

Lecture 23

CMOS LOGIC

Lectures 22:

- Review of Energy in Switching
- Efficiency
- Dynamic Power in CMOS
- Other Power

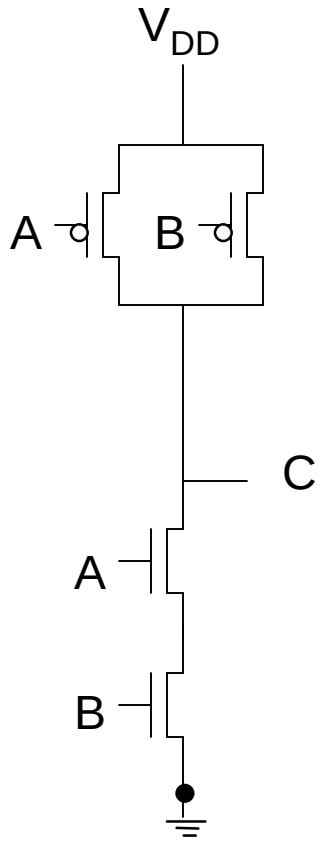
TODAY:

- CMOS Logic Gates: NAND, NOR
- Delay in Logic Gates
- More on Capacitance -
  - Diffusion
  - Interconnect
- Interconnect Scaling

CMOS DIGITAL LOGIC

NAND gate

A	B	A B	C= $\overline{A\,B}$
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0



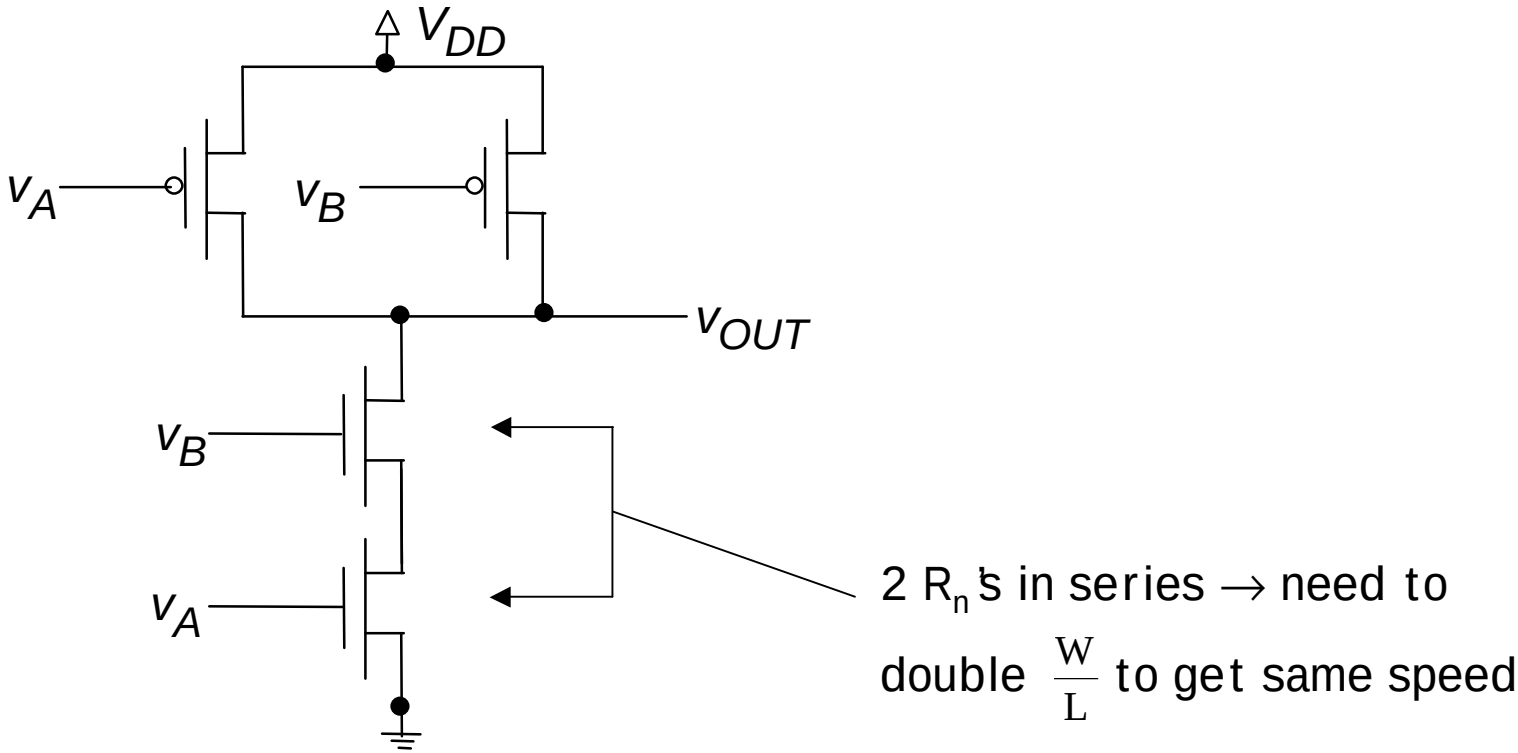
Making a NAND gate:

NMOS portion: both switches need to be closed for output to be low → series

PMOS portion: either switch can be closed for output to be high → parallel

CMOS NAND GATE

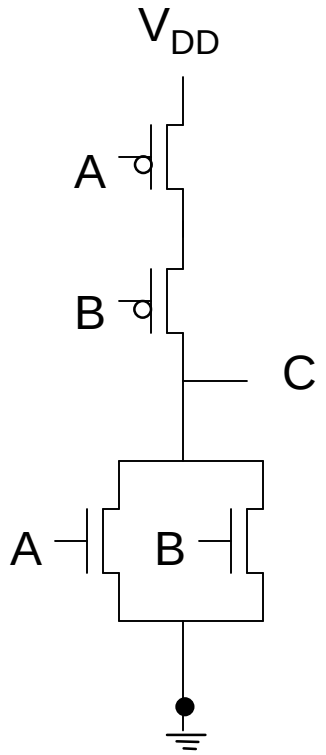
NMOS switches in series from output to ground; PMOS switches in parallel from output to the supply



CMOS NOR GATE

NOR function (two inputs)

A	B	A +B	$C=\overline{A+B}$
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0



Output is high only if both inputs are low →

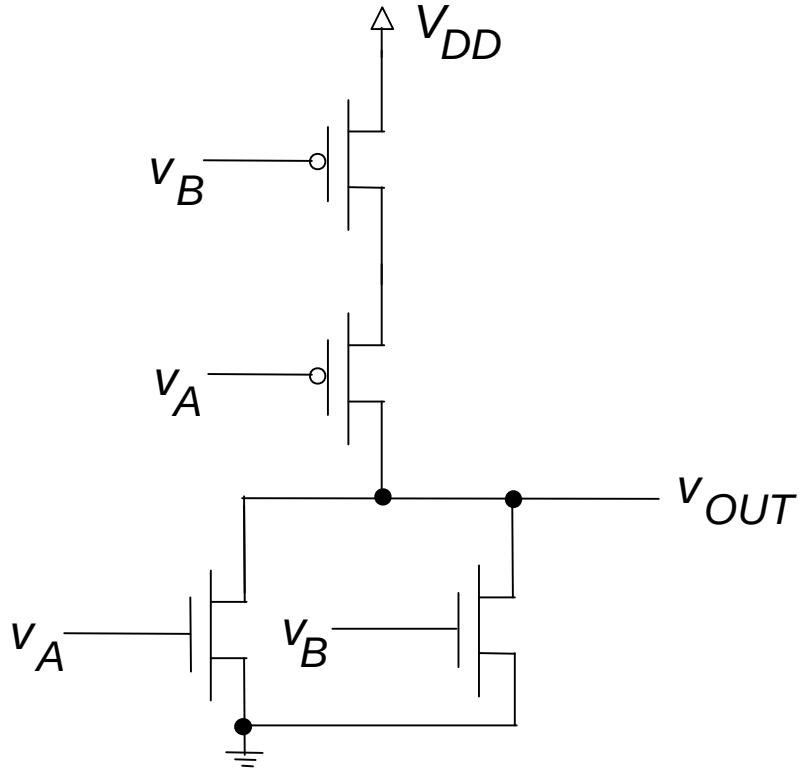
PMOS switches (between the supply and the output) in series

Output is low if either input is high →

NMOS switches (between ground and the output) in parallel

CMOS NOR GATE

"Complementary" configuration to the NAND gate

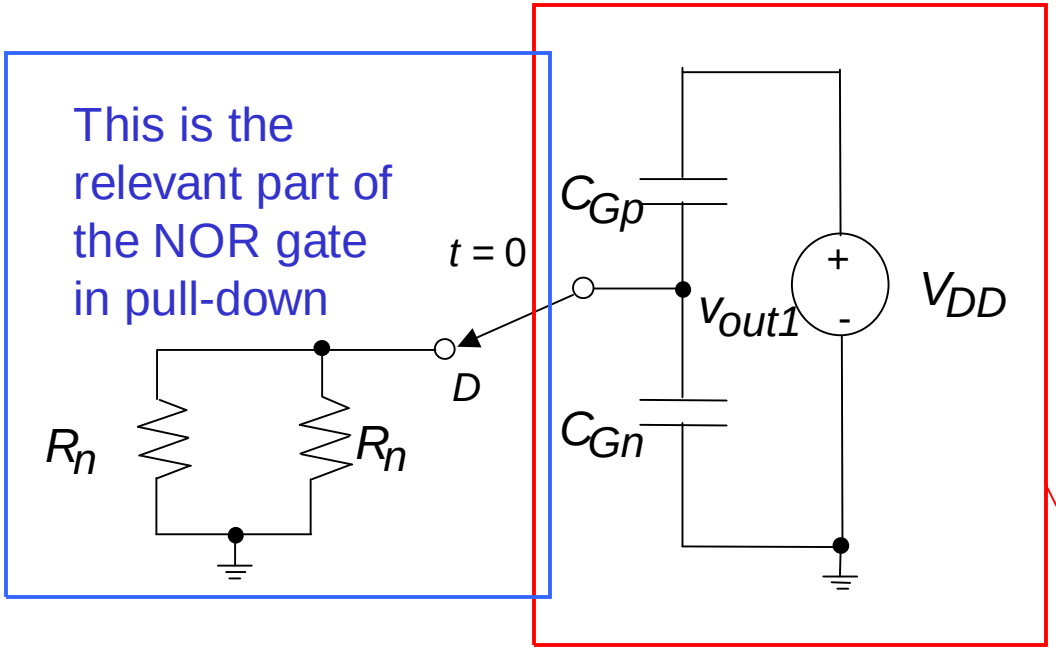


PMOS are in series (they are slower than NMOS) – already larger than NMOS. To double them in $\frac{W}{L}$ to get same speed requires huge PMOS devices (and high capacitance).

NOR Gate Pull-Down Transient Model

Depends on values of A and B inputs

If both A and B are high, then the NMOS resistances are in parallel



If only one of the inputs is high, then only one transistor pulls the output down ($R = R_n$)

This is the load on the NOR gate -- for simplicity we assume a simple inverter.

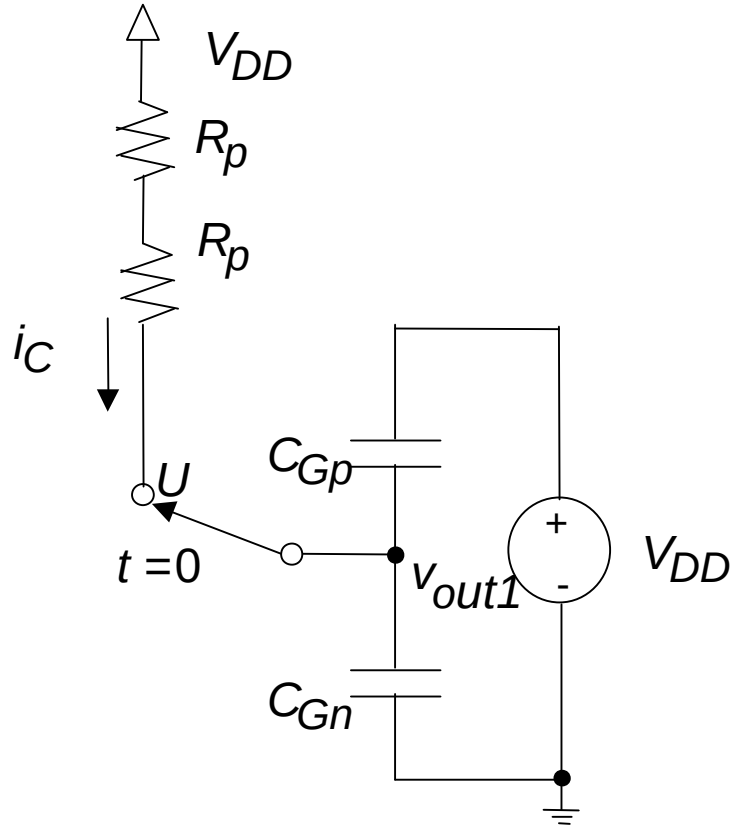
Worst case: $RC = R_n (C_{Gp} + C_{Gn})$

NOR Gate Pull-UP Transient Model

PMOS switches are in series → resistors are in series

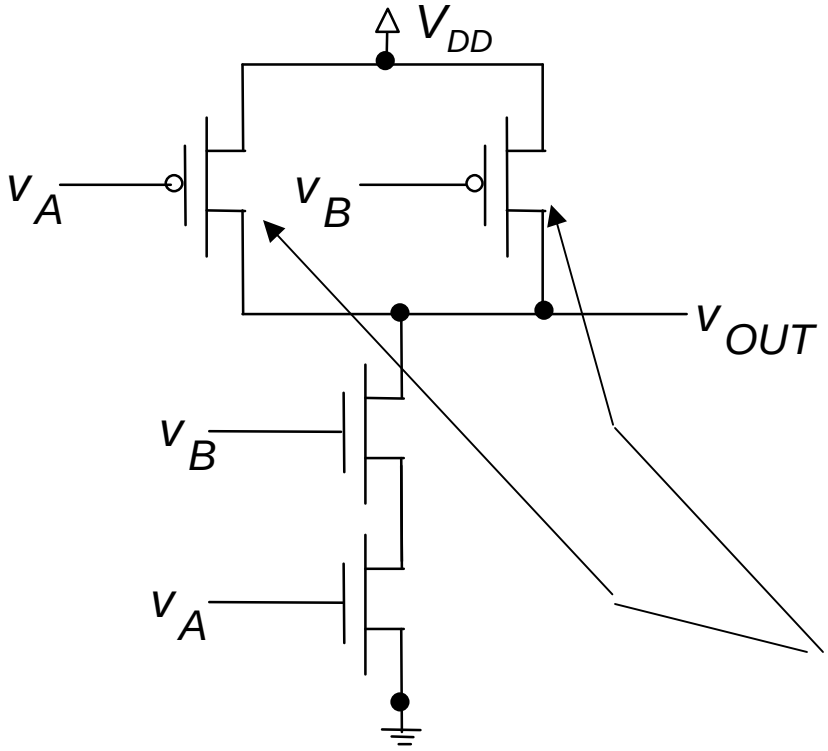
Pull-up - both inputs switch low together

$\tau = (R_p + R_p)(C_{Gn} + C_{Gp})$

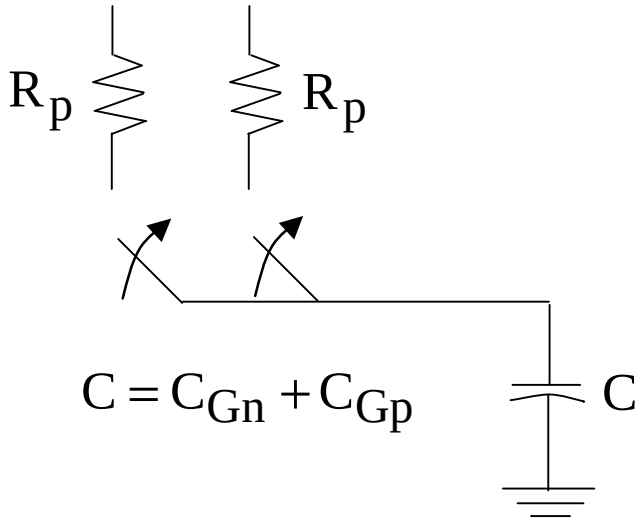


Need to size PMOS so that $R_p = \frac{1}{2} R_n \Rightarrow \left.\frac{W}{L}\right|_p \approx 4 \left.\frac{W}{L}\right|_n$

NAND Gate Pull-Up Model



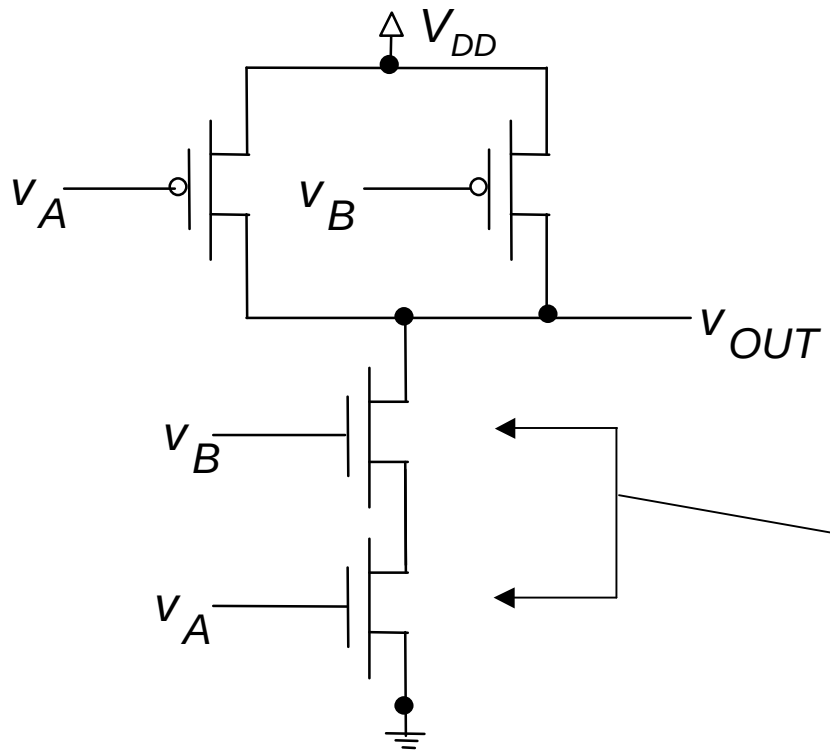
$\Rightarrow$

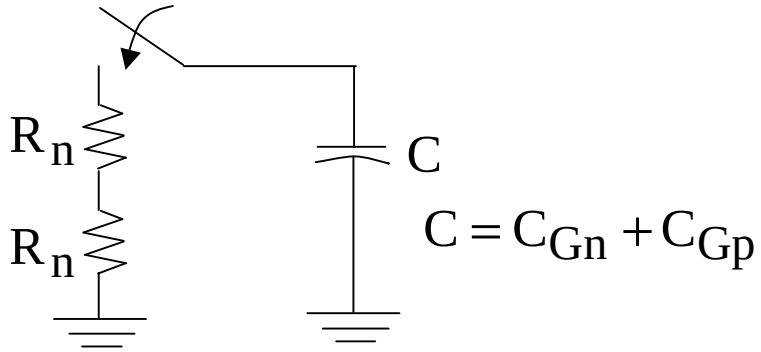


One or both switches closed
(worse case: one switch)

$$\tau = RC = R_pC \quad = R_p(C_{Gn} + C_{Gp})$$

NAND Gate Pull-Down Model





$\Rightarrow$

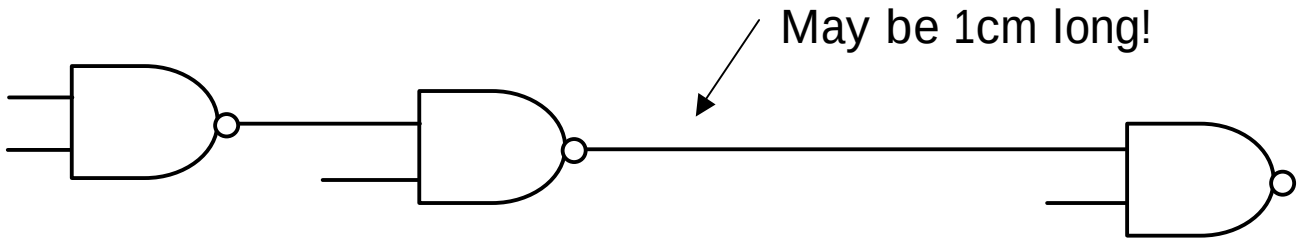
Two R_n's add in series, so RC = $2R_nC = 2R_n (C_{Gn} + C_{Gp})$

Since $2R_n \sim R_p$, this circuit is "automatically" balanced for equal rise and fall times.

Interconnect Models

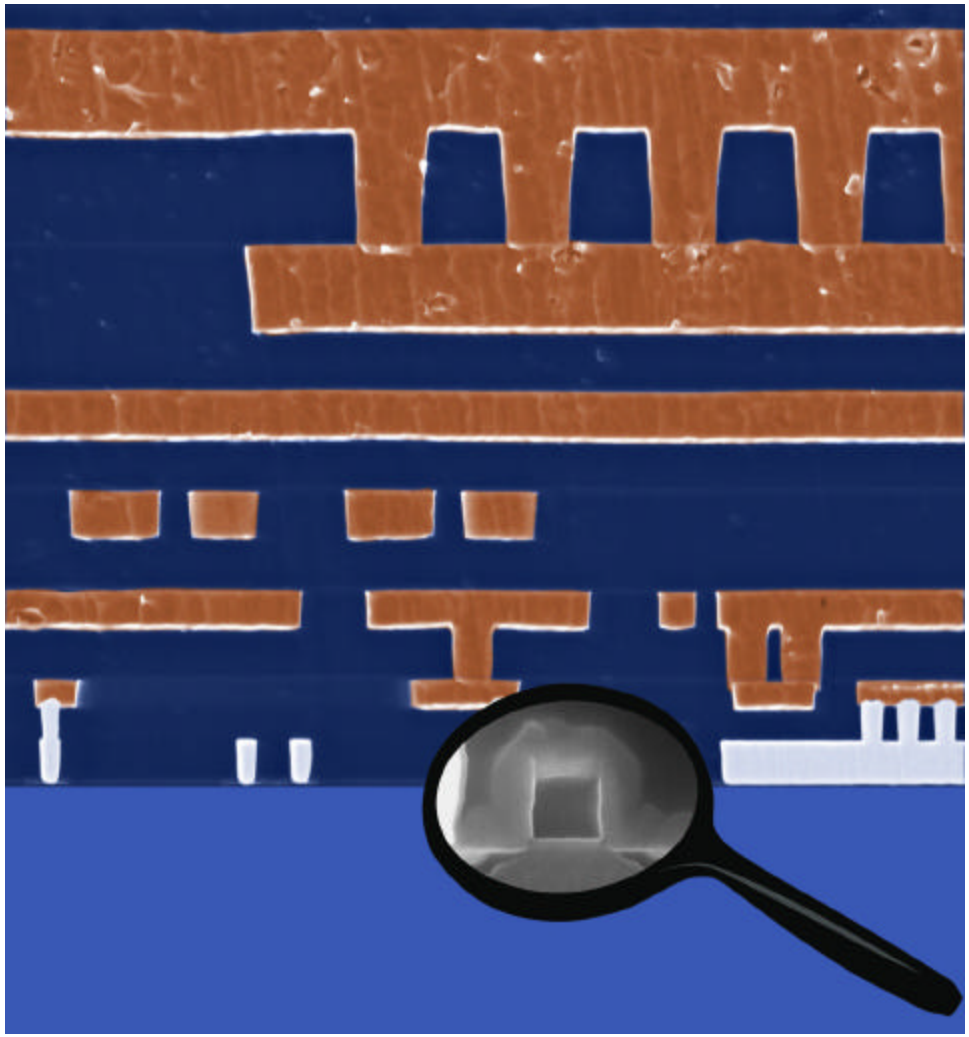
Interconnect = thin-film “wire” connecting gates; typically aluminum

Significant source of capacitance (and resistance if the line is long)



Resistance can be MUCH bigger if the wire happens to be polysilicon (~10 - 20 Ω/sq versus 0.1 for aluminum) .

We can also use the source-drain doped region for interconnect, but it has sheet resistance similar to polysilicon.



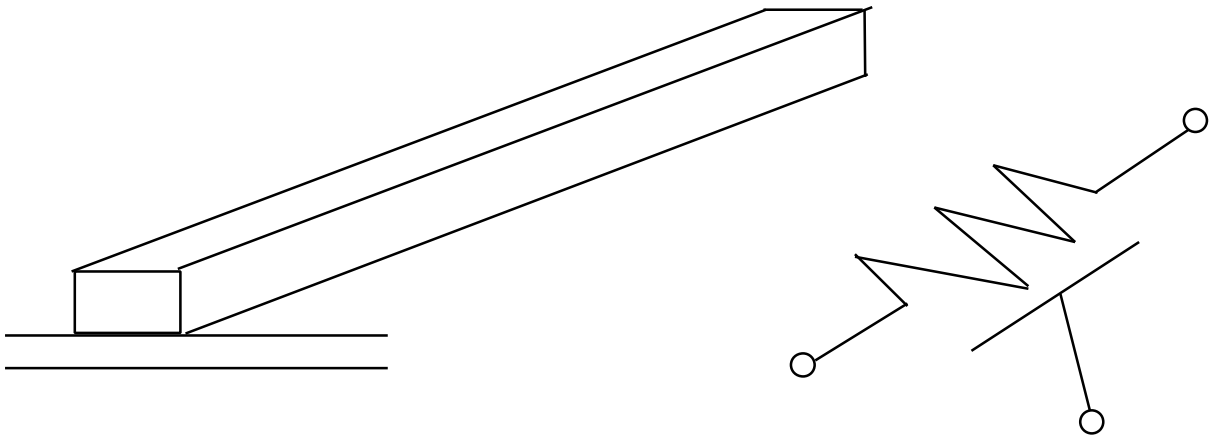
Copper wiring + expanded cross-section of a transistor

Note that we use many levels of wiring.

Interconnect Models

Problem: Capacitance to underlying substrate (or well) is "smeared" along the length of the wire, along with the resistance

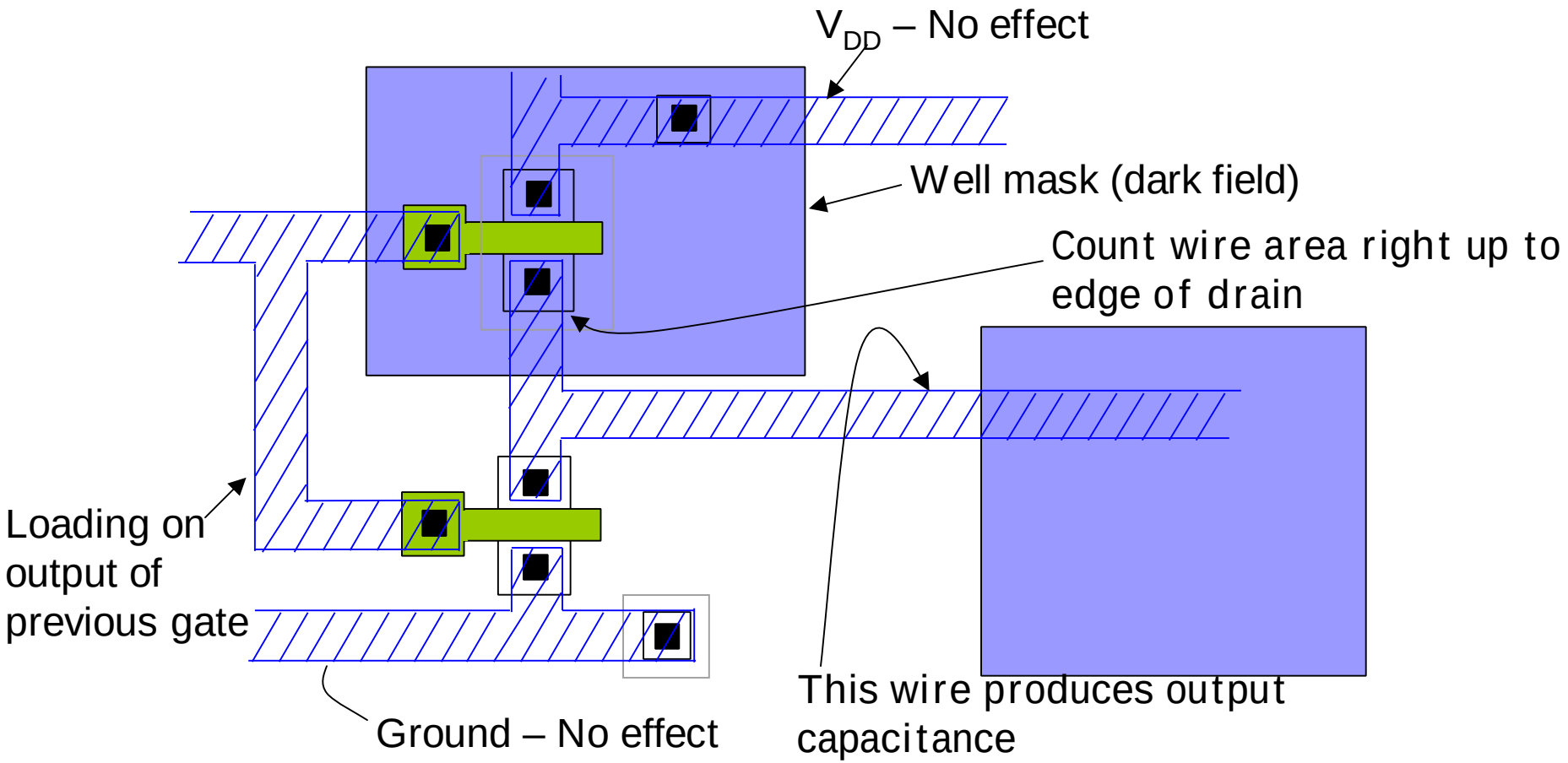
Distributed RC line – We will look at this next time



Let's first look at simple extra capacitance.

Interconnect Layout: Determining Wire Capacitance

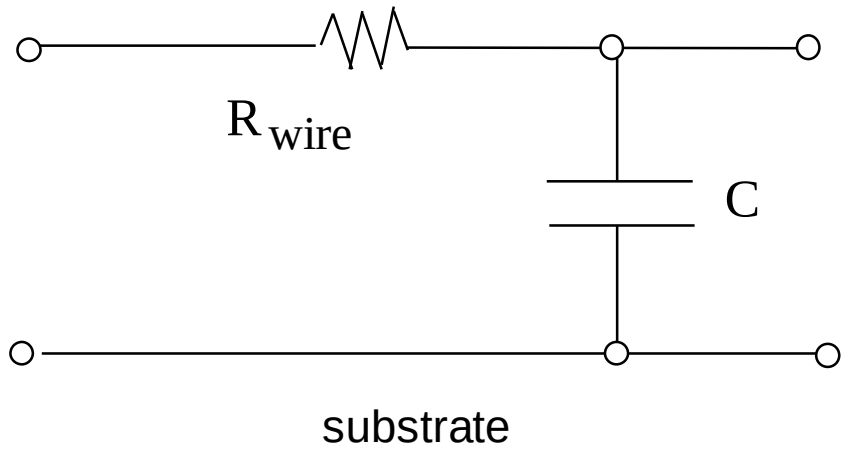
Metal layer overlaps substrate (grounded) and well (connected to V_{DD})



Approximate Models

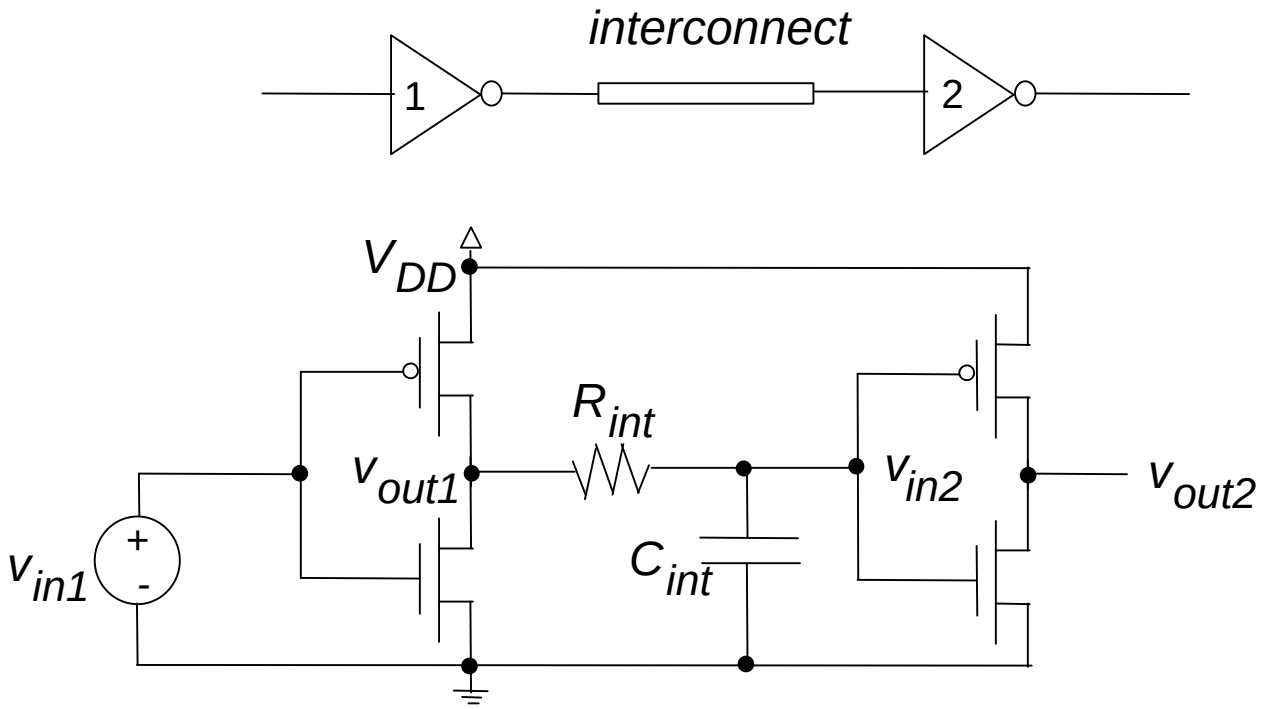
"One-lump" model: put all of capacitance at the end of the interconnect, where it adds to the gate capacitance of the load

The series resistance adds to the MOSFET equivalent resistance



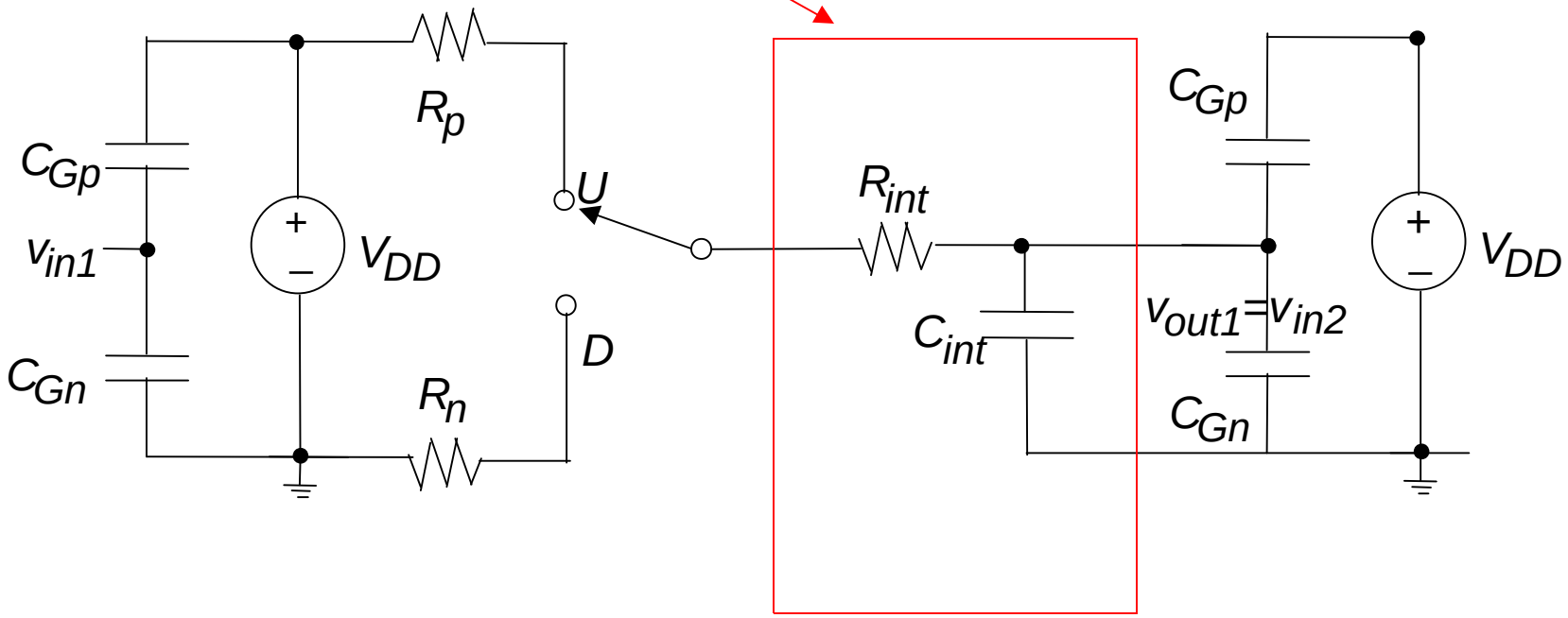
CMOS Inverter Pair with Interconnect

Use "one-lump" R-C approximation to interconnect

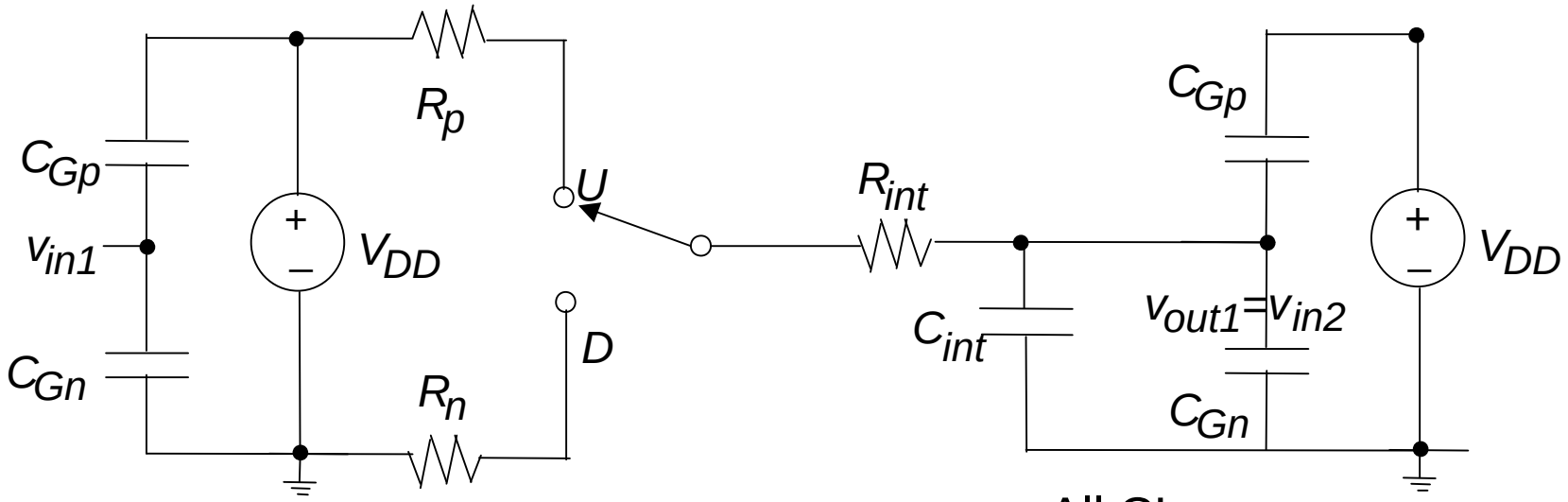


Effect of Interconnect on Inverter Pull-Up Transient

Insert one-lump model for interconnect in between the output of inverter 1 and the input gate capacitances of inverter 2



Effect of Interconnect on Inverter Pull-Up Transient



All C's are "virtually" parallel

Assessing the effect of interconnect:

$$RC=(R_p+R_{int})(C_{Gn}+C_{Gp}+C_{int})$$

Can also be significant, especially in poly

Biggest effect

Interconnect Parameters

Typical wire capacitance:

$$C_w = \frac{\varepsilon_{ox}}{t_{thox}}\left(W_w L_w\right) = \left(\frac{\varepsilon_{ox} W_w}{t_{thox}}\right) L_w$$

thick oxide = deposited oxide … about 500 nm, wire width = 1 μm (for lowest metal level)

$$C_w = \frac{\varepsilon_{ox}}{t_{thox}}\left(W_w L_w\right) = \frac{(3.45\text{x}10^{-13}\,\text{F/cm})(10^{-4}\,\text{cm})}{5\text{x}10^{-5}\,\text{cm}} L_w$$

$$= (70\text{aF}/\mu\text{m}) L_w$$

Long interconnects are needed before C_w becomes significant compared with the gate capacitance –

Example: L_w = 15 μm → C_w = 1.05 fF (not that big compared with C_G)

If L_w = 15 mm, C_w = 1.05 pF (very big)

Interconnect Resistance

Typical MOSFET "on" resistances R_n and R_p: 1 K for small gates (w ~ few μm), but R_n, R_p much less for large "line drivers"

Compare with aluminum:

$R_\square = \rho / t = (2.7\ \mu\Omega\text{-cm}) / (1\ \mu m) = 2.7 \times 10^{-2}\ \Omega/\square$

Example: $L_w = 1000\ \mu m \rightarrow R = R_\square (L_w / W_w) = (2.7 \times 10^{-2}\ \Omega/\square)(1000/1)$

$R = 27\ \Omega$, fairly small

But for polysilicon, if $R_\square \sim 20\ \Omega/\square$

$\Rightarrow$ Keep poly lines short!

Interconnect Scaling

Take S to be the scaling factor - Dimensions shrink as S

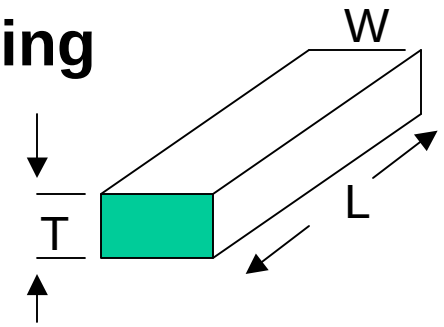
Every 2-3 years, a new technology is introduced with minimum feature size reduced by S (~ 1/1.4)

How do relevant interconnect features shrink?

Relevant parameters: width W, spacing Sp , thickness T, inter-level dielectric (ILD) thickness H, line length L, resistivity ρ, dielectric constant ε

(Also Voltage)

Interconnect Scaling



Ideally L, W,T,Sp,H (ILD) all scale down by S

Length scales by S for "local" wires, and is

equal or slightly *longer* for global wires

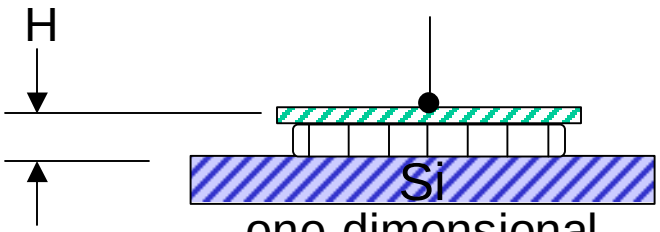
Resistivity and dielectric constant are reduced somewhat with new materials (Cu vs. Al, polyamides vs. SiO_2) (at least we *try* to reduce)

Net Results: Effect of dimensional scaling on **Local wires:**

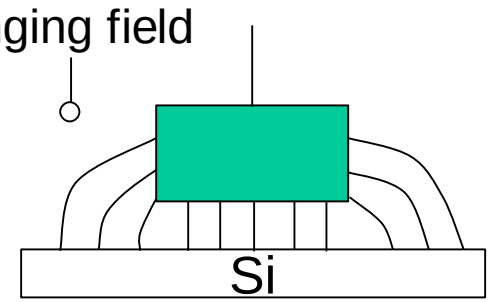
Resistance: $\mathbf{R=\rho L/A_w}$ Capacitance to substrate: $\mathbf{C= \epsilon A_C/H}$

$$RC = \rho L/WT * (\varepsilon WL/H) \quad = \rho\varepsilon L^2 / TH \sim \rho\varepsilon S^2 / S^2 \sim \rho\varepsilon$$

So, we expect local wires to become faster as new materials are introduced with lower ρ and ε .



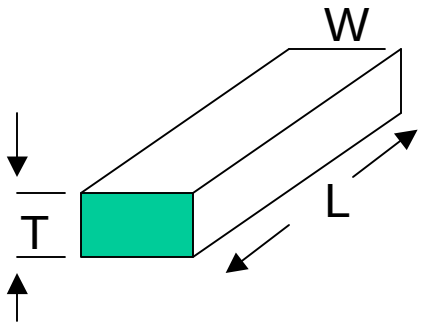
one-dimensional capacitor



real capacitor

(fringing fields will degrade this, however)

Interconnect Scaling



Global wires: $RC = \rho\varepsilon L^2 / TH$ but now L doesn't scale as S (chip size increasing)

If L is constant then RC scales as $\rho\varepsilon / S^2$ That's bad.

Assume L rises as $S^{-0.5}$ (20% longer / generation) Then RC scales as: $\rho\varepsilon / S^3$

If S = 0.7 then RC delay of a global wire increases nearly 3x per generation!

The real interconnect bottleneck is in the global interconnects

How to handle?

The way to combat rising global RC delays is to maintain a larger cross-sectional area -- Not possible with one layer of wiring.

If we DON'T scale global wires down in size, WT product stays relatively flat. The only way to do this is run them in separate planes of wiring where there is more room.

Other approaches/ 2nd Order effects

Fringing field capacitance adds to wire capacitance beyond parallel-plate model AND

In modern processes, ~80% of capacitance is to neighboring wires, not underlying ground planes

This means capacitance doesn't scale down with W, making things worse than we projected earlier

Can model the coupling capacitance as a sideways parallel plate

Other approaches/ 2nd Order effects (cont.)

In order to reduce resistance without losing area, we can make the wires taller (T scales more slowly than S)

Cross-sectional area only reduces as S then, not S^2

Problems: Hard to manufacture tall and thin wires (define aspect ratio = T / W) High aspect ratios are tough to make.

Also, more capacitance to neighboring wires -- *noise*! (We need to briefly treat this next time.)

Other approaches/ 2nd Order effects (cont.)

Adding more metal layers

10 years ago, we had 1 or 2 metal layers in a process

Today we have 6

Why? We use fat wires for long runs and thus do not get big resistance. Moreover, we reduce the average wire length because we don't need to go around obstacles so much.

Also, chip area is reduced since wires can be put on top of each other

Why not 10 or 15 levels? Cost – each added layer adds to # of masks and reduces yield, but that's definitely the direction we are going.

What about "diffusion capacitance"

(ie drain to substrate pn-junction capacitance.)

That's easy to estimate: Just multiply drain area A_D by the diffusion capacitance per unit area C_{JA}

Typical values of C_{JA} = 0.5fF/μm^2 or 5 x 10^{-8} F/cm2

$C_{JA} = \varepsilon_{Si}\, \varepsilon_o$ /d where ε_{Si} is 12 and d is the junction depletion layer thickness, about 100-400nm.