The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

310/1780-16

ICTP-INFN Advanced Tranining Course on
FPGA and VHDL for Hardware Simulation and Synthesis
27 November - 22 December 2006

Basic CMOS Technology

Jorgen CHRISTIANSEN
PH-ED
CERN
CH-1221 Geneva 23
SWITZERLAND

These lecture notes are intended only for distribution to participants

CMOS Technology

Paulo Moreira &
Jorgen Christiansen
CERN - Geneva, Switzerland

This part is compressed set of transparencies
from Paulo Moreira:
http://paulo.moreira.free.fr/

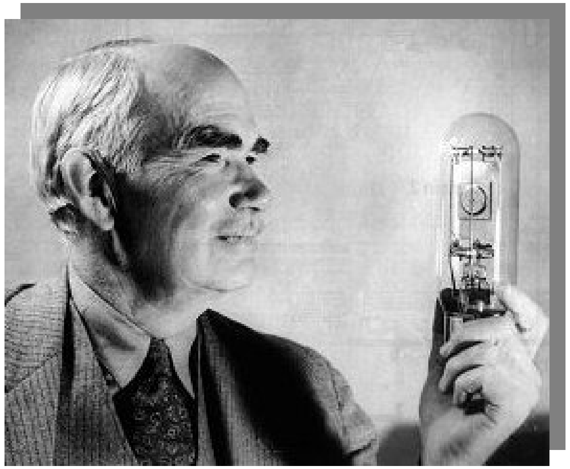
Outline

- Part 1: Basic CMOS technology (from Paulo Moreira)
  - How it all started
  - CMOS Transistors
  - Parasitics
  - The CMOS inverter
  - Technology
  - Scaling

Introduction

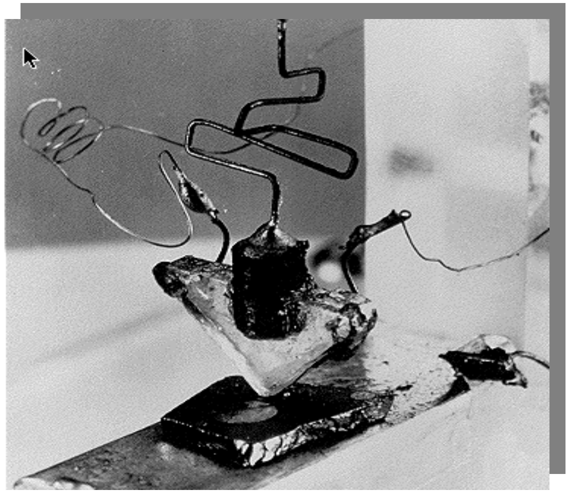
1906

Audion (Triode), 1906
Lee De Forest



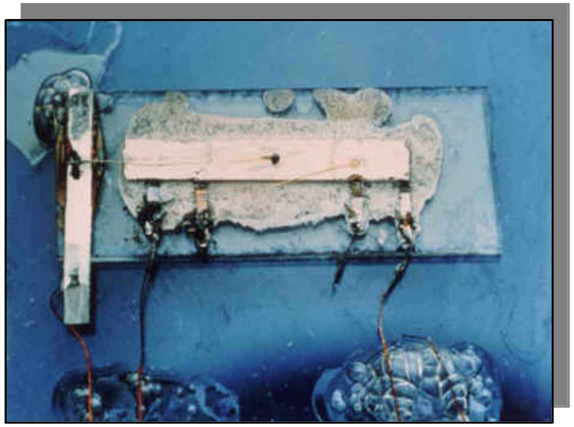
1947

First point contact transistor (germanium), 1947
John Bardeen and Walter Brattain
Bell Laboratories



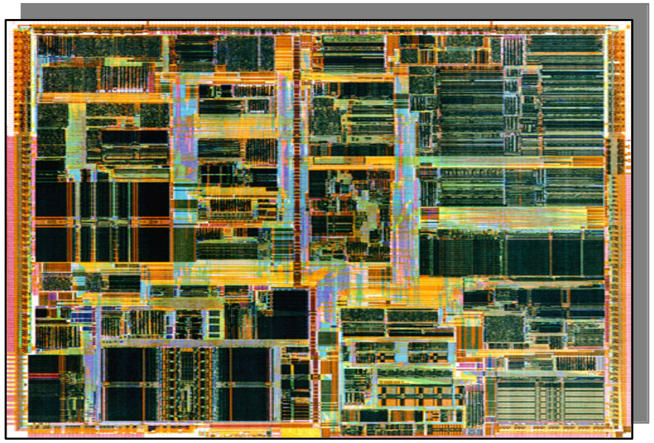
1958

First integrated circuit (germanium), 1958
Jack S. Kilby, Texas Instruments
transistors resistors and capacitors



1997

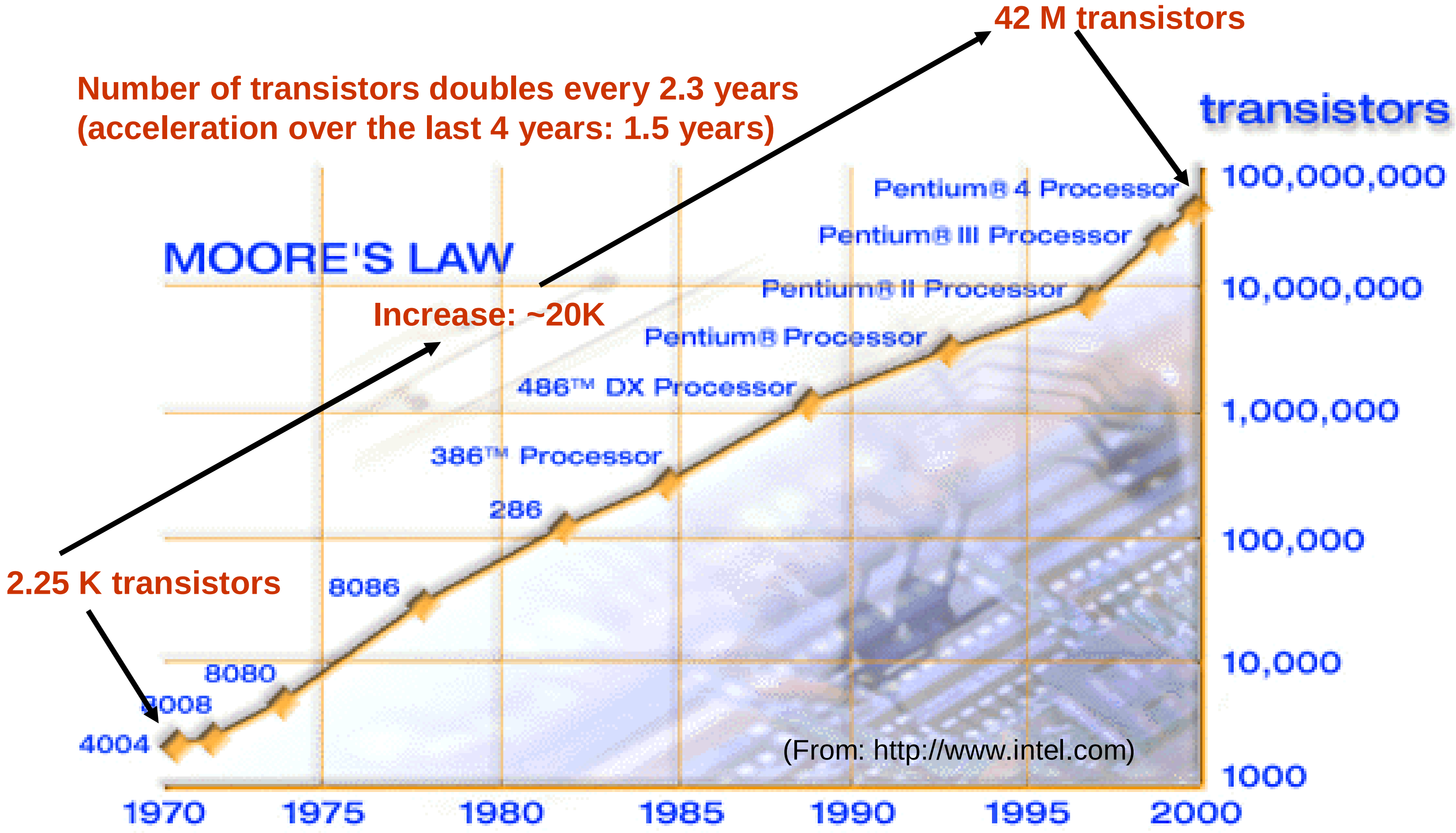
Intel Pentium II, 1997
Gate Length: 0.35, Clock: 233MHz
Number of transistors: 7.5 M



The world is becoming digital

- Digital processing is taking over:
  - Computing, DSP
  - Instrumentation
  - Control systems
  - Telecommunications
  - Consumer electronics
- But analog is still needed in critical parts:
  - Amplification of weak signals
  - A/D and D/A conversion
  - Radio Frequency (RF) communications
- As digital systems become faster and circuit densities increase the analog side of digital circuits are becoming important
  - Crosstalk, Delays in R-C wires, Jitter, matching, substrate noise, etc.

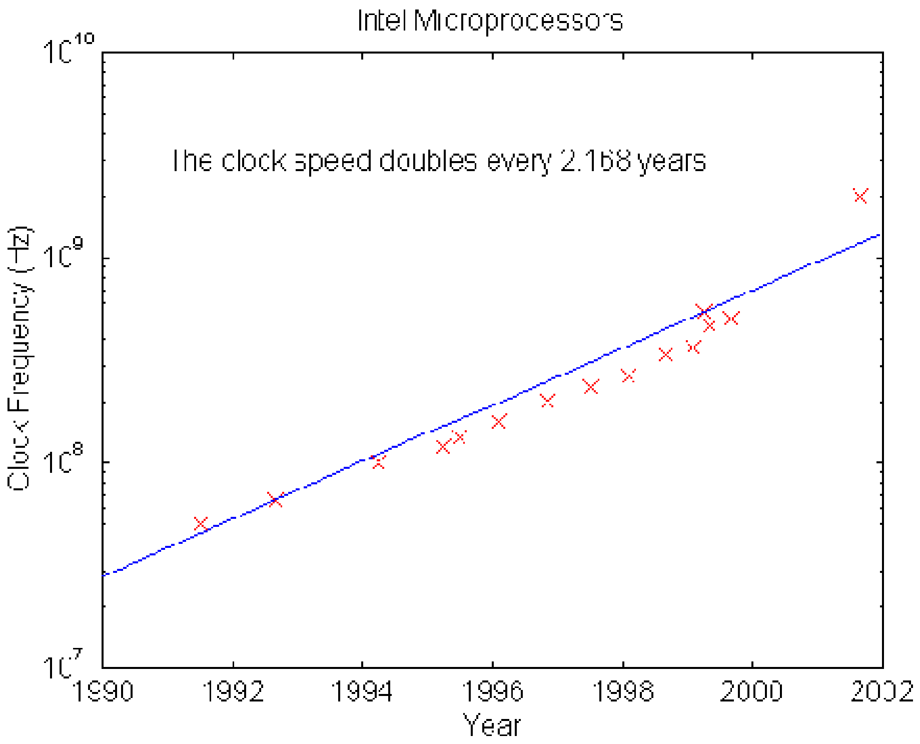
"Moore's Law"

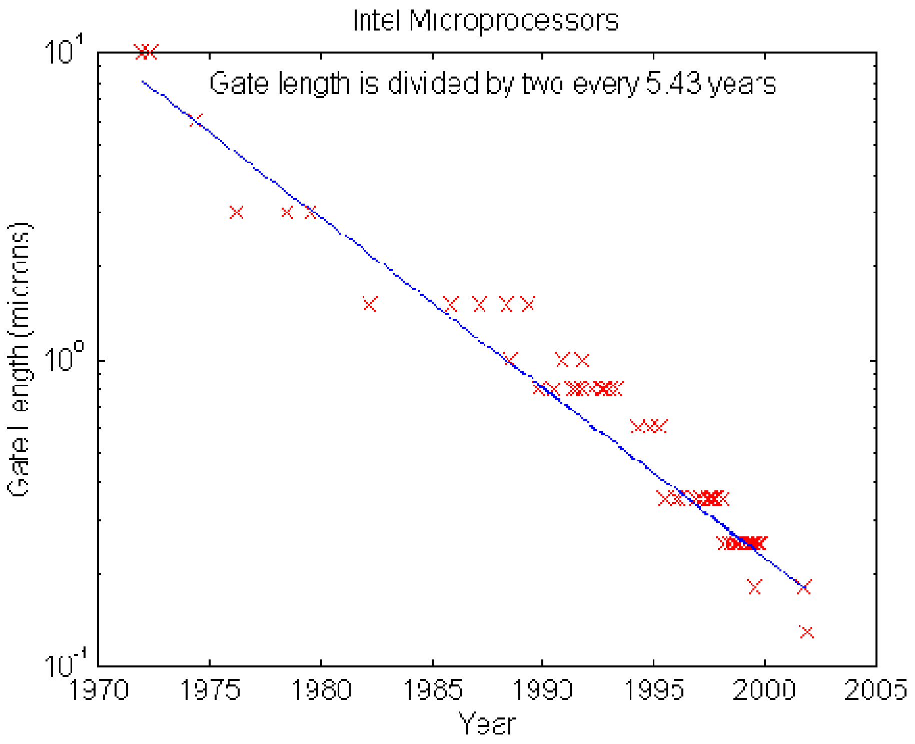


(From: http://www.intel.com)

"Integration complexity doubles every three years"
Gordon Moore, Fairchild1965

Trends


Intel Microprocessors
The clock speed doubles every 2.168 years
Clock Frequency (Hz)
10^{10}
10^{9}
10^{8}
10^{7}
1990
1992
1994
1996
1998
2000
2002
Year


Intel Microprocessors
Gate length is divided by two every 5.43 years
Gate Length (microns)
10^{1}
10^{0}
10^{-1}
1970
1975
1980
1985
1990
1995
2000
2005
Year

Driving force: Economics

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

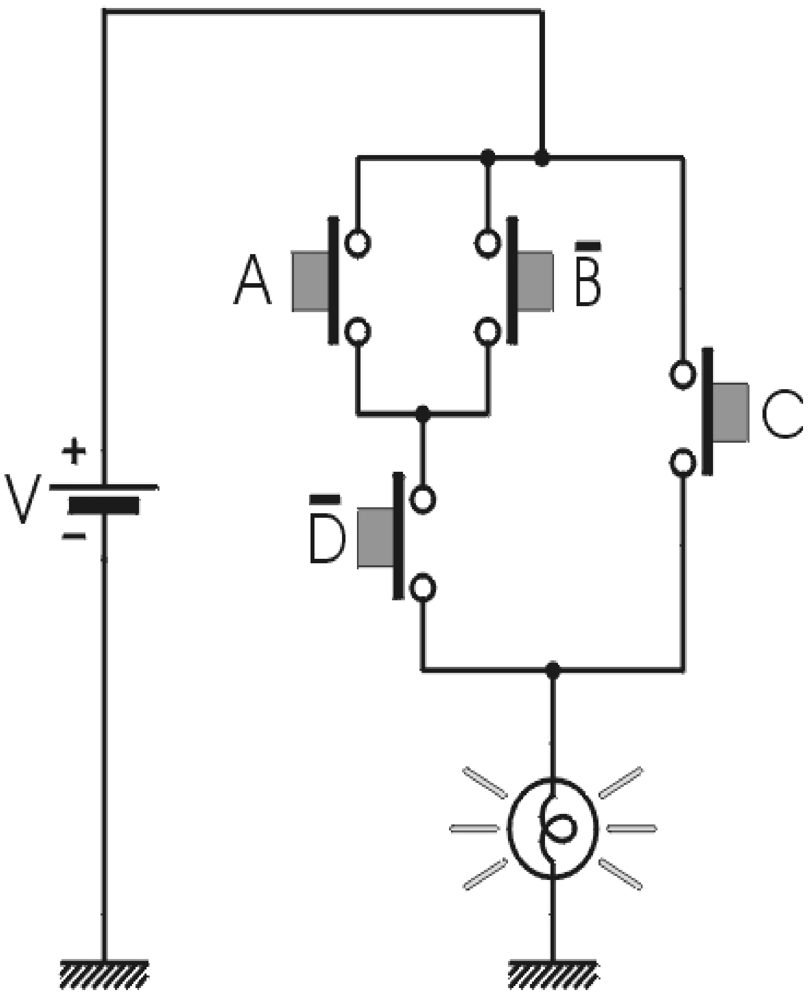
“CMOS building blocks”

- “Making Logic”
- Silicon switches:
  - The NMOS
  - Its mirror image, the PMOS
- Electrical behavior:
  - Strong inversion
    - Model
    - How good is the approximation?
  - Weak inversion
  - Gain and inversion

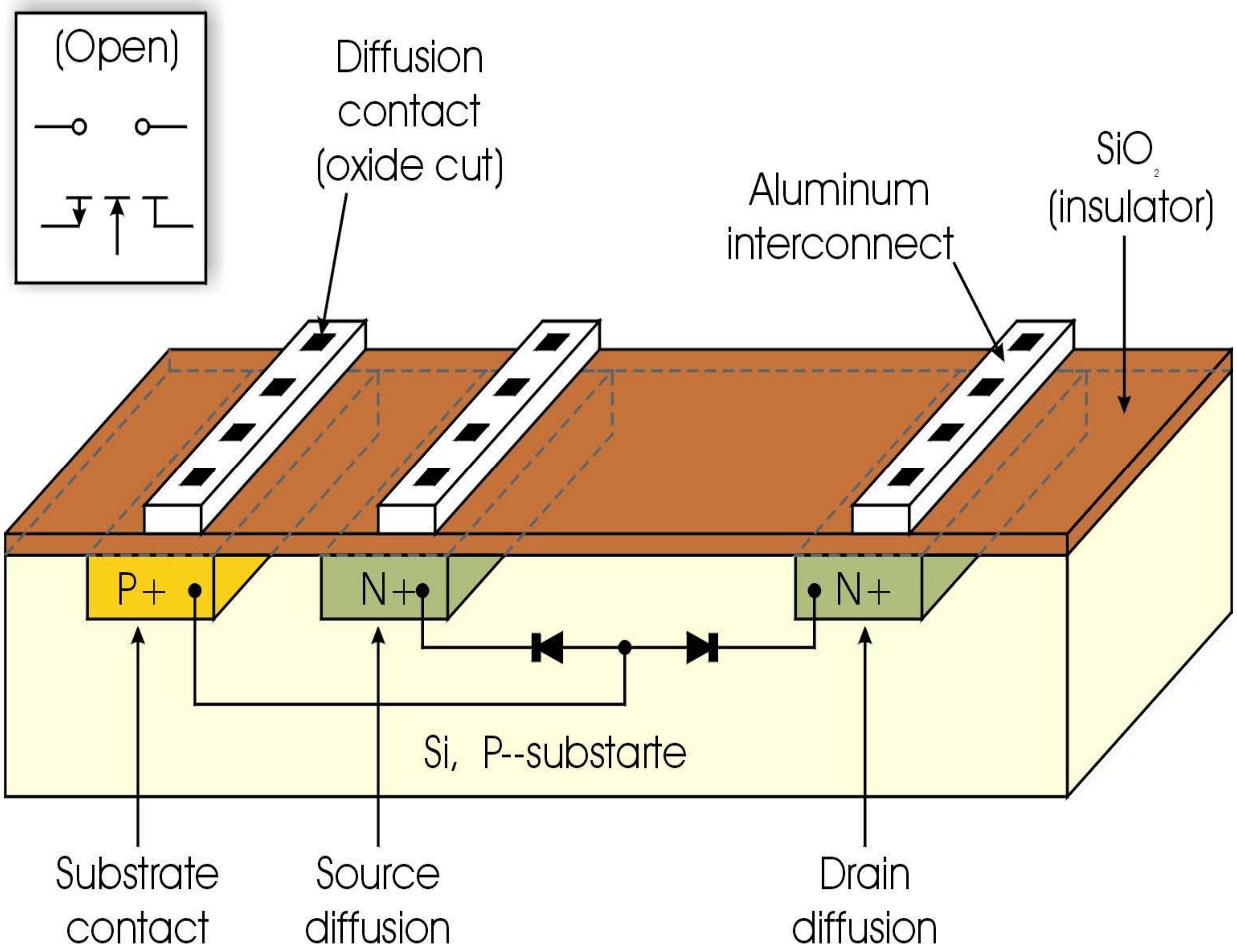
“Making Logic”

- Logic circuit “ingredients”:
  - Power source
  - Switches
  - Power gain
  - Inversion
- Power always comes from some form of external EMF generator.
- NMOS and PMOS transistors:
  - Can perform the last three functions
  - They are the building blocks of CMOS technologies!

$$\text{Light ON} = (A + \overline{B})\,\overline{D} + C$$



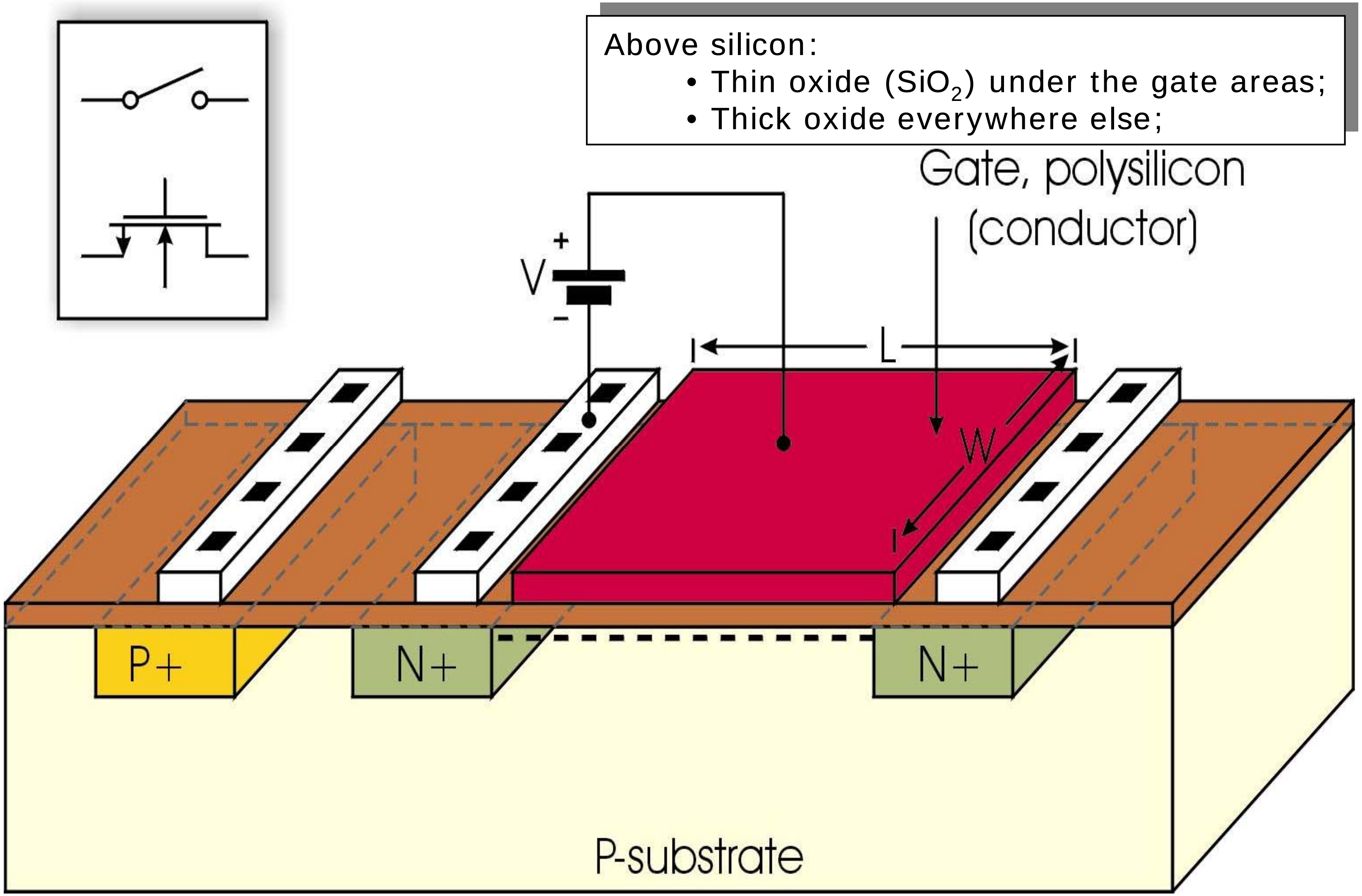
Silicon switches: the NMOS



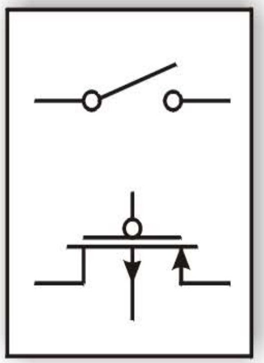
Silicon switches: the NMOS

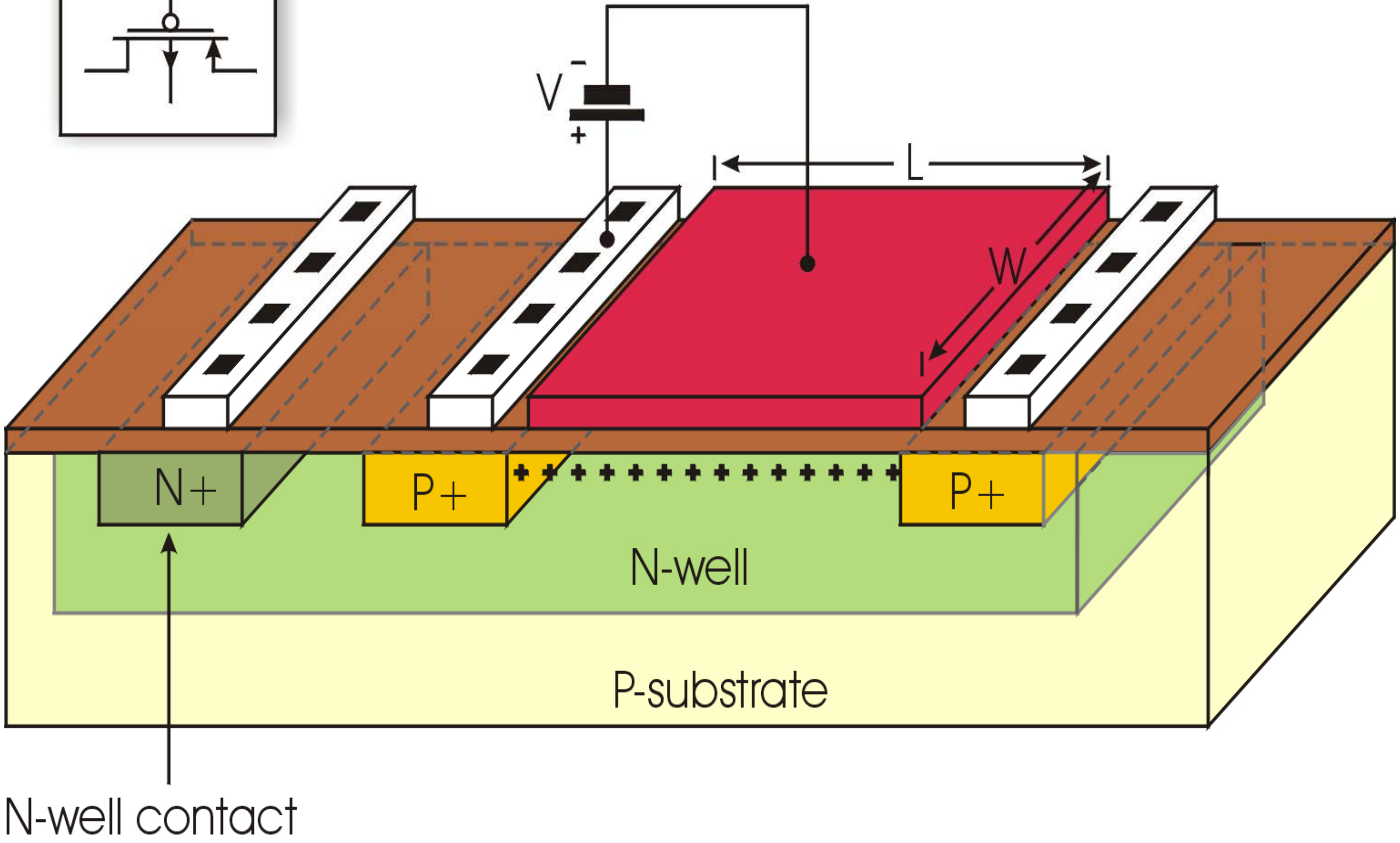
Above silicon:

- Thin oxide (SiO_2) under the gate areas;
- Thick oxide everywhere else;



Silicon switches: the PMOS





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”

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

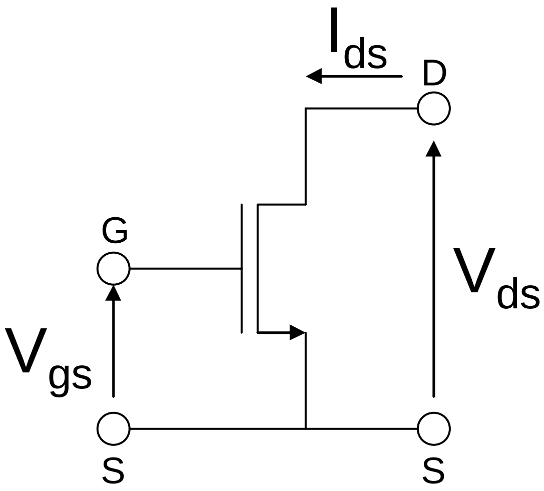
0.24μm process

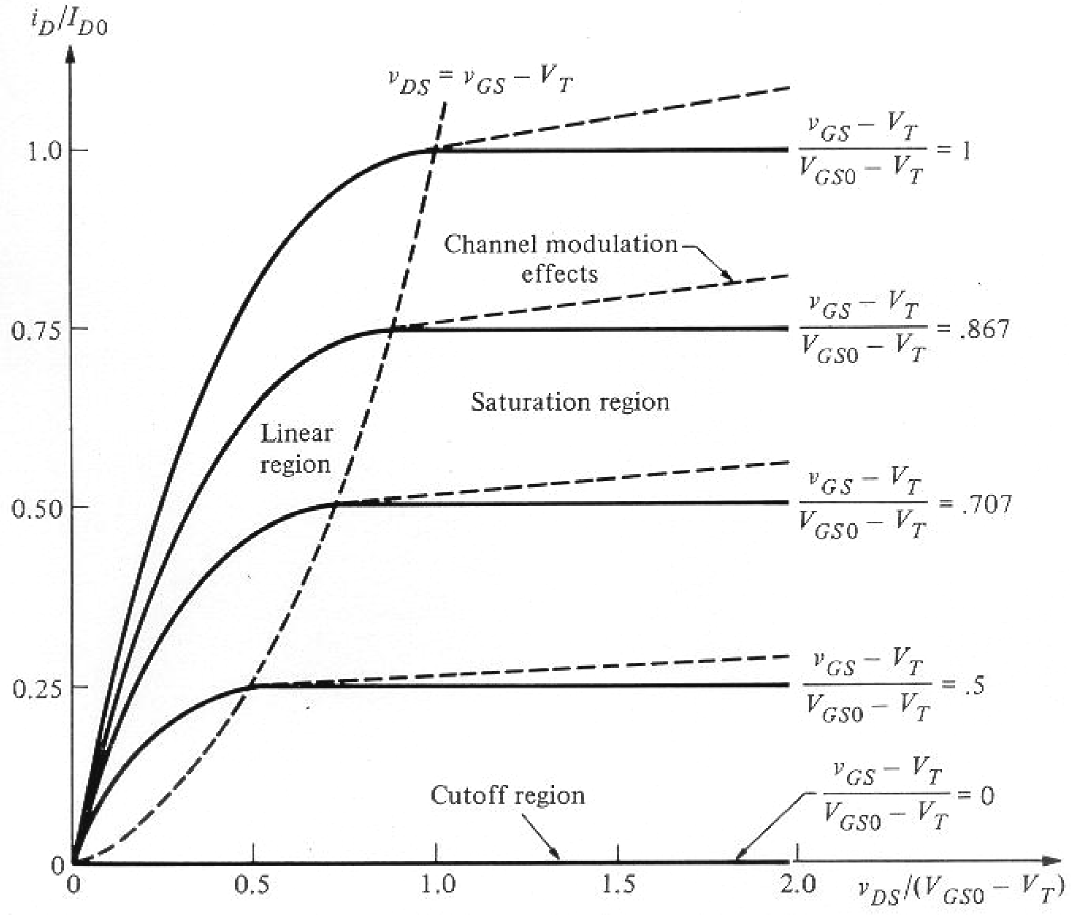
t_{ox} = 5nm (~10 atomic layers)

C_{ox} = 5.6fF/μm^2

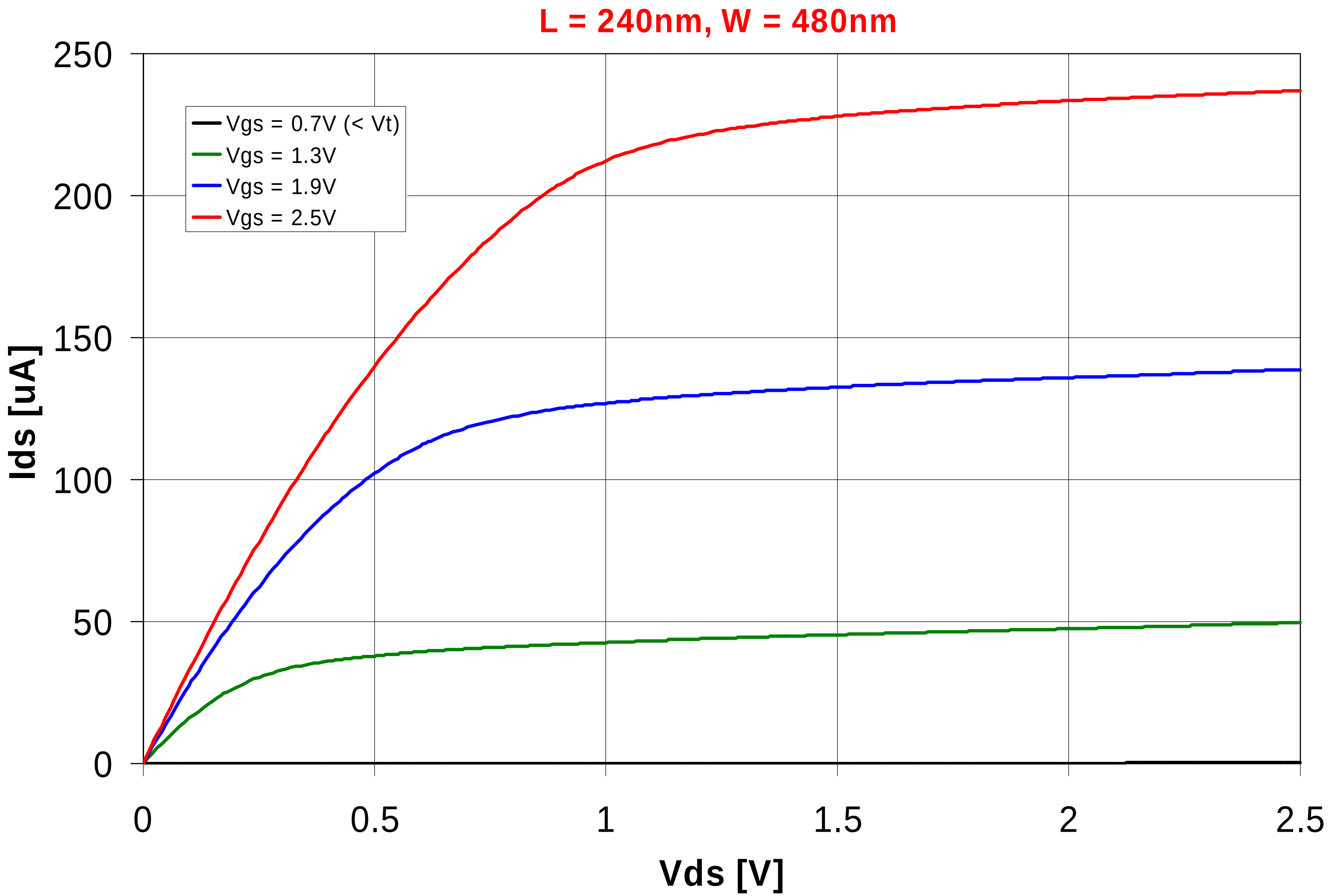
MOS output characteristics

- **Cut off: Vgs < Vt**
- **Linear region**: $V_{ds}<V_{gs}-V_T$
  - Voltage controlled resistor
- **Saturation region**: $V_{ds}>V_{gs}-V_T$
  - Voltage controlled current source
  - Deviation from the ideal current source caused by channel modulation

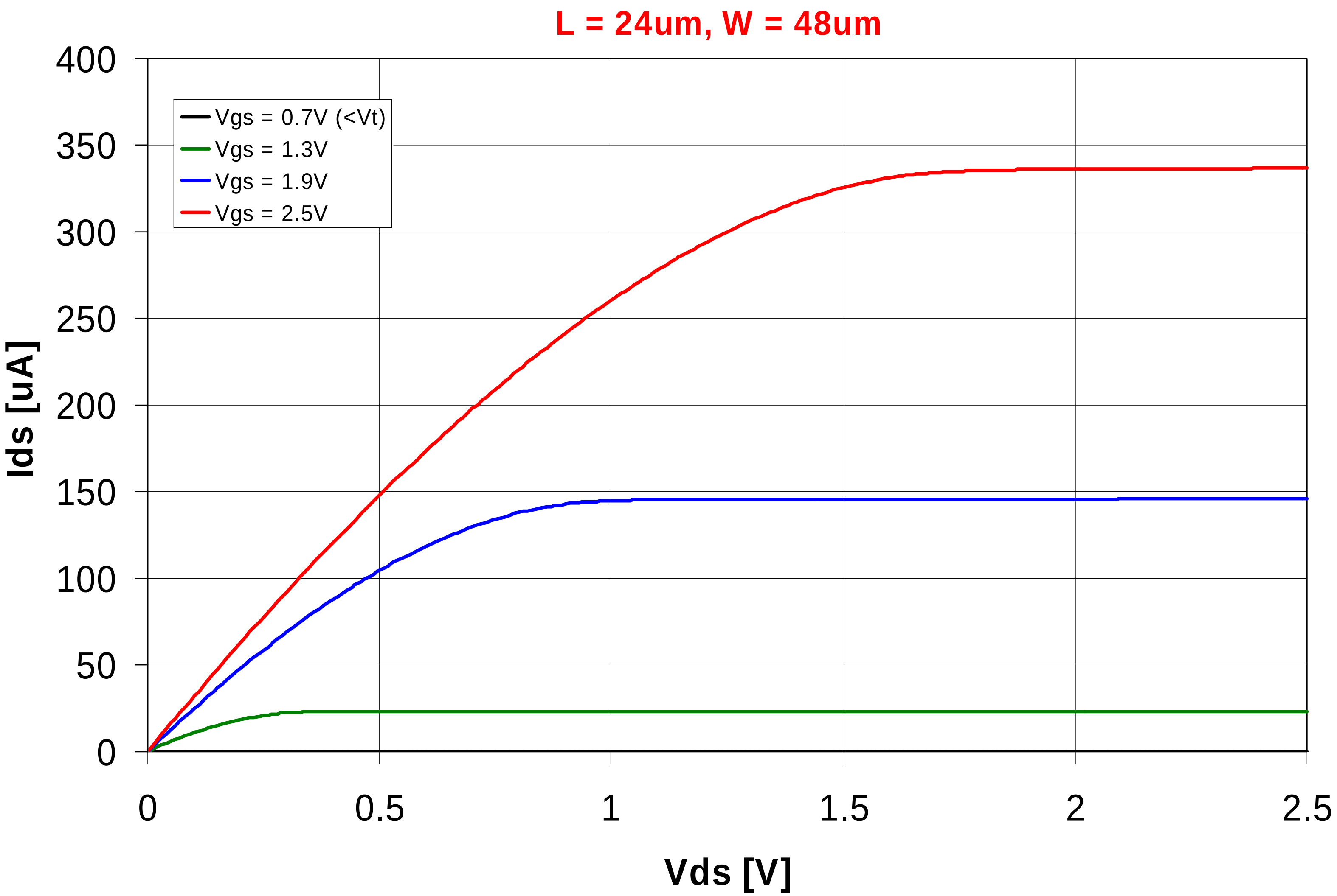




MOS output characteristics

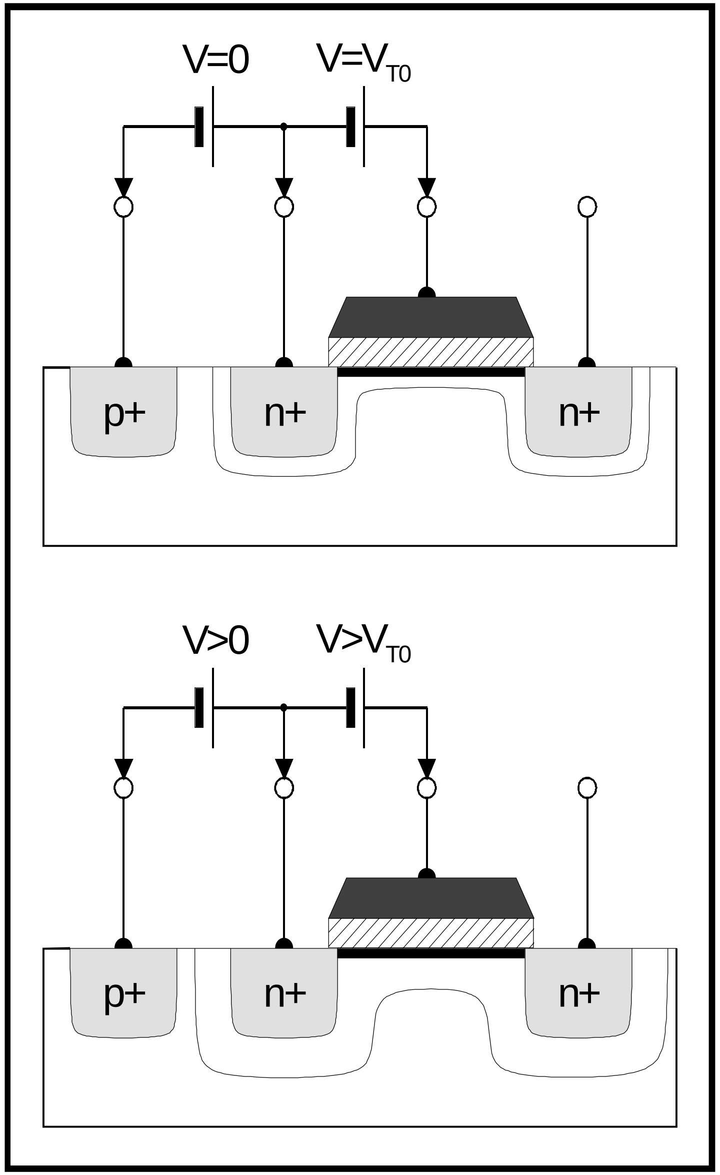


MOS output characteristics

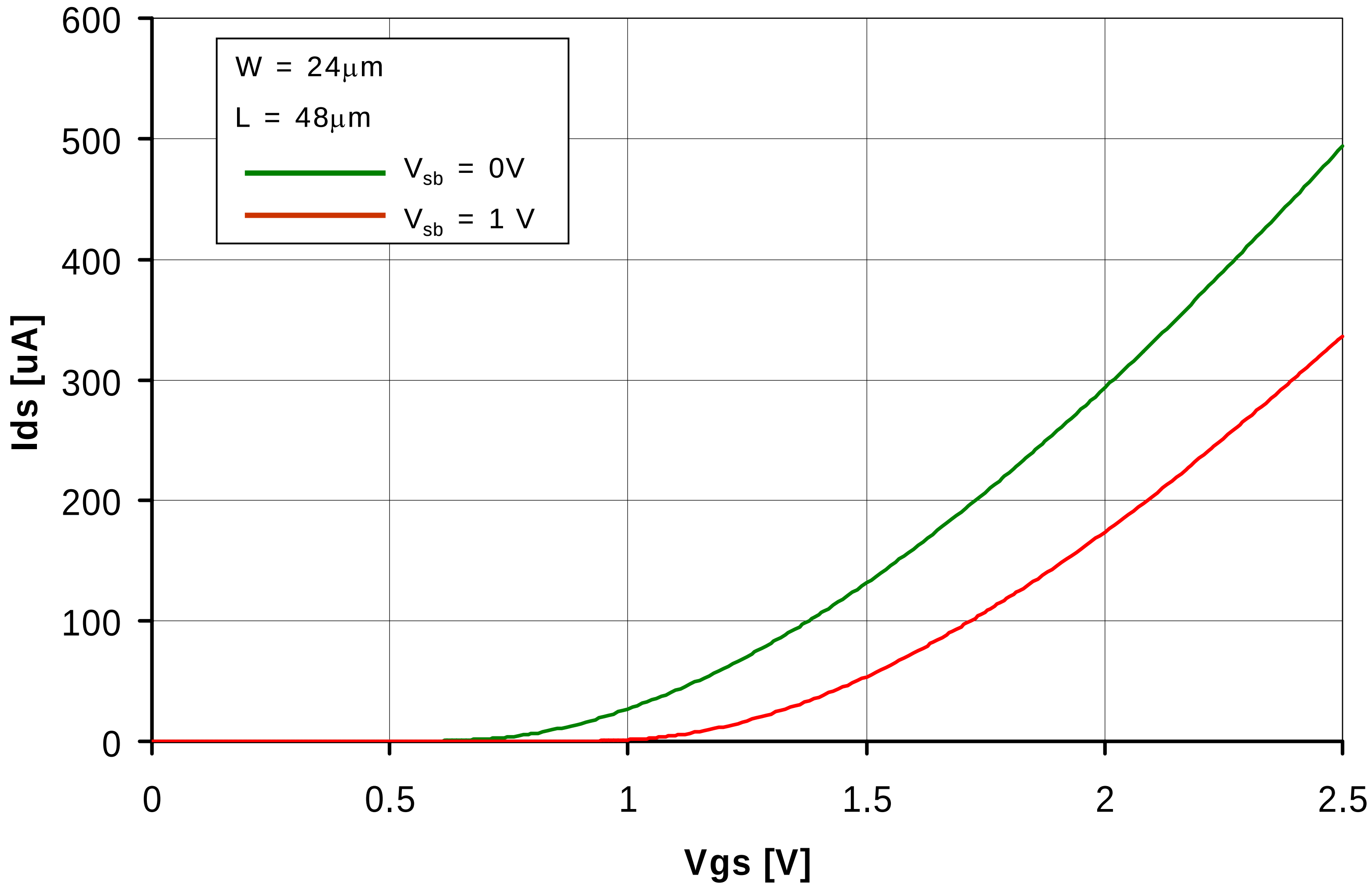


Bulk effect

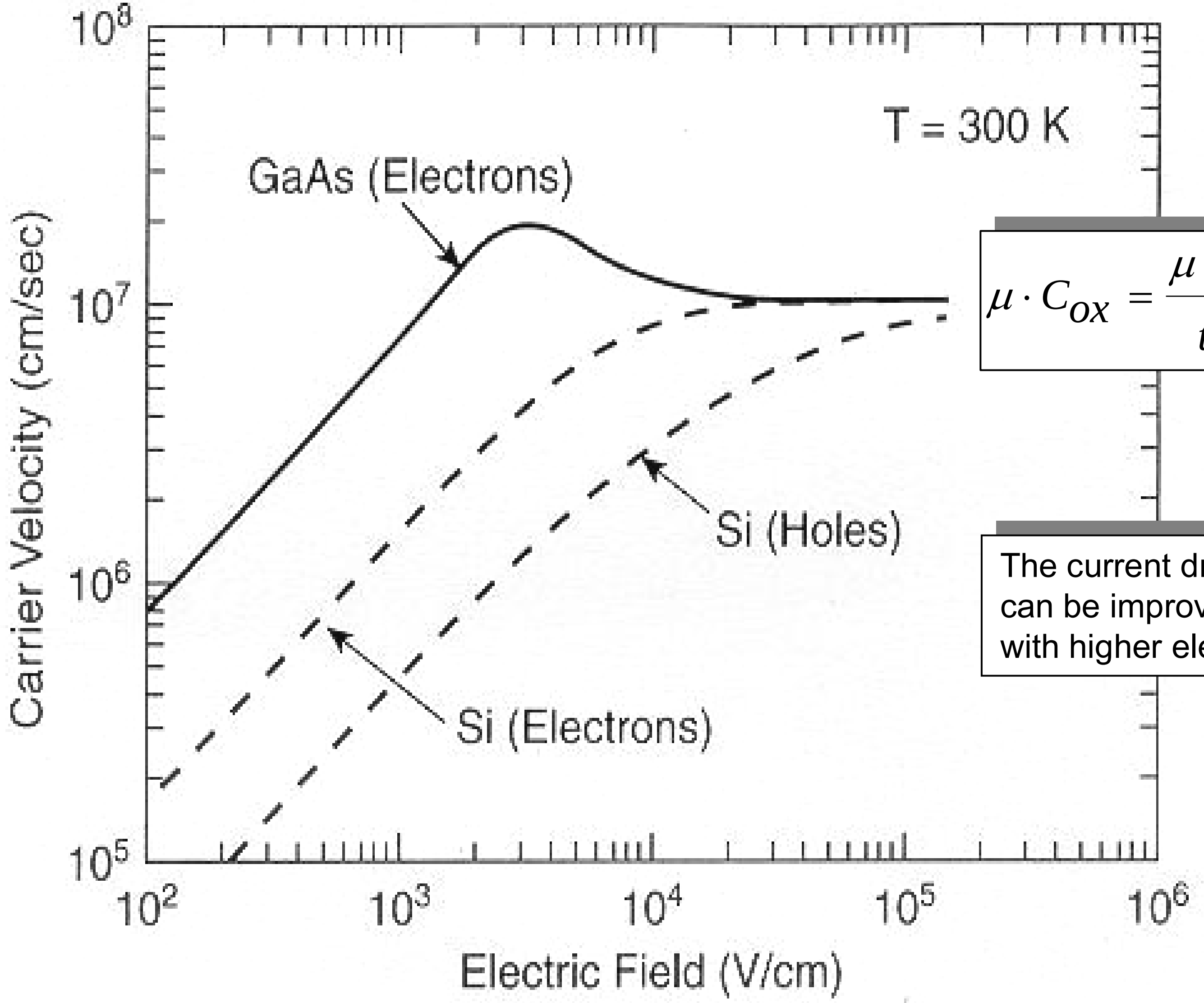
- The threshold depends on:
  - Gate oxide thickness
  - Doping levels
  - Source-to-bulk voltage
- When the semiconductor surface inverts to n-type the channel is in “strong inversion”
- $V_{sb} = 0 \Rightarrow$ strong inversion for:
  - surface potential > $-2\phi_F$
- $V_{sb} > 0 \Rightarrow$ strong inversion for:
  - surface potential > $-2\phi_{F} + V_{sb}$



Bulk effect


W = 24μm
L = 48μm
V_{sb} = 0V
V_{sb} = 1 V
Ids [uA]
600
500
400
300
200
100
0
0
0.5
1
1.5
2
2.5
Vgs [V]

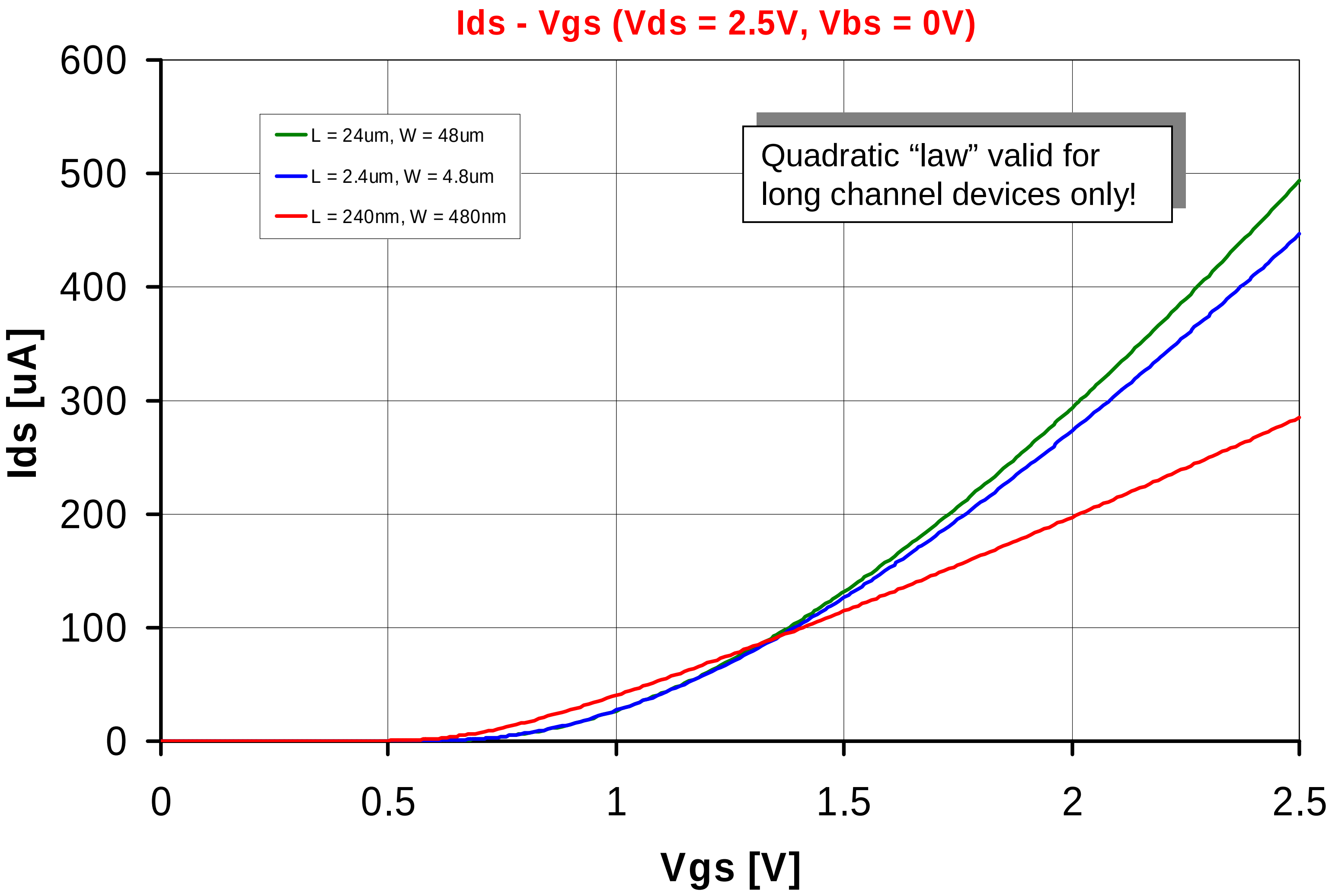
Mobility



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

The current driving capability can be improved by using materials with higher electron mobility

Is the quadratic law valid?



Quadratic "law" valid for long channel devices only!

Weak inversion

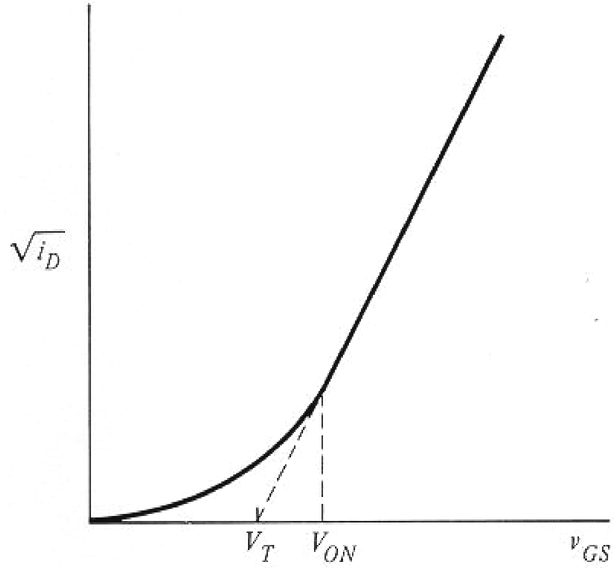
- Is $I_d=0$ when $V_{gs}<V_T$?
- For $V_{gs}<V_T$ the drain current depends exponentially on V_{gs}
- In weak inversion and saturation (V_{ds} > ~150mV):

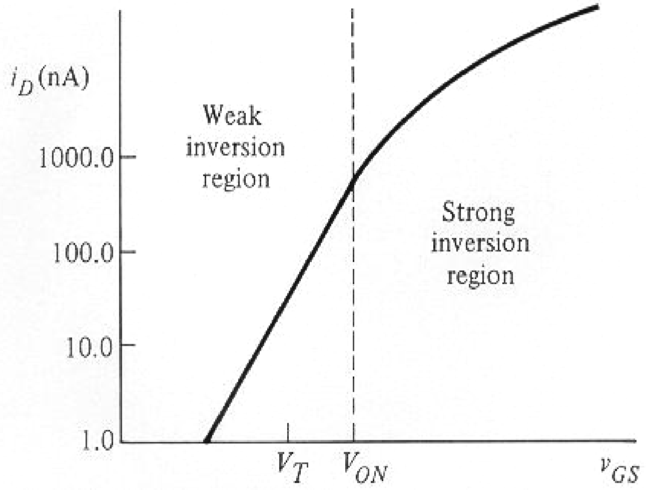
$$I_d \cong \frac{W}{L} \cdot I_{do} \cdot e^{\frac{q \cdot V_{gs}}{n \cdot k \cdot T}}$$

where

$$I_{do} = e^{-\frac{q \cdot V_T}{n \cdot k \cdot T}}$$

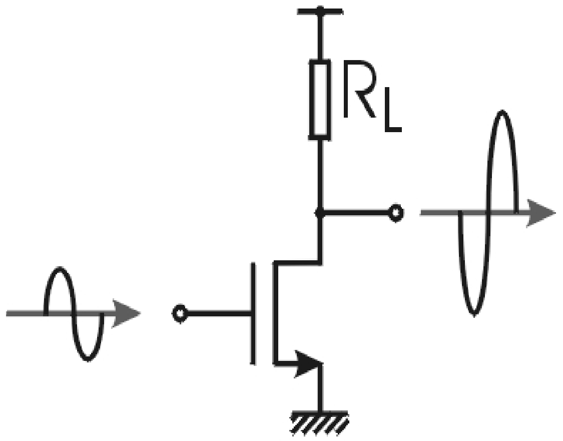
- Used in very low power designs
- Slow operation



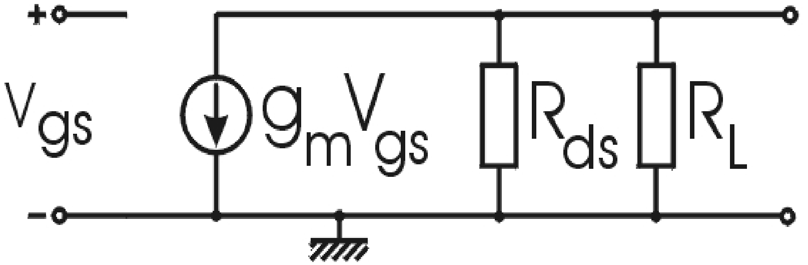


Gain & Inversion

- Gain:
  - Signal regeneration at every logic operation
  - “Static” flip-flops
  - “Static” RW memory cells
- Inversion:
  - Intrinsic to the common-source configuration
- The gain cell load can be:
  - Resistor
  - Current source
  - Another gain device (PMOS)

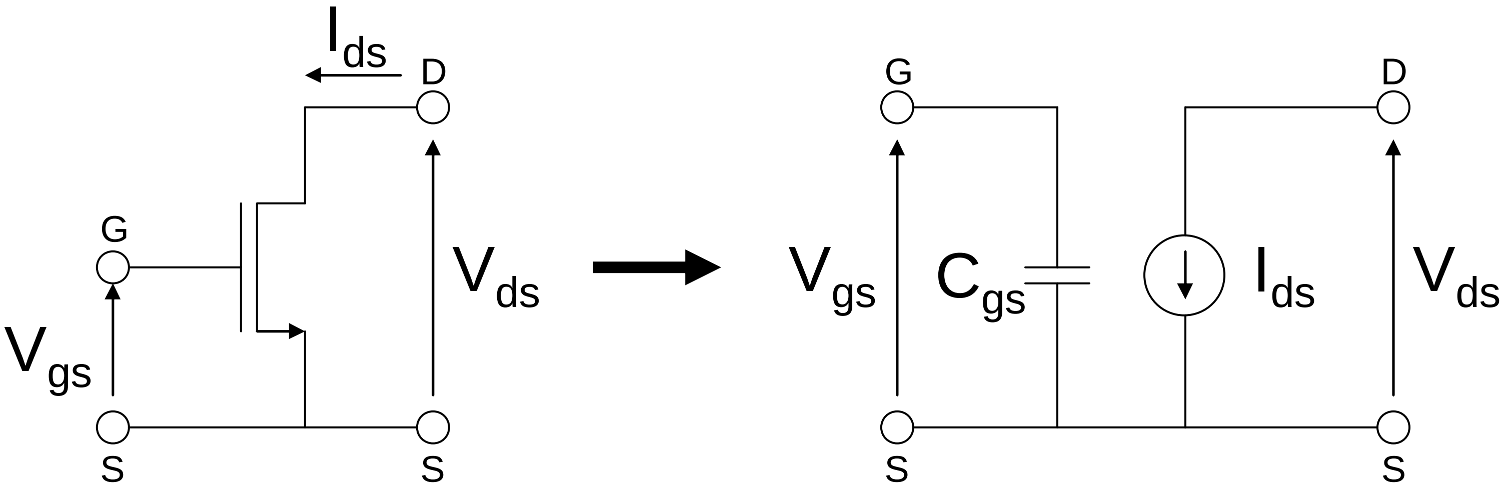


$$V_{out} = -\,g_m \times (R_{ds} \,\|\, R_L) \times V_{in}$$



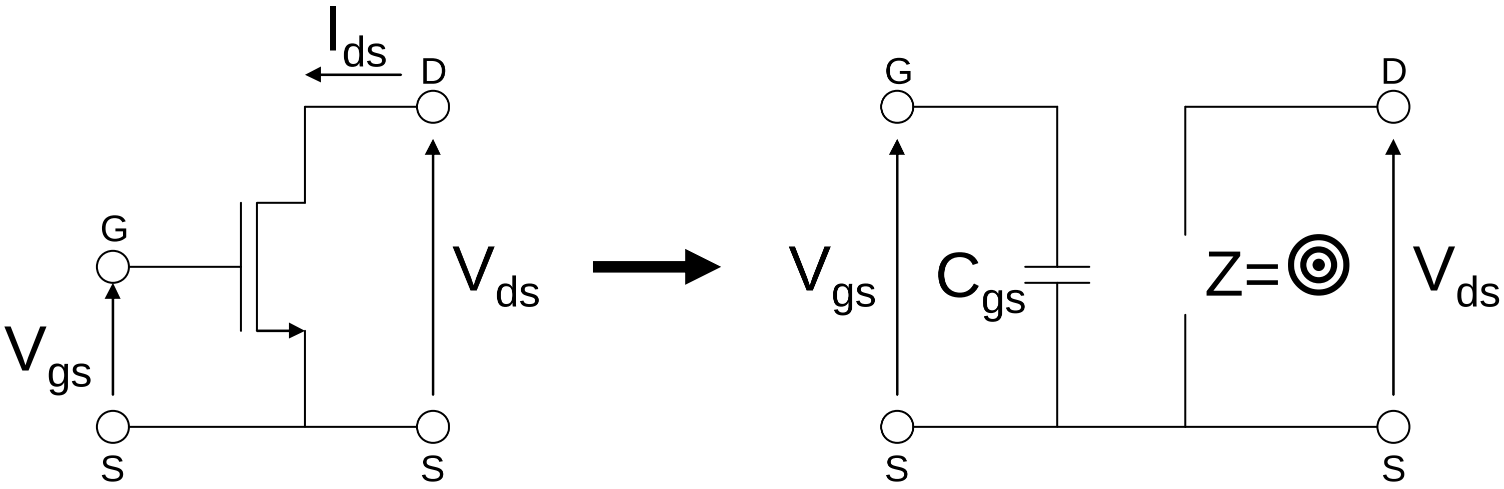
Simple MOS model for digital designers

- The MOS transistor "is" a capacitor whose voltage controls the passage of current between two nodes called the ***source*** and the ***drain***.
- One of the electrodes of this capacitor is called the ***gate***, the other the ***source***.
- The "way" the current flows between the source and the drain depends on the gate-to-source voltage (V_{gs}) and on the drain-to-source voltage (V_{ds}).



Simple model

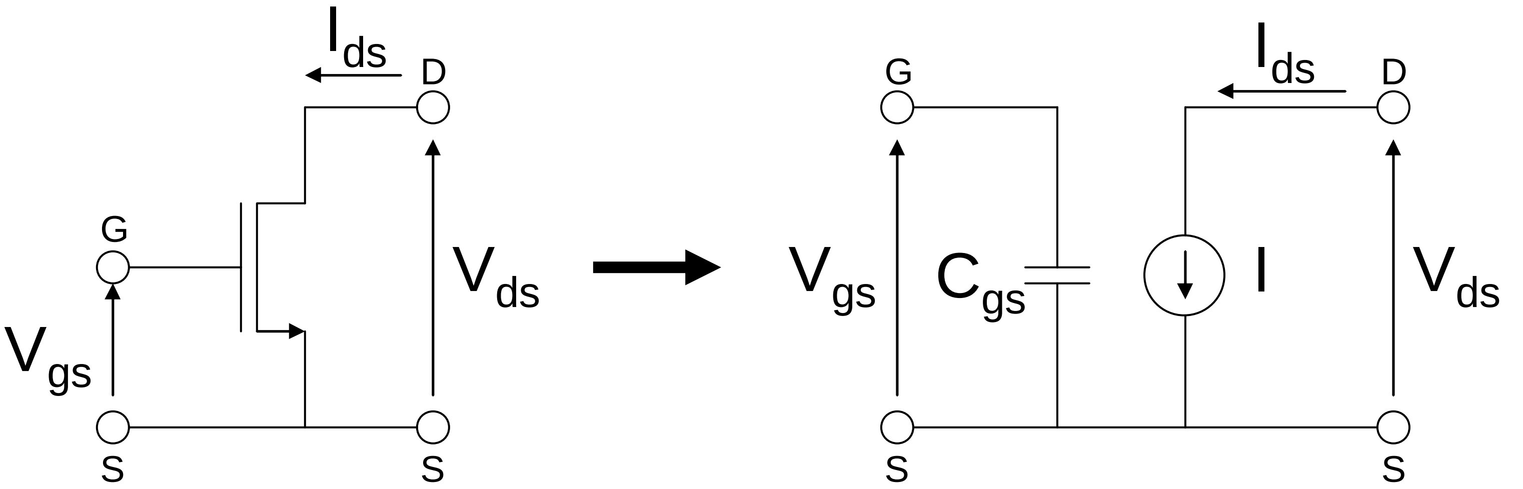
- If the gate-to-source voltage (V_{gs}) is less than a certain voltage, called the threshold voltage (V_{th}), no current flows in the drain circuit no matter what the drain-to-source voltage (V_{ds}) is!
- This is the actual definition of ***threshold voltage V_{th}.***
- That is, $I_{ds} = 0$ for $V_{gs} < V_{th}$
- This is the same as saying that the drain circuit is an infinite impedance (an open circuit)!



Simple model

- If the gate-to-source voltage (V_{gs}) is bigger than the threshold voltage (V_{th}) ***<u>and</u>*** the drain-to-source voltage (V_{ds}) is <u>bigger</u> than $V_{gs} – V_{th}$ then the drain current only depends on the ***<u>gate overdrive voltage</u>*** ($V_{gs} – V_{th}$)
- That is, the drain circuit behaves as an “ideal” voltage controlled current source:

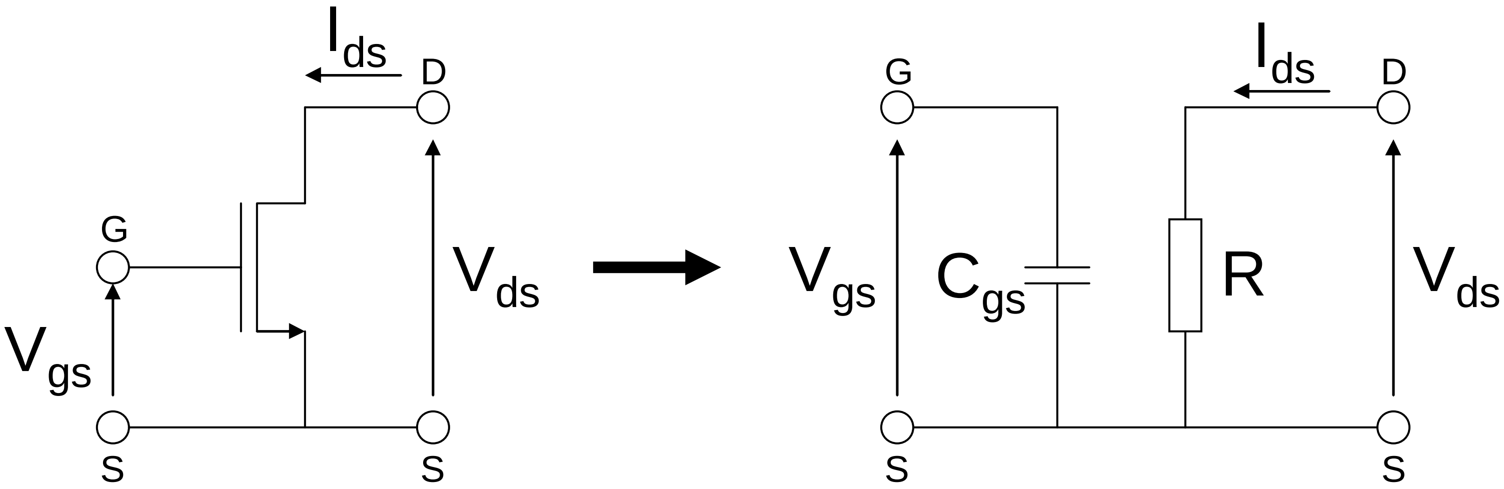
$$I_{ds} = \frac{\mu \cdot C_{ox}}{2} \cdot \frac{W}{L} \cdot \left(V_{gs} - V_T\right)^2$$



Simple model

- If the gate-to-source voltage (V_{gs}) is bigger than the threshold voltage (V_{th}) ***and*** the drain-to-source voltage (V_{ds}) is smaller than $V_{gs} – V_{th}$ then the drain current depends both on the ***gate overdrive voltage*** ($V_{gs} – V_{th}$) and the ***drain-to-source voltage*** (V_{ds})
- That is, the drain circuit behaves as a voltage controlled resistor:

$$I_{ds} = \mu \cdot C_{ox} \cdot \frac{W}{L} \cdot \left(V_{gs} - V_T\right) \cdot V_{ds}$$



Simple model

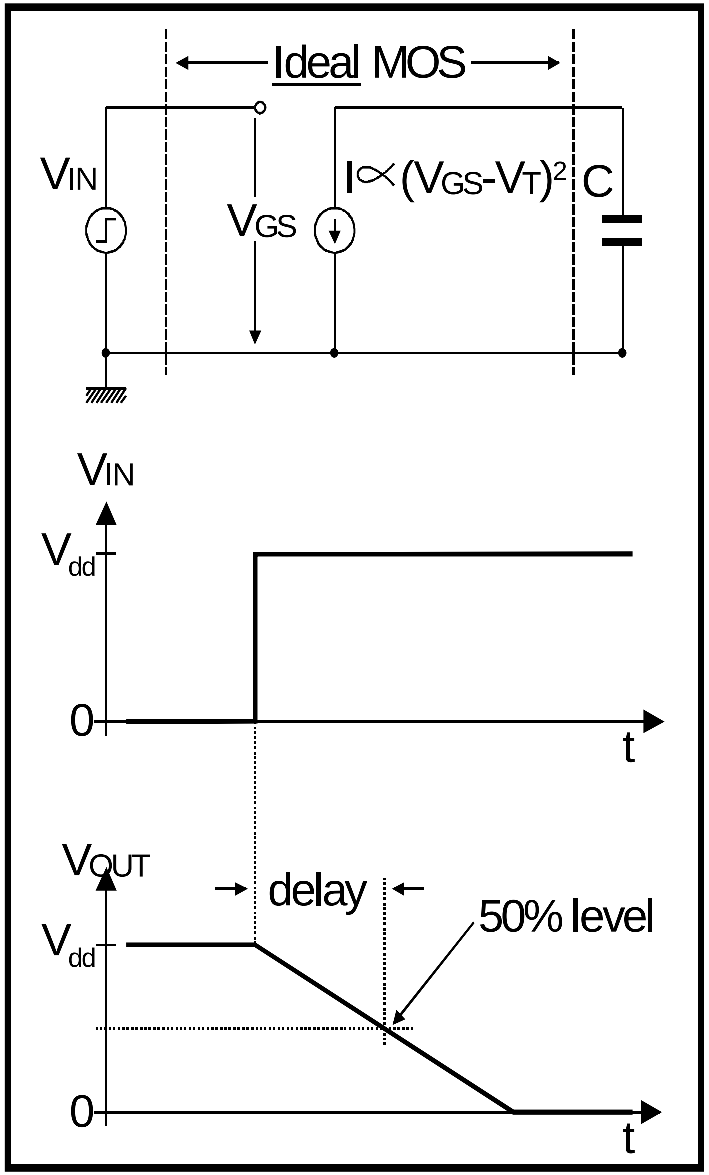
- For PMOS transistors the same concepts are valid except that:
  - All voltages are negative (including V_{th})
  - Were we used bigger than you should use smaller than
  - The drain current actually flows out of the transistor instead of into the transistor.
- REMEMBER!
  - This is a very simplistic model of the device!
  - For detailed analog simulations one relies on complicated SPICE models with many parameters defined by the technology and transistor size
- However, it will allow us to understand qualitatively the behaviour of CMOS logic circuits!
  - Even some conclusions will be based on such a simple model.

We digital designers care about delays

- In MOS circuits capacitive loading is the main cause. (RC delay in the interconnects will be addressed latter)
- Capacitance loading is due to:
  - Device capacitance
  - Interconnect capacitance

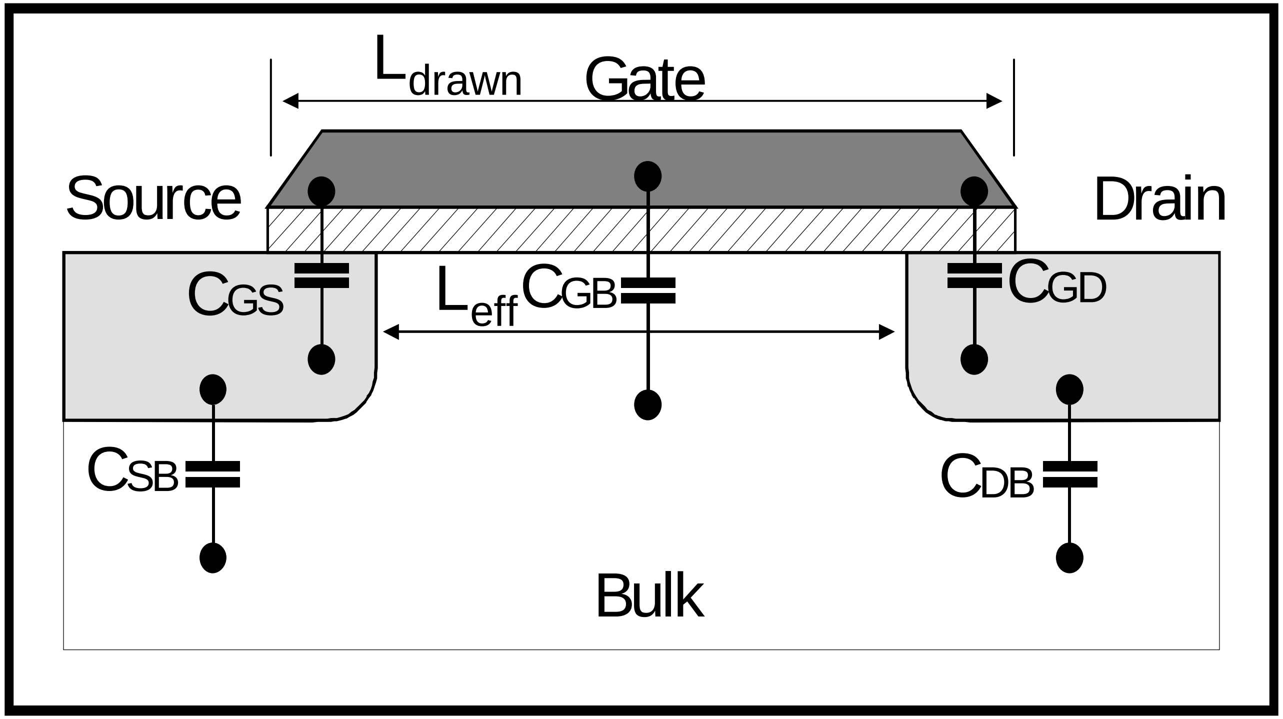
$$\Delta t = C \cdot \frac{\Delta V}{I} \approx \frac{C}{\mu \cdot C_{ox} \cdot V_{dd}} \cdot \frac{L}{W}$$

Assuming $V_T = 0$



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

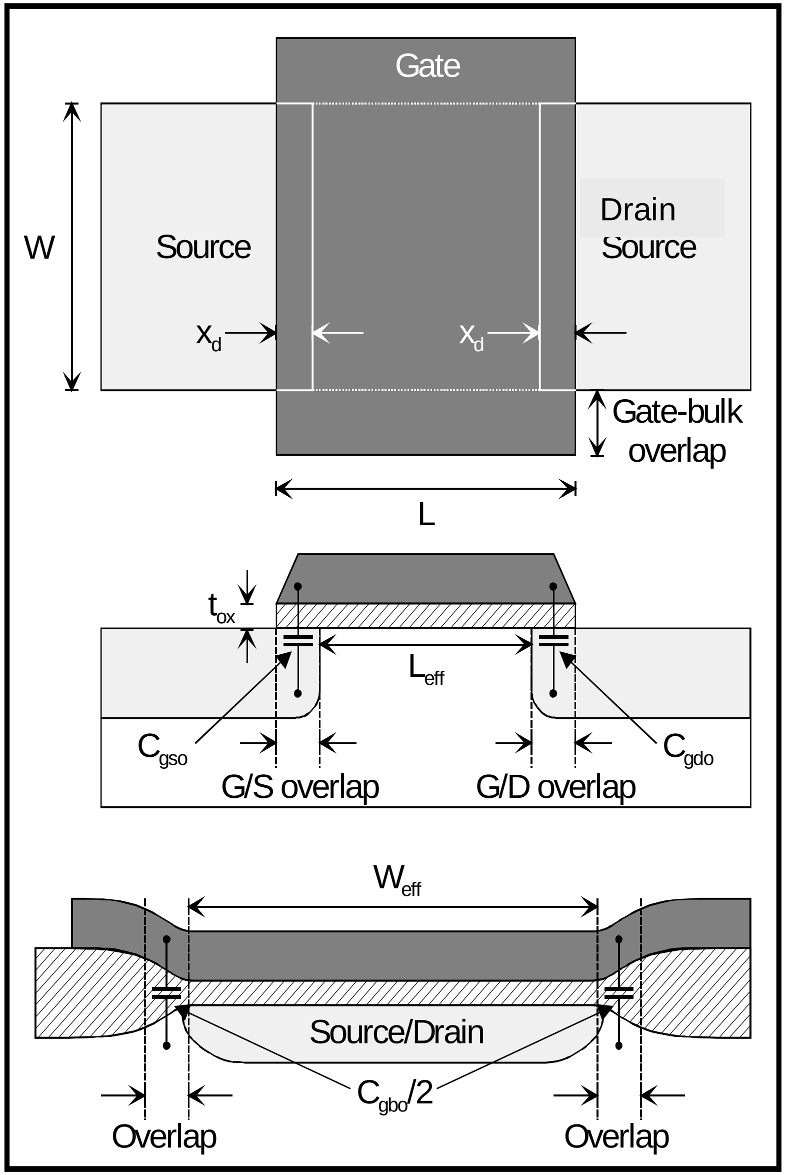
- This gives origin to the source/drain overlap capacitances:

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

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

- Gate-bulk overlap capacitance:

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



MOS structure capacitances

0.24 μm process

NMOS

L(drawn) = 0.24 μm

L(effective) = 0.18 μm

W(drawn) = 2 μm

C_o (s, d, b) = 0.36 fF/μm

$C_{ox} = 5.6$ fF/μm^2

$C_{gso} = C_{gdo} = 0.72$ fF

$C_{gbo} = 0.086$ fF

$C_g = 2.02$ fF

Note that: ***For small L devices*** the overlap capacitances are becoming as important as the “intrinsic” gate capacitance ($C_g = W_{eff} \times L_{eff} \times C_{ox}$)

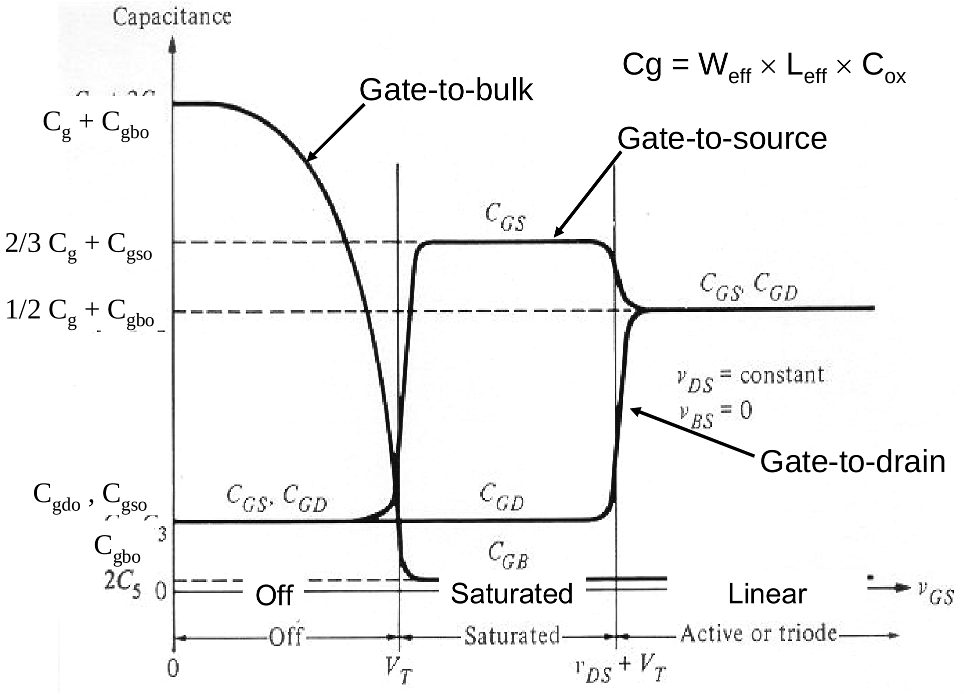
Channel capacitance

- 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



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

Operation region	$\mathbf{C_{gb}}$	$\mathbf{C_{gs}}$	$\mathbf{C_{gd}}$
Cutoff	$C_{ox}\ W\ L_{eff}$	0	0
Linear	0	$(1/2)\ C_{ox}\ W\ L_{eff}$	$(1/2)\ C_{ox}\ W\ L_{eff}$
Saturation	0	$(2/3)\ C_{ox}\ W\ L_{eff}$	0

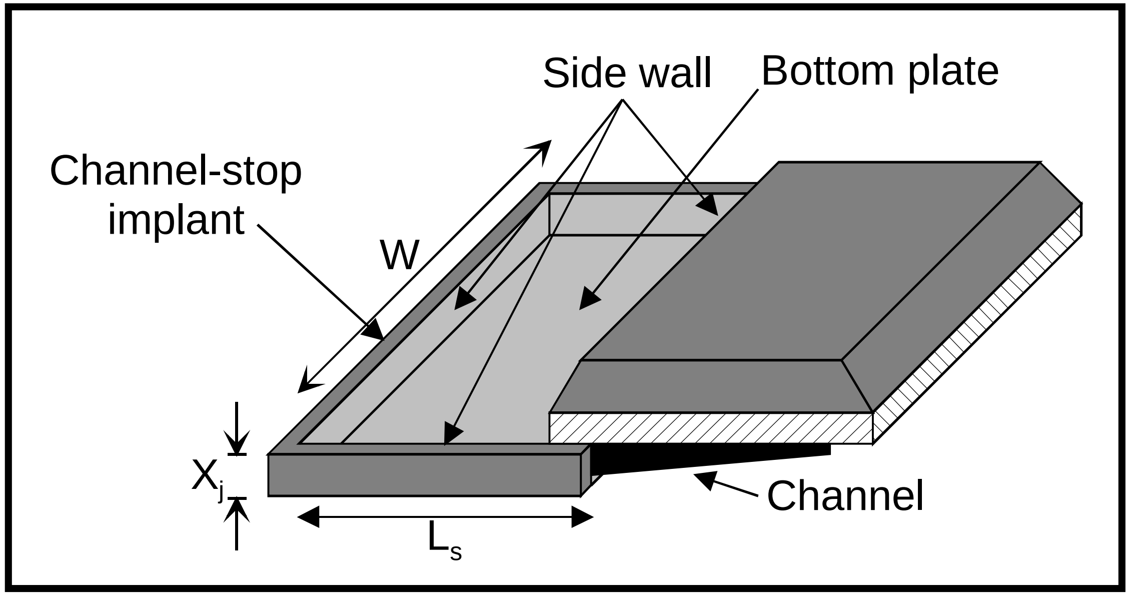
Junction capacitances

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

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

  - Side-wall capacitance:

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



<u>0.24 µm process</u>

NMOS
L(drawn) = 0.24 µm
L(effective) = 0.18 µm
W(drawn) = 2 µm
L_s = 0.8 µm
C_j (s, d) = 1.05 fF/µm^2
C_{jsw} = 0.09 fF/µm

C_{bottom} = 1.68 fF
C_{sw} = 0.32 fF

C_g = 2.02 fF

Note that: ***<u>For small L devices</u>*** the junction capacitances are becoming as important as the “intrinsic” gate capacitance ($C_g = W_{eff} \times L_{eff} \times C_{ox}$)

“Building a full MOS model

- MOS process parasitics
- pn-Junction diodes
- Depletion capacitance
- Source/drain resistance
- MOS Model
- Parasitic bipolars

MOS Parasitics

In a CMOS process the devices are:

- PMOS FETs
- NMOS FETs

+ unwanted (but ubiquitous):

- pn-Junction diodes
- parasitic capacitance
- parasitic resistance

and

- parasitic bipolars
- *parasitic inductance*

Parasitics or useful?

- Resistors
- Capacitors
- Inductors
- Diodes
- Bipolar transistors

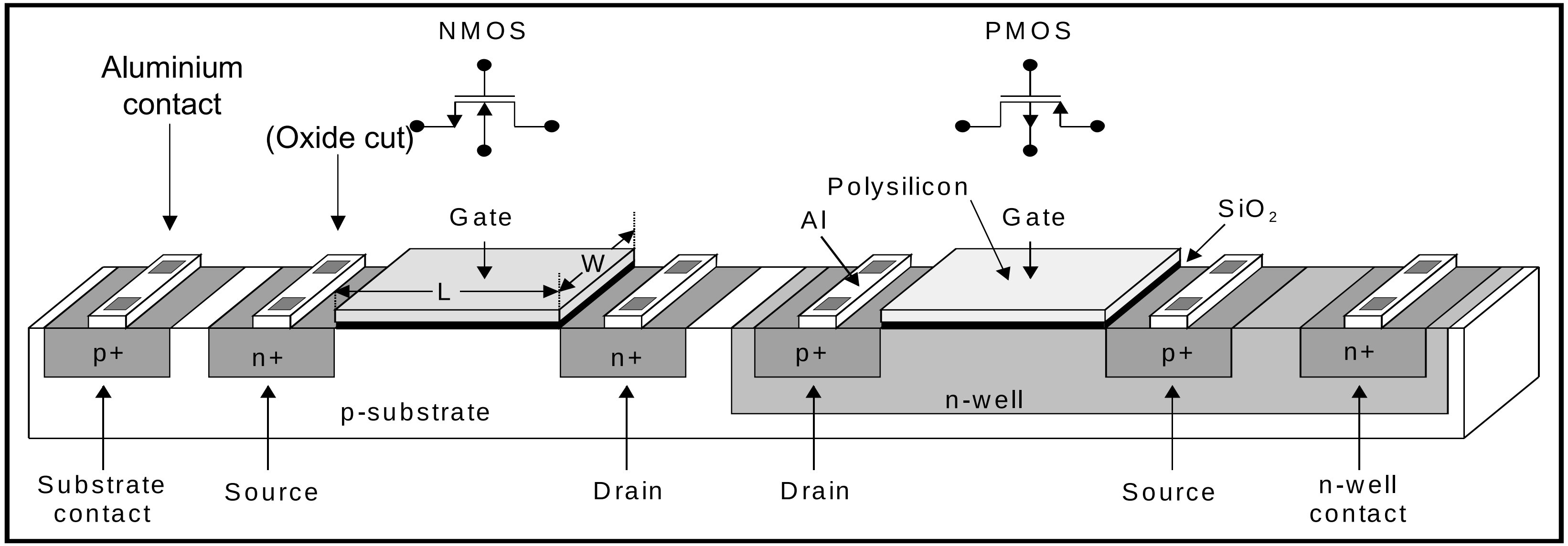
Are useful circuit elements for analogue circuit design. Some technologies offer the possibility of manufacture such devices under controlled conditions.

pn-Junction diodes

- pn – Junction diodes:
  - – Provide isolation between devices (if reversed biased)
  - – Can be used to implement:
    - band-gap circuits (if forward biased)
    - variable capacitors
    - clamping devices
    - level shifting
  - – Are extremely useful as Electro Static Discharge (ESD) protection devices.

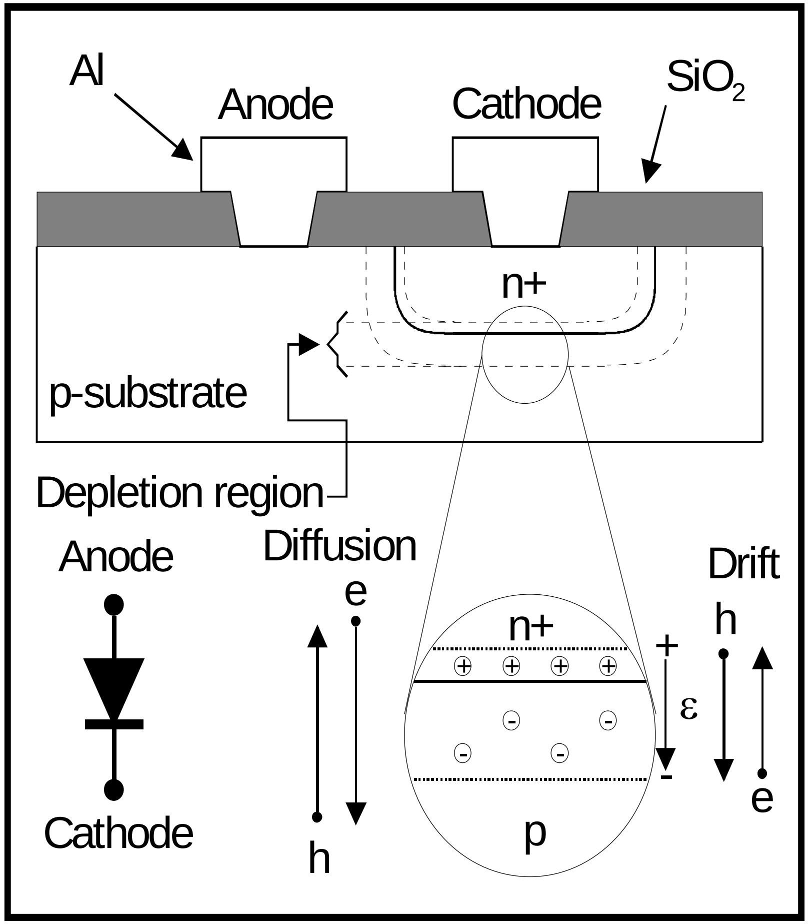
CMOS devices

- Remember:
  - Every source and drain creates a pn-junction
  - pn-junctions must be reversed biased to provide isolation between devices
  - Reversed biased pn-junctions display parasitic capacitance



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

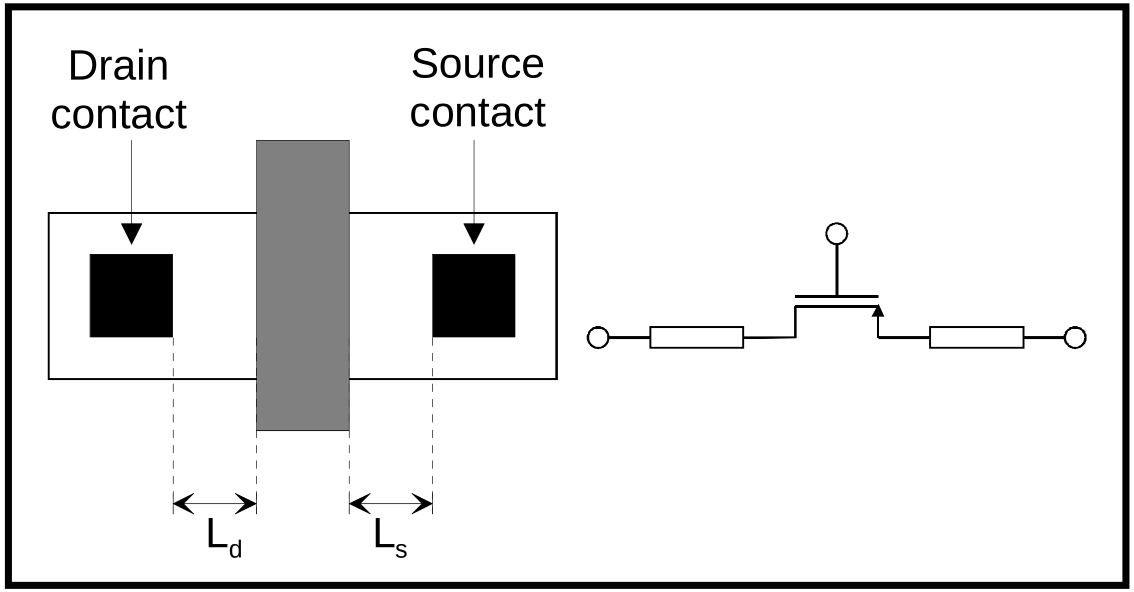


Source/drain resistance

- 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

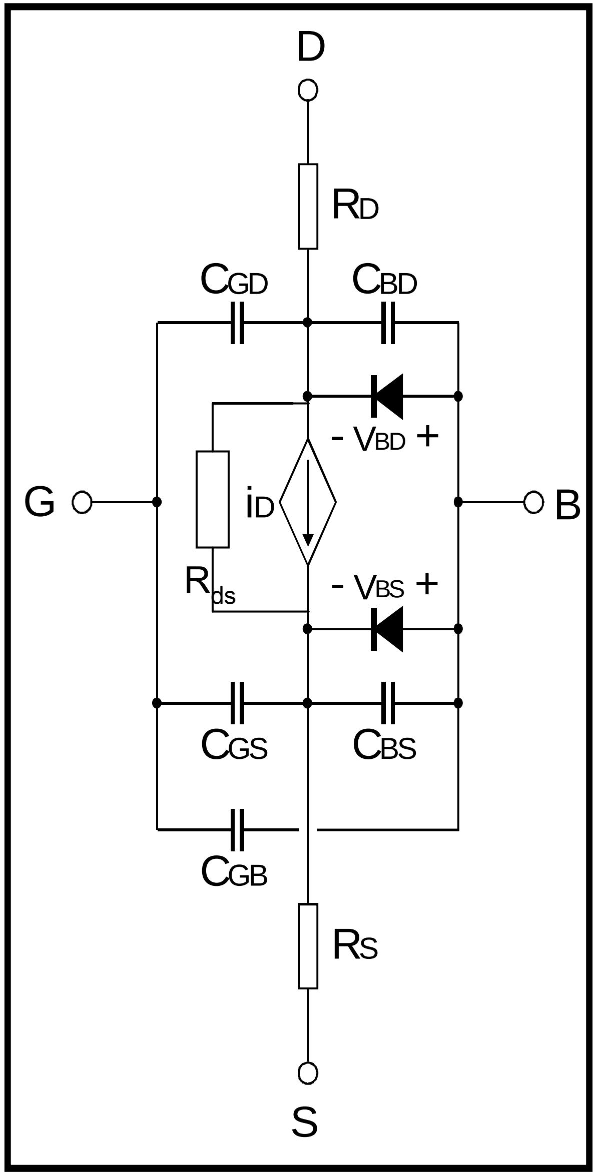


0.24 μm process

R (P+) = 4 Ω/sq
R (N-) = 4 Ω/sq

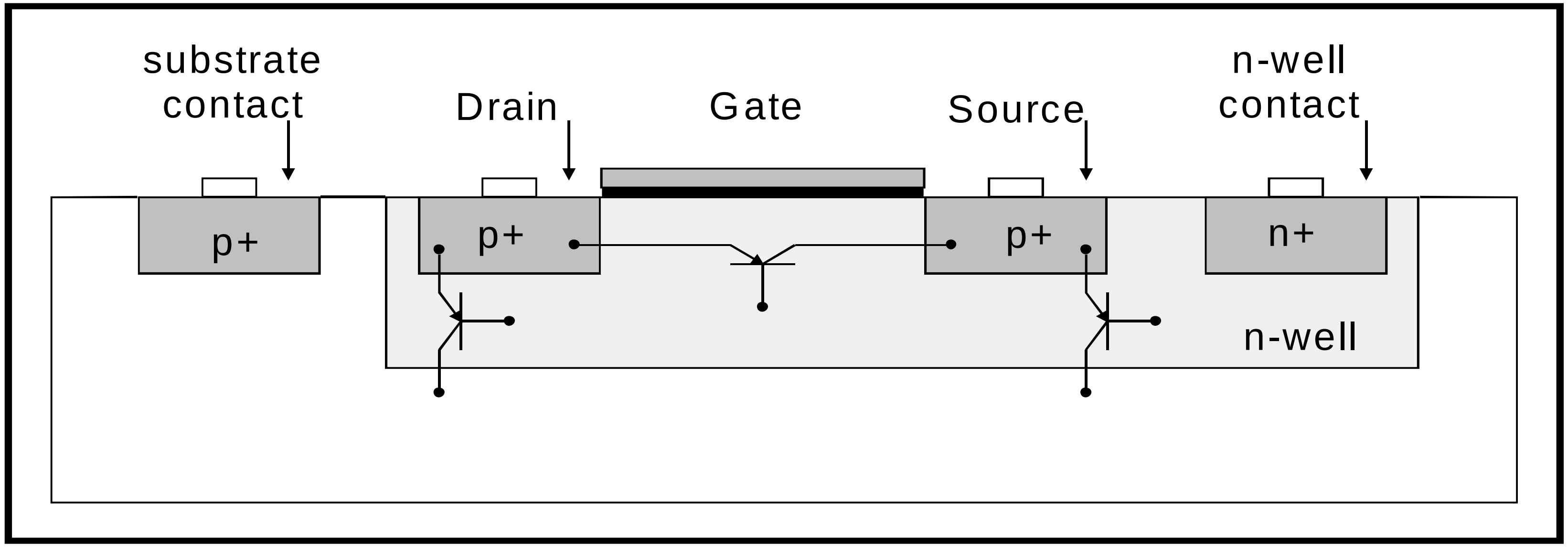
Basic MOSFET model

- For designing we rely on simulators (SPICE) with appropriate models and parameters given by technology supplier.



CMOS parasitic bipolar

- Every p-n-p or n-p-n regions form parasitic bipolar transistors.
- In standard MOS circuits these devices must be turned off.
  - If not a latchup (short circuit) can occur
- For some applications (like bandgap circuits) these devices can be used. But, better know what you are doing…
- For digital designs we "forget" about this

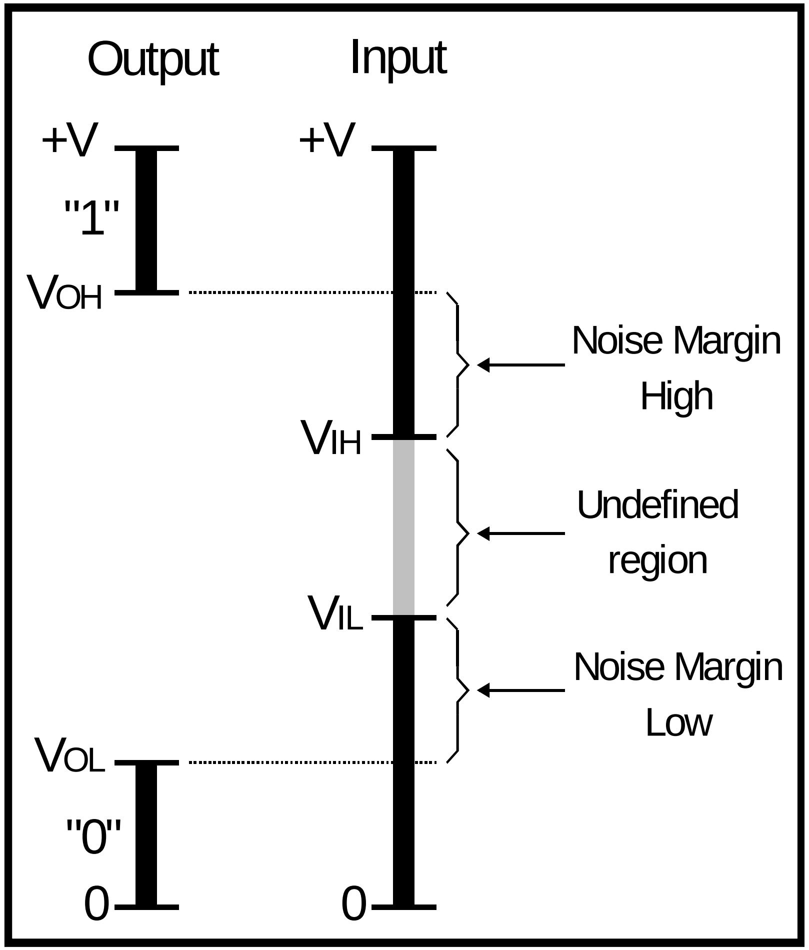


How do we make digital from this ?

- Logic levels
- MOST – a simple switch
- The CMOS inverter:
  - DC operation
  - Dynamic operation
  - Propagation delay
  - Power consumption
  - Layout

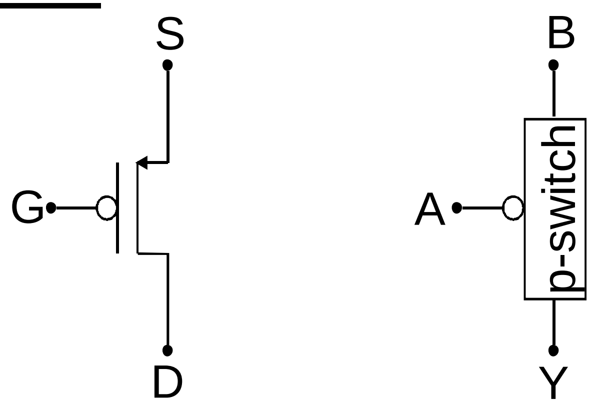
CMOS logic: “0” and “1”

- Logic circuits process Boolean variables
- Logic values are associated with voltage levels:
  - $V_{IN} > V_{IH} \Rightarrow$ “1”
  - $V_{IN} < V_{IL} \Rightarrow$ “0”
- Noise margin:
  - $NM_H = V_{OH} - V_{IH}$
  - $NM_L = V_{IL} - V_{OL}$



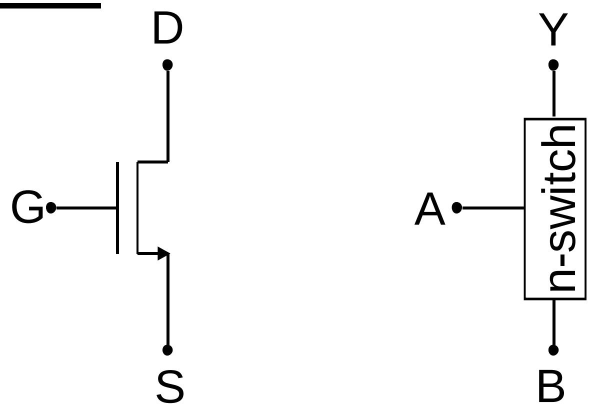
The MOST - a simple switch

p-switch



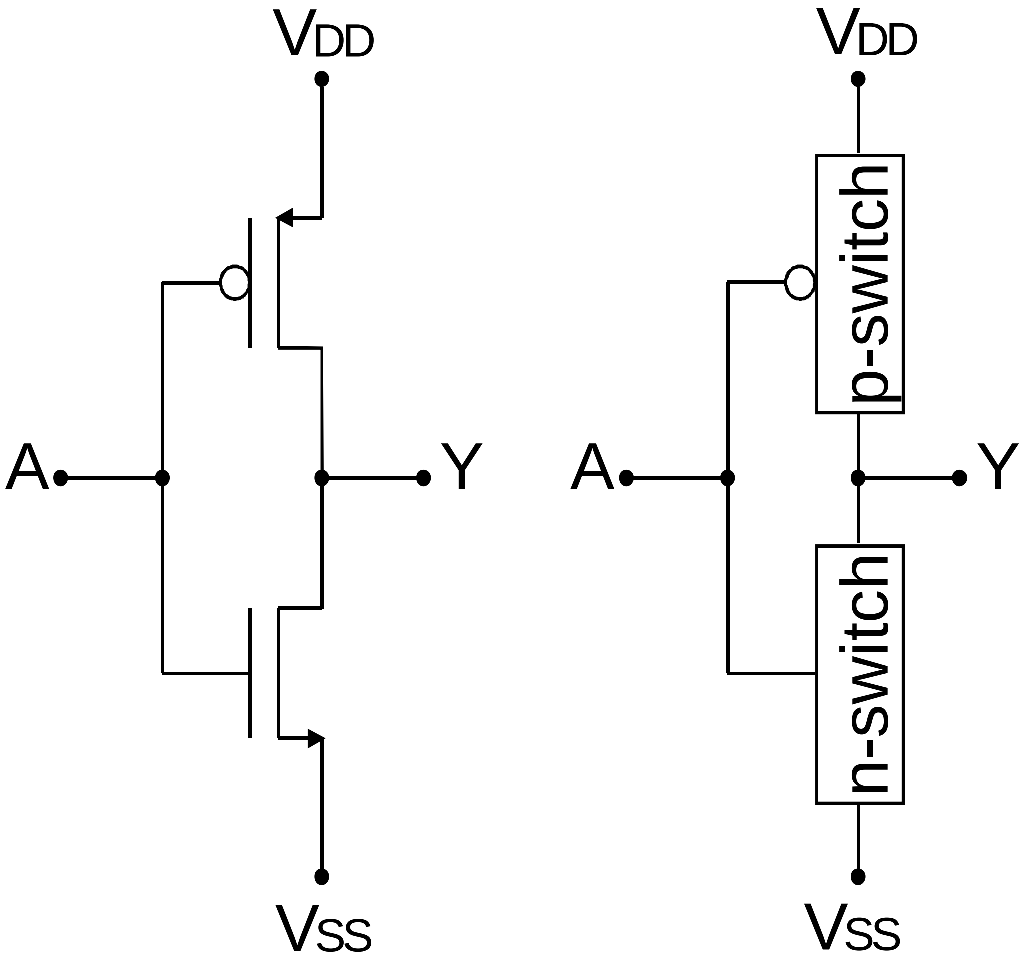
A	B	Y	
0	0	bad 0	(source follower)
0	1	**good 1**	
1	0	?	(high Z)
1	1	?	(high Z)

n-switch



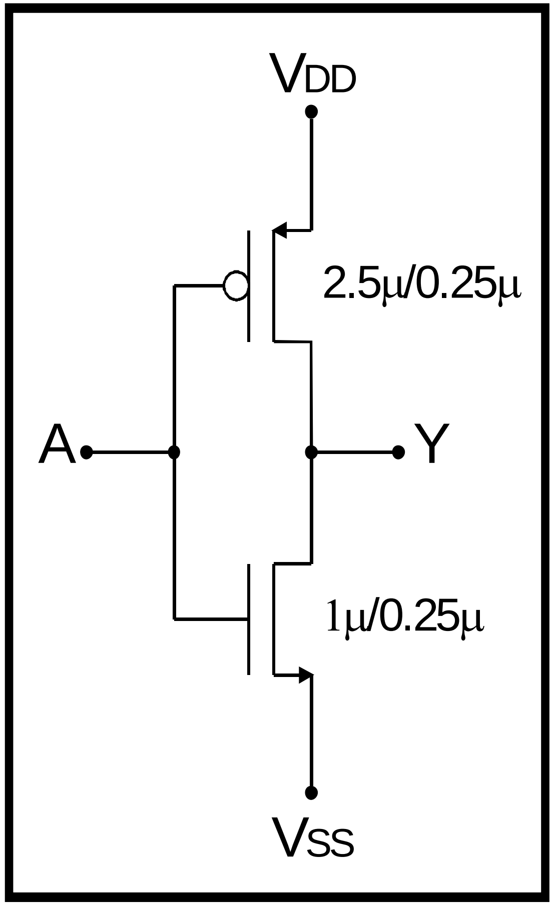
A	B	Y	
0	0	?	(high Z)
0	1	?	(high Z)
1	0	**good 0**	
1	1	bad 1	(source follower)

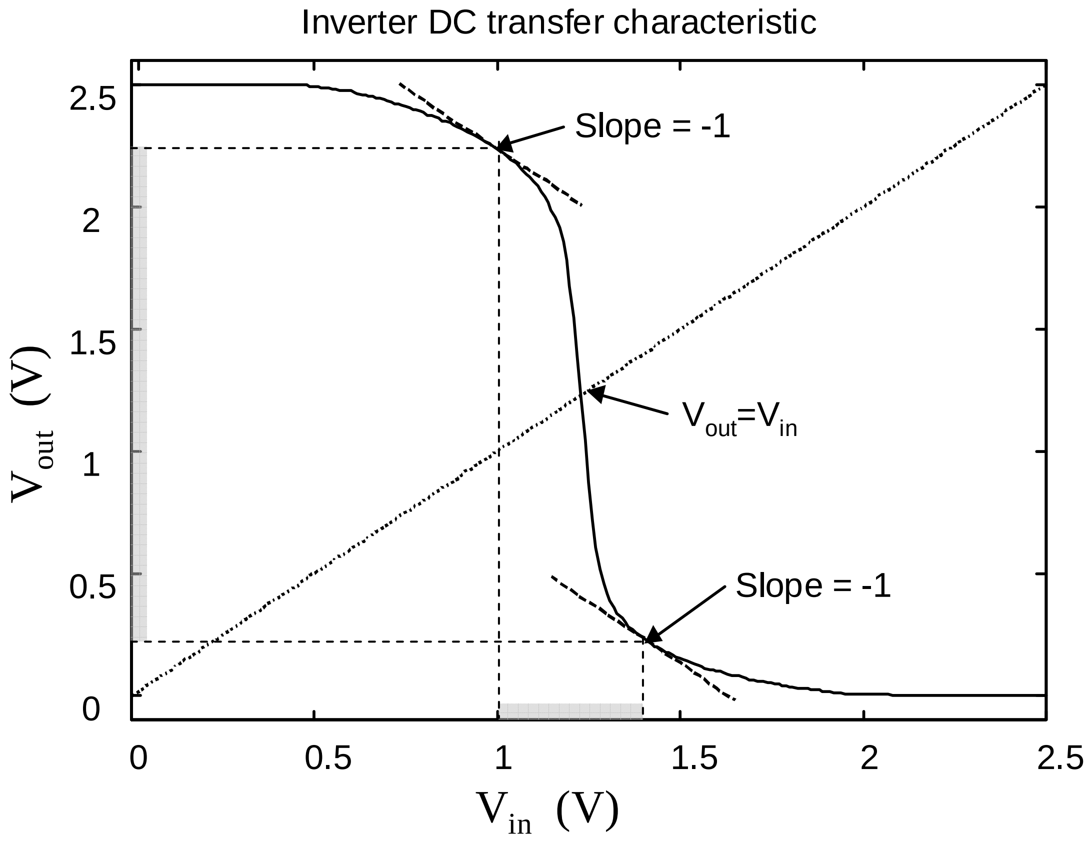
The CMOS inverter



A	Y
0	good 1
1	good 0

The CMOS inverter



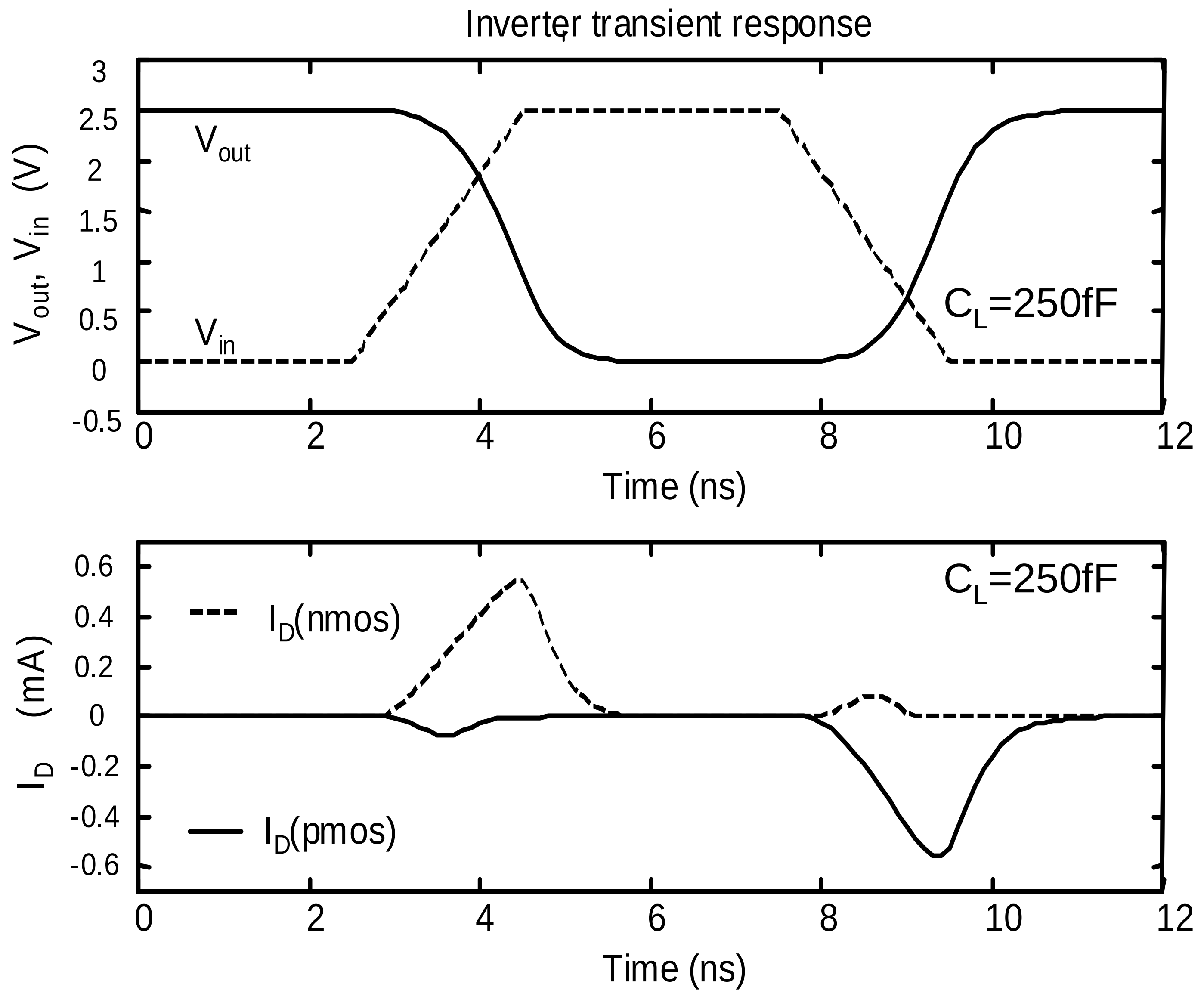


The CMOS inverter

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	$V_{dd}/2$
$V_{dd}-\|V_{TP}\|>V_{in}>V_{dd}/2$	linear	saturation	~ 0
V_{dd}	linear	cut-off	0

The CMOS inverter



The CMOS inverter

- Propagation delay
  - Main origin: load capacitance

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

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

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

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

The CMOS inverter

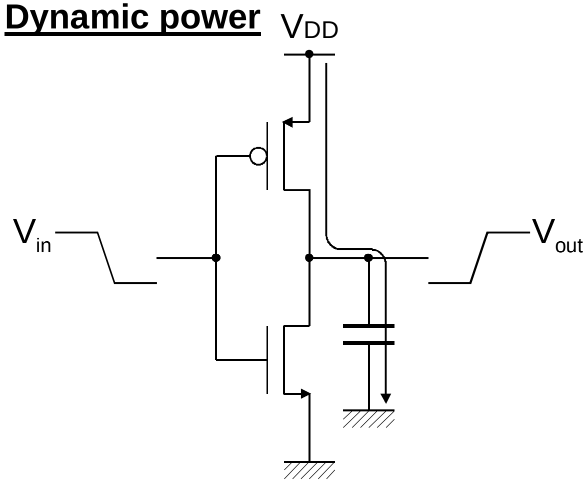
- CMOS power budget:
  - – Dynamic power consumption:
    - Charging and discharging of capacitors
  - – Short circuit currents:
    - Short circuit path between power rails during switching
  - – Leakage
    - Leaking diodes.
    - Leaking transistors:
      - – Sub-threshold currents
      - – In the future devices gate leakage current!?

The CMOS inverter

- The dynamic power dissipation is a function of:
  - Frequency
  - Capacitive loading
  - Voltage swing
- To reduce dynamic power dissipation
  - Reduce: C_L
  - Reduce: f
  - Reduce: V_{dd} ⇐ The most effective action

CMOS logic:
no static power
consumption!

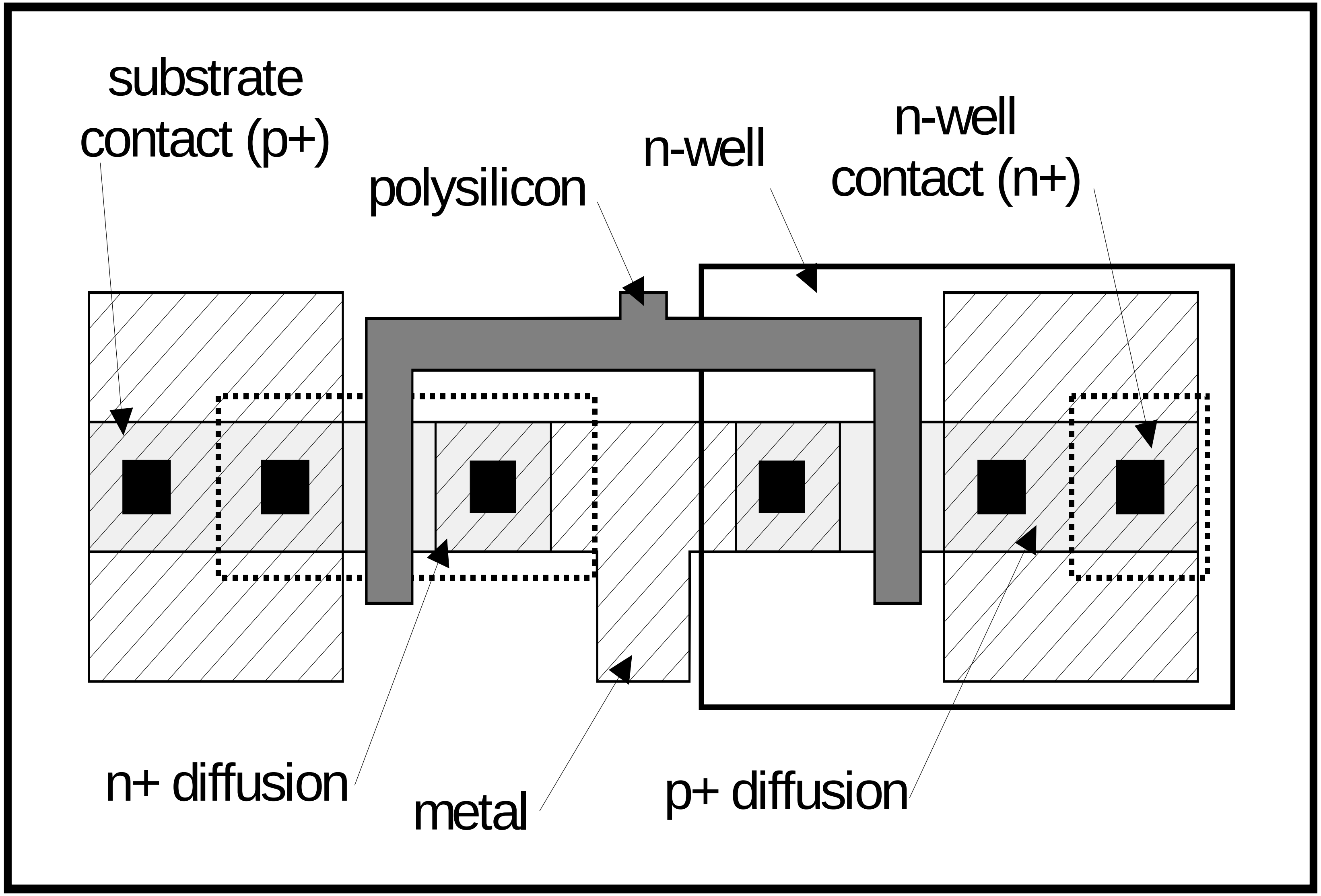
Dynamic power



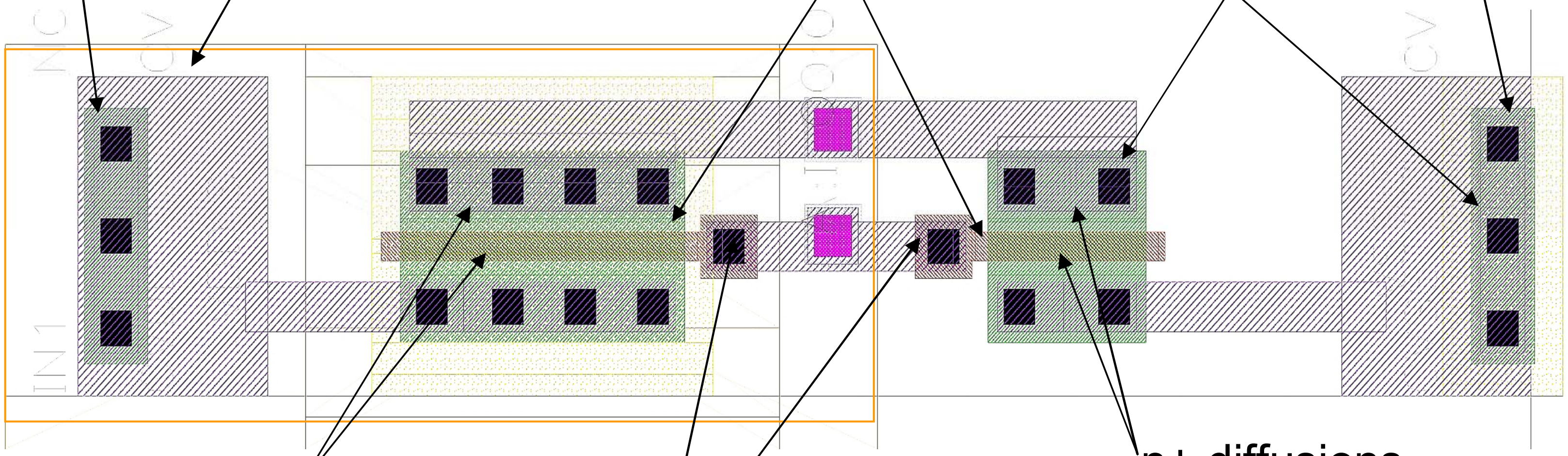
$$E = \text{Energy / transition} = \frac{1}{2} \cdot C_L \cdot V_{dd}^2$$

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

The CMOS inverter



The CMOS inverter



n-well contact (n+)

n-well

polysilicon

substrate contact (p+)

diffusion contacts

p+ diffusions

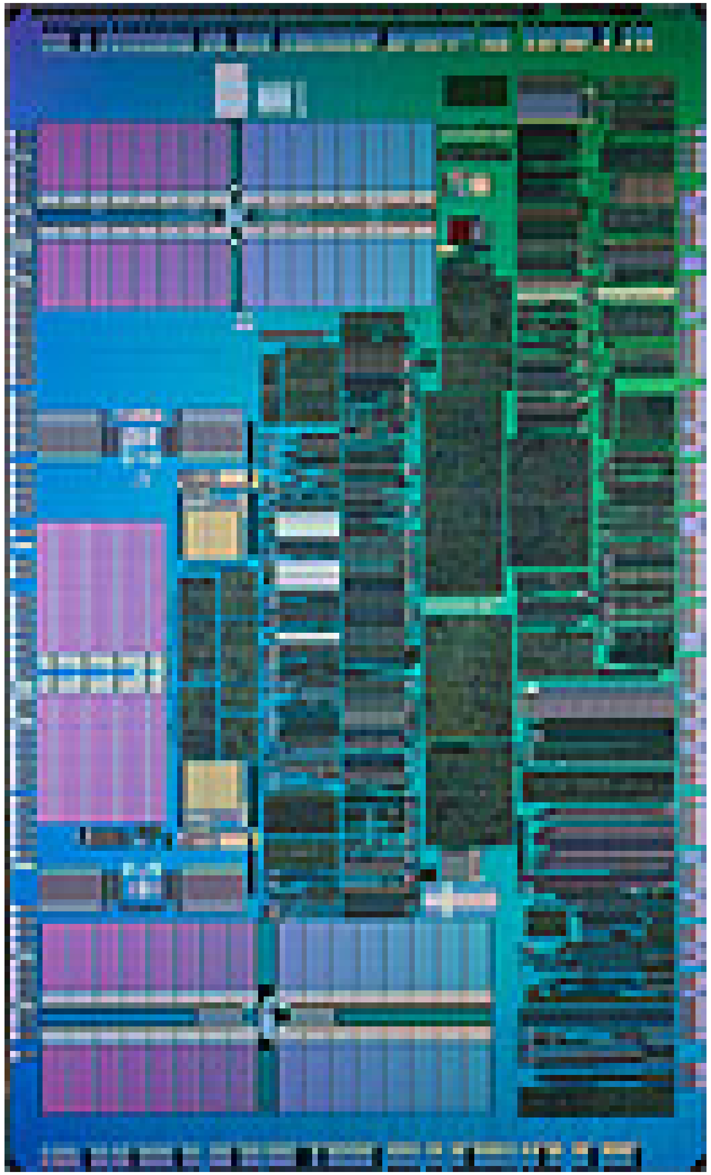
n+ diffusions

polysilicon contacts

- Scale between N and P MOS
- How to make a buffer

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

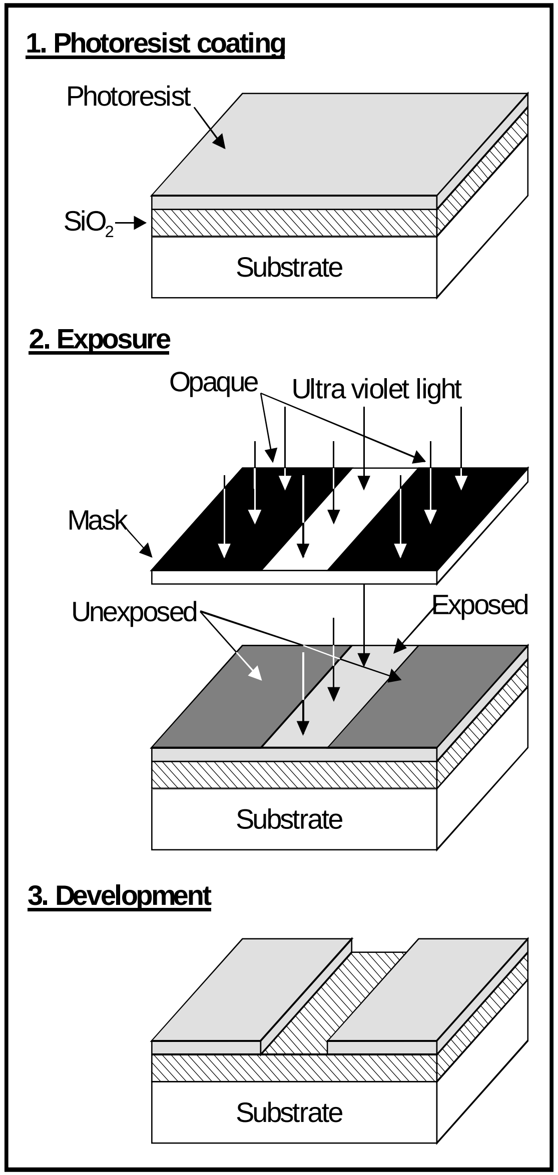
Lithography sequence steps:

- Designer:
  - Drawing the “layer” patterns on a layout editor
- Silicon Foundry:
  - Masks generation from the layer patterns in the design data base
    - Masks are not necessarily identical to the layer patterns but they are obtained from them.
  - Printing: transfer the mask pattern to the wafer surface
  - Process the wafer to physically pattern each layer of the IC

Lithography

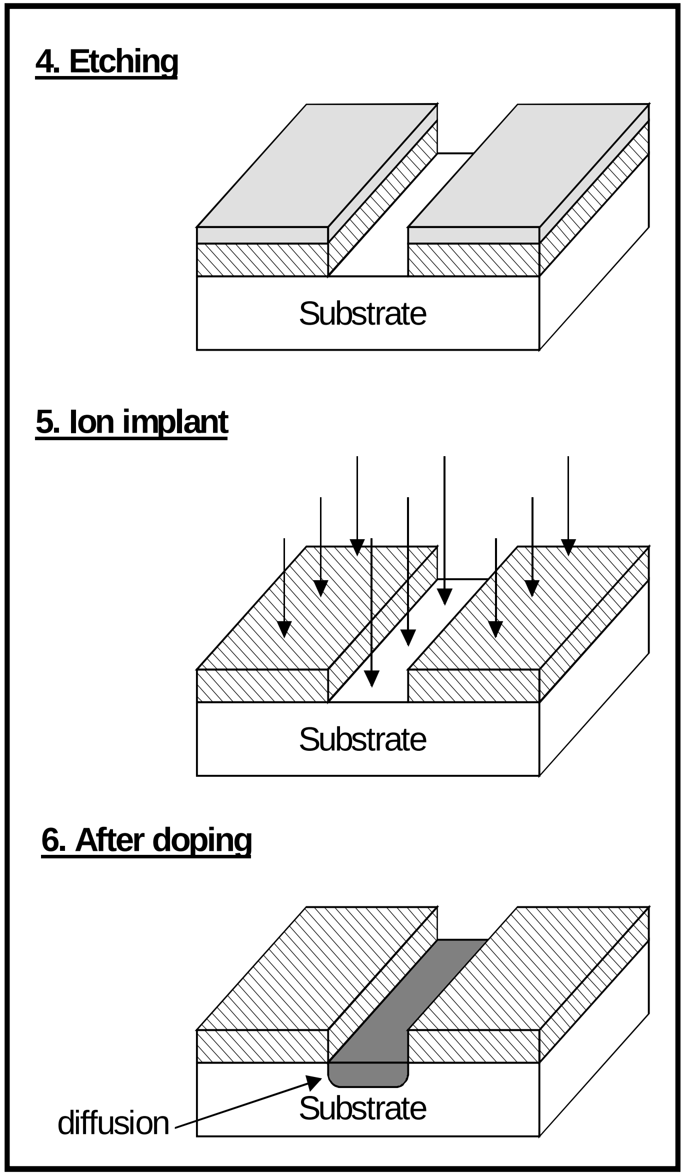
Basic sequence

- 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

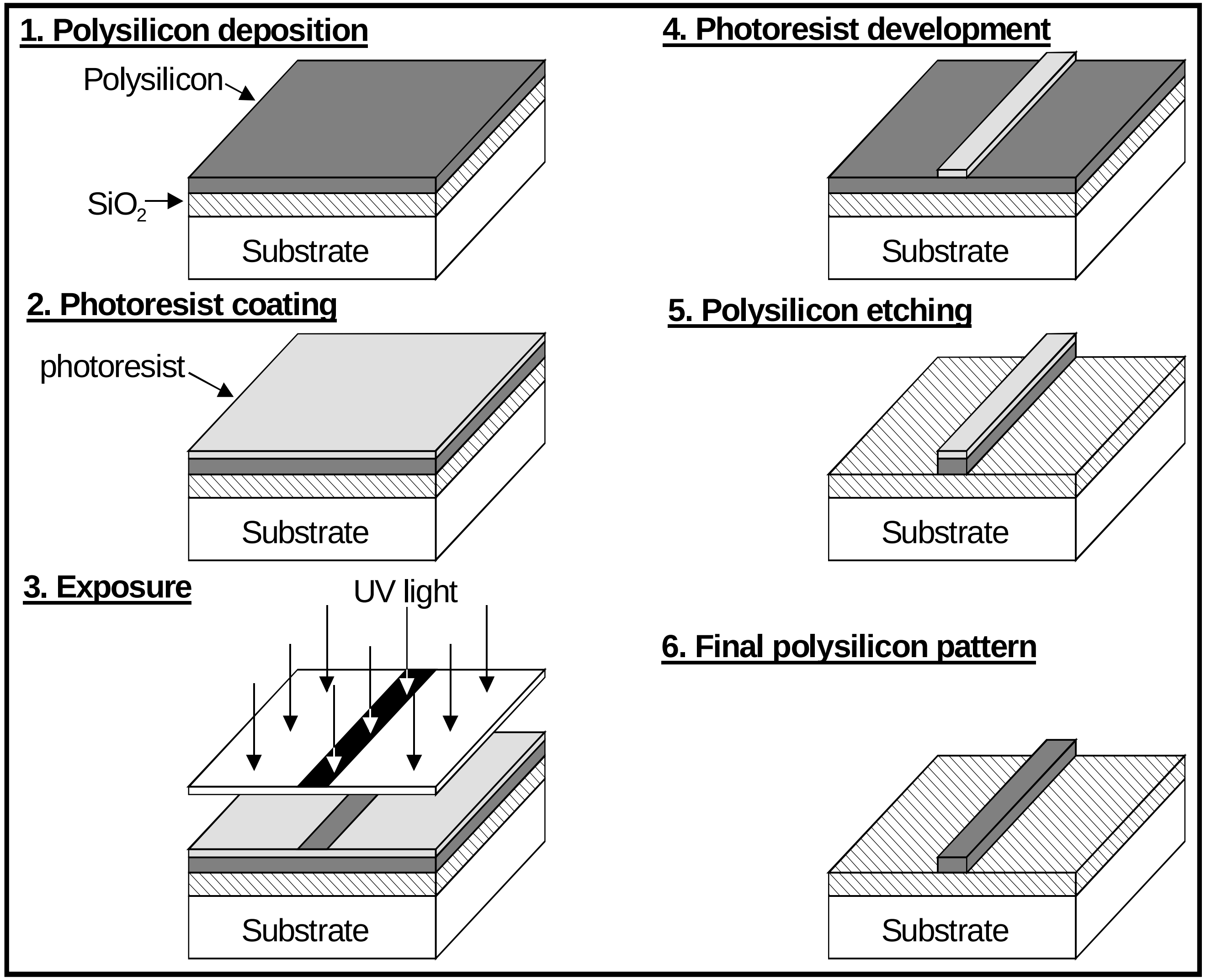


Lithography

- 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

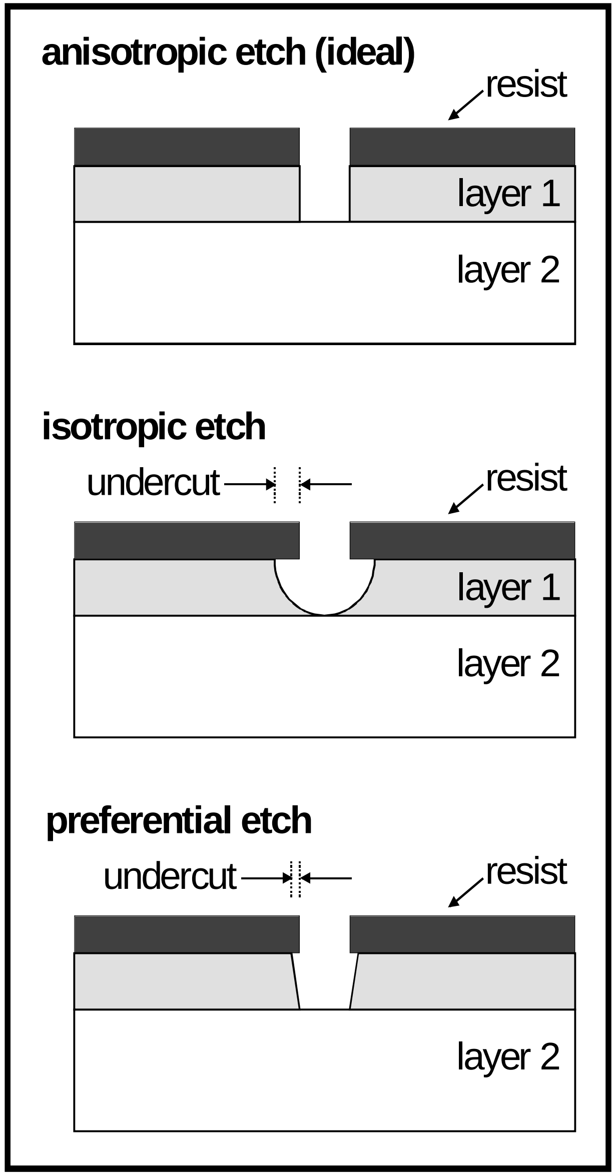
Lithography

Patterning a layer above the silicon surface

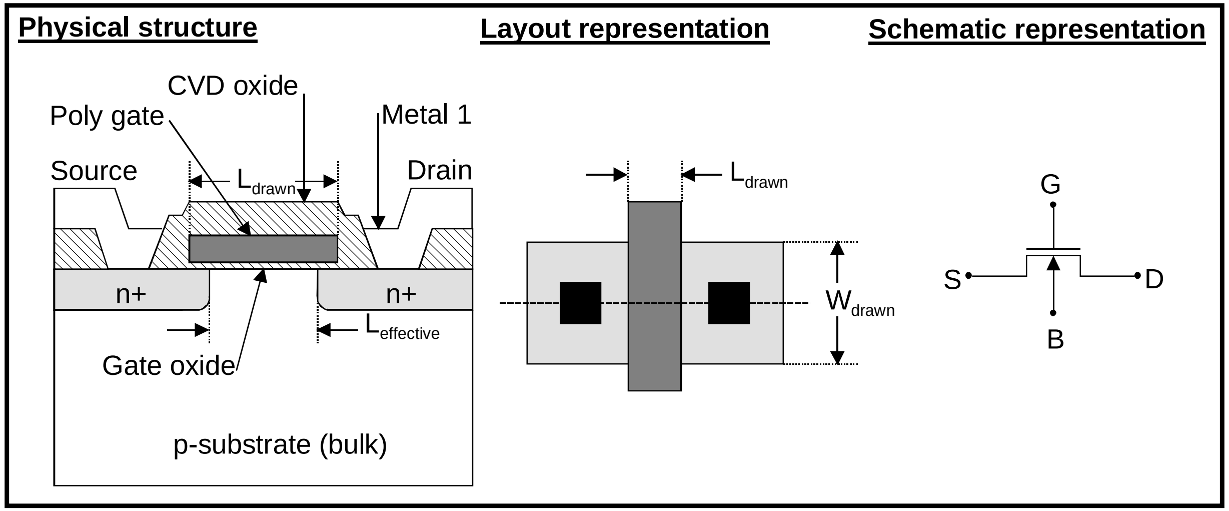


Lithography

- 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



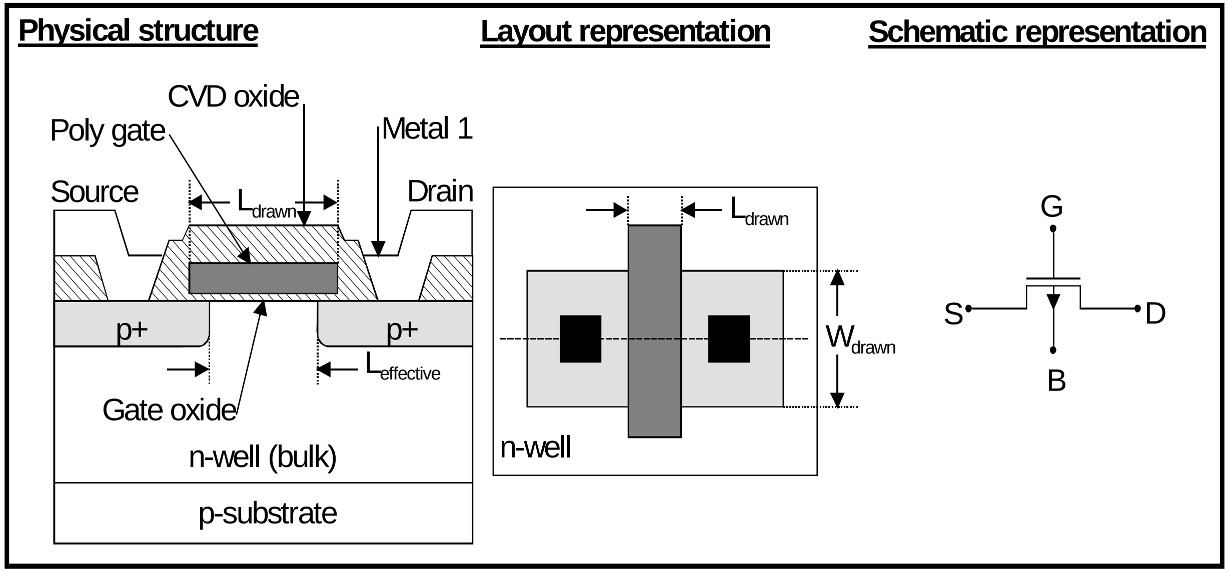
NMOS physical structure:

- 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

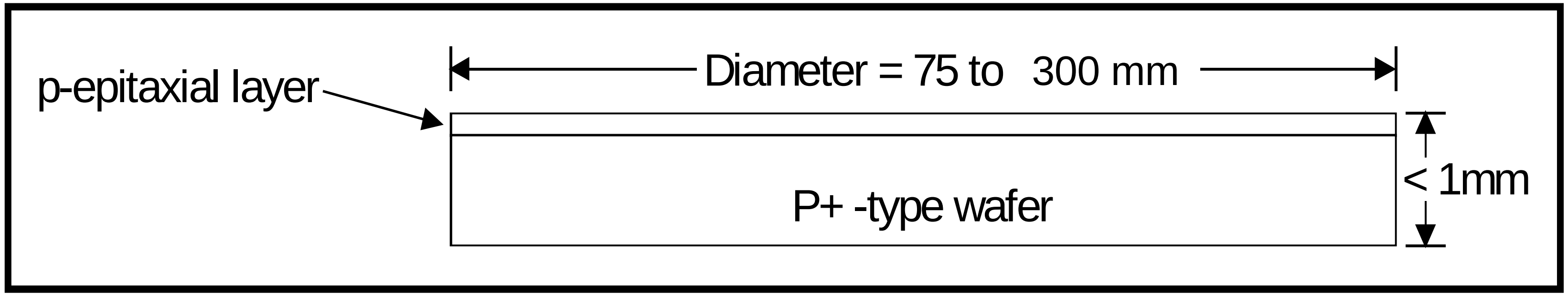
CMOS fabrication sequence

0. Start:

- For an n-well process the starting point is a p-type silicon wafer:
- wafer: typically 75 to 300mm in diameter and less than 1mm thick

1. Epitaxial growth:

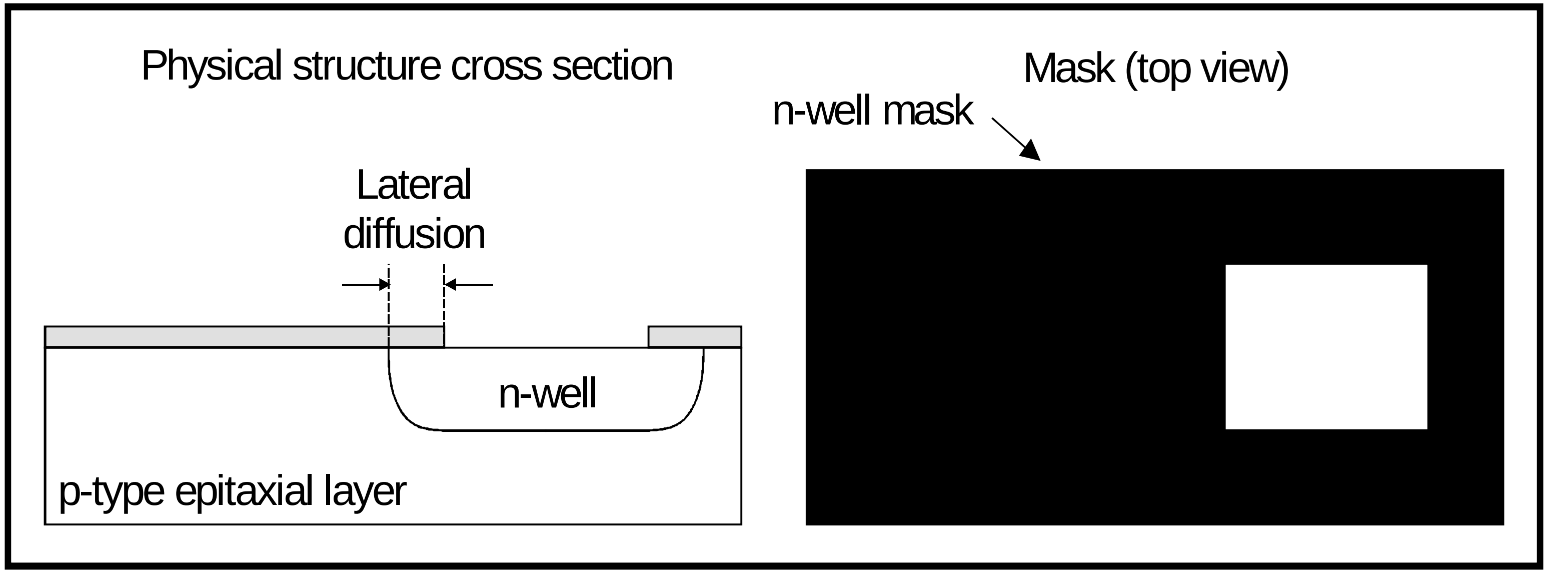
- A single p-type single crystal film is grown on the surface of the wafer by:
  - subjecting the wafer to high temperature and a source of dopant material
- The epi layer is used as the base layer to build the devices
- Advanced technologies use high resistively substrates (non-epi)



CMOS fabrication sequence

2. N-well Formation:

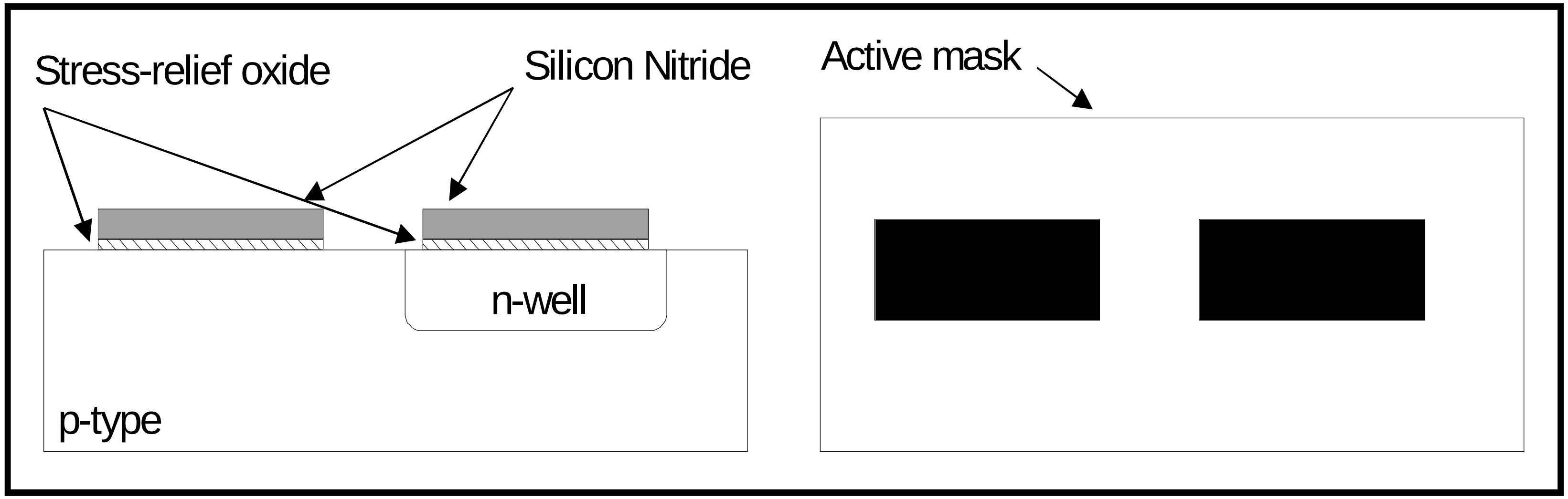
- PMOS transistors are fabricated in n-well regions
- The first mask defines the n-well regions
- N-well’s are formed by ion implantation or deposition and diffusion
- Lateral diffusion limits the proximity between structures
- Ion implantation results in shallower wells compatible with today’s fine-line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



CMOS fabrication sequence

3. Active area definition:

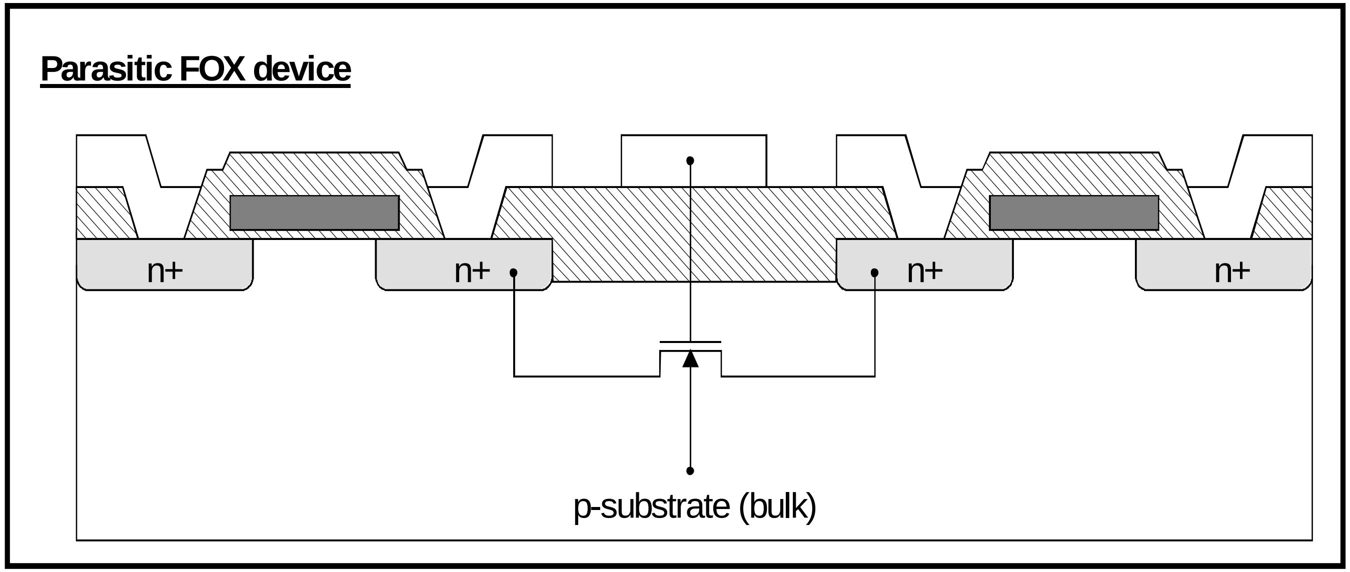
- Active area:
  - planar section of the surface where transistors are build
  - defines the gate region (thin oxide)
  - defines the n+ or p+ regions
- A thin layer of SiO_2 is grown over the active region and covered with silicon nitride



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

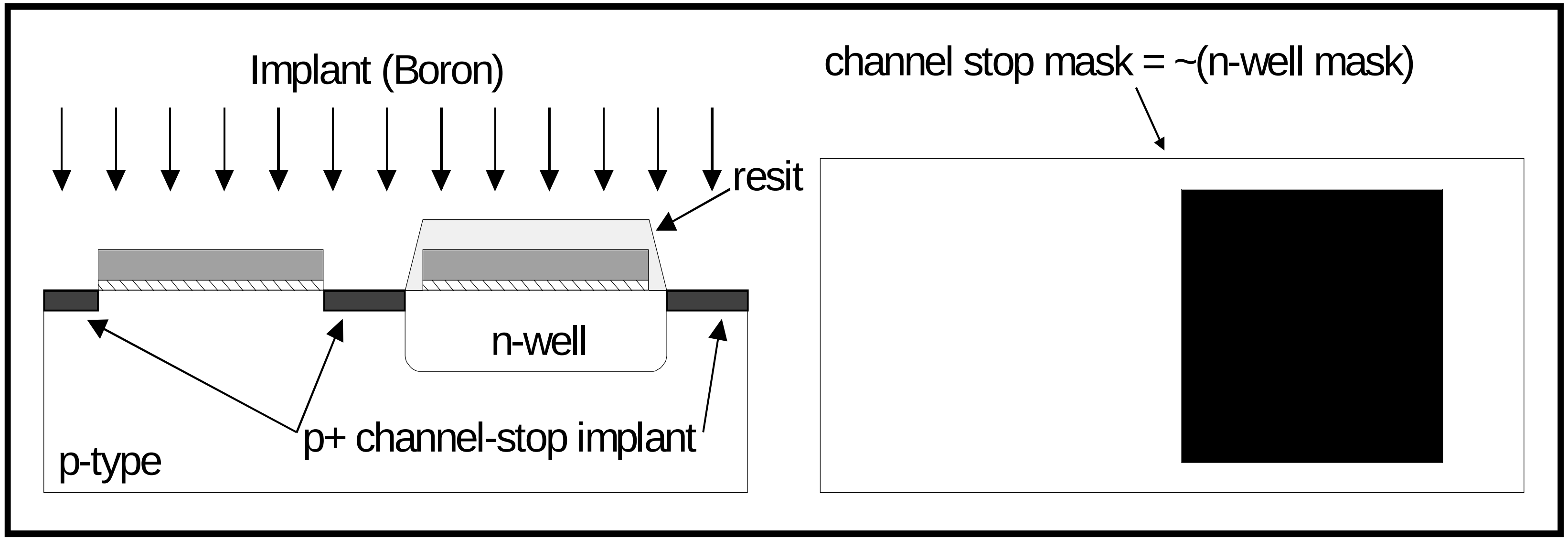


CMOS fabrication sequence

– 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

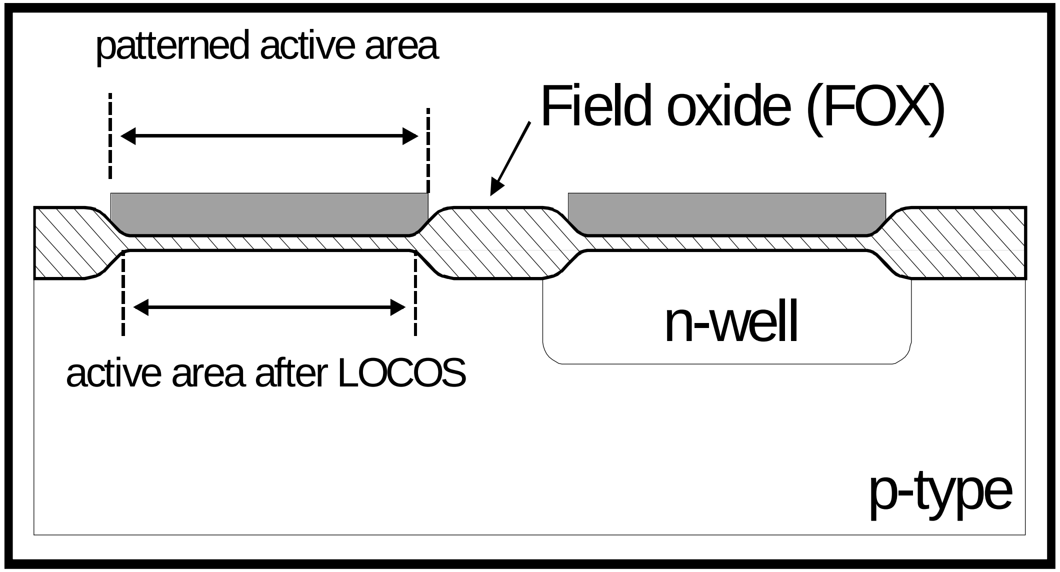
– The silicon nitride (over n-active) and the photoresist (over n-well) act as masks for the channel-stop implant



CMOS fabrication sequence

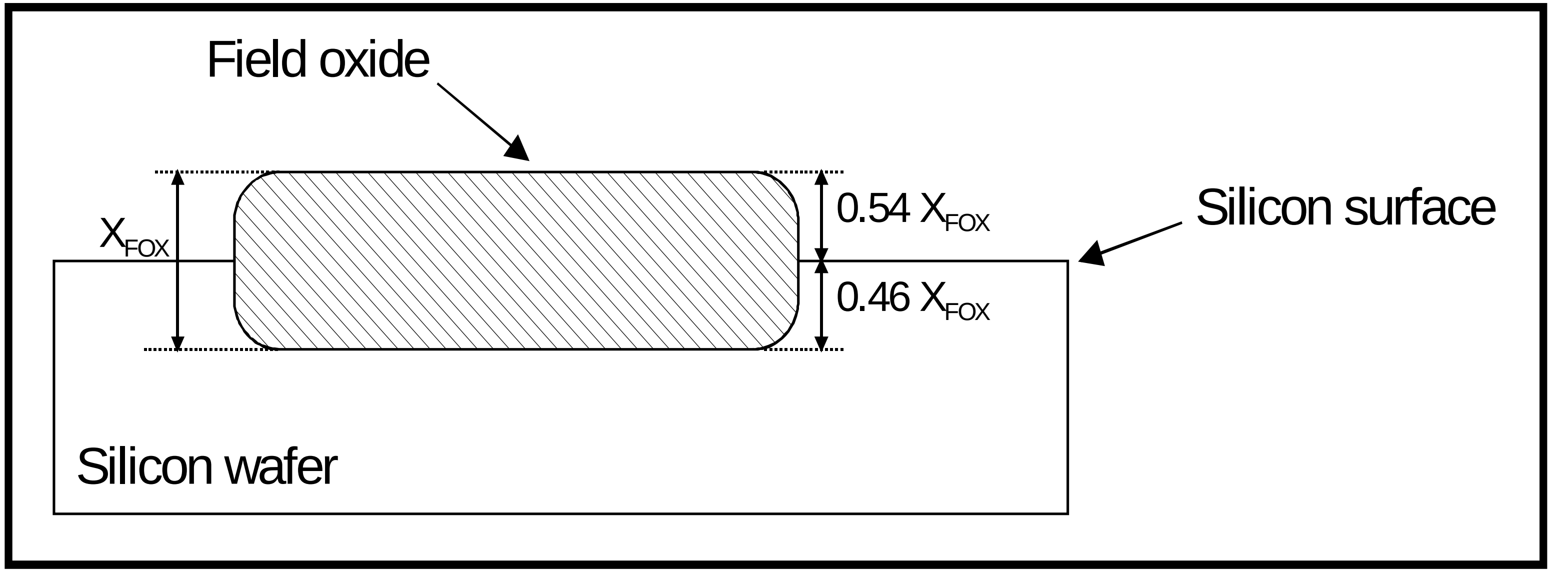
4.2 Local oxidation of silicon (LOCOS)

- 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

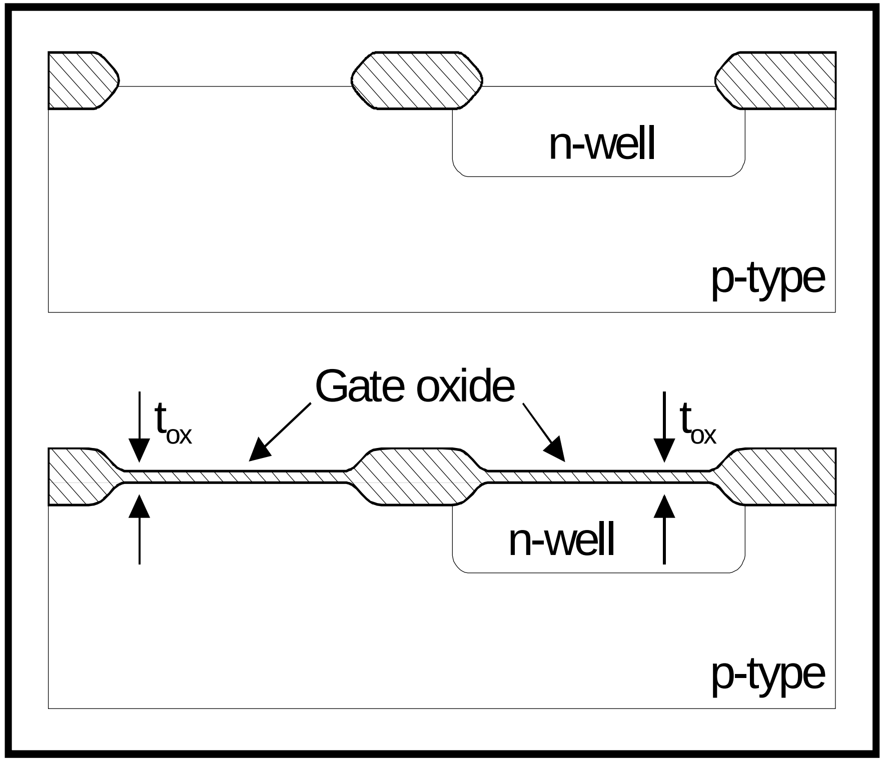
- Silicon oxidation is obtained by:
  - Heating the wafer in a oxidizing atmosphere:
    - Wet oxidation: water vapor, T = 900 to 1000ºC (rapid process)
    - Dry oxidation: Pure oxygen, T = 1200ºC (high temperature required to achieve an acceptable growth rate)
- Oxidation consumes silicon
  - SiO_2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



CMOS fabrication sequence

5. Gate oxide growth

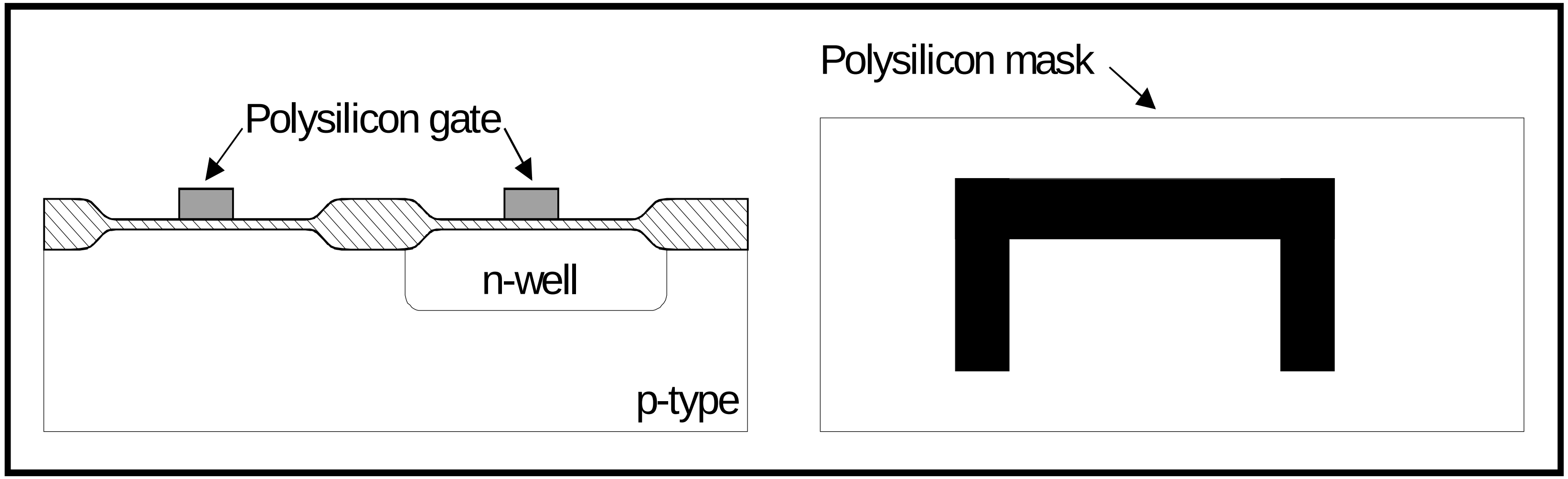
- The nitride and stress-relief oxide are removed
- The devices threshold voltage is adjusted by:
  - adding charge at the silicon/oxide interface
- The well controlled gate oxide is grown with thickness t_{ox}



CMOS fabrication sequence

6. Polysilicon deposition and patterning

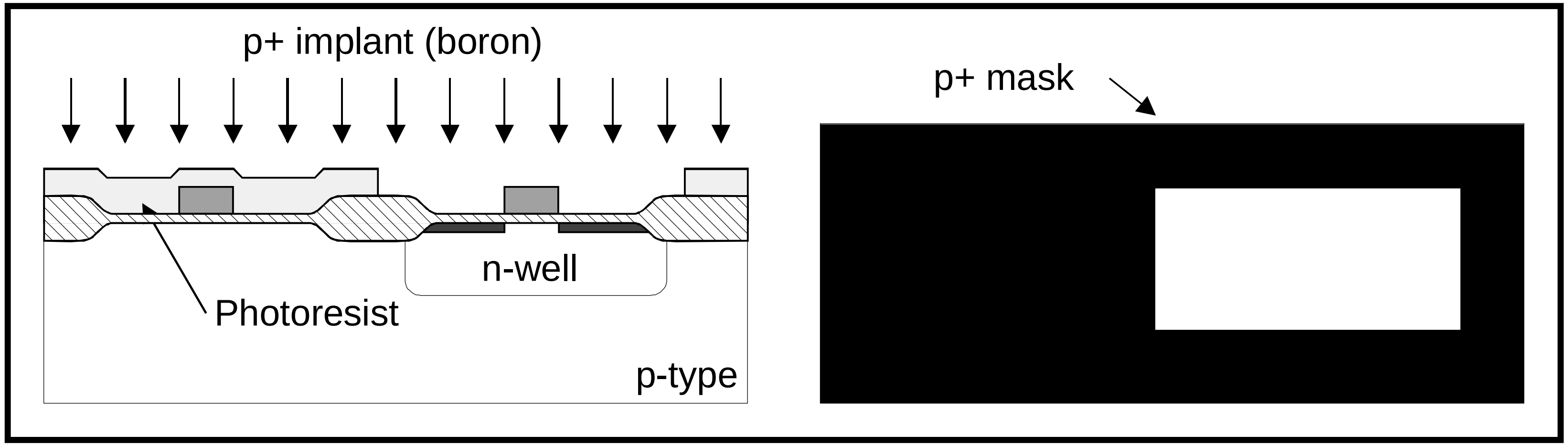
- A layer of polysilicon is deposited over the entire wafer surface
- The polysilicon is then patterned by a lithography sequence
- All the MOSFET gates are defined in a single step
- The polysilicon gate can be doped (n+) while is being deposited to lower its parasitic resistance (important in high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



CMOS fabrication sequence

7. PMOS formation

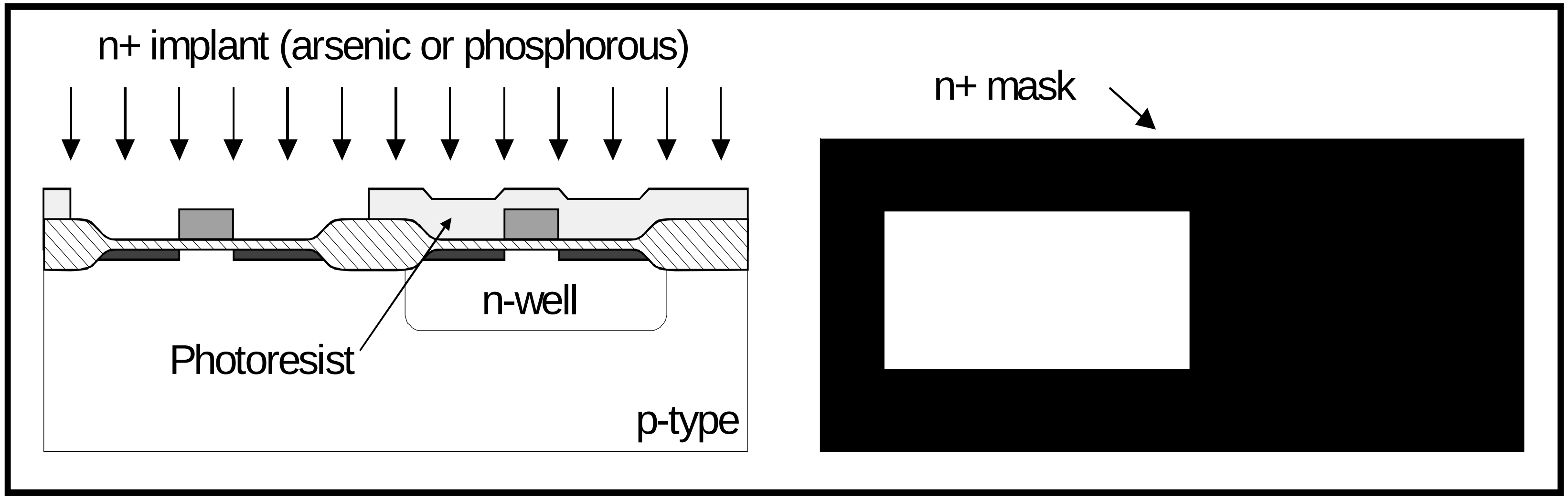
- Photoresist is patterned to cover all but the p+ regions
- A boron ion beam creates the p+ source and drain regions
- The polysilicon serves as a mask to the underlying channel
  - This is called a self-aligned process
  - It allows precise placement of the source and drain regions
- During this process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

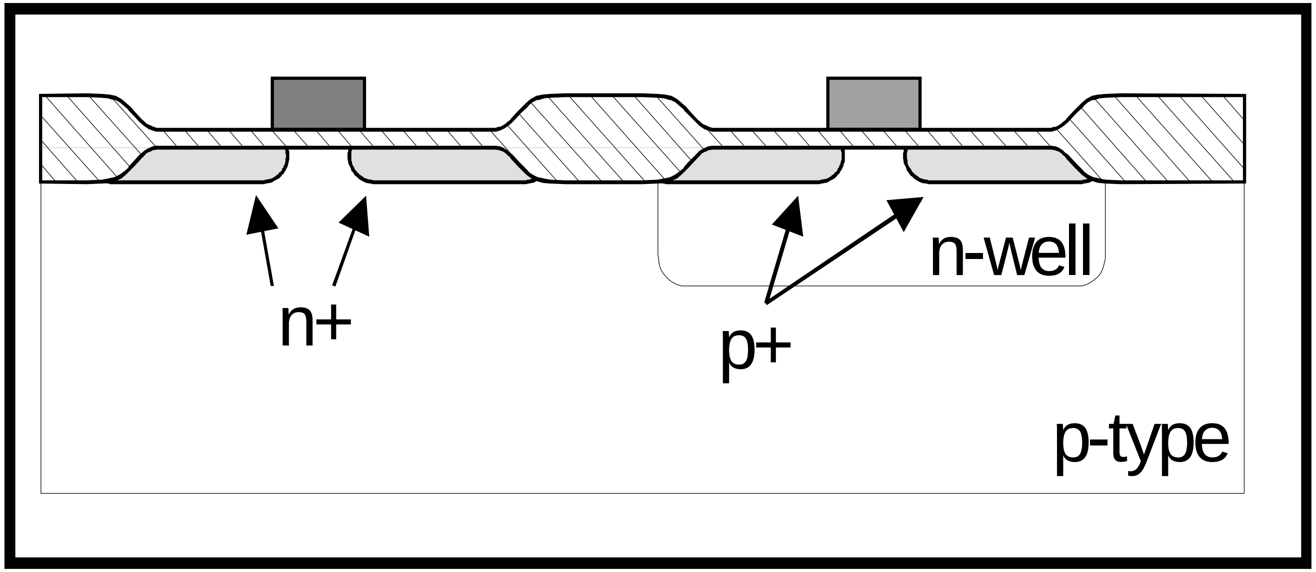
- Photoresist is patterned to define the n+ regions
- Donors (arsenic or phosphorous) are ion-implanted to dope the n+ source and drain regions
- The process is self-aligned
- The gate is n-type doped



CMOS fabrication sequence

9. Annealing

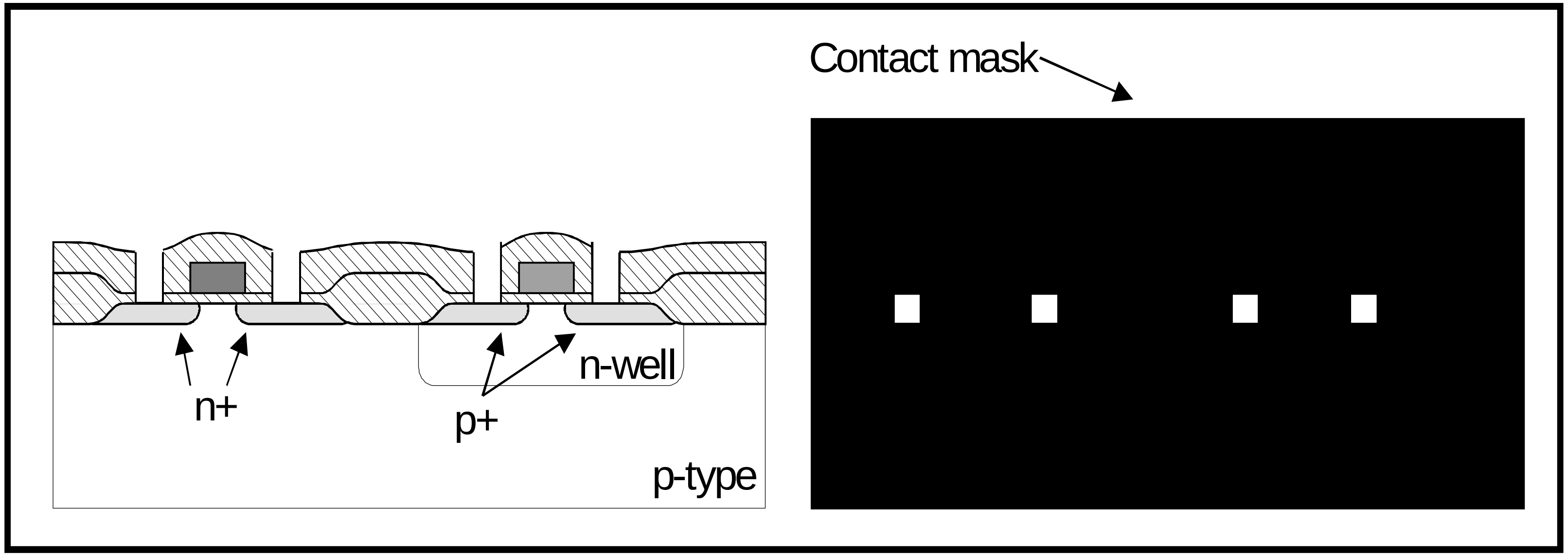
- After the implants are completed a thermal annealing cycle is executed
- This allows the impurities to diffuse further into the bulk
- After thermal annealing, it is important to keep the remaining process steps at as low temperature as possible



CMOS fabrication sequence

10. Contact cuts

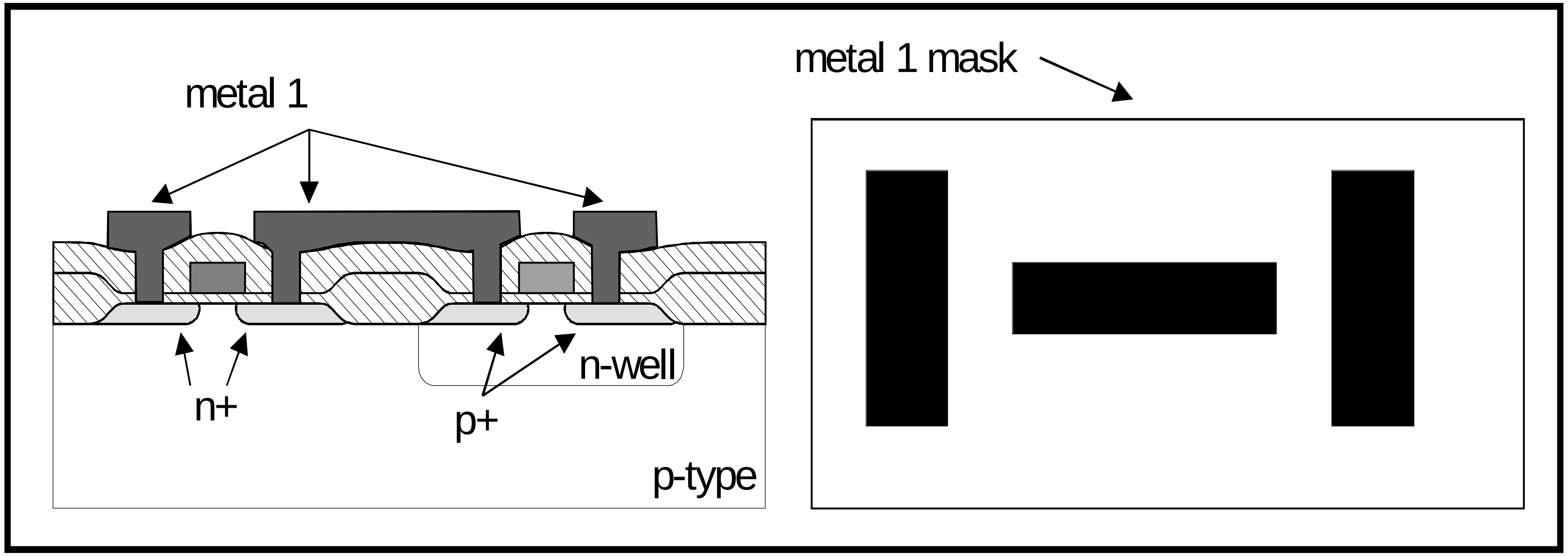
- The surface of the IC is covered by a layer of CVD oxide
  - The oxide is deposited at low temperature (LTO) to avoid that underlying doped regions will undergo diffusive spreading
- Contact cuts are defined by etching SiO_2 down to the surface to be contacted
- These allow metal to contact diffusion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



CMOS fabrication sequence

11. Metal 1

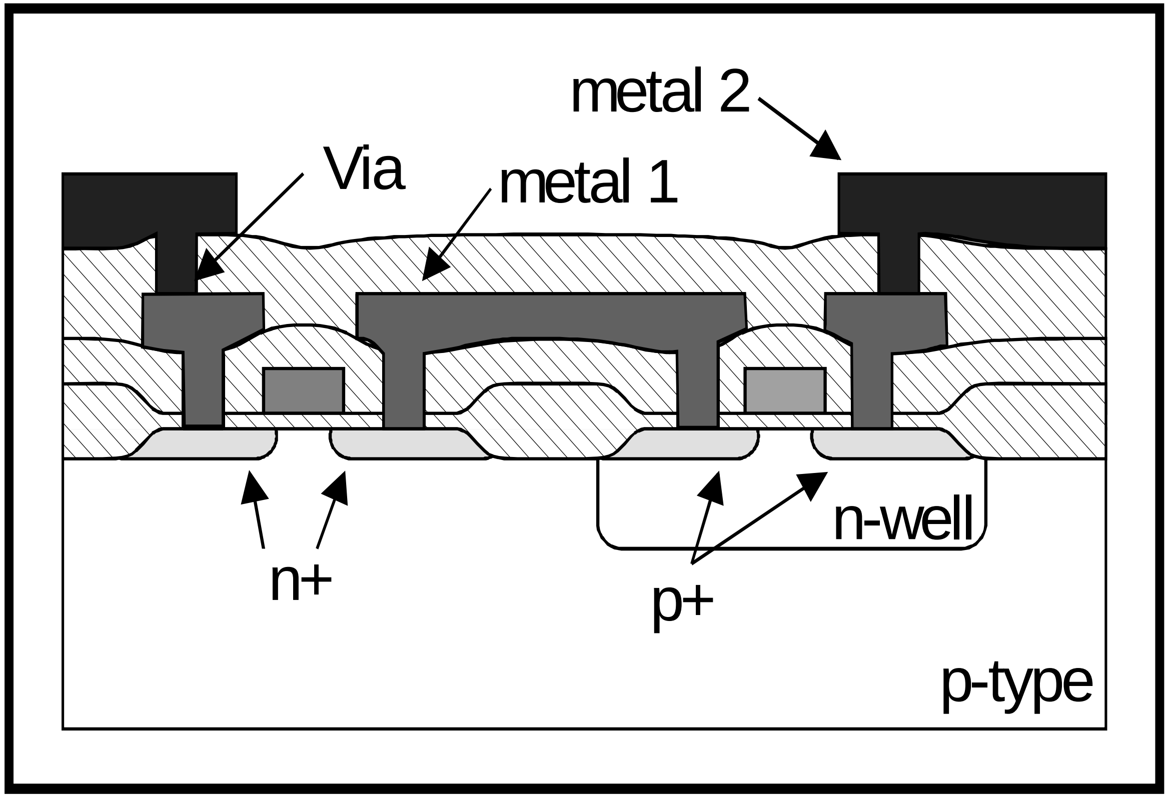
- A first level of metallization is applied to the wafer surface and selectively etched to produce the interconnects



CMOS fabrication sequence

12. Metal 2

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



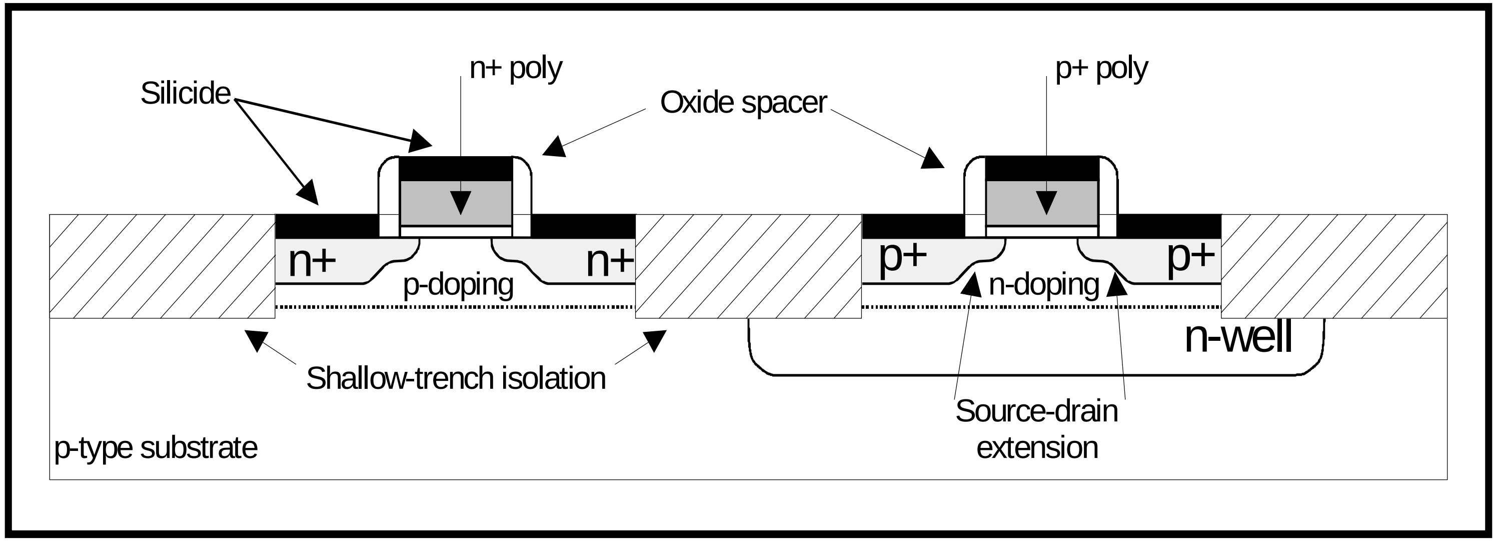
CMOS fabrication sequence

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.

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 interconnections
- Stacked contacts and vias
- Chemical Metal Polishing for technologies with several metal levels
- For analogue applications some processes offer:
  - capacitors
  - resistors
  - bipolar transistors (BiCMOS)

Yield

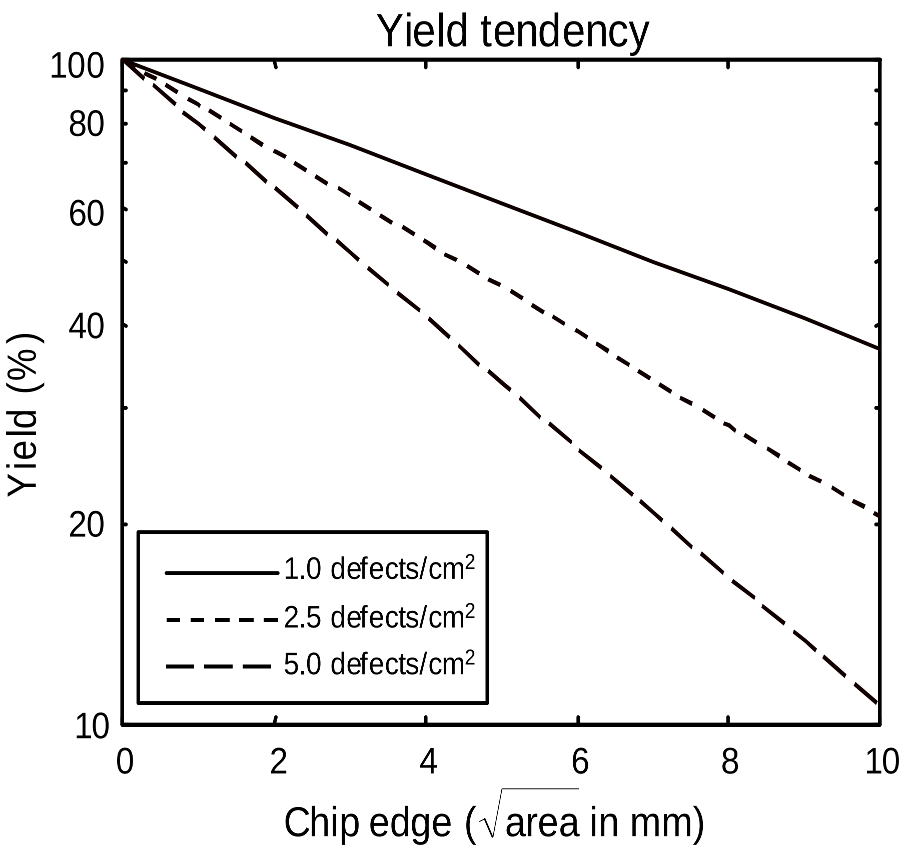
- Yield

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

- 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)

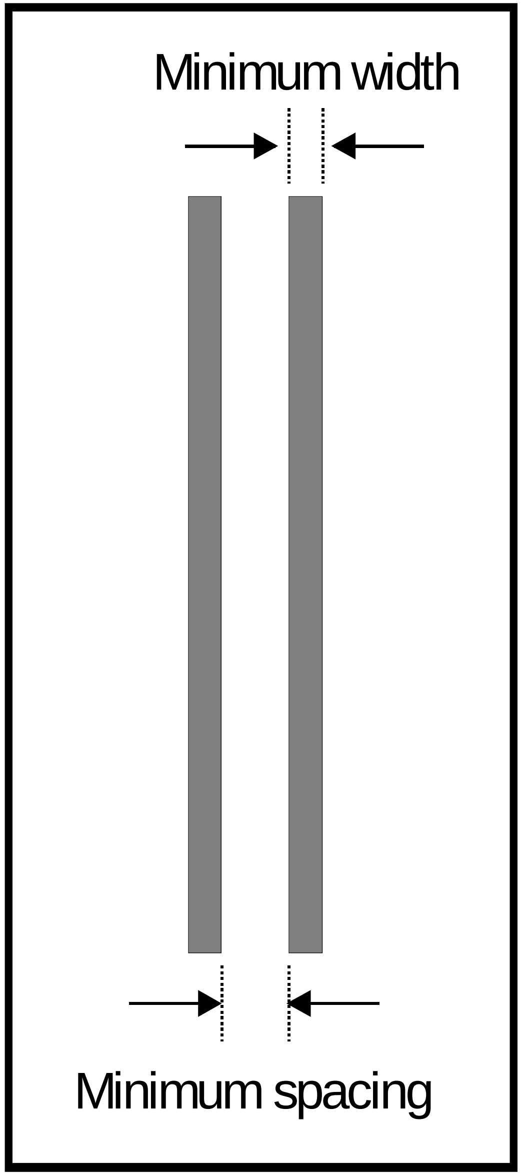


Design rules

- 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

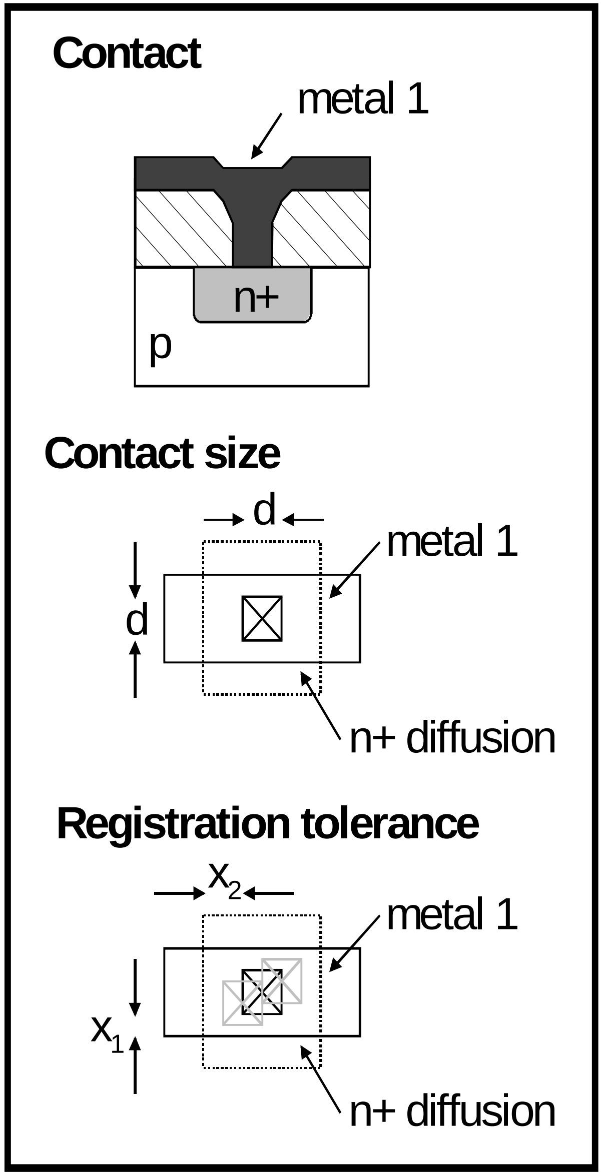
Design rules

- 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



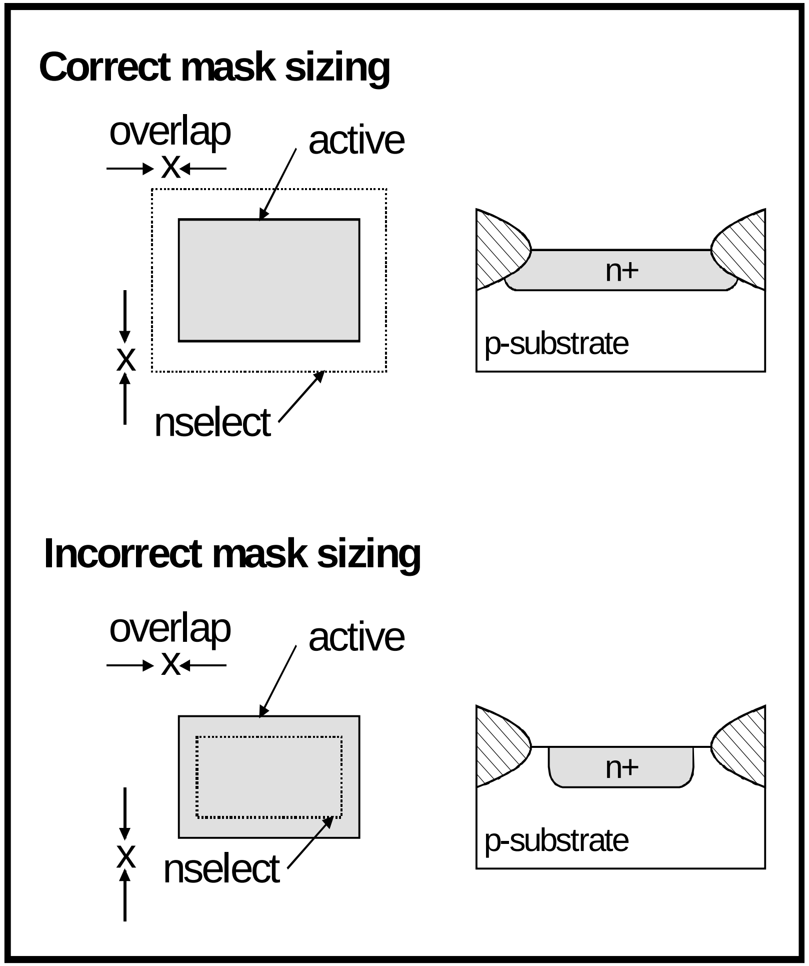
Design rules

- 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)



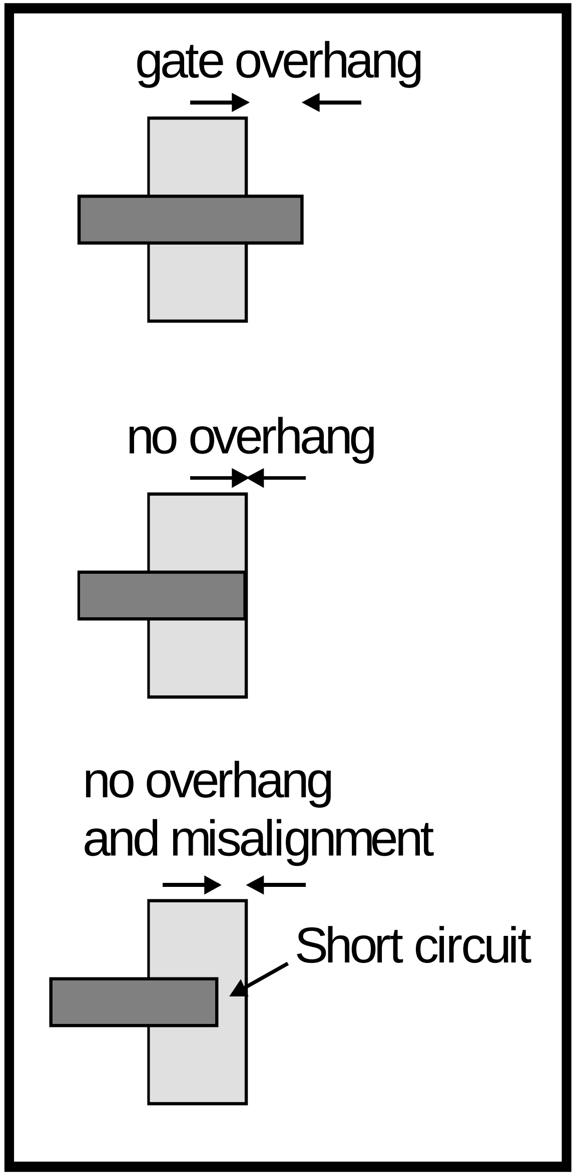
Design rules

- 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

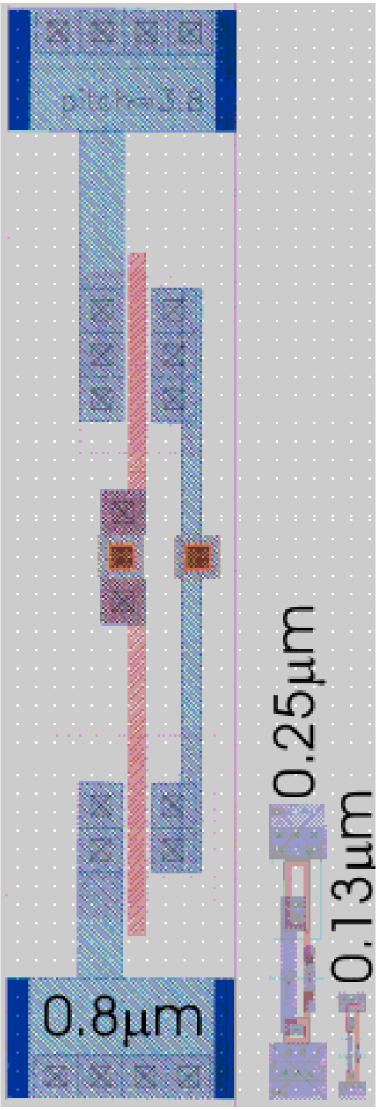
- Gate overhang:
  - The gate must overlap the active area by a minimum amount
  - This is done to ensure that a misaligned gate will still yield a structure with separated drain and source regions
- A modern process have thousands of rules to be verified
  - Programs called Design Rule Checkers assist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



Technology scaling

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

Scaling, why is it done?



Technology scaling

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

Technology scaling

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

Technology scaling

- The **scaling variables** are:

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

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

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

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

Scaling consequences

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

- Device/die area:

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

  - In practice, microprocessor 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…)

Scaling consequences

- Gate capacitance:

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

- Drain current:

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

- Gate delay:

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

  Frequency $\rightarrow \alpha = 1.43$

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

Scaling consequences

- Power:

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

- Power density:

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

  - In practice due to the faster increase in frequency power density has also been increasing

- Active capacitance/unit-area:
  Power dissipation is a function of the operation <u>frequency</u>, the power <u>supply voltage</u> and of the <u>circuit size</u> (number of devices).
  If we normalize the power density to $V^2 \times f$ we obtain the <u>active capacitance per unit area</u> 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: new microprocessors integrate relatively more memory that the previous generation.

Scaling consequences

- 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

Scaling consequences

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