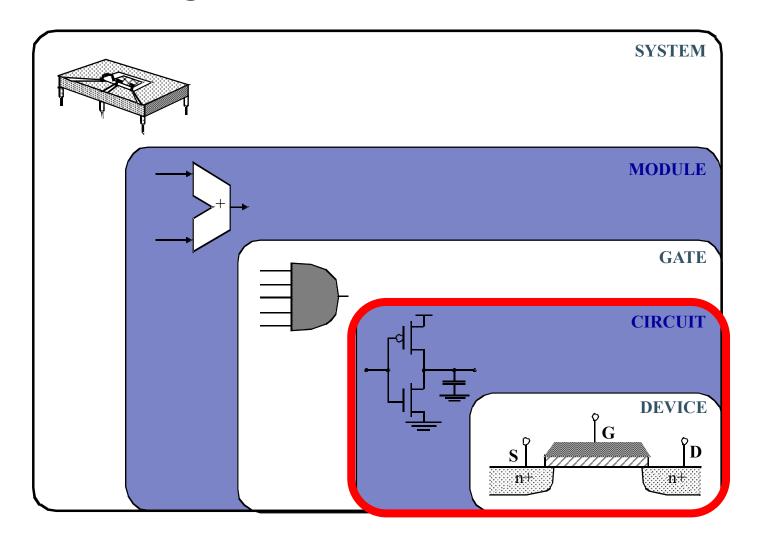
CG2027 Transistor-Level Digital Circuits

Handout #2: CMOS Inverter

National University of Singapore Kelvin Fong



Design Abstraction Levels





Lecture Overview

In this lecture, you will learn about

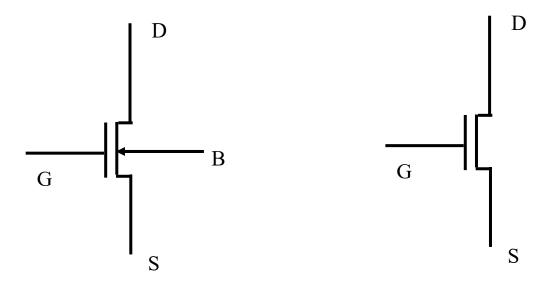
- Basic CMOS inverter
- Analyze switching and speed of the inverter
- Designing an inverter cell



Basic Notations

- In CMOS technologies, both n-type (n-channel) and p-type (p-channel)

 Metal-Oxide-Semiconductor field-effect transistors (MOSFETs) are used.
- For MOSFETs, S denotes <u>Source</u>, D the Drain, G the Gate and B the body or substrate terminals. Small or capital letters are equivalent for voltages.

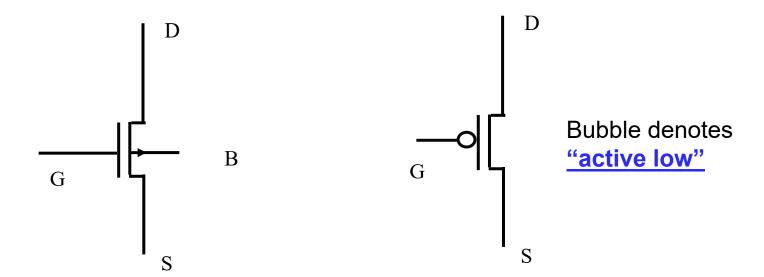


- n-type transistor : substrate is p-doped (acceptor type) and is connected to -ve supply V_{SS} (usually 0V or ground denoted as GND).
- For simplicity, as substrate bias is usually fixed, removal of substrate (body) contact leads to the compact notation on the right.



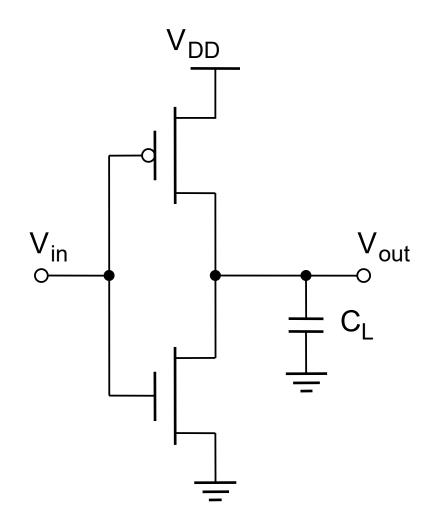
Basic Notations

- p-type transistor : substrate is n-doped (donor type) and is connected to +ve supply (usually power supply voltage V_{DD}).
- For simplicity, as substrate bias is usually fixed, removal of substrate (body) contact leads to the compact notation on the right.



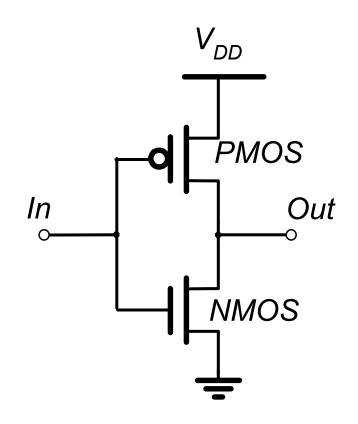


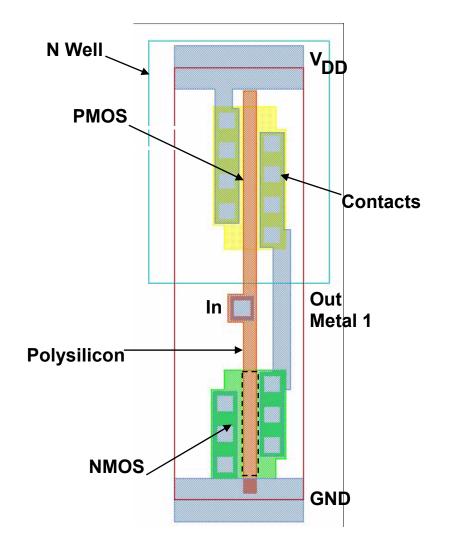
The CMOS Inverter: A First Glance





CMOS Inverter

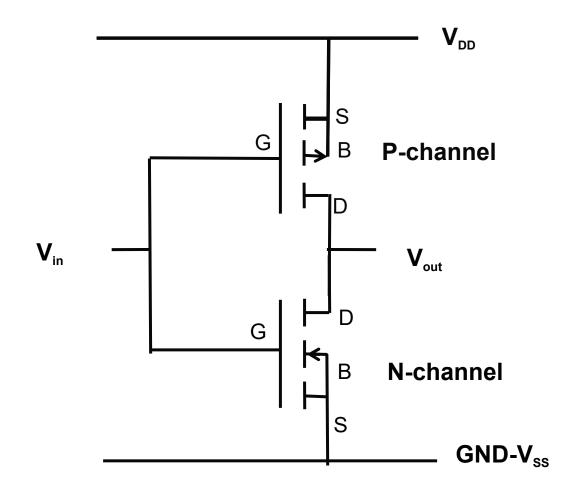




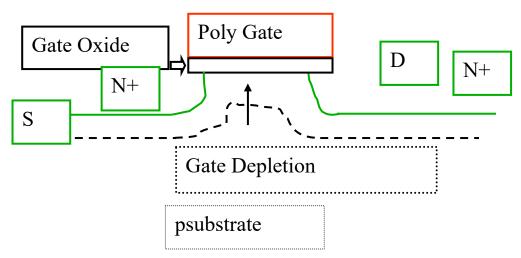


Basic Notations

- The following circuit shows another way to represent 4 terminal MOS device (MOSFET).
- Basic simplest circuit block in CMOS technology is an inverter that uses both devices. The schematic of CMOS inverter is shown below.

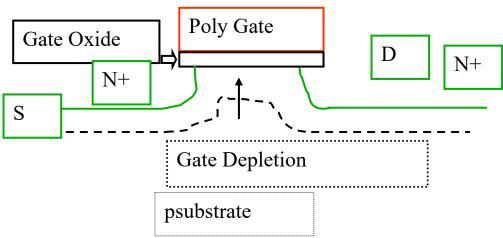






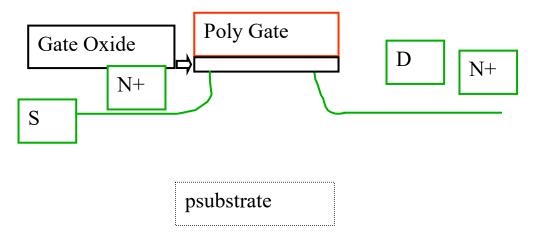
- Physical structure of MOSFET is very different as seen in earlier lectures and is schematically shown as above. The carriers are holes below the gate naturally and are positively charged.
- However, MOSFET conduction is done by electrons. Thus, the surface below the gate
 is inverted and electrons will be present when MOSFET is turned on. The MOSFET
 turns on when V_{qs} is above a predefined threshold voltage V_t.
- Source and Drain have a large number of electrons indicated with N+ symbol.
- If V_{gs} is negative, positively charged holes in the substrate will be attracted to Si-SiO₂ interface. Accumulation.





- As V_{GS} increases, the interface comes out of accumulation and holes are repelled and depleted below the gate oxide. The gate depletion (or space charge region) shown by dashed line under the gate becomes deeper with increasing V_{GS} as more holes will now be depleted.
- If V_{GS} increases further, eventually the device enters the regime of <u>strong</u> <u>inversion</u> where there is a large electron concentration (opposite to the substrate carriers holes) below the oxide. In effect strong electric field of the gate attracts electrons from the source into the channel. The gate depletion now stops deepening. The threshold voltage V_t of MOS device is defined as the V_{gs} that leads to strong inversion. After this V_{gs}, the device conducts with the help of the electrons in source, below the oxide and drain for positive V_{ds}.





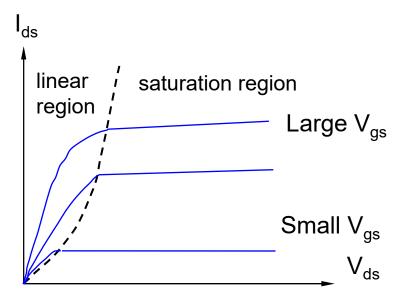
- Thin gate oxide always desirable for better gate control as lower thickness brings the gate closer to the conduction channel. As thin as 1.2 nm for technologies below 0.1 micron (13nm today).
- Polysilicon gate helped to achieve better process integration in the past. Metal gate and high-k get oxide help in the current technologies.
- Large electron concentration source drain regions for low resistance.
- Low resistance metal connections for source, drain, body (same as substrate or well or tub - different notations).



- P-channel device has opposite relative voltages, has P+ (heavily p doped) source-drains, and n-type substrate. Both devices are used in CMOS technology which is our focus. P-channel device normally carries less current for similar dimension and voltages compared to n-channel due to lower mobility of holes.
- For designers, it is important to know how the <u>drain current ID</u> changes when device dimensions, V_{GS} and V_{DS} change. Under simplified assumptions, analytic expressions for this current can be derived which we will use for design
- I-V plots and analytic expressions will now be discussed followed by inverter cell design and timing calculations with parasitic (unwanted capacitors and resistors which cannot be eliminated) estimation.
- Much later, we will introduce basic knowledge on what determines device limitations and how to logically develop a digital a circuit to implement it in Silicon.



MOS Transistor Characteristics



ullet Simple Model: cut-off region : $V_{gs} - V_t < 0$

$$I_{ds} = 0$$

(V_t is the threshold voltage, also denoted as V_{TN} for n-channel MOSFET and V_{TP} for p-channel MOSFET. Normally, V_{TN} >0 and V_{TP} <0).

In the linear region where $0 \le V_{ds} \le (V_{gs} - V_t)$, Also called <u>Triode</u> region.

$$I_{ds} = \beta \left[\left(V_{gs} - V_t \right) V_{ds} - \frac{V_{ds}^2}{2} \right]$$



MOS Transistor Characteristics

In the saturation region where $V_{ds} \ge (V_{gs} - V_t) > 0$,

$$I_{ds} = \frac{\beta}{2} \left(V_{gs} - V_t \right)^2$$

where
$$\beta = \frac{\mu \varepsilon}{t_{ox}} \left(\frac{W}{L} \right)$$

μ: effective surface mobility of charge carrier.

ε: permittivity of gate insulator.

t_{ox}: thickness of gate insulator.

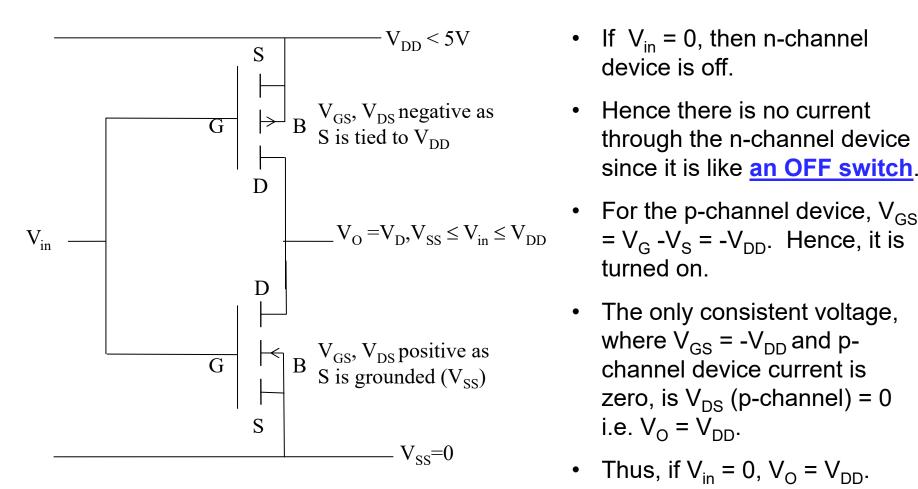
W: channel width.

L: channel length.



Basic Operation of an Inverter

The schematic of CMOS inverter is repeated below.



- If $V_{in} = 0$, then n-channel
- since it is like an OFF switch.
- For the p-channel device, V_{GS}
- The only consistent voltage, where $V_{GS} = -V_{DD}$ and pchannel device current is zero, is V_{DS} (p-channel) = 0 i.e. $V_0 = V_{DD}$.
- Thus, if $V_{in} = 0$, $V_O = V_{DD}$.



Basic Operation of an Inverter

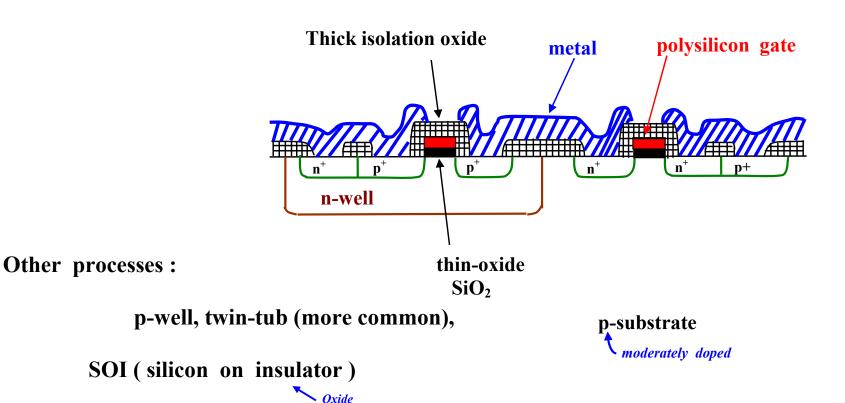
- On the other hand, if $V_{in} = V_{DD}$, then $V_{GS} = 0$ for the p-channel device.
- Since V_{GS} ≤ V_{TP} is a necessity for it to conduct, the device is off.
- Hence, there is no current through the n-channel device which is turned on by V_{in} = V_{DD} ≥ V_{TN}.
- Again, for this to be consistent, V_{DS} (n-channel) = 0 or V_O = 0.
- Hence if $V_{in} = V_{DD}$, $V_{O} = 0$.
- Thus it is clear that the circuit here performs a job of an inverter with $V_{in} = 0$ or $V_{in} = V_{DD}$. When input switches from say 0 to V_{DD} , output switches from V_{DD} to 0.
- The devices are in different regions of operation during the switch. The switching of the inverter will be discussed later.



Cross-section of CMOS Inverter

• Using simple isolation process, the inverter cross section including both NMOS and PMOS is as below. Two types of substrates.

n-well process





- Earlier, we drew a circuit of CMOS inverter and its cross section without worrying about how to make these devices and connect them on Silicon.
- Let us now discuss how many different layers will be needed on the layout at the minimum to achieve the cross section of inverter. Normally, many more layers are needed as there are additional steps and more than one metal layer.
- The starting substrate will normally have <u>one type of doping</u>. Here we use p-type substrate. Hence to make both p- and n-channel devices, we need to create an n-type regions using implant/diffusion in p-type substrate. This level is labeled as NWEL colored in brown.
- There should be a mask that defines gate regions using selective etching. This will normally be called as GATE or POLY mask named after polysilicon material normally used to form the gate.
- There should be a mask layer to form the N+ doped regions around the gate to make n-channel devices using implant/diffusion.



- There should be a mask layer to form the P+ doped regions around the gate to make p-channel devices using implant/diffusion.
- We need to connect these Source, Drain, Gate regions depending on the circuits to be implemented using <u>metal lines</u>. These metal lines are isolated from each other by an insulator such as oxide. For this step, oxide is deposited over whole silicon substrate wafer and holes are made in it over S/D/G by etching to make actual connections by filling these holes with the metal. This is achieved using contact mask.
- We still need to connect these Source, Drain, Gate regions depending on the circuits to be implemented using metal lines. For this step, metal is deposited over whole silicon substrate wafer and the metal is etched from unwanted areas. This is achieved using metal mask.
- There can be several layers of contacts and metal mask pairs up to 8 in advanced technologies. However, for simplicity, we will just confine to maximum 2 layers of metal.



- Also, note that many other considerations may add some process complexities or masks to the above basic set.
- The color convention is now given in the table for the minimum set. The convention for poly, N type active (S/D) area, contact and metal is universal. Others may vary a lot.

Layer Color

NWEL Brown

POLY (Gate) Red

N+ (NSD-N type Source-Drain) Green

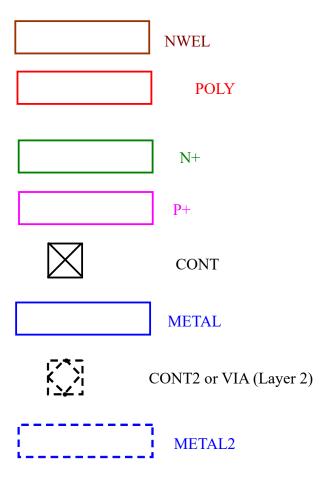
P+ (PSD-P type Source-Drain) Pink/Purple

CONT Black

METAL Blue

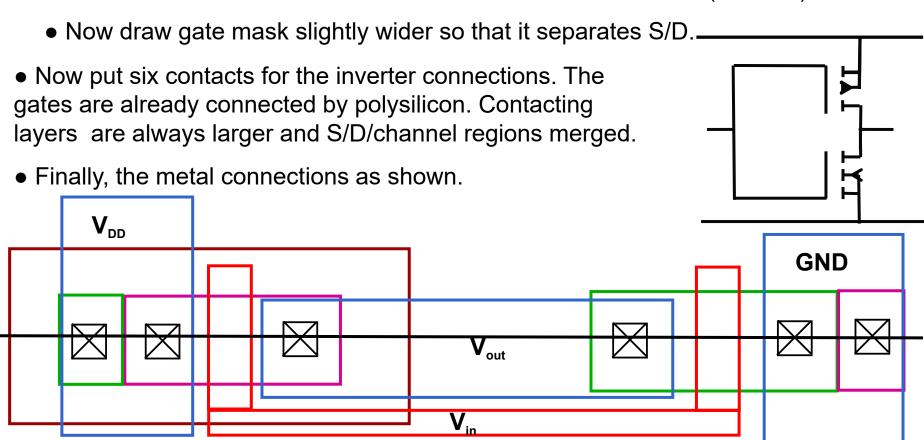


Layout diagram colour code in the process sequence order





- Let us now evolve a primitive, stand alone, layout of the inverter.
- First form n-well which gives n-type substrate within p-type wafer. This has to accommodate all of p-type MOSFETs.
- Now form S/D regions for both devices. They should be wide enough to accommodate connections for the inverter and are continuous (slide-40).



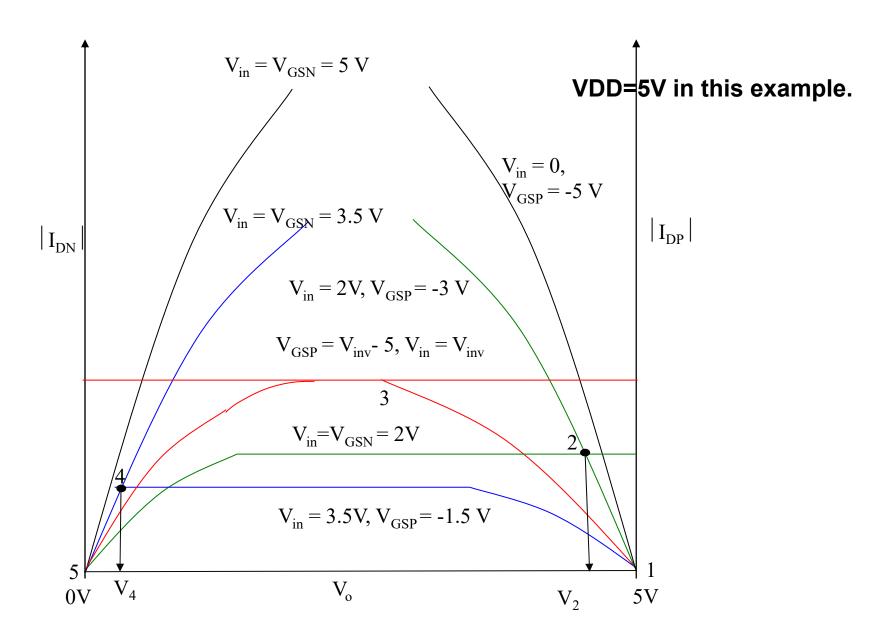
Now we study how regions of operations changes when inverter switches.

As V_{in} changes, these devices go through five different combinations of <u>regions of operation</u>. These combinations are indicated below.

- 1. $0 \le V_{in} \le V_{TN}$ n-channel off, p-channel on
- 2. $V_{TN} < V_{in} < V_{inv}$ p-channel linear, n-channel saturation
- 3. $V_{in} = V_{inv}$ Both in saturation
- 4. $V_{inv} < V_{in} < V_{DD} V_{TP}$ n-channel linear, p-channel saturation
- 5. V_{DD} $V_{TP} \le V_{in} \le V_{DD}$ p-channel off, n-channel on

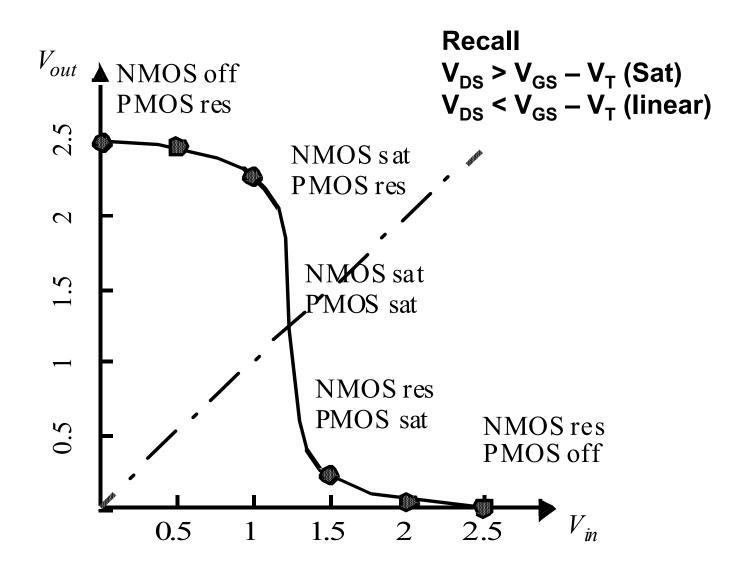
Where V_{inv} is called as <u>switching voltage</u>. The characteristics are derived based on the fact that both MOSFETs in the inverter here carry the same amount of current.



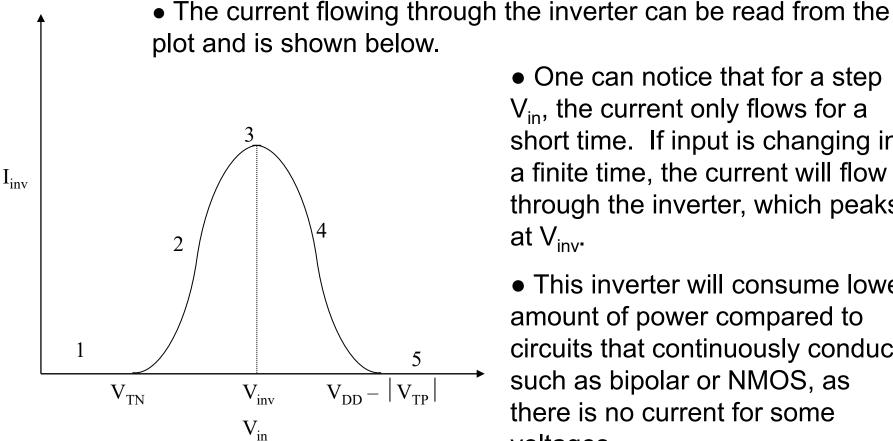




CMOS Inverter VTC



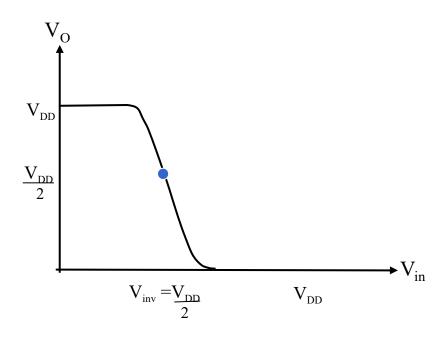




- - One can notice that for a step V_{in}, the current only flows for a short time. If input is changing in a finite time, the current will flow through the inverter, which peaks at V_{inv} .
 - This inverter will consume lower amount of power compared to circuits that continuously conduct such as bipolar or NMOS, as there is no current for some voltages.
- If there is no switching (Idle), there is no power consumption, in principle, a huge benefit for VLSI which made CMOS dominant.
- In reality, there is some power consumption due to sub-threshold currents, reverse biased diode currents and continuous running clocks.



• A simplified Voltage Response or voltage transfer characteristics (VTC) is as shown.



• Most change occurs near V_{inv} as the current here is the largest and aids in a fast transition. The slope at any point on the curve is the <u>gain</u> and is much larger than <u>unity</u> in magnitude at the blue switching voltage point.



Switching voltage of the inverter.

- At switching voltage, the <u>saturation</u> current of n- and p- channel devices are equal.
- Assume that the threshold voltages of n- and p- transistors have the same magnitudes. Then

For n-transistor in saturation,
$$I = \frac{\beta_n}{2} (V_{in} - V_t)^2$$

For p-transistor in saturation,
$$I = \frac{\beta_p}{2} (V_{DD} - V_{in} - |V_t|)^2$$

For switching at
$$V_{in} = \frac{V_{DD}}{2}$$
, $I = \frac{\beta_n}{2} \left(\frac{V_{DD}}{2} - V_t \right)^2 = \frac{\beta_p}{2} \left(\frac{V_{DD}}{2} - |V_t| \right)^2$

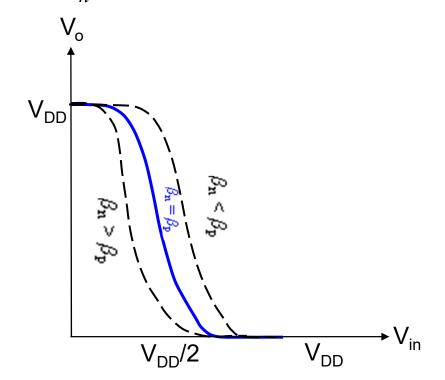
i.e.
$$\beta_n = \beta_p$$



$$\beta_n = \beta_p \Rightarrow \frac{\mu_n \varepsilon}{t_{ox}} \frac{W_n}{L_n} = \frac{\mu_p \varepsilon}{t_{ox}} \frac{W_p}{L_p} \Rightarrow \frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}$$

as normally $L_n = L_p$ in CMOS digital design.

 $\therefore \frac{W_p}{W_n} \approx 2 \text{ based on electron and hole mobilities.}$

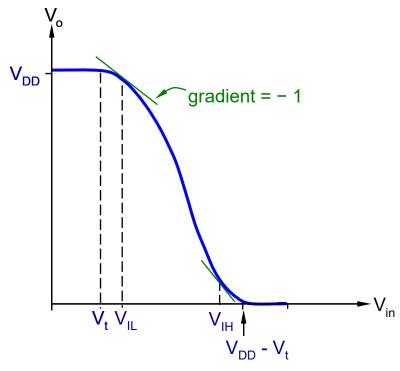


• The effect of sizes is best depicted in the figure.

Any observations?



- Noise margin in digital circuits is very important and should be as large as possible. We have studied it earlier. Here $V_{OL} = 0$, $V_{OH} = V_{DD}$.
- It can be easily deduced from output response as the output is not likely to switch if magnitude of dV_o/dV_{in} is <1 as the output voltage V_o changes slowly with the input voltage V_{in} .

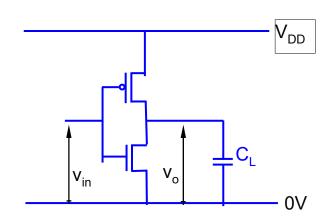


- V_{IL}= maximum LOW input voltage
- V_{IH} = minimum HIGH input voltage
- Noise margin (low) = V_{IL}
- Noise margin (high) = V_{DD} V_{IH}

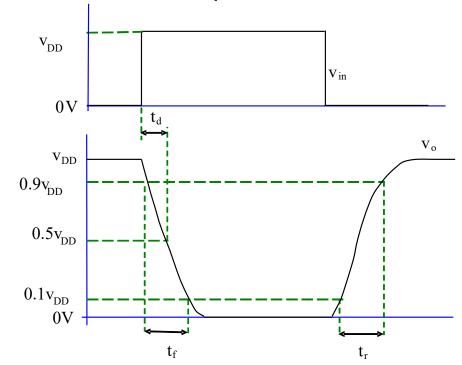


Loaded CMOS inverter

- The above conditions in switching studies are all ideal. Normally, CMOS inverter will have parasitic resistances and capacitances attached to the output node. Hence principle of operation of inverter alters a bit.
- The capacitor loaded CMOS inverter and its response is shown below.



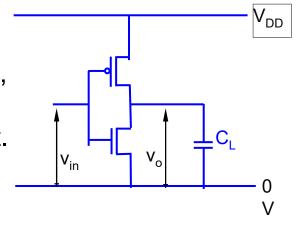
• We define: t_d : delay time, t_f : fall time and t_r : rise time as before





ullet If input was at low voltage, the output is high and the capacitor C_L holds the charge to maintain this high voltage.

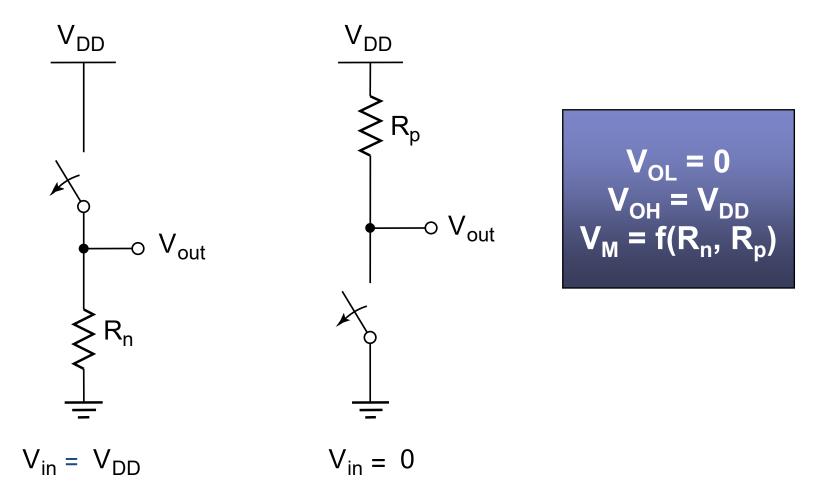
• If input now switches instantaneously to high, the capacitor voltage does not change instantaneously as this requires infinite current. Instead, at the switching time, both the gate voltage and the drain voltage are high for the NMOS transistor which turns on.



- The NMOS device current will discharge the capacitor to zero after a delay time. This is why NMOS device is called as a <u>pull down device</u>.
- Once the output voltage reaches zero, NMOS transistor stops conducting as $V_{DS} = 0$. The output stays at zero volts after this delay and there is no current flowing through any part of the circuit.
- Exactly opposite cycle occurs when input becomes low and PMOS device charges the capacitor to raise output voltage to the power supply voltage. Similarly, PMOS device is called as a <u>pull up device</u>.

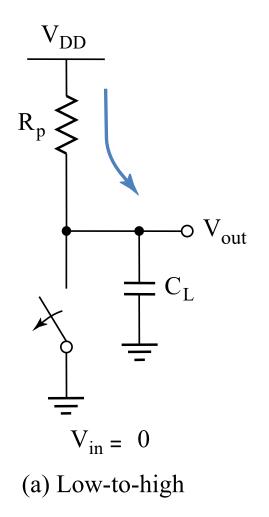


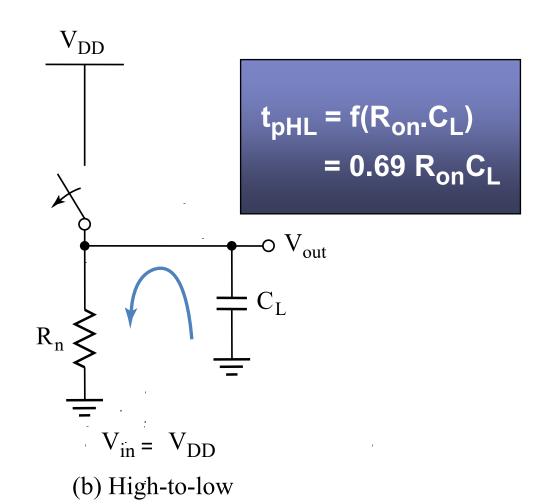
CMOS Inverter: First-Order DC Analysis





CMOS Inverter: Transient Response







POWER DISSIPATION



Where Does Power Go in CMOS?

Dynamic Power Consumption

Charging and Discharging Capacitors

Short Circuit Currents

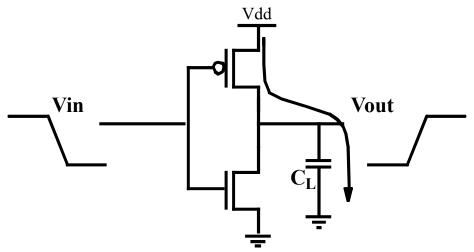
Short Circuit Path between Supply Rails during Switching

Leakage

Leaking diodes and transistors



Dynamic Power Dissipation



$$E_{VDD} = \int_0^\infty i_{VDD}(t) * V_{DD} dt$$

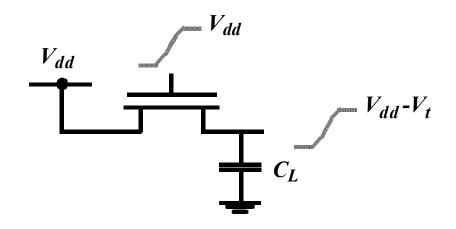
$$E_{C_L} = \int_0^\infty i_{VDD}(t) * v_{out} dt$$

Energy/transition = $C_L * V_{dd}^2$

Power = Energy/transition * f =
$$C_L * V_{dd}^2 * f$$

- Not a function of transistor sizes if C_L dominated by connection metal!
- Need to $\frac{\text{reduce}}{\text{c}} C_L$, V_{dd} , and f to reduce power.

Modification for Circuits with Reduced Swing



$$E_{0 \to 1} = C_L \cdot V_{dd} \cdot (V_{dd} - V_t)$$

• Can exploit reduced swing to lower power (e.g., reduced bit-line swing in memory)



Node Transition Activity and Power

• Consider switching a CMOS gate for N clock cycles

$$E_{N} = C_{L} \cdot V_{dd}^{2} \cdot n(N)$$

 E_N : the energy consumed for N clock cycles

n(N): the number of 0->1 transition in N clock cycles

$$P_{avg} = \lim_{N \to \infty} \frac{E_N}{N} \cdot f_{clk} = \left(\lim_{N \to \infty} \frac{n(N)}{N}\right) \cdot C_L \cdot V_{dd}^2 \cdot f_{clk}$$

$$\alpha_0 \to 1 = \lim_{N \to \infty} \frac{n(N)}{N}$$

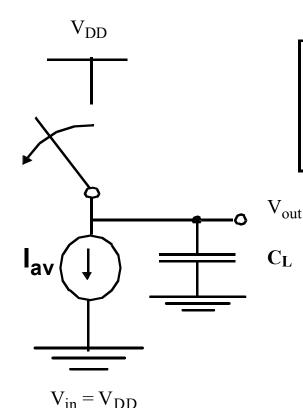
$$P_{avg} = \alpha_{0 \to 1} \cdot C_{L} \cdot V_{dd}^{2} \cdot f_{clk}$$



CMOS Inverter Propagation Delay Approach 1 (Current Avg)

$$t_p = C_L \int_{v_1}^{v_2} \frac{dv}{i(v)}$$

$$t_p = \frac{C_L(v_2 - v_1)}{I_{av}}$$



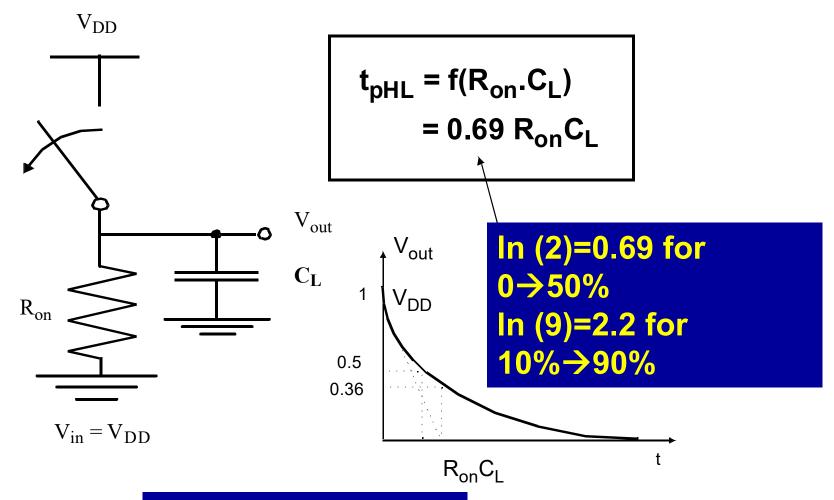
$$t_{pHL} = \frac{C_L V_{swing}/2}{I_{av}}$$

$$\sim \frac{C_L}{\beta_n V_{DD}}$$

Note) t_p: 50% of full swing



CMOS Inverter Propagation Delay Approach 2 (RC relaxation)





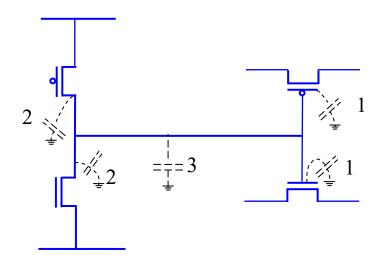
41

LOAD (CAPACITANCE / RESISTANCE)



Load capacitance estimation

- When a CMOS gate drives another CMOS gate, the driver "sees" a capacitive load. The load capacitance consists of
- (1) gate capacitance of the load from two MOSFETs on the right,
- (2) junction capacitance from common Drain of the driver inverter on the left, and
- (3) the parasitic routing capacitance from the metal line used for connection.





• Gate capacitances actually depend on terminal voltages. The simplest estimation in different regions is below. However many times a constant value is used.

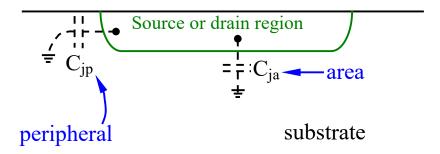
• In summary

C_{gd}	C_{db}
C_{gs}	$\mathcal{C}_{\mathrm{sb}}$
	C _{gb}

	Capacitance						
Parameter	Off	Linear	Saturation				
C_{gb}	$\frac{\varepsilon_{sio_2}A}{t_{ox}}$	0	0				
C_{gs}	0	$\frac{\varepsilon_{sio_2}A}{2t_{ox}}$	$\frac{2\varepsilon_{sio_2}A}{3t_{ox}}$				
$C_{ m gd}$	0	$rac{arepsilon_{sio_2}A}{2t_{ox}}$	0				



Source or drain junction capacitance estimation (2)



- <u>Peripheral</u> capacitance is determined by the perimeter of Source or Drain whereas the area capacitance is determined by the area. Both area and perimeter are determined from the geometry of source or drain in the circuit layout which we have seen earlier.
- The total junction capacitance C_i is given by

$$C_{j} = area * C_{ja0} \left(1 - \frac{V_{j}}{\phi}\right)^{-ma} + perimeter * C_{jp0} \left(1 - \frac{V_{j}}{\phi}\right)^{-mp}$$



- Source or drain junction capacitance estimation (2)
- The total junction capacitance C_i is given by

$$C_{j} = area * C_{ja0} \left(1 - \frac{V_{j}}{\phi}\right)^{-ma} + perimeter * C_{jp0} \left(1 - \frac{V_{j}}{\phi}\right)^{-mp}$$

• where C_{ja0} and C_{jp0} zero voltage junction capacitance per unit area and perimeter tabulated below, respectively.

 V_j is the voltage across junction. (+ve for forward bias. -ve for reverse bias (MOS case)).

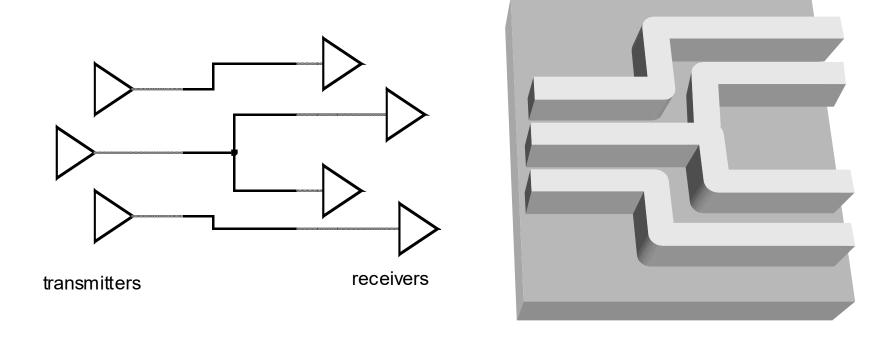
 ϕ = built-in junction potential \approx 0.6 \sim 0.9 V.

ma, mp (Grading coefficients) = $0.3 \sim 0.5$

Parameter	n-diffusion	p-diffusion
$C_{ m ja0}$	$0.1\mathrm{fF/\mu m^2}$	$0.1\mathrm{fF/\mu m^2}$
C_{jp0}	0.9 fF/μm	$0.8\mathrm{fF}/\mathrm{\mu m}$



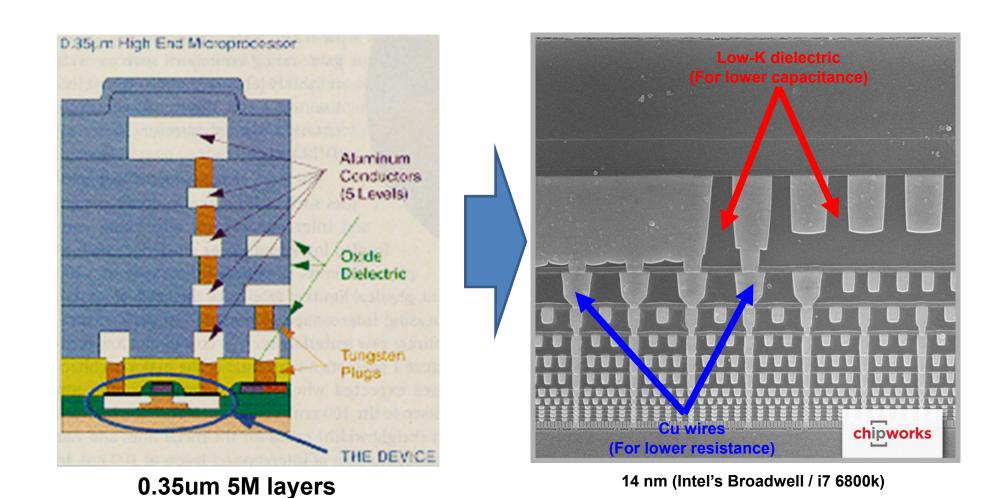
Load Estimation: The Wire



Schematics

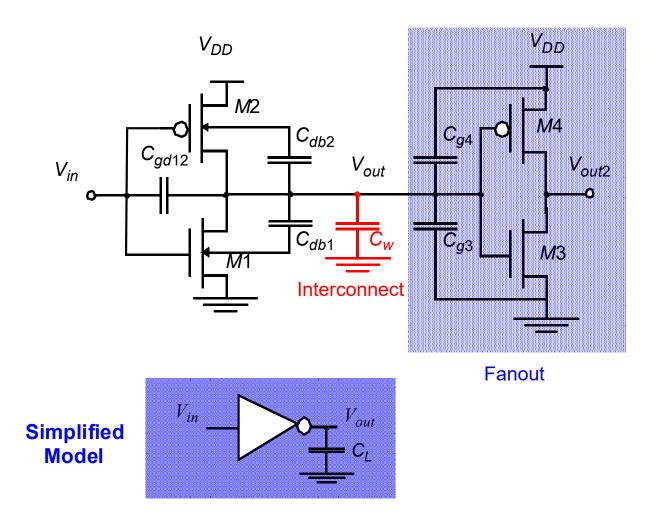
Physical

Interconnect Impact on Chip

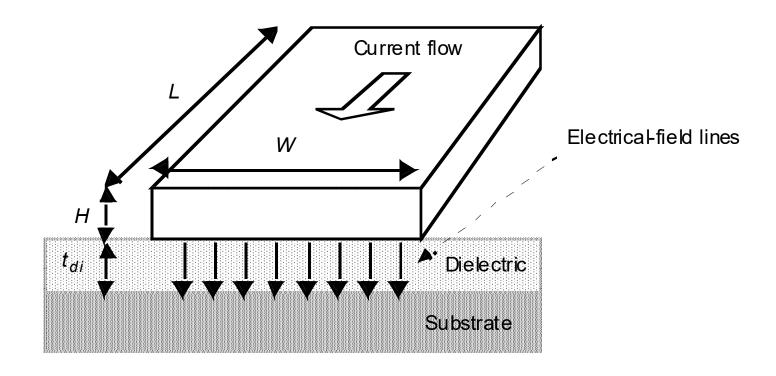


48

Capacitance of Wire Interconnect



Capacitance: The Parallel Plate Model



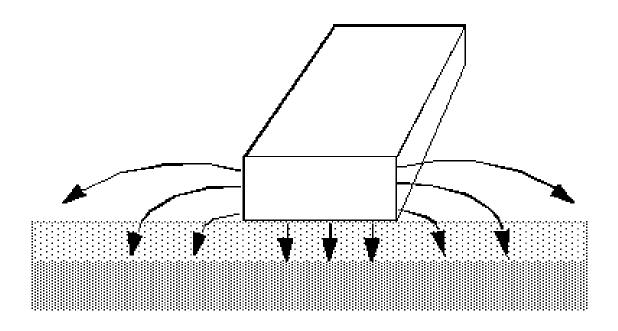
To calculate the parallel plate wire capacitance (C_{pp}) for this segment (length L, width w)

$$C_{pp} = \frac{\varepsilon_{di}}{t_{di}} WL$$

Permittivity

Material	ϵ_r
Free space	1
Aerogels	~1.5
Polyimides (organic)	3-4
Silicon dioxide	3.9
Glass-epoxy (PC board)	5
Silicon Nitride (Si ₃ N ₄)	7.5
Alumina (package)	9.5
Silicon	11.7

Fringing Capacitance

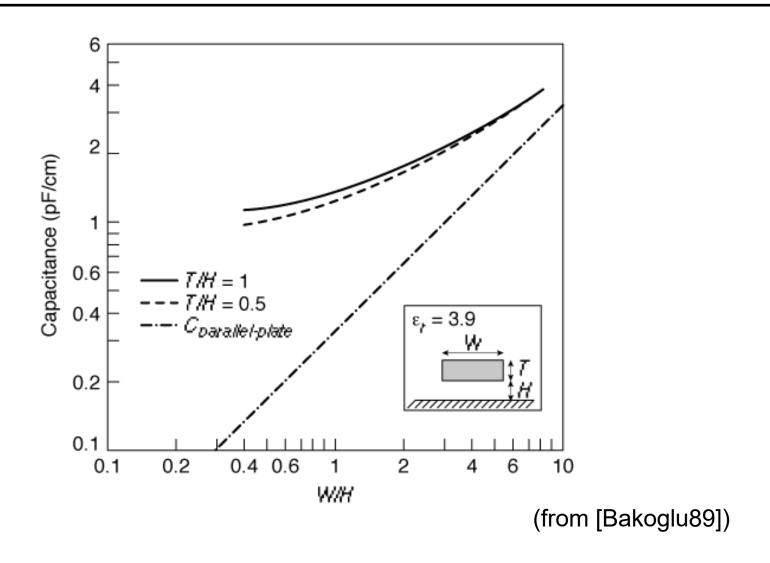


Wire capacitance per unit length:

$$c_{wire} = c_{pp} + c_{fringe} = \frac{w \varepsilon_{di}}{t_{di}} + \frac{2\pi \varepsilon_{di}}{\log(t_{di}/H)}$$

To calculate the total wire capacitance, you can multiply c_{wire} by the length L.

Fringing vs. Parallel Plate



Wiring Capacitances

CAPACITANCE PARAMETERS Area (substrate) Area (N+active) Area (P+active)	N+ 963	P+ 1159	POLY 110 8556 8309		м2 19 21	м3 13 14	м4 8 11	M5 8 9	мб 3 8	D_N_W	N_W 125	UNITS aF/um^2 aF/um^2 aF/um^2
Area (poly) Area (metal1) Area (metal2) Area (metal3) Area (metal4) Area (metal5) Area (r well) Area (no well)	987 143	202			17 41	10 15 36	7 10 14 39	5 7 9 15 36	4 5 7 9 14 37			aF/um^2 aF/um^2 aF/um^2 aF/um^2 aF/um^2 aF/um^2 aF/um^2
Fringe (substrate) Fringe (poly) Fringe (metal1) Fringe (metal2) Fringe (metal3) Fringe (metal4) Fringe (metal5) Overlap (P+active)	248	203	636	13 64	61 38 61	55 29 34 53	43 24 36 54	24 20 23 26 34 57	17 19 22 28 35 53			aF/um aF/um aF/um aF/um aF/um aF/um aF/um aF/um

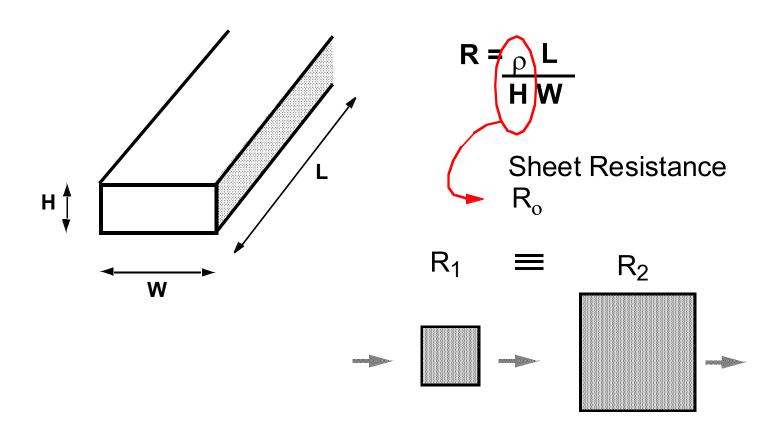
Example



Area Capacitance

Fringe Capacitance

Wire Resistance



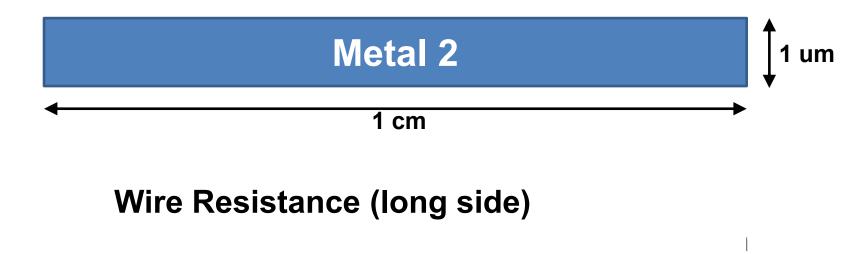
Sheet Resistance

Material	Sheet Resistance (Ω/□)
n- or p-well diffusion	1000 - 1500
n^+ , p^+ diffusion	50 - 150
n^+ , p^+ diffusion with silicide	3 – 5
n^+ , p^+ polysilicon	150 - 200
n^+ , p^+ polysilicon with silicide	4 – 5
Aluminum	0.05 - 0.1

PROCESS PARAMETERS Sheet Resistance Contact Resistance Gate Oxide Thickness	N+ 6.6 10.4 40	P+ 7.6 10.9	POLY 7.7 9.7	N+BLK 60.1	PLY+BLK 312.5	M1 0.08	M2 0.08 4.36	UNITS ohms/sq ohms angstrom
PROCESS PARAMETERS Sheet Resistance Contact Resistance	м3 0.08 9.08	POLY_H 1304.	6	м4 0.07 3.85	M5 0.07 19.37	м6 0.04 22.28	N_W 913	UNITS ohms/sq ohms

0

Example



Interconnect Resistance

Material	ρ (Ω-m)
Silver (Ag)	1.6×10^{-8}
Copper (Cu)	1.7×10^{-8}
Gold (Au)	2.2×10^{-8}
Aluminum (Al)	2.7×10^{-8}
Tungsten (W)	5.5×10^{-8}

Dealing with Resistance

- Selective Technology Scaling
- Use Better Interconnect Materials
 - reduce average wire-length
 - e.g. copper, silicides
- More Interconnect Layers
 - reduce average wire-length

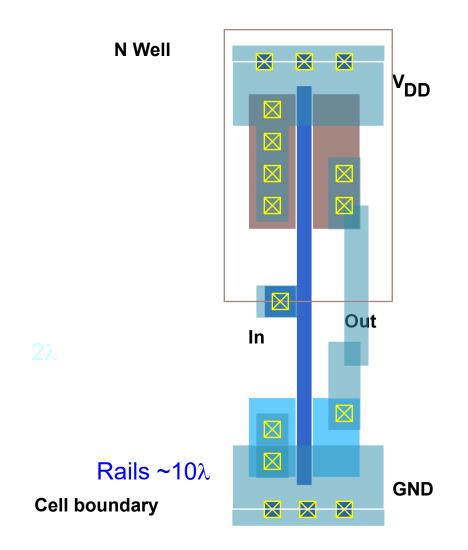
IR Drop

- Power ring design
 - Metal width
 - Choosing a metal layer

CMOS INVERTER LAYOUT



Inverter Layout: Standard Cells



Cell height 12 metal tracks

Metal track is approx. $3\lambda + 3\lambda$

Pitch = repetitive distance between objects

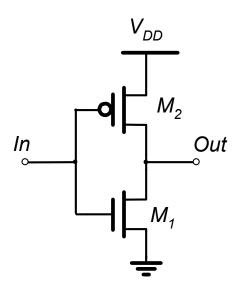
Cell height is "12 pitch"

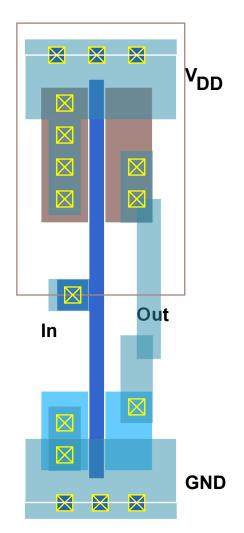
The design rules specify minimum dimensions and spacing. They are technology independent using a parameter λ which shrinks as technology scales.



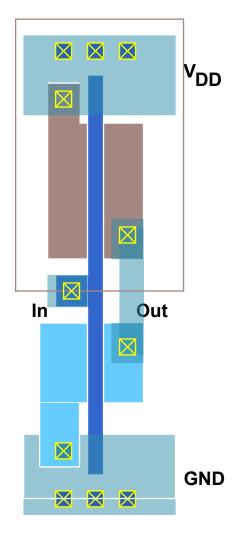
Standard Cells

With minimal diffusion routing



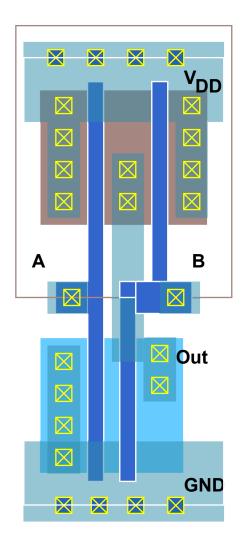


With silicided diffusion

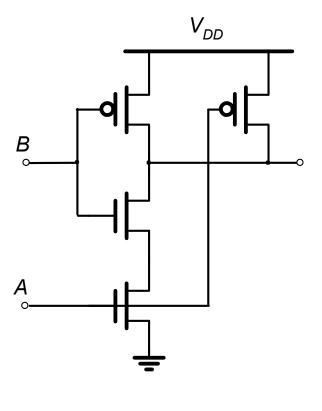




Standard Cells



2-input NAND gate

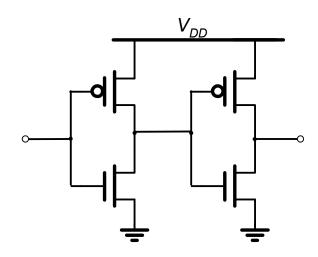


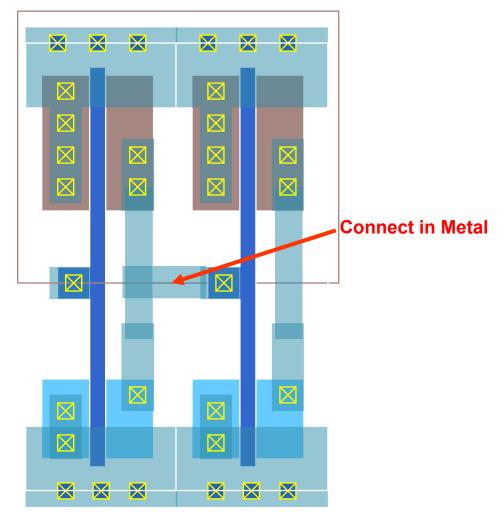


Two Inverters

Share power and ground

Abut cells

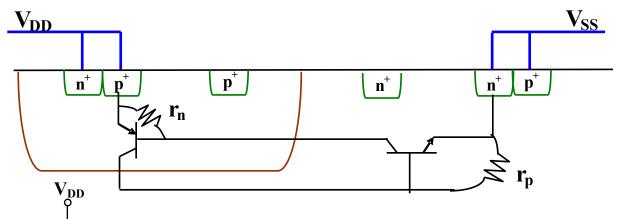






CMOS Latch-up

• It is clear that we have got many parasitic devices along with the MOSFET in the cross-section. Here is one such parasitic p-n-p-n device which is responsible for latch-up reliability issue in CMOS. It has one npn and one pnp bipolar transistor and 2 resistors as shown.



- The pnpn parasitic circuit above can be re-arranged as shown on the left.
- Due to <u>positive feedback</u>, this circuit can conduct large current between two power voltages and gate is immaterial.
- What can we do to suppress this?
- Gives rise to large design rules and extra features. The design rules are technology independent using a parameter λ which shrinks as technology scales.



r₁≥

 V_{SS}