Outline

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

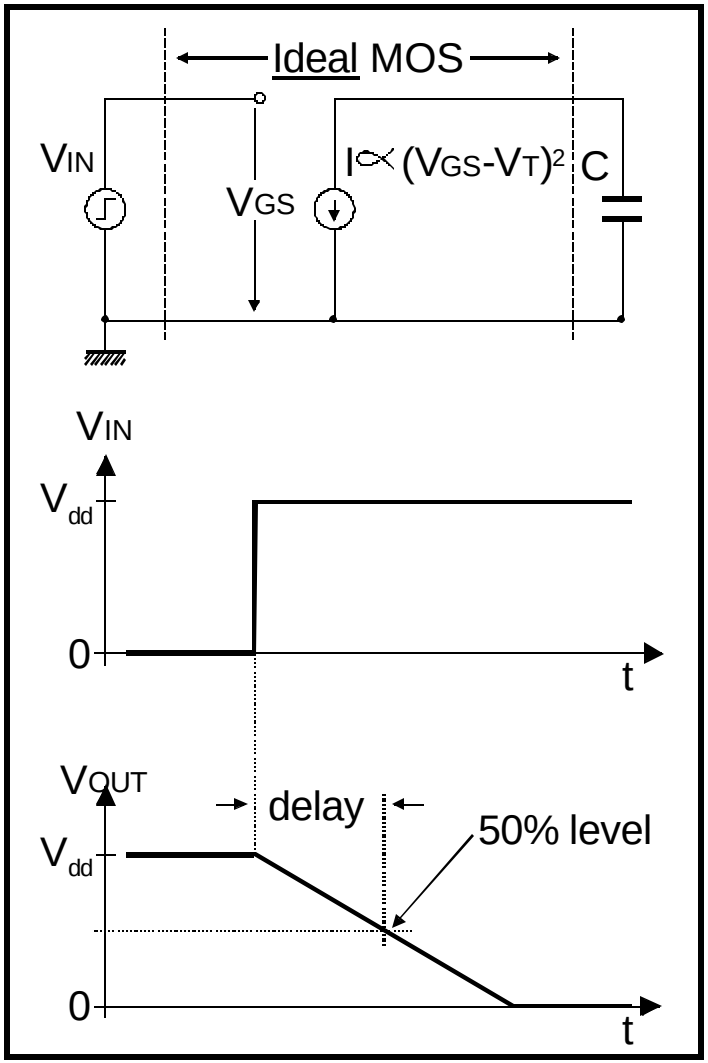
“The [un]desirables”

- Delay in CMOS circuits
- MOSFET capacitances:
  - Physical structure
  - Channel
  - p-n junctions
- MOS resistances
- MOS equivalent circuit

What causes delay?

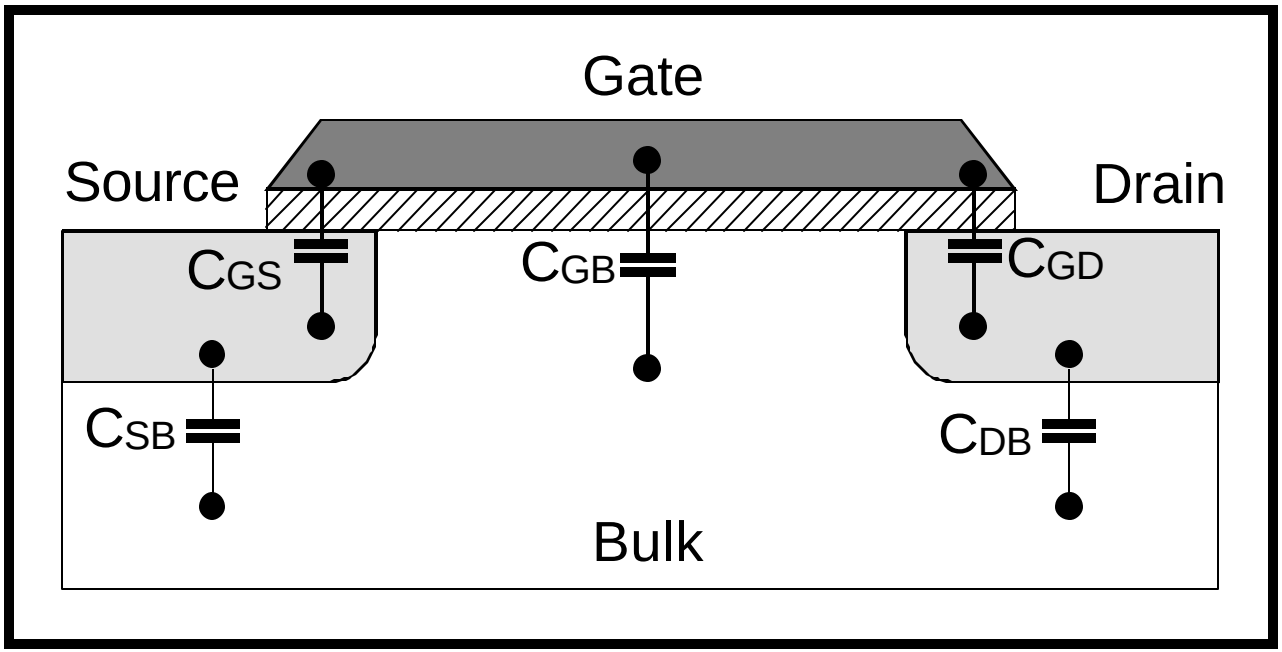
- In MOS circuits capacitive loading is the main cause
- Due to:
  - Device capacitance
  - Interconnect capacitance

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



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

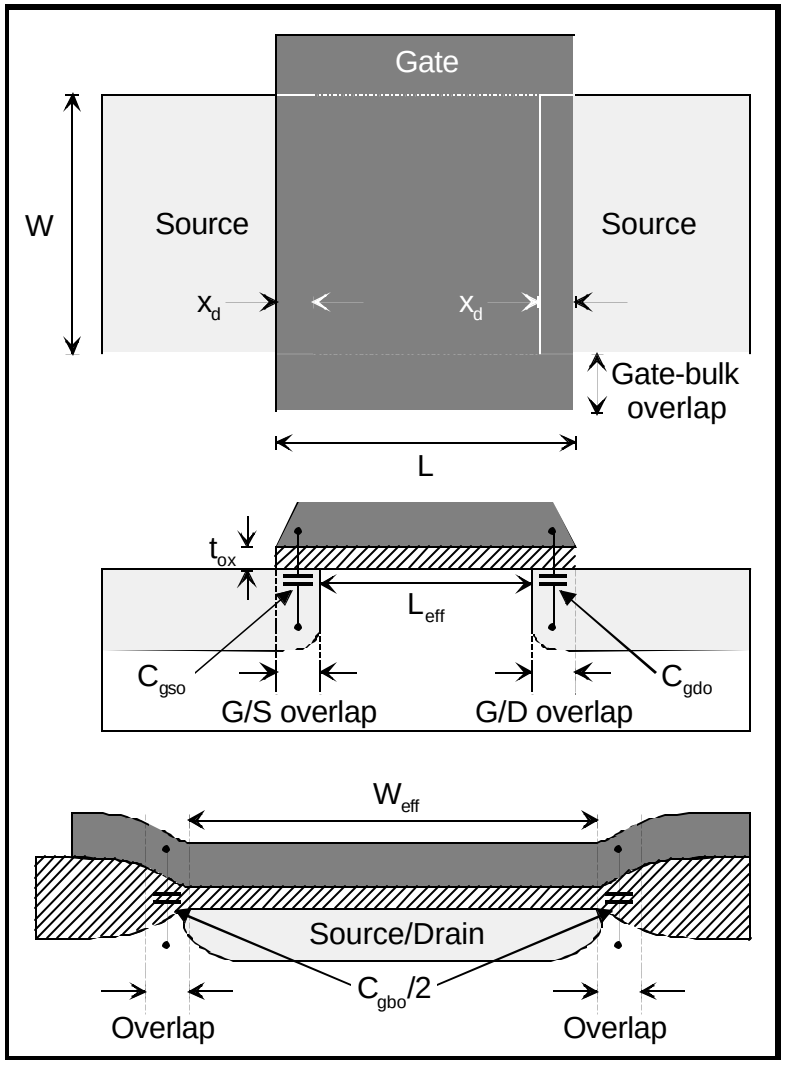
- This gives origin to the source/drain overlap capacitances:

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

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

- Gate-bulk overlap capacitance:

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



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

Channel capacitance

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

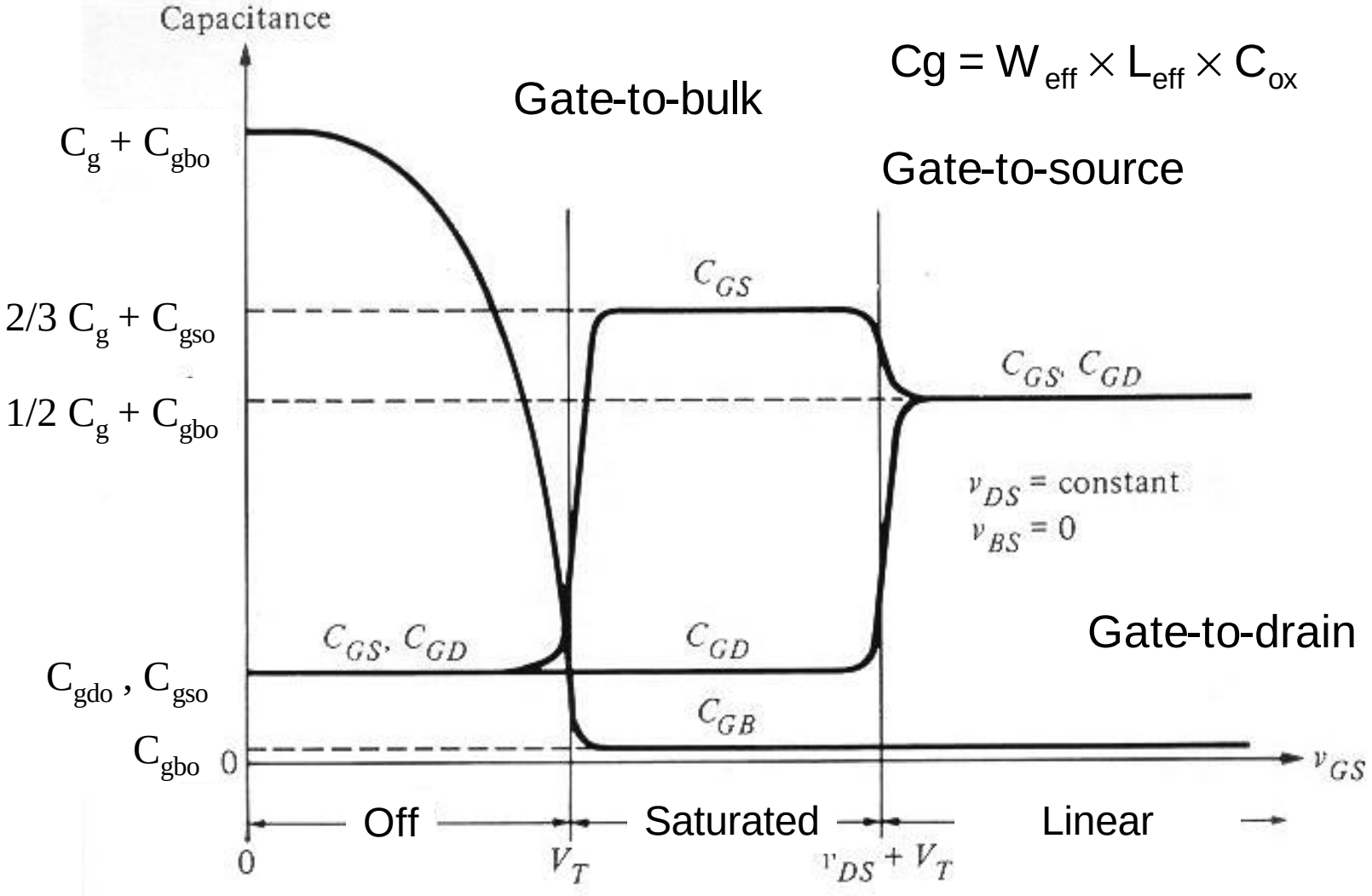
  C_{gb} - gate-to-bulk capacitance

  C_{gs} - gate-to-source capacitance

  C_{gd} - gate-to-drain capacitance

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

Channel capacitance



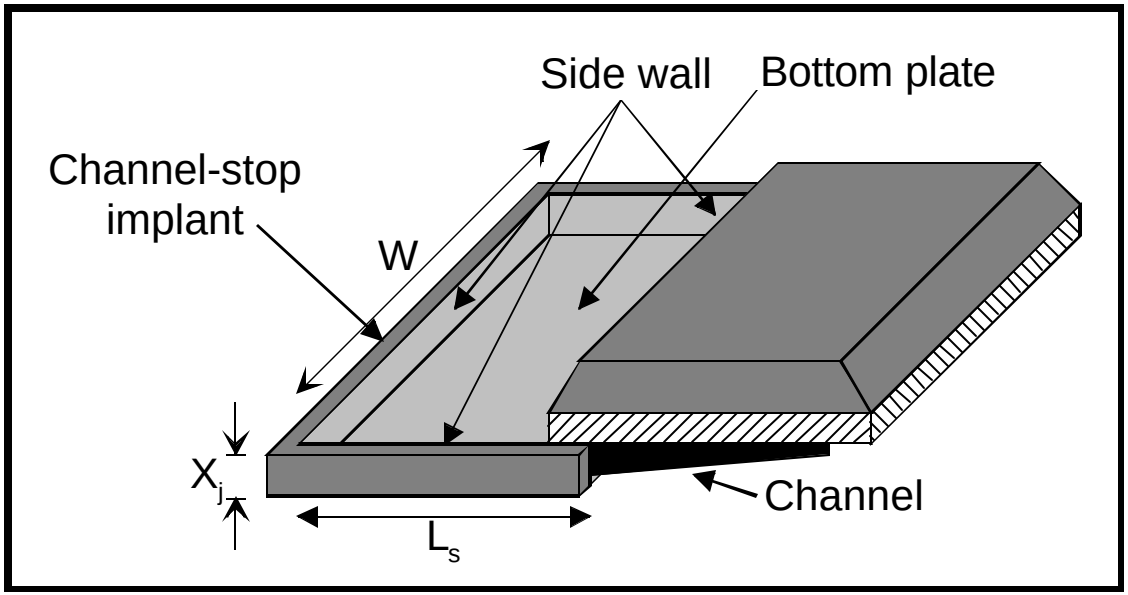
Junction capacitances

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

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

  - Side-wall capacitance:

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



Junction capacitances

0.24 μm process

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

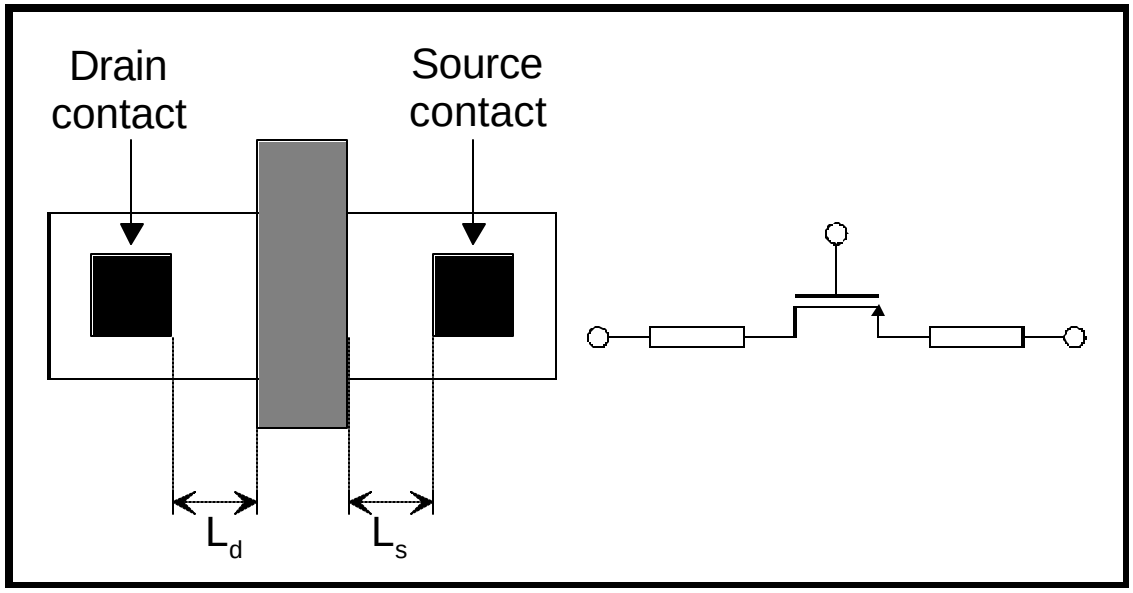
C_g = 2.02 fF

Source/drain resistance

- Scaled down devices $\Rightarrow$ higher source/drain resistance:

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

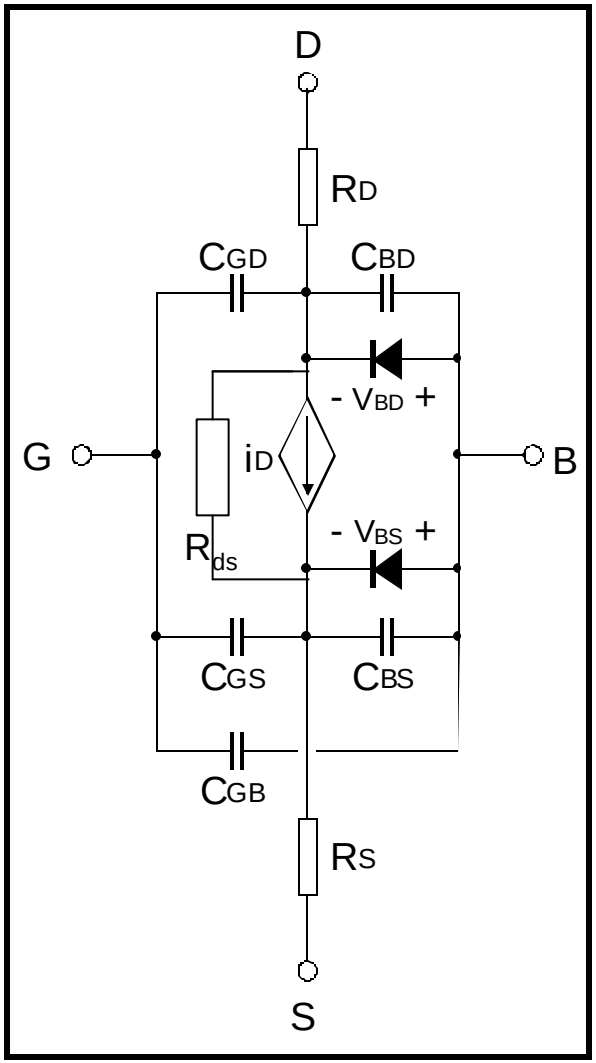
- In sub-μ processes *silicidation* is used to reduce the source, drain and gate parasitic resistance



0.24 μm process

R (P+) = 4 Ω/sq
R (N-) = 4 Ω/sq

MOSFET model



Outline

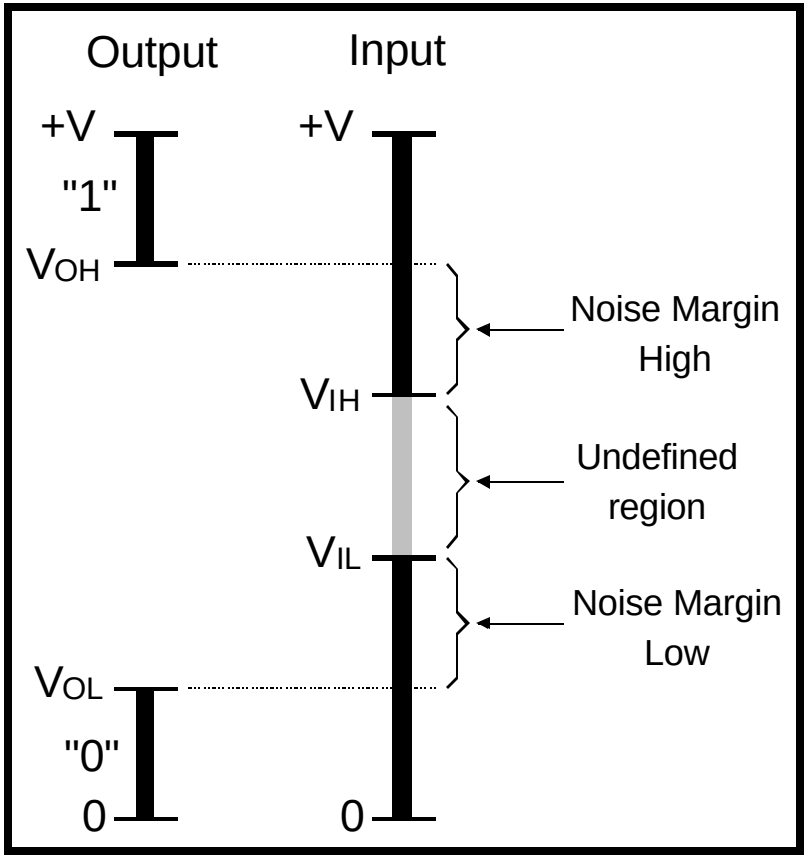
- Introduction – *"Is there a limit?"*
- Transistors – *"CMOS building blocks"*
- Parasitics – *"The [un]desirables"*
- The CMOS inverter – *"A masterpiece"*
- Gates – *"Just like LEGO"*
- Sequential circuits – *"Time also counts!"*
- Storage elements – *"A bit in memory"*
- Technology scaling – *"…,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

“A masterpiece”

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

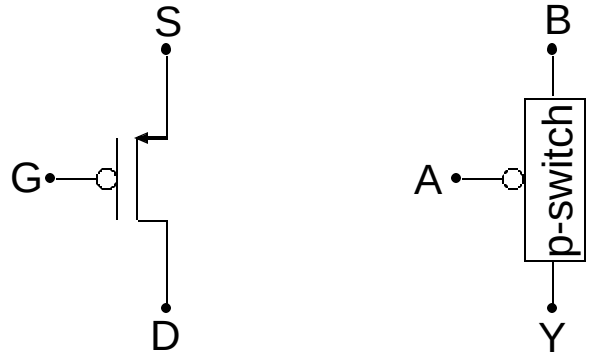
CMOS logic: “0” and “1”

- 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:
  - $NM_H = V_{OH} - V_{IH}$
  - $NM_L = V_{IL} - V_{OL}$



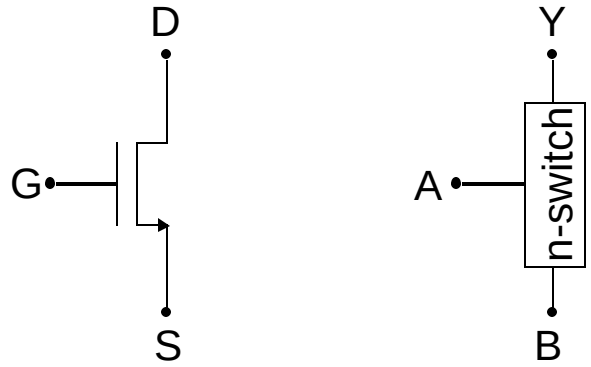
The MOST - a simple switch

p-switch



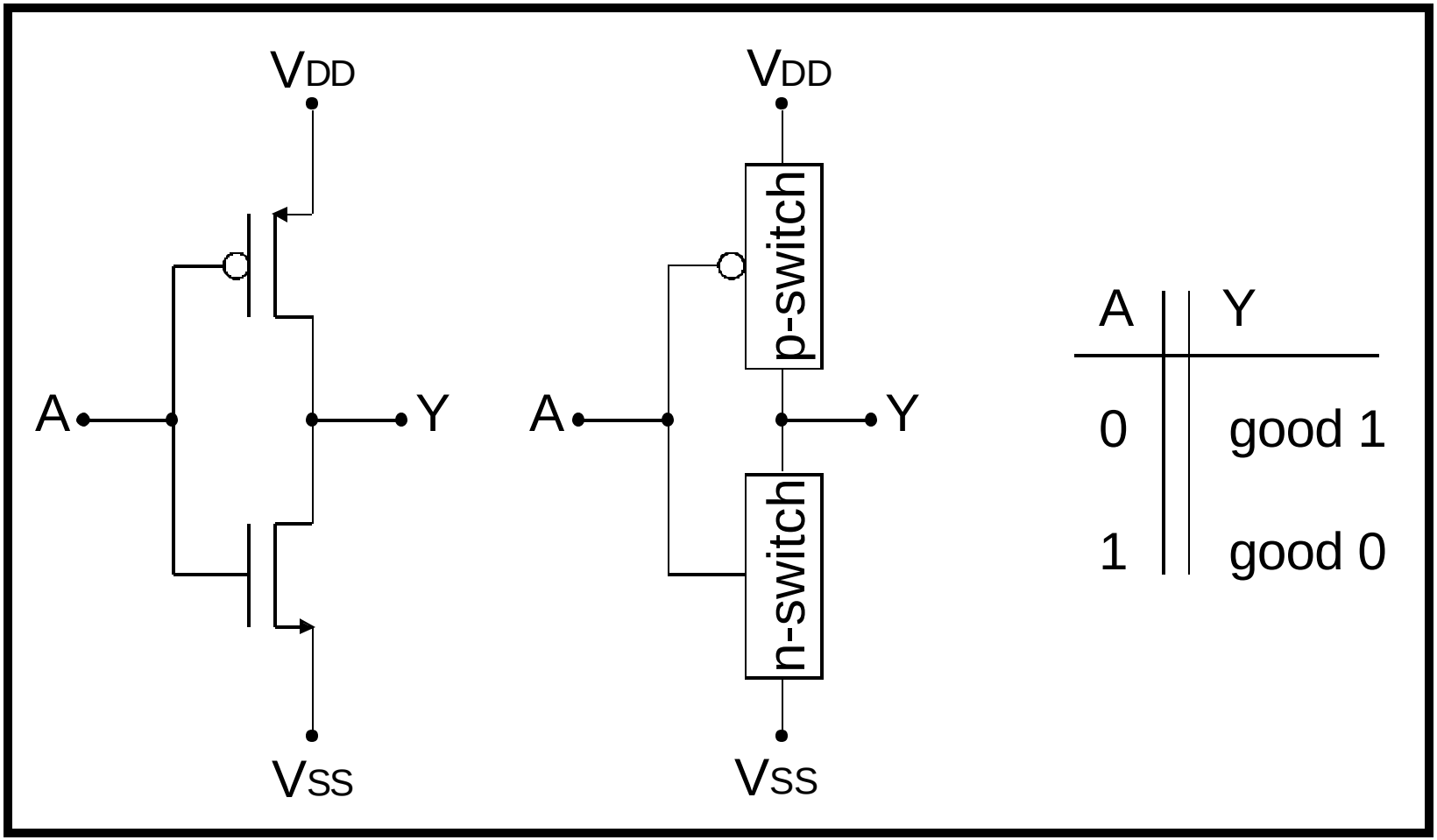
A	B	Y	
0	0	bad 0	(source follower)
0	1	**good 1**	
1	0	?	(high Z)
1	1	?	(high Z)

n-switch

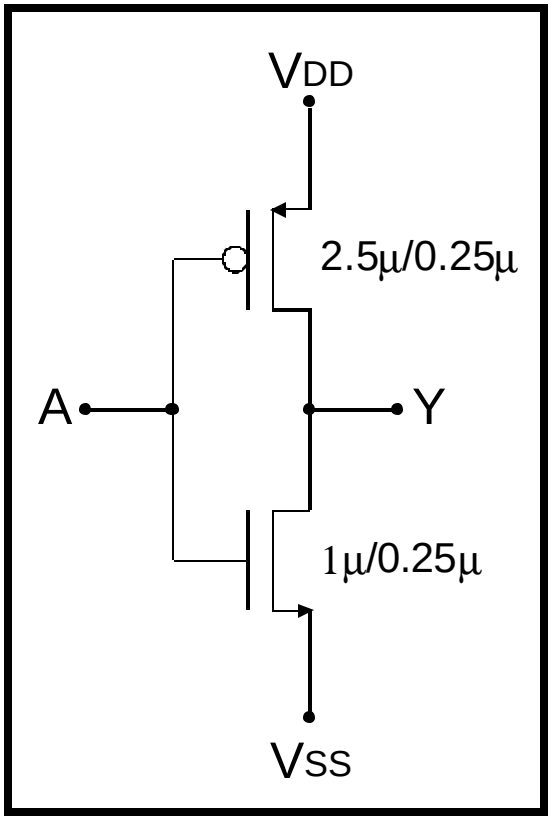


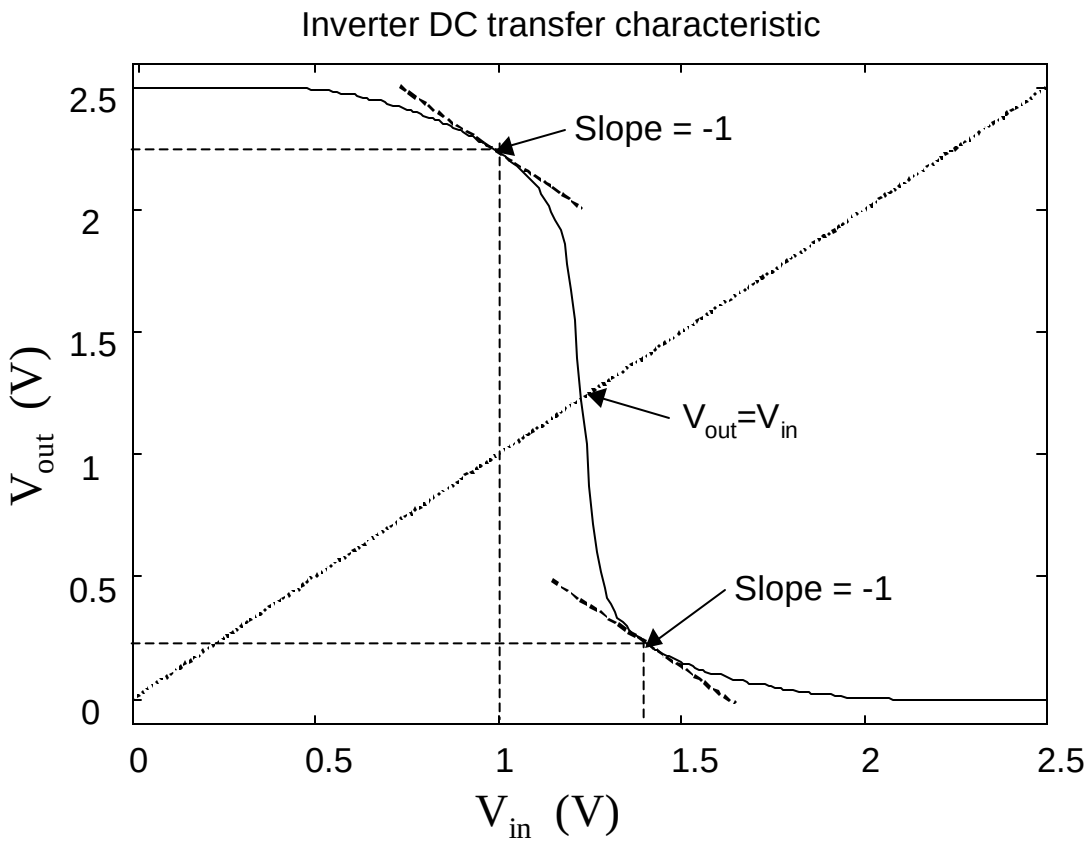
A	B	Y	
0	0	?	(high Z)
0	1	?	(high Z)
1	0	**good 0**	
1	1	bad 1	(source follower)

The CMOS inverter



The CMOS inverter



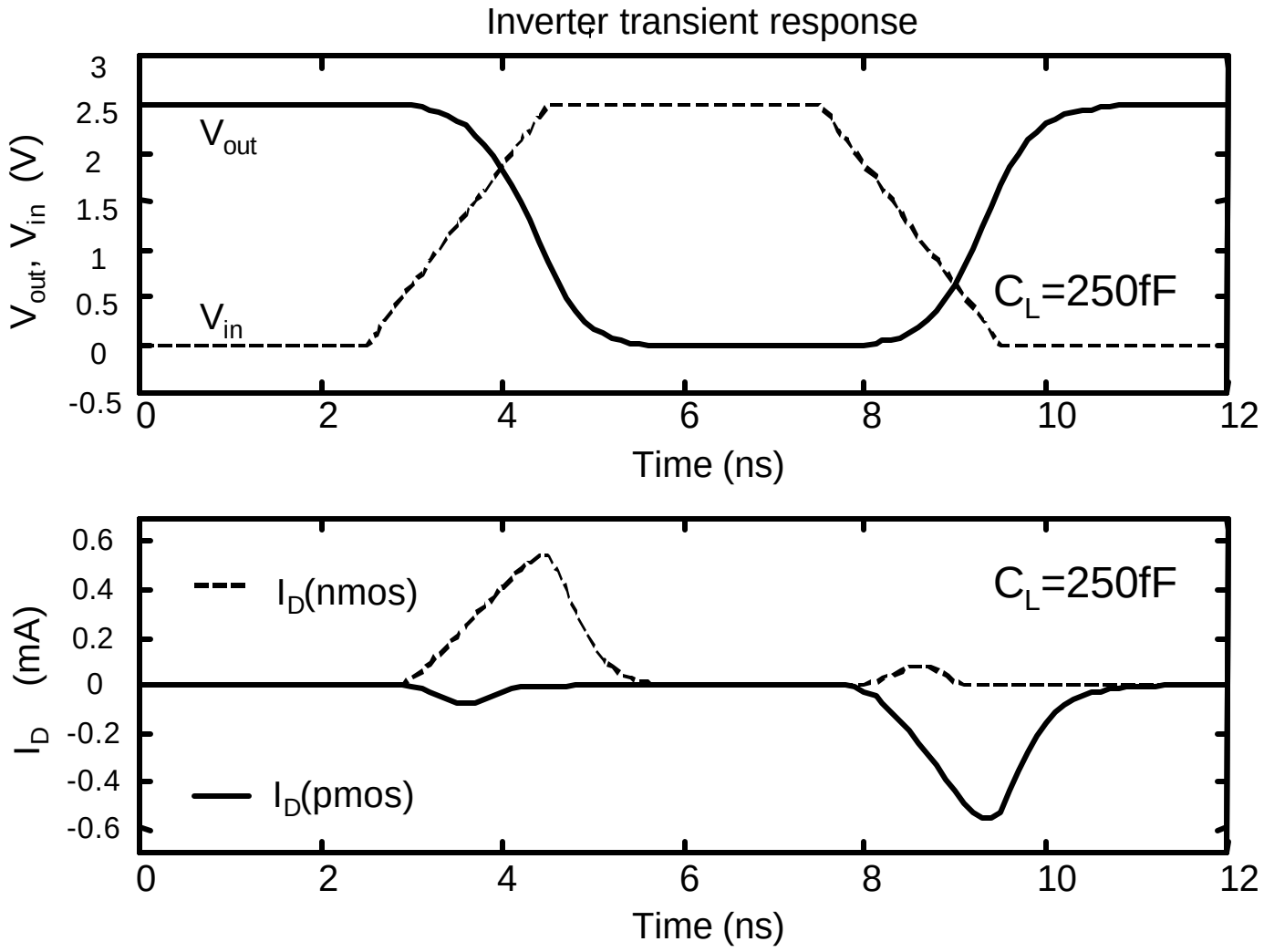


The CMOS inverter

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

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

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

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

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

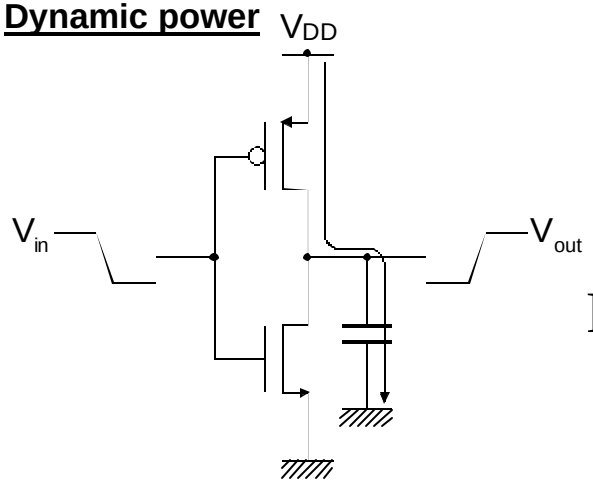
The CMOS inverter

- 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

- 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

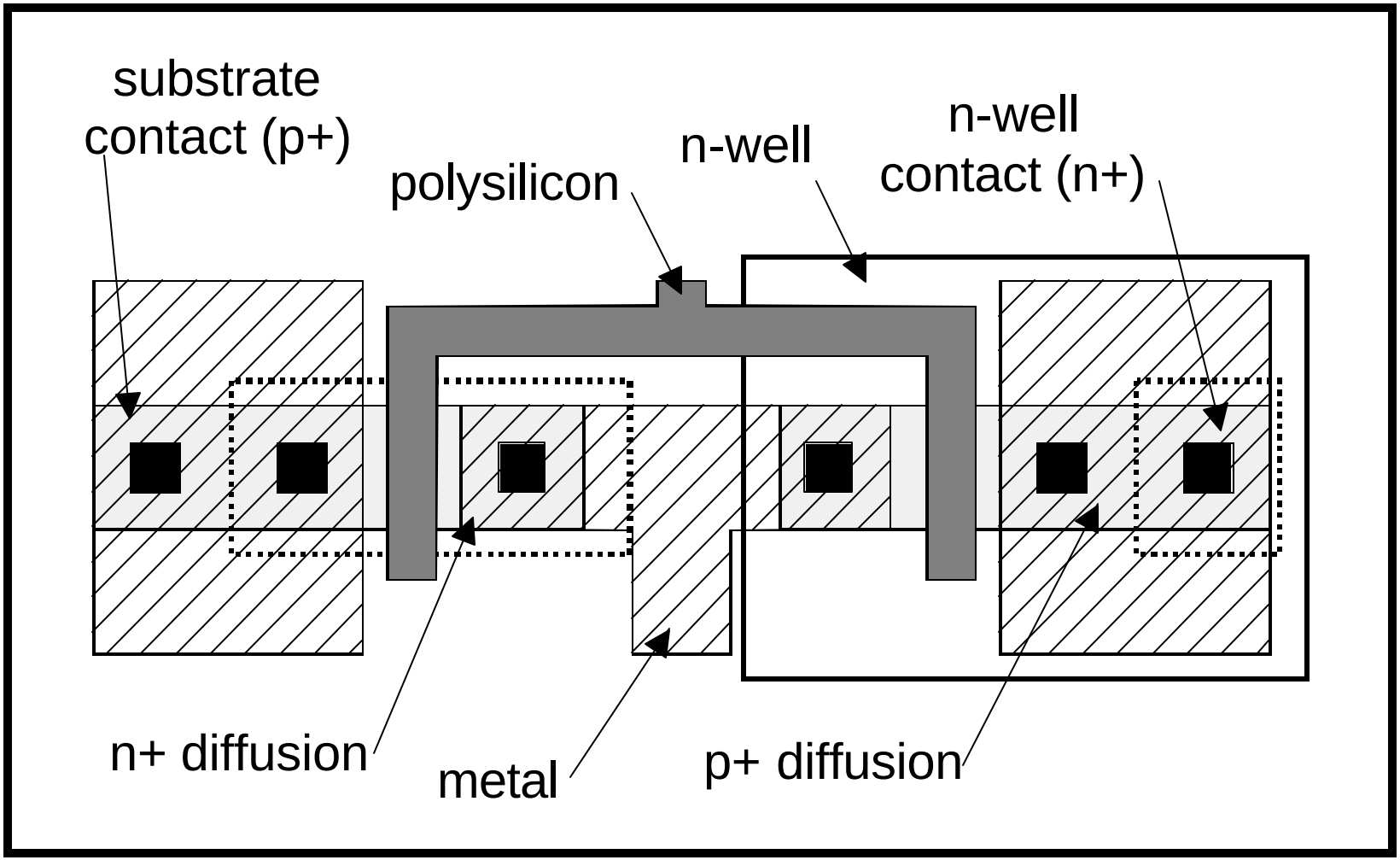
Dynamic power



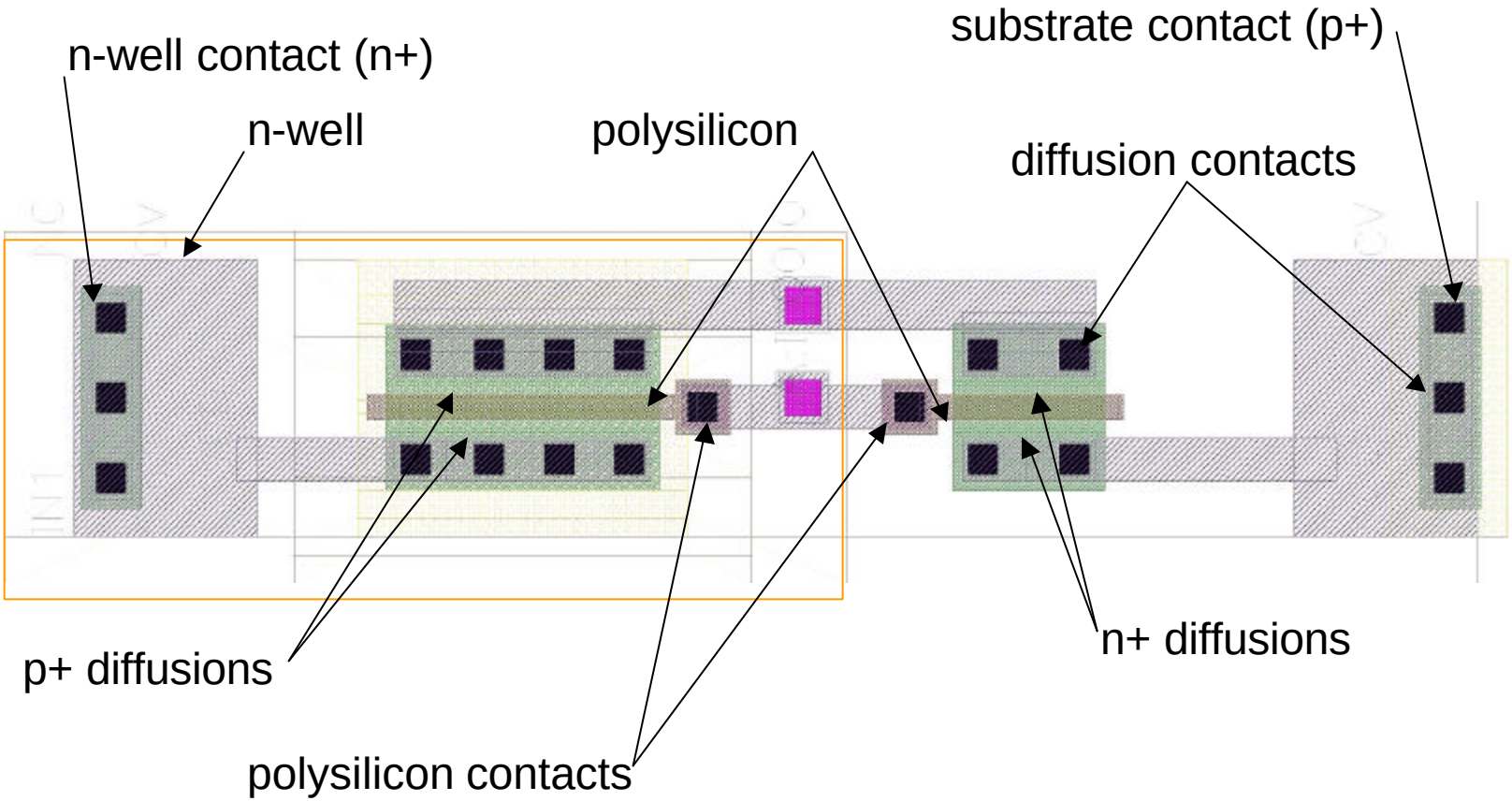
$$\text{E} = \text{Energy / transition} = \frac{1}{2} \cdot C_L \cdot V_{dd}^2$$

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

The CMOS inverter



The CMOS inverter



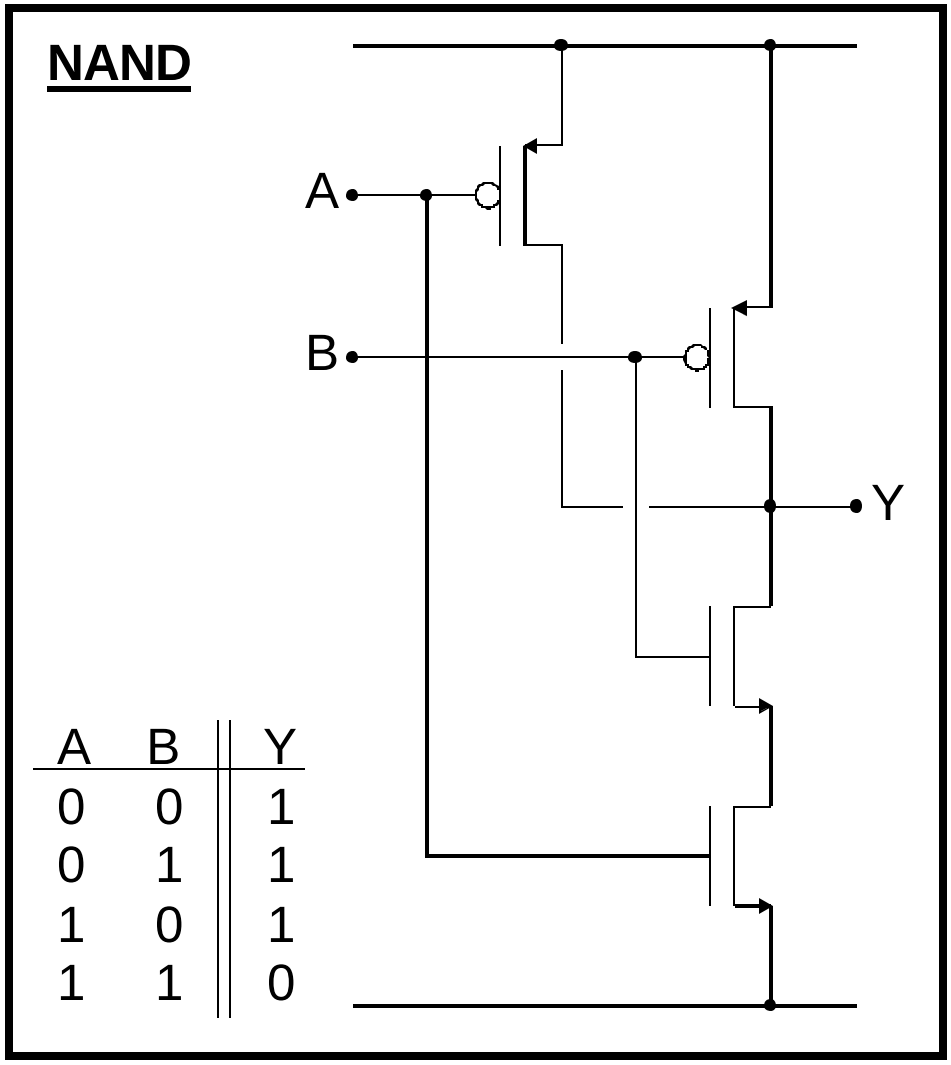
Outline

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

“Just like LEGO”

- The NAND gate
- “Reading” CMOS gates
- Designing CMOS gates

NAND 2-inputs

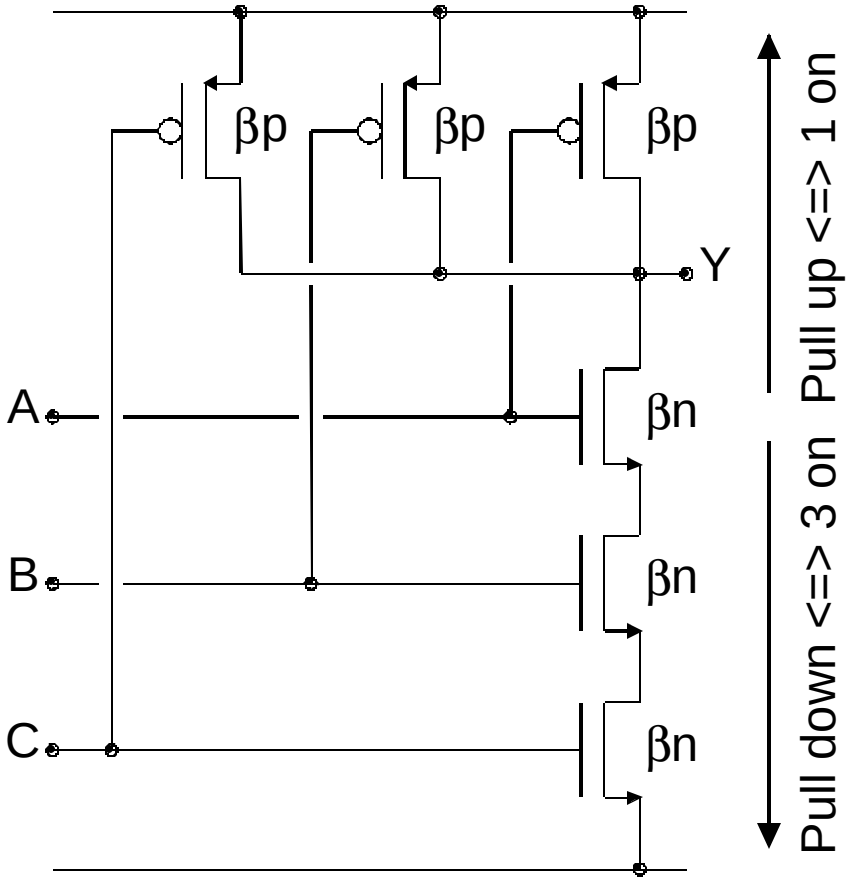


A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

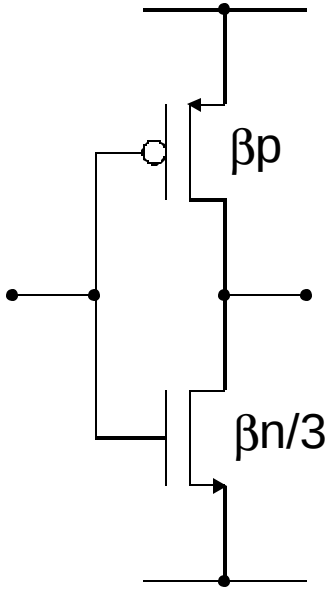
"Gates are inverters in disguise!"

NAND 3-inputs

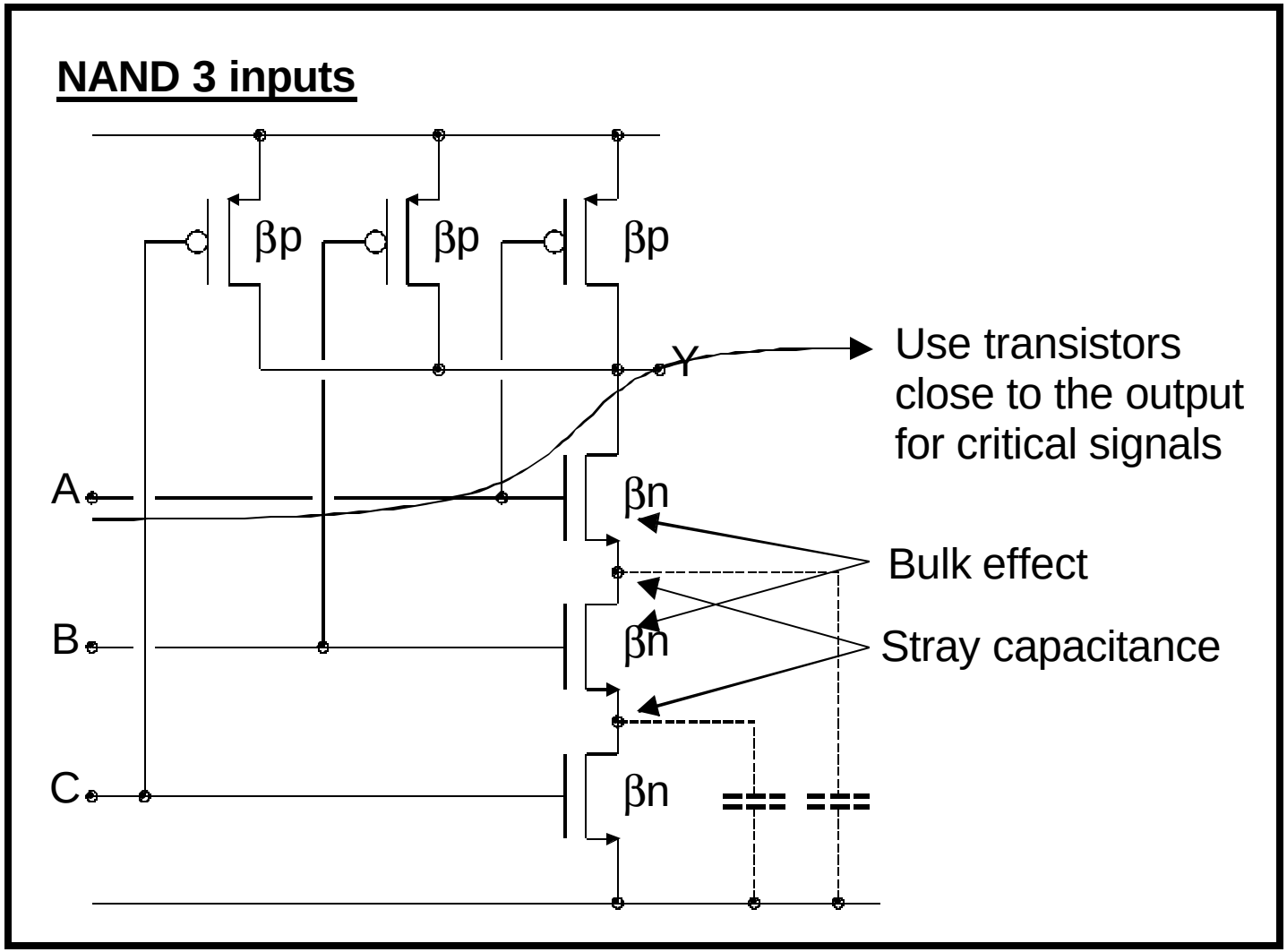
NAND 3 inputs



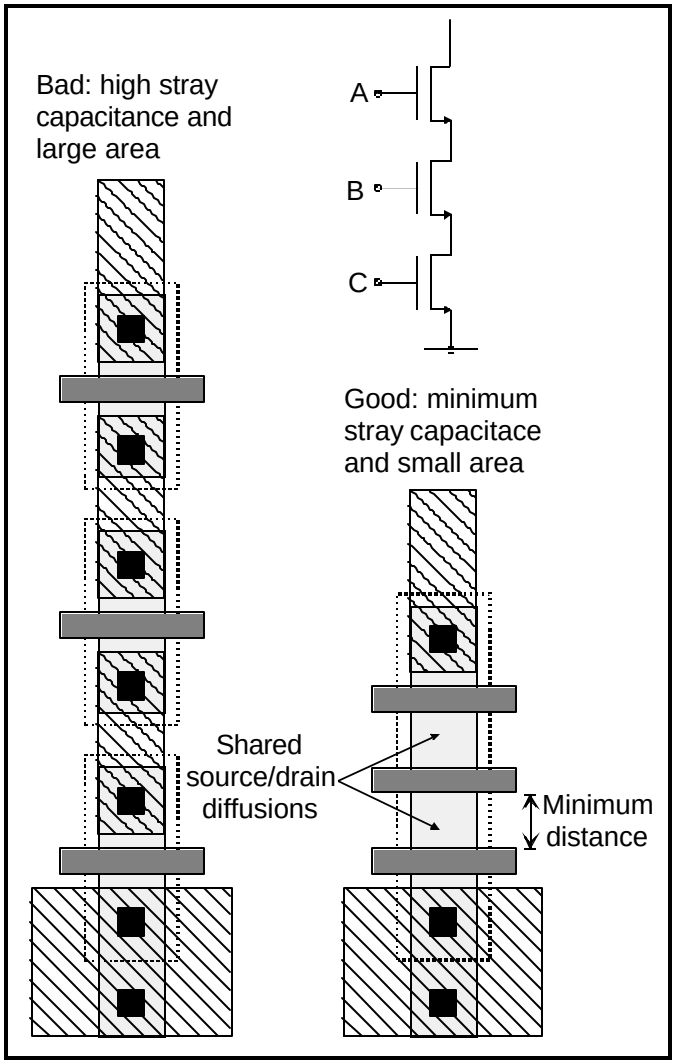
"Delay equivalent" inverter



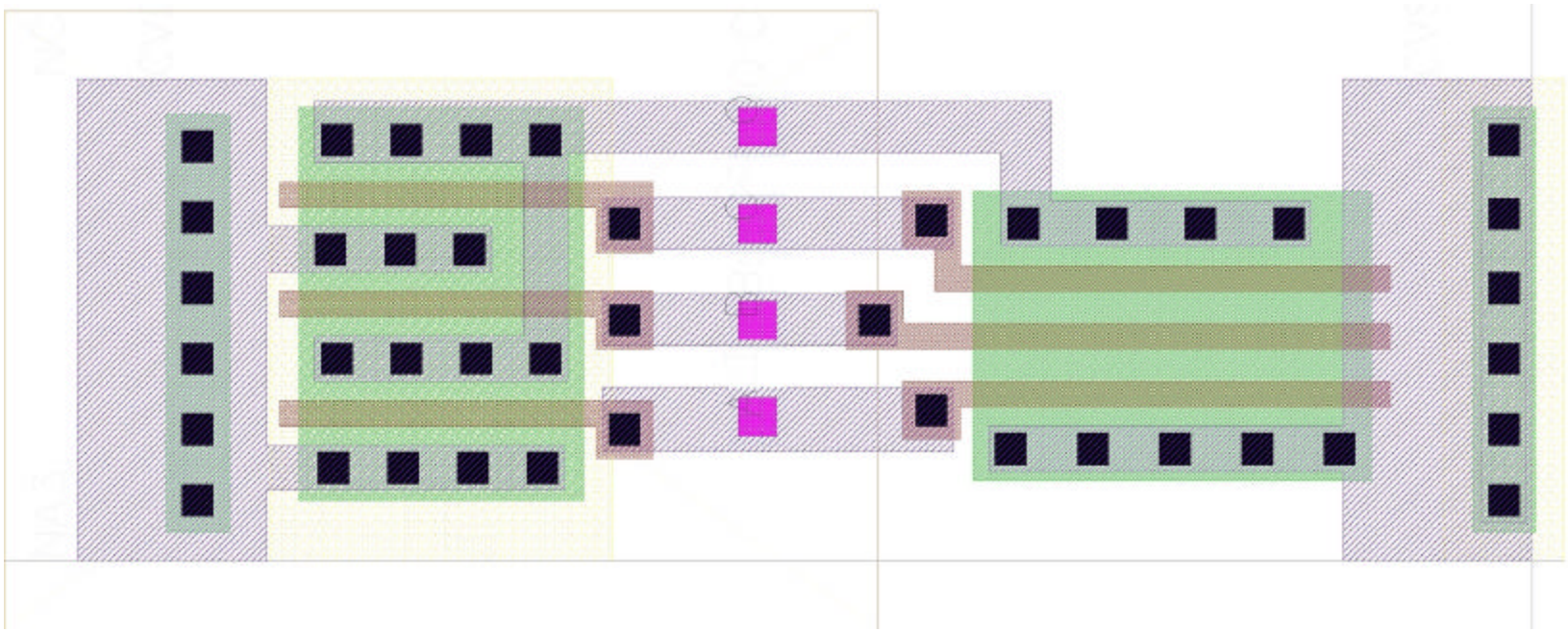
NAND 3-inputs



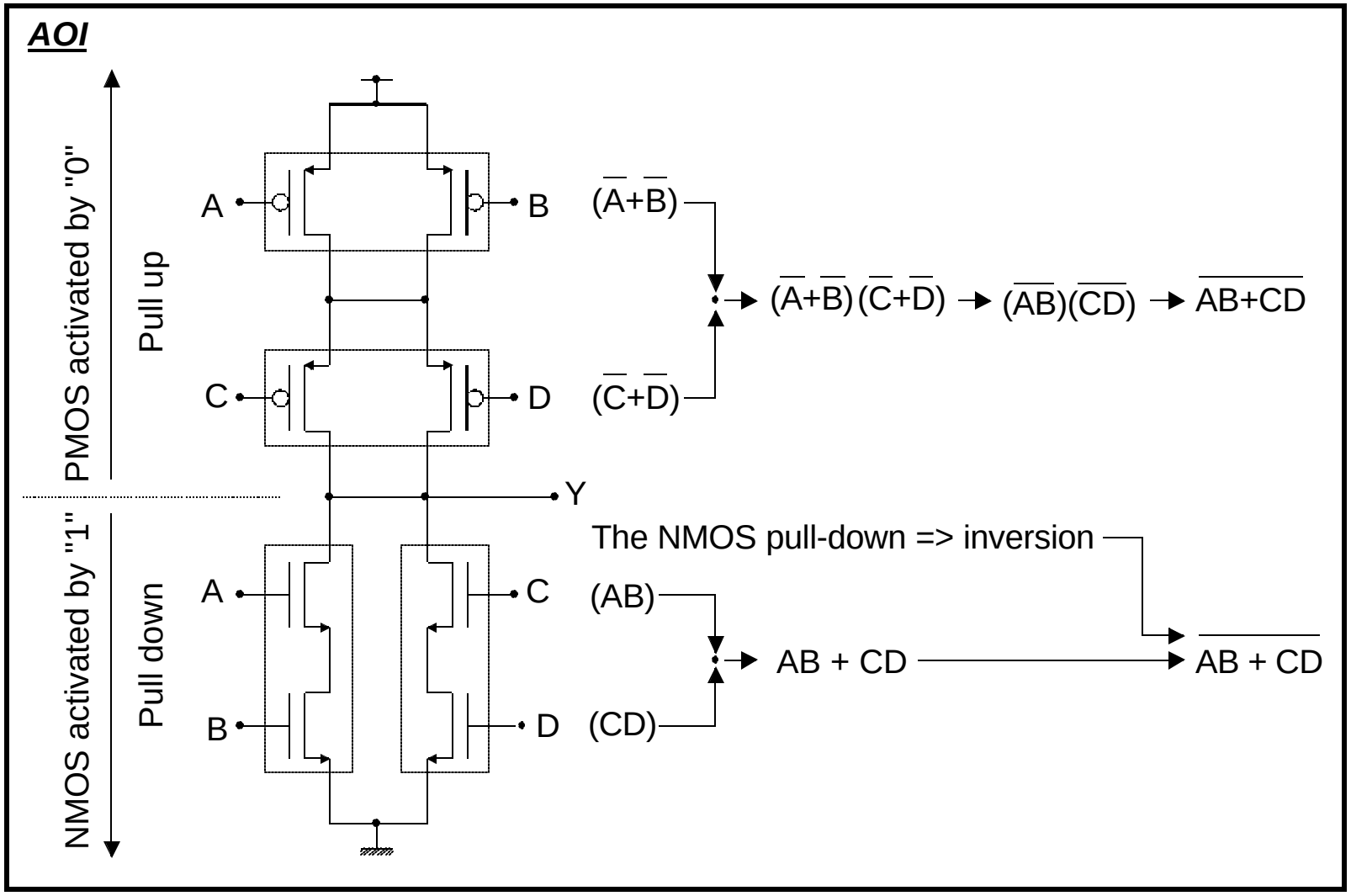
NAND 3-inputs



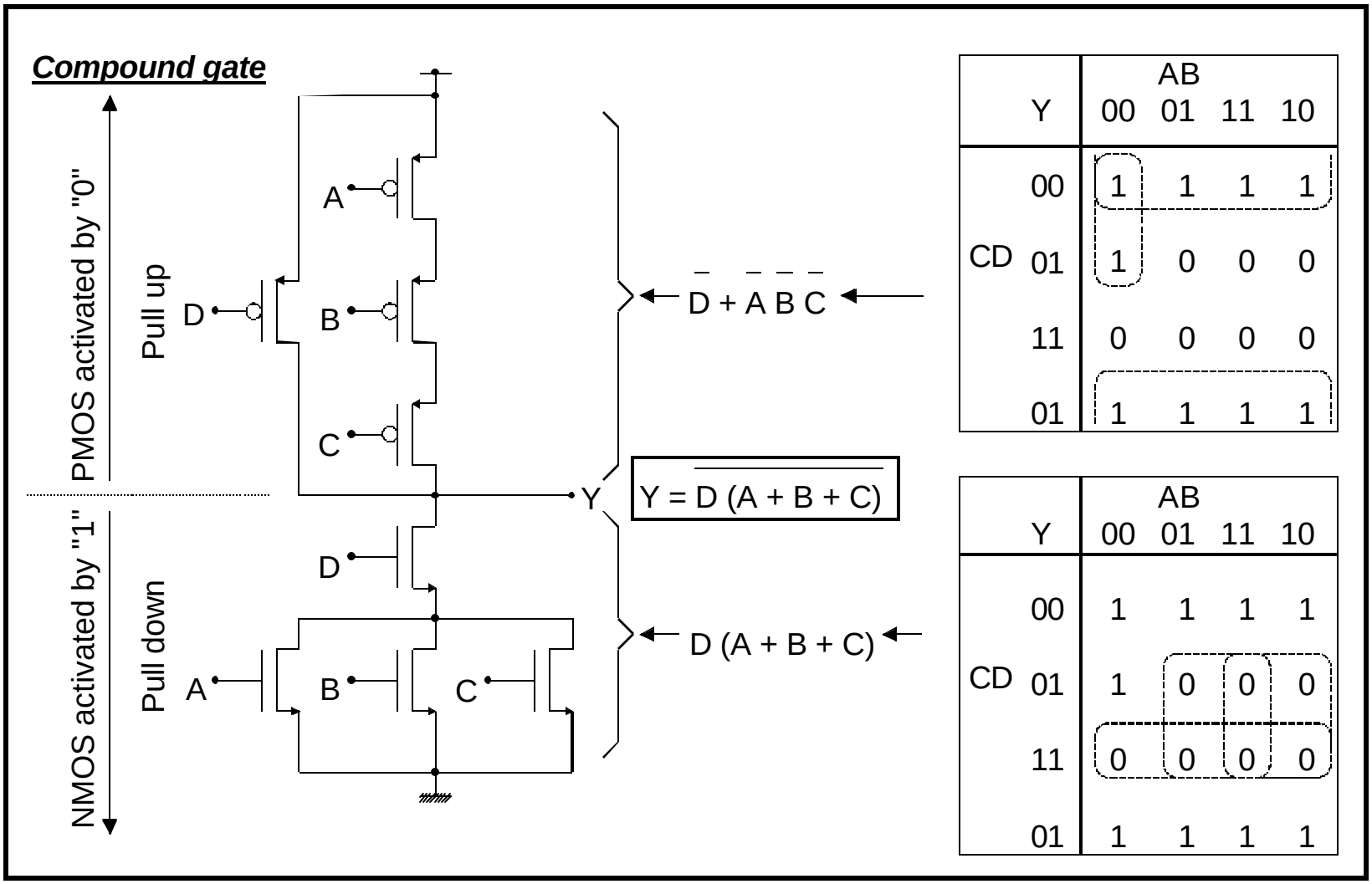
NAND 3-inputs



“Reading” CMOS gates



Designing CMOS gates



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

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)