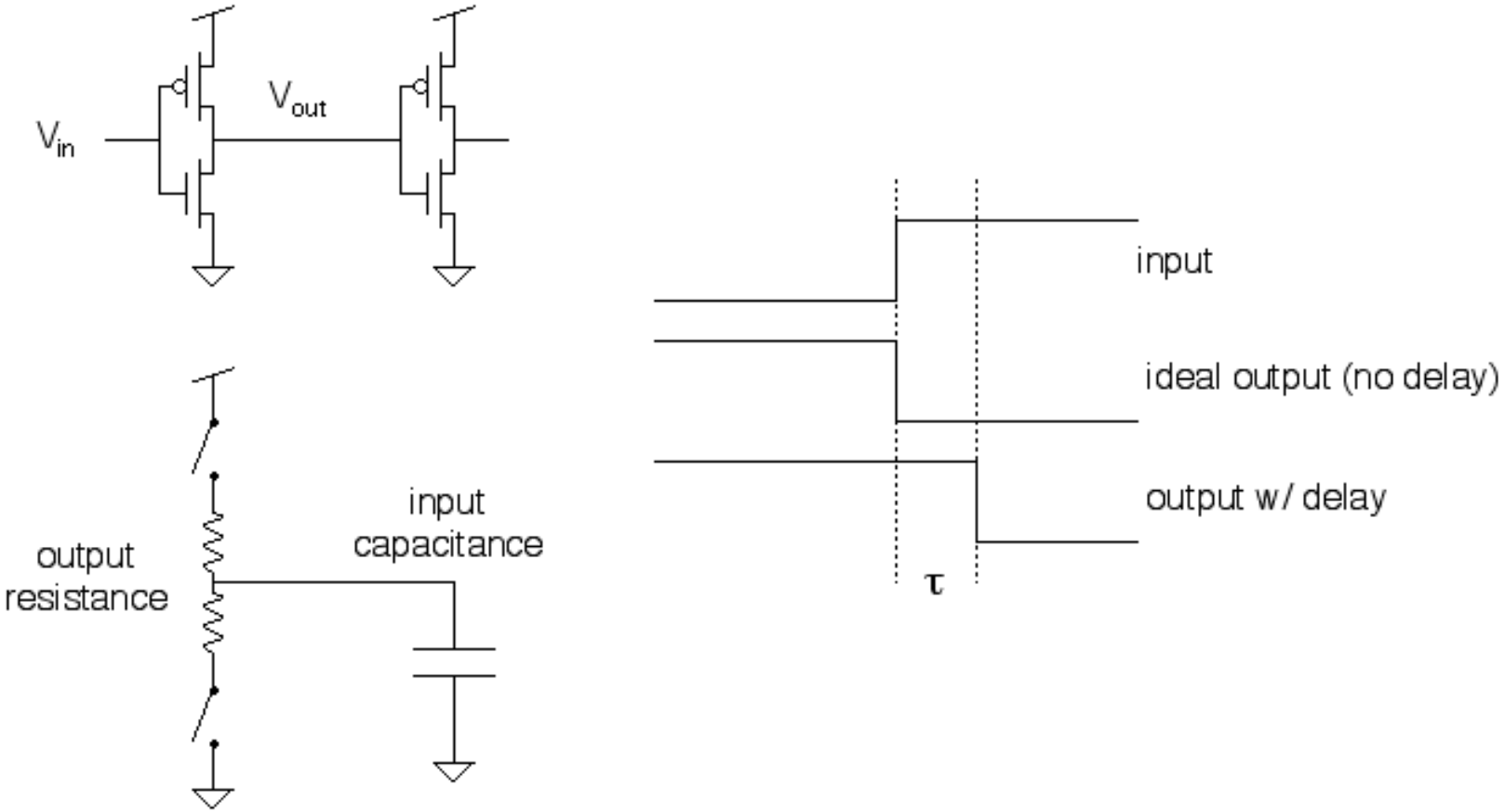
ECE 3060
VLSI and Advanced Digital Design

Lecture 6

Gate Delay and Logical Effort

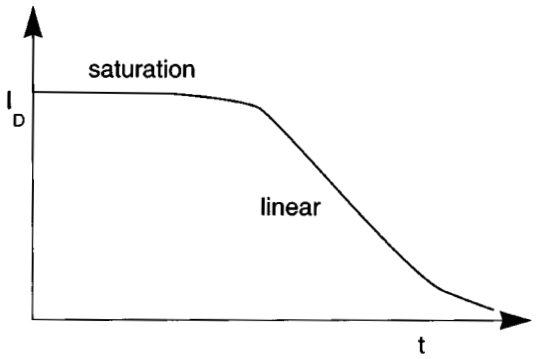
First Model of Gate Delay

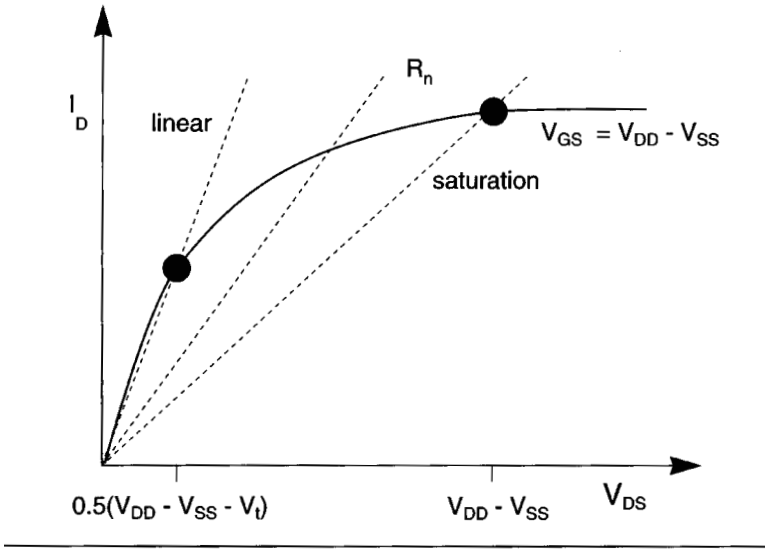


- **This model will be refined shortly**

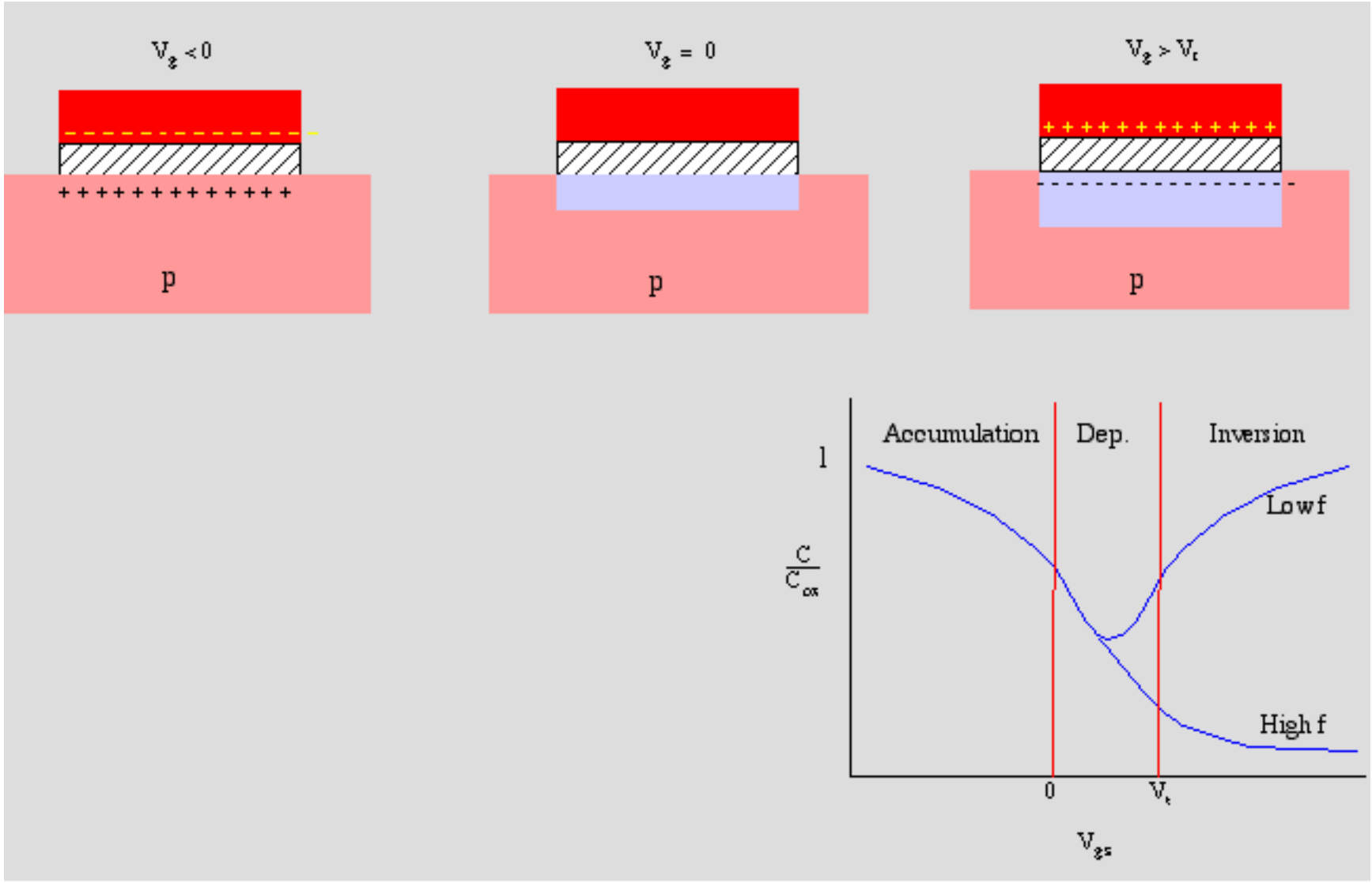
Equivalent R

- **The average resistance of a MOSFET is someplace between the linear region, and saturation**





MOS Capacitor



Capacitance Equations

- **Capacitors store charge**
  - **Q = CV charge is proportional to the voltage on a node**
- **The equation can be put in a more useful form**

$$i = \frac{dQ}{dt} \Rightarrow i = C\frac{dV}{dt} \Rightarrow C\frac{\Delta V}{i} = \Delta t$$

  - **i = dQ/dt => i = C*(dV/dt) => (C*dV)/i = t**
- **Thus, to change the node's voltage (e.g., from 0 to 1), the transistor or gate driving that node must charge (up in our example) the capacitance associated with that node. The larger the capacitance, the large the required charge, and the longer it will take to switch the node.**
- **Since** i **of a transistor is approximately** V/R_{trans}

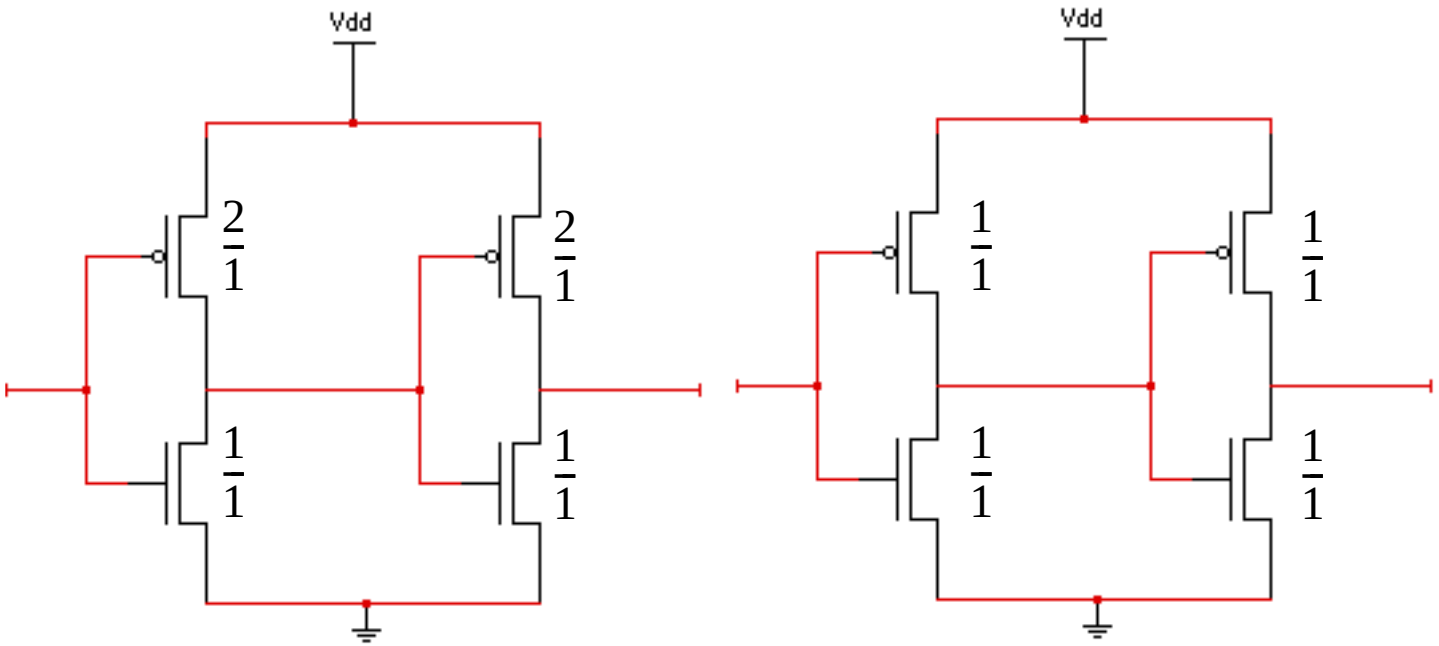
$$\Delta t = \frac{C\Delta V}{i} = \frac{C\Delta V}{\left(\frac{V}{R_{\text{trans}}}\right)} = R_{\text{trans}} C$$

Calculating R and C

- pFET vs. nFET
  - Mobility (μ) of electrons twice mobility of holes
  - pFET resistance is twice nFET resistance
- Series and Parallel Configurations
  - Series resistances add
  - Parallel: worst case is one transistor on
- Transistor Sizing
  - Resistance Inversely Proportional to W / L
  - Gate Capacitance Proportional to W x L

Symmetric Rise/fall?

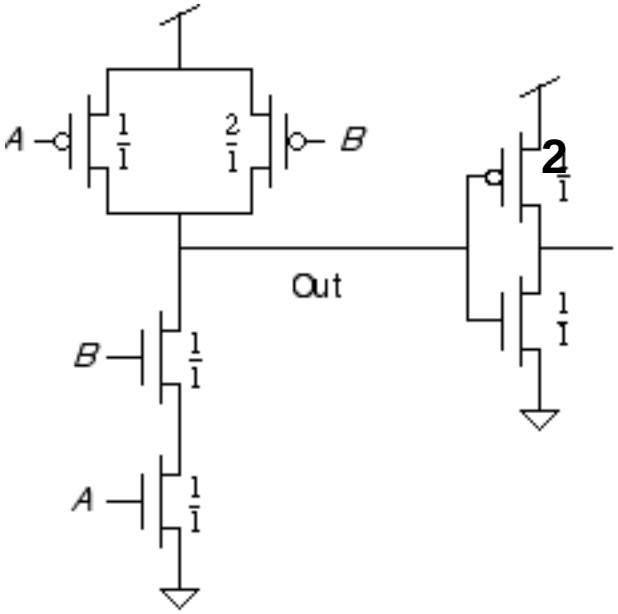
- **Make $\beta_n = \beta_p$, but then gate capacitance increased**

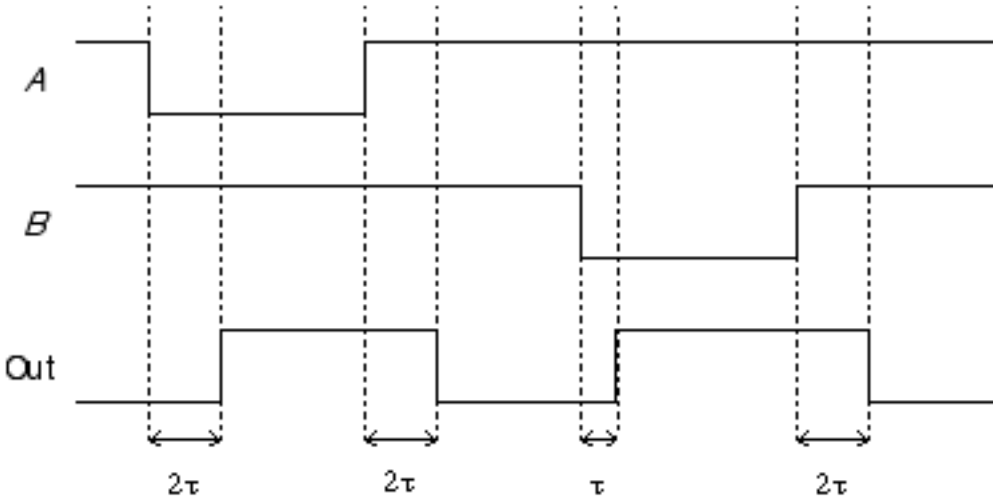


Tau Metric

- **We can normalize delay to technology independent units: for example $\tau = RC$ may be defined by**
  - ***R* is the resistance of a minimum size nFET**
  - ***C* is the gate capacitance of a minimum size inverter with equal rise and fall time: a minimum size nFET plus a double sized pFET**

Delay Example

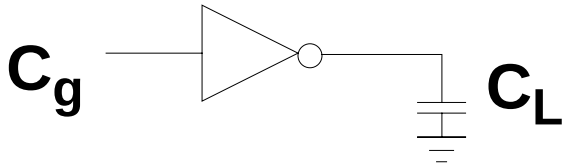




- **What is the delay for different values of A and B?**
  - **AB=01: 2τ due to pFET mobility**
  - **AB=11: 2τ due to series resistance**
  - **AB=10: 1τ due to pFET mobility and W/L=2**

Driving Large Loads

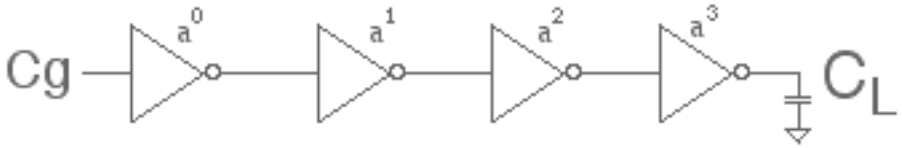
- **Suppose we wish to drive a large load, say equivalent to 100 inverters. What is the delay?**



- **If the delay driving one inverter is δ, then the delay driving C_L is 100δ.**
- **Suppose we scale up the inverter, so the $\frac{W}{L}$ of each FET in the inverter increased by a factor of 100?**
- **The delay driving the load goes down to δ, but the delay driving the inverter itself goes up to 100δ.**

Minimizing Delay

- **Let's consider driving the load with a chain of inverters**



- **Each inverter has $\frac{W}{L}$ ratios a times the previous stage.**
- **Each intermediate stage has delay $a\tau$. If this is true for the last stage, we have $a^n = \frac{C_L}{C_g}$ or $n = \frac{1}{\ln a}\ln\frac{C_L}{C_g}$**
- **Then total delay $\Delta = na\tau = \ln\left(\frac{C_L}{C_G}\right)\frac{a}{\ln a}\tau$**

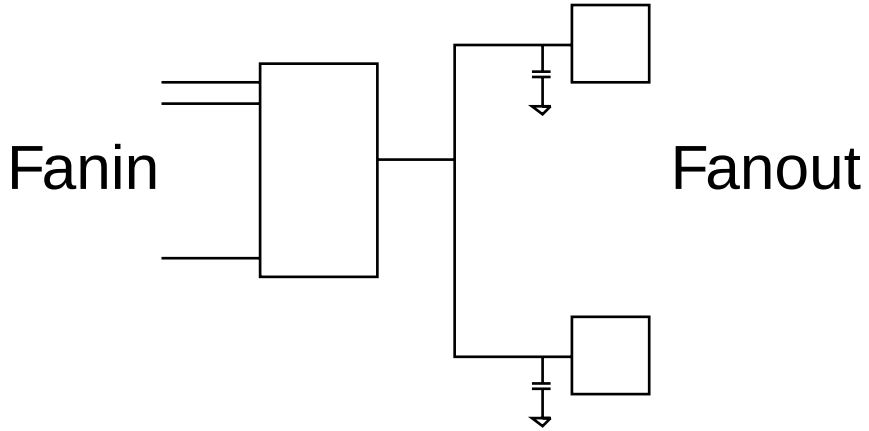
Minimizing Delay (cont)

- **To find the value of a which minimizes Δ, solve $\frac{d\Delta}{da} = 0$ which is minimized for $a \approx e \approx 3$**
- **In practice $2 \leq a \leq 10$, and $a = 4$ is often used.**
- **Example: Using $a = 3$, design a circuit to drive 81 inverters.**

What to do when n is not integral: $n = \left\lceil \frac{1}{\ln a} \ln \frac{C_L}{C_g} \right\rceil$

Fanin and Fanout

- **Fanin is the number of inputs to a complex gate**
  - **High fanin may imply long chains of FETs which will affect rise/fall times**
  - **Best results: chain lengths between two and five**
- **Fanout is the number of gates driven by a gate**
  - **Output rise/fall times are proportional to output capacitance**



- **For the next week, we will investigate the effect of fanin and fanout on gate delay in great detail.**

Goals of Method of Logical Effort

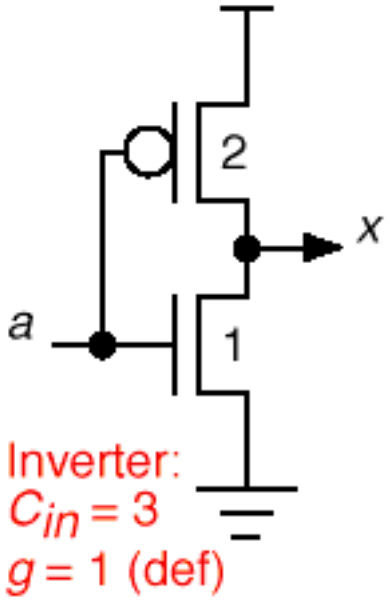
- Learn how to design a combinational logic network with minimum delay
- Learn how to take load into account
  - Example: a decoder output may drive hundreds of inverter equivalent loads
  - How many levels of logic are correct?
  - Which gates to use?
  - So many choices, so little time
- Basic idea:
  - logical effort will describe the gate's contribution to delay
  - electrical effort (fanout) will describe the capacitive load
- Read Chapter 1 of Sutherland, Sproul and Harris

Refining Calculation of Delay

- **As before, we will measure delay in units of $\tau = RC$, so that $d_{\text{abs}} = d\tau$.**
- **As before, we will use $R = R_{\text{inv}}$ and $C = C_{\text{inv}}$, so that τ is the delay of a minimum size inverter (with equal rise and fall times) driving a minimum size inverter.**
  - **Note: $\tau \approx 20\text{ps}$ in a 0.25 micron process**
- **For now we will assume symmetric rise/fall times are required for all of our gates**
- **Observe that so far we have not accounted for output capacitance of the logic gate itself in our delay calculations. That is we have assumed that the delay of a gate with zero fanout is zero. This is about to change.**

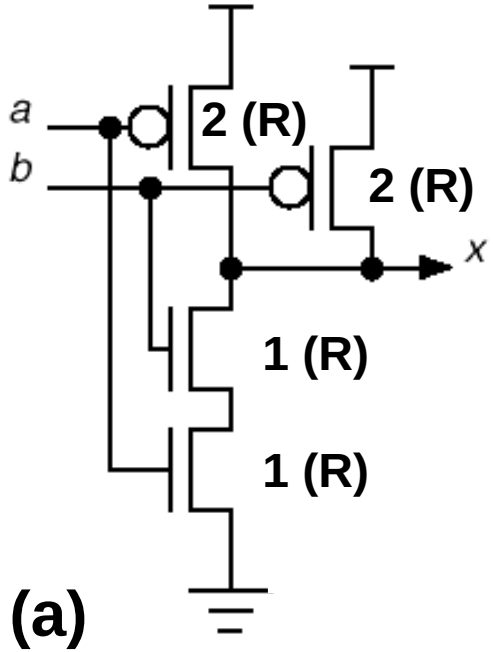
Logical Effort of Inverter

- **Consider the effect of a particular choice of logic gate on (worst case) delay.**
- **The logical effort g of an inverter is defined to be 1.**

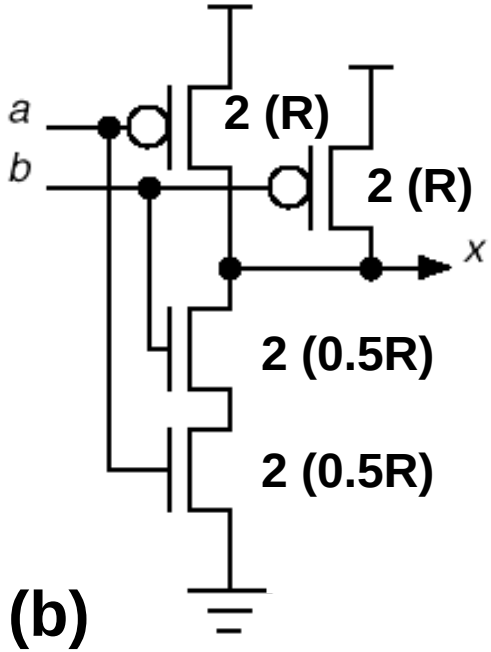


Logical Effort of NAND

- If we size transistors just like the inverter, the delay in the pull-down network will be twice that of pull-up (a)



(a)

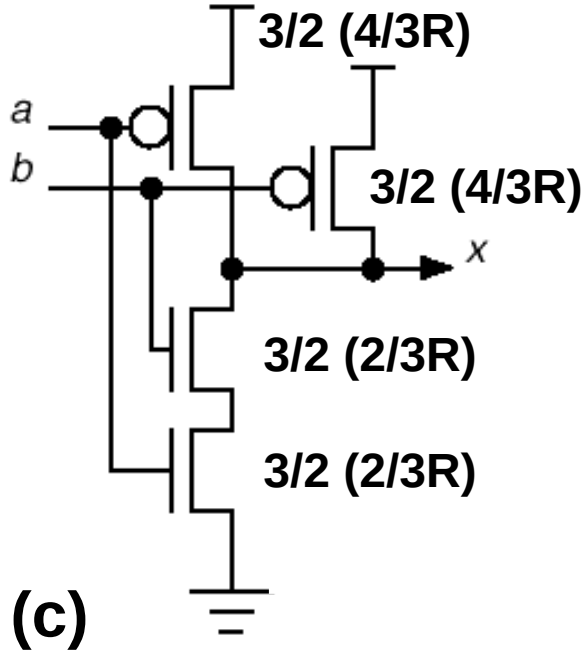


(b)

- If we resize the pulldown nFETs, we can drop the fall time back to τ, but we have increased the input capacitance by one (to 4) (b)

Logical Effort of NAND

- If we scale (b) back so that the input capacitance is the same as the inverter, the delay rises by a factor $g = 4/3$ (c).
- LE DEFINITION 1: So logical effort is the current delivery ability (R_{out}) of a gate with the same C_{in} as an inverter compared to the inverter.



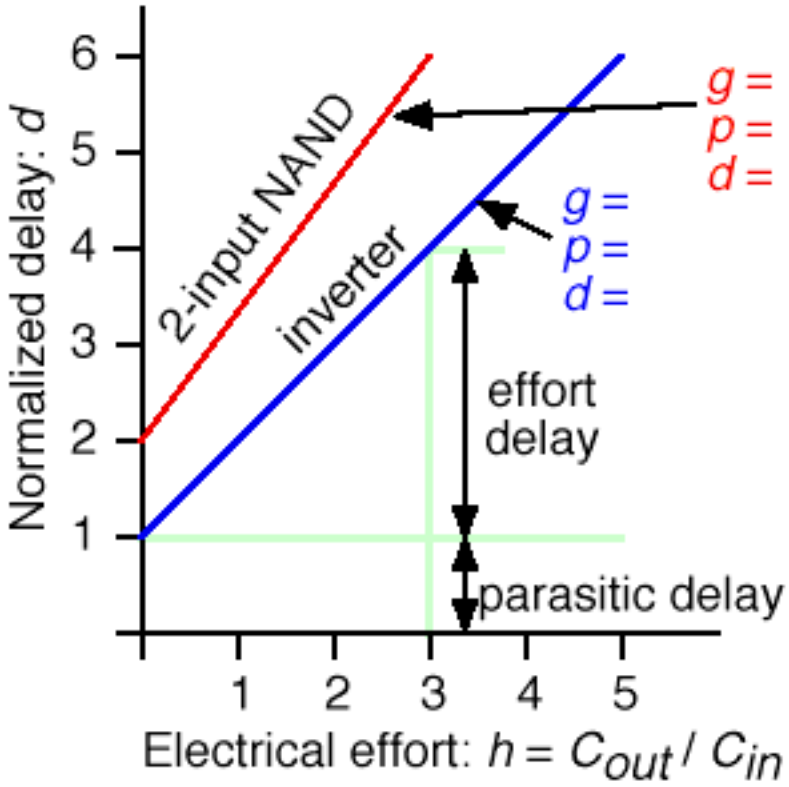
(c)

- LE DEFINITION 2: Alternatively, logical effort is the ratio of the input capacitance of a gate with delay τ (i.e., $R_{\text{out}} = R$) to the input capacitance of the inverter

Electrical Effort and Delay

- **Electrical effort is defined to be the contribution of fanout to delay** $h = C_{\text{load}} / C_{\text{in}}$**.**
- **So we have separated delay into three components:**
  - **logical effort** g
  - **electrical effort** h
  - **parasitic delay** p
- **Now we can write** $d = gh + p = f + p$
- $f = gh$ **is called the stage effort**
- **Parasitic delay is due primarily to the drain capacitance of the FETs connected to the output.**

Delay Plot



How about a 2-input NOR?

Example: Three input NAND

- **What is g?**
- **Suppose we are driving 10 loads. What is f?**

Table of LE

Table 1: Logical effort of static CMOS gates

Gate type	Number of inputs					
	1	2	3	4	5	*n*
inverter	1					
NAND		4/3	5/3	6/3	7/3	(*n*+2)/3
NOR		5/3	7/3	9/3	11/3	(2*n*+1)/3
multiplexer		2	2	2	2	2
XOR, XNOR		4	12	32		

Table 2: Parasitic delay of static CMOS gates

Gate type	Parasitic delay
inverter	p_{inv}
n-input NAND	np_{inv}
n-input NOR	np_{inv}
n-way multiplexer	$2np_{inv}$
2-input XOR, XNOR	$4np_{inv}$

$p_{inv} \approx 1$

parasitic delays depend on diffusion capacitance