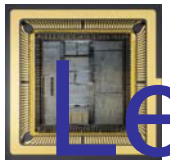
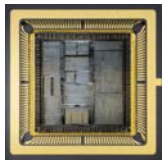


Lecture 6 : Delay Estimation

CSCI 5330
Digital CMOS VLSI Design

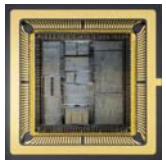
Instructor: Saraju P. Mohanty, Ph. D.

NOTE: The figures, text etc included in slides are borrowed from various books, websites, authors pages, and other sources for academic purpose only. The instructor does not claim any originality.



Lecture Outline

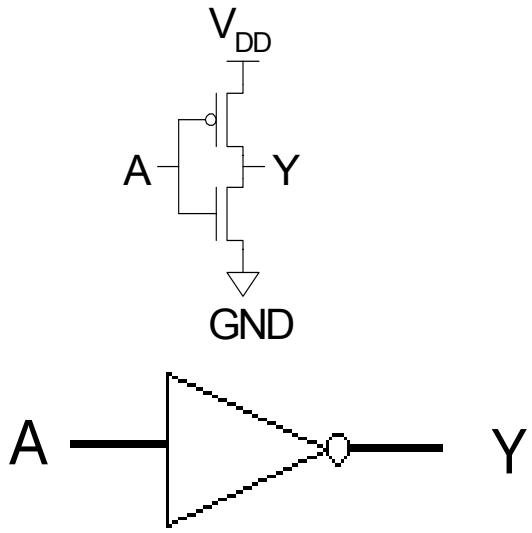
- Delay Definitions
- Switch-level RC Delay Models
- Effective Resistance and Capacitance
- Diffusion Capacitance and Layout Effects
- Elmore Delay Model
- Linear Delay Model
- Parasitic Delay
- Logical Efforts

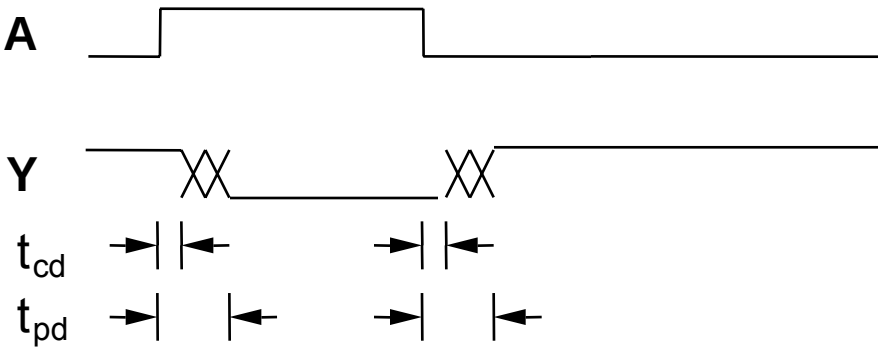


Delay Definitions

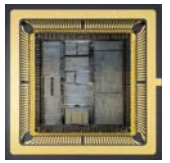
- Combinational logic has two types of delay:
  - Propagation
  - Contamination
- When the input changes, output retains it old value for at least the contamination delay and take on its new value in at most the propagation delay.
- The gate that charges or discharges a node is called driver, and the gates and wires that being driven are called load.

Delay Definitions : Prop. Vs Contamination

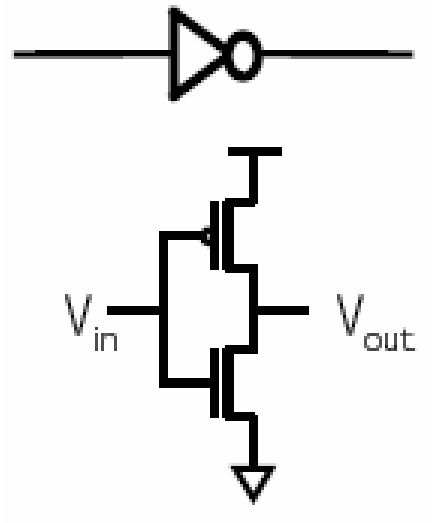


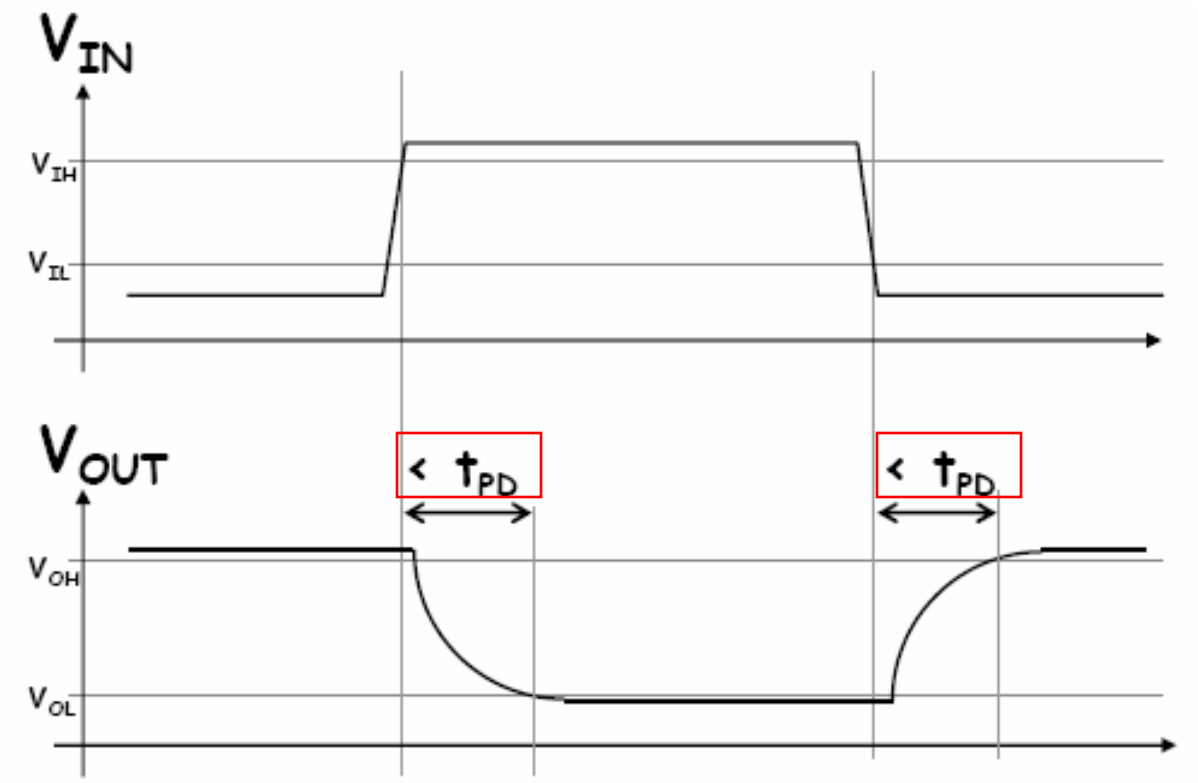


- The output remains unchanged for a time period equal to the contamination delay, t_{cd}
- The new output value is guaranteed to valid after a time period equal to the propagation delay, t_{pd}

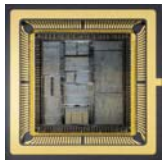


Delay Definitions : Propagation

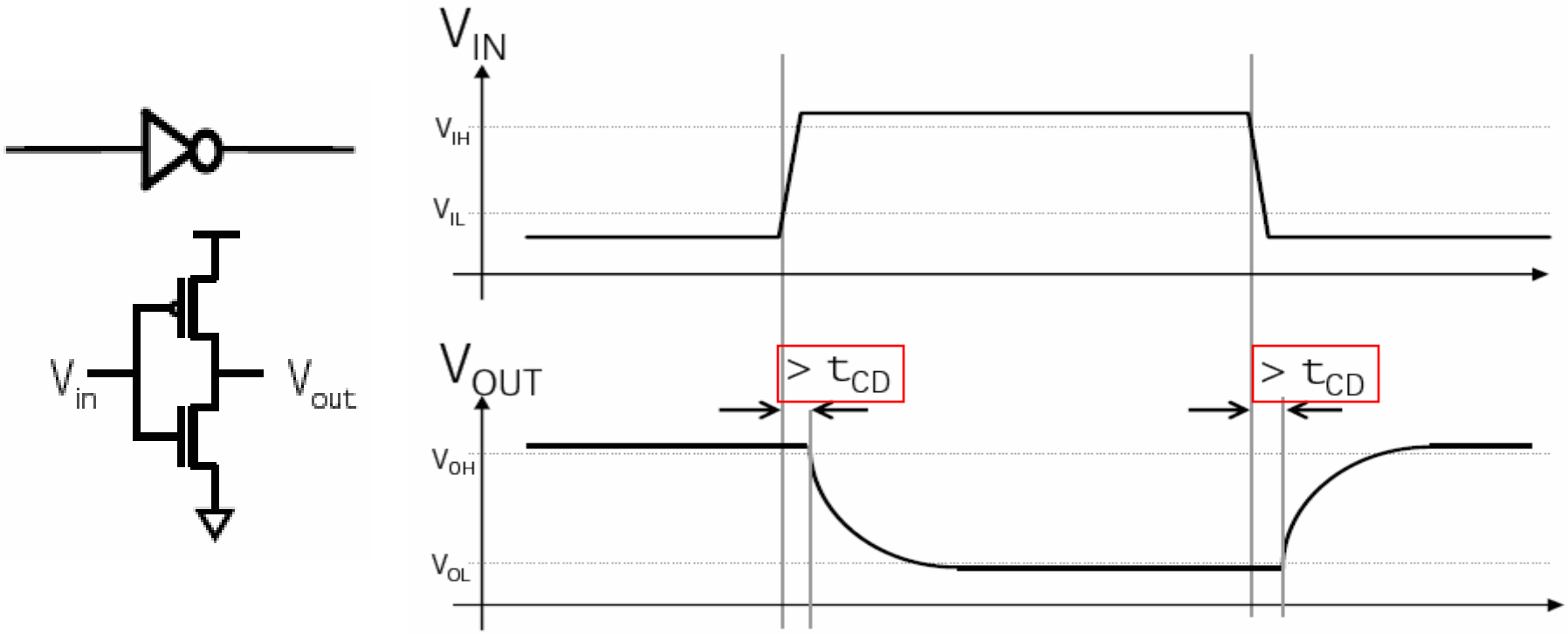




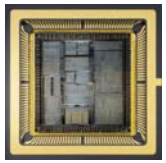
Source: http://www.unc.edu/courses/2003fall/comp/120/001/handouts/Lecture04.pdf



Delay Definitions : Contamination



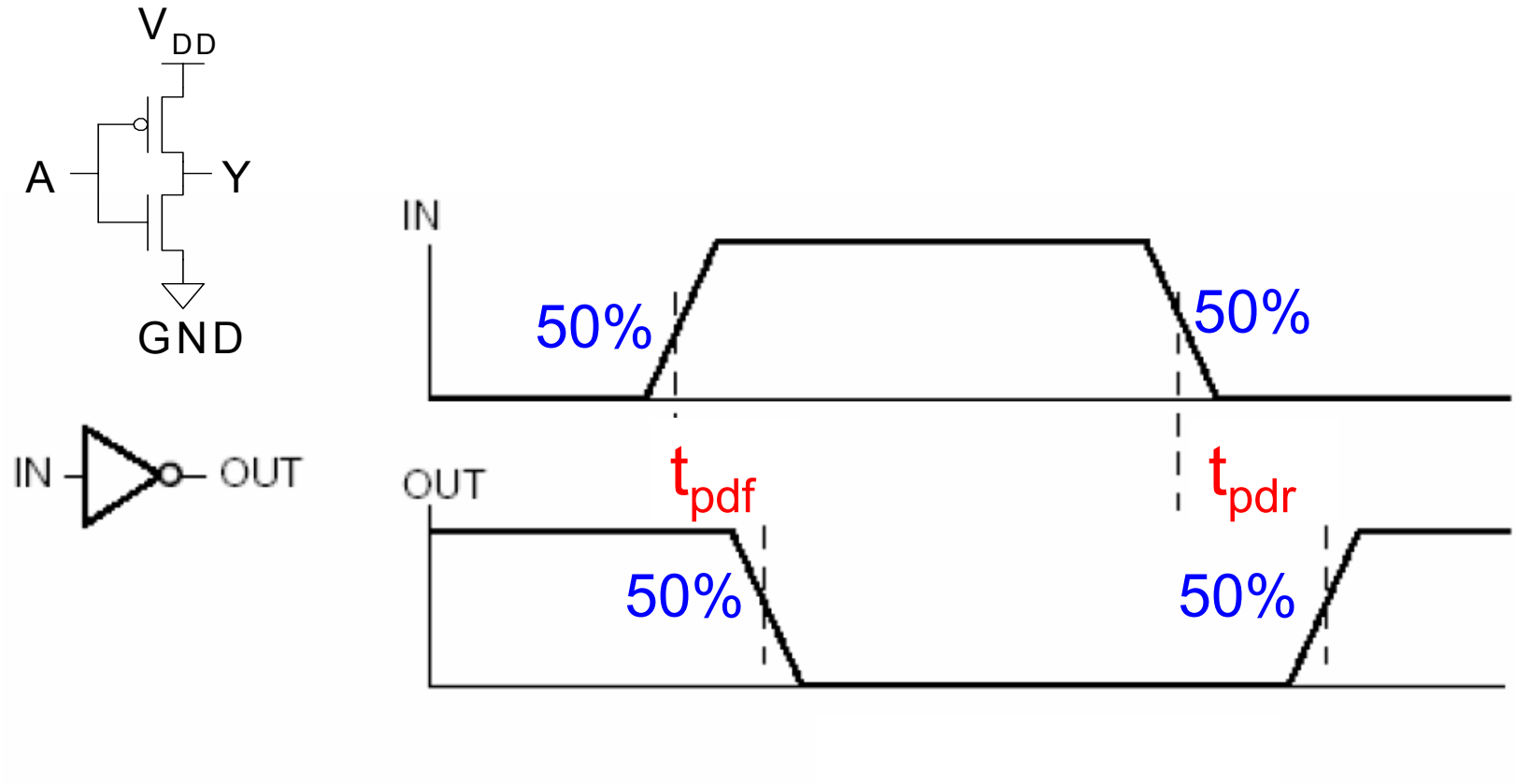
Source: http://www.unc.edu/courses/2003fall/comp/120/001/handouts/Lecture04.pdf

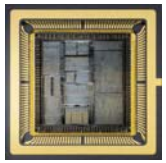


Delay Definitions : Propagation

- **t_{pdr}**: rising propagation delay
  - Time delay from the reference voltage ($V_{DD}/2$) at the input to the reference voltage at the output, when output voltage is going from low-to-high.
- **t_{pdf}**: falling propagation delay
  - Time delay from the reference voltage ($V_{DD}/2$) at the input to the reference voltage at the output, when output voltage is going from high-to-low.
- **t_{pd}**: (average) propagation delay (also max-time)
  - defined in two ways: (maximum or average of two)
    - maximum (t_{pdr}, t_{pdf})
    - $t_{pd} = (t_{pdr} + t_{pdf})/2$

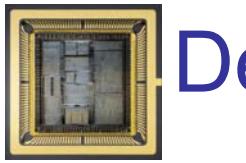
Delay Definitions : Propagation …



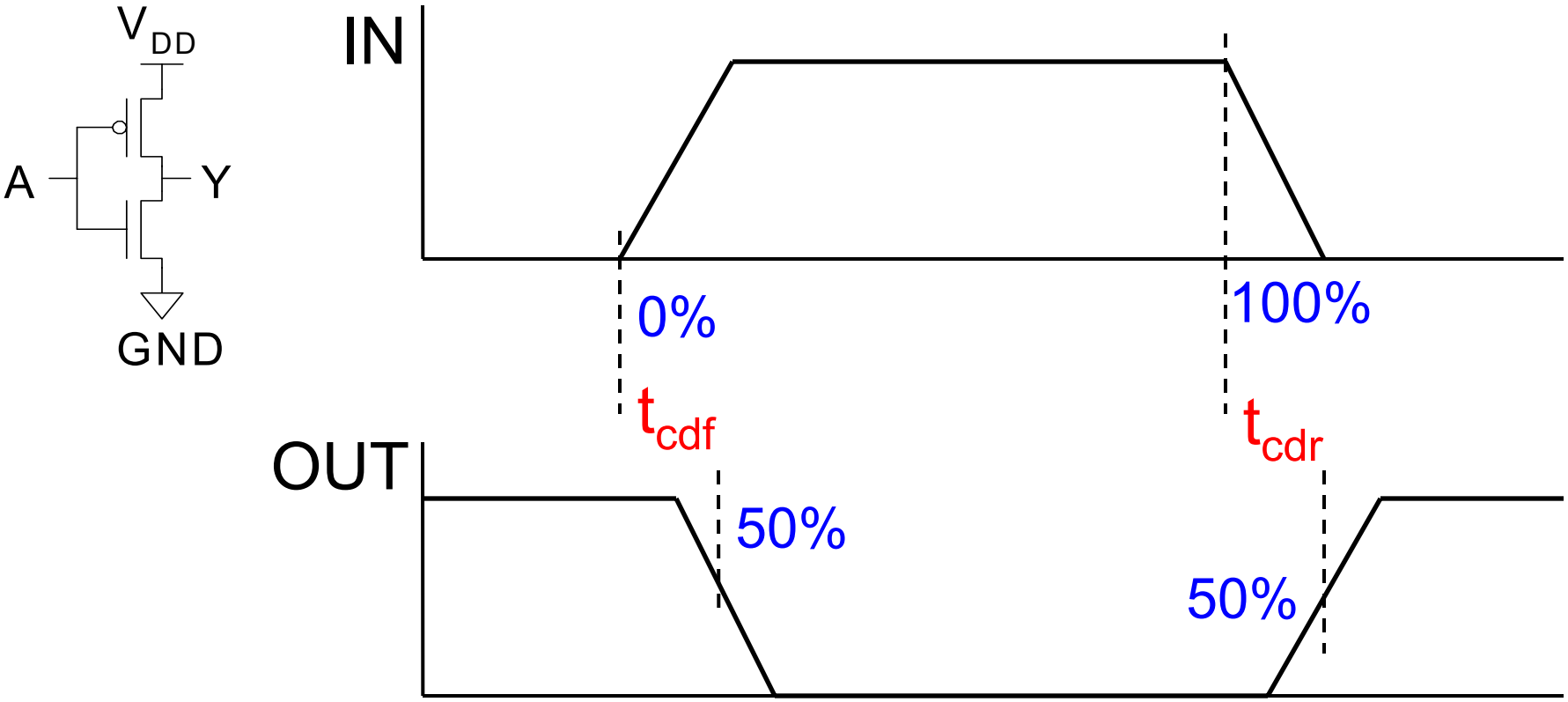


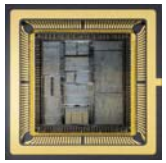
Delay Definitions : Contamination

- **t_{cdr}**: rising contamination delay
  - From input to rising output crossing $V_{DD}/2$
- **t_{cdf}**: falling contamination delay
  - From input to falling output crossing $V_{DD}/2$
- **t_{cd}**: average contamination delay (also min-time)
  - $t_{pd} = (t_{cdr} + t_{cdf})/2$



Delay Definitions : Contamination ….

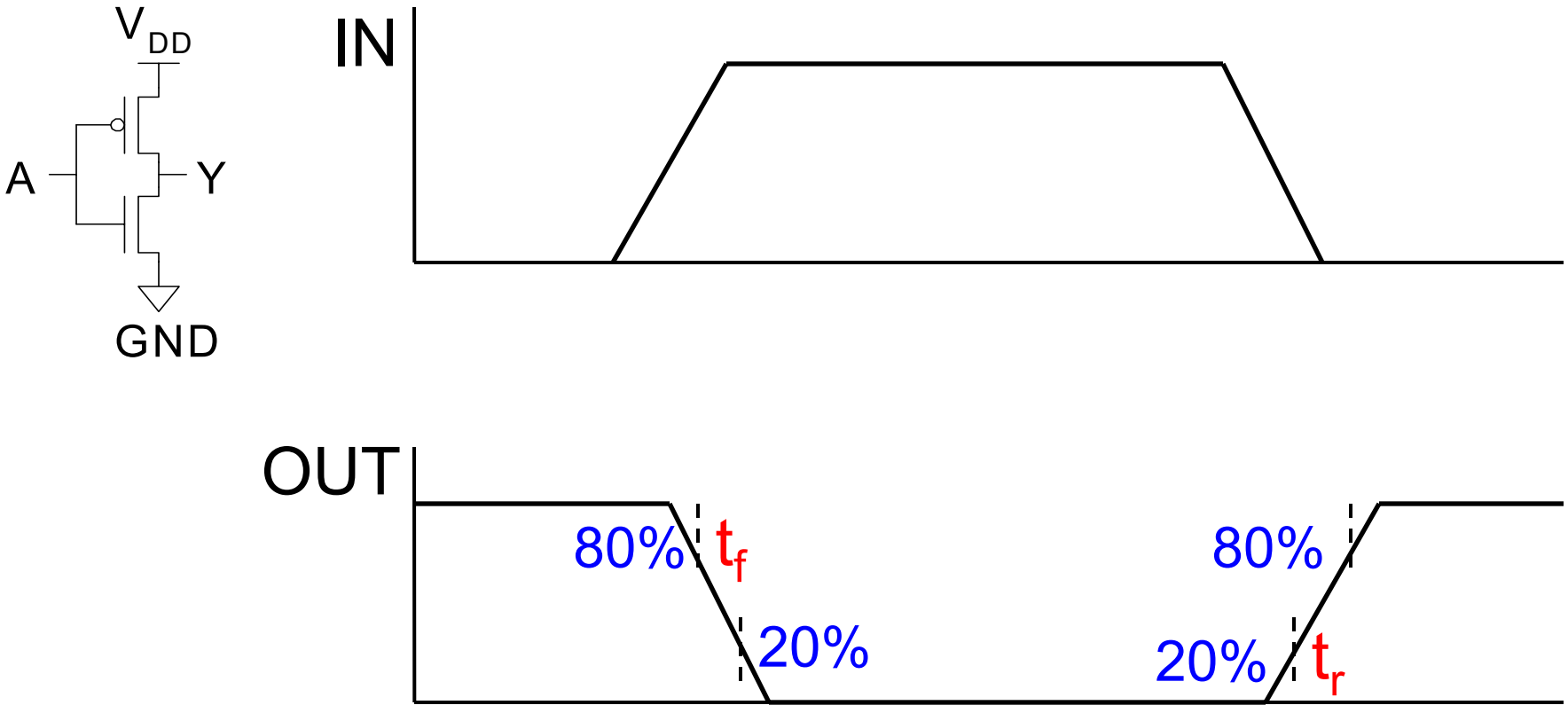


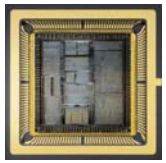


Delay Definitions : Rise and Fall

- **t_r**: rise time
  - From output crossing 0.2 V_{DD} to 0.8 V_{DD}
- **t_f**: fall time
  - From output crossing 0.8 V_{DD} to 0.2 V_{DD}
- Rise / Fall times are also called slope or edge rates.
- Edge Rate: $t_{rf} = (t_r + t_f)/2$

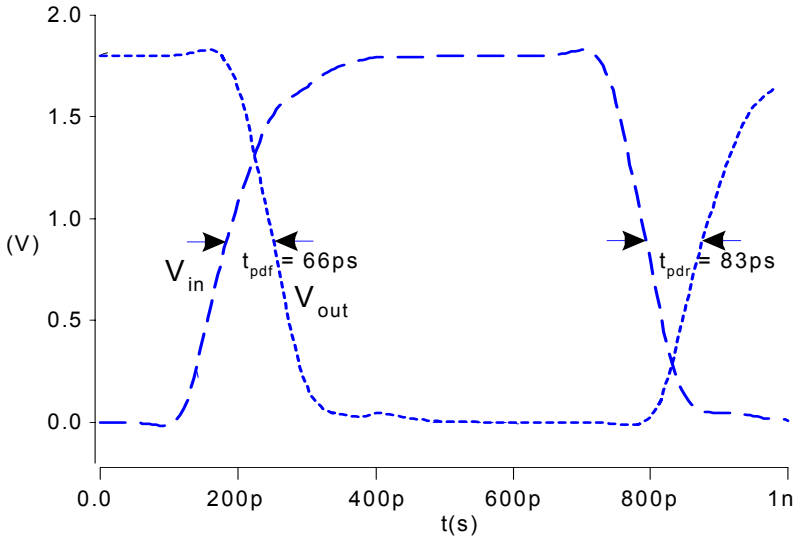
Delay Definitions : Rise and Fall ….

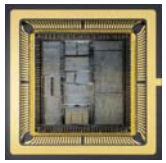




Simulated Inverter Delay

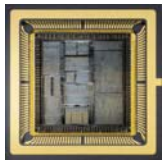
- Solving differential equations by hand is too hard
- SPICE simulator solves the equations numerically
  - Uses more accurate I-V models too!
- But simulations take time to write





Delay Estimation

- We would like to be able to easily estimate delay
  - Not as accurate as simulation
  - But easier to ask “What if?”
- The step response usually looks like a 1st order RC response with a decaying exponential.
- Use RC delay models to estimate delay
  - C = total capacitance on output node
  - Use effective resistance R
  - So that t_{pd} = RC
- Characterize transistors by finding their effective R
  - Depends on average current as gate switches



Switch-level RC Delay Models

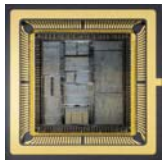
- RC models treat MOSFETs as switches in series with resistors.
- Unit effective resistance R can be obtained from any operating point of I-V characteristics as:

$$R = 1 / (\partial I_{ds} / \partial V_{ds})$$

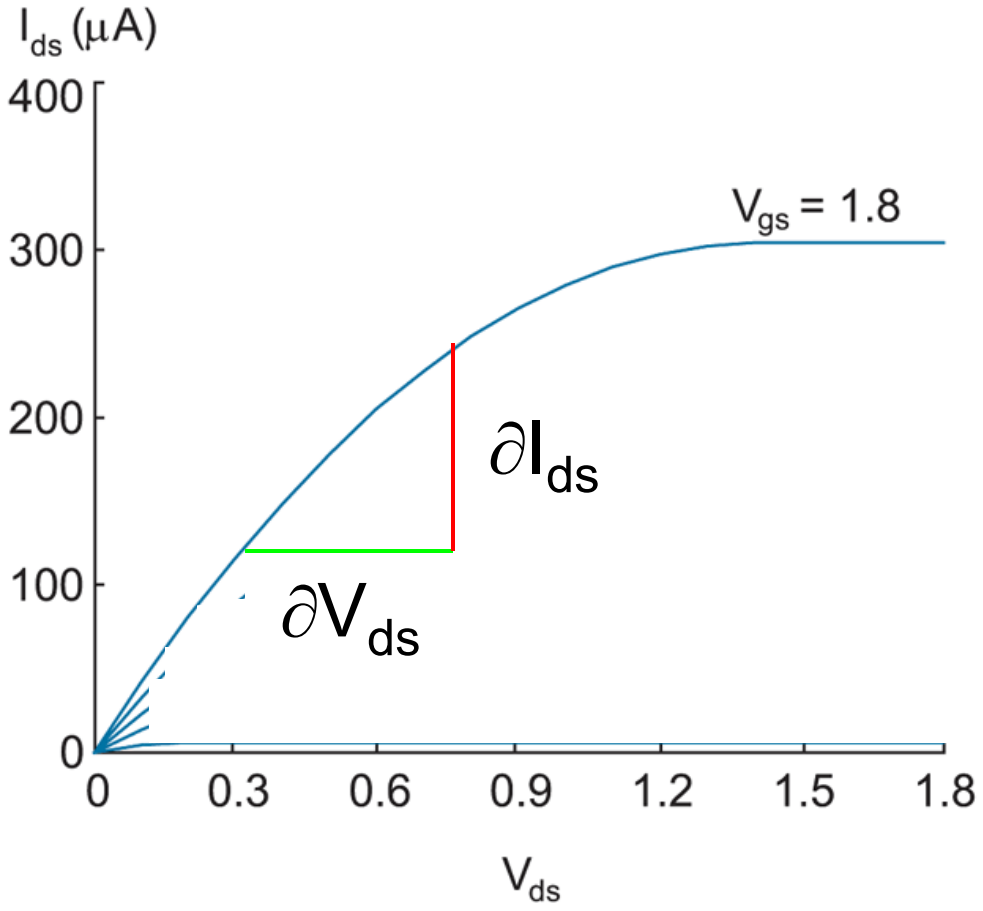
- When ∂V_{ds} is small the resistance R can be obtained by differentiating the I_{ds} equation:

$$R = 1 / [\beta (V_{gs} - V_t)]$$

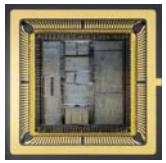
- NOTE: The above way of calculating resistance is not practically accurate as the non-ideal effects (velocity saturation) have strong impact on it.



Switch-level RC Delay Models …



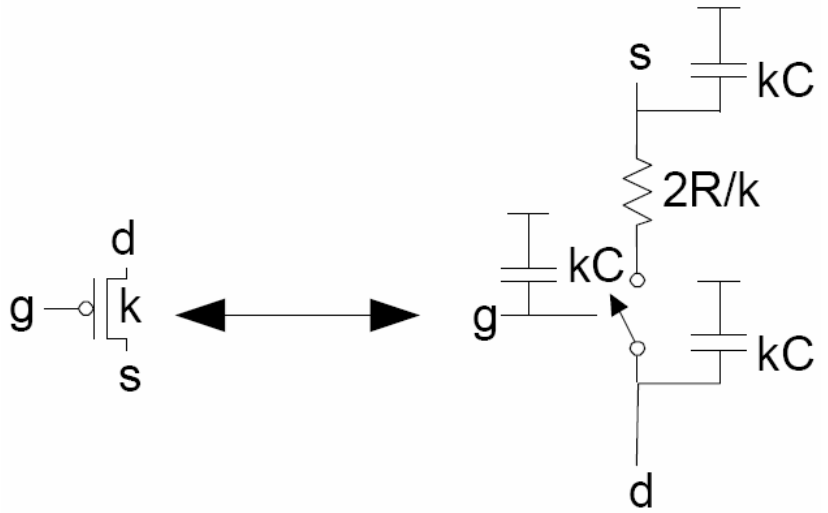
Slope of a curve gives conductance, inverse of which is resistance.



Switch-level RC Delay Models …

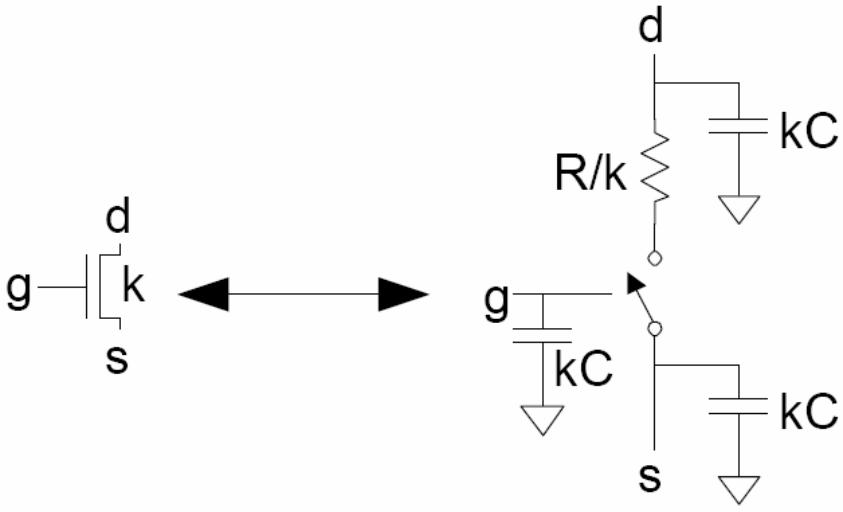
- Use equivalent circuits for MOS transistors
  - Ideal switch + capacitance and ON resistance
  - Unit NMOS has resistance R, capacitance C
  - Unit PMOS has resistance 2R, capacitance C
- Capacitance proportional to width: If unit effective resistance is R, then the transistor of width *k* units has resistance R/*k*.
- Resistance inversely proportional to width: If C is the capacitance of a unit transistor, then the transistor of width *k* units has capacitance *k*C.

RC Delay Models: Inverter



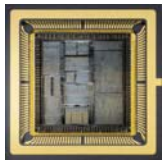
PMOS equivalent RC model:

- Width of transistor is *k* units
- Both gate and diffusion capacitances shown
- One terminal is shown connected to V_{dd} (n-well)

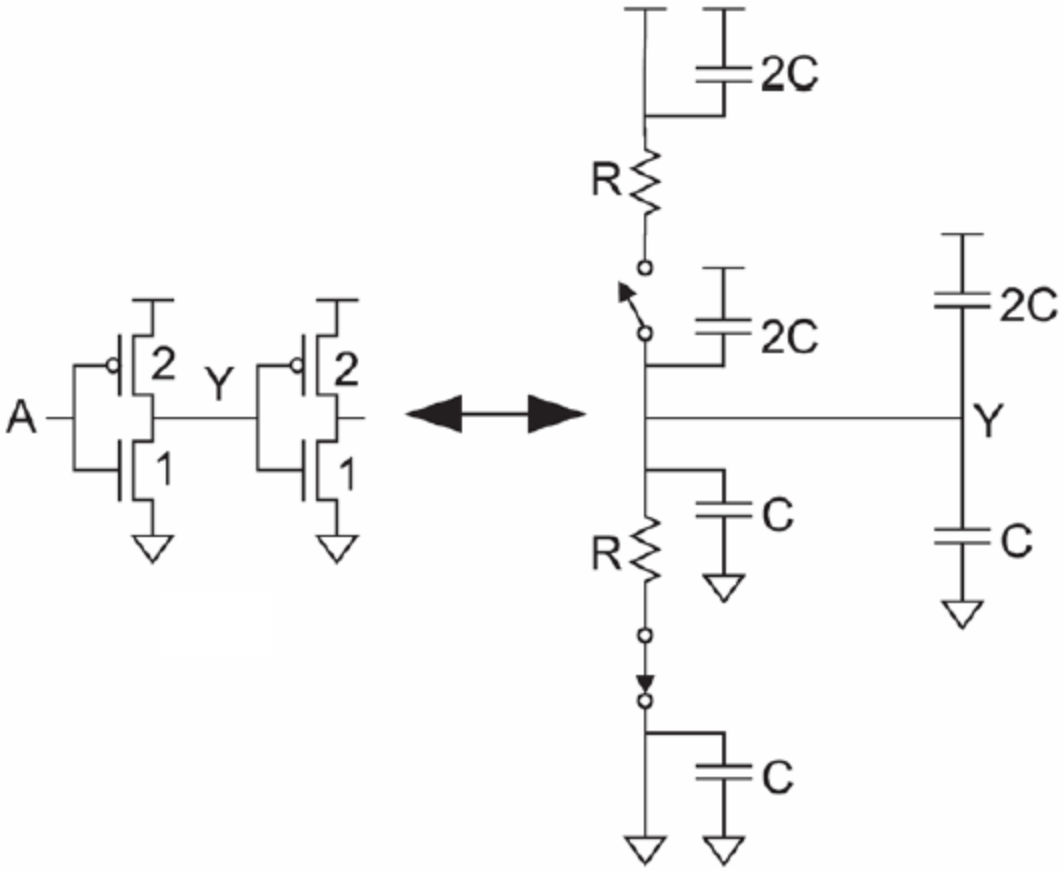


NMOS equivalent RC model:

- Width of transistor is *k* units
- Both gate and diffusion capacitances shown
- One terminal is shown connected to GND (substrate)



RC Delay Models: Inverter …

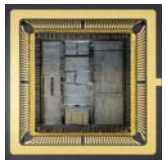


Estimation of delay of a fanout-of-1 inverter.

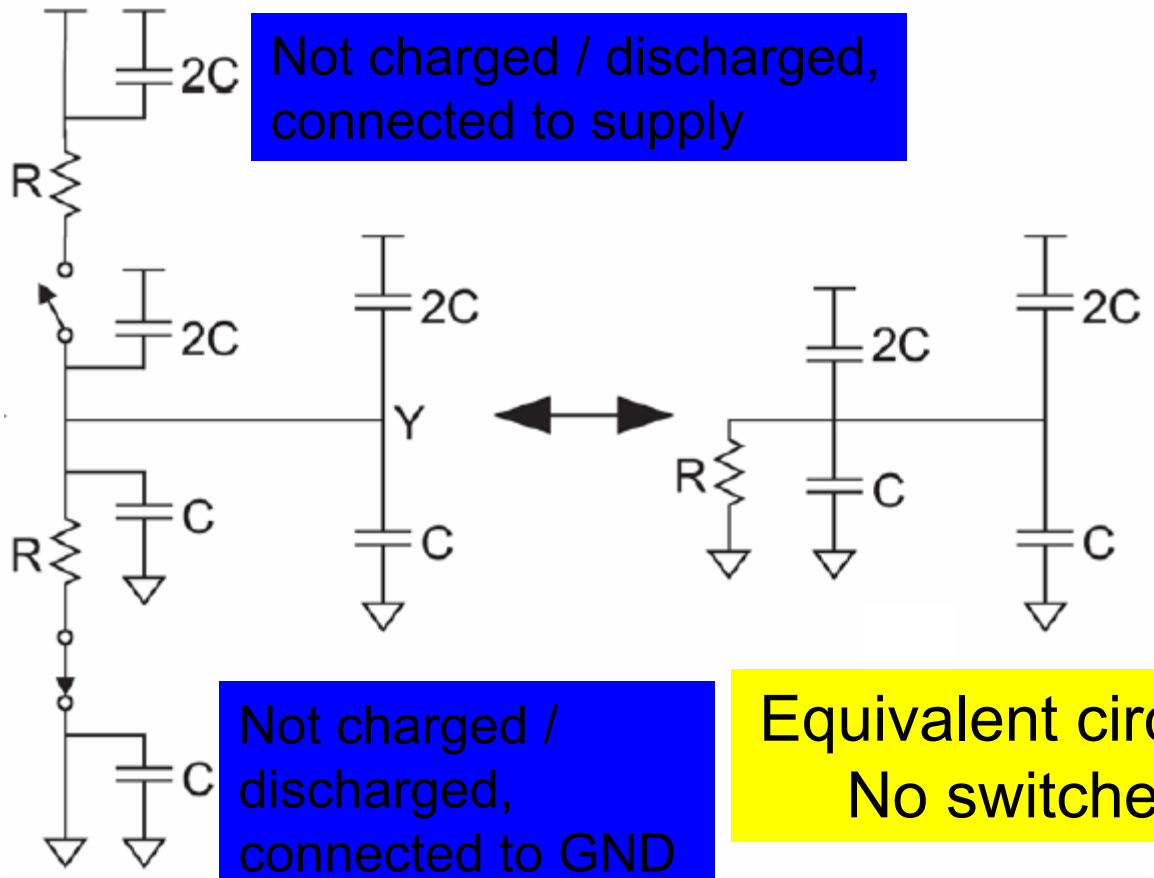
NMOS is of 1-unit width and PMOS is of 2-unit width to achieve equal fall / rise resistance.

Inverter fanout-of-1

Equivalent circuit : 1st inverter driving 2nd



RC Delay Models: Inverter …



Not charged / discharged, connected to supply

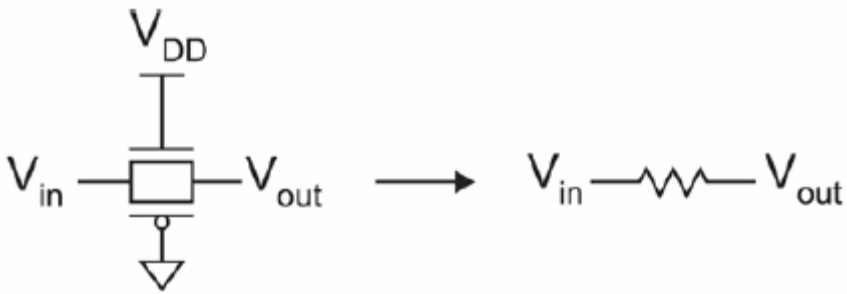
Not charged / discharged, connected to GND

Equivalent circuit : No switches

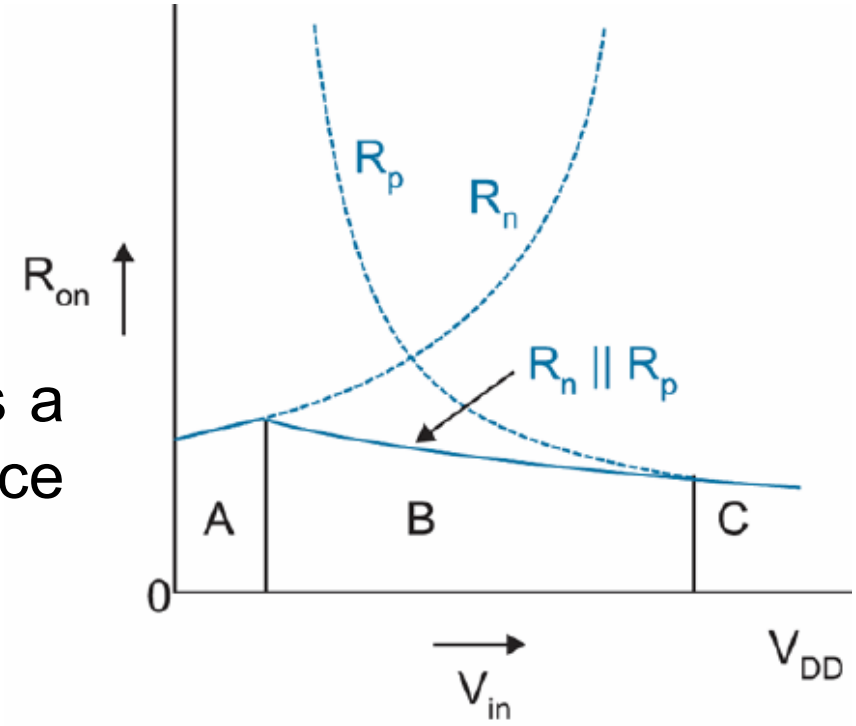
Equivalent circuit : 1st inverter driving 2nd

- T_{pd} =R*(6C) =6RC
- Time constant τ = RC

Switch-level RC Delay Models: Tx gate

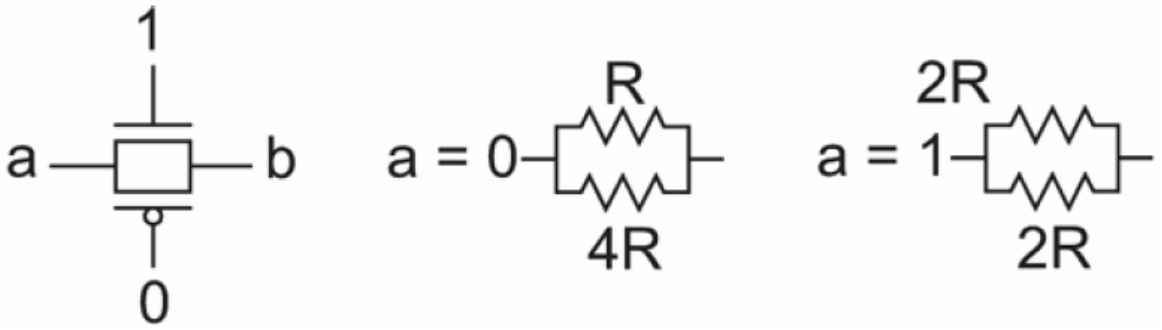


Trans gate effectively works as a voltage controlled resistance between input and output.



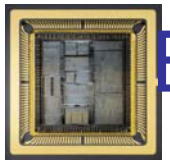
- A – NMOS in linear, PMOS in cut-off
- B – NMOS in linear, PMOS in linear
- C – NMOS in cut-off, PMOS in linear

Switch-level RC Delay Models: Tx gate



Effective resistance of a unit transmission gate

- R for a MOS is greater in its poor direction.
- NMOS passing ‘1’ – effective resistance is 2R
- PMOS passing ‘0’ – effective resistance is 4R
- When a = 0 : R_{Tx} = R parallel with 4R = (4/5)R
- When a = 1 : R_{Tx} = 2R parallel with 2R = R

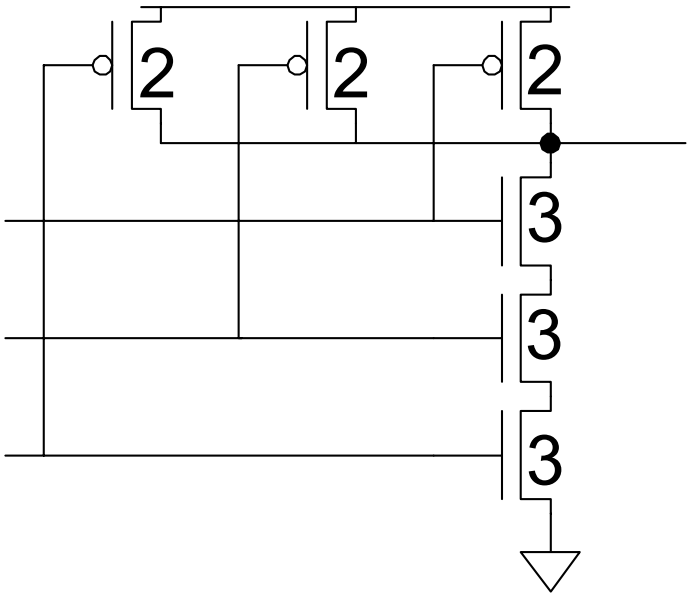


Effective Resistance and Capacitance

- Parallel and series transistors combine like conventional resistors.
- When in series: Total resistance is the sum of all
- When in parallel: Total conductance is the sum of conductance, inverse of which is the total resistance.
- Resistance is low if they are in parallel.
- Worst case delay → when only one of several parallel transistors is ON.

Effective R and C : 3-input NAND Example

Question: Sketch a 3-input NAND with transistor widths chosen to achieve effective rise and fall resistances equal to a unit inverter (R).

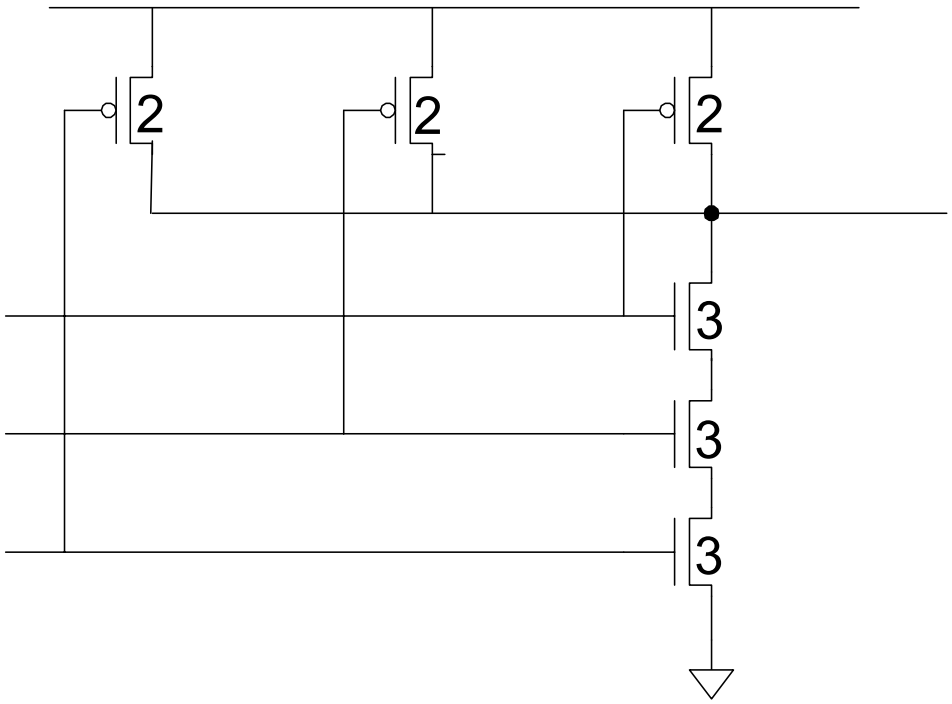


3-input NAND

- Each NMOS should have R/3 resistance
- Each PMOS should have R resistance (worst case one even one ON should provide R resistance).
- Since 1-unit NMOS has R resistance, so its W/L is 3.
- Also 1-unit PMOS has 2R resistance, so its W/L is 2.

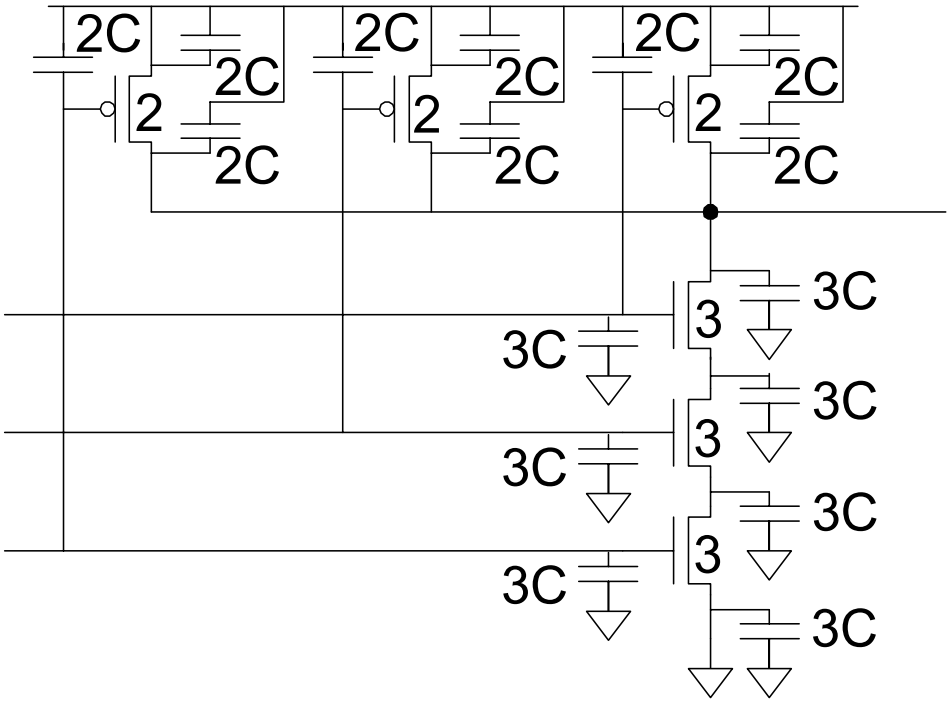
Effective R and C : 3-input NAND Capacitance

Question: Annotate the 3-input NAND gate with gate and diffusion capacitance.



3-input NAND

Effective R and C : 3-input NAND Capacitance

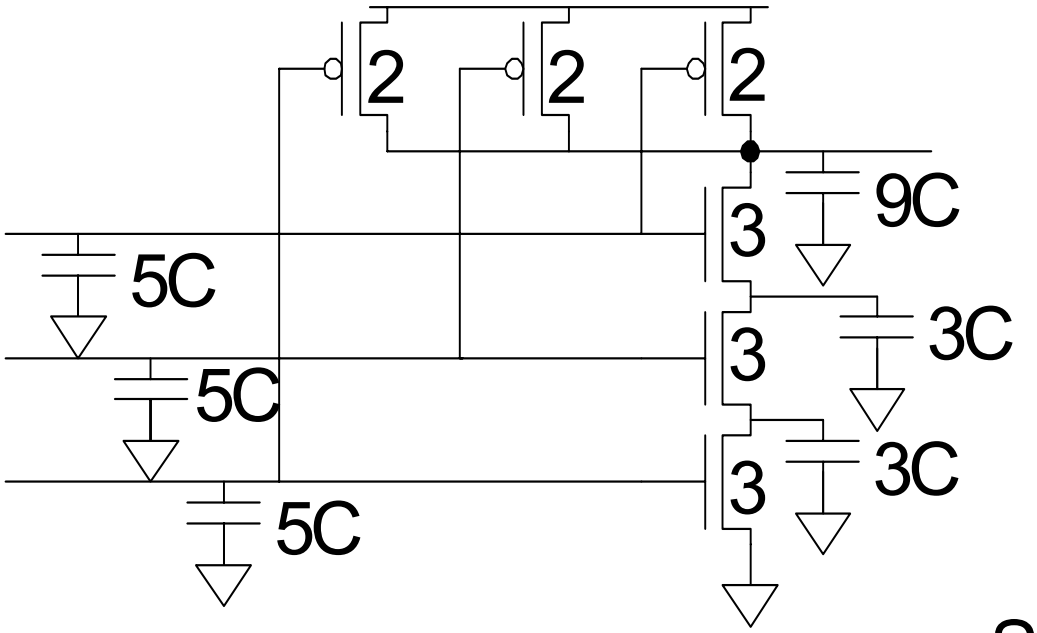


3-input NAND

Recall

- Unit NMOS has resistance R, capacitance C
- Unit PMOS has resistance 2R, capacitance C
- *k* units has capacitance *k*C.

Effective R and C : 3-input NAND Capacitance

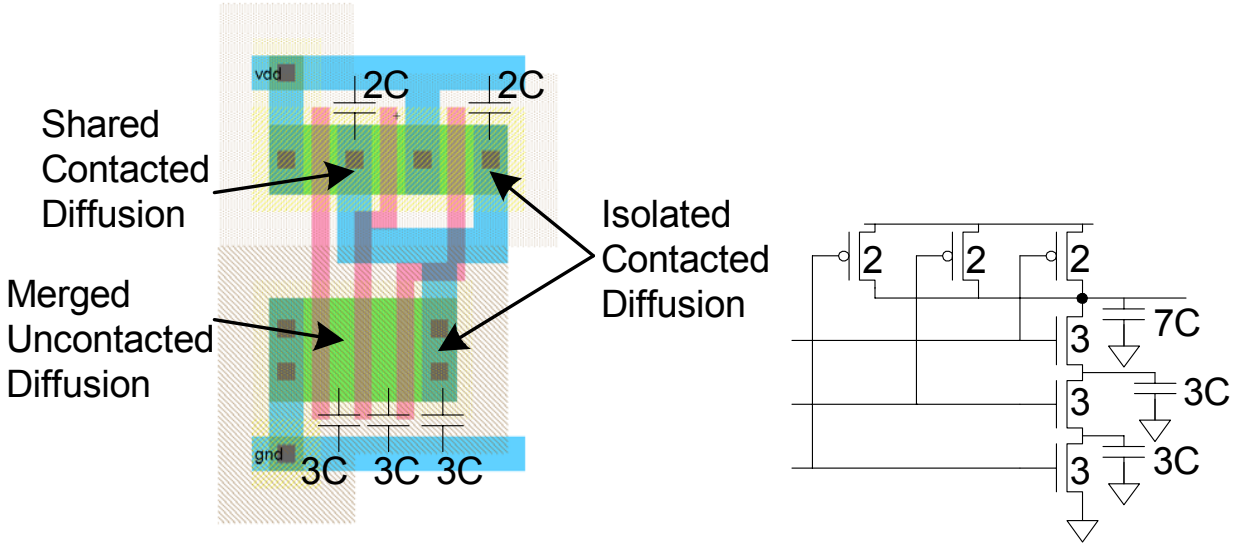


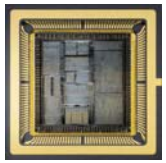
3-input NAND

Shorted capacitances deleted and remaining capacitances lumped.

Effective R and C : Diffusion Cap Example

- We assumed contacted diffusion on every s / d.
- Good layout minimizes diffusion area
- Example: NAND3 layout shares 1 diffusion contact
  - Two of the PMOS share a single diffusion region
  - Reduces output capacitance by 2C



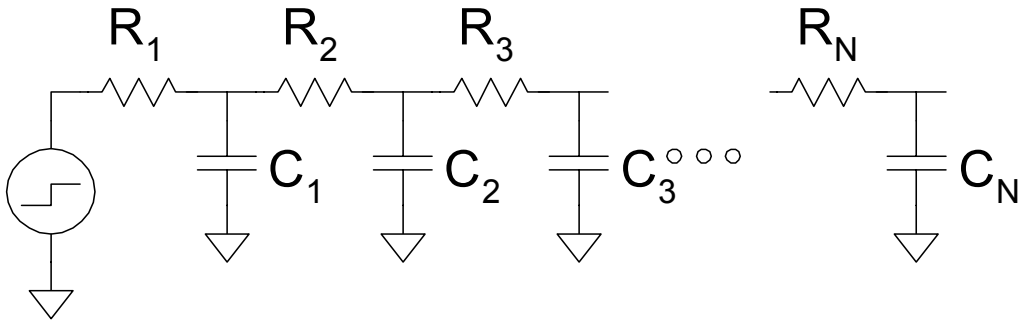


Elmore Delay Model

- ON transistors look like resistors
- Pullup or pulldown network modeled as *RC ladder*
- Elmore delay of RC ladder

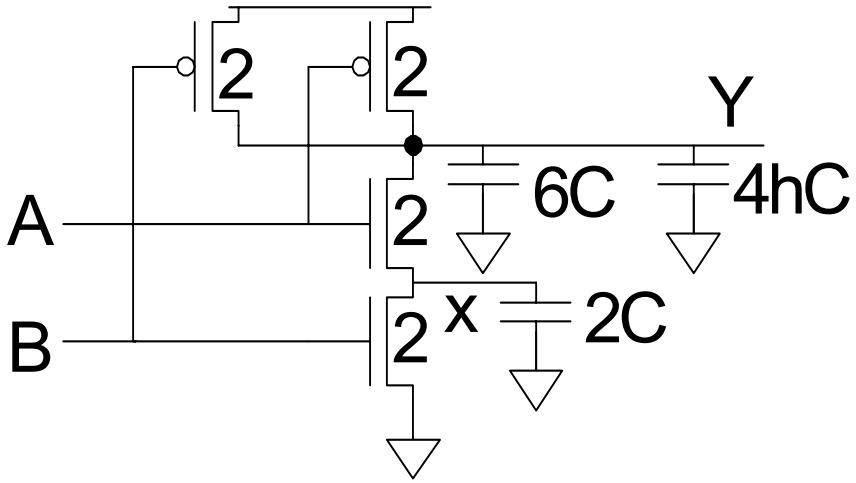
$$t_{pd} \approx \sum_{\text{nodes } i} R_{i-to-source} C_i$$

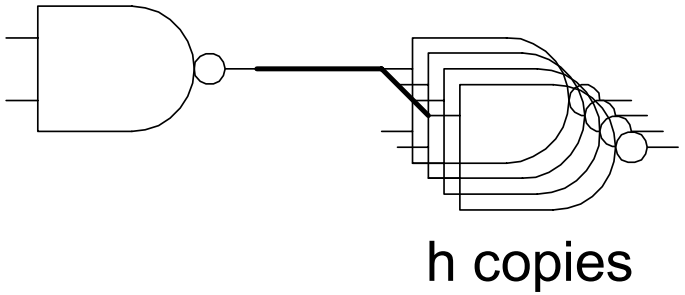
$$= R_1 C_1 + (R_1 + R_2) C_2 + \ldots + (R_1 + R_2 + \ldots + R_N) C_N$$



Rising and Falling Delay Example

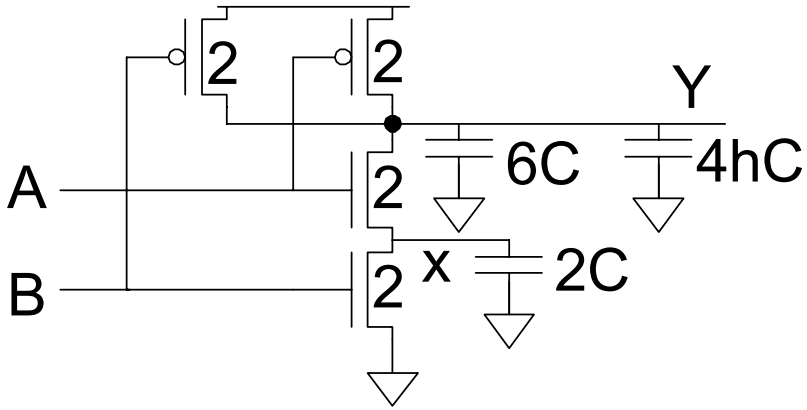
Example: Estimate rising and falling propagation delays of a 2-input NAND driving *h* identical gates.

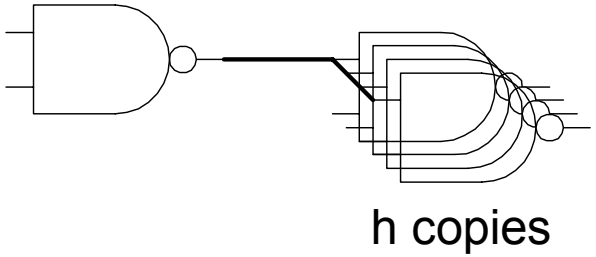


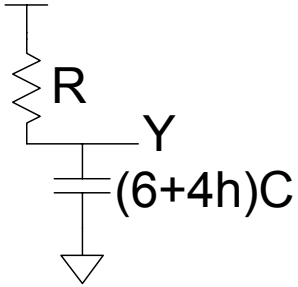


Rising and Falling Delay Example

- Rising propagation delay :



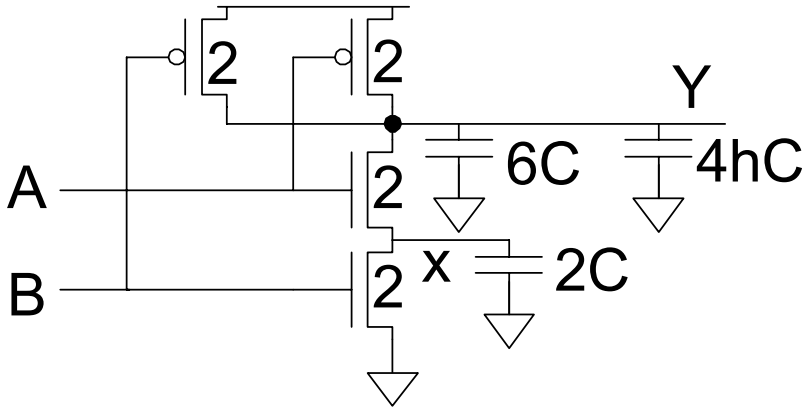


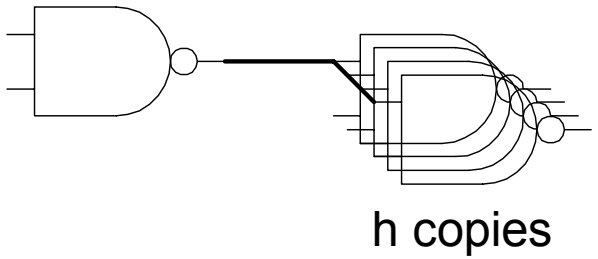


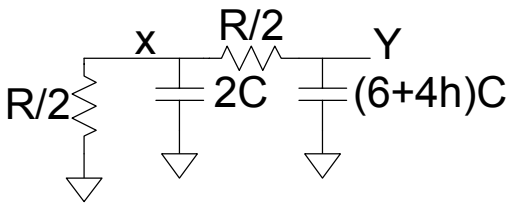
$$t_{pdr} = (6 + 4h)RC$$

Rising and Falling Delay Example

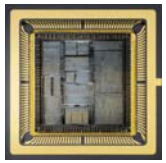
- Falling propagation delay :





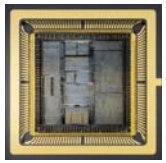


$$t_{pdf} = (2C)\left(\frac{R}{2}\right) + \left[(6+4h)C\right]\left(\frac{R}{2}+\frac{R}{2}\right)$$
$$= (7+4h)RC$$



Two Components of Delay

- Delay has two parts
  - Parasitic delay (determined bygate driving its own diffusion capacitance)
    - 6 or 7 RC
    - Independent of load
  - Effort delay (determined by load capacitance)
    - 4h RC
    - Proportional to load capacitance
- The capacitance ratio is called the electrical effort or fanout.



Linear Delay Model

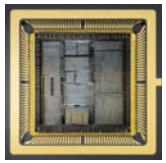
- Express delays in process-independent unit

$$d = \frac{d_{abs}}{\tau}$$

- Delay has two components

$$d = f + p$$

- Effort delay $f = gh$ (a.k.a. stage effort)
  - Again has two components
- g: logical effort
  - Measures relative ability of gate to deliver current
  - $g \equiv 1$ for inverter



Linear Delay Model ….

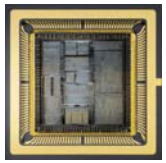
- Express delays in process-independent unit

$$d = \frac{d_{abs}}{\tau}$$

- Delay has two components

$$d = f + p$$

- Effort delay $f = gh$ (a.k.a. stage effort)
  - Again has two components
- h: electrical effort = C_{out} / C_{in}
  - Ratio of output to input capacitance
  - Sometimes called fanout



Linear Delay Model ….

- Express delays in process-independent unit

$$d = \frac{d_{abs}}{\tau}$$

- Delay has two components

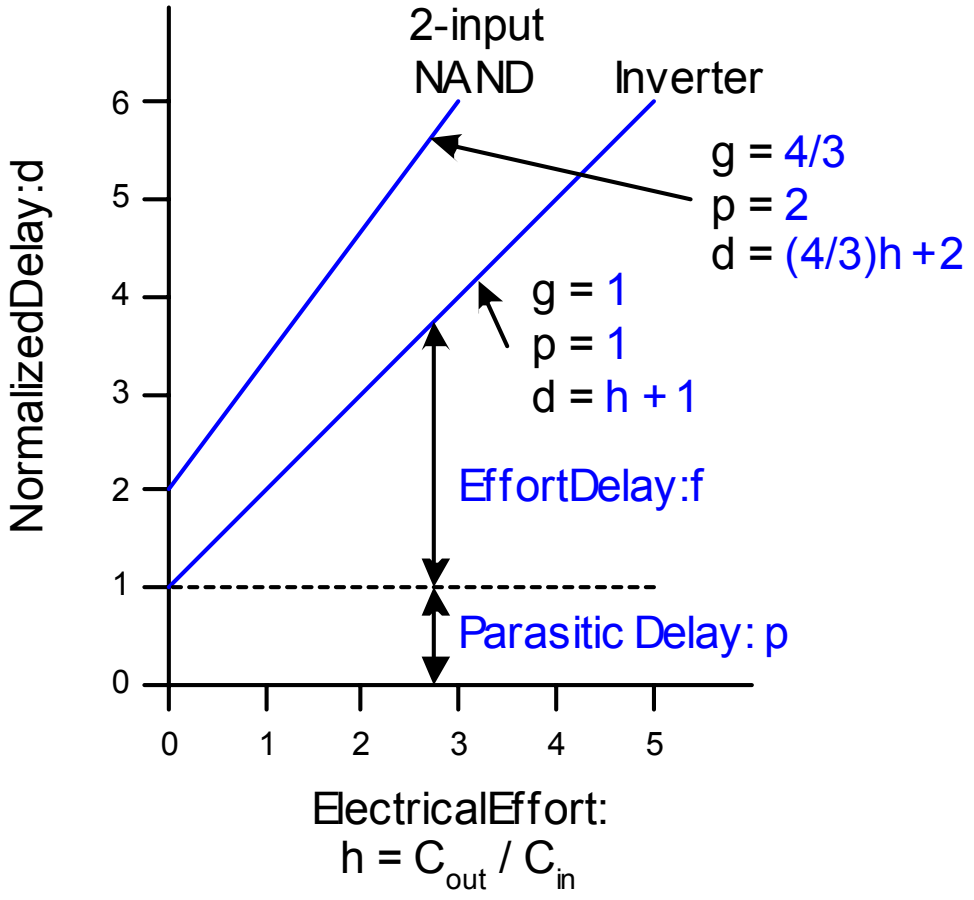
$$d = f + p$$

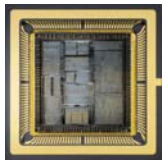
- Parasitic delay *p*
  - Represents delay of gate driving no load
  - Set by internal parasitic capacitance

Linear Delay Model : Delay Vs Fanout

$$d = f + p$$
$$= gh + p$$

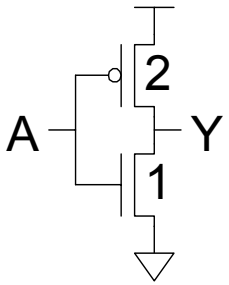
- What about NOR2?



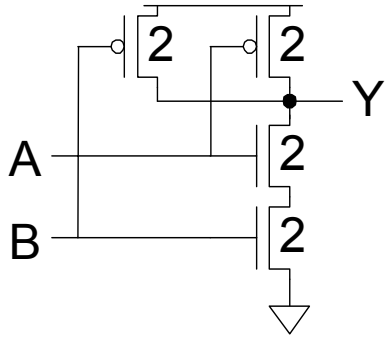


Computing Logical Effort

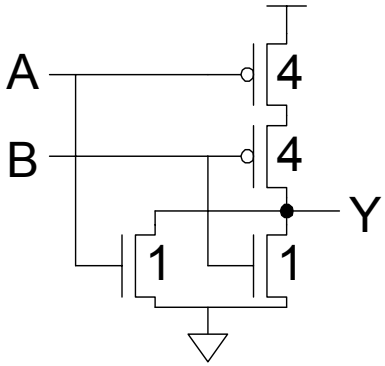
- Definition: Logical effort is the ratio of the input capacitance of a gate to the input capacitance of an inverter delivering the same output current.
- Measure from delay vs. fanout plots
- Or estimate by counting transistor widths



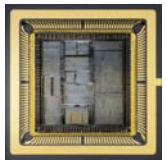
$C_{in} = 3$
$g = 3/3$



$C_{in} = 4$
$g = 4/3$



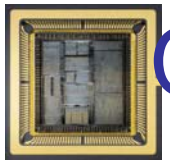
$C_{in} = 5$
$g = 5/3$



Catalog of Gates : Logical Effort

- Logical effort of common gates

Gate type	Number of inputs				
	1	2	3	4	n
Inverter	1				
NAND		4/3	5/3	6/3	(n+2)/3
NOR		5/3	7/3	9/3	(2n+1)/3
Tristate / mux	2	2	2	2	2
XOR, XNOR		4, 4	6, 12, 6	8, 16, 16, 8	



Catalog of Gates : Parasitic Delay

- Delay of a gate when it drives zero load.
- Parasitic delay of common gates
  - In multiples of p_{inv} (≈ 1)

Gate type	Number of inputs				
	1	2	3	4	n
Inverter	1				
NAND		2	3	4	n
NOR		2	3	4	n
Tristate / mux	2	4	6	8	2n
XOR, XNOR		4	6	8	