6.012 - Microelectronic Devices and Circuits

Lecture 15 - Digital Circuits: CMOS - Outline

- **Announcements**
  - One supplemental reading on Stellar
  - Exam 2 - Thursday night, Nov. 5, 7:30-9:30
- **Review - Inverter performance metrics**
  - **Transfer characteristic:** logic levels and noise margins
  - **Power:** $P_{ave,\ static} + P_{ave,\ dynamic}$ $(= I_{ON}V_{DD}/2 + f\ C_L V_{DD}^{\ 2})$
  - **Switching speed:** charge thru pull-up, discharge thru pull-down
    - **If can model load as linear C:** $dv_{OUT}/dt = i_{CH}(v_{OUT})/C_L$; $= i_{DCH}(v_{OUT})/C_L$
    - **If can say i_{CH}, i_{DCH} constant:** $\tau_{HI\text{-}LO} = C_L(V_{HI}\text{-}V_{LO})/I_{CH}$; $\tau_{HI\text{-}LO} = C_L(V_{HI}\text{-}V_{LO})/I_{DCH}$
  - **Fan-out, fan-in** (often only 10 to 90% swings)
  - **Manufacturability**
- **CMOS**
  - Transfer characteristic
  - Gate delay expressions
  - Power and speed-power product
- **Velocity Saturation**
  - General comments
  - Impact on MOSFET and Inverter Characteristics

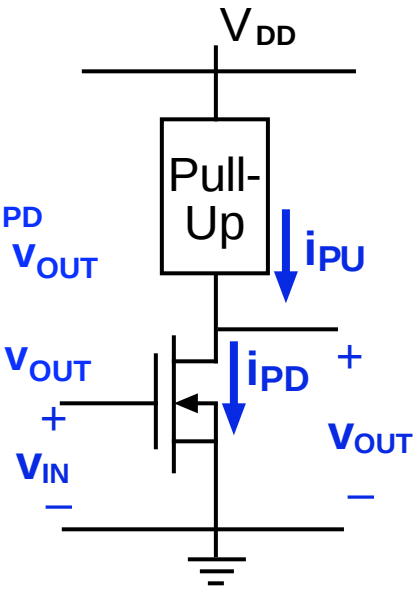
Transfer characteristic

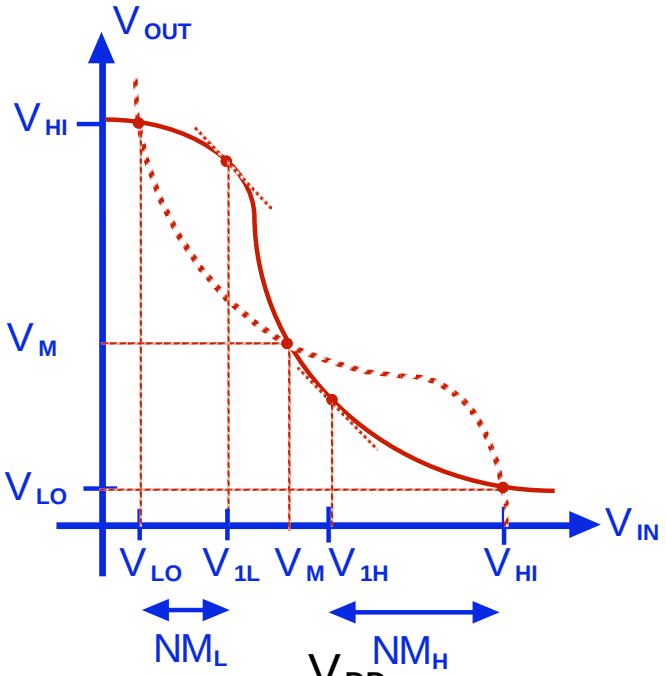
Node equation: $i_{PD} = i_{PU}$

$$i_{PD} = \begin{cases} 0 & \text{for } v_{IN} < V_{T,PD} \\ K_{PD}(v_{IN}-V_{T,PD})^2/2 & \text{for } v_{IN}-V_{T,PD} < v_{OUT} \\ K_{PD}(v_{IN}-V_{T,PD}-v_{OUT}/2)\,v_{OUT} & \text{for } v_{IN}-V_{T,PD} > v_{OUT} \end{cases}$$

i_{PU}: Depends on the device used

Gives us: V_{HI} and V_{LO}
NM_L and NM_H





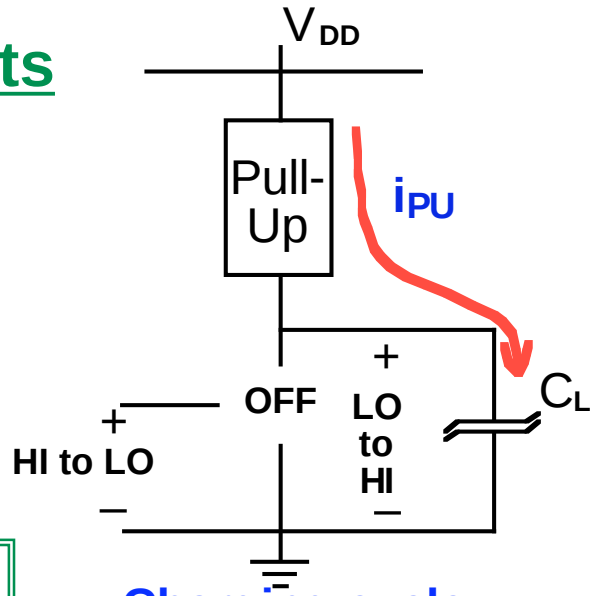
Switching transients

General approach:

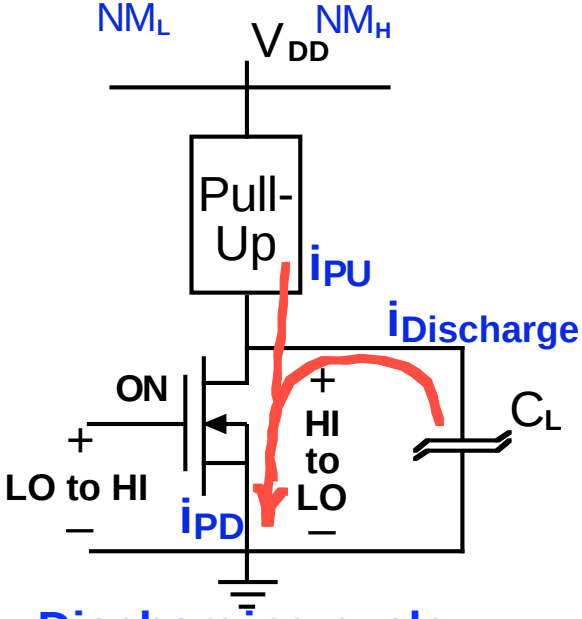
The load, C_L, is a non-linear charge store, but for MOSFETs it is fairly linear and it is useful to think linear:

$$dv_{out}/dt \approx i_{CL}/C_L$$

Bigger current → faster v_{OUT} change



Charging cycle:
$i_{Charge} = i_{PU}$

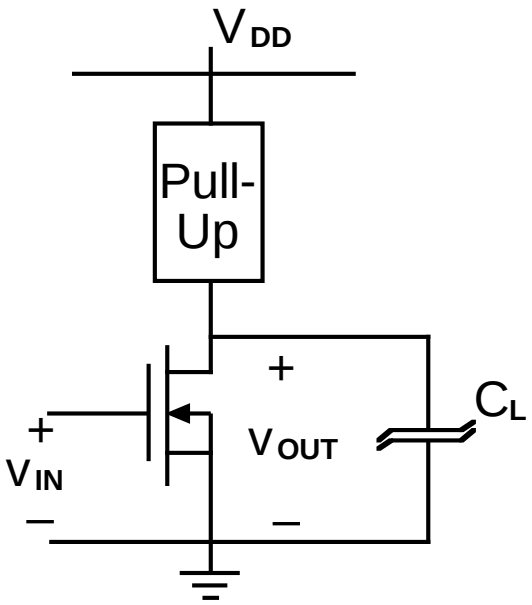


Discharging cycle:
$i_{Discharge} = i_{PD} - i_{PU}$

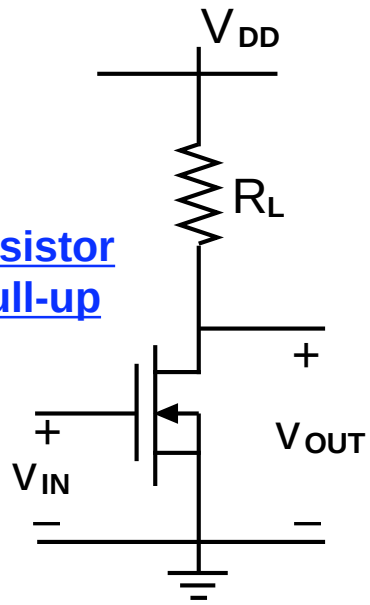
MOS inverters:

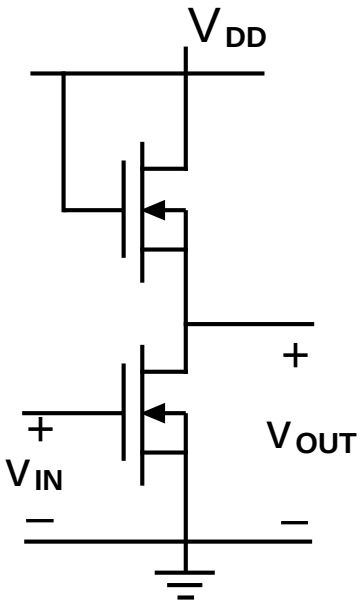
5 pull-up choices

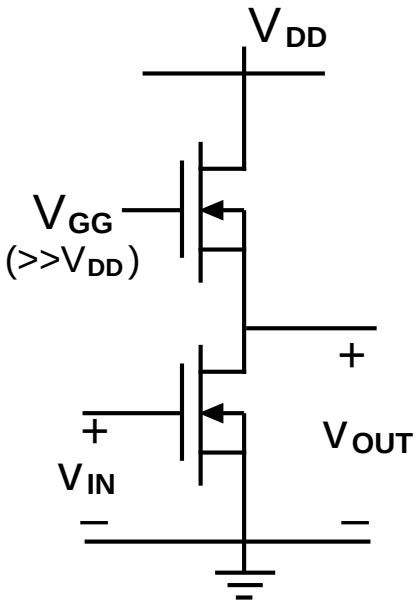
Generic inverter

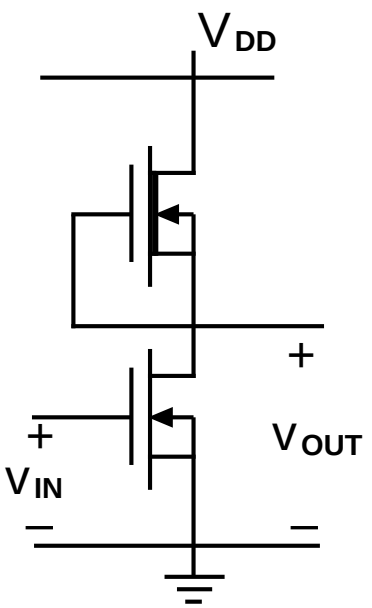


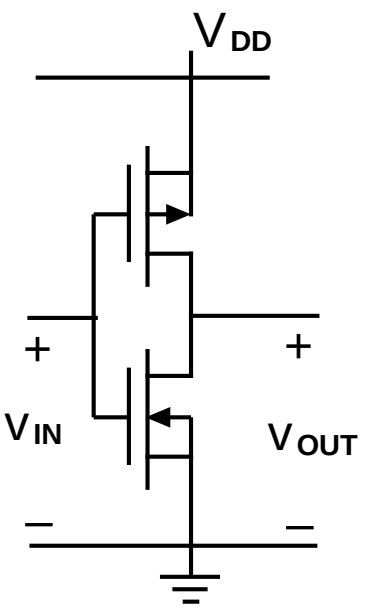
Resistor pull-up











n-channel, e-mode pull-up

V_{DD} on gate **V_{GG} on gate**

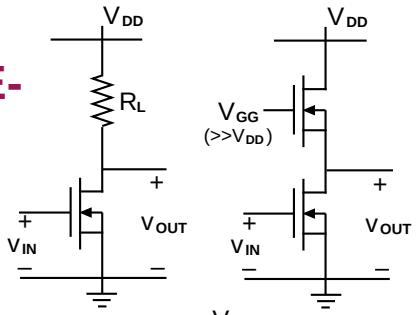
n-channel, d-mode pull-up (NMOS)

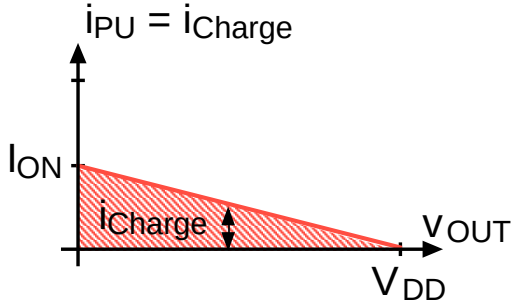
Active p-channel pull-up (CMOS)**

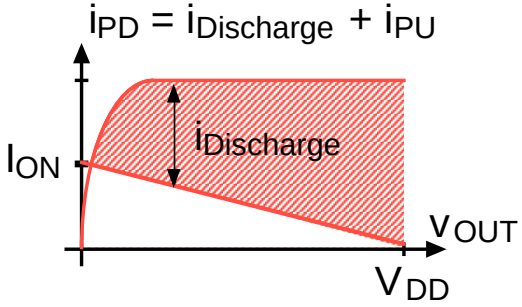
* Known as PMOS when made with p-channel. ** Notice that CMOS has a larger (~3x) input capacitance.

Switching transients: summary of charge/discharge currents

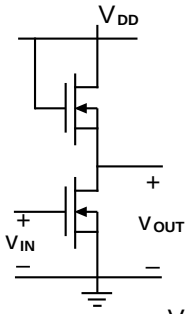
Resistor and E-mode pull-up (V_{GG} on gate)

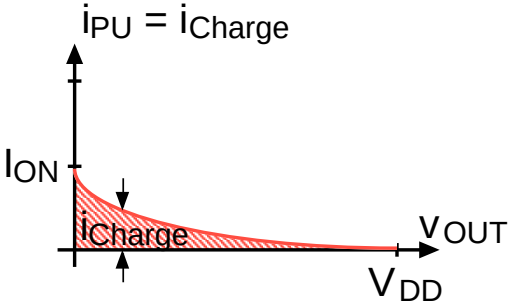


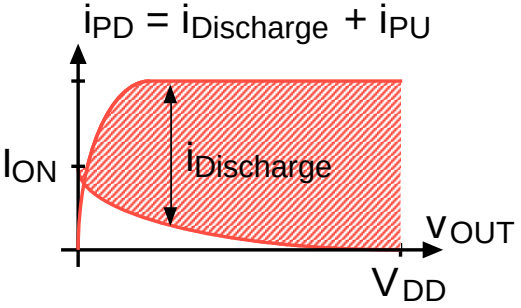




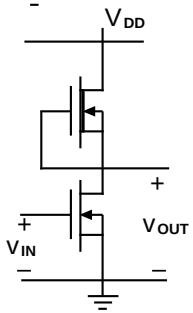
E-mode pull-up (V_{DD} on gate)

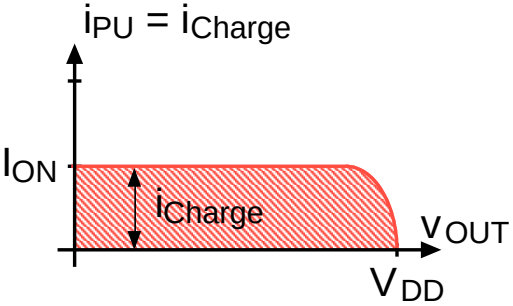


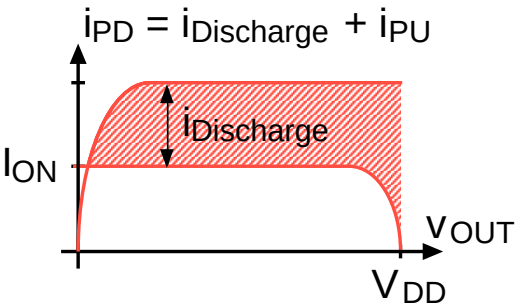




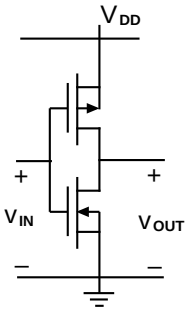
D-mode pull-up (called "NMOS")

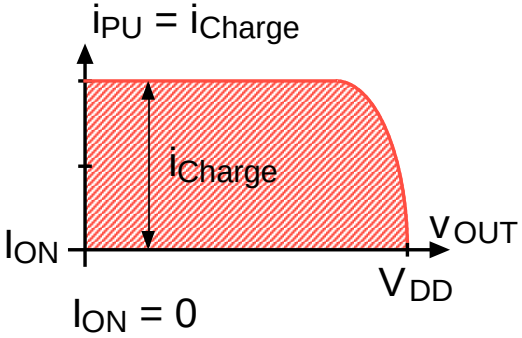


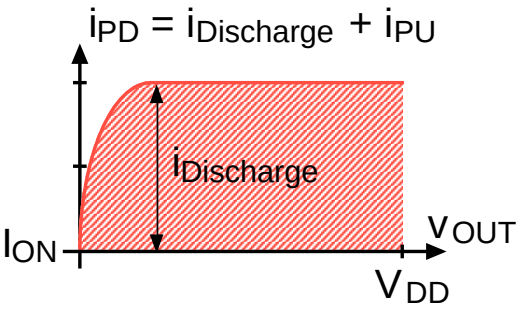




CMOS

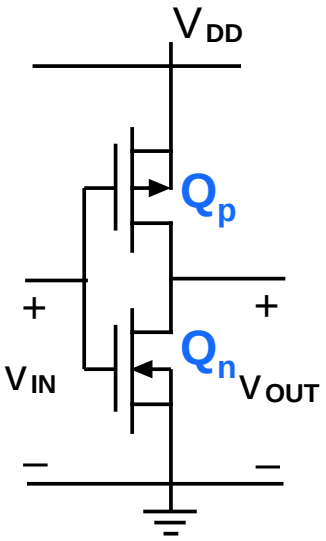


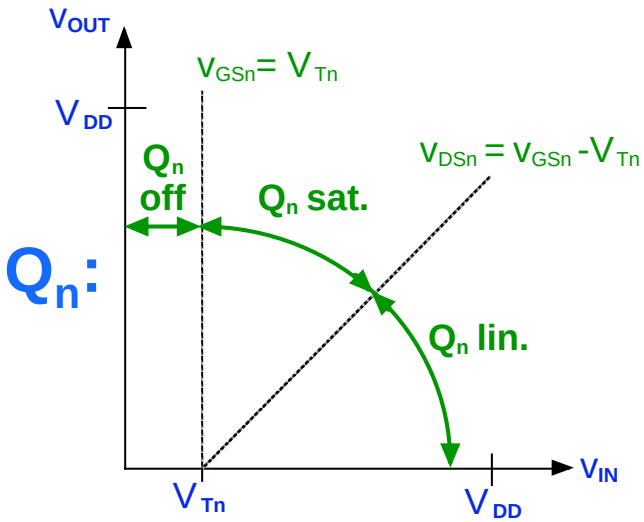


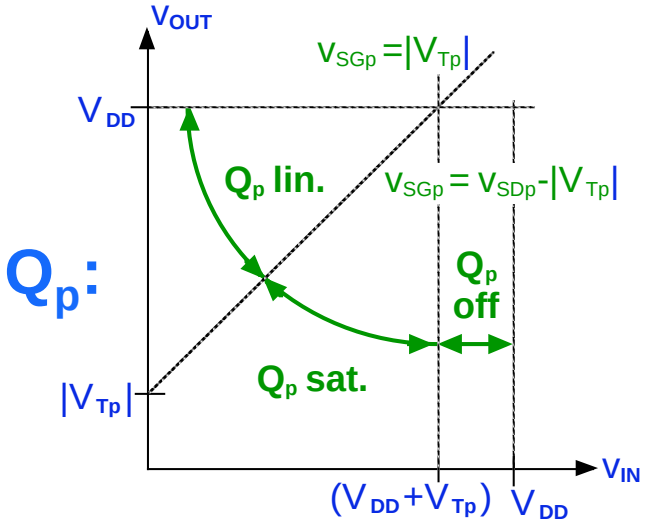


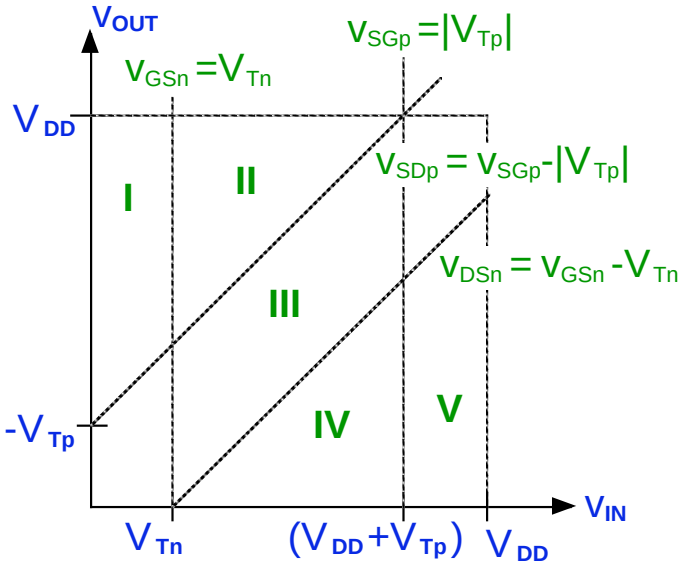
- **Comparisons made with same pull-down MOSFET, V_{HI}, and I_{ON}.**

CMOS: transfer characteristic calculation









Transistor operating condition in each region:

Region	Q_n	Q_p
I	cut-off	linear
II	saturation	linear
III	saturation	saturation
IV	linear	saturation
V	linear	cut-off

CMOS: transfer characteristic calculation, cont.

Region I:

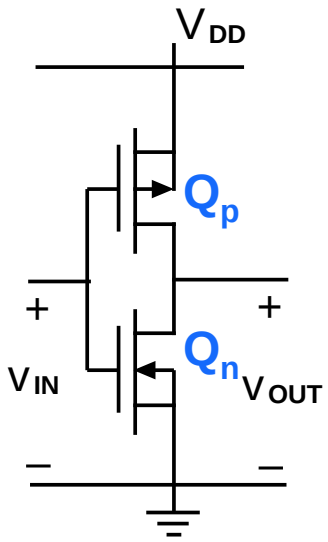
$$i_{Dn} = 0 \quad \text{and} \quad i_{Dp} = K_p\left[V_{DD} - v_{IN} - |V_{Tp}| - \frac{(V_{DD} - v_{OUT})}{2}\right](V_{DD} - v_{OUT})$$

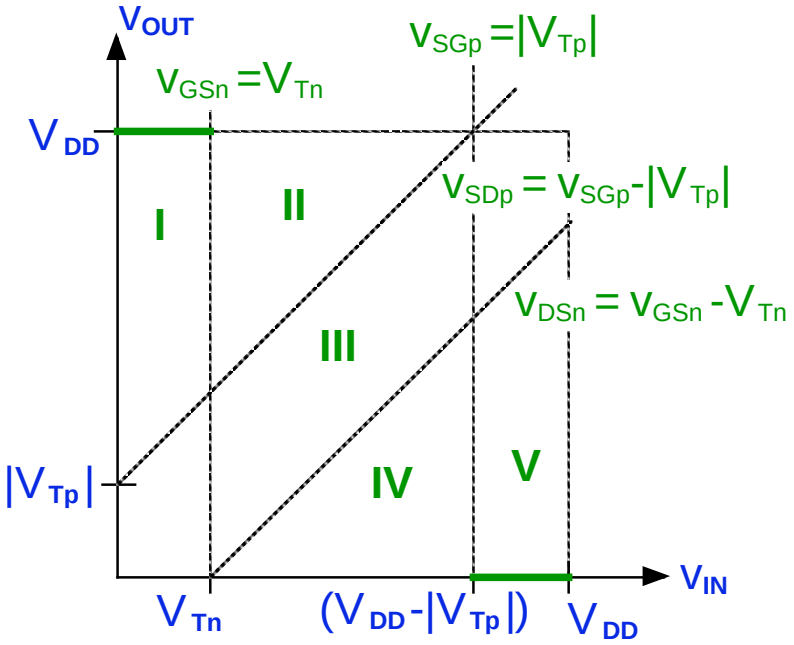
$$\text{so} \quad i_{Dn} = i_{Dp} \quad \Rightarrow \quad v_{OUT} = V_{DD}$$

Region II:

$$i_{Dn} = K_n\left[v_{IN} - V_{Tn} - \frac{v_{OUT}}{2}\right]v_{OUT} \quad \text{and} \quad i_{Dp} = 0$$

$$\text{so} \quad i_{Dn} = i_{Dp} \quad \Rightarrow \quad v_{OUT} = 0$$





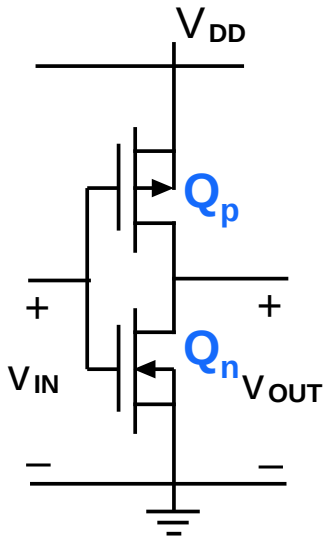
CMOS: transfer characteristic calculation, cont.

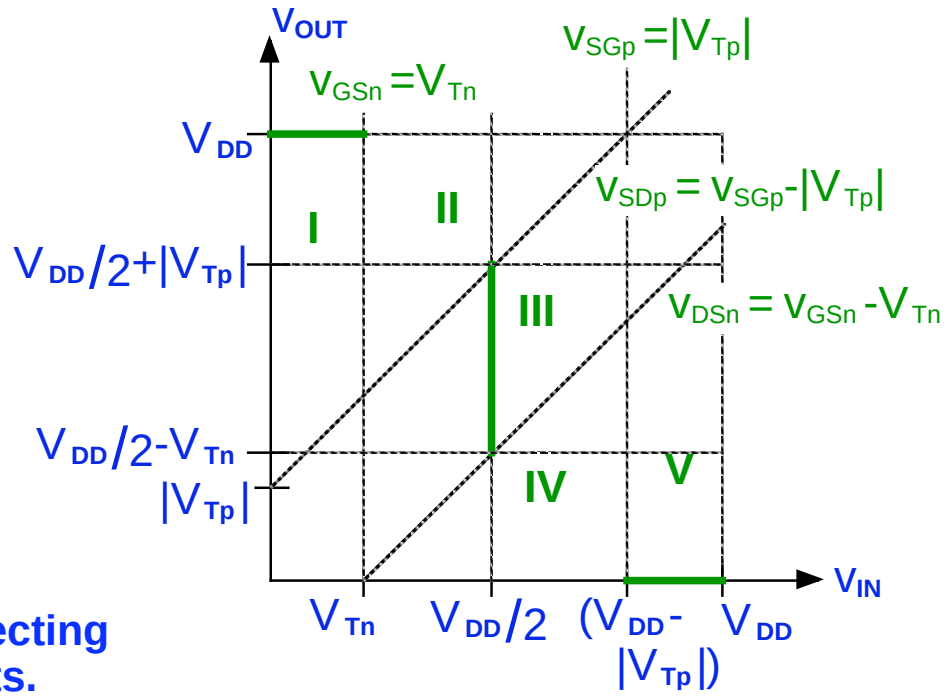
Region III:

$$i_{Dn} = \frac{K_n}{2}\left[v_{IN} - V_{Tn}\right]^2 \quad \text{and} \quad i_{Dp} == \frac{K_p}{2}\left[V_{DD} - v_{IN} - \left|V_{Tp}\right|\right]^2$$

so $i_{Dn} = i_{Dp} \Rightarrow v_{IN} = \dfrac{V_{DD} - \left|V_{Tp}\right| + V_{Tn}\sqrt{K_n/K_p}}{1+\sqrt{K_n/K_p}}$. To achieve symmetry we make

$K_n = K_p$, and $\left|V_{Tp}\right| = V_{Tn}$. With this: $v_{IN} = \dfrac{V_{DD}}{2}$ and $\dfrac{V_{DD}}{2} - V_{Tn} \le v_{OUT} \le \dfrac{V_{DD}}{2} + \left|V_{Tp}\right|$



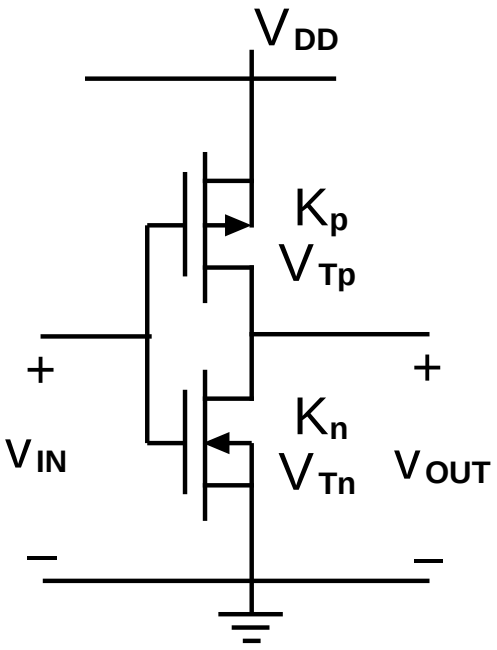


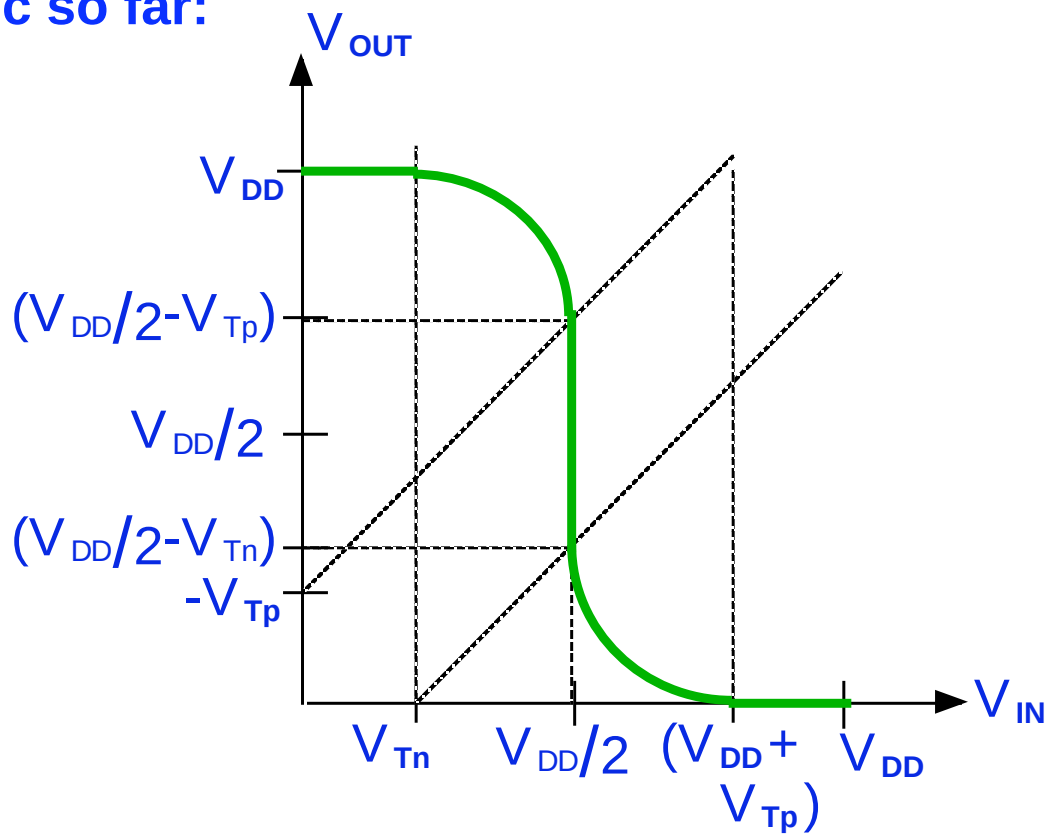
Regions II and IV:

Parabolic segments connecting the three straight segments.

CMOS: transfer characteristic calculation, cont.

Complete characteristic so far:





NOTE: We design CMOS inverters to have $K_n = K_p$ and $V_{Tn} = -V_{Tp}$ to obtain the optimum symmetrical characteristic.

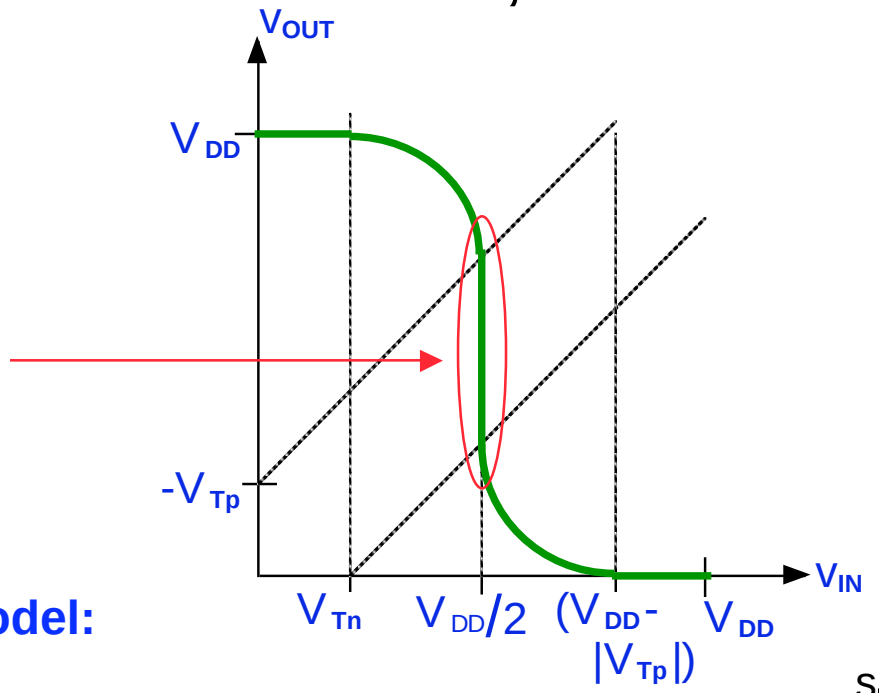
CMOS: transfer characteristic calculation, cont.

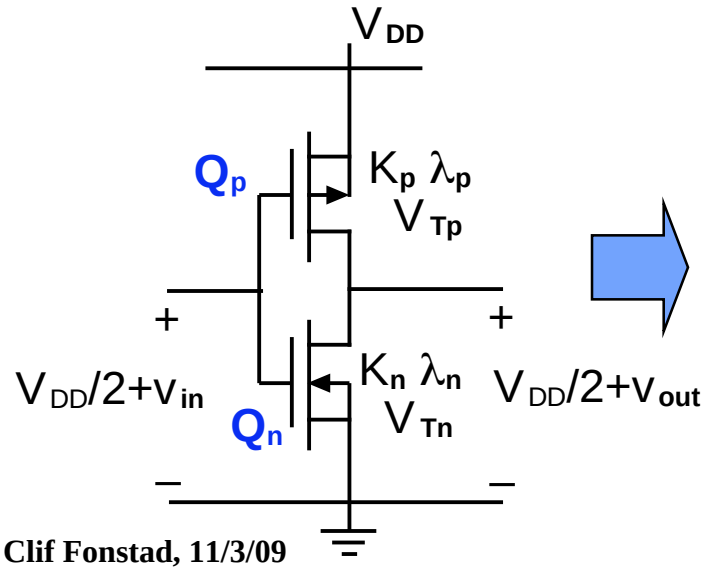
Our calculation says that the transfer characteristic is vertical in Region III.

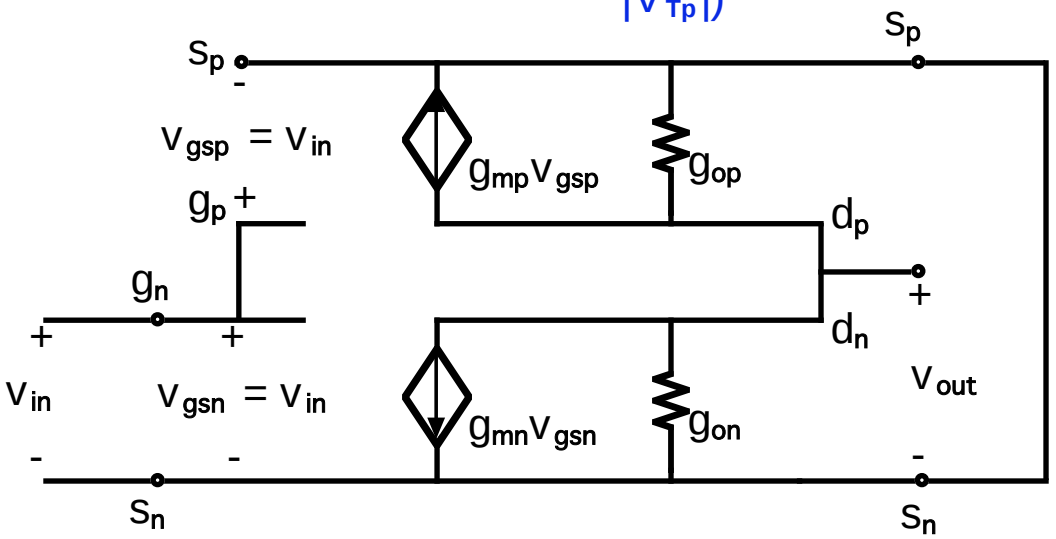
We know it must have some slope, but what is it?

To see, calculate the small signal gain about the bias point: $V_{IN} = V_{OUT} = V_{DD}/2$

Begin with the small signal model:

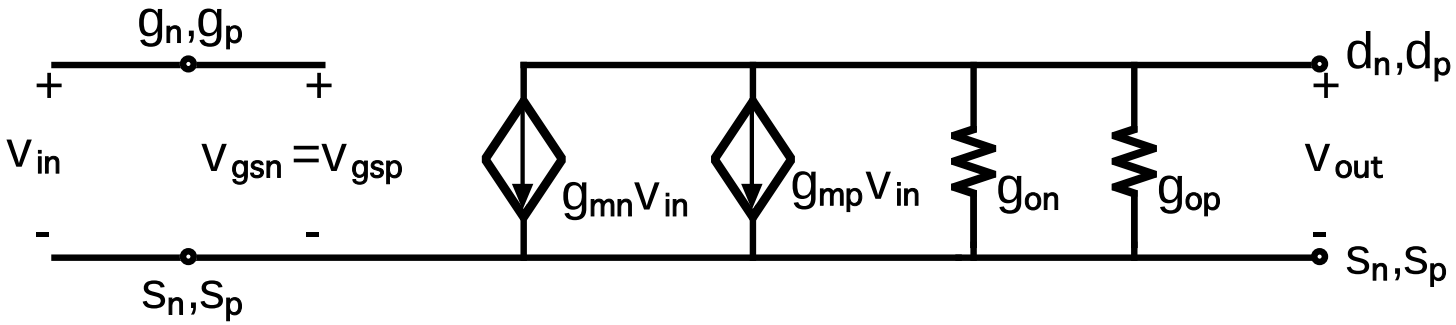






CMOS: transfer characteristic calculation, cont.

Redrawing the circuit we get



from which we see immediately

$$A_v \equiv \left.\frac{\partial v_{OUT}}{\partial v_{IN}}\right|_Q = \frac{v_{out}}{v_{in}} = -\frac{\left[g_{mn}+g_{mp}\right]}{\left[g_{on}+g_{op}\right]}$$

In Lecture 13 we learned how to write the conductances in terms of the bias point as

$$g_{mn} = \sqrt{2K_n I_{Dn}}, \quad g_{mp} = \sqrt{2K_p \left|I_{Dp}\right|} = g_{mn}, \quad g_{on} = \lambda_n I_{Dn}, \quad g_{op} = \lambda_p \left|I_{Dp}\right| = \lambda_p I_{Dn}$$

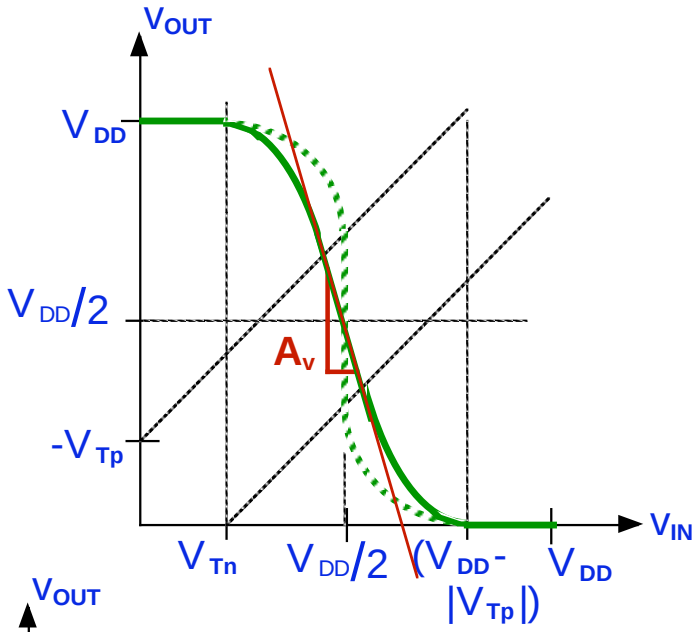
which will enable us to express the gain in terms of the bias point, I_{Dn} ($= |I_{Dp}|$), and MOSFET parameters

$$A_v \equiv \left.\frac{\partial v_{OUT}}{\partial v_{IN}}\right|_Q = -\frac{2\sqrt{2K_n I_{Dn}}}{\left[\lambda_n+\lambda_p\right] I_{Dn}} = -\frac{2\sqrt{2K_n}}{\left[\lambda_n+\lambda_p\right]\sqrt{I_{Dn}}}$$

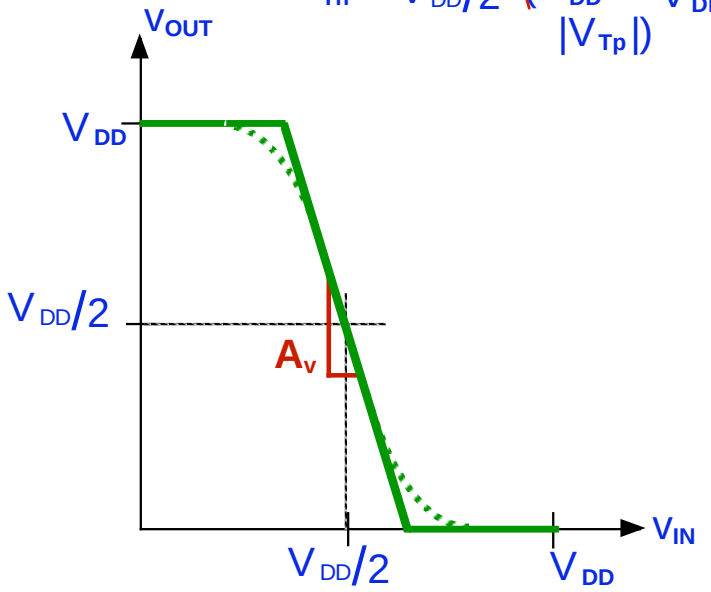
CMOS: transfer characteristic calculation, cont.

Returning to the transfer characteristic, we see that the slope in Region III is not infinite, but is instead:

$$A_v \equiv \left.\frac{\partial v_{OUT}}{\partial v_{IN}}\right|_Q = -\frac{\left[g_{mn} + g_{mp}\right]}{\left[g_{on} + g_{op}\right]}$$



Final comment: A quick and easy way to approximate the transfer characteristic of a CMOS gate is to simply draw the three straight line portions in Regions I, III, and V:



CMOS: switching speed; minimum cycle time

The load capacitance: C_L

- Assume to be linear
- Is proportional to MOSFET gate area
- In channel: $\mu_e = 2\mu_h$ so to have $K_n = K_p$ we must have $W_p/L_p = 2W_n/L_n$
  Typically $L_n = L_p = L_{min}$ and $W_n = W_{min}$, so we also have $W_p = 2W_{min}$

$$C_L \approx n\left[W_n L_n + W_p L_p\right]C_{ox}^* = n\left[W_{\min} L_{\min} + 2W_{\min} L_{\min}\right]C_{ox}^* = 3nW_{\min} L_{\min} C_{ox}^*$$

Charging cycle: v_{IN}: HI to LO; Q_n off, Q_p on; v_{OUT}: LO to HI

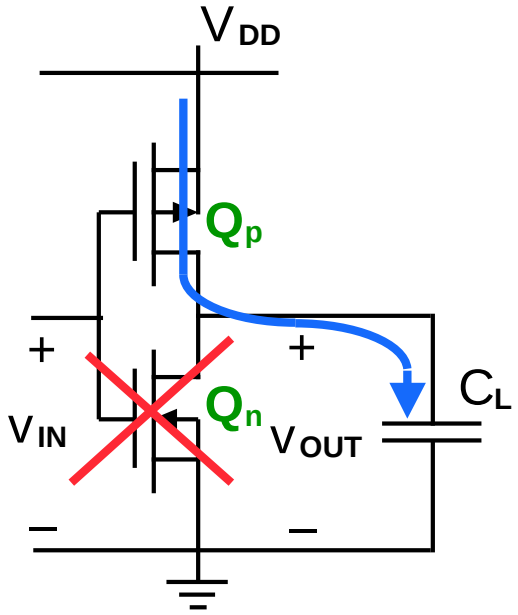
- Assume charged by constant $i_{D,sat}$

$$i_{Charge} = -i_{Dp} \approx \frac{K_p}{2}\left[V_{DD} - |V_{Tp}|\right]^2 = \frac{K_n}{2}\left[V_{DD} - V_{Tn}\right]^2$$

$$q_{Charge} = C_L V_{DD}$$

$$\tau_{Charge} = \frac{q_{Charge}}{i_{Charge}} = \frac{2C_L V_{DD}}{K_n\left[V_{DD} - V_{Tn}\right]^2}$$

$$= \frac{6nW_{\min} L_{\min} C_{ox}^* V_{DD}}{\frac{W_{\min}}{L_{\min}}\mu_e C_{ox}^*\left[V_{DD} - V_{Tn}\right]^2} = \frac{6nL_{\min}^2 V_{DD}}{\mu_e\left[V_{DD} - V_{Tn}\right]^2}$$



CMOS: switching speed; minimum cycle time, cont.

Discharging cycle: v_{IN}: LO to HI; Q_n on, Q_p off; v_{OUT}: HI to LO

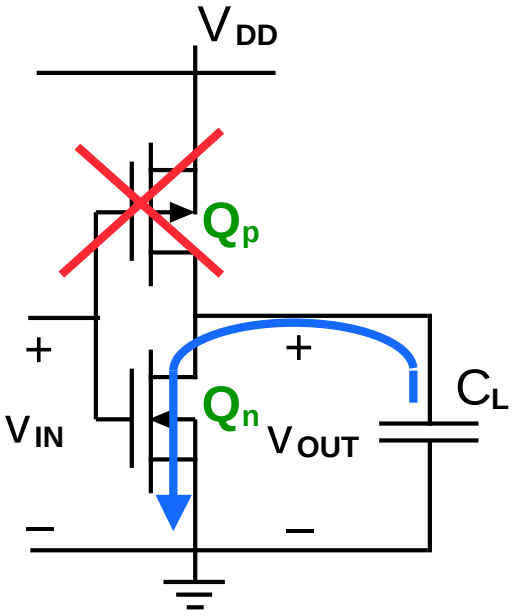
- Assume discharged by constant $i_{D,sat}$

$$i_{Disch\,arg\,e} = i_{Dn} \approx \frac{K_n}{2}\left[V_{DD} - V_{Tn}\right]^2$$

$$q_{Disch\,arg\,e} = C_L V_{DD}$$

$$\tau_{Disch\,arg\,e} = \frac{q_{Disch\,arg\,e}}{i_{Disch\,arg\,e}} = \frac{2C_L V_{DD}}{K_n\left[V_{DD} - V_{Tn}\right]^2}$$

$$= \frac{6nW_{\min}L_{\min}C_{ox}^{*}V_{DD}}{\frac{W_{\min}}{L_{\min}}\mu_e C_{ox}^{*}\left[V_{DD} - V_{Tn}\right]^2} = \frac{6nL_{\min}^2 V_{DD}}{\mu_e\left[V_{DD} - V_{Tn}\right]^2}$$



Minimum cycle time: v_{IN}: LO to HI to LO; v_{OUT}: HI to LO to HI

$$\tau_{Min.Cycle} = \tau_{Ch\,arg\,e} + \tau_{Disch\,arg\,e} = \frac{12nL_{\min}^2 V_{DD}}{\mu_e\left[V_{DD} - V_{Tn}\right]^2}$$

CMOS: switching speed; minimum cycle time, cont.

Discharging and Charging times:

What do the expressions tell us? We have

$$\tau_{Min\,Cycle} = \frac{12nL_{\min}^2 V_{DD}}{\mu_e \left[V_{DD} - V_{Tn}\right]^2}$$

This can be written as:

$$\tau_{Min\,Cycle} = \frac{12nV_{DD}}{\left(V_{DD} - V_{Tn}\right)} \cdot \frac{L_{\min}}{\mu_e \left(V_{DD} - V_{Tn}\right)/L_{\min}}$$

The last term is the channel transit time:

$$\frac{L_{\min}}{\mu_e \left(V_{DD} - V_{Tn}\right)/L_{\min}} = \frac{L_{\min}}{\mu_e \mathrm{E}_{Ch}} = \frac{L_{\min}}{\bar{s}_{e,Ch}} = \tau_{Ch\,Transit}$$

Thus the gate delay is a multiple of the channel transit time:

$$\tau_{Min\,Cycle} = \frac{12nV_{DD}}{\left(V_{DD} - V_{Tn}\right)} \tau_{Channel\,Transit} = n' \tau_{Channel\,Transit}$$

CMOS: power dissipation - total and per unit area

Average power dissipation

All dynamic

$$P_{dyn,ave} = E_{Dissipated\ per\ cycle}\, f = C_L V_{DD}^2 = 3nW_{\min}L_{\min}C_{ox}^* V_{DD}^2 f$$

Power at maximum data rate

Maximum f will be $1/\tau_{\text{Gate Delay Min.}}$

$$P_{dyn@\,f_{\max}} = \frac{3nW_{\min}L_{\min}C_{ox}^* V_{DD}^2}{\tau_{Min.Cycle}} = 3nW_{\min}L_{\min}C_{ox}^* V_{DD}^2 \cdot \frac{\mu_e\left[V_{DD}-V_{Tn}\right]^2}{12nL_{\min}^2 V_{DD}}$$

$$= \frac{1}{4}\frac{W_{\min}}{L_{\min}}\mu_e C_{ox}^* V_{DD}\left[V_{DD}-V_{Tn}\right]^2$$

Power density at maximum data rate

Assume that the area per inverter is proportional to $W_{min}L_{min}$

$$PD_{dyn@\,f_{\max}} = \frac{P_{dyn@\,f_{\max}}}{\text{InverterArea}} \propto \frac{P_{dyn@\,f_{\max}}}{W_{\min}L_{\min}} = \frac{\mu_e C_{ox}^* V_{DD}\left[V_{DD}-V_{Tn}\right]^2}{L_{\min}^2}$$

CMOS: design for high speed

Maximum data rate

Proportional to $1/\tau_{\text{Min Cycle}}$

$$\tau_{Min.Cycle} = \tau_{Ch\,arg\,e} + \tau_{Disch\,arg\,e} = \frac{12nL_{\min}^2 V_{DD}}{\mu_e\left[V_{DD} - V_{Tn}\right]^2}$$

Implies we should reduce L_{min} and increase V_{DD}.

Note: As we reduce L_{min} we must also reduce t_{ox}, but t_{ox} doesn't enter directly in f_{max} so it doesn't impact us here

Power density at maximum data rate

Assume that the area per inverter is proportional to $W_{min}L_{min}$

$$PD_{dyn@\,f_{\max}} \propto \frac{P_{dyn@\,f_{\max}}}{W_{\min}L_{\min}} = \frac{\mu_e \varepsilon_{ox} V_{DD}\left[V_{DD} - V_{Tn}\right]^2}{t_{ox}L_{\min}^2}$$

Shows us that PD increases very quickly as we reduce L_{min} unless we also reduce V_{DD} (which will also reduce f_{max}).

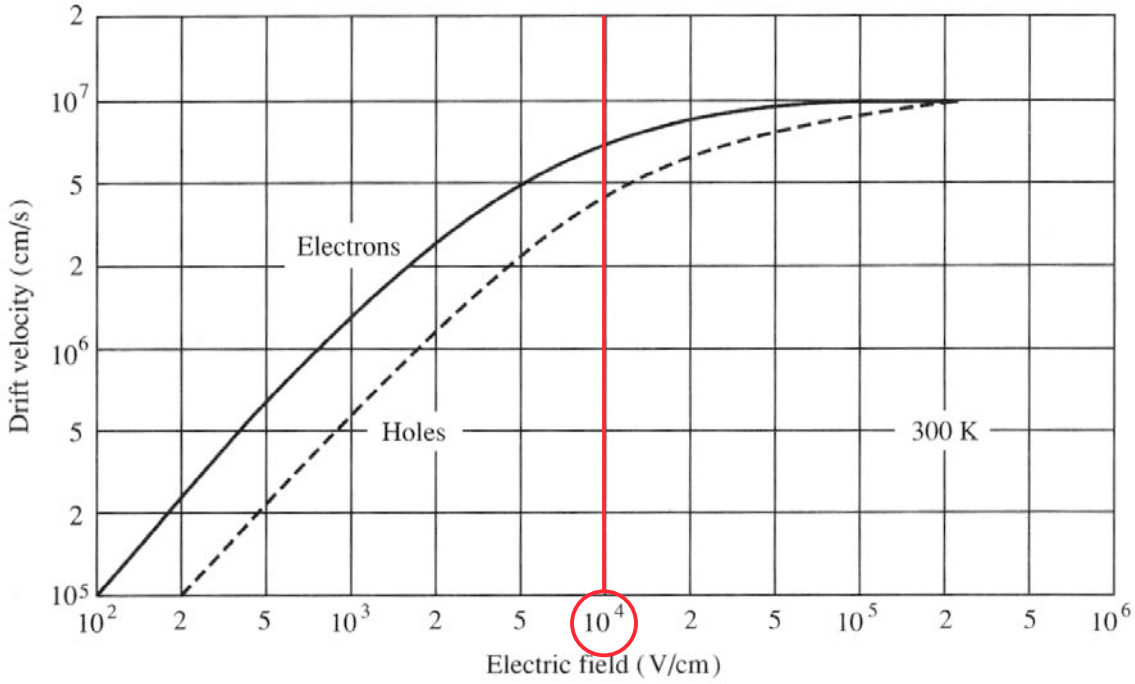
Note: Now t_{ox} does appear in the expression, so the rate of increase with decreasing L_{min} is even greater because t_{ox} must be reduced along with L.

How do we make f_{max} larger without melting the silicon?
By following CMOS scaling rules - the topic of Lecture 16.

CMOS: velocity saturation

Sanity check

CMOS gate lengths are now under 0.1 μm (100 nm). The electric field in the channel can be very high: $E_y \geq 10^4$ V/cm when $v_{DS} \geq 0.1$ V.



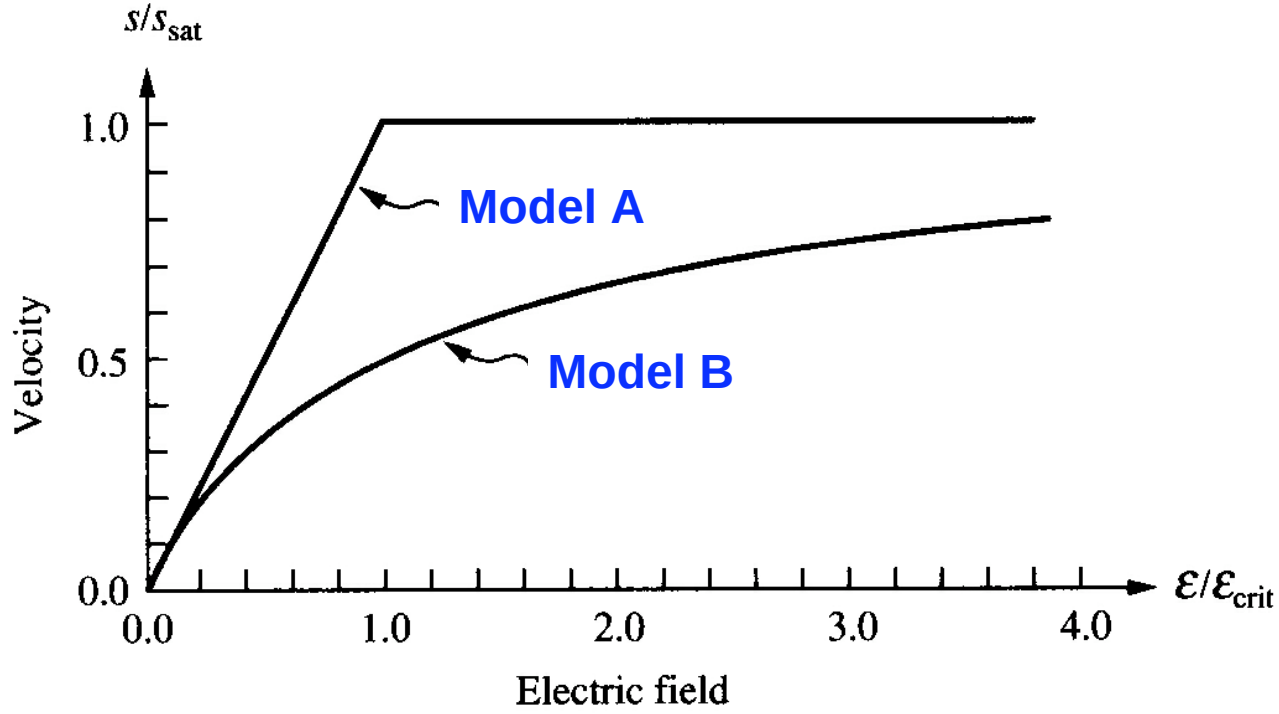
Clearly the velocity of the electrons and holes in the channel will be saturated at even low values of v_{DS}!

What does this mean for the device and inverter characteristics?

CMOS: velocity saturation, cont.

Models for velocity saturation*

Two useful models are illustrated below. We'll use Model A today.



Model A

$$s_y(E_y) = \mu_e E_y \quad if \quad E_y \le E_{crit}$$
$$= \mu_e E_{crit} \equiv s_{sat} \quad if \quad E_y \ge E_{crit}$$

Model B

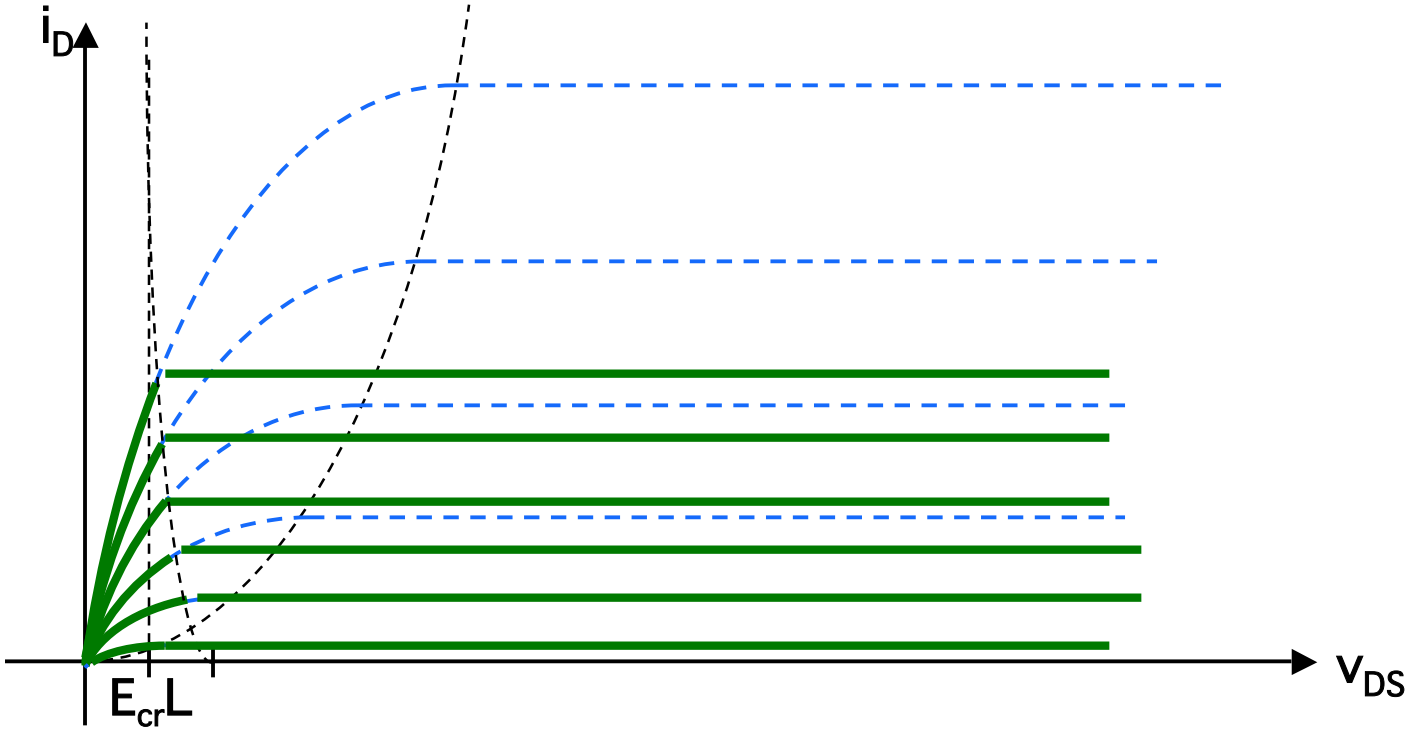
$$s_y(E_y) = \frac{\mu_e E_y}{1 + \frac{E_y}{E_{crit}}}$$

* See pp 281ff and 307ff in course text.

CMOS: velocity saturation, cont.

Drain current: $i_D(v_{GS},v_{DS},v_{BS})$

With Model A*, the low field i_D model, $s = \mu E$, holds for increasing v_{DS} until the velocity of the electrons at some point in the channel reaches s_{sat} (this will happen at the drain end). When this happens the current saturates, and does not increase further for larger v_{DS}.



* Model A: $s_y(E_y) = \mu_e E_y$ if $E_y \le E_{crit}$

$= \mu_e E_{crit} \equiv s_{sat}$ if $E_y \ge E_{crit}$

CMOS: velocity saturation, cont.

If the channel length, L, is sufficiently small we can simplify the model even further because the carrier velocity will saturate at such a small v_{DS} that for $v_{DS} \geq E_{crit}L$ the inversion layer will be uniform and all the carriers will be drifting at their saturation velocity. In this situation (the saturation region) we will have:

$$i_D(v_{GS},v_{DS},v_{BS}) \approx -W\, q_N(v_{GS},v_{BS})\, s_{sat} = W\ s_{sat}\, C_{ox}^{*}\left[v_{GS} - V_T(v_{BS})\right]$$

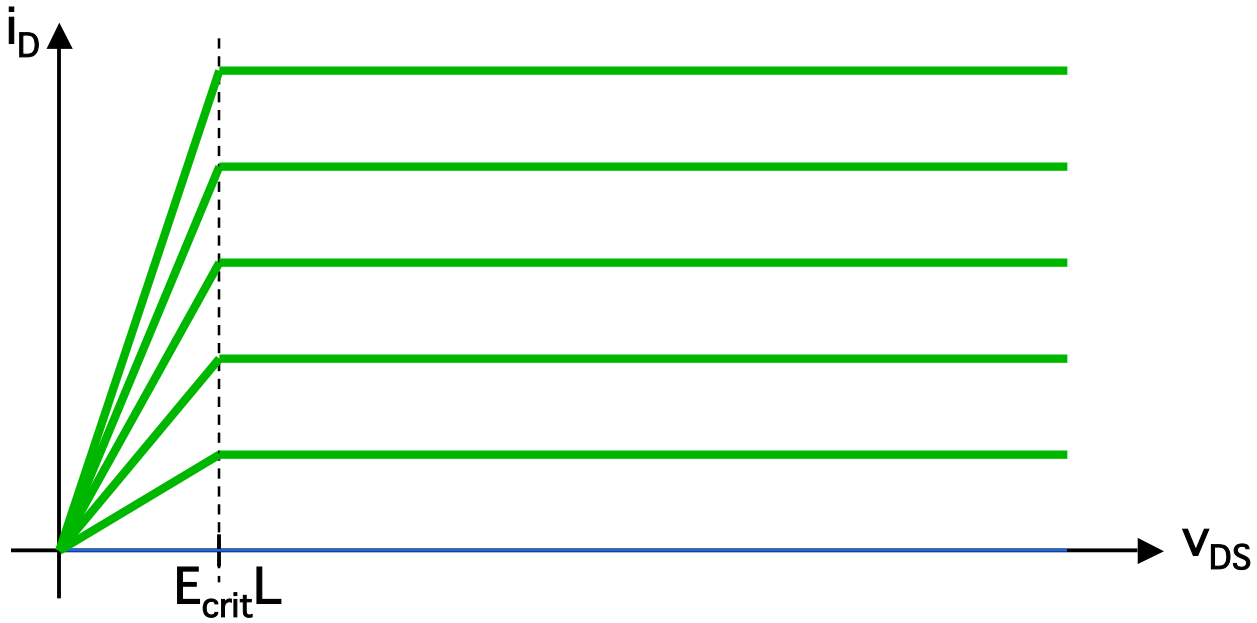
For smaller v_{DS}, prior to the onset of velocity saturation, the linear region model we had earlier will hold. The entire characteristic, neglecting the $v_{DS}/2$ factor in the linear region expression, is

$$i_D(v_{GS},v_{DS},v_{BS}) \approx \begin{cases} 0 & \text{for} \quad (v_{GS} - V_T) < 0 < v_{DS} \\ W\ s_{sat}\, C_{ox}^{*}\left[v_{GS} - V_T(v_{BS})\right] & \text{for} \quad 0 < (v_{GS} - V_T),\ E_{crit}L < v_{DS} \\ \dfrac{W}{L}\,\mu_e\, C_{ox}^{*}\left[v_{GS} - V_T(v_{BS})\right]v_{DS} & \text{for} \quad 0 < (v_{GS} - V_T),\ v_{DS} < E_{crit}L \end{cases}$$

Note that the current in saturation increases linearly with $(v_{GS} - V_T)$, rather than as its square like it did then the gate was longer.

CMOS: velocity saturation, cont.

This simple model for the output characteristics of a very short channel MOSFET (plotted below) provides us an easy way to understand the impact of velocity saturation on MOSFET and CMOS inverter performance.



Note first that in the forward active region where $v_{DS} \geq E_{crit}L$, the curves in the output family are evenly spaced, indicating a constant g_m:

$$g_m \equiv \partial i_D / \partial v_{GS}\big|_Q = W\, s_{sat}\, C_{ox}^*$$

CMOS: velocity saturation, cont.

Charge/discharge cycle and gate delay:

The charge and discharge currents, charges, and times are now:

$$i_{Disch\,arg\,e} = i_{Ch\,arg\,e} = W_{\min}\, s_{sat}\, C^*_{ox}\left(V_{DD} - V_{Tn}\right)$$

$$q_{Disch\,arg\,e} = q_{Ch\,arg\,e} = C_L V_{DD} = 3W_{\min} L_{\min} C^*_{ox} V_{DD}$$

$$\tau_{Disch\,arg\,e} = \tau_{Ch\,arg\,e} = \frac{q_{Disch\,arg\,e}}{i_{Disch\,arg\,e}} = \frac{3W_{\min} L_{\min} C^*_{ox} V_{DD}}{W_{\min}\, s_{sat}\, C^*_{ox}\left(V_{DD} - V_{Tn}\right)} = \frac{3nL_{\min} V_{DD}}{s_{sat}\left(V_{DD} - V_{Tn}\right)}$$

CMOS minimum cycle time and power density at f_{max}:

$$\tau_{Min.Cycle} = \tau_{Ch\,arg\,e} + \tau_{Disch\,arg\,e} = \frac{6\,n\,L_{\min} V_{DD}}{s_{sat}\left[V_{DD} - V_{Tn}\right]}$$

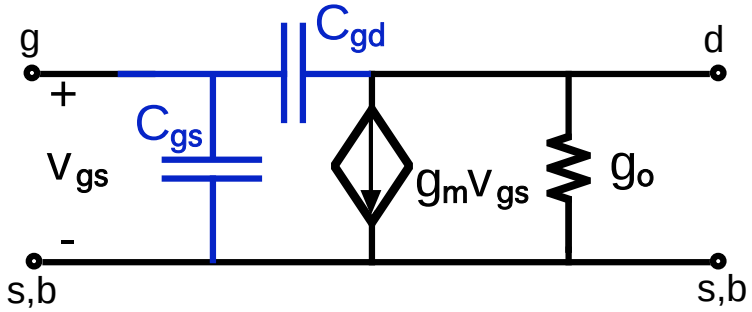
Note: $\tau_{ChanTransit} = \dfrac{L}{s_{sat}}$

$$\tau_{Min.Cycle} \propto \frac{L_{\min} V_{DD}}{s_{sat}\left[V_{DD} - V_{Tn}\right]} = n'\tau_{ChanTransit} \qquad PD_{dyn\,@\,f_{\max}} \propto \frac{s_{sat}\,\varepsilon_{ox} V_{DD}\left[V_{DD} - V_{Tn}\right]}{t_{ox} L_{\min}}$$

Lessons: Still gain by reducing L, but not as quickly.
Scaling of both dimensions and voltage is still required.
Channel transit time, L_{min}/s_{sat}, still rules!

MOSFETs: LEC w. velocity saturation

Small signal linear equivalent circuit: g_m and C_{gs} change



$$g_m \equiv \left.\frac{\partial i_D}{\partial v_{GS}}\right|_Q = W\, s_{sat}\, C_{ox}^*$$

$$C_{gs} = W\, L\, C_{ox}^*$$

One final model observation: Insight on g_m

We in general want g_m as large as possible. To see another way to think about this is to note that g_m can be related to $\tau_{Ch\text{-}Transit}$:

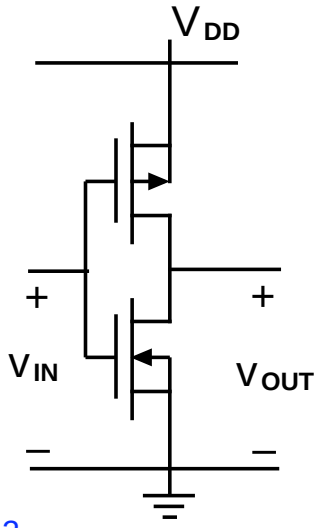
No velocity saturation →

Full velocity saturation →

$$g_m = \begin{cases} \frac{W}{L}\mu_e C_{ox}^*\left(v_{GS}-V_T\right) = \frac{W\, L\, C_{ox}^*}{L^2/\mu_e\left(v_{GS}-V_T\right)} \\ W\, s_{sat}\, C_{ox}^* = \frac{W\, L\, C_{ox}^*}{L/s_{sat}} \end{cases} \propto \frac{C_{gs}}{\tau_{Ch\,Transit}}$$

C_{gs} is a measure of how much channel charge we are controlling, and $1/\tau_{Ch\text{-}tr}$ is a measure of how fast it moves through the device. We'd like both to be large numbers.

Lecture 15 - Digital Circuits: CMOS - Summary



- **CMOS**

  Transfer characteristic: symmetric

  $V_{LO} = 0$, $V_{HI} = V_{DD}$, $I_{ON} = 0$

  $N_{ML} = N_{MH}$ implies $K_n = K_p$, $|V_{Tp}| = V_{Tn} \equiv V_T$

  $L_n = L_p = L_{min}$, $W_p = (\mu_e/\mu_h)W_n$

  Gate delay expressions

  $\tau_{LO\text{-}HI} = \tau_{HI\text{-}LO} = 2V_{DD}C_L/K_n(V_{DD} - V_T)^2$

  Gate delay (GD) $= \tau_{LO\text{-}HI} + \tau_{HI\text{-}LO} = 4V_{DD}C_L/K_n(V_{DD} - V_T)^2$

  If $C_L = n(W_nL_n + W_pL_p)C_{ox}^* = 3n\, W_nL_{min}C_{ox}^*$ **(Assumes $\mu_e = 2\mu_h$)**

  then GD $= 12\, n\, L_{min}^2\, V_{DD}/\, \mu_n(V_{DD} - V_T)^2$ **(Motivation for reducing L_{min})**

  Power and speed-power product

  $P_{ave} = f\, C_L V_{DD}^{\,2}$

  $P_{dyn}@\, f_{max} \propto C_L V_{DD}^2/GD = K_n V_{DD}\,(V_{DD} - V_T)^2/4$ **(Motivation for reducing V_{DD})**

- **Velocity Saturation**

  Gate delay; Power and speed-power product:

  Scales as $1/L_{min}$, rather than $(1/L_{min})^2$

6.012 Microelectronic Devices and Circuits

Fall 2009