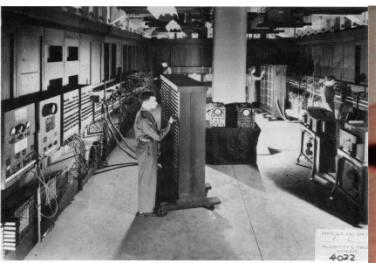
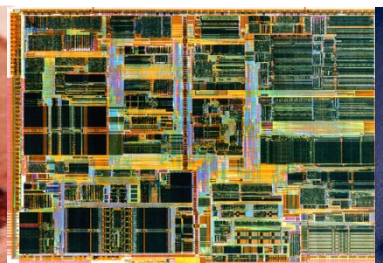
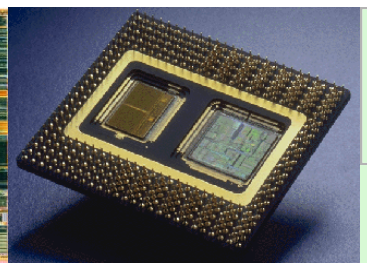
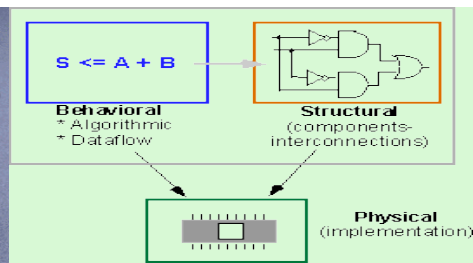
CpE 690: Introduction to VLSI Design

Lecture 7
CMOS Transistor Theory and DC Response

Bryan Ackland
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030

Adapted from Lecture Notes, David Mahoney Harris CMOS VLSI Design



Introduction

- So far, we have treated transistors as ideal switches
- An ON transistor passes a finite amount of current
  - Depends on terminal voltages
  - Derive current-voltage (I-V) relationships
- Transistor gate, source, drain all have capacitance
  - $I = C (\Delta V/\Delta t) \rightarrow \Delta t = (C/I)\ \Delta V$
  - Capacitance and current determine speed
- Also revisit what a “degraded level” means

nMOS

pMOS

MOS Capacitor

- Gate and body form MOS capacitor:

Accumulation:

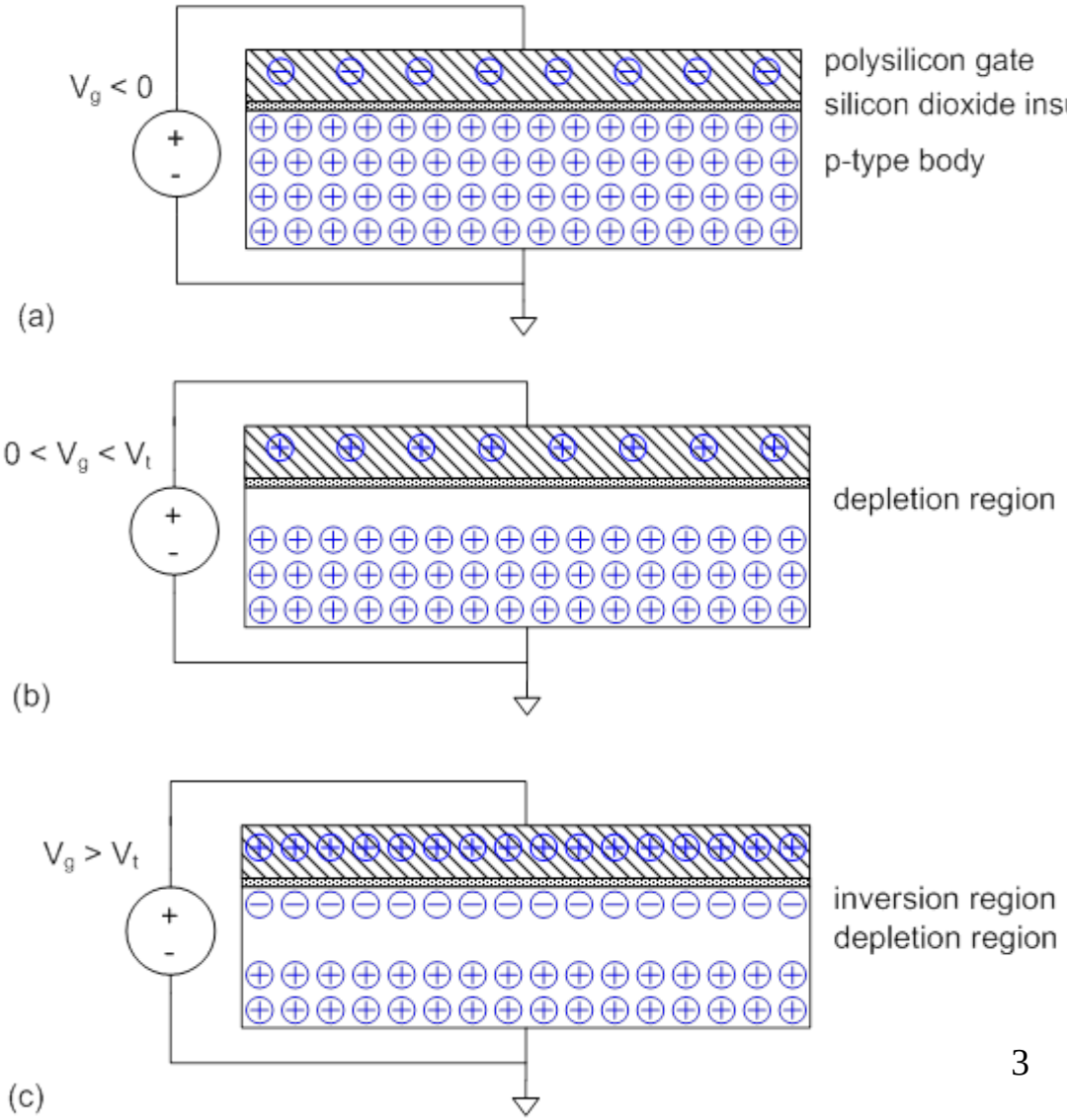
$V_g < 0$

Depletion:

$0 < V_g < V_t$

Inversion:

$V_g > V_t$



nMOS Terminal Voltages

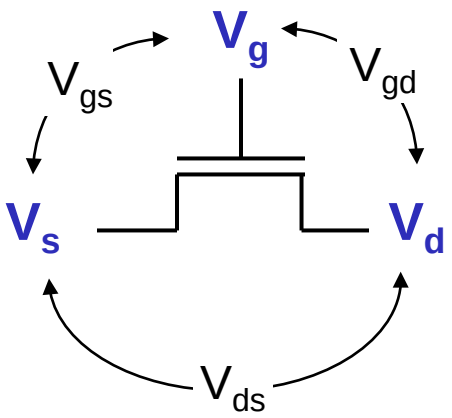
- Mode of operation depends on V_g, V_d, V_s

  $V_{gs} = V_g - V_s$

  $V_{gd} = V_g - V_d$

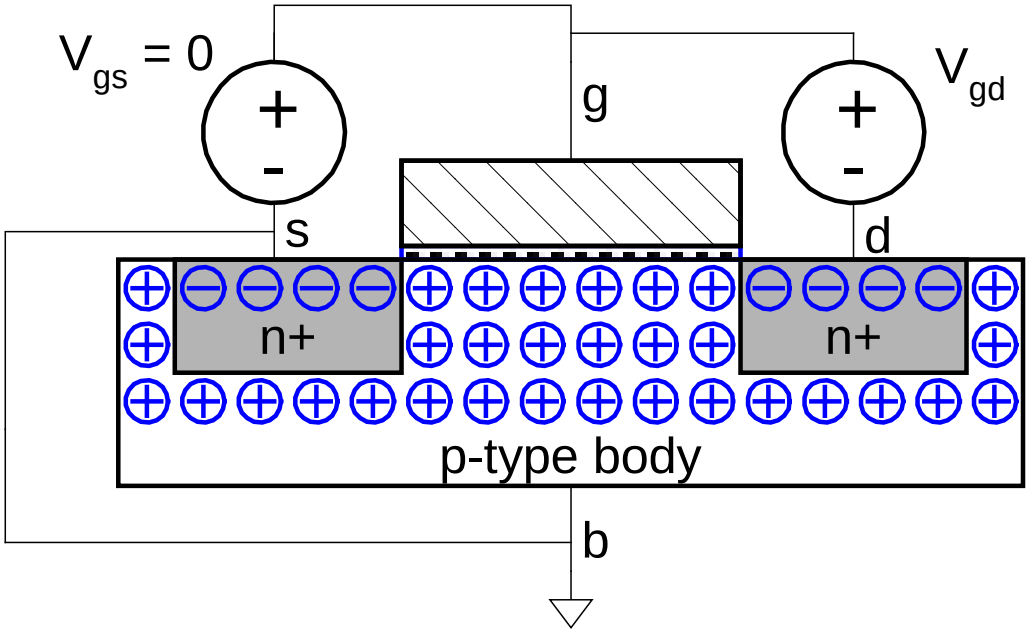
  $V_{ds} = V_d - V_s = V_{gs} - V_{gd}$

- Source and drain are (physically) symmetric terminals
  - By convention, nMOS source is terminal at lower voltage
  - Hence $V_{ds} \geq 0$
- nMOS body is grounded (0 volts).
  - For now, assume source is grounded too.
- Three regions of operation
  - Cutoff
  - Linear
  - Saturation



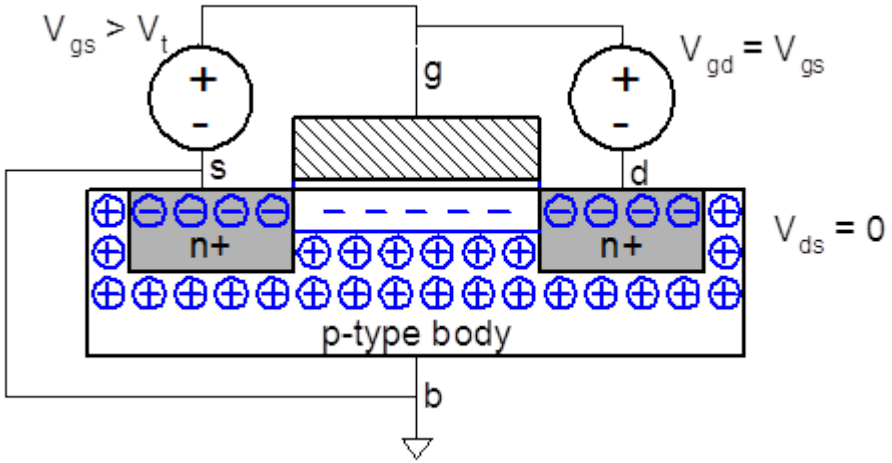
nMOS Cutoff

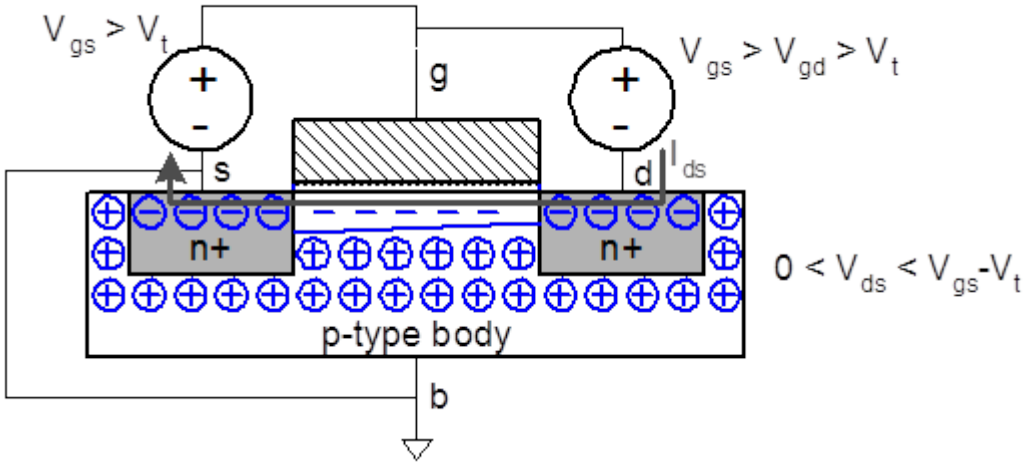
- $V_{gs} < V_t$: No channel
- Source-body and drain-body junctions are reverse biased
- $I_{ds} \approx 0$



nMOS Linear

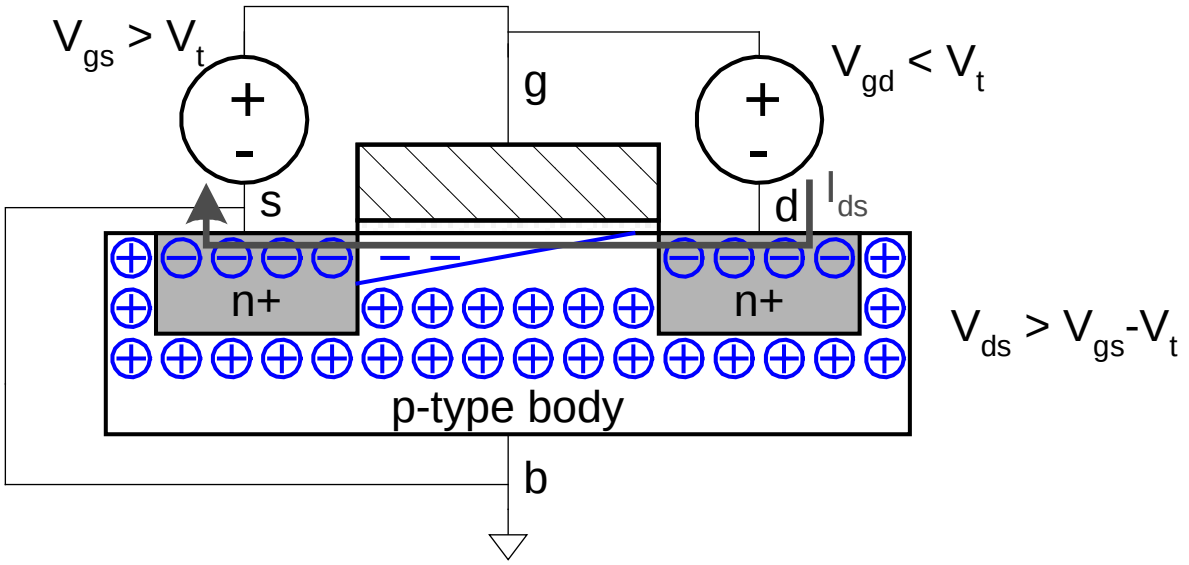
- $V_{gs} > V_t$: Channel forms
- Current flows from drain to source
  - electrons go from source to drain
- I_{ds} increases with V_{ds}
  - Similar to linear resistor
- Also called:
  - resistive
  - triode
  - non-saturated





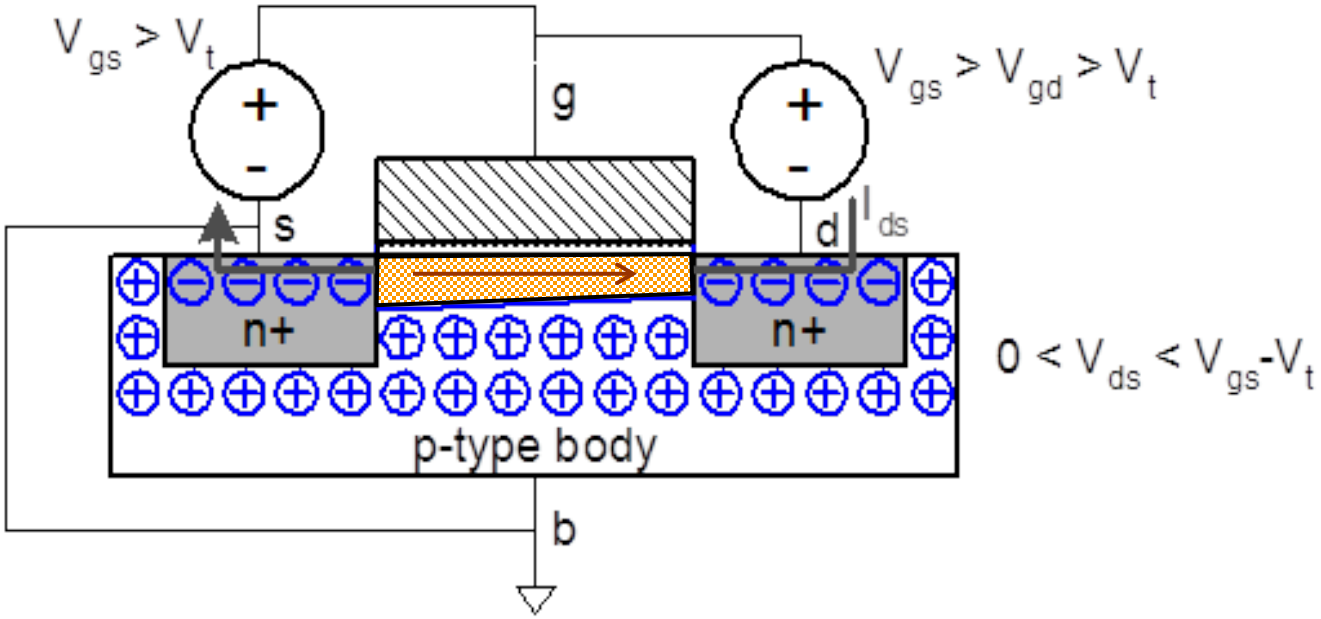
nMOS Saturation

- If $V_{ds} > V_{gs} - V_t$ then $V_{gd} < V_t$: channel "pinches off"
- Conduction due to drift induced by positive drain voltage
- I_{ds} independent of V_{ds}
- We say channel current "saturates"
- Similar to current source



Linear I/V Characteristics

- What is I_{ds} (V_{gs}, V_{ds}) ?
- In linear region, I_{ds} depends on
  - How much charge is in the channel?
  - How fast is the charge moving?

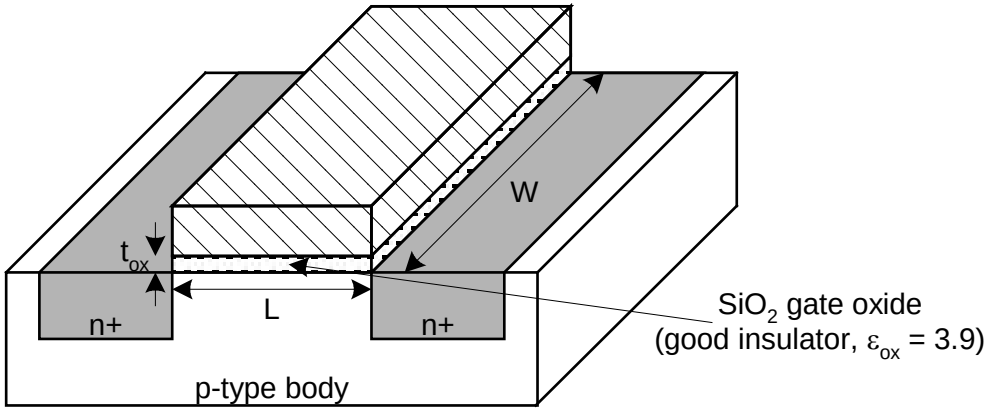


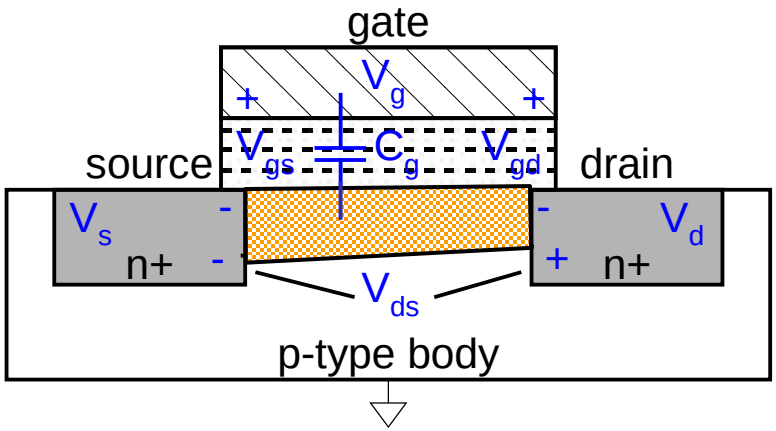
Channel Charge

- MOS structure looks like parallel plate capacitor while operating in inversion
- Gate : oxide : channel
- $Q_{channel}$ = C.V
- $C = C_g = \varepsilon_{ox}.W.L/t_{ox} = C_{ox}.W.L$
- $V = V_{gc} - V_t = (V_{gs} - V_{ds}/2) - V_t$

where $C_{ox} = \varepsilon_{ox} / t_{ox}$

C_{ox} is gate capacitance per unit area





Carrier Velocity

- Charge is carried by e-
- Electrons are propelled by the lateral electric field between source and drain

  $E = V_{ds}/L$

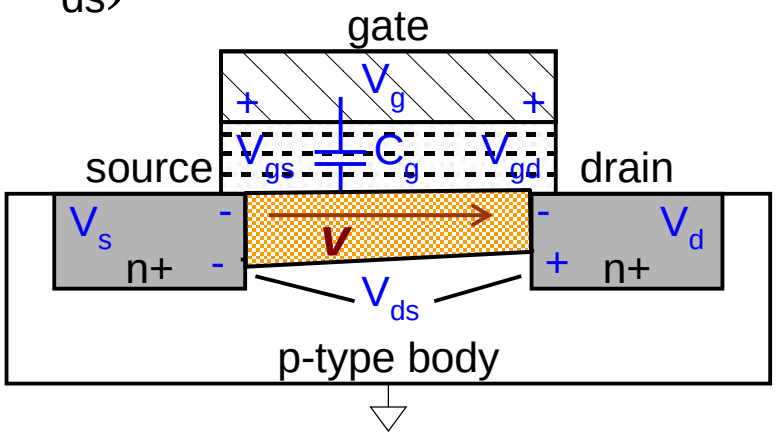
- Carrier velocity v proportional to lateral E-field

  $v = \mu.E$ μ called (electron) mobility

  (~ 500-600 $cm^2/V.s$ in heavily doped channel)

- Time for carrier to cross channel:

  $\tau = L / v = L^2/(\mu.V_{ds})$



nMOS Linear I/V

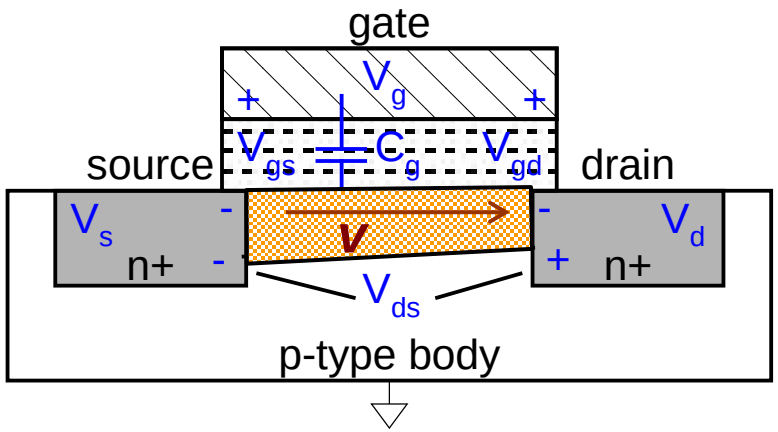
- Now we know
  - How much charge $Q_{channel}$ is in the channel
  - How much time τ each carrier takes to cross

$$I_{ds} = Q_{channel} / \tau$$

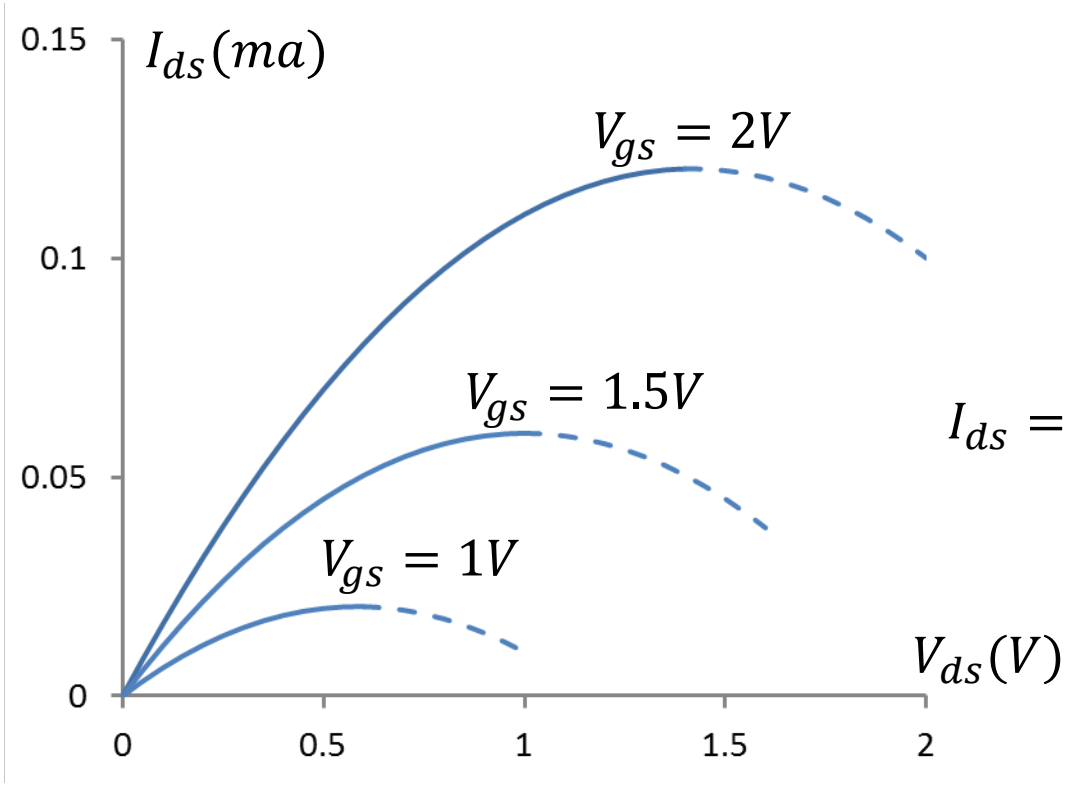
$$= (C_{ox}.W.L) \,.\, (V_{gs} - V_t - V_{ds}/2) \,/\, (L^2/(\mu.V_{ds}))$$

$$= \mu.\, C_{ox}.(W/L).\, (V_{gs} - V_t - V_{ds}/2).\, V_{ds}$$

$$= \beta.\, (V_{gs} - V_t - V_{ds}/2).\, V_{ds} \qquad \text{where } \beta = \mu.\, C_{ox}.(W/L)$$



nMOS IV – Linear Region



$$I_{ds} = \beta.\left(V_{gs} - V_t - \frac{V_{ds}}{2}\right).V_{ds}$$

- For small V_{ds}, I_{ds} increases linearly – behaves as a resistor
- As V_{ds} increases, charge in channel decreases
  - as a result: dI_{ds}/dV_{ds} decreases
- What happens when I_{ds} reaches it maximum ?

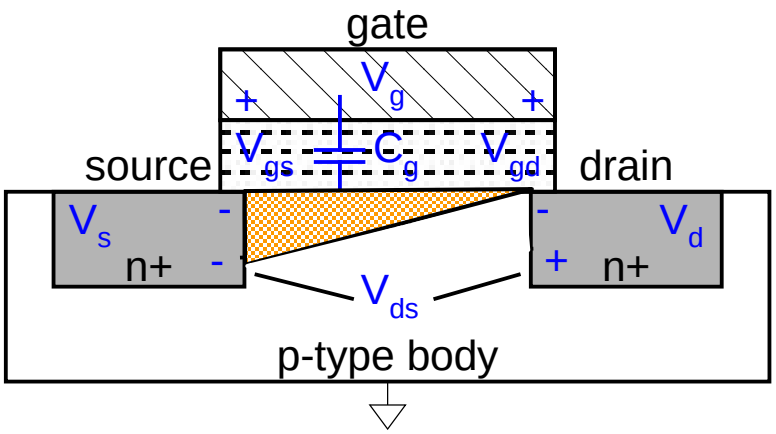
nMOS Saturation

- Suppose we increase V_{ds} until $V_{ds} = V_{gs} - V_t$
- Then $V_{gd} = V_{gs} – V_{ds} = V_t$
- The channel pinches off near drain
- We call this value of V_{ds} the saturation voltage:

$$V_{dsat} = V_{gs} – V_t$$

- At the point of saturation:

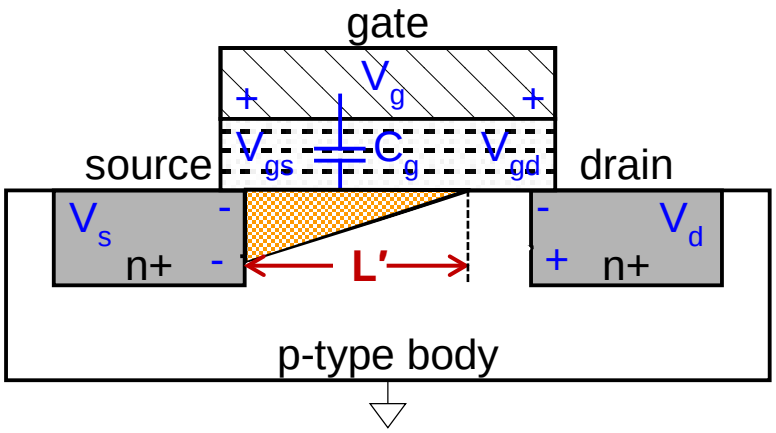
$$I_{dsat} = \beta . (V_{gs} – V_t – V_{dsat}/2) . V_{dsat} = (\beta/2) . (V_{gs} - V_t)^2$$



nMOS Saturation Channel Charge

- What happens if $V_{ds} > V_{dsat}$
- Pinch off extends from the drain towards the source
- Now, length of inverted channel is L′ (< L)
- Gate to channel capacitance is now

  $C = C_{ox}.W.L'$ (no inversion charge for x > L′)
- Average voltage across capacitor is $(V_{gs}-V_t)/2 = V_{dsat}/2$
- So $Q_{channel} = C_{ox}.W.L'.(V_{dsat}/2)$



nMOS Saturation I/V

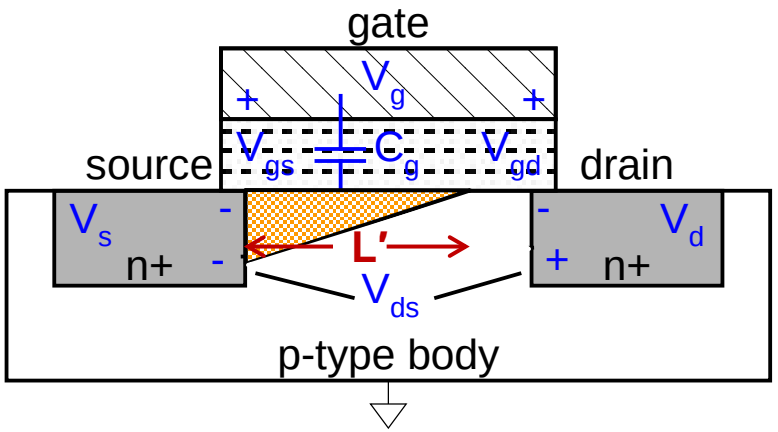
As before: $I_{ds} = Q_{channel} / \tau$

$= (C_{ox}.W.L'.(V_{dsat}/2) \bullet ((\mu.V_{dsat})/L'^2)$

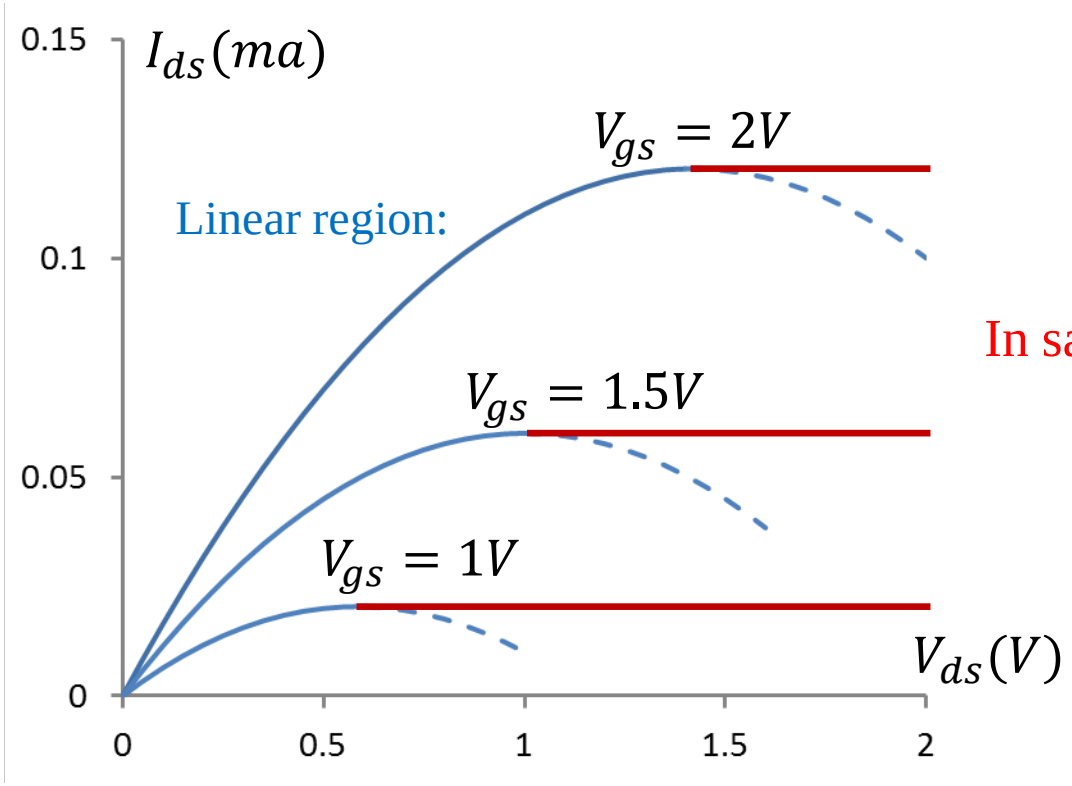
$= (\mu. C_{ox}.(W/L')/2).(V_{dsat})^2$ *but* $L' \approx L$ *for long channel*

$= (\beta/2). (V_{gs} - V_t)^2 = I_{dsat}$

- Note that $I_{ds} = I_{dsat}$ and is now independent of V_{ds}
  - MOS transistor in saturation behaves like a constant current source (with respect to V_{ds})
  - I_{dsat} has a square law dependence on V_{gs}



nMOS IV: Linear + Saturation Region



In saturation:

$$I_{ds} = (\beta/2).(V_{gs} - V_t)^2$$

- For $V_{ds} > V_{sat} = (V_{gs} - V_t)$, channel saturates
- Transistor behaves as a constant current source
  - I_{ds} independent of V_{ds}

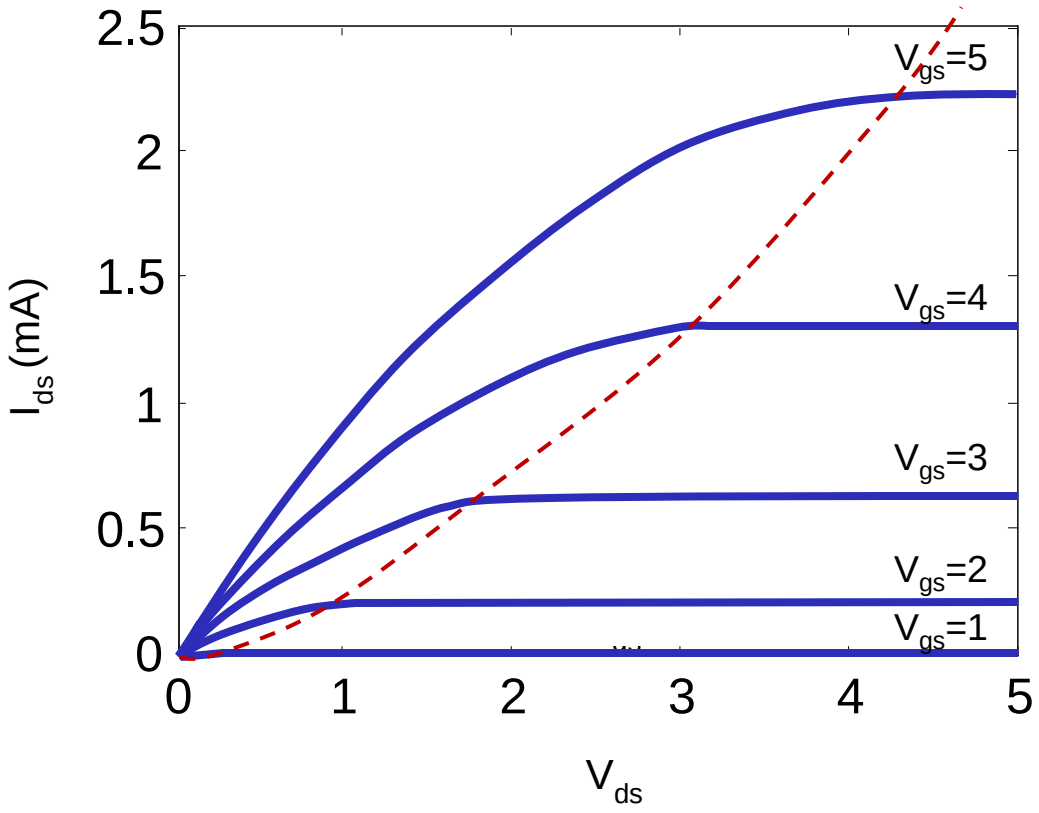
nMOS I/V Summary

- Shockley first-order model:
  - also known as ideal, long-channel model

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t \quad \text{cutoff} \\ \beta\left(V_{gs} - V_t - V_{ds}/2\right)V_{ds} & V_{ds} < V_{dsat} \quad \text{linear} \\ \frac{\beta}{2}\left(V_{gs} - V_t\right)^2 & V_{ds} > V_{dsat} \quad \text{saturation} \end{cases}$$

Example: 0.6μm process

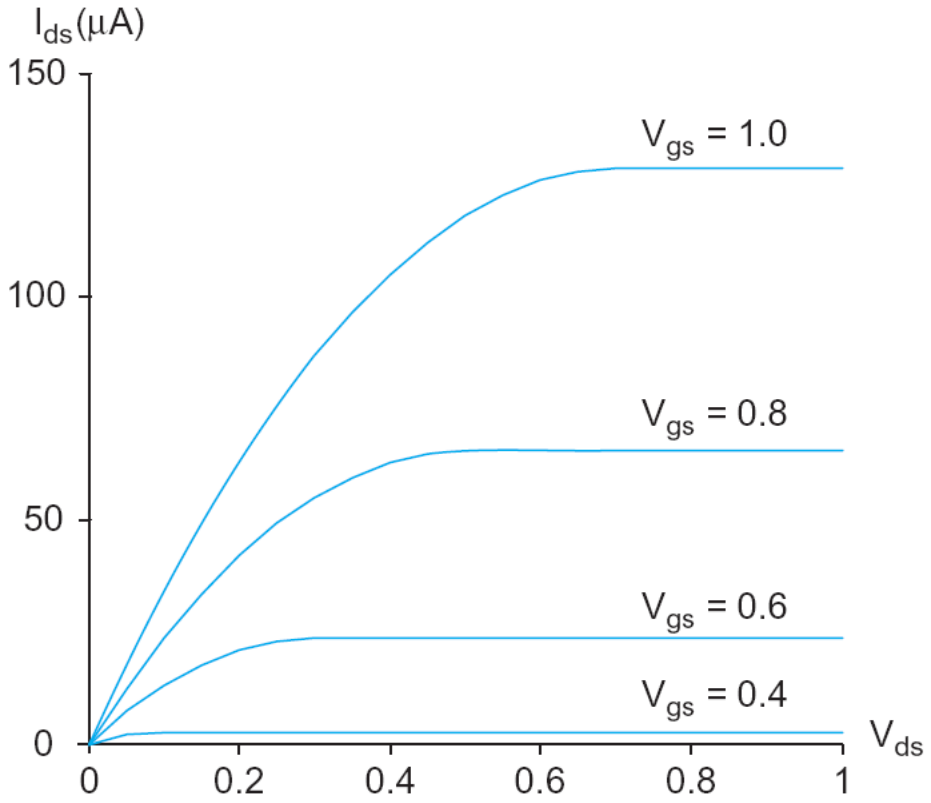
- t_{ox} = 100 Å
- μ = 350 cm²/V.s
- V_t = 0.7 V
- Use W/L = 4/2 λ



$$\beta = \mu C_{ox}\frac{W}{L} = (350)\left(\frac{3.9 \bullet 8.85\cdot 10^{-14}}{100\cdot 10^{-8}}\right)\left(\frac{W}{L}\right) = 120\frac{W}{L}\,\mu A/V^2$$

Example: 65 nm process

- t_{ox} = 10.5 Å
- μ = 80 $cm^2/V.s$
- V_t = 0.3 V
- Use W/L = 4/2 λ



$$\beta = 262 \,.\, (W/L) \;\; \mu A/V^2$$

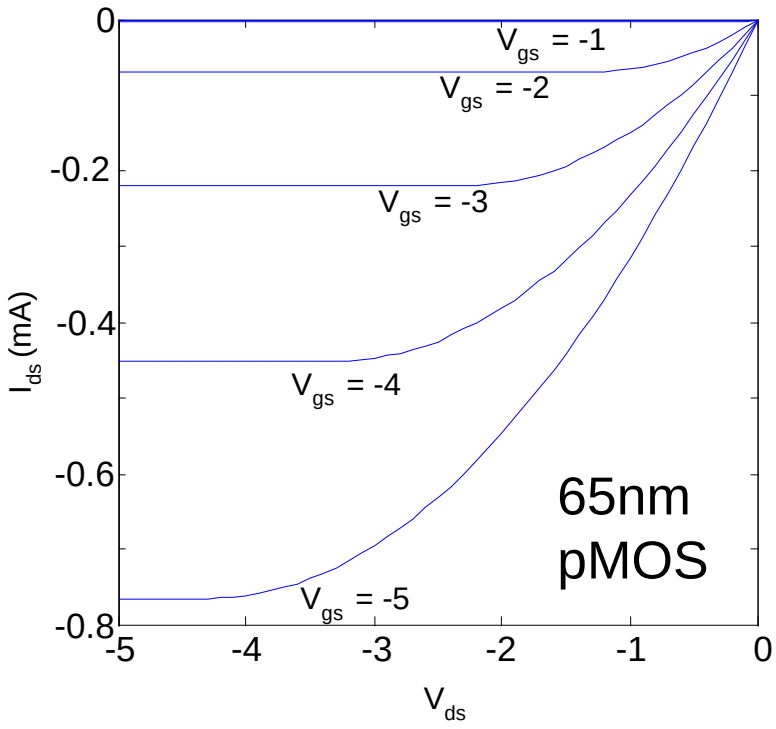
pMOS I/V Summary

- Shockley first-order model:
  - also known as ideal, long-channel model

$$I_{ds} = \begin{cases} 0 & V_{gs} > V_t \quad \text{cutoff} \\ -\beta\left(V_{gs} - V_t - {}^{V_{ds}}/_{2}\right)V_{ds} & V_{ds} > V_{dsat} \quad \text{linear} \\ -\dfrac{\beta}{2}\left(V_{gs} - V_t\right)^2 & V_{ds} < V_{dsat} \quad \text{saturation} \end{cases}$$

pMOS I/V

- All dopings and voltages are inverted for pMOS
  - Source is the more positive terminal
- Mobility μ_p is determined by holes
  - Typically 2-3x lower than that of electrons μ_n
  - 120 $cm^2/V.s$ in 0.6 μm process
- Thus pMOS must be wider to provide same current
- In this class:
  - assume $\mu_n / \mu_p = 2$

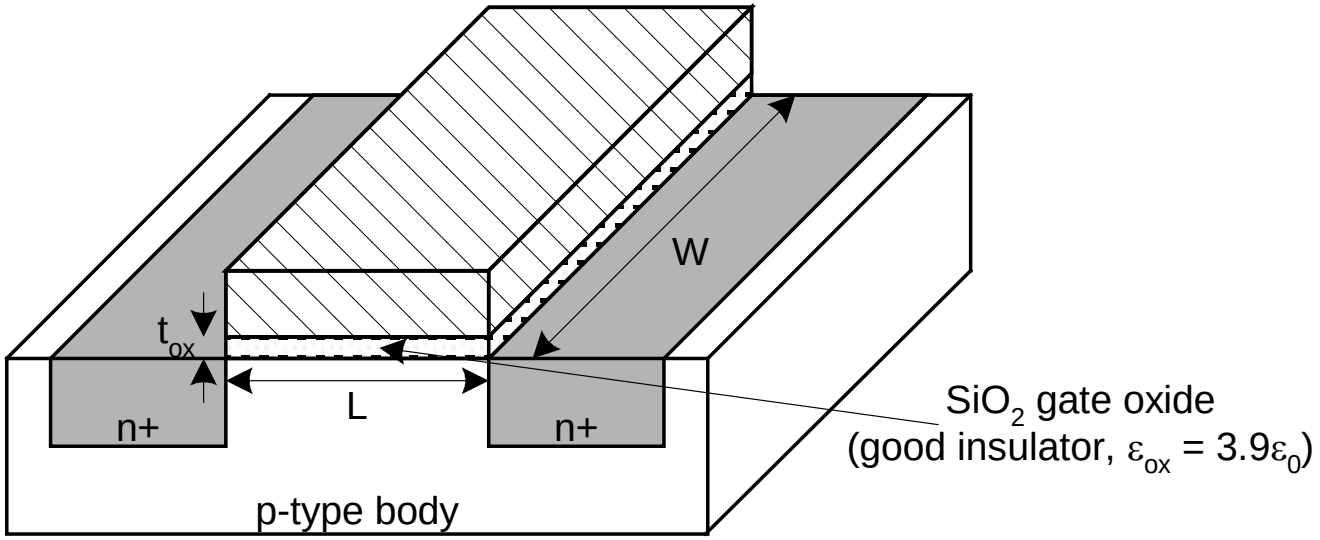


Capacitance

- Input to CMOS gate presents effectively infinite input resistance
- The dominant load in CMOS circuits is capacitance
- Capacitance exists wherever there are two conductors separated by a thin insulator
- Gate to channel capacitor is very important
  - Creates channel charge necessary for operation
- Source and drain have capacitance to body
  - Parasitic capacitance across reverse-biased diodes
  - Depletion region (insulator) separates N & P type conductors
  - Called diffusion capacitance because it is associated with source/drain diffusion

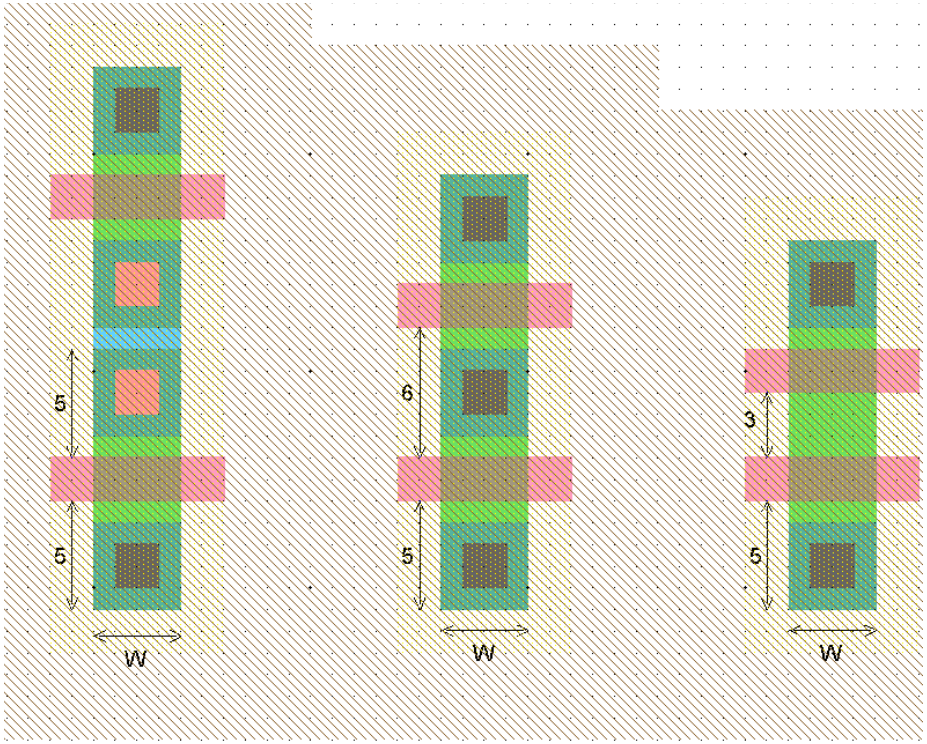
Gate Capacitance

- Gate is top plate of capacitor
- Assume bottom plate is source
  - In cut-off, bottom plate is actually the body
  - In linear mode, bottom plate is channel which is connected to source and drain
  - In saturation, bottom plate is channel connected to source
- $C_{gs} = \varepsilon_{ox}.W.L/t_{ox} = C_{ox}.W.L = C_{permicron}.W$
- $C_{permicron}$ is typically about 1-2 fF/μm of width



Diffusion Capacitance

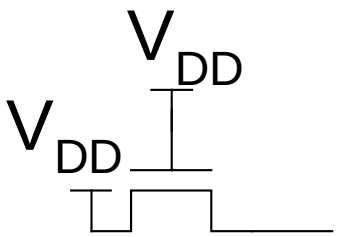
- C_{sb}, C_{db}
- Diffusion region is resistive and capacitive (to body)
- Capacitance depends on area and perimeter
- Use small as possible diffusion nodes
- Comparable to C_g for contacted diffusion
- Use $C_g/2$ for merged
- Varies with process



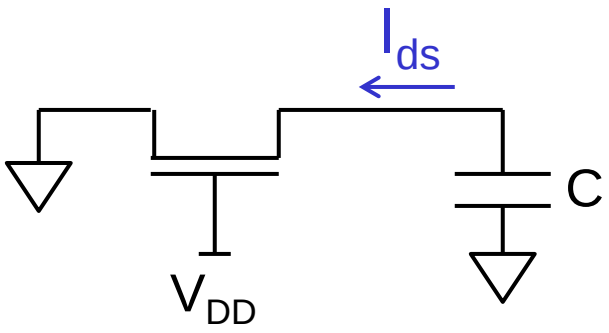
Isolated Diffusion	Shared Diffusion	Merged Diffusion
≈ G_g	≈ G_g	≈ $G_g/2$
$C_{node} = 2.C_g$	$C_{node} = C_g$	$C_{node} = C_g/2$

Pass Transistors

- We have assumed source is grounded
  - or at least close to ground, pulling drain down
- What if source >> 0?
  - e.g. nMOS pass transistor passing V_{DD}
- $V_g = V_{DD}$
  - If $V_s > V_{DD}-V_t$, $V_{gs} < V_t$
  - Hence transistor would turn itself off
- nMOS pass transistors pull no higher than $V_{DD}-V_{tn}$
  - Produces a degraded “1”
  - Approach degraded value slowly (low I_{ds})
- pMOS pass transistors pull no lower than $|V_{tp}|$
  - Transmission gates are needed to pass both “good” 0 and “good” 1

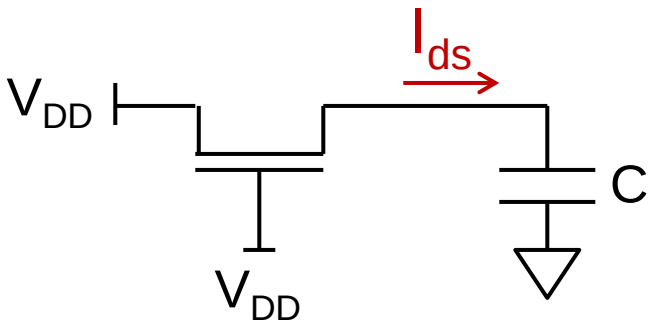


Degraded Time Constant



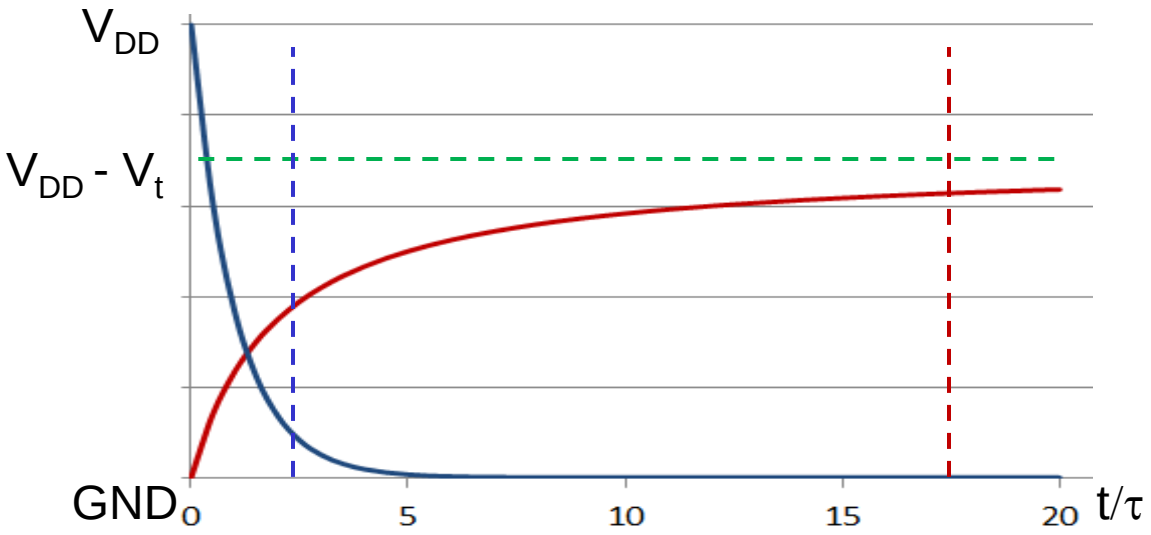
pull-down device is in linear (mostly):

$I_{ds} \alpha V_s$



pull-up device is in saturation:

$I_{ds} \alpha ((V_{DD} - V_t) - V_s)^2$

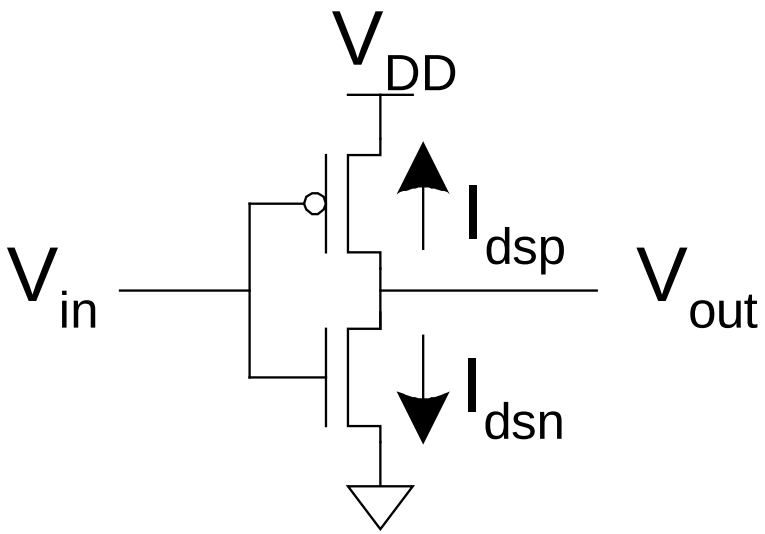


Pull-up time constant (to 90% final value) can be 6x pull-down time constant!

DC Response: Inverter

- Digital circuits are merely analog circuits used over a constrained portion of their range
- Derive DC transfer function for static CMOS inverter
- When $V_{in} = 0 \Rightarrow V_{out} = V_{DD}$
- When $V_{in} = V_{DD} \Rightarrow V_{out} = 0$
- In between, both transistors may be conducting
- By KCL, V_{out} must settle so that

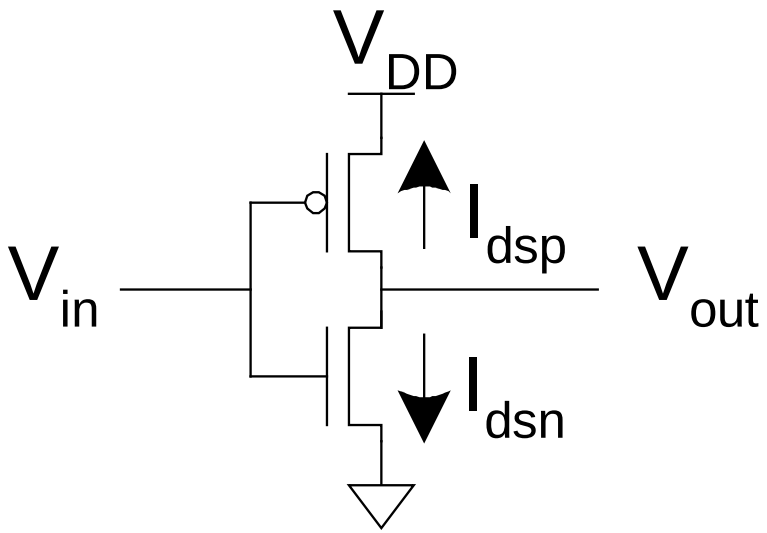
  $I_{dsn} = |I_{dsp}|$



- We could solve equations, but …
- Graphical solution gives more insight

Transistor Operation

- Current (I_{dsn}, I_{dsp}) depends on region of transistor behavior
- For what V_{in} and V_{out} are nMOS and pMOS in
- Cutoff?
- Linear?
- Saturation?

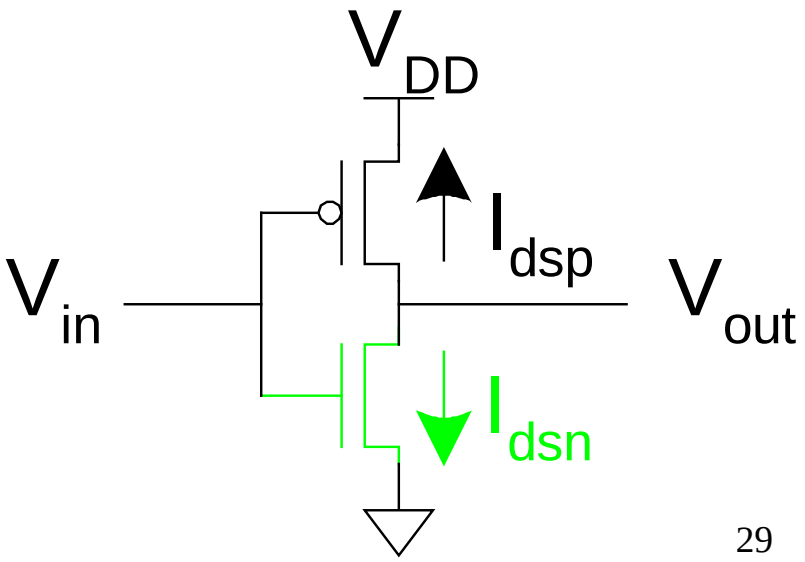


Inverter: nMOS Operation

Cutoff	Linear	Saturated
$V_{gsn} < V_{tn}$	$V_{gsn} > V_{tn}$	$V_{gsn} > V_{tn}$
$V_{in} < V_{tn}$	$V_{in} > V_{tn}$	$V_{in} > V_{tn}$
	$V_{dsn} < V_{gsn} - V_{tn}$	$V_{dsn} > V_{gsn} - V_{tn}$
	$V_{out} < V_{in} - V_{tn}$	$V_{out} > V_{in} - V_{tn}$

$V_{gsn} = V_{in}$

$V_{dsn} = V_{out}$



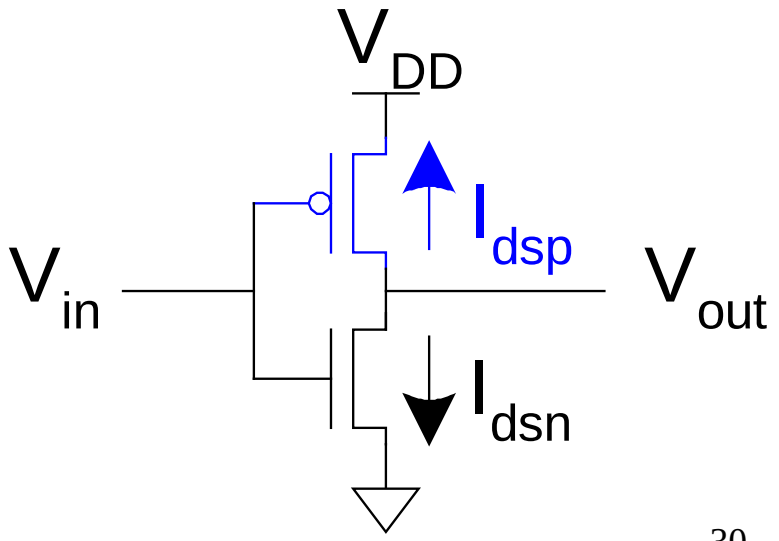
Inverter: pMOS Operation

Cutoff	Linear	Saturated
$V_{gsp} > V_{tp}$	$V_{gsp} < V_{tp}$	$V_{gsp} < V_{tp}$
$V_{in} > V_{DD} + V_{tp}$	**$V_{in} < V_{DD} + V_{tp}$**	**$V_{in} < V_{DD} + V_{tp}$**
	$V_{dsp} > V_{gsp} - V_{tp}$	$V_{dsp} < V_{gsp} - V_{tp}$
	$V_{out} > V_{in} - V_{tp}$	**$V_{out} < V_{in} - V_{tp}$**

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

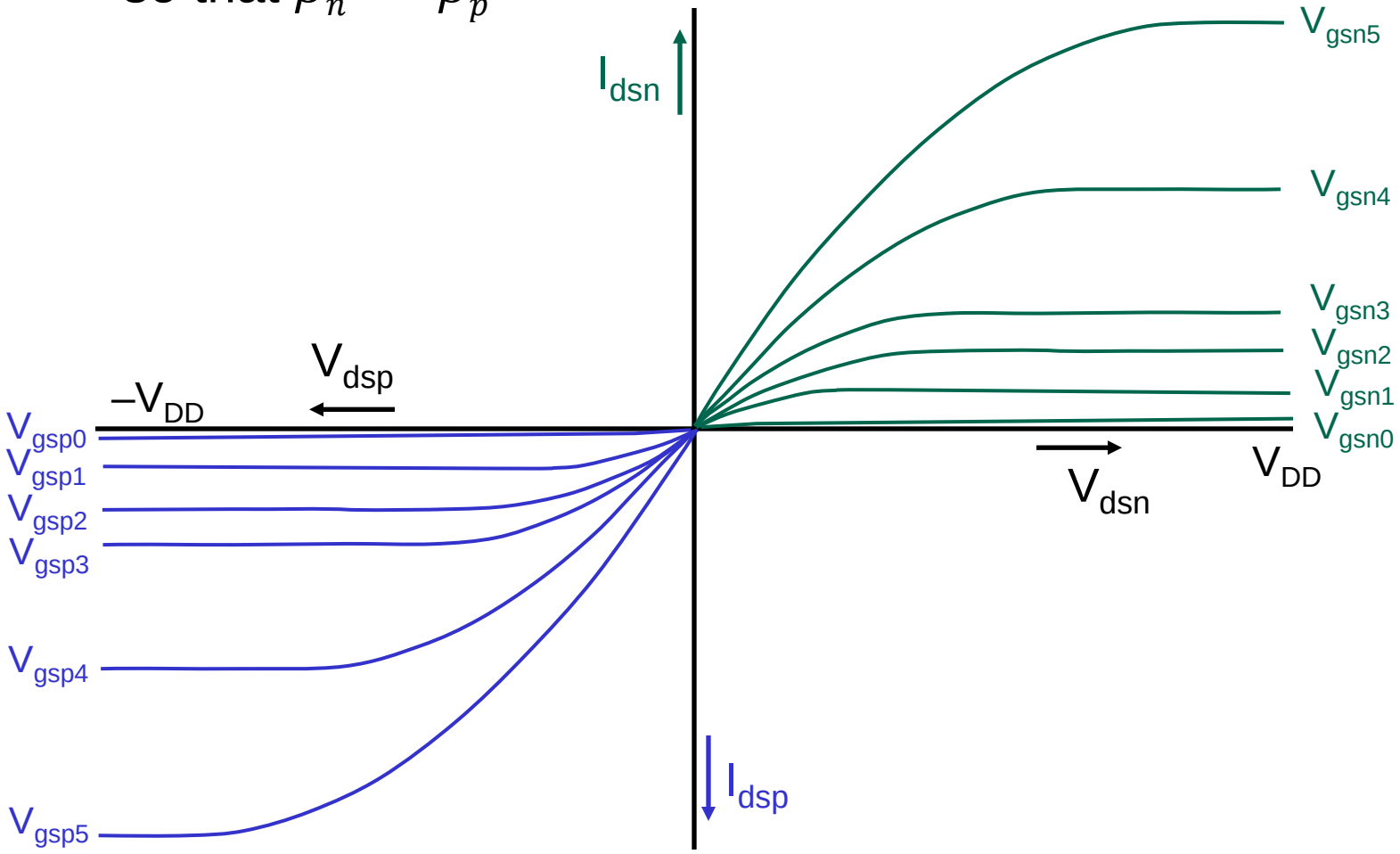
$$V_{dsp} = V_{out} - V_{DD}$$

(remember: V_{gsp}, V_{dsp} and $V_{tp} < 0$)

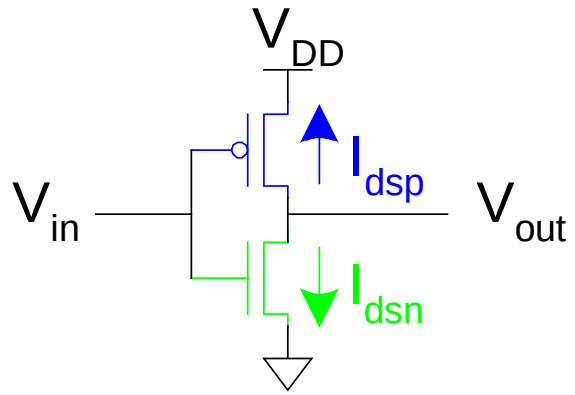


I-V Characteristics

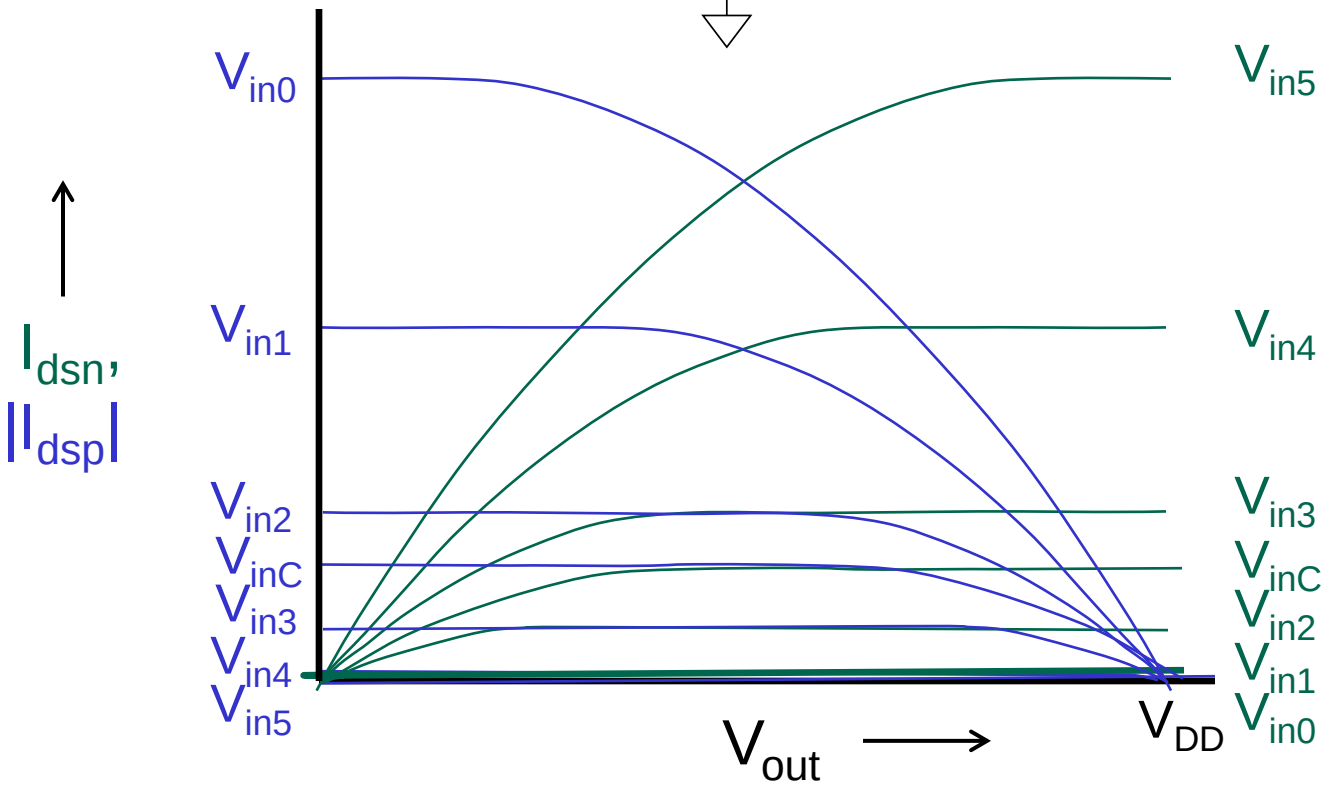
- Mobility of holes is 2-3x less than mobility of electrons
- Usually make pMOS 2x wider than nMOS
  - so that $\beta_n \approx \beta_p$



Replot I-V as function of V_{out} & V_{in}

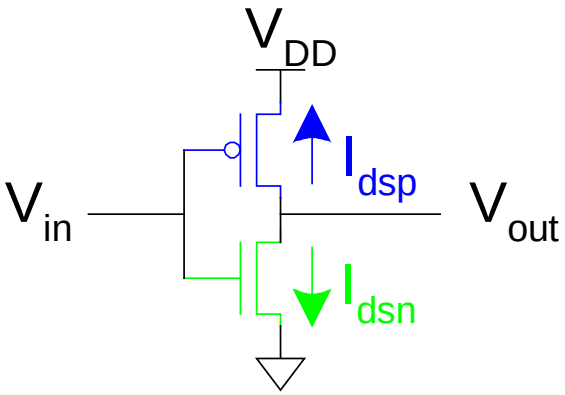


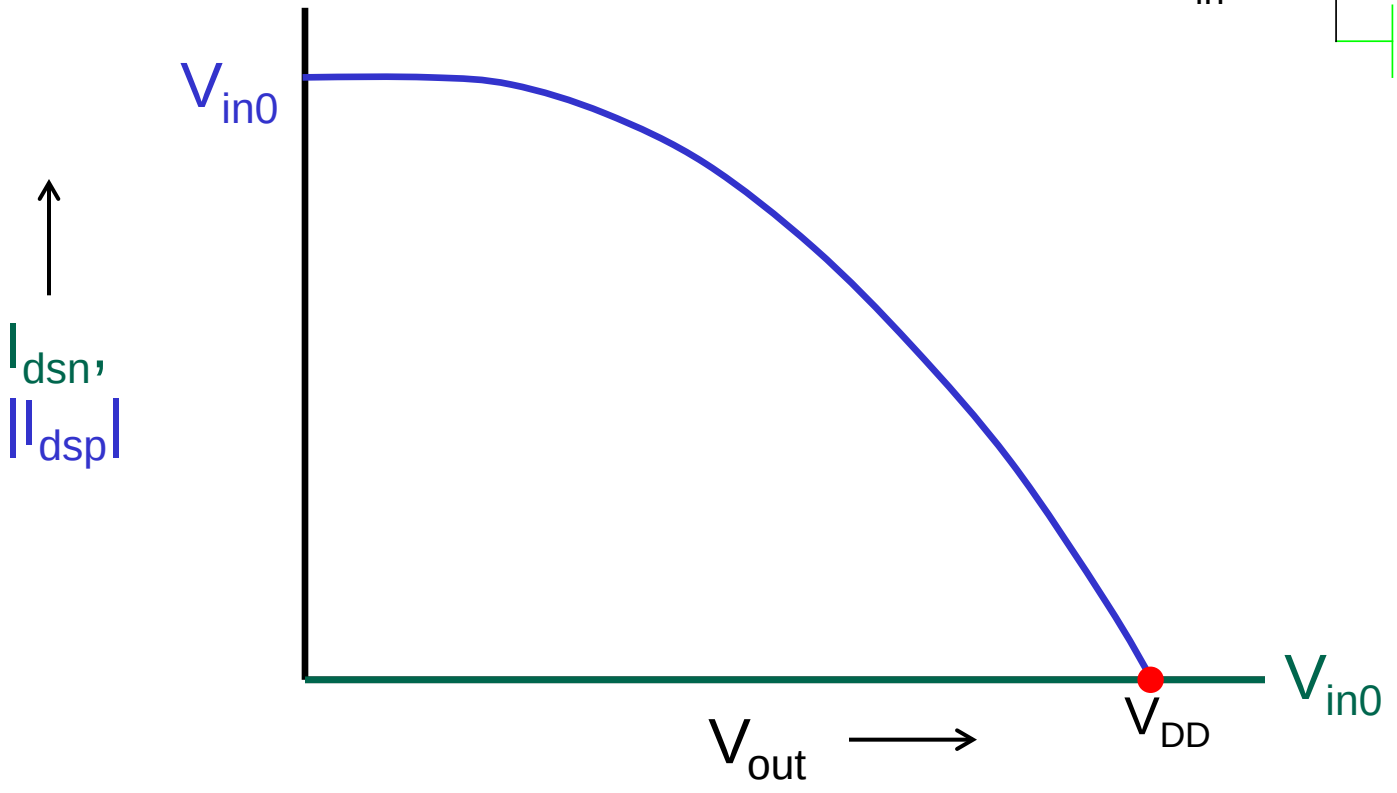
V_{out} is where $I_{dn} = -I_{dp}$



Load Line Analysis

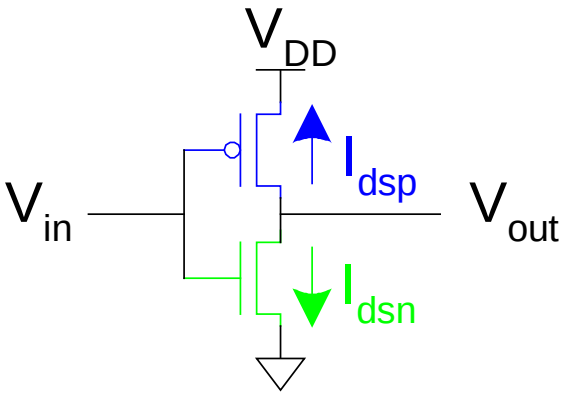
- $V_{in} = 0$

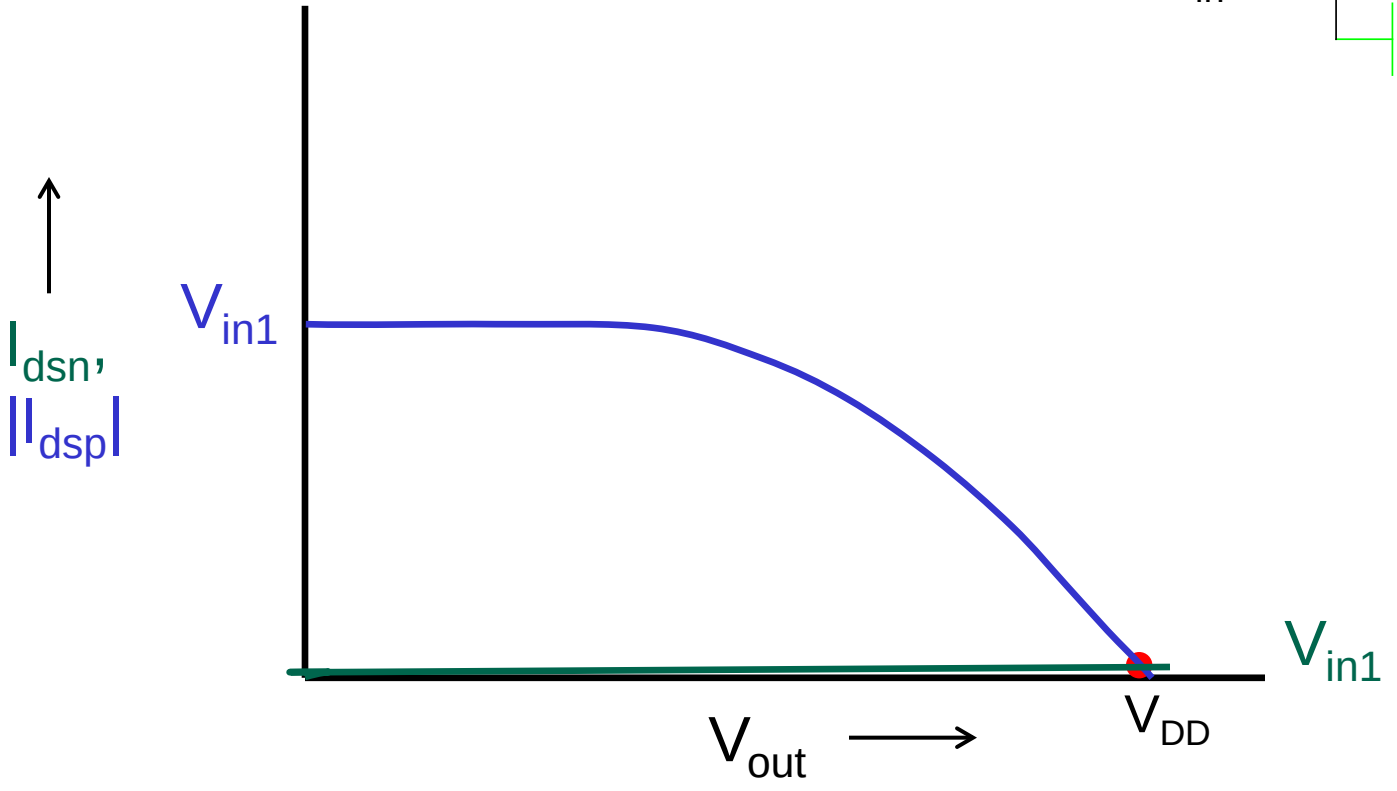




Load Line Analysis

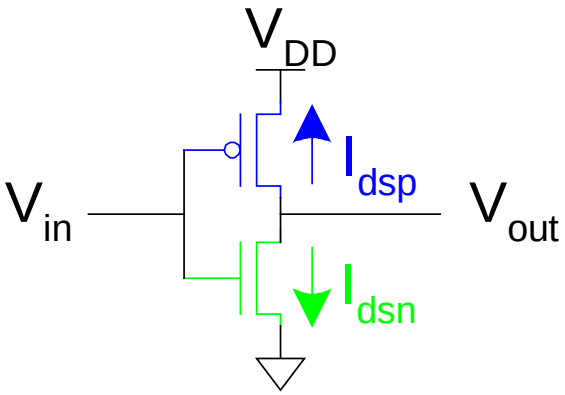
- $V_{in} = 0.2V_{DD}$

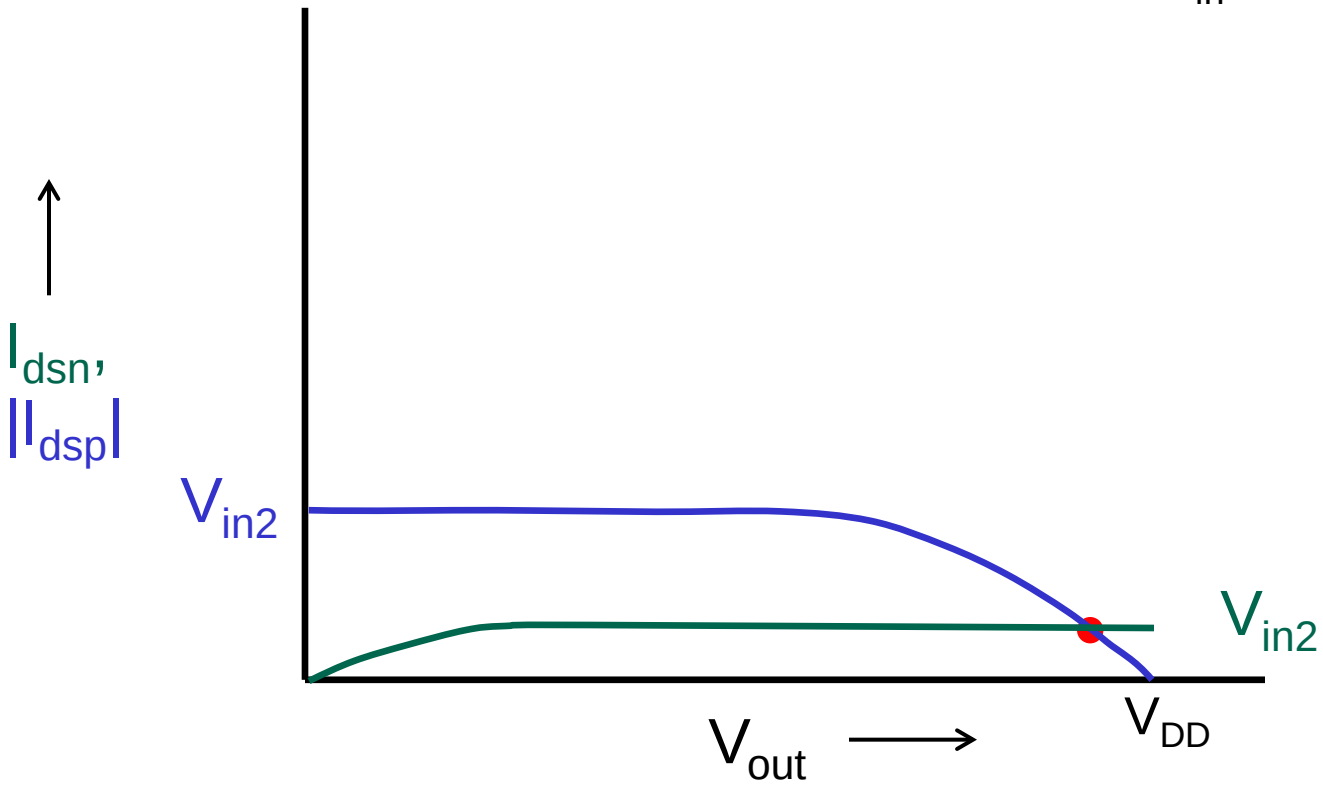




Load Line Analysis

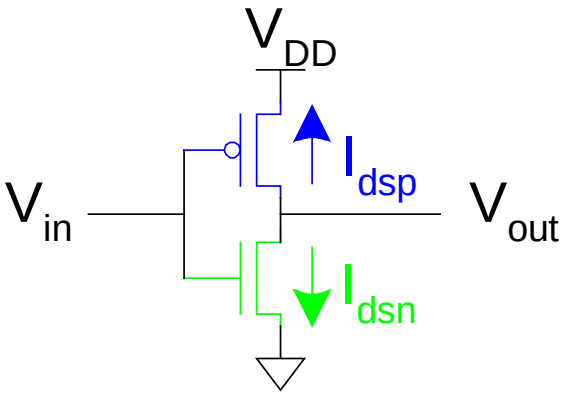
- $V_{in} = 0.4V_{DD}$

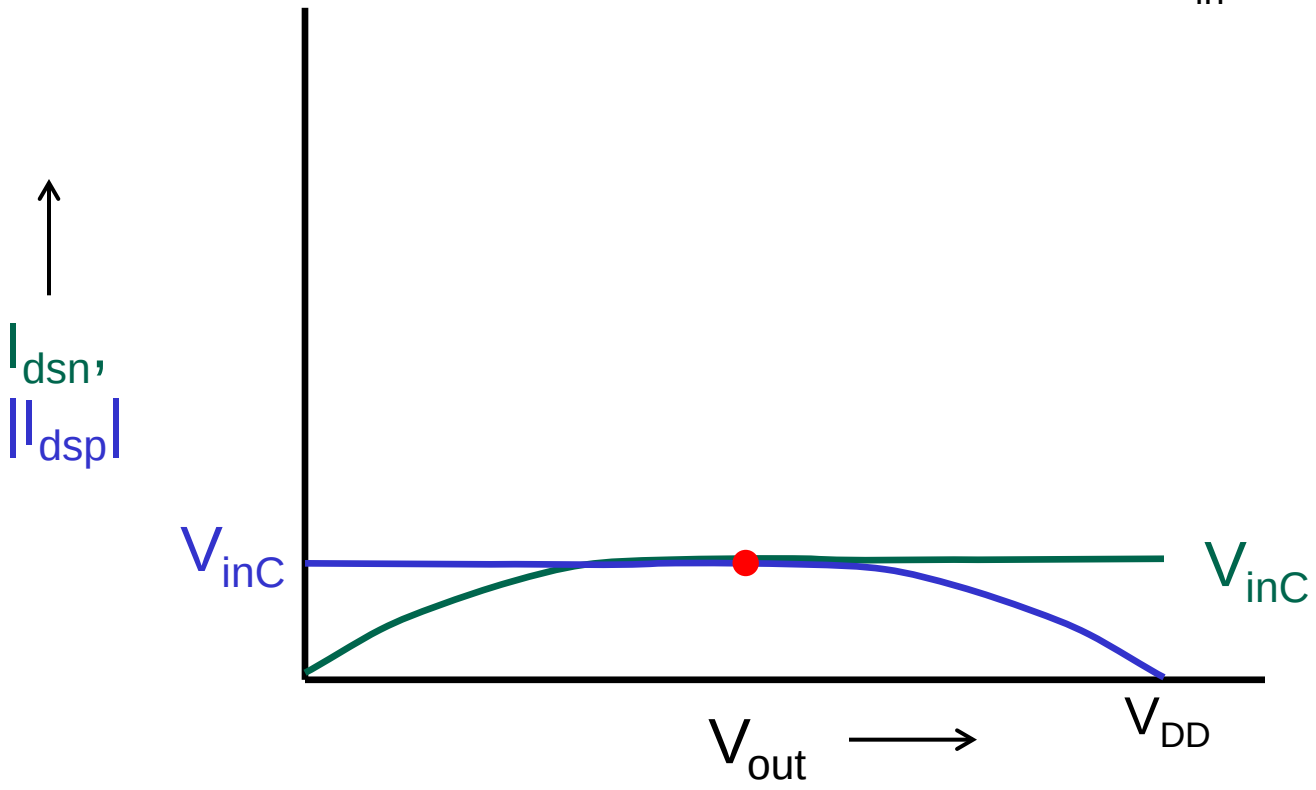




Load Line Analysis

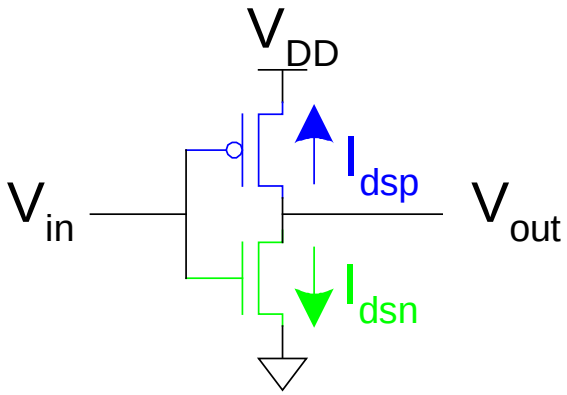
- $V_{in} = 0.5V_{DD}$

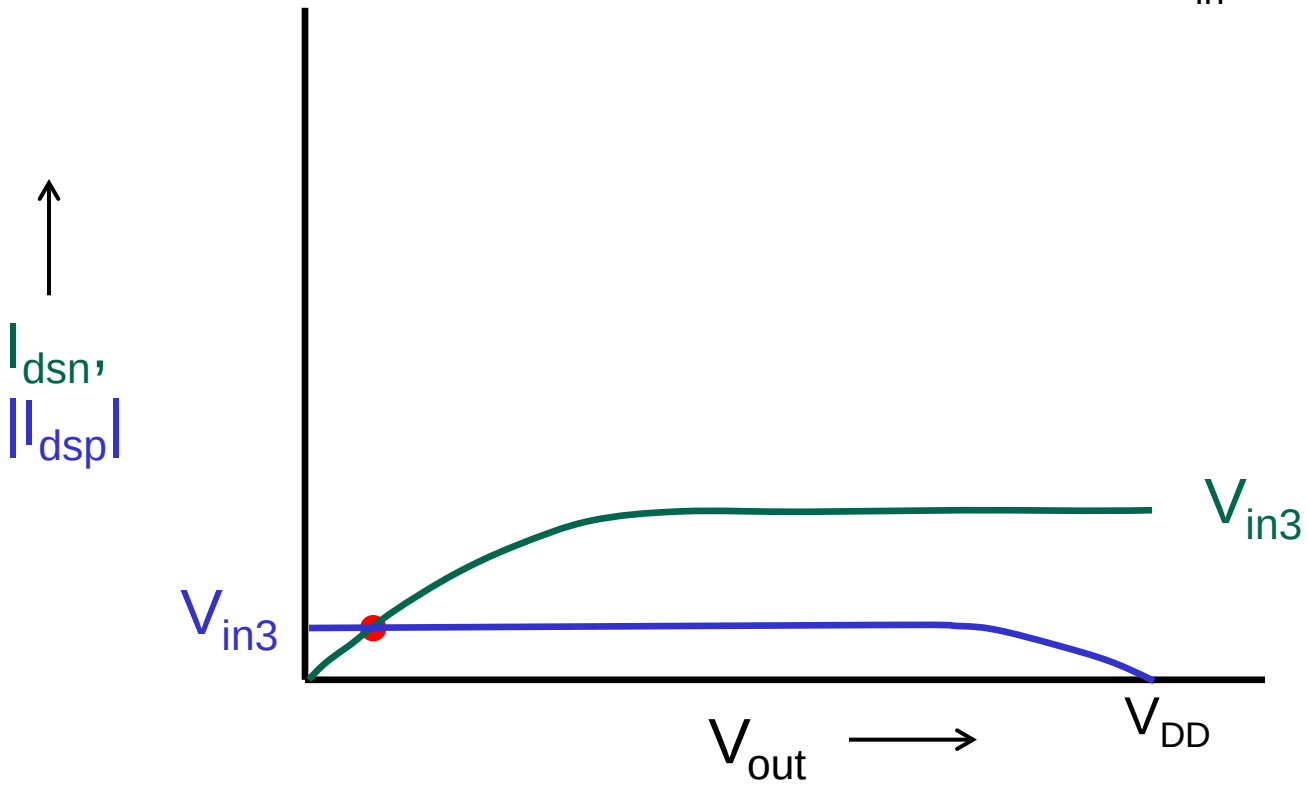




Load Line Analysis

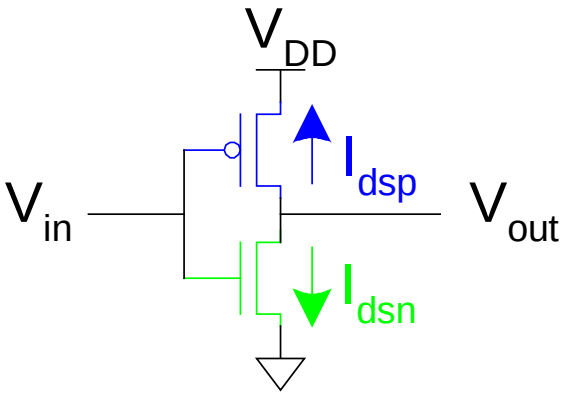
- $V_{in} = 0.6V_{DD}$

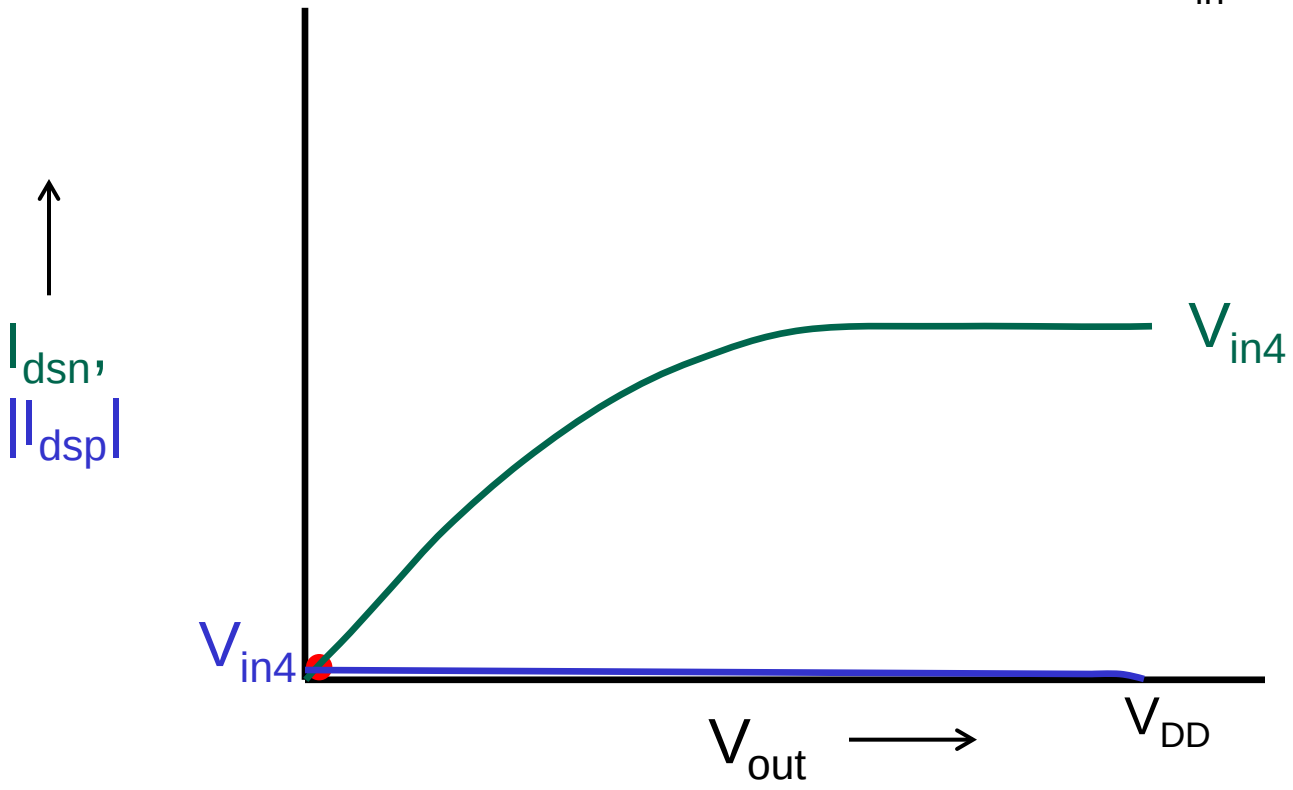




Load Line Analysis

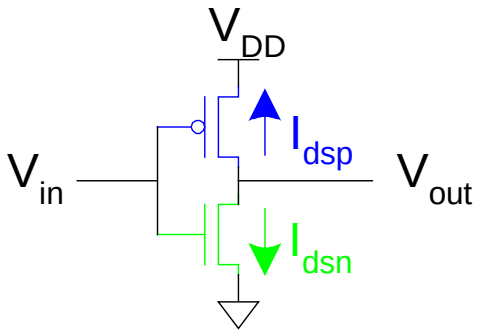
- $V_{in} = 0.8V_{DD}$

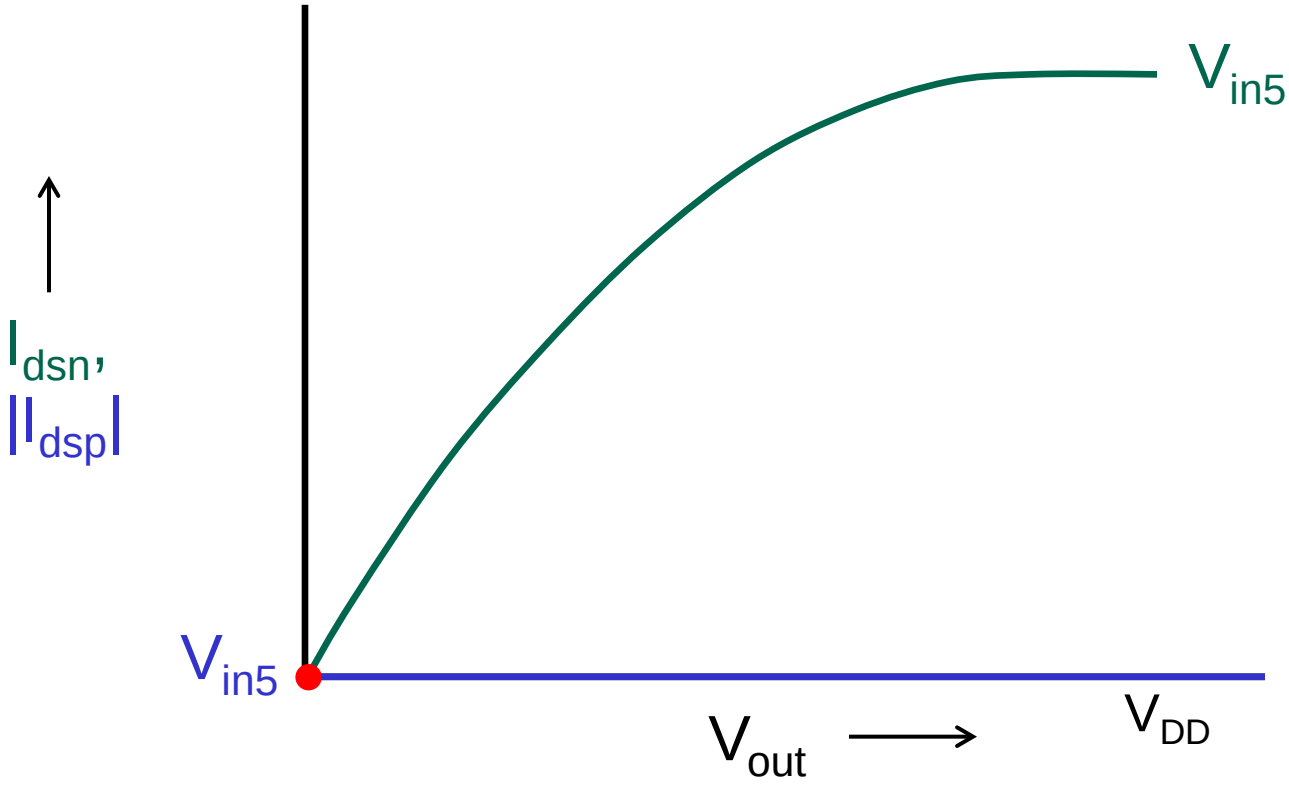




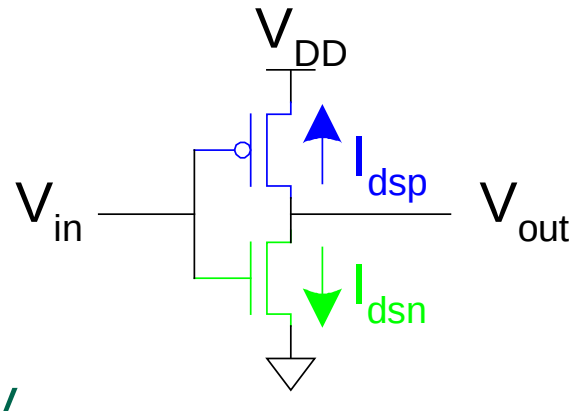
Load Line Analysis

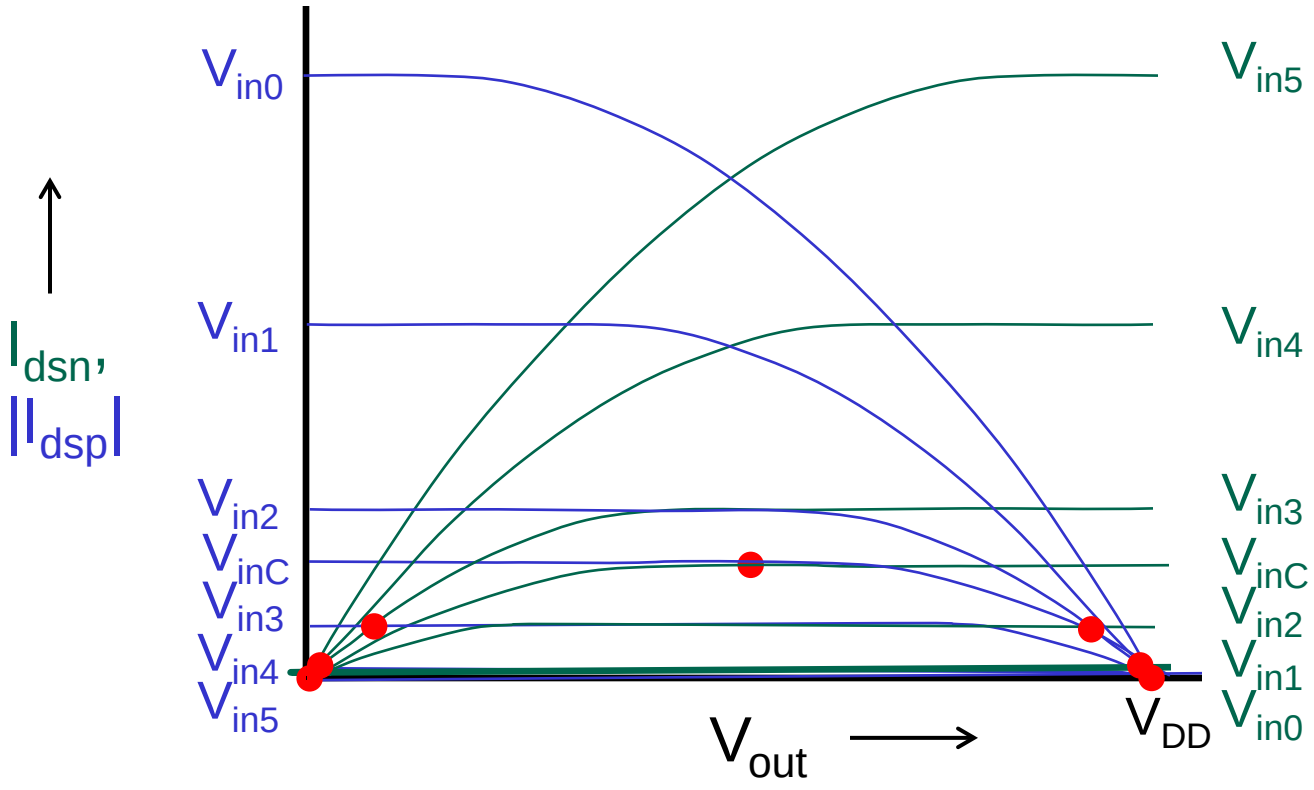
- $V_{in} = V_{DD}$





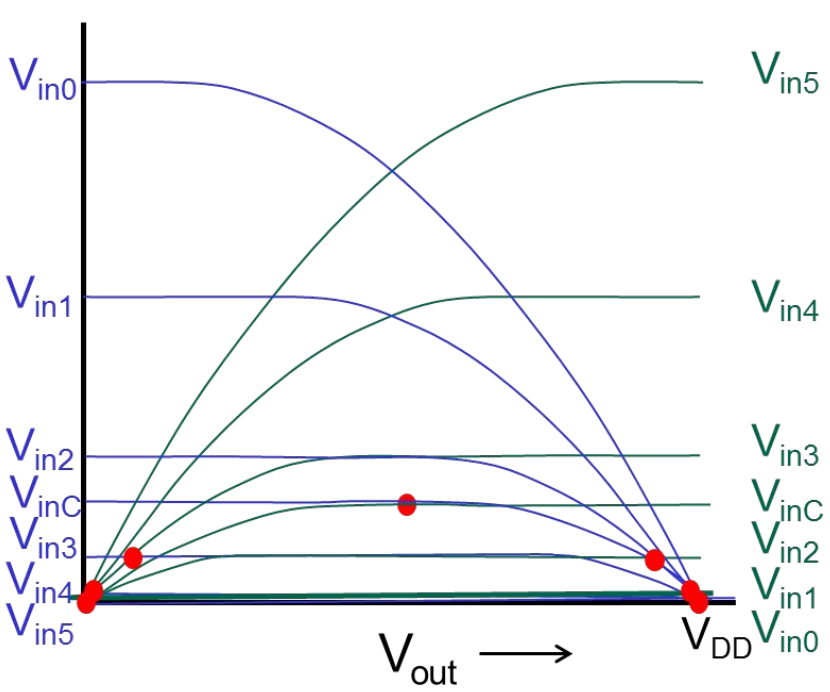
Load Line Analysis

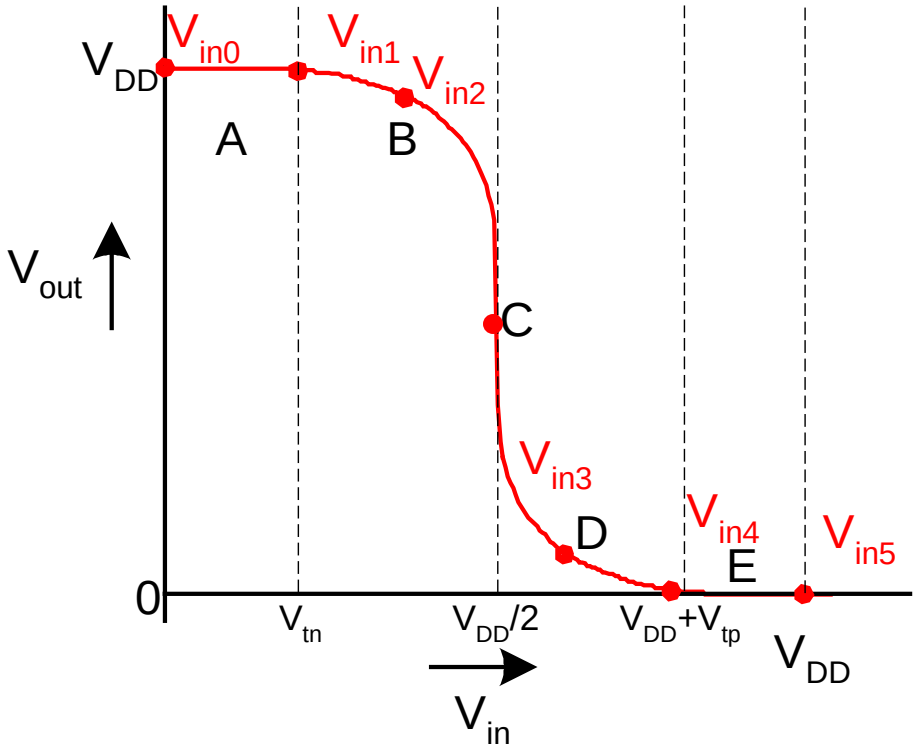




DC Transfer Curve

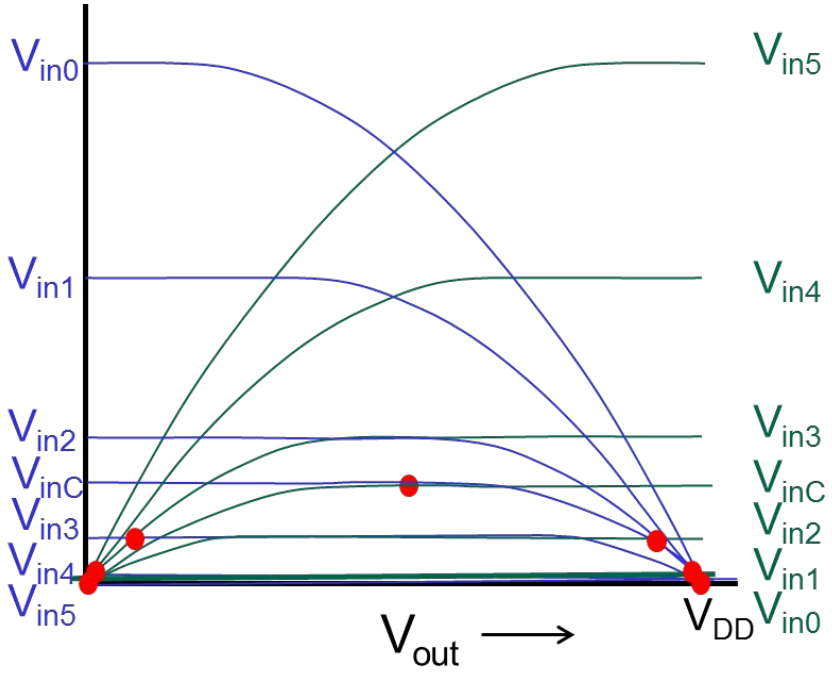
- Trans-scribe points onto V_{in} vs. V_{out} plot

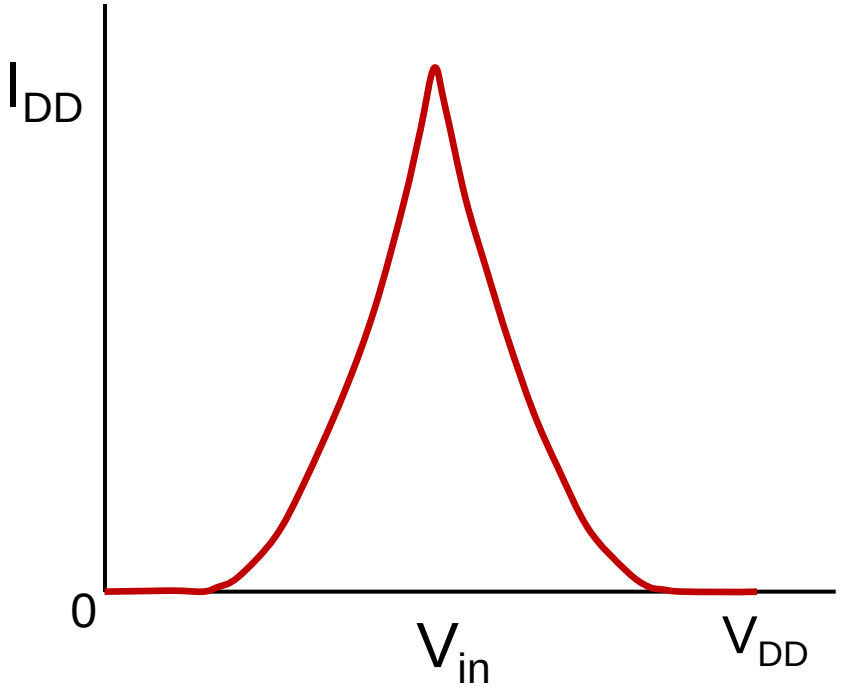




Supply Current

- $I_{DD} = I_{dsn} = -I_{dsp}$

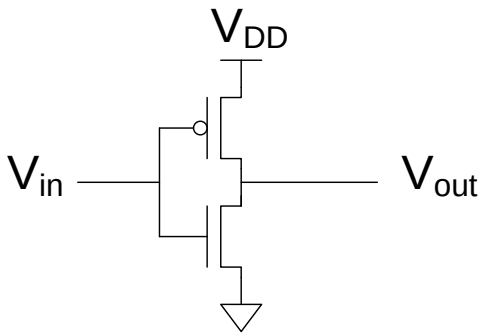




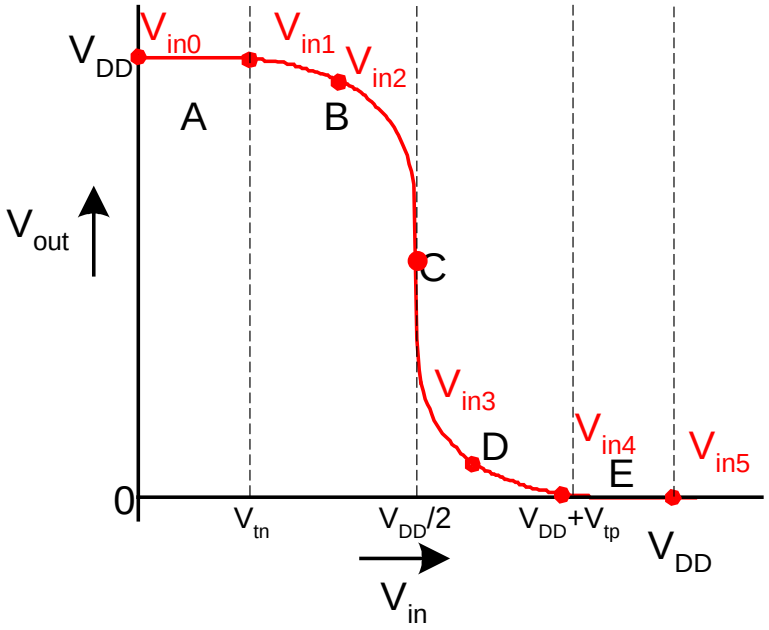
- Zero current when in normal logic range
- Transient current pulse drawn from V_{DD} supply on each switching event

Operating Regions

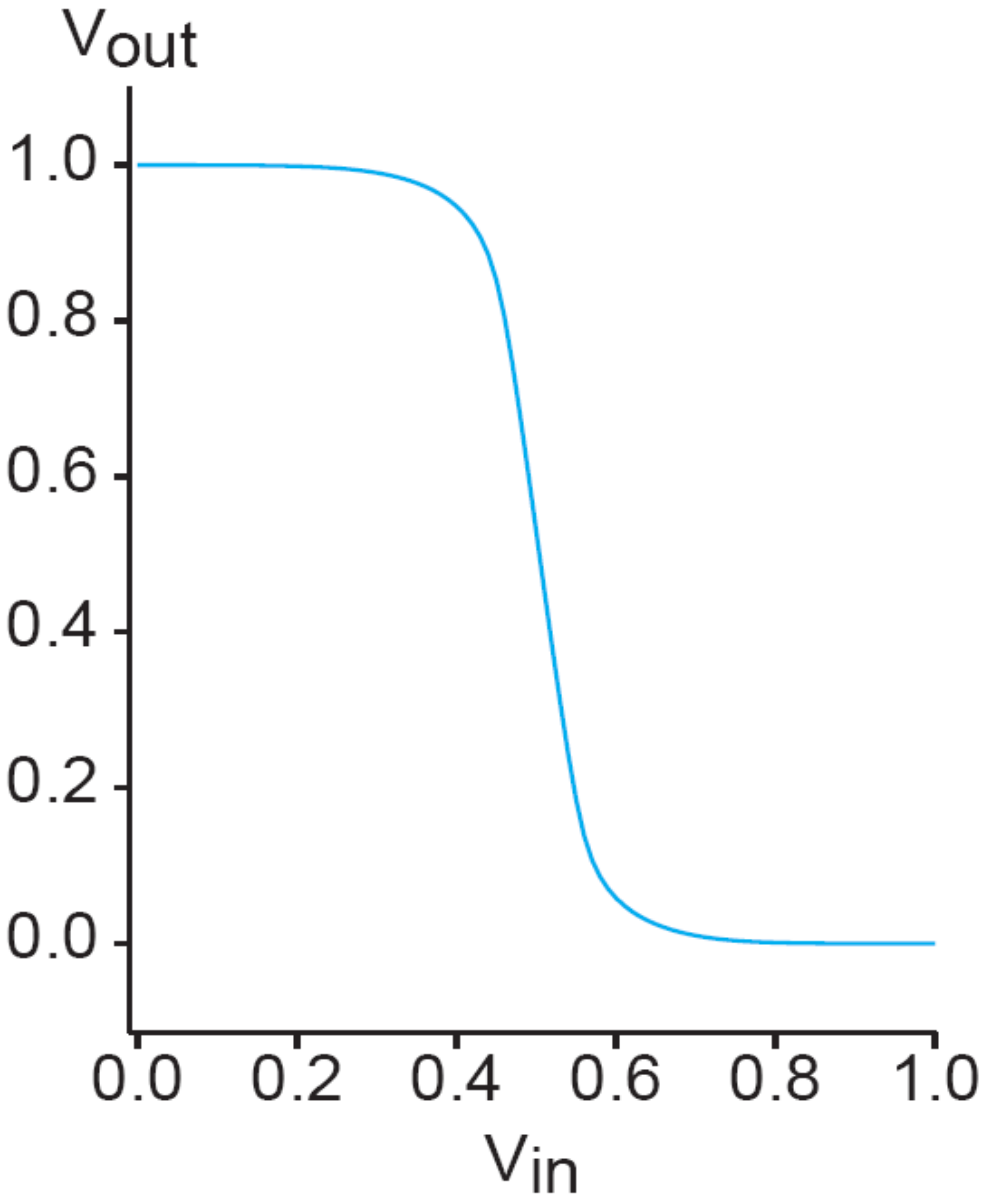
- Re-visit operating regions



Region	nMOS	pMOS
A	Cutoff	Linear
B	Saturation	Linear
C	Saturation	Saturation
D	Linear	Saturation
E	Linear	Cutoff

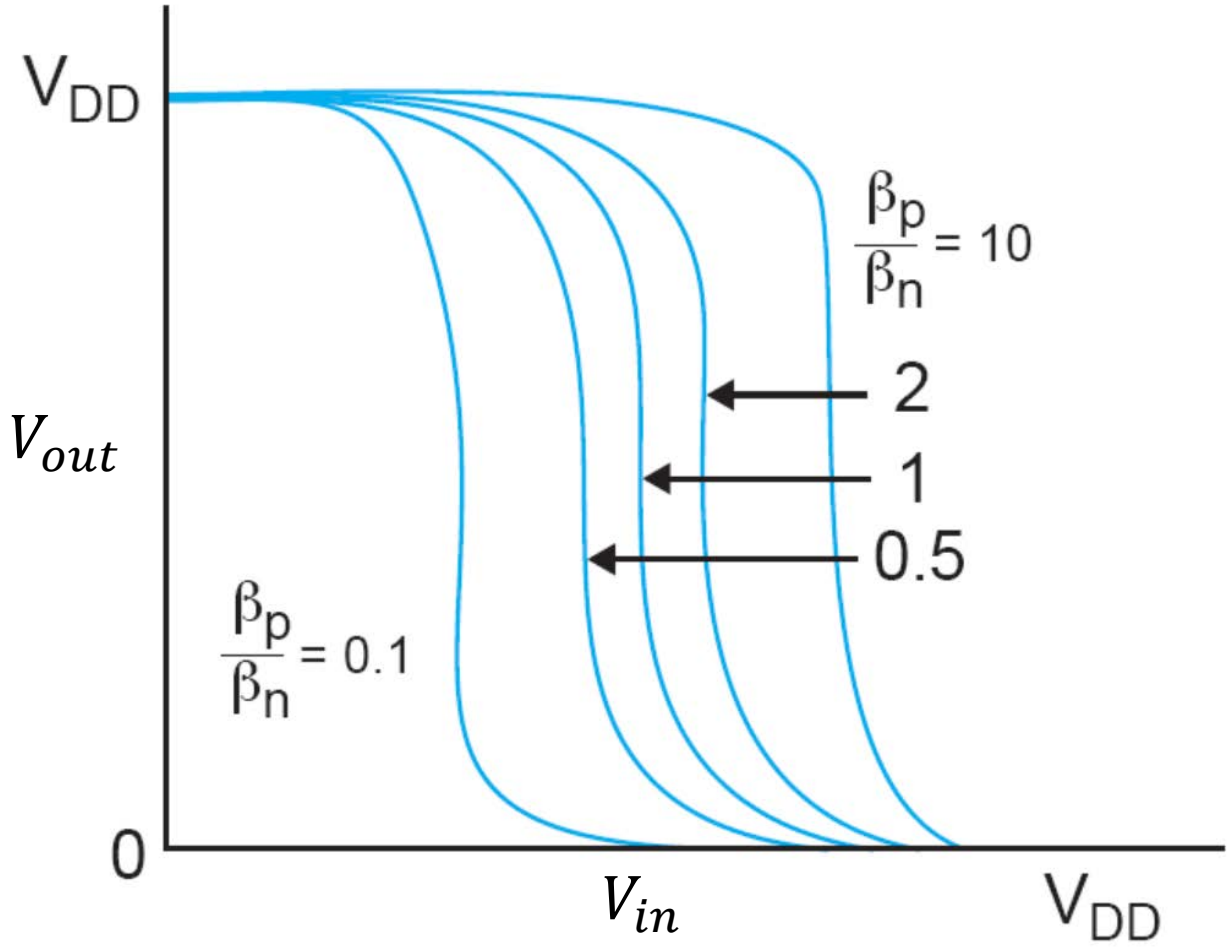


Simulated 65nm DC Characteristic



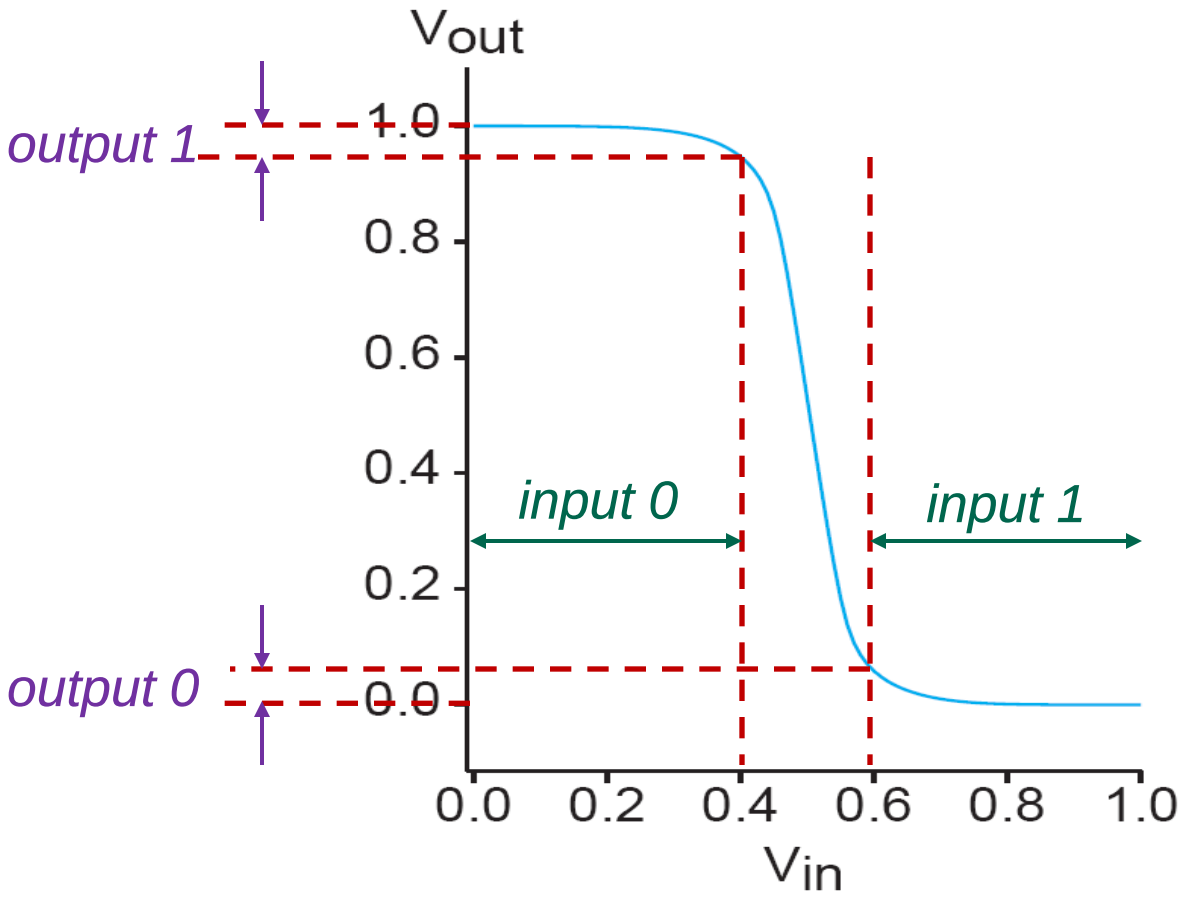
Beta Ratio

- If $\beta_p / \beta_n \neq 1$, switching point will move from $V_{DD}/2$
- Called *skewed* gate
- Other gates: collapse into equivalent inverter



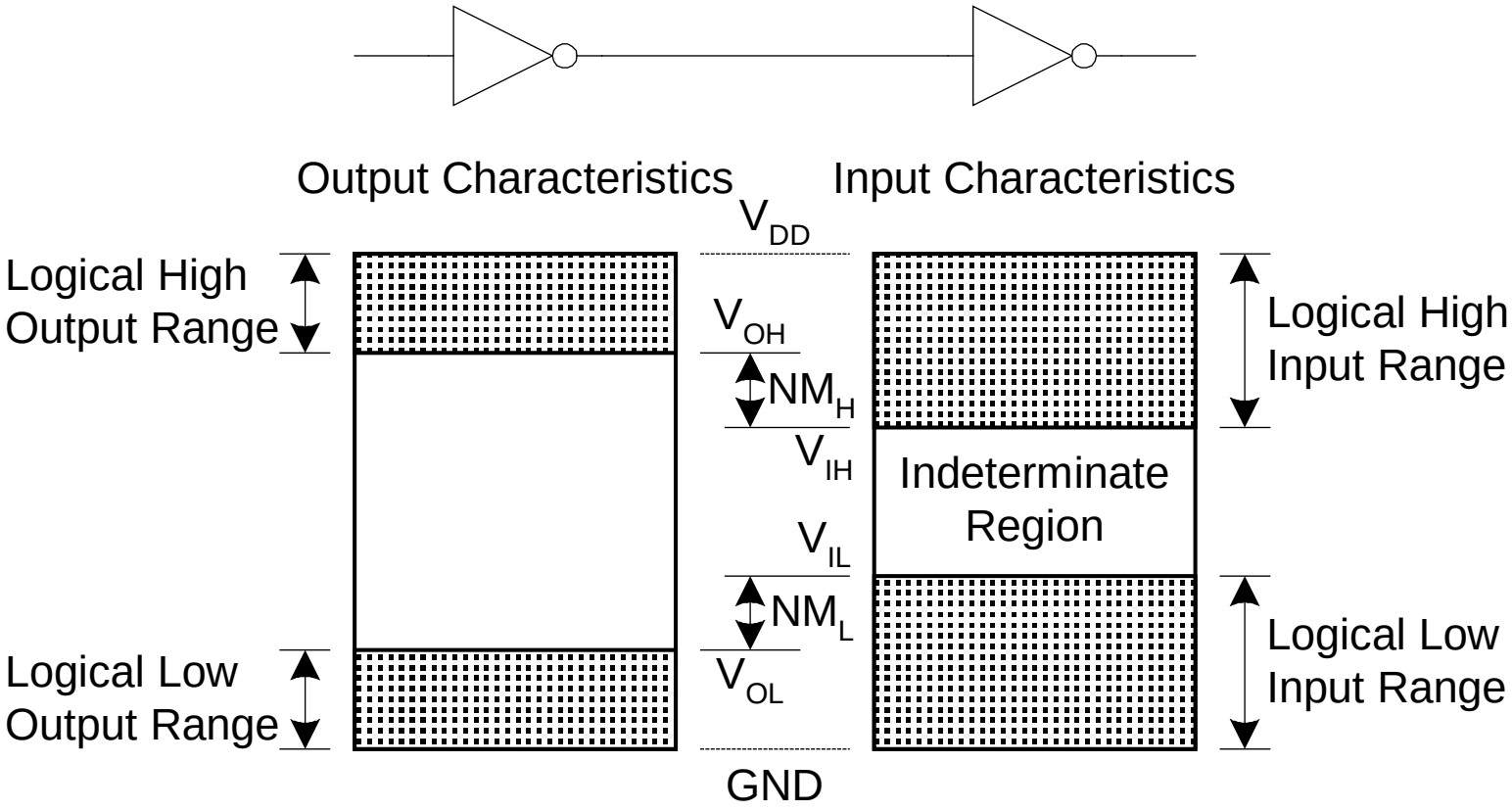
Restoring Logic

- Reason that we can build digital circuits with millions of gates and always get same answer is:
- Most CMOS logic gates are “restoring”
  - output logic level is better than input logic level



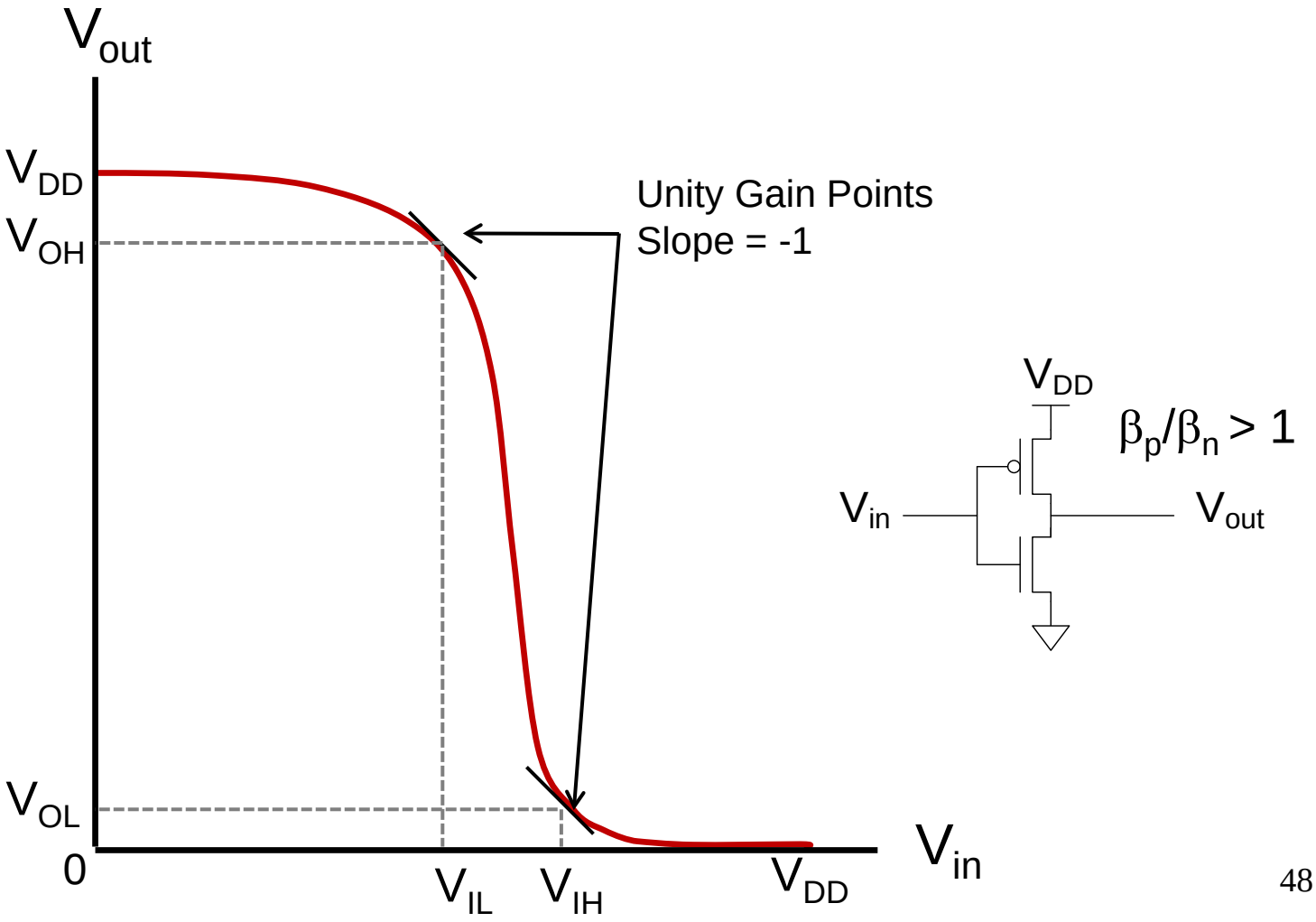
Noise Margins

- How much noise can a gate input see before it does not recognize the input?



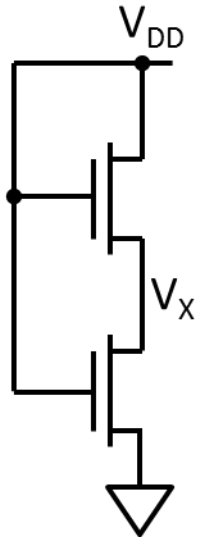
Nominal Logic Levels

- To maximize noise margins, select worst case logic levels at
  - unity gain point of DC transfer characteristic



Example: MOS IV Formula

Suppose we connect two identical nMOS devices in series between VDD and GND and connect the gates of each to VDD:



Assuming $V_{DD} > V_T$,

1. In which region is the upper transistor operating? Why?
2. In which region is the lower transistor operating? Why?
3. Derive an expression for the voltage V_x at the intermediate node