Chapter 1: MOS Devices and Circuits

Sections:

In this chapter we begin with a discussion of the basic properties of the n-channel, metal-oxide-semiconductor (MOS), field effect transistor (FET). We then describe and analyze a number of circuits composed of interconnected MOS field effect transistors. The circuits described are typical of those we will commonly use in the design of integrated systems. The analysis, though highly condensed, is conceptually correct and provides a basis for the solution of most system problems typically encountered.

Integrated systems in MOS technology contain three levels of conducting material separated by intervening layers of insulating material. Proceeding from top to bottom, these levels are termed the *metal*, the *polysilicon*, and the *diffusion* levels respectively. Patterns for paths on these three levels, and the locations of contact cuts through the insulating material to connect certain points between levels, are transferred into the levels during the fabrication process from *masks* similar to photographic negatives. The details of the fabrication process will be discussed in chapter 2.

In the absence of contact cuts through the insulating material, paths on the metal level may cross over paths on the polysilicon or diffusion levels with no significant functional effect. However, wherever a path on the polysilicon level crosses a path on the diffusion level, a transistor is created. Such a transistor has the characteristics of a simple switch, with a voltage on the polysilicon level path controlling the flow of current in the diffusion level path. Circuits composed of such transistors, interconnected by patterned paths on the three levels, form our basic building blocks. With these basic circuits, we will architect integrated systems, to be fabricated on the surface of monolithic crystalline chips of silicon.

The MOS Transistor

An MOS transistor will be produced on the integrated system chip wherever a polysilicon path crosses a diffusion path, as shown in figure 1a. The electrical symbol used to represent the MOS transistor in our circuit diagrams is shown in figure 1b, along with symbols and polarities of certain voltages of interest. Note that the source and drain terminals of the device are physically symmetrical. For the n-channel MOSFETs, these terminal labels are assigned such that V_{ds} is normally positive. A more detailed view of the rectangular region called the gate, where the polysilicon (poly) crosses the diffusion, is given in figure 1c. During fabrication the diffusion paths are formed after the poly paths are formed, as explained more fully in chapter 2. The poly gate, and the thin layer of oxide beneath it, mask the region under the gate during diffusion. Therefore, no diffusion path forms under the gate, and there is no direct connection on the diffusion level between the source and drain terminals of the transistor. Notice in this discussion that metal, poly, and diffusion paths all conduct electricity well enough to be considered "wires" until further notice.

In the absence of any charge on the gate, the drain to source path through the transistor is like an open switch. The gate, separated from the substrate by the layer of thin oxide, forms a capacitor. If sufficient positive charge is placed on the gate so that V_{gs} exceeds a *threshold voltage* V_{th}, electrons will be attracted to the region under the gate to form a conducting path between drain and source. Most of the transistors we will use in our systems have threshold voltages greater than zero. These are called *enhancement mode* MOSFETs, and their threshold voltage typically equals ~ 0.2(VDD), where VDD is the positive supply voltage for the particular technology.

The basic operation performed by the MOS transistor is to use charge on its gate to control the movement of negative charge between source and drain through the channel under the gate. The current from source to drain equals the charge induced in the channel divided by the transit time or average time required for an electron to move from source to drain. The transit time itself is the distance the electron has to move divided by its average velocity. In semiconductors under normal conditions, the velocity is proportional to the electric field driving the electrons. The relationship between drain to source current I_{ds}, drain to source voltage V_{ds}, and gate to source voltage V_{gs} is sketched in figure 1d. For small V_{ds}, the transit time τ is given by equation 1.

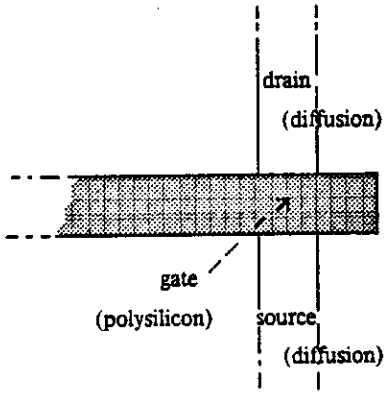


Fig. 1a. MOS Transistor
(top view)

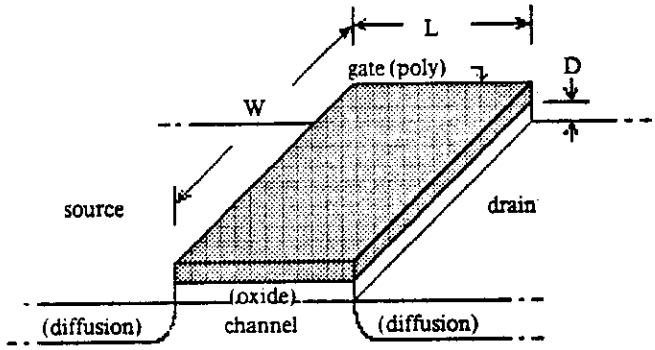


Fig. 1c. MOSFET Gate Dimensions

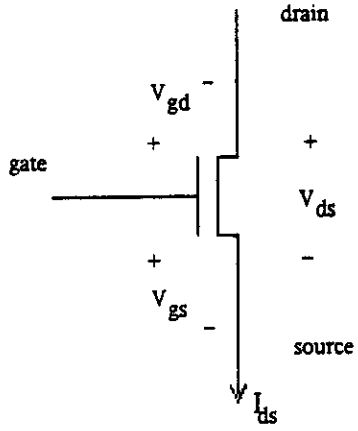


Fig. 1b. MOS Transistor Symbol
(subscripts in + to - direction sequence)

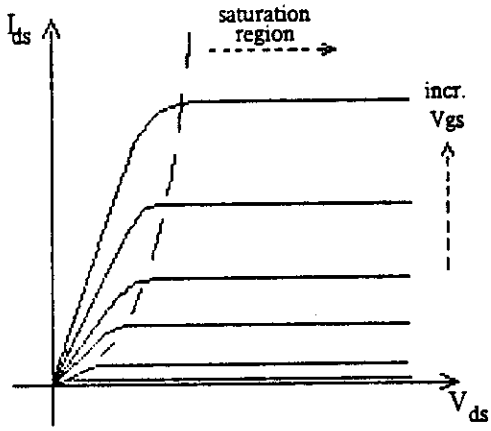


Fig. 1d. Current vs Voltage

Transit time: $\tau = L/\text{velocity} = L/\mu E = L^2/[\mu V_{ds}]$ [eq.1]

The proportionality constant μ is called the *mobility* of the charge carriers, in this case electrons, under the influence of an electric field in the conducting material of the channel region. It is a velocity per unit electric field (cm^2/volt-sec). We shall see that the transit time is the fundamental time unit of the entire integrated system.

The amount of negative charge in transit is just the gate capacitance times the voltage on the gate in excess of the threshold voltage. The capacitance of two parallel conductors of area A, separated by insulating material of thickness D, equals $\varepsilon A/D$. The proportionality constant ε is called the permittivity of the insulating material, and has a simple interpretation. It is the capacitance of parallel conductors of area A = 1 cm^2, separated by a thickness D = 1 cm of the insulator material, and is in the units farad/cm. Therefore, the gate capacitance equals $\varepsilon WL/D$. Thus the charge in transit is given by eq. 2, and the current is given by eq. 3.

Charge in transit: $$Q = - C_g(V_{gs} - V_{th}) = - \frac{\varepsilon WL}{D}(V_{gs} - V_{th}) \quad \text{[eq.2]}$$

Current: $$I_{ds} = - I_{sd} = - \frac{\text{charge in transit}}{\text{transit time}} = \frac{\mu\varepsilon W}{LD}(V_{gs} - V_{th})(V_{ds}) \quad \text{[eq.3]}$$

Note that for small V_{ds}, the drain current is proportional to the source-drain voltage and also to the gate voltage above threshold. Any device with a current through it proportional to the voltage across it, may be viewed as a resistor, and in the case of an MOS device with *low* drain to source voltage, the resistance is controlled by the gate voltage as given in eq. 3a.

$$V_{ds}/I_{ds} = R = L^2/[\mu C_g(V_{gs} - V_{th})] \quad \text{[eq.3a]}$$

In both equations 2 and 3a, C_g is the gate to channel capacitance of the turned on transistor. In the simple case where this transistor is driving the gate of another one identical to it, the time response of the system will be an exponential with a time constant RC_g, given in equation 4. This time constant is identical to the transit time τ given in equation 1.

$$RC_g = L^2/[\mu(V_{gs} - V_{th})] = \tau \qquad \text{[eq.4]}$$

Although the above equations are greatly simplified, they provide sufficient information to make many design decisions which we will face, and also give us insight into the scaling of devices to smaller sizes. In particular, the transit time τ can be viewed as the basic time unit of any system we shall build in the integrated technology. In almost all situations, the fastest operation which we can perform is to transfer a signal from the gate of one MOS transistor onto the gate of another. The transit time is the minimum time in which a charge placed on the gate of one transistor results in the transfer of a similar charge through that transistor's channel onto the gate of a subsequent transistor. For example, to transfer a charge from one transistor onto two transistors identical to it requires a minimum of two transit times. Thus, the transit time of the basic transistor in an integrated system can be viewed as the unit of time in which all other times in the system are scaled. Although it is a somewhat optimistic approximation, we will use τ as the primary time metric in calculating the delay through elementary inverting logic stages. More accurate predictions of circuit behavior can be produced using any one of a number of available circuit simulation programs.[5,6]

As V_{ds} is increased, not all of the drain to source voltage is available for reducing the transit time. Drain voltage in excess of one threshold below the gate voltage creates a short region of high electric field adjacent to the drain which the carriers cross very quickly. The electric field in the major portion of the channel from the source up to this region is proportional to $V_{gs} - V_{th}$, as shown in figure 1e. For $V_{ds} > (V_{gs} - V_{th})$, the drain current becomes independent of V_{ds}. Further increases in V_{ds} neither increase I_{ds} nor decrease the transit time. This range of V_{ds} values is known as saturation.

In saturation: $$I_{ds} = \frac{\mu\varepsilon W(V_{gs} - V_{th})^2}{2LD} \qquad \text{[eq.5]}$$

With the exception of the factor of 2 in the denominator, this equation is similar to equation 3, with the V_{ds} factor in that equation replaced by its maximum effective value, $V_{gs} - V_{th}$. The factor of 2 in equation 5 arises from the non-uniformity of the electric field in the channel region when in saturation[1,R4].

The Basic Inverter

The first logic circuit we will describe is the basic digital inverter. Analysis of this circuit is then extended to analysis of basic NAND and NOR logic gates. The inverter's logic function is to produce an output which is the complement of its input. When describing the logic function of circuits in integrated systems, we assign the value logic-1 to voltages equaling or exceeding some defined logic threshold voltage, and logic-0 to voltages less than this threshold voltage.

Were there an efficient way to implement resistors in the MOS technology, we could build a basic digital inverter circuit using the configuration of figure 2a. Here, if the inverter input voltage V_{in} is less than the transistor threshold voltage $V_{th,}$ then the transistor is switched off, and V_{out} is "pulled-up" to the positive supply voltage VDD. In this case the output is the complement of the input. If V_{in} is greater than $V_{th,}$ the transistor is switched on and current flows from the VDD supply through the resistor R to GND. If R were sufficiently large, V_{out} could be "pulled-down" well below V_{th}, thus again complementing the input. However, the resistance per unit length of minimum width lines of various available conducting elements is far less than the effective resistance of the switched on MOSFET. Implementing a sufficiently large inverter pullup using resistive lines would require a very large area compared to that occupied by the transistor itself.

To circumvent this problem a *depletion mode* MOSFET is used as a pullup for the basic inverter circuit, symbolized and configured as shown in figure 2b. In contrast to the usual enhancement mode transistor, the depletion mode transistor has a threshold voltage, V_{dep}, that is less than zero. During fabrication, one of the masks is used to select any desired subset of transistors in the integrated system for processing as depletion mode transistors. For a depletion mode transistor to turn off, it requires a voltage on its gate relative to its source that is more negative than V_{dep}. But the depletion mode pullup transistor's gate is connected to its source, and thus it is always turned on. Hence, when the enhancement mode transistor is turned off, for example by connecting zero voltage to its gate, the output of the inverter will be equal to VDD. We will find that for reasonable ratios of the gate geometries of the two transistors, input voltages above a defined logic threshold voltage, V_{inv}, will produce output voltages below that logic threshold voltage, and vice versa.

The top view of the layout of an inverter on the silicon surface is sketched in figure 2c. It

consists of two polysilicon regions overhanging a path in the diffusion level which runs between VDD and GND. This arrangement forms the two MOS transistors of the inverter. The inverter input A is connected to the poly forming the gate of the lower of the two transistors. The pullup is formed by connecting the gate of the upper transistor to its source. The fabrication details of such connections are described in chapter 2. The output of the inverter is shown emerging on the diffusion level, from between the drain of the pulldown and the source of the pullup. The pullup is a depletion mode transistor, and it is usually several times longer than the pulldown, to achieve the proper inverter logic threshold.

Figures 3a and 3b show the characteristics of a typical pair of MOS transistors used to implement an inverter. The relative locations of the saturation regions of the pullup and pulldown differ in these characteristics, due to the difference in their threshold voltages.

We can use a graphical construct to determine the actual transfer characteristic, V_{out} vs V_{in}, of the inverter circuit. From figure 2b we see that the V_{ds}(enh) of the enhancement mode transistor equals VDD minus V_{ds}(dep) of the depletion mode transistor. Also, V_{ds}(enh) equals V_{out}. In a steady state and with no current drawn from the output, the I_{ds} of the two transistors are equal. Since the pullup has its gate connected to its source, only one of its characteristic curves is relevant, namely the one for V_{gs}(dep) = 0. Taking these facts into account, we begin the graphical solution (fig. 3c) by superimposing plots of I_{ds}(enh) vs V_{ds}(enh), and the one plot of I_{ds}(dep) vs [VDD - V_{ds}(dep)]. Since the currents in both transistors must be equal, the intersections of these sets of curves yields V_{ds}(enh) = V_{out} versus V_{gs}(enh) = V_{in}. The resulting transfer characteristic is plotted in figure 3d.

Studying figures 3c and 3d, consider the effect of starting with V_{in} = 0 and then gradually increasing V_{in} towards VDD. While the input voltage is below the threshold of the pulldown transistor, no current flows in that transistor, the output voltage is constant at VDD, and the drain to source voltage across the pullup transistor is equal to zero. When V_{in} is first increased above the enhancement mode threshold, current begins to flow in the pulldown transistor. The output voltage decreases slowly as the input voltage is first increased above V_{th}. Subsequent increases in the input voltage rapidly lower the pulldown's drain to source voltage, until the point is reached where the pulldown leaves its saturation region and becomes resistive. Then as V_{in} continues to increase, the output voltage asymptotically approaches zero. The input voltage at which V_{in}

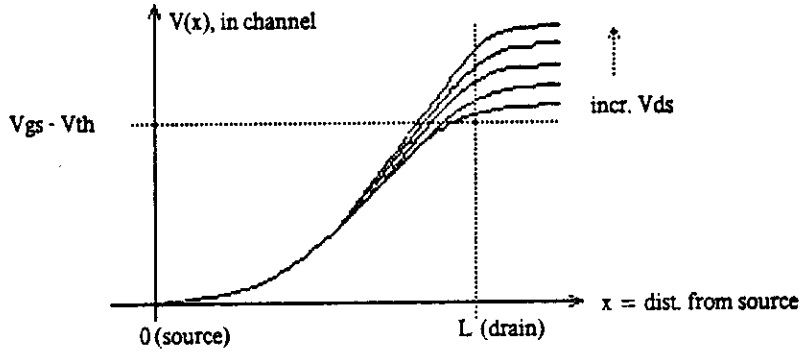


Fig. 1e. Voltage Profile Across Channel

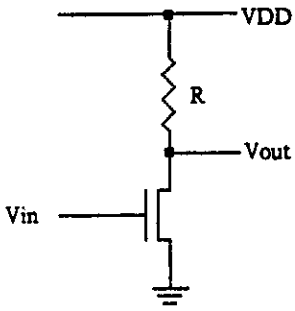


Fig. 2a. An Inverter

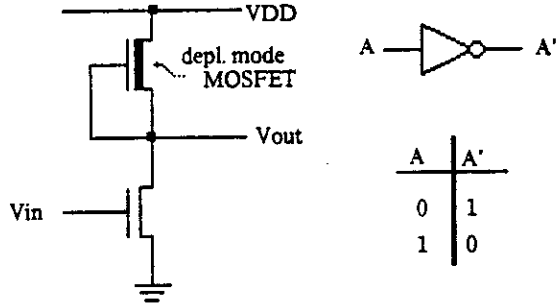


Fig. 2b. The Basic Inverter Circuit Diagram, Logic Symbol, Logic Function

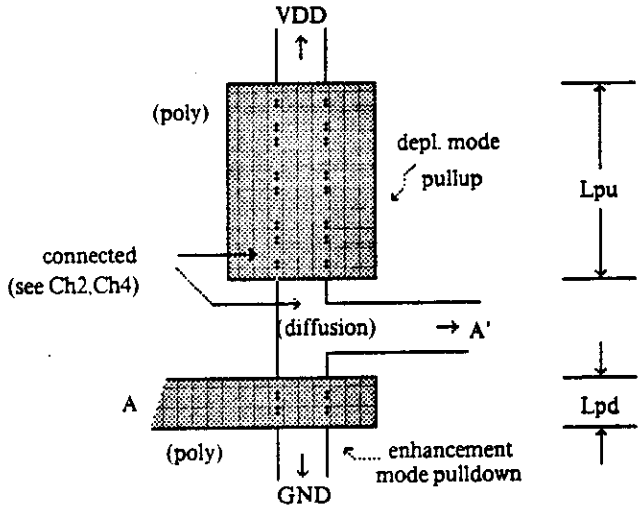


Fig. 2c. Basic Inverter Layout

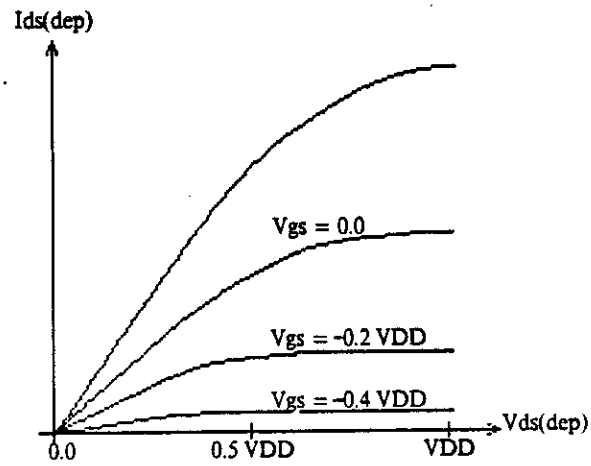


Fig. 3a. Inverter Pullup Characteristics

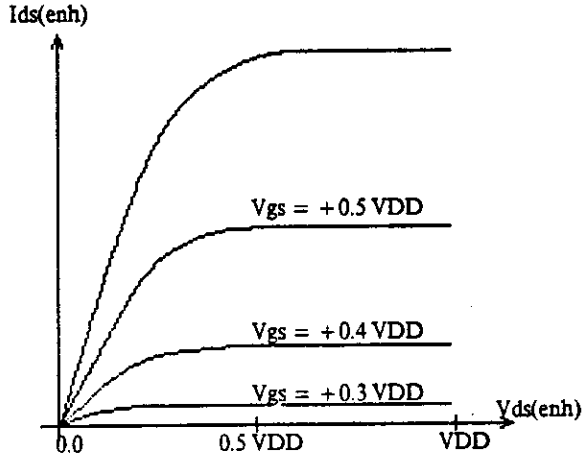


Fig. 3b. Inverter Pulldown Characteristics

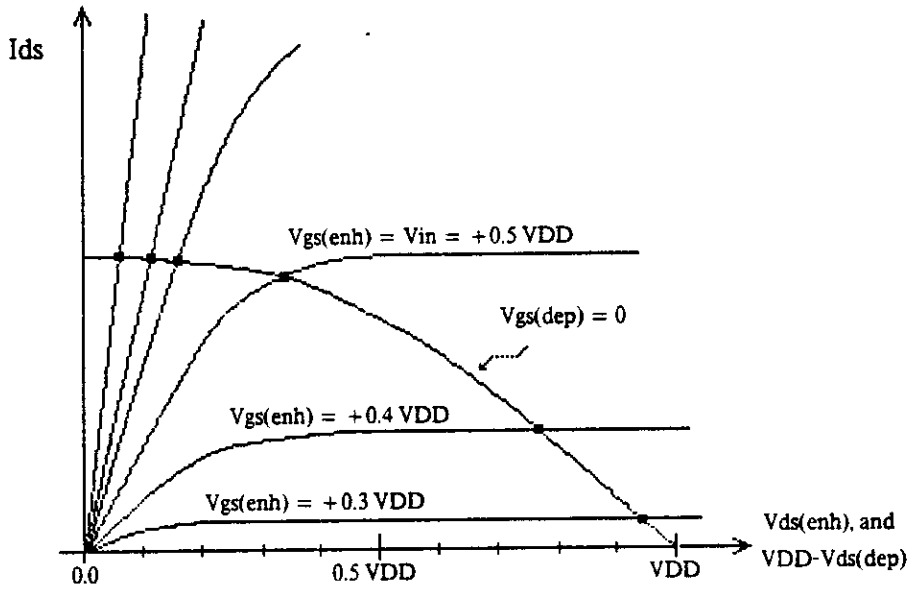


Fig. 3c. Ids(enh) vs Vds(enh), and Ids(dep) vs [VDD - Vds(dep)]

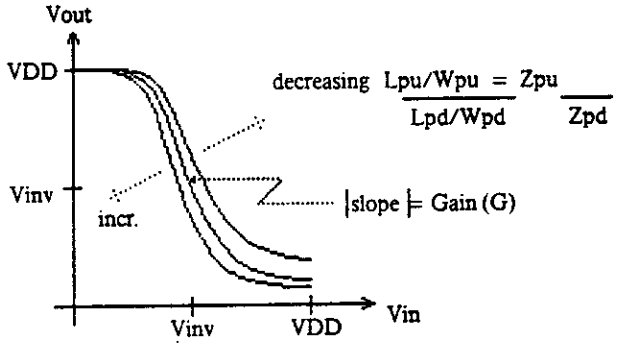


Fig. 3d. Vout vs Vin for the Basic Inverter

equals V_{out} is known as the logic threshold voltage V_{inv}. Figure 3d also shows the effect of changes in the transistor length to width ratios on the transfer characteristics and on the logic threshold voltage. The resistive impedance of the MOS transistor is proportional to the length to width ratio Z of its gate region. Using the subscript pu for the pullup transistor and pd for the pulldown transistor: If $Z_{pu} = L_{pu}/W_{pu}$ is increased relative to $Z_{pd} = L_{pd}/W_{pd}$, then V_{inv} decreases, and vice-versa. The gain, or negative slope of the transfer characteristic near V_{inv}, increases as Z_{pu}/Z_{pd} increases. The gain G must be substantially greater than unity for digital circuits to function properly.

Inverter Logic Threshold Voltage

The most fundamental property of the basic inverter circuit is its logic threshold voltage, V_{inv}. The logic threshold here is *not* the same as V_{th} of the enhancement mode transistor, but is that voltage on the input of the enhancement mode transistor which causes an equal output voltage. If V_{in} is increased above this logic threshold, V_{out} falls below it, and if V_{in} is decreased below V_{inv}, V_{out} rises above it. The following simple analysis assumes that both pullup and pulldown are in saturation, so that equation 3 applies. Usually the pullup is not quite in saturation, but the following is still nearly correct. V_{inv} is approximately that input voltage which would cause saturation current through the pulldown transistor to be equal to saturation current through the pullup transistor. Referring to eq.5, we find the condition for equality of the two currents given in eq.6.

Currents equal when: $$\frac{Wpd}{Lpd}(V_{inv} - V_{th})^2 = \frac{Wpu}{Lpu}(-V_{dep})^2 , \qquad \text{[eq.6]}$$

or thus when: $$V_{inv} = V_{th} - V_{dep}/[Zpu/Zpd]^{1/2} \qquad \text{[eq.6a]}$$

Here we note that the current through the depletion mode transistor is dependent only on its geometry and threshold voltage V_{dep}, since its $V_{gs} = 0$. Note that V_{inv} is dependent upon the thresholds of both the enhancement and depletion mode transistors, and also the square root of the ratio of the Z = L/W of the enhancement mode transistor to that of the depletion mode transistor.

Tradeoffs are possible between these threshold voltages and the areas and current driving capability of transistors in the system's inverters. To maximize (V_{gs} - V_{th}) and increase the pulldowns' current driving capability for a given area, V_{th} should be as low as possible. However, if V_{th} is too low, inverter outputs won't be driveable below V_{th}, and inverters won't be able to turn off transistors used as simple switches. The original choice of V_{th} ~ 0.2VDD is a reasonable compromise here.

Similarly, to maximize the current driving capability of pullups of given area, we might set the system's V_{dep} as far negative as possible. However, eq. 6a shows that for chosen V_{inv} and V_{th}, decreasing V_{dep} requires an increase in L_{pu}/W_{pu}, typically leading to an increase in pullup area. The compromise made in this case is usually as follows. The negative threshold of depletion mode transistors is set during fabrication such that with gate tied to source, they turn on approximately as strongly as would an enhancement mode transistor with VDD connected to its gate and its source grounded. In other words, depletion mode and enhancement mode transistors of equal gate dimensions would have equal drain to source currents under those conditions. Applying eq.6 in those conditions we find that:

$$(-V_{dep})^2 \sim (VDD - V_{th})^2.$$

Therefore, $-V_{dep} \sim (VDD - V_{th})$, and $V_{dep} \sim -0.8VDD$. While adjustments in the details of this choice are often made in the interest of optimization of processes for a particular product, we will assume here this approximate equality of turn-on voltages of the two transistor types for the sake of simplicity. Substituting this choice of V_{dep} into eq.6a, we find that for V_{th} small compared to VDD:

$$V_{inv} \sim VDD/[Zpu/Zpd]^{1/2} \qquad \text{[eq.7]}$$

In general it is desirable that the margins around the inverter threshold be approximately equal, i.e., that the inverter threshold, V_{inv}, lie approximately midway between VDD and ground. We see from eq.7 that this criterion is met by a ratio of pullup Z to pulldown Z of approximately 4:1. We will see later that the choice of $V_{dep} \sim VDD - V_{th}$, producing a ratio of 4:1 here, will lead to a balancing of performances in certain other important circuits.

Inverter Delay

A minimum requirement for an inverter is that it drive another identical to itself. Let us analyze the delay through a string of inverters of identical dimensions. This is the simplest case in which we can estimate performance. Inverters connected in this way are shown in Fig. 4a. We define the inverter ratio k as the ratio of Z of the pullups to Z of the pulldowns. We will sometimes use the alternative "resistor with gate" pullup symbol, as in fig.4a, to clarify its functional purpose.

Let us assume that prior to t = 0, the voltage at the input of the first inverter is zero, and hence, the voltage output of the second inverter will be low. At time t=0, let us place a voltage equal to VDD on the input of the first inverter and follow the sequence of events which follows. The output of the first inverter, which leads to the gate of the second inverter, will initially be at VDD. Within approximately one transit time, the pulldown transistor of the first inverter will remove from this node an amount of charge equal to VDD times the gate capacitance of the pulldown of the second inverter. The pullup transistor of the second inverter is now faced with the task of supplying a similar charge to the gate of the third inverter, to raise it to VDD. Since it can supply at most only 1/k'th of the current that can be supplied by the pulldown transistor, the delay in the second inverter stage is approximately k times that of the first.

It is thus convenient to speak of the *inverter pair delay* which includes the delay for one lowgoing transition and one highgoing transition. This inverter pair delay is approximately (1+k) times the transit time, as shown in figure 4a. The fact that the rising transition is slower than the falling transition by approximately the inverter transistors' geometry ratios is an inherent characteristic of any ratio type logic. It is not true of all logic families. For example, in families such as complementary MOS where there are both pMOS and nMOS devices on the same silicon chip and both types operate strictly as pulldown enhancement mode devices, any delay asymmetry is a function of the difference in mobilities of the p and n type charge carriers rather than of the transistor geometrical ratios.

Fig. 4b shows an inverter driving the inputs of several other inverters. In this case, for a fanout factor f, it is clear that in either the pullup or pulldown direction, the active device must supply f times as much charge as it did in the case of driving a single input. In this case, the delay both in the up and downgoing directions is increased by approximately the factor f. In the case of the

downgoing transition, the delay is approximately f times the transit time of the pulldown transistor, and in the case of the upgoing transition, the delay is approximately the inverter ratio k times the fanout factor times the pulldown transit time.

In the discussions of transit time given earlier, it was assumed that both the depletion mode pullup device and the enhancement mode pulldown device were operating in the resistive region. It was also assumed that all capacitances were constant, and not a function of voltage. These conditions are not strictly met in the technology we are discussing. Delay calculations given in this text are based on a "switching model" where individual stages spend a small fraction of their time in the mid-range of voltages around V_{inv}. This assumption introduces a small error of the order of 1/G. Because of these and other second order effects, the switching times actually observed vary somewhat from those derived.

Parasitic Effects

In integrated systems, capacitances of circuit nodes are due not only to the capacitance of gates connected to the nodes, but also include capacitances to ground of signal paths connected to the nodes and other stray capacitances. These other capacitances, sometimes called parasitic or stray capacitances, are not negligible. While gate capacitances are typically an order of magnitude greater per unit area than the capacitances of the signal paths, the signal paths are often much larger in area than the associated gate regions. Therefore, a substantial fraction of the delay encountered may be accounted for by stray capacitance rather than by the inherent properties of the active transistors. In the simplest case where the capacitance of a node is increased by the presence of parasitic area attached to the node, the delays can be accounted for by simply increasing the transit time by the ratio of the total capacitance to that of the gate of the transistor being driven. Time is required to supply charge not only to the gate itself but also to the parasitic capacitance.

There is one type of parasitic, however, which is not accounted for so simply. All MOS transistors have a parasitic capacitance between the drain edge of the gate and the drain node. This effect is shown schematically in figure 4c. In an inverter string, this capacitance will be charged in one direction for one polarity of input, and in the opposite direction for the opposite polarity input. Thus, on a gross scale its effect on the system is twice that of an equivalent parasitic capacitance to ground. Therefore, gate to drain capacitances should be approximately

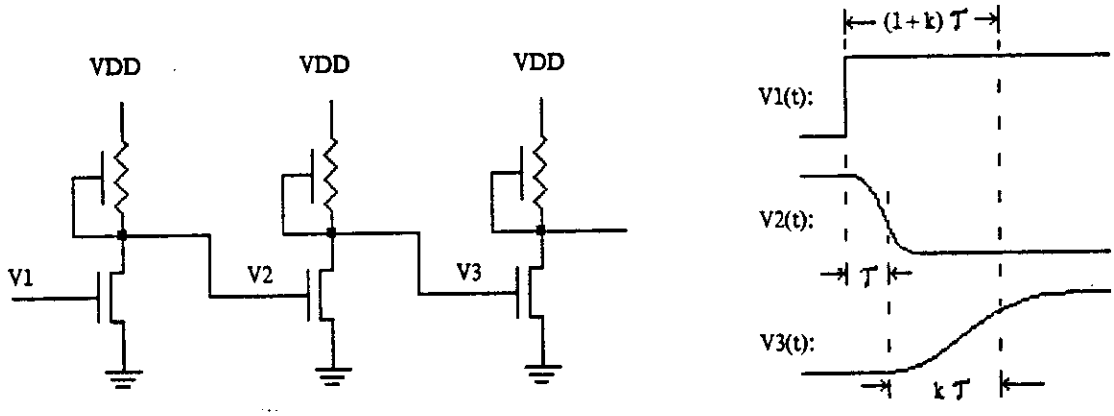
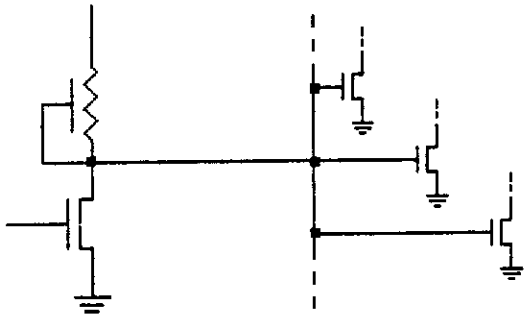


Fig. 4a. Inverter Delay

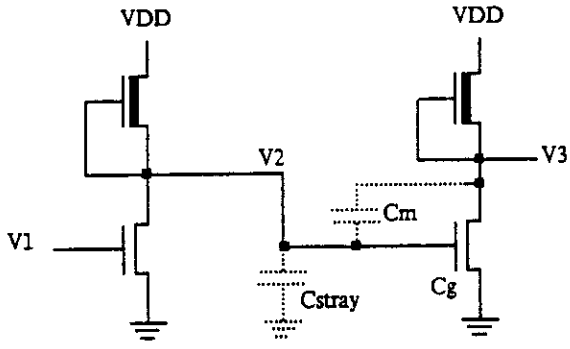


For fanout of f:

down delay ~ f T

up delay ~ kf T

Fig. 4b. Fanout



Start:

V2 = 0, V3 = VDD

Qg = 0, Qm = -CmVdd

Qwire = 0

Finish:

V2 = VDD, V3 = 0

Qg = CgVDD, Qm = +CmVDD

Qwire = CstrayVDD

Fig. 4c. The Miller Effect

$$\text{total effective input capacitance} = \frac{dQ}{dV} = \frac{(Q\text{finish} - Q\text{start})}{(V\text{finish} - V\text{start})} = Cg + 2Cm + Cstray$$

doubled, and added to the gate capacitance C_g and the stray capacitances, to account for the total capacitance of the node and thus for the effective delay time of the inverter. The effective inverter pair delay then is equal to $\tau(1+k)C_{total}/C_g$.

Driving Large Capacitive Loads

As we have seen, the delay per inverter stage is multiplied by a fanout factor. The overall performance of a system may be seriously degraded if it contains any large fanouts, where one circuit within the system is required to drive a large capacitive load. As we shall see, this situation often occurs in the case of control drivers required to drive a large number of inputs to memory cells or logic function blocks. A similar and more serious problem is driving wires which go off the silicon chip to other chips or input/output devices. In such cases the ratio of the capacitance which must be driven to the inherent capacitance of a gate circuit on the chip is often many orders of magnitude, causing a serious delay and a degradation of system performance.

Consider how we may drive a capacitive load C_L in the minimum possible time given that we are starting with a signal on the gate of an MOS transistor of capacitance C_g. Define the ratio of the load capacitance to the gate capacitance, C_L/C_g, as Y. It seems intuitively clear that the optimum way to drive a large capacitance is to use our elementary inverter to drive a larger inverter and that larger inverter to drive a still larger inverter until at some point the larger inverter is able to drive the load capacitance directly. Using an argument similar to the fanout argument it is clear that for one inverter to drive another inverter, where the second is larger in size by a factor of f, results in a delay f times the inherent inverter delay, τ. If N such stages are used, each larger than the previous by a factor f, then the total delay of the inverter chain is $Nf\tau$, where f^N equals Y. Note that if we use a large factor f, we can get by with few stages, but each stage will have a long delay. If we use a smaller factor f, we can shorten the delay of each stage, but are required to use more stages. What value of N minimizes the overall delay for a given Y? We compute this value as follows:

Since $f^N = Y$, $\ln(Y) = N \ln(f)$

Delay of one stage $= f\tau$,

Thus total delay is $= Nf\tau = \ln(Y)[\, f/\ln(f)\,]\tau$ [eq.8]

Notice that the delay is always proportional to ln(Y), a result of the exponential growth in successive stages of the driver. The multiplicative factor, f/ln(f), is plotted as a function of f in figure 5, normalized to its minimum value (e). Total delay is minimized when each stage is larger than the previous one by a factor of e, the base of natural logarithms. Minimum total delay is the elementary inverter delay τ times e times the natural logarithm of the ratio of the load capacitance to the elementary inverter capacitance.

$$\text{Min. total delay} \quad \sim \quad \tau e[\ln(C_L/C_g)] \qquad \text{[eq.9]}$$

Minimum delay through the driver is seldom the only design criterion. The relative time penalty introduced by the choice of other values of f can be read directly from figure 5.

Space vs Time

From the results of the sections on inverter delay, parasitic effects, and driving large capacitances, we see that areas and distances on the silicon surface trade off against delay times. For an inverter to drive another inverter some distance away, it must charge not only the gate capacitance of the succeeding inverter but also the capacitance to ground of the signal path connecting the two. Increasing the distance between the two inverters will therefore increase the inverter pair delay. This effect can be counterbalanced by increasing the area of the first inverter, so as to reduce the ratio of the load capacitance to the gate capacitance of the first inverter. But the delay of some previous driving stage is then increased. There is no way to get around the fact that transporting a signal from one node to another some distance away requires charging or discharging capacitance, and therefore takes time. Note that this is not a velocity of light limitation as is often the case outside the chip. The times are typically several orders of magnitude longer than those required for light to traverse the distances involved. To minimize both the time and space required to implement system functions, we will tend to use the smallest possible circuits and locate them in ways which tend to minimize the interconnection distances

The results of a previous section can be used here to illustrate another interesting space vs time effect. Suppose that the minimum size transistors of an integrated system have a transit time τ and gate capacitance C_g. A minimum size transistor within the system produces a signal which is then passed through successively larger inverting logic stages and eventually drives a large

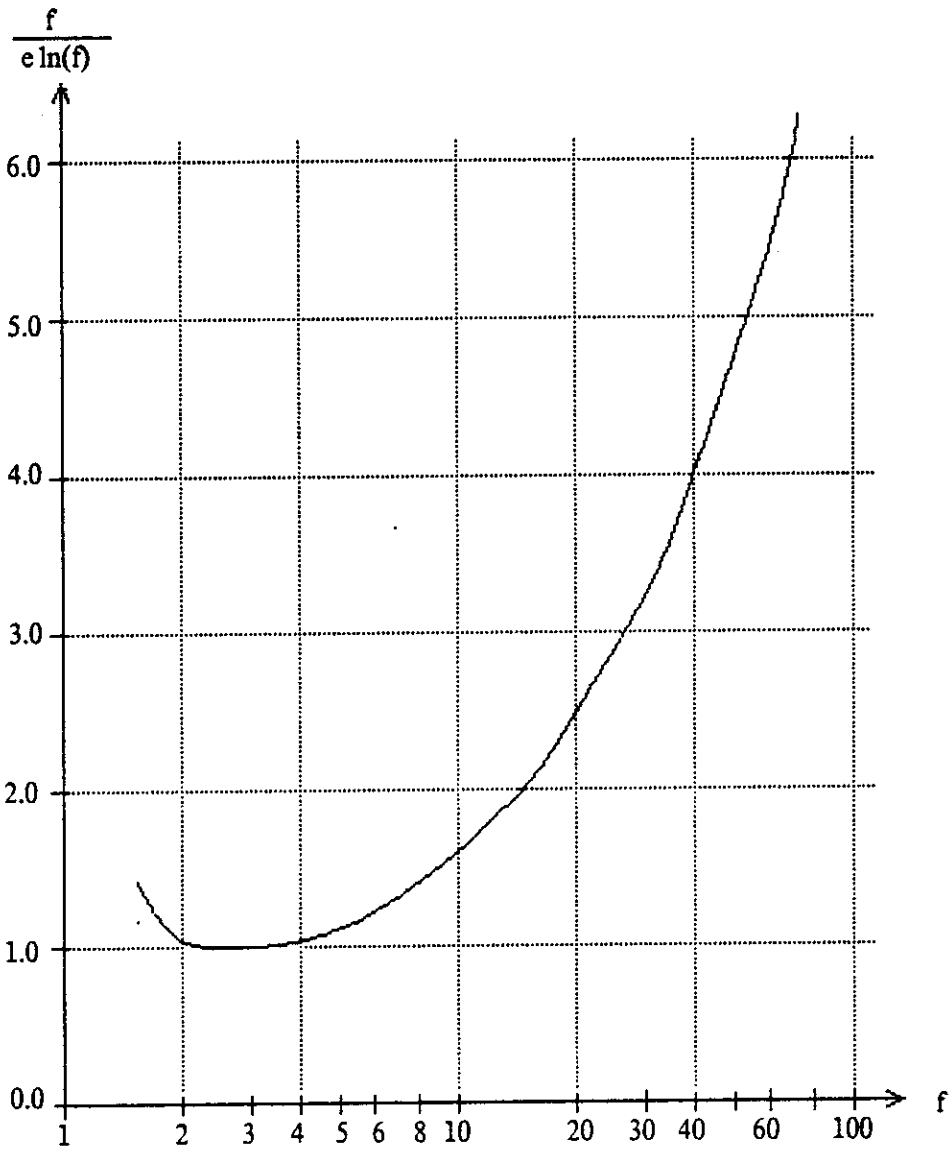


Fig. 5. Relative Time Penalty, $\frac{f}{e \ln(f)}$, vs Size Factor f

capacitance C_L with minimum total delay equal to t_{min}. With the passage of time, fabrication technology improves. We replace the system with one in which all circuit dimensions, including those vertical to the surface, are scaled down in size by dividing by a factor α , and the values of V_{dd} and V_{th} are also scaled down by dividing by α. The motivation for this scaling is clear: the new system may contain α^2 as many circuits. As described in a later section, we will find that the transit times of the smallest circuits will now be $\tau' = \tau/\alpha$, and their gate capacitance will be $C_g' = C_g/\alpha$. The new ratio of load to minimum gate capacitance is: $Y' = \alpha Y$. Referring to equation 9., we find that the new minimum total delay, t_{min}', to drive C_L scales as follows:

$$t_{min}' = t_{min} \ (1/\alpha) \ [1+(\ln\alpha/\ln Y)]$$

Therefore, as the inverters scale down and τ gets smaller, more inverting logic stages are required to obtain the minimum offchip delay. Thus the relative delay to the outside world becomes larger. However, the absolute delay becomes smaller.

Basic NAND and NOR Logic Circuits

NAND and NOR logic circuits may be constructed in integrated systems as simple expansions of the basic inverter circuit. The analysis of the behavior of these circuits, including their logic threshold voltages, transistor geometry ratios and time delays, is also a direct extension of the analysis of the basic inverter.

The circuit layout diagram of a two input NAND gate is shown in figure 6a. The layout is that of a basic inverter with an additional enhancement mode transistor in series with the pulldown transistor. NAND gates with more inputs may be constructed by simply adding more transistors in series with the pulldown path. The electrical circuit diagram, truth table and logic symbol for the two input NAND gate are shown in figure 6b. If either of the inputs A or B is a logic-0, the pulldown path is open and the output will be high, and therefore a logic-1. For the output to be driven low, to logic-0, both inputs must be high, at logic-1. The logic threshold voltage of this NAND gate is calculated in a similar manner to that of the basic inverter, except equation 7 is rewritten with the length of the pulldowns replaced with the sum of the lengths of the two pulldowns (assuming their widths are equal) as follows:

$$V_{thNAND} \sim VDD/[(Lpu/Wpu)/((Lpd_a + Lpd_b)/Wpd)]^{1/2}$$

This equation indicates that as pulldowns are added in series to form NAND gate inputs, the pullup length must be enlarged to hold the logic threshold voltage constant.

The logic threshold voltage of an n-input NAND gate, assuming all the pulldowns have equal geometries, is:

$$V_{thNAND} \sim VDD/[(Lpu/Wpu)/(nLpd/Wpd)]^{1/2}$$

As inputs are added and pullup length is increased, the delay time of the NAND gate is also correspondingly increased, both for rising and falling transitions.

$$\tau_{NAND} \sim n\tau_{inv}$$

The circuit layout diagram of a two input NOR gate is shown in figure 6c. The layout is that of a basic inverter with an additional enhancement mode transistor in parallel with the pulldown transistor. Additional inputs may be constructed by simply placing more transistors in parallel with the pulldown path. The circuit diagram, truth table and logic symbol for the two input NOR gate are shown in figure 6d. If either of the inputs A or B is a logic-1, the pulldown path to ground is closed and the output will be low, and therefore a logic-0. For the output to be driven high, to logic-1, both inputs must be low, at logic-0. If one of its inputs is kept at logic-0, and the other swings between logic-0 and logic-1, the logic threshold voltage of the NOR gate is the same as that of a basic inverter of equal pullup to pulldown ratio. If this ratio were 4:1 to provide equal margins, then $V_{thNOR} \sim VDD/2$ with only one input active. However, if both pulldowns had equal geometries, and if both inputs were to move together between logic-0 and logic-1, V_{thNOR} would be reduced to $\sim VDD/(8)^{1/2}$. The logic threshold voltage of an n-input NOR circuit decreases as a function of the number of active inputs (inputs moving together from logic-0 to logic-1). The delay time of the NOR gate with one input active is the same as that of an inverter of equal transistor geometries, except for added stray capacitance. Its delay time for falling transitions is decreased as more of its inputs are active.

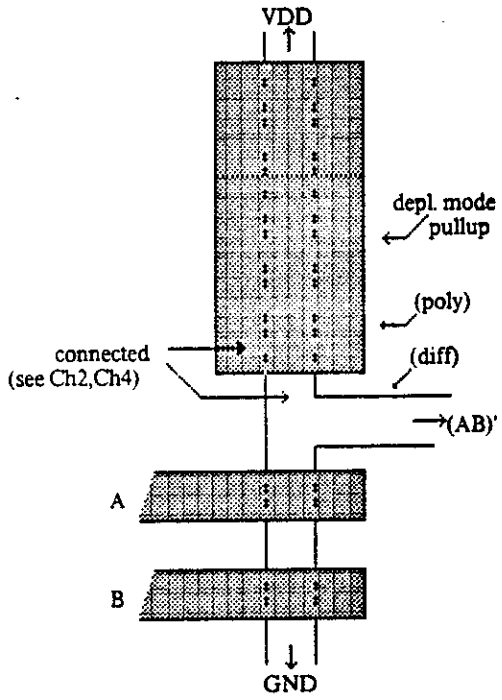
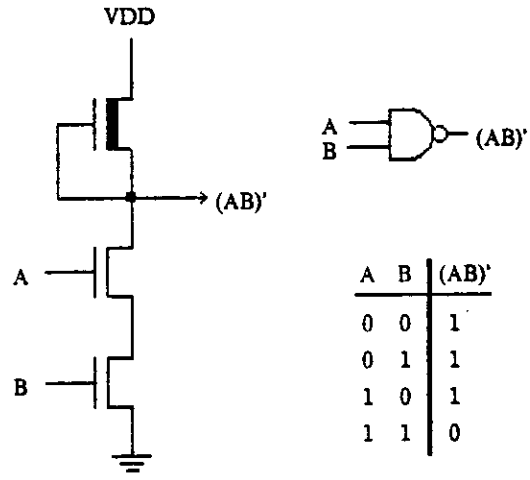


Fig. 6a. NAND Gate
[top view of layout]



A	B	(AB)'
0	0	1
0	1	1
1	0	1
1	1	0

Fig. 6b. NAND Gate Circuit Diagram,
Logic Symbol, Logic Function

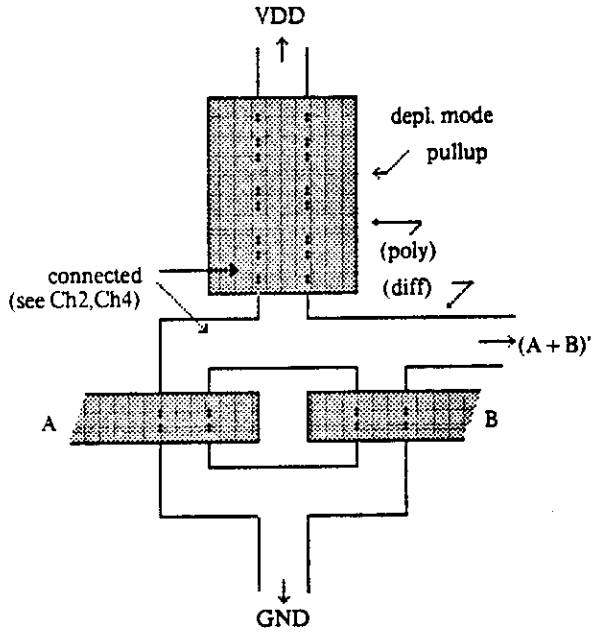
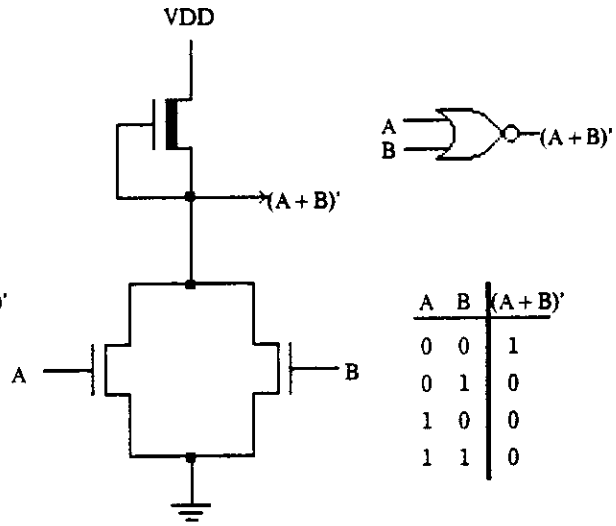


Fig. 6c. NOR Gate
[top view of layout]



A	B	(A + B)'
0	0	1
0	1	0
1	0	0
1	1	0

Fig. 6d. NOR Gate Circuit Diagram,
Logic Symbol, Logic Function

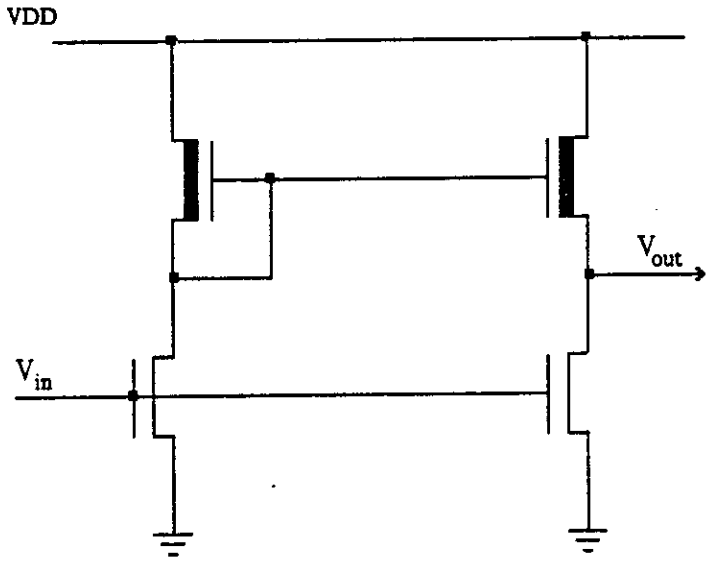


Fig. 7a. Inverting Super Buffer

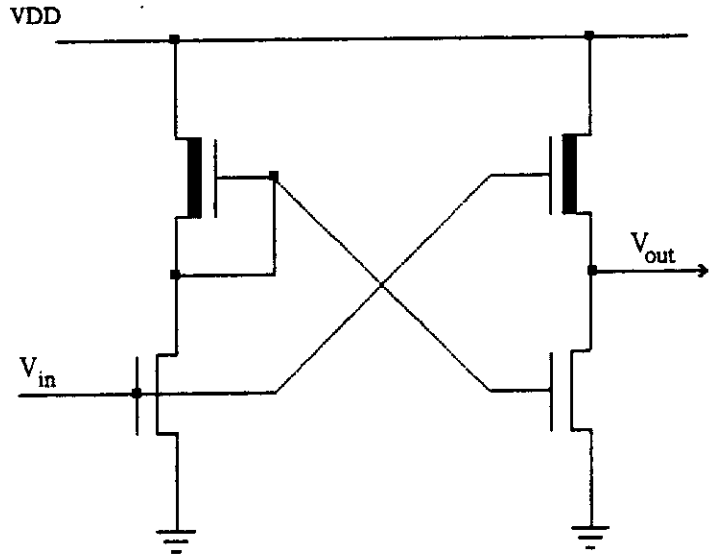


Fig. 7b. Non-Inverting Super Buffer

Super Buffers

As we have noted, ratio type logic suffers from an asymmetry in its ability to drive capacitive loads. This asymmetry results from the fact that the pullup transistor has of necessity less driving capability than the pulldown transistor. There are, however, methods for avoiding this asymmetry. Shown in figures 7a and 7b are circuits for inverting and non-inverting drivers which are approximately symmetrical in their capability of sourcing or sinking charge into a capacitive load. Drivers of this type are called *super buffers.*

Both types of super buffer are built using a depletion mode pullup transistor and an enhancement mode pulldown transistor, with a ratio of Z's of approximately 4:1 as in the basic inverter. However, the gate of the pullup transistor, rather than being tied to its source, is tied to a signal which is the complement of that driving the pulldown transistor.

When the pulldown transistor gate is at a high voltage, the pullup transistor gate will be approximately at ground, and the current through the super buffer will be similar to that through a standard inverter of the same size. However, when the gate of the pulldown transistor is put to zero, the gate of the pullup transistor will go rapidly to VDD since it is the only load on the output of the previous inverter, and the depletion mode transistor will be turned on at approximately twice the drive which it would experience if its gate were tied to its source. Since the current from a device in saturation goes approximately as the square of the gate voltage, the current sourcing capability of a super buffer is approximately four times that of a standard inverter. Hence, the current sourcing capability of its pullups is approximately equal to the current sinking capability of its pulldowns, and wave forms from super buffers driving capacitive loads are nearly symmetrical.

The effective delay time, τ, of super buffers is thus reduced to approximately the same value for highgoing and lowgoing wave forms. Needless to say, when large capacitive loads are to be driven, super buffers are universally used. The arguments used in the last section to determine how many stages are used to drive a large capacitive load from a small source apply directly to super buffers. For that reason we have not explicitly indicated an inverter ratio k in that section.

A Closer Look at the Electrical Parameters

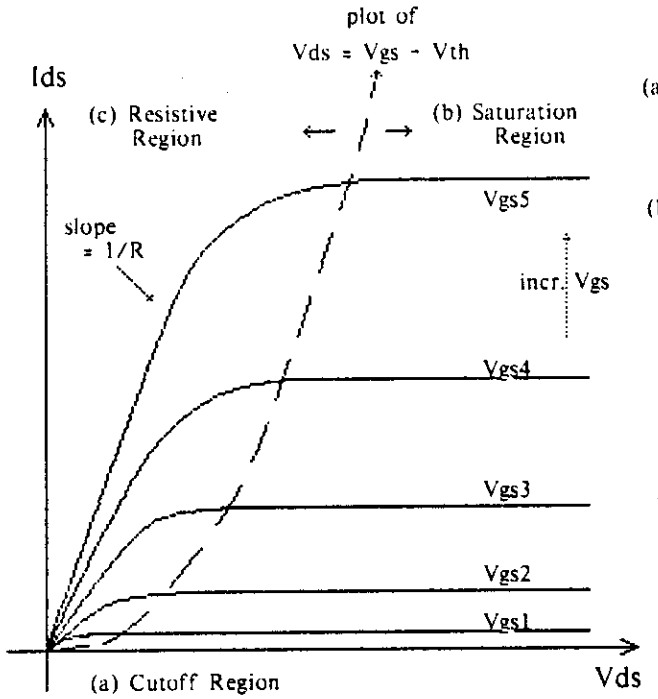
Up to this point we have talked in very simple terms about the properties of the MOS transistors. They have a capacitance associated with their gate input and a transit time for electrons to move from the source to the drain. We have given simple expressions for the drain to source current. For very low V_{ds}, the MOS transistor's drain to source path acts as a resistor whose conductance is directly proportional to the gate voltage above threshold, as given in equation 3. For values of V_{ds} larger than $V_{gs} - V_{th}$, the device acts as a current source, with a current proportional to $(V_{gs} - V_{th})^2$, as given in equation 5. As V_{ds} passes through the intermediate range between these two extremes, there is a smooth transistion between the two types of behavior[1], as given in the following equation:

$$I_{ds} = Q/\tau = \mu C_g[(V_{gs} - V_{th})\, V_{ds} - (V_{ds}^2)/2]/L^2 \qquad \text{[eq.10]}$$

Figure 9a plots I_{ds} vs V_{ds}, summarizing the various regions of MOS transistor operation.

There is another electrical characteristic we may occasionally have to take into account. The threshold voltage of an MOS transistor is not a constant, but varies slightly as a function of the voltage between the source terminal of the transistor and the silicon substrate, which is usually at ground. This so called *body effect* is illustrated in figure 9b. If the source to bulk (substrate) voltage, V_{sb} , equals zero, then V_{th} is at its minimum value of approximately 0.2 VDD. As V_{sb} is increased, V_{th} increases slightly.

For enhancement mode transistors fabricated using typical processes, V_{th} reaches a maximum value of about 0.3 VDD when V_{sb} is increased to ~ VDD. The value of the depletion mode transistor threshold, V_{dep}, is similarly affected, ranging from about -0.8 VDD to -0.7 VDD as V_{sb} is raised from zero to VDD volts. As shown in figure 9b, it is possible to insert a fixed bias voltage between the circuit ground and the substrate, rather than just connect them. Such a *substrate bias* provides an electrical mechanism for setting the threshold to an appropriate value.



(a) Cutoff Region:

Vgs < Vth , Ids = 0

(b) Saturation Region:

Vgs ≥ Vth , Vds sufficiently high so Vgd < Vth . (i.e.: Vds ≥ Vgs-Vth)

MOSFET acts as current source, with Ids proportional to $(Vgs - Vth)^2$

(c) Resistive Region:

Vgs ≥ Vth , Vds sufficiently low so Vgd ≥ Vth . (i.e.: Vds < Vgs - Vth)

MOSFET acts as resistor, with resistance inversely proportional to (Vgs - Vth)

Fig. 9a. Summary of MOS Transistor Characteristics

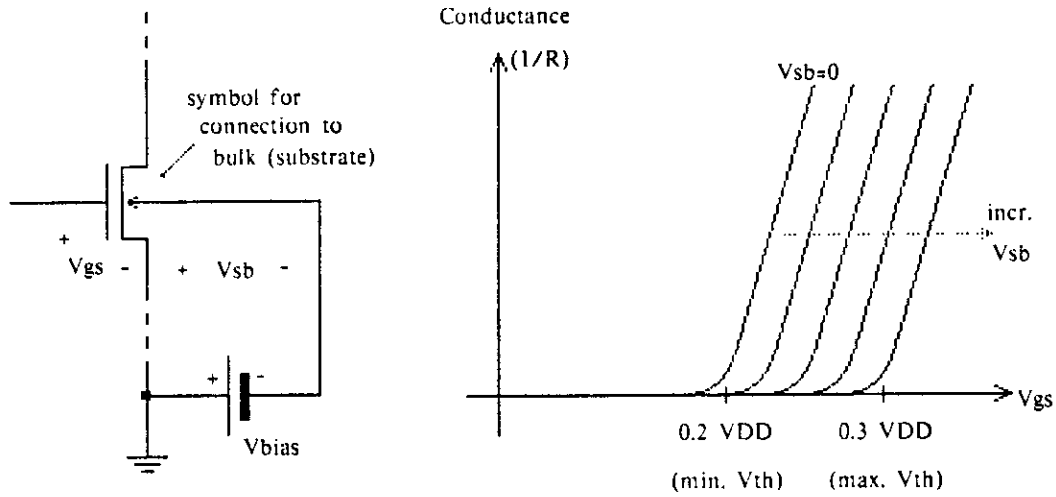


Fig. 9b. The Body Effect

Depletion Mode vs Enhancement Mode Pullups

With its gate tied to VDD, an enhancement mode transistor will be on for all $V_{ds} > V_{th}$, and thus can be used for a pullup device in inverting logic circuits. Early MOS processes used pullup devices of exactly this type.

In this section we will make a comparison of the rising transients of the two types of pullup circuits. As noted earlier, rising transients in ratio type logic are usually slower than falling transients, and thus rising transients generally have greater impact on system performance. In the simplest cases, this asymmetry in the transients results from the current sourcing capability of the pullup transistor being less than that of its pulldown counterpart. The simple intuitive time arguments given earlier are quite adequate for making estimates of system performance in most cases. However, there are situations in which the transient time may be much longer than a naive estimate would indicate. The rising transient of the enhancement mode pullup is one of these.

A depletion mode pullup transistor feeding a capacitive load is shown schematically in figure 10a. Since $V_{gs} \geq V_{th}$ and $V_{gd} \geq V_{th}$, the pullup transistor is in the resistive region. The final stages of the rising transient are given by the following exponential:

$$V(t) = VDD[1 - e^{-t/(RC_L)}]$$

For an inverter ratio k, pulldown transit time τ, and gate capacitance C_g, the time-constant of the rising transient is given by:

$$RC_L = k\tau C_L/C_g$$

A somewhat more complicated situation is presented by an enhancement mode transistor sourcing charge into a capacitive load. This situation is shown schematically in Fig. 10b. Note that since $V_{gd} = 0$, the transistor is in saturation whenever $V_{gs} > V_{th}$. The problem with sourcing charge from the enhancement mode transistor is that as the voltage at the output node gets closer and closer to one threshold below VDD, the amount of current provided by the enhancement mode transistor decreases rapidly.

The dependence of the enhancement mode pullup current, I_{ds}, upon output voltage, V, is given in equation 11:

$$Q = - \frac{\varepsilon WL}{D} [(VDD - V_{th}) - V]$$

$$\tau = 2L^2/\mu[(VDD - V_{th}) - V]$$

$$I_{ds} = - Q/\tau = \frac{\mu\varepsilon W}{2LD} [(VDD - V_{th}) - V]^2 \qquad \text{[eq.11]}$$

The fact that the pullup current decreases as the output voltage nears its maximum value causes the rising transient from such a circuit to be of qualitatively different form than that of a depletion mode pullup. Equating $I_{ds} = C_L dV/dt$ with the expression in equation 11, and then solving for V(t), we find the rising voltage transient, for large t:

$$V(t) = VDD - V_{th'} - C_L \frac{LD}{\mu\varepsilon Wt} \qquad \text{[eq.12]}$$

Note that in this configuration, the threshold voltage Vth' of the pullup is near its maximum value as V(t) rises towards VDD, due to the body effect.

A comparison of the rising transients of the preceding two circuits, assuming the same load capacitance and the same pullup source current at zero output voltage, is shown in Fig. 10c. The rising transient for the depletion mode pullup transistor is crisp and converges rapidly towards VDD. However, the rising transient for the enhancement mode pullup transistor, while starting rapidly, lags far behind and within the expected time response of the system, never even comes close to one threshold below VDD. Even for very large t, $V(t) < VDD - V_{th'}$.

The practical effect of this property of enhancement mode transistors is that circuits designed to work from the output of such a circuit should be designed with an inverter threshold V_{inv} considerably lower than that of circuits designed to work with the output of a depletion mode pullup circuit. In order to obtain equal inverter margins without sacrificing performance, we will normally use depletion mode pullups.

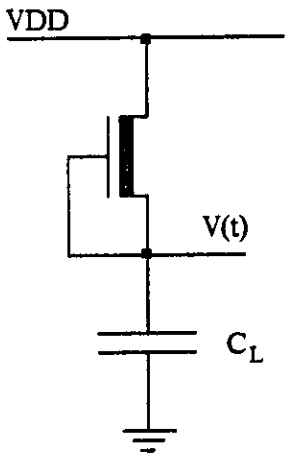


Fig. 10a. Depletion Mode MOSFET Pulling Up Capacitive Load

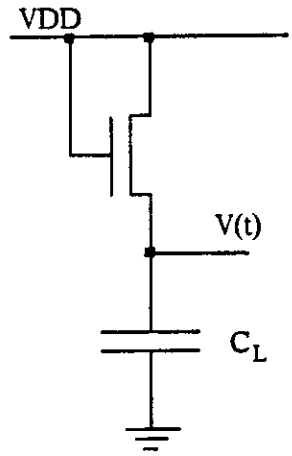


Fig. 10b. Enhancement Mode MOSFET Pulling Up Capacitive Load

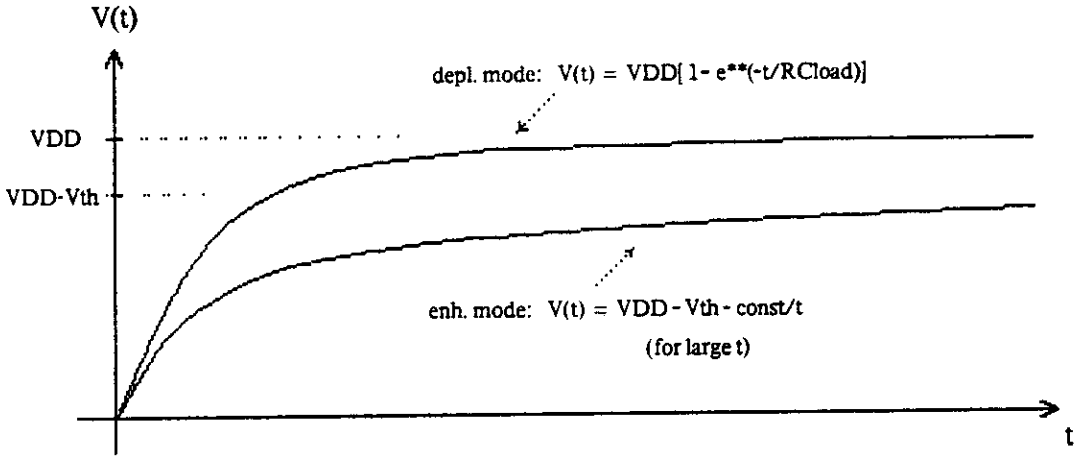


Fig. 10c. Comparisons of Rising Transients for the Two Types of Pullups

Delays in Another Form of Logic Circuitry

Enhancement mode transistors, when used in small numbers and driving small capacitive loads, may often be used as switches in circuits of simple topology to provide logic signal steering functions of much greater complexity than could be easily achieved in ratio type inverting logic. These circuits are reminiscent of relay switching logic, and transistors used in this way are referred to as "pass transistors" or "transmission gates". Example circuits using this type of design are given in Chapter 3. A particularly interesting example is the Manchester carry chain[4a,b], used for propagating carry signals in parallel adders. In each stage of the adder a carry propagate signal is derived from the two input variables to the adder, and if it is desired to propagate the carry, this propagate signal is applied to the gate of an enhancement mode pass transistor. The source of the transistor is carry-in to the present stage, and the drain of the transistor is carry-out to the next stage. In this way, a carry can be propagated from less to more significant stages of the adder without inserting a full inverter delay between stages. The circuit is shown schematically in Fig. 11a.

The delay through such a circuit does not involve inverter delays but is of an entirely different sort. A voltage along the chain divides into V_{ds} across each pass transistor. Thus V_{ds} is usually low, and the pass transistors operate primarily in the resistive region. We can think of each transistor as a series resistance in the carry path, and a capacitance to ground formed by the gate to channel capacitance of each transistor, and the strays associated with the source, drain, and connections with the following stage. An abstraction of the electrical representation is shown in Fig. 11b. The minimum value of R is the turned on resistance of each enhancement mode pass transistor, while the minimum value of C is the capacitance from gate to channel of the pass transistor. Strays will increase both values, especially that of C. The response at the node labelled V_2 with respect to time is given in eq. 13. In the limit as the number of sections in the network becomes large, eq. 13 reduces to the differential form shown in eq. 14 where R and C are now the resistance and capacitance per unit length, respectively.

$$C\,dV_2/dt = [(V_1 - V_2) - (V_2 - V_3)]/R \qquad \text{[eq.13]}$$

$$RC\,dV/dt = d^2V/dx^2 \qquad \text{[eq.14]}$$

Equation 14 is the well-known diffusion equation, and while its solutions are complex, in general the time required for a transient to propagate a distance x in such a system is proportional to x^2. One can see qualitatively that this might be so. Doubling the number of sections in such a network doubles both the resistance and the capacitance, and therefore causes the time required for the system to respond to increase by a factor of approximately four. The response of a system of n stages to a step function input is shown in Fig. 11c.

If we add one more pass transistor to such a chain of n pass transistors, the added delay through the chain is small for small n, but very large for large n. Therefore, it is highly desirable to group the pass transistors used for steering, multiplexing, and carry-chain type logic into short sections and interpose inverting logic between these sections. This approach applied to the carry chain is shown in figure 11d. The delay through a section of n pass transistors is proportional to RCn^2. Thus the total delay is ~ RCn^2 plus the delay through the inverter τ_{inv} . The average delay per stage is given in eq. 15. To minimize the delay per stage, chose n such that the delay through n pass transistors equals the inverter delay:

$$\begin{aligned} &\text{Total delay} \quad \sim \quad RCn^2 + \tau_{inv}\,, \\ &\text{Average delay/stage} \quad \sim \quad RCn + \tau_{inv}/n \\ &\text{Min. delay when:} \quad RCn^2 \sim \tau_{inv} \end{aligned} \qquad \text{[eq.15]}$$

Since logic done by steering signals with pass transistors does not require static power dissipation, a generalization of this result may be formulated. It pays to put as much logic into steering type circuits as possible until there are enough pass transistors to delay the signal by approximately one inverting logic delay. At this point, the level of the signal can be restored by an inverting logic stage.

The pass transistor has another important advantage over an inverting logic stage. When used to control or steer a logic signal, the pass transistor has only an input, control, and output connections. A NAND or NOR logic gate implementing the same function, in addition to containing two more transistors and thus occupying more area, also requires VDD and GND connections. As a result, the topology of interconnection of pass transistor circuits is far simpler than that of inverting logic circuits. This topological simplicity of pass transistor control gates is an important factor in the system design concepts developed in later chapters.

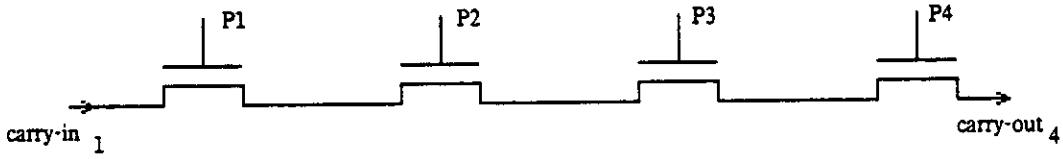


Fig. 11a. Pass Transistor Chain Propagating a Carry Signal

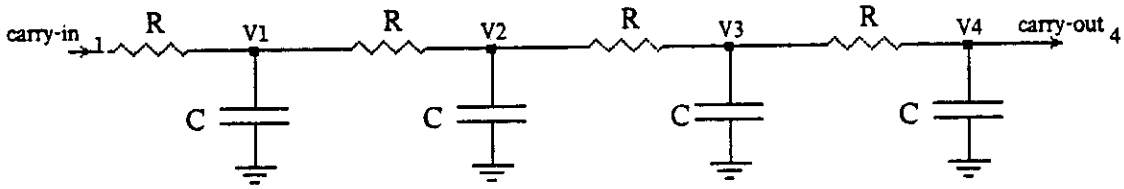


Fig. 11b. Equivalent Circuit

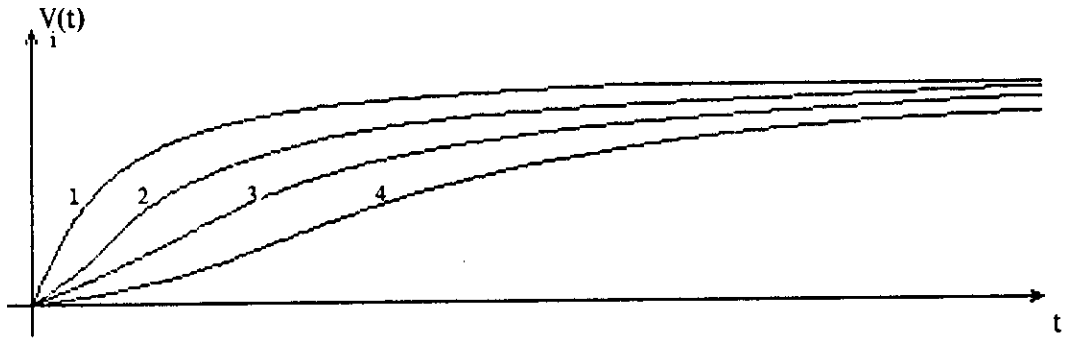


Fig. 11c. Response to Step Function Input

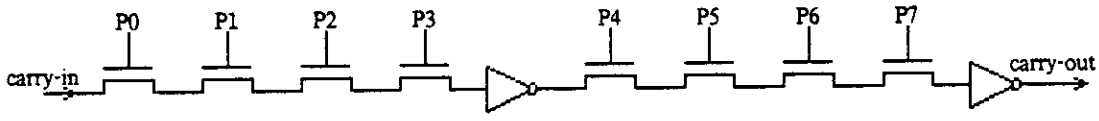


Fig. 11d. Minimizing Delay by Interposing Inverters

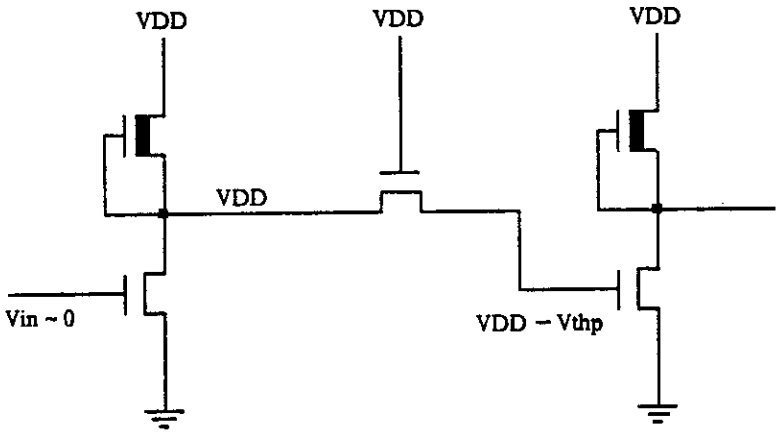


Fig. 12a. Inverters Coupled by Pass Transistor

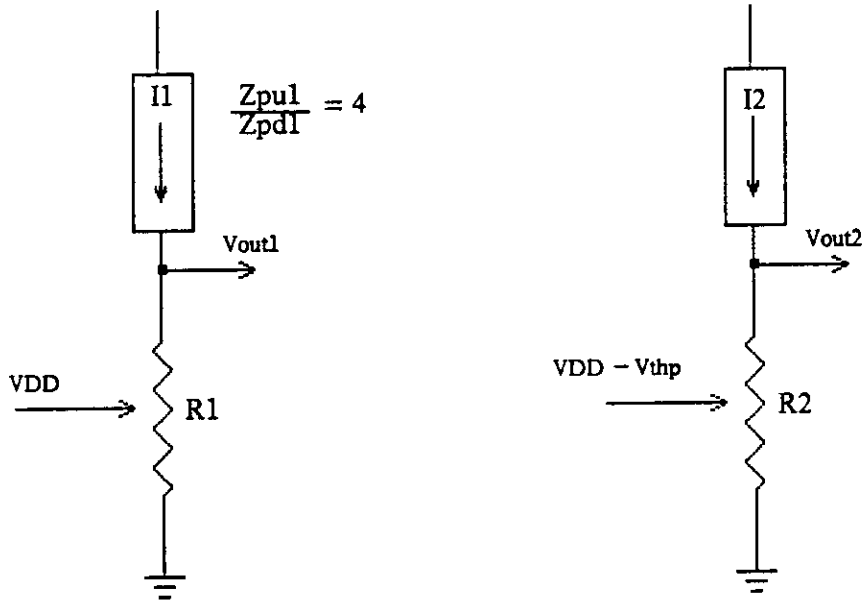


Fig. 12b. For Vout2 = Vout1, Zpu2/Zpd2 = 8

Pullup/Pulldown Ratios for Inverting Logic Coupled by Pass Transistors

Earlier we found that when an inverting logic stage directly drives another such stage, a pullup to pulldown ratio $Z_{pu}/Z_{pd} = (L_{pu}/W_{pu})/(L_{pd}/W_{pd})$ of 4:1 yields equal inverter margins, and also provides an output sufficiently less than V_{th} for an input equal to VDD. Rather than coupling inverting logic stages directly, we often couple them with pass transistors for the reasons developed in the preceding section, thus affecting the required pullup to pulldown ratio.

Figure 12a shows two inverters connected through a pass transistor. If the output of the first inverter nears VDD, the input of the second inverter can rise at most to (VDD - V_{thp}), where V_{thp} is the threshold of the pass transistor. Why does this effect occur? Consider the following: The output of the first inverter is at or above (VDD - V_{thp}), the pass transistor gate is at zero volts, and the input gate of the second inverter is also at zero volts. The pass transistor's gate voltage is now driven quickly to VDD, turning on the pass transistor. As current flows through the pass transistor, from drain to source, the input gate voltage of the second inverter rises and the gate to source voltage of the pass transistor falls. When the gate voltage of the second inverter has risen to (VDD - V_{thp}), the pass transistor's gate to source voltage has fallen to its threshold value, and the pass transistor will switch off.

If the second inverter is to have its output driven as low with an input of (VDD - V_{thp}) as would a standard inverter with an input of VDD, then the second inverter must have a pullup to pulldown ratio larger than 4:1. This larger ratio is calculated as follows: With inputs near VDD, the pullups of inverters are in saturation, and the pulldowns are in the resistive region. Figure 12b shows equivalent circuits for two inverters. VDD is input to one, and (VDD - V_{thp}) to the other. For the output voltages of the two inverters to be equal under these conditions, I_1R_1 must equal I_2R_2. Referring to equations 3a and 5, we find:

$$(Z_{pu1}/Z_{pd1})(VDD - V_{th}) = (Z_{pu2}/Z_{pd2})(VDD - V_{th} - V_{thp})$$

Since V_{th} of the pulldowns is approximately 0.2VDD, and V_{thp} of the pass transistor is approximately 0.3VDD due to the body effect, then $Z_{pu2}/Z_{pd2} \sim 2Z_{pu1}/Z_{pd1}$. Thus a ratio of $(L_{pu}/W_{pu})/(L_{pd}/W_{pd}) = 8$ is usually used for inverting logic stages placed as level restorers between sections of pass transistor logic.

Transit Times and Clock Periods

In chapter 3 we will develop a system design methodology in which we will be able to construct and estimate the performance of arbitrarily complex digital systems, using only the basic circuit forms presented in the preceding sections. The basic *system* building block in the design methodology is a register to register transfer through combinational logic, implemented with pass transistors and inverting logic stages. Using the basic ideas already presented, we may anticipate the results of that chapter in order to estimate the maximum clocking frequency of such systems.

The design methodology uses a two-phase non-overlapping clock scheme. During the first clock phase, data passes from one register, through combinational logic stages and pass transistors to a second register. During the second clock phase, data passes from the second register through still more logic and pass transistors to a third (or possibly back to the first) register. The data storage registers are implemented by using charge stored on the input gates of inverting logic stages, the charge being isolated by pass transistors controlled by clock signals, as described in chapter 3.

Since pass transistors are used to connect inverting logic stages, inverter ratios of $k \sim 8$ are required. If the combinational logic between registers is implemented using only pass transistors, and if the delays through the pass transistors have been carefully matched to those of the inverting logic stages, the total delay will be twice that of the simple $k = 8$ inverter. In the absence of strays, the $k = 8$ inverters have a maximum delay (in the case of the output rising towards VDD) of 8τ, and hence a minimum of 16τ must be allowed for the inverter plus logic delay. However, in most designs the stray capacitance is at least equal to that inherent in the circuit. Thus the minimum time required for one such operation is $\sim 30\tau$. Control lines to the combinational logic and pass transistors each typically drive the gates of 10 to 30 transistors. Even when using a super buffer driver, the delay introduced by this fan out is at least the minimum driving time for a capacitive load. With $Y = 30$, this time is $\sim 9\tau$. To this we must add an 8τ inverter delay for operation of the drivers.

Thus the total time for one clock phase is $\sim 50\tau$. Since two clock phases are required per cycle, a minimum clocking period of $\sim 100\tau$ is required for system designed in this way. In 1978, τ ~0.3 nanoseconds, and clocking periods of 30 to 50 ns are achievable in carefully structured integrated systems where successive stages are in close physical proximity. If it is necessary to communicate data over long distances, longer periods are required.

Properties of Cross-Coupled Circuits

In many control sequencing and data storage applications, memory cells and registers are built using two inverters driving each other, as shown in figure 13a. This circuit can be set in either the state where V_1 is high and V_2 is low, or in the state where V_1 is low and V_2 is high. In either case, the condition is stable and will not change to the other condition unless it is forced there through some external means. The detailed methods of setting such cross-coupled circuits into one state or another will be discussed in detail later. However, it is important at the present time to understand the time evolution of signals impressed upon cross-coupled circuits, since they exhibit properties different from circuits not having a feedback path from their output to an input.

We have seen that there exists a voltage at which the output of an inverter is approximately equal to its input voltage. If a cross-coupled circuit is inadvertently placed in a situation where its input voltage is equal to this value, then an unstable equilibrium condition is created where voltages V_1 and V_2 are equal. Since the net current flowing onto either gate is now zero, there is no forcing function driving the system to any voltage other than this equilibrium one, and the circuit can stay in this condition for an indefinite period. However, if either voltage changes, even very slightly, the circuit will leave this unstable equilibrium. For example, if the voltage V_1 is increased from its unstable equilibrium value V_{inv} by a slight amount, this will in time cause a lowering of voltage V_2 , as net current flows from gate 1. This lowering of V_2 will at some later time cause V_1 to increase further. As time goes on, the circuit will feedback on itself until it rests in a stable equilibrium state.

The possibility of such unstable equilibria in cross coupled circuits has important system implications[2], as we will later see. For this reason, we will make a fairly detailed analysis of this circuit's behavior near the metastable state. While it is not essential that the reader follow all the details of the analysis, the final result should be studied carefully. The time constant of the final result depends in detail on the regions of operation of the transistors near the metastable state, as given in the following analysis. However, the exponential form of the result follows simply from the fact that the forcing function pushing the voltage away from the metastable point is proportional to the voltage's distance away from that point. This general behavior is characteristic of bistable storage elements in any technology. However, more complex waveforms are observed in logic families having more than one time constant per stage.

The time evolution of this process can be traced as follows. At the unstable equilibrium, the current in the pullups equals that in the pulldowns, and is some constant, k_1, times $(V_{inv}-V_{th})^2$. If V_1 is then changed by some small ΔV_1 to V_{init}, I_{pu2} remains constant but I_{pd2} changes immediately, producing a non-zero I_{g1}:

$$I_{g1} = I_{pu2} - I_{pd2} = k_1[(V_{inv}-V_{th})^2 - (V_{inv}+\Delta V_1-V_{th}^2)]$$

For small ΔV_1, $I_{g1} = -2k_1(V_{inv}-V_{th})\Delta V_1$. More precisely, since I_{g1} = function(V_1,V_2), then near V_{inv}:

$$\partial I_{g1}/\partial V_1 = -2k_1(V_{inv}-V_{th})$$

Noting that the pullups are not quite in saturation, but are in the resistive region, and:

$$\partial I_{g1}/\partial V_2 = -1/R_{pu} ,$$

where R_{pu} = effective resistance of the pullup near V_{inv}. Noting that $I_{g1} = C_g dV_{g2}/dt$, we find that:

$$dI_{g1}/dt = -2k_1(V_{inv}-V_{th})[dV_1/dt] - (1/R_{pu})[dV_2/dt] = C_g[d^2V_2/dt^2]$$

Evaluating the constants in this equation yields $-k_1(V_{inv}-V_{th}) = C_g/\tau_o$, where τ_o is the saturation transit time of the pulldowns for t near zero. Assume a pullup/pulldown Z ratio of 4:1, and consider the operating conditions near t = 0. Evaluating the effective resistance of the pullups in terms of the parameters of the pulldowns yields $1/R_{pu} \sim C_g/\tau_o$.

Therefore: $$-(2/\tau_o)[dV_1/dt] -(1/\tau_o)[dV_2/dt] = d^2V_2/dt^2$$

Similarly: $$-(2/\tau_o)[dV_2/dt] -(1/\tau_o)[dV_1/dt] = d^2V_1/dt^2$$

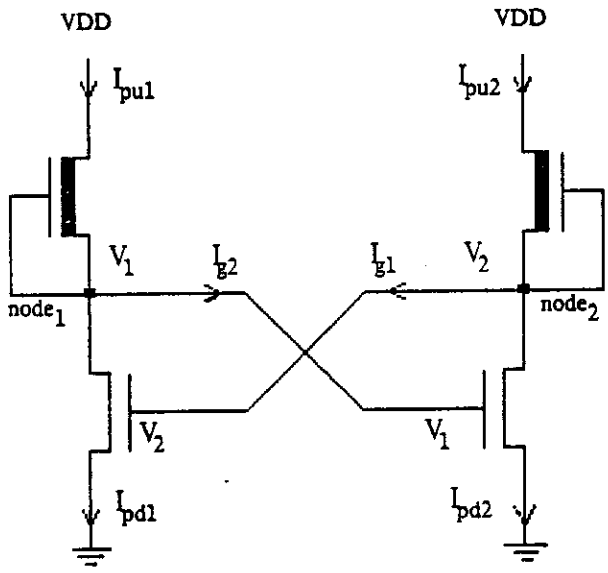


Fig. 13a. Cross Coupled Inverters

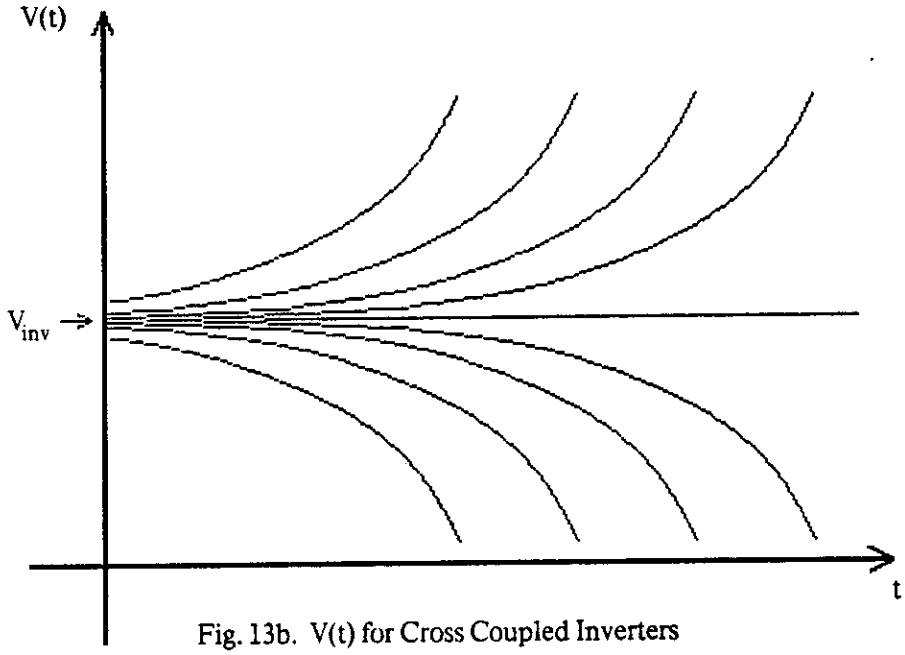


Fig. 13b. V(t) for Cross Coupled Inverters

(mos13ab.press)

Near time $t = 0$, dV_1/dt approximately equals $- dV_2/dt$, and therefore:

$$d^2V_1/dt^2 = -(1/\tau_o)\, dV_2/dt = (1/\tau_o)^2\, V_1 + \text{const.} \qquad \text{[eq.16a]}$$

The solution to eq. 16a is an exponential diverging from the equilibrium voltage V_{inv}, with a time constant $\tau_o/2$ equal to 1/2 the pulldown delay time. Note that the solution given in eq.16b satisfies the conditions that $V(0) = V_{init}$, and that V(t) is constant, if $V_{init} = V_{inv}$:

$$V_1(t) = V_{inv} + (V_{init} - V_{inv})\, e^{t/\tau_o} \qquad \text{[eq.16b]}$$

The above analysis applies to cross coupled circuits in the absence of noise. Noise unavoidably present in the circuit spreads the input voltage into a band from which such an unstable equilibrium can statistically be initiated. The width of this band is equal to the noise amplitude. Any timing condition which causes the input voltage to settle in this band has some probability of causing a balanced condition, from which the circuit may require an arbitrarly long time to recover. The time evolution of such a system is shown in Fig. 13b, for several initial voltages near V_{inv}. The time for the cross-coupled system to reach one of its equilibria is thus logarithmic in the displacement from V_{inv}, and is given approximately by eq. 16c:

$$t \sim \tau_o \ln[V_{inv}/(V_{init}-V_{inv})] \qquad \text{[eq.16c]}$$

A Fluid Model for Visualizing MOS Transistor Behavior

[Section Contributed by Carlo H. Sequin, U. C. Berkeley]

When designing circuits and systems, it is often useful to have some method for visualizing the physical behavior of the devices used as basic building blocks. This section develops such a method for the MOS transistor. Some readers of this text may be unfamiliar with semiconductor device physics, and would have difficulty visualizing what is going on inside an active semiconductor device, if device behavior were described in purely analytical terms. However, it is possible to construct a simple but very effective model of the behavior of certain charge controlled devices, such as MOS transistors, charge coupled devices (CCD's), and bucket brigade devices (BBD's)[8], without referring to the details of device physics.

This model will be developed using two basic ideas: We think of electrical charge as though it were a fluid, and we mentally map the relevant electrical potentials into the geometry of a "container" in which the charge is free to move around. One can then apply one's intuitive understanding of, say, water in buckets of various shapes towards a visualization of what is going on inside the devices. Often a design guided by a good intuitive understanding of how a fluid would behave in the designed structure may show superior performance over designs based on complicated but possibly inadequate two-dimensional analytical modelling.

The MOS Capacitor

The basic element of MOS transistors or charge transfer devices is the MOS capacitor. The notions of a fluid model will first be introduced using this elementary building block.

In physical space an MOS capacitor is a sandwich structure of a metal or polysilicon electrode on a thin insulator on the surface of a silicon crystal (fig. fm-1a). A suitable voltage applied to the electrode, i.e. positive for a p-type silicon substrate as used in nMOS, will repel the majority carriers in the substrate under the electrode, generating a depletion region which is at first free of any mobile charge carriers. Minority carriers, in this case electrons, can be injected electrically into this area, or generated by incident light, and subsequently stored underneath the MOS electrode. Applying the notions of a fluid model, the same situation can be described as follows:

The positive voltage applied to the MOS electrode generates a pocket in the surface potential of the silicon substrate. This can be visualized as a container, where the shape of the container is

Fig. fm1a:

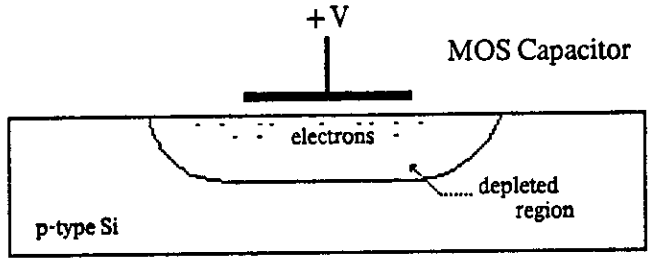


Fig. fm1b:

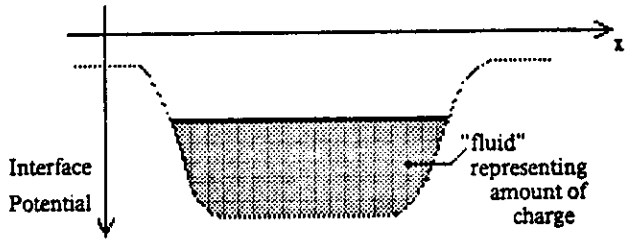


Fig. fm2a:

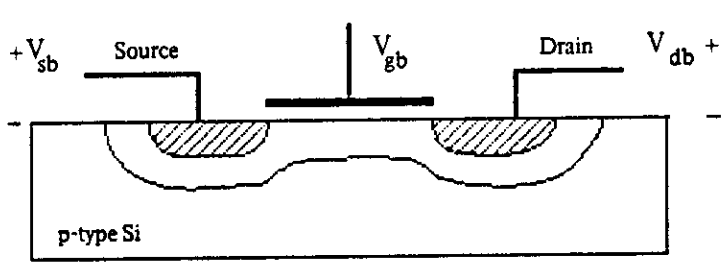
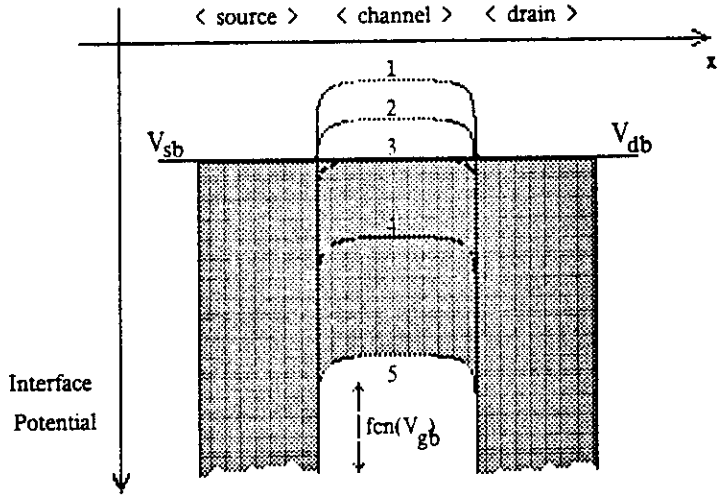


Fig. fm2b:



defined by the potential along the silicon surface, as plotted by the dashed line in figure fm-1b. Note that in fig. fm-1b, *increasing positive potential* is plotted in the *downward direction.* The presence of minority charge carriers in an inversion layer changes the surface potential: an increase in this charge decreases the positive surface potential under the MOS electrode. The potential profile in the presence of inversion charge is indicated by the solid line in fig. fm-1b. The area between the dashed and solid lines in fig. fm-1b is hatched to indicate the presence of this charge. This representation shows charge sitting at the bottom of the container, just as a fluid would reside in a bucket. Of course the surface of the fluid (solid line) must be level in an equilibrium condition; if it were not, electrons would move under the influence of the potential difference until a constant surface potential has been established.

This model allows one to visualize easily the amount of charge present (hatched area), the fact that the charge tends to sit in the deepest part of the potential well, and the fact that the capacity of the bucket is finite and dependent upon the applied electrode voltage. The higher this voltage, the deeper the bottom of the bucket and the more charge that can be stored. It should be kept in mind that this fluid model differs from the physical reality in so far as in reality the minority carriers in the inversion layer reside directly at the silicon surface.

The MOS Transistor

The same kind of model can be used to describe MOS transistor behavior. Figure fm-2a shows the physical cross section through an MOS transistor. Source and drain diffusions have been added to the simple MOS capacitor. For the moment we consider these two diffusions to be connected to two identical voltage sources, $V_{sb} = V_{db}$, which thus define the potential of the source and drain regions.

In the potential plot these diffusions are represented by exceedingly deep buckets, filled with charge carriers up to the levels of the source and drain region potentials. Whether the MOS transistor is conducting, or is isolating the two diffused regions from one another, now depends on the potential underneath the MOS gate electrode. If the applied gate potential is chosen so that the potential underneath is less than V_{sb}, then there exists a potential barrier between source and drain regions (case 1 and 2 in fig. fm-2b). However, if the potential of an "empty bucket" under the gate electrode would be higher than V_{sb}, then the transistor is turned on (case 4 and 5). Of course, in cases 4 and 5, carriers from the source and drain regions will spill underneath the gate electrode so that a uniform surface potential exists throughout the whole transistor. The

conductivity of the channel area depends on the thickness of the inversion layer, which can readily be visualized in fig. fm-2b. Channel conductivity goes to zero at the turn-on threshold of the transistor (case 3), when the "empty bucket" potential under the gate electrode is equal to the source and drain potential. Thus, the region under the gate can be viewed as a movable barrier of variable height which controls the flow of charge between the source and drain areas.

The same model enables us to visualize what happens when source and drain regions are biased to different potentials, as is usually the case in normal operation of MOS transistors. Figure fm-3a again shows a physical cross section through an MOS transistor, as a reference for the following figures. Figure fm-3b reviews the case of equal source and drain potentials with the channel turned on fairly strongly, thus readily allowing charge to move between source and drain. Figure fm-3c shows the situation when a small voltage difference, ΔV, has been applied between source and drain. Since the potential difference is maintained by external voltage sources, electrons will be forced to move from source to drain under the influence of the potential gradient, just as a liquid would flow from the higher to the lower level.

As the potential difference between source and drain is made larger, the variation in the "depth" of the fluid along the channel becomes significant (fig. fm-3d). Continuity in the fluid requires that the charge move faster in the areas where the layer is thinner. This implies that the potential increases more rapidly closer to the drain region. With increasing drain potential the amount of charge flowing from source to drain per unit time increases, since the product of charge layer depth and local gradient increases. However there is a limit. Once the drain potential exceeds the empty channel potential the rate of charge flow will be limited by the drain-side edge of the barrier under the gate electrode. The MOS transistor has now reached saturation (fig. fm-3e). The drain current density now is determined by the potential difference between the source and the empty channel and by the length of the channel (or the width of the barrier over which the charge has to flow), and is to first order independent of the drain voltage V_{db}.

Even in simple transitor circuits the above fluid model helps one quickly develop a feeling for device and circuit operation. However, the real power of this intuitive model emerges when it is applied to complex structures where closed form solutions describing charge motion can no longer be found. The empty potential under the various electrodes can first be plotted as in the above examples, and the flow of charge then visualized using the analogy to the behavior of a fluid.

Fig. fm3a:

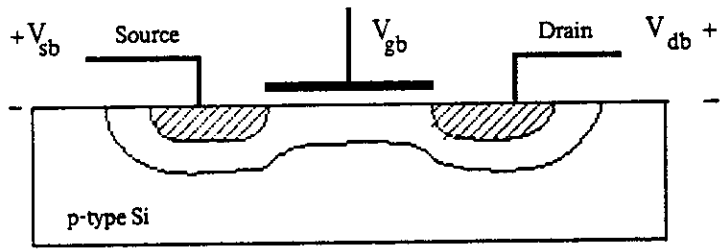


Fig. fm3b:

Vsb = Vdb

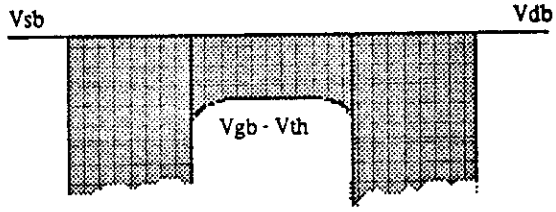


Fig. fm3c:

Vdb = Vsb + $\triangle$V

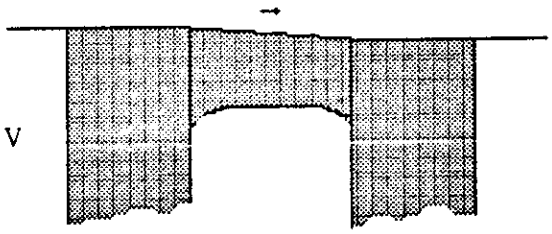


Fig. fm3d:

Vdb > Vsb

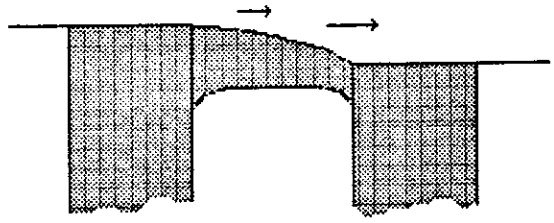
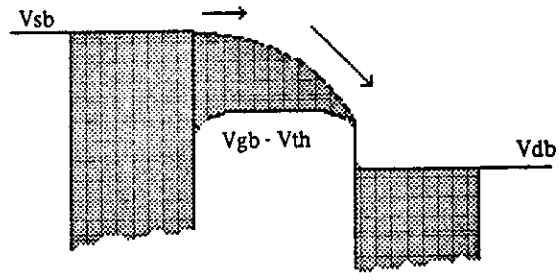


Fig. fm3e:

Vdb > Vgb - Vth



Effects of Scaling Down the Dimensions of MOS Circuits and Systems

This section examines the effects on major system parameters resulting from scaling down all dimensions of an integrated system, including those vertical to the surface, by dividing them by a constant factor α. The voltage is likewise scaled down by dividing by the same constant factor α. Using this convention, all electric fields in the circuit will remain constant. Thus many non-linear factors affecting performance will not change as they would if a more complex scaling were used.

Figure 14a. shows a MOSFET of dimensions L, W, D, with a (Vgs - Vth) = V. Figure 14b. shows a MOSFET similar to that in figure 14a., but of dimensions $L' = L/\alpha$, $W' = W/\alpha$, $D' = D/\alpha$, and $V' = V/\alpha$. Refer to equations 1., 2., and 3. From these equations we will find that as the scale down factor α is increased, the transit time, the gate capacitance, and drain to source current of every individual transistor in the system scale down proportionally, as follows:

$$\tau \propto L^2/V, \quad \tau'/\tau = [(L/\alpha)^2/(V/\alpha)]/[L^2/V], \text{ therefore, } \tau' = \tau/\alpha$$

$$C \propto LW/D, \quad C'/C = (L/\alpha)(W/\alpha)/(D/\alpha)]/[LW/D], \quad \text{and} \quad C' = C/\alpha$$

$$I \propto WV^2/LD, \quad I'/I = [(WV^2/\alpha^3)/(LD/\alpha^2)]/[WV^2/LD], \quad \text{and} \quad I' = I/\alpha$$

Switching power, P_{sw}, is the energy stored on the capacitance of a given device divided by the *clock period*, or time between successive charging and discharging of the capacitance. A system's clock period is proportional to the τ of its smallest devices. As devices are made smaller and faster, the clock period is proportionally shortened. Also, the dc power, P_{dc}, dissipated by any static circuit equals I times V. Therefore, P_{sw} and P_{dc} scale as follows:

$$P_{sw} \propto CV^2/\tau \propto WV^3/DL, \quad \text{and} \quad P_{sw}' = P_{sw}/\alpha^2$$

$$P_{dc} = IV, \quad \text{and} \quad P_{dc}' = P_{dc}/\alpha^2$$

Both the switching power and static power per device scale down as $1/\alpha^2$. The average dc power for most systems can be approximated by adding the total P_{sw} to one-half of the dc power which would result if all level restoring logic pulldowns were turned on. The contribution of pass transistor logic to the average dc power drawn by the system is due to the switching power consumed by the driving circuits which charge and discharge the pass transistor control gates.

The switching energy per device, E_{sw} , is an important metric of device performance. It is equal to the power consumed by the device at maximum clock frequency multiplied by the device delay, and scales down as follows:

$$E_{sw} \propto CV^2 , \qquad \text{and} \quad E_{sw}' = E_{sw}/\alpha^3$$

The following table summarizes values of the important system parameters for current technology, and for a future technology near the limits imposed by physical law:

	1978	19XX
Minimum Feature Size:	6 μm	0.3 μm
τ :	0.3 to 1 ns	~0.02 ns
Esw:	$\sim 10^{-12}$J	$\sim 2 \times 10^{-16}$J
System Clock Period: [see earlier section]	~30 to 50 ns	~2 to 4 ns

A more detailed plot of the channel conductance of an MOS transistor near the threshold voltage is shown in figure 15. Below the nominal threshold, the conductance (1/R) is not in reality zero, but depends on gate voltage and temperature as follows:

$$1/R \propto e^{(V_{gs}-V_{th})/(kT/q)},$$

where T is the absolute temperature, q is the charge on the electron, and k is Boltzmann's constant. At room temperature, kT/q ~ 0.025 volts. At present threshold voltages, as in the

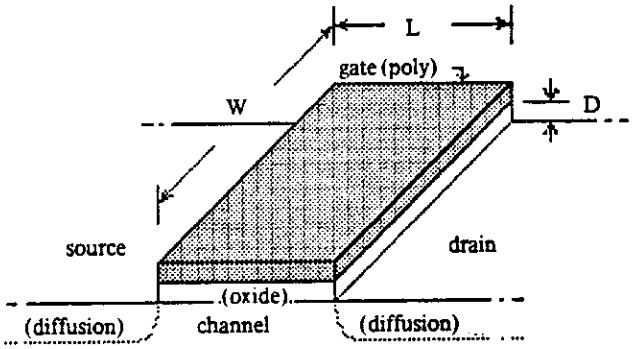


Fig. 14a. MOSFET, 1978

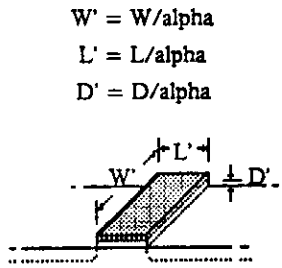


Fig. 14b. MOSFET Scaled Down by Alpha, 19XX

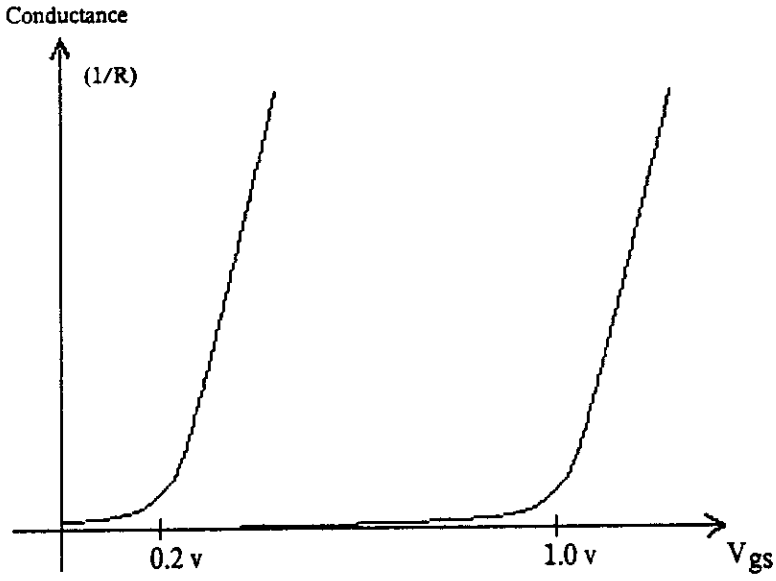


Fig. 15. Conductance as a Function of Threshold Voltage

rightmost curve in figure 15., an off device is below threshold by perhaps 20 kT/q, i.e. by about 0.5 volts, and its conductance is decreased by a factor of the order of ten million. Said another way, if the device is used as a pass transistor, a quantity of charge which takes a time τ to pass through the on device, will take a time on the order of $10^7\tau$ to "leak" through the off device.

The use of pass transistors switches to isolate and "dynamically store" charge on circuit nodes is common in many memory applications using 1978 transistor dimensions. However, if the threshold voltage is scaled down by a factor of perhaps 5, as shown in the leftmost curve in figure 15., then an off transistor is only 4kT/q below threshold. Therefore, its conductance when "off" is only a factor of 100 or so less than when it is "on". For such relatively large values of subthreshold conductance, charge stored dynamically on a circuit node by the transistor when "on" will safely remain on that node for only a few system clock periods. The charge will not remain on the node for a very large number of periods as it does in present memory devices using this technique. One way of possibly coping with this problem, as device dimensions and threshold voltages are scaled down, is to reduce the temperature of device operation[7].

Suppose we scale down an entire integrated system by a scale down factor of $\alpha = 10$. The resulting system will have one hundred times the number of circuits per unit area. The total power per unit area remains constant. All voltages in the system are reduced by the factor of 10. The current per unit area is increased by a factor of 10. The time delay per stage is decreased by a factor of 10. Therefore, the power-delay product decreases by a factor of 1000.

This is a rather attractive scaling in all ways except for the current density. The delivery of the required average dc current presents an important obstacle to scaling. This current must be carried to the various circuits in the system on metal conductors, in order that the voltage drop from the off-chip source to the on-chip subsystems will not be excessive. Metal paths have an upper current density limit imposed by a phenomenon called metal migration, discussed further in chapter 2. Many metal paths in today's integrated circuits are already operated near their current density limit. As the above type of scaling is applied to a system, the conductors get narrower, but still deliver the same current on the average to the circuits supplied by them.

Therefore, it will be necessary to find ways of decreasing system current requirements to approximately a constant current per unit area relative to the present current densities. In n-channel silicon gate technology, this objective can be partially achieved by using pass transistor

logic in as many places as possible and avoiding restoring logic except where it is absolutely necessary. Numerous examples of this sort of design are given later in this text. This design approach also has the advantages of tending to minimize delay per unit function and to maximize logic functions per unit area. However, when scaled down to submicron size, the pass transistors will suffer from the subthreshold current problem. It is possible that when the fabrication technologies have been developed to enable scaling down to sub-micron devices, a technology such as complementary MOS, which does not draw any dc current, may be preferable to the nMOS technology used to illustrate this text. However, even if this occurs, the methodology developed in the text can still be applied in the design of integrated systems in that technology.

The limit to the kind of scaling described above occurs when the devices created are no longer able to perform the switching function. To perform the switching function, the ratio of transistor on to off conductance must be $>>1$, and therefore the voltage operating the circuit must be many times kT/q. For this reason, even circuits optimized for operation at the lowest possible supply voltages still require a VDD of ~ 0.5 volts. Devices in 1978 operate with a VDD of approximately five volts and minimum channel lengths of approximately six microns. Therefore, the kind of scaling we have envisioned here will take us to devices with approximately one half micron channel lengths and current densities approximately ten times what they are today. Power per unit area will remain constant over that range. Smaller devices might be built but must be used without lowering the voltage any further. Consequently the power per unit area will increase. Finally, there appears to be a fundamental limit[3] of approximately one quarter micron channel length, where certain physical effects such as the tunneling through the gate oxide, and fluctuations in the positions of impurities in the depletion layers, begin to make the devices of smaller dimension unworkable.

References

1. W. M. Penney, L. Lau, Eds., "MOS Integrated Circuits", Van Nostrand, 1972, pp.60-85.

2. T. J. Chaney, C .E. Molnar, "Anomalous Behavior of Synchronizer and Arbiter Circuits", *IEEE Transactions of Computers*, April 1973, pp. 421-422.

3. B. Hoeneisen, C. A. Mead, "Fundamental Limitations in Micro-electronics--I. MOS Technology", *Solid-State Electronics*, Vol.15, 1972, pp. 819-829.

4a. T. Kilburn, D. B. G. Edwards, D. Aspinall, "A Parallel Arithmetic Unit Using a Saturated Transistor Fast-Carry Circuit", Proc. IEE, Pt. B, vol. 107, pp.573-584, Nov. 1960.

4b. Staff of the Computation Lab, "Description of a Relay Calculator", Annals of the Harvard Computation Lab, vol. 24, Harvard University Press, 1949.

5. T. K. Young, R. W. Dutton, "MINI-MSINC - A Minicomputer Simulator for MOS Circuits with Modular Built-in Model", Stanford Electronics Laboratories, Technical Report No. 5013-1, March 1976.

6. L. Nagel, D. Pederson,"Simulation Program with Integrated Circuit Emphasis (SPICE)", 16th Midwest Symposium on Circuit Theory, Waterloo, Ontario, Apr. 12, 1973.

7. F. H. Gaensslen, V. L. Rideout, E. J. Walker, J. J. Walker, "Very Small MOSFETs for Low-Temperature Operation", IEEE Transactions on Electron Devices, March 1977.

8. C. H. Sequin, M. F. Tompsett, "Charge Transfer Devices", Academic Press, 1975.

Reading References

R1. J. F. Gibbons, "Semiconductor Electronics", McGraw-Hill, 1966, is a classic text containing a good introduction to basic semiconductor theory.

R2. W.M. Penney, L. Lau, Eds., "MOS Integrated Circuits", Van Nostrand Reinhold,1972, is a good, general text on MOS devices and circuits.

R3. R. W. Keyes, "Physical Limits in Digital Electronics", *Proceedings of the IEEE*, Vol. 63, No. 5, May 1975, pp. 740-767, is an excellent invited survey paper on this topic.

R4. P. Richman, "MOS Field-Effect Transistors and Integrated Circuits", Wiley, 1973, provides an excellent discussion of the physics and fabrication of MOS devices.