E&CE 437

Integrated VLSI Systems

The Inverter

M. Sachdev

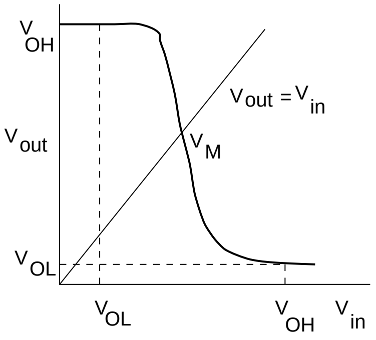
Inverter: Introduction

- **The inverter is the simplest of all digital logic gates**
  - ❍ However, building understanding for its properties and operation is crucial for the design and analysis of larger/complexer logic gates
  - ❍ We will discuss,
  - ❍ General properties of an inverter (and logic gates)
  - ❍ Inverter implementation issues in MOS and bipolar technologies

General Properties: Area and Complexity

- **Small area is a desirable property for a digital logic gate**
  - Larger packing density
  - Small parasitic capacitances
  - Shorter interconnects,
  - smaller chip area, hence higher number of devices per wafer (lower cost)
- **Fewer transistors for a logic gate usually results into smaller area**
  - Hence, minimum possible number of transistors for a give gate are important

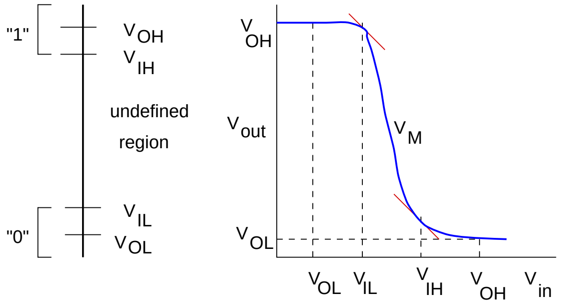
Inverter: Static Behavior



- **Static behavior of an inverter can be described by voltage transfer characteristics**
  - V_{OL} and V_{OH} are low and high logic values of an inverter
  - V_M is the logic (gate, or switching) threshold of an inverter

Noise Margin

- **In real life applications, output voltage of a gate may not have the nominal value**
  - Owing to load, high switching speed, etc.
  - Hence, it is desirable to define an acceptable voltage range for logic 1 and logic 0, respectively



Noise Margin, Regenerative Property & Fan-in, Fan-out

- For logic robustness large noise margin is desirable
- $NM_L = V_{IL} - V_{OL}$
- $NM_H = V_{OH} - V_{IH}$

- **Logic gates have the property to restore the proper output logic values despite of non-ideal input levels (regenerative action)**
- **Fan-out is the number of logic gates that can be driven from a given logic gate (maximum fan-out)**
- **Fan-in is the number of inputs to a logic gate, large fan-in results in poorer performance**

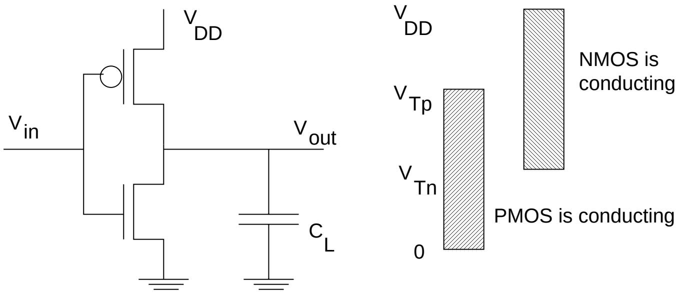
Dynamic Behavior

- **Performance is an important attribute to any logic gate and determines its dynamic behavior**
  - Performance is measured by propagation delay through the logic gate (t_{plh} and t_{phl}) and its rise and fall times
- **Performance is a strong function of the output load**
  - Often ring oscillator is used to compare different technologies objectively
  - However, ring oscillator performance has little correlation to actual logic gate or IC performance
- **Exercise**
  - *Simulate a 21 stage CMOS ring oscillator and measure its frequency*

Power & Energy Consumption

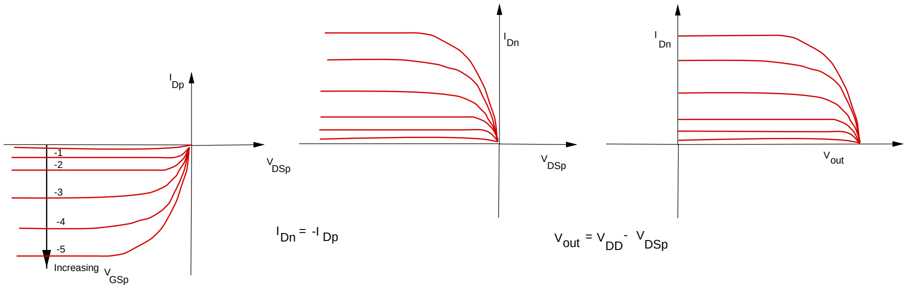
- **Power consumption of a gate conveys how much heat it dissipates and how much energy is consumed per cycle**
  - Power consumption influence many critical decisions in the design of an IC (e.g., packaging, cooling, long term reliability, etc.)
  - $P_{peak} = i_{peak}V_{supply}$
  - $$P_{av} = \frac{1}{T}\int_0^T p(t)dt = \frac{V_{supply}}{T}\int_0^T i_{supply}(t)dt$$
- **Power consumption has dynamic as well as static components**
  - Dynamic part is associated with charging and discharging of capacitance and is proportional to frequency
  - Static part is owing to sub-threshold leakage

CMOS Inverter: Static Behavior



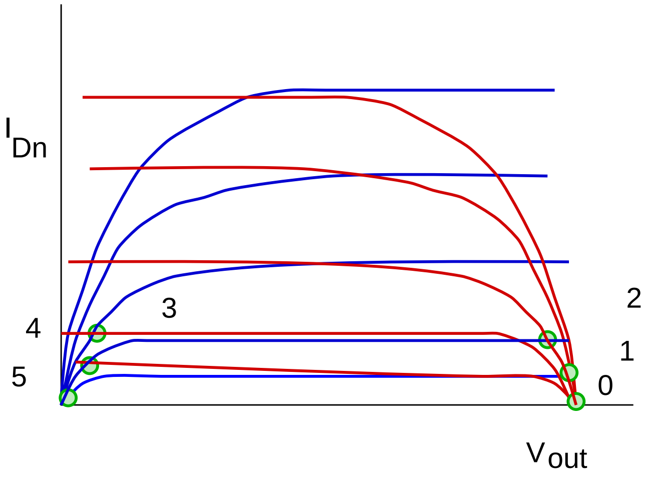
- **A transistor has on resistance, R_{on}, when conducting, other wise almost infinite resistance**
  - ❍ In steady state, either PMOS or NMOS provides a low impedance path from output to VDD or GND
  - ❍ Infinite input impedance for the inverter

Voltage Transfer Characteristics

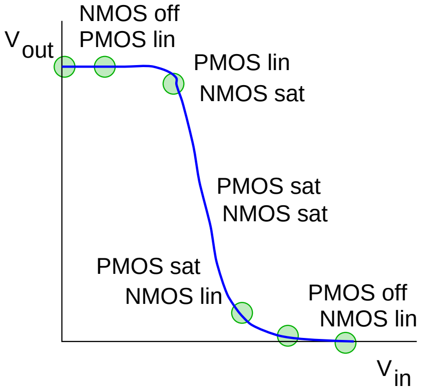


- **Voltage transfer characteristics for an inverter can be deduced from the load lines**
  - ❍ Current characteristics of PMOS and NMOS should be superimposed on each other
  - ❍ I-V characteristics of PMOS should be transformed

Load Lines



Voltage Transfer Characteristics



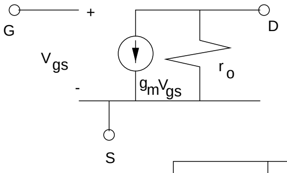
- **Exercise**
  - *Simulate a voltage transfer characteristics of an inverter*

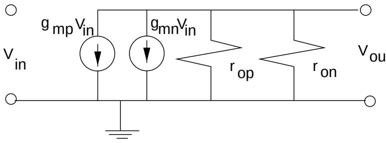
CMOS Inverter: An Amplifier

- In VTC of inverter, VOL and VOH were defined where

$$\frac{\partial V_{out}}{\partial V_{in}} = -1$$

- At this point, small signal gain, g, of the amplifier (inverter) is equal to -1





	g_m	r_o
Linear	kV_{DS}	$[k(V_{GS} - V_T - V_{DS})]^{-1}$
Saturation	$k(V_{GS} - V_T)$	$1/\lambda I_D$

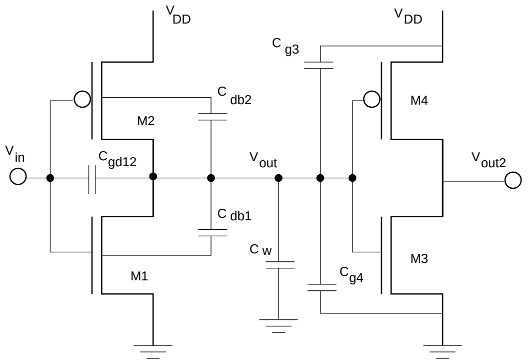
- Gain, g, is $g = \frac{V_{out}}{V_{in}} = -(g_{mn} + g_{mp})(r_{on} \parallel r_{op}) = -1$ when V_{in} = V_{IH} and V_{IL}
- At V_{in} =V_{IH}, PMOS and NMOS transistors can be assumed to be in saturation and linear regions, respectively
- g_{mn} = $k_n V_{out}$ and g_{mp} = $k_p(V_{DD}$ -V_{IH} - $|V_{Tp}|)$
- $\lambda_p = 0$, ignoring the channel length modulation
- $r_{on} = \frac{1}{k_n(V_{IH} - V_{out} - V_{Tn})}$ and $r_{op} = \infty$

- **Putting these formulas in equation**
  - $g = -(k_n V_{out} + k_p(V_{DD} - V_{IH} - |V_{Tp}|))\left(\frac{1}{k_n(V_{IH} - V_{out} - V_{Tn})}\right) = -1$
- **Also, the static current through PMOS and NMOS should be the same**
  - $k_n\left[\langle V_{IH} - V_{Tn}\rangle V_{out} - \frac{V^2_{out}}{2}\right] = \frac{k_p}{2}\langle V_{DD} - V_{IH} - |V_{Tp}|\rangle^2$
- **With these two equations, equations for V_{out} and V_{IH} can be found**

- **Similarly, equations can be derived for $V_{in} = V_{IL}$**
- **In the same fashion we can find out an analytical expression for the inverter threshold (V_M)**
  - V_M is the point where $V_{out} = V_{in}$; in this region both transistors are in saturation
  - Equating their currents
  - $$\frac{k_n}{2}[\langle V_M - V_{Tn}\rangle^2] = \frac{k_p}{2}\langle V_{DD} - V_{IH} - |V_{Tp}|\rangle^2$$
  - Or
  - $V_M = \frac{r(V_{DD} - |V_{Tp}|) + V_{Tn}}{1 + r}$ where $r = \sqrt{\frac{k_p}{k_n}}$
- **If threshold voltages are equal and $k_p = k_n$, then $V_M = V_{DD}/2$**
  - Under these conditions, both PMOS and NMOS have equal strength and the inverter is balanced

CMOS Inverter: Dynamic Behavior

- **Dynamic behavior (delay) of an inverter is determined by the time it takes to (dis)charge output capacitance**



  - We are interested in determining transient response of V_{out} assuming V_{in} is driven by an ideal voltage source

- **Several capacitances contribute to overall capacitance**
  - **C_{gd12}:** M1, M2 in steady state are either in cut-off or in saturation; overlap capacitance of M1, M2 contributes to this capacitance; the gate capacitance is either between gate and bulk (cut-off) or gate and source (saturation)
  - **C_{db1} and C_{db2}:** Capacitance between drain and bulk is due to reverse biased pn-junction. This is a non-linear capacitance which is approximated as linear
  - **C_w** : This is the capacitance due to interconnect
  - **C_{g3}** and **C_{g4}** : We assume that the gate capacitance of loading gates is between V_{out} and V_{DD} (GND). Overlap and gate capacitances are clustered into a single component

Propagation Delay

- **The propagation delay can be computed by integrating the capacitor (dis)charge current**
  - $$t_p = C_L \int_{v1}^{v2} \frac{1}{i(v)}(dv) = \frac{C_L(V_2 - V_1)}{I_{av}}$$
- **Propagation delay is defined as time between input reaching 50% to output reaching 50% of full value**
  - $$t_p = \frac{C_L(V_{OH} - V_{OL})/2}{|I_{av}|}$$
  -
  - This equation holds for both t_{plh} and t_{phl} transitions
  - Assuming that the input transition (V_{DD} -> 0) is abrupt then only PMOS contributes to the current
  - The PMOS is in saturation so long $V_{out} < |V_{Tp}|$ and after that it is in linear mode, therefore currents in respective regions

- $$I(V_{out} = 0) = \frac{k_p}{2}\langle V_{DD} - |V_{Tp}|\rangle^2$$
- and
- $$I\left(V_{out} = \frac{V_{DD}}{2}\right) = k_p\langle(V_{DD} - |V_{Tp}|)\frac{V_{DD}^2}{2} - \frac{V_{DD}^2}{8}\rangle$$
- The average current, I_{av}, can be computed as
- $$I_{av} = \left(\frac{I(V_{out} = 0) + I\left(V_{out} = \frac{V_{DD}}{2}\right)}{2}\right) = \frac{k_p}{2}\langle\frac{7V_{DD}^2}{8} + \frac{|V_{Tp}|^2}{2} - \frac{3V_{DD}|V_{Tp}|}{2}\rangle$$
- A simpler equation is arrived at if we assume that PMOS remains in saturation from V_{out} =0 to V_{DD}/2; In this case
- $$t_{plh} = \frac{C_L(V_{DD})}{k_p\langle V_{DD} - |V_{Tp}|\rangle^2} \approx \frac{C_L}{k_pV_{DD}}$$ If $V_{DD} >> V_{Tp}$

- Similarly for t_{phl}
- $$t_{phl} = \frac{C_L(V_{DD})}{k_n\langle V_{DD} - V_{Tn}\rangle^2} \approx \frac{C_L}{k_nV_{DD}}$$
- therefore,
- $$t_p = \frac{1}{2}(t_{plh} + t_{phl}) = \frac{C_L}{2V_{DD}}\left(\frac{1}{k_p} + \frac{1}{k_n}\right)$$

- **Design Challenges, how to reduce t_p?**
  - Increase k_p and k_n
  - Reduce C_L
  - Increase V_{DD}
  - .....

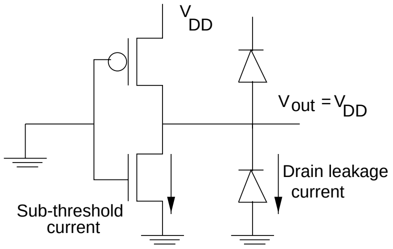
Second Order Performance Issues

- **Previously, we assumed that input transition is abrupt and only one transistor is on during (dis)charging process**
  - Signals have finite rise/fall times and for a brief period both PMOS and NMOS are on
  - t_{plh} increases as input fall time is increased
  - Smaller rise/fall times are also desirable for low power consumption
- **We assumed that the maximum (dis)charge current is saturation current of transistors (proportional to V^2_{DD})**
  - In small geometries, owing to velocity saturation, I_{av} is proportional to V_{DD}
  - Therefore, if $V_{DD} >> V_T$ $\quad t_p \approx \frac{C_L}{2}\left(\frac{1}{k_p} + \frac{1}{k_n}\right)$

- **In the timing analysis, we have so far ignored the source and drain resistance of (dis)charging device**
  - The source resistance affects the performance
(i): Effective V_{GS} is reduced (i.e., lowering the saturation current)
(ii): The source is no longer grounded, V_T of the transistor is increased due to body effect
  - $I_{n,\,sat,\,R} = \frac{k_n}{2}\langle V_{DD} - V_S - V_S V_{Tn}\rangle^2$
  -
  - where $V_S = R_S.I_{n,sat,R}$

Power Consumption and Power-Delay Product

- **Static power consumption**
  - Ideally, there should be no static power consumption since both transistors are never on simultaneously

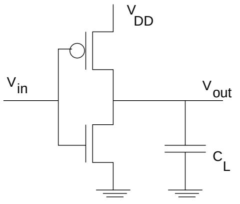


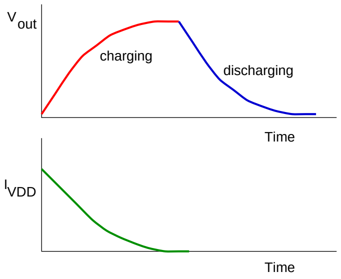
  - A small component is contributed by reverse biased diode leakage current

  - As mentioned before, at $V_{GS} = 0$, the transistor current is not absolutely 0 but a small component (sub-threshold current)
  - As we scale the technology, V_T is also scaled resulting in exponentially higher sub-threshold current
  - $P_{stat} = I_{leakage} V_{DD}$

- **Majority of power is consumed during switching; two components**

- **1. Charging of load capacitance, C_L**





- Energy, EVDD taken from supply during the transition and EC is stored on the capacitor at the end of transition

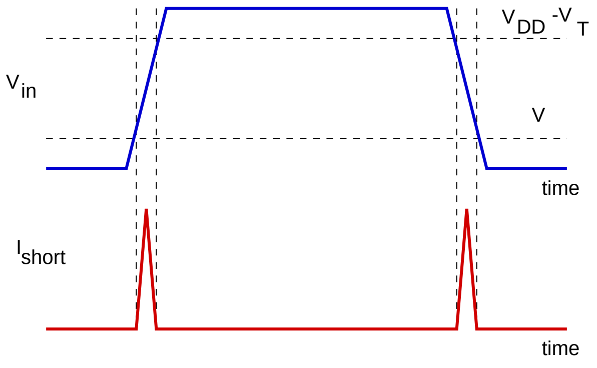
$$E_{VDD} = \int_0^{\infty} i_{VDD}(t)V_{DD}dt = V_{DD}\int_0^{\infty} C_L\frac{dv_{out}}{dt}dt = C_LV_{DD}\int_0^{V_{DD}} dv_{out} = C_LV_{DD}^2$$

$$E_C = \int_0^{\infty} i_{VDD}(t)v_{out}dt = \int_0^{\infty} C_L\frac{dv_{out}}{dt}v_{out}dt = C_L\int_0^{V_{DD}} v_{out}dv_{out} = \frac{C_LV_{DD}^2}{2}$$

- Energy stored on the capacitor is only half of what is drawn from the supply, rest of it dissipated in the PMOS transistor
- Energy dissipation is independent of transistor dimensions
- During the discharge phase, charge is removed from the capacitor and its energy is dissipated in the NMOS transistor
- Hence, in each switching cycle fixed energy $C_LV_{DD}^2$ is taken from the supply, to compute the power, we must multiply energy by frequency, $P_{dyn} = C_LV_{DD}^2f$

■ 2. Direct-Path Currents

- Owing to finite rise and fall times, both transistors are on for a brief period of time during transitions



- If we assume that the current spikes can be approximated as triangular waveform and V_{DD} >>$|V_T|$; then energy consumed per period

$$E_{dp} = V_{DD}\frac{I_{peak}t_r}{2} + V_{DD}\frac{I_{peak}t_f}{2} = V_{DD}I_{peak}\left(\frac{t_r + t_f}{2}\right)$$

Total Power Dissipation of an Inverter

- **Total power is the sum of three components**
  - $P_{tot} = P_{dyn} + P_{dp} + P_{stat} = C_L V_{DD}^2 f + V_{DD} I_{peak} \left(\frac{t_r + t_f}{2}\right) f + V_{DD} I_{leak}$
- **Energy per operation or Power-Delay Product**
  - Power-Delay Product (PDP) is a quality measure for a logic gate and is defined as amount of energy consumed in each cycle
- **Exercise**
  - *Optimize a CMOS inverter for minimum PDP. Also optimize the inverter for minimum power and minimum delay.*

Technology Scaling &CMOS

- **Ever since ICs were invented, dimensions are scaled to**
  - Integrated more transistors in the same area
  - Allow higher operational speed
- **Scaling has profound impact on many aspects of ICs**
- **Constant Voltage Scaling**
  - All device dimesions are scaled by a factor S
  - Voltage (i.e., V_{DD}) after the scaling is same as before
  - This method of scaling is followed till 0.8 micron
  - However for lower geometries, higher electric field resulted in poor device reliability

- **Therefore, for advanced technologies today Constant Field Scaling is followed**
  - All dimensions including power supply is scaled by a factor S

Parameter	Relation	CVS	CFS
W,L, t_{ox}		1/S	1/S
V_{DD}, V_T		1	1/S
Area	WL	$1/S^2$	$1/S^2$
C_{ox}	t_{ox}	S	S
C_L	$C_{ox}WL$	1/S	1/S
k_n, k_p	$C_{ox}W/L$	S	S
I_{av}	$k_{n,p}V^2$	S	1/S
J_{av}	I_{av}/Area	S^3	S
t_p (intrinsic)	C_LV/I_{av}	$1/S^2$	1/S
P_{av}	C_LV^2/t_p	S	$1/S^2$
PDP	C_LV^2	1/S	$1/S^3$