ECE520 – VLSI Design

Lecture 6: Dynamic Behavior of CMOS Inverter

Payman Zarkesh-Ha

Office: ECE Bldg. 230B
Office hours: Wednesday 2:00-3:00PM or by appointment
E-mail: pzarkesh@unm.edu

Review of Last Lecture

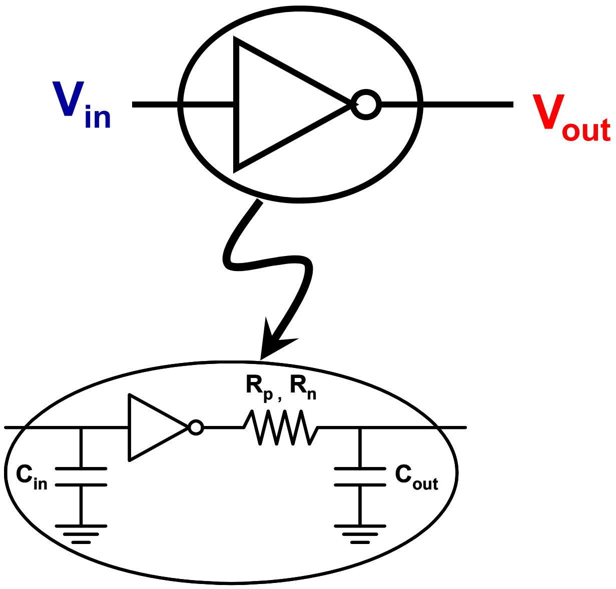
- ☐ **Static Behavior of CMOS Inverter**
  - **Switching threshold**
  - **Noise margin**
  - **CMOS Voltage-Transfer Characteristic (VTC)**
- ☐ **CMOS Inverter Robustness**
  - **Device variations**
  - **V_{dd} scaling (minimum supply voltage)**

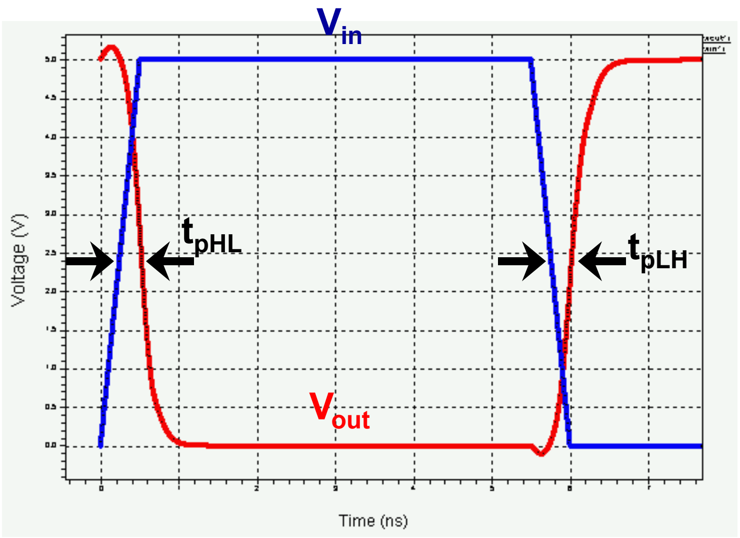
Today’s Lecture

- ❑ **Dynamic Behavior of CMOS Inverter**
    - **Computing the capacitances**
    - **Propagation delay model**
    - **Dynamic power**
    - **Power due to direct-path current**
    - **Leakage power**
    - **Some design techniques**

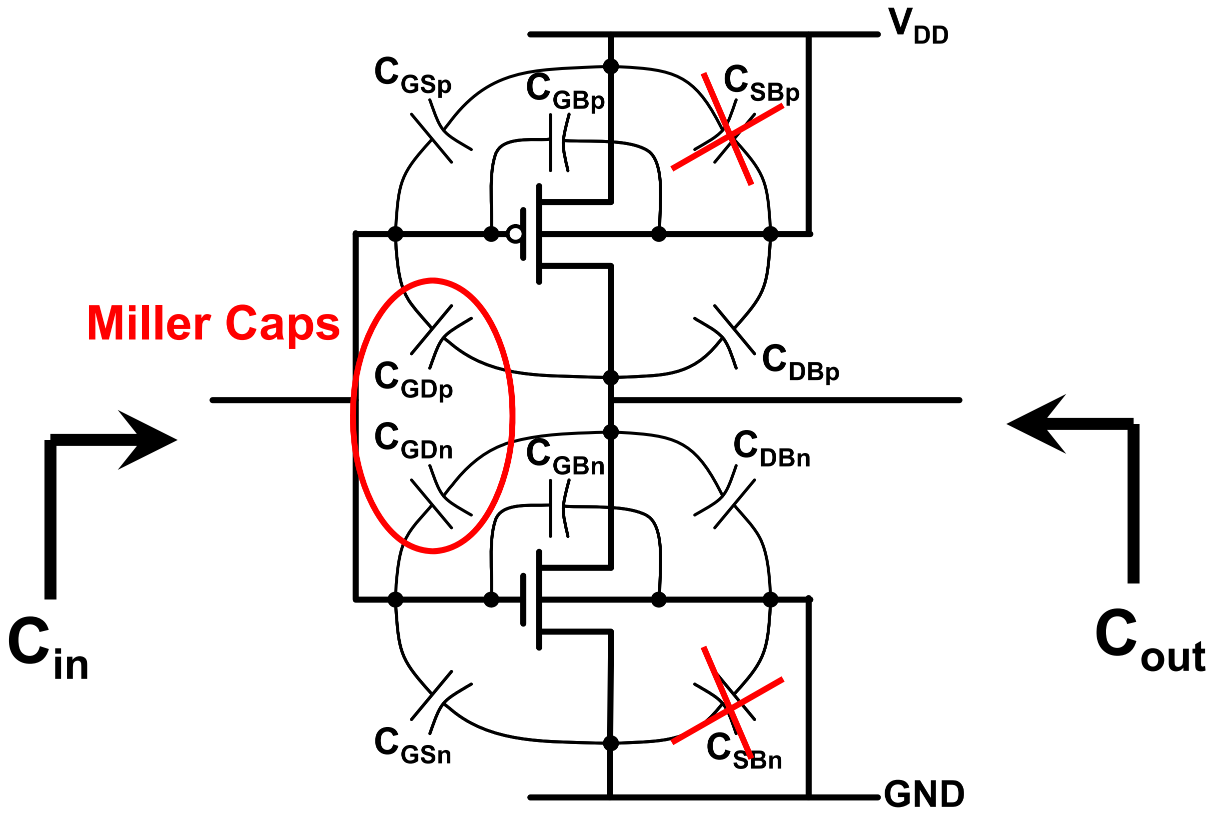
Definition: Propagation Delay

- ☐ **Definition of propagation delay**
  - **is delay from where input crosses 50%Vdd to where output crosses 50%Vdd**
  - **Remember: the value of 50%Vdd is from switching threshold voltage (V_M)**
  - **t_{pHL} is propagation delay when *<u>output</u>* switches from “High to Low”**
  - **t_{pLH} is propagation delay when *<u>output</u>* switches from “Low to High”**
- ☐ **To compute delay, the inverter must be simplified**





Parasitic Capacitances Components

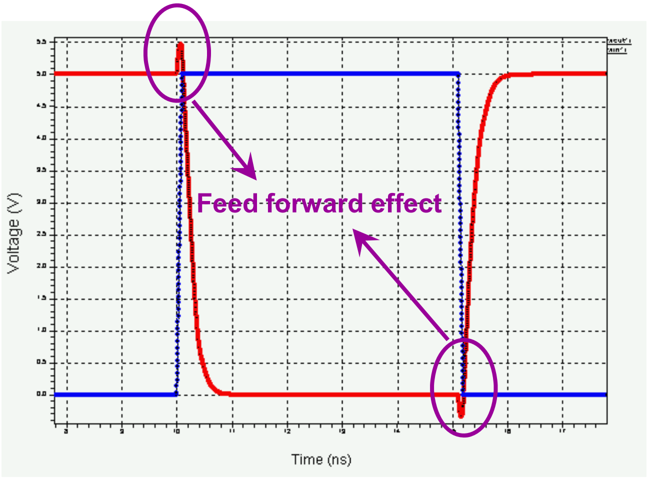


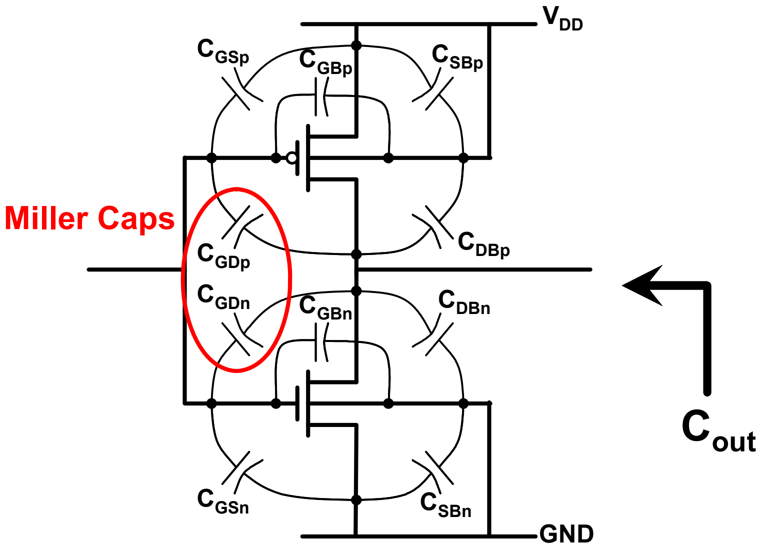
$$C_{in} = (C_{GDp} + C_{GSp} + C_{GBp}) + (C_{GDn} + C_{GSn} + C_{GBn})$$

$$C_{out} = (2C_{GDp} + C_{DBp}) + (2C_{GDn} + C_{DBn})$$

Output Caps: Miller Capacitances

- **C_{GD} is often the most important component, because**
  - **They are often large**
  - **Since both the input and output are moving in opposite direction, the effective capacitance is doubled (miller effect)**
  - **This is also a source of overshoot in the output waveform (feed forward effect)**
  - **Assuming very fast input rise time, the devices are either cutoff or saturation (for the most part). In either case C_{GD} is only the overlap capacitance i.e. $C_{GD} = C_{GDOV} = WC_{ox}X_d$ (refer to lecture 3)**





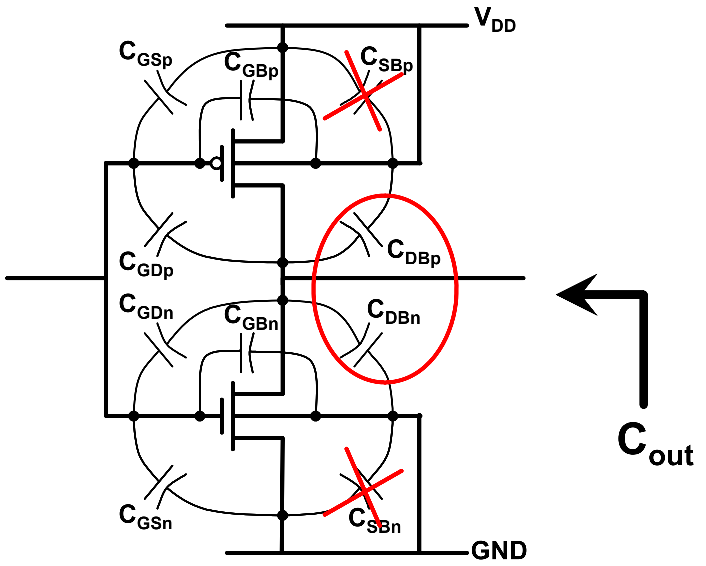
Output Caps: C_{DB} Junction Capacitances

- ☐ C_{DB} is drain ***junction*** capacitances of NMOS and PMOS
  - They are non-linear voltage dependent capacitances (refer to lecture 3)
  - To simplify calculations, we assume a constant value for C
  - A linearized approximation can be used to find average C_{DB} over the range of interest (V_{low} to V_{high}) :

$$C_{eq} = K_{eq} C_{j0}$$

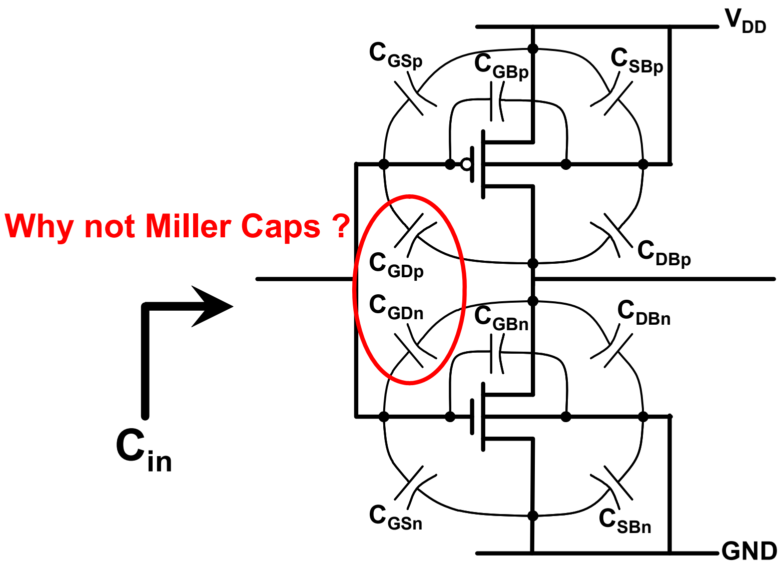
$$C_j = \frac{C_{j0}}{\left(1 - V_{SB}/\varphi_0\right)^m} \qquad \varphi_0 = \frac{KT}{q} Ln\left(\frac{N_A N_D}{n_i^2}\right)$$

$$K_{eq} = \frac{2\sqrt{\varphi_0}}{V_{high} - V_{low}}\left(\sqrt{\varphi_0 + V_{high}} - \sqrt{\varphi_0 + V_{low}}\right)$$



Input Caps: C_{GD} Capacitances

- **C_{GD} is computed similar to what we used for out put cap**
  - **The devices are either cutoff or saturation (for the most part). In either case C_{GD} is only the overlap capacitance i.e. $C_{GD} = C_{GDOV} = WC_{ox}X_d$ (refer to lecture 3)**
  - **But the impact is not doubled for C_{in} (no Miller effect)**
  - **Why?**

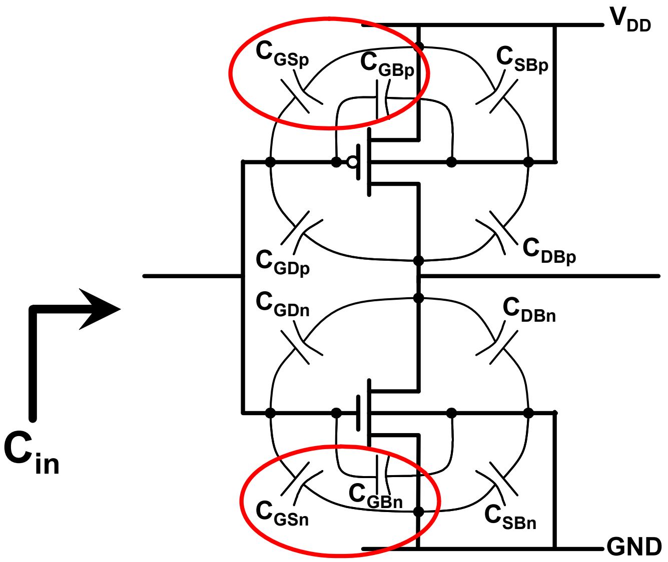


Why not Miller Caps ?

Input Caps: Gate Capacitances

- Input gate capacitances comprised of C_{GS} and C_{GB}
  - Assuming very fast input rise time, the devices are either cutoff or saturation (for the most part)
  - Referring to lecture 3:
    - In cutoff $C_{GS} = C_{GSCH} + C_{GSOV} = 0 + WC_{OX}X_d$
    - Also, in cut off C_{GB} is $C_{GB} = C_{GBCH} = WC_{OX}L_{eff}$
  - Referring to lecture 3:
    - In saturation $C_{GS} = C_{GSCH} + C_{GSOV} = 2/3\ WC_{OX}L_{eff} + WC_{OX}X_d$
    - Also, in saturation C_{GB} is $C_{GB} = C_{GBCH} = 0$
- Therefore:

$$C_{Gate} = C_{GS} + C_{GB} \approx C_{GSOV} + WC_{ox}L_{eff}$$

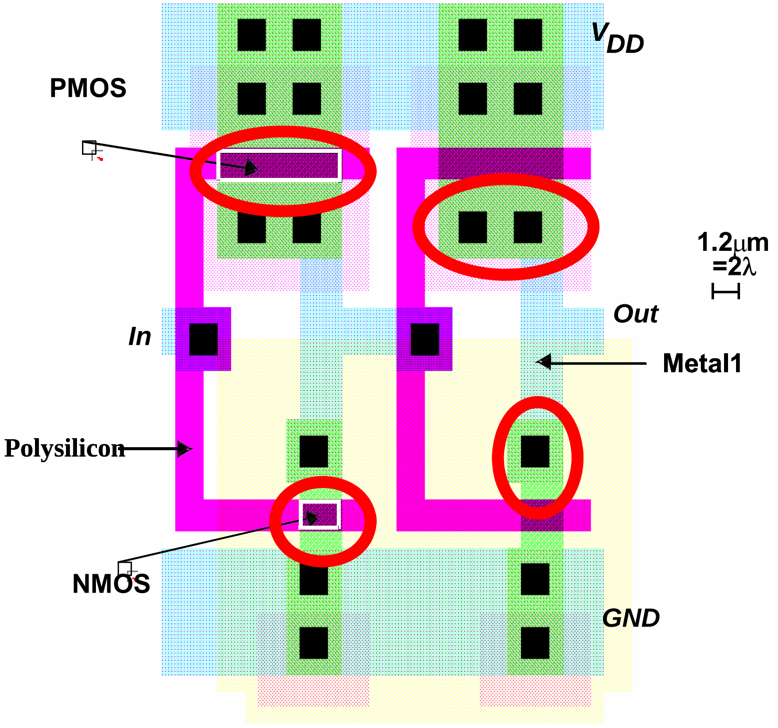


Summary of Parasitic Capacitances

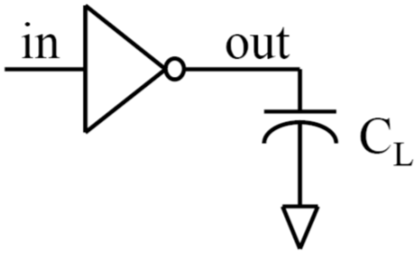
- Input and output capacitances of an inverter can be approximated by a linear capacitance for hand calculations:

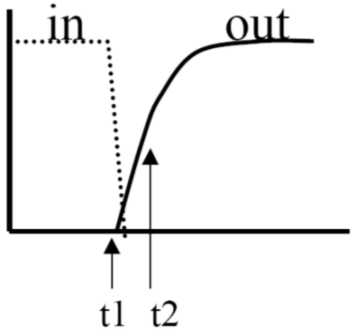
$$C_{in} \approx \left(C_{GDOVn} + C_{GSOVn} + W_n C_{ox} L_{effn}\right) + \left(C_{GDOVp} + C_{GSOVp} + W_p C_{ox} L_{effp}\right)$$

$$C_{out} \approx \left(2C_{GDOVn} + K_{eqn} C_{jn}\right) + \left(2C_{GDOVp} + K_{eqp} C_{jp}\right)$$



Propagation Delay Model

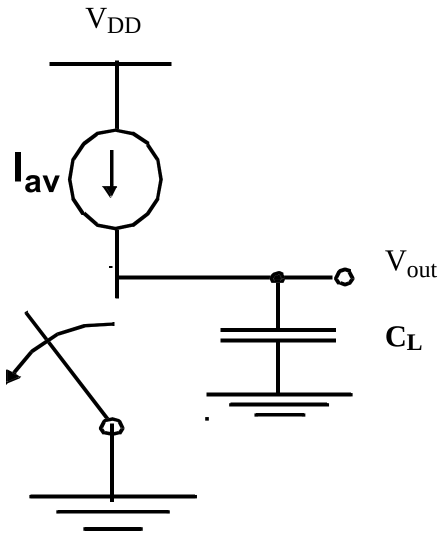




$$Q = I \cdot \Delta t = C \cdot \Delta V$$

$$\Delta t = \frac{C \cdot \Delta V}{I}$$

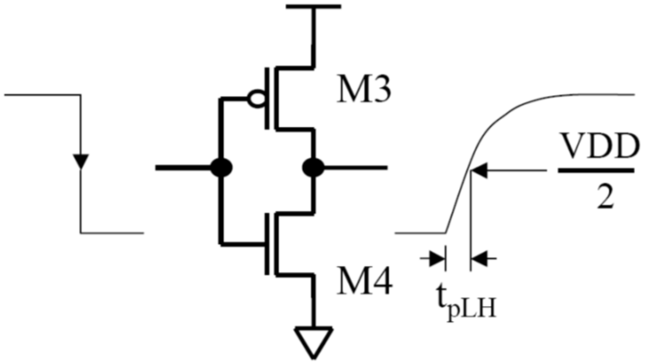
- ☐ **Propagation delay is defined as the time between the input reaching $V_{DD}/2$ and the output reaching $V_{DD}/2$**
- ☐ **To simplify the model, let's assume I is a constant I_{av}**



$$t_{pLH} = t_2 - t_1 = \frac{C_L \cdot (V_{DD}/2)}{I_{av}}$$

Propagation Delay Model

- How to compute I_{av} ?
  - Assume step input
  - NMOS goes into cutoff and stays there
  - PMOS goes into saturation at first because $|V_{DS}| > |V_{GS}| - |V_T|$
  - PMOS will transition to linear, however, before V_{out} reaches $V_{DD}/2$



$$I_{av} = \frac{I_{DS}\left(V_{out} = 0\right) + I_{DS}\left(V_{out} = V_{DD}/2\right)}{2}$$

$$I_{av} = \left(\frac{K'_p}{2}\right)\left(\frac{W_p}{L_p}\right)\left(\frac{\left(V_{DD} - |V_{tp}|\right)^2}{2} + \frac{V_{DD}\left(V_{DD} - |V_{tp}|\right)}{2} - \frac{V_{DD}^2}{8}\right)$$

Propagation Delay Model

- ☐ A simpler model for I_{av} can be obtained by assuming that the PMOS stays in saturation the whole time, therefore acts as an ideal current source

$$I_{av} = \left(\frac{K'_p}{2}\right)\left(\frac{W_p}{L_p}\right)\left(V_{DD} - |V_{tp}|\right)^2$$

$$t_{pLH} = t_2 - t_1 = \frac{C_L \cdot (V_{DD}/2)}{I_{av}} \quad \Rightarrow \quad t_{pLH} = \frac{C_L \cdot V_{DD}}{K'_p\left(\frac{W_p}{L_p}\right)\left(V_{DD} - |V_{Tp}|\right)^2}$$

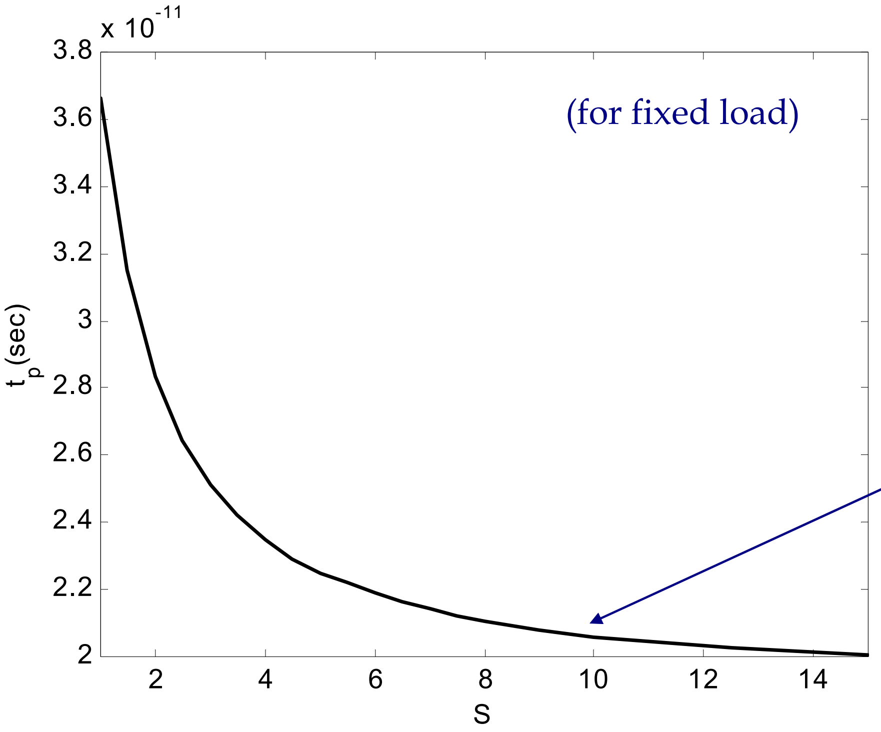
- ☐ Assuming $V_{DD} >> V_{Tp}$

$$t_{pLH} = \frac{C_L}{K'_p\left(\frac{W_p}{L_p}\right)V_{DD}}$$

- ☐ Same arguments hold for t_{pHL}

Propagation Delay Versus Transistor Width

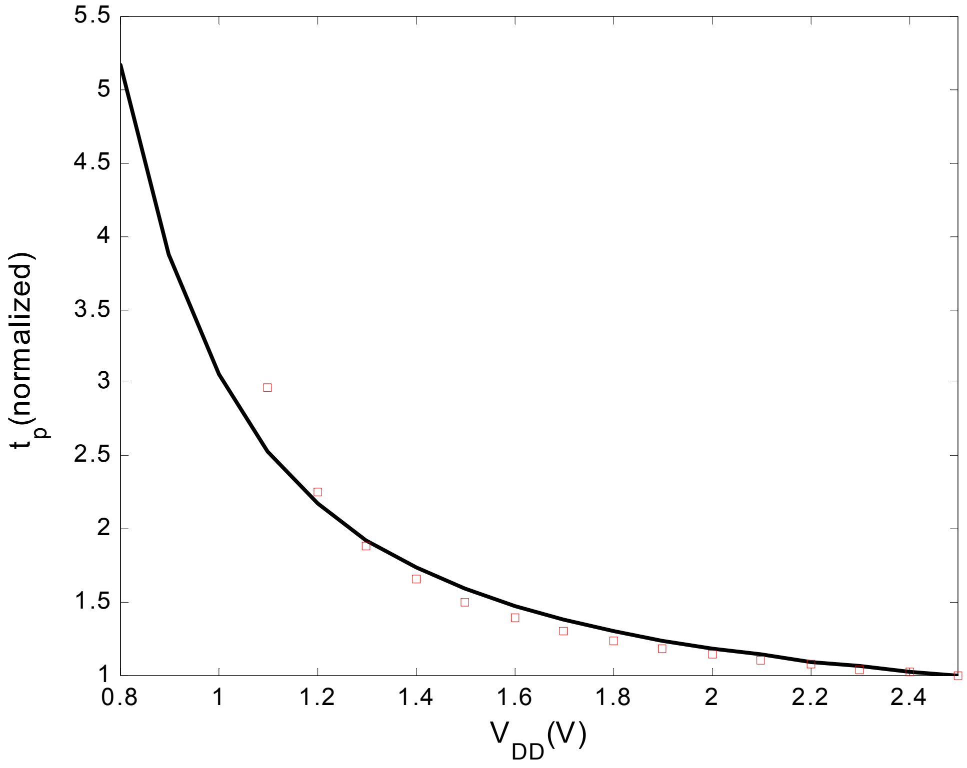
- ☐ **Delay should decrease by increasing W_p**
- ☐ **However, C_{out} has components (C_{GD} and C_{DB}) that also increase with W_p**
- ☐ **As a result diminishing returns for increased device width**



Self-loading effect:
Intrinsic capacitances
dominate

Propagation Delay Versus V_{DD}

- ❑ **Delay should decrease by increasing V_{DD}**
- ❑ **However, when V_{DD} goes beyond V_{DSAT} delay become insensitive to supply voltage**



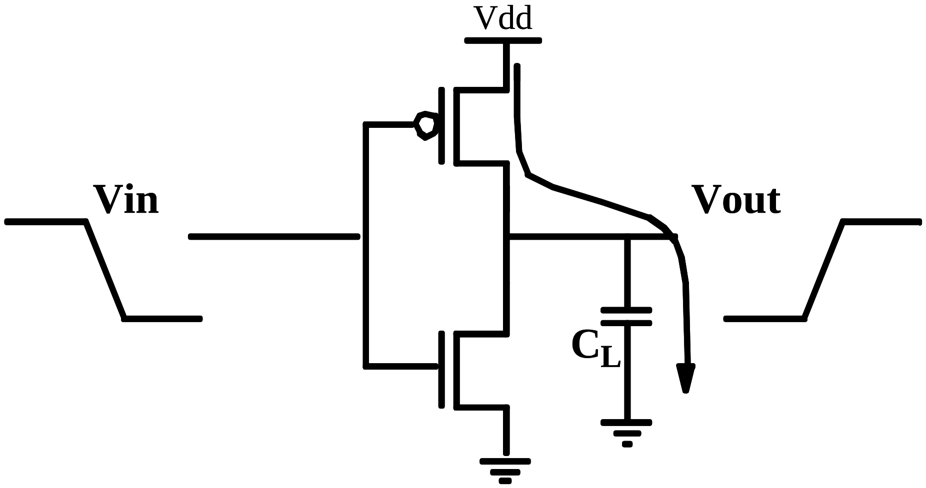
Minimum Delay Design Techniques

- **Reduce C_{in} and C_{out} – Careful layout, keep drain diffusion area as small as possible**
- **Reduce wiring capacitance – Careful layout, keep devices as close as possible**
- **Increase (W/L) of devices – Need to be careful not to get into self-loading effect**
- **Increase V_{DD} – Need to be careful not to get into V_{DSAT} or velocity saturation**

CMOS Power Consumption

- ☐ **Dynamic Power**
- ☐ **During switching there is an interval in which both NMOS and PMOS conduct resulting in "short-circuit", or "direct path" currents**
  - **Power consumption due to this is usually about 5% of active power**
  - **Mainly important because it causes large instantaneous currents, which cause considerable noise on power bussing (IR drops) especially for slow inputs**
  - **Essentially the VDD and VSS power rails are tied together through the R_{on} of the transistors**
  - **Can be dealt with by specifying a maximum bus driver size—as we'll see later, a bigger driver is not the answer to faster wire delay**
- ☐ **Leakage Power (no steady state power, only leakages)**
  - **Subthreshold leakages are becoming very significant**
  - **Now even gate leakages (tunneling)**

CMOS Dynamic Power Consumption



Energy/transition = $C_L * V_{dd}^2$

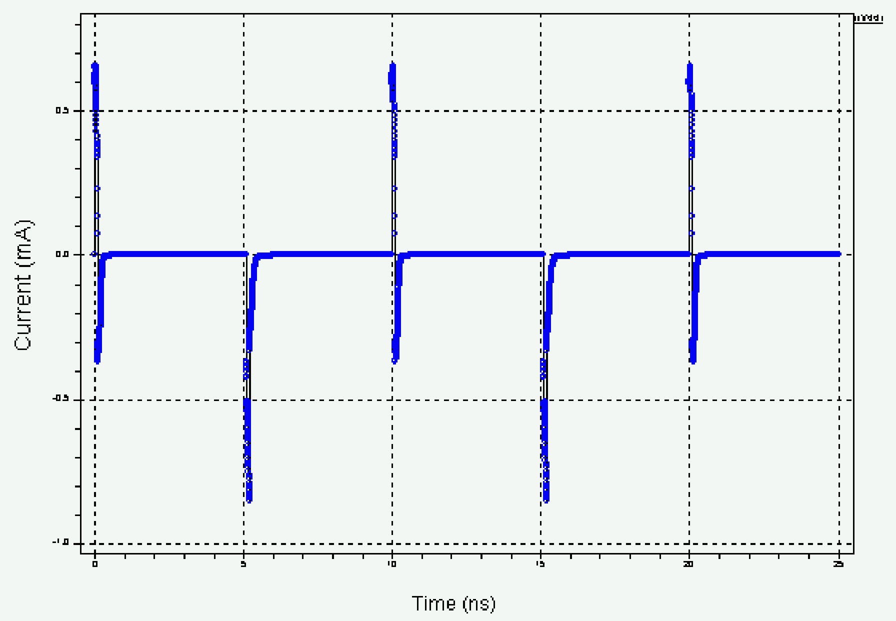
Power = Energy/transition * f = $C_L * V_{dd}^2 * f$

$$P_d = \frac{1}{T}\int_0^T i_{VDD}(t)\cdot V_{DD}dt = \frac{V_{DD}}{T}\int_0^T C_L\frac{dV_{out}}{dt}dt = \frac{C_L V_{DD}}{T}\int_0^{V_{DD}} dV_{out} = \frac{C_L V_{DD}^2}{T} = C_L V_{DD}^2 f$$

- **Need to reduce C_L , V_{DD}, or f to reduce power**

Power Due to Direct-Path Current

- ❑ **Generally is about 5% of the active power**
- ❑ **Really care about this when driving large busses or other large loads**
- ❑ **Basically, the NMOS and PMOS both on are shorting the supplies!**



Computing Active Power

- ☐ **Estimation of active power is not easy**
  - **Need to know average switching rate of the transistors**
  - **Generally only the clocks toggle on each cycle**
  - **Even with clock gating technique, the clock doesn't toggle every cycle**

$$P_{active} = C_{eff} V_{DD}^{2} f = \alpha C_{total} V_{DD}^{2} f$$

  - **α is the "average activity factor" of the circuit**
  - **This can only be understood architecturally, i.e., on an average**
  - **clock cycle, how much of the total capacitance is switched?**
  - **α is low for memories, higher for logic, highest for clocks**
  - **After a design is back, C_{eff} can be found by measuring power**

- ☐ **Power depends strongly on what the chip is doing**
- ☐ **Just assume 5% crowbar (multiply P_{active} by 1.05)**

CMOS Leakage Power Consumption

- ☐ **Leakage is easier to estimate**
- ☐ **Just sum the transistor and assume ½ on and ½ off**
  - **The on transistors don't contribute leakage, the off transistors do (that's why it's called I_{off}!)**

$$P_{leakage} = \left(\sum \frac{W_n}{2} I_{off(n)} + \sum \frac{W_p}{2} I_{off(p)} \right) V_{DD}$$

  - **Other terms (I_{gate}) can be important on sub 90nm processes**
- ☐ **Note that both active and leakage power are dependent on transistor width**
  - **Keep widths down**
  - **Designers tend to widen transistors during the timing effort no matter how ineffective they know it to be!**

Minimum Power Design Techniques

- **Prime choice: Reduce voltage!**
  - **Recent years have seen an acceleration in supply voltage reduction**
  - **Design at very low voltages still open question (0.6 … 0.9 V by 2010!)**
- **Reduce switching activity**
- **Reduce physical capacitance**
- **Reduce clock frequency, but use multi-core architecture to enhance performance**