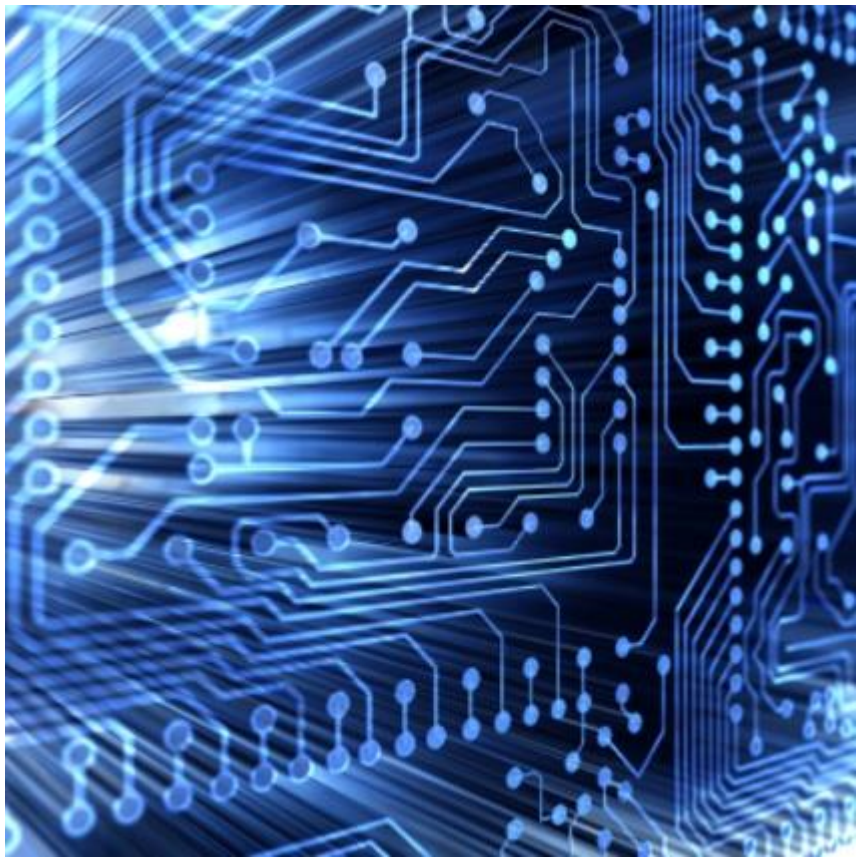
Digital Microelectronic Circuits (361-1-3021)



Presented by: Adam Teman

Lecture 4:

The CMOS Inverter

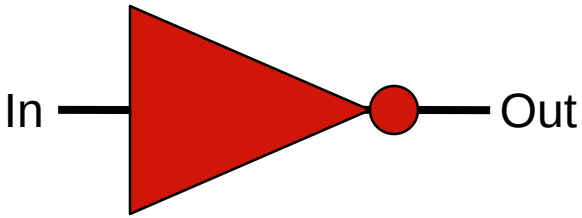
Last Lectures

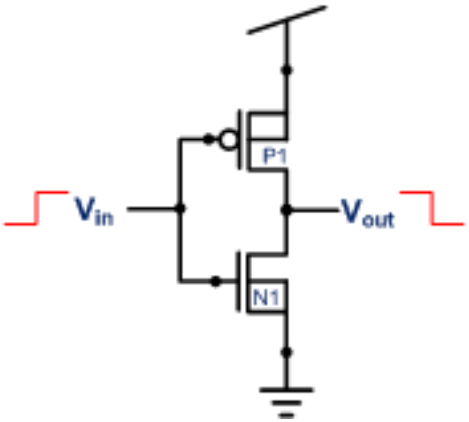
- Moore's Law
- Terminology
  - Static Properties
  - Dynamic Properties
  - Power
- The MOSFET Transistor
  - Shockley Model
  - Channel Length Modulation, Velocity Saturation, Body Effect

This Week - Motivation

- The ***Inverter***, or ***NOT*** gate, is truly the nucleus of all digital designs.
- We will analyze the inverter and find its characterizing parameters.
- Once its operation and properties are clearly understood, designing and analyzing more intricate structures, such as ***NAND*** gates, adders, multipliers and microprocessors is greatly simplified.
- This lecture focuses on the ***static CMOS inverter*** – the most popular at present and the basis for the ***CMOS digital logic*** family.

In	Out
0	1
1	0





What will we learn today?

- 4.1 An Intuitive Explanation
- 4.2 Static Operation
  - 4.2.1 The Inverter's VTC
  - 4.2.2 Operating Regions
  - 4.2.3 Switching Threshold
  - 4.2.4 Noise Margins
- 4.3 Dynamic Operation
  - 4.3.1 Parasitic Capacitances
  - 4.3.2 Propagation Delay
  - 4.3.3 Device Sizing - β
  - 4.3.4 Device Sizing – S
  - 4.3.5 Sizing a Chain of Inverters
- 4.4 Power Consumption
  - 4.4.1 Dynamic Power
  - 4.4.2 Short Circuit Power
  - 4.4.3 Static Power
  - 4.4.4 Total Power Consumption
- 4.5 Summary

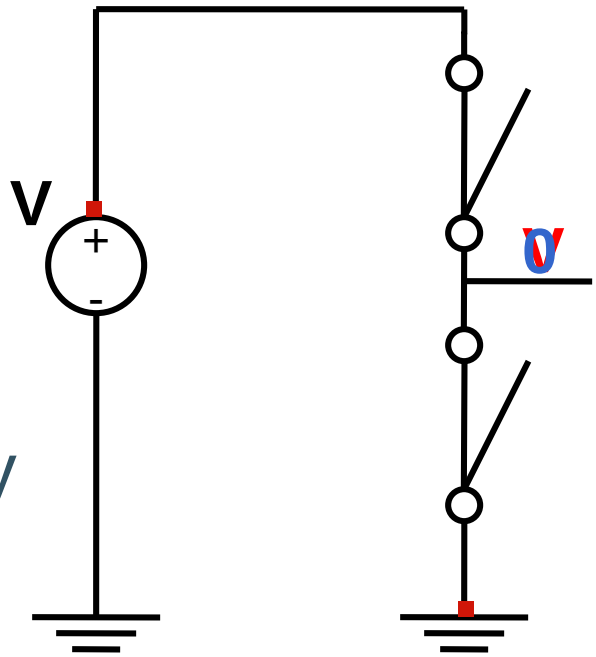
4.1

As usual, we'll start with

AN INTUITIVE EXPLANATION

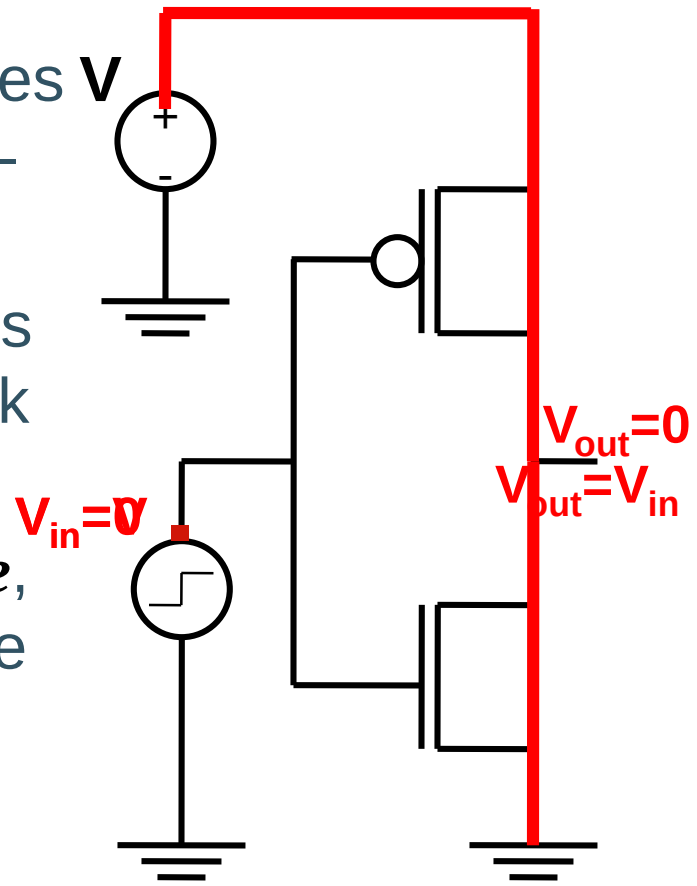
An Intuitive Explanation

- A ***Static CMOS Inverter*** is modeled on the double switch model.
- The basic assumption is that the switches are ***Complementary***, i.e. when one is ***on***, the other is ***off***.
- When the ***top switch*** is ***on***, the supply voltage propagates to the ***output node***.
- When the ***bottom switch*** is ***on***, the ***ground voltage*** is propagated out.

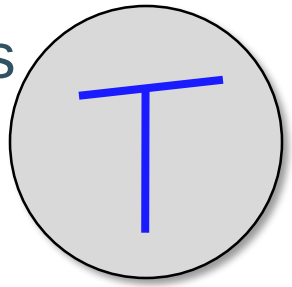


An Intuitive Explanation

- Now we will replace the model switches with real voltage controlled switches – ***MOS Transistors***.
- We will use complementary transistors – one ***nMOS*** and one ***pMOS***, and hook them up to the ***same input voltage***.
- Now, when we set a ***high input voltage***, the ***nMOS*** is ***on*** and the ***pMOS off***. The ***ground voltage*** propagates.
- When we put a ***low input voltage***, the ***pMOS*** is ***on*** and the ***nMOS*** is ***off***. The ***supply voltage propagates***.
- We've built an ***inverter***!



An Intuitive Explanation



- The voltage connected to the Source of the pMOS is known as the "***Supply Voltage***" or V_{DD}.*
- We mark the connection to V_{DD} with a horizontal or slanted bar.
- Accordingly, V_{DD} represents a logical *'1'* and ***GND*** represents a logical *'0'*.
- Inputting V_{DD} to the CMOS inverter will present ***GND*** at the output. Inputting ***GND*** will present V_{DD} at the output.
- This characteristic is non-trivial and is one of the advantages of CMOS design. It is known as ***"Rail to Rail Swing"***.**

* It can also be called V_{CC}, regarding the *Collector* of BJT transistors.

** "***Rails***" are the supply voltages, i.e. V_{DD} and ***GND***. If a voltage is connected to the nMOS Source instead of ***GND***, we refer to this voltage as V_{SS}.

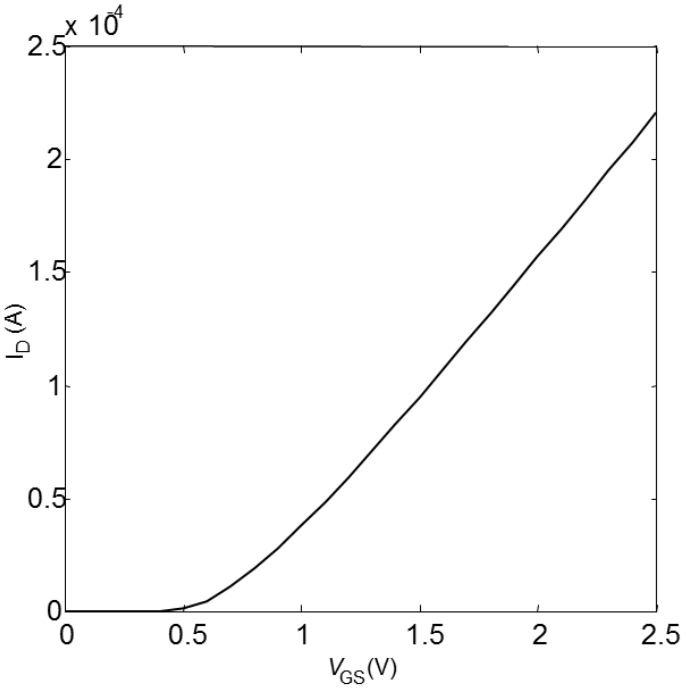
4.2

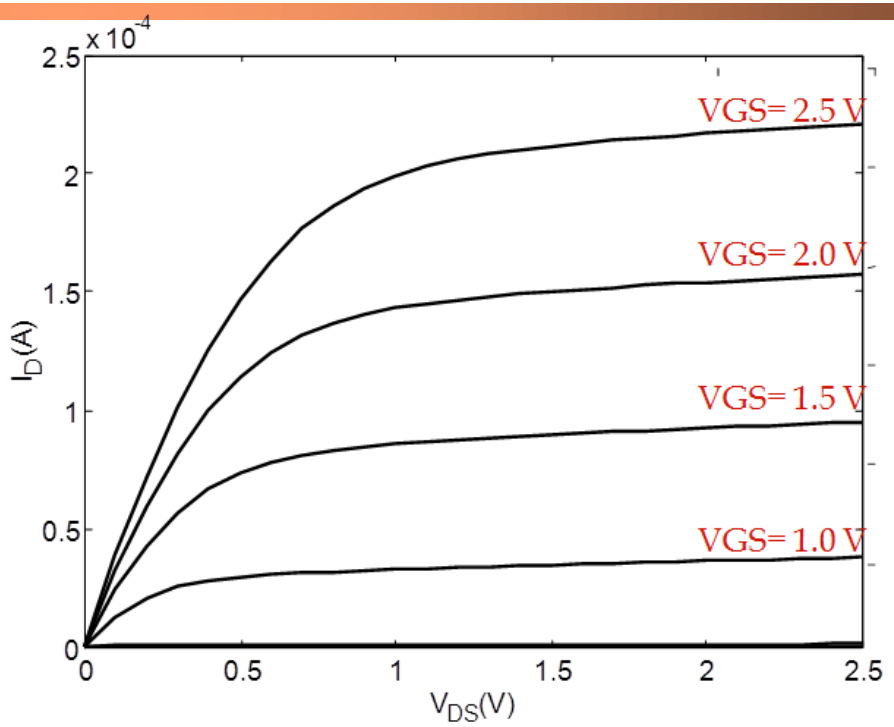
- 4.1 An Intuitive Explanation
- 4.2 Static Operation
- 4.3 Dynamic Operation
- 4.4 Power Consumption
- 4.5 Summary

Now that we understand the principles, we'll analyze

STATIC OPERATION

Reminder: The Unified MOSFET Model





$$I_{DSn} = k_n' \frac{W_n}{L_n}\left[\left(V_{GSn} - V_{Tn}\right)V_{DSeff} - \frac{V_{DSeff}^{\ 2}}{2}\right]\left(1+\lambda_n V_{DSn}\right)$$

$$V_{DSeff} = \min\left(V_{GSn} - V_{Tn}, V_{DSn}, V_{DSATn}\right)$$

Reminder: Static Properties

- VTC
- Noise Margins

The Inverter's 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.
- We can see that:

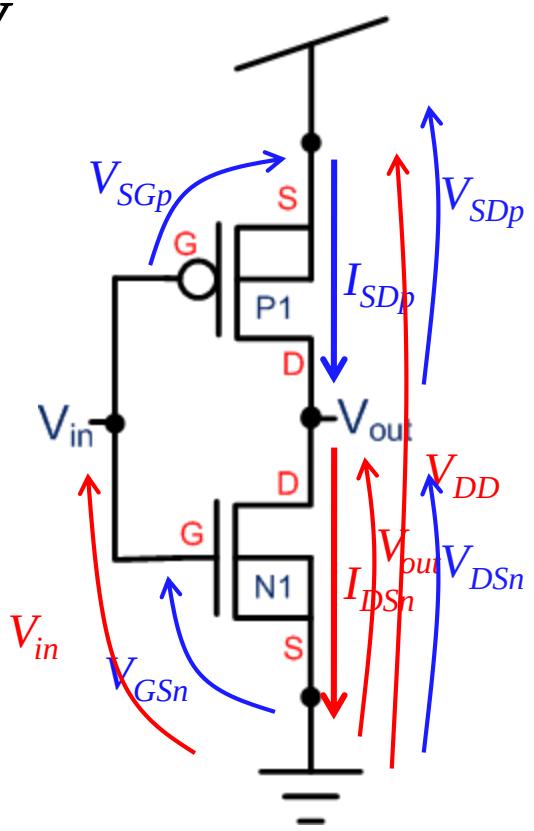
$$I_{SDp} = I_{DSn}$$

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

$$V_{SGp} = V_{DD} - V_{in}$$

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

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



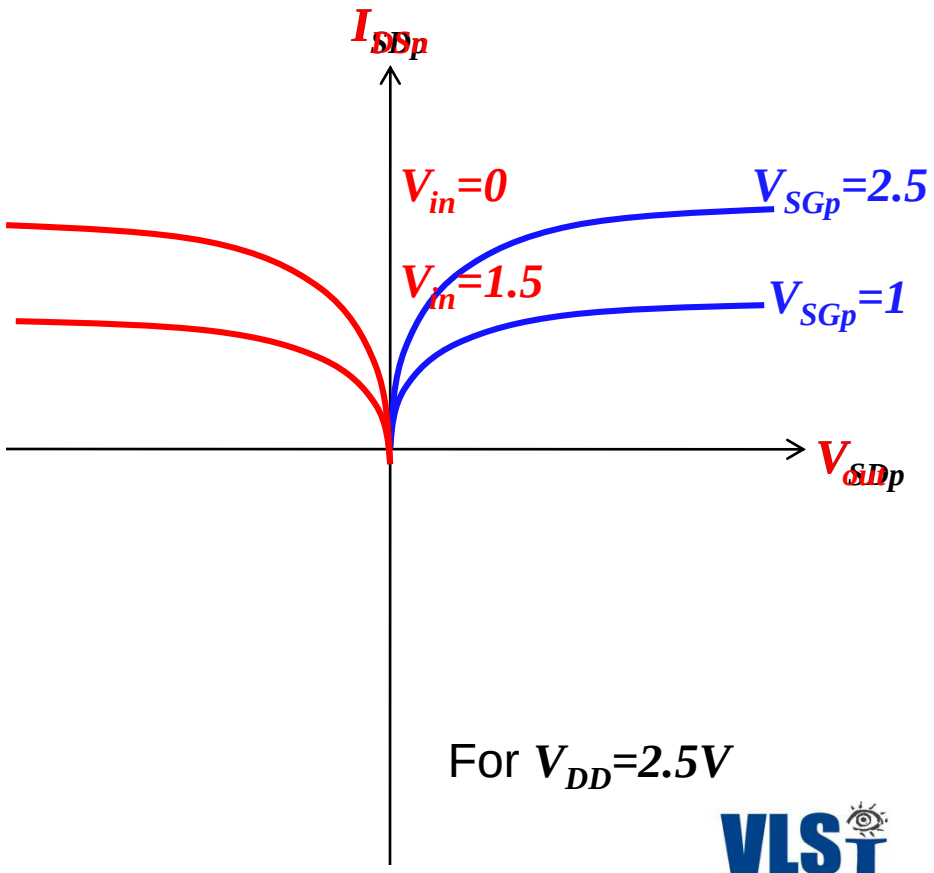
The Inverter's VTC

- Since V_{in} and V_{out} are the input and output voltages of the ***nMOS*** transistor, we will change the coordinates of the ***pMOS***.

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

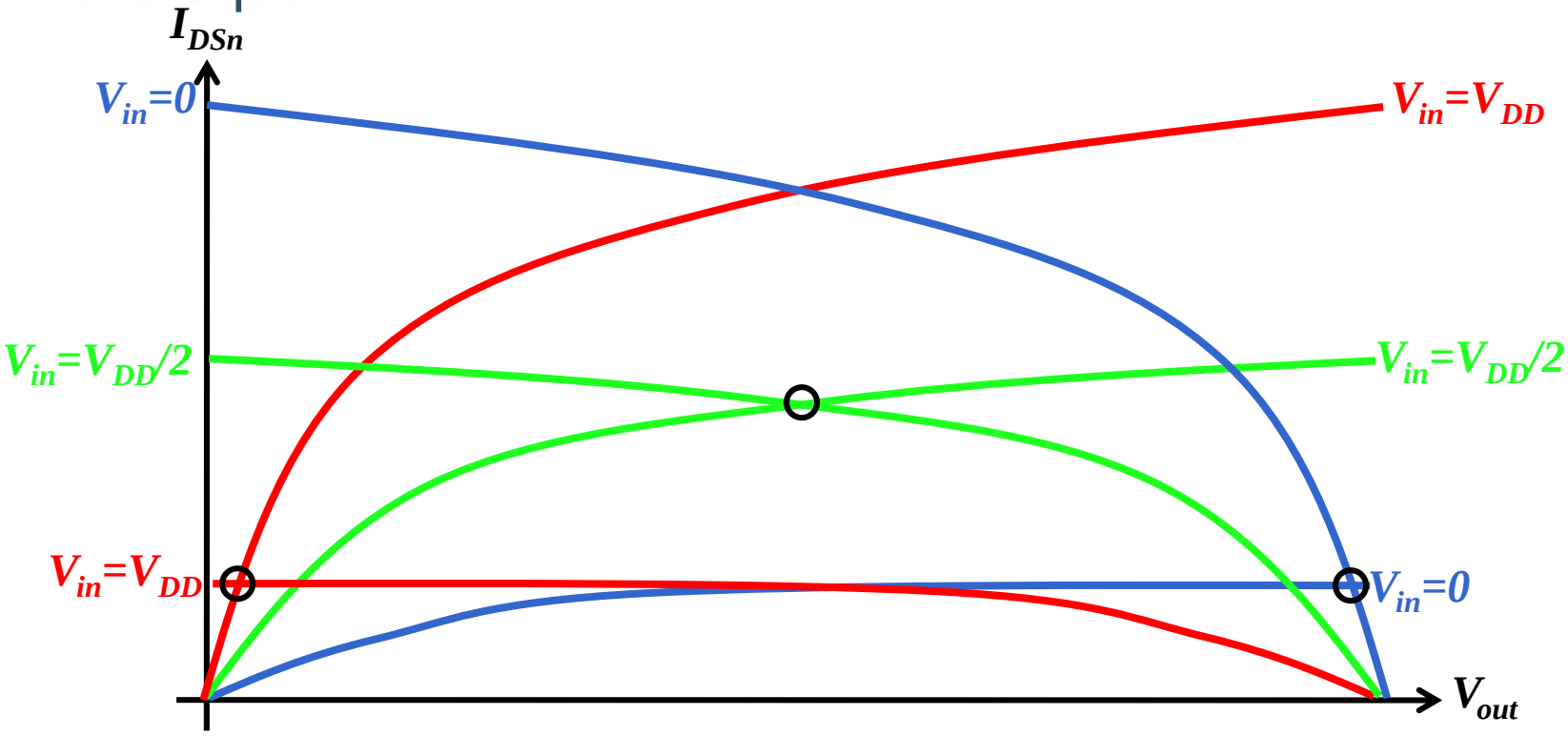
$$V_{in} = V_{DD} - V_{SGp}$$

$$I_{DSn} = I_{SDp}$$



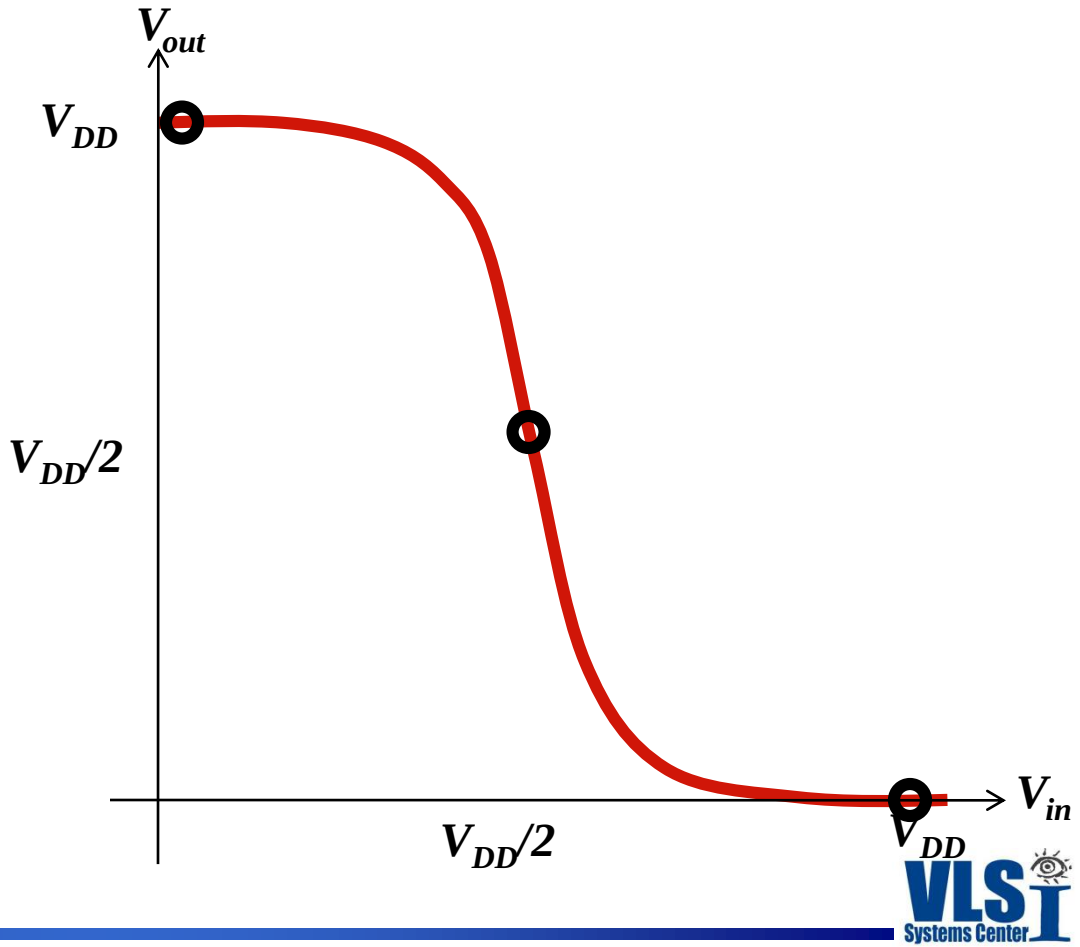
The Inverter's VTC

- Now, we will superimpose the modified ***nMOS*** I-V curves on the ***pMOS*** I-V curves, where the currents of the ***nMOS*** and ***pMOS*** are equal.

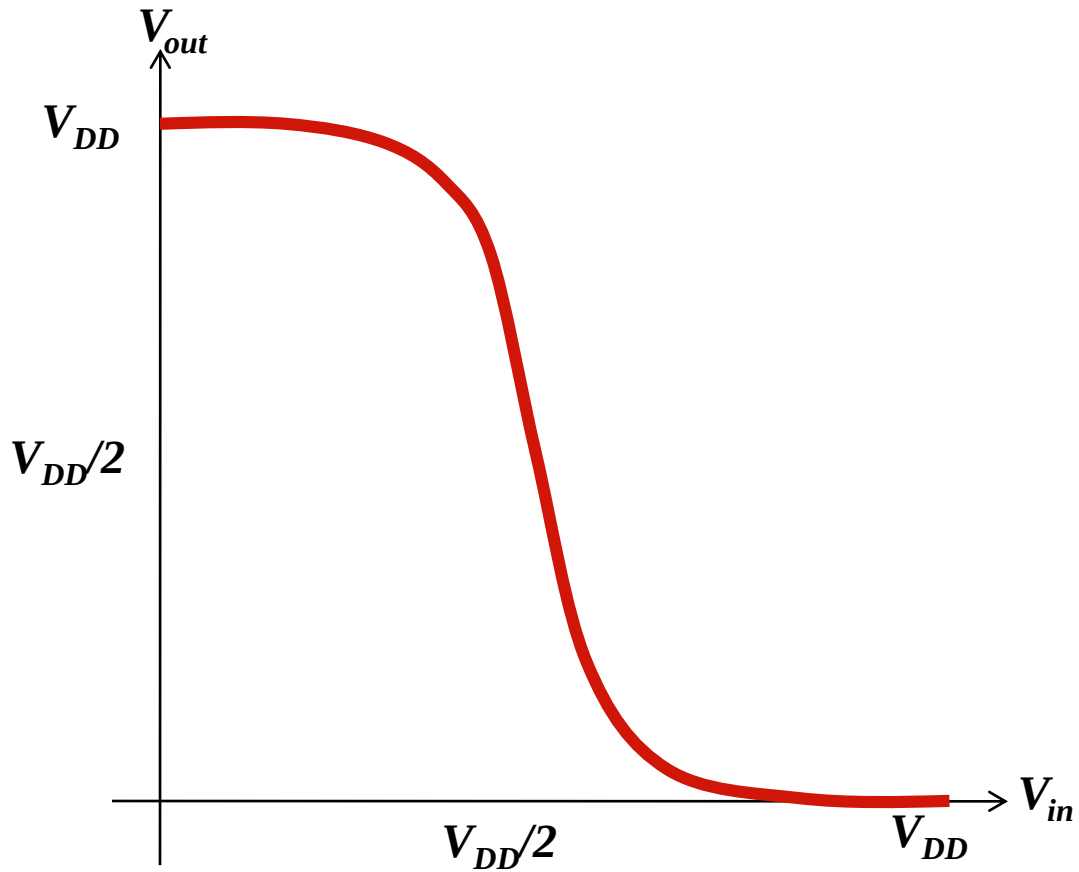


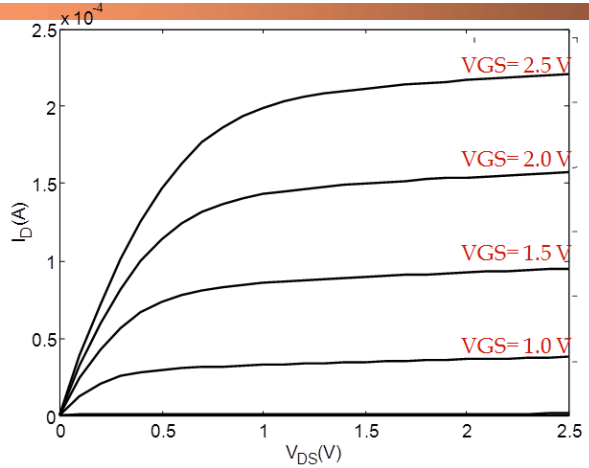
The Inverter's VTC

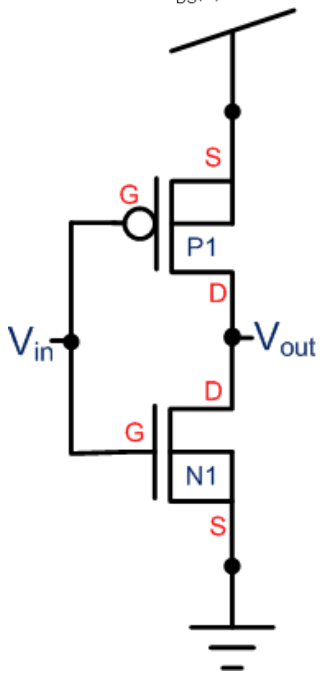
- Putting all the intersection points on a graph with the corresponding output voltage will give us the CMOS inverter's ***VTC***:



Intuitive Operating Regions



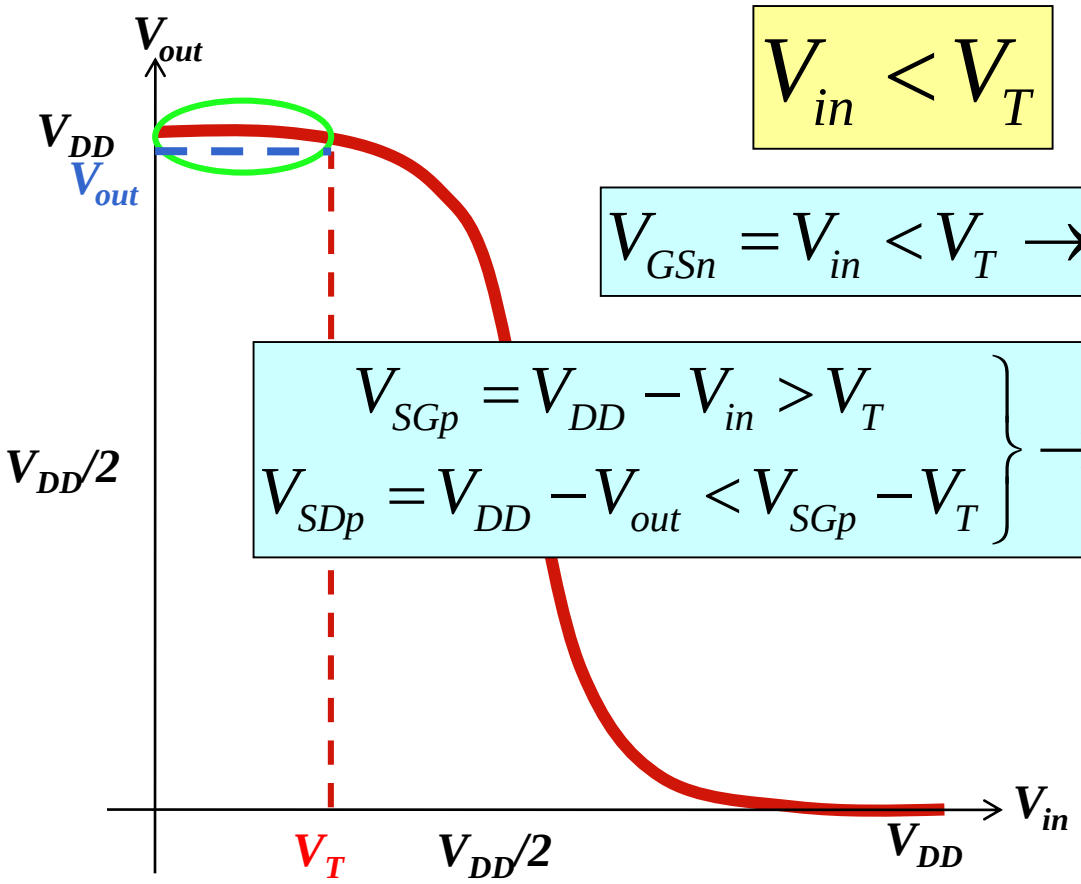




Operating Regions

- Let's figure out what ***region of operation*** each transistor is in throughout the ***VTC*** curve.*

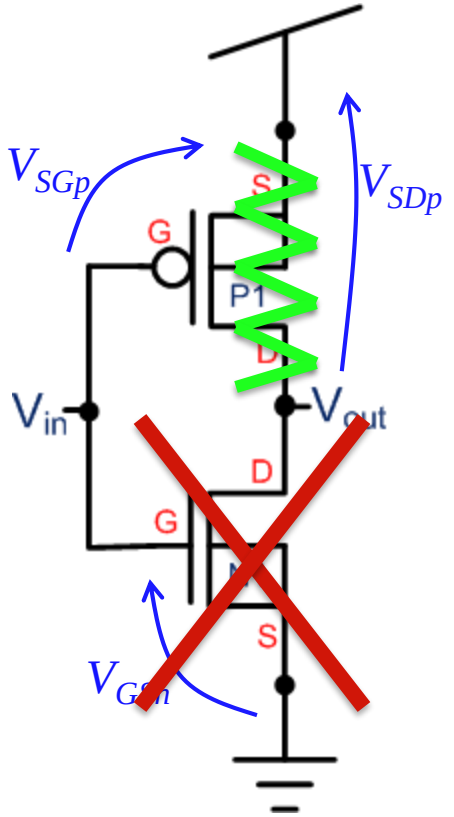
* Considering Long Channel Transistors With $V_T < V_{DD}/2$



$$V_{in} < V_T$$

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

$$\left.\begin{aligned} V_{SGp} &= V_{DD} - V_{in} > V_T \\ V_{SDp} &= V_{DD} - V_{out} < V_{SGp} - V_T \end{aligned}\right\} \rightarrow Linear$$



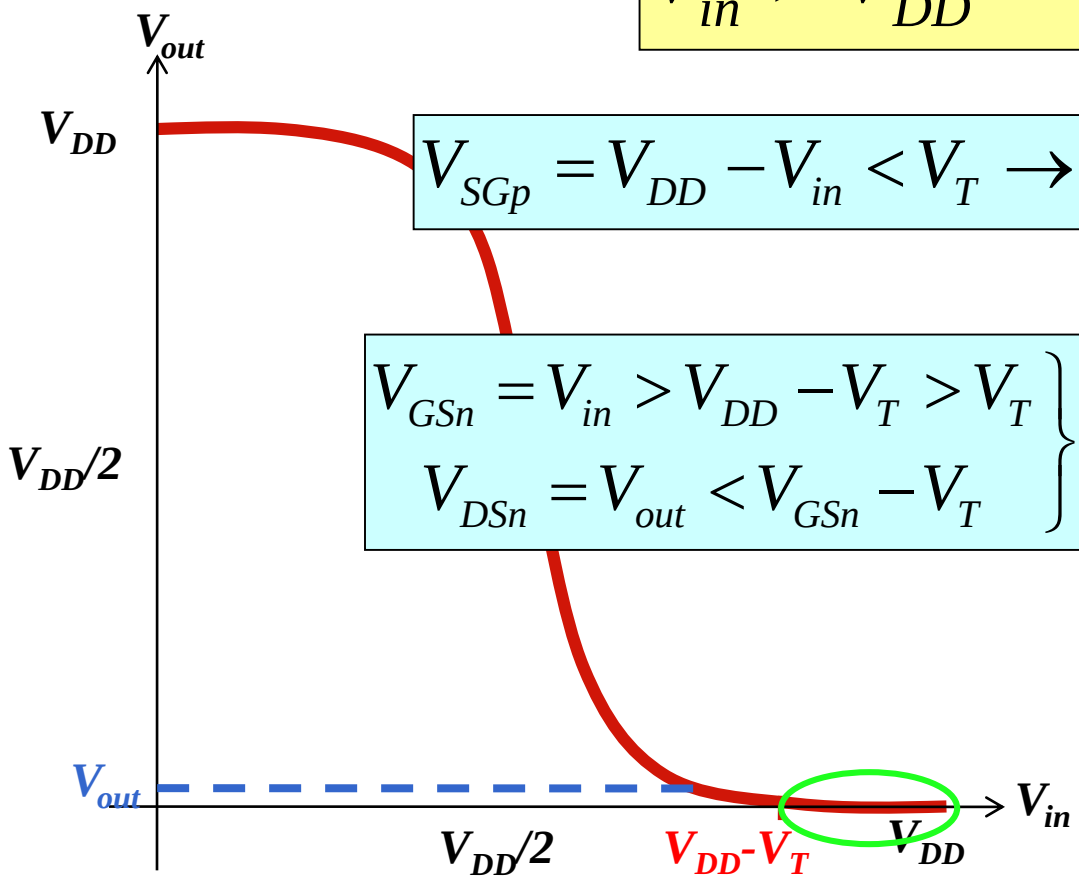
Operating Regions

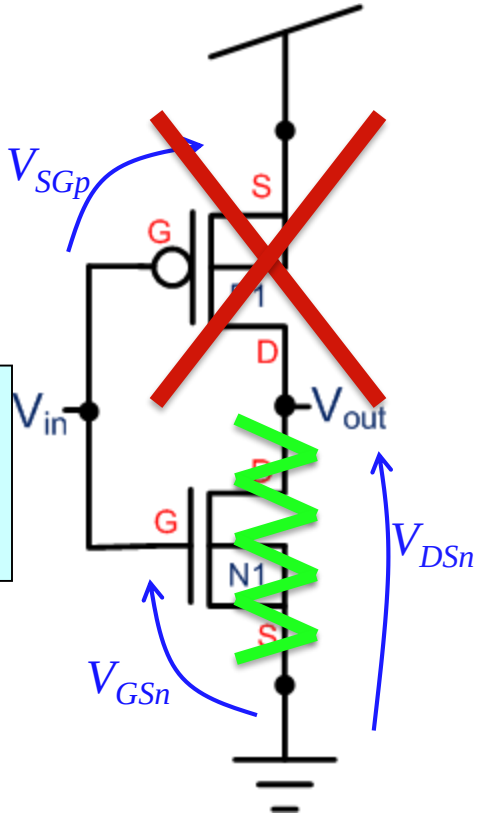
- So now, let's jump to the other side of the ***VTC***.

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

$$V_{SGp} = V_{DD} - V_{in} < V_T \rightarrow Cutoff$$

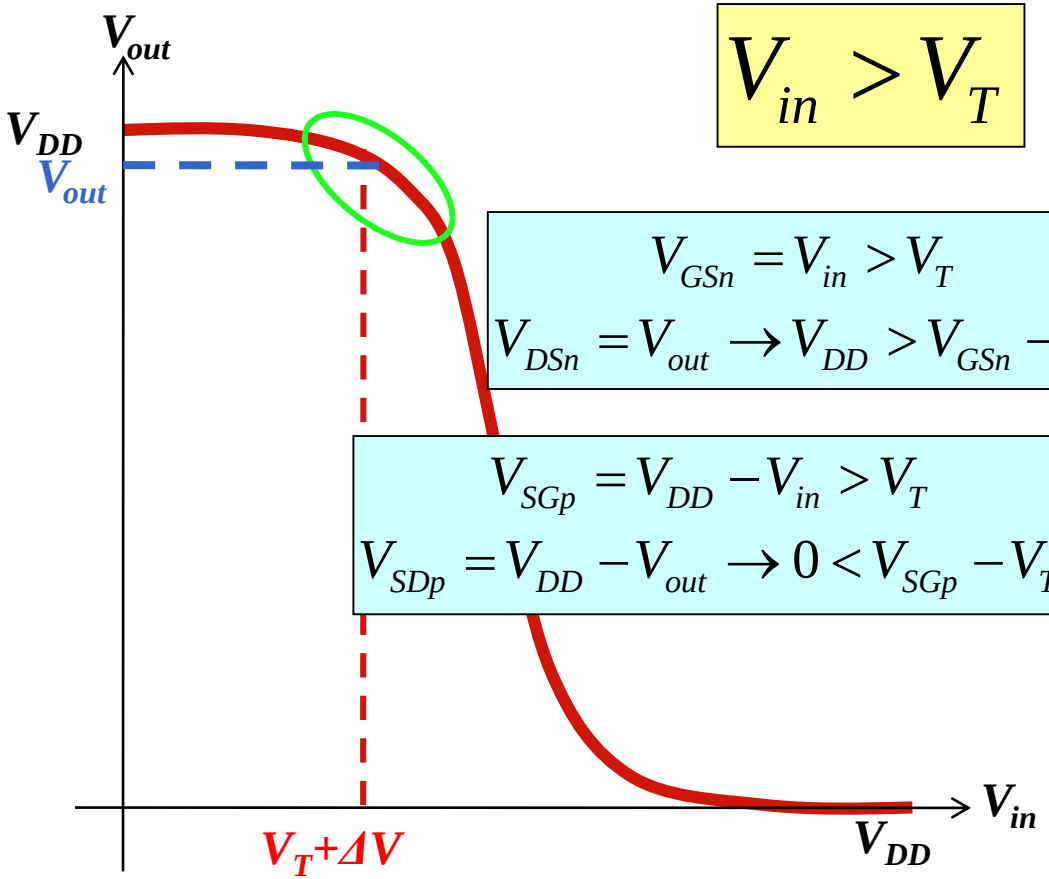
$$\left.\begin{aligned} V_{GSn} = V_{in} > V_{DD} - V_T > V_T \\ V_{DSn} = V_{out} < V_{GSn} - V_T \end{aligned}\right\} \rightarrow Linear$$





Operating Regions

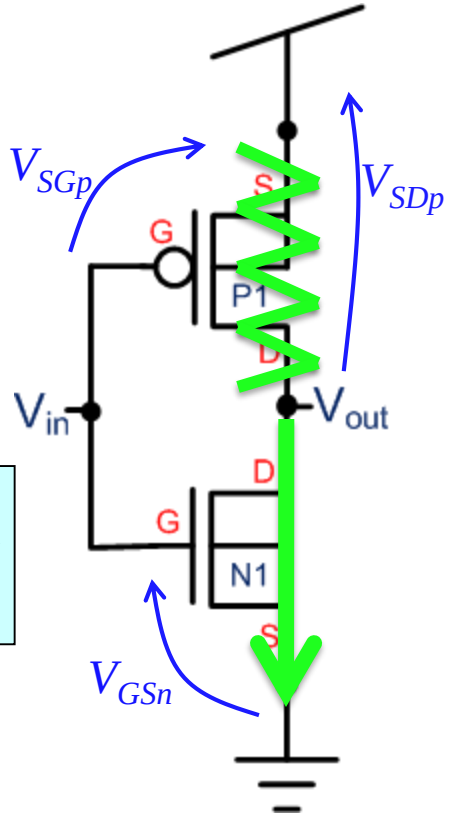
- Now, back to the ***VTC*** regions. Let's see what happens when we raise the input voltage slightly past V_T.



$$V_{in} > V_T$$

$$\left.\begin{aligned} V_{GSn} &= V_{in} > V_T \\ V_{DSn} &= V_{out} \rightarrow V_{DD} > V_{GSn} - V_T \end{aligned}\right\} \Rightarrow Sat$$

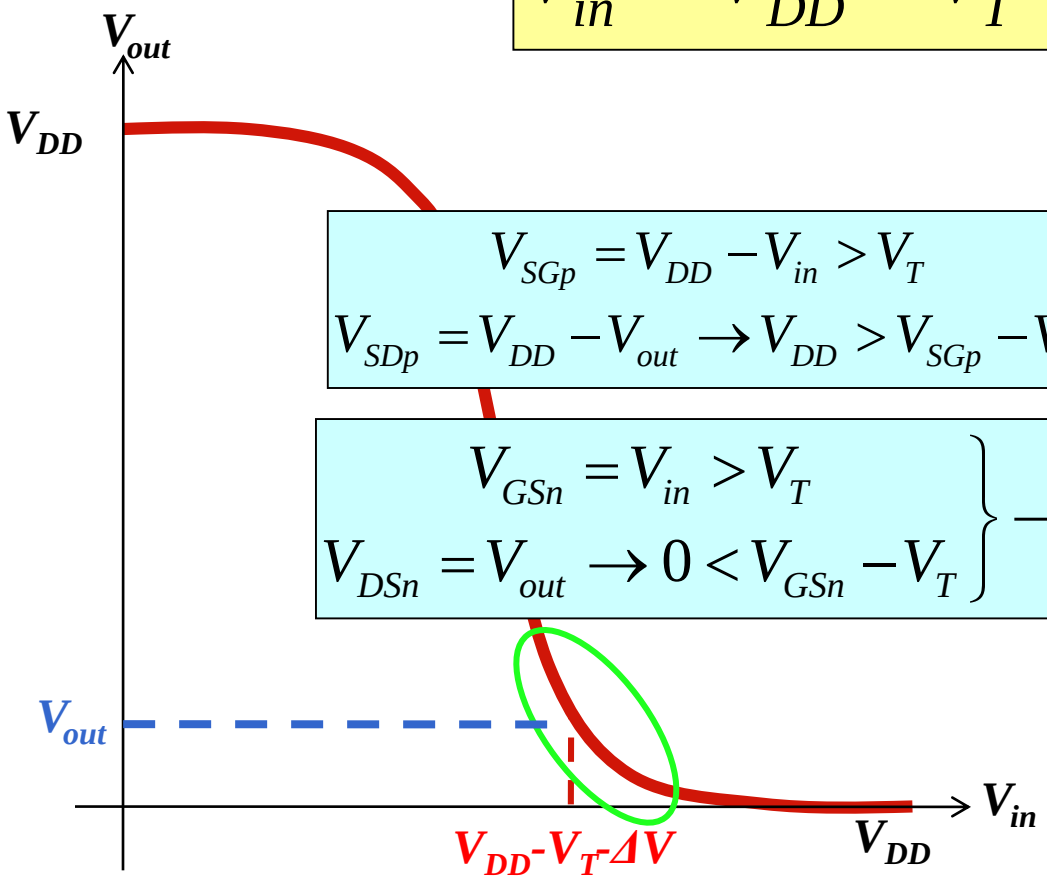
$$\left.\begin{aligned} V_{SGp} &= V_{DD} - V_{in} > V_T \\ V_{SDp} &= V_{DD} - V_{out} \rightarrow 0 < V_{SGp} - V_T \end{aligned}\right\} \rightarrow Linear$$



Operating Regions

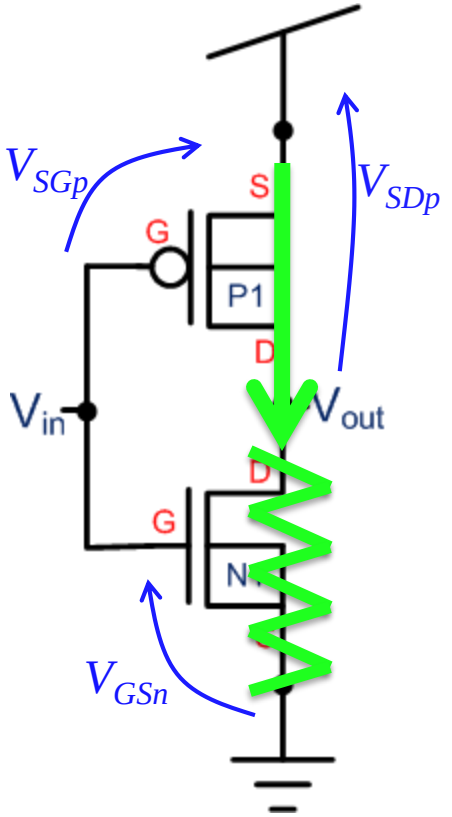
- The same when V_{in} is a bit more than V_T lower than V_{DD}.

$$V_{in} = V_{DD} - V_T - \Delta V$$



$$\left.\begin{aligned} V_{SGp} &= V_{DD} - V_{in} > V_T \\ V_{SDp} &= V_{DD} - V_{out} \rightarrow V_{DD} > V_{SGp} - V_T \end{aligned}\right\} \Rightarrow Sat$$

$$\left.\begin{aligned} V_{GSn} &= V_{in} > V_T \\ V_{DSn} &= V_{out} \rightarrow 0 < V_{GSn} - V_T \end{aligned}\right\} \rightarrow Linear$$



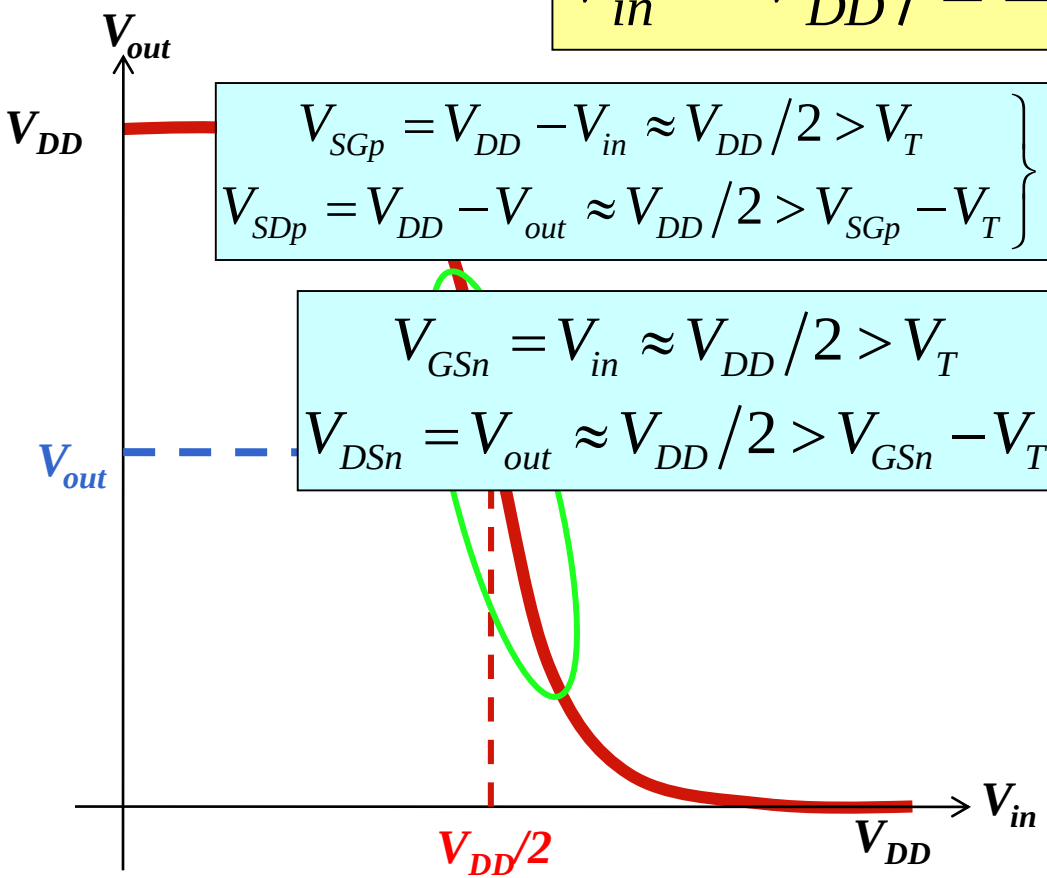
Operating Regions

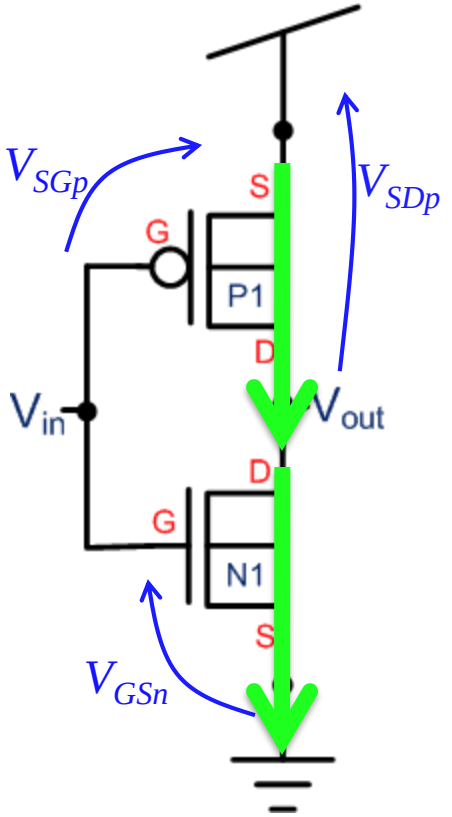
- Finally, we have the middle area, where:

$$V_{in} = V_{DD}/2 \pm \Delta V$$

$$\left.\begin{aligned} V_{SGp} &= V_{DD} - V_{in} \approx V_{DD}/2 > V_T \\ V_{SDp} &= V_{DD} - V_{out} \approx V_{DD}/2 > V_{SGp} - V_T \end{aligned}\right\} \Rightarrow Sat$$

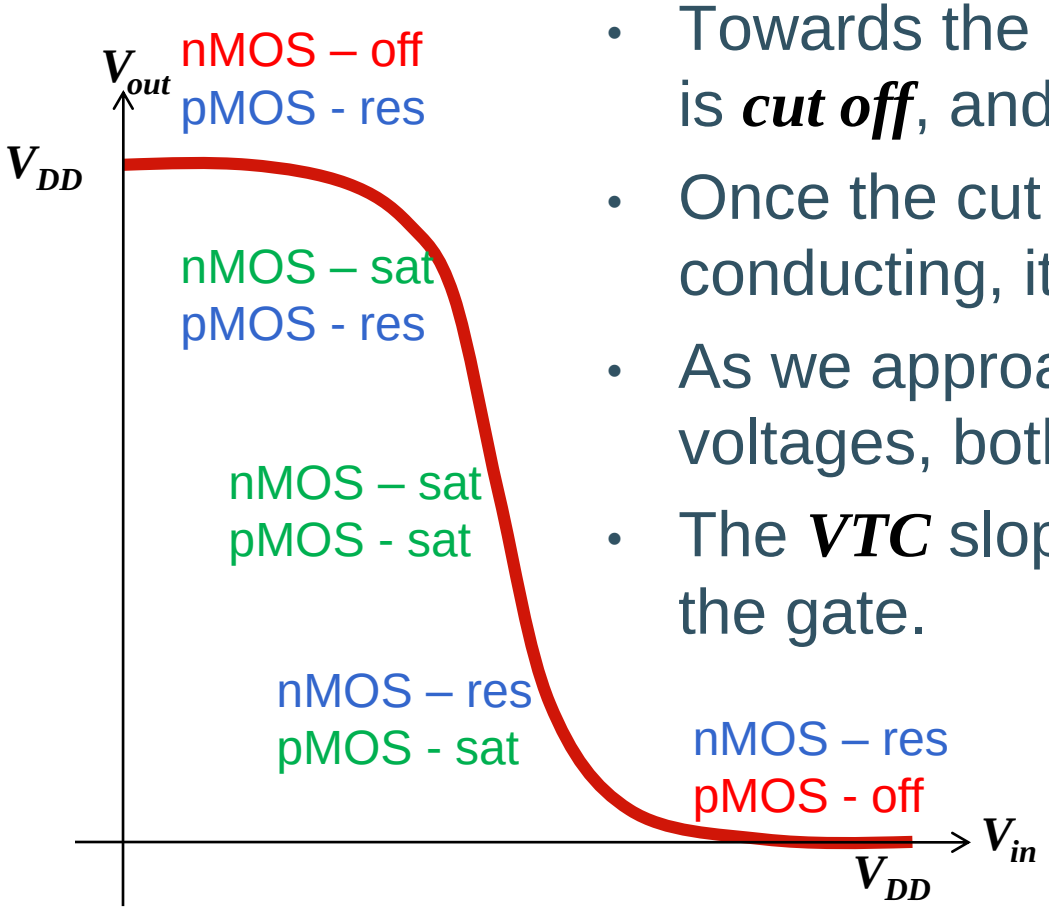
$$\left.\begin{aligned} V_{GSn} &= V_{in} \approx V_{DD}/2 > V_T \\ V_{DSn} &= V_{out} \approx V_{DD}/2 > V_{GSn} - V_T \end{aligned}\right\} \rightarrow Sat$$





Operating Regions

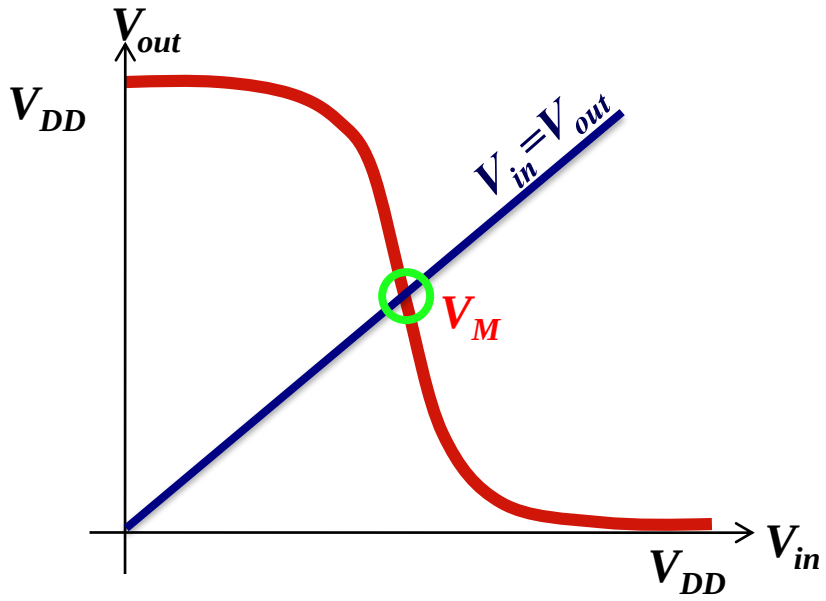
- To Sum it up:



- 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.

Switching Threshold

- The ***Switching Threshold***, V_M, is the point where $V_{in}=V_{out}$.
- This can be calculated:
  - » Graphically, at the intersection of the ***VTC*** with $V_{in}=V_{out}$



  - » Or analytically, by equating the ***nMOS*** and ***pMOS*** saturation currents with $V_{in}=V_{out}$.

Switching Threshold

- But let's start with the intuitive approach…

Switching Threshold

- Let's analytically compute V_M.
  - Remember, the saturation current for a MOSFET is given by:

$$I_{DS} = \frac{k}{2}\left(V_{GS} - V_T\right)^2\left(1+\lambda V_{DS}\right)$$

  - Lets assume $\lambda=0$ and we'll equate the two currents:

$$I_D = \frac{k_n}{2}\left(V_{GSn} - V_{Tn}\right)^2 = \frac{k_p}{2}\left(V_{SGp} - V_{Tp}\right)^2$$

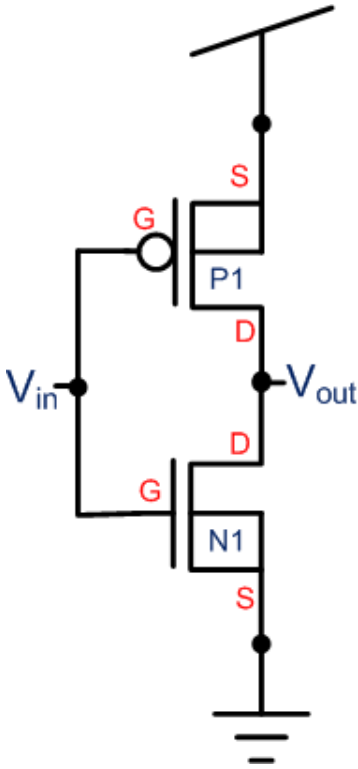
  - Now we'll substitute:

$$V_{GSn} = V_{in} = V_M \qquad V_{SGp} = V_{DD} - V_{in} = V_{DD} - V_M$$

  - And we'll arrive at:

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

$$r \triangleq \sqrt{\frac{k_p}{k_n}}$$

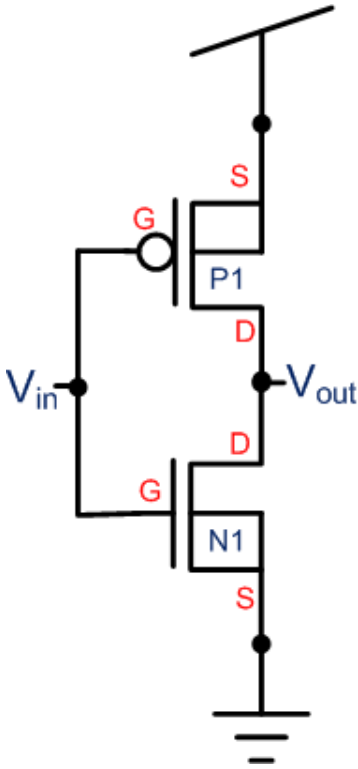


Switching Threshold

- As we can see, ***r*** is an important factor in setting the ***switching threshold***.
- ***r*** is a design parameter, that is set by the ***drive strength ratios*** of the ***nMOS*** and ***pMOS***:

$$r_{long\_channel} \triangleq \sqrt{\frac{k_p}{k_n}}$$

$$k \triangleq k' \frac{W}{L} = \mu C_{ox} \frac{W}{L}$$



- Using the current equations again, we can find the ***drive strength ratio*** for a desired V_M:

$$\frac{k_p}{k_n} = \left(\frac{V_M - V_{Tn}}{V_{DD} - V_M - V_{Tp}} \right)^2$$

Switching Threshold

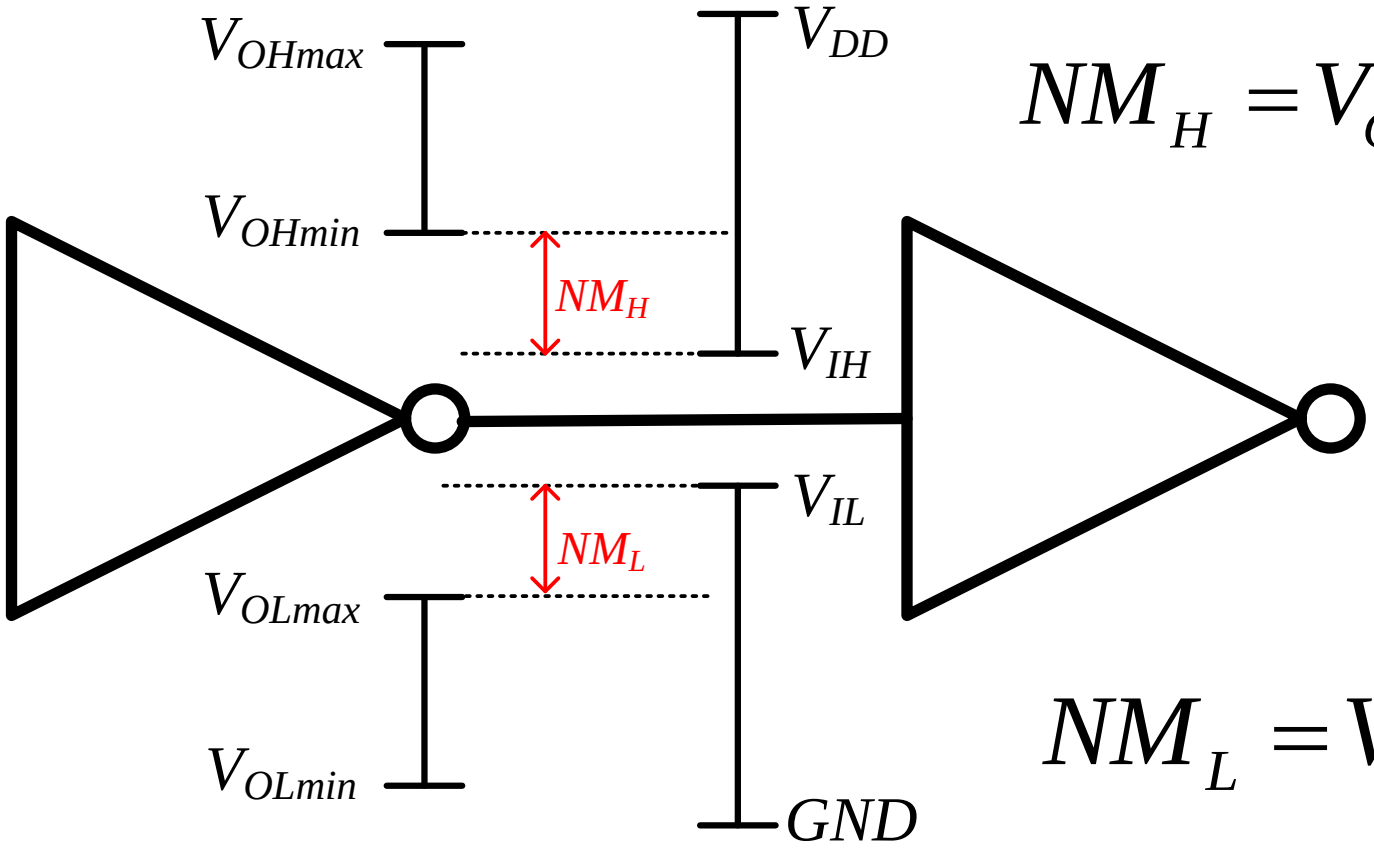
- A symmetric ***VTC*** ($V_M=V_{DD}/2$) is often desired. In this case:

$$V_M = \frac{V_{DD}}{2} = \frac{V_{Tn} + r\left(V_{DD} - V_{Tp}\right)}{1+r} \quad \rightarrow \quad \left(\frac{W}{L}\right)_p = \frac{\mu_n}{\mu_p}\left(\frac{W}{L}\right)_n$$

- Generally, the same length (L_{min}) is taken for all transistors in digital circuits, and so for a symmetric ***VTC***:

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

Reminder: Noise Margins



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

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

$$NM = \min\left(NM_L, NM_H\right)$$

Noise Margins

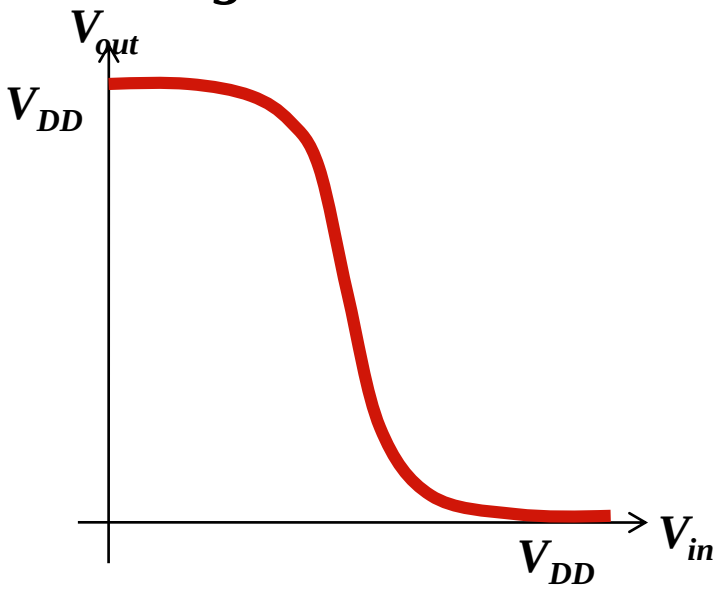
- One of the CMOS logic family's advantages is a ***Full Rail to Rail Swing***. In other words:

$$V_{OH\max} = V_{DD}$$

$$V_{OL\min} = GND$$

- To calculate the ***Noise Margins***, we will need to find V_{IL} and V_{IH}. These are the points where the ***gain*** is ***-1***.

- To do this we will equate the currents:
  - » V_{IL}→nMOS ***sat***, pMOS ***res***
  - » V_{IH}→nMOS ***res***, pMOS ***sat***



Noise Margins

- Let's calculate V_{IH}:

$$I_{DSn}(res) = k_n\left[(V_{GSn} - V_{Tn})V_{DSn} - \frac{V_{DSn}^2}{2}\right] = I_{DSp}(sat) = \frac{k_p}{2}(V_{GSp} - V_{Tp})^2$$

- Assuming matching devices ($k_n = k_p$, $V_{Tn} = V_{Tp}$):

$$(V_{IN} - V_T)V_{out} - \frac{V_{out}^2}{2} = \frac{1}{2}(V_{DD} - V_{IN} - V_T)^2$$

- Differentiating and equating **-1**, we reach:

$$V_{IH} = \frac{1}{8}(5V_{DD} - 2V_T)$$

- Doing the same for V_{IL} or using symmetry, we reach:

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

- Accordingly, for a matched ***long-channel*** device, and assuming $V_{OHmin} \rightarrow V_{OHmax}$ and $V_{OLmax} \rightarrow V_{OLmin}$ in ***CMOS***, we get:

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

Noise Margins

- The previous analysis “assumed” many things and therefore SHOULD NOT be memorized.
- Let us look at the noise margins intuitively to try and understand trade offs:

Summary of Static Properties

- When Vin<VT or Vin>VDD-VT, one of the networks (PUN/PDN) is off, providing Rail to Rail Swing.
- The *skew* of the VTC is set by the sizing ratio between the PUN and PDN.
- Analytic Noise Margin calculation is rigorous and approximation should be used when possible.

4.3

Now that we see how the inverter behaves in steady state, we will analyze it's transient:

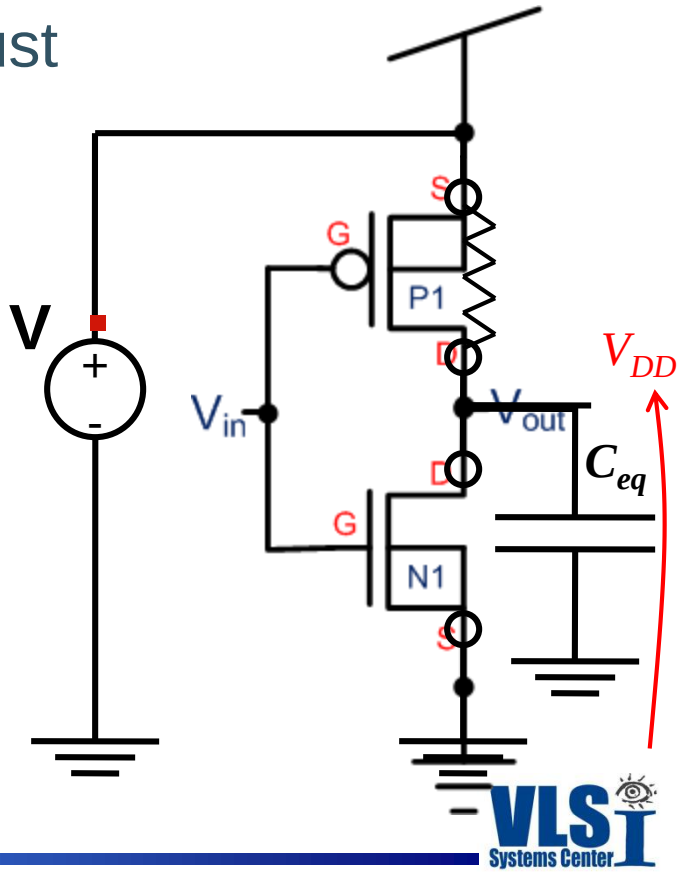
DYNAMIC OPERATION

Reminder: Dynamic Properties

- Propagation Delay
- Rise/Fall Time

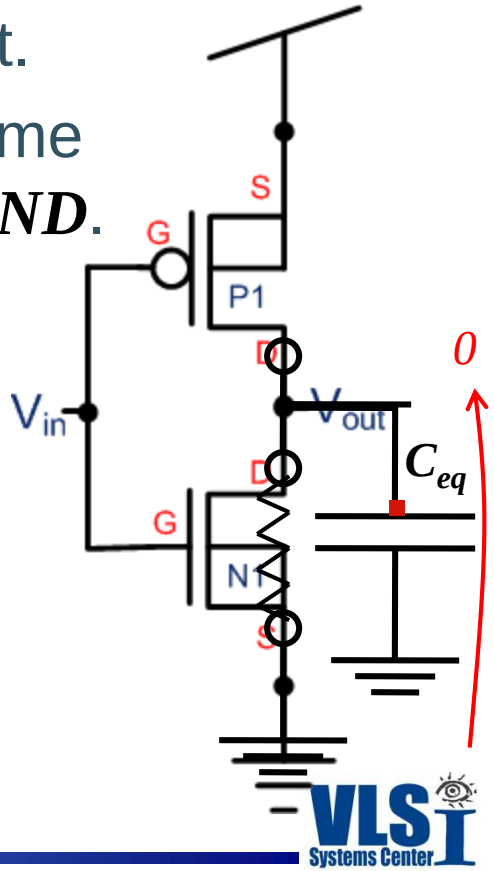
Parasitic Capacitances

- Remember that our transistors have ***capacitance*** connected to the ***output node***.
- We'll calculate the capacitance values in the next lecture, but for now, let's just use and ***equivalent output capacitance***.
- When the input is low our ***pMOS*** is a non-linear ***resistor*** and our ***nMOS*** is ***cut off***, so we get a ***simple RC circuit***.
- Our capacitance is ***charged***, bringing the output voltage to V_{DD}.



Parasitic Capacitances

- When the input is high, we essentially have closed the top switch and opened the bottom one.
- This creates a ***resistive path*** from the capacitor to ***GND***, and blocks the path from the supply to the output.
- Again we have an ***RC network***, though this time we are just ***discharging*** the capacitance to ***GND***.
- We end up with an output equal to ***GND***.



Parasitic Capacitances

- So we saw that, a switching CMOS inverter ***charges and discharges*** a parasitic output capacitance.
- During the switching process, we can create a model that will transform the circuit into a ***simple RC network***.
- In this way, we can easily derive a first order analysis of the CMOS dynamic operation for ***propagation delay*** and ***power consumption*** calculation.

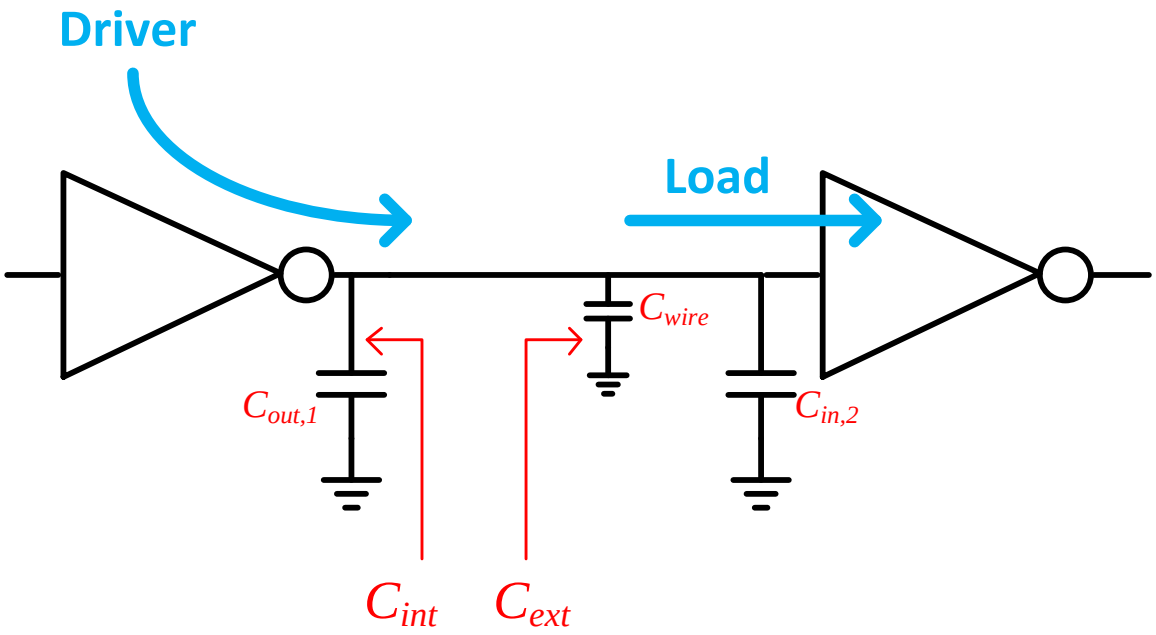
Parasitic Capacitances

- For now, let us make the following assumptions:
  - A transistor has a gate capacitance that is proportional to its area (W*L)
  - A transistor has a diffusion capacitance that is proportional to its width (W)

Parasitic Capacitances

- For now we will use a very simple model to represent all the parasitic capacitances between the output node and ground.

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



Propagation Delay

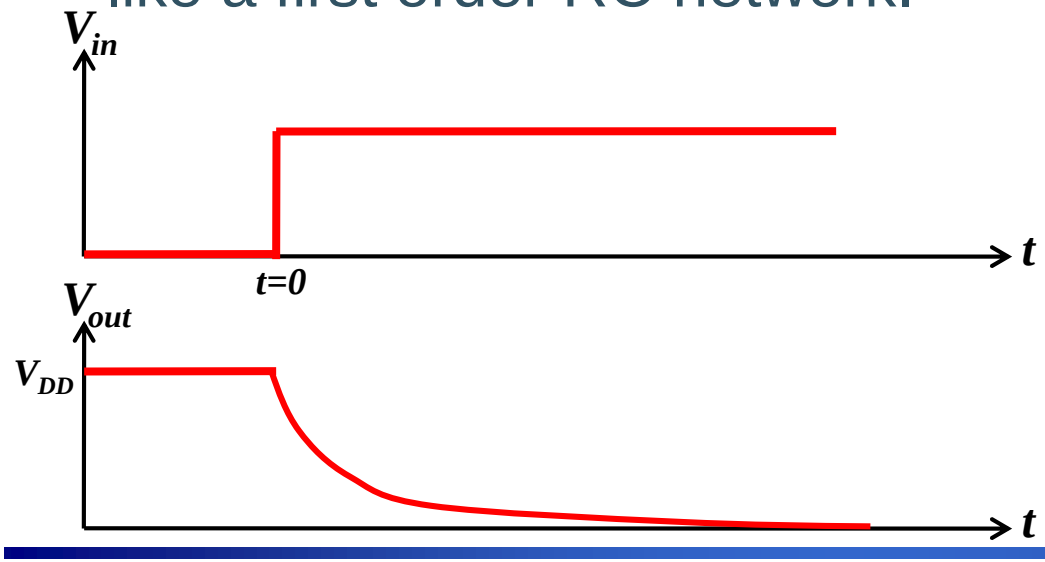
- Using our simple model for ***load capacitance***, we can write:

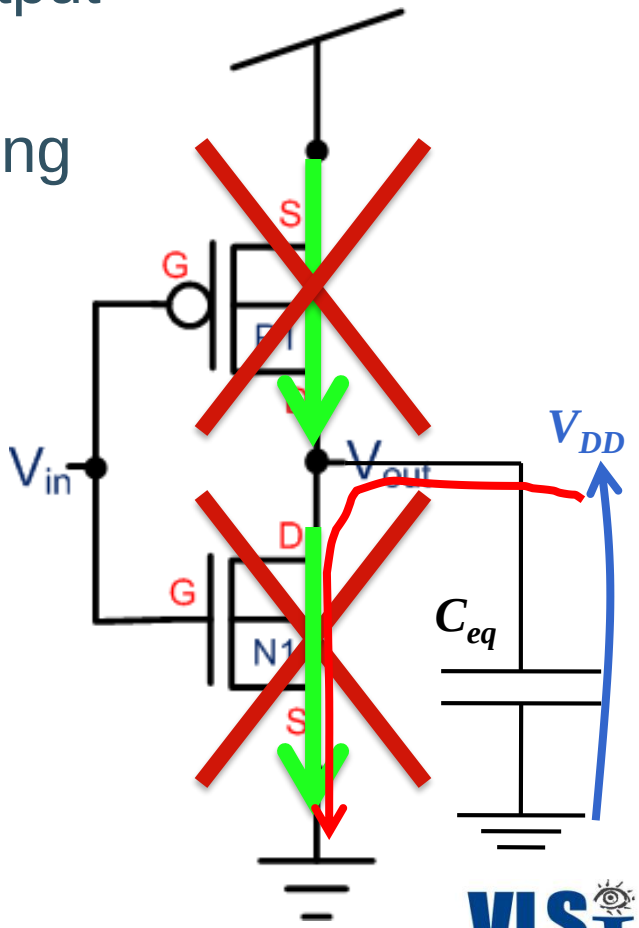
$$\int_{t_1}^{t_2} dt = \int_{v_1}^{v_2} \frac{C_{load}\left(v_{out}\right)}{i\left(v_{out}\right)} dv_{out}$$

- Assuming an ***ideal step*** at the input, the ***propagation delay***, t_{pd} is the time it takes the output to (dis)charge ***50%*** of its voltage.
- We will look at three ways to calculate the propagation delay:
  - By solving the integral above.
  - By approximating the average current
  - By using equivalent resistance

t_{pd} – solving the integral

- We'll start with $V_{in}=0$. Accordingly, ***P1*** is open and ***N1*** is closed, causing the output voltage to be held at $V_{out}=V_{DD}$
- At $t=0$, V_{in} changes from 0 to V_{DD}, closing ***P1*** and opening ***N1***.
- This causes the output to discharge like a first order RC network.





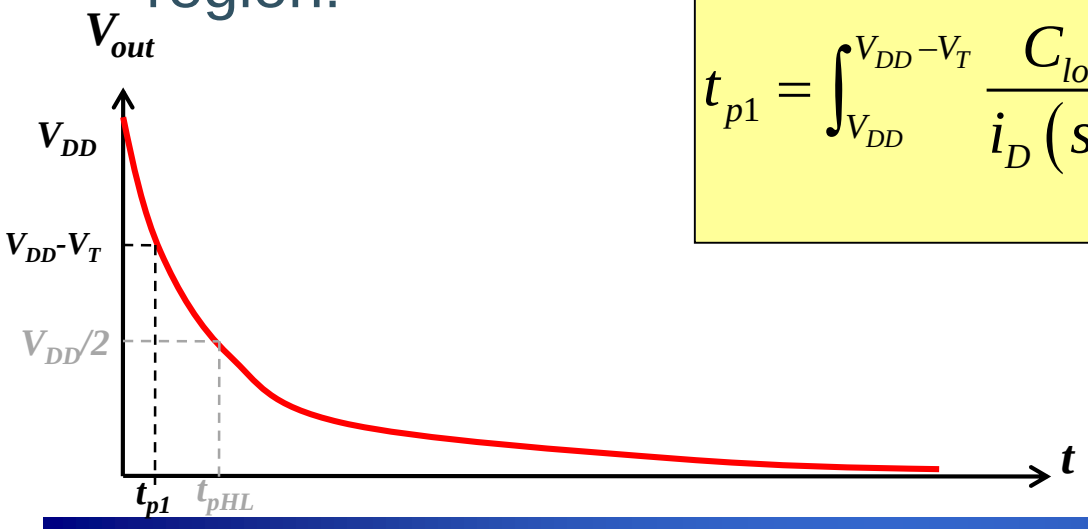
t_{pd} – solving the integral

- At this point $V_{DSn}=V_{DD}>V_{GS}-V_T$, so ***N1*** is in ***saturation*** and the discharge current is given by:

$$I_{DSn}=\frac{k_n}{2}\left(V_{In}-V_T\right)^2\left(1+\lambda V_{out}\right)$$

- Assuming $\lambda=0$ and integrating until ***N1*** enters the ***linear*** region:

$$t_{p1}=\int_{V_{DD}}^{V_{DD}-V_T}\frac{C_{load}}{i_D\left(sat\right)}dv=\frac{C\left[V_{DD}-\left(V_{DD}-V_T\right)\right]}{\frac{k_n}{2}\left(V_{DD}-V_T\right)^2}$$



t_{pd} – solving the integral

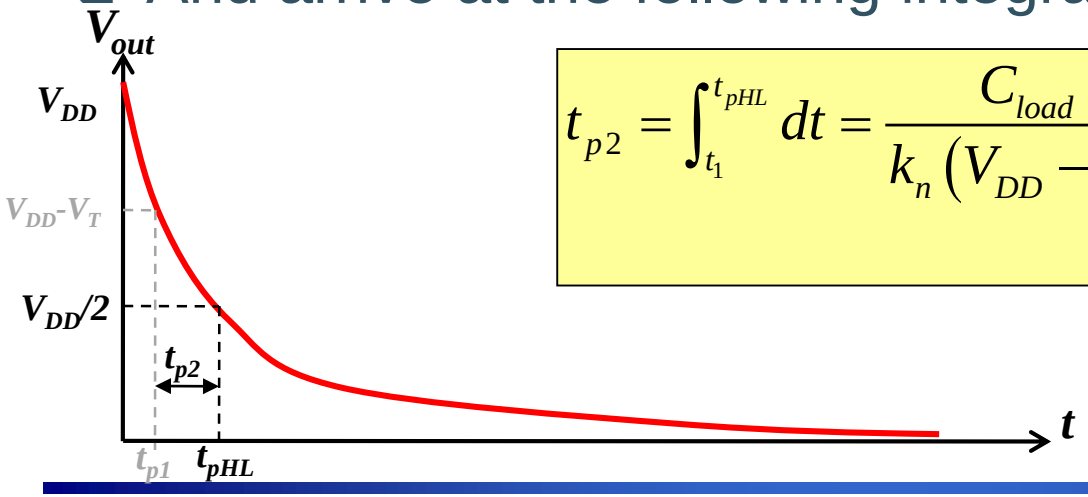
- Now $V_{DSn} < V_{GS} - V_T$ and so ***N1*** goes into ***linear*** operation with:

$$I_{DSn}(res) = k_n\left[(V_{in} - V_T)V_{out} - \frac{V_{out}^{\ 2}}{2}\right]$$

- We'll write an expression for the change in voltage:

$$dV_{out} = -\frac{i_{DSn}}{C_{load}}dt = -\frac{k_n}{C_{load}}\left[(V_{in} - V_T)V_{out} - \frac{V_{out}^{\ 2}}{2}\right]$$

- And arrive at the following integral:



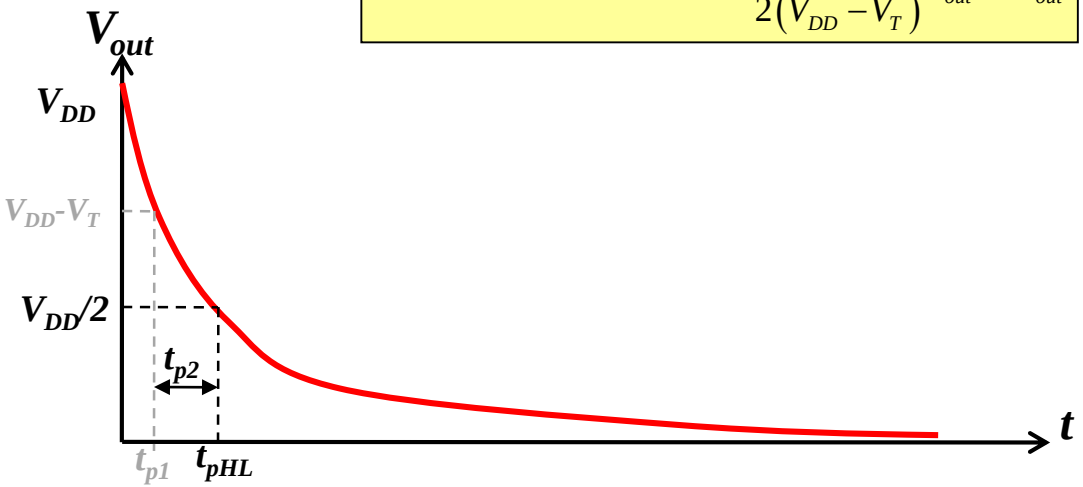
$$t_{p2} = \int_{t_1}^{t_{pHL}} dt = \frac{C_{load}}{k_n(V_{DD} - V_T)} \int_{V_{DD}-V_T}^{V_{DD}/2} \frac{dV_{out}}{\frac{1}{2(V_{DD} - V_T)}V_{out}^{\ 2} - V_{out}}$$

t_{pd} – solving the integral

$$t_{p2} == \frac{C_{load}}{k_n\left(V_{DD}-V_T\right)}\int_{V_{DD}-V_{TH}}^{V_{DD}/2}\frac{dV_{out}}{\frac{1}{2\left(V_{DD}-V_T\right)}V_{out}^{\ 2}-V_{out}}$$

- Using the equality:

$$\int\frac{dx}{ax^2-x}=\ln\left(1-\frac{1}{ax}\right)$$



- We get:

$$t_{p2}=\frac{C_{load}}{k_n}\ln\left(\frac{3V_{DD}-4V_T}{V_{DD}}\right)$$

- And putting together the two parts, we get:

$$t_{pHL}=t_{p1}+t_{p2}=\frac{2C_{load}}{k_n\left(V_{DD}-V_T\right)}\left[\frac{V_T}{V_{DD}-V_T}+\frac{1}{2}\ln\left(\frac{3V_{DD}-4V_T}{V_{DD}}\right)\right]$$

- For $V_T=0.2V_{DD}$, we get:

$$t_{pHL}=\frac{1.6C_{load}}{k_nV_{DD}}$$

t_{pd} – average currents

- Instead of solving the integral, we can sometimes assume a linear change in current to make life much easier.
- This assumption just lets us find the average current between ***t=0*** and ***t=t_{pHL}*** and calculate:

$$t_{pHL} \cong \frac{C_{load}\Delta V}{i_{DSn}\big|_{avg}} = \frac{C_{load}\dfrac{V_{DD}}{2}}{\dfrac{1}{2}\left[i_{DSn}(0) + i_{DSn}(t_{pHL})\right]}$$

t_{pd} – average current

- Example
  - Step input into a CMOS inverter with a minimum sized nMOS driving a 1.5fF load.
- t_0
  - VDS=VDD, VGS-VT=VDD-VT

$$V_{DSeff} = \min\left(V_{DD}, V_{DD} - V_{Tn}, V_{DSatn}\right) = V_{DSat}$$

$$I_{DS}\left(t_0\right) = k_n\left[V_{GTn}V_{DSatn} - 0.5V_{DSatn}^2\right]\left(1+\lambda V_{DSn}\right)$$

$$= 115\mu\frac{0.18\mu}{0.18\mu}\cdot\left(1.37\cdot 0.63 - 0.5\cdot 0.63^2\right)\left(1+0.06\cdot 1.8\right) = 84.7\mu A$$

- t_{pd}
  - VDS=VDD/2, VGS-VT=VDD-VT

$$V_{DSeff} = \min\left(V_{DD}/2, V_{DD} - V_{Tn}, V_{DSatn}\right) = V_{DSat}$$

$$I_{DS}\left(t_0\right) = k_n\left[V_{GTn}V_{DSatn} - 0.5V_{DSatn}^2\right]\left(1+\lambda V_{DSn}\right)$$

$$= 115\mu\frac{0.18\mu}{0.18\mu}\cdot\left(1.37\cdot 0.63 - 0.5\cdot 0.63^2\right)\left(1+0.06\cdot 0.9\right) = 80.6\mu A$$

t_{pd} – average current

» The average current is:

$$I_{Avg} = 0.5\left(I_{DS}\left(t_0\right) + I_{DS}\left(t_{pd}\right)\right) = \frac{84.7\mu + 80.6\mu}{2} = 82.65\mu A$$

» So we can find the delay:

$$t_{pd} = C_L \frac{V\left(t_0\right) - V\left(t_{pd}\right)}{I_{Avg}} = 1.5 f \frac{0.9}{82.65\mu} = 16.33 ps$$

t_{pd} – equivalent resistance

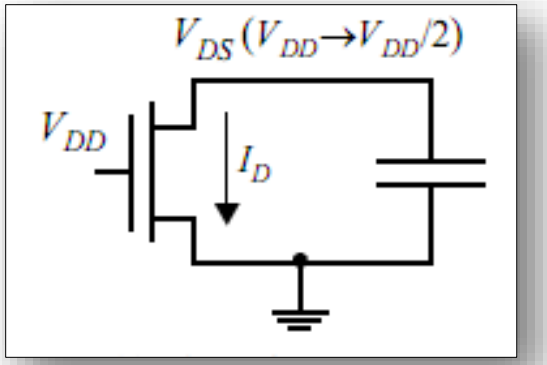
- A good way to estimate the propagation delay is by finding the resistance of the MOSFET during the transition and using this resistance in quick calculations.
- The primary approach to deriving such an ***equivalent resistance*** is to calculate the transistor's average resistance throughout its operation (***ON***) period.

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

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

t_{pd} – equivalent resistance

- Now, we will calculate the ***propagation delay*** of a ***short channel*** inverter, by using the equivalent resistance to discharge a capacitor from V_{DD} to $V_{DD}/2$.



- Assuming $V_{DD} >> V_{DSATn}$, we can assume that throughout the propagation delay, the transistor is velocity saturated.

- We can write:

$$R_{eq} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{V_{DS}}{I_{DSAT}\left(1+\lambda V_{DS}\right)} dV_{DS} \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)$$

$$I_{DSAT} = k_n\left[\left(V_{DD}-V_T\right)V_{DSAT} - \frac{V_{DSAT}^{\,2}}{2}\right]$$

t_{pd} – equivalent resistance

- For example, if VDD=1.8V:

$$I_{DSATn} = k_n \left[V_{GTn} V_{DSatn} - 0.5 V_{DSatn}^2 \right]$$

$$= 115\mu \frac{0.18\mu}{0.18\mu} \cdot \left(1.37 \cdot 0.63 - 0.5 \cdot 0.63^2 \right) = 76.43\mu A$$

$$R_{eqn} \approx \frac{3}{4} \frac{V_{DD}}{I_{DSAT}} \left(1 - \frac{7}{9} \lambda V_{DD} \right)$$

$$= \frac{3}{4} \frac{1.8}{76.43\mu A} \left(1 - \frac{7}{9} 0.06 \cdot 1.8 \right) = 16.18k\Omega$$

t_{pd} – equivalent resistance

- So all we need is to use our "magic" equation:

$$t_{pdHL} = 0.69 R_{eqn} C_L = 0.69 \cdot 16.18k \cdot 1.5f = 16.75ps$$

- Remember that for t_{pd}, we also need R_{eqp} for:

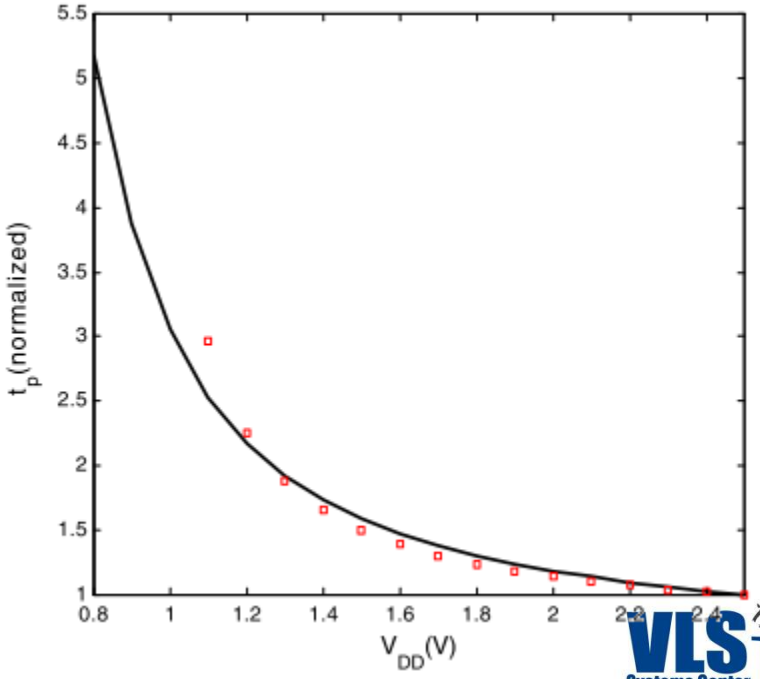
$$t_{pd} \triangleq \frac{t_{pLH} + t_{pHL}}{2} = \frac{0.69 C_{load} \left(R_{eqp} + R_{eqn} \right)}{2}$$

t_{pd} – equivalent resistance

- To calculate and analyze the parameters that affect the propagation delay, we will take $\lambda=0$ and get:

$$t_{pHL} = 0.69 C_{load} \frac{3}{4} \frac{V_{DD}}{I_{DSATn}} = 0.52 \frac{V_{DD} C_{load}}{k_n V_{DSATn} \left(V_{DD} - V_{Tn} - V_{DSAT}/2 \right)}$$

- Accordingly, we can minimize the delay in the following ways:
  - » Minimize C_{load}.
  - » Increase W/L
  - » Increase V_{DD}



Last Lecture

- CMOS Inverter
  - Intuitive Explanation
  - VTC
  - VM
  - Noise Margins
  - Propagation Delay

Affect of Device Sizing

- So we saw that to reduce the propagation delay, we need to increase the device sizes (W/L).
- But how much should we increase them? What are the tradeoffs?
- For this, we will discuss two sizing parameters:
  - Beta Ratio (β)
  - Upsizing Factor (S)

What happens when we upsize a transistor?

- Effective resistance decreases:

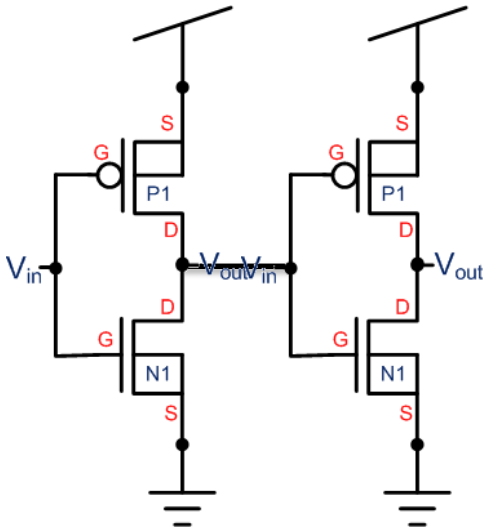
- Gate and Drain Capacitance increase:

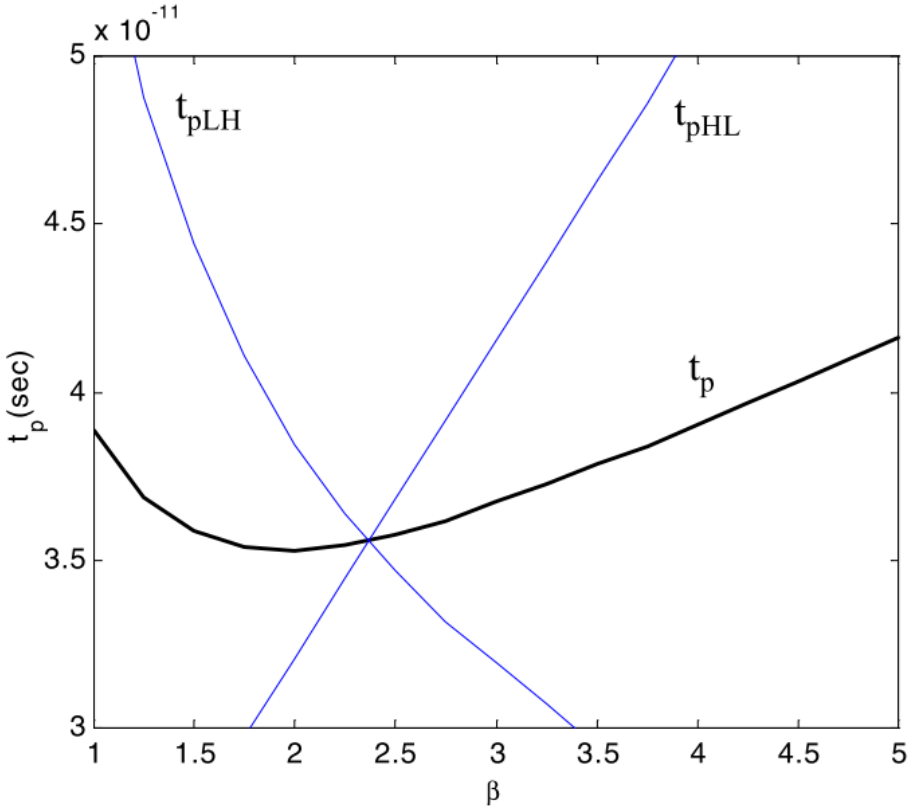
Device Sizing - β

- Device Sizing is the Width to Length ratio (W/L) of the transistor.
- When discussing a CMOS logic gate, we relate to the ***pMOS/nMOS*** ratio $((W_p/L_p)/(W_n/L_n))$.
  - » We will call this ratio β.
- To get a *balanced* inverter (i.e. $V_m=V_{DD}/2$) we usually will need $\beta=3\text{-}3.5$, mainly due to the mobility ratio of holes and electrons.
- 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.

Device Sizing - β

- How can this be?
  - To get a faster t_{pLH}, we need to enlarge the ***pMOS*** width.
  - This increases the parasitic capacitance (C_{load}), degrading t_{pHL}.

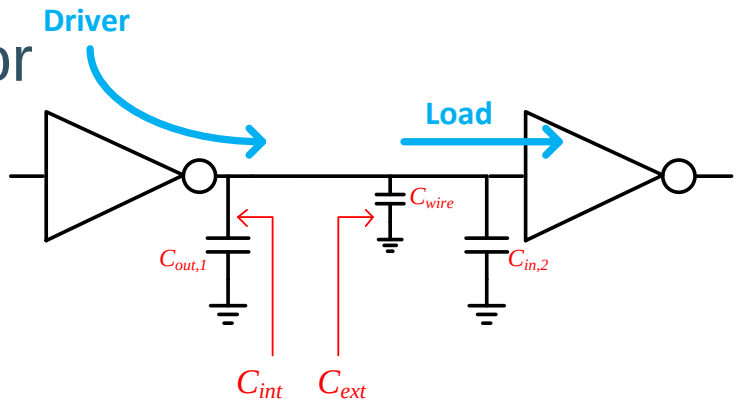




Device Sizing - β

Device Sizing - β

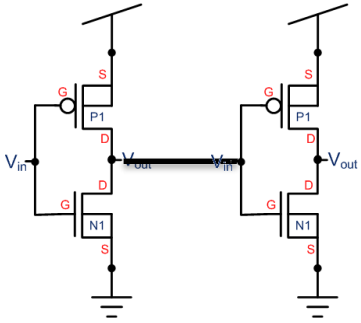
- We will now find the optimum ratio for sizing an inverter, considering two identical cascaded CMOS inverters.



- The load capacitance of the driving gate is:

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

- Assuming the input capacitance is the *gate capacitance* of the transistors (C_g) and the output capacitance is the *drain capacitance* (C_d), we can write:



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

- Assuming a linear dependence on device size, we get:

$$C_{load} = (1+\beta)\left(C_{dn1} + C_{gn2}\right) + C_{wire}$$

$$\beta \triangleq \frac{(W/L)_p}{(W/L)_n}$$

Device Sizing - β

- Noting that we have reduced the equivalent resistance of the ***pMOS*** by ***β***, we can write the first order ***RC*** propagation delay:

$$t_{pd} = \frac{0.69C_{load}}{2}\left(R_{eqn} + R_{eqp}\right) = 0.345\left[(1+\beta)\left(C_{dn1} + C_{gn2}\right) + C_{wire}\right]\left(R_{eqn} + \frac{R_{eqp}}{\beta}\right)$$

- Now we just need to find the minimum:

$$\frac{dt_{pd}}{d\beta} = 0$$

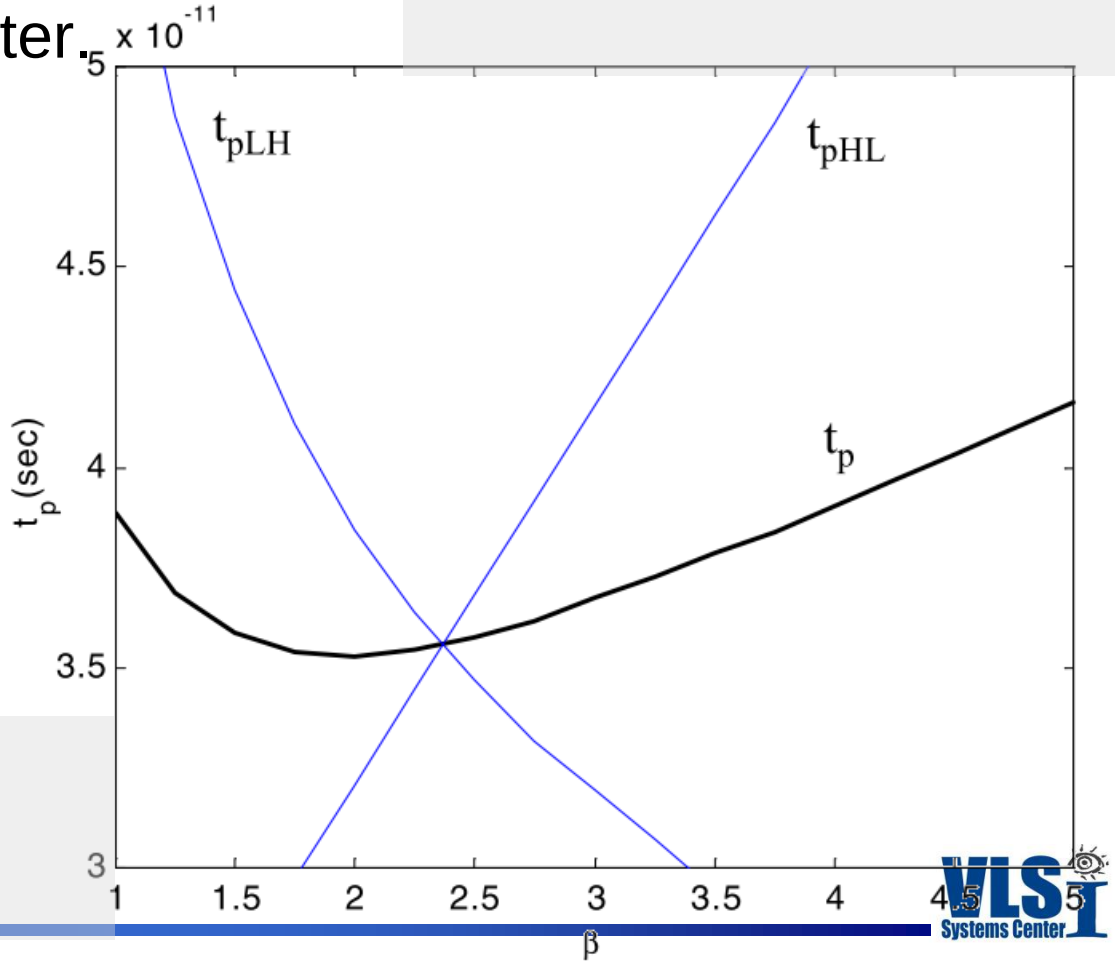
$$\beta_{opt} = \sqrt{\frac{R_{eqp}}{R_{eqn}}\left(1 + \frac{C_{wire}}{C_{dn1} + C_{gn2}}\right)}\Bigg|_{\left(C_{dn1}+C_{gn2} >> C_{wire}\right)} \rightarrow \sqrt{\frac{R_{eqp}}{R_{eqn}}}$$

- A typical optimum for ***β*** is usually around ***2***.

Device Sizing - β

- Conclusions:
  - A balanced inverter isn't usually the fastest possible inverter.
  - A typical optimal ***pMOS/nMOS*** ratio for performance is given by:

$$\beta_{opt} \approx \sqrt{\frac{R_{eqp}}{R_{eqn}}} \approx 2$$



Device Sizing - S

- We saw how the ratio between the ***pMOS*** and ***nMOS*** can be optimized to improve performance.
- Now, we will take a balanced inverter and see how upsizing affects the ***intrinsic*** or ***unloaded delay***.
- We will start by writing the delay as a function of the ***intrinsic capacitance*** (diffusion and overlap) and the ***extrinsic capacitances*** (fanout and wiring):

$$C_{load} = C_{int} + C_{ext}$$

$$t_{pd} = 0.69 R_{eq} C_{load} = 0.69 R_{eq} \left(C_{\text{int}} + C_{ext} \right)$$

Device Sizing - S

- Now, we will mark the minimal intrinsic delay as t_{p0}. This is the delay of a minimum sized balanced inverter only loaded by its own intrinsic capacitance ($C_{ext}=0$):

$$t_{p0} \triangleq 0.69 R_{ref} C_{ref}$$

- We will now mark the ***sizing factor***, S.
  This is the relative upsizing of the inverter,
  i.e. $C_{int}=SC_{ref}$ and accordingly $R_{eq}=R_{ref}/S$.

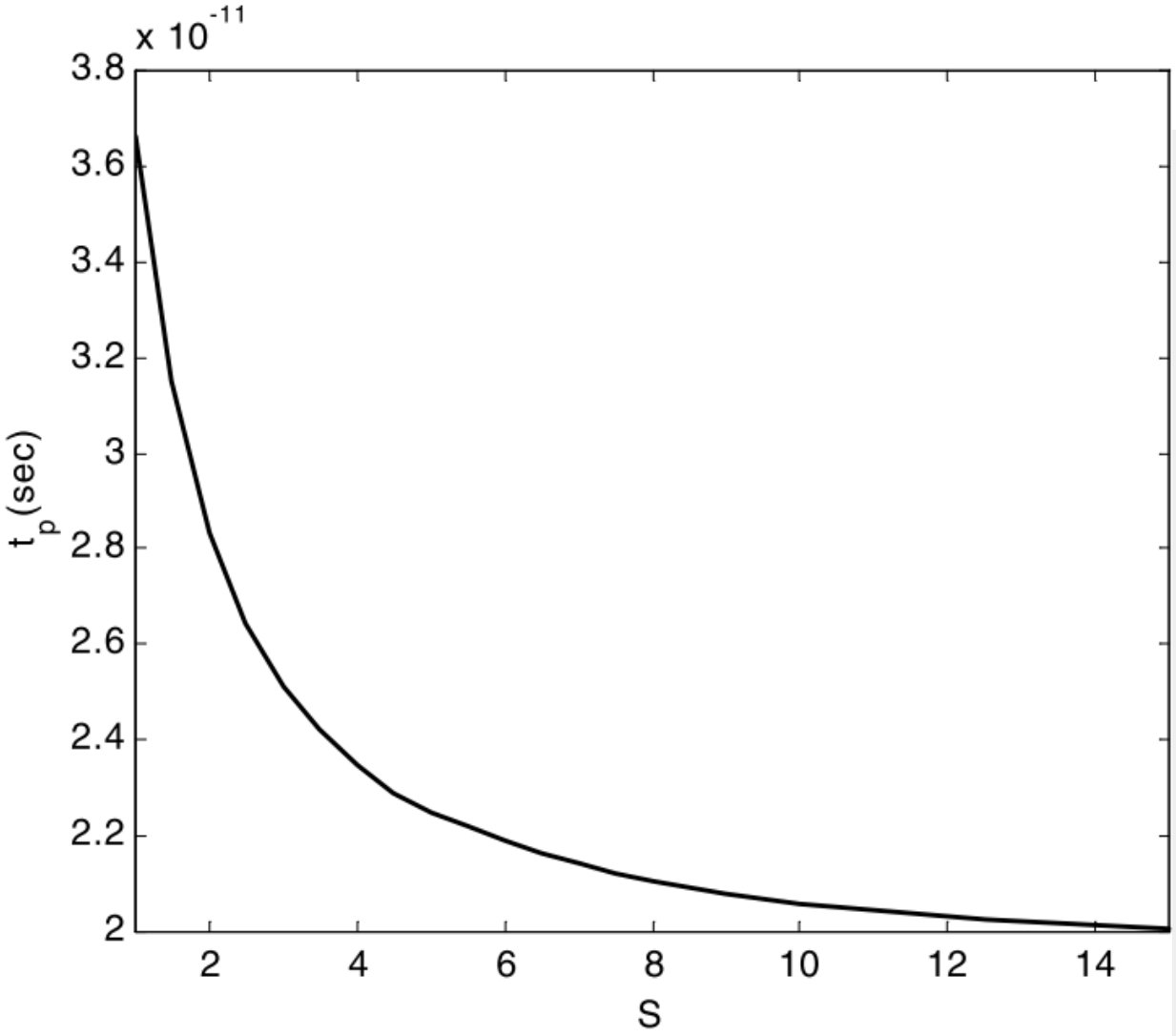
- Now we can write the delay of an upsized inverter:

$$t_{pd} = 0.69 R_{eq}\left(C_{\text{int}} + C_{ext}\right) = 0.69 \frac{R_{ref}}{S}\left(SC_{ref} + C_{ext}\right)$$

$$= 0.69 R_{ref} C_{iref}\left(1 + \frac{C_{ext}}{SC_{ref}}\right) = t_{p0}\left(1 + \frac{C_{ext}}{SC_{ref}}\right)$$

Device Sizing - S

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



Device Sizing - S

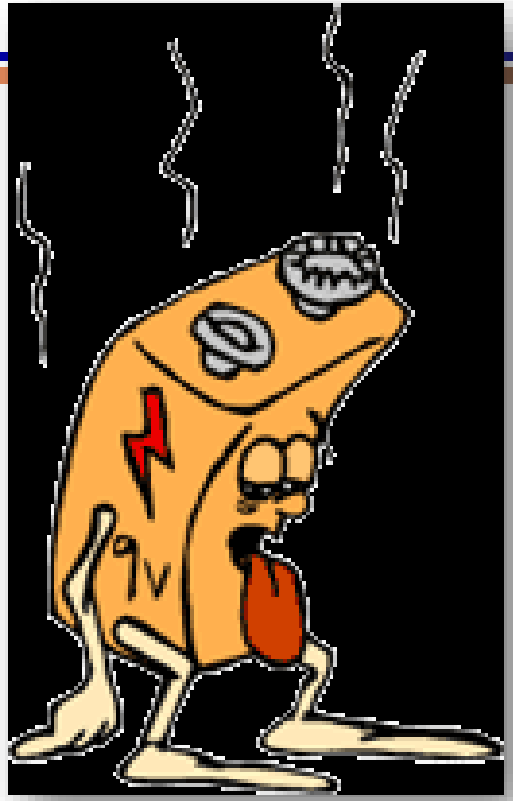
$$t_{pd} = t_{p0}\left(1 + \frac{C_{ext}}{SC_{ref}}\right)$$

- Conclusions:
  - The intrinsic delay of an inverter (t_{p0}) 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.
  - To minimize a loaded inverter's delay, S should be enlarged, but at the expense of a substantial gain in area.

Summary of Dynamic Parameters

- We can calculate t_{pd} in several ways, but the easiest is to measure the equivalent resistance during a typical transition.
- One of our main techniques to improve the delay is through transistor sizing, which we discussed in two fashions:
  - Setting the optimal ratio between the PUN/PDN.
  - Upsizing the gate to deal with a large output load.

4.4

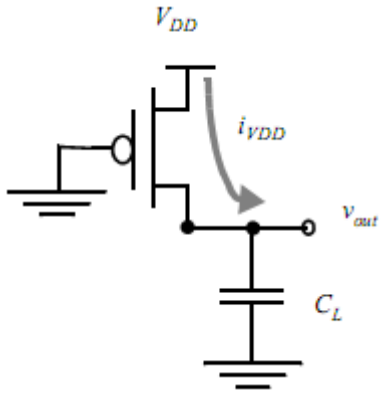


And now that we fully understand the static and dynamic operation of the CMOS Inverter, it's time to take a look at

POWER CONSUMPTION

Dynamic Power

- Assuming an ideal step input, we can analyze the energy consumed from the supply of the equivalent circuit during a ***Low-to-High*** transition is given by:



$$E_{V_{DD}} = \int_0^\infty i_{V_{DD}}(t) V_{DD} dt = V_{DD} \int_0^\infty C_{load} \frac{dV_{out}}{dt} dt = C_{load} V_{DD} \int_0^{V_{DD}} dV_{out} = C_{load} {V_{DD}}^2$$

- Now, looking at the energy stored in the load capacitance, we get:

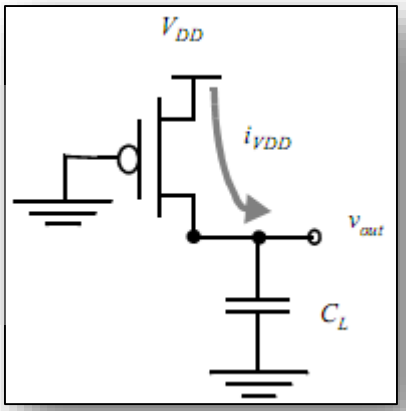
$$E_{Charge} = \int_0^\infty i_{V_{DD}}(t) V_{out} dt = \int_0^\infty C_{load} \frac{dV_{out}}{dt} V_{out} dt = C_{load} \int_0^{V_{DD}} V_{out} dV_{out} = \frac{C_{load} {V_{DD}}^2}{2}$$

Dynamic Power

$$E_{V_{DD}} = C_{load} V_{DD}^{\ 2}$$

$$E_{Charge} = \frac{C_{load} V_{DD}^{\ 2}}{2}$$

- Analyzing these results, we see that the energy required to charge the output capacitance is ***twice*** the energy stored on the capacitor at the end of the transition.



- This is relatively surprising and very important. It means that ***half of the energy*** was wasted on the ***pMOS*** resistance ***independent of its size***!

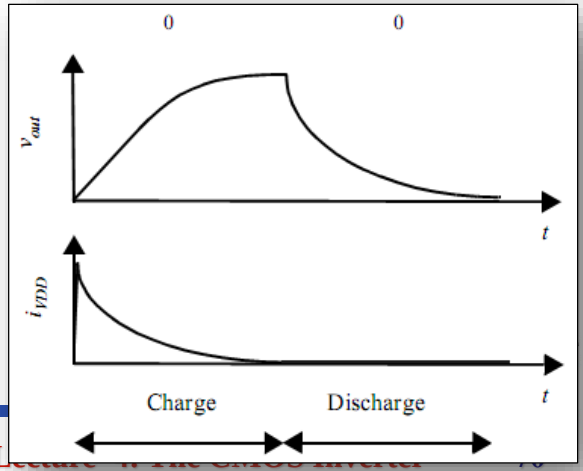
Dynamic Power

- Assuming the input changes now from ***zero*** to ***one***, we now get a ***High-to-Low*** transition.
- Here, the supply is disconnected, and the charge stored on the capacitance flows through the ***nMOS*** to the ground.
- The energy dissipated is the total energy stored on the output capacitance, as no charge is left:

$$E_{discharge} = \int_0^{\infty} i_{DSn}(t) V_{out} dt = \int_0^{\infty} C_{load} \frac{dV_{out}}{dt} V_{out} dt = C_{load} \int_{V_{DD}}^{0} V_{out} dV_{out} = \frac{C_{load} V_{DD}^{\ 2}}{2}$$

- The sum of the charge and discharge energy is obviously equal to the energy supplied:

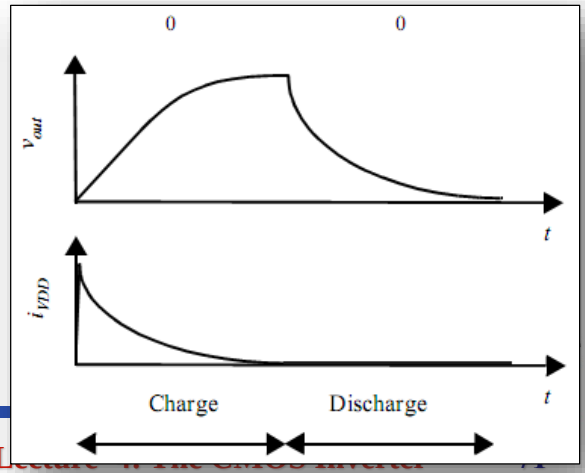
$$E_{charge} + E_{discharge} = E_{V_{DD}} = C_{load} V_{DD}^{\ 2}$$



Dynamic Power

- Conclusions:
  - » Each charge and discharge cycle dissipates a fixed amount of energy, independent of the size of the device.
  - » The effective energy is the charge stored on the capacitance. The rest of the energy is wasted as heat burned on the ***pMOS*** (charge) and ***nMOS*** (discharge) resistance.
- To compute the ***Power Dissipation***, we calculate the total energy wasted per one second.
- For a circuit that completes a ***Low-to-High*** transition $f_{0\to1}$ times per second (and therefore a ***High-to-Low*** transition as well…), the dynamic power consumption is:

$$P_{dyn} = C_{load} V_{DD}^{\ 2} \cdot f_{0\to1}$$



Dynamic Power

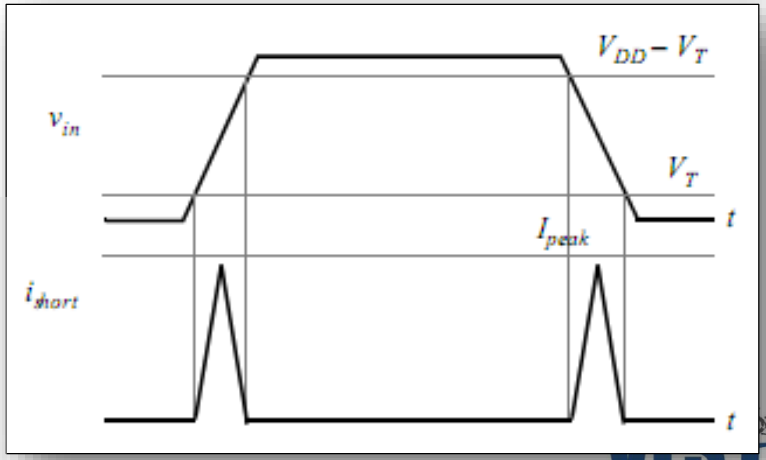
- We said that the ***Power Dissipation*** is a factor of the switching frequency of the gate.
- But the gate only switches when its input changes. In other words, the ***switching activity*** is smaller than the circuit frequency.
- We can rewrite the ***Dynamic Power*** expression using the ***activity factor***, α of the inverter, expressing the probability of the output to switch:

$$P_{dyn} = C_{load} V_{DD}^{\ 2} \cdot f \cdot \alpha = C_{eff} V_{DD}^{\ 2} f$$

- C_{eff} is the ***effective capacitance*** of a complex circuit, describing the average capacitance that actually switches each cycle.

Short Circuit Power

- During the above analysis, the input was an ***ideal step function***, immediately closing one transistor when the other was opened.
- In a real circuit, the input signal has a ***non-zero rise/fall time***, resulting in a time interval with both the ***pMOS*** and ***nMOS*** transistors open.
- This provides a direct path from V_{DD} to ***GND***, with a current known as ***Short Circuit*** current.
- This is shown in the following waveforms:



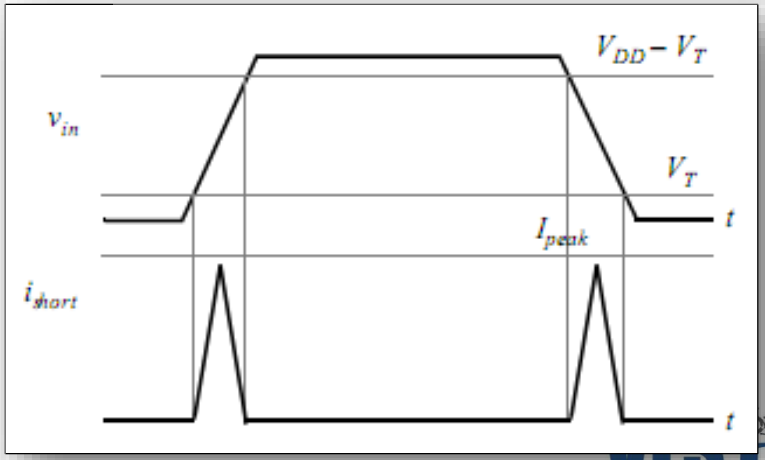
Short Circuit Power

- The energy dissipated via ***short circuit power*** is the area under the triangles:

$$E_{sc} = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak}$$

- And the ***average power consumption*** is:

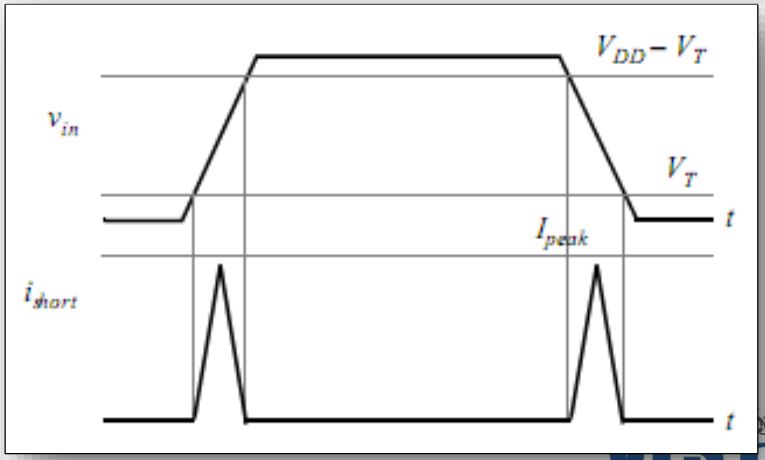
$$P_{sc} = t_{sc}V_{DD}I_{peak}f$$



Short Circuit Power

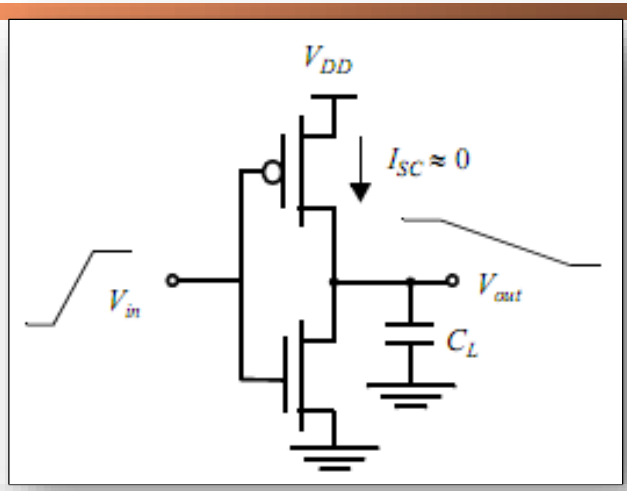
- The short circuit interval, t_{sc}, is the margin between the threshold voltages of the transistors.
- Assuming a linear input transition:

$$t_{sc} = \frac{V_{DD} - 2V_T}{V_{DD}} \cdot \frac{t_{rise(fall)}}{0.8}$$



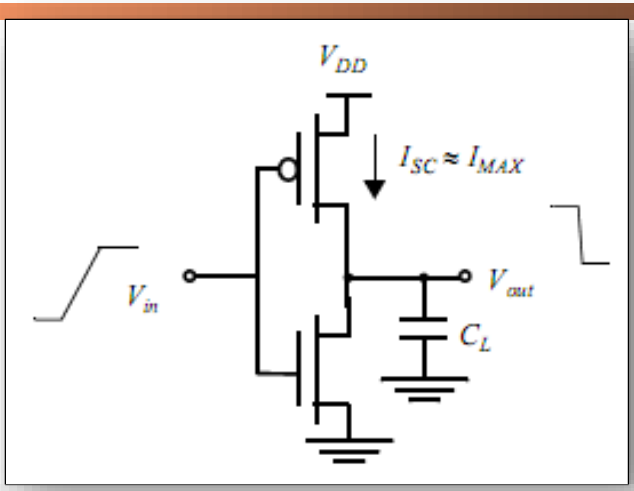
Short Circuit Power

- What affects the short circuit power?
  - » A ***short input rise time*** with ***a large output capacitance*** (large fall time) minimizes short circuit power, as the peak current is very small.



Short Circuit Power

- What affects the short circuit power?
  - ***small output capacitance*** relative to the ***input rise time*** causes extensive short circuit power, as the peak current is maximal (saturation current of the transistors).



- Conclusion:
  - Try to keep the input and output rise/fall times similar to maximize performance and minimize short circuit power.

Static Power

- Ideally, a MOSFET transistor is a ***perfect switch***.
- In such a case, a CMOS inverter never has a conducting path in steady state, resulting in ***no static power dissipation***.
- However, in reality, MOSFET transistors ***never completely turn off***.

Static Power

- Since static power is ***constantly consumed***, the power dissipation can be simply expressed as:

$$P_{static} = I_{static} V_{DD}$$

- Sources of static power are beyond the scope of this course, however they are quickly becoming the ***dominant source of power*** in advanced sub-micron technologies.

- For further probing, see these subjects:
  - Subthreshold current
  - Hot Electrons
  - DIBL
  - Punchthrough

Total Power Consumption

- As we saw above, there are three components to the ***power dissipation*** of a CMOS inverter:
  - ***Dynamic Power***
  - ***Short Circuit Power***
  - ***Static Power***

- Putting them all together, we get the ***total power consumption*** of a CMOS logic gate:

$$P_{total} = P_{dyn} + P_{sc} + P_{static} = \alpha f C_{load} V_{DD}^{\ 2} + \alpha f V_{DD} I_{peak} t_{sc} + V_{DD} I_{static}$$

Total Power Consumption

- We previously learned that the ***power-delay-product*** (***PDP***) measures the average energy of a switching event:

$$PDP = P_{dyn} \cdot t_{pd} = C_{load} V_{DD}^{\ 2} f_{\max} t_{pd} = \frac{1}{2} C_{load} V_{DD}^{\ 2}$$

- Since both ***Power*** and ***PDP*** give a clear advantage to energy reduction versus performance, we measure the ***energy-delay-product*** (***EDP***) as a combined measurement of the two:

$$EDP = PDP \cdot t_{pd} = \frac{1}{2} C_{load} V_{DD}^{\ 2} t_{pd}$$

4.5

Okay, enough with the inverter. But before we go on, let's go over a short

SUMMARY

Summary

- The CMOS inverter is characterized by:
  - A ***pMOS*** Pull-Up device and an ***nMOS*** pull down device.
  - The ***pMOS*** is usually wider due to inferior current driving.
  - An almost ideal ***VTC*** with a ***full rail to rail swing***.
  - ***Noise margins*** of a balanced inverter are close to $V_{DD}/2$
  - The steady state response is not affected by ***fanout***.
- Propagation delay:
  - Can be approximated as:

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

  - Small loads make faster drivers.
  - Widening the transistors improves the delay.
- Power Dissipation:
  - Dominated by ***dynamic power***, given by:

$$P_{dyn} = \alpha f C_{load} V_{DD}^{\ 2}$$

  - ***Short circuit power*** can be reduced by equating input/output slopes.
  - ***Static Power*** is a problem out of the scope of this course.