UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

The CMOS Inverter

A.Carini – Digital Integrated Circuits

The CMOS Inverter

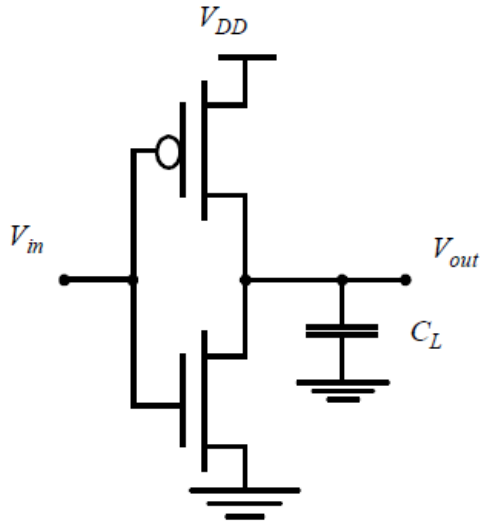
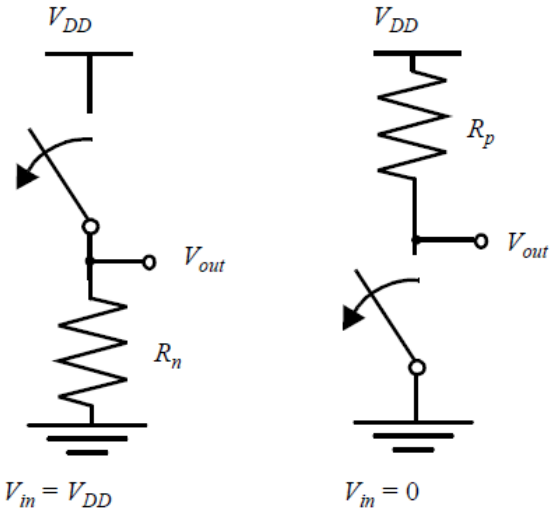


Figure 5.1 Static CMOS inverter. V_{DD} stands for the supply voltage.

The CMOS Inverter



(a) Model for high input (b) Model for low input

Figure 5.2 Switch models of CMOS inverter.

The CMOS Inverter

- The high and low output levels equal V_{DD} and *GND*, respectively. The voltage swing is equal to the supply voltage.
- The logic levels are not dependent upon the relative device sizes, so that the transistors can be minimum size. Gates with this property are called *ratioless.*
- In steady state, there always exists a path with finite resistance between the output and either V_{DD} or *GND*. A CMOS inverter has a *low output impedance*, which makes it less sensitive to noise and disturbances.
- The *input resistance* of the CMOS inverter is extremely high, as the gate of an MOS transistor is a virtually perfect insulator and draws no dc input current.
- No direct path exists between the supply and ground rails under steady-state operating conditions. The gate does not consume any static power.

The voltage-transfer characteristic

- The form of the voltage-transfer characteristic (VTC) can be graphically deduced by superimposing the current characteristics of the NMOS and the PMOS devices.
- Such a graphical construction is traditionally called *a load-line plot*.
- It requires that the *I-V* curves of the NMOS and PMOS devices are transformed onto a common coordinate set.
- Consider the input voltage V_{in}, the output voltage V_{out} and the NMOS drain current I_{DN} as the variables of choice.

$$I_{DSp} = -I_{DSn}$$

$$V_{GSn} = V_{in} \; ; \; V_{GSp} = V_{in} - V_{DD}$$

$$V_{DSn} = V_{out} \; ; \; V_{DSp} = V_{out} - V_{DD}$$

The voltage-transfer characteristic

- The load-line curves of the PMOS device are obtained by a mirroring around the *x*-axis and a horizontal shift over V_{DD}.

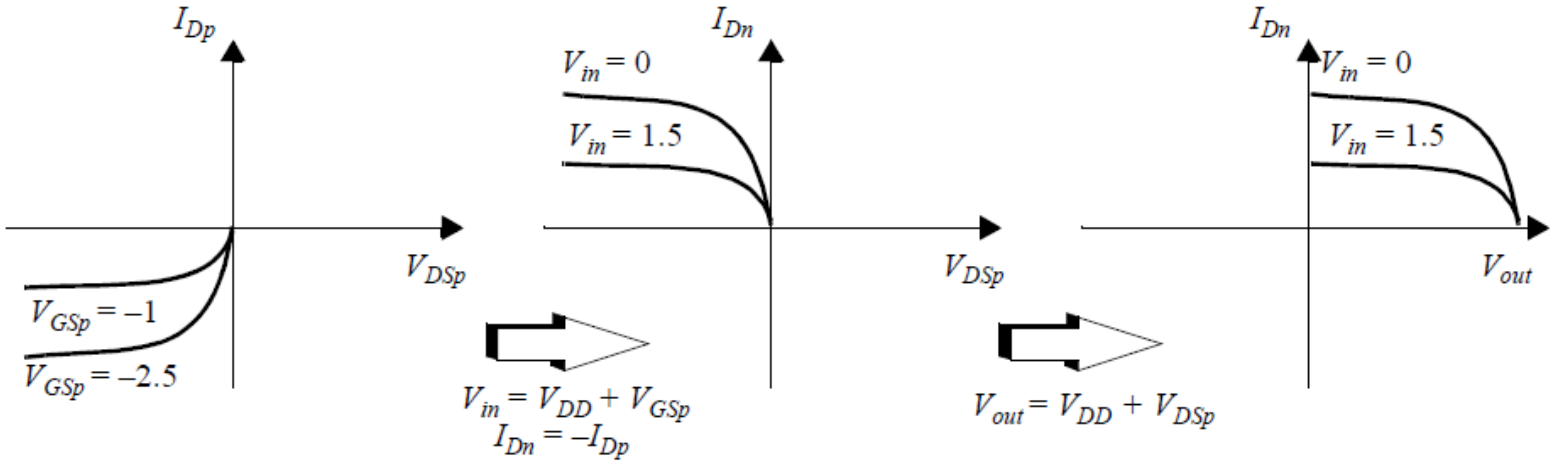


Figure 5.3 Transforming PMOS *I-V* characteristic to a common coordinate set (assuming VDD = 2.5 V).

The voltage-transfer characteristic

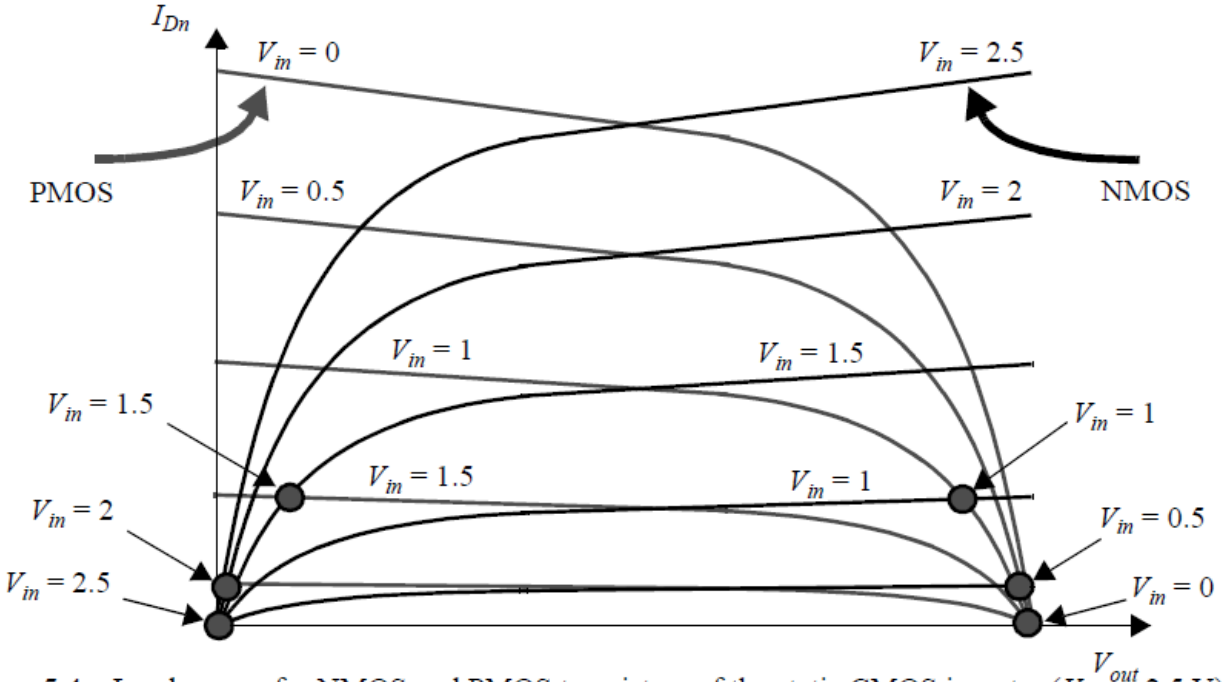


Figure 5.4 Load curves for NMOS and PMOS transistors of the static CMOS inverter ($V_{DD} = 2.5$ V). The dots represent the dc operation points for various input voltages.

The voltage-transfer characteristic

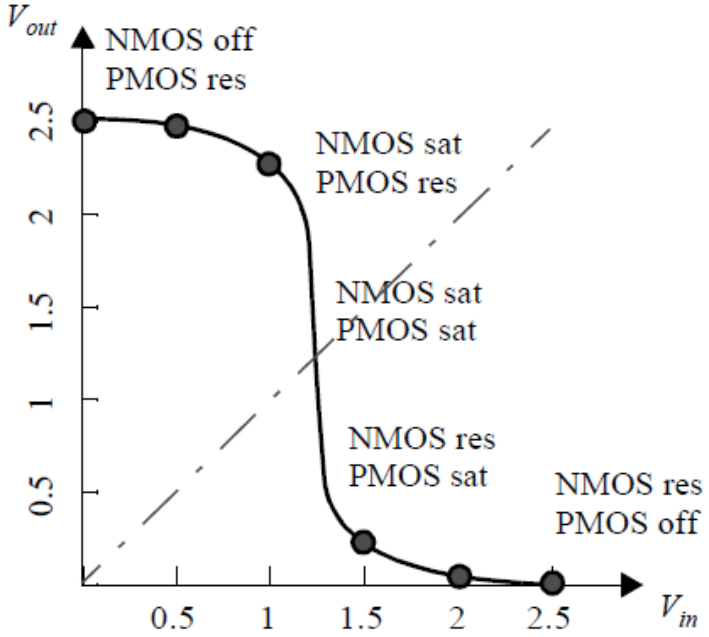


Figure 5.5 VTC of static CMOS inverter, derived from Figure 5.4 (V_{DD} = 2.5 V). For each operation region, the modes of the transistors are annotated — off, res(istive), or sat(urated).

Qualitative analysis of the transient behavior

- This response is dominated mainly by the output capacitance of the gate, C_L, which is composed of the drain capacitances of the NMOS and PMOS, the capacitance of the connecting wires, and the input capacitance of the fan-out gates.
- Assuming that the transistors switch instantaneously:

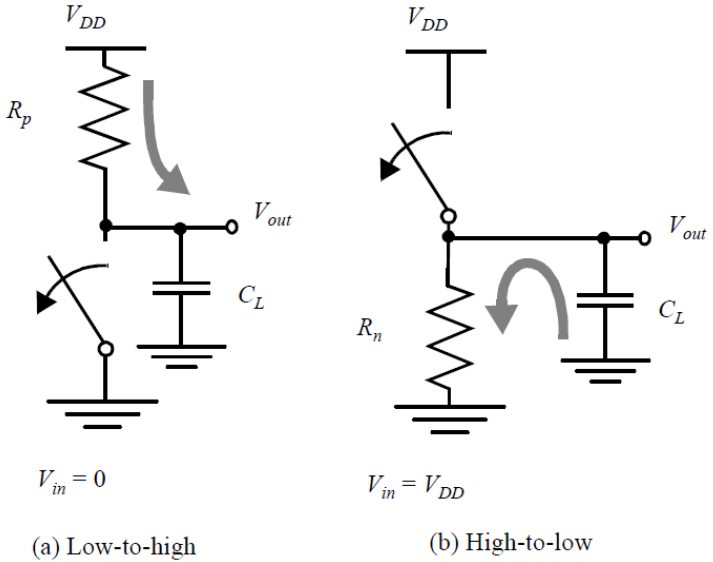


Figure 5.6 Switch model of dynamic behavior of static CMOS inverter.

Qualitative analysis of the transient behavior

- For a low-to-high transition, the gate response time is simply determined by the time it takes to charge the capacitor C_L through the resistor R_p.
- The time constant is $R_p C_L$.
- **Hence, a fast gate is built either by keeping the output capacitance small or by decreasing the on-resistance of the transistor.**
- The latter is achieved by increasing the W/L ratio of the device.
- Similar considerations are valid for the high-to-low transition, which is dominated by the $R_n C_L$ time-constant.

The Static Behavior: Switching Threshold

- The switching threshold, V_M, is defined as the point where $V_{in} = V_{out}$.
- Can be obtained graphically from the intersection of the VTC with the line $V_{in} = V_{out}$
- Both PMOS and NMOS are always saturated, since $V_{DS} = V_{GS}$.
- By equating the currents through the transistors, assuming the MOS are velocity-saturated, i.e., $V_{DSAT} < V_M - V_T$

$$k_n V_{DSATn}\left(V_M - V_{Tn} - \frac{V_{DSATn}}{2}\right) + k_p V_{DSATp}\left(V_M - V_{DD} - V_{Tp} - \frac{V_{DSATp}}{2}\right) = 0$$

$$V_M = \frac{\left(V_{Tn} + \frac{V_{DSATn}}{2}\right) + r\left(V_{DD} + V_{Tp} + \frac{V_{DSATp}}{2}\right)}{1 + r} \quad \text{with} \quad r = \frac{k_p V_{DSATp}}{k_n V_{DSATn}} = \frac{\upsilon_{satp} W_p}{\upsilon_{satn} W_n}$$

- For large values of V_{DD} : $V_M \approx \frac{r V_{DD}}{1 + r}$

The Static Behavior: Switching Threshold

$$V_M \approx \frac{r V_{DD}}{1 + r}$$

- It states that the switching threshold is set by the ratio *r*, which compares the relative driving strengths of the PMOS and NMOS transistors.
- It is generally desirable for V_M to be located around $V_{DD}/2$.
- This requires *r* to be approximately 1, $r \approx 1$,

$$(W/L)_p = (W/L)_n \times (V_{DSATn} k'_n)/(V_{DSATp} k'_p).$$

- To move V_M upwards, a larger value of *r* is required, which means PMOS wider.
- For a desired value V_M

$$\frac{(W/L)_p}{(W/L)_n} = \frac{k'_n V_{DSATn}(V_M - V_{Tn} - V_{DSATn}/2)}{k'_p V_{DSATp}(V_{DD} - V_M + V_{Tp} + V_{DSATp}/2)}$$

The Static Behavior: Switching Threshold

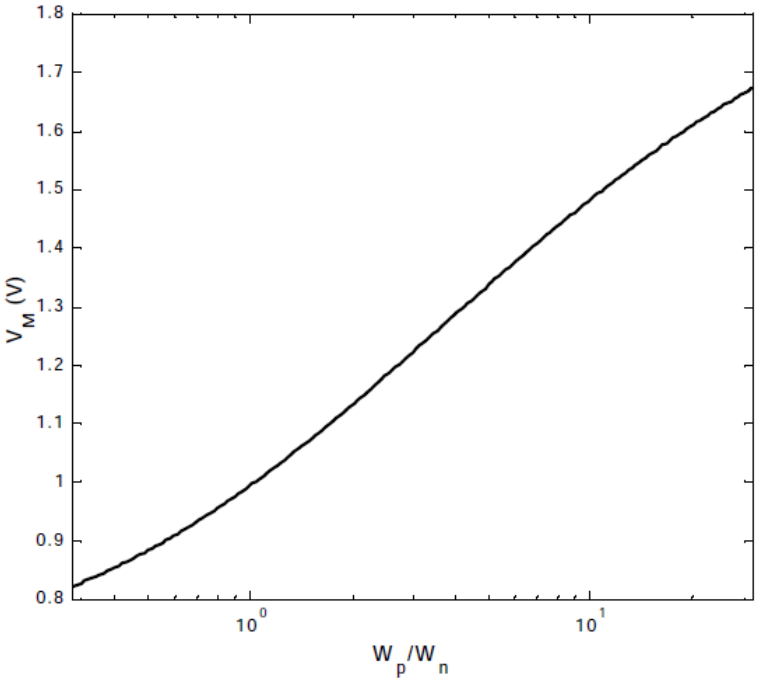


Figure 5.7 Simulated inverter switching threshold versus PMOS/NMOS ratio (0.25 μm CMOS, V_{DD} = 2.5 V)

The Static Behavior: Noise Margins

- By definition, V_{IH} and V_{IL} are the operational points of the inverter where $\frac{dV_{out}}{dV_{in}} = -1$
- We perform an approximate analysis.
- We use a piece wise linear approximation for the VTC.
- The transition region is approximated by a straight line, the gain of which equals the gain g at the switching threshold V_M.

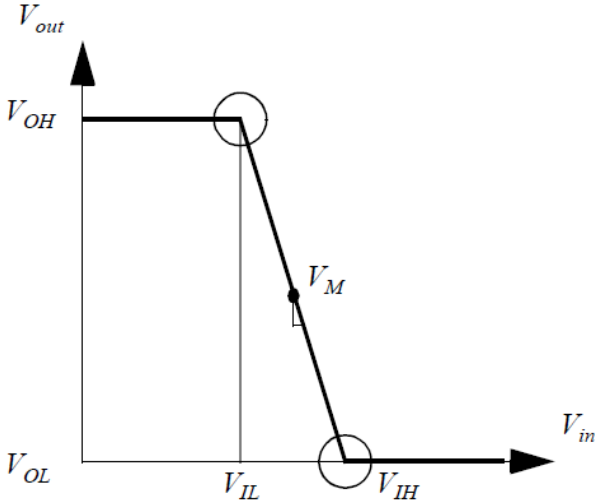


Figure 5.9 A piece-wise linear approximation of the VTC simplifies the derivation of V_{IL} and V_{IH}.

The Static Behavior: Noise Margins

- This approach yields the following expressions:

$$V_{IH} - V_{IL} = -\frac{(V_{OH} - V_{OL})}{g} = \frac{-V_{DD}}{g}$$

$$V_{IH} = V_M - \frac{V_M}{g} \qquad V_{IL} = V_M + \frac{V_{DD} - V_M}{g}$$

$$NM_H = V_{DD} - V_{IH} \qquad NM_L = V_{IL}$$

- Remains us to determine *g,* assuming both PMOS and NMOS are velocity-saturated.
- The channel-length modulation factor cannot be ignored in this analysis.
- The gain can be derived by differentiating the current equation

$$k_n V_{DSATn}\left(V_{in} - V_{Tn} - \frac{V_{DSATn}}{2}\right)(1 + \lambda_n V_{out}) +$$

$$k_p V_{DSATp}\left(V_{in} - V_{DD} - V_{Tp} - \frac{V_{DSATp}}{2}\right)(1 + \lambda_p V_{out} - \lambda_p V_{DD}) = 0$$

The Static Behavior: Noise Margins

$$\frac{dV_{out}}{dV_{in}} = -\frac{k_n V_{DSATn}(1+\lambda_n V_{out}) + k_p V_{DSATp}(1+\lambda_p V_{out} - \lambda_p V_{DD})}{\lambda_n k_n V_{DSATn}(V_{in} - V_{Tn} - V_{DSATn}/2) + \lambda_p k_p V_{DSATp}(V_{in} - V_{DD} - V_{Tp} - V_{DSATp}/2)}$$

- Ignoring some second-order terms, and setting $V_{in} = V_M$ results in

$$g = -\frac{1}{I_D(V_M)} \frac{k_n V_{DSATn} + k_p V_{DSATp}}{\lambda_n - \lambda_p}$$

$$\approx \frac{1+r}{(V_M - V_{Tn} - V_{DSATn}/2)(\lambda_n - \lambda_p)}$$

- The gain is almost purely determined by technology parameters, especially the channel length modulation.
- It can only in a minor way be influenced by the designer through the choice of supply and switching threshold voltages and form factors.

Device Variations

- The actual operating temperature might vary over a large range, and the device parameters after fabrication deviate from the nominal values of the design.
- Fortunately, the dc-characteristics of the static CMOS inverter are rather insensitive to these variations, and the gate remains functional over a wide range of operating conditions.

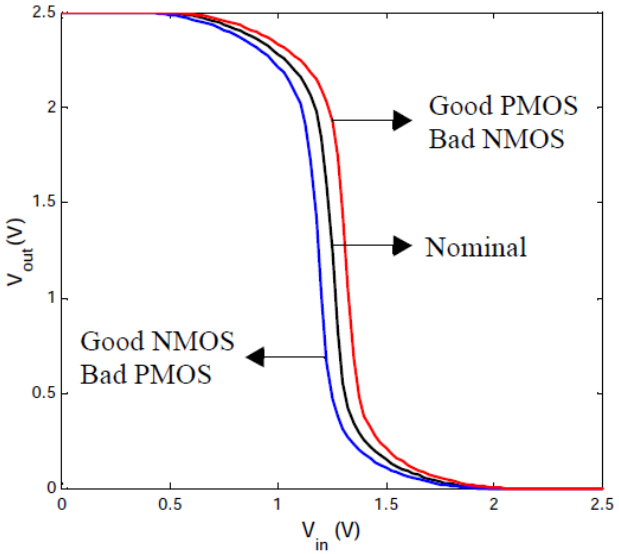
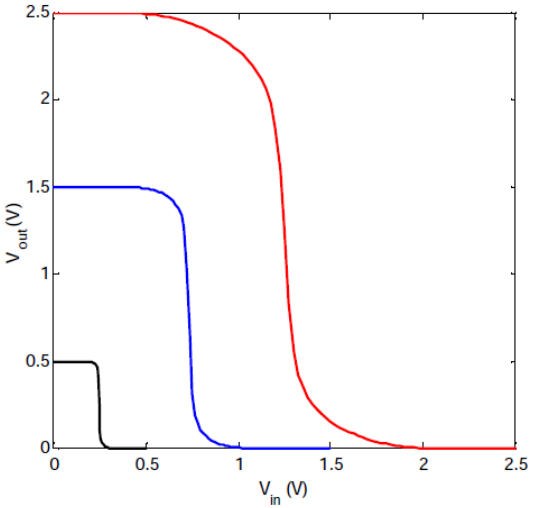


Figure 5.11 Impact of device variations on static CMOS inverter VTC. The "good" device has a smaller oxide thickness (- 3nm), a smaller length (-25 nm), a higher width (+30 nm), and a smaller threshold (-60 mV). The opposite is true for the "bad" transistor.

Scaling the Supply Voltage

- In the years, the supply voltages have reduced at rates similar to the device dimensions. At the same time, device threshold voltages are virtually kept constant.
- What is the impact of this trend on the integrity parameters of the CMOS inverter?
- The equation of the gain *g* of the inverter in the transition region actually increases with a reduction of the supply voltage!



(a) Reducing V_{DD} improves the gain...

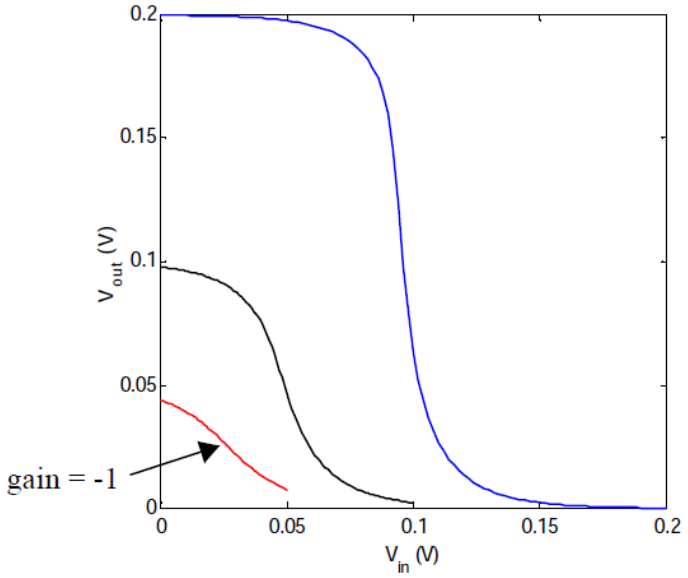
Scaling the Supply Voltage

- Why do we not operate all our digital circuits at these low supply voltages?

- We will learn that reducing the supply voltage indiscriminately has a positive impact on the energy dissipation, but is absolutely detrimental to the performance on the gate.
- The dc-characteristic becomes increasingly sensitive to variations in the device parameters such as the transistor threshold.
- Scaling the supply voltage means reducing the signal swing. While this typically helps to reduce the internal noise in the system (such as caused by crosstalk), it makes the design more sensitive to external noise sources.

Scaling the Supply Voltage

- Amazingly enough, we still obtain an inverter characteristic when the supply voltage is not even large enough to turn the transistors on!
- The sub-threshold currents are sufficient to switch the gate between low and high levels, and provide enough gain to produce acceptable VTCs.



The Dynamic Behavior

- The qualitative analysis concluded that the propagation delay of the CMOS inverter is determined by the time it takes to charge and discharge the load capacitor C_L through the PMOS and NMOS transistors, respectively.
- This observation suggests that **getting C_L as small as possible is crucial to the realization of high-performance CMOS circuits.**
- To make the analysis tractable, we assume that all capacitances are lumped together into one single capacitor C_L, located between V_{out} and *GND*.

The Dynamic Behavior

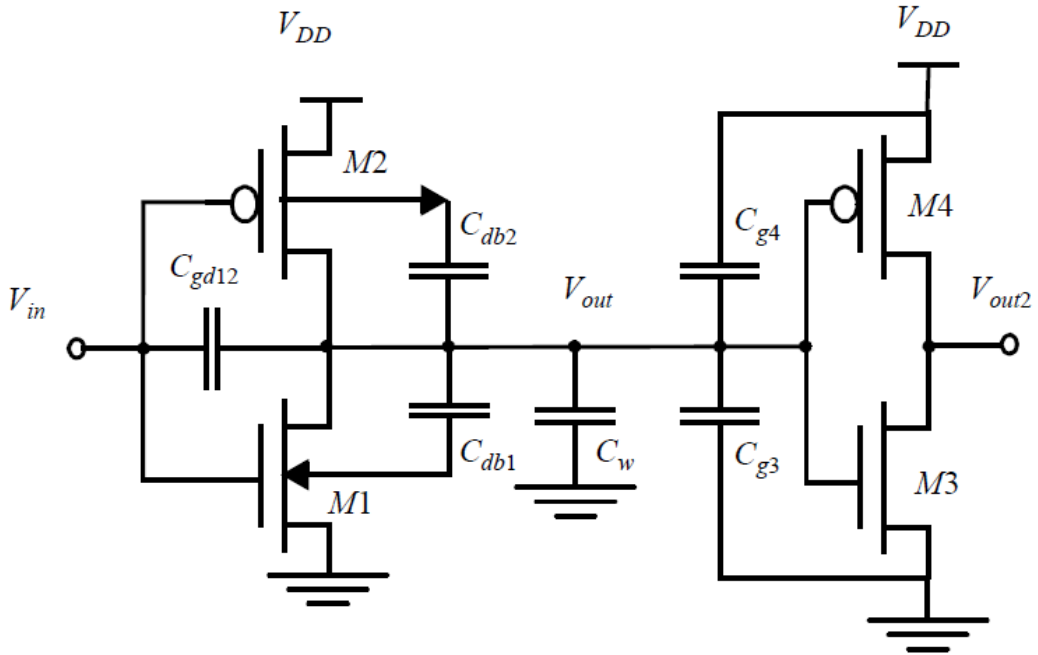


Figure 5.13 Parasitic capacitances, influencing the transient behavior of the cascaded inverter pair.

Gate-Drain Capacitance C_{gd12}

- M1 and M2 are either in cut-off or in the saturation mode during the first half (up to 50% point) of the output transient. Under these circumstances, the only contributions to C_{gd12} are the overlap capacitances of both M1 and M2.
- The lumped capacitor model now requires that this floating gate-drain capacitor be replaced by a capacitance-to-ground.
- This is accomplished by taking the so-called **Miller effect** into account.
- During a low-high or high-low transition, the terminals of the gate-drain capacitor are moving in opposite directions. The voltage change over the floating capacitor is hence twice the actual output voltage swing.
- To present an identical load to the output node, the capacitance-to-ground must have a value that is twice as large as the floating capacitance.

Gate-Drain Capacitance C_{gd12}

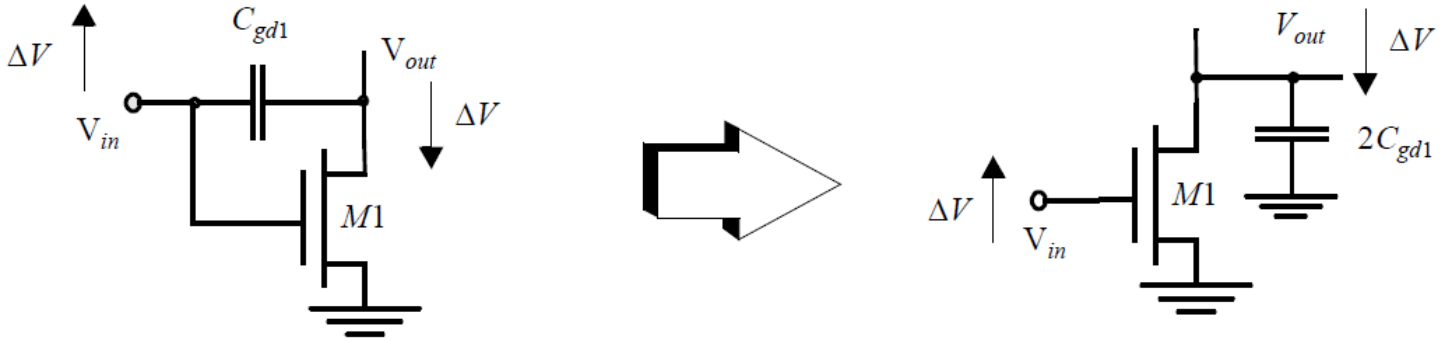


Figure 5.14 The Miller effect—A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground, whose value is two times the original value.

Other capacitances

- Junction capacitances C_{db1} and C_{db2}
- Wiring Capacitance C_w
- Gate Capacitance of Fanout C_{g3} and C_{g4}

$$
\begin{aligned}
C_{fanout} &= C_{gate}(\text{NMOS}) + C_{gate}(\text{PMOS}) \\
&= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox})
\end{aligned}
$$

Capacitances

Table 5.2 Components of C_L (for high-to-low and low-to-high transitions).

Capacitor	Expression	Value (fF) (H→L)	Value (fF) (L→H)
C_{gd1}	$2\ CGD0_n\ W_n$	0.23	0.23
C_{gd2}	$2\ CGD0_p\ W_p$	0.61	0.61
C_{db1}	$K_{eqn}\ AD_n\ CJ + K_{eqswn}\ PD_n\ CJSW$	0.66	0.90
C_{db2}	$K_{eqp}\ AD_p\ CJ + K_{eqswp}\ PD_p\ CJSW)$	1.5	1.15
C_{g3}	$(CGD0_n+CGSO_n)\ W_n + C_{ox}\ W_n\ L_n$	0.76	0.76
C_{g4}	$(CGD0_p+CGSO_p)\ W_p + C_{ox}\ W_p\ L_p$	2.28	2.28
C_w	From Extraction	0.12	0.12
C_L	Σ	6.1	6.0

Propagation Delay: First-Order Analysis

- One way to compute the propagation delay of the inverter is to integrate the capacitor (dis)charge current.

$$t_p = \int_{v_1}^{v_2} \frac{C_L(v)}{i(v)} dv$$

- We derive an approximation of the propagation delay with the switch-model.
- The voltage-dependencies of the on-resistance and the load capacitor are addressed by replacing both by a constant linear element with a value averaged over the interval of interest.

Propagation Delay: First-Order Analysis

$$R_{eq} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)$$

$$\text{with} \quad I_{DSAT} = k'\frac{W}{L}\left((V_{DD}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right)$$

- Deriving the propagation delay of the resulting circuit is now straightforward, and is nothing more than the analysis of a first-order linear *RC*-network.

$$t_{pHL} = \ln(2)R_{eqn}C_L = 0.69R_{eqn}C_L$$

$$t_{pLH} = 0.69R_{eqp}C_L$$

$$t_p = \frac{t_{pHL}+t_{pLH}}{2} = 0.69C_L\left(\frac{R_{eqn}+R_{eqp}}{2}\right)$$

Propagation Delay: First-Order Analysis

- How can we manipulate and/or optimize the delay of a gate?
- It is necessary to make the parameters governing the delay explicit by expanding R_{eq} in the delay equation.
- Assuming that the channel-length modulation factor λ is ignorable

$$t_{pHL} = 0.69\frac{3}{4}\frac{C_L V_{DD}}{I_{DSATn}} = 0.52\frac{C_L V_{DD}}{(W/L)_n k'_n V_{DSATn}(V_{DD} - V_{Tn} - V_{DSATn}/2)}$$

- In the majority of designs, $V_{DD} >> V_{Tn} + V_{DSATn}/2$ and the delay becomes virtually independent of V_{DD}

$$t_{pHL} \approx 0.52\frac{C_L}{(W/L)_n k'_n V_{DSATn}}$$

- This is a first-order approximation: increasing V_{DD} yields an observable, albeit small, improvement due to a non-zero λ

Propagation Delay: First-Order Analysis

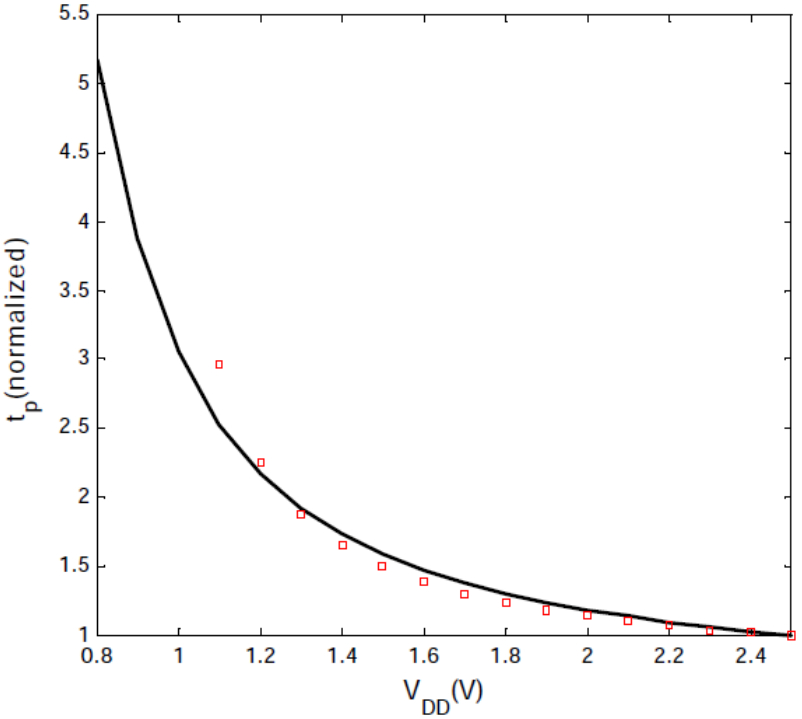


Figure 5.17 Propagation delay of CMOS inverter as a function of supply voltage (normalized with respect to the delay at 2.5 V). The dots indicate the delay values predicted by Eq. (5.21). Observe that this equation is only valid when the devices are velocity-saturated. Hence, the deviation at low supply voltages.

Design techniques

The propagation delay of a gate can be minimized in the following ways:

- ***Reduce C_L*** Three major factors: the internal diffusion capacitance, the interconnect capacitance, and the fanout. *Good design practice requires keeping the drain diffusion areas as small as possible.*
- ***Increase the W/L ratio of the transistors*.** Most powerful and effective tool, but proceed with caution. Increasing the transistor size also raises the diffusion capacitance and hence C_L. Once the intrinsic capacitance starts to dominate the extrinsic load increasing the gate size does not longer help: *"self-loading".*
  In addition, wide transistors have a larger gate capacitance, which increases the fan-out factor of the driving gate.
- ***Increase V_{DD}*.** The delay of a gate can be modulated by modifying the supply voltage.

Sizing Inverters for Performance

- We assume a symmetrical inverter, i.e., PMOS and NMOS sized such that the rise and fall delays are identical.
- The load capacitance of the inverter can be divided into $C_L = C_{int} + C_{ext}$
- C_{int} represents the self-loading or intrinsic output capacitance of the inverter, associated with diffusion capacitances and gate-drain overlap (Miller) capacitances.
- C_{ext} is the extrinsic load capacitance, attributable to fanout and wiring capacitance.
- Considering R_{eq} the equivalent output resistance of the gate:

$$t_p = 0.69 R_{eq}(C_{int} + C_{ext})$$
$$= 0.69 R_{eq} C_{int}(1 + C_{ext}/C_{int}) = t_{p0}(1 + C_{ext}/C_{int})$$

- $t_{p0} = 0.69\ R_{eq} C_{int}$ the delay of the inverter only loaded by its own intrinsic capacitance, called the *intrinsic or unloaded delay*.

Sizing Inverters for Performance

- How transistor sizing impacts the performance of the gate ?
- Let us denote with R_{ref} and C_{iref} the resistance and the intrinsic capacitance of a minimum size inverter and with S the sizing factor:

$$C_{int} = SC_{iref} \qquad R_{eq} = R_{ref}/S$$

$$t_p = 0.69(R_{ref}/S)(SC_{iref})(1 + C_{ext}/(SC_{iref}))$$

$$= 0.69R_{ref}C_{iref}\left(1 + \frac{C_{ext}}{SC_{iref}}\right) = t_{p0}\left(1 + \frac{C_{ext}}{SC_{iref}}\right)$$

- The intrinsic delay of the inverter t_{p0} is independent of the sizing of the gate and is purely determined by technology and inverter layout.
- Making S infinitely large yields the maximum obtainable performance gain, eliminating the impact of any external load, and reducing the delay to the intrinsic one.

Sizing A Chain of Inverters

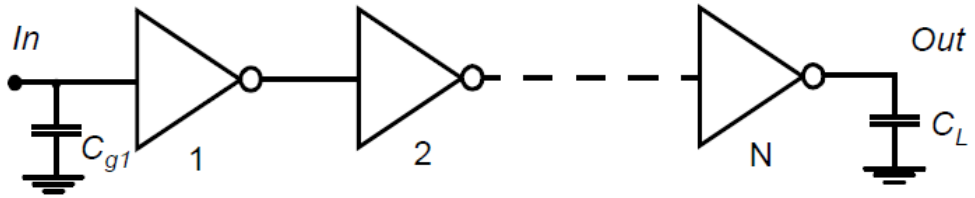


Figure 5.20 Chain of N inverters with fixed input and output capacitance.

- Let us establish the relationship between the input gate capacitance C_g and the intrinsic output capacitance of the inverter. Both are proportional to the gate sizing.

$$C_{int} = \gamma C_g$$

- With γ a proportionality factor, which is only function of technology.

$$t_p = t_{p0}\left(1 + \frac{C_{ext}}{\gamma C_g}\right) = t_{p0}(1 + f/\gamma)$$

- It is only function of the ratio between the external load capacitance and input capacitance, called the *effective fanout f*.

Sizing A Chain of Inverters

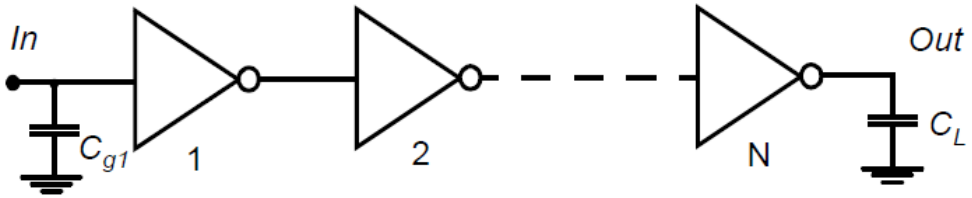


Figure 5.20 Chain of N inverters with fixed input and output capacitance.

- Out goal is to minimize the delay through the inverter chain, with the input capacitance of the first inverter C_{g1} (typically a minimally-sized device) and the load capacitance C_L fixed.
- Given the delay expression for the j-th inverter stage,

$$t_{p,j} = t_{p0}\left(1 + \frac{C_{g,j+1}}{\gamma C_{g,j}}\right) = t_{p0}(1 + f_j/\gamma)$$

- we can derive the total delay of the chain

Sizing A Chain of Inverters

$$t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0} \sum_{j=1}^{N} \left(1 + \frac{C_{g,j+1}}{\gamma C_{g,j}}\right), \text{ with } C_{g,N+1} = C_L$$

- This equation has N-1 unknowns, $C_{g,2}, C_{g,3}, \ldots, C_{g,N}$.
- The minimum delay can be found by imposing the N-1 partial derivatives to be zero.
- $\partial t_p / \partial C_{g,j} = 0$ results in the set of constraints $C_{g,j+1}/C_{g,j} = C_{g,j}/C_{g,j-1}$
- the optimum size of each inverter is the geometric mean of its neighbors' sizes

$$C_{g,j} = \sqrt{C_{g,j-1} C_{g,j+1}}.$$

- Each inverter is sized up by the same factor f with respect to the preceding gate, has the same effective fanout ($f_j = f$), and hence the same delay.

Sizing A Chain of Inverters

- With $C_{g,1}$ and C_L given, we can derive the sizing factor,

$$f = \sqrt[N]{C_L / C_{g,1}} = \sqrt[N]{F}$$

- and the minimum delay through the chain,

$$t_p = N t_{p0}(1 + \sqrt[N]{F}/\gamma) .$$

- F represents the *overall effective fanout* of the circuit, and equals $C_L/C_{g,1}$.
- The relationship between t_p and F is a function of the number of stages N.
- How to choose N ?
- When N is too large, the first component of the equation, representing the intrinsic delay of the stages, becomes dominant.
- If N is too small, the effective fanout of each stage becomes large, and the second component is dominant.

Sizing A Chain of Inverters

- Differentiating the minimum delay expression by *N,*

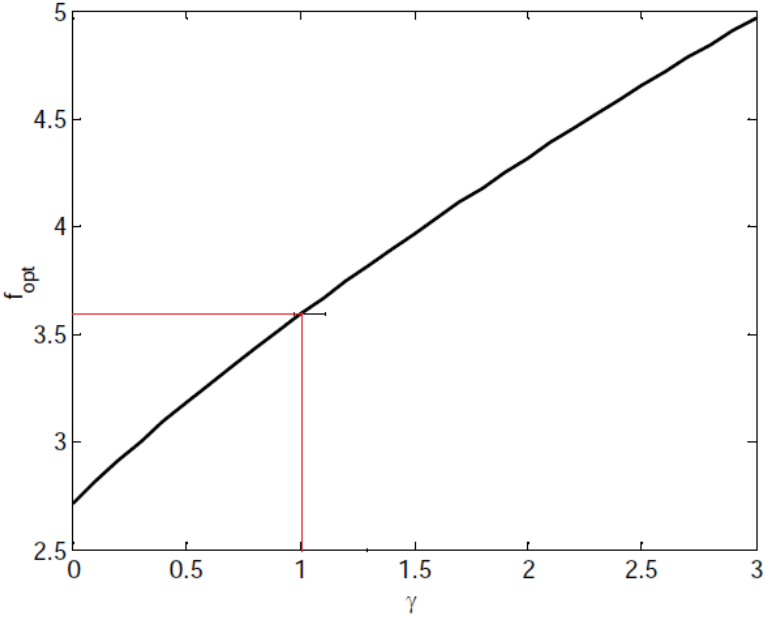
$$\gamma + \sqrt[N]{F} - \frac{\sqrt[N]{F}\ln F}{N} = 0$$

or equivalently

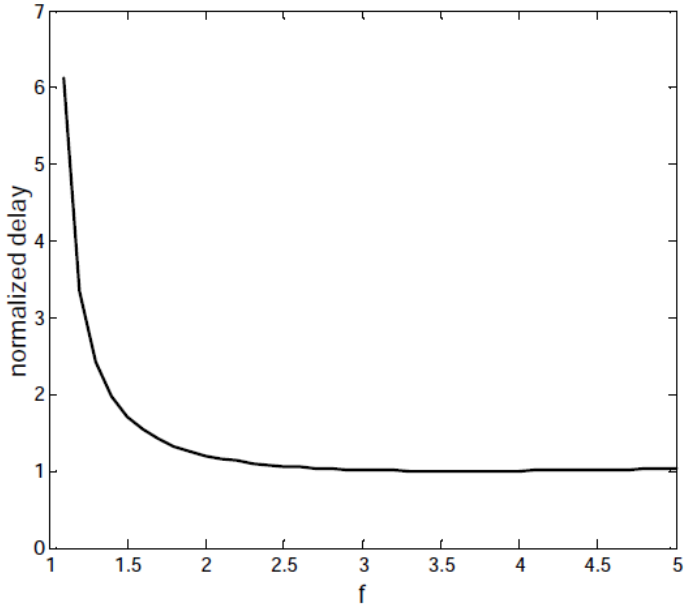
$$f = e^{(1+\gamma/f)}$$

- This equation only has a closed-form solution for $\gamma = 0$, i.e., when the self-loading is ignored and the load capacitance only consists of the fanout.
- In that case $N = \ln(F)$ and $f = 2.71828 = e$.
- When self-loading is included, the equation can only be solved numerically…

Sizing A Chain of Inverters



(a) Optimum effective fanout f (or inverter scaling factor) as a function of the self-loading factor γ in an inverter chain.



(b) Normalized propagation delay ($t_p/(t_{popt})$ as a function of the effective fanout f for γ=1.

Sizing A Chain of Inverters

Table 5.3 t_{opt}/t_{p0} versus x for various driver configurations.

F	Unbuffered	Two Stage	Inverter Chain
10	11	8.3	8.3
100	101	22	16.5
1000	1001	65	24.8
10,000	10,001	202	33.1

Delay in the Presence of (Long) Interconnect Wires

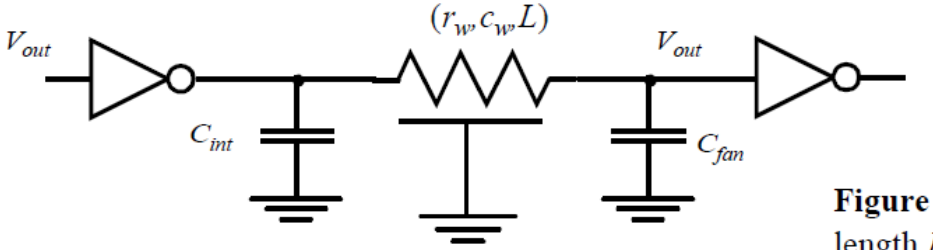


Figure 5.24 Inverter driving single fanout through wire of length L.

- The driver is represented by a single resistance R_{dr} (average between R_{eqn} and R_{eqp}).
- C_{int} and C_{fan} account for the intrinsic capacitance of the driver, and the input capacitance of the fanout gate, respectively.
- t_p can be obtained by applying the Elmore delay expression:

$$t_p = 0.69R_{dr}C_{int} + (0.69R_{dr} + 0.38R_w)C_w + 0.69(R_{dr} + R_w)C_{fan}$$

$$= 0.69R_{dr}(C_{int} + C_{fan}) + 0.69(R_{dr}c_w + r_w C_{fan})L + 0.38r_w c_w L^2$$

- (The 0.38 factor accounts for the fact that the wire represents a distributed delay.)

Dynamic Power Consumption

- Each time the capacitor C_L gets charged through the PMOS transistor, its voltage rises from 0 to V_{DD}, and a certain amount of energy is drawn from the power supply.
- Part of this energy is dissipated in the PMOS device, while the remainder is stored on the load capacitor.
- During the high-to-low transition, this capacitor is discharged, and the stored energy is dissipated in the NMOS transistor.
- A precise measure for this energy consumption can be derived.
- We assume, initially, that the input waveform has zero rise and fall times, i.e, the NMOS and PMOS devices are never on simultaneously.

Dynamic Power Consumption

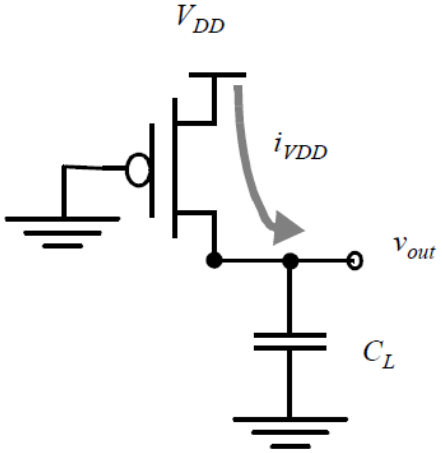


Figure 5.25 Equivalent circuit during the low-to-high transition.

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

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

Dynamic Power Consumption

- These results can also be derived by observing that during the low-to-high transition, C_L is loaded with a charge $C_L V_{DD}$.
- Providing this charge requires an energy from the supply equal to $C_L V_{DD}^2$ (= Q V_{DD})
- The energy stored on the capacitor equals $C_L V_{DD}^2/2$.
- Only half of the energy supplied by the power source is stored on C_L.
- The other half has been dissipated by the PMOS transistor.
- ***This energy dissipation is independent of the size (and hence the resistance) of the PMOS device!***
- During the discharge phase, the charge is removed from the capacitor, and its energy is dissipated in the NMOS device.
- ***Once again, there is no dependence on the size of the device*.**
- In summary, each switching cycle (consisting of an L→H and an H → L transition) takes a fixed amount of energy, equal to $C_L V_{DD}^2$

Dynamic Power Consumption

- If the gate is switched **on and off** $f_{0\to1}$ times per second, the power consumption is

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1}$$

- Advances in technology result in ever-higher of values of $f_{0\to1}$ (as t_p decreases).
- At the same time, the total capacitance on the chip (C_L) increases as more and more gates are placed on a single die.
- Consider for instance a 0.25 μm CMOS chip with a clock rate of 500 Mhz and an average load capacitance of 15 fF/gate, assuming a fanout of 4.
- The power consumption per gate for a 2.5 V supply equals approximately 50 μW.
- For a design with *1 million gates* and assuming that a transition occurs at every clock edge, *this would result in a power consumption of 50 W!*
- This evaluation presents, fortunately, a pessimistic perspective.
- In reality, not all gates in the complete IC switch at the full rate of 500 Mhz.
- **The actual activity in the circuit is substantially lower.**

Dynamic Power Consumption

- Computing the dissipation of a complex circuit is complicated by the $f_{0\to1}$ factor, also called the *switching activity.*
- The switching activity is function of the nature and the statistics of the input signals: **If the input signals remain unchanged**, no switching happens, and the dynamic **power consumption is zero!**
- But, rapidly changing signals provoke plenty of switching and hence dissipation.
- Other factors influencing the activity are the overall network topology and the function to be implemented.

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} = C_L V_{DD}^2 P_{0\to1} f = C_{EFF} V_{DD}^2 f$$

- where *f* now presents the maximum input rate (often the clock rate) and $P_{0\to1}$ the probability that a clock event results in a 0→1 event at the output of the gate.
- $C_{EFF} = P_{0\to1} C_L$ is called the *effective capacitance* and represents the average capacitance switched every clock cycle.

Low Energy/Power Design Techniques

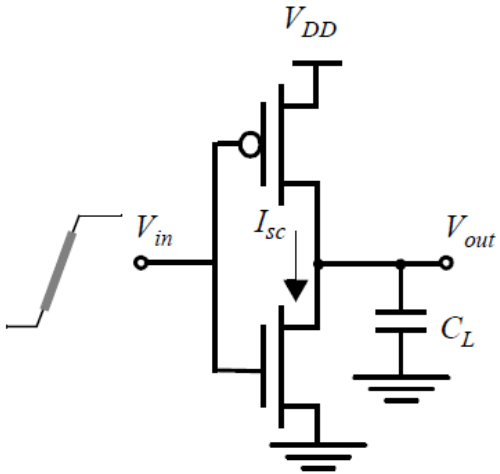
- With the increasing complexity of the digital integrated circuits, it is anticipated that the power problem will only worsen in future technologies.
- Lower supply are more and more attractive.
- **Reducing V_{DD} has a quadratic effect on P_{dyn} .**
- This assumes that the same clock rate can be sustained.
- This assumption is not that unrealistic as long as the supply voltage is substantially higher than the threshold voltage.
- An important performance penalty occurs once V_{DD} approaches 2 V_T.
- When a lower bound on the supply voltage is set by external constraints or when the performance degradation due to lowering the supply voltage is intolerable, the only means of reducing the dissipation is by lowering the effective capacitance.
- This can be achieved by addressing both of its components: the *physical capacitance* and the *switching activity*.

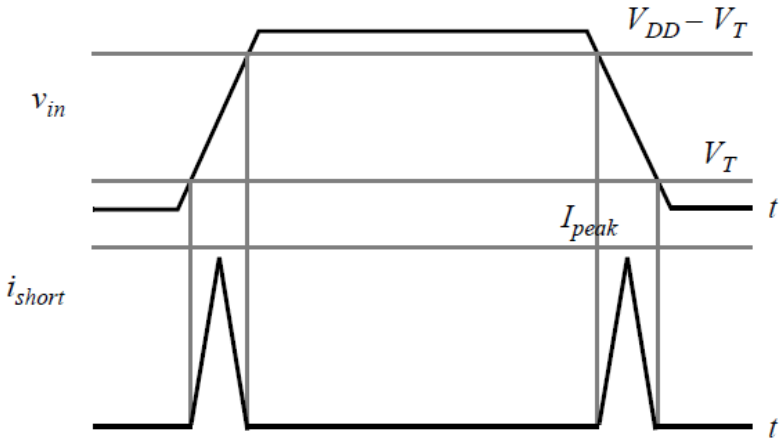
Low Energy/Power Design Techniques

- A research topic is devoted to study how to reduce the *switching activity*.
- *Lowering the physical capacitance* is an overall worthwhile goal, which also helps to improve the performance of the circuit.
- As most of the capacitance is due to transistor capacitances (gate and diffusion), it makes sense to keep those to a minimum when designing for low power.
- This means that transistors should be kept to *minimal size* whenever possible or reasonable.
- The only instances where transistors should be sized up is when the load capacitance is dominated by extrinsic capacitances (such as fan-out or wiring capacitance).

Dissipation Due to Direct-Path Currents

- The assumption of the zero rise and fall times of the input wave forms is incorrect.
- The finite slope of the input signal causes a direct current path between V_{DD} and *GND* for a short period of time during switching, while the NMOS and the PMOS transistors are conducting simultaneously.





Dissipation Due to Direct-Path Currents

- Under the assumption that the current spikes can be approximated as triangles and that the inverter is symmetrical in its rising and falling responses, we can compute the energy consumed per switching period and the power consumption

$$E_{dp} = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak}$$

$$P_{dp} = t_{sc}V_{DD}I_{peak}f = C_{sc}V_{DD}^2 f$$

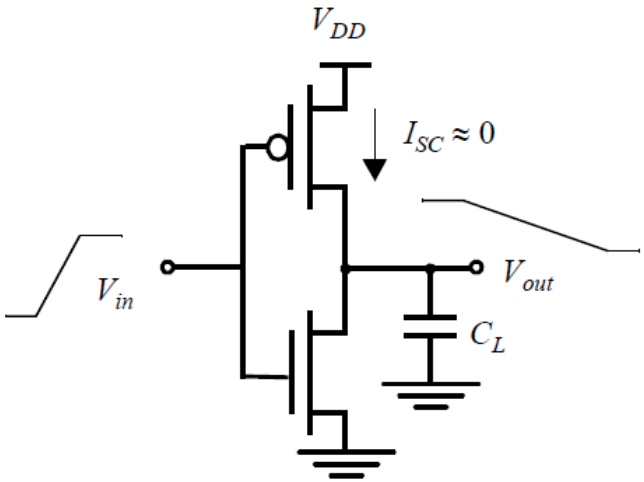
- For a linear input slope, the time both devices are conducting t_{sc} is approximated by

$$t_{sc} = \frac{V_{DD} - 2V_T}{V_{DD}}t_s \approx \frac{V_{DD} - 2V_T}{V_{DD}} \times \frac{t_{r(f)}}{0.8}$$

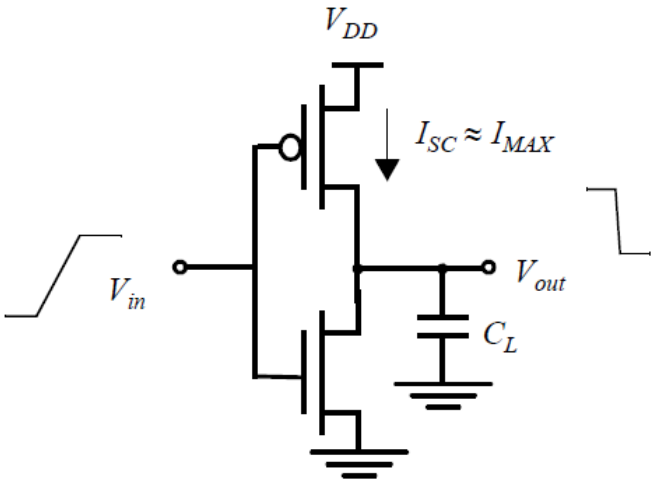
- where t_s represents the 0-100% transition time.
- I_{peak} is determined by the saturation current of the devices and is hence directly proportional to the sizes of the transistors.

Dissipation Due to Direct-Path Currents

- The peak current is also **a strong function of the ratio between input and output slopes**.



(a) Large capacitive load



(b) Small capacitive load

Dissipation Due to Direct-Path Currents

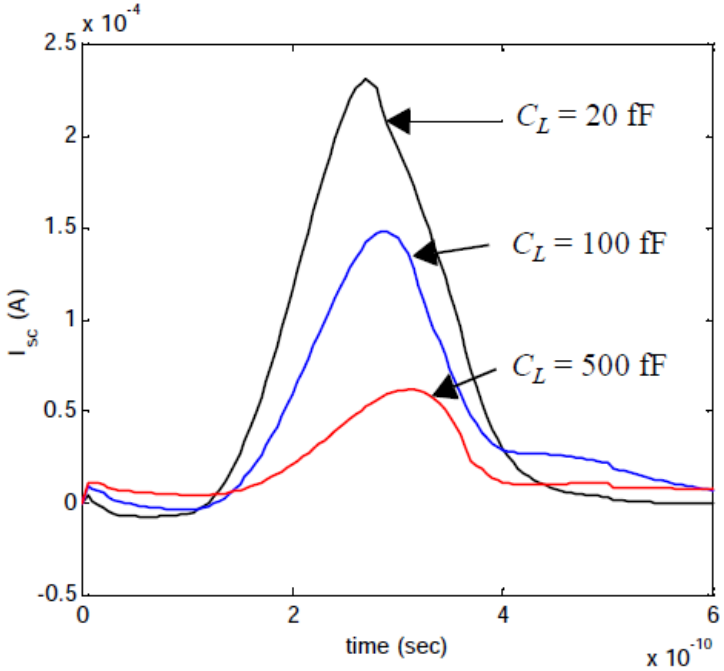


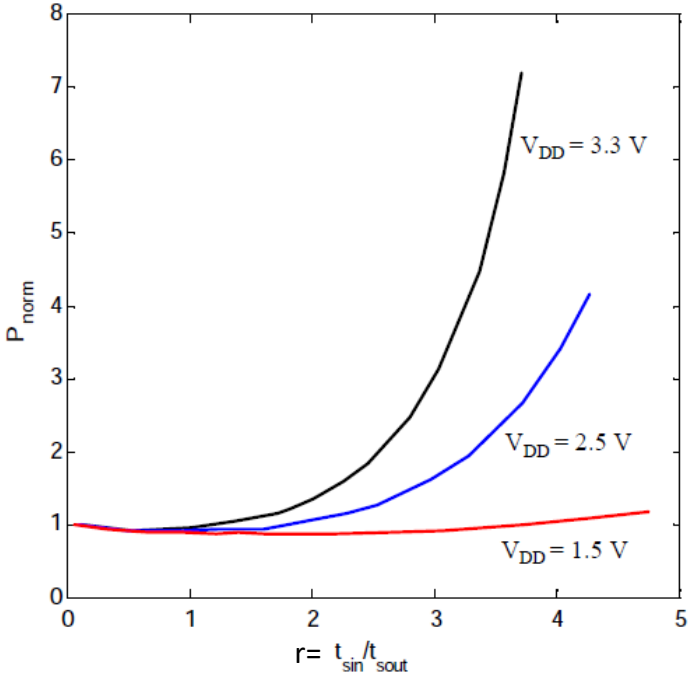
Figure 5.32 CMOS inverter short-circuit current through NMOS transistor as a function of the load capacitance (for a fixed input slope of 500 psec).

Dissipation Due to Direct-Path Currents

- This analysis leads to the conclusion that the short-circuit dissipation is minimized by making the output rise/fall time larger than the input rise/fall time.
- On the other hand, making the output rise/fall time too large slows down the circuit and can cause short-circuit currents in the fan-out gates.
- A more practical rule, which optimizes the power consumption in a global way:

  The power dissipation due to short-circuit currents is minimized by matching the rise/fall times of the input and output signals. At the overall circuit level, this means that rise/fall times of all signals should be kept constant within a range.

Dissipation Due to Direct-Path Currents



$W/L|_P = 1.125\ \mu m/0.25\ \mu m$

$W/L|_N = 0.375\ \mu m/0.25\ \mu m$

$C_L = 30$ fF

Figure 5.33 Power dissipation of a static CMOS inverter as a function of the ratio between input and output rise/fall times. The power is normalized with respect to zero input rise-time dissipation. At low values of the slope ratio, input-output coupling leads to some extra dissipation.

Dissipation Due to Direct-Path Currents

- When the load capacitance is too small for a given inverter size (r > 2-3 for V_{DD} = 2.5-3.3 V), the power is dominated by the short-circuit current.
- For very large capacitance values, all power dissipation is devoted to charging and discharging the load capacitance.
- When the rise/fall times of inputs and outputs are equalized, most power dissipation is associated with the dynamic power and only a minor fraction (< 10%) is devoted to short-circuit currents.
- Observe also that the impact of **short-circuit current is reduced when we lower the supply voltage.**

Static Consumption

- The static (or steady-state) power dissipation of a circuit is expressed by

$$P_{stat} = I_{stat} V_{DD}$$

- Where I_{stat} is the “static” current.
- Ideally, the static current of the CMOS inverter is equal to zero.
- There is, unfortunately, a *leakage current* flowing through the reverse-biased diode junctions of the transistors.
- This contribution is, in general, *very small and can be ignored*.
- However, be aware that the junction leakage currents are caused by thermally generated carriers. Their value increases with increasing junction temperature, and this occurs in an exponential fashion.
- An emerging source of leakage current is the *subthreshold current* of the transistors.

Static Consumption

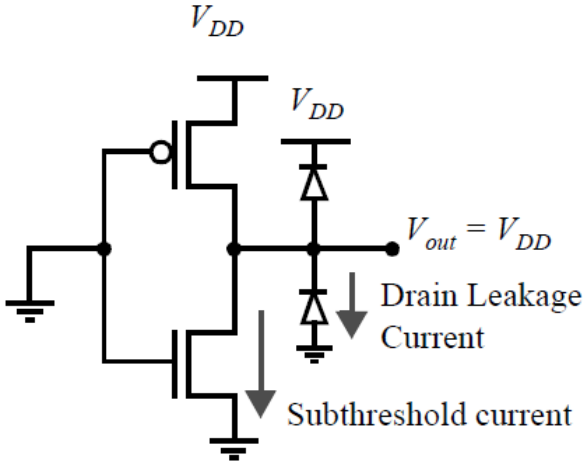


Figure 5.34 Sources of leakage currents in CMOS inverter (for $V_{in} = 0$ V).

- An MOS can experience a drain-source current, even when $V_{GS} < V_T$
- The closer V_T is to zero volts, the larger the leakage current at $V_{GS} = 0$ V and the larger the static power consumption.
- To offset this effect, V_T has generally been kept high enough.
- This approach is being challenged by the reduction in supply voltages.

Static Consumption

- Scaling V_{DD} while keeping V_T constant results in an important loss in performance, especially when V_{DD} approaches 2 V_T.
- One approach to address this performance issue is to scale the device thresholds down as well.
- Unfortunately, the threshold voltages are lower-bounded by the amount of allowable subthreshold leakage current.
- The choice of the threshold voltage hence represents a trade-off between performance and static power dissipation.
- The continued scaling of V_{DD} forces V_T ever downwards and makes subthreshold conduction a dominant source of power dissipation.
- Process technologies that contain devices with sharper turn-off characteristic will therefore become more attractive.
- An example of the latter is the SOI (Silicon-on-Insulator) technology whose MOS transistors have slope-factors that are close to the ideal 60 mV/decade.

The total power consumption and power delay product

$$P_{tot} = P_{dyn} + P_{dp} + P_{stat} = (C_L V_{DD}^2 + V_{DD} I_{peak} t_s) f_{0\to 1} + V_{DD} I_{leak}$$

- In typical CMOS circuits, the capacitive dissipation is by far the dominant factor.
- The *power-delay product* (PDP) is a quality measure of the gate.

$$PDP = P_{av} t_p$$

- The PDP presents a measure of energy, as is apparent from the units (W s = J).
- Assuming that the gate is switched at its maximum possible rate of $f_{max} = 1/(2t_p)$, and ignoring the contributions of the static and direct-path currents to the power consumption,

$$PDP = C_L V_{DD}^2 f_{max} t_p = \frac{C_L V_{DD}^2}{2}$$

- The PDP stands for the **average energy consumed per switching event**

Energy-Delay Product

- A more relevant metric that combines a measure of performance and energy is the **energy-delay product** (EDP)

$$EDP = PDP \times t_p = P_{av} t_p^2 = \frac{C_L V_{DD}^2}{2} t_p$$

- It is worth analyzing the voltage dependence of the EDP. An optimum V_{DD} exists.
- Assuming that NMOS and PMOS transistors have comparable threshold and saturation voltages

$$t_p \approx \frac{\alpha C_L V_{DD}}{V_{DD} - V_{Te}}$$

where $V_{Te} = V_T + V_{DSAT}/2$, and α technology parameter.

$$EDP = \frac{\alpha C_L^2 V_{DD}^3}{2(V_{DD} - V_{TE})} \longrightarrow V_{DDopt} = \frac{3}{2} V_{TE}$$

See:

- J. M. Rabaey, A. Chandrakasan, B. Nikolic, «Digital Integrated Circuits: A Design Perspective», Pearson, 2003
  - Cap. 5