < 0, can be used to compensate a critical data path BUT this results in more positive skew for the next data path!

Clock distribution

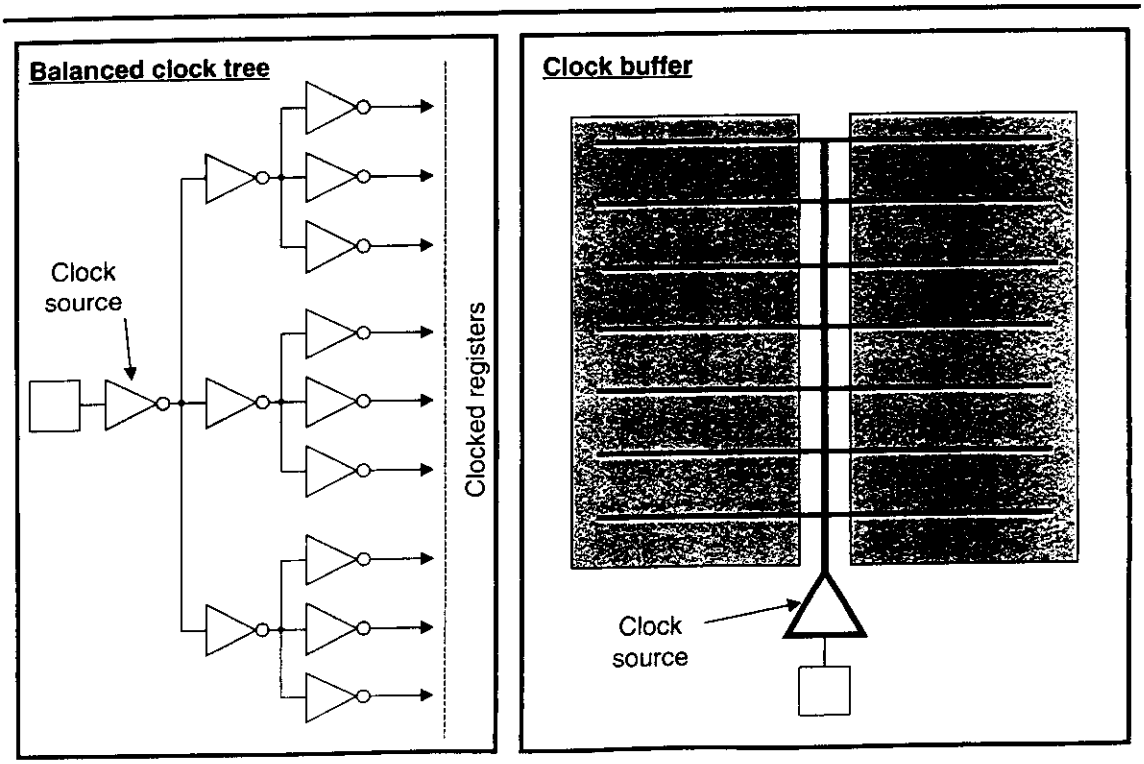
- Different clock paths can have different delays due to:
  - Differences in line lengths from clock source to the clocked registers
  - Differences in delays in the active buffers within the clock distribution network:
    - Differences in passive interconnect parameters (line resistance/capacitance, line dimensions, ...)
    - Differences in active device parameters (threshold voltages, channel mobility)
- In a well designed and balanced clock distribution network, the distributed clock buffers are the principal source of clock skew

Clock distribution

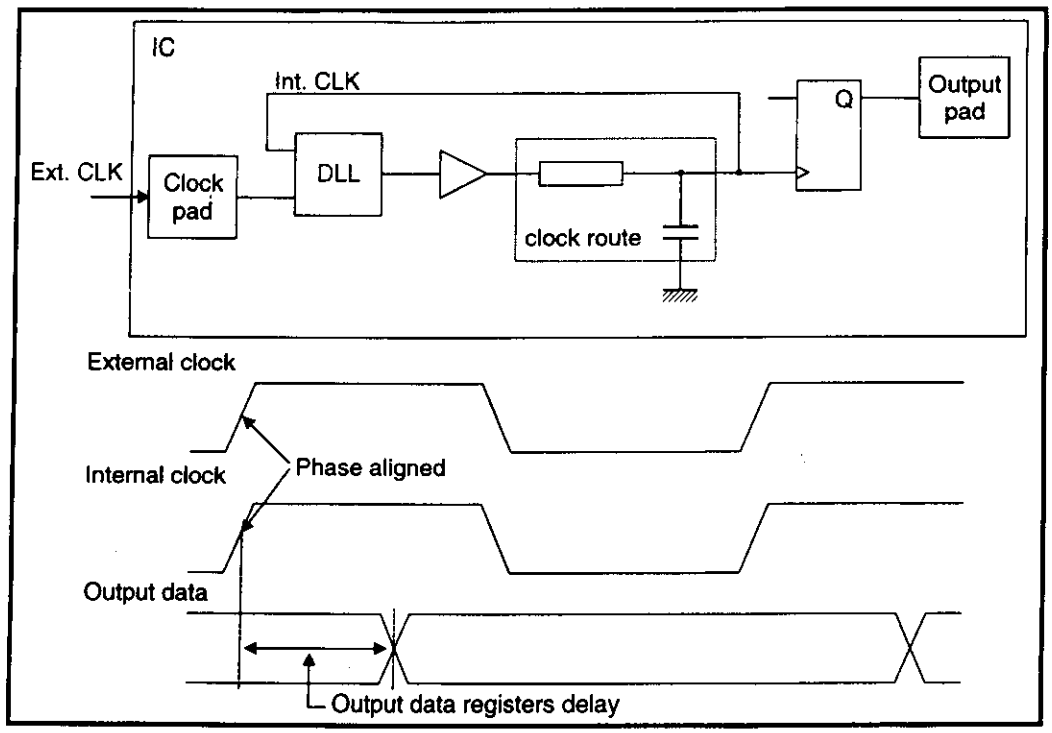


- Clock buffers:
  - Amplify the clock signal degraded by the interconnect impedance
  - Isolate the local clock lines from upstream load impedances

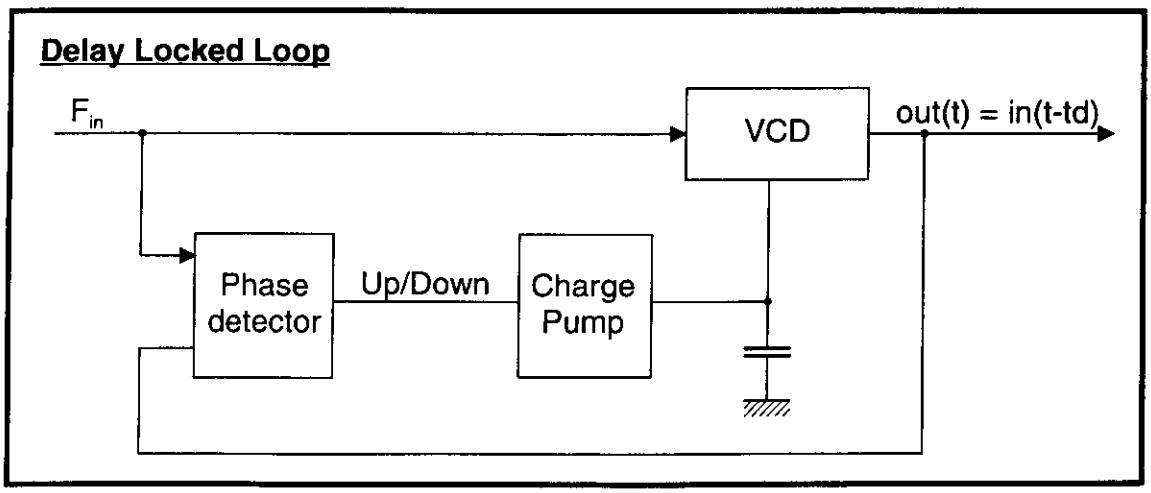
Clock distribution



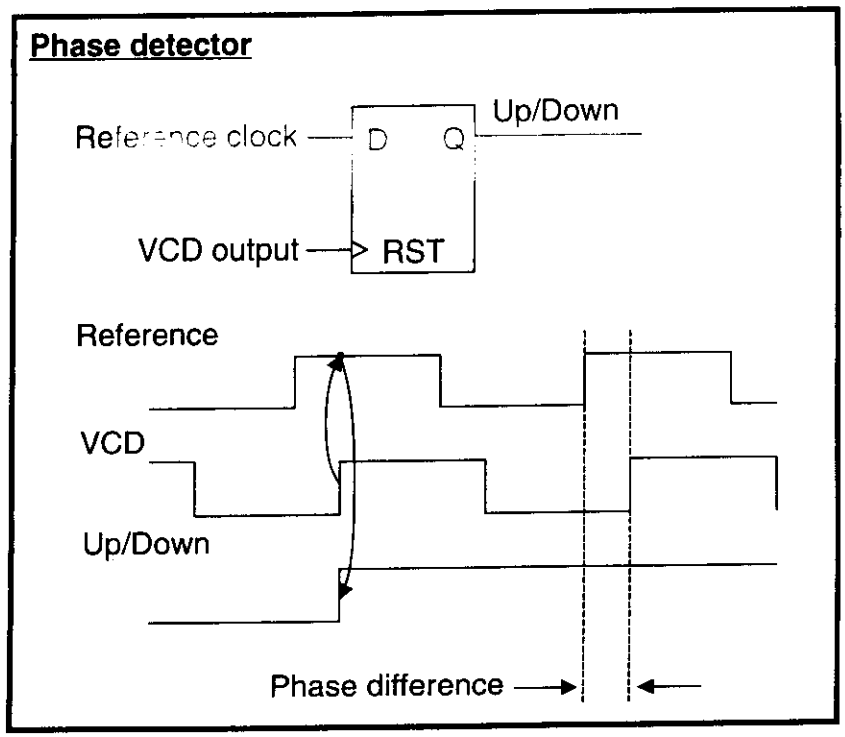
Delay locked loops



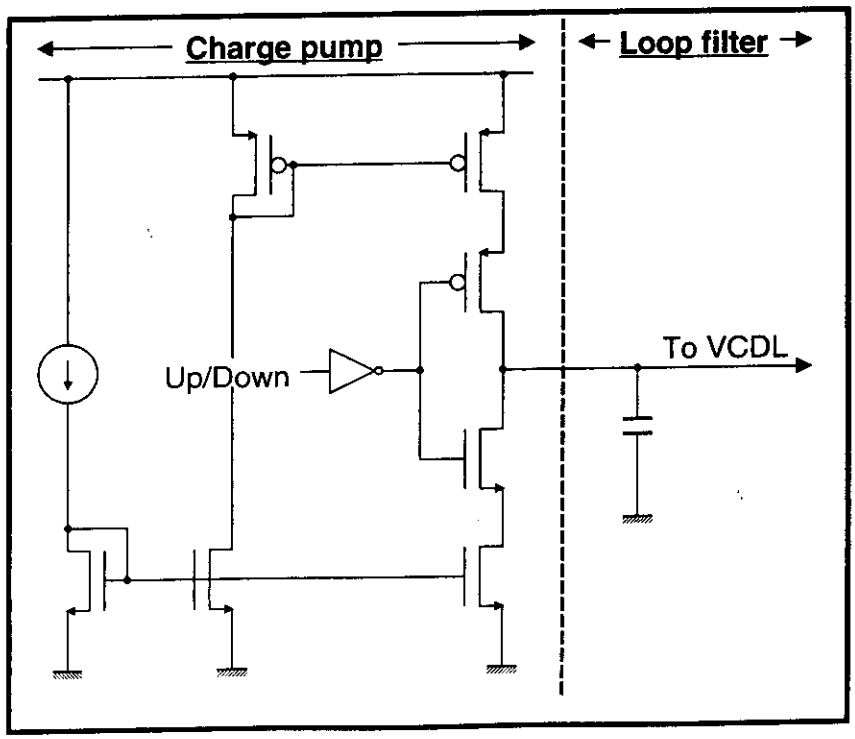
Delay locked loops



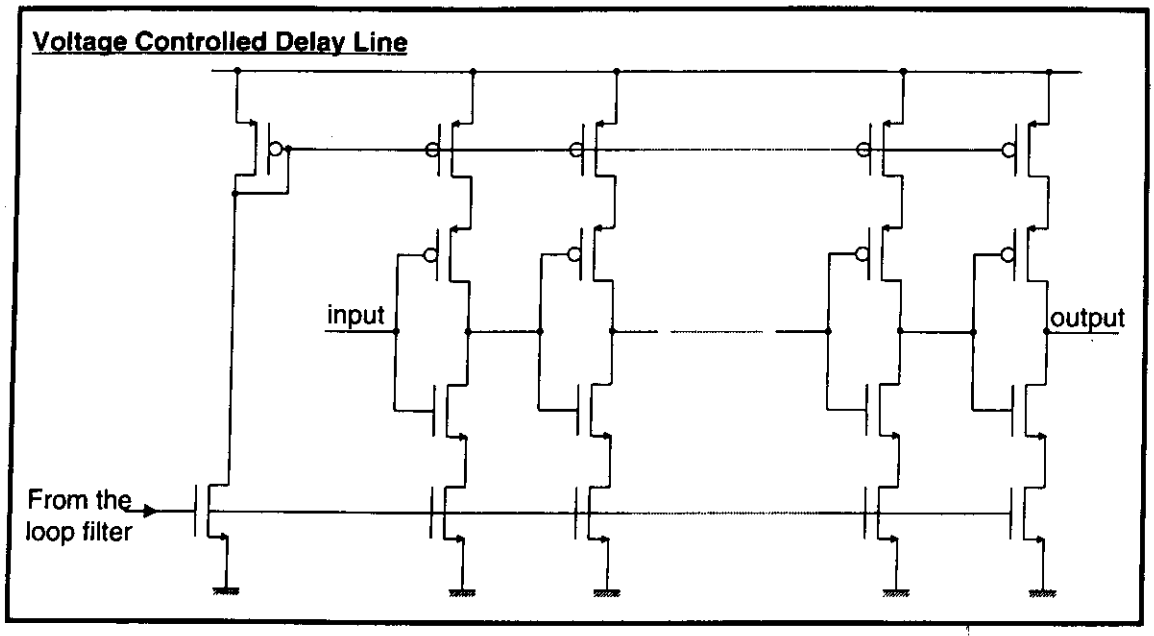
Delay locked loops



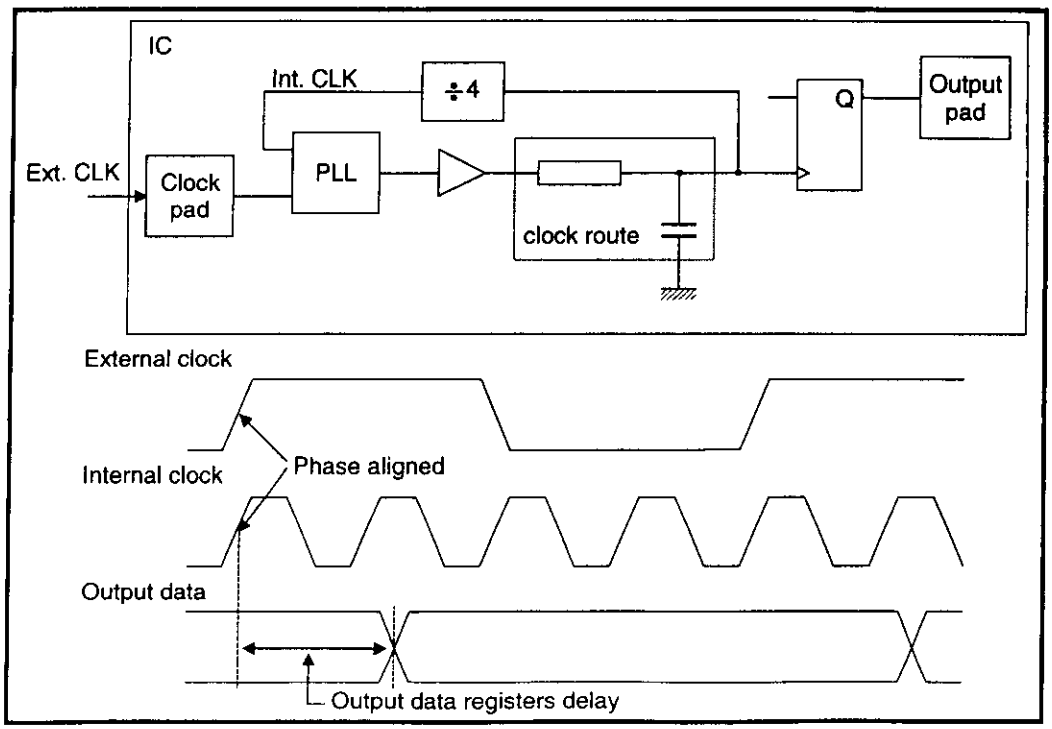
Delay locked loops



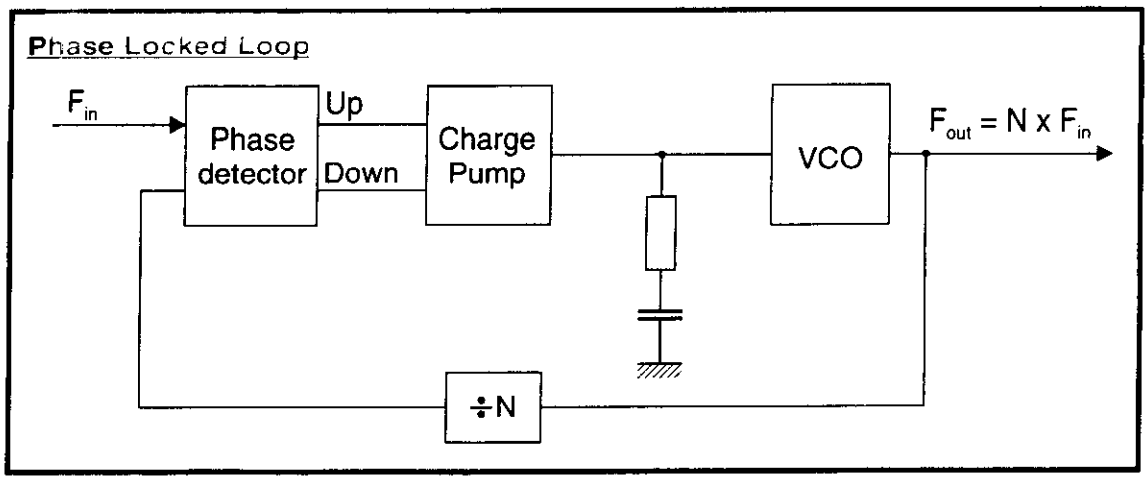
Delay locked loops



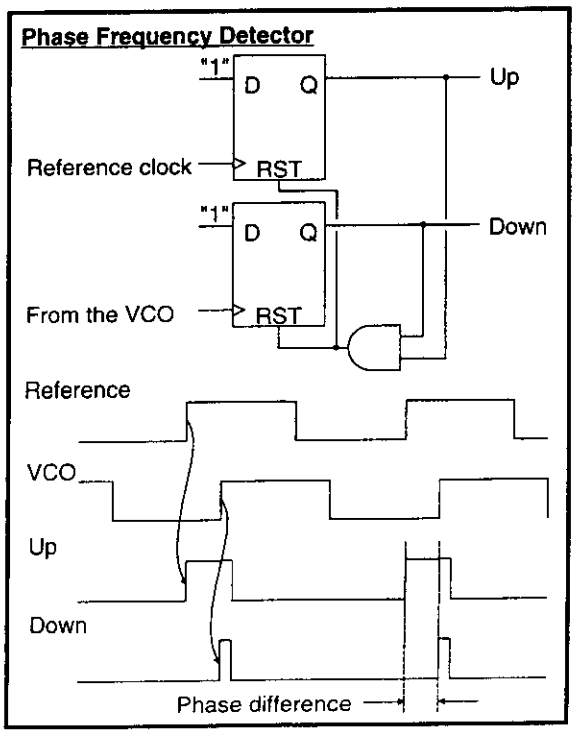
Phase Locked Loops



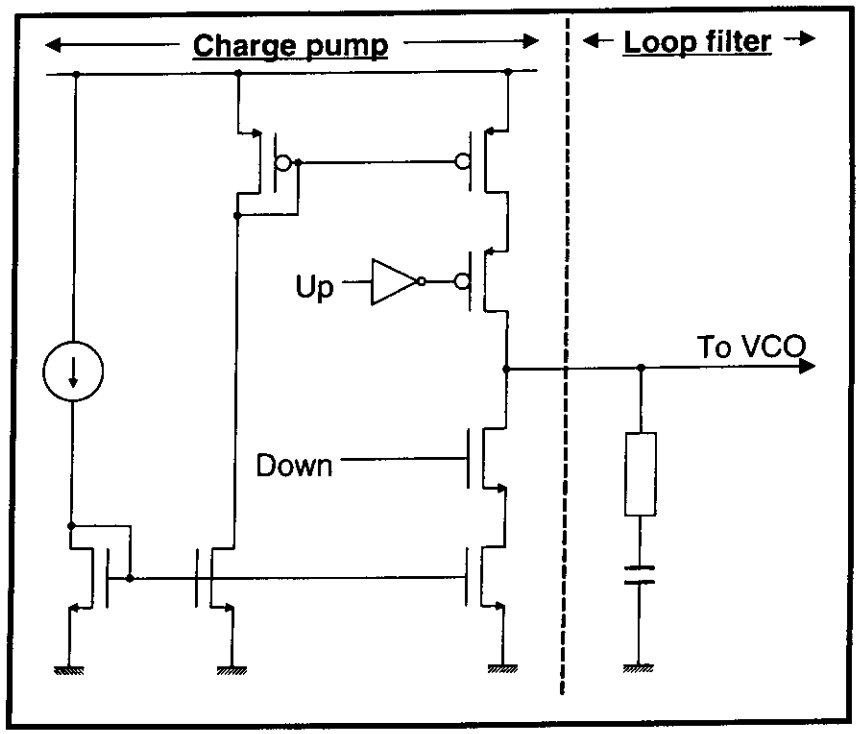
Phase locked loops



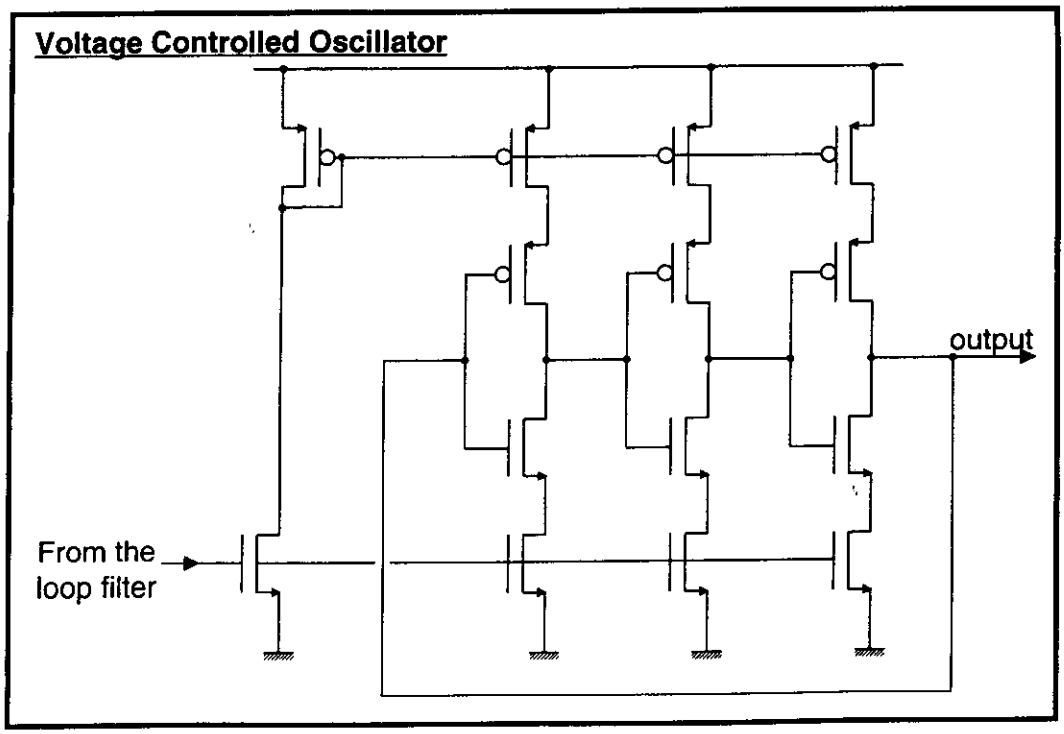
Phase locked loops



Phase locked loops



Phase locked loops



Outline

- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS regular structures

- Memory classification
- Write/read cycle
- Memory architecture
- Read-only memories
- Nonvolatile read-write memories
- Read-write memories
- Sense amplifiers

Memory classification

- Memory: logic element where data can be stored to be retrieved at a later time
- Read-Only Memory (ROM)
  - The information is encoded in the circuit topology
  - The data cannot be modified: it can only be read
  - ROM's are not volatile. That is, removing the power source does not erase the information contents of the memory.

Memory classification

- Read Write Memories (RWM)
  - RWM's allow both reading and writing operations
  - RWM can be of two general types:
    - Static: the data is stored in flip-flops
    - Dynamic: the data is stored as charge in a capacitor
  - Both types of memories are volatile, that is, data is lost once the power is turned off
  - Dynamic memories require periodic "refresh" of its contents in order to compensate for the charge loss caused by leakage currents in the memory element

Memory classification

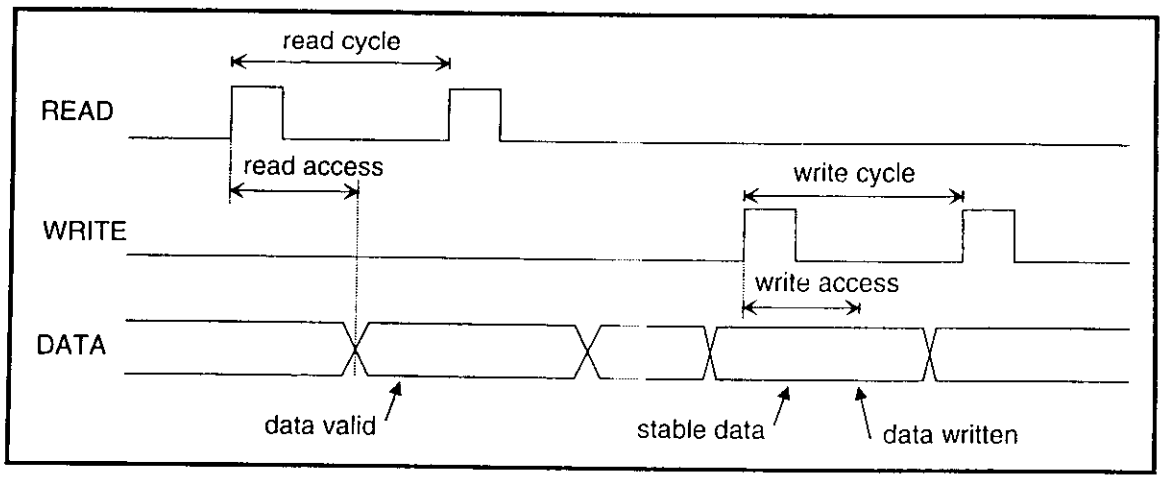
- Nonvolatile Read-Write Memories (NVRWM)
  - These are non volatile memories that allow write operations
  - However:
    - The write operation takes substantially more time than the read operation
    - For some types of NVRWM's, the write operation requires special lab equipment
  - Examples of such memories are:
    - EPROM (Erasable Programmable Read-Only memory)
    - E^2PROM (Electrically Erasable Programmable Read-Only Memory)

Memory classification

- Memories can also be classified according to the way they allow access to the stored data:
  - Random Access: memory locations can be read or written in a random order
  - First-In First-Out (FIFO): The first word to be written is the first word to be read
  - Last-In First-Out (LIFO): The last word to be written is the first word to be read (stack)
  - Shift Register: information is streamed in and out. It can work either as a FIFO or as a LIFO

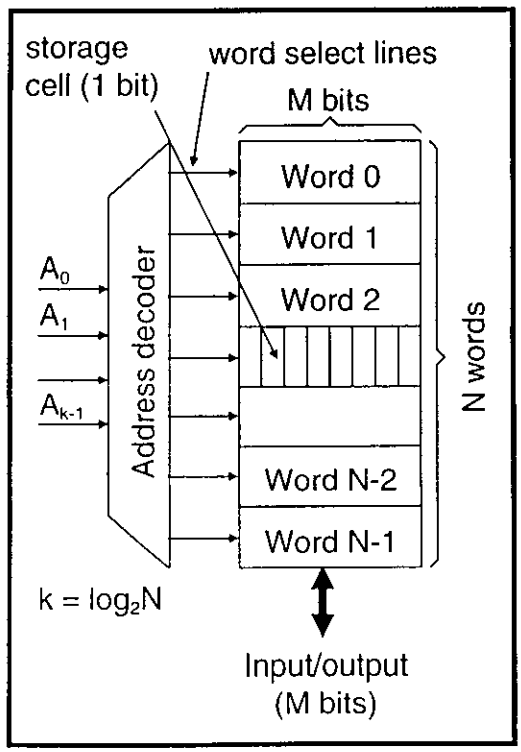
Write/read cycle

- Read-access time: delay between read request and data valid
- Write-access time: delay between write request and the actual writing
- Read or write cycle time: minimum time required between successive read or write operations



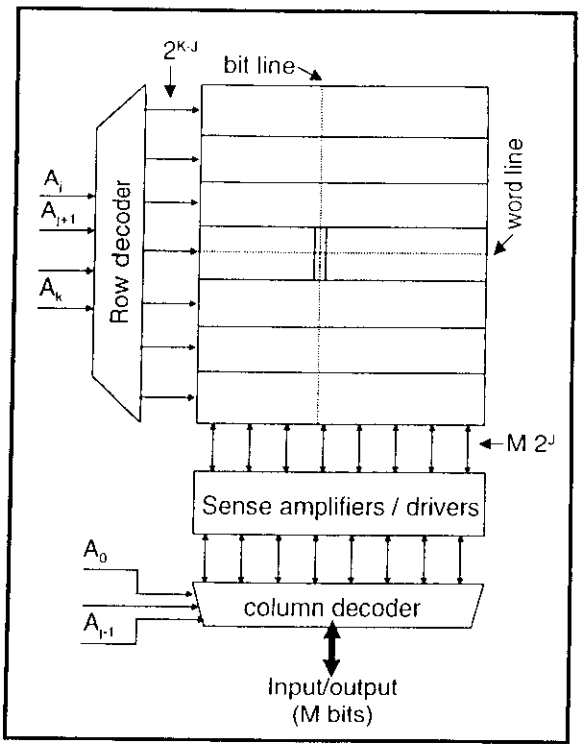
Memory architecture

- The memory is organized in N words, each of M bits wide
- One word at a time is selected for read/write using a select signal
- A decoder is used to convert a binary encoded address into a single active word select line
- This structure is not practical, it results in very big aspect ratios



Memory architecture

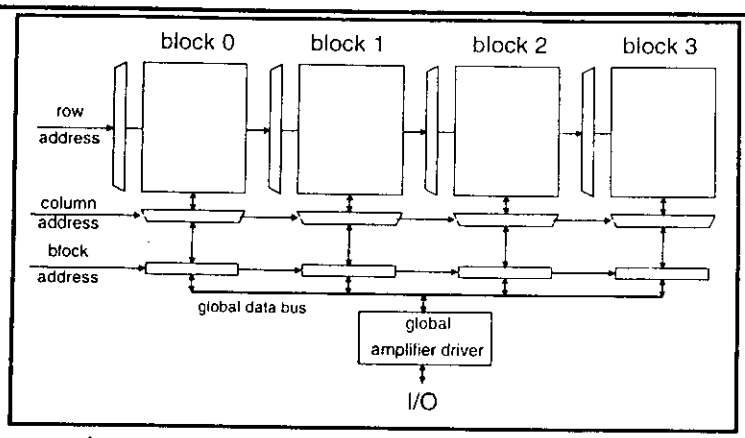
- Memories are organized to be almost square in layout:
  - Multiple words are stored in the same row and selected simultaneously
  - The correct word is then selected by the column decoder
  - The word address is split in two fields:
    - row address: enables one row for R/W
    - column address: selects a word within a row
  - Even this structure is impractical for memories bigger than 256Kbits



Memory architecture

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

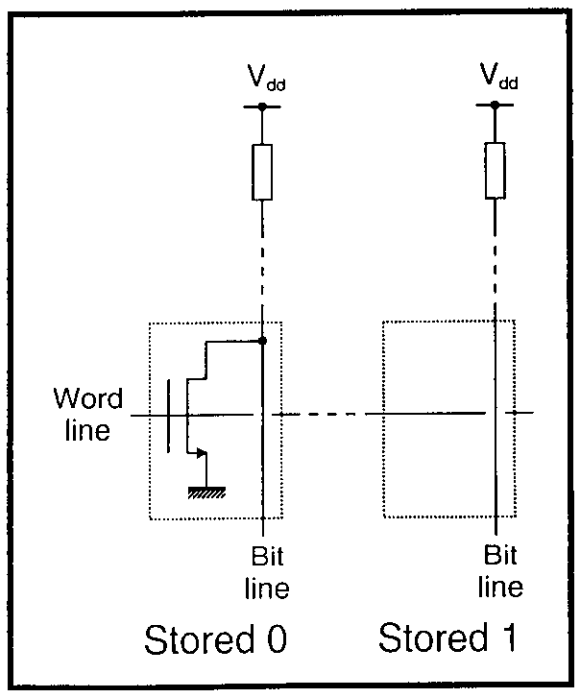
Memory architecture



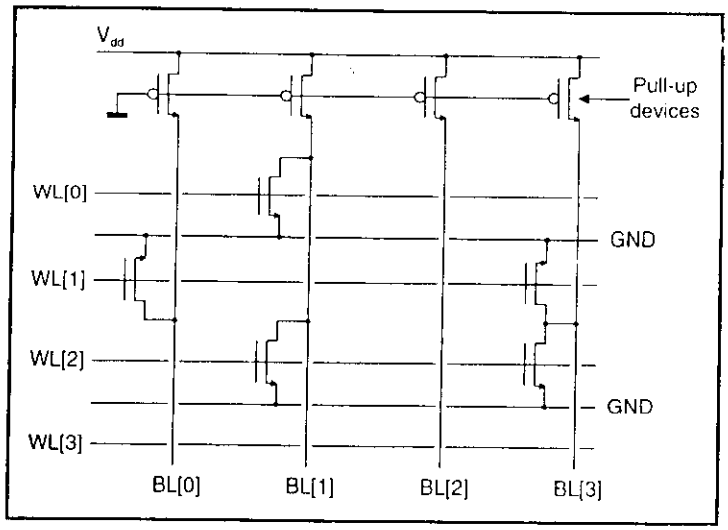
- Large memories start to suffer from speed degradation due to wire resistance and capacitive loading of the bit and word lines
- The solution is to split the memory into "small" memory blocks
- That allows to:
  - use small local word and bit lines ⇒ faster access time
  - power down sense amplifiers and disable decoders of non-active memory blocks ⇒ power saving

Read-only memories

- 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
- The NMOS isolates the bit from the word line



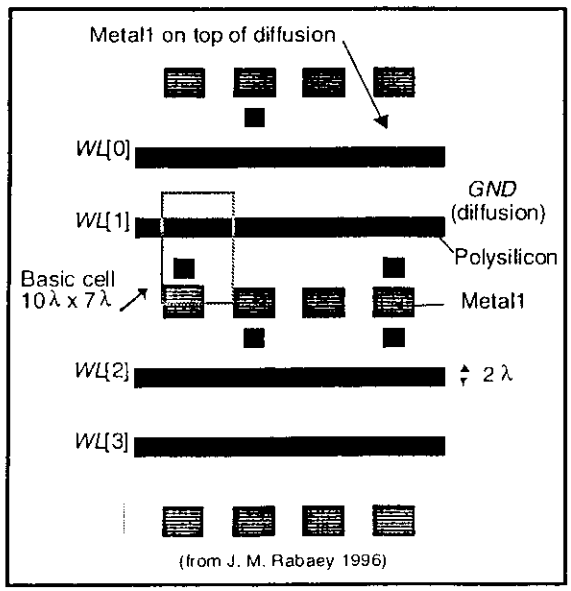
Read-only memories



- A ground contact has to be provided for every cell
  - a ground rail has to be routed through the cell
  - the area penalty can be shared between two neighbor cells:
    - the odd rows are mirrored around the horizontal axis

Read-only memories

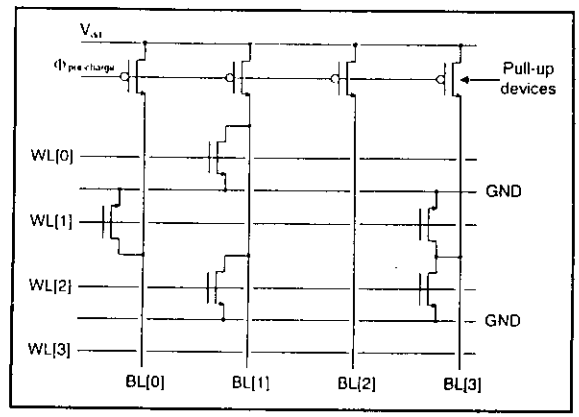
- Use close to minimum size pull-down devices to:
  - make the cell size small
  - reduce the bit line capacitance
- R(pull-up) > R(pull-down) to:
  - ensure adequate low level
- Since for large memories the bit line capacitance can be of the order of pF's, low to high transitions will be slow
- A wider pull-up device can be used resulting in a higher V_{OL}
  - this reduces the noise margin but speeds the low-to-high transition
  - to interface with external logic, a sense amplifier is required to restore the logic levels
  - an inverter with adjusted switching threshold can be used as a sense amplifier



- 0 ⇒ metal-to-diffusion contact
- 1 ⇒ no metal-to-diffusion contact
- only the contact mask layer is used to program the memory array

Read-only memories

- Disadvantages:
  - V_{OL} depends on the ratio of the pull-up/pull-down devices
  - A static current path exists when the output is low causing high power dissipation in large memories
- Solution:
  - Use pre-charged logic
  - Eliminates the static dissipation
  - Pull-up devices can be made wider
  - This is the most commonly used structure



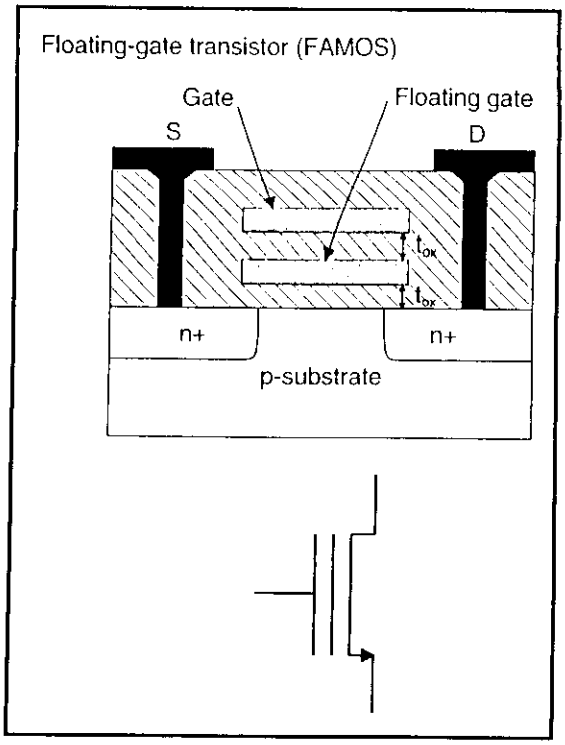
- The bit lines are first pre-charged by the pull-up devices
  - during this phase the word lines must be disabled
- Then, the word lines are activated (word evaluation)
  - during this phase the pull-up devices are off

Nonvolatile read-write memories

- 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)

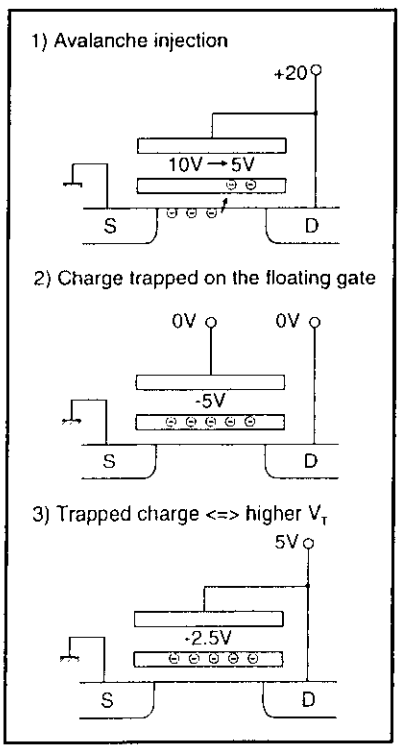
Nonvolatile read-write memories

- 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



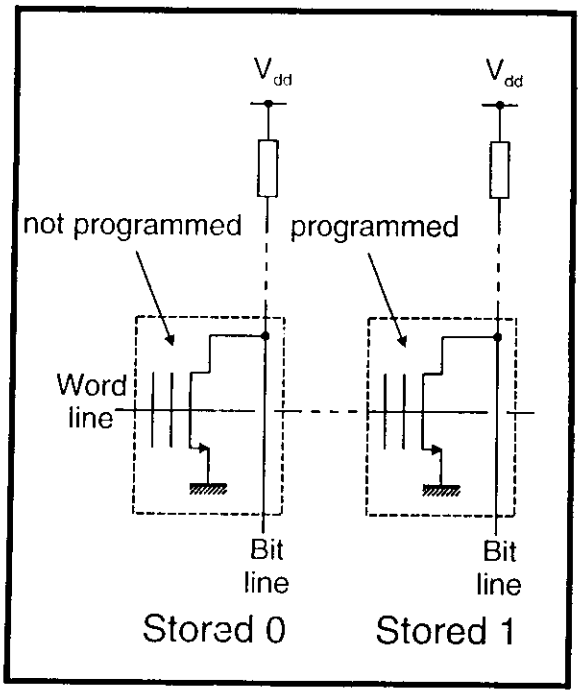
Nonvolatile read-write memories

- 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 read-write memories

- 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

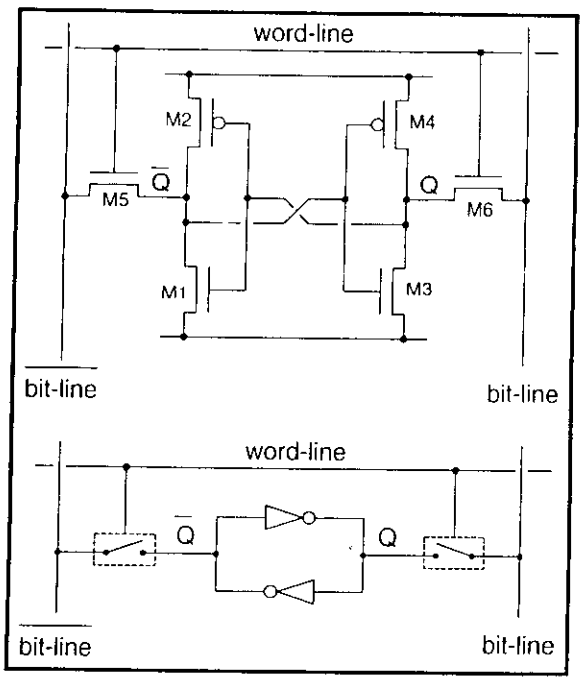


Nonvolatile read-write memories

- 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

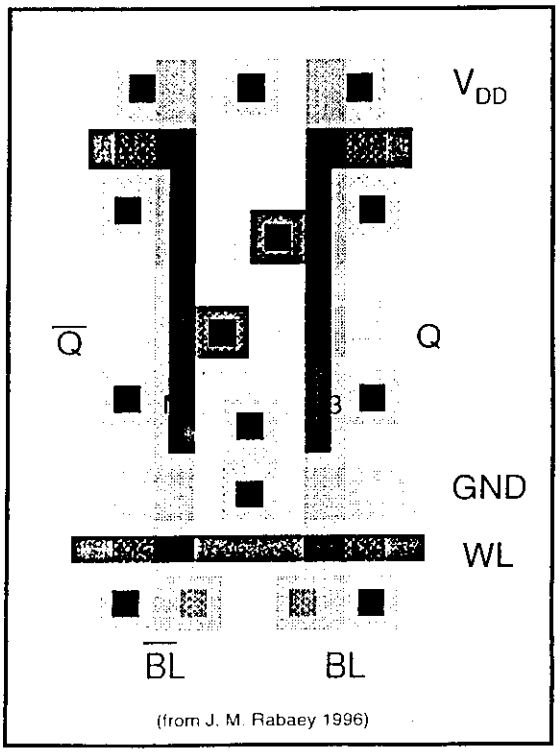
Read-write memories

- Static Read-Write Memories (SRAM):
  - data is stored by positive feedback
  - the memory is volatile
- The cell use 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



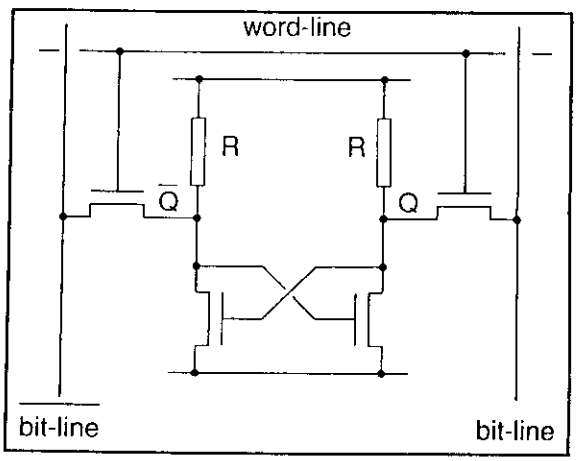
Read-write memories

- SRAM performance:
  - The read operation is the critical one:
    - It involves discharging or charging the large bit-line capacitance through the small transistors of the cell
  - The write time is dominated by the propagation delay of the cross-coupled inverter pair
  - The six-transistor cell is not area efficient:
    - It requires routing of two power lines, two bit lines and a word line
    - Most of the area is taken by wiring and interlayer contacts



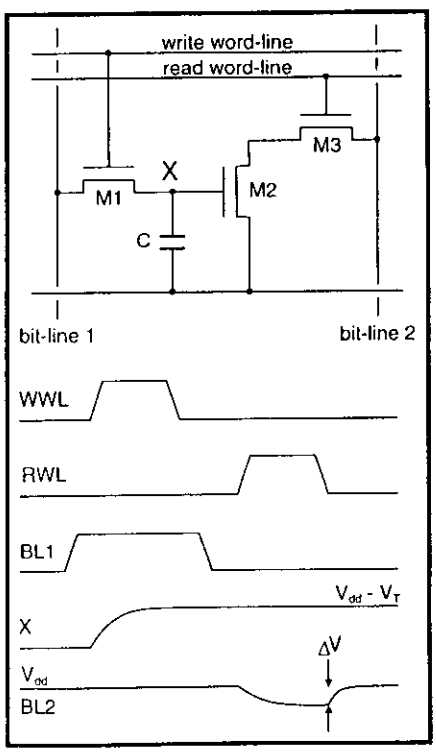
Read-write memories

- 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^{-15A})
    - they must be made as high as possible to minimize static power dissipation
    - undoped polysilicon $10^{12}\Omega/$
  - 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



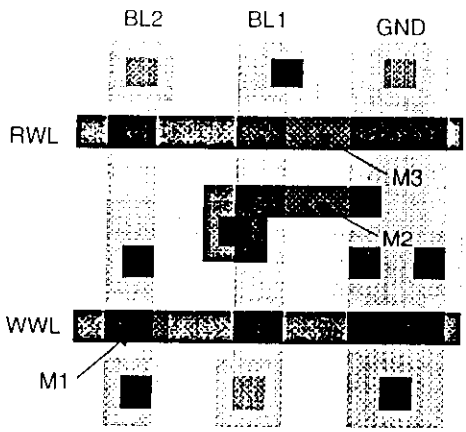
Read-write memories

- 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



Read-write memories

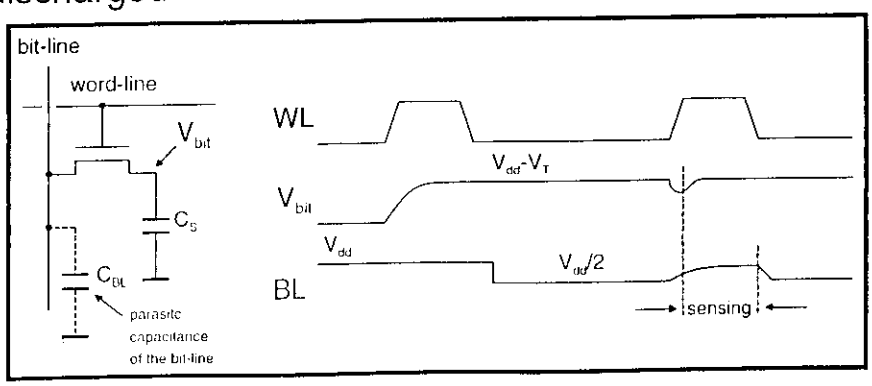
- The cell is inverting
- Due to leakage currents the cell needs to be periodically refreshed (every 1 to 4ms)
- Refresh operation:
  - read the stored data
  - put its complement in BL1
  - enable/disable the WWL
- Compared with an SRAM the area is greatly reduce:
  - SRAM $\Rightarrow$ 1092 λ^2
  -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



(from J. M. Rabaey 1996)

Read-write memories

- 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

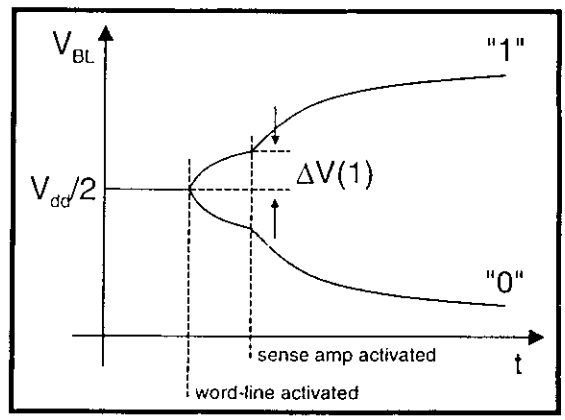


Read-write memories

- 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 $C_S \Rightarrow$ $\Delta V \cong 250mV$

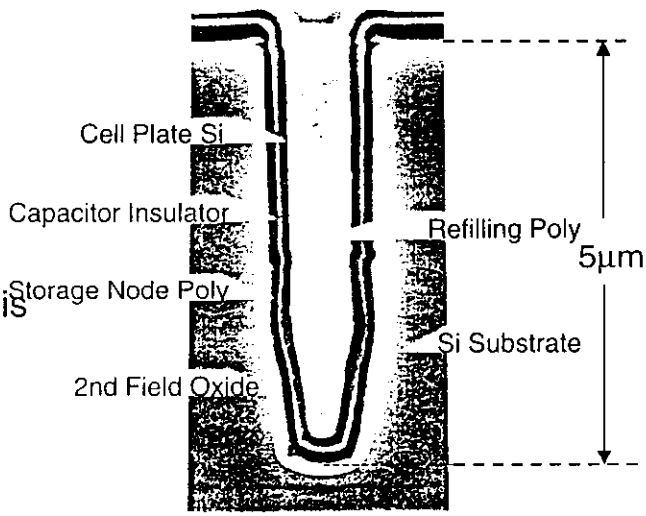


  - The amount of charge stored in the cell is modified during the read operation
  - However, during read, the output of the sense amplifier is imposed on the bit line restoring the stored charge

Read-write memories

- 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

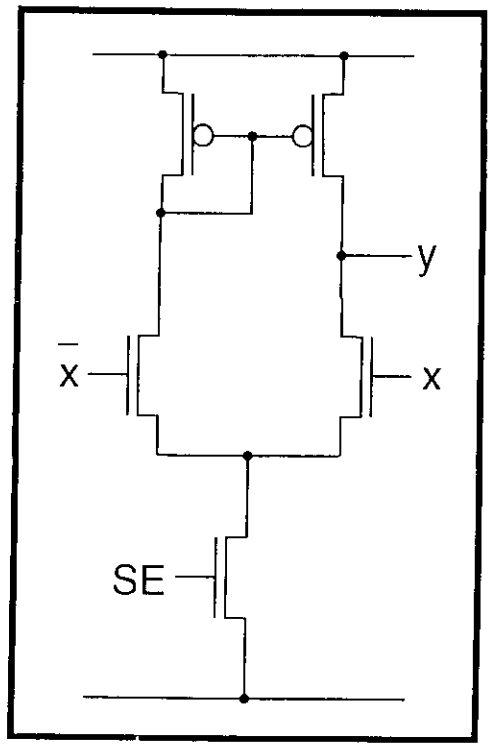
Trench capacitor



(from T. Mano et al., 1987)

Sense amplifiers

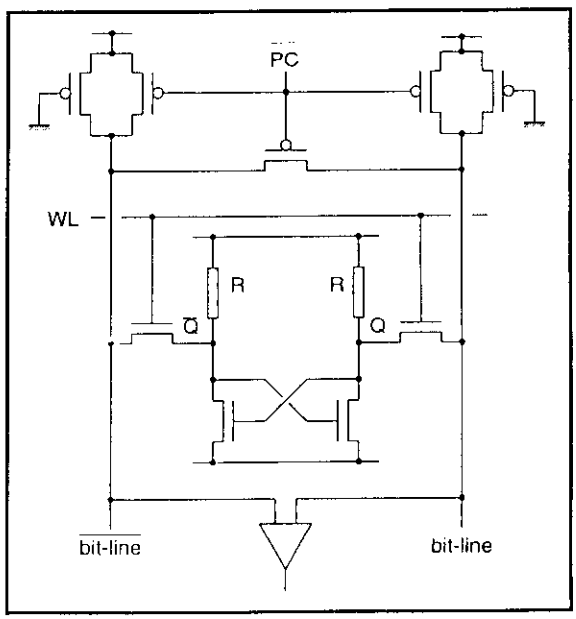
- Sense amplifiers improve the speed performance of the memory cell:
  - they compensate for the low driving capability of the cells
- Contribute to power reduction by allowing to use low signal swings on the heavily capacitive bit-lines
- They perform signal restoration in the refresh and read cycles of 1T dynamic memories
- They can be differential or single ended



Sense amplifiers

SRAM read cycle:

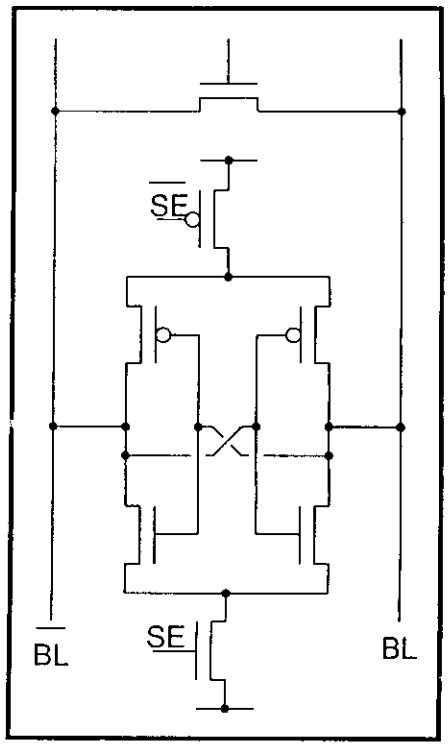
- pre-charge:
  - pre-charge the bit-lines to V_{dd} and make their voltages equal
- Reading:
  - disable the pre-charge devices
  - enable the word lines
  - once a minimum (≅0.5V) signal is built up in the bit-lines the sense amplifier is turned on
- The grounded PMOS loads limit the signal swing and facilitate the next pre-charge



Sense amplifiers

A cross-coupled inverter pair can be used as a sense amplifier

- To act as a sense amplifier:
  - The bit-lines are equalized: this initializes the flip-flop in its metastable point
  - A voltage is built over the bit lines by the selected cell
  - The sense amplifier is activated once the voltage is large enough
  - The cross-coupled pair then toggles to one of its stable operating points
  - The transition is fast due to positive feedback
- Ideal for an 1T DRAM: inputs and outputs are merged



Sense amplifiers

- The memory array is divided in two: the sense amplifier in the middle
- On each side "dummy" cells are added
- These cells serve as a reference during the reading
- EQ is asserted and both halves pre-charged to $V_{dd}/2$
- The dummy cells are also pre-charged to $V_{dd}/2$
- If a cell in one of the halves of the bit line is selected, the dummy cell on the other half is used as a reference for the sense amplifier