UNIT-1 Fundamentals of Low Power VLSI Design

Need for Low Power Circuit Design:

The increasing prominence of portable systems and the need to limit power consumption (and hence, heat dissipation) in very-high density ULSI chips have led to rapid and innovative developments in low-power design during the recent years. The driving forces behind these developments are portable applications requiring low power dissipation and high throughput, such as notebook computers, portable communication devices and personal digital assistants (PDAs).In most of these cases, the requirements of low power consumption must be met along with equally demanding goals of high chip density and high throughput. Hence, low-power design of digital integrated circuits has emerged as a very active and rapidly developing field of CMOS design.

The limited battery lifetime typically imposes very strict demands on the overall power consumption of the portable system. Although new rechargeable battery types such as Nickel-Metal Hydride (NiMH) are being developed with higher energy capacity than that of the conventional Nickel-Cadmium (NiCd) batteries, revolutionary increase of the energy capacity is not expected in the near future. The energy density (amount of energy stored per unit weight) offered by the new battery technologies (e.g., NiMH) is about 30 Watt-hour/pound, which is still low in view of the expanding applications of portable systems. Therefore, reducing the power dissipation of integrated circuits through design improvements is a major challenge in portable systems design.

The need for low-power design is also becoming a major issue in high-performance digital systems, such as microprocessors, digital signal processors (DSPs) and other applications. Increasing chip density and higher operating speed lead to the design of very complex chips with high clock frequencies. If the clock frequency of the chip increases then the power dissipation of the chip, and thus, the temperature, increase linearly. Since the dissipated heat must be removed effectively to keep the chip temperature at an acceptable level, the cost of packaging, cooling and heat removal becomes a significant factor. Several high-performance microprocessor chips designed in the early 1990s (e.g., Intel Pentium, DEC Alpha, PowerPC) operate at clock frequencies in the range of 100 to 300 MHz, and their typical power consumption is between 20 and 50 W.

ULSI reliability is yet another concern which points to the need for low-power design. There is a close correlation between the peak power dissipation of digital circuits and reliability problems such as electro migration and hot-carrier induced device degradation. Also, the thermal stress caused by heat dissipation on chip is a major reliability concern. Consequently, the reduction of power consumption is also crucial for reliability enhancement.

The methodologies which are used to achieve low power consumption in digital systems span a wide range, from device/process level to algorithm level. Device characteristics (e.g., threshold voltage), device geometries and interconnect properties are significant factors in lowering the power consumption. Circuit-level measures such as the proper choice of circuit design styles, reduction of the voltage swing and clocking strategies can be used to reduce power dissipation at the transistor level. Architecture-level measures include smart power management of various system blocks, utilization of pipelining and parallelism, and design of bus structures. Finally, the power consumed by the system can be reduced by a proper selection of the data processing algorithms, specifically to minimize the number of switching events for a given task.

Sources of Power Dissipation

The average power dissipation in conventional CMOS digital circuits can be classified into three main components, namely,

(1) The dynamic (switching) power dissipation

(2) The short-circuit power dissipation and

(3) The leakage power dissipation.

If the system or chip includes circuits other than conventional CMOS gates that have continuous current paths between the power supply and the ground, a fourth (static) power component should also be considered

(1)Switching Power Dissipation

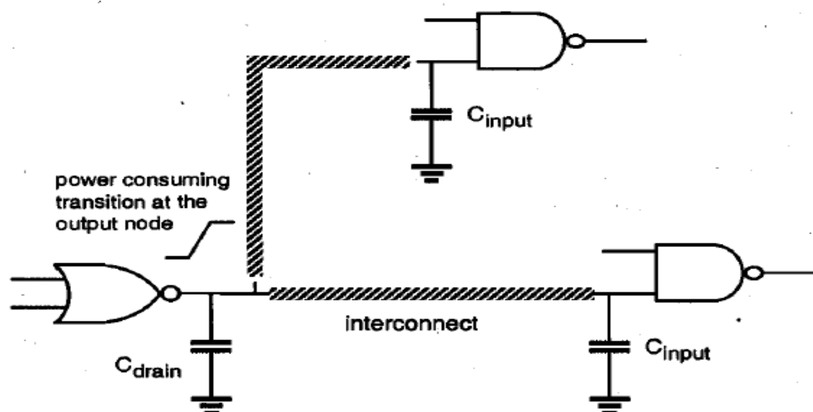
Switching Power Dissipation represents the power dissipation during a switching event. This means that the output node voltage of a CMOS logic gate makes a power consuming transition. In digital CMOS circuits, dynamic power is dissipated when energy is drawn from the power supply to charge up the output node capacitance. During the charge-up phase, the output node voltage typically makes a full transition from 0 to V_{DD}, and the energy used for the transition is relatively independent of the function performed by the circuit.

To explain the dynamic power dissipation during switching, consider the circuit given in Fig. below. Here, a two - input NOR gate drives two NAND gates, through interconnection lines. The total capacitive load at the output of the NOR gate consists of

(1) The output capacitance of the gate itself

(2) The total interconnect capacitance, and

(3) The input capacitances of the driven gates.



A NOR gate driving two NAND gates through interconnection lines.

Output capacitance:

The output capacitance of the gate consists mainly of the junction parasitic capacitances, which are due to the drain diffusion regions of the MOS transistors in the circuit. The important feature to highlight here is that the amount of capacitance is approximately a linear function of the junction area. So, the size of the total drain diffusion area determines the amount of parasitic capacitance.

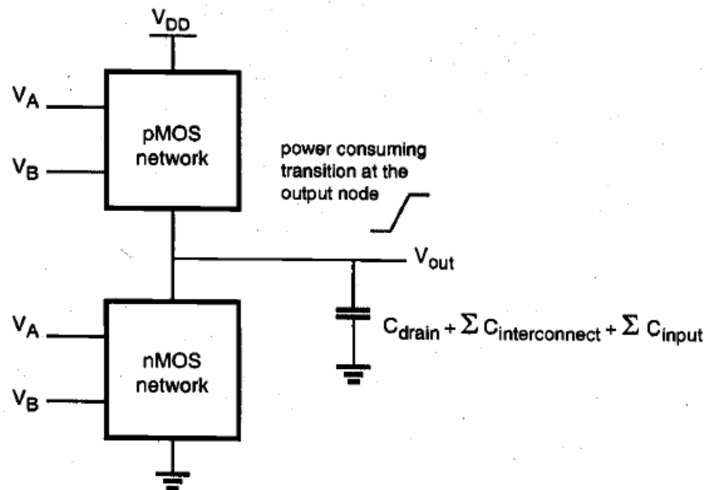
Total interconnect capacitance:

The interconnect lines between the gates contribute to the total interconnect capacitance. **Note** Particularly in sub-micron technologies, the interconnect capacitance can become the dominant component, compared to the transistor-related capacitances.

Input capacitances:

The input capacitances are mainly due to gate oxide capacitances of the transistors connected to the input terminal. Again, the amount of the gate oxide capacitance is determined primarily by the gate area of each transistor.

Generic representation of a CMOS logic gate



Generic representation of a CMOS logic gate for switching power calculation.

Any CMOS logic gate making an output voltage transition can thus be represented by its nMOS network, pMOS network, and the total load capacitance connected to its output node, as seen in above Fig. The average power dissipation of the CMOS logic gate, driven by a periodic input voltage waveform with ideally zero rise- and fall-times, can be calculated from the energy required to charge up the output node to V_{DD} and charge down the total output load capacitance to ground level.

$$P_{avg} = \frac{1}{T}\left[\int_0^{T/2} V_{out}\left(-C_{load}\frac{dV_{out}}{dt}\right)dt + \int_{T/2}^{T} (V_{DD}-V_{out})\left(C_{load}\frac{dV_{out}}{dt}\right)dt\right]$$

.......(1)

Simplifying this integral gives the well-known expression for the average dynamic (switching) power consumption in CMOS logic circuits.

$$P_{avg} = \frac{1}{T} C_{load} V_{DD}^2$$

........(2)

$$P_{avg} = C_{load} \cdot V_{DD}^2 \cdot f$$

.........(3)

Note the average switching power dissipation of a CMOS gate is essentially independent of all transistor characteristics and transistor sizes. Hence, given an input pattern, the switching delay times have no relevance to the amount of power consumption during the switching events as long as the output voltage swing is between 0 and V_{DD}.

Equation (3) shows that the average dynamic power dissipation is proportional to the square of the power supply voltage; hence, any reduction of V_{DD} will significantly reduce the power consumption.

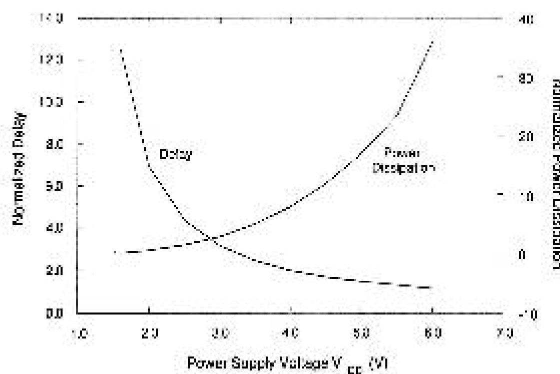
Another way to limit the dynamic power dissipation of a CMOS logic gate is to reduce the amount of switched capacitance at the output.

Effect of reducing the power supply voltage V_{DD} on switching power dissipation

Although the reduction of power supply voltage significantly reduces the dynamic power dissipation, the expected design trade-off is the increase of delay. This can be seen by examining the following propagation delay expressions for the CMOS inverter circuit.

$$\tau_{PHL} = \frac{C_{load}}{k_n\left(V_{DD} - V_{T,n}\right)}\left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + ln\left(\frac{4\left(V_{DD} - V_{T,n}\right)}{V_{DD}} - 1\right)\right] \quad \text{.............(4)}$$

Assuming that the power supply voltage is being scaled down while all other variables are kept constant, it can be seen that the propagation delay time will increase. Figure below shows the normalized variation of the delay as a function of V_{DD}, where the threshold voltages of the nMOS and the pMOS transistor are assumed to be constant, $V_{T,n} = 0.8$ V and $V_{T,p} = -0.8$ V, respectively. The normalized variation of the average switching power dissipation as a function of the supply voltage is also shown on the same plot.



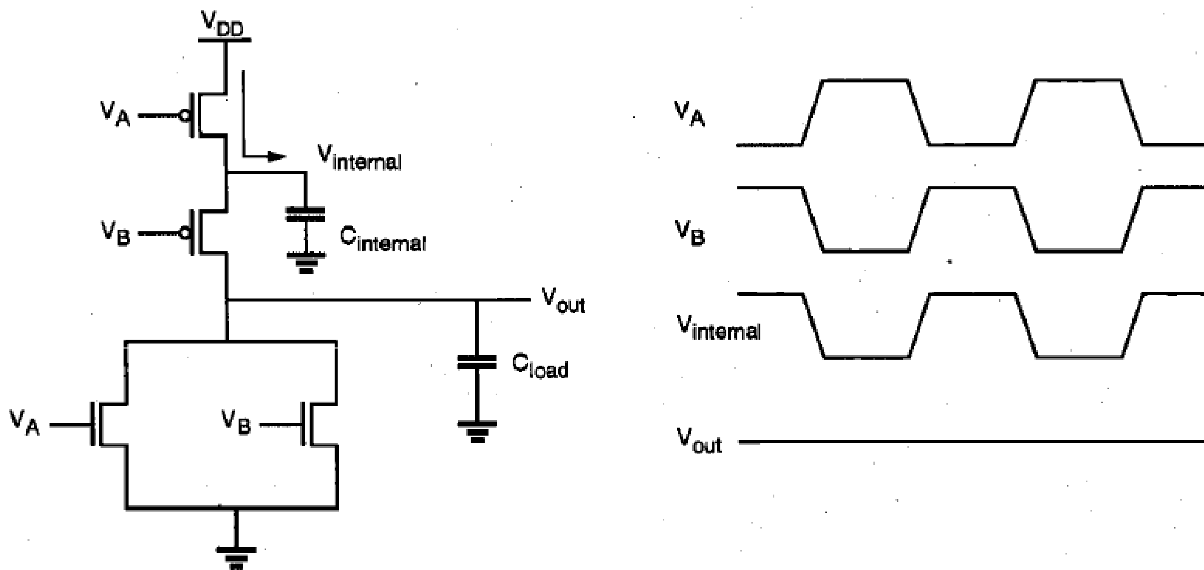
Note the dependence of circuit speed on the power supply voltage may also influence the relationship between the dynamic power dissipation and the supply voltage. Equation (3) suggests a quadratic improvement (reduction) of power consumption as the power supply voltage is reduced. However, this interpretation assumes that the switching frequency (i.e., the number of switching events per unit time) remains constant. If the circuit is always operated at the maximum frequency allowed by its propagation delay, on the other hand, the number of switching events per unit time (i.e., the operating frequency) will obviously drop as the propagation delay becomes larger with the reduction of the power supply voltage. The net result is that the dependence of switching power dissipation on the power supply voltage becomes stronger than a simple quadratic relationship.

The analysis of switching power dissipation presented above is based on the assumption that the output node of a CMOS gate faces one power-consuming transition (0-to-V_{DD} transition) in each clock cycle. This assumption, however, is not always correct; the node transition rate can be smaller than the clock rate, depending on the circuit topology, logic style and the input signal statistics.

To better represent this behavior, we will introduce αT (node transition factor), which is the effective number of power-consuming voltage transitions experienced per clock cycle. Then, the average switching power dissipation becomes

$$P_{avg} = \alpha_T \cdot C_{load} \cdot V_{DD}^2 \cdot f_{CLK}$$

Note that in most complex CMOS logic gates, a number of internal circuit nodes also make full or partial voltage transitions during switching. Since there is a parasitic node capacitance associated with each internal node, these internal transitions contribute to the overall power dissipation of the circuit. In fact, an internal node may face several transitions while the output node voltage of the circuit remains unchanged, as illustrated in below Fig.



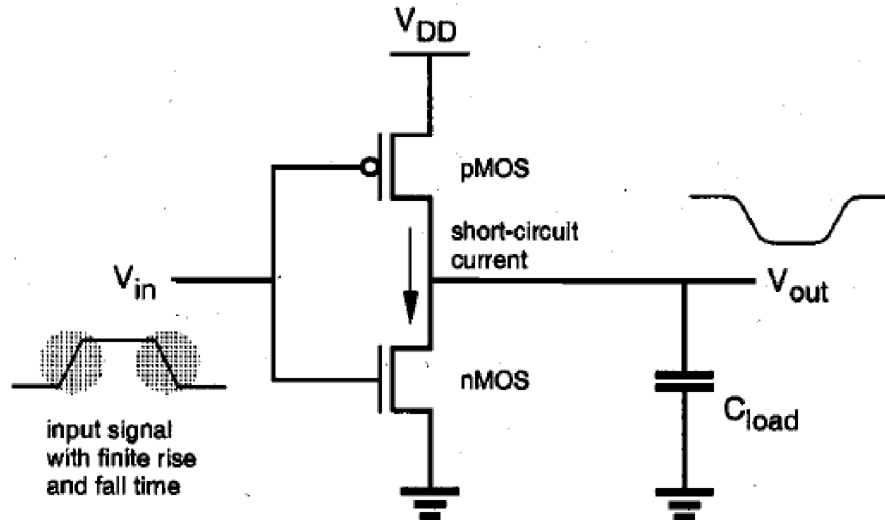
Switching of the internal node in a two-input NOR gate results in dynamic power dissipation even if the output node voltage remains unchanged.

In the most general case, the internal node voltage transitions can also be partial transitions, i.e., the node voltage swing may be only Vi which is smaller than the full voltage swing of V_{DD}. Taking this possibility into account, the generalized expression for the average switching power dissipation can be written as

$$P_{avg} = \left(\sum_{i=1}^{\#\ of\ nodes} \alpha_{Ti} \cdot C_i \cdot V_i \right) \cdot V_{DD} \cdot f_{CLK}$$

Where C_i represents the parasitic capacitance associated with each node and αTi represents the corresponding node transition factor associated with that node.

(2)Short-Circuit Power Dissipation:
The switching power dissipation discussed above is purely due to the energy required to charge up the parasitic capacitances in the circuit, and the switching power is independent of the rise and fall times of the input signals. Now, if a CMOS inverter (or a logic gate) is driven with input voltage waveforms with finite rise and fall times, both the nMOS and the pMOS transistors in the circuit may conduct simultaneously for a short amount of time during switching, forming a direct current path between the power supply and the ground, as shown in below Fig.



Both nMOS and pMOS transistor may conduct (simultaneously) a short-circuit current during switching.

Short-circuit current component is the current component which passes through both the nMOS and the pMOS devices during switching. It does not contribute to the charging of the capacitances in the circuit.

This component is particularly powerful if the output load capacitance is small, and/or if the input signals rise and fall times are large, as shown in below Fig. Here, the input/output voltage waveforms and the components of the current drawn from the power supply are illustrated for a symmetrical CMOS inverter with small capacitive load. The nMOS transistor in the circuit starts conducting when the rising input voltage exceeds the threshold voltage $V_{T,n}$.

The pMOS transistor remains on until the input reaches the voltage level (V_{DD} - $|V_{T,p}|$). Thus, there is a time window during which both transistors are turned on. As the output capacitance is discharged through the nMOS transistor, the output voltage starts to fall. The drain-to-source voltage drop of the pMOS transistor becomes nonzero, which allows the pMOS transistor to conduct as well. The short circuit current is terminated when the input voltage transition is completed and the pMOS transistor is turned off. A similar event is responsible for the short- circuit current component during the falling input transition, when the output voltage starts rising while both transistors are on.

Note the magnitude of the short-circuit current component will be approximately the same during both the rising-input transition and the falling-input transition, assuming that the inverter is symmetrical and the input rise and fall times are identical. The pMOS transistor also conducts the current which is needed to charge up the small output load capacitance, but only during the falling-input transition (the output capacitance is discharged through the nMOS device during the rising-input transition). This current component, which is responsible for the switching power

dissipation of the circuit (current component to charge up the load capacitance), is also shown in Fig. The average of both of these current components determines the total amount of power drawn from the supply.

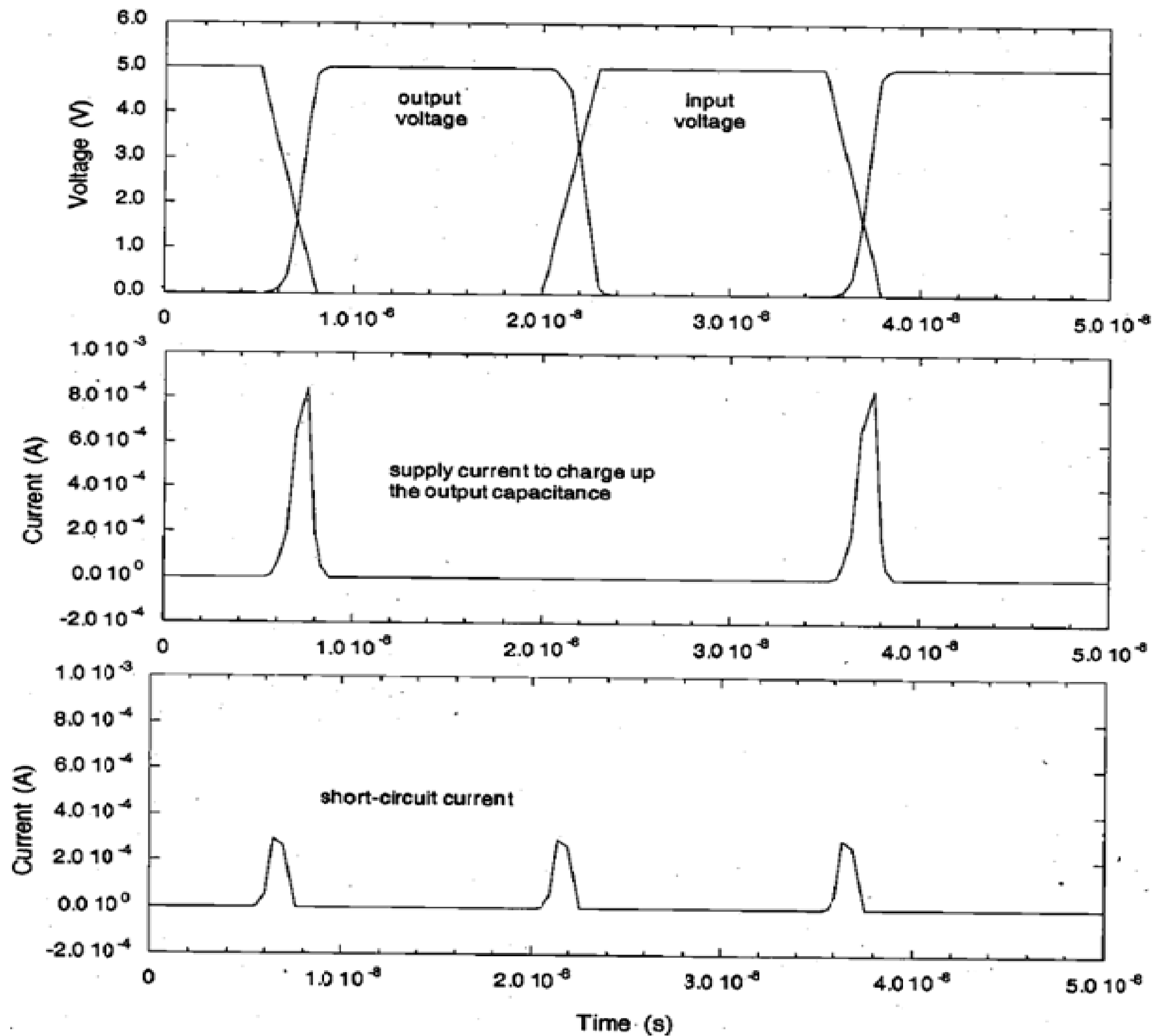
For a simple analysis consider a symmetric CMOS inverter with k = kn = kp and $V_T = V_{T,n} = |V_{T,p}|$, and with a very small capacitive load. If the inverter is driven with an input voltage waveform with equal rise and fall times (t = trise = tfall), it can be derived that the time-averaged short circuit current drawn from the power supply is

$$I_{avg}(short-circuit) = \frac{1}{12} \cdot \frac{k \cdot \tau \cdot f_{CLK}}{V_{DD}} (V_{DD} - 2V_T)^3$$

Hence, the short-circuit power dissipation becomes

$$P_{avg}(short-circuit) = \frac{1}{12} \cdot k \cdot \tau \cdot f_{CLK} \cdot (V_{DD} - 2V_T)^3$$

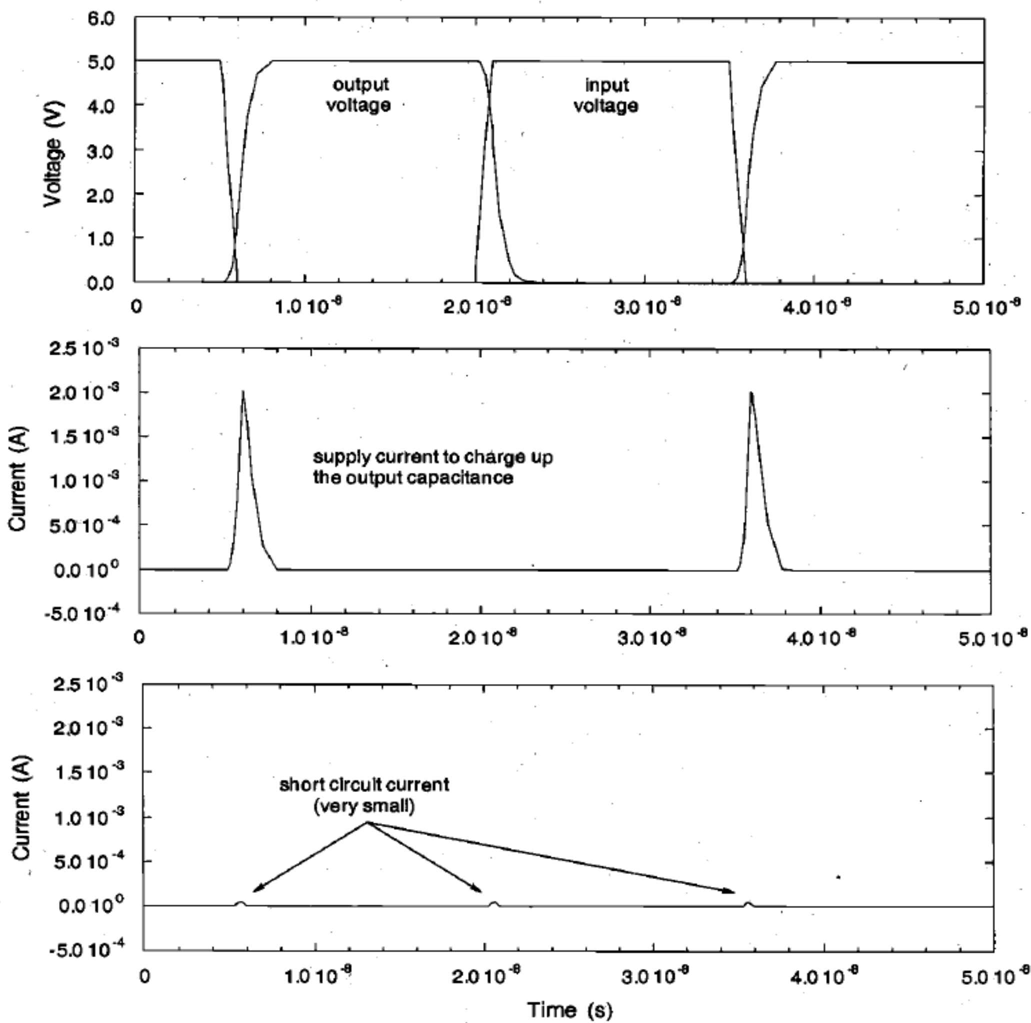
Note that the short-circuit power dissipation is linearly proportional to the input signal rise and fall times, and also to the transconductance of the transistors. Hence, reducing the input transition times will obviously decrease the short circuit current component.



Now consider the same CMOS inverter with a larger output load capacitance and smaller input transition times. During the rising input transition, the output voltage will effectively remain at V_{DD} until the input voltage completes its swing and the output will start to drop only after the input has reached its final value.

Although both the nMOS and the pMOS transistors are on simultaneously during the transition, the pMOS transistor cannot conduct a significant amount of current since the voltage drop between its source and drain terminals is approximately equal to zero. Similarly, the output voltage will remain approximately equal to 0 V during a falling input transition and it will start to rise only after the input voltage completes its swing. Again, both transistors will be on simultaneously during the input voltage transition, but the nMOS transistor will not be able to conduct a significant amount of current since its drain-to-source voltage is approximately equal to zero. The fig below shows the input and output voltage waveforms of the inverter and short-circuit and dynamic current components.

Note that the peak value of the supply current to charge up the output load capacitance is larger in this case. The reason for this is that the pMOS transistor remains in saturation during the entire input transition, as opposed to the previous case where the transistor leaves the saturation region before the input transition is completed.



The short-circuit power dissipation can be reduced by making the output voltage transition times larger and/or by making the input voltage transition times smaller. Yet this goal should be balanced carefully against other performance goals such as propagation delay, and the reduction of the short-circuit current should be considered as one of the many design requirements that must satisfied by the designer.

(3)Leakage Power Dissipation

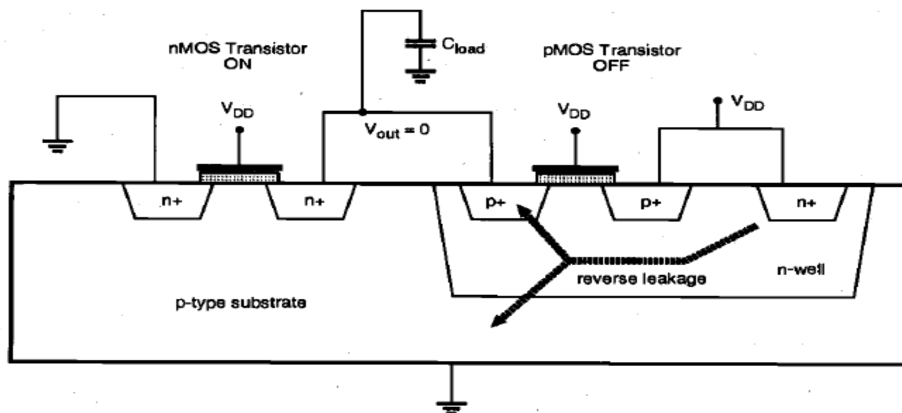
The nMOS and pMOS transistors used in a CMOS logic gate generally have nonzero reverse leakage and sub threshold currents. In a CMOS VLSI chip containing a very large number of transistors, these currents can contribute to the overall power dissipation even when the transistors are not undergoing any switching event. The magnitude of the leakage currents is determined mainly by the processing parameters.

Two main leakage current components found in a MOSFET are

(1) Reverse diode leakage current
(2) Sub threshold leakage current

Reverse diode leakage current:

The reverse diode leakage occurs when the pn-junction between the drain and the bulk of the transistor is reversely biased. The reverse-biased drain junction then conducts a reverse saturation current which is eventually drawn from the power supply. Consider a CMOS inverter with a high input voltage, where the nMOS transistor is turned on and the output node voltage is discharged to zero. Although the pMOS transistor is turned off, there will be a reverse potential difference of V_{DD} between its drain and the n-well, causing a diode leakage through the drain junction. The n-well region of the pMOS transistor is also reverse-biased with V_{DD}, with respect to the p-type substrate. Therefore, another significant leakage current component exists due to the n-well junction



A similar situation can be observed when the input voltage is equal to zero, and the output voltage is charged up to V_{DD} through the pMOS transistor. Then, the reverse potential difference between the nMOS drain region and the p-type substrate causes a reverse leakage current which is also drawn from the power supply (through the pMOS transistor).

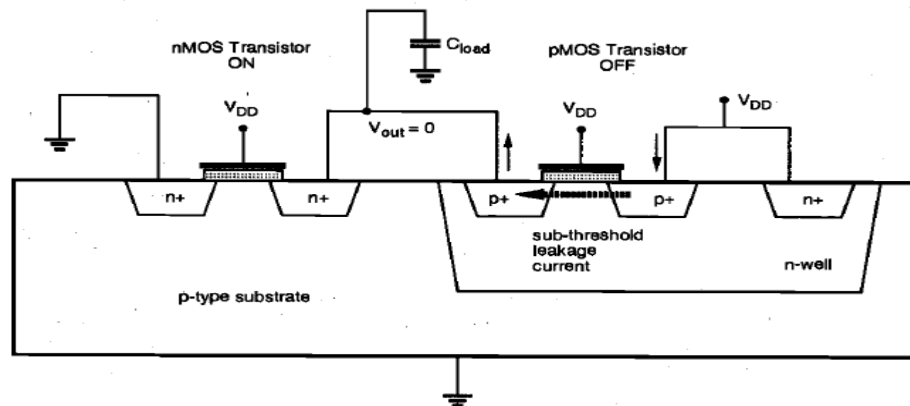
The magnitude of the reverse leakage current of a pn-junction is given by the following expression

$$I_{reverse} = A \cdot J_S \left(e^{\frac{q V_{bias}}{kT}} - 1 \right)$$

Where V_{bias} is the magnitude of the reverse bias voltage across the junction, J_S is the reverse saturation current density and the A is the junction area. The typical magnitude of the reverse saturation current density is 1 - 5 pA/mm^2, and it increases quite significantly with temperature. **Note** that the reverse leakage occurs even during the stand-by operation when no switching takes place. Hence, the power dissipation due to this mechanism can be significant in a large chip containing several million transistors.

Subthreshold leakage current:

Another component of leakage currents which occur in CMOS circuits is the subthreshold current, which is due to carrier diffusion between the source and the drain region of the transistor in weak inversion. An MOS transistor in the subthreshold operating region behaves similar to a bipolar device and the subthreshold current exhibits an exponential dependence on the gate voltage. The amount of the subthreshold current may become significant when the gate-to source voltage is smaller than, but very close to the threshold voltage of the device. In this case, the power dissipation due to subthreshold leakage can become comparable in magnitude to the switching power dissipation of the circuit. The subthreshold leakage current is shown in Fig. below.



Subthreshold leakage current path in a CMOS inverter with high input voltage.

Note the sub threshold leakage current also occurs when there is no switching activity in the circuit, and this component must be carefully considered for estimating the total power dissipation in the stand-by operation mode. The sub threshold current expression is given below, in order to illustrate the exponential dependence of the current on terminal voltages.

$$I_D(subthreshold) \cong \frac{qD_n W x_c n_0}{L_B} \cdot e^{\frac{q\phi_r}{kT}} \cdot e^{\frac{q}{kT}(A \cdot V_{GS} + BV_{DS})}$$

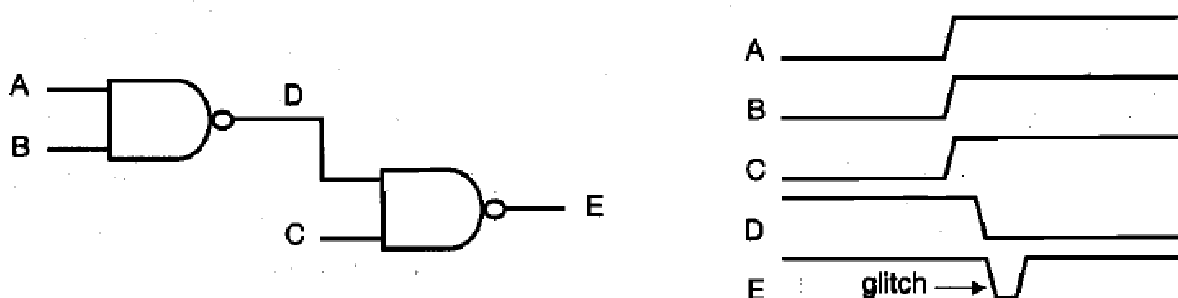
One relatively simple measure to limit the subthreshold current component is to avoid very low threshold voltages, so that the V_{GS} of the nMOS transistor remains safely below $V_{T,n}$ when the input is logic zero, and the $|V_{GS}|$ of the pMOS transistor remains safely below $|V_{T,p}|$ when the input is logic one.

In addition to the three major sources of power consumption in CMOS digital integrated circuits discussed here, some chips may also contain components or circuits which actually consume static power. One example is the pseudo-nMOS logic circuits which utilize a pMOS transistor as the pull-up device. The presence of such circuit blocks should also be taken into account when estimating the overall power dissipation of a complex system.

GLITCH POWER DISSIPATION:

The glitching power dissipation occurs due to finite delay. This Power dissipated in the intermediate transitions during the evaluation of the logic function of the circuit.

In multi-level logic circuits, the finite propagation delay from one logic block to the next can cause spurious signal transitions, or glitches as a result of critical races or dynamic hazards. In general, if all input signals of a gate change simultaneously, no glitching occurs. But a dynamic hazard or glitch can occur if input signals change at different times. Thus, a node can exhibit multiple transitions in a single clock cycle before settling to the correct logic level. In some cases, the signal glitches are only partial, i.e., the node voltage does not make a full transition between the ground and V_{DD} levels, yet even partial glitches can have a significant contribution to dynamic power dissipation.

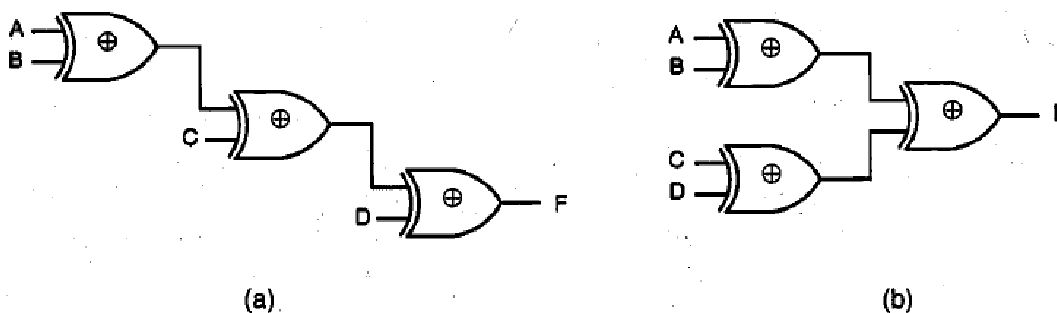


Signal glitching in multi-level static CMOS circuits.

Glitches occur primarily due to a mismatch or imbalance in the path lengths in the logic network. Such a mismatch in path length results in a mismatch of signal timing with respect to the primary inputs.

Example, consider the simple parity network shown in Fig below. Assuming that all XOR blocks have the same delay, it can be seen that the network in Fig.(a) will suffer from glitching due to the wide disparity between the arrival times of the input signals for the gates.

In the network shown in Fig.(b), on the other hand, all arrival times are identical because the delay paths are balanced. Such redesign can significantly reduce the glitching transitions, and consequently, the dynamic power dissipation in complex multi-level networks. Also notice that the tree structure shown in Fig. (b) Results in smaller overall propagation delay. Finally, it should be noted that glitching is not a significant issue in multi-level dynamic CMOS logic circuits, since each node undergoes at most one transition per clock cycle.



(a) Implementation of a four-input parity (XOR) function using a chain structure.

(b) Implementation of the same function using a tree structure which will reduce glitching transitions.

Short-Channel Effects

Short-Channel Devices:

A MOSFET device is considered to be short when the channel length is the same order of magnitude as the depletion-layer widths (x_{dD}, x_{dS}) of the source and drain junction. As the channel length L is reduced to increase both the operation speed and the number of components per chip, the so-called short-channel effects arise.

Short-Channel Effects

The short-channel effects are attributed to two physical phenomena:

1. The limitation imposed on electron drift characteristics in the channel,
2. The modification of the threshold voltage due to the shortening channel length.

In particular five different short-channel effects can be distinguished:

1. Drain-induced barrier lowering and punch through
2. Surface scattering
3. Velocity saturation
4. Impact ionization
5. Hot electron effect

Drain-induced barrier lowering and punch through:

The expressions for the drain and source junction widths are:

$$x_{dD} = \sqrt{\left(\frac{2\varepsilon_{Si}}{qN_A}\right)(V_{DS} + \phi_{Si} + V_{SB})}$$

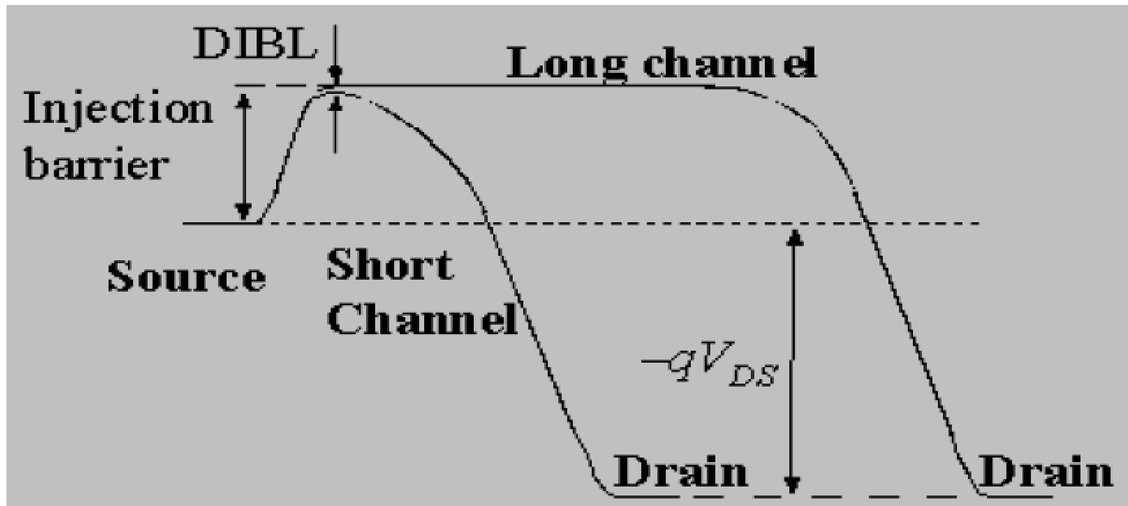
And

$$x_{dS} = \sqrt{\left(\frac{2\varepsilon_{Si}}{qN_A}\right)(\phi_{Si} + V_{DB})}$$

Where V_{SB} and V_{DB} are source-to-body and drain-to-body voltages

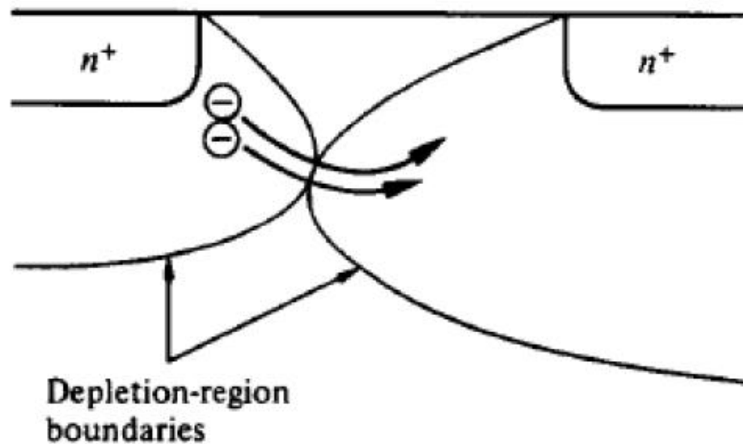
There exists a potential barrier between source and drain which is to be lowered by applying gate voltage. In short channel devices in addition to the gate voltage, drain voltage also has a significant effect on reducing this barrier. As the source & drain get closer, they become electrostatically coupled, so that the drain bias can affect the potential barrier to carrier flow at the source junction. As a result, sub threshold current increases. As the drain depletion region continues to increase with the bias, it can actually interact with the source to channel junction and hence lowers the potential barrier. This problem is known as Drain Induced Barrier Lowering (DIBL).

When the source junction barrier is reduced, electrons are easily injected into the channel and the gate voltage has no longer any control over the drain current. In long channel devices, the source and drain are separated far enough that their depletion regions have no effect on the potential or field pattern in most part of the device. Hence, for such devices, the threshold voltage is virtually independent of the channel length and drain bias. DIBL is enhanced at high drain voltages and shorter channel lengths.



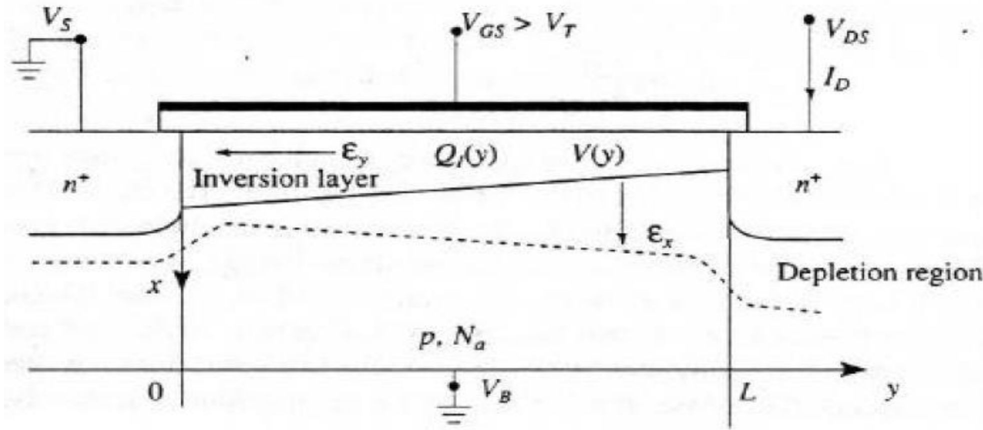
Punchthrough:

When the drain is at high enough voltage with respect to the source, the depletion region around the drain may extend to the source, causing current to flow irrespective of gate voltage (i.e. even if gate voltage is zero). This is known as Subsurface Punchthrough as it takes place away from the gate oxide and substrate interface. So when channel length L decreases (i.e. short channel length case), punch through voltage rapidly decreases. In short-channel devices, due to the proximity of the drain and the source, the depletion regions at the drain-substrate and source-substrate junctions extend into the channel. As the channel length is reduced, if the doping is kept constant, the separation between the depletion region boundaries decreases. An increase in the reverse bias across the junctions also pushes the junctions nearer to each other. When the combination of channel length and reverse bias leads to the merging of the depletion regions, punchthrough is said to have occurred.



Surface scattering:

As the channel length becomes smaller due to the lateral extension of the depletion layer into the channel region, the longitudinal electric field component ε_y increases, and the surface mobility becomes field-dependent. Since the carrier transport in a MOSFET is confined within the narrow inversion layer, and the *surface scattering* (that is the collisions suffered by the electrons that are accelerated toward the interface by ε_x) causes reduction of the mobility, the electrons move with great difficulty parallel to the interface, so that the average surface mobility, even for small values of ε_y, is about half as much as that of the bulk mobility.



Velocity saturation:

The performance short-channel devices are also affected by *velocity saturation*, which reduces the transconductance in the saturation mode. At low ε_y, the electron drift velocity v_{de} in the channel varies linearly with the electric field intensity. However, as ε_y increases above 10^4 V/cm, the drift velocity tends to increase more slowly, and approaches a saturation value of $v_{de(sat)}=10^7$ cm/s around ε_y =105 V/cm at 300 K.

Note that the drain current is limited by velocity saturation instead of pinchoff. This occurs in short channel devices when the dimensions are scaled without lowering the bias voltages. Using $v_{de(sat)}$, the maximum gain possible for a MOSFET can be defined as

$$g_m = WC_{ox}v_{de(sat)}$$

Impact ionization

Another undesirable short-channel effect, especially in NMOS, occurs due to the high velocity of electrons in presence of high longitudinal fields that can generate electron-hole (e-h) pairs by *impact ionization*, that is, by impacting on silicon atoms and ionizing them.

It happens as follow: normally, most of the electrons are attracted by the drain, while the holes enter the substrate to form part of the parasitic substrate current. Moreover, the region between the source and the drain can act like the base of an npn transistor, with the source playing the role of the emitter and the drain that of the collector. If the aforementioned holes are collected by the source, and the corresponding hole current creates a voltage drop in the substrate material of the order of 0.6V, the normally reversed-biased substrate-source pn junction will conduct appreciably. Then electrons can be injected from the source to the substrate, similar to the injection of electrons from the emitter to the base. They can gain enough energy as they travel toward the drain to create new electron hole pairs. The situation can worsen if some electrons generated due to high fields escape the drain field to travel into the substrate, thereby affecting other devices on a chip.

Hot electron effect:

Another problem, related to high electric fields, is caused by so-called *hot electrons*. These high energy electrons can enter the oxide, where they can be trapped, giving rise to oxide charging that can accumulate with time and degrade the device performance by increasing V_T and affect adversely the gate's control on the drain current.