EN1600
Design and Implementation of VLSI Systems
Fall 2016

Lecture 8, 9: Sizing and Layout of Complex CMOS Gates

Reading: Chapter 4, sections 4.3-4.5
Chapter 1, section 1.5.5
Weste & Harris

October 3 & 5, 2016
Prof. R. Iris Bahar

BROWN

Homework #2

- Available on course webpage
- After today's lecture you should be able to do all the problems.
- Prof. Bahar will be holding office hours Tuesday from 10-noon to make up for missed office hours today
- Note that HW#2 is due this Friday by 5pm

Today's lecture

- Review of last lecture
  - Sizing the inverters in a chain
    - Optimal scaling factor
    - Optimal number of inverters
  - Sizing NAND/NOR gates
    - Based on topology of gate
    - balance rise/fall delay
- New material for today
  - Topology and transistor sizing of complex gates
  - Optimal layout configuration or complex gates

Impact of fanout on delay

$$t_p = t_{p0}\,(1 + C_{ext}/C_{int})$$

- Extrinsic capacitance, C_{ext} , is a function of the gates being driven by the gate under question (i.e. the fanout)

  larger fanout ➡ larger external load.

- Re-express the intrinsic capacitance (C_{int}) in terms of input gate capacitance:

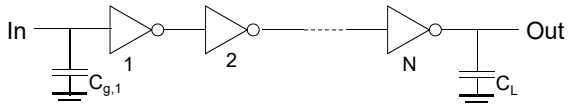
$$C_{int} = \gamma C_g \,, \quad \text{where } \gamma \approx 1$$

$$t_p = t_{p0}\,(1 + C_{ext}/\gamma C_g) = t_{p0}\,(1 + f/\gamma)$$

$f = C_{ext}/C_g$ is the effective fanout

Inverter chain

- Goal is to minimize the delay through an inverter chain



- The delay of the j^{th} inverter stage is

$$t_{p,j} = t_{p0}\,(1 + C_{g,j+1}/(\gamma C_{g,j})) = t_{p0}(1 + f_j/\gamma)$$

and $t_p = t_{p,1} + t_{p,2} + \ldots + t_{p,N}$

so $t_p = \sum t_{p,j} = t_{p0} \sum (1 + C_{g,j+1}/(\gamma C_{g,j}))$

- If C_L and $C_{g,1}$ are given, we have 2 different optimizations
  - *How should the inverters be sized to minimize delay?*
  - *How many stages are needed to minimize the delay?*

Sizing the inverters in the chain

- After a bit of calculus, we find that for minimum delay:

$$C_{g,j+1}/C_{g,j} = C_{g,j}/C_{g,j-1} \quad \text{for } j=2\ldots N$$

- What does this imply?
  - All gates have the same effective fanout, *f*
  - Each gate should be scaled up by the same factor w.r.t. its preceding gate
- What is the effective fanout for a gate given C_L and $C_{g,1}$?
  - With a bit of algebra and inductive reasoning we find that:

$$f = \sqrt[N]{C_L / C_{g,1}} = \sqrt[N]{F}$$

  - $F = C_L/C_{g,0}$ is the overall effective fanout
- What is the minimum delay through the chain?

$$t_p = N t_{p0}(1 + \sqrt[N]{F}/\gamma)$$

Optimal number of inverters

- What is the optimal value for *N* given F? (where $F = f^N$)
  - if the number of stages is too large, the intrinsic delay of the stages dominates
  - if the number of stages is too small, the effective fan-out of each stage dominates

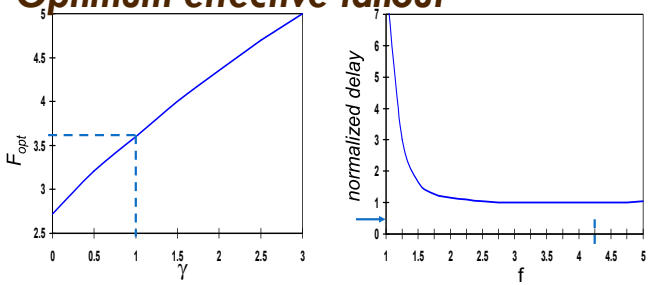
$$t_p = N t_{p0}(1 + \sqrt[N]{F}/\gamma)$$

$$\partial t_p / \partial N = \gamma + \sqrt[N]{F} - \frac{\sqrt[N]{F}\ln F}{N} = 0$$

$$\rightarrow f = e^{(1+\gamma/f)}$$

- For $\gamma = 0$ (ignoring self-loading) $N = \ln(F)$ and the effective-fan out (tapering factor) becomes $f = e = 2.718$
- For $\gamma = 1$ (the typical case) the optimum effective fan-out can be solved numerically and turns out to be close to 3.6

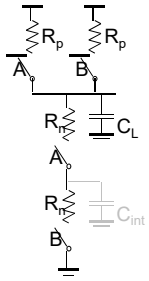
Optimum effective fanout



- Too many stages has a substantial negative impact on delay
- Choosing f slightly larger than optimum has little effect on delay and reduces the number of stages (and area).
  - Common practice to use f = 4 (for $\gamma = 1$)

➡ Fanout of 4 (FO4) rule of thumb delay metric is based on this result

Input pattern effects on delay

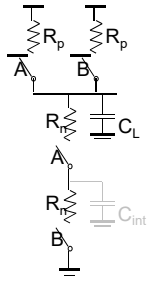


- Delay is dependent on the pattern of inputs
- 1st order approximation of delay:

  $t_p \approx 0.69\ R_{eff}\ C_L$

- R_{eff} depends on the input pattern

Input pattern effects on delay



- 0→1 transition on output: 2 possibilities
  - one input goes low: what is R_{eff} ?
    - delay is 0.69 R_p C_L
  - both inputs go low: what is R_{eff} ?
    - delay is 0.69 $R_p/2$ C_L since two p-resistors are on in parallel
- 1→0 transition on output: 1 possibility
  - both inputs go high
    - delay is 0.69 $2R_n$ C_L

$t_p \approx 0.69\ R_{eff}\ C_L$

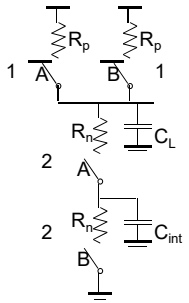
Adding transistors in series (without sizing) slows down the circuit

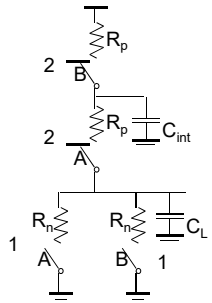
The rest of today's lecture

- *How do we develop design rules for sizing CMOS gates in general?*
  - The 2:1 ratio for an inverter doesn't necessary work best for other types of gates
- *How should we go about planning the layout of these more complex CMOS gates?*
  - Gates with multiple inputs means more complex routing
  - How to you order inputs and draw out the active area to minimize total area?

Transistor sizing

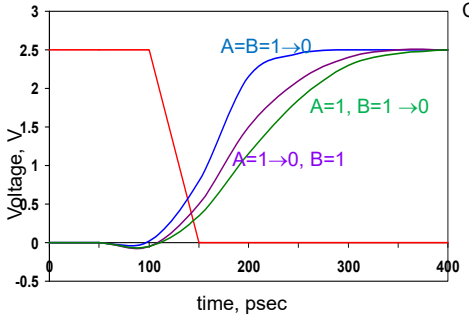
- How should NMOS and PMOS devices be sized relative to an inverter with equal rise/fall times?





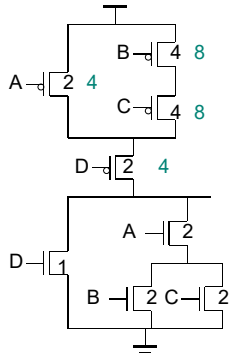
Delay dependence on input patterns

2-input NAND with
NMOS = 0.5µm/0.25 µm
PMOS = 0.75µm/0.25 µm
C_L = 10 fF



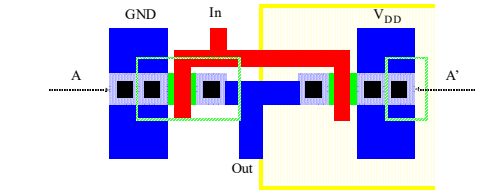
Input Data Pattern	Delay (psec)
A=B=0→1	69
A=1, B=0→1	62
A= 0→1, B=1	50
A=B=1→0	35
A=1, B=1→0	76
A= 1→0, B=1	57

Transistor sizing a complex gate

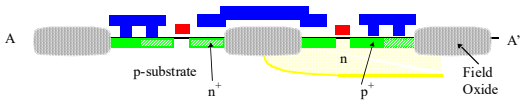


OUT = !(D + A • (B + C))

CMOS inverter layout



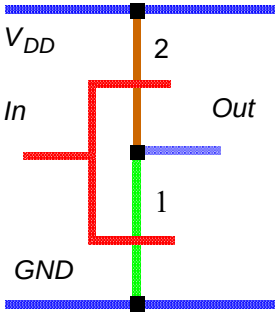
(a) Layout



(b) Cross-Section along A-A'

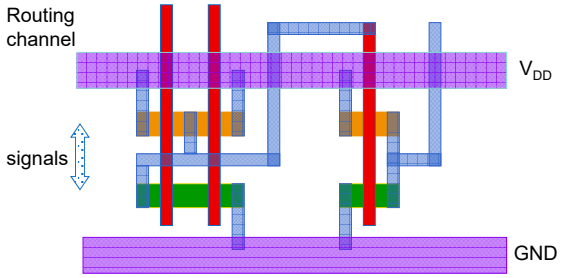
Symbolic layout

- Dimensionless
- Positioning is most important



Stick diagram of inverter

Standard Cell Layout Methodology

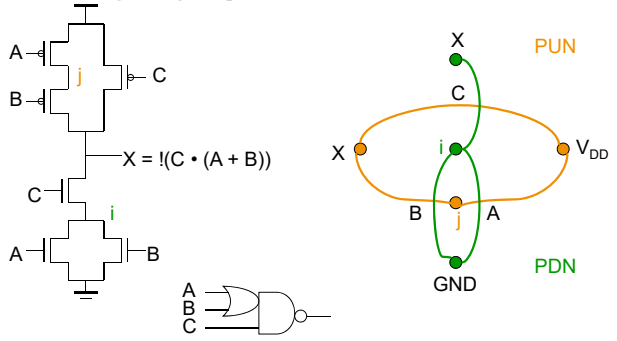


What logic function is this?

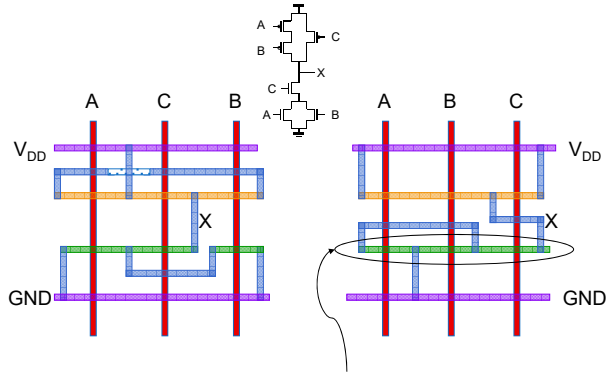
Layout planning for complex gates

- Want layout to be as dense (area efficient) as possible.
- Try to realize all NMOS and PMOS transistors in unbroken row of devices
  - Requires only single strip of diffusion in both wells
- Careful ordering of inputs is important to help achieve this
- Use a systematic approach identifying Euler paths

OAI21 logic graph



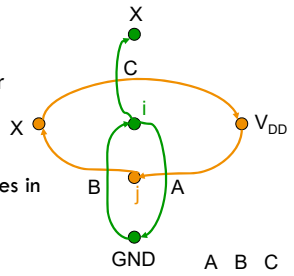
2 stick layouts of !(C • (A + B))



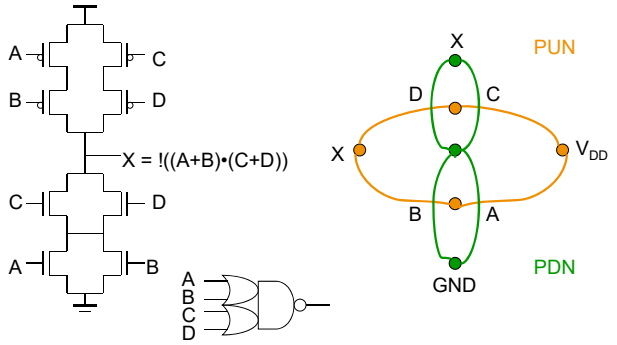
uninterrupted diffusion strip

Consistent Euler path

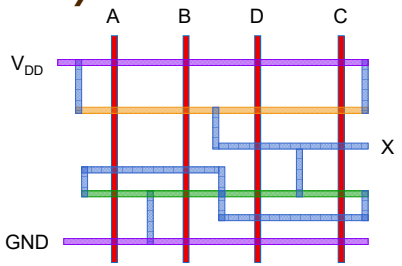
- An uninterrupted diffusion strip is possible only if there exists a Euler path in the logic graph
  - Euler path: a path through all nodes in the graph such that each edge is visited once and only once.
  - For a single poly strip for every input signal, the Euler paths in the PUN and PDN must be consistent (the same)



OAI22 logic graph



OAI22 Layout



- X = !((A+B)•(C+D))
- Chosen Euler path: ABDC

Draw the Euler paths for this function

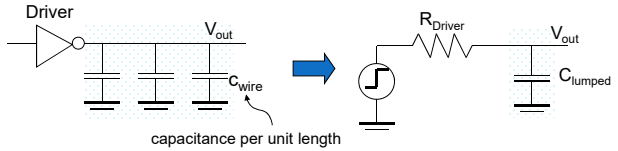
- x = !(a + bc + de)
- x = !(bc + a + de)
- Some functions have no consistent Euler path!

Homework #2

- Should now have all the background to do all the problems for this homework
- Due Friday, by 5pm
  - Hardcopy due to me or Marc
  - /gpfs/data/engn1600 directory needs to contain all relevant files
- Let me know if you have any issues meeting this deadline

Wire delay models

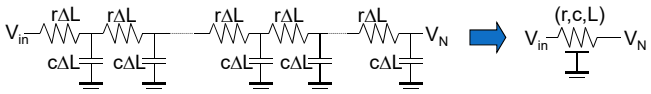
- Ideal wire
  - same voltage is present at every segment of the wire at every point in time - at equi-potential
  - only holds for *very short* wires, i.e., interconnects between *very* nearest neighbor gates
- Lumped C model
  - when only a single parasitic component (C, R, or L) is dominant the different fractions are lumped into a single circuit element
    - When the resistive component is small and the switching frequency is low to medium, can consider only C; the wire itself does not introduce any delay; the only impact on performance comes from wire capacitance



  - good for short wires; pessimistic and inaccurate for long wires

Wire Delay Models, con't

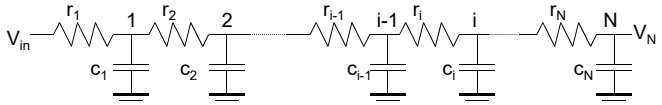
- Lumped RC model
  - total wire resistance is lumped into a single R and total capacitance into a single C
  - good for short wires; pessimistic and inaccurate for long wires
- Distributed RC model
  - circuit parasitics are distributed along the length, L, of the wire
    - c and r are the capacitance and resistance per unit length



- Delay is determined using the Elmore delay equation:

$$\tau_{Di} = \sum_{k=1}^{N} c_k r_{ik}$$

Chain network Elmore delay

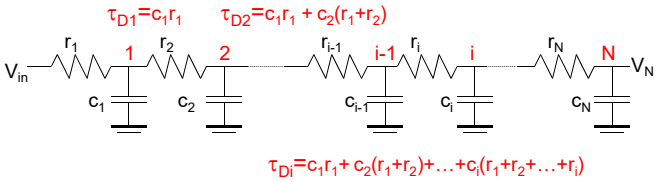


- A typical wire is a chain network with (simplified) Elmore delay of

$$\tau_{DN} = \sum c_i r_{ii} = \sum_{}^{N} c_i \sum_{}^{i} r_j$$

- Where $\sum_{}^{i} r_j = r_1 + r_2 + \ldots + r_i$

Chain Network Elmore Delay

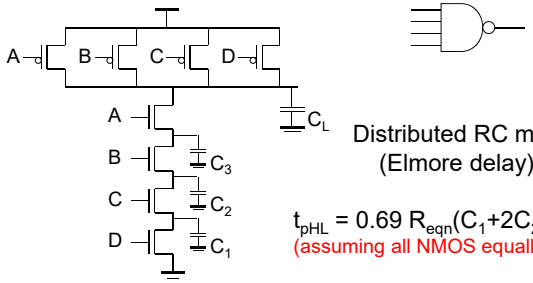


Elmore delay equation $\tau_{DN} = \sum c_i r_{ii} = \sum_{}^{N} c_i \sum^{i} r_j$

If all resistors are equal size,

$$\tau_{Di} = c_1 r_{eq} + 2c_2 r_{eq} + 3c_3 r_{eq} + \ldots + ic_i r_{eq}$$

Fanin considerations

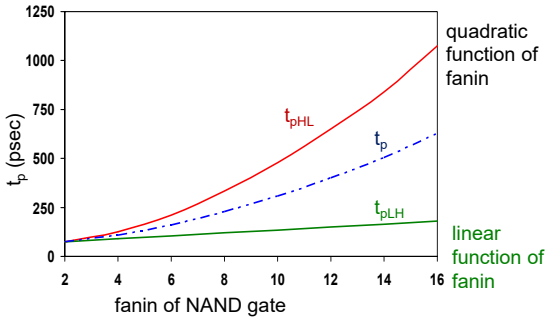


Distributed RC model (Elmore delay)

$t_{pHL} = 0.69\, R_{eqn}(C_1+2C_2+3C_3+4C_L)$
(assuming all NMOS equally sized)

Propagation delay deteriorates rapidly as a function of fanin: quadratically in the worst case.

t_p as a function of fanin



- Gates with a fan-in greater than 4 should be avoided.