united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

The United Nations
University

SMR.1073-3

ICTP-UNU-Microprocessor Laboratory
Fifth Course on Basic VLSI Design Techniques

9 November - 4 December 1998

INTRODUCTION TO VLI ASIC DESIGN AND TECHNOLOGY

Paulo Rodrigues S. MOREIRA
CERN
EP Division
Building 14-6-014, F18300
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"

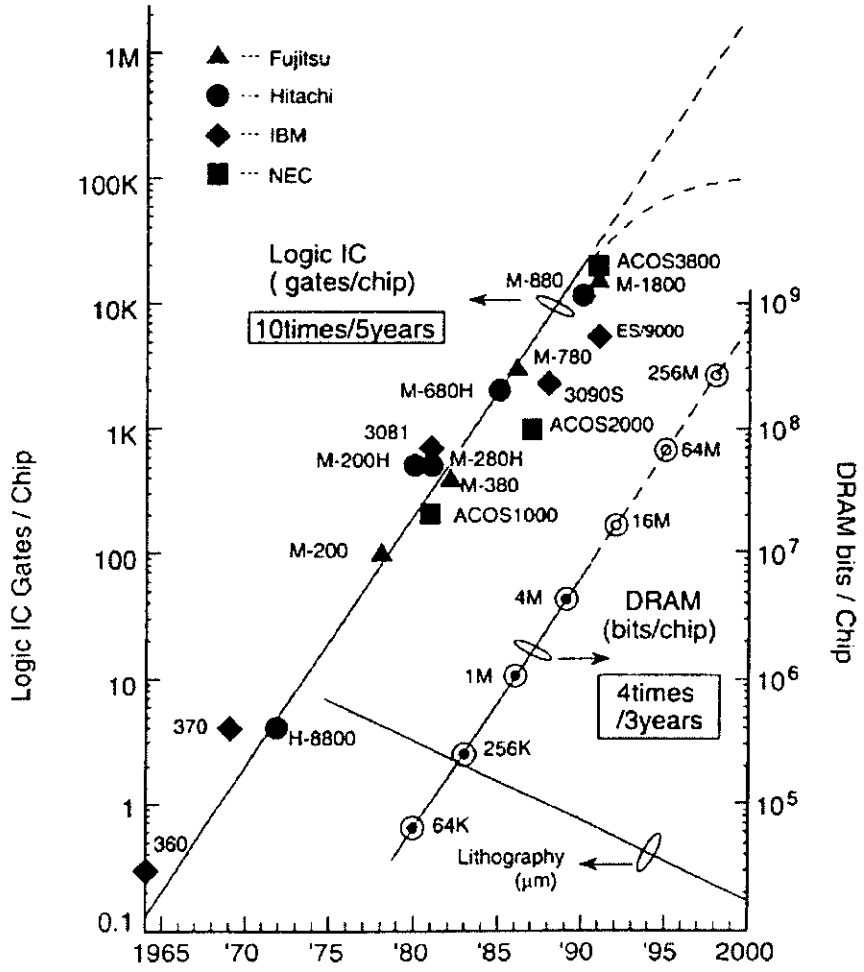
- Amplification of very week signals
- A/D and D/A conversion
- Very high frequency amplification
- Very high frequency signal processing
- As digital systems become faster and faster and circuits densities increase:
  - Analogue phenomena are "creeping" in digital systems

"Moore's Law"

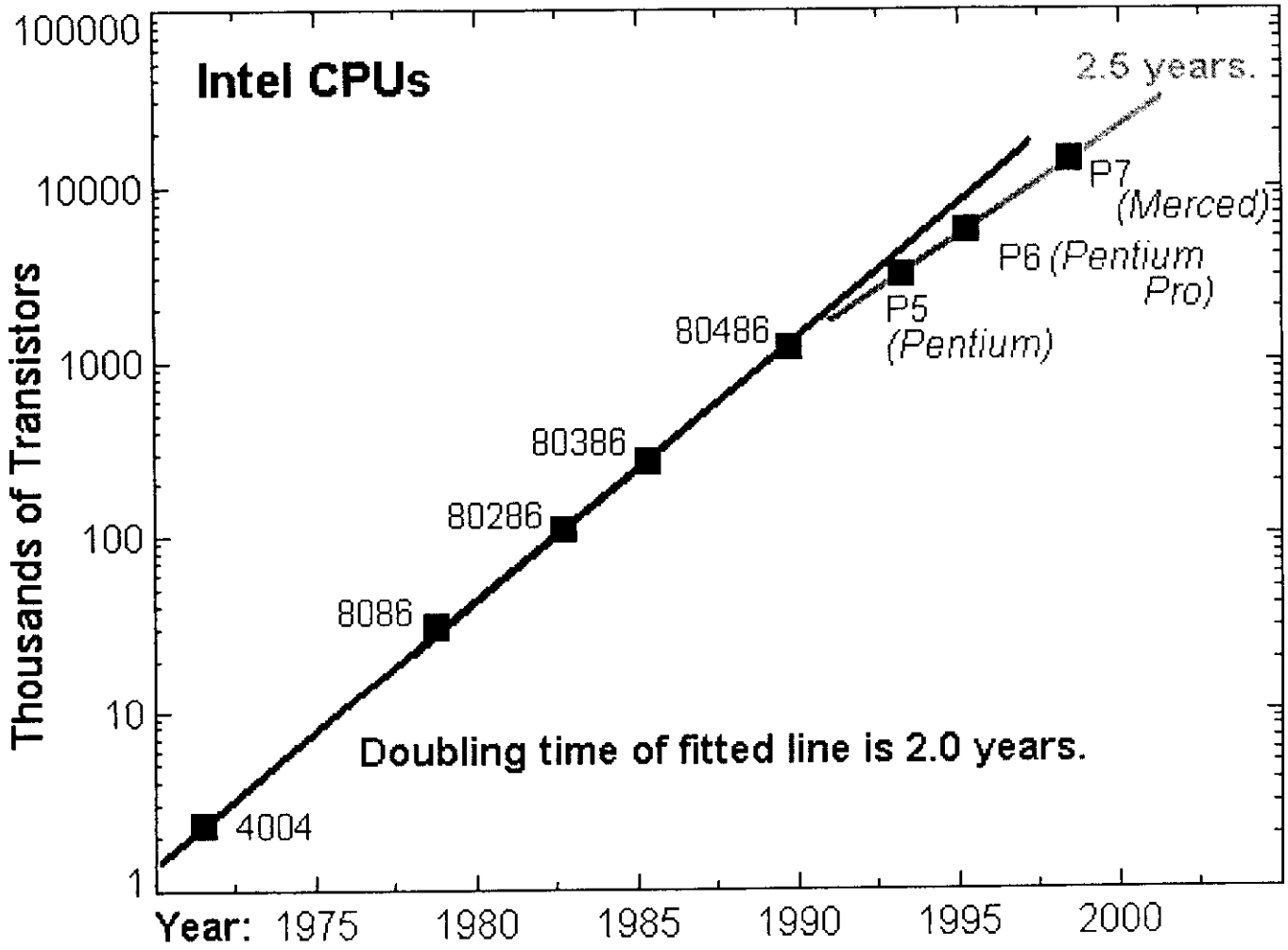
The number of transistors that can be integrated on a single IC grows exponentially with time

(*integration complexity doubles every three years*)

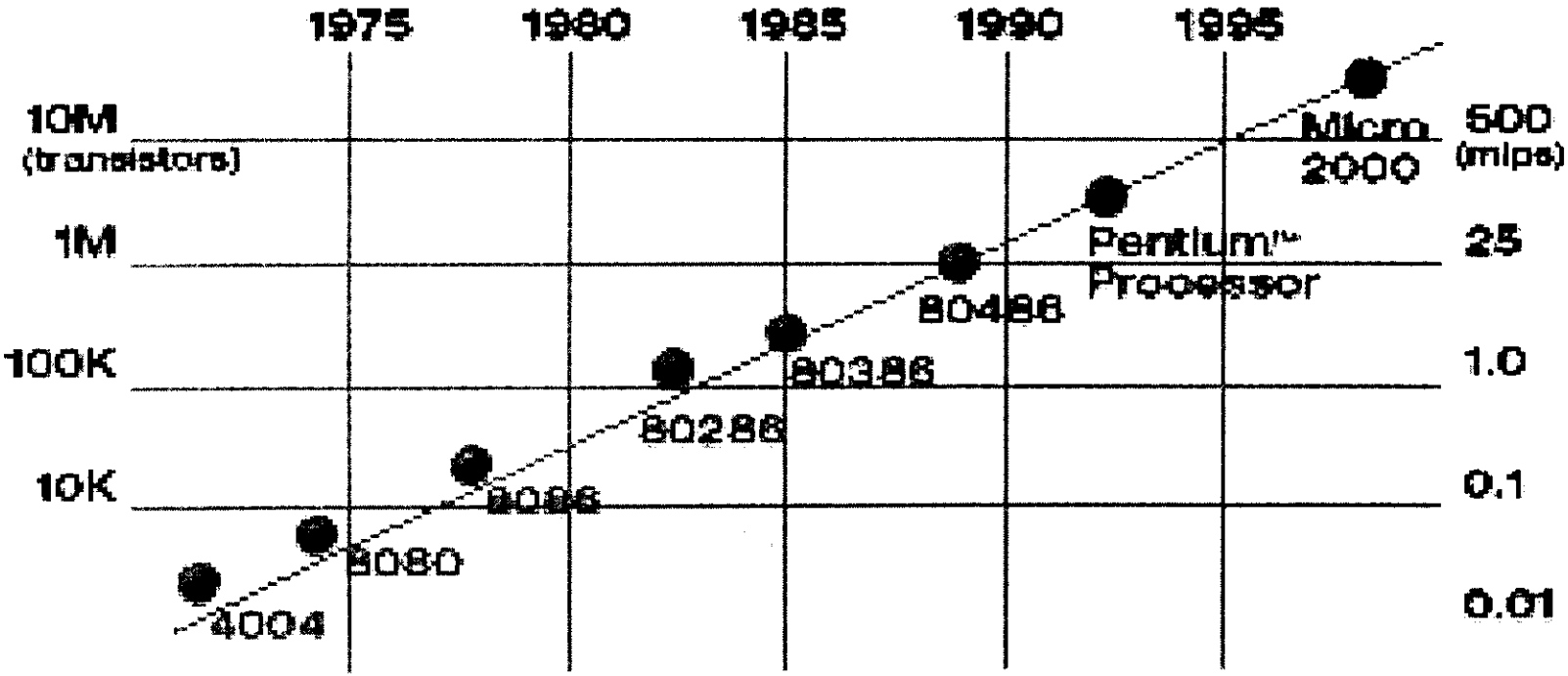
Trends in IC complexity



Trends in transistor count



Trends in clock frequency



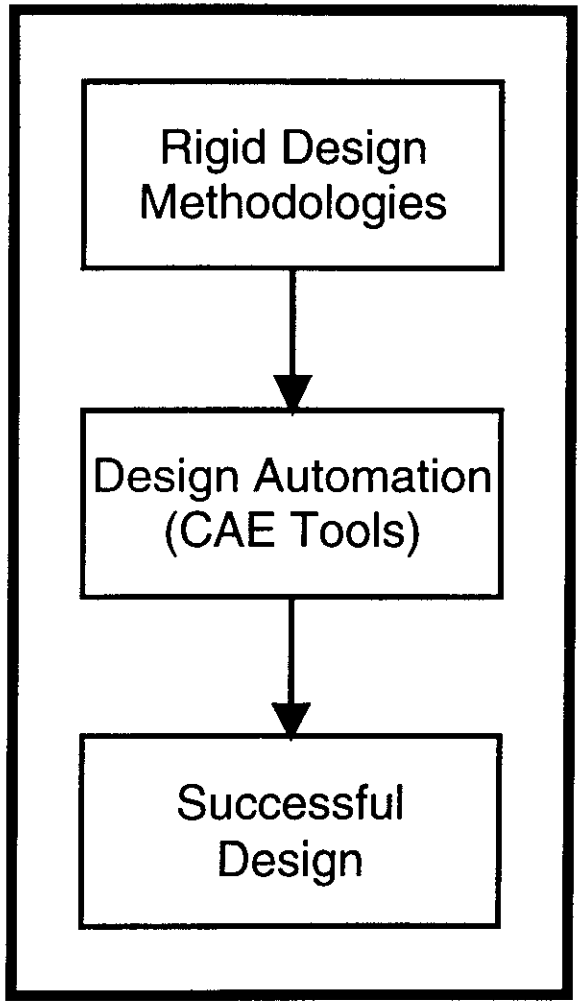
2001 and beyond ?

1997 Semiconductor Industry Association (SIA) Road Map

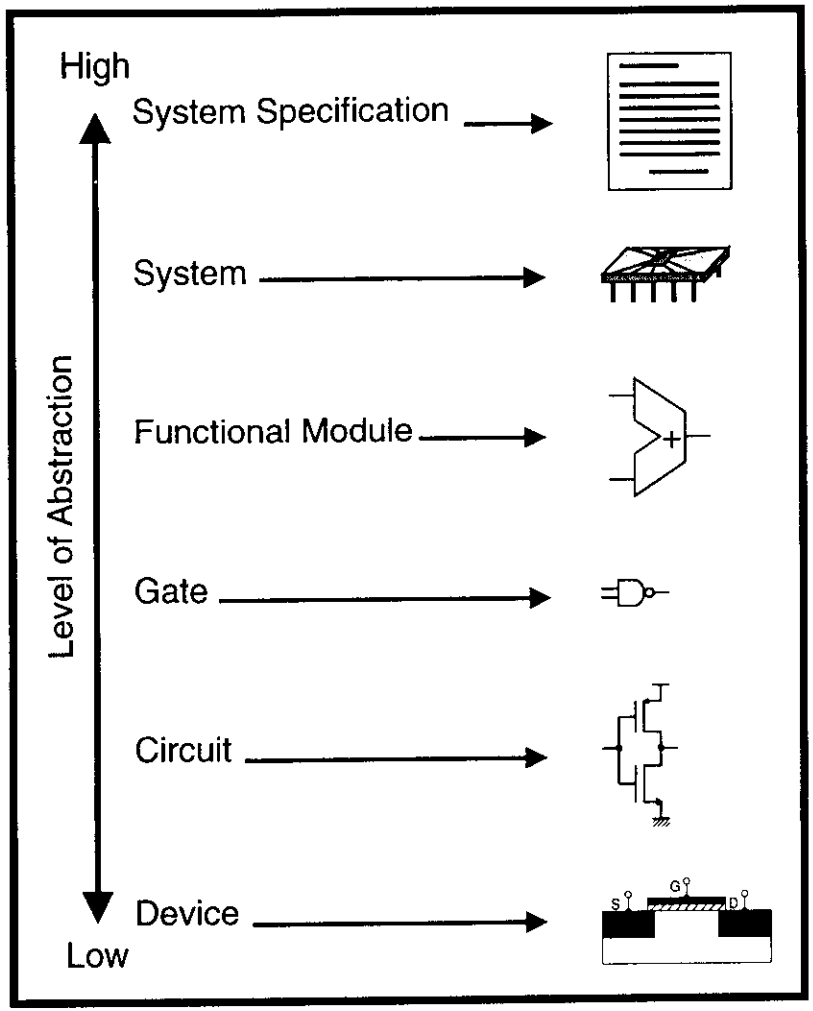
	1997	***2001***	***2012***
Technology (nm)	250	150	50
Minimum mask count	22	23	28
Wafer diameter (mm)	200	300	450
Memory	256M	1G	256G
Transistors/cm^2	3.7M	10M	180M
Wiring levels (maximum)	*6*	*7*	*9*
Clock, local (MHz)	750	1500	10000
Chip size: DRAM (mm^2)	280	445	1580
Chip size: µP *(mm^2)*	300	385	750
Power supply (V)	1.8-2.5	1.2-1.5	0.5-0.6
Maximum Power (W)	*70*	*110*	*175*
Number of pins	600	900	2700

How to cope with complexity?

- By applying:
  - Rigid design methodologies
  - Design automation



Design abstraction levels



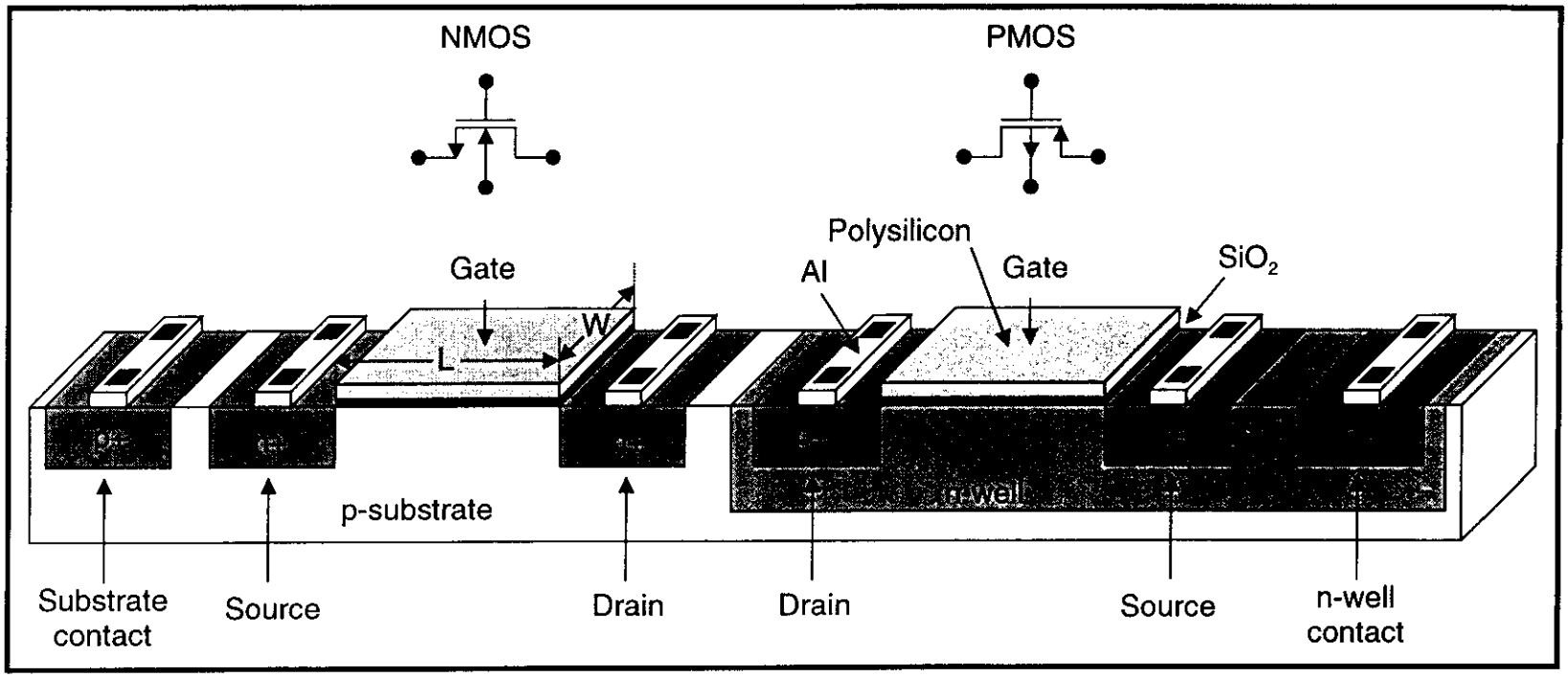
Outline

- 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):

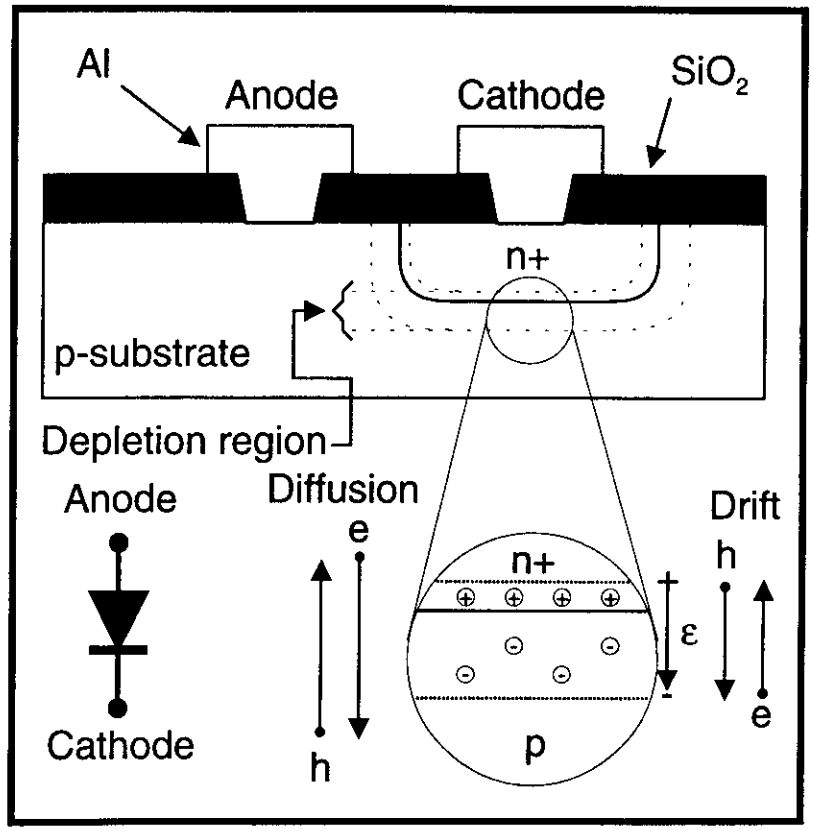
- pn-Junction diodes
- parasitic capacitance

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



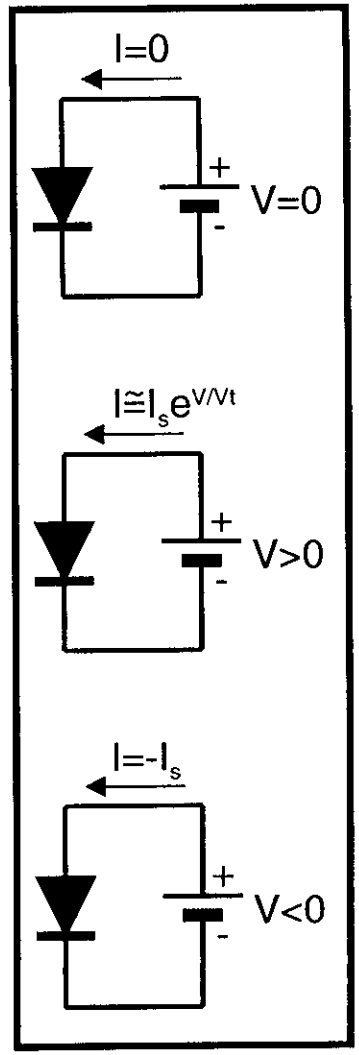
pn-Junction diodes

- 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 $V > \phi_T$ (forward bias)

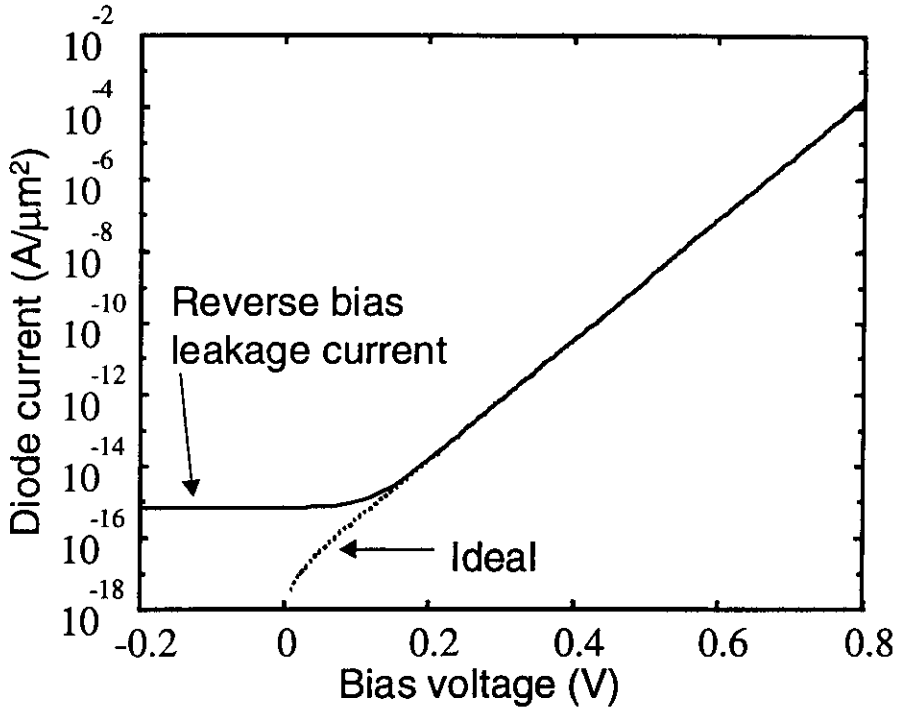
$$I_F \cong I_s \cdot e^{V/\phi_T}$$

- For $V < 0$ (reversed bias)

$$I_R \cong -I_s$$

- In practical diodes due to thermal generation

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



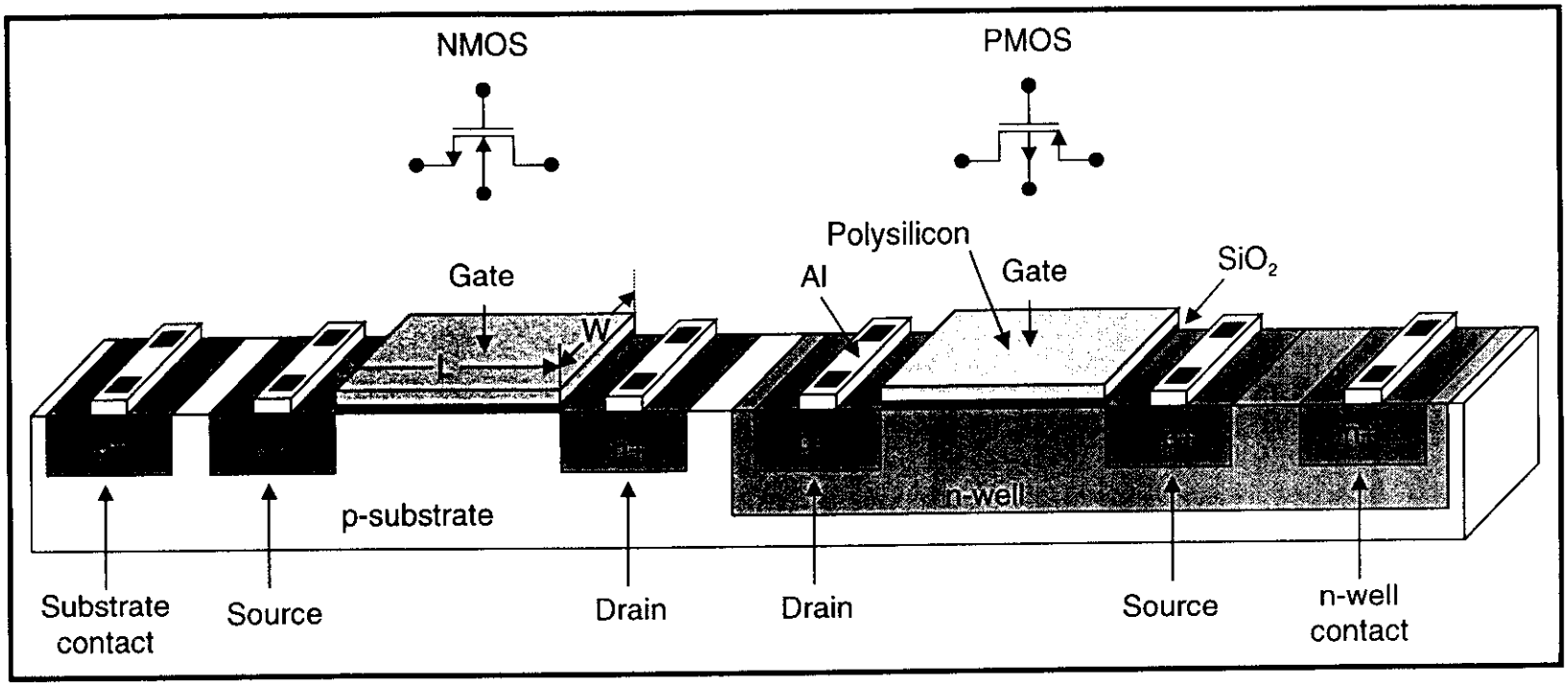
Outline

- 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):

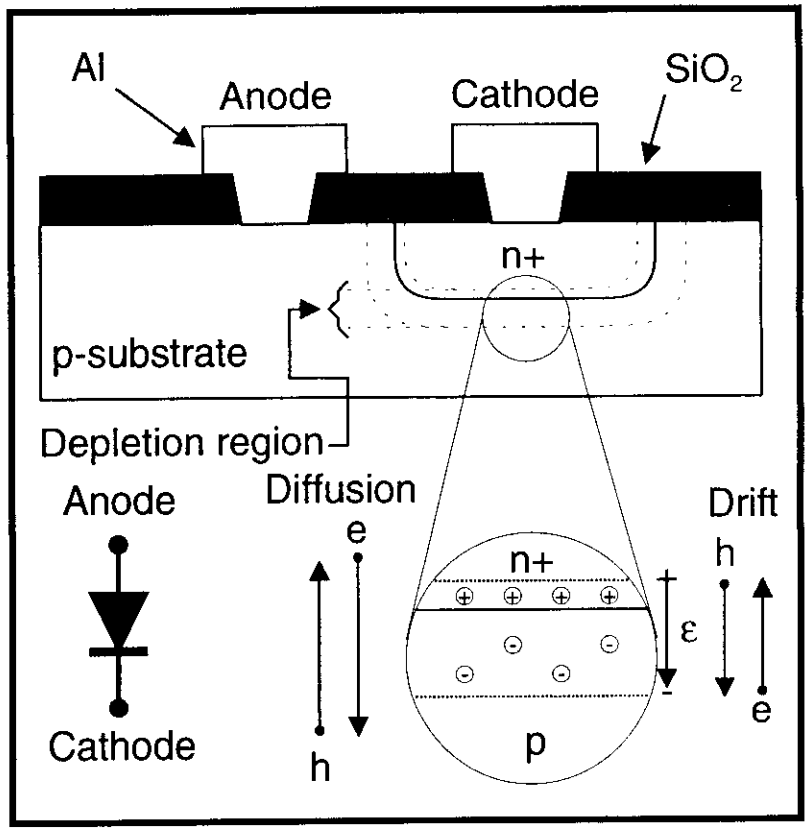
- pn-Junction diodes
- parasitic capacitance

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



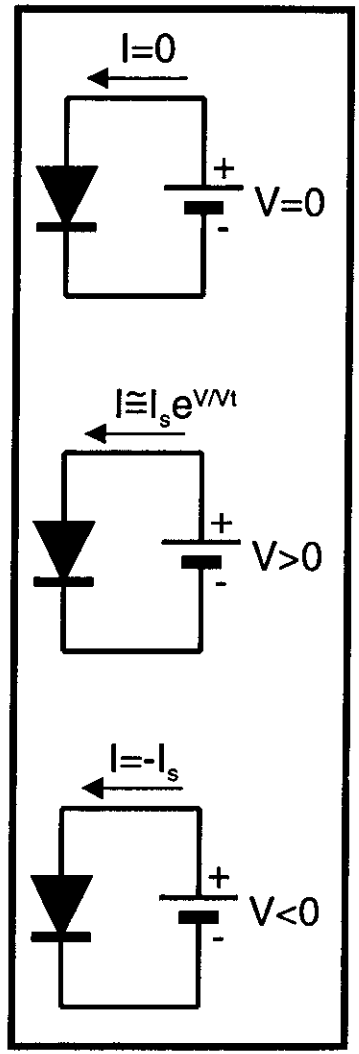
pn-Junction diodes

- 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)

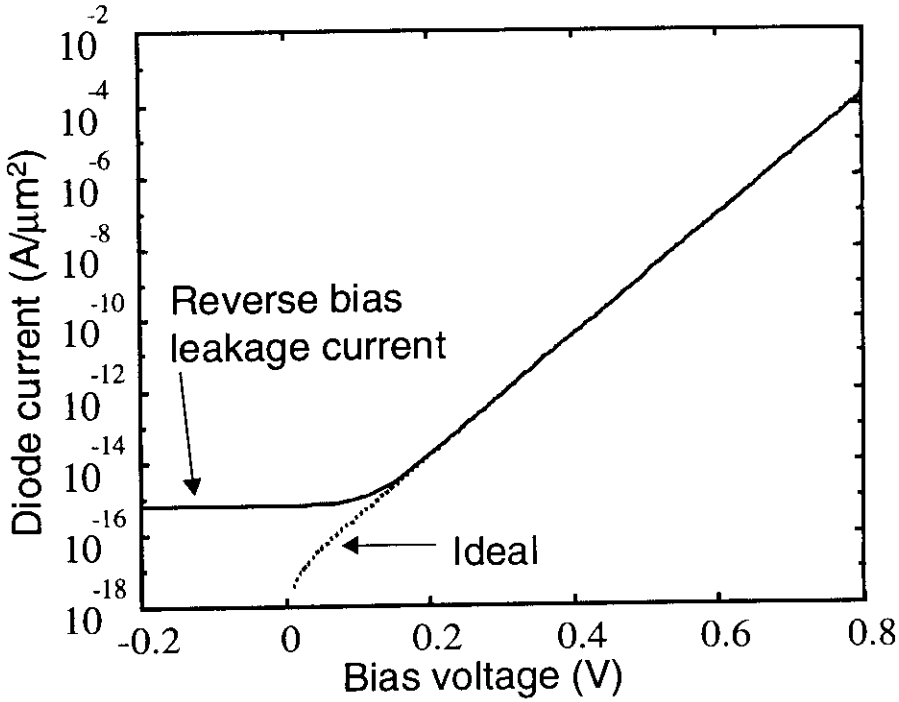
$$I_F \cong I_s \cdot e^{V/\phi_T}$$

- For V<0 (reversed bias)

$$I_R \cong -I_s$$

- In practical diodes due to thermal generation

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



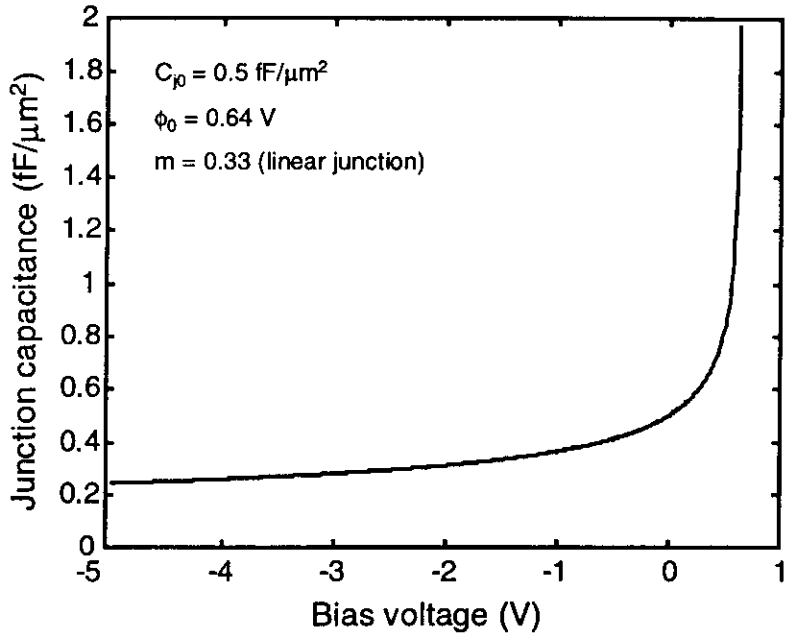
Depletion capacitance

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

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

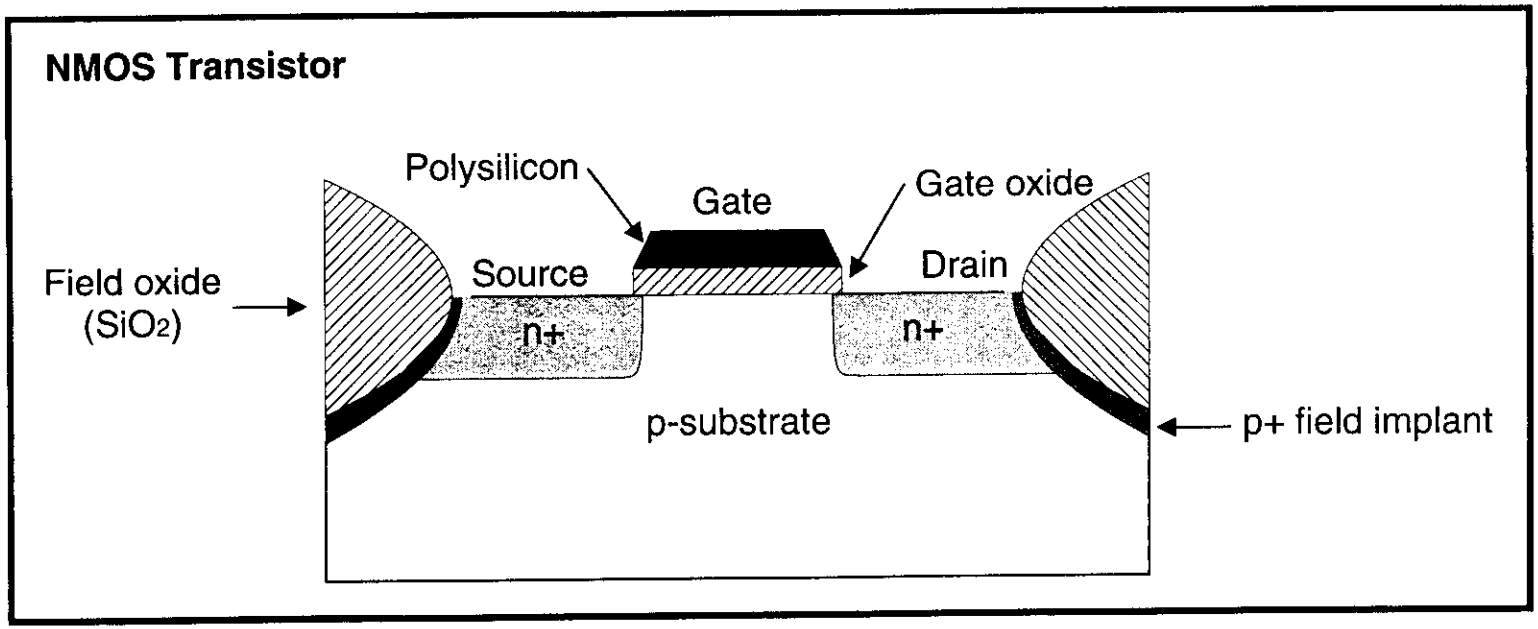
- Simplification: for V<0

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



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

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

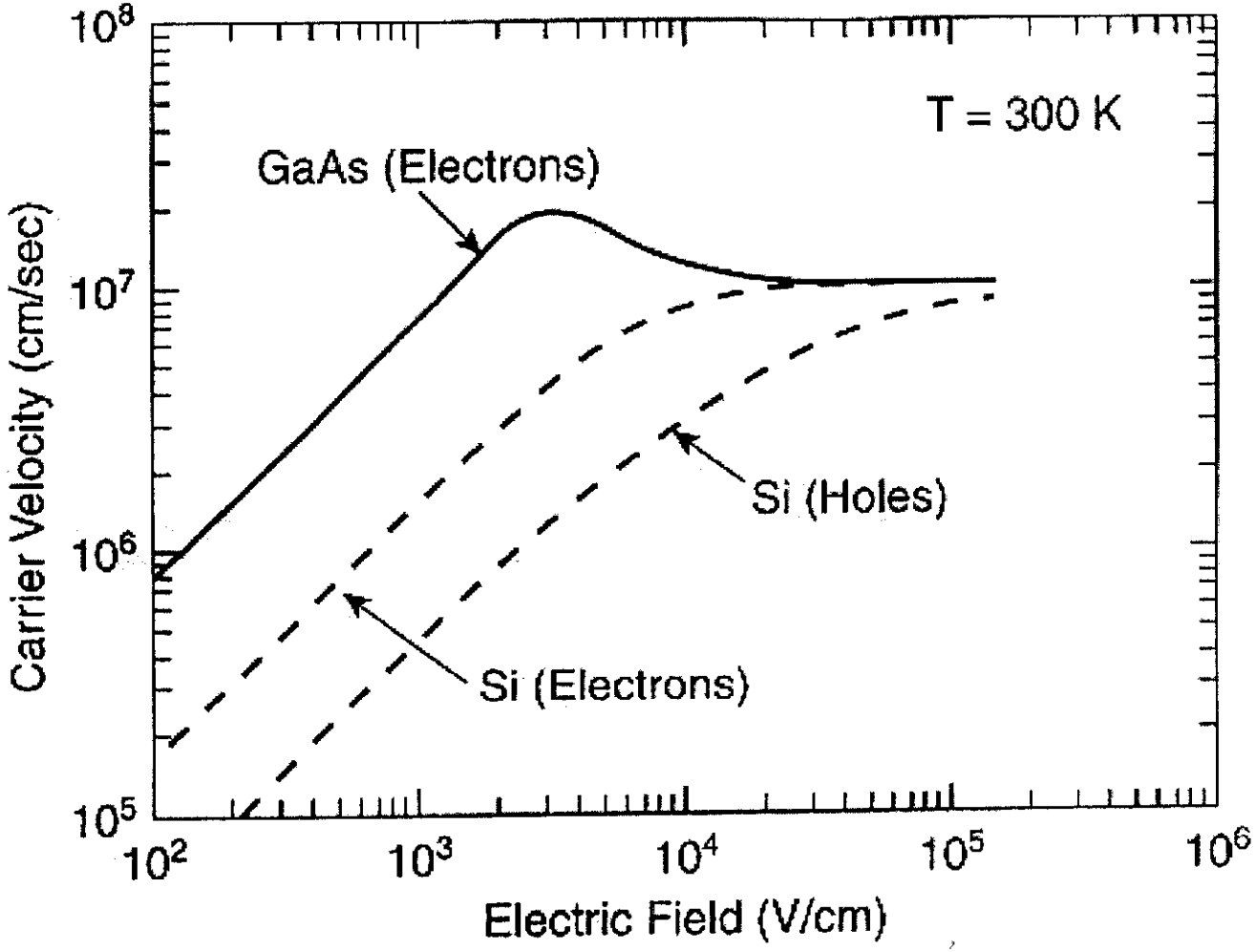
- Oxide capacitance

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

- Process "transconductance"

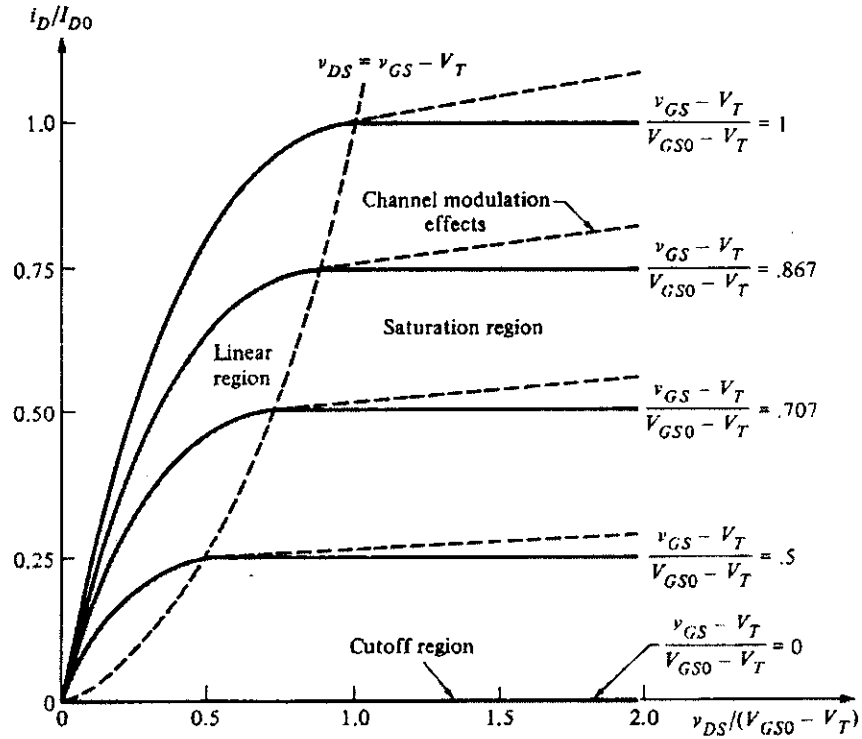
$$\mu \cdot C_{ox} = \frac{\mu \cdot \varepsilon_{ox}}{t_{ox}} \quad \left(\mathrm{A / V^2} \right)$$

Mobility


T = 300 K
GaAs (Electrons)
Si (Holes)
Si (Electrons)
Carrier Velocity (cm/sec)
Electric Field (V/cm)
10^8
10^7
10^6
10^5
10^2
10^3
10^4
10^5
10^6

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

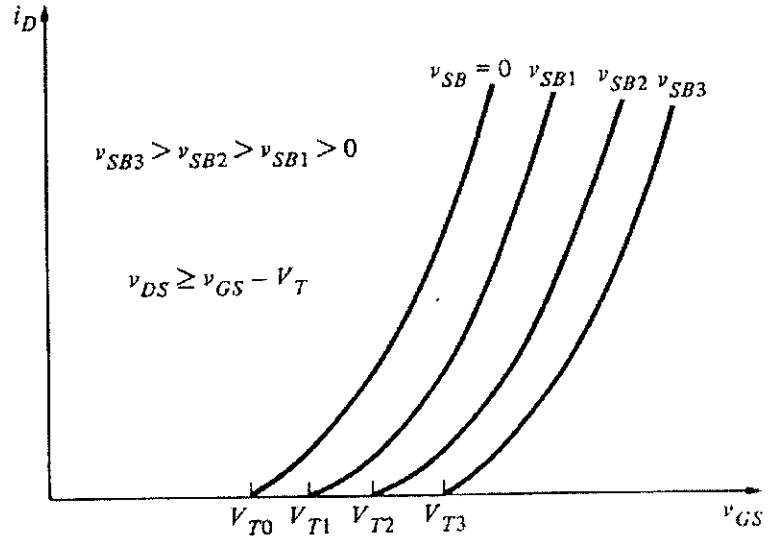
- The threshold depends on the:
  - Doping levels
  - Source-to-bulk voltage
  - Gate oxide thickness

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

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

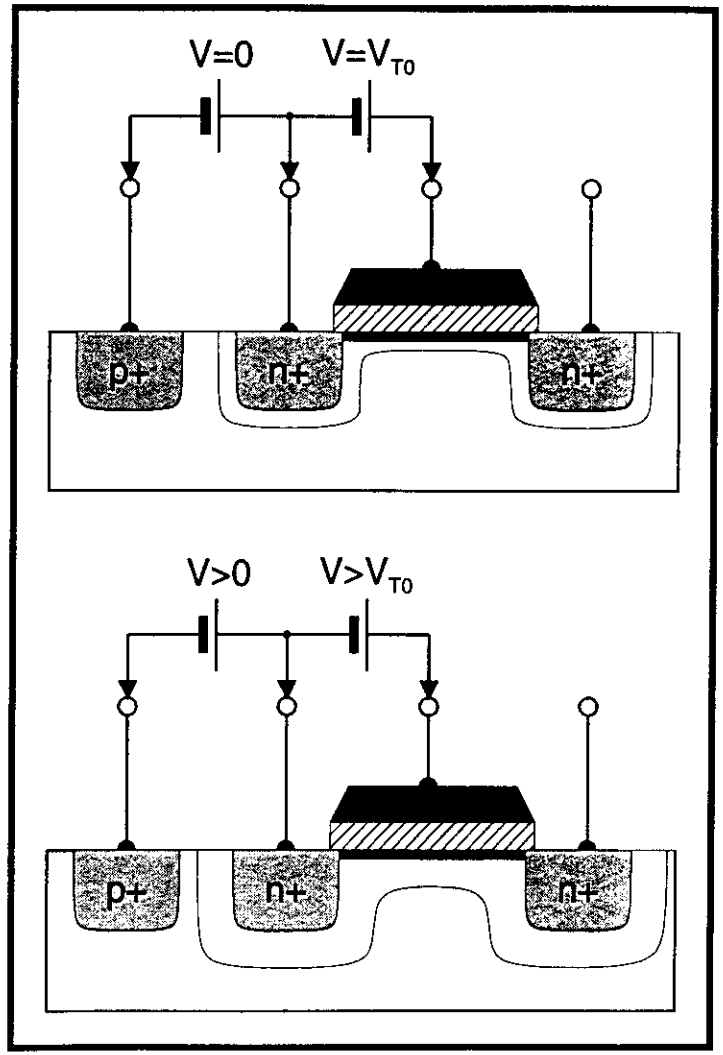
$$\gamma = \frac{\sqrt{2\, q\, \varepsilon_{si}\, N_A}}{C_{ox}}$$

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



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 > $-2\phi_F + V_{sb}$

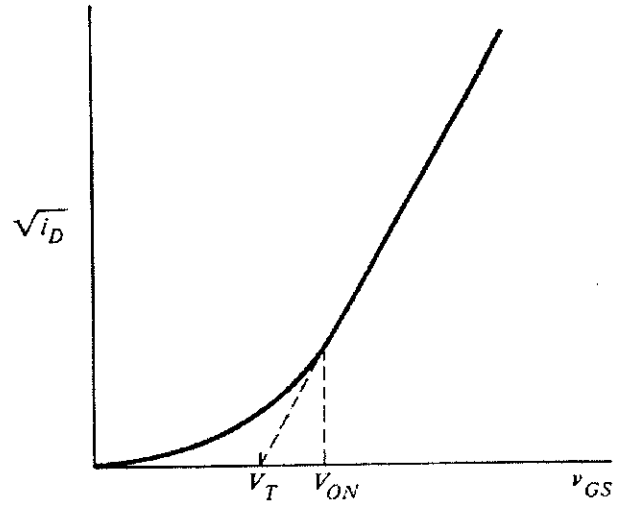


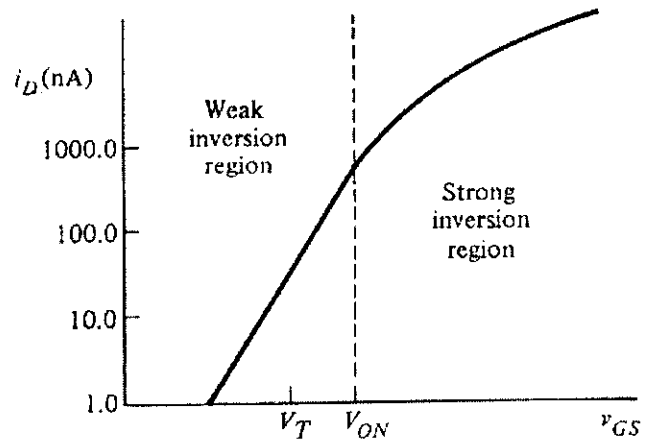
Weak inversion

- 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

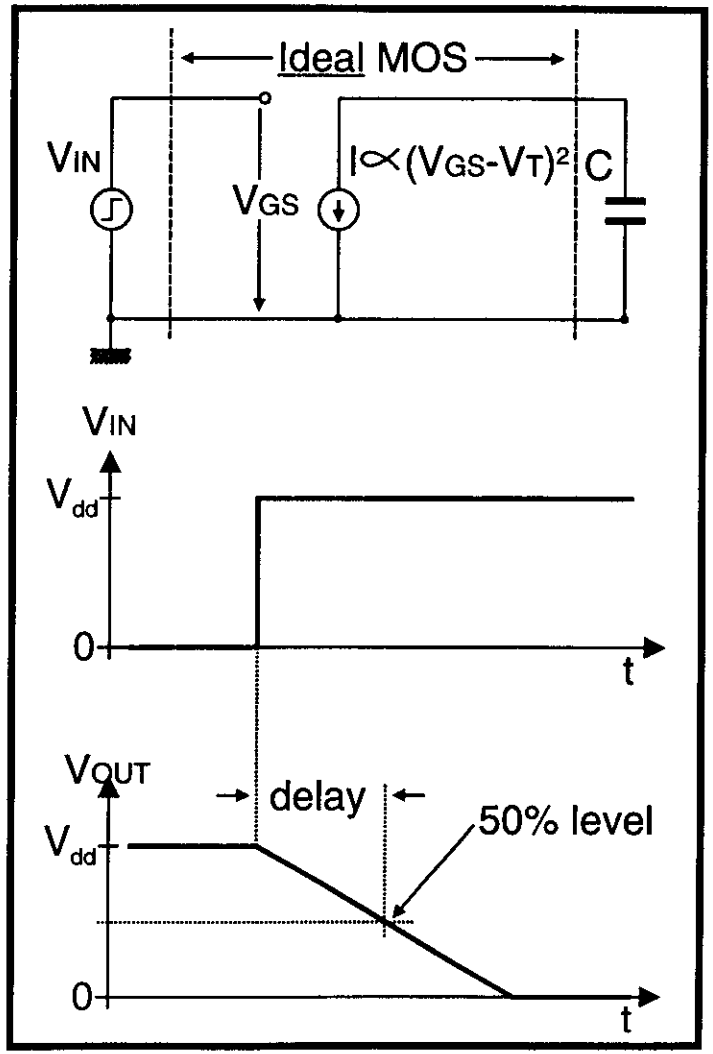




What causes delay?

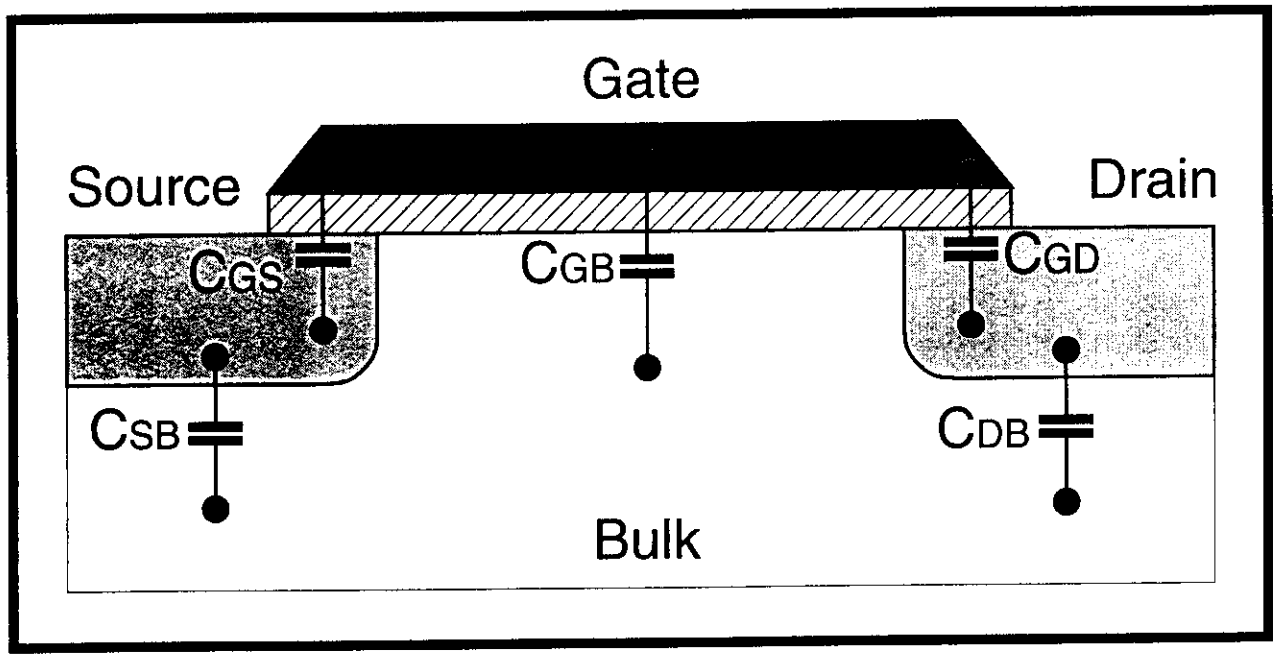
- In MOS circuits capacitive loading is the main cause
- Due to:
  - Device capacitance
  - Interconnect capacitance

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



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

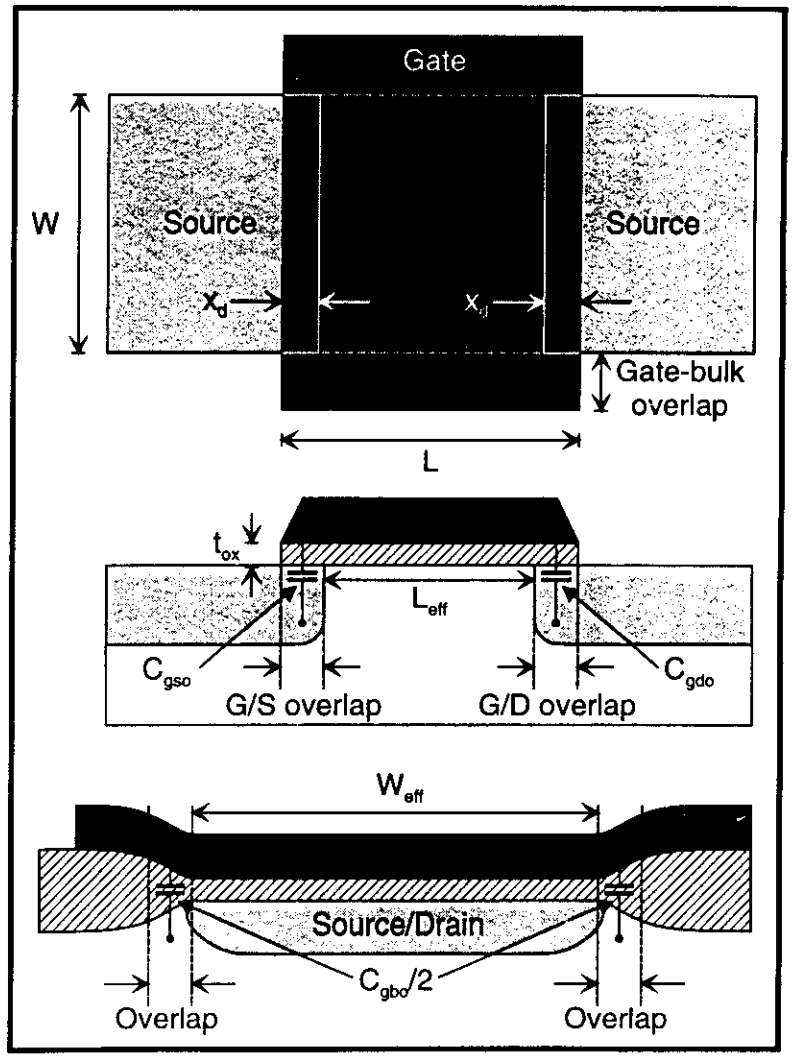
- This gives origin to the source/drain overlap capacitances:

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

$$C_o \ (\mathrm{F/m})$$

- 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:

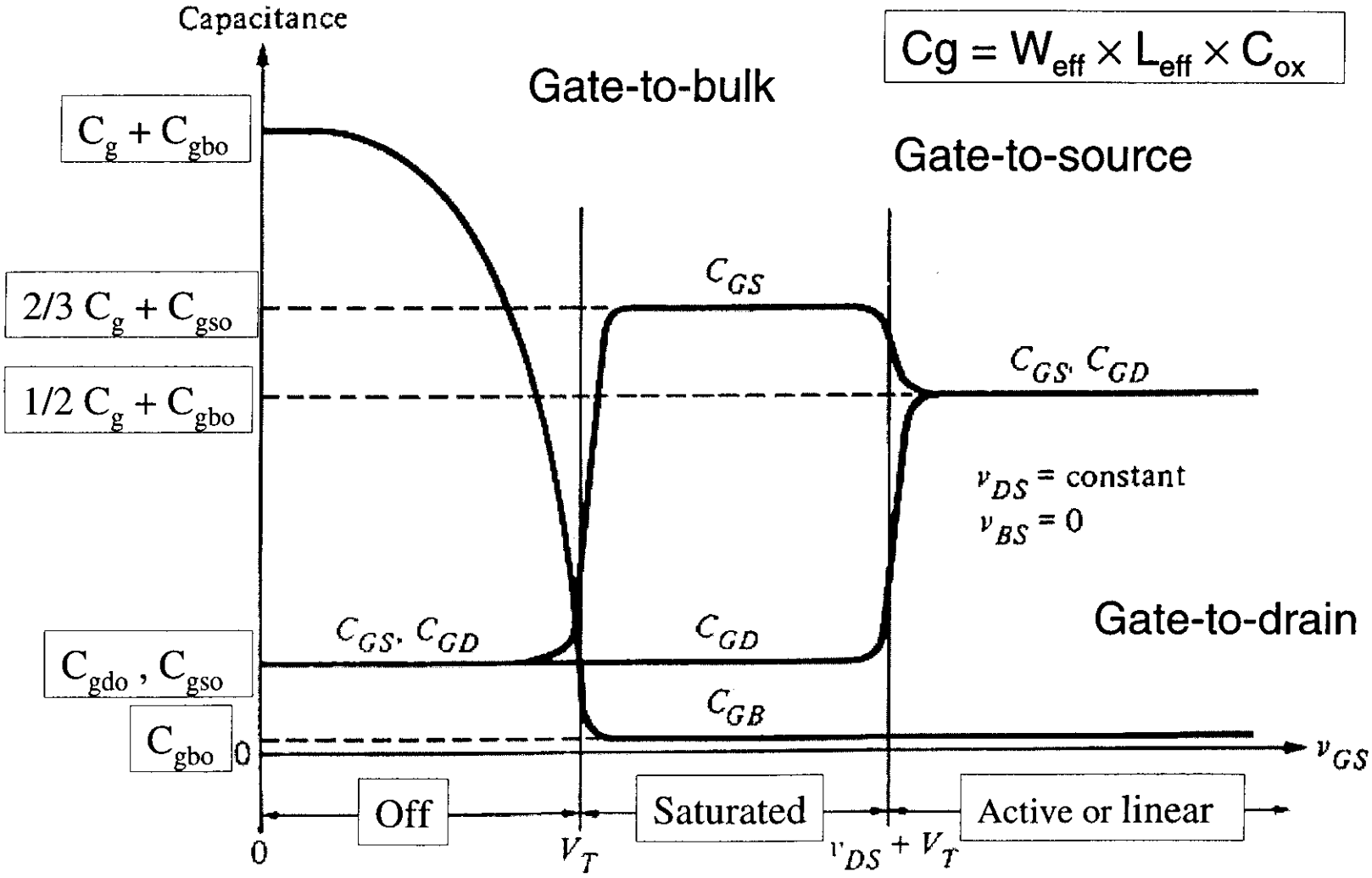
  C_{gb} - gate-to-bulk capacitance

  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



$$Cg = W_{eff} \times L_{eff} \times C_{ox}$$

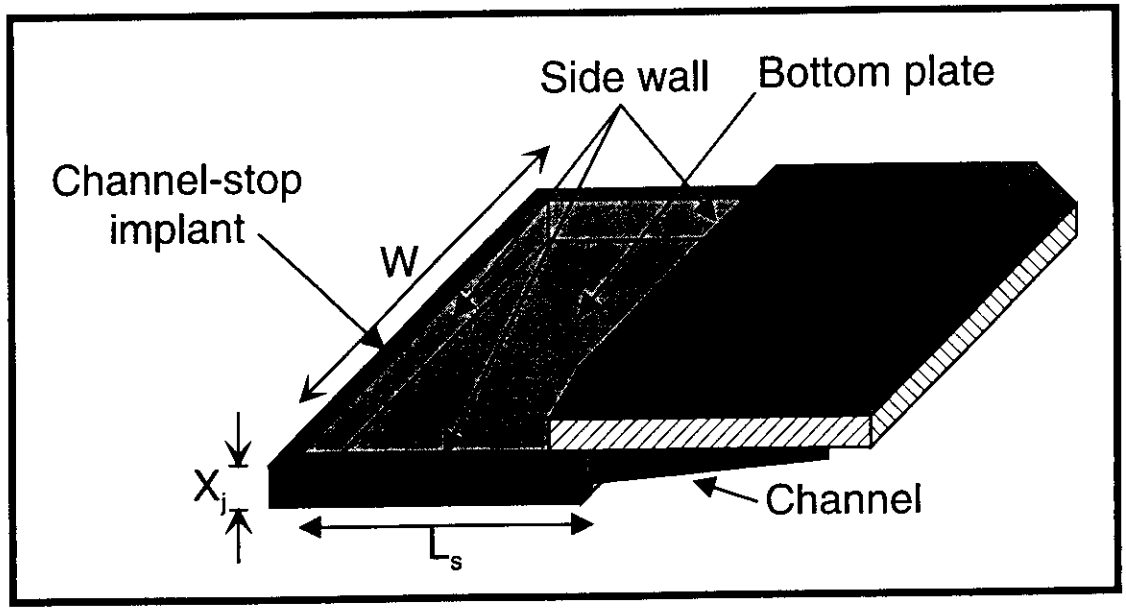
Junction capacitances

- 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:

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

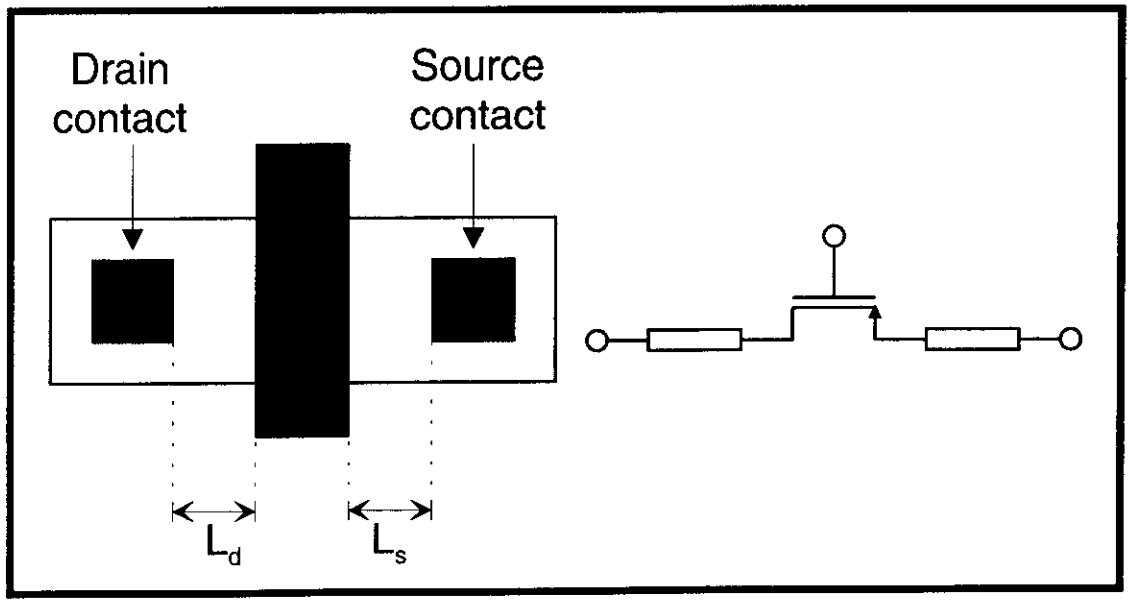


Source/drain resistance

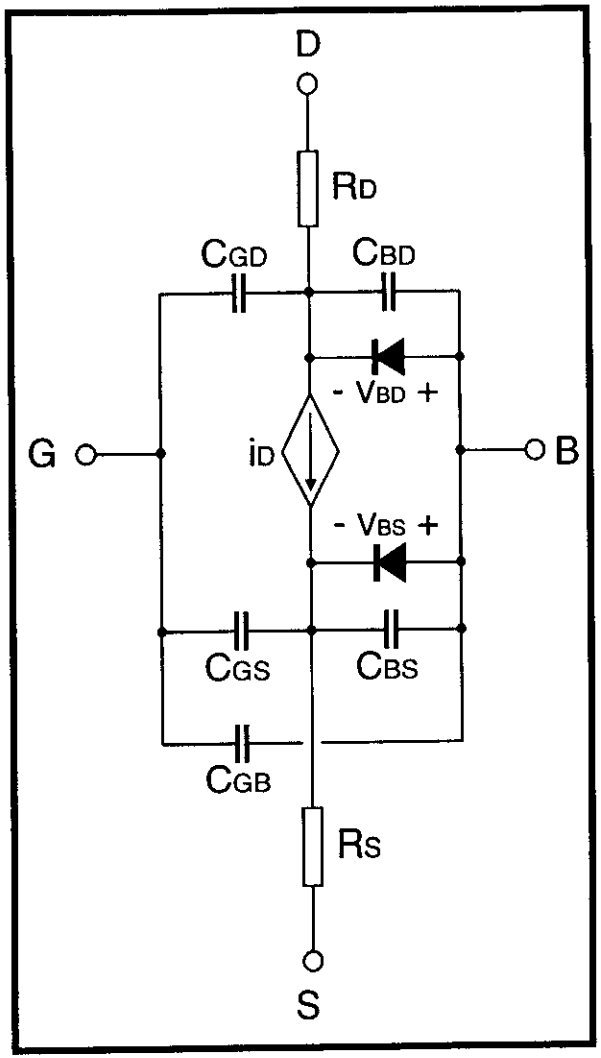
- Scaled down devices ⇒ higher source/drain resistance:

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

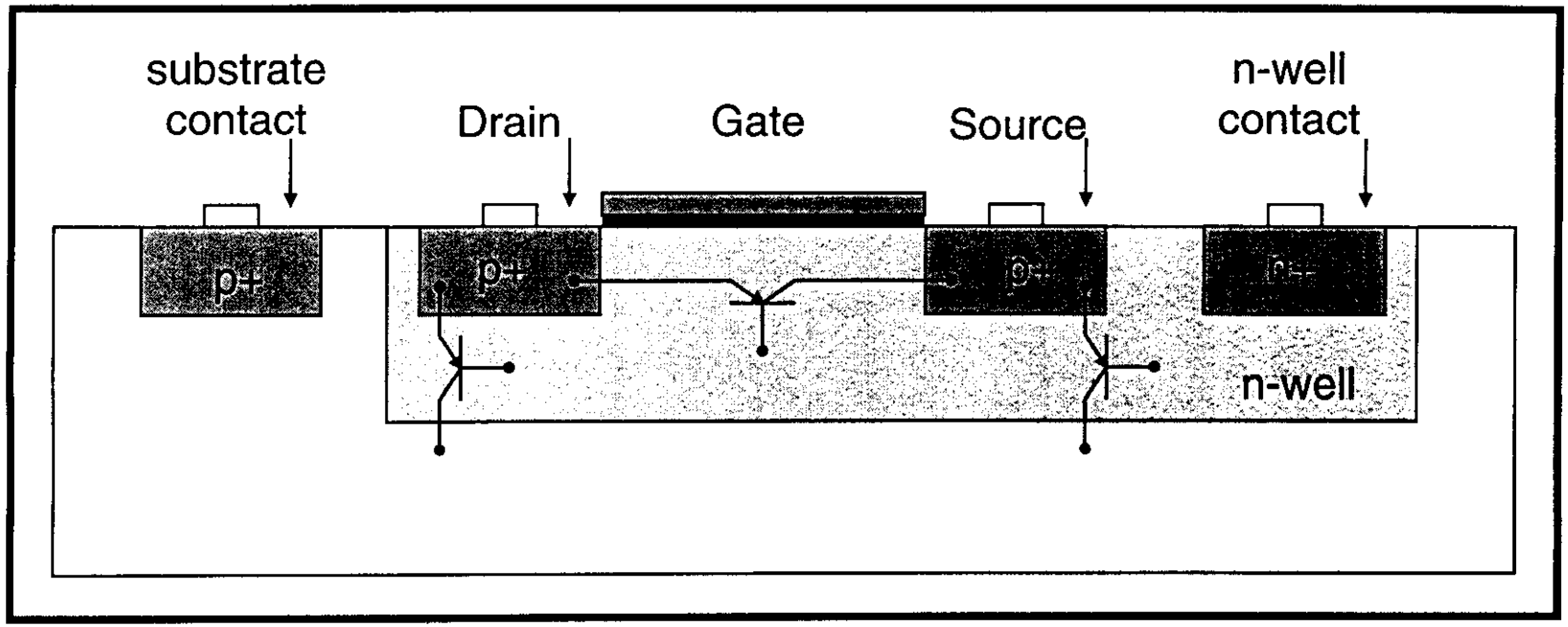
- In sub-μ processes *silicidation* is used to reduce the source, drain and gate parasitic resistance



MOSFET model

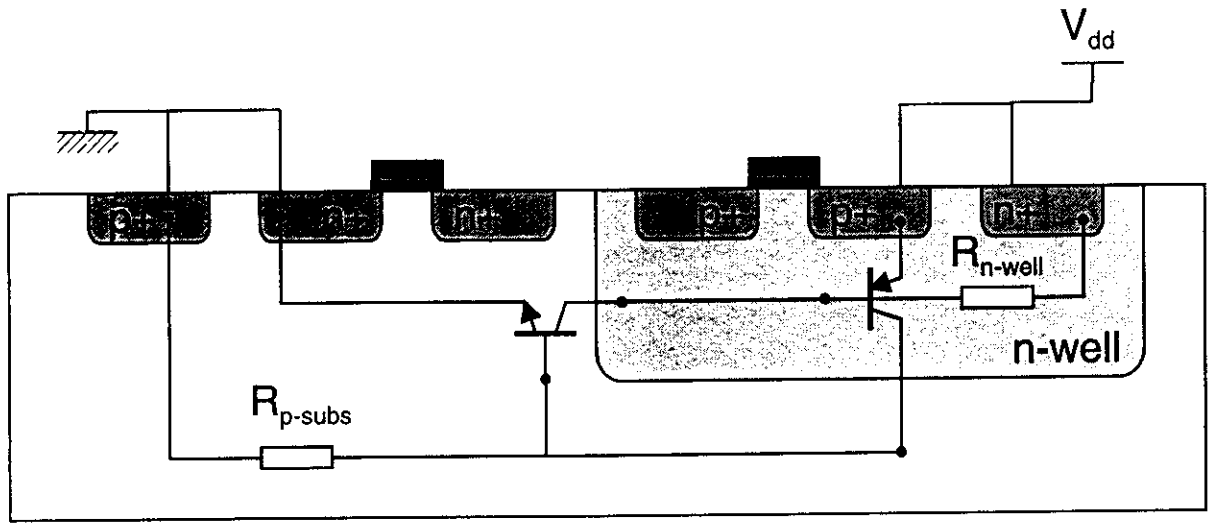


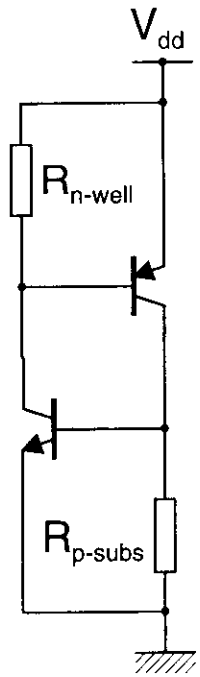
CMOS parasitic bipolar



CMOS device hazards

CMOS latchup origin

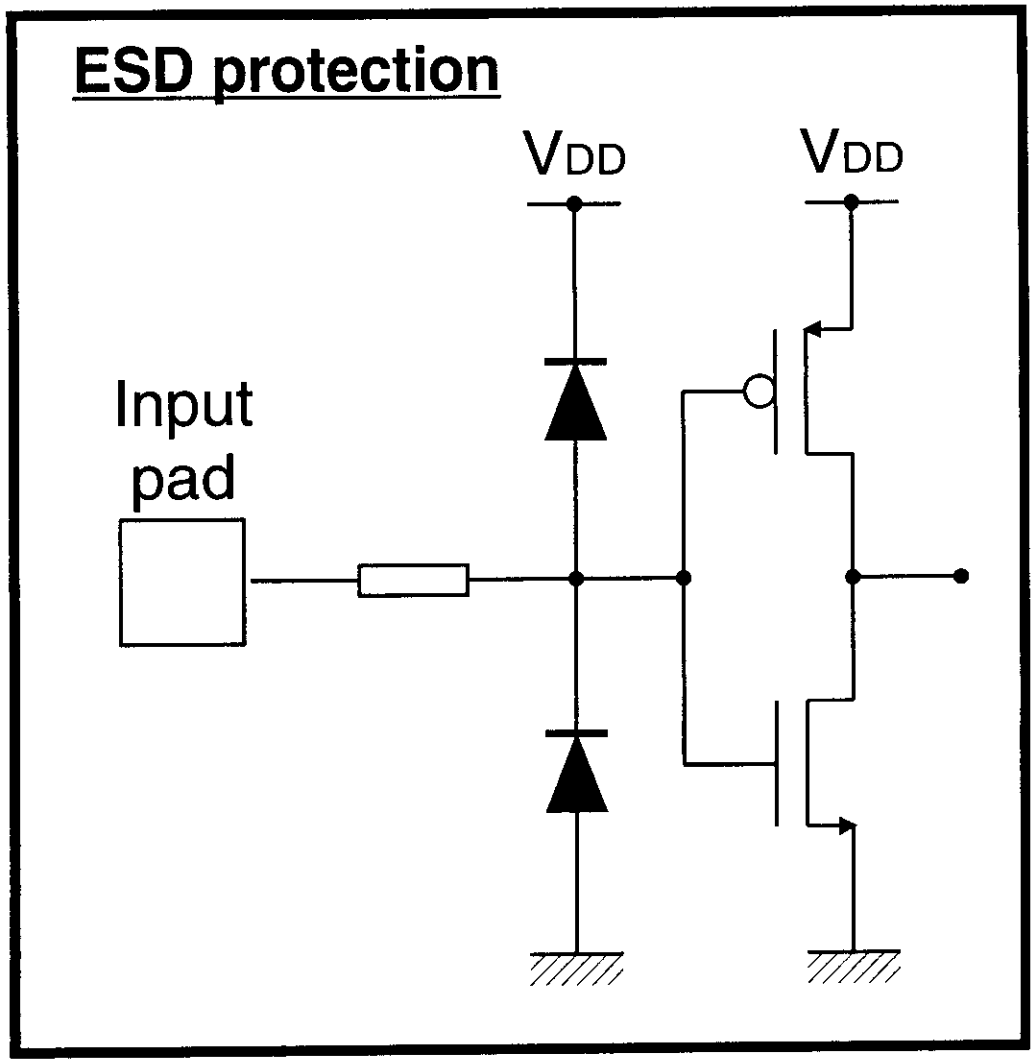




CMOS device hazards

- 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

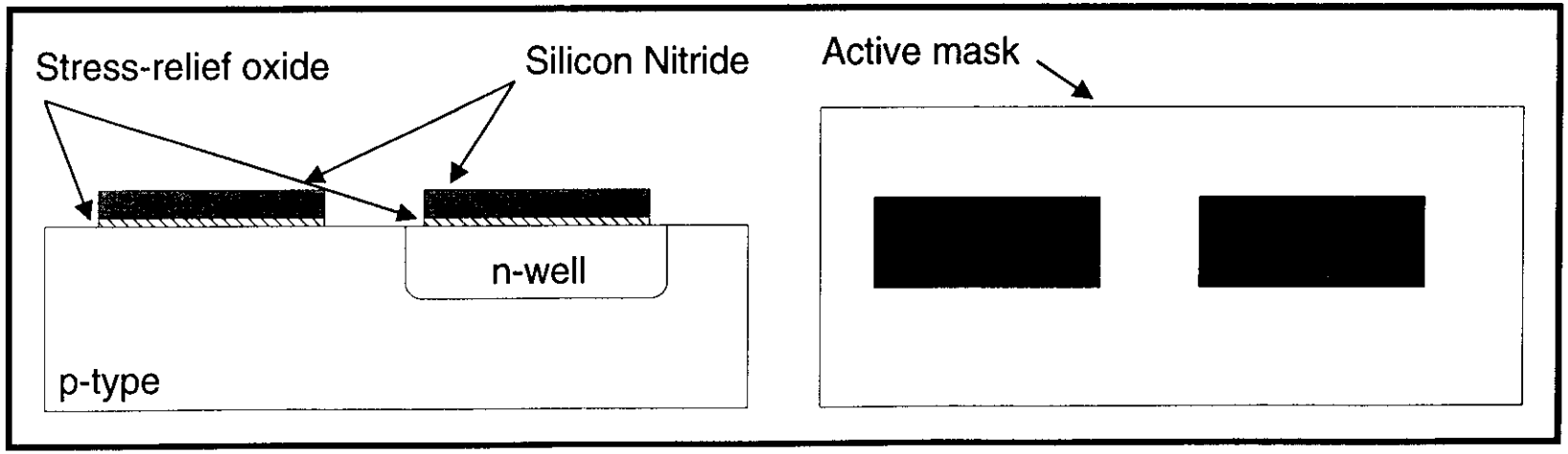
CMOS device hazards



CMOS fabrication sequence

3. Active area definition:

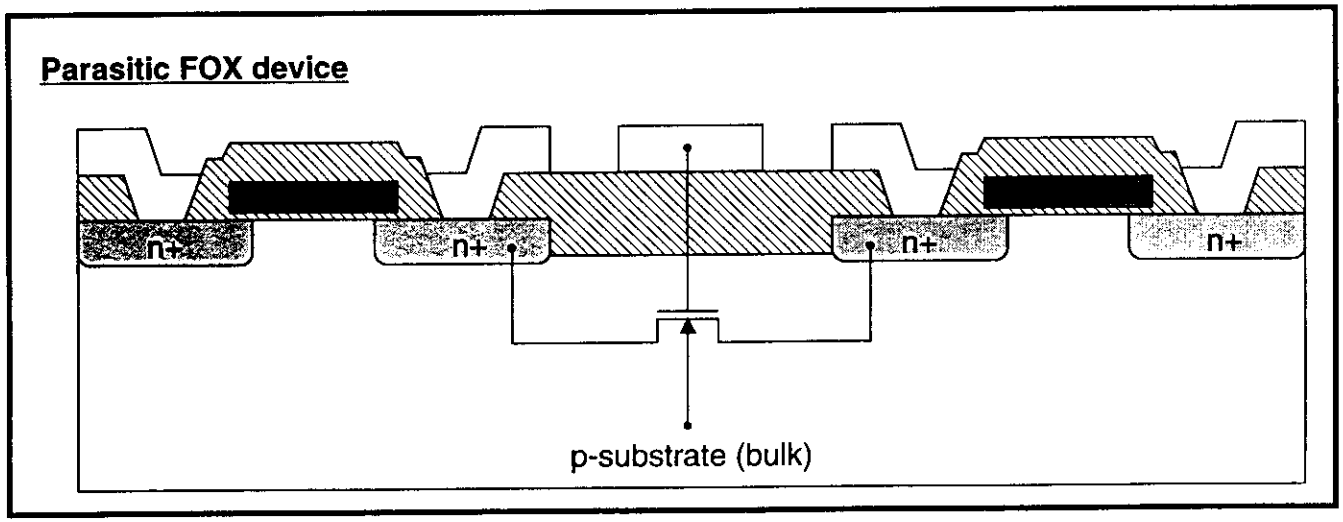
- Active area:
  - planar section of the surface where transistors are build
  - defines the gate region (thin oxide)
  - defines the n+ or p+ regions
- A thin layer of SiO_2 is grown over the active region and covered with silicon nitride



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

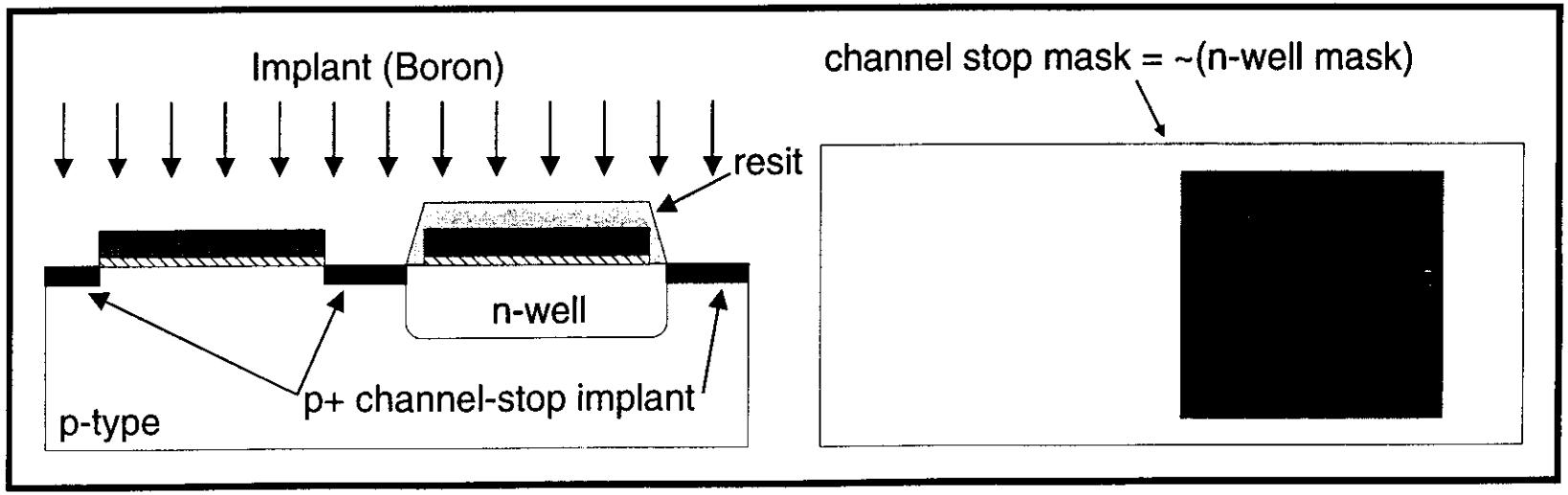


CMOS fabrication sequence

- 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

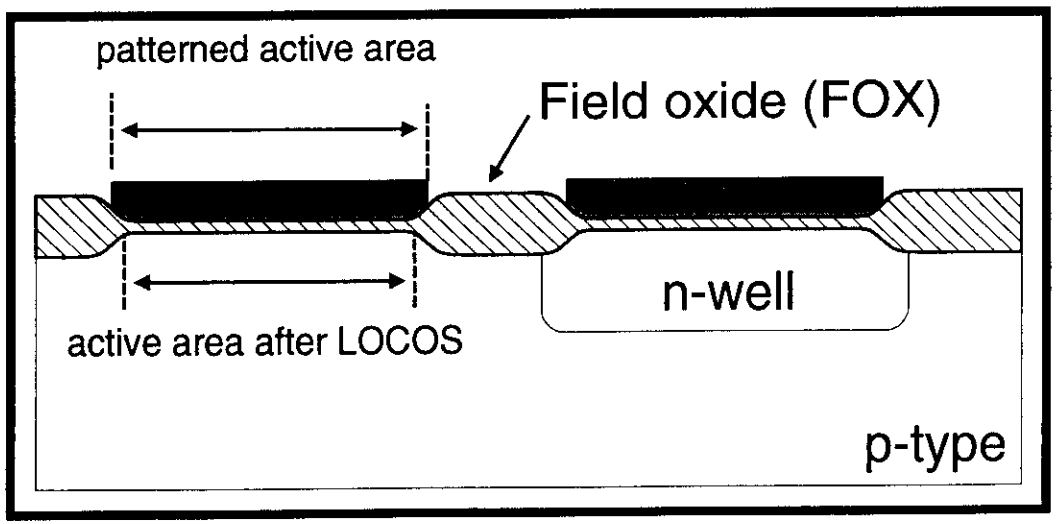
- The silicon nitride (over n-active) and the photoresist (over n-well) act as masks for the channel-stop implant



CMOS fabrication sequence

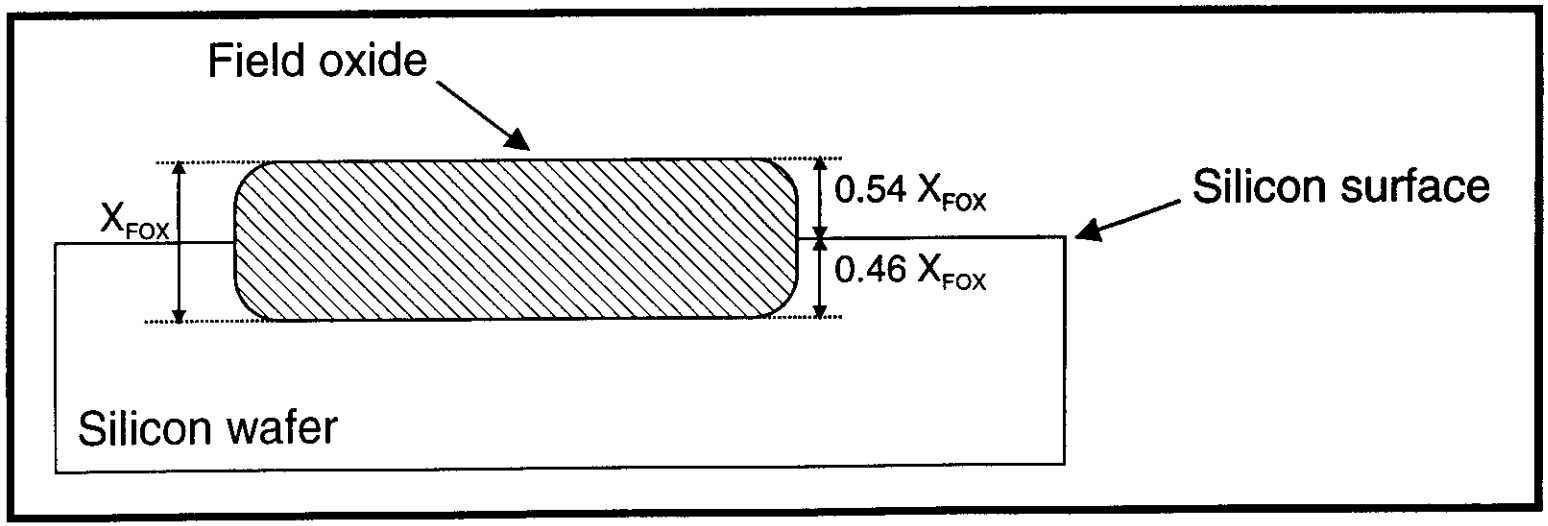
4.2 Local oxidation of silicon (LOCOS)

- 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 1000ºC (rapid process)
    - Dry oxidation: Pure oxygen, T = 1200ºC (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 $0.46X_{FOX}$



Outline

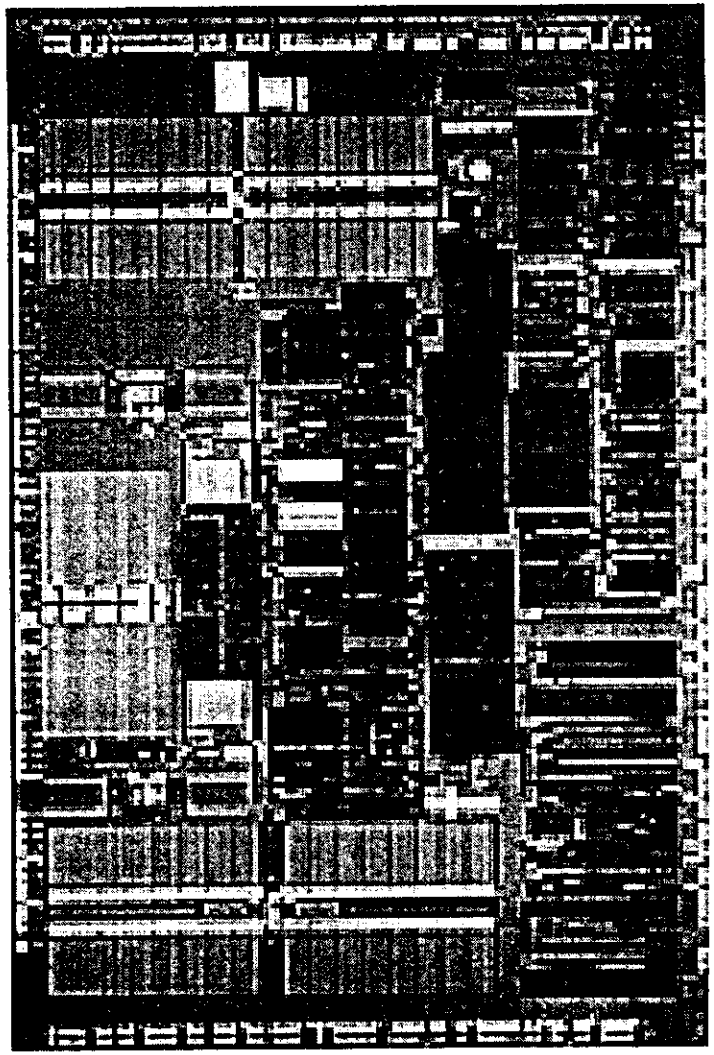
- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS technology

- Lithography
- Physical structure
- CMOS fabrication sequence
- Yield
- Design rules
- Other processes
- Process enhancements
- 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

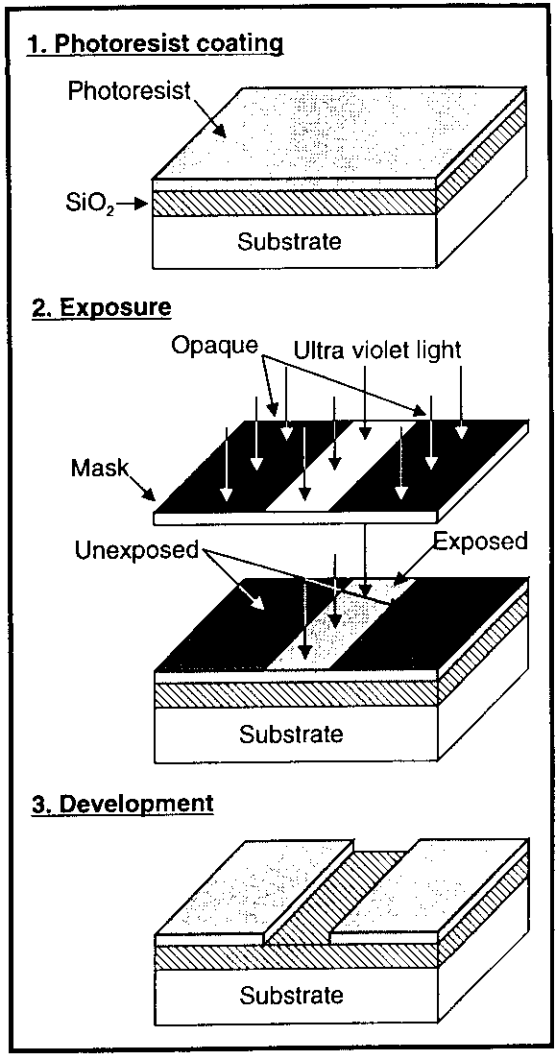
Lithography sequence steps:

- 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

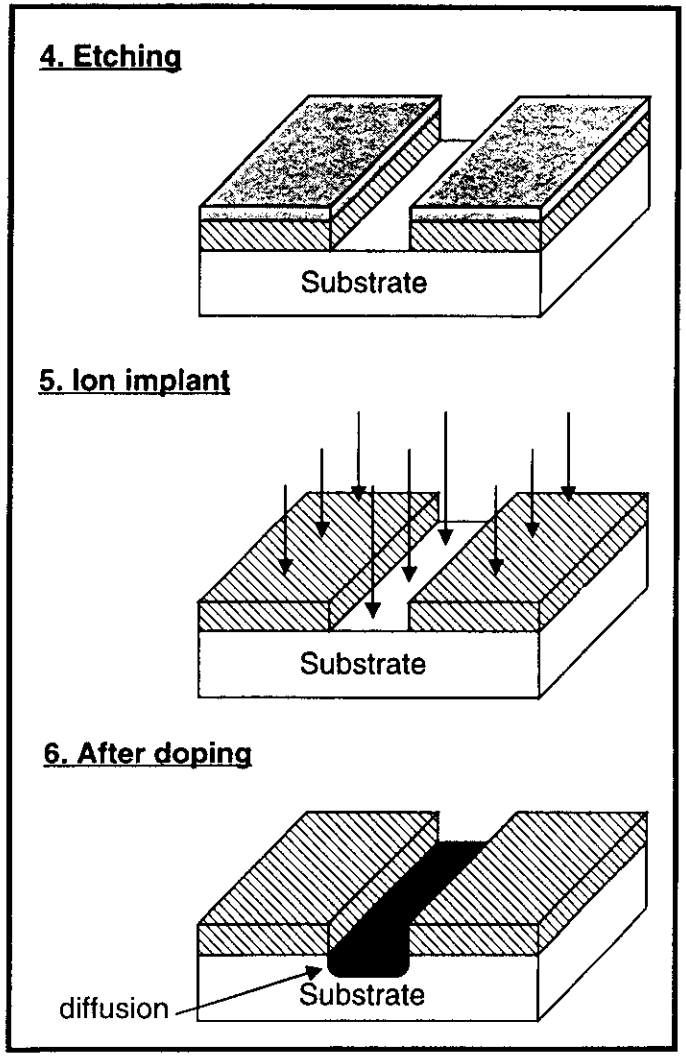
Basic sequence

- 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

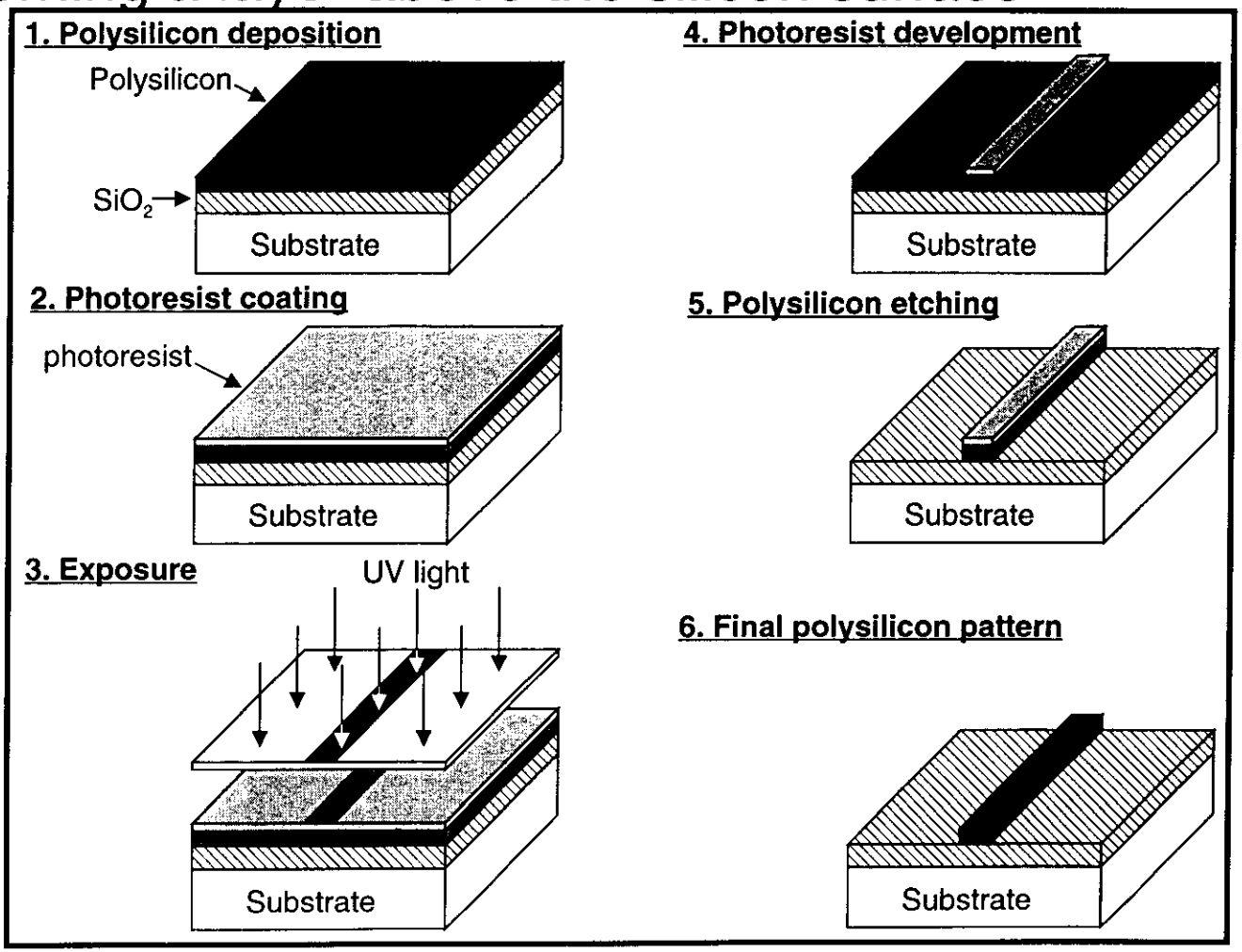


Lithography

- 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

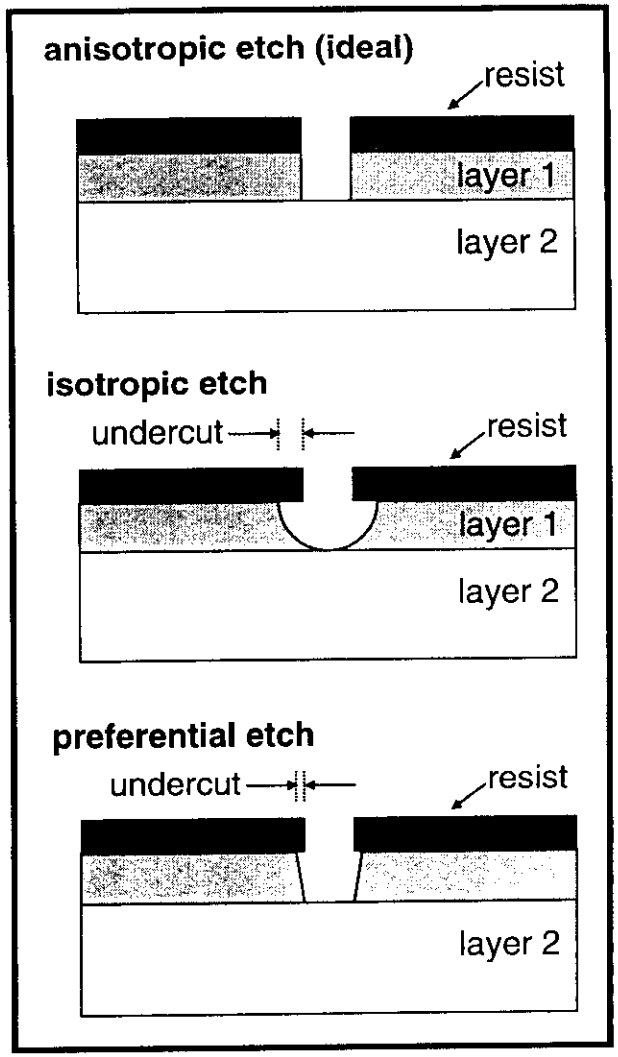
Lithography

Patterning a layer above the silicon surface

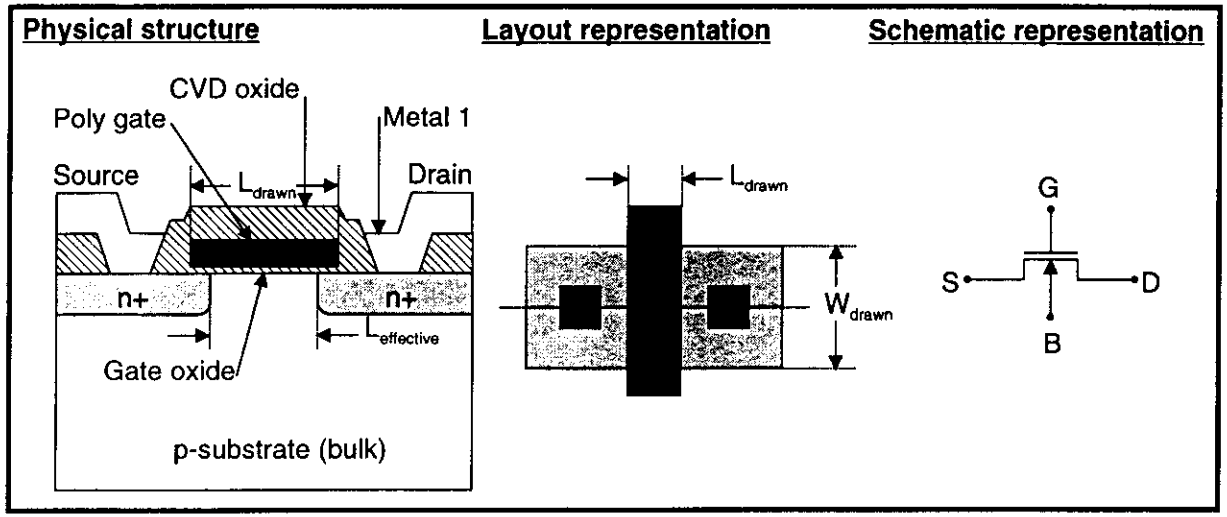


Lithography

- 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



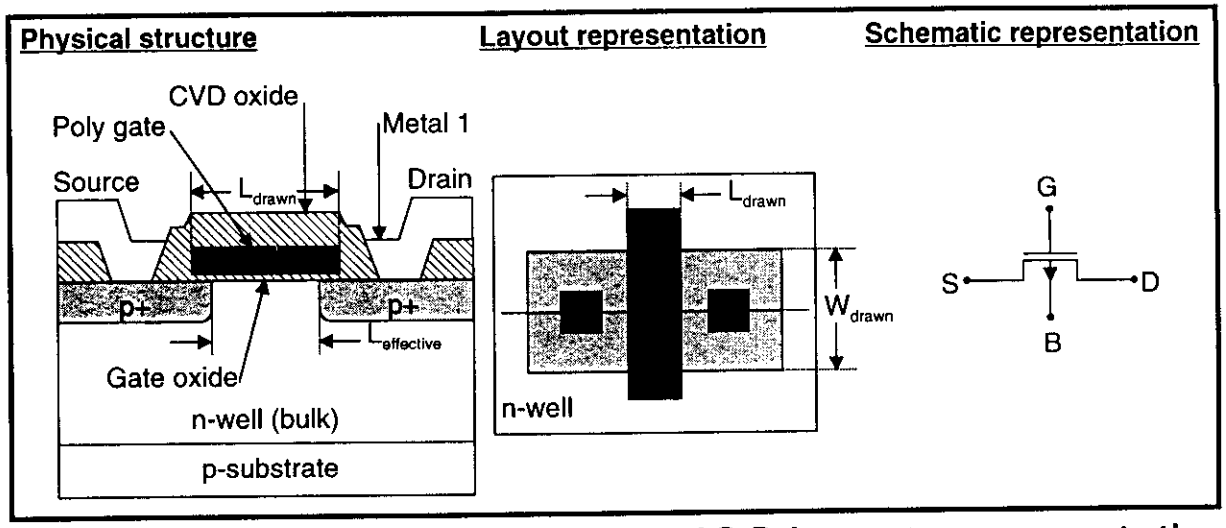
NMOS physical structure:

- 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

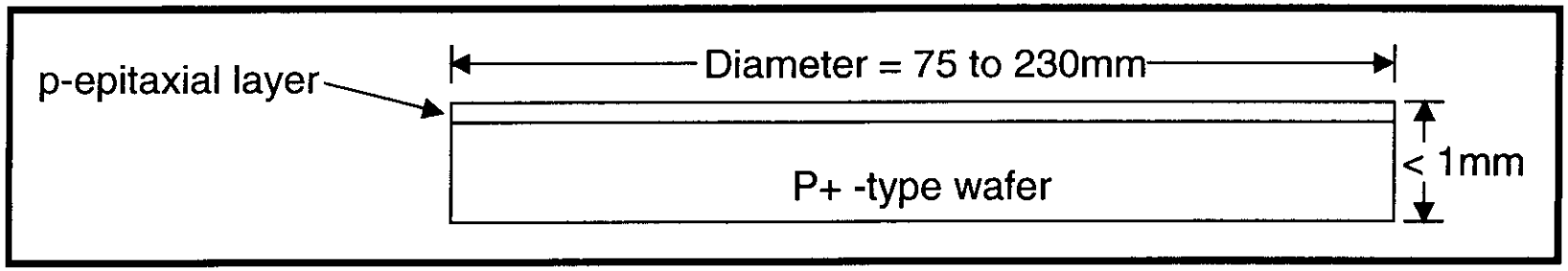
CMOS fabrication sequence

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:

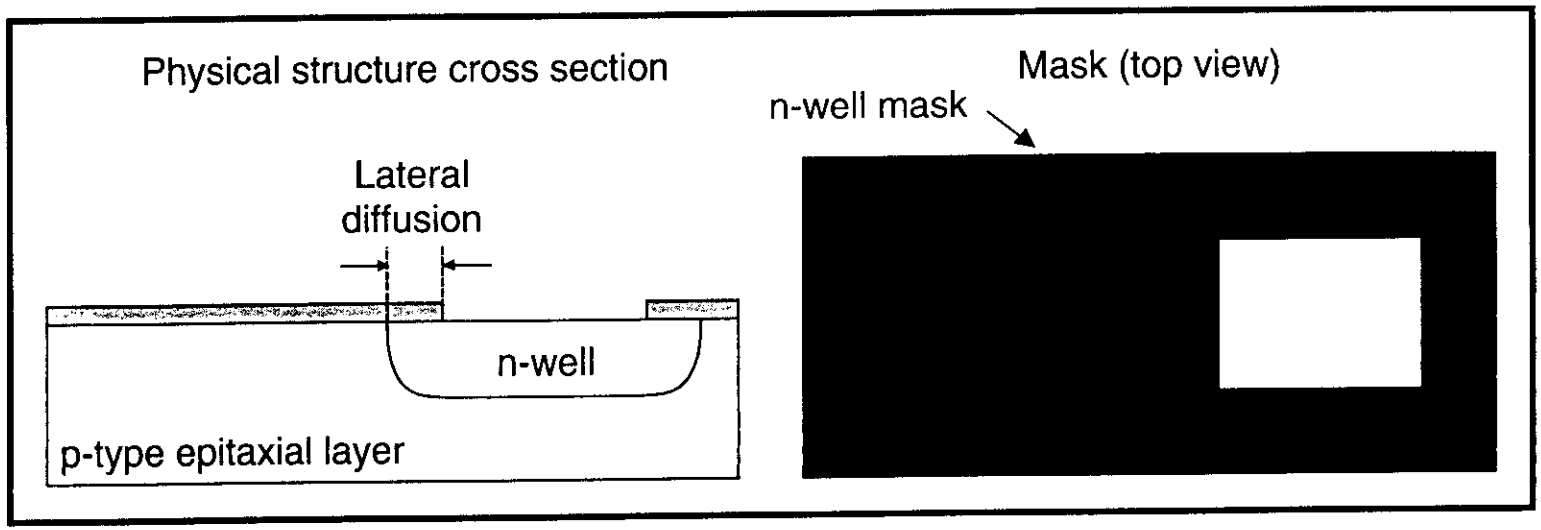
- A single p-type single crystal film is grown on the surface of the wafer by:
  - subjecting the wafer to high temperature and a source of dopant material
- The epi layer is used as the base layer to build the devices



CMOS fabrication sequence

2. N-well Formation:

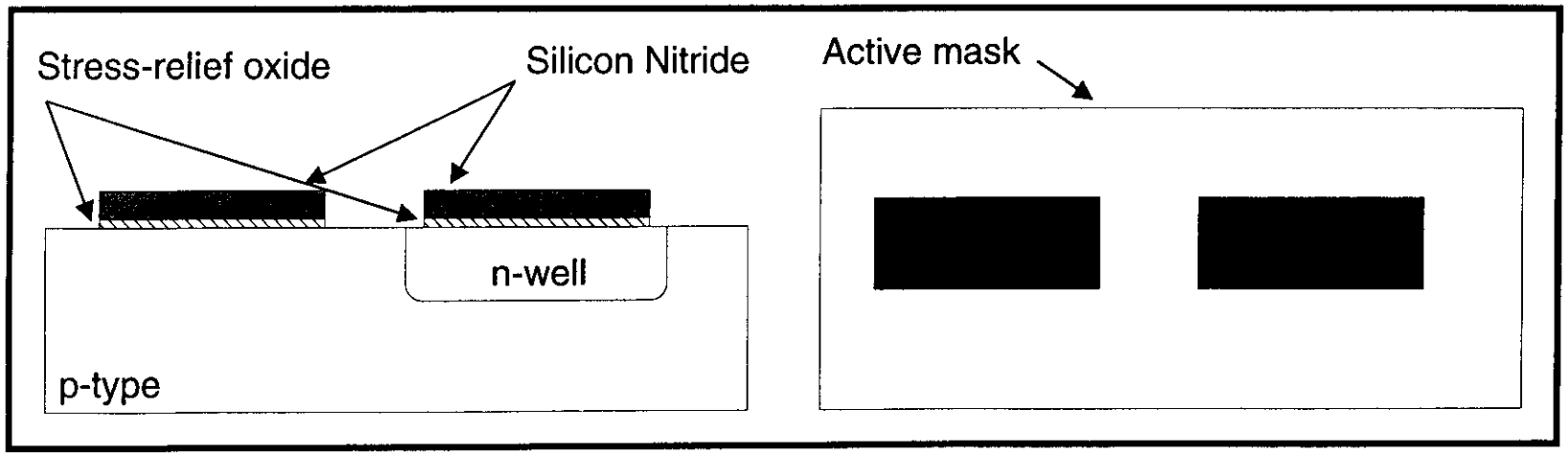
- PMOS transistors are fabricated in n-well regions
- The first mask defines the n-well regions
- N-well's are formed by ion implantation or deposition and diffusion
- Lateral diffusion limits the proximity between structures
- Ion implantation results in shallower wells compatible with today's fine-line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



CMOS fabrication sequence

3. Active area definition:

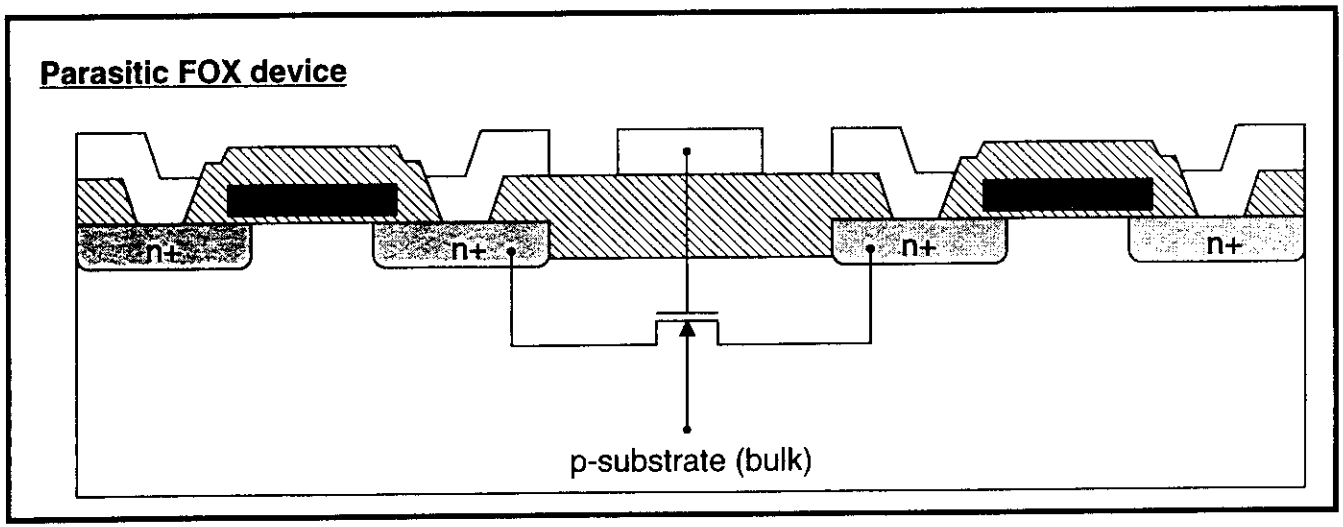
- Active area:
  - planar section of the surface where transistors are build
  - defines the gate region (thin oxide)
  - defines the n+ or p+ regions
- A thin layer of SiO_2 is grown over the active region and covered with silicon nitride



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

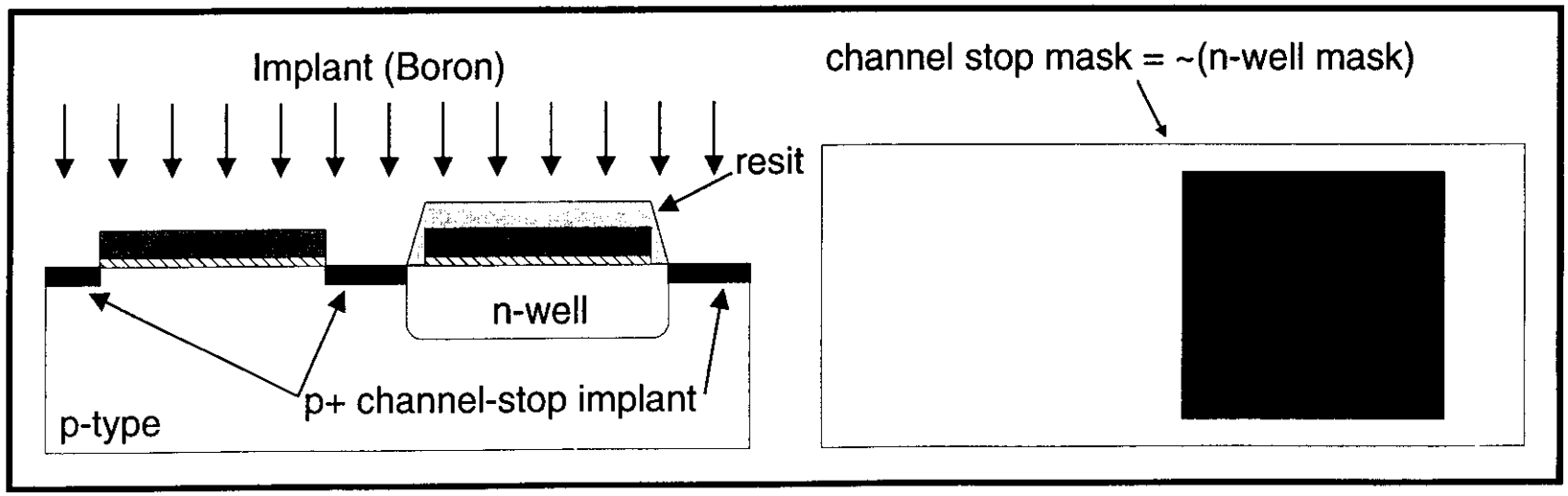


CMOS fabrication sequence

- 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

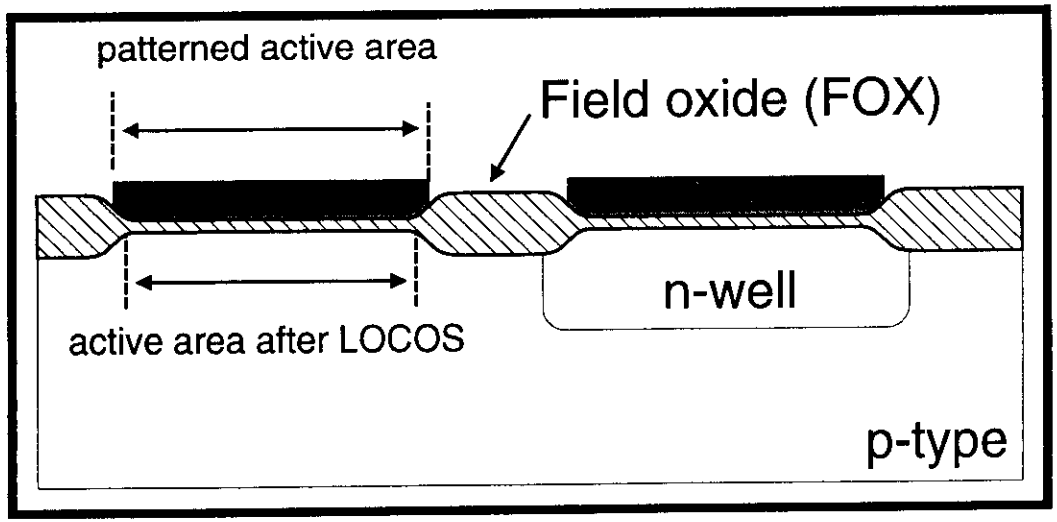
- The silicon nitride (over n-active) and the photoresist (over n-well) act as masks for the channel-stop implant



CMOS fabrication sequence

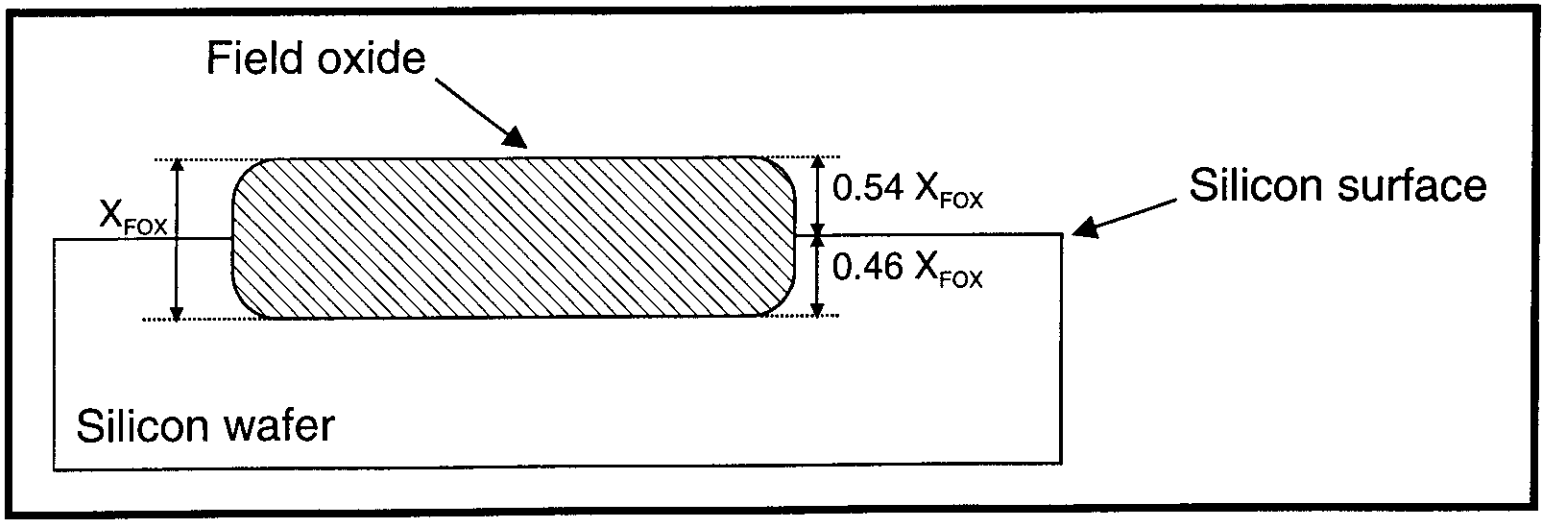
4.2 Local oxidation of silicon (LOCOS)

- 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

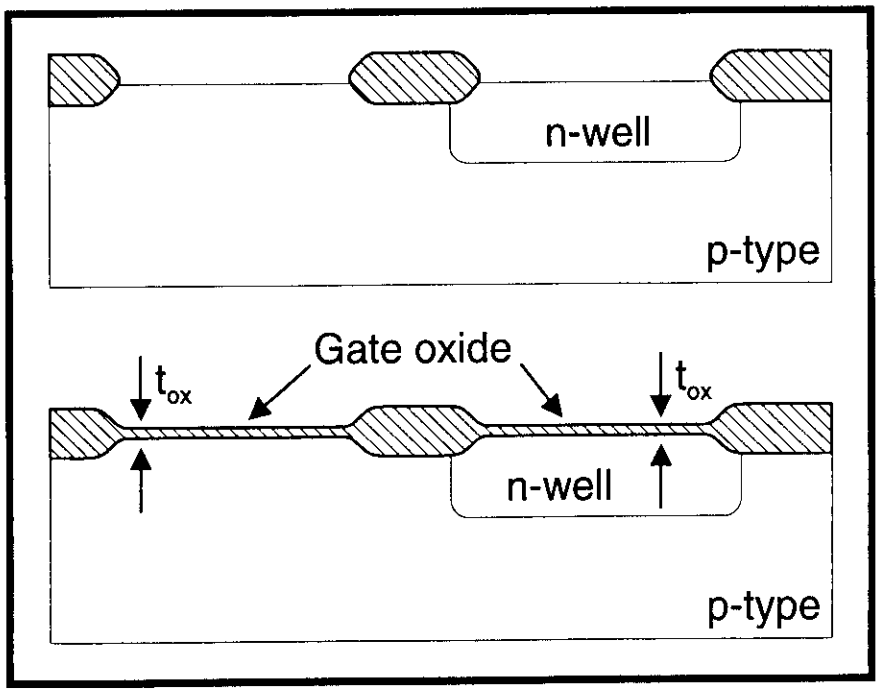
- Silicon oxidation is obtained by:
  - Heating the wafer in a oxidizing atmosphere:
    - Wet oxidation: water vapor, T = 900 to 1000ºC (rapid process)
    - Dry oxidation: Pure oxygen, T = 1200ºC (high temperature required to achieve an acceptable growth rate)
- Oxidation consumes silicon
  - SiO_2 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

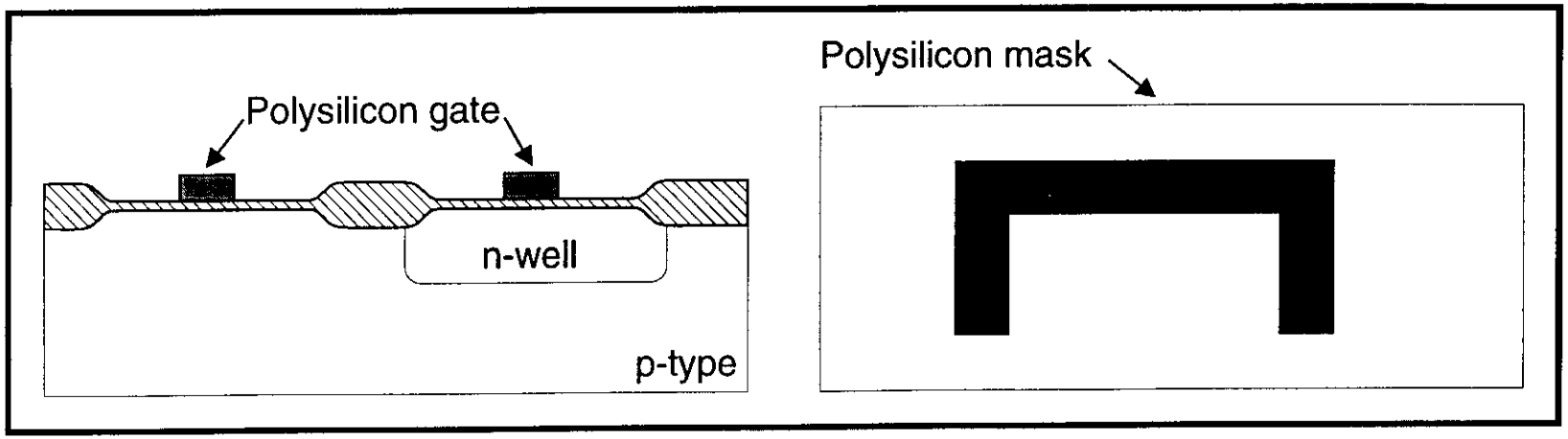
- The nitride and stress-relief oxide are removed
- The devices threshold voltage is adjusted by:
  - adding charge at the silicon/oxide interface
- The well controlled gate oxide is grown with thickness t_{ox}



CMOS fabrication sequence

6. Polysilicon deposition and patterning

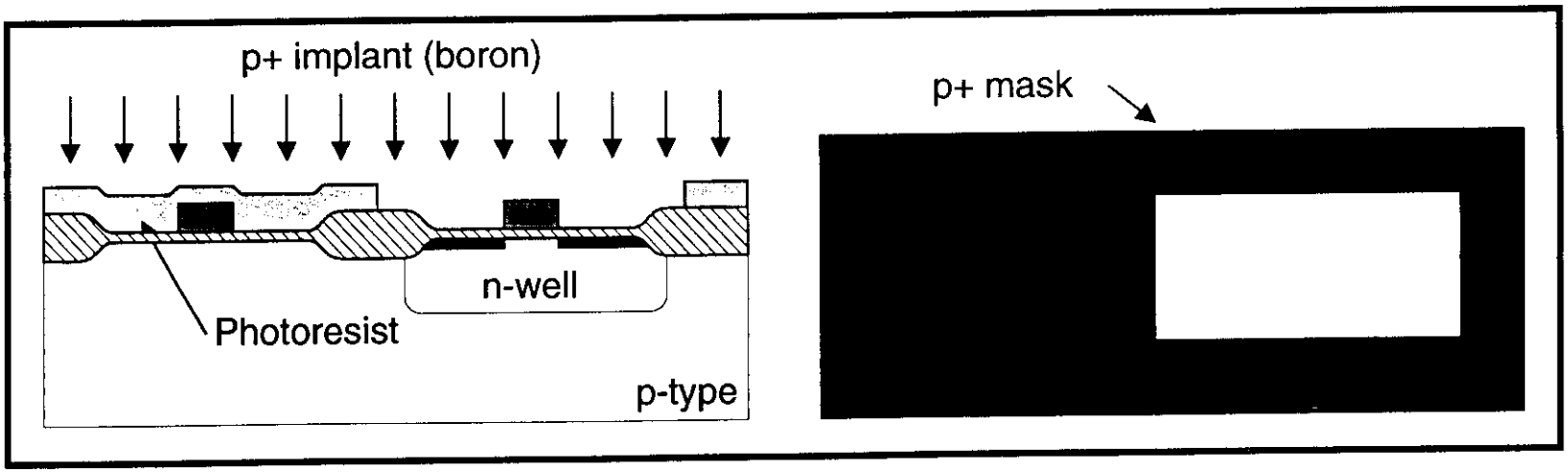
- A layer of polysilicon is deposited over the entire wafer surface
- The polysilicon is then patterned by a lithography sequence
- All the MOSFET gates are defined in a single step
- The polysilicon gate can be doped (n+) while is being deposited to lower its parasitic resistance (important in high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



CMOS fabrication sequence

7. PMOS formation

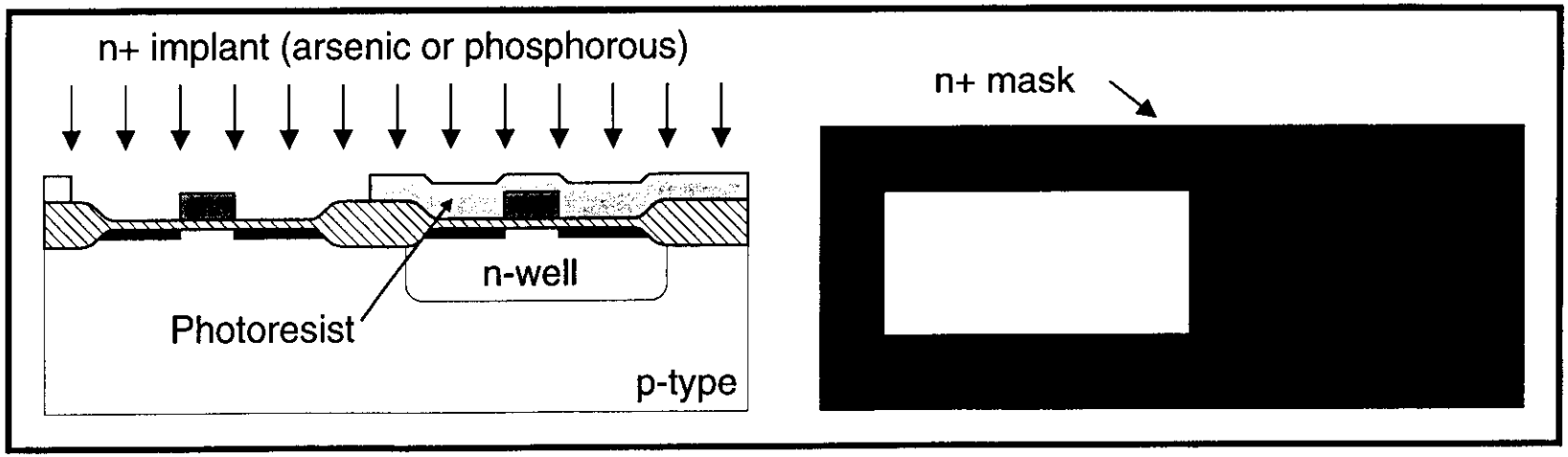
- Photoresist is patterned to cover all but the p+ regions
- A boron ion beam creates the p+ source and drain regions
- The polysilicon serves as a mask to the underlying channel
  - This is called a self-aligned process
  - It allows precise placement of the source and drain regions
- During this process 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

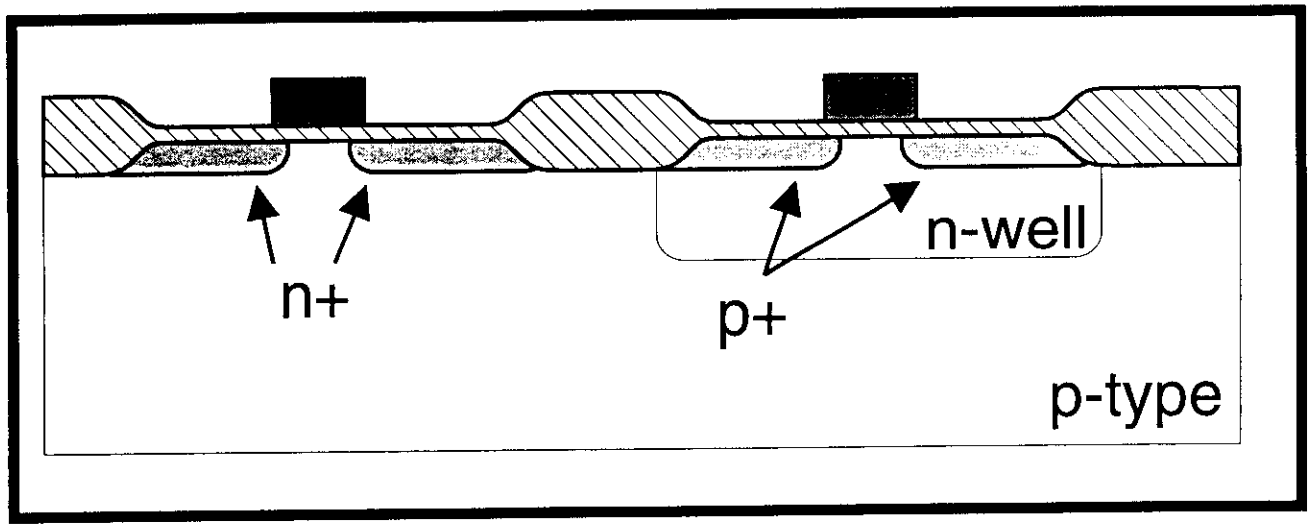
- Photoresist is patterned to define the n+ regions
- Donors (arsenic or phosphorous) are ion-implanted to dope the n+ source and drain regions
- The process is self-aligned
- The gate is n-type doped



CMOS fabrication sequence

9. Annealing

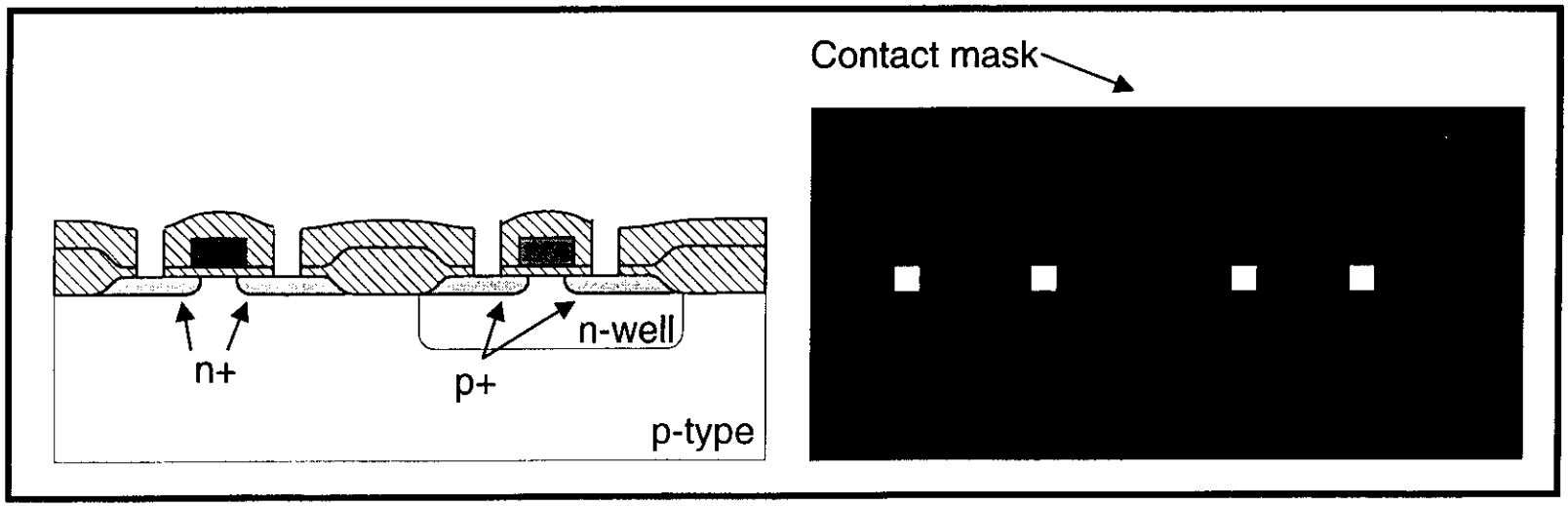
- After the implants are completed a thermal annealing cycle is executed
- This allows the impurities to diffuse further into the bulk
- After thermal annealing, it is important to keep the remaining process steps at as low temperature as possible



CMOS fabrication sequence

10. Contact cuts

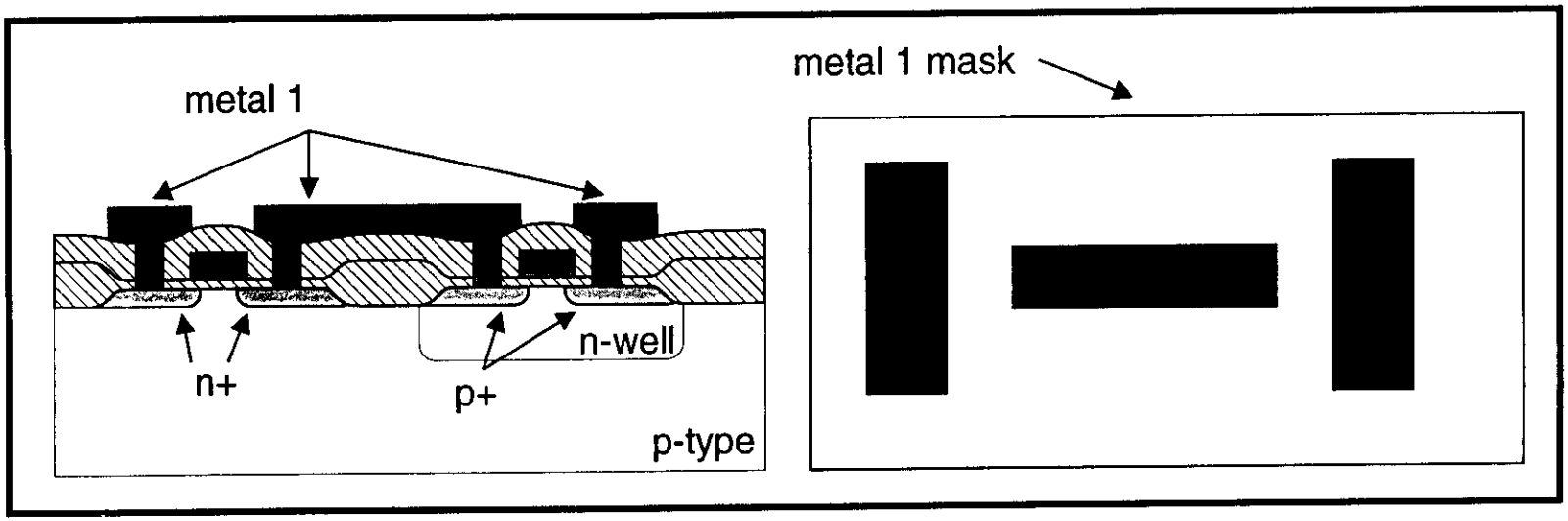
- The surface of the IC is covered by a layer of CVD oxide
  - The oxide is deposited at low temperature (LTO) to avoid that underlying doped regions will undergo diffusive spreading
- Contact cuts are defined by etching SiO_2 down to the surface to be contacted
- These allow metal to contact diffusion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



CMOS fabrication sequence

11. Metal 1

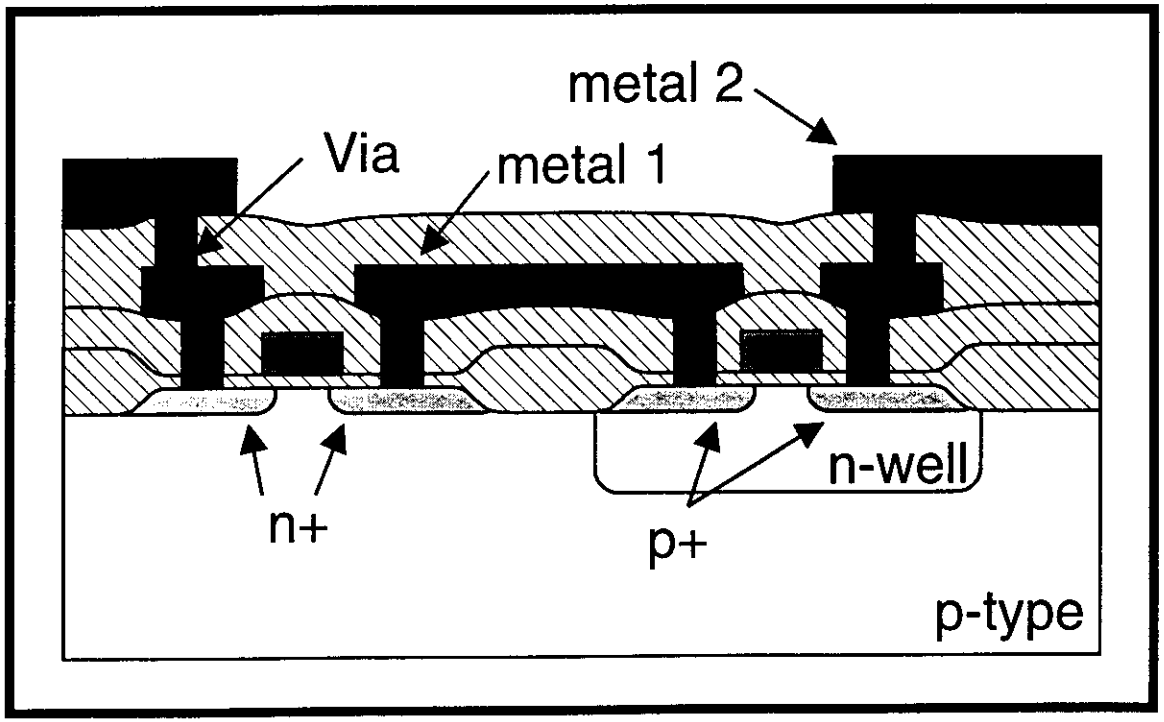
– A first level of metallization is applied to the wafer surface and selectively etched to produce the interconnects



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

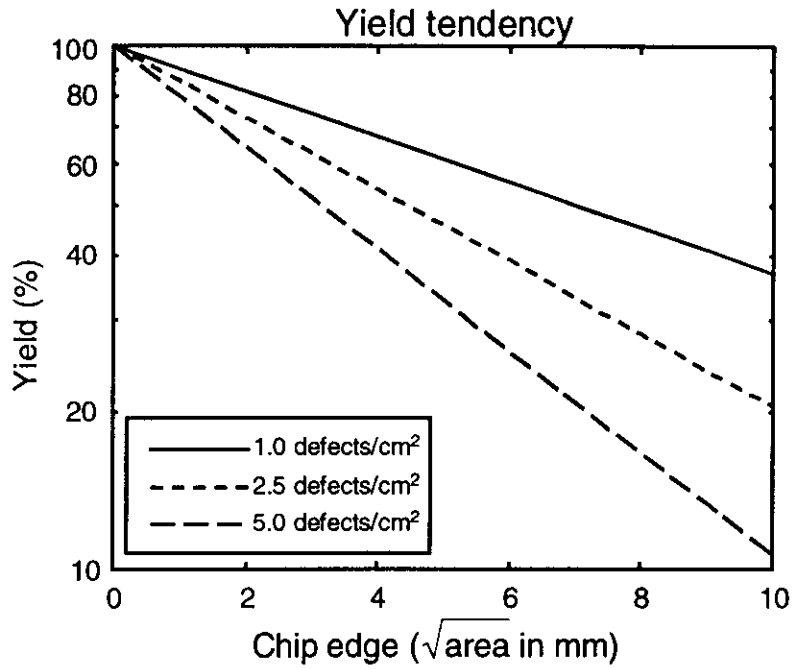
- Yield

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

- 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: $Y = e^{-\sqrt{A \cdot D}}$

  A - chip area (cm^2)

  D - defect density ($defects/cm^2$)

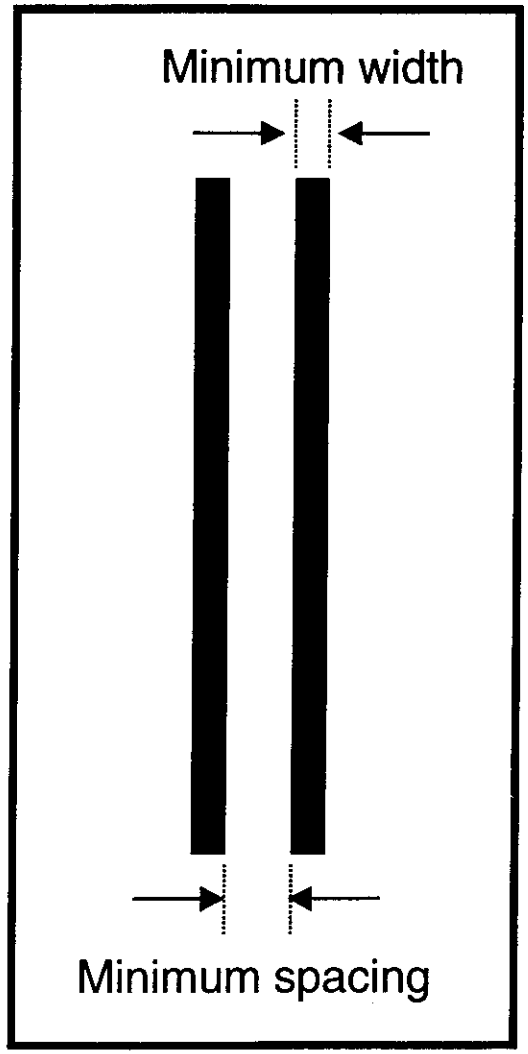


Design rules

- 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

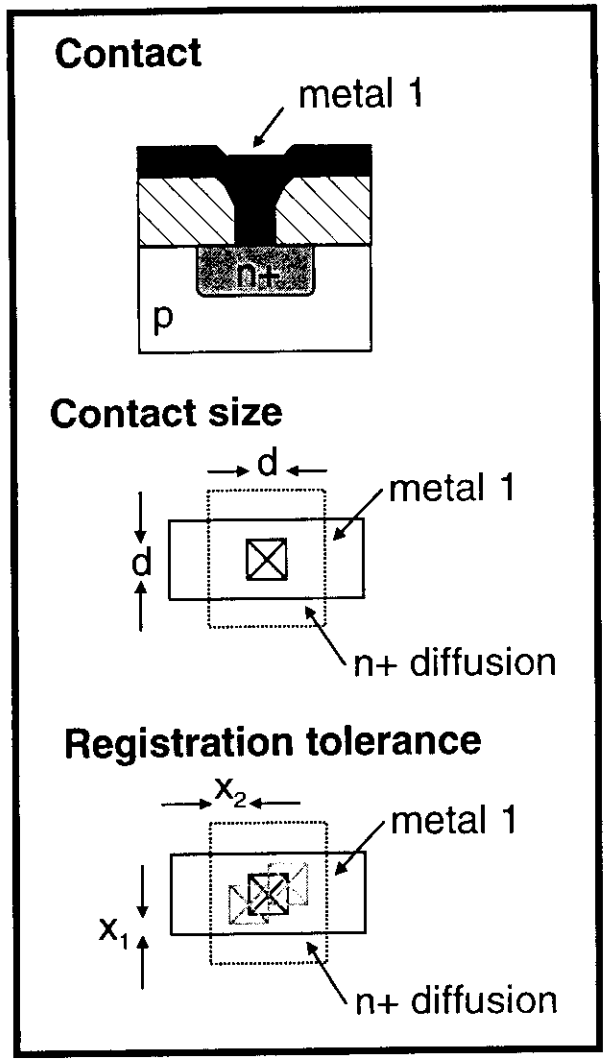
Design rules

- 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



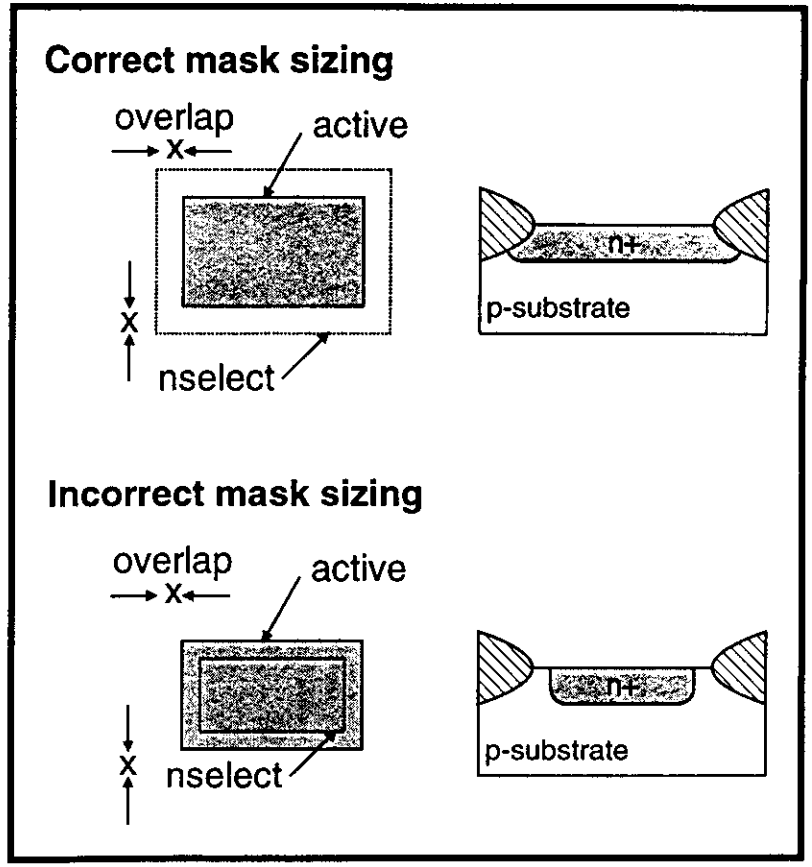
Design rules

- 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)



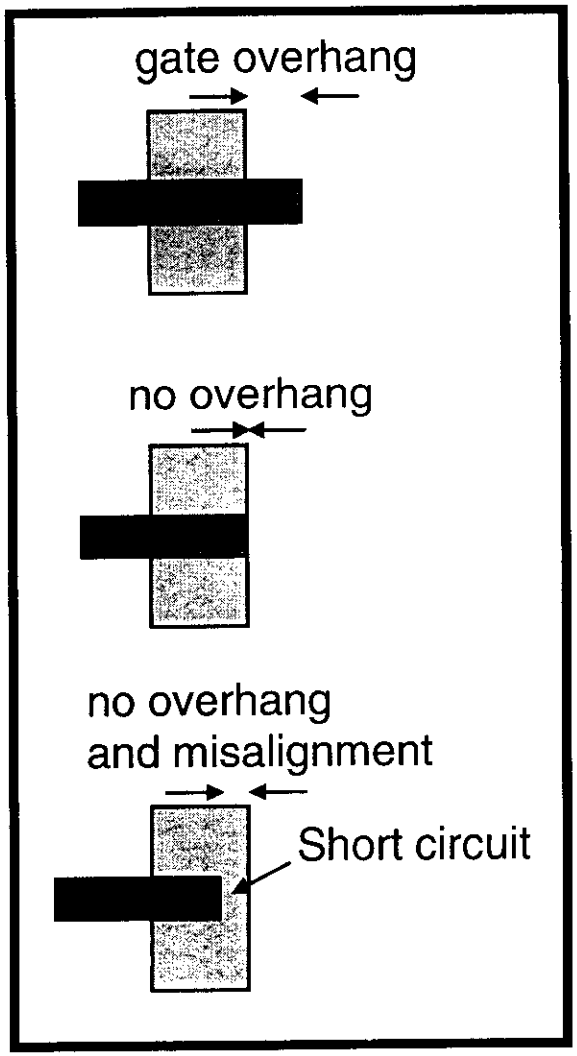
Design rules

- 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



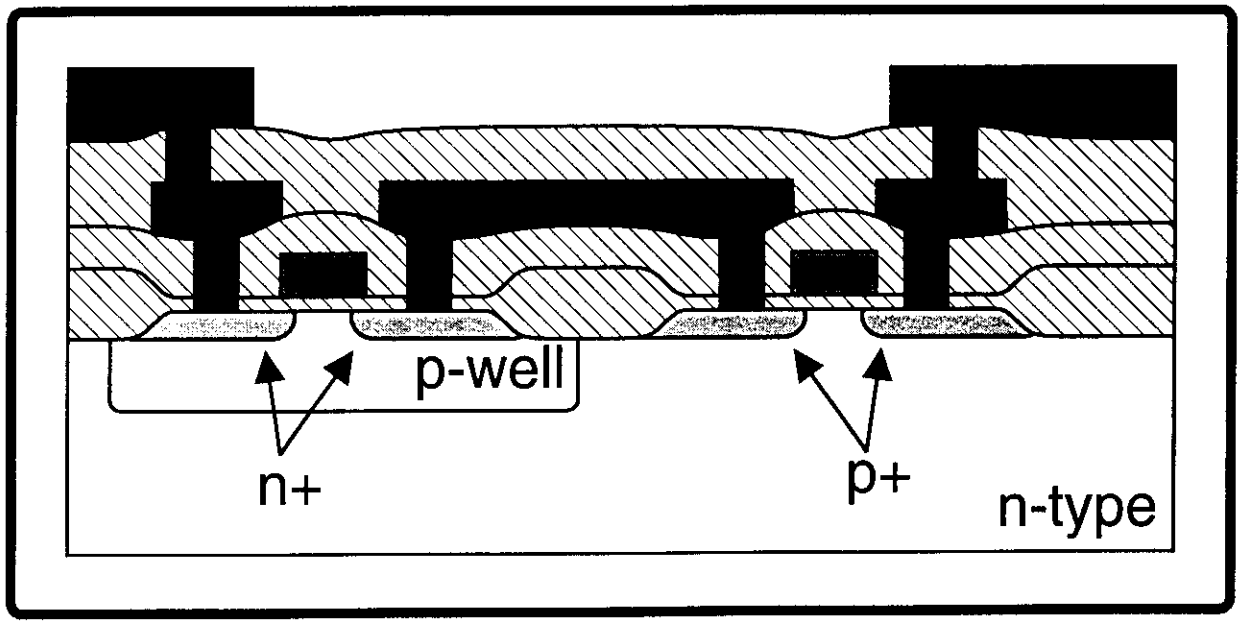
Design rules

- 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



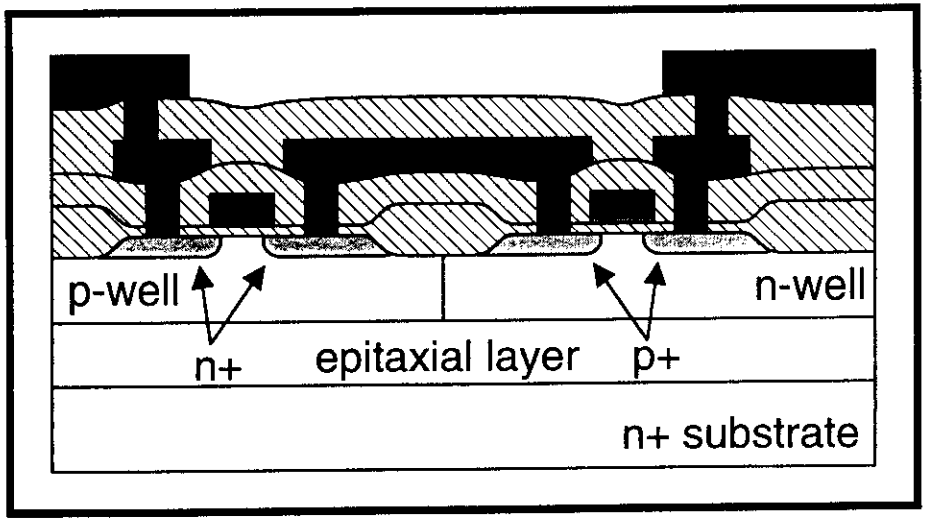
Other processes

- **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



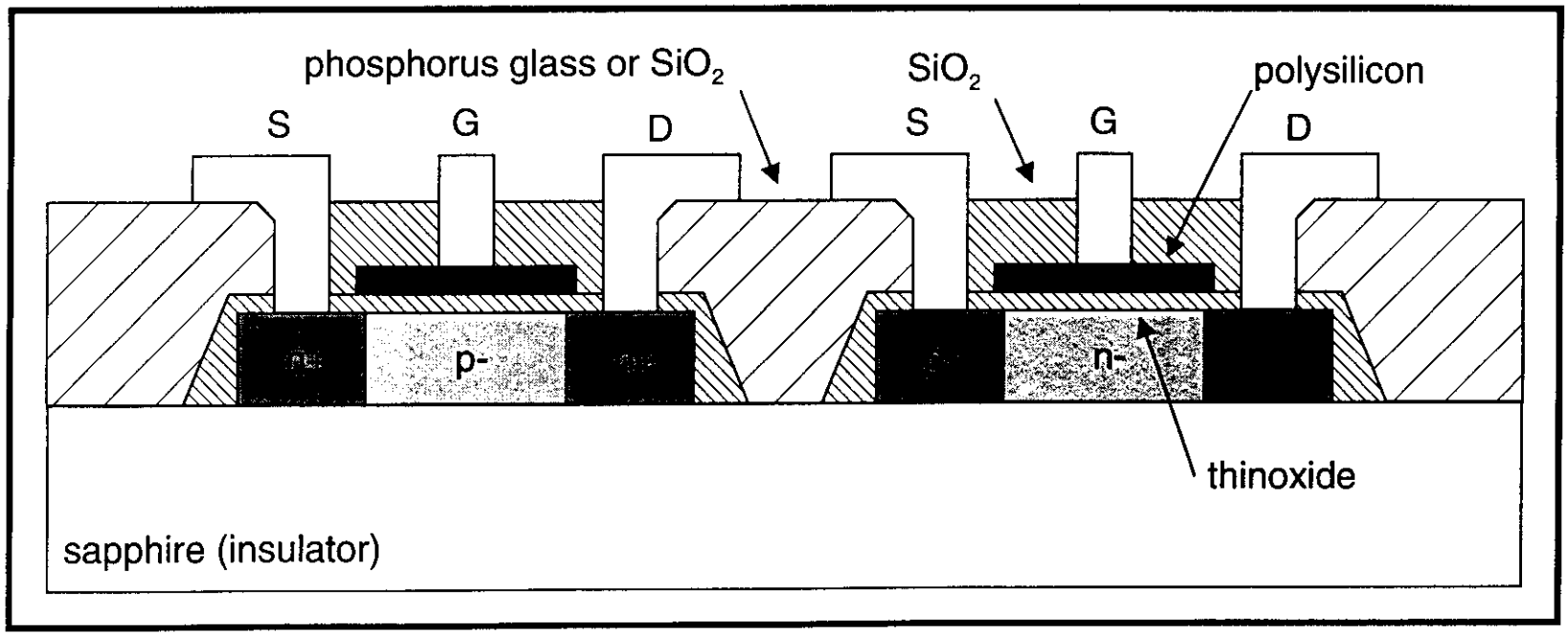
Other processes

- **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

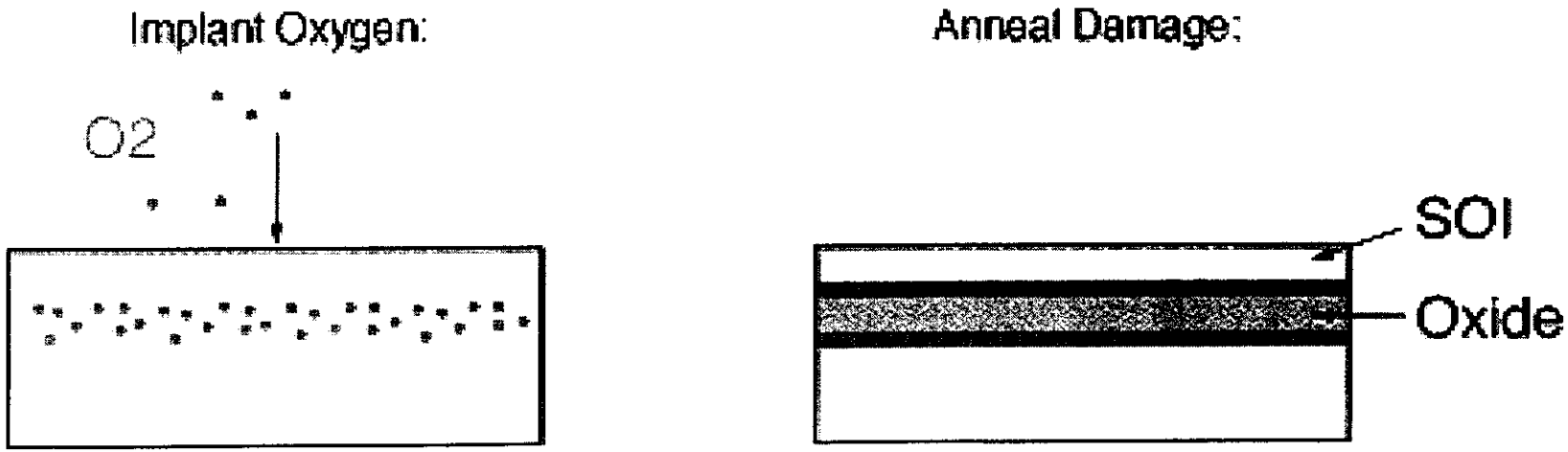


Other processes

  - 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 $\Rightarrow$ 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



Process enhancements

- Up to five metal levels in modern processes
- Copper for metal level 2 and higher
- Silicided polysilicon (e. g. silicon & tantalum)
- Silicided source and drain
- Stacked contacts and vias
- Chemical Metal Polishing for technologies with several metal levels
- Shallow trench isolation
- Lightly Doped Drain (LDD) structure
- For analogue applications some processes offer:
  - capacitors and resistors structures

Process enhancements

- Lithography:
  - Deep UV, $\lambda = 193$nm
  - X-rays
  - 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

Scaling

Parameter	constant field	constant voltage
length (L)	$1/\alpha$	$1/\alpha$
width (W)	$1/\alpha$	$1/\alpha$
supply voltage	$1/\alpha$	1
gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$
current ($I=(W/L)(1/t_{ox})V^2$)	$1/\alpha$	α
transconductance (g_m)	1	α
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	$1/\alpha$	$1/\alpha$
load capacitance ($C=WL/t_{ox}$)	$1/\alpha$	$1/\alpha$
gate delay (VC/I)	$1/\alpha$	$1/\alpha^2$
dc & dynamic power dissipation (Ps)	$1/\alpha^2$	α
power-delay product	$1/\alpha^3$	$1/\alpha$
gate area	$1/\alpha^2$	$1/\alpha^2$
power density	1	α^3
current density	α	α^3

Outline

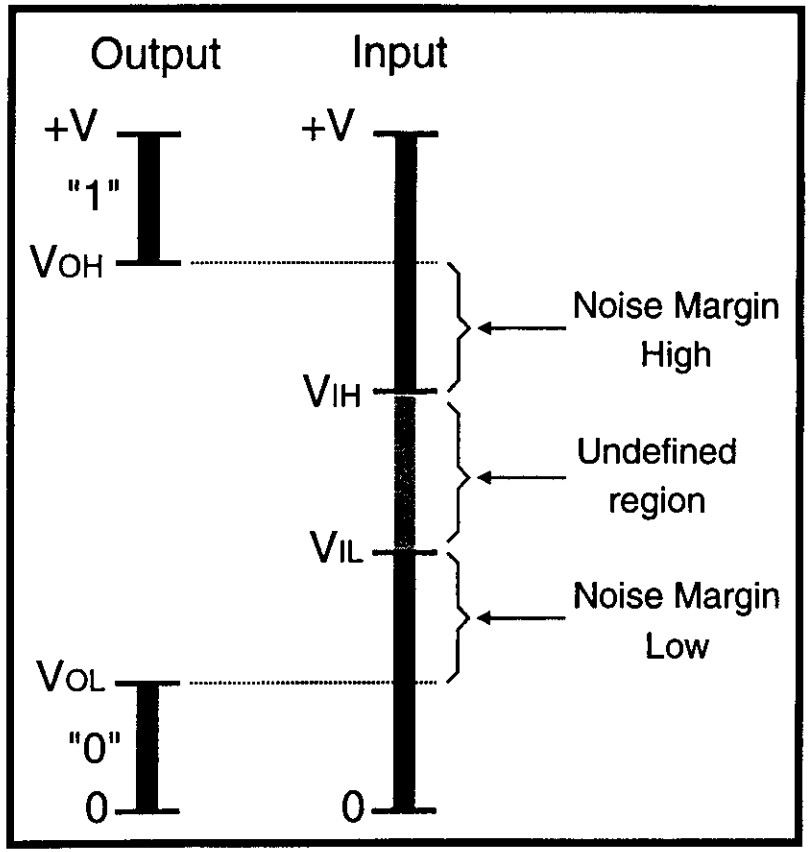
- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

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”

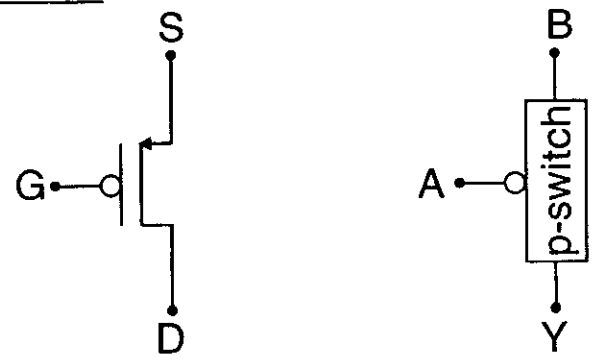
- Logic circuits process Boolean variables
- Logic values are associated with voltage levels:
  - $V_{IN} > V_{IH} \Rightarrow$ “0”
  - $V_{IN} < V_{IL} \Rightarrow$ “0”
- Noise margin:
  - $NM_H = V_{OH} - V_{IH}$
  - $NM_L = V_{IL} - V_{OL}$



9r

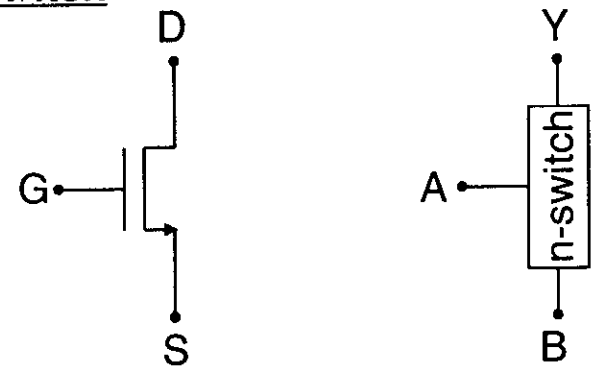
The MOST - a simple switch

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

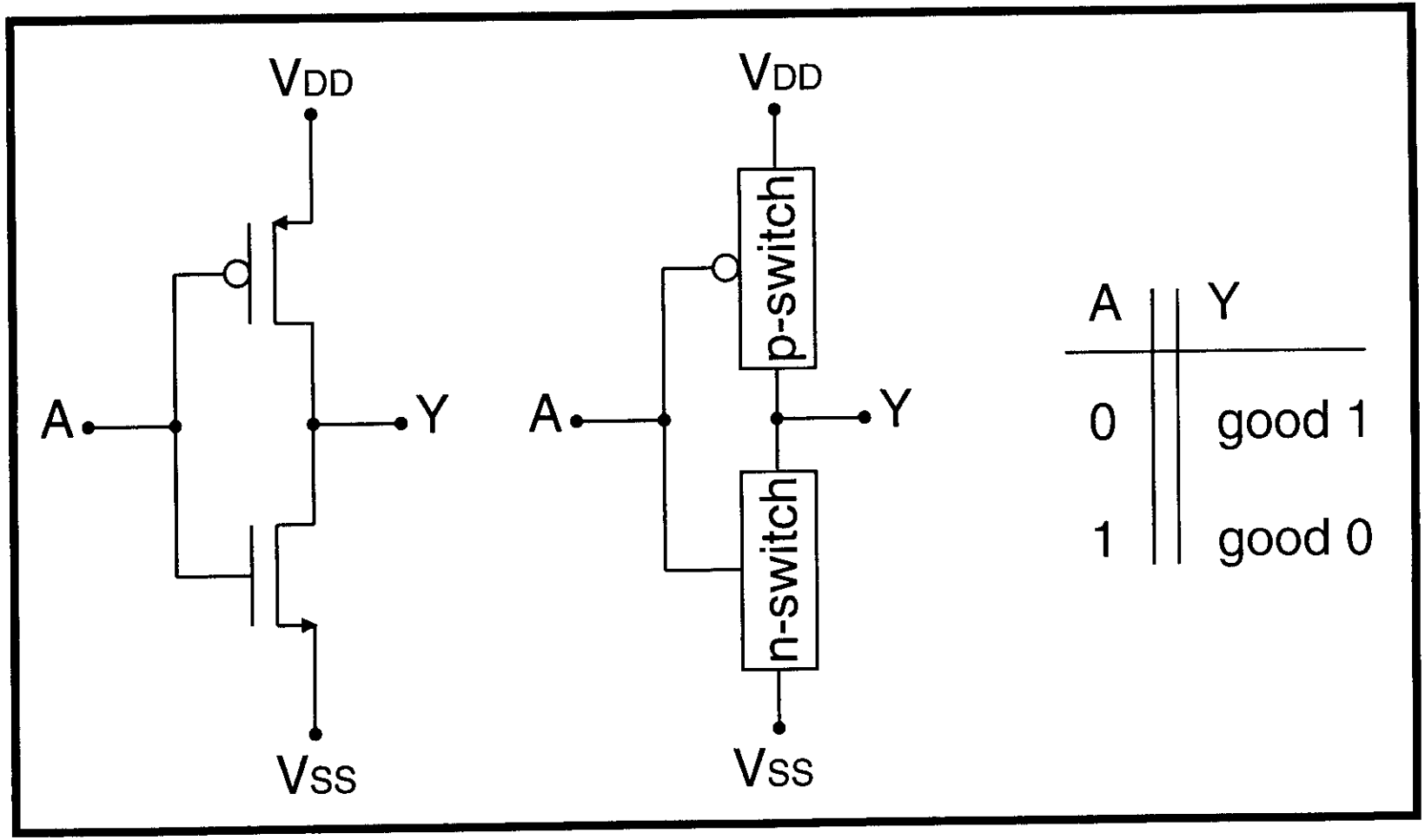


A	B	Y	
0	0	?	(high Z)
0	1	?	(high Z)
1	0	**good 0**	
1	1	bad 1	(source follower)

MOSFET's in digital design

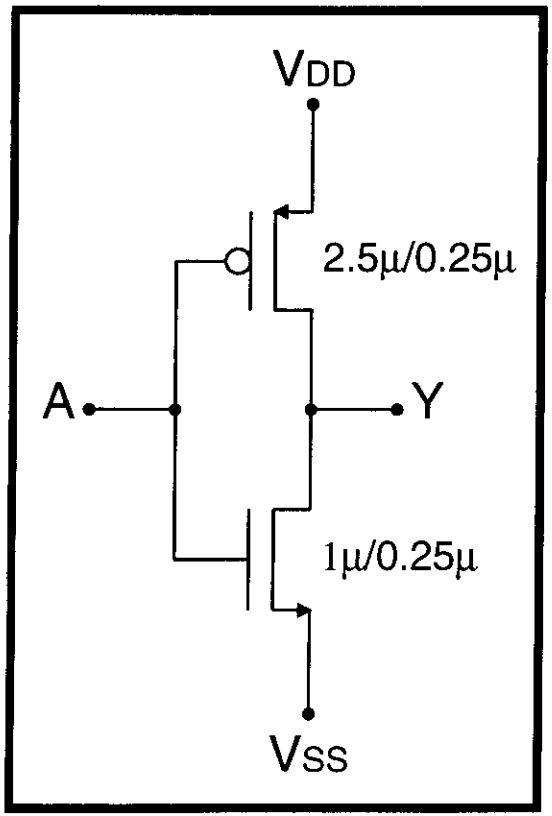
- Important characteristics:
  - It is an unipolar device
    - NMOS - charge carrier: electrons
    - PMOS - charge carrier: holes
  - It is a symmetrical device
    - Source = drain
  - High input impedance (Ig=0)
    - Low standby current in CMOS configuration
  - Voltage controlled device with high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

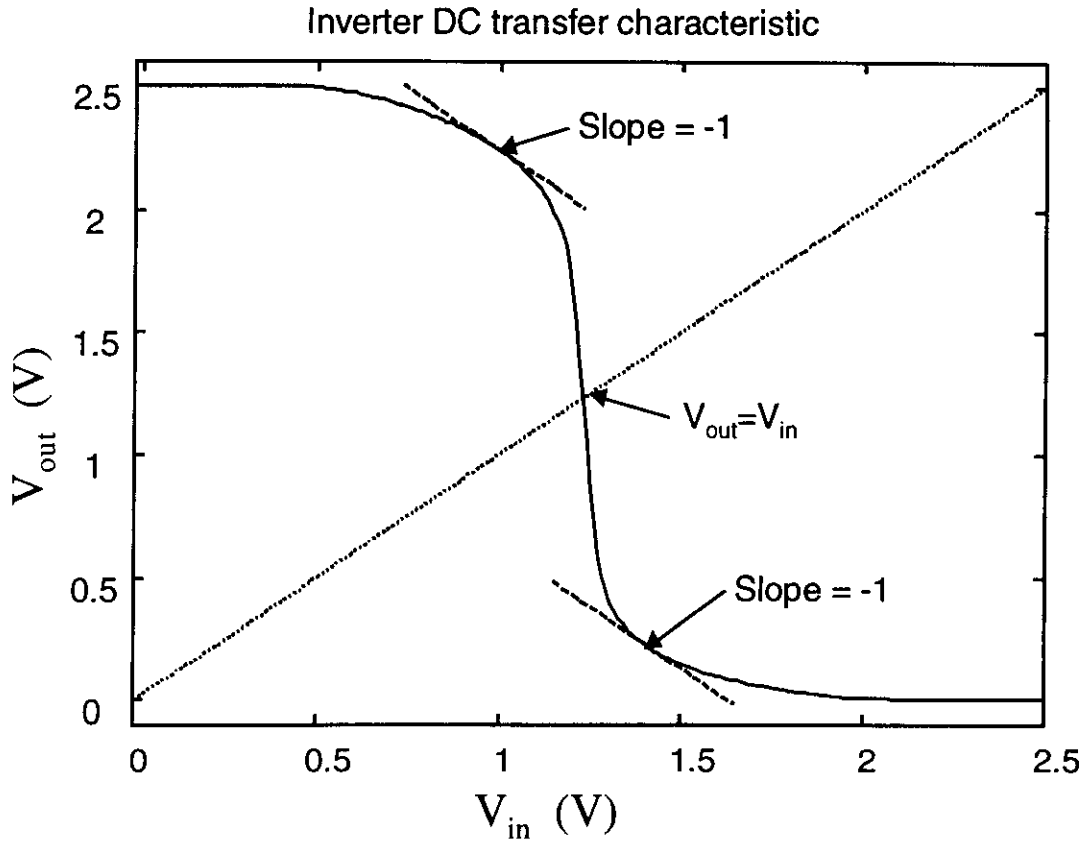
The CMOS inverter



A	Y
0	good 1
1	good 0

The CMOS inverter



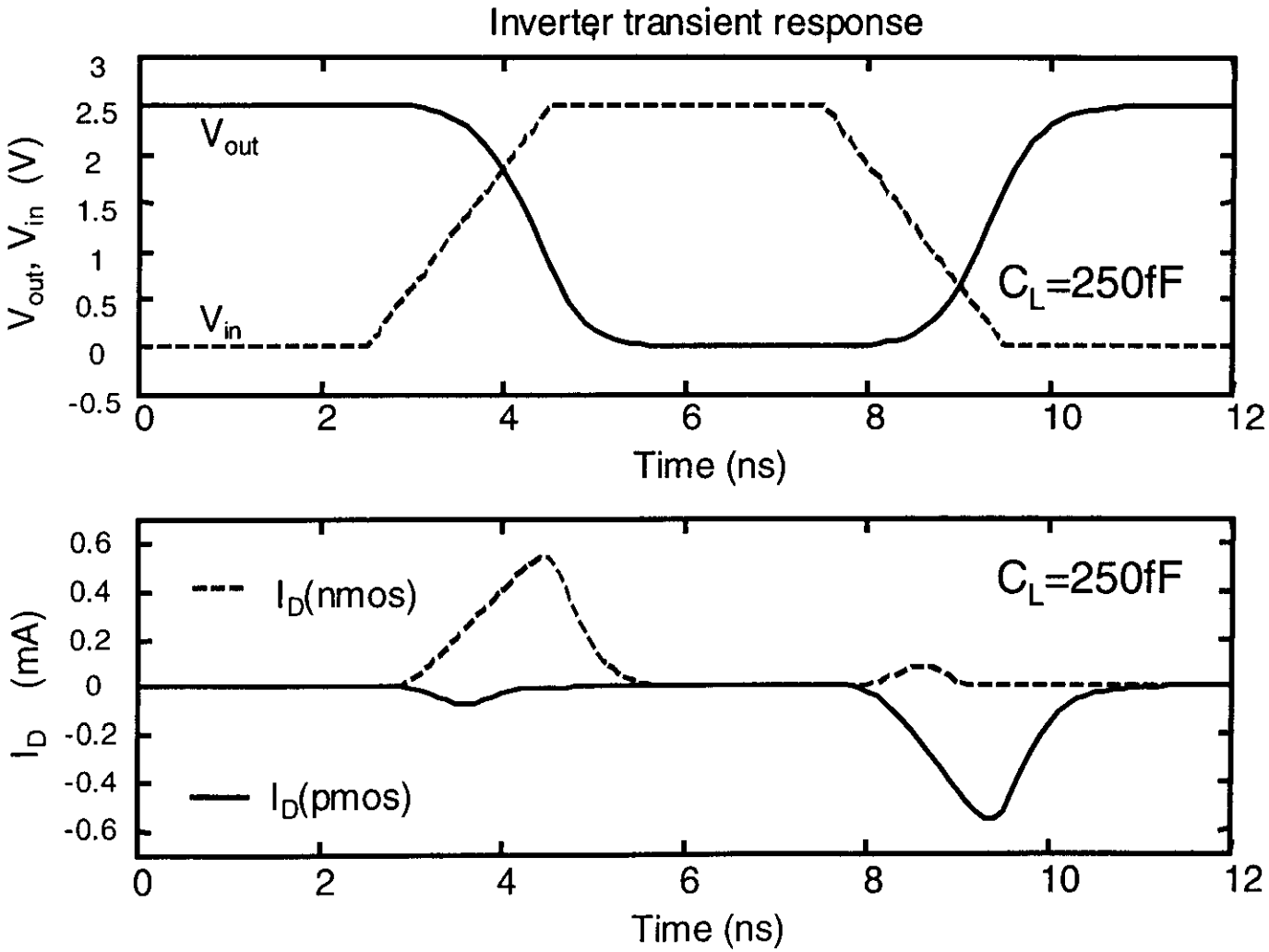


The CMOS inverter

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

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

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

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

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

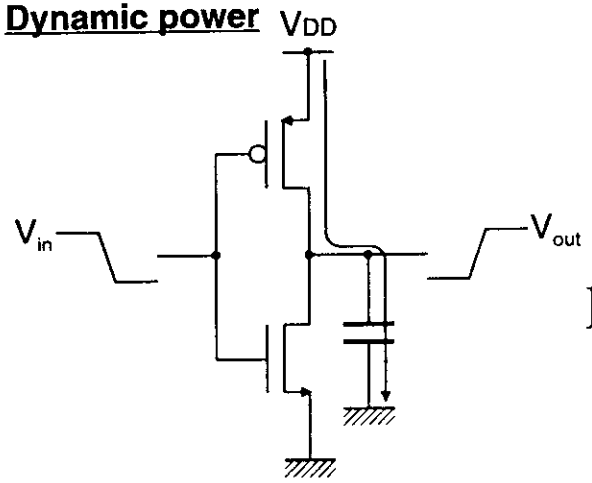
The CMOS inverter

- 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

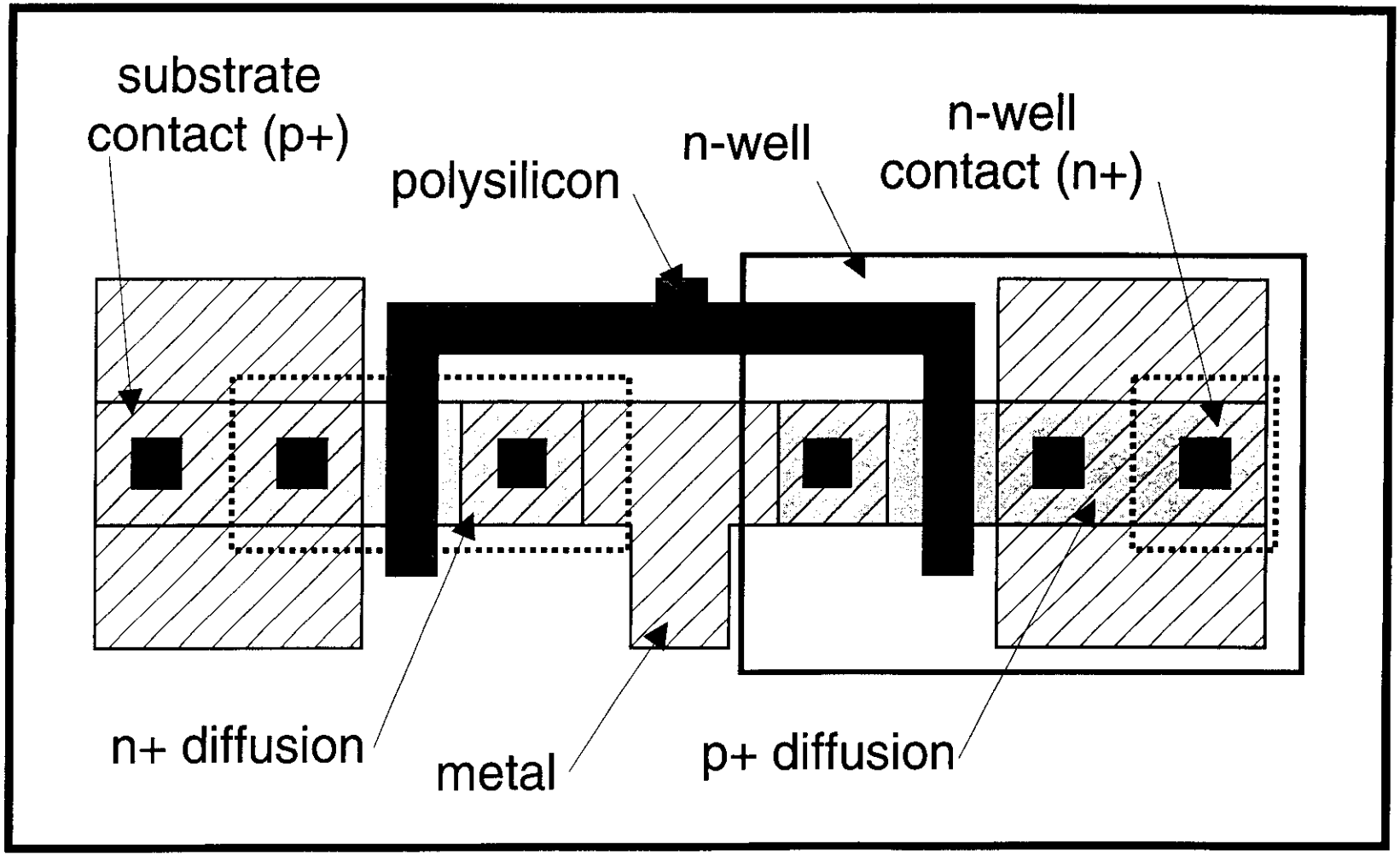
Dynamic power



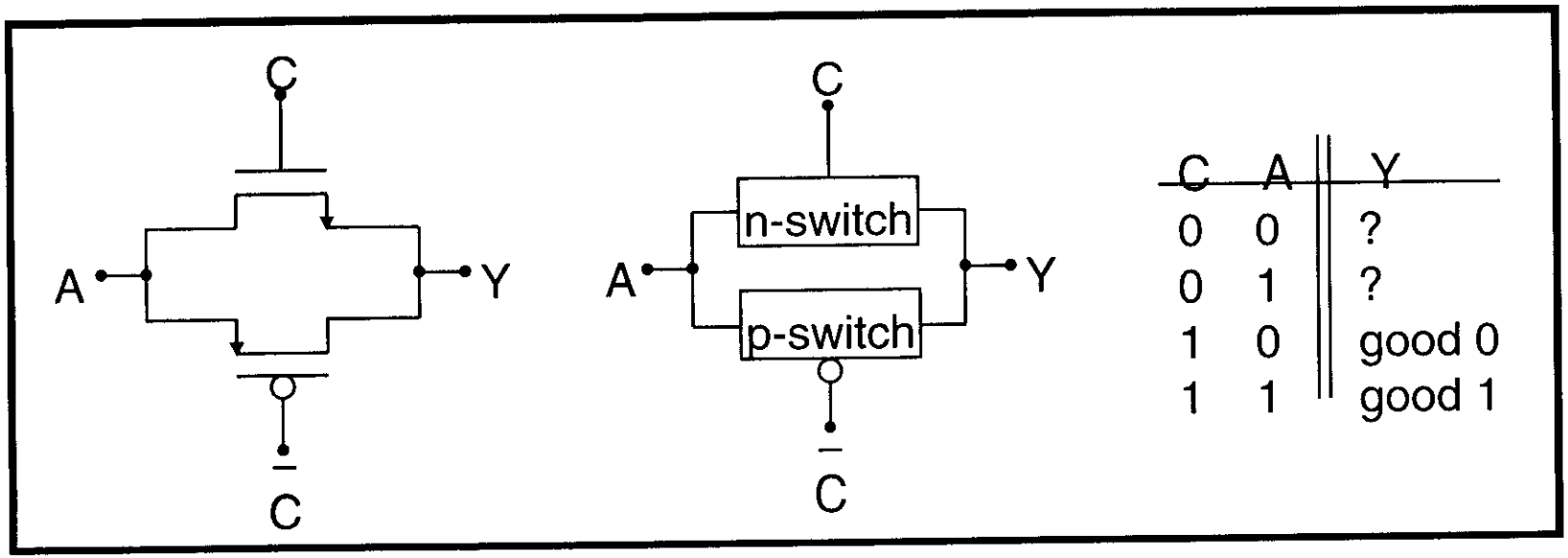
$$E = \text{Energy / transition} = \frac{1}{2} \cdot C_L \cdot V_{dd}^2$$

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

The CMOS inverter



The CMOS pass gate

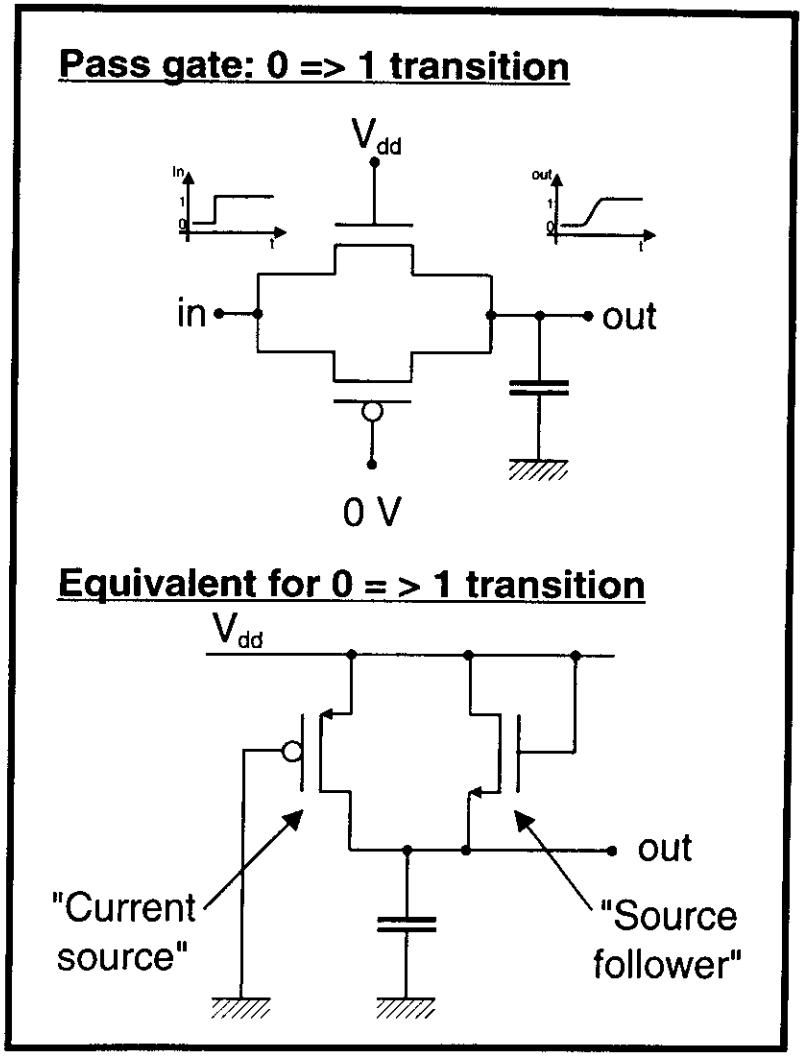


C	A	Y
0	0	?
0	1	?
1	0	good 0
1	1	good 1

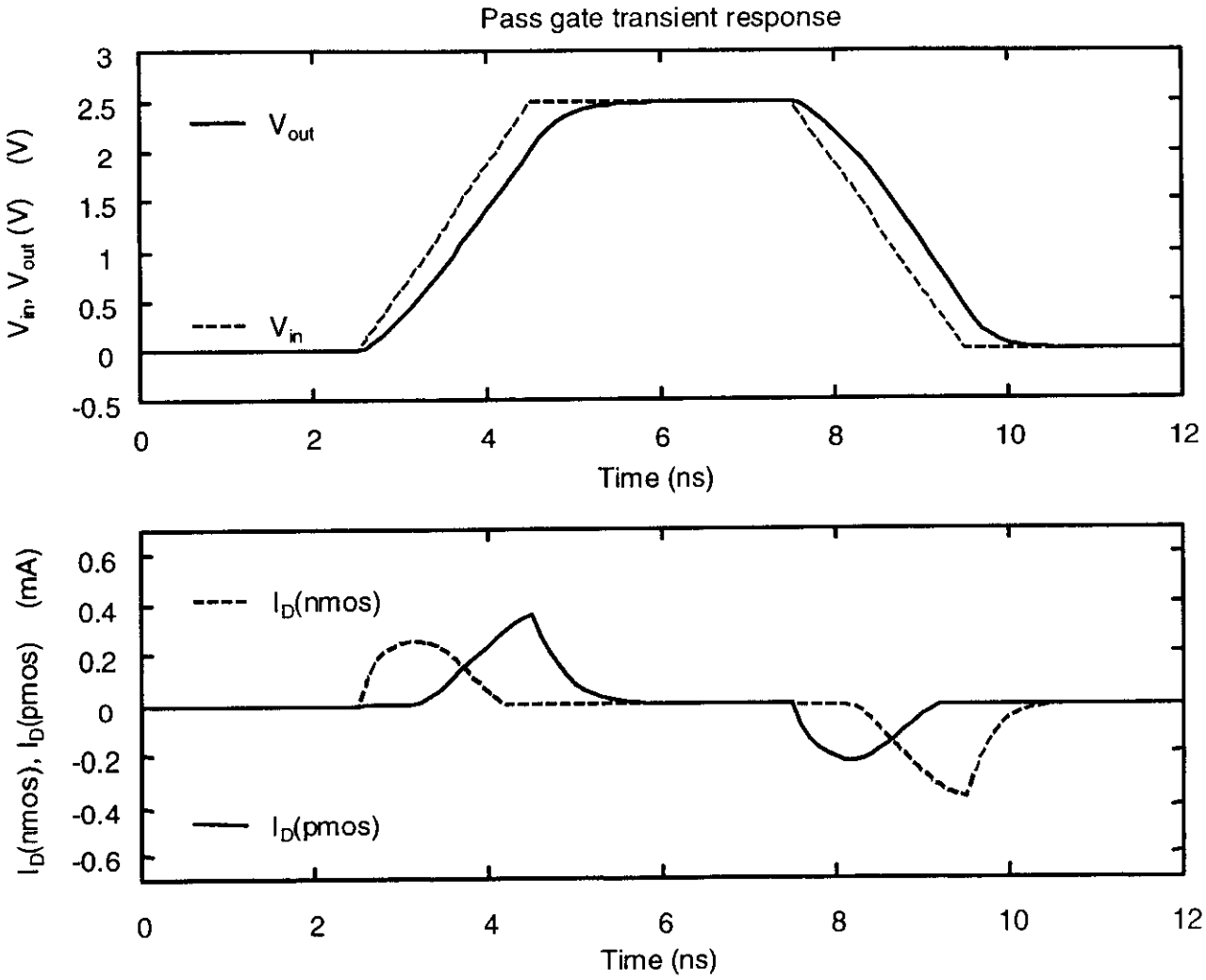
The CMOS pass gate

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

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

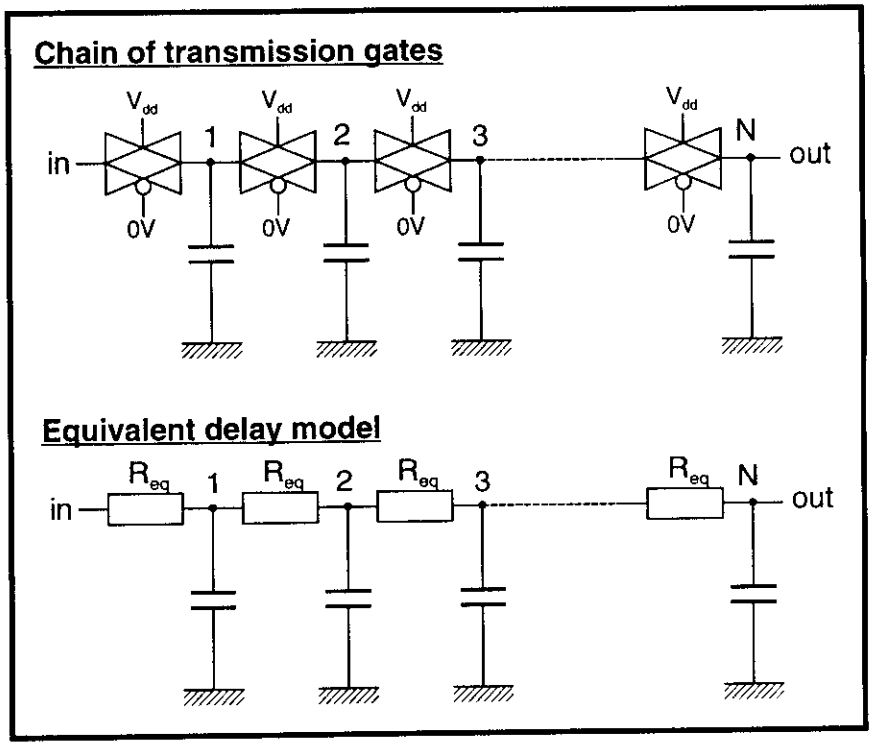
- Both devices combine to form a good switch

The CMOS pass gate

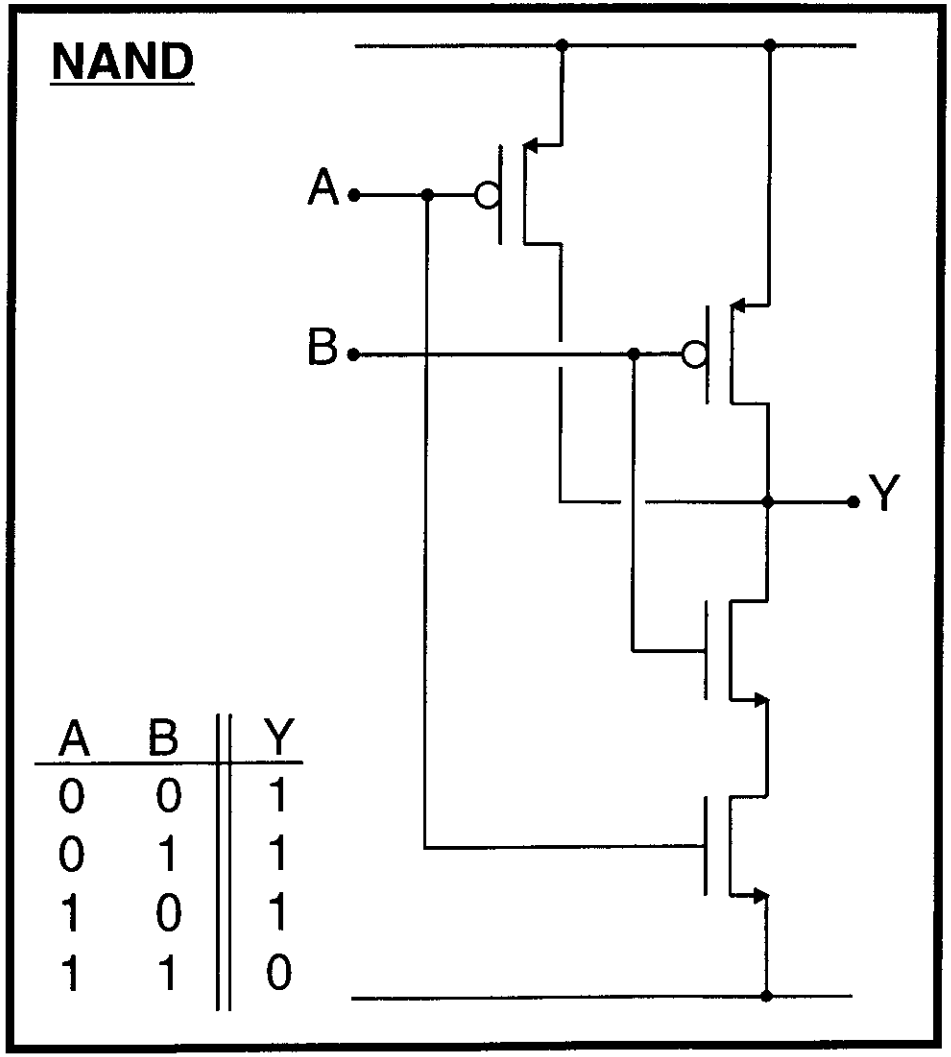
- Delay of a chain of pass gates:

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

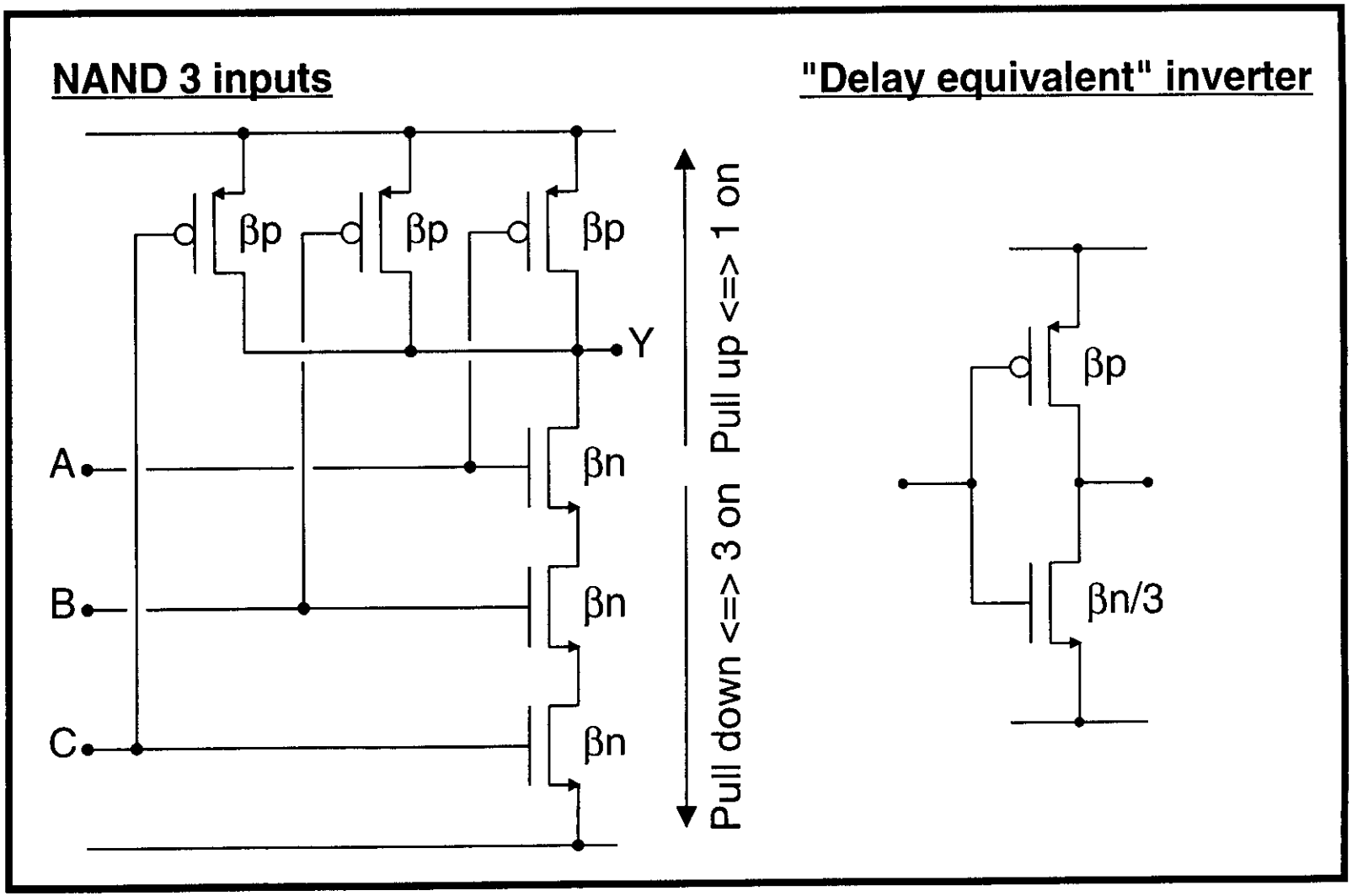
- Delay proportional to N^2
- Avoid N large:
  - Break the chain by inserting buffers



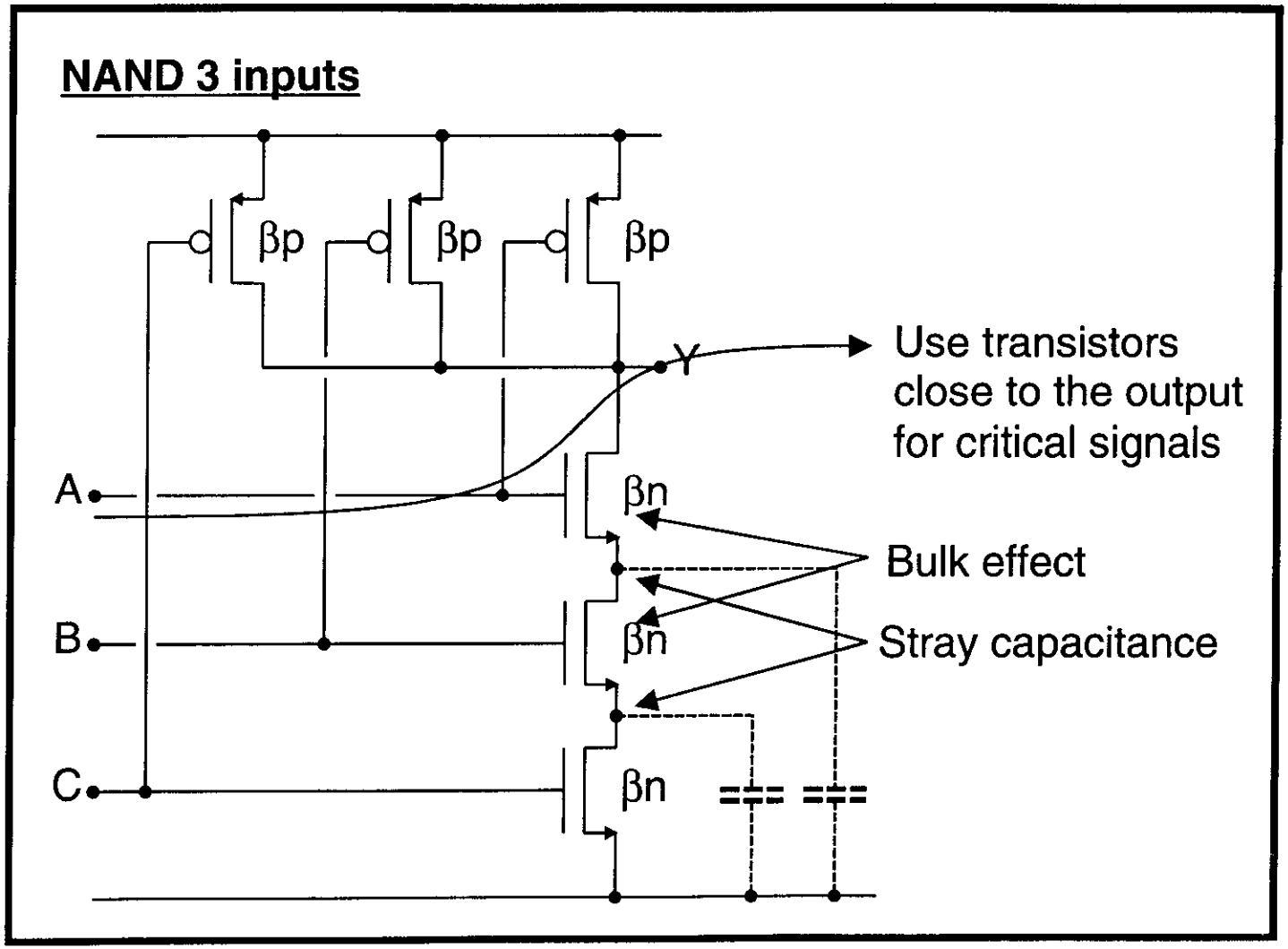
Simple CMOS gates



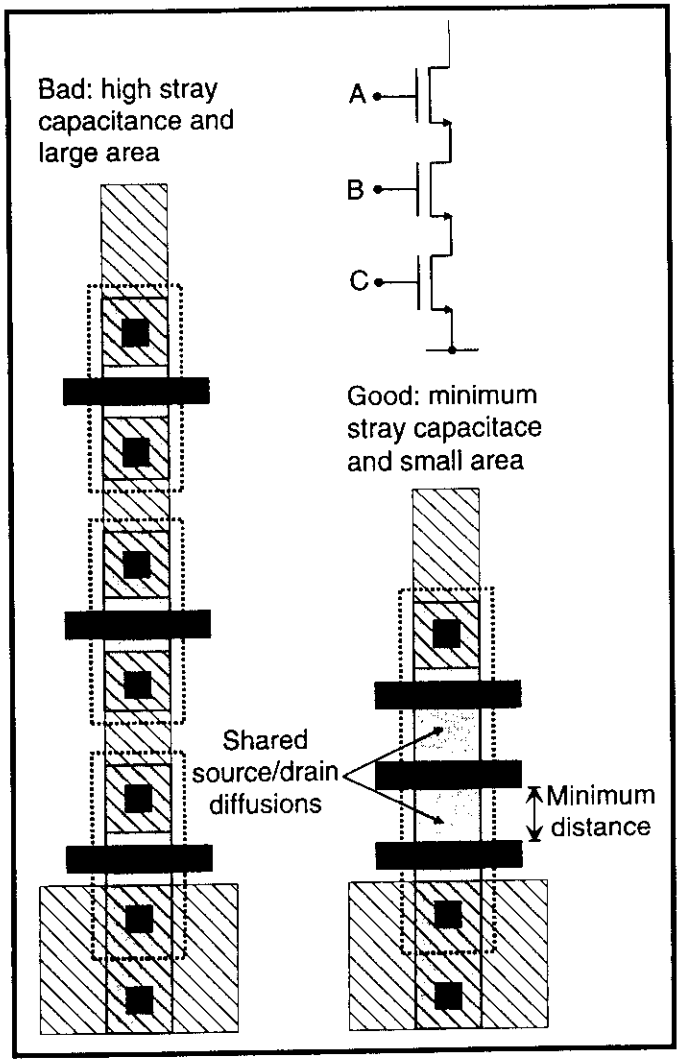
Simple CMOS gates



Simple CMOS gates



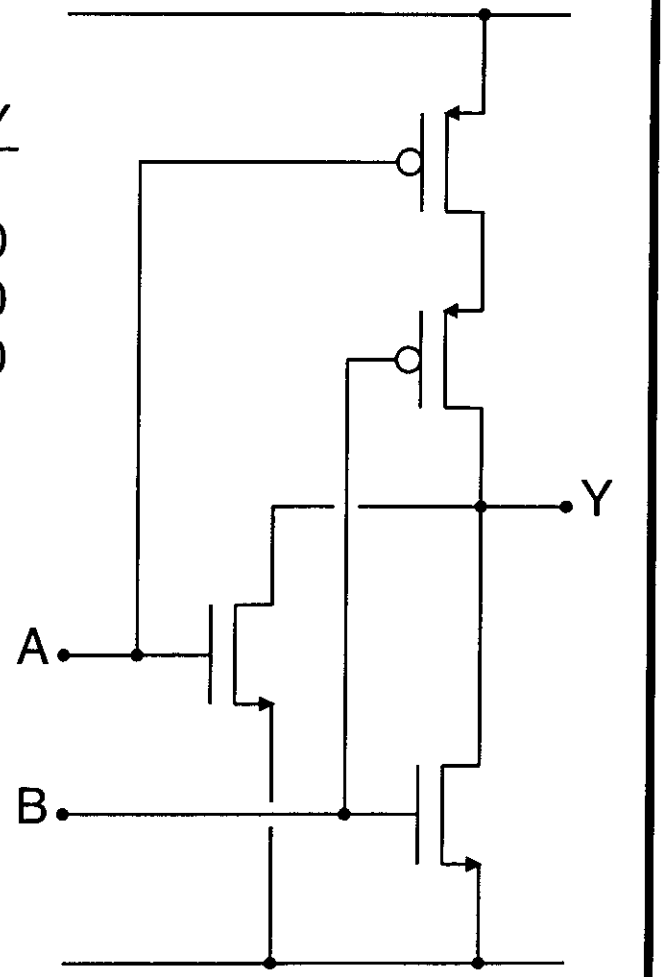
Simple CMOS gates



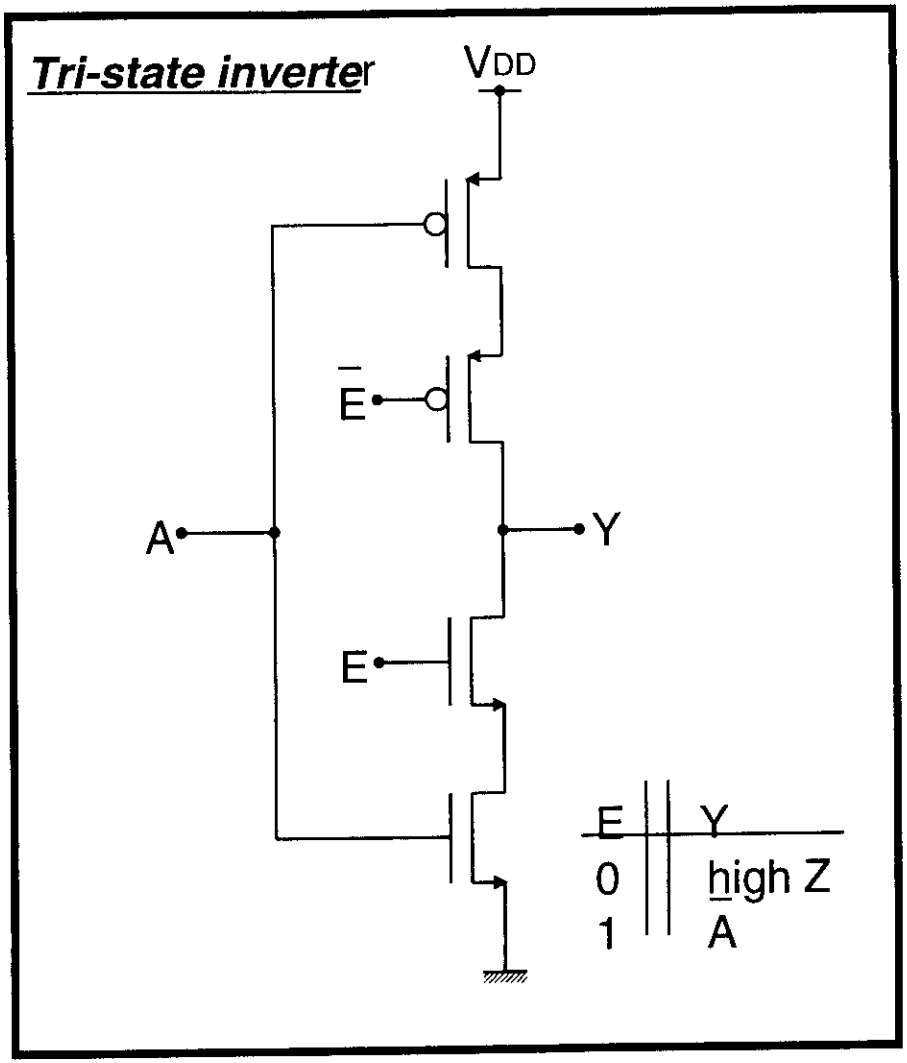
Simple CMOS gates

NOR

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

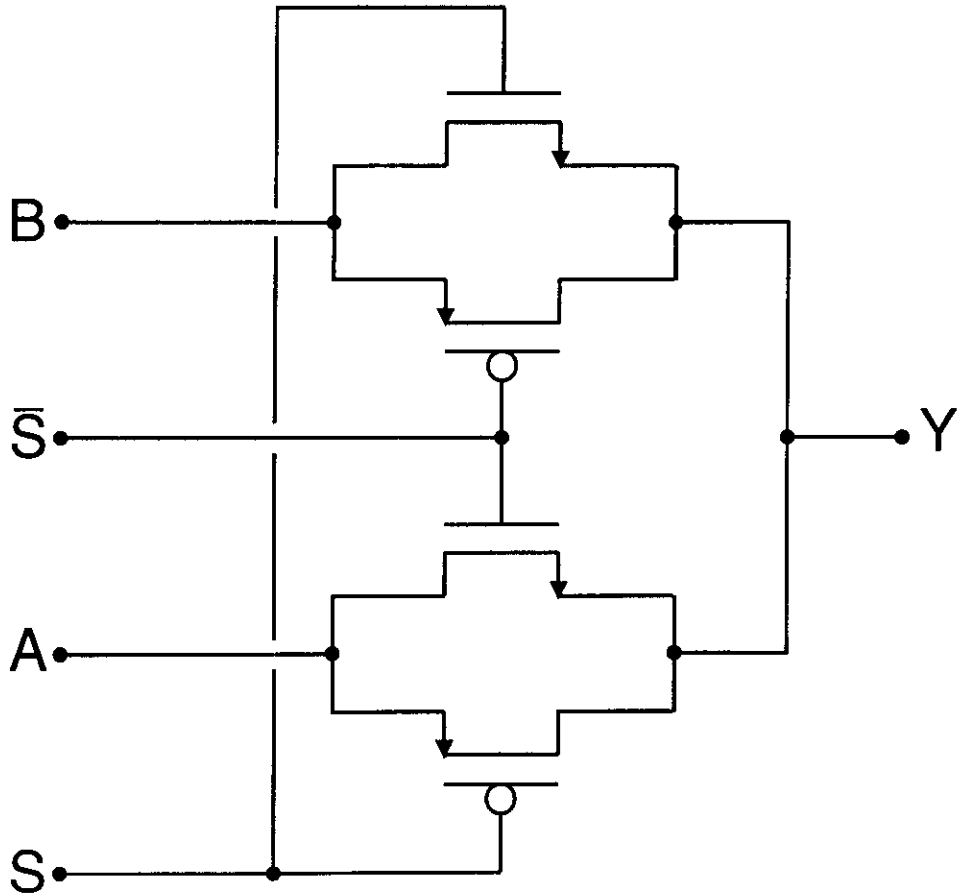


Simple CMOS gates



Complex CMOS gates

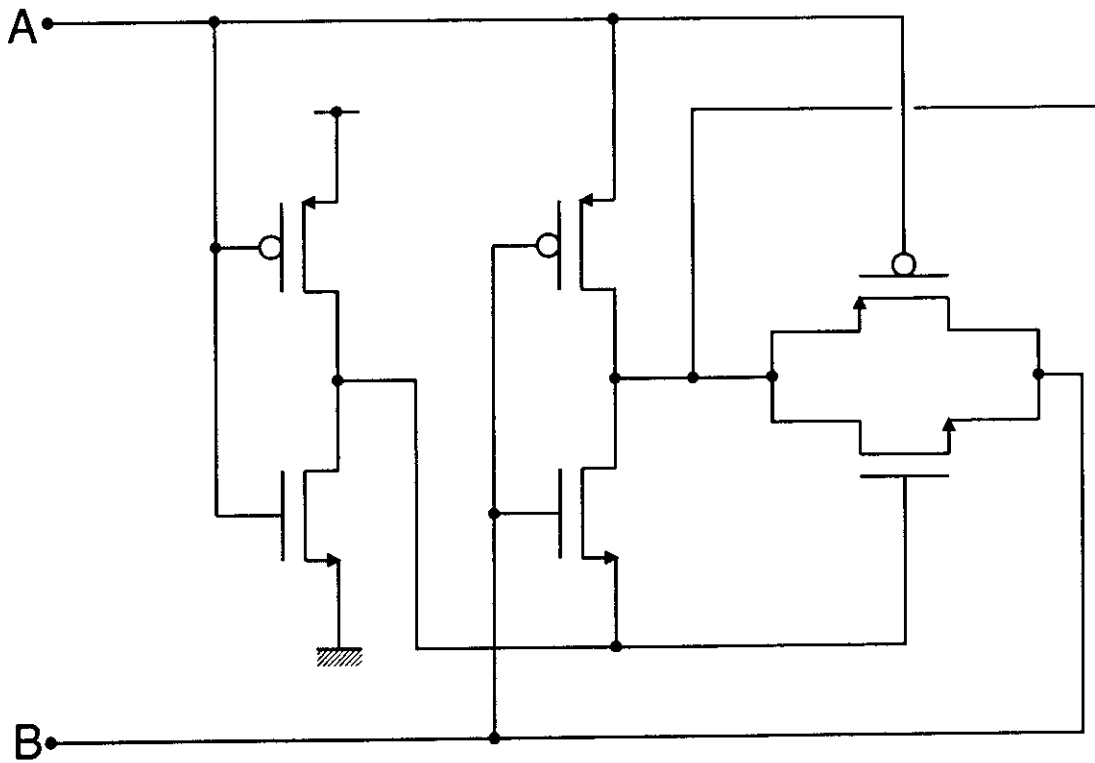
Multiplexer



S	Y
0	A
1	B

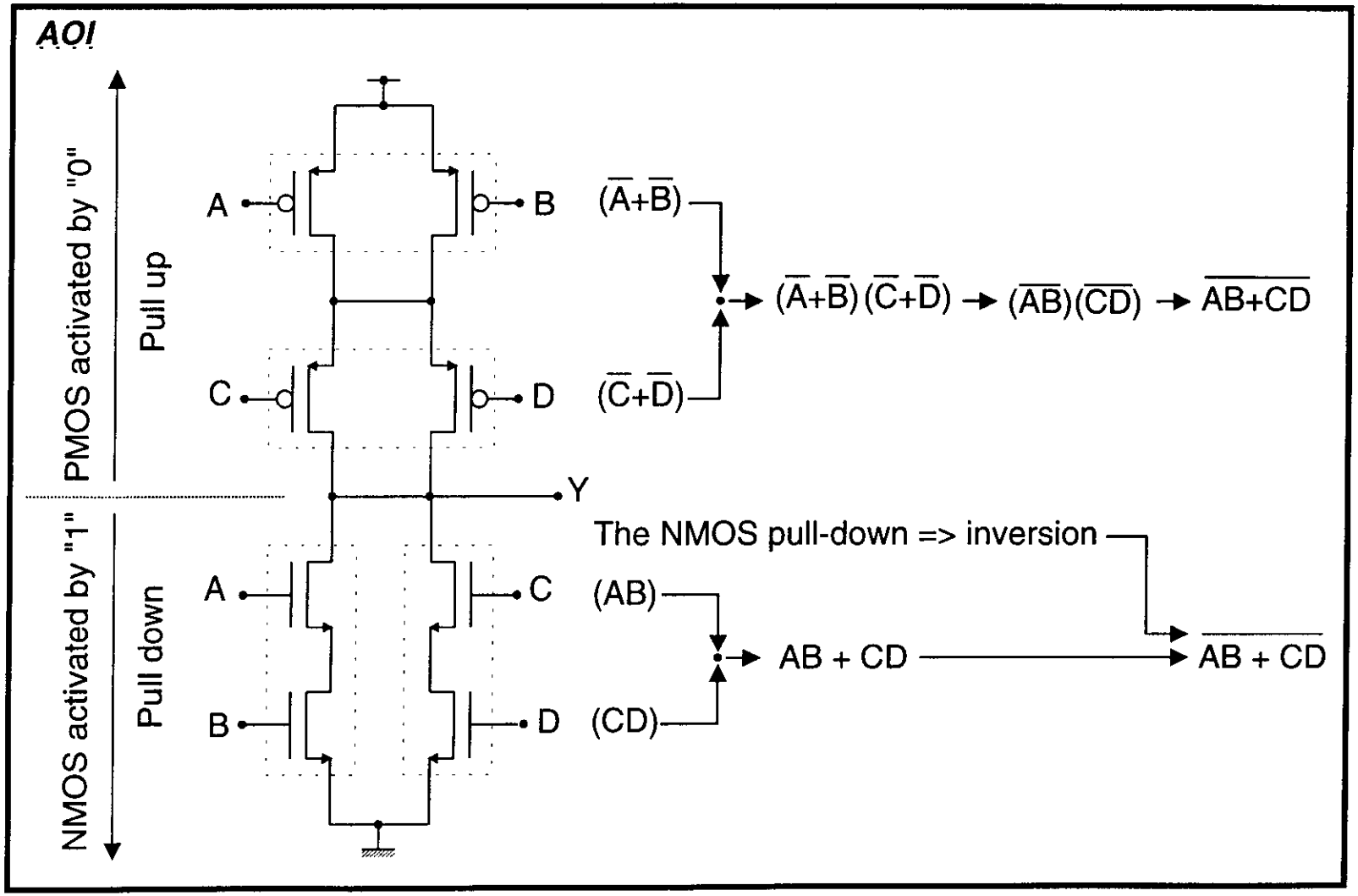
Complex CMOS gates

Exclusive OR

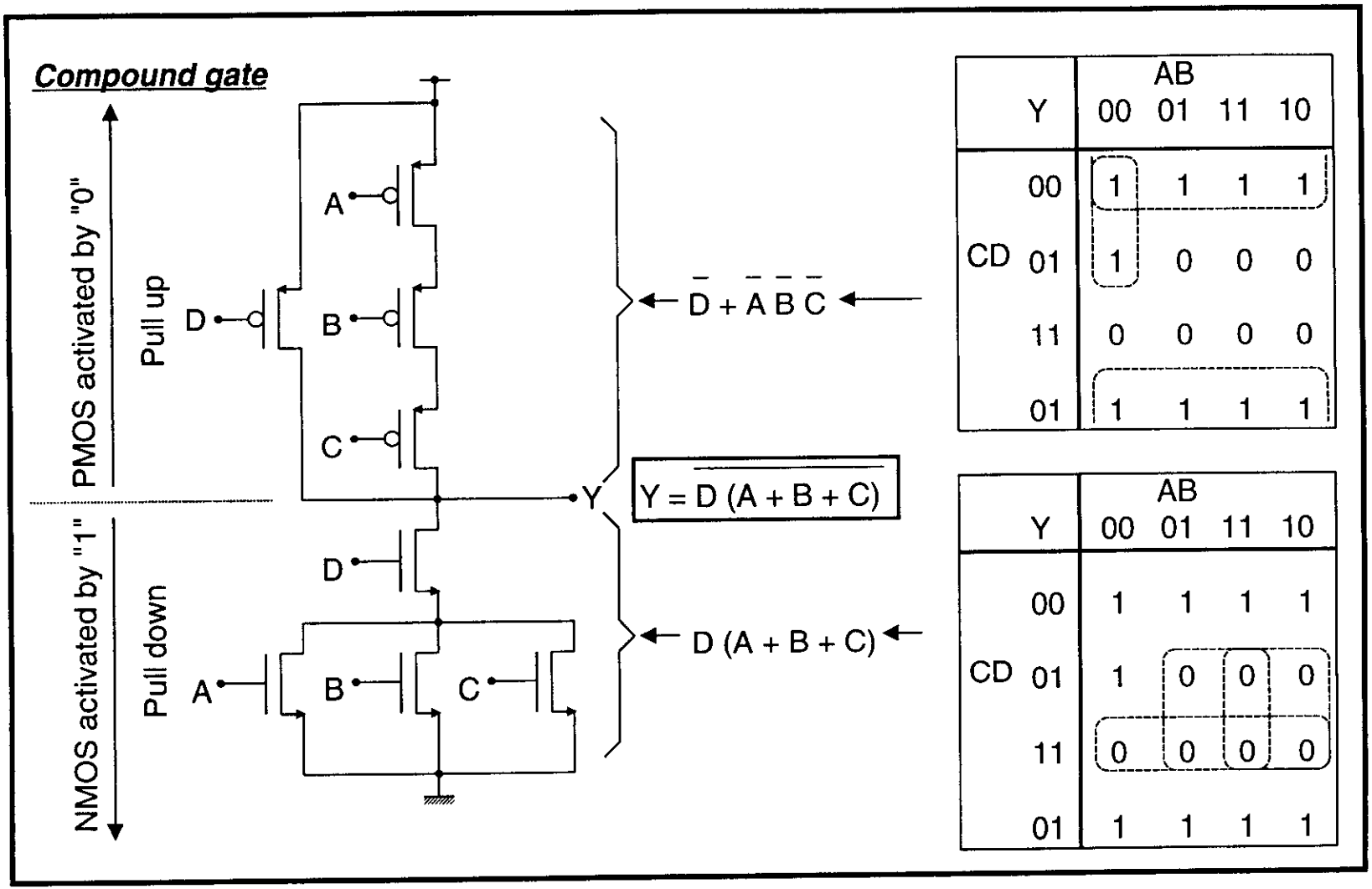


A	B	Y
0	0	0
0	1	1
1	0	1
1	1	0

Complex CMOS gates



Complex CMOS gates



Y	CD \ AB	00	01	11	10
	00	1	1	1	1
	01	1	0	0	0
	11	0	0	0	0
	01	1	1	1	1

Y	CD \ AB	00	01	11	10
	00	1	1	1	1
	01	1	0	0	0
	11	0	0	0	0
	01	1	1	1	1

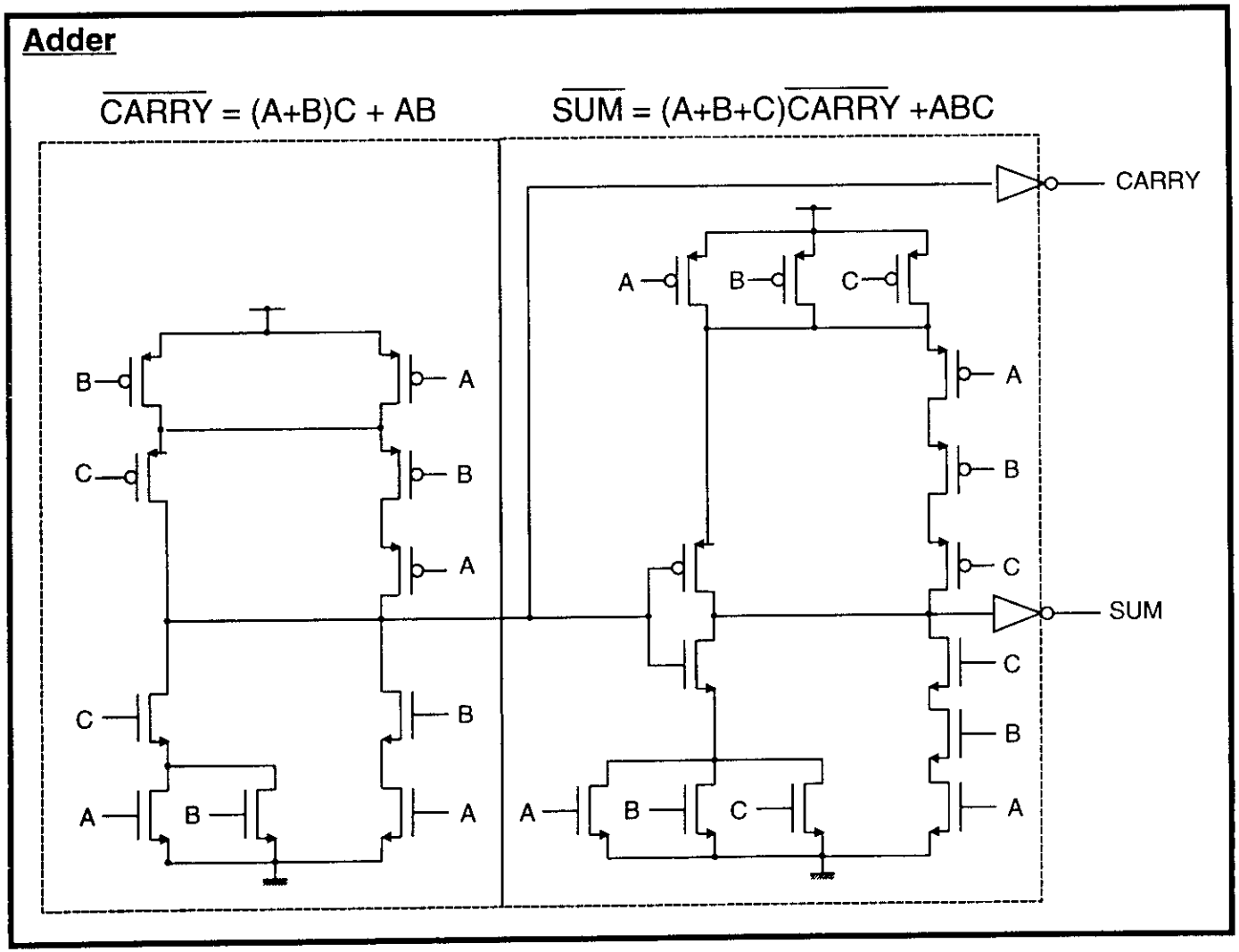
Complex CMOS gates

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

Complex CMOS gates



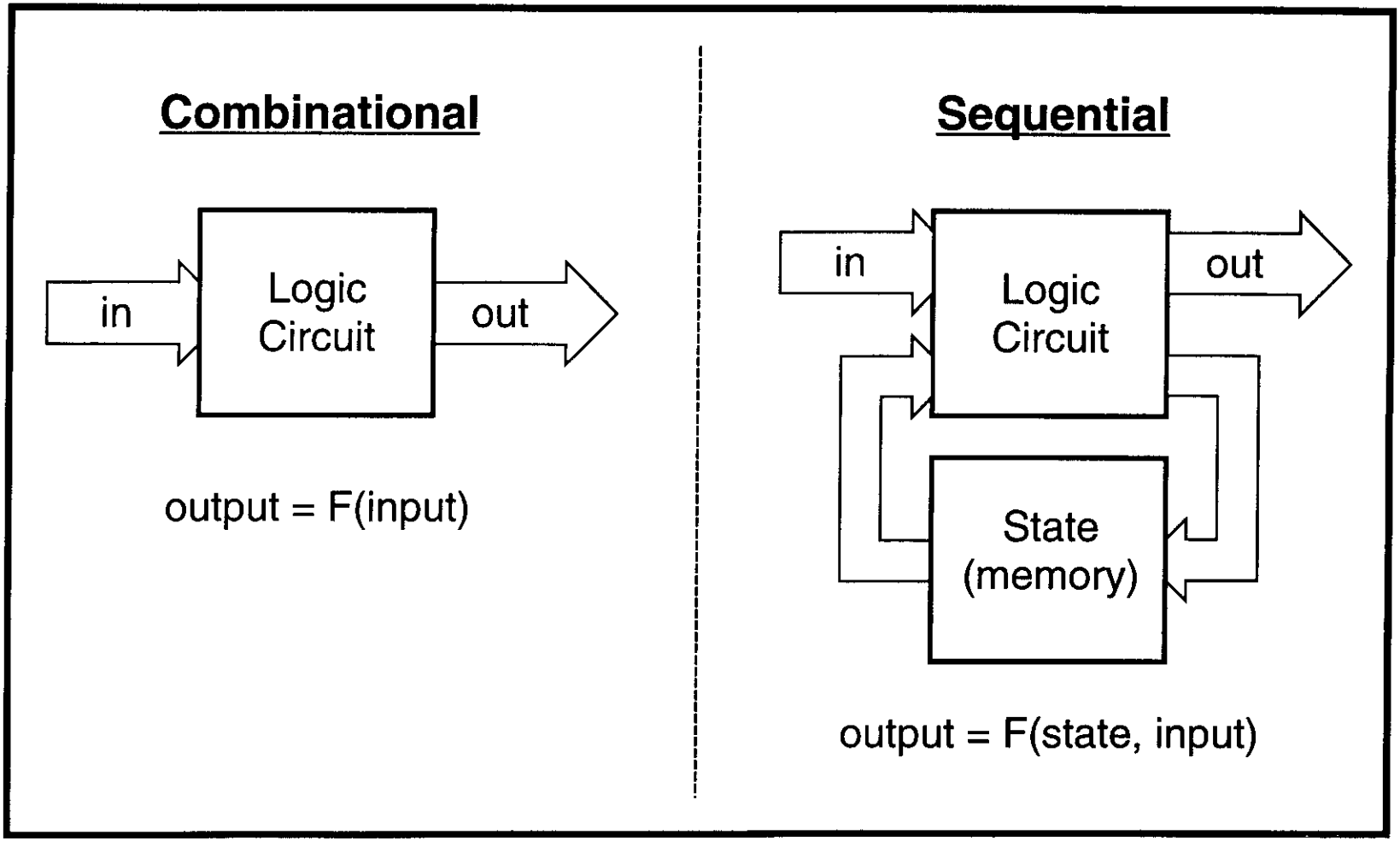
Outline

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

CMOS sequential circuits

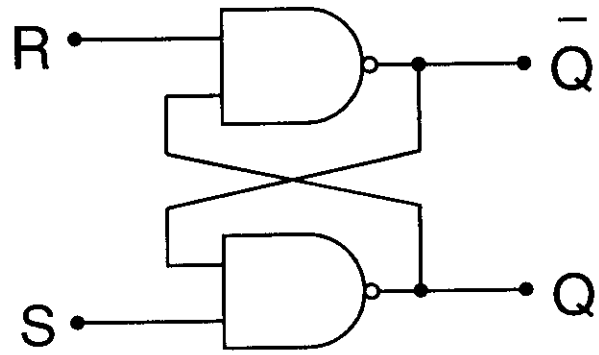
- Sequential circuits
- Interconnects
- Clock distribution
- DLL's and PLL's

Sequential circuits



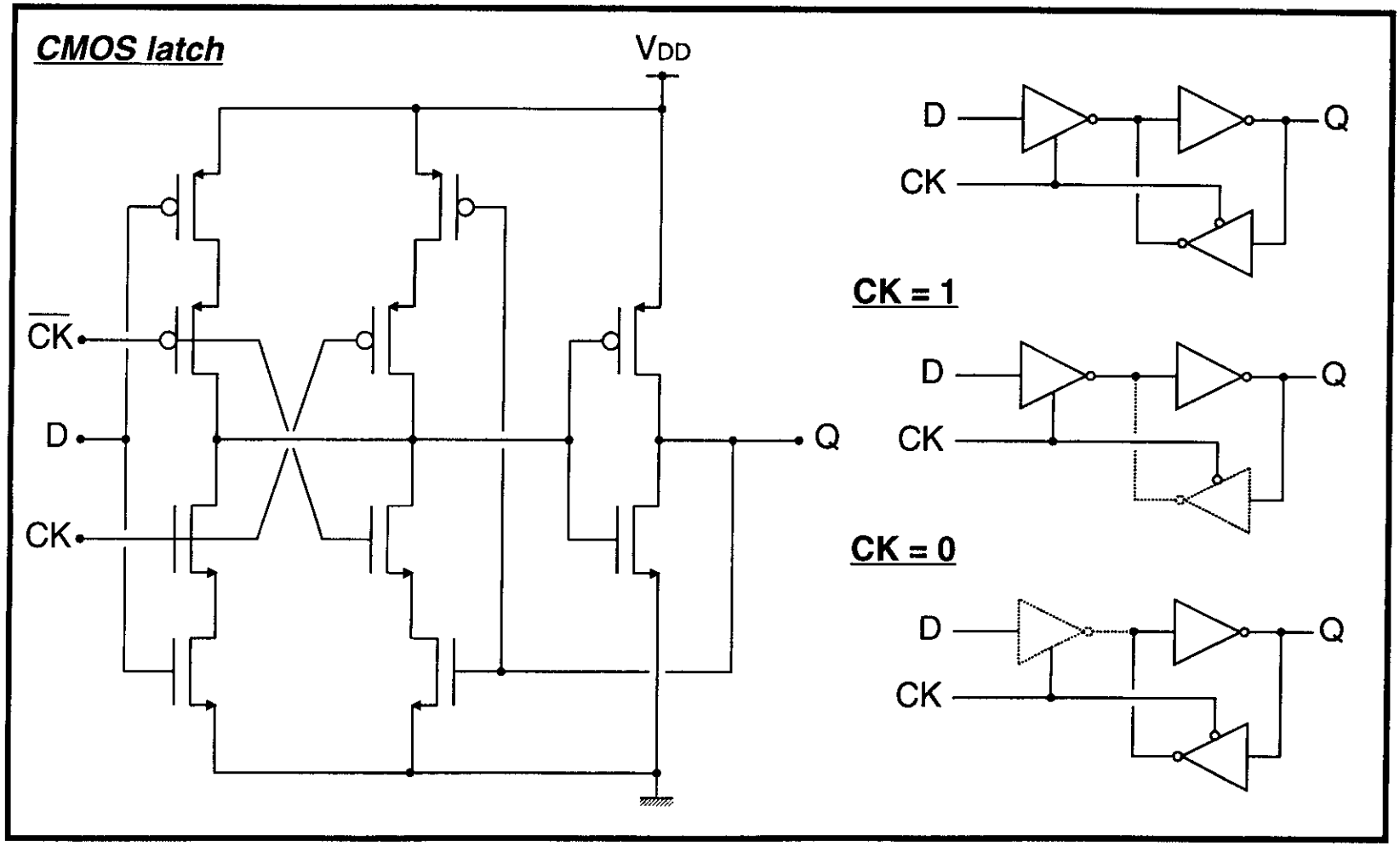
Sequential circuits

RS latch

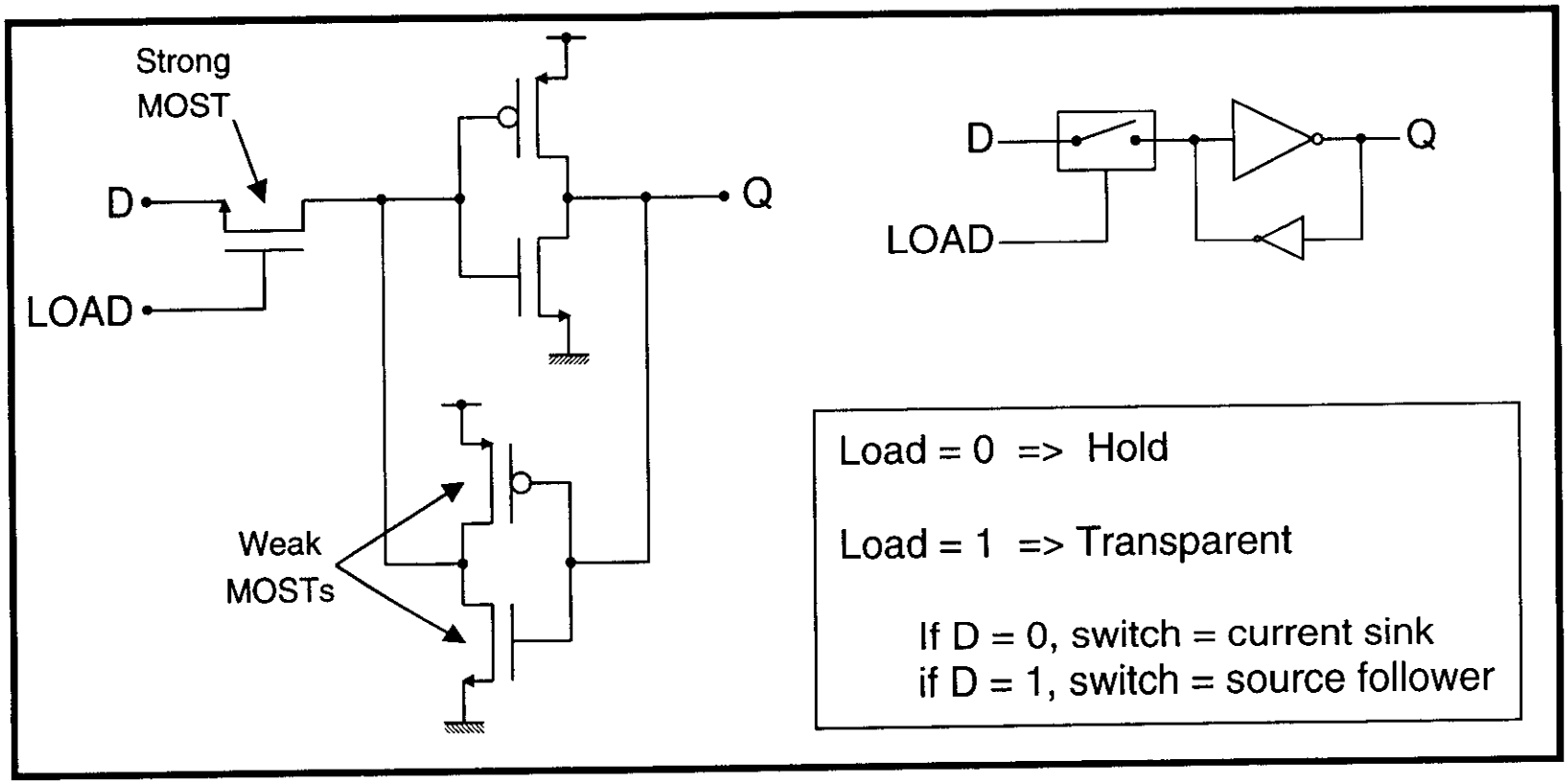


R	S	Q_{n+1}	$\bar{Q}_{n+1}$	
0	0	1	1	(ilegal)
0	1	0	1	
1	0	1	0	
1	1	Q_n	$\bar{Q}_n$	(memory)

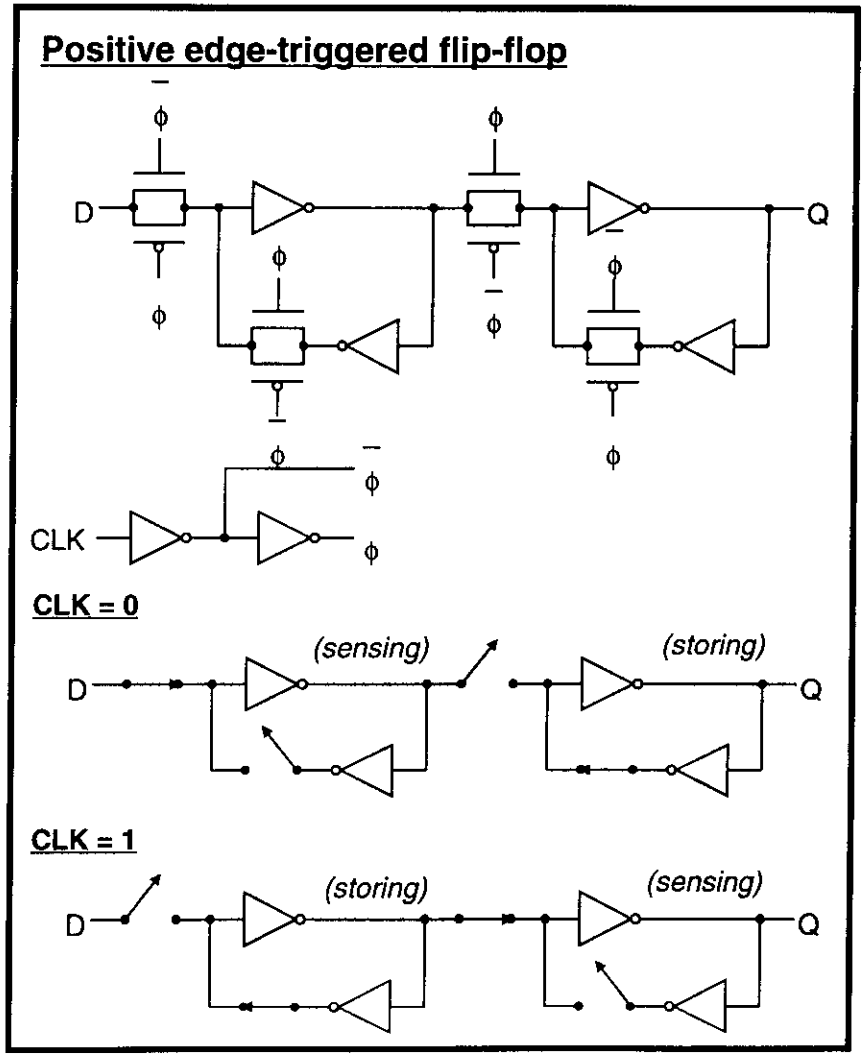
Sequential circuits



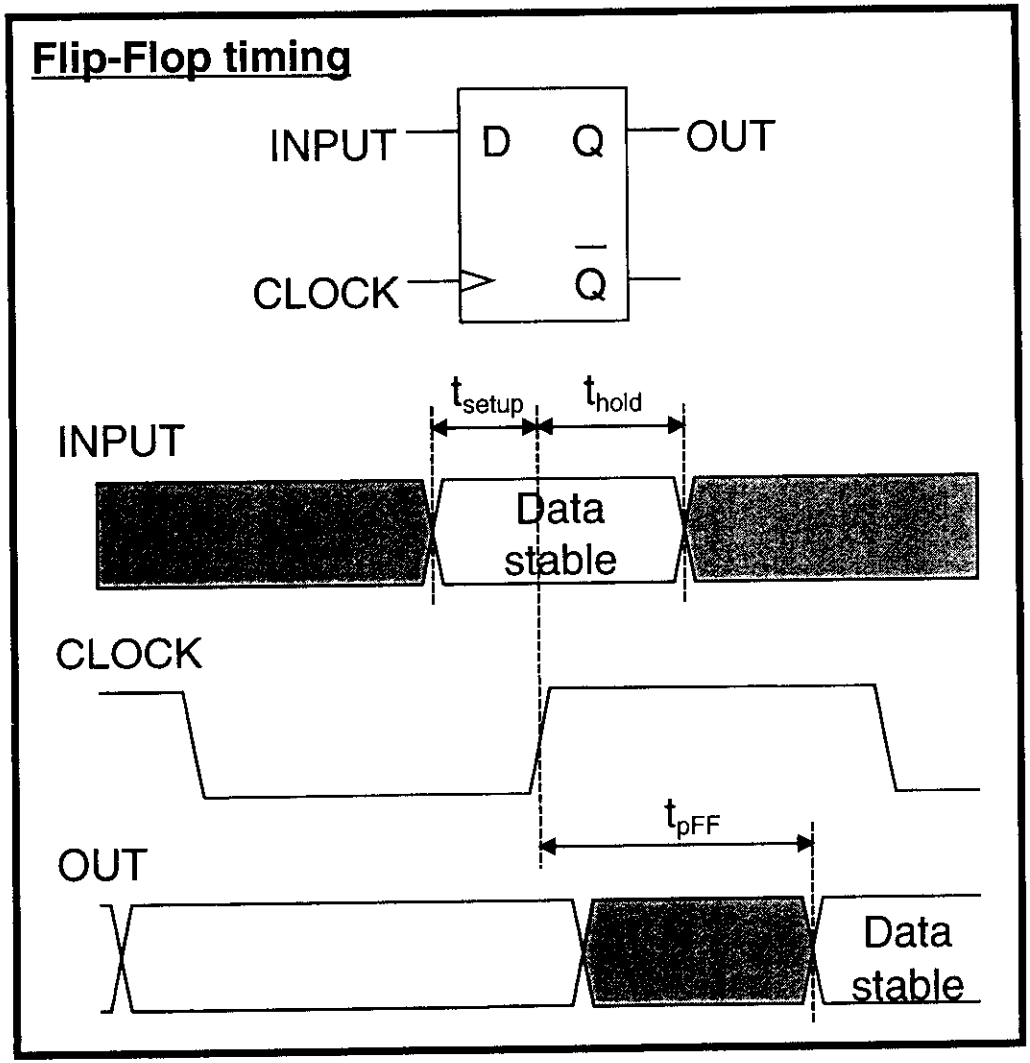
Sequential circuits



Sequential circuits

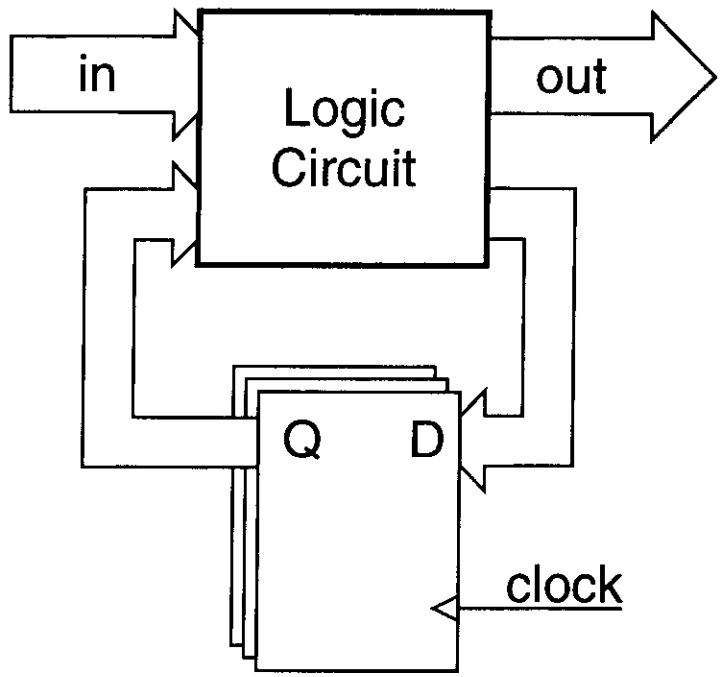


Sequential circuits



Sequential circuits

Maximum clock frequency



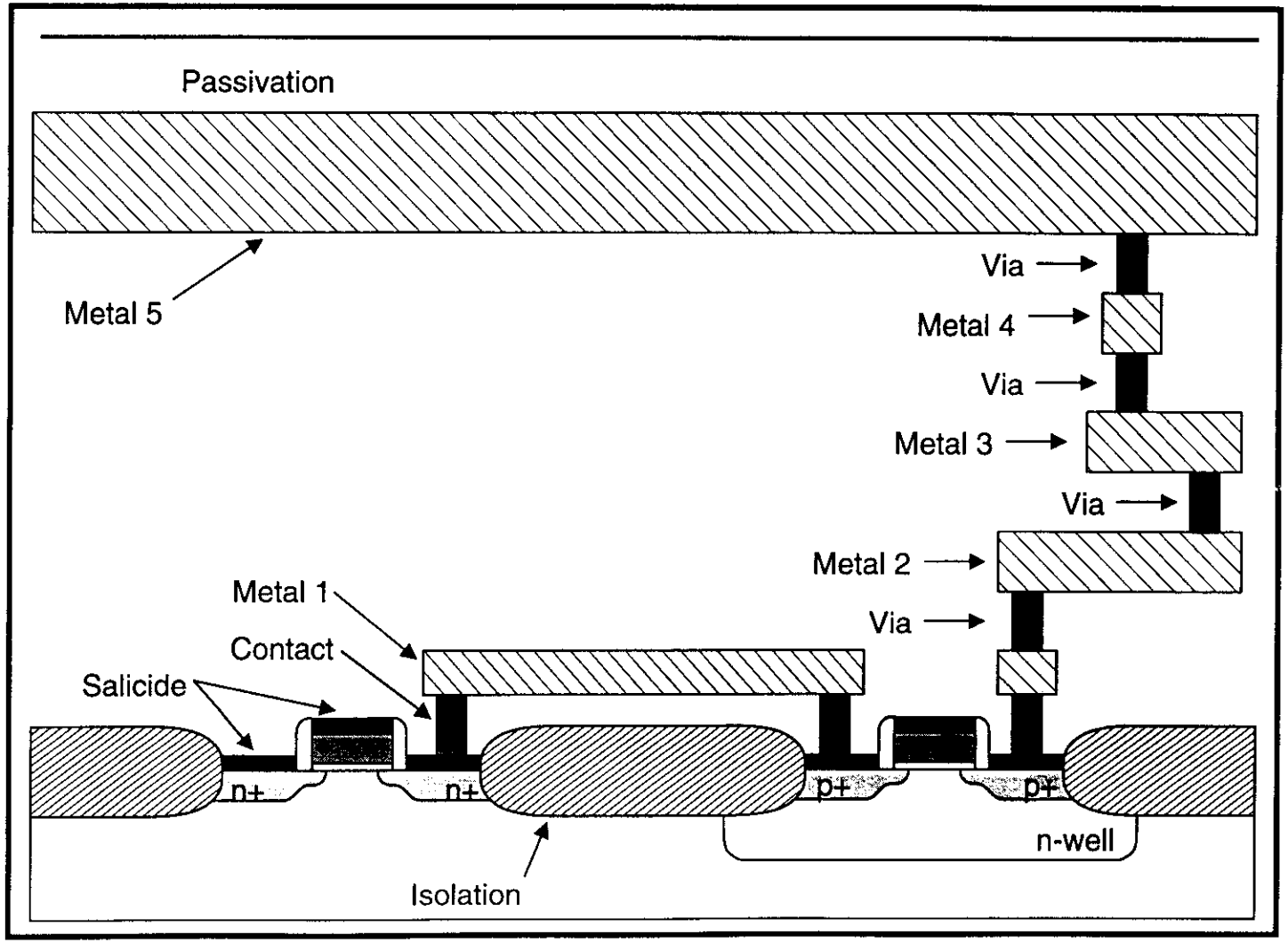
$$T_{min} > t_{pFF} + t_{p,comb} + t_{setup}$$

$$f_{max} = 1/T_{min}$$

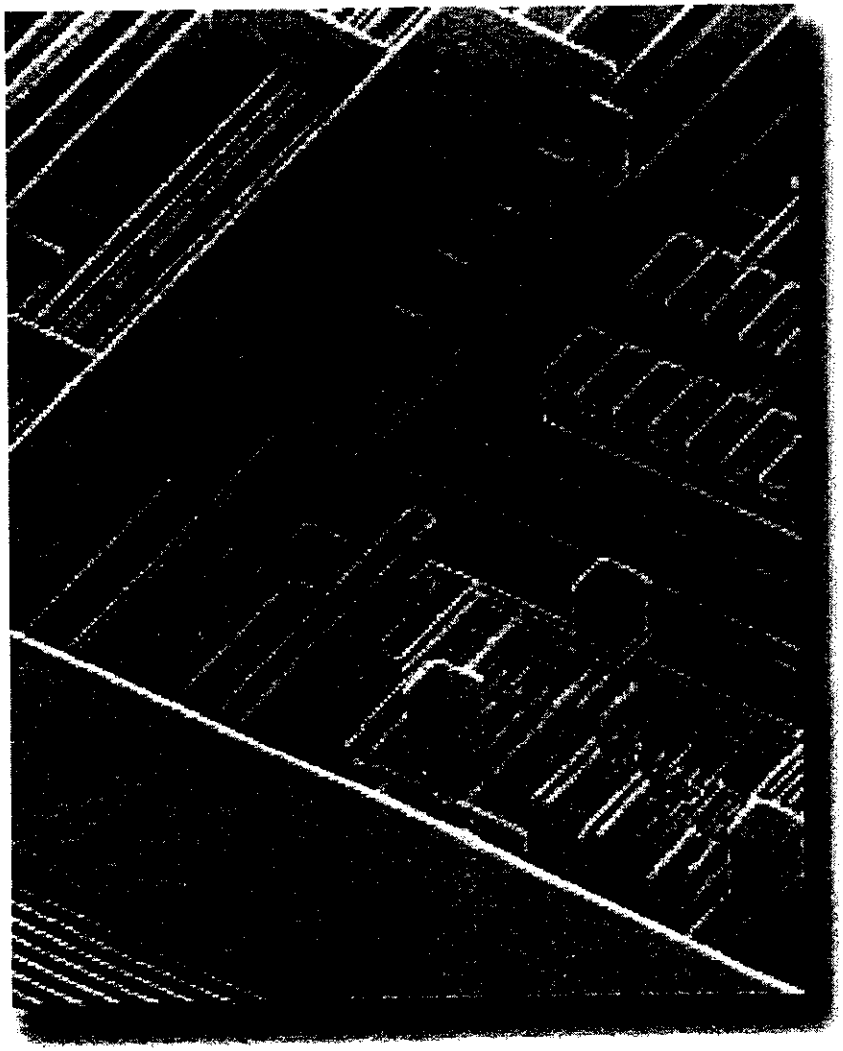
Interconnects

- 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

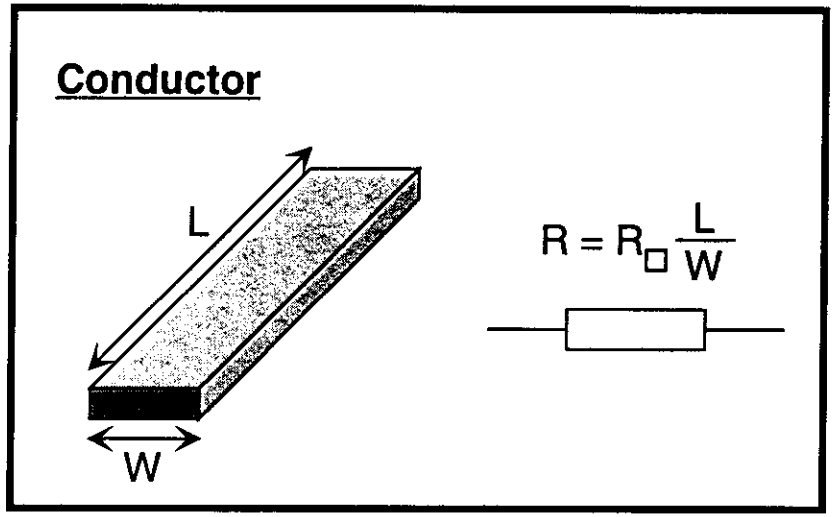
Interconnects



Interconnects



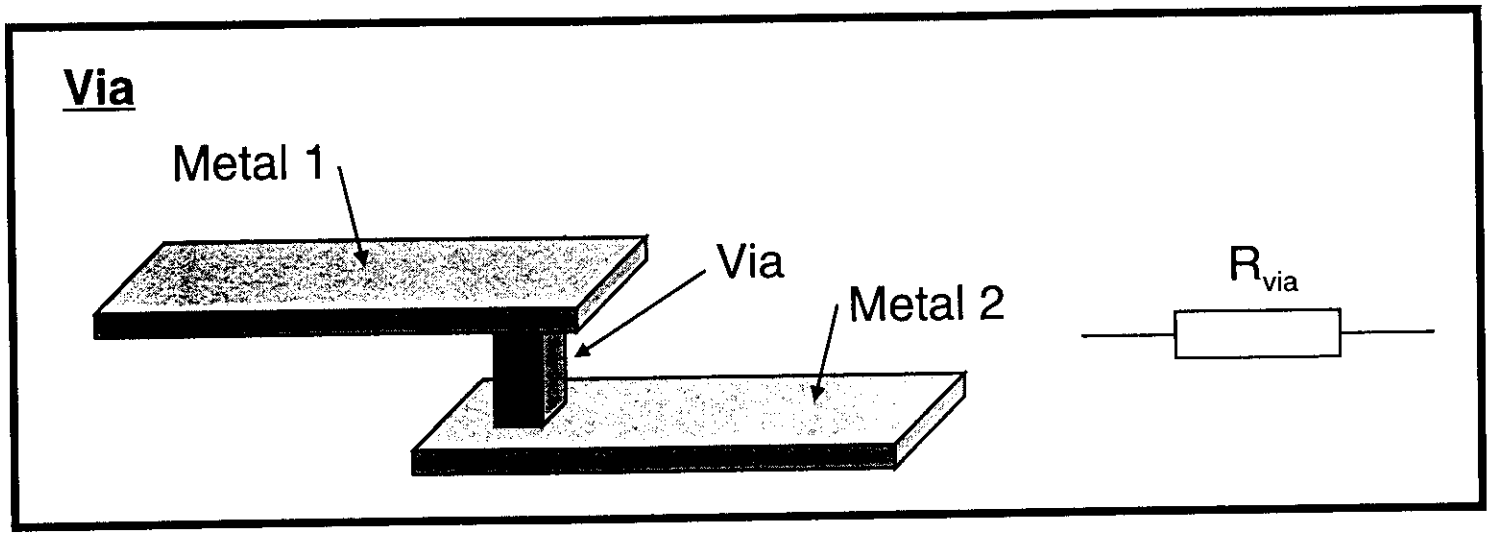
Interconnects



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

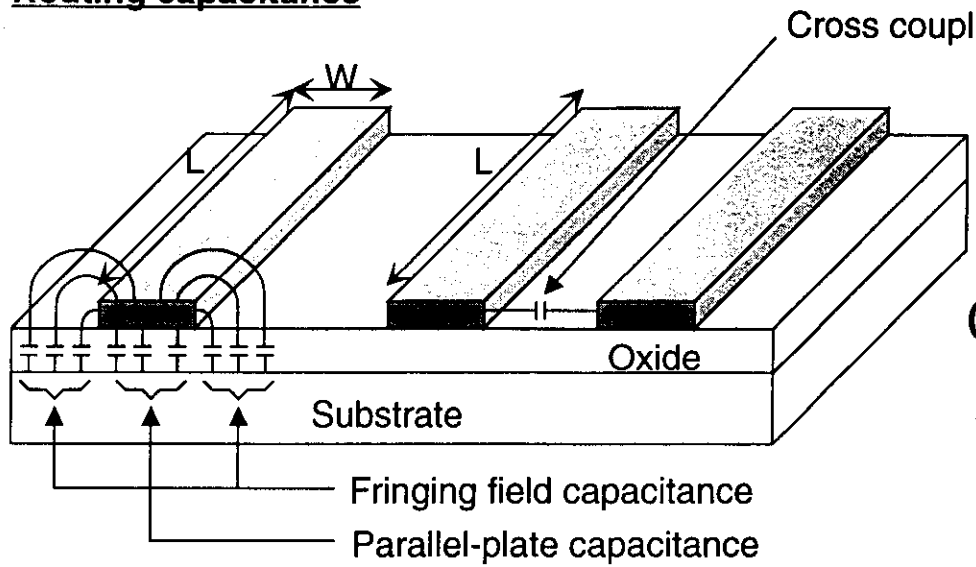


- Via or contact resistance depends on:
  - The contacted materials
  - The contact area

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

Interconnects

Routing capacitance

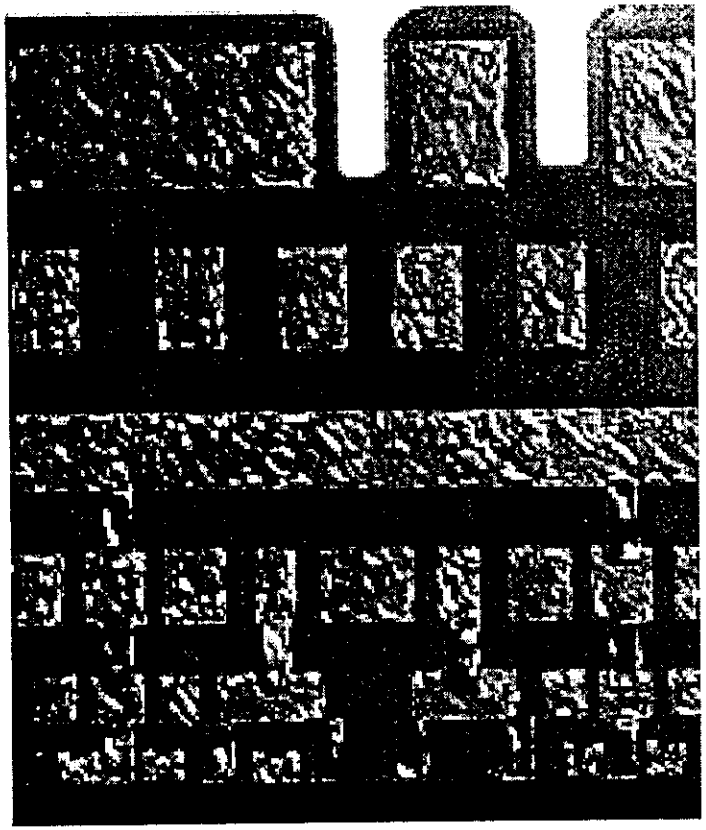


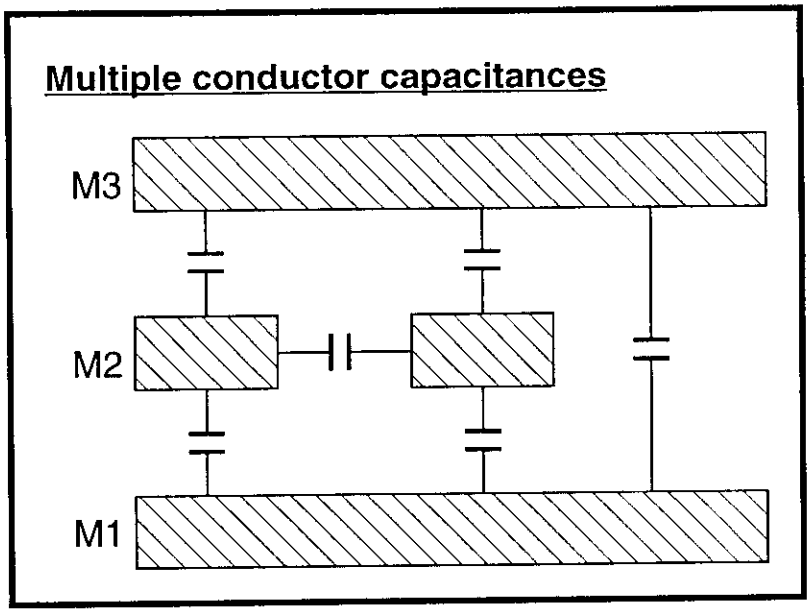
$$C = C_{f0}\, 2\, L + C_{p0}\, W\, L$$

$$C_x = C_{x0}\, L$$

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

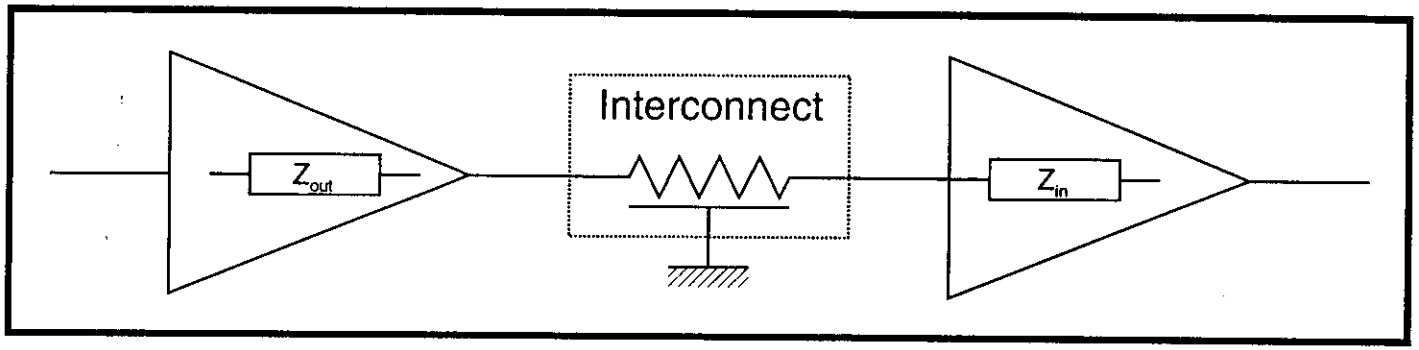
Interconnects





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

Interconnects

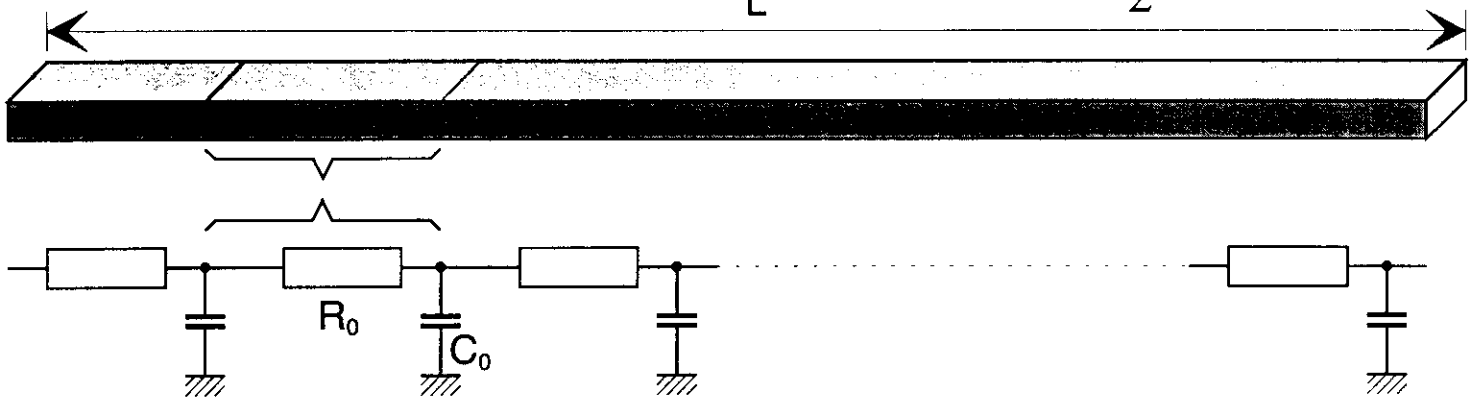


- Delay depends on:
  - Impedance of the driving source
  - Distributed resistance/capacitance of the wire
  - Load impedance
- Distributed RC delay:
  - Can be dominant in long wires
  - Important in polysilicon wires (relatively high resistance)
  - Important in salicided wires
  - Important in heavily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

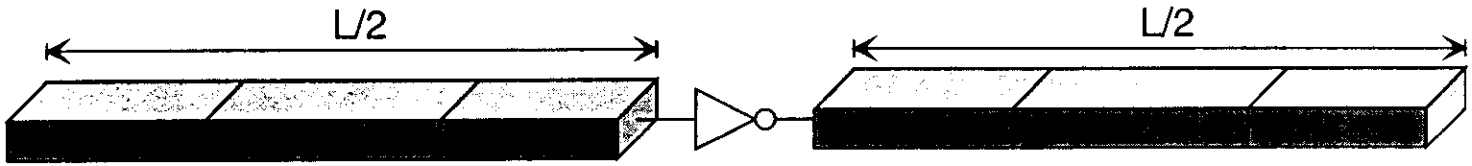
Interconnects

Long line

$$t_d = \frac{1}{2} \cdot R_0 \cdot C_0 \cdot L^2$$



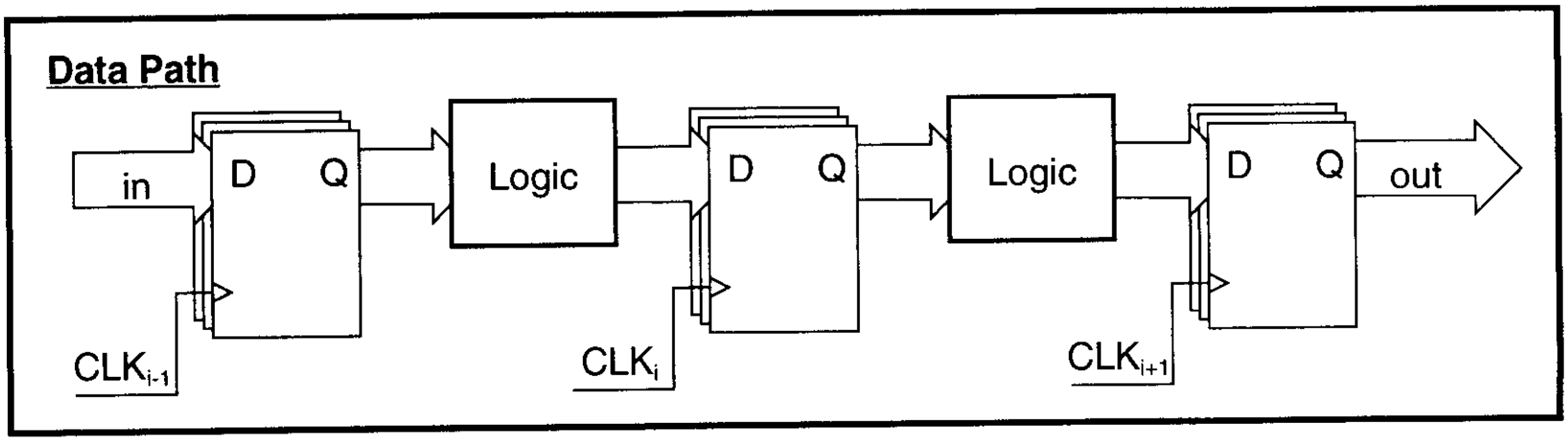
Delay optimization



Clock distribution

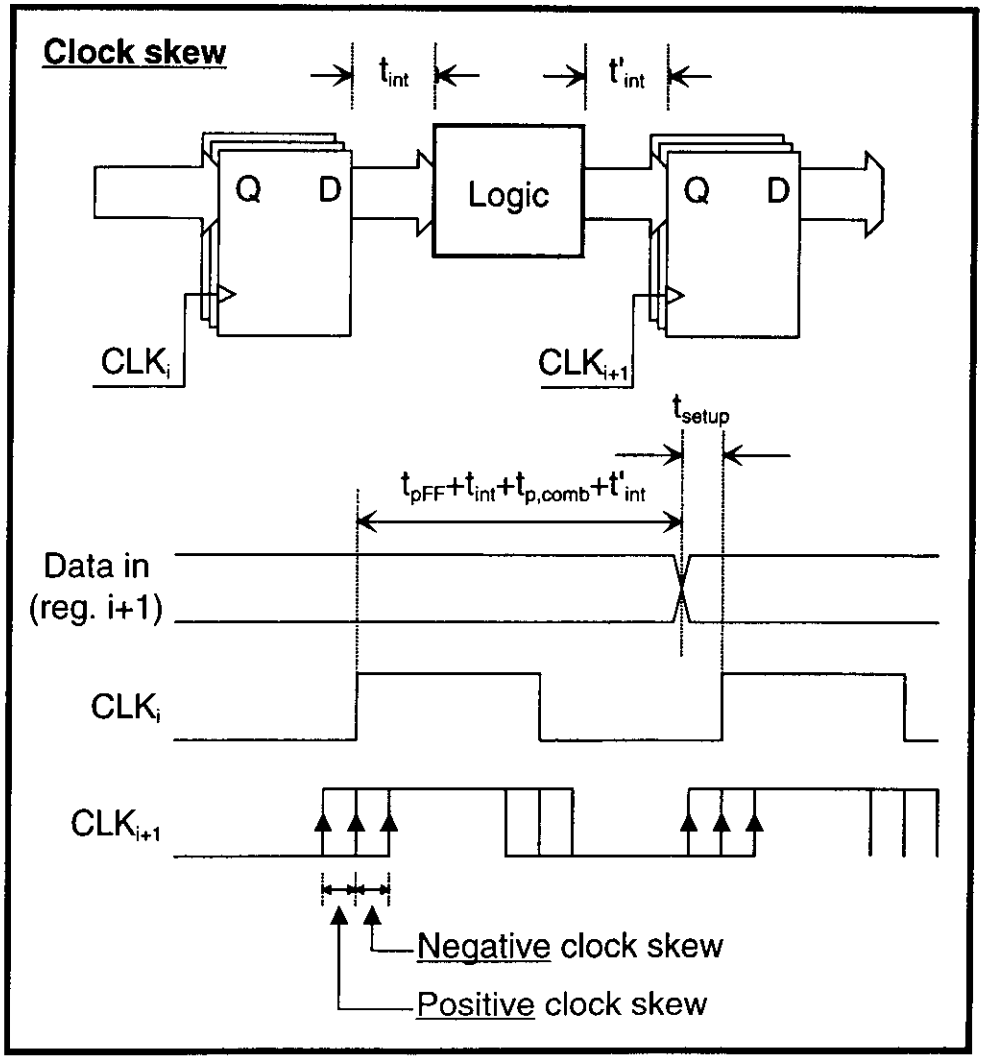
- Clock signals are “special signals”
- Every data movement in a synchronous system is referenced to the clock signal
- Clock signals:
  - Are typically loaded with high fanout
  - Travel over the longest distances in the IC
  - Operate at the highest frequencies

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 <u>simultaneously</u> ... = 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:

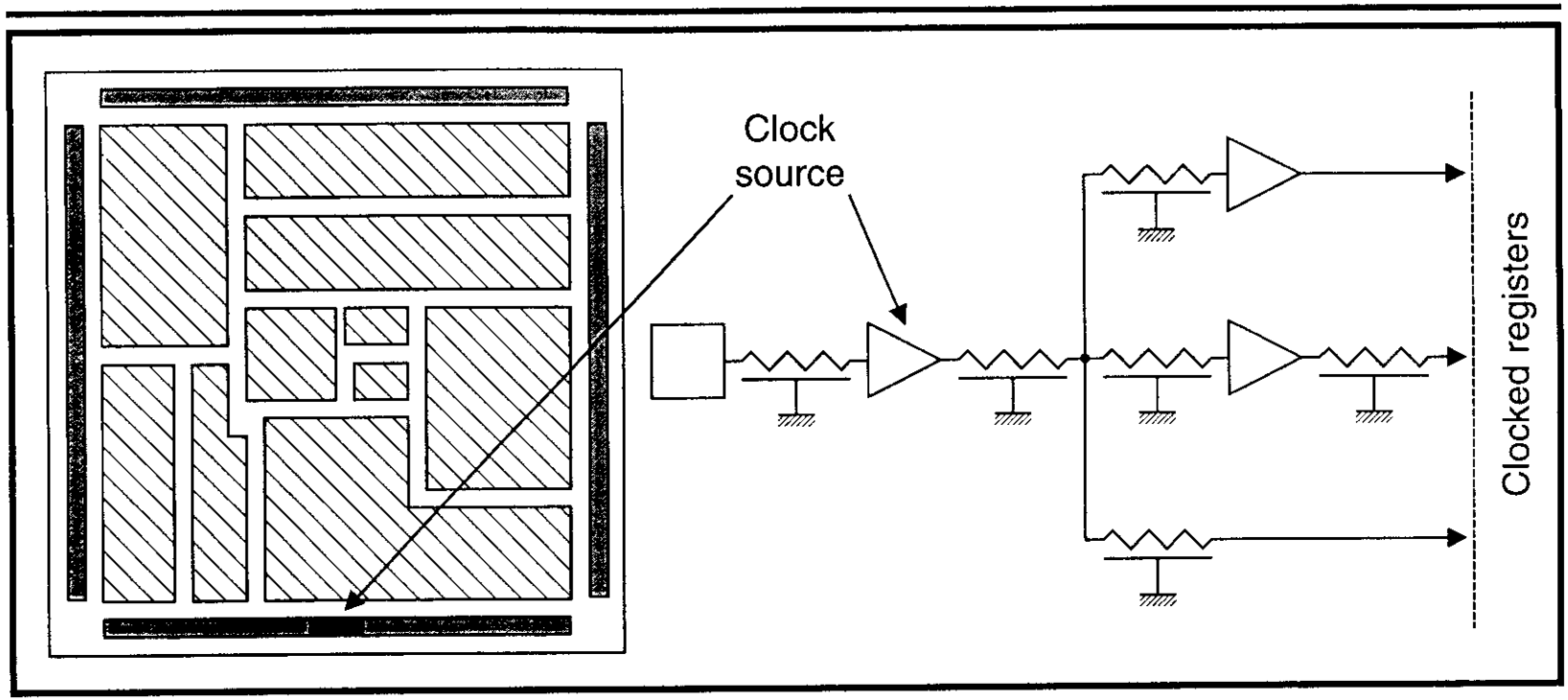
$$T_{\min} = \frac{1}{f_{\max}} = t_{dFF} + t_{\text{int}} + t_{p,comb} + t'_{\text{int}} + t_{setup} + t_{skew}$$

- 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

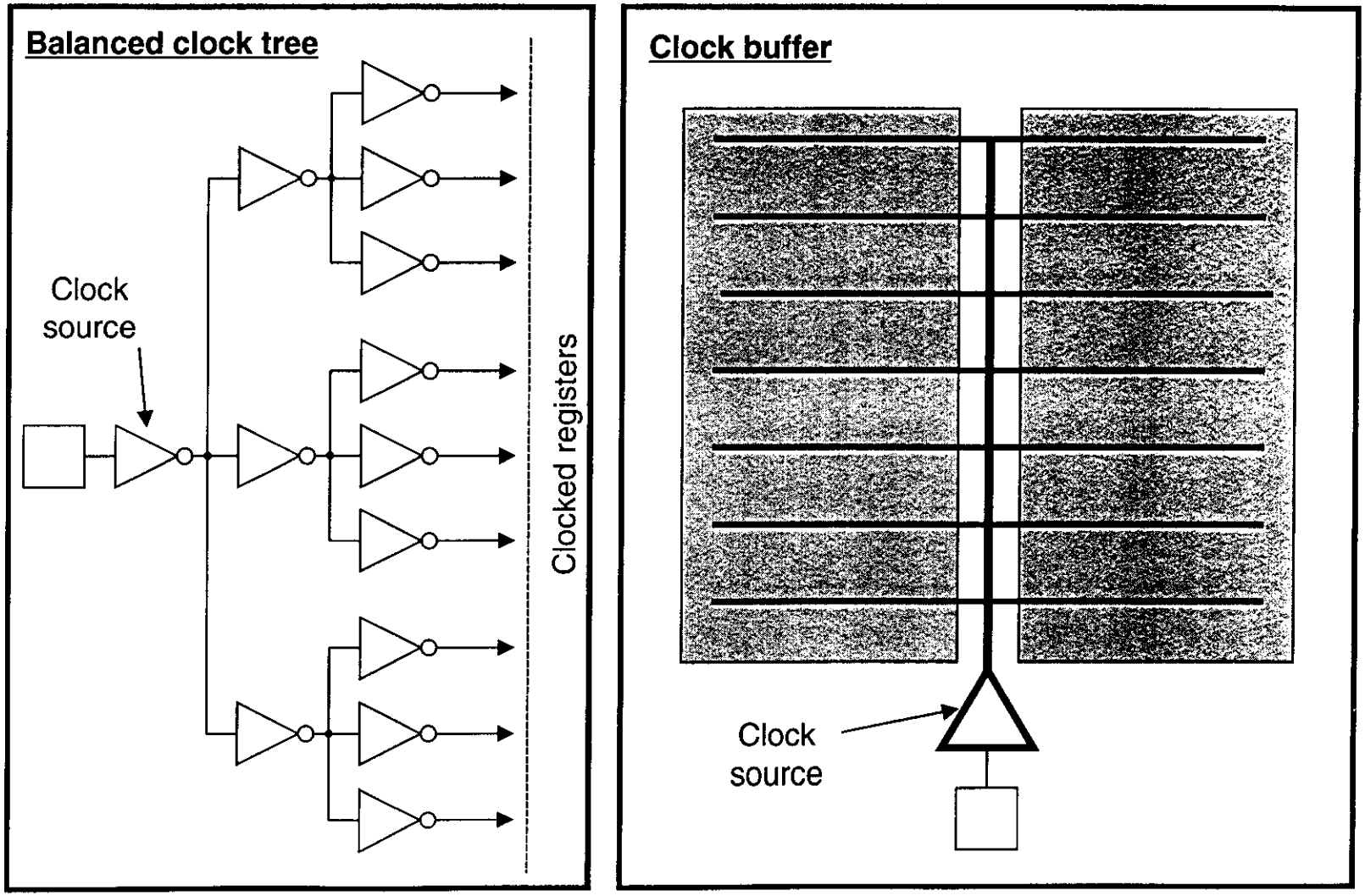
- Different clock paths can have different delays due to:
  - Differences in line lengths from clock source to the clocked registers
  - Differences in delays in the active buffers within the clock distribution network:
    - Differences in passive interconnect parameters (line resistance/capacitance, 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

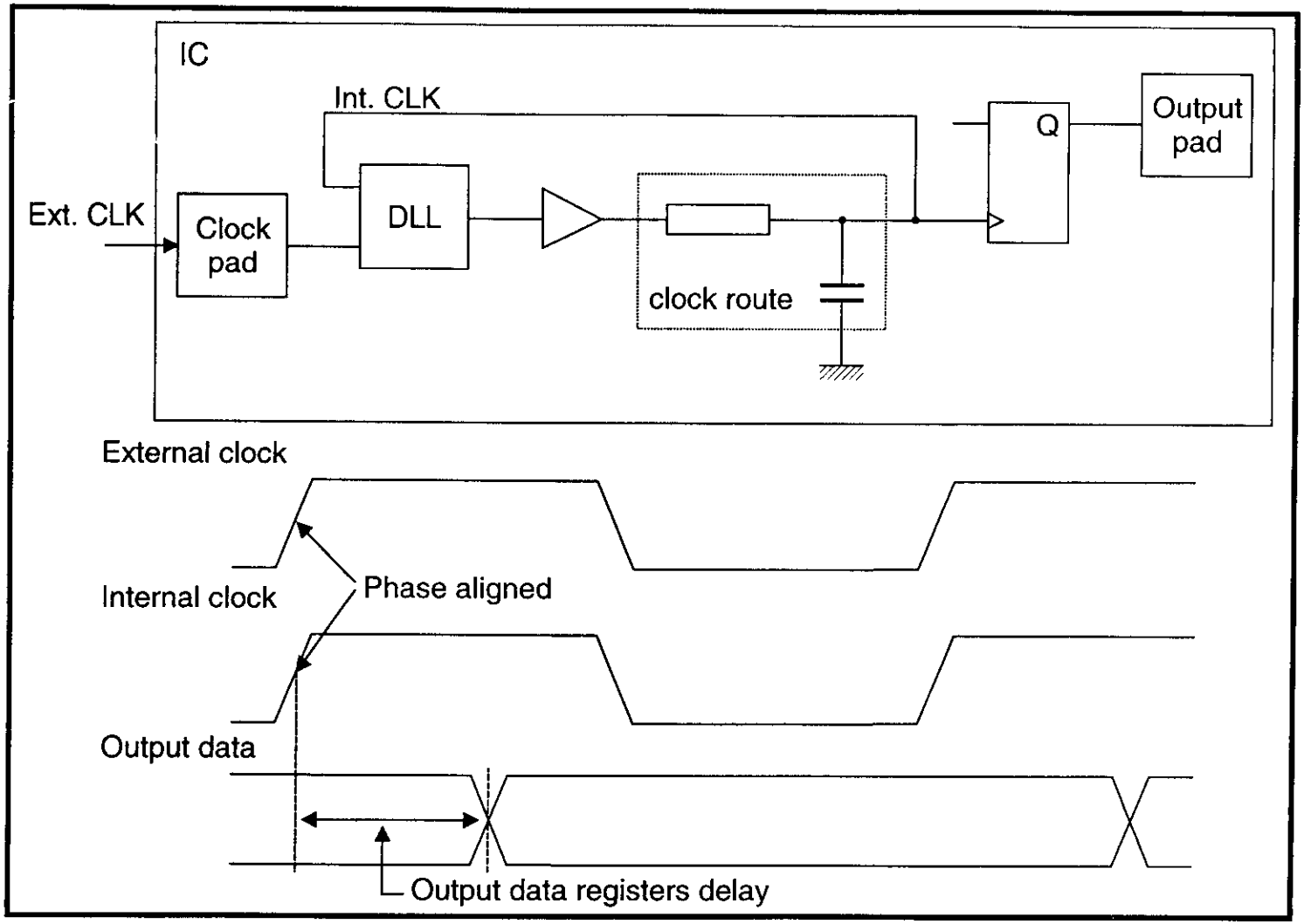


- Clock buffers:
  - Amplify the clock signal degraded by the interconnect impedance
  - Isolate the local clock lines from upstream load impedances

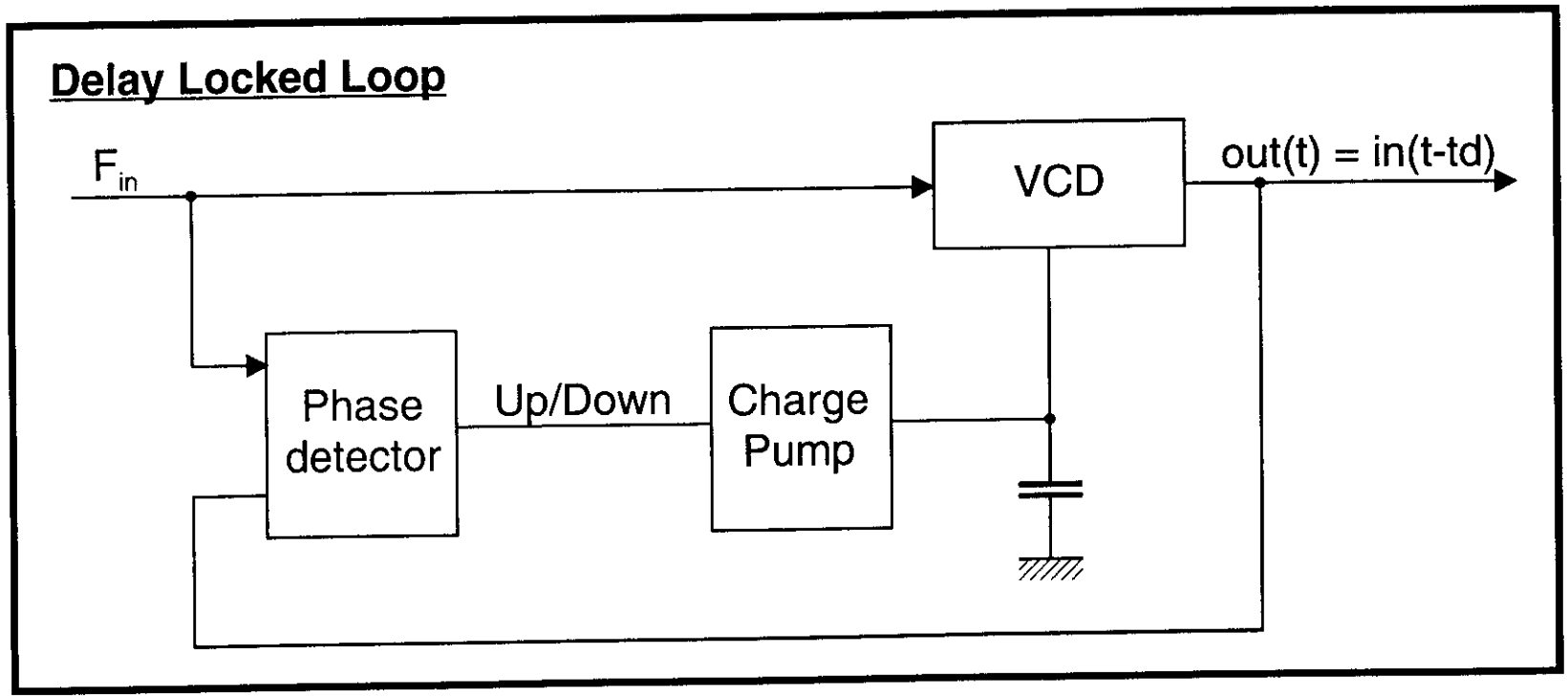
Clock distribution



Delay locked loops

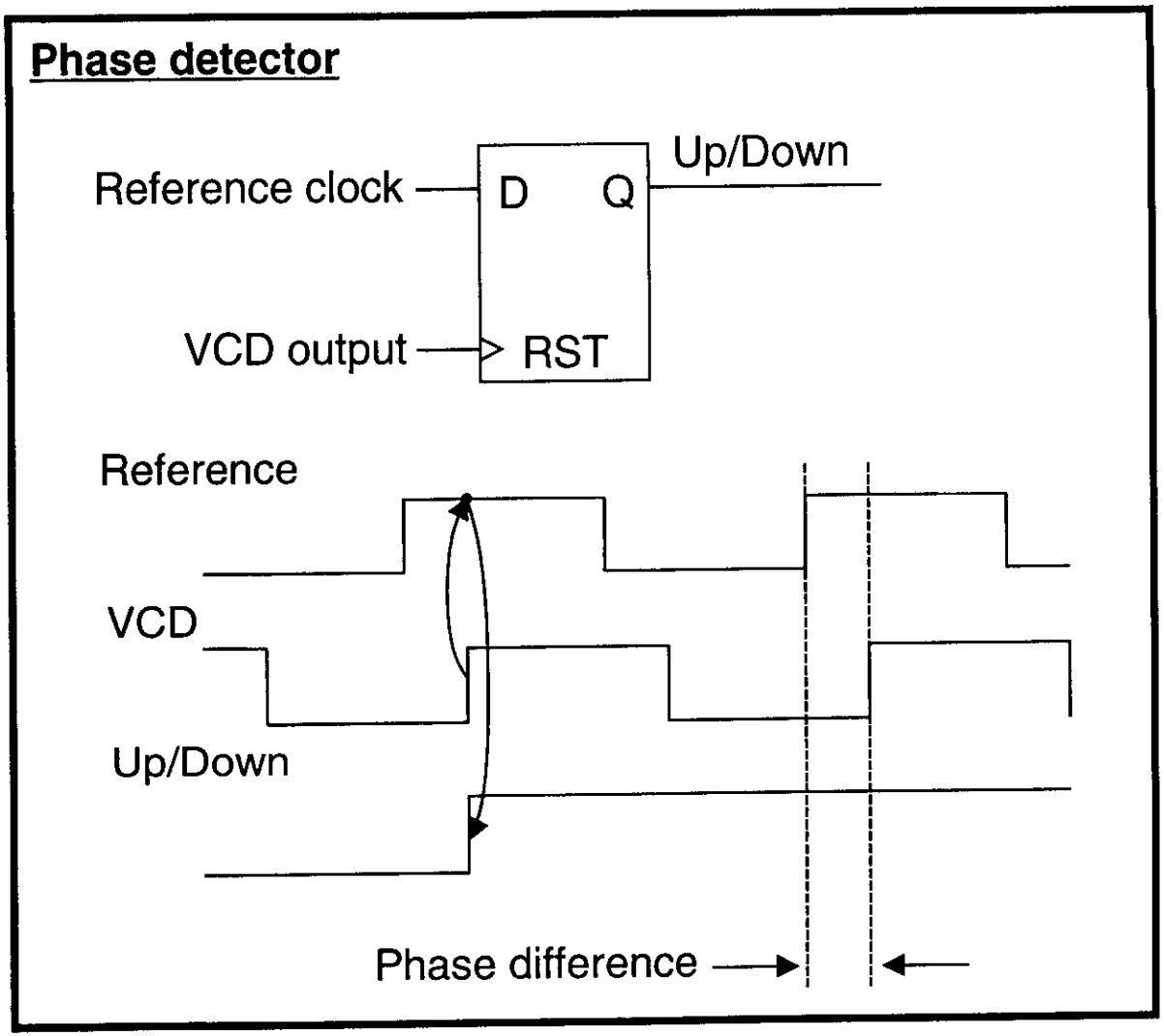


Delay locked loops

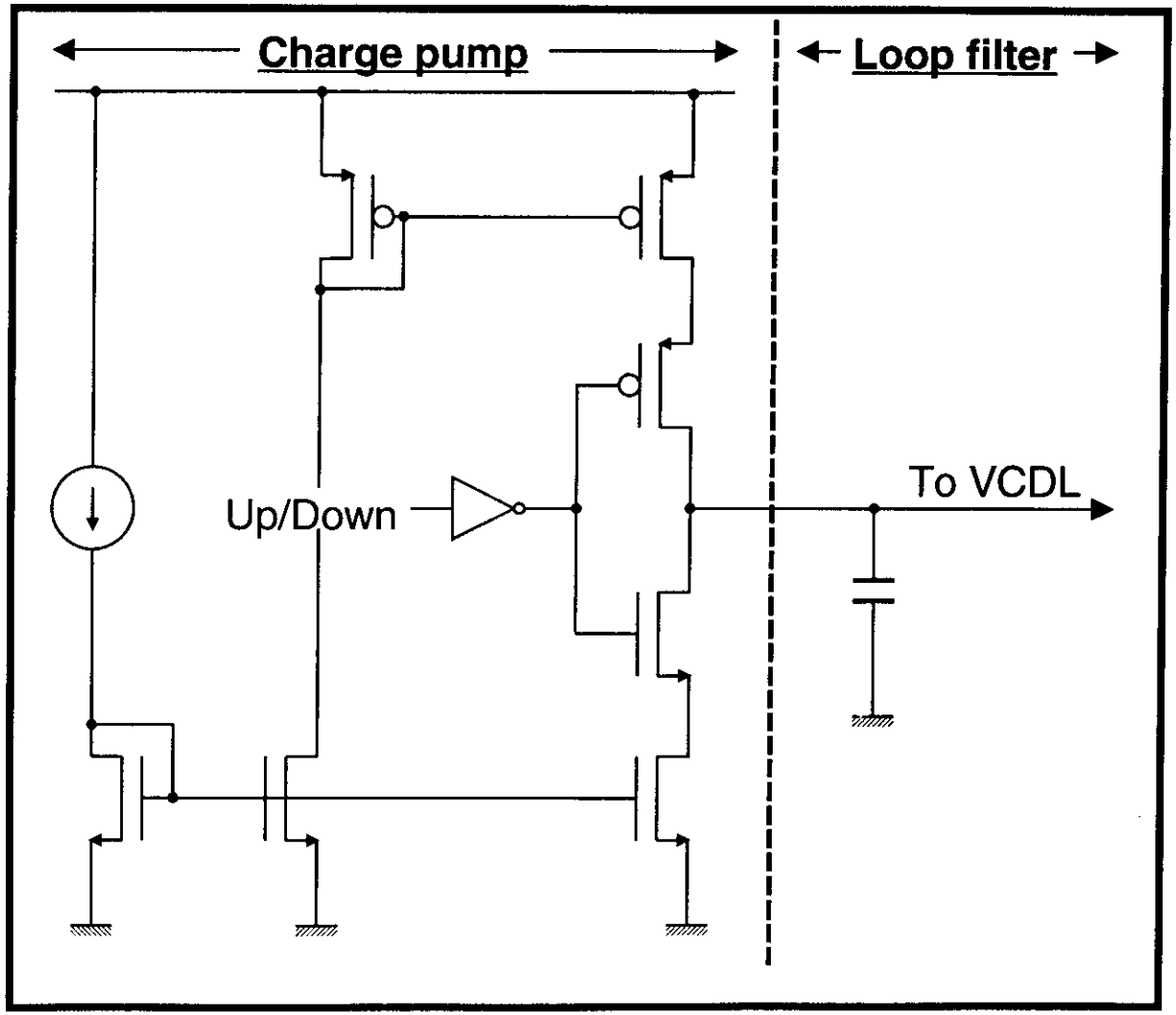


Delay locked loops

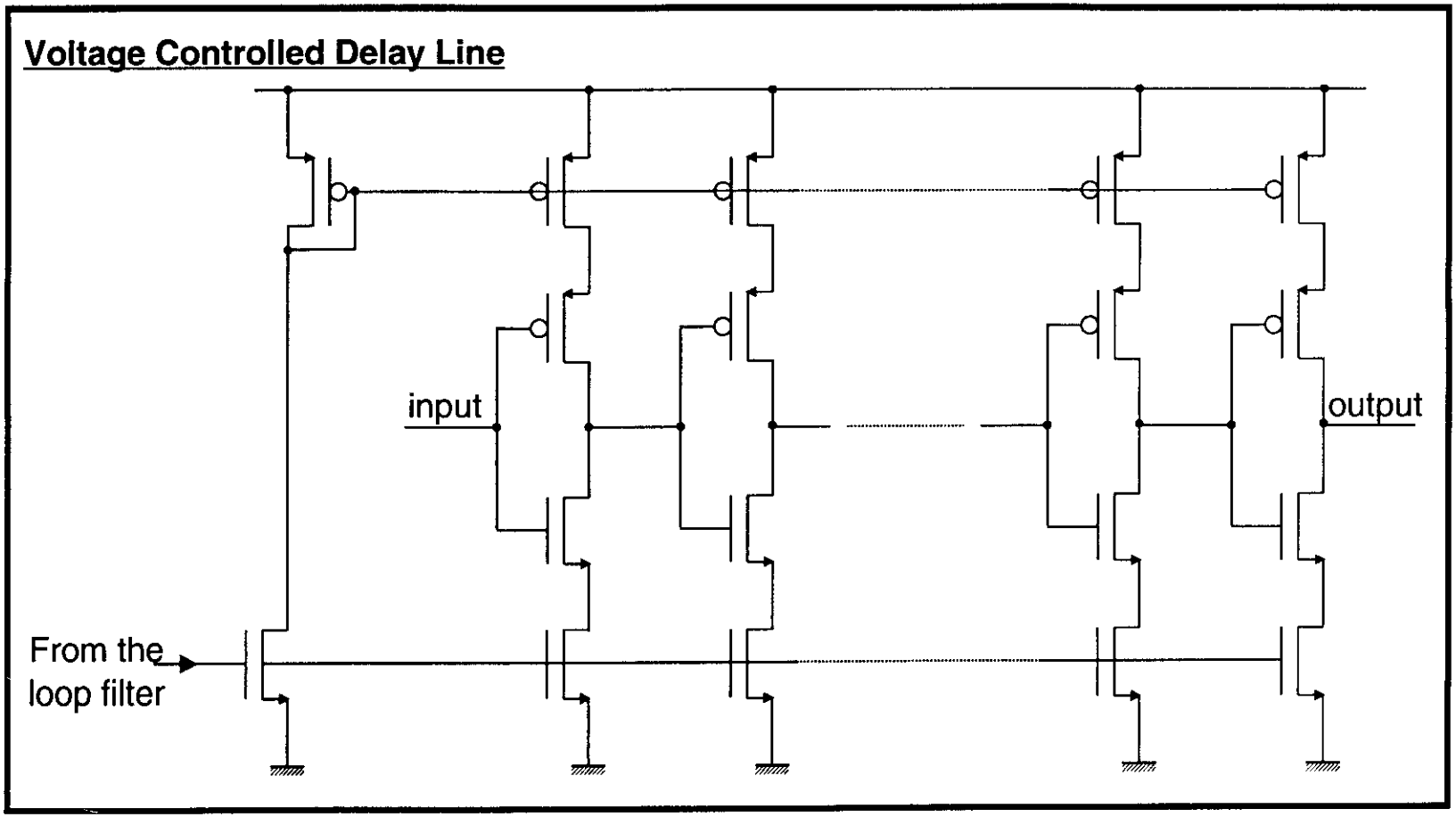
Phase detector



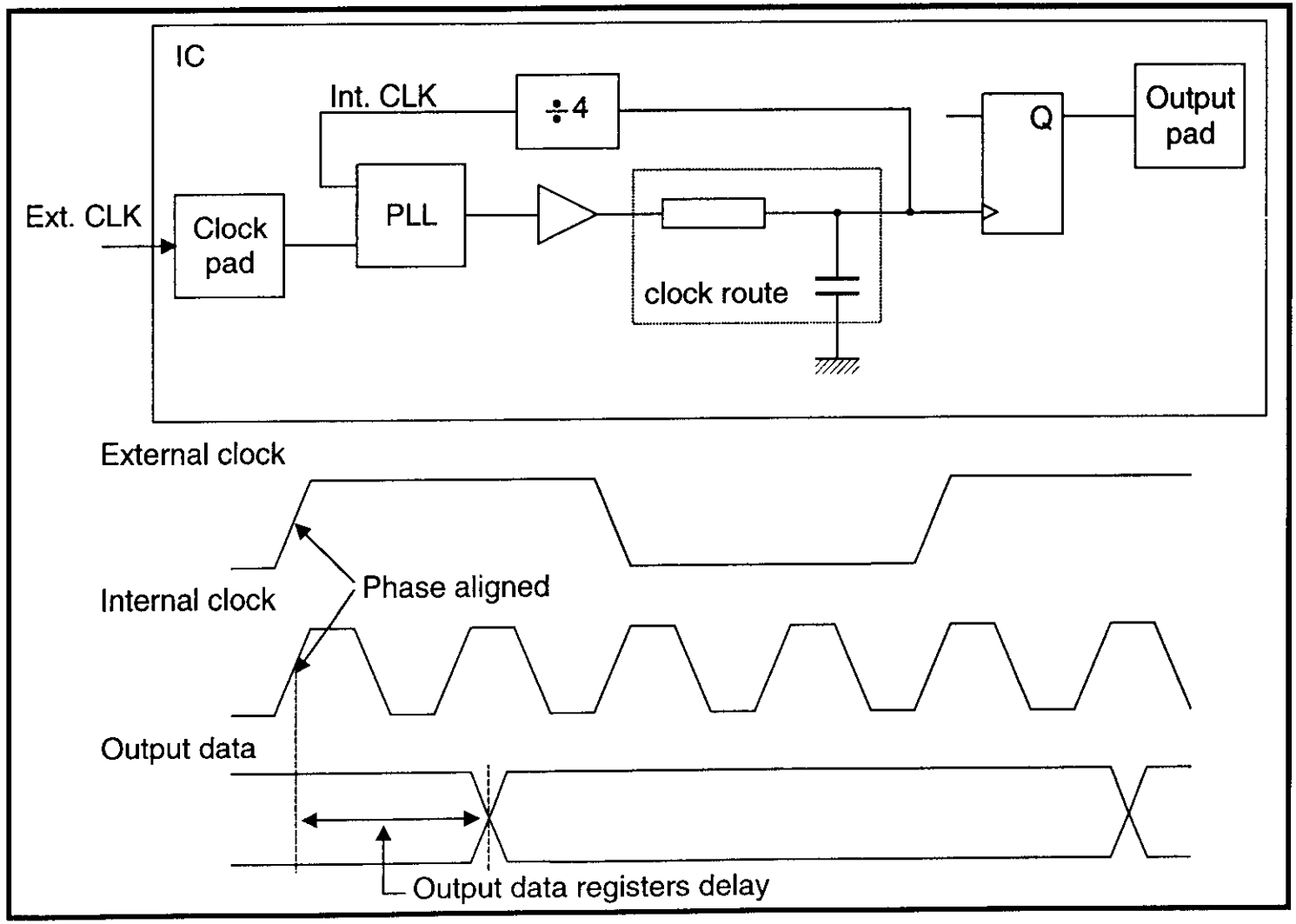
Delay locked loops



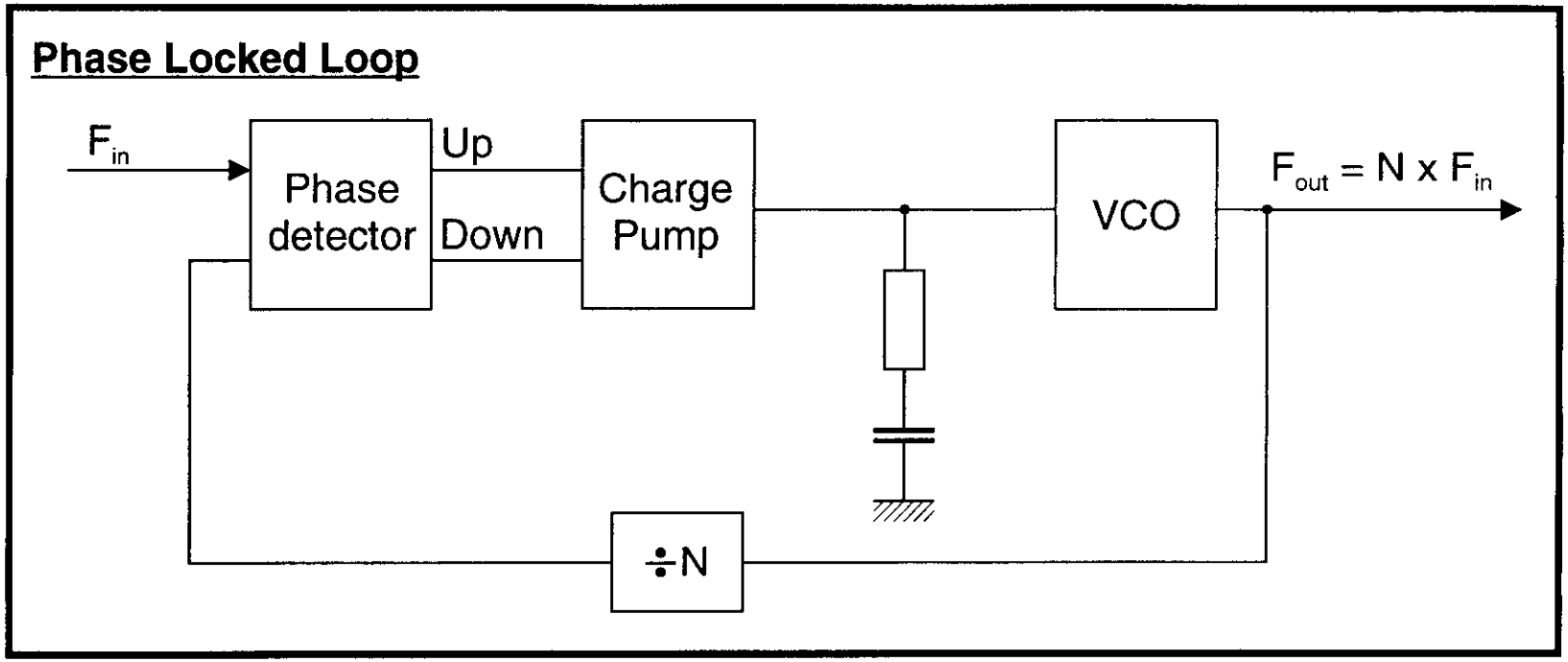
Delay locked loops



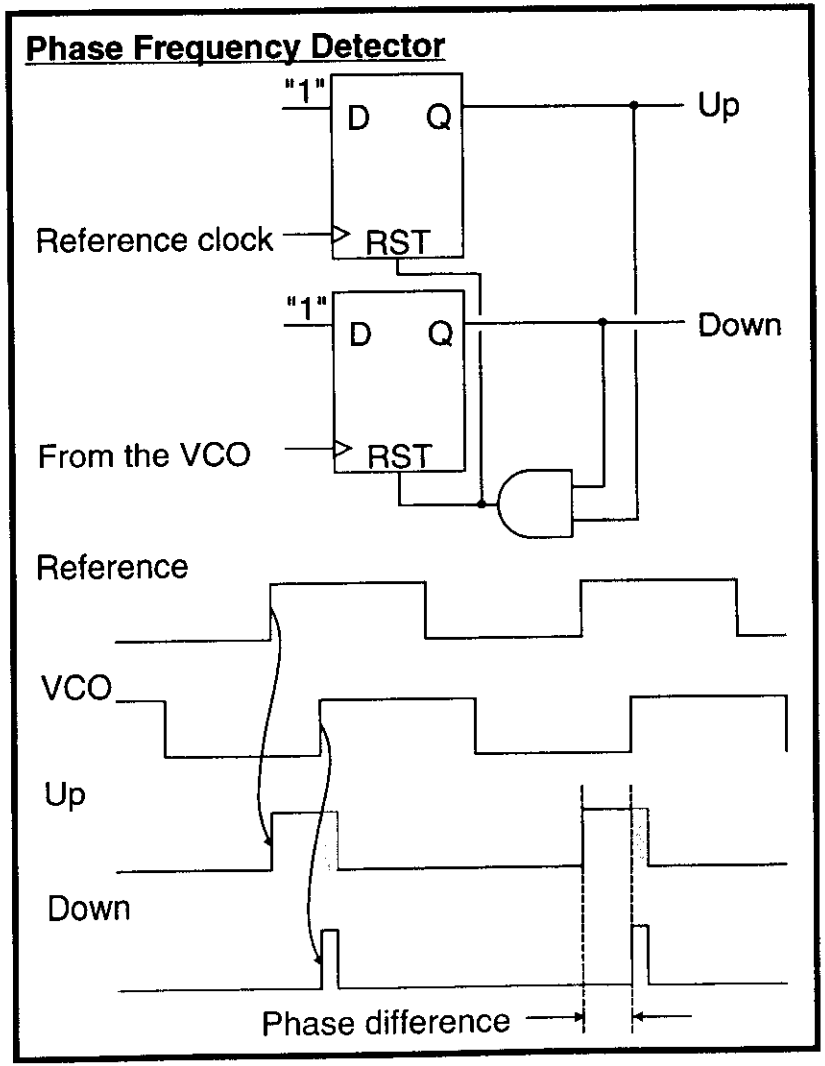
Phase Locked Loops



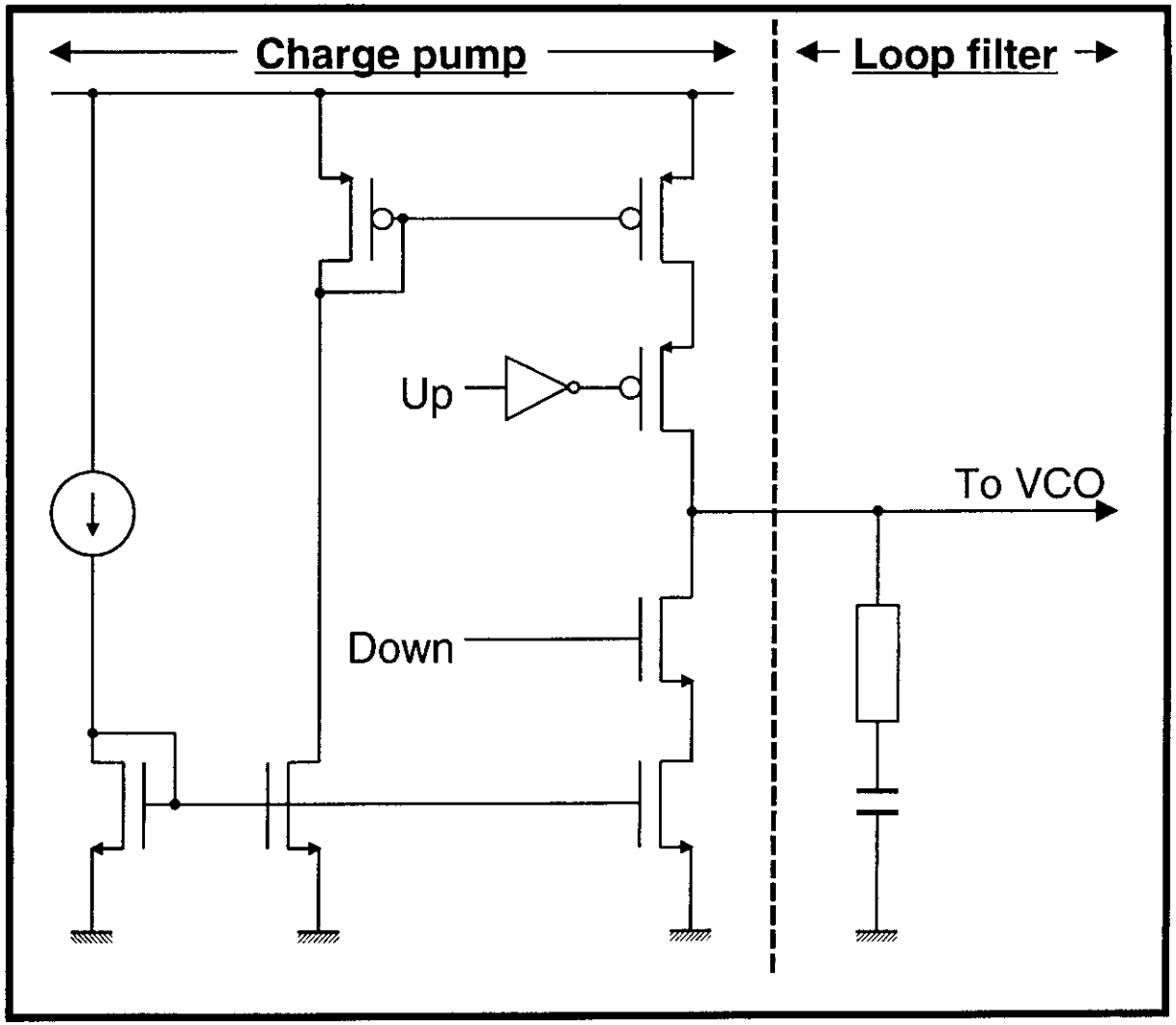
Phase locked loops



Phase locked loops



Phase locked loops



Phase locked loops