Module

2

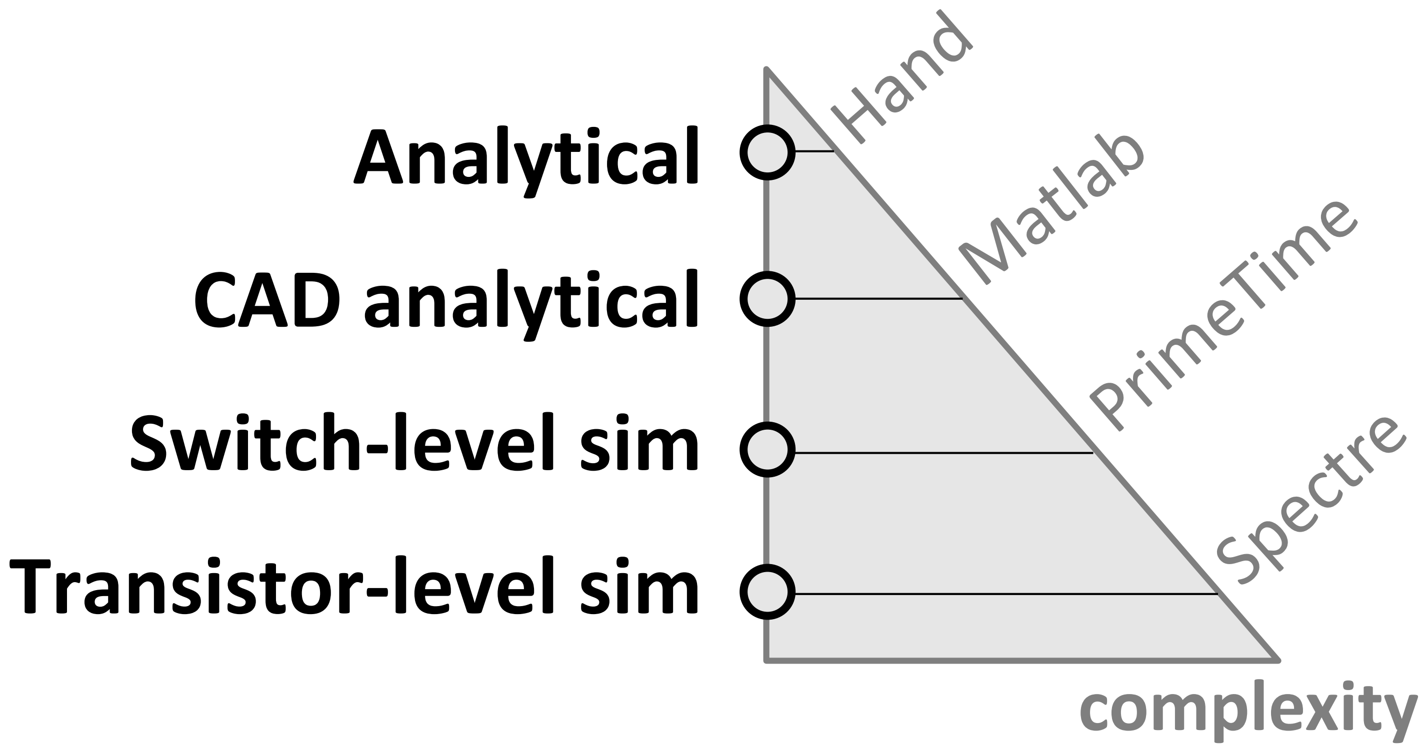
ECE M216A

MOS and Delay Models

Prof. Dejan Marković

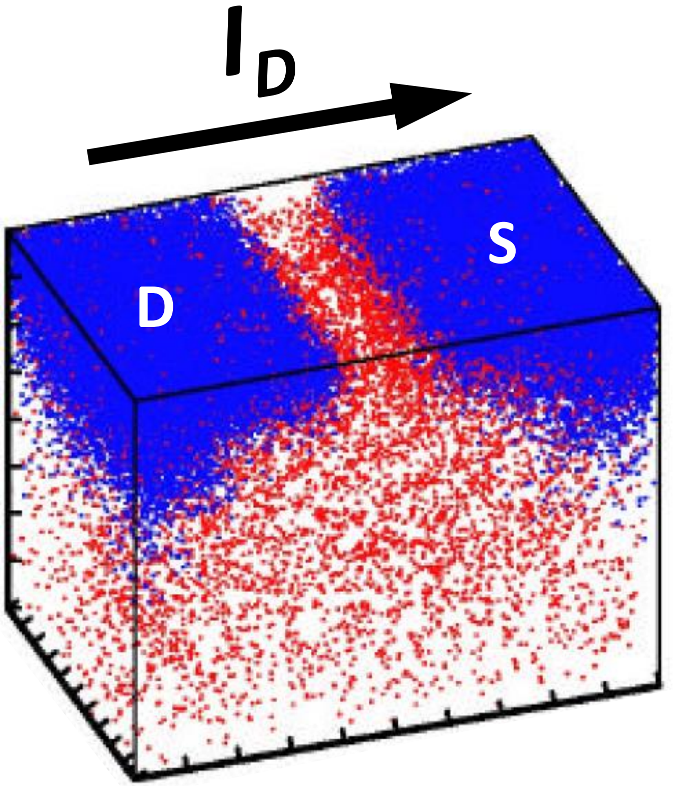
ee216a@gmail.com

Levels of Modeling



Different complexity, accuracy, speed of convergence…

MOS Transistor Modeling



Our goal is to model delay and energy

not current

But have to start with current

MOSFET, Notations

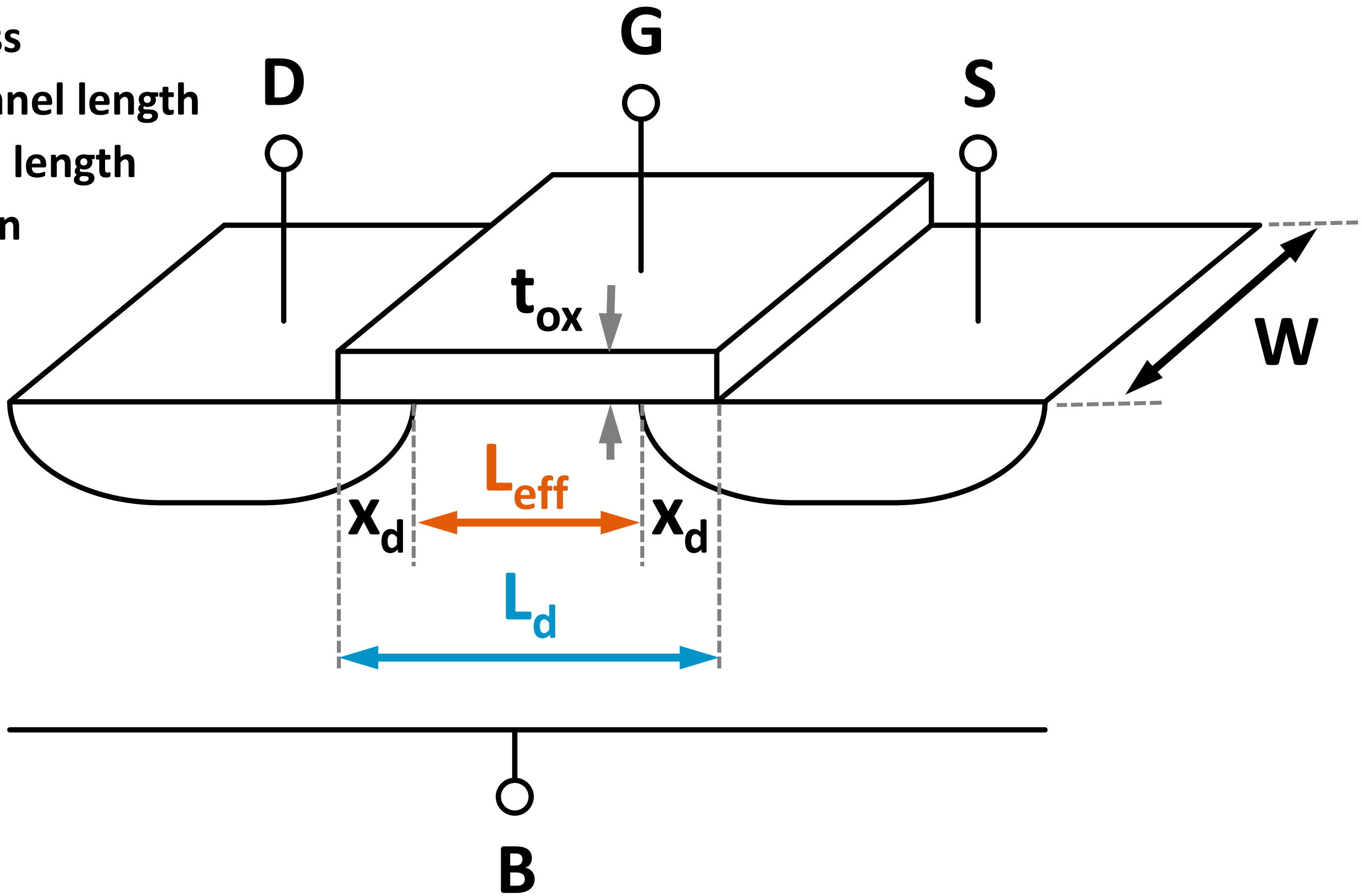
W: transistor width

t_{ox}: oxide thickness

L_{eff}: effective channel length

L_d: drawn channel length

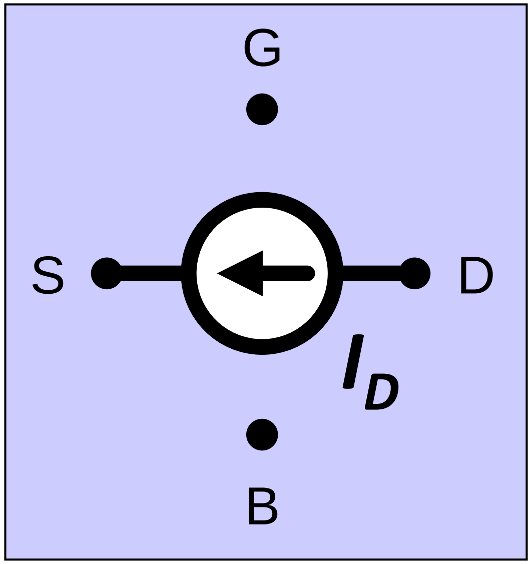
x_d: lateral diffusion



Hand-analysis I-V formulas: $\boxed{L = L_{eff}}$

MOS I-V Model

$$V_{GT} = V_{GS} - V_T$$



Subthreshold region ($V_{GT} \leq 0$)

$$I_D = I_0 \cdot \frac{W}{W_0} \cdot 10^{\frac{V_{GS} - V_T + \gamma_D \cdot V_{DS}}{S}}$$

Active region ($V_{GT} \geq 0$) Lin, Sat, V-Sat

$$I_D = k' \cdot \frac{W}{L} \cdot \left(V_{GT} \cdot V_{min} - \frac{V_{min}^2}{2}\right) \cdot (1 + \lambda \cdot V_{DS})$$

$$V_{min} = min(V_{DS}, V_{GT}, V_{DSAT})$$

Lin **Sat** **V-Sat**

Model Parameters: Active Region

V_{T0} : Threshold voltage

γ : Body effect

V_{DSAT} : Velocity saturation

k' : Transconductance ($k' = \mu \cdot C_{ox}$)

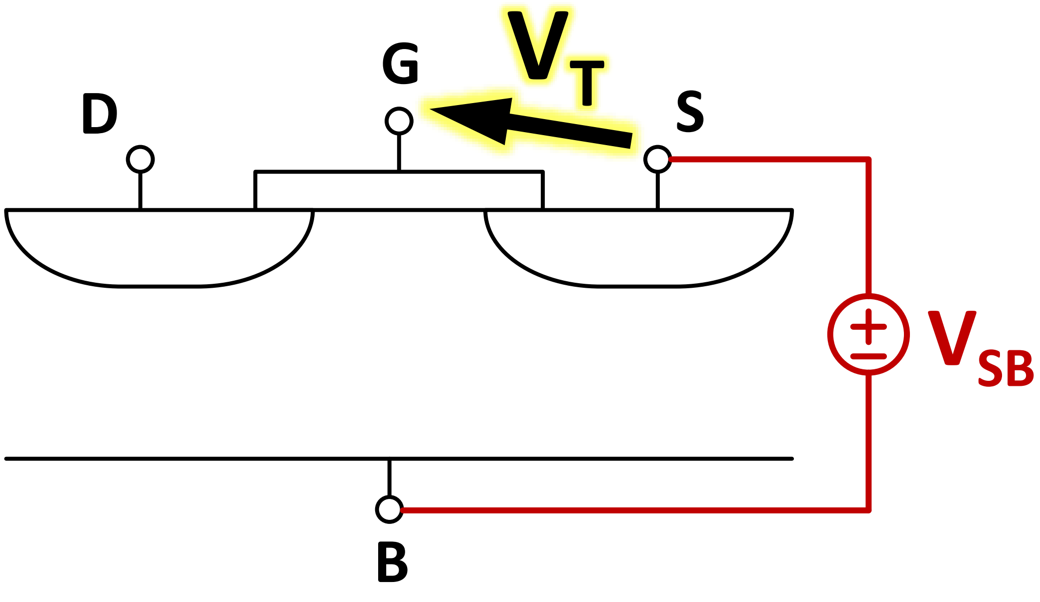
λ : Channel-length modulation (CLM)

- **CLM term $(1 + \lambda V_{DS})$ also included for linear region**
  - Empirical, no physical justification

Threshold Voltage, V_T

$V_{SB} = 0$

$$V_T = V_{T0} + \gamma \cdot (\sqrt{2\Phi_F + V_{SB}} - \sqrt{2\Phi_F})$$



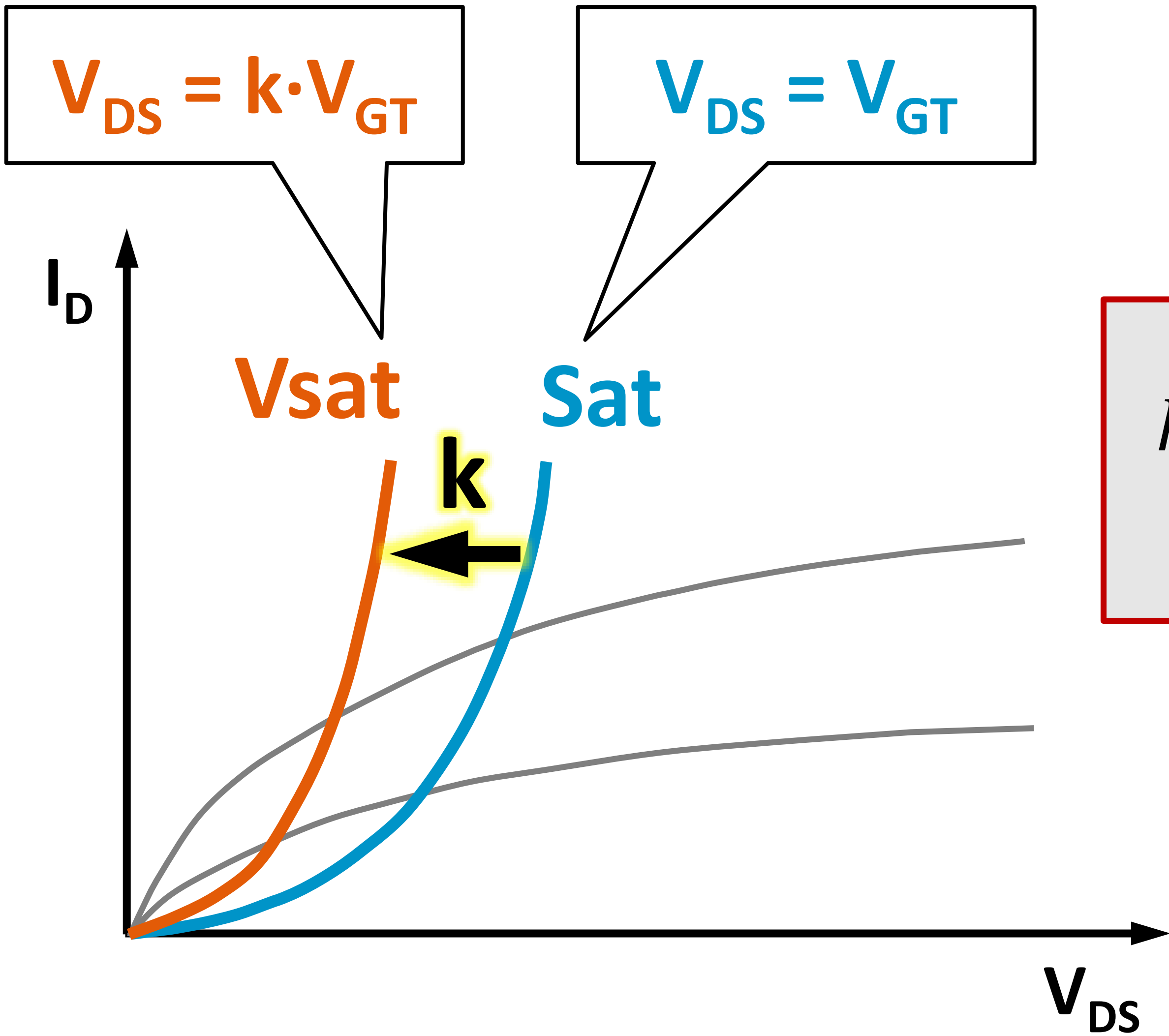
NMOS:

- $V_{SB} > 0$ (RBB)
- $V_{SB} < 0$ (FBB)

PMOS:

- $V_{SB} > 0$ (FBB)
- $V_{SB} < 0$ (RBB)

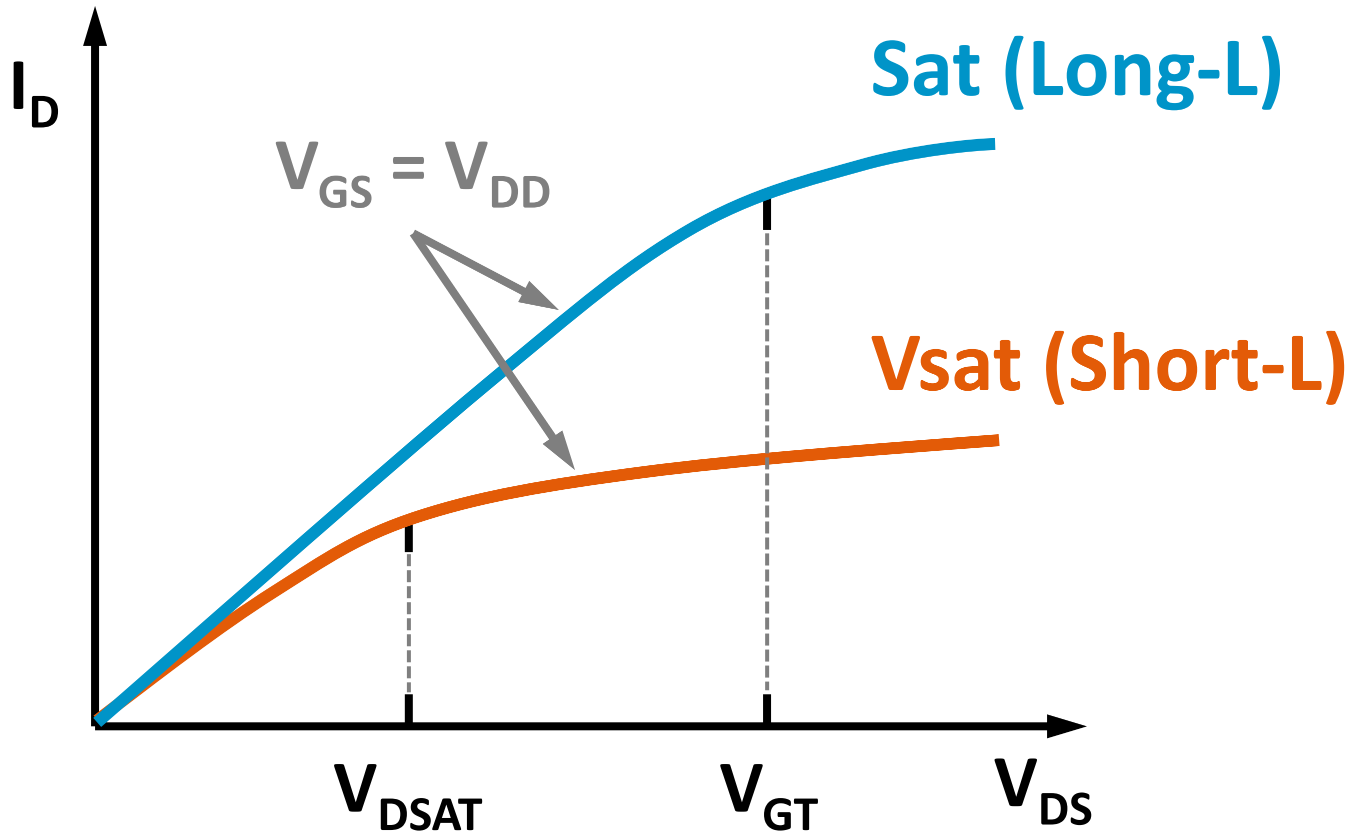
Vsat Occurs at LOWER V_{DS} than Sat



$$k = \frac{1}{1 + \frac{V_{GT}}{\epsilon_c \cdot L}}$$

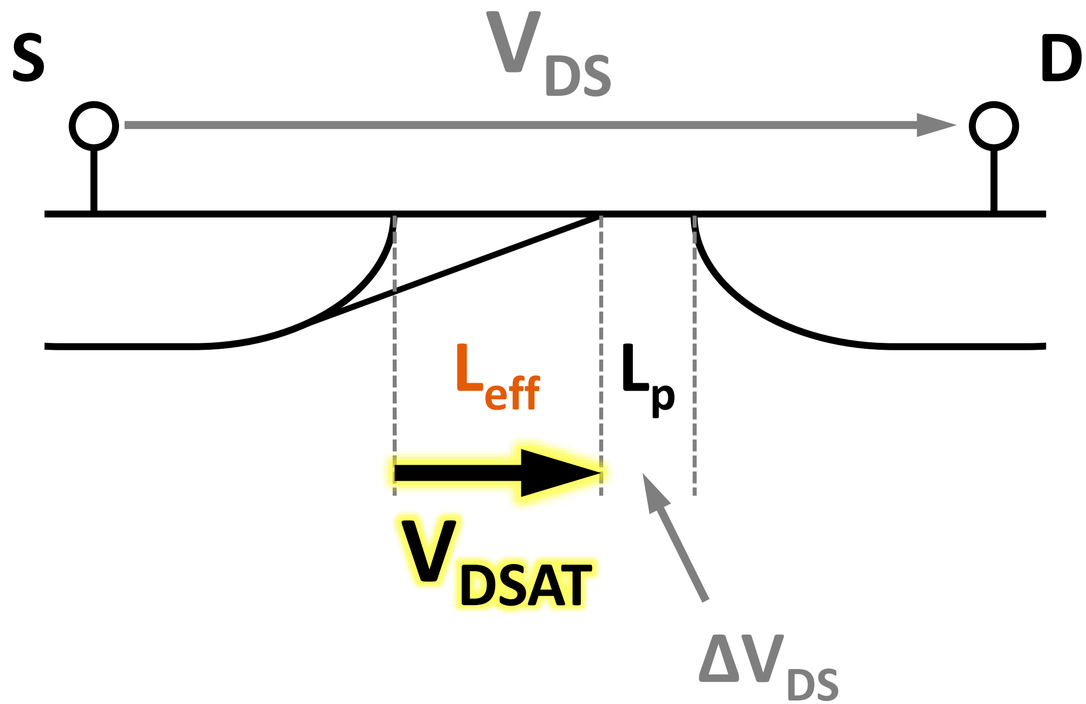
$k = k(V_{GT})$

Vsat: Less Current for Same V_{GS}



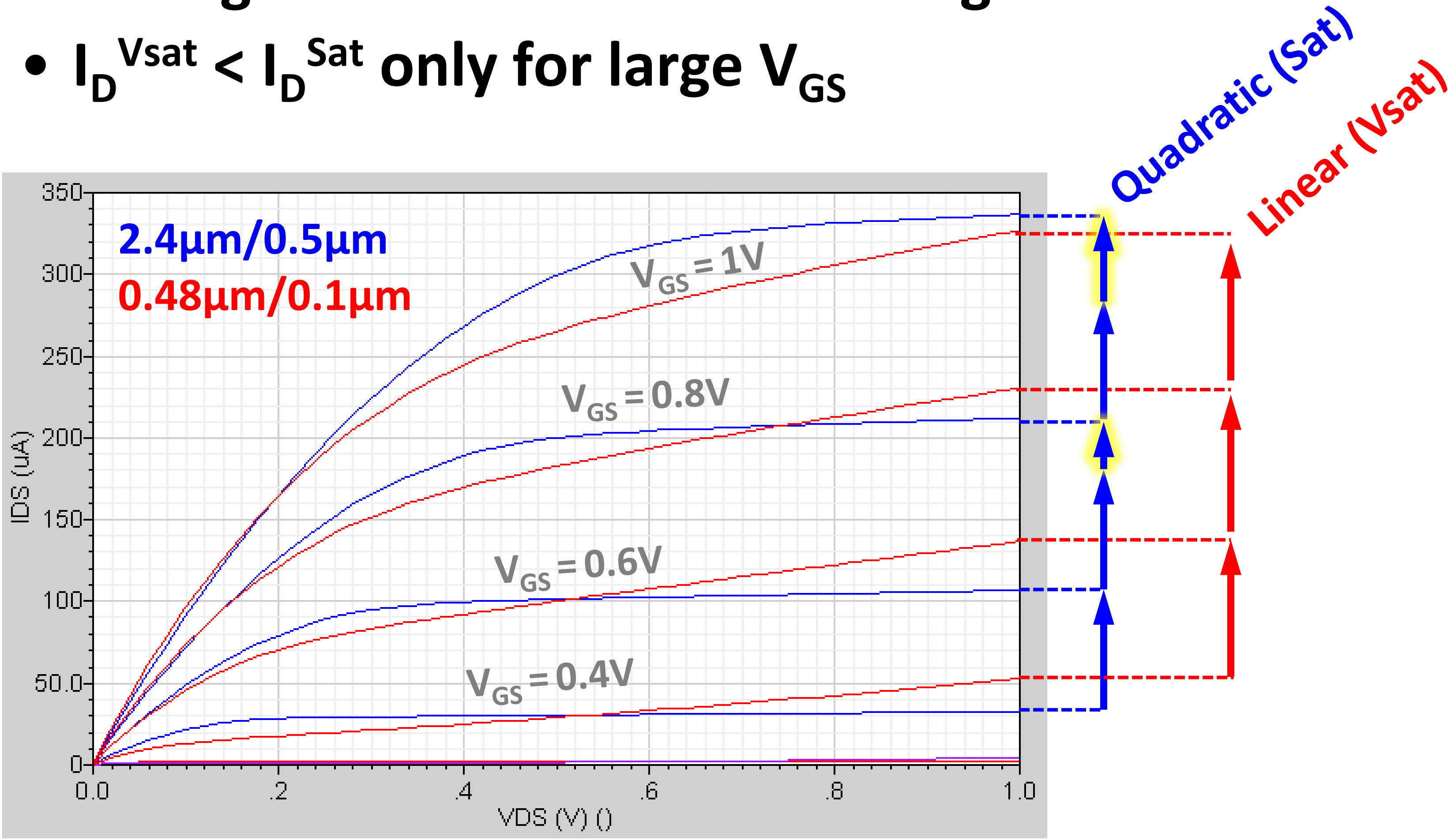
CLM Holds in Vsat

$V_{DS} > V_{DSAT}$

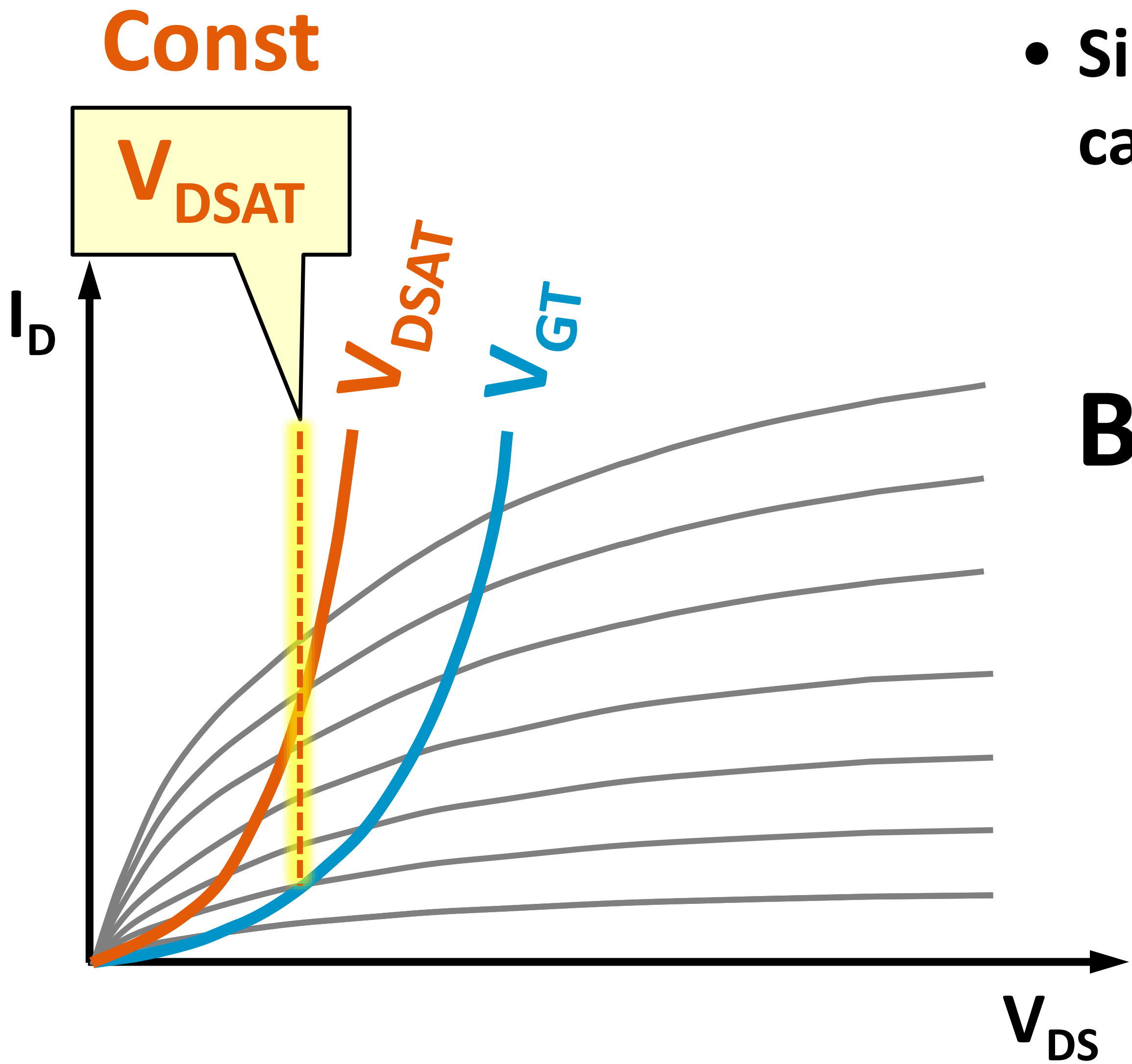


Simulation: Long vs. Short Channel (90nm)

- $I_D^{Sat}(V_{GS})$ quadratic, $I_D^{VSat}(V_{GS})$ linear
- Stronger CLM in short-L than long-L
- $I_D^{Vsat} < I_D^{Sat}$ only for large V_{GS}



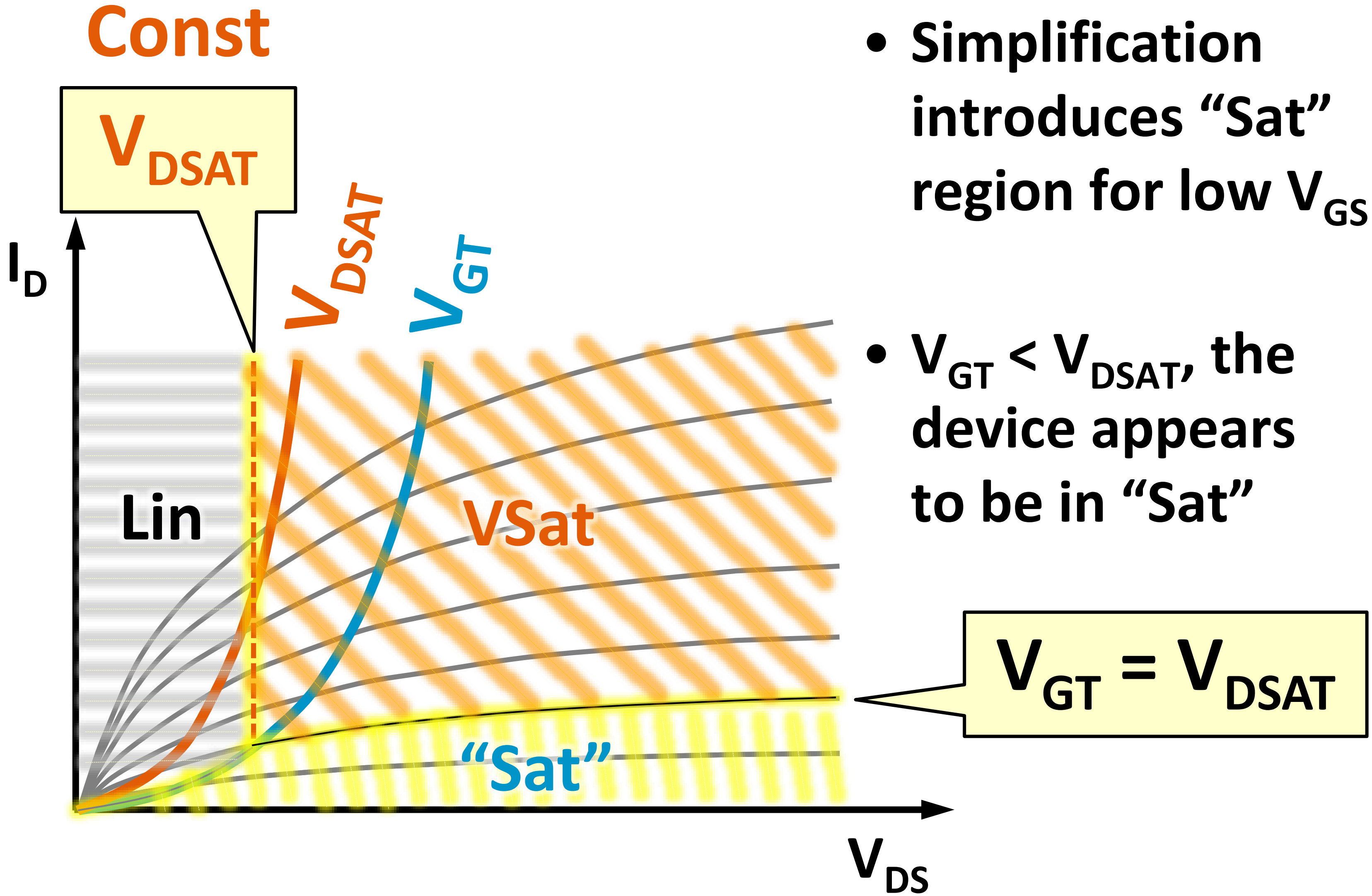
Simplification: V_{DSAT} = Constant



- Simplifies hand calculations

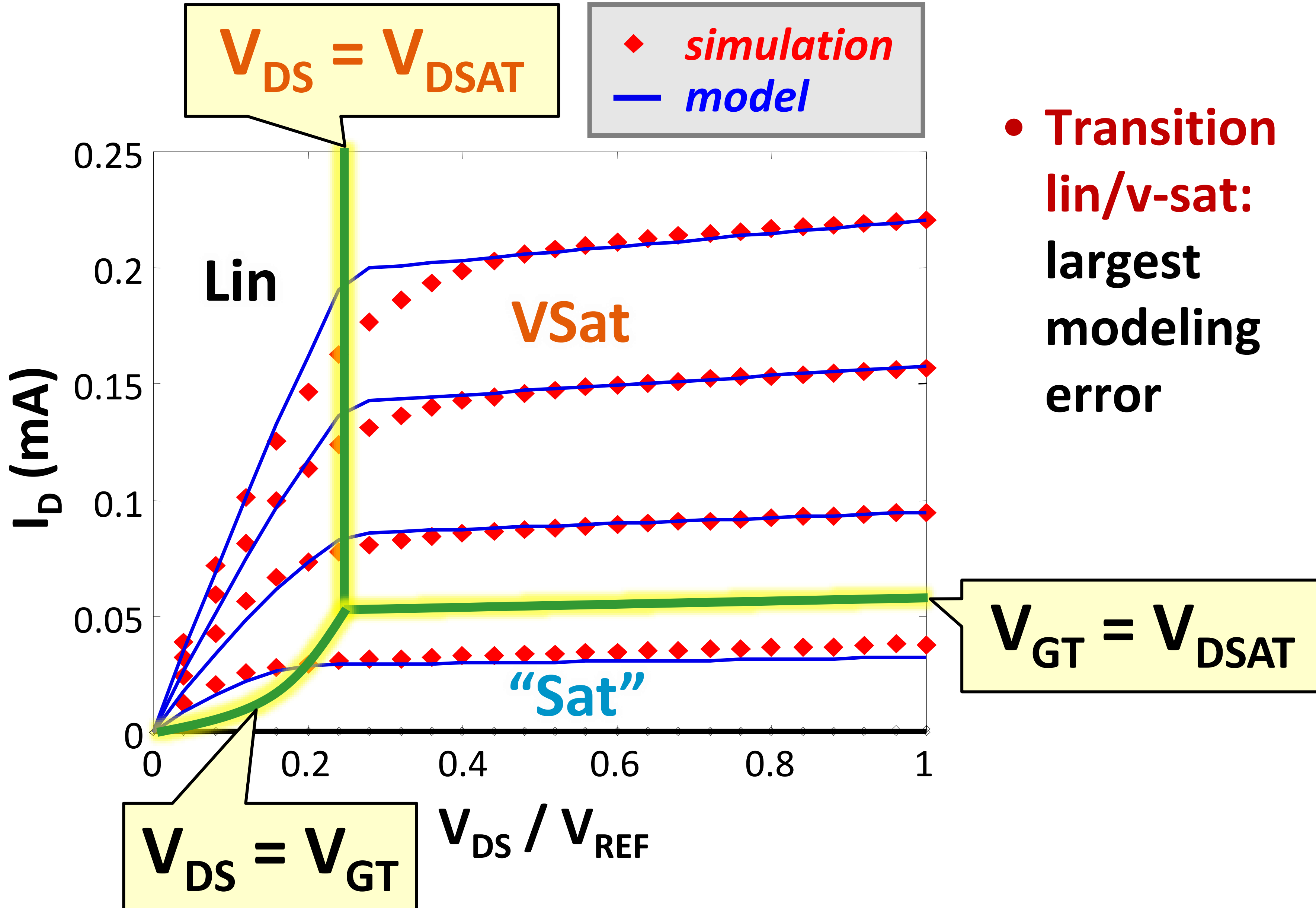
BUT…

Regions of Operation



- Simplification introduces "Sat" region for low V_{GS}
- $V_{GT} < V_{DSAT}$, the device appears to be in "Sat"

Unified Model vs. SPICE Simulation



- **Transition lin/v-sat: largest modeling error**

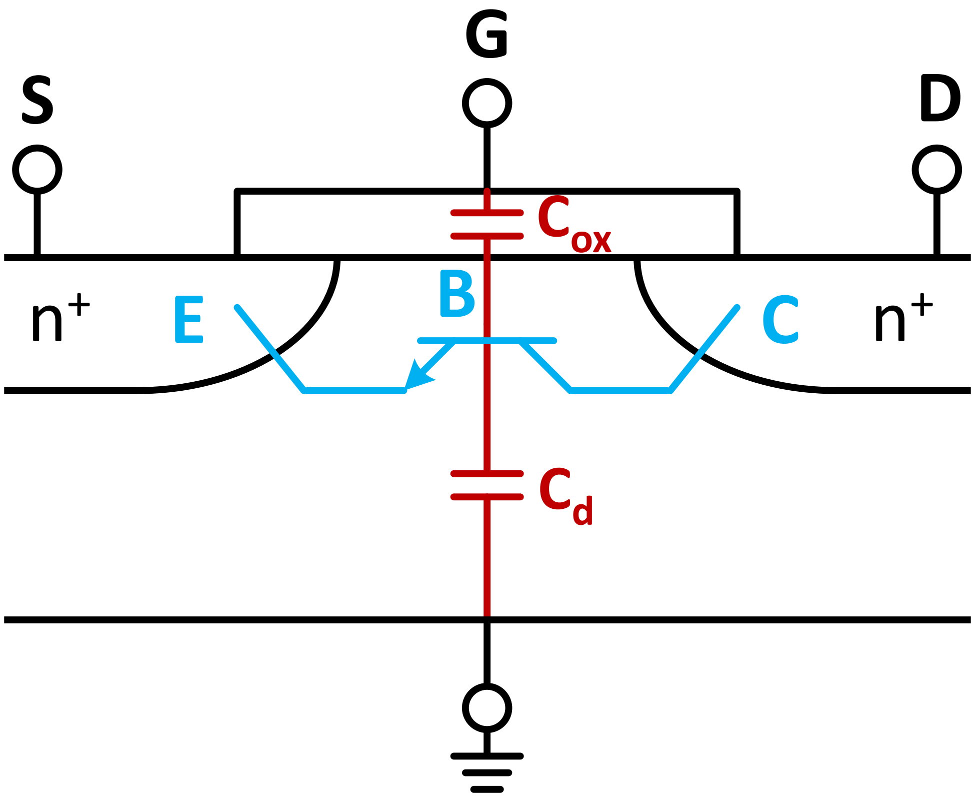
Model Parameters: Subthreshold

I_0 : Nominal leakage current

S : Subthreshold slope

γ_D : DIBL factor

Modeling the Sub-threshold Behavior



Parasitic BJT

$$I_C = I_0 e^{\frac{V_{BE}}{\Phi_t}}$$

$$\boxed{V_{BE} = \frac{V_{GS}}{1 + \frac{C_d}{C_{ox}}}}$$

$$n = 1 + \frac{C_d}{C_{ox}}$$

$$\Phi_t = \frac{kT}{q}$$

$$\boxed{I_D = I_0 \cdot e^{\frac{V_{GS}}{n\Phi_t}} \cdot \left(1 - e^{-\frac{V_{DS}}{\Phi_t}}\right)}$$

Sub-threshold I_D vs. V_{GS}

Physical model

$$\boldsymbol{I_D = I_0 \cdot e^{\frac{V_{GS}}{n\Phi_t}} \cdot \left(1 - e^{-\frac{V_{DS}}{\Phi_t}}\right)}$$

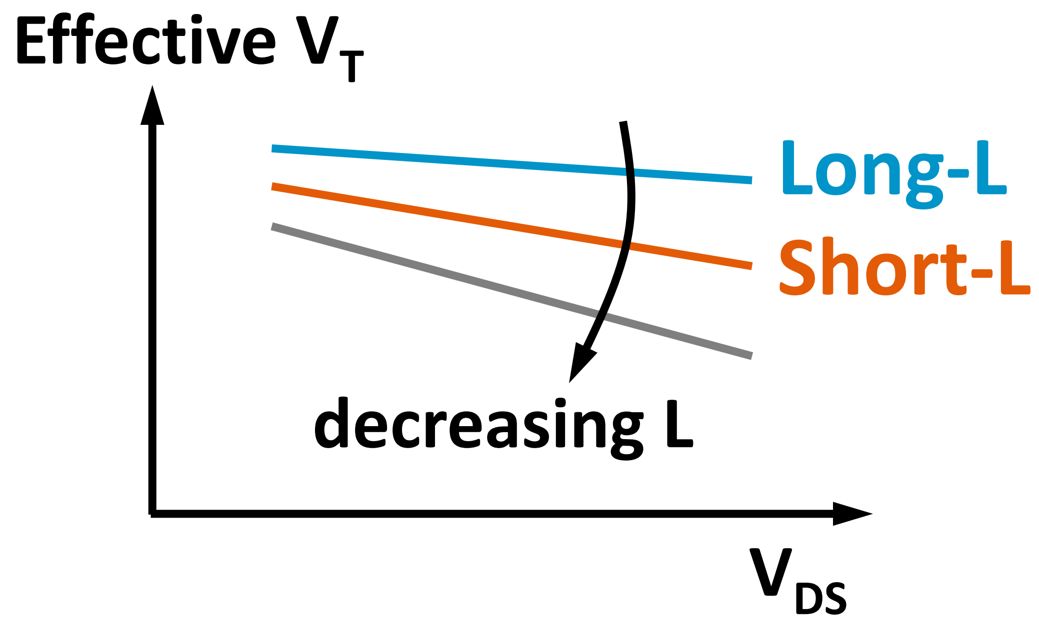
$$I_0 = \mu \frac{W}{L} {\Phi_t}^2 e^{-\frac{V_T}{n\Phi_t}}$$

Empirical model

$$\boldsymbol{I_D = I_0 \cdot \frac{W}{W_0} \cdot 10^{\frac{V_{GS} - V_T + \overbrace{\gamma_D V_{DS}}^{\text{DIBL}}}{S}}}$$

$$\boldsymbol{S = n\Phi_t \ln(10)} \quad \textbf{[mV/dec]}$$

Drain Induced Barrier Lowering (DIBL)



- **Field lines from the drain affect charge in the channel**
- **Typically derived for small V_{DS}, holds for large V_{DS}**
  - Even if we neglect CLM, I_{DS} will increase b/c of V_T drop
  - Device turned off by V_{GS} (below V_T) may turn on by V_{DS}

The Sub-threshold Slope Parameter

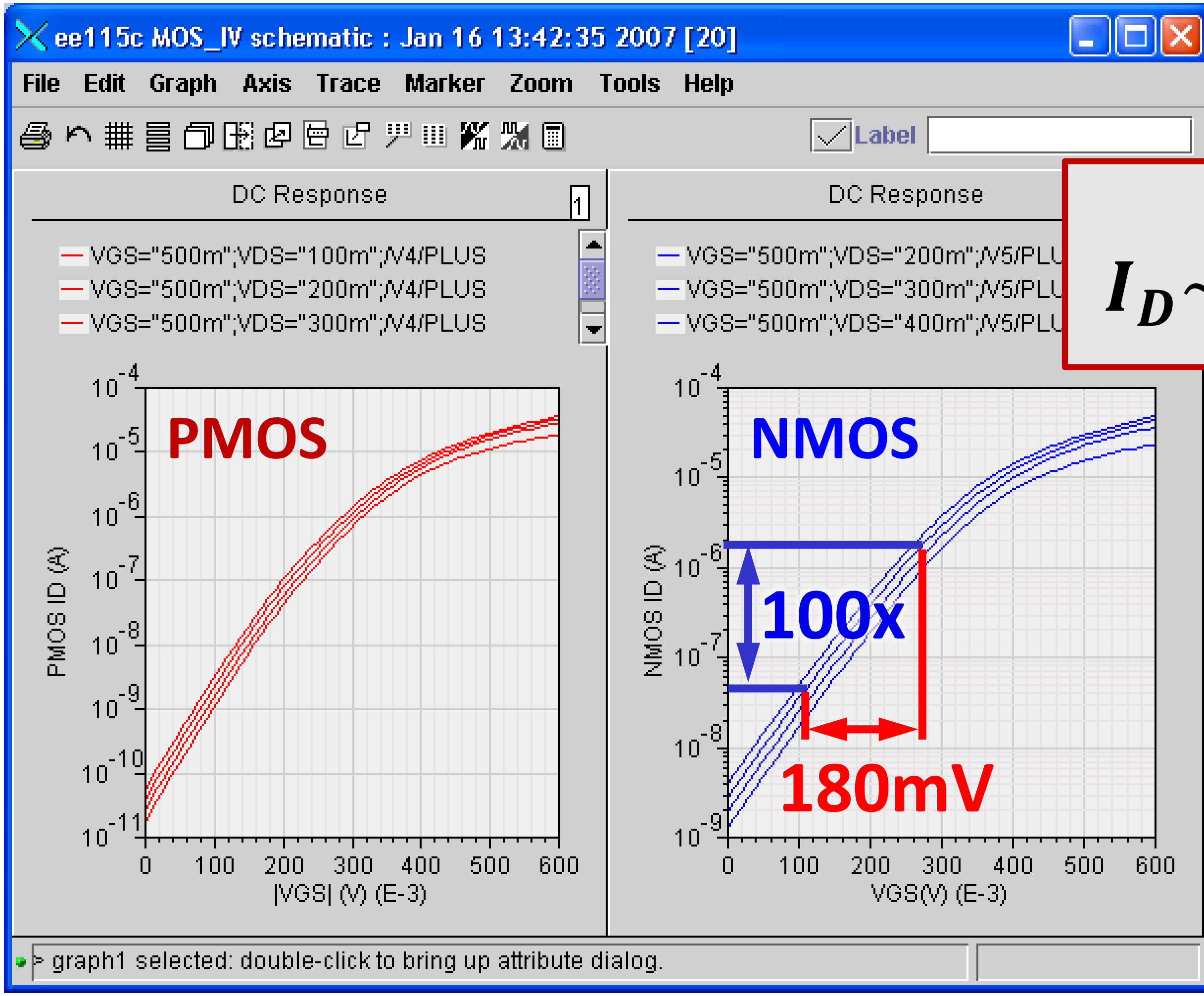
$$\boxed{S = n\Phi_t \ln(10)} \quad [\text{mV/dec}]$$

Change in V_{GS} that gives 10x change in I_{DS}

- **n = 1 60 mV/dec (ideal)**
- **n = 1.5 90 mV/dec (typical)**

- **S:** increases with temperature (Φ_t)
- **n:** intrinsic to device topology / structure

90nm Simulation: Sub-threshold I_D vs. V_{GS}



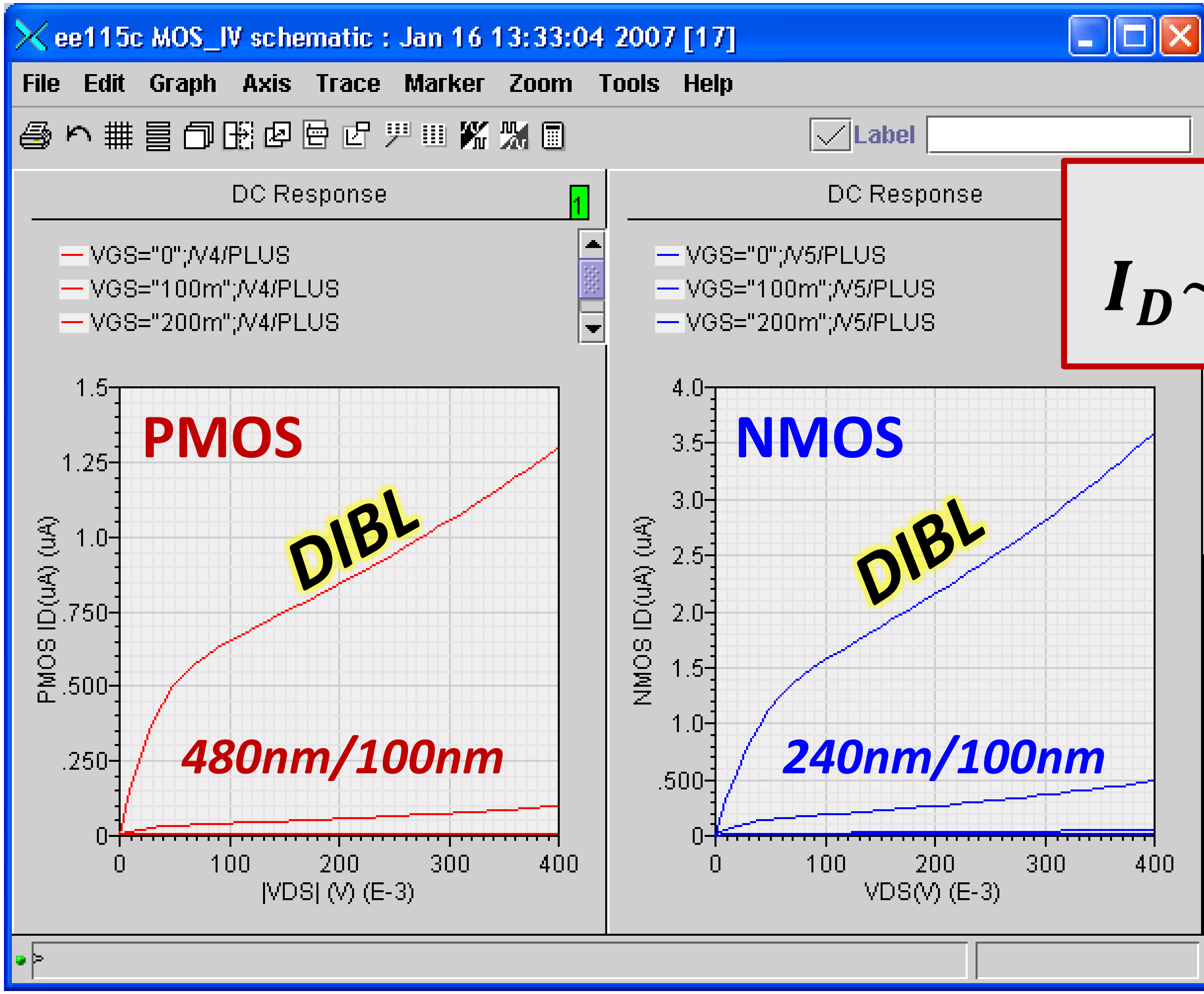
$$I_D \sim 10^{\frac{V_{GS}-V_T+\gamma_D V_{DS}}{S}}$$

V_{DS} : 0 to 0.4V

$$S = n\Phi_t \ln(10)$$

90mV/dec

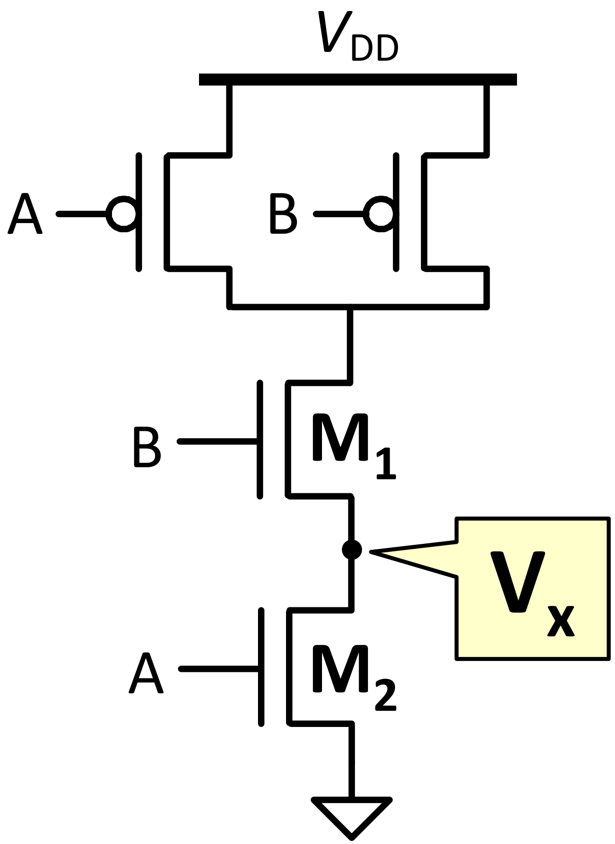
90nm Simulation: Sub-threshold I_D vs. V_{DS}



$$I_D \sim 10^{\frac{V_{GS}-V_T+\gamma_D V_{DS}}{S}}$$

V_{GS} : 0 to 0.3V

Transistor Stacks Reduce Leakage

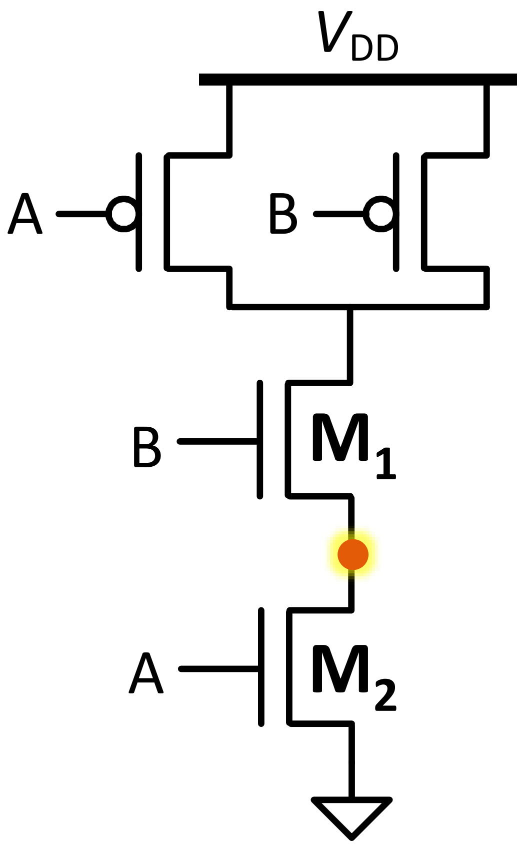


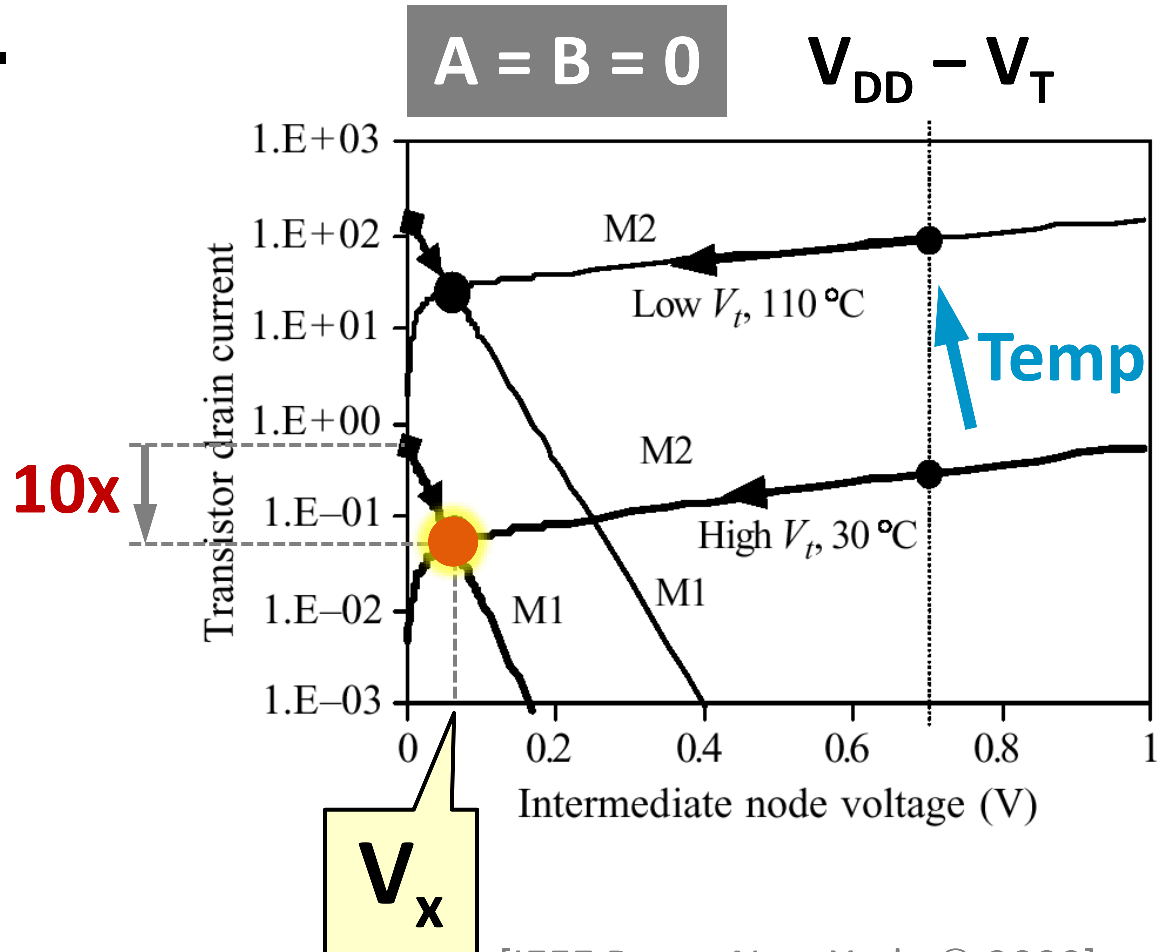
A = B = 0

V_x @ $I_{D1} = I_{D2}$?

- **$V_{T1} > V_{T10}$ (RBB)**
  - $I_{D1} \propto 10^{-\frac{V_{T1}}{S}}$
- **Large ΔV_{DS1} required**
  - V_x very small

~10x Lower Leakage for a Stack of 2





[IEEE Press, New York, © 2000]

Practically Stack 2 or 3 Transistors

Leakage Power Reduction

	High V_t	Low V_t
2 NMOS	10.7X	9.96X
3 NMOS	21.1X	18.8X
4 NMOS	31.5X	26.7X
2 PMOS	8.6X	7.9X
3 PMOS	16.1X	13.7X
4 PMOS	23.1X	18.7X

[IEEE Press, New York, © 2000]

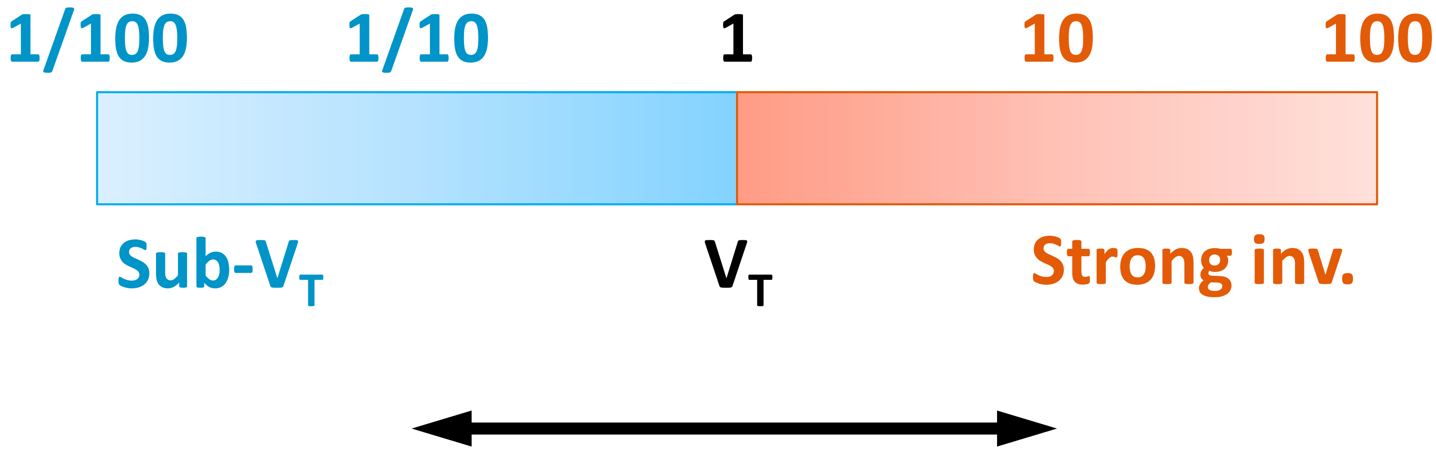
Near-V_T Region

($V_T + \Delta V$ Region)

Definition: Inversion Coefficient (IC)

Inversion coefficient indicates proximity to V_T

IC = 1 (@ V_T), IC < 1 (sub-V_T), IC > 1 (above-V_T)



a.k.a. $V_T + \Delta V$ region

Current Model

Start from I_S

$$I_S = 2n \cdot \mu \cdot C_{ox} \cdot \frac{W}{L} \cdot {\Phi_t}^2$$

Sub-V_T ($V_{GS} = 0$)

$$I_D = I_S \cdot e^{\frac{\sigma V_{DD} - V_T}{n\Phi_t}}$$

Above-V_T

$$I_D = I_S \cdot IC / k_{fit}$$

$$IC = \left(ln\left(e^{\frac{(1+\sigma)V_{DD} - V_T}{2n\Phi_t}} + 1\right)\right)^2$$

Calculate V_{DD} from IC

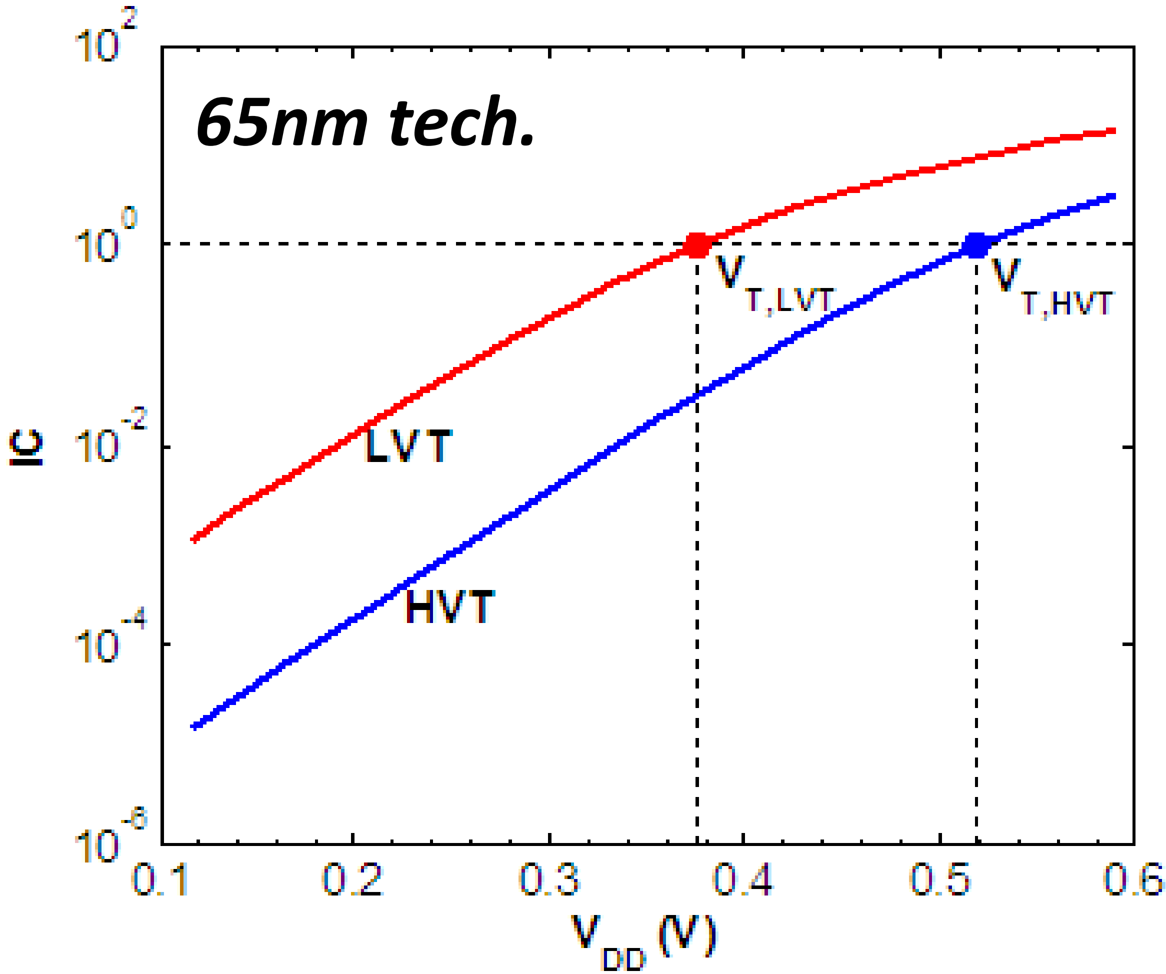
Useful for optimizations

$$\boldsymbol{V_{DD} = \frac{V_T + 2n\Phi_t ln\left(e^{\sqrt{IC}} - 1\right)}{1 + \sigma}}$$

Given IC, find V_{DD} for LVT and HVT?

Fitting the IC Parameter

Constrain MMSE-based curve fit with IC = 1 @ V_T



Toward Delay Model:

Alpha-Power-Law Model

Alpha-Power Model of the Drain Current

Basis for delay calculation, useful for hand analysis

$$\boxed{I_D = \frac{1}{2}\mu \cdot C_{ox} \cdot \frac{W}{L} \cdot (V_{GS} - V_T)^{\alpha}}$$

Neglects CLM

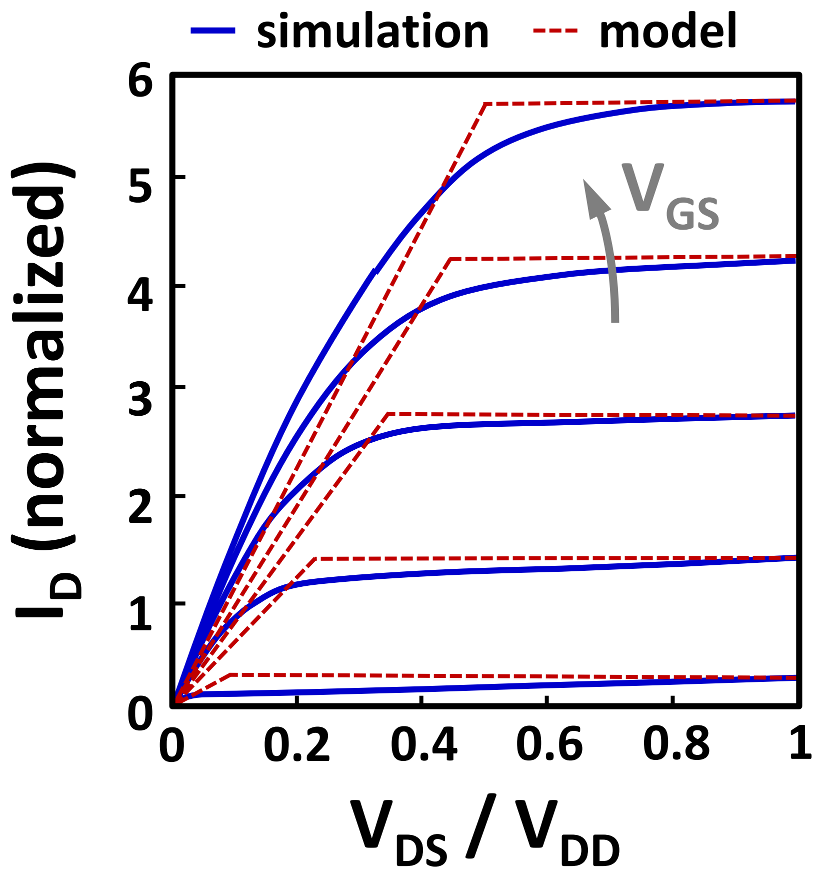
Empirical model

α: vel. sat index

$1 < \alpha < 2$

T. Sakurai and R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584-594, Apr. 1990.

α-Power Model: Curve Fitting (MMSE)



How to fit the model?

- $1 < \alpha < 2$
  - Degree of v-sat
- α depends on V_T
  - Many combinations
  - Use V_{T0} (your tech.)

Simulation Models

Physical + empirical parameters (100+ parameters)

- **Spectre 45nm Cadence GPDK**

  /w/apps/public.2/tech/cadence
  /45nm/gpdk045_v_3_5/models
  /spectre/**gpdk045_mos.scs**

- **HSPICE 32-28nm Synopsys EDK**

  /w/apps/public.2/tech/synopsys
  /32-28nm/SAED32_EDK/tech
  /hspice/**saed32nm.lib**

MOSFET Behavior: Summary

- **MOSFET: a 4-terminal device**
  - Body impacts performance (V_T)

- **The current in (V)Saturation depends on V_{DS}**
  - **CLM:** L_{eff} is a function of V_{DS}
  - **DIBL:** High E_{DS} lowers V_T

MOSFET: Modes of Operation

1 • **Velocity saturation**
  - Charge velocity saturates at high E_{DS}

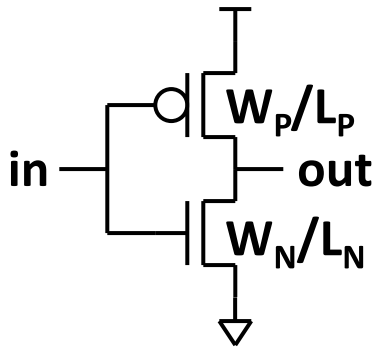
2 • **Subthreshold**
  - Current still flows when $V_{GS} < V_T$

3 • **Linear**
  - Not interesting in digital design

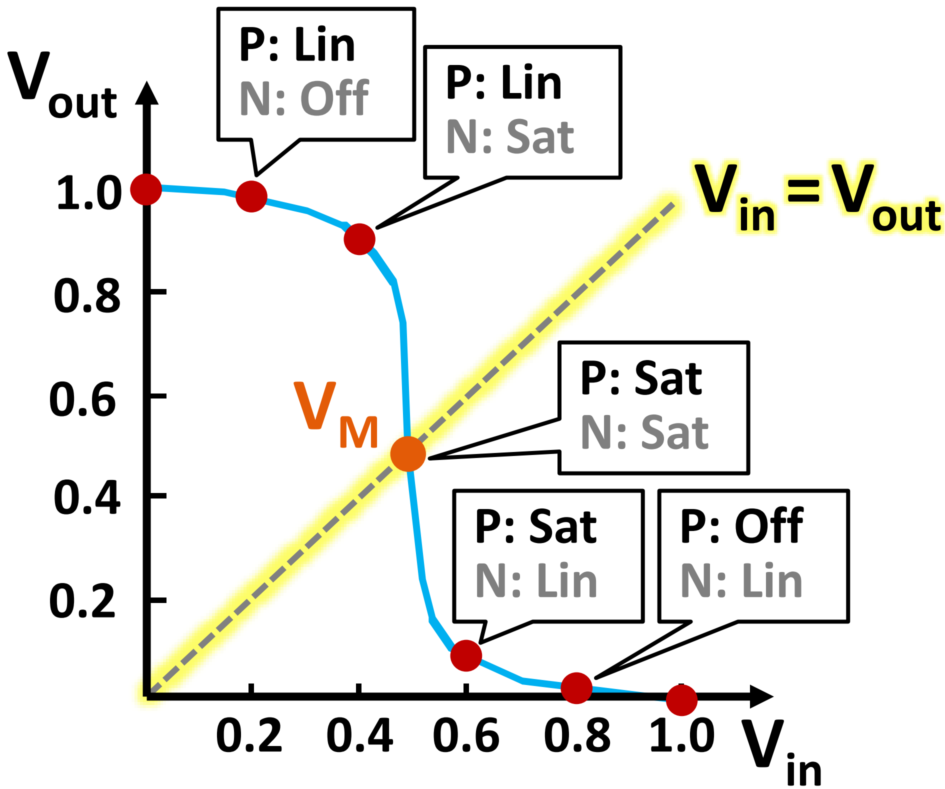
4 • **Weak (near-V_T) inversion**
  - Crucial for ultra-low-power design

Modeling
Gate Delay

Review: CMOS Inverter VTC

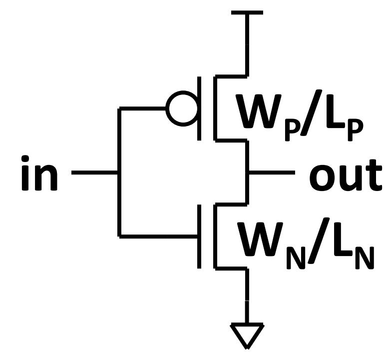


- **Inverter DC response**
- **5 regions of operation**
- **Logical threshold**
  - $V_{in} = V_{out}$



Logical Threshold Voltage

- **Set $I_{DP} = I_{DN}$ and solve**
  - Dependence on P:N sizing and mobility ratio
  - Slight dependence on $V_{TP/N}$

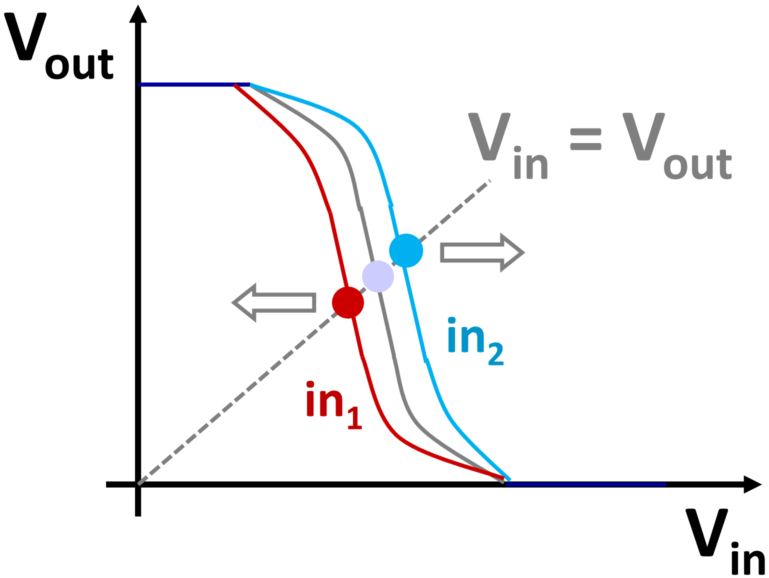


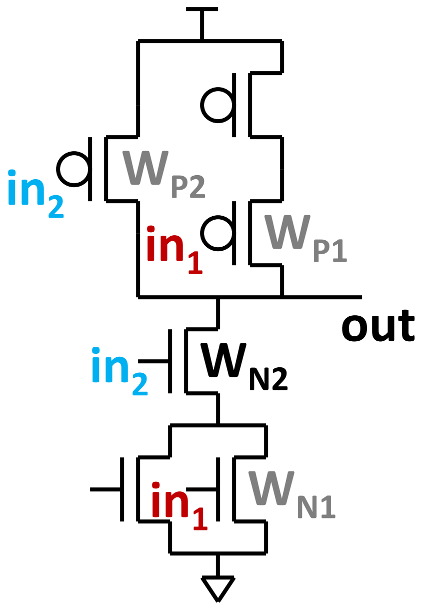
$$V_M = \frac{\left(V_{TN} + \frac{V_{DSATN}}{2}\right) + r \cdot \left(V_{DD} + V_{TP} + \frac{V_{DSATP}}{2}\right)}{1 + r}$$

$$r = \frac{k_p \cdot V_{DSATP}}{k_n \cdot V_{DSATN}}$$

Use $V_M = V_{DD}/2$ Unless Severely Skewed

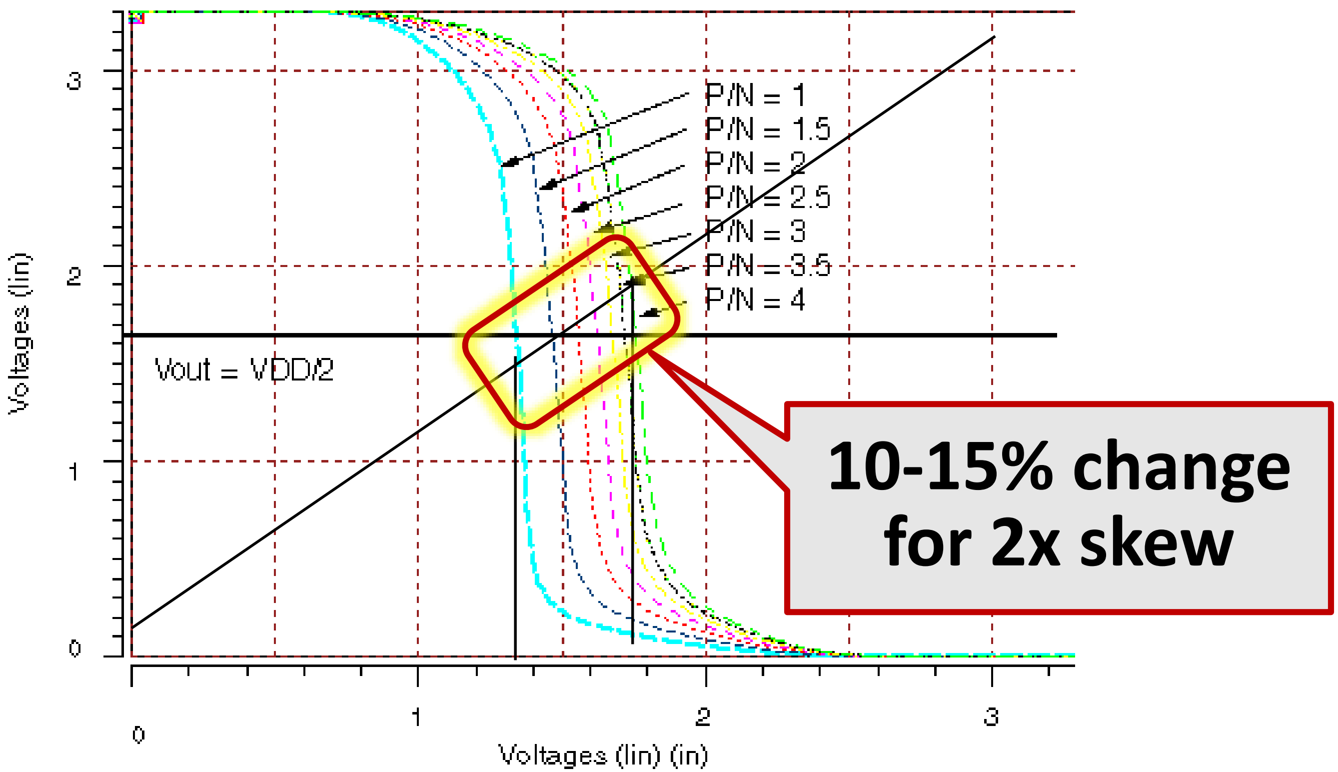
- **Not so easy if not an inverter**
  - Depends on which input the gate is driving
    - In_1 to Out VTC can be different from In_2 to Out
  - Use $V_{DD}/2$ as average case
    - Unless severely skew the P:N ratio



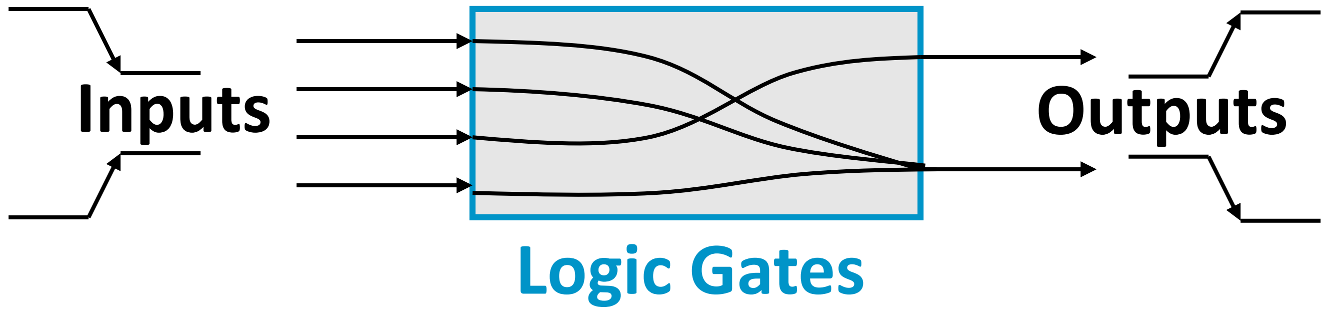


(In)Sensitivity of VTC to P:N Ratio

- **Fortunately, V_M is not very sensitive to P:N ratio (skew)**
  - Ranges from 1.35V to 1.75V (for a 3.3-V V_{DD})
  - $V_M = V_{DD}/2$ is quite reasonable



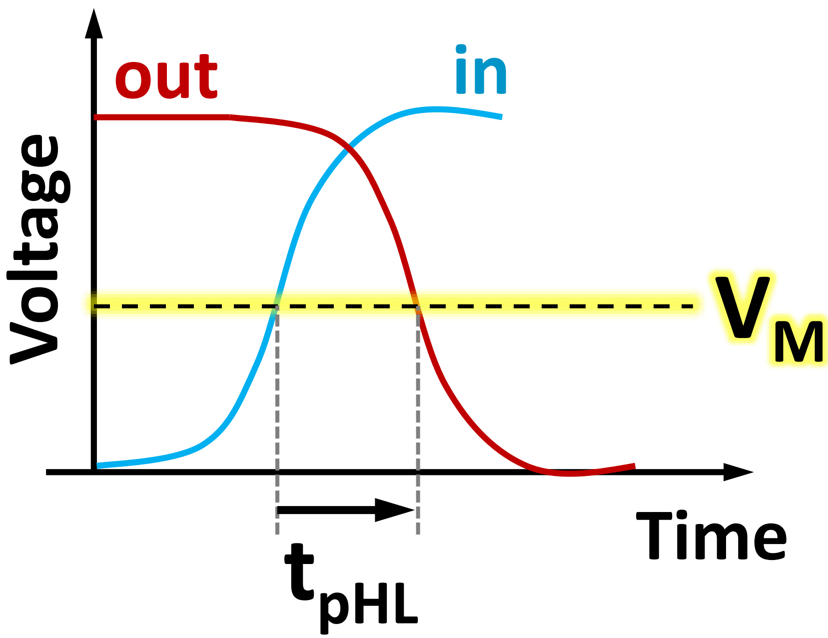
Gate Delay



- **Time b/w an input transition and an output transition**
  - Different delays for different input to output paths
  - Different for an upward or downward transition

Logic Transition

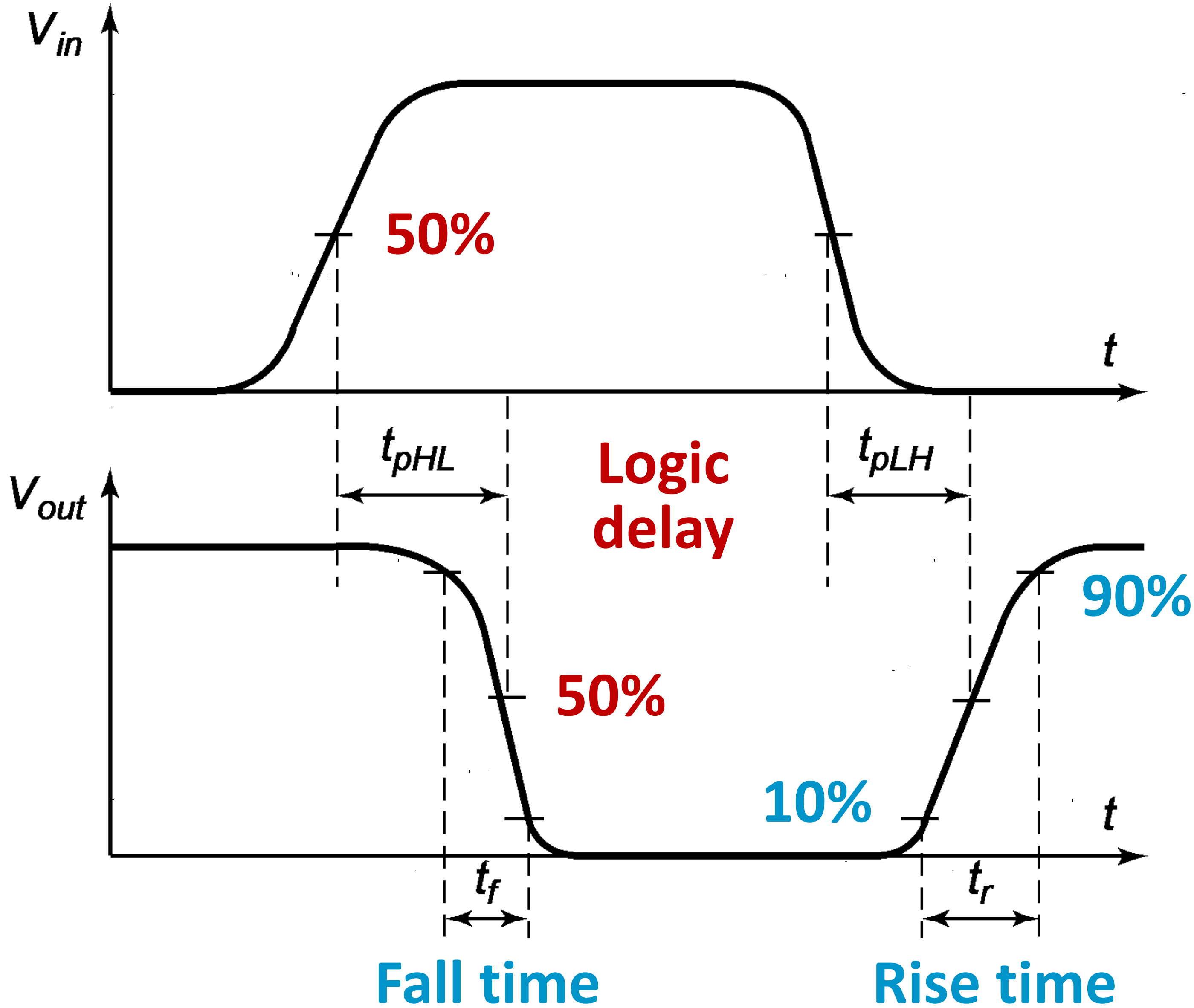
- **Time at which a signal crosses logical threshold voltage**
  - Digital abstraction for 1 and 0
  - Often use $V_{DD}/2$



t_{pHL}

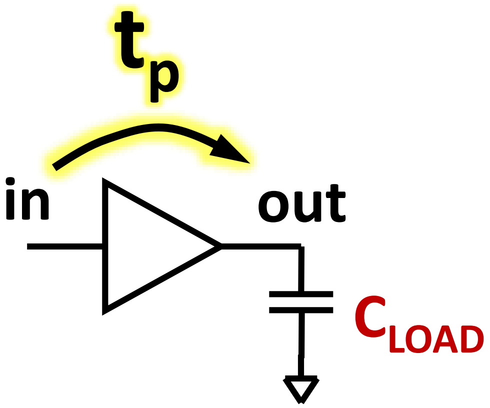
High-to-Low
Output Transition

Delay Definitions



Static CMOS Gate Delay

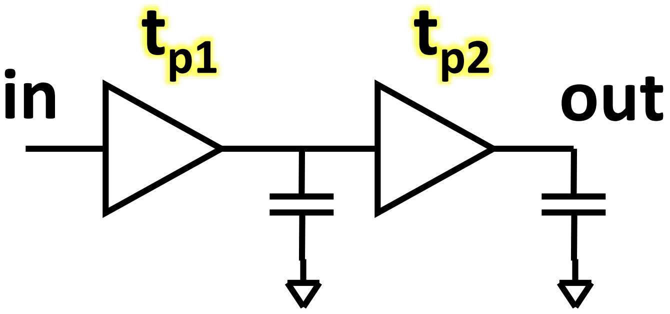
- **Gate output drives the inputs to other gates (+ wires)**
  - Only pull-up or pull-down, not both
  - Capacitive loads (C_{LOAD})



$$t_p = t_{pLH} \text{ or } t_{pHL}$$

Multi-Stage Logic

The delay of **each stage treated separately**



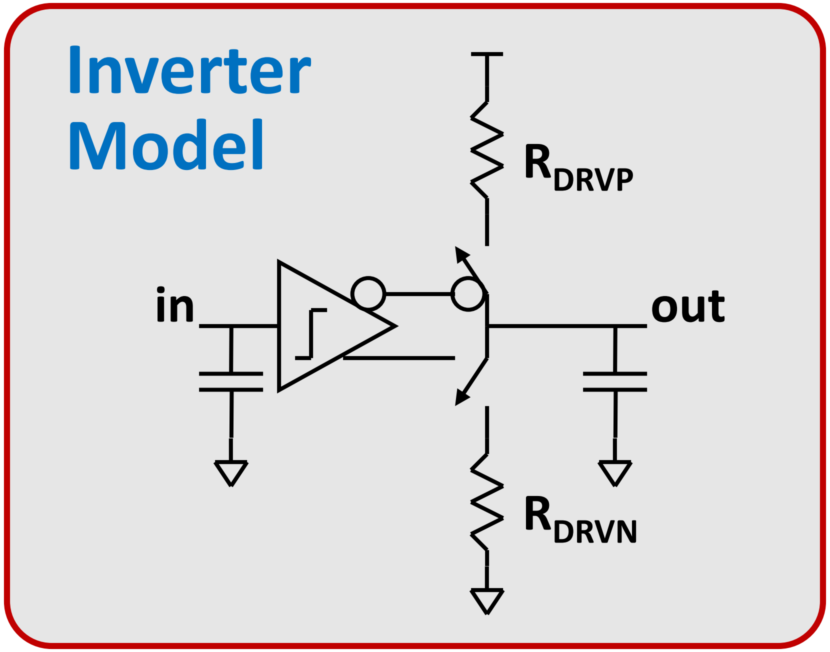
$$\boxed{t_p = t_{p1} + t_{p2}}$$

RC Delay Model

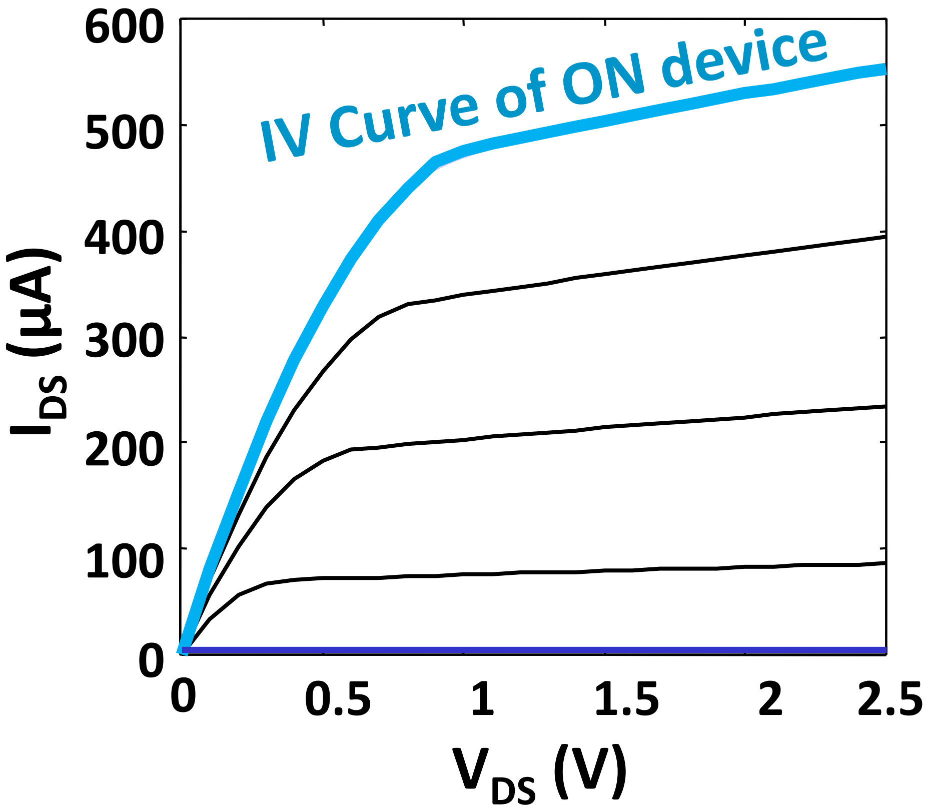
- **R: we can use the resistor model of a transistor**
  - Take into account the different regions of operation
  - Use a realistic slope to model an input switching

- **C: take the average capacitance of a transistor as well**

- **The easy model (one we'll primarily use)**
  - Delay ~ $R_{DRV}C_{LOAD}$
  - R_{DRV} ~ L/W

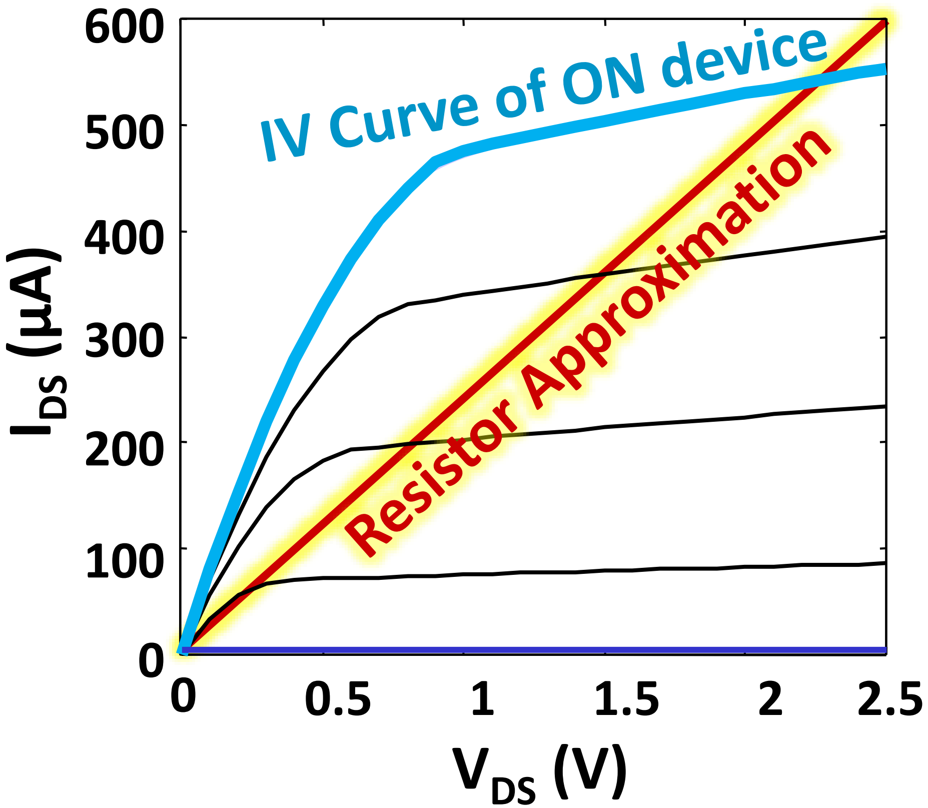


Switched Resistor Model

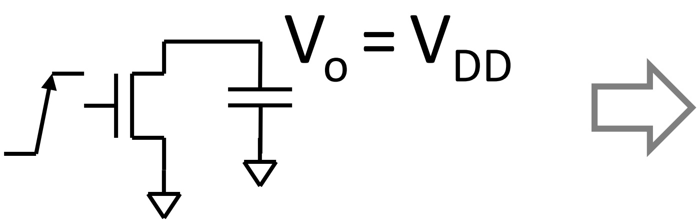


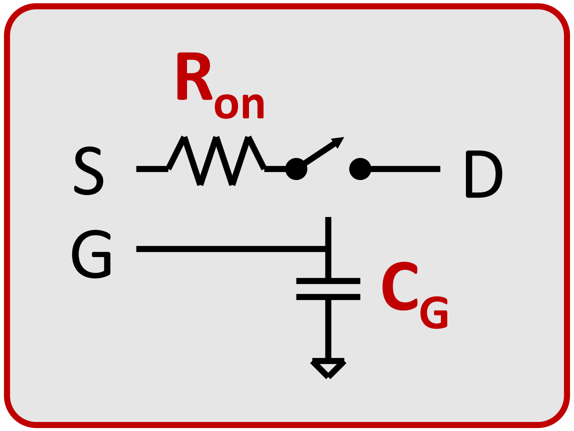
- Switch model insufficient
- Regions of operation matter
- With digital input on gate, device is either ON or OFF

Resistor Approximation

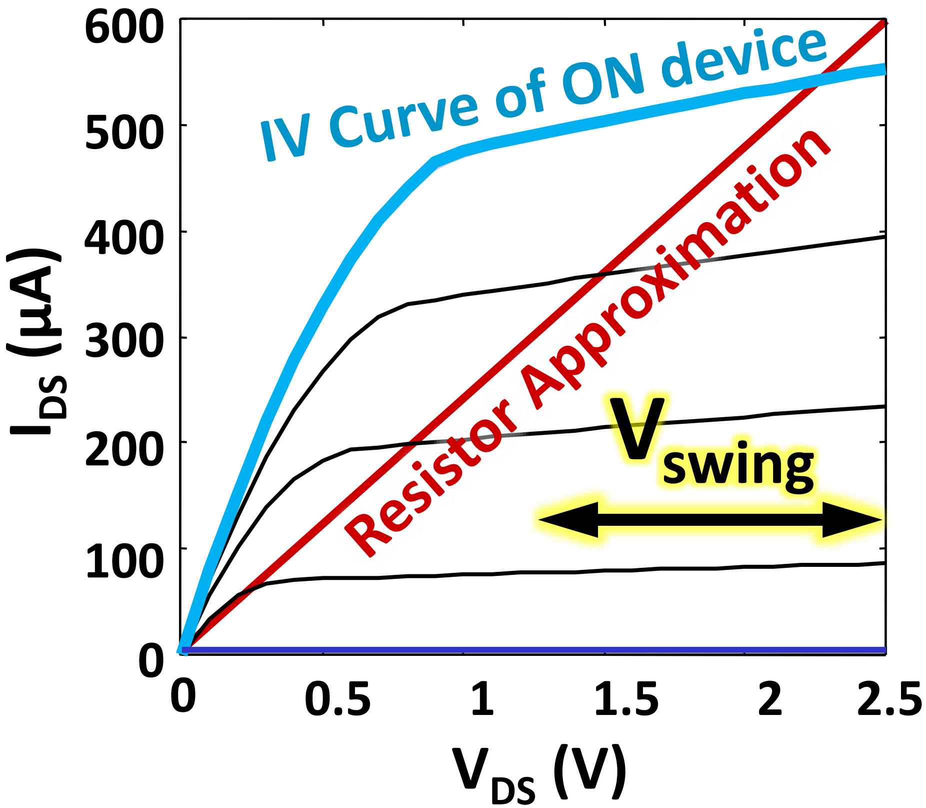


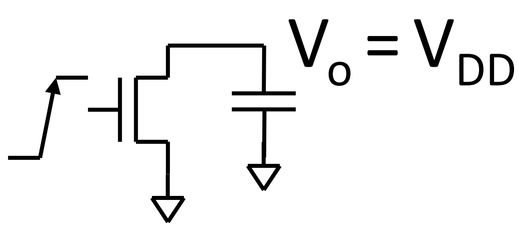
- **Linear R approximation**
- **With digital input on gate, device is either ON or OFF**
  - Approx. **ON** device with **R_{on} (red line)**





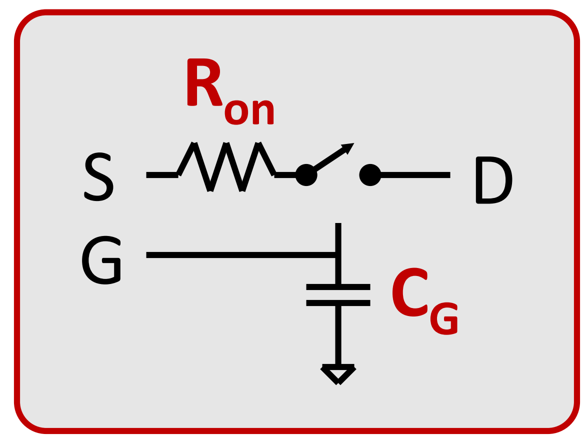
Range of $V_{DS} = V_{swing}$





<u>Assumptions:</u>

- **Saturation region**
- **V_{DS} : $V_{DD} \rightarrow V_M$**
- **$V_{swing} = V_{DD} - V_M$**



Calculating the Resistance

- R_{on} is an "effective" resistance that is averaged

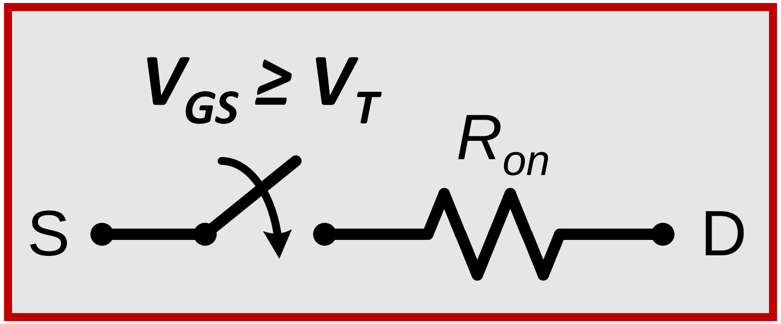
$$R_{on} = R_{avg} = \frac{R\left(V_{DS} = \frac{V_{DD}}{2}\right) + R(V_{DS} = V_{DD})}{2}$$

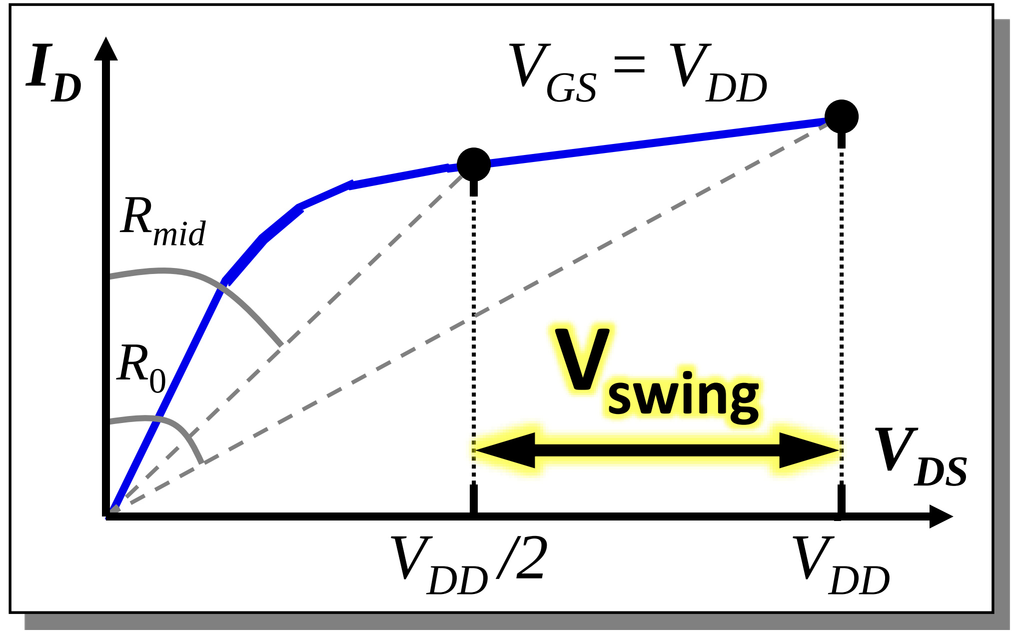
- R is large-signal resistance

$$R(V_{DS} = V_{DS0}) = \frac{V_{DS0}}{I_D(V_{DS0})}$$

- **Input transition dependent**
  - Input is not a perfect step

Calculating (Effective) R_{on}





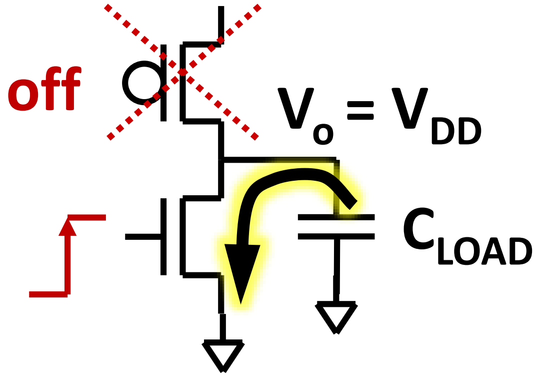
$$R_{on} = \frac{1}{2} \cdot (R_{mid} + R_0)$$

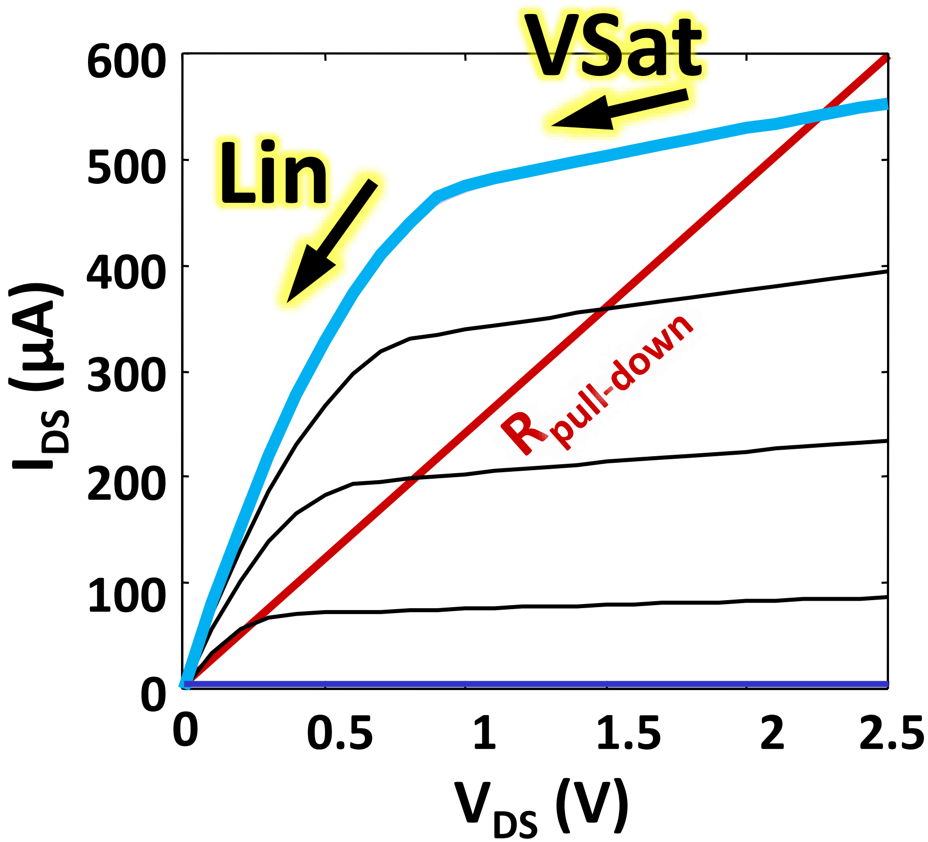
$$R_{on} = \frac{1}{2} \cdot \left(\frac{V_{DD}}{I_{DSAT} \cdot (1 + \lambda \cdot V_{DD})} + \frac{V_{DD}/2}{I_{DSAT} \cdot (1 + \lambda \cdot V_{DD}/2)} \right)$$

$$R_{on} \approx \frac{3}{4} \cdot \frac{V_{DD}}{I_{DSAT}} \cdot \left(1 - \frac{5}{6} \cdot \lambda \cdot V_{DD} \right)$$

[EE115C stuff]

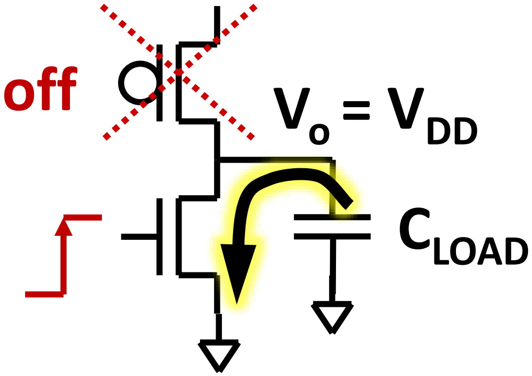
0^{th} Order Model: Step Input





- **NNOS and PMOS drive with maximum $|V_{GS}|$**
- **$I = CdV/dt$, $\Delta t = C\Delta V/I$**
  - Discharge C_{LOAD} in VSat
  - Discharge in Triode

0^{th} Order Model: Discharge Model



- I = CdV/dt
- Δt = CΔV/I

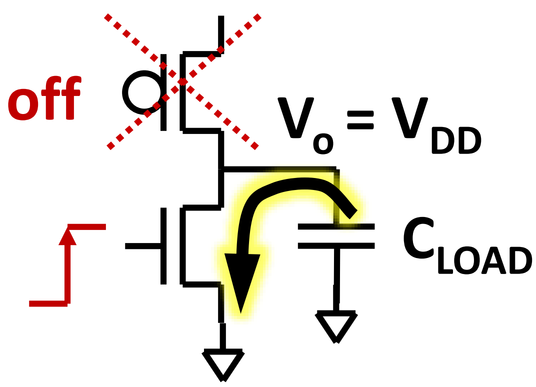
- Discharge in VSat
  - $V_{DSAT} < V_{out} < V_{DD}$

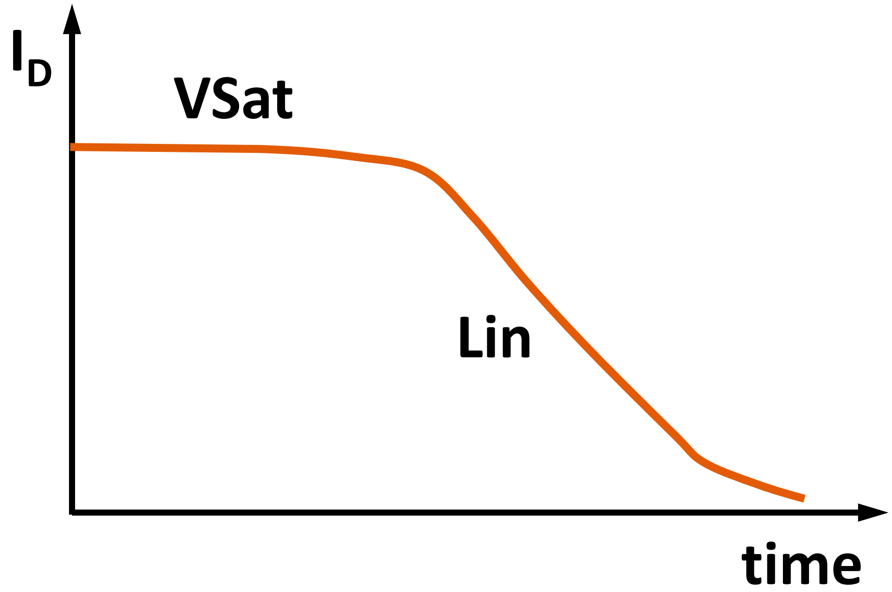
$$\Delta t_1 = C_{LOAD} \cdot \frac{V_{DD} - V_{DSAT}}{I_{DSAT,avg}}$$

- Discharge in Triode
  - Remainder of the way

$$\Delta t_2 = C_{LOAD} \cdot \int_{V_{DSAT}}^{0} \frac{dV}{I_D}$$

Output Transition of 0^{th} Order



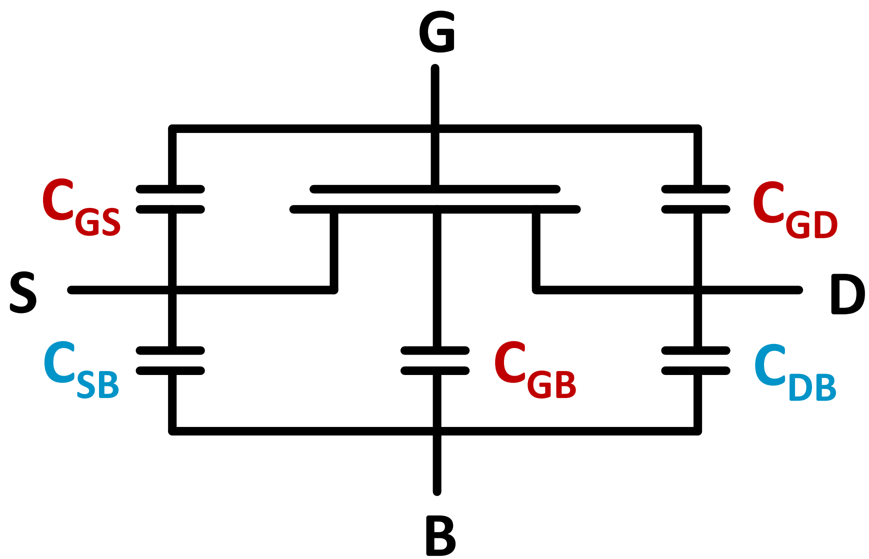


- **Solvable equations, BUT**
  - Unrealistic input
  - C_{LOAD} not linear
  - Only Vsat matters

- **Empirical model**
  - Slope correction
  - Effective C_{LOAD}
  - Linear resistance

Calculating the Capacitance

- **Like R, MOS capacitances are voltage-dependent**
- **Many capacitance models, here's a common one:**



Too detailed for designers

- **For delay analysis, we linearize gate and diffusion caps**
  - Gate capacitance (G-Ch, G-overlap)
  - S/D capacitance (Diffusion)

MOS Capacitances: Summary

- **Gate-Channel Capacitance**
  - $C_{GC} = C_{ox} \cdot W \cdot L_{eff}$ (Off, Lin)
  - $C_{GC} = (2/3) \cdot C_{ox} \cdot W \cdot L_{eff}$ (VSat)
- **Gate Overlap Capacitance**
  - $C_{GSO} = C_{GDO} = C_O \cdot W$ (All)

⇨ C_{gate}

- **Junction/Diffusion Capacitance**
  - $C_{diff} = C_j \cdot L_S \cdot W + C_{jsw} \cdot (2L_S + W)$ (All)

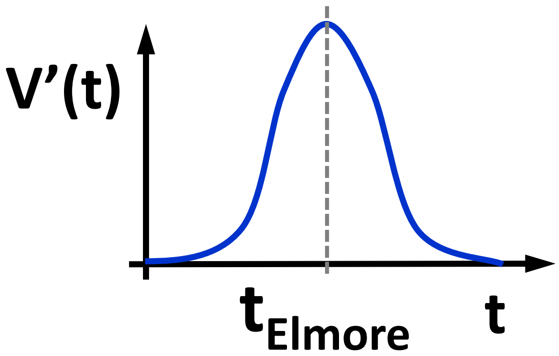
⇨ $C_{parasitic}$

$\boxed{\gamma = C_{par} / C_{gate} < 1}$ $\boxed{C \propto W}$ **(fF / µm)**

Elmore Delay

Elmore Delay (1948)

- **Defined as the first moment of the impulse response**
  - Derivative of the unit step response, V'(t)



$$\boldsymbol{t_{Elmore} = \int_0^{\infty} t \cdot v'(t) \cdot dt}$$

when $\boldsymbol{\int_0^{\infty} v'(t) \cdot dt = 1}$

- **Works for monotonic waveforms**
- **Works well with symmetric impulse response**
  - Reasonable for an output transition of a gate

The ln(2) Issue

- **Ideal RC response has a non-symmetric V'(t)**
  - This results in a positive skew (overestimated delay)
  - Dominant-pole approximation:

$$\boldsymbol{V(t) = 1 - e^{-\frac{t}{RC}}}$$

- **The 50% point delay**
  - For ½ = $e^{t/RC}$, t = t_{Elmore}ln(2) → a factor of 0.69
- **t_{Elmore} is the upper bound on gate delay**
  - If we use a slow input transition (instead of a step), the factor approaches 1

The ln(2) Issue: Another Look

$$\boxed{t_p = 0.69 \cdot t_{Elmore} = 0.69 \cdot RC}$$

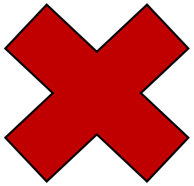
- **Account for the error by characterizing gate resistance**
  - Use RC delay to calculate $R_{effective}$
  - $R_{effective}$ already includes the ln(2) factor

$$\boxed{t_p = R_{effective}C}$$

Slope dependent

The Impact of Input Slope

- **Model the delay as t_p = 0.69RC** (step response)
  - **Non-step input:** rise/fall time is absorbed in R
  - **R is different** than the one extracted from I-V



Too many R's to keep track of…

Input Slope: A Better Model

Delay is linearly dependent on input rise/fall time:

$$\boxed{t_p = 0.69RC + \eta \cdot t_{slope}}$$

- **η is the slope factor** (typical values: 0.1 – 0.2)
- **The model is limited to a range of fanouts**

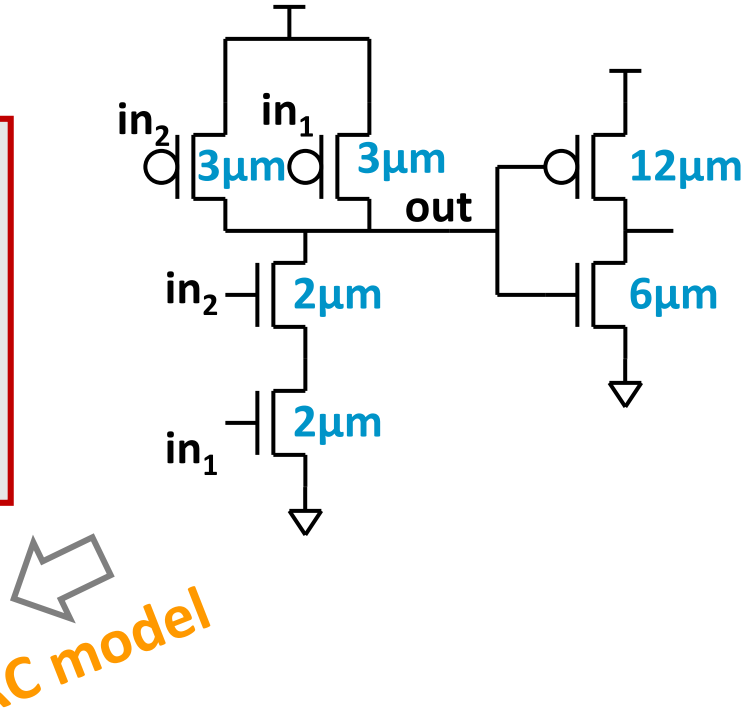
Another version of this model (stage n):

$$\boxed{t_{p(n)} = t_{p(n),step} + \beta \cdot t_{p(n-1),step}}$$

- **β is the slope factor** (typical values: find by simulation)
- **Slope is proportional to step-delay of previous stage**

Example 2.1: RC Gate Delay

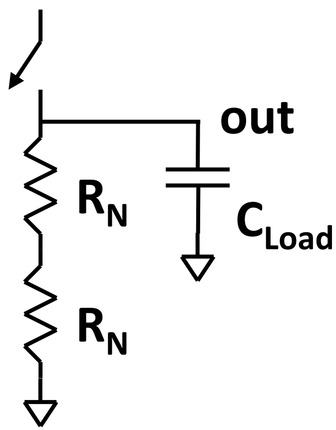
- **NAND2 driving an Inverter**



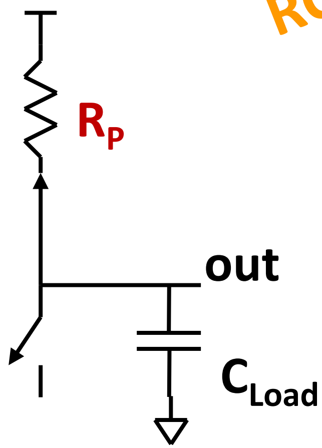
Assumptions (Sat):

- R_N = 3 kΩ-μm, R_P = 7.5 kΩ-μm
- $C_{GN} = C_{GP}$ = 2 fF/μm
- C_{DN} = 1.5 fF/μm, C_{DP} = 2 fF/μm

Pull-Down



Pull-Up



Calculate τ_{PU} and τ_{PD}:

- R_N = 1.5 kΩ, R_P = 2.5 kΩ [?]
- C_{Load} = 36 fF
- τ_{PU} = 90 ps, τ_{PD} = 108ps

Accounting for Velocity Saturation

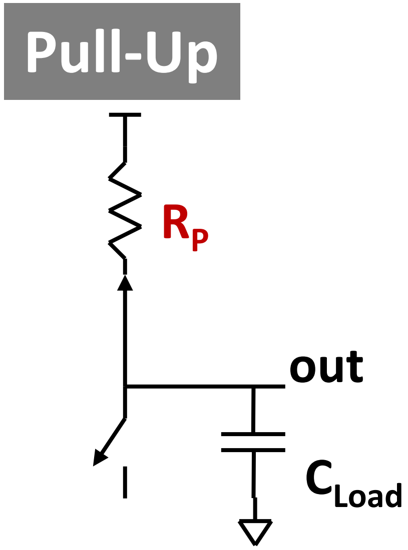
- **PMOS (no stack) is VSat**
  - $R_{P,no\text{-}stack} = \mathbf{6/5} \cdot R_{P,stack} = \mathbf{6/5} \cdot R_P$ (Sat)

- **VSat : less current → higher R**

Calculate R_P in VSat:

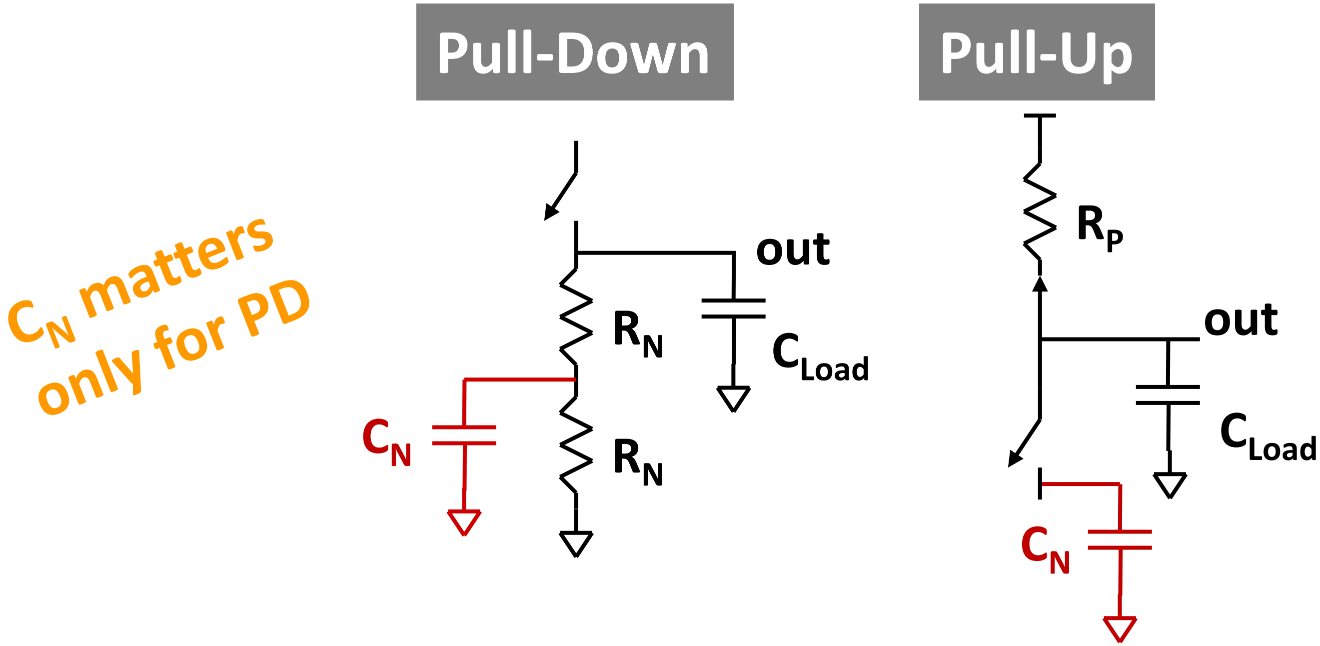
$R_P = 6/5 \cdot 2.5\ k\Omega = 3\ k\Omega$

- **C_{Load} = 36fF**
- **τ_{PU} = 108 ps (instead of 90ps)**



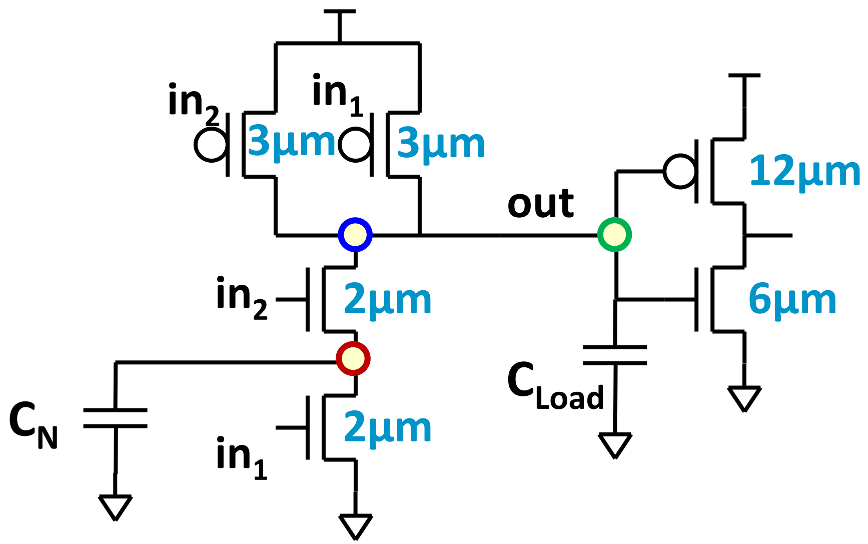
Including Self-Loading Capacitance

- **C_N: diffusion cap** (depends on the layout and sharing)



- **Model is now RC network and depends on input**
  - In_1 switching assumed

Finding the Capacitances



Assumptions (Sat):

- $C_{GN} = C_{GP} = 2$ fF/µm
- $C_{DN} = 1.5$ fF/µm
- $C_{DP} = 2$ fF/µm

Calculate C_{Load} and C_N:

- $C_{Load} = C_{inv} + C_{par} = 51$ fF
  - $C_{inv} = 2\cdot(12 + 6) = 36$ fF
  - $C_{par} = 2\cdot3 + 2\cdot3 + 1.5\cdot2 = 15$fF
- $C_N = 1.5 \cdot 2 = 3$ fF

Components:

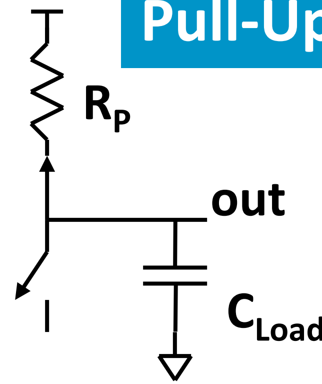
- **Gate**
- **Diffusion**
- **Shared diff**

Calculate RC Time Constants

τ_{PU} = 127.5 ps

Worst-case RC (In_1, In_2)

Pull-Up

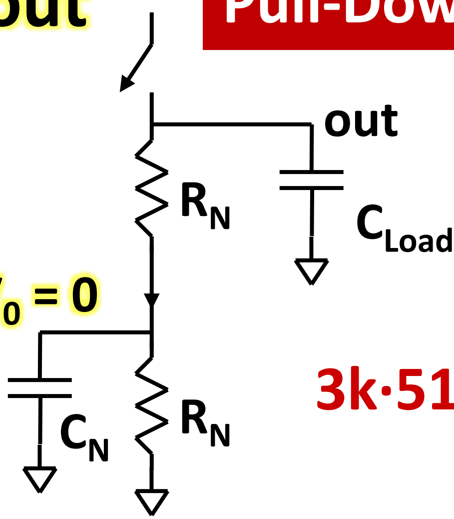


2.5k·51f

τ_{PD} = 153 ps

In_2-out

Pull-Down

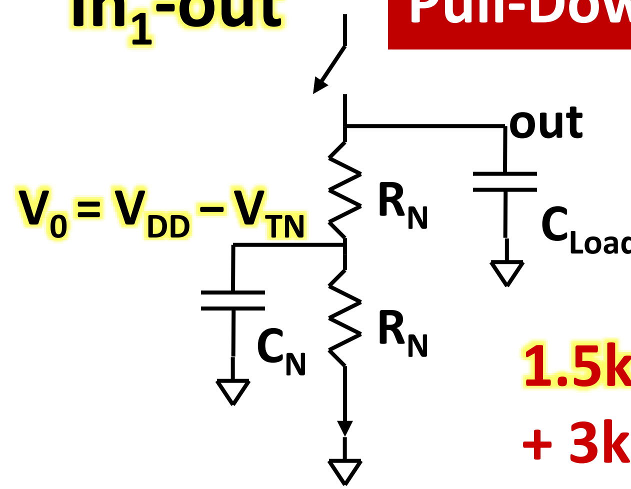


3k·51f

τ_{PD} = 157.5 ps

In_1-out

Pull-Down



1.5k·3f

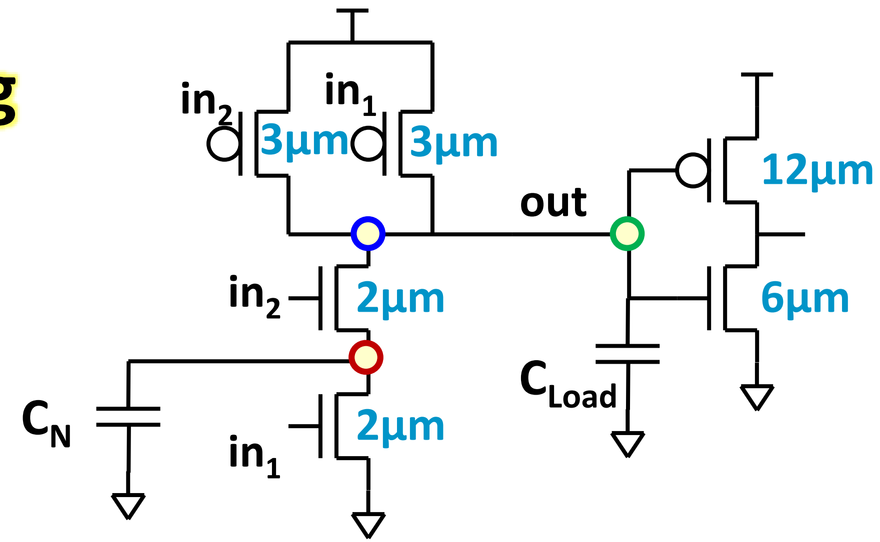
+ 3k·51f

Two Components of Delay

1 • Delay due to **self-loading**
  - Blue and red capacitances

2 • Delay due to **gate loading**
  - Green capacitances

$$C_{Load} = C_{self} + C_{gate}$$



Write delay as 2 parts:

$$\tau_{PU} = R_P C_{self} + R_P C_{gate}$$

$$\tau_{PD} = R_N(C_N + 2C_{self}) + 2R_N C_{gate}$$

Note the high self-loading delay

$C \cdot \Delta V/I$ Delay Model

- **Based on the capacitance charging and discharging**
- **ΔV is the voltage to the transition ($\sim V_{DD}/2$)**

- **Similar except we are breaking R into 2 components**
  - Averaging of V/I
  - I is an average drive current

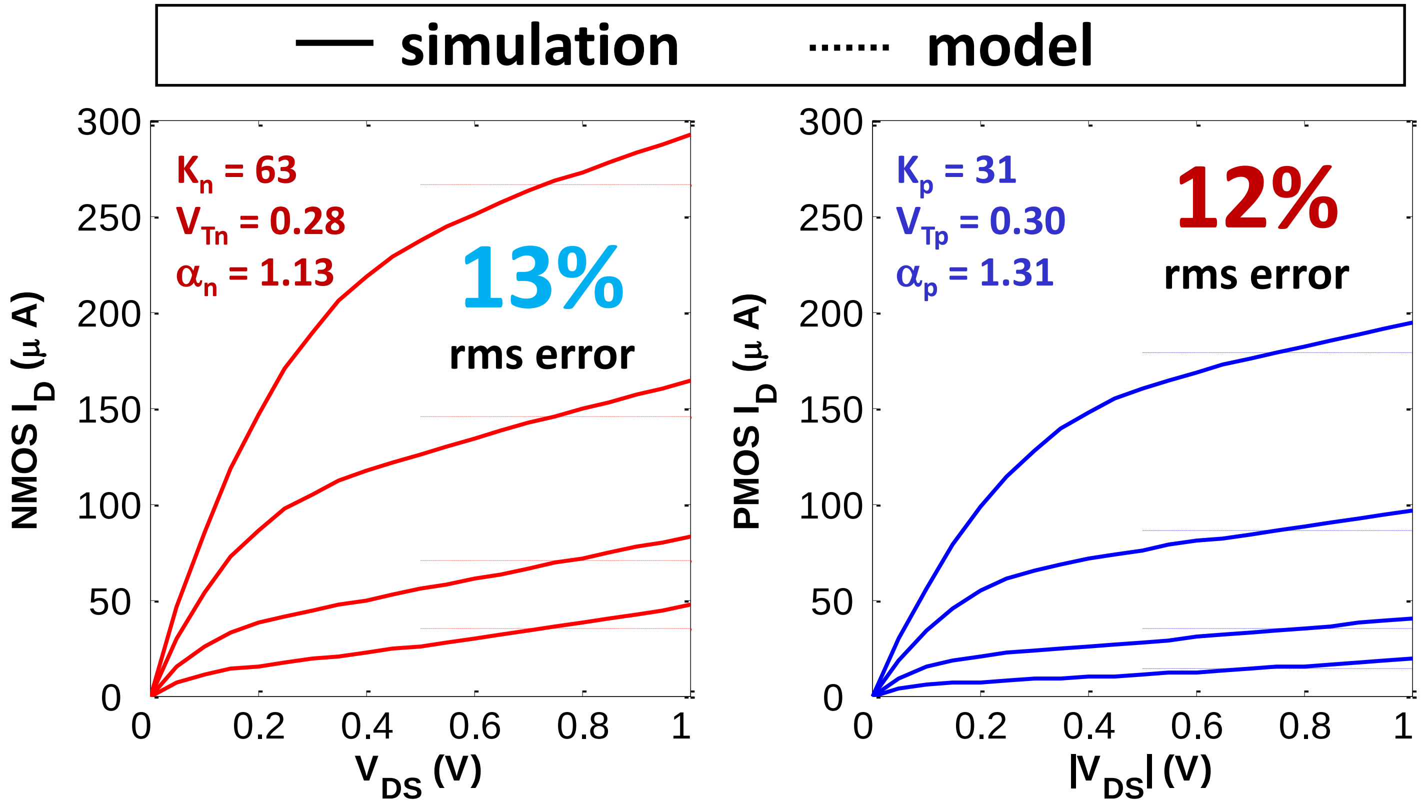
- **Helps understand what determines R**
  - $I \propto$ mobility and W/L
  - $I \propto (V_{GS} - V_T)$, $V_{GS} \propto V_{DD}$
  - Can anticipate what might happen if V_{DD} drops

Alpha-Power-Law Model

- **Bad for current**
- **Good for delay**

Alpha-Power Model: Saturation Current

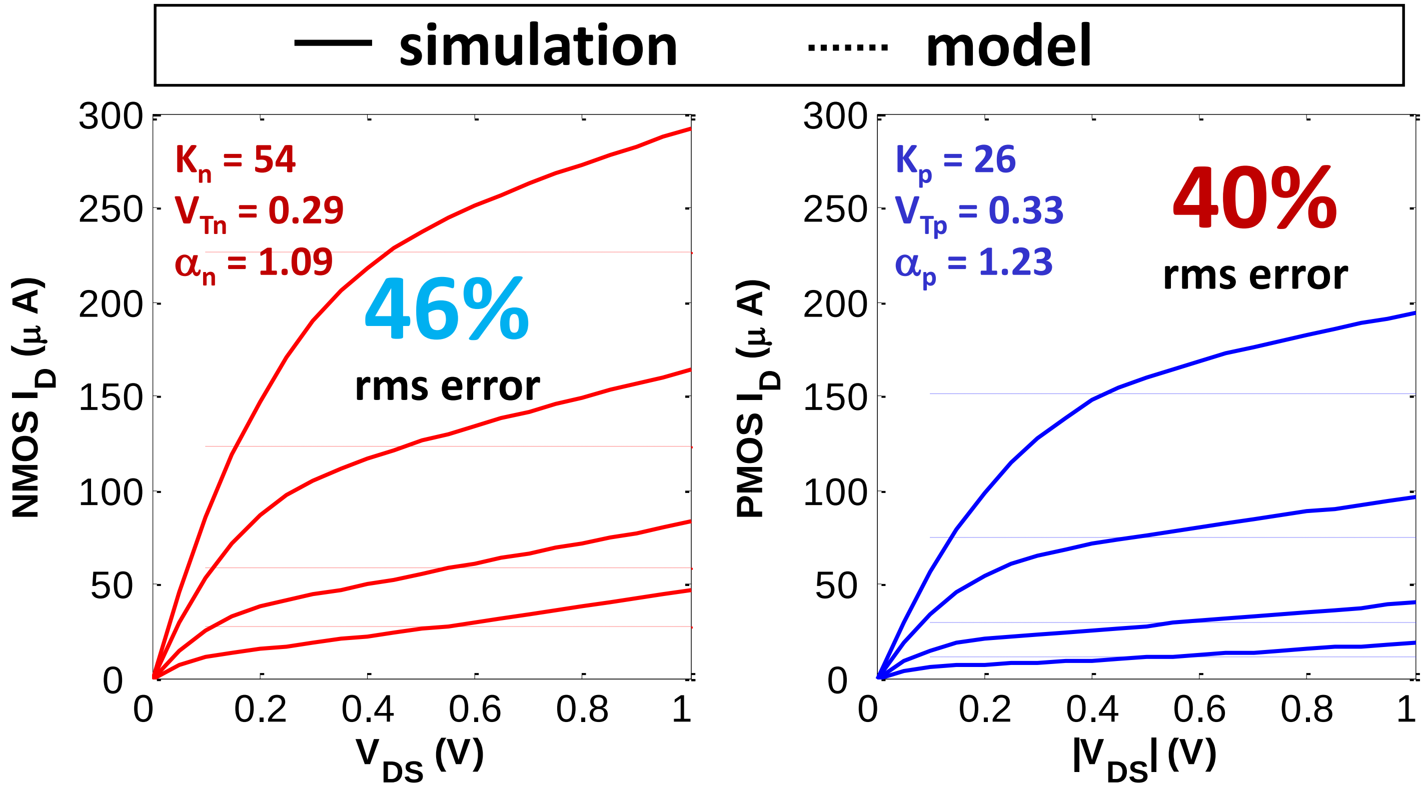
- $|V_{DS}| > 0.5V$



The model could be refined to include CLM

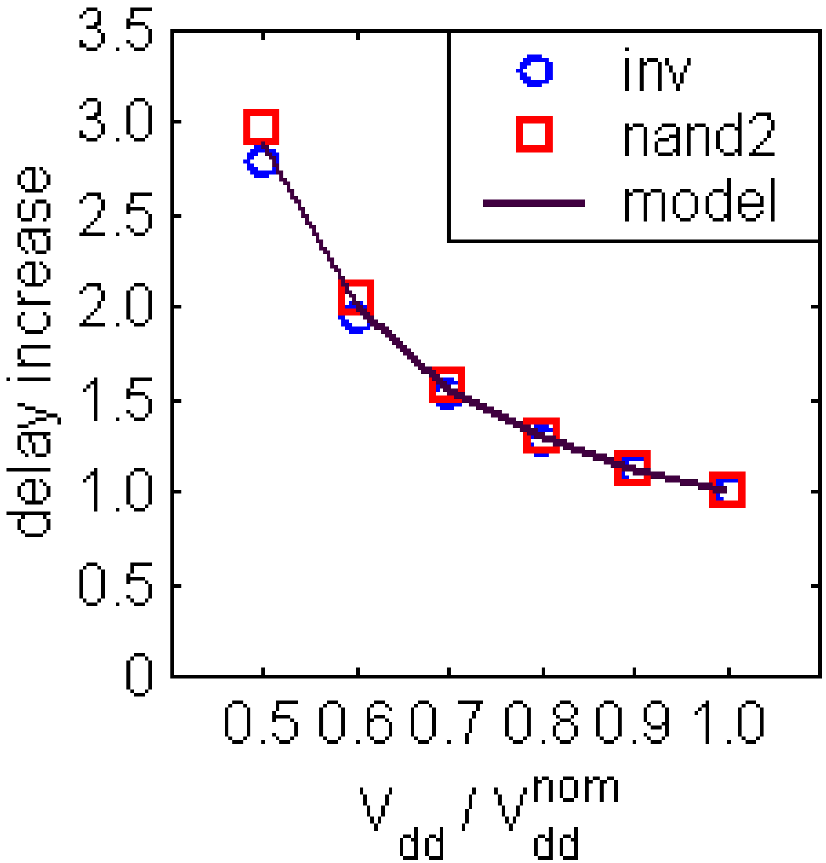
Saturation + Linear: Error Increases

- $|V_{DS}| > 0.1V$ 13% → 40%+ error



Alpha-power model does not fit well in linear region

Alpha-Power Model: Great for Delay



- **Start from 1st principles**

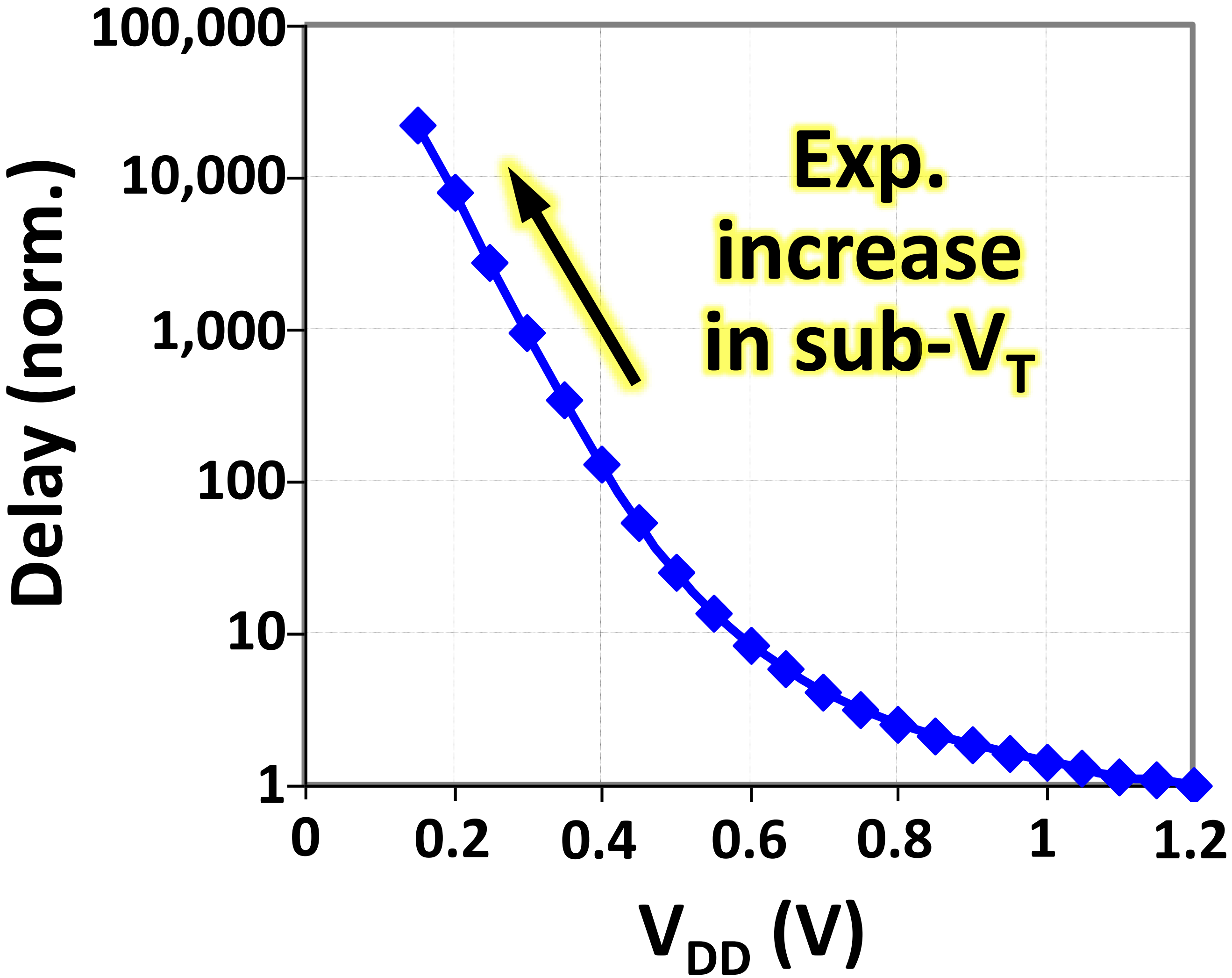
$$Delay = C \cdot \frac{\Delta V}{I_{avg}}$$

- **Delay = f (W, V_{DD})**

Fitting parameters:
V_{on}, α_d, K_d

$$Delay = \frac{K_d \cdot V_{DD}}{(V_{DD} - V_{on} - \Delta V_T)^{\alpha_d}} \cdot \left(\frac{W_{out}}{W_{in}} + \frac{W_{par}}{W_{in}} \right)$$

Gate Delay as a Function of V_{DD}



Summary

- **Device R and C determine circuit performance**
- **Elmore delay (approximation): initial insight into design**
  - Step response, does not account for signal slopes
  - Several models to account for slope (+ more coming)
  - Simulation-based parameter extraction most accurate (next lecture)

Next lecture:

- **Logic design concepts**
- **Simulation-based models**
- **Gate vs. wire delay**
- **Gate sizing basics**