EE-429
Fundamentals of VLSI Design

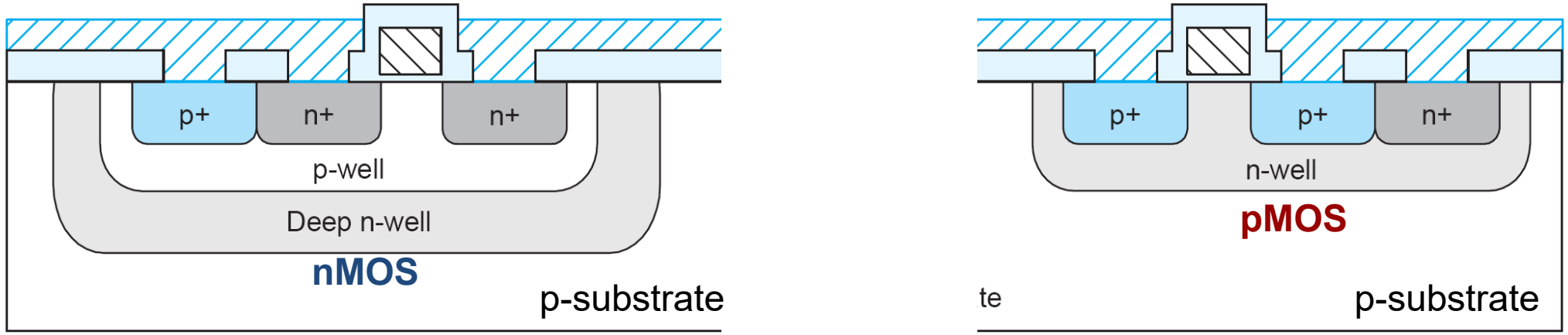
Quick Review of the MOS Transistor for Digital Designers

Andreas Burg, Alexandre Levisse

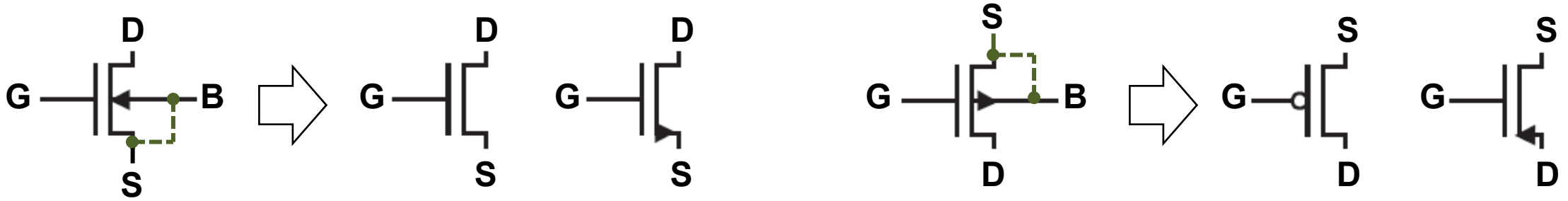
Acknowledgement: Prof. Adam Teman (BIU)

Fundamental Element: nMOS & pMOS Transistors

- Complementary Metal Oxide Semiconductor (**CMOS**) **circuits** are **build from** complementary transistors: **nMOS and pMOS transistors**



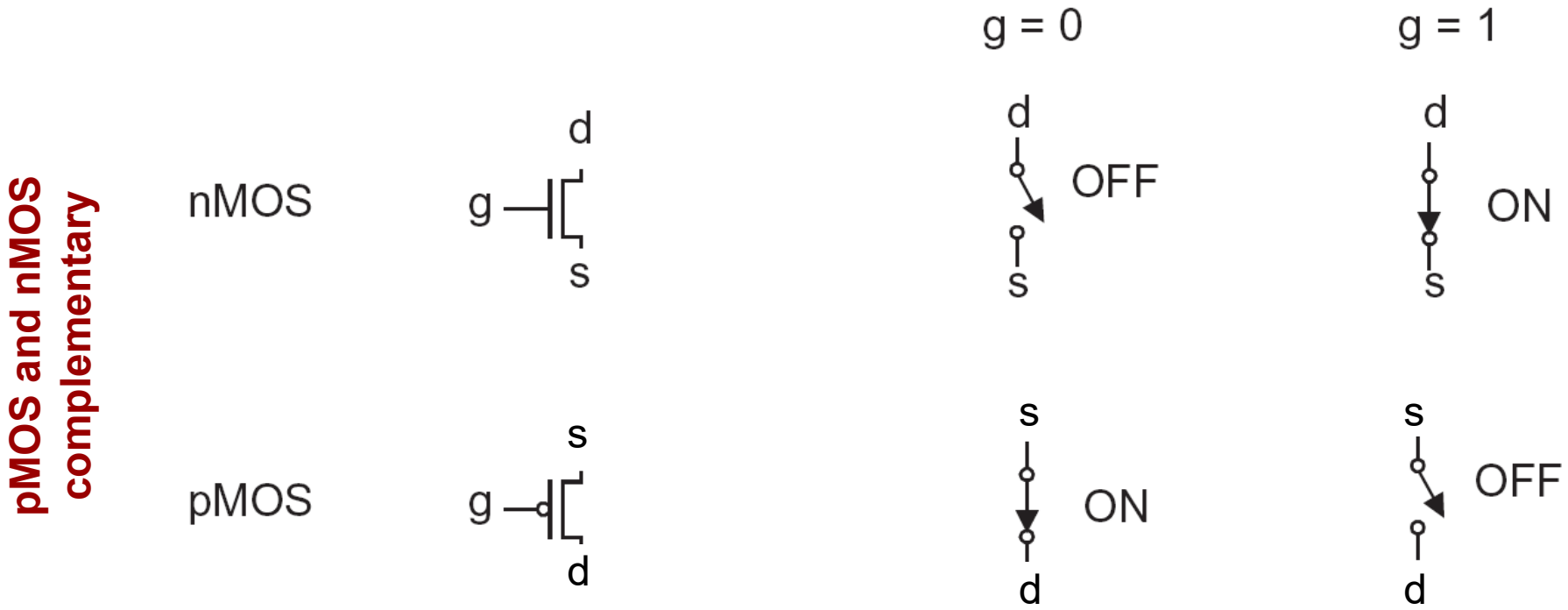
- **CMOS transistors have 4-terminals**: source (S), Drain (D), Gate (G), Bulk (B)
  - **We will often connect Bulk (B) to Source (S)** and omit the Bulk (B) terminal



The Switch Model

- **Boolean** electric **model** for **functional considerations**
  - Logic ‘0’ = GND Logic ‘1’ = VDD
  - Boolean terminals S → s, D → d, G → g

- **Gate** terminal (g) **controls current flow** between source (s) and drain (d)

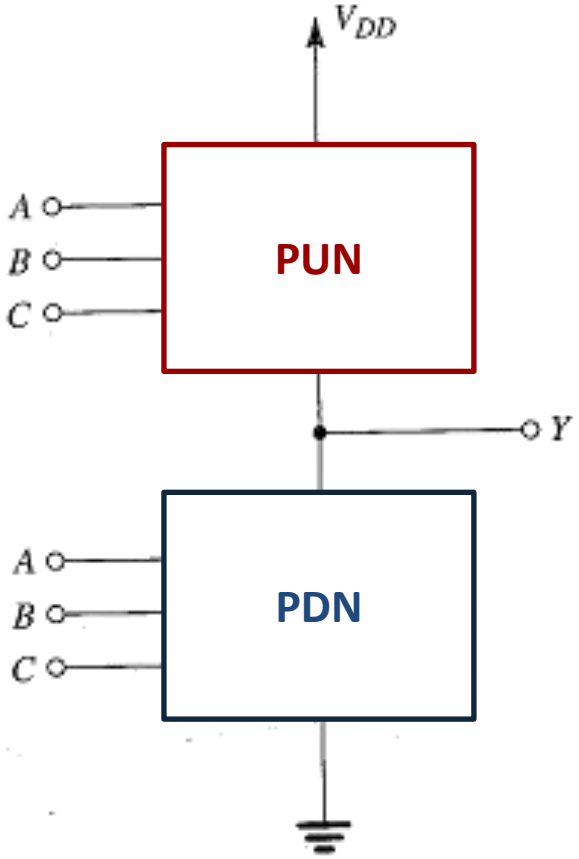


Basic CMOS logic Principal

- **With the switch model, we can construct the basic CMOS gates**

- **Basic idea:**
  - Start from the truth table (input/output relationship)
    - Pull-Up network (PUN):
      Connect OUT to VDD for input combinations that lead to a '1'
    - Pull-Down network (PDN):
      Connect OUT to GND for input combinations that lead to a '0'

A/B/C	000	001	010	011	100	101	110	111
Y	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1

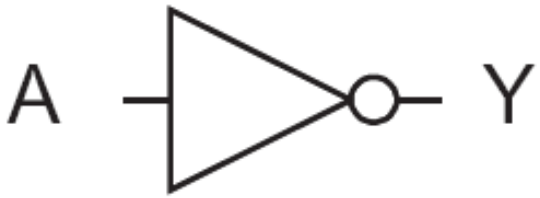


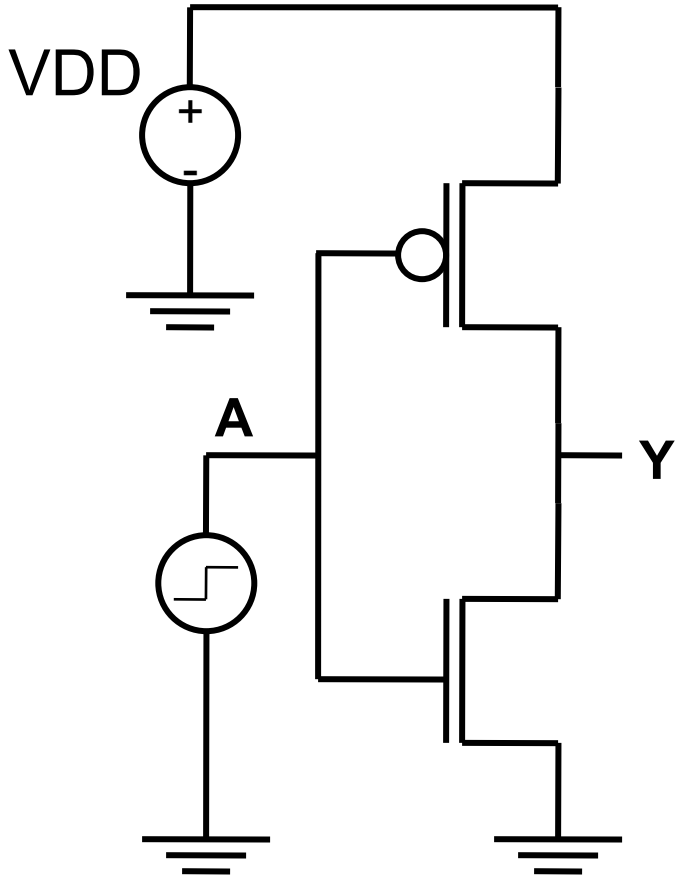
A Simple Inverter

- **The most simple combinational CMOS logic gate is the INVERTER (INV)**

A	Y
0	1
1	0

INV truth table

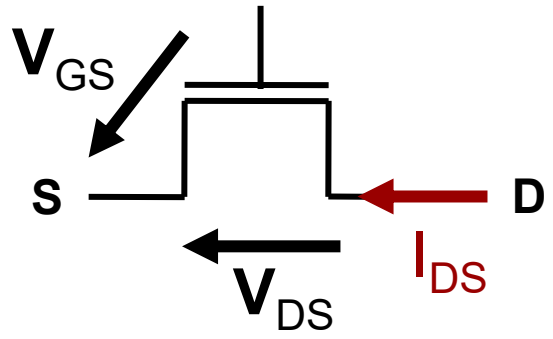




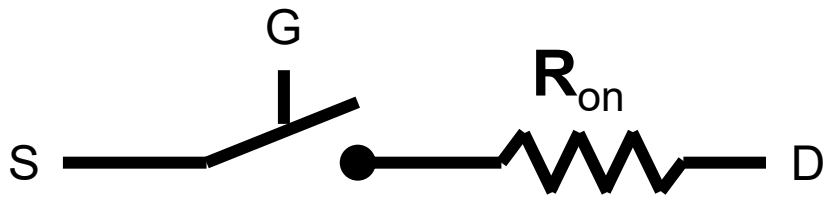
- PUN: pMOS
- PDN: nMOS

A Simple Linear Model

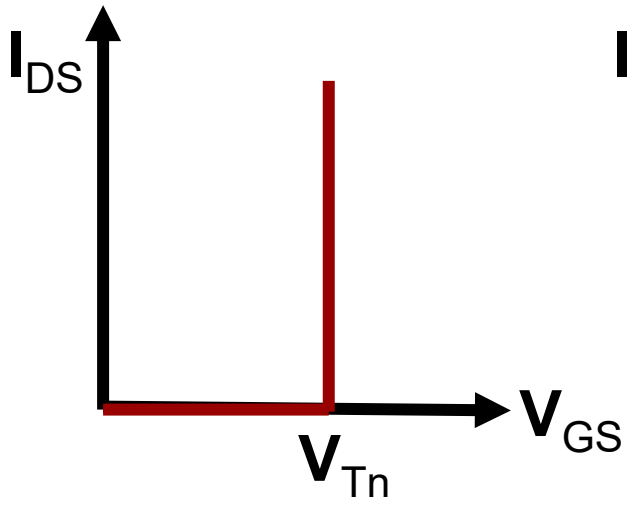
- **Controlled switch with** an **internal resistance** between source and drain

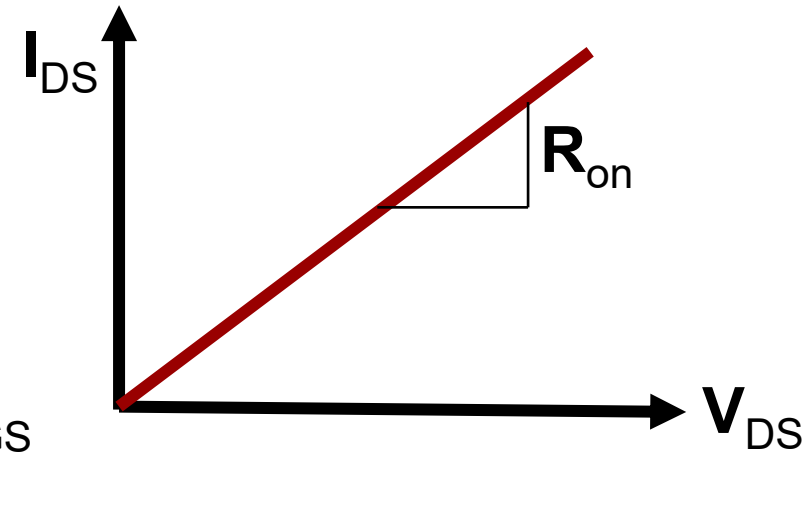


- Characterized by two transfer functions
  - I_{DS} vs V_{GS} characteristics
  - I_{DS} vs V_{DS} characteristics



- **Small signal model**: only accurate over a small V_{DS} range
  -
- Still **often used** even for a large swing of V_{DS} **with a properly fitted R_{on}**





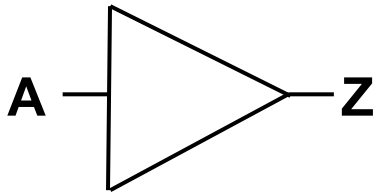
So why not this one?

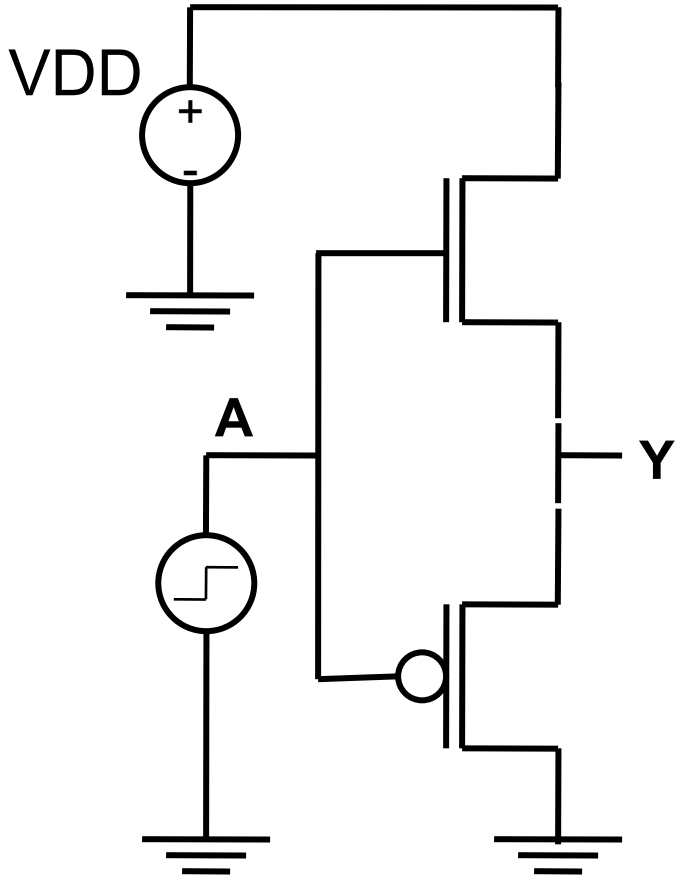
- **Can we also build a BUFFER (BUF) in a similar way?**

A	Z
0	**0**
1	**1**

INV truth table

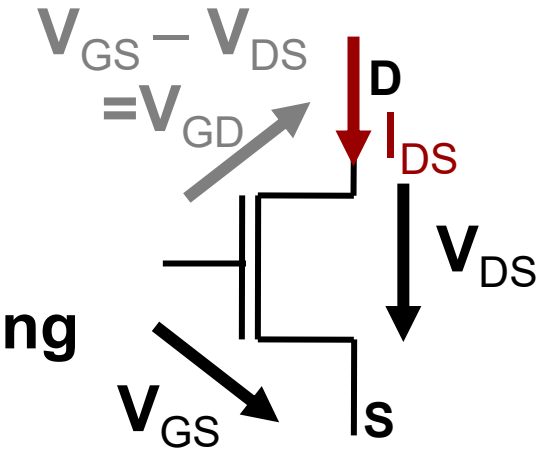
- PUN: nMOS
- PDN: pMOS





nMOS Operation

- **MOS transistor has three operating regions depending on V_{GS} and V_{DS}**

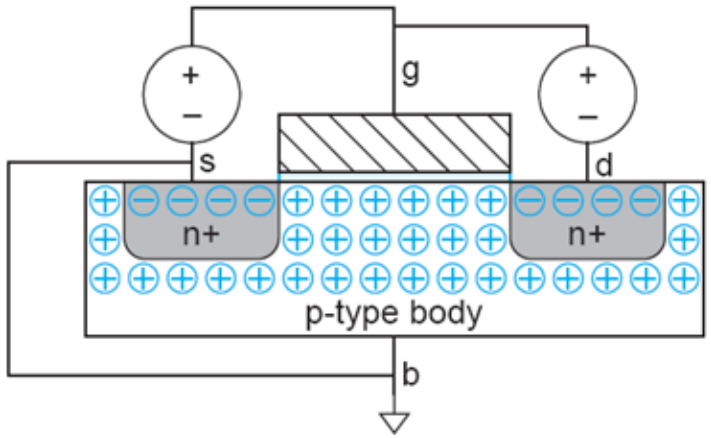


$V_{Tn} > 0$: **Threshold voltage of the nMOS transistor**

SOURCE: most negative terminal

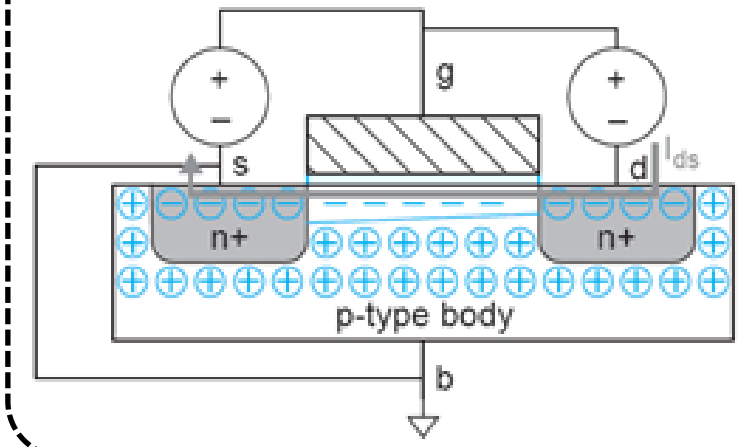
Cutoff: $V_{GS} < V_{Tn}$

$I_{DS} = 0$



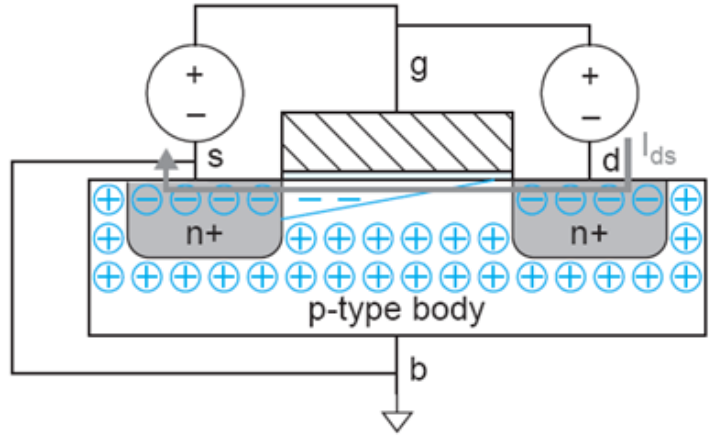
Linear: $V_{GS} > V_{Tn}$
(Resistor) $V_{DS} < V_{GS} - V_{Tn}$
$V_{GD} > V_{Tn}$

I_{DS} = linear in V_{DS}



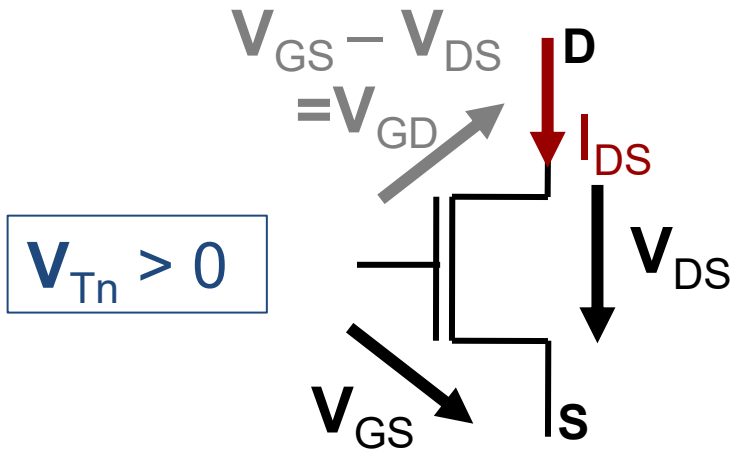
Saturation: $V_{GS} > V_{Tn}$
(Current source) $V_{DS} > V_{GS} - V_{Tn}$
$V_{GD} < V_{Tn}$

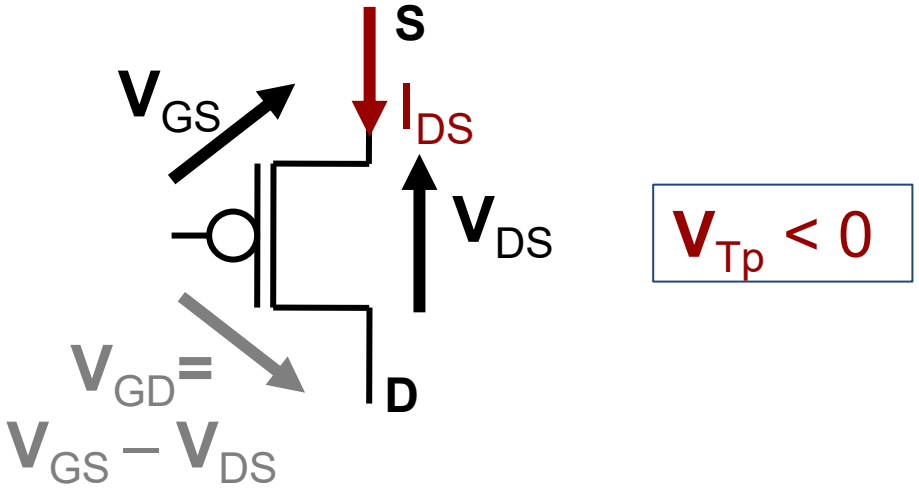
I_{DS} = ~independent of V_{DS}



pMOS vs. nMOS Operation

- **pMOS and nMOS operation are symmetric**
  - Source and drain swapped (nMOS: most negative terminal ←→ pMOS: most positive terminal)
  - All inequality conditions swapped





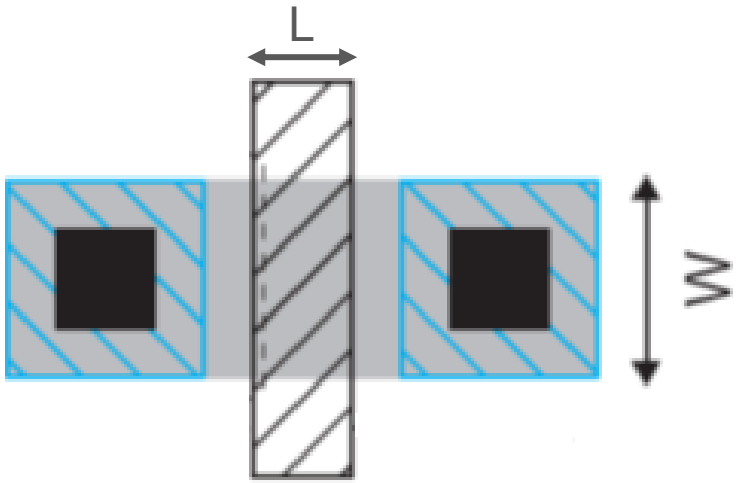
$V_{GS} < V_{Tn}$			**Cutoff**	$V_{GS} > V_{Tp}$		
$V_{GS} > V_{Tn}$	$V_{DS} < V_{GS} - V_{Tn}$	$V_{GD} > V_{Tn}$	**Linear**	$V_{GS} < V_{Tp}$	$V_{DS} > V_{GS} - V_{Tp}$	$V_{GD} < V_{Tp}$
$V_{GS} > V_{Tn}$	$V_{DS} > V_{GS} - V_{Tn}$	$V_{GD} < V_{Tn}$	**Saturation**	$V_{GS} < V_{Tp}$	$V_{DS} < V_{GS} - V_{Tp}$	$V_{GD} > V_{Tp}$

Transistor Parameter

Design parameters

L : channel length

W : channel width



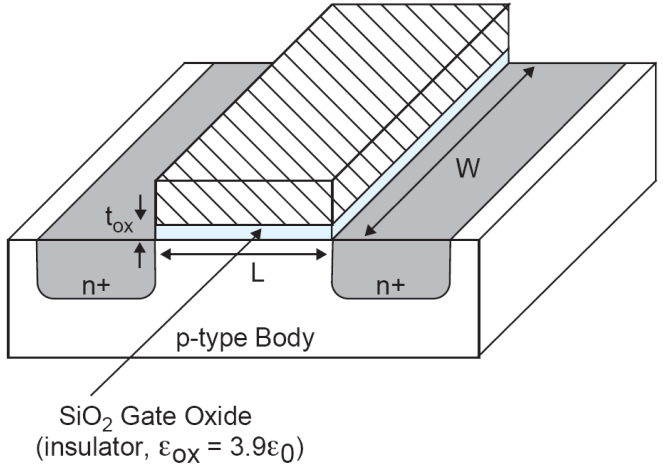
Technology parameters

V_T : Threshold voltage

C_{ox} : Oxide capacitance

μ_n, μ_p : Carrier mobility

$k'_{n/p} = \mu_{n/p} * C_{ox}$: Transconductance

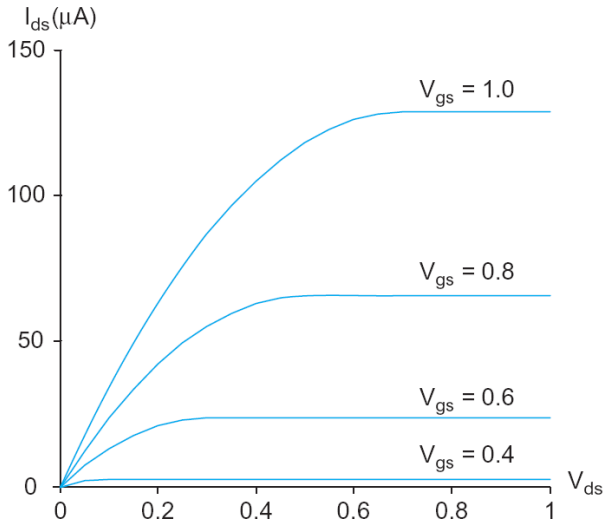


$$\beta_{n/p} = k'_{n/p} \frac{W}{L} \text{ : Drive strength}$$

Shockley 1st Order nMOS and pMOS Models

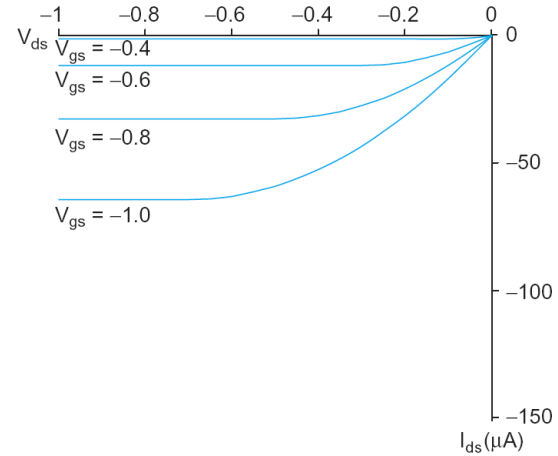
- **nMOS model:**

$$I_{DSn} = \begin{cases} 0 & V_{GS} < V_{Tn} & & & cutoff \\ \beta_n \left(V_{GS} - V_{Tn} - \frac{V_{DS}}{2}\right) V_{DS} & V_{GS} > V_{Tn} & V_{DS} < V_{GS} - V_{Tn} & V_{GD} > V_{Tn} & linear \\ \frac{\beta_n}{2}(V_{GS} - V_{Tn})^2 & V_{GS} > V_{Tn} & V_{DS} > V_{GS} - V_{Tn} & V_{GD} < V_{Tn} & saturation \end{cases}$$

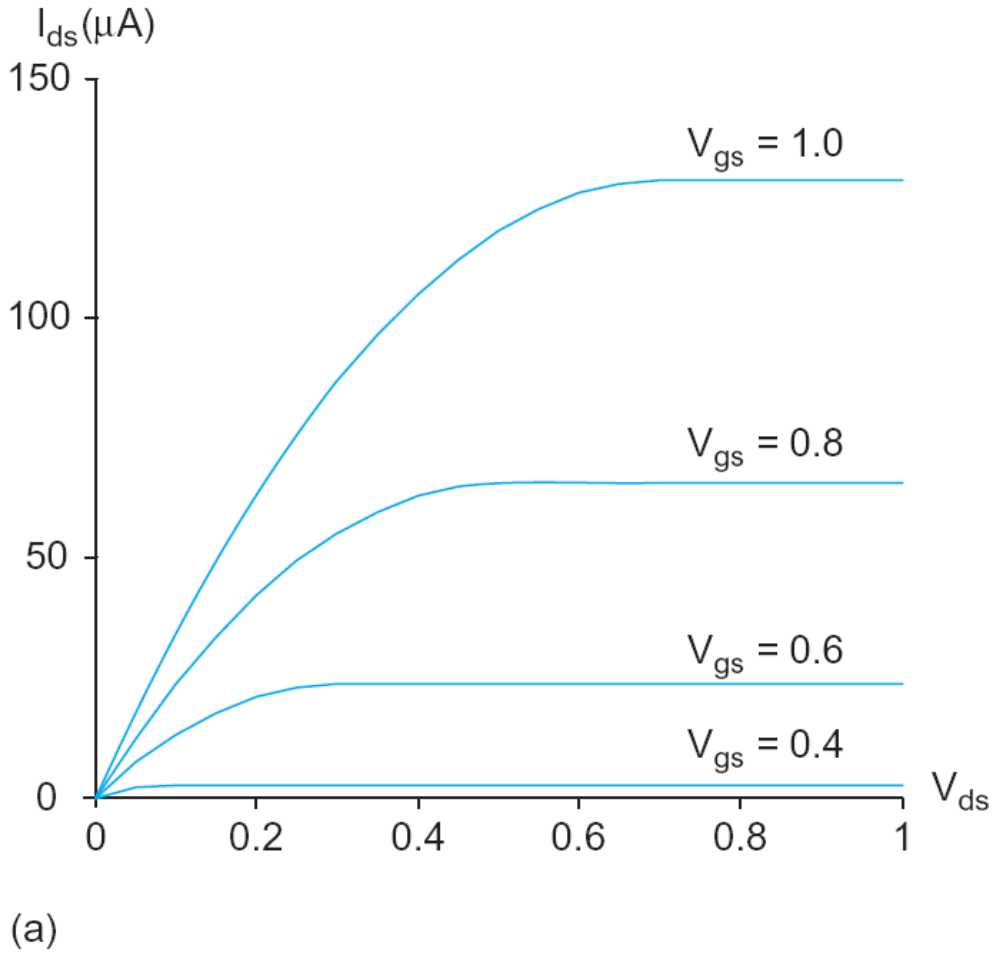


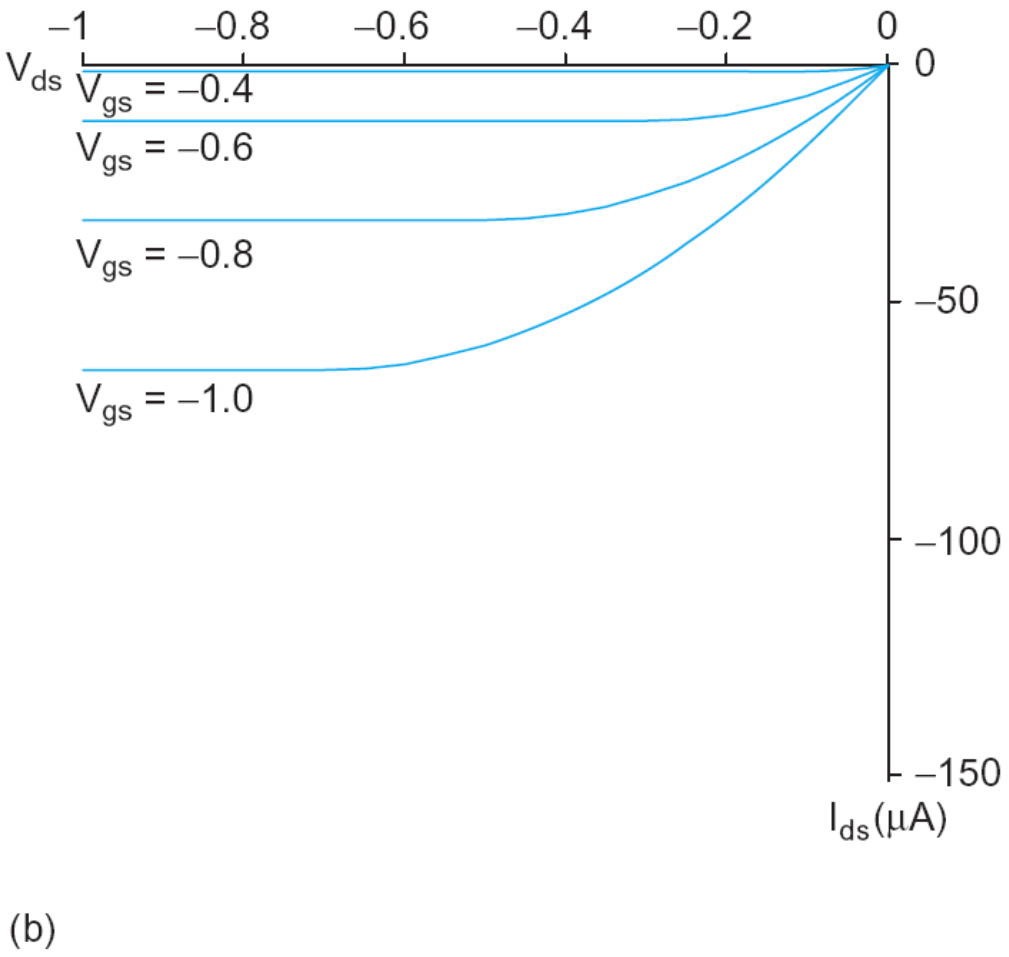
- **pMOS model:**
  - Currents are reversed (since source and drain are swapped), $I_{DSp} < 0$
  - Relation operators for operating region are swapped

$$I_{DSp} = \begin{cases} 0 & V_{GS} > V_{Tp} & & & cutoff \\ -\beta_p \left(V_{GS} - V_{Tp} - \frac{V_{DS}}{2}\right) V_{DS} & V_{GS} < V_{Tp} & V_{DS} > V_{GS} - V_{Tp} & V_{GD} < V_{Tp} & linear \\ -\frac{\beta_p}{2}(V_{GS} - V_{Tp})^2 & V_{GS} < V_{Tp} & V_{DS} < V_{GS} - V_{Tp} & V_{GD} > V_{Tp} & saturation \end{cases}$$



nMOS and pMOS I-V Characteristics

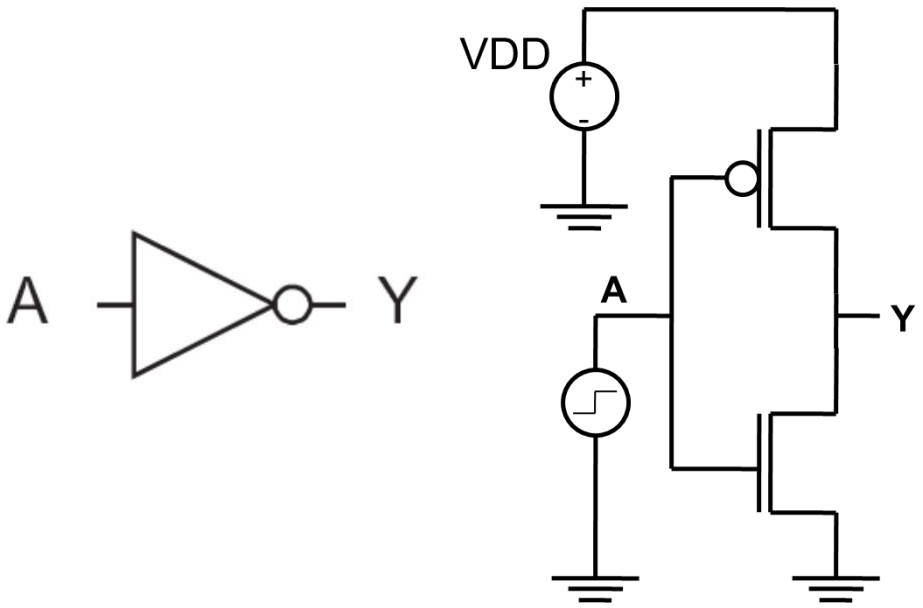




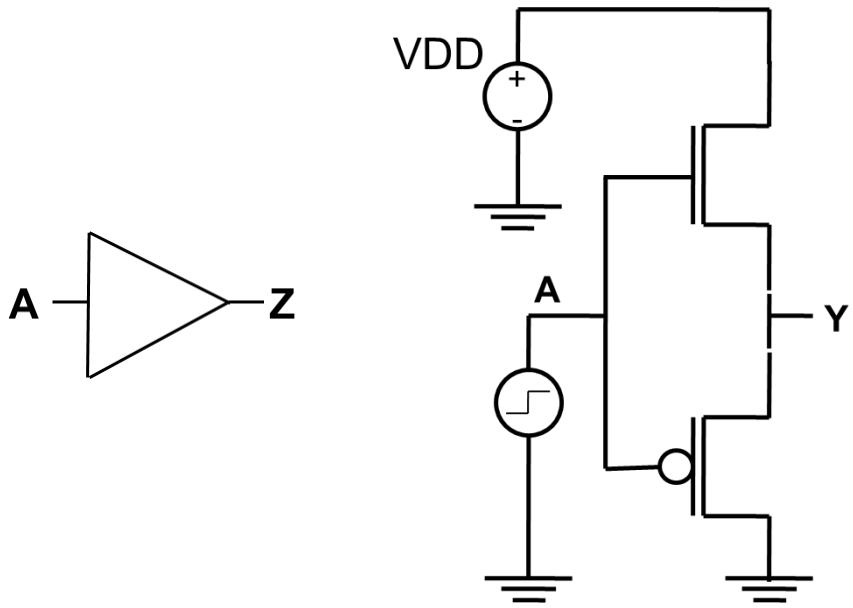
Implications of the MOS Transistor on CMOS Logic

- Realization of **PUN and PDN from nMOS and pMOS determines** if a function is
  - **Positive unate**: rising input causes rising output
  - **Negative unate:** rising input causes falling output
  - **Non unate:** rising input can cause both falling or rising output

negative unate



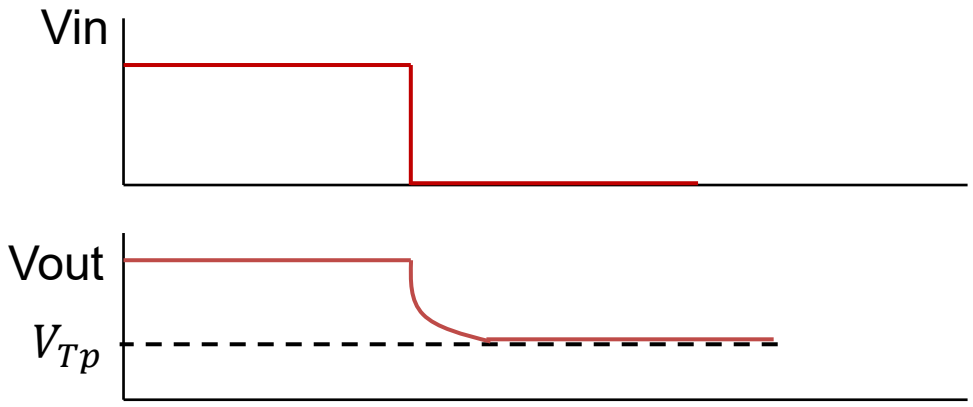
positive unate

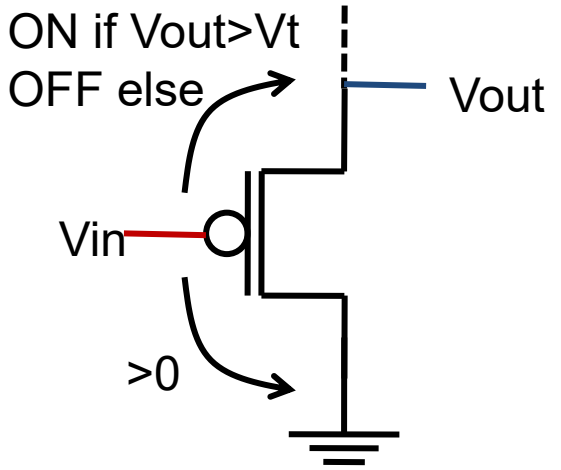


Problem of Positive Unate Functions in CMOS

- **Positive unate** (or non unate) functions **would imply** driving the output
  - **to '0' with an pMOS** or
  - **to '1' with an nMOS**

- **Unfortunately, for a pMOS**
  - A Gate-Source voltage <Vt is needed to turn on the PDN and pull output low (non-inverting)
  - PDN network turns off before output reaches GND level

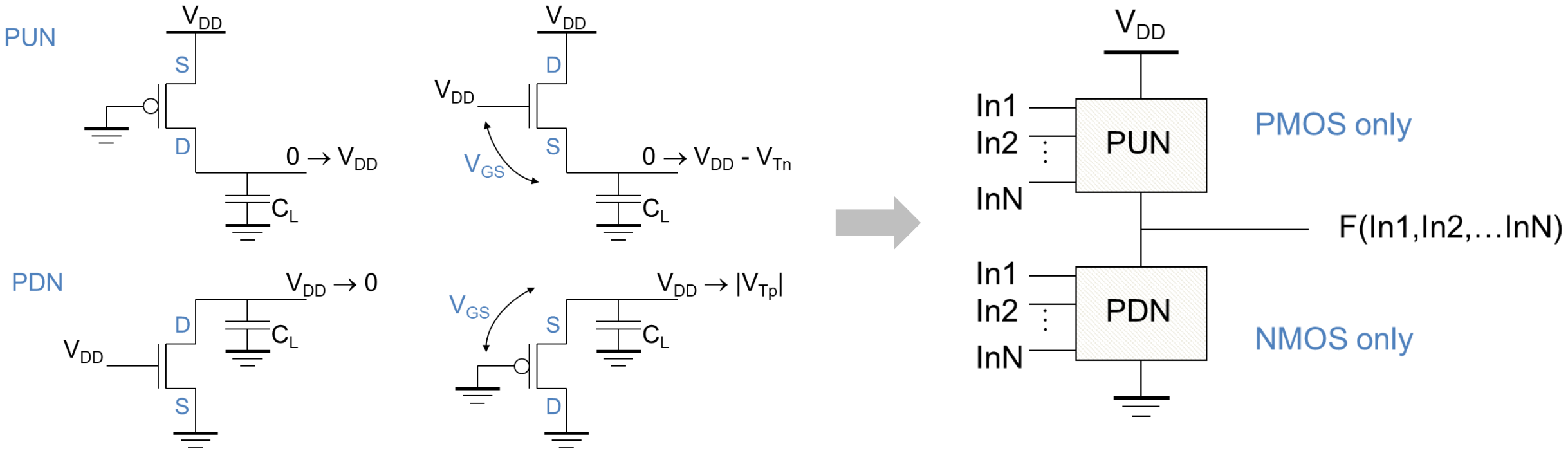




  - The nMOS shows the opposite behavior

CMOS is Limited to Negative Unate Functions

- **Driving a '0' with a pMOS or a '1' with an nMOS prevents rail-to-rail outputs**
  - pMOS is a good driver for a '1'
  - nMOS is a good driver for a '0'



EE-429
Fundamentals of VLSI Design

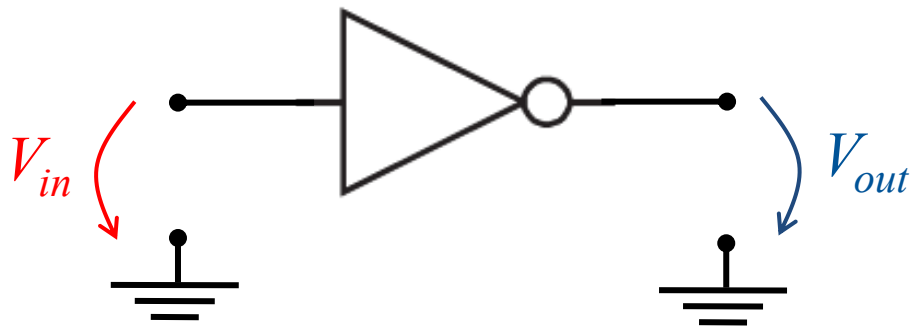
Inverter Voltage Transfer and DC Characteristic

Andreas Burg

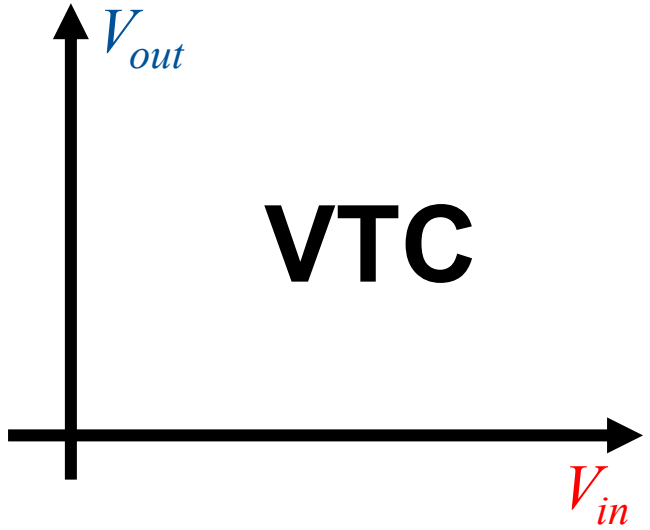
CMOS Inverter VTC

- **Voltage Transfer Characteristic**
  - Input-output relationship of a component (e.g., digital gate)
  - DC characteristic (i.e., not immediately useful for dynamic considerations – no time aspect)

- **The VTC provides important information on**
  - Useful input range and achievable output range
  - Voltage gain and sensitivity
  - Voltage margins



$$V_{out} = f(V_{in})$$



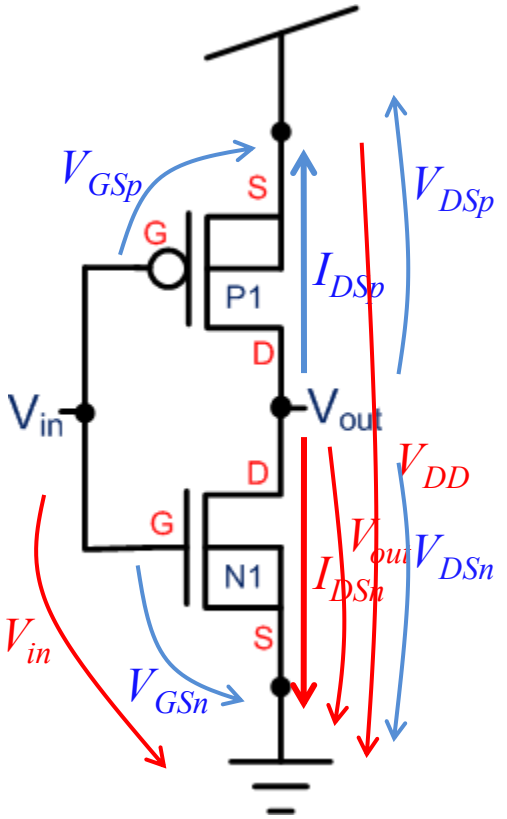
CMOS Inverter VTC

- **To construct the VTC of the CMOS inverter, we need to graphically superimpose the I-V curves of the nMOS and pMOS onto a common coordinate set.**
- **Basic idea:** equal pMOS and nMOS current

$$I_{DSp} = -I_{DSn}$$

$$V_{GSn} = V_{in} \qquad V_{GSp} = V_{in} - V_{DD}$$

$$V_{DSn} = V_{out} \qquad V_{DSp} = V_{out} - V_{DD}$$

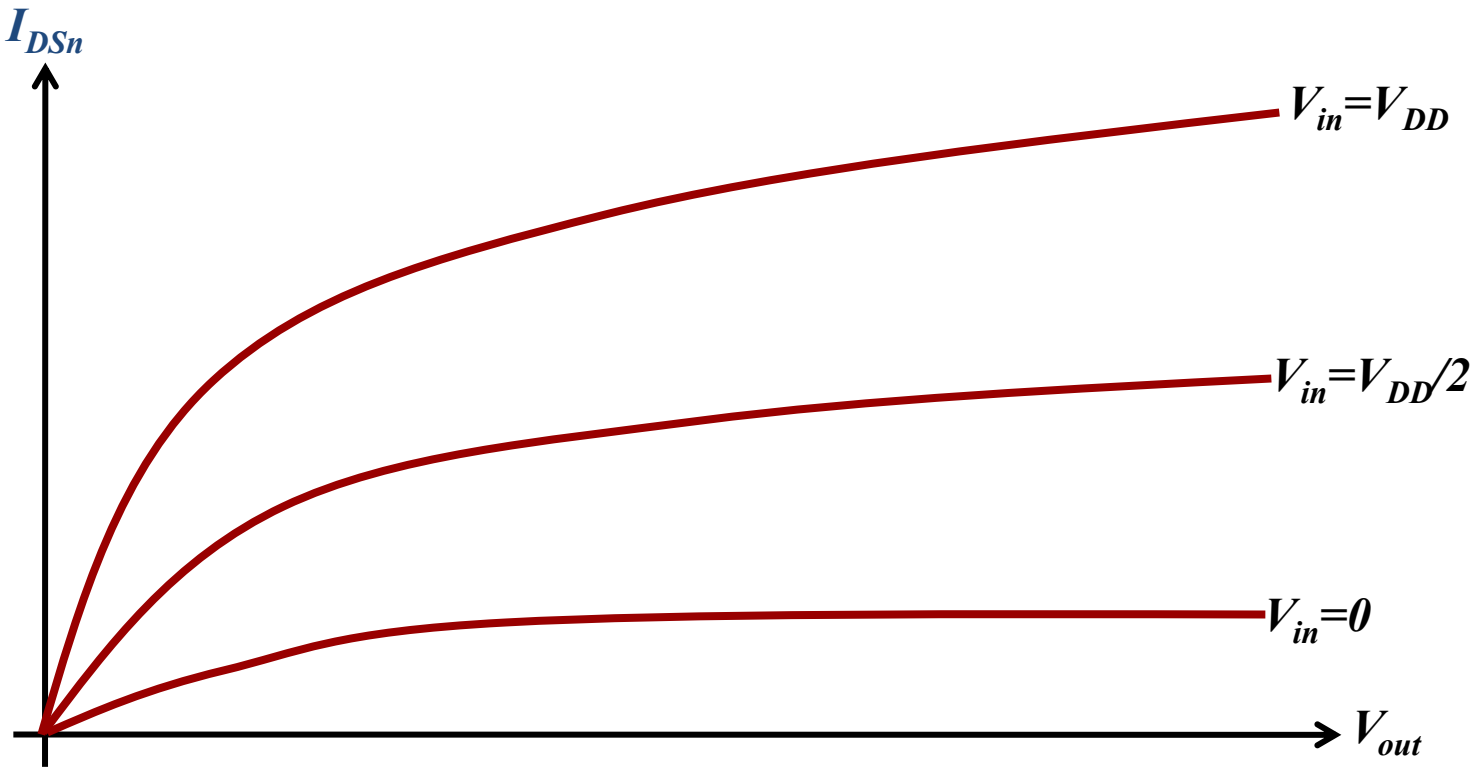


CMOS Inverter VTC

- **Graphical approach:** transfer I-V curves of nMOS and pMOS into the same coordinate system
  - Start with the I-V characteristics of the nMOS transistor, parameterized on V_{in}

$$V_{out} = V_{DSn}$$

$$V_{in} = V_{GSn}$$



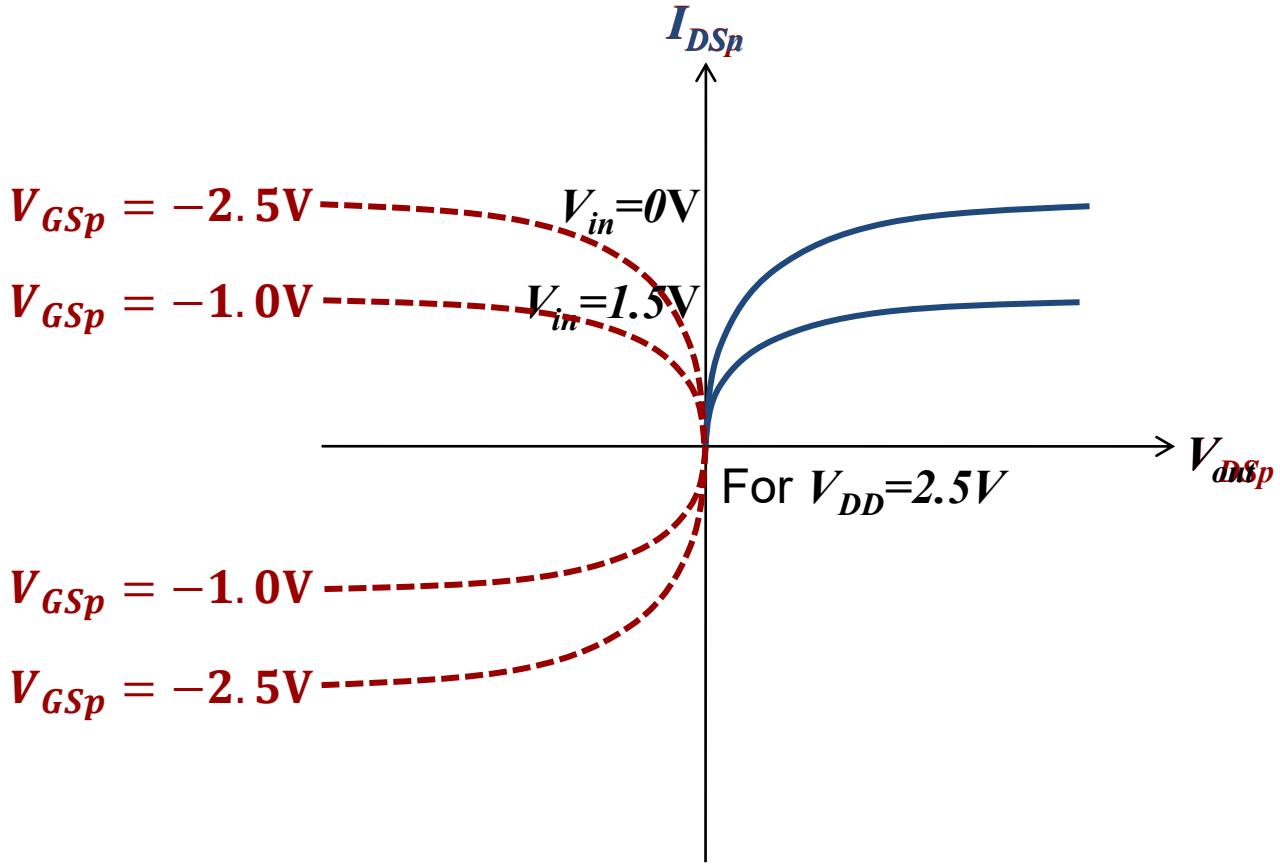
CMOS Inverter VTC

- **Graphical approach:** transfer I-V curves of nMOS and pMOS into the same coordinate system
  - More complex for pMOS since V_{in}, V_{out}, and I_{DSn} are only indirectly related to the pMOS I-V

$$I_{DSn} = -I_{DSp}$$

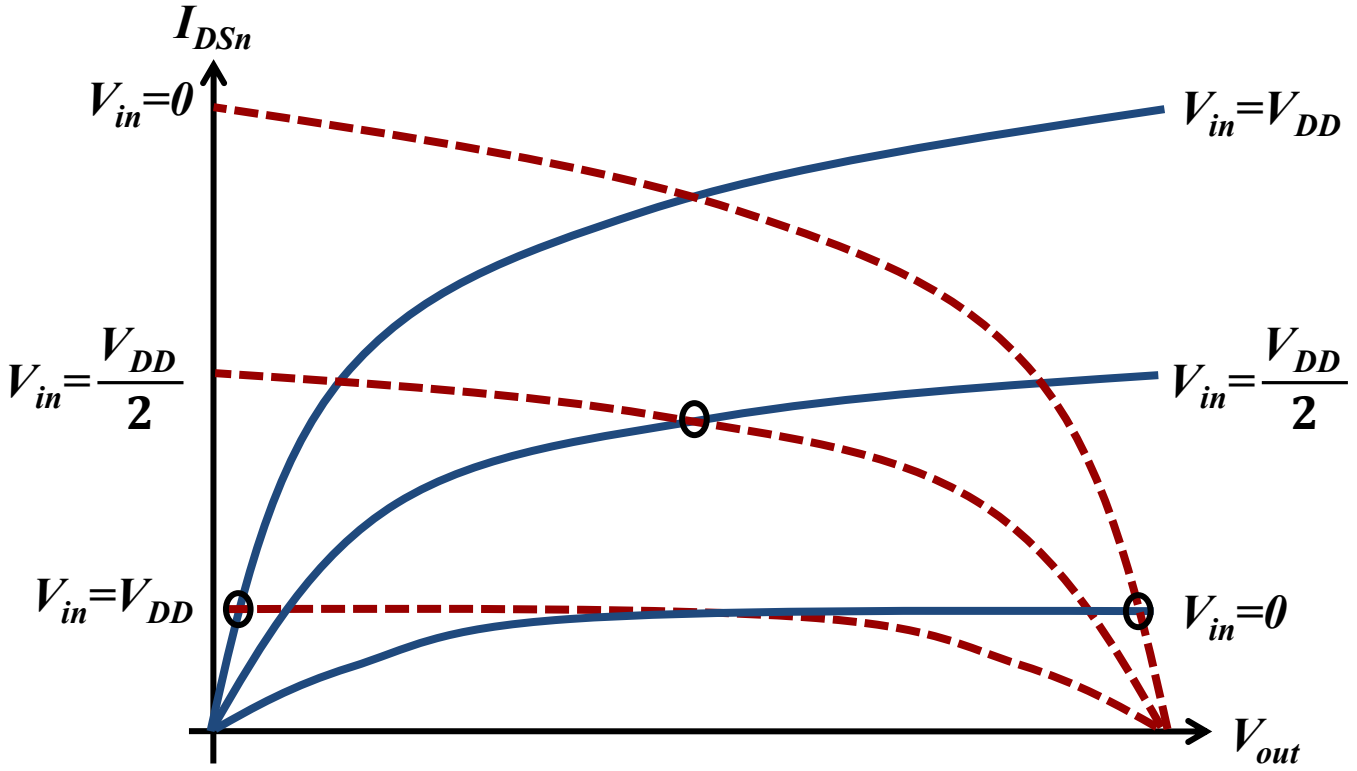
$$V_{out} = V_{DD} + V_{SDp}$$

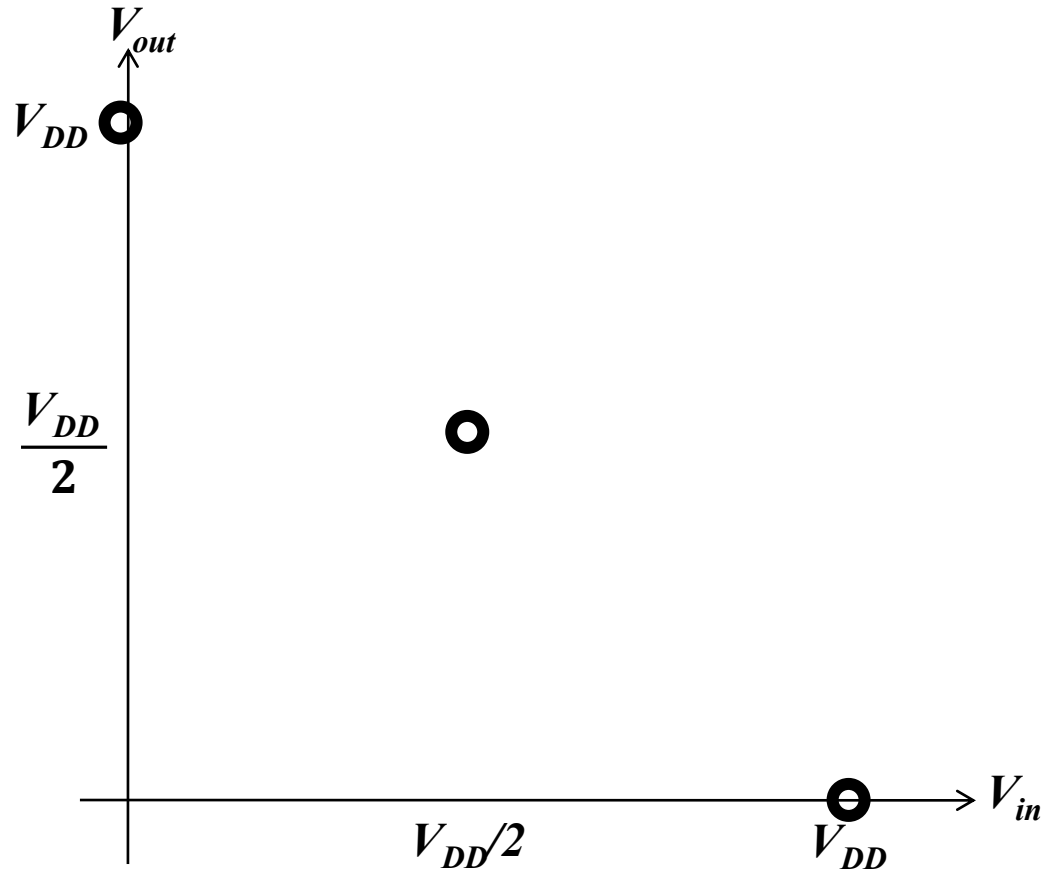
$$V_{in} = V_{DD} + V_{GSp}$$



CMOS Inverter VTC

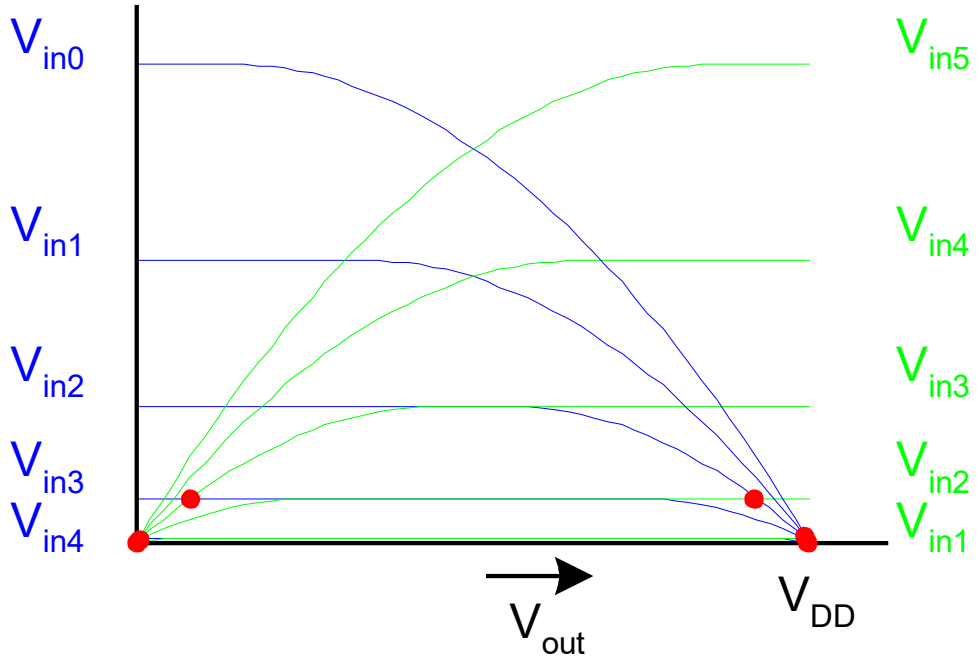
- The **intersection of corresponding load lines**, where the currents of the nMOS and pMOS are equal, **shows the DC operating points.**
  - Putting all the intersection points on a graph with the corresponding output voltage will give us the CMOS inverter's VTC

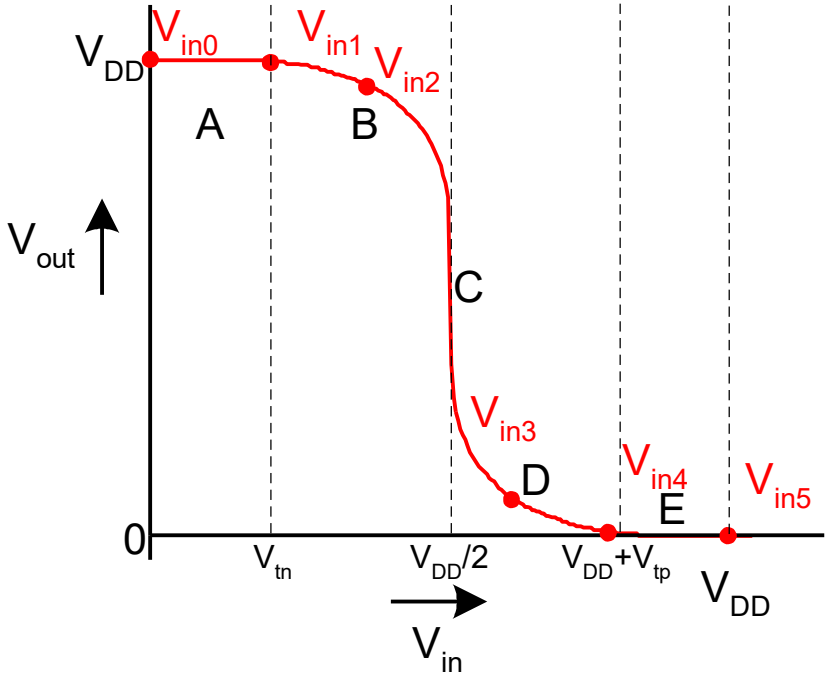




CMOS Inverter VTC

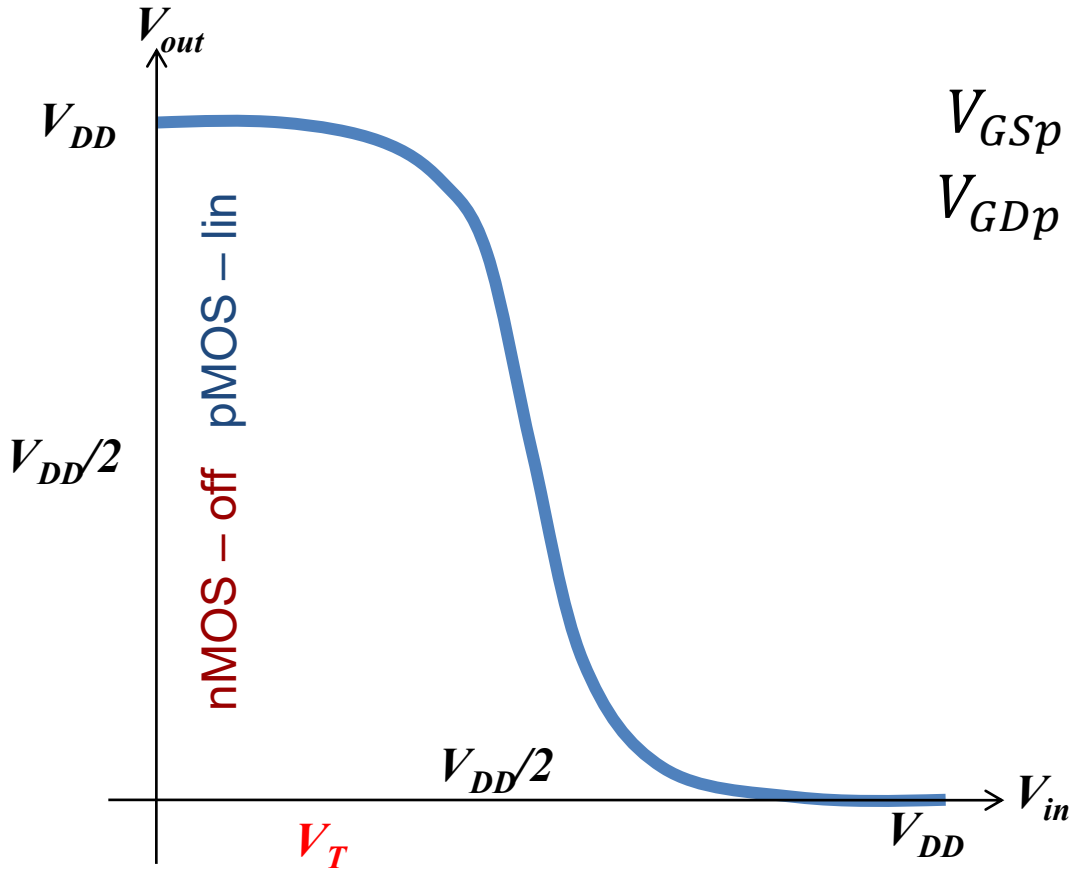
- The **intersection of corresponding load lines**, where the currents of the nMOS and pMOS are equal, **shows the DC operating points.**
  - Putting all the intersection points on a graph with the corresponding output voltage will give us the CMOS inverter's VTC
- **Different points on the VTC fall into different n/pMOS operating regions**





CMOS Inverter VTC: Operating Regions

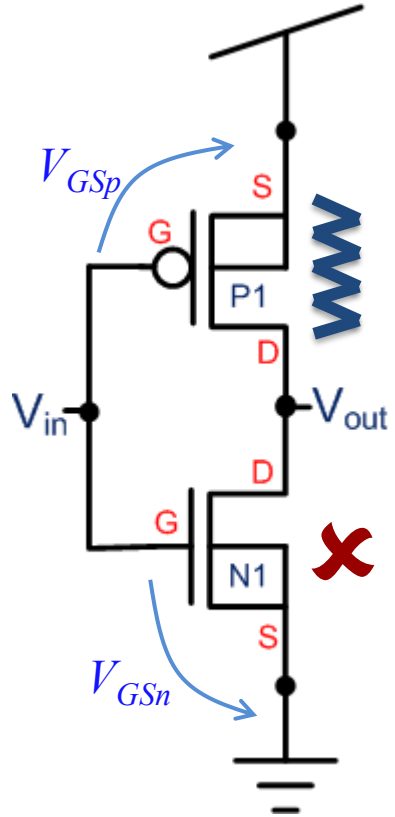
- Analytical analysis requires **different equations for each operating region** due to the non-linear model



$$V_{in} < V_{Tn}$$

$$\left.\begin{aligned} V_{GSp} &= V_{in} - V_{DD} < V_{Tp} \\ V_{GDp} &= V_{in} - V_{out} < V_{Tp} \end{aligned}\right\} \rightarrow Linear$$

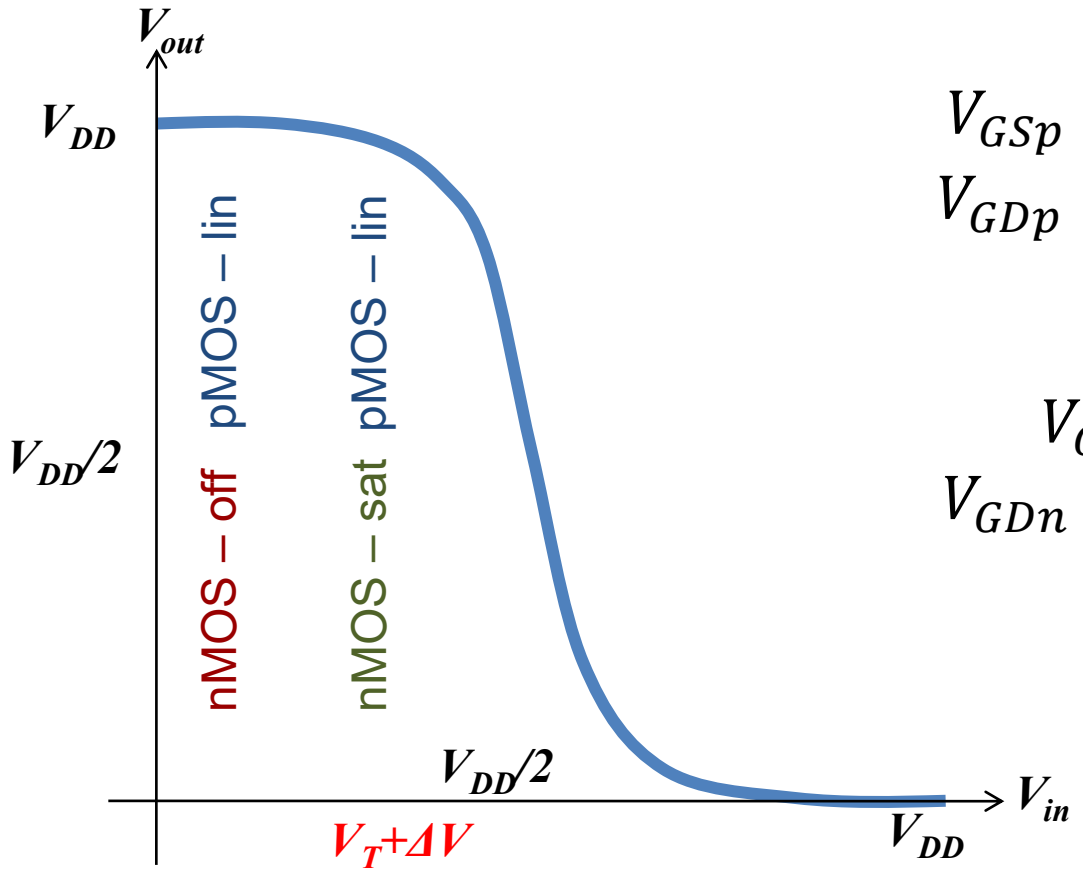
$$V_{GSn} = V_{in} < V_T \rightarrow Cutoff$$



Considering Long Channel Transistors with $V_T < V_{DD}/2$

CMOS Inverter VTC: Operating Regions

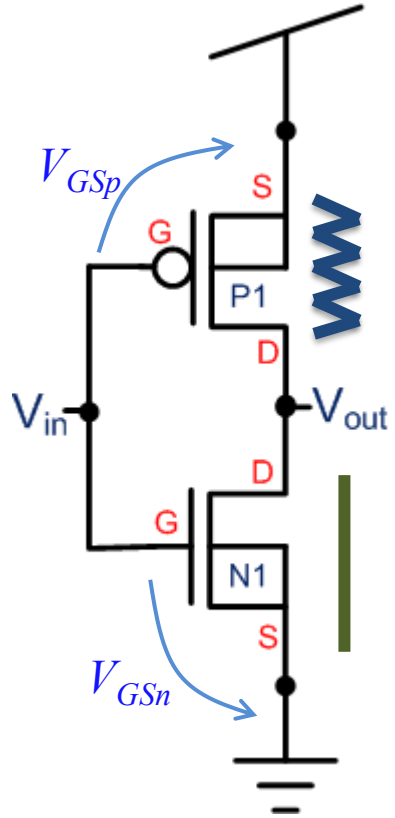
- Analytical analysis requires **different equations for each operating region** due to the non-linear model



$$V_{in} > V_{Tn}$$

$$\left.\begin{aligned} V_{GSp} &= V_{in} - V_{DD} < V_{Tp} \\ V_{GDp} &= V_{in} - V_{out} < V_{Tp} \end{aligned}\right\} \rightarrow Linear$$

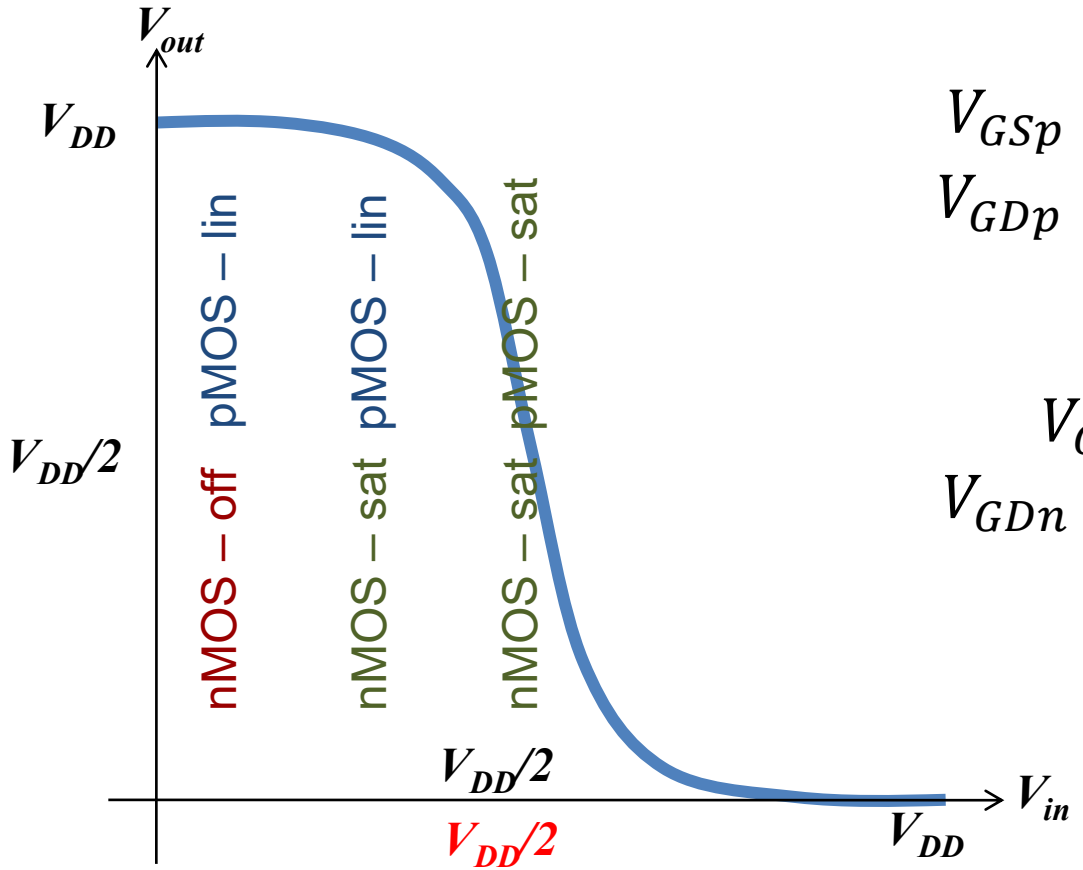
$$\left.\begin{aligned} V_{GSn} &= V_{in} > V_{Tn} \\ V_{GDn} &= V_{in} - V_{out} < V_{Tn} \end{aligned}\right\} \rightarrow Sat$$



Considering Long Channel Transistors with $V_T<V_{DD}/2$

CMOS Inverter VTC: Operating Regions

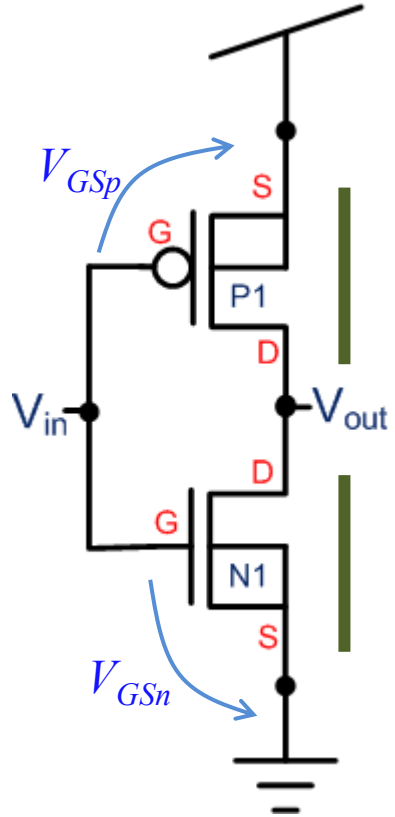
- Analytical analysis requires **different equations for each operating region** due to the non-linear model



$$V_{in} > V_{DD}/2$$

$$\left.\begin{aligned} V_{GSp} &= V_{in} - V_{DD} < V_{Tp} \\ V_{GDp} &= V_{in} - V_{out} > V_{Tp} \end{aligned}\right\} \rightarrow Sat$$

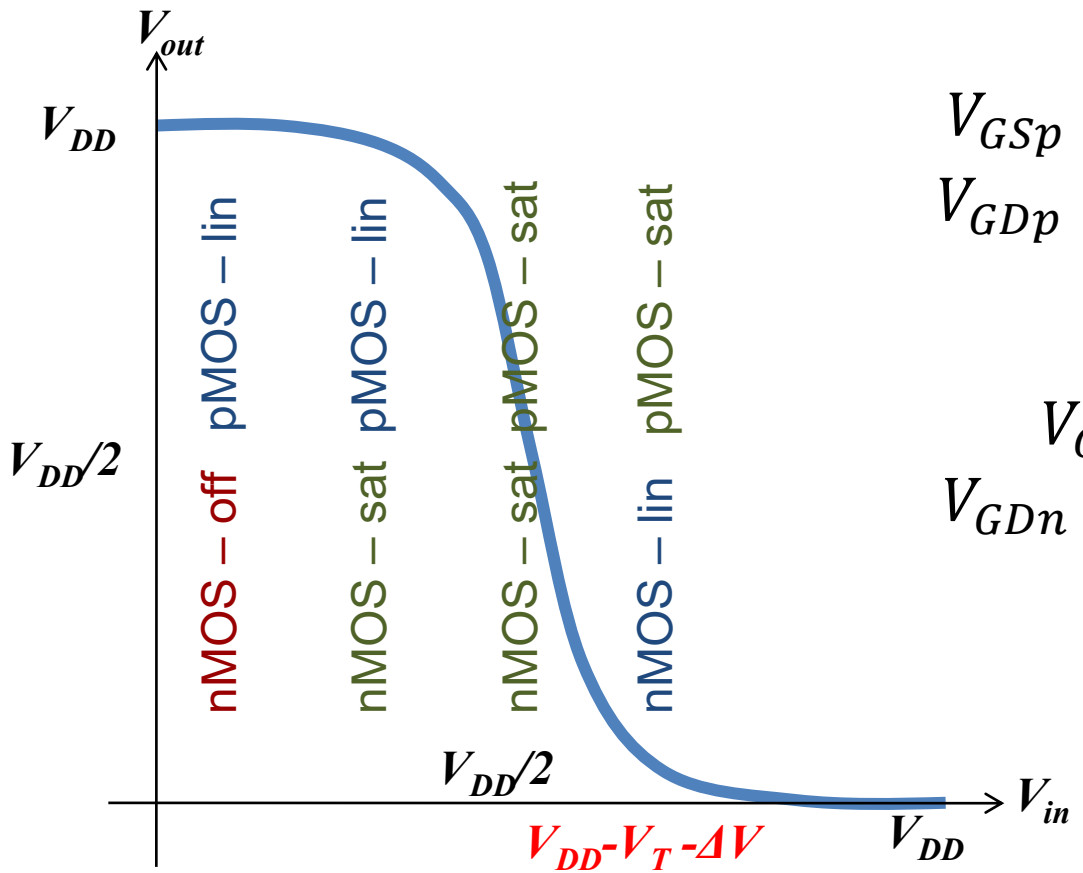
$$\left.\begin{aligned} V_{GSn} &= V_{in} > V_{Tn} \\ V_{GDn} &= V_{in} - V_{out} < V_{Tn} \end{aligned}\right\} \rightarrow Sat$$



Considering Long Channel Transistors with $V_T < V_{DD}/2$

CMOS Inverter VTC: Operating Regions

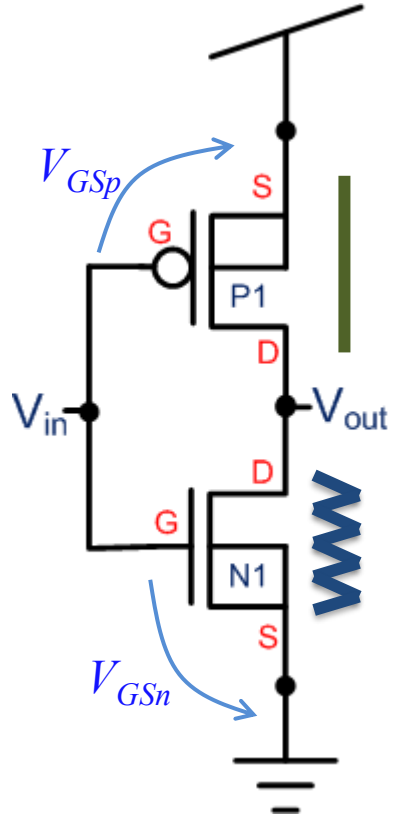
- Analytical analysis requires **different equations for each operating region** due to the non-linear model



$$V_{in} < V_{DD} - V_{Tp}$$

$$\left.\begin{aligned} V_{GSp} &= V_{in} - V_{DD} < V_{Tp} \\ V_{GDp} &= V_{in} - V_{out} > V_{Tp} \end{aligned}\right\} \rightarrow Sat$$

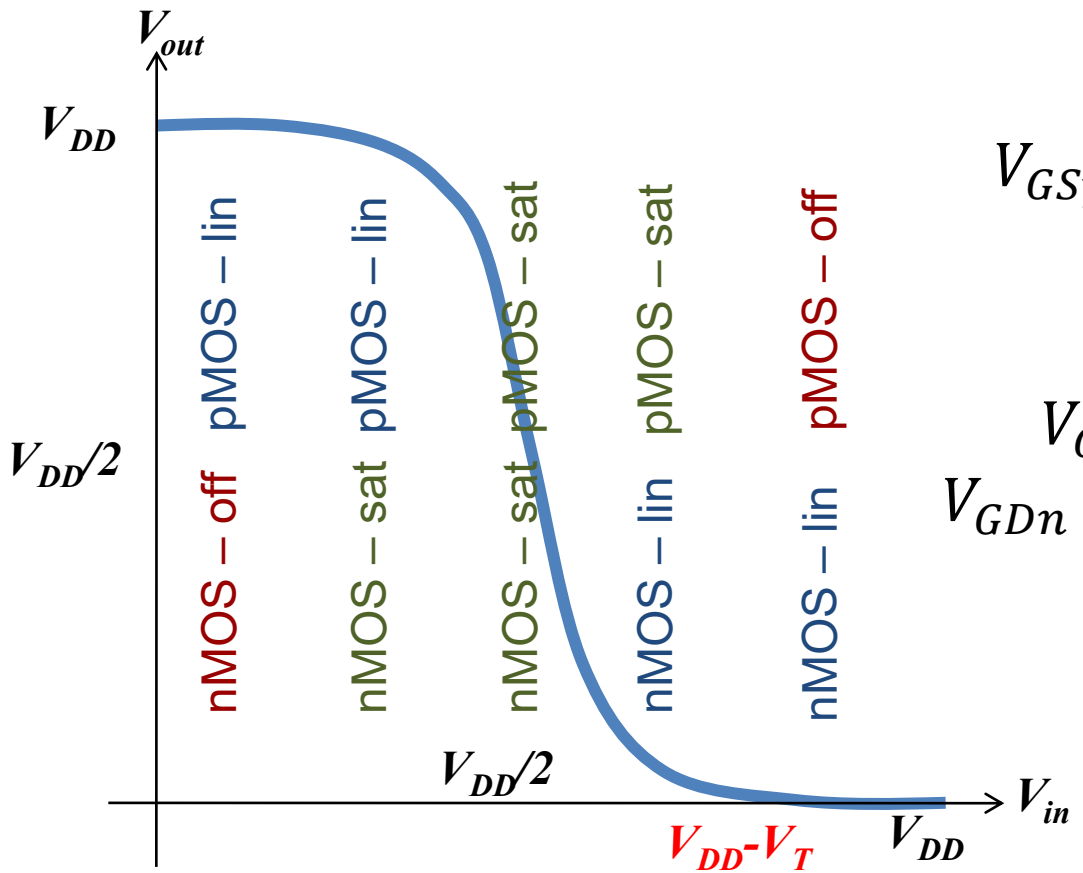
$$\left.\begin{aligned} V_{GSn} &= V_{in} > V_{Tn} \\ V_{GDn} &= V_{in} - V_{out} > V_{Tn} \end{aligned}\right\} \rightarrow Linear$$



Considering Long Channel Transistors with $V_T < V_{DD}/2$

CMOS Inverter VTC: Operating Regions

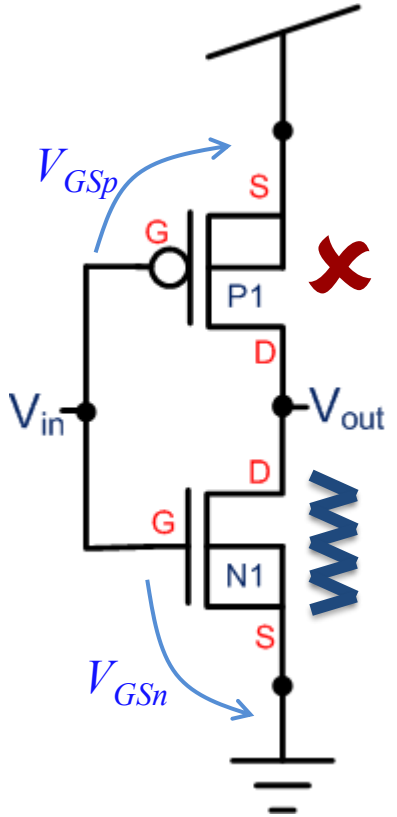
- Analytical analysis requires **different equations for each operating region** due to the non-linear model



$$V_{in} > V_{DD} - V_{Tp}$$

$$V_{GSp} = V_{in} - V_{DD} > V_{Tp} \rightarrow Cutoff$$

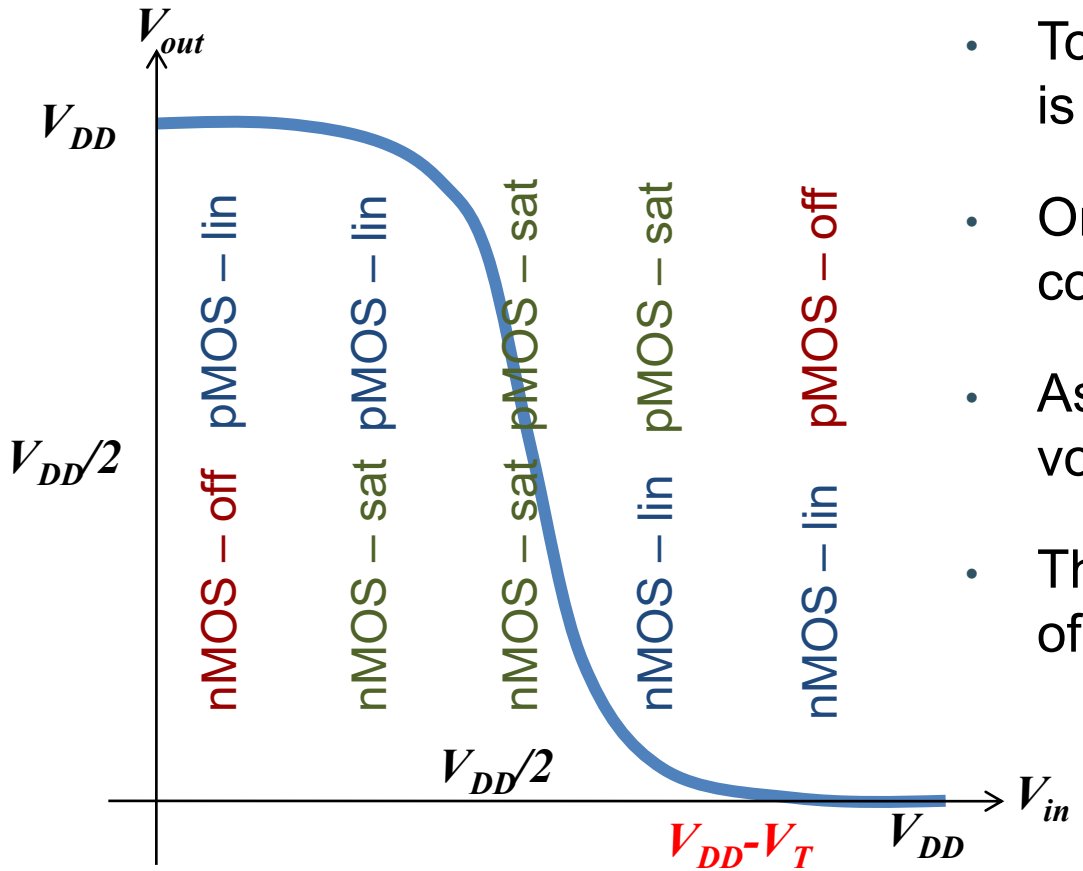
$$\left.\begin{aligned} V_{GSn} = V_{in} > V_{Tn} \\ V_{GDn} = V_{in} - V_{out} > V_{Tn} \end{aligned}\right\} \rightarrow Linear$$



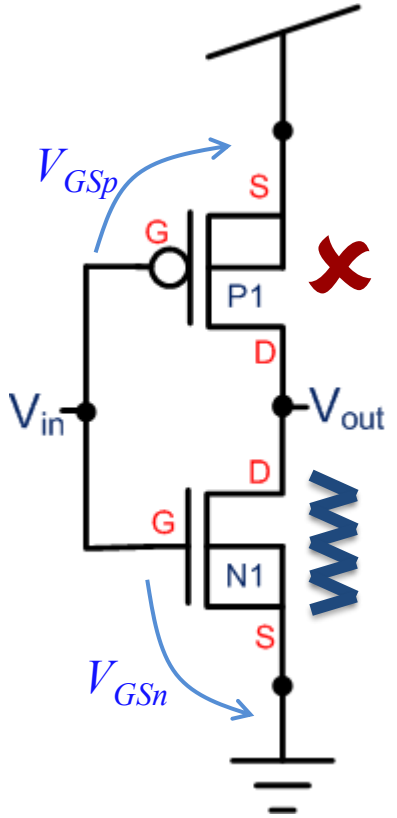
Considering Long Channel Transistors with $V_T < V_{DD}/2$

CMOS Inverter VTC: Operating Regions

- Analytical analysis requires **different equations for each operating region** due to the non-linear model



- Towards the rails, one of the transistors is cut off, and the other is resistive.
- Once the cut off transistor starts conducting, it immediately is saturated.
- As we approach the middle input voltages, both transistors are saturated.
- The VTC slope is known as the Gain of the gate.



Considering Long Channel Transistors with $V_T<V_{DD}/2$

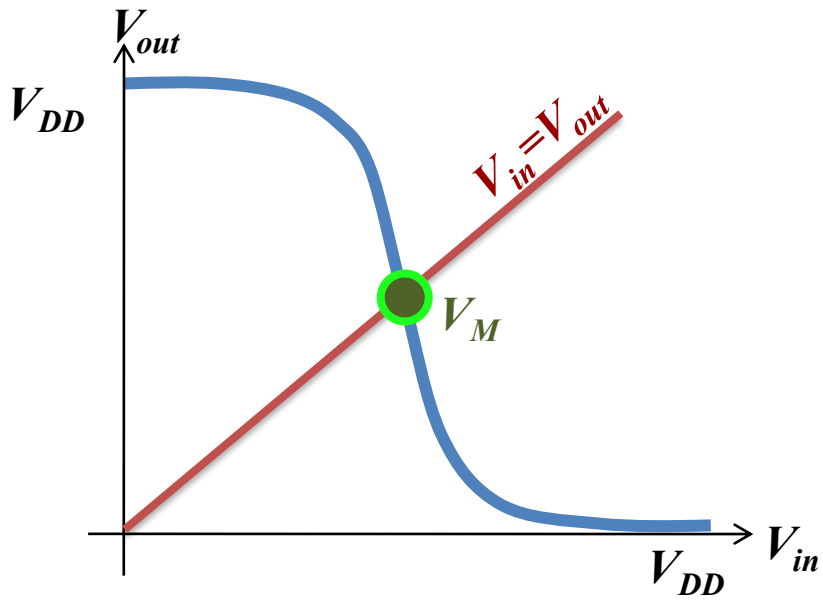
EE-429
Fundamentals of VLSI Design

Inverter DC Characteristic

Andreas Burg

Switching Threshold

- **The Switching Threshold, V_M, is the point where $V_{in}=V_{out}$**
  - Inverter has the largest gain at this point
  - Knowing V_M is desired to define threshold for delay measurements
  - Skewing V_M skews the delay for rising or falling
  - Placing V_M at $V_{DD}/2$ maximized noise margins

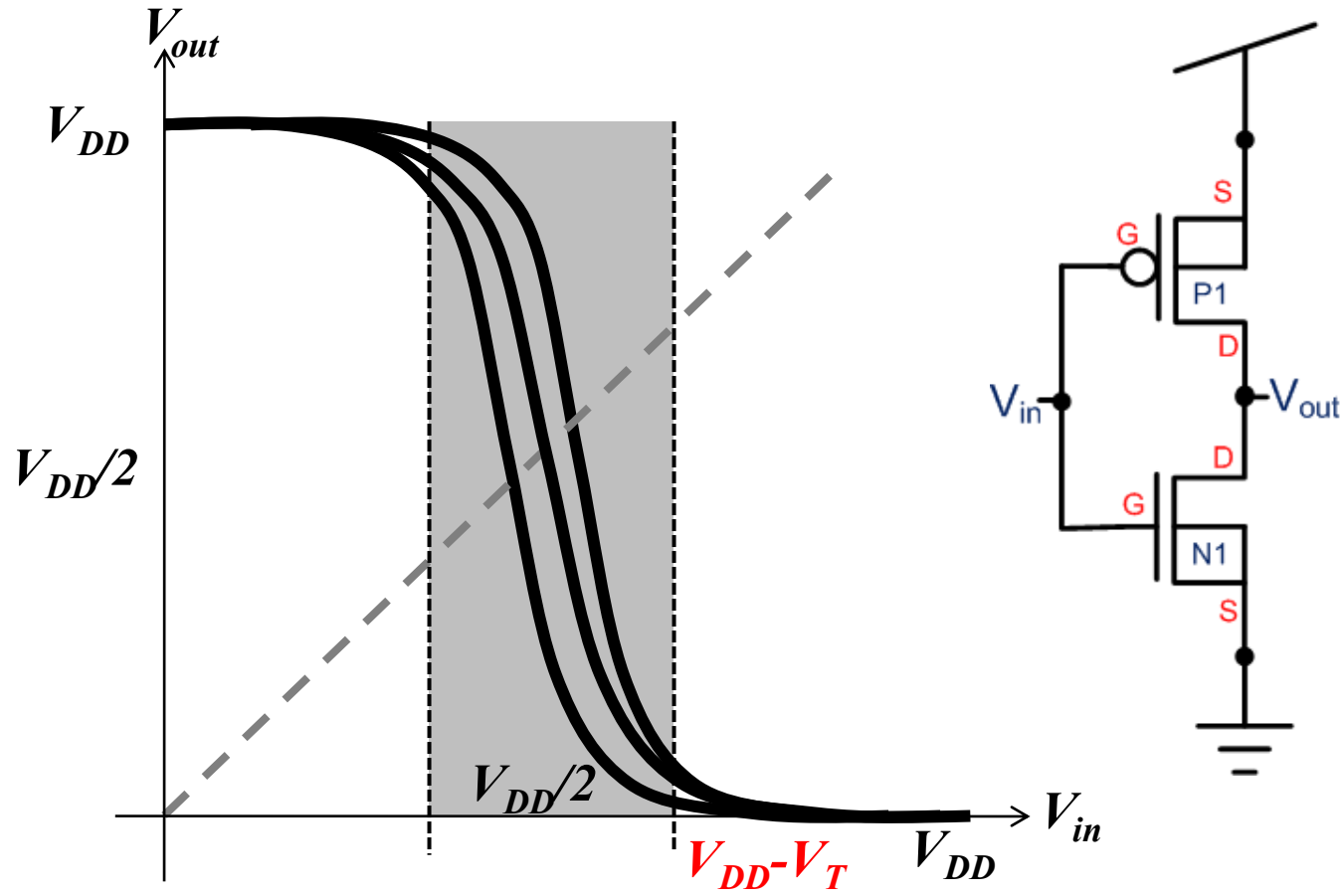


- **Graphical calculation: intersection of the VTC with $V_{in}=V_{out}$**

- To **understand the impact of design parameters** on **V_M** we are interested in an **analytical solution**

Switching Threshold

- **Analytical computation of V_M** : equating the currents through pMOS and nMOS
  - Switching threshold **V_M lies around the** point of **largest gain** (center of the VTC)



Switching Threshold

- **Analytical computation of V_M** : equating the currents through pMOS and nMOS
  - Switching threshold **V_M lies around the** point of **largest gain** (center of the VTC)
  - Both pMOS and nMOS in saturation

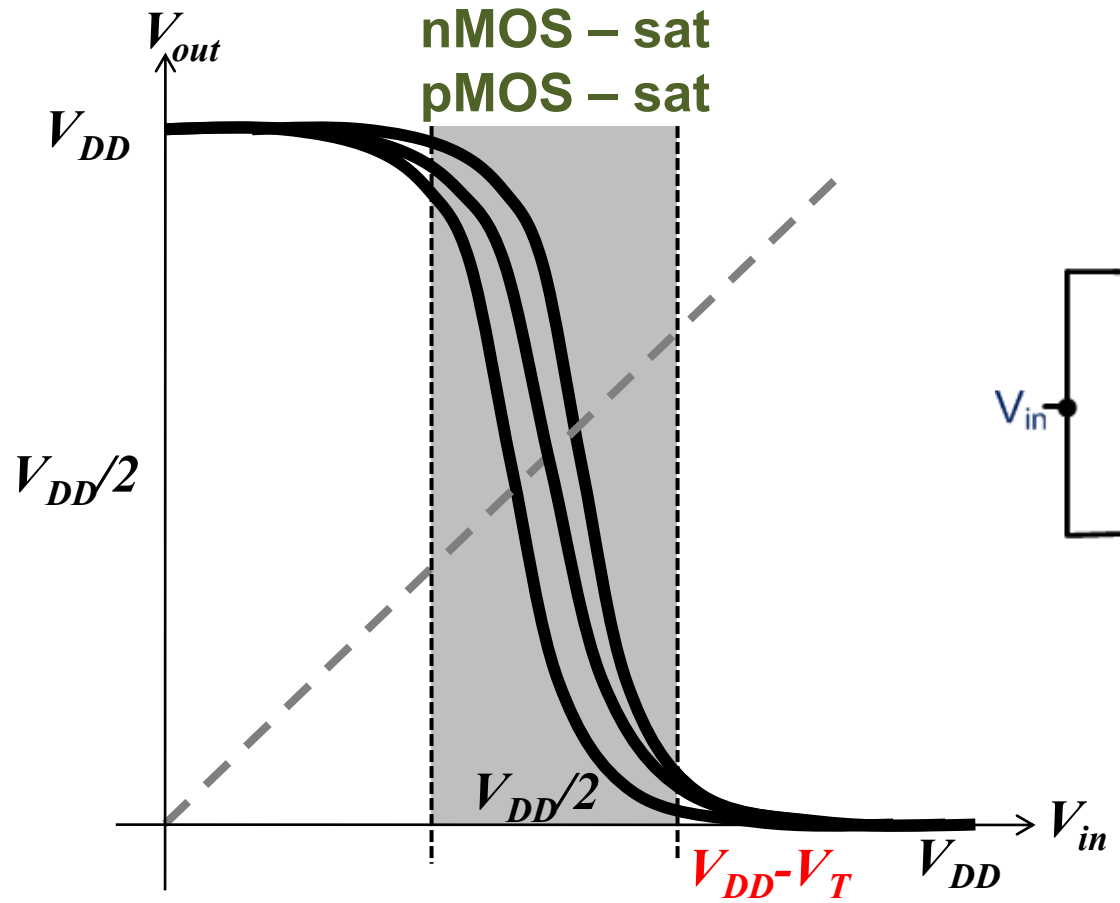
$$I_{DSn} = \frac{\beta_n}{2}(V_{in} - V_{Tn})^2$$

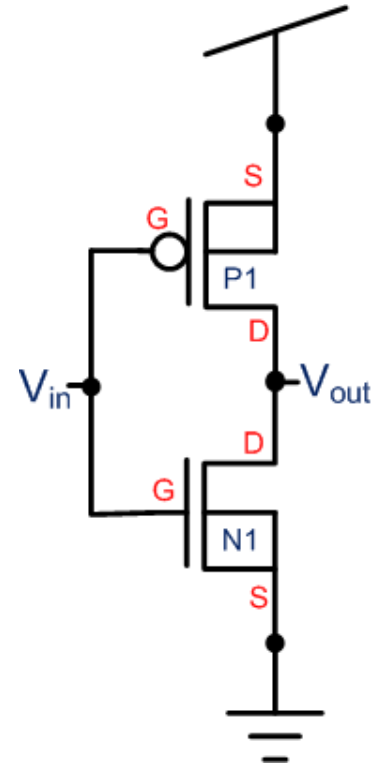
$$I_{DSp} = -\frac{\beta_p}{2}(V_{in} - V_{DD} - V_{Tp})^2$$

$$I_{DSn} = -I_{DSp}$$

  - Threshold defined as **V_{in}=V_M**

$$\frac{\beta_n}{2}(V_M - V_{Tn})^2 = \frac{\beta_p}{2}(V_M - V_{DD} - V_{Tp})^2$$





Switching Threshold

- **Analytical computation of V_M**
  - Switching threshold lies around the point of largest gain (center of the VTC)
  - Both pMOS and nMOS in saturation

$$I_{DSn} = \frac{\beta_n}{2}(V_{in} - V_{Tn})^2$$

$$I_{DSp} = -\frac{\beta_p}{2}(V_{in} - V_{DD} - V_{Tp})^2$$

$$I_{DSn} = -I_{DSp}$$

  - Threshold defined as $\mathbf{V_{in}}$=$\mathbf{V_M}$

$$\frac{\beta_n}{2}(V_M - V_{Tn})^2 = \frac{\beta_p}{2}(V_M - V_{DD} - V_{Tp})^2$$

Solving for V_M

$$V_M = \frac{V_{Tn} + r(V_{DD} + V_{Tp})}{1 + r}$$

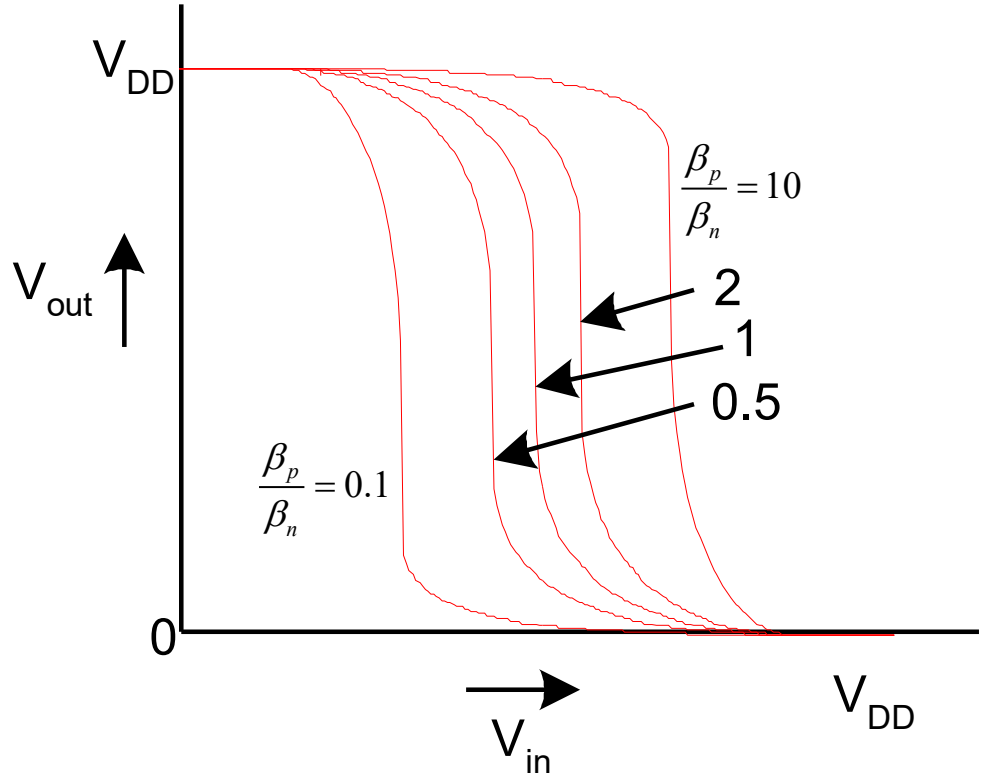
$$r = \sqrt{\frac{\beta_p}{\beta_n}}$$

Switching Threshold

- ***r* is an important factor** in setting the switching threshold

- ***r*** is **set by** the **drive strength ratio** of nMOS and pMOS

$$r = \sqrt{\frac{\beta_p}{\beta_n}} \qquad \beta_{n/p} = \mu_{n/p} C_{ox} \frac{W_{n/p}}{L_{n/p}}$$



- **Using the current equations again, we can find the drive strength ratio for a desired V_M:**
  - $\mu_{n/p}$ are technology parameters
  - $W_{n/p}$ and $L_{n/p}$ are controlled by the designer

$$\frac{\beta_p}{\beta_n} = \left(\frac{V_M - V_{Tn}}{V_{DD} - V_M + V_{Tp}} \right)^2$$

Switching Threshold Optimization

- A **symmetric VTC (V_M=V_{DD}/2)** is often desired.
  - In practice, **we often find** that $\mu_n > \mu_p$ and hence

- **Adjust the switching threshold with W and L** of nMOS and pMOS
  - Consider $V_{Tn} = -V_{Tp} = V_T$

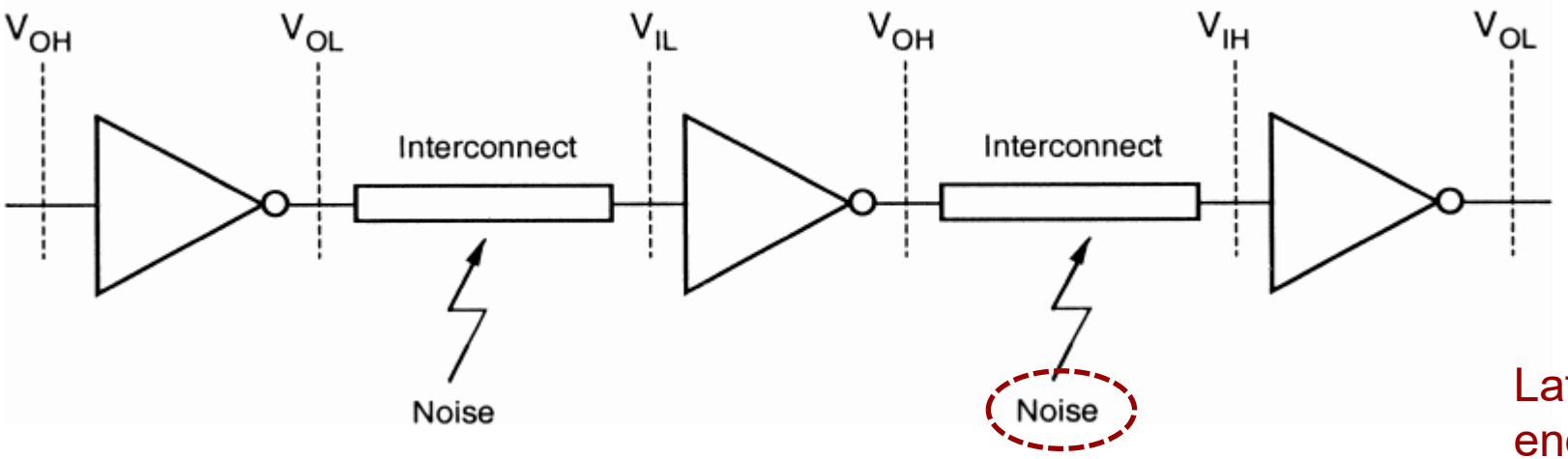
$$\frac{\beta_p}{\beta_n} = \left(\frac{\frac{V_{DD}}{2} - V_T}{V_{DD} - \frac{V_{DD}}{2} - V_T} \right)^2 = 1 \longrightarrow \frac{W_p}{L_p} = \frac{\mu_n}{\mu_p} \left(\frac{W_n}{L_n} \right)$$

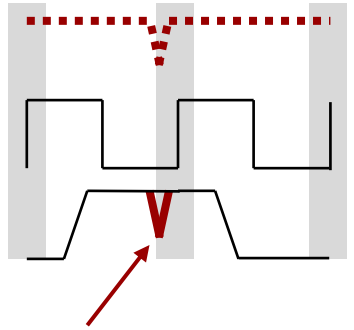
- For **digital circuits**, we often **use minimum length (L_{min})** to maximize drive
  - **Adjust** only the **width of the weaker pMOS** to compensate the drive mismatch

$$\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p} \approx 2 \ldots 4 \qquad W_p = W_n \frac{\mu_n}{\mu_p}$$

Impact of Noise on Synchronous Digital Circuits

- In practice, **digital circuits are affected by noise** (e.g., coupling or supply noise)
  - Other non-idealities affecting the output levels of a gate may also be understood as noise





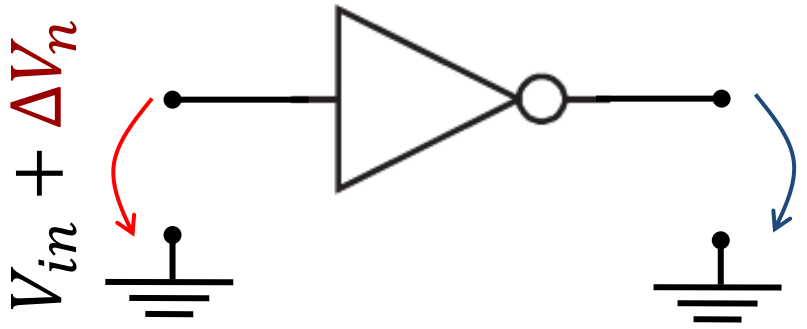
Late glitch induced at the end of a cycle by noise

- **Noise may increase** the **delay OR may cause failures in synchronous systems** due to transitions from noise late in the clock cycle

- **Reducing** the impact and **propagation of noise** improves stability

Propagation of Noise

- **Inject noise at the input** of a gate.
- **How does** this **noise propagate** to the output of the gate?



$$V_{out} = f(V_{in}) + \frac{\partial V_{out}}{\partial V_{in}} \Delta V_n$$

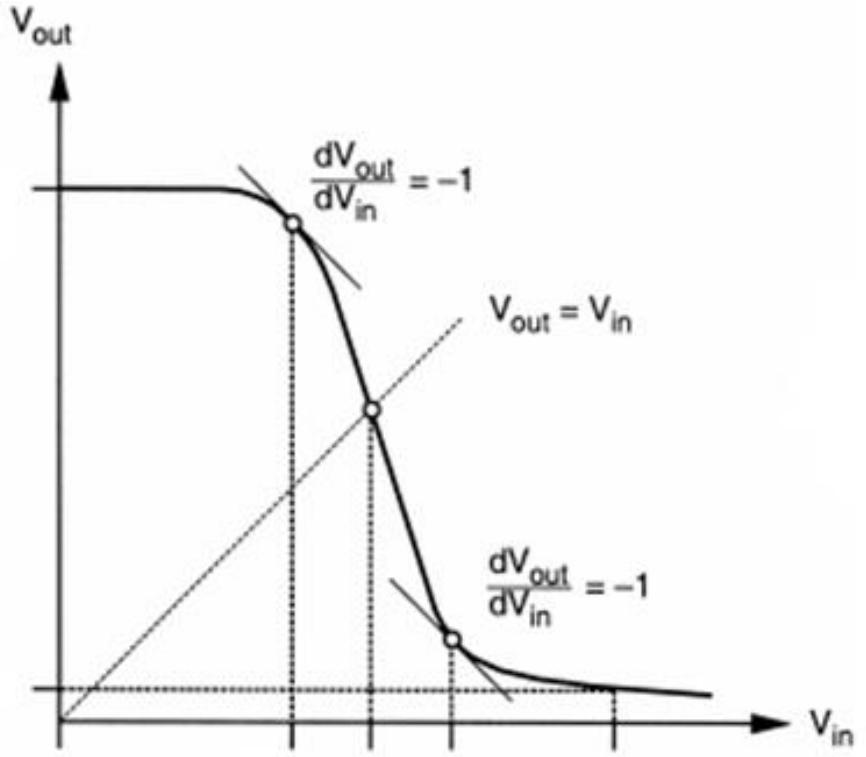
Impact of noise on output

- Large gain in the transition region amplifies the noise

$$\left.\frac{\partial V_{out}}{\partial V_{in}}\right|_{V_{in}=V_{DD}/2} \gg 1$$

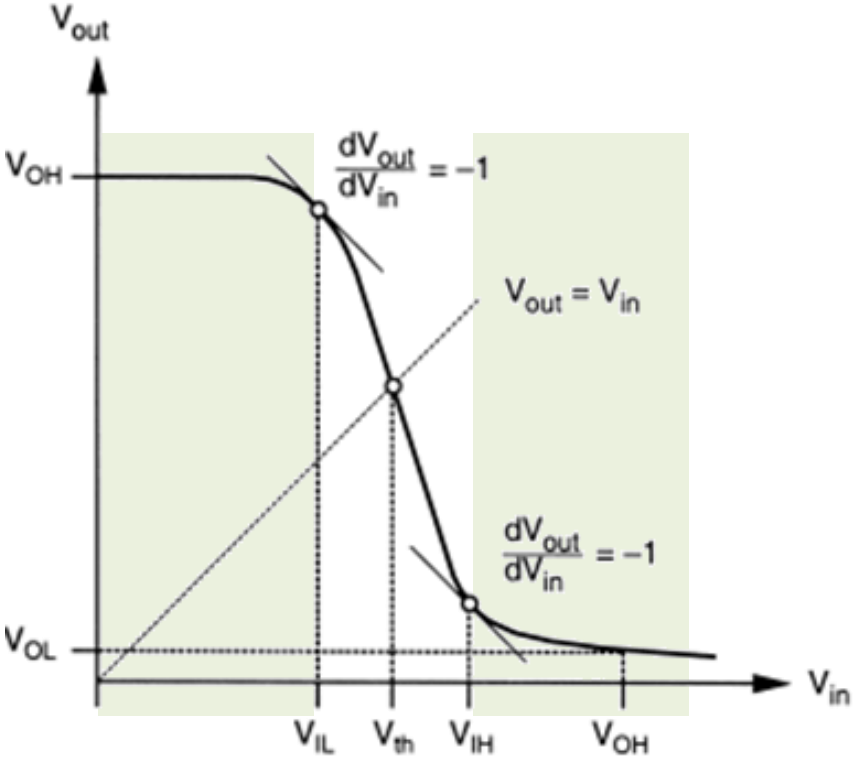
- Saturation to the rails limits gain and attenuates the noise

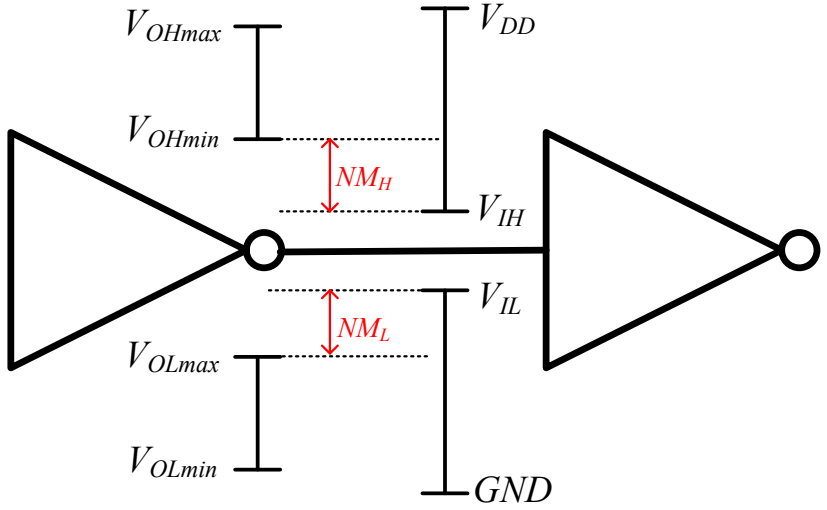
$$\left.\frac{\partial V_{out}}{\partial V_{in}}\right|_{V_{in}=0,V_{DD}} \ll 1$$



Noise Margins

- **Noise insensitive region** of a gate: $\frac{\partial V_{out}}{\partial V_{in}} \ll 1$
- **Noise margin**: margin between the min/max high/low output of one gate and the noise insensitive region of the next gate

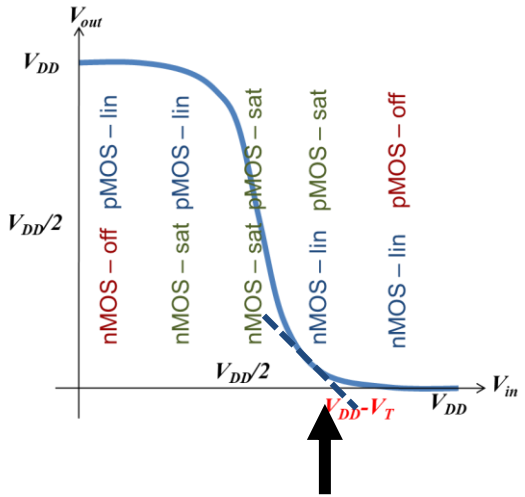




$$NM_H = V_{OH\,\min} - V_{IH}$$

$$NM_L = V_{IL} - V_{OL\,\max}$$

Analytical Derivation of Noise Margins



- Assume **balanced inverter:** $\beta_n = \beta_p = \beta$ and $V_{Tn} = -V_{Tp} = V_T$
- **Find** the **unit-gain points** of the VTC
  - Consider only V_{IH}
  - nMOS linear, pMOS saturation

$$I_{DSn}(res) = \cancel{\beta}\left[(V_{in} - V_T)V_{out} - \frac{{V_{out}}^2}{2}\right] = -I_{DSp}(sat) = \frac{\cancel{\beta}}{2}(V_{in} - V_{DD} + V_T)^2$$

  - Note that $\boldsymbol{\beta}$ **does not impact the NM if** the inverter is **balanced**
  - Obtain $\frac{\partial V_{out}}{\partial V_{in}}$ through implicit derivation (without the need to first solve for $V_{out} = \cdots$).
  - Set $\frac{\partial V_{out}}{\partial V_{in}} = -1$ and solve for V_{in} ➔ $V_{IH} = \frac{1}{8}(5V_{DD} - 2V_T)$

➔ $V_{IL} = \frac{1}{8}(3V_{DD} + 2V_T)$

$$NM_H = NM_L \approx V_{DD} - V_{IH} = V_{IL} - 0 = \frac{1}{8}(3V_{DD} + 2V_T)$$

EE-429
Fundamentals of VLSI Design

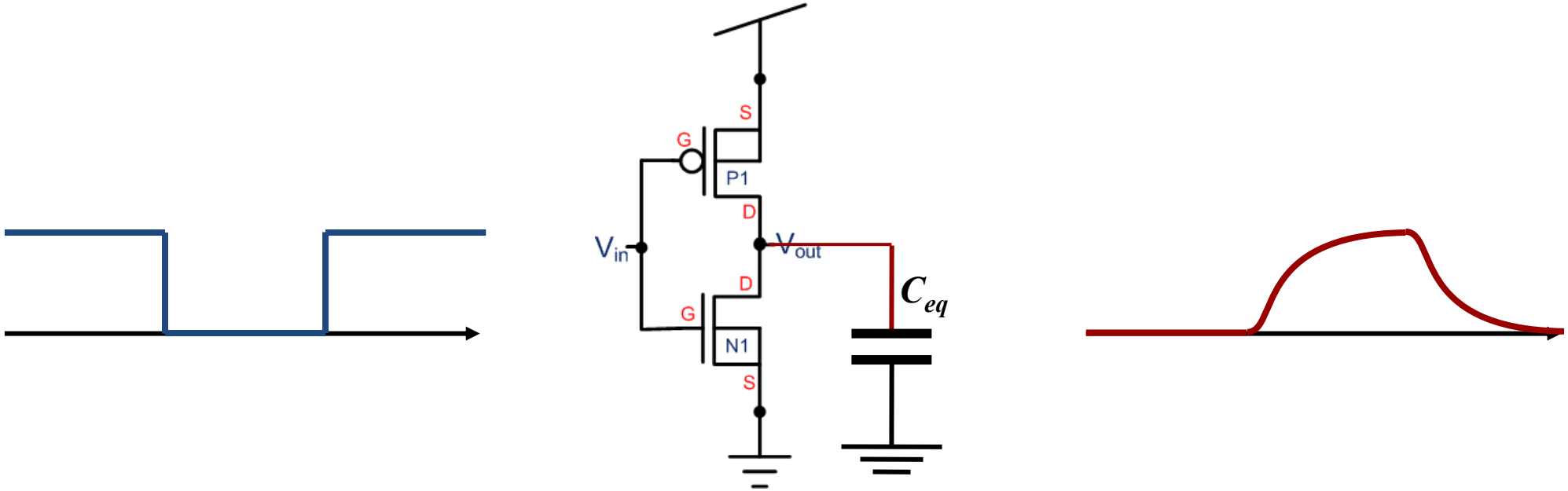
Inverter Dynamic Characteristics

Andreas Burg

Origin of Delay in Digital Circuits

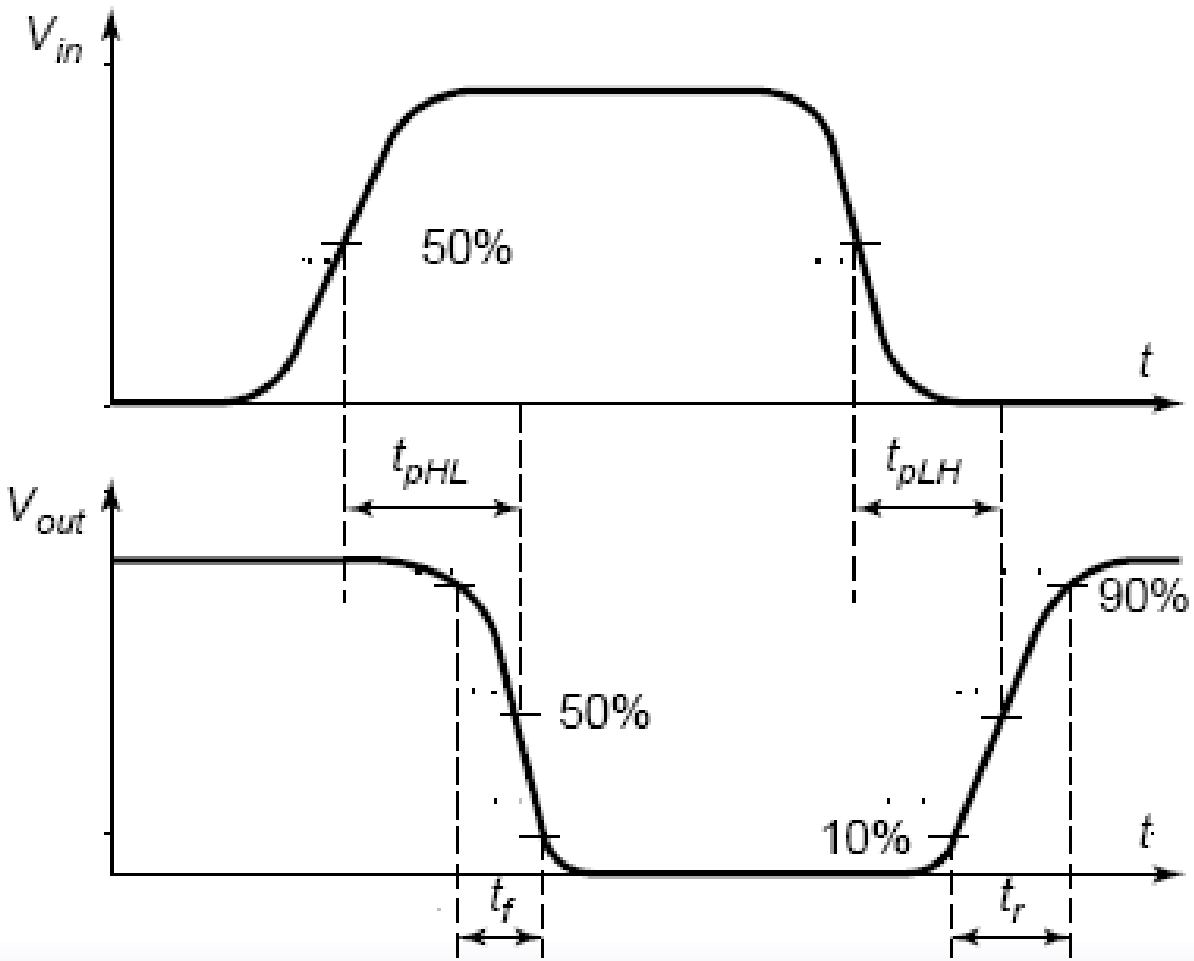
- DC characteristic only indicates the stead-state behaviour
  - Capacitances are removed (replaced by an open circuit)

- **For the dynamic behaviour** (speed) of a circuit, **capacitances play a** major **role**
  - **Time to charge/discharge the load** (output capacitance) of a circuit **determines the speed**



Delay Parameters

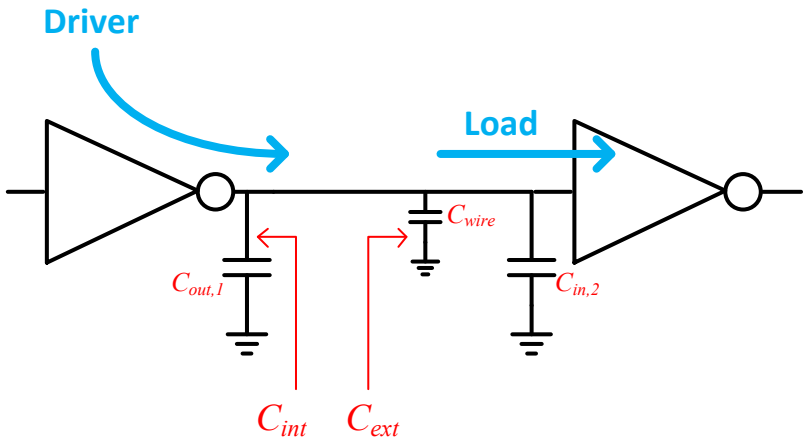
- **Delays impact the speed of a digital circuit**
- Propagation Delay:
  - **Measures** the **delay from** 50% transition point at the **input to** the 50% transition point at the **output**
  - Often different for rise-rise, rise-fall, fall-rise, fall-fall
  - Measured between 2 signals
- Rise/Fall Time:
  - **Time for a transition** from 10%-to-90% or 90%-to-10% of the full swing
  - Measured on one signal only

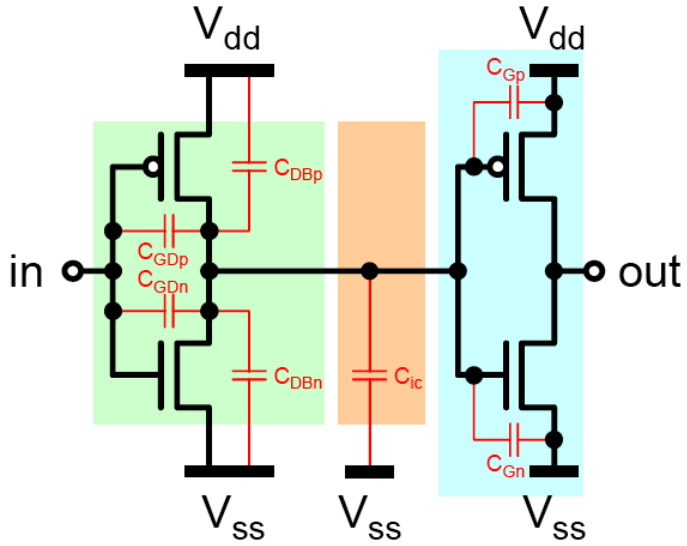


Parasitic Capacitances

- The **capacitive load C_{load}** on a gate results **from three main origins**
  - **Intrinsic capacitance C_{INT}** from the **transistors of the driving gate**
  - **Wire capacitance** from the **connected routing wires**
  - **Input load** of the connected (driven) **gate(s)**

Extrinsic: C_{EXT}

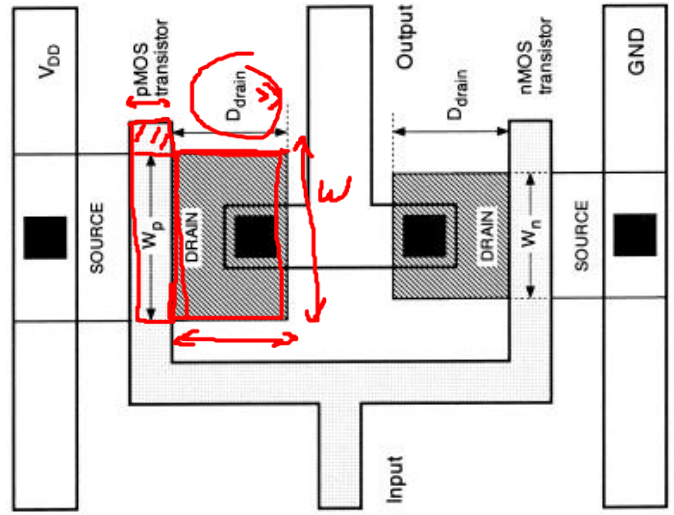




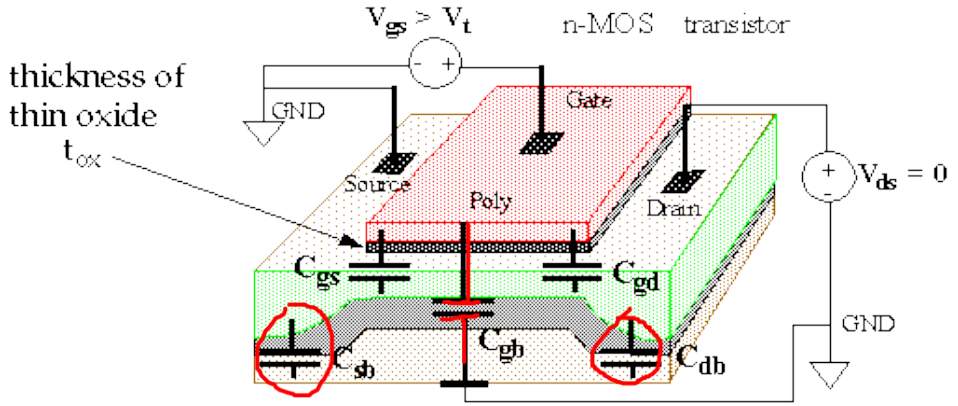
$$C_{load} = C_{out} + C_{wire} + N \cdot C_{in}$$

Parasitic Capacitance of Transistors

- **Intrinsic capacitance and input load** are caused by MOS transistors and **depend on transistor parameters**
  - Unfortunately, MOS capacitors are highly non-linear and depend on terminal voltages (non-constant)
  - For better intuition, we use **a simple constant capacitance model that** only **scales with** transistor **geometries**



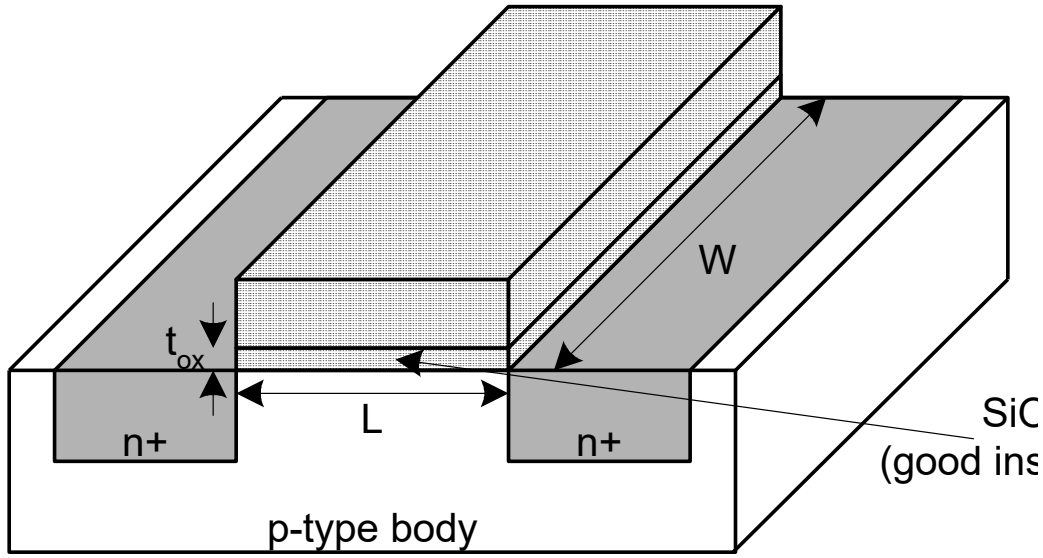
- Dominant capacitances:
  - **Gate/Channel Capacitance**: capacitance caused by the insulating oxide layer under the gate $\sim W \cdot L$
  - **Junction Capacitance**: pn-Junction capacitance between the diffusions and the substrate $\sim W \cdot D_{drain}$



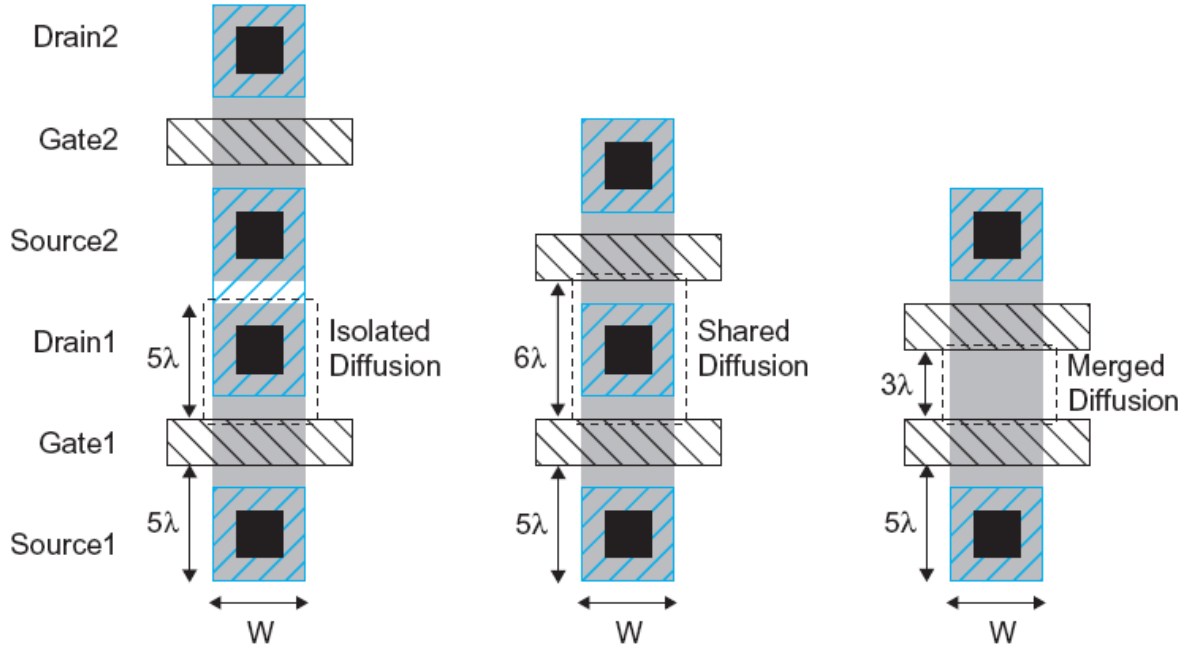
- **Larger transistors also bring larger load**

Gate and Diffusion Capacitances

- **Affects delay of previous gate**
- **Approximate channel as connected to source**
- $C_{gs} = \varepsilon_{ox}WL/t_{ox} = C_{ox}WL \sim W \cdot L$

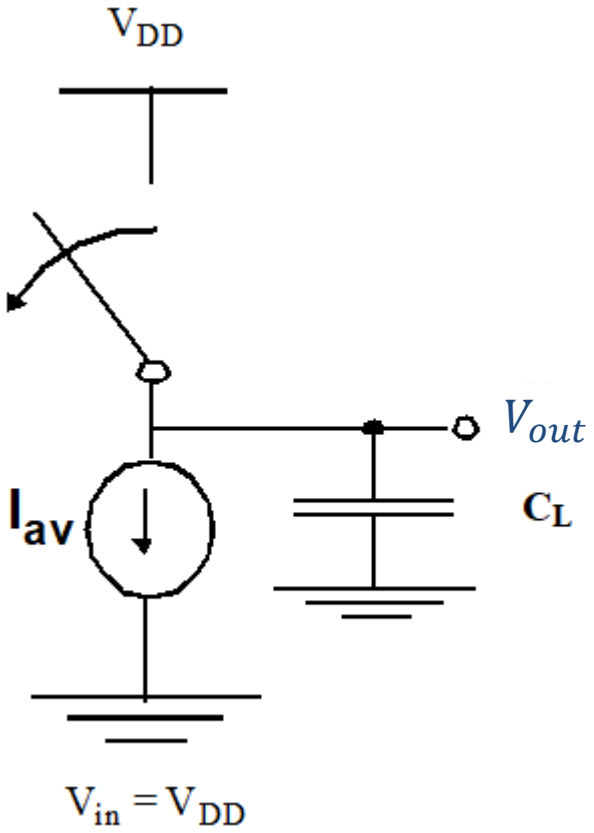


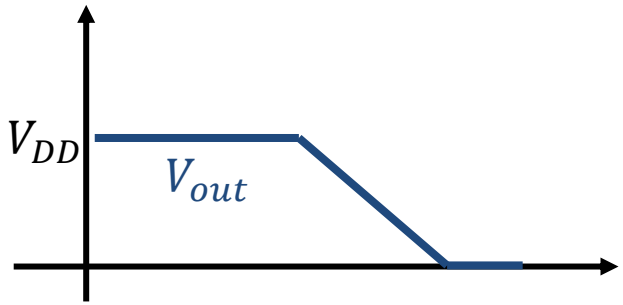
- **Affects delay of same gate**
- **Depends on area $\sim W \cdot D_{drain}$ & perimeter**
  - Use small diffusion nodes
  - Comparable to C_g for contacted diff
  - ½ C_g for uncontacted



A Simple Delay Model: Current Source Model

- **Simplest** delay **model** considers **transistor as a current source** assuming it operates in saturation during the entire transition
  - Transition measured from V_{DD} to $V_{DD}/2$



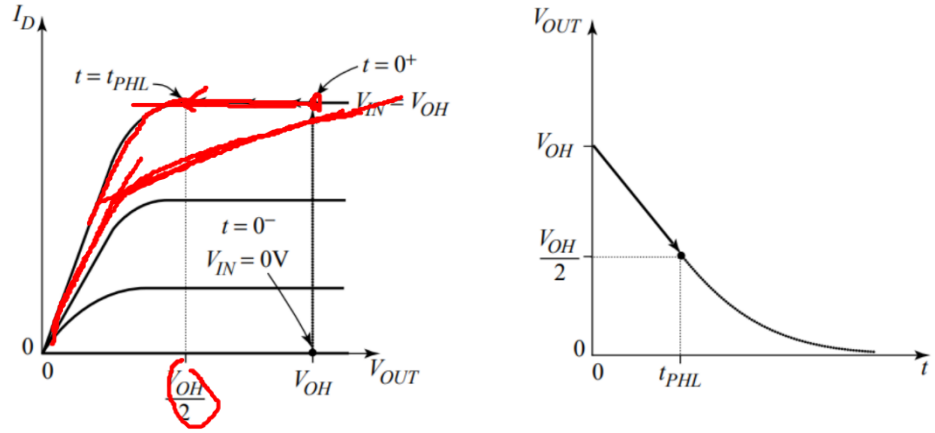


$$I_{DSn} = \frac{\beta}{2}(V_{in} - V_{Tn})^2$$

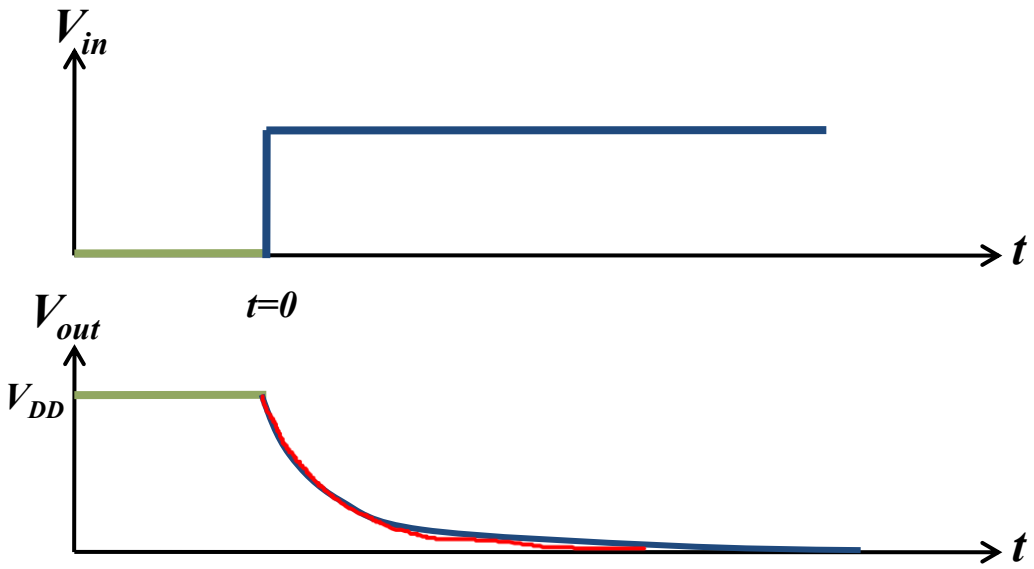
$$t_{HL} = \frac{C_L \cdot \frac{V_{DD}}{2}}{I_{DS}} = \frac{C_L \cdot V_{DD}}{\beta(V_{DD} - V_{Tn})^2}$$

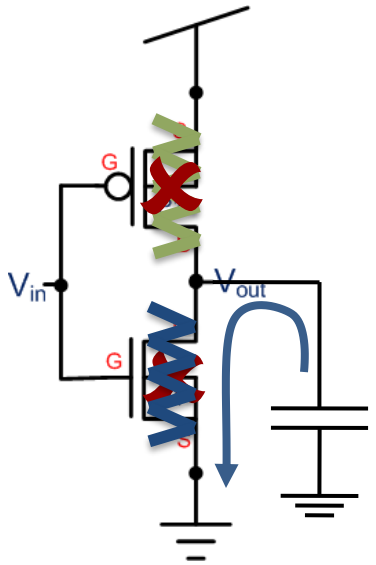
The RC Delay Model

- Unfortunately, the **current source model**
  - Produces a rather inaccurate waveform
  - Becomes inaccurate with more sophisticated transistor models (e.g., long channel models)



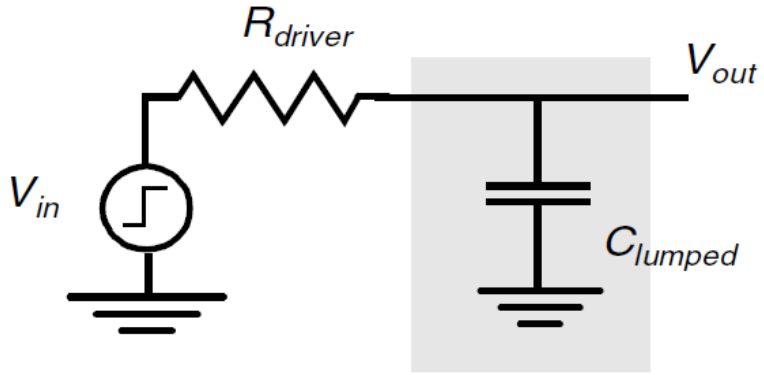
- **Better model: replace driving transistor with an equivalent resistor**





Delay of an RC Network

- **Delay can be derived from well-known equations from RC networks**



$$V_{out}(\uparrow) = V_{DD}\left(1 - e^{-\frac{t}{RC}}\right)$$

$$V_{out}(\downarrow) = V_{DD}e^{-\frac{t}{RC}}$$

Time required to reach x%:

$x = e^{-\frac{t}{RC}}$ ➔ $t = -R \cdot C \cdot \ln x$

50% rise/fall delay:

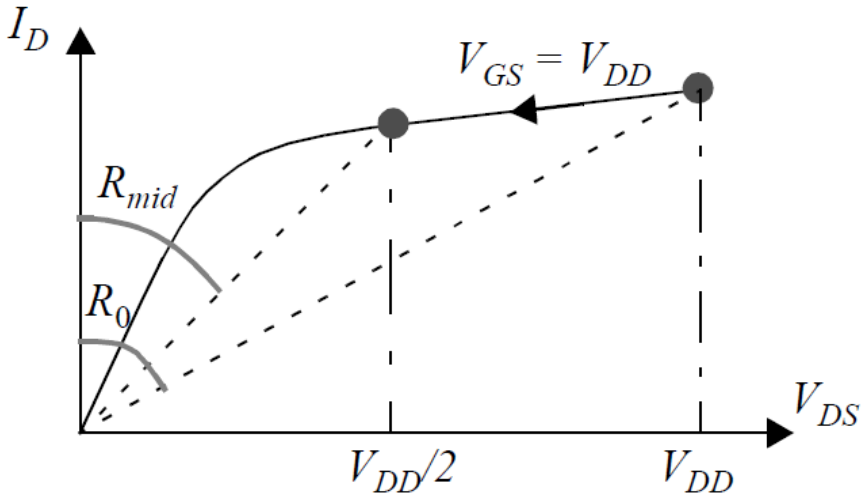
$t_{50\%} = R \cdot C \cdot 0.69$

Computing an Equivalent Resistance

- **On-resistance** of the MOSFETS **changes** during transition

- **Use** an **equivalent resistance** reflecting a good average of the transition region
  - Start from onset of the transition: $V_{out} = V_{DD}$
  - Up to the midpoint: $V_M \approx V_{DD}/2$

$$R_{eq} = average_{t=t_1 \ldots t_2}\left(R_{on}(t)\right) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} R_{on}(t)\,dt = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{V_{DS}(t)}{I_{DS}(t)}\,dt$$



$$\approx \frac{1}{2}\left[R_{on}(t_1) + R_{on}(t_2)\right]$$

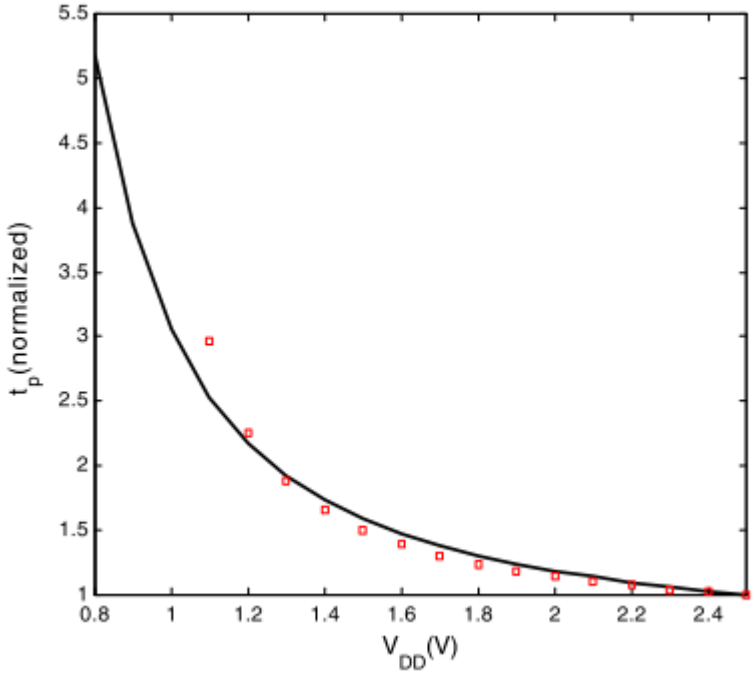
$$R_{eqn} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{2V_{DS}}{\beta_n (V_{DD} - V_T)^2}\,dV_{DS} \approx \frac{3}{2}\frac{V_{DD}}{\beta_n (V_{DD} - V_T)^2}$$

Propagation Delay with Equivalent Resistance

- Analyze the parameters that affect the propagation delay:

$$t_{pHL} = 0.69 \cdot \frac{3}{2} \frac{V_{DD}}{\beta_n (V_{DD} - V_T)^2} C_{load} = 1.035 \frac{\boldsymbol{L}}{\boldsymbol{W}} \frac{1}{k_n} \frac{V_{DD}}{(V_{DD} - V_T)^2} C_{load}$$

- Accordingly, we can minimize the delay in the following ways:
  - Minimize C_{load}
  - Increase **W/L** typically by making W wider
  - Increase V_{DD}



Effect of Device Sizing

- To reduce propagation delay, it is tempting to increase the device width **W**

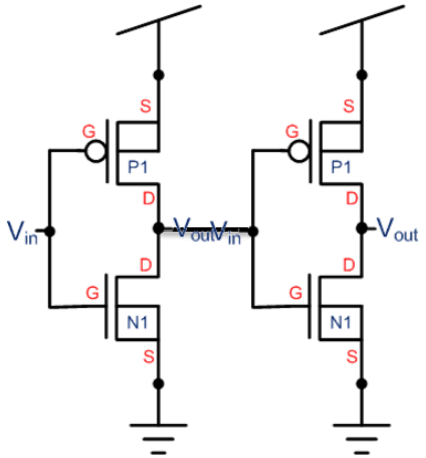
- **At the same time we usually need $\beta_p > \beta_n$ to compensate for the mobility ratio of holes and electrons for a balanced inverter**
  - This generally equates the propagation delay of High-to-Low and Low-to-High transitions.
  - However, this does not imply that this ratio yields the minimum overall propagation delay.

- **For this, we will discuss two sizing parameters:**
  - Beta Ratio ($\beta = \beta_p/\beta_n$): ratio between pMOS and nMOS drive strength
  - Upsizing Factor (**S**): ratio of the nMOS to a minimum-size nMOS

pMOS/nMOS Ratio

- Consider the **average propagation delay for rise and fall**

$$t_{\mathrm{pd}} = \frac{t_{\mathrm{pLH}} + t_{\mathrm{pHL}}}{2} = \frac{0.69 C_{\mathrm{load}} \left(R_{eqp} + R_{eqn} \right)}{2}$$



  - With $W_p = W_n$, we get an unbalanced inverter since $\mu_p < \mu_n$ which leads to
    - $V_M < V_{DD}/2$
    - $R_{eqp} > R_{eqn}$ and therefore $t_{\mathrm{pdr}} > t_{\mathrm{pdf}}$

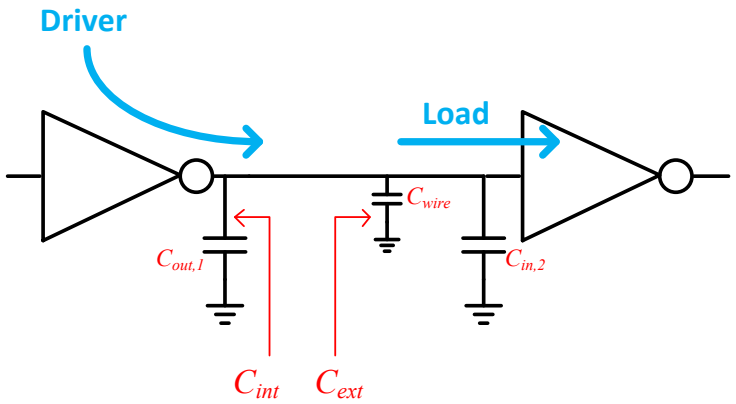
  - We usually **enlarge** the **pMOS to get a "balanced" inverter setting** $W_p = W_n \gamma$ with $\gamma \approx 2 \ldots 4$ **for** $V_M = V_{DD}/2$

- It is worth asking: **is the balanced inverter also the fastest?**

Optimizing pMOS/nMOS Ratio for Minimum Delay

- Consider **two** identical **cascaded** CMOS **inverters**:
  - The load capacitance on the driving gate is:

$$C_{load} = C_{out1} + C_{in2} + C_{wire}$$



  - The input capacitance of the 2nd gate C_{in2} and the output capacitance of the driving gate C_{out1} **contain** both **pMOS and nMOS parasitics**

$$C_{load} = (C_{dp1} + C_{dn1}) + (C_{gp2} + C_{gn2}) + C_{wire}$$

  - Gate and drain **capacitances scale linear in the width** of the transistors

$$C_{dn1} \sim W_n \qquad C_{dp1} \sim W_p = \gamma W_n$$

$$C_{gn1} \sim W_n \qquad C_{gp1} \sim W_p = \gamma W_n$$

$$C_{load} = (1 + \gamma)(C_{dn1} + C_{gn2}) + C_{wire}$$

Optimizing pMOS/nMOS Ratio for Minimum Delay

- Formulate the **propagation delay**, **including** the **impact of γ on the parasitics and on the pMOS** $R_{eqp}(\boldsymbol{W_n}\gamma) = R_{eqp}(\boldsymbol{W_n})/\gamma$

$$t_{pd} = \frac{0.69C_{load}}{2}\left(R_{eqn} + R_{eqp}\right) = \frac{0.69}{2}\left[(1+\gamma)\left(C_{dn1} + C_{gn2}\right) + C_{wire}\right]\left(R_{eqn} + \frac{R_{eqp}(\boldsymbol{W_n})}{\gamma}\right)$$

- Find $\boldsymbol{\gamma}$ to minimize the delay: $\frac{\partial t_{pd}}{\partial \gamma} = 0$

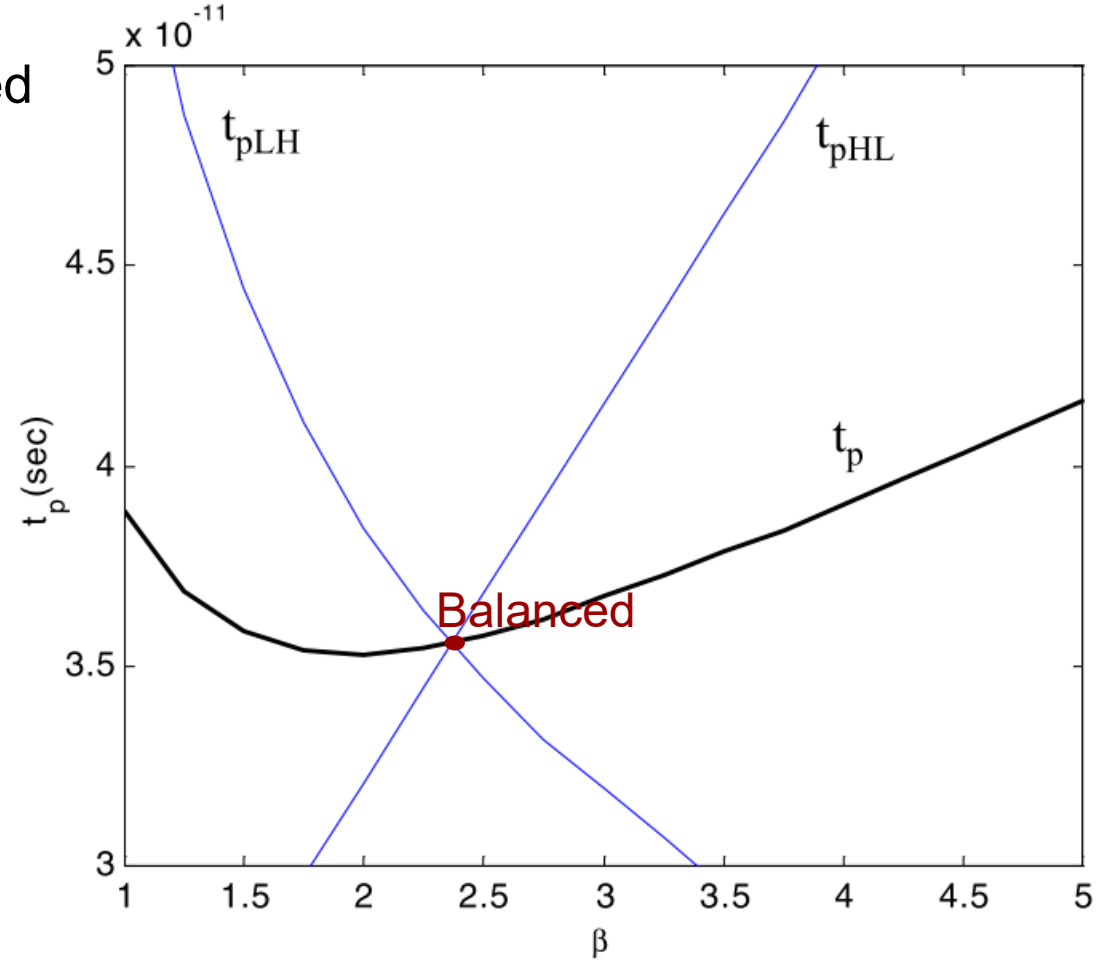
$$\gamma_{opt} = \sqrt{\frac{R_{eqp}(\boldsymbol{W_n})}{R_{eqn}}\left(1 + \frac{C_{wire}}{C_{dn1} + C_{gn2}}\right)}\Bigg|_{(C_{dn1}+C_{gn2} >> C_{wire})} \rightarrow \sqrt{\frac{R_{eqp}}{R_{eqn}}}$$

- A **typical optimum** γ_{opt} is usually **around 2**

Optimizing pMOS/nMOS Ratio for Minimum Delay

- **Balanced inverter usually does not provide the shortest delay**

  - Delay penalty compared to a delay-optimized inverter is often small.



Impact of Device Sizing on Delay

- **Recap**: pMOS/nMOS ratio can be optimized to minimize delay
  - The balanced inverter is not always optimal, but often close

- See how **upsizing both pMOS and nMOS** affects the delay?
- Consider a **minimum size balanced** (γ_0 for $V_M = V_{DD}/2$) **inverter as reference**
  - $W_n^0 = W_{min}, W_p^0 = \gamma_0 W_{min}, L_{n/p}^0 = L_{min}$

Common upsizing factor S

$$W_p = SW_p^0 \qquad W_n = SW_n^0$$

$$C_{int}(S) = S \cdot C_{int}^0$$

$$R_{eq}(S) = R_{eq}^0/S$$

Impact of Device Sizing on Delay

- Start by **writing the delay as** a **function of** the **intrinsic capacitance** and the **external capacitances** (fanout and wiring)
  - **Intrinsic capacitance: scales with S** compared to that of a minimum size inverter
  - **External capacitance: constant**

$$C_{load} = C_{int}(S) + C_{ext}$$

$$t_{pd} = 0.69 \cdot R_{eq}(S) \cdot C_{load} = 0.69 \cdot R_{eq}(S) \cdot [C_{int}(S) + C_{ext}]$$

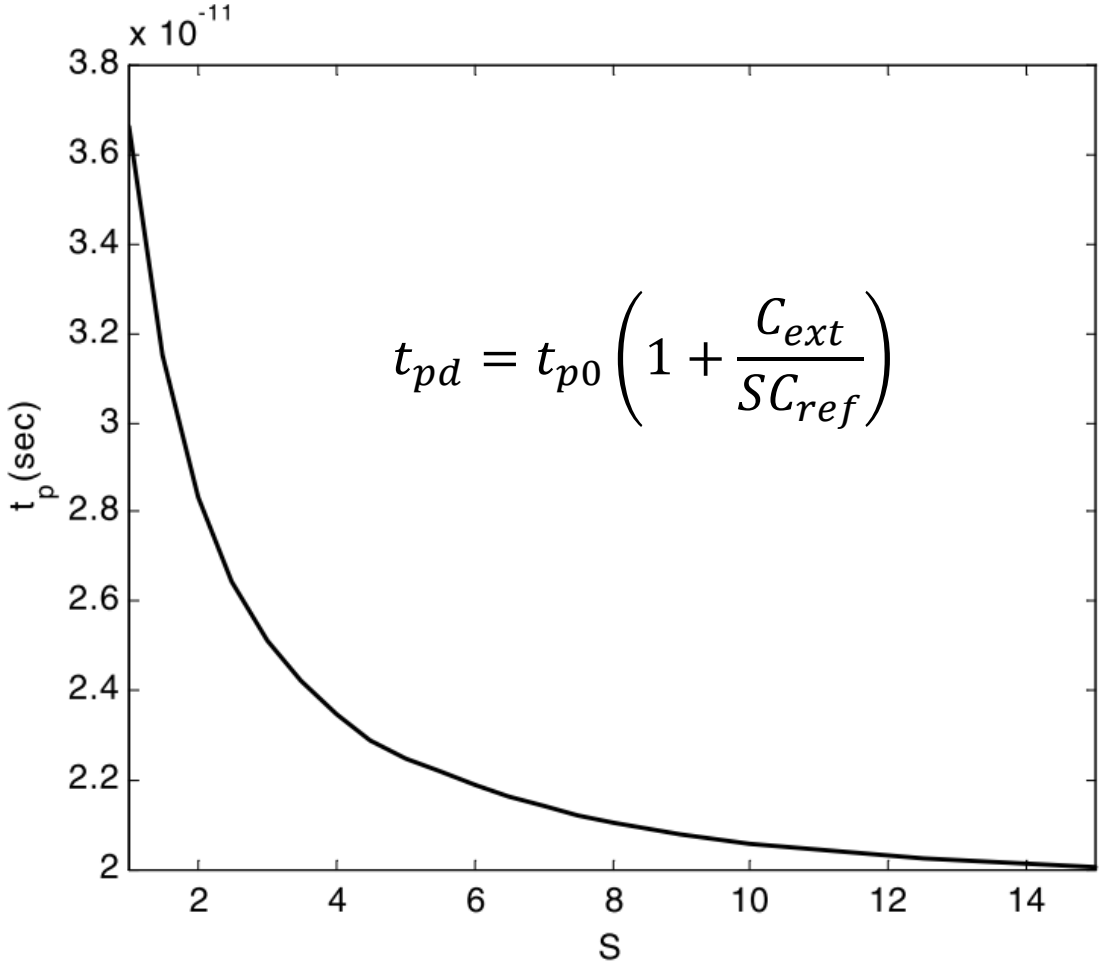
$$= 0.69 \cdot \frac{R_{eq}^0}{S}\left[S \cdot C_{int}^0 + C_{ext}\right]$$

$$= \underbrace{0.69 \cdot R_{eq}^0 \cdot C_{int}^0}_{t_{p0}} \cdot \left(1 + \frac{C_{ext}}{S \cdot C_{int}^0}\right) = t_{p0}\left(1 + \frac{C_{ext}}{SC_{ref}}\right)$$

Impact of Device Sizing on Delay

- The intrinsic delay of an inverter (tp0) is independent of the sizing of the gate and is purely determined by technology.
- When no load is present, an increase in the drive of the gate is totally offset by the increased capacitance.

- **Upsizing becomes relevant once the external capacitance starts to dominate**



$$t_{pd} = t_{p0}\left(1 + \frac{C_{ext}}{SC_{ref}}\right)$$